Big Data Analysis using Hadoop

Nikhitha Cyril
PG student,Dept of Information Technology
Rajagiri School of Engineering and Technology,
Kochi, India

Arun Soman
Asst. Professor, Dept of Information Technology
Rajagiri School of Engineering and Technology,
Kochi, India

ABSTRACT
In the present world, where more and more users upload data to the internet, the overall size of data that need to be stored and analyzed exceeds the capacity of traditional storage and analysis techniques. Hence, it is necessary to introduce new and efficient methods for the analysis of Big Data in order to extract useful information from them. Big Data has tremendous importance in almost all areas such as education, healthcare, etc. Big data is defined by its three main characteristics which are high volume, high velocity and huge variety. Hadoop is provides an efficient platform for the analysis of Big Data. It consists of a distributed storage system, HDFS for the storage of large volumes of data and a programming framework, MapReduce for the analysis of data.

General Terms
Big data, Hadoop.

Keywords
HDFS, MapReduce

1. INTRODUCTION
Most people today have accounts in different social networking sites such as Facebook, Twitter, Instagram, LinkedIn, etc. In Facebook alone there are 1.393 billion monthly active users among which 890 million are daily active users. On an average, 350 million photos are uploaded and 4.75 billion items are shared in Facebook on a daily basis. As per the statistics of November 2014, Facebook stores 300 petabytes of data and takes in an average amount of 600 terabytes of data each day [14]. This data never gets deleted. Instead, it increases in such a way that the rate of increase in the data itself gets increased. Such large amounts of data are termed as Big Data.

"Big Data is the territory where our existing traditional relational database and file systems processing capacities are exceeded in high transactional volumes, velocity responsiveness, and quantity and or variety of data. The data are too big, move too fast, or don't fit the structures of RDBMS architectures" [2].Big Data usually consists of three types of data. They are the traditional enterprise data, machine generated/sensor data and the social data. The enterprise data includes information from Customer Relationship Management (CRM) Systems, Enterprise Resource Planning (ERP), web store transactions, etc. Examples for machine generated data are Call Detail Records (CDR), weblogs and smart meters. Customer feedback streams, social media platforms, blogs, etc. make up the social data [5].Big Data has great importance in today's world from healthcare to large-scale analytics. Aggregation of all related healthcare information from different sources helps tremendously in the treatment of a patient. The doctor can easily obtain information relating to a disease from different parts of the world. Also integration of data from different areas such as clinical data, cost involved, claims available, administrative data, pharmaceutical data, research and development data, patient behavior and sentiment data, etc. helps in efficiently treating a patient[7].The Champions of the FIFA World Cup 2014, Germany had Big Data to provide a competitive edge to the team. A German software company named SAP built a tool named Match Insights that was capable of analyzing video data from on-field cameras. The tool produced thousands of data points including the player's position and speed. These data was sent to the SAP database where it was analyzed. Based on the result the coach was able to provide feedback to the player's cellphone. Not only football but also other sports like basketball and tennis have also started using Big Data for the same purpose [11]. Big Data enables revenue agencies to handle mass data and instant requirements. It helps in choosing the correct collection strategy based on the analysis of tax payers risk and to detect fraud in real time. It also helps tax auditors to search through mass data in real time [15]. In case of business, Big Data can be used in analytics to draw important conclusions and insights on business plans as well as in the development of applications and real time services that benefit the customers [8].Similarly, Big Data can be used in a wide range of areas such as intelligent transportation, Financial Market Trading and surveillance, crowd control, military decision making, early warning for natural disasters, etc.[3].

2. CHARACTERISTICS
The characteristics of Big Data are specified by three 'V's which stands for Volume, Velocity and Variety. However, in addition to these, Big Data has other characteristics such as value, veracity, variability and virality but they are not used to specifically define Big Data as they apply to normal data as well.

2.1 Volume
The word Big Data itself specifies the volume of Big Data. When the terabytes of data generated in the social media each day gets added to the existing petabytes of data, soon the data will be in the range of zettabytes. As the size of data increases, so does the difficulty in analyzing those data using traditional methods. This brings about the need for introducing better and efficient methods to analyze Big Data [1].

2.2 Velocity
Velocity specifies the data that is in motion. For example, velocity specifies the rate at which the data from sensors or web logs are collected, i.e. the speed of incoming data as well as the speed at which the data is captured, stored and analyzed, i.e. the speed at which the data flows. Another dimension of velocity is by the lifetime of data utility. It specifies how long the will be valuable, i.e. whether the data is permanently valuable or loses its importance rapidly. In addition to this, velocity has a third dimension which specifies the speed at which the data should be stored and retrieved [10].

2.3 Variety
Variety implies that data that forms Big Data may vary in terms of type as well as the source from which the data is

obtained. Big Data may not contain only structured data as in traditional database systems. Instead they also contain semi structured and unstructured data such as emails, documents, text messages, videos, images, audio, graphs, output from sensors, etc. The challenges involved in using such varieties of data are in storing and retrieving the data quickly and efficiently as well as in extracting data related to a single event so as to analyze them together [10].

2.4 Value

Big Data provides a much more personalized results compared to traditional database which provides group data. Big Data enables us to rank the data based on its value. Big Data stores data before understanding the data. Also, Big Data is found to have low value density, i.e. a large amount of data provides comparatively small information [10].

2.5 Veracity

Veracity describes the provenance of data. It shows whether the data is from a reliable source. Accordingly, we can say whether the data is accurate and complete retrieved [10].

2.6 Variability

Variability shows whether the data is consistent in terms of availability and interval of reporting. Thus, we could know if the data is important or just noisy data retrieved [10].

2.7 Virality

The rate at which data spreads is defined as virality. It shows how often data is picked up and repeated by other events retrieved [10].

3. ANALYZING BIG DATA

Even though the importance of Big Data is real and significant, there are some technical challenges that should be analyzed to completely realize its full potential. These challenges mainly involve dealing with the large volume, high velocity and different varieties of data. In addition to these, Big Data also has to meet the privacy and usability requirements of data.The analysis of Big Data involves distinct phases such as data acquisition and recording, information extraction, data aggregation, analysis and interpretation. These are described below [6].

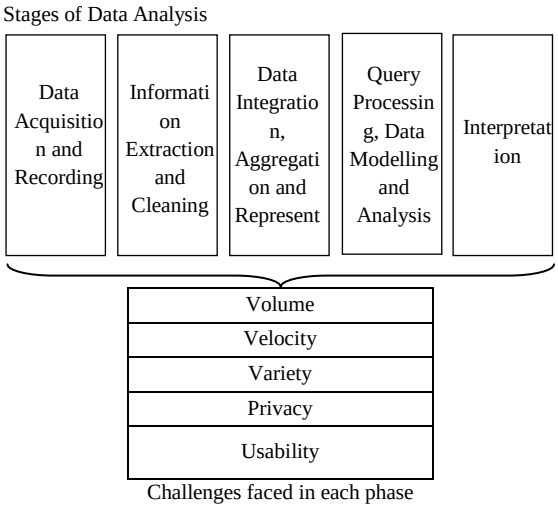

Fig 1: Stages of Big Data Analysis [6].

3.1 Data Acquisition and Recording

As mentioned before Big Data has low value density. Hence most of the data are noise data that can be filtered and compressed by large scale. These filters must be defined such that useful information is preserved. Intelligent methods are to be introduced in data reduction so that the haystack is removed without losing the needle in it. Also, real-time analysis to process streaming data on the fly should be incorporated [6].Another major challenge is to generate the right metadata that describes what data is recorded and how the data is recorded. Efficient data systems should also be utilized to carry the data provenance along with the metadata through the data analysis pipeline. This is because the error in data analysis at one step may render all the subsequent analysis useless and data provenance helps in identifying the subsequent processing that depend on this step [6].

3.2 Information Extraction and Cleaning

The data acquired usually will be in an unstructured form that is not suitable for the data analysis. Hence, we have to perform an information extraction process that extracts the information and converts it into a structured format that supports data analysis. Big Data does not always provide valid information. Sometimes the data may contain wrong information that may be caused deliberately or accidently. In either case, the data should be cleaned by well recognized constraints or well understood error models [6].

3.3 Data Integration, Aggregation and Representation

Data obtained from different sources will be heterogeneous in nature. These data cannot be simply stored into a repository. Instead, the differences in data structure and semantics should be expressed in computer understandable form, so that they can be robotically resolvable. Another challenge is in designing the database format that decides how the data should be stored. Professionals such as domain scientists should be entrusted to create effective database designs either by devising tools to assist them in the design process or by developing techniques by which databases can be used effectively in the absence of efficient database design [6].

3.4 Query Processing, Data Modeling and Analysis

Methods for querying Big Data are different from that of traditional methods. Big Data in itself is noisy, dynamic, heterogeneous, interrelated and untrustworthy. However, data mining requires clean, integrated, trustworthy and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms and big data computing environments. Still, mining itself can be used to improve the quality and trustworthiness of data, understand its semantics, and provide important querying functions. Scaling complex query processing to terabytes of data interactive response time is a major challenge in Big Data. The problem is lack of coordination between the databases that stores data and provides SQL querying from the analytical packages that perform non-SQL processing. Hence, to perform non-SQL processing, the data should be exported from the databases, and stored back after the processing [6].

