Van der Wee, Marlies; Bahreini, Samaneh; Verbrugge, Sofie

Conference Paper

How to deal with big data? Techno-economic analysis of different storage, processing and analysis alternatives

27th European Regional Conference of the International Telecommunications Society (ITS): "The Evolution of the North-South Telecommunications Divide: The Role for Europe", Cambridge, United Kingdom, 7th-9th September, 2016

Provided in Cooperation with:

International Telecommunications Society (ITS)

Suggested Citation: Van der Wee, Marlies; Bahreini, Samaneh; Verbrugge, Sofie (2016) : How to deal with big data? Techno-economic analysis of different storage, processing and analysis alternatives, 27th European Regional Conference of the International Telecommunications Society (ITS): "The Evolution of the North-South Telecommunications Divide: The Role for Europe", Cambridge, United Kingdom, 7th-9th September, 2016, International Telecommunications Society (ITS), Calgary

This Version is available at:
http://hdl.handle.net/10419/148712

How to deal with big data? Techno-economic analysis of different storage, processing and analysis alternatives

Marlies Van der Wee, Samaneh Bahreini, Sofie Verbrugge
Department of Information Technology, Ghent University, Ghent, Belgium
Marlies.vanderwee@intec.ugent.be

Abstract – It is beyond doubt that the exponential growth in available data will serve many purposes and will be provided for and to many users. To meet the wide range of needs, a business handling data system must meet different criteria. This paper provides an a both qualitative and quantitative assessment of five different data handling systems: a data warehouse, a database, a data portal, a data lake and a single point of contact. The qualitative assessment relies on the PEST framework to determine the strengths and weaknesses of each option, while the quantitative assessment uses the Equipment Cost Modeling Notation (ECMN) to give a high-level estimation of the implementation cost. The paper concludes that there is a clear trade-off between adding functionality and adding cost, and that the most important decision parameters are the need for storage capacity, the need for a standardized structure and data format, the need for integrated analytics and the requested degree of scalability.

Keywords *– big data, techno-economic analysis, data storage, data management, data processing*

1 Introduction

Driven by an ongoing digital evolution, the amount of collected data is growing exponentially. This growth can be found in the range of application domains that invest in collecting data for information and analytic purposes, as well as the number of collection methods and tools, which span a wide variety of input sources: from online surveys to connected, intelligent devices and sensors [1]. The Boston Consulting Group identifies four global trends that drive this growth: (i) Social Media, (ii) the Internet of Things, (iii) Online Data Transactions and (iv) Digital Services & Media [2]. According to Chen, Mao, & Liu [3], Cloud Computing should also be added as a driving factor. These developments made the global amount of created and copied data rise to 1.8 ZB in 2011. It is expected that this growth factor has since been at least doubled every two years. Both companies and governments have increasing amounts of data, while studies show that uncontrolled data growth is slowing down the deployment of new

applications based on data and information [4]. In the specific case of the Internet of Things, McKinsey calculated in its 2015 report that less than 1% of available data is actually used [5]. On the other hand, the ability to extract value from data is recognized as a powerful competitive factor in different sectors. Businesses in different industries increasingly rely on data to make critical managerial decisions. Most of the leading companies (e.g. IBM, Google, SAS) are investing in intelligent data management systems and advanced data analytics to enhance their capabilities, manage the risks and to meet the growing needs for agile solutions. Hence, one of the biggest emerging challenges for data scientists now is how to keep up with this fast growing amount of data in order to subtract sufficient added value from these data: how should this huge amount of data be stored, processed and analyzed as efficiently as possible [6]?

An important decision for organizations is to choose the most suitable system for handling their data, such that they can adapt their decision-making process to their needs and the characteristics of the available data [7]. Management, storage and retrieval systems should be carefully selected and designed 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.

