Artificial Intelligence in Customs Risk Management for e-Commerce

Design of a Web-crawling Architecture for the Dutch Customs Administration

A design science research approach

Alessandro Giordani

< confidential >

TU Delft

IBM

Artificial Intelligence in Customs Risk Management for e-Commerce

Design of a Web-crawling Architecture for the Dutch Customs Administration

By

Alessandro Giordani

in partial fulfilment of the requirements for the degree of

Master of Science
in **Management of Technology**

at Delft University of Technology
Faculty of Technology, Policy and Management

to be defended publicly on Tuesday August 14, 2018 at 10:00 AM.

Graduation Committee

Chairperson:	Prof. Dr. Y. Tan, Section ICT (TPM)
First Supervisor:	Dr. S. W. Cunningham, Section Policy Analysis (TPM)
Second Supervisor:	Dr. M. Y. Maknoon, Section Transports and Logistics (TPM)
External Supervisor:	Dr. B. D. Rukanova, Section ICT (TPM)
External Supervisor:	Ben van Rijnsoever, International Business Machines (GBS)

Student number: 4634330

This thesis has been done in collaboration with International Business Machines Corporation (IBM) Netherlands, and the Dutch Customs Administration (Belastingdienst).

This thesis is confidential and cannot be made public until August 31, 2019.

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

IBM

Abstract

The last decade saw the rise of e-commerce trade and the shift of the manufacturing industry to the emerging economies, China first of all. In this context, the European Customs Authorities experienced an explosion of small parcels coming from e-commerce websites, often from China, and faced difficulties to detect fiscal frauds and security threats using their conventional risk management systems. To address this problem, the European project PROFILE brings together the customs administrations of Netherlands, Belgium, Sweden, Norway, and Estonia, aiming to provide the EU with a shared platform for: (1) accurately assessing customs risks; (2) optimizing operation and logistics by integrating multiple sources of information; (3) developing a shared data platform to share customs risk management (CRM) practices.

As part of this project, the Dutch Customs Administration (DCA) and International Business Machines (IBM) Corporation are collaborating to deploy the cutting-edge technologies of artificial intelligence to automatically cross-check the customs declarations coming from Chinese e-commerce against online information. Through a Design Science approach, I carried out this research for the Delft University of Technology, written in collaboration with IBM Netherlands, aiming to deliver a preparatory study for the developing team before the PROFILE project begins. This includes knowledge brokering between the Dutch Customs Administration and IBM Netherlands so that a more precise problem scope can be defined, and the requirements elicited. In particular, this research focuses on the first part of the project: the development of an adaptive web-crawler for e-commerce, able to compare the declarations documents against online information.

According to the Dutch Customs Administration, the web-crawling system should gather the description of the goods from declarations, search the product on the web, find its price of sale on the e-commerce platforms, compare it with the value declared in the declaration, and return a risk indicator of green/red flag to the targeting officer. The design process of this system follows approaches coming from the systems engineering discipline, starting with the requirement analysis, addressing them with the state-of-the-art big data analytics, and finally deriving the logical components of the system, whose design is presented through a logical architecture.

First, the application domain is investigated. When goods entry the Netherlands need an entry declaration. These goods arrive at the harbor of Rotterdam or airport of Schiphol, where some of these are imported into the country and become import/export, and others stop temporarily as transit waiting to be shipped somewhere else. The Dutch Customs Administration monitors these processes through risk management systems aiming to stop non-compliant goods. This research describes these practices, with a higher focus on the e-commerce risk targeting. About the e-commerce world, a study of the e-commerce processes behind an online purchase is also carried out through a real purchase on Chinese e-commerce. This was used to observe how the Chinese sender described the item, and how the Dutch Customs assessed the risk and decided on the duties to be paid. This led to reflect on the possible frauds scenarios and how to address them. Finally, the Dutch Customs also reported that the products descriptions are often vague and ambiguous, and a more accurate formulation of the problem is described.

Secondly, an in-depth literature on the fields of web-crawling and big data analytics techniques is carried out. The possible technologies that could be useful to address the requirements and the problem formulation are investigated. Starting with an analysis of the existing literature on the field of big data analytics, this research also covers the recent trends of machine learning and artificial intelligence. To avoid reporting a too big literature, the topics reported have been accurately chosen, for instance describing only the techniques for web analytics and text analytics.

This literature on big data analytics is further broken in two sub-topics, one more theoretical, which classifies the types of analytics methods and defines the technology of machine learning and natural language processing, including the last paradigms of deep learning and reinforcement learning, and one more practical, where guidelines for the design, development, and implementation of machine learning techniques are proposed. It is here that a theoretical framework to systematically reflect on the challenges of the field of big data analytics has been identified. This framework is then used to systematically collect the main technological challenges of the use case under analysis and translate them into non-functional requirements.

Finally, the last part of the literature describes what a web-crawler is and what web-crawling/web-craping means. This later extends to the concepts of focused web-crawling and smart, intelligent, adaptive web-crawling, where machine learning techniques are deployed to improve performance. The literature concludes by providing related works of machine learning techniques implemented in smart web-crawling of the e-commerce websites and stating the knowledge gap that needs to be bridged to address the use case under analysis.

After the application domain and the literature review, the knowledge from these previous phases combines in a continuous iterative process according to the design science methodology (Hevner, 2014). Through unstructured interviews with the DCA and IBM experts, the requirements elicitation is carried out. The approach by Armstrong and Sage (2000) deriving from the field of systems engineering is used. The main objective of the system to be developed is broken down into a series of sub-activities that must be carefully structured to formulate the requirements. About the non-functional requirements, instead of reflecting on the different domains – technological, environment, law compliance, etc. – as it is proposed by the same systems engineering approach mentioned earlier, this research uses the framework identified in the literature review about the main challenges of big data project (Sivarajah, 2016).

To derive the components of the architecture from the requirements and customer needs, the methodology proposed by Suh (1998) called Axiomatic Design has been used, mapping the requirements into architectural components in a rigorous manner. In this way, the design domains proposed by this methodology – customer, functional, physical and process domains – are taken as the reference point for the design process: first, the business needs are identified, then these are translated into requirements, which are mapped into design features. The process domain is left out of this research and will be addressed by the IBM development team in Ireland.

The design cycle leads to the design of a web-crawling system represented through a service-oriented architecture (SOA). Its block diagram and black-box description of each application service are provided. Furthermore, the architecture functionality is described with an architecture walk-through and a sequence diagram in the unified modeling language (UML). The result is an innovative real-time web-crawling system to identify the value of a given product on the e-commerce websites. It deploys natural language process models to filter the non-relevant search results, and other machine learning models to best matching the remaining relevant results with a given item description.

The design and architecture description of this innovative web-crawling system is the main artifact of this research, while the mixed methodology of systems engineering methodologies and big data frameworks is another important scientific contribution.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CBRA	Cross-border Research Association
BCA	Belgium Customs Administration
DCA	Dutch Customs Administration
NCA	Norwegian Customs Administration
SCA	Swedish Customs Administration
ECA	Estonian Customs Administration
TNO	Netherlands Organization for Applied Scientific Research
TUD	Delft University of Technology
IBM	International Business Machines
JRC	Joint Research Centre
ENS	Entry Summary Declaration
ICS	Import Control System
DMF	Douane ManiFest
SAD	Single Administrative Document
DMS	Declaration Management System
AGS	Aangifte Systeem

1 Introduction

This research is carried out for the Delft University of Technology from March to August 2018 and has been written in collaboration with the Department of Global Business Service (GBS) at the International Business Machines (IBM) in the Netherlands. It is a preparatory study for the part which will be carried in the Netherlands of the European project PROFILE starting in August 2018 and aiming to improve the Customs Risk Management (CRM) among five European customs authorities.

Working closely with the department of IBM Global Business Service, the researcher acts as a knowledge broker between the experts at IBM and at the Dutch Customs Administration (DCA) to deliver a requirements analysis and architecture design of a web-crawling system. Thus, the aim of this research is thus to shape the scope of the project with the Dutch Customs to define precise requirements of the web-crawling system. This is furthered by the high-level design of the system architecture so that the IBM developers can take this preparatory analysis and immediately start the development of the system.

This chapter provides an overview of this research, including its main objective and scientific-social relevance. It introduces the concept of customs risk management and explains how customs agencies could deploy artificial intelligence to address problems related to the rise of the e-Commerce trade. This chapter thus introduces the main research question and sub-questions that leads this research. Finally, the research approach is explained in detail, and a research strategy for each research question is described. In the last section of this chapter, the thesis structure is presented with a short description of each chapter of this research.

1.1 Problem Statement

With the rising of the e-commerce trades among countries, the customs authorities are experiencing an increasing number of parcels to inspect and less consolidated shipments (Delfmann, Albers, Gehring, 2002). This is because the e-commerce business model eliminates the intermediate steps of agglomeration (disintermediation) and deliver goods more directly to the end-users. This leads to an increase of smaller parcels shipped to different non-standard destinations, and to an explosion of customs declarations to inspect.

In addition, small parcels often benefit from tax and duties exemptions, because when the customs law was idealized small parcels were seen as non-business exchanges among privates. This leads to a competitive disadvantage for traditional enterprises which are subjected to higher taxes and pushes businesses to prefer smaller parcels to the conventional consolidated shipments. As a result, the valuation of e-commerce shipments has become a major challenge and customs authorities experienced an increasing number of fiscal frauds.

Finally, in such a chaotic environment, customs also have more difficulties in detecting counterfeit articles, fiscal contraband or illegal dangerous products. In particular, the European Union is forecasting about 15% increase in customs declarations over the next five years. In this scenario, manual cross-checking is no longer possible because of the massive quantity of declarations, and the pricing databases currently available to the EU customs are almost useless for cross-border e-commerce characterized by a huge product diversity, large number of online sellers, and fast-changing prices.

For this reason, the EU plans to implement the latest data analytics (DA) technologies to innovate its customs risk management (CRM) practices – defined as "the systematic identification of risk, including through random checks, and the implementation of all measures necessary for limiting exposure to risk.16" (UCC, art.5, 2017).

CRM can be measured through the hit-rate effectiveness, i.e. ordering inspections that were actually to be executed. It includes both false positive – inspections executed but resulting in legal shipments – and false negative – missed inspections to illegal shipments. The idea is to improve the hit-rate effectiveness of CRM practices among Europe using the latest cutting-edge artificial intelligence technologies.

1.2 The PROFILE Project

Five European customs of the countries the Netherlands, Sweden, Norway, Belgium, and Estonia are involved in the European project PROFILE which aims to improve the data sources, analytics and common architecture of the European customs; design more effective indicators to assess the risk related to the new parcels environment; improve the operations, supply chain and logistics within the European Union.

The ultimate objective is to more effectively and efficiently monitor the EU export and import – which adds up to a trade value of 3.5 trillion euros in 2015 (European Commission, 2018) – and to lower the risk of illegal goods to reach the EU citizens. The PROFILE stakeholders are organized in several Living Labs (LL) across Europe in charge of specific working packages (WP).

Living labs are research concept of an ecosystem that is primarily user-centered and open-innovation and that is regularly operating in a territorial context (e.g. region, agglomeration, city), integrating innovation processes and concurrent research within a public-private-people partnership. Based on Living Labs common European innovation system is officially supported by the European Union by stimulating projects to coordinate, accelerate and promote it (Dutilleul et al., 2010).

Table 1: PROFILE Living Labs

LL	**Title**	**Leader**
LL I	Dutch Living Lab	DCA
LL II	Belgian Living Lab	BCA
LL III	Sweden-Norway Living Lab	SCA
LL IV	PROFILE Risk Data Sharing Architecture Living Lab	JRC

Table 2: PROFILE Working Packages

WP	**Main Task**	**Responsible**
1	Project Management and Coordination	CBRA
2	Technical Support and Exploration	TNO
3	Postal Parcels Targeting 1. Smart Web-crawling System	DCA (LL I)

	2. Machine Learning for Products Historical Information	
4	Containers Targeting - Machine Learning for Traders Historical Information	BCA (LL II)
5	System for Automatic Exchange of Declarations Information	SCA (LL III)
6	Common Data Sharing Architecture	JRC (LL IV)
7	Data Governance Policies	TUD
8	Dissemination, Education, and Exploitation	CBRA

As reported in the table above, the Dutch Customs Administration and IBM Netherlands are in charge of the working package 3 and aim first, to develop a smart web-crawling system for able to compare the declarations documents against online information on e-commerce websites; and second, a machine learning model able to recognize a fraudulent parcel coming from e-commerce websites on the base of historical data. The Belgium living labs aims to develop a similar technology but for the container trading.

Finally, the Swedish living lab aims to address the problem of finding a policy framework or operational mechanism to share information among customs of different countries. This is particularly critical in the case of the Swedish-Norwegian border, but once developed it can be applied to other European countries such as Belgium and Netherlands. The idea is that each living lab develops the technology that is most useful in their use case, but with the ultimate goal to share the findings with the other participant of the PROFILE consortium.

1.3 The scope of the Research

This research is carried out for the Delft University of Technology and has been written in collaboration with IBM Netherlands, focuses on the Dutch Living Lab (LL1) and the third working package (WP3). The main purpose is targeting e-commerce trade, which means postal parcels targeting. This is slightly different from the classical trade and customs risk management practices: because of the high number of different traders and dynamic environment (traders are fast changing), customs risk management practices cannot rely on the trader profiles for their analysis.

The approach to e-commerce postal parcel targeting that the Dutch Customs together with IBM wants to use can be summarized in two main use cases: the automate cross-checking of customs declaration data against online information, and parcel inspection decision making based on the analysis of historical data.

These are two different technologies which will be developed one after the other. This research only focuses on the first technology to be developed: a web-crawling system which can analyze the customs declarations, retrieve information on the e-commerce websites, compare the information it finds online with the declarations, and decide whether a package should be inspected or not. An important note: the PROFILE project is a research project, which means that its aims are to validate prototypes and verify that the suggested approached and technologies actually can solve the problems of the customs administrations.

After this research, an engineering phase of these technologies would be necessary to operationalize their use in the customs environment. This would address for instance problems of scalability, such as the response times from the e-commerce websites which may be too slow.

Thus, storing the retrieved information into a cache – so that the same lookup is not repeated – and doing off-line crawling to gather most used info would be necessary.

This is an important consideration to keep in mind, otherwise part of the design that will be proposed later might appear not optimal. For instance, the solution will not be an automatic system-to-system tool that can handle a high transaction rate, because making a complete working solution is not the objective of the project. The real purpose of the project is to see whether developing a web-crawling system to cross-validate the price information in e-commerce platforms is feasible and if it improves the customs risk management.

For this reason, given that there are no databases or offline supporting structure, the web-crawling will be done real-time. In this sense, it can be better defined as a look-up to the Web to see if the product described in the declarations exists and what its value is. This is the main difference with a traditional web-crawler that usually stores and indexes web pages for further analysis (see literature review, section 3.2).

An additional restriction to the scope of the research is to consider only the fiscal fraud detection and to leave out the security threats, which are much more difficult to detect. This again comes from the research nature of the PROFILE project, which first wants to consider an easier scope, and then adding more complicated features. Addressing the fiscal frauds means to check whether the price of the product declared in the declaration document is the actual value of the good or not, and so make sure that the traders pay the fair duties and taxes.

Thus, the focus of the research is on the reduction of the false positive, meant as those inspections which were ordered and resulted in conform products. In other words, a false positive is a wrong result of the customs risk management system according which a compliant package is assessed as not-conform good, and thus to be inspected, when in reality it should be considered as "free to go". This research thus leaves out the other – more complicated – category of wrong results by false-negatives: a not-conform product is missed because considered as free to go by the CRM system. Also, much more data is available about the false positive than for the false negative, and this makes the false positive easier to tackle with data analytics technologies.

In addition, the Dutch Customs wants to begin its PROFILE workflow scoping the data analytics techniques to address the trade with China, as it represents the EU's biggest source of imports and its second-biggest export market (The European Commission, 2017). China and Europe trade on average over €1 billion a day, and the number of parcels coming from China is forecasted to increase, especially given the growth of Chinese global e-commerce such as Alibaba. Although the focus on the Dutch-Chinese trade, the technologies to be developed will work only using the English language. This is a pure operative constraint to avoid adding complexity to the problem since the artificial intelligence technologies are fine-tuned to work with the English language.

In summary, this research will scope down to:

- LL1 and WP3 of PROFILE.
- E-commerce postal parcel targeting.
- Dutch-China trade.
- Design guidelines for the adaptive web-crawling technology (only).
- Fiscal fraud detection only (not security threats).
- Focus on the reduction of the false positive only (not false-negatives).
- English language and English websites only

1.4 Research Objective

The objective of this research is to investigate the design and development of an Information System (IS) to improve customs risk management for e-commerce at the Dutch Customs Administration. This objective can be seen also as improving the hit-rate effectiveness of the CRM system at the DCA. However, because of the complexity of the problem, the research objective is focused on *"improving the cross-validation of price information between the declaration and the online information in e-commerce platforms"*, which eventually can improve the customs risk management. This means eventually finding the price deviations between the value declared on the customs documents and what it is reported in the e-commerce websites, and consequently decide whether the package should be inspected or not (red or green flag).

This can be resembled in a practical and academic purpose. The practical result is to provide a preparatory study to the PROFILE project. The academic result is to provide one artifact: a high-level architecture which describes the smart web-crawling system to be developed, including the data analytics techniques, and which address the requirement analysis carried out at the Dutch Customs Administration. Other final outcomes and useful contributions are the schemes of the CRM practices at the DCA, the process of requirement analysis for crawling e-commerce websites, and the literature review of the BDA techniques and other related works.

To reach the research objective, the following research questions are defined:

> *What design of a web-crawling architecture can deploy data analytics techniques to improve the cross-validation of price information for e-commerce at the Dutch Customs Administration?*

To clearly understand this research approach, the main research question is broken down into sub-questions which will be answered in a respective section of this master thesis project. The following sub-questions are set down:

1. What are the current customs risks management practices for e-commerce at the Dutch Customs and their limitations?
2. What is the state-of-the-art of web-crawling and big data analytics technologies relevant for the web-crawling architecture?
3. What is the most suitable design of a web-crawling architecture to improve the cross-validation of price information for e-commerce at the DCA?

1.5 Scientific and Social Relevance

First of all, the social relevance of this research has to be linked to the importance of the PROFILE project. It aims to improve the customs risk management of several European countries, which eventually means more safety on one hand, and more fair taxes and business rules on the other hand. The Dutch Customs Administration, in particular, will have new tools and technologies to increase its CRM practices, and the Netherlands as a country would benefit from it. Finally, this research could be used by other customs around the world which want to implement big data analytics techniques in their customs risk management practices.

Secondly, the scientific relevance is also important, as there is almost no literature on the topic, both for the application domain of the customs and for the new technology of artificial intelligence. Thus, an application of the latter in this specific domain provides a useful case study for both the fields of research of artificial intelligence and customs risk management. In addition,

the use of design science for the requirements analysis and the design of the architecture in the field of big data analytics is also relevant for the scientific community.

1.6 Research Methodology

This section defines the Research Methodology and the Theoretical Framework used to investigate the objectives of the research. The Research Diagram of this research is also presented (figure 1). In addition, the research strategies for each research question are summarized in table 4, and the collected data and collection methods for each question are explained.

The research approach used in this research is known in the literature as Design Science. It is a problem-solving oriented methodology which aims to foster innovations or artifacts concerning practices, ideas, products, practices, or technical capabilities to improve the design, analysis, management, implementation, or use of information systems (Denning, 1997; Tsichritzis, 1998).

In this case, the innovation consists of an IT artifact which implements data analytics into a broader information system. Questions which will be researched are for example: is it possible to develop such an artifact? Will it perform as desired? Will be relevant to the solution of the problem at stake? (Hevner, March, Park, & Ram, 2004). Through interviews with IBM and Dutch Customs experts, these questions are investigated. At IBM the technology characteristics and the design possibilities will be studied, while at the Dutch Customs the main requirements and the critical factors of the use case will be defined.

The research consists of an iterative and heuristic process of design and evaluates requirements and techniques to discover an effective solution to the customs' problem. The interviews with experts will be integrated with an on-going literature review of the new issues for the design. Below, the most recognized scheme of the design science methodology adapted to this case is reported as research diagram of this research.

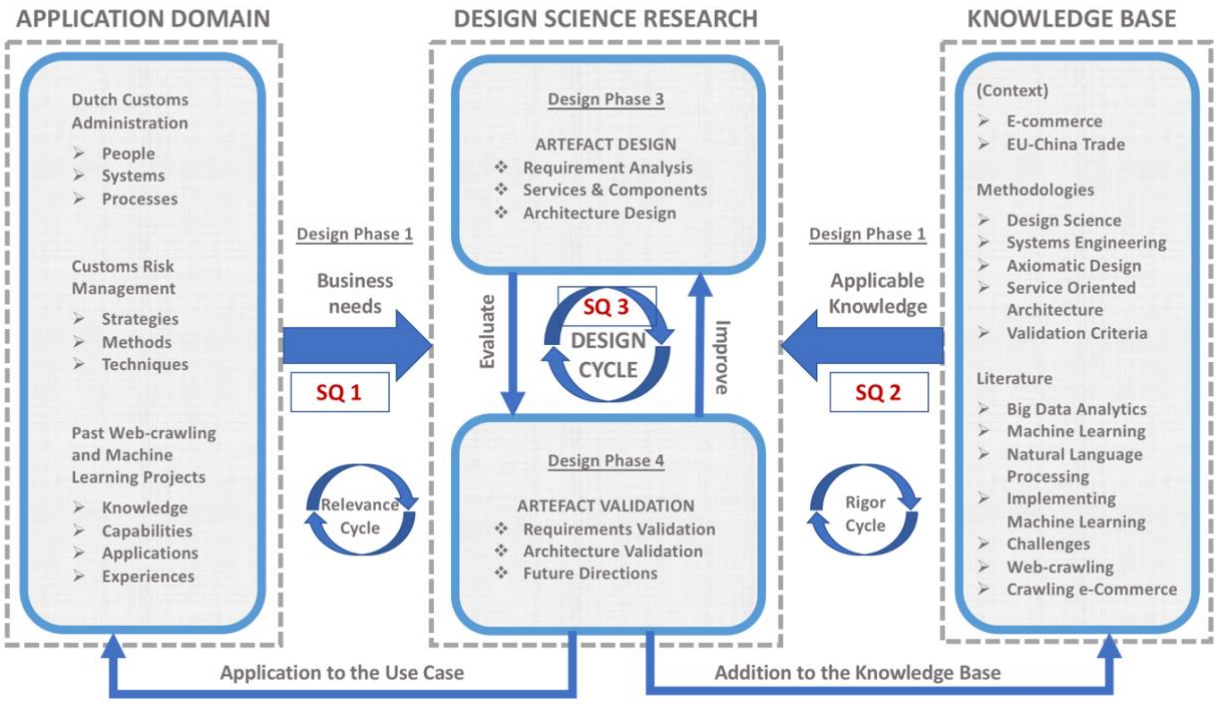


Figure 1: Information Systems Research Framework (adapted from Hevner, 2007)

In this case, the environment is represented by the Dutch Customs Administration, its systems to handle the risk management, and the e-commerce environment. At the other hand, the Knowledgebase is represented by the literature review on web-crawling and machine learning technologies carried out by the researcher, the expertise brought by the experts of International Business Machines Corporation, and the design science methodology which guides this research. These two forces are merged in the design cycle leading to the artifact design, in this case, a web-crawling architecture, which refers to the knowledge base through the "rigor cycle", and to the environment through the "relevance cycle". In this way, the design of the artifact is guided by the academic rigor and business relevance (Hevner, March, Park, & Ram, 2004).

As reported in the article of Hevner (2013), the result of a design science research is an artifact which can be one of the following types:

Table 3: Outputs of Design Science Research (taken from Hevner, 2013)

	Output	**Description**
1	Constructs	The conceptual vocabulary of a domain
2	Models	Sets of propositions or statements expressing relationships between constructs
3	Frameworks	Real or conceptual guides to serve as support or guide
4	Architectures	High-level structures of systems
5	Design Principles	Core principles and concepts to guide the design
6	Methods	Sets of steps used to perform tasks – how-to knowledge
7	Instantiations	Situated Implementations in certain environments that do or do not operationalize constructs, models, methods, and other abstract artifacts; in the latter case, such knowledge remains tacit.
8	Design Theories	A prescriptive set of statements on how to do something to achieve a certain objective. A theory usually includes other abstract artifacts such as constructs, models, frameworks, architectures, design principles, and methods.

The artifact result of this research will be an architecture type, as it will be a high-level design of the web-crawling system to be developed. This high-level architecture describes how the web-crawling system can address the requirements of the Dutch Customs with the state-of-the-art data analytics techniques.

The design process used in this research is an iterative process made of four phases:

1. Analysis of the application domain and the business needs of the Dutch Customs;
2. Research on the applicable knowledge, including research/design methodologies and an in-depth literature review in the field of big data analytics;
3. Design of the architecture artifact which addresses the requirements;
4. Evaluation of the architecture artifact with the customer and technical experts.

Phase 1: application domain

In this phase, the e-commerce environment and the relation with the customs authorities in general are described. Then, the customs risk management practices at the Dutch Customs Administration are described, and how the web-crawling architecture could be implemented is discussed. Finally, the past experiences of web-crawling and machine learning that are carried

out at the DCA are presented. From this, a first sketch of the requirements and the use case is drawn, including a further scope of the problem.

Phase 2: applicable knowledge
In this phase, the field of big data analytics and machine learning are investigated, and the web-crawling technology is explained using the existing academic literature. After having investigated and understood the theory behind these technologies, more practical guidelines for the implementation are researched, and relevant related works of web-crawling deploying advanced analytics techniques to process e-commerce data are provided. This knowledge base and scientific methodologies are then used to guide the next phase of the design cycle.

Phase 3: design cycle
The third phase is the design building. It is made of four sub-phases:

1. Gathering of the requirements of the technology;
2. Addressing the requirements with state-of-the-art big data analytics techniques;
3. Deriving the technological components of the architecture;
4. Defining the web-crawling architecture.

Though these four sub-phases, the design cycle transforms the business needs of the Dutch Customs into the design of the smart web-crawling system represented with an architecture.

Phase 4: results evaluation
The final phase of the design cycle is the validation and evaluation of the architecture. First, the artifact validity is addressed. This is made on two sides, external and internal validity. The external validity is addressed through accurate documentations of interviews with experts at both sides DCA and IBM to assure the repeatability of the research, while the internal validity is done evaluating the correctness of the scientific methodologies. Validation and evaluation are finally divided into efficacy, quality and utility, including criterion, content and construct validity.

The following table shows the four research questions, the four phases of the design process, the research strategy used in each phase, and the deliverable of each research question.

Table 4: Research Questions Strategies

Research Question		Research Strategy	Deliverables
SQ1	What are the current risks management practices for e-commerce and their limitations at the Dutch Customs?	Design Phase 1 Desk Research, Unstructured Interviews (DCA & IBM)	A better understanding of the business domain, business needs, and use case requirements
SQ2	What is the state-of-the-art of web-crawling and big data analytics technologies?	Design Phase 2 Literature Review, Unstructured Interviews (IBM)	A general overview of the existing knowledge about the technologies that are to be implemented, and a set of scientific methodologies to guide the architecture design
SQ3	What is the most suitable design of a web-crawling architecture to improve the cross-validation of price information for e-commerce at the DCA?	Design Phase 3 Semi-structured Interviews (DCA & IBM) Design Phase 4 Structured Interviews (DCA)	The artifact of this research: a business service architecture describing the web-crawling system. To arrive in the architecture design, also a requirements analysis is made

1.7 Systems Engineering Approach

The Engineering breakthroughs of today, due to their technological complexities, requires a powerful engineering process for the promotion of successful products, systems or software development (Snyder and Khalid, 2013). To guide the design of the web-crawling architecture, an approach from systems engineering will be adopted throughout the design cycle. This choice has been taken because the technology to be developed can be considered a system encompassing multiple technologies and made of different architectural components.

In addition, system engineering is also the field concerning the discipline of requirements engineering, which is the initial and main step of the design cycle, together with the representation of the high-level architecture. In this section, the systems engineering and requirements engineering methodologies are described, including the system engineering approach called Axiomatic Design. This methodology is explained and combined with the main systems engineering approach presented by Armstrong and Sage (2000), and the design science methodology described earlier.

What is a system? Every system might be defined as a collection of sub-systems, software and hardware components and actors that are designed to accomplish a number of tasks by satisfying particular functional and non-functional requirements (constraints) (Suh, 1998). Systems engineering is an integrative field of engineering management and engineering that focuses on designing and managing complex systems during their life cycles. It does not only design the system's components but does design the comprehensive architecture of the system. It sets priorities for the requirements of the system in conjunction with the client to guarantee that the various attributes of the system are properly weighted when balancing different technical efforts.

Systems engineering processes include requirements analysis, validation, functional and design verification, synthesis, and trade and assessment studies (IEEE Computer Society, 2005). Within the field of systems engineering, requirement engineering is the discipline that concerns with the requirement analysis phase of a system design. It is the first action towards the design of a system. In software engineering and systems engineering, it encompasses the activities of analyzing, validating, managing and even documenting software or system requirements. Also known as specifications, requirements determine the needs and conditions that are goals to reach when developing or modifying a new product or project. This also includes the possibility of conflicting requirements and the necessity of setting trade-offs.

Conceptually, requirements analysis includes three types of activities (Bijan et al., 2012):

- Eliciting requirements: commonly called requirements gathering, or requirements discovery, it is the process concerning stakeholder interviews and requirements documentation. This is the main activity of the requirement analysis process, and often a source of mistakes. For this reason, the product development practices adopted the new trend so-called "agile development", where numerous "sprints" (rounds of interactions with the client) are carried out throughout the entire project (Hazzan, Dubinsky, 2008).
- Analyzing requirements: determining whether the formulation of the requirements is clear and complete, consistent and unambiguous, and addressing any conflicts with appropriate trade-offs;
- Recording requirements: documenting the requirements. It can be made in various forms, such as a summary list, use cases, user stories, process specifications and a variety of models. A common practice in software engineering is, for instance, the use of the (Osis & Donins, 2017).

The elicitation of requirements is the activity that is perhaps most often seen to be the first step in the process of requirement engineering. The term "elicitation" is more accurate than the term "capture" it allows to avoid the assumption that the requirements can be collected simply by asking the proper questions. It is important to interpret, analyze, model and validate the information collected during the requirement elicitation before the requirements engineer can consider the set of requirements complete enough (Kaur and Singh, 2010). That is why, also in this research, the requirements collected during the interviews with the Dutch Customs are first re-written in a different and organized manner, and then validated with additional interviews.

Before the development of the requirements, it is necessary to understand the desires and needs of the customer and understand the context for the system's operation. During this phase, engineers, analysts, and the client need to guarantee that requirements are implementation-free. The implementation details can be captured as constraints if needed. Even though methods for prioritizing and facilitating customer needs are not applied consistently, they are well understood (Bijan et al., 2012).

The customer interview process to develop and understand the initial requirements is the first active step to undertake. As such, the success or failure of the interview weighs heavily on the outcome of the project, when the true need is understood, the engineers can record the proper requirements and choose the best design (Bijan et al., 2012). The requirements to develop must be actionable, traceable, documented, testable, measurable, related to determined business needs or opportunities, and detailed at the sufficient level for system design (Snyder and Khalid, 2013).

In software engineering and systems engineering, two types of requirements are distinguished: functional and non-functional requirements. A functional requirement mostly addresses the question "what a software system should do", whereas non-functional requirements set constraints on how the software system will do so. Repeated more formally, a functional requirement specifies a particular function of a system or component of this system, where the definition of function lies in a specification of behavior between inputs and outputs. Rather, a non-functional requirement (NFR) is a requirement that defines the criteria for the judgment of the operation of a system, but not the specific behaviors. In general, non-functional requirements (or quality requirements) are more complicated for the expression in a measurable way. This fact makes them more complex for the analysis. Particularly, NFRs are more likely to be properties of a whole system, and therefore will not be verified for individual components.

The definition and formulation of the requirements is the first sub-phase of the design cycle. It is carried out following the approach by Armstrong and Sage (2000), in particular using a method called Functional Decomposition and Structural Analysis. From this step onward instead, this research makes use of another systems engineering approach called Axiomatic Design. This approach was proposed by Suh in the 1998 and presents a rigid mapping between the architecture requirements and the architectural components. In this sense, it is a methodology to pass from the requirements analysis to the design of the features of the architecture. While the systems engineering approach guides the requirements formulation, this methodology guides the design process toward the architecture description.

According to the axiomatic design, the world of design is made up of four domains: the customer domain, the functional domain, the physical domain and the process domain (Suh, 1998). The customer domain is described through the customer needs, which can be explained as the attributes that the customer is looking for in the product to be developed. From the customer domain, the following one is called the functional domain: the customer needs are specified in terms of Functional Requirements (FRs) and constraints (Cs), which are also known as non-functional requirements (Brace, Cheutet, 2011). To satisfy the FRs, design parameters (DPs) are the next step described in the physical domain. Finally, to produce the product described by the DPs, its functionality must be described through process variables (PVs) explained in the process

domain. According to Suh (1998), "many seemingly different designs tasks in many different fields can be described in terms of the four design domains" and "all designs fit into these four domains".

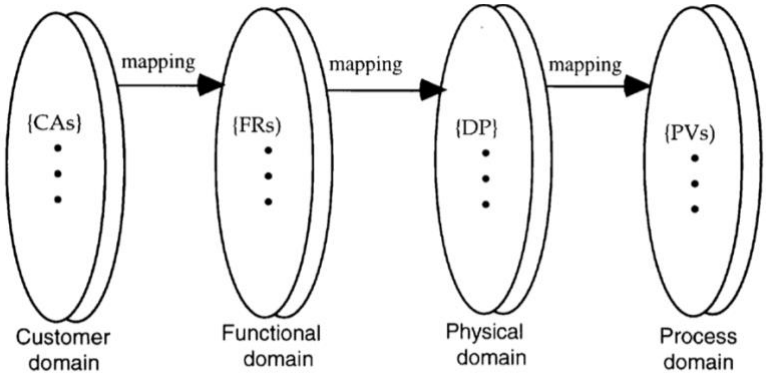


Figure 2: The 4 Design Domains according to the Axiomatic Design Methodology (Suh, 1998)

In the figure above, the domain on the left relative to the domain on the right represents 'what I want to achieve', while the domain on the right represents the design solution for 'how I propose to satisfy the requirements specified in the left domain'. In this sense, the axiomatic design is a methodology which guides the designer from the customer domain to the process domain in a systematic and rigorous manner. It is a top-down approach: it starts from the requirements analysis and gets to the architecture design.

Besides the four domains described earlier, the axiomatic design also consists of mapping matrixes between one domain and another. These mapping mechanisms are fundamental in reaching the system design in a rigorous manner. In the case of this research, the mapping of the requirements of the functional domain into the architecture components is the most relevant and focal center of this research. The following scheme summarizes the design methodology used referring to the axiomatic design and the four design domains:

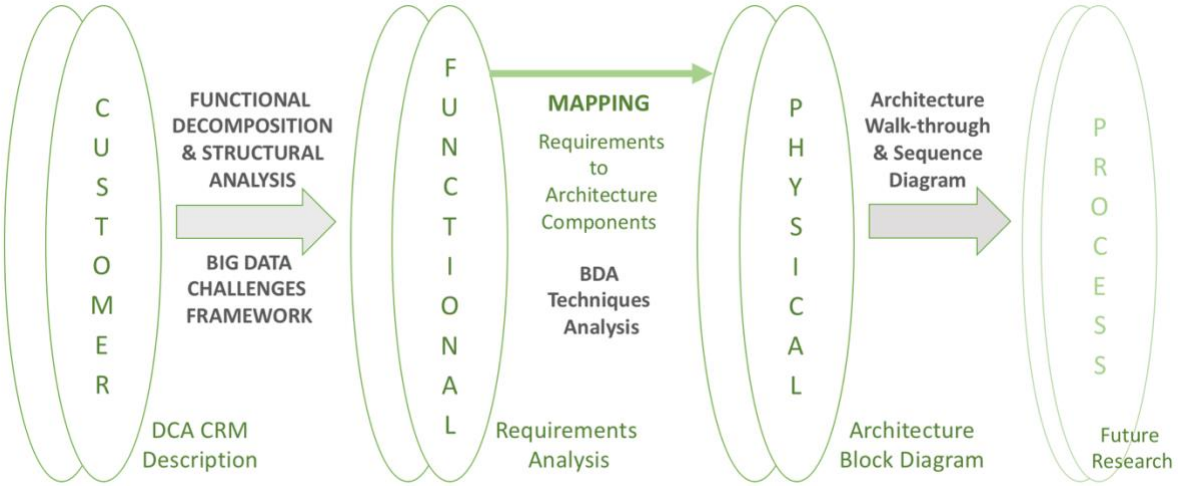


Figure 3: Design Domains adjusted to this research (adjusted from Suh, 1998)

As it is shown in the scheme (figure 3), to support the sub-phase (2) of the design cycle, also a framework collecting the challenges of big data projects is used. This will be explained further in the literature review. In addition, it is explained as the fourth and last domain of the design is left to future research. Once the DPs are chosen, designers must go to the process domain and identify the Process Variables (PVs) based on the creation of a new process or the use of an existing

process. This is a more highly detailed description of the system which is possible when the specific algorithms are decided, and it will be the main task of the technical team in the IBM Research Lab, Ireland.

However, the mapping toward the fourth design domain is addressed. After the description of the physical domain, a further step is described toward the process domain. This is made describing the architecture walk-through and its sequence diagram in the Unified Modeling Language (UML). This is a diagram to show the interaction and the behaviors of the architecture components within a single use case (Osis & Donins, 2017). As these description techniques aim to explain the flow of information through the architecture, they are used to map the physical domain to the process domain.

In these representations, the web-crawling architecture will be sketched using a block diagram, where each block represents an architecture component. A block diagram is, in fact, a diagram of a system in which the principal parts or functions are represented by blocks connected by lines that show the relationships of the blocks. In particular, when describing the architecture, each block of the architecture is considered from the black-box perspective, which is a typical approach in the systems engineering field (Mendez Fernandez, Penzenstadler, 2014). A black box can represent any object or components, and it can be viewed only in terms of its function, inputs, and outputs. No any knowledge about its internal workings is available. Its implementation is said "opaque" (black).

1.8 Service Oriented Architecture

Since the main delivery of this research is an architecture-type artifact, this section of the literature review addresses the approaches in literature used to describe an IT architecture. In particular, it has been chosen the design style of service-oriented architecture (SOA) as it is a modern approach which enhances a flexible design of the architecture by focusing on the description of independent components called “services”. Given the high variability of the solution design, it is wise to choose an architecture representation which allows to be flexible and focus on the technology, in this case the big data analytics techniques.

The use of service-oriented architecture (SOA) approach for applications’ building is one of the latest trends in the evolution of the way to deliver IT functionality in the last decade, together with end-to-end business processes and applications (Dahl, 2007). SOA might be presented as a flexible IT architecture style of software design with different application services, which can be integrated. These services might be offered by service providers or developed internally (Butler, 2008). Application components provide services to other components, using communication protocols. Thus, the discrete function of application is organized and integrated into interoperable services to combine them and re-use for specific consumers’ business needs. Independency of products, technologies and vendors is one of the main design principles of SOA (Draheim, 2010).

In SOA environment, applications are basically “collection of services” that might communicate with each other and they are mutually connected. Reassembly of these applications is possible only if services are equally and universally discoverable, accessible and clear to understand to any other virtually service application independently from its location (Butler, 2008). The means of communication together with method of connection are fundamental components as far as services mostly built around interactions. The envelope (defines communication protocol) and the payload (specification of the message) are used by the services in SOA environment to communicate reliably and it applies open standards towards enabling of data exchange and operation instructions to be enabled (McIntosh, 2004).

The main four properties were derived with accordance of various definitions of SOA in previous studies (The Open Group, 2018):

- Specified outcome of a business activity is logically represented;
- SOA is self-contained;
- For end consumers SOA uses principles of a black box;
- Various underlying services might be a part of SOA.

In addition, every application service of a SOA is a building block and must play one of three roles:

- Service provider

Service provider creates particular web service and transfer its information to the service registry. Each of these providers discus upon a various of whys and hows, for example which service is displayed, what to give more importance: security or ease of use, what price to propose for the service for and others. The vendor must also decide in which category the service for this broker should be specified, and which agreements about the trading partners should use this service.

- Service registry, service broker or service repository

The main functionality of this role is to make web services' the information available to every possible requester. Whoever enforce the broker creates the capacity of the broker. Availability of public brokers must be anywhere and everywhere; however private brokers are of limited availability only to a specific public. Previously, UDDI was a supported attempt to arrange Web services detection.

- Service requester/consumer

This role finds records in the broker registry using different search operations, and then attaches to the service provider to call one of his Web services. Whichever service consumers demand, they should take it to the brokers, link it to the appropriate service, and afterwards use it. They can access several services if the service arranges several services.

