Opportunities to manage big data efficiently and effectively

A study on big data technologies, commercial considerations, associated opportunities and challenges

---

Zeituni Baraka

Zeituni Baraka

2014-08-22

Dublin Business School, tunibaraka@yahoo.com

Word count 20,021

Dissertation MBA

Acknowledgements

I would like to express my gratitude to my supervisor Patrick O'Callaghan who has taught me so much this past year about technology and business. The team at SAP and partners have been key to the success of this project overall.

I would also like to thank all those who participated in the surveys and who so generously shared their insight and ideas.

Additionally, I thank my parents for proving a fantastic academic foundation on which I've leveraged on at post graduate level. I would also like to thank them for modelling rather than preaching and for driving me on with their unconditional love and support.

TABLE OF CONTENT

Abstract

Research enquiry: Opportunities to manage big data efficiently and effectively

Big data can enable part-automated decision making. By by-passing the possibility of human-error through the use of advanced algorithm, information can be found that otherwise would be hidden. Banks can use big data analytics to spot fraud, government can use big data analytics for cost cuts through deeper insight, the private sector can use big data to optimize service or product offering as well as targeting of customers through more advanced marketing.

Organization across all sectors and in particular government is currently investing heavily in big data (Enterprise Ireland, 2014). One would think that an investment in superior technology that can support competitiveness and business insight should be of priority to organization, but due to the sometimes high costs associated with big data, decision makers struggle to justify the investment and to find the right talent for big data projects.

Due to the premature stage of big data research, the supply has not been able to keep up with the demand from organizations that want to leverage on big data analytics. Big data explorers and big data adopters struggle with access to qualitative as well as quantitative research on big data.

The lack of access to big data know-how information, best practice advice and guidelines drove this study. The objective is to contribute to efforts being made to support a wider adoption of big data analytics. This study provides unique insight through a primary data study that aims to support big data explorers and adopters.

Background

This research contains secondary and primary data to provide readers with a multidimensional view of big data for the purpose of knowledge sharing. The emphasis of this study is to provide information shared by experts that can help decision makers with budgeting, planning and execution of big data projects.

One of the challenges with big data research is that there is no academic definition for big data. A section was assigned to discussing the definitions that previous researchers have contributed with and the historical background of the concept of big data to create context and background for the current discussions around big data, such as the existing skills-gap. An emphasis was placed on providing use cases and technical explanations to readers that may want to gain an understanding of the technologies associated with big data as well as the practical application of big data analytics.

The original research idea was to create a like-for-like data management environment to measure the performance difference and gains of big data compared to traditional database management systems (DBMS). Different components would be tested and swapped to conclude the optimal technical set up to support big data. This experiment has already been tried and tested by other researchers and the conclusions have been that the results are generally biased. Often the results weigh in favor of the sponsor of the study. Due to the assumption that no true conclusion can be reached in terms of the ultimate combination of technologies and most favorable commercial opportunity for supporting big data, the direction of this research changed.

An opportunity appeared to gain insight and know-how from big data associated IT professionals who were willing to share their experiences of big data project. This dissertation focuses on findings from a surveys carried out with 23 big data associated professionals to help government and education bodies with the effort to provide guidance for big data adopters (Yan, 2013).

Big data definition, history and business context

To understand why big data is an important topic today it's important to understand the term and background. The term big data has been traced back to discussions in the 1940's. Early discussions where just like today about handling large groups of complex data sets that were difficult to manage using traditional DBMS. The discussions were led by both industry specialists as well as academic researchers. Big data is today still not defined scientifically and pragmatically however the efforts to find a clear definition for big data continue (Forbes, 2014).

The first academic definition for big data was submitted in a paper in July 2000 by Francis Diebold of University of Pennsylvania, in his work in the area of econometrics and statistics. In this research he states as follows:

"Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. In this new and exciting world, sample sizes are no longer fruitfully measured in "number of observations," but rather in, say, megabytes. Even data accruing at the rate of several gigabytes per day are not uncommon."

(Diebold.F, 2000)

A modern definition of big data is that it is a summary of descriptions, of ways of capturing, containing, distribute, manage and analyze often above a petabyte data volume, with high velocity and that has diverse structures that are not manageable using conventional data management methods. The restrictions are caused by technological limitations. Big data can also be described as data sets that are too large and complex for a regular DBMS to capture, retain and analyze (Laudon, Laudon, 2014).

In 2001, Doug Laney explained in research for META Group that the characteristics of big data were data sets that cannot be managed with traditional data management tools. He also summaries the characteristics into a concept called the ‘’Three V’s’’: volume (size of datasets and storage), velocity (speed of incoming data), and variety (data types). Further discussions have led to the concept being expanded into the "*Five V's*": volume, velocity, variety, veracity (integrity of data), value (usefulness of data) and complexity (degree of interconnection among data structures), (Laney.D, 2001).

Research firm McKinsey also offers their interpretation of what big data is:

"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don't define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes)''

(McKinsey&Company,2011)

The big challenge with big data definition is the lack of measurable matrix associated, such as a minimal data volume or a type of data format. The common understanding today, is that big data is linked with discussions around data growth which is linked with data retention law, globalization and market changes such as the growth of web based businesses. Often it's referred to data volumes above a petabyte of Exabyte, but big data can be any amount of data that is complex to manage and analyze to the individual organization.

Why is big data research important?

The reason why this research is relevant is because big data has never been as business-critical as today. Legal pressures and competition is adding to pressures to not just retain data, but leverage on it for smarter, faster and more accurate decision making. Having the ability to process historical data for analysis of patterns trends and to gain previously unknown facts, provides a more holistic view for decision makers.

Decision makers see value in having the ability to leverage on larger sets of data which will give them granule analysis to validate decision. The sort of information that organization look for can be also contrasting information, discrepancies in data, evidence of quality and credibility. The rationale behind the concept of big data is simple, the more evidence gathered from current and historical data; the easier it is to turn a theory into facts and the higher the probability is that what the data shows is conclusive. It sounds like a simple task; simply gather some data and use a state of the art Business Intelligence solution (BI) to find information. It has proven to not be easy as management of larger sets of data is often time consuming, resource heavy and in many cases expensive (Yan, 2013).

Big data can help to improve prediction, improve efficiency and create opportunities for cost reductions (Columbus 2012). The inability to find information in a vast set of data can sometimes affect competitiveness and halter progression as decision makers don't have facts to support justification. Yet, organizations struggle to justify the investment in big data despite awareness of the positive impact that big data analytics can have.

Big data issues

Decision makers find it difficult to decide on budget allocation for big data. Is big data an IT matter that should be invested in using IT budget? Or is big data a marketing and sales matter? Perhaps big data is a leadership matter that should be funded by operations management budget? There is no wrong or right answer. Another issue that decision makers are struggling with is defining the measurement and key performance indicators (KPI's) to assess potential and results. What defines return on investment (ROI) can be difficult to establish and results can often not be proven before the investment is already made (Capgemini, 2014).

Performance demanding applications and internet based applications, along with data retention laws has forced software developers to rethink the way software is developed and the way data management is carried out. Business processes are today often data driven and business decisions often rely on business intelligence and analytics for justification. There are global pressures around accountability and emphasis on the importance of historical documentation for assurance of compliance and best practice.

Governments are currently working to narrow technology knowledge gaps associated with big. They're also working to provide guidelines, policies, standards and to enforce regulations for use of big data technologies (Yan, 2013). Moral issues around big data are mainly around legislations and privacy laws. Experts worry that if lagers data volumes are retained, the risk is higher would the data be compromised. There are few internationally agreed standards in terms of data management. The lack of legislations around web data in particular can lead to misuse.

Big data is subject to laws like the Data Protection (Amendment) Act 2003 and ePrivacy regulations Act 2011. However they don't give much guidance in terms of best practices. Data sources such as social media are also very loosely regulated (Data Protection Commissioner, 2014). New legislations around security and accounting law require

organization's to retain email archives for longer than before, in the US for example it's 5 years (Laudon, Laudon, 2014)

Organizations are challenged with preparing legacy systems and traditional IT environments for big data adoption. If for example a company struggle with data quality or poor implementation results of previous hardware and software, a big data investment would be ineffective. To ensure success, knowledgeable management is needed. McKinsey state that that there is a need for 1.5 million data knowledgeable managers in the US to take advantage of the potential that big data brings along with a need for 140,000-190,000 analytical professionals (McKinsey&Company,2011).

In a study in 2002 commissioned by ITAC, the top most sought after IT skills were identified. SQL Server, SQL Windows, IT security skills, Windows NT Server, Microsoft Exchange and wide area networks skills topped the list (McKeen, Smith, 2004). Just 12 years later, the demand looks very different with advanced analytics, cloud, mobility technology, and web skills being at the forefront of discussions. All of these skills are relevant for big data projects.

Big data opportunities

Researcher David J.Teece talks about competitive advantage in his publication *Managing Intellectual Capital* in a chapter called *The Knowledge Economy*. He points out that competitive advantage has transformed as a concept with the emergence of advanced information technology. Already in 1981, he stated that ‘’economic prosperity rests upon knowledge’’ and it’s fair to say that today, 33 years later, that history shows that he was accurate in his statement (Teece, 2002). The complex business issues that have been solved through big data analytics is testament to the importance of using technology for innovation and innovation for business gains.

Steve Ellis, explained in 2005 that knowledge-based working is when intellectual assets is used collectively to create a superior ability to meet market challenges before the competition. The emphasis is to move away from tacit-knowledge which is knowledge only held by an individual for individual task completion (Ellis, 2005). It’s been proven that explicit knowledge benefits organizations as it leaves originations less vulnerable to staff turnover and change management issues when intelligence is widely accessible (Dalkir, 2005). Knowledge-based working requires a change of approach to organizational operations. This change can be supported only through faster, deeper and more accurate intelligence gathering, which is something that big data analytics can provide. With the use of big data, knowledge-based-workings can be applied optimally.

Organizations seek predictability for stability and sustainability. The ability to see ahead provides security and the ability to introduce initiatives that can help avoid risks as well as initiatives that leverages on opportunities that change can bring. The demands for insight due to web traffic, growth of email massaging, social media content and information from connected machines with sensors such as GPS products, mobile devices, and shop purchasing devices drives data growth. The constant flow of large volumes of data drives organization’s to invest in new data management tools to be able to capture, store and gain business intelligence through analytics of larger sets of data.

Big data has many known use-cases. Most commonly it's used by government's or associated agencies to provide things like national statistics, weather forecasting, traffic control, fraud prevention, disaster prevention, finance management, managing areas around national education, national security, health care and many other use cases in the private sector such as retail, banking, manufacturing, wholesale, distribution, logistics, communications industry, and utilities. In short, there's a use case in most sectors (Yan, 2013).

Gartner Inc, estimated in 2012, that organizations would spend 28 billion USD on big data that year and that the number would rise to 56 billion USD by 2016. Market revenues are projected to be 47.5 billion USD by 2017. According to Gartner, a general profile of a big data user is an organization with a database larger than 1.5TB and that has a data growth rate of 20% per year.

Research McKinsey estimate a 40% projected data growth generated by the US year on year. Adoption of big data and improved data management could help for example reduce US health care expenditure by 8% and retailers can expect a possible 60% increase in operational margin if they adopt big data (McKinsey&Company,2011).

Peter Sondergaard, Senior Vice President and Global Head of Research at research company Gartner, stated in 2013 that "By 2015, 4.4 million IT jobs globally will be created to support big data'' (Gartner, 2012). Another demonstration of the growing interest can be found in statistics provided by Google that shows a 90% increase of big data searches has been recorded between 2011 and March 2014 (Google, 2014).

Use case- US Government

The US Government formed a new organization called the *Big Data Senior Steering Group* in 2010, consisting of 17 agencies to support research and development. Two years later the *Big Data Research and Development Initiative* was provided a 200 million USD budget to accelerate the technical science and engineering effort of big data for the purpose of improved national security.

In 2013 the *Big Data Community of Practice* was founded, which is a collaboration between the US government and big data communities. This was followed by the *Big Data Symposium* which was founded to promote big data awareness. Furthermore, significant investments have been made to support higher education programs to train data scientist to cover the existing knowledge gap around big data (The White House, 2012).

An example of benefits that have been seen is the case of the Internal Revenue Service in the US. They have documented a decrease of time spent on loading tax returns from over 4 months in 2005, to 10 hours through big data initiatives in 2012 (Butler.J, 2012).

Big data from a technical perspective

To understand big data it's helpful to understand data from a corporate management point of view and associated concerns. When decisions makers review analytics one of the main tasks is also to review the accuracy, completeness, validity and consistency of the data (Chaffey,Wood, 2005). One big threat to any analytics initiative is the lack of access to usable data. Poor data is a threat to companies and inhibits organizations from leveraging on analytics. Big data depends on high quality data but can also be used to spot data discrepancies.

Historically businesses have been strained by lack of processing power, but this has changed today due to decreasing costs for hardware and processing. The new strain is growing data volumes that hamper efficient data management. To be able to leverage on big data organization's need to prepare systems to be able to take full advantage on the opportunity to big data brings. Data needs to be prepared and existing IT systems need to have the capability to not just handle the data volume but also maintain the running of business applications. Organizations worry about latency, faultiness, lack of atomicity, consistency, isolation issues, durability (ACID) security and access to skilled staff that can manage the data and systems (Yan, 2013)

Data management issues

One common cause for IT issues is architectural drift, which is when implementation of software deviates from the original architectural plan over time and causes complexity and confusion. Experts point to that complexity can be caused by lack of synchronization, standardization and awareness as new code is being developed and data models change. Furthermore, architects are often reluctant to revisiting problematic areas due to time constraints, sometimes lack of skills and demotivation.

For data to be used efficiently the data modelling structure needs to consists of logic that determines rules for the data. Experts talk about entity, attribute and relationship. Data modelling enables identification of relationships between data and the model defines the logic behind the relationship and the processing of the data in a database. One example is data modelling for data storage. The model will determine which data will be stored, how it will be stored and how it can be accessed (Rainer, Turban, 2009).

Data is often managed at different stages and often in several places as it can be scattered across an organization and be managed by multiple individuals, leaving room for compromise of data integrity. One big issue is data decay, which is an expression used to explain data change. An example could be a change of a customer surname, change of address or an update of product pricing.

There are multiple conditions that can affect data management such as poorly written software, software compatibility, hardware failures can be caused by insufficient storage space affecting the running of software. Data can be affected by operator errors caused by for example the wrong data being entered or a script could be instructing the computer to do the wrong thing, which can affect the mainframe and mini computers which can cause issues with batch jobs. Multiple issues can cause down-time and disruption to business operations.

Hardware also affects data management. A computer can fail to run a back-up or install new software or struggle with multiple tasks such as processing real-time data at the same time as restoring files to a database. This might affect multiple simultaneous tasks causing confusion and time loss (Chartered Institute of Bankers, 1996)

Big data discussions are directly linked to data management and database discussions. Databases enable management of business transactions from business applications. Using DBMS enables data redundancy which helps with avoidance of losing data through data storage at different locations. It also helps with data isolation as a precaution to enable assignment of access rights for security. By using one database inconsistencies can be avoided. It can act as a point of one truth rather than having different sets of the same data that can be subject to discrepancies.

Previously, transactional data was the most common data and due to its simple structured format it could with ease be stored into a row or column of a relational database. However with the introduction of large volumes of web data which is often unstructured or semi-structured, traditional relational databases do no longer suffice to manage the data. The data can no longer be organized in columns and rows and the volume adds additional strain on traditional database technologies. Big data enables management of all types of data formats, including images and video, which makes it suitable for modern business analytics (Harry, 2001)

1.1 Data structures

One issue associated with big data is management of different data structures. The lack of standardization makes data difficult to process. As mentioned in the section above, the introduction of large volumes of data makes it difficult for traditional DBMS to organize the data in columns and rows. There are several types of data formats such as structured transactional data, loosely structured data as found in social media feeds, complex data as what can be found in web server log files and unstructured data such as audio and video files. The data mix can also be referred to as an enterprise mashup, basically integration of heterogeneous digital data and applications from multiple sources, used for business purposes.

The difficulty with web data is that the data was inserted without following rules, like for example the rules that a database administrator would follow as standard. The data can generally be divided into three categories; Structured, unstructured and semi structured. Structured data is often described as data that is modelled in a format that makes it easy to

shape and manage. The reason it may be easier to manage is because formal data modelling technique has been applied that are considered standard. A great example of a solution based on structured data is an excel spread sheet.

The opposite to structured data is unstructured data which is difficult to define as it is both language based and non-language based like for example pictures, audio and video. Popular websites like Twitter, Amazon and Facebook contain a high volume of unstructured data which can make reporting and analysis difficult due to the mixture of data and difficulty to translate image and video for example into text language, to make the items easier to search for (Laudon, Laudon, 2014)

Semi structured data has a combination of structured and unstructured data. Semi-structured data is when the data does not fit into fixed fields but do contain some sort of identifier, tag or markers that give it a unique identity. In a scenario of building a database with this sort of data set, part of it would be easier to manage than other parts. The online companies mentioned above along with the likes of Linkedin, Google, and Yahoo.com will all have databases containing this sort of data. XML and HTML tagged text is an example of semi-structured data (McKinsey, 2011)

To give an example of the importance of data structure the following scenario can be considered. If UK retailer Tesco's would want to release a new product on their site, they would decide on the order and structure of associated data for that product in advance to enable quick search and reporting relating to that product. The product would have attributes like example color, size, salt level and price, inserted in a structure, order and format that make associated data easier to manage than data from a public online blog input for example. The ability to identify the individual product is critical to being able to analyze sales and marketing associated with the product. If the product is not searchable, opportunities can be missed (Laudon, Laudon, 2014)

1.2 Data warehouse and data mart

Traditional transactional DBMS is the core of big data but it does not allow retrieval of optimal analytics in the same way as data warehousing. Data warehouses have been used for over 20 years to consolidate data from business applications into a single depository for analytical purposes. Many businesses use it as the source of truth for validation and data quality management. Data warehouses provide ad-hoc and standardized query tools, analytical tools and graphical reporting capability.

Data warehouse technologies started to be widely promoted in the 1990, a little while before ERP systems were introduced. Just like today, the associated data consisted of feeds from a transactional database. The addition today is that the data can also be feed from an analytical database and faster than ever before using in-memory transactional capability. Previously, data warehousing has not been used for daily transactions. This is shifting with the introduction of real-time data processing.

Data warehouse management requires significant upfront development and effort to be able to provide value. A common implementation project scenario would be a follows:

- *Create a business model of the data*
- *Create logical data definition (schema)*
- *Create the physical database design*
- *Create create-transform-load (ETL) process to clean, validate and integrate the data*
- *Load data it into the data warehouse*
- *Ensure format conforms to the model created*
- *Create business views for data reporting*

(Winter Corporation, 2013)

One of the most common data sources for data warehouses is ERP data. The ERP systems feeds data warehouses and vice versa. Many organization's use Enterprise Resource Planning solutions (ERP) to consolidate business applications onto one platform. An ERP system provides the ability to automate a whole business operation and retrieve reports for business strategy. The system also provides a ready-made IT architecture and is therefore very relevant to big data.

The most important data that makes up a data warehouse is Meta data, which can be described as data about the data. It provides information about all the components that contributes to the data, relationships, ownership, source and information about whom can access the data. Meta data is critical as it gives the data meaning and relevance, without it, a data warehouse is not of value. Data warehouse data needs to be readable and accurate to be of useful and in particular in relation to big data analytics as it would defeat the purpose of the use case if the information provided was questionable (McNurlin, Sprague, 2006)

Users use ETL (extract, transform, and load) tools for data uploading. This uploading process or the opposite, data extraction can be a tedious process. However the biggest associated issues around data warehouses are search times due to the query having to search across a larger data set. Sometimes organizations want to have segmented data warehouses for example to enable faster search, minimizing data access, and to separate divisions or areas of interest. In those cases data marts can be used. It is a subset of a data warehouse, stored on a separate database. The main issue around data marts is to ensure that the Meta data is unified with the Meta data in the data warehouse so that all the data uses the same definition otherwise there will be inconsistencies in the information gathered (Hsu, 2013).

As the data volume grows and becomes big data and new tools are introduced to manage the data, data warehousing remains part of the suite of tools used for processing of big data analytics.

Big data management tools

The process flow of big data analytics is data aggregation, data analysis, data visualization and then data storage. The current situation is such that there is not a package or single solution that tends to suffice to fulfill all requirements and therefore organizations often use solutions from multiple vendors to manage big data. This can be costly, especially is the decision makers don't have enough insight about cost saving options.

The key tools needed to manage big data apart from a data warehouse are tools that enable semi-structured and unstructured data management and that can support huge data volumes simultaneously. The main technologies that need consideration when it comes to big data in comparison to traditional DBMS are storage, computing processing capability and analytical tools.

The most critical element of a big data system is data processing capability. This can be helped using a distributed system. A distributed system is when multiple computers communicate through a network which allows division of tasks across multiple computers which gives superior performance at a lower cos. This is because lower end clustered computers, can be cheaper than one more powerful computer. Furthermore, distributed systems allow scalability through additional nodes in contrast to replacement of central computer which would be necessary for expansion in the scenario where only one computer is used. This technology is used to enable cloud computing.

Big data analytics tools and Hadoop

To enable advanced statistics, big data adopters use a programming languages pbdR and R. for development of statistical software. The R language is standard amongst statisticians and amongst developers for development of statistical software. Another program used is Cassandra which is an open source DBMS that is designed to manage large data sets on a distributed system. Apache Software foundation is currently managing the project however it was originally developed by Facebook.

Most important of all big data analytics tools is Hadoop. Yahoo.com originally developed Hadoop but it is today managed by Apache Software foundation. Hadoop has no proprietary predecessor and has been developed through contributions in the open-source community. The software enables simultaneous processing of huge data volumes across multiple computers by creating sub sets that are distributes across thousands of computer processing nodes and then aggregates the data into smaller data sets that are easier to manage and use for analytics.

Hadoop is written in Java and built on four modules. It's designed to be able to process data sets across multiple clusters using simple programming models. It can scale up to thousands of servers, each offering local computation and storage, enabling users to not have to rely extensively on hardware for high-availability. The software can library itself, detect and handle failures at application layer which means that there is a backup for clusters (McKinsey&Company,2011).

Through Hadoop data processing, semi-structured and unstructured data is converted into structured data that can be read in different format depending on the analytics solution used (Yan, 2013). Each Hadoop cluster has a special Hadoop file system. A central master node spreads the data across each machine in a file structure. It uses a hash algorithm to cluster data with similarity or affinity and all data has a three-fold failover plan to ensure processing is not disrupted in case the hardware fails (Marakas, O'Brien, 2013)

The Hadoop system captures data from different sources, stores it, cleanses it, distributes it, indexes, transforms it, makes it available for search, analyses it and enables a user to visualize it. When unstructured and semi-structured data gets transformed into structured data, it's easier to consume. Imagine going through millions of online video and images in an effort to uncover illegal content, such as inappropriately violent content or child pornography and be able to find the content automatically rather than manually. Hadoop enables ground breaking ability to make more sense out of content.

Hadoop consist of several services: the Hadoop Distributed File System (HDFS), MapReduce and HBase. HDFS is used for data storage and interconnects the file systems on numerous of nodes in a Hadoop cluster to them turn them into one larger file system. MapReduce enables advanced parallel data processing and was inspired by Google File System and Google MapReduce system which breaks down the processing and assigns work to various nodes in a cluster. HBase is Hadoops non-relational database which provides access to the data stored in HDFS and it's also used as a transactional platform on which real-time applications can sit.

From a cost perspective, Hadoop is favorable. Its open source and runs on clusters of most cheap servers and processors can be added and removed if needed. However one area that can be costly is the tools used for inserting and extracting data to enable analytics within Hadoop (Laudon, Laudon, 2014)

As mentioned above, a Hadoop license is free of cost and only requires hardware for the Hadoop clusters. The administrator only needs to install HDFS and MapReduce, transfer data into a cluster and begin processing the data in the set up analytic environment. The area that can be problematic is the configuration and implementation of the cluster. This can be costly if an organization does not have in-house skills.

Technical limitations relating to Hadoop

As Hadoop is not a data management platform and does not include data schema layer, indexing and query optimizing, making Hadoop analytics management very manual. If any data in a Hadoop file changes, the entire file needs to be rewritten as HDFS does not change files once they are written. To ease this process a language called HiveQL can be used which allows writing simple queries without programming however HiveQL is not effective for complex queries and has less functionality than an advanced data warehouse platform (Winter Corporation, 2013)

As mentioned previously in this report, the critical element that enables big data is processing capability. One major difference between traditional database management and new techniques is that traditionally structured data is often used after collection. Another expression for this is that the data aggregates first before being used. New technology enables management of data that has not yet been aggregated, which is what's often referred to as in-memory technology, which enables real-time data capturing, reporting and analytics. This capability is critical to big data analytics processing.

In-memory computing

In memory computing relies mainly on RAM memory which is the computers main memory for data storage. When data is stored in the computers main memory, bottlenecks from reading and retrieving data can be eliminated. Traditional database technologies mainly use disk storage which makes query response times longer. With in-memory computing data warehouses and data marts can exist in memory, enabling reporting without data aggregation. Complex calculations can be done in seconds rather than days. Commercial offerings for such technology are provided by SAP with their solution HANA and Oracle with their solution Exalytics.

To leverage fully on in-memory capability an advanced analytics platform can be used that uses both relational and non-relational technology, like NoSQL for analyzing large data sets. IBM offers such a solution called Netezza and is competing with Oracle Exadata. In-memory solutions are also often provided with supportive hardware technology that enables optimized processing of transactions, queries and analytics (Laudon, Laudon, 2014)

In-memory capability is enabled through Online Analytical Processing (OLAP). It enables multidimensional data analysis which provides the ability to view data in different ways. OLAP analytics enables the difference between viewing yearly profit figures and the capability of compare results with projected profits across divisions and other segmentations. The different dimensions represented could be country, city, type of service, product, customer characteristics or time frames. In short, OLAP enables comparative views and relational views which provide much more depth to an analysis (Marakas, O'Brien, 2013).

There are certain data that cannot be analyzed using OLAP. For example patterns, relationships in bug databases and predictive analytics. Data mining can provide this information as it uses, associations, sequences, classification, clusters and forecast to conclude a query result.

Data mining solutions can help with discovery of previously unidentified groups through patterns or affinity and associations that symbolize occurrences linked to an event. A scenario could be; when customers at a supermarket buy bread, they are 50% more likely to also buy cheese and tend to spend no longer than 12 minutes in the retail store.

A sequence represents the time frames associated with a particular event. For example; when customers have bought cheese, they are likely to return within one week to buy bread again. Classifications can for example describe the characteristics of the person buying, ex; the person buying cheese the first time is likely to be a mother but the person buying the bread the second time is likely to be a father.

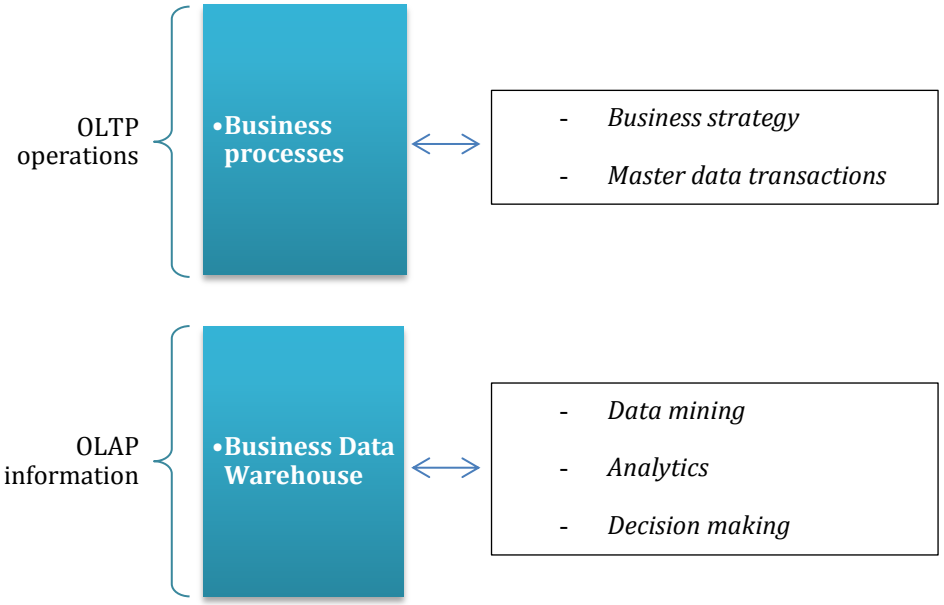
Organizations use classification for example for targeted marketing campaigns. Another way of discovering groups is clusters. One cluster can for example be; all identified customers that are mothers in Dublin (Jessup, Valacich, 2003).

Data mining is often used for forecasting and all the different dimensions helps with determining probability and the estimated level of accuracy. The end goal is predictability and the ability to foresee events, change and impact. Data mining becomes difficult when dealing with unstructured data which today represents 80% of organizational information globally. Blogs, email, social media content, call center transcripts, videos, images all required text mining solutions for analysis. The text mining solutions can extract elements from big data sets that are unstructured for pattern analysis, analysis of affiliation and collate the information into a readable report.

Another big data tools that leverage on text mining is sentiment analysis tools which provide insight about what's being said about for example a brand or about the government across multiple channels for example online. Analysis of web information can be done with web mining. Search services like Google Trends and Google Insight use text mining, content mining, structure mining and usage mining to measure popularity levels of words and phrases.

