Time Complexity Comparison for Hadoop and Spark

**Problem Statement**

The objective of this project is to determine the minimum and maximum temperatures of cities in European countries by scraping the data using two big data processing frameworks: Hadoop and Spark. The goal is to analyze and compare the efficiency, execution time, and resource utilization of both frameworks when handling large-scale temperature data. By evaluating these implementations, we aim to determine the most optimal framework for temperature-based data processing.

**Execution Time Results**

* Hadoop Execution Time: 1.12 seconds
* Spark Execution Time: 0.0042 seconds
* Speedup Factor (Hadoop/Spark): ~265x faster

**Key Observations**

1. Spark is significantly faster (265 times) compared to Hadoop.
   * This is because Spark processes data in-memory, while Hadoop relies on disk-based processing.
   * Hadoop’s architecture involves multiple read/write cycles to HDFS, introducing overhead.
2. Hadoop’s slower execution is due to its batch processing nature, while Spark utilizes in-memory computations for better efficiency.
3. Real-World Implications:
   * For large-scale data processing tasks requiring high speed (e.g., real-time analytics), Spark is the better choice.
   * If the use case involves storing and processing huge datasets in a distributed environment, Hadoop remains relevant.

Submitted by,

Sadeck Jaffal ,

Reshma Banu Rasool