2016

ANALYSIS REPORT ON USAGE OF VDI MACHINES IN CMU

DATA ANALYSIS OF USER AND SYSTEM LOG DATA

ARUN CHALLA

Table of Contents

[Problem: 3](#_Toc457808468)

[Solution: 3](#_Toc457808469)

[Keywords: 3](#_Toc457808470)

[Methodology: 3](#_Toc457808471)

[Questions: 1.How many users were on the system total? 3](#_Toc457808472)

[3.What is the highest number of users per day? 3](#_Toc457808473)

[4. Top 5 users by length of time logged in? 4](#_Toc457808474)

[5.Top 5 applications by number of times run? Click here for Query 4](#_Toc457808475)

[Analysis and Visualization: 5](#_Toc457808476)

[Usage of VDI Machines per Month by average users- 5](#_Toc457808477)

[Average Users Vs Total logins per month- 6](#_Toc457808478)

[Top 5 Users by Length of time logged in: 7](#_Toc457808479)

[Top 5 Applications by number of times run: 8](#_Toc457808480)

[Number of users logged in per day: 9](#_Toc457808481)

[Technical Information: 10](#_Toc457808482)

[Tables: 10](#_Toc457808483)

[vdi\_serverlogs table: 10](#_Toc457808484)

[vdi\_apps table: 10](#_Toc457808485)

[Views: 10](#_Toc457808486)

[slogs view: 10](#_Toc457808487)

[slogs\_avg view: 10](#_Toc457808488)

[Query 1: Total number of users on the system. 10](#_Toc457808489)

[Query 2: Average number of users per day. 10](#_Toc457808490)

[Query 3: Highest number of users per day. 11](#_Toc457808491)

[Query 4: Top 5 users by length of time logged in. 11](#_Toc457808492)

[Query 5: Top 5 Applications by number of times run. 11](#_Toc457808493)

[Query 6: Usage of VDI Machines per Month by average users 11](#_Toc457808494)

[Query 7: Average Users Vs Total logins per month 11](#_Toc457808495)

[Query 8: Number of users logged in per day 11](#_Toc457808496)

[Analysis and Visualization tools: 12](#_Toc457808497)

[Errors / Outliers/ Bad data: 12](#_Toc457808498)

[Limitations/ Drawbacks: 13](#_Toc457808499)

[References: 13](#_Toc457808500)

# Problem:

Identify the user statistics and application statistics of VDI Machines at CMU based on the server logs and application statistics data captured on CMU servers for the year 2015.

# Solution:

A detailed analysis report is achieved by filtering and analyzing the data from the logs and statistics and then visualizing the data required. Such graphical and tabular visualizations provide a clear picture on the data analysis of system and user data.

# Keywords:

Data Analysis, log file analysis, application statistics, Data Visualization.

# Methodology:

Raw data containing huge datasets is given. Tables are created and all the data is loaded into these tables to make the analysis. Hive Query Language (HQL) is used to find answers to various business questions and detailed analysis is made using the visualization methodology in Excel worksheet. Each visualization is clearly presented to make the business decisions. The two tables are compared using “joins” to filter the data wherever necessary. There is a lot of information captured but only the required fields are selected for analysis. All the rows containing bad data is eliminated.

## Questions: 1.How many users were on the system total?

Total 19 users have used the CMU VDI machines in the year 2015.  
Click here for  [Query](#_Query_1:_Total)  
2. What is the average number of users per day?  
Average number of users per day in 2015 using CMU VDI machines are 14.  
Click here for  [Query](#_Query_2:_Average)

### 3.What is the highest number of users per day?

Highest number of users per day are 18.  
Click here for  [Query](#_Query_3:_Highest)  
4. Top 5 users by length of time logged in?

Click here for  [Query](#_Query_4:_Top)

|  |  |
| --- | --- |
| User\_ID | Length of Time Logged in  (seconds) |
| userid25355 | 30267480 |
| userid3341 | 24841500 |
| userid26603 | 21888420 |
| userid24832 | 20495520 |
| userid12 | 19032900 |

### 5.Top 5 applications by number of times run? Click here for [Query](#_Query_5:_Top)

|  |  |
| --- | --- |
| Application | No.of Times Run |
| Iexplore | 27247 |
| Devenv | 23201 |
| WINWORD | 20192 |
| Firefox | 16208 |
| SnippingTool | 8399 |

# Analysis and Visualization:

To filter and analyze the data according the questions and objectives, we have used the Sandbox- Hive online tool. Data is summarized according to the requirement and downloaded the respective filtered data in .csv formats.  
Now, to visualize this data, we have imported to MS Excel so as to present the tabular data more graphically and pictorially. Below are the few analytical and interpreted data visualizations which explains the statistics of the applications usage and Machines usage by users in CMU for the year 2015.