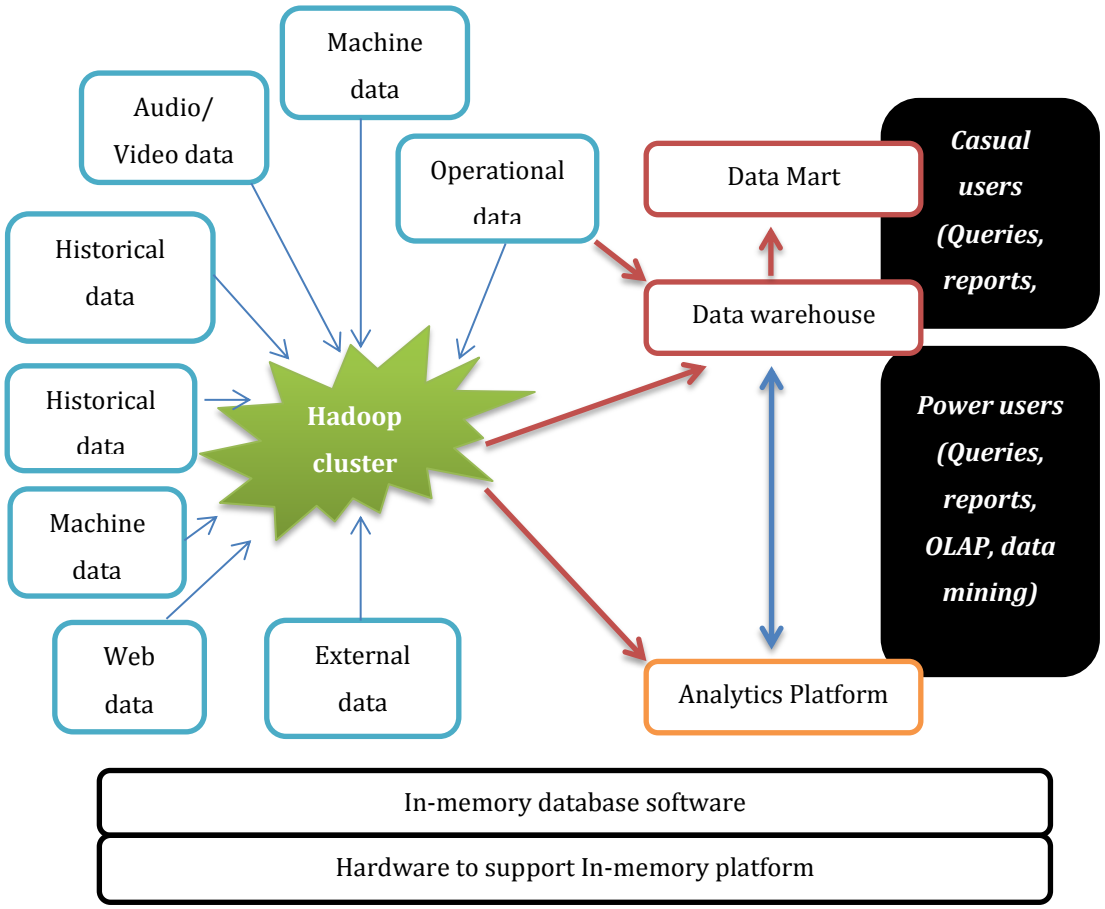
You may have come across sentiments poll machines at air ports or in super markets but also at industry events and you might have participated in surveys. All that data is sentiment analysis relevant information. (Laudon, Laudon, 2014). See example of the difference between OLTP and OLAP below.

1.3 *Table 1. View of the difference between OLTP and OLAP*



OLAP should be separated from Online Transfer Process OLTP which is the business process operation. It can also be described as the process before data aggregates into a data warehouse. When for example online data is inserted in a web platform it is OLTP process until the data is compiled in the data warehouse. The process of managing the data in the data warehouse is OLAP. See view of a modern data warehouse using in-memory technology below.

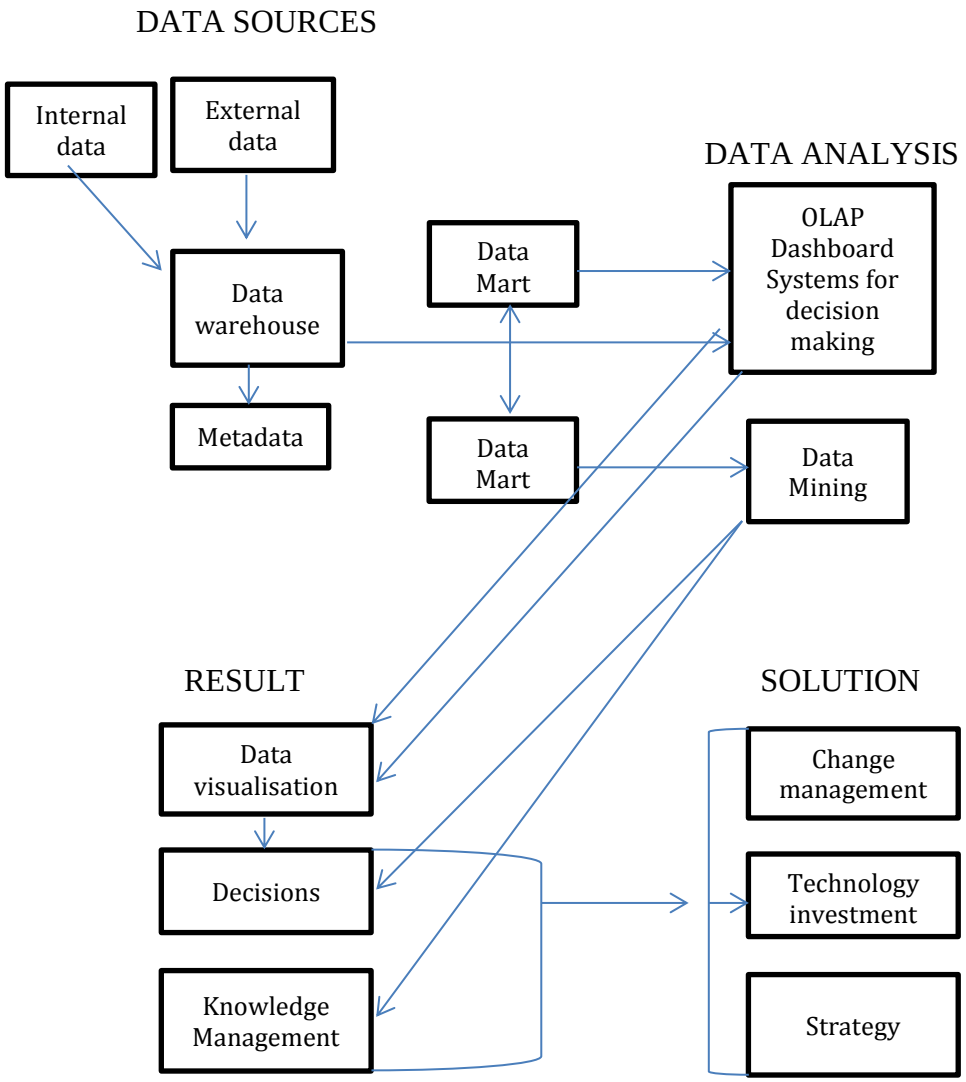
1.4 Table 2. View of a modern data warehouse using big data and in-memory technology



The image above shows the different components of a data warehouse and how different users can have be assigned user rights. A casual user may not need advanced analytics and may only want to access a selection of data in a data mart, whilst a power user may want more advanced features.

The previous table shows data warehousing using in-memory technology and below is a more holistic view of a simplistic data model of an IT infrastructure that shows data flow from insert to data being translated into actionable information.

1.5 Table 3. Data life cycle- An example of a basic data model



Differences between big data analytics and traditional DBMS

The main benefit of big data is that it eliminates threats of data decay. As data changes over time, it only adds to the depth of the insight that can be gained through big data analytics. To give an example, if manufacturer changes product attributes, or if customers change their direct debit details it only adds to the depths of the analytics gained. As the information accumulates, organizations can see trends, patterns and impacts of changes over time.

Big data technologies enable extraction from multiple operational systems and processing of previously hard-to-manage data formats and volumes. Hadoop clusters the big data for data warehousing use that can in cases be segmented into data marts. If required, the data can be processed using an advanced analytics platform that enables real-time reporting before the data is aggregated. The out-put is presented in reports and dashboards (Laudon, Laudon, 2014)

Big data analytics uses genetic algorithms which help with solving non-linear problem solving. It improves the data processing. McKinsey describes it as follows:

'' A technique used for optimization that is inspired by the process of natural evolution or "survival of the fittest." In this technique, potential solutions are encoded as "chromosomes" that can combine and mutate. These individual chromosomes are selected for survival within a modeled "environment" that determines the fitness or performance of each individual in the population. Often described as a type of "evolutionary algorithm," these algorithms are well-suited for solving nonlinear problems''

(McKinsey&Company,2011)

(See appendix A for examples of big data analysis methods)

Big data entails two main platform architectures, the data warehouse and the analytics solution, Hadoop. Hadoop is significantly cheaper than a data warehouse platform, the on-going queries and analytics is more expensive using Hadoop than traditional data warehousing. The main big data cost is similarly to traditional DBMS the appliance and professional services. See table 4 below for a comparison of big data analytics costs compared to traditional data warehousing.

1.6 Table 4: View of cost difference between data warehousing costs in comparison to Hadoop

	Data Warehouse Platform	Hadoop
Volume of Data	500 TB	500 TB
System Cost[6]	$44.6	$1.4
Initial Acquisition Cost	$10.8[7]	$0.2[8]
Upgrades at 26% CAGR	$16.4	$0.3
Maintenance/Support[9]	$15.9	$0.2
Power/Space/Cooling	$1.5	$0.6
Admin	$7.7	$8.5
Application Development	$16.5	$36
ETL	$18.4	$0
Complex Queries	$88.7	$475
Analysis	$88.7	$219
Total Cost of Data (TCOD)	**$265 million**	**$740 million**

(Winter Corporation,2013, p7).

See table 5 below for a view of differences between traditional DBMS characteristic and big data characteristics.

1.7 Table 5. Major differences between traditional database characteristics and big data characteristics

	Traditional DBMS characteristics	Big data database characteristics
Data characteristics	• Weak in handling non-structured data • Does not learn from user access behavior	• Real time live data • Environment supports all types of data and from all types of sources • Appropriate for handling petabytes and exabytes of data • Learns from user access behavior
Considerations for analytics	• Is appropriate for analysis of data containing information that will answer to information gaps that are known • Stable data structure • Passive user experience- the user simply retrieves information • Focus on attribute search • Data management simplified • Historical data management 2 dimensional • Limited inter organizational data access • Analytical capability expectations are minimal	• A lot of parallel processing often strains supporting systems • Can be truly ground-breaking for organization's as previously completely unknown gaps of information can be revealed randomly rather than just providing information about what is known- is not know • Emergent data structure • Active user experience- the system may initiate discourse that may need attention • Focuses on historical pattern, trend, multi-dimensional search • Rich multi-channel knowledge access • Real-time analytics technologies plays a vital role • Analytical capabilities expectations may be too high
Relevant technologies	• Implementation straight forward • SQL is most common language • Relational database function model • Not open source • Analytics is done on batch jobs containing aggregated data which is historical data rather than real-time data	• Implementation more complex due to lack of industry standards and direction • No SQL but near on SQL compliant • Hadoop framework • Open Source is appropriate • Stream processing capability is often relevant

(Galliers, Leidner, 2003), (Yan, 2013)

Big data costs- findings from primary and secondary data

One of the biggest obstacles for organizations that want to invest in big data analytics is cost. The research findings in the primary data in relation to big data costs, was similar to the information found in secondary data. Although big data analytics is based on traditional data warehouse set up, it can require additional investments to enable processing, storage and analytics capability beyond the capacity of the traditional data warehouse.

Majority of the experts that took part in the survey support the notion of additional investments in IT being necessary for big data projects. One participant shared that 50TB of additional storage had been bought for £450,000. Another stated that 64 cores of processing were added at a cost of £150,000. The cost of time spent on development of analytics applications was stated to be £100,000 in one case and £170,000 in another case. The time spent on development of analytics applications was between 1 week and 6 months.

26% of the participants choose to not share information about the amount of storage that had been required for the big data project. 13% of the participants stated that no additional storage needed to be bought for the big data project and 61% responded that there had been a storage investment. The storage volumes that were stated ranged from 250 gigabytes to 50 terabytes and it was stated that the storage was needed for disaster recovery and data persistence as well as for the production system.

One expert stated that although additional hardware was needed the extra cost was counteracted by ease of maintenance which had led to a reduction of headcount in the IT department. Another participant advised to not refrain from investing in big data due to high initial costs and that big data did not only bring increased speed but also simplification the of the IT environment, which was the biggest benefit and brought a huge cost saving.

39% of the participants could not respond to whether any additional processing capability had been needed for the big data project however 17% responded that it had not been required. One participant shared that less resources had been required than before and that they managed to get almost double the processing speed compared to before the big data solution roll out. 44% of the participants stated that additional processing capability had been needed for the big data project.

In one case, there had been an investment in 80 cores of SAP Sybase ASE a traditional, relational database technology. Participants stated that additional 8X4 cores were added; another stated that 64 cores were added. Another participant stated that an investment in SAP HANA had been made with 40 cores. One participant shared that an investment in in-memory capability for improved processing and real-time insight was made.

Data integration costs and time spend on implementation were stated to be between £20,000 and £250,000 and took between 3 weeks and 7 months. One expert stated that although the project took time and had an upfront cost associated with it, the total cost of ownership (TCO) was reduced.

The estimated time spent on runtime disaster recovery of big data management system was stated by the specialists to be between 2 hours every 6 months and 2 weeks. One expert advice that a stress test can take up to 2 days and that there will be a need to short-list the high resource processes and use that as a test.

Runtime analysis and regression of big data management systems can take between 4 hours every 6 months and 1 month. The expert's experiences were that migration and testing takes between 10 hours every 6 months and 17 months. It was also advised to assign appropriately enough of time for migration and testing. The main objective of the migration and testing is to ensure smooth running of analytic queries.

Archiving and recovery of big data system were stated by the experts to take between 2-4 weeks. Patching and upgrades were stated to take between 4 days and 4 weeks and it is also advised to implement a big data project in phases, as it saves time & resource. Doing it in stages will also ease monitoring of return on investment.

The big data investment numbers varied from £30,000 to £2,7 million showing the great variation between big data investment and also hopefully giving confidence to those with pre assumptions about big data costs being high.

The software investments that the experts shared were between £400,000 and £500,000 and included a cloud solution, Hadoop, big data analytics solution from Oracle, SAP HANA, Teradata, new BI front end, NoSql, and new ERP solutions.

(See appendix B for survey results)

Many studies have been carried out on cost analysis of big data analysis comparing it to traditional DBMS systems. See the table below for a summary of the findings from secondary data. Note that cost for traditional DBMS is included as it's the foundation for big data management.

