Programming Assignment #4.

Building a HTTP Log Analyzer Using Pig Latin

Due date: May 7 2013, Noon
URL: http://www.cs.colostate.edu/~cs480

1 Introduction

The HTTP log analyzer provides a fast and accurate data analysis for public web pages. Your static log file should be loaded into HDFS (Hadoop Distributed File System), and specific data analysis features will be executed through customized Pig Latin scripts. A key goal of the HTTP log analyzer is to provide a basic understanding of the pattern of Web access to a Web server hosted by NASA. Apache Pig [1] will provide scalable environment to access and analyze the log data.

2 Objective

In this assignment, you will build several scripts to analyze Web logs. If it is needed, you can develop your own UDFs and invoke them from your script file.

3 Description of Task

1. Preparation

You should install Pig in your CS account. Hadoop is installed in clusters in the CS department.

2. Understanding Access Patterns

3.1 Visitor Recency

Build a script to generate a sorted list of users (without duplicate) based on the frequency of visits within specified time period. Your script should take arguments from the command line:

- Time period (e.g. ***2:*** recent 2 hours, ***6***: recent 6 hours)
- Input File Name
- Output File Name

Your output file should include, visitor's URL and the number of visits within specified hours.

3.2 Peak time analysis: Number of access
Build a script to figure the peak time of accesses based on the total number of accesses. Your script should count the total number of accesses every hour and sort them in a descending order. In this script, you do not consider duplicate detection.

- Input File Name
- Output File Name

Your output file should include, a sorted list of the total number of accesses and time periods accordingly.

3.3 Peak time analysis: Size of reply
Build a script to figure the peak time of accesses based on the size of the replies performed by this host. Your script should add up the size of the replies every hour and sort them in descending order. In this script, you do not consider duplicate detection.

- Input File Name
- Output File Name

Your output file should include, a sorted list of the total size of the replies performed by current host generated hourly. Your output should show the time period associated with these values.

3.4 Error analysis
Build a script to figure out which resource have related most errors. From the unique list of your resources, count the number of non-successful accesses (non-200 code) and provide a list of the resources and number of errors sorted in descending order.

- Input File Name
- Output File Name

Your output file should include, a sorted list of the total number of the non-successful access. Your output should show the URLs of resources associated with these values.

4 Input File

We will provide one month's worth of all HTTP requests to the NASA Kennedy Space Center WWW server [3]. Please download the data file from, ftp://ita.ee.lbl.gov/traces/NASA_access_log_Jul95.gz. This file is 20.7MB gzip compressed, 205.2MB uncompressed.

Format
The logs are in an ASCII file with one line per request, with the following columns:

1. User
   The URL of the user making the request. A hostname when possible otherwise the IP address if the name could not be looked up.

2. Timestamp in the following format “Day Month DD HH:MM:SS YYYY” Where Day is a day of a week and Month is a name of the month, DD is the day of the month, HH:MM:SS is the time of day using 24-hour clock, and YYYY is the year.
3. Request
4. HTTP reply code
5. Bytes in the reply

This log was collected from 00:00:00 July 1, 1995 through 23:59:59 July 31, 1995 a total of 31 days. For more information, please visit [3].

5 Evaluation

This assignment will account for 10% of your final grade. The grading will be done on a 100 point scale. You are required to work alone on this assignment.

6 Late Policy

Please check the late policy available from the course web page.

7 Useful Links

[1] Apache Pig, http://pig.apache.org/
[2] NASA-HTTP public data, http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html