**Assignment 9.1**

**Question:** Why MapReduce program is needed in Pig Programming?

* In simple terms Map Reduce is low level of programming and Pig is a high-level language for expressing data analysis programs which internally create sequence of Map Reduce Programs.
* Pig is simple to learn and use as compared to Map Reduce.
* Pig data flow language i.e pig Latin. For MapReduce, Java is by default supported programming language. However support for other language is also available.
* Pig provides inbuilt optimization for MR jobs whereas in map reduce developer needs to take care of optimization.

**Question:** What are advantages of pig over MapReduce?

|  |  |
| --- | --- |
| **Apache Pig** | **MapReduce** |
| Apache Pig is a data flow language. | MapReduce is a data processing paradigm. |
| It is a high level language. | MapReduce is low level and rigid. |
| Performing a Join operation in Apache Pig is pretty simple. | It is quite difficult in MapReduce to perform a Join operation between datasets. |
| Any novice programmer with a basic knowledge of SQL can work conveniently with Apache Pig. | Exposure to Java is must to work with MapReduce. |
| Apache Pig uses multi-query approach, thereby reducing the length of the codes to a great extent. | MapReduce will require almost 20 times more the number of lines to perform the same task. |
| There is no need for compilation. On execution, every Apache Pig operator is converted internally into a MapReduce job. | MapReduce jobs have a long compilation process. |

**Question:** What is pig engine and what is its importance?

* The language used to analyze data in Hadoop using Pig is known as Pig Latin. It is a highlevel data processing language which provides a rich set of data types and operators to perform various operations on the data.
* 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, and Embedded). After execution, these scripts will go through a series of transformations applied by the Pig Framework, to produce the desired output.
* Internally, Apache Pig converts these scripts into a series of MapReduce jobs, and thus, it makes the programmer’s job easy. The architecture of Apache Pig is shown below.



* Finally the MapReduce jobs are submitted to Hadoop in a sorted order. Finally, these MapReduce jobs are executed on Hadoop producing the desired results.

**Question:** What are the modes of Pig execution?

There are various components in the Apache Pig framework. Let us take a look at the major components.

**Parser**

* Initially the Pig Scripts are handled by the Parser. It checks the syntax of the script, does type checking, and other miscellaneous checks. The output of the parser will be a DAG (directed acyclic graph), which represents the Pig Latin statements and logical operators.
* In the DAG, the logical operators of the script are represented as the nodes and the data flows are represented as edges.

**Optimizer**

* The logical plan (DAG) is passed to the logical optimizer, which carries out the logical optimizations such as projection and pushdown.

**Compiler**

* The compiler compiles the optimized logical plan into a series of MapReduce jobs.

**Execution engine**

* Finally the MapReduce jobs are submitted to Hadoop in a sorted order. Finally, these MapReduce jobs are executed on Hadoop producing the desired results.

**Question:** What is grunt shell in Pig?

* The Grunt shell of Apache Pig is mainly used to write Pig Latin scripts.
* This document describes commands supported by grunt that can be used in interactive shell as well as in batch mode. The supported commands include DFS commands, pig commands as well as a few others. All of them are discussed in the document.
* This section describes currently available commands. The commands in each section are listed in alphabetical order. All commands are case insensitive and white spaces are not significant.

**Question:** What are the features of Pig Latin language?

* Apache Pig is a generic framework which consists of implementation of many MapReduce Design Pattens.
* Apache Pig is implemented in Java Programming Language.
* Instead of providing Java Based API framework, Pig provides its own scripting language which is called as Pig Latin.
* 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.

**Question:** Is Pig Latin commands case sensitive?

Unfortunately, Pig Latin cannot decide whether it is case-sensitive. Keywords in Pig Latin are not case-sensitive; for example, LOAD is equivalent to load. But relation and field names are. So A = load 'foo'; is not equivalent to a = load 'foo';. UDF names are also case-sensitive, thus COUNT is not the same UDF as count.

**Question:** What is a data flow language?

* In computer programming, **dataflow** programming is a programming paradigm that models a program as a directed graph of the **data flowing** between operations, thus implementing **dataflow** principles and architecture.
* Pig provides developers many operators which can be applied on data one after another to get final output.
* Once data is loaded, it flows through all Pig operators.

This is the reason Pig is called as data flow language.