Big Data: Tools and Technologies in Big Data

Jaskaran Singh
Student
Lovely Professional University, Punjab

Varun Singla
Assistant Professor
Lovely Professional University, Punjab

ABSTRACT

Big data can be defined as data which requires latest technologies other than the traditional tools and techniques to handle it, due to its unstructured nature. Huge volume, various varieties and high velocity create lots of other challenges and issues regarding its management and processing. Big data is growing at an exponential rate so it becomes important to develop new technologies to deal with it. This paper covers the leading tools and technologies for big data storage and processing. Hadoop, Map Reduce and No SQL are the major big data technologies. These technologies are very helpful in big data management. Technologies based on Hadoop called Hadoop Eco system have also discussed. This Paper also throws some light on other big data emerging technologies. There are so many areas from which big data is being generated, this paper covered those areas and provide solutions for dealing with that data.

Keywords

Hadoop, Map Reduce, No SQL, Unstructured Nature, Hadoop Eco System.

1. INTRODUCTION

“Big data” word itself describes that it is a huge amount of data, but this is not the complete explanation of big data, if you want to understand it properly. For a complete understanding of big data you have to study all the basic properties of it. The main thing in big data is that it has no structure. This is the main difficulty to deal with it. Big data is beyond the structured data. It is very tough task to manipulate the big data due to its unstructured form. There are three basic properties of big data, Volume, Velocity and Variety.

1.1 Volume

The huge amount is the basic property of big data. These days there is an exponential growth in big data. Data is everywhere and it is generated very fast. Social media, Server logs are generating huge amounts of data on a daily basis.

1.2 Velocity

Velocity is another basic property of big data. This is the speed at which data is generated. Big data is generated very fast now days due to social media and the digital world. Thousands GB’s of data is generated everyday over the digital world.

1.3 Variety

Variety is one of the most important properties of big data. The basic identification of big, data is based on a variety of data. There are two basic types of data, structured and unstructured. There is also a semi structured data which is an extension of structured data. Unstructured data is difficult to handle with traditional relational database.

So no structure of data is a big challenge in itself. It is a big challenge for the traditional technologies and databases to deal with big data because it completely unstructured. Traditional databases are basically made for structured data. In this paper some new technologies which help in storage and processing of big data and Extension of traditional database tools are discussed.

2. RELATED WORK

Avita Katal, Mohammad Wazid, R H Goudar in “Big Data: Issues, Challenges, Tools and Good Practices” [1] discussed about big data issues, challenges and tools. What is big data and what are the basic properties of it like volume, velocity, variety, complexity, value has been discussed. Sources from which big data is coming and generated are also mentioned. Big data has a great importance in various projects like social media, sensor data, and log storage and risk analysis. In social media there is great importance of big data. Facebook is generating Terabytes of data on every day. This data is huge, fast and very complex means follow all the properties of big data. Many projects are going on in the market like in the private sector, government sector and in science field which deals with big data. Privacy and security, data access and sharing of information, storage and processing are the main issues and challenges in big data. The authors discussed that privacy and security are very sensitive issue and it can disturb someone’s personal life because the data that has taken for analysis, it is personal data of any individual, in case of social media, after discovery of some pattern may be that person does not want that it should be known to someone. Some technical challenges like fault tolerance and scalability are also discussed. Hadoop and map reduce has been discussed as the main tools and technologies to process and deal with big data. Hadoop is basically a framework on which map reduce works as a programming model. It works in batch processing means it divides the task into smaller units and then executes them parallel. At the end of the paper a comparison between Hadoop and grid computing tools is also shown.

Major issues related [2] to big data storage, management and processing have been discussed by Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money. Challenges that may be faced by big data in future due to exponential growth of big data are explained. Beyond volume, velocity and variety, data value and complexity are the other characteristics of big data which define how important data is and its complexity respectively. Storage and transport, management and processing of big data are the major issues highlighted by authors. Analytics is the biggest challenge under big data processing. Analytics in big data environment is same as finding needle in haystack.

