D10.1 -
Dynamic Supply Chain Networks Pilots Design

<table>
<thead>
<tr>
<th>Project Acronym</th>
<th>TT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title</td>
<td>Transforming Transport</td>
</tr>
<tr>
<td>Grant Agreement number</td>
<td>731932</td>
</tr>
<tr>
<td>Call and topic identifier</td>
<td>ICT-15-2016-2017</td>
</tr>
<tr>
<td>Funding Scheme</td>
<td>Innovation Action (IA)</td>
</tr>
<tr>
<td>Project duration</td>
<td>30 Months [1 January 2017 – 30 June 2019]</td>
</tr>
<tr>
<td>Coordinator</td>
<td>Mr. Rodrigo Castiñeira (INDRA)</td>
</tr>
<tr>
<td>Website</td>
<td><a href="http://www.transformingtransport.eu">www.transformingtransport.eu</a></td>
</tr>
</tbody>
</table>

Funded by the European Union’s H2020 GA - 731932
Document fiche

Authors: Eleni Zampou [AUEB], Christina Milioti [AUEB], Fraidaki Katerina [AUEB], Aggelos Liapis [AUEB], Ioanna Kelepouri [AUEB], Lefteris Manousakis [AUEB], Vega Rodrigálvarez [ITAINNOVA], Juan José Navamuel Castillo [ITAINNOVA], Alberto Capella Garcés [ITAINNOVA], George Bravos [ITML], Giorgos Panagiotopoulos [INTRASOF], Florian Flocke [FRAUNHOFER], Julian Eggemann [FRAUNHOFER]

Internal reviewers: Henning Hasemann [TOMTOM], Iker Guinea [CI3]

Work Package: WP10

Task: T10.1

Nature: R

Dissemination: PU

Document History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Contributor(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>31/01/2017</td>
<td>AUEB</td>
<td>ToC</td>
</tr>
<tr>
<td>0.2</td>
<td>28/02/2017</td>
<td>All partners</td>
<td>A first consolidated version</td>
</tr>
<tr>
<td>0.3</td>
<td>15/03/2017</td>
<td>All partners</td>
<td>A draft version to be submitted for internal review</td>
</tr>
<tr>
<td>0.4</td>
<td>20/03/2017</td>
<td>All partners</td>
<td>An updated consolidated version to be reviewed by pilot partners</td>
</tr>
<tr>
<td>0.5</td>
<td>25/03/2017</td>
<td>All partners</td>
<td>An updated consolidated version that incorporates pilot partners comments</td>
</tr>
<tr>
<td>0.6</td>
<td>28/03/2017</td>
<td>All partners</td>
<td>An updated consolidated version that incorporates internal reviewers comments</td>
</tr>
<tr>
<td>0.7</td>
<td>31/03/2017</td>
<td>AUEB</td>
<td>Final version to be submitted</td>
</tr>
</tbody>
</table>
Keywords: e-commerce logistics, shared logistics, collaboration, big data, business analytics

Abstract (few lines):
The TT pilot main idea is to embrace key aspects of e-commerce logistics and the various stakeholders' views (customers, online retailers, couriers, 3PLs) and demonstrate how the deployment of Big Data technologies can transform them in order to provide better logistics services and enhance the logistics processes performance.

Following a coherent method of work, as analysed in Section 2.1.1, we have clarified in sufficient detail the current practices and the envisioned ones that rely on collaboration in e-commerce logistics and will, to a large extent, be supported by the big data technologies. Hence, the aim of this deliverable is to assimilate novel knowledge from the pilot regarding users’ views and requirements for e-commerce logistics in order to frame the pilot’s implementation and also structure the requirements of the big data infrastructure and the analytics techniques applied.

This deliverable presents in detail the Dynamic Supply Chain Networks Pilots Design process and provides the conceptual backbone of the overall pilot, enabling common understanding among project partners for the different use cases and supporting requirement acquisition and analysis.

DISCLAIMER
This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content. This document may contain material, which is the copyright of certain TT consortium parties, and may not be reproduced or copied without permission. All TT consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.
Neither the TT consortium as a whole, nor a certain party of the TT consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

ACKNOWLEDGEMENT

This document is a deliverable of TT project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 731932
# Table of Contents

TABLE OF CONTENTS .................................................................................................................. 5  
TABLE OF FIGURES .................................................................................................................... 6  
DEFINITIONS, ACRONYMS AND ABBREVIATIONS .................................................................. 6  
EXECUTIVE SUMMARY ............................................................................................................. 7  
1 MOTIVATION AND AMBITION ............................................................................................. 10  
2 DESIGN OF PILOT .................................................................................................................. 13  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 REQUIREMENTS</td>
<td>13</td>
</tr>
<tr>
<td>2.1.1 User Requirements Capturing and Elicitation Approach</td>
<td>13</td>
</tr>
<tr>
<td>2.1.2 Literature review insights</td>
<td>16</td>
</tr>
<tr>
<td>2.1.3 In-depth interview insights</td>
<td>20</td>
</tr>
<tr>
<td>2.1.4 3PL user insights</td>
<td>22</td>
</tr>
<tr>
<td>2.1.5 Key outcomes and pilot requirements</td>
<td>26</td>
</tr>
<tr>
<td>2.2 OBJECTIVES</td>
<td>26</td>
</tr>
<tr>
<td>2.3 USE CASES AND SCENARIOS</td>
<td>28</td>
</tr>
<tr>
<td>2.4 DATA ASSETS</td>
<td>30</td>
</tr>
<tr>
<td>2.5 BIG DATA TECHNOLOGY, TECHNIQUES AND ALGORITHMS</td>
<td>32</td>
</tr>
<tr>
<td>2.5.1 IDS Connectors</td>
<td>32</td>
</tr>
<tr>
<td>2.5.2 Moriarty Analytics and Data Storage</td>
<td>34</td>
</tr>
<tr>
<td>2.5.3 Kajal Routing</td>
<td>38</td>
</tr>
<tr>
<td>2.5.4 Kajal Forecasting</td>
<td>41</td>
</tr>
<tr>
<td>2.5.5 DISMOD</td>
<td>43</td>
</tr>
<tr>
<td>2.5.6 Inventory Routing Algorithm</td>
<td>45</td>
</tr>
<tr>
<td>2.5.7 ITML Optimisation Solution</td>
<td>48</td>
</tr>
<tr>
<td>2.5.8 INTRASOFT Data Services</td>
<td>49</td>
</tr>
<tr>
<td>2.6 POSITIONING OF PILOT SOLUTIONS IN BDVA REFERENCE MODEL</td>
<td>51</td>
</tr>
<tr>
<td>2.7 BIG DATA INFRASTRUCTURE</td>
<td>53</td>
</tr>
<tr>
<td>2.8 ROADMAP</td>
<td>56</td>
</tr>
<tr>
<td>3 CONCLUSIONS</td>
<td>58</td>
</tr>
<tr>
<td>4 REFERENCES</td>
<td>59</td>
</tr>
<tr>
<td>5 ANNEX I: INTERVIEW PROTOCOL</td>
<td>62</td>
</tr>
</tbody>
</table>
Table of Figures

Figure 1. European B2C e-Commerce Sales ................................................................. 10
Figure 2. European B2C e-Commerce Growth Rate ....................................................... 10
Figure 3. User Requirements Capturing and Elicitation Approach .................................. 14
Figure 4. Companies Profile and Logistics Characteristics .............................................. 15
Figure 5. Big data applications in logistics .................................................................. 17
Figure 6. e-Commerce Logistics Process up to Courier .................................................. 23
Figure 7. WMS Procedure ......................................................................................... 24
Figure 8. Data Flow from e-shop to WMS .................................................................... 25
Figure 9. Pilot Requirements, Objectives and Use Cases ................................................ 30
Figure 10. Connector Communication ........................................................................... 34
Figure 11. Moriarty Structure ....................................................................................... 35
Figure 12. Moriarty WF working with Spark ................................................................. 37
Figure 13. Moriarty Machine learning algorithm example .............................................. 38
Figure 14. Kajal Routing Example ................................................................................. 39
Figure 15. Service Level Analysis ................................................................................. 45
Figure 16. Sentiment Analysis Outcomes .................................................................... 49
Figure 17. Big Data Value Reference Model ................................................................. 52
Figure 18. Big Data Infrastructure ............................................................................... 54
Figure 19. Big Data Infrastructure Implementation Approach ....................................... 57

Definitions, Acronyms and Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>3PLs</td>
<td>Third party logistics</td>
</tr>
<tr>
<td>POS</td>
<td>Point of sales</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Executive Summary

Digital evolution has changed Europeans shopping habits and expectations and the number of internet users that become internet shoppers increases steadily (57% of European Internet users shop online (Ecommerce Europe, 2016)). According to a recent expert study by GfK, the driver that is currently most influential and is having most impact on the retail scene and its success is to meet shoppers’ convenience expectation. At the same time, the omni-channel growth has created new challenges for the retailers. Whether their business is online, in physical stores or both, all retailers face similar challenges. They have to deliver a seamless customer experience at every touchpoint, maximize sales across every channel and device, and live up to their promises regarding product availability and delivery. Therefore, the e-commerce trend has created a dynamic and turbulent environment where the density of the distribution network consisted by many delivery points (orders and returns from various consumers at different points), multiple delivery channels and last-mile delivery requirements significantly increases the distribution cost.

While we witness this significant growth of e-Commerce market and the logistics is considered to be a basic part of e-commerce (Bask et al., 2012), it is undoubtedly an area where big data can be extremely fertile (Swaminathan, 2012). Personalized services, dynamic pricing, predictive analytics, supply chain optimization and visibility are some of the core big data application areas in these fields (Akter et al., 2016; Koutsabasis et al., 2008; Mehra, 2013). In the context of this pilot, we are going to investigate the role of big data in transforming the current e-commerce logistics practices and bringing value on all the involved stakeholders 3PLs, online retailers and final consumers. More specifically, we will investigate the role of collaboration in e-commerce and explore how the big data can enable the collaboration in this context. We will also try to provide alternative shipping methods to the consumer in order to respond to this growing e-commerce consumers' demand. By taking into account the high competition and the under exploitation of big data in the e-commerce logistics industry, we will also investigate how the application of advanced business analytics techniques that exploit the various types of data can transform and enhance the existing e-commerce logistics processes.

The ambition is to provide generic guidelines about the role of big data in e-commerce logistics and provide a roadmap that will describe how the big data solutions can be applied for tackling specific requirements in the e-commerce logistics. In order to materialize our ambition, we need firstly to capture the requirements for this pilot and reveal the needs of the various big data solutions in e-commerce logistics context. Hence, the aim of this deliverable is to assimilate novel knowledge from the pilot regarding users’ views and requirements for e-
commerce logistics in order to frame the pilots’ implementation and also structure the requirements of the big data infrastructure and the analytics techniques applied.

Following a coherent method of work, that employs a literature review method, individual in-depth interviews with selected industry experts, on-site observation and discussions with the pilot 3PL user, we have clarified, in sufficient detail, the current practices and the envisioned ones that rely on collaboration in e-commerce logistics and can, to a large extent, be supported by the big data technologies. More specifically, we conclude to a set of pilot requirements that express the needs towards the TT big data solutions from the side of the (end-) users of the e-commerce domain. Then, we translate them into a set of pilot objectives where we discuss their implementation process and, then, formulate a set of application scenarios-use cases. The following figure summarizes the pilot requirements, objectives and use cases that came out.

