Business Analytics Using Hadoop

Project - Apache Log Analysis

Group 6:
I008 - Anusha Buch
I020 - Rishita Jain
N077- Shruti Golchha
N087 - Shivani Pundhir
N092 - Vishal Shah
N100 - Simran Chauhan

Project Overview

Big Data refers to a large set of data that can be analysed by means of computational techniques to draw patterns and reveal the common or recurring points that would help to predict the next step.

The project is about parsing the apache log file and reading its contents in the dataframe. This was done to achieve a prototype of the main project.

Thereafter, the main data i.e. the web logs from NASA website was rendered as an Apache dataset. This dataset was analysed on a monthly basis of server hits, page requests, and data downloaded.

The analysis of the dataset was done in Cloudera Hadoop using :-

- Apache Pig, which is a high-level platform for creating programs that run on Apache Hadoop and can also execute jobs in MapReduce.
    - Hive, which is a data warehouse used for summarisation, querying and analysis of data

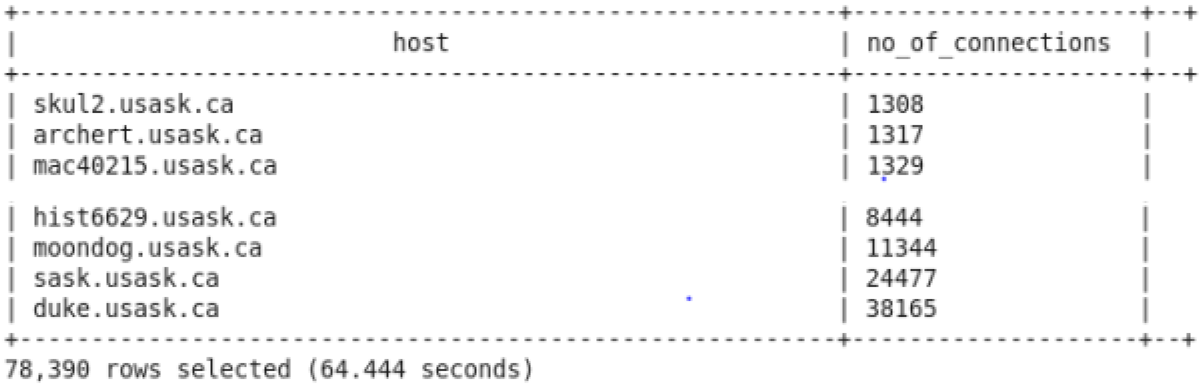
Commands Used

```
hdfs dfs -put   /home/cloudera/apache/apache_dataset.log  /mypig/apache/input/
```

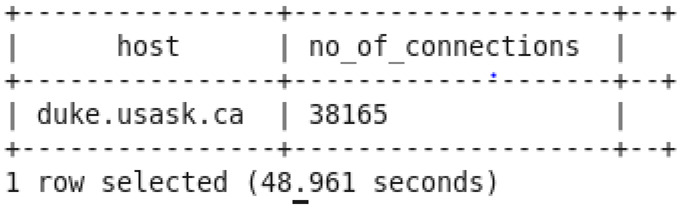
//

```
REGISTER /home/cloudera/apache/piggybank.jar;
DEFINE          ApacheCommonLogLoader
org.apache.pig.piggybank.storage.apachelog.CommonLogLoader();
DEFINE LogLoader org.apache.pig.piggybank.storage.apachelog.CombinedLogLoader();


LOGLINES = LOAD '/mypig/apache/input/apache_dataset.log' USING
ApacheCommonLogLoader AS (host, hclient, userid, logtime, method, pagerequest, protocol,
serverstatus, sentbytes);

b = foreach LOGLINES generate host as host:chararray, hclient as hclient:chararray, userid as
userid:chararray, ToDate(logtime,'dd/MMM/yyyy:HH:mm:ss Z') as (logtime:DateTime),method as
method:chararray, pagerequest as pagerequest:chararray, flatten(STRSPLIT(protocol,'/')) as
(protocol:chararray,version:chararray), serverstatus as serverstatus:chararray, sentbytes as
sentbytes:int;

c = foreach b generate host as host:chararray, hclient as hclient:chararray, userid as
userid:chararray, ToString(logtime, 'yyyy-MM-dd') as (logdate:chararray),ToString(logtime,
'HH:mm:ss') as (logtime:chararray), method as method:chararray, pagerequest as
pagerequest:chararray, protocol as protocol:chararray, version as version:chararray,serverstatus
as serverstatus:chararray, sentbytes as sentbytes:int;

STORE c into '/apache/hiveinput/apache_dataset_full.log' using PigStorage(',');

beeline -u jdbc:hive2:// ###to be used in hdfs command prompt

create database apachelogs;
use apachelogs;

create table nasalogs (host string, hclient string, userid string, logdate string,logtime string,
method string, pagerequest string, protocol string, version string, serverstatus string, sentbytes
int) row format delimited fields terminated by ',';
load data inpath '/apache/hiveinput/apache_dataset_full.log' into table nasalogs;
```

