International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST229656

A Study of Hadoop and Mapping Approach Techniques on Big Data Strategies

[1]Ms. Kamal Verma, [2]Prof. (Dr.) R. K. Bathla
[1]Ph.D Research Scholar, Computer Science and Applications, Desh Bhagat University, Punjab India
[2]Professor, Computer Science and Applications, Desh Bhagat University, Punjab India

Article Info
Volume 9, Issue 6
Page Number : 377-383

Publication Issue
November-December-2022

Article History
Accepted : 20 Nov 2022
Published : 07 Dec 2022

ABSTRACT

Research is an art of scientific examination. The advance learner's vocabulary of current English lays down the meaning of research as "A careful exploration and enquiry especially through search for new facts in any branch of knowledge. Bradman and Morry define research as "A standardize efforts to increase new knowledge". Research is, thus an original contribution to existing stock of knowledge making for its advancement. It is detection of truth with the help of study, observation, comparison, and experiments. The technologies that give support to the entire process of cost-effectively storing and processing data, and utilize internet technologies in a scattered way have arisen in the past few years. NoSQL and Cloud computing are the renowned ones that improve the potential offered by Big Data Technologies. Map Reduce is a software manufacture introduced by Google to act upon parallel processing on large datasets supercilious that large dataset storage is distributed over a large number of machines. Each machine computes data stored locally, which in turn contributes to distribute and parallel processing. This paper focuses on the Big data and Cloud services using impact of Map Reduce Algorithm and very advantageous for the researchers and corporate sectors who are using Map Reducing System technology.

Keywords :- Big Data, Cloud Services, Map Reduce Algorithm, Mapper, Reducer.

I. INTRODUCTION

Big data is a well-liked period used to describe the exponential expansion and ease of use of data, both structured and unstructured. Big data may be imperative to business – and society – as the Internet has become. Big Data is so outsized that it's easier said than done to process using long-established database and software techniques. More data may lead to more precise investigation. More accurate analyses may guide to more positive decision making. It better decisions can suggest superior outfitted efficiencies, cost decreases and reduced risk. Analyzing big data is one of the faces up for researchers system and academicians that needs extraordinary analyzing techniques. Big data analytics is the progression of

examining big data to uncover hidden patterns, unknown correlations and other useful in order that can be used to make better decisions. Big data analytics refers to the process of collecting, organize and analyzing large sets of data ("big data") to discover patterns and other useful information. Not only will big data analytics help you to comprehend the information controlled within the data, but it will also assist recognize the data that is the majority imperative to the business and future business decisions. Big data analysts on the whole want the acquaintance that comes from analyzing the data. HDFS, the Hadoop Distributed File System, is a distributed file system designed to run on article of trade hardware. It is encouraged by the Google File System. Hadoop is based on a simple data model, any data will fit. HDFS planned to hold especially large amounts of data (terabytes or petabytes or even zettabytes), and provide high-throughput right of entry to this information. Hadoop Map Reduce is a technique which analysis big data. Map Reduce has in recent times emerged as a new paradigm for large-scale data analysis due to its high scalability, fine-grained fault acceptance and easy training model. The term Map Reduce in point of fact refers two take apart and dissimilar tasks map and reduce that Hadoop programs perform.

The beginning of the digital age has led to a rise in different types of data with every passing day. In fact, it is projected that half of the total data will be on the cloud by around 2016. This data is multifaceted and needs to be stored, processed and analyzed for in sequence that can be used by organizations. Cloud computing provides an pertinent platform for big data analytics in view of the storage and computing requirements of the end. This makes cloud-based investigative a viable research field. However, several issues need to be addressed and risks need to be mitigated before practical applications of this synergistic model can be popularly used. This paper explores the existing research, challenges, open issues and future research direction for this field of study.[1],[2]

II. Hadoop Architecture

Big Data have plentiful other applications. Taking social network data analysis for example, massive amount of social network data is being created by Twitter, Facebook, LinkedIn and YouTube. These data expose numerous individual's characteristics and have been oppressed in various fields. In addition, the societal media and Internet enclose enormous amount of information on the end user first choices and confidences, most important economic indicators, business cycles, political attitudes, and the economic and social states of a society. It is predictable that the communal network data will persist to explode and be exploited for a lot of novel applications. More than a few other new applications that are becoming possible in the Big Data era include:

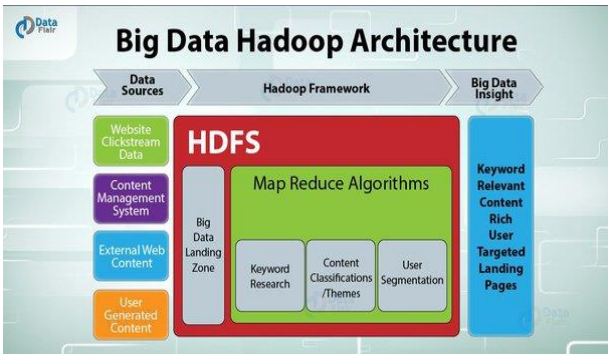

Fig:1 Big Data Hadoop Architecture

A. Personalized services.

With additional personal data collected, profit-making enterprises are able to provide personalized services adapt to individual preferences. For example, Target a retailing company in the United States of America is an intelligent to predict a customer's need by analyzing the collected operation records.

B. Internet security System.

When a set of connections-based attack takes place, historical data on network traffic may permit us to

professionally recognize the foundation and targets of the attack.

C. Personalized medicine.

More and more healthiness related metrics such as individual's molecular characteristics, human behavior, human habits and ecological factors are now obtainable. Using these pieces of information, it is potential to diagnose an individual's disease and select individualized treatments.

D. Digital humanities Right.

Now day's frequent archives are creature digitized. For example, Google has scanned millions of books and recognized about every word in every one of those books. This bring into being massive amount of data and make possible addressing topics in the humanities, such as mapping the transportation system in ancient Roman, visualizing the economic connections of ancient China, studying how natural languages evolve over time, or analyzing historical events.

III. Cloud Based Services Methodology

The cloud computing upbringing recommends development, setting up and implementation of software and data applications 'as a service'. Three multi-layered infrastructures namely, platform as a service (PaaS), software as a service (SaaS), and infrastructure as a service (IaaS), exist. Infrastructure-as-a-service is a model that make available computing and storage resources as a service. On the other hand, in case of PaaS and SaaS, the cloud services make available software platform or software itself as a service to its clients. The cost of storage space has significantly reduced with the advent of cloud-based solutions. In addition, the 'pay-as-you-go' model and the conception of product hardware allow effective and timely processing of large data, giving rise to the concept of 'big data as a service'. An example of one such podium is Google Big Query, which provides real-time impending from big data in the cloud environment Shakil, Sethi, and Alam make obvious the application of cloud for supervision of Big Data in educational institutions which special focus on University-level data. However, there have not been many practical applications of big data analytics that make use of the cloud. This has escort to an increasing shift of research focal point towards cloud-based big data analytics. A problem that is evident in this arrangement is information security and data privacy. As part of the cloud services, confidence in data is also defined as a service. There shall be a substantial decrease in trust in outlook of the fact that the chances of security breaches and privacy violation will significantly move up upon implementation of big data strategies in the cloud. In adding together, an extra important issue of possession and control will also exist.

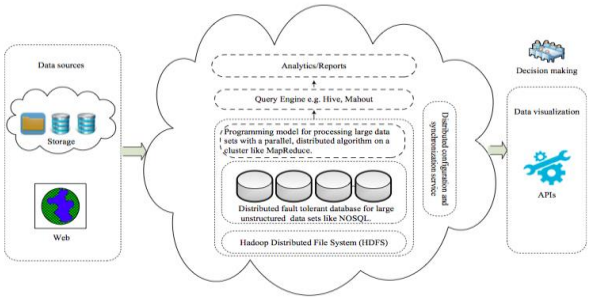

Fig:2- Make use of Cloud Computing in Big Data

IV.BIG DATA ANALYTICS

Big data analytics refers to the progression of collecting, organizing and analyzing large sets of data ("big data") to determine patterns and other constructive information. Not only will big data analytics help you to comprehend the in sequence restricted within the data, but it will also help make out the data that is most important to the business and future business decisions. Big data analysts fundamentally want the knowledge that comes from analyzing the data. Traditional data management tools and data dealing out or data mining techniques cannot be used for Big Data Analytics for the large volume and complexity of the datasets that it includes. Conventional commerce intelligence applications make use of methods, which are based on customary

