Big Data Analytics: Challenges, Tools and Limitations

K. Siddardha, Ch. Suresh

***Abstract*— Data has become an essential part of every economy industry, organization and every individual. The amount of data generating daily on internet is not only increasing rapidly, but also complex, this lead to the contribution to big data. The big data is a term for massive data sets which have complex structure and cannot be handled by standard software. Big data is challenging in terms of effective storage, effective computation and analysis. In this paper, we discuss about the big data challenges, key tools and the limitations of big data analytics.**

***Index Terms*— Big Data, Hadoop, MapReduce, Security**

I. INTRODUCTION

Just like, Internet Big data is also part of our lives today. From search, online shopping, video on demand, to editing, Big data always plays an important role behind the scenes. Some people claim that the Internet of Things (IOT) will take over big data as the most hyped technology. But IOT cannot come alive without big data. In this paper we dive into the big data challenges, technologies and limitations. But we need to understand big data first.

Compared with traditional database, big data had greater volume; it is more varied; it drives from a greater range of sources. The key to big data is deriving valuable information form a mass of data. The features of big data can be summarized as the **"THE FIVE V's"**: volume, variety, velocity, value and veracity [1].

1.1. Volume: Volume indicates more data; it is granular in nature and unique. Big data requires processing high volumes of low density unstructured Hadoop data- that is, data of unknown value such as Twitter data feeds, click streams on web page, mobile app and many more. It is the task of big data to convert such Hadoop data into valuable information. For some organizations it might be tens of terabytes for other it might be hundreds of petabytes [2].

1.2. Velocity: The fast at which data is received and acted upon. The highest velocity data normally streams directly into memory. Some IOT applications have health and safety ramifications that require real time evaluation and action.

Operationally mobile phone users have large user populations, increased network traffic and expectations for immediate response [2].

K. Siddardha, Student, Department of CSE, GITAM University. Visakhapatnam. India

Ch. Suresh, Assistant Professor, Department of CSE, VidyaJoythi Institute of technology. Hyderabad. India.

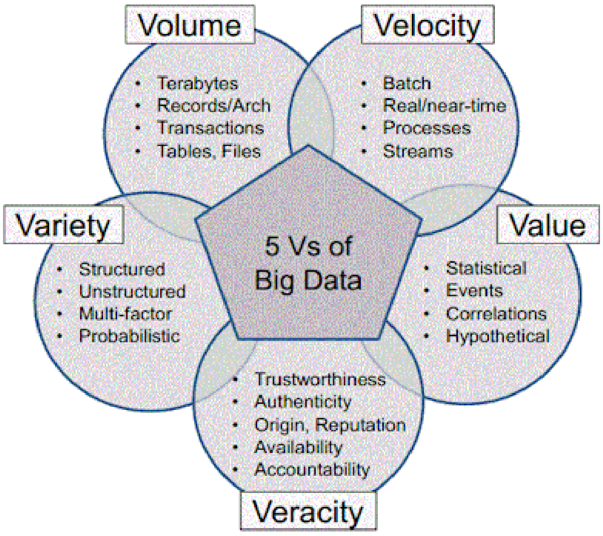

Fig: 1. V's of Big Data

1.3. Variety: Unstructured and semi structured data types such as audio and video. These types of data require different types of analysis or different type tools to use in order to process it. Unstructured data has many of the same requirements as structured data such as summarization, lineage, auditability and privacy. Complexity arises when data source changes without notice [2].

1.4. Value: There are a range of quantitative and investigative techniques to derive value from data. Big data involves in extracting valuable information from mass data to predict future trends and make decisions [2].

1.5. Veracity: It refers to messiness and trustworthiness of data. With many forms of Big data, quality and accuracy are less controllable (Twitter hash tags, abbreviations etc.) Big data analytic technologies will help to work with this type of data. Volumes often make up with lack of quality and accuracy [2].

II. CHALLENGES OF BIG DATA

The data originated from single source has less volume. Storing, managing and analyzing that kind of data does not present greater challenges and most processing is done through database and data warehouse. But in Big data environment the data is of large volume and cannot be processed through database because of unstructured data. We discuss the problems of big data-analytics [4].

2.1. Huge data sources and poor data quality: Big data is characterized by heterogeneous data sources like images, videos and audios. A traditional method for describing structure of data is not suitable for big data. Big data is also affected by noise and data may be inconsistent. So, efficient storing and processing are prerequisites for Big data.

