**Cloud Computing for Data Analysis**

**VIDEO CASE 04: Spark**

|  |
| --- |
| **Submitted by:** Abdullah Al Raqibul Islam  **UNCC ID:** 801151189 |

Watch following videos:

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

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

**Video Case Questions:**

1. What is Spark?

|  |
| --- |
| Apache Spark is an open-source, highly fault tolerant data analytics and cluster computing framework which runs on top of HDFS. In contrast with Hadoop MapReduce, Apache Spark is not tied to two-stage of MapReduce paradigm. For certain applications, it performs 100X faster than Hadoop MapReduce. Spark provides primitives for in-memory cluster computing, meaning, it allows user programs to load data into a cluster’s memory, so that, they could be queried repeatedly. This makes Apache Spark well suited for machine learning algorithm. |

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

|  |
| --- |
| Apache Spark packages,   * Spark SQL: SQL interface, similar to Hive SQL used in Apache Hadoop ecosystem. * Spark Streaming: Allows processing of live data stream. * MLlib: Machine learning part like Mahout in Apache Hadoop. * GraphX: Allows graph processing and parallel execution on graph. |

1. Why does the Spark runs faster than Hadoop?

|  |
| --- |
| Apache Spark is faster than Hadoop because Spark does all the processing in-memory whereas the Hadoop MapReduce is not doing like that.  Hadoop MapReduce persists the full dataset to HDFS after running each job, which is very expensive. Spark overcome this by utilizing the technique of operation pipeline, where it can feed the output of one operation into another one without writing it back to a persistent storage. As Spark relies on in-memory processing, it can utilize the in-memory caching as well. |