To build any SOA it is necessary to respect standard principles, which were stated in various researches and manifesto. Among them:

- Standardized service contract: there is a standard communications protocol for all the services.
- Service reference autonomy: each service knows about the other services just of their existence, and not about their functions.
- Service location transparency: services can be called from anywhere within the architecture network.
- Service longevity: services should be designed to be last long.
- Service abstraction: the services act as black boxes, which means that the consumer of that service does not see the inner logic of that service.
- Service autonomy: services are independent and in full control of their functionalities.
- Service granularity: all the services should have an adequate size and scope.
- Service normalization: the services should be designed to minimize redundancy.
- Service composability: services can be used to compose other services.
- Service reusability: logic is divided into various services, to promote reuse of code.

Now that the type of architecture has been described (SOA), it is left to be chosen how to describe it. For this, the existing literature offers an approach based on architecture viewpoints, which are basically the description of the architecture under different perspectives. Views and viewpoints are central to the standard's way of describing architectures. A view model or viewpoint framework in systems engineering and software engineering, is a framework which defines a coherent set of views to be used in the construction of a system architecture or software architecture (or even enterprise architecture). A view is a representation of a whole system from the perspective of a related set of concerns. A viewpoint is instead defined as a collection of patterns, templates, and conventions for constructing one type of view (IEEE Standard 1471).

A viewpoint can be for instance, the analysis of the architecture from the user perspective, or focused on the relation between the architecture and its datasets. Otherwise, a viewpoint can also be a layer of the architecture such as logical – focused on the architecture components – or physical – focused on the hardware. Philippe Kruchten proposed in 1995 the 4+1 Architectural View Model, which is a conceptual framework that defines five different viewpoints to describe an architecture: logical, physical, process, development and scenario.

The description of the architecture provided in this research is focused on its architectural components. In this sense, it is a logical viewpoint, designed according the SOA principles. This choice has been made because the logical viewpoint and the choice of the architecture components is the core problem of the design, in this case. The final user of this manuscript – i.e. the technical team that will develop the system – knows how to physically develop a system architecture and what hardware is best to deploy for each case. What it is important is that the solution which should be adopted best satisfies the requirements of the Dutch Customs Administration. Thus, the focus of this research is to investigate the technologies and techniques that should be deployed. And from an architectural point of view, this is better described through the logical and process viewpoint.

This has not to be confused with the design domains of the Axiomatic Design (Suh, 1998) described in the methodology section (see 1.7). These domains represent the steps of the design, which from the customer domains goes through the functional analysis and then to the architecture representation, which is divided in first the physical domain – the parameters of the system – and second the process domain – the interactions and functionalities of the system (Suh, 1998). This is a simplified description because there are also the mapping layers, but it is important to consider that for instance the physical viewpoint is not the same as the physical domain of the Axiomatic Design methodology.

Finally, these representations and viewpoints of the architecture must be done in a standard manner so that engineers and managers around the world can easily exchange design ideas. For this reason, I will use in my research the the Unified Modeling Language (UML), a general-purpose graphic language used by software professionals for specifying, visualizing, constructing, and documenting the artifacts of a software intensive system. It can be defined as the standard language for writing software blueprints (Boosch, Rambaugh, Jacobson, 2005).

The UML has three main models: the User Model, the Object Model, and the Dynamic Model (Lodderstedt, Basin, & Doser, 2002). As my focus is not to represent the entire architecture, but to collect the requirements and give a guideline for the best design that matches these requirements, this research will mainly use the UML standard as a formal mean to describe the architecture functionality. This is done through the so-called in literature "Sequence Diagram", a scheme to show the interaction and the behavior of the architecture components within a single use case (Osis & Donins, 2017), which in this case I consider the most complete scenario.

The reader could opt that the UML use case diagram would be useful for the gathering of the requirements, and thus it should be used in this research. But since the use case under analysis

does not present many scenarios with different requirements, but instead it has a clear main objective, I rather preferred to use the approach of Amstrong and Sage (2000) coming from the systems engineering discipline, instead of the software engineering approach largely based on the UML representations.

1.9 Thesis Outline

After this first chapter of introduction to the problem and explanation of the research methodology, the next chapter will describe the application domain of the e-commerce environment and the customs risk management practices used by the Dutch Customs Administration. Here a real experience of an e-commerce purchase is described, and the past experiences of the DCA in web-crawling and machine learning are presented. Chapter three rather addresses the existing literature review on the fields to set the knowledge base of the research. It ranges from the state-of-the-art of the web-crawling technologies and machine learning to systems engineering and requirements analysis.

After these preparatory chapters, chapter four is where the design of the artifact is explained. First, the requirements analysis is carried out, then the big data analytics techniques which could address these requirements are discussed. This leads to some functional components of the architecture and finally to the ultimate design of the high-level architecture of the system.

Chapter five explains the validation of the artifact, which in this case is the design of the architecture. This is carried out through interviews on the requirements analysis, and with a reflection on the used methodology. In addition, guidelines to conduct validations tests once the first prototype will be built are proposed.

Finally, in chapter six the conclusions, expected results and limitations are outlined. This chapter also gives recommendations to the Dutch Customs Administration and IBM, explains the main practical and scientific contribution of this research, and eventually reports the reflection of the researcher.

2 Application Domain

This chapter describes the domain in which the technology to be developed should operate. It corresponds to the first phase of the design cycle described by Hevner (2004) and reported in figure 1 (section 1.6). It thus answers to the first research question of "what are the current customs risks management practices for e-commerce at the Dutch Customs and their limitations". Answering to this question, the business needs of the Dutch Customs Administration are investigated and brought to the design phase of the design cycle (phase 3) as valuable input. As it is explained in the table of the research strategies (table 4), the research strategy of this phase is desk research and expert interviews, at both sides DCA and IBM. These interviews are written down and reported in the appendixes.

As explained, the focus is on the Dutch Customs Administration (DCA) and its system for customs risk management. Here the systems to handle the declarations documents and assess the associated risk are explained. It is thus described how the DCA decides to inspect a shipment in both container and parcel targeting. It is important to distinguish these two because one focuses mostly on the logistics connected to the harbor of Rotterdam, and the other is linked to e-commerce trade and it is mostly handled at the airport of Schiphol.

In addition, the e-commerce world is described, including an initial analysis of possible search queries on both search engines and e-commerce platforms. To further understand the application domain, this section also describes a real case purchase of a drone on the Chinese website AliExpress, including how the package was described in the declaration document and what duties have been asked to pay by the Dutch Customs.

Finally, after the application domain has been explored, the problem statement is better defined and formulated. Here is also described the past experiences of the DCA about web-crawling and data analytics, so that valuable learned lessons can be shared. This chapter is made reporting the numerous interviews (see Appendix B, C, D, E, F) with the Dutch Customs Administration, and in collaboration with the industry experts at International Business Corporation.

2.1 E-Commerce & Customs Administrations

This section describes the logistics behind the e-commerce purchases, and then it gives important information for the customs authorities, including the differences among the most relevant e-commerce platforms.

When the consumer buys something on the e-commerce platform, the e-commerce asks the supplier to ship, the package arrives at the e-commerce warehouse, the e-commerce gives it to the courier, the courier brings it to the final consumer. This is how a purchase on e-commerce is handled from a logistic point of view. When a product enters a foreign country – meant as a different country from where the product is manufactured – it must be checked by the local customs administration to not break the local laws and pay the import duties.

In this scheme, the couriers act as e-fulfillment service providers, which mean that they pay the duties to the customs authorities before the products arrive in the country where must be imported. This is true for every business-to-consumer (B2C) e-commerce platforms, where the consumers pay these duties directly when selecting the shipment option. When it is business-to-business (B2B), it should be the end of business that should fill the declaration at the customs through customs brokers. Customs authorities require specific procedures and documents and

set consolidated paths with certified traders like UPS, DHL or FedEx to increase the speed of clearance of goods and facilitate the international trade. The taxes to be paid are usually import VAT, import duty, and excise duty.

What are the main e-commerce platforms? Besides Amazon or eBay, this research addresses particularly the Chinese platforms which delivery to the European Union. The most common Chinese e-commerce websites are: Taobao, which is consumer-to-consumer (C2C) and operates in China and sometimes shipping abroad (its counterpart can be considered to be eBay); AliExpress, which is B2C and is made specially to ship outside of China (it can be considered similar to Amazon); finally, Alibaba is the B2B Chinese e-commerce thought to connect businesses around the world with the Chinese manufacturing industry. In Alibaba, it is possible to order almost anything, and the Chinese factories will produce it in a minimum quantity and ship it anywhere in the world in a reasonable time. Among the Chinese e-commerce, there are also Tmall and JD, which are B2C (like AliExpress) but mostly operating within China. However, it is useful to consider them because they can provide real information about products prices, and they have an English version of their platforms. JD is especially good for electronics products.

From the Dutch Customs Administration perspective, the most relevant e-commerce websites are Alibaba and AliExpress, as they are the ones which mostly ship to the EU and to the Netherlands. What's the difference between Alibaba.com and AliExpress.com? As said before, Alibaba is B2B, while AliExpress is B2C. In practice, this means that most of the members on Alibaba.com are manufacturers, trading companies or resellers who trade in large order quantities, while AliExpress is a global retail marketplace offering quality products at factory prices in small quantity. Analyzing the e-commerce platforms characteristics before developing the web-crawling technology can be useful to understand what differences should be detected by the web-crawler. For instance, if an e-commerce website includes taxes or shipment costs in the value of the product, and another e-commerce does not, the web-crawler should be able to adjust the products prices accordingly.

2.2 Risk Management at the Dutch Customs

In this section, it is described how the system for customs risk management currently works at the Dutch Customs Administration. As the IBM industry expert Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management, explained during an interview on the customs risk management practices of the Dutch Customs, there are four different scenarios: entry, import/export, transit, and e-commerce.

Before goods are shipped to the Netherlands, they need an *"Entry Summary Declaration"* (ENS) which is the EU standard declaration format for entry declarations. These declarations are managed by the system *"Import Control System*" (ICS), or in Dutch *"Douane ManiFest"* (DMF). This scenario concerns with the EU security and logistics.

When the goods arrive in the harbor of Rotterdam, some goods are imported in the Netherlands and some other stopped at the harbor as transit to be shipped to other countries. In this process, the declarations are not connected directly with the physical containers, as an entry declaration can result in multiple import and transit declarations.

When a goods item is imported it is reported with a *"Single Administrative Document"* (SAD) declaration, which is the EU standard that is used for all variations of Import and of Export Declarations. These declarations are processed by the *"Declaration Management System"* (DMS), in Dutch "*Aangifte Systeem"* (AGS). In this scenario, the imported goods are subjected to import

and fiscal duties to be paid, and the Dutch Customs Administration is in charge of detecting fiscal frauds.

After conducting desk research and interviews with experts of both the Dutch Customs Administration, in particular Marcel Molenhuis, Senior Advisor for Data Analytics, and IBM (as mentioned earlier, Ben van Rijnsoever), it has been possible to recap these four scenarios and their systems in the simplified scheme below (figure 4).

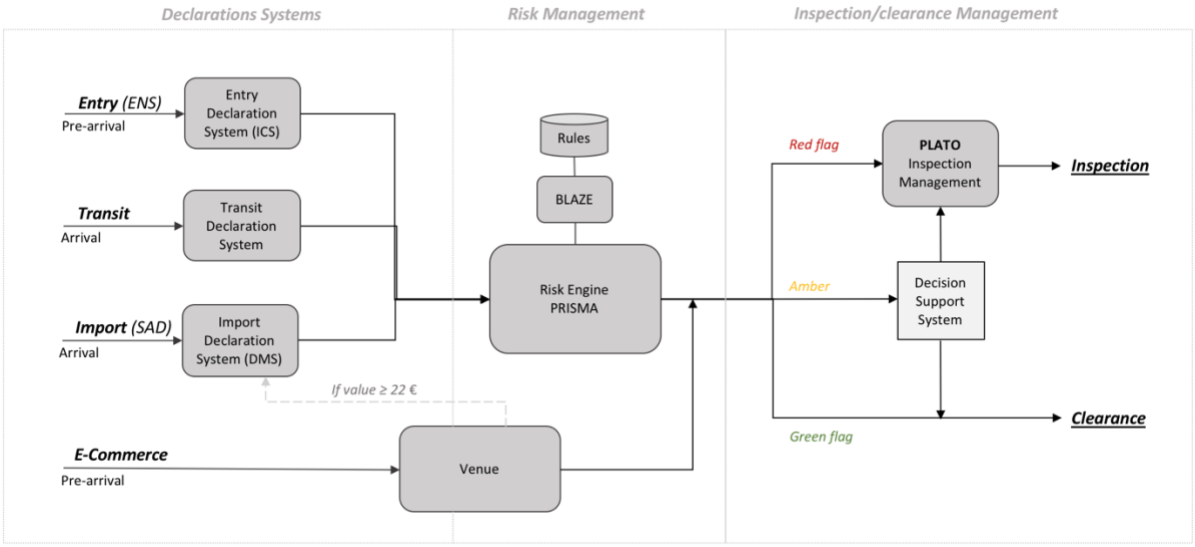


Figure 4: Scheme of the CRM system at the DCA

As the figure above shows, the customs risk management processes are nagged through three types of software systems (columns in grey in figure 4): a system to handle the declarations documents; a risk engine to assess the risk related to each declaration and decide if that package should be inspected or not (green or red flag); a system to handle the inspections for those package that have been a red flag.

The ICS and DMS systems mentioned above are of the first type, i.e. to handle the declarations documents. The second type of system is the risk engine to assess the risk. This is managed in the following steps:

1. The declarations information is processed by the system called *PRISMA*.
2. Many "if-then" rules are applied by the system, using a business rules engine called *BLAZE*.
3. Packages are flagged as red (to inspect), green (free to go) or amber (need further supervision).
4. For the amber flags, the targeting officers use the management dashboard CRIS to collect external information and have a 360-degrees-view to better decide whether to flag as green or red.

The packages that have been classified as the ones to be inspected are then processed by the system *PLATO*. This system is in charge of assisting the DCA officers during the inspections, for instance recording the results of the inspections.

2.3 E-Commerce at the Dutch Customs

As the reader can see from the previous scheme (figure 4), the CRM process for e-commerce is a bit different from the other three scenarios. In particular, the declaration processing and the risk assessment is done with a different system called *"Venue"*, and not by PRISMA as described earlier. Moreover, the DCA department for e-commerce is located in Schiphol, the Airport of Amsterdam, and not in Rotterdam, as the e-commerce packages are shipped mostly by plane and not by ships in containers.

This section explains the CRM process specifically for e-commerce packages, thus for parcel risk targeting, instead of container risk targeting. As explained earlier in the introduction chapter, the main difference is that the common risk assessment done on the basis of the traders is less effective in the case of e-commerce because the traders are many and fast-changing.

The DCA National Coordinator for e-Commerce Han Bosch reported that "about one out of three e-commerce declarations is wrong" (3rd May 2018, see appendix C). This makes the e-commerce risk assessment the most critical one for the Dutch Customs Administration. By interviewing the same expert Han Bosch and Ben Schmitz, the DCA E-Commerce System Coordinator, the DCA CRM process for e-commerce is described:

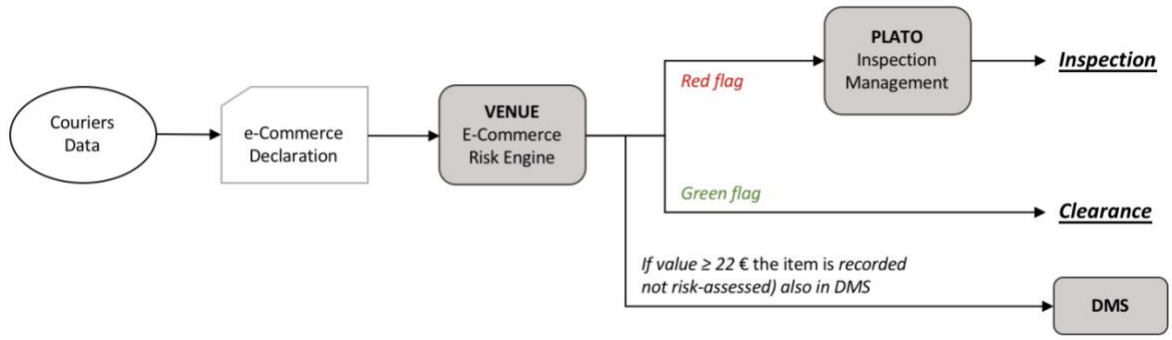


Figure 5: Scheme of the e-Commerce CRM system at the DCA

The DCA department for e-commerce receives files from the couriers that are already structured and ready to be processed by the DCA e-commerce system *Venue*. Ben Schmitz, the Venue E-Commerce System Coordinator at the DCA, reported in an interview on 6th June 2018 (see appendix D) that each item is signed as:

- "A" if the item value is below 22 euros for the VAT and 150 euros for the customs duties, and thus it is free to go;
- "B" if the item is a special type of product, and thus free to go;
- "C" if the item value is above 22 euros, and thus it a customs declaration for that item has to be forwarded to the DMS system;
- "D" if the item is to be stored in the warehouse because it will depart again.

In addition, Ben Schmitz also reported that the Venue system (1) formats the files by the couriers when they are not correctly structured, (2) does the risk assessment, and (3) gives three outputs:

- Sends a reply to the couriers with the output of the risk assessment for each item (thus what item must be inspected and what is free to go);
- Sends a list of items to inspect to the system PLATO, the same system used to handle the inspections in the other scenarios.

- Sends a log file to be added to the history archive. The Venue System Coordinator Ben Schmitz reported as the DCA has collected data for the last six years for a total of around 30 million items.

Finally, the items that require taxes to be paid (must be above 22 euros) are forwarded to the declarations management systems (DMS), the same system used in the import scenario, because they must be recorded as imported goods. The "DMS will not execute the PRISMA risk assessment for these items since the risk assessment is already done in Venue" (Ben Schmitz, 6th June 2018).

2.4 Real e-Commerce Purchase

After having understood how the Dutch Customs Administration carries out its processes of risk management, I wanted to see how a real purchase of an e-commerce product from China would have been processed.

This section describes a real purchase experience made on the e-commerce platform AliExpress in March 2018. It has been bough Drone on the AliExpress platform and shipped to the Netherlands. The drone was bought at a price of 1244.90 euros. When the packaged arrived at the house of the final consumer, the item description on the box was "toy model", while in the e-commerce website was: "EU version DJI Mavic Air drone and Mavic Air fly more combo drone with 3-Axis Gimbal 4K Camera and 8 GB Internal Storage". Also, the declared value was different from the one on the e-commerce. The reported value on the declaration was 80 euros, and the duties asked to be paid were 53 euros.

This example shows as without asking anything to the Chinese sender, the description of the good and its value were on purpose modified to make the buyer paying fewer taxes. In addition, from the product description on the e-commerce website, it is possible to understand how long a product description might be, and how complicated it might be to make the right query online. Finally, the company sender was not the same as the one in the e-commerce platform. This means that it is possible that online there is the main company producing the product, but on the declaration, another company is listed as the sender, for instance, the local warehouse company, or even a retailer. Thus, it makes it ineffective to query the product on the e-commerce platforms by looking for the sender, because the one on the declaration might not exist on the e-commerce platforms.

I did some further analysis investigating the e-commerce environment. I noticed that trying to search for a product on the Web is not that easy. The first problem is what to type in the search bar of the browser to find the desired product. It is hard to decide a standard query that would be well-working for every product. For example, it is easy to type just the category of the product, such as "drone" plus some keywords like "e-commerce" or "price" and find e-commerce websites that sell drones. But then it might not be possible to find that exact type of drone. On the other hand, inserting as search query the entire description of the product could lead to misleading results or even null results. From here, it is clear as a first problem to be addressed when trying to find a product on e-commerce platforms is how formulating the right search query.

Another problem is after having submitted the query, the results showed by Google are not just e-commerce, but also advertisement, official stores, or even completely irrelevant results (depending on the search query and the product). Another problem is thus how to recognize e-commerce platforms among the search results. And going further on the same line, the following question could be: should the web-crawler check any e-commerce it finds after the search query on Google or should select the most relevant? And what would be the selection criteria?

2.5 Problem Formulation

In this section, the problem is more precisely defined given the knowledge of the Dutch Customs domain and analysis of the e-commerce environment, with the aim of further scoping the problem and deriving a more concrete formulation. Let's start with the scoping already defined.

As it has been mentioned in the previous chapter (section 1.3) that the web-crawling system will focus only on English information on the Web and it will address only the fiscal frauds with the purpose of decreasing only the false positive. In addition, given the complexity of the problem, it has been agreed to scope the research even more to only the five most critical categories of products.

The Senior Advisor for Data Analytics at the Dutch Customs Marcel Molenhuis provided me a list of the 10 most critical products and the explanations of why they are critical for the DCA. I report his excel file sent to me by email on email on 19th June 2018 in the figure below:

Goods description	Non-conformity
SYNTHETIC HAIR	Low value
WATCH	Low value (75%) and counterfeit (25%)
AIRSOFT GUN BELOW 2 JOULES	Low value, repcipient of the goods must be a member of the NABV (Dutch Airsoft Intereste Association)
LEATHER JACKET	Low value (95%) and counterfeit (5%)
CAMERA LENS	Low value
HARD DISK DRIVE	Low value
CAR CD (MUSIC) PLAYER	Low value
(RAW) TOBACCO/TABAK	Excise duty
IPHONE	Most counterfeit, others low value
USB CABLES	Low value and in one case product safety

Figure 6: Excel File by the DCA (2018) showing the most critical 10 products

Because as mentioned earlier, this research focuses on the fiscal problem only, the following five products have been chosen from this list for this research:

Watch	Leather jackets	Camera lens	Hard disk drive	Car CD player

This selection has been done also because some of the provided products are not of high value (below the 22 euros) and so could be exempt from import duties (at least in part).

Marcel Molenhuis also provided me with an example of the files used in Venue and Plato. This was requested as it is important to see the available data that could be used when deciding on the design of the system. But also, to see what a product description looks like. The product description I received by email on 15th May 2018 by Marcel Molenhuis was:

" Toestellen voor het ontvangen, omzetten en zenden of regenereren van spraak - INV 76382821 ",

which in English means: "devices for receiving, converting and transmitting or regenerating speech". Besides being in Dutch, this product description does not say anything for example about the brand of these devices, or any detail (e.g. technical specifications) to identify the right price online. Furthermore, Marcel Molenhuis reported me as this example which he provided is one of the most complete cases.

This reality makes the problem to be solved really complex. After few manual tests on the Web and e-commerce platforms, it has been clear that if the product descriptions on the declarations are this vague, crawling the Web and returning a price deviation that actually makes sense

becomes difficult, since such vague descriptions could lead to crawling products online that might have price spans too large to carry a meaningful analysis of the price deviation.

However, there could be still declarations for which this analysis makes sense, either because accurate enough or because of the nature of the product. For this reason, it has been agreed to assume a sufficient level of accuracy of the declarations descriptions for this research, so that it is possible to proceed toward the design of the web-crawling system which is not too complex and that can still address part of the declarations.

In addition, to make the problem solvable, I need to specify what kind of fiscal fraud I am going to tackle first. As it is shown in the previous section, the fiscal fraud was made under different aspects: the product value was much lower; the product description was fake; the sender was not the same seller of the product. Finding a mechanism to tackle these frauds altogether is challenging and could lead the research out of addressing the main purpose: verifying that a web-crawling technology for e-commerce platforms would help to improve the cross-validation of price information. Trying to detect all the frauds in the declaration could instead leading to use of all the resources without reaching this objective.

This research thus needs to be further scoped. According to the Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Service (GBS) at IBM Netherlands Ben van Rijnsoever, it may be possible to partially address these frauds by checking the weight information. The Dutch Customs, in fact, has this information for each package since it is provided by the couriers – which accurately measure it to consequently price their service (the more the weight of a package, the more is expensive to ship). According to Ben van Rijnsoever, estimating the number of products inside a package could result in an expected value to be paid, and thus to the inspection of that package – even if the real case could be just a different (heavier) product, but still a fraud.

Listening to the opinion of the DCA expert Han Bosch, National Coordinator for e-Commerce, the approach proposed above would also detect frauds scenarios such as a declaration describing an "iPhone cover" when in reality there is also a proper iPhone (inside the cover). After having the approval of the DCA experts, it has been agreed that this research will assume that the declarations correctly describe what is in the packages, but that the value declared is much lower than the actual one.

Summarizing, the problem scenario needs to be better defined by two assumptions about the product descriptions on the declarations in order to simplify the use case:

1. The descriptions are sufficiently informative about the products.
2. The descriptions are not fake, i.e. not misleading or describing false products.

These assumptions were agreed on the meeting of 18th June 2018 (appendix E). Once the problem has been accurately scoped, the problem statement of "*how to cross-validate the price information of the declaration with the online information on the e-commerce platforms"* can be addressed.

As the DCA National Coordinator for e-Commerce Han Bosch reported during the interview on May 3rd 2018, "the targeting officers have much more red flags to inspect than what they can physically check" (see appendix C). For this reason, it is useful that the web-crawling system would be used in this section of the process.

In the scheme below, I recap the customs risk management processes at the Duct Customs Administration and show where the web-crawling architecture would be implemented.

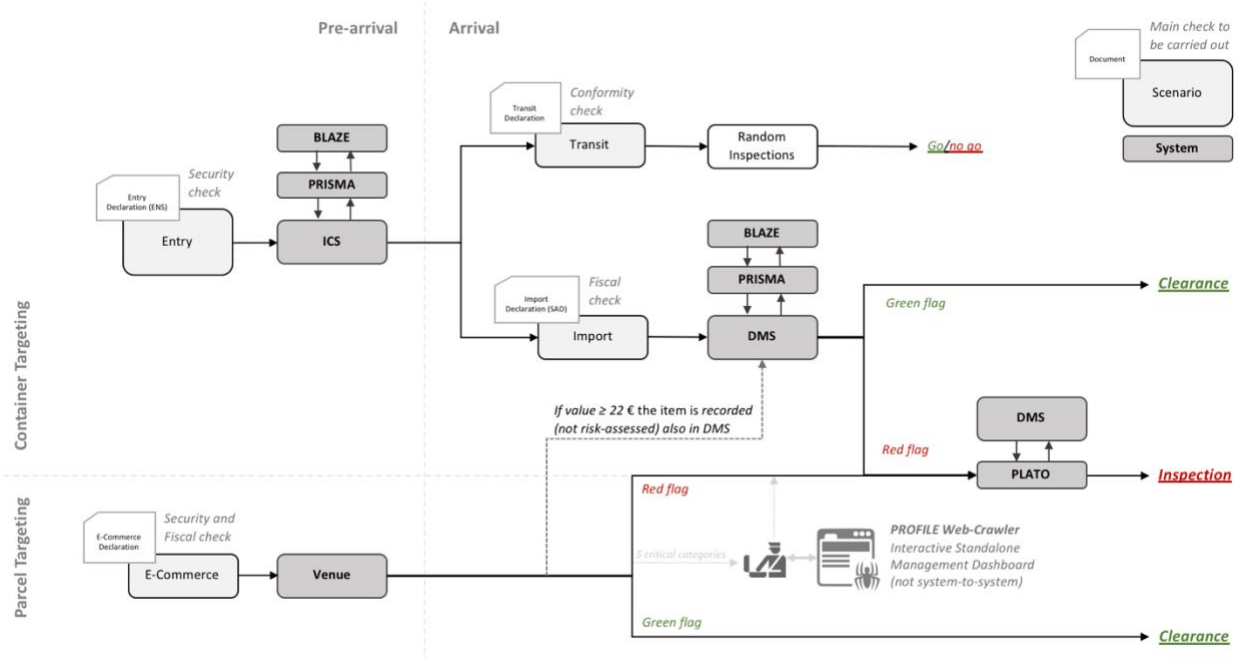


Figure 7: Scheme of the CRM processes at the DCA with the PROFILE web-crawling system

When having too many red flags to inspect, the targeting officers can use the web-crawling tool to decide which one of these packages they should inspect. Thus, the web-crawling system to be developed will be an interactive standalone management dashboard used by the targeting officers working with the PLATO system and in charge of managing the e-commerce inspections. The targeting officers will then be the users of this web-crawling tool.

Besides showing where the web-crawling system will operate within the DCA systems, the figure 7 shows as transit and import are after the entry declaration, which is pre-arrival. In addition, there is an evident distinction between parcel targeting and container targeting, and it shows that the transit scenario does not use the BLAZE-PRISMA risk engine to assess the risk.

2.6 Web-Crawling Projects at the Dutch Customs

In this section, past experiences about web-crawling carried out by the Dutch Customs Administration are presented. This is valuable information for PROFILE and can avoid reinventing the wheel. This information has been reported after having interviewed the DCA Senior Advisor for Data Analytics Marcel Molenhuis, and the Open Source Intelligence Expert and Web-crawling Lead for the DCA Jo Bootsma during two interviews on 3rd May and 28th June 2018 (see meetings notes at the appendixes C and F).

According to these experts, the DCA developed two projects, one for web-crawling (just indexing) and one for web-scraping (retrieving information; this difference will be explained better in the literature review in the next chapter). The web-crawling project has been abandoned because it was a too old technology, while the web-scraping tool is currently used.

About the web-crawling project not in use anymore, its first version was called Xenon, and it was a project by the British and Dutch Customs 10 years ago. There has been an updated version 3 years ago called Tafeic with also the Swedish and Belgium customs involved, but also this project has been abandoned. The main reason is that the "technology deployed can only handle text-

based web content and is not able to retrieve information in a more dynamic web populated with multimedia data as is often used today" (Jo Bootsma, 28th June 2018, appendix F).

The system takes as input a list of websites to crawl and returns a list of relevant words with their weight. It was meant to make investigations on request of the business intelligence department. Today is not used. This crawler also had the possibility of being trained through feedbacks to improve its accuracy. Finally, they did not consider possible problems related to privacy or terms and conditions of websites which might not allow robots to crawl their information. For the DCA legal department it was enough to consider that the data were stored for investigations just temporarily, and thus the requirements of data privacy were no applying.

The web-scraping project is more recent and still in use. The DCA currently uses Visual Web Ripper (http://visualwebripper.com) to scrape all the information starting from an URL and save it in a database. After the URL is inserted, the software goes to that page (as a normal browser) and the user can select the elements of the page that the software should save in the database (thus it recognizes the page layout). The DCA is currently working on making a database with information about 5/10 chosen products. This could be useful to create a database with personal information which would be hard to be used by externals companies such as IBM.

The DCA expert Jo Bootsma also shared his past experiences with the e-commerce platforms they crawled. In particular, they reported that Alibaba does not show the shipping cost at the first generation (thus one further crawl is required); eBay has the shipping cost shown below in the same page; AliExpress is slower than Alibaba in terms of response time; considering 22000 results for USB chargers on Alibaba, only 400 had the weight information. This is useful to have an idea of how many products have the weight information in the e-commerce platform.

2.7 Machine Learning Projects at the Dutch Customs

In this section, past experiences about machine learning carried out by the Dutch Customs Administration are presented. This information has been reported after having interviewed the DCA Senior Advisor for Data Analytics Marcel Molenhuis, and the Data Scientist and Data Analytics Expert Jetze Baumfalk on 18th June 2018 (see appendix E).

According to these experts, the DCA is already deploying machine learning technologies as a decision-making system to choose which of the red and amber flags packages to inspect. Given the result of the risk engine, and the limited number of inspectors, the machine learning model helps to choose what package should be inspected.

This machine learning model is applied thus after the PRISMA/BLAZE risk engine as a de-risking tool. As Jetze Baumfalk explained, this is also necessary from a technical point of view. Because "the model needs the dataset with the inspection results, it can only be applied on those packages that have a history of inspections results", thus only those red/amber flags that have been inspected and "that have the label Y/N anomaly" (Jetze Baumfalk, 18th June 2018, appendix E). Of this dataset, 75% of this data set is used to train the model. 25% is used to test it.

"The result of the machine learning model is a number between 0 and 1 according to the relevance of the risk" (Jetze Baumfalk, 18th June 2018, appendix E). In the end, the final decision on whether to inspect or not is still completely on the targeting officers. Below it is reported the figure provided to us by the DCA experts where they explain the datasets used for the machine learning project. In this figure, the grey area within the "all data" (figure 8) is the declarations with the label of the inspections results, already cleaned and pre-processed.

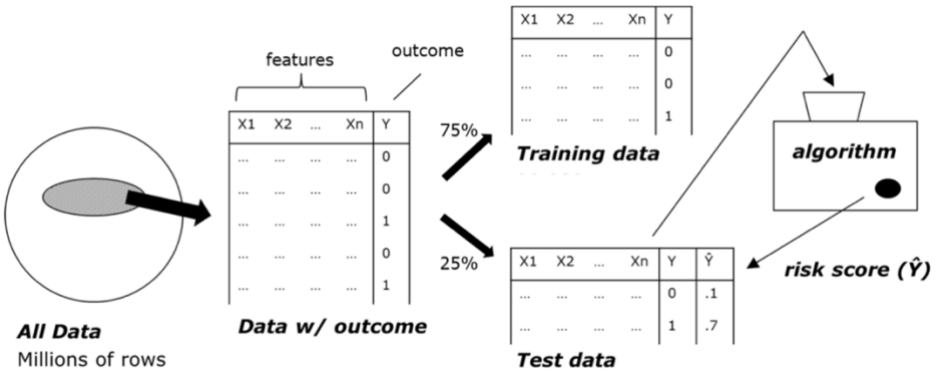


Figure 8: Scheme of the dataset used by the DCA in their machine learning project (DCA, 2018)

The DCA experts reported that usually decreasing the false positive consequently leads also to an increase of false negative. A receiver operating characteristic (RoC) curve is used to see the tradeoff between false positive reduction and false negative increase for any de-risking practice, and thus also to track the accuracy of the machine learning. Given the packages targeted by BLAZE, the machine learning model leaves out some of these packages (de-risking), decreasing the false positive but also increasing the false negative (because fewer inspections are carried out). The RoC curve shows how much decreasing of false positive it is possible to have by allowing an increase of some false negative. In the case of the DCA, if the machine learning model increases the false negative of 5%, it also decreases the number false positive by 10%. Jetze Baumfalk also reported us that without this machine learning model, the average hit rate of rules (PRISMA & BLAZE) is 5%, and the one by the targeting officers is 10%.

The machine learning algorithm used is a random forest, because it is well performing when finding non-linear correlations. The model was created using Python. It can also be loaded in the BLAZE system, part of the risk engine with PRISMA (see sections above). When they deployed the ML model, they had a validation period of 3-month shadow-running, which means that the DCA let the ML model running in parallel with the existing solution, with real data, so that they could compare the actual findings to assess the model. They are using data recorded since 2014.

Jetze Baumfalk explained to us as "the biggest challenge is to track the results of the machine learning model to the declarations parameters so that it is possible to update the business rules of the risk engine" (Jetze Baumfalk, 18th June 2018, appendix E). Finally, because the results of the inspections are free text, it is hard to understand them correctly and label them. Thus, also these data need to be clean and pre-processed, in particular labeling the inspections in a standard way and removing the declarations that have more than one type of product. From a 100% of the dataset that is useful for the machine learning (already labeled), only an 85% was left as good to use.

3 Literature Review

This chapter investigates the state-of-the-art of big data analytics and the techniques that could be deployed to address the use case of the Dutch Customs Administration. This corresponds to the second phase of the design cycle described by Hevner (2004) and reported in figure 1 (section 1.6). It thus answers to the second research question of "what is the state-of-the-art of web-crawling and big data analytics technologies relevant to the web-crawling architecture".

Answering this question, the most suitable big data analytics techniques and methods of web-crawling in the e-commerce domain are brought as useful knowledge to the design phase of the design cycle (phase 3). It consists of two main parts, one for each main topic of this research: big data analytics and web-crawling. The part on the BDA is further broken in two sub-topics, one more theoretical to define the technology, and one more practical where development guidelines are proposed.

Thus, the literature review has been divided into three main parts. Within the BDA part, the new disciplines of artificial intelligence and machine learning will be also explained, and the implementation of machine learning projects becomes the second topic of this literature. In the third and final part, first the web-crawling/web-scraping process is described, and then useful applications of ML in web-crawling and web-crawling in the domain of the e-commerce are investigated.

As it is explained in the table of the research strategies (table 4), the research strategy of this phase is an accurate literature review of the available literature. This is done the following strategy has been used: the topics have been divided into sections, namely "big data analytics", "machine learning" and "web-crawling". For each of this section, a first exploratory research on google and TU Delft library portal has been carried out to identify the most relevant keywords. The relevance has been evaluated by the number of useful results obtained, and by the number of citations of these results. Given this research, the following keywords have been chosen for each part of the literature review:

Table 5: Search Keywords for each Literature Section

Big Data Analytics	**Implementing Machine Learning**	**Web-Crawling**
[big data analytics] [data analytics] [advanced analytics] [advanced data analytics] [big data] [artificial intelligence] [machine learning] [data mining] [pattern recognition] [deep learning] [natural language processing]	[machine learning implementation] [developing machine learning] [machine learning projects] [designing machine learning] [machine learning architecture] [machine learning systems] [machine learning algorithm] [choosing machine learning] [machine learning requirements] [machine learning guidelines] [scaling up machine learning]	[web-crawling] [web-scraping] [web-crawlers] [web-scrapers] [web-crawling system] [web data retrieval] [web information retrieval] [web data mining] [adaptive web-crawling] [smart web-crawling] [intelligent web-crawling] [web-crawling e-commerce] [search e-commerce product] [products look-up]

These keywords and their combinations have been searched in the most famous knowledge databases, such as Scopus, Springer, Elsevier, Emerald, and Google Scholar. Finally, a systematic review of the existing articles, books, and conference papers has been collected.

Given the iterative nature of the design science research, the literature review is not only a preparatory phase at the beginning of the actual study, neither it has to be placed correctly as chapter three, but it is an on-going process which starts at the beginning of the research and continues aside the other phases during the entire duration of the design cycle. At the end of this chapter, the knowledge gap is defined by collecting all the topics of missing literature among the three parts that have been investigated.

3.1 Big Data Analytics

With big data analytics (BDA) is meant the application of specific analytics techniques on big data. It is thus necessary to define the concept of big data. The chosen definition of big data for this research is the one provided by Hu, Wen, Chua, and Li (2014), which explains the difference between big data with respect to traditional data.

Table 6: Differences between Traditional Data and Big Tada (taken from Hu et al., 2014)

Characteristic	**Traditional Data**	**Big Data**
Volume	GB	TB, PB
Generated Rate	Per hour, day	Per minute, second
Structure	Structured	Semi-structured, unstructured
Data Source	Centralized	Fully distributed
Data Integration	Easy	Difficult
Data Store*	RDBMS	HDFS, NoSQL
Access Interactive	Batch	Near real-time

The main difference coming from the name, it is about the higher volume of information in big data with respect to the traditional data, and about the speed of its generation, including that they are provided "near real-time" – i.e. as continuous flow – and not through batches – i.e. pieces of information, discontinues flow (Wu et al., 2014). The other differences are related to the complexity of processing this type of data because they are not organized (structured) and ready to use, and they are distributed in multiple locations. Thus, the integration of big data and its storage require more advanced techniques.

*The term RDBMS stays for a relational database management system, or simply a relational database, and it refers to traditional databases like MySQL for instance, which represent and store data in tables and rows. They're based on a branch of an algebraic set theory known as relational algebra. Meanwhile, non-relational databases are the database that is "not only SQL" (from here the name NoSQL), so programmed with different logic according to the application. The main advantages are that it is possible to scale the database also horizontally and not just vertically, and they have parallel processing capabilities which means it is possible to run jobs in parallel to process large volumes of data.
As said earlier, big data analytics is the application of statistical techniques to analyze and discover knowledge from big data. To understand it more in detail, a presentation of the different types is provided in the following section.

3.1.1 Types of Big Data Analytics

Big data can reinforce the decision-making and enlarge output of the organizations; this became possible through the use of advanced analytical methods applied to extract sense from this data. Sivarajah et al. (2016) present an interesting classification of the main types of big data analytics techniques by their purpose. I summarized them below as:

- Descriptive Analytics: analysis of data to describe or define what is represented by the dataset.
- Inquisitive Analytics: this type of data analysis aims to verify or deny a certain proposition. It is to reject or accept the hypothesis.
- Predictive Analytics: concerned with forecasting and statistical modeling to determine future possibilities.
- Prescriptive Analytics: it is about optimization and randomized testing to provide advises on a certain topic.
- Pre-emptive Analytics: it is data analysis aiming to take precautionary actions against negative scenarios.

