Shape

Description automatically generated with medium confidence



**Essay / Assignment Title:** **Advancing Insights in Big Data Analytics**

**Programme title:MSc Data Analytics**

**Name: Turan Can Gün (Q1055616)**

**Year:2024 (2023 October Intake)**

**CONTENTS**

[INTRODUCTION 3](#_Toc577863401)

[1. INTRODUCTION to BIG DATA 4](#_Toc1743260901)

[2. BIG DATA TECHNOLOGY 8](#_Toc1062699612)

[2.1 Overview of Big Data Technology 8](#_Toc280537252)

[2.2 Big Data Technologies 9](#_Toc1969705120)

[2.2.1 HADOOP 9](#_Toc737480297)

[2.2.2 Hadoop Distributed File System (HDFS) 9](#_Toc1610075289)

[2.2.3 MapReduce 10](#_Toc1962606660)

[2.2.4 Yet Another Resource Negotiator (YARN) 12](#_Toc1623003566)

[3. REAL WORLD APPLICATIONS 13](#_Toc1472618690)

[3.1 Clinical Big Data and COVID-19 13](#_Toc1380529277)

[3.2 Finance Sector 14](#_Toc2003152294)

[3.3 Social Media Platforms 15](#_Toc1468161744)

[4. CRITICAL ANALYSIS and EVALUATION 17](#_Toc1702828471)

[CONCLUDING REMARKS 20](#_Toc1452849038)

[BIBLIOGRAPHY 21](#_Toc503838737)

[APPENDIX (if necessary) 23](#_Toc529804805)

**Statement of compliance with academic ethics and the avoidance of plagiarism**

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname (Capital letters):TURAN CAN GÜN

.......................................................................................................................................

Date: 03 / 07 /2024

# INTRODUCTION

Big Data has enormous potential for accelerating efficiency and growth, making it an indispensable tool in today's corporate environment. Companies use big data to improve decision-making, optimize operations, and obtain deeper insights into consumer behavior. Businesses can find hidden patterns, trends, and correlations by analyzing massive datasets, which leads to more informed plans and competitive advantages.

A number of essential parts make up Hadoop, a core technology in the Big Data ecosystem. Huge volumes of data can be stored across numerous workstations with the Hadoop Distributed File System (HDFS), which guarantees high availability and fault tolerance. The resource management layer of a Hadoop cluster, known as YARN (Yet Another Resource Negotiator), is in charge of assigning system resources to different applications. The MapReduce programming model divides jobs into smaller subtasks and runs them over numerous nodes in order to process huge datasets in a distributed setting.

# 1. INTRODUCTION to BIG DATA

Big data can be considered as a huge size dataset and which is able to used for distributed computation , SQL operations , data mining-machine learning and data visualization.