The use cases data needs will be served by a combination of Logika available data sets and a set of various open data; and a big data infrastructure will be also deployed. Different technologies and algorithms, such as Kajal Routing, Kajal Forecasting and DISMOD, will be used to support the pilot scenarios over the different layers of the architecture.
The pilot will be implemented in 3 stages. The Stage 1 (end of month 6) will outline the scenarios in a smaller scale with basic big data technologies running on the cluster and by integrating a small amount of data manually. The second stage (end of month 12) of the big data infrastructure required the first implementation version of the IDS Connector Reference Model. During Stage 2, additional data modules will be integrated in the model. Stage 3 refers to the fully implemented big data infrastructure, which consists of several data sources and a multi-functional infrastructure to share, store and communicate data based on the defined architecture. By employing the iterative approach adopted in the overall project, we will refine the scenarios at each stage by taking into account their results.
1 Motivation and Ambition

Digital technology changes the way consumers shop and the way consumers wish to receive their purchases. Shopping habits have changed rapidly during the last decade and a high percentage of consumers now shop online, following the spread of IT systems such as laptops, tablets and smartphones. Today, 57% of European Internet users shop online (Ecommerce Europe, 2016). The e-commerce sector is booming as almost all growth in retail comes from e-commerce. The European B2C e-commerce sales have been growing steadily since 2011. Still, the growth rate has decreased the last few years, from 18.4% in 2011 to 13.3% in 2015. This trend is expected to continue in 2016, as a growth rate of 12.0% has been forecasted, resulting in a European B2C ecommerce turnover of €509.9bn (Ecommerce Europe, 2016) as shown in Figure 1 and 2. Greece is included among the TOP 10 countries in terms of B2C e-Commerce growth rate with a growth rate of 18.8%.

However, the full potential of the European e-commerce market has not yet been reached as the e-commerce penetration can still be increased; only 16% of SMEs sell online – and less than half of those sell online across borders (7.5%). According to a recent expert study by Glaxo SmithKline, the driver that is currently most influential and is having most impact on the retail scene is shoppers’ convenience expectation: a very demanding issue in an increasingly complex retail and shopping environment (Glaxon SmithKline, 2016).

Moreover, customers today are demanding a high level of service from retailers, regardless of how they buy products or receive delivery. The omni-channel growth has created new challenges for the retailers. Whether their business is online, in physical stores, or both, all retailers face similar challenges. They have to deliver a seamless customer experience at every touch point, maximize sales across every channel and device, and live up to their promises regarding product availability and delivery. Taking a look at the UK and Germany, we see that more than 65% of shoppers aged 21 or younger prefer to have both a physical and an online experience. Consumers expect shopping to be convenient, and expectations are still rising: 40%
are looking for even easier shopping across on- and offline channels, 41% expect improved customer service, 45% want improved delivery service, and 46% ask for easier return and refund (E-commerce Europe, 2016).

The outcomes of the annual Greek e-Commerce survey, which ELTRUN (ELTRUN, 2016) conducts annually reinforce the above ones. The survey was conducted in the beginning of 2016 and almost 1000 questionnaires were collected. The survey insights revealed the problems that customers deal with at the delivery process and highlighted the main customers’ requirements concerning the delivery process. Indicatively, only 80% declared that they received the product in the right condition. Moreover, only 26% of the consumers recognize that the total procedure of delivery is simple and only 23% declares that e-shops can react properly in urgent situations. Order traceability is recognized as one of the major requirements of the delivery process, since a significant part of the consumers cannot track their order. Therefore, online retailers and logistics companies should meet the e-commerce customers’ requirements of more qualitative, traceable delivery processes.

Responding to the main emerging trend triggered by e-commerce growth, the transportation industry is undergoing a major transformation nowadays. The e-commerce trend has created a dynamic and turbulent environment, where the density of the distribution network consisted by many delivery points, multiple delivery channels and last-mile delivery requirements significantly increases the distribution cost. Therefore, the reduction of delivery costs can become a significant competitive advantage. In this context, complexity has also raised by the variety of distribution requests, e.g. orders and returns from various consumers at different points. Therefore, e-commerce logistics turned to be a key area of innovation and one of the European Commission’s key policy areas (European Union Report, 2016). Retail formats that deliver on this complexity will therefore be more successful in the future (E-commerce Europe, 2016).

E-commerce for physical goods generates a significant demand for dedicated delivery services that results in increasing fragmentation of shipments in the “last mile” (ARCEP, 2012; AT Kearney, 2012; Morganti, et al., 2014). In particular, home delivery services, which are usually preferred by the online consumers, contribute to the atomization of parcel flows thus causing particular problems within the urban areas. However, alternative delivery solutions such as pick-up points and lockers are growing fast, especially in urban areas. Stores are becoming fulfillment centers, serving as pick-up locations for online orders and customers require more dynamic delivery services for the last mile rendering this process a challenging task.

This pilot is also motivated and justified by evidence collected in the context of the U-TURN EU-funded project regarding the benefits, the implementation challenges and various aspects that
need to be taken into account for the successful implementation of online grocery collaboration practices in logistics. More specifically, the feasibility and the impact of alternative urban last mile shared logistics scenarios for grocery products in the area of London have been explored. Sharing trucks, routes, picking up points etc. might be the case for reducing last mile delivery costs of e-logistics and increase customers’ satisfaction levels. As consumer turns to be an important stakeholder in the e-commerce area, a consumers’ survey in the beginning of 2016 focusing on e-Commerce deliveries in Greece was conducted in order to get more insights about the alternative delivery methods consumers prefer. Home delivery is by far the most preferable delivery method for consumers according to the survey results (80%), but, at the same time, participants show a positive attitude in using potential alternative delivery methods, such as pick up points and that the solution of pick up points would be more preferable if it was offered by more e-shops (ELTRUN, 2016).

While we witness a significant growth of e-Commerce market and the logistics is considered to be a basic part of e-commerce (Bask et al., 2012), this is undoubtedly an area where big data can be extremely fertile (Swaminathan, 2012). Big data collection and analysis can make activities in e-commerce logistics much more efficient (Akter et al., 2016). Personalized services, dynamic pricing, predictive analytics, supply chain optimization and visibility are some of the core big data application areas in these fields (Akter et al., 2016; Koutsabasis et al., 2008; Mehra, 2013). In the e-commerce context, big data can be used to improve decision making in all activities involving infrastructure and operation on one hand, and consumers’ behavior on the other, thus achieving a better matching between supply and demand.

Building up on the aforementioned evidence, this pilot will investigate the role of big data in transforming the current e-commerce logistics practices for bringing value to all the involved stakeholders i.e. 3PLs, online retailers and final consumers. More specifically, we will investigate how the big data exploitation can enable collaborative practices that will enhance efficiency of the distribution process in e-commerce. We will also try to provide alternative shipping methods to the consumer in order to respond to this growing e-commerce consumers' demand. By taking into account the immature collaboration culture and the under exploitation of big data in the e-commerce logistics industry, we will also investigate how advanced business analytics techniques that exploit the various types of data can transform and enhance the existing e-commerce logistics processes.

Due to the continuously changing e-commerce environment and the limited adoption of big data solutions at the e-commerce logistics, we initially utilize a user requirements elicitation approach that involves the pilot user, as well as representatives of the broader industry in our pilot. The aim is to capture the broader sector’s requirements and, then, to further elaborate on them in the context of our 3PL user. Thus, we formulate a broader set of requirements and respective scenarios that will be tested in order to examine their applicability and quantify their
potential impact in the first stage of the pilot execution. By employing the iterative approach adopted in the overall project, we will refine the scenarios at each stage by taking into account their pilot results. Our ambition is to provide generic guidelines about the role of big data in e-commerce logistics and provide a roadmap for applying big data solutions with the purpose to tackle specific requirements in the e-commerce logistics. In comparison with the other pilot, the e-commerce pilot will not have a replication pilot. Due to the more complex horizontal data integration necessary for the e-Commerce pilot, we refrained from pushing for a replication pilot due to the time constraints of the project. On the contrary, we will give more emphasis on examining the phenomenon on a broader horizontal scope than focusing on vertical examination and covering specific user needs.

The remainder of this document is structured according to the structure described in the D2.1 Methodology Handbook. More detailed explanation of the overall methodology and, also, the structure of the design deliverables can be found in this deliverable.

2 Design of Pilot

2.1 Requirements

2.1.1 User Requirements Capturing and Elicitation Approach

This section describes the general approach we have used in this document for capturing the pilot requirements and identifying the needs of the various big data solutions in the e-commerce logistics context.

Initially, a meeting was held at 21/12/2016 with an industrial working group, comprised by the major Greek e-commerce retailers and 3PLs representatives. The industrial working group focuses on e-commerce logistics and is organized with the support of the Greek e-Commerce association. During the meeting, we presented the overall TT pilot objectives and, then, we discussed the current problems, challenges in the e-commerce area and tried to explore any potentials for collaboration and application of big data. The discussion revealed the importance of enhancing last mile logistics processes by incorporating big data solutions. Since the immature collaboration culture was also one of the key findings, we decided to arrange a series of in-depth interviews with the e-commerce logistics stakeholders in order to explore not only collaborative scenarios, but also the more traditional logistics aspects.

More specifically, to delve into the needs of e-commerce logistics sector and, further, elaborate on the user requirements, our research methodological approach encompassed three phases (Figure 3), which we detail in the following paragraphs.

- A literature review regarding e-commerce logistics and big data in the logistics era.
- Individual in-depth interviews with selected industry experts.
- On-site observation and discussions with the pilot 3PL user.

In the first phase, a comprehensive literature review of both journal and practitioners articles dealing with ‘big data’- "e-commerce"- "logistics" and "supply chain" related topics was conducted.

In the second phase, a series of in-depth interviews were conducted in order to enrich our insights about industry needs. The interviews were conducted from 20/2/2017 to 10/3/2017 with four leading companies of their sector. We selected the Interviewees based on two criteria: the nature of their business (pure online vs click and mortal companies) and the cooperation or not with a 3PL company. We developed an “interview protocol” (Appendix) to collect information on the following areas:

- Demographics characteristics of the companies,
- Overview of the current situation in e-commerce logistics,
- The attitude towards collaborative logistics practices in e-commerce area,
- The impact of big data technologies in e-commerce logistics and
- Future scenarios in e-commerce logistics.
The interviewees consist of three retailers in the sector of electronics (one of them pure online and two of them click and mortar) with presence in all country, as well as a courier company which is the leading courier company in Greece.

![Companies Profile and Logistics Characteristics](image)

The interviewees presented the profile of their company together with some facts concerning their current logistics practices. Further, they were asked to express their opinion about the applicability of collaborative logistics in their firm and their industry and make related comments, even if they had not had a relevant experience. Subsequently, an interviewer provided a description of different collaborative logistics distribution models that can be implemented using big data techniques and the interviewees offered their perspectives, thoughts and their interest about them. Through this method the project team was able to identify a number of problematic areas in the current practices, accompanied by some drivers and barriers towards the adoption of collaborative logistics practices, and many challenges that must be addressed or taken into consideration for both the design and the implementation of big data solutions. In the third phase an analysis of an in-depth case study of an Greek 3PL user which is currently interested in using ‘big data’ for improving its logistics processes and formulating innovative services for the end -consumer is realized.

