# A picture containing text Description automatically generated

**J.E. CAIRNES SCHOOL OF BUSINESS & ECONOMICS**

**Mid-Term Assignment1**

**Module Name and Code: Data Science & Big Data Analytics MS5016**

**Student Name and ID: Dilip Venkatesan Sankar 22225743**

In submitting this assignment, I am aware that it is my responsibility to adhere to the submission guidelines. Please tick (double click… or Yes/No) for the following:

|  |  |  |
| --- | --- | --- |
|  | **Yes** | **No** |
| I am aware of what the NUI Galway plagiarism policy entails. |  |  |
| I have named the assignment file (MS Word docx , doc, or jar ) to contain my student ID and the module code and assignment number (e.g., 1187404\_MS220\_A1.docx or 1187404\_MS220\_A1.jar). |  |  |

Text

Description automatically generated

**Overview**:

Eurotrade is an international corporation that sells its products and services online, and its website is very important to them because they have branches all over the world. Eurotrade has reached out to me for assistance due to the recent resignation of their business analyst. All the logs are stored in a quasi-structured format in the below S3 bucket,

**s3://us-east-1.elasticmapreduce.samples**

We are provided access to the above path to access log files and also provided with the hive script which contain the below commands,

1. Create statement

2. Insert statement

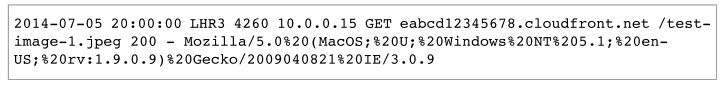
3. Select statement

**Question1:** **Code walk-through, Query Analyses and Explanation**

**1.1 Code Walk-through:**

Hive is a data warehouse program that streamlines the reading, writing, and management of massive SQL-stored databases. It provides a SQL-like query interface for accessing data stored in a variety of Hadoop-integrated databases and file systems. Hive may be used to arrange Hadoop data, and HiveQL, a language comparable to SQL, can be used to query the data.

**AWS's Amazon S3 (Simple Storage Service)** is a cloud storage system that helps enterprises manage and archive data. It provides secure, cost-effective data storage with high availability and durability. The Hive\_CloudFront is stored in a quasi-structure where each line contains details about a single user and the sample is shown below,



**Elastic MapReduce (Amazon EMR),** a managed Hadoop framework-based cloud solution from Amazon Web Services, provides big data processing. And we will be utilizing the same to process the given queries on the CloudFront\_logs stored in S3.

Below is the explanation of the hive statements,

1.***"CREATE EXTERNAL TABLE IF NOT EXISTS cloudfront\_logs*”** is used to check if a table exists in the database. If there is a table by the same name then an exception will be thrown called***“Table already exists”.*** If not an external table will be created.

2. The statement contains 13 columns, and the attributes are provided with the below datatypes,

* DateLog – Data of the order (Data Type - DATE)
* Time - Time of the order (Data Type - STRING)
* Location – Location of the order (Data Type - STRING)
* Bytes – Total bytes of the request (Data Type – INT)
* RequestIP – IP address of the requestor (Data Type - STRING)
* Method - Type of request placed (Data Type - STRING)
* Host - Name of the Host (Data Type - STRING)
* Uri – Uniform resource identifier (Data Type - STRING)
* Status – Return code (Data Type – INT)
* Referrer - Name of the page (Data Type - STRING)
* OS – Name of Operating System (Data Type - STRING)
* Browser – Web Browser Name (Data Type - STRING)
* BrowserVersion – Web browser version (Data Type - STRING)

3. The next part of the query contains the below functions,

***“ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'***

***WITH SERDEPROPERTIES (***

***"input.regex" = "^(?!#)([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+[^\(]+[\(]([^\;]+).\*\%20([^\/]+)[\/](.\*)$"***

***) LOCATION '${INPUT}/cloudfront/data'”***

The first part of the statement which contains ***“ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe' ”*** constitutes the below information,

* The log files in the s3 bucket are kept in a semi-structured format, and the hive can only process and store files in the form of rows and columns, implying that the hive cannot process this unstructured data.
* The SerDe ***“ROW FORMAT SERDE”*** procedure, which stands for Serialization and Deserialization, comes into play here. The Apache Hive SerDe library is a powerful and adaptable data serialization and deserialization framework for the Apache Hadoop platform.
* It allows programmers to quickly build data processing pipelines that may take data from various sources, perform transformations, and return the results in a format suitable for HiveQL searches.

The process of transforming an object or data structure into a series of bytes that can be stored or sent is known as serialization. Deserialization is the process of rebuilding a data item from a string of bytes for storage.

In the second part of the statement where the SerDe function is merged with the ***Regex function and SERDEPROPERTIES***, is a library function that assists hive in identifying and reading any pattern of data and storing it in columns, and the regex expression is stored under input.regex which is used to identify the columns in the log file which is 14 in our case.

The third part of the statement contains the input location of the file and in our case, the log file is the input file,

**s3://us-east-1.elasticmapreduce.samples**

**1.2 Query Analysis and Explanation:**

The insert overwrite query is used to load the files into the specified output directory by overwriting any existing data,

***“INSERT OVERWRITE DIRECTORY '${OUTPUT}/os\_requests/'”,*** where the ${OUTPUT} is created as a sub-folder in the s3 bucket.

The following select query is used to select the OS name and the count of requests made with respect to the OS name between the dates of 07/05/2014 and 08/05/2014 from CloudFront logs, then group by the values based on the OS column,

***“SELECT OS,' -> ', COUNT (\*) FROM cloudfront\_logs WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05' GROUP BY OS;”***

In order to execute the steps, we are creating an S3 bucket and EMR Cluster for query execution:

1. Under S3 we are creating a bucket called **dilipassignment1.bucket** which contains the output and script folders,

Graphical user interface, text, application, email

Description automatically generated

***Fig 1.1 Amazon S3 bucket***

where the output folder has a sub-folder that in turn contains the output files and the script folder houses the hive script.

2. The EMR cluster and steps are added in order to execute the steps in the HIVE file. While creating the cluster we use the below EC2 subnet,



We name the cluster **dilipassignment1.Cluster** and we disable the **Termination protection** option and once the cluster is created we add the steps below,

Graphical user interface, text, application, email

Description automatically generated

***Fig 1.2 EMR steps***

Where the S3 script location, Input log file location, and output folder location are specified.

Graphical user interface, text, application

Description automatically generated

***Fig 1.3 Cluster showing a HIVE problem Running.***

**Question2: Script modification, Analysis, business value and explanation**

The CIO of Eurotrade is interested in seeing some more information based on the statistical analysis done on the log file, even though the logs include valuable information. It is also possible to get the results simply by altering the current query. The log file also includes columns for Location, Internet browser, and Status code that can be utilized for analysis and may have business value in addition to the OS column.

Below are a few Hive script examples, executed to extract valuable information from the log data,

1. Total Requests based on Operating System.
2. Total Requests based on the Location.
3. Total Requests based on the Internet browser.
4. Analysis of Status code.
5. Total count of browsers used in different locations.

The below query is used to create the external table called cloudfront\_logs before the insertion and selection operation,

-- Creates a Table cloudfront\_logs if it does not already exist

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

)

-- Converts rows into column using the RegEx expression

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'

WITH SERDEPROPERTIES (

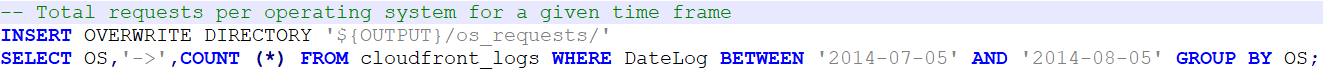
"input.regex" = "^(?!#)([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+[^\(]+[\(]([^\;]+).\*\%20([^\/]+)[\/](.\*)$"

