**Cloud Computing for Data Analysis**

**VIDEO CASE 04: Spark**

**Student Name: Balasundaram Avudai Nayagam**

**Student ID: 801104903**

Watch following videos:

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

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

**Video Case Questions:**

1. What is Spark?
2. - Spark is a cluster based Data Analytics tool developed by AMP Labs at UC Berkeley.

* Spark is primarily used for large scale processing of data with high efficiency.
* Spark comes with higher-level libraries, like support for SQL queries, streaming data, machine learning and processing. These libraries increase developer productivity and can be seamlessly combined to create complex workflows.
* Spark framework features an interface for programming complete clusters with implicit data parallelism and fault tolerance which makes it more efficient than Map Reduce operation.

1. What are all the layers or packages that come along with Spark? And what they are used for?
2. Layers with spark:
3. **Spark SQL**: It is an interface provided for queries to be executed.
4. **Graph X**: It is a package which performs ETL (Extract, Transform and Load), exploratory data analysis and iterative processing.
5. **Spark Streaming**: It enables continuous stream of data resulting in high throughput for processing live data streams.
6. **MLlib**: It consists of set of Machine Learning packages that can be implemented in Hadoop.
7. Why does the Spark runs faster than Hadoop?
8. - Spark is implemented in Scala which is comparatively faster than other programming languages such as Java.

* Spark does not follow the Map – Reduce two step mechanism since it is built on top of Hadoop File System.
* The data processing is done within the memory in case of Spark where as in Map- reduce, the data is transferred to the disk once the processing is done.
* It supports parallel processing which makes its operation to be performed faster.