**Assignment 9.6**

1. **Explain about the different complex data types in pig.**

Pig data types are classified into two types. They are:

1. Primitive

2. Complex

Primitive Data Types: The primitive datatypes are also called as simple datatypes. The simple data types that pig supports are:

a. int : It is signed 32 bit integer. This is similar to the Integer in java.

b. long : It is a 64 bit signed integer. This is similar to the Long in java.

c. float : It is a 32 bit floating point. This data type is similar to the Float in java.

d. double : It is a 63 bit floating pint. This data type is similar to the Double in java.

e. chararray : It is character array in unicode UTF-8 format. This corresponds to java's String object.

f. bytearray : Used to represent bytes. It is the default data type. If you don't specify a data type for a filed, then bytearray datatype is assigned for the field.

g. boolean : to represent true/false values.

2. Complex Types: Pig supports three complex data types. They are listed below:

a.Tuple : An ordered set of fields. Tuple is represented by braces. Example: (1,2)

b.Bag : A set of tuples is called a bag. Bag is represented by flower or curly braces. Example: {(1,2),(3,4)}

c.Map : A set of key value pairs. Map is represented in a square brackets. Example: [key#value] . The # is used to separate key and value.

Pig allows nesting of complex data structures. Example: You can nest a tuple inside a tuple, bag and a Map

1. **How can you interact with the shell in Apache pig .**

The shell in apache pig is called grunt shell.

The grunt can be launched by using simple command “pig” or “pig -x local”.

To enter Grunt, invoke Pig with no script or command to run. Typing:

pig -x local

will result in the prompt:

grunt>

This gives you a Grunt shell to interact with your local file system. If you omit the -x local and have a cluster configuration set in PIG\_CLASSPATH, this will put you in a Grunt shell that will interact with HDFS on your cluster.

1. **Explain how pig differs from Map reduce.**

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.

Here are detail definition from official documentation:

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject)

1. **Explain how pig differs from sql.**

Structured Query Language (SQL) has been a programmer’s companion for decades. It was the de-facto solution for extracting data for further processing. Big Data has changed how we visualize and process data. SQL’s demand of storing data in a strict relational database schemas and its declarative nature often deflects focus from the ultimate purpose – to extract data for analysis. For all its popularity, advent of Big Data, challenged SQL’s ability and performance.

SQL programmers required languages that were relatively easy to learn for someone having SQL background and at the same time was –

1. Free of SQL’s excess baggage mentioned above and
2. Could easily handle large data sets.

Originally developed at Yahoo Research in 2006, Pig addressed all these issues and provided better optimization scope and extensibility. Apache Pig also allows developers to follow multiple query approach, which reduces the data scan iterations. It has provisions for a number of nested data types (Maps, Tuples and Bags) and commonly used data operations such as Filters, Ordering and Joins. These advantages have seen Pig being adopted by a large number of users around the globe. Its simplicity has resulted in Yahoo and Twitter resorting to Pig for the majority of their MapReduce operations.

1. **Explain the scalar data types in pig.**

The scalar data types in pig are as below:

a.int : It is signed 32 bit integer. This is similar to the Integer in java.

b.long : It is a 64 bit signed integer. This is similar to the Long in java.

c.float : It is a 32 bit floating point. This data type is similar to the Float in java.

d.double : It is a 63 bit floating pint. This data type is similar to the Double in java.

e.chararray : It is character array in unicode UTF-8 format. This corresponds to java's String object.

f.bytearray : Used to represent bytes. It is the default data type. If you don't specify a data type for a filed, then bytearray datatype is assigned for the field.

g.boolean : to represent true/false values.