REPORT

PROJECT: Current and Future Use of (Big) Mobility Data

Bilateral agreement between Vlaams Instituut voor Mobiliteit (VIM) and iMinds-UGent

Current and Future Use of (Big) Mobility Data: To data warehouse or not to data warehouse?

Authors:

Samaneh Bahreini (IBCN)

Marlies Van der Wee (IBCN)

Table of Contents

1 Introduction and motivation

The digital evolution is accompanied by an exponential growth of available data from different sources at different levels. These data will be of great importance for the logistics and mobility sector in Flanders in the near future, as they can be used to analyze problems related to accessibility, safety and traffic congestion and improve the competitiveness of these sectors at European and international level.

The purpose of this project is to investigate the current and future availability of mobility data in Flanders, how this data can be inventoried - in a dynamic and practical way - and finally, what the value of these data could be for the different stakeholders and how it can contribute to future innovations. The research questions can therefore be summarized as follows:

- How can the large amount of available mobility data be handled in a dynamic and service-friendly manner?
- What is the value, or what could be the added value of this data to businesses, governments and research institutions?

This report provides a summary of project, and builds on the results gathered throughout the different tasks. For more detailed information on the followed methodologies and specific results, we refer to the wiki.

The approach taken in the project was an iterative one, where input from project partners, literature and interviews with experts was used to further detail the overview and improve the results. For the quantitative analysis, we relied on the ECMN tool for cost modelling and a bottom-up effects estimation for the calculation of the benefits.

The next section provides an overview about categorizing mobility data based on different parameters and investigate the current approaches for handling mobility data around the world. Moreover, in this section, different alternatives for handling data are described in terms of definition, advantages and disadvantages and cost estimation. Section 3 presents an overview of the most important political, economic, social and technical issues that are relevant to (big) mobility data (PEST framework). Both qualitative and quantitative assessment of costs and benefits for different handling option/business area combinations is the topic of section 4, while section 5 concludes the report and provides some concrete recommendations.

2 Categorizing, storing and analyzing mobility data

A first step in this project consisted of composing a sufficiently detailed overview of the available mobility data and potential categorization axes, which in a second step, served as an input to investigating the possible data handling methods. This section of this report will summarize the categorization and handling methods that were identified.

2.1 Categorizing mobility data

Data, and more specifically mobility data, can be categorized in different ways. Based on a literature review, different parameters were chosen: the type of data, source of data, level of data, collection method, etc. The results are summarized below.

- The **type** of data predicate on being *structured* (e.g. spreadsheets based on columns and rows), *unstructured* (e.g. text, images, emails) or *semi-structured* (e.g. Markup Language (XML), a textual language for exchanging data on the web).
- Data can be collected through internal or external **sources**. *Internally* collected data is collected from the available sources inside the organization or company and *external* sources of data collection indicate the use of data published by external agencies.
- Mobility data has from different **levels** as well. *Primary* data refers to dynamic data, directly related to 'moving' or 'moving objects' (e.g. data collected from probes, mobile phones, travel behavior studies, sensors in vehicles), *secondary* data is static data (e.g. road sign databases, speed maps, road information of the physical structure) and *tertiary* data does not have a direct relation to transport and mobility but might have an impact on the mobility and transport of goods and people (e.g. delivery times, costs, meteorological data, population growth, employment figures or crowdsourcing data (Twitter, Facebook, Foursquare) [1].

2.2 Current approaches for handling mobility data

In the mobility sector, different parties operate and maintain various facilities and systems that should work together in an integrated way. Each of them has different responsibilities and operational procedures. Consequently, each part of this complex system uses and stores different levels of data for different applications. This section provides a couple of examples of different data handling systems used by different transport and mobility authorities around the world, focusing on their needs and requirements.

In the Netherlands, the National Data Warehouse for Traffic Information (NDW) started to store traffic data about the most important roads in the Netherlands in the middle of 2009. By now, it has become an ideal and still growing source for conducting traffic and mobility analyses based on integrated data. The NDW is more than a valuable database containing both real-time and historical traffic data: "NDW is a unique alliance of 24 public authorities with the purpose of working together, and learning from each other" [2], and unify their data with other sources of traffic data. NDW is recognized as a "databank" for municipalities, information service providers, universities and research centers and the road users that want to get a better and more clear vision about the traffic situation along the roads. The road authorities, for example, use the data in order to monitor the traffic situation and redirect it by using ramp meters or other traffic management instruments. Service providers use media tools (TV, radio, etc.), websites, apps and navigation systems to inform and advise travelers about the traffic congestion before and during their trips. The data can also be used for scientific and research purposes like traffic policy and traffic simulation [2].

In Florida, the Central Data Warehouse (CDW) system is used for both real-time and archived traffic data [3]. In Singapore, the Land Transport Authority had the goal to improve transportation planning to be more agile by making fact-based and fast decisions and taking a long-term view; therefore, it needed a transport data warehouse that could analyze billion records of transport data in minutes rather than hours [4].

Belgium does not own a fully operational data warehouse (yet), but there are data portals on different levels (e.g. city-level: Antwerp, Ghent and Kortrijk, Flemish level: Informatie Vlaanderen, etc.). The city of Ghent for example allows everyone to use their collected data (e.g. underground parking occupancy, travel times during events, taxi locations, etc.). The city's open data platform monitors the number of apps that make use of the offered open data: there currently are about 20 apps for mobility data, ranging from finding the nearest parking spot to finding the nearest public toilet. As an example of open mobility data, Dynacity is a project implemented by VIM (Vlaams Instituut voor Mobiliteit) in Ghent as a test site to figure out from the end-user point of view what mobility data and information within a city must consist of to make urban dynamic mobility management useful and desirable [5]. Apart from the municipalities mentioned above, the federal government also provides more than 3200 datasets via their portal. Their transport data is categorized in different datasets, based on file type, publisher (Federal level, Flanders or Wallonia, city-level) and license. On a regional level, the Brussels Open Mobility Data Portal provides access to various data on mobility and public works in Brussels. These data are made available in different format and may be used under the conditions specified in the open license of the Brussels-Capital Region [6]-[7].

Following that different countries worldwide approach the structuring and storage of mobility data in different ways, this report also not focuses solely on the data warehouse option, but considers different alternatives for handling mobility data: ranging from a single point of contact to a full data warehouse.

2.3 Data handling systems: different alternatives

In this part of the report, the different systems to handle mobility data are described in terms of definition, advantages and disadvantages and implementation cost estimation.

2.3.1 Data warehouse: a full-scale solution

"A data warehouse, recognized as organization's "single source of truth" is a database designed to enable business intelligence activities: it exists to help users understand and enhance their organization's performance" [8]. The purpose of designing a data warehouse is to be able to query the data and use them for analysis and reports that might help to gain a better understanding of the business rather. As such, a data warehouse goes beyond the simple storage of data and information from different sources.

The large number of benefits of implementing a data warehouse for a business or an organization are based on its name: a data warehouse is a centralized warehouse to store data and information. Its users can use the output to analyze both real-time and historical data and detect patterns or links between data, which help them to make important business decisions. In general, data warehouses can save time by storing important information in the same location instead of keeping data in several different places. As a result, they can use this centralized data to optimize strategic decisions. Data warehouses help to save money as owners and executives can search the data without extensive assistance from the IT department. Companies can benefit from storing their data in the same format, which ensures data quality and uniformity. However, a data warehouse has a number of disadvantages that need to be mentioned as well. The disadvantages are explained below:

- The data must be must be cleaned, loaded, or extracted before it can be stored in a data warehouse, which can take a long time.
- Users who want to work with the data warehouse must be trained to use it.
- Accessibility via internet could lead to security problems.
- There is a significant investment cost related to setting up a data warehouse. Furthermore, the cost of a data warehouse is not just about creating it; there is also a maintenance cost that cannot be neglected.

As a data warehouse comprises the most functionalities, it will also entail the highest investment, both in terms of technical as well as human capital requirements. In this part of the report, the cost for deploying a data warehouse based on ECMN[1] is estimated, using the amount of required storage as a main driver. We furthermore assume that all historical data is also kept in the system.

It is necessary to mention that the model explained below is only a high-level estimation of the cost for a data warehouse because of a lack of specific inputs. To have a concrete estimation of the cost of handling data, the exact amount of data is needed, as well as cost input data from official vendor quotes. As determining the exact amount of data strongly depends on strategic decisions made by the different stakeholders, it is considered beyond the scope of this project. We hence make the assumption of two types of data growth curves (linear for conventional data and exponential for future mobility data), and an initial storage requirement of 2000TB of data in the first year.

Figure 1 represents the yearly amount of data to store and process in data warehouse. The amount of conventional data (e.g. inductive loops) follows a linear function because we assume no (significant) increase in the number of data sources (loops), such that the increase is only represented by the historical data. The exponential function on the other hand represents the growth in the amount of future mobility data. The increasing volume and detail of data captured by sensors and the rise of multimedia individuals with smartphones, social media, and the Internet of Things (IoT) will fuel exponential growth of data in the future. There is a broad agreement that data generation has been growing exponentially and this in every sector of the global economy [10].

[1] ECMN is a tool for calculating the cost of equipment that should be installed for a specific project [9]. The model draws from a hierarchical structure that allows determining the amount and cost of each equipment type to be installed. This hierarchical structure documents how equipment types are linked to each other and what the constraints on later calculations will be. By installing only the equipment that is needed at each point in time (based on the amounts of drivers), the costs of equipment are spread out and the investing firm receives a direct payoff that can be used to pay back the investment in equipment.

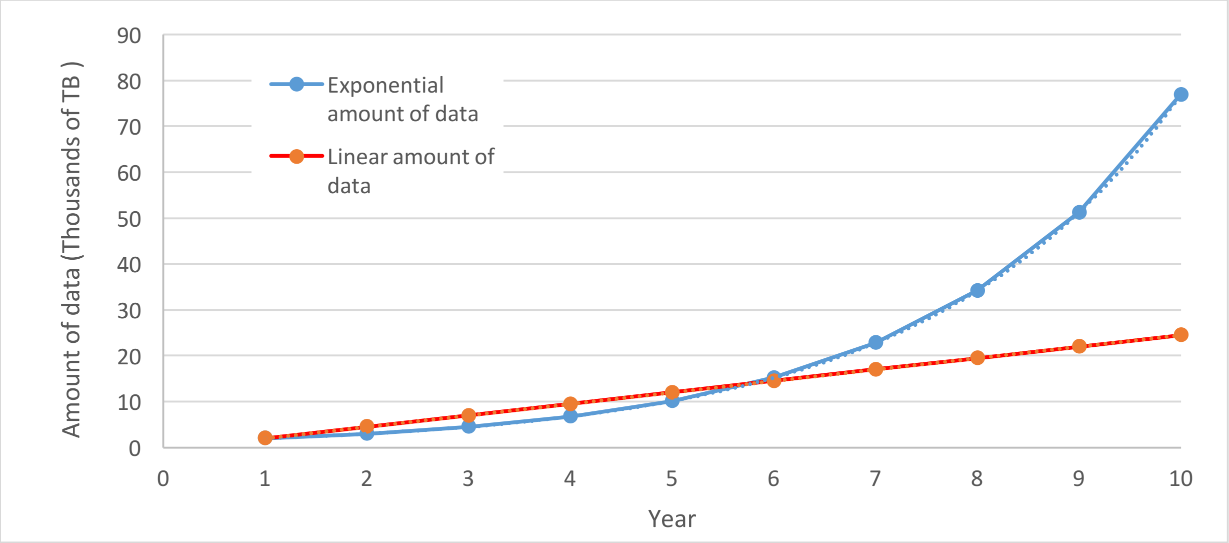

Figure 1:Yearly amount of data to store and process in data warehouse

Apart from the input on the amount of data, input of the types and costs of equipment is needed. These assumptions are summarized in Table 1. Please note that we consider two types of servers: storage servers and processing servers (the latter responsible for generating reports and being able to analyze the data). The system needs IT staff to operate the system and process the data. For each 100 servers, it is assumed that five FTEs (full-time equivalents) are needed.

Table 1: Equipment costs input

Equipment	**Price (Euro)**
Server, capacity = 2TB	3000
Power	0.25 per kWh
Cabling	5 per meter
Switch	619
Rack (contains up to 40 servers and 2 switches)	600
Operational system	3500
Software license	2000
IT staff salary	2000 per month, yearly increase of 5%

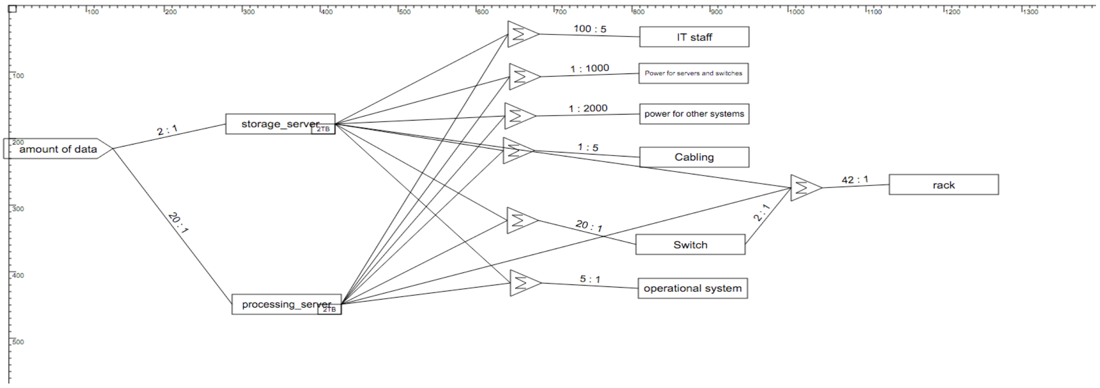

Figure 2: Cost model for data warehouse based on ECMN model

The cost of a data warehouse is hence estimated using the EMCN tool (Figure 2), and visualized for the two growth assumption curves in Figure 3. The total cost reaches about €40 million after 5 years for both growth assumptions. The total cost of storing data when it follows an exponential function is lower in the first years. The explanation for this fact is that, the amount of IoT or sensor data (e.g. data produced by autonomous vehicles and connected vehicles) is lower than the stored data collected by traffic loops. However, based on different views this will change in the next years as the amount of sensor data will increase exponentially [10].

It should be noted that, if the data warehouse needs a backup system for security purposes, then this requires doubling the entire installation, which will of course also double the total cost.

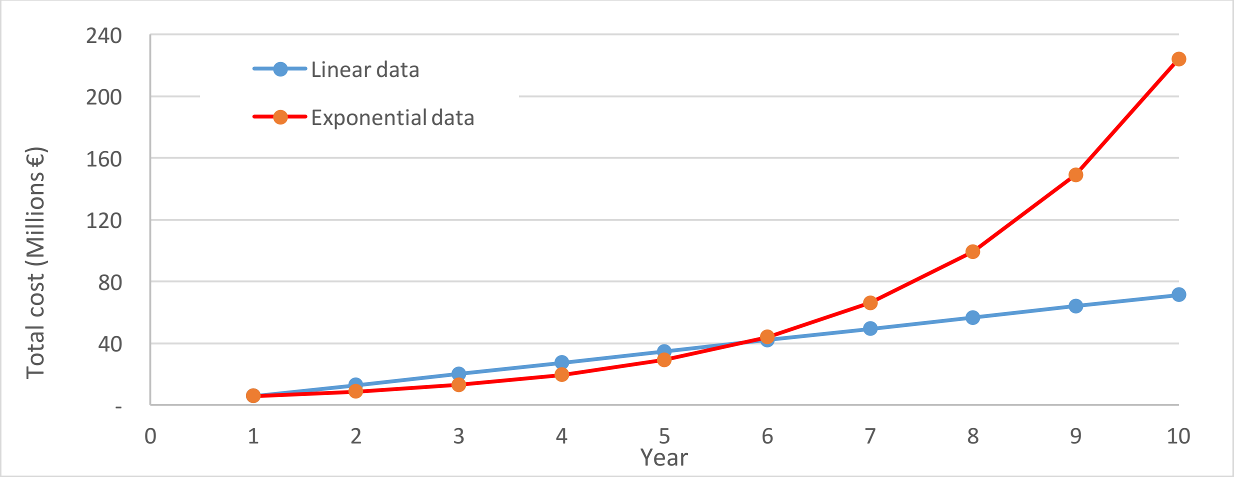

Figure 3: Total cost of implementing a data warehouse

2.3.2 Database: a structured collection of standardized data

A database is a collection of information that is organized in a way that a computer program can very quickly search and select desired information and data. Databases are designed to manage large amounts of data by storing and retrieving that information. Databases normally consist of rows and columns. Data is entered into a row, which build a "record". A database may contain millions of these records. Once the records are created in the database, they can be sorted in different ways.

Databases can be very important and helpful tools for managing large amounts of data, but they also have their own limitations. They can be programmed to help to organize the data, maintain it and even allow users to find exactly what they need without searching in information they do not care about. Databases can also provide security for the data because of their ability to store and keep data in one place. Another important point about databases is that they can be used by various application programs.

On the other hand, creating them can be a long and costly process, but once they are made, they can save money and time. Similar to a data warehouse, databases require a significant investment in hardware and software for startup. The most important disadvantage of a database however is that a database does not lend itself to analytics, it is not organized to "facilitate reporting and analysis.

As the main difference between a data warehouse and a database is that databases do not have the ability to analyze the data, we assume for the cost estimation that the equipment is the same as for a data warehouse, but less servers and IT staff will be needed (no processing server, one FTE for each 100 servers). The yearly cost for implementing a database hence is around 40% lower than the cost of implementing a data warehouse, as was anticipated.

2.3.3 Data portal: integrated, centralized website with search engine

Data portals are basically websites that provide different customized facilities to their users. They are designed to be used by different applications. The first web portals were online services that provided access to the information on the web, but by now most of the traditional search engines have transformed into web portals to attract and keep a larger amount of users. As defined by IBM, an Internet portal is “a single integrated, ubiquitous, and useful access to information (data), applications and people” [11].

The advantages of a data portal are the possibilities to search, browse, review and access all data in the platform through a single familiar interface. By providing customizable features and development tools, data portals increase productivity for the end user and increase interaction between data providers and data users. The disadvantages of data portal are the complexity of the search platform, additional testing efforts and related costs. There is a need to customize the portal and integrate applications. Furthermore, no direct analysis is possible because a data portal does not store the actual data, only a link to their source location.

There is a range of costs attached to a data portal. Generally, from $0 (if no one is hired and no additional money is needed for technology) to $500,000 (for hiring new staff and adopting new technology) may be budgeted. However, this cost depends on the type of server and host which will be applied.

2.3.4 Data lake: give me whatever you have

A data lake is a large storage repository that provides massive storage for any kind of data, and has enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. The concept of a data lake is developing as an interesting way to organize and build the next generation of systems, which can deal with new big data challenges. Companies are investing in data lakes because lakes have the ability to store data with increased volume, variety and velocity, which was not an issue in data management in the past. Data lakes are recognized as an evolution in existing data handling systems architecture.

A data lake holds raw and un-processed data in its native format. The main important advantages of a data lake are based on the fact that a data lake is more suitable for big data since it is not limited to specific, static structures that characterize a database or warehouse. However, a data lake has limitations or disadvantages as well. The most important disadvantage relates to searching for specific data through a pool of unfiltered data when this data has no category or identifier. Another issue is that a data lake does not provide explicit guarantees about the quality or usefulness of the data. In other words, gaining direct value from a data lake is difficult.

A data lake is a cost-effective tool to store big data, yet includes a lot less functionalities in comparison to a data warehouse or database. A data lake minimizes the storage costs but still allows accessing the data on the long run, which might be more cost-effective than investing in a full data warehouse. We estimate the cost for a data lake by including the only storage servers (without any IT staff) and compare it to a cloud alternative (Google Cloud Storage, Microsoft Azure and Amazon S3).

Table 2: Cost estimation for a data lake

Storage option	Euro (per TB per year)
Microsoft Azure	333.12
Google Cloud	300.96
Amazon Web Service (Amazon S3)	343.92
Own deployment (ECMN)	1500

Table 2 summarizes the monthly and yearly cost for these different cloud storage options assuming that the price of data storage for every month is the same. Though the comparison learns that the most cost-effective option for storing data are cloud services, a fair comparison between public cloud and a physical, own deployment of a data lake is a complex issue (for example when taking into account utilization, as on the public cloud, businesses pay only for what they use, while in their own system they pay the full cost - whether it is completely busy or not).

2.4 Single point of contact: one responsible within the company

A single point of contact (SPOC) is a person or a department serving as the coordinator of information concerning an activity or program. A SPOC is used when information is time-sensitive and accuracy is important. Although there may only one technician assigned per company as a full-time staff, the costs of all its other technicians that work related and close to the assigned technician, sales staff, and office staff will raise the cost of its services.

The specialized IT staff is paid to know everything about the data. This person is responsible to update the data and the corresponding technical equipment. We estimate the cost for this SPOC by using European levels of salaries for IT staff. Of course, the salary will be increased every year. For example, in the Netherlands (Amsterdam), the average pay for a data analyst is €35,290 per year (2015). A skill in SQL (Structured Query Language) is associated with the high wage for this job. People in this job generally do not have more than 10 years of experience. A business analyst in IT in Belgium (Brussels) earns an average salary of €40,020 per year . Experience has a moderate effect on income for this job.

