ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 1, pp: (67-75), Month: January - March 2016, Available at: www.researchpublish.com

A Comparative Study on Data Analytics and Big Data Analytics

[1]Ananta Chandra Das, [2]Santosh Kumar Pani, [3]Sachi Nandan Mohanty

[1]M.Tech School of Computer Engineering, KIIT University, Bhubaneswar, 751024, India
[2]Associate Professor at School of Computer Engineering, KIIT University,
[3]Assistant Professor at School of Computer Engineering, KIIT University, Bhubaneswar, 751024, India

Abstract: **Data is a part of information that is formatted in some special approach. In general way data can be exist in various forms like numbers or text on a paper, as bits or bytes in computer memory, or as facts in the brain of human beings. Moreover, data is present each and every place in the universe. There is no limit of the data present all over the universe. Since the day when human civilization starts, we always exist with some data. Our day to day life deals with some data with or without any reason. The data can be of any type like remembering someone's name or saving a playlist of thousands of songs. But, with the development of the human civilization the rapid growth of the data has already come into the existence. As we live in the twenty first century we exist with a new concept of data which is known as "BIG DATA". In last century the human civilization faced a new concept of manipulating data known as Data Analytics in order to make the business profitable and the life more simple that deals with large volume of data. In this paper we have discussed about different types of data, the problems in the traditional data analytics process. After that we discussed about what is big data and different parameters of big data. How the problems in traditional data analytics is being resolved by big data analytics also discussed in this paper.**

Keywords: **Big data; data analytics; data storage procedure; big data landscape; hadoop; Hadoop Distributed File System(HDFS); Business Intelligence (BI).**

1. INTRODUCTION

The sudden and excessive growth of data in the beginning of the twenty first century has become a challenge for human civilization. In the initial stage that mans around the latter part of the 20th century the concept of relational database came into the lime light[8]. Introduction of relational database was revolutionary step in the world of data where data can be stored in a table and can be easily processed according to the need. Then to make the life easier, some analysis over the data was needed which solved a lot of problems of the day to day life. But the real problem came into the picture just after the internet was introduced. Due to the vast use of internet, we got some data having huge-volume, high-velocity, wide-variety[2]. The relational database could not able to handle and process that data. Hence, a new type of data known as Big Data was introduced with different concepts and different technologies. According to the behaviour, the data can be categorized into the following types.

1.1. Structured Data:

The data which can be stored in the relational database table in a row - column format [6].

As the name suggests, structured data is having some specific structure and that structure is defined by the data model which is created by the organization. The organization first defines the data model i.e., a model which will allows to store, process and access the organizational data. The model has to define the properties of the data that will be stored. The property of the data includes : data type (numeric, alphabetic, name , date, etc.) and some restrictions on the data (size of the data, number of characters, etc.). The major advantage of structured data is that, it can be easily stored and analyzed. Due to the high cost and limitations of the storage space and processing techniques, relational database is the only way to store and process the structured data effectively.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 1, pp: (67-75), Month: January - March 2016, Available at: www.researchpublish.com

Managing Structured Data:

Basically structured data is managed by using Structured Query Language (SQL). SQL is a programming language for managing the data in the RDBMS[6]. It was developed by IBM in the year 1970 and later it was developed commercially by Relational Software, INC. (Presently Oracle Corporation).

1.2. Semi-Structured Data:

The data which is in the form of structured data but that does not fit with the data models defined for the structured data is known as semi-structured data[6]. The semi - structured data can't be stored in the relational databases or other forms of a data table, rather it is stored in some specific type of files which contain some tags. The tags or the markers are separated with some semantic rules and enforce the data to be stored with a hierarchy.

This type of data is increasing rapidly after the web is introduced where different forms of the data and completely different kinds of applications need mediums for exchanging the information like XML and JSON.

1.3. Unstructured Data:

The data which doesn't have any specific structure, therefore, can't be stored in a row-column format of a traditional database is known as unstructured data[6]. As the name suggests, the unstructured data is the opposite of structured data. Hence, it can't be stored in fields in a database.

Example : text files, image files, audio files, video files, web pages etc.