In order to support this major business decision, this paper identifies and compares different alternatives for managing, storing and analyzing big data: a data warehouse, a database, a data portal, a data lake and a single point of contact. 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, without storing 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 the "human approach": using a *single point of contact*, which is basically one specialist that has an overview of the available data in a specific region and/or on a specific domain. This is an introductory paper and aims at providing an overview of available options for data handling as well as guidelines for different interested stakeholders, in form of different decision characteristics. The paper can furthermore serve as a definition guideline, as not all literature sources adopt the same definition for the different data handling options (e.g. the definition of a data lake in one source may correspond to the definition of a data warehouse in another one).

The paper uses both a qualitative as well as quantitative approach to tackle this comparison, the underlying frameworks and modelling languages will be presented in section 2. The qualitative comparison aims at identifying the functionalities (e.g. storing data, querying data) and advantages, disadvantages and issues related to each of the five options, relying on the PEST framework. The quantitative part is based on a techno-economic calculation of the cost for each alternative, using the ECMN (Equipment Coupling Modeling Notation). Section 3 describes and assesses the five data handling options in detail, whereas the different options are compared in section 4. Finally, section 5 concludes this paper and gives some recommendations for different types of data users.

2 Framework for analysis

This section shortly introduces the frameworks and modelling languages used for assessing the different data handling alternatives, both in a qualitative and quantitative manner.

2.1 Qualitative assessment supported by the PEST framework

The qualitative assessment aims at identifying the different functionalities, advantages and disadvantages of the data handling systems. In order to do this in a structured manner, we opted to use the PEST framework. PEST is a simple and widely used framework that aims to identify the parameters that have an impact on the organization or business by recognizing them as Political, Economic, Social and Technological influences. PEST analysis helps to understand the "big picture" and forces of change that businesses are exposed to, and from this, take advantage of the opportunities that they present [9].

The main factors that spring to mind when performing a PEST analysis for data management are:

- Political aspects: legislation and regulation, funding and support, privacy, ownership and security
- Economic aspects: international and national trends, costs for collection and storage, revenues, business models
- Socio-cultural aspects: environmental issues, value for the user, value for society
- Technological aspects: technical advances (new technologies), quality, accuracy, reliability

The growing importance of environmental or ecological factors have given rise to green business and also legal issues encouraged widespread use of an updated version of the PEST framework or PESTEL model. Because most of legal and environmental issues will be covered in the political and social parts, this paper sticks to the PEST framework.

2.2 Quantitative cost assessment relying on the ECMN model

The quantitative assessment focuses on the deployment and operational cost for the different data handling options. The model used for estimating the deployment cost for each alternative is the Equipment Coupling Modeling Notation (ECMN) [10]. ECMN is a tool that can be applied to different areas of study to calculate the cost of equipment that should be installed for a specific project. 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 only installing 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. We refer to the specific models in section 3 for examples of this modelling language.

For all of the options in this paper, the same assumptions with regards to the amount of data will be used. We assume 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. We furthermore assume that all historical data is also kept in the system. The linear growth function is based on conventional data collection (e.g. traffic counting loops in the road

surface, employee administrative data) as no (significant) increase in the number of data sources is expected in the future (the yearly data 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, such as data captured by sensors, smartphones, social media, and the Internet of Things (IoT) [11]. Figure 1 represents both growth curves.

