Study and Analysis of Big Data Analytics frameworks and challenges

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

SOFTWARE ENGINEERING

Submitted By

Saurabh Singh Shekhawat

2K21/SWE/23

Under the supervision of

Mr. Rahul

(Assistant Professor)

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, New Delhi - 110042

MAY, 2023

M.Tech (Software Engineering) Saurabh Singh Shekhawat 2023

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CANDIDATE'S DECLARATION

I, Saurabh Singh Shekhawat, Roll No. 2K21/SWE/23 student of M. Tech (Software Engineering), hereby declare that the project Dissertation titled **"Study and Analysis of Big Data Analytics frameworks and challenges"** which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Saurabh Singh Shekhawat

(2K21/SWE/23)

Place: Delhi

Date: 31/05/2023

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Study and Analysis of Big Data Analytics frameworks and challenges"** which is submitted by Saurabh Singh Shekhawat, 2K21/SWE/23 Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

31/5/22

Place: Delhi
Date: 31/05/2023

Mr. Rahul
(SUPERVISOR)
Assistant Professor
Department of Software Engineering
Delhi Technological University

ACKNOWLEDGMENT

The success of this project requires the assistance and input of numerous people and the organization. I am grateful to everyone who helped in shaping the result of the project.

I express my sincere thanks to my project manager **Mr. Rahul,** for providing me with the opportunity to undertake this project under her guidance. Her constant support and encouragement have made me realize that it is the process of learning which weighs more than the end result. I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout with new ideas, provided information necessary and pushed me to complete the work.

I also thank all my fellow students and my family for their continued support.

Saurabh Singh Shekhawat

(2K21/SWE/23)

ABSTRACT

Every day as we can see around us that data is generating exponentially. We are the reason the reason for that amount of data today, an individual generating on an average 40 Exabyte's of data daily. The data can be come from any sources like social media, online transactions, IOT's, digital media, records, different sensors etc. handling this huge amount of data nowadays becoming a challenging task. The data can be big or small in the size and can be of any form like unstructured, semi-structured or structured. We can't handle these amount of big data withthe traditional techniques.

Therefore, in order to handle such large amounts of unstructured data, we need methods and mechanisms that are simple to use, quick to process, and effective. The two main technological advancements that can manage any type of information are Hadoop and Spark. for storing, processing, and analysing the data, there are many tools and techniques are used in the Hadoop and spark. The Hadoop framework data is processing the data in distributing manner. The two basis elements of Hadoop are HDFS for storage, MapReduce and yarn for parallel processing in distributed manner, scheduling the data(tasks) and analalyzing the data. The second one spark uses resilient distribute data sets for fast processingfor overcome computational complexity. In this report we will see what is the Hadoop architecture, how it stores and process the data using MapReduce, how spark is better than Hadoop, how sparks done the job, what is the Apache spark technology and Hadoop and spark's comparative analysis.

Index Terms- Big data, Architecture, HDFS, Hadoop, Spark

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND NOMENCLATURE

1. HDFS- Hadoop distributed file system
2. SQL- Structured query language
3. RDBMS- Relational database management system
4. ML- Machine Learning
5. HQL-Hive Query Language

CHAPTER 1
INTRODUCTION

1.1 Overview

Large amount of information is generating all over the world daily. Mostly the data which is coming are the unstructured data. The world population is very high that's why the generation of data is also very huge and coming from many sources. So every day we need to store that data and analysis it after processing. Moreover, new advancements have been arisen to detonating volumes of complicated data, including web traffic, web-based media content, machine generated information, sensor information and system information . In the early days when internet become popular Hadoop used to store the data, managing it and process on large data sets. Hadoop MapReduce has come up as a exceptionally powerful and productive tool for examination Big Data. MapReduce is likewise great for checking authenticinformation and performing analytics [1].

But MapReduce was slow to give response on large set of data, so for that type of unstructured data the processing is done by spark which gives fast results and utilised extensively in large data. Spark was created in 2009, and in 2010 it will be made available to everyone as open source.. Apache spark is done sophisticated analytics in very fast speed. nowadays, as associations confronting the development need for ongoing information investigation to accomplish advantage, another open source engine, Apache Spark, has entered in the field.

1.2 What is Big Data

Big data is nothing but the data which is huge and other term we can say the small data which is bigger in size.

Here the aim is to solve the new problem or older one in the efficient and effective way. There are five v's of big data that are-

