**Cloud Computing for Data Analysis**

**VIDEO CASE 03: Cloud Tools – Pig, Hive and HBase**

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

Watch following videos:

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

**Video 2:** <https://youtu.be/uY7Rr7ru9E4>

**Video 3:** <https://youtu.be/kN01ELCAsn8>

Video 1, 2 and 3 gives you a basic knowledge about Pig, Hive and HBase respectively

**Video Case Questions:**

1. Briefly explain the execution steps followed by Pig.

|  |
| --- |
| * Load the input data from HDFS   + A = LOAD ‘datafile.txt’; * Run the program through a set of transformation, which under the cover translated to a set of map and reduce tasks * Dump/store the results while execution done   + DUMP C;   + STORE C into ‘Results’; |

1. What is the purpose of Hive? Mention some of the advantages of Hive.

|  |
| --- |
| Purpose of Hive: Though the Pig is quite powerful and easy to use, the downside is it has a learning curve to be master on that. It’s easier than writing MapReduce program, but it’s not easy as like SQL is. Hive comes to fill this gap. Hive is a Hadoop runtime component that allows anyone already mastered in SQL to utilize the Hadoop platform without an additional learning curve.  Some major advantages of using Hive is:   * As mentioned earlier, it is easy to use Hive due to it’s similarity with SQL. * It is possible run the Hive queries in a few different ways (i.e. command line interface named HiveShell, JDBC/ODBC driver, Hive Thrift Client etc), which brings elasticity for the developers who want to utilize Hadoop for their application. * The Hive Thrift client can be used with applications written in C++, Java, PHP, Python, Ruby etc. This means, developers from a diversified domain can utilize the power of Hadoop in their application with a very small change. |

1. Give some similarities of architectures of HBase and HDFS and MapReduce.

|  |
| --- |
| HBase follows master-slave architecture. Just like HDFS which has a NameNode and SlaveNode, and MapReduce that has job tracker and task tracker slaves, HBase is built on similar concepts. In HBase, the master node manages the cluster and resolves servers store portion of the tables and perform the work on the data. Similar to HDFS’s concern to the availability of NameNode, HBase is also sensitive to the loss of it’s master node. |