In the following subsections, we present the insights obtained by the aforementioned process and we conclude with the main requirements of this pilot.
2.1.2 Literature review insights

Many studies show that customers consider the logistics performance as an important factor of e-commerce (Espé et al. 2003; Agatz et al. 2008). Distribution of goods to customers and in particular, last mile operation is probably the most demanding process in e-commerce. This is because in e-commerce customer distribution deals with frequent personalized orders resulting in high costs and difficult to manage processes. Outbound delivery costs are high because customer orders do not fill a truck resulting in low utilization of delivery trucks and personnel and longer routes. In addition, re-distribution, or returns system, is considered to be one of the most problematic and costly activities in e-commerce logistics. An efficiently organized return process, as part of the distribution system, can help to retain the customer by minimizing his inconvenience and reducing pick up and return handling costs.

Logistics is considered to be a basic part of e-commerce (Bask et al., 2012) and, according to literature findings, it is undoubtedly an area where big data can be extremely fertile (Swaminathan, 2012). Big data collection and analysis can make activities in e-commerce logistics much more efficient (Akter et al., 2016). Personalized services, dynamic pricing, predictive analytics, supply chain optimization and visibility are some of the core big data application areas in these fields (Akter et al., 2016; Koutsabasis et al., 2008; Mehra, 2013). In the e-commerce context, big data can be used to improve decision making in all activities involving infrastructure and operation on one hand, and consumers’ behavior on the other, thus archiving a better matching between supply and demand.

Exploitation of big data in logistics requires matching data sources, analytic procedures and business understanding. Three main areas of logistic applications, where big data analytics can provide effective decision support, are identified in the literature (Figure 5): operational efficiency and network planning, customer experience and new business models (Jeseke et al.,
Operational efficiency: Using big data analytics, there is a huge potential for improving operational efficiency. Operational efficiency includes both last mile optimization, as well as network planning. Last mile efficiency can be considerably improved by real time optimization and consolidated pick-up and delivery. A second area of big data applications in logistics regarding operational efficiency field is network planning, which includes decisions concerning warehouses, distribution centers and custom-built vehicles.

Business Models: Using the huge amount of data, logistics providers can develop new business models and provide additional services (Jeseke et al., 2013). New business models include financial demand and supply chain analytics, address verification and correct geocoding. All these applications call for big data analysis methodological approach and can help enterprises gain additional revenue sources.

Customer experience: Big data analytics help to maximize customer satisfaction and understand customer demand (Choi, 2016). Use cases include both customer value management and supply chain risk management (Jeseke et al., 2013). Merging data from multiple sources (sales information, data derived from social media, discussion forums and e-commerce catalogues) valuable information for effective customer relation management can be obtained that results in assessing customer satisfaction, improving service quality and increasing customer loyalty. Moreover, the use of big data analytics can help towards capturing consumer behavior and restricting supply chain risk management.
Specific use cases that exploit the value of big data analytics regarding distribution process and contribute to operational efficiency were identified and presented in this section. These use cases fall in two categories; last mile optimization and network planning.

Operational Efficiency - Last mile optimization cases

Real-time (dynamic) route optimization: Last mile is the most expensive distribution leg (Punakivi, et al., 2001) and it can be considerably improved if we use real time information from different sources (traffic data / sensors / real time events) to dynamically optimize routing and provide drivers with directions on the spot (Pillac et al., 2013; Fabian & Christian, 2012). In e-commerce deliveries, dynamic routing is even more challenging due to the fact that orders change unpredictably and dynamically. (Du et al., 2005). Dynamic Vehicle Routing allows for taking into account both the delivery and the new return requests. As indicated earlier, re-distribution or Returns system is considered to be one of the most problematic and costly activities in e-commerce logistics and thus, dynamically rerouting by combining deliveries and new return requests in real time could improve operational efficiency.

Forecast of failed or problematic deliveries: Problematic deliveries are also recognized as one of the major problems in e-commerce delivery since they contribute to increased costs, emissions, and wasted time and effort (Gevaers et al., 2009). According to Duin et al. (2016), delivery efficiency can be improved by identifying problematic zip code areas and applying changes in last mile delivery regarding time, location and route.

Crowd based pick-up and delivery: An additional application that optimizes last mile delivery is crowd –based pick up and delivery that employs different classes of commuters and taxi drivers to undertake paid deliveries (Jeseke et al., 2013). These operations are based on real time data streams involving event processing and geo-correlation big data techniques (Zeiler, 2013).

Consolidated pick-up and delivery: Based on recent professional reports and academic publications, (Tyan et al., 2003; Danloup et al., 2014) delivery efficiency can be considerably improved by combining routes of multiple supply chain stakeholders (e.g. e-commerce retail players, 3PLs). Consolidating routes from different companies can help in making more efficient the routing process and also increase the utilization capacity of vehicles. This requires exploiting and consolidating information from multiple data sources. This use case appears to be a novel feature in the practice of collaborative logistics in the context of e-commerce deliveries.

Operational Efficiency - Network and Operational planning Cases

Network planning: Big data techniques support network planning by analyzing historic data, seasonal trends, emerging freight flows as well as more general information concerning regional growth. Accurate long- term demand forecasts based on detailed data are generated in order to improve strategic decisions and support investments concerning network planning.
such as location of distribution points and warehouses. According to Wang et al. (2016) demand fluctuations or uncertainties constitute the most challenging problem that network design has to face.

Dynamic inventory routing: Inventory deliveries which are based on sales data in a Vendor Managed Inventory (VMI) framework can benefit from real time dynamic routing. In this case, routing is organized on a short-term basis. Special purpose models can provide considerable help in improving the routing schedule, increasing the precision of the future demand estimations and decreasing the requirement of extra (non-prescheduled) vehicle routes and their respective costs (Wang et al., 2016; Sage, 2013).

Location of click and collect points or pick up points: Click and collect service is an alternative delivery method that can offer flexibility and cost reduction and thus increase customers' satisfaction. Xu et al. (2014), argue that pick up points appear to be a novel way for limiting the costs of last mile e-commerce delivery. According to McKinsey global institute report (2013) the future of online grocery in Europe is closely related to successful operation of the click and collect service. This model gives retailers easier entry into the online-grocery space, since it has much less daunting economics than home-delivery service. In the case of e-commerce, adopting a hybrid system of retail stores and click and collect points by adding click and collect points to an existing distribution network on one hand offers an additional service to the consumers, but on the other entails the risk of creating points of unutilized capacity. For these reasons making the right mix of physical retail stores and click and collect points is a strategic decision concerning the distribution network that should be carefully evaluated (Gulati and Garino, 2000). Location of click and collect points needs to take into account both transportation flows and strategic planning of the distribution centers. There is no doubt that the development of convenient pickup points is crucial for the successful implementation of such a service and has attracted attention in e-commerce delivery processes (Xu et al., 2014). According to Tanskanen et al. (2002) customer density appears to be a crucial factor regarding the correct location of these points.

In summary, Big Data Analytics (BDA) play an essential role in improving the efficiency and effectiveness of E-commerce logistics (Yu et al., 2016). Although there are some existing industrial business use cases, implementation of big data techniques appears to be an untapped asset across industries (Addo-Tenkorang and Helo, 2016). Based on the existing literature, one of the most problematic areas of supply chain logistics in general and e-commerce logistics in particular, is distribution of goods because it is mainly performed on a personalized basis, resulting in high costs and difficult to manage processes. Outbound delivery costs are high because customer orders do not fill a truck resulting in low utilization of delivery trucks and personnel and longer routes. According to Yu et al. (2016), future logistics models
should focus on providing high service quality in last mile distribution and improving operational efficiency for companies.

Literature review findings indicate that some effective ways for improving efficiency in e-commerce distribution are: providing alternative methods of delivery and augmenting the distribution network with click and collect points, forecasting problematic deliveries, implementing efficient returns management and routing, combining deliveries and returns in real time. Moreover, there is room for action in the field of shared logistics in e-commerce (consolidating distribution by arranging cooperation of different logistics providers). Although shared logistics have been used in some cases, this type of collaboration has not yet been implemented in e-commerce logistics and needs further investigation.

2.1.3 In-depth interview insights

In order to gain a better understanding of the e-commerce sector and more specifically of the specificities of the logistics in this era, a series of interviews were conducted with stakeholders and potential users of big data solutions as described above. This subsection includes the evidence obtained towards the in-depth interviews.

During the interviews, the consortium was able to become aware of the main characteristics and the problematic areas of e-commerce and logistics market in Greece. Almost from the very first discussions it was evident that the competition in the market of e-commerce is harsh, which constitutes perhaps one of the most significant barrier for the adoption and use of collaborative logistics practices. The geographic diversity of the country and its small population density, the problematic zip code system, the national orientation of most of the retailers companies, some special characteristics of the most 3PL companies and couriers and a number of legislative restrictions and the effects of the on-going fiscal and economic crisis are the main problematic areas that have been identified. The aforementioned aspects result in a high distribution cost that is even higher at the last mile part of the delivery process (as last mile is defined the distance from the warehouse to the consumer).

Aligned with the literature review insights, the majority of the companies has also recognized the last mile delivery as the most expensive part of the delivery process. The last mile delivery cost in Greek market is also increased by the phenomenon of “cash on delivery”. “Cash on delivery” is a payment method which is preferred by almost 75% of online consumers and gives the consumer the right not to accept the delivery. Therefore, except for the cost of "cash on delivery" option, the delivery cost is also increased by the number of undelivered orders that needs to be returned. The last mile delivery is also recognized as the most challenging one as it directly affects the consumers' perceptions about a brand. For example, companies declare that most of customers' complaints are related to the wide time window delivery that forces then to wait at home until they receive the product. As the time window in home deliveries
cannot yet be decreased, retailers investigate the possibility of alternative delivery options. Current delivery options include home deliveries, deliveries at physical stores or couriers offices and deliveries at pick up points e.g. gas stations. All the interviewees expressed their interest for providing more advanced delivery options at their customers and they mentioned that they have already explored various options e.g. port baggage delivery options. However, the limited traceability of the orders and the restrictive process of deliveries that do not allow consumers to alter the point of delivery seem to be responsible for the lack of dynamically changed delivery options and processes. Interviewees stressed the fact that the new big data technologies that can collect data from various partners, from various sources (e.g. social media data, sales data, locations data and geolocation data) can provide solutions that can decrease the delivery cost and also improve the efficiency of the distribution process.

