The Main Components of Hadoop Frameworks

Author, Thisal Avishka Wijayasekara

A Data Science Foundation Blog

May 2020

---

www.datascience.foundation

Hadoop is an open-source software package for storing and processing large amounts of #data in small hardware clusters. When it comes to the main components of Hadoop, we recognise two components, which are **Data Storage and Management** and **Processing and Computation**. (DataFlair, 2019)

Data Storage and Management - Hadoop Distributed File System (HDFS)

This is the most important component of the Hadoop ecosystem. HDFS is Hadoop's primary storage system. Hadoop Distributed File System (HDFS) is a Java-based file system that provides reliable, fault tolerance and accessible data storage for the big data. HDFS is a distributed file system that runs on conventional hardware. HDFS is already configured with the default settings for many installations. Typically, a large cluster configuration is required. Hadoop interacts directly with HDFS using commands. Wen comes to HDFS, there are also two components can be identified, which are known as **Name Node** and **Data Node**. *(DataFlair, 2019)*

Name Node

It is also known as Master node. Here, it does not store actual data or datasets. Name Node stores the Meta data, for an example, the number of calls transform from a tower, their position, where the end users are getting the call, the Data node data and other details. Basically, this contains files and directories. The tasks of Name node can be recognized as follows. *(DataFlair, 2019)*

- Managing file system namespace
- Controlling the access of clients to files
- Executing file system through naming, opening, closing files and directories

Data Node

Data node is called as Slave. Data node is responsible for the effective storage of data in HDFS. The data node completes read and write operations on customer request. Replica Block of Data node consists of two files in the file system. The first file is for data and the second for registry metadata. HDFS metadata contains a data control. At startup, each Data node is connected to the appropriate Name node and grasp. The ID of the Data Node namespace and the software version are controlled by the handshake. If a discrepancy is detected, Data Node is automatically disabled. When comes to tasks of Data node, those can be detailed as follows. *(DataFlair, 2019)*

- This is consisting of operations like block replica creation, deletion, and replication according to the instruction of Name node
- Managing data storage of the system

Processing and Computation – Hadoop MapReduce

When comes to Hadoop MapReduce, that is the main component of the Hadoop, that provides data processing. MapReduce is can be identified as an easy-to-write application framework that processes the large amount of structured and unstructured data stored in the Hadoop distributed file system. *(DataFlair,*

2019)

MapReduce programs are parallel, so they are very useful for large-scale data analysis using multiple clusters. Therefore, this parallelism increases the speed and reliability of the cluster. In MapReduce, there are two functions, Map function and Reduce function. *(DataFlair, 2019)*

Two functions can be identified, map function and reduce function.

- The map function retrieves a data set and converts it to another data set. Each element is divided into processing (key / value pairs).
- The Reduce function accepts the Map output as an input and integrates these data nodes based on the key and changes the key value accordingly.

Using Hadoop for Processing Large Datasets such as Data Record (CRD) or Customer Transaction Data

Hadoop is used for processing the large data sets such as Call Data Record (CDR) or Customer Transaction Data.

CDR is a record that contains detailed information on a telecommunications transaction, such as the start time of the call, the end time of the call, the duration, parties of the call, the phone ID, the websites requested, the type of remote access or Internet data and so on. *(Sinha, 2019)*

Map reduce is a programming model for examining and processing large data sets. Apache Hadoop is an effective framework and most popular implementation of the map reduce model.

Reasons Big Data Engineers are Moving to Lamba (Λ) Architecture to Analyze Data

Big data, Internet of Things (IoT), Machine learning, and various other systems today are coming to the stage these days. People from all walks of life begin to interact with data warehouses and servers as part of their daily lives. So, we can say that investing in the best way becomes the main area of interest for companies, scientists and as well as big data engineers and people. *(Samizadeh, 2018)*

There are various data processing architectures and one of them is Lambda Architecture. Big data engineers are moving for Lambda architecture because of various reasons. So, from now onwards, here is a brief discussion to prove the statement.

The Lambda architecture is a data processing method that can efficiently process large amounts of data. The efficiency of this architecture is manifested in terms of high efficiency, low latency, and small errors. The Lambda architecture can be considered a near real-time data processing architecture. Use the package and stream features to keep adding new data to main storage and protect existing data. Companies such as Twitter, Netflix and Yahoo use this architecture to meet quality of service standards.

(Samizadeh, 2018)

The Batch layer of Lambda architecture manages historical data in distributed memory with fault tolerance, reducing the possibility of errors in the even in a system failure. It also balances speed and reliability. In addition, it is a scalable and fault tolerant architecture for data processing. *(Samizadeh, 2018)*

So, if you need an architecture, that is more reliable for updating your data lake, and want to succeed in developing a machine learning model to keeps the efficiency of the data, you can move to Lambda architecture to reduce errors and increase speed through batch layer and speed layer. *(Samizadeh, 2018)* Because of these kind of reasons, big data engineers are moving to the Lambda Architecture for data processing.

References

Ibm.com. (2020). [online] Available at: https://www.ibm.com/downloads/cas/ZO0AOLQX.

EDUCBA. (2020). *Big Data vs Data Science - How Are They Different ?* [online] Available at: https://www.educba.com/big-data-vs-data-science/.

Edwards, J. (2020). *Predictive analytics: Transforming data into future insights*. [online] CIO. Available at: https://www.cio.com/article/3273114/what-is-predictive-analytics-transforming-data-into-future-insights.html.

Framework, B. (2019). *The Four V's of Big Data | Big Data Framework©*. [online] Big Data Framework©. Available at: https://www.bigdataframework.org/four-vs-of-big-data/.

DataFlair. (2019). *Hadoop Ecosystem and Their Components - A Complete Tutorial - DataFlair*. [online] Available at: https://data-flair.training/blogs/hadoop-ecosystem-components/.

Kaminskiy, D. (2017). *What's the difference between 'Big Data' and 'Data'?*. [online] Digital Leaders. Available at: https://digileaders.com/whats-difference-big-data-data/.

Pierce, B. (2018). *Prescriptive Analytics Use Cases for Sales and Marketing*. [online] Blog.riverlogic.com. Available at: https://blog.riverlogic.com/use-cases-for-prescriptive-analytics-in-sales-marketing.

Samizadeh, I. (2018). *A brief introduction to two data processing architectures—Lambda and Kappa for Big Data*. [online] Medium. Available at: https://towardsdatascience.com/a-brief-introduction-to-two-data-processing-architectures-lambda-and-kappa-for-big-data-4f35c28005bb.

Sharma, H. (2019). *What Is Data Science? A Beginner's Guide To Data Science | Edureka*. [online] Edureka. Available at: https://www.edureka.co/blog/what-is-data-science/.

Sinha, S. (2019). *Hadoop Tutorial | Getting Started With Big Data And Hadoop | Edureka*. [online] Edureka. Available at: https://www.edureka.co/blog/hadoop-tutorial/.

Sisense. (2020). *What is Prescriptive Analytics? | Sisense*. [online] Available at: https://www.sisense.com/glossary/prescriptive-analytics/.

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email: admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation