International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue (6) NCRTCST -2013, ISSN 2249–071X

National Conference on Recent Trends in Computer Science and Technology (NCRTCST)-2013

Challenges and Opportunities of BIG DATA ANALYTICS in Business Applications

Mr. M. Joseph Rajakumar
Associate Professor, Department of Computer Science
St. Joseph's PG College, Hyderabad, India
Email:josephrajakumar@rediffmail.com

Mrs. E. Sushma
Assistant Professor, Department of Computer Science
St. Joseph's PG College, Hyderabad, India
Email: sushma@josephspgcollege.ac.in

***Abstract* – Big data analytics is to discover a hidden data from the massive and messy data. Analytics is a sophisticated technique to extract actionable knowledge and insight from the data. Big Data is not just about size. The amount of data available is growing faster than our ability to deal with it, and more is coming [1]. Business people are aware of problem but aren't aware of how data can help them for their potential use. The big data can come from sources such as runtime information about traffic, tweets during the Olympic Games, stock market updates, usage information of an online game [2], or the data from any other rapidly growing data-intensive software system.**

In this paper we throw light on the impact of big data and its analytics which helps in business applications. The big data involved in Business organization for skills, leadership, organizational structures, technologies and architectures developments. The potential of Big Data is in its ability to solve business problem and provide business opportunities. Big data requires high-performance analytics to process and figure out what's important and what's not. The value of data in only realised through insight and visualizations play a key role turning data into insight.

***Keywords* – Big Data, Analytics, Cloud, Hadoop, Grid Computing, Mapreduce.**

I. INTRODUCTION

The name "big data" originated as technology with roots in high-performance computing, as pioneered by Google in the early 2000s. Today, the big data market is expanding quickly and includes new technologies, such as distributed file and database management tools led by the Apache Hadoop project and integration technology for exposing data to other systems and services.

The problems that start right away during data acquisition, when the data like natural disaster requires us to make decisions, currently in an ad- hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are not structured enough text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modelling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analysed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge.

Data comes from many sources – sensor data is usually fairly well structured and uniform. Application logs are often a valuable source of information. Data in web logs can be used to improve the user experience of the application and also suggest information for new features and product. Data stored in organizations are documents such as word-processing, pdfs, spreadsheets, and presentation; they are not properly structured but contain critical information.

Big Data as the three Vs.: Volume, Velocity, and Variety. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon.

The fast rise of open source technologies such as Hadoop and other NoSQL ways of storing and manipulating data. Big Data is a collection of large and complex data sets that cannot be handled by regular tools. The best way to get big data flowing in real-time is with middleware that takes care of message queuing and delivery so publishing applications and sensors can send data without worrying about where it needs to go or how it needs to get there.

II. HOW TO MANAGE THE BIG DATA

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. Big data analytics and the cloud are almost a perfect solutions. The ability to elastically provision the number of processing nodes necessary for the analytics job while paying only for their actual use is a prime example of the real benefits the cloud offers. Big data refers to the size of a dataset that has grown too large to be manipulated through traditional methods. These methods include capture, storage, and processing of the data in a tolerable amount of time.

Although the term big data was once applied to the concept of data warehouses, it now refers to large-scale processing architectures that focus on capacity, throughput, and genericity of processing.

In order to understand how to combat big data's faults, we must first understand its nature. We can define big data as a blend between three things: technology, analysis, and myth. First we employ extreme computational power to gather, link, and analyse large data sets. Then we analyse and draw patterns to make claims including society, economics, finance and technology. Lastly, the myth that more data will grant us higher acumen, and award us the power to generate better insights that were previously impossible with exactitude.

Another key objective involving time reduction is to be able to interact with the customer in real time, using analytics and data derived from the customer experience. If the customer has "left the building," targeted offers and services are likely to be much less effective. This means rapid data capture, aggregation, processing, and analytics.