1.8 Table 6: Estimated project cost for 40TB data warehouse system –big data investment

Estimated project cost for 40TB data warehouse system -big data project

Hardware/Software costs	Support/Maintenance cost
Administration costs (year 1): **Costs= USD 375,000**	**Data growth** Estimated 26% annual growth
Development costs (year 1): **Costs= USD 700,000**	**Line of Java/Mapreduce code, larger (100K lines)** Costs= USD 24 /line
Professional services costs/hourly rate: **Cost= USD 150** **Estimated project time= 1,500 hours**	**Lines of JAVA/MapReduce code for proxy for acquired and developed applications** Cost= USD 300,000
Hadoop developer/administrator **Cost= USD 150,000** **(Only 1 staff required normally)**	**Line of JAVA/SQL code, SQL developer (100,000 lines)** Cost= USD 22
Database server **Production:** **8 CPU= 8 XUSD 25,000** **64GB memory** **50TB storage= 50X USD 8,000** **Non-production:** **2 CPU= 2X USD 15,000** **32GB memory** **3X20TB storage= 20X USD 5,000**	**Maintenance/support costs:** Costs= USD 217,800 /annually
Application platform **Ex. SAP Netweaver** **Production:** **4 CPU= 4X USD 25,000** **64GB memory** **5TB storage= 5X USD 8,000** **Non-production:** **4 CPU= 2X USD 15,000** **32GB memory** **2X 5TB storage= 5X USD 5,000 (2X)**	**Maintenance/support costs:** Costs= USD 55,000 /annually
Operational reporting server (40TB) **Production:** **4 CPU= 4X USD 25,000** **64GB memory** **40TB storage= 40X USD 8,000**	**Maintenance/support costs:** Costs= USD 92,400 /annually
Hardware for big data	**Hardware maintenance cost:**

Production: **1 CPU/8 cores= USD 25,000** **Storage/1 TB= USD 8,000** **Non-production:** **1 CPU/8 cores= USD 15,000** **Storage/ 1TB= USD 5,000**	22% of hardware cost **Software maintenance cost:** 22% of software cost
Hardware for non- big data relevant system **Production:** **2 CPU/16 cores= 2X USD 25,000** **1 TB memory= USD 8,000** **8TB storage** **Non-production:** **Disaster recovery (same as production)** **1 development** **QA testing (2CPU)= 2X USD 15,000** **Storage 8TB= 8X USD 5,000**	**Hardware maintenance cost:** Cost= USD 65,650 **Vendors discount for data warehouse appliance:** 40%
Data warehouse appliance space, cooling and power supply per TB/year **Cost= USD 291**	
Analytics appliance **Ex. SAP HANA** **HANA appliance= 9X USD 150,000** **9X 1TB scale out (based on 2Tb compressed - 40TB system)** **4X production** **1 standby high-availability** **Secondary- 4X disaster recovery** **Test and development**	**Hardware maintenance cost:** Cost= USD 297,700
Analytics Platform **Ex: SAP Netweaver** **Production:** **4 CPU= 2X USD 25,000** **64GB memory** **5TB storage= 5X USD 8,000** **Non-Production:** **2 CPU= 2X USD 15,000** **32GB memory** **2X 5TB storage= 2X USD 5,000 (2X)**	**Maintenance cost:** Cost= USD 55,000
Data compression **4 X**	
Big data analytics software **Analytics software= Percentage of ERP cost**	**Maintenance cost:** Cost= USD 35,200

<u>Estimates for big data analytics platform solutions implementation (such as SAP HANA) year 1</u> **Internal implementation cost= USD 108,173** **Professional implementation services- costs year 1= USD 157,500** **Administration resources= USD 600,000** **Development resources= USD 1,120,000**	
<u>Hadoop</u> **1 TB= USD 1,000** **(compression of X2 applies)** Hadoop cluster space/power/cooling/data refining **1TB/annum= USD 301**	**Vendor discounts** Commonly vendors provide 30% discount for data warehouse platforms and 10% discounts for Hadoop cluster **Maintenance support** Commonly maintenance and support costs are 10% of acquisition cost for Hadoop and 20% for data warehouse per annum. Furthermore general support is estimated USD 100/TB/year

(Forrester Research, 2014), (Winter Corporation,2013

Research Objective

The objective of this research was to collate ideas and advice from big data professionals to support spread of information to help big data explorers and decision makers. The aim is to encourage wider adoption. Analysts and education bodies may also find the research findings interesting. I hope that Dublin Business School will be able to use this research in initiatives relating to IT and business related education

As mentioned in the chapter *Background*, the initial idea was to set up two like-for like-DBMS environments, one traditional and one big data environment and work-backwards to see what can be removed or replaced to create the most commercially and technically favorable big data set-up. However multiple cost analyses have already been concluded by multiple other researchers. Some findings I've shared in the chapter *Big data costs- findings from primary and secondary data*.

The conclusion is that it is near-on impossible to create a like for like environment, comparing traditional data warehousing to big data and results from such experiment are questionable. However, a cost analysis can be a valid tool to use for decision makers to be able to draw assumptions about the estimated costs, resources and time that needs to be allocated for a big data project.

The direction of this research was changed as an opportunity was presented to gain valuable insight big data experienced IT professionals who expressed a willingness to share their experiences and insight of big data projects for the purpose of information sharing and to help big data explorers and decision makers. The direction is unique from the perspective of a quantities research as qualitative research is more often used for more in depth big data projects information gathering.

Research methodology

The informatics philosophy used for this research is to use methodology, as the purpose is to find valid arguments to support the ideas gathered and to provide ideas that can be used practically. Theories such as epistemology or ontology for example do not support the practical approach that many organizations apply in relation to IT and are better suited for Social research.

As big data is a physical thing, a combination of technologies which are physically tangible, positivism has been used for this research as it commonly is used for other IT related research. Big data in itself can be considered relevant for an interpretivism stand point as the concept in itself has not been coined yet. The same goes for the benefits of big data. They can also be considered to be more appropriately approached from an interpretivism perspective; however the objective of this research is practical appliance of theory, which makes positivism more appropriate.

The applied thinking for this research is practical to be of use to readers that are looking to apply a big data strategy. Part of the purpose of this study, is to encourage wider adoption of big data analytics and therefore an emphasis has been on simple language and avoidance of industry jargon (Beynon-Davies, 2002)

The research design, commonly referred to as the blueprint for the research process, functions as a way to combine the research objectives and the methodology chosen to fulfil them. A quantitative survey research design has been chosen for this study in order to gain a deeper understanding of the existing skills gaps amongst big data related professionals. After an analysis of the secondary data the research objectives were formulated and a qualitative form of primary data was deemed to be the most appropriate to fulfil them.

As mentioned in the previous chapter, *Research objective*, the vast majority of previous big data studies follow a qualitative route. A verifications paradigm was chosen to get a wider overview. Besides, it is common to use quantitative research for topics related to IT amongst research firms like Gartner and Forrester.

The objective of the research is to conclude possibilities rather than prove or disprove a theory. A hypothesis is therefore inappropriate for this research. The research query is ‘‘Opportunities to manage big data efficiently and effectively’’. The title indicates that there is an assumption that big data is not currently managed as efficiently and effectively as it can be. This is not a theory that needs disproving as the premature stage of big data analytics in itself indicates that there could be additional possibilities that are not yet explored. Other experts also support this theory, as can be seen in the secondary data.

This research does not aim to be empirical or conceptual but theoretical as the purpose of this research is to gain insight from the big data community. Neither is this research longitudinal, due to the restrictive project time.

Data collection

The data collected for this research consists of results from surveys and interviews with big data professionals. A cross sectional survey was developed as the big data community is not vast. The survey was designed using interpretivism perspective which is unusual for a quantitative study, however, it was appropriate as this research is based on individual's experiences. Focus groups are more commonly looked at from a subjectivism and interpretivist perspective. IT is more often looked at from a scientific objectivist however this research focuses subjectivist perspectives as the research data focuses on observation, interpretation and sharing of ideas (The Marketing Review, 2004).

As mentioned above, the initial idea was to carry out a laboratory experiment or action research, using positivist philosophy. The idea was to build a like for like DBMS system, one that supports big data and one traditional one and then work backwards to review the necessity of components to see how costs for big data can be reduced. Proof-of-concepts and simulations are known for not being representative of live production systems. I used secondary data to conclude this fact and therefore started focusing on an area where I could add more value and provide a more unique contribution.

Another alternative was to carry out a field experiment using positivist philosophy, however due to confidentially issues and data protection laws it was difficult to get access to decision makers that were willing to share confidential information.

An interesting experience would have been to sit down with big data administrators to experience a day-in-the-life of a big data IT professional. A subject argumentative research method could have been a hindrance to finding non-bias and objective insight from big data professionals. Furthermore, working with just one organization's can make the findings questionable as the participants may be hesitant to express opinions and ideas freely as they're representing their organizations.

An idea was to suggest a case-study to selected organizations for possible mutual gains, but the proposition had little appeal amongst those who were approached. Some expressed that their big data practice was not mature enough and others saw a competitive threat in sharing

too much information about their internal practices. Similar arguments were used to reject the opportunity to create a documentary analysis or an observation study in collaboration. A case study would add interpretivist philosophy to this study however, it could also been limiting to capture fewer views which may not be representative of the wider big data practicing community (International Journal of Social research Methodology, 2005).

Literary review

People who work with big data management are more likely to have an understanding of the skills, experience and natural talent that can add value for organizations in the area as well as associated heuristics (Chaffey,Wood, 2005). The survey that was carried out was designed to gather information that I had failed to find in my secondary data studies as well as to validate previous findings from secondary data.

One issue with the survey was the volume chapter *Big data issues*, there is currently a skills gap in the market and a shortage of big data specialists. The limited number of participants makes the findings questionable as a generalization cannot be made based on a small number of participants.

One of the main issues of big data is the lack of academic research that has been carried out in the topic. Most of the existing material is online. The lack of resources supports the quantitative research method used for this study and the use of triangulation by combining secondary and primary data (Beynon-Davies, 2002).

Research survey

Originally 54 participated in the survey named ‘’Opportunities to manage big data more cost efficiently and more effectively’’. 23 survey responses were selected after review due to credibility issues. The survey solution used was Survey Planet- https://www.surveyplanet.com. It worked very well as it was inexpensive and enables participants to answer in essay style as well as other styles. The approach of using open questions enabled the participants to provide rich information.

This research questions focuses on a realist, scientific approach. An inductive perspective is inappropriate as this research has a specific objective and needs to maintain focused on the subject matter (Saunders, et al, 2011). A risk relating to this research is bias opinions. A possible scenario could be if a participant had ulterior motive to promote a specific technology. After careful consideration the following questions were chosen for the research survey:

1.9 Table 7: Survey questions

Q1 - Your profile Name
Q1 - Your profile Title
Q1 - Your profile Email address
Q1 - Your profile Phone number
Q1 - Your profile Company
Q1 - Your profile Number of employees
Q1 - Your profile Company annual turnover
Q1 - Your profile Recent big data project/company, customer or reference name

Q2 - Please provide information about your existing IT landscape Existing IT systems data volume (GB/TB)
Q2 - Please provide information about your existing IT landscape Main ERP/Business Applications
Q2 - Please provide information about your existing IT landscape Database
Q2 - Please provide information about your existing IT landscape Data warehouse
Q2 - Please provide information about your existing IT landscape Existing middleware
Q2 - Please provide information about your existing IT landscape BI solution
Q2 - Please provide information about your existing IT landscape Big data analytics platform

Q3 - Please answer the business questions on big data below? Please elaborate on the top 5 external data sources related to your big data analytics project? (Ex. customer data, financial data etc.)
Q3 - Please answer the business questions on big data below? Describe business issues and technical issues that drove the big data investment?
Q3 - Please answer the business questions on big data below? What is the total spend on big data analytics to date?

Q4 - Did the big data project require additional storage?

Q5 - Did the big data project require additional processing capability (CPU/cores)?

Q6 - Please elaborate on your experiences relating to hardware for big data?

Q7 - Was there a software investment associated with the big data project?

Q8 - How long did the big data implementation take and what did it entail?

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How do you measure financial gains from big data?

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How do you measure operational gains from big data?

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How do you measure internal user gains from big data?

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How has ROI for big data been presented?

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? Has there been any officially acknowledged efficiency improvements following big data investment? (elaborate please)

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How will big data help in terms of competitiveness and what tip can you give?

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? Has any new services or products been introduced as an effect of big data adoption? (elaborate please)

Q10 - Please answer the technical questions on big data implementation below. Have you experienced any internal challenges following adoption of big data analytics, if so, what advise can you give?

Q10 - Please answer the technical questions on big data implementation below. What is the estimated time spent on patching/upgrades of big data management system and what advise can you give?

Q10 - Please answer the technical questions on big data implementation below. What is the estimated time spent on archiving/recovery of big data system and what advise can you give?

Q10 - Please answer the technical questions on big data implementation below. What is the estimated time spent on migration/testing of big data management system and what advise can you give?

Q10 - Please answer the technical questions on big data implementation below. What is the estimated time spent on runtime analysis/regression of big data management systems and what advise can you give?

Q10 - Please answer the technical questions on big data implementation below. What is the estimated time spent on runtime disaster recovery of big data management system and what advise can you give?

Q10 - Please answer the technical questions on big data implementation below. What is the estimated time spent on data/software architecture and coding of big data management system and what advise can you give?

Q10 - Please answer the technical questions on big data implementation below. What were the data integration costs and time spend and what advise can you give? (Ex. cost of ETL)

Q10 - Please answer the technical questions on big data implementation below. What was the time spend on developing queries and what advice can you give?

Q10 - Please answer the technical questions on big data implementation below. What was the cost and time spend on development of analytics applications and what advise can you give?

Q11 - What do you do to get around systems faultiness, atomicity issues, lack of data consistency, isolation issues, lack of data durability (ACID), caused by big data roll out and what advise can you give?

The survey consisted of 43 questions. The reason for the large number of questions is to establish the validity of the participant's experiences to ensure the integrity of the study. In hind sight the survey questions cold have been reviewed in more detail before the survey was published. The following changes to the survey are recommended:

Concern 1.

Professionals with the following titles participated: 2 SAP Specialist, 2 Sales for Big Data, Cloud Specialist, Analyst, Account manager, EPM & GRC Sales, SAP BI commercial specialist, 2 Support engineer, IT Manager, Sales Executive, IT Director, Data Analyst, Senior solution engagement manager. This mixture of professionals makes it difficult to avoid biases reflecting in the responses and it makes the responses uneven and even at times questionable. Is the participant responding with facts or perhaps responding with information of something they've only heard about rather than what they can confirm?

Concern 2.

The other issues with the job titles are that they don't provide any clarity on what skills the participant has or the participant's remits. The job title ''Support engineer'' could have been further clarified through improved phrasing, for example; ''explain what your role entails'' or ''what are your remits''.

Concern 3.

The number of employees varied from 10 to 67,000 staff and the company turnover numbers varied from 1 million to 76.84billion. However specific details about currency were requested. Therefore there may be a variation of currencies that applies to the numbers provided. Similarly on the questions; '' Company annual turnover'' and '' Was there a software investment associated with the big data project?'', no specifics is being requested in terms of currency.

Concern 4.