2.5 Comparison of alternatives

As Table 3 shows, not all of the options have the same functionalities. Data warehouses, databases and data lakes have storage ability. Data warehouses furthermore also have the ability to integrate data from different sources, and of courses integrates possibilities for reporting and analyzing. When it comes to analyzing data, a static list is insufficient. There is a need for aggregating, summarizing, and extracting vision from data. A data warehouse enables to perform many types of analysis, and also enables users to mine the data to extract value and knowledge. In databases, this reporting is typically limited to types that are more static, for example one-time lists in PDF format. These reports are helpful - particularly for real-time reporting - but they do not allow in-depth analysis. Since portals do not store raw data, they also have no analysis or reporting functionality.

Table 3: Comparing different data handling systems based on functionalities

	Data warehouse	Web portal	Database	Data lake	Single point of contact
Storage of massive data sets	x	/	x	x	/
Integration from multiple sources	x	/	x	x	x
Automated reporting	x	/	x	/	/
Automated analysis	x	/	/	/	/
Data mining	x	/	/	/	/
Handling real-time data	x	x	x	/	/
Searchable	x	x	/	+/-	/
Handling historical data	x	x	x	x	x
Privacy	x	/	x	/	/
Security	x	/	x	/	/
Data quality	x	/	/	/	/
Cleaning data	x	/	/	/	/

Furthermore, comparing the cost of handling data per TB for each data handling alternatives can give a better overview about the difference between these different options. Figure 4 shows the average cost of handling 1 terabyte of data by using a data warehouse, database and data lake and cloud storage systems. Although the cost for a data warehouse is higher than the other options, it offers more functionalities: cleaning, processing and standardizing of data to prepare it for analyzing and reporting.

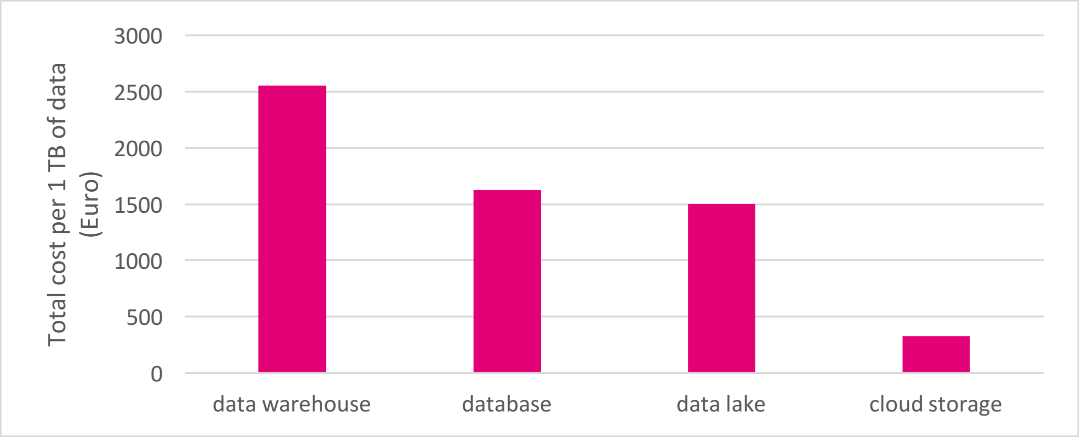

Figure 4: Comparing the total cost per one TB of data for each of data handling systems

3 PEST analysis for mobility data

As a third step in the analysis towards the best data handling process for mobility data, this report presents an overview of the most important political, economic, social and technical issues that are relevant to the big mobility data (PEST framework). Identifying such issues can assist in a better understanding of areas for potential growth and development within the transport industry and assist in boosting the digital economy. The report first discusses the PEST analysis for mobility data in general, then adds some specific issues for the data handling systems described in section 2.3.

3.1 PEST for mobility data in general

Figure 5 visualizes the main impact factors for mobility data in general, they are shortly explained below.

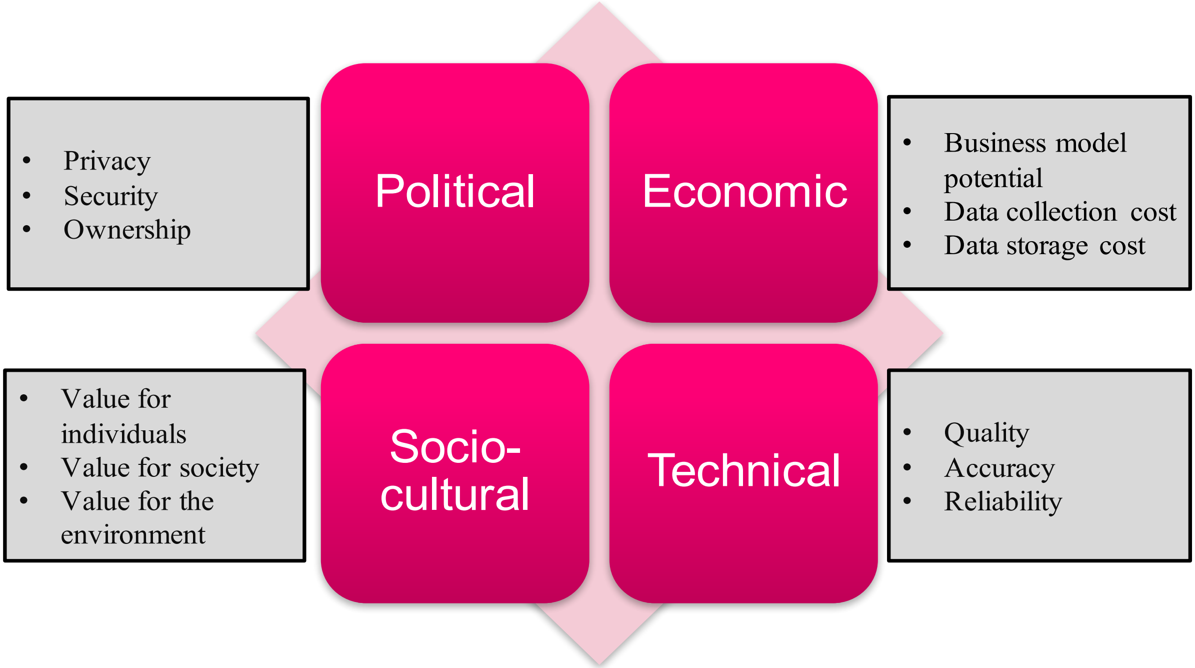

Figure 5: PEST for mobility data in general

3.1.1 Political factors

- **Privacy**
  Mobility data is mostly about people, "where they have been, at what times, how often, and with whom". Therefore, privacy is an important concern for mobility data which needs to be addressed before considering economic opportunities. Determining whether and to what extent information can be collected, processed and stored, by whom and for how long the controller of the file may/must keep it, is an issue for legislation. The collector of data should only collect the data necessary for the achievement of a goal (e.g. location data from GPS being used only to estimate traffic situation).
- **Security**
  Failures to protect the security of personal data could result in data being shared with unauthorized users. As the amount of data being collected continues to grow, more and more companies are willing to aggregate and extract meaning from data. Security management challenges are an important issue because

big data repositories will likely include information deposited by various sources across the company. This variety of data makes secure access management a challenge.

Authorities should make the real benefits more clear: better data, better access to data, and better use of data can lead to equity and environmental impacts. These benefits must be balanced against protection of privacy and security of data [12].