) LOCATION '${INPUT}/cloudfront/data';

**2.1. Total Requests based on Operating System.**

The below query returns the OS name and the total number of requests based on the OS name between the date range of 05-07-2014 and 05-08-2014,

**Query Used:**



**S3 output:**

After the execution of the steps the below files are generated in the s3 os\_requests output path and have the below data representing the name of the OS along with the number of requests per OS,

Text, letter

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

**Inferences:**

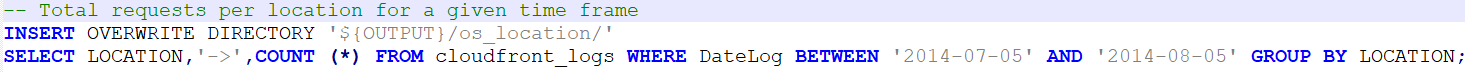
We can observe from the output that Windows OS has the most requests (883), followed by Android and macOS (855 and 852, respectively).

Although OSX and iOS are at the bottom of the list with 799 and 794 queries, respectively, the data flow between all operating systems between 05-07-2014 and 05-08-2014, is essentially identical. This implies that all platforms and systems can access Eurotrade.

**2.2. Total Requests based on Location.**

The below query returns the LOCATION name and the total number of requests based on the LOCATION name between the date range of 05-07-2014 and 05-08-2014,

**Query Used:**



**S3 Output:**

Graphical user interface, text, application, email

Description automatically generatedThe following files are created in the s3 location output directory after the steps have been completed and include the data below, which represents the name of the LOCATION and the number of requests per LOCATION,

Text, letter

Description automatically generated

**Inferences:**

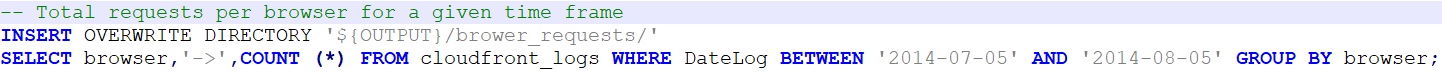
It is clear from the study that MIA3 location, which has the most requests overall with 385, and LAX1 location, which has the fewest, have roughly comparable numbers of requests.

So, it is crucial for Eurotrade to conduct an analysis and take the appropriate action to draw more clients to the LAX1 site.

**2.3. Total Requests based on the Internet browser.**

The below query returns the INTERNET BROWSER name and the total number of requests based on the INTERNET BROWSER name between the date range of 05-07-2014 and 05-08-2014,

**Query used:**



**S3 Output:**

**Text, table

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated**

**Inferences:**

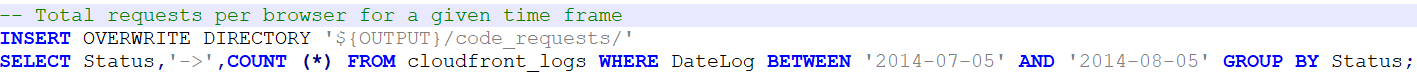
From the result, we can see that Lynx has the most requests (889), and on the other hand, IE has the fewest requests (774), but there isn't a lot of variation, and the data flow is split nearly evenly among all browsers.

Thus, Eurotrade can be accessed using any internet browser.

**2.4. Analysis of Status Code**

The below query returns the STATUS CODE name and the total number of STATUS CODES transferred based on the server between the date range of 05-07-2014 and 05-08-2014,

**Query used:**

****

**S3 output:**

**Text

Description automatically generated with low confidenceGraphical user interface, text, application, email

Description automatically generated**

**Inferences:**

From the output, we can see there are 2 types of status codes,

1. 304 **(Not Modified)** = 984 requests

2. 200 **(OK)** = 4012 requests

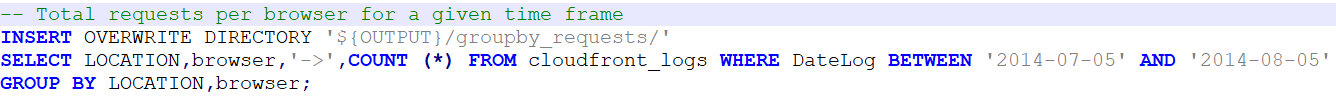
Success code 200 denotes a successful request, while status code 304 implies no changes were made and there is no need to send the request again.

This implied that the server connections were successful and that it operated flawlessly over the specified time frame. In addition, we did not encounter a 404 status code **(Not Available)**, which denotes that the website did not have any downtime.

**2.5 Total count of browsers used in different locations.**

The below query returns the log entry for unique location name and the count of different browsers used in locations between the date range of 05-07-2014 and 05-08-2014,

**Query Used:**

****

**Table

Description automatically generated**

**Inferences:**

Six browsers are utilized in total; however, it appears from the s3 output that not all of them are being used in all 14 nations. For instance, AMS1 used Firefox, Internet Explorer, and Lynx but not Opera, Chrome, or Safari. Therefore, it is essential for EuroTrade to investigate any problems with accessing browsers in certain places.

**Question3:** **Log management for Big Data Projects**

As the world progresses, the data also increases and as stated by **Stephen Marsland, “if data had mass, the earth would be a black hole.”.**  Every hour, terabytes of data are produced, and many businesses are struggling to store and process these enormous volumes of data. Big data has grown in importance for organizations over the past few years. Companies need to be able to handle data quickly and accurately, which calls for the usage of specialized software. The solutions also need to be flexible enough to keep up with the rapidly evolving data landscape.

Text, letter

Description automatically generated

Two things are clear from the aforementioned assertion,

1. More than 50% of organizations are working on big data initiatives.

2. A technique known as log management is utilized by more than half of businesses.

**3.1 What is big data?**

Big data is just an umbrella phrase for the accumulation of enormous and intricate data collections. Big data technologies are used by businesses to manage data, obtain business insights, analyze data, and enhance decision-making. They are applied throughout industries to gather insightful customer trends and behaviour data. Big data's primary benefit is its ability to analyze organized, semi-structured, and unstructured data sources; as a result, this growth in data can present both opportunities and challenges that can be avoided by making use of big data's unique properties,

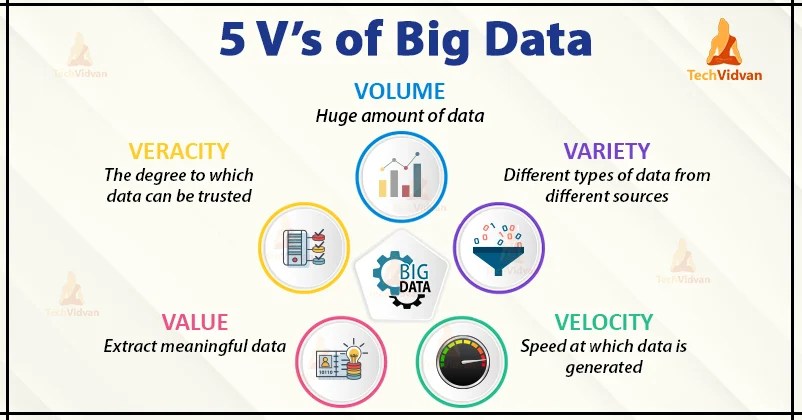
**1. Volume** - As the name says, it refers to the size and it establishes how much data an organization handles.

**2. Value** - This term refers to the advantages or understandings a corporation derives from the data to boost its commercial worth.

**3. Variety** - Managing many data kinds, including structured, semi-structured, and unstructured data

**4. Velocity** - The rate at which the business receives, controls, transforms and saves data is referred to as this.

**5. Veracity** - This term refers to the data's accuracy. It is the most crucial quality because it determines the company's reputation.



***Fig 1.4 The 5 Vs of Big Data***

**3.2 What is a log management and how can it be used in big data?**

According to the aforementioned assertion, adopting log management technology to examine the logs is crucial for businesses using or planning to use big data.

