A REVIEW OF BIG DATA ENVIRONMENT, TOOLS AND CHALLENGES

Chandra Shekhar Gautam,Dr. Prabhat Pandey

Research Scholar A.P.S University Rewa , Prof. Physices & OSD

A.P.S University Rewa (M.P)(India)

Abstract:

This is the era of big data .it refers to large amount of data set whose size is growing at a vast speed and it is collection of dataset 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. To handle such large amount of data using traditional software tools available. This paper reviews the technologies and challenges related to big data analytics. First introduce general definition of big data second essential technologies related big data analytics like hadoop, map reduce framework etc. and third challenges arising from big data analytics.

Keywords: big data, Hadoop, Map Reduce, analytics, decision making.

1 Introduction

Historically data is one of the most important assets for firms and governments and data analysis is the first elements needed for creating strategies and measures their effects, so big data processing and archival technologies may bring great opportunities with them in order to improve performance and competitiveness of modern organizations and resulting in significant benefits for the overall society.

The most important thing that can be true about big data is that it does not have one single definition in fact .it refers to large amount of datasets whose size is growing at a vast speed making it difficult to handle such large amount of data using traditional software tools available data mining techniques and database measurement tools, In general can define this data as a huge size dataset which hide any information in its massive volume which require a new data mining techniques or algorithms to explore.

Imagine a world without data storage; a place where every detail about a person or organization, every transaction performed. The contribution of this paper is to provide an analysis of the available literature on big data analytics. Accordingly, some of the various big data tools, methods, and technologies which can be applied are discussed, and their applications and opportunities provided in several decision domains are portrayed. The literature was selected based on its novelty and discussion of important topics related to big data, in order to serve the purpose of our research. This is due to big data being a recently focused upon topic. Furthermore, our corpus mostly includes research from some of the top journals, conferences, and white papers by leading corporations in the industry. Due to long review process of journals, most of the papers discussing big data analytics, its tools and methods, and its applications were found to be conference papers, and white papers. While big data analytics is being researched in academia, several of the industrial advancements and new technologies provided were mostly discussed in industry papers.

2 Big Data Analytics:

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 [12]. Big data sizes are constantly increasing, currently ranging from a few dozen tera bytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn’t know before [17]. Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage [17].

In this section, we will start by discussing the characteristics of big data, as well as its importance. Naturally, business benefit can commonly be derived from analyzing larger and more complex data sets that require real time or near-real time capabilities; however, this leads to a need for new data architectures, analytical methods, and tools. Therefore the successive section will elaborate the big data analytics tools and methods, in particular, starting with the big data storage

and management, then moving on to the big data analytic processing. It then concludes with some of the various big data analyses which have grown in usage with big data.

2.1 Characteristics of Big Data:

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize big data: volume, variety, and velocity, or the three V's. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data [9].

Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as extensible Markup Language (XML) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume.

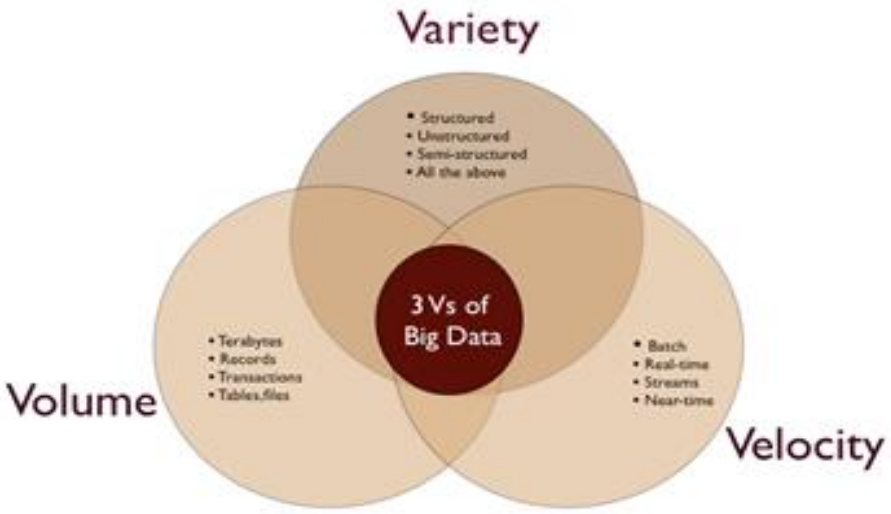

Figure1: characteristics of big Data

Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites [17]. Some researchers and organizations have discussed the addition of a fourth V, or veracity. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations [22].