Now-a-days, the volume of unstructured data is growing so rapidly that, it is very difficult to handle and analyze the data. So, to analyze the unstructured data it requires more knowledge with some advanced technologies.

2. TRADITIONAL DATA STORAGE PROCEDURE

In traditional data storage procedure the data model defines some properties. The structured data needs to satisfy all the properties defined by the data model in order to be stored in the database. So, the data will only be accepted if and only it satisfies all the properties defined by the data model. If the data doesn't satisfy at least a single property then the data will be rejected. The data basically stored in the row column format in relational databases. SQL is used to handle the data and to process it.

3. PROBLEMS IN TRADITIONAL DATA ANALYTICS PROCEDURE

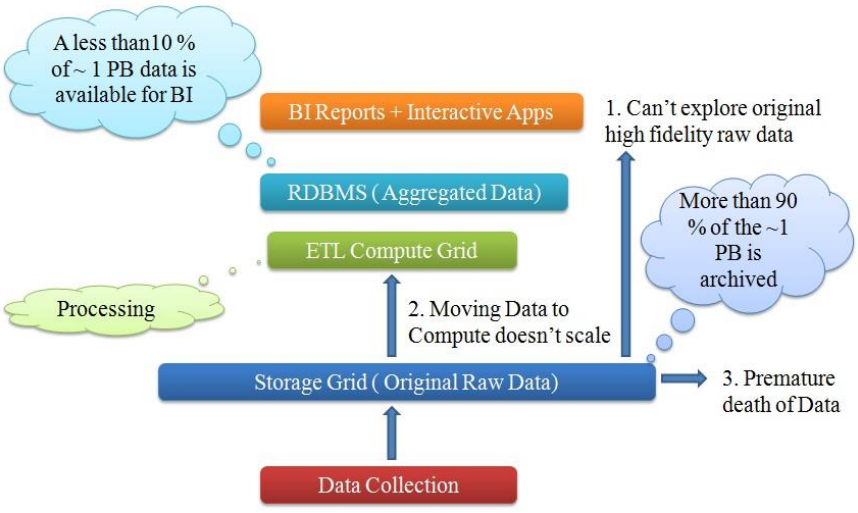

Fig 1: Block diagram showing Problems in Traditional Data Analytics Procedure

In Traditional data analytics procedure it not so easy to process the data correctly and efficiently. In traditional system, the data is first collected from various sources, then it is stored in the database (Storage grid). The problem arises when large amount of data with high velocity comes to the storage grid. The storage grid can't handle the huge amount of data, even if it stores the data then most of the data gets archived. Suppose 1 PB of data comes to the storage grid in order to analyze,

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 1, pp: (67-75), Month: January - March 2016, Available at: www.researchpublish.com

then 90 % of the data gets archived due to shortage of storage space. Now, only 10 % of the data remains for the analysis and the entire 90 % of the data gets archived considering the premature death of the data. The remaining 10 % of the data goes to the ETL compute grid[4]. The ETL (Extraction, Transformation, Load) consists of the following three steps :

i. Extraction : In this step, the source system gets connected and the necessary data is selected and collected for analysis and processing.

ii. Transformation : In this step a series of rules applied to the data which is extracted and after that the data is converted into a standard format.

iii. Load : In this step, the data which is extracted and transformed is imported to the desired and targeted data storage.

After that, data goes to the RDBMS where the BI(Business Intelligence) reports are generated. The reports generated by this process are not efficient so unable to provide the effective business solution. So, some advanced tool is required which can analyze the entire data.

4. INTRODUCTION TO BIG DATA

In a general meaning Big data is the huge amount of data. But the only parameter i.e., amount can't express the definition of big data completely. So, basically it is identified according to the large-volume, high-velocity and wide-variety of information.

4.1 3 V's of Big Data:

i. Volume:

In this case, the amount of data is taken into consideration. If large-volume of the data comes, then it will be difficult for the RDBMS to store and manage that amount of data. The name itself indicates that, the size of the data is the major parameter.

ii. Velocity:

It this case, the rate of speed at which the data is captured is the major concern. If large-volume of data comes at fraction of time, then that will be very difficult for the RDBMS to capture and process the data.

iii. Variety:

In this case, the type of the data becomes the major parameter. the RDBMS can process only structured the data. So, if the data is unstructured, then that can't be captured and process by the RDBMS.