2.2.Efficient Storage of Big data: The way Big data stored effects not only cost but also analysis and processing. To meet service and analysis requirements in Big data realible, high performance, high avalibility and low cost storage need to be developed. As data come from different sources it may causes redundancy. Detecting and eliminating redundancy may imporove storage area.

2.3.Efficiently processing Unstructured and Semi-Structured data: Databases and warehouses are unsatisifactory for processing of unstructured and semi structured data. With Big data read/write operations are highly concurrent for large number of users. As the size of database increases, algorithm may become insufficient. The CAP theorem states that it is impossible for a disturbuted system to have all the three, we can choose only 2 out of 3. Fig 2 shows the CAP theorem . Because consistency is required

2.4.Mass Data mining: Reaserch has shown that larger the data sets, machine algorithm is more accurate. Most traditional data mining algorithmsinvalid by Big data sets. When the size of data reaches petabytes, serial algorithm may fail to compute within timeframe. Effective machine learning and data mining algorithms need to be developed.

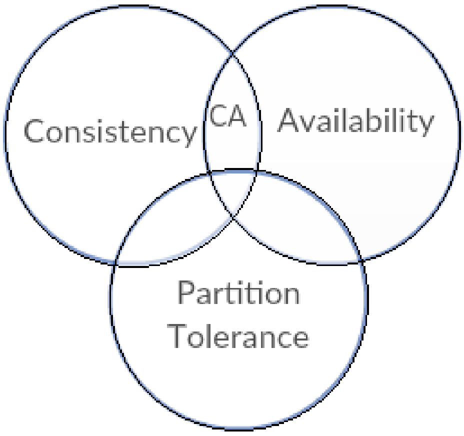

Fig: 2. CAP Theorem

III. BIG DATA TOOLS

3.1. Apache Hadoop: It is open source framework for storing and distributed processing of large data sets. Hadoop enables distributed parallel processing of large data sets. Hadoop is cost effective solution for storing of large volumes of data and it does not require any format [5]-[6].

3.1.1. Hadoop Architecture: Hadoop architecture [7] is master slave architecture shown in figure 3. Master being the namenode and slaves are the datanodes. The namenode controls the access to the data by clients. The datanodes manages the storage of data on the nodes that are running. Hadoop splits into one or more blocks and these blocks are stored in datanodes. Each data block is replicated to different datanodes to provide the high availability of Hadoop system. The JobTracker is responsible for scheduling the client jobs. JobTracker creates a map and reduces tasks and schedules them to run on datanodes (TaskTrackers). TaskTrackers runs on datanodes. The job of TaskTracker is to run the map and reduce the task assigned by namenode and report status of task to namenode.

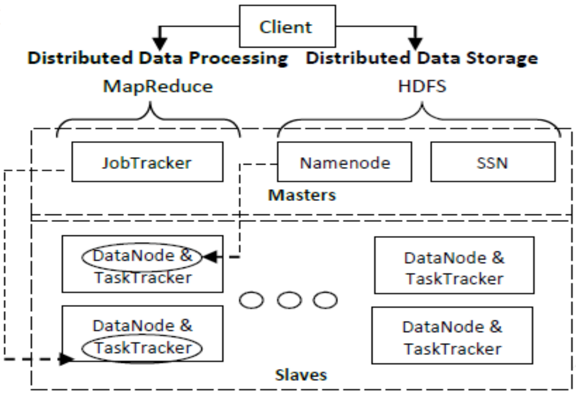

Fig: 3.Hardware implementation of Master/Slave Apache Framework consists of

1. Hadoop Common: It Contains libraries needed by other Hadoop modules.

2. Hadoop Distributed File System: It is a java based distributed file system used to store large volumes of data.

3. Hadoop YARN: It provides resource management and central platform to deliver consistent operations, security and governance tools across Hadoop clusters.

4. Hadoop MapReduce: It is a programming model for processing large scale of data.

3.1.2. Why Hadoop Big data?

One of the biggest challenges organizations have had is handling the unstructured data due to lack of technology in the past, but now Hadoop big data had changed that way and brings value to the unstructured data which is used for decision making process. Hadoop big data is famous because of [9]-[12]

1. Hadoop brings flexibility in data processing.

2. Hadoop is easily scalable.

3. Hadoop is faster in data processing.

4. It is very cost effective.

When it comes to processing of large data sets, Hadoop's map reduce programming allows processing these data sets in a completely safe and cost-effective manner.

