**ASS-9.6**

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

**Atom:**

• Any single value in Pig Latin, irrespective of their data or type is known as an Atom.

• It is stored as bytearray by default and can be used as string or number like int, long, float, double, chararray, and bytearray are the atomic values of Pig.

• A piece of data or a simple atomic value is known as a field. Example − ‘raja’ or ‘30’

**Tuple:**

• A record that is formed by an ordered set of fields is known as a tuple, the fields can be of any type.

• A tuple is similar to a row in a table of RDBMS

**Bag:**

• A bag is an unordered set of tuples.

• In other words, a collection of tuples (non-unique) is known as a bag.

• Each tuple can have any number of fields (flexible schema).

• A bag is represented by ‘{}’.

• It is similar to a table in RDBMS, but unlike a table in RDBMS, it is not necessary that every tuple contain the same number of fields or that the fields in the same position (column) have the same type. Example − {(Raja, 30), (Mohammad, 45)}

**Map :**• A map (or data map) is a set of key-value pairs.

• The key needs to be of type chararray and should be unique.

• The value might be of any type. It is represented by ‘[]’ Example: [name#Raja, age#30]

Relation :• A relation is an outer bag of tuples.

• The relations in Pig Latin are unordered (there is no guarantee that tuples are processed in any particular order).

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

It can by two ways :

1**.grunt shell** -Interactive Shell for executing Pig Commands

• It is used when script file is not provided

• It can execute scripts from Grunt

2**.Using script files**- Executes Commands in a file

• Pig commands are executed using script files as batch Jobs

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

|  |  |
| --- | --- |
| **PIG** | **MAP REDUCE** |
| It can be used for Structured and semi structured data | It can be used for Structured ,semi structured and un structured data |
| It divides the data into series of simple map reduce jobs | It performs map and reduce for entire data set. |
| It can be useful for effectively performing joins, querying etc. | It can be used for only map reduce |
| Development time is less (code development) | Development time is high (code development) |
| Processing Time is high since it runs on top of map reduce | Processing Time is comparatively less |
| Useful for performing analysis | Useful for complex data analysis |

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

• In SQL, when users want to do several data operations together, they must either write separate queries, storing the intermediate data into temporary tables, or write it in one query using subqueries inside that query to do the earlier steps of the processing.

• Pig, however, is designed with a long series of data operations in mind, so there is no need to write the data pipeline in an inverted set of subqueries or to worry about storing data in temporary tables.

• SQL is designed for the RDBMS environment, where data is normalized and schemas and proper constraints are enforced (that is, there are no nulls in places they do not belong, etc.).

• Pig is designed for the Hadoop data-processing environment, where schemas are sometimes unknown or inconsistent. Pig does not require data to be loaded into tables first. It can operate on data as soon as it is copied into HDFS.

• Pig Latin is the native language of parallel data-processing systems.

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

**1.Int**

It is like integer datatype in java. They store a four-byte signed integer. Constant integers are expressed as integer numbers, for example, 4.

**2.long**

It is like Long datatype in java. They store an eight-byte signed integer. Constant longs are expressed as integer numbers with an L appended, for example, 5000000000L.

**3.float**

It is like integer datatype in java a floating-point number and use four bytes to store their value. Since this is a floating-point number, in some calculations it will lose precision. For calculations that require no loss of precision, you should use an int or long instead. Constant floats are expressed as a floating-point number with an f appended. Floating-point numbers can be expressed in simple format, 3.14f, or in exponent format, 6.022e23f.

**4.double**

A double-precision floating-point number. It use eight bytes to store their value. You can find the range of values representable by Java’s Double type. Note that because this is a floating-point number, in some calculations it will lose precision. For calculations that require no loss of precision, you should use an int or long instead. Constant doubles are expressed as a floating-point number in either simple format, 2.71828, or in exponent format, 6.626e-34.

**5.chararray**

A string or character array. Chararrays are represented in interfaces by java.lang.String. Constant chararrays are expressed as string literals with single quotes, for example, 'fred'. In addition to standard alphanumeric and symbolic characters, you can express certain characters in chararrays by using backslash codes, such as \t for Tab and \n for Return. Unicode characters can be expressed as \u followed by their four-digit hexadecimal Unicode value. For example, the value for Ctrl-A is expressed as \u0001.

**6.bytearray**

A blob or array of bytes. Bytearrays are represented in interfaces by a Java class DataByteArraythat wraps a Java byte[]. There is no way to specify a constant bytearray.