- **Data ownership**
  Local authorities have invested a significant amount of money into data collection. However, different ownerships of data lead to the situation that data collected by one company is not shared with other companies or institutions, and hence part of its value gets lost. Moreover, because of various mobility data standards, data collected for different purposes or by different organizations rarely have the same formats and quality, so may be difficult to share (see further in the technological factors).

3.1.2 Economic factors

- **Potential business models for mobility data**
  Using big data in any sector can lead to important changes in how business processes run, in both the private and the public sector. The explosion of the big data has led policy-makers, industry actors, private companies and academics to consider data as a “resource” which has “value” which can provide a boost to the economy [13]. Business models that rely on personal data (such as cell phone data) as a key input, are becoming more and more common. Attaining significant market valuations by applying business models predicated on the successful use of personal data within the existing legal and regulatory frameworks, is a new trend. There are different potential business models using mobility data in the terms of apps, different mobility and transport services and mobility data analytics for different business usages (e.g. retail sector).
- **Data collection cost**
  The share of mobility data collected by the private sector is growing. The private sector collects millions and millions of data in the context of commercial activity or as a by-product of location-based services [12]. The multiple tasks of collecting, processing and cleaning data have significant cost for authorities and private companies. Mobility data should be collected from different sources by using different techniques which have their own requirements.
- **Data storage cost**
  Above, the cost of data storage is discussed in detail. Storing data for different purposes needs different handling systems, each with their own functionalities and related cost.

3.1.3 Social/societal factors

- **Value for the society**
  Big transport data provides considerable benefits for citizens (e.g. better transport services), for society in general (e.g. a smoother traffic flow), for the public sector (optimization of traffic management), and for service providers (business opportunities). Big and open data also play an important role in how smart cities deploy and use ICT to enhance their transportation networks.
- **Value for individuals**
  If organizations provide individuals with access to their data, creative users can start building applications on their data for new innovative uses. Moreover, the promise of benefits and value sharing propositions will incentivize individuals to share their own data for more profits .

- **Value for the environment**
  The increasing mobility demand has led to severe transportation problems, such as traffic congestions, which in turn are the cause of environmental problems, economic damages and social problems, all of which negatively affect everyday life. One solution for these transportation issues is upgrading the existing or deploying new infrastructures, but this is very costly. Traffic management applications using big data analytics on the other hand can utilize the existing infrastructures more efficiently, leading to better traffic management, less traffic jams and congestion, in turn leading to reduced air pollution.

3.1.4 Technological factors

- **Quality of data**
  With the advent of big data, data quality management has become more important. In the mobility and transport sector, with the huge amount of generated data and the fast velocity of newly arriving data, the quality of data is far from perfect. It has been estimated that erroneous data costs US businesses 600 billion dollars annually. In most data center projects, data cleaning accounts for 30 − 80% of the development time and entire data center budget for improving the quality of the data [14].
- **Accuracy**
- Accuracy (defined as the minimal allowed level of mismatch between the system's estimate and the actual value) of collected, stored and processed data is a critical issue to ensure the proper operations of the transportation system. A wrong estimation of network congestion because of inaccuracy in the measurements can cause to wrong estimates of traffic situation, which in turn lead to ineffective system operations and other consequences such as significant errors in the estimation of total traveled time.
- **Reliability**
  Making mobility data reliable and usable, requires time, effort, and expenses. When collecting data from sensors, Internet of Things, social media or GPS, reliability and continuity is often seen as an important issue.

3.2 PEST analysis for mobility data handling systems

Apart from the general PEST analysis, we shortly add some specific factors important to data handling systems.

3.2.1 Political factors

The absence of well-defined policies in building a data warehouse will lead to conflicts in the Information System roles. Data handling system projects are always potentially political because they change both the terms of data ownership and data access. Who should develop and administrate a data warehouse? Who owns the data? Who controls access to the data? Who maintain the data and ensure its quality? How should we deal with the user needs?

3.2.2 Economic factors

Handling, storage and retrieval systems need careful consideration to ensure that all the relevant data is stored in such a way that it maintains data reliability and allows easy access, retrieval, and updating of the data. Hence, making the right decision about the best data handling system is critical for businesses. It depends on the size of the company, the resources it has and its performance needs.

3.2.3 Social factors

The explosion of data is making data centers one of the fastest-growing users of electricity. Data center electricity consumption is projected to increase to roughly 140 billion kilowatt-hours annually by 2020, the equivalent annual output of 50 power plants, costing American businesses $13 billion per year in electricity bills and causing the emission of nearly 150 million metric tons of carbon pollution annually [15].

3.2.4 Technological factors

The ability to blend data from different sources and aggregate it into an enterprise data center will be critical to deriving value from all of the data collected. Poor architecture, poor infrastructure planning, inadequate components, inadequate or too much data to handle are some of the technical issues related to data handling systems.

4 Business models and cases for mobility data

Mobility data serve many purposes and are provided for and to many users. To meet the wide range of needs, a business handling data system must meet different criteria. Ideally, data should be collected and made accessible in a way that helps business owners answer some important business questions: Where are my customers and potential customers located? Where is the competition located? Where do my employees and potential employees live? Where should I locate my stores, offices, and plants? How much should I produce? How much should I order? How much should I hold in inventory? How should I set my prices? What is the best way to promote my products and services?

Not all stakeholders can use raw data as a material because they do not have specific tools and experts for data analytics. Table 4 summarizes different data users and their actual needs in terms of data requirements.

Table 4: Potential data users and their needs of different data requirements

	Pre-processing (cleaned, missing data points filled)	Standardized	Searchable (categorized, structured)	Minimal delay of data (latency)
Policy/authorities	×	×	×	-
Real estates	×	×	×	-
Citizens	×	×	×	-
Logistics	×	×	×	-
Business app developers	-/×	-	-	×
Research	-	-	-/×	×
Automotive manufacturers	×	×	×	-
Emergency services	×	×	×	-
Private mobility companies	-	-/×	-	×
Consultants	×	×	×	-
local businesses	×	×	×	-

These specific data needs can be linked to the different alternatives for handling mobility data (or data in general) that were investigated before (see section 2.3). Each of these options has its own characteristics to meet the different data functionalities which are discussed above (summarized in Table 3). Following the link between business areas (data users) and data requirements, and the link between these requirements and the different data handling systems, the users can be linked to the right choice of handling system, which in turn provides input for both the qualitative and quantitative cost-benefit analysis. Figure 5 summarizes the steps that we followed to complete this part of report.

