International Journal of Modern Computer Science and Applications (IJMCSA) ISSN: 2321-2632 (Online)
Volume No.-3, Issue No.-2, April, 2015

Big Data Analytics – Business Benefits and Technical Challenges

Neerja Kulkarni
Symbiosis Centre for Information Technology
Pune, India
E-mail: neerja2205@gmail.com

Dr. Priti Puri, Assistant Professor
Symbiosis Centre for Information Technology
Pune, India
E-mail: priti@scit.edu

Abstract: Since Big Data is a vague and new concept, it is necessary to know what big data is and what big data analytics is. Big Data is defined as: "extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions."
Big data analytics refers to: "the process of examining large data sets containing a variety of data types -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information."
The objective of this paper is exploring more about big data and big data analytics taking into considerations its business benefits and technical challenges faced in Industries. Problem analysis is done particularly by targeting speed and security concern while handling big data.

Keywords: Big Data, Security, Hadoop, BI

I. INTRODUCTION

The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid components of big data analytics applications.

Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases. These technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems. In some cases, Hadoop clusters and NoSQL systems are being used as landing pads and staging areas for data before it gets loaded into a data warehouse for analysis, often in a summarized form that is more conducive to relational structures. Increasingly though, big data vendors are pushing the concept of a Hadoop data lake that serves as the central repository for an organization's incoming streams of raw data. In such architectures, subsets of the data can then be filtered for analysis in data warehouses and analytical databases, or it can be analyzed directly in Hadoop using batch query tools, stream processing software and SQL on Hadoop technologies that run interactive, ad hoc queries written in SQL. [4]

II. OBJECTIVES

Improving Organization's Performance: Responding to the growing importance of data-based initiatives on business strategy and success, this program investigates the critical role that business analytics play in decision making and

product/service strategies at the highest levels. As a participant, we can explore the proper balance between analytical and non-analytical skills and emerge better preparation to make decisions that build competitive advantage for our organization. Moreover, we can gain an understanding of the role that big data plays in organization's strategy.

Taking our Skills to the Next Level

- Establish and sustain advanced analytics capabilities in the organization
- Attract and develop the talent who will take our company's analytics to the next level
- Understand how analytics and big data affect various functions including marketing and the supply chain, both now and in the future
- Appreciate the impact of analytics and big data on the information industry and the external ecosystem for analytical and data services

III. METHODOLOGY

Work Outline:

1. Firstly understanding big data and what is it all about.
2. Understanding and studying the steps in big data analytics
3. Exploring how big data analytics can fetch business benefits to industries.
4. The business benefit according to us would be, analyzing the huge amount of unstructured data and getting pattern to make sense out of it to use it for generating customer base, trends. Exploring what can be more business benefits.
5. As the stream is new, the technical challenge would be that there are not much of techniques available to analyze big data in proper manner. The awareness amongst the people is not satisfying when it comes to big data; people often think it is for only huge industries. But in actual, big data analytics is applicable wherever there is huge amount of unstructured data and people need to make sense out of it using certain techniques and identify patterns and trends.
6. Basic idea has taken on Big Data from published material, industry report available on internet, study of traditional BI approaches and its inadequacy in arriving at business decisions for non-SQL database like MapReduce. Defining the challenges in the Big Data Analytics and suggested approach for handling Analytics in smaller organizations.

BIG DATA IN DEPTH

Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. Big data has the potential to help companies improve operations and make faster, more intelligent decisions.

Is Big Data a Volume or a Technology?

While the term may seem to reference the volume of data, that isn't always the case. The term big data, especially when used by vendors, may refer to the technology (which includes tools and processes) that an organization requires to handle the large amounts of data and storage facilities. The term big data is believed to have originated with Web search companies who needed to query very large distributed aggregations of loosely-structured data.

An Example of Big Data

An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact centre, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

Big Data and Types of Business Datasets

When dealing with larger datasets, organizations face difficulties in being able to create, manipulate, and manage big data. Big data is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyze massive datasets.

As research from Webopedia parent company QuinStreet demonstrates, big data initiatives are poised for explosive growth. QuinStreet surveyed 540 enterprise decision-makers involved in big data and found the datasets include traditional structured databases of inventories, orders, and customer information, as well as unstructured data from the Web, social networking sites, and intelligent devices.

This data, when captured, formatted, manipulated, stored, and analysed can help a company to gain useful insight to increase revenues, get or retain customers, and improve operations. [5]

Trends and Patterns used:

Big data has been getting lots of attention for what it can possibly reveal about customers, market trends, marketing programs, equipment performance and other business elements. For many IT decision makers, big data analytics tools and technologies are now a top priority. These stories highlight trends and perspectives in the rapidly expanding world of big data analytics. [4]

IV. EVOLUTION OF BIG DATA

The story of how data became big starts many years before the current buzz around big data. Already seventy years ago we encounter the first attempts to quantify the growth rate in the volume of data or what has popularly been known as the "information explosion" (a term first used in 1941, according to theOxford English Dictionary). The following are the major milestones in the history of sizing data volumes plus other "firsts" in the evolution of the idea of "big data" and observations pertaining to data or information explosion.

1944 Fremont Rider, Wesleyan University Librarian, publishes The Scholar and the Future of the Research Library. He estimates that American university libraries were doubling in size every sixteen years. Given this growth rate, Rider speculates that the Yale Library in 2040 will have "approximately 200,000,000 volumes, which will occupy over 6,000 miles of shelves a cataloguing staff of over six thousand persons."

1961 Derek Price publishes Science since Babylon, in which he charts the growth of scientific knowledge by looking at the growth in the number of scientific journals and papers. He concludes that the number of new journals has grown exponentially rather than linearly, doubling every fifteen years and increasing by a factor of ten during every half-century. Price calls this the "law of exponential increase," explaining that "each [scientific] advance generates a new series of advances at a reasonably constant birth rate, so that the number of births is strictly proportional to the size of the population of discoveries at any given time."

1971 Arthur Miller writes in The Assault on Privacy that "Too many information handlers seem to measure a man by the number of bits of storage capacity his dossier will occupy."

April 1980 I.A. Tjomsland gives a talk titled "Where Do We Go From Here?" at the Fourth IEEE Symposium on Mass Storage Systems, in which he says "Those associated with storage devices long ago realized that Parkinson's First Law may be paraphrased to describe our industry—'Data expands to fill the space available'…. I believe that large amounts of data are being retained because users have no way of identifying obsolete data; the penalties for storing obsolete data are less apparent than are the penalties for discarding potentially useful data."

1981 The Hungarian Central Statistics Office starts a research project to account for the country's information industries, including measuring information volume in bits. The research continues to this day. In 1993, Istvan Dienes, chief scientist of the Hungarian Central Statistics Office, compiles a manual for a standard system of national information accounts.

August 1983 Ithiel de Sola Pool publishes "Tracking the Flow of Information" in Science. Looking at growth trends in 17 major communications media from 1960 to 1977, he concludes that "words made available to Americans (over the age of 10) through these media grew at a rate of 8.9 percent per year… words actually attended to from those media grew at just 2.9 percent per year…. In the period of observation, much of the growth in the flow of information was due to the growth in broadcasting… But toward the end of that period [1977] the situation was changing: point-to-point media were growing faster than broadcasting." Pool, Inose, Takasaki and Hurwitz follow in 1984 with Communications Flows: A Census in the United States and Japan, a book comparing the volumes of information produced in the United States and Japan.

1996 Digital storage becomes more cost-effective for storing data than paper according to R.J.T. Morris and B.J. Truskowski, in "The Evolution of Storage Systems," IBM Systems Journal, July 1, 2003.

October 1997 Michael Cox and David Ellsworth publish "Application-controlled demand paging for out-of-core visualization" in the Proceedings of the IEEE 8th conference on Visualization. They start the article with "Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources." It is the first article in the ACM digital library to use the term "big data."

October 1999 Bryson, Kenwright and Haimes join David Banks, Robert van Liere, and Sam Uselton on a panel titled "Automation or interaction: what's best for big data?" at the IEEE 1999 conference on Visualization.

February 2001 Doug Laney, an analyst with the Meta Group, publishes a research note titled "3D Data Management: Controlling Data Volume, Velocity, and Variety." A decade later, the "3Vs" have become the generally-accepted three defining dimensions of big data, although the term itself does not appear in Laney's note.

September 2005 Tim O'Reilly publishes "What is Web 2.0" in which he asserts that "data is the next Intel inside." O'Reilly:

"As Hal Varian remarked in a personal conversation last year, 'SQL is the new HTML.' Database management is a core competency of Web 2.0 companies, so much so that we have sometimes referred to these applications as 'infoware' rather than merely software."

March 2007 John F. Gantz, David Reinsel and other researchers at IDC release a white paper titled "The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010" It is the first study to estimate and forecast the amount of digital data created and replicated each year.