analytics methods and techniques and make use of OLAP, BPM, Mining and database systems like RDBMS. It was in the 1980s that artificial intelligence-based algorithms were developed for data mining. One of the majority popular models used for data processing on cluster of computers is MapReduce. Jackson, Vijayakumar, Quadir and Bharathi provide a survey on the programming models that support big data analytics. It identifies Map Reduce and Hadoop as the most productive model for Big Data Analytics yet mentions that languages and addition like HiveQL, Latin and Pig have overpowering benefits for this use. Hadoop is simply an open-source implementation of the Map Reduce framework, which was in the beginning created as a distributed file system. According to Neaga and Hao the evolution of Hadoop as an absolute ecosystem or communications that works alongside Map Reduce components and includes a diversity of software systems like Hive and Pig languages, a harmonization service called Zookeeper and a distributed table store called HBase. For cloud-based big data analytics, several frameworks like Google Map Reduce, Spark, Haloop, Twister, Hadoop Reduce and Hadoop++ are available. These frameworks are used for storing and processing of data. In order to store this data, which may be of any structure, databases like HBase, BigTable and HadoopDB may be used. When it comes to data processing, the Pig and Hive technologies move toward into the picture.[17]

V. MAPREDUCE TECHNIQUE

Map Reduce is a framework for professionally processing the analysis of big data on a outsized number of servers. It was developed for the back end of Google's search engine to make possible a large number of article of trade servers to efficiently process the analysis of huge numbers of WebPages collected from all over the world. Apache developed an assignment to implement MapReduce, which was published as open source software (OSS), this enabled many organizations, such as businesses and universities, to tackle big data analysis. It was in the beginning developed by Google and built on well-known principles in parallel and distributed processing. Since then Map Reduce was broadly adopted for analyzing large data sets in its open source give flavor to Hadoop.[8]

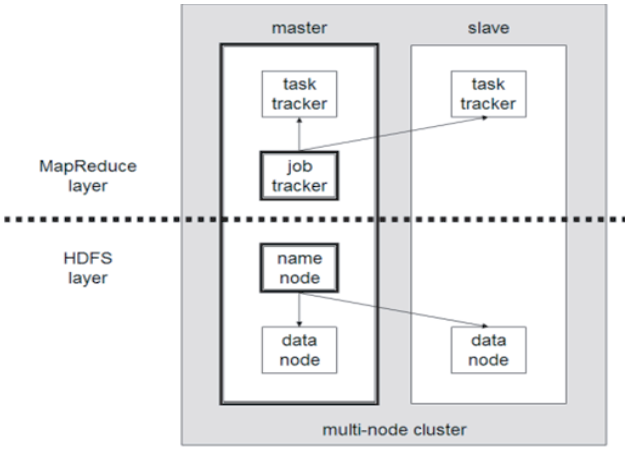

Figure 3 : Map Reduce Master/slave

Map Reduce is a straightforward programming model for processing massive data sets in parallel. Map Reduce have master/slave architecture this is shown in figure 2. The basic conception of Map Reduce is to split a task into subtasks, hold the sub-tasks in analogous, and aggregate the results of the subtasks to form the final output. Programs written in Map Reduce are robotically parallelized: programmers do not need to be concerned about the accomplishment details of parallel processing. As an alternate for, programmers write two functions: map and reduce. The map phase reads the input (in parallel) and distributes the data to the reducers. Auxiliary phases such as sorting, partitioning and combining values can also take place between the map and reduce phases.

Map Reduce programs are generally used to practice large files. The input and output for the map and reduce functions are expressed in the form of key-value pairs. A Hadoop Map Reduce program also has a element called the Driver. The driver is answerable for initializing the job with its arrangement details, specifying the mapper and the reducer classes for the

job, informing the Hadoop platform to perform the code on the specified input file(s) and controlling the location where the output files are placed.[14],[15]

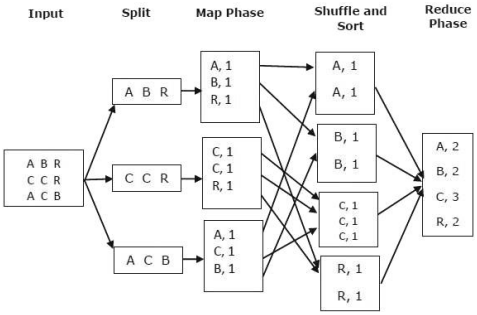

Fig 4 Map Reduce Example

VI. WORDS COUNT ALGORITHM OF USING MAP REDUCE

Algorithm 1. Word count

The mappers produce an intermediate key-value pair for each word in a document.

The reducer sums up all counts for each word.

1: class Mapper
2: method Map(docid a, doc x)
3: for all term t ∈ doc x do
4: produce (term t, count 1)
1: class Reducer
2: method Reduce (term t, counts [c1, c2 . . .])
3: sum ← 0
4: for all count c ∈ counts [c1, c2 . . .] do
5: sum ← sum + c
6: produce (term t, count sum)

Algorithm 2. Word count mapper using associative arrays

1: class Mapper
2: method Map (docid a, doc x)
3: R ← new Associative Array
4: for all term t ∈ doc x do
5: R{t} ← R{t} + 1 . Tally counts for entire document
6: for all term t ∈ R do
7: produce (term t, count R{t})

Algorithm 3. Word count mapper using the"in-mapper combining"

1: class Mapper
2: method Initialize
3: R ← new Associative Array
4: method Map (docid a, doc x)
5: for all term t ∈ doc x do
6: R{t} ← R{t} + 1 . Tally counts across documents
7: method close
8: for all term t ∈ R do
9: produce (term t, count R{t})

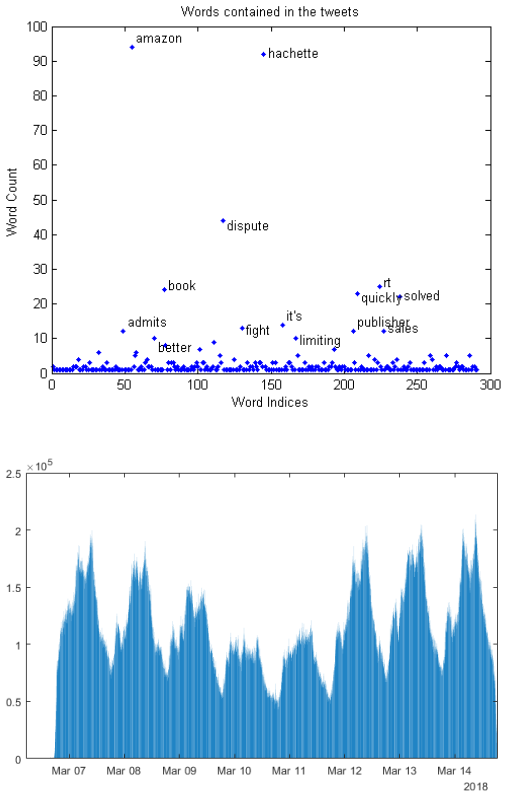

Fig 5.Map Reduce Technique to Words Count

VII. MAP REDUCE COMPONENTS

a. Name Node: Manages HDFS metadata, doesn't agreement with files directly.

b. Data Node: stores blocks of HDFS—defaults reproduction level for each block.

c. Job Tracker: schedules, distributes and observes job execution on slaves—Task Trackers.

d.Task Tracker: Runs Map Reduce operations[14[

VIII. REDUCE FRAMEWORK

Map Reduce is a software structure for scattered processing of outsized data sets on computer gathers. It is first developed by Google. Map Reduce is projected to make easy and simplify the processing of vast amounts of data in parallel on large clusters of article of trade hardware in a reliable, fault tolerant manner.

XI. HIVE

Hive is a distributed negotiator platform, a spread out system for building applications by networking local system resources. Apache Hive data warehousing module, an element of cloud-based Hadoop ecosystem which recommend a query language called Hive QL that translates SQL-like queries into Map Reduce jobs robotically. Applications of apache hive are SQL, oracle, IBM DB2.

X. NO-SQL

No-SQL database is an move toward to data management and data propose that's useful for very large sets of distributed data. These databases are in common part of the real-time events that are detected in process deployed to inbound channels but can also be seen as a facilitate technology subsequent analytical capabilities such as family member search applications developer in move forward. It is valuable when venture necessitate to access enormous quantity of unstructured data. There are more than one hundred No SQL approaches that concentrate in management of different multimodal data types (from structured to non-structured) and with the aspire to solve very specific challenges. Data Scientist, Researchers and Business Analysts in specific pay more concentration to agile come within reach of that leads to prior insights into the data

Sets that may be obscured or constrained with a more official development process.[12]

X.HPCC

HPCC is an unlock source platform used for computes and that make available the tune-up for managing of massive big data work flow. HPCC system is a single platform having a single architecture and a single programming language used for the data reproduction. HPCC system was premeditated to analyze the enormous amount of data for the purpose of solving multifaceted problem of big data. HPCC system is based on enterprise control language which has the declarative and on-procedural nature programming language the main components of HPCC are:

HPCC Data Refinery: Use parallel ETL engine by and large.