Usage of VDI Machines per Month by average users-

* The average number of users using VDI machines for every month in the year of 2015. The below visual explains that, in 2015 for the month of January, on an average 18 users used 193 CMU VDI machines.
* We can observe that the usage statistics are decreased during the month of May, June, July and August. This could be due to the summer vacations the students and faculty have left the college. As the summer term is not a regular academic year. So, the usage would probably decrease.
* From the statistics, we can forecast that the CMU machines usage may decrease in the month of December and January probably due the winter holidays.   
    
  Click here for  [Query](#_Query:_Usage_of)

## Average Users Vs Total logins per month-

* The total number of logins made by average number of users with CMU machines are visualized below in bar graph.
* From this analysis, we can see maximum number of logins are made in the month of February, i.e. 466 logins by an average number of 17 users.
* Total number of logins recorded are also recorded as 3195.
* According the observation from statistics, logins recorded in August are least. This could be due to students and faculty leaving the campus during summer holidays.
* The average number of users using the CMU VDI machines are least during August for the same reason.

Click here for  [Query](#_Query:_Average_Users)

## Top 5 Users by Length of time logged in:

* The top 5 users by length of time logged in are visualized below in a pie chart.
* From the table we can see the longest duration a user logged in to CMU machines is by userid25355 as 30267480 seconds.
* It can be calculated as approximately 350 days of logged in. But the actual number of dates logged in to the VDI machines is 223 days.
* This discrepancy is an outlier and is understood that, the records could be due to multiple logins of a user in multiple machines.

Click here for  [Query](#_Query_4:_Top_1)

|  |  |
| --- | --- |
| User\_ID | Length of Time Logged in  (seconds) |
| userid25355 | 30267480 |
| userid3341 | 24841500 |
| userid26603 | 21888420 |
| userid24832 | 20495520 |
| userid12 | 19032900 |

## Top 5 Applications by number of times run:

* Top 5 applications used by users in the year 2015 for CMU VDI machines are visualized below in the pie chart and table.
* Top used app is Internet explorer with 27247 times run by users.
* Next to that, Devenv which belongs to Visual Studio, a Microsoft development suite.
* From data, it can be seen snipping tool is also used highly. On this observations, we can interpret that most students are using this application for assignments to paste the screenshots of their work.
* Also, as IE is the default browser in VDI machines, it encounters most frequently to open any linked file in web browser.Click here for  [Query](#_Query_5:_Top_1)

|  |  |
| --- | --- |
| Application | No.of Times Run |
| Iexplore | 27247 |
| Devenv | 23201 |
| WINWORD | 20192 |
| Firefox | 16208 |
| SnippingTool | 8399 |

## Number of users logged in per day:

* Here, we can see the number of users logged in for each day in the year 2015 in CMU-VDI machines.
* From the below visual, we can observe the maximum number of users logged in a day is 18.
* During the start of January month and end of August month the minimum users are logged in to the systems.
* From all the user login observations, an observation made is during the summer the number of users using machines and applications are less than the regular academic year.
* From maintenance point of view most of the updates or upgrades on machines can be performed during the month of August as minimum number of users access the systems.

Click here for  [Query](#_Query:_Number_of)

Visualizations of the above mentioned analysis can be viewed by double clicking the below excel icon.

# Technical Information:

## Tables:

### vdi\_serverlogs table:

create external table vdi\_serverlogs (vdi\_id int, comp\_name string,  
user\_id string, logon string, logout string, connection\_server string,

remote\_ip string, remote\_od string, avg\_cpu float)

row format delimited fields terminated by ',' location '/user/root/vdi2/';

vdi\_apps table:  
create external table vdi\_apps(appid int, vdi\_id int,   
appname string, appstart string, appstop string)

row format delimited fields terminated by ',' location '/user/root/vdi2/';