5. SOLVING THE PROBLEMS OF TRADITIONAL DATA ANALYTICS WITH BIG DATA

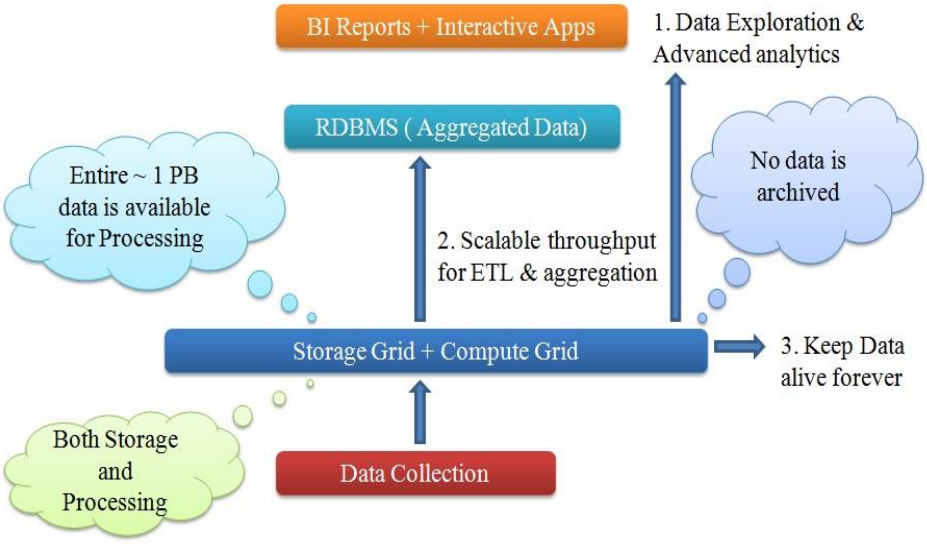

Fig 2: Block diagram showing the Solution of the problem in Traditional Data Analytics Procedure

If we can replace traditional storage grid with an advanced storage grid which itself used as a compute grid then the problem can be easily resolved. In place of traditional data storage grid if we will use some big data tools like Hadoop, then our problem will be solved. Hadoop stores the data in a distributed file system known as HDFS(Hadoop Distributed File System) due to which the storage and processing become easier and faster. Hive (DW system) is the ETL tool in Hadoop. After analysis the entire data stored in the RDBMS where BI reports are generated which is the most accurate and provide the effective business solution

7. BIG DATA LANDSCAPE

7.1. Big data infrastructures:

- **Hadoop-** Hadoop is a tool which provides an ecosystem to store, process and analyze the data. It works with distributed file system which breaks a large file into small files and distributes the data across different nodes in order to process the data faster.

- **NoSQL -** It Stands for Not-Only-SQL. It is used to process large volumes of unstructured or semi-structured data. Hbase is one of the popular NoSQL database which can work with Hadoop.

- **Massively Parallel Processing Database-** This database distributes the data in different segments across multiple nodes, and then process them in parallel by using SQL. Basically it runs on expensive hardware. The difference between MPP and Hadoop is that, MPP runs on expensive hardware whereas hadoop runs on cheaper commodity hardware.

7.2. BIG data analytics technology:

Some of the big data infrastructure technologies provide the data analysis in some manner. The big data analytics technologies specifically provide the facility for analysis. The followings are some of the sub-categories.

- **Analytical Platforms-** It integrates and analyses data to discover some new knowledge, which helps organizations for better decision making. The main focus in this case is to provide the solution of the users need in timely manner.

- **Visualizations -** The main objective is to visualize the data. It takes the data and presents the data in some visual forms in order to extract the information from it.

- **Business Intelligence -** It is specifically used to provide business solutions. It collects, integrates, and analyses data that is required for a business to get more profitable solutions. It enables users to build applications that help organizations to learn and understand their business.

7.3. Applications:

Applications are generally used to analyze big data and offer optimized insights to the end-users. Following are some of the application fields :

- **Ad Optimization -** MediaMath is the first demand-side platform (DSP), changing the way digital media is purchased, and creating a new, more efficient way for advertisers to reach consumers, individually, at scale.

- **Publisher Tools-** Visual Revenue is a real-time predictive analytics platform developing a suite of tools providing decision support for editors content.

- **Energy-** AutoGrid takes the data from smart meters, voltage regulators, thermostats to assist customers track the amount of power used, scale back waste, balance the grid, increase the system operations and forecasts the future consumption.

8. MAJOR INDUSTRIES USING BIG DATA

Table 1: Type of industries that use big data in different fields

Industry	Area where big data is used
Retail	Supply Chain Analysis, Dynamic Pricing, Sentiment Analysis
Banking	Modeling true risks, Fraud Detection, Threat Analysis, Trade Surveillance, Credit Scoring and Analysis
Advertising	Ad targeting, Recommendation Engine, Click Fraud Detection
Telecommunications	Customer churn Prevention, Network Optimization, Calling data record analysis
Healthcare	Gene Sequencing, Bioinformatics, Pharmaceutical Research, Prediction of diseases
Manufacturing	Product Research, Engineering Analysis, Quality Analysis

9. RAPID GROWTH OF THE GLOBAL DATA

According to a survey conducted by CMC, the production of the data is expanding at an astonishing pace[15]. It predicts there will be 4300 % increase in annual data generation by 2020. According to that survey, data production will be 44 times greater in 2020 than it was in 2009.

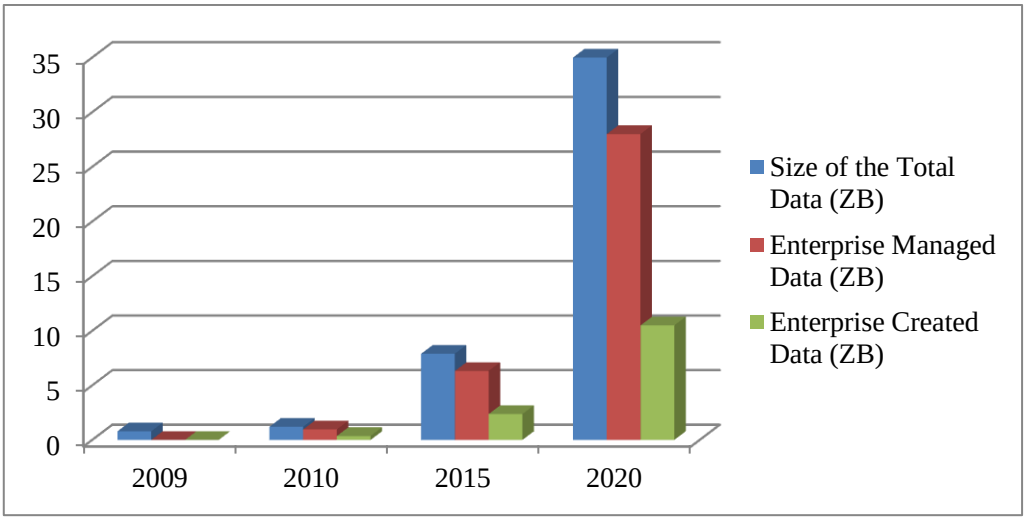

Fig 1 : Graph showing the rapid growth of the size of the data

10. OPEN SOURCE TOOLS FOR BIG DATA

10.1. Big Data Analysis Platforms and Tools:

10.1.1. Hadoop:

It is developed by Apache Foundation. As discussed earlier A whole ecosystem of technologies designed for the storing, processing and analyzing the data. To extend the capabilities of hadoop, the Apache organization also provides some related technologies and projects. Some technology providers provide support for Hadoop and supported technologies.

Platforms : Windows, Linux, OS X.

10.1.2. MapReduce:

It is originally developed by Google. It is described as a programming model which processes huge amounts of data on large clusters of computer nodes very quickly. Mostly this software framework is used by Hadoop. It is also used by different applications which are used for data processing.

Platforms : OS Independent.

10.1.3. Storm:

It is a product of Twitter. It provides distributed real-time computation facilities and is described as the "Hadoop of realtime". It is highly scalable, robust and works with a lot of programing languages.

Platforms : Linux.

