IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 10, October 2020

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

Hadoop Tools for Big Data Analytics

Anu M S

Lecturer, Department of Computer Engineering, SCMS College of Polytechnics, Kochi, India[1]

ABSTRACT: The internet application and communication have a lot of development and reputation in the field of Technology. These internet applications and communication are continually generating the large size, different variety of structure data called big data. As a consequence, we are handle massive automatic data collection, systematically obtaining many measurements, not knowing which one will be relevant to the phenomenon For example, E-commerce transactions include activities such as online buying, selling or investing. Thus they generate the data which are high in dimensional and complex in structure. The traditional data storage techniques are not suitable to store and analyses those high volume of data. Many researchers are doing their research in dimensionality reduction of the big data for effective and better analytics report and data visualization.

KEYWORDS: Big data, Analytics.

I. INTRODUCTION

Imagine an international without statistics garage; an area in which each element approximately someone or employer, each transaction performed, or each thing which may be documented is misplaced at once after use. Organizations could for that reason lose the potential to extract precious facts and knowledge, carry out specific analyses, in addition to offer new possibilities and advantages. Anything starting from client names and addresses, to merchandise available, to purchases made, to personnel hired, etc. has come to be crucial for everyday continuity. Data is the constructing block upon which any employer thrives. Now consider the volume of info and the surge of statistics and facts furnished in recent times thru the improvements in technology and the internet.

With the growth in garage abilities and techniques of statistics collection, large quantities of statistics have come to be effortlessly available. Every second, increasingly more statistics is being created and wishes to be saved and analyzed if you want to extract cost. Furthermore, statistics has come to be less expensive to store, so companies want to get as a great deal cost as feasible from the large quantities of saved statistics. The size, variety, and speedy extrade of such statistics require a brand new sort of huge statistics analytics, in addition to distinct garage and evaluation techniques. Such sheer quantities of huge statistics want to be nicely analyzed, and pertaining facts have to be extracted.

The purpose of this report is to reflect knowledge and understanding in the intriguing field of big data acquired through various methods. It aims to focus on the importance of understanding big data, envisioning the transformation from traditional analytics into big data analytics, data storage.

II. BIGDATA

Big Data is a large sets of complex data, structured unstructured and semi structured. In which traditional processing techniques or algorithms are unable to operate on these data. Well, the first thing to understand this data is and not all data is created equal. This means the data generated from social media apps are completely different from the data generated by point-of-sales or supply chain systems. All data are different. Some data is structured, but most of data is in a unstructured format. The way in which this type's data is collected, processed, and analysed are depends on its format.

Structured data is most often categorized as quantitative data, and it's the type of data most of us are used to working with. The data is arranged in a fixed fields and columns like relational databases and spreadsheets. The data that have a structure and organized well either in the form of tables etc and that can be easily operated is known as structured data. in a structured data searching and accessing the information is very easy.For example, data stored in the relational database in the form of tables having multiple records and each record contains rows and columns. The spreadsheet is another example of structured data.

Unstructured data is a qualitative data; This form of data cannot be processed and analysed using conventional tools and methods. This type of data is unstructured or unorganized. Such type of data becomes too difficult to process, and

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

itrequires advance tools and software to access information about the data. Examples are images and graphics, pdf files, word document, audio, video, emails, power point presentations, web pages and web contents,

Semi-structured data is a structured data that is unorganised. Web data such JavaScript Object files, text files, XML and other markup languages are the examples of Semi-structured data found on the web. Due to unorganized structure format, the semi-structured is too difficult to retrieve, analyze and store as compared to structured data. It requires software framework for performing operations.

III. CHARACTERISTICS OF BIG DATA

The characteristics of big data are discussed below.

Volume – Current data existing is in petabytes, which it's predicted that in the next few years it's to increase to zettabytes (ZB) [39]. This is due to increase use of mobile devices and social networks primarily.

Velocity – Refers to the rate at which data is captured and the rate of data flow. Increased dependability on live data cause challenges for traditional analytics as the data is too large and continuously in motion.

.

Variety – As data collected is not of a specific category or from a single source, there are numerous raw data formats, obtained from the web, texts, sensors, e-mails, etc. which are structured or unstructured. The traditional analytic methods is failed to manage the big data due to its large size

Veracity – Ambiguity within data is the primary focus in this dimension – typically from noise and abnormalities within the data.