2.2 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 such data. Having piles of data on hand is no longer enough to make efficient decisions at the right time.

Such data sets can no longer be easily analyzed with traditional data management and analysis techniques and infrastructures. 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. Accordingly, the emergence of big data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions.

Consequently, [8] proposed the Big – Data, Analytics, and Decisions (B-DAD) framework which incorporates the big data analytics tools and methods into the decision making process [8]. The framework maps the different big data storage, management, and processing tools, analytics tools and methods, and visualization and evaluation tools to the different phases of the decision making process. Hence, 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. Each area will be further discussed in this section. However, since big data is still evolving as an important field of research, and new findings and tools are constantly developing, this section is not exhaustive of all the possibilities, and focuses on pro-viding a general idea, rather than a list of all potential opportunities and technologies.

Big Data Analytics software is widely used in providing meaningful analysis of a large set of data. This software helps in finding current market trends, customer preferences, and other information.

Here are the 5 Top Big Data Analytics Tools with key feature and download links.

i)**Microsoft HDInsight:**

Azure HDInsight is a Spark and Hadoop service in the cloud. It provides big data cloud offerings in two categories, Standard and Premium. It provides an enterprise-scale cluster for the organization to run their big data workloads.

Features:

- Reliable analytics with an industry-leading SLA.
- It offers enterprise-grade security and monitoring.
- Protect data assets and extend on-premises security and governance controls to the cloud.
- High-productivity platform for developers and scientists.
- Integration with leading productivity applications.
- Deploy Hadoop in the cloud without purchasing new hardware or paying other up-front costs.

ii)Skytree:

Skytree is a big data analytics tool that empowers data scientists to build more accurate models faster. It offers accurate predictive machine learning models that are easy to use.

Features:

- Highly Scalable Algorithms.
- Artificial Intelligence for Data Scientists.
- It allows data scientists to visualize and understand the logic behind ML decisions.
- Sky tree via the easy-to-adopt GUI or programmatically in Java.
- Model Interpretability.
- It is designed to solve robust predictive problems with data preparation capabilities.
- Programmatic and GUI Access.

iii) Talend:

Talend is a big data tool that simplifies and automates big data integration. Its graphical wizard generates native code. It also allows big data integration, master data management and checks data quality.

Features:

- Accelerate time to value for big data projects.
- Simplify ETL & ELT for big data.
- Talend Big Data Platform simplifies using MapReduce and Spark by generating native code.
- Smarter data quality with machine learning and natural language processing.
- Agile DevOps to speed up big data projects.
- Streamline all the DevOps processes.

iv) Splice Machine: Splice Machine is a big data analytic tool. Their architecture is portable across public clouds such as AWS, Azure, and Google.

Features:

- It can dynamically scale from a few to thousands of nodes to enable applications at every scale.

- The Splice Machine optimizer automatically evaluates every query to the distributed HBase regions.
- Reduce management, deploy faster, and reduce risk.
- Consume fast streaming data, develop, test and deploy machine learning models.

v) Spark: Apache Spark is a powerful open source big data analytics tool. It offers over 80 high-level operators that make it easy to build parallel apps. It is used at a wide range of organizations to process large datasets.

Features:

- It helps to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk.
- It offers lighting Fast Processing.
- Support for Sophisticated Analytics.
- Ability to Integrate with Hadoop and Existing Hadoop Data.
- It provides built-in APIs in Java and Python.

3. Big Data Analytic Processing

According to [10], there are four critical requirements for big data. The first requirement is fast data loading. Since the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time. The second requirement is fast query processing. In order to satisfy the requirements of heavy workloads and real-time requests, many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increase. Additionally, the third requirement for big data processing is the highly efficient utilization of storage space. Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing, and issues on how to store the data so that space utilization is maximized be addressed. Finally, the fourth requirement is the strong adaptively to highly dynamic workload patterns. As big data sets are analyzed by different applications and users, for different purposes, and in various ways, the underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns [10].

