International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023,   pp: 693-702www.ijaem.net   ISSN: 2395-5252

Maintaining the quality of big data using the Hadoop platform

Sakshi Sanjay Chalke[1], Shivani Saha[2], Krutika Prakash Bhagane[3]

[1,2]*Department of Computer Engineering and Information Technology Terna Engineering College Nerul West, Navi Mumbai*

[3]*Department of Computer Engineering Pillai HOC College of Engineering and Technology, Rasayani, Navi Mumbai*

Date of Submission: 12-01-2023   Date of Acceptance: 24-01-2023

ABSTRACT: The amount of data within today's high-tech society has grown dramatically over time. We have big data, which works with all data formats, to handle this massive volume of data. This study highlights the seven v's of big data and also the necessity of processing the data in a usable manner. It provides a brief overview of the Hadoop tool and how it aids in the pre-processing of massive amounts of data. Data mining is a component of data pre-processing that helps numerous companies make informed decisions based on the data at hand by identifying comparable or useful patterns in data.

KEYWORDS: Big data, Hadoop platform, Data mining, Seven V's of big data.

I. INTRODUCTION

Quality in data is very important while we are working with big data. Since Big data can store large amount of data and in multiple formats, however it does not guarantee's that the stored data is in usable form. Infect, the data which is directly store in the memory location offered by big data is often called as raw data or unprocessed data.

A set of information that was delivered to the data provider from a particular data entity but has not yet been processed by a machine or a person is called raw data. In order to provide a comprehensive understanding of users' online behaviour, this data is gathered from online sources. With this information, marketers can easily create customized online campaigns and send the right message to the right people at the right time to their target audience.

It is important to acknowledge that raw data, when not processed by algorithms, is of little use. Most of the time, it is just a bunch of code, like a user cookie, that does not give much information. However, when this data is combined with the right user profiles, it really helps marketers or business analysts.

So, how can we pre-process such large amount of data very effectively?

Clearly, it impossible for us human to manually pre-process such large volume of data, it would be very tedious as well as very time-consuming task and in the end the data which we get would definitely contain errors due to human mistakes. Manually pre-processing is not at all efficient enough. Instead, we need a software that will help us to pre-process big data in less time and less efforts.

The traditional method of data management deals with restricted data storage and works with structured and semi-structured data. It does not process unstructured data, which might include audio, video, and other types of data. We have big data, which stores an incredibly vast amount of data. There are several tools available to pre-process this data, but Hadoop is one of the best tool among them.

Hadoop is an open-source framework created by Apache that is used to store, process, and analyse data that is extremely large in volume. It provides us with enormous data storage capacity, enormous computational power, and the capacity to manage almost countless jobs, including running jobs, waiting jobs, and tasks. Its primary crucial function is to enable developing big data technologies, which in turn supports cutting-edge analytics including predictive analytics, machine learning, and data mining. Hadoop can manage a variety of data types, including structured, unstructured, and semi-structured data. It allows us the flexibility to gather, process, and research data that the outdated data warehouses idea lacked.

Why Hadoop is considered as best tool for processing bigdata?

Hadoop has the following features which make it better than any other tools, these features are:-

High scalability - We can add unlimited number of nodes, substantially increasing performance.

High availability - Hadoop makes data highly available even when hardware malfunctions. If a system or a piece of hardware fails, we can still access data via another method.

Reliable - In spite of machine failure, data is reliably kept on the cluster.

Cost-effectiveness- Hadoop uses a cluster of inexpensive commodity hardware to run.

II. NEED FOR DATA QUALITY IN BIG DATA

2.1. Unauthorized data collection

Traditional data collection is often performed or monitored by scientific researchers, research institutes or government agencies organ. Data collected by these competent scientific bodies often have high authenticity and reliability of data because researchers from these institutions generally adhere to research ethics and follow good scientific practices. These establishments also have more resources and power to complete these missions. These scientific data collection tasks are what these researchers, research institutions and government agencies are hired and get paid to do.