Sivarajah et al. (2016) map these methods just explained in the following figure 9:

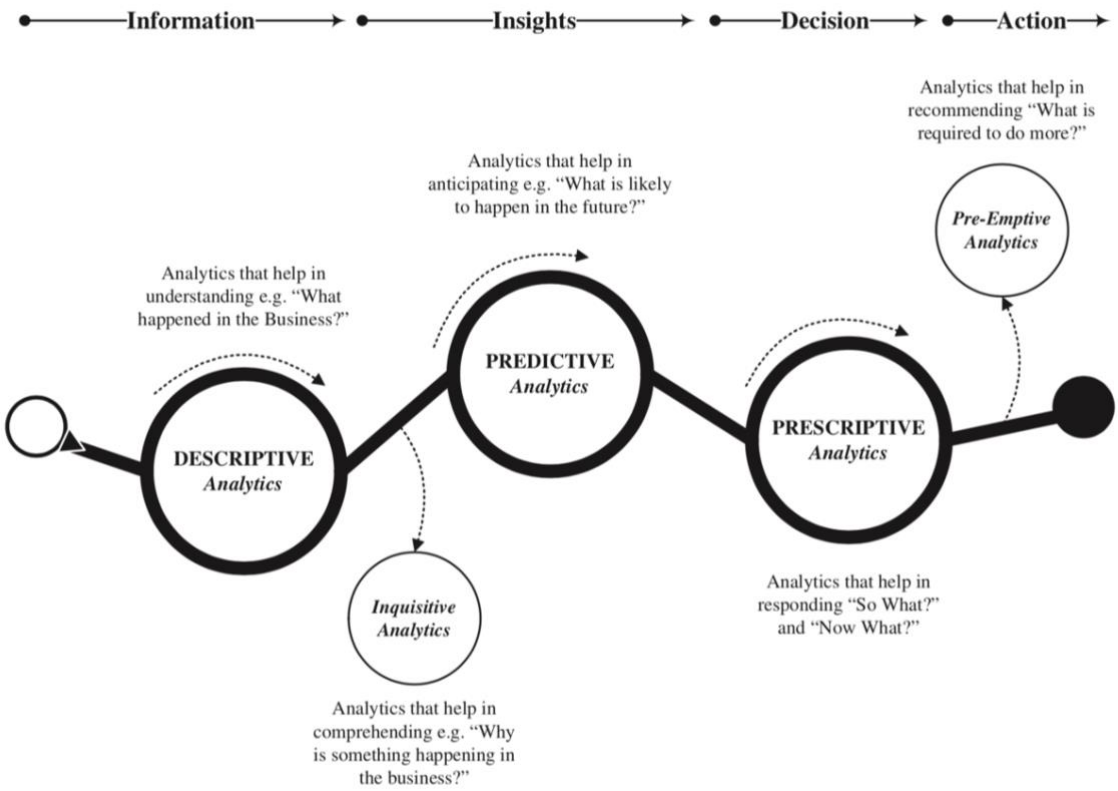


Figure 9: Classification of types of Big Data Analytics Methods (taken from Sivarajah et al., 2016)

Of these five categories, it is appropriate to analyze the main three (biggest circles in figure 9) more in detail. Descriptive analytics is the simplest form of BDA method and involves the summarization and description of knowledge patterns using simple statistical methods, such as mean, median, mode, standard deviation, variance, and frequency measurement of specific events

in BD streams (Rehman et al., 2016). Part of descriptive analytics is also reporting, dashboards, scorecards, and data visualization. These techniques are about explaining a dataset, which in business means, for instance, to monitor a process by analyzing its description over time by setting monitoring metrics. In this sense, descriptive analytics are considered backward-looking and revealing of what has already occurred. Most of the BDA techniques are descriptive (exploratory) and use descriptive statistical methods also known as data mining tools.

Predictive analytics is at the contrary focused on forecasting and determine future possibilities through statistical modeling. These techniques are based on advanced statistical methods which seek to discover patterns and relationships among data. For this reason, Gandomi and Haider (2015) associate the predictive techniques with regression techniques and the new trend of analytics of machine learning, since they all aim to predict the future by analyzing current and historical data. The following section addresses the field of machine learning in detail.

Prescriptive analytics focuses on investigating cause-effect relationships in order to provide advises on different topics. It usually concerns problems of optimization or decision-making. A direct business application is the deployment of these techniques to answer strategy questions. Because of its vast venerability, there are very limited examples of good prescriptive analytics in the real world (Sivarajah et al., 2016).

Gartner (2017) reports these types of analytics in a graph with the value a technique brings to an enterprise on one axis and the complexity of developing such a technique on the other:

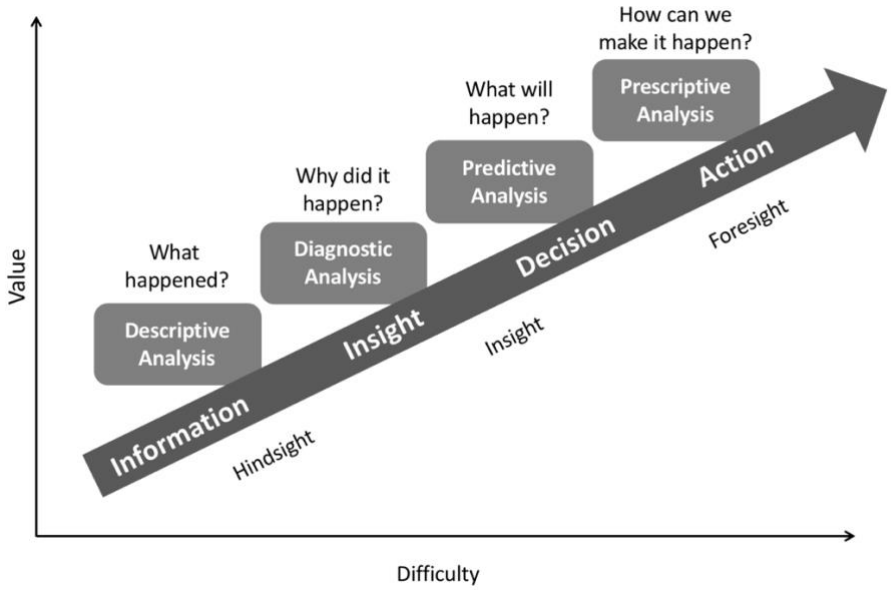


Figure 10: Big Data Analytics Methods by Complexity and Value (taken from Gartner, 2017)

In this graph, Gartner (2017) names "diagnostic" the type of analytics that Sivarajah (2016) calls "inquisitive". As one would expect, the more the complexity of the technique, the more the value that it adds to the company.

3.1.2 Big Data Analytics Value Chain

This framework presents a value chain for big data analytics broken into four stages (generation, acquisition, storage, and processing), together with a technology map that associates the leading technologies in this domain for each of this stage. Through this framework is thus immediate put

in relation to the two main concepts of this master thesis: big data analytics and web-crawling. Since the aim of this research is to design an architecture of a web-crawling system which deploys big data analytics techniques to address the DCA issue within the e-commerce risk targeting, this framework is used to position the web-crawling technology within the field of big data analytics. In particular, it places the web-crawling as technology to be deployed within the phase of "data acquisition" of a BDA project.

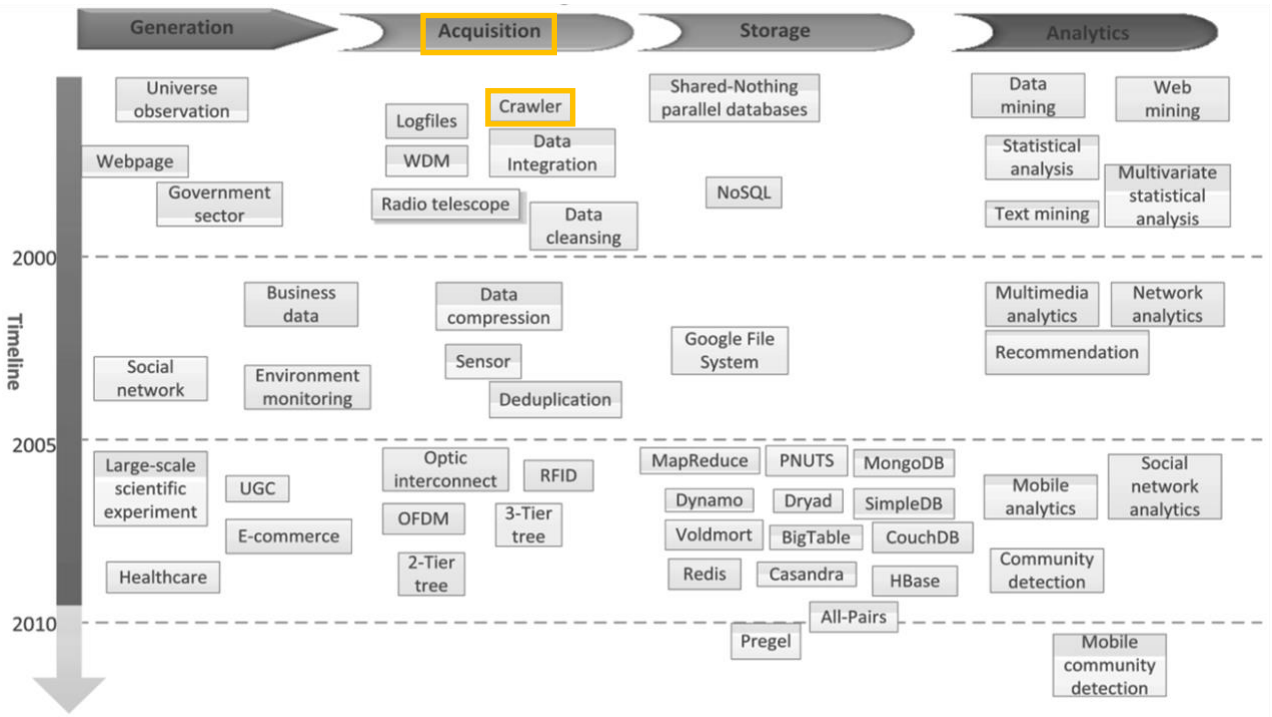


Figure 11: Big Data Analytics Value Chain (taken from Hu et al., 2014)

The framework adopts a systems-engineering approach to describe a big data analytics project. It divides a typical big data system into four consecutive phases. These are namely data generation, data acquisition, data storage, and data analytics. It has to be noticed that there is not a phase of data visualization, but it is considered part of the data analysis phase, differently from other frameworks where it is considered a separate phase, for instance, the one of Curry et al. (2014). Let's go through each phase. The first stage is data generation, which simply concerns how data are generated.

The data acquisition phase comes after and refers to the process of obtaining information and is subdivided into data collection, data transmission, and data pre-processing. This process starts with acquiring raw data and then checking whether this data is meaningful or not. Collecting meaningless data unnecessarily increases the amount of storage and resources that must be deployed to conduct the analysis. Thus, data pre-processing operations must be done to eliminate data redundancy and filtering out useless information (Hu et al., 2014).

Data storage is the next phase. It is about the efficient storing and management of large datasets. This phase is a challenge for both the aspect of hardware, where always more complicated IT infrastructure is needed to support big data activities and software, as advanced algorithms and file systems are required.

Finally, data analytics leverages analytical methods or tools to inspect, transform, and model data to extract value. Six critical technical areas can be identified: structured data analytics, text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics (Hu et al.,

2014). The analytics phase will be explained in detail further in this research, as it represents the core of the technological solution for the problem under analysis.

This last phase of the big data analytics value chain provides another way of categorizing BDA techniques by considering the type of application, instead of the purpose of usage. Hu et al. (2014) suggest six types of applications organized by data type: “structured data analytics, text analytics, web analytics, multimedia analytics, network analytics, and mobile analytics”. For the use case of this research, the most relevant are web analytics and text analytics.

Web analytics concerns those techniques that allow the retrieval, extraction, and evaluation of information from the Web. This includes the fields of information retrieval and web data mining, which is divided itself into three categories: web content mining, web structure mining, and web usage mining (Chen & Chau, 2005). These topics will be better addressed in the next sections about web-crawling. Similarly, text analytics or text mining refers to the process of extracting useful information from unstructured text. This is the field also of the recent AI technology of Natural Language Processing (NLP). An accurate section on NLP is provided more ahead in this chapter (section 3.1.6).

Analyzing these BDA techniques, the terminologies of data mining, machine learning, AI and NLP came out as the latest development in the field. In the next sections, these terminologies will be explained and organized. Since an accurate description of the entire existing literature on the topic would be too ambitious because of its complexity and volume, I will only describe its history and development, addressing more in details specific topics that might be useful in the use case of this research.

3.1.3 Machine Learning

Machine learning is sometimes defined as a subset of data mining – meant as the computational process of discovering patterns in large data sets – which itself is a subset of data analytics. It can be defined as that field of computer science that uses algorithms coming from the discipline of statistics to give computers the ability of "learning". This happens through the analysis of data and leads to the progressive improvement of the algorithm's performance on a specific task without the need to be explicitly programmed. Simon already in 1983 was defined as machine learning any process where a system improves its performance. Few years after in 1997, Mitchell defined a machine learning algorithm as "any computer algorithm that improves its performance at some tasks through experience." In addition to data mining, the machine learning is also considered to be really close to the field of pattern recognition which as the same word explains, focuses on the recognition of patterns meant as regularities in data (Ivanovic & Radovanovic, 2015).

In these terms, this is a radical change in addressing IT problems. In the conventional approach, software programs are hard-coded by developers with specific instructions for the tasks that need to be executed. This can work well in most of the cases, but it has big limitations. It assumes that the human programmers can imagine every scenario and code instructions for any possible state of the world. But If the environment changes in an unpredicted state, the hard-coded software will not work well anymore and will stop working. By contrast, the idea of the machine learning approach is that ideally, it is possible to create algorithms that "learn" from data automatically. Thus, in case of changes in the environment, they can adapt to the new circumstances without needing to be explicitly programmed by human programmers. The idea is to give these algorithms "experiences" (training data) and a general strategy for learning, and finally let them identify patterns, associations, and insights from the data. In short, machine learning systems are trained instead of programmed.

As this topic is new and "hot" at the moment, I could not find a generally accepted literature about the machine learning field, neither on its exact definition nor about its relationship with similar topics such as data mining or pattern recognition, or even artificial intelligence. But In this section of this literature review, I integrate the sources I analyzed to make an order in this discipline. After having defined the relationship between machine learning and data mining and pattern recognition, where does artificial intelligence should be placed with respect to machine learning?

Artificial intelligence is the concept of intelligence demonstrated by machines, in contrast with natural intelligence which is characteristic of humans and animals. In computer science, AI can be defined as the study of "intelligent agents" which are devices able to perceive the surrounding environment and taking actions to maximize the possibilities of achieving their goals. Thus, an artificial intelligence technology must satisfy the characteristics of intelligence, which has been defined as the capacity for logic, understanding, self-awareness, learning, emotional knowledge, reasoning, planning, creativity, and problem-solving. If a technology has just some of these capabilities, it can be defined partially intelligent.

To recap what it has been described so far, I propose the following conceptual map to describe how the different disciplines are connected (figure 12). As mentioned earlier, I could not find a generally accepted definition of machine learning or its relationship with the other disciplines of big data analytics. Thus, I derived this scheme from the literature I analyzed during this research and should be considered as the personal view of the author.

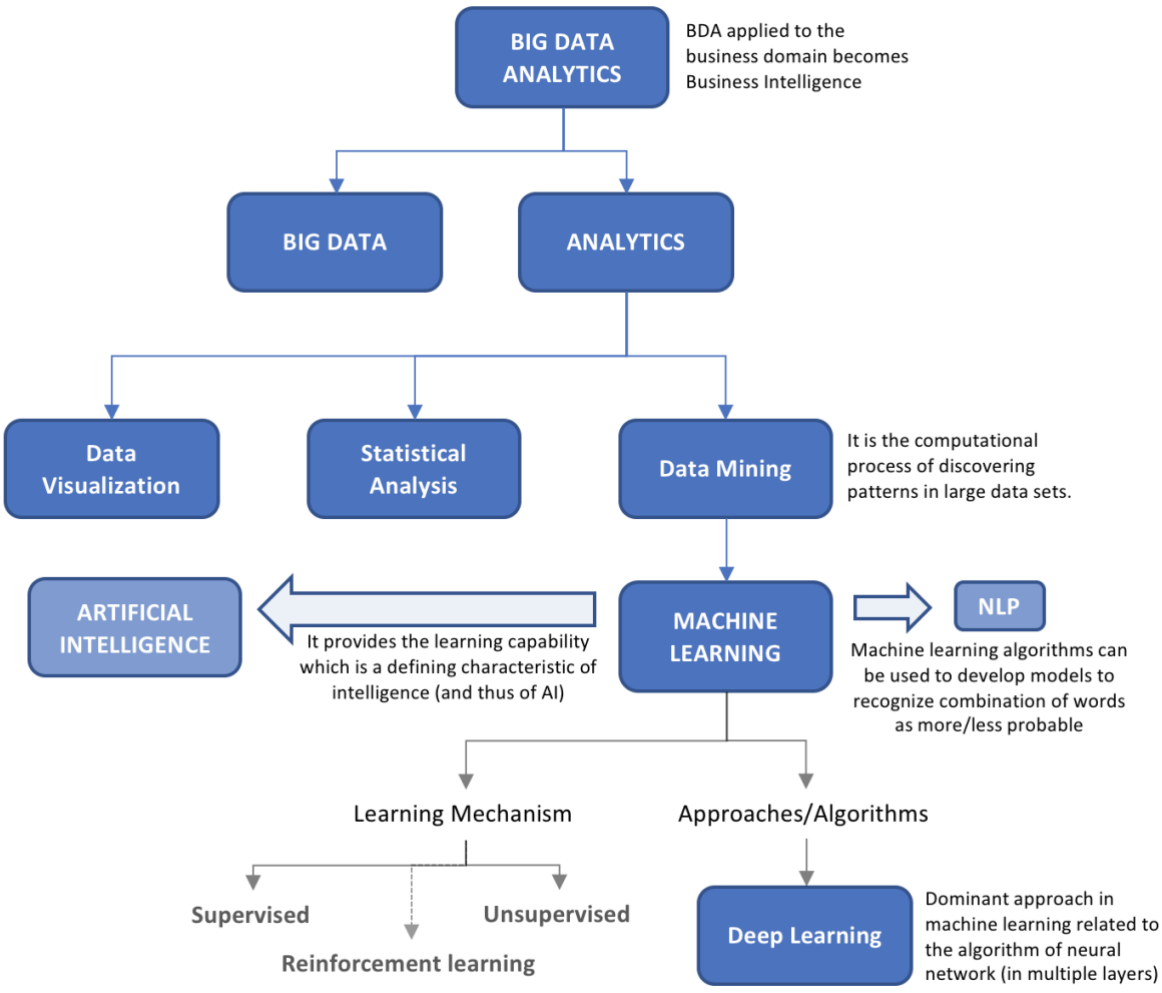


Figure 12: Relationship among the fields of Big Data Analytics, Machine Learning, and AI

As it is shown in this figure, machine learning algorithms are categorized according to their learning mechanism. In most of the cases, they are trained from labeled "training" data, where each input as a pair output. An algorithm working with a labeled dataset is known as supervised learning, and its goal is to predict the output values of new examples, based on their input values (Chen & Chau, 2005). When no labeled data are available instead, it is referred to as unsupervised learning. In this latter case, the training examples contain only the input patterns (without any explicit output associated with each input) and the learning algorithm needs to discover the values of output by generalizing rules from the input data. Unsupervised learning can be a goal in itself – as given a set of data the algorithm discovers and learn hidden patterns – or it can be a mean towards a specific goal – and in this case, it is known as feature learning (Ivanovic & Radovanovic, 2015). In any case, the unsupervised learning begins with the exploration of the data, usually carried out through clustering algorithms which understand the dataset by dividing it into classes.

This distinction between supervised and unsupervised learning is key to understand machine learning applications. In a few words, supervised learning is about classification problems, while unsupervised learning is about clustering problems, or also pattern recognition (as a discovery of the data). This is because in the first case, the task is to classify textual documents into predefined categories. The fact that these categories are predefined means that they are known, and thus that the machine learning model is provided with training examples which defined them. On the contrary, text clustering groups documents into categories defined dynamically on the basis of their similarities. The algorithm receives a dataset which has to explore and categorize (make categories within the dataset) according to its understanding of the data.

Another maybe more intuitive way to understand this difference between supervised and unsupervised learning is to check whether there is a learning "feedback" available to a learning system or not. Thus, in supervised learning, the examples of inputs and desired outputs are like feedbacks given by a "teacher", as they show the correct mapping of inputs to outputs. When this input feedback is only partially available or restricted to special feedback, supervised learning can be further classified in semi-supervised learning – if the training dataset/feedback is incomplete – or active learning – if the training labels are limited to a set of instances (based on a budget). In this latter case, the algorithm has to optimize its choice of objects for which acquiring labels. Finally, another approach considered almost unsupervised learning, but not completely yet, is the so-called reinforcement learning, where the training data (in form of rewards or punishments) is given only as feedback to the program's actions in a dynamic environment. An accurate section on reinforcement learning is provided more ahead in this chapter (section 3.1.5).

Besides the learning mechanisms, machine learning has different approaches, or also called methods or techniques, which themselves have different algorithms. These can be defined as indeed the approach used to solve a given machine learning problem. For instance, a classification problem can be addressed using a decision tree approach, which means to split the classification question into different sub-problems. In every node, the model chooses the best split among all features in order to maximize a certain function. Then, there are different algorithms of decision trees. For instance, the algorithm "random forest", used by the DCA, is one of these. Another approach that became popular recently is the so-called artificial neural network (ANN), and further the deep learning approach, which tries to imitate the neurons in the human brain.

3.1.4 Deep Learning

As I introduced it in the previous section, deep learning is a frontier area of research within machine learning which uses artificial neural networks with many layers, hence the label “deep” (LeCun, Bengio, & Hinton, 2015). Falling hardware prices and the development of GPUs for

personal use in the last few years have contributed to the development of the deep learning approach since the training of multi-layered neural networks requires a huge computational power and complexity. Data scientists working in this field have recently made breakthroughs that enable machines to recognize objects and faces, to beat humans in challenging games such as chess and Go, to read lips, and even to generate natural language.

But what is an artificial neural network (ANN)? It is a machine learning method which gets its inspiration from how the neurons in the human brain. A neural network is a graph of many nodes (neurons) connected to each other through weighted links (synapses), also called edges. The signals transmitted through a connection between artificial neurons (nodes) are real numbers, and the output of each artificial neuron is computed functions of the sum of the inputs.

The weights of each link change and adjust as the learning process continues, and this is how the ANN algorithm learns. These weights are responsible to increase or decrease the influences of the signals when passing through the links. Artificial neurons may have thresholds such that the signals are sent only if they cross these thresholds. While in the case of the decision trees knowledge is represented by an organized structure of questions, for ANN knowledge is learned through the network of interconnected neurons, weighted synapses, and threshold logic units (Lippmann, 1987; Rumelhart, Hinton, & McClelland, 1986).

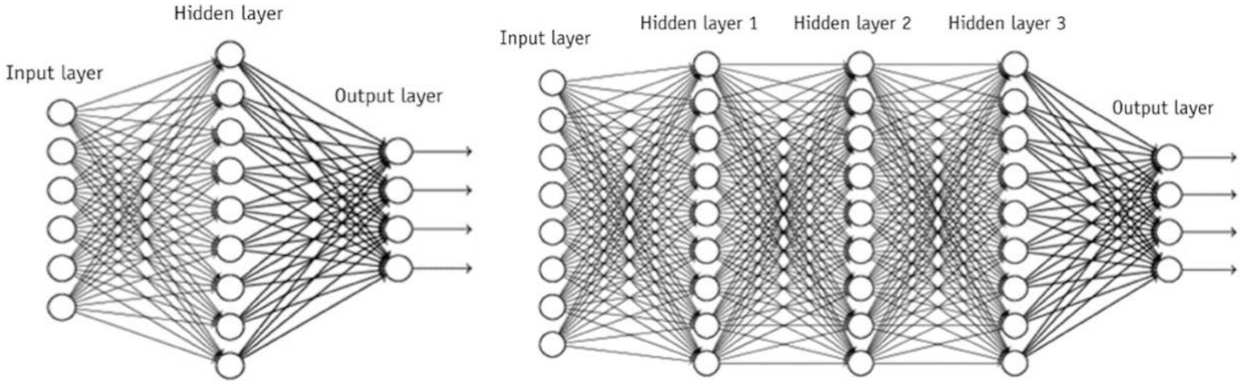


Figure 13: Artificial Neural Network and Multi-layered ANN (taken from Nielsen, 2018)

ANN became popular with the development of the so-called deep learning approach, which consists of using multiple hidden layers in an artificial neural network, as it is shown in figure 13 above on the right. The difference with a single layer ANN is that each layer adds its own level of non-linearity that cannot be contained in a single layer. Each layer's inputs are only linearly combined and hence cannot produce the non-linearity that can be seen through multiple layers.

Among the numerous types of neural networks that have been developed, maybe the most commonly used is the so-called feed-forward-hack-propagation model. Backpropagation networks are fully connected, layered, feed-forward networks. In the beginning, the network has a set of random weights and after each training example, it adjusts its weights.

The nodes are activated when learning examples are input into the network. The final output of the network is compared with the desired output and the error deviations are sent back as input to the input and hidden layers. According to these errors, the network is able to update the weights information until the network gets stable (low error deviation). ANN can be used for pattern recognition, clustering, or unsupervised learning, for instance using the Self-organizing Maps (Kohonen, 1995).

In the case of image recognition, for instance, convolutional deep neural networks showed to perform the best. In this case, the nonlinearities are represented by convolutional and pooling layers, capable of capturing the features of images. This approach is also successful when working with text analytics. For this purpose, recurrent neural networks (RNN) are well performing (Russell, Norvig, & Davis, 2010). RNNs are of two different types, namely long-short term memories (LSTM) and gated recurrent units (GRU). These are described more in detail in the next section.

3.1.5 Natural Language Processing

Natural language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation. In our case, it is relevant only natural language understanding (NLU), since we do not have to generate any text, but only process products descriptions in e-commerce platforms. NLU can be further divided into five domains: phonology (sound), morphology (word formation), syntax (sentence structure), semantics (meanings) and pragmatics (context) (Chomsky, 1965).

The standard technology used before the development of the NLP is the parsing analysis. Parsing, syntax analysis is defined as the process of analyzing a string of symbols, either in natural language, computer languages or data structures, conforming to the rules of a formal grammar. For instance, HTML parsing is the analysis of the HTML tags that structure the web pages (Martin, J., 2004). More recently, with the development of machine learning, it is possible to use machine learning models to predict the next word that should be placed in a sequence (language generation) or understanding what product a given piece of text is describing (classification), which is the relevant use case for this research.

In the field of NLP, it is important the distention between machine learning models generative or discriminative. Generative methods create rich models of probability distributions. Discriminative methods have posterior estimating probabilities and are based on observations (Khurana, Koli, Khatter, & Singh, 2017). In other words, with an input data x to classify into labels y, a generative model learns the joint probability distribution, while a discriminative model learns the conditional probability distribution (the probability of y given x). An example of discriminative methods is Logistic Regression, while a generative method is Naive Bayes.

According to Ng & Jordan (2002), overall, discriminative models generally outperform generative models in classification tasks. For this reason, an algorithm of this type will be probably chosen during the development of the web-crawling system at the DCA. In addition, neural networks can be used as both generative or discriminative. In particular, for text classification, it has been noticed as the Long-Short Term Memory Neural (LSTM) Network perform well (Yogatama, Dyer, Ling, & Blunsom, 2017).

LSTM networks are a special kind of Recurrent Neural Network (RNN), an evolution of the standard neural networks with loops in them, allowing information to persist. In particular, LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior. They were introduced by Hochreiter & Schmidhuber (1997) and were refined and popularized by many people in the following work.

3.1.6 Reinforcement Learning

Finally, in this section, I want to come back to talk about Reinforcement learning (RL) which is an area of machine learning inspired by behaviorist psychology concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. The main difference with the machine learning methods described earlier is that a reinforcement learning algorithm does not need to know the correct input/output pairs, but instead, it learns by balancing the exploration of an uncharted territory and the exploitation of the current knowledge. For this reason, it is said that RL algorithms use the resources efficiently (Sutton & Barto, 2015).

In other words, the setup of Reinforcement Learning consists of two elements, an agent, and an environment. The environment is the space where the agent acts, whereas the agent is the algorithm of reinforcement learning. The agent takes actions according to its current state and knowledge of the environment. In return, the environment responses with the next state and reward to the agent in case of positive action. The agent updates his state and knowledge, and through the rewards learns the right actions (Sutton & Barto, 2015).

The environment of the reinforcement learning algorithm can be model-free or model-based according to whether the environment is described by a model of its dynamics or not. A model-free algorithm relies on trial-and-error to update its knowledge. As a result, it does not require space to store all the combination of states and actions. Another differential characteristic among reinforcement learning models is on-policy or off-policy. An on-policy agent learns the value based on its current action a derived from the current policy, whereas its off-policy counterpart learns on the basis of the action obtained from another policy (Ivanovic & Radovanovic, 2015)

In the use case under of the e-commerce risk targeting and products identification, the web would be the environment, or more, in particular, the e-commerce platforms, and the agent would be the web-crawler looking for the product. When it finds a product that matches the item description in the declaration, it gets a reward. However, this approach could be difficult to realize because it requires a perfect knowledge of the web, which is the environment, and might require to develop an own index of the web, which is time consuming and maybe not feasible in the use case of this research. This topic will be better explained when describing the web-crawling process (section 3.3).

3.2 Implementing Machine Learning

In this section of machine learning, I want to explore if there are any guidelines in the existing literature that could be useful in implementing projects of machine learning. This ranges from analyzing the main challenges of implementing such projects and scaling them up, choosing the right algorithm, and finally investigating the architectural demands that the machine learning technology requires. First of all, I earlier described the many existing approaches to machine learning and explained as each approach has different algorithms. Then following natural question is thus, how to choose the most appropriate algorithm for a considered problem? The next section tries to address this question and give the answer offered by the current literature.

3.2.1 Algorithm Choice

Choosing the right algorithm may be very complicated: there are dozens of machine learning algorithms considering both unsupervised and supervised, and each has a specific approach to learning. The best method or unified fit do not exist. Identifying the most appropriate algorithm is done most of all through trial and error. Even qualified data scientists cannot determine a priori whether the algorithm will work well without testing it. Flexible models are likely to overfit data by simulating minor changes that can be interference. Simple models can be easily interpreted but may have a lower accuracy (Oladipupo, 2010). Thus, to choose the right algorithm it is necessary to trade one advantage against the other, including speed, accuracy, and complexity. Trial-and-error is thus the main practice to choose the most appropriate machine learning method: if one algorithm/approach does not work, try the next one.

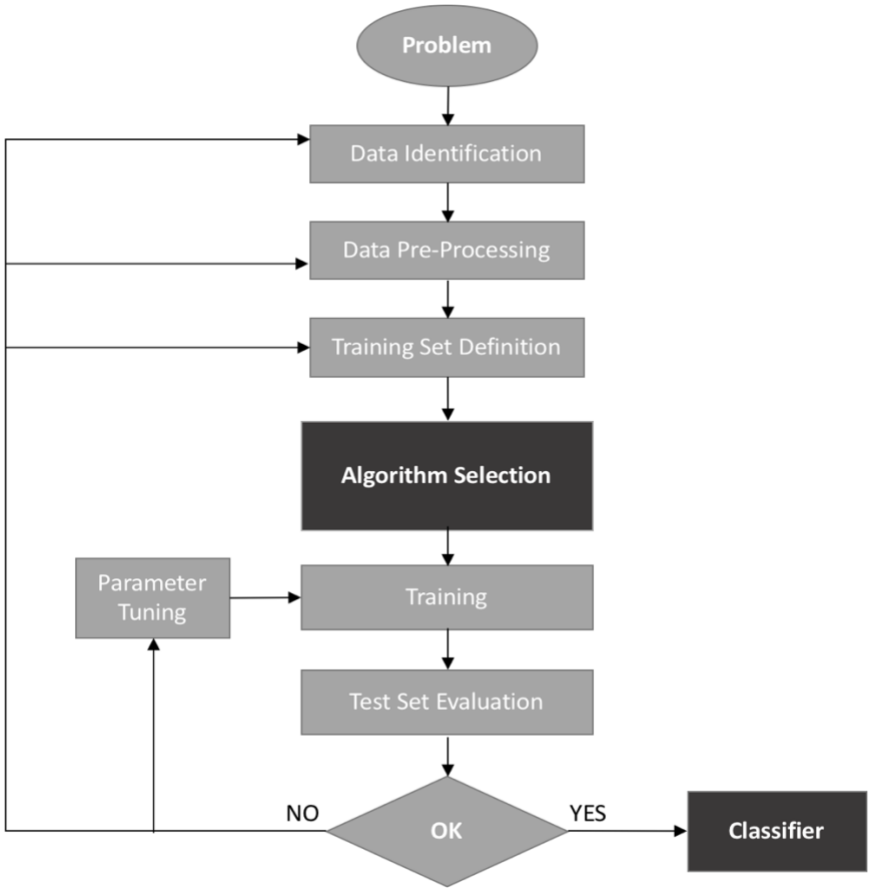


Figure 14: Trial-and-error for choosing a Classification Algorithm (taken Oladipupo, T., 2010)

Besides this premise, it is possible to know the group of algorithms or the machine learning approach that would address a considered problem. As I mentioned earlier, the first distinction is about the data set, thus supervised or unsupervised learning, and then about the type of problem: classification, regression or clustering.

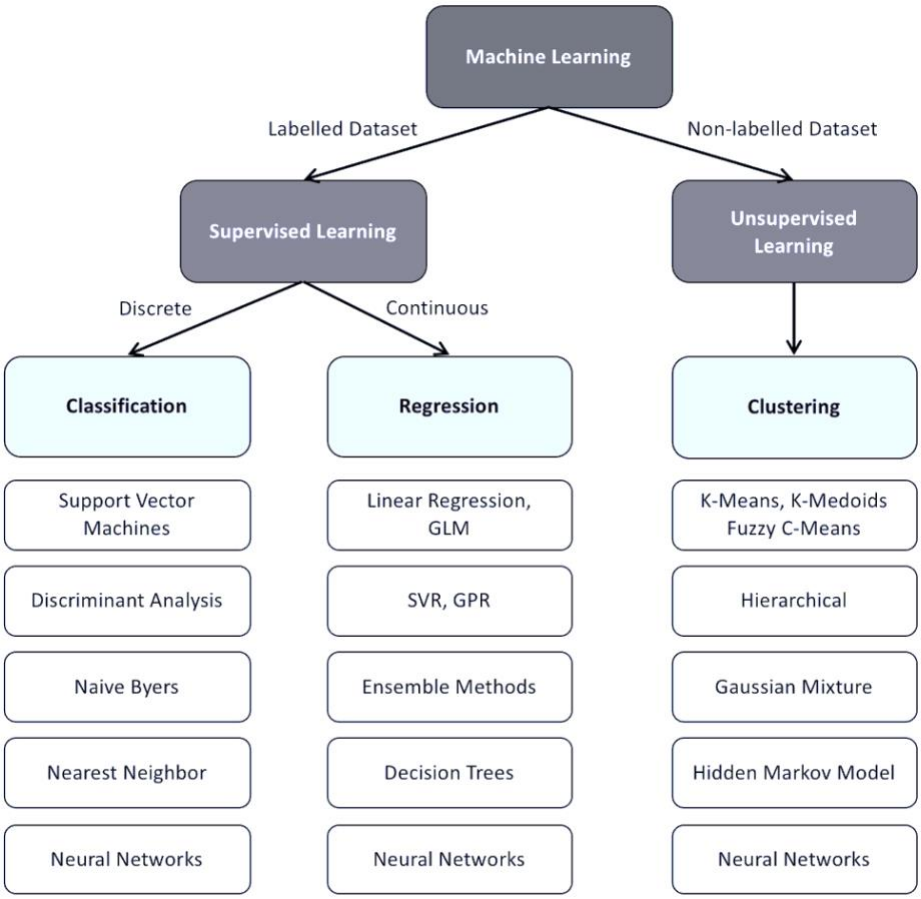


Figure 15: Classification of Machine Learning Techniques (taken from MATLAB & Simulink, 2018)

The figure above shows that the labels in supervised learning can be discrete or continuous, which are handled by classification and regression algorithms respectively. Classification is used mostly for prediction, pattern recognition, and outlier detection, whereas regression is used for prediction and ranking. Unsupervised instead are generally clustering algorithms.

The picture also shows that the random forest algorithm, used by the DCA and within the algorithms of issue trees, is a technique with mostly regression purposes, but that it can also be used for classifications. In particular, the random forest technique operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the mean prediction of the individual trees.

Once the right algorithm has been chosen, there is still the problem of whether this algorithm will continue to outperform its competitors when passing from prototyping to the full implementation of the project, which usually means a higher volume of data, different datasets, and more constraints to take into account. In the next section, I want to investigate this issue and the solutions proposed by the current literature.

3.2.2 Production Scale Analytics

In this part of this more practical section of the literature review, I want to investigate the issues and challenges of scaling up machine learning and BDA techniques. What happens when from the prototype it becomes a full production system? Do the same algorithms work the same way with higher volumes of data? This is a relevant problem because the design choices that are taken now must continue to work also for future implementations and full-scale systems. Otherwise, it would be useless to build this prototype and research project.

From the interview with the IBM experts, I have been told that usually machine learning projects are developed firstly looking for the right algorithm, then testing the scaling up and finally implementing the details. In the case of this research, a similar developing plan has been drawn (see appendix j). In this case, first, the pure web-crawling system will be developed. Then the machine learning components identified in the previous section (3.2.3) are developed. Here the machine learning algorithms will be tested on samples datasets – e.g. data on e-commerce platforms about a specific product – and the most appropriated ones will be chosen. When this is done, finally the supporting database for a better operationalization will be connected.

Unfortunately, about this topic, the literature available on the topic is not sufficient. The only guideline that can be taken is to use the trial-and-error approach also considering the scalability. Thus, when choosing the right algorithm, this should be tested also with bigger datasets, which do not have to be real data but can be an example used just to test the scalability of the algorithm.

3.2.3 Machine Learning Common Challenges

In this section the common problems of machine learning are provided: the bias-variance trade-off, under/over-fitting, high dimensionality, and big data.

When deploying supervised learning algorithms, the error that an algorithm makes can be broken down into three components: bias, variance and irreducible error (K.-Z. Huang, 2008). While the last component cannot be controlled, the first two can be influenced by tuning the algorithm parameters. Bias is about how consistently the model is "right" or "wrong," compared to the truth. On the other hand, the variance expresses how "smooth" the model is: larger variance indicates that small changes in the dataset can lead to radical changes of the outcomes. Usually, to try to increase the accuracy of a supervised learning model, it is possible to reduce the bias, but this will also tend to increase the variance and vice versa. Thus, the ultimate goal is to find the optimal balance between the two variables of error.

Overfitting is also related to the bias-variance trade-off within supervised learning and refers to the idea that a machine learning model can be trained "too much" so that its optimal performance on the training set may result in suboptimal performance on a separate test set and real-life data (Ivanovic & Radovanovic, 2015). This is because the model became overcomplex compared to the reality. It may be a consequence of a small or large number of training instances, noisy data, and/or high dimensionality. Some algorithms are more naturally prone to overfitting than others, and many of them have already complex strategies to avoid it in their formulas. On the other hand, underfitting is the opposite extreme, where the derived model is too simple compared to the reality, and thus not able to accurately predict the right outcome in real situations. In this setting, the variance is low since the model is simple, but the bias is high.

The last common challenge when developing machine learning systems is called high dimensionality. Often datasets have a large number of rows – representing the instances – and/or a large number of columns – representing the features of the model (Ivanovic & Radovanovic,

2015). High dimensionality refers to the high number of columns since the rule of thumb is to have at least 5 training examples for dimension. With a fixed number of training samples, the performances of an algorithm first increase as a number of dimensions/features used increases but then decreases sharply. Because of these phenomena, it is required to be careful to the right proportion of training data and number of features to consider.

Finally, big data and its processes are the main challenges for machine learning projects and in general big data analytics projects. For this reason, the next section describes these challenges in a systematic manner.

3.2.4 Big Data Challenges Framework

Sivarajah et al. (2016) propose his big data challenges framework to classify the main obstacles that are faced when developing projects that deal with big data (figure 13). This section describes each of this challenge in a systematic manner.

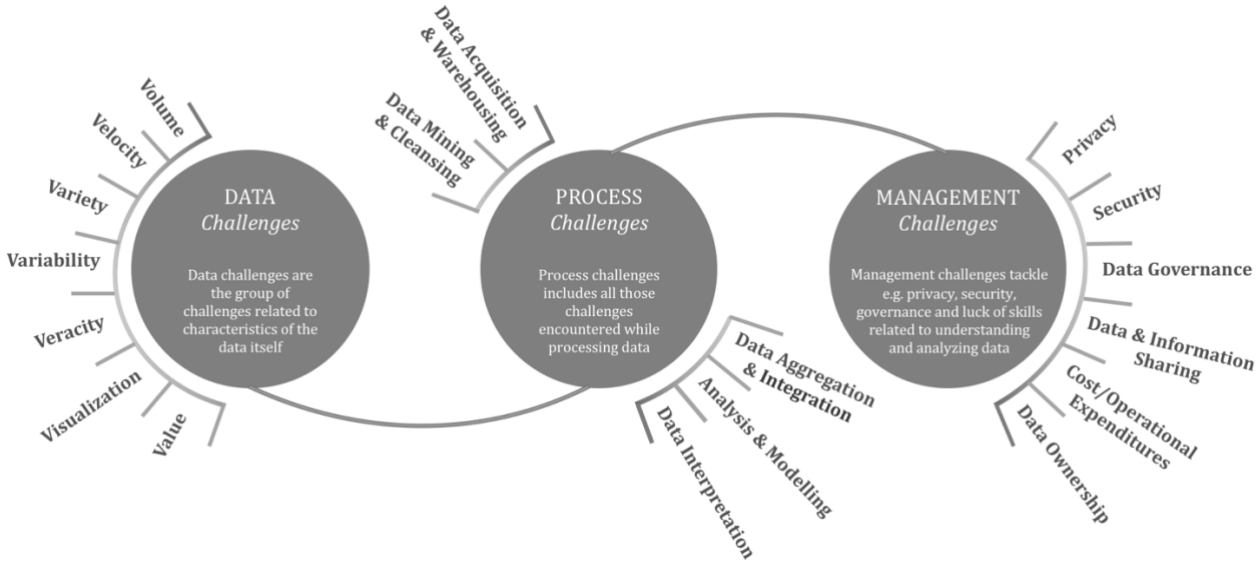


Figure 16: Big Data Challenges Framework (taken from Sivarajah et al., 2016)

In figure 13 above framework for the main challenges of Big Data are presented. The framework is the result of various research studies, which not only addresses these key challenges but also explores opportunities for novel theories or emerging applications. According to Sivarajah, Kamal, Irani, and Weerakkody (2016), it is possible to group the broad challenges of BD into three main categories in the data lifecycle:

- Data challenges that are related to the attributes (characteristics) of the data itself (e.g. data veracity, variety, volume, velocity, discovery, quality, volatility, and dogmatism).
- Process challenges address a series of how techniques. The main question here starts with "How to…": How to capture, integrate, or transform data? How to choose the suitable model and how to provide the results of the analysis?
- Management challenges cover mostly management aspects, such as security, governance, privacy, and ethics.

As it was previously stated, data challenges address the group of the challenges with the characteristics of the data itself. However, the understanding of which data characteristic should be taken into consideration varies in different studies. Shah et al. (2015) state the importance of 3V (velocity, volume and variety), Liao et al. (2015) claim it to be 4V (adding variability to the previous list), by the inclusion of veracity and value we will get 6V framework of Gandomi and Haider (2015). Among all these versions, the 7Vs framework has been taken as a referral point for the data challenges. The seven data characteristics are variety, volume, veracity, value, velocity, visualization, and variability.

Table 7: Data Challenges Description

Volume	Collecting, cleaning, storing, analyzing large scale of data (because of the computing power, storing capacity, time to process and analyze, etc.).
Variety	Comprehending and managing heterogeneous data in multiple formats with unstructured and structured text / image / multimedia content / audio/ video/ sensor data/ noise.
Veracity	Understanding integral discrepancies among data such as inconsistency, anonymities, increasingly complex structure, or imprecision.
Value	It is about deriving value or knowledge from large amounts of unstructured and structured data without its loss for end users. It might be that lot of data are available but extracting knowledge out of it is complicated.
Velocity	It is requisite to manage the high influx rate of non-homogenous data (Sivarajah et al., 2016). This is mainly because it needs evidence-based planning and real-time analytics (Lu, Zhu, Liu, Liu, & Shao, 2014).
Visualization	Presenting the data in a readable manner can be a challenge, especially for big data. Visualization is about an ability to present key information and knowledge more effectively and instinctively by using different visual formats. As an example, it might be the pictorial or graphical layout.
Variability	It is when the meaning of data is constantly and rapidly changing (to not be confused with the variety which is the challenge of having different types of data). For instance, the same word might have completely different meanings depending on its' context. In order to conduct a proper analysis, context should be understood by the algorithm that should decode the exact meaning of a word in a particular context (Hu et al., 2014).

While analyzing and processing the data another group of challenges might be faced. It might happen at any stage of the process starting with capturing the data and to presenting and interpreting the results. Some of the challenges related to data processing might be arranged into five steps. They are data acquisition and warehousing, data mining and cleansing, data integration and aggregation, data analysis and modeling, and data interpretation. From the literature, data mining and cleansing proves to be a vital step in the processing of high-volume unstructured data.

Table 8: Process Challenges Description

Data Acquisition & Warehousing	This step is about acquiring and storing data, which was gathered from different sources and needs to generate value while being stored. For the purpose of capturing valuable and related information, there is a need for smart filters. They should be intelligent and robust to capture only that

	information that would be useful and does not contain inconsistencies or imprecision. It is necessary to have efficient analytical algorithms to understand the provenance of data and make clear the process for the vast streaming data, in addition to reducing data prior to storing (Zhang et al., 2015).
Data Mining & Cleansing	Due to its diverse, interrelated, vibrant, strident and unreliable features, the mining, cleansing, and analysis prove to be very challenging (Chen et al., 2013). For the meaningful use of this huge data, an extraction method is needed. This method should mine out the necessary information from unstructured BD and articulate it in a structured and standard, easy-understandable form. Labrinidis and Jagadish (2012) admitted that the process of development and maintenance of this extraction method is most of the time a continuous challenge.
Data Aggregation & Integration	The challenge addresses the process of aggregation and integration of the clean data which was mined from large unstructured data.
Analysis & Modelling	This process challenge is about delivering the business value through the data analysis.
Data Interpretation	It is about the process (not the data itself) of visualizing data and making it comprehensible for users so they can interpret the findings and extract sense and knowledge.

These are challenges related to Big Data are a group of challenges that are encountered while managing, accessing, and governing the data.

Table 9: Management Challenges Description

Privacy	The prime challenge for BD in the digital age is concerns towards privacy and ways to preserve it.
Security	Lu at al. (2014) identified security as the major issue. He argues that in case security challenges cannot be appropriately addressed then the BD as a phenomenon will never receive great acceptance globally.
Data Governance	Categorizing, mapping and modeling the data same as it is captured and stored is a significant challenge in the process of governing BD. It happens due to the complex and unstructured nature of data.
Data & Information Sharing	For the distant organizations sharing data and information is a challenge. The biggest question is how to make sure not to cross the fine line between BD collection and usage together with guaranteeing the privacy rights of the user.
Cost/Operational Expenditures	The data-intensive operations of handling a massive amount of complex data result in high storage and data processing costs. In this sense, the most emerged challenge is cost minimization.
Data Ownership	Ownership of data is a complicated issue. Claiming ownership for data presents a continuing and critical challenge. Who owns that data? It is not always easy to agree with it. Data ownership might be considered to be a deep social issue.

Being the online information and data the focal point of the web-crawling technology to be developed during the PROFILE project, the big data challenges framework could be used to be aware of the most common challenges when approaching a BD project. In addition, given these challenges described, the requirements analysis must provide the necessary resources to

overcome these challenges. In this sense, the big data challenges framework is used to guide the requirements analysis, and thus the initial phase of the architecture design.

As the next chapter addresses the design of the technology to be developed starting from its requirement analysis, it is important to investigate an innovative practice to the analysis of the requirements in the big data analytics field. With this purpose, the big data challenges framework could be deployed to structure the Dutch Customs needs from the technology perspective, and thus help to derive the non-functional requirements concerning the technology domain.

3.2.5 Architecture for Machine Learning

In this section, it is analyzed the technology requirements that machine learning, and in general big data analytics, require to be developed. In particular, it is also investigated what architectural demands are requested by these technologies in the architecture design, which means what components or application services should be included in an SOA design (see section 1.8).

What is needed by every machine learning system is the capability of running the model, and the capability of updating the model with log files or feedbacks collected. This leads to the architectural need of two corresponding architecture components. The need to have two distinct components is because the running model one simply inserts the values in existing equation and computes the results, while the updating model one is responsible for the process of the feedbacks and the activity of data mining and learning. From this, it is obvious to derive that these machine learning projects need a log database where it is possible to save every result for further processing and analysis.

Usually, machine learning projects require an accurate design of the data collection strategy, but in this case, this is not relevant, since the data to gather are already existing in e-commerce platforms. About also a more hardware matter, even if it is out the scope, I want to report some guidelines that have been found during the literature review. This choice depends obviously on the size of the data. A cluster is a better option if data does not fit in RAM. When optimizing for speed or for throughput, GPUs and FPGAs can reach enormous speeds. Finally, training a model and applying a model usually does not require particular requirements, since it is done offline.

This section of the literature review also considers the relation between software and humans, in particular, what are the guidelines to consider when designing a machine learning tool that interacts with humans, in this case, the target officers. This is important because as seen in the previous literature (section 3.1.4), machine learning techniques often need feedbacks to improve their performances, and thus this could lead to architectural demands to consider. This can go from an obvious need of a user interface to the control mechanism which ensures the good quality of the feedbacks.

Besides this, no any further literature is provided on this topic. In the next section, I wanted to investigate machine learning challenges because it could be useful to understand the most relevant issues that are usually encountered while implementing machine learning methods. This is the conclusion of the literature review on the machine learning implementation, and in general on the big data analytics field. A lot of knowledge gap has been identified, especially in this last more practical section (3.2). In the next section, the web-crawling process is described.

3.3 Web-crawling

In this section, I explain what a web-crawler and what web-crawling means. Later on, the two sub-questions presented report the existing literature on big data analytics and machine learning techniques applied to web-crawling, and on web-crawling in the specific domain of e-commerce.

A Web crawler is a software that browses the World Wide Web in a methodical and automatic manner. Web crawlers are also known as the Web spider or Web robot, but also ants, automatic indexers, bots, worm. Web-crawling is also known as spidering. A main purpose of web-crawling is of collecting web pages from the Web and arranging them in such a way that the search engines can use to faster reach web contents. The critical objective is to do it efficiently and without interfering with the functioning of the servers. This is, for instance, the main purpose of search engines (e.g. Google), which largely use web-crawlers to index the Web (Brin & Page, 1998).

Other common purposes of crawlers are the automatic maintenance of tasks on a website, such as checking links or validating HTML code. But in general, all these purposes are linked to the activity of gathering specific types of information from the Web, such as harvesting e-mail addresses (usually for spam), on as in this case, extracting products' prices from e-commerce platforms. This is the reason why the web-crawling technology is commonly considered in the field of information retrieval, and it is placed in the acquisition phase of Hu et al. (2014).

A web-crawler starts its analysis with a URL (or a list of URLs), called seeds. The crawler visits the URL at the top of the list and in visits that web page, looking for hyperlinks to other web pages. It then adds them to the existing list of URLs in the list. This process is repeated until the crawler decides to stop. A web-crawling system can adopt different strategies (ordering metrics) to crawl the web. These strategies define how to crawl the next website or referring to the previous scheme (figure 9), how to compile the URL stack. The most relevant are six: Breadth-First, Depth-First, Backlinkcount, Best-First, PageRank, Shark-Search (M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, S. Ur, 1998). The most traditional ones are breadth-first, where the web-crawling analyzes the current level of depth among all the possible paths, and depth-first, where the crawler analyzes first each branch path until its end before continuing to the next path.

PageRank is instead the algorithm invented by Google in 1998: a web page inherits a high PageRank if it is being pointed by pages that themselves have a high PageRank. It is a way for "bringing order to the web" (Brin & Page, 1998). With this strategy, Google spiders can sort the next links that they should crawl, instead of just crawling the everything they find. Thresholds can be set so that it is possible to leave out not-enough important web pages from the next generation of crawling.

The web-crawling activity refers only to indexing the information, whereas web-scraping is about extracting data from the Web in an automated manner and storing this information for further use. In case of just web-crawling, the software simply stores in a database the structure of interconnections of the web pages, building an index of links. In case of web-craping, it is saved the information in the crawled web pages (not just the link to the web page URL). This is because the aim is to retrieve information from those web pages and use it for further analysis. In my research, the system to be developed is both a web-crawler and a web-scraper, as it has to find the products described in the declaration and extract their values from their web pages. But for the simplicity of writing, I will refer to the technology to be developed simply as web-crawling.

Every crawler that crawls the internet must have the same basic features (Manning, Raghavan, Schutze, 2008), for instance, robustness. This is important because the Web is populated of so-called spider traps, which are loops stuck crawlers in crawling a particular domain without indefinitely. A good crawler must be resilient to such traps. These traps are not always appositely

designed to stop web-crawlers but can be the result of mistakes in websites developments. Another important characteristic of web-crawlers is the politeness. Since web servers have policies to regulate the visits of web-crawlers and can decide to ban too aggressive crawlers, an effective web-crawler must be able to respect these policies.

A good crawler should be able to work in a distributed manner, coordinating its activities with other crawlers that are crawling in parallel. Distributed web-crawling is fundamental when it is needed to crawl the Web quickly. For instance, parallelization could be the solution when the response time of servers is slow, or when servers' visiting policies stops high-frequency requests. Furthermore, a good crawler should be scalable and able to add new machines and extra bandwidth whenever necessary. Finally, the web-crawler should also be extensible, which means being able to adapt to new data formats popping out in a dynamic web environment. Same holds also for new protocols used by innovative servers.

A good web-crawler should be performant and efficient. The system resources like processing power, network bandwidth and storage should be used efficiently without wastes. At the same time, a quality of a web-crawler can be defined as its ability to understand what information that is useful and what is not. Finally, the last feature of a web-crawler to consider is the known as "freshness", and it defines how often a crawler re-visits the same page already visited. This is the case because many web pages update their content over time, order to get new content from the old page. These features can be defined as policies of the crawling process: e.g. selection policy to decide the next page to crawl; re-visit policy to decide the frequency with which re-visiting already processed web pages; politeness policy or parallelization policy (Hu, Wen, Chua, Li, 2014).

3.3.1 Web-crawling Components

After this description of the characteristics that evaluate a web-crawling, it is appropriate to describe how a traditional web-crawler is made. Many versions are found in the literature, and all have small differences. My aim in this section is to give an as more general as a possible structure of a web-crawler trying to show the most relevant functionality, also considering the case of this research. Thus, a web-crawler can be described by the following components:

- Data Extractor: it is the main component of a web-crawling system. It is the component that in the first iteration gets the URL from the user and visits the related server. This is done sending an HTTP request to the remote server which returns the requested information.
- Extracted URL stack: it stores the URLs found by the Data Extractor on the web pages it processed. These URLs a structured in a queue of next web pages to visit.
- Domain filter: it checks that the URL it gets from the URL stack belongs to a given domain or not. In such way, this component can restrict the crawling activity to a specified domain.
- URL classifier: it is the component that checks whether the URL is worth it to be crawled or not. In the case of URLs to files like jpeg, css, img, js, etc., it might be not worth it for the crawler to visit them because looking for different type of information.
- Valid URL list: in this list, the URLs that have been approved for future visits are stored.
- HTML Parser: since web pages are written in HTML and structured with HTML tags so that can be rendered by the browser, it is important to remove this information and focus on the content of a web page.
- Script Pruning: This component is the parallel of the HTML parser for other scripts in other languages, first of all, JavaScript.

- Data Files Creation: it gets the information left by this screening and creates the data file that was wanted to be scraped from the Web.
- Data Storage: finally, this last component stores the scraped files in a database. In case of a standard web-crawling for indexing (no web-scraping or information retrieval), the stored data is an index of web pages: what web page is linked to what others. This is what Google does, crawling and saving the URLs in a database, and then organizing them in a structure sorted by relevance. When browsing on Google, the search is following this offline (the path we follow is the one stored in a database, it is not decided by real-time crawling). As in this research we focus on a real-time look-up crawling, the part of creating an index, including storing and sorting the information is not addressed in this literature review.

The figure 10 below summarizes the components of a web-crawler just mentioned. In the figure, I also made the distinction between web-crawling and web-scraper, so that the reader can visualize as web-craping concerns with the extraction and storage of the data, while crawling is just visiting links in a continuous iteration and mapping these paths in an index.

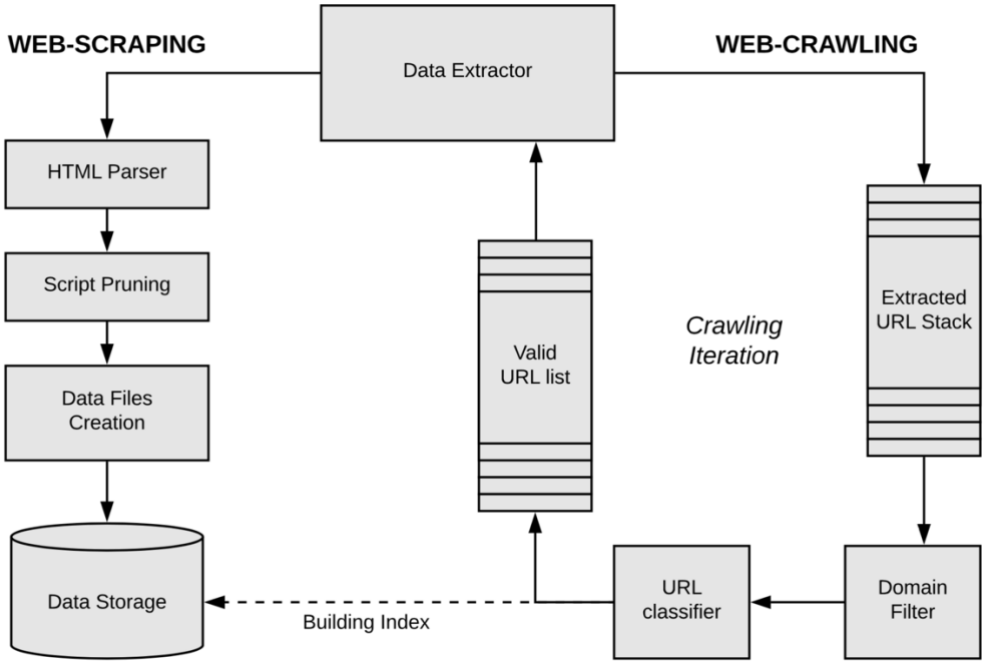


Figure 17: Scheme of Traditional Web-crawling/scraping Components

(This figure has been drawn during the research as a merge of the several sources cited earlier and taking an example from their web-crawling/web-scraping architectures)

In the case of this research, the web-crawler does not need to store the information, because it would be a real look-up on the web. Its scheme could be similar to the one in figure 10 above, but instead of web-storage, there would be a component which compares the value extracted on the Web with the one on the declaration.

Earlier I defined the quality of a crawler as the capacity of understanding what information is important and what is not. In this direction, focused crawlers are built in such a way that can crawl and download only pages that are related to a certain interested topic. For this reason, they are also called Topic Crawlers. They do this by determining the relevance of the document within the candidate web page before crawling it, thus saves hardware and network resources. For instance,

the scheme presented above (figure 10) is already an example of the web-crawler is filtering out the URLs by their domains or classifying if an URL is directing to just an image or a valuable new web page, or also at the scraping side by analyzing the content of the web page (HTML parser). These are already examples of focused web-crawlers. But with the new development of the big data analytics field, new techniques are available to the web-crawling technology.

An interesting approach is to consider the problem of focused crawling as the process of exploring a graph iteratively, focusing on parts of the graph relevant to a given topic (Gouriten, Maniu, & Senellart, 2014). This is a well-known problem in optimization and statistics. Otherwise, another advance of the web-crawling technology derives from the development of the big data analytics field. Here not only the traditional techniques of the statistics discipline are deployed, but also the state-of-the-art machine learning techniques. These more advanced types of web-crawling systems are known as Smart Crawlers, or also Adaptive or Intelligent crawlers (Menczer, Gasparetti, 2004). The main novelty of these crawlers is in the strategy used to decide the next generation of crawling. While the strategies mentioned for traditional crawlers are static – in the sense that they do not learn from experience or adapt to the context of a topic in the course of crawl (Eliassi-Rad & Shavlik, 2003) – smart crawlers use adaptive learning models to assign priorities to the URLs in the frontier.

3.3.2 Smart Web-crawlers

Smart crawlers can be considered as machine learning algorithms, or at least advanced analytics techniques, to the field of web-crawling or information retrieval in general. They started to become popular with the concept of focused crawlers since the amount of data available is growing exponentially every day and it is thus fundamental to be able to efficiently collect data. One solution, it is to collect only relevant data. That's why focused crawlers became important. And do that, the recent big data analytics techniques developed to analyze and process these data are deployed also in this context.

As said earlier, while the traditional strategies of crawling are static because do not learn from experience or adapt to the context of a topic in the course of the crawl, smart crawlers use adaptive learning models to assign priorities to the next URLs. In the literature, there exist at least three adaptive crawling approaches: InfoSpiders, ant-based crawling and HMM-supported crawling (Batsakis, E. G. Petrakis, and E. Milios, 2009). While HMM-supported crawling utilizes Hidden Markov Models for learning paths leading to relevant pages, InfoSpiders and ant-based crawling are inspired by evolutionary biology studies and models of social insect collective behaviour correspondingly.

HMM stays for Hidden Markov Model. It is an advanced machine learning algorithm good for sequential objects, and it can also be used for natural language processing. In this case, the HMM offer an approach to predict the important links to relevant web pages given a learned user model. Firstly, the web pages that a user visits during a learning session and specifically marks as relevant are collected. Then, the semantic content of these pages is exanimated to construct a concept graph which is used to learn the dominant content and link structure leading to target pages using a Hidden Markov Model (HMM). Experiments show that with learned HMM from a user's browsing, the crawling performs better than Best-First strategy (Hongyu Liu, Milios, & Janssen, 2004). The main drawback of this technique is that the computation cost for large document collections is high.

The ant-based crawling is different from this approach because it does not focus on how a single agent crawling the web. It is, in fact, a multi-agent system based on the idea of Ants. The difference between agents and the intelligent systems described above is the social ability that agent could

communicate and coordinate with other agents (Zhang, Du, & Li, 2009). As natural ants communicate with pheromone to find food, our agents work with two kinds of "pheromone" to communicate with others, one is "food pheromone" denoting the values of the importance of pages; the other is "visits pheromone" denoting the visits numbers of pages in recent time. The "food pheromone" value of a page is decided by the importance of itself and the pages it links to.

This approach suggests the use of the state-of-the-art machine learning technique called reinforcement learning to the problem of web-crawling, in which the crawler is regarded as an agent and the Web database as the environment (Wu, Wen, Liu, Ma, 2006). The agent perceives its current state and selects an action (query) to submit to the environment (the web database) which responds by giving the agent some (possibly zero) reward (new records) and changing the agent into the successor state.

For this reason, reinforcement learning would fit the approach of the ant-based crawler. An Ant is considered to be an autonomous living entity that is equipped with a certain amount of energy, moving and communication abilities. Ants are motivated to find useful content in their search to maintain a higher energy level. Therefore, they are rewarded with energy increase for indexing useful information; conversely, they are penalized with energy reduction for wasting bandwidth in useless sites (Zhang et al., 2009, p.).

The other approach mentioned earlier is the one of the InfoSpiders. They are numerous crawlers that form together a multi-agent system for online web search. These agents autonomously check its own information neighborhood by hyperlinks to search for relevant documents according to a user's query (Menczer, 2000). InfoSpiders are able to autonomously evaluate the relevance of the web content with respect to the user's query, and autonomously choose the most convenient future actions, exactly as human users would do. Furthermore, InfoSpiders can adapt, both at individual and population levels, by using reinforcement learning algorithms. The goal is to maintain diversity within the population, but at the same time trying to achieve a good coverage of all the query topics.

InfoSpiders usually rely on traditional search engines to obtain the starting URLs. These links are supposed to be relevant to the query submitted by the user. A crawler is then positioned on each of these URLs. Thee agents start to analyze the current page where they are positioned and evaluate what is the most appropriate link to go next. The analysis consists in looking at a small set of words around each possible next link, and the choice is made by counting the frequencies of query matching terms. The score of each possible next hyperlink is computed by a neural network. After a website has been visited, its relevance to the query is evaluated and then used as a learning signal to update the weights of the neural network (Menczer, 2000).

3.3.3 Crawling the E-commerce

In this section, I try to collect useful experiences existing in the current academic literature on web-crawling and big data analytics applied to the domain of the electronic commerce. For this purpose, it is useful to study the literature of the web data mining discipline, defined as "the use of data mining techniques to automatically discover web documents and services, extract information from Web resources, and uncover general patterns on the Web" (Chen & Chau, 2005). To extract the most detailed information about a product on the e-shop, we saw earlier that there are two are the main sub-problems: finding relevant e-commerce websites, and finding the relevant products that best match a certain description.

Huang, Zhang, Zhang, & Zhu, (2009) propose an approach to recognize e-commerce websites given a comparison of an e-commerce candidate with an ontology domain describing e-commerce

platforms. In this case, the ontology domain is defined as an organized structure of the knowledge background, meant as a scheme of keywords and the relations among them. In addition, to keep this ontology domain updated, they propose reinforcement learning which continuously updates the ontology graph of keywords offline. This could be a possible way to classify an e-commerce website, but it must be experimented to see if it could actually work. The main critic I have for this approach is that it is not for granted that all e-commerce is presented with similar keywords. If their description would vary considerably, this approach could fail.

Verma, Malhotra, Malhotra, & Singh (2015) proposed instead an approach to rank the e-commerce web pages through a supervised back-propagation neural network. The input layer of the neural network gets five variables as input: the content priority, the time spent priority, the recommendation semantic, the explicit and implicit users' feedbacks, and the biased input. The first one is a frequency count within the e-commerce candidate of keywords stored in an e-commerce dictionary. In this sense, this point is similar to the approach proposed by Huang, Zhang, Zhang, & Zhu, (2009), as this dictionary can be seen as the ontology domain. But then there are the other inputs. The time spent priority, for instance, counts the time spent by users in the considered e-commerce through analyzing the log of the users. The recommendation semantic is instead a ranking variable of the e-commerce candidate given the NLP analysis of the user profiles. Although this approach could be interesting for this research, as it provides an innovative solution to rank e-commerce also on the basis of how much is used by their users, it has to be checked whether IBM could benefit of data such as access log files or users' information.

Once decided how to select the e-commerce platforms, the crawler also needs to navigate itself to the product's detail page. Detail page usually contains the product name, price, photos, product properties, product description, customer reviews, etc. A useful property of detail pages is, that they usually have a uniform design, different from any other page on the e-shop. Then it has somehow to find the product it is looking forward (i.e. the product described in the declaration). If the description is not perfectly complete and unique, the crawler probably can only look for the best matching products among the results or rank them for the best to the least matching. To do that, the crawler must be able to compare the products on the e-commerce platforms. This is not an easy task.

Given a set of products from the same category of a same online store, where each product is described in a catalog by a number of attributes (e.g., general characteristics, technical specifications, etc.). This problem, which at a first glance may be seen as straightforward or even mundane, is, in fact, challenging and intriguing. In fact, any automatic solution for it requires techniques for comparing tens of different attributes, whose semantics are often very technical and specific (e.g., the shutter speed of a camera) and also requires dealing with hundreds of products in the category. To be generic, such a solution must also deal with several distinct product categories.

Considering products of the same category from an online store where the product descriptions are made of a list of attributes collected in a catalog and trying to compare them and recognize what is similar to what, might seem a quite straightforward problem, but in reality, it might be extremely complicated. Comparing two products means dealing with a great number of different products and attributes. A possible solution is using a specific similarity function for each group of attributes. Thus, before the comparison, each attribute is classified into to a group that is handled by a specific similarity function. These functions compare the products based on their attributes and rank them from the most similar to the less one (Hoffmann, Silva, & Carvalho, 2018).

3.4 Knowledge Gap

After conducting this literature review, I identified that there is not enough knowledge base in each of the sections presented above. In particular, there is missing literature on how to choose the right machine learning algorithm, how to implement machine learning techniques, and finally about similar applications of these techniques in web-crawling systems such as the use case under analysis. These topics represent the knowledge gap of this research.

The first part of the knowledge gap concerns the big data and big data analytics fields of research, as they are still evolving and not yet systematically organized. Thus, a comprehensible understanding of the trends and new disciplines, their definitions and classifications are yet to be fully established. Nevertheless the fast progress made in BD and BDA, there is a clear lack of management researches and theoretical framework in these fields. For instance, there is no a systematic framework which provides support in the choice of the machine learning algorithm. It is basically trying the best candidates and seeing what it works better.

In addition, there is almost null literature on how to implement machine learning techniques and design machine learning systems. In particular, there is little literature on what an architecture design should take into account in case of a machine learning system, or what architectural components or architectural requirements are necessary in case of such systems. Same is in the case of a knowledge base on how to scale up machine learning projects. As a technique is chosen for its implementation, it is important to know if this choice will work also on a production scale, and not only when prototyping.

Also, there is no formal guideline on how to conduct the requirements analysis in machine learning projects. Since they are mostly focused on data, they have different priorities from classical engineering projects, and also the software engineering approach is not a perfect fit. The only framework specific for such projects is the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Shafique & Qaiser, 2014), but it is more focused on how develop a machine learning project, and not on how to gather its requirements.

Finally, even if there is an existing literature on how machine learning algorithms are used to improve the activities of web-crawling, I could not find specific examples similar to the case of this research. This might be because of the real-time nature of the crawling. In fact, there are no similar examples of web-crawling/web-scraping systems that use machine learning techniques to look up products on e-commerce. There is a lot of literature about analyzing e-commerce data, but they are almost all focused on how to recommend the best products to users or analyzing big quantity of data to understand customers behaviors, while there is no literature on how to find a specific product on the e-commerce platforms, which is the use case under analysis in this research.

Analyzing the smaller sub-problem identified earlier in this research, it is possible to find useful examples, but these usually do not satisfy the use case completely. In particular, the web-crawling system to be developed should be able to recognize the e-commerce platforms and then find the considered product. The first step is thus to recognize e-commerce platforms. There is existing literature on how to do this through an ontology domain, but this might require some computation and waiting time since the crawler needs to visit the website and analyze the content. Same is to analyze the list of products results and discard products that are not the one the crawler is looking for. This can be done maybe by applying NLP models on the product description, but there is no existing literature that could provide useful lessons learned, especially about the computational time that such an approach would require.

Furthermore, there is very limited, if not null literature about the application of machine learning or web-crawling technologies in the specific domain of the Customs Administration and Customs Risk Management, forcing the researcher to look for useful examples in different domains, such as quality management etcetera. However, also in these cases, the so specific and uncommon nature of the use case under analysis made it difficult to leverage knowledge from other domains.

4 Architecture Design

Since the objective of this research is to investigate how data analytics could be implemented in customs risk management to solve the problem of how to monitor an increasing number of e-commerce parcels, the knowledge of the previous chapters must be combined in the third phase of the design cycle to tackle the problem and find a solution (Hevner, 2004). Thus, this chapter combines the knowledge base coming from the e-commerce environment and the Dutch Customs risk management practices with the literature review on the technological aspects, trying to answer to the research question number 3: what is the most suitable design of a web-crawling architecture to improve the cross- validation of price information for e-commerce at the DCA?

To answer this question, the design process and the final artifact of this research are described. As it has been described in the methodology section (1.6), the design process is divided in four sub-phases, starting with the requirement analysis. This is broken down in the definition of the functional requirements – derived from the problem statement – of the non-functional requirements and constraints. These last ones are investigated applying the big data challenges framework described in the literature chapter (figure 16). Finally, addressing the non-functional requirements with the state-of-the-art big data analytics techniques leads to the choice of the technical solutions, which is the second sub-phase of the design process.

In addressing the requirements with the state-of-the-art big data analytics techniques, the research planned to work closely with the data analytics experts at the IBM Research Lab in Ireland, but due to contractual issues during the PROFILE project start-up phase, they have not been available during the research period (only one conference call interview was possible). Hence Ben van Rijnsoever, the Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Services (GBS) at IBM Netherlands, has been the only main IBM interviewee, even if he is an Executive Chief IT Architect Consultant and not a DA expert.

Once requirements and big data analytics techniques are sorted, the components of the web-crawling architecture are derived (sub-phase 3). As mentioned in the methodology section, if when deriving the functional requirements, I followed the systems engineering approach by Armstrong and Sage (2000), when mapping the requirements to the components of the architecture, I follow the Axiomatic Design approach (Suh 1998) which offers a more systematic approach.

Finally, the fourth and final sub-phase of the design process is deriving the design of the web-crawling architecture. This can be defined as a physical/logical design since according to the axiomatic design, I describe the design parameters, and thus the physical domain (Suh, 1998). At the other hand, I represent the high-level design of the architecture with its main activities. According to traditional architecture description, this representation should be classified as rather a logical architecture, with some representing also of the physical parts of hardware (like the application BUS, or the data platform and databases.

Referring more precisely to the Armstrong-Sage (2000) and Axiomatic Design (Suh, 1998) approaches, the definition of the requirements analysis (section 4.1) is the application of the "structured analysis" and “functional decomposition” methods (Armstrong, Sage, 2000); the totality of the functional and non-functional requirements is the functional domain described by Suh (1998); the BDA techniques (section 4.2) and the definition of the architecture components (section 4.3) are the mappings to the physical domain; the Web-crawling architecture (section 4.4) is the physical domain of the axiomatic design (Suh, 1998).

It is appropriate to remind the reader that this is a project for a public institution (DCA), and thus with a fixed budget and a fixed scope. A more agile approach to the requirements – i.e. with more elastic design and interactions with the client after each prototype – is not possible. That's why a structured and accurate requirements analysis is necessary before the development begins. This entire analysis is carried out through interviews with experts from the Dutch Customs and IBM, as it is reported in table 4 (section 1.6). The requirements are addressed with big data analytics techniques found in the literature review and discussion with Ben van Rijnsoever from IBM.

4.1 Requirements Analysis

This section represents the sub-phase 1 of the design process (see section 1.6). Following the approach suggested by Armstrong and Sage (2000), the functional requirements are derived from the analysis of the problem statement. Given the problem of: *"cross-validation of price information between the declaration and the online information in e-commerce platforms"* by the DCA, this is broken down in a sequence of sub-activities to identify the functional requirements. This phase is also known in the literature as a functional breakdown or functional decomposition (Fiorineschi, Frillici, & Rotini, 2018). The list below lists the high-level sequence of steps required from the architecture to be performed to address the stated problem:

1. Gathering the declaration description from the targeting officer, including the package description, package value, and package weight
2. Understanding the number of elements inside the package by comparing the weight of the package with the standard weight of the product
3. Searching the product on the Web
4. Finding the e-commerce platforms which sell that product
5. Finding the product on the e-commerce platforms
6. Extracting the price information of the product from the e-commerce websites.
7. Computing the minimum, average and maximum prices of the products found on the e-commerce websites
8. Comparing these prices found online with the value of the package, also considering the number of products that are calculated to be in the package, and computing the price deviation in percentage, with respect to the minimum, average, and maximum prices
9. Returning these price deviations and a risk indicator of green/red flag to the targeting officer

In addition, interviewing the DCA experts Frank Heijman, the Head of Trade Relations of Dutch Customs, Maarten Veltman, the Chairman of the Innovation Committee of Dutch Customs, and Marcel Molenhuis, the Senior Advisor for Data Analytics, the long-term vision of the project was defined. In particular, they said that the Dutch Customs ambition is to look for price information on arbitrary e-Commerce websites, and not just on the well-known Alibaba, AliExpress or Amazon. In addition, they expressed as, in the future, the DCA aims to generalize this research prototype to every product coming from every part of the world, and not only for the five most critical categories of products coming from the Chinese e-commerce platforms.

This led to one important principle that the design of the web-crawling architecture has to take into account: generalizability. It can be summarized in one more functional requirement: the architecture design must allow the architecture functionality to be extended to any category of products, countries of origin or website.

The same DCA experts also expressed the DCA will to have an elastic architecture that could follow and capture the expertise of the targeting officers and evolve with the dynamics of the web.

This constraint comes from the past experiences that the DCA had in the web-crawling field. By interviewing the DCA Open Source Intelligence Expert and responsible for the DCA web-crawling projects Jo Bootsma, it has been explained that the past projects of web-crawling failed because the technology could not work anymore in a more multimedia-populated internet.

At the same time, if a new e-commerce platform would come out and start to be relevant for the e-commerce trade, the architecture design should be able to recognize it and integrate these different e-commerce scenarios among the options of analysis.

From the sequence of activities listed initially from the problem formulation, and from these general long-term visions, it is possible to derive the functional requirements of the architecture:

Table 10: Functional Requirements of the Architecture

FR1	The architecture must be able to interact with the targeting officer (to gather the information and to present the results of price deviation and risk indicator)
FR2	The architecture must able to retrieve the weight information of the product
FR3	The architecture must be able to interact with/search on the web
FR4	The architecture must be able to find the product and its price online
FR5	The architecture must be able to return a green/red flag given the comparison of values
FR6	The architecture functionality must be generalizable to different categories of products, countries of origin, and e-commerce website
FR7	The architecture functionality must be able to evolve with the dynamics of the Web and the expertise of the targeting officers to different categories of products, countries of origin

Each requirement is written following the structure suggested by Armstrong-Sage (2000):

" To (action word) + (object) + (qualifying phrase) "

For instance, the FR1 is:

To (cross-validate the price on the e-commerce platforms with the value on the declaration) + (the architecture) + (must be able to interact with the targeting officer).

Within this list, the functional requirement number [FR4] is the one that requires to be decomposed in further activities of more technical nature. By interviewing the Lead Architect for Public Safety, Customs and Border Management at IBM Netherlands, he reported that this requirement is to be broken in sub-activities. Finding the product and its price online means:

(1) Creating a search query

(2) Executing a call search engine

(3) Obtaining websites results (e-commerce platforms to look for the product)

(4) Filtering and choosing the most relevant results

(5) Finding the right product inside these websites

(6) Parsing the results to remove non-relevant information (e.g. layout, multimedia, adds)

(7) Filtering and choosing the most relevant results

(8) Extracting the minimum, average and maximum prices

Structuring these requirements in a hierarchical issue three of needs, objectives, and activities following the structured analysis by Armstrong & Sage (2000), the following scheme is obtained:

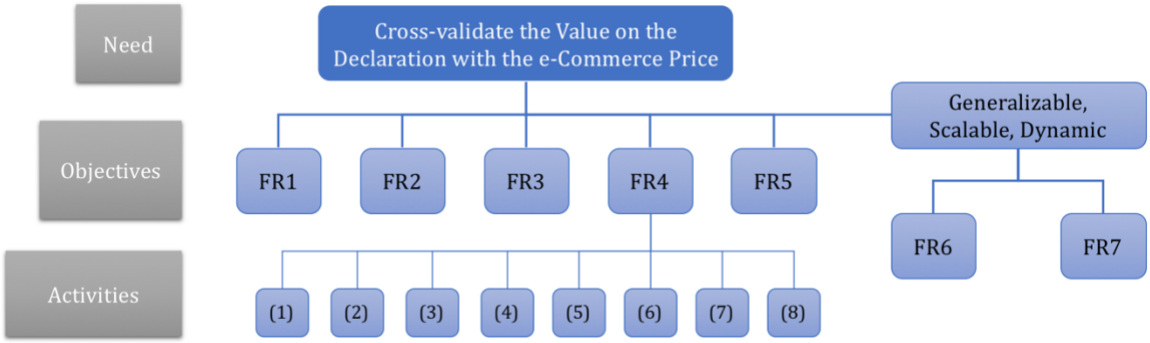


Figure 18: Hierarchical Issue Three of the Problem Statement and Functional Requirements

Thus, what technologies or techniques to use to best perform these activities has to be further investigated. As I mentioned earlier, this project falls into the big data analytics field, as the aim of the project is to solve the problem at stake with the state-of-the-art BDA techniques. For this reason, I believe it is useful to consider the most common challenges when facing big data projects (Sivarajah, Kamal, Irani, & Weerakkody, 2017).

The big data challenges framework (figure 16 in the literature review, chapter 3) is used to analyze the information collected during the numerous interviews with the experts from the DCA mentioned early, from the perspective of big data and its requirements in a systematic manner. This can help to understand more insights into the requirements that the web-crawling architecture needs to address and reflect on the technologies or techniques that could be deployed.

4.2 Non-Functional Requirements

As the big data challenges framework addresses the problem not from the perspective of the problem statement, but from the technology one, the requirements deriving from it are to be considered non-functional requirements and constraints. These are thus investigated from three categories of challenges: data, processes or management challenges.

During this systematic analysis, some requirements will be already addressed with possible solutions, and others will be addressed in the later sections with the design of the architecture and its components.

Table 11: Requirements related to Data Challenges

Challenge	Relevant?	Requirements Analysis
Volume	NO	For the use case considered for the architecture to be developed within the PROFILE project, the data volume is not a constraint

		as only five categories of products are selected. In addition, there is not system-to-system communication, meaning that the officer inputs the declaration information manually through the keyboard. For this reason, the volume of data is not considered a challenge for the architecture to be developed, and thus not a requirement.
Variety	YES	The web-crawling architecture must be able to crawl from general google searches to e-commerce websites, being able to analyze the results of the search engine and within the e-commerce platforms. These data are different in terms of types of information, layouts or multimedia files. In addition, the architecture must be able also to correctly understand the packages descriptions inserted by the targeting officers. Being able to interpret these different types of data (products descriptions, websites results, and products catalogs) is to be considered a requirement.
Veracity	YES	Imprecision and inconsistency are common among both the items descriptions on the declarations and the online descriptions on the e-commerce platforms. The National Coordinator for e-Commerce at the DCA Han Bosch reported that the same product can be described in different ways in different declarations; moreover, during the real purchase done during this research, it has been observed as the same product can be described in many ways on the e-commerce platforms. Finally, some e-commerce websites might show prices with VAT or customs duties included, and some others not. Thus, understanding these discrepancies within the data is a challenge and thus a requirement.
Value	YES	As the Lead Architect for Public Safety, Customs & Border Management at IBM Netherlands Ben van Rijnsoever explained, it is hard to extract the knowledge from unstructured data. In this case, thus, it is hard to extract the right meaning from both the products' description in the declarations and the one in the e-commerce websites. Extracting the right value from data is thus a requirement of the architecture design.
Velocity	NO	The data in the PROFILE research concept are inserted manually by the targeting officer who types the package data through the keyboard. For this reason, handling a high-speed flux of data is not a requirement to be considered.
Visualization	NO	The visualization of data is part of the architecture to be developed, as it has to interact with the targeting officer to show the final results of the risk assessment. But in this case, it is not about showing a high volume of information in a few interfaces, neither showing complicated information. For this reason, the data visualization is not considered a challenge.
Variability	YES	The declarations data is not changing meaning, as well as the information on the e-commerce websites. However, the way how information on the Web is presented might change over time. As reported earlier, the DCA past experience with the web-crawling failed because the Web evolved towards a more multimedia

environment. In addition, the e-commerce platforms might change layouts over time. Thus, in this sense, data variability must be considered a challenge, and thus being able to evolve together with the Web dynamics is a requirement for the architecture to be developed.

