JECET; March 2015-May 2015; Sec. B; Vol.4.No.2, 202-209. E-ISSN: 2278-179X

Journal of Environmental Science, Computer Science and Engineering & Technology

An International Peer Review E-3 Journal of Sciences and Technology

Available online at www.jecet.org

Section B: Computer Science

Research Article

Big Data: Problems, Challenges and Techniques

Piyush Gupta[1], Pardeep Kumar Mittal[2], Girdhar Gopal[3]

Research Scholar[1], Assistant Professor[2,3] Dept. of Comp. Sci. and Appl., KUK, Haryana[1,2,3]

Received: 1 April 2015**; Revised:** 16 April 2015**; Accepted:** 20 April 2015

Abstract: Today a large amount of data is generated every second by the users in the form of blog posts, status messages, photographs and audio/video files. To manage this big data, which can be structured, unstructured or semi-structured, new techniques, algorithms and analytics are required. Hadoop is a processing framework designed for structuring big data that can be tens or hundreds of terabytes and even petabytes in size. Hadoop has two main components - HDFS and Map Reduce. The data is loaded into the Hadoop Distributed File System (HDFS) and this massive data is operated by doing horizontal scaling across very large number of servers through an approach called Map Reduce.

Keywords: Big Data, Data Analytics, Hadoop, HDFS, Map Reduce

INTRODUCTION

For many years, companies have been building data warehouses to analyze business activity and produce insights for decision makers to act on to improve business performance. These traditional analytical systems are often based on a classic pattern where data from multiple operational systems is captured, cleaned, transformed and integrated before loading it into a data warehouse. many companies have built up multiple data warehouses and data marts in various parts of their business. Not only is the amount of data being generated increasing, but the rate of increase is also accelerating. From emails to Facebook

posts, from purchase histories to web links, there are large data sets growing everywhere. The challenge is in extracting from this data the most valuable aspects; sometimes this means particular data elements, and at other times, the focus is instead on identifying trends and relationships between pieces of data[1].

This paper is organized in the following sections. First the big data arises will be discussed in section 2. Then in Section 3 Big data Analytics will be discussed. The techniques to approach the big data analytics will be covered in section 4 and section5. In Section 6 conclusion is given about the problems, challenges and solution techniques of Data Analytics.

Big Data: Big data is data that is too large, complex and dynamic such that it exceeds an organization's storage or compute capacity for accurate and timely decision making. Big data is not just about data volumes, It may be the case that data volumes are moderate but data variety (data and analytic complexity) are significant. Big data can be associated with both structured and multi-structured data.

In the last twenty years, the data is increasing day by day across the world .some facts about the data are, there are 277,000 tweets every minute, 2 million searching queries on Google every minute, 72 hours of new videos are uploaded to YouTube, More than 100 million emails are sent, 350 GB of data is processing on facebook and more than 570 websites are created every minute.

Dimensions of Big data: As the data is too big from various sources in different form, it is characterized by the 3 dimensions. These dimensions of Big data generally called **3 V's of Big data**[1, 2, 3]**.**

- **Volume–** The rate at which data is gathered by the companies. It includes emails, audio/video files, weblog data, machine generated data, photographs and other web contents. It varies from tera bytes to peta bytes and up.
- **Variety–** New data types and sources are now being captured by the organizations. These include- Semi-structured data(e.g. email, e-forms, HTML, XML), Unstructured data(e.g. document collections, social interactions, images, video) and sound Sensor and machine generated data. These collection of new data types is reffered to as multi- structured data. Main problem with multi-structured data is that it is dynamic. So we need to do some investigative analysis to identify data.
- **Velocity–** It refers to the rate at which data is being created. Example is Financial markets data where data is being generated and emitted at very high rates and where there is a need to analyze it immediately to respond to market changes in a timely manner. Other examples include sensor and machine generated data where the same requirement applies, or cameras requiring video and image analyses.

Additional Vs in Big data that IT business and data scientists need to be concerned[4]:

- **Veracity-** It refers to the uncertainty of data. Is the data that is being stored, and mined meaningful to the problem being analyzed. veracity in data analysis is the biggest challenge when compares to things like volume and velocity.
- **Validity-** It refers to the accuracy and correctness of the data for the intended use.

Challenges in Big data: To take full advantages of data analytics lot of challenges are there related. Some of them are discussed below.

- **Meeting the need for speed:** In today's hyper competitive business environment, companies not only

have to find and analyze the relevant data they need, they must find it quickly. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed[5].

- **Understanding the data:** It takes a lot of understanding to get data in the right shape so that you can use visualization as part of data analysis. For example, if the data comes from social media content, you need to know who the user is in a general sense – such as a customer using a particular set of products – and understand what it is you're trying to visualize out of the data[5].
- **Addressing data quality:** Even if you can find and analyze data quickly and put it in the proper context for the audience that will be consuming the information, the value of data for decision-making purposes will be jeopardized if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the volumes of information involved in big data projects, it becomes even more pronounced[5].

Big data Analytics: Big data analytics is the process of examining large data sets containing a variety of data types to uncover hidden patterns, market trends, customer preferences and other business information. The primary goal of big data analytics is to analyze the mix of structured, semi-structured and unstructured data to help organizations make more informed business decisions. Relational databases can't handle the semi-structured and unstructured data, so organizations turned to a newer class of technologies that includes Hadoop and related tools such as YARN, Map Reduce and NoSQL databases[6, 16, 17]

Some examples of industry use cases for Big Data analytics:

EBay customer journey: With 50TB of machine-generated data produced daily and the need to process 100PB of data all together, eBay's data challenge is truly astronomical. The eBay site has 100 million customers who list items in 30,000 categories. In terms of transactions, the site processes thousands of dollars per second[7]. In 2002, eBay built a 13TB Teradata enterprise data warehouse, which effectively provides a massive parallel relational database. This has now grown to 14PB, with the system built on hundreds of thousands of nodes. The enterprise data warehouse gives tremendous performance on standard structured queries, but it is unable to meet eBay's needs for storage and processing flexibility. These systems are fairly expensive, so when you are looking at adding 50TB of data every day, costs are prohibitive. The company has split its data analytics across three platforms, the first of which is a traditional enterprise data warehouse from Teradata. This core transactional system must be extremely reliable. "The system can't go down. Every day we process 50TB of data, accessed by 7,000 analysts with 700 [concurrent users]." The company worked with Teradata to develop a custom appliance built with several hundred user-defined functions. The system was built on commodity hardware, with proprietary software to process all the customer journey data and store it cheaply. The end result is a custom data warehouse called Singularity. Along with the enterprise data warehouse and Singularity, eBay is also using Hadoop, which completes the third side of its data analytics triangle. The auction site has built two 20,000-node Hadoop clusters with 80PB of capacity. These work alongside the Teradata data warehouse and Singularity custom data analytics appliance to give eBay the tools it needs to use data analysis to follow the customer journey.

Health Care: storing and processing Medical Records in Cloudera Hadoop Distribution(CDH): A health IT company instituted a policy of saving seven years of historical claims and remit data, but its in-house database systems had trouble meeting the data retention requirement while processing millions of claims every day[8,9].

A Hadoop system allows archiving seven years' claims and remits data, which requires complex processing to get into a normalized format, logging terabytes of data generated from transactional systems daily, and storing them in CDH for analytical purposes.

Use Case: One of the big advantages of Hadoop has been to be able to segregate big data from transactional processing data and allow smoother processing of information[8]. Two areas that were a strong fit for Hadoop. Firstly, Archiving seven years' claims and remit data, which requires complex processing to get into a normalized format? Secondly, Logging terabytes of data generated from transactional systems daily, and storing them in CDH for analytical purposes.

Impact: Low Cost + Greater Analytic Flexibility: Because Hadoop uses industry standard hardware, the cost per terabyte of storage is, on average, 10x cheaper than a traditional relational data warehouse system. "One of my pet peeves is: you buy a machine, you buy SAN storage, and then you have to buy licensing for the storage in addition to the storage itself," explained the manager of software development. "You have to buy licensing for the blades, and it just becomes an untenable situation. With Hadoop you buy commodity hardware and you're good to go. In addition to the storage, you get a bigger bang for your buck because it gives you the ability to run analytics on the combined compute and storage. The solutions that we had in place previously really didn't allow for that. Even if the costs were equivalent, the benefit you get from storing data on a Hadoop type solution is far greater than what you'd get from storing it in a database."

Hadoop: A technique to process Big data: It all started with Google, which in 2003 and 2004 released two academic papers describing Google technology: the Google File System (GFS) [10] and Map Reduce[11] The two together provided a platform for processing data on a very large scale in a highly efficient manner.

Doug Cutting was working on the Nutch open source web search engine. He had been working on elements within the system that resonated strongly once the Google GFS and Map Reduce papers were published. Doug started work on the implementations of these Google systems, and Hadoop was soon born, firstly as a subproject of Lucene and soon was its own top-level project within the Apache open source foundation. Doug Cutting was hired by Yahoo in 2006 and Yahoo quickly became one of the most prominent supporters of the Hadoop project[15, 16].

Hadoop is a open source Programming framework used to store and process large data sets. Hadoop uses Map Reduce parallel programming model to handle big data. It is used by a number of leading internet service companies, including Facebook, Amazon, Yahoo, and others. Hadoop framework has two main components: HDFS and Map Reduce.

Hadoop Distributed File System (HDFS) is used to store the large data sets and Map Reduce is used to

process that data sets. Highly parallelized nature of the Map Reduce increase application throughput.

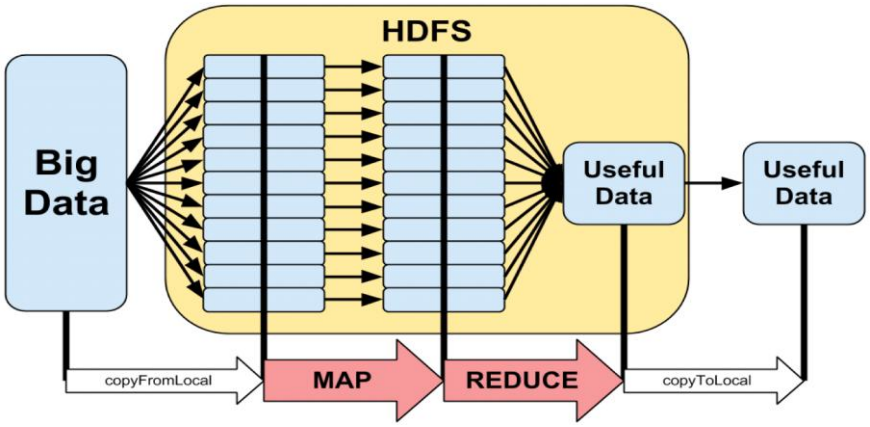

Figure 1: Hadoop components[12]

Hadoop Distributed File System (HDFS): Hadoop uses HDFS to store files efficiently in the cluster. When a file is placed in HDFS it is broken down into blocks, minimum of 64 MB block size which is far more in comparison to other file systems like FAT, NTFS where the block size is typically 4-32 KB.
Each block is assigned a number blk_xxxxxxx, where xxxxxxx is a long number depending on the size of data. Each of these blocks is then saved on some nodes in the cluster called Data Nodes. Meta Data of these files is stored on a specific node called Name Node in the cluster. To prevent the failure at data nodes, these blocks are replicated across the different nodes in the cluster. The default replication value is 3, i.e. there will be 3 copies of the same block in the cluster. A Hadoop cluster can comprise of a single node (single node cluster) or thousands or millions of nodes. Figure 2 shows th architecture of HDFS file system.