In big data, the data that is collected may not be monitored by the researchers or may not be collected ethically. For example, social media data is collected from Twitter, Meta and other social media platforms. These organizations are commercial companies not established for the purpose of scientific research but rather commercial platforms for profit. Commercial platforms have no obligation or incentive to guarantee the authenticity and validity of the data they collect. A good example is social media like Twitter and Meta that have many bots or machine-run "zombie" accounts. These companies have no desire to "discard" these substandard models. Instead, these companies rely on these accounts to make money.

Since big data is used in the commercials companies mentioned previously, the data these companies have are raw, that is means the data collected may be of different format it can be text, image, pdf etc.it becomes difficult to work in these data, therefore pre-processing of these collected data is very much necessary.

2.2. Noise and insufficient comprehensive Information

When we are dealing with big data which is like ocean for information since big data consist of large volume of information from different sources. These varied sources can produce data which have no meaning at all, it cannot even be interpreted by the machines either these data are referred as noise. Storing noisy data is not necessary, these data should be removed as soon as possible for this pre-processing of data is highly recommended.

When it commercial companies they need information about their subscribers to know what they want from latter’s establishment not only that the commercial companies like Flipkart, Amazon etc. , they use their subscriber information to recommend them products according to subscriber’s searches in different browsers .This helps in selling products more efficiently.

2.3. Data collection may not be consistent or reliable

Real – time data is used in many iot devices, the data quality for iot analytics is affected due to:-

i. Lossy networks result from constrained devices. Missing or inconsistent data are frequently the outcome for analytics. The missing data is frequently not random. The environment might have an effect. Firmware, the operating system that powers devices, may not be consistent from one site to the next. This could indicate variations in reporting frequency or value presentation. Data loss or corruption may occur as a result.

ii. IoT device data messages frequently require the recipient to understand how to interpret the message being transmitted. Messaging and data records can become jumbled as a result of software flaws.

iii. Values are absent when communications are lost in translation or are never sent because of dead batteries. Not all values accessible on the device are frequently supplied at once due to battery conservation. Since the device sends some numbers consistently every time it reports while sending other values less frequently, the resulting datasets frequently contain missing values.

To overcome all the above issues mentioned above we need to pre-process the data to overcome the hindrances that will affect working with big data. Hadoop is a widely used tool for pre-processing large volume raw data and converting it to a usable form.

Looking back on this project, the overall outcome of results to be observed. This can be evaluated by looking at how well our objectives were met. Our first objective is to control the engine valve of an engine, select a linear actuator that meets specifications, and construct an electronic control system, deal with the design aspect of our project and were all almost achieved. More specifically, next objective, the electronic control system we constructed is able to read engine speeds from 0 to 3600 rpm and vary the valve timing depending on engine speed and operator inputs. However, our final objective, to obtain gains in horsepower, torque, and efficiency of 2% was not met because of not setting up in an engine but theoretically it should be done. We are confident though that this objective of installing in an engine can be met if more time for testing and facilities is given. There is a lot we could say about the need for variable valve timing. This design is very realistic for the future of the automotive industry as well as our education.

Some of the Advanages from the Above Results

a) Eliminated Mechanical Linkages
b) It can make Engine clean , efficient and responsive
c) ECU can control the valve velocity acceleration and deceleration of valve
d) Reduction in size and weight
e) Fuel economy Increases
f) Power and Torque increase

III. BIG DATA

If the data has expanded to such an extent that now a single computing system is unable to store and process it, then we call this data “Big Data”. Big data refers to large datasets which are very complex to process by the traditional system. The traditional system is not able to handle the huge data getting generated and ultimately causing Denial of Service hence a scientist John Marshey in the 1990s analysed this situation and the need for huge processing for managing the data. The concept of Big data can be analysed by using the concept of the 7 V’s. The main characteristics which define big data are volume, velocity, and variety. The sources through which data is generated have become complex as this big data have been extracted from social media, artificial intelligence(AI), mobile devices, and the Internet of things(IoT). In today’s scenario, big data plays an important role in analysing any organization and field and help for better decision making it also enhances and adds meaning to the organization. Big data help administrations to represent multiple operations like analysing the data, storing data, processing the data to extract value, updating the data, and visualization on a single platform. The most important advantage of big data is to Improve Efficiency.