Table 12: Requirements related to Process Challenges

Challenge	**Relevant?**	**Requirements Analysis**
Data Acquisition & Warehousing	YES	As the system to be developed is a web-crawling architecture to retrieve information online and compare it with the declarations data, the data acquisition is the crucial process to be performed. Acquiring the right data in the correct manner is thus a main requirement of the design. Storing the data (data warehousing) is instead not relevant in this case, because as said earlier, the architecture does not manage a high volume of data.
Data Mining & Cleansing	YES	This process is meant as developing and maintaining an extraction method that mines out the required information from unstructured data. While in this case, the mining method is coinciding with the data acquisition, the maintaining process is vital to keep the pace of change of the web, and with the expertise of the targeting officers. In this sense, this challenge is similar to the required functional requirement number [FR7], and it is thus a requirement.
Data Aggregation & Integration	NO	In this case, there are not the big quantity of data to be aggregated or integrated, as the only data available are the products information found online compared with the declarations. Thus, this process is not a challenge.
Analysis & Modelling	YES	It is a challenge to analyze the semi-structured online information available in the e-commerce websites, and at the same create a model able to filter the results that have been analyzed and recommend the best one to the targeting officer. As the Lead Architect for Public Safety, Customs & Border Management at IBM Netherlands (GBS) explained earlier when describing the phases of the functional requirements [FR4], the search for the right product on the e-commerce platforms concern both the analysis of the results of the search query and the filtering of the results to recommend the best matching result to the targeting officer. Thus, analysis and modeling are to be considered a requirement of the architecture design.
Data Interpretation	NO	Interpreting the data and finally recommending to the targeting officers to open or not a package is not considered a difficult issue because the recommendation can be done with simple rules on the value deviation. For instance, Ben van Rijnsoever, the Lead Architect for Public Safety, Customs & Border Management at IBM Netherlands (GBS) proposed that if the minimum price found online among the

products approved previously is 50% higher than the value declared in the description (also including the weight analysis and thus considering the quantity of products in the package), then the architecture should send a red flag notification.

Table 13: Requirements related to Management Challenges

Challenge	Relevant?	Requirements Analysis
Privacy	YES	Data privacy is certainly one of the main issues for this project, if not the most critical one. By law, the tax agency is not allowed to expose any information about traders to third parties. Because of this strict requirement, either the technology to be developed must not use any personal information, or special IT and a legal solution must be found to address this requirement. In addition, in the middle of the research (May 25th, 2018), the new European General Data Protection Regulation (GDPR) came into force and made this constraint even more problematic. Unless in the future (from when the PROFILE project will start in August 2018) a policy solution will be found, the web-crawling architecture must be able to find the right product on the e-commerce platforms (and operate in general) without the sender/receiver information. This has been repeatedly confirmed by Marcel Molenhuis, Senior Advisor for Data Analytics at the Dutch Customs Administration.
Security	NO	Security is certainly a big concern and top requirement for the DCA. However, as the Lead Architect for Public Safety, Customs & Border Management at IBM Netherlands (GBS) stated during an interview IBM is planning to use its cybersecurity technology to prevent any possible breach. For this reason, the security issue is not considered a requirement in this research.
Data Governance	NO	Governing big data, categorizing, modeling and mapping them is in this case not a challenge or requirement for the architecture because it is not within the objective of the research. To note that in this framework, the term "data governance" does not mean data security or accountability, but it refers as stated earlier to the management of data, including categorizing, modeling and mapping them.
Data & Information Sharing	YES	As it is for the privacy challenge, sharing data is for the DCA a real issue. To address this requirement, the Lead Architect for Public Safety, Customs & Border Management at IBM Netherlands (GBS) explained that IBM will eventually install most of the software solution in the DCA facilities (even if at the beginning, the initial prototype will be developed on IBM cloud).
Cost/Operational Expenditures	NO	Because within the PROFILE project the prototype considers only five categories of products and the data are manually inserted by the targeting officers, big cots for data centers or

		other operational costs are not estimated to be a problem. Thus, this is not a requirement.
Data Ownership	NO	In this case, the data of the declarations are clearly of DCA ownership, and thus there is not the challenge in defining the data ownership.

From this analysis and reflection on the non-functional requirements and constraints taken from the most common challenges faced during big data projects (Sivarajah et al., 2016), the list of the following non-functional requirements is derived (table 13). Not every "YES" in the previous table are listed because some challenges are redundant and can be reduced to the same non-functional requirement. For instance, the "variability" and the "data mining and cleansing" are redundant with the functional requirements [FR7] of elasticity and dynamicity of the architecture, thus they are not listed as non-functional requirements, but they are still useful to see the same problem from the data perspective (Sivarajah et al., 2016).

Table 14: Non-functional Requirements of the Architecture

NFR1	The architecture must be able to correctly interpret different types of data (products descriptions, websites results, and products catalogs)
NFR2	The architecture must able to correctly interpret vague and inconsistent information (same products described in different ways/in ambiguous ways)
NFR3	The architecture must be able to correctly extract the right knowledge from the data
NFR4	The architecture must be able to correctly analyze and filter the search results
NFR5	The architecture must be able to choose the right websites and products among the search results

From the same analysis, some voices such as the one of data privacy and data sharing are described more appropriately as constraints rather than non-fictional requirements.

Table 15: Constraints of the Architecture

C1	The architecture must function without using the sender/receiver information
C2	The architecture must function with software and hardware installed at the DCA facilities

The data privacy issue has already been addressed shortly in the previous table, but the theme requires a more accurate reflection. Since there has not be found a solution to this constraint, the architecture will work without using the sender/receiver information. However, someone could argue that this information could be very useful in detecting fiscal frauds. This might be true, but as it has been demonstrated in chapter 2 when a real e-commerce purchase has been described, the sender on the declaration is not always the same which is o the e-commerce platforms, making the search query by the sender ineffective. There is not an estimation about how many senders would appear differently on the declarations, thus it is hard to judge if this would be useful or not.

In any case, as it has been described in chapter 2, the DCA has already an active web-crawling tool that could already make a dataset of the five categories of products with their senders and

minimum, average and maximum values. In this way, this tool could be used as the first step: if the sender on the declaration is found online, and if the minimum price among the products sold online by the sender is higher than the price on the declaration, the package should be presented as a red flag.

About the constraint two of data and information sharing, it has been proposed previously to install everything at the DCA facilities. Unfortunately, it is hard to be sure at this point in the project that everything could be set at the DCA, but there is also another option. Ben van Rijnsoever, the Lead Architect for Public Safety, Customs & Border Management at the Department of Global Business Services (GBS) at IBM Netherlands, said that to overcome the challenge of sharing the DCA data with external organizations, the IBM experts usually recur to techniques of anonymization. However, in this case, this is not an option, as these data must be used to find online information and be compared.

The solution could be a safe room: a virtual room that is part of the DCA internal network and where all people who have access are subject to the same procedures as currently used by DCA to give local people access to their data. This safe room might be a physical place in Dublin, but security will be controlled by DCA, or in DCA location with a remote VPN for IBM. This Safe Room must have a computer where IBM can copy the data on and install its software.

4.3 Big Data Analytics Techniques

In this section, the functional/non-functional requirements and the constraints are discussed with the IBM expert Ben van Rijnsoever – the Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Service (GBS) at IBM Netherlands – and the data analytics experts from the IBM Research Lab in Ireland Gavin Shorten – Manager for the Innovation Exchange – and Bora Caglayan – Applied Researcher – with the aim to investigate the most appropriate big data analytics techniques to address the architecture requirements.

However, as mentioned earlier, the experts at the IBM Research Lab in Ireland have not been available during the research period (only one conference call interview was possible) due to contractual issues. Hence Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management, has been the main technical reference point from IBM.

In particular, it will be addressed the functional requirements [FR6] and [FR7], and all the non-functional requirements, as they need further explanation also concerning the technical solutions. Further below, these requirements are discussed and addressed one by one. To refer to the design cycle of Hevner (2004), this represents the sub-phase 2 of the design process.

Let's start with the [FR6] and [FR7]. The first is about the architecture to be able to process in the future every product on every website. Even if now the research has been scoped to five most critical categories of products, it is in the current scope to be able to analyze every e-commerce website and not only the most common ones. If these would not to be the case, according to Ben van Rijnsoever (IBM, 2018) the approach would be: either contact the site owner and request to receive their products/prices data directly (e.g. possible with Amazon); or build a hard-coded solution (i.e. manually writing the steps to be performed with lines of code in a sequential manner) that gathers the information from their website (what DCA did with Alibaba, see section 2.6).

But if the requirement is a more generable solution which can work with any arbitrary e-commerce websites, it is needed a system smart enough to understand any layout and content in such a way that it can extract the necessary information. Therefore, artificial intelligence and machine learning models could be the solution. These models learn by example (positive and

negative examples) collected in so-called "training sets" (Kashyap, 2017), and that is why Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management, advises designing an architecture able to capture the user's feedbacks, in this case, the targeting officer. Besides the training sets of the past experiences, the targeting officers' feedbacks would continue to train these models while in action and improve their accuracy.

This type of technology could also address the [FR7] of building an architecture that can evolve over time following the dynamics of the Web and learn the expertise of the targeting officers. In fact, if the system captures the corrections and the feedbacks of the targeting officers, it will learn from them absorbing their expertise (Ivanović and Radovanović, 2017). Of course, this is not given for granted, and an in-depth research is needed to verify that the information available is enough to allow the system to capture this expertise (Banna et al., 2006). But if this would be the case, then the architecture would also be able to evolve and learn to consider new websites because suggested by the targeting officers.

This approach using artificial intelligence and machine learning models also addresses the [NFR5] which concerns the capability of the architecture to choose the right websites to analyze and the right products among the results. In fact, if these models are trained by the targeting officers about what the right e-commerce platforms are for each product category, and what product on these platforms better match a certain description, these models have the potential of giving accurate recommendations (Arel and Karnowski, 2010). However, it is to be demonstrated that this could actually work, given the available datasets and information. This will be investigated in the research.

There would be thus two different models based on machine learning algorithms to recommend the best e-commerce platforms (e.g. if the country of origin is China, the best websites would be AliExpress and Alibaba, now, but they might change in the future), and the best products within those e-commerce websites (i.e. those products which best match the description inserted by the targeting officers).

From this first analysis, I derive that the design of the architecture must include a component to create these models, a user interface to show the recommendations to the targeting officers and to collect their feedback. Also, these models need to be updated with the corrections and the previous results. It is better to distinguish this functionality in two architectural components, one to run the model how it currently is – called "Model Run" – and one to create/update the model on the background called "Model Calculation". I called the first one "model run" and the second one "model calculation".

These two architecture components would be responsible also for the model which at the end returns the risk indicator of a green or red flag as output to the targeting officer. However, according to the Lead Architect for Public Safety, Customs & Border Management Ben van Rijnsoever, this model will be made with hard-coded rules, and thus without a machine learning algorithm because it is not needed. In fact, this model should follow a much simpler logic.

Ben van Rijnsoever proposed an example of the rule: if the minimum price found online among the products approved through feedbacks is 50% higher than the value declared in the description, also including the weight analysis and thus taking into account the number of products in the package, then it should be a red flag. Another rule, maybe to be applied in parallel could be done considering the average value to avoid that incorrect extreme low values as it could be the price of an accessory instead of the real product (for instance, an iPhone cover instead of the proper iPhone).

But how would this architecture search the products on the e-commerce platforms?

According to Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Services (GBS) at IBM Netherlands, to find an e-commerce website that offers a specific product, an internet search must be performed. This can be done in several ways:

- Either building an index of the Web by crawling all websites in a certain country and later using it to find the websites that mention that given product;
- Or using an existing internet search service, such as Google or Bing, that already indexed the web.

Building an own index would require large systems (also in terms of hardware and internet bandwidth) and thus a large investment, and therefore Ben van Rijnsoever reported that the system to be developed within PROFILE should use an already existing internet search service.
Whether the web-crawling architecture builds its own index or uses an already existing one, in both cases the list of websites resulting from an internet search will likely contain many other web pages (so-called hits) that are not from e-commerce websites and thus need to be filtered out. To address this, Ben van Rijnsoever advised while interviewed, to use the same approach as a human would do:

- Formulating the search query as accurate as possible for instance by adding keywords that are typically for e-commerce websites such as the word "price" plus the country of interest, in this case “China”.
- Reading the extracts of the web pages that are returned and assessing whether this is an e-commerce website or not.

The second step requires the ability to "understand" these extracts and use this information to classify a website result as e-commerce or not. Since these extracts are free text, Ben van Rijnsoever mentioned the possibility of using the artificial intelligence technology of natural language processing (NLP) to process this text and understand whether this description is about e-commerce websites or not.

In this sense, this NLP capability of the architecture would allow the system to filter out the results not relevant, and thus NLP would address the [NFR4]. However, to better filter the results, Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management, also advises using HTML parsing techniques first to filter advertisements and other elements from the search results. This component is also the general element that processes the result pages, which are HTML formatted. Though the HTML parser, the HTML is removed as well as the sections on the page that handle layout, navigation, etcetera.

Thus, all the components not useful to the recommendations models would be removed, and this combination of HTML parsing and NLP addresses the [NFR4] of how to analyze and filter the search results, and the [NFR1] of how to correctly interpret different types of data, as the architecture can then understand text, multimedia, page layout, etcetera.

Finally, the same NLP capability can be used to better understand the package descriptions inserted by the targeting officers. As the NLP technology can interpret the content of the natural language (Russell and Norvig, 2010, p. 860), it can (1) classify the description of the package in one of the five categories of product, (2) give a different level of relevance to the words in the description, and (3) recommend what words could be necessary to complete an insufficient description. I believe that these characteristics are useful to formulate a better query on the Web and match the right product.

For the functionalities above described, the NLP component would address the [NFR3] because it would make the architecture able to understand the meaning of a product description and

create the best query which is not a copy and paste of that description. This means that the architecture can interpret data and extract the right meaning, thus the right value. Same it is for the description of the products on the e-commerce platforms. NLP can make the architecture understand if that description really describes the products it is looking for (LeCun, Bengio, and Hinton, 2015).

Finally, NLP would also address the [NFR2] for the same reason. Being able to understand the true meaning of a package/product description, it would help the architecture to correctly interpret vague or incomplete descriptions.

From the perspective of the architecture design, both these capabilities of NLP analysis and HTML parsing can be defined as two distinct architectural components with the same names: "NLP" and "HTML Parser". The architecture design must have these two components in order to perform the activities above described, together with other components that allow the architecture to interact with the Web – making the search queries – and storing the web results.

Now I addressed all the requirements that were needed, but do the technologies discussed set any further constraint?
Reflecting on this with the Lead Architect for Public Safety, Customs & Border Management Ben van Rijnsoever, he pointed out that these big data analytics techniques work best if they are updated with the results of each analysis. For this reason, the architecture should be able to save every result (of queries, HTML parsing, NLP, recommendations, and feedbacks) to improve the performances of these techniques. Thus, the new non-functional requirement is added:

NFR6	The architecture must be able to save every result of analysis to improve its performances.

Finally, given that the acquisition phase of this web-crawling architecture would be made of such sophisticated big data analytics techniques, it becomes computationally heavier to be performed. For this reason, Ben van Rijnsoever – Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Service (GBS) at IBM Netherlands – reported that it is useful to place a check on the historical data to see whether that declaration description has been already processed or not.

This would be useful to have more precise risk indicators made of a triangulation of three sources instead of two (the current declaration and the online value). In addition, in case of strange differences in prices between historical prices and current values found online, another risk indicator would be created so that the targeting officers can check the causes of such anomaly.

Ben van Rijnsoever also added that this check on the historical data would also give an idea of how many declarations are not aligned with the historical ones and thus actually need a web-crawling search. If for instance, the price on the current declaration would be the same as on the historical declaration, the architecture would not need to perform a search query on the web, but it could already return the green flag. This could also be a validation mechanism on the utility of the web-crawling architecture.

This check mechanism is translated in a last non-functional requirement that must be considered to guide the design of the architecture:

NFR7	The architecture must be able to check whether the product of the current declaration was recently processed or not.

This dataset with the historical declarations about any products belonging to the five categories of products considered should be made available by DCA since they are owners of the data and they have strong constraints in sharing them with third parties.

To recap the requirement analysis so far, the first seven functional requirements have been derived from the problem statement and the information gathered through the interviews at the Dutch Customs Administration. Then, applying the big data challenges framework, five non-functional requirements common to big data projects have been listed. Finally, addressing the non-functional requirements with the state-of-the-art big data analytics techniques, a first idea of the architectural components came out. In addition, two more non-functional requirements derived from the choice of these technologies. The next step is to derive the components of the logical/physical architecture from this list of functional and non-functional requirements and reflections on the appropriate big data analytics techniques.

4.4 Architecture Components

This section completes the mapping process from the functional requirement to the physical domain, describing the components of the architecture, and it represents the third sub-phase of the design process (Hevner, 2004). Every component of the architecture is explained and a precise addressing of each requirement, also functional, is made.

From the previous section, I suggested deploying the technologies of NLP and machine learning to address the critical requirements. As explained earlier these technologies will be represented as three different components in the logical block diagram of the architecture. The NLP component for the natural language processing capability, and the Model Run and Model Calculation to create/update and run the machine learning models.

In addition, I explained that to perform the data acquisition on the web, the architecture also needs a component to interact with the Web (placing the queries and collecting the search results) and the HTML Parser component working together with the NLP one for the filtering and analysis of the results.

As explained, the machine learning models need feedbacks to improve their accuracy, and thus the architecture must be designed with a user interface component to show the results and capture the feedbacks from the targeting officers. This architectural component would be the same that allows the architecture to get the first input with the package description and to show the final results with the risk indicators (green/red flags).

Also, the Model Calculation component needs to have access to every result (of queries, HTML parsing, NLP, recommendations, and feedbacks) to improve the accuracy of the models. These results are saved in the dataset called “Log Dataset”. The architecture is designed with three datasets in total: the log dataset to save the results to update the machine learning models; a dataset to store the historical declarations and inspections results of the five categories of products considered, called "History Dataset"; and a dataset to store the weight information of these five products so that it is possible to estimate the number of elements in the package. This last component is called "Weight Dataset".

From this analysis, the following architecture components are derived. Each component is described through its function, input, and output according to the SOA paradigm (Erl, 2008).

- **Web Interact**
  - **Function**: Send an "HTTP get" request to a web location and gather the resulting page. In this sense, it is the standard part of web-crawling – meant as indexing pages results from a certain search query in the web.
  - **Input**: The URI consisting of the Website and search query.
  - **Output**: The HTML web page that is obtained.

- **HTML Parser**
  - **Function**: Process an HTML-formatted web page; the HTML code is removed as well as the sections on the page that handle layout, navigation, adds, and other non-relevant information.
  - **Input**: an HTML document which can be web pages or search results.
  - **Output**: Text blocks which are left taking out the HTML and layout from web pages.

- **NLP**
  - **Function**: Extracts entities and values from free text. An example of an entity is "category" that will get a value derived from the free text. As explained earlier, it can better understand the package description and classify it into one of the five categories of products; process the textual extract of the websites results to understand if they are e-commerce platforms or not; recognize discounted, second-hand or non-relevant products and filter them out.
  - **Input**: Text blocks.
  - **Output**: List of entities with values. These entities include: product category, e-commerce (Y/N), discounted Y/N, second-hand Y/N.

- **UI**
  - **Function**: Dynamic web pages to support all user interactions, e.g. to collect the package description from the targeting officer, provide him/her recommendations (list of websites and list of products), and capture his/her feedbacks (on both the websites and the products). This component thus also generates the dynamic web pages for the user interface to present the results, including the pages for the intermediate steps that allow the user to make corrections.
  - **Input**: Static HTML content (e.g. stylesheet, graphical symbols) and index of navigation.
  - **Output**: Web pages selections and navigation path of the user.

- **Model Run**
  - **Function**: Execute the algorithms of the recommendation models that filter and select the appropriate match from the results that are appropriate e-commerce pages or appropriate products.

- **Input**: Models, parameters, web pages and values of the model parameters, which are related to the product data on the declaration, including its description, value, weight, and country of origin (this can be more specific when the algorithm to be used will be chosen by the technical experts).
- **Output**: Classification or decision-making, presented as a recommendation list or a risk indicator.

- **Model Calculation**
  - **Function**: It determines the parameters of the machine learning models – and thus creates and updates them – based on the data of the previous results and the corrections by the targeting officers. This is done offline.
  - **Input**: Log data, including historical values of parameters and feedbacks by the targeting officers.
  - **Output**: updated models, which means updated values of the coefficients of the parameters.

- **Log Dataset**
  - **Function**: Store all executed web queries, results of the data acquisition analysis (HTML parsing and NLP values), and user feedback, so that is can be used to improve the analytics and modeling. This information is loaded when the model calculation updates the models.
  - **Input**: Log data to be saved (most of the time it is only input).
  - **Output:** None.

- **Weight Dataset**
  - **Function**: Predefined dataset with weight information of every product category. This is created by the DCA using their crawling tools and their expertise.
  - **Input**: Product category.
  - **Output**: Product weight.

- **History Dataset**
  - **Function**: Information about the historical declarations related to the product categories considered. This is to be created by the DCA because they are the owner of these data and they are the only ones who can access them.
  - **Input**: Product description, value, weight, and country of origin.
  - **Output**: Already processed yes or not, and if any the historical values.

In the table below, all the components are shown to address all the requirements, functional and non-functional, of the architecture to be developed. This is formally called mapping between the functional domain – made of requirements – and the physical domain – made of design parameters, which in our case are the architecture components (Suh, 1998).

Table 16: Mapping between Architecture Components and Architecture Requirements

	ARCHITECTURE COMPONENTS								
	UI	Weight Dataset	Web Interact	HTML Parser	NLP	Model Run	Model Calc.	Log Dataset	History Dataset
FR1	**X**								
FR2		**X**							
FR3			**X**						
FR4				**X**	**X**	**X**			
FR5						**X**			
FR6							**X**		
FR7							**X**		
NFR1				**X**	**X**				
NFR2					**X**				
NFR3					**X**				
NFR4				**X**	**X**				
NFR5						**X**	**X**		
NFR6								**X**	
NFR7									**X**

Table 17: Justification of the Functional-Physical Mapping

[FR1]	The architecture must be able to interact with the user, i.e. the targeting officer.	The component UI (user interface) allows the architecture to present the results in a comprehensible way for the targeting officer.
[FR2]	The architecture must able to retrieve the weight information of the product.	The Weight Dataset contains the weight information for every product category.
[FR3]	The architecture must be able to interact with/search on the web.	The Web Interact can place search queries and return the websites result.
[FR4]	The architecture must be able to find the product and its price online.	The combination of HTML Parser, NLP and Model Run allows the architecture to filter the results first, and then to match the remaining ones with the declaration data.
[FR5]	The architecture must be able to return a green/red flag given the comparison of values.	The Model Run executes a risk-assessing model and its return is the risk indicator of green/red flag.

[FR6]	The architecture functionality must be generalizable to different categories of products, countries of origin, and e-commerce website.	The Model Calculation updates the models with the corrections of the targeting officers so that it can extend the architecture functionality to other general products.
[FR7]	The architecture functionality must be able to evolve with the dynamics of the Web and the expertise of the targeting officers.	The Model Calculation updates the models with the corrections of the targeting officers so that the architecture can evolve following the dynamics of the Web and learning from the targeting officers.
[NFR1]	The architecture must be able to correctly interpret different types of data.	The combination of HTML Parser and allows the architecture to process products descriptions, websites results, and products catalogs.
[NFR2]	The architecture must able to correctly interpret vague and inconsistent information.	Natural Language Processing can understand the category of the product described and complete/correct vague or ambiguous declarations.
[NFR3]	The architecture must be able to correctly extract the right knowledge from the data.	Natural Language Process can give different relevance to the words in the items descriptions and thus better acquire the knowledge necessary to formulate effective search queries.
[NFR4]	The architecture must be able to correctly analyze and filter the search results.	The HTML Parser can take out the layout, the multimedia, the adds, and the NLP can recognize non-e-commerce platforms and discounted, second hand or accessories products to filter out.
[NFR5]	The architecture must be able to choose the right websites and products among the search results.	The Model Run executes machine learning models able to choose the best matching website and product given the package description. In addition, the Model Calculation updates these models to keep them up-to-date.
[NFR6]	The architecture must be able to save every result of analysis to improve its performances.	The Log Dataset collects every result of queries, HTML parsing, NLP, recommendations, and feedbacks.
[NFR7]	The architecture must be able to check whether the product of the current declaration was recently processed or not.	The History Database collects every historical declaration of the five categories of products chosen and their historical values.

4.5 Web-crawling Architecture

This section completes is the phase 4 of the design process and completes the third phase of the design cycle (Hevner, 2004), leading to the description of the Web-crawling system through a service-oriented architecture (SOA). This is the final artifact of this master thesis project (figure 19).

The web-crawling architecture is described with a block diagram combining all the architecture components expl;ained in the previous section. Every component is an application service and is represented as a black box with inputs/outputs, according to the approach coming from systems engineering (Fernández and Penzenstadler, 2015).

Besides the components described earlier, I added to the architecture some architectural components standard of a software architecture (Visnyakov and Orlov, 2015) and according to the service-oriented architecture style of representation. I am referring to the "Orchestration Layer" and the "Data Access Services". The orchestration layer is in computer science jargon, the software that orchestrates the logic flow of the application and calls the application services that are needed. It is the layer that handles the IF statuses and takes decisions consequently.

As described previously, the application services are functional blocks in charge of a specific function and described through their inputs and outputs. The NLP, HTML Parser, and all the architecture components described earlier are all application services. Whether every application service passes its outputs always through the orchestration layer is a design choice. Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Service (GBS) at IBM Netherlands, proposed the architecture principle: "every possible branch in the flow should be controlled by the orchestration layer".

In this way, the architecture components become simpler and therefore easier to test. For example, the Web Interact can return a page, but can also return an error (e.g. internet connection not available, website not available). Thus, following this principle, the execution of the next component (HTML Parser) is done by the orchestration layer, which can also design to call again the web-interact with a different search query.

Finally, the data access services are instead a software layer which is in charge of querying the databases and thus storing and loading the information. For instance, it is the information layer which supplies and saves the values of the parameters for the machine learning models in the log database. For simplicity of representation, the same IBM expert Ben van Rijnsoever, proposed an additional principle: "only exchange data that is required". In this way, the application services of NLP, HTML Partner, etc. can send their output directly to the data access services (because the passage through the orchestration layer is not necessary as it is an obvious operation).

In the scheme below in the next page (figure 19), the block diagram of the service-oriented architecture of the web-crawling system to be developed is provided:

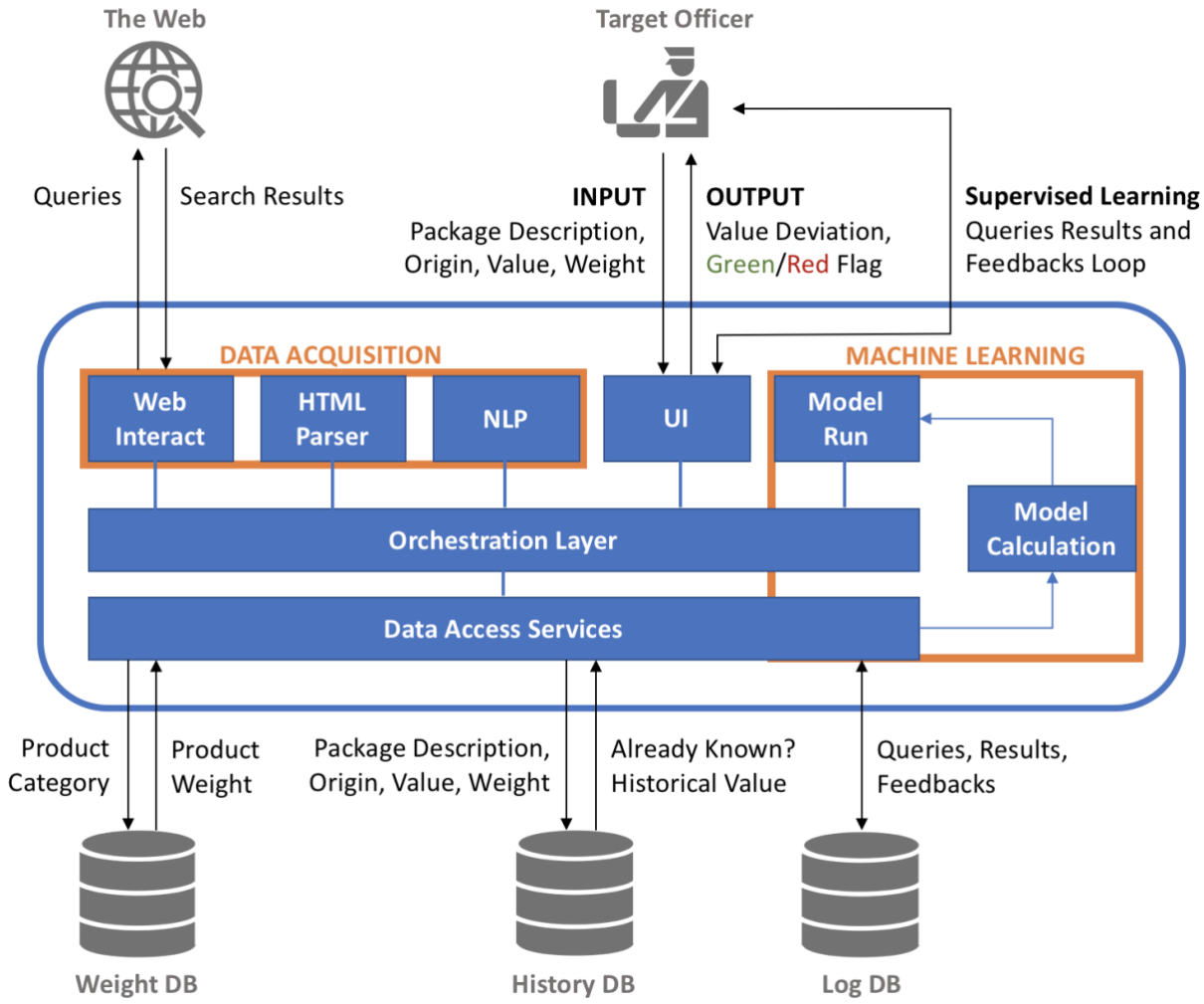


Figure 19: PROFILE Web-crawling High-level Architecture

The web-crawling system to be developed will be an interactive standalone management dashboard where the targeting officer inputs manually (through a keyboard) the description on the declaration of the package that needs to be checked. This can be the e-commerce declaration or the import declaration if the targeting officer wants to check only the products that have been assessed with a value above 22 euros. Because the targeting officer inputs the package description manually through a keyboard, the web-crawling system can be used for any package and any declaration. This design choice was agreed upon as the best option during interviews with the IBM Lead Architect for Public Safety, Customs & Border Management Ben van Rijnsoever, and the DCA Senior Advisor for Data Analytics Marcel Molenhuis, in line with the scope of the project which focuses on demonstrating whether such technology would work or not – and it is not oriented to the development of an autonomous operational tool.

The crawler will use the natural language processing (NLP) technology to process this description and give the right importance to the different words of the description. This is useful to understand what category of product the inserted description is about, and thus understanding what weight a single element should have, and what is the quantity of products inside the considered package. After having understood these details, the system can perform a check on the historical declarations databases and see whether that description is already been processed or not. This is useful to have more precise risk indicators made of a triangulation of three sources instead of two (the online value, the declaration value, and the historical value). In addition, in

case of strange differences in prices between historical prices and current values found online, another risk indicator would be created so that the officers can check the causes of such anomaly.

The targeting officer can press the button “crawl” to start the online research. The crawler (Web Interact) will perform a Google/Bing/Baidu search and show the list of websites that are the most likely to be e-commerce platforms selling that product. The selection of these websites will be helped by the HTML parsing and NLP which can process extract only the relevant information (no HTML, layout, adds, etc.) and can recognize the description of an e-commerce platform from the short piece of text shown in the list of websites results.

The machine learning models are executed by the Model Run which recommends a list of websites to the targeting officers through the user interface (UI). The targeting officers will then give their feedback on what websites they believe should be crawled. These feedbacks are captured and saved in the log dataset which will be used to improve the machine learning models by the Model Calculation component.

The same will be repeated about the right products inside each website has been approved by the targeting officers through the feedbacks. The crawler will look into the chosen websites and will return a list of products that best match the declaration descriptions. Again, these choices are made through machine learning models run by the Model Run component. And also, in this case, the officers are called to give their feedback on what products they believe should be taken into account to compute the minimum, average and maximum values.

Besides the found prices, the user interface used by the targeting officers will show also some confidence indicators and some more information such as the deviation price, minimum/average value, etc., so that it can be easier for the officers to give their feedbacks. Finally, after the officers confirmed the right products among the list of recommended ones, the crawler will return the maximum, minimum and average price, and a risk indicator of green red or flag.

When computing these risk indicators, the system also considers the weight information retrieved before. Each declaration reports the weight of the package which is compared to the weight stored in the weight database. It must be that the value of the package is the value found online times the number of elements estimated to be inside the package. This estimation is computed taking the weight of the package divided by the weight of the product which is in the weight database.

4.6 Architecture Walk-through

With this section, I want to explain the architecture more in detail and moving forward toward the process domain described by Suh (1998). After a general description of the architecture in the previous section, the architecture functionality is described step by step, showing the input and output of each block from when the target officer inputs the product description until the architecture returns the price deviation and the risk indicator.

The used notation expresses the values that are exchanged with variables denoted with a "_" sign (e.g. package_description), as it is done in the coding languages. Each functional step is represented by a sorting number in the architecture block diagram. This type of representation is called an architecture walk-through (figure 20, 21, 22 and 23).

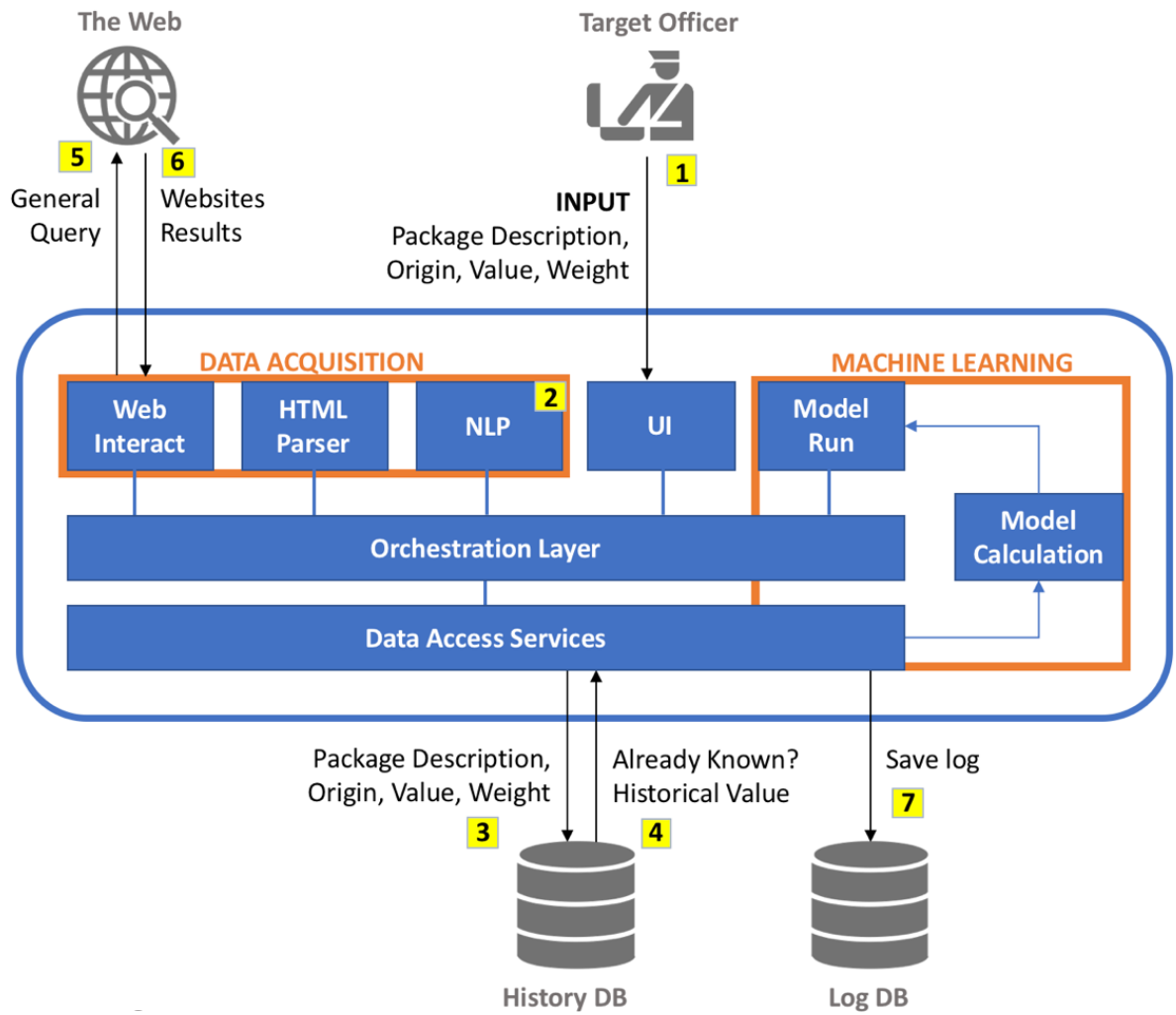


Figure 20: PROFILE Web-crawling Architecture, functionality steps 1 to 7

As it is shown in figure 20, the targeting officer inputs into the user interface (UI) if the system four variables from the package declaration document: package_description, package_origin, package_value, package_weight [1]. These variables are taken by the user interface to the Application BUS which transfers the variable package_description to the natural language processing (NLP) application service. It analyses it and returns the the product_category and a different package_description* which is better appropriate for a search query [2]. This information is then transferred to the Application BUS and to the Data Platform so that it can interact with the historical declarations dataset to check whether the same package has already been processed or not in the past. The Data Platform inputs the four variables of package_description, package_origin, package_value, package_weight to the History Dataset [3]. If the declaration has already been processed in the past, the dataset returns the historical_value variable that will be considered as the element for the risk assessment [4]. At this point, the application service web Interact (the crawler) calls the data processed by the NLP earlier and stored in the Data Platform. It receives the variable package_description* and performs the search query [5]. The results of the search websites_list is returned by the Web to the Web Interact application service [6]. These results are saved in the Log Database [7] passing through the Application BUS and the Data Platform.

