Question 1

The script is used to analyse the total number of requests per operating system over a specific interval. Script is developed using HiveQL, which is a SQL-like scripting language for data warehousing and analysis.

The log file is stored in Amazon S3 at

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

Initially, a table is created with the name **cloudfront_logs**. Subsequent statements following the create table statement define the columns that are to be present in the table.

Column Description

1) date of the request- DateLog;
2) time of the request- Time;
3) location of the request- Location;
4) number of bytes used during a request- Bytes;
5) IP address of a request- RequestIP;
6) method by which the request is made- Method;
7) host, otherwise known as a system identifier- Host;
8) Uri- Uniform Resource Identifier stores the URL that the user is pointing to and the information being transferred;
9) Status of the request- Status;
10) Referrer, which identifies the webpage URL;
11) OS which dictates the operating system from which the request had come through;
12) Browser, which shows the browser from which the request came from;
13) BrowserVersion which holds the version of the browser stored in the Browser string.

Reads the CloudFront log files from Amazon S3 using EMRFS and parses the CloudFront log files using the regular expression serializer/deserializer (RegEx SerDe).Writes the parsed results to the Hive table cloudfront_logs.Submits a HiveQL query against the data to retrieve the total requests per operating system for a given time frame.Writes the query results to the Amazon S3 output bucket.

Following is the hive code that parses the log files using the RegEx SerDe:

```
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/';
```

The regex pattern is thus applied to the rows read to split them up into the appropriate columns.

Following is the HiveQL query:

INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT OS, COUNT(*) count FROM cloudfront_logs WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05' GROUP BY OS;

The location of the INPUT and OUTPUT is described by the Amazon S3 locations used when creating the step.

Question 2

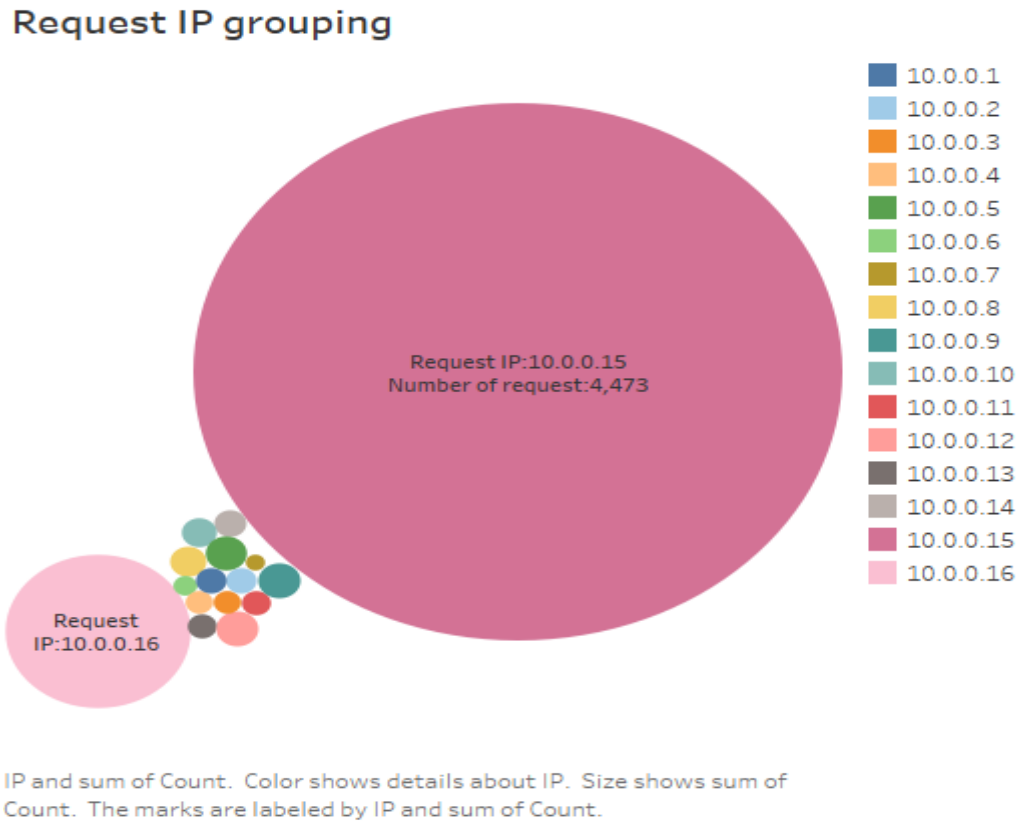
Analysis 1:

Request IP grouping to identify any suspicious activity or IP Spoofing

Query 1

```
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT
RequestIP,COUNT(*) AS NUMBER  FROM cloudfront_logs
WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05'
GROUP BY RequestIP
ORDER BY NUMBER DESC;
```

Result



From the above figure it can be seen that the number of request from a specific id address 10.0.0.15 is about 4,473 which is exceedingly high when compared to other values.
Please find the below analysis to confirm the suspicious request count for 10.0.0.15

Query 2

```
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT Location,COUNT(*) AS NUMBER  FROM cloudfront_logs
WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05'
AND RequestIP = '10.0.0.15'
GROUP BY Location
ORDER BY NUMBER DESC;
```

Result

MIA3->374
IAD2->370
AMS1->367
FRA2->366
LHR3->366
SFO4->364
HKG1->364
EWR2->363
STL2->360
DFW3->360
SEA4->360
NRT4->335
LAX1->124

From the above result it can be observed that the request IP -'10.0.0.15' is sending request to the server from multiple locations over such a short time interval of two days. Hence it can be inferred that it is an instance of IP spoofing – method to increase the server load by sending invalid requests.

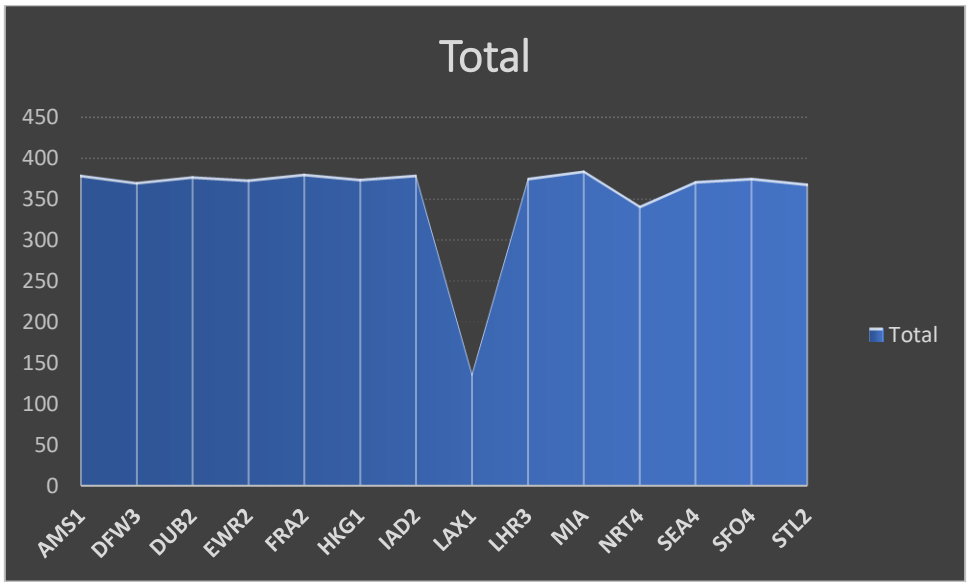
Analysis 2:

Location wise grouping of requests

Query 3

```
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT Location,COUNT(*) AS NUMBER FROM cloudfront_logs
WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05'
GROUP BY Location
ORDER BY NUMBER DESC;
```

Result



From the above figure it can be inferred that location LAX1 is having the lowest number of requests . *Hence Eurostar can focus on campaigns of its products and services at LAX1 to increase it's user base.*

