**ASSIGNMENT 9.1**

Give a brief answer to the questions below:

1. Why MapReduce program is needed in Pig Programming?

Ans) Pig Programming is a high level language whereas map reduce program are written in java.

In pig programming most of the functions are inbuilt so number of lines in a code is reduced to few lines .

When pig program is executed these inbuilt functions are converted in map reduce actual code.

Pig’s programming language referred to as Pig Latin is a coding approach that provides high degree of abstraction for MapReduce programming but is a procedural code not declarative.

Pig Latin code can be extended through various user defined functions that are written in Python, Java, Groovy, JavaScript, and Ruby.

Pig has tools for data storage, data execution and data manipulation.

Pig Latin is highly promoted by Yahoo as all the data engineers at Yahoo use Pig for processing data on the biggest hadoop clusters in the world.

1. What are advantages of pig over MapReduce?

ANS)

* Hadoop MapReduce is a compiled language whereas Apache Pig is a scripting language and Hive is a SQL like query language.
* Pig and Hive provide higher level of abstraction whereas Hadoop MapReduce provides low level of abstraction.
* Hadoop MapReduce requires more lines of code when compared to [Pig and Hive](https://www.dezyre.com/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79). Hive requires very few lines of code when compared to Pig and Hadoop MapReduce because of its SQL like resemblance.
* [Hadoop](https://www.dezyre.com/Hadoop-Training-online/19)MapReduce requires more development effort than Pig and Hive.
* Pig and Hive coding approaches are slower than a fully tuned Hadoop MapReduce program.
* When using Pig and Hive for executing jobs, Hadoop developers need not worry about any version mismatch.
* There is very limited possibility for the developer to write java level bugs when coding in Pig or Hive.
* Pig has problems in dealing with unstructured data like images, videos, audio, text that is ambiguously delimited, log data, etc.
* Pig cannot deal with poor design of XML or JSON and flexible schemas.

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

ANS) [Pig](http://en.wikipedia.org/wiki/Pig_%28programming_tool%29) is an application that works on top of MapReduce, Yarn or Tez. Pig is written in Java and compiles Pig Latin scripts into to MapReduce jobs. Think of Pig as a compiler that takes Pig Latin scripts and transforms them into Java.

Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with **Hadoop**; we can perform all the data manipulation operations in Hadoop using Apache Pig.

To write data analysis programs, Pig provides a high-level language known as **Pig Latin**. This language provides various operators using which programmers can develop their own functions for reading, writing, and processing data.

To analyze data using **Apache Pig**, programmers need to write scripts using Pig Latin language. All these scripts are internally converted to Map and Reduce tasks. Apache Pig has a component known as **Pig Engine** that accepts the Pig Latin scripts as input and converts those scripts into MapReduce jobs. To perform a particular task Programmers using Pig, programmers need to write a Pig script using the Pig Latin language, and execute them using any of the execution mechanisms (Grunt Shell, UDFs, Embedded). After execution, these scripts will go through a series of transformations applied by the Pig Framework, to produce the desired output.

It’s acts as interpreter between Pig Latin script and Map Reduce Jobs.It creating environment to execute Pig scripts into series of mapreduce jobs in parallel manner.

**QUES 4) What are the modes of Pig execution?**

ANS)

You can run Pig in two modes:

**MapReduce/Hadoop Mode**: Here Pig jobs run as a series of MapReduce jobs picking the input and output paths from HDFS. Note: Input file has to be copied in HDFS in case of Map reduce mode.

Type the command pig or pig –x mapreduce to run Pig in MapReduce Mode. While running pig in MapReduce mode, make sure the job history server is running. mr-jobhistory-daemon.sh start historyserver Job history server helps us to view previous MR job details

Eg: <http://localhost:19888/jobhistory/tasks/job_1462381858094_0001/r>

Local Mode: Here the entire Pig job runs as a single JVM picking the local Unix path for execution.

Note: Input file has to be kept in local file system in case of local mode. Type the command pig or pig -x local to run Pig in Local Mode.

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

ANS)

Grunt Shell

• Interactive Shell for executing Pig Commands

• Used when script file is not provided

• Can execute scripts from Grunt via run or exec commands

1. What are the features of Pig Latin language?

ANS) Pig Latin uses a lot fewer lines of code than the Java MapReduce script.

* The Pig Latin script was is easier to read for someone without a Java background.
* Pig Latin is a very simple scripting language. It has constructs which can be used to apply different transformation on the data one after another.

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

ANS)

The names (aliases) of relations and fields are case sensitive. The names of Pig Latin functions are case sensitive. The names of parameters (see [Parameter Substitution](https://pig.apache.org/docs/r0.10.0/cont.html#Parameter-Sub)) and all other Pig Latin keywords (see [Reserved Keywords](https://pig.apache.org/docs/r0.10.0/basic.html#Reserved-Keywords)) are case insensitive.

In the example below, note the following:

* The names (aliases) of relations A, B, and C are case sensitive.
* The names (aliases) of fields f1, f2, and f3 are case sensitive.
* Function names PigStorage and COUNT are case sensitive.
* Keywords LOAD, USING, AS, GROUP, BY, FOREACH, GENERATE, and DUMP are case insensitive. They can also be written as load, using, as, group, by, etc.
* In the FOREACH statement, the field in relation B is referred to by positional notation ($0).

grunt> A = LOAD 'data' USING PigStorage() AS (f1:int, f2:int, f3:int);

grunt> B = GROUP A BY f1;

grunt> C = FOREACH B GENERATE COUNT ($0);

grunt> DUMP C;

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

Ans)

In a dataflow language, you have a stream of data which is passed from instruction to instruction to be processed. Conditional execution, jumps and procedure calls route the data to different instructions. This could be seen as data flowing through otherwise static instructions like how electrical signals flow through circuits or water flows through pipes. A dataflow "if" statement would route the data to the correct branch.