BIG DATA ANALYTICS – AN OVERVIEW OF RESEARCH OPPORTUNITIES AND CHALLENGES

Aftab Yaseen[1] , P. M. Khan[2]

[1]*Assistant Professor, Department of CSE, Integral University Lucknow, (India)*

[2]*Director, Computer Centre, Aligarh Muslim University Aligarh, (India)*

ABSTRACT

For quite some time, processing capabilities and architectural constraints of conventional database technologies continued to be a limiting factor, in making it a viable option to handle Big Data in real sense. With recent advancements of technology and increasing ability to process large volumes of data, there is an ever growing interest of industry and researchers to explore the hidden treasures of Big Data. While a few companies like Google and Facebook are forerunners in Big Data research and development projects, but the potential application domains are virtually countless - like healthcare, banking, finance, biotechnology, life sciences, engineering & technology etc. While the data types and structures are diverse in Big Data, but application of Big Data Analytics has potential to reveal patterns and clues that can help organizations and businesses to address underlying causes of problems and take informed decisions in real-time that can be strategic for businesses. This paper is the result of a study undertaken to provide an overview of Big Data, technology options, research issues & challenges ahead in this field.

Keywords: Big Data, Big Data Analytics, Hadoop, Map Reduce

I. INTRODUCTION

The term Big Data refers to the massive collections of datasets which cannot be processed or analyzed with the help of conventional data management tools. The volume of big data is growing exponentially every year due to the generation of large amount of data by the IT companies, social networking sites, industrial and health care systems. The nature of big data can be characterized by its volume, velocity and variety. The limitation of existing storage and processing architectures differentiate the volume of data to be considered as big data. The most immediate challenge to the conventional processing architectures is the volume of big data. This volume questions about the scalability of existing storage, and the processing of data in distributed manner. The velocity of big data refers to the increasing rate of generation of data.

The variety is the diversity in different forms of the source data including text, images, audio, video and sensor data generated by modern IT, industrial and other systems. Integrating such type of diverse data for processing is a challenge for companies. There are two processing options available to deal with these massive volumes of data. The first one is the data warehousing approach which involves the predetermined schemas, best suited for datasets which are evolving with a lower velocity. The Apache Hadoop-based solution is the other approach as it can process the data irrespective of the structure of the data. In the following sections we discuss the big data analytics techniques including the Hadoop MapReduce and the underlying challenges in implementing them. We will also discuss the various approaches for classification of big data.

II. BIG DATA ANALYTICS

Big data analytics provides deep insights hidden by big data that go beyond the processing capability of existing systems. In the past decades, the data was mostly used to only record and report business activities and scientific events. In future data will be used also to gain new insights, to influence business decisions and to speed up scientific innovations. There are several steps involved in big data analytics like data preprocessing which include data cleaning, integration, transformation and reduction etc. and data visualization for decision making.

The latest technologies such as cloud computing, parallel processing and distributed processing frameworks has enabled the big data analytics as an emerging field of research. For distributed and large scale computations like in cloud computing environment Hadoop MapReduce programming model is widely used. The key challenge in big data analytics is to provide the right platforms and tools to make reasoning of big data easy and simple.

III. APACHE HADOOP-BASED SOLUTIONS

The Apache Hadoop is an open source framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. This framework is so designed that it can be scaled from a single server to thousands of machines. All these machines offer local computation and storage. Without depending on the hardware for delivering high-availability, the library is designed in such a manner that it detects and handles failures at the application layer resulting in a high availability of services on top of a cluster of computers, each of which may be prone to failures [1]. The projects surrounding Apache Hadoop consists of following modules:

- **Hadoop Common:** The common utilities that support the other Hadoop modules. It has utilities and scripts for starting Hadoop components and interfaces to access the file system supported by Hadoop [2].
- **Hadoop Distributed File System (HDFS):** The Hadoop Distributed File System (HDFS) is a highly fault tolerant distributed file system designed to provide high throughput access to application data [3].
- **Hadoop YARN:** A framework for job scheduling and cluster resource management. YARN is the prerequisite for Enterprise Hadoop, providing resource management and a central platform to deliver consistent operations, security, and data governance tools across Hadoop clusters [4].
- **Hadoop MapReduce:** A YARN-based distributed programming model for parallel processing of large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key [5].

Other Hadoop related projects at Apache are summarized in **Table I:**

Table I: Hadoop related projects at Apache [1]

Ambari™	A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters.
Avro™	A data serialization system.
Cassandra™	A scalable multi-master database with no single points of failure.
Chukwa™	A data collection system for managing large distributed systems.
HBase™	A scalable, distributed database that supports structured data storage for large tables.

Hive™	A data warehouse infrastructure that provides data summarization and ad hoc querying.
Mahout™	A Scalable machine learning and data mining library.
Pig™	A high-level data-flow language and execution framework for parallel computation.
Spark™	A fast and general compute engine for Hadoop data.
Tez™	A generalized data-flow programming framework, built on Hadoop YARN.
ZooKeeper™	A high-performance coordination service for distributed applications.