Map Reduce is a parallel programming model, inspired by the “Map” and “Re-duce” of functional languages, which is suitable for big data processing. It is the core of Hadoop, and performs the data processing and analytics functions [6]. According to EMC, the Map Reduce paradigm is based on adding more computers or resources, rather than increasing the power or storage capacity of a single computer; in other words, scaling out rather than scaling up [9]. The fundamental idea of Map Reduce is breaking a task down into stages and executing the stages in parallel in order to re-duce the time needed to complete the task [6].

The first phase of the Map Reduce job is to map input values to a set of key/value pairs as output. The “Map” function accordingly partitions large computational tasks into smaller tasks, and assigns them to the appropriate key/value pairs [6]. Thus, unstructured data, such as text, can be mapped to a structured key/value pair, where, for example, the key could be the word in the text and the value is the number of occurrences of the word. This output is then the input to the “Reduce” function [9]. Reduce then performs the collection and combination of this output, by combining all values which share the same key value, to provide the final result of the computational task [6].

The Map Reduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results [9]. The Map Reduce job starts by the Job-Tracker assigning a portion of an input file on the HDFS to a map task, running on a node [13]. On the other hand, the Task Tracker nodes actually run the jobs and communicate results back to the Job Tracker. That communication between nodes is often through files and directories in HDFS, so inter-node communication is minimized [9].

Figure 1 Shows how the map reduce nodes and HDFS work together. At step 1 there is very large dataset including log files, and sensor data or anything of sorts.

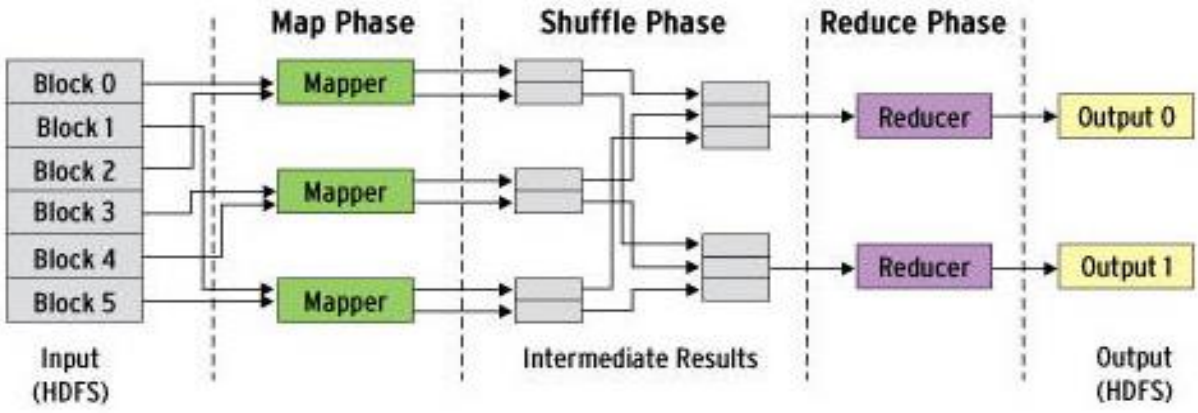

Figure2. MapReduce and HDFS

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. [4].

Moreover, organizations are currently accustomed to analyzing internal data, such as sales, shipments, and inventory. However, the need for analyzing external data, such as customer markets and supply chains, has arisen, and the use of big data can provide cumulative value and knowledge. With the increasing sizes and types of un-structured data on hand, it becomes necessary to make more informed decisions based on drawing meaningful inferences from the data [7].

Accordingly, [8] developed the B-DAD framework which maps big data tools and techniques, into the decision making process [8]. Such a framework is intended to enhance the quality of the decision making process in regards to dealing with big data. 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.

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, such as those previously discussed, is selected and planned, and then applied, and finally analyzed.

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. Finally, the last phase in the decision making process is the implementation phase, where the proposed solution from the previous phase is implemented [8].

As the amount of big data continues to exponentially grow, organizations through-out the different sectors are becoming more interested in how to manage and analyze such data. Thus, they are rushing to seize the opportunities offered by big data, and gain the most benefit and insight possible, consequently adopting big data analytics in order to unlock economic value and make better and faster decisions. Therefore, organizations are turning towards big data analytics in order to analyze huge amounts of data faster, and reveal previously unseen patterns, sentiments, and customer intelligence. This section focuses on some of the different applications, both proposed and implemented, of big data analytics, and how these applications can aid organizations across different sectors to gain valuable insights and enhance decision making.