Cloudera [3] figure out Hadoop able problems in their paper. This is the white paper by Cloudera Company. In this all the areas are discussed where Hadoop can be useful to implement and they also provide a solution to a specific problem with Hadoop. Banks need Hadoop to perform risk analysis on their customer’s profile. There is a different type of data present in the banking and financial institutes. This data is very sensitive and important due to its privacy. So this data needs better management and Hadoop is capable of managing this data. There are other areas also like advertisement targeting which helps companies to target perfect customer for their products and Hadoop is very much efficient in this area. Point of sale transaction analysis is now in huge demand, which figure out the customer’s buying pattern. Fraud analysis, analysing network data to predict failure, threat analysis or other areas in which Hadoop can be very efficient and helpful.

Kala Karun. A, Chitharanjan. K [4] discussed about Hadoop and Hadoop distributed file system infrastructure extensions. In this major enhancement in Hadoop in data storage, data processing and placement are also reviewed. The Authors have made a Comparison of Hadoop Infrastructure Extensions on the basis of scalability, fault tolerance, load time, data locality, data compression, etc. Hadoop is widely accepted in many areas, but its extensions, which are the improvements of Hadoop, can also be very helpful. HadoopDB, Hadoop++, Co-Hadoop, Hail, Dare, Cheetah, etc. are the main extensions of Hadoop and these are considered for comparison.

Min Chen, Shiwen Mao, Yunhao Liu [5] reviews the big data background and all the related technologies. Hadoop has also discussed as a big data processing tool. In this paper technical challenges faced by big data has been discussed. Applications of big data have also reviewed. Four major areas, data generation, acquisition, storage and analysis have been discussed widely. Background and technical challenges faced by each area has discussed in detail.

Characteristics and need of big data have been discussed by Sachchidanand Singh, Nirmala Singh [6]. Unstructured data and how to deal with it, is also discussed. Results shown in this paper infers that big data analytics is very much important to make business intelligent. Kapil Bakshi [7] mainly discussed about the analysis of unstructured data. Map reduce and Hadoop are the major tools for the analysis of unstructured data and it is widely discussed in this paper. Demchenko,Y, de Laat, C., Membrey, P. [8] Discussed about the basic definition of big data and also focused on the importance of the Hadoop Eco system. 5 V’s Volume, Velocity, Variety, Value and Veracity have been discussed as the main properties of big data. Big data analytics, security, data structure and models are the main components of the Hadoop Eco system. The author reviewed these core architectural components of the Hadoop Eco system. These components are very important in big data challenges.

3. PROBLEM SOLVING AREAS WITH BIG DATA

3.1 Customer Behaviour Analysis

When one company provides a better service to its users, then other companies’ users mostly shifted to that company due to convenience of service over their previous company. So to avoid this company have to analyse its customer’s behaviour with time to time that what user really want. Telecommunication companies, mostly do this by tracking log calls and other data and trying to analyse something to avoid churn. HBase helps to store large amount of data in a very efficient way because it is a non-relational column oriented database and it is built at the top of Hadoop.

3.2 Advertisement Targeting

Money through Advertisements becomes best business nowadays for websites. Social networking websites target user by tracking their activities and by using his profile information. They collect user data and try to predict the behaviour of users and show advertisements to the user according to his taste. Hadoop makes very important role in this type of analysis. Facebook is the biggest social networking site which uses Hadoop for the big data analysis and target users for their business. Facebook also developed hive which is a data warehouse built on the top of Hadoop for doing analysis.

3.3 Recommendation to Customer

Nowadays e-commerce websites become so popular. People are buying stuff online instead of going to market physically. All companies want to make more profit so they recommend their customer after buying some stuff into their profile by using their previous activities of shopping. Hadoop is very useful in this type of scenario. Many E-commerce websites are using Hadoop for this type of analysis, such as Amazon because data presents in this scenario are very huge in amount because of a lot of users.