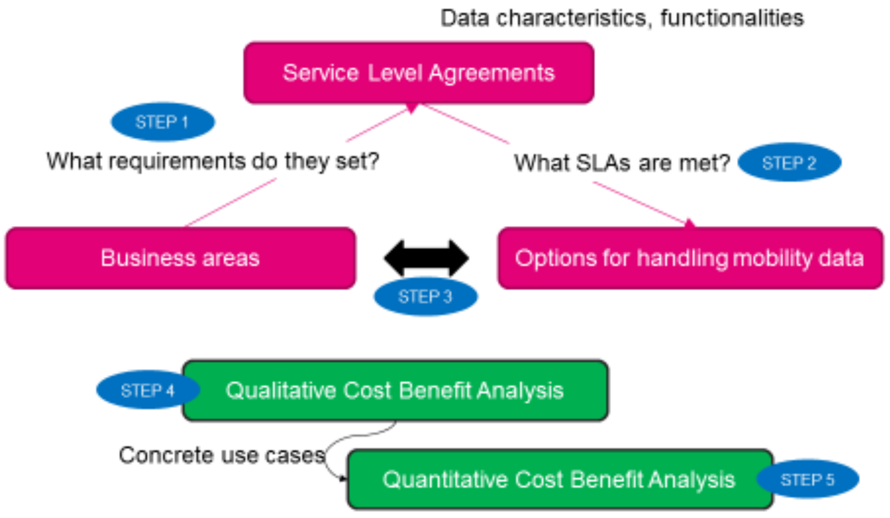
Data characteristics, functionalities
Service Level Agreements
STEP 1
What requirements do they set?
What SLAs are met?
STEP 2
Business areas
Options for handling mobility data
STEP 3
STEP 4
Qualitative Cost Benefit Analysis
Concrete use cases
Quantitative Cost Benefit Analysis
STEP 5

Figure 6: General scheme of followed steps towards cost-benefit analysis for mobility data

4.1 Qualitative cost-benefit analysis for different combinations of handling options and business areas

Generally speaking, the cost of implementing data handling system (data warehouse) is based on the cost of collecting data (which we did not consider in this project) combined with the cost of setting up a physical data center that prepares and processes mobility data for providing information (see section 2.3). The mobility data handling system benefits society by offering information that can be used to e.g. more effectively manage traffic flows on the road network system. This leads to shorter and more reliable travel times that, in turn, result in lower time that drivers spend in traffic congestion. It helps travelers to have reliable estimation on their arrival times and destination. The potential benefits of a data warehouse for different business areas can be summarized as below [2], [16], [17]:

- Quicker and more effective responses to incidents which in turn reduce congestion (Policy/authorities)
- Resolving the discrepancy between all sources of information that can be collected, i.e. different municipalities, different collection systems, different refreshing rates etc. (Policy/authorities)
- Providing reliable and fast access to data concerning the current state of traffic for efficient traffic management and information systems and enabling predicting actual traffic phenomena (Authorities, private companies)
- Improving the traffic flow overall on the network, by restricting access to roads that operate close to their capacity and redirecting traffic to alternative routes (Traffic authorities and logistics)
- Allowing more effective environmental management and control of the air quality (Policy/authorities)
- Contributing to improve traffic safety by monitoring all high-risk locations with high quality data levels (Policy/authorities, private companies)
- Improving the scheduling of management and maintenance of the roads (Policy/authorities)

- Improving the integration and scheduling of public transport (Policy/authorities)
- Developing long-term traffic policies (Policy/authorities, traffic managers)
- Providing opportunities for more effective planning and design of new mobility management strategies, new infrastructures and improvement of the existing ones (Policy/authorities)
- Providing a valuable data center for assessment and research analyses (Research/universities)
- Using information about traffic situation with could of interest to the real estate market (Real estates)
- Vehicle fleet planning (Logistics)
- Using pre-processed data for business (Retailers)
- Business planning, customer management (SMEs, private companies)

4.2 Quantitative cost and benefit analysis for concrete cases

Following the more general, qualitative description of the business potential and benefits of mobility data above, this section discusses a quantitative cost-benefit analysis for two very concrete business cases. The purpose of this analysis is to figure out what could be the monetary value of a business model based on mobility data and different data handling systems. For the monetization of the benefits, a bottom-up approach is used, which will be explained first.

4.2.1 Bottom-up benefit modelling

This model proposes an approach to quantify the value potential of a certain application or product, and consists of two steps [18]. In a first step, the different effects that the application of product will entail should be identified. This can be done using literature review, desk research or interviews or workshops with experts.

In a second step, the quantification model includes three calculation steps per effect. The first calculation step quantifies the total value potential (TVP) per service, which indicates the maximum monetary value that a certain effect could entail, without considering the market that adopted it. It is a simple multiplication of four parameters: the *population group* experiencing influence of the effect, the *benefit expressed in units U* (e.g. amount of km saved by avoiding commuting), the *conversion factor* (e.g. 1 km equals €0.5), and the *occurrence* (e.g. the amount of times the app is used per year).

$$TVP_i(t) = population\ group_i \times unit\ benefit\ [U] \times conversion\ \left[\frac{€}{U}\right] x\ occurence\ (t)$$

Secondly, we calculate the total value per actor per time period by taking in to account the share of the actor and the adoption curve of the service and the TVP for each service. The *Adoption Curve (AC)* of the specific service reflects how fast the service is adopted over time [18]. Typically, four types of actors are taken into account: governments and authorities (G), individuals (I), companies (C) and the society in general (S).

$$TVP_a(t) = \sum\nolimits_i TVP_i(t) \times \frac{share_{ia}}{\sum_k share_{ik}} \times AC_i(t)$$

Finally, the Total Value (TV) for the app can then be calculated by summing all TVPa.

$$TV(t) = \sum\nolimits_a TVP_a(t)$$

4.2.2 Case one: Carambla

Nowadays, more and more drivers drive around the city looking for a free place. This causes traffic congestion to such a level that people can lose a lot of time. In addition, searching for parking causes start-stop driving, which in turn is a source of a lot of pollution and frustration [19]. Carambla offers a solution to this problem: it is a unique, free app based on open data sets which helps drivers to always find the nearest and cheapest parking in Ghent, Antwerp and Brussels. Carambla is the only one in Belgium that can offer this accurate, complete and valuable information on real-time availability, pricing and opening hours of both private and public parking places. These parking spaces are offered by individuals or companies at times when the owners themselves do not use them.

Using Carambla as a service will have different effects on the different actors involved. Using Carambla improves the traffic flow, reduce the traffic jam caused by cars which are searching for a parking place and has also other societal and environmental effects. Table 5 summarizes the social effects of using Carambla, while Table 5 gives an overview of the input parameters used.

Table 5: The effects of using Carambla

Outcomes	Types of benefits	Quantified?
Economic competitiveness	▪ Travel time saving ▪ Vehicle operation saving ▪ Improve local and regional transit ▪ Increased city parking occupancy	▪ Yes ▪ Yes ▪ No ▪ No
Quality of life	▪ Reduction in the commute time, thus more satisfaction ▪ Decrease stress level through providing more convenient way of finding parking in the city	▪ No ▪ No
Environmental impact	▪ Reduction of emissions	▪ Yes
Safety	▪ Reduction of vehicles looking for a parking spot within the network as result of improved guidance on parking availability in the area	▪ No

Table 6: Input data for Carambla in Ghent

Service	Effects	Population group	Unit benefit[U]	Conversion [€/U]	Occurrence (year (X/	Actor's share(%)			
						G	C	I	S
App for finding nearest available parking	Reduced travel time	186289	0.25	10€/h	40	0	0	100	0
	Reduced travel costs	186289	1	0.5€/km	40	0	8.75	91.25	0
	Decreased traffic jams and road accidents	186289	0.98	1€/€	40	0	0	0	100
	Reduced CO2 emission and other harmful gases	186289	120g	0.135€/g	40	0	0	0	100

The total value representing the effects due to using Carambla for a period of 5 years is estimated to amount to about €180 million, divided amongst the society and the individuals (as well as a small percentage for companies, see Figure 7). By spending about €400,000, a data portal can be made which provides data sets for app developers and service providers to offer services like Carambla for better mobility management. As it is clear, the benefit of this service is much higher than the cost of making the data available. It should furthermore be noted that Carambla is not only one app that is made based on the data portal, in reality there could be a lot of these services which can be implemented. On the other hand, the data collection cost nor the cost for developing and maintaining the app is included here, only the cost of opening the data is counted.