Value-data abstraction is done by a mechanism to bring out the correct meaning out of data. First, need to mine the data, i.e., a process to turn raw data into useful data. Then, an analysis is done on the useful data that you have cleaned or retrieved out of the raw data. Then make sure whatever analysis you have done to get benefits your business such as in finding out insights, results, etc.

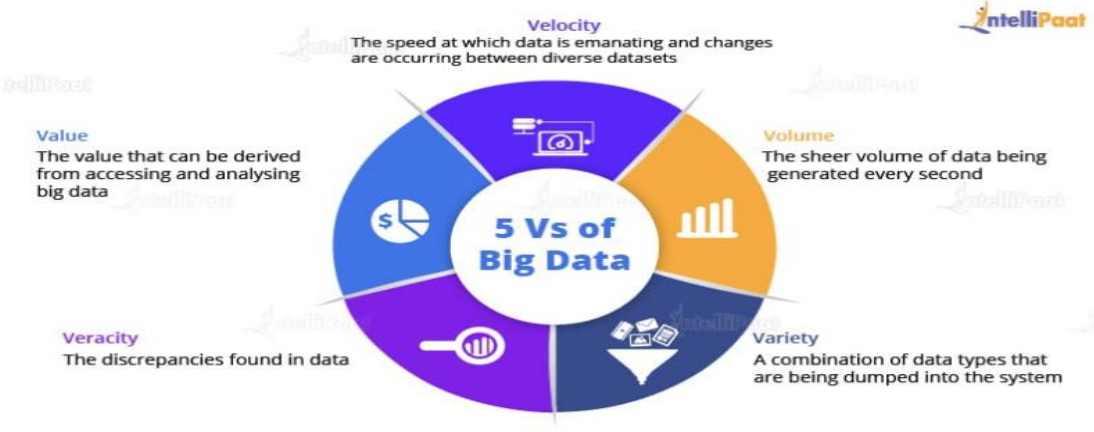

Fig 1: Characteristics of big data

IV. CHALLENGES IN BIG DATA ANALYTICS

Big data has been used in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches purposes. Web-based applications frequently use the big data facilities, such as social network analysis, internet text and documents, and internet search etc. Existing algorithms may not always respond in an proper time when dealing with high dimensional data. Developing new machine learning algorithms to ensure consistency is a major challenge [5] in big data. In addition to all these Clustering of large datasets that help in analyzing the big data is of prime concern. The key challenge is that how to effectively analyze these data for obtaining better results. A standard process to transform the semi structured or unstructured data into structured data, and then apply data mining algorithms to extract knowledge.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

Another challenge [6] with Big Data analysis is related to the diversity of data. With the growing of datasets, data mining tasks have been significantly increased. Additional methods such as data reduction, data selection, and feature selection are an essential task used when dealing with large datasets. The major challenge is to give more attention for designing storage systems and to elevate efficient data analysis tool that provide guarantees on the output when the data comes from different sources. Furthermore, the design of machine learning algorithms to analyse data is essential for improving efficiency and scalability**.**

V. BIG DATA ANALYTICS TRANSFORMATION

In the past decade we have to deal with large volume of data, and at the same time, the price of data storage has systematically reduced. All Private companies and research institutions needs terabytes of data about their users' communications, business purposes, social media, and sensors from devices such as mobile phones and automobiles. The challenge of this to make sense understand the large volume of data. This is where big data analytics becomes more important in the coming days. Big Data Analytics [3] involves collecting large amount of raw data from different sources, combine it in a way that it becomes available to the analysts and finally deliver data products useful to the organization business. The process of converting large amounts of unstructured raw data, collected from different sources to a data product useful for organizations business is the core part of Big Data Analytics [4].

Traditional analytical methods include structured data sets that are periodically queried for specific purposes. A common data model used to manage and process commercial applications using the relational model; this Relational Database Management Systems (RDBMS) provides "user-friendly query languages" and provides simplicity for processing data rather than network or hierarchical models are not able to deliver. In this system data are represented as tables, each with a unique name, where data is stored in rows and columns.

These streams of data are obtained through integrated databases and use that data set in an intended purpose. Frequent usage of mobile devices, Web applications, and growth in the Internet of Things (Io T) are among a few to mention in reasoning behind organizations looking to transform analytical processes. Organizations are attracted to big data analytics as it provides a means of obtaining real time data, suitable for enhancing business operations Data storage is now extremely complex. The data-sets are mapped into knowledge-space in order to drive some value to a business as well as the costs involved.

VI. TOOLS FOR BIG DATA PROCESSING

Several tools are available to process big data. In this section, we discuss some current techniques [7] for analysing big data with emphasis on emerging tools namely Hadoop and Mapreduce . The most established software platform for big data analysis is Apache Hadoop and Mapreduce. It consists of hadoop kernel, mapreduce, hadoop distributed file system (HDFS). Map reduce is a programming model for processing large datasets. It is based on divide and conquers method. The divide and conquer method is implemented by using two steps such as Map step and Reduce Step. Hadoop works based on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step based on divide and conquers method. Thereafter the master node combines the outputs of all the subproblems in reduce step. Moreover, Hadoop and MapReduce work as a powerful software framework for solving big data problems. Its also helpful in fault-tolerant storage and high throughput data processing.

Hadoop Architecture

Apache Hadoop was developed with the goal of having an inexpensive, redundant data store that would enable organizations to leverage Big Data Analytics economically and increase the profitability of the business. A Hadoop architectural design needs to have several design factors in terms of networking, computing power, and storage. Hadoop provides a reliable, scalable, flexible, and distributed computing Big Data framework.Hadoop follows master–slave architecture for storing data and data processing. This master–slave architecture has master nodes and slave nodes as shown in the image below:

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

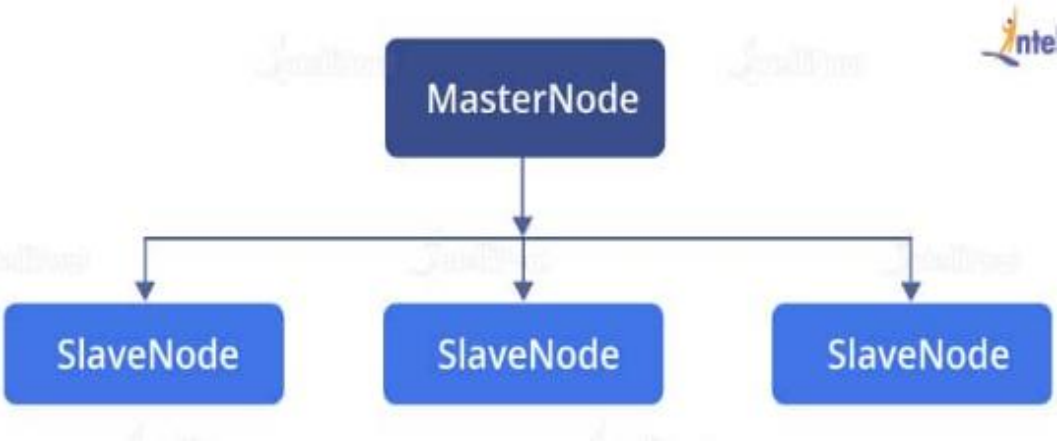

Fig2: Hadoop master-slave architecture

What is Apache Hadoop?

Apache Hadoop was used to enhance the usage and solve major issues of big data. The web applications generate loads of information on a daily basis, and it was very difficult to manage the data of around one billion pages of content. In that purpose, Google invented a new methodology for processing large volume of data popularly known as MapReduce. . Considering the original case study, Hadoop was designed with simpler storage infrastructure facilities.

Apache Hadoop is the important framework for working with Big Data. Hadoop biggest strength is scalability. It upgrades from working on a single node to thousands of nodes without any issue in a seamless manner.

