PYSPARK

ASSESSMENT

S.R.TAANUSRI

22.12.2023

Introduction to Apache Spark

* Industries rely on Hadoop for data analysis due to its:
* Scalability
* Flexibility
* fault tolerance
* cost-effectiveness.
* Framework based on the MapReduce programming.
* Main concern is the need to maintain speed in processing large datasets.
* Apache Spark was introduced by the Apache Software Foundation to enhance the computational capabilities of Hadoop.
* Spark is not a modified version of Hadoop.
* Spark has own cluster management and can use Hadoop for storage purposes only.
* The main feature of Spark is its lightning-fast, in-memory cluster computing.

Evolution of Apache Spark

* Spark originated as a subproject of Hadoop in 2009 at UC Berkeley's AMPLab, developed by Matei Zaharia.
* It was open-sourced in 2010,
* In 2013, it was donated to the Apache Software Foundation, becoming a top-level Apache project in February 2014.

Features of Apache Spark

1. Speed:

* Spark run applications 100 times faster in memory and 10 times faster on disk
* achieved by minimizing read/write operations to disk and storage.

1. Language Support:

* Allows developers to write applications in different languages.
* It also offers 80 high-level operators for interactive querying.

1. Advanced Analytics:

* Spark Supports :
* batch applications,
* iterative algorithms,
* interactive queries,
* streaming,
* machine learning (ML),
* graph algorithms.



Spark Deployment

There are three ways to deploy Spark:

1. Standalone:

Occupies a place on top of Hadoop Distributed File System (HDFS), running alongside MapReduce to cover all Spark jobs on the cluster.

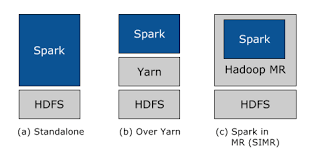
1. Hadoop Yarn:

Spark runs on Yarn without requiring pre-installation or root access,

1. Spark in MapReduce (SIMR):

Used to launch Spark jobs in addition to standalone deployment.

Allows users to start Spark and use its shell without administrative access.



Components of Spark

1. Spark Core:

General execution engine providing in-memory computing and referencing datasets in external storage systems.

1. Spark SQL:

Introduces a new data abstraction called SchemaRDD.

1. Spark Streaming:

Performs streaming analytics by applying RDD transformations.

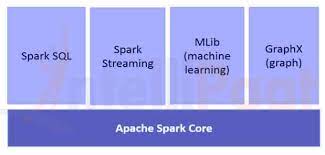
1. MLlib (Machine Learning Library):

A distributed machine learning framework above Spark, offering faster performance than Hadoop-based alternatives.

1. GraphX:

A distributed graph-processing framework on top of Spark.

Apache Spark –



RDD (Resilient Distributed Datasets)

* RDD is a fundamental data structure in Spark
* Immutable distributed collection of objects
* Support fault-tolerant parallel.
* Spark utilizes RDDs to achieve faster and more efficient MapReduce operations.

Data Sharing using Spark RDD

* Data sharing in MapReduce is slow due to
* Replication
* Serialization
* disk IO
* Spark addresses this issue with Resilient Distributed Datasets (RDD)
* Data sharing 10 to 100 times faster than network and disk.

Iterative and Interactive Operations on Spark RDD

* Iterative Operations:
* Intermediate results are stored in distributed memory
* Enhance system speed.
* If memory is insufficient, results are stored on disk.
* Interactive Operations:
* Data for repeated ad-hoc queries on the same subset
* faster execution times