Another issue relates to the questions ''Number of employees'' and ''Company annual turnover''. The questions do not specify whether the participant should put the numbers for

their existing company or the numbers of the organizations that they're working with on the big data project. To give an example, almost all the SAP specialists have put different values on the questions ''Number of employees'' and ''Company annual turnover''. The same issue occurs on the question; '' Please provide information about your existing IT landscape- Existing IT systems data volume (GB/TB)''. Hopefully the answers relate to the big data end-user.

Concern 5.

Another issue is that although 90% of the participants responded to the question; '' Please provide information about your existing IT landscape- Main ERP/Business Applications'', the answers provided were not particularly detailed making it difficult to generalize on associated types of IT landscapes that may commonly form the foundation for big data. In hind sight, it may have been naïve to believe that this information would be provided without asking for specific types of business applications. For example; existing web platform and existing finance solutions. An alternative could have been to allow participants to draw an image of the existing landscape.

Concern 7.

There were also answers that were questionable, as participants may have misinterpreted or misread the question. One participant for example referred to Business Objects as the database platform; when Business Objects is a suite of analytics solutions, not a database technology. One of the participants stated that big data had cost 0 which is highly unlikely. It can be assumed that the person was meant to put not applicable as an answer, or used the character 0 to show that the number was not known.

Concern 8.

The question ''Please provide information about your existing IT landscape- Existing middleware'' should have been better phrased to ensure that the middleware was associated with the big data project specifically and not systems that are irrelevant to the big data project.

Summary of key research findings

The purpose of this study was to encourage sharing of big data know-how. Recommendations were shared by participants of a survey that could be of importance to big data explorers and adopters. Responses from 23 IT professionals were shortlisted for this study based on a survey called; ‘’ Opportunities to manage big data more cost efficiently and more effectively’’. The participants are in IT roles where they carry out tasks relating to big data projects. Some participants are technical specialists who carry out hands on big data administration and some are commercial specialists that are involved with big data from a business strategy point of view.

The participants worked for the following companies; Software Placements, SPP, SAP, N.M.E, Edenhouse, BMW, Anonymous customer, Bubbleroom.se, JMG Ltd, Codeit and Birchman.

Some participants could not share the names of the organizations that they were working with due to confidentiality agreements. The participants that could share the names referred to the following big data project names or companies that they were working with in regards to big data:

Consolidation of Customer Database, Addison Lee, Sales and marketing acceleration, ARCO, DFB, Edenhouse Solutions, BMW data, Professional Services providing data as a service to their customers; use case around financial management, SAP, Linking Digital Data- Kings and Marsh, Codeit, PWC and Birchman.

The data sources that the participants stated were the main data sources were as follows:

Client and customer data, candidate data, financial Data, sales data, supply chain data, social media, web server logs, cloud solutions, logistics, CRM, ERP, performance data, retail data, marketing data, distribution data, partner data, stock exchange data, asset associated data,

government data, agency data, archives, e-commerce data, inventory data, SAP internal time monitoring system and SAP finance data.

Positive experiences were shared by the experts in the study about new services and products being introduced as a consequence of the big data adoption. One participant shared that the organization had been able to enter into a new market as the result of the big data adoption, another shared that they'd been able to enhance their products following bid data analytics adoption. Another benefit that was mentioned was the ability to better predict risk management and loan evaluation, increased efficiency and improved forecasting.

Following big data analytics adoption the organizations have been able to respond to client requirements quicker than our competition. The biggest competitive gain has come from real time sales insights enabling shorter waiting times for access to performance statistics form sales campaigns, reduced marketing costs and customer sentiment analysis enabling quick changes of strategy.

Big data has helped with spotting trends before the competition, to be more flexible, be more responsive and increase customer satisfaction. Following big data analytics adoption, it has been noted that processing speed improved, budget assigned to resourcing and time spend on reporting reduced by 70% in one of the cases. Another participant stated that they had been able to quantify that the big data investment had lead to 2% increase of sales revenue.

The KPI's that the participant stated had been used to measure the success of the big data roll out were speed, user satisfaction rates, ability to react faster and make faster decision, growth in numbers of deliveries and orders, improved billing accuracy and project profitability, enable just-in-time-logistics and improved employee satisfaction. Other participants shared that success had been measured on improved alignment between headquarters and field, smoother and more efficient running of warehouses, less excess, less waste and ability to detect, prevent and remediate financial fraud.

The participants also talk about improved sales operation. One example is faster sales cycles, improvement s seen in returns from TV advertisement, reduced stock, increased billable days and improved asset management.

The experts that participated in this study shared experiences of multiple internal changes following big data adoption. In one case staff reductions had been made however in other cases there had been need for additional staff due to skills gaps.

The most sensible tactic is to align investment with business strategy. The rationale is; what is needed to achieve the targets and goals that the organization's is aiming for. However this can be restrictive as this may eliminate opportunity for innovation. If IT investment is only aligned to strategy it relies on the strategy to be innovative and forward thinking.

The variation of organizations that the participants were employed was satisfactory. It provides depth to this study. The fact that the organizations varied in size, turnover and sector supports the notion that big data can be suitable for a large variation of sectors and sizes of organizations. The variation of data volume associated with the big data projects that the participants shared, was telling of the reasons why big data is not defined academically yet. The volumes varied from 3 terabytes to 500 terabytes. This tells that big data simply means different volumes to different organizations and that the more accurate measurement is any volume that is difficult to manage by the traditional DBMS that the organization has. This is supported by the findings in secondary data.

Another interesting outcome from the survey was the variation of database platforms that the participants referred to as being used in association with the big data projects. Sybase ASE, DB2, SAP HANA, MS SQL and Oracle database were noted. The same goes for the supporting data warehousing solution, SAP BW, Oracle Exadata, Terradata. Secondary shows that this is common.

The analytics platforms that were used were SAP Business Objects (BOBJ) and Predictive Analytics, SAP Business Warehouse, Oracle Business Intelligence Suite, QlickTech and SAS for the traditional DBMS and Hadoop, SAP HANA, Terradata and Oracle Big Data Analytics. All these solutions are mentioned as common solutions in secondary data.

The use cases that were the participants stated were the main drivers for big data adoption were varied. One use case was the ability to match client requirements with candidate requirements. Other use cases were real time cash management, cleaner access to customer data from multiple sources, easier maintenance of customer data, real time sales analysis, improve sales effectiveness, marketing campaigns, just-in-time logistics, closing the books, need for speed and scale, enhanced performance, enhance business analytics by leveraging on historical data, hardware investment, knowledge investment, software investment, lack of insight, more precise planning and schedule, sourcing and consolidating data.

Drivers were mentioned such as, increasing data volumes by 20 TB plus per annum, data availability, issues with scalability of current systems, asset management, risk management, forecasting accuracy, marketing, end of month/year operational financial management, lack of real time insight, need for rapid, up to date and accurate management information on projects and fiscal control of projects.

Many uses cases echoed generalizations being made by other researchers in secondary data such as; issues with linking digital data, lack of 360 degree view of customers across all digital channels, not understanding customers behaviors**,** data growth, and reporting latency. However, other uses cases were more unusual such as the use case where the big data investor wanted to deliver an application nationwide for taxi operations management. Another interesting use case was the big data investor who wanted to improve football team management by leveraging on big data analytics.

The future of big data is impossible to predict and there is no regression analysis or series analysis that can be applied to forecast what is to come for big data analytics practice. The only assumption that can be drawn from a positivists point of view is that big data technology will become more advanced with time and larger data sets will be able to get analyzed faster. Another assumption that can be drawn is that adoption of big data will make organizations even more vulnerable to security threats as the repercussions can be even more critical would systems be compromised by intruders (Alter, 2002)

Recommendations

Business strategy recommendations

This research analysis on opportunities to manage big data efficiently and effectively highlights the obvious skills gap in the market relating to big data. The difficulty to collect credible data indicates that information sharing efforts need to accelerate to meet the needs of the market demand. It is recommended to carefully review advice provided by experts, such as service providers, software vendors and hardware vendors, as there could be a case of experts capitalizing on the premature stages of big data analytics adoption and the general inexperience amongst big data investors.

Today a lot of organizations use out-sourcing partners that are responsible for part or full running of their IT. In many cases the partners are also morally responsible for introducing new technology and ideas that can help with performance improvement. The problem is that many times the partner lead engagement may be driven by individuals that may be less or more interested in new technical possibilities. The partner may be limited by existing partnerships which means that they can only introduce certain vendor's technologies and sometimes partners may be too unskilled to realize potential.

Organizations need to ensure that they do not place all the responsibility for IT strategy in the hands of outsiders and that they have resources in place for internal investigators that have relevant skills and interest to realize potential when it's presented (Irani, Love, 2008).

Experts advised that it is important to get feedback from employees to ensure greater take up of big data and acceptance of the new platform. It's also recommended to review measurements for success so that results can be quantified. It's also recommend carrying out preliminary research before the start of the implementation and a review of qualified personnel and resources. Multiple participants refer to the importance of data management and other comments were that attention should be paid to data integrity.

Often strategies are two dimensional. Ignorance is limiting and therefor it's important for IT professionals to lead the way in terms of thinking about IT from an opportunistic point of view and look out for new use-cases and return on investment (ROI) opportunities. Particular care should be paid to intangible benefits such brand awareness and customer sentiment.

Technical recommendations

The participants of the study stated that the big data implementation entailed roll out of a new database platform, implementing hardware, blue print, consultant services, integration, testing, production deployment, system preparation, data cleansing, archiving and took between 1 month and 3 years.

An expert revealed that system faultiness that can be experienced in association with big data projects can be caused by integration issues relating to integration of big data systems and the legacy database systems. Another participant advised to use indexed storage relational implementations.

The hardware experiences shared, included investment in cloud technology to cut down cost of hardware cost. SAP HANA Enterprise Cloud and SAP HANA appliance and Dell hardware were mentioned. It was also mentioned in one case that synchronous replication requirements increased. It was advised to create a competitive situation by having discussions with multiple hardware vendors simultaneous to achieve a strong negotiation position. Another advice was to sync the big data adoption with hardware refreshes and to leverage on financial support services to ease up front costs.

Review of the existing architecture as well as streamlined processes is an associated exercise that many of the participants mentioned to in the study. Decision makers should consider that on development of analytics applications can take 6 months. The time spent on developing queries was between 2 weeks and 3 month. Time spent on data/software architecture and coding of big data management system should also be considered and the participants of the survey refer to the importance of good consultancy and to allocate between 2 weeks and 2 months for the task.

Self-reflection

Thoughts on the projects

The topic of big data is brought up daily in my work place. The topic intrigued me very much because I couldn't figure out why it was such a big deal. Surely, managing a large set of data should not be a too difficult of a task, I thought to myself. As I started to ask database professionals, architects and BI specialists I noticed, that they too were uncertain. I started to wonder whether it was just yet another hyped up concept made up to prompt more hardware and software sales in the industry. I started to bring up the topic in conversations with my customers and to my surprise I noticed that it was a very relevant topic for them. Many of them were already using big data analytics and some had made significant investments in big data management tools and personnel.

I decided to finally retrieve some of the many white papers on big data from my Dropbox and have a read. The biggest surprise was the lack of definition, the lack of research that was available and the fact that most of the information talked about big data from a business point of view rather than a technical point of view. I started to understand why so many talked about it and the business benefits of big data but few seemed to understand what it required, how it worked and how to practically manage big data.

I started to wonder why that was, why did so few have a technical understanding of big data and why was the know-how not more widely shared? Was there an industry pact created to secure consulting revenue by limiting access to technical know-how and insight?

As I continued my research I concluded that there is no conspiracy theory around big data, it's simply management of larger sets of data than what most technologies can handle and due to the rarity of solutions that enables management of larger sets of data, it becomes something unique.

This lead me on to thinking about what the effect could be if more people were provided with deeper insight, shared by big data professionals that they could use in their own exploration, adoption or in the daily running of their big data systems?

My suspicions are that big data technology eventually can catapult a new wave of data oriented product and services that can make a significant impact in many sectors.

The biggest issue that I foresaw already at the beginning of this research was that it would be difficult to find professionals that work on big data projects daily. I choose to turn to the top IT firms and associated partners for insight. The technical team at SAP, Salesforce, Edenhouse Solutions and Birchman Solutions were of great help however it did not mount up to a large volume of potential participants. It leads me on to think that perhaps it was too limiting to look for specifically big data professionals.

I started to look for participants amongst technical support professionals that deal with ERP, data warehouse and database technologies daily. I thought that by turning to the professionals that big data access daily through IT support contracts, that I'd get access to a wider community of everyday big data administrators. I was wrong. I found that the experience amongst technical support professionals in regards to big data was very varied and although many were exposed to big data related enquiries and tasks daily, they were not involved enough in the projects to have a more holistic view.

The difference in the level of big data skills was particularly surprising amongst some of the technical professionals that participated in the survey as some were working at the same company, doing the same job and carrying the same job title.

The conclusion was that most people that had grown a deeper understanding for big data had only done so as it had been required in their day job. One SAP Support specialist was particularly insightful as he had been tasked by several customers to support with data inconsistencies caused by big data and therefore had been driven to read up on the topic and associated technologies to be able to help them technically with their queries.

IT consultancy firms that participated in this study also expressed that they had only just started to invest in big data staff and resources such as testing labs and proof of concept environments as customers had started to demand big data support.

Given the government legislations that have been introduced in recent years in regards to data retention as well as the data boom relating to the web, cloud computing and mobility, organizations will miss the opportunity to leverage on competitive edge through big data investment if they do not start to reflect on possibilities before the competition does.

Government and education bodies seem to get it but can efforts to drive more focus, find talent, and develop new tools materialize fast enough to drive more adoption before organizations operations are hampered by the steady data growth that's due to legislations?

It appears as if government implemented new legislations without providing sufficient guidelines for how to handle the expected data growth. A thought relating to web data and mobility data is that perhaps programmers and developers should look at new data formats with more compression capability that enables more efficient storage and faster processing than the large file formats that are common to use today.

I eventually managed to get enough participants, however I choose to just use 23 surveys for this study as the other surveys were questionable. The results were very interesting and gave fruit for thought.

What was also very educational was the actual creation of the survey itself. As I drafted the survey I knew that I wanted to ask questions that I hadn't seen other researchers ask. I knew that I wanted to focus on requirements for big data and details of implementation and every day administration to gain insight. It was not surprising that only a few participants completed all the questions in the survey with solid answers rather than notes such as ''not applicable''.

It is understandable that not everyone is exposed to all elements of big data management and what my research shows is just how big the skills gap is. In particular the gap between business professionals and technical professionals but also the fact that top tier IT consultancy firms at the level of SAP, Oracle and Salesforce that all promote big data adoption, still struggle with knowledge sharing in regards to big data. The skills gap amongst the participants of this study was apparent and it was surprising how many IT professionals that participated failed to answer even the most basic questions associated to big data.