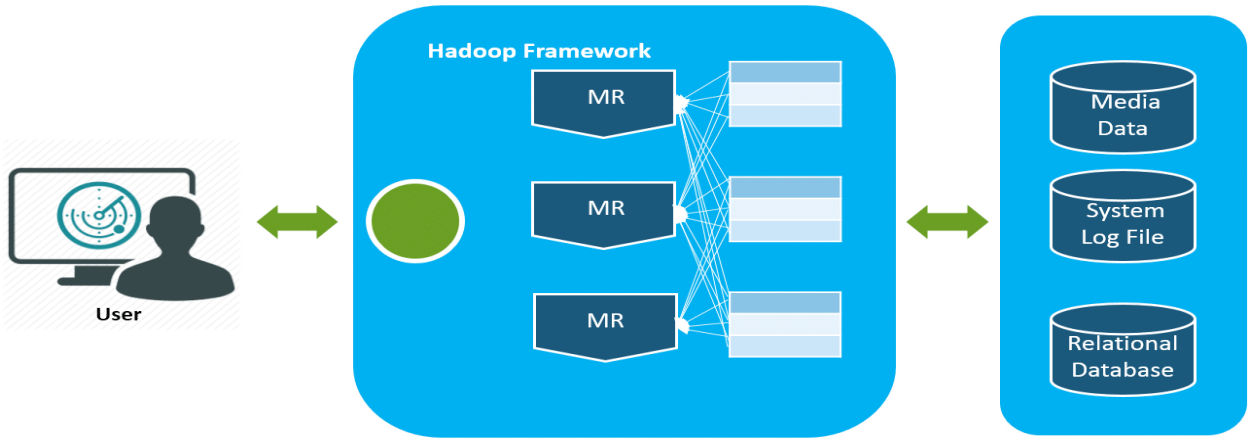

Fig3: Hadoop framework

The different domains of Big Data means we are able to manage the data's are from videos, text medium, transactional data, sensor information, statistical data, social media conversations, search engine queries, ecommerce data, financial information, weather data, news updates, forum discussions, executive reports, and so on Google's Doug Cutting and his team members developed an Open Source Project namely known as HADOOP which used to handle the very large amount of data. Hadoop runs the applications on the basis of MapReduce where data is processed in parallel and provides the entire statistical analysis on large volume of data.It is a framework which is based on java programming. It is intended to work on a single server to thousands of machines each offering local computation and storage. It supports the large collection of data set in a distributed computing environment.The Apache Hadoop software library based framework gives permissions to distribute large volume of data sets processing across clusters of

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

computers using easy programming models.Hadoop high level Architecture based on the two main components namely MapReduce and HDFS.

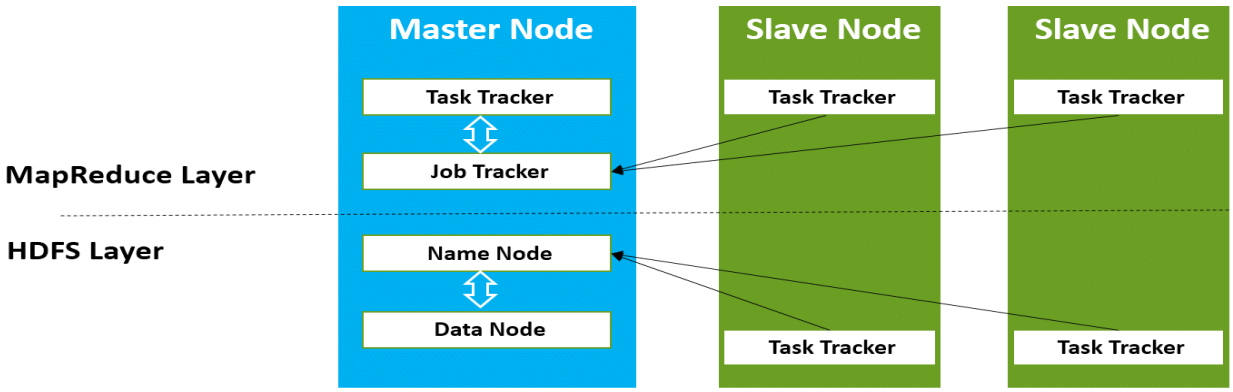

Fig4: Hadoop high level architecture

- **HDFS:** Hadoop Distributed File System provides unrestricted, high-speed access to the data application.
- **MapReduce:** This is a highly efficient methodology for parallel processing of huge volumes of data.

- Hadoop Mapreduce (Processing/Computation layer) –MapReduce is a parallel programming model used for writing large volume of data distribution applications devised from Google for efficient processing of large volume of datasets, on large group of clusters.
- Hadoop HDFS (Storage layer) –Hadoop Distributed File SystemorHDFS is based on the Google File System (GFS) which provides a distributed file system that is designed to run on commodity hardware. It reduces the faults or errors and helps to incorporate low-cost hardware. It gives high level processing throughput to application data and is suitable for applications deals with large datasets.
- Task Tracker –It is a node which is used to accept the tasks such as shuffle and Mapreduce form job tracker.
- Job Tracker –It is a service provider which runs Mapreduce jobs on cluster.
- Name Node –It is a node where Hadoop stores all file location information (data stored location) in Hadoop distributed file system.
- Data Node – The data is stored in the Hadoop distributed file system.