- Volume

- Velocity
- Variety
- Value
- Veracity

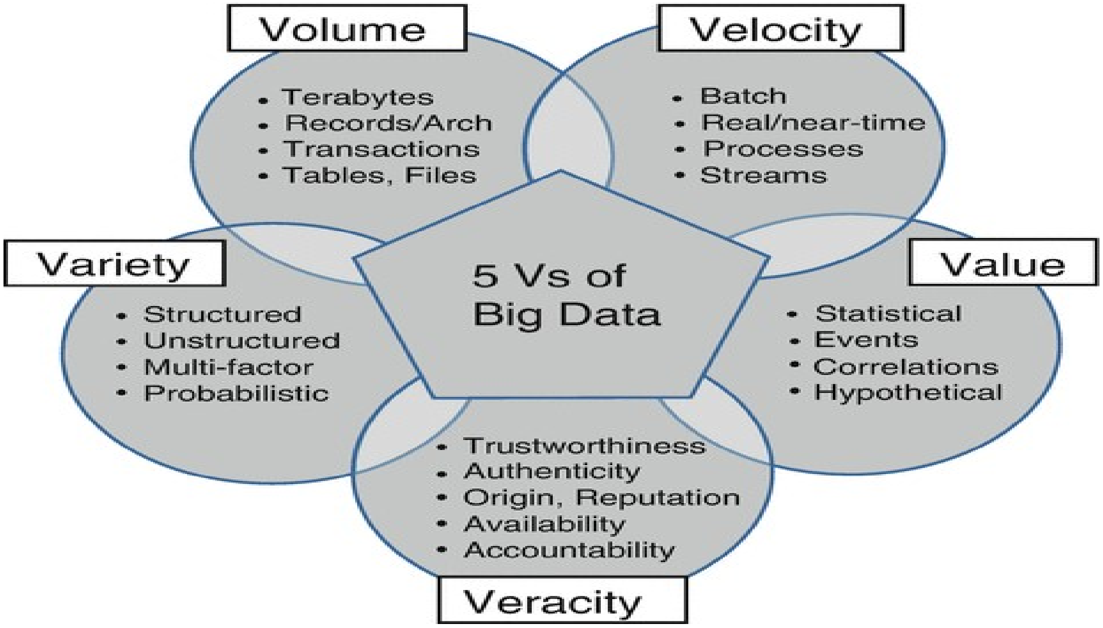


Fig 1.1 : 5 v's of big data [2]

Variety means Big data is not only the numbers, string or dates. It is also the media, audio, video, 3-D data, unstructured text etc. so here are varies type of data in the market. Traditionaldata base system can handle only small set of structured data, consist able data.

Volume means the amount of data which is generated. Data can be from online or offline transections. The data is saved in tables, records or files. A commonplace PC may have had 10 gigabytes of capacity in 2000.

Velocity denotes the speed of generating the data, means how frequently data has been generating. Data generated in streams, batches or bits and it can in both in online and offline mode. [2]

Veracity means the is the coming data is truthful and worthiness or not. Today data is such a large that the difficult control the quality of the big data and accuracy of it [2].

1.3 Big Data Analytics

In order to find rediscovered patterns, correlations, and other priceless insights, enormous and complicated datasets—often referred to as "big data"—are examined and assessed. It entails using creative approaches, algorithms, and technologies to extract valuable knowledge and information from enormous amounts of data.

Big data analytics has grown in value across a variety of industries, helping businesses to make informed decisions that will improve operations and keep them competitive in the data-driven world of today. It has uses in a variety of fields, including Analytics for business and marketing like organizations may better understand consumer preferences and purchasing trends with the use of big data analytics. This enables personalised advice, targeted marketing campaigns, pricing optimisation, demand forecasting, and market trend research. Other one may be in Healthcare: sector by enhancing patient care, streamlining hospital operations, and advancing medical research, big data analytics has completely changed the healthcare industry . There are many more sectors like social media, Cyber security, business management etc.

1.4 Frameworks of Big Data

There are a number of well-liked frameworks for working with big data that offer the infrastructure and resources needed to manage and handle enormous datasets. Some of the well-known frameworks employed in the big data ecosystem include the following:

One of the most well-known frameworks for massive data processing is Apache Hadoop. The Hadoop Distributed File System (HDFS) for distributed storage and the MapReduce programming language for distributed processing make up its two main parts. Massive datasets can be processed using Hadoop across a cluster of commodity hardware in a fault-tolerant and scalable manner [4].

MapReduce can be replaced by Apache Spark, an open-source distributed computing platform that is quicker and more adaptable. Various data processing activities, such as batch processing, real-time streaming, and machine learning are supported by Spark.

1.5 Challenges of Big Data

While big data analytics brings many advantages, it also has a number of problems that businesses must solve. These difficulties include massive volumes of data that are beyond the capacity of conventional data processing technologies are dealt with by big data analytics. Scalable infrastructure and effective data storage and retrieval methods are needed for managing, storing, and processing such massive datasets. Big data is produced quickly, frequently in real-time or very close to real-time. Streaming data analysis and fast insight extraction necessitate sophisticated methods. There are some challenges as follows

- Data Integrity: Big data may contain biases, inaccuracies, and other types of poor data quality. To make wise decisions, data accuracy and dependability must be guaranteed.
- Scalability: Businesses need to make sure that their big data analytics infrastructure can grow as data volumes rise. To managData Integrity: Big data may contain biases, inaccuracies, and other types of poor data quality
- To make wise decisions, data accuracy and dependability must be guaranteed.
- To manage the growing workload, scaling horizontally by introducing more computing power and dividing data processing over numerous nodes is essential.
- Scalability: As data volumes rise, businesses must make sure their big data analytics infrastructure can grow with them. To managData Integrity: Big data may contain biases, inaccuracies, and other types of poor data quality.

Data security and privacy Working with private and sensitive data is a common part of big data analysis. To guard against unauthorised access, breaches, and misuse of data, organisations must handle privacy and security issues. In such circumstances, adherence to data protection standards becomes essential.

It takes a combination of technical solutions, organisational tactics, and qualified employees to overcome these obstacles. To effectively utilise big data analytics, organisations must invest in solid infrastructure, adopt adequate data management and processing frameworks, establish data governance practises, and promote a data-driven culture [5].

CHAPTER 2

LITERATURE REVIEW

2.1 Hadoop

Hadoop is open-source for all, it is the execution of MapReduce tool and generally utilized for lot of data handling. Hadoop has mainly two components one is HDFS and another one is map reduce. Hadoop is developed by the doug cutting in 2006 after the google published the paper on GFS that how to process the data using GFS map reduce . Hadoop is open source. It takes more time then spark in real time. It is purely disk type storage means every time when it performs the task it will touch the disk.