Logs are produced by numerous systems and may or may not contain important data that can be used to gain business insights. Log management is a method for gathering, examining, and managing log data from many sources inside an organization.

Although it has been theoretically demonstrated, combining, manipulating, and altering complex data sets is quite difficult. Therefore, with these kinds of requirements, log management enters the picture. With this approach, the software arranges log files in a particular order to satisfy search criteria, making it simple to find the right log file for SysAdmins, SecOps, DevOps analysis software, or for a developer to review and also troubleshoot issues.

Diagram

Description automatically generated

***Fig 1.5 Log management process***

Log Management contains 5 main elements:

**1. Log Collection and aggregation**

At any given time, logs are produced, and they are stored in various places. The log collector (Apache Flume, Logstach), which is responsible for buffering, parsing, and enriching the data, receives these logs either directly from the source or from user scripts. The data will now be collected and stored in a centralized storage system. Having a dedicated server for parsing, buffering, and enriching the data is also crucial. Doing the collecting and aggregation close to the source can aid in scaling. This data can be stored in a Hadoop distributed file system or a NoSQL database like MongoDB or Hbase.

**2. Log Search and Analysis**

Once the log files have been collected and saved, they may be searched for and analyzed, which can aid in locating security concerns, resolving problems, and improving performance. We use this stage in log management to identify insights, information that may be used to take action, and the main cause of problems because of the enormous volume of data created by many sources, which includes both structured and unstructured data. For rapid processing, the processing stage can be completed using the Spark framework, which runs on top of the Hadoop framework.

**3. Log Monitoring and Alerting**

Monitoring and alerting assist in delivering data regarding system performance as well as reports regarding security lapses and anomalies, if any. In a big data context, log monitoring entails gathering, analyzing, and interpreting logs produced by numerous systems and applications. Alerting is a further essential element of log management with huge data. The process of alerting consists of setting up automated notifications that are sent out when specific system events occur. For instance, if a Kerberos Ticket expires in a Hadoop environment, an alert message might be issued to the application support team.

**4. Log Visualization and reporting**

The final and most important part of log management is visualization and reporting, which will allow all members of the internal and external teams to access the same data. This enables people to learn more and take more knowledgeable judgments.

A significant addition to the business metrics is made by visualization and reporting, which uses graphs and charts like time series line charts and pie charts to provide a clear image of trends and patterns that are frequently difficult to see in raw data. Out of all the available types, dashboards like Tableau or PowerBI will have the biggest impact because they display a combination of graphs and charts that provide a comprehensive view of how the data is behaving.

Some of the benefits of log management are,

* Increased security and compliance
* Optimizing the performance of the system
* Detecting issues
* Supporting troubleshooting

**3.3 Conclusion:**

Every organization needs to adapt to new technologies and solutions in order to avoid becoming a bottleneck because the amount of data being generated is growing daily. Businesses also cannot afford to make any form of wrong analysis or decision. Due to the importance of big data, it is essential to store, manage, and analyze it. To handle such a massive number of data, it is needed to think about new and potent technologies. Through improving operational performance, security, and compliance, log management aids enterprises in better understanding their systems. On the basis of the aforementioned ideas, it is clear that log management needs to be employed in all big data initiatives, including Eurotrade, in order for them to be able to manage the large data issue when they go worldwide.

**4. References**

1. Carly Fiorina, ex CEO of Hewlett-Packard available at (online)

<https://techvidvan.com/tutorials/5-vs-of-big-data/> [Accessed February 24, 2023]

2. RAJESH KUMAR (MARCH 23, 2017) available at (online)

<https://www.devopsschool.com/blog/top-5-log-monitoring-tools/> [Accessed February 24, 2023]

3. Arfan Sharif - December 21, 2022 available at (online)

<https://www.crowdstrike.com/cybersecurity-101/observability/log-management/> [Accessed February 24, 2023]