# Big Data Hadoop Training

Session 7 Assignment 1 Solution:

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

A. MapReduce is a programming model, or framework for processing large data sets in distributed manner, using large number of computers, i.e. nodes.

This algorithm consists of two steps – map and reduce. During the mapping phase, master node takes the input, creates smaller sub-problems out of it and distributes those to computers that are actually performing the processing – worker nodes. After the data was processed, it is being sent back to the master node. That is when reduce step begins: master node aggregates all the responses and combines them and creates the answer to the original problem.

The advantages of MapReduce programming are:

* Scalability
* Cost-effective solution
* Flexibility
* Fast
* Security and Authentication
* Parallel processing
* Availability and resilient nature
* Simple model of programming

Because of these reasons, Pig sits on top of Hadoop framework for processing large data sets and makes use of both the Hadoop Distributed File System (HDFS) and Hadoop’s processing system (MapReduce).

Apache Pig programs are written in a query language known as Pig Latin that is similar to the SQL query language. To execute the query, there is a need for an execution engine. The Pig engine converts the queries into MapReduce jobs and thus MapReduce acts as the execution engine and is needed to run the programs.

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

**A.** Advantages of Pig over MapReduce:

* Hadoop MapReduce is a compiled language whereas Apache Pig is a scripting language.
* Pig provide higher level of abstraction whereas Hadoop MapReduce provides low level of abstraction.
* Hadoop MapReduce requires more lines of code when compared to Pig.
* Hadoop MapReduce requires more development effort than Pig.
* When using Pig 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.

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

**A.**Apache Pig has a component known as Pig Engine that accepts the Pig Latin scripts as input and converts those scripts into MapReduce jobs. It acts as interpreter between Pig Latin script and MapReduce jobs. It creates environment to execute Pig scripts into series of MapReduce jobs in parallel manner.

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

**A.** Pig has two execution modes or exectypes:

* **Local Mode** - To run Pig in local mode, you need access to a single machine; all files are installed and run using your local host and file system. Specify local mode using the -x flag (pig -x local).
* **MapReduce Mode** - To run Pig in MapReduce mode, you need access to a Hadoop cluster and HDFS installation. MapReduce mode is the default mode; you can, *but don't need to*, specify it using the -x flag (pig OR pig -x MapReduce).

You can run Pig in either mode using the "pig" command (the bin/pig Perl script) or the "java" command (java -cp pig.jar ...).

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

**A.** The Grunt shell of Apache Pig is mainly used to write Pig Latin scripts. After invoking the Grunt shell, we can run our Pig scripts in the shell. In addition to that, there are many useful shell and utility commands provided by the Grunt shell.

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

**A.** Features of Pig Latin language:

**Apache Pig comes with the following features** −

* **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?**

**A. Case Sensitivity:**

The names (aliases) of relations and fields are case sensitive. The names of Pig Latin functions are case sensitive. The names of parameters and all other Pig Latin keywords are case insensitive.

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

**A.** 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. Dataflow programming languages share some features of functional languages, and were generally developed in order to bring some functional concepts to a language more suitable for numeric processing. Some authors use the term DataStream instead of Dataflow to avoid confusion with Dataflow Computing or Dataflow architecture, based on an indeterministic machine paradigm