Sanket Kshirsagar
A20399862

CSP 554 - Project

Visualisation of Website's Visitors' Data Using Apache Hadoop Tools

Literature Review

Overview:

Having a website in today's ever-evolving online world is a must- especially for small businesses selling products and services. If you plan on having a lot of customers, you need an online presence to give your clients information at the click of a button. With evolution of tools like Google Analytics, Adobe Web Analytics, etc. website traffic can be monitored via various visual methods.

Need of web log analysis:

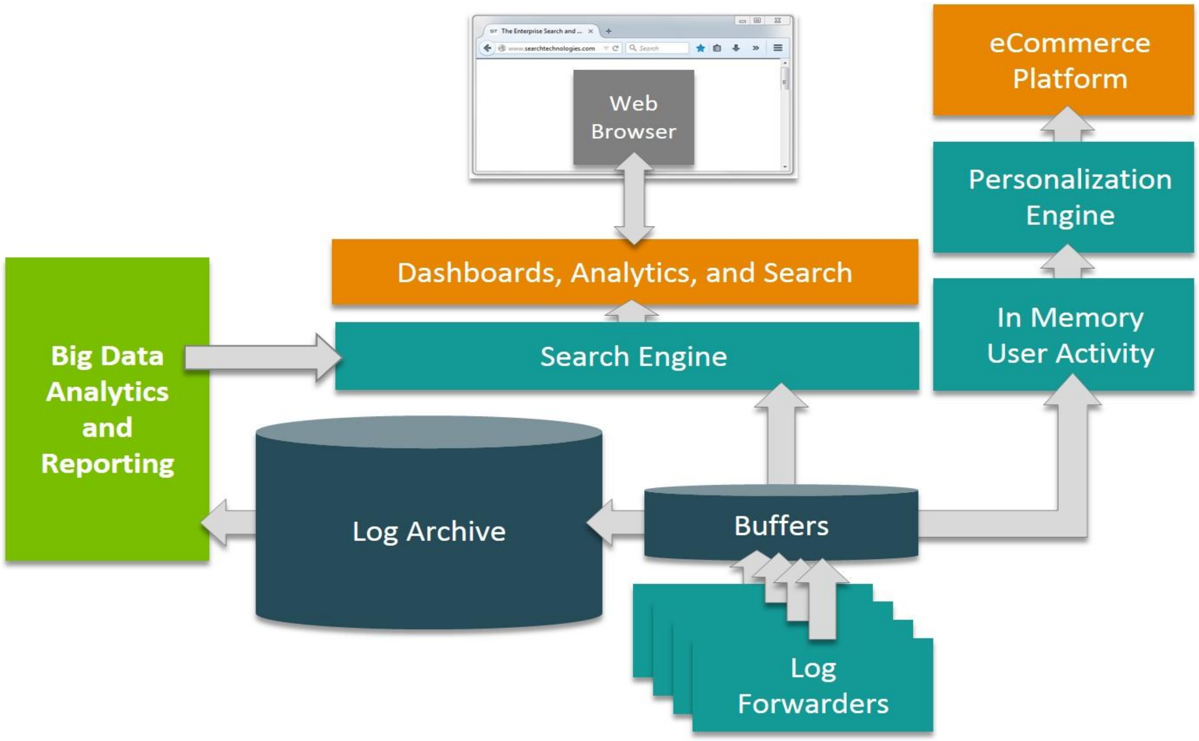
For any company, specially company which has their own eCommerce website(s), for selling products, analysis of web logs is one the most important source of information about online business patterns and user preferences on their website. If a company isn't setting aside time each week or month to look through their log data, they have no idea how well their website is performing as a marketing tool, no idea how well they're doing on the search engines, no idea how well their AdWords are performing, and no idea of how easily they might be able to improve their sales. [1].

Traditionally, companies used to rely on web analysis provided by host providers or traditional web log analysis software like Loggly, GoAccess, etc.

Need of using Big Data Tools for web log analysis:

With data amounting to terabytes, even petabytes, it's virtually impossible for traditional log analysis software to quickly and accurately discern patterns and pinpoint trends. Without an efficient and automated process to make sense of this data, organizations would face the danger of dumping valuable data in an unrefined "data lake," and eventually lose the ability to discover data-driven competitive advantages [2]. Using Big Data tools such a Hadoop we can extract useful data such as clickstream data, which will mainly help us realise the website visitor's clickstream patterns (i.e. if website have pages A-Z and want to see how many people land on Page G and then go to Page B). Segmenting, and analyzing this data from web logs using Big Data tools will give you a more refined look at your customer's behavior patterns - from the time they land on your website till the time they either buy your product or leave without buying.

