# 2004-NPcrestHori_B-W School of InfoComm Technology

Distributed Data Pipelines

Diploma in Data Science (DS) October 2022 Semester

INDIVIDUAL ASSIGNMENT 1

(30% of Distributed Data Pipelines Module)

**Deadline for Submission:**

16th Dec 2022 (Saturday), 2359 Hours

|  |  |  |
| --- | --- | --- |
| Student Name | : | Radchanon Don Sukkram |
| Student Number | : | S10223354J |

#### Penalty for late submission:

10% of the marks will be deducted every day after the deadline.

**NO** submission will be accepted after 23rd Dec 2022, 23:59.

Contents

[Cover Page 1](#_Toc122106051)

[1. Introduction 3](#_Toc122106052)

[2. Similarity 3](#_Toc122106053)

[3. Differences 4](#_Toc122106054)

[3.1 Capability 4](#_Toc122106055)

[3.2 Data Processing 4](#_Toc122106056)

[3.3 Performance 5](#_Toc122106057)

[3.4 Costs 5](#_Toc122106058)

[3.5 Scalability 6](#_Toc122106059)

[3.6 Fault Tolerance 6](#_Toc122106060)

[3.7 Ease of Use 6](#_Toc122106061)

[References 8](#_Toc122106062)

# 1. Introduction

Hadoop and Apache Spark are both similar and different in many ways. There are many aspects of similarities and differences that will be discussed.

# 2. Similarity

Hadoop and Apache Spark are similar as both are open-source frameworks. They both can conduct data processing in the data pipeline stage. They also enable us to deal with huge collections of data no matter its format — from Excel tables to images and video files. They both also process data in a distributed environment. Furthermore, Both of them also have the same level of fault tolerance. However, the approach is different. More will be covered below.

# 3. Differences

Despite having a common function, both Hadoop and Apache have many differences.

## 3.1 Capability

First, in terms of capability. As mentioned, both Hadoop and Apache can conduct data processing. However, Hadoop can conduct more functions like storage and resource management. In contrast, Apache Spark can only conduct data processing as it was designed to replace Hadoop MapReduce. (AltexSoft, 2021).

## 3.2 Data Processing

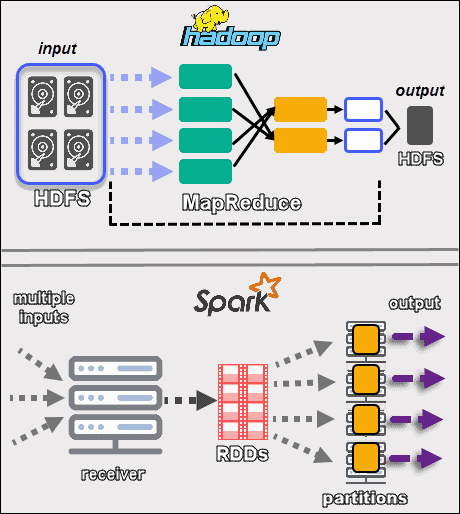


Figure : Hadoop vs Apache Spark Framework. The following visual compares the two frameworks in detail.

In terms of processing. Hadoop cannot process data in real-time while Apache Spark can process data near real-time. This is due to the differences in processing methods. Hadoop conducts batch processing using a hard disk to read or write data. Thus, Hadoop is suitable for batch processing and linear data processing. (IBM, 2021). On the other hand, Apache Spark conducts batch and micro-batch processing in RAM, making it ideal for real-time processing and processing live unstructured data streams. (IBM, 2021).

## 3.3 Performance

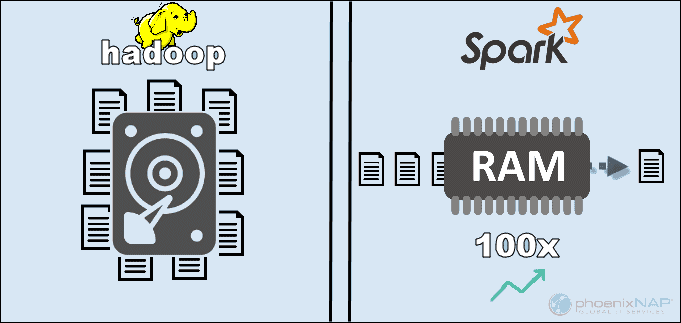


Figure : Shows why Apache Spark is 100x faster than Hadoop by displaying the difference in what is used by each one.