IV. SEVEN V’S OF BIG DATA (CHARACTERISTICS OF BIG DATA)

1. Volume:

It refers to the amount of data generated per second. Huge Data coming from different sources would need good storage and computing power. It is one of the important characteristics which tells about the size of big data which has been generated. If the volume is large enough, we can say that the data generated is big data. Size play important role in handling big data. “Big Data” itself refers to data that is huge in size. So depending on the volume we can say whether the data generated is big data or not. For example- number of tweets on Twitter, it was observed that in August 2014 there were 661 million tweets that were recorded in a 30-day sampling period.

2. Velocity:

The speed at which the data is generated. The huge volume of data collected from various sources like social media, mobile phones, industries, etc. has a massive and continuous flow of data. Rapid generation and the rate at which data is generated leads to a sudden spike in data volume, which also consist of real-time data flow. This is one of the important characteristics of big data which gives us an idea about the data generated continuously. For example- 2.1 million tweets per second on Twitter

3. Variety:

It refers to the nature of data that is generated, different types of big data are generated whether structured, unstructured, or semi-structured. Data is collected from heterogeneous sources. A huge amount of data is generated from the internal and external parts of the enterprise can be in different types of big data

Structured Data – this type of data follows a strict structure, approximately 5% of the data which is generated in huge volumes consists of structured data. E.g. Relational Data Base.

Unstructured Data – this type of data does not have any definite structure or pattern, approximately 80% of the data generated is in unstructured form. E.g. Mongo DB

Semi-structured Data - some data is structured and some are unstructured, which only refers to unorganized data. Approximately 15% of the data which is generated is in this form. E.g. Email.

4. Veracity:

It defines the uncertainty of the data. Huge data which is collected or generated from different sources can lead to data inconsistency. The huge volume of the data available can be in a complex form which can be difficult to maintain the accuracy and the quality of big data. This characteristic of big data helps to build robust systems. For example – In some form, a 10-digit number is asked to mention and the entered data is “A1B2C3D4E5”.

5. Value(Validity):

It refers to how much correct information is retrieved. The huge volume of data that is generated but consists of no value turns out to be useless for that organization. The main advantage of extracting value from big data is strategic decision-making. Many business decision-making can proceed with the help of value and consist of Competitive advantages. Big data itself is of no use if the proper value from it is not extracted. For example – ask a 10-digit number and the user enters data “0123456789”.

6. Variability:

It refers to the way the meaning of data keeps changing. This characteristic shows how dynamic the behaviour of data can be, if the meaning changes continuously it will impact the data standardization. Variability leads to a change in perspectives with a change in a variable. For example- the word “drowning” in both sentences has a completely different meaning

1) Student was drowning while learning to swim.
2) Student is drowning in pressure due to a lot of assignments.

7. Volatility:

It determines how fast the data which is acquired will expire. It also refers to the validity of the lifetime of particular data. Data generated in huge volume and increasing velocity need to take the information(value) from it as soon as possible as each data has its lifecycle.

For example – Important news headlines should be processed and presented to the real world on priority.

V. MINING IN BIG DATA:

Mining in big data means collective extraction techniques that are performed on large volumes of data or what we call big data. Big data mining enables to obtain useful information from databases that are huge in terms of Big Data V’s like volume, velocity and variety. It is done on all types of data especially unstructured data. It works mostly on data searching, extraction, comparing algorithms and many more. It requires support from various computing devices, different operating systems or performing various operations or queries on the huge chunk of data.

These techniques are used in big data for analytics, business intelligence to give relevant information and perform various processes. Nowadays data is being sent to the global network not only by people but also by various sensors and with the use of cloud computing where huge chunk of data gets stored. The main functions of data mining are predictive functions such as classification, time series analysis, etc. and descriptive functions such as clustering, association and pattern mining. These functions differ from each other mainly in terms of temporal reference, descriptive functions mainly focuses on present and settled dependencies, and the predictive refer to the study of future and tentative dependencies. It is most important for social media for hidden customer insights. Data mining techniques are most used ones in big data analytics. The success of big data analytics depends on factors as top management support, organizational change, technical infrastructure, the data science skillset, data availability and quality, security and privacy. The big data mining in real life may make decision making processes in organization because it deals with real time uncertainties. There are many data mining techniques namely clustering, association, data cleaning, data visualization, classification, and many more.

5.1 Clustering:

Clustering is a process of grouping a series of different type of data point based on their characteristics.

There are various methods of data clustering:

- **Partitioning method**: This is dividing a data set into group of specific clusters for evaluation based on criteria of individual cluster.
- **Hierarchical method**: Data points are a single cluster, based on grouped based on similarities.
- **Density based method**: A ML method where data points are plotted together are analysed, but the data is discarded.
- **Grid based method**: This is dividing data into cells on a grid, which then can be clustered by individual cells rather than the entire database.
- **Model based method**: Models are created for each data cluster to locate the best data to fit in a model.

5.2 Association:

Association rules are used to find correlations, or associations between points in a data set.

Method for data mining association:

- **Single dimensional association:** This involves looking for one repeating instance of a data point or attribute.
- **Multi-dimensional association :** This involves looking for more than one data point in a dataset.

5.3Data cleaning:

Data cleaning is the process of preparing data to be mined.

Methods for data cleaning

- **Verifying the data:** It involves checking each data point in the data set for proper format.
- **Converting data types:** This ensures data is uniform across dataset.
- **Removing errors:** This eliminates typing mistake, spelling errors, and input errors that could negatively affect analysis outcomes.

5.4 Data visualization:

Data visualization is the translation of data into graphic form to illustrate its meaning to business stakeholders.

- **Comparison charts:** Charts and tables express relationships in the data, such as monthly product sales over a one-year period.
- **Maps:** Data maps are used to visualise data pertaining to specific locations.
- **Density plots:** These visualizations track data over a period of time, creating a lookalike mountain range. Easy to represent occurrences of single event over time.
- **Histograms:** Like density plots but are represented by bar on a graph instead of linear form.
- **Scatter plots:** Represent data points relationships on a two-variable axis. Scatter points can be used to compare unique variables.

5.5 Classification:

Is a fundamental technique in data mining and can be applied to every industry. It is a form of clustering that I useful for extracting comparable points of data for analysis.

Methods for classification:

- **Logistic regression:** This algorithm attempts to show the probability of outcome.
- **Naïve Bayes:** This uses historical data to predict whether similar events will occur based on different dataset.
- **Support vector machine:** This algorithm is used to define the line that best divides a dataset into two classes.

VI. HADOOP:

It is an open-source platform developed by the Apache software foundation based on java implementation. Hadoop is the platform where big data can be handled efficiently by storing and processing it. Hadoop is the framework that is divided into two phases which are Hadoop Distributed File System (HDFS) and MapReduce. The main aim is that it distributes the data on multiple locations and makes sure that all the data are processed without denying any service for a user it appears as a coherent system. A large amount of data whose size ranges from gigabytes to petabytes can be stored and processed. Hadoop allows the assembling of multiple systems to examine huge volumes of datasets in parallel more quickly.

The core components of Hadoop include MapReduce, YARN (Yet Another Resources Negotiator), HDFS (Hadoop distributed File System), Hadoop Common (includes in others).

6.1 Components of Hadoop:

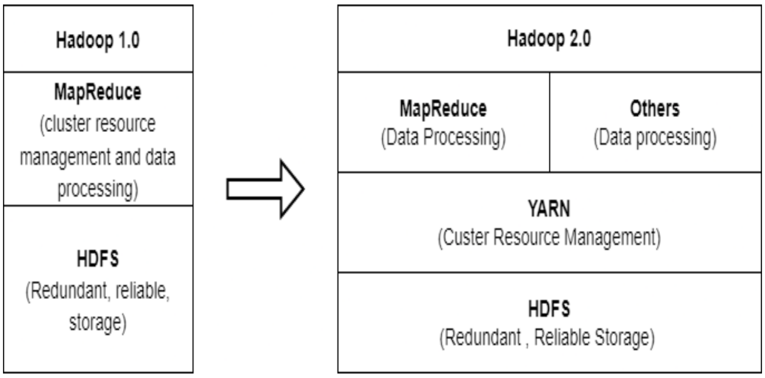
Hadoop 1.0
MapReduce
(cluster resource management and data processing)
HDFS
(Redundant, reliable, storage)
Hadoop 2.0
MapReduce
(Data Processing)
Others
(Data processing)
YARN
(Custer Resource Management)
HDFS
(Redundant , Reliable Storage)

Fig 6.1.1 Component of Hadoop

- **HDFS (Hadoop Distributed File System**): This component is used for managing storage. It is also referred as the storage unit of Hadoop. The data which is collected from the client is referred to as the big data Hadoop takes this input data and precedes it to HDFS, the main aim of this component is to break down the big data into small chunks of data and store them into different nodes. The name node present in HDFS divides the data into blocks and then stores them in Data Node
- **MapReduce:** It manages resources, job scheduling, and performs data processing. The data collected in the data node acts as a slave handover data to MapReduce for further data processing. Parallel processing takes place in this component. MapReduce consists of various internal processes which includes splitting, mapping, shuffling, and reducing. The data in this component is reduced in the final result. This phase act as the important phase for processing the big data.
- **YARN (Yet Another Resource Negotiator):** In Hadoop 2.0 the task of MapReduce which is resources management is given to YARN. The two main components of YARN are job tracker and job scheduling/monitoring.YARN supports batch processing, stream processing, interactive processing, and graph processing of the data which is stored in the Data node in HDFS.
- **Hadoop Common:** These components consist of java files and libraries which are required for the Hadoop framework also the java scripts which we require for all the other components are maintained here.

6.2Simple Block Diagram of Hadoop:

Data is submitted to Hadoop. There are two phases in Hadoop which include HDFS and MapReduce. The name node (NN) which is present in the HDFS distributes that data among the data node (DN). HDFS is the component in Hadoop that is used for storing purposes the main function is working in distributed file system design. Big data is collected from different sources consisting of different types of data in form structured and unstructured manners hence HDFS only allows the data in a large chunk of the block. Only the Big data is supported by the Hadoop framework. HDFS also includes fault-tolerance and high availability of data present as the architecture of HDFS consists of its replication of data. The Name node and Data node are present in HDFS. Here the Name node acts as the master, which guides the Data node to store the data. The Name node only stores the metadata (data about the data) and instructs the Data node for the operations like delete, create, etc. Data node act as a slave in the process, which follows the instruction given Name node. The data node stores the data. To store more data more Data nodes are required. Here the storage of big data depends upon the number of Data nodes present in HDFS. As shown

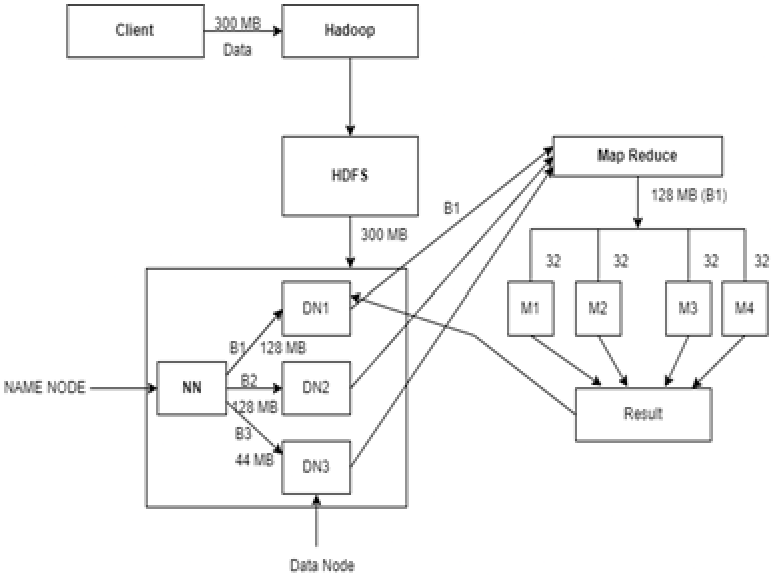

Fig 6.2.1 Block diagram of Hadoop

in the figure above the data is given as input of 300MB and is distributed to the Data node with the help of the Name node. The default value at which it is distributed is 128MB hence the data is distributed in (128MB + 128MB + 44MB) among the Data node (DN1, DN2, DN3). After the storage MapReduce comes into the picture, where this component works in two phases named as Mapping phase and the Reducing phase. MapReduce processing works in a parallel manner and hence this process is done fast. MapReduce takes the block from the Data node and processes it parallelly. The data is first mapped and at the file stage, it is reduced with the help of key values. The 128MB in this phase is split among the machines in (32MB + 32MB + 32MB +32 MB). And then the final result is merged and stored back in the respected data node in HDFS.

VII. HDFS ARCHITECTURE

Managing and storing the big data is done by Hadoop distributed file system. As shown in the above diagram the massive data which is given as the input to the Hadoop platform, HDFS splits the file into small data blocks of size 128 MB, therefore the data in the distributed manner will be B1=128 MB, B2 = 128 MB and B3 = 4 MB (the input data is 260 MB). Now the most important component of HDFS is the Name node which is the primary component for managing the files and directories. The name node is responsible to decide in which data node the block should be stored. The data blocks are distributed and stored in data nodes. The number of data nodes decides the storage capacity. The main advantage of using the Hadoop tool is the quality of the data is preserved. The secondary name node is also called as checkpoint node. Since false tolerance and reliability is the matter of concern, the Name node performs the replication of the block in HDFS (default replication factor is 3). Therefore, the data blocks present consist of the replication of the data in different data nodes. Data nodes are considered the logical hard disk of Hadoop.

Certain measures are taken in HDFS to maintain the quality of data.

- **What if the Name node fails?**

If the name node fails, then the entire HDFS will collapse. The solution for this failure can be improved with the help of a secondary name node. Here the secondary name node comes into the picture, the primary node continuously sense logs to the secondary name node when the primary node fails a standby node temporarily takes the charge of the primary node, it takes the entire current state information from the secondary name node.

- **What if the Data node fails?**

The data node periodically sends a heartbeat message to the name node. If the name node doesn't receive the heartbeat message from the data node for ten minutes it will assume that the data nose has failed. The name node will check metadata and migrate the data of failed node to the other data node and will make sure the replication factor is maintained. In this manner, the data in the Hadoop platform is maintained and stored properly without dropping any data block from the data node.

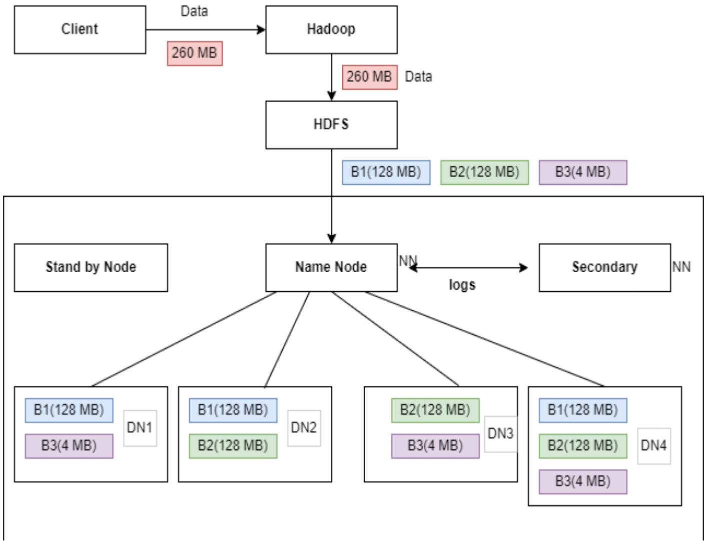

Fig 7.1 HDFS Architecture

7.1 HDFS Read:

Whenever the read request comes for a particular file, the Name node refers to metadata to find the number of blocks of that file and where that blocks are stored. It then requests the respected data node for fetching that block. After this process, it joins all the blocks and returns the file to the requested client. If the error occurs during read operations that the data block is corrupted, then the solution is to replace that data block with its replica.

7.2HDFS Write:

The client sends a request for the HDFS client for writing a file which is forwarded to the name node for creating the file. The name node creates the file and sends a leasing message. Now client starts sending the file which is stored in the HDFS client buffer, once it reaches the specified block size, it sends that block to the name node. Name node after receiving all the blocks list out on which the data node should be stored. HDFS client sends those blocks in the data queue and further to one of the data node. E.g. DN2. The DN2 forwards the block to the Acknowledgement queue. Once all the acknowledgments are received for the data block from the data node, those blocks are dequeued (removed) from the acknowledgment queue and written a complete message to the Name node.

If the data node fails while writing the data?

When the name node realizes (eg DN3 has failed), it will move all the blocks from the acknowledgment queue to the data queue and send those blocks to DN2 to update those in DN4. As well as move blocks in the acknowledgment queue. When all the acknowledgments for all the blocks are received, the blocks will be dequeued from the acknowledgment queue, and write complete message will be given to the name node.

VIII. MAPREDUCE ARCHITECTURE:

MapReduce is used for processing big data. MapReduce is the component that receives input data and is responsible for processing it as per the feature designed by the programmer. HDFS data node output is given as the input to the MapReduce. MapReduce works in two phases that are Map phase and Reduce phase.

As shown, MapReduce is internally divided into four parts which include splitting, mapping, shuffling, and reducing. MapReduce is a massively parallel processing application the input file is divided into small numbers of chunks of data file, this is called splitting. These data chunks are parallelly executed by the mapper and the output of the mapper is the intermediate key-value pair. Each value in the input data has been defined by a certain key in the mapper phase. The intermediate output is to be further given to the reducer concerning the keys for which the shuffle and sort operation is performed. Finally, the reducers join the final output and present it in the real world. Combiners can be also used in MapReduce. It is an optional phase and is also called a semi-reducer, by using combiner it will add all the values of the particular kays and give it to the reducer. The quality of data can be maintained in MapReduce by coping with node failure.

- If the master mapper fails, this will bring down the entire Map reducing process.
- If the mapper worker (slave) fails then this will be understood by the master mapper, which will migrate all the tasks of that mapper to another mapper and will also inform that reducer which takes input from the failed mapper.

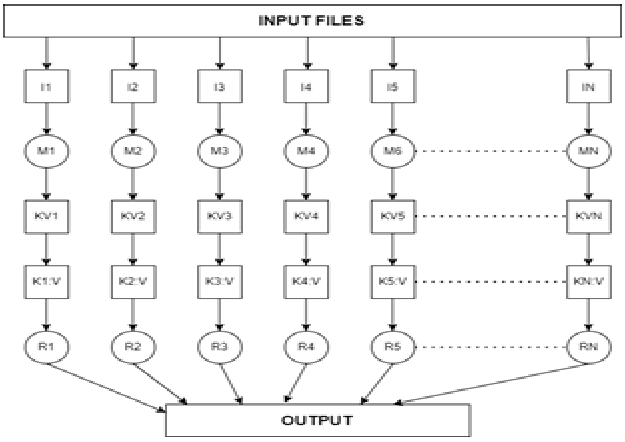

Fig 8.1 MapReduce Architecture

- If the reducer worker fails which will be only known to the master reducer, the task of the reducer will berescheduled on another reducer.

IX. IMPROVING THE QUALITY OF QUALITY OF DATA ON HADOOP:

Hadoop is one of the most important framework of the big data. Data quality on Hadoop is critical as the data is stored there. As the volume and variety of data on Hadoop increases, data management becomes a necessity.

Big data is receiving same treatment as relational data so that the quality of data is maintained.

