MS5106 DATA SCIENCE & BIG DATA ANALYTICS

Mid-Term Assignment 1

Big Data Analytics: Log Management

Assignment:

Eurostar is an international company with many branches all over the world. The company's web site is of great importance for them as most of the products and services they offer online.
Recently, the company's business analyst resigned, and the CIO John O'Leary decided to hire you in order to help them.
You are given access to the web server logs and the Hive script used for analysis before. Details as follows:

The log files are stored in AWS S3 at:

s3://us-west-2.elasticmapreduce.samples

Each entry in the log files is a quasi-structured record, which provides details about a single user request in the following format:

```
2014-07-05 20:00:00 LHR3 4260 10.0.0.15 GET eabcd12345678.cloudfront.net /test-
image-1.jpeg 200 - Mozilla/5.0%20(MacOS;%20U;%20Windows%20NT%205.1;%20en-
US;%20rv:1.9.0.9)%20Gecko/2009040821%20IE/3.0.9
```

The Hive script that calculates the total number of requests per operating system over a specified timeframe is provided as file. Code is below:

```
-- Summary: This sample shows you how to analyze CloudFront logs stored in S3 using Hive

-- Create table using sample data in S3.  Note: you can replace this S3 path with your own.
CREATE EXTERNAL TABLE IF NOT EXISTS cloudfront_logs (
  DateLog DATE,
  Time STRING,
  Location STRING,
  Bytes INT,
  RequestIP STRING,
  Method STRING,
  Host STRING,
  Uri STRING,
  Status INT,
  Referrer STRING,
  OS STRING,
  Browser STRING,
  BrowserVersion STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
  "input.regex" = "^(?!#)([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+
([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+[^\(]+[\(]([^\;]+).*\%20([^\/]+)[\/](.*)$"
) LOCATION '${INPUT}/cloudfront/data';

-- Total requests per operating system for a given time frame
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT OS, COUNT(*) FROM
cloudfront_logs WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05' GROUP BY OS;
```

You have been asked to write a report to John, which includes the following:
1. Comment on the script providing a brief explanation of what it does.
2. Modify the script in order to provide additional log analysis of your choice. John will be happy to see any interesting findings, particularly those with potential business value. To run your scripts, you can use a cloud AWS EMR cluster with Hadoop and Hive on it. Include the code and findings to your report.
3. John came across the following statement:
   "According to a survey made among more than 200 professionals in IT, security and compliance, 55% of organisations have deployed big data projects and more than a half of them use the technology for log management. In addition, of those who plan to deploy big data in the next two years, 58% said that log management is a priority."
   You have been asked to do a research on the topic and present a brief summary of your findings, explaining to John what that means in practice.

Requirements:

- Work individually.
- Submit a report to the Blackboard drop box provided.
- Submit also a hard copy of the report, a document up to 3000 words, printed in Times New Roman font, 12-point, single line spacing.
- Submit also an individual assignment submission form, signed and stapled to the document.
- The assignment is due on or before **Fri, 8th Mar 2019**, and carries 20% of the total marks.