**Name: Nikhil Jadhav**

**Student ID: 801075504**

**Cloud Computing for Data Analysis**

**VIDEO CASE 04 : Spark**

Watch following videos:

**Video 1:** <https://youtu.be/PiJGa26OHFM>

**Video 2:** <https://youtu.be/eMGjuK-Pk9g>

**Video Case Questions:**

1. What is Spark?
2. What are all the layers or packages that come along with Spark? And what they are used for?
3. Why does the Spark runs faster than Hadoop?

Answers:

1. Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. With Spark running on Apache Hadoop YARN, developers everywhere can now create applications to exploit Spark’s power, derive insights, and enrich their data science workloads within a single, shared dataset in Hadoop. Spark can give performance 100 times faster compared to that of Hadoop for certain applications. Also, spark is not tied to two stage map-reduce paradigm like Hadoop.
2. Following are the layers that come along with Spark
3. Spark SQL: Spark SQL is a Spark module for structured data processing. Unlike the basic Spark RDD API, the interfaces provided by Spark SQL provide Spark with more information about the structure of both the data and the computation being performed. Internally, Spark SQL uses this extra information to perform extra optimizations.
4. Spark Streaming: Spark Streaming is an extension of the core SparkAPI that enables scalable, high-throughput, fault-tolerant stream processing of live data streams. Spark Streaming provides a high-level abstraction called discretized stream or DStream, which represents a continuous stream of data.
5. MLlib: MLlib is Spark's scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as underlying optimization primitives, as outlined below: summary statistics.
6. GraphX: GraphX is a new component in Spark for graphs and graph-parallel computation. At a high level, GraphXextends the Spark RDD by introducing a new Graph abstraction: a directed multigraph with properties attached to each vertex and edge.
7. Spark perform the data processing in in-memory, while Hadoop requires series of access to data on the disk after the map or reduce actions.