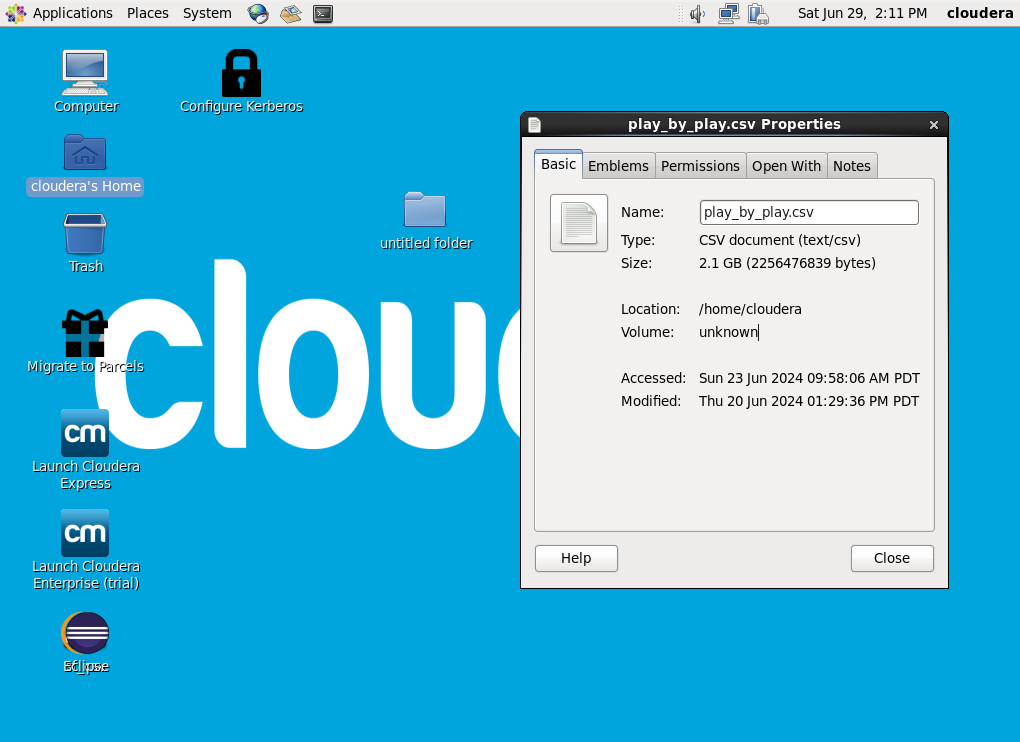


Figure 1.1 Properties of the Sample Dataset

The dataset is taken from Kaggle also it is updated by daily.The dataset consists 30 teams , over 4800 players and 65.000+ games and the dataset is built up from over 13.000.000 rows.As Figure 1.1 demonstrates the dataset has size as 2.1 GB so which can be considered as Big Data.

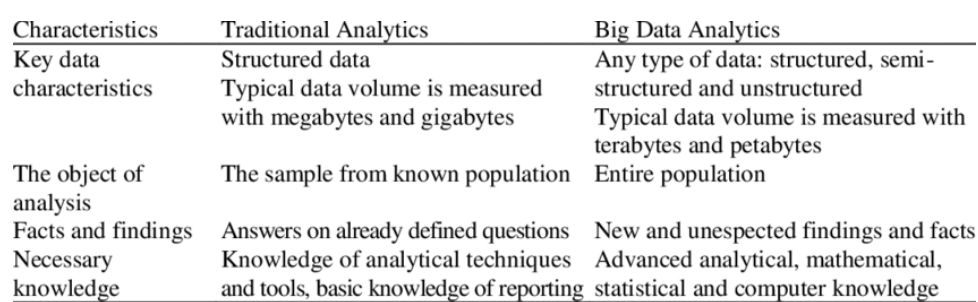


Figure 1.2 Comparision of Traditional and Big Data (Yan,J. 2013 Big Data,Bigger Opportunities)

Traditional analytics uses megabytes or gigabytes of structured data, analyzing samples from known populations to provide predetermined answers to questions. This type of analysis requires a foundational understanding of analytics. In contrast, Big Data Analytics handles any data type, including gigabytes , terabytes and petabytes, analyzing entire populations to uncover unexpected insights, necessitating advanced analytical, mathematical, and computational expertise.

According to World Economic Forum between 2023-2027 Big Data Analyst and related occupations will be grown by %58 as a net growth (WEF 2023 , Future of Jobs Survey).

  
Figure 1.3 Expected Impact of Techonolgy Adoption on Jobs (World Economic Forum, Future of Jobs Survey 2023)

Big Data concept is built up on 5V’s rules as following at the below;

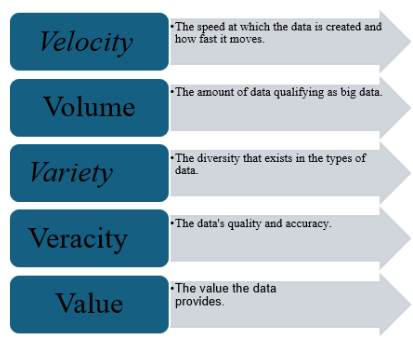


Figure 1.4 5V’s Rule of Big Data (My Knowledge)

The significance of Big Data lies in its potential to provide big data is important because it has the ability to inform decisions and offer profound insights. Businesses are producing enormous volumes of data from a variety of sources, including social media, transactions with customers and mobile devices. By analyzing this data, patterns, trends that were previously hidden can be found. These insights can lead to improved business strategies, enhanced customer experiences, and increased operational

In the business context, big data analytics offers a competitive advantage by enabling companies to harness the power of their data to make informed decisions. For example, online shopping corporations could use big data to understand customer preferences and optimize their inventory.On the other side financial institutions can detect abnormal activities in real-time. Healthcare providers can analyze patient data to improve diagnostics and treatment plans, and manufacturers can enhance production processes by analyzing data from machinery and equipment.

Traditional methods to either handle or process Big Data isn’t efficient choice because;

Typical database systems weren’t designed to handle big data.Also to improving this type system across big data isn’t effective solution as well as not cheap.

