**NAME : SAI KRISHNA MANNAVA**

**STUDENT ID : 801136361**

**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?

ANSWER:

* Spark is a data analytics cluster computing framework
* It is built on the top of HDFS.
* It is developed using java, SCALA and python in AMPlab at UC Berkeley.
* It is a fast and general engine for large scale data processing.
* It performs 100 times faster than the Hadoop map reducer functions.
* It allows the users to upload the data into the cluster, so that they can be queried repeatedly and makes it run faster.

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

ANSWER:

* Spark-SQL- Package that works with structured data and HQL as its feature and can work with SQL interface rather than java, python etc.
* MLLib(machine learning)- **It** is **Spark's** machine learning (ML) library. At a high level, it provides tools such as common learning algorithms such as classification, regression, clustering, and collaborative filtering
* GraphX(graph)- A library that aides in controlling diagrams, performing chart level calculations and manipulating graph
* Spark-Streaming- Streaming provides API for manipulating data streams and also enables to process livestreaming data.

1. Why does the Spark run faster than Hadoop?

The Hadoop map reducer processes the data present on the disk while the Spark processes data present In-Memory. It is the main reason for which the spark can beat Hadoop effectively in data processing. Therefore, it can run faster than Hadoop.