Q1-Explain about the different complex data types in pig.

Answer:

1. **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.

Eg: ‘pqr’ or ‘21’

2. **Tuple:**

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

Eg: (dog, 3)

3. **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 ‘{}’

Eg: {(a, 1), (z, 2)}

4. **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 ‘[]’

Eg: [name#Raja, age#30]

5. **Relation:**

• A relation is an outer bag of tuples.

Q2-How can you interact with the shell in Apache pig?

Answer:

1. After invoking the Grunt shell, you can run your Pig scripts in the shell.

2. In addition to that, there are certain useful shell and utility commands provided by the Grunt shell.

3. The Grunt shell of Apache Pig is mainly used to write Pig Latin scripts. Prior to that, we can invoke any shell commands using **sh** and **fs.**

4. The Grunt shell provides a set of utility commands. These include utility commands such as **clear, help, history, quit,** and **set**; and commands such as **exec, kill,** and **run** to control Pig from the Grunt shell.

Q3-Explain how pig differs from Map reduce.

Answer:

1. 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.

2. Pig is simple to learn and use as compared to Map Reduce.

3. Pig has its’ own data flow language i.e pig Latin. For MapReduce, Java is by default supported programming language. However support for other language is also available.

4. Pig provides inbuilt optimization for MR jobs whereas in map reduce developer needs to take care of optimization.

Q4-Explain how pig differs from sql.

Answer:

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.

Q5-Explain the scalar data types in pig.

Answer:

Int Signed 32-bit integer 10

Long Signed 64-bit integer 10L or 10l

Float 32-bit floating point 10.5F or 10.5f or 10.5e2f or 10.5E2F

Double 64-bit floating point 10.5 or 10.5e2 or 10.5E2

Chararray Character array (string) hello world

In Unicode UTF-8 format

Bytearray Byte array (blob)