According to Manyika et al.'s research, big data can enable companies to create new products and services, enhance existing ones, as well as invent entirely new business models. Such benefits can be gained by applying big data analytics in different areas, such as customer intelligence, supply chain intelligence, performance, quality and risk management and fraud detection [14]. Furthermore, Cebr's study highlighted the main industries that can benefit from big data analytics, such as the manufacturing, retail, central government, healthcare, telecom, and banking industries [4].

4. Challenges with Big Data

Recent year's big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, and Thomson Reuters etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges.

- Collection of distributed data.
- Recognition of useful versus irrelevant Data.
- Accuracy, completeness and Timeliness of Data.
- Efficient storage and transfer.
- Privacy and security of Data.
- Fault Tolerance.
- Scalability and economic impact of implementation.
- Intelligent analysis
- Insightful and flexible presentation.

5. Conclusion

This review generate innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. In the information era we are currently living in, voluminous varieties of high velocity data are being produced daily, and within them lay intrinsic details and patterns of hidden knowledge which should be extracted and utilized. Hence, big data analytics can be applied to leverage business change and enhance decision making, by applying advanced analytic techniques on big data, and revealing hidden insights and valuable knowledge.

Accordingly, the literature was reviewed in order to provide an analysis of the big data analytics concepts which are being researched, as well as their importance to decision making. Consequently, big data was discussed, as well as its characteristics and importance. Moreover, some of the big data analytics tools and methods in particular were examined. Thus, big data storage and management, as well as big data analytics processing were detailed. In addition, some of the different advanced data analytics techniques were further discussed.

By applying such analytics to big data, valuable information can be extracted and exploited to enhance decision making and support informed decisions. Consequently, some of the different areas where big data analytics can support and aid in decision making were examined. It was found that big data analytics can provide vast horizons of opportunities in various applications and areas, such as customer intelligence, fraud detection, and supply chain management. Additionally, its benefits can serve different sectors and industries, such as healthcare, retail, telecom, manufacturing, etc.

References

1. Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2016)
2. Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)
3. Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)
4. Cebr: Data equity, Unlocking the value of big data. in: SAS Reports, pp. 1–44 (2012)
5. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analy-sis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)

6. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2017)
7. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1–24 (2012)
Big Data Analytics: A Literature Review Paper 227
8. Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)
9. EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)
10. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208 (2013)
11. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innova-tive Data Systems Research, pp. 261–272 (2011)
12. Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28 (2012)
13. Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011)
14. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: McKinsey Global Institute Reports, pp. 1–156 (2011)
15. Mouthami, K., Devi, K.N., Bhaskaran, V.M.: Sentiment Analysis and Classification Based on Textual Reviews. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 271–276 (2013)
16. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2017)
17. Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 (2011)
18. Sanchez, D., Martin-Bautista, M.J., Blanco, I., Torre, C.: Text Knowledge Mining: An Al-ternative to Text Data Mining. In: IEEE International Conference on Data Mining Work-shops, pp. 664–672 (2008)
19. Serrat, O.: Social Network Analysis. Knowledge Network Solutions 28, 1–4 (2009)
20. Shen, Z., Wei, J., Sundaresan, N., Ma, K.L.: Visual Analysis of Massive Web Session Da-ta. In: Large Data Analysis and Visualization (LDAV), pp. 65–72 (2018)
21. Song, Z., Kusiak, A.: Optimizing Product Configurations with a Data Mining Approach. International Journal of Production Research 47(7), 1733–1751 (2009)
22. TechAmerica: Demystifying Big Data: A Practical Guide to Transforming the Business of Government. In: TechAmerica Reports, pp. 1–40 (2012)
23. Van der Valk, T., Gijsbers, G.: The Use of Social Network Analysis in Innovation Studies: Mapping Actors and Technologies. Innovation: Management, Policy & Practice 12(1), 5–17 (2010)
24. Zeng, D., Hsinchun, C., Lusch, R., Li, S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010)
25. Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., Keim, D.: Visual Analytics for the Big Data Era—A Comparative Review of State-of-the-Art Commercial Systems. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182 (2014)