How does Hadoop work?

Hadoop helps to execute large volume of data processing where the user can connect together by multiple computers to a single-CPU, as a single functional distributed system and have a particular set of clustered machines that reads the large data set in parallel and provide intermediate desired output.

Hadoop runs code across a cluster of computers and performs the following tasks:

- First data are divided into files and directories. Files are then divided into consistent sized blocks ranging from 128M and 64M.
- Then the files are distributed across various cluster nodes for further processing of data.
- Job tracker starts its scheduling programs applied on individual nodes.
- After all the nodes are done with scheduling then the output is return back.

What is MapReduce in Hadoop?

Now that you know about HDFS, it is time to talk about MapReduce. So, in this section, we're going to learn the basic concepts of MapReduce. MapReduce [1] in Hadoop are nothing but the processing model in Hadoop. The programming model of MapReduce is designed to process huge volumes of data parallelly by dividing the work into a set of independent tasks. As we learned in the Hadoop architecture, the complete job or work is submitted by the user to the master node, which is further divided into small tasks, i.e., into slave nodes.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

MapReduce.

MapReduce[2] in Hadoop is a distributed programming model for processing large datasets. This concept was conceived at Google and Hadoop adopted it. It can be implemented in any programming language, and Hadoop supports a lot of programming languages to write MapReduce programs.

You can write a MapReduce program in Scala, Python, C++, or Java. MapReduce is not a programming language; rather, it is a programming model. So, the MapReduce system in Hadoop manages data transfer for parallel execution across distributed servers or nodes.

How does MapReduce in Hadoop work?

Let's explore a scenario.

Imagine that the governor of the state assigns you to manage the upcoming census drive for State A. Your task is to find the population of all cities in State A. You are given four months to do this with all the resources you may require.

So, how would you complete this task? Estimating the population of all cities in a big state is not a simple task for a single person. A practical thing to do would be to actually divide the state by cities and assigning each city to separate individuals. They would calculate the population of the respective cities.

Here is an illustration for this scenario considering only three cities, X, Y, and Z:

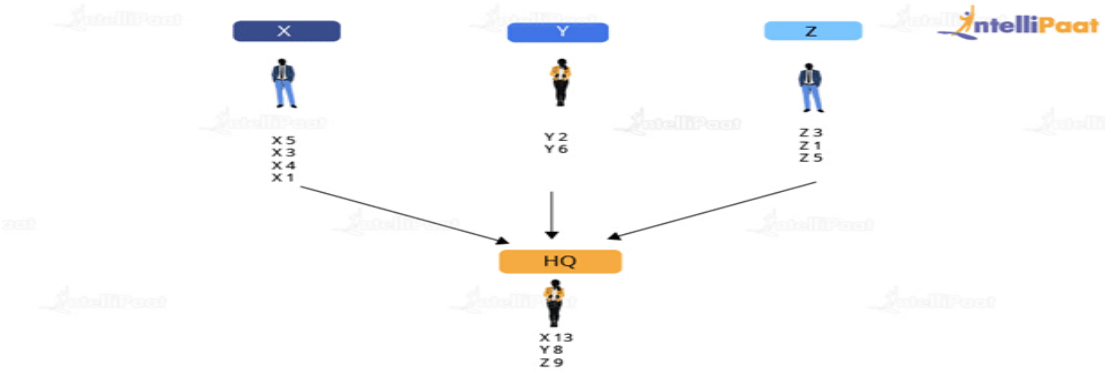

Fig5: Hadoop working scenario.

Person 1, Person 2, and Person 3 will be in charge of the X, Y, and Z cities, respectively. So broken down State A into cities where each city is allocated to different people, and they are responsible for figuring out the population of their own respective cities. Now, you provide some specific instructions to each of them. You will ask them to go to home and find out how many people live there.

Assume, there are five people in the home Person 1 first visits in City X. He notes down, X 5. It's a proper divide and conquer approach .The same set of instructions would be carried out by everyone associated with that. That means that Person 2 will go to City Y and Person 3 will go to City Z and will do the same things. At the end, they would need to submit their results to the state's headquarters where someone would aggregate the results. Hence, by using this strategy, you will be calculating the population of State A in four months.