To understand the user distribution based on OS at location LAX1 which will enable the business to focus on the specific category of users (mobile/Desktop), I ran the below query

Query 4

```
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT OS,Location,COUNT(*) AS NUMBER FROM cloudfront_logs
WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05'
AND Location = 'LAX1'
GROUP BY OS,Location
ORDER BY NUMBER DESC;
```

Result

OS	Count
MacOS	30
Android	24
OSX	22
iOS	22
Linux	21
Windows	18

Analysis 3:

Time wise grouping of requests

Query 5

```
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT
SUBSTR(Time,1,2),COUNT(*) AS NUMBER  FROM cloudfront_logs
WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05'
GROUP BY SUBSTR(Time,1,2)
ORDER BY NUMBER DESC;
```

Result

Time	Count
14	2000
20	1998
15	998

From the above table it can be inferred that majority of the customers of Eurostar products and services are active during 14:00-14:59 and 20:00-20:59. *Hence the company can focus to increase customer support and server monitoring active during these timeframes.*

Analysis 4:

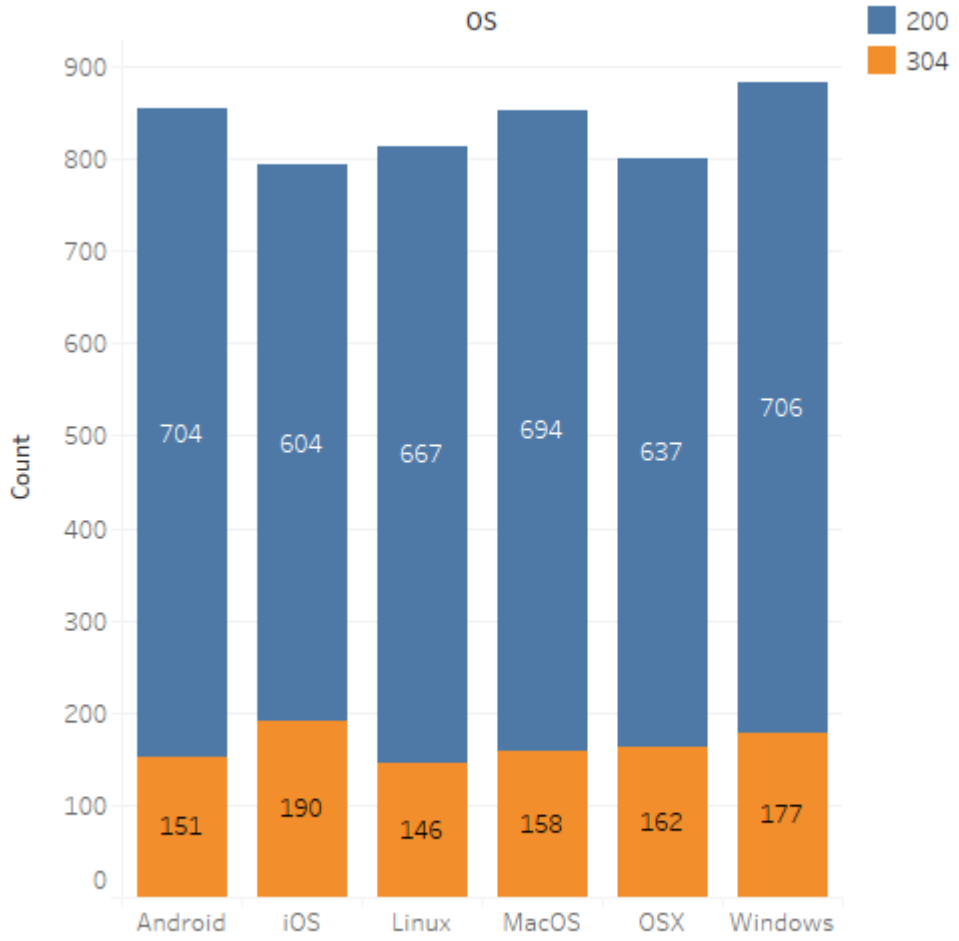
Status wise grouping of request

Analysis based on web server status code aids to provide insights on user activity through different interfaces.
304 - **Not Modified** client redirection response code indicates that there is no need to retransmit the requested resources.
200- **OK response**- entity will be transferred based on request type.

Query 6

```
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT OS,
Status,COUNT(*) AS NUMBER  FROM cloudfront_logs
WHERE DateLog BETWEEN '2014-07-05' AND '2014-08-05'
GROUP BY OS,Status
ORDER BY NUMBER DESC;
```

Result



Sum of Count for each OS. Color shows details about Status. The marks are labeled by sum of Count.

To increase or engage the user business needs to focus on attractive offers or better user experience through apps . Measures like adding reviews for products and provision to return the products incase of issues will help the company to build a good relationship with the customer

Question 3

Big data plays a pivotal role in the growth of modern day organizations. Data is increasing at an exponential rate. Hence log management is one of leading use case for big data. Log management is the collective processes and policies used to administer and facilitate the generation, transmission, analysis, storage, archiving and ultimate disposal of the large volumes of log data created within an information system. [1]

There are two common log types:

- Event logs

  Provide an overview on how your systems and its related components are performing at any given point of time
  e.g. Any network failures or server downtimes etc.

- User logs

  User logs are more aligned to an individual which helps organizations to study user search pattern and behaviour. It is also termed as user profiling which is used to provide the customer with a customized set of options according to his interest.

  e.g. Google Analytics

Major focus of my research is to analyse the significance of log management which is insisting companies to invest in big data. Log Analytics is the study of events which an organization uses to understand user behaviours, identify problems, audit security activities or ensure compliance with established rules, and plan for capacity or IT infrastructure changes.[2]. Event is any activity under the umbrella of an information system.

Organizations rely on tools, such as Windows Event Viewer for Windows OS, or the application SolarWinds Event & Log Manager, to access, view and analyse logs. Log analytics software aids organizations to parse through error logs and locate the same. These software also helps to determine trends in an application, aggregate data from different sources to provide a comprehensive view on the performance of the entire system.

Log analytics software collects information on an activity or event. As the number of events are too high traditional approaches will take days or even weeks for analysis on specific incident. Bigdata offers a cost effective scalable solution to collect this data, but also provides analytics tools to look for long-term and subtle patterns that might be undetected by traditional rules-based and signature-based approaches.

Sensitive data , much of the data stored in big data systems is sensitive data . Hence security of these systems becomes a critical issue. Log management aids this with the ability to generalize trends from past events . Hence any suspicious activity can be seen as a threat and alerts could be generated.

Listed below are few of the manager advantages of log management[3]