So what is Hadoop?

- It is not the big data
- Hadoop is not a database
- Hadoop is only a stage or structure which permits client to compose and test appropriated framework and Hadoop is additionally exceptionally productive to naturally disseminate the information and assignments across the machines.

There are a few new executions of Hadoop to conquer its presentation issues, for example, gradualness to stack information and the absence of reuse of information. For example, Starfishis a Hadoop-based system, which meant to work on the exhibition of MapReduce occupations utilizing information lifecycle in examination. It likewise utilizes work profiling and work process enhancement to lessen the effect of unbalance information during the work execution. Starfish is a self-tuning framework dependent on client prerequisites and framework responsibilities with practically no need from clients to design or change the settings or boundaries [6]. In addition, Starfish's Elasticized can mechanize the decision making for making advanced Hadoop groups utilizing a blend of recreation and model-based assessment to track down the most appropriate responses for consider the possibility that inquiries concerning responsibility execution.

2.2 HDFS

HDFS is a dispersed document framework which was developed to run on product equipment. HDFS is similar to DFS in which also data stores in distributed fashion and one node taken care of all the distributions. There is one gateway node which is responsible for the mapping of data distributed and MapReduce process done on top of it for process the data. Hadoop cluster or room created to store Hadoop room .

2.2.1 Hadoop Cluster Creation

Cluster means group of nodes, laptops, virtual machines. Every laptop which have Hadoop cluster will have an “Edge Node”.

We can get IP address of edge node. We need IP address to communicate the edge node. There are total 4 nodes in Hadoop cluster, they are-

- Edge node
- Name node
- Secondary Name node
- Data nodes

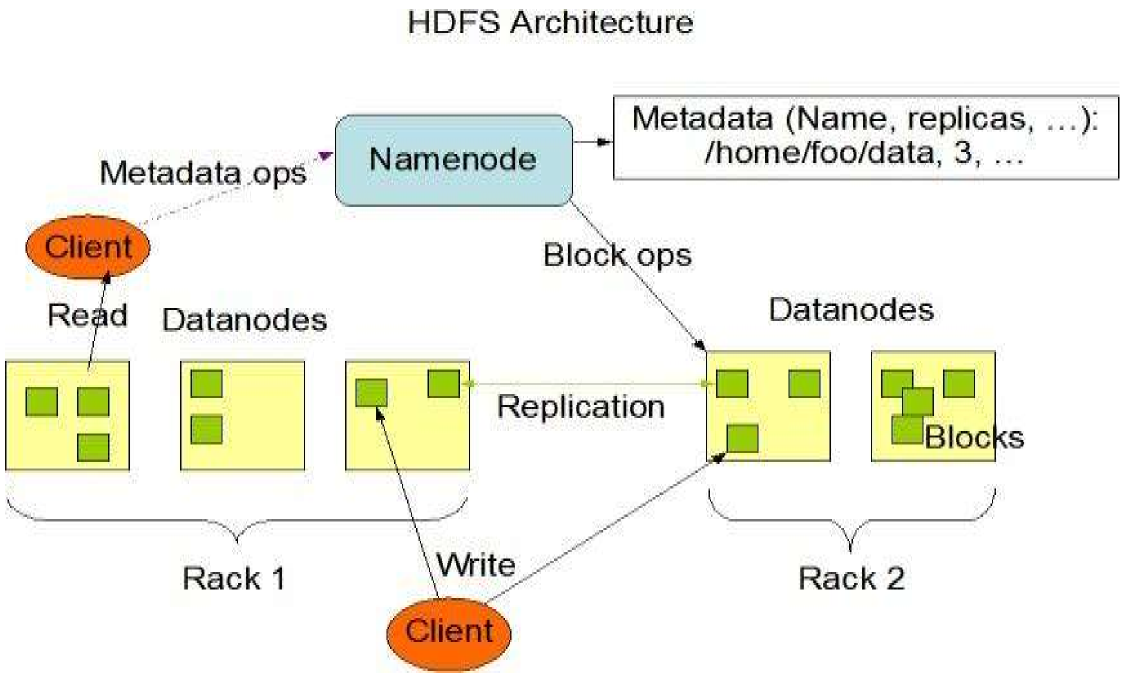


Fig 2.1 : HDFS architecture [7]

With the putty tool anyone can communicate to edge node. We can have many number of edge node for same cluster. To talk to the Hadoop nodes user should communicate to edge node in Hadoop language, so that edge node will pass that inside the room. Ambari on Microsoft azure will show you the cluster details. every node has IP address/host name, ram and hard disk.

2.3 MapReduce

It is the processing the big data using map () and reduce () functions. It works on the data locality means process will go towards to data and process it, early model DFS fails to process the data efficiently because DFS don't use data locality.

MapReduce …... Data locality   Distributed process

Big data ordinarily put away in a huge number of ware servers so conventional programming models, for example, message passing interface (MPI) can't deal with them successfully. Along these lines, new equal programming models are used to work on the exhibition of NoSQL information bases in datacenter's. MapReduce is proposed by the google. These coverings can give a superior command over the MapReduce code and help in the source code improvement [8].

