**Assignment 9.6**

**Problem Statement:**

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

The different complex data types in pig are as below:

* Tuple

Tuple is an ordered set of fields.

For example,

(16,1)

* Bag

Bag is a collection of tupes.

For example,

{(16,1),(3,9)}

* Map

Map is a set of key value pairs.

For example,

[key#value]

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.**

PIG is a data flow language, the key focus of Pig is manage the flow of data from input source to output store. As part of managing this data flow it moves data feeding it to process1, taking output and feeding it to process2. The core features are preventing execution of subsequent stages if previous stage fails, manages temporary storage of data and most importantly compresses and rearranges processing steps for faster processing. While this can for any kind of processing tasks Pig is written specifically for managing data flow of Map reduce type of jobs. Most if not all jobs in a Pig are map reduce jobs or data movement jobs. Pig allows for custom functions to be added which can be used for processing in Pig, some default ones are like ordering, grouping, distinct, count etc.  
    Map reduce on the other hand is a data processing paradigm, it is a framework for application developers to write code in so that its easily scaled to PB of tasks, this creates a separation between the developer that writes the application vs the developer that scales the application. Not all applications can be migrated to Map reduce but good few can be including complex ones like k-means to simple ones like counting uniques in a dataset.be done

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:

int :

Signed 32-bit integer

Long :

Signed 64-bit integer

float :

32-bit floating point

double :

64-bit floating point