

# School of InfoComm Technology

**Distributed Data Pipelines**

Diploma in Data Science (DS)

October 2023 Semester

**INDIVIDUAL ASSIGNMENT 1**

(30% of Distributed Data Pipelines Module)

**Submission Deadline:**

**15th Dec 2023 (Friday), 2359 Hours**

|  |  |  |
| --- | --- | --- |
| **Student Name** | **:** | **Eng Zhen Ye** |
| **Student Number** | **:** | **S10243329J** |

**Penalty for late submission:**

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

**NO** submission will be accepted after 22nd Dec 2023, 23:59.

**Section A**



(Chittora & Chittora, 2023)

Hadoop and Apache Spark are both frameworks used for big data processing but differ significantly in their approaches and functionalities across various stages of the data pipelining process.

They are considered competitors in the industry by many and are few of the most looked-for technologies and platforms for big data analytics. However, in recent times companies that have been using Hadoop have also started implementing Spark and are even used together. (Chittora & Chittora, 2023)

Let us explore the similarities and differences of both to figure out why.

Similarities

|  |  |
| --- | --- |
| Data Preprocessing | Can handle large-scale data processing |
| Parallel Processing | Supports parallel processing, allow tasks to be split across multiple nodes in a cluster for faster computation |
| Distributed Computing | Operate on distributed systems, distributing data across clusters for processing |
| Fault Tolerance | Designed to handle failures within a cluster environment, ensures fault tolerance and data reliability |

Differences

|  |  |
| --- | --- |
| Hadoop Ecosystem - Analytics Vidhya | Apache Spark - Wikipedia |
| Data Sources Data Sources icon PNG and SVG Vector Free Download | |
| Typically uses HDFS (Hadoop Distributed File System) and MapReduce for storage and processing. | Supports various data sources and formats (including HDFS). Offers connection from Kafka, databases using Spark’s Dataframe API |
| Data Processing Data processing icon vector isolated on white background, logo concept of Data  processing sign on transparent background, black filled symbol Stock Vector  | Adobe Stock | |
| Slower processing | Faster processing |
| Primarily uses the batch processing model, MapReduce | Utilizes in-memory processing, taking advantage of RDDs (Resilient Distributed Datasets) |
| Involving mapping data, reducing it, shuffling it and sorting the results. | Supports various operations including SQL, machine learning and streaming. |
| Data Transformation and AnalysisData Analytics Search Business Money Svg Png Icon Free Download (#563461) -  OnlineWebFonts.COM | |
| MapReduce’s batch processing makes real-time analytics challenging | Better real-time analytics due to in-memory processing capabilities |
| Suitable for historical data analysis | Suitable for both historical and real-time processing |
| PerformancePerformance icon - download free icons | |
| Generally slower iterative computations and interactive data analysis | Performs better for iterative tasks, machine learning algorithms and interactive analysis |
| More reliance on disk I/O | In-memory computing model |

|  |  |
| --- | --- |
| Hadoop Ecosystem - Analytics Vidhya | Apache Spark - Wikipedia |
| Ecosystemecosystem" Icon - Download for free – Iconduck | |
| Well-Established ecosystem | Expanding ecosystem |
| Have tools including Hive, Pig, and Hbase. | Integrating various tools like Spark SQL, Spark Streaming, MLlib (machine learning library), and GraphX (graph processing) |

(*Apache Hadoop*, n.d.)

(*Apache SparkTM - Unified Engine for Large-scale Data Analytics*, n.d.)

Conclusion

Both Hadoop and Apache Spark focus on big data processing but both their approaches in the data pipelining process differ significantly.

Due to Hadoop relying on HDFS and MapReduce, Hadoop suits batch processing a lot more but tends to lag in real-time analytics due to disk I/O.

Spark on the other hand offers faster performance and support for real-time analytics due its in-memory computing and its versatile ecosystem suits it for various data processing tasks.

**References**

Chittora, D., & Chittora, D. (2023, March 11). *Hadoop vs Spark: Major Differences Explained*. Intellipaat. <https://intellipaat.com/blog/hadoop-vs-spark-choosing-the-right-big-data-software/>

*Apache Hadoop*. (n.d.). <https://hadoop.apache.org>

*Apache SparkTM - Unified Engine for large-scale data analytics*. (n.d.). <https://spark.apache.org>