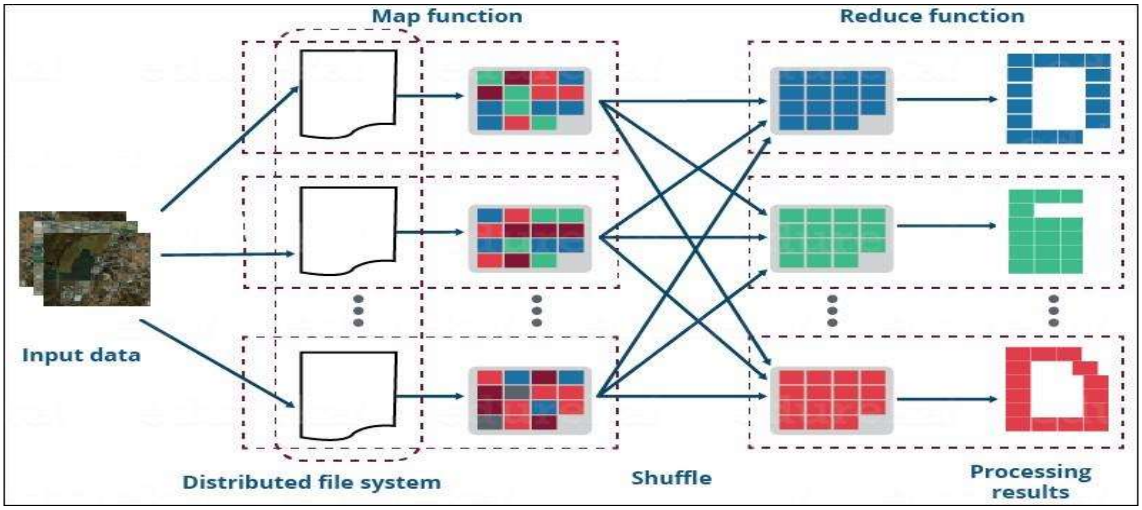


Fig 2.2 : Map and Reduce phases[8]

As we can see in above figure there are an Edge node or master node which split the data to the all name nodes or slave nodes in a distributed manner. Now when we processed on the stored data then with the help of data locality (when the process is going toward the data) at each node processed result will have generated. This is called Map phase. After the processeddata now we need to combine them at one node so master decides to store the processed data in selected one name node and there all the data gets merged, and this process is call Reduce.

2.4 Data Analysis Process

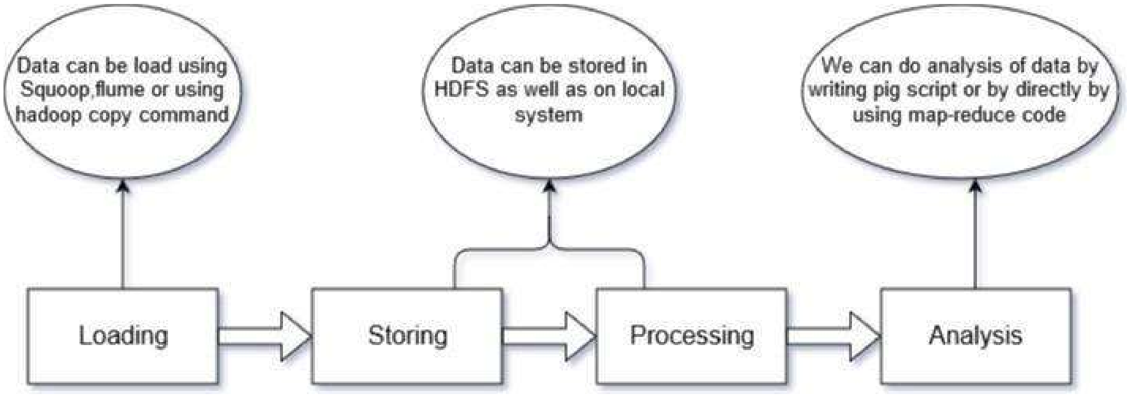

Data can be load using Squoop,flume or using hadoop copy command
Data can be stored in HDFS as well as on local system
We can do analysis of data by writing pig script or by directly by using map-reduce code
Loading
Storing
Processing
Analysis

Fig 2.3 : Data analysis process [9]

As above figure shows generally cycle to dissect any sort of information in hadoop environment. At first information can be anyplace however we need to stack it into HDFS or neighborhood document framework which we should be possible utilizing flume, Sqoop, or Hadoop commands. When information get stacked, it is put away and handled in HDFS or nearby document framework. To examine this store information we can use Pig or MapReduce [9]. overall process that how any kind of data can be analyze in Hadoop framework document framework with the assistance of instruments like sqoop, Hadoop order or flume. When the information stacked, it will be put away and handled on the neighborhood document framework or HDFS. For dissect the handled information we can utilize MapReduce or pig innovation to investigate this put away information.

2.5 Sqoop

Sqoop tool is nothing but the combination of SQL and the Hadoop. Scoop provides the faster imports like multi-threading imports. Sqoop is basically used to GET the data like import and export. In 2009 Sqoop is introduced by the Apache foundation. Sqoop is nothing but java commands. Sqoop meta store help with import and export with RDBMS. It helps with incremental data imports and capture the changed data also. Semi structured imports and controlling imports also done by sqoop [9] .

2.6 Hive

Hive is a data processing tool. It is like a data warehouse that allows users to interact with HDFS. Initially in Utilising Hadoop MapReduce, the data was processed. That was based on java code framework and MapReduce java jar deployment was very hard to do Facebook created Hive in 2008, but the Apache Foundation eventually acquired it and made it available as open source in 2010 as Apache hive.