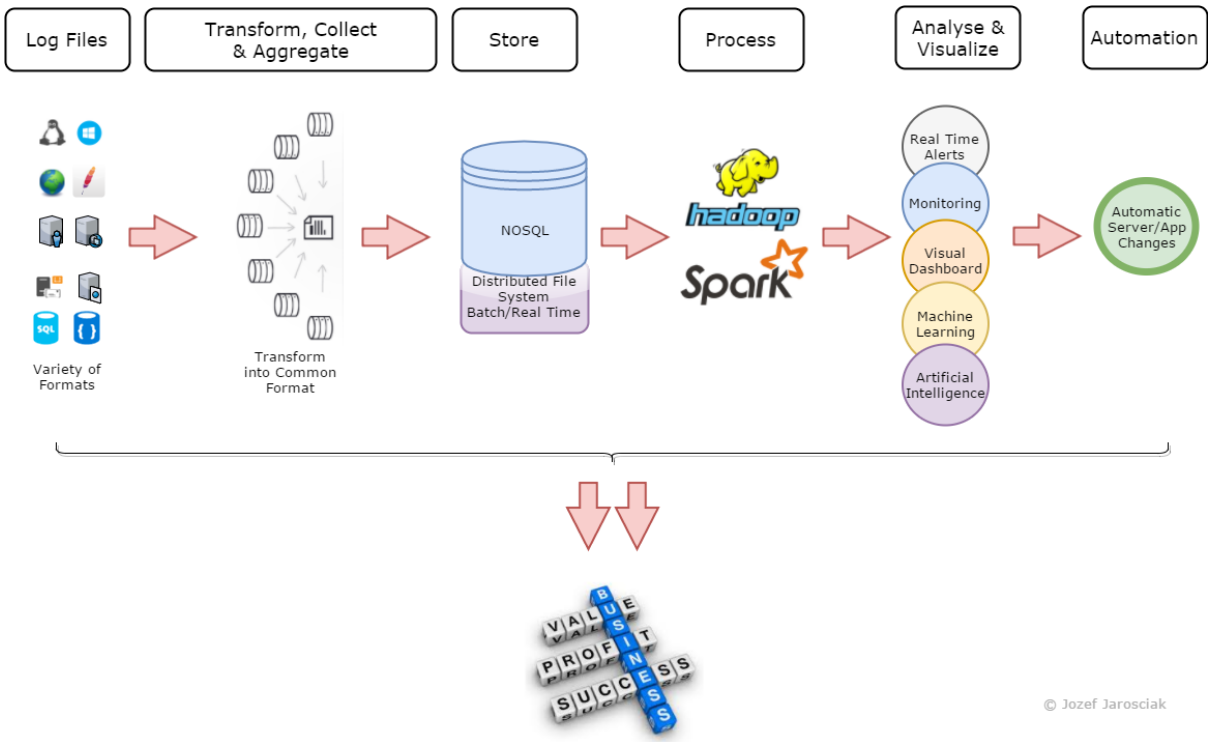
ADVANTAGES

- Log collection and Centralization
  When numerous applications are running over geographically spread locations it becomes a very tedious process to go to each file and manually check logs for each location. A centralized log management system will aggregate and store logs allowing access to search and manage logs from a single location.
- Parse Logs and Analyze Events
  Logs are obtained in different formats and layout based on the collection method. Log management tools automatically identify certain formats and extract individual fields in real time.
- Early problem detection and alerting
  Log management helps to identify trends and which is then analysed to create alerts if any suspicious activities are observed.
- Security and threat detection
  Real-time monitoring and parsing lets you scan logs for signs of security vulnerabilities, cyber threats, and malicious activity. Pinpoint the time and origin of a cyberattack by filtering events by time range, host system, or application.
- Complete Log Management for Your Entire Infrastructure

  Log management tools give you the ability to manage your logs from a single location regardless of the size of your infrastructure.

Applying Big data analytics to Logging

Architecture



Log Transformation, Collection, and Aggregation

Logs from different sources and combined into a single centralized source. Log files are transformed into a single common format during or after the process. This is the crucial initial step what allows the data contained in the logs to be later processed and analysed.[4]

Some of the tools used for log aggregation are *Scribe, Apache Flume, logstash,* that support shipping, parsing and indexing logs as well as moving large amounts of data into storage.

Log Storage

Once the data is logged it is moved to NoSQL database that sits on top of a distributed file system such as HDFS. Selection of database is case driven by the data store type for which NoSQL is designed.

