**ACD\_BDD\_Session\_7\_Assignment\_1\_Main**

**Problem Statement**

Give a brief answers to the questions below:

**1. Why Map-reduce program is needed in Pig Programming?**

Answer: Pig is an application that works on top of MapReduce or Yarn and abstracts the java mapreduce job. It is written in java. Pig translates the Pig Latin script into MapReduce so that it can be executed within YARN for access to a single dataset stored in the Hadoop Distributed File System (HDFS)

**2. What are advantages of pig over MapReduce?**

Answer:

* Development cycle is long in mapreduce
* Pig Latin uses a lot fewer lines of code than the Java MapReduce script.
* It is easier for someone without Java knowledge to understand the Pig Latin script
* It is difficult to join large datasets in mapreduce. Pig uses sql like query which helps in performing various tasks like union, filter by, group by etc easily and faster.

**3. What is pig engine and what is its importance?**

Answer: Pig Engine is a component of Apache Pig which acts as interpreter between Pig Latin and mapreduce jobs. It converts Pig Latin scripts into a series of MapReduce jobs.

**4. What are the modes of Pig execution?**

Answer: Local mode and MapReduce/Hadoop mode.

**5. What is grunt shell in Pig?**

Answer: Grunt shell is used to execute Pig latin programs in Interactive mode.

**6. What are the features of Pig Latin language?**

Answer:

* Pig latin includes operators for many of the traditional data operations.
* Pig latin is extensible so that users can develop their own functions for reading, processing and writing data.
* Pig latin script is made up of a series of operations, or transformations, that are applied to the input data to produce output.
* Pig latin programs can be executed either in Interactive mode through Grunt shell or in Batch mode via Pig latin scripts.

**7. Is Pig latin commands case sensitive?**

Answer: Yes.

**8. What is a data flow language?**

Answer: Dataflow languages are ones that focus on the state of the program and cause operations to occur according to any change in the state. Dataflow languages are inherently parallel; because the operations rely on inputs that when met will cause the operation to execute. This means unlike a normal program where one operation is followed by the next operation, in dataflow program operations will execute as long as the inputs are met and thus there is no set order.

Pig being a dataflow language can easily process conditions, loops, jumps thus processing data in more efficient manner.