Another important aspect in the procedure of e-commerce logistics is the “reverse logistics” as the European law gives to customers the opportunity to return the product within 14 days from the order date. The total return percentage varies per industry. It is not surprising that specialty stores and food have a low return rate. Consumers already know what they are looking for in a specialty store, and fresh products such as food are hard to return most of the time. This is also the case for our interviewees' companies that represent electronics industry. However, in fashion the return rate is much higher than in other industries. The main reason for this is that clothes have different sizes, different fabrics, etcetera. For example, the total return percentage in fashion industry is 16,50%. However, some companies believe it is not a bad thing to have a high return rate and that it is related to the overall company strategy. However, their collection adds complexity in the distribution processes and increases the distribution cost. A major problem that many 3PLs companies also face is related to the reverse logistics and the empty runs. Even if the load factor of a vehicle is high and very satisfying, the same vehicle returns without any goods at the end of the day, and this is translated only to costs (regardless of whether it is paid by the companies themselves or not). Therefore, collaboration in the returns area was recognized as a potential scenario of applying big data solutions.

Another aspect that needs to be taken into account in the Greek market is the oligopoly structure of couriers business where one of the key players holds 40% of the market share and the other five couriers share the rest. Couriers play a significant role in the e-commerce logistics chain as they are responsible for the last mile distribution and the respective cost. Moreover, they can provide traceability data of their orders that can support a series of advanced analytics in e-commerce logistics and can release a series of dynamic delivery services that can be used by consumers. Indicatively, the interviewees proposed the following services:

- Use of POS by couriers in order to replace the cash on delivery method.
 Sending SMS, which gives to customers the opportunity to change the place of delivery or the time of delivery.

 Sending SMS, which gives to customers the opportunity to change at the same time the place and the time of delivery.

 A unified system which endorses the procedure of reverse logistics.

The above services can be also supported by a service that exploits logistics data and provide a service where a customer can record its new requests (change of delivery point, return requests) and then modify the routing schedule. Moreover, as alternative solutions for reducing the cost came out the following: (1) cooperation between different warehouses, especially in non-urban areas, (2) cooperation between couriers (a solution which takes into consideration the legal framework of every country), (3) cooperation between e-shops in order to share the same pick-up points/lockers. However, the market has still an immature collaboration culture and the high competition make some companies resistant to collaborate even if they recognize a potential value. Therefore, they suggested to us the creation of a working group for e-commerce logistics that involves couriers in order to engage market representatives in discussions that affects their industry.

2.1.4 3PL user insights

In this specific section, we focus on the case of the 3PL company that is our data provider and we further elaborate on the requirements in order to turn them into pilot objectives.

Logika is a 4th party logistics services company, surpassing the firm’s context and aiming to offer innovative solutions covering the whole range of a company’s supply chain activities. The consultant team identifies company’s needs to support products and customers, throughout the supply chain and implement a solution for business, to create an Asset light, Dynamic company with Flow of its supply chain. Various meetings were held with Logika personnel in order to further elaborate the user requirements and also to get a better understanding about Logika internal processes and available data.
Logika’s key activities include: Receipt & Storage of products, Picking, Stock Replenishment, Packing, Orders execution, Storage of goods, Raw and packing materials and Lot traceability. It covers the activities that take place in specific part of e-commerce logistics processes as depicted in the following figure. In order to fulfill the previous activities, Logika uses its own WMS system. In this system, the company collects and analyzes data via its customized algorithm, and at the same time it designs the proper services based on its customers’ needs. The WMS (Figure 6) function may be described via the analysis of: (a) the notion of data that are collected and (b) data flow.
The WMS system consists of data which are sent by two different sources. On the one hand, it receives data by the e-shop and on the other hand it receives – and on parallel sends- data to couriers companies. To be more specific, Figure 7 describes the first part of data flow; via the ftp server, the e-shop sends the orders to Logika; this file is saved to WMS and the system starts all the internal procedures of checking the availability of the product. The second part of data flow comes from the courier procedure. As it is described on Figure 8, Logika combines courier’s data with internal data in order to identify the optimal solution.
The data that are kept represent the following entities:

- Customer
- Customer Branch
- Order
- Order Item
- Product
- Receipt
- Receipt Item
- Return
- Return Item
- Stock
- Supplier

Until now the previous data are stored in the database and are used only in order to transact the orders. The basic objective of this pilot is the analysis of the previous data and the design of algorithms which may combine the tables and extract the points which are problematic. These
points will be the first point in order to develop the new services and ameliorate logistics procedure.

2.1.5 Key outcomes and pilot requirements

In the previous section, we presented the general requirements as they came up from an extensive literature review and interaction with the industry. We elaborated on them and we conclude here with some more specific requirements:

- **Req. 1:** To explore how collaboration could be applied in the e-commerce logistics domain in order to address the current challenges and support firms to decrease cost and improve the overall distribution process performance.
- **Req. 2:** To identify potential synergies among the e-commerce stakeholders at the reverse logistics.
- **Req. 3:** To analyze current distribution processes in order to depict the current distribution patterns and forecast future problems.
- **Req. 4:** To provide alternative shipping methods at the consumer in order to increase customers' satisfaction and provide lower prices.
- **Req. 5:** To enable dynamically changed supply chains that take into consideration various routing and customer preference characteristics in order to decrease logistics costs and to increase customers' satisfaction.

2.2 Objectives

In the previous section, we presented the requirements we elicited from an extensive literature review and interaction with the industry. Here, we mainly focus on the 3PL company, which is our data provider, and further elaborate on the aforementioned requirements in order to convert them into specific pilot objectives. Our aim is to use the context of the 3PL company in order to implement the aforementioned requirements and provide a proof-of-concept for them. There is one to one relationship between requirements and objectives. For example, **Objective 1** realizes the first requirement. Next, we discuss how the aforementioned requirements will be implemented.

- **Objective 1:** To quantify the impact of shared logistics scenarios among 3PLs in e-commerce by taking into account data from various supply chain stakeholders (e.g. e-commerce retail players, 3PLs) in order to decrease both cost and environmental burden.

Despite the fact that this scenario imposes the need of data from additional 3PL companies, it can be supported for our purposes by Logika data alone. We will achieve this by splitting the transactions in more than one segments using random sampling and using additional
virtual 3PL companies with different locations. The impact of shared logistics scenarios will be outlined by using the predictive analytics approach and descriptive analytics. For example, prescriptive analysis using Kajal Routing can be applied in order to optimize and plan different routes. Moreover, Moriarty Analytics will be available to gather information from different data sources and will provide algorithms to work in near real time.

- **Objective 2:** To expand the shared logistics concept at the reverse logistics supply chain of our 3PL.

Logika maintains a record of all returned orders from various retailers and we will test this scenario sufficiently, in terms of routing optimization and we will quantify any synergies exist. To gain an additional impact the predictive analytics in objective one will be extended based on the requirements of the reversed logistics. Moreover, Kajal Routing can be applied in order to optimize and plan different routes.

- **Objective 3:** To identify patterns of problematic deliveries based on previous data and forecast problematic situations.

Problematic deliveries fall into different categories and Logika maintains data regarding failed deliveries. In order to perform a good forecast, a record is required that contains both good deliveries and problematic deliveries; such a record exists in Logika data. In this scenario, the first step is to determine the exact reasons for failed deliveries and “bundle” them into categories. The next step is an analytical approach to identify failed deliveries both in terms of a specific recipient and in terms of geographic areas that tend to yield a lot of failures. As this approach requires the analysis of previous data to gain a better base for decision-making, the descriptive analytics approach is used. Moreover, Predictive analysis using Kajal Forecasting in order to know what will happen and how to proceed. It allows developing and configuring new and already existing forecasting methods in order to identify potential failed deliveries in advance.

- **Objective 4:** To explore alternative shipping methods and, more specifically, the click and collect option at the Attika urban area.

This approach requires the analysis of previous data to gain a better base for decision making. The Logika historical data about deliveries can be used in order to determine the best locations for these micro-hubs and having the deliveries demand. In this context, the descriptive analytics approach is used and the DISMOD tool will provide several techniques to analyse the logistics network structure of the partners and to optimize the locations of micro hubs.

- **Objective 5:** To dynamically identify alternative delivery options by taking into consideration various routing and customer preference characteristics, such as maximum
travel time diversion, multi-destination, neighbouring customer or friend, return requests in order to combine various deliveries and decrease delivery and returns collection cost.

Logika distribution data can be used in order to support this scenario. The impact of a combined delivery and returns will be analysed by using the predictive analytics approach. The previous data has to be analysed, then aggregated with forecasting the future deliveries and returns. Here, the predictions can be used to offer “optimized” routing. Later, prescriptive analytics might be a further development in objective four. Also, an algorithm that will take the aforementioned preferences into account will be deployed.

Overall, the five objectives will be approached by the following analytics techniques described by the BDVA Reference Model (see Figure 18). The described objectives focus on Predictive and Descriptive Analytics first. In future developments, the prescriptive analytics might be required to enable user advice based on additional functionality for an optimized shopping experience for the client.

### 2.3 Use Cases and Scenarios

In this specific section, we describe the application scenarios that will address the previous section pilot objectives. There is one to one relationship between objectives and use cases. For example, Use Case 1 realizes the Pilot Objective 1. Below, we describe in more detail the use cases and the scenarios that will be implemented in order to cover the requirements and the objectives mentioned:

- **Use case 1: Shared routings / trucks among 3PLs from 3PLs picking locations to the customer delivery**: This use case supports the concept of collaboration among 3PL companies. The aim of this scenario is to minimize delivery cost (e.g. in terms of time, fuel consumption etc.) and CO$_2$ emissions while maximizing loading factor by incorporating the distribution flows from 3PLs picking locations to the customer delivery points and sharing a common vehicle in order to serve the common or adjacent delivery points.

- **Use case 2: Shared routings / trucks among online retailers for the returns**: This use case supports the concept of collaboration among online retailers. The aim of this scenario is to reduce the cost of collecting the returns and CO$_2$ emissions by incorporating the distribution flows from customer delivery points and sharing a common vehicle in order to collect them by the common or adjacent delivery points. This scenario will also incorporate the notion of dynamic re-routing in order to provide alternatives where the product is collected when a courier is nearby in order to deliver products.

- **Use case 3: Forecasting the problematic deliveries coming from e-commerce retailers at the Logika case**: In this use case, we will try to identify the delivery patterns of specific e-commerce retailers in order to enhance inventory management and optimize the delivery
process in general. Problematic deliveries fall into different categories: faulty products returned, recipient not found, refusal to pay for the product at the time of delivery etc. Sometimes a delivery is delayed due to not having enough stock. After identifying the categories for failed deliveries, we will identify failed deliveries both in terms of customer characteristics and in terms of geographic areas and we will establish a process of forecasting them in advance.

- **Use case 4: Shared Micro-Hubs (in terms of space) where customers collect their online orders:** An alternative way of distribution models in the e-commerce market involves the micro-hubs or click and collect points. In this use case, consumers shop online but, then, travel to the collection point to collect their purchases. Today, the retailers use their own stores (collection points) when they offer Click & Collect services; so there is still a need to have a trip but the consumer saves the picking time. In the pilot, retailers will use micro-hubs points in central locations of the town (e.g. in bus or train stations or fuel stations) where the customers could collect their orders. The idea about the Micro-hubs points is that these points could be small warehouses with lockers containing mainly small packets or specialized areas in some of the existing shopping centers. The customer’s orders could be delivered in off-peak hours by the retailers/couriers to these points from which the customers could collect their orders.

