**Cloud Computing for Data Analysis**

**VIDEO CASE 04 : Spark**

**Video Case Questions:**

1. What is Spark?

**Answer:**

Apache Spark is a general-purpose distributed data processing engine which is open source and was developed in UC Berkeley.

Libraries for machine learning, SQL, graph computation and stream processing are available on top of Spark; these can be used together in an application. Java, Scala, R and Python are the programming languages supported by Spark. Spark is incorporated (by data scientists) into several applications to rapidly analyze, query, and transform data.

Although it’s built on top of HDFS, it’s not tied to MapReduce’s two-stage implementation. It’s also, 100 times faster than MapReduce in case of some applications. also provides primitives for in-memory cluster computing.

Tasks that are frequently associated with Spark are: IoT, machine learning, processing of streaming data from sensors, ETL and SQL batch jobs across large data sets.

1. What are all the layers or packages that come along with Spark? And what they are used for?

**Answer:**

Following are the Spark layers and their usage:

1. Spark SQL: Spark SQL is Spark’s package for working with structured data. Allows querying data via SQL and HQL.
2. MLlib: Provides several types of machine learning algorithms, including regression, classification, clustering, and collaborative filtering, as well as supporting functionality such as model evaluation and data import.
3. Spark Streaming: Enables processing of live streams of data.
4. GraphX: Library for manipulating graphs and performing graph-parallel computations.
5. Why does the Spark run faster than Hadoop?  
   **Answer:**Spark runs in-memory on the cluster and is not tied to Hadoop’s MapReduce two-stage paradigm. On the other hand, Hadoop MapReduce persists back to the disk after a map or reduce action. Because of this, in case of Spark, repeated access to the same data is much faster. Spark can run as a standalone or on top of Hadoop YARN, where it can read data directly from HDFS. Also, Spark ecosystem has established a versatile stack of components to handle SQL, ML, Streaming, Graph Mining task on the other hand, all the packages have to be installed in Hadoop.