Next year, you're do the same job but in two months. What would you do? So, you divide City X into two parts and would assign one part to Person 1 and have one more person to take charge of the other part. Then, the same thing would be done for Cities Y and Z, and the same set of instructions would be given again.

Also there would be two headquarters HQ1 and HQ2. So, you will ask the census takers at X1 and X2 to send their results to HQ1 or HQ2. Similarly, you would instruct census takers for Cities Y and Z and tell them that they should either send the results to HQ1 or HQ2. Problem solved. So, with twice the force, you would be able to achieve the same within two months.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

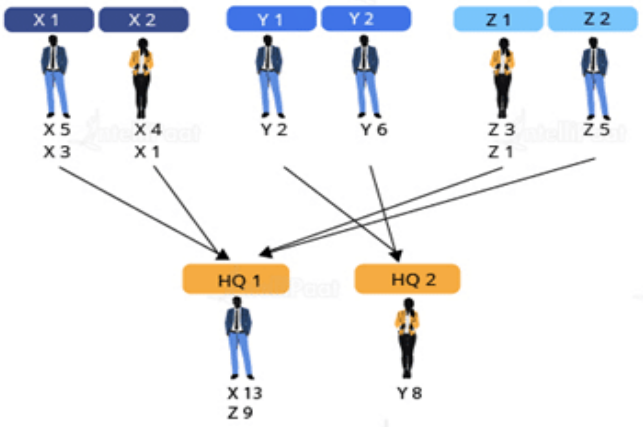

Fig6: Hadoop working scenario.

This model is called MapReduce.MapReduce in Hadoop is a distributed programming model for processing large amount of datasets. This concept was conceived at Google and Hadoop adopted it. This can be implemented by using any programming language, Hadoop supports a lot of programming languages to write MapReduce programs.You can write a MapReduce program in Python, C++, or Java. MapReduce is not a programming language; rather than it is a programming model. The MapReduce concept in Hadoop manages data transfer for parallel execution across distributed servers or nodes.

Phases in MapReduce

There are mainly three phases in MapReduce: the map phase, the shuffle phase, and the reduce phase.

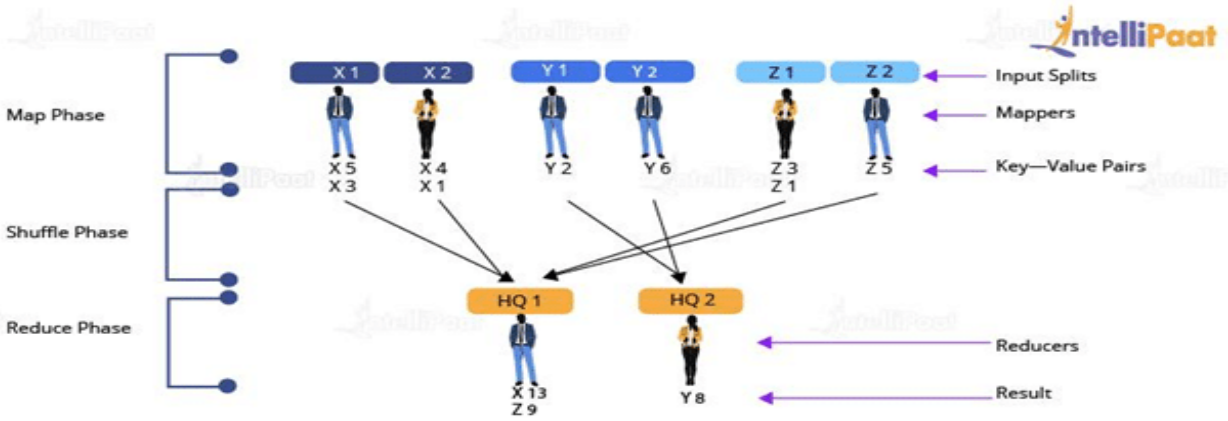

Fig7: Hadoop working scenario.

Map Phase

The phase where individuals count the population of their assigned cities is called the map phase. There are some terms related to this Map phase.

Mapper: The individual (census taker) involved in the actual calculation is called a mapper.

Input split: The part of the city each census taker is assigned with is known as the input split.

Key–Value pair: The output from each mapper is a key–value pair. The key is X and the value is, say, 5.

Reduce Phase

The large data set have been broken down into various input splits and the instances of the tasks have been processed. Then the reduce phase comes into place. Like the map phase, the reduce phase processes each key separately.