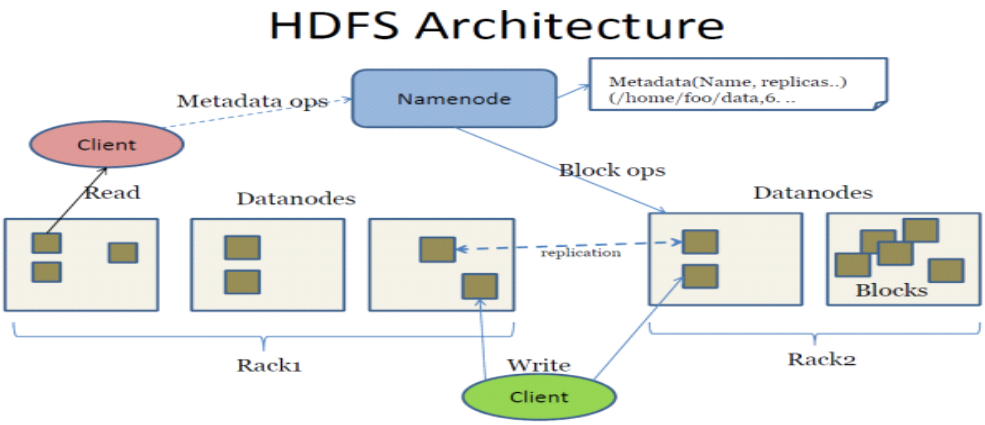

Figure 2: HDFS architecture[14]

Name Node: The Name Node in Hadoop is the node where Hadoop stores all the location information of the files in HDFS. In other words, it holds the Meta data for HDFS. Whenever a file is placed in the cluster a corresponding entry of it location is maintained by the Name Node[13].

Data Node: The Data Node is responsible for storing the files in HDFS. It manages the file blocks within the node. It sends information to the Name Node about the files and blocks stored in that node and responds to the Name Node for all file system operations[13].

Map Reduce: Map Reduce is a data processing paradigm that takes a specification of how the data will be input and output from its two stages (called map and reduce) and then applies this across arbitrarily large data sets. Map Reduce integrates tightly with HDFS, ensuring that wherever possible, Map Reduce tasks run directly on the HDFS nodes that hold the required data. Another key underlying concept is that of "divide and conquer", where a single problem is broken into multiple individual subtasks. This approach becomes even more powerful when the subtasks are executed in parallel; in a perfect case, a task that takes 1000 minutes could be processed in 1 minute by 1,000 parallel subtasks.

Unlike traditional relational databases that require structured data with well-defined schema, Map Reduce and Hadoop work best on semi-structured or unstructured data. Instead of data conforming to rigid schema, the requirement is instead that the data be provided to the map function as a series of key value pairs. The output of the map function is a set of other key value pairs, and the reduce function performs aggregation to collect the final set of results. As the name implies, map/reduce jobs are principally comprised of two steps: the map step and the reduce step[12].

The Map step: Once a map/reduce job is initiated, the map step launches a number of parallel mappers across the computer nodes that contain chunks of your input data. For each chunk, a mapper then "splits" the data into individual lines of text on newline characters (\n). Each split (line of text that was terminated by \n) is given to mapper function. Mapper function is expected to turn each line into zero or more key-value pairs and then "emit" these key-value pairs for the subsequent reduce step.
The map step's job is to transform raw input data into a series of key-value pairs with the expectation that these parsed key-value pairs can be analyzed meaningfully by the reduce step.