In terms of performance. Apache Spark has a much better performance as it is faster. This is because Apache Spark uses random access memory (RAM) unlike Hadoop which uses a hard disk to read and write data. (PhoenixNAP, 2020). Moreover, Hadoop stores data in various sources and processes it in batches using MapReduce which further slows its performance. (EDUCBA).

## 3.4 Costs

In terms of costs. It is more expensive to run Apache Spark compared to running Hadoop as infrastructure maintenance and development costs are higher. (Ksolves, 2021). The main cost is from the hardware required. Apache Spark relies on in-memory computations for real-time data processing which requires high quantities of RAM usage to spin up nodes, hence, increasing ownership costs. (IBM, 2021). In contrast, for data processing, Hadoop depends on any disk storage type. (PhoenixNAP, 2020). Compared to RAM, disks are a cheaper commodity. Hence, it is cheaper to run.

## 3.5 Scalability

In terms of scalability, Hadoop is easier to scale compared to Apache Spark as nodes and disks can simply be added for storage to accommodate the demand of rapidly growing volumes of data through the Hadoop Distributed File System (HDFS). (IBM, 2021). This is because HDFS is part of Hadoop’s architecture that allows for data storage. On the contrary, Apache Spark is hard to scale as it relies on RAM. If data is too large, Apache Spark relies on HDFS as shown in the visual. (IBM, 2021).

In terms of security, Hadoop is more secure than Apache Spark as it provides more authentication than Apache Spark by using multiple authentication and access control methods revolving around its five pillars of security which include administration, authentication perimeter security, authorisation, auditing, and data protection. (Ksolves, 2021). In contrast, Apache Spark is not secure as the security is turned off by default. To achieve necessary security, it has to integrate with Hadoop. Even with shared secret or event logging, it is not secure enough for production workloads. (PhoenixNAP, 2020).

## 3.6 Fault Tolerance

In terms of fault tolerance, both are similar but the approach is different. Hadoop splits each file and replicates it. If an issue occurs, the system resumes the work by creating the missing blocks from other locations. (PhoenixNAP, 2020). This ensures that rebuilding takes place even when the machine is down. Apache Spark provides fault tolerance through its building block, resilient distributed dataset (RDD). RRDs can operate parallelly and refer to any dataset in the HDFS. (Ksolves, 2021). The system also tracks how the immutable dataset is created. On top of that, Spark can use DAG tracking of the workflows to rebuild data in a cluster. This data structure allows Apache Spark to handle failures. (PhoenixNAP, 2020)

## 3.7 Ease of Use

Lastly, in terms of ease of use, Apache Spark trumps Hadoop due to its easy-to-use and straightforward system. In addition, 80 high-level operators are used by the resilient distributed dataset (RDD) to process data. Spark also provides multiple APIs such as Java, Scala, Python, R, and SQL. (Ksolves, 2021). Hadoop on the other hand is a complicated and long framework because of its MapReduce code. Moreover, Hadoop requires low-level API and a huge amount of hand-coding. (Ksolves, 2021). Thus, poorer ease of use.

# References

AltexSoft. (2021, June 07). Hadoop vs Spark: Main Big Data Tools Explained. Retrieved from https://www.altexsoft.com/blog/hadoop-vs-spark/

EDUCBA. (n.d.). Hadoop vs Spark. Retrieved from https://www.educba.com/hadoop-vs-spark/

IBM. (2021, May 27). Hadoop vs. Spark: What's the Difference? Retrieved from https://www.ibm.com/cloud/blog/hadoop-vs-spark#:~:text=Like%20Hadoop%2C%20Spark%20splits%20up,use%20cases%20that%20Hadoop%20cannot.

Ksolves. (2021, July 13). A Comparison between Spark vs. Hadoop. Retrieved from https://www.ksolves.com/blog/big-data/spark/a-comparison-between-spark-vs-hadoop#:~:text=Security%20and%20Fault%20Tolerance%20Comparisons,building%20block%20of%20Apache%20Spark.

PhoenixNAP. (2020, June 4). Hadoop vs Spark – Detailed Comparison. Retrieved from https://phoenixnap.com/kb/hadoop-vs-spark#ftoc-heading-5