January 2008 Bret Swanson and George Gilder publish "Estimating the Exaflood" in which they project that U.S. IP traffic could reach one zettabyte by 2015 and that the U.S. Internet of 2015 will be at least 50 times larger than it was in 2006.

February 2010 Kenneth Cukier publishes in The Economist a Special Report titled, "Data, and data everywhere." Writes Cukier: "…the world contains an unimaginably vast amount of digital information which is getting ever vaster more rapidly… The effect is being felt everywhere, from business to science, from governments to the arts. Scientists and computer engineers have coined a new term for the phenomenon: 'big data.'"

May 2011 James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers of the McKinsey Global Institute publish "Big data: The next frontier for innovation, competition, and productivity." They estimate that "by 2009, nearly all sectors in the US economy had at least an average of 200 terabytes of stored data (twice the size of US retailer Wal-Mart's data warehouse in 1999) per company with more than 1,000 employees" and that the securities and investment services sector leads in terms of stored data per firm. In total, the study estimates that 7.4 exabytes of new data were stored by enterprises and 6.8 exabytes by consumers in 2010.

April 2012 The International Journal of Communications publishes a Special Section titled "Info Capacity" on the methodologies and findings of various studies measuring the volume of information. In "Tracking the flow of information into the home" Neuman, Park, and Panek (following the methodology used by Japan's MPT and Pool above) estimate that the total media supply to U.S. homes has risen from around 50,000 minutes per day in 1960 to close to 900,000 in 2005. And looking at the ratio of supply to demand in 2005, they estimate that people in the U.S. are "approaching a thousand minutes of mediated content available for every minute available for consumption." In "International Production and Dissemination of Information" Bounie and Gille (following Lyman and Varian above) estimate that the world produced 14.7 exabytes of new information in 2008, nearly triple the volume of information in 2003.

V. BENEFITS AND LIMITATIONS OF USING BIG DATA ANALYTICS

Having a lot of data pouring into your organization is one thing, being able to store it, analyze it and visualize it in real-time is a whole different ball game. More and more organization want to have real-time insights in order to fully understand what is going on within their organization. What are the advantages of Real-Time Big Data Analytics and what are the challenges and which tools can be used for real-time processing of Big Data?

The advantages of processing Big Data in real-time are many:

Errors within the organization are known instantly. Real-time insight into errors helps companies react quickly to mitigate the effects of an operational problem. This can save the operation from falling behind or failing completely or it can save your customers from having to stop using your products.

New strategies of your competition are noticed immediately. With Real-Time Big Data Analytics you can stay one step ahead of the competition or get notified the moment your direct competitor is changing strategy or lowering its prices for example.

Service improves dramatically, which could lead to higher conversion rate and extra revenue. When organizations monitor the products that are used by its customers, it can pro-actively respond to upcoming failures. For example, cars with real-time sensors can notify before something is going wrong and let the driver know that the car needs maintenance.

Fraud can be detected the moment it happens and proper measures can be taken to limit the damage. The financial world is very attractive for criminals. With a real-time safeguard system, attempts to hack into your organization are notified instantly. Your IT security department can take immediately appropriate action.

Cost savings: The implementation of a Real-Time Big Data Analytics tools may be expensive, it will eventually save a lot of money. There is no waiting time for business leaders and in-memory databases (useful for real-time analytics) also reduce the burden on a company's overall IT landscape, freeing up resources previously devoted to responding to requests for reports.

Better sales insights, which could lead to additional revenue. Real-time analytics tell exactly how your sales are doing and in case an internet retailer sees that a product is doing extremely well, it can take action to prevent missing out or losing revenue.

Keep up with customer trends: Insight into competitive offerings, promotions or your customer movements provides valuable information regarding coming and going customer trends. Faster decisions can be made with real-time analytics that better suit the (current) customer.

The Limitations of Real-Time Big Data Analytics

Of course, Real-Time Big Data Analytics is not only positive as it also offers some challenges.

It requires special computer power: The standard version of Hadoop is, at the moment, not yet suitable for real-time analysis. New tools need to be bought and used. There are however quite some tools available to do the job and Hadoop will be able to process data in real-time in the future.

Using real-time insights requires a different way of working within your organization: if your organization normally only receives insights once a week, which is very common in a lot of organizations, receiving these insights every second will require a different approach and way of working. Insights require action and instead of acting on a weekly basis this action is now in real-time required. This will have an effect on the culture. The objective should be to make your organization an information-centric organization. [1]

VI. PROBLEM ANALYSIS

Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights. Big data analytics is the process of examining large data sets containing a variety of data types (i.e., big data as qualified for its contents and nature –Volume, Velocity and Variety) with appropriate technologies to confirm the existence of hidden patterns and /or unknown correlations so as to establish market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.

One face of the problem space in the big data analytics problem in the growing data volumes from countless sources is essentially that of real-time informatics that is use of the business data extracted from the IT systems in real time for strategic business decisions. Otherwise there are risks of being swamped by data deluge without a decision. Meanwhile, competitors that use data to deliver better insights to decision-makers stand a better chance of thriving through the difficult economy and beyond. [6]

VII. PROPOSED SOLUTION

So far we have seen the basic approach to tackle Big Data Analytics is through the path of Hadoop /Map Reduce. The slight variation of approach for Big Data Analytic is by integrating other technologies with Hadoop. Big data, analytics should have the capability not only more data to work with, but also the processing power to handle large numbers of records with many attributes, the ability to do very large numbers of records and very large numbers of attributes per record. To enable real-time analysis and predictive modelling out of the same Hadoop core, that's where the interest of today's big data analytics. The problem has been speed, with Hadoop taking up to 20 times longer to get questions answered than did more established technologies. In this regard perhaps best promise for big data analytics come from In Memory Analytics, which is described in brief.

In-memory analytics

Unlike conventional business intelligence (BI) software that runs queries against data stored on server hard drives, in-memory technology queries information loaded into RAM, which can significantly accelerate analytical performance by reducing or even eliminating disk I/O bottlenecks.

The use of in-memory databases to speed up analytic processing is increasingly popular and highly beneficial in the right setting. In fact, many businesses are already leveraging hybrid transaction/analytical processing (HTAP) -- allowing transactions and analytic processing to reside in the same in-memory database.

But there's a lot of hype around HTAP, and businesses have been overusing it. For systems where the user needs to see the same data in the same way many times during the day -- and there's no significant change in the data -- in-memory is a waste of money.

And while we can perform analytics faster with HTAP, all of the transactions must reside within the same database. The problem is that most analytics efforts today are about putting transactions from many different systems together. Just putting it all on one database goes back to this disproven belief that if we want to use HTAP for all of our analytics, it requires all of our transactions to be in one place. We still have to integrate diverse data.

"The biggest benefit to in-memory analytics is speed of analysis and exploration,, The data latency that often bogs down traditional BI querying "interrupts the whole thought process" for business users, analytical flexibility as another in-memory analytics plus: With in-memory tools, users can ask business questions they could never ask before because the technology was too slow

Moreover, bringing in an in-memory database means there's another product to manage, secure, and figure out how to integrate and scale. Consultants and experienced users say the resulting speed boost is particularly compelling for big data analytics applications involving complex what-if scenarios and large amounts of information from a variety of data sources. Far more flexibility for creating queries on the fly and joining together information from disparate data sources to get answers to their business questions.

VIII. CONCLUSION

This paper describes various applications of big data analytics and its benefits. We have provided studies on big data, big data analytics, advantages and limitations of big data analytics in real time scenario.[3] Rather than improve methods of simulating various operations, scenarios, and environments as big data analytics has traditionally done, we suggested deploying In-memory analytics. We dynamically instrument the code to inspect scenarios of interest. Once this data has been collected, we can break it down into constituent dimensions—by usage scenario, location, and machine configuration— and present the results in ways that can help project stakeholders to make decisions. An analysis dashboard allows developers to investigate performance data. More and improved algorithms could be used for the performance would help in faster analytics and speedy decision making with the availability of more data

REFERENCES

[1] Big Data Big Mess- Sound Risk Intelligence through Complete Context
[2] Big Data Analytics-E-Book
[3] Operationazing-the-buzz-sas
[4] 7 top tools for taming big data _ InfoWorld
[5] data-insights-peer-research-report
[6] big-data-meets-big-data-analytics

AUTHOR'S BIOGRAPHY

Prof. Dr. Priti Puri

Dr. Priti Puri is a Doctorate in Computer Science from Kurukshetra University and has an MTech degree in Computer Science from Kurukshetra University. She has over 8 years of experience including academics, research and as a Patent Analyst for Microsoft. She has published and presented many research papers at various refereed/indexed International journals and Conferences.

Neerja Kulkarni

Ms. Neerja Kulkarni has completed her BCA degree from S.N.D.T University. She has also completed MBA-ITBM from Symbiosis Centre for Information Technology, Constituent of Symbiosis International University.

.

.