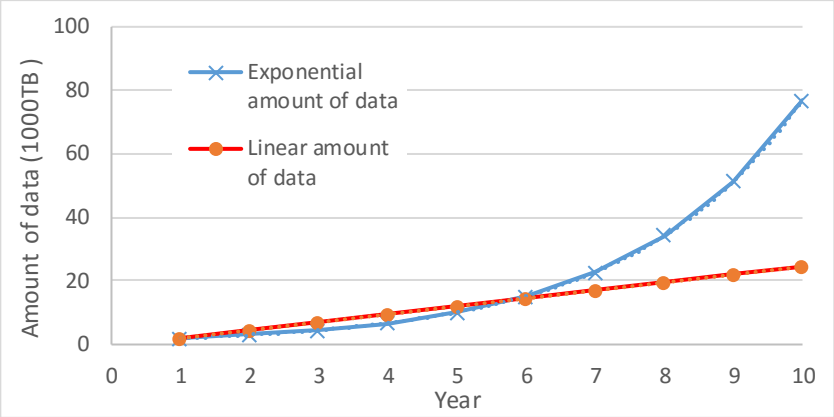

Figure 1: Yearly amount of data to store and process in data warehouse

3 Introducing and assessing the different data handling alternatives

Using the qualitative and quantitative framework described above, this section defines and assesses the different identified data handling options. For each of the systems, a definition will be given first, followed by its functionalities and issues identified using PEST, to end of with a quantitative cost assessment using the ECMN modelling.

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" [12]. 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. As such, a data warehouse goes beyond the simple storage of data and information from different sources: it is a centralized warehouse to store, process and analyze data and information [13]. 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 centralize important information in the same location instead of keeping data in several different places. This data can then be used to optimize strategic decisions.

3.1.1 Qualitative PEST assessment

As mentioned above, data warehouses are the most extensive solution for storing, managing and analyzing data. This wide range of functionality has a large number of benefits, but also comes at high costs and risks. This section will analyze the strengths and weaknesses of data warehouses based on the PEST framework as introduced above. Please note that the PEST analysis of the other

data handling options (see further) will be based on a comparison to the analysis of data warehouses here.

Political

Data handling system projects are always potentially political because they change both the terms of data ownership and data access. As data warehouses store data from different sources into one platform, privacy and security of these data should be safeguarded. Not only should the data be stored centrally, also access and control to the data should be managed from one control point or at least uniformly across users. It should furthermore be formally contracted who has the ownership of the data, as well as the responsibility.

Economic

Establishing a data warehouse requires a significant investment cost for storage capacity, processing power and initial setup. Data warehouses have the highest functionality (storage, processing and analytics), hence also require the highest cost. Storage costs relate to physical server space (can be own server racks or an external cloud-based option), while processing power is needed for formatting, cleaning and checking for accuracy and quality. As continuous availability of data frequently is a hard requirement, costs are increased to ensure redundancy.

These costs are of course made to increase functionality: data warehouses have a high revenue potential because the data is easily searchable and because the platform offers integrated analytics and reporting tools.

Social

The social value of data handling systems in general is implicitly recognized, though not yet easily quantifiable. It is clear that the greatest value of data warehouses is for improving the efficiency and productivity of (big) data storage and analytics, but it is yet to be seen how this value can generate spillover effects to society.

The explosion of data on the other hand already has visible effects on the environment; it 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 [14].

Technical

As mentioned above, data warehouses have the highest functionality and hence most complex technical implementation. Data stored in a data warehouse is always available in a standardized, consistent format, and its users can be assured of its quality and accuracy. As a downside, the process of cleaning, loading and checking before storing can take a long time and caused unwanted delays (especially for data that should be available real-time).

The standardized format however makes a data warehouse easy to use (limited training required before being able to use the searching and analytics functionalities), as well as scalable towards storing large amounts of data or integration of multiple data sources.

3.1.2 Quantitative cost estimation

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. Its cost is estimated using the ECMN model, visualized in Figure 2. In this figure, one can identify one driver, the amount of data, that determines the amounts of all types of equipment that are needed. In a first step, the amount

of servers (both storage and processing) are calculated using the granularities depicted on the connection lines. For example, one storage server is needed for every 2TB of data, hence a new server is installed each time the previous one is full, so when exceeding an amount of 2 TB (i.e. 2TB, 4 TB, 6 TB, 8 TB, etc.). The amount of servers then acts as a driver for calculating the amount of IT staff, power, cabling, switches and operational systems. Note the summation sign that indicates that the total amount of both servers has to be taken into account for dimensioning. Finally, servers and switches are stored in racks, as indicated on the right side of the figure.

