Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2016, 3(1):21-24**

Review Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Big Data- A Review on Analysing 3Vs

Abhinandan Banik[1], Samir Kumar Bandyopadhyay[2]

[1]IBM India Pvt. Ltd.
[2]Lincoln University, Malaysia

Abstract Big Data is comprised of large data sets that can't be handle by traditional systems. Big data includes structured data, semi-structured and unstructured data. The data storage technique used for big data includes multiple clustered network attached storage (NAS) and object based storage. Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This paper aims to analyse3Vs of big data which can be applied to big.

Keywords Big Data, Data Mining, Data Classification, Mining Techniques

Introduction

Big data is the need for new techniques and tools in order to be able to process huge database. Prominent examples include social media network analyzing their members' data to learn more about them and connect them with content and advertising relevant to their interests, or search engines looking at the relationship between queries and results to give better answers to users' questions. wo of the largest sources of data in large quantities are transactional data, including everything from stock prices to bank data to individual merchants' purchase histories; and sensor data, much of it coming from what is commonly referred to as the Internet of Things (IoT). This sensor data might be anything from measurements taken from robots on the manufacturing line of an auto maker, to location data on a cell phone network, to instantaneous electrical usage in homes and businesses, to passenger boarding information taken on a transit system.

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, we must choose an alternative way to process it. The hot IT buzzword of 2012, big data has become viable as cost-effective approaches have emerged to tame the volume, velocity and variability of massive data.

Extracting knowledge from Big Data is a high-touch business today, requiring a human expert who deeply understands the application domain as well as a growing ecosystem of complex distributed systems and advanced statistical methods. These experts are hired in part for their statistical expertise, but report that the majority of their time is spent scaling and optimizing the relatively basic data manipulation tasks in preparation for the actual statistical analysis or machine learning step: identifying relevant data, cleaning, filtering, joining, grouping, transforming, extracting features, and evaluating results.

In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data.

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition.

Review Literature

Researchers describe the big data content, its scope, methods, samples, advantages and challenges of Data [1]. The critical issue about the Big data is the privacy and security. Big data samples describe the review about the atmosphere, biological science and research. Life sciencesetc. By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets.

The overall Evaluation describe that the data is increasing and becoming complex. The challenge is not only to collect and manage the data also how to extract the useful information from that collected data.

According to the Intel IT Centre, there are many challenges related to Big Data which are data growth, data infrastructure, data variety, data visualization, data velocity.

Grid Computing offered the advantage about the storage capabilities and the processing power andthe Hadoop technology is used for the implementation purpose. Grid Computing provides the concept of distributed computing. The benefit of Grid computing centre is the high storage capability and the high processing power. Grid Computing makes the big contributions among the scientific research, help the scientists to analyse and store the large and complex data [2].

Big data analytics define the analysis of large amount of data to get the usefulinformation and uncover the hidden patterns. Big data analytics refers to the Mapreduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model.In this the performance of SF-CFS is compared with the HDFS using the SWIM by the facebook job traces .SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns [3].

Some researchers reportthe experimental work on the Big data problems. It describe the optimal solutions using Hadoopcluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets [4].

Over the last many years, there are many researchers has completed their work successfully on big data. Hundreds of articles have appeared in the general business press (For example Forbes, Fortune, Bloomberg, Business week, The Wall street journal, The Economist). National Institute of Standards and Technology [NIST] said that Big Data in which data volume, velocity and data representation ability to perform effective analysis using traditional relational approaches.

Analysis of Big Data

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analysed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data.

Big data can be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Data is generated from various different sources and can arrive in the system at various rates. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance.

Analyzing new and diverse digital data streams can reveal new sources of economic value, provide fresh insights into customer behaviour and identify market trends early on. But this influx of new data creates challenges for IT departments. To derive real business value from big data, you need the right tools to capture and organize a wide variety of data types from different sources, and to be able to easily analyze it within the context of all enterprise data.

Big Data storage infrastructures must also be flexible, so that they can support multiple file types and client-side operating systems on the front end and multiple storage platforms on the back end. In reality these large infrastructures can eventually become a consolidation of many disparate storage devices as they accumulate more diverse data sets. New problems and growing computing power will spur the development of new analytical techniques. There is also a need for ongoing innovation in technologies and techniques that will help individuals and organizations to integrate, analyze, visualize, and consume the growing torrent of big data.

We have all heard of the 3Vs of big data which are Volume, Variety and Velocity, yet other Vs that IT, business and data scientists need to be concerned with, most notably big data Veracity.

Data Volume

Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it. As data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors.

Data Variety

Data variety is a measure of the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl.

Data Velocity

Data velocity measures the speed of data creation, streaming, and aggregation. Ecommerce has rapidly increased the speed and richness of data used for different business transactions (for example, web-site clicks). Data velocity management is much more than a bandwidth issue; it is also an ingest issue.

Data Veracity

Data veracity refers to the biases, noise and abnormality in data. Is the data that is being stored and mined meaningful to the problem being analysed. Veracity in data analysis is the biggest challenge when compares to things like volume and velocity.

Big data analytics refers to the process of collecting, organizing and analysing large sets of data ("big data") to discover patterns and other useful information. Not only will big data analytics help you to understand the information contained within the data, but it 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 analysing the data.

The Benefits of Big Data Analytics

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.

The Challenges of Big Data Analytics

For most organizations, big data analysis is a challenge. Consider the sheer volume of data and the many different formats of the data (both structured and unstructured data) collected across the entire organization and the many different ways different types of data can be combined, contrasted and analysed to find patterns and other useful information.

The first challenge is in breaking down data silos to access all data an organization stores in different places and often in different systems. A second big data challenge is in creating platforms that can pull in unstructured data as easily as structured data. This massive volume of data is typically so large that it's difficult to process using traditional database and software methods.

Big Data Requires High-Performance Analytics

To analyse such a large volume of data, big data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, and forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high-performance analytics. Using big data tools and software enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analysed to drive better business decisions in the future.

How big data analytics is used today?

As technology to break down data silos and analyse data improves, business can be transformed in all sorts of ways. Big Data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Face book. The business cases for leveraging Big Data are compelling. For instance, Netflix mined its subscriber data to put the essential ingredients together for its recent hit House of Cards, and subscriber data also prompted the company to bring Arrested Development back from the dead.

Conclusions

In big data analytics, researchers divided generated data into various big data application such as structured data analytics, text analytics, web analytics, multimedia analytics and mobile analytics. Many challenges in the big data system need further research attention. Research on typical big data application can generate profit for businesses, improve efficiency of government sectors.

References

[1]. Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013),”Big Data: A Review”

[2]. Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ;,(17-19 Jan. 2013),”A Big Data implementation based on Grid Computing”, Grid Computing

[3]. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , “Shared disk big data analytics with Apache Hadoop”

[4]. Aditya B. Patel, Manashvi Birla, UshmaNair ,(6-8 Dec. 2012),“Addressing Big Data Problem Using Hadoop and Map Reduce”.

[5]. Y. Low, J. Gonzalez, A. Kyrola, D. Bickson,C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In Conference on Uncertainty in Artificial Intelligence (UAI), Catalina Island, California, July 2010.