Initially data quality on Hadoop data quality was overlooked. However not all Hadoop uses cases are for analytics.

Hadoop obviously shatters the limits of data storage, not only in terms of data volume and variety as well as in terms of structure. One way that data quality is maintained in a conventional data warehouse is by imposing strict limits on the volume, variety, and structure of data.

A Hadoop data quality tool is a data integration tool with data quality components and capabilities; it takes data from Hadoop, cleanses it and puts it back. This method involves a lot of performance overhead, but an off Hadoop tool makes sense if you’re moving data off your Hadoop cluster and into other data stores anyway.

Handling big data is to automate conventional data quality and processes to detect, correct data quality issues.

Big data quality issues can lead on to many inaccurate algorithms and various data quality issues which can lead to misleading real world system outcomes which can an affect many businesses users to trust data sets.

Some best practices to improve data quality is:

- **Make data accessible to all your users:**

Data should be accessible to all the types of users across the organization. The value of data can be realized when all the data is accessible to all the users.

- **Use the right data:**

The best data combines multiple sources to create a broad and complete view of the organization, so you can answer the most complex questions. By using different types of data so it will be beneficial for the users.

- **Secure your data:**

It is very important to secure your data so that the quality of data is not tampered with. To keep the data secure while allowing access to the organizations, so that it can handle vulnerabilities,encryption, and authentication as well as access and controls.

- **Treat data quality as a process:**

Creating value of data is more than just the insights. Businesses need to cleanse and maintain the data through their practices. Everything from data storage to application need to be streamlined managed and automated.

X. CONCLUSION:

Maintaining the quality of data is a need of any organization as it helps in a better decision-making process and also finding the value from the big data. The quality of data depends upon the data which is given as input in Hadoop. As using Hadoop, the limitation on the volume of the data is not the issue, it helps us to process big data with the help of distributed file system and reducing phase. A huge amount of data is generated every second which results in big data so by using the Hadoop platform there is the advantage of the cluster's massive parallel performance. In the Hadoop platform, the data is taken, stored, and reduce and store it back. It helps to maintain the quality of data, as it corrects, detects, and rectified data quality issues. The architecture of Hadoop is also capable of data cleaning.

REFERENCES

[1]. Jeffrey Dean and Sanjay Ghemawa, "MapReduce: Simplified Data Processing on Large Clusters", Communications of the ACM, vol. 51, Jan 2008.

[2]. iddaraju, C. L. Sowmya, K. Rashmi and M. Rahul, "Efficient Analysis of Big Data Using Map Reduce Framework", International Journal of Recent Development in Engineering and Technology, vol. 2, June 2014.

[3]. Dominique A. Heger, "Hadoop Design Architecture &MapReduce Performance"

[4]. AmolBansod, "Efficient Big Data Analysis with Apache Spark in HDFS", IJEAT, vol. 4, August 2015.

[5]. Harshwardhan S. Bhosale et al, “Review paper on Big Data using Hadood”, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

[6]. Poonam S. Patil et al. “Survey Paper on Big Data Processing and Hadoop Components” International Journal of Science and Research, Volume 3, Issue 10, October 2014.

[7]. Ms. GurpreetKaur,Ms. Manpreet Kaur, “Review Paper on Big Data using Hadoop”, International Journal of Computing Engineering and Technology, Volume 6, Issue 12, Dec 2015, pp. 65-71.

[8]. Kambatla, K., Kollias, G., Kumar, V., &Grama, A. (2014). Trends in big data analytics. Journal of parallel and distributed computing, 74(7), 2561-2573.

[9]. M. Sogodekar, S. Pandey, I. Tupkari and A. Manekar, "Big data analytics: hadoop and tools," 2016 IEEE Bombay Section Symposium (IBSS), Baramati, India, 2016, pp. 1-6, doi: 10.1109/IBSS.2016.7940204.

[10]. Sowmya R and Suneetha K R, "Data Mining with Big Data," 2017 11th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, 2017, pp. 246-250, doi: 10.1109/ISCO.2017.7855990.