International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

A Review on Big Data Challenges and Techniques Used to Overcome from IT

Bibhudutta Jena

School of Computer Engineering, KIIT University

Abstract: ***In today's era of software industry "big data" is a common term used in each and every field. As the name suggests big data means large or vast amount of data which are either in the form of structured, unstructured or in semi-structured format. The major sources of big data generation are social media data, search engine data, Black box data etc. Now-a-days big data is generally used in different domains for decision making process, Hence analyzing big data is a biggest challenge in trending time and different researches are going on how to analyze and process big data so that it can be used in different application domains for decision making process. This paper provides a summary about different challenges that are generally associated with big data analysis process and different application of big data in different domains of software industry. This paper also provides a basic description of different framework which is generally used to process the big data. The main aim of this paper is to find out some optimization techniques to overcome from different challenges which are occurred during big data analysis process. In general optimization means to get the possible results under the given circumstances and big data optimization means to optimize the analysis process of big data.***

Keywords: Big Data, Optimization, Hadoop, Tableau, Processing Capabilities

1. Introduction

Concurrent research in multiple directions from social science to computer science, mathematics and physics, is distinguished by the availability of large amounts of data. Data is a necessary requirement on each and every field of computer application, as the users of computers gradually increasing day by day huge amount of data is generated from multiple sources. These huge amount of data are generally treated as big data. The enlarged volume of concurrent system allows the assembling, storage and analysis of vast amounts of data which only a few years ago would have been not possible. These new data are producing huge quantities of information, and storing it by using new computing methods and databases. There are many matters are generally rising from big data, from computational capacity to data manipulation techniques, all of which are creating challenges for analysis process. Big data is a docket that generally refers a large dataset that cannot be kept in a traditional Structured Query Language (SQL) database and needs a NoSQL (not only SQL) database which can be handle the large amounts of data. These big data are generated from social media, telco industry, transport industry etc.

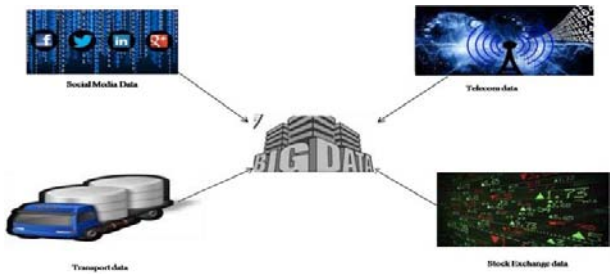

Figure 1: Big data Sources

The above figure [1] describes the sources where the big data are generally generated and are generally consider for big data analysis process. The rest of the paper are organizing as follow, Section 2 generally describes the literature survey, Section 3 gives a brief idea about the challenges on big data analysis, Similarly section 4 describes an idea about Nosql data bases which are generally used to store the big data and finally section 5 concludes the paper.

2. Literature Survey

As James Kobielus states [1], results from big data projects may be materialize either as revenue gains, cost reductions, or risk mitigation which have an easy and measurable Return on Investment (ROI). Similarly A survey by Forrester Research [2] indicated that most companies are relying on a mix of different technologies to enable the keep and operating of big data. In the below the paper has suggested some basic application of big data in different domain. Big data technologies are predominant in giving analysis more accurately, which may consider to more tangible decision-making process. Currently big data is widely used to optimize business processes. Retailers are able to optimize their stock on basis of predictions generated from social media data, web search trends and weather forecasts [3]. We can now also get the benefit from the data generated from wearable devices such as smart watches or smart bracelets. Take the Up band from Jawbone as an example: the armband gathers data on the basis of our calorie consumption, activity levels. Analyzing such volumes of data will bring entirely new insights that it can feed back to individual users [4]. The computing power of big data analytics enable us to detect entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns [5], In most of the sports have now held big data analytics. We have the IBM SlamTracker tool used for tennis tournaments; video analytics is generally used to track the performance of every player in a football or baseball game [6], Big data analytics also used to improving science & research in different domains of science. Science and research is currently being modified by the different possibilities big data escorts. Consider, the example, CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the

world's largest and most powerful particle accelerator [7]. In similar manner big data is applied rapidly in enhancing security and enabling law enforcement [8].

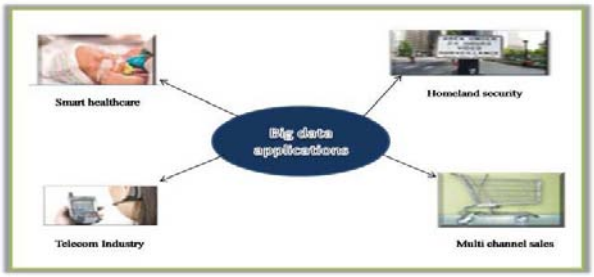

Figure 2 .Application of Big data

The above figure [2] describes some basic application of big data, although there is a wide range of application of big data are present some of them are discussed in the above.

3. Challenges

The role of benchmarking becomes even more apposite as a process for putting and recognizing better the internals of a particular platform. On the other hand, benchmarks are used to estimate different systems using both technical and economic metrics that can be able to monitor the user in the process of finding the right platform according to their needs. The basic challenges generally faced during the analysis of big data are described in the following part.

Capturing data: As big data refers to the vast amount of data so to capture and live streaming of this type of data is a very difficult task.

Storage: Storage is also a biggest issue during the analysis process of big data, How to store this large amount of data is a challenge for us.

Searching: Searching this vast amount in the search engine is also a biggest issue raised in current research domain.

Transfer: How to transfer the large amount of data from data storage place to hdfs during analysis process is a biggest challenge for us.

Presentation: Once the analysis process is over to show the result is also an biggest challenge .Different algorithms, framework are used to present this analysed result.

Basically the challenges of big data can be described by 5 V'S and it also specifies the characteristic of big data analysis. Different optimization techniques and tools are used to overcome from these challenges and helpful foe the decision making process.

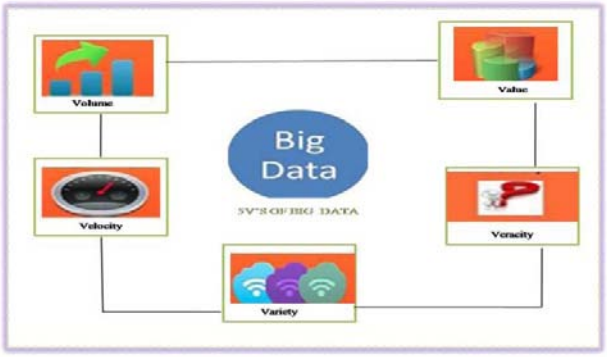

Figure 3: 5V'S Of big data

The above figure describes the 5V'S of big data that are consider as the major challenges in big data analysis.

Volume: Volume signifies large amount of data. As the paper already describes in the above that big data concept deals with large amount of data.

Velocity: Velocity means the rate at which data will spread in different domain, for example consider a photo post in social media will viral in few seconds now.

Variety: Big data consist of different types of data as we described in the earlier section. These type of data are structured data, unstructured data, semi-structured data.

a) Structured data: These data are organized in the predefined format and the data that resides in fixed fields within a record or file. The formatted data has their own entities and mapped attributes. The example of these type of data are traditional dataset etc.

b) Unstructured data: Unstructured data is a set of data that might or might not have any logical or repeating patterns. It generally consist typically of metadata, i.e., the additional information related to data. It consists of data obtained from files.The example of unstructured data are social media data, email data.

c) Semi-structured data: Semi structured data also known as having a schema-less or self-describing structure, refers to a form of structured data. Examples of these type of data are: Json files,log files etc.

Veracity: Veracity refers to uncertainty of data or trustworthiness of data i.e. the obtained data is correct or not.

Veracity plays an important role for decision making process during big data analysis.

Value: Then there is another V to take into account when looking at Big Data: Value! It is all well and good having access to big data but unless we can turn it into value it is useless. So you can safely argue that 'value' is the most important V of Big Data.

4. Tools & Techniques Used To Overcome Challenges

NOSQL is a term basically refered to define a class of non-relational databases that can scale horizontally to very large data sets but never give ACID guarantees . Nosql data keeps data vary widely in their offerings and have some definite features of itsown. The CAP Theorem provided by Eric Brewer in 2000 describes that it is impossible for a distributed environment to be consistent, available, and partition-tolerant at the same time. Consistency meaning is that all copies of data in the system show the same to the outside observer at all times. Availability defines that the system as a whole continues to run in case of a node failure [9]. Partition-tolerance generally defines that system must be work during the arbitrary message loss.

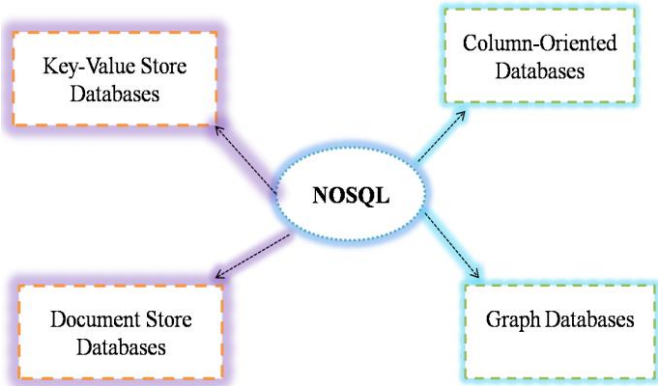

Figure 4: Nosql database types

The above figure [4] describes different types of databases are present in NOSQL which are generally used to store the unstructured and semi-structured data.