IV. HADOOP MAPREDUCE PROGRAMMING MODEL

The Hadoop MapReduce provides a programming model for large scale data processing and an execution environment for MapReduce jobs. A MapReduce job is executed in two phases: the Map phase and the Reduce phase. The input data to these two phases are in the form of a key/value pairs. In the Map phase, data is read from a distributed file system. This data is then partitioned among a set of computing nodes in the cluster and send to the nodes as a set of key/value pairs. After processing these partitioned data independently an intermediate result is produced by the map task as key/value pairs. These intermediate results are stored on the local disk of the node running the Map task. After completing the Map tasks, the Reduce phase begins whose task is to aggregate the intermediate data with the same key. The advantage of MapReduce programming model is that it does the computations to where the data is located which results in decreasing the transmission of data and hence improving efficiency. This model is well suited for parallel processing of large scale data in which the data analysis tasks can be accomplished by independent Map and Reduce tasks. Google has used the MapReduce programming model successfully for many different purposes. The reason behind this is that this model is easy to use, even for programmers without experience with parallel and distributed systems, since it hides the details of parallelization, fault-tolerance, locality optimization, and load balancing. There is another reason of using this model is that a large variety of problems can be easily modeled into MapReduce programming model.

V. RESEARCH OPPORTUNITIES AND CHALLENGES

The sources of Big Data can be finance and business where huge amount of stock exchange, banking, and online shopping data flows via Internet every day and are then confined and stored for knowledge discovery in inventory analysis, customer and market behavior. The field of life sciences is another major producer of large and massive datasets including genome sequencing, clinical data and patient data. These data are analyzed and used to advance breakthroughs in medical sciences & research. Astronomy, oceanography, and engineering are the other areas of research where Big Data is of essential importance [6]. In the area of drug discovery large volume of structured and unstructured biomedical data stemming from a wide range of experiments and surveys collected by hospitals, laboratories, pharmaceutical companies or even social media. The challenge is to develop such algorithms to discover the hidden patterns in such data for predictions that may be used to determine possible drug structures with different desirable properties. In this field big data analytics may contribute to better drug efficacy and safety for pharmaceutical companies and regulators [7]. There is a number of emerging research areas relating to the use of Big Data Analytics for evidence based medicine (EBM). El-Gayar, Timsina [8] presented a study that provides a research agenda for health informatics researchers and data scientists to

address issues of reducing the cost and improving the cost of healthcare by broadening the practice of evidence based medicine through the applications of business intelligence big data analytics [8]. Cardenas et al. [9] highlighted the challenges related to the security issue in big data. These challenges are privacy, data provenance problem and human computer interaction. The privacy depends largely on the technological constraints on the ability to extract, analyze, and correlate potentially sensitive datasets. With the advancement in big data analytics different tools were developed to extract and correlate this data, and hence making the privacy violations easier. For data provenance related issues the authenticity and integrity of data used in such tools should be reconsidered in order to produce accurate results. Some machine learning and statistical techniques can be used to identify the maliciously inserted data and to deal with it. The use of human computer interaction or visual analytics helps in analyzing the query results. Although human-computer interaction in big data has received less attention but it is one of the primary tools of big data analytics, because its purpose is to provide information in a most effective manner [9]. Big data analytics use data mining algorithms which are computationally intensive and require efficient high performance processors for producing results in a given time frame. For computational and data storage requirement in big data analytics, cloud computing infrastructures can serve as an effective platform. Advanced data mining techniques and associated tools for knowledge discovery can help in extracting information from massive and complex datasets. Hence big data analytics and knowledge discovery techniques with scalable computing systems can be combined to give useful results in timely manner. The challenging areas in which the cloud based data analytics can be used are the development of scalable higher-level models and tools. Interoperability of data and tools is another major issue in large scale applications [10].

VI. CONCLUSION

It is evident from this study that Big Data Analytics approaches are increasingly being used to obtain valuable information from big data. Organizations are trying to develop multiple analytic platforms that can synthesize traditional structured data with semi-structured and unstructured sources of information. This paper focused on the thrust areas of research in the field of big data analytics. Based on the study of various research papers and articles on big data analytics, it is found that a lot of research is needed to develop tools for big data analytics that can help organizations. Findings from this study have also confirmed that discovery of hidden patterns in data, privacy and security issues of organizations and development of a cost effective system are major challenges in the field of big data analytics. Scalability and efficiency are other issues identified with cloud based systems. Volumes of research possibilities in these areas needs to be explored further, and will continue to attract attention of researchers and practitioners, globally.

REFERENCES

[1]. http://hadoop.apache.org/, Retrieved 2015.

[2]. A. Bagha,V. Madisetti, *cloud computing: a hands-on approach* (University Press India Private Limited, 2014).

[3]. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, Retrieved 2015.

[4]. http://hortonworks.com/hadoop/yarn/, Retrieved 2015.

[5]. J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, In Proc. of the 6th Symposium on Operating Systems Design and Implementation, San Francisco CA, Dec. 2004, 137-149.

[6]. H. Gali, M. Henk, The Evolution of Big Data as a Research and Scientific Topic, Research Trends Special Issue on Big Data, Elsevier 30 Sep 2012, 03-06.

[7]. Chan, K.C.C., Big data analytics for drug discovery, IEEE International Conference on Bioinformatics and Biomedicine, 18-21 Dec. 2013.

[8]. El-Gayar O., Timsina P., Opportunities for Business Intelligence and Big Data Analytics In Evidence Based Medicine, 47th Hawaii International Conference on System Science, 6-9 Jan. 2014, 749-757.

[9]. Cardenas, A.A. ; Manadhata, P.K. ; Rajan, S.P., Big Data Analytics for Security, Security & Privacy, IEEE, Vol. 11,Issue 6, Nov-Dec. 2013, 74-76.

[10]. Talia D., Clouds for Scalable Big Data Analytics, Computer, IEEE, Vol. 46, Issue 5, May 2013, 98-101.