q1.
select host, count(*) as no_of_connections from nasalogs group by host order by no_of_connections;
select host, count(*) as no_of_connections from nasalogs group by host order by no_of_connections desc limit 1;

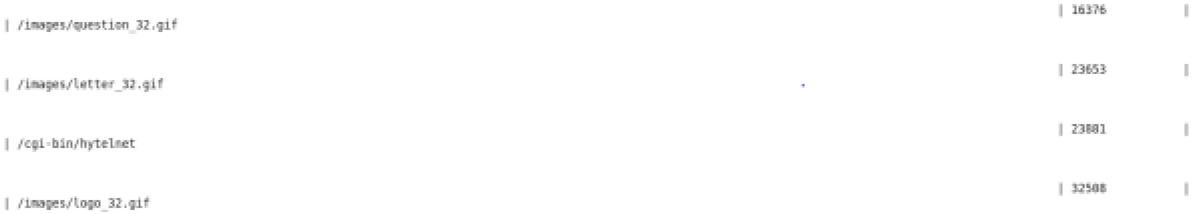


```
+---------------------------------------------------+-------------------+--+
|                       host                        | no_of_connections |
+---------------------------------------------------+-------------------+--+
| skul2.usask.ca                                    | 1308              |
| archert.usask.ca                                  | 1317              |
| mac40215.usask.ca                                 | 1329              |

| hist6629.usask.ca                                 | 8444              |
| moondog.usask.ca                                  | 11344             |
| sask.usask.ca                                     | 24477             |
| duke.usask.ca                                     | 38165             |
+---------------------------------------------------+-------------------+--+
78,390 rows selected (64.444 seconds)
```

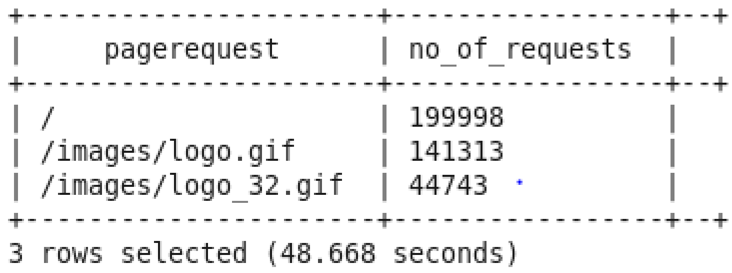


```
+----------------+--------------------+--+
|      host      | no_of_connections  |
+----------------+--------------------+--+
| duke.usask.ca  | 38165              |
+----------------+--------------------+--+
1 row selected (48.961 seconds)
```

q2
select pagerequest, count(*) as no_of_requests from nasalogs group by pagerequest order by no_of_requests;
select pagerequest, count(*) as no_of_requests from nasalogs group by pagerequest order by no_of_requests desc limit 3;



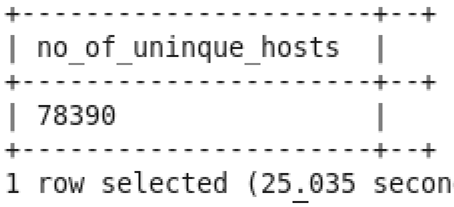
```
| /images/question_32.gif                                                    | 16376          |
| /images/letter_32.gif                                                      | 23653          |
| /cgi-bin/hytelnet                                                          | 23881          |
| /images/logo_32.gif                                                        | 32508          |
```



```
+---------------------+---------------+--+
|     pagerequest     | no_of_requests  |
+---------------------+---------------+--+
| /                   | 199998          |
| /images/logo.gif    | 141313          |
| /images/logo_32.gif | 44743           |
+---------------------+---------------+--+
3 rows selected (48.668 seconds)
```

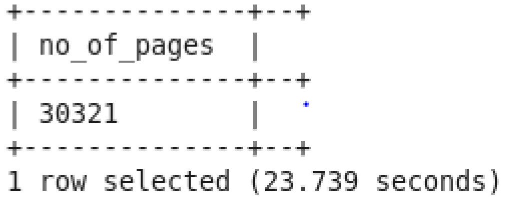
;

q3
select count(distinct(host)) as no_of_uninque_hosts from nasalogs



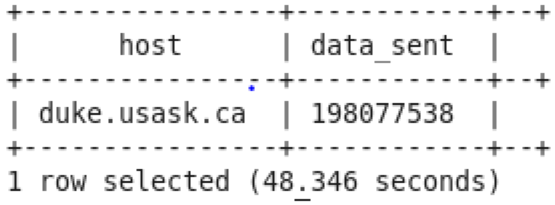
```
+---------------------+--+
| no_of_uninque_hosts  |
+---------------------+--+
| 78390                |
+---------------------+--+
1 row selected (25.035 seconds)
```

q4
select count(distinct(pagerequest)) as no_of_pages from nasalogs;



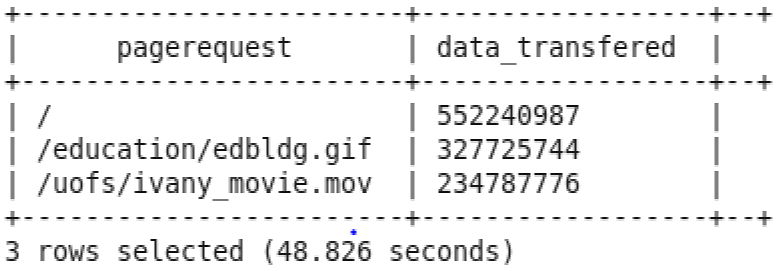
```
+--------------+--+
| no_of_pages  |
+--------------+--+
| 30321        |
+--------------+--+
1 row selected (23.739 seconds)
```

q5
select host, sum(sentbytes) as data_sent from nasalogs group by host order by data_sent desc limit 1;



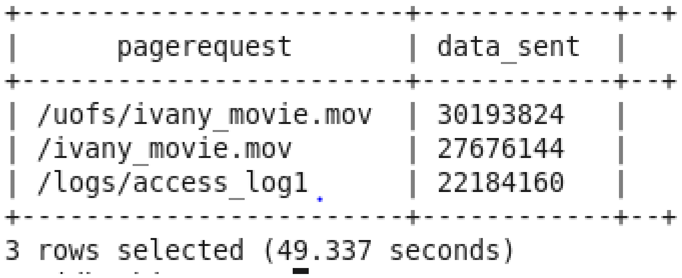
```
+---------------+------------+--+
|     host      | data_sent  |
+---------------+------------+--+
| duke.usask.ca  | 198077538  |
+---------------+------------+--+
1 row selected (48.346 seconds)
```

q6
select pagerequest, sum(sentbytes) as data_transfered from nasalogs group by pagerequest order by data_transfered desc limit 3;