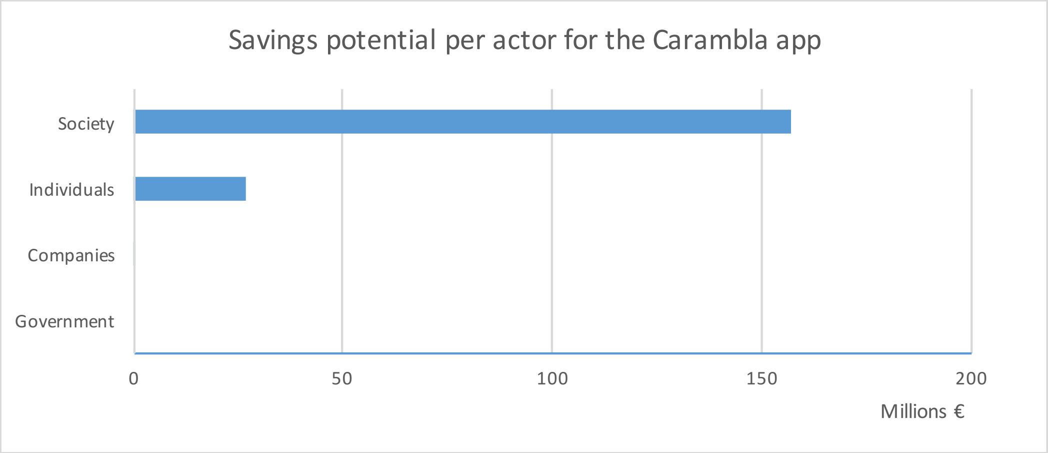

Figure 7: Savings potential per actors for the Carambla app

4.2.3 Case two: Proximus' data analytics

Businesses active in all sectors and industries collect and store data and perform data analysis, to better understand their business processes, reduce expenditures and gain competitive advantage. For this reason, Business Intelligence tools – that report, analyze and process these huge volumes of organizational data into understandable and actionable information – are growing in popularity and importance. As an example in the data analytics industry, Proximus' data analytics improves market insights and helps to know the customer better. Proximus' analytics offers businesses those insights thanks to analyses and reports based on location data [20]. The advantages of these analytics are:

- Obtain new insights to understand the customers and their behavior better: how many visitors attend the specific region/event? Where do they come from?
- Develop new strategies or adjust the sales and marketing tactics based on independent, objective and accurate data;
- Make use of location data to gain a competitive advantage or make adjustments where necessary.

Today, retailers have a wide range of tools available in order to find out what will be this season's "must have" items, whether that be children's toys or designer dresses. Trend forecasting algorithms monitor social media posts and web browsing habits to figure out the shopping patterns and this data can be used to accurately predict what the top selling products in a category are likely to be. The primary purpose of data analytics is to improve the quality of decision-making as better decisions directly impact the business. Below, some of important benefits of data analytics in retail sector are discussed in more detail.

- **Dialogue with consumers:** Nowadays, consumers look around a lot before they buy a service or product, they talk to their entire social network about their purchases. Data analytics allows companies to profile these potential customers increasingly and engage in a one-on-one, almost real-time, conversation with them.

- **Re-develop your products:** Data analytics can also help businesses understand how others perceive their products so that they can adapt them, or their marketing, if it is needed.
- **Perform risk analysis:** Success not only depends on how businesses run their company. Social and economic factors are crucial for their success as well. Predictive analytics allows them to monitor and analyze social media feeds so that they permanently keep up to speed on the latest developments in their industry and its environment.
- **Create new revenue streams:** The insights that companies gain from analyzing their market and its consumers with data analytics are not just valuable to them. They could sell them as non-personalized trend data to large industry players operating in the same segment as the businesses and create a whole new revenue stream.

Cost and benefit analysis for using data analytic techniques in business is not an easy task because advanced data analytics is completely a new tool in the business world. In this section, we make a first estimate of the benefit potential of (big) data analytics and compare it to the cost estimations for a data warehouse (see section 2.3.1).

It is clear that more data allows more concrete analysis about the business, hence increasing the profit potential of the company. Marketing is one way of increasing revenues, its general purpose is to attract more customers by informing more people about a specific product or service. Applying technical tools such as data analytics allows spending this budget more efficiently, as only the right customers will be targeted. To be more precise, knowing what techniques and what marketing tools work best for each customer segmentation lets companies carefully target them rather than randomly bombarding all potential consumers with the same advertising. With data analytics, marketers can determine where to advertise and how to direct communications, which results in fewer wasted marketing budget .

To estimate the benefit of Proximus' data analytics in this report, the same calculation as was used in the previous case is used, assuming that the average budget of marketing per SME is about €30,000 [21] and that using data analytics can save up to 50% of the marketing budget. The estimated benefit of data analytic for retail stores in Belgium over 5 years is about €140 million.

Table 7: Input values for benefit calculation of case 3

Input	Value	Reference
Number of stores in Belgium	7820	[22]
Average spend per SME on marketing	About €30,000 per year	[21]
Number of cellphone users in Belgium	11.8 million	[23]
Proximus' market share	45%	[24]

To compare this benefit with the cost of implementing data warehouse to provide data analytics, we need to estimate the cost of data warehouse for the amount of data that the operator should store. To calculate the cost for the data warehouse, we need an estimate of the amount of data that should be stored, processed and analyzed. Following the amount of cellphone users in Belgium and Proximus' mobile market share, we estimate about 5 million smartphone devices collecting data (numbers from 2013). Furthermore, assuming one cellphone produces one GB of useful location data each year, we estimated the cost for data warehouse for 5000 TB of data (in year 1). The cost for implementing data warehouse for the specific case (Proximus) is about €80 million after 5 years and €140 million after 10 years of implementation.

Comparing the yearly benefit to the yearly cost already learns that this marketing budget reduction is sufficient for the deployment of a (limited) data warehouse. On the other hand, the cost detailed here is only the cost for the data warehouse, and does not take into account the cost for collecting the data and performing the actual analytics (service towards the specific retailers).

5 Conclusion and recommendations

The way data is collected, processed, and stored will fundamentally change in the near-term future from how it is done today. Mobility data (e.g. traffic information) plays an important role in the decision making for road users, traffic planners and other businesses. Because of the growing number of application of ITS technologies, the need for collecting, storing, and manipulating huge amounts of transportation data is becoming more and more significant. Both companies and decision makers now have the opportunity to shape this development. New forms of data collection and new data types lead to new opportunities and business models, but at the same time require a new way of handling. This project aimed to help in tackling this second issues, by investigating two research questions:

- How can the large amount of available mobility data be handled in a dynamic and service-friendly manner?
- What is the value, or what could be the added value of this data to businesses, governments and research institutions?

In a first step, the mobility data are categorized based on different parameters, in order to get to a general overview about the most important mobility data owned by governments, research centers, private companies and individuals. To get an idea about how authorities can deal with mobility data in Flanders, this report presented an overview of mobility data warehouses worldwide, and included different data archiving systems as well. Some countries like the Netherlands implemented a complete data warehouse that archives, stores and processes data in order to receive a general and wide overview about their country's traffic situation, while in some other cases (e.g. Oregon in United States), the portal limits to archived data and a limited set of real-time data.