## Views:

### slogs view:

create view if not exists slogs as select vdi\_id, comp\_name,

user\_id,

to\_date(from\_unixtime(unix\_timestamp(logon,"MM/dd/yyyy ”))) as login,

to\_date(from\_unixtime(unix\_timestamp(logout,"MM/dd/yyyy"))) as logoff,

avg\_cpu from serverlogs where comp\_name LIKE 'CMU%' AND

year(from\_unixtime(unix\_timestamp(logon,"MM/dd/yyyy"))) = 2015 AND

year(from\_unixtime(unix\_timestamp(logout,"MM/dd/yyyy"))) = 2015 AND

user\_id != '#N/A';

### slogs\_avg view:

create view if not exists slogs\_avg as select count(distinct(user\_id))   
as total,login from slogs2015 group by login order by login;

## Query 1: Total number of users on the system.

select count(distinct(user\_id)) from slogs;

## Query 2: Average number of users per day.

select round(avg(total)) from slogs\_avg;

## Query 3: Highest number of users per day.

select max(total) from slogs\_avg;

## Query 4: Top 5 users by length of time logged in.

select user\_id,sum((unix\_timestamp(logout,'MM/dd/yyyy HH:mm') - unix\_timestamp(logon,'MM/dd/yyyy HH:mm'))) as total\_loggedin  
from vdi\_serverlogs where comp\_name LIKE 'CMU%' AND year(from\_unixtime(unix\_timestamp(logon,"MM/dd/yyyy"))) = 2015 AND year(from\_unixtime(unix\_timestamp(logout,"MM/dd/yyyy"))) = 2015 AND   
user\_id != '#N/A' group by user\_id order by total\_loggedin DESC limit 5;

## Query 5: Top 5 Applications by number of times run.

SELECT apps.appname, count(apps.appname) as appcount   
FROM vdi\_apps apps JOIN vdi\_serverlogs slogs ON (apps.vdi\_id = slogs.vdi\_id)   
WHERE apps.appstart LIKE '2015%' and apps.appstop LIKE '2015%' and slogs.comp\_name LIKE 'CMU%'   
GROUP BY apps.appname ORDER BY appcount DESC LIMIT 5;

## Query 6: Usage of VDI Machines per Month by average users

select count(distinct(user\_id)), count(distinct(comp\_name)), month(login) from slogs group by month(login);

## Query 7: Average Users Vs Total logins per month

select round(avg(total)), sum(total), month(login) from slogs\_avg group by month(login);

## Query 8: Number of users logged in per day

Select date(login), count(distinct(user\_id)) from slogs group by date(login);

# Analysis and Visualization tools:

Sandbox- Ambari Hive, MS Excel, Prezi.

# Errors / Outliers/ Bad data:

* During the analysis of first query : Total number of users in the system for CMU VDI machines in 2015 year; there was a lot of bad data included for user\_id: #N/A.  
    
  With bad data the number of users are 20.  
  select count(distinct(user\_id)) from vdi\_serverlogs serverlogs where comp\_name LIKE 'CMU%' AND year(from\_unixtime(unix\_timestamp(logon,"MM/dd/yyyy"))) = 2015 AND year(from\_unixtime(unix\_timestamp(logout,"MM/dd/yyyy"))) = 2015;   
    
  After filtering the data without user\_id: #N/A, the number of users are 19.  
  select count(distinct(user\_id)) from vdi\_serverlogs serverlogs where comp\_name LIKE 'CMU%' AND year(from\_unixtime(unix\_timestamp(logon,"MM/dd/yyyy"))) = 2015 AND year(from\_unixtime(unix\_timestamp(logout,"MM/dd/yyyy"))) = 2015 AND  
  user\_id != ‘#N/A’;
* According to the vdi\_serverlogs data, the total length of time logged in by top 5 users are calculated to be in around 300 days per user in the year 2015 with VDI machines. Practically, the total number of days VDI machines used are 223. This result has probably achieved by multiple logins of a user in multiple machines. This way a user records as the sum of time logged in all machines in total. This is outlying data analyzed from the server logs.

# Limitations/ Drawbacks:

* With huge chunks of data, the data analysis varies because of appropriateness in data is recorded but not accurately. The data may have been collected with respect to specific research question or objective.
* Data quality issues are not guaranteed with these kind of system logs. For example, the user id error, difference in time logged error for some users may be the discrepancies.

# References:

* Confluence Wiki- Getting started with Apache Hive  
  <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Cli>
* Hadoop Online Tutorial  
  <http://hadooptutorial.info/category/hive/>
* W3Schools Tutorial  
  <http://www.w3schools.com/sql/>