3.4 Retail Sector

The retail sector has a very important role in our society. People are going there for purchasing their basic needs. Now to provide better service and to predict which item is in more demand, the retailer uses customer’s bill data and tries to predict some useful information for business intelligence. HDFS (Hadoop distributed file system) makes very important role for storing this kind of data and also helpful for processing and analysing this data to extract some useful information.

3.5 Search Quality

There are so many unsuccessful search attempts on the web and this is due to lower quality of search service by various providers. To avoid this kind of problem companies can use the user’s queries and can predict what the user actually wants. This will help in improving the search quality. This big data can be handled via Hadoop in a very efficient manner.

4. HADOOP ARCHITECTURE

Apache Hadoop [9] is an open source framework used to store and analyse big data which is present in Hadoop cluster. Hadoop always runs on a cluster means on homogeneous environment. Moreover homogeneous environment means all the systems which are present in cluster, their all components must be same in terms of RAM, CPU etc. Primarily Hadoop has two major components:

HDFS (Hadoop distributed File system)

Map Reduce

4.1 Hadoop Distributed File System

HDFS is a file system based on the master slave architecture. It breaks the large files into default 64 MB blocks and stored them into large cluster in a very efficient manner. In this architecture there are two basic nodes in a Hadoop cluster, data node and name node. There is also secondary name node present on Hadoop cluster.

Name Node is the master node which controls all the data nodes and it contains Meta data. It manages all the file operations like read, write etc.

Data nodes are the slave nodes present in Hadoop cluster. All the file operations performed on these nodes and data is actually stored on these nodes as decided by name nodes.

Secondary name node is the back up of name node. Name node is the master node so it becomes very important to make its backup. If failure would happen secondary name node will be used as name node.

4.2 Map Reduce

Map reduce is a core technology developed by Google and Hadoop implements it in an open source environment. It is a very important component of Hadoop and very helpful in dealing with big data. The basic meaning of map reduce is dividing the large task into smaller chunks and then deal with them accordingly. Map Reduce has four core components: input, mapper, reducer, and output.

4.2.1 Input

Input means that data which gets for processing and it is divided into further smaller chunks which are further allocated to the mappers.

4.2.2 Mapper

Mappers are the individuals that are assigned with the smallest unit of work for some processing.

4.2.3 Reducer

Mappers output become input for the reducers to aggregate the data in form of final output.

4.2.4 Output

Reducers' jobs are finally collected in the form of aggregated output.

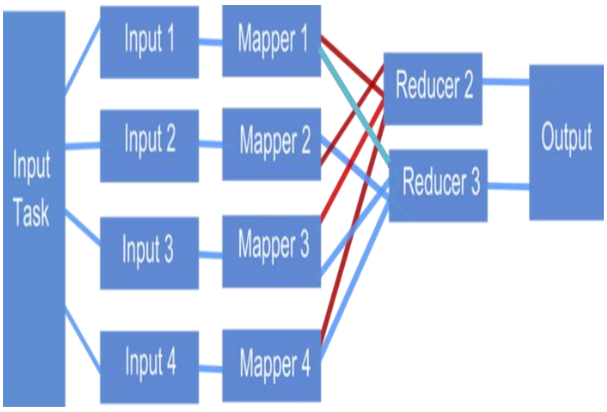

Fig 1: Map Reduce Architecture

5. TECHNOLOGIES BASED ON HADOOP

There are many technologies which are built on the top of the Hadoop [9] by Apache which means that Hadoop is not a single project as it includes other projects also. These technologies or projects have been designed for increasing the efficiency and functionality of Hadoop. These all technologies specially designed for dealing with big data and these all are along with HDFS and Map reduce. Hadoop along with these set of technologies also known as Hadoop Eco system. Primarily Hadoop Eco system consists of following technologies:

Apache PIG

Apache HBase

Apache Hive

Apache Sqoop

Apache Flume

Apache Zookeeper

Table 1. Hadoop Technologies

Hadoop Technologies	**Description**
Apache PIG	PIG is a scripting language which is used for writing programs for processing of large data set present in the Hadoop cluster. This language is known as PIG Latin. PIG runs those programs and convert them into map reduce jobs and then execute those jobs. It was created by Yahoo and now it is under Apache software foundation.
Apache HBase	HBase is non-relational or column oriented database which runs on the top of HDFS (Hadoop distributed file system).It also comes under Apache Software Foundation. It is open source and written in Java. Apache HBase allows reading and writing data on HDFS (Hadoop distributed file system) on the real-time scenario. HBase can deal with petabytes of data.
Apache Hive	Hive is SQL like language called HiveQL. It was developed by Facebook, but now it is used by many companies for data analysis. Moreover, it is a data warehouse infrastructure which provides all these functionalities. Hive allows querying of data from HDFS (Hadoop distributed file system) and these queries are converted into map reduce jobs.
Apache Sqoop	Sqoop is an application which helps in moving data in and out from any Relational database management system to Hadoop. So it is data management application built on the top of Hadoop by Apache Software Foundation.
Apache Flume	Flume is an application which is built on the top of Hadoop. It allows moving streaming data i.e. web log files into Hadoop cluster or HBase. Basic components of Apache Flume are source, channel, sink, agent, interceptor etc.
Apache Zookeeper	Zookeeper is open source project by Apache which provides centralized infrastructure that helps in synchronization across the Hadoop cluster. Naming services and configuration management are also provided by ZooKeeper across the Hadoop cluster.

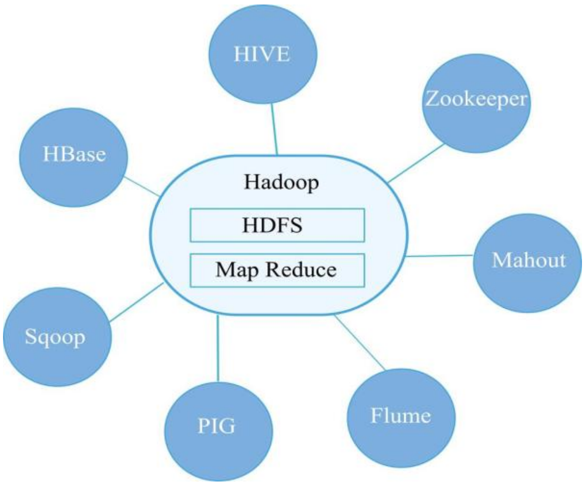

Fig 2: Hadoop Eco System

6. NO SQL (NOT ONLY SQL) TECHNOLOGY

No SQL is another technology which is widely used to handle the big data mostly unstructured data. It basically provides flexibility and scalability. There is no schema used in No SQL databases which is very helpful for dealing with huge amount of unstructured data.

Main databases under No SQL technology:

1) Key value pair store

2) Document oriented databases

3) Graph databases

4) Extended Relational databases

6.1 Key Value Pair Store

Key value pair is used to store the huge unstructured amount of data. There is no scheme used in the key value pair store like in RDBMS which provide great flexibility. There is one unique key and a particular value corresponding to that. The key is an entity and value is attributed in key value. Key value pair is also used in map reduce by map function. After loading of data map function fetch data in key/value pair in Hadoop environment. Most data is stored in string in key value pair store.

Example:

Key	Value
Facebookuser7778_name	Jas
Twitteruser4566_name	Rohit

Riak is the key value pair database which implements the No SQL technology and it is widely used among the social networking websites to handle its data.

6.2 Document Oriented Database

Document oriented is one of the famous No SQL database. It uses the document to handle the data. Mongo dB is the open source famous No SQL database. Mongo dB is used JSON (Java script object notation) and BJSON (Binary JSON). JSON is used for writing the queries. It is used to handle a document. In Mongo dB data are stored in a document with no schema and also without the concept of normalization. In mongo dB tables are automatically created after inserting data into documents via JSON. Tables are known as collection in mongo db.

