Objectives

Display understanding of the Hadoop Ecosystem

Display an understadning of how to use and process data using Sqoop

Display an understanding of writing MapReduce Jobs

Display an understanding of using the HIVE metastore and writing HIVEQL queries

Display an understanding of using Pig and writing Pig Latin queries

Show and understanding of the use of standard MySQL based SQL queries

Outcomes

At the outcome of this assignment you will have installed, applied, and built your own hadoop-based components for Big Data queries. This exercise will create a comparative grid of some of the different tools availalbe. You will compare and contrast these solutions as well.

You need to create a Hadoop Pseudo-Cluster on your local Vagrant Box

You need to take the provided Webserver Logs (Blackboard and insert them in to HDFS) - Where applicable, all work will be done in HDFS

You will need to insert the Webserver Logs into a MySQL table (you provide the schema of all fields)

Create a week 13 folder in your private github repo and include the following code and ReadMe.md where indicated

You will execute the following jobs:

Create a sub-folder item-one

Write a Map Reduce Job (Java Classes to find) (You can use Reducers, Chaining, or any other topic we covered)

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per month

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per day

Create a ReadMe.md file with your answers and any assumptions made during the job run

Create a sub-folder item-two

Write a SQL command or commands to achieve: This can be run directly from the MySQL prompt or via .sql file (only if you wish to write a Java program NOT required)

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per month

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per day

Create a ReadMe.md file with your answers and any assumptions made during the job run

Create a sub-folder item-three

Write a Sqoop job that will query the webserver-log dataset, generate java classes (see book widget examples and reuse any logic from your MapReduce code) and write the Sqoop Job to find:

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per month

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per day

Create a ReadMe.md file with your answers and any assumptions made during the job run

Create a sub-folder item-four

Write a HiveQL query to find: (jsut include the HIVEQL code in a file named logs.hive)

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per month

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per day

Create a ReadMe.md file with your answers and any assumptions made during the job run

Create a sub-folder item-five

Write a Pig Latin Script to find: (include the pig latin script logs.pig)

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per month

the most frequently visted WebPage (URL) with a 200 return code that is not index.\* per day

Create a ReadMe.md file with your answers and any assumptions made during the job run

Create a sub-folder item-six

Provide a written comparison and contrasting of all of these 5 elements in two categories: Place all items in a ReadMe.md

1. Ease of gathering, processing, and wrtting code to retrieve data

2. Explain your personal choice based on item 1, and if you have any scenarios from work or past work experience where any of these would help (extra)

Create a ReadMe.md file with your answers and any assumptions made during the job run