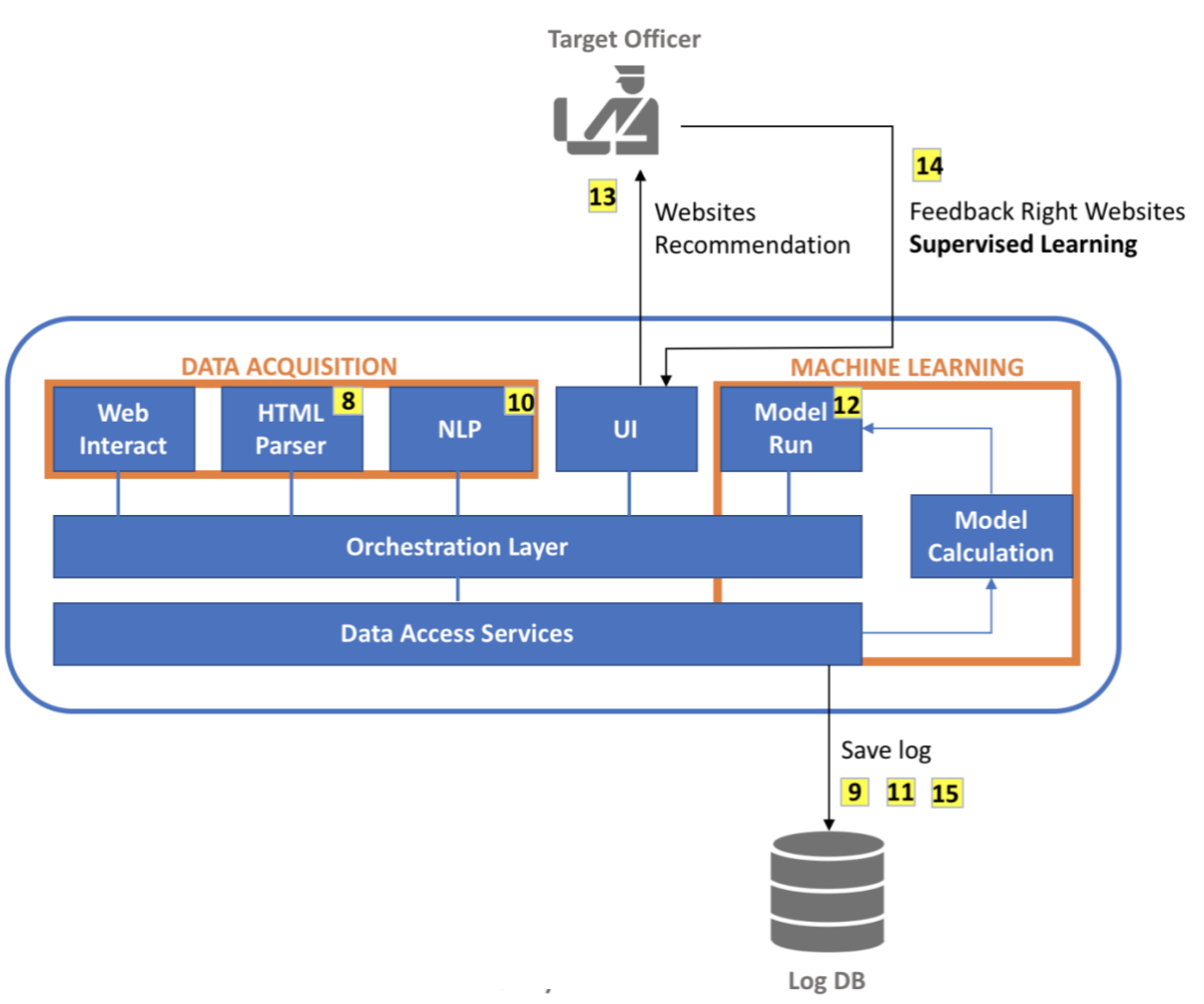


Figure 21: PROFILE Web-crawling Architecture, functionality steps 8 to 15

The HTML Parser asks the Application BUS to provide the websites_list so that it can be analyzed [8]. Then the HTML Parser passes its results websites_list* to the BUS and then to the Data Platform so that it stores them in the Log DB [9]. The same is repeated with the NLP application service which first analyzes the websites_list* [10], and then returns its results websites_list** to be saved in the Log DB [11]. At this point, the Model Run service can take the website_list** and the other information about the product (the four variables) and run the machine learning model which recommends the products [12]. A recommendation of websites to be crawled recommendation_websites is shown to the targeting officer through the user interface [13]. The targeting officer gives his/her feedbacks_websites about what he/she believes they are the most appropriate websites [14]. The feedbacks_websites are also saved in the Log Database [15] to that later the Model Calculation can use them to improve the recommendation model.

The figure 22 shows how the application runs new queries within the right websites defined by the targeting officer using the Web Interact service [16], and it returns results about products on e-commerce platforms [17]. These results are saved in the log database [18]. Now, as previously, the HTML scraper analyzes these results [19] and this analysis is saved in the log database [20]. The same is for the NLP service [21], [22]. Again, as previously, the model run service makes the recommendation list of products [23] which is showed to the targeting officer through the user interface application [24]. The officer gives his/her feedbacks [25] which are also saved in the log database [26].

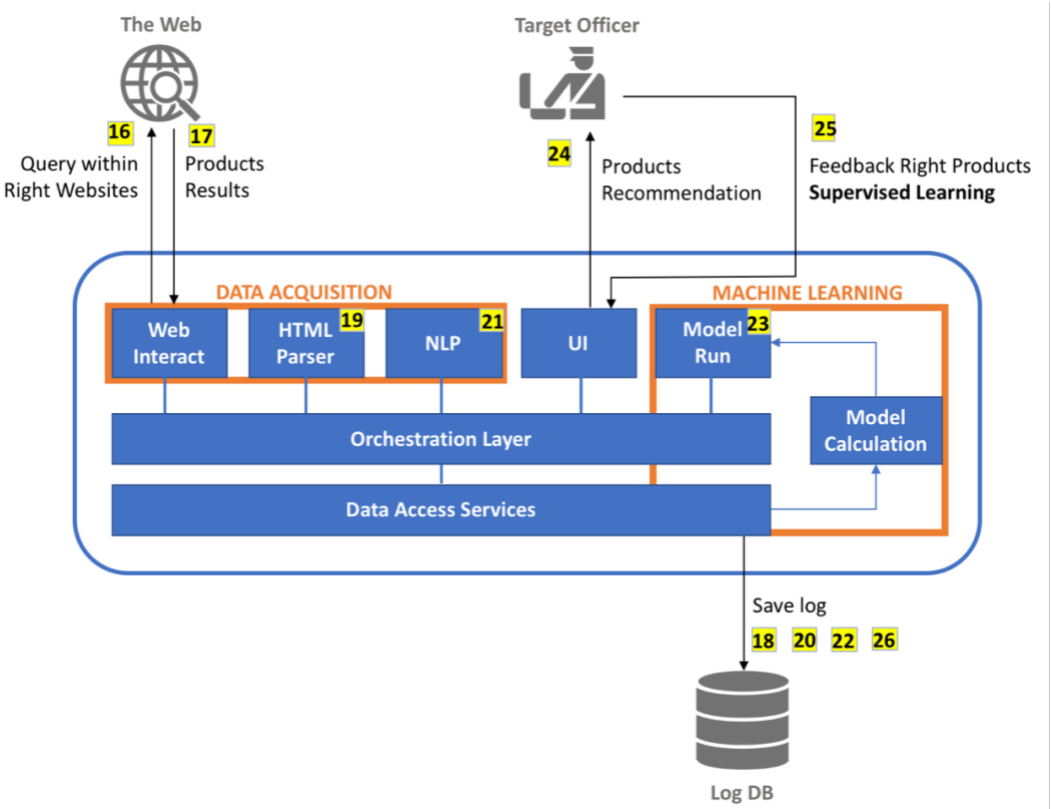


Figure 22: PROFILE Web-crawling Architecture, functionality steps 16 to 26

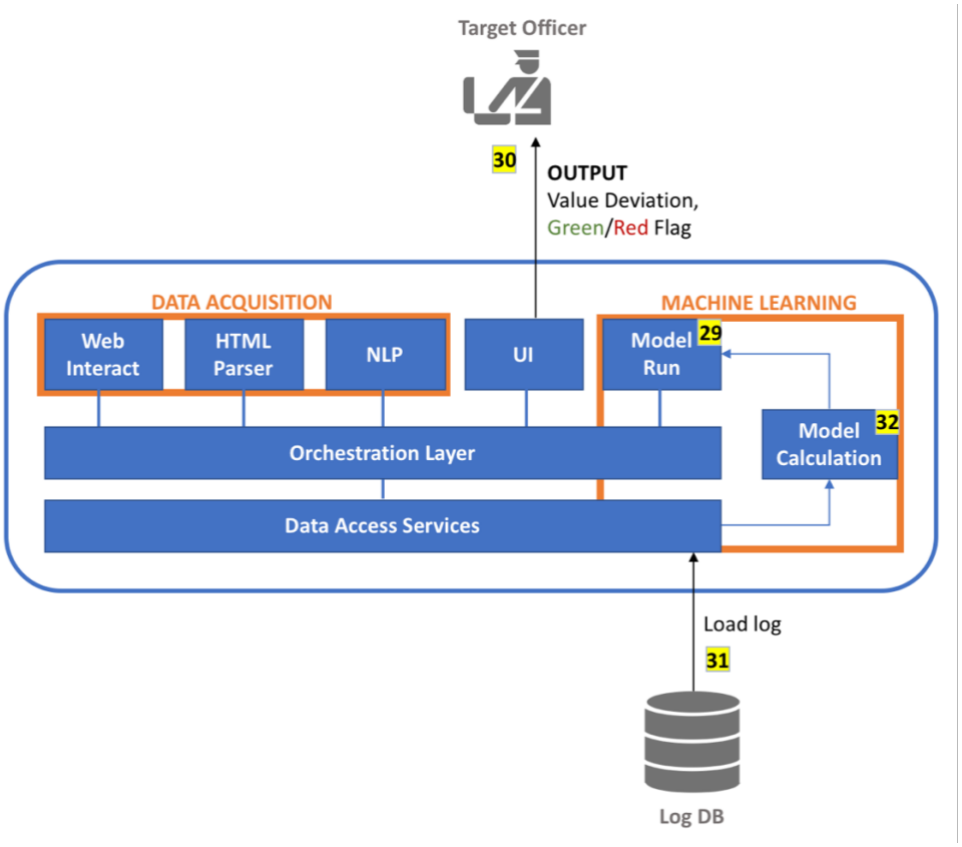


Figure 23: PROFILE Web-crawling Architecture, functionality steps 16 to 26 and 27 to 31

At this point (figure 23), the data platform asks the weight database the weight information for the product category considered [27], [28]. This information can be used together with the rest of the data collected so far to perform the risk assessment [29] and show the results green/red flag to the targeting officer [30]. Finally, when offline, the web-crawling system loads all the data log from the database [31] and use them to improve its recommendations models [32]. This operation is done offline because might require some time to be performed, so it is better to be done when the management dashboard is not used by the officers.

4.7 Architecture Sequence Diagram

Another and more formal mean to describe the architecture functionality is the so-called in literature Sequence Diagram in the Unified Modeling Language (UML). It is a diagram to show the interaction and the behavior of the architecture components within a single use case (Osis & Donins, 2017), which in this case I consider the most complete scenario. As I did for the architecture walk-through, I divided the sequence diagram of the architecture for reasons of space. In this case, this is done in two steps: until the websites recommendation (1), and until the product's recommendations and update of the models (2).

In this diagram, the architecture components are shown as vertical lines with the messages as horizontal lines between them. The sequence of messages is indicated by reading down the page. The vertical axis of the diagram is a sort of timeline: if a component has a more or less long bar means that it stays active more or less time respectively. Let's go through the diagram.

As explained early, also the sequence diagram will consider the following design principles. These principles have to be taken into account while reading the following diagrams.

- Every possible branch in the flow should be controlled by the orchestration layer.
- Only exchange data that is required.

In this way, the architecture components become simpler and therefore easier to test (e.g. the Web Interact can return a web page, but also an error). Thus, following this principle, the execution of the next component (HTML Parser) is done by the orchestration layer, which can also decide to call again the web-interact with a different search query. In addition, for simplicity of representation, the application services of NLP, HTML Partner, etc. can send their output directly to the data access services (because the passage through the orchestration layer is not necessary as it is an obvious operation).

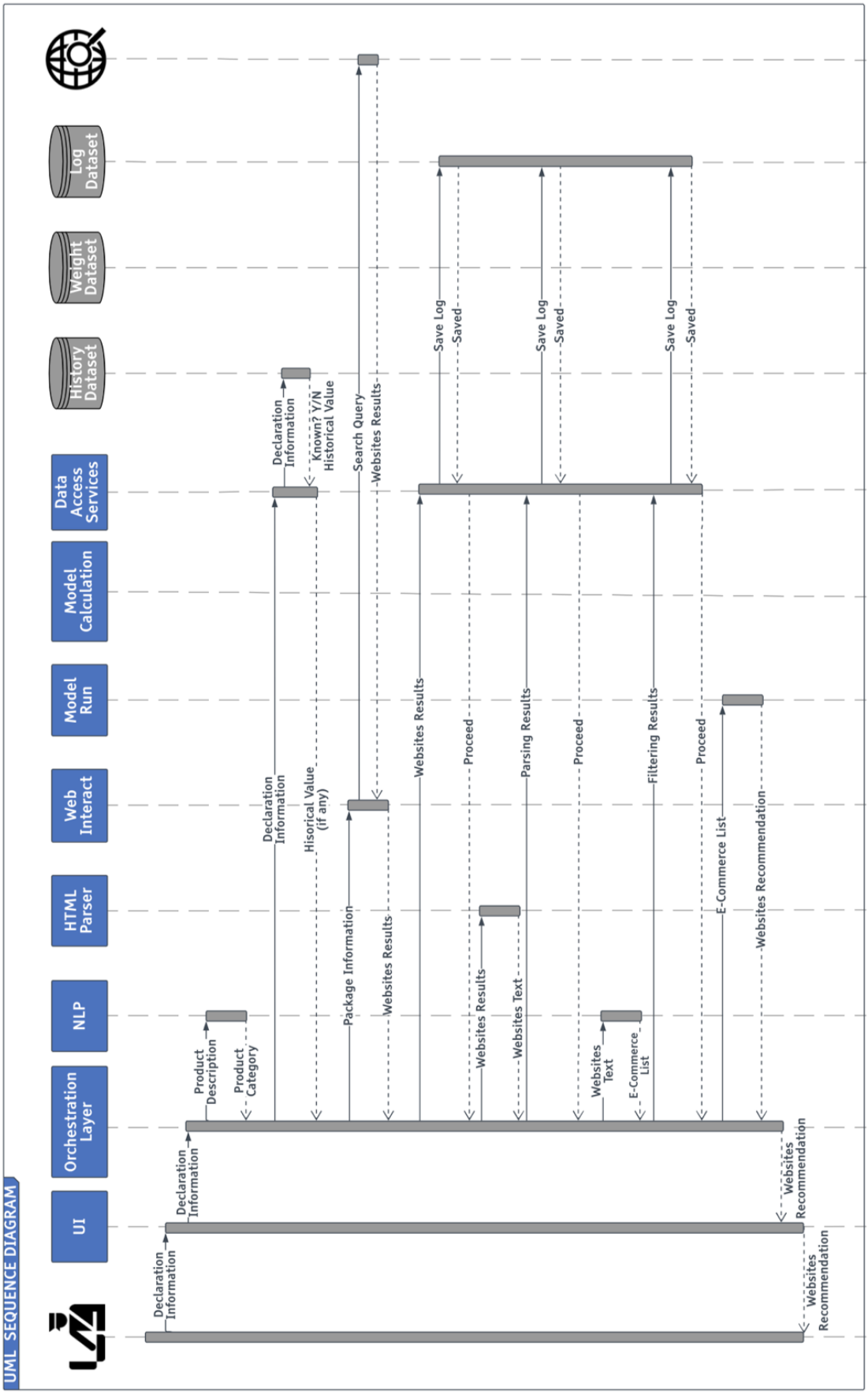


Figure 24: Architecture UML Sequence Diagram, until the Websites Recommendation

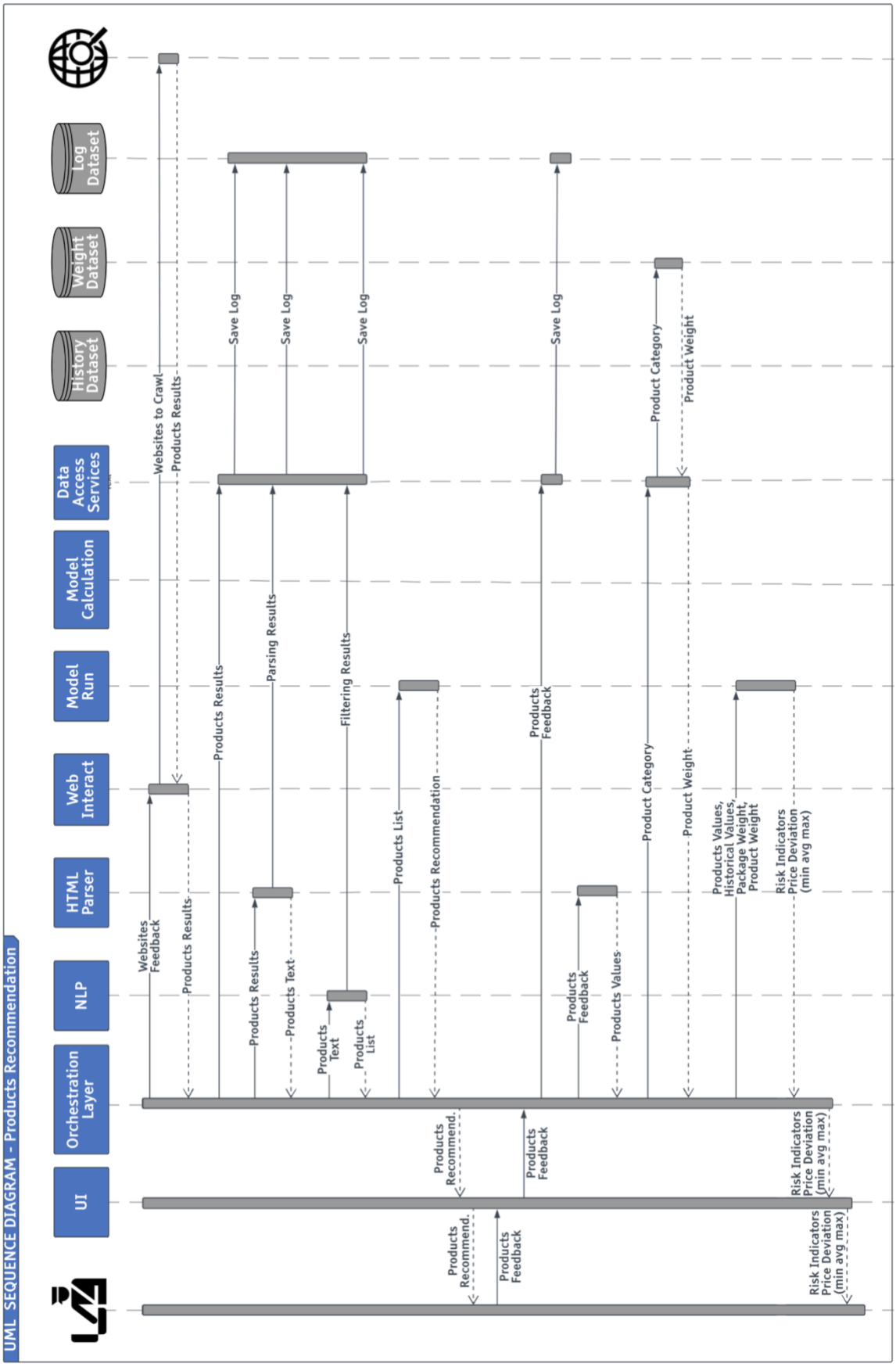


Figure 25: Architecture UML Sequence Diagram, final part

5 Architecture Validation

This chapter presents the validation and evaluation analysis as the fourth and last phase of the design science cycle described by Hevner (2004) and reported in figure 1 (section 1.6). As the figure shows, this process completes the answer to the third research question of "what is the most suitable design of a web-crawling architecture to improve the cross-validation of price information for e-commerce at the DCA". To answer this question, is in fact necessary to validate and evaluate the designed proposed, and make sure that it is indeed the most suitable for the use case under analysis. As it is explained in the table of the research strategies (table 4), the research strategy of this phase is structured interviews with expert at the DCA. These interviews have been written down and reported in this chapter (see also the appendix G and H).

Validation and evaluation do not have to be confused: while evaluation is the process to assess the worth of an artifact, validation is about checking whether or not a certain design is (1) appropriate for its purpose, (2) meets all constraints and (3) will perform as expected (March & Smith, 1995). As the artifact of this research is a web-crawling architecture, the objective of this chapter is to validate and evaluate the designed architecture and the methodologies which brought the research from the interviews with the DCA experts and requirements analysis to the architecture components and finally the architecture design. In other words, the objective of an architecture review is to assess an architecture's ability to deliver a system capable of fulfilling the formulated requirements with the user of such system.

According to the discipline of software engineering, this step is known as Validation and Verification (V&V). The architecture "validation" assurances that a product, service, or system meets the needs of the customer and other identified stakeholders. It often involves acceptance and suitability with external customers. This is indeed the main validation which will be carried out in this research (Chemuturi, 2013).

On the contrary, its "verification" is the evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition. For this reason, it is often an internal process. In this case, the verification would be for instance if the architecture is compliant with the DCA laws in the Dutch jurisdiction, and if the GDPR are respected. These compliances about the data privacy are carried out, but other verification topics are left out of the scope of this research, such as whether the terms and conditions of the e-commerce platforms that are crawled are respected.

Rather, in the field of the design science research literature, Hevner (2004) mention three types of artifact validations: utility, efficacy and quality. The utility of an artifact is about its economical convenience, while its efficacy is how well the proposed design is effective in solving the problem. Citing Hevner, "a design artifact is complete and effective when it satisfies the requirements and constraints of the problem it was meant to solve" (2004). Thus, according to Hevner, the efficacy of an artifact is what is more similar to the validation meant in the software engineering discipline. Finally, the quality of an artifact can be mathematically "computed" by setting key performance indicators (KPIs) on attributes such as completeness, consistency, accuracy, performance, reliability, usability, fit with the organization, etcetera.

Summarizing, the artifact can be assessed by its validity, efficacy, quality and utility. Utility and quality are about artifact evaluation, while validity and efficacy are about validation. The artifact validity is not mentioned by Hevner, who just consider it between efficacy and quality, but it is cited by other design science researchers (Straub, Boudreau, & Gefen, 2004), (Lukyanenko, Evermann, & Parsons, 2014). In this way, it is discussed also common topics of validation for

other research approaches, in particular internal and external validity, and construct, criterion and content validity (Sekaran & Bougie, 2016).

5.1 Artifact Validity

In this section, every aspect of the artifact validity mentioned above will be discussed. First of all, the external and internal validity. The external validity is the level of generatability of the research, and in this case is given by the fact that scientific and rigorous design science approach, together with theories from the systems engineering field, have been used to guide this research. Besides the Axiomatic Design theory developed by Suh (1998) and the common theories of requirements engineering (Armstrong and Sage, 2000), also a rigorous writing down and documenting of the interviews with the experts at the Dutch Customs Administration and International business Machines have been carried out throughout the entire duration of the research, so that future researches can replicate my work (Sekaran & Bougie, 2016).

Besides its replicability, this study refers and largely uses the existing literature and builds upon previous researches. However, the external validity is limited to the specific application domain and to a narrowly defined problem formulation which was necessary in order to developing an effective solution. Further reflections on the limitations of this work are at the section 6.3.

About the internal validity instead, which is about how the scientific methods have been used to indeed address the Dutch Customs requirements, there are the criterion, content and construct types of validity. Criterion validity is the level for which an outcome is related to its input. In the case of this research, the Axiomatic Design paradigm by Suh (1998), maps the requirements to the architectural components in a precise and rigorous manner. This methodology, together with the other scientific research approaches address the criterion validity.

Concerning the content validity, it has to be proved that the problem has been investigated in every possible aspect, and thus that the artifact has a valid knowledge base of content to derive its design (Straub et al., 2004). This also means to validate the fact that among every possible alternative, it has been chosen the most appropriate one. In the case of this research, there are two factors to be considered: firstly, the artifact of this research is a service oriented architecture, and thus its content validation must be done on the possible alternatives for its application services; and secondly, a content validation on the data analytics technique is not relevant, since the choice of the machine learning approach and algorithm is mostly based on trial-and-error, and thus its content validation must wait the development phase.

Finally, the last part of the validation is the construct validation, which is validating that the artifact addresses the problem that was supposed to do (Sekaran & Bougie, 2016). This is the most important validation since it is related to the requirements analysis: given a correct requirements analysis, the interviews with the IBM experts, in particular with Ben van Rijnsoever – Lead Architect for Public Safety, Customs & Border Management at the department of Global Business Service in the Netherlands – can assure that the proposed design matches those requirements. This is thus the most important form of validity. The next section addresses this validity in detail, and it refers to it as efficacy validation, like Hevner does in his publications (2004).

5.2 Efficacy Validation

To examine the validity of the requirements, the interview methodology is described: multiple rounds of semi-structured interviews (see appendixes B, C, D, E, F) have been carried out. The

requirements were asked to the experts at the Dutch Customs during the several meetings related to the project. Every requirement was asked in different ways in the different meeting to see if they were confirmed. Thus, the requirements have been structured and formalized in a different way than how they have been written down during their collection.

Finally, a formal meeting was set up to validate the requirements analysis at the DCA site in Rotterdam. A structured interview has been carried out with Marcel Molenhuis, Senior Advisor for Data Analytics and main responsible for the PROFILE project at the DCA, Ben Schmitz, Venue E-Commerce System Coordinator, and Han Bosch, National Coordinator for e-Commerce, to validate especially the functional requirements, but also the non-functional ones with the support Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Service (GBS) at IBM Netherlands, which participated to the meeting.

Every functional and non-functional requirement and constraints were asked to be confirmed the above mentioned DCA experts. A hand out has been distributed to the interviewees (see appendix G) to facilitate the interview, as each interviewee was able to read the requirements by himself and more easily understand them. The hand out reported the main "goal" of the architecture and the list of its requirements and constraints, all collected in a table with two extra columns to mark each line as confirmed, "Yes" or "Not". I read out-loud each statement and I asked the following questions to the three interviewees:

- Is this requirement clear?
- Is this requirement necessary?
- Would you write it differently?

Each line has been openly discussed, and the interview lasted almost two hours, which was enough to address everything completely. In case I would have been in short of time I would have made sure that at least the functional requirements and the goal would have been addressed properly (they were also the first to be described in the handout). The results of this validation interview are reported in the table below:

Table 18: Requirements Validation by the DCA Experts (carried out on July 31st, 2018)

	Marcel Molenhuis		**Han Bosch**		**Ben Schmitz**	
	Clear	Necessary	Clear	Necessary	Clear	Necessary
Goal	Yes	Yes	Put a focus on the risk indicators	Yes	Yes	Yes
FR1	Yes	Yes	Yes	Yes	Yes	Yes
FR2	Yes	Yes	Yes	Yes	Needed further explanation	Yes
FR3	Yes	Yes	Yes	Yes	Yes	Yes
FR4	Yes	Yes	Yes	Yes	Yes	Yes
FR5	Yes	Yes	Yes	Yes	Yes	Yes
FR6	Yes	Yes	Yes	Yes	Yes	Yes
FR7	Needed further explanation	Yes	Needed further explanation	Yes	Needed further explanation	Yes

NFR1	Yes	Yes	Needed further explanation	Yes	Yes	Yes
NFR2	Yes	Yes	Yes	Yes	Needed further explanation	Yes
NFR3	Needed further explanation	Yes	Needed further explanation	Yes	Needed further explanation	Yes
NFR4	Yes	Yes	Yes	Yes	Yes	Yes
NFR5	Yes	Yes	Yes	Yes	Yes	Yes
NFR6	Yes	Yes	Needed further explanation	Yes	Needed further explanation	Yes
NFR7	Yes	Yes	Yes	Yes	Yes	Yes
C1	Yes	Yes	Yes	Yes	Yes	Yes

As the table above shows, the functional requirements were clear almost for everyone, while the non-functional requirements generally needed further explanation, as they concern more technical aspects. This was expected, as the DCA is more familiar with the functional requirements, while the non-functional requirements have been derived mostly from the application of the big data challenges framework (figure 13) and the interaction with IBM.

While expressing the goal of the architecture, Han Bosch, the DCA National Coordinator for e-Commerce, expressed to put more focus on the return of the risk indicators, and not only on the cross-validation of values itself. I then specified that the architecture will return a risk indicator and a price deviation, and this would be useful to the targeting officers in their customs risk management practices, even if the main goal of the research is to see whether this approach would improve the cross-validation of online values with the declarations.

The [FR7] needed more explanation for every interviewee. This is the one concerning the machine learning technology which would make the crawler smart and able to improve over time. The interviewees needed also to know what "web dynamic" means. For the non-functional requirements, also the [NFR3] needed to be explained to all the participants. This is related to the technology of the NLP which is able to extract the right knowledge from a text block. After explaining this, all the interviewees agreed on the necessary condition of this requirement.

When every requirement has been discussed, I asked the three interviewees the following additional questions:

- Do these requirements describe the right product?
- Would you add any further requirement?

On the first of these questions, the three interviewees agreed unanimously. I then asked if they would have liked to add some other features or requirements that were not addressed by the list in the hand-out. Han Bosch answered presenting the problem of the terms and conditions of the e-commerce platforms that will be crawled and asking if the design of the architecture was addressing this issue. I answered that this issue will be described in the thesis manuscript but not addressed by this research, as it is another complicated issue and will be addressed by the IBM technical experts. Please see section 6.4.2 among the recommendation for IBM for further information.

Other issues expressed by the interviewees did not concern with these requirements, but with the scope of this research. We then repeated and agreed on the scope, so that the requirements analysis was complete. The interviewees agreed again on these following points:

- Dutch-China trade.
- English language and English websites only.
- Fiscal fraud detection only (not security threats).
- Focus on the reduction of the false positive only (not false-negatives).

Finally, Ben Schmitz, the DCA Venue E-Commerce System Coordinator, expressed a final concern about the waiting time that the targeting officer would bare from when pressed the "craw bottom" until the first answer of the system. Ben van Rijnsoever – who was present at the meeting – answered that this time will not be higher that one or two minutes. We, therefore, agreed on writing this down as the further requirement for the architecture. As it is about the technology/operational side and describes "how" the architecture should perform its activities, we set it as the non-functional requirement.

NFR8	The architecture must be able to respond to the user input in less than a minute.

A similar but less structured process was carried throughout the entire process to validate the technical solution and design of the architecture. Numerous interview sections with the IBM expert Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Service (GBS) at IBM Netherlands, were carried out to validate the architecture components and the architecture design, and on the procedure that has been followed toward the design by the researcher.

Finally, to conclude the section on the efficacy validation, the role of the people that have been interviewed must be explained. In particular, it must be explained why their roles are appropriate to confirm and validate the requirements and the architecture design.

5.3 Interviewees Roles

The experts interviewed at the Dutch Customs Administration were chosen because firstly they are the more directly connected to the PROFILE project and they will be responsible to follow its development from August 2018 on (when the project will officially start). Secondly, I tried to cover enough wide range of expertise to analyze the problem from every point of view. In this research, I interviewed some business-oriented roles like Frank Heijman (Head of Trade Relations of Dutch Customs), Maarten Veltman (Chairman of the Innovation Committee of Dutch Customs), and Marcel Molenhuis (Senior Advisor for Data Analytics); some technical roles like Jetze Baumfalk (Data Scientist and Data Analytics Expert) and Jo Bootsma (Open Source Intelligence Expert and Web-crawling Lead for the DCA); and finally some e-commerce related experts like Han Bosch (National Coordinator for e-Commerce) and Ben Schmitz (Venue E-Commerce System Coordinator).

Altogether, these experts' profiles guaranteed the research to have a 360-degrees view of the problem. Concerning the requirements analysis, I believe that the business-oriented positions and the e-commerce related experts were the most appropriate profiles to be interviewed about the functional requirements and the constraints, as they can define accurately the problem

statement and the business needs, the future vision of the project, and the current context of the risk management practices for e-commerce.

For the non-functional requirements, interviewing the technical experts at the DCA and at IBM, combined with the literature review is also, according to us, the most appropriate mean I had available through this research. The technical experts at the DCA Jo Bootsma and Jetze Baumfalk reported valuable insights from the past experiences that have been carried out at the DCA about web-crawling and machine learning respectively (see chapter 2, sections 2.6 and 2.7). These past experiences of the DCA in similar projects, their limitations and the best practices they learned added valuable information to define the non-functional.

Finally, the IBM expert Ben van Rijnsoever, Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Service (GBS) at IBM Netherlands, supported the researcher on filling in the big data challenges framework to derive the non-functional requirements, and was involved in the process of design of the web-crawling architecture from the beginning of the research.

I believe Ben van Rijnsoever is the most appropriate IBM expert among the IBM personnel in the Benelux region to involve in this research. Besides being the Lead Architect for Public Safety, Customs & Border Management from the Department of Global Business Service (GBS) at IBM Netherlands, and thus with an in-depth knowledge of the industry and multiple years of experience in the field, he is also an Executive Chief IT Architect expert of requirements analysis and architecture design. He worked often aside from the researcher and has been interviewed multiple times through the entire duration of the project.

Besides him, also the technical experts who will develop the web-crawling architecture from the IBM Research Lab in Dublin, Ireland, have been interviewed once by Skype to investigate the more technical aspects of the technologies to deploy within the architecture. However, as mentioned earlier, they have been not available as it was planned due to contractual issues.

5.4 Dry-run Test

The so-called "dry-run" test is a sort of feasibility study that was advised to me by Ben van Rijnsoever – the IBM lead architect for public safety, customs and border management at the department of global business services in the Netherlands – and that I was supposed to carry out during this research, since it does not require a working prototype. But because of the strict data privacy regulations and a complex legal framework that entered in place during the research, the exchange of data became an extremely sensitive issue, and I could not perform this test before the end of this study. However, I highly recommend carrying it out in the near future phase of the project.

In this dry-run test, the researcher together with some DCA officers will perform a manual search online of 100 declarations selected randomly. The idea underneath this test is that a web-crawling robot only automates the online search of a human in a faster and more systematic manner, but it cannot succeed when a human is not able to find the product online. Thus, the purpose of this feasibility study is to check up-front how effective this technology could be in finding the right values of products in e-commerce websites. In this sense, this test validates the proposed functionalities and the artifact effectiveness.

The DCA provides the researcher an excel file with 100 item descriptions, values, and weights (no full declaration), taken from 100 declarations, which are selected randomly among the last month of arrived parcels. The dry-run then is performed to try to find the right value on e-commerce

websites as the web-crawling would do and see in how many cases this would detect fiscal frauds. To the excel documents, it will be added three columns to report if the research is successful or not (item found/not found), the price range, and finally if this could be useful or not in the risk assessment. For this specific task, it is important that the dry-run test is carried out in presence of DCA officers who can express their "expert" opinion on this last column.

In these terms, this test is another mean to validate the efficacy and the construct validity of the artifact, because if the problem, as it has been formulated, is not resolvable through this approach (or it is solvable for a very limited number of declarations), then the artifact that is presented in this master thesis would not solve the problem that was meant to solve, and thus would not be valid under the construct validity. In other words, it would not be effective.

5.5 Utility Validation

In this and the next section, I want to give guidelines on how to evaluate the artifact. Since the development of the first prototype of the artifact will start only a few months after the conclusion of this research, the validation of the architecture has only been possible on its efficacy and construct validity. About the artifact utility and quality, I will only propose several guidelines and test design that can be applied once the first prototype will be developed.

As defined earlier, the utility of an artifact is how this is economically convenient. In other words, it is the method to evaluate an artifact from an economic perspective, to make conscious decisions on whether the first prototype should be brought to full development and implementation, or on what level of accuracy it should be developed, for instance.

Taking this last example, I propose a model which links the level of accuracy of the web-crawling architecture to its economic return. As mentioned early, this method can be applied only once the first prototype will be developed. This method is a general method that can be used for many types of artifact and systems design. Knowing the estimation of the total loss that the DCA has for fiscal frauds and collecting some inspections results conducted with the web-crawling, it is possible to link the level of accuracy of the web-crawler – defined as how many false positive it avoids – with the economic return to the DCA. Simply putting it, zero percent accuracy is equal to zero percent return. 100% accuracy (i.e. the web-crawling avoids all the false positive and identifies all the frauds given as input) is equal to 100% economic return (considering the maximum number of inspections possible in one day). Knowing how expensive a certain level of accuracy of the web-crawling system is, it is possible to validate the artifact utility.

Another type of utility validation considering the implication on the logistics and supply chain is not possible in this case. This is because the DCA targeting officers are in a limited number, and they are always overcome by a much bigger number of inspections for their capacity. For this reason, the designed artifact (web-crawling architecture) would not decrease the number of inspections carried on every day, but it only aims to improve the risk targeting by improving the cross-validation of price information.

5.6 Quality Validation

As explained earlier, the quality of an artifact is related to how well the artifact addresses the problem which it is meant to solve. In this case, the problem is the lack of information about the price information on the declarations of products purchased in e-commerce platforms, and the proposed solution is a web-crawling architecture that improves the cross-validation of online price information, and eventually increase the risk targeting for e-commerce.

As for the utility validation, since it is about the artifact evaluation, the methods to validate the artifact quality can be performed only once having a working prototype. For instance, by comparing the historical declarations with the results obtained with the prototype, it is possible to see if the risk targeting performed with the help of the web-crawling system helps to detect the right red-flags or not.

As Hevner explains (2004), the quality valuation of an artifact can be carried out though key performance indicators that measure different attributes of the artifact. Considering the previous example, some KPIs could be the number of frauds detected, or the reduction of the false positive inspections. Another KPI could be also the computing time required, or the capacity of easily adopt to different products and country of origin.

These statistics would be also useful to better estimate the artifact utility and the business case of the artifact. In normal contexts, these data are very useful for example to commercialize and sale the technology or the system to other possible clients. But in this context, it is appropriate to remind few factors described in the application domain (chapter 2): (1) the e-commerce trading is growing every day in an exponential manner; (2) the Dutch Customs has a really limited capacity of inspectors, and the inspections that can be eventually carried out are far less than the ones that are ordered by VENUE; (3) the number of false positive is extremely high.

In this context, any sufficiently working system or technology that can improve the parcel risk targeting has a positive business case. The real challenge is to eventually design and develop a working system able to address this complex problem and respect all the requirements expressed by the Dutch Customs Administration.

In addition, I want to remind the reader that the objective of this research is not to improve the risk targeting, but just to improve the cross validation of prices, and eventually prove that the current technology is mature enough to tackle this problem. For this reason, the real KPIs to measure the quality of the web-crawling system to be developed would be to compare the price deviation found by the system with the estimation provided by the target officer instinctively and count how better, or worse, the smart web-crawler would perform.

6 Discussion and Conclusion

In this section, the conclusions of the findings of this research are drawn, including the contributions of this master thesis project. Later, the limitations of this study are defined, and the recommendations to the parties involved are outlined. Finally, the chapter concludes with the reflections related to this project, including how the methodology has been used, as well as how the available academic knowledge has been applied and what it has been learned through this experience.

6.1 Recap Research Questions

The main research question of this research is "How can data analytics techniques be applied in the design of a web-crawling architecture to improve the cross-validation of price information for e-commerce at the Dutch Customs Administration." This research question has been answered through an iterative design process consisting of understanding the application domain and the customers' needs (the DCA requirements), and how these could be addressed with the state-of-the-art big data analytics. This second part has been investigated through an in-depth literature review in many fields – web-crawling, big data analytics, and machine learning – and through interviews with experts at IBM. This process resulted in a structured version of the requirements and the design of a high-level architecture which address them. This design has eventually been validated through interviews with the same experts who participated throughout the previous analysis. To recap the conclusions of this research I will respond to each sub-research question in details.

1. What are the current customs risks management practices for e-commerce at the Dutch Customs and their limitations?

This question has been addressed starting in chapters 1 and mostly in chapter 2. Initially, the concept of customs risk management is defined – as "the systematic identification of risk including random checks and implementation all the necessary measures for limiting exposure to risk" – operationalized through the hit-rate effectiveness – i.e., the rate of inspections that were actually to be executed.

Consequentially the application domain of the Dutch Customs is described. It is made of four different scenarios: entry, import/export, transit, and e-commerce. Entry is about those goods entering the Netherlands. When they arrive at the harbor of Rotterdam or airport of Schiphol, some of these are imported into the country and become import/export, while other stops at the harbor as transit waiting to be shipped somewhere else.

Finally, it is explained as the Dutch Customs Administration manages its risk management processes through three types of software systems: a system to handle the declarations documents (DMS); a risk engine to assess the risk related to each declaration and decide if that package should be inspected or not (PRISMA and BLAZE or VENUE); a system to handle the inspections of those package that have been targeted as red flag (PLATO).

During these interactions with the experts at the DCA, the five most critical products that will be used in the initial use case have been provided: watch, leather jackets, camera lens, hard disk drive, car CD player (from the main DCA reference person Marcel Molenhuis through email on 19th June 2018). In addition, a sample of a declaration has been provided (Maecel Molenhuis, email on 15th May 2018). While at one hand, it was positive to notice that the DCA is collecting

enough and useful data for every inspection, at the other hand it was shown that the product description is often ambiguous and can be complicated.

After this general overview, the e-commerce domain is investigated more in detail. A study of the e-commerce processes behind an online purchase has been supported with a real purchase of a drone on the Chinese e-commerce AliExpress. The drone was bought at a price of 1244.90 euros, but the declaration document reported a value of 80 euros and the item description of "toy model". Finally, also the sender on the declaration was different from the seller on AliExpress.

This led to reflect on the possible frauds scenarios and how to address them. After further interviews with DCA and IBM experts, it has been agreed to use the weight information as general check for detecting fraud. In case of the real experience presented, they could have written “toy model” of 80 euros, but the weight of the package containing a drone would have been much higher than the estimated weight of a toy model, so that the crawler would multiply the value of an average toy model (found online) for the estimated quantity and return red/green flag according to whether this value is lower/higher than 80 euros.

After these interviews and real purchase experience, I jointly agreed with the DCA and IBM experts that the problem needed to be further scoped in order to make it feasible. It has been shown and confirmed by the Dutch Customs that the products descriptions are often vague and ambiguous, as well as often misleading (from the interview with the national coordinator for e-commerce Hans Bosch on 3rd May 2018; appendix C). Two further assumptions have been derived consequently, in order to further scope the problem and making it feasible: (1) the descriptions are sufficiently informative about the products; (2) the descriptions are not fake, i.e. not misleading or describing false products.