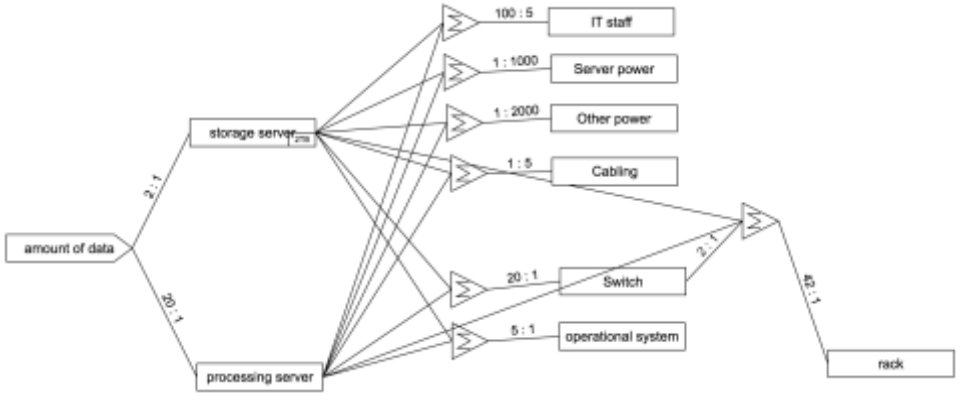

Figure 2: Cost model for data warehouse based on ECMN model

The amount of data is the main driver of the cost: the more data needs to be handled, the higher the cost for the system. Because a data warehouse allows both storing and processing of the data, two types of servers are considered: 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. All cost assumptions are listed in Table 1 below.

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 monthly salary	2000, yearly increase of 5%

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, after 5 years the exponential growth curve overtakes the linear one in terms of total cost.

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.

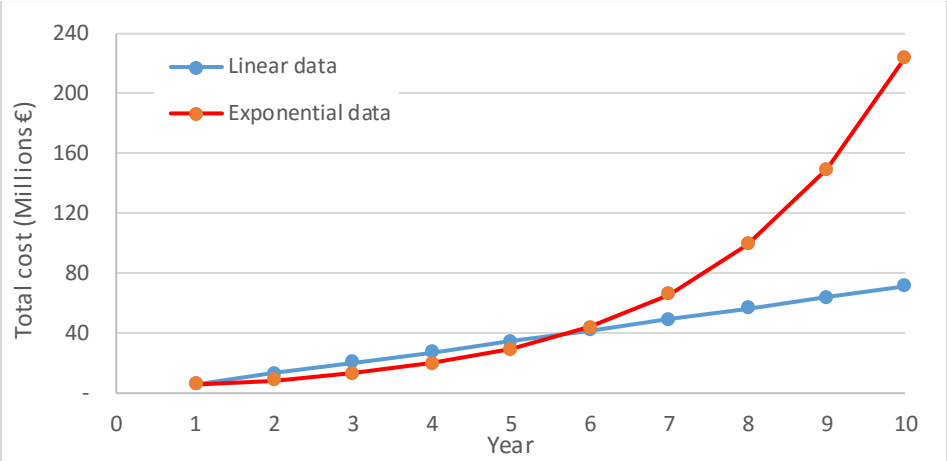

Figure 3: Total cost of implementing a data warehouse

3.2 Database: a structured collection of standardized data

A database is a collection of information that is organized such that transactions are handled efficiently: a computer program can very quickly search and select desired information and data [15]. Databases are designed to manage large amounts of data by storing and retrieving that information. The analysis functionality of a data warehouse is however not included in a database, hence databases are often referred to as the "data store" of a data warehouse [16]. Or reversely, a data warehouse is often defined as "a database designed to enable business intelligence activities" [17].