From everywhere massive amount of data is produced during the everytime so these datasets must be handled as fast as possible.However traditional dataset systems aren’t capable to realize this demand.

As a natural feature of big data which contains unstructured-semi structured data and to handle big data wtih this situation by standard methods might not be easy for instutions.

# 2. BIG DATA TECHNOLOGY

## 2.1 Overview of Big Data Technology

Business operations are too dependent on the thorough analysis of big data to continue progressing. The analysis of big data assumes a significant part in the decision-making process to enhance the advancement and prosperity of the association (Elgendy,N. , 2016). In any case, sheer computing measures of information with the guide of customary information computing devices fail to provide productive results, and the device is not attainable. Therefore, several big data tools have been developed in the past years that assist associations and data researchers in inferring the information-driven choices productively and cost-effectively.

## 2.2 Big Data Technologies

### 2.2.1 HADOOP

Hadoop is a system which consist various libraries to handle enormous amount of data. The main benefit of the Hadoop system, where the distributed repository is known separately as the Hadoop Distributed File System (HDFS), is its versatility. This Hadoop framework is a highly accessible analytical tool independent of the hardware equipment. Furthermore , Hadoop system entails the normal libraries, stockpiling libraries, Hadoop YARN, and Hadoop MapReduce to compute the extremely large datasets (Abdalla, H.B. , 2022).

### 2.2.2 Hadoop Distributed File System (HDFS)

HDFS is a system which can be considered as stockpilling structure.Moreover it has a good advantage it is related to handling both structured and unstructured information of big datasets effectively.However it isn’t designed as general purpose framework since main target across to high-idleness activities cluster computing also HDFS isn’t able to give quick record queries in documents.The biggest advantage of HDFS that is versatility towards to heterogeneous equipments as well as programmimg steps. It allocates huge amounts of information to the group. Indeed, the group has a novel expert (NameNode) that controls document framework activities and many slaves (DataNodesTo give information accessibility, Hadoop is located on information replication (Abdalla, H.B. , 2022).

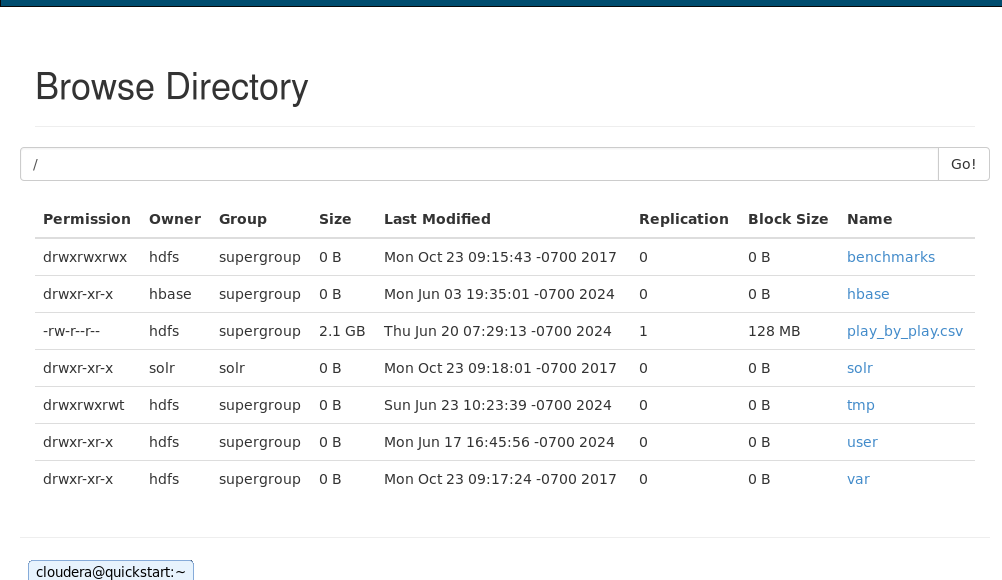


Figure 2.1 HDFS System of Cloudera (The Sample Dataset)

As can be seen at the above our dataset has been transferred to HDFS then which is divided into around 17 block (2100 MB / 128 MB) then each block has 128 MB size.

### 2.2.3 MapReduce

MapReduce is a scheduling framework that responsible from creating software framework and its execution.MapReduce can be defined as a new generation fundamental improvement for BigData control devices and manipulaiton and which can handle enormous amount of data by its productive instruments as well as this techonolgy is able to allows realizing parallel computing.

