Date: -22/12/2023

Topics:

* Apache Spark: Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.
* Use: PySpark is very well used in the Data Science and Machine Learning community as there are many widely used data science libraries written in Python including NumPy, and TensorFlow. Also used due to its efficient processing of large datasets. PySpark has been used by many organizations like Walmart, Trivago, Sanofi, Runtastic, and many more.
* Features:
* 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 & 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 and Kafka.
* 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
* Cluster Manager type:
* Standalone – a simple cluster manager included with Spark that makes it easy to set up a cluster.
* Apache Mesos – Mesons is a Cluster manager that can also run Hadoop MapReduce and PySpark applications.
* Hadoop YARN – the resource manager in Hadoop 2. This is mostly used as a cluster manager.
* Kubernetes – an open-source system for automating deployment, scaling, and management of containerized applications.
  + **Modules & packages**
* PySpark RDD (pyspark.RDD)
* PySpark DataFrame and SQL (pyspark.sql)
* PySpark Streaming (pyspark.streaming)
* PySpark MLib (pyspark.ml, pyspark.mllib)
* PySpark GraphFrames (GraphFrames)
* PySpark Resource (pyspark.resource) It’s new in PySpark 3.0

