Assignment 7.1

5. Problem Statement

Give a brief answers to the questions below:

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

Pig runs on Hadoop. It makes use of both the Hadoop Distributed File System, HDFS, and Hadoop’s processing system, MapReduce. That is why map reduce program is needed in Pig Programming.

2. What are advantages of pig over MapReduce?

Pig has the ability to perform computations which cannot be done by MapReduce.

Pig is a high level language. Learning and using Pig Latin is comparatively easier than learning to write mapreduce program.

Pig does not require programming language knowledge, whereas Map reduce requires programmers. Programmers should think in terms of map and reduce functions. And mostly knowledge of Java programming is required.

Number of lines of code for the same functionality in Pig Latin is much smaller than a mapreduce program.

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

Pig provides an engine for executing data flows in parallel on Hadoop. This is known as the pig engine.

4. What are the modes of Pig execution?

Local

Mapreduce

5. What is grunt shell in Pig?

It is an interactive Shell for executing Pig Commands. It is used when script file is not provided.

We can execute scripts from Grunt via run or exec commands.

6. What are the features of Pig Latin language?

1. Pig can operate on data with or without metadata.

2. Pig is not environment dependent

Pig is intended to be a language for parallel data processing. Pig is not tied to any particular framework.

3. Pig processes data very fast-no need to go through a time-consuming data import process prior to running queries, as in conventional databases.

4. Pig allows integration of user code where ever possible, so it supports user defined:

> field transformation functions

> user defined aggregates

> user defined conditionals.

7. Is Pig latin commands case sensitive?

Keywords in Pig Latin are not case sensitive.

Field names and relations are case sensitive.

UDF names are also case sensitive.

8. What is a data flow language?

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. Pig is a dataflow language in which, you provide instructions in pig variables to process data. Pig provides an engine(pig engine) for executing the data flows in parallel on Hadoop.