- **Use case 5: Dynamic identification of alternative delivery options by taking into account various alternatives of the delivery points and new delivery requests.** This use case scenario will explore how the delivery options and routing will be dynamically changed by taking into account the customer preferences or return requests. For example, just before the check-out, an e-commerce online customer is provided with an additional delivery option on top of the standard ones: one built dynamically in real time, after requesting all other collaborating e-commerce players whether they have to make a delivery at the same delivery point or area or by taking into account customer preferences (e.g. delivery at a neighbouring customer or friend, delivery at in multi-destination options)

The following figure summarizes the pilot requirements, objectives and use cases that came out.
2.4 Data Assets

This section presents the data assets that are available in this Pilot.

<table>
<thead>
<tr>
<th>Name of Data Asset</th>
<th>Short Description</th>
<th>Initial Availability Date</th>
<th>Data Type</th>
<th>Link to Data ID Card (in basecamp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logika WMS Data</td>
<td>The company data of Logika regarding warehouse and transportation derived from Mantis lvision WMS (e.g. ID order, ID receipt, locations, customers, etc.)</td>
<td>01/04/2017</td>
<td>CSV, XML, TXT, ORACLE 11g</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/432596697">https://3.basecamp.com/3320520/buckets/1429164/uploads/432596697</a></td>
</tr>
<tr>
<td>Company Ecommerce Purchases</td>
<td>Enterprises having purchased online (at least 1%) - % of</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/upl">https://3.basecamp.com/3320520/buckets/1429164/upl</a></td>
</tr>
<tr>
<td>Name of Data Asset</td>
<td>Short Description</td>
<td>Initial Availability Date</td>
<td>Data Type</td>
<td>Link to Data ID Card (in basecamp)</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>-------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Company Ecommerce Sales</strong></td>
<td>Enterprises having received orders online (at least 1%) - % of enterprises with at least 10 persons employed</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/432596676">https://3.basecamp.com/3320520/buckets/1429164/uploads/432596676</a></td>
</tr>
<tr>
<td><strong>Individual Ecommerce Purchases</strong></td>
<td>Individuals using the internet for ordering goods or services - % of individuals aged 16 to 74</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/432596642">https://3.basecamp.com/3320520/buckets/1429164/uploads/432596642</a></td>
</tr>
<tr>
<td><strong>Individual Ecommerce Purchases within three months</strong></td>
<td>Individuals having ordered/bought goods or services for private use over the internet in the last three months - % of individuals</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/432596653">https://3.basecamp.com/3320520/buckets/1429164/uploads/432596653</a></td>
</tr>
<tr>
<td><strong>Individual Ecommerce Purchases from EU countries</strong></td>
<td>Individuals using the internet for ordering goods or services from other EU countries - % of individuals aged 16 to 74</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/432596665">https://3.basecamp.com/3320520/buckets/1429164/uploads/432596665</a></td>
</tr>
<tr>
<td><strong>Disposable income of private households</strong></td>
<td>Disposable income of private households by NUTS 2 regions - PPS (based on final consumption) per inhabitant</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/432596622">https://3.basecamp.com/3320520/buckets/1429164/uploads/432596622</a></td>
</tr>
<tr>
<td><strong>Primary income of private households</strong></td>
<td>Primary income of private households by NUTS 2 regions -</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/432596642">https://3.basecamp.com/3320520/buckets/1429164/uploads/432596642</a></td>
</tr>
</tbody>
</table>
### 2.5 Big Data Technology, Techniques and Algorithms

The present chapter explains different technologies and algorithms, which have been used to support the pilot scenarios over the different layers of the architecture.

#### 2.5.1 IDS Connectors

The Industrial Data Space is an initiative, developed by Fraunhofer, designed to create a secure data space that enables companies of various industries and sizes to manage their data assets confidently. The main goal of the Industrial Data Space is to facilitate the exchange between Data Providers and Data Users.

In Transforming Transport, this concept is used to enable a secure exchange over the defined architecture. The main component in this approach is the base on the Core IDS Connector. The connector architecture is designed to exchange data between two or more industrial companies. The data will only be exchanged between the certified and linked partners. The data owner - the company - determines who is allowed to use the data. As a result, partners in a value-added chain can jointly access certain data to create new business models, develop their own processes more efficiently, or otherwise initiate additional value-added processes.

The architecture components are the based on **IDS connector**, which have to be implemented on each partners’ IT-infrastructure to fit the requirements on the Companies’ environment, as well as the requirements of the IDS Reference Architecture. Through the Connector, the data applications (like databases and/or it-systems) can be securely connected to the Big Data Infrastructure. Any kind of transformation into another structure or standard like XML or JSON can be implemented as within a container application. Basic technologies of the IDS connector

---

<table>
<thead>
<tr>
<th>Name of Data Asset</th>
<th>Short Description</th>
<th>Initial Availability Date</th>
<th>Data Type</th>
<th>Link to Data ID Card (in basecamp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPS per inhabitant</td>
<td>Population on 1 January by NUTS 2 region - Persons</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td>oads/432596770</td>
</tr>
<tr>
<td>Level of household internet access</td>
<td>Level of internet access - households - %</td>
<td>01/04/2017</td>
<td>TSV file</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/432596748">https://3.basecamp.com/3320520/buckets/1429164/uploads/432596748</a></td>
</tr>
</tbody>
</table>
architecture are Docker (application container), Spring boost (micro services), Apache ActiveMQ as the Message Bus of the connector. Default-language for implementation of any data transformation or further data processing is Java. Different connector implementations may use diverse kinds of communication and encryption technologies depending on the requirements.

The Industrial Data Space Connectors realizes the transport of data through different communication components. A solution should consider data volume, technology and security levels. The Connectors enable a higher security level by combining multiple security mechanisms.

In general, an IDS Connector consists of the following containers:

- **Message Bus**: The Message Bus stores data between services or connectors. Usually the Message Bus provides the simplest method to exchange data between connectors. The IDS Reference Architecture recommends the implementation named Apache ActiveMQ. If required, the use of a different implementation is possible to meet the requirements of the operator.

- **Message Router**: The Message Router is the execution engine inside a Connector. It executes multiple workflows and invokes Data Services as included workflow steps. It enables the communication with the Message Bus. The exchange of the Message Router component by alternative implementations of different vendors is possible to support the demands of different participants. If a Connector in a limited or embedded platform only consists of a single Data Service or a fixed workflow (e.g. on a sensor device), the Message Router may be replace by a hardcoded workflow or the Data Service is exposed directly.

- **Data Service(s)**: A Data Service provides the executable activities for a workflow. The Message Router executes this workflow. Furthermore, a Data Service defines a public API, which is invoked from a Message Router. The duties of Data Services vary in a broad range. A different programming language is possible to meet the requirements of the partners IT-Infrastructure. The use of standard technologies like REST, Web-services or custom protocols with extended security features is possible.
Figure 10 shows the described Connector communication.

The full Infrastructure layout describes the connector in a two parts-component: the company-side connector and the framework-side connector between which the data is exchanged. Before the data transfer, several security and homogenisation services are implemented within the connectors.

2.5.2 Moriarty Analytics and Data Storage

Moriarty is an advanced Artificial Intelligence and near real time analytics software solutions framework for big data, developed by ITAINNOVA. It allows understanding and structuring information, identifying hidden patterns and correlations in the data, and inducing knowledge, as well as building learning systems. Its capacity to be flexibly adapted to develop new functionalities through the definition of workflows, which address the complexity of scientific and business applications, new Semantic Models and the rapid development of cloud services gives it a differentiating value. With Moriarty, it is possible to obtain highly useful information through the capture, storage, process and analysis of a massive and variety amount of data.

In addition, the way in which data scientists (that work with algorithms and data transformations using a visual interface) and software engineers (that work with the idea of
services to be invoked) easily collaborate in order to reduce the “time to market” in the development of complex big data and artificial intelligence applications is another innovative and distinctive feature of Moriarty framework.

In general, Moriarty maximizes the use of libraries, generates rapidly decoupled applications, allows integrating in a simple way new functionalities by combining basic ones and/or existing complex processes, adding new algorithms easily and also making changes in the generated processed and, overall, invoking external services in an easy way for both data scientist and software engineers. Figure 11 shows the structure of the Moriarty Analytics platform.

Moriarty Analytics platform is also fully integrated and tested with the provided Big Data Infrastructure for WP10 pilot, offering pre-existent services for the capture, storage, process and analysis of data on the aforementioned infrastructure.
ITAINNOVA will use Moriarty Analytics for fast and easy development of big data services that includes the different technologies and platforms available within WP10 pilot.

**Collecting and storing data:** Moriarty has implemented interfaces to extract information from different data sources, being able to work with structured and non-structured databases. These interfaces will be used along the project to collect information which has been delivered through IDS connectors. The following storage technologies, fully supported by Moriarty platform, will be used along the project to support WP10 pilots:

- **Apache HBase:** An Open Source Distributed NoSQL database built on HDFS (Hadoop Distributed File System), scalable to support big amounts of data with a high performance. One single table can have thousands of columns. HBase will work as an intermediate service layer to spread out data loading in different files and servers. Data will be sent to Hbase through IDS connectors and will be processed and stored in Solr and MongoDB.

- **Apache Solr:** An Open Source Search Engine written in Java. Main functionalities are whole text searches, facet searches, real-time indexing and dynamic clustering. Solr is a powerful text search engine.

- **MongoDB:** An Open Source NoSQL Database system. Instead of saving data in tables as relational DBs do, it saves data structures as documents with dynamic schemes, enables easy and fast integration of data in different applications.

- **Apache Hadoop Yarn:** Open Source Software for managing computing and storage cluster. It enables distributed execution of Big Data Analytics optimizing resource consumption.

**Data Analytics and processing in near real time:** In order to process information and execute algorithms in near real time Moriarty Analytics takes benefit of the data processing tools offered by the underlying infrastructure. These tools enable the shared use of memory and processors to allow parallel execution of algorithms and data processing. Thus, Moriarty Analytics offers services to manage parallelization and near real time processing.

For example, algorithms can be built using Moriarty Platform to load, clean, process information from HDFS to Solr and MongoDB.
The following tools, together with Moriarty Analytics, will be used in order to process information used in WP10 pilots:

- **Apache Spark**: It is an open-source framework which provides an API for the development of Java & Scala applications for data analytics that are executed in a distributed way for optimizing computing and storage resource in a cluster. Algorithms implemented in Moriarty or Kajal will run over Apache Spark to take advantage of available resources.

- **Apache Hive**: Open source tool which enables the analysis of big amount of data stored in Hadoop HDFS by using a SQL-based language. Initially developed by Facebook, Apache Hive is now used and evolved by many companies.

- **Python**: Open source general-purpose language that is frequently used for data engineering tasks. It allows quick pre-processing of any type of data either natively or through an API that implements Apache Spark that can be used in case distributed computing is required.

- **Cloudera Impala**: Open source tool similar to Hive, Impala is an Open Source SQL engine for massive parallel processing of data stored in Hadoop HDFS. Aimed for data analysts and data scientists.