Another factor that drove my interest for this topic is my growing interest in furthering my technical skills in my day job. I work as an Account Executive for SAP and I look after database sales and projects for all our SME customers in the UKI with a turnover below £300 million. Within my product portfolio is SAP HANA which is SAP's in memory database technology that promises real-time analytics without any need for data aggregation due to the in memory processing capability. Its biggest rival is Oracle's Exalytics. These are also two common analytics platforms used for big data analytics.

The reason why I had not been exposed to technical tasks associated with big data projects was due to the general size of the organizations that I work with. Small and medium sized enterprises do not adopt big data technology as widely as larger enterprises. In my conversations with customers about big data, they expressed that the main factor that hindered them from adopting a big data practice was the associated cost.

However many studies such as the one carried out by Winter Corporation in 2013, shows that there is not a significant upfront cost associated with big data in comparison to a traditional data warehouse roll out. In fact, figures shows that the upfront cost is lower and that the main costly areas are more around report writing and data processing. I hope that studies such as this one that I have carried out can help organizations with their reviews and that it will give decision makers deeper insight that may still the fears associated with investing in big data.

Formulation

It's fair to say that this study could become even more useful if more participants would have been found. In hindsight I would have wished to find the angle for my study far earlier to be able to get access to big data professionals who were willing to participate in an observation study so that I could get even deeper insight to what they do daily and the challenges they face. Another idea could have been to choose a qualitative study instead and gain a narrower but deeper understanding through a case study with a big data practicing company.

There were moral issues with getting access of this kind as many originations that I asked worried about data protection and confidentiality issues associated with allowing an outsider access to their data, systems, and internal intelligence.

I believe that what I have produced is new and fresh as many of the other studies that I've read have been very focused on the commercial side of big data such as cost analysis's and comparisons to traditional data warehousing. This study addresses the technical concerns with weight on the existing skills gap and the lack of know-how. I have not seen any study yet that addresses the issue in a practical way, providing advice from professionals that can be used by the wider community.

Most of the information that a big data administrator needs can be found online, but it's not common for non-technical professionals to seek to very technical communities including myself, as the language used and the jargons can create a feeling of alienation. I aimed to use simple language to make this reading as easy to understand as possible to those that may not have had exposure to the topic before for this very reason.

I hope that I've been able to cover most of the basic topics that should be relevant for big data explorers of practitioners as a way to whet their appetite for further discovery.

Main learnings

The survey that I designed was created in such a way that it would be very apparent who had hands on experience with big data and who didn't but the result showed that even those involved in big data projects were often not exposed to anything else but their specific job task and therefore could not answer general questions around the project that they had been part of.

I can relate to this as I sit and sell cutting edge technology every day that I have no idea how to implement or administer. In fact it's not uncommon for me to be completely overwhelmed by the questions customers ask me about basic database related matters and yet I'm considered to be a database specialist.

The learnings from this research has expanded my IT skills and I feel more assured in my role that is currently very commercially focused, to carry more technical discussions rather than relying on technical resources around me. I also feel more confident about what I don't know. I've learned that the learning curve is never ending and I've been reminded of the empowering feeling it gives to get that ''aha'' moment when pieces of a puzzle come together and the understanding of a topic appears clear.

I'm pragmatic, when people told me that there was a skills gap relating to big data I tested the theory by creating a survey with both business and technical questions to test the theory. What I also noticed was a new found respect amongst my peers and management as they witnessed my passion and personal development throughout the dissertation project time.

Although access to big data specialist was difficult to gain, I'm happy with the resources that I did find and I feel that the participants the study were reflective of the sort of resources that organizations generally have available for big data initiatives, which makes this study very reflective of a scenario where a decision maker is trying to source talent for their big data roll out. I can say that I've gained a little bit of insight into the challenges of the task.

I’ve also experienced a moment when someone asked me about something technical that I would not have known the answer to just a few months ago, and I truly enjoyed the experience of being seen as a source of knowledge. However I realize that I still have a big skills gap to fill within myself in terms of hands-on experience and software administration. This experience has enticed me to want to learn basic technical administration skills to develop my skills further.

Word count 20,021

Bibliography

1. Alter S, 2002, Information Systems- The foundation of E-Business, fourth edition, pp 33-34,
2. Beynon-Davies P, 2002, Information Systems- An introduction to informatics and organizations, pp 558-567
3. Boddy D, Boonstra A, Kennedy G, 2008, Managing Information systems- Strategy and Organization, third edition, pp 8-10, pp 271-273
4. International Journal of Social research Methodology, Taylor&Francis Ltd, Vol 8, No 3, July 2005, pp 173-184
5. Chaffey D, Wood S, 2005, Business Information Management-Improving Performance using Information Systems, pp 122- 123, pp 508- 509, pp 516- 518, pp 542-551
6. Chartered Institute of Bankers, 1996, Management of Information Technology, pp 208-217
7. Dalkir K, 2005, Knowledge Management in Theory and Practice, pp 52-56
8. Ellis S, 2005, Knowledge –Based Working- Intelligent Operating for the Knowledge Age, pp 2-20
9. Galliers R.D, Leidner D.E, 2003, Strategic Information Management- Challenges and strategies in managing information systems, third edition, pp 406- 411, pp 582-583
10. Harry M, 2001, Business Information: A systems approach, pp 223-225,
11. Hsu C, 2013, Information Systems- the connection of people and resources for innovation, pp 184-185
12. Irani Z, Love P, 2008, Evaluating Information Systems – Public and private sector, pp 1-3, pp 8-10, pp 18-21,
13. Jessup L, Valacich J, 2003, Information Systems today, Second edition, pp 97-99
14. Laudon, K.C and Laudon, J.P (2014), Management Information Systems, Managing the digital firm, 13th edition, pp 36-37, pp498-499, pp254-269
15. Marakas G.M, O'Brien J.A (2013), Introduction to information systems, 16th edition, pp 425-427, pp 199-214
16. McKeen J.D, Smith H.A, 2004, Making IT Happen- Critical Issues in IT Management pp 265- 279
17. McNurlin B.C, Sprague R.H.Jr, 2006, Information Systems Management In Practice, 7th Edition, pp 269-277
18. Pearlson K.E, 2001, Managing and using information systems- A strategic approach, pp 21-23
19. Rainer R.Kelly, Turban E, 2009, Introduction to Information Systems, Second edition, pp 106-125
20. Teece D.J, 2000, Managing Intellectual Capital, pp 3-13
21. The Marketing Review- Waterford Institute of Technology, Holden M.T, Lynch P., 2004, Choosing the appropriate methodology understanding research philosophy, pp 397-409

Web resources

22. Butler. J. 2012, An IBM-Tech America Event, Techamerica a big data commission, Demystifying Big Data, Big Data and analytics at the IRS, retrieved 21st April 2014 from https://www-950.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/Jeff%20Butler%27s%20Presentation/$file/Jeff%20Butler%27s%20Presentation.pdf
23. Capgemini, 2014, Deciding Factor: Big Data & Decision Making, retrieved 16th April 2014, from http://www.capgemini.com/thought-leadership/the-deciding-factor-big-data-decision-making
24. Columbus, L. (August 16, 2012). Roundup of Big Data Forecasts and Market Estimates, 2012, Forbes. Retrieved 30th August 2014 from http://www.forbes.com/sites/louiscolumbus/2012/08/16/ roundup-of-big-data-forecasts-and-market-estimates-2012
25. Diebold.X.F, 2000, ‘’Big Data’’ Dynamic Factor Models for Macroeconomic Measurement and Forecasting, retrieved 16th April 2014 from http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF
26. Forbes, 2014, A very short History of Big Data, retrieved 20th April from http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/2/
27. Forrester,Shaheen Parks, 2014, Projected cost analysis of SAP HANA- cost savings enabled by transitioning to HANA, retrieved 8th June 2014 from http://www.sap.com/bin/sapcom/en_us/downloadasset.2014-04-apr-14-22.projected-cost-analysis-of-the-sap-hana-platform-cost-savings-enabled-by-transitioning-to-hana-pdf.bypassReg.html
28. Gartner Inc, 2012, Gartner Says Big Data Will Drive $28 Billion of IT Spending in 2012, retrieved 19th April 2014 from http://www.gartner.com/newsroom/id/2200815
29. Gartner Inc 2012, Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015. , retrieved 19th April 2014 from http://www.gartner.com/newsroom/id/2207915.
30. Google.com, 2014, Google Trends, Topics- search term ‘’big data’’, Interest over time, retrieved 21st April 2014 from http://www.google.com/trends/explore#q=big%20data
31. Laney. D. 2001, on behalf of META Group, 3D Data Managemnt Controlling Data Volume Velocity and Variety, retrieved 20th April 2014 from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
32. McKinsey&Company, Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung-Byers A, 2011, Big data: the next frontier for innovation, competition and productivity, retrieved 11th July 2014 from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
33. The White House, 2012, Kalil.T., Big Data is a Big Deal, retrieved 20th April 2014 from http://www.dailytech.com/Obama+Admin+Plans+200M+USD+Big+Data+Spending+Spree/article24350.htm
34. Yan, J, 2013, Big Data Opportunities- Data.gov’s roles: Promote, lead, contribute, and collaborate in the era of big data, retrieved 28th July 2014 from http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf

Other recommended readings

1. Anderson, J. & Rainie, L. (July 2012). The future of big data, the Pew Research Center's Internet & American Life Project Series. PewInternet. http://www.pewinternet.org/Reports/2012/Future-of-Big-Data.aspx
2. Bakshi, K. (Oct. 11, 2012). Big Data Trends: Architectural & Strategy Guidance. Presentation at 2012 Big Data Conference. http://www.digitalgovernment.com/media/Downloads/asset_upload_file41_4504.pdf.
3. CTOLABS.COM (2012). The winner of the 2012 Government Big Data Solutions Award is the National Cancer Institute. http://ctolabs.com/the-winner-of-the-2012-government- big-data-solutions-award-is-the-national-cancer-institute/
4. Dow Jones Newswires (February 28, 2013). IBM Sets Higher Bar for Revenue from Big Data. http://www.foxbusiness.com/technology/2013/02/28/ibm-sets-higher-bar-for-revenue-from-big-data/#ixzz2NA10gNDC.
5. Floyer, D. Kelly, J., Vellante, D., & Miniman, S. (Feb. 25, 2013). Big Data Database Revenue and Market Forecast 2012-2017. http://wikibon.org/wiki/v/Big_Data_Database_Revenue_and _Market_Forecast_2012-2017
6. Franzen, C. (November 2, 2012). Charting agency big data progress. FCW. http://fcw.com/Articles/2012/11/02/big-data-agency-progress.aspx?Page=1.
7. Gabriel, A.R. (November 1, 2012). Getting a Handle on Big Data, Fall 2012 issue of FedTech Magazine. http://www.fedtechmagazine.com/article/2012/11/getting-handle-big-data.
8. Gil Press (February 20, 2013). Graduate Programs in Big Data Analytics and Data Science. http://whatsthebigdata.com/2012/08/09 /graduate-programs-in-big-data-and-data-science
9. Goldman, T. (Oct. 11, 2012). A Practical Approach to Big Data. http://www.digitalgovernment.com/media/Downloads/asset_upload_file389_4504.pdf
10. Konkel, F. (March 15, 2013). Of policy and petabytes: Shaping the use of big data. http://fcw.com/articles/2013/03/15/big-data-policy.aspx.
11. Konkel, F. (March 11, 2013). Big data's big hurdle: Federal policy. http://fcw.com/articles/2013/03/11/big-data-policy.aspx
12. NetApp (May 7, 2012). The big data gap. http://www.meritalk.com/bigdatagap
13. SAS (2013). Big Data: Lessons from the Leaders, Economist Intelligence Unit Report. http://www.sas.com/reg/gen/corp/1774120
14. Smith, J.A. (December 19, 2012). FIELD NOTE: What Makes Big Data Big – Some Mathematics Behind Its Quantification, Data Scientist Insights. http://datascientistinsights.com/2012/12/19/ field-note-what-makes-big-data-big-some-mathematics-behind-its-quantification/.
15. Taleb, N.N. (February 8, 2013). Beware the Big Errors of 'Big Data'. Wired. http://www.wired.com/ opinion/2013/02/big-data-means-big-errors-people/.
16. TATA Consultancy Services (2012). Big Data ATCS 2013 Global Trend Study. Big Data Investment: Which industries should be investing more in Big Data? http://sites.tcs.com/big-data-study/industries-big-data-investment
17. The White House, Office of Science and Technology Policy. (March 29, 2012). Big data is a big deal. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.
18. The White House, Executive Office of the President. (March 29, 2012). Big data across the Federal government (fact sheet). http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet.pdf

Appendices

Appendix A: Examples of big data analysis methods

The main analytics methods that big data supports are as follows:

Big data analysis method	Explanation
Visualization	To capture a true picture of data, visualization tools can be used that makes data easier to understand such as heat maps, tag clouds, cluster-gram, history flow, spatial information flow, and performance dashboards. Data can be translated into images, diagrams and animations for ease of interpretation and to ease communications.
Data mining	As mentioned previously, data mining is analysis of an extract of a larger data set that is analyzed using statistics, ensemble and machine learning, association rule learning, cluster analysis, classification and regression.
Statistics	Statistics is crucial to big data analytics. Statistics can for example guide prioritization and create sense of urgency for action. Statistics entails capturing and interpreting data from multiple sources such as surveys, tests and experiments to draw assumptions and justify assumptions about relationships between things. It helps with eliminating errors or false information It also helps with predictability and can be used for A/B testing
A/B testing	Tests on variables of relative differences to establish best options in comparison to multiple other options. To give an example, a test group that views cars while wearing sensors that can pick up reaction. A/B testing would analyses reactions between views of different cars and from a multi-dimensional perspective, such as change of car features, also taking into account attributes such as demographic, background and stereotype of the test person. Big data analysis enables this multi-dimensional analysis technique as it can process a large set of data simultaneously.
Natural language processing	Natural language processing is machine learning technology. Machine learning enables computers to develop behaviors

	based on empirical data. It enables recognition of complex data patterns. Machine learning fits into the category of artificial intelligence science.
Optimization	Genetic algorithms are used for optimization amongst a few other techniques. Optimization is used to improve data processing and data modelling to improve for example processing speed.
Association rule learning	This is the analytics most commonly used for data mining. Big data enables views of relationships that normally would not be apparent to the human eye as it can process large sets of historical data that can show patterns that otherwise would not be obvious. This analytics method allows retailers to view correlations between purchases of multiple products for example a scenario where there is a clear pattern that when customers buy instant barbecues they also buy matches and are more likely to buy chicken or meat. Retailers could use such information for product placement in shops and online to make sure to maximize the up sale opportunity per customer transaction.
Classification	Classification fits in to the category of pattern recognition analytics. This method is also commonly used for data mining as explained in the section earlier about data mining. Its' used to determine the most likely characteristics in relation to a particular outcome. If we continue the example scenario with the barbecue purchase. Classification would describe the characteristics or conditions associated with the person who makes the purchase. For example; more barbeques are sold when it's sunny outside, and more barbeques are sold in the month of July in comparison to January and the barbecues are most likely to be bought by men in the age group between 25-55 years old whilst the chicken in most likely to be bought by women with more than one child in the household.
Regression	Regression is a predictive analytics modelling technique. Predictive analytics is a method to measure probability through multiple matrix such as historical data, pattern, trend and characteristics of an item, person or an event. It's also used for data mining.
Cluster analysis	Cluster analysis can be used to compare a smaller more focused set of data for example to compare the barbecue purchases during one week in July to one week's purchase in January or to look at the scenario from a demographic point of view; for example only purchases in a particular city.