Common databases are organized by fields, records and files. A field is a piece of information, a record is one complete set of fields and a file is a collection of records. Once the records are created in the database, they can be sorted in different ways, or can be linked by relationships (relational database).

3.2.1 Qualitative PEST assessment

As databases have similar functionalities as data warehouses (apart from the integrated analytics), their PEST evaluation is very comparable as well. Differences can be noted in the economic part, where processing and energy costs will be lower, which can be opposed by higher costs needed for manual intervention for analytics and reporting. Lower processing power might also lead to lower control on the quality of the data, though this can be a separate decision investment. Due to their standardized way of structuring and visualizing data, the ease of use and searching in databases can be even higher than in data warehouses.

3.2.2 Quantitative cost estimation

As the main difference between a data warehouse and a database is that databases do not have the ability to analyze the data (no integrated analytics), 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). This is reflected in Figure 4. The yearly cost for implementing a database hence is around 40% lower than the cost of implementing a data warehouse.

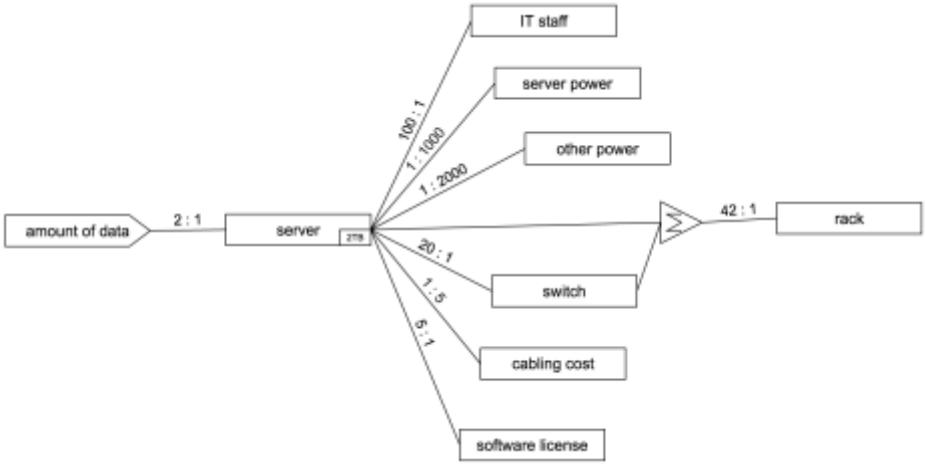

Figure 4: Cost model for database based on ECMN model

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" [18].

A web portal is most often defined as a one special designed website that manages information from diverse sources in an integrated way. Usually, each information source gets its dedicated area on the page for displaying information. A portal may look like a website, but it is much more [19]:

- **Single access point**: a single gateway or logon to identify approved users, making it unnecessary to sign onto each of the different systems that provide portal content
- **Internet tools**: site search and navigation tools to provide users with easy access to information.
- **Collaboration tools**: e-mail, chat, etc. that offer a whole range of ways to communicate and share information.
- **User customization**: When that user authenticates to the portal, this information determines what he/she will see on the home page immediately after login.
- **User personalization**: A portal enables the end user to take customization one step further, e.g. to subscribe and unsubscribe to channels and alerts, set application parameters, create and edit profiles, add or remove links.

One good example of a data portal is given by Dzemydienė et al. [20]: they designed one overarching web portal for water resource management that links to different EU Member States' data warehouses and allows combining data from these different sources.

3.3.1 Qualitative PEST assessment

The main difference between a data portal and a database or warehouse is the fact that the actual data is not stored on the platform itself. A portal rather directs its users to the storage source of the data. This fact significantly reduces the political issues of privacy, security and data ownership. The web-based interface (search engine) that a portal frequently is based on, furthermore simplifies users access and control. By providing customizable features and development tools, data portals increase productivity for the end user and increase interaction between data providers and data users.