**Machine Learning and Artificial Intelligence**: Moriarty Analytics offers Machine Learning components (work items and pre-existent workflows) that can be used to create advance Artificial Intelligence services. New algorithms could be developed along the project in order to adapt the system to the pilot requirements. Figure 13 shows an example of Moriarty machine learning.
2.5.3 Kajal Routing

Kajal is an integral framework owned by ITAINNOVA oriented to supply chain management and data analysis that includes different software modules such as demand forecasting, multilevel inventory optimization, production planning, resource scheduling and dynamic routes calculation. More specifically, ITAINNOVA will provide KajalRouting software and will develop new algorithms and functionalities for dynamic route calculation within use case 1 and use case 2 to optimize direct and reverse LSP flows to and from customers, and ITAINNOVA will also provide KajalForecasting software and will develop new and configure existing forecasting methods for use case 3 in order to identify potential failed deliveries in advance. Kajal routing is described in this section in more detail and KajalForecasting in the following section.

KajalRouting will create optimal routes for both, direct and reverse flows, to optimize the distribution process of different LSPs, allowing a shared logistics scenario. The general data used by KajalRouting include:

- Locations of depots: sources of transports (direct flow) and destinations of transports (reverse flow)
- Location of customers: delivery points (direct flow) and pick up points (reverse flow)
- Shipment Data: weight, volume.
- Available fleet of transport: number, size, type of vehicles.
- Shipments restrictions in terms of “time windows”
- Function objective to optimize: number of vehicles, distance (number of kms), time.

Additionally open data can be added to increase the value of the optimization results.

In order to do this, KajalRouting will implement solutions for two principal problems: the vehicle routing problem (VRP) and the vehicle routing problem with time windows (VRPTW).

The VRP is a combinatorial optimization and integer programming problem which tries to provide the optimal set of routes for a fleet of vehicles to traverse in order to deliver to a given set of customers. The context is that of delivering goods located at a central depot to customers who have placed orders for such goods. The objective of the VRP is to minimize the total route cost. Determining the optimal solution is an NP-hard problem in combinatorial optimization, so the size of problems that can be solved optimally is limited. KajalRouting solver therefore will use heuristics and genetic algorithms due to the size & frequency of real world VRPs they need to solve.

The VRPTW is the same problem that VRP with the additional restriction that in VRPTW a time window is associated with each customer, defining an interval wherein the customer has to be supplied. The interval at the depot is called the scheduling horizon.

The VRPTW is, regarding to VRP, characterized by the following additional restrictions:
A solution becomes infeasible if a customer is supplied after the upper bound of its time window.

A vehicle arriving before the lower limit of the time window causes additional waiting time on the route.

Each route must start and end within the time window associated with the depot.

In the case of soft time windows, a later service does not affect the feasibility of the solution, but is penalized by adding a value to the objective function.

Some variations of these two principal problems will be taken into account:

- **Multiple Depot VRP:** A company may have several depots from which it can serve its customers. If the customers are clustered around depots, then the distribution problem should be modelled as a set of independent VRPs. However, if the customers and the depots are intermingled then a Multi-Depot Vehicle Routing Problem should be solved.

- **VRP with Pick-Up and Delivering:** The Vehicle Routing Problem with Pick-up and Delivering (VRPPD) is a VRP in which the possibility that customers return some commodities is contemplated. So in VRPPD it is needed to take into account that the goods that customers return to the delivery vehicle must fit into it. This restriction makes the planning problem more difficult and can lead to bad utilization of the vehicles capacities, increased travel distances or a need for more vehicles.

In order to implement these routing problems KajalRouting will make use of different computer science and mathematical optimization techniques: heuristics, metaheuristics (variable neighborhood search) and genetic algorithms.

- **Heuristics:** heuristic is a technique designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution. This is achieved by trading optimality, completeness, accuracy, or precision for speed.

- **Metaheuristics:** is a higher-level procedure or heuristic designed to find, generate, or select a heuristic (partial search algorithm) that may provide a sufficiently good solution to an optimization problem, especially with incomplete or imperfect information or limited computation capacity. Metaheuristics sample a set of solutions which is too large to be completely sampled. Metaheuristics may make few assumptions about the optimization problem being solved, and so they may be usable for a variety of problems. Variable neighbourhood search (VNS) is a metaheuristic method for solving a set of combinatorial optimization and global optimization problems. It explores distant neighbourhoods of the current incumbent solution, and moves from there to a new one if and only if an
improvement was made. The local search method is applied repeatedly to get from solutions in the neighbourhood to local optima.

- Genetic Algorithms (GAs): are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomised, GAs are by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution, especially those follow the principles first laid down by Charles Darwin of "survival of the fittest". Since in nature, competition among individuals for scanty resources results in the fittest individuals dominating over the weaker ones. Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection.

2.5.4 Kajal Forecasting

KajalForecasting will implement and configure forecasting methods in order to identify potential failed deliveries in advance. Thus, the delivery process will be improved from an holistic view since customers will perceive a better quality service reducing the current annoying process of waiting for long time at home or going to the transport agency to collect the parcel and transport companies will become more efficient and competitive in terms of costs due to the reduction of failed deliveries.

The general process followed by KajalForecasting is based in the following steps:

- Analyse the purpose of the forecast: it determines the accuracy and power required of the techniques, and hence governs selection. Deciding whether to enter a business may require only a rather gross estimate of the size of the market, whereas a forecast made for budgeting purposes should be quite accurate. The appropriate techniques differ accordingly.

- Identify and analysis of explanatory variables: the objective of this task is to evaluate the available information that may have correlation with the problem, in our case the failed deliveries, that is likely to be included in the multivariate forecast models.

- Definition of statistical methods to be used: once the explanatory variables are identified, the objective of this task is to select the forecasting methods and models that have a greater precision for each type delivery.
Forecasting system configuration: In this task, the forecasting system will be configured with the models selected in the previous tasks, and the rest of the data necessary to obtain forecast (forecasting periods, aggregation level, etc.) will be parameterized.

Deployment: The deployment of the forecasting system will be deployed in the different in a staggered way, based in: urban areas, transport companies or users profiles.

In order to implement this forecasting challenge, KajalForecasting will make use of different forecast methods: statistical, econometric and soft-computing methods. The selection of a method will depend on many factors—the context of the forecast, the relevance and availability of historical data, the degree of accuracy desirable, the time period to be forecast, the cost/ benefit (or value) of the forecast to the company or the time available for making the analysis.

Statistical methods: time series methods use historical data as the basis of estimating future outcomes. Methods for statistical methods may be divided into two classes: frequency-domain methods and time-domain methods. The former include spectral analysis and wavelet analysis; the latter include auto-correlation and cross-correlation analysis. In the time domain, correlation and analysis can be made in a filter-like manner using scaled correlation, thereby mitigating the need to operate in the frequency domain. Additionally KajalForecasting divides time series analysis techniques into parametric and non-parametric methods. The parametric approaches assume that the underlying stationary stochastic process has a certain structure which can be described using a small number of parameters (for example, using an autoregressive or moving average model). In these approaches, the task is to estimate the parameters of the model that describes the stochastic process. By contrast, non-parametric approaches explicitly estimate the covariance or the spectrum of the process without assuming that the process has any particular structure. Both, univariate and multivariate methods are considered within the system.

Econometric methods: some forecasting methods try to identify the underlying factors that might influence the variable that is being forecast. For example, including information about climate patterns might improve the ability of a model to predict potential failed or successful deliveries. Forecasting models often take account of regular seasonal variations. In addition to climate, such variations can also be due to holidays or other related events.

Several informal methods used in causal forecasting do not employ strict algorithms, but instead use the judgment of the forecaster. Some forecasts take account of past relationships between variables: if one variable has, for example, been approximately linearly related to another for a long period of time, it may be appropriate to extrapolate such a relationship into the future, without necessarily understanding the reasons for the relationship.
- Soft-computing methods: The Soft Computing techniques are based on the information processing in biological systems. To overcome the restriction of the linear models and to account for certain nonlinear patterns observed in real problems, some nonlinear models have been proposed. However, since these nonlinear models were developed for specific nonlinear patterns, they are not capable of modeling other types of nonlinearity in time series. In recent years, to overcome these issues, efficient soft computing techniques such as artificial neural networks, fuzzy time series and some hybrid models have been used to forecast any kind of real life time series.

In addition, conventional models require some assumptions such as linearity and normal distribution that cannot be utilized efficiently for some real time series since this kind of series contain some uncertainty in itself. However, when soft computing methods are used to forecast time series, there is no need to satisfy any assumption and the time series uncertainty can be forecasted efficiently. Unlike statistical methods, soft computing methods are more tolerant to imprecision, uncertainty, partial truth, and approximation in time series.

ITAINNOVA will configure KajalForecasting for a dynamic forecasting method selection behavior. A benchmark system will compare the results of all implemented techniques and will select, in an unattended way, the best forecasting technique every time the forecasting process is launch. So far, different key factors have been identified that could influence the accuracy of forecasting models and methods and they make it necessary this smart benchmark module, such as customer profile, time of the day of the delivery, weather, specific events.

2.5.5 DISMOD
Fraunhofer IML will provide several techniques to analyse the logistics network structure of the partners and to optimize the locations of micro hubs within Use case 4. For that case DISMOD®, an analysis and planning software for an established and transparent decision-making support in logistics will be used. DISMOD® offers a broad selection of possibilities, e.g. regarding the extrapolation of shipping volumes, the determination of optimal warehousing sites, route planning etc.

DISMOD® Locate will find optimal logistics locations based on distances, cost information and shipment volumes. The transports will be recorded precisely at SKU-level and will be optimised based on real freight rates. General savings in transport costs due to optimised network structures will become visible. Since DISMOD® Locate will consider existing structures and multistage logistics networks, an optimized network approach regarding the location and structure of micro hubs will be found.

In detail, the optimization of the network and locations is done in three steps:
First, the database is prepared and filled with data from the partners or open data. The quality of the data is assured through consistency checks. The general data used by DISMOD® contains:

- Locations and LocationTypes
- Shipment Data
- Shipment and Transport types
- Article Masterdata
- Tariffs
- Warehousing costs
- Sourcing information
- Forecasting information

For each partner, all data that can be provided will be used to optimize the logistics network but it is necessary to include location and shipment data for all partners.

Additionally open data can be added to increase the value of the optimization results. For the initial pilot design, Fraunhofer IML will include E-Commerce Sales & Purchases data from Eurostat and match the data to regional boundaries (e.g. zip codes). The additional data will be used to adapt the micro-hub locations depending on the historical regional purchasing power. Any suitable data that is found within the project lifetime will also be used for optimization.

After establishing a DISMOD® database, an initial look at the as-is scenario is done. With DISMOD®, Fraunhofer will create a baseline scenario and analyse it regarding Performance KPIs and network structure. These values will later on be used for comparison to the shared logistics scenario with micro hubs.

For scenario analysis, there are several topics that can be addressed:

- Quantity structures
- Amount analysis
- Service level analysis (cf. Figure 15)
- Cost analysis
- Shipment structure analysis
- Demand analysis (with the help of open data)
When the as-is scenario is evaluated, Fraunhofer will start to optimize the logistics networks of the partner with two approaches: Single Optimization and Shared Logistics Optimization.

In the shared logistics use-case, the location of micro-hubs that combine goods of different logistics partners is optimized given the single networks. The optimization algorithms can be adjusted to scenario and use-case and will combine a Center-Of-Gravity approach with hub costs, sales and purchases data and possible market predictions.

