CS5229

Big Data Analytics Technologies

Senduran R. 239357C

Introduction to MapReduce and Spark

Big Data?
Big data is primarily defined by the volume of a data set or complying to 4V. Traditional databases won't be able to process them. Therefore we need Big Data frameworks.

Frameworks like

- **MapReduce**
- **Apache Spark**

They help us to get deep insights into this huge amount of **structured, unstructured, and semi-structured** data.

Hadoop and Spark, both developed by the Apache Software Foundation, are widely used open-source frameworks for big data architectures. Each framework contains an extensive ecosystem of open-source technologies that **prepare, process, manage and analyze big data sets**

Apache Hadoop MapReduce

MapReduce was developed in 2006. It also processes structured and unstructured data that are **stored in HDFS**. Hadoop MapReduce is designed in a way to process a large volume of data on a cluster of commodity hardware. Also MapReduce can process data in **batch mode**

The MapReduce programming model is based on the idea of **dividing a large data set into smaller chunks and processing** them in parallel on the cluster of computers

Apache Spark

Apache Spark was originally developed in 2009 at UC Berkeley. It provides a faster and more general purpose data processing engine. Spark is basically designed for fast computation. It also covers a wide range of workloads for example batch, interactive, iterative and streaming

Spark's core abstraction is the **Resilient Distributed Dataset (RDD)**, a fault-tolerant collection of elements that can be processed in parallel across a cluster. RDDs can be created from Hadoop Distributed File System (HDFS) files or any other data source, and can be cached in memory for faster access

Demo Steps

Step 1 : Create EMR cluster in aws (Hadoop + Hive , Spark)

Step 2 : Upload the csv file to S3

Step 3 : Connect to EMR cluster master node,

Step 4 : Load data into hive , spark from s3

Step 5 : Run SQLs

Step 6 : Collect Stats.

Comparing MapReduce & Spark

1. Easy of Use

Apache spark provides APIs for Scala, Java and Python and Spark SQL for SQL users.

Apache Spark basic building blocks that enable users to create their own custom functions with ease

Hadoop MapReduce was created using Java and can be challenging to program. Unlike Apache Spark, it doesn't have an interactive mode. However, **Hive** has a command-line interface that allows users to issue commands through the hive cli.

Comparing MapReduce & Spark

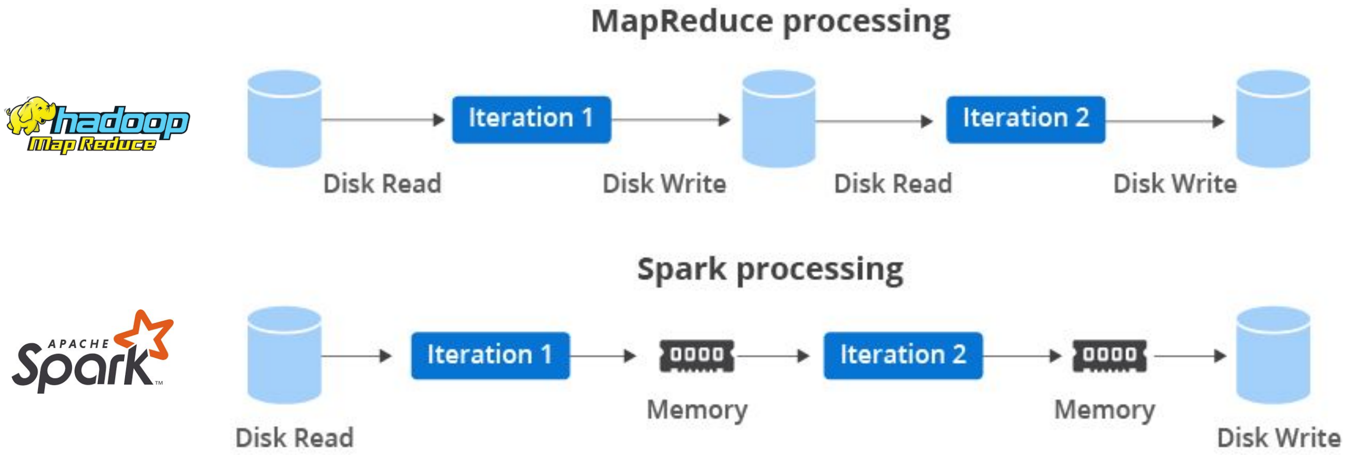
1. Fast Process

Apache Spark is benchmarked to execute batch processing jobs nearly 10 to 100 times faster than the hadoop mapreduce framework by cutting down the number of reads and writes to disk.

MapReduce, there are these Map and Reduce tasks, after which there is a synchronization barrier, and one needs to preserve the data in the disc

Apache Spark, the concept of RDDs (Resilient Distributed Datasets) lets you save data on memory and preserve it to the disc if and only if it is required, and it does not have any synchronization barriers that could slow down the process. Thus the general execution engine of Spark is much faster than Hadoop MapReduce with memory

Data Processing of MapReduce & Spark



As we can see, MapReduce involves at least 4 disk operations whereas Spark only involves 2 disk operations. This is one reason for Spark is much faster than MapReduce

(https://www.knowledgehut.com/blog/big-data/apache-spark-and-mapreduce-comparison)