Business Analytics – Project – Apace Log Analysis

Objectives

This program enables the participants to review the learnings of the Business Analytics Using R Workshop.

The primary objective of the project is to enhance the participant's knowledge of R & develop exploratory analysis & visualization skills.

Procedure

- View Apache Sample Log
  Refer apache_sample.pdf
- Understand Apache Logs
  Refer apache_desc.pdf
  Source: http://httpd.apache.org/docs/2.2/logs.html
- Use Dataset as given
  Refer section Apache Data Sets
- Parse & Analyze
  Refer section Procedure
- Analytics Requirement
  Refer section Analytics Requirement
- Generate Project Report
  Refer section Project Report

Apache Data Sets

- apache_http.log - small apache log to create your prototype
- usask_access_log.gz - compressed file containing "UofS_access_log"; an apache log of approx 233 MB

Note:

- "UofS_access_log" to be renamed as "apache_dataset.log"
- Site: Web logs from NASA Web Site
- Source: http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

Procedure

- Copy log_file to your working directory of your choice
- Parse log file & read into data frame
- Store csv format data in an working directory
- The csv file should have
  - date field in yyyy-mm-dd format (time zone to be ignored)
  - time field in hh:mm:ss format (time zone to be ignored)
  - protocol, page visited & http-version should be separate cols
- Provide analysis results as per section Analytics Requirement below
- Copy results to local file system or local MySQL as per requirement.

Analytics Requirement

Required As Data Frame

- For each month, how many times each individual host has connected to our server? Store data month wise & highest count first.
- For each month, how many times each individual page has been requested from our server? Store data month wise & highest count first.
- For each month, how much data has been downloaded by each individual host that has connected to our server? Store data month wise & by highest count first.
- How much data was sent out as each individual page was downloaded from our server? Store data month wise & by highest count first.

Required As Visualization

- For each data set generated above, prepare suitable visualization (giving reason why the graph is chosen). Limit data to suitable significant number if graph is looking too cluttered.
- Time Series Graph for total hits per day with each month being shown as separate line.
- Time Series Graph for total download size per day with each month being shown as separate line.
- How would you show top "10 most popular pages" per day as Time Series Graph with each month being shown as separate line. Explain how you find "top 10 most popular pages"

Answers Required

Using the above results and carrying out any other analysis as may be required, provide answers to the following questions.

- Which host has connected the maximum number of times to our server? Give the host name & count of connections from that host.
- Which page that has been requested the maximum number of times from our server? Give the page name & count of the times the page was requested.
- How many unique hosts have connected to our server? Give counts.
- How many unique pages have been requested from our server? Give counts.
- Which host has caused maximum data transfer from our server? Give host name & the data transfer for the host.
- Which page has caused maximum data transfer from our server? Give page name & the data transfer for the page.

- Which page has maximum download size from our server? Give page name & the size for the page.
- What is the download count of the page that has maximum download size from our server? Give page name & download count
- Which page has minimum download size from our server? Give page name & the size for the page.
- What is the download count of the page that minimum download size from our server? Give page name & the size for the page.

Project Report Using RMD (for R) OR IPYNB (for Python)

- Project Overview
- Commands / Code Section
- Results Section
- Summary - How you used R for Data Analytics

Project Overview

- Brief Overview Of The Project
- Learning Objective

Commands / Code Section

- Should contain all R / Python commands used to transfer logs to data-frames and sample data-frame using head (data-frame, 10)
- Should contain all R / Python commands used in visualization of data-frames and the output as desired.
- Should contain all R commands used to answer the specific queries raised in problem definition along with the output as desired.

Summary

- Describe your experience of using R / Python for Data Analytics.