International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

Big Data Analysis

Mohd Suhail Khan

Department of Management, M. Phil Student, Mewar University, Chittorgarh, Rajasthan, India

Abstract: ***A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. \Enormous amounts of data have become available on hand to decision makers. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. 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. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. The basic objective of this paper is to explore the characteristics of big Data, Identify its challenges, various tools associated with it & exploring the advantages of Data Analytics in Decision making.***

Keywords: Big data analytics; Hadoop; Massive data; Structured data; Unstructured Data; Decision making

1. Introduction

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes(TB) to many petabytes (PB) of data in a single data set. They are difficult to process using traditional database management tools or data processing applications

Characteristics of Big Data

Data are available in structured, semi-structured, and unstructured format in petabytes and beyond. Four main features characterize big data: volume, velocity, variety & veracity. Volume refers to the huge amount of data that are being generated everyday whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi structured etc. The fourth V refers to veracity that includes availability and accountability.

The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques

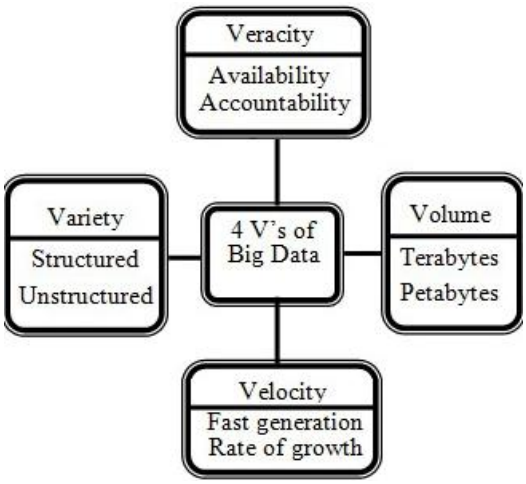

Figure 1: Characteristics of Big Data

Big Data Analytics Tools and Methods

With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analyzing data. Having piles of data on hand is no longer enough to make efficient decisions at the right time. Therefore, there arises a need for new tools and methods specialized for big data analytics, as well as the required architectures for storing and managing such data.

The changes associated with big data analytics are reflected in three main areas: big data storage and architecture, data and analytics processing, and, finally, the big data analyses which can be applied for knowledge discovery and informed decision making.

Large numbers of tools are available to process big data. In this section, we discuss some current techniques for analyzing big data with emphasis on three important emerging tools namely Map Reduce, Apache Spark, and Storm. Most of the available tools concentrate on batch processing, stream processing, and interactive analysis. The typical work flow of big data project discussed by Huang et al is depicted below figure

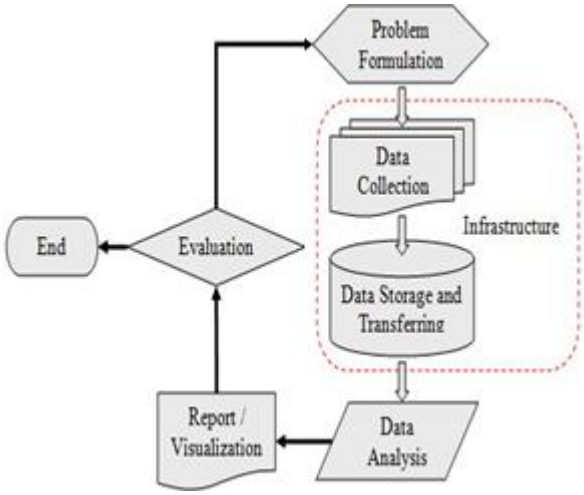

Figure 2: Workflow of Big Data Project

Apache Hadoop and MapReduce

The most established software platform for big data analysis is Apache Hadoop and Mapreduce. It consists of hadoop kernel, map reduce, hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the sub problems in reduce step. Moreover, Hadoop and MapReduce work as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing

Apache Spark

Apache spark is an open source big data processing framework built for speed processing, and sophisticated analytics. It is easy to use and was originally developed in 2009 in UC Berkeleys AMP Lab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, scala, or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of existing hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing in the form of tasks. Each application will have a set of processes called executors that are responsible for executing the tasks. The major advantage is that it provides support for deploying spark applications in an existing hadoop clusters. Below figure depicts the architecture diagram of Apache Spark