Hive is like a SQL layer runs on top of HDFS which triggers MapReduce at the back end to perform distributed processing on the HDFS data. Hive doesn’t required java or understanding MapReduce framework also. Hive uses HQL (hive query language) that are similar to SQL but not SQL for querying the data. The purpose of HQL is ELT (Extract Load Transform) that means first extract the raw data after that loading has been done and then transform it. But in the SQL, the purpose of SQL is ETL (Extract Transform Load). SQL works on load time parsing means at the loading phase data type validation has been done but HQL works on Query time parsing. HQL supports structured and semi-structured data and it is good for large data, it is not good for small data. For small data SQL is preferable .

CHAPTER 3

SPARK

Apache spark is the very fast cluster computing technology. The significant element of Spark that makes it one of a kind is its capacity to perform in memory calculations. It permits the information to be reserved in memory, in this way dispensing with the Hadoop's plate overhead limit for iterative assignments [10].

Spark introduced RDD's which is its fundamental architecture, RDD is the Resilient Distributed Dataset.

- Resilient – When memory data is lost, it can be regenerated.
- Distributed – The information is distributed throughout a cluster.
- Dataset- starting file can be generated either manually or automatically from a file.

In Spark data is distributed in the memory as in Hadoop firstly we need to store it then can distribute it. Spark is very frequently using these days in the organizations to process the data. Spark is RAM memory type computing means it does not use disk frequently, it uses it partially. Spark map reduce is similar to the Hadoop map reduce but it is 100 times faster than Hadoop map reduce .

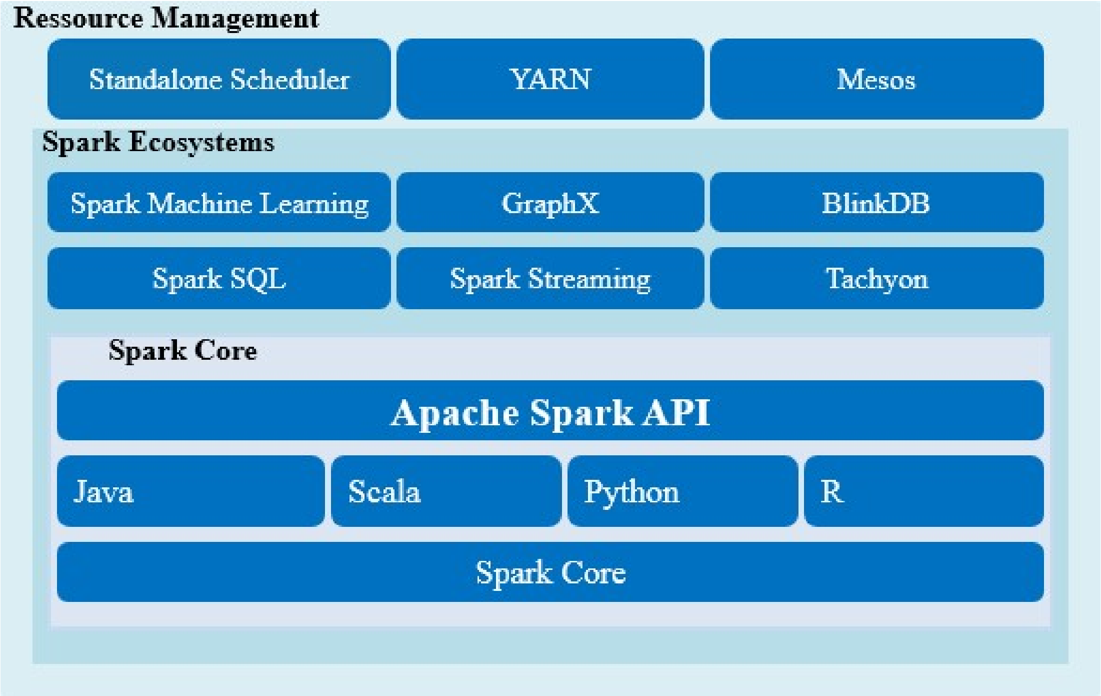


Fig 3: Apache spark architecture[10]

Essentially, there are other proposed procedures for profiling of MapReduce applications to track down potential bottlenecks and re-enact different situations for execution examination of the adjusted applications. This pattern uncovers that utilizing basic Hadoop arrangement would not be effective for huge information examination, and new devices and procedures to mechanize provisioning choices ought to be planned and created. This can be another help (i.e., large information investigation as-a-administration) that ought to be given by the Cloud suppliers to programmed huge information examination on datacenter's. Not with standing MapReduce, there are other existing programming models that can be utilized for large information handling in datacenter's, for example, Dryad and Pregl [12]. Dryad is a disseminated execution motor to run enormous information applications as coordinated non- cyclic chart (DAG).

Activity in the vertexes will be run in bunches where information will be moved utilizing information channels including archives, transmission control convention (TCP) associations,and shared memory. Additionally, any sort of information can be straightforwardly moved between hubs. While MapReduce just help single info and yield set, clients can utilize quite a few information and yield information in Dryad. Pregl is utilized by Google to deal with hugescope diagrams for different purposes, for example, investigation of organization charts and long range interpersonal communication administrations. Applications are acquainted as coordinated diagrams with Pregl where every vertex is modifiable, and client characterized worth and edge show the source and objective vertexes.

