**Cloud Computing for Data Analysis**

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

**Video Case Questions:**

1. Briefly explain the execution steps followed by Pig.

**Answer:**

Following execution steps are followed by Pig:

1. **LOAD:** The load command takes the data file on which the operations are to be performed as an input. This input file could be a directory or an HDFS file. If the input file specified is a directory, all the files in that directory are loaded. But, if the file is not in the native Pig format, then the USING command can be used in tandem with the LOAD command. The USING command can point to a custom program which has the code to read the specific file.
2. **Transformation Logic:** This execution step is where data manipulation takes place. It makes use of the following commands:
3. GROUP
4. FILTER

This way, unnecessary rows can be filtered out, two datasets can be joined, data can be grouped to build aggregations.

1. **DUMP / STORE:** To output data to the screen, the DUMP command is used. The STORE command can be used to store all the data in HDFS format. The stored data file can further be used for analysis/processing.
2. What is the purpose of Hive? Mention some of the advantages of Hive. **Answer:**  
   The purpose of Hive is to leverage the Hadoop platform using the existing SQL syntax It allows usage of Hive Query Language aka HQL statements which are pretty similar to SQL statements.

HQL queries are broken down in to Map-Reduce operations. We can implement Hive through:

1. Hive Shell
2. JDBC/ODBC Client
3. Hive Thrift Client

**Advantages of Hive:**

* Provides easy and simple data summarization, query, and analysis.
* Supports partitioning of data at the level of tables for performance improvement.
* Has a rule-based optimizer for optimizing logical plans.
* Supports external tables which make it possible to process data without actually storing in HDFS.
* Apache Hive fits the low-level interface requirement of Hadoop perfectly.

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

**Answer:**🡪Just like MapReduce, HBase applications are written in Java. However, they can be written using THRIFT, REST APIs. It is not a relational DBMS.   
🡪HBase is sensitive to the loss of the master node.

🡪Similar to HDFS, HBase is very scalable as it allows adding columns at any time.

🡪Like HDFS and MapReduce, HBase has master node which is responsible to manage the cluster, and the slave nodes do the work.