Session 7: Pig

Assignment 1

1. Why Map-reduce program is needed in Pig Programming?

**Ans :** Pig is just a tool that is used to access the data that is stored in HDFS. It is a High level platform for MapReduce programming. All the pig scripts are converted to map-reduce statements. To run this map-reduce statements we need map-reduce programming. This MapReduce statements access the HDFS data and return the output to the user.

1. What are advantages of pig over MapReduce?

**Ans :** 1. Writing a pig script is very easy and a non-programmer can also write it with much ease.

2. The way in which tasks are encoded permits the systems to optimize their execution automatically.

3. The user just have to focus on semantics rather than efficiency.

4. It’s very easy to debug a pig script in compare to a map-reduce program.

1. What is pig engine and what is its importance?

**Ans:** Pig engine consists of below components:

1. Parser: Checks the syntax of the script. The output of the parser is a DAG (directed acyclic graph). DAG represents the pig Latin statements and logical operators.

Logical operators are represented as the nodes and the dataflow are represented as edges.

2. Optimizer: DAG is the input for optimizer, which carries out the logical optimizations such as projection and pushdown.

3. Compiler: The output of optimizer is the input for this phase. The compiler compiles the optimized logical plan into a series of MapReduce jobs.

4. Execution engine: MapReduce jobs are submitted to Hadoop in a sorted order. These jobs are executed on Hadoop producing the output.

1. What are the modes of Pig execution?

**Ans :** There are two modes of pig execution

1. Local mode : pig or pig -x local
2. Mapreduce/Hadoop mode : pig or pig -x mapreduce
3. What is grunt shell in Pig?

**Ans :** Pig has provided us a default shell that is grunt. Grunt shell is used to run pig commands from console. We can also run the pig script from grunt shell by using exec command.

1. What are the features of Pig Latin language?

**Ans:** Pig Features:

Rich set of operators − It provides many operators to perform operations like join, sort, filer, etc.

Ease of programming − Pig Latin is similar to SQL and it is easy to write a Pig script if you are good at SQL.

Optimization opportunities − The tasks in Apache Pig optimize their execution automatically, so the programmers need to focus only on semantics of the language.

Extensibility − Using the existing operators, users can develop their own functions to read, process, and write data.

UDF’s − Pig provides the facility to create User-defined Functions in other programming languages such as Java and invoke or embed them in Pig Scripts.

Handles all kinds of data − Apache Pig analyzes all kinds of data, both structured as well as unstructured. It stores the results in HDFS.

1. Is Pig latin commands case sensitive?

**Ans :** Keywords are case insensitive whereas functions are case sensitive.

1. What is a data flow language?

**Ans:** Dataflow language is a language where data is passed from instruction to instructions for processing.