CHAPTER 4

COMPARATIVE ANALYSIS OF FRAMEWORKS

4.1 RDBMS vs Hadoop

Before Hadoop, RDBMS used for the data storing and processing, but unstructured data is the problem in RDBMS. Hadoop can process both structured and unstructured data. Fault tolerance in Hadoop is the main advantage compare to RDBMS. Hadoop is the highly scalable framework compared [14] to traditional databases system. Relational DB used to most popular on early 80s in rdbms information is stored in row and Colum fashion in table format.And to process the data mostly use SQL for query and writing the data. we know that data becomes increased so DBMS also not in the situation to help you. That time we got internet into the market, internet got emerging in 1980s.

Because of internet explosion in 1980s, data analytics emerging a new tool as RDBMS. Remotely we started to accessing the data, it is not necessary to save the data on my personal computer. While by all account not the only proposed game around for this sort of inquiry, SQL came to win out. Unquestionably, many variables contributed. Be that as it may, an ostensibly executioner component of SQL was its explanatory nature. Application developers need just indicate which records they need rather than how to recover them. The "how" turned into an execution detail of the RDBMS.

In Hadoop at whatever point the information is ship off the hub then that is repeated over the all hubs in the bunch that implies in the event of disappointment of any hub happened than the information won't be lost in light of the fact that the duplicate of the information accessible to the another hub. So to deal with huge information Hadoop enjoys an upper hand over the customary information bases to wipe out adaptation to non-critical failure .

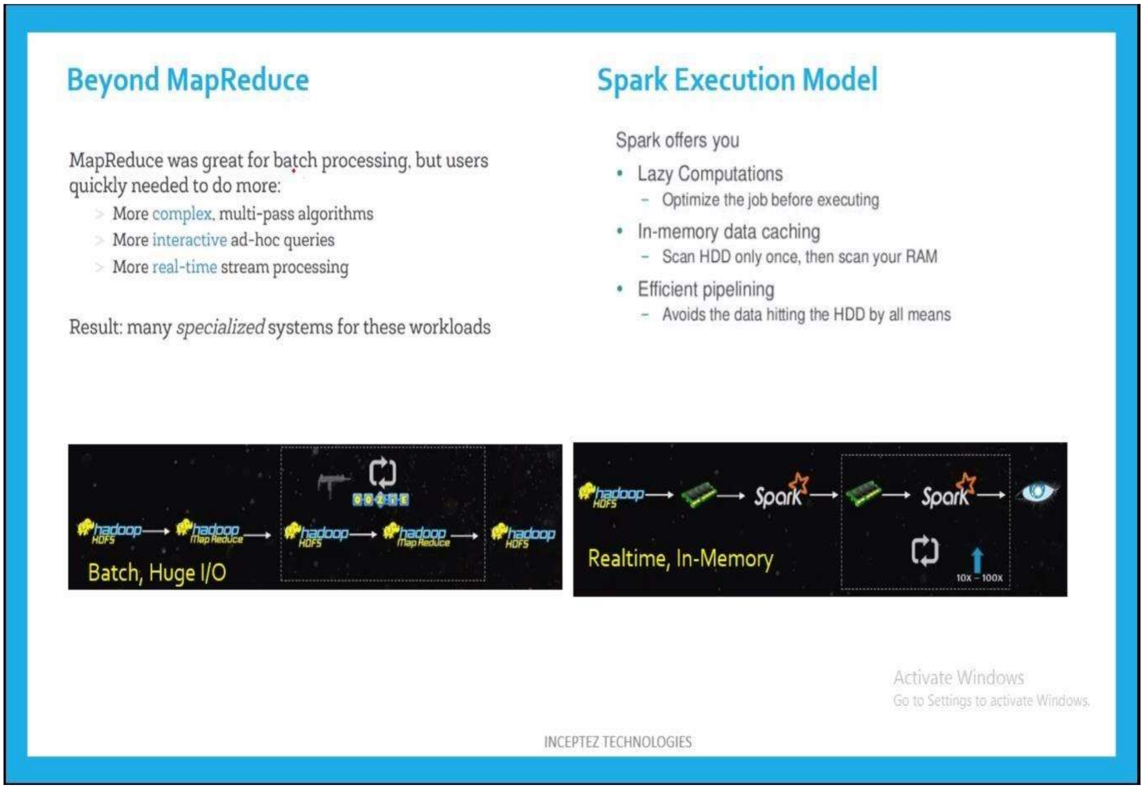
Table 4.1 Hadoop vs RDBMS

S.no	Hadoop	RDBMS
1	The type of data can be anything processed by Hadooop like unstructured, semi structured or structured	In RDBMS mostly structured data processed
2	Here processing is coupled with the data storage	Mainly data storage
3	It is open source	For software license cost required
4	Ideal for massive data processing or storage	ideal for an OLTP environment
5	Low integrity	Follows ACID properties that’s why integrity is high
6	provides greater throughput than RDBMS	Provides less throughput

4.2 Hadoop vs Spark

Apache spark done the cluster handling faster than the Hadoop MapReduce and the speed is close to 10 to 100 times, the main thing is that Hadoop MapReduce's number of read and writes to the hard disk is more [21] .

Spark offers lazy computation means it optimizes the job before execution but in Hadoop is optimize after the execution. In spark processed result after every iteration will save in RAM parallel but in Hadoop is store in disk file that's why Hadoop is 10 to 100 times slower than Spark.



Beyond MapReduce

MapReduce was great for batch processing, but users quickly needed to do more:

- More complex, multi-pass algorithms
- More interactive ad-hoc queries
- More real-time stream processing

Result: many *specialized* systems for these workloads

Spark Execution Model

Spark offers you

- Lazy Computations
  - Optimize the job before executing
- In-memory data caching
  - Scan HDD only once, then scan your RAM
- Efficient pipelining
  - Avoids the data hitting the HDD by all means

INCEPTEZ TECHNOLOGIES

Fig 4.1: Spark vs Hadoop frameworks [21]

.

