Session 7: Pig

Assignment 1

Session 7: Pig

Assignment –PIG

Table of Contents

Big Data and Hadoop Development

1. Introduction

You will work on the concepts of Pig.

2. Objective

This assignment will help you to understand pig concepts.

3. Prerequisites

Acadgild's VM , or Linux operating system with Hadoop and Pig installed in it.

4. Associated Data Files

N/A

5. Problem Statement

Give a brief answers to the questions below:

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

• ANSWER: cuz Each pig command (script) run a map-reduce program which is already defined in pig. Pig is application that runs on top of MapReduce and abstracts Java MapReduce jobs away from developers.

2. What are advantages of pig over MapReduce?

• ANSWER: Pig Latin uses a lot fewer lines of code than the Java MapReduce script. The Pig Latin script was is easier to read for someone without a Java background. MapReduce jobs can written in Pig Latin.

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

ANSWER: Pig Engine – parses, optimizes, and automatically executes PigLatin scripts as a series of MapReduce jobs on a Hadoop cluster

4. What are the modes of Pig execution?

ANSWER: There are three modes of Pig Execution

1. Local Mode

2. MapReduce Mode

3. Embadded Mode

5. What is grunt shell in Pig?

ANSWER: grunt shell in pig gives a plateform where we can execute pig script and get the result back .

- 6. What are the features of Pig Latin language?

1. Pig is application that runs on top of MapReduce and abstracts Java MapReduce jobs away from developers.
2. Pig Latin uses a lot fewer lines of code than the Java MapReduce script.
3. The Pig Latin script was is easier to read for someone without a Java background.
4. MapReduce jobs can written in Pig Latin.

7. Is Pig latin commands case sensitive?

yes

8. What is a data flow language?

A programming paradigm in which computation is modelled as a directed graph (which may or may not contain cycles), the nodes of which are either **data** sources (producers of **data**), **data** sinks (consumers), or "processing elements" which compute some function; and the arcs of which represent **dataflow** between nodes.

6. Expected output

N/A

Big Data and Hadoop Development

7. Approximate Time to Complete Task

30 min