Finally, the DCA past experiences concerning web-crawling and machine learning have been investigated, and relevant recommendations came out from their lessons learned. The DCA is currently working on an internal web-crawling project to web-crawling for the business intelligence department that could be deployed to create the support database of the web-crawling architecture designed in this research.

2. What is the state-of-the-art of web-crawling and big data analytics technologies?

An extensive literature review has been done to investigate the possible technologies that could have been used to address the requirements and the problem formulation. First, a broad analysis of the field of big data analytics is investigated, starting with the definition of big data and continuing with the categorization of the analytics techniques. To avoid reporting too big literature, the topics reported have been accurately chosen – e.g. describing only the techniques for web analytics and text analytics. Finally, the study led to the most recent advances in the field of machine learning and natural language processing, including the last paradigms of deep learning and reinforcement learning.

After this introductory section where the technology is described, the literature review investigates frameworks and design guidelines which can be useful for the design phase. In particular, it is investigated how to choose the most appropriate machine learning algorithm and what are the main challenges to be aware of. The review also investigates how to scale up machine learning projects, and what the architectural demands that machine learning algorithms require in terms of what application services or components are needed. Here, I also wanted to investigate if the literature on machine learning development and implementation provided any guideline on how to carry out the requirements analysis specifically in machine learning projects.

This question mark, unfortunately, has not been answered, since I did not find any literature available on this specific topic. But a recent theoretical framework which collects the main

challenges of big data projects (Sivarajah et al., 2017) to support the requirements analysis from the field of big data analytics has been considered and used to systematically reflect on the non-functional requirements, since big data is the main focus of big data analytics, machine learning and this project in particular. Furthermore, another useful takeaway from this section is that machine learning projects are developed using a trial-and-error approach where the approaches that are believed to perform the best given the past experience of the data scientist are tested and compared.

Finally, in the last part of the literature review, the web-crawling technology described. In the big data analytics section of the literature, the web-crawler is positioned within the literature as data acquisition technology (see the big data analytics value chain by Hu et al., 2014, figure 11). As follows, it is explained what a web-crawler is and what web-crawling/web-craping means. This later extends to the concepts of focused web-crawling and smart, intelligent and adaptive web-crawling, where big data analytics techniques are deployed to improve the crawling performance.

Here that the main pieces of literature of big data analytics and web-crawling are combined to provide useful examples of existing applications. For instance, Huang, Zhang, Zhang, and Zhu (2009) propose an approach to recognize e-commerce websites given a comparison of an e-commerce candidate with an ontology domain describing e-commerce platforms, while Verma, Malhotra, Malhotra, and Singh (2015) provide an approach to rank the e-commerce web pages though supervised back-propagation neural networks. Finally, it is explained as the problem of comparing products on e-commerce platforms is not so straight-forward as it looks like since it is often about comparing many different products with many different attributes. In these terms, this final part of the literature was useful to give an insight of the complexity of the problem.

3. What is the most suitable design of a web-crawling architecture to improve the cross-validation of price information for e-commerce at the DCA?

This sub-question is the proper design process. Here the knowledge from the previous design phases of the problem domain and the literature review is applied in a continuous iterative process according to the design science methodology. The design process starts with the requirements analysis, which has been carried out following the structured analysis method by Armstrong and Sage (2000), thus performing a functional decomposition of the main objective in a series of sub-activities that must be accurately structured to formulate the requirements. These have been collected from the interviews with the experts at both the DCA and IBM, where my role was to mediate between the two parties of customer (DCA) and developer (IBM) to scope the problem in the right direction, and to jointly agree on the most critical factors to be considered for the design of the architecture.

About the non-functional requirements, instead of reflecting on the different domains – technological, environment, law compliance, etc. – as it is proposed by the same systems engineering approach mentioned earlier (Armstrong, Sage, 2000), this research uses a framework which collects the main challenges related to big data project (Sivarajah, 2016). This choice has been made because the web-crawling architecture is the design of a system mainly concerning big data analytics techniques (or at least this is its main innovation). For this reason, reflecting on the challenge related to big data in a systematic manner allowed to cover the technology, management and law/compliance domain of the non-functional requirements. In addition, this framework also hints the discussion of the technological solutions that could be deployed to address the problem. Finally, to link the requirements deriving from the customer needs (DCA) to the components of the architecture, the methodology proposed by Suh (1998) called Axiomatic Design has been used, mapping the requirements into architectural components in a rigorous manner.

These components are then described as application services, according to the service-oriented architecture (SOA) design style and principles. Each component of the architecture is seen as a standard and independent service defined through its function, input, and output, as it was a black-box. This allows the comparability and reusability of these services, which are other design principles of the SOA design style. Each component is thus described in detail. It is described how this system has to have a web-crawling service, an HTML parsing service, a NLP service for multiple use case – categorization of products into categories, classification of a website into e-commerce or not, recognition of second hand products or discounted ones – a model run service to create recommendation models, and finally a model calculation service to update these models.

After the architectural components are defined, the design cycle leads to the design of the web-crawling architecture represented with a block diagram and black-box services. Thus, the high-level functionality is described through the input/output exchanged among the application services and through the external interactions with the user and the Web. Furthermore, the architecture functionality is described with an architecture walk-through and a sequence diagram in the unified modeling language (UML). This design and architecture description are the main artifacts of this design science research, while the mixed methodology encompassing classical systems engineering, software engineering, and big data is its main scientific contribution.

At the end of this research, a validation of the artifact in terms of its requirements analysis is carried out. Being the artifact an architecture, the validation process focused on how to assess the architecture's ability to deliver a system capable of fulfilling the formulated requirements. Based on the assessment done on 31st July 2018, all the requirements have been confirmed, and one requirement was added to complete the set of business needs of the Dutch Customs. Furthermore, the architecture was derived following the Axiomatic Design method by Suh (1998) and confirmed by the IBM expert Ben van Rijnsoever, Executive Chief IT Architect at the department of Global Business Service in the Netherlands. Being the requirements analysis and the architecture design solid, I can state that the artifact is valid, meant that it addresses the problem that was supposed to address (Hevner, 2004). Concerning its quality, however, only future guidelines have been given, since this analysis requires a working prototype to be investigated.

6.2 Recap Knowledge Gap

Altogether, the three research questions answer the main research question of "what design of a web-crawling architecture can deploy data analytics techniques to improve the cross-validation of price information for e-commerce at the Dutch Customs Administration". The first research question brings in the knowledge of the application domain, and what is the problem, the objective, and the limitations of the current system. The second research question brings the knowledge base available in the literature, and here the knowledge gap of this research is identified.

In particular, it is identified that there is missing literature on how to choose the right machine learning algorithm, implement machine learning techniques, design machine learning systems, scale up machine learning projects, on what the architectural demands of machine learning techniques are, and finally on how to conduct the requirements analysis in machine learning projects. Concerning the web-crawling and the use-case, missing literature is also about applications of these techniques in web-crawling systems such to tackle similar problems such as the use-case under analysis.

In front of this knowledge gap, the research question is answered pooling knowledge base from other disciplines, in particular systems engineering and software engineering, and using their approaches for requirements analysis and architecture design. The result is a combination of multiple classical approaches (Armstrong & Sage, 200) (Suh, 1998) at different stages of the design cycle and the use of the Big Data Challenges Framework (Sivarajah et al., 2017). This combination of design science, the functional decomposition and structural analysis for the functional requirements, the big data challenges framework for deriving the non-functional requirements, and the Axiomatic design to go from the requirements to the design features, is my response to the lack of literature review to guide the requirements engineering in complex web-crawling systems deploying machine learning techniques.

Finally, the design of the architecture and final artifact of this research shows how a web-crawling system can deploy big data analytics techniques to improve the cross-validation of price information for e-commerce, answering the main research question of this thesis project and addressing the knowledge gap in this topic. The approach proposed is to deploy natural language processing for the classification of a given product description into one of the five categories (and here it is nothing new). Then performing the query on the search engines and filter out the results not concerning e-commerce platforms applying again NLP on the textual abstract of the search result.

In addition, these results are processed by another machine learning model that ranks them by similarity with the product description, or further filters out non-matching results. Then, the web-crawler would continue its research among the products results and would again use NLP to filter out non-relevant results, and other machine learning models to better match the results with the considered product.

This solution is novel in the literature and bridge this gap in the literature for future researches, and it is thus to be considered a scientific contribution of this research. In the next chapter, these are systematically presented.

6.3 Research Contribution

This thesis project has both practical and scientific/academic contributions. The practical contribution is obviously the result of the applied nature of this research, as it aims to solve a real-life problem at the Dutch Customs Administration. The scientific contribution is instead in both the approach and methodology used and on the combination of the web-crawling and big data analytics techniques. In the following section, these are outlined.

6.3.1 Practical Contribution

The main practical contribution of this research is a preparatory research for the team in the IBM Research Lab in Ireland and the other parties involved in the further development of this project. This manuscript has two main contributions on this side. First, it makes each party involved to agree on the goal, requirements to be satisfied, and solution design. Second, it provides an accurate definition of the problem and a high-level description of the system to be developed for the IBM technical team. They can, in fact, use the service-oriented architecture description of the system provided by this research – in particular, its block diagram and sequence diagram – to

understand what they have to develop. In this description, it is described how this system has to have a web-crawling service, an HTML parsing service, an NLP service for multiple use case – categorization of products into categories, classification of a website into e-commerce or not, recognition of second hand products or discounted ones – a model run service to create recommendation models, and finally a model calculation service to update these models. Given this description of these services and their interactions among them, the IBM technical team can start the development of this web-crawling system.

Another practical contribution is the design of such a system to address a real-life problem at the Dutch Customs Administration. This can be generalized as a guideline case study for customs around the world that want to implement such a technology to tackle a similar problem. It gives an idea of how complicated this problem is, and how important it is to scope the research to a more restricted problem. In particular, it is important to scope one single country, one single language, and fix a certain number of most critical products to begin the research. This is shown later during the design solution when it is described that the natural language processing service categories the products in categories so that an estimation of the product quantity within a package can be done through the weight information of the product categories. This research also shows to customs authorities around the world what type of data is important to have for the machine learning technologies. It shows that it is extremely valuable to collect the data about the historical declarations and the results of the inspections.

This research finally shows how customs authorities can address the problem of cross-validation of declarations' values with online values without the use of personal information, which is an always more relevant constraint (see GDPR on 25th May 2018). The solution proposed in this research is a multi-step approach which collects similar products on e-commerce platforms and returns the minimum, maximum and average price deviation, and in a base of this, it recommends a risk indicator of a green or red flag. The search on the Web is made with a standard web-crawler with an HTML parser service, plus the use of natural language processing to filter the results given their description, and machine learning models that match the other parameters of the products with the online information. In case of fake (misleading) description of the item, the system performs a check on the weight of the package with is a real data provided by the couriers. This proposed solution and architecture design is a practical contribution not only for customs which want to tackle a similar problem but also for any other use cases that aim to retrieve information on e-commerce platforms (or even online in general) without using personal information.

6.3.2 Scientific Contribution

The first scientific contribution of this research is given by addressing the lacks in the existing literature presented in the knowledge gap section (3.5). The approach described in this research of web-crawling the e-commerce in two steps combining NLP and machine learning techniques – in particular NLP for filtering the results and machine learning models to best match the results with the items description – to recommend what package to inspect or not in the domain of the Customs authorities is something new and innovative, not described in the existing literature. For this reason, the proposed design represents an addition to the academic literature. Furthermore, according to the design science methodology, the artifact result of the design cycle represents a scientific contribution itself as it solves a novel problem not existing in the current literature (Hevner, 2004). The design of the architecture, including the application of the state-of-the-art techniques and the requirements analysis in this domain (the customs administration), can represent a relevant scientific contribution, as a similar system with similar requirements can be replicated in other academic contexts or different purposes.

But how it has been mentioned earlier in the recap of the knowledge gap, another main scientific contribution of this research is the process of how the design has been reached, how the requirements have been gathered from the customer needs, including what theoretical framework has been used, and how these requirements have been addressed and converted into the architecture design. This is a novel approach, resulting from the combination of design science with two main approaches to the discipline of systems engineering, and with the additional use of a framework from the big data field. During this research, I, in fact, used the Armstrong-Sage approach to derive the functional requirements from the business domain by breaking down and structure the problem formulation. Then, the axiomatic design is used to map the requirements to the architecture components. In the middle, I use the big data challenges framework to derive the non-functional requirements and to reflect on the big data analytics techniques that could be deployed. This also helped to guide the interviews with the IBM experts.

This is a novel approach that worked well for the case of this research and could be taken as an example for future requirements analysis and architecture design. I am not stating here to have found a new rigorous methodology for requirements analysis of machine learning projects, since I did not compare it to the classical methodologies, neither made an accurate evaluation, but I am proposing this approach as useful example that could be re-used, at least partially, in future similar researches.

Another theoretical scientific contribution is an observation related to the big data analytics value chain framework described by Hu et al. (2014), one of the main authors cited in this research also concerning the classification of the big data analytics techniques. According to Hu, the web-crawling technology has to be placed in the phase of data acquisition, when observing a BDA project. But what it is shown in this research, both by the web-crawling architecture I propose and by the numerous related works, e.g. (Huang, Zhang, Zhang, & Zhu, 2009), (Verma, Malhotra, Malhotra, & Singh, 2015), (Menczer, 2000), is that the techniques of crawling are used together with analytics techniques in order to retrieve the desired information.

In the case of this research, a web-crawling system is, in fact, merging a classic service of HTML parsing with the innovative application services of NLP and machine learning models. Once the crawler service returns the websites results, the HTML parsing service extracts text blocks from them that are processed by the NLP service and the model run service (ML model) before acquiring the information and showing it to the targeting officer. The NLP analysis is in fact carried out before retrieving and storing the results to compute the price deviation. While the analytics phase proposed by Hu et al. (2014) occurs only after data is collected and stored, in this research is thus performed simultaneously with the phase of acquisition. NLP techniques are used to understand the relevant information that should be acquired. Finally, the system described in this research is provided with the capacity of improving its data acquisition by learning from the experts' feedbacks. This shows not only that the analytics phase is often incorporated within the acquisition phase, but also that nowadays analytics is supporting almost every activity and process, as it can fit numerous different applications. This necessity of combination in the e-commerce web-crawling of HTML parsing and text analytics is also a theoretical contribution of this research.

Finally, it is to be noticed as contrary to the many existing projects deploying machine learning techniques, this specific use case of the technology does not benefit of the big quantity of historical data that the Dutch Customs could provide. This is because the data processed are the online information obtained by each time different queries. The only information that is used to increase the accuracy of the models is the correction given by the targeting officers. That is why the interactions with the user is so important in this use case. In addition, the support of the feedbacks is maybe a forced choice given the complexity of the problem. Because of the huge variability of products and attributes in the e-commerce platforms, making an automatic and accurate comparison without the human supervision could not be feasible.

6.4 Research Limitations

Two are the factors to be avoided in order for this research to be of limited application. First, this study aims to provide an artifact to enable data analytics in customs risk management in the case of the Dutch-Chinese trade. Therefore, the reader should be careful with the generalization of case-specific findings, both because of the environment of the Dutch customs – which might be different in other customs – and because of the Chinese e-commerce landscape: other countries under analysis might have different characteristics. Also, another limitation is given by the Chinese language: considering only the English language might lead to a lack of accuracy and effectiveness in the final outcome. Obviously, future research should be done to include the Chinese information into the IS artifact. This would require the inclusion of other data analytics techniques such as image recognition and would increase the complexity of the final artifact. Therefore, considering the Chinese language from the beginning could cause a relevant delay in the implementation, and thus is left out for future research.

Second, given the novelty of the technologies at stake and their continuous innovation, the reader should check for new techniques and updates of the literature and integrate them with this research. In addition, these data analytics techniques are very dependent on the data that are available (Hu, 2014). If different use cases in a different domain have different datasets available, with different variables, other algorithms or approaches could be more appropriate. The same is true if it is possible to use personal information for the analysis. Also, in this case, other approaches could perform better at detecting frauds and providing more accurate risk indicators.

6.5 Recommendations

This master thesis has recommendations for every party involved in the project, thus both the Dutch Customs Administration and International Business Machines. In addition, this section also provides guidelines for future researches and how to take this research further.

6.5.1 For the Dutch Customs Administration

The Dutch Customs Administration should start preparing a dataset about historical declarations and inspections for the five categories of products that have been chosen for this research. This means to start collecting the historical data of the declarations from Venue, and the results of the inspections from Plato, and merge the two datasets. In addition, this data should be divided into the five categories of products, and this is likely to be done almost only manually, since the item descriptions are high variable and vague. It is not needed to divide the entire dataset available, but the most data is available, the most it would be possible to better tackle the problem. This is a process that might take time and it is recommended to act in advance.

As seen in the architecture design the other database to be prepared is the dataset with the weight information of the products considered: Watch, Leather jackets, Camera lens, Hard disk drive, Car CD player. For each type and sub-type of the considered products, the DCA should know the average weight. As explained in the previous chapters, this is useful to compute the quantity of the product type is likely to be inside the package under analysis.

For this purpose, they could use the web-scraping tool which is currently in use. This tool has been described in chapter 2, section 2.6 (also see appendix F). Furthermore, as the DCA can use the personal information of senders and receivers, they could use the web-scraping tool as a first

check: they could make a dataset where for each sender it is stored the minimum, average and maximum values of the products that they sell; if the minimum price among the products sold online by that sender is higher than the price on the declaration, then that package should be presented as red flag.

Finally, the DCA should prepare its teams and employees that will interact with the tool to be developed in order to have both the best results and to make sure that this application will be used. In particular, it is important that the DCA explains the feedbacks role that the targeting officers have to give to the system. It is important for instance that the targeting officers know that they do not have to look for the exact product, but that they have to select products that have a similar description and more importantly, a similar value – and if they can, also a similar weight, but that is a hard task. This is extremely important for the success of the project, and the DCA should take this recommendation as a priority.

6.5.2 For the International Business Machines Corporation

The recommendations for IBM are about the next research questions that are necessary for the development of the technology. These questions arose during the design process and choice of the architecture but have been not addressed as they are out of scope. After the design has been reached, two are the main questions that the IBM researchers must investigate when the project begins:

- How to best place the query?

The IBM developers have to investigate what search query is best to find the right product in the right e-commerce website. For instance, should the query just contain the category of the product plus "price" plus "China", or it should have the entire description of the product reported on the declaration? It should probably have the product category plus important words if available, like a brand or specific type or dimensions, but this very depends on the type of product. There are also "smart" techniques to real-time decide the most appropriate query, but this also would require more computational power and probably more waiting time. In addition, it should be investigated which existing search engine (or combination of multiple ones) should be used. It could be Google, or its Chinese counterpart Baidu, or even Yahoo or Bing.

- How to best parse the results using NLP and machine learning?

With this question, the researcher wants to investigate how to best use NLP and machine learning models to interpret the input and analyze the search results. For instance, what parameters to use in the machine learning models, and what algorithms should be used. This is hard to define up-front, as the choice of the best machine learning algorithm is made by testing every algorithm and see what works best (Ivanovic & Radovanovic, 2015).

Another important recommendation for IBM is about the feedbacks and interaction between the tool to be developed and the DCA targeting officers. As said previously in the recommendation for the Dutch Customs, the feedbacks are a critical point for the success of the project. From the perspective of IBM, I believe it important to set some KPIs to evaluate the feedbacks given by the targeting officers. For instance, it might be that an officer did not understand the mechanism properly and involuntarily gives wrong feedbacks. The system should then be able to detect this wrong behavior and give a notification. In addition, IBM should reflect on how the system should behave in case of all negative feedbacks – for instance, if the targeting officer thinks that all the suggested products are not appropriate. Given the numerous vague or not complete enough descriptions on the declarations, I believe IBM should define these scenarios accurately.

Finally, a complicated technical-legal problem remains unsolved. In crawling the web, there might be the problem of the robot captcha. In addition, there not might be the robot captcha, but still, the terms and conditions could declare that it is illegal to use the information on that website for other purposes different from selling/buying items for instance. Then the solution might be to make a database of websites that are ok to be crawled, and before a website can be inserted into this database, an officer must give his approval. Then the crawler would regularly check that the terms and conditions are not changed. However, this might be quite challenging because the terms and conditions might be hard to find.

6.5.3 For Future Research

As explained earlier, this project is a research project to test whether this approach to the problem would work or not and whether it is possible to improve the parcel risk targeting by crawling the Web and comparing the values found on the e-commerce platforms with those on the declarations. In case it would be successful, a completely different study should be carried out in the future to operationalize this system. Future researches should investigate how to integrate this system into the DCA existing information system, or what IT architecture should be needed to allow the full scalability of this system. For instance, when the volume of declarations becomes high enough, the response times from the e-commerce websites may be too slow to process all the necessary information. Thus, a scalable approach would already consider storing retrieved information into a cache – so that the same lookup is not repeated – and doing off-line crawling to gather the most used info.

In addition, this approach is not so robust against misleading descriptions of the products. In fact, this is one of the assumptions taken in section 2.5 and agreed with the Dutch Customs Administration to simply the problem. However, detecting fiscal frauds of any kind is the main goal of the adoption of this technology. Thus, one of the main areas of future research is possible to parallel approaches that could be implemented to cover every type of fiscal frauds related to e-commerce trading. One possible approach could be to build an own index (not use an existing one such as Google) and deploy an algorithm of reinforcement learning to find the product described on the declaration. In this case, the crawler would act as an agent and the web database (index) would be the environment. The agent perceives its current state and selects an action (query) to submit to the environment which responds by giving the agent some (possibly zero) reward (new records) and changing the agent into the successor state. The crawler would go on until finding the exact product online – also considering the weight and all the value. If it is not able to find, then it would be a red flag, otherwise green. This is an alternative approach that could be more suitable in case of a complete automatic tool, as the reinforcement learning model does not need any interaction with the user to receive feedbacks. At the other hand, it would not be able to sort the fraudulent packages from the biggest fraud to the smallest one (as it is possible to do with the design I proposed, as it computes the value deviations). As it would be useful to detect some frauds in an automatic manner, it would be useful to investigate this approach in a future research.

Finally, an additional recommendation for future research is a more in-depth study on the best NLP algorithms and machine learning models that should be deployed in this case. In general, this field misses of concrete criteria to choose the appropriate algorithm for a given use case. This is probably due to the novelty of the topic. For this reason, additional studies in this field that could give more applicable knowledge would be needed. Finally, the same is true for the application domain of customs authorities and the field of customs risk management.

6.6 Reflections

In this section, I reflect on this experience of research from many points of views. Firstly, the timeline of this research is walked through, and the main obstacles are reflected. Then, the use of the scientific methodology is reflected.

6.6.1 On this Research

The initial idea of this project was to interact with the university, the Dutch Customs Administration and the International Business Machines Corporation to understand what the problem is, what the solution should look like, and how the state-of-the-art big data analytics techniques could be deployed to address these requirements. I understood the customs domain and the customs risk management practices used at the Dutch customs through interviews at the DCA, and I did an extensive literature review on big data analytics techniques, machine learning, natural language processing, and web-crawling, supported with expert interviews at IBM.

Making a requirement analysis was however not a possible deliverable of the design science research approach. How the final artifact should have looked like, and what the scientific contribution would have been was not clear. The reflection on what type of artifact the deliverables should have been crafted around shifted the research toward the design of an architecture, more than the requirement analysis or the BDA algorithms that could be deployed.

The scientific contribution would have been in this case, a critical extension to the big data analytics value chain which makes a clear distinction between the acquisition and analytics phase within BDA projects. This is indeed not the case for the technology solution that is proposed to the Dutch customs, as the web-crawling system needs to use advanced analytics techniques to acquire the right data. But as the research is realized with a much more practical orientation rather than theoretical, the scientific contribution has been re-shaped again towards a more design science related one. This continuous change of direction caused a considerable loss of time and drawbacks throughout the design cycle.

However, this has been a source also of continuous learning and development. First of all, I learned about many fields of research that I am interested in. Big data analytics, information technology and in particular the web-crawling technologies, as well as newer fields of machine learning, artificial intelligence, and natural language processing, they all have been fields touched and investigated in our research. In addition, scientific methodologies such as design science and requirement engineering for systems engineering have been used and considerably contributed to the learning experience of this research.

Another main constraint that, as mentioned multiple times in this manuscript, has considerably influenced this research was that the experts at the IBM Research Lab in Ireland, the ones who will develop the web-crawling system and that are part of the team which will receive this manuscript as input for their work, have not been available during the research period (only one conference call interview was possible) due to contractual issues. The second unexpected issue was the entry in force of the new European General Data Protection Regulation (GDPR) in the middle of the research (25th May 2018).

According to these new regulations, an enterprise can be fined up to its 10% of the worldwide revenue (not of the single branch) in case these laws would be violated. Being IBM a multinational company operating all over the world, such a fine represents a big risk that must be addressed with the right measures. This slowed down the flow of this research considerably, as every activity concerning data (especially data sharing among the DCA and IBM) must be taken extra

seriously. About the GDPR, I feel to share my personal view on these new regulations and stand beside those voices that think that GDPR risk to slow down the innovation in the European Union, since all the projects related to artificial intelligence or machine learning would suffer of huge bureaucratic anchors and companies would refuse such projects just for fear of these possible huge fines.

6.6.2 On the Methodology

Another important part of the reflections is about the scientific methodology, and thus how the design science approach has been applied. I indeed experienced the continuous iterative as discovery process and improvement of the design. During the design process, there have been numerous jumps between knowledge base and application domain or returns to the literature review after the interviews with the experts. Numerous interviews have been carried, both with experts from the DCA and IBM who participated in a continuous improvement of the design. To structure, this iterative process in a static document also has been a challenge, because of the continuously evolving and not perfectly sequential sequence of events.

During the numerous interviews with experts, the requirements were gathered in a general description, and it has been a responsibility of the researcher to apply scientific methodologies to structure them and systematically question the data analytics techniques to best address these requirements. Stating with the approach by Armstrong and Sage (2000), I noticed that it was not suggesting a clear way to derive the architecture components. That is why I used the Axiomatic Design by Suh (1998), which maps each requirement to a specific design feature of the architecture, which in the SOA paradigm means an application service.

After the functional decomposition (Amrstrong and Sage, 2000), a first attempt of mapping the requirements into architecture components was made, but it was clear that a further decomposition was needed, in particular for the [FR4]. This was addressed by deriving non-functional requirements with the use of the Big Data Challenges Framework by Sivarajah et al. (2017). This further decomposition has been included in the structural analysis by Armstrong and Sage (2000) deriving a more complete list of requirements which have been mapped into the service components of the SOA architecture.

According to the design science approach (Hevner, 2004), knowledge from the previous design phases of the problem domain and the literature review have applied continuously to suggest alternative solutions and better designs. For example, the idea of doing two different models for the filtering of the ranking of the results have been the result of this process.

During the interviews, it has also been difficult to mediate between the customer and the developer side to agree on common objectives and design specifications. Many alternatives were proposed and confronted with the list of the requirements to be validated and choose the most appropriate design. As new requirements were coming out continuously it has been hard to scope them systematically and give them a systematic structure enabling an effective design process. Another example of the active contribution of the researcher into the design of the architecture has been the study and analysis of the e-commerce platforms, that being different are themselves sources of constraints and push towards certain requirements instead of others.

Finally, I want to finish this reflection section with a self-critic: I definitely underestimated the numerous scenarios of the use case, in terms of possible frauds or different characteristic of the products and e-commerce. This made the choice of a general design and algorithms which can work in every scenario a much more challenging task. I should have addressed this complexity earlier and more in detail during the interviews, for instance, preparing more structured

questions and brainstorming. Knowing more about the scenarios of the five products could have led to a more accurate choice of the machine learning algorithms with a probable better performance, even if in exchange of freedom of generalizability: in case of different products, the same algorithms might not maintain the same performance.

6.6.3 On my Role as Researcher

In this research, my role was to be the knowledge broker between the client and developer side, working within the business service department of IBM. As knowledge broker, my role has been to mediate between the expectations of the client, the Dutch Customs, and the real possibilities that could have been developed. Covering this role, I noticed that often it is under-evaluated, but it is rather important because often different professionals involved in the project have different understanding of the problem, and their knowledge is not aligned (Waheed, 2018). Being a knowledge broker, I could notice this, and made me reflect on the problem more accurately, and making all the parties agree on necessary requirements.

As it is reported by Oluikpe (2015), projects success and shared decision-making on results and deliverables depend most of all on relational processes developed in the project environment, rather than formal documentation. For this reason, projects can be seen as a social network of Individuals and organizations, where knowledge brokers who bound the multiple actors together are extremely important (Waheed, 2018), since they create links between individuals or departments that possess the knowledge and those who need it

In addition, I learned how important the professional service roles are, meant as key figures between the technical experts and the industry focus. Since the technical experts are indeed experts on a single technology, they often lack the true understanding of the problem, while the industry experts can make sure that the technology which is developing is indeed what the client requires and needs. In big organizations such as IBM, there are technology experts and industry experts, and both are equally important to implement successful projects.

6.6.4 On the Management of Technology Master's Program

This reflection section is about my master's program Management of Technology at Delft University of Technology, in the Netherlands. I believe this type of research is appropriate for master students of this program, since the management of technology has its focus to solve societal problems with the use of technology and innovations. In these terms, the PROFILE web-crawling system that will be developed at the Dutch Customs Administration is perfectly matching this description.

However, I felt MoT did not prepare me to manage complex and innovative IT systems like the use case of this research. Given my technical background in engineering I knew about software engineering, systems engineering, and requirements analysis, and thanks to MoT I felt I had the general management and business knowledge to evaluate the business needs of the DCA and acting as knowledge broker between the client and developer side, but I would have liked MoT to provide me with some specific courses of management in IT settings. Courses about information technology architecture design or IT management topics, or even more specific about big data analytics/machine learning, would have been useful in such a research.

Bibliography

Arel, I., Rose, D., & Karnowski, T. (2010). Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]. *IEEE Computational Intelligence Magazine*, *5*(4), 13-18. doi: 10.1109/mci.2010.938364.

Arsanjani, A., Booch, G., Boubez, T., Brown, P., Chappell, D., & deVadoss, J. et al. (2018). SOA Manifesto. Retrieved from http://www.soa-manifesto.org

Bapna, R., Goes, P., Gopal, R., & Marsden, J. (2006). Moving from Data-Constrained to Data-Enabled Research: Experiences and Challenges in Collecting, Validating and Analyzing Large-Scale e-Commerce Data. *Statistical Science*, *21*(2), 116-130. doi: 10.1214/088342306000000231.

Batsakis, E. G. Petrakis, and E. Milios, "Improving the performance of focused web crawlers," Data & Knowledge Engineering, vol. 68, no. 10, pp. 1001–1013, 2009.

Bhute, A., Bhute, H., & Meshram, D. (2010). Intelligent Web Agent for Search Engines. In *International Conference on Trends and Advances in Computation and Engineering, TRACE-2010*. Freiburg, Germany.

Bijan, Y., Yu, J., Stracener, J., & Woods, T. (2012). Systems Requirements Engineering - State of the Methodology. *Systems Engineering*, *16*(3), 267-276. doi: 10.1002/sys.21227.

Booch, G., Rambaugh, J., & Jacobson, I. (2005). The Unified Modeling Language User Guide(2nd ed.). Addison-Wesley Professional.

Brace, W., & Cheutet, V. (2012). A Framework to Support Requirements Analysis in Engineering Design. *Journal Of Engineering Design*, *23*(12), 876-904. doi: 10.1080/09544828.2011.636735.

Braganza, A., Brooks, L., Nepelski, D., Ali, M., & Moro, R. (2017). Resource Management in Big Data Initiatives: Processes and Dynamic Capabilities. *Journal Of Business Research*, *70*, 328-337. doi: 10.1016/j.jbusres.2016.08.006.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X.

Butler, R. (2008). An Analysis of Service-Oriented Architecture (SOA) to Determine the Impact on the Activities Performed by the External Auditor of a SOA Service Consumer. Meditari Accountancy Research, 16(2), 13-30. doi: 10.1108/10222529200800010.

Caracciolo, A., Lungu, M., & Nierstrasz, O. (2014). How Do Software Architects Specify and Validate Quality Requirements?. In *European Conference on Software Architecture (ECSA 2014)* (pp. 374-389). Bern, Switzerland: Software Composition Group.

Carayannis, E., & Campbell, D. (2009). 'Mode 3' and 'Quadruple Helix': toward a 21st century fractal innovation ecosystem. *International Journal Of Technology Management*, *46*(3/4), 201-234. doi: 10.1504/ijtm.2009.023374.

Cavanillas, J., Curry, E., & Wahlster, W. (2016). *New Horizons for a Data-Driven Economy. A Roadmap for Usage and Exploitation of Big Data in Europe* (pp. 29-37). Springer International Publishing AG Switzerland.

Chandarana, P., & Vijayalakshmi, M. (2014). Big Data Analytics Frameworks. In *International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)* (pp. 430-434). IEEE Xplore.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0*. [Place of publication not identified]: SPSS.

Chemuturi, M. (2013). Understanding Requirements. *Requirements Engineering And Management For Software Development Projects*, 13-32. doi: 10.1007/978-1-4614-5377-2_2.

Chen, H., & Chau, M. (2005). Web mining: Machine learning for web applications. Annual Review of Information Science and Technology, 38(1), 289–329. https://doi.org/10.1002/aris.1440380107.

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence And Analytics: From Big Data To Big Impact. *Business Intelligence Research. MIS Quarterly*, *36*(4), 1165-1188.

Chomsky, Noam, 1965, Aspects of the Theory of Syntax, Cambridge, Massachusetts: MIT Press.

Chuang, H., Chang, C., & Kao, T. (2014). Effective Web Crawling for Chinese Addresses and Associated Information. In *International Conference on Electronic Commerce and Web Technologies* (pp. 13-25). Munich, Germany: EC-Web 2014: E-Commerce and Web Technologies.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal Of Machine Learning Research*, *12*, 2493-2537.

Curry, E. (2016). The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. In J. M. Cavanillas, E. Curry, & W. Wahlster (Eds.), New Horizons for a Data-Driven Economy (pp. 29–37). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-21569-3_3.

Dahl, D. (2007). SOA and SaaS: Getting the Best of Both Worlds. Springcm, (Reference Series).

Denning, P. J. A New Social Contract for Research, Communications of the ACM (40:2), February 1997, pp. 132-134.

DeSanctis, G., & Poole, M. (1994). Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory. *Organization Science*, *5*(2), 121-147. doi: 10.1287/orsc.5.2.121.

Dias, J., & Ferreira, H. (2017). Automating the Extraction of Static Content and Dynamic Behaviour from e-Commerce Websites. *Procedia Computer Science*, *109*, 297-304. doi: 10.1016/j.procs.2017.05.355.

Doest, H., Iacob, M., Lankhorst, M., van Leeuwen, D., & Slagter, R. (2004). *Viewpoints Functionality and Examples* (pp. 1-92). Telematica Instituut/ArchiMate Consortium.

Draheim, D. (2010). Service-Oriented Architecture. Business Process Technology, 221-241. doi: 10.1007/978-3-642-01588-5_8.

Eliassi-Rad, T., & Shavlik, J. (2003). Intelligent Web Agents that Learn to Retrieve and Extract Information. In P. S. Szczepaniak, J. Segovia, J. Kacprzyk, & L. A. Zadeh (Eds.), Intelligent Exploration of the Web (Vol. 111, pp. 255–274). Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-1772-0_16.

Erl, T. (2012). SOA: Principles of Service Design. Upper Saddle River, NJ: Prentice Hall.

Fiorineschi, L., Frillici, F., & Rotini, F. (2018). Enhancing Functional Decomposition and Morphology with TRIZ: Literature Review. *Computers In Industry*, *94*, 1-15. doi: 10.1016/j.compind.2017.09.004.

Gouriten, G., Maniu, S., & Senellart, P. (2014). Scalable, generic, and adaptive systems for focused crawling (pp. 35–45). ACM Press. https://doi.org/10.1145/2631775.2631795

Hazzan, O., & Dubinsky, Y. (2008). Agile software engineering (Undergraduate topics in computer science). London: Springer. doi:10.1007/978-1-84800-198-5

Hasan, S., Shamsuddin, S., & Lopes, N. (2014). Machine Learning Big Data Framework and Analytics for Big Data Problems. *International Journal Of Advances In Soft Computing And Its Applications*, *6*(2), 1-14.

Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur, "The shark-search algorithm. an application: tailored web site mapping," Computer Networks and {ISDN} Systems, vol. 30, no. 1-7, pp. 317–326, 1998.

Hevner, A. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal Of Information Systems*, *19*(2).

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science In Information Systems Research. *Design Science In IS Research. MIS Quarterly*, *28*(1), 75-105.

Hoffmann, U., Silva, A., & Carvalho, M. (2018). Finding Similar Products in E-commerce Sites Based on Attributes. Retrieved from https://www.semanticscholar.org/paper/Finding-Similar-Products-in-E-commerce-Sites-Based-Hoffmann-Silva/ab960bd8095e243d15a631326c4013afafbea2f7.

Hongyu Liu, Milios, E., & Janssen, J. (2004). Focused Crawling by Learning HMM from User's Topic-specific Browsing (pp. 732–732). IEEE. https://doi.org/10.1109/WI.2004.10057

Hu, H., Wen, Y., Chua, T., & Li, X. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, *2*, 652-687. doi: 10.1109/access.2014.2332453.

Huang, W., Zhang, L., Zhang, J., & Zhu, M. (2009). Focused Crawling for Retrieving E-commerce Information Based on Learnable Ontology and Link Prediction (pp. 574–579). IEEE. https://doi.org/10.1109/IEEC.2009.127

IEEE Computer Society. (1998). IEEE Guide for Developing System Requirements Specifications. IEEE Std 1233- 1998. *The Institute Of Electrical And Electronics Engineers*. doi: 10.1109/ieeestd.1998.88826.

Ivanović, M., & Radovanović, M. (2015). Modern Machine Learning Techniques and Their Applications. In *5th International Conference on Electronics, Communications and Networks*. Shanghai, China.

Kashyap, P. (2017). Industrial Applications of Machine Learning. *Machine Learning For Decision Makers*, 189-233. doi: 10.1007/978-1-4842-2988-0_5.

Kaur, R., & Singh, T. (2010). Analysis and Need of Requirements Engineering. *International Journal Of Computer Applications*, *7*(14), 27-32. doi: 10.5120/1328-1653.

Kobayashi, M., & Takeda, K. (2000). Information Retrieval on the Web. *ACM Computing Surveys*, *32*(2), 144-173.

Kouamou, G. (2011). A Software Architecture for Data Mining Environment. *New Fundamental Technologies In Data Mining*. doi: 10.5772/13351.

Kruchten, P. (1995). Architectural Blueprints – The "4+1" View Model of Software Architecture. *IEEE Software*, *12*(6), 42-50.

Kune, R., Konugurthi, P., Agarwal, A., Chillarige, R., & Buyya, R. (2015). The Anatomy of Big Data Computing. *Software: Practice And Experience*, *46*(1), 79-105. doi: 10.1002/spe.2374.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges, 25.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436-444. doi: 10.1038/nature14539.

Lewis, D., & Jones, K. (1996). Natural Language Processing for Information Retrieval. *Communications Of The ACM*, *39*(1), 92-101. doi: 10.1145/234173.234210.

Liakos, P., Ntoulas, A., Labrinidis, A., & Delis, A. (2015). Focused Crawling For The Hidden Web. *World Wide Web*, *19*(4), 605-631. doi: 10.1007/s11280-015-0349-x.

Liu, H., & Milios, E. (2012). Probabilistic Models For Focused Web Crawling. *Computational Intelligence*, *28*(3), 289-328. doi: 10.1111/j.1467-8640.2012.00411.x.

Lodderstedt, T., Basin, D., & Doser, J. (2002). SecureUML: A UML-Based Modeling Language for Model-Driven Security. In J.-M. Jézéquel, H. Hussmann, & S. Cook (Eds.), ≪UML≫ 2002 – The Unified Modeling Language (Vol. 2460, pp. 426–441). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45800-X_33.

MacKenzie, C., Laskey, K., McCabe, F., Brown, P., & Metz, R. (2006). Reference Model for Service Oriented Architecture 1.0. OASIS Open, 1-31.

Manning, C., Raghavan, P., & Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Massimino, B. (2016). Accessing Online Data: Web-Crawling and Information-Scraping Techniques to Automate the Assembly of Research Data. *Journal Of Business Logistics*, *37*(1), 34-42. doi: 10.1111/jbl.12120.

Machine Learning in MATLAB- MATLAB & Simulink- MathWorks India. (2018). Retrieved from https://in.mathworks.com/help/stats/machine-learning-in-matlab.html?w.mathworks.com.

Martin, J. (2004). Study and Supervision Guide on Natural Language Processing, page 80.

McKinsey Global Institute (MGI). (2016). The Age of Analytics: Competing in a Data-Driven World (pp. 1-123). McKinsey & Company.

McIntosh, R. (2004). Open-Source Tools for Distributed Device Control Within a Service-Oriented Architecture. Journal Of The Association For Laboratory Automation, 9(6), 404-410. doi: 10.1016/j.jala.2004.08.011.

Menczer, G. Pant, and P. Srinivasan, "Topical web crawlers: Evaluating adaptive algorithms," ACM Trans. Internet Technol., vol. 4, no. 4, pp. 378–419, 2004.