Figure 2.2 indicates that Framework of Mapreduce in Hadoop. It consists of map as well as reduce operations. The multiple inputs are obtained, divided, and shuffled during the map stage. During this reduction stage, all reduce tasks procedure the input in-between data with a decrease function and create output (Rajendran, S. ,2021).

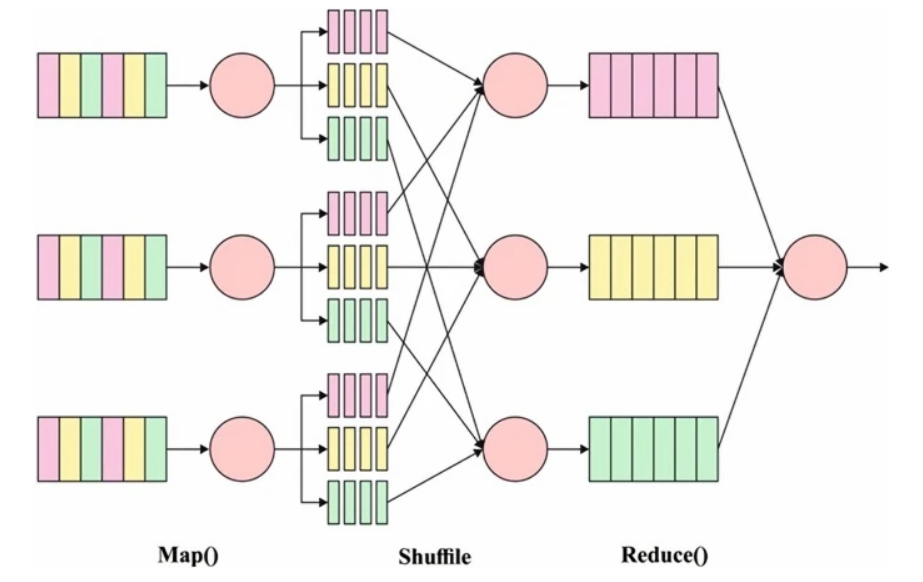
.

Figure 2.2 Illustration of MapReduce (Rajendran, S. ,2021)

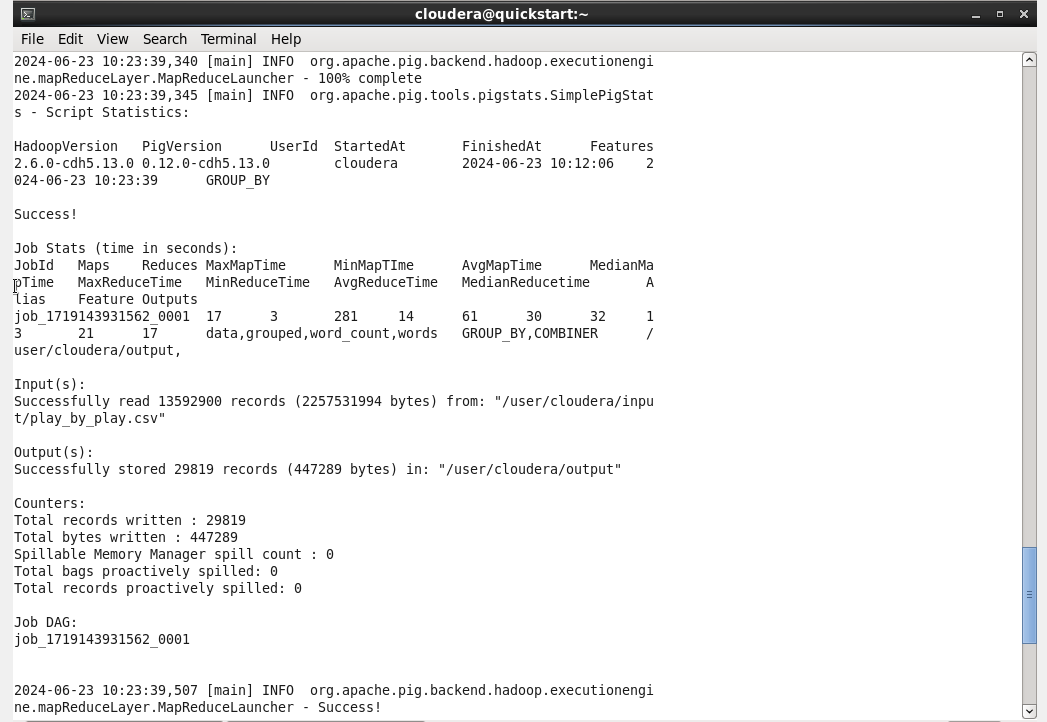


Figure 2.3 Sample MapReduce Application of Our Dataset

As Fig 2.3 implies we are able to apply MapReduce onto our dataset then as a result of this situation which can store 29819 records from 13592900 records.Moreover although the most popular method to realize MapReduce is Java in this sample application Pig method is used as shown at the below of Figure 2.3 .On the other hand from “Job Stats” part operation time can be obtained by splitted version like Max Reduce and Min Reduce time.

### 2.2.4 Yet Another Resource Negotiator (YARN)

The YARN framework is more conventional than MapReduce. It gives priority versatility, equality, and progressed asset administration in contrast with MapReduce. It transforms working framework capacities for Big Data in terms of scientific solicitation. The YARN Resource is integrated into the modified Hadoop structure. Specifically, the applicaiton of YARN takes place at the peak of HDFS. This position aloows the parallel computation of multiple applications. It allows handling both batch processing and real-time interactive processing. YARN is deferent with the Application Programming Interface (API) of MapReduce. Users have just to recompile MapReduce jobs to run them on YARN.

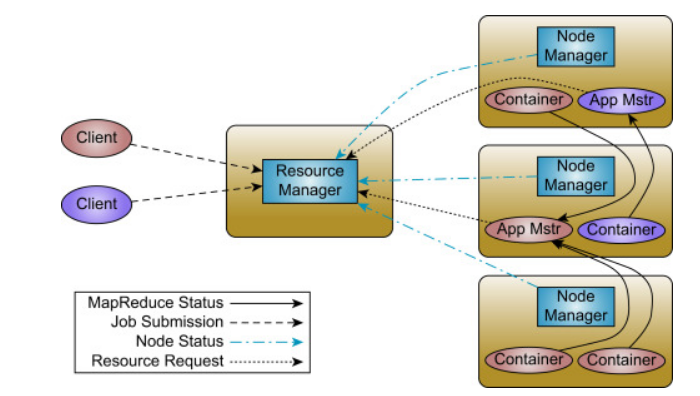


Figure 2.3 YARN Architecture (Krishnan, K. 2020)

YARN basicallyis built up form 2 main parts as follwing ;

The Resource Manager (RM) allocates and manages resources across the cluster.

Application Master (AM) is a system comprised of huge archives that allocates specific tasks, which are equalized with TaskTrackers to track their enhancement. Getting more books, keeping up counters, failing at self-detrainment, or being late are all tasks that are empowered by the AM (Krishnan, K. 2020) .

# 3. REAL WORLD APPLICATIONS

Hadoop is a suite of open-source software tools designed to facilitateLarge-scale data processing in distributed computing environments can be made easier with the help of the open-source Hadoop software tool suite. MapReduce, HDFS (Hadoop Distributed File System), and YARN (Yet Another Resource Negotiator) are some of its essential parts. These technologies have been used in numerous real-world applications, proving how reliable and effective they are at managing large amounts of data.

## 3.1 Clinical Big Data and COVID-19

In a clinical setting such datasets are emerging from large-scale laboratory information system (LIS) data, test utilization data, electronic medical record (EMR), biomedical data, biometrics data, gene expression data, and in other areas. Large datasets are very challenging to query and analyze with conventional methods, particularly when the queries are complex. analyse and query using traditional mechanisms, especially when the queries themselves are quite complicated. In effect, a MapReduce algorithm maps both the query and the dataset into constituent parts. The mapped components of the query can be processed simultaneously – or reduced – to rapidly return results (Mohammed, E.A 2014) .

To investigate the possibility of using the MapReduce framework to speed up biomedical data mining tasks using this pharmacovigilance case as one specific example. The results demonstrated that the MapReduce programming framework could improve the performance of common signal detection algorithms for pharmacovigilance (Wang , W. et. Al. , 2011) in a distributed computation environment at approximately linear speedup rates.

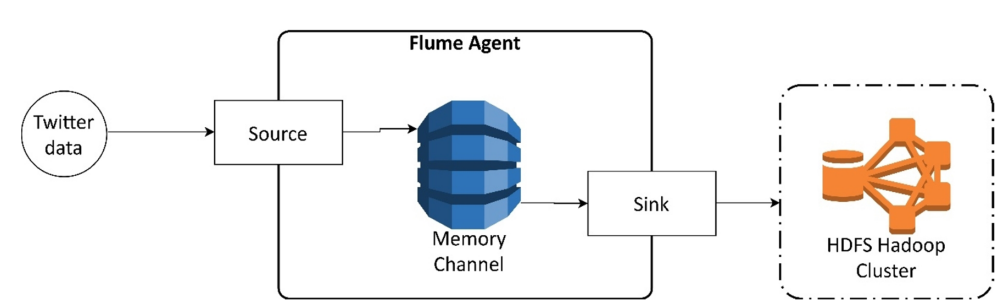
Second scneario is that distributed systems can analyze the enormous datasets derived from genomic data very well. Single-nucleotide polymorphism (SNP) selection algorithms have a new and improved tag thanks to the MapReduce framework (Chen, W-P et al., 2014). Based on HBase (Wynden, R., 2011) for data management and MapReduce jobs for computation, a genome sequence comparison algorithm (Zhang, K. et al., 2003) has been implemented on top of Hadoop. e comparison algorithm (Zhang , K. et al. , 2003) has been implemented on top of Hadoop while relying on HBase (Wynden , R. , 2011) for data management and MapReduce jobs for computation. The system performance has been tested with real-life genetic sequences on the level of single genes as well as artificially generated test sequences (Zhang , K. et al. , 2003). While the initial test runs clearly illustrated the feasibility of the approach, more work is needed to improve the applicability of the solution. Moreover additional tuning of the local Hadoop configuration towards the genome comparison is expected to yield additional performance benefits tuning of the local Hadoop configuration towards the genome comparison is expected to yield em and the Pig Latin data flow language (Nordberg , H. et al. , 2013).

## 3.2 Finance Sector

As was previously noted , MapReduce splits data processing activities into two steps inside the HADOOP ecosystem: the reduction phase, which involves aggregating the results of the mapping phase to produce the final output, and the mapping phase, which involves splitting data into smaller pieces and processing them in parallel. This capacity to analyze data in parallel is very helpful for financial data analysis, where efficiency and speed are critical. Financial firms can use MapReduce to quickly and accurately do intricate calculations, portfolio analysis, and risk assessment on their data. Moreover, the Apache Hadoop ecosystem has a number of parts that improve its financial data management capabilities. For example, Hive is a SQL-like query language and data warehousing system developed on top of Hadoop.. Hive facilitates the use of common SQL syntax for financial data interaction by data scientists and analysts, hence streamlining the process of performing queries, aggregations, and reporting. According to S. Wadkar (2014), this characteristic expedites the analytical process, facilitating speedier insights and well-informed decision-making in the financial sector. On the other hand ,Pig is an important part of the Hadoop ecosystem; it's a high-level framework for developing MapReduce applications. Pig makes it easier to construct data processing applications by offering Pig Latin, a scripting language that encapsulates the difficulties involved in writing low-level MapReduce code. Pig may be used by financial institutions to design bespoke data processing pipelines that meet their unique needs, making processes like data enrichment, transformation, and cleansing easier (Ali ,M. 2022) .

## 3.3 Social Media Platforms

One of the sutiable example for this section might be related to Twiiter big data architecture and which can be explained as following ;

Figure 3.1 Illustration of Flume Architecture (Podhoranyi , M 2021)

|  |
| --- |
| Source Type --> Twitter – source was connected to the Twitter interface to download tweets continuously. All data were serialized by AVRO serialization system. AVRO stores the data definition in JSON format |
| Channel Type --> Memory Channel – The events were stored into an in-memory queue. The drawback of this channel is the inability of data recovery in the case of agent failure. It acts like a buffer and a bridge between source and sink. |
| Sink Type --> HDFS Sink – The sink delivered data into the Hadoop cluster/HDFS. HDFS was the final destination for ingested data |

Table 3.1 Configuration Details of Twitter Big Data System (Podhoranyi , M 2021)

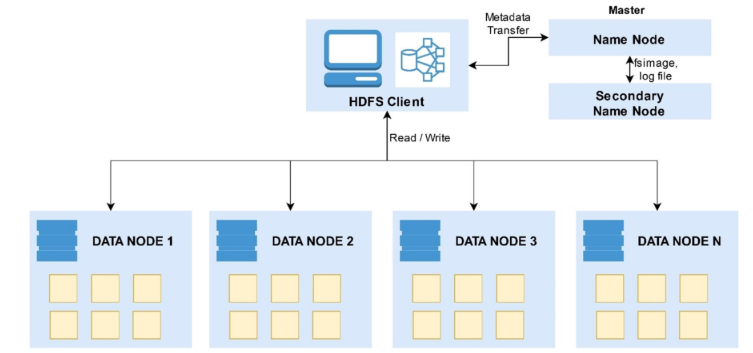


Fig 3.2 HDFS Hadoop Cluster Architecture ( Podhoranyi , M 2021)

The DataNodes mainly workThe actual data that is physically stored on the nodes is primarily managed by the DataNodes. We used the Flume agent to move data from HDFS to Twitter API and into the pre-configured cluster. Consequently, the data were distributed across the nodes in the cluster and split into large blocks (size 128 MB) and independently replicated at multiple data nodes (Replication factor = 3). Replication factor is the quantity of times that each Data Block is replicated by the Hadoop framework. The system can self-recover after a cluster node fails because all blocks are kept on the local file system, which serves as a fault tolerance mechanism.the to provide a fault tolerance mechanism to the system, and therefore the system is able to recover yourself after a node in the cluster fails (Podhoranyi , M 2021).

Effective resource management that offers quick access to resources when they are needed is a must for all data processing systems. These resources could be disk, network, memory, or CPU. disk or network. The proposed architecture made use of Hadoop Yarn as a resource management tool. While HDFS is the storage layer of Hadoop, Yarn is the resource management layer of Hadoop architecture (Podhoranyi , M 2021) .

# 4. CRITICAL ANALYSIS and EVALUATION

|  |  |  |
| --- | --- | --- |
| Comparison Field | HADOOP | YARN |
| Resource Management | Hadoop support programming model which support parallel processing that we know as Map Reduce. Before Hadoop 2.0, Hadoop 1.0 already support Mapreduce. | Hadoop 2.0 using YARN for resource management |
| Recourses Retrieval | When user submit Mapreduce Job. Resource will be back to free. | Resource manager will give MR master all resources it needs or it is according to cluster computing capabilities |
| Cluster Resources | In Hadoop 1.0, there is tight coupling between Cluster Resource Management and Map Reduce programming model. | In Hadoop 2.0, this is totally based on Cluster Resource Management |
| Cluster Management | Map Reduce job is divided between number of tasks called mappers and reducers. Each task runs on one of the machine (Data Node) of the cluster, and each machine has a limited number of predefined slots (map slot, reduce slot) for running tasks concurrently. | Job Tracker is responsible for both managing the cluster’s resources and driving the execution of the Map Reduce job. It reserves and schedules slots for all tasks, configures, runs and monitors each task |
| Resources Sharing | Multiple applications can run on Hadoop via Map Reduce and all application could not share common resource management | Multiple applications can run on Hadoop via YARN and all application could share common resource management. |

Table 4.1 Comparision of Hadoop and YARN (Abdallah ,F. 2017)

From the above table, it can be concluded that Yarn platform is better than Hadoop for several applications, Yarn has central resource manager component resources and allocates the resources to the application, while Hadoop, does not have this type of resources to the application.

|  |  |  |  |
| --- | --- | --- | --- |
| Map Reduce Implementations | HADOOP | YARN | SPARK |
| Scalability | Max 5000 nodes | Exceeds 4000 nodes | Over 8000 nodes |
| Data Input/Output | On disk | Memory – CPU -Disk | In memory |
| Real Time Process | Batch process | Master application job processing | 10x faster on disk or 100x faster on memory |
| Fault Tolerance | Yes | Yes | Yes |
| Data Size Supported | Pet bytes | Pet bytes | From mega to pet bytes |
| Iterative Tasks Support | Slow iterative (HDFS) | Central resource manager | Support iterative tasks by Scala interpreter |

Table 4.2 General Comparison between Mapreduce Implementations (Abdallah ,F. 2017)

Table 4.2 shows the comparison between Hadoop, spark and yarn, starting with Hadoop, we see that Hadoop maximum scalability size reaches 5,000 nodes in cluster but Yarn exceeds 4,000 nodes that’s a massive scale compared with Hadoop, also spark scale reached over 8,000 nodes all of them supports data size up to petabyte. Table 4.2 presents a comparative analysis of Hadoop, Spark, and Yarn. In particular, Hadoop can only support up to 5,000 nodes in a cluster, whereas Yarn can support up to 4,000 nodes, suggesting a much bigger scale than Hadoop. Furthermore, all three platforms support petabytes of data, and Spark's scale reached over 8,000 nodes. maximumAdditionally, Spark's scale reached over 8,000 nodes, and all three platforms support data sizes up to petabytes. More data I/O reveals that Hadoop writes back to disk after map or reduce operations, whereas Spark processes its data in memory. Fault tolerance all have fault tolerance but differ in the methodology using. Spark and Hadoop Mapreduce both have acceptable failure tolerance, but Hadoop Mapreduce is a little more tolerant (Abdallah ,F. 2017).

# CONCLUDING REMARKS

In conclusion, Big Data has revolutionized the business landscape by providingIn summary, big data has completely changed the business environment by enabling previously unheard-of insights and influencing strategic decision-making. Businesses can effectively manage and analyze massive datasets with the integration of technologies such as Hadoop, HDFS, YARN, and MapReduce. Scalable and fault-tolerant solutions are guaranteed by the Hadoop ecosystem thanks to the powerful processing power of MapReduce and the sturdy storage capabilities of HDFS.MapReduce ensures scalable and fault-tolerant solutions. Resource management is further improved by YARN, enabling the handling of multiple applications at once. The useful uses of these technologies in industries like social media, healthcare, and finance highlight how important they are. Learning Big Data technologies will continue to be essential for gaining competitive advantages and promoting innovation as long as businesses generate and rely on enormous amounts of data.

# BIBLIOGRAPHY

1. Abdalla, H.B. A brief survey on big data: technologies, terminologies and data-intensive applications. *J Big Data* 9, 107 (2022).
2. Ali, M. and Iqbal, K., 2022. The role of Apache Hadoop and Spark in revolutionizing financial data management and analysis: A comparative study.
3. Al-Alem, F.A., 2017. Big data analysis: Implementations of Hadoop MapReduce, YARN, and Spark. *Journal of Information Engineering and Applications*, 7(3). pp. 10.
4. Elgendy, N.S. and Elragal, A., 2016. Big data analytics in support of the decision making process. *Procedia Computer Science*, 100, pp.1071-1084.
5. Elgendy, N. and Elragal, A., 2016. Big data analytics in support of the decision making process. *Procedia Comput Sci.*, 100, pp.1071-84.
6. Walsh ,W. 2023 NBA Dataset Available at: <https://www.kaggle.com/datasets/wyattowalsh/basketball> (Accessed 19 July 2024)
7. Krishnan, K., 2020. In *Building Big Data Applications*.
8. Lukić, J., 2017. The impact of big data technologies on competitive advantage of companies. *Facta Universitatis, Series: Economics and Organization*, 14(3), pp.255-264.
9. Mohammed, E.A., Far, B.H. & Naugler, C. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. *BioData Mining* 7, 22 (2014).
10. Nguyen, A.V., Wynden, R., Sun, Y., 2011. HBase, MapReduce, and Integrated Data Visualization for Processing Clinical Signal Data. *AAAI Spring Symposium: Computational Physiology*, 2011.
11. Nordberg, H., Bhatia, K., Wang, K., Wang, Z., 2013. BioPig: a Hadoop-based analytic toolkit for large-scale sequence data. *Bioinformatics*, 29(23), pp.3014-3019.
12. Podhoranyi, M., 2021. A comprehensive social media data processing and analytics architecture by using big data platforms: a case study of Twitter flood-risk messages.
13. Rajendran, S., Khalaf, O.I., Alotaibi, Y. et al., 2021. MapReduce-based big data classification model using feature subset selection and hyperparameter tuned deep belief network. *Sci Rep* 11, 24138 (2021).
14. Wadkar, S. and Siddalingaiah, M., 2014. *Pro Apache Hadoop*. Apress.
15. World Economic Forum, 2023. *Future of Jobs Report 2023*. INSIGHT REPORT, May 2023.
16. Yang, X., Huang, K., Yang, D., Zhao, W., Zhou, X., 2023. Biomedical Big Data Technologies, Applications, and Challenges for Precision Medicine: A Review. *Glob Chall.*, 8(1):2300163.
17. Zhang, K., Sun, F., Waterman, M.S., Chen, T., 2003. Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology: 2003. Dynamic programming algorithms for haplotype block partitioning: Applications to human chromosome 21 haplotype data, pp.332-340.

# APPENDIX (if necessary)

AM: Application Master

API: Application Programming Interface

CPU: Central Processing Unit

EMR: Electronic Medical Record

HDFS: Hadoop Distributed File System

LIS: Laboratory Information System

RM: Resource Manager

SQL: Structured Query Language

WEF: World Economic Forum

YARN: Yet Another Resource Negotiator