On the economic side, a data portal has much lower cost: there is no need for storage capacity and processing power can hence also be significantly reduced. It is yet still perfectly searchable, though cannot provide any guarantees about the quality, format or continuous availability of the data. In terms of effects for the environment, a data portal is more energy-friendly, especially if the alternative would be duplicating data from different sources to be stored on one platform.

3.3.2 Quantitative cost estimation

The cost for a data portal mainly consists of two parts: a limited cost for hosting and setting up the platform, and a significant cost for adding the information and keeping it up to date. The main benefit is that there is no need for high storage or processing costs since a portal only links to the source data, but does not provide storage itself. Hosting costs can be limited to $500 per year or less, maintenance costs strongly depend on the amount of data sources, but will typically not need more than 1 FTE to manage.

Overall, the cost of a data portal of significant size will be limited to $25,000 - $30,000 per year.

3.4 Data lake: give me whatever you have

A data lake is a large storage repository that holds raw and un-processed data in its native format. It hence stores different types of data (ranging from pure text to video, or audio files) while ignoring almost everything around. In other words, there are no predefined rules about how or when its data should be used, governed, defined or secured [21].

Companies are investing in data lakes because lakes have the ability to store data with increased volume, variety and velocity. Big Data initiatives have begun to use data lakes because they have the ability of store all data in an unstructured, unorganized format. The data is not specialized with the specific format, meaning that it can be transformed in a variety of ways. This might be the biggest benefit of data lakes because it is difficult for data scientists to uncover insights in when data is pre-processed and pre-organized.

3.4.1 Qualitative PEST assessment

The main distinguishing factor of a data lake is the lack of standardization, which leads to a lower implementation cost (virtually no cost for processing, cleaning, checking) compared to a database or warehouse but on the other hand also significantly reduces the reliability and direct usability of the data: there are no guarantees towards data quality, accuracy, or format. The most important consequence of this pool of unfiltered data is that searching for specific data becomes very difficult and time-consuming. On the other hand, it can be a good solution to store data temporarily, such that the structure of the data can be defined at the time it will actually be used, or it can be used for storing different types of data in the same storage location.

3.4.2 Quantitative cost estimation

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 only storage servers in our ECMN model (without any IT staff) and compare it to a couple of cloud, Infrastructure as a Service (IaaS) alternatives (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
Google Cloud	301
Amazon Web Service (Amazon S3)	344
Own deployment (ECMN)	1500

Table 2 summarizes the 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).

3.5 Single point of contact: one responsible person/department 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 (or persons) is (are) responsible to update the data and the corresponding technical equipment.

3.5.1 Qualitative PEST assessment

As a SPOC can be seen as a human version of a portal, the political risks regarding privacy, security and data ownership are reduced, be it that an important social issue arises: trust. Since the SPOC has access to and control over almost all data and information, it is important that he or she displays a proper behavior. Whereas a data portal is a digital link to the relevant data sources, the SPOC represents a manual search. As such, one of the most important disadvantages of a SPOC is the potential delays in information retrieval (e.g. only during business hours), hence impacting the continuous availability of the data. As the amount of information one FTE can handle is limited, this solution is also not very scalable.

3.5.2 Quantitative cost estimation

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, [22]). 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 [23]. Experience has a moderate effect on income for this job.

4 Comparison of the five options

This section compares the five data handling options described above, again using the PEST framework. As Table 3 shows, not all of the options have the same functionalities, strengths and risks. Data warehouses, databases and data lakes have storage ability. Data warehouses furthermore also have the ability to integrate data from different sources, and integrate possibilities for reporting and analyzing. 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.

This extended functionality offered by a data warehouse also comes at a cost: not only an economic cost for storage capacity, processing power and setup, but also a political cost in ensuring privacy and security and managing data ownership. Storing data in the platform itself (database, warehouse, lake) versus referring to external sources (portal, SPOC) clearly influences both economic and political cost and risk. Furthermore, duplicating data to a new platform (database or warehouse) significantly increases the impact on the environment, as data centers are a huge energy consumer.

Standardization in terms of data structure and format is another important influencing factor: the offered platform becomes much easier searchable and manageable towards analytics and reporting, but also requires a significant investment in data cleaning, processing and checking.

Finally, digital systems are more scalable than manual intervention.

Table 3: Comparing different data handling systems based on functionalities and issues

	Data warehouse	Database	Data portal	Data lake	Single point of contact
POLITICAL					
Privacy issues	Yes	Yes	No	Yes	Limited
Security issues	Yes	Yes	No	Yes	No
Data ownership issues	Yes	Yes	No	Yes	Limited
Access and control issues	Yes	Yes	Minimal	Yes	No, but delay
ECONOMIC					
Investment in storage capacity	High	High	Low	High	Low
Investment in processing power	High	Medium	Low	Medium	None
Investment in platform set-up	High	Medium	Medium	Low	None
Searchable	Yes	Yes	Yes	No	Manual
Integrated analytics	Yes	No	No	No	No
SOCIAL					
Value for general society	-	-	Yes	-	-
Trust issues	No	No	No	No	Yes
Impact on environment (energy)	High	High	Low	High	Low
TECHNICAL					
Data quality control	Yes	Yes	No	No	Limited
Data accuracy	Yes	Limited	No	No	Limited
Continuous availability	Yes	Yes	No	Yes	No
Scalability	Yes	Yes	Yes	Yes	No
Standardization (consistency)	Yes	Yes	No	No	Depends
Ease of use (upfront knowledge)	Medium	Medium	High	Low	High

These functionalities should be combined to a cost comparison in order to clarify the trade-off that is involved. Figure 5 shows the average cost of handling 1 terabyte of data by using a data warehouse, database and data lake and cloud storage systems. These costs were determined taking into account the linear growth curve and an operational period of 10 years. This graph makes it clear that offering more functionality also increases the cost. Please note that the costs for data portal and SPOC are not included here, because they do not depend (as much) on the amount of data. Their costs lie in the same order of magnitude (€25,000 to €35,000 per year).

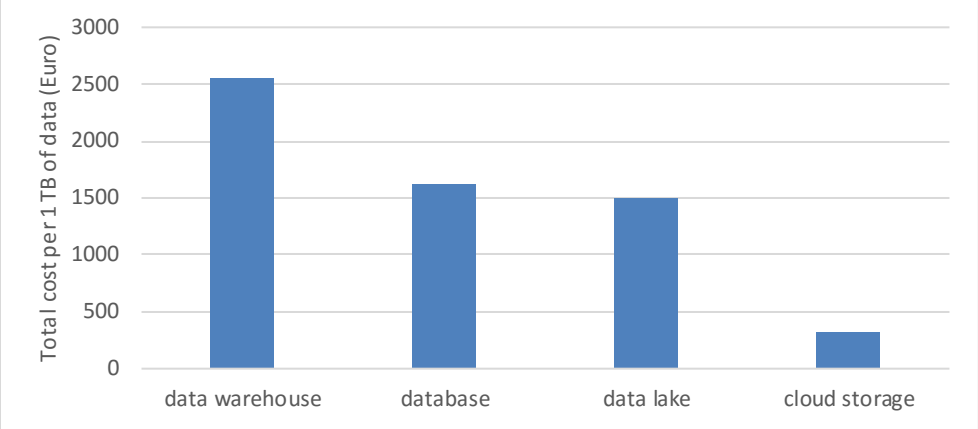

Figure 5: Comparing the total cost per one TB of data for each of data handling systems

5 Conclusion and recommendations

This paper aimed at giving an overview of different data handling options: a full-scale data warehouse, structured storage in a database, a searchable data portal that links to the right data source, a data lake of huge amounts of different data types and the "manual" alternative of the Single Point of Contact. Based on both a qualitative assessment of each option's strengths and weaknesses (using the PEST framework), as well as a quantitative high-level assessment of the implementation cost, the paper showed that the different options provide different functionalities, but also come at a very different cost.

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. 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. Finally, governments and other authorities are not interested in the data as such, but rather value the information and knowledge resulting from these data, and hence are more interested in analyzed data.

Each of these stakeholders could use the analysis provided in this paper to make a decision on the best suited data handling system for their needs. Important parameters to take into account are the need to include storage capacity (database, warehouse, lake) versus linking to data sources (data portal, SPOC), the need for standardized structure and format of the data (included in a database and warehouse) but hence also facing the extra processing and cleaning cost, the need for integrated analytics (data warehouse) and the scalability of a digital system versus manual intervention.

References

[1] Cisco. (2015). Cisco Visual Networking Index: Forecast and Methodology, 2014–2019. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

[2] BCG. (2012). The value of our digital identity.

[3] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171–209. doi:10.1007/s11036-013-0489-0

[4] White, C. (2006) A roadmap to enterprise data integration. IBM. ftp://public.dhe.ibm.com/software/emea/de/db2/A-Roadmap-To-Enterprise-Data-Integration.pdf

[5] Manyika, J. et al. (2015) The Internet of Things. Mapping the value beyond the hype. McKinsey Global Institute.

[6] Xiaofeng, M., & Xiang, C. (2013). Big data management: concepts, techniques and challenges [J]. *Journal of Computer Research and Development*, *1*, 98.

[7] Enricq (2008) Data Management Alternatives for Intelligent Devices. TechOnline. http://www.techonline.com/electrical-engineers/education-training/tech-papers/4134482/Data

[8] Casier, K., Van der Wee, M., & Verbrugge, S. (2014, July). Cost evaluation of innovative offers using detailed equipment, process and network modeling languages. In *Transparent Optical Networks (ICTON), 2014 16th International Conference on* (pp. 1-4). IEEE.

[9] PEST Analysis - Strategy Tools From MindTools.com. https://www.mindtools.com/pages/article/newTMC_09.htm

[10] Spruytte, J., Van der Wee, M., Verbrugge, S., Colle, D. (2016) Introduction of BEMES, a webtool to simplify business process and equipment cost modelling. *BMSD, June 2016, Rhodes, Greece.*

[11] McKinsey Global (2011) Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

[12] Introduction to Data Warehousing Concepts. https://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG001

[13] Krishnan, K. (2013). Data warehousing in the age of big data. Newnes.

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

[15] Cardon, D. (2014). Database vs data warehouse: A comparative review.

[16] Watson, H. J. (2002). Recent developments in data warehousing. Communications of the Association for Information Systems, 8(1), 1.

[17] Oracle. (2016) Database Data Warehousing Guide. Available: https://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG001

[18] Daigle, S. L., & Cuocco, P. M. (2002). Portal technology opportunities, obstacles, and options: a view from the California State University. *Web portals and higher education technologies to make IT personal. San Francisco, CA: Jossey-Bass*, 109-123.

[19] Benefits and Limitations of Portals. Available: http://what-when-how.com/portal-technologies-and-applications/benefits-and-limitations-of-portals/.

[20] Dzemydienė, D., Maskeliūnas, S., & Jacobsen, K. (2008). Sustainable management of water resources based on web services and distributed data warehouses. Technological and Economic Development of Economy, 14(1), 38-50.

[21] CITO Research (2014) Putting the Data Lake to Work. A Guide to Best Practices.

[22] PayScale (2015) Data Analyst Salary (the Netherlands). Available: http://www.payscale.com/research/NL/Job=Data_Analyst/Salary

[23] PayScale (2015) Business Analyst, IT Salary (Belgium). Available: http://www.payscale.com/research/BE/Job=Business_Analyst,_IT/Salary/