3.1.3. Hadoop and Big data: With 90% of data being unstructured and growing rapidly, Hadoop is required to put the right big data workloads in the right system and optimize data management structure in an organization. The cost effectiveness is the major factor that makes it more necessary for organization to store and process big data.

3.2. Apache Spark: Apache spark [10] is a lightning-fast computer technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more type of computations. The main feature of Spark is it is 100x faster than the Hadoop MapReduce in memory.

3.2.1. Features of Apache Spark [10]:

1. Speed: Spark enables applications in Hadoop clusters to run upto100x faster in memory [11], and 10x faster even when running on disk. Spark makes it possible by reducing the number of read/write disc. It stores the intermediate processing data in-memory. It uses a concept of Resilient Distributed Dataset (RDD), which allows it to transparently store the data on memory and persist it to disc only it's needed. This helps to reduce the most of disc read and write-the main time consuming factors –of data processing.

2. Ease of use: Spark helps developer to create and run their applications on their familiar programming languages like java, Python etc. It comes with built-in set of over 80 high-level operators. We can use it to query data within the shell too.

3. Generality: In addition to simple "map" and "reduce" operations, spark supports SQL queries, streaming data and complex analytics such as machine learning and graph algorithms out-of-box. Not only can that user also combine all these capabilities seamlessly in a single workflow.

4. Runs Everywhere: Apache Spark runs on Hadoop, standalone or in cloud. It also can access from a variety of sources including HDFS, Cassandra, HBase and S3 etc.

3.2.2. Spark's major use cases over Hadoop:

1. Iterative algorithms in Machine Learning.
2. Interactive Data mining and Data processing.
3. Spark is fully Apache Hive-compatible data warehousing system that can run 100x faster than Hive [11].
4. Stream processing: Log processing and fraud detection in live streams for alerts, aggregates and analysis.
5. Sensor data processing: Where data is fetched and joined from multiple sources, in memory data set are very helpful as they are easy and fast to process.

3.3. Apache Hive: Hive is a tool which is built on top of Hadoop. This is used to process structured data that is present in Hadoop. It is an open source data warehouse system which is used for querying large sets of data stored in Hadoop.
Hive was developed by Facebook
Hive is not a relational database nor a online transactional process.

3.3.1. Hive Key Features:

1. Familiar SQL interface: Use the existing SQL skills to run batch queries on data stored in Hadoop. Queries are written using a SQL-like language, HiveQL, and are executed through either MapReduce or Apache Spark, making it simple for user to process and analyze unlimited amounts of data.

2. Shared Data Structures: Using HCatalog, a table and storage management layer for Hadoop, Hive Meta data is exposed to other data processing tools, including Pig and MapReduce. As well as through a REST API. This allows users to easily read and write data without worrying about where data is stored, what format it is.

3. Faster Batch Processing: Hive-on-spark features the next generation of batch processing for Hive. With queries executed through Apache Spark, a powerful data processing tool, users will see dramatic performance improvements compared to MapReduce.

3.3.2. Work flow Between Hive and Hadoop:

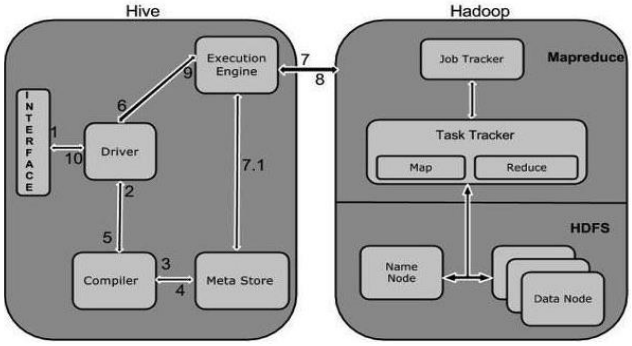

Fig: 4.Work flow between Hive and Hadoop

1. Hive interface sends query to any of the driver (i.e., database driver) to execute the query.
2. The driver takes the help of query compiler that parses the query to check the syntax and query plan or the requirement of query.
3. The compiler sends metadata request to Metastore (any database).
4. Metastore sends metadata as a response to the compiler.
5. Parsing and compiling of a query is completed based on the plan sent to driver by the compiler.
6. The driver sends the execute plan to the execution engine.
7. The process of execution job is a MapReduce job. The execution engine sends the job to JobTracker, which is in Name node and it assigns this job to TaskTracker, which is in Data node. Here, the query executes MapReduce job.
8. During the execution, the execution engine can execute metadata operations with Metastore.
9. The execution engine receives the results from Data nodes.
10. The execution engine sends those resultant values to the driver.
11. The driver sends the results to Hive Interfaces.

