SPARK

Systems are working with massive amounts of data in petabytes or even more and it is still growing at an exponential rate. Big data is present everywhere around us and comes in from different sources like social media sites, sales, customer data, transactional data, etc. And I firmly believe, this data holds its value only if we can process it both interactively and faster.

Apache Spark is an open-source, fast cluster computing system and a highly popular framework for big data analysis. This framework processes the data in parallel that helps to boost the performance. It is written in Scala, a high-level language, and also supports APIs for Python, SQL, Java and R.

Apache Spark is a unified analytics engine for large-scale data processing including built-in modules for SQL, streaming, machine learning and graph processing. Unlike MapReduce, it can deal with real-time processing. It is an open-source framework used for faster data processing.

Apache Spark is a lightning-fast cluster computing designed for fast computation. It was built on top of Hadoop MapReduce and it extends the MapReduce model to efficiently use more types of computations which includes Interactive Queries and Stream Processing.

Apache Spark is an integrated processing engine that can analyze big data using SQL, graph processing, machine learning, or real-time stream analysis.

Apache Spark has Spark SQL as its very own query language. Spark is a fast and general processing engine compatible with Hadoop data. It can run in Hadoop clusters through YARN or Spark's standalone mode, and it can process data in HDFS, HBase, Cassandra, Hive. Many organizations run Spark on clusters of thousands of nodes.

Apache Hadoop framework is divided into two layers.

- Hadoop Distributed File System (HDFS)

- Processing Layer (MapReduce)

Apache Spark is a big data solution that has been proven to be easier and faster than Hadoop MapReduce. Apache Spark is a lightning-fast and cluster computing technology framework, designed for fast computation on large-scale data processing. Apache Spark is a distributed processing engine but it does not come with inbuilt cluster resource manager and distributed storage system. You have to plug in a cluster manager and storage system of your choice. Apache Spark consists of a Spark core and a set of libraries similar to those available for Hadoop. The core is the distributed execution engine and a set of languages. Apache Spark supports languages like Java, Scala, Python and R for distributed application development.

Additional libraries are built on top of the Spark core to enable workloads that use streaming, SQL, graph and machine learning. Apache Spark is data processing engine for batch and streaming modes featuring SQL queries, Graph Processing, and Machine Learning. Apache Spark can run independently and also on Hadoop YARN Cluster Manager and thus it can read existing Hadoop data. You can choose Apache YARN or Mesos for cluster manager for Apache Spark. You can choose Hadoop Distributed File System (HDFS), Google cloud storage, Amazon S3, Microsoft Azure for resource manager for Apache Spark.

DIFFERENCE BETWEEN MAPREDUCE VS APACHE SPARK

- MapReduce is a Disk-Based Computing while Apache Spark is a RAM-Based Computing. MapReduce is strictly disk-based while Apache Spark uses memory and can use a disk for processing.

- The primary difference between MapReduce and Spark is that MapReduce uses persistent storage and Spark uses Resilient Distributed Datasets.

- Hadoop MapReduce is meant for data that does not fit in the memory whereas Apache Spark has a better performance for the data that fits in the memory, particularly on dedicated clusters.

- Hadoop MapReduce can be an economical option because of Hadoop as a service and Apache Spark is more cost effective because of high availability memory

- Spark is able to execute batch-processing jobs between 10 to 100 times faster than the MapReduce

- Map reduce uses replication for fault tolerance. Apache Spark uses RDD and other data storage models for fault tolerance.

- Map reduce supports Sql through Hive Query Language. Apache Spark supports Sql through Spark Sql