Table 4.2 Spark vs Hadoop

Parameter	Spark	Hadoop
Storage of Data	In spark Data stored in memory	Stores on the disk
Fault tolerance	Fault tolerance occurs due to resilient distributed datasets	For the fault tolerance it uses replication
Line of code	The project apache spark contains 20,000 lines of code.	The Hadoop 2.0 contains more than one lakh line of code
Speed	Due to in memory calculations, it is quicker.	Slower than spark
OS support	Windows, Linux, Mac	Linux
Streaming data	Can be used for modification of real time data as well as process this data	With the help of MapReduce we can processing a group of saved data
Security	Less secure	More secure because of ACL's

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

In this paper we different kind of Hadoop technologies, how Hadoop is used for the data processing and storage, how spark is different from Hadoop, why spark is faster than Hadoop framework. We studied that MapReduce is most efficient than traditional techniques. In the word count problem, we saw that how Spark done the job in less time compare to Hadoop. Because in spark data is distributed in memory and in Hadoop distribute the data when its stored and Spark store it in RAM instead of disk so that's all the reason why Spark is faster.

Many of the studies shows that the data in future will generated so huge and exponentially. Recent research published that by the year of 2025 the big data will reach to the 175 zettabytes. Future work will be definitely performed on the distributed networks, because in coming years'cloud computing will also use by drastically. New tools and frameworks always be need to thebetter change on existing

REFERENCES

[1] M. M. Rathore, S. A. Shah, D. Shukla, E. Bentafat, and S. Bakiras, “The Role of AI, Machine Learning, and Big Data in Digital Twinning: A Systematic Literature Review, Challenges, and Opportunities,” *IEEE Access*, vol. 9, pp. 32030–32052, 2021, doi: 10.1109/ACCESS.2021.3060863.

[2] J. Liu, F. Liu, and N. Ansari, “Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop,” *IEEE Netw.*, vol. 28, no. 4, pp. 32–39, 2014, doi: 10.1109/MNET.2014.6863129.

[3] A. Jain and V. Bhatnagar, “Crime Data Analysis Using Pig with Hadoop,” *Phys. Procedia*, vol. 78, no. December 2015, pp. 571–578, 2016, doi: 10.1016/j.procs.2016.02.104.

[4] S. R. Salkuti, “A survey of big data and machine learning,” *Int. J. Electr. Comput. Eng.*, vol. 10, no. 1, pp. 575–580, 2020, doi: 10.11591/ijece.v10i1.pp575-580.

[5] A. W. Khan *et al.*, “Analyzing and Evaluating Critical Challenges and Practices for Software Vendor Organizations to Secure Big Data on Cloud Computing: An AHP-Based Systematic Approach,” *IEEE Access*, vol. 9, pp. 107309–107332, 2021, doi: 10.1109/ACCESS.2021.3100287.

[6] M. Humayun, “Role of emerging IoT big data and cloud computing for real time application,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 494–506, 2020, doi: 10.14569/IJACSA.2020.0110466.

[7] M. Supriya and A. Deepa, “Machine learning approach on healthcare big data: a review,” *Big Data Inf. Anal.*, vol. 5, no. 1, pp. 58–75, 2020, doi: 10.3934/bdia.2020005.

[8] A. Wakde, P. Shende, S. Waydande, S. Uttarwar, and G. Deshmukh, “Comparative Analysis of Hadoop Tools and Spark Technology,” *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–4, 2018, doi: 10.1109/ICCUBEA.2018.8697577.

[9] P. P. Talan, K. U. Sharma, P. P. Nawade, and K. P. Talan, *An Overview of Hadoop MapReduce, Spark, and Scalable Graph Processing Architecture*, vol. 740. Springer

Singapore, 2019. doi: 10.1007/978-981-13-1280-9_3.

[10] P. P. Talan and K. U. Sharma, “An overview and an Approach for Graph Data Processing using Hadoop MapReduce,” *Proc. 2nd Int. Conf. Comput. Methodol. Commun. ICCMC 2018*, no. Iccmc, pp. 59–63, 2018, doi: 10.1109/ICCMC.2018.8487681.

[11] A. V. Hazarika, G. Jagadeesh Sai Raghu Ram, and E. Jain, “Performance comparision of Hadoop and spark engine,” *Proc. Int. Conf. IoT Soc. Mobile, Anal. Cloud, I-SMAC 2017*, pp. 671–674, 2017, doi: 10.1109/I-SMAC.2017.8058263.

[12] S. Neethirajan, “The role of sensors, big data and machine learning in modern animal farming,” *Sens. Bio-Sensing Res.*, vol. 29, no. July, p. 100367, 2020, doi: 10.1016/j.sbsr.2020.100367.

[13] K. Singh and R. Kaur, “Hadoop: Addressing challenges of Big Data,” *Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014*, pp. 686–689, 2014, doi: 10.1109/IAdCC.2014.6779407.

[14] P. J. Charles, S. T. Bharathi, and V. Susmitha, “BIG DATA – CONCEPTS , ANALYTICS , ARCHITECTURES – OVERVIEW,” pp. 125–129, 2018.

[15] M. Liroz-Gistau, R. Akbarinia, D. Agrawal, and P. Valduriez, “FP-Hadoop: Efficient processing of skewed MapReduce jobs,” *Inf. Syst.*, vol. 60, pp. 69–84, 2016, doi: 10.1016/j.is.2016.03.008.

