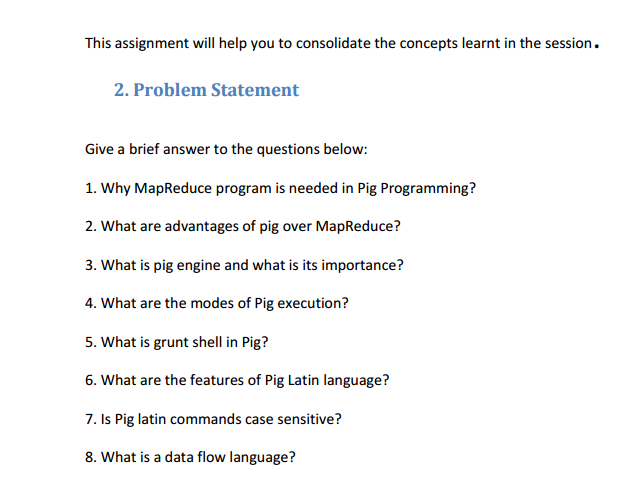
**Assignment 9.1**

****

1 >> MapReduce program is needed in Pig Programming as :

* 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 provides inbuilt optimization for MR jobs whereas in map reduce developer needs to take care of optimization.
* Pig’s programming language referred to as Pig Latin is a coding approach that provides high degree of abstraction for MapReduce.
* MapReduce is a powerful programming model for parallelism based on rigid procedural structure. Hadoop MapReduce allows programmers to filter and aggregate data from HDFS to gain meaningful insights from big data. The Map and Reduce algorithmic functions can also be implemented using C, [Python](https://www.dezyre.com/data-science-in-python/36)and Java. The only drawback to use the coding approach of [Hadoop MapReduce](https://www.dezyre.com/article/hadoop-mapreduce-vs-apache-spark-who-wins-the-battle/83) is that hadoop developers need to write several lines of basic java code requiring extra effort and time for code review and QA.
* Thus, to simplify this Apache offers other options like Pig Latin and Hive SQL languages that help in constructing MapReduce programs easily.

2 >> 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 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. |

3 >> Pig Engine and its importance :

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.

Pig Latin is a command-based scripting language and designed specifically for data transformation and flow.

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.

**Importance :**

* Pig Latin queries are converted to Map and Reduce jobs and hence they take advantage of parallel processing.
* It has the ability to perform computations which can not be done by MapReduce.
* Pig Latin provides all of the standard data-processing operations, such as join, filter, group by, order by, union, etc.
* Internally, Apache Pig converts these scripts into a series of MapReduce jobs, and thus, it makes the programmer’s job easy.
* In Pig Latin joins and ordering codes comprise of 8-9 lines of code and take few minutes to write and debug. The same code in MapReduce will span hundred lines of code and takes hours to develop.

4>> Modes of Pig Execution :

We can run Pig in two modes:

1. **MapReduce/Hadoop Mode:**

Here Pig jobs run as a series of MapReduce jobs picking the input and output paths from HDFS.

* 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>

1. **Local Mode:**

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

* 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.

5>> **Grunt shell in Pig :**

* The Grunt shell of Apache Pig is mainly used to write Pig Latin scripts. Prior to that, we can invoke any shell commands using **sh** and **fs** .
* Grunt shell is an Interactive Shell for executing Pig Commands
* Pig with Grunt is used when script file is not provided
* Running Pig with Grunt shell can execute scripts from Grunt via run or execute commands.
* The Grunt shell provides a set of utility commands. These include utility commands such as **clear, help, history, quit,** and **set**; and commands such as **exec, kill,** and **run** to control Pig from the Grunt shell.

6>> **Features of Pig latin 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.
* Pig Latin is a command-based scripting language .
* It is designed specifically for data transformation and flow.
* Pig Latin statements can span multiple lines and must end with a semi-colon ( ; ).
* Pig Latin statements are generally organized in the following manner:
* A LOAD statement reads data from the file system.
* A series of "transformation" statements process the data.
* A STORE statement writes output to the file system; or, a DUMP statement displays output to the screen.

**7>> Is Pig latin commad is 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.

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) and all other Pig Latin keywords are case insensitive.

In the example below, note the following:

1. The names (aliases) of relations A, B, and C are case sensitive.
2. The names (aliases) of fields f1, f2, and f3 are case sensitive.
3. Function names PigStorage and COUNT are case sensitive.
4. 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.
5. In the FOREACH statement, the field in relation B is referred to by positional notation ($0).

**8>> Data Flow Language :**

In computer programming, **dataflow** programming is a programming 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.