Comparative study of analytic tools of BigData

Barkha M.Joshi
Computer Engineering Dept.
Sardar Vallabhbhai Patel
Institute of
Technology,vasad,India.

***Abstract*—** BigData is the advanced topic of the datamining. It contains the data may be Tera byte, Zetta byte and bronto bytes. Traditional application is insufficient to analysis such type of structured, unstructured and semi structured data. Analyze these BigData with advance tools like Hadoop, MongoDB, Cassendra and Elasticsearch. These tools are used to identify the knowledge from large and complex type of data. In this paper shows the different dimensions and those dimension data handle by the different tools and also show the comparison of those different analytical tools. From the study of tools try to identify the which method is best for handle the BigData in the distributed environment.

***Keywords*—** *BigData, Dimension, BigData Analytics tools*

INTRODUCTION

BigData is the developing discipline. Traditional applications are inadequate to handle the BigData. Some organization take smart and effective decision based on the BigData. It provide the predictive and effective decision making ,cost effectiveness and marketing effective ness. There are different fields that comes under the BigData title like black box data, social media data, stock exchange data, power grid data, transport data and search engine data. These data are operational and analytical type of BigData.

1) What is BigData?: BigData is collection of complex and large number of data that may be in Terabyte, zetta byte or bronto byte and those data are not handle by the traditional database processing application so this data is called a BigData. "BigData" tends to refer to the use of predictive analytics, use behaviours analytics and certain other advanced data analytics methods that extract the values from the data and stored in the particular size of data.

2) Applications: BigData contribute to Education process like generation of grading ,refresh the courses and also it is use in the online classroom activities. BigData contribute the healthcare and medicine field to early stages of disease diagnosis and evidence of the medicines. BigData contribute to the public sector like food and drug administration, social media, weather forecasting processes. It also contribute to transportation like route planning and congestion management by traffic planning. It contribute to finance and crime detection like misuse of credit and debit card ,risk management and money laundering.

3) Types: There are two type of BigData 1) Operational BigData and 2) Analytical BigData.

   1) Operational BigData: It's real-time, interactive workload where data is primarily captured and stored. This database is easy to manage, cheaper and faster to implement. Technology used by these data are NoSQL.
   Examples : MongoDB

   2) Analytical BigData: It's a Massively Parallel Processing data base system .generally use this data by data scientist and use technology like MapReduce and MPP database.
   Example : Hadoop

I. DIMENSION

BigData uses the 5,Vs dimensions concept to understand the different terminologies.. BigData initially consisted of three dimensions namely volume**,** velocity, and variety**.** These three attributes pretty much gave the essence of the definition of BigData. There are another attribute that were added in the list, termed as veracity and value.

1) Volume: It's a best characteristic of BigData. It represent the size of data or can say quantity of data. volume of data can be large like terabytes, zettabytes and brontobytes. These type of data can be generated by the emails, sensor data and message, photos and video clips of social networking site.

2) **Velocity:** It represent the growth of data or we can say the speed of movement of data. It refers to the speed at which data is being generated, produced, created, or refreshed. Generally these type of data we can collect from the social networking site. Every time user wants to read the

new messages instead of the old messages. It discard the old messages and pay attention to recent updates.

3) **Variety:** It means the different type of data it may be structured, unstructured and semi structured data. Data may collect from the financial environment like sale data. Variety of data means it may be different format like video, audio, images, text etc. These data are divide in to three parts like

Structured data: Relational data.
Semi Structured data: XML data.
Unstructured data: Word, PDF, Text, Media Logs.

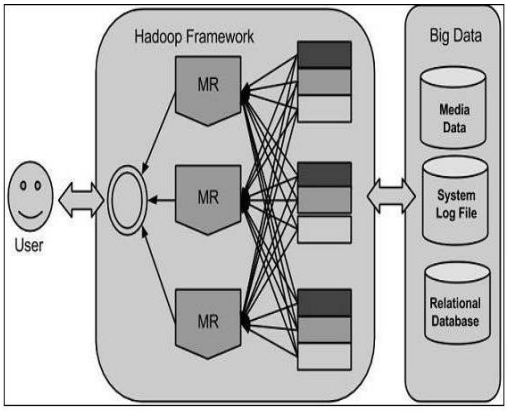
4) **Veracity:** It just about data quality. Data must be collect from the reliable data source and it must be understandable by the user. Large number of data may store in the data warehouse, so only select those data which is required for the further process so we can say that the data always the accurate and reliable.

