Spark working

→ It has a small code base and the system is divided into several layers and the layers are independent to each other.

Apache Spark :-

→ It is an open-source which is used for large-scale data processing.

→ Spark is fast, flexible and easy to use.

→ Spark can run on single-node machines or multi-node machine (clusters)

→ Apache Spark can also Process real-time streaming.

→ It provides libraries like Java, Scala, Python etc

By Spark features

① In-memory computation
② Distributed processing using parallelize
③ used with many cluster managers
④ fault-tolerant
⑤ Immutable
⑥ Lazy evaluation
⑦ Cache & persistence
⑧ In built optimization
⑨ Supports ANSI SQL

Advantages of Pyspark

→ It is distributed processing engine that allows us to process data efficiently

→ It is 100X faster :

→ It has better benefits for data ingestion pipelines

→ We can process data from Hadoop HDFS, AWS S3 etc

→ Process real-time data using streaming and (kafka)

Modules & Packages

→ Pyspark RDD (pyspark.RDD)

→ Pyspark DataFrame SQL (pyspark.sql)

→ Pyspark Streaming (pyspark.streaming)

→ Pyspark MLib (pyspark.ml, pyspark.mllib)

→ Pyspark GraphFrames (GraphFrames)