

Hadoop

Epuret M. Obuya, 2021/HD05/2291U, *Makerere University*

1 INTRODUCTION

Over the years, data produced and stored in different databases have grown exponentially leading to the terminology big data. Big data is small data collected in big sizes with exponential growth. This big data is so voluminous that the traditional relational databases can't efficiently store and manage it[1]. This big data is of great benefit to organisations and businesses but there must be sustainable ways to manage it. Big Tech giants like Meta, Amazon, Google and Yahoo make use of data they collect from clients to generate insights and patterns and make business decisions that aim to maximise profits and minimise costs. For these giants to be able to store and process these large volumes of data, they employ Apache Hadoop.

2 HADOOP

2.1 What is Hadoop

Hadoop is an open-source framework that helps and is used to store and efficiently process large data sets that range in sizes of gigabytes to as big as petabytes[2]. Hadoop stores data in cluster servers and processes across these distributed clusters in parallel. It provides building blocks on which other applications can be built. As time has gone by, the Hadoop ecosystem has grown and it's made up of several components to facilitate the collection, storage, processing, analysis and management of big data. Hadoop is based on the paradigms of fault tolerance, reliability, availability, security, efficiency and accuracy[3].

2.2 Components of Hadoop

Hadoop is made up of the following components

Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) is a distributed file system that provides high throughput and efficient access to data. HDFS replicates data to sustain availability at all times and in case of failures, the process doesn't stop[3]. The HDFS also stores system metadata and application data separately[4]. The HDFS is made up of the following components.

1. Master (Name) Node – this node contains the meta-data about the system. It's a hierarchy of files and directories which are represented by inodes and store attributes like permissions, modifications and access times, namespaces and disk space quotas[4].
2. Data Nodes – these are the actual storage nodes where the data needed is held and can be read and written. Each block on the data node is represented by 2 files in the local file system with one file containing the data itself and the other containing the metadata of the block.
3. Secondary Name Node – these are the nodes that are used as helpers to the Master node where any action performed by the Master creates a checkpoint and this is stored in the secondary node[3].

MapReduce

This is another component of Hadoop and this is responsible for performing the processing of the data kept in the HDFS. MapReduce performs operations in key/value pairs and returns results as a value or values or key/value pairs as well. The major functions of MapReduce are Map and Reduce and these are the only two functions that perform operations on data[3].

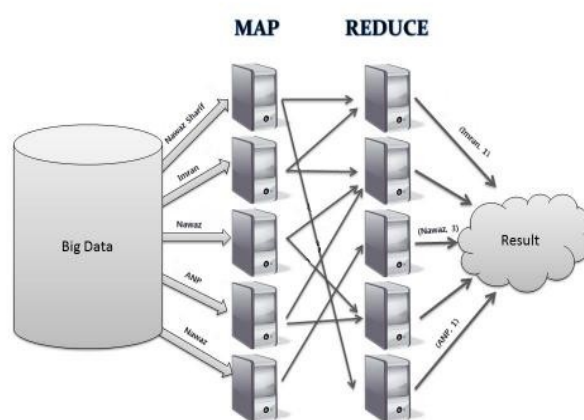


Fig. 1. MapReduce Architecture

Yet Another Resource Negotiator (YARN)

YARN manages and monitors resource usage and cluster nodes while also scheduling jobs and tasks. YARN consists of the scheduler, resource manager, application master, container and node manager. The node manager sends periodical updates to the resource manager on its aliveness. The application master sends a resource request to the resource manager and the scheduler is responsible for scheduling jobs and providing resources to the application master as the container. The resource manager contains the scheduler and application master[5].

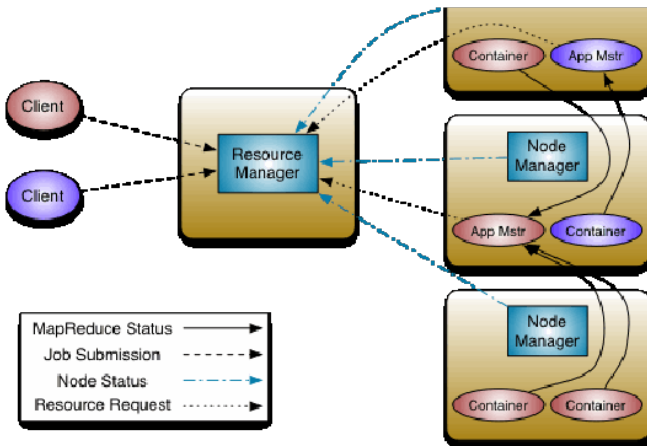


Fig. 2. YARN Architecture

Hadoop Common

This provides a set of libraries that are usable in all in all modules these are a set of utilities that support the other Hadoop modules[6].