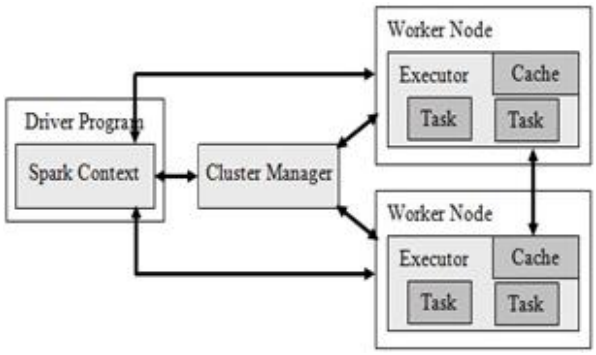

Figure 3: Architecture of Apache Spark

Big data analytics

Now a day, people don't just want to collect data, they want to understand the meaning and importance of data, and use it to aid them in decision making. Data analytics is a process of applying algorithms in order to analyze sets of data and extract useful and unknown patterns, relationships and information's.

Data analytics are used to extract previously useful, unknown, valid and hidden patterns and information from large data sets, as well as to detect important relationships among them. Therefore analytics has a significant impact on research and technologies as most of the decision makers are interested in learning from previous data, thus gaining competitive advantage.

For example, social media has recently become important for social networking and content sharing. Yet, the content that is generated from social media websites is enormous and remains largely unexploited. However, social media analytics can be used to analyze such data and extract useful information and predictions. Social media analytics is based on developing and evaluating informatics frameworks and tools in order to collect, monitor, summarize, analyze, as well as visualize social media data. Furthermore, social media analytics facilitates understanding the reactions and conversations between people in online communities, as well as extracting useful patterns and intelligence from their interactions, in addition to what they share on social media websites.

Big Data Analytics and Decision Making

From the decision maker's perspective, the significance of big data lies in its ability to provide information and knowledge of value, upon which to base decisions. The managerial decision making process has been an important and thoroughly covered topic in research throughout the years.

Big data is becoming an increasingly important asset for decision makers. Large volumes of highly detailed data from various sources such as scanners, mobile phones, loyalty cards, the web, and social media platforms provide the opportunity to deliver significant benefits to organizations. This is possible only if the data is properly analyzed to reveal valuable insights, allowing for decision makers to capitalize upon the resulting opportunities from the wealth of historic and real-time data generated through supply chains, production processes, customer behaviors, etc.

Phases of Decision making through Data Analytics

Phase I:
The first phase of the decision making process is the intelligence phase, where data which can be used to identify problems and opportunities is collected from internal and external data sources. In this phase, the sources of big data need to be identified, and the data needs to be gathered from different sources, processed, stored, and migrated to the end user. Such big data needs to be treated accordingly, so after the data sources and types of data required for the analysis are defined, the chosen data is acquired and stored in any of the big data storage and management tools previously discussed. After the big data is acquired and stored, it is then organized, prepared, and processed, This is achieved across a high-speed network using ETL/ELT or big data processing tools, which have been covered in the previous sections.

Phase II:
The next phase in the decision making process is the design phase, where possible courses of action are developed and analyzed through a conceptualization, or a representative model of the problem. The framework divides this phase into three steps, model planning, data analytics, and analyzing. Here, a model for data analytics is selected and planned, and then applied, and finally analyzed.

Phase III:
Consequently, the following phase in the decision making process is the choice phase, where methods are used to evaluate the impacts of the proposed solutions, or courses of action, from the design phase.

Phase IV:
Finally, the last phase in the decision making process is the implementation phase, where the proposed solution from the previous phase is implemented.

2. Conclusion

In this research, we have examined the innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. Analyzing big data is challenging for a general man. In this paper, we survey the various issues, challenges, and tools used to analyze these big data. We also saw how the data analytics plays a important role for decision makers in decision making. From this survey, it is understood that every big data platform has its individual focus. Some of them are designed for batch processing whereas some are good at real-time analytic. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing. We believe that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently. We also believe that big data analytics is of great significance in this era of data overflow, and can provide unforeseen insights and benefits to decision makers in various areas. If properly exploited and applied, big data analytics has the potential to provide a basis for advancements, on the scientific, technological, and humanitarian levels.

References

[1] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management.
[2] Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management(2015).
[3] Lynch, Big data: How do your data grow?, Nature
[4] X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research
[5] R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society
[6] C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275 (2014).
[7] K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing
[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of mapreduce for imbalanced big data using random forest, Information Sciences
[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence