BIG DATA ANALYTICS "Better Decision Making"

Wale Ayandiran W.

The Federal Polytechnic Offa

Offa Kwara State

P.M.B 420.

Email: walexrvp@gmail.com

@walebrity121

November 06, 2015

ABSTRACT

The term Big Data is used almost anywhere these days; from news articles to professional magazines, from tweets to YouTube videos and blog discussions. The term coined by Roger Magoulas from O'Reilly media in 2005, refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools—due to their size, but also their complexity. Big Data can be seen in the finance and business where enormous amount of stock exchange, banking, online and on-site purchasing data flows through computerized systems every day and are then captured and stored for inventory monitoring, customer behavior and market behavior.

There is so much enthusiasm currently about the possibilities created by new and extensive sources of Data to better understand, manage and aid decision making. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data (called big data) to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Big data analysts basically want the knowledge that comes from analyzing the data. More Processed amount of Data leads to better and more accurate decision making.

Keywords: *Big Data, Big data Analytics, Volume, Velocity, Variety, Complexity, Cloud computing, Decision, Data Mining, Knowledge based Computing.*

INTRODUCTION

What is Big Data?

According to Wikipedia, (the free encyclopedia) Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate to handle. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.

As far back as 2001, industry analyst Doug Laney (currently with Gartner) articulated the now mainstream definition of big data as the three Vs of big data: volume, velocity and variety.

Research trends articulated that the term was coined by Roger Magoulas from O'Reilly media in 2005, refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools—due to their size, but also their complexity.

A SAS Institute Inc. White-paper Defined Big Data as "a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making".

Characteristics of Big Data

Big data can be described by the following characteristics:

Volume

The quantity of generated data is important in this context. The size of the data determines the value and potential of the data under consideration, and whether it can actually be considered big data or not. The name 'big data' itself contains a term related to size, and hence the characteristic.

Variety

The type of content, and an essential fact that data analysts must know. This helps people who are associated with and analyze the data to effectively use the data to their advantage and thus uphold its importance.

Velocity

In this context, the speed at which the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development.

Variability

The inconsistency the data can show at times—which can hamper the process of handling and managing the data effectively.

Veracity

The quality of captured data, which can vary greatly. Accurate analysis depends on the veracity of source data.

Complexity

Data management can be very complex, especially when large volumes of data come from multiple sources. Data must be linked, connected, and correlated so users can grasp the information the data is supposed to convey.

BIG DATA ARCHITECTURE

In 2000, Seisint Inc. developed a C++-based distributed file-sharing framework for data storage and query. The system stores and distributes structured, semi-structured, and unstructured data across multiple servers. Users can build queries in a modified C++ called ECL. ECL uses an "apply schema on read" method to infer the structure of stored data at the time of the query. In 2004, LexisNexis acquired Seisint Inc. and in 2008 acquired ChoicePoint, Inc. and their high-speed parallel processing platform. The two platforms were merged into HPCC Systems and in 2011, HPCC was open-sourced under the Apache v2.0 License. Currently, HPCC and Quantcast File System are the only publicly available platforms capable of analyzing multiple exabytes of data.

In 2004, Google published a paper on a process called MapReduce that used such an architecture. The MapReduce framework provides a parallel processing model and associated implementation to process huge amounts of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). The framework was very successful, so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named Hadoop.

MIKE2.0 is an open approach to information management that acknowledges the need for revisions due to big data implications identified in an article titled "Big Data Solution Offering".The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records.

Recent studies show that the use of a multiple-layer architecture is an option for dealing with big data. The Distributed Parallel architecture distributes data across multiple

processing units, and parallel processing units provide data much faster, by improving processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end user by using a front-end application server.

Big Data Analytics for Manufacturing Applications can be based on a 5C architecture (connection, conversion, cyber, cognition, and configuration).

The data lake allows an organization to shift its focus from centralized control to a shared model to respond to the changing dynamics of information management. This enables quick segregation of data into the data lake, thereby reducing the overhead time.

Big Data Technologies

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. A 2011 McKinsey report suggests suitable technologies include;

A/B testing

Crowdsourcing

Data fusion and integration,

Genetic algorithms,

Machine learning,

Natural language processing,

Signal processing,

Simulation,

Time series analysis

Visualisation.

Multidimensional big data can also be represented as tensors, which can be more efficiently handled by tensor-based computation, such as multilinear subspace learning. Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data mining, distributed file systems, distributed databases, cloud-based infrastructure (applications, storage and computing resources) and the Internet.

Some but not all MPP relational databases have the ability to store and manage petabytes of data. Implicit is the ability to load, monitor, back up, and optimize the use of the large data tables in the RDBMS.

The practitioners of big data analytics processes are generally hostile to slower shared storage, preferring direct-attached storage (DAS) in its various forms from solid state drive (SSD) to high capacity SATA disk buried inside parallel processing nodes. The perception of shared storage architectures—Storage area network (SAN) and Network-attached storage (NAS) —is that they are relatively slow, complex, and expensive. These qualities are not consistent with big data analytics systems that thrive on system performance, commodity infrastructure, and low cost.

Real or near-real time information delivery is one of the defining characteristics of big data analytics. Latency is therefore avoided whenever and wherever possible. Data in memory is good—data on spinning disk at the other end of a FC SAN connection is not. The cost of a SAN at the scale needed for analytics applications is very much higher than other storage techniques.

There are advantages as well as disadvantages to shared storage in big data analytics, but big data analytics practitioners as of 2011 did not favor it

BIG DATA ANALYTICS

Big Data Analysis

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills. This brings us to the question: what exactly is Big data Analytics?

Big data analytics refers to the process of collecting, organizing and analyzing large sets of data (called big data) to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Big data analysts basically want the knowledge that comes from analyzing the data.

Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on. Scientists, business executives, practitioners of media, and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter

limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research.

Data-driven decision-making is now being recognized broadly, and there is a growing enthusiasm for the notion of “Big Data”. While the promise of Big Data is real, for example; it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009...there is currently a wide gap between its potential and its realization.

We are awash in flood of data today, in a broad range of application areas, data is being collected at an unprecedented scale. Decisions that were previously based on guessbook or painstakingly constructed models of reality, can now be made based on data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences and physical sciences.

The Idea is to use analytics to determine relevance instead of always putting all data in the storage before analyzing.

Challenges of Big Data Analytics

Challenges to consider:

Many organizations are concerned that the amount of amassed data is becoming so large that it is difficult to find the most valuable pieces of information.

What if your data volume gets so large and varied you don't know how to deal with it?

Do you store all your data?

Do you analyze it all?

How can you find out which data points are really important?

How can you use it to your best advantage?

Until recently, organizations have been limited to using subsets of their data, or they were constrained to simplistic analyses because the sheer volumes of data overwhelmed their processing platforms. But, what is the point of collecting and storing terabytes of data if you can't analyze it in full context, or if you have to wait hours or days to get results? On the other hand, not all business questions are better answered by bigger data. You now have two choices:

Incorporate massive data volumes in analysis. If the answers you're seeking will be better provided by analyzing all of your data, go for it. High-performance technologies that extract value from massive amounts of data are here today. One approach is to apply high-performance analytics to analyze the massive amounts of data using technologies such as grid computing, in-database processing and in-memory analytics.

Determine upfront which data is relevant. Traditionally, the trend has been to store everything (some call it data hoarding) and only when you query the data do you discover what is relevant. We now have the ability to apply analytics on the front end to determine relevance based on context. This type of analysis determines which data should be included in analytical processes and what can be placed in low-cost storage for later use if needed.

Benefits of Big Data

Enterprises are increasingly looking to find actionable insights into their data. Many big data projects originate from the need to answer specific business questions. With the right big data analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management.

Benefits of big data cannot be over emphasized without highlighting some of the successful implementation of big data in some enterprise. This includes;

UPS: is no stranger to big data, having begun to capture and track a variety of package movements and transactions as early as the 1980s.The company now tracks data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. The company stores more than 16 petabytes of data. Much of its recently acquired big data, however, comes from telematics sensors in more than 46,000 vehicles. The data on UPS trucks, for example, includes their speed, direction, braking and drive train performance. The data in not only used to monitor daily performance, but to drive a major redesign of UPS drivers' route structures. This initiative, called ORION (On-Road Integration Optimization and Navigation), is arguably the world's largest operations research project. It also relies heavily on online map data, and will eventually reconfigure a driver's pickups and drop-offs in real time. The project has already led to savings in 2011 of more than 8.4 million gallons of fuel by cutting 85 million miles off of daily routes. UPS estimates that saving only one daily mile per driver saves the company $30 million, so the overall dollar savings are substantial. The company is also attempting to use data and analytics to optimize the efficiency of its 2,000 aircraft flights per day.

Manufacturing: Based on TCS 2013 Global Trend Study, improvements in supply planning and product quality provide the greatest benefit of big data for manufacturing. Big data provides an infrastructure for transparency in manufacturing industry, which is the ability to unravel uncertainties such as inconsistent component performance and availability. Predictive manufacturing as an applicable approach toward near-zero downtime and transparency requires vast amount of data and advanced prediction tools for a systematic process of data into useful information.[65] A conceptual framework of predictive manufacturing begins with data acquisition where different type of sensory data is available to acquire such as acoustics, vibration, pressure, current, voltage and controller data. Vast amount of sensory data in addition to historical data construct the big data in manufacturing. The generated big data acts as the input into predictive tools and preventive strategies such as Prognostics and Health Management (PHM).

Technology:

eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay's 90PB data warehouse

Facebook handles 50 billion photos from its user base.

As of August 2012, Google was handling roughly 100 billion searches per month.

Oracle NoSQL Database has been tested to past the 1M ops/sec mark with 8 shards and proceeded to hit 1.2M ops/sec with 10 shards.

Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.

CLOSING THOUGHTS

Big Data why it should matter to you

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And big data may be as important to business – and society – as the Internet has become. Why? More data may lead to more accurate analyses.

More accurate analyses may lead to more confident decision making. And better decisions can mean greater operational efficiencies, cost reductions and reduced risk.

The real issue is not that you are acquiring large amounts of data. It's what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyze it to find answers that enable cost reductions, time reductions, new product development and optimized offerings, and smarter business decision making.

For instance, by combining big data and high-powered analytics, it is possible to:

Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.

Optimize routes for many thousands of package delivery vehicles while they are on the road.

Analyze millions of SKUs to determine prices that maximize profit and clear inventory.

Generate retail coupons at the point of sale based on the customer's current and past purchases.

Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.

Recalculate entire risk portfolios in minutes.

Quickly identify customers who matter the most.

Use clickstream analysis and data mining to detect fraudulent behavior.

Where is big data coming from?

Before you begin to make sense of your data, it’s important to know its origins. The sources of big data are increasing every year, but they generally fall into one of three categories.

Streaming data. Also called the Internet of Things, this includes data that reaches your IT systems from a web of connected devices. Your organization can analyze this data as it arrives and make decisions on what data to keep, what not to keep and what requires further analysis.

Social media data. The data on social interactions is an increasingly attractive set of information, particularly for marketing, sales and support functions. This data is often in unstructured or semi-structured forms, so besides the sheer size of the data, it poses a unique challenge when consuming and analyzing this information.

Publicly available sources. Massive amounts of data is available through open data sources like US government's data.gov, the CIA World Factbook or the European Union Open Data Portal. Learn how SAS is helping people visualize 300+ million rows of global UN trade data.

Big Data in Big Companies: How New

Big data burst upon the scene in the first decade of the 21st century, and the first organizations to embrace it were online and startup firms. Arguably, firms like Google, eBay, LinkedIn, and Facebook were built around big data from the beginning. They didn't have to reconcile or integrate big data with more traditional sources of data and the analytics performed upon them, because they didn't have those traditional forms. They didn't have to merge big data technologies with their traditional IT infrastructures because those infrastructures didn't exist. Big data could stand alone, big data analytics could be the only focus of analytics, and big data technology architectures could be the only architecture. Consider, however, the position of large, well-established businesses. Big data in those environments shouldn't be separate, but must be integrated with everything else that's going on in the company.

Analytics on big data have to coexist with analytics on other types of data. Hadoop clusters have to do their work alongside IBM mainframes. Data scientists must somehow get along and work jointly with mere quantitative analysts.

Big data may be new for startups and for online firms, but many large firms view it as something they have been wrestling with for a while. Some managers appreciate the innovative nature of big data, but more find it "business as usual" or part of a continuing evolution toward more data. They have been adding new forms of data to their systems and models for many years, and don't see anything revolutionary about big data. Put another way, many were pursuing big data before big data was big. When these managers in large firms are impressed by big data, it's not the "bigness" that impresses them. Instead it's one of three other aspects of big data: the lack of structure, the opportunities presented, and low cost of the technologies involved. This is consistent with the results from a survey of more than fifty large companies by NewVantage Partners in 2012. It found, according to the survey summary:

"It's About Variety, not Volume: The survey indicates companies are focused on the variety of data, not its volume, both today and in three years. The most important goal and potential reward of Big Data initiatives is the ability to analyze diverse data sources and new data types, not managing very large data sets"

Firms that have long handled massive volumes of data are beginning to enthuse about the ability to handle a new type of data voice or text or log files or images or video. A retail bank, for example, is getting a handle on its multi-channel customer interactions for the first time by analyzing log files. A hotel firm is analyzing customer lines with video analytics. A health insurer is able to better predict customer dissatisfaction by analyzing speech-to-text data from call center recordings. In short, these companies can have a much more complete picture of their customers and operations by combining unstructured and structured data. There are also continuing if less dramatic advances from the usage of more structured data from sensors and operational data-gathering devices. Companies like GE, UPS, and Schneider National are increasingly putting sensors into things that move or spin, and capturing the resulting data to better optimize their businesses. Even small benefits provide a large payoff when adopted on a large scale. GE estimates that a 1% fuel reduction in the use of big data from aircraft engines would result in a $30 billion savings for the commercial airline industry over 15 years. Similarly, GE estimates that a 1% efficiency improvement in global gas-fired power plant turbines could yield a $66 billion savings in fuel consumption. UPS has achieved similarly dramatic savings (see the“Big Data at UPS” case study) through better vehicle routing.

CONCLUSION

Big data is not just about helping an organization be more successful, to market more effectively or improve business operations. It reaches to far more socially significant issues as well. Could we have foreseen the mortgage meltdown, the financial institution crisis and the recession, if only we had gotten our arms around more data and done more to correlate it? Could we trim millions of dollars in fraud from government programs and financial markets? Could we improve the quality and cost of health care and save lives?

Even though it hasn't been long since the advent of big data, these attributes add up to a new era. It is clear from our research that large organizations across industries are joining the data economy. They are not keeping traditional analytics and big data separate, but are combining them to form a new synthesis. Some aspects of Analytics 3.0 will no doubt continue to emerge, but organizations need to begin transitioning now to the new model. It means change in skills, leadership, organizational structures, technologies, and architectures. It is perhaps the most sweeping change in what we do to get value from data since the 1980s. It's important to remember that the primary value from big data comes not from the data in its raw form, but from the processing and analysis of it and the insights, products, and services that emerge from analysis. The sweeping changes in big data technologies and management approaches need to be accompanied by similarly dramatic shifts in how data supports decisions and product/service innovation. There is little doubt that analytics can transform organizations, and the firms that lead the 3.0 charge will seize the most value.

REFERENCES

Big Data: The Next Frontier for Innovation, Competition, and Creativity," McKinsey Global Institute, 2011

Google driverless car: https://en.wikipedia.org/wiki/Google_driverless_car

Hilbert, Martin (2013) Big Data for development. A review of promises and challenges . Development Policy review.

IBM Big Data – What is Big Data https://www.ibm/en/us/IBM Big Data – What is Big Data – United States.html

SAS Institute inc. - What is Big Data? Big data what it is and Why it matters. http://www.sas.com/en/insights/what-is-big-data%3F|SAS.html

SAS Institute inc. White-paper Big Data meets Big Data Analytics http://www.sas.com/en/insights/big-data-meets-big-data-analytics-105777.pdf

Gali Halexi, MLS, PHD, & Dr Henk F, Moed. (2012) The evolution of Big data as a Research and Scientific topic.

http://www.researchtrends.com/The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature - Research Trends.html

Luis M.A Bettencourt (september 17, 2013) The Use of Big Data in Cities

Webopedia Definition – What is Big Data Analytics?

http://www.webopedia.com/TERM/B/Big-Data-Analytics-Expert-Predictions.html

https://en.wikipedia.org/wiki/Big_data – wikipedia, the free wncyclopedia.html