This worldwide overview suggests that are different types of data handling systems, each with different features, advantages, disadvantages and implementation cost. A data warehouse is the most complex system with the most functionalities, but also requires the largest investment to set up. A database allows storing and querying data, but does not include any analysis capabilities. A data portal provides access to different sources of data through an accessible platform, but does not actually store the raw data. A data lake allows storing a lot of information, but does not impose any storage formats, nor analysis capabilities. The last alternative to handling data is using a single point of contact, which is basically one specialist that has an overview of the available data in a specific company and/or on a specific domain.

Apart from the more technological analysis into different data handling systems, this report also used the PEST framework to investigate political, economic, social and technical issues for mobility data in general and mobility data handling systems specifically. For mobility data in general, privacy, security and data ownership are the most important political issues to consider. The economic issues relate to finding the right business models that can ensure value from data, and of course considers the cost for collection and storage. This means that changes to business models in the mobility and transport sector can be made to increase the value of the data as an asset. The value of data for individual, society and environment are discussed as social issues and finally quality, accuracy and reliability are described as the most important technical issues.

It should be mentioned that, balancing the needs of different data users is not an easy task. Data analysts, for example, prefer access to raw data because of reliability issues, while web and mobile developers often want the pre-processed results from an Application Programming Interface (API) to allow them to quickly and easily build an application, without setting up a custom data processing process. Therefore, the attractiveness of a data handling system depends on its ability to meet the needs of its users. Different mobility users have different requirements regarding data and should hence opt for a different data handling system. By matching the data users with their requirements on the one hand, and the data handling systems with their functionalities on the other, this report could make a qualitative recommendation for the best combination of data user and handling system.

For a selection of concrete cases (an app based on open governmental data, data analytics resulting from a private company's data warehouse), a quantitative cost-benefit analysis was performed. The purpose of this analysis was to determine the monetary value of a business model based on mobility data and different data handling systems. The results clearly showed that in both cases, the effort and investment spent on the deployment of a data portal or data warehouse was worth it for the overall society. It should however be noted that these calculations are based on estimations. In order to use these results as a basis for a real investment analysis, a more detailed study should be undertaken. Furthermore, the here calculated benefits are not necessarily flowing directly to the same actor as the one who is implementing the data handling system (e.g. Proximus deploys a data warehouse, but the main benefits flow to the retailers). Therefore, a more detailed value network analysis including detailed cost and revenue allocation should be performed before using the results for real-life strategic decisions.

In short, this report concludes that each business should use the right data handling system according to its needs and purposes, and that this decision is critical. It depends on the size of the company, its resources and its performance needs. For example, while implementing a data warehouse is the costliest option, it also offers the most functionalities. Therefore, choosing the right data handling system depends strongly on the specific case or application, a full data warehouse is definitely not the best way forward in all situations!

6 References

[1] Sven Vlassenroot, Koen Valgaeren, and Peter Defreyne, "A mobility and transport data warehouse for Belgium: Meta-analyses on different aspects concerning the development of a data warehouses," presented at the 22nd ITS World Congress, Bordeaux, France, 2105.

[2] "Home - Nationale Databank Wegverkeersgegevens." [Online]. Available: http://www.ndw.nu/. [Accessed: 26-May-2016].

[3] "Florida Department of Transportation." [Online]. Available: http://www.dot.state.fl.us/trafficoperations/ITS/RITIS.shtm. [Accessed: 26-May-2016].

[4] "Overview: Singapore Empowers Land Transport Planners With Data Warehouse." [Online]. Available: https://www.gartner.com/doc/1825914/overview-singapore-empowers-land-transport. [Accessed: 26-May-2016].

[5] "VIM | DYNAcity." [Online]. Available: http://www.vim.be/projects/dynacity. [Accessed: 29-May-2016].

[6] "Data.gov.be | open data metadata." [Online]. Available: http://data.gov.be/en. [Accessed: 26-May-2016].

[7] "Data Portals in Belgium | Open Knowledge Belgium." [Online]. Available: http://www.openknowledge.be/belgian-open-data-portals/. [Accessed: 26-May-2016].

[8] "Introduction to Data Warehousing Concepts." [Online]. Available: https://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG001. [Accessed: 26-May-2016].

[9] Lynn Hendrickx, "Techno-economic evaluation of a blow molding production plant: the impact of production time,energy and inventory optimization," Department of Information Technology,Ghent university, 2014.

[10] "Big data: The next frontier for innovation, competition, and productivity | McKinsey & Company." [Online]. Available: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation. [Accessed: 26-May-2016].

[11] "Benefits and Limitations of Portals." [Online]. Available: http://what-when-how.com/portal-technologies-and-applications/benefits-and-limitations-of-portals/. [Accessed: 26-May-2016].

[12] "Data-driven Transport Policy, International Transport Forum." OECD, 2016.

[13] "Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities, Deliverable D2.1: Report on legal, economic, social, ethical and political issues." 2104.

[14] "Data Center Efficiency Assessment Scaling Up Energy Efficiency Across the Data Center Industry:
Evaluating Key Drivers and Barriers." Anthesis, 2014.

[15] "America's Data Centers Consuming Massive and Growing Amounts of Electricity | NRDC." [Online]. Available: https://www.nrdc.org/media/2014/140826. [Accessed: 26-May-2016].

[16] "Sustainable Transport Data Collection and Application: China Urban Transport Database." [Online]. Available: http://www.hindawi.com/journals/mpe/2013/879752/. [Accessed: 26-May-2016].

[17] Francesco Viti, Serge P. Hoogendoorn, Chris M.J. Tampère, Lambertus H. (Ben) Immers, and Sascha Hoogendoorn Lanser, "NATIONAL DATA WAREHOUSE: HOW THE NETHERLANDS IS CREATING A RELIABLE, WIDESPREAD AND ACCESSIBLE DATA BANK FOR TRAFFIC INFORMATION, MONITORING AND CONTROL OF ROAD NETWORKS." 2008.

[18] Marlies Van der Wee, Sofie Verbrugge, and Bert Sadowski, Menno Driesse, Mario Pickavet, "Identifying and quantifying the indirect benefits of broadband networks for e-government and ebusiness: A

bottom-up approach." [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030859611300205X?np=y. [Accessed: 02-May-2016].

[19] "Carambla: Helping you find the nearest and cheapest parkings in Belgium - iMinds." [Online]. Available: https://www.iminds.be/en/business/portfolio/carambla. [Accessed: 02-May-2016].

[20] "Proximus analytics - Proximus." [Online]. Available: http://www.proximus.be/en/id_cl_analytics/large-companies-and-public-sector/solutions/orphans/proximus-analytics.html. [Accessed: 26-May-2016].

[21] "How SMEs make marketing add up | Marketing Week." [Online]. Available: https://www.marketingweek.com/2015/04/01/how-smes-can-make-marketing-add-up/. [Accessed: 26-May-2016].

[22] "List of countries by number of mobile phones in use - Wikipedia, the free encyclopedia." [Online]. Available:

[23] https://en.wikipedia.org/wiki/List_of_countries_by_number_of_mobile_phones_in_use. [Accessed: 11-May-2016].

[24] "Market dynamics." [Online]. Available: https://annualreport.proximus.com/market-dynamics. [Accessed: 11-May-2016].