During the optimization progress, Fraunhofer will create KPIs that show the transfer possibilities to other transport sectors and also keep an eye on the possibilities to support the idea of a physical internet with shared logistics hubs.

2.5.6 Inventory Routing Algorithm
The Inventory Routing Problem (IRP) is an Operational Research problem that has risen relatively recently in the context of Vendor Managed Inventory (VMI). In the classical formulation of the problem, a supplier delivers goods to the customers that are geographically dispersed. The supplier has the full responsibility for not causing stock-out to any of the customers and handles their inventory and stock levels. This means that the supplier is
responsible for making the replenishment decisions according to the demands and the imposed constraints of all customers. VMI can be easily identified as a win-win situation for both the vendors and the customers:

- Vendors benefits: the vendors can more carefully schedule the production plan and manage their stock by minimizing inventory costs. Also, they are responsible for the routing which allows minimizing the goods transportation costs as well.

- Customers benefits: The outsourcing of inventory management allows the customer to not allocate effort and resources on inventory control, thus reducing the inventory costs.

The decisions that should be taken in a classic version of IRP are the following:

1. The time to serve each customer
2. The quantity that is delivered to each customer in each visit
3. The routing planning for serving all customers.

Each customer can determine the minimum and maximum stock capacity and the rate at which the product is being sold. The vendor can also determine the production rate. The problem is solved with respect to both mentioned sets of constraints to ensure that the customers do not run out of stock while optimizing transportation and inventory costs. Thus, a satisfactory solution to IRP implies lower operation cost, improved efficiency and increased customer satisfaction.

The IRP is a realistic mathematical formulation of the B2C model that is being studied. The e-commerce logistics procedure can be parallelized with the IRP as described above. In the use case that is being studied the vendors are the e-stores and the customers are the individuals who place orders to the e-shops. The problem is to decide when the customer will be served and which routing solution will be adopted (e-stores own fleet or a fleet of another e-commerce player). A solution to IRP is expected to explore and evaluate the possible benefits of the shared logistics scenario where the e-stores collaborate in both specific cases of interest:

Case A - Collaborating e-stores use “shared” stock

In this case, a synergetic vehicle routing algorithm can be developed, in order to minimize the total transportation costs. The transportation model in this case will require that customers of all collaborating e-stores are jointly served, and at the same time vehicles should visit the various e-stores depots to pick-up the various purchased items. The algorithm can quantify the benefits of the collaborators (transportation costs reduction, customer satisfaction increase, etc.) in this scenario and in the no-collaboration scenario.

Case B – Each store has its own unique product stock
This case investigates the installation of commonly operated intermediate depots to store products later to be pushed to the customers. These intermediate depots will be fed with products from the various collaborating e-stores and then vehicles will be loaded with products from these intermediate facilities and proceed to the various customers. This two-echelon distribution process should be able to save many kilometers travelled in urban areas and thus bring great economic and environmental savings, because the necessity of visiting every e-store depot along the generated routes will be eliminated. Given the intermediate depots and the demand forecast, the depot stock replenishments times and quantities can be optimized to reduce the transportation costs. The IRP can be solved with the intermediate depots as vendors to further reduce the transportation costs to the actual customers.

In order to develop an effective algorithm, the following data are required: Distances between all customers, Customer selling rate, Customer minimum and maximum stock, Vendor production rate, Vendor fleet (number, capacity), Intermediate depots locations and distances to all customers. Minimum and maximum capacities (for Case B), Demand predictions for Intermediate depots (for Case B)

The applicability of IRP in real life and its ability to reduce transportation and inventory costs for both partners in the supply chain implies the complexity of the problem. The problem is classified as NP-hard, which means that cannot be solved to optimality in an acceptable period of time (polynomial time relatively to the scale of the problem). Well known and commonly used mathematical optimization approaches fail to deal with the complexity of the problem. Especially, with the contribution of big data the scale of the IRP problem becomes unsolvable with common techniques. Thus, an advanced algorithmic approach will be developed to deal with the problem. The main objective is to provide sufficient cost reducing solutions in short computational time. This will enable us to deal with the real-time orders that should be served, constantly changing the routing decisions and requiring to resolve the problem.

The solution originates from the joint of computer science and mathematical optimization fields. The algorithm which will be developed will be a metaheuristic approach to the problem. Metaheuristics are high-level heuristic algorithms able to find sufficiently good solutions to optimization problems. Their special structure enables them to solve an impressively wide range of problems with real-life complicated objectives and constraints. More specifically, different intelligent greedy constructive algorithms will be designed and tested to achieve satisfactory performance. Even more advanced local search algorithms can be utilized to further improve the solution. Local search algorithms are intelligent, often nature-inspired, procedures to effectively explore “promising” solution of the optimization space. This allows adopt a “clever” exploration trajectory, only examining good solution and not wasting time with exhaustive search. All previous are considered state-of-the-art techniques for complex real life
problems and can produce satisfactory solutions in occasions where common techniques fail. Thus, we expect this approach to effectively assist the examination of the two cases.

2.5.7 ITML Optimisation Solution

ITML will provide a two-dimensional solution for the needs of the aforementioned pilot use cases. The first is directly related to Use cases (1) and (2); it focuses on the optimisation (and visualisation of results) of shared routings and trucks among 3PLs both from 3PLs picking locations to the customer and from the customer back to 3PLs (for returned products). The optimisation algorithms and visualisation tools will focus mainly on (i) costs, in terms of fuel and time and (ii) COx emissions, in order to maximise the pilots’ impact to the environment.

The envisioned module will comprise features for:

- Fleet management
- Route Management
- Data collection (using sensors and/or based on manual input)

The second dimension focuses on a data aggregation and sentiment analysis tool based on information available in social media, to be exploited within Use case 4 towards efficient business exploitation of the proposed shared Micro-Hubs solution: insights will be captured and analysed regarding the end-users’ / customers’ opinions (an example is illustrated in Figure 16) about the services provided by the Micro-Hubs, and provide specific business operation solutions / strategies.
This specific tool can also be utilised in Use Case (3); it can enhance the forecasting capabilities through the exploitation of social media: Analyse social media data, provide insights regarding the users’ perceptions in specific geographical areas (also including events and/or special cases likely to create delays and problems) and enhance accuracy of forecasting problematic deliveries.

2.5.8 INTRASOFT Data Services
INTRASOFT provides several services, which are used to organize and model the data stored in the big data infrastructure.

Annotation/Curation Service: This service will deal with organizing and annotating data utilized for the pilot. The service will be developed on top of the Hadoop Cluster, using appropriate tools depending on the available Hadoop distribution. For each dataset (either raw format ingested or processed format), appropriate metadata will be kept in a formalized manner, including dataset description, dataset utilization, version control etc.
Data engineering: This service will be developed using either Python or Apache Spark, or a combination of both. These tools have been selected having flexibility in mind as well as providing scalability in a distributed manner to handle big data challenges.

- Python is an open source interpreted language with a design philosophy that emphasizes code readability and a simple syntax that allows programmers to express concepts in relatively few lines of code. Python is a general purpose language which means that every type of software can be created using it and therefore is considered very flexible. In recent years, Python has become one of the most popular languages for data engineering tasks such as data visualization, cleansing, wrangling and homogenization. Extensive libraries have been developed that allow handling of structured, semi-structured and unstructured data using a single tool.

- Apache Spark is an open-source cluster-computing framework. Originally developed at the University of California, Berkeley's AMPLab, the Spark codebase was later donated to the Apache Software Foundation, which has maintained it since. Spark provides an interface for programming entire clusters with implicit data parallelism and fault-tolerance. Apache Spark provides programmers with an application programming interface centered on a data structure called the resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way.

- Spark SQL is a component on top of Spark Core that introduced a data abstraction called DataFrames, [a] which provides support for structured and semi-structured data. Spark SQL provides a domain-specific language (DSL) to manipulate DataFrames in Scala, Java, or Python. It also provides SQL language support, with command-line interfaces and ODBC/JDBC server.

Data engineering service will provide descriptive analytics on each and every data set (data visualization). It will also handle the task of converting each data element to the appropriate equivalent form defined by the data model (data homogenization). Furthermore, data will be cleansed and wrangled to avoid having null values wherever possible. Finally, this service will transform all data in the appropriate forms required by the higher level analytical services.

Data modelling service: For the implementation of this service, an appropriate data model will be created based on the available data and the use cases. The data model will be defined from scratch and grow as the pilot evolves and more data sources are added. For each element in the data model, specific mappings will be created in order to be used by different data sources.

The data model will be implemented in one of the open source databases included in the infrastructure, and data will be loaded to be utilized by the analytics algorithms.
2.6 Positioning of Pilot Solutions in BDVA Reference Model

TT is part of the Big Data Value PPP (e.g., see http://www.bdva.eu/) and as such will benefit and contribute to collaboration among related PPP projects. To provide a common framework for understanding potential points of common interests among PPP projects, the so called Big Data Value Reference Model will be employed as part of the technical collaboration activities.

To this end, each TT pilot is asked to position its solutions in the Big Data Value Reference Model. Ideally, each of the solutions developed is given a unique number in the deliverable, which is referred to in the description of the solution (i.e., Section 2.5 or 3.5 in D*.1 respectively). This number should then also be used for positioning the solution graphically in the below visualization of the Big Data Value Reference Model\(^1\). The BDVA Reference Model covers the most important big data technical areas (shown horizontally). The BDVA Reference Model also covers key cross-cutting concerns, such as data protection, cyber security, development and operations, and standards (shown vertically).

Figure 17 depicts the Big Data Value Reference Model and illustrates an abstract pilot contributing to 2D visualisation of predictive analytics solutions based on streaming analytics, where data linking is done using existing solutions.

\(^1\) Available as editable version (PPT) to TT members from https://3.basecamp.com/3320520/buckets/1429164/vaults/338367019
Figure 17. Big Data Value Reference Model

1 = Industrial Data Space Connector
2 = Moriarty Analytics ® and Data Storage
3 = Kajal Routing©
4 = Kajal Forecasting©
5 = DISMOD ®
6 = Inventory Routing Algorithm
7 = ITML Optimisation Solution (Social media analytics and emissions and fuel management/monitoring analytics)

8 = INTRASOFT Data Services

2.7 Big Data Infrastructure

The big data infrastructure used within this pilot consists of four different layers (as shown in Figure 18). On the side of the logistics partners, the company data that is used within the use-case is stored in several databases or platforms depending on data and company (data sources layer). On top of the data sources and origins the Industrial Data Space Connector by Fraunhofer IML is implemented to ensure safe data exchange with our central big data infrastructure or different companies.
Figure 18. Big Data Infrastructure
The full Infrastructure layout describes the connector in a two parts-component: the company-side connector and the framework-side connector between which the data is exchanged. Before the data transfer, several security and homogenisation services are implemented within the connectors.

The framework-side connector is set on top of the real big data storage systems (run by ITAINNOVA). ITAINNOVA provides a big data storage and computation infrastructure for the pilot, as well as different tools and applications for Big Data analytics tasks.

This infrastructure is basically a set of network connected servers with regular storage and computation resources, arranged in cluster mode for sharing resources and enhancing performance.