3.4. NoSQL Databases: NoSQL databases have grown in popularity. These Not Only SQL databases are not bound by traditional schema. The Flexibility of NoSQL databases like MangoDB, Cassandra and HBasemake them a popular option for big data analytics.

3.5. Benefits of Big Data tools:

Big data tools helps in storing and processing of unstructured data which helps for many organizations to make right decisions. The main reasons that we are relaying on big data tools are [6].

1. Big data tools like Hadoop; spark can reduce the cost of storing the data.
2. These tools help for big organizations to take faster and better decisions.
3. Big data tools also help to create new products and service for customers.

IV. .LIMITATIONS OF BIG DATA

Although big data tools had benefits they have limitations too. The main Big data limitation are discussed in the below [13].

1.1 Prioritizing correlations:

"**Correlation is not simply causation.**"

Big data is used to tease out correlations. Correlation exists when one variable is linked with another, just because two variables linked doesn't mean that causative relationship exists between them.

1.2 Security: Security is foremost aspect for every technology. Big data is prone to data breaches. The important information that is provided to some third party may get leaked to customers. Proper encryptions must be made in order to protect the data.

1.3 Large growth in data: data is growing faster than the processing power. Large volumes of data are being exploded in past years. We need some new machines to work; otherwise we will get over run by data. Large Data centers can solve this problem.

1.4 Inconsistency in data collection: Sometimes the tools we use to gather big data sets are imprecise. This will happen when the data is collecting for example, consider Google search the results of the search on one day will be different from other day, this is mainly due to inconsistency in data collection [13].

V. CONCLUSION

Big data has become a new era for every economy, industry and organizations. Through the scrutiny of huge volume of data that are becoming available, there is the possibility for making faster advances in various technical areas. Big data analysis is becoming essential for automatic discovering of intelligence for the occurrence of patterns and hidden rules. Big data analysis makes it easier for companies in decision making, predicting and identifying the new opportunities. In this paper we have discussed about the issues and challenges of big data and also big data analysis tools which help researchers for extracting useful knowledge out of big data.

REFERENCES:

[1.] http://www.excelacom.com/resources/blog/the-5-vs-of-big-data-predictions-for-2016

[2.]https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know.

[3.] Complete Introspection on Big Data and Apache Spark http://www.ijsdr.org/papers/IJSDR1604006.pdf

[4.] Big data challenges and Opportunities http://blogs.wsj.com/experts/2014/03/26/six-challenges-of-big-data/

[5.] Open source Tools for Big data http://www.happiestminds.com/blogs/top-10-open-source-big-data-tools/

[6.] Big data Analytics Toolshttps://www.qubole.com/big-data-analytics/

[7.] Hadoop Architecture https://hadooptutorial.wikispaces.com/Hadoop+architecture

[8.] Tools that ditch the big datahttp://www.infoworld.com/article/3128344/analytics/7-big-data-tools-to-ditch-in-2017.html

[9.] Nobody tells you- 5 things big data 'CAN' and 'CANNOT' DOhttps://www.analyticsvidhya.com/blog/2015/11/5-big-data-can-cannot/

[10.] Apache Spark Features http://spark.apache.org/

[11.] Five Things you need to know about Hadoop vs. Apache Spark http://www.infoworld.com/article/3014440/big-data/five-things-you-need-to-know-about-hadoop-v-apache-spark.html

[12.] The documentation of big data by Oracle https://docs.oracle.com/cd/E63064_01/

[13.] The Limitations of big data. https://www.elevatedthird.com/article/big-data-limitations.

K. Siddardha, Department of CSE, GITAM University. His Areas of interests includes Big data analytics, Web Technologies.

CH. Suresh, Assistant professor, Department of CSE at Vidya Jyothi Institute of Technology, Hyderabad. His Research areas include Data warehousing and Mining, Distributed Computing etc