HPCC Data Delivery*:* It is massively based on structured query engine used. Enterprise Control Language deal outs the workload between the nodes in appropriate even load.

XI. CONCLUSION

Due to put in to in the capacity of data in the pasture of genomics, meteorology, biology, environmental research, it be converted into problematical to handle the data, to find organizations, patterns and to analyze the outsized data sets. In this paper we have also discussed the about the Big data (volume, variety, velocity, value, veracity) and Map reduce Technologies .This paper discussed an architecture using Map Reduce distributed data storage, real-time No SQL databases, and Map Reduce distributed data processing over a cluster of commodity servers. The commercial impacts of the Big data have the prospective to engender considerable productivity intensification for a number of vertical sectors. The main goal of our paper was to make a survey of various big data treatment techniques those handle a gigantic amount of data from different sources and improves overall presentation of systems. Growing talent and building teams to make analytic-based decisions is the key to realize the value of Big Data. As we have entered an era of Big Data, dealing out large

volumes of data has never been greater. Through better Big Data analysis tools like Map Reduce over Hadoop and HDFS, assurance faster advances in many scientific disciplines and improving the productivity and success of many enterprises. Map Reduce has received a lot of attentions in many fields, including data mining, information retrieval, image retrieval, machine learning, and pattern recognition. On the other hand, as the quantity of data that require to be processed grows, many data handing out methods have become not appropriate or imperfect.

III. REFERENCES

[1]. Hadoop, "Poweredby Hadoop,"http://wiki.apache.org/hadoop/PoweredBy.

[2]. Hadoop Tutorial,YahooInc., https://developer.yahoo.com/hadoop/tutorial/index.html

[3]. Apache: Apache Hadoop, http://hadoop.apache.org

[4]. Hadoop Distributed File System (HDFS), http://hortonworks.com/hadoop/hdfs/

[5]. Jianqing Fan1, Fang Han and Han Liu, Challenges of Big Data analysis, National Science Review Advance Access published February, 2014.

[6]. Haddop MapReduce-http://hadooptutorial.wikispaces.com/MapReduce

[7]. Amazon Simple Storage Service (Amazon S3). http://aws.amazon.com/s3/

[8]. Apache Hive, http://hive.apache.org/

[9]. Jens Dittrich JorgeArnulfo Quian´eRuiz, Efficient Big Data Processing in Hadoop MapReduce.

[10]. Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, Big Data Processing in Cloud Computing Environments, 2012 International Symposium on Pervasive Systems, Algorithms and Networks.

[11]. Y. Kim and K. Shim. TWITOBI: A recommendation system for twitter using probabilistic modeling. In ICDM, 2011.

[12]. Y. Kim and K. Shim. Parallel top-k similarity join algorithms using Map Reduce. In ICDE, 2012.

[13]. H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Chang. PFP: Parallel FP-Growth for query recommendation. ACM Recommender Systems, 2008.

[14]. A. Okcan and M. Riedewald. Processing theta-joins using MapReduce. In SIGMOD, 2011. [20] B. Panda, J. S. Herbach, S. Basu, and R. J. Bayardo. Planet: Massively parallel learning of tree ensembles with MapReduce. In VLDB, 2012.

[15]. K. Zhai, J. L. Boyd-Graber, N. Asadi, and M. L. Alkhouja. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In WWW, 2012.

[16]. "Big Data for Development: Challenges and Opportunities", Global Pulse, May 2012 Yuri Demchenko The Big Data Architecture Framework (BDAF) Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.

Cite this article as :

Ms. Kamal Verma, Prof. (Dr.) R. K. Bathla, "A Study of Hadoop and Mapping Approach Techniques on Big Data Strategies", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 9 Issue 6, pp. 377-383, November-December 2022. Available at doi : https://doi.org/10.32628/IJSRST229656
Journal URL : https://ijsrst.com/IJSRST229656