"Management of Large Data: A Research upon Analytics and Better Decision Making"

Vittal S

Lecturer, Mahatma Jyotiba Phule Rohilkhand University, Barely, UP

Abstract – Evolving technologies in the energy and utilities industry, including smart meters and smart grids, can provide companies with unprecedented capabilities for forecasting demand, shaping customer usage patterns, preventing outages, optimizing unit commitment and more. At the same time, these advances also generate unprecedented data volume, speed and complexity.

To manage and use this information to gain insight, utility companies must be capable of high-volume data management and advanced analytics designed to transform data into actionable insights. For example, designing effective demand response programs requires that utilities execute advanced analytics across a combination of data about customers, consumption, physical grid dynamic behavior, generation capacity, energy commodity markets and weather.

Data generated by financial transactions, different types of sensors and meters, social media networks and numerous other sources are increasing exponentially in terms of their volume, variety and velocity. These "3 Vs" are making datasets increasingly difficult to capture, manage and process through conventional means. This phenomenon is known as "big data".

Deriving value out of the huge volumes of data created by users on a day-to-day basis has become popularised by companies like Google and Facebook that are increasingly applying analytics and decision making solutions to capture, manage and process data. In doing so, companies benefit from real-time market intelligence that empowers company decision-making which, in turn, may result in increased revenues and reduced costs.

Businesses benefitting from the double-digit growth in the big data market are currently taking advantage of low barriers to entry for start-ups, whether in terms of infrastructure and capital requirements or the reduced need for big data firms to be located in close proximity to their clients.

----------♦----------

INTRODUCTION

Becoming an analytics-driven organization helps companies reduce costs, increase revenues and improve competitiveness, and this is why business intelligence and analytics continue to be a top priority for CIOs. Many business decisions, however, are still not based on analytics, and CIOs are looking for ways to reduce time to value for deploying business intelligence solutions so that they can expand the use of analytics to a larger audience of users.

Companies are also interested in leveraging the value of information in so-called big data systems that handle data ranging from high-volume event data to social media textual data. This information is largely untapped by existing business intelligence systems, but organizations are beginning to recognize the value of extending the business intelligence and data warehousing environment to integrate, manage, govern and analyze this information.

Big data is a popular buzzword, but it is important to realize that this data comes in many shapes and sizes. It also has many different uses – real-time fraud detection, web display advertising and competitive analysis, call center optimization, social media and sentiment analysis, intelligent traffic management and smart power grids, to name just a few. All of these analytical solutions involve significant (and growing) volumes of both multi-structured3 and structured data.

Most of these analytical solutions were not possible

previously because they were too costly to implement, or because analytical processing technologies were not capable of handling the large volumes of data involved in a timely manner. In some cases, the required data simply did not exist in an electronic form.

New and evolving analytical processing technologies now make possible what was not possible before.

Big data involves more than simply the ability to handle large volumes of data. Instead, it represents a wide range of new analytical technologies and business possibilities. The challenge is how to deploy these technologies and manage the many extreme analytical processing workloads involved, while at the same time providing faster time to value.

One may question whether the surveyed firms are as "data-driven" as their executives say. The research also shows that organisations are struggling with the enormous volumes of data and often with poor quality data, and many are struggling to free data from organisational silos. The same share of respondents who say their firms are data-driven also say there is not enough of a "big data culture" in their organisation; almost as many – 55% – say that big data management is not viewed strategically at senior levels of their organisation.

When it comes to integrating big data with executive decision-making, there is clearly a long road to travel before the results match the optimism. This report will examine how far down that road firms in different industries and regions are, and will shed light on the steps some organisations are taking to make big data a critical success factor in the decision-making process.

THE ISSUES OF LARGE DATA

An enterprise data warehouse can be used to handle big data and extreme workloads, but often it is more efficient to preprocess the data before loading it into the warehouse. Event data from hardware sensors, for example, has more business value if it is filtered and aggregated before using it for analysis. Usually the raw event data is not required for historical purposes, and storage and processing costs are reduced if it is not kept in the warehouse.

The best way of handling big data and supporting the extreme processing involved is to deploy optimized hardware and software solutions for processing different types of big data workloads, and then combine these solutions with the existing enterprise data warehouse to create an integrated information supply chain (see Figure 1). The objectives of an information supply chain are to consume and integrate the many varieties of raw source data that exist in organizations, analyze that data, and then deliver the analytical results to business users. This information supply chain enables the data component of what IBM calls Smarter Computing.