10.1.4. GridGain:

Being compatible with the HDFS, GridGain is the alternative to MapReduce used in Hadoop. For faster analysis of the data, it provides in-memory processing.

Platforms: Windows, Linux, OS X.

10.1.5. HPCC (High Performance Computing Cluster)

It is a product of LexisNexis Risk Solutions. It offers better performance than Hadoop. It is available in free community versions as well as paid enterprise versions.

Platforms: Linux.

10.2. Database/ Data warehouse:

10.2.1. Cassandra:

It was a product of Facebook. But, now it is now maintained by the Apache Organization. Many organizations like Netflix, Twitter, Urban Airship, Reddit etc. use this.

Platforms : OS Independent.

10.2.2. HBase:

It is a product of Apache organizations. It is a non-relational data storage for Hadoop. Some of the features include modular scalability, consistent read and write, fail-over support etc.

Platforms: OS Independent.

10.2.3. MongoDB:

It supports wide range of databases. It is a NoSQL database and it provides document oriented data storage, duplication, support for full index and highly available, etc.

Platforms: Linux, Windows, Solaris, OS X.

10.2.4. Neo4j:

Now it is a leading graph database in the world. The performance of Neo4j is around 1000 more than the relational database.

Platforms: Linux, Windows.

10.2.5. CouchDB:

It is basically developed for the Web. In this the data is stored in the JSON documents in order to access from the web or using JavaScript query.

Platforms: Linux, Android, OS X, Windows.

10.2.6. Hive:

It was initially developed by Facebook. But now it is used and developed by other companies. It is known as the data warehouse for Hadoop. It uses HiveQL for queries.

Platforms: OS Independent.

10.2.7. Hypertable:

It is developed by Zvents Inc. It is a NoSQL database that provides efficiency and performs faster which results in cost savings.

Platforms: Linux, OS X.

10.2.8. FlockDB:

It is developed by Twitter. It is also popularly known as database of. Social graphs are stored here. i.e., the information of followings and followers and blocked users etc.

Platforms: OS Independent.

10.2.9. Hibari:

Many of the telecom industries use this. It's an ordered key-value, storage of big data and guarantees high bandwidth and reliable.

Platforms: OS Independent.

10.2.10. Riak:

It is developed by Basho Technologies. It is a distributed NoSQL, key-value data store that offers high availability, fault tolerance and scalability.

Platforms: Linux, OS X.

10.3. Business Intelligence:

10.3.1. Knime:

It is known as Konstanz Information Miner, or KNIME. It offers very user-friendly data integration, processing, analysis, and exploration.

Platforms: Windows, Linux, OS X.

10.3.2. Palo BI Suite:

It includes an OLAP Server, Palo Web, Palo ETL Server and Palo for Excel.

Platforms: OS Independent.

10.3.3. BIRT:

It is the short form of "Business Intelligence and Reporting Tools". It is an Eclipse-based tool that adds reporting features to Java applications.

Platforms: OS Independent.

10.3.4. Pentaho:

It is used by more than thousands of companies. It offers BI tools and big data analytics tools with data mining, reporting and dashboard capabilities.

Platforms: Windows, Linux, OS X.

10.4. Data Mining:

10.4.1. Rapid Miner / Rapid Analytics:

It is the world-leading open-source system for data and text mining. RapidAnalytics is a server version of RapidMiner.

Platforms: OS Independent.

10.4.2. Mahout:

It is a project of Apache foundation and it offers algorithms for clustering, classification and batch-based collaborative filtering that runs on the top of Hadoop. The is used to build scalable machine learning libraries.

Platforms: OS Independent.

10.4.3. SPMF:

It is a Java-based data mining framework. It is basically used in sequential pattern mining, but also includes tools for association rule mining, sequential rule mining and frequent item set mining.

Platforms: OS Independent.

10.4.4. Weka:

It stands for "Waikato Environment for Knowledge Analysis." It offers a set of algorithms for data mining that can be applied directly on data or can be used in another Java application. It is sponsored by Pentaho.

Platforms: Windows, Linux, OS X.

10.5. File Systems:

10.5.1. Gluster:

It was developed by Red Hat. It offers object storage for very large datasets. It can be used to extend the capabilities of Hadoop beyond the limitations of HDFS .

Platforms: Linux.

10.5.2. Hadoop Distributed File System:

It is popularly known as HDFS. It is the primary storage system for Hadoop. It quickly distributes the data into several nodes in a cluster and also replicates the data. It provides reliable, fast performance.

Platforms: Windows, Linux, OS X.

10.6. Programming Languages

10.6.1. Pig:

It is an Apache Big Data project. It is a data analysis platform that uses a textual language called Pig Latin and it produces sequences of Map-Reduce programs.

Platforms: OS Independent.

10.6.2. R:

R is developed by R core team. R is a programming language and an environment for statistical computing and graphics and is provided by R foundation for statistical computing. It provides a set of tools that make it easier to manipulate data, perform calculations and generate charts and graphs.

Platforms: Windows, Linux, OS X.

10.7. Data Aggregation and Transfer

10.7.1. Sqoop:

It is developed by Apache Software Foundation. It is used to transfer the data between Hadoop and RDBMS. It imports individual tables or entire databases to HDFS. It provides the ability to import from SQL databases straight into the hive data warehouse.

Platforms: OS Independent.

10.7.2. Flume:

It is a project by Apache organization. It collects, aggregates and transfers large amount of log data from applications to HDFS. It is written in java and robust and fault-tolerant.

Platforms: Windows, Linux, OS X.

11. FUTURE WORK

- Detecting the problems in Big data analytics
- Solution of the detected problem
- Defining a universal model for all types of data analytics

12. CONCLUSION

After this survey we got that data is growing rapidly irrespective of the type and size. So, we can conclude that in the near future we may deal with some new data definition having more advanced characteristics and there will be some more advanced tools which can solve the problems caused by that new type of data.

REFERENCES

[1] S. Sruthika, N. Tajunisha, A Study On Evolution Of Data Analytics To Big Data Analytics and Its Research Scope. in: Paper presented in *IEEE Sponsored International Conference on Innovations in Information Embedded and Communication Systems ICIIECS, 2015*.

[2] Y. Demchenko, P. Membrey (2014), Defining Architecture Components of the Big Data Ecosystem. in: Paper published in Collaboration Technologies and systems (CTS), *pp*-104 - 112. doi : 10.1109/CTS.2014.6867550.

[3] D. Singh, CK. Reddy (2014), A survey on platforms for big data analytics. In: paper published in Journal of Big Data. doi : 10.1186/s40537-014-0008-6.

[4] H. Hu, Y. Wen, TS. Chua, X. Li (2014), Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. In: Access, IEEE (Volume:2), *PP- 652 - 687. doi :* 10.1109/ACCESS.2014.2332453.

[5] RJT. Morris, BJ. Truskowski (2003), The Evolution of Storage Systems. In : IBM Systems Journal (Volume:42 , Issue: 2). PP - 205-217. doi: 10.1147/sj.422.0205.

[6] RIDER, FREMONT. The Scholar and the Future of the Research Library. 236 pp. New York, Hadham Press, 1944.

[7] Automatic data compression, published in Communications of the ACM, Volume 10 Issue 11, Nov. 1967Pages 711-715, doi:10.1145/363790.363813.

[8] Y. Chen , S. Alspaugh, and R. Katz, Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads, *Proc.VLDB Endowment*, vol. 5, no. 12, pp. 1802_1813, 2012.

[9] J. Gantz and D. Reinsel, The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, In : Proc. IDC iView,IDC Anal. Future, 2012.

[10] B. Franks, Taming the Big Data TidalWave: Finding Opportunities in Huge Data Streams With Advanced Analytics, vol. 56. New York, NY, USA:Wiley, 2012.

[11] N. Tatbul, Streaming data integration: Challenges and opportunities, In : *Proc. IEEE 26th Int. Conf. Data Eng. Workshops (ICDEW)*, Mar. 2010, pp. 155_158.

[12] *What is Big Data*, IBM, New York, NY, USA [Online].(2013) Available: http://www-01.ibm.com/software/data/bigdata/.

[13] Wikibon. (2013). *A Comprehensive List of Big Data Statistics* [Online]. Available: http://wikibon.org/blog/big-data-statistics/.