- Key-Value Store Databases

The key-value database are quiet simplistic, but are efficient and powerful model. The data have two parts, a string which presents the key and the actual data which is known as value thus creating a "key-value‟ pair. These contains are uniform to hash tables where the keys are used like indexes, to making it faster than RDBMS .Some examples of key value store databases are: Amazon Dynamo DB, RIAK etc.

- Column-Oriented Databases

In this type of database, the data is also involved with a “key” but it is organized by “columns” and the columns can be gathered by “family”. With the column notion, this type of database is not so far of the relational databases but there’s no specific column definition (no schema) hence the storage is more flexible .Some examples of column-oriented databases are: Bigtable , Cassandra etc .

- Document Store Databases

In this kind of database each key is associated with a “document” (usually formatted in JSON or XML). Obviously it’s possible to store everything in each “document” so this is the best solution to combine structured data (JSON or XML) with data store flexibility .Some examples of Document Store Databases are Mongo DB, Couch DB etc.

- Graph Databases

The databases which store data in the form of a graph is generally called as graph database. It constructed by nodes and edges, where nodes are considered as the objects and edges are considered as the relationship between the objects. The graph also contains of some properties related to nodes.

4.1 Tableau Software

Visualizing data is important regardless of the size of the data because it translates information into insight and action. The approach to visualizing Big Data is especially important because the cost of storing, preparing and querying data is much higher. Therefore, organizations must leverage well-architected data sources and rigorously apply best practices to allow knowledge workers to query Big Data directly.

5. Conclusion

The increased capacity of contemporary computers allows the gathering, storage and analysis of large amounts of data which only a few years ago would have been impossible. To work with large volume of data gathered through various applications in multiple locations is a challenging and rewarding task. Accessing valuable information from data means to combine qualitative and quantitative analysis techniques [10].

These new data are providing large quantities of information, and enabling its interconnection using new computing methods and databases. This paper has aimed, to describe some of the main issues that are relevant when setting up and optimizing a big data project and concentrated the attention first on the managerial task of setting up a big data project using insights from industry leaders, then it described in some detail at the available data management and processing technologies for big data.

References

[1] D. Beaver, S. Kumar, H. C. Li, J. Sobel and P. Vajgel, “Finding a needle in Haystack: facebook's photo storage”, Proceedings of the 9th USENIX conference on Operating systems design and implementation, Vancouver, BC, Canada, **(2010)** October 4–6, pp. 1-8.
[2] J. Cui, T. S. Li and H. X. Lan, “Design and development of the mass data storage platform based on Hadoop”, Journal of Computer Research and Development, vol. 49, **(2012)**, pp. 12-18.
[3] F. Mu, W. Xue, J. W. Shu and W. M. Zheng, “Metadata management mechanism based on route directory”, Journal of Tsinghua University (Sci & Tech), vol. 49, no. 8, **(2009)**, pp. 1229-1232.
[4] L. N. Song, H. D. Dai and Y. Ren, “A learning method of hot-spot extent in multi-tiered storage medium based on huge data storage file system”, Journal of Computer Research and Development, vol. 49, **(2012)**, pp. 6-11.
[5] G. G. Zhang, C. Li, Y. Zhang and C. X. Xing, “A kind of cloud storage model research based on massive information processing”, Journal of Computer Research and Development, vol. 49, **(2012)**, pp. 32-36.

[6] S. Yu, X. L. Gui, R. W. Huang and W. Zhuang, "Improving the storage efficiency of small files in cloud storage, Journal of Xi'An Jiao Tong University, vol. 45, no. 6, **(2011)**, pp. 59-63.
[7] Nicolas Sicard, B´en´edicte Laurent, Michel Sala, Laurent Bonnet, " REDUCE, YOU SAY: What NoSQL can do for Data Aggregation and BI in Large Repositories " , 2011 22nd International Workshop on Database and Expert Systems Applications. .
[8] Karamjit Kaur and Rinkle Rani , " Modelling and Querying Data in NoSQL Databases ", 2013 IEEE International Conference on Big Data.
[9] Richard K. Lomotey and Ralph Deters , " Terms Mining in Document-Based NoSQL: Response to Unstructured Data " , 2014 IEEE International Congress on Big Data.
[10] S. Ghemawat, H. Gobioff and S. T. Leung, "The Google file system", Proceedings of the 19th ACM symposium on Operating systems principles, Bolton Landing, NY, USA, **(2003)** October 19–22, pp. 29- 43.

Author Profile

Bibhudutta Jena, is a CSI Accredited Student. Currently pursuing M. Tech (Computer Science and Engineering) at the School of Computer Engineering, KIIT University, Bhubaneswar. His area of interest Data Analytics, Data mining etc. He can be reached at bibhuduttajena728[at]gmail.com