Crowdsourcing	Capturing online data is a challenge for many organizations. Crowdsourcing enables analysis of a larger set of data such as social media, blogs, online communities or analysis of data captured at events such as concerts, industry events or festivals.
Data fusion and data integration	What makes big data particularly unique in comparison to traditional DBMS analytics is the capability to analyze data from multiple sources in parallel through signal processing, using natural language processing in real time which can for example be used for analysis of internet-of-thing data or customer sentiment. Signal processing enables processing of subtle and continuous signals. This can be used for time series analysis, which is popular to use for analysis of radio, sound and image. It enables differentiation of signal and noise. Time series analysis uses statistics and signal processing. It looks at sequence to extract characteristics from data. Stock markets use this type of analysis as it provides the ability to predict future values and structural modeling which is when data used to decompose a series into trends, patterns and for example seasonal information this Is another popular technique for forecasting.
Neural networks	Neural networks analytics shows linear as well as non-linear patterns and can be used to identify discrepancies in patterns that can be used to for example identify bank fraud or insurance fraud.
Network analysis	Network analysis is often used to analyze social media. It can be used to look at relationships between individuals or organizations. Organizations can this kind of analysis to monitor how information travels, for example news and associated influences.
Ensemble learning	Ensemble learning is a way of using predictive analytics supported by statistics. It fits into the category of supervised learning and machine learning. Ensemble learning is applicable in classification too.
Unsupervised learning	Cluster analysis is an example of unsupervised learning. It is a machine learning technique that can find structure in unlabeled data.
Spatial analysis	Spatial analysis looks at geometric and geographic properties encoded in a set of data. Latitude, longitude, addresses are

	some examples of the sort of information that it provides and this can then be used for regression analysis such as an analysis showing consumer behavior in a particular geographical location or in simulation analysis, showing performance of different shops for example. A simulation models behaviors of a process and is often used for forecasting and scenario planning.

(McKinsey&Company,2011)

Appendix B: Survey results

See results from survey research on the following page, extracted in excel format from website *https://www.surveyplanet.com*

Survey Participant	Date Taken	Q1 - Your profile Name
brennsgk1@gmail.com	August 12th 2014, 4:02:05 am	Steve Brennan
luke.roche@hotmail.co.uk	August 12th 2014, 4:22:23 am	Luke Roche
colin.morrissey@sap.com	August 12th 2014, 4:49:58 am	Colin Morrissey
al_weir@hotmail.com	August 12th 2014, 5:24:34 am	Alex
kieran.o-riordan@sap.com	August 12th 2014, 5:28:04 am	Kieran O'Riordan
david888@eircom.net	August 12th 2014, 5:26:51 am	Dave
chris.freeney@sap.com	August 12th 2014, 5:50:44 am	Chris F
thomas.siddons@gmail.com	August 12th 2014, 6:03:18 am	Thomas Siddons
eamonn.doheny@sap.com	August 12th 2014, 6:23:13 am	Eamonn Doheny
andre.delaney@sap.com	August 12th 2014, 6:27:39 am	Andre Delaney
gamonal6@yahoo.co.uk	August 12th 2014, 7:48:00 am	raul
rlmetcalfe@hotmail.co.uk	August 12th 2014, 9:13:15 am	Richard Metcalfe
attilio.petrini@sap.com	August 12th 2014, 9:45:48 am	Attilio Petrini
adam.gilbey@edenhousesolutions.co.uk	August 13th 2014, 4:35:53 am	Adam Gilbey
charlie.proctor@bmw.com	August 13th 2014, 5:51:09 am	Charlie Proctor
suresh.chinnasamy@sap.com	August 13th 2014, 6:06:49 am	suresh
brian.patterson@sap.com	August 13th 2014, 6:40:57 am	Brian Patterson
thomas.kagiri@gmail.com	August 13th 2014, 7:06:33 am	Thomas Kagiri
raymondasare@gmail.com	August 13th 2014, 8:50:10 am	Raymond Asare
jmgahaya@gmail.com	August 13th 2014, 1:54:43 pm	jm
newlyby@gmail.com	August 14th 2014, 6:10:48 am	Ade
roy.ferrari@sap.com	August 18th 2014, 1:03:28 am	roy ferrari
sudesh.lourdes@birchmangroup.com	August 19th 2014, 7:04:30 am	Sudesh Lourdes

Q1 - Your profile Title	Q1 - Your profile Email address	Q1 - Your profile Phone number
SAP Specialist	brennsgk1@gmail.com	3.54E+11
Sales	luke.roche@hotmail.co.uk	862412764
Sales for Big Data	colin.morrissey@sap.com	858464210
Cloud Specialist	al_weir@hotmail.com	831800960
Analyst	Kieran.o-riordan@sap.com	877757826
Account manager	david888@eircom.net	003531471 9038
EPM & GRC Sales	chris.freeney@sap.com	14719036
Mr	thomas.siddons@gmail.com	+353 87 233-1203
SAP BI commercial specialist	eamonn.doheny@sap.com	1628761394
SAP specialist	ANDRE.DELANEY@SAP.COM	872110697
support engineer	gamonal6@yahoo.co.uk	n/a
IT Manager	richard@networksmadeeasy.com	3.53E+11
Sales Executive	attilio.petrini@sap.com	N/A
Mr	adam.gilbey@edenhousesolutions.co.uk	07826 541097
IT Director	charlie.proctor@bmw.com	4.48E+11
Mr	suresh.chinnasamy@sap.com	872225196
Mr	brain.patterson@gmail.com	086 8128 608
Support Engineer	thomas.kagiri@gmail.com	831706218
Data Analyst	raymondasare@gmail.com	4.68E+11
Mr	jmgahaya@	419847536
Mr	newlyby@gmail.com	899558858
senior solution engagement manager	roy.ferrari@sap.com	7966975066
Mr	sudesh.lourdes@birchmangroup.com	7445771607

Q1 - Your profile Company	Q1 - Your profile Number of employees	Q1 - Your profile Company annual turnover
Software Placements	20	â‚¬1m
SPP	200	80m
SAP	26000 worldwide	not sure
SAP	60,000	10 Billion
SAP	60000	N/A
SAP	35000	$19B
SAP	60000	16.8BN
SAP	67k+	â‚¬16.81 bn
SAP	50000+	n/a
sap	100,000	na
SAP	25000	n/a
N.M.E	50	5 mil
SAP	> 1000	N/A
Edenhouse Solutions	150	26m
BMW	101,000	$76.84 Billion
SAP	1300	billion
I'm working for an anonymous customer	12000	2 billion euro
SAP SSC Ireland Ltd.	1000+	N/A
Bubbleroom.se	22	Â£10 Million
jmg Ltd	250	50 million
Codeit	10	25,000,000
SAP	50,000+	â‚¬20bn
Birchman Group	250	N/A

Q1 - Your profile Recent big data project/company, customer or reference name	**Q2 - Please provide information about your existing IT landscape Existing IT systems data volume (GB/TB)**
Consolidation of Customer Database	GB
n/a	3tb
Addison Lee	40TB
Sales and maketing accelaration for customer x	SAP ERP, CRM 5
N/A	500TB
Arco	120 TB
N/A	N/A
DFB	175
Company confidential	Bobj
Customer X using big data for sales and marketing acceleration	5TB
n/a	n/a
n/a	90tb
N/A	250TB
Edenhouse Solutions	4TB
BMW Data	Tens of Thousands of Users, Tens of Millions of Records
NA	TB
Professioanl Services providing data as a service to their customers; use case around financial management.	35 TB
SAP	Big, N/A
Linking Digital Data- Kings and Marsh	35TB
data warehouse	tb
codeit	!TB
PWC	3tb+
N/A	N/A

Q2 - Please provide information about your existing IT landscape Main ERP/Business Applications	Q2 - Please provide information about your existing IT landscape Database	Q2 - Please provide information about your existing IT landscape Data warehouse
SAP Business One	Oracle	Oracle Exadata
SAP	MS SQL	SAP BW
NA	ASE	NA
SAP ERP	Business Objects	I don't know all the details
SAP	Oracle	Terradata
SAP	oracle	oracle
SAP	HANA	HANA
SAP	HANA	HANA
Oracle	Oracle	Oracle
SAP	MS SQL	SAP BW
SAP	Oracle	n/a
Microsoft Dynamix	Sql	n/a
SAP	sybase ASE	HANA
SAP	MS Sequel	BW
SAP - Multiple	SAP HANA	n/a
BW	DB6 and HANA	BW
SAP	Oracle	SAP B/W
SAP ERP	SAP HANA	SAP Netweaver
Oracle E-Business Suite	Oracle Database 10g	Oracle Data Warehousing
no	relational	data warehouse
FI/CO	DB2	SAP BW
SAP	HANA	SAP HANA
N/A	All	N/A

Q2 - Please provide information about your existing IT landscape Existing middle wear	Q2 - Please provide information about your existing IT landscape BI solution	Q2 - Please provide information about your existing IT landscape Big data analytics platform
Oracle Fusion	Business Objects	Hadoop
sap	SAP BOBJ	HANA
not sure	SAP Business Objects	SAP
I don't know all the details	I don't know all the details	SAP Hana
Netweaver	Business Objects	Hadoop & HANA
fusion	Buisness Onjects	Hana
N/A	Bobj	HANA
HANA	Crystal Reports	HANA
n/a	Business Objects	Oracle
ETL	Bi SUITE	SAP HANA
n/a	n/a	n/a
n/a	QlickTech	n/a
N/A	SAP Predictive Analytics	SAP Business Objects
SAP Net Weaver	Business Objects	Hana
n/a	BOBJ	SAP
NA	BOBJ	NA
SAP CRM with qRFC/ALE	SAP Business Objects and SAP Businwess Warehouse	Teradata
SOA	SAP BW	SAP HANA in-memory Analytics
Oracle Big Data SQL	Oracle Business Intelligence Suite	Oracle Big Data Analytics
no	sas	no
SOA	SAP	SAP HANA
SAP PI	BOBJ	HANA
N/A	N/A	N/A

Q3 - Please answer the business questions on big data below? Please elaborate on the top 5 external data sources related to your big data analytics project? (Ex. customer data, financial data etc.)
Client Data / Candidate Data / Financial Data /
Financial data, customer data, sales data
Murex
External data, financial data and supply chain data
Customer Data, Social Media, web server logs, Financial, Cloud solutionsTelemetry
CRM, ERP, BI, Financial Planning, Logistics
n/a
performance, movement, results data
Web, financial data, retail data, marketing data and customer data
Logistics Data, Distribution, Web Platform, Online Social Media, Customer Data, Partner Data
n/a
Customer Data, Transactional Data
Customer Data, Financial Data, Market sentiment,
Financial Data, Utilisation Data, Customer Data, Social Media
Supplier, Customer, Labour, Financial, Logistical
DB
Customer data, stock exchange data, asset assoicated data, government data, agency data
N/A
Transactional Data, Archives, Activity Generated Data, Social Media Data and E-Commerce Data
n/a
customer, financial, logistic, inventory, transactional
SAP, Internal time monitoring system, SAP Finance, customer data from CRM
Master data, financial data, production data

Q3 - Please answer the business questions on big data below? Describe business issues and technical issues that drove the big data investment?
We did not possess the ability to match client requirements with candidate requirements
Real time cash management, cleaner access to customer data which came from multiple sources, easier maintenance of cusomter data, real time sales analysis
Customer wanted to deliver an app nationwide for the largest taxi company
To improve sales effectiveness and marketing campaigns
Unorganised and unformated data in large volumes at real time. No platform to support the data types or volumes.
Just in time logistics, closing the books
speed and scale
how to improve game play of football team through shorter pass times
Enhanced performance, enhance business analytics by leveraging on historical data
Lack of insight, more precise planning and schedule, Sourcing and Consolidating Data
hardware investment, knowledge investment, software investment
Increasing Data Volumes. 20 TB plus per annum. We were having issues around Data Availability and the scalability of current systems
lack of real time insight
Need for rapid, up to date and accurate management information on projects and fiscal control of projects
GP to increase 0.8% by Y/E - in line with this
Data volume maintanance
Asset management, risk management, forecasting accuracy, marketing, end of monh/year operational financial management
At the core of the SAP HANA real-time platform is the SAP HANA database. Unlike other database management systems on the market today, the SAP HANA database processes both transactional and analytical workloads fully in-memory. By consolidating two landsc
We had an issue with linking Digital Data, Lack of 360Â° view of our customers across all digital channels, not understanding customersâ€™ behaviors.
n/a
Data growth, management and analysis
reporting took too long
Complex architecture set up, streamline processes, scalable for HANA

Q3 - Please answer the business questions on big data below? What is the total spend on big data analytics to date?
â‚¬40,000
Â£1m
500k GBP
0
$500M
Â£700k
n/a
n/a
company confidential
750k Estimate what im aware of.
n/a
30,000 Euros
N/A
Â£1.5m
N/A
NA
I heard something in the region of 2.7 million euro upfront cost; however, I know there has since been more investment.
N/A
Â£500K
n/a
250,000
Â£2.5m
N/A

Q4 - Did the big data project require additional storage?
No
Yes - extra cost needed on hardware for storage - however this extra cost is counteracted by ease of maintenance which has led to a reduction of headcount in the IT department. Don't be put off by high initial cost thinking big data tools only bring increased speed. The simplification the of the IT environment was the biggest benefit to me and brought a huge cost saving
No additional storage was required
Yes but I don't know all the details
Yes. 10's of millions. Standard RDBMS systems are not suitable for such projects.
Yes, 50TB. Â£450k
n/a
n/a
Yes. Company confidential
Yes but details are unavailable
n/a
20 TB, Dell Compellent was great for powering our critical workloads with self optimized, auto-tiered storage.
yes additional storage was required at a considerable investment
Yes for DR and data persistance. Approx 2 TB storage. Tip - dont underestimate the growth potential of data.
N/a
Compressed data storage
Yes, however I was not involved in this.
N/A
15TB extra storage bought.
N/a
250GB
no
Storage was already there in the SAN so not sure of the cost

Q5 - Did the big data project require additional processing capability (CPU/cores)?
No
n/a
80 cores of ASE
Yes but I don't know all the details
No
Yes, 64 cores, Â£150k. tip is to get clean data
n/a
n/a
Company confidential
Yes but details are unavailable
n/a
N/A
move to in-memory for processing capability to allow for real-time insight and analysis
Yes 8 additional 4 core CPU's were required.
Unsure - beginning of project
yes, nearly double
Yes, but again I was not directly involved. I know extensive preparation and investment was required to prepare the project, a lot of work from a data management perspective, there were multiple hardware refreshes and I know it was quite problematic.
No, if anything we required less resources than before while getting almost double the processing speed
Yes we bought new servers worth Â£200K with higher CPU capacity.
No
ITB . 520 physical memory
HANA appliance - 40 Cores
No

Q6 - Please elaborate on your experiences relating to hardware for big data?
Yes, we decided that a Cloud model was suitable for us so we are exploring this possibility but this investment is more transactional than capital and thus more manageable.
Yes - Hana enterprise cloud - via our partner
With ASE, Hardware is not required as part of this purchase , unless they are purchasing Hana. They will require hardware if they explore this option
Yes but I don't know all the details
Yes there was a data center requirement.
yes, synchronous replication requirements increased.
n/a
n/a
My tip would be to shop around, always have 2 hardware vendors at the same time
We synced the big data project with the clients hardware refresh, and also looked at the financial services support offerings to enable branching out of hardware costs to drive the project forward and within Budget
n/a
A Data Centre investment was needed. We had a work shop with the Dell Big Data specialists and had a solution tailored for our needs.
we preferred to keep in house
No this was managed in house.
Current landscape capable.
I don't have much experience on big data
See above.
Cloud hosting, no local data center maintained.
No hosting or data center investment needed, we are a very small company with on premise servers.
No
We invested on cloud hosting and also use dedicated servers in some data centres
HANA applaicne was quite expensive
Review of the architecture was required.

