Assignment 7.1

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

PIG is just an abstraction over mapreduce which gives query like syntax known as PIG Latin language. PIG makes use of HDFS and Mapreduce in order to get results from datasets. Hence MapReduce Program is needed in Pig Programming.

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

1. Pig Latin provides all of the standard data-processing operations, such as join, filter, group by, order by, union, etc.

2. MapReduce requires programmers:

• Programmers must think in terms of map and reduce functions.

• Most probably Java programmers are required.

3. 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. Pig provides high-level language that can be used by

• Data Analysts

• Data Scientists

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

Pig provides an engine for executing data flows in parallel on Hadoop. It includes a language, Pig Latin, for expressing these data flows.

• Pig Latin includes operators for many of the traditional data operations (join, sort, filter, etc.), as well as the ability for users to develop their own functions for reading, processing, and writing data.

• Pig runs on Hadoop. It makes use of both the Hadoop Distributed File System, HDFS, and Hadoop’s processing system, MapReduce.

• It allows users to describe how data from one or more inputs should be read, processed, and then stored to one or more outputs in parallel.

• Pig Latin script describes a directed acyclic graph (DAG), where the edges are data flows and the nodes are operators that process the data.

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

1. **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 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

**2. 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.

19

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

* Interactive Shell for executing Pig Commands.
* Used when script file is not provided.
* Can execute scripts from Grunt via run or exec commands

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

* **Rich set of operators** − It provides many operators to perform operations like join, sort, filer, etc.
* **Ease of programming** − Pig Latin is similar to SQL and it is easy to write a Pig script if you are good at SQL.
* **Optimization opportunities** − The tasks in Apache Pig optimize their execution automatically, so the programmers need to focus only on semantics of the language.
* **Extensibility** − Using the existing operators, users can develop their own functions to read, process, and write data.
* **UDF’s** − Pig provides the facility to create **User-defined Functions**in other programming languages such as Java and invoke or embed them in Pig Scripts.
* **Handles all kinds of data** − Apache Pig analyzes all kinds of data, both structured as well as unstructured. It stores the results in HDFS.

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

**NO**

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

dataflow programming is a [programming paradigm](https://en.wikipedia.org/wiki/Programming_paradigm) that models a program as a [directed graph](https://en.wikipedia.org/wiki/Directed_graph) of the data flowing between operations, thus implementing [dataflow](https://en.wikipedia.org/wiki/Dataflow) principles and architecture. Dataflow [programming languages](https://en.wikipedia.org/wiki/Programming_language) share some features of [functional languages](https://en.wikipedia.org/wiki/Functional_language), and were generally developed in order to bring some functional concepts to a language more suitable for numeric processing.