3.5 Intepretation

The users must be able to understand the result of Big Data analysis. Hence supplementary information provenance of the data should be provided to the user. Provenance of the result

should explain and based on what inputs how the result was obtained. Capturing, storing and querying provenance along with the metadata creates an infrastructure that enables user to not only interpret the results but also to repeat the analysis with different assumptions, parameters and datasets [6].

4. HADOOP

Hadoop is an open source framework provided by Apache Software Foundation and was created by Doug Cutting to meet the large scale data storage and analysis. Hence, it is an important tool for storing and analyzing Big Data. For this, Hadoop mainly consists of a distributed storage unit named Hadoop Distributed File System and a software framework called MapReduce. In addition, to these two components Hadoop has many other components such as HBase, Pig, Hive, etc. HBase is a distributed column oriented database that uses HDFS for its underlying storage. It supports both batch style computations using MapReduce and point queries. Hive is a distributed data warehouse used for managing HDFS data using an SQL based query language. Pig runs on HDFS and MapReduce and provides a data flow language and environment to process large datasets.Hadoop version 0.20 and the 1.x versions use Kerberos authentication to authorize access to the Hadoop data. The Hadoop version 0.22 and the 2.x versions uses a new MapReduce named MapReduce 2 that runs on YARN (Yet Another Resource Negotiator) [4].

4.1 Hadoop Distributed File System (HDFS)

HDFS is the storage unit of Hadoop. It is a distributed file system that is used to store large size, unstructured, streaming data. HDFS has high fault tolerance and is designed to be deployed on low cost hardware. HDFS also provides high throughput access to application data with large size. In order to enable streaming access to file system data, HDFS relaxes some of the POSIX requirements.

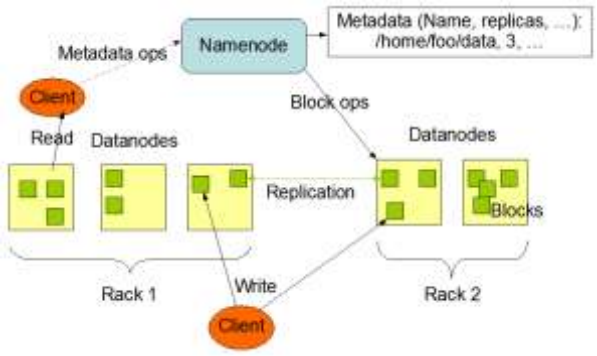

Fig 2: HDFS Architecture [12].

HDFS has master/slave architecture with two types of nodes. It consists of a namenode acting as the master and a number of datanodes acting as the slaves. The namenode manages the file system namespace and regulates access to the files. There are usually one datanode per node in the cluster. The datanode manages the storage of data in the node that it runs on. Once HDFS exposes a file system, users can store data to the files. Each file is split internally into one or more blocks and is stored to the datanodes. The default size of each block is 64MB. Namenode performs file system namespace operations such as opening, closing and renaming files and directories as well as mapping blocks to datanodes. The datanodes performs the read and write requests from the clients. Datanodes are also used to perform block creation, deletion and replication based on the instructions received from the namenode.HDFS is built using java. Hence namenode and datanode can run on any system supporting java [9,12].

4.2 MapReduce

Hadoop MapReduce is a programming paradigm used for writing applications to process large datasets on large clusters in parallel way. As the name suggests, MapReduce programs execute in two phases which are Map phase and Reduce phase. MapReduce consists of a Job Tracker and many Task Trackers to perform the tasks. The Job Tracker is the master that allots mapping task and reducing tasks to different Task Trackers. The Task Trackers perform the tasks that are allotted to it by the Job Tracker.The MapReduce job begins when the client program submits the job configuration to the Job Tracker. The job configuration specifies the map and reduce functions as well as the input and output paths for data retrieval and storage. From the input path, the Job Tracker determines the number of input splits and selects a bunch of Task Trackers that are closer to the data source to perform the task.During the map phase, the Task Trackers read the data from the input splits and convert them to <key, value> pairs and are stored to the memory buffer. A periodic wakeup process sorts the memory buffer and stores it to R local files based on the key value (R is the number of reducer nodes). Once the map task is done, the Task Tracker informs the Job Tracker that the task is completed.

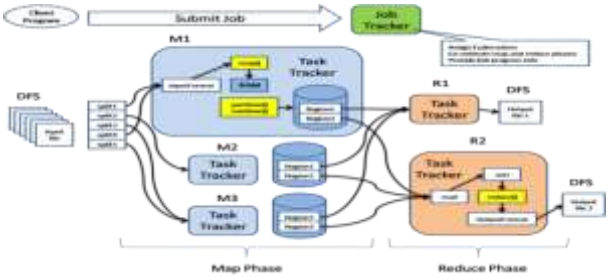

Fig 3: MapReduce Architecture[13].