When data volumes get into the multi-terabyte and multi-petabyte range, they require different treatment. Algorithms that work fine with smaller amounts of data are often not fast or efficient enough to process larger data sets, and there's no such thing as infinite capacity, even with storage media and management advances. But volume is only the first dimension of the big data challenge; the other two are velocity and variety. Velocity refers to the speed requirement for collecting, processing, and using the data. Many analytical algorithms can process vast quantities of information. Variety signifies the increasing array of data types—audio, video, and image data, as well as the mixing of information collected from sources as diverse as retail transactions, text messages, and genetic codes. Traditional analytics and database methods are excellent at handling data that can easily be represented in rows and columns and manipulated by commands such as select and join. But many of the artifacts that describe our world can neither be shoehorned into rows and columns, nor easily analysed by software that depends on performing a series of selects, joins, or other relational commands.

Fig.1. Representation of Large Data Analytics

Adding volume, variety, and velocity together does not provide better data. And as a result, dealing with big data demands a level of database agility and changeability that is difficult or impossible to achieve using today's techniques alone. "In a traditional database, design is everything," says Tom Deutsch, IBM Information Management program director. "It's all about structure. If the data changes or if what you want to know changes—or if you want to combine the data with information from another stream or warehouse—you have to change the whole structure of the warehouse. With big data, you're often dealing with evolving needs—and lots of sources of data, only some of which you produce yourself—and you want to be able to change the job you're running, not the database design."

The real issue is not that you are acquiring large amounts of data. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyse it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) 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. Analyse millions of stock keeping units 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 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 behaviour.

More often than not, we are too trusting of statistics, and fail to examine the data with a critical eye. Oftentimes, we are quick to conclude that the data presented to us is factual, which is entering risky waters in the context of big data.

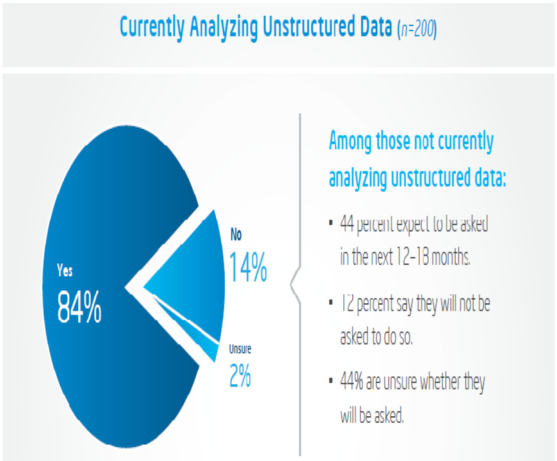
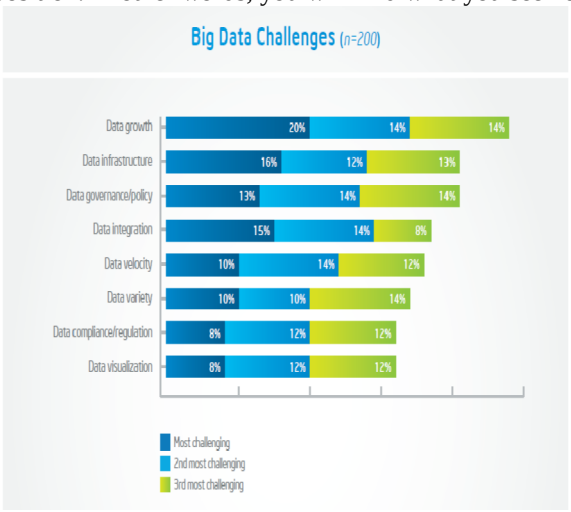


Fig.2. Analyses of Unstructured Data

The first problem with big data is it's so vast and unorganized, that organizing it for analysis is no easy task. Dan Ness, principal research analyst at MetaFacts, states, "A lot of big data today is biased and missing context, as it's based on convenience samples or subsets." [6]. The so-called experts have given people the illusion that since they've come up with an algorithm, the data you plug in will always be correct. But that'll only work if the assumptions that went into the algorithm are correct. As a result, people have a false sense of confidence in the data, especially as the data sets get larger and larger. If the "experts" produced an algorithm with the Twitter data that showed Manhattan as the centre of disaster, they would have come to the wrong conclusion.

Which leads us to our second problem: the sheer amount of data! No wonder we are more prone to "signal error" and "confirmation bias." Signal error is when large gaps of data have been overlooked by analysts. If places like Coney Island and Rockaway were overlooked in Hurricane Sandy, like they were in the Twitter study, we could be looking at a higher death toll today. Confirmation bias is the phenomenon that people will search within the data to confirm their own pre-existing viewpoint, and disregard the data that goes against their previously held position. In other words, you will find what you seek out.

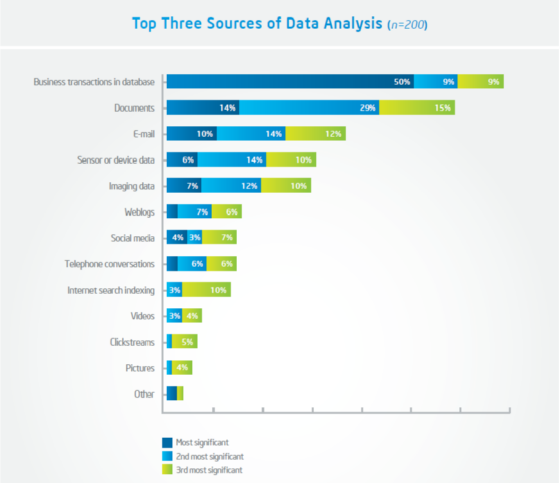


Incorporate massive data volumes in analysis. If the answers you're seeking will be better provided by analysing all of your data, go for it. High-performance technologies that extract value from massive amounts of data are here today. One approach is to apply high-performance analytics to analyse the 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.

III. SOLUTIONS

To achieve speed and scalability, Hadoop relies on MapReduce, a simple but powerful framework for parallel computation. MapReduce breaks down a problem into millions of parallel computations in the Map phase, producing as its output a stream of key-value pairs. Then MapReduce shuffles the map output by key and does another parallel computation on the redistributed map output, writing the results to the file system in the Reduce phase of the computation. For example, when processing huge volumes of sales transaction data to determine how much of each product was sold, Hadoop would do a Map operation for each block of a file containing transactions, add up the count of each product sold in each transaction, and then "reduce" as it returned an answer.



More often than not, we are too trusting of statistics, and fail to examine the data with a critical eye. Oftentimes, we are quick to conclude that the data presented to us is factual, which is entering risky waters in the context of big data.

Hadoop is a Java-based framework that supports data-intensive distributed applications, enabling applications to work with thousands of processor nodes and petabytes of data. Optimized for the sequential reading of large files, it automatically manages data replication and recovery. Even if a failure occurs at a particular processor, data is replicated and processing continues without interruption or loss of the rest of a computation, making the system somewhat fault-tolerant and capable of sorting a terabyte of data very quickly.

Because it's so simple to understand and use this technology—since it relies so heavily on just two steps, Map and Reduce—Hadoop-based systems have been used to handle a wide variety of problems, particularly in social media.

IV. Related Work

There are two related work in this areas.

A) One clear example of Big Data is the Square Kilometre Array (SKA) (www.skatelescope.org) planned to be constructed in South Africa and Australia. When the SKA is completed in 2024 it will produce in excess of one exabyte of raw data per day (1 exabyte = 1018 bytes), which is more than the entire daily internet traffic at present [3]. The SKA is a 1.5 billion Euro project that will have more than 3000 receiving dishes to produce a combined information collecting area of one square kilometre, and will use enough optical fibre to wrap twice around the Earth. Another example of Big Data is the Large Hadron Collider, at the European Organisation for Nuclear Research (CERN), which has 150 million sensors and is creating 22 petabytes of data in 2012 (1 Petabyte = 1015 bytes, see Figure 1). In biomedicine the Human Genome Project is determining the sequences of the three billion chemical base pairs that make up human DNA. In Earth observation there are over 200 satellites in orbit continuously collecting data about the atmosphere and the land, ocean and ice surfaces of planet Earth with pixel sizes ranging from 50 cm to many tens of kilometres.

