MANDAPALLI HEMA

ASSESSMENT-17

APACHE SPARK:-

Industries are using Hadoop extensively to analyze their data sets.Hadoop introduce a new way to data processing called distributed processing.Instead of single machine we can use multiple computers to get a multiple computers to get a final result.

Key Components:-

\*Hadoop distributed file system.

\*Mapreduce-it is a smart way of processing all of data together. Mapreduce helps in processing all of the data in parallel.

Apache Spark is a lightning-fast cluster computing technology, designed for fast computation.RDD is the backbone of the apache spark.it allows data to be stored in memory and enables faster data access and processing.instead of writing of writing and reading from the disk.

Spark also gave the ability to write code in various programming languages such as python,java ans scala.

Components of Spark:Apache Spark consists of Spark Core Engine, Spark SQL, Spark Streaming, MLlib, GraphX and Spark R.

SPARKCORE:-

Spark Core is the heart of the Apache Spark framework. Spark Core provides the execution engine for the Spark platform which is required and used by other components which are built on top of Spark Core as per the requirement.

SPARK SQL:-

It ingests data in mini-batches and performs RDD transformations on those mini-batches of data.

SPARK STREAMING:-

This is a very popular Spark library as it takes Spark’s big data processing power and cranks up the speed. Spark Streaming has the ability to Stream gigabytes per second.

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.

GRAPHX:

It provides an API for expressing graph computation that can model the user-defined graphs by using Pregel abstraction API.

APACHE SPARK ARCHITECTURE:

The Apache Spark framework uses a master-slave architecture that consists of a driver, which runs as a master node, and many executors that run across as worker nodes in the cluster. Apache Spark can be used for batch processing and real-time processing as well. Learn Pyspark from industry experts