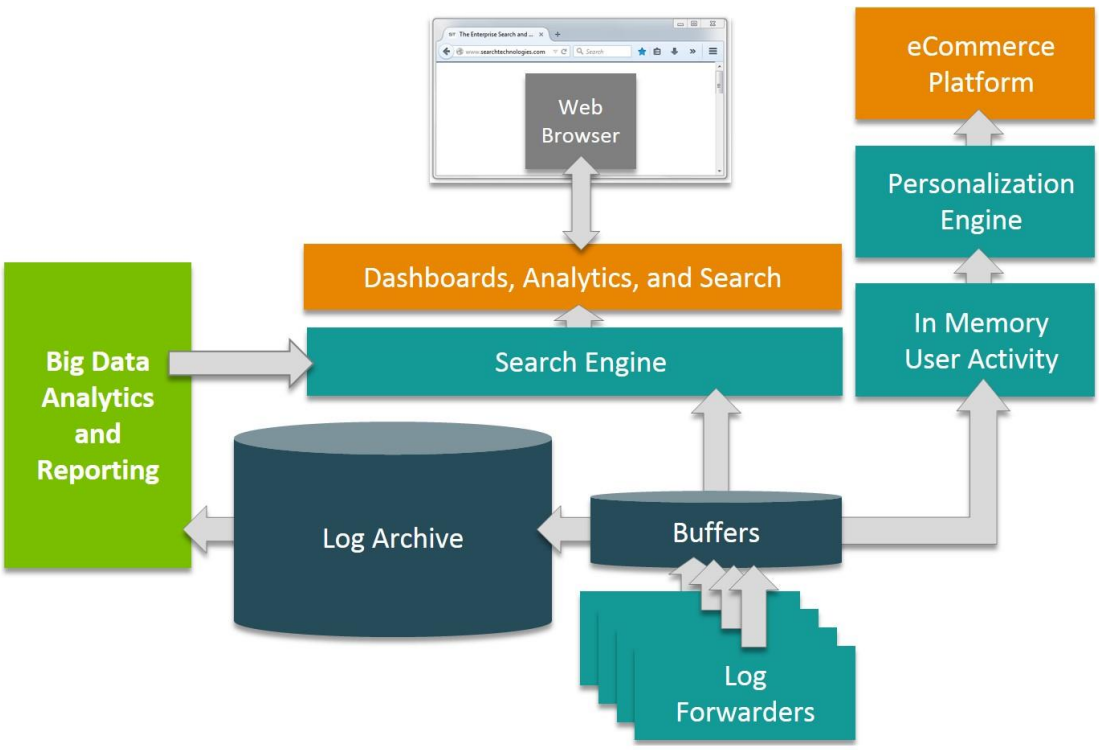
Log Processing

Various big data technologies like Hadoop , Apache Spark provide efficient ways to process the logs and gain an overview of what is happening on all systems.These tools helps to reduce the information which aids faster analyses.

Log Analysis and Visualization

Log analysis phase uses advanced big data analytics techniques like data mining, machine learning , pattern recognition and applied statistics. Visualization helps in detecting DDoS(distributed denial of service) attacks efficiently by classifying machines as authorized and unauthorized.

Open source approach to log analysis using big data



Big data log analytics platform goes through the following steps[5]

1.Gathers data in raw form from multiple business systems

2.Runs the data through buffers

3.Loads the data into log analytics stack for query parsing , search indexing and trend visualization.

4.Allows businesses to perform large-scale analysis of user trends, clustering, clustering trends, market trends, etc.

Below are some of the tools used in big data- log analytics

- Elasticsearch: import of log files into a search engine for indexing and access through search
- Logstash: log gathering, storage, and parsing
- Kibana: intuitive browser interface for trend visualization and analysis

Conclusion

Implementing a Big Data solution to log management not only gives the ability to visually monitor entire hardware and software system through the visual portal and various analytic dashboard, but also allows us to see the overall health of all systems, use custom reporting, get the status of automated monitoring[4].Big data solution helps to build improved solutions prediction and recommendation models, use continual tuning (automatically change hardware and software configuration in response to results of analytics), and generally to drive more automation into the entire process of working with logs. Hence resulting in better business insights which paves way for more revenue generation.

Log analysis is being used for monitoring the servers and applications to improve business and customer intelligence, prevent fraud thereby enhancing the overall security of the system. Big data solutions enhance log management through automation which reduces the manual effort thereby generating more profit.

References

[1] https://searchitoperations.techtarget.com/definition/log-management

[2] https://searchitoperations.techtarget.com/definition/log-analytics

[3] https://logdna.com/benefits-of-log-management

[4] https://www.joe0.com/2017/02/05/applying-big-data-analytics-to-logging/

[5] https://www.searchtechnologies.com/blog/big-data-open-source-log-analytics