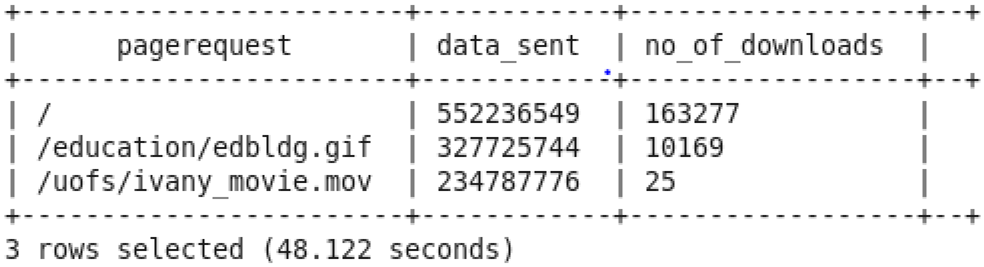
```
+------------------------+------------------+--+
|      pagerequest       | data_transfered  |
+------------------------+------------------+--+
| /                      | 552240987        |
| /education/edbldg.gif  | 327725744        |
| /uofs/ivany_movie.mov  | 234787776        |
+------------------------+------------------+--+
3 rows selected (48.826 seconds)
```

q7
select pagerequest, max(sentbytes) as data_sent from nasalogs where serverstatus >= 200 and serverstatus < 300 group by pagerequest order by data_sent desc limit 3;



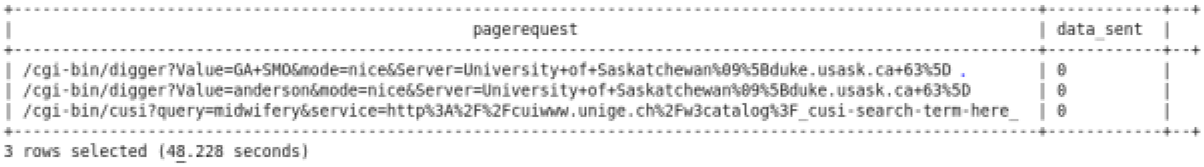
```
+------------------------+------------+--+
|      pagerequest       | data_sent  |
+------------------------+------------+--+
| /uofs/ivany_movie.mov  | 30193824   |
| /ivany_movie.mov       | 27676144   |
| /logs/access_log1      | 22184160   |
+------------------------+------------+--+
3 rows selected (49.337 seconds)
```

q8
select pagerequest, max(sentbytes) as data_sent, count(*) as no_of_downloads from nasalogs where serverstatus >= 200 and serverstatus < 300 group by pagerequest order by data_sent desc limit 3;



```
+------------------------+------------+------------------+--+
|      pagerequest       | data_sent  | no_of_downloads  |
+------------------------+------------+------------------+--+
| /                      | 552236549  | 163277           |
| /education/edbldg.gif  | 327725744  | 10169            |
| /uofs/ivany_movie.mov  | 234787776  | 25               |
+------------------------+------------+------------------+--+
3 rows selected (48.122 seconds)
```

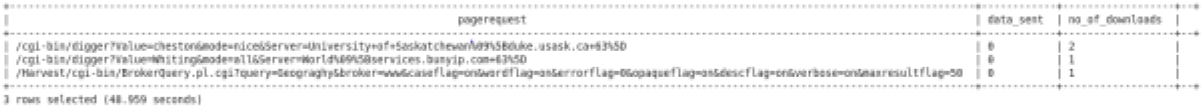
q9
select pagerequest, min(sentbytes) as data_sent from nasalogs where sentbytes >= 0 group by pagerequest order by data_sent limit 3;



```
+----------------------------------------------------------------------------------------------------------+------------+--+
|                                                pagerequest                                               | data_sent  |
+----------------------------------------------------------------------------------------------------------+------------+--+
| /cgi-bin/digger?Value=GA+SMO&mode=nice&Server=University+of+Saskatchewan%09%5Bduke.usask.ca+63%5D        | 0          |
| /cgi-bin/digger?Value=anderson&mode=nice&Server=University+of+Saskatchewan%09%5Bduke.usask.ca+63%5D      | 0          |
| /cgi-bin/cusi?query=midwifery&service=http%3A%2F%2Fcuiwww.unige.ch%2Fw3catalog%3F_cusi-search-term-here_ | 0          |
+----------------------------------------------------------------------------------------------------------+------------+--+
3 rows selected (48.228 seconds)
```

q10

select pagerequest, min(sentbytes) as data_sent, count(*) as no_of_downloads from nasalogs where serverstatus >= 200 and serverstatus < 300 group by pagerequest order by data_sent limit 3;