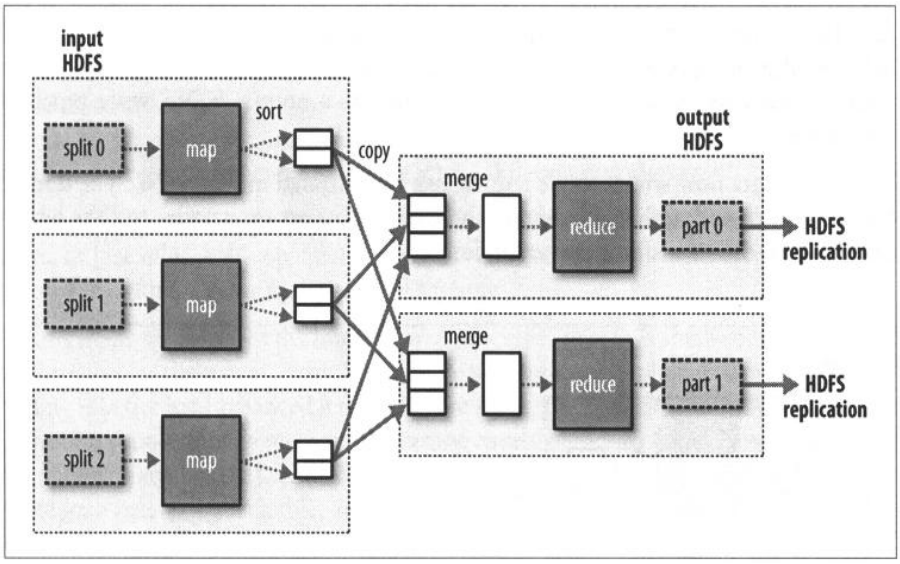

Figure 3: Map Reduce data flow[14]

The Reducer step: Once all of the mappers have finished digesting the input data and have emitted all of their key-value pairs, those key-value pairs are sorted according to their keys and then passed on to the reducers. The reducers are given key-value pairs in such a way that all key-value pairs sharing the same key always go to the same reducer. Reducer function then does some sort of calculation based on all of the values that share a common key. The reducers then emit key-value pairs back to HDFS where each key is unique, and each of these unique keys' values are the result of the reducer function's calculation.
There are lot of other tools to program with if programming with Map-Reduce is not handy for someone. Hive is a tool providing HiveQL for writing Structured Query Language alike queries, which then converted to map reduce and executed on Hadoop. Pig is another language which allow to write scripts in writing the map reduce programming. Hadoop also provides Hadoop Streaming which allows to write the map reduce programs other than Java language.

CONCLUSION

Everyday a large amount of data is generated, to handle this large data called big data a different approach or algorithm is required by the organizations. This analysis of big data can result in finding those patterns from data which are otherwise hidden. So nowadays it is required by a company to process such a massive data to take the decisions. Hadoop is a technique used to handle big data. It can store tens or hundreds of terabytes and even petabytes of data and is able to process it. It mainly contains two components HDFS and Map Reduce. HDFS is used to store the data in chunks on many nodes in the cluster. Map Reduce is then used to process the whole data on the cluster nodes where the data is stored. HDFS itself takes care of failure of nodes and replication. This paper describes the problems of Big Data and challenges and also the techniques used by HDFS and Map Reduce are also discussed.

REFERENCES

1. http://www.genalice.com/technology/big-data-challenge/3vs-of-big-data-5/, Retrieved 2015-02-26
2. H.S.Bhosale, D.P.Gadekar; “A Review Paper on Big Data and Hadoop”, International Journal of Scientific and Research Publications; October 2014, 4(10)
3. Ferguson Mike; Architecting A Big Data Platform for Analytics, Intelligent Business Strategies for IBM; October 2012
4. http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/, Retrieved 2015-02-22.
5. http://www.sas.com/resources/asset/five-big-data-challenges-article.pdf, Retrieved 2015-02-22.
6. http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics, Retrieved 2015-02-22.
7. http://www.computerweekly.com/news/2240219736/Case-Study-How-big-data-powers-the-eBay-customer-journey, Retrieved 2015-02-25
8. Cloudera, Streamlining Healthcare Connectivity with Big Data, http://hadoopilluminated.com/hadoop_illuminated/cached_reports/Cloudera_Case_Study_Healthcare.pdf, 2013.
9. http://hadoopilluminated.com/hadoop_illuminated/Hadoop_Use_Cases.html, Retrieved 2015-01-18.

10. http://research.google.com/archive/gfs.html
11. http://research.google.com/archive/mapreduce.html.
12. http://www.glennklockwood.com/di/hadoop-overview.php, Retrieved 2015-01-07.
13. http://www.rohitmenon.com/index.php/introducing-hadoop-part-ii/, Retrieved 2015-01-205.
14. https://searchenginedeveloper.wordpress.com/tag/hadoop/
15. Garry Turkington, "Hadoop: Beginner's Guide", PACKT Publishing, 2013.
16. “Big Data for Development: Challenges and Opportunities”, Global Pulse, May 2012
17. Jules J. Berman, “Principles of Big Data”, Elsevier India.

Corresponding Author: Piyush Gupta

Dept. of Comp. Sci. and Appl., KUK, Haryana