Menczer, F. (2000). Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web, 40.

Méndez Fernández, D., & Penzenstadler, B. (2014). Artifact-Based Requirements Engineering: the AMDiRE Approach. *Requirements Engineering*, *20*(4), 405-434. doi: 10.1007/s00766-014-0206-y.

Micarelli and F. Gasparetti, "Adaptive focused crawling," in The Adaptive Web, ser. Lecture Notes in Computer Science, 2007, vol. 4321, pp. 231–262.

Michael, J., Riehle, R., & Shing, M. (2009). The Verification and Validation of Software Architecture for Systems of Systems. In *International Conference on System of Systems Engineering (SoSE 2009)*. Monterey, CA, USA: IEEE.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Morandini, M., Penserini, L., Perini, A., & Marchetto, A. (2015). Engineering Requirements for Adaptive Systems. *Requirements Engineering*, *22*(1), 77-103. doi: 10.1007/s00766-015-0236-0.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In *European Simulation and Modelling Conference - ESM'2011* (pp. 117-121). Guimaraes, Portugal.

Ng, A. Y., & Jordan, M. I. (n.d.). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes, 8.

Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2018). Outline Of A Design Science Research Process. In *4th International Conference on Design Science Research in Information Systems and Technology, DESRIST 2009*. Philadelphia, Pennsylvania, USA.

Oladipupo, T. (2010). Types of Machine Learning Algorithms. In Y. Zhang (Ed.), New Advances in Machine Learning. InTech. https://doi.org/10.5772/9385.

Oluikpe, P.I. (2015), "Knowledge creation and utilization in project teams", Journal of Knowledge Management, Vol. 19 No. 2, pp. 351-371.

Olston, C., & Najork, M. (2010). *Web Crawling*. Hanover: Now Publishers.

Osis, J., & Donins, U. (2017). Unified Modeling Language: A Standard for Designing a Software. Computer Science Reviews And Trends, 3-51. doi: 10.1016/b978-0-12-805476-5.00001-0.

Pavel, R., & Gurský, P. (n.d.). Focused Web Crawling of Relevant Pages on e-Shops, 5.

Pham, K., Santos, A., & Freire, J. (2018). Learning to Discover Domain-Specific Web Content. In *ACM International Conference on Web Search and Data Mining - WSDM 2018*. Los Angeles, California, USA.

Philip Chen, C., & Zhang, C. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314-347. doi: 10.1016/j.ins.2014.01.015.

Proper, E., & Greefhorst, D. (2010). The Roles of Principles in Enterprise Architecture. *Trends In Enterprise Architecture Research*, 57-70. doi: 10.1007/978-3-642-16819-2_5.

Osis, J., & Donins, U. (2017). Unified Modeling Language: A Standard for Designing a Software. Computer Science Reviews And Trends, 3-51. doi: 10.1016/b978-0-12-805476-5.00001-0.

Quarteroni, S. (2012). Natural Language Processing for the Web. *Lecture Notes In Computer Science*, 508-509. doi: 10.1007/978-3-642-31753-8_57.

Rungsawang, Angkawattanawit, "Learnable topic- specific web crawler", Journal of Network and Computer Applications, 2005, pp.97-114.

Russell, S., & Norvig, P. (2010). *Artificial intelligence. A Modern Approach* (3rd ed.). Pearson Education.

Sage, A., & Armstrong, J. (2000). *Introduction to Systems Engineering* (1st ed.). New York: Wiley-Interscience.

Sarnovsky, M., Bednar, P., & Smatana, M. (2018). Big Data Processing and Analytics Platform Architecture for Process Industry Factories. *Big Data And Cognitive Computing*, *2*(1), 3. doi: 10.3390/bdcc2010003.

Sekaran, U., & Bougie, R. (2016). Research Methods For Business: A Skill Building Approach Seventh Edition. John Wiley & Sons.

Service-Oriented Architecture Standards | The Open Group. (2018). Retrieved from http://www.opengroup.org/standards/soa.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal Of Innovation And Scientific Research*, *12*(1), 217-222.

Shestakov, D. (2013). Intelligent Web Crawling (WI-IAT 2013 Tutorial). *IEEE Intelligent Informatics Bulletin*, *14*(1), 5-7.

Shkapenyuk, V., & Suel, T. (2002). Design and Implementation of a High-Performance Distributed Web Crawler. In *18th International Conference on Data Engineering*. San Jose, CA, USA: IEEE.

Simon, H. (1983). Reason in human affairs (Harry camp lectures at stanford university, 1982). Stanford, Calif.: Stanford University Press.

Singh, S., & Singh, N. (2012). Big Data Analytics. In *International Conference on Communication, Information & Computing Technology (ICCICT)*. Mumbai, India: IEEE.

Sivarajah, U., Kamal, M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal Of Business Research*, *70*, 263-286. doi: 10.1016/j.jbusres.2016.08.001.

Snyder, K., & Khalid, A. (2013). Improving the Systems Engineering Requirements Analysis Process: A Few Tools and Techniques. In *5th Polytechnic Summit Wentworth Institute of Technology*. South Marietta Parkway, SE, Marietta, USA: Southern Polytechnic State University.

Spitkovsk, V., Jurafsky, D., & Alshawi, H. (2010). Profiting From Mark-Up: Hyper-Text Annotations For Guided Parsing. In *48th Annual Meeting of the Association for Computa- tional Linguistics (ACL 2010)*. Uppsala, Sweden.

Suh, N. (1998). Axiomatic Design Theory for Systems. *Research In Engineering Design*, *10*(4), 189-209. doi: 10.1007/s001639870001.

Sunil Kumar, M., & Neelima, P. (2011). Design and Implementation of Scalable, Fully Distributed Web Crawler for a Web Search Engine. *International Journal Of Computer Applications*, *15*(7), 8-13. doi: 10.5120/1963-2629.

Sutton, R. (1992). Introduction: The challenge of reinforcement learning. Machine Learning, 8(3-4), 225-227. doi: 10.1007/bf00992695.

Tsai, C., Lai, C., Chao, H., & Vasilakos, A. (2016). Big Data Analytics. In B. Furht & F. Villanustre, *Big Data Technologies and Applications* (pp. 13-52). Springer International Publishing.

Tsichritzis, D. The Dynamics of Innovation, in Beyond Calculation: The Next Fifty Years of Computing, P. J. Denning and R. M. Metcalfe (eds.), Copernicus Books, New York, 1998, pp. 259-265.

Vishnyakov, A., & Orlov, S. (2015). Software Architecture and Detailed Design Evaluation. *Procedia Computer Science*, *43*, 41-52. doi: 10.1016/j.procs.2014.12.007.

Waheed, Z. (2018). Translating customer needs into project decisions: identifying knowledge brokers in project networks. Development and Learning in Organizations: An International Journal. https://doi.org/10.1108/DLO-03-2018-0034.

Wieringa, R. (1996). *Requirements engineering*. Chichester: Wiley.

Wieringa, R. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer Heidelberg New York Dordrecht London: Springer-Verlag Berlin Heidelberg.

Xiao, S., Lin, Z., Guang-hong, G., Yan-Qiang, D., & Peng-fei, Y. (2009). Digital Product Data Exchange in Semantic Service-Oriented Architecture. COMPEL - The International Journal For Computation And Mathematics In Electrical And Electronic Engineering, 28(6), 1560-1578. doi: 10.1108/03321640910992083.

Wu, P., Wen, JR., Liu, H., Ma, WY.: Query Selection Techniques for Efficient Crawling of Structured Web Source. In Proceedings of ICDE2006, Atlanta GA, pp. 47--56 (2006).

Yogatama, D., Dyer, C., Ling, W., & Blunsom, P. (2017). Generative and Discriminative Text Classification with Recurrent Neural Networks. ArXiv:1703.01898 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1703.01898.

Yu, C., & Lin, S. (2010). Web crawling and filtering for on-line auctions from a social network perspective. *Information Systems And E-Business Management*, *10*(2), 201-218. doi: 10.1007/s10257-010-0135-3.

Zhang, Z., Du, Y., & Li, C. (2009). AntCrawlers: Focused Crawling Agents Based on the Idea of Ants (pp. 250–253). IEEE. https://doi.org/10.1109/CASE.2009.9.

Zheng, Q., Wu, Z., Cheng, X., Jiang, L., & Liu, J. (2013). Learning to crawl deep web. *Information Systems*, *38*(6), 801-819. doi: 10.1016/j.is.2013.02.001.

Zicari, R., Rosselli, M., Ivanov, T., Korfiatis, N., Tolle, K., Niemann, R., & Reichenbach, C. (2016). Setting Up a Big Data Project: Challenges, Opportunities, Technologies and Optimization. *Big Data Optimization: Recent Developments And Challenges. Studies In Big Data*, *18*, 17-47. doi: 10.1007/978-3-319-30265-2_2.

APPENDIXES

Appendix A: Table of the Interviews

Name	Organization	Role	Date
Ben van Rijnsoever	IBM Global Business Service (GBS), Netherlands	Lead Architect for Public Safety, Customs & Border Management	March 16th, 2018 April 6th, 2018 April 13th, 2018 April 25th, 2018 May 18th, 2018 June 1st, 2018 June 15th, 2018 July 6th, 2018 July 20th, 2018 August 8th, 2018
Gavin Shorten	IBM Research Lab, Ireland	Manager for the Innovation Exchange	June 7th, 2018
Bora Caglayan	IBM Research Lab, Ireland	Applied Researcher	June 7th, 2018
Marcel Molenhuis	Dutch Customs Administration	Senior Advisor for Data Analytics	May 3rd, 2018 June 6th, 2018 June 18th, 2018 June 28th, 2018 July 31st, 2018
Frank Heijmen	Dutch Customs Administration	Head of Trade Relations	April 17th, 2018
Maarten Veltman	Dutch Customs Administration	Chairman of the Innovation Committee	April 17th, 2018
Han Bosch	Dutch Customs Administration	National Coordinator for e-Commerce	May 3rd, 2018 July 31st, 2018
Ben Schmitz	Dutch Customs Administration	Venue E-Commerce System Coordinator	June 6th, 2018 July 31st, 2018
Jo Bootsma	Dutch Customs Administration	Open Source Intelligence Expert and Web-crawling Reference Person	May 3rd, 2018 June 28th, 2018
Jetze Baumfalk	Dutch Customs Administration	Data Scientist and Data Analytics Expert and Machine Learning Reference Person	May 3rd, 2018 June 18th, 2018

Appendix B: DCA 1st Meeting, Kick-off

Date: 17th April 2018
Location: Rotterdam, Laan op Zuid 45

Participants:

- Frank Heijmann (DCA)
- Maarten Veltman (DCA)
- Ben van Rijnsoever (IBM)
- Yao-Hua (TU Delft)

- From the Dutch Customs:
  - They are okay with searching/processing goods data on the internet.
  - They want to make something that can be used by everyone, also the other customs in the EU, because they want their solution to have political support.
  - One of the main problems is: in the E-commerce the buyer does not know the seller; in particular on e-commerce platforms (e.g. Alibaba).
  - In the long term, their goal is to have no officer anymore, they would like the risk assessment process of declarations to be fully automated – in the future PROFILE data analytics methods will fed directly PRISMA.
  - Any declaration-related data must be stored on servers within the EU, because of strict EU privacy legislation (GDPR).

- Technology to be developed:
  - Web-crawling of e-commerce sites
  - Machine Learning for recommendations of prices and websites to crawl
  - (the system shows the results of each step of the crawling analysis and provides a user interface which allows the DCA officers to make corrections. This information is captured to improve the system recommendations)

- Scoping:
  - Only the fiscal risk assessment, not security risk assessment.
  - Customs will select top-5 list of e-commerce products that have most inaccurate declarations with respect to fiscal risk assessment
  - Only reduction of the false positive, not false-negatives.
  - Only English language websites.

- Data to crawl online: mainly Price, if possible, also Size and Weight. In this way a fraudulent sender has to falsify and match more information.

- Future Maintenance (and already during the development):
  - The Web changes during the time, so sometimes experts should give new feedbacks to the system.
  - Technical experts might be needed to do actual web crawls in case the security/privacy measures to stop robots on the Web would become stronger in the future and would block automated web-crawling.

- Feedback:
  - The officers would train the system through a dashboard where they can like/dislike the recommendations they receive by the system.
  - New feedbacks will be needed when new products are added.

In the meeting I discussed the differences between the research part within PROFILE, and the engineering phase (scaling up and full implementation). According PROFILE, the web-crawling prototype will be implemented as a standalone dashboard. The customs expressed the desire to implement it in the risk assessment engine (PRISMA) in future, but this would require a different, more complicated approach because of the high declaration volume that have to be risk-assessed (and thus a different IT architecture would be required).

- Performance measurement system:
  - I have to discuss KPIs to measure the effectiveness of the web-crawling prototype that will be developed in PROFILE. I could take the declarations related to the 5 most critical goods and measure how the number of false positives is decreasing using the web-crawling technology with respect to the current method.
  - Knowing the saving costs for reducing a false positive, I could estimate its economic benefit. Also, a secondary research, I will study how this reduction of false-positive could have positive effect on logistic optimization of the supply chain.

- For the next meeting:
  - National Statistics Center.
    They already get all the declarations information from the customs without the identifiers. It could be useful to hear their best practices and how deal with the privacy and legal issues. (ACTION: Frank will send contact person info).
  - There is an existing database of historical goods data called XENON. At the next meeting there will be exert from Dutch Customs who explain what data are collected in the database and how they are used. I have to investigate how XENON could interact with the web-crawling technology (e.g. if XENON is used to forecast the future price of goods, it could call the web-crawling technology to verify the declared information which don't match the forecasts).

- Proposed roadmap for implementation of the web-crawling technology:
  - step 1: only data on the web.
  - step 2: combine with historical data (Xenon).
  - step 3: machine learning anonymous.
  - step 4: "safe room" (*) of Dutch Customs in Ireland.
  - step 5: sharing with other customs (when legislation is fixed).

(*) Safe Room: a virtual room that is part of the DCA internal network and where all people who have access are subject to the same procedures as currently used by DCA to give local people access to their data. This safe room might be a physical place in Dublin, but security will be controlled by DCA; or it might be in DCA with a remote VPN for us the IBM team in Ireland. This Safe Room must have a computer where I can copy the data on and where I can install our software.

Appendix C: DCA 2nd Meeting, E-commerce Scenario

Date: 3rd May 2018
Location: Rotterdam, Laan op Zuid 45

Participants:

- Marcel Molenhuis (DCA)
- Han Bosch (DCA)
- Jo Bootsma (DCA)
- Jetze Baumfalk (DCA)
- Ben van Rijnsoever (IBM)
- Yao-Hua (TU Delft)

From Han Bosch (detail information about e-commerce and their system):

- About 1 out of 3 declarations is wrong.
- There are so many false positive that the targeting officers have much more red flags to inspect than what they can physically check.
- They have a separate eCommerce import system called "Venue".
- Below 22 euros, no duty has to be paid and HS code is not mandatory. They only have goods description (no brand, no type, no model nr, etc).
- The good description can vary for the same product: e.g. "mobile phone", "gsm", "cell phone".
- The declarations also have the Weight information. This is useful because comparing the weight on the declaration with the crawled online can help to detect fiscal frauds. Example that was given: description declared as "phone cover" but there is also a phone inside, so the actual weight (on the declaration) is of course higher than the one crawled on the Web (which is only the cover).
- They consider requiring the Amazon Standard Identification Number (ASIN) – the amazon unique product identifier – as mandatory field, with which they can retrieve all details from Amazon (and same for e.g. Alibaba).
- There are 2 different teams in DCA, Venue and AGS (Venue is for all eCommerce pre-arrival and AGS for imports of the goods that have a value above 22 euro).
- The fiscal value and the duties to pay is decided before the parcels arrive at the customs (pre-arrival information). Usually the parcels are not risk-assessed in AGS (since this is done already via Venue). But AGS can do "manual" assessment (maybe post clearance audit, not sure).
- They also have a high percentage of false positives in the security issue, not the fiscal side (but I confirmed I won't address those ones in this part of the project).

The analytics team explained about other projects that have been already done, including good practices to take away and the technology which has been used: this might be useful to the team in Ireland to avoid duplication of work and research:

- They scrape Alibaba using Visual Web Ripper – a visual tool used for automated web scraping/harvesting and content extraction from the web. Alibaba allows to scrape products info as much as you want, but not the vendors info.
- It takes 3 weeks to get all products data from Alibaba with their methods.
- How do they retrieve the data? The scraper system is taught how to recognize the layout once and it works specifically for that web page (only Alibaba layout). This means the technology is not scalable to other e-commerce sites or for general search engine.
- The Xenon database has been discontinued. Tafeeic – a crawler, only text based and not looking to layout – was a project using the Xenon database. At the beginning it was working fine but then the Web changed, became more dynamic with less text, and the project has been abandoned.

- For some products they have reference prices (it is done for general products, not specifically for e-commerce). They also have a "small" database which comes from EU Brussels and contains reference prices for certain goods, but it is not used often.
- They already have some machine learning to reduce false positives, but they use it in the de-risking phase (i.e. to choose what parcel to inspect among all the red flags given by Prisma). Using machine learning this way, the justifications and instructions as produced by Prisma remain available. The machine learning system uses only a select number of features of declarations. It is planned to be integrated with Prisma within this year (2018).

I agreed that they will:

- Provide us a list of the 5 most critical categories of products, and the explanation of why they chose them. This is useful to reflect if those products are actually the right ones for the web-crawling prototype. For instance, if they choose a product which is always to be inspected for an if-then rule coming from EU directions, it would be useless to do web-crawling to reduce its false positives. Besides being critical products, they should also be some for which the web-crawling has good chances to be effective. Han Bosch is in charge to provide us this list.
- Provide us the contact of the national center of statistics reference person.
- Provide us the characteristics of the EU database, and a contact of reference.
- Confirm us that the solutions I proposed for the data privacy issues (anonymization and safe room, etc.) are feasible. They will ask their legal department.

Appendix D: DCA 3rd Meeting, the Venue System

Date: 6th June 2018
Location: Schiphol, Evert van de Beekstraat 384. Outlook Complex, Building F

Participants:

- Marcel Molenhuis (DCA)
- Ben Schmitz (DCA)

Relevant points about the Venue System:

- The DCA department for e-commerce receives files from the couriers where there are 3-4k items. These files are already structured and ready to be processed by Venue (DCA system for e-commerce). In particular each item is signed as:
    - "A" if it is below 22 euros (in reality there are more options according the type of product: it can also be below 45, or 150, etc.; they mentioned the law "commission regulation (EC) No 1126/2009"), and free to go;
    - "B" if it is a special product which should be below 1000 euros, and free to go as well;
    - "C" if above 22 euros, and thus to be forwarded to AGS (DMS); in this case, DMS will not execute the PRISMA risk assessment for these items since the risk assessment is already done in Venue;
    - "D" if to be stored in the warehouse because will be depart again.

- The Venue system (officially called "ProcessVenue") formats the files by the couriers if they do not comply, does the risk assessment, and gives 3 outputs:
    - Sends a reply to the couriers with the output of the risk assessment for each item (thus what item must be inspected and what is free to go);
    - Sends a list of items to inspect to another system called "PLATO" which cares of the inspections;
    - Sends a file to be added to the history archive. They have collected data for the last 6 years, and they should have data of around 30 millions items (it has to be confirmed).

- Thus, the dataset which would be useful for PROFILE would be a merge of two datasets: the final output of PLATO (inspection results) which is not in Venue (but it is in PLATO), and the history archive dataset of Venue.

- Marcel said that he has to ask the department in charge whether it would be ok for IBM to have the results of the inspections. Also he mentioned that the previously proposed solution of the "Safe Room" might have problems, but I did not discuss this further as he is waiting for further information by the right people.

- Venue does not interface to PRISMA or BLAZE, but it does the risk-assessment itself. The risk assessment is done mainly on the base of the traders (similar to the containers trade; obviously it is less effective in the case of the e-commerce because the traders are many and fast-changing).

- Venue runs automatically only for a Dutch courier (I think is ACC but I am not sure), otherwise it usually calls another system called "SelectieTool" used by the targeting officers (the same officers could use here the web-crawler and give feedbacks).

- The DCA wants to move everything on AGS within 2 years (thus Venue will be abandoned), and they want to do a new AGS system in 2019.

Confirmations about the web-crawling tool:

- The main problems are fake value on the declarations, and products that cannot enter the EU (because of illegal/dangerous material or not respecting the European laws).
- Sometimes the description of the object item is not clear: the specific type or model is missing, or the number of pieces inside a box is not declared (it might be there are 10 leather jackets inside the package and the declaration just says "leather jacket").
- For each item, they don't know the e-commerce website, but they do know the manufacturer (although this is a separate column, and the description of the product does not have the brand of the producer).
- They have only the gross weight and it is provided by the couriers (DCA does not measure the weight of the packages; the couriers measure the gross weight carefully because according to the weight of the package they price their service).
- They do not have something like the "tweakers pricewatch", and for the e-commerce parcels, it would not be much useful to have the historical prices because the e-commerce deliveries are fast (around a weak or little more). It might be useful to have a similar database with the average price of each item in each e-commerce and company website.
- About the idea of the web-crawling tool to be used to support the inspection only (i.e. using the tool once the packages are open only), they said it would be useful, but that they would prefer to use it for the risk assessment as well (i.e. finding information online and deciding whether to order an inspection or not). I repeated that with only the description of the item and its weight, without the manufacturer, it might be extremely hard to identify the product and its right value. Marcel said that the DCA might be willing to provide IBM also the senders' information (but Marcel has to ask the legal department).

I concluded the meeting scheduling another meeting on Monday 18 June focused on the previous web-crawling and machine learning projects carried out within the DCA (the ones mentioned during our previous meeting; the same experts will attend). I thought this would be very useful after talking to the IBM experts in Ireland.

Appendix E: DCA 4th Meeting, Machine Learning Project

Date: 18th June 2018
Location: Rotterdam, Laan op Zuid 45

Participants:

- Marcel Molenhuis (DCA)
- Jetze Baumfalk (DCA)
- Ben van Rijnsoever (IBM)
- Yao-Hua (TU Delft)
- Boriana Rukanova (TU Delft)

Notes on the initial use case: the web-crawler will be an interactive tool with focus on algorithms and not on the engineering phase:

- For PROFILE the goal is not to cover the engineering phase which addresses issues for the production environment (issues like performance etc.).
- The focus in PROFILE is to define the algorithms, not to design a production system.
- The prototype will need to:
    - Show whether it is possible for the web-crawler to find relevant eCommerce websites.
    - Show whether it is able to find relevant price information.
- The machine learning will be based on the above, and not related to the engineering phase.
- Type of product – the 5-10 product categories that were identified by Dutch Customs.

I focus on valuation and not on misleading goods description (for the initial phase).

- The goal is to find an answer to the question is this feasible to find 1) eCommerce websites and 2) the right price information based on goods descriptions that are found in the customs declaration (yes/no). For the price it may be possible also to look at range of prices etc. (e.g. price plus confidence information). If this is not feasible other scenarios will be explored.
- It is assumed that the goods description is not misleading.

Envisaged methods to be used:

- Natural language processing;
- Machine learning;
- Reinforcement learning;
- Supervised learning.

The process of how the machine learning/ reinforcement learning will take place. For both the finding of eCommerce websites, as well as for finding price information the same approach:

- First an expert will start with supervised learning and Natural Language Processing by doing a search using key words and URL information. The expert would need to identify which are eCommerce websites and try to define reasons why he thinks this is an eCommerce website. This reflection will serve as a basis for Reinforced learning.
- Experts from Dublin will do themselves the 1st prototype using NLP, supervised learning and reinforcement learning; At a second steps officers will be used to do the supervised learning (reinforcement learning).
- Same approach will be used for both web-crawling to find eCommerce websites, as well as for finding the price information.

Legal aspects to be checked/ considered:

- To check whether it is possible to use brand name form the goods description (whether this is legally allowed).
    - Working assumption for the moment that needs to be checked is:
    - Information about goods description (including brand name) is allowed to use;
    - Information about buyer/ seller (e.g. manufacturer name) is not allowed under GDPR.
- How to check whether websites that are visited allow robots to collect information
    - It needs to be investigated how this could be done, normally the first time a website is visited a human would need to read the legal conditions on the website.

Further Notes on the Machine learning methods (discussed during the presentation of Jetze):

- Machine learning methods used: Random forest; Radom forest good in finding non-linear correlations.
- Belgium wants to look at behaviour of a trader. This can be seen as anomaly detection. You have different techniques, partially unsupervised; you want to model what is normal and what not. Normal is average on the data. It may be possible to use Bayesian modelling (conditions, probabilities); e.g. a company normally visits these ports, now another port. You can give probability but also human readable text to explain how you got to this probability.
- Presentation Jetze (Marcel already shared the slides).

More notes about the Machine Learning Mode they are using at the DCA:

- The machine learning model is created using Python. It can also be loaded in Blaze.
- Without the machine learning model, the average hit rate of rules is 5%, and the one by the targeting officers is 10%.
- The result of the machine learning model is a number between 0 and 1 according to the relevance of the risk. This allows to make one more choice in case the officers are not enough (i.e. inspecting the packages with the highest score).
- The machine learning model is "after" the PRISMA/BLAZE risk engine (de-risking phase) because it needs the dataset with the inspection results. After PRISMA/BLAZE denote a product as to inspect or not, PLATO records the result of these inspections as Y/N compliant. Only the data with this final Y/N label can be used by the machine learning model.
- Results from Plato (called Labeling) is free text, so understanding this was a challenge. Needed to clean / preprocess the data (label in inspections, remove declarations that have more than one type of product). From an 100% of dataset, 85% was left as good to use.
- 75% of this data set is used to train the model. 25% is used to test it.
- They are using data recorded since 2014.
- When they deployed the ML model, they had a validation period of 3-month shadow-running, which means that the DCA let the ML model running in parallel with the existing solution, with real data, so that they could compare the actual findings to assess the model.
- Having less inspection leads to have less false positive but also having more false negative. A Receiver Operating Characteristic (RoC) curve is used to see the accuracy of the machine learning model. Curve: %reduction inspection vs %missed hits. It is to plot a curve to see how much decreasing of false positive it is possible to have by letting go some false negative. In our case it was 5% more misses leaded to a decrease of the number of inspections by 10%.
- At the end, it is all a human decision whether to inspect or not.
- The challenge is to track the results of the machine learning model to the declarations parameters so that it is possible to update the business rules of the risk engine.

TO-DO list

- Marcel will check whether historic data can be released for analysis after the legal issues have been resolved. [Marcel]
- The confidentiality requirements to be checked:
    - Marcel to check what are the confidentiality requirements for TUD [Marcel]
    - Boriana to check with TUD what is the non-disclosure agreement that is used for a Master thesis [Boriana]
    - The Master Thesis of Alessandro will be confidential [Alessandro]
- To conform the working assumption that information related to goods description (e.g. brand name) can be used for the web-crawler and information about people and companies cannot at this stage due to GDPR [Marcel has to confirm that]

Appendix F: DCA 5th Meeting, Web-crawling Project

Date: 28th June 2018
Location: Rotterdam, Laan op Zuid 45

Present at the meeting:
- Marcel Molenhuis (DCA)
- Jo Bootsma (DCA)
- Boriana Rukanova (TU Delft)

The DCA developed two projects, one for web-crawling (just indexing) and one for web-scraping (retrieving information). The web-crawling is not used anymore, because too old technology, while the web-scraping is currently used and could be useful in future.

The interview has been carried out with also a live demonstration of the web-crawling currently in use by the DCA.

About the Web-crawling project:
- The first version of the project was Xenon, made for the British and Dutch Customs 10 years ago by an external company
- There has been an updated version 3 years ago called Tafeic with also Sweden and Belgium involved
- Its technology deployed can only handle text-based web content and is not able to retrieve information in a more dynamic web populated with multimedia data as is often used today
- The system takes as input a list of websites to crawl written on a txt text and returns a list of relevant words with their weight. The example of a search for medicines it has been showed. After a manual search, a list of useful websites is identified and written on a txt file. This file is given to the crawler which returns for instance "Viagra" and its weight, and same for other common medicines, after it crawled all the provided websites.
- It was meant to do investigation on request of the business intelligence department. Today is not used.
- This crawler also had the possibility of being trained through feedbacks to improve its accuracy.
- They did not consider the problem of the terms and conditions of websites which might not allow robots to crawl their information. The excuse was that if the data are stored just temporarily for investigations there is no problem.

About the Web-scraping project:
- It is a more recent project only by DCA.
- The DCA uses Visual Web Ripper () to scrape all the information starting from an URL and save it in a database. After the URL is inserted, the software goes to that page (as a normal browser) and the user can select the elements of the page that the software should save in the database (thus it recognizes the page layout).
- Right now, they are working on making a database with information about 5/10 chosen products.
- It could be useful to create useful database with personal information (that can't be used by externals), or weight information.
- Visual Web Ripper can also recognize discounted prices and other features.

About the E-commerce experiences:

- Alibaba does not show the shipping cost at the first generation (one further crawl is required).
- eBay has the shipping cost showed below in the same page.
- AliExpress is slower than Alibaba in terms of response time.
- Considered 22000 results for chargers on Alibaba, only 400 had the weight information.

Appendix G: Requirements Validation Hand-Out

Requirements Validation

Tuesday, July 31st

Dutch Customs Administration
Rotterdam Laan op Zuid 45, 12e Verdieping

Interviewee: _____

Title: _____

This document is to validate the requirements for the design of the architecture.
With "architecture" it is meant the representation of the PROFILE web-crawling system.
The question to be investigated is: are we building the right product?

Please cross "YES" or "NOT" in the table below whether you consider these statements about the requirements of the architecture true or not.

- "FR" stays for Functional Requirement: what functionality the architecture should have.
- "NFR" stays for Non-functional Requirement: how the architecture should function (from a technical point of view).
- "C" stays for Constraint: architecture limitation deriving from external factors (e.g. jurisdiction).

OBJECTIVES & REQUIREMENTS VALIDATION		**Confirmed?**	
		Yes	No
Goal	Cross-validating the Packages' Values on the Declarations with the Products' Prices found on e-Commerce Platforms		
FR1	The architecture must be able to interact with the user which is the targeting officer		
FR2	The architecture must be able to retrieve the weight information of the product		
FR3	The architecture must be able to interact with and search on the web		
FR4	The architecture must be able to find products and their prices online		
FR5	The architecture must be able to compute prices deviations and return a risk indicator of green or red flag accordingly		
FR6	The architecture functionality must be in future generalizable to any categories of products, countries of origin, and e-commerce website		
FR7	The architecture functionality must be able to adapt to the dynamics of the web and expertise of the targeting officers		

NFR1	The architecture must be able to interpret different types of data		
NFR2	The architecture must be able to interpret vague and inconsistent information		
NFR3	The architecture must be able to extract the right knowledge from the data		
NFR4	The architecture must be able to analyze and filter the search results		
NFR5	The architecture must be able to choose the right websites and products among the search results		
NFR6	The architecture must be able to save every result of analysis to improve its performances		
NFR7	The architecture must be able to check whether the product of the current declaration was recently processed		
C1	The architecture must function without using the sender and receiver information		

Please write below if there is anything missing that you think it should be included:

_____

_____

_____

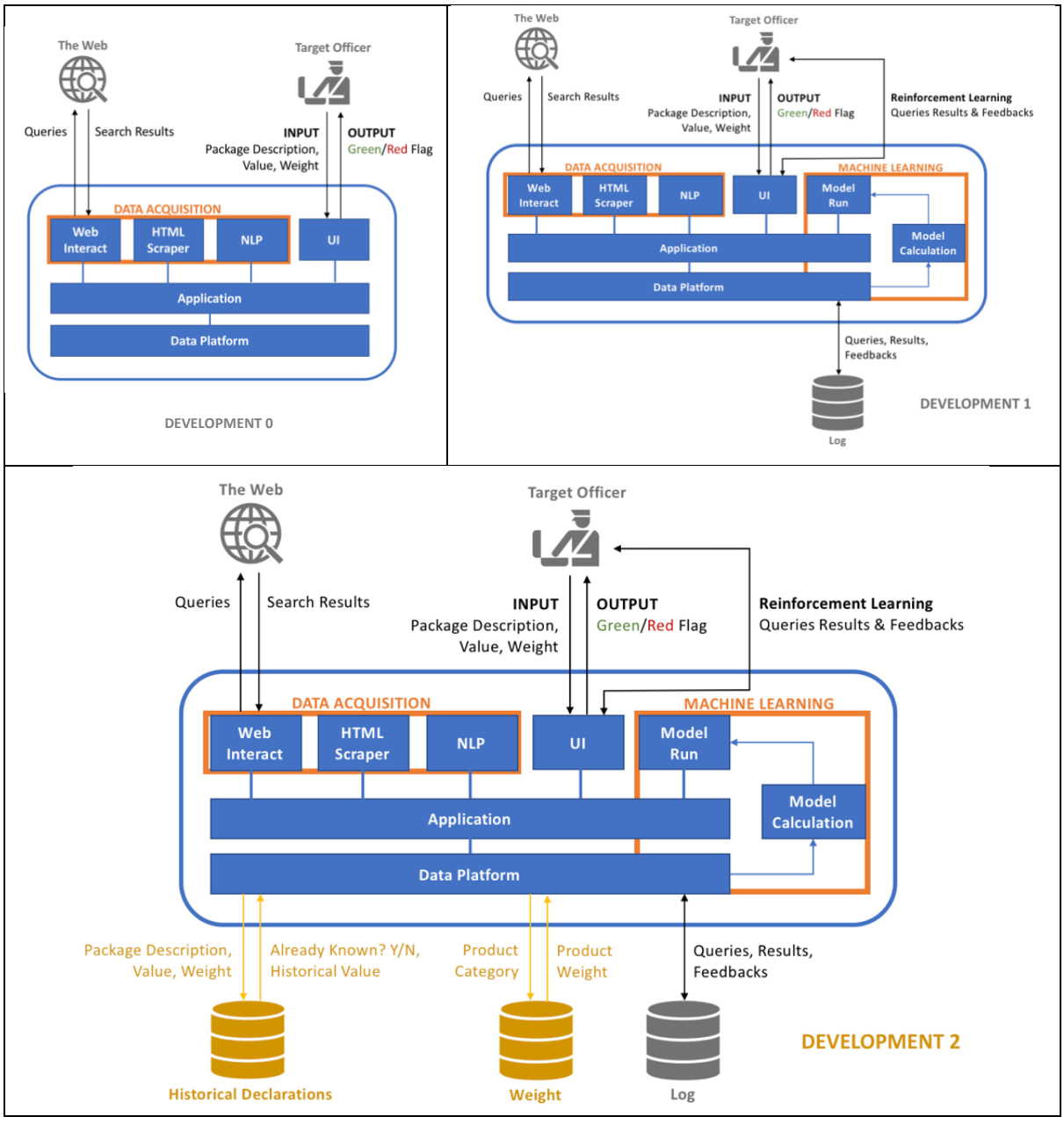
_____

THANK YOU

Appendix H: Requirements-Interviewees Map

Requirements and main pro		Interviewee
FR1	The architecture must be able to interact with the user which is the targeting officer	Project Spec
FR2	The architecture must be able to retrieve the weight information of the product	Project Spec
FR3	The architecture must be able to interact with and search on the web	Project Spec
FR4	The architecture must be able to find products and their prices online	Project Spec
FR5	The architecture must be able to compute prices deviations and return a risk indicator of green or red flag accordingly	Project Spec
FR6	The architecture functionality must be in future generalizable to any categories of products, countries of origin, and e-commerce website	Marcel Molenhuis, Frank Heijmann, Maarten Veltman
FR7	The architecture functionality must be able to adapt to the dynamics of the Web and expertise of the targeting officers	Marcel Molenhuis, Frank Heijmann, Maarten Veltman
NFR1	The architecture must be able to interpret different types of data	Ben van Rijnsoever
NFR2	The architecture must be able to interpret vague and inconsistent information	Han Bosch
NFR3	The architecture must be able to extract the right knowledge from the data	Research
NFR4	The architecture must be able to analyze and filter the search results	Ben van Rijnsoever
NFR5	The architecture must be able to choose the right websites and products among the search results	Research
NFR6	The architecture must be able to save every result of analysis to improve its performances	Ben van Rijnsoever
NFR7	The architecture must be able to check whether the product of the current declaration was recently processed	Ben van Rijnsoever
C1	The architecture must function without using the sender and receiver information	Marcel Molenhuis
NFR8	The architecture must response to the user in less than one minute (this has been added during the validation interview)	Ben Schmitz

Appendix I: Development Plan



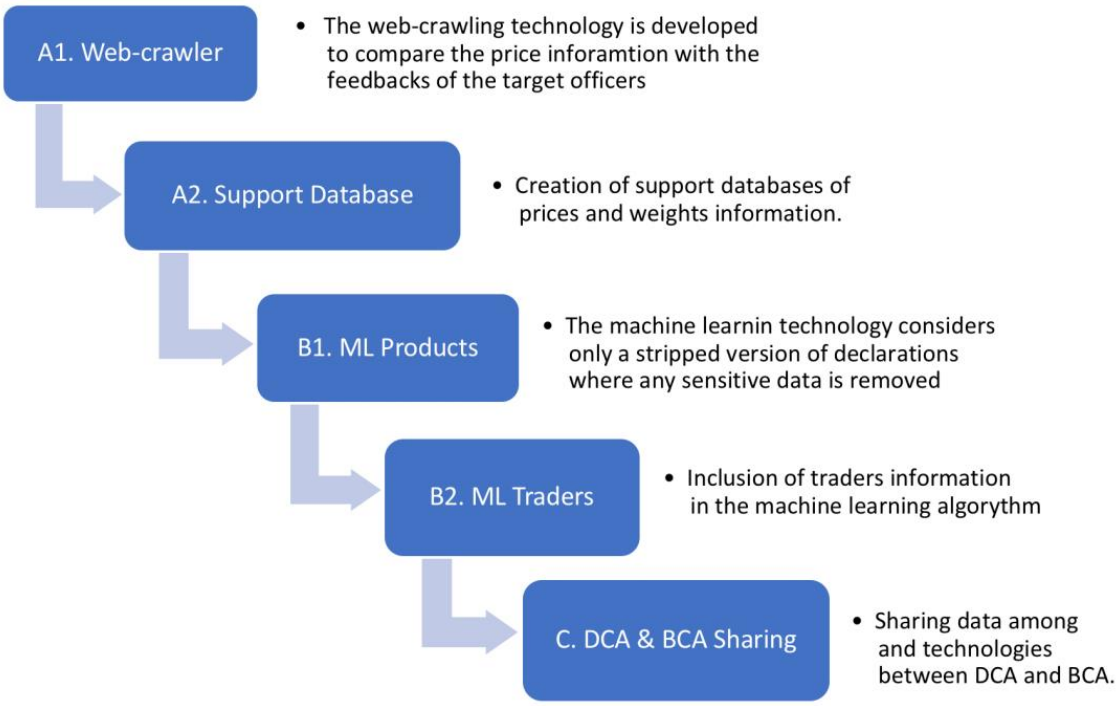
The development of the web-crawling system will be made in two steps. This choice is driven by the requirements of modularity and data sharing constraint. The modular development allows more interaction with the client and more effective solution delivery. At the same time, a multi-step approach allows to immediately start the development of the core functionalities and providing time for the policy makers to find the right policy framework to satisfy the data sharing constraint.

In the first step, the core functionalities of the web-crawler will be developed. This means only (1) the user interface that the targeting officers will use to input the declaration data and (2) the acquisition capability of the web-crawling, including the HTML parsing and the NLP capability. This core development is called “development 0”.

Then the recommendations models will be developed, including the feedbacks loops and the logic which makes the risk indicators. In this phase, the application services to support the machine learning models are developed, including the log dataset to store every result and improve the models. This phase is called “development 1”.

Finally, the last step focuses on the development of the two datasets to operationalize the web-crawling architecture functionalities: the one with the weight information, and the one with the historical declarations. As explained early, both of them will include only the information of the five categories of products that have been described in the use case section. The development of these two datasets has to be done by the experts from the DCA, since they possess the data. The database with the historical information will have the maximum/minimum and average values of the historical declarations. The database with the weight information will be made with an accurate research of weight information for each item considered within historical and external sources.

Appendix J: PROFILE Netherlands Roadmap



Given the multi-step approach used for the web-crawling technology, a similar approach will be used for the rest of the working package of PROFILE in the Netherlands. In particular, a research roadmap is provided. The web-crawling technology is here broken in two steps instead of three for simplicity. In this case, the step "A1" concerns both the "development 0" and "development 1" described earlier. The step "A2" is the about the "development 2".

In this multi-step approach, each technology is developed in two steps, thus first the core functionalities, and then the support and improvement elements. The order is obviously before the web-crawler, as it is the core objective for the working package of the Netherlands, and then the machine learning on the historical data, as it will be already developed in Belgium. Finally, the final step is to merge the DCA and BCA data and technologies. The ideal situation would be a European shared platform among all the EU countries, because the more data the better, and because it would be obviously wise to share valuable experience, but this is far out the scope of this research.

From the step B2 on, the Dutch living lab would benefit of the developments from the other living labs in Belgium – which focuses on the machine learning on the traders and so it needs to use the personal information – and the one in Sweden – which focuses on the data sharing between Sweden and Norway.