[16] A. Verma, A. H. Mansuri, and N. Jain, “Big data management processing with Hadoop MapReduce and spark technology: A comparison,” *2016 Symp. Colossal Data Anal. Networking, CDAN 2016*, 2016, doi: 10.1109/CDAN.2016.7570891.

[17] K. El Bouchefry and R. S. de Souza, *Learning in Big Data: Introduction to Machine Learning*. Elsevier Inc., 2020. doi: 10.1016/B978-0-12-819154-5.00023-0.

[18] R. Marlow *et al.*, “A phase III, open-label, randomised multicentre study to evaluate the immunogenicity and safety of a booster dose of two different reduced antigen diphtheria-tetanus-acellular pertussis-polio vaccines, when co-administered with measles-mumps-rubella vaccine in 3 and 4-year-old healthy children in the UK,” *Vaccine*, vol. 36, no. 17, pp. 2300–2306, 2018, doi: 10.1016/j.vaccine.2018.03.021.

[19] A. K. Sandhu, "Big Data with Cloud Computing: Discussions and Challenges," *Big Data Min. Anal.*, vol. 5, no. 1, pp. 32–40, 2022, doi: 10.26599/BDMA.2021.9020016.

[20] G. T. Reddy *et al.*, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2020.2980942.

[21] Z. Lu, "Computational discovery of energy materials in the era of big data and machine learning: A critical review," *Mater. Reports Energy*, vol. 1, no. 3, p. 100047, 2021, doi: 10.1016/j.matre.2021.100047.

[22] A. Singh, M. Mittal, and N. Kapoor, *Data Processing Framework Using Apache and Spark Technologies in Big Data*, vol. 43. Springer Singapore, 2019. doi: 10.1007/978-981-13-0550-4_5.

[23] I. Chebbi, W. Boulila, N. Mellouli, M. Lamolle, and I. R. Farah, "A comparison of big remote sensing data processing with Hadoop MapReduce and Spark," *2018 4th Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2018*, pp. 1–4, 2018, doi: 10.1109/ATSIP.2018.8364497.

[24] B. Hosseini and K. Kiani, "A big data driven distributed density based hesitant fuzzy clustering using Apache spark with application to gene expression microarray," *Eng. Appl. Artif. Intell.*, vol. 79, no. January 2018, pp. 100–113, 2019, doi: 10.1016/j.engappai.2019.01.006.

[25] Shyam R., B. Ganesh H.B., S. Kumar S., P. Poornachandran, and Soman K.P., "Apache Spark a Big Data Analytics Platform for Smart Grid," *Procedia Technol.*, vol. 21, pp. 171–178, 2015, doi: 10.1016/j.protcy.2015.10.085.

[26] Y. Cui, S. Kara, and K. C. Chan, "Manufacturing big data ecosystem: A systematic literature review," *Robot. Comput. Integr. Manuf.*, vol. 62, no. January 2019, p. 101861, 2020, doi: 10.1016/j.rcim.2019.101861.

[27] N. L. Bragazzi, H. Dai, G. Damiani, M. Behzadifar, M. Martini, and J. Wu, "How big data and artificial intelligence can help better manage the covid-19 pandemic," *Int. J. Environ. Res. Public Health*, vol. 17, no. 9, pp. 4–11, 2020, doi: 10.3390/ijerph17093176.

[28] M. Assefi, E. Behravesh, G. Liu, and A. P. Tafti, "Big data machine learning using apache

spark MLlib," *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-January, pp. 3492–3498, 2017, doi: 10.1109/BigData.2017.8258338.

[29] D. Saidulu and R. Sasikala, "Machine learning and statistical approaches for big data: Issues, challenges and research directions," *Int. J. Appl. Eng. Res.*, vol. 12, no. 21, pp. 11691–11699, 2017.

[30] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0151-6.

[31] Y. Xu, H. Liu, and Z. Long, "A distributed computing framework for wind speed big data forecasting on Apache Spark," *Sustain. Energy Technol. Assessments*, vol. 37, no. November 2019, p. 100582, 2020, doi: 10.1016/j.seta.2019.100582.

[32] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang, "Big data analytics on Apache Spark," *Int. J. Data Sci. Anal.*, vol. 1, no. 3–4, pp. 145–164, 2016, doi: 10.1007/s41060-016-0027-9.

[33] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, 2014, doi: 10.1007/s11036-013-0489-0.

[34] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0206-3.

[35] J. Camacho-Rodríguez *et al.*, "Apache hive: From mapreduce to enterprise-grade big data warehousing," *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 1773–1786, 2019, doi: 10.1145/3299869.3314045.

[36] M. Juez-Gil, Á. Arnaiz-González, J. J. Rodríguez, C. López-Nozal, and C. García-Osorio, "Approx-SMOTE: Fast SMOTE for Big Data on Apache Spark," *Neurocomputing*, vol. 464, pp. 432–437, 2021, doi: 10.1016/j.neucom.2021.08.086.

[37] L. Zhao, H. L. Ciallella, L. M. Aleksunes, and H. Zhu, "Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling," *Drug Discov. Today*, vol. 25, no. 9, pp. 1624–1638, 2020, doi: 10.1016/j.drudis.2020.07.005.

[38] D. E. O'Leary, "Artificial Intelligence and Big Data," in IEEE Intelligent Systems, vol. 28, no. 2, pp. 96-99, March-April 2013, doi: 10.1109/MIS.2013.39.