Q7 - Was there a software investment associated with the big data project?
Only on expertise and consultancy
Yes - 400k - match software and big data vendor for better integration, one single maintenace instance and if possible one single services provider
Cost 500k for overall Project
Yes - Cloud for Sales
There was investment in the Hadoop
No
n/a
n/a
Hadoop. Had to buy additional analytics solution (cannot say more)
Yes, No specifics given but Hadoop was there. NO SQL. A new analytics platform from Oracle. A new BI Front end from SAP
n/a
Not as of yet. We are looking at the possibility of using a system to leverage the amount of dark data we have and turn it into useful data with which we will make informed decisions. We understand our amount of data has been an issue but we see it as a huge opportunity to leverage information as an organisation we have collected in the past.
yes, we implemented SAP
Yes additional skills were required around the HANA implementation.
SAP HANA purchased last year in preparation. Costings cannot be disclosed.
yes
Yes there was. We were bidding against Teradata but lost on the basis that thy reached the customer first and we got into the loop too late. Also, they had a global relationship with Teradata at management level which enabled them to offer a more competitive bid.
SAP HANA
We invested in Hadoop
No
Yes. SAP HANA being the main investment
Â£500k
The same software was used

Q8 - How long did the big data implementation take and what did it entail?
The process is still ongoing but it started by analysing our requirements, what we would like to see and we are currently about to choose a vendor.
1 year (elapsed time)
Testing of ASE systems with Technical Consultant Migration of current licences to new platform Additional installation of new licences Took 6 months
9 months
Implementation is on going - 18 months
6 months, blue print, consultant services, integration, testing, production deployment
n/a
3+ years
We were working on this for over 18 months
Project ongoing for the last 14 months with continued investment.
n/a
We are still implementing our big data strategy across the organisation. We now have the hardware in which we would like to base a "big data analytics" software solution on.
3 months
4 months for the pilot and 4 months for the live environment.
Not finished implementation yet.
nearly a month
It's ongoing having started in March 2013. I know it entailed new h/w and s/w, extensive work from external consultants to prepare the systems and a lot of work was done on data cleansing, archiving. At one point we brought in a specialist on data modelling who helped them and I personally have been part of the technical extensive support team helping out with daily queries.
It takes less than a year depending on your specific systems
We still haven't completed the implementation completely but we been working on it for 15 months.
N/a
Several Months 1. Collect Data, 2 Collate and move data 3. Analyse data
9 months
N/A

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How do you measure financial gains from big data?
Increased speed in matching client and candidate requirements
n/a
Performance increase on client site
Improvements in sales effectiveness, forecasting, marketing campaugns and reduction in sales cycle
Unquantified gains through increased customer retention and loyalty
Book closing, staff reductions
N/A
Results, fans buying tickets / merchandise, tv ad revenue - all resulting from improved performance through Big Data
Not my remit but I guess better reporting
Being able to deliver faster, more efficently, reducing excess stock and being able to plan projects more efficently
Time
n/a
time saved, sales opportunities improvement, shorter sales-cycle
This is still work in progress - hoping to see an increase in utilisation and billable days,
N/A
NA
The company expressed loss of money on poor asset management and some of their parties were very upset with unfulfilled deliveries. The company had to invest in big data to maintain competitiveness as they had been underperforming for several years.
Increased revenue
no direct way of measuring ROI but we see an increase in customer retention
n/a
Saves time on reporting and risk analysis
N/a
Managing an easier landscape and processes reduced labour hours and cost

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How do you measure operational gains from big data?	Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How do you measure internal user gains from big data?
NA	Spped
time spend on prcoesses	user satisfaction
deliver faster and better client experience	ROI
I don't know all the details	I don't know all the details
N/A	N/A
just in time logistics	end user survey
N/A	N/A
n/a	n/a
Not my remit	We needed to be able to make faster decisions which we are now. We are more reactive as a company and more accurate in our forecasting
Headquarters and field are now much more aligned, warehouses run more efficently, less excess, less waste.	Growth of delivers and orders have increased, customer satisfaction has also increased
Time	Time
n/a	n/a
time saved, propductivity	N/A
See above	Utilisation, greater billing accuracy and project profitability
N/A	N/A
NA	NA
I don't know, this is outside my competency.	As above
Improved efficiency in business processes	Improved KPIs & employees satisfaction
No way to measure but we see increased efficiency in the way we operate.	N/A
n/a	n/a
Detect, prevent and remediate financial fraud.	Quick and easy data processing and retrieval
n/a	n/a
N/A	N/A

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How has ROI for big data been presented?
NA
1.5 years
SAP Realtime Strategy Presentation
I don't know all the details
N/A
no
N/A
n/a
Sales figures are up by 2% and we believe that big data has contributed to that
Management is torn. Most see the value while some are still to be convinced as it is a difficult project to quanity
n/a
n/a
N/A
As above.
Currently working through
NA
I can assume that imrpoved financial management is the key objective as they are a financial management company.
Imporoved customer satisfaction
N/A
n/a
faster processing and turnaround of tasks and activities
n/a
N/A

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? Has there been any officially acknowledged efficiency improvements following big data investment? (elaborate please)
Yes, clients have mentioned that speed has improved
Yes - finance - time spend on reporting/budgets- 70%
faster performance with client
Yes the company has seen improvements in the sales effectiveness
Yes - Improved and faster analytics of cloud systems telemetry
yes
N/A
Winning the world cup
It has been hard to quantitfy as so may different contributory factors such as acquisitions, entry into new markets etc
Yes the general consenus is that Big Data has helped through the current financial instability of the market in a period of downturn which started in 2008 with the financial crisis.
n/a
n/a
N/A
Work in progress
N/A
NA
All the contacts I speak to from the customer's sdie seem to be happy with the investment; feedback is positive.
Yes, employees performance has improved tremendously
360Â° view of our customers across all digital channels, faster reporting on digital data
n/a
It is too soon to conclude
n/a
N/A

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? How will big data help in terms of competitiveness and what tip can you give?
We now can respond to client requirements quicker than our completition
The biggest competitive gain is real time sales insights - we now no longer wait for weeks after to see the results of sales campagins but have the flexibility to see results as they come in so we can quickly change our strategy if unsuccessful or boost r
na
I don't know all the details
Most gains will be seen through customer sentiment analysis
faster to market, cost reductions, cost saving led to investment in marketing
N/A
Improve performance going forward
We see it as a strategic tool to ensure that we have accurate insight and that in itself helps us to predict changes to avoid surprises
See Above
Be ahead of the business competitors
n/a
spot trends before the competition, more flexibility and responsiveness
increased customer satisfaction
Insight into business KPI's in real time, unable to do previously - can act much quicker if KPI's aren't being hit.
NA
The customer talks a lot about fast delivery. I don't know the details.
We are able to predict the behaviour of our consumers & the market, helping us to improve on our products & production
By understanding our customers purchase behaviour we will be able to deliver products and services suiting their needs.
n
Saves time on reporting and risk analysis
n/a
N/A

Q9 - Please answer the questions about what value big data has brought to your organization or your customer's organization? Has any new services or products been introduced as an effect of big data adoption? (elaborate please)
NA
no
Yes , services as part of the software acqusition
I don't know all the details
N/A
no
N/A
n/a
Yes - we have been able to enter into new markets to to new found confidence in our forecasting and more accurate estimations of opportunities
We have proposed a couple of options to the client, which are ongoing and under review.
n/a
n/a
N/A
N/A
N/a
NA
Not that I'm aware of.
Yes, with increased efficiency & customer awareness, we have new enhancement products
No
n
Risk management and loan evaluation
n/a
N/A

Q10 - Please answer the technical questions on big data implementation below Have you experienced any internal challenges following adoption of big data analytics , if so, what advise can you give?
Staff Engagement
Employees knew headcount reduction would come as result - introduced negativity
No
I don't know all the details
More awareness of information in non core systems (outside our datawarehouse and ERP)
no
N/A
Translating findings into improvements
Yes - we have had difficulties with skill gaps but this is not my remit.
Yes, we had a difficult time finding the right staff
n/a
No the consensus is that this is an issue we need to tackle
N/A
None to speak of
Commercials - who's budget to come out of.
NA
The customer has struffled with finding the right skills for th eproject, they've used mutlitple 3rd parties which have led to mistakes, redundancy of effort etc.
Yes, it is important to involve all stake-holders with the project from the word go! mix two project management approaches, e.g. lean
No
n
None
n/a
No challenges. That is why it is important to start in a test environment.

Q10 - Please answer the technical questions on big data implementation below What is the estimated time spent on patching/upgrades of big data management system and what advise can you give?
NA
handled by partner in hosted environment
weeks depending on applications
I don't know all the details
Unknown
4 weeks
N/A
n/a
n/a
NA
n/a
n/a
N/A
N/A
n/a
NA
Not applicable.
Yes, it is advisable to implement a big data project in phases, as it saves time & reasource (incremental development). It can also give you time to see benefits as you implement
Dont Know
n
2weeks
n/a
4 days and the advise is to check pre-requisites first

Q10 - Please answer the technical questions on big data implementation below What is the estimated time spent on archiving/recovery of big data system and what advise can you give?	**Q10 - Please answer the technical questions on big data implementation below What is the estimated time spent on migration/testing of big data management system and what advise can you give?**
NA	10 hours every 6 months
handled by partner in hosted environment	handled by partner in hosted environment
2 weekly task	4 months
I don't know all the details	I don't know all the details
Unknown	1 Month
4 weeks, godd consultancy needed with specialist hire to manage	5 months
N/A	N/A
n/a	n/a
n/a	n/a
NA	NA
n/a	n/a
n/a	n/a
N/A	budget an appropriate amount of time
Automated	N/A
n/a	n/a
NA	NA
How long is a piece of string?	They've jsut recently started to get the analytics to work after aprrox 17 months.
It depends on specific systems & size of the organization	It depends on specific systems & size of the organization
Dont Know	1,5 Month
n	n
2weeks	1week
n/a	n/a
N/A	Testing time can be around 3 to 5 days and it is importan to have a proper test plan

Q10 - Please answer the technical questions on big data implementation below What is the estimated time spent on runtime analysis/regression of big data management systems and what advise can you give?	**Q10 - Please answer the technical questions on big data implementation below What is the estimated time spent on runtime disaster recovery of big data management system and what advise can you give?**
4 hours every 6 months	2 hours every 6 months
minimal	handled by partner in hosted environment
na	na
I don't know all the details	I don't know all the details
Unknown	Almost no time
1 month	2 week s
N/A	N/A
n/a	n/a
n/a	n/a
NA	NA
n/a	n/a
n/a	n/a
N/A	N/A
N/A	Automated
n/a	n/a
Na	NA
N/A	N/A
It depends on specific systems & size of the organization	It depends on specific systems & size of the organization
Dont Know	Dont Know
n	n
4days	1week
n/a	n/a
Stress test can take up to 2 days. Just need to short-list the high resource processes and use that as a test.	Proper tested DR must be in place before any project can start

Q10 - Please answer the technical questions on big data implementation below What is the estimated time spent on data/software architecture and coding of big data management system and what advise can you give?	**Q10 - Please answer the technical questions on big data implementation below What were the data integration costs and time spend and what advise can you give? (Ex. cost of ETL)**
NA	They have proved to be negligible
handled by partner in hosted environment	75k
Na	Not relevent
I don't know all the details	I don't know all the details
Unknown	3 Months
2 months, good consultancy required	Â£250k
N/A	N/A
n/a	n/a
n/a	n/a
NA	NA
n/a	n/a
n/a	n/a
N/A	N/A
N/A	3 Weeks
n/a	n/a
NA	NA
N/A	N/A
It depends on specific systems & size of the organization	N/A
Dont Know	Dont know the cost but i took 7 Months
m	n
2weeks	20,000. It has reduced greatly TCO(Total cost of ownership)
n/a	n/a
N/A	N/A

Q10 - Please answer the technical questions on big data implementation below What was the time spend on developing queries and what advise can you give?	Q10 - Please answer the technical questions on big data implementation below What was the cost and time spend on development of analytics applications and what advise can you give?
NA	I would say that it is key to include feedback from employees to ensure greater take up and acceptance of new platform
handled by partner in hosted environment	170k
2 months	100k
I don't know all the details	I don't know all the details
2 Months	6 Months
1 month,	6 months, blueprint to user requirements
N/A	N/A
n/a	n/a
n/a	n/a
NA	NA
n/a	n/a
n/a	n/a
N/A	N/A
On going	on Going
n/a	n/a
NA	NA
N/A	As mentioned above we have only recenlty started to get it operative.
It depends on specific systems & size of the organization	It depends on specific data size, resources & the organization
3 Months	N/A
n	n
2weeks	1week
n/a	n/a
N/A	N/A

Q11 - What do you do to get around systems faultiness, atomicity issues, lack of data consistency, isolation issues, lack of data durability (ACID), caused by big data roll out and what advise can you give?
Platform still relatively new so we have yet to experience this
n/a
NA
I don't know all the details
System faultiness caused by integration of newer Big data systems with our legacy systems and databases.
constant development of data integrity planning and cleaning of data prior
N/A
n/a
This is not my remit. However the DBA's have expressed some data management issues relating to accuracy which is not due to the Big Data project but the readiness of our systems
NA
n/a
n/a
N/A
N/A
N/a
NA
The customer turns to me once in a while regarding questions relating to data consistency, qRFC management, resource management. That's all I know.
The system should have less or no system failures, downtime should be very minimal. We carried out a very well though preliminary research before the start of the implementation. We had qualified personnel, enough resources and great management. We also have efficient redundancy systems
N/A
N
We use Indexed storage and Relational implementations
n/a
No issues faced.