**Session 9**

**Assignment 6**

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

* Pig supports these complex data types. They are listed below:
* **Atom:** Single value in Pig Latin, irrespective of the data type.

**Example:** ‘raja’ or ‘10’

* **Tuple:** A record formed by an ordered set of fields. The fields can be of any type. Tuple is represented by braces.

**Example:** (1, 2)

* **Bag:** A bag is an unordered set or collections of tuples. Bag can have duplicate tuples. It can have tuples with fields that have different data types. Bag is represented by flower or curly braces.

**Example:** {(1, 2), (3, 4)}

* **Map:** A set of key value pairs. The **key** should be of char array and unique. Map is represented in a square brackets.

**Syntax:** [key#value]. The ‘#’ is used to separate key and value.

**Example:** [name#ram]

**Note:** 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 can be interacted in two ways. They are,

1. With Grunt Shell.
2. Using script files.

* **With Grunt Shell**
* Shell is highly interactive for executing Pig Commands.
* Used when script file is not provided.
* Grunt workspace can be used to execute scripts via run or exec commands.
* **Using script files**
  + Commands in a file are executed
  + **Syntax:** pig ScriptFile.pig
  + Script files executes the pig commands as batch jobs.

1. Explain how Pig differs from MapReduce.

* **PIG:**

1. Apache Pig is an abstraction over MapReduce.
2. It is a high level language known as **Pig Latin**.
3. It is a tool or platform which is used to analyze larger sets of data representing them as data flows and thus known as data flow language.
4. Any programmer with a basic knowledge of SQL can work conveniently with Apache Pig.
5. Apache Pig uses multi-query approach, thereby reducing the length of the codes to a great extent.
6. There is no need for compilation. On execution, every Apache Pig operator is converted internally into a MapReduce job.

* **MapReduce:**

1. MapReduce is a data processing model.
2. It is a low level and rigid language.
3. In MapReduce, it is difficult to perform a Join operation between the datasets.
4. Exposure to Java is must to work with MapReduce.
5. MapReduce will require almost 20 times more the number of lines to perform the same task.
6. MapReduce jobs have a long compilation process because of the great length of codes.
7. Explain how pig differs from SQL.

* **PIG:**

1. Pig Latin is a procedural language.
2. In Apache Pig, schema is optional. We can store data without designing a schema. Values are stored as $01, $02 etc.
3. The data model in Apache Pig is nested relational.
4. Apache Pig provides limited opportunity for Query optimization.

* **SQL:**

1. SQL is a declarative language.
2. Schema is mandatory in SQL because it does not store values as default.
3. The data model used in SQL is flat relational.
4. There is more opportunity for query optimization in SQL.
5. Explain the scalar data types in pig.

* The scalar datatypes are also called as primitive datatypes. The simple data types that pig supports are:
* **int:** It is signed 32 bit integer. This is similar to the Integer datatype in java. Constant integers are expressed as integer numbers.

**Example:** 40, -20.

* **long:** It is a 64 bit signed integer. This is similar to the Long in java. Constant longs are expressed as integer values with an L appended.

**Example:** 1000000000L.

* **float:** It is a 32 bit floating point number. This data type is similar to the Float in java. During calculations it will lose precision. To avoid this you should use int or long instead. Constant floats are expressed as floating point values with f appended.

**Example:** 3.14f, 3.8e8f (for exponent format).

* **double:** It is a 63 bit floating point number. This data type is similar to the Double in java. Constant doubles are expressed as a floating point value either in a simple format or in an exponent format.

**Example:** 2.167, 7.368e4

* **chararray:** It is character array in unicode UTF-8 format. This corresponds to java's String object. Constant chararray are generally expressed as string literals with single quotes.

**Example:** ‘hadoop’

* **bytearray:** It is a blob or array of bytes. 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. There is no way to specify a constant bytearray.