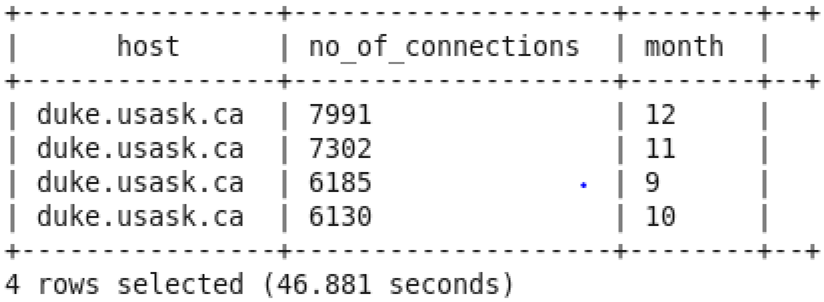


```
+----------------------------------------------------------------------------------------------------------------------------------------------------+------------+------------------+--+
|                                                                    pagerequest                                                                     | data_sent  | no_of_downloads  |
+----------------------------------------------------------------------------------------------------------------------------------------------------+------------+------------------+--+
| /cgi-bin/digger?Value=cheston&mode=nice&Server=University+of+Saskatchewan%09%5Bduke.usask.ca+63%5D                                                 | 0          | 2                |
| /cgi-bin/digger?Value=Whiting&mode=all&Server=World%09%5Bservices.bunyip.com+63%5D                                                                 | 0          | 1                |
| /Harvest/cgi-bin/BrokerQuery.pl.cgi?query=Geography&broker=www&caseflag=on&wordflag=on&errorflag=0&opaqueflag=on&descflag=on&verbose=on&maxresultflag=50 | 0          | 1                |
+----------------------------------------------------------------------------------------------------------------------------------------------------+------------+------------------+--+
3 rows selected (48.959 seconds)
```

q11

select host, count(*) as no_of_connections, month(logdate) as Month from nasalogs group by host, month(logdate) order by no_of_connections DESC,month;

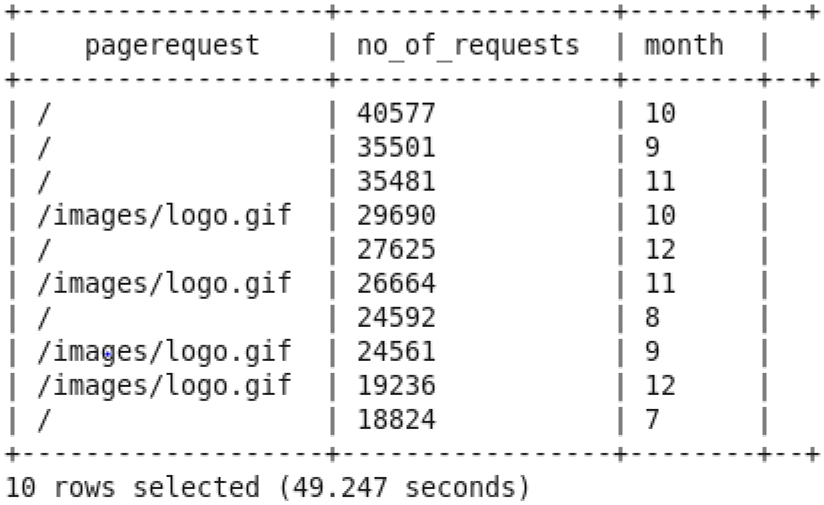
select host, count(*) as no_of_connections, month(logdate) as Month from nasalogs group by host, month(logdate) order by month, no_of_connections DESC limit 4;



```
+----------------+--------------------+--------+--+
|      host      | no_of_connections  | month  |
+----------------+--------------------+--------+--+
| duke.usask.ca  | 7991               | 12     |
| duke.usask.ca  | 7302               | 11     |
| duke.usask.ca  | 6185               | 9      |
| duke.usask.ca  | 6130               | 10     |
+----------------+--------------------+--------+--+
4 rows selected (46.881 seconds)
```