The cluster is configured as an Open Source Apache Hadoop Ecosystem. Apache Hadoop is an ecosystem of open source components that fundamentally changes the way enterprises store, process, and analyze data. Unlike traditional systems, Hadoop enables multiple types of analytic workloads to run on the same data, at the same time, at massive scale on industry-standard hardware.

In this project Cloudera's open source platform (CDH distribution) is used for deploying the Apache Hadoop Ecosystem. CDH is the most popular distribution of Hadoop and related projects in the world.

For providing hardware, two different alternatives are being considered:

- using ITAINNOVA servers
- using Cloud Infrastructure provided by local partners (BIFI) or global as Amazon or Azure

Choosing one or another will depend on the resource needs of the pilot scenarios. First stage will be covered using a set of ITAINNOVA servers, but as long as the scenarios increase their complexity more storage and computing resources will be added.

Nevertheless, one of the reason for being deploying an Open Source Apache Hadoop Ecosystem for the pilot is its scalability: new elements (nodes or servers) can be added effortlessly to the cluster in order to keep up with the increasing resource needs of the pilot.

The Apache Hadoop Ecosystem also provides an integrated set of open-source tools and applications for running Big Data Analytics.

ITAINNOVA will provide to the consortium a pre-configured environment that can be customized to meet the requirements of the different scenarios.
The last layer of the infrastructure is the data management and analytics layer, where several frameworks and algorithms use the stored data for the single use-cases.

Moriarty and Kajal will be placed on the Data Management & Analysis layer in order to execute algorithms use in WP10 pilots. Data will be taken from different databases and predictive and prescriptive analysis will be done over there in order to calculate routes dynamically and forecast inventories.

The DISMOD® toolset by Fraunhofer IML will also be located within the analytics layer and transfer the needed data from the infrastructure to a local database used for optimization. DISMOD® is used as a desktop software system and the optimization results are transferred back into the storage systems to be a possible input for the other analytics software.

The Open Data Application has to provide an interface to various open data platforms. The Application should allow the user-friendly integration of data sources into the Big Data Infrastructure of the presented use case. Open Data is primarily defined as data from public sources, which are provided for free availability and usability by generally official services. Since the data are information from governments or government-related organizations, they are generally secure and of high quality. The inclusion of these data allows additional data analyzes and increases the reliability of the statements. With the Open Data Application, a number of data sources are preconfigured. The user should define the inclusion of additional sources. The Open Data Application will transfer the retrieved data to the Big Data Infrastructure via an IDS connector.

### 2.8 Roadmap

At this section we define what will be available as results of each of the three Stages 1-3 – please use the below table.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Delivery Date (Project Month)</th>
<th>Features / Objectives Addressed</th>
<th>Embedding in Operational Environment</th>
<th>Big Data Infrastructure Used</th>
<th>Scale of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: Technology Validation</td>
<td>M6</td>
<td>All use cases except the 5th one will be applied in a smaller scale</td>
<td>Historical data uploaded manually</td>
<td>The infrastructure will be implemented in 3 phases (See the figure below)</td>
<td>1 year historical data from Logika</td>
</tr>
<tr>
<td>S2: Large-scale experimentation and</td>
<td>M12</td>
<td>The use cases that came up that will have a biggest</td>
<td>Historical data uploaded automatically</td>
<td>The infrastructure will be implemented</td>
<td>1 year historical data from Logika,</td>
</tr>
<tr>
<td>Stage</td>
<td>Delivery Date (Project Month)</td>
<td>Features / Objectives Addressed</td>
<td>Embedding in Operational Environment</td>
<td>Big Data Infrastructure Used</td>
<td>Scale of Data</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>demonstration</td>
<td></td>
<td>potential impact will be expanded by using more data</td>
<td>in 3 phases (See the figure below)</td>
<td>Open data, Social Media data</td>
<td></td>
</tr>
<tr>
<td>S3: In-situ trials</td>
<td>M21</td>
<td>Based on the outcomes of the previous phases, three use cases will be applied in the operational environment</td>
<td>Real data in order to support daily processes</td>
<td>The infrastructure will be implemented in 3 phases see the figure below</td>
<td>Daily data from Logika, Social Media data</td>
</tr>
</tbody>
</table>

The Figure 19 below describes the three stages in which the big data infrastructure used at the use cases implementation. The Stage 1 will outline the use cases in a smaller scale with basic big data technologies running on the cluster and by integrating a small amount of data manually. The stage one is due by End of Month 6.

![Figure 19. Big Data Infrastructure Implementation Approach](image)
The stage two of the big data infrastructure required the first version implementation of the IDS Connector Reference Model. The data transport/delivery is taking place via the implemented connectors by end of the first year. The extended stage also includes the integration of a second data module, such as an additional partner or a defined application.

The fully implemented big data infrastructure consists of several data sources and a multi-functional Infrastructure to share, store and communicate data based on the defined architecture. The advanced approach consists of additional IDS Connectors for other companies, but also a number of internal or external resources, made available. The Architecture in Stage 3 implements all the required needs for the big data infrastructure.

3 Conclusions

The TT pilot main idea is to embrace key aspects of e-commerce logistics and the various stakeholders' views (customers, online retailers, couriers, 3PLs) and demonstrate how the deployment of Big Data technologies can transform them in order to provide better logistics services and enhance the logistics processes performance.

Following a coherent method of work, as analysed in Section 2.1.1, we have clarified in sufficient detail the current practices and the envisioned ones that rely on collaboration in e-commerce logistics and will, to a large extent, be supported by the big data technologies. Hence, the aim of this deliverable is to assimilate novel knowledge from the pilot regarding users’ views and requirements for e-commerce logistics in order to frame the pilot’s implementation and also structure the requirements of the big data infrastructure and the analytics techniques applied.

This deliverable presents in detail the Dynamic Supply Chain Networks Pilots Design process and provides the conceptual backbone of the overall pilot, enabling common understanding among project partners for the different use cases and supporting requirement acquisition and analysis. In the next couple of months, we envision to have a first version of the big data infrastructure (M6) and some first evidence about the impact and the applicability challenges of the selected use cases. In the meantime, we will build the Pilots Performance Assessment Plan that will be based on the first stage results and will also cover the impact aspects that we envision for our pilot. A strategy of pilot dissemination has started to be formulated in order to disseminate the pilot outcomes, validate pilot results, increase the expected impact and support the post project replication phase.
4 References


Danloup et al., 2014. Improving sustainability performance through collaborative food distribution International Conference on Green Supply Chain, France.


Duin, JHR; De Goffau, W; Wiegmans, B; Tavasszy, LA; Saes, M (2016). Improving home delivery efficiency by using principles of address intelligence for B2C deliveries. Transportation Research Procedia 12 (2016) 14 – 25


ELTRUN. (2016). Mapping the e-commerce logistics scene.


Glaxo SmithKline Annual Report (2016), Measuring Europe in eCommerce


McKinsey global institute report (2013). Perspectives on retail and consumer goods


5 Annex I: Interview Protocol

Introduction

The e-commerce retailing sector is experiencing rapid growth in the omni-channel offer with customers ordering online for home delivery or collection from the stores. Serving multiple sales channels - selling via web shops, own retail shops, wholesale customers, outlet stores, international stores and event-driven sales (including pop-ups) increases complexity in the supply chain. The needs of this rapidly growing market turned e-Commerce logistics as a key area of innovation and the provision of new delivery options to the global e-commerce customer stands out as a source of competitive advantage. The new reality in the online retail market worldwide demands new services and approaches in order to meet new customer requirements. The double-digit market growth and respective exponential increase of delivery points, make e-commerce logistics one of the most prominent and challenging areas, calling for collaborative logistics practices and innovative transportation methods for the companies to remain sustainable and competitive.

The general objective of the project is to demonstrate how the deployment of Big Data technologies can transform the e-commerce logistics in the following aspects:

(1) enable and support shared logistics that will improve the supply chain performance by merging and aggregating data from various sources,
(2) investigate how shared logistics could further be extended by using product availability information and incorporating an inventory routing approach,
(3) provide to the customer access to more delivery options and information concerning the route, cost and convenience of each delivery.

In this interview, the main aim is to depict the state of the art in the e-commerce logistics and to identify barriers and future challenges. The insights gathered towards interviews will enable us to formulate future e-commerce logistics use cases and to identify the logistics processes that Big Data Technologies can disrupt in order to provide new innovative services.

The interview length is approximately 60 minutes. A summary of the interviews will be sent to the respondents.
Interview Questions

Part A: Demographics

- What is your position within [case organization]? Which are your main responsibilities?
- How could you describe your company’s position in the overall supply chain?
- Could you please describe the main activities of your company?
- Could you please mention the countries of your company interest?
- Could you please describe the importance of e-commerce for your company?
- How many e-commerce orders do you deliver per day? (approximately)
- How many of these deliveries refer to Greece, and how many to other countries? (percentage)

Part B: Overview of the current situation in e-commerce logistics

- Could you please describe the current practices about e-commerce logistics?
  - e.g. a typical delivery process, the flow of information, the data sent by e-shops, the way these data are received
- Do you know how these processes are supported by information systems?
  - Could you please mention the main systems involved?
  - Could you describe the way that you design the delivery routes?
  - Could the existence of more data ameliorate the procedure of delivery?
- Which are these parts of the logistics process that augment the cost of e-commerce delivery?
  - Which are the main issues lead to the increased cost?
- What are the KPIs that measure the success of cooperation with an e-shop?
- What are the main problems that you have to face with?
- How familiar are the online consumers with the delivery process?
  - What are the main complaints that you receive by them?
  - How important is the “time window” variable for them?
- Could you describe us the reverse logistics process?
  - What percentage of your activity is relevant with the reverse logistics process?
  - What are the problems that you face within it?

Part C: The attitude towards collaborative logistics practices in e-Commerce area

- Do you know any collaborative logistics practices in e-commerce?
• Do you think that there are parts of the logistics procedure where e-shops could cooperate? Can you please name specific examples?
• Do you think that there are parts of the logistics procedure where couriers or 3PL companies could cooperate? Can you please name specific examples?
• Which are the potential obstacles of implementing collaborative logistics processes in e-commerce industry?
• The high delivery cost is recognized as a main problem in e-commerce. Do you think that collaborative logistics could provide a solution? If yes, please elaborate
• How important could the collaborative logistics process be for the procedure of reverse logistics?

**Part D: The impact of Big Data Technologies in e-commerce logistics**

• Are you familiar with the term “Big Data Technologies”?

The proposed Transforming Transport concept will be presented at this point.

• Do you think that the aforementioned shared logistics practices could be applied in your organization?
• Could you prioritize them in terms of applicability / efficiency?
• Which are the potential obstacles of their implementation in your [case organization] and the industry?
• Which are the potential benefits of their implementation?
• What type of data could enable you to improve your current distribution processes?
• Under which circumstances, would you be willing to share your data? (e.g. if the data are collected by a third company? If the data are collected by a “union”? etc)

**Part E: Future scenarios in e-commerce logistics**

• Could you please describe us the ideal scenario of a straightforward delivery process?
• Could you please describe us the ideal scenario for the reverse logistics process?

**Part F: What else?**

• Did we forget anything? Is there anything else you would like to discuss?
• Could we get back to you in case we have some further questions from our data analysis?