MapReduce Vs. Apache Spark

STUDENT NO :239356X

STUDENT NAME: T.M.D.SAUMYA

Introduction to MapReduce

- MapReduce is a programming engine for processing and generating large data sets with a parallel, distributed algorithm on a cluster of the computer.
- MapReduce is composed of several components, including :
  - ***JobTracker*** *— The master node that manages all jobs and resources in a cluster*
  - ***TaskTrackers*** *— Agents deployed to each machine in the cluster to run the map and reduce tasks*
  - ***JobHistoryServer*** *— A component that tracks completed jobs, and is typically deployed as a separate function or with JobTracker*

Introduction to Apache Spark

- Apache Spark is an open-source data processing engine built for fast performance and large-scale data analysis. It uses RAM for caching and data processing .
- Apache spark consists of five main components –
  - ✓ Spark core,
  - ✓ Spark SQL (to gather information about structured data),
  - ✓ Spark streaming (for live data streams),
  - ✓ ML Library,
  - ✓ GraphX.

Demonstration

Comparison Of MapReduce & Apache Spark

1. **Ease of Use**

- Spark code is easy to write and maintain compared to MapReduce pipeline writing with more verbose and length.
- Spark does not need any abstraction where as MapReduce needs abstractions

2. **Fast Processing**

- Spark enables applications running on Hadoop to run up to 100x faster in memory and up to 10x faster on disk. But MapReduce doesn't provide the capability of fast processing compared to Spark and Map Reduce can be used for generating reports that help find answers to historical queries.
- Programmers can modify the data in real time through Spark streaming in comparison MapReduce allow to process batch of stored data.

Conclusion

- Map Reduce is ideal for linear data processing and batch processing, while Spark is best suited for projects requiring live unstructured data streams and real-time data processing.
- The decision on choosing one over the other depends on various aspects. For instance, Spark is potentially 100 times faster than MapReduce in terms of speed. In comparison, MapReduce is less expensive than Spark when it comes to cost.
- The decision is with the users to pick the technology depending on the requirement .

Thank You !