6.3 Graph Database

Graph database is used to store the data in form of nodes and edges. It is a very easy and better approach than RDBMS due to its processing convince because graph database is very fast in terms of performance as compared to RDBMS. It is very helpful in performing graph like queries.

6.4 Extended Relational Database

Nowadays unstructured data is growing very fast. No SQL database, HDFS and other databases come into the picture to handle this data. Now to compete with these databases, companies like Microsoft and Oracle also extending their databases to remain in the market. Microsoft's file table is a good example of it. They are basically trying to make their software compatible with unstructured data.

7. WHY TO USE HADOOP

7.1 Load Balancing

Hadoop is basically used on the cluster, which is nothing but a group of homogeneous machines. There are various Linux machines in that cluster. Each machine is referred as one node. So the data are distributed among various nodes present in that cluster. It is very helpful to distribute load among the cluster node which also increases the processing power.

7.2 Flexibility

Flexibility is the greatest attribute for Hadoop. In Hadoop cluster you can easily add or remove a node according to our requirements. If sometimes load or processing of data increases then you can easily add some more nodes to overcome this problem and you can also delete any node if not required at any small level scenario.

7.3 Cost Effective Solution

Using a supercomputer or one machine with very high processing power for big data analysis is very costly. Alternate to this is making the homogenous cluster with various group of machines is more beneficial. It decreases the overall cost and moreover, in failures you can easily replace one node with another at very less cost. Hadoop is also open source so no need to pay for it for the commercial uses.

8. CONCLUSION

The world of big data has just started. Time has been started when the world is capable of generating data in terabytes and petabytes every day. In 2012 2.5 Exabyte data was generated on each day. Social networking business such as Facebook, Twitter, Tumblr, Google+ have completely changed the view of data.This data has a great impact on the business. It is very helpful to make the business intelligent if used properly by extracting information from it.

So there are lots of challenges and issues in big data technologies today. Making a sense out of a lot of data is a big challenge. Today lot of platforms and challenging situations are there ahead of this world to make new technologies for both processing and storage. This paper has covered the latest technologies which are made to deal with big data. Extensions of traditional database technologies are also discussed to deal with big data. In this paper platform from which big data can be generated are also discussed. So more technologies can be made to deal with big data processing and storage to make it more concise and meaningful.

9. ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my mentor Prof. Varun Singla for continuous provision of my M.Tech study and research, for his patience, motivation, enthusiasm, and immense knowledge. His supervision helped me in writing of review paper. Last I would thank my family for supporting me throughout my life.

10. REFERENCES

[1] Avita Katal, Mohammad Wazid, R H Goudar "Big Data: Issues, Challenges, Tools and Good Practices". In IEEE, Contemporary Computing (IC3), Sixth International Conference, pages 404-409, Noida, 2013.

[2] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money. "Big Data: Issues and Challenges Moving Forward". In IEEE, 46th Hawaii International Conference on System Sciences, pages 995-1004, 2013.

[3] Cloudera White paper,"Ten Common Hadoop able Problems", 2011.

[4] Kala Karun. A, Chitharanjan. K, "A Review on Hadoop – HDFS Infrastructure Extensions". In IEEE, Information & Communication Technologies (ICT), pages 132-137, 2013.

[5] Min Chen, Shiwen Mao, Yunhao Liu, "Big Data: A Survey". In Springer US, Mobile Networks and Applications, Volume 19, Issue 2, pp 171-209, 2014.

[6] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics". In IEEE, International Conference on Communication, Information & Computing Technology (ICCICT) pages 1-4, 2012.

[7] Kapil Bakshi, "Considerations for Big Data: Architecture and Approach". In IEEE, Aerospace Conference, pages 1-7 2012.

[8] Demchenko,Y, de Laat, C., Membrey, P.," Defining architecture components of the Big Data Ecosystem".In Collaboration Technologies and Systems (CTS),pages 104-112,2014.

[9] Apache Hadoop Project, http://hadoop.apache.org/.