After all the Task Trackers complete their job, the Job Tracker allocates reduce tasks for the selected Task Trackers. During the reduce phase, each Task Tracker reads the <key, value> pairs from a local file and performs the reduce function once for each key. It generates a <key, aggregated value> pair for each key and stores it to the output file.The Job Tracker tracks the progress of each phase. It also sends periodic heartbeat signals to each Task Tracker to obtain their health status. If any of the Task Tracker crashes, the Job Tracker will assign the task to a different Task Tracker [9,13].

5. RESULT

The experiment was performed on a hadoop cluster set up using the Amazon EC2 Servers. Ubuntu Server 14.04 LTS was used.The instance type used was t2.micro. Hadoop 1.2.1 was installed on each server and the multinode setup was formed. The master node was used extensively as the Namenode and JobTracker. The 3 worker nodes acted as the DataNodes and TaskTrakers. One of the worker node was also set as the Secondary NameNode. The Hadoop cluster was tested using word count and pi calculation problems and the results were obtained efficiently.

6. CONCLUSION

Big Data has proved its importance in different areas from sports, education and healthcare to large scale data analytics. If properly stores and analysed, Big Data can provide fruitful results in almost all these areas. The analysis of Big Data involves different phases which are data acquisition and recording, information extraction and cleaning, data integration, aggregation and recording, query processing, data modelling and analysis and interpretation. Apache Hadoop with its HDFS and MapReduce provides an efficient framework to analyse Big Data. As a future scope we can write different MapReduce programs for analysis of Big Data. We can also edit the source code of Hadoop to enhance the performance of Hadoop.

7. REFERENCES

[1] Avita et al, "Big Data: Issues, Challenges, Tools and Good Practices," in IEEE Sixth International Conference on Contemporary Computing, 2013, pp. 404-409.

[2] Raymond Gardiner Goss and KousikanVeeramuthu, "Heading Towards Big Data Building a Better Data Warehouse for more data, more speed and more users," in IEEE 24th Annual SEMI Advanced Semiconductor Manufacturing Conference, 2013, pp. 220-225.

[3] Nader Mohamed and Jameela Al-Jaroodi, "Real-Time Big Data Analytics: Applications and Challenges," in IEEE International Conference on High Performance Computing and Simulation, 2014, pp. 305-310.

[4] Tom White, "Meet Hadoop, " in Hadoop: The definitive guide, 3rd Edition, California: O'Reilly Media, 2012, ch. 1, pp. 9-15.

[5] Jean Pierre Dijcks, "Oracle: Big Data for the Enterprise," in Oracle White Paper, 2013© Oracle Corporation.

[6] Divyakant Agrawal et al., "Challenges and Opportunities with Big Data," Cyber Center Technical Reports, Purdue e-Pubs, Purdue University, 2011.

[7] Sonja Zillner et al, "Towards a Technology Roadmap Big Data Applications in the Healthcare Domain," in IEEE 15th International Conference on Information Reuse and Integration, California, 2014, pp. 291-296.

[8] F. CananPembeMuhtaroglu et al, "Business Model Canvas Perspective on Big Data Applications," in IEEE International Conference on Big Data, 2013, pp. 32-37.

[9] HimanshuRathod and Tarulata Chauhan, "A Survey on Big Data Analysis Techniques," in International Journal for Scientific Research and Development, 2013, vol, 1, issue. 9, pp. 1806-1808.

[10] Bill Vorhies. (2013, October 13). *How many "V"s in Big Data- The Characteristics that define Big data.* [Online]. Available: http://data-magnum.com/how-many-vs-in-big-data-the-characteristics-that-define-big-data/

[11] Steven Norton. (2014, July 10). *Germany's 12th man at The World Cup: Big Data.* [Online]. Available: http://blogs.wsj.com/cio/2014/07/10/germanys-12th-man-at-the-world-cup-big-data/

[12] DhrubaBorthakur. (2013, April 8). *HDFS Architecture Guide.* [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[13] Ricky Ho. (2008, December 16). *How Hadoop Map/Reduce works.* [Online]. Available: http://architects.dzone.com/articles/how-hadoop-mapreduce-works

[14] Craig Smith. (2015, March 19). *By the Numbers: 200+Amazing Facebook User Statistics (February 2015).* [Online].Available: http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats

[15] Hannes Venter, "Case Study Mauritius: Successful Implementation of Innovative Public Revenue Management Solutions," in IST- Africa Conference Proceedings, 2014.