Supporting extreme workloads is not a new challenge for the computing industry. Business transaction processing systems have supported extreme transaction workloads ever since the advent of data processing. As new business needs arose, organizations employed custom and optimized transaction processing systems to handle application workloads that pushed the boundaries beyond what could be handled by more generalized technologies. Airline reservation systems, bank ATM systems, retail point-of-sale terminals, financial trading systems and mobile phone systems, are all examples of these types of applications. More recently, applications for handling sensor networks that track items with RFID tags can be added to the list.

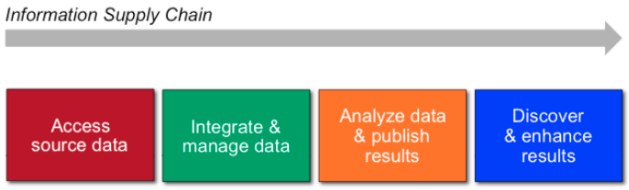

Figure 1 : The information supply chain.

In recent years there has been a similar trend toward analytical processing reaching the performance limitations of more generalized systems. Today, multi-terabyte data warehouses are no longer the exception and analytical processing workloads are becoming increasingly more complex. The result is that once again organizations

require optimized systems to meet the challenges of extreme workloads. The industry response has been to offer packaged hardware and software solutions that are optimized for analytical processing.

Satisfying business agility requirements is another important factor in supporting analytical processing. In today's fast paced business environment, organizations need to make faster decisions, and this agility is important to business success. In the case of fraud detection, for example, real time action is required. Not all business decisions have to be made in real time, but for many organizations the ability to act in a few seconds or minutes, rather than hours or days, can be a significant financial and competitive advantage.

GETTING LARGE DATA TO LARGE UTILIZE

"A lot of people will say data is important to their business, but I think it's incredibly important to healthcare and it's probably getting more and more important," says Lori Beer executive vice president of executive enterprise services at WellPoint, an American healthcare insurer. Ms Beer compares data in healthcare with "oxygen"—without it, the organization would die.

There is near consensus across industries as to which big data sets are most valuable. Fully 69% of survey respondents agree that "business activity data" (eg, sales, purchases, costs) adds the greatest value to their organization. The only notable exception is consumer goods and retail where point-of-sale data is deemed to be the most important (cited by 71% of respondents). Retailers and consumer goods firms are arguably under more pressure than other industries to keep their prices competitive. With smartphone apps such as Red Laser and Amazon's Price Check, customers can scan a product's barcode in-store and immediately find out if the product is available elsewhere for less.

To keep customers loyal, retailers have to target customers with personalized loyalty bonuses, discounts and promotions. Today, most large supermarkets micro-segment customers in real time and offer highly targeted promotions at the point of sale.

Office documentation (emails, document stores, etc) is the second most valued data set overall, favored by 32% of respondents. Of the other major industries represented in the survey, only healthcare, pharmaceuticals and biotechnology differ on their second choice. Here social media are viewed as the second most valuable data set, possibly because reputation is vitally important in this sector, and "sentiment analysis" of social media is a quick way to identify shifting views towards drugs and other healthcare products.

Over 40% of respondents agree that using social media data for decision-making has become increasingly important, possibly because they have made organizations vulnerable to "brand damage". Social media are often used as an early warning system to alert firms when customers are turning against them. In December 2011 it took Verizon Wireless just one day to make the decision to withdraw a $2 "convenience charge" for paying bills with a smartphone, following a social media-led consumer backlash. Customers used Twitter and other social media to express their anger at the charge. Verizon Wireless was prompt in responding to the outcry, possibly forestalling customer defection to rival mobile operators.

A possible reason for this is that today's business intelligence tools are good at aggregating and analyzing structured data whilst tools for unstructured data are predominantly targeted at providing access to individual documents (eg search and content management). It may be a while before the more advanced unstructured data tools, such as text analytics and sentiment analysis, which can aggregate and summarise unstructured content, become mass market. This may be why 40% of respondents say they have too much unstructured data to support decision-making, as opposed to just 7% who feel they have too much structured data.