5) **Value:** Value starts and ends with the business use case. It is a most important characteristic. Accurate data are useful for the business process. Large amount of data and variety of data may access to using the different algorithm or analytics tool.

II. BIGDATA ANALYTICS TOOLS

BigData analytics refers to the process of collecting, analyzing and organizing the large set of data to find the patterns and useful information. These knowledge we can generate through different analytical tools like Hadoop, MongoDB, Cassandra, Elasticsearch.

1. Hadoop



Hadoop uses the MapReduce concept to divide the data in parallel way to different location and collect the result from the different places and perform the integration on the centre location. Hadoop is Open Source Project. Complete statistical analysis perform on the huge amount of the data. Hadoop is an Apache open source framework. It is written in java, and allows distributed processing of large datasets across clusters of computers using simple programming models. Hadoop is provide the efficient, fault tolerance high availability and distributed data automatically on the different machines.

2. MongoDB
MongoDB is an open-source document database. It leading NoSQL database. It is written in C++. It provide to create and deployed scalable and performance-oriented database. It is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. It works on concept of collection and document. MongoDB contain less schema,no complex join and it support the dynamic queries.

3. Cassandra
Cassandra is a distributed database from Apache. It is that is highly scalable and designed to manage very large amounts of structured data. It is scalable, fault tolerant and consistent and it a column oriented database. Cassandra is being used by some of the biggest companies such as Facebook, Twitter, Cisco, Rackspace, ebay, Twitter, Netflix.

4. Elasticsearch:
Elasticsearch is a real-time distributed and open source full-text search and analytics engine. is a real-time distributed and open source full-text search and analytics engine. It is used in Single Page Application (SPA) projects. Elasticsearch is open source developed in Java and used by many big organizations around the world. It is accessible from RESTful web service interface and uses schema less JSON (JavaScript Object Notation) documents to store data. It is built on Java programming language, which enables Elasticsearch to run on different platforms. It enables users to explore very large amount of data at very high speed.it is scalable upto petabytes of structured and unstructured data.

III. COMPARISION

	Hadoop	**MongoDB**	**Cassandra**	**Elasticsearch**
Discriptio	Distributed Open	Distributed NOSQL	Wide column	Distributed RESTful

n	Source Java Framework	type database	store Dynemo DB	morden search
Database Model	Java Framework	Document store	Wide Column store	Search engine
Operting System	Linux	Linux Windows OS X	Linux Windows OS X	All OS with Java VM
Impleme ntation Language	Java	C++	Java	Java
Method	MapReduce MPP	MapReduce	MapRedu ce	ES Hadoop connector

TABLE 1 COMPARISION OF BIGDATA TOOLS

CONCLUSION

Study shows that the used analytical tools work for distributed environment and finding the useful information from the complex and large dataset. The comparison shows that the Hadoop an open source apache based software tool provide useful information from distributed large dataset in parallel way using MapReduce concept.

REFERENCES

[1]. 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)
[2]. Bakshi, K.: Considerations for BigData: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)
[3]. Cebr: Data equity, Unlocking the value of BigData. in: SAS Reports, pp. 1–44 (2012)
[4]. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for BigData. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)
[5]. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The BigData Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)
[6]. Economist Intelligence Unit: The Deciding Factor: BigData & Decision Making. In: Capgemini Reports, pp. 1–24 (2012) BigData Analytics: A Literature Review Paper 227
[7]. Elgendy, N.: BigData Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)
[8]. EMC: Data Science and BigData Analytics. In: EMC Education Services, pp. 1–508 (2012)
[9]. 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 (2011)
[10].Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A Self-tuning System for BigData Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272 (2011)
[11].Kubick, W.R.: BigData, Information and Meaning. In: Clinical Trial Insights, pp. 26–28 (2012)
[12].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)
[13].V.Maria Antoniate Martin,Dr. K.David, A.Vignesh Big Data and Its Challenges,International Journal of scientific Research in Computer Science, Engineering and Information Technology