22/12/2023

Chakradhar Bhogapurapu, BIBA

Pyspark

Apache Spark

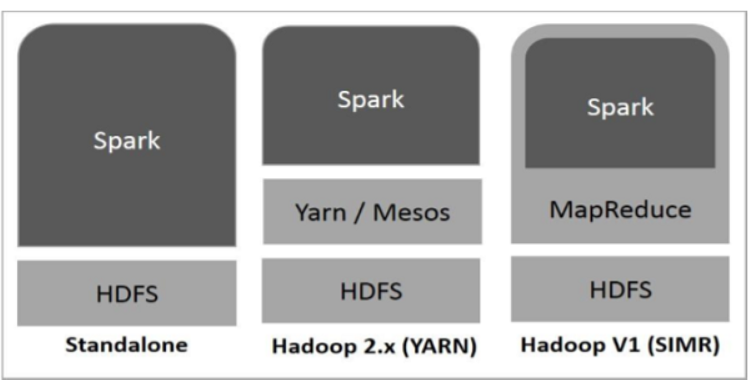
Hadoop is used to analize their data sets. Hadoop is based on map reduce. It computes large data sets. Spark was introduced by Apache Software Foundation for speeding up the Hadoop computational computing software process.

Spark is not modified version of Hadoop.

Apache spark is lightening-fast cluster computing technology. The main feature of spark is its in-memory cluster computing and increases process speed of an application. Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming.

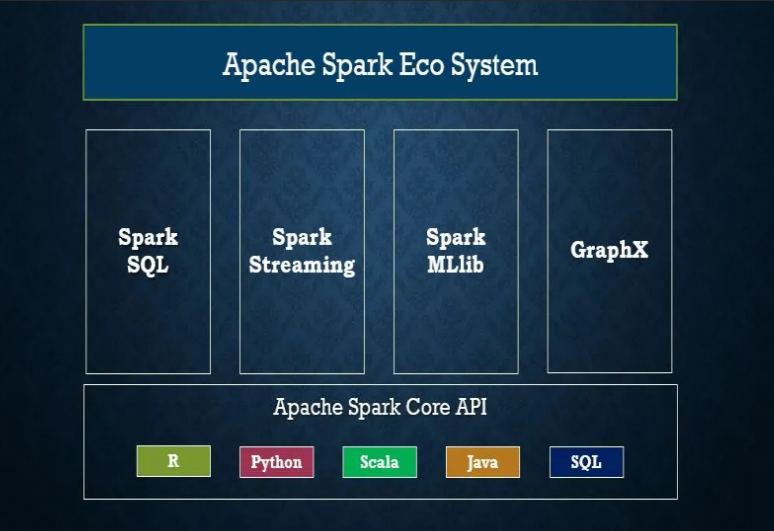
Features of apache spark

* Speed
* Support Multi Languages
* Advanced Analytics.



* **Standalone** − Spark Standalone deployment means Spark occupies the place on top of HDFS(Hadoop Distributed File System) and space is allocated for HDFS, explicitly. Here, Spark and MapReduce will run side by side to cover all spark jobs on cluster.
* **Hadoop Yarn** − Hadoop Yarn deployment means, simply, spark runs on Yarn without any pre-installation or root access required. It helps to integrate Spark into Hadoop ecosystem or Hadoop stack. It allows other components to run on top of stack.
* **Spark in MapReduce (SIMR)** − Spark in MapReduce is used to launch spark job in addition to standalone deployment.

Components of Spark



Components of Spark:

* Apache Spark Core: Spark Core is the underlying general execution engine for spark platform that all other functionality is built upon. It provides In-Memory computing and referencing datasets in external storage systems.
* Spark SQL: Spark SQL is a component on top of Spark Core that introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.
* Spark Streaming: Spark Streaming leverages Spark Core, fast scheduling capability to perform streaming analytics. It ingests data in mini-batches and performs RDD(Resilient Distributed Datasets) transformations on those mini-batches of data.
* MLlib (Machine Learning Library): MLlib is a distributed machine learning framework above Spark because of the distributed memory-based Spark architecture.
* It is, according to benchmarks, done by the MLlib developers against the Alternating Least Squares (ALS) implementations. Spark MLlib is nine times as fast as the Hadoop disk-based version of Apache Mahout (before Mahout gained a Spark interface).
* GraphX: GraphX is a distributed graph-processing framework on top of Spark. It provides an API for expressing graph computation that can model the user-defined graphs by using Pregel abstraction API. It also provides an optimized runtime for this abstraction.

Pyspark Apache RDD

It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

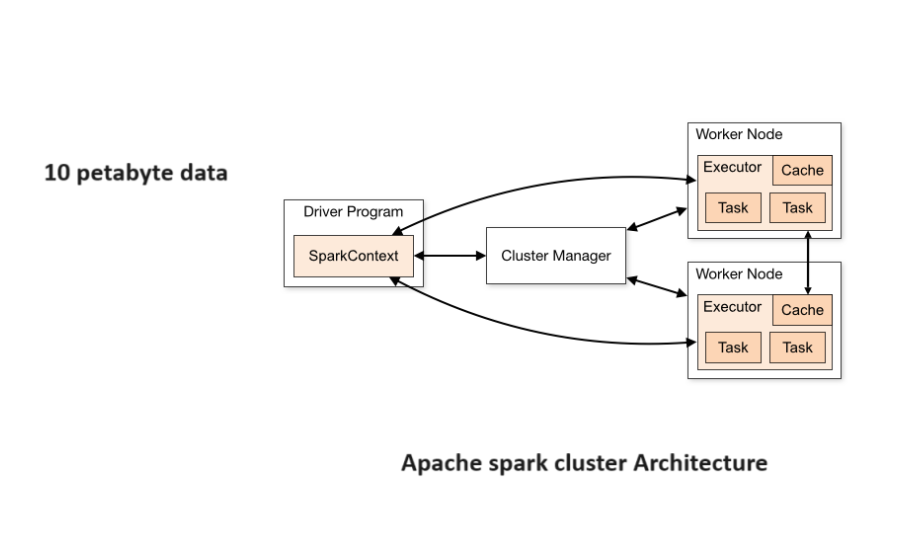
There are two ways to create RDDs – parallelizing an existing collection in your driver program, or re in an external storage system, such as a shared file system, HDFS, HBase, or any data source offering a Hadoop Input Format.

How spark streaming works

* Gathering
* Processing
* Data Storage

Data sharing is slow in map reduce

Architecture



Features of PySpark:

 In-memory computation

 Distributed processing using parallelize

 Can be used with many cluster managers (Spark, Yarn, Mesos e.t.c)

 Fault-tolerant

 Immutable

 Lazy evaluation

 Cache &amp; persistence

 Inbuild-optimization when using DataFrames

 Supports ANSI SQL

Advantages:

 PySpark is a general-purpose, in-memory, distributed processing engine that allows you to process data efficiently in a distributed fashion.

* Applications running on PySpark are 100x faster than traditional systems
* You will get great benefits from using PySpark for data ingestion pipelines.
* Using PySpark we can process data from Hadoop HDFS, AWS S3, and many file
* systems.
* PySpark also is used to process real-time data using Streaming
* Using PySpark streaming you can also stream files from the file system and also stream from the socket.

PySpark natively has machine learning and graph libraries.