3 HOW HADOOP WORKS

Hadoop uses Master/Slave architecture where the Master is responsible for the NameNode and Resource manager. The Resource Manager initiates tasks, and tracks and dispatches their implementation. The Application Manager handles the processing of local data and collection of results and reports to the resource manager. The Data Node and Name Node fulfill HDFS tasks while Resource Manager and Application Manager fulfill MapReduce tasks[7]. Hadoop basically works in such a way that the data processing is distributed across multiple clusters and processed simultaneously. HDFS performs the storage, MapReduce performs the data processing and YARN performs the division of the tasks across the existing resources.

4 STRENGTHS OF HADOOP

- Hadoop is suitable for deployment in low-cost machines since its HDFS is a subordinative construction system[7].
- Hadoop is highly fault-tolerant because of the existence of secondary name nodes thus in case of failure of one node, there is always a backup available to help avoid a pause in service[3].
- There is always data availability because of the replication of data on various nodes.

5 CHALLENGES AND WEAKNESSES OF HADOOP

- Hadoop employs strictly batch processing and this makes it hard to process data in real-time applications.
- Hadoop is not suitable for incremental computing in that for any small change in the data, MapReduce has to run all over again.
- With the growth of data over time, there is a need to rewrite the application's particular algorithms which increases algorithm and code complexity[3].
- The results of MapReduce can not be used until it has

been saved to disk.

- With data volumes growing exponentially, there are innate design issues that prevent Hadoop from scaling to the extreme scales[8].
- HDFS is not suitable for small files with Name Node going to require a lot of space in case there is a large number of small files thus bringing about scalability issues.
- With many clusters being set up in different data centres, there is a possibility of energy efficiency issues arising[9].
- HDFS handles different data classifications and has no appropriate role-based access for controlling security issues[10].

6 APPLICATIONS OF HADOOP

Hadoop is widely used in large data processing which data is used in different sectors to help with fraud detection and marketing through analysis of transactions and fraud patterns and targeted marketing, security and law enforcement, sentiment analysis by businesses, and media among others. The meteorological organisations use Hadoop on the big data to forecast weather over the next coming days for any place in the world. This has been used to predict hurricanes, storms, monsoons, heavy rains, flash floods and landslides among others. Media and entertainment industries use Hadoop with the big data from the industry to perform targeted advertising, recommend new products and content, perform sentiment analysis of customers and help them understand why a client changed or terminated their subscription through analysis of data of previous clients who changed their subscriptions. This analysis helps them understand what causes the clients to most likely terminate thus they tailor the services in order to keep the clients.

7 CONCLUSION

With the advent of the digital age, new technologies have and are still emerging increasing the size of the data pool. There is still exponential growth of big data in the different industries and disciplines thus adopting Hadoop for data storage and analysis. There may be limitations existing in the Hadoop environment but there is active research and development of tools to fix the shortcomings. Thus, as this data grows as companies adopt Hadoop and related technologies, there is going to be an increase in the need for experts in the field of big data with skills in Hadoop thus having many making careers out of Hadoop.

REFERENCES

- [1] B. Saraladevi, N. Pazhaniraja, P. Victor Paul, M.S. Saleem Basha, P. Dhavachelvan, Big Data and Hadoop-a Study in Security Perspective, Procedia Computer Science, Volume 50, 2015, Pages 596-601, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2015.04.091>.
- [2] "What is Hadoop?," Amazon Web Services, Inc. <https://aws.amazon.com/emr/details/hadoop/what-is-hadoop/>

- [3] A. Alam and J. Ahmed, "Hadoop Architecture and Its Issues," 2014 International Conference on Computational Science and Computational Intelligence, 2014, pp. 288-291, doi: 10.1109/CSCI.2014.140.
- [4] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010, pp. 1-10, doi: 10.1109/MSST.2010.5496972.
- [5] B. J. Mathiya and V. L. Desai, "Apache Hadoop Yarn Parameter configuration Challenges and Optimization," 2015 International Conference on Soft-Computing and Networks Security (ICSNS), 2015, pp. 1-6, doi: 10.1109/ICSNS.2015.7292373.
- [6] Apache Software Foundation, "Apache Hadoop," *Apache.org*, 2019. <https://hadoop.apache.org/>
- [7] H. Lu, C. Hai-Shan and H. Ting-Ting, "Research on Hadoop Cloud Computing Model and its Applications," 2012 Third International Conference on Networking and Distributed Computing, 2012, pp. 59-63, doi: 10.1109/ICNDC.2012.22.
- [8] K. Wang et al., "Overcoming Hadoop Scaling Limitations through Distributed Task Execution," 2015 IEEE International Conference on Cluster Computing, 2015, pp. 236-245, doi: 10.1109/CLUSTER.2015.42.
- [9] J. Leverich and C. Kozyrakis, "On the energy (in)efficiency of Hadoop clusters," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 1, pp. 61-65, Mar. 2010, doi: 10.1145/1740390.1740405.
- [10] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, "Big Data and Hadoop-a Study in Security Perspective," *Procedia Computer Science*, vol. 50, pp. 596-601, 2015, doi: 10.1016/j.procs.2015.04.091.