B) Hadoop Log Analysis: Hadoop typically runs on large-scale clusters of machines with hundreds or even thousands of nodes. As a result, large amounts of log data are generated by Hadoop. To collect and analyse the large amounts of log data from Hadoop, Boulon et al. built Chukwa [4]. This framework monitors Hadoop clus- ters in real-time and stores the log data in Hadoop's distributed file system (HDFS). By leveraging Hadoop's infrastructure, Chukwa can scale to thousands of nodes in both collection and analysis. However, Chukwa focuses more on collecting logs without the ability to perform complex analysis. Tan et al. introduced SALSA, an approach to automatically analyze Hadoop logs to construct state-machine views of the platform's execution [5]. The derived state-machines are used to trace the data-flow and control-flow executions. SALSA computes the histograms of the durations of each state and uses these histograms to estimate the Probability Density Functions (PDFs) of the distributions of the durations. SALSA uses the difference between the PDFs across machines to detect anomalies. Tan et al. also compare the duration of a state in a particular node with its past PDF to determine if the duration exceeds a determined threshold and can be flagged as an anomaly. Another related work to this paper is the approach of Xu et al. in, which uses the source code to understand the structure of the logs. They create features based on the constant and variable parts of the log messages and apply the Principal Component Analysis (PCA) to detect the abnormal behaviour. All the above approaches are all designed for system administrators in managing their large clusters. Our approach, on the other hand, aims to assist developers in comparing the deployed system on such large clusters against the development cloud.

V. Conclusion

Big Data offers a frontier of opportunities that allow businesses across all industries to improve everything from their marketing and customer service to their manufacturing and product development. The amount of enterprise data, and the rate at which it's being accumulated, is rising exponentially.

Making forward-looking, proactive decisions requires proactive big analytics like optimization, predictive modelling, text mining, forecasting and statistical analysis. They allow you to identify trends, spot weaknesses or determine conditions for making decisions about the future. But although it's proactive, big analytics cannot be performed on big data because traditional storage environments and processing times cannot keep up.

Lastly, by using big data analytics you can extract only the relevant information from terabytes, petabytes and exabytes, and analyze it to transform your business decisions for the future. Becoming proactive with big data analytics isn't a one-time endeavor; it is more of a culture change – a new way of gaining ground by freeing your analysts and decision makers to meet the future with sound knowledge and insight.

References

[1] https://cs.uwaterloo.ca/~hhemmati/pubs/ICSE13Preprint.pdf
[2] N. Wingfield, "Virtual product, real profits: Players spend on zynga's games, but quality turns some off," Wall Street Journal.
[3] http://www.researchtrends.com/wp-content/uploads/ 2012/09/ Research_Trends_Issue30.pdf
[4] J. Boulon, A. Konwinski, R. Qi, A. Rabkin, E. Yang, and M. Yang, "Chukwa, a large-scale monitoring system," in CCA '08: Proc. of the first workshop on Cloud Computing and its Applications, 2008, pp. 1–5.
[5] J. Tan, X. Pan, S. Kavulya, R. Gandhi, and P. Narasimhan, "Salsa: analyzing logs as state machines," in WASL'08: Proceedings of the First USENIX conference on Analysis of system logs. Berkeley, CA, USA: USENIX Association, 2008, pp. 6–6.
[6] http://www.techpageone.com/technology/storage/big-data-seen-as-problematic-by-some-analysts/#.UljUmdL0Bv4