THE PROSPECTIVE OF BIG DATA

The big data market is at a nascent stage and is expected to develop as organizations seek to enhance their competitive advantage. In doing so, firms seek to better understand the ever-growing amounts of data, through analytic and decision making solutions. Employing this software may involve a variety of techniques, technologies and visualization tools. Today, there is a growing market of companies offering hardware, software and professional services solutions.

The benefits of data-driven analytics - Business intelligence is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.

As the pace of business increases with globalization and the spread of communication technologies, important management decisions need to be taken more quickly. Detailed big data analytics can provide critical business intelligence on customer behavioral trends or consumer profiling to a level of detail never before imagined. As a result, analytics for decision making is becoming a valuable tool for both private and public entities. There is an increasing demand for business intelligence to be provided in real-time in order to be able to react to it as rapidly as possible.

Understanding the big data market - Quantifying the exact size of the big data analytics and decision making market is difficult and the estimates vary amongst recent reports. According to the International Data Corporation (IDC), the global revenue of players involved in big data grew by 35% to €6.1 billion in 2012 and is expected to continue rising at a similar rate until 20169. Whilst, according to the Economist, "In recent years Oracle, IBM, Microsoft and SAP between them have spent more than €11.3 billion on buying software firms specializing in data management and analytics. This industry is estimated to be worth more than €56 billion and growing at almost 10% a year, roughly twice as fast as the software business as a whole".

The application of big data analytic solutions is also believed to offer €250 billion in annual savings to Europe's public sector administration, as a result of improved productivity and increased efficiency and effectiveness.

Applications of big data analytics tools can already be found across a wide range of sectors including retail, utilities, health care, media and telecoms. A wide range of different techniques and technologies have been developed to collect, process, analyse and visualise big data in all these industries.

Applying the prospective of big data analytics - Analytics and decision making solutions offer a variety of benefits, but these fundamentally can be boiled down to the provision of business intelligence. Clients can benefit from analytic solutions of their internal data and benefit from reductions in costs or help themselves reduce inventory. Certain clients benefit from analytics solutions which take into account external data which can, for example, help retailers tweak their offerings based on newly discovered trends in consumer behaviour, adjust advertising strategies to suit identified consumer profiles, or even discover new market niches.

The predictive analytics offered by AiRPX benefitted its client airline companies in that it allowed them to reduce their costs through multiple avenues. Its top-down approach gave users an overview of the plane in its entirety providing a complete evaluation of the plane's status in real-time and in-flight. The solution is compatible with planes of different brands, allowing the consolidation of plane statuses for their entire fleets rather than per-brand. The system also being predictive, allows for a reduction in both maintenance equipment and staff, as plane maintenance can be foreseen and resources allocated in advance. It also improves aircraft availability and increases schedule reliability.

Market adoption - The interviewed companies were for the most part created to capitalize on a wide variety of demands and applications in different sectors. What their solutions have in common is that they help improve business performance through a better understanding of data and the ability to make more relevant and timely decisions. Wipro Promax Analytic Solutions, Digital Route and AiRPX help their clients improve their business performance by providing analytic solutions of their customers' internal data. Trendiction, Neodata and Quiterian offer their customers the benefits of analytics services from structured and unstructured external data.

Client perspectives - There is still a cultural barrier to be overcome before the trend becomes widespread. This is because embracing data-driven decision making involves moving away from conventional decision making processes. These conventional processes involve the preparation of reports provided by IT staff after their own analytics process. On the one hand there is a degree of skepticism from the decision makers themselves about this new data-driven decision making. On the other hand, these new solutions would effectively cut out the "middle-man" and empower company management to make decisions themselves. Obviously such a disruptive change in process takes time before the market accepts it.

There is also a fear that the implementation of the new solutions offered within this trend might be highly disruptive. A company that invested in software to provide their decision makers with both the means to analyses big data and to present the salient results in real-time may mean that it would make their current analytics system obsolete. This would effectively make a prior investment worthless, and is a contributing factor to the reluctance of companies to uptake the trend.

HANDLING AND EXAMINING BIG DATA

The main challenges of big data and extreme workloads are data variety and volume, and analytical workload complexity and agility. Each of these can be viewed from the perspective of the information supply chain.