Reducers: The individuals who work in the headquarters are known as the reducers. This is because they reduce or consolidate the outputs from many different mappers.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

After finishing the task of reducer, a results file is generated, which is stored in HDFS. Then, HDFS replicates these results.

Shuffle Phase

I this phase in which the values from different mappers are copied or transferred to reducers is known as the shuffle phase.The shuffle phase comes in-between the map phase and the reduce phase.

The architectural view of how map and reduce work together:

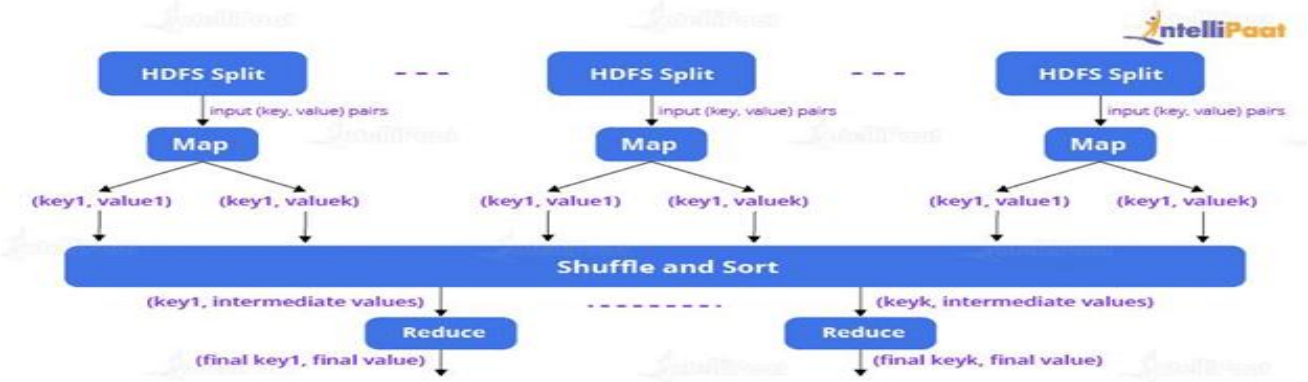

Fig8: Hadoop working scenario.

When the input data is given to a mapper, it is processed through some user-defined functions that are written in the mapper. The output is generated by the mapper, i.e., the intermediate data. This intermediate data is the input for a reducer.The output from the reducer is then processed by some user-defined functions that are written in the reducer, and then the final output is produced. Further, this output is saved in HDFS and the replication is done.

VII. CONCLUSION

The Big Data analytics concept is continually growing. Its environment demonstrates great opportunities for Organizations within various sectors to compete with competitive advantages, as shown in the examples mentioned earlier. The future of the fields such as medical science is changing dramatically due to this concept, scientist are able to access data rapidly on a global scale via the cloud, and these analytics contribute to the development of predictive analytics tools .Due to inconsistencies and challenges within Big Data privacy, sufficient encryption algorithms to conceal raw data or analysis, reliability & integrity of Big Data. This paper shed light on conceptual ideologies about big data analytics and displayed through a few scenarios how it's beneficial for organizations within various sectors if analytics are conducted correctly. Further areas to research to grasp knowledge of Big Data are data processing & data transfer techniques.

REFERENCES

[1] S Grolinger, K, 2014. Challenges for MapReduce in Big Data. 2014 IEEE 10th World Congress on Services, 1, 182-183..

[2] IBM. 2015. IBM - What is MapReduce. [ONLINE] Available at:https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/.

[3] Park, K, 2015. Web-based Collaborative Big Data Analytics on Big Data as a Service Platform. Web-based Collaborative Big Data Analytics on Big Data as a Service Platform, 1, 564-566.

[4] Hu, H, 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, 1, 658-659, 665.

[5] Katal, A, 2013. Big Data: Issues, Challenges, Tools and Good Practices.Big Data: Issues, Challenges, Tools and Good Practices, 1, 404.

[6] Bhardwaj, V, 2015. Big Data Analysis: Issues and Challenges. Big Data Analysis: Issues and Challenges, 1, 1-3.

[7] Simon Robinson. 2012. The Storage and Transfer Challenges of Big Data. [ONLINE] Available at: http://sloanreview.mit.edu/article/thestorage- and-transfer-challenges-of-big-data/. [Accessed 29 December

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details