q12

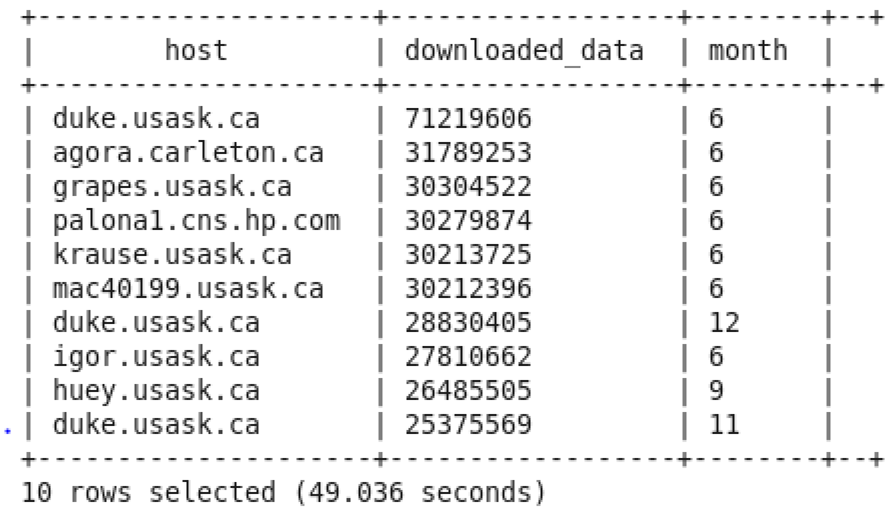
select pagerequest, count(*) as no_of_requests, month(logdate) as Month from nasalogs group by month(logdate),pagerequest order by no_of_requests DESC,month limit 10;



```
+-------------------+-----------------+--------+--+
|    pagerequest    | no_of_requests  | month  |
+-------------------+-----------------+--------+--+
| /                 | 40577           | 10     |
| /                 | 35501           | 9      |
| /                 | 35481           | 11     |
| /images/logo.gif  | 29690           | 10     |
| /                 | 27625           | 12     |
| /images/logo.gif  | 26664           | 11     |
| /                 | 24592           | 8      |
| /images/logo.gif  | 24561           | 9      |
| /images/logo.gif  | 19236           | 12     |
| /                 | 18824           | 7      |
+-------------------+-----------------+--------+--+
10 rows selected (49.247 seconds)
```

q13

select host,sum(sentbytes) as downloaded_data, month(logdate) as Month from nasalogs where serverstatus >= 200 and serverstatus < 300 group by host, month(logdate) order by downloaded_data desc;



```
+--------------------+----------------+-------+--+
|        host        | downloaded_data | month  |
+--------------------+----------------+-------+--+
| duke.usask.ca      | 71219606        | 6      |
| agora.carleton.ca  | 31789253        | 6      |
| grapes.usask.ca    | 30304522        | 6      |
| palona1.cns.hp.com | 30279874        | 6      |
| krause.usask.ca    | 30213725        | 6      |
| mac40199.usask.ca  | 30212396        | 6      |
| duke.usask.ca      | 28830405        | 12     |
| igor.usask.ca      | 27810662        | 6      |
| huey.usask.ca      | 26485505        | 9      |
| duke.usask.ca      | 25375569        | 11     |
+--------------------+----------------+-------+--+
10 rows selected (49.036 seconds)
```

q14
select pagerequest,sum(sentbytes) as data_sent, month(logdate) as Month from nasalogs group by pagerequest, month(logdate) order by Month desc, data_sent desc limit 10;

Summary

Dealing with a ‘big data’ like the above mentioned dataset could only be possible in Hadoop due to its capability of storing and processing large amounts of data of various kinds. There is no need to preprocess the data before storing it. Hadoop is highly scalable as it can store and distribute large data sets over several machines running in parallel. This framework is free and uses cost-efficient methods.