Apache Hadoop vs Spark

Index no. - 248233U

J. A. A. M. JAYAWEERA

CS5229 - Big Data Analytics Technologies

ABOUT ME

Hey!

I'm Amesh Jayaweera

B.Sc. (Hons) in IT (UOM)

Software Engineer @ Sysco LABS

INTRODUCTION TO MAPREDUCE

MapReduce is a programming model and processing framework for processing large datasets in parallel across a distributed cluster.
Developed by Google and popularized by Apache Hadoop.

How does it work?

Input data is divided into smaller chunks and processed in parallel across multiple nodes in a cluster.

Two main phases:

1. Map phase :

- Applies a function to each input key-value pair and generates intermediate key-value pairs.

2. Reduce phase :

- Aggregates and processes the intermediate key-value pairs to produce the final output.

INTRODUCTION TO APACHE SPARK

Apache Spark is an open-source, distributed computing system for big data processing and analytics.

Developed at UC Berkeley's AMPLab and later donated to the Apache Software Foundation.

Key features:

In-memory computation:

- Utilizes in-memory caching to speed up iterative and interactive computations.

DAG execution engine:

- Optimizes task execution through a directed acyclic graph of operations.

Wide range of APIs:

- Supports multiple programming languages including Scala, Java, Python, and R.

Unified platform:

- Integrates various modules for batch processing, streaming, SQL, machine learning, and graph processing.

DEMOSTRATION

EASE OF USE:

MapReduce vs Apache Spark

MapReduce

- Requires developers to write more low-level code for each stage of processing.
- Complex programming model with explicit handling of map and reduce functions.
- Steeper learning curve, especially for developers new to distributed computing.

Apache Spark

- Offers a more intuitive and higher-level API, reducing the amount of boilerplate code needed.
- Provides a wide range of built-in higher-level abstractions like DataFrames and Datasets.
- Spark's APIs are generally more developer-friendly and easier to learn, especially for those familiar with functional programming.

FAST PROCESSING
MapReduce vs Apache Spark

MapReduce

- Disk-based processing, leading to slower performance due to frequent disk I/O operations.
- Limited in-memory caching capabilities, impacting the speed of iterative algorithms.
- Generally slower for iterative and interactive processing tasks.

Apache Spark

- Leverages in-memory computing for faster data processing, especially for iterative algorithms.
- Optimized task execution through DAG (Directed Acyclic Graph) engine.
- Offers superior performance, particularly for iterative and interactive workloads, compared to MapReduce.

CONCLUSION

- MapReduce requires more low-level coding and has slower processing speed due to disk-based operations.
- Apache Spark provides a more user-friendly API and significantly faster processing speed, primarily due to its in-memory computing capabilities.

THANK YOU!