Input to the information supply chain consists of the raw source data required for analysis. For the past two decades most business analytics have been created using structured data extracted from operational systems and consolidated into a data warehouse. Big data dramatically increases both the number of data sources and the variety and volume of data that is useful for analysis. A high percentage of this data is often described as multi-structured to distinguish it from the structured operational data used to populate a data warehouse. In most organizations, multi-structured data is growing at a considerably faster rate than structured data.

There are two main techniques for analyzing big data – the store and analyze approach, and the analyze and store approach.

Store and Analyze Approach - The store and analyze approach integrates source data into a consolidated data store before it is analyzed. This approach is used by a traditional data warehousing system to create data analytics. In a data warehousing system, the consolidated data store is usually an enterprise data warehouse or data mart managed by a relational or multidimensional DBMS. The advantages of this approach are improved data integration and data quality management, plus the ability to maintain historical information. The disadvantages are

additional data storage requirements and the latency introduced by the data integration task.

Two important big data trends for supporting the store and analyze approach are relational DBMS products optimized for analytical workloads (often called analytic RDBMSs, or ADBMSs) and non-relational systems (sometimes called NoSQL systems) for processing multi-structured data. A non-relational system can be used to produce analytics from big data, or to preprocess big data before it is consolidated into a data warehouse. Certain vendors in the search and content management marketplaces also use the store and analyze approach to create analytics from index and content data stores.

Analytic RDBMSs (ADBMSs) - An analytic RDBMS is an integrated solution for managing data and generating analytics that offers improved price/performance, simpler management and administration, and time to value superior to more generalized RDBMS offerings.

Performance improvements are achieved through the use of massively parallel processing, enhanced data structures, data compression, and the ability to push analytical processing into the DBMS. ADBMSs can be categorized into three broad groups:4 packaged hardware and software appliances, software-only platforms, and cloud-based solutions.

Packaged hardware and software appliances fall into two sub-groups: purpose built appliances and optimized hardware/software platforms. The objective in both cases is to provide an integrated package that can be installed and maintained as a single system. Depending on the vendor, the dividing line between the two subgroups is not always clear, and this is why in this article they are both categorized as appliances.

A purpose-built appliance is an integrated system built from the ground up to provide good price/performance for analytical workloads. This type of appliance enables the complete configuration, from the application workload to the storage system used to manage the data, to be optimized for analytical processing. It also allows the solution provider to deliver customized tools for installing, managing and administering the integrated hardware and software system.

Many of these products were developed initially by small vendors and targeted at specific high-volume business area projects that are independent of the enterprise data warehouse. As these appliances have matured and added workload management capabilities, their use has expanded to handle mixed workloads and in some cases support smaller enterprise data warehouses. Large and well-established vendors have acquired several of these solutions. An example is the IBM Netezza TwinFin appliance.

The success of these purpose-built appliances led to more traditional RDBMS vendors building packaged offerings by combining existing products. This involved improving the analytical processing capabilities of the software and then building integrated and optimized hardware and software solutions. These solutions consist of optimized hardware/software platforms designed for specific analytical workloads.

The level of integration and optimization achieved varies by vendor. In some cases, the vendor may offer a choice of hardware platform. An example of this type of approach is the IBM Smart Analytics System.

The IBM Smart Analytics System is intended for extending the performance and scalability of an enterprise data warehouse, whereas the IBM Netezza TwinFin appliance is used to maintain a separate data store in situations where it is unnecessary, impractical, or simply not cost effective to integrate the data into an enterprise data warehouse for processing. The IBM Netezza TwinFin is also used to analyze data extracted from an enterprise data warehouse, either to offload an extreme processing workload, or to create a sandbox for experimental use and/or complex ad hoc analysis.

A software-only platform is a set of integrated software components for handling analytical workloads. These platforms often make use of underlying open source software products and are designed for deployment on low-cost commodity hardware. The tradeoff for hardware portability is the inability of the product to exploit the performance and management capabilities of a specific hardware platform. Some software platforms are available as virtual images, which are useful for evaluation and development purposes, and also for use in cloud-based environments.

Cloud-based solutions offer a set of services for supporting data warehousing and analytical application processing. Some of these services are offered on public clouds, while others can be used in-house in private cloud environments. The underlying software and hardware environment for these cloud-based services may be custom built, employ a packaged hardware and software appliance, or use the

capabilities of a software-only platform. The role of cloud computing for business intelligence and data warehousing is discussed in more detail later in this paper.

Non-Relational Systems - A single database model or technology cannot satisfy the needs of every organization or workload. Despite its success and universal adoption, this is also true for RDBMS technology. This is especially true when processing large amounts of multi-structured data and this is why several organizations with big data problems have developed their own non-relational systems to deal with extreme data volumes. Web-focused companies such as Google and Yahoo that have significant volumes of web information to index and analyze are examples of organizations that have built their own optimized solutions. Several of these companies have placed these systems into the public domain so that they can be made available as open source software.

Non-relational systems are useful for processing big data where most of the data is multi-structured. They are particularly popular with developers who prefer to use a procedural programming language, rather than a declarative language such as SQL, to process data.6 These systems support several different types of data structures including document data, graphical information, and key-value pairs.

One leading non-relational system is the Hadoop distributed processing system from the open source Apache Software Foundation. Apache defines Hadoop as "a framework for running applications on a large hardware cluster built of commodity hardware." This framework includes a distributed file system (HDFS) that can distribute and manage huge volumes of data across the nodes of a hardware cluster to provide high data throughput. Hadoop uses the MapReduce programming model to divide application processing into small fragments of work that can be executed on multiple nodes of the cluster to provide massively parallel processing. Hadoop also includes the Pig and Hive languages for developing and generating MapReduce programs. Hive includes HiveQL, which provides a subset of SQL.

CONCLUSION

Most of the executives polled for this report are also optimistic about the cost reductions and efficiencies that can be had from automating decision-making using big data. While there is certainly much scope for decision-automation in heavy industry, especially in areas such as energy production and distribution ("smart grids") and transportation ("smart cars", etc), excessive automation of business processes can hamper flexibility. Besides, the growing post-financial-crisis regulation calling for greater accountability requires humans to ultimately make the decisions. Prosecutors cannot put an algorithm in the dock.

The financial crisis has also led to calls for greater transparency. As the survey shows, people are increasingly wary of business decisions based purely on intuition and experience. Even if a sizeable minority agree that business managers have a better feel for business decisions than analytics will ever provide, managers will increasingly need to show how they arrived at their decision. And big data will provide a post-decision review—was it a good decision or not? As one of the survey participants puts it, using big data for decision-making will lead to "better decisions; better consensus; better execution".

Big data adds several new high-volume data sources to the information supply chain. Several new and enhanced data management and data analysis approaches help the management of big data and the creation of analytics from that data. The actual approach used will depend on the volume of data, the variety of data, the complexity of the analytical processing workloads involved, and the responsiveness required by the business. It will also depend on the capabilities provided by vendors for managing, administering, and governing the enhanced environment. These capabilities are important selection criteria for product evaluation. An enhanced information supply chain, however, involves more than simply implementing new technologies. It requires senior management to understand the benefits of smarter and timelier decision making, and what IBM calls Smarter Computing. It also requires the business to make pragmatic decisions about the agility requirements for analyzing data and producing analytics given tight IT budgets. The good news is that many of the technologies outlined in this paper not only support smarter decision making, but also provide faster time to value.

REFERENCES

- Brynjolfsson, Erik, "Riding the Rising Information Wave– Are you swamped or swimming?", MIT Sloan Experts.
- Brynjolfsson, Erik, Hitt, Lorin M. and Kim, Heekyung Hellen, "Strength in Numbers: How Does Data-Driven Decision making Affect Firm Performance?" (April 22, 2011).
- Davenport, T.H. and Patil, D.J. 2012. Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review.
- Economist Intelligence Unit. 2012. Harnessing a game-changing asset.
- Gartner, 'Gartner IT Glossary – Big data', available

at: http://www.gartner.com/it-glossary/big-data/

- Gartner, 'Gartner IT Glossary - Business Intelligence (BI)'.

- International Data Corporation (IDC), 2012, Worldwide Big Data Technology and Services 2012–2016 Forecast.

- Kelly, J. 2013. Big Data Market Size and Vendor Revenues. Wikibon Article.

- McKinsey Global Institute, 2011, Big data, the next frontier for innovation, competition and productivity.

- McKinsey Global Institute, 2011, Big data, the next frontier for innovation, competition and productivity.

- PwC Australia, 2012, Big Data – The next frontier for innovation.

- Yiu, C. 2011. The Big Data Opportunity. Policy Exchange

-