Big Data Architecture for log analytics [3]



Big data architecture for log analytics

Tools and technologies used:

- **Adobe Omniture:** This is web analytics tool which is used to create web logs for monitoring clickstream data of user. Logs files generated by this tool usually has fields for website visitor data such as IP address, timestamps, date, browser info, location info, clicked links, etc. We're going to use this important data generated from logs to analyse data (such as "popular shopping category according to geolocation", "average age of user in each shopping category") in this project.

- **Hortonworks VM:** It is an open source framework for distributed storage and processing of large, multi-source data set. It has Apache Hadoop tools in built in it including HDFS, Hive and Ambari, which are key 3 tools we're going to use in this project.

- **Apache Ambari:** This tool is used to make Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters. We'll be using this tool for creating director, copying web log files in it and using Hive View (1.0) to execute HiveQL queries.

- **Apache Hadoop HDFS:** Hadoop HDFS is file system to store data in Hadoop environment. We will be storing log files on HDFS in this project.

- **Apache Hadoop Hive:** This technology is used for reading, writing, and managing large datasets residing in distributed storage using SQL. We'll be throwing HiveQL queries to join tables and extract useful information.

- **ODBC:** An ODBC driver uses the Open Database Connectivity (ODBC) interface by Microsoft. We'll be using this driver to connect Database created in Hadoop environment to Microsoft Excel.

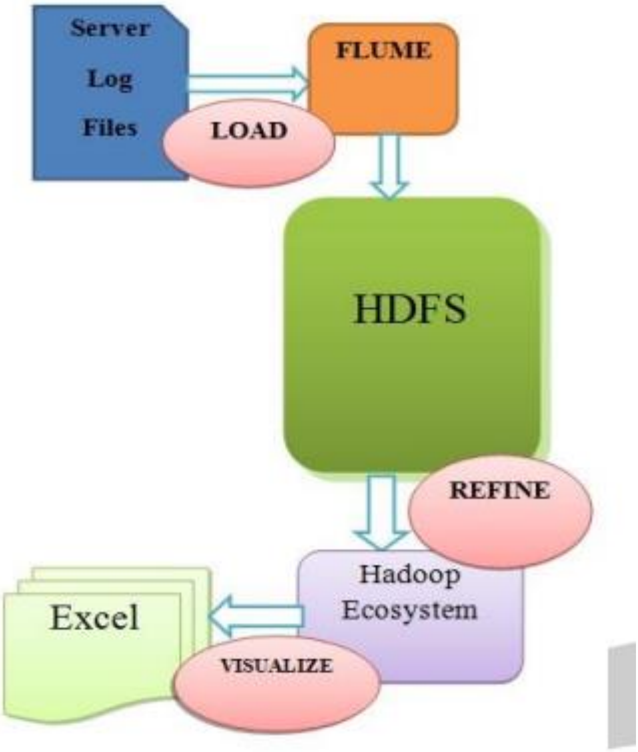
- **Microsoft Excel 2016:** Microsoft Excel is a powerful tool to visualise the data in tabular, charts, graphs and maps format. It has add-ins such as pivot charts and power view to analyse data by applying different functions on fields in table(s). Due to its ability to import data from external sources (such as ODBC data sources, Microsoft Azure) it makes Microsoft Excel a very useful tool in this project.

Project Description

Goal:

To extract and visualise the useful data, through tables, charts, maps and graphs, about the website visitors extracted from huge web logs using Apache Hadoop tools.

Basic Overview of Web Server Log Processing Using Hadoop Architecture [4]



Note: Flume is a framework for populating Hadoop with data. Agents are populated throughout one's IT infrastructure – inside web servers, application servers and mobile devices, for example – to collect data and integrate it into Hadoop. It is not directly used in the project. Here in this project Flume is not directly used because it needs a special web server to extract data and generate web logs.

Step By Step Execution:

1. **Setting up Hortonworks HDP 2.6.5:**
   Hortonworks HDP is installed on VMware Workstation locally. 8 GB of memory and 120 GB of HDD space is assigned on for this OS to run virtually.

2. **Web logs selection:**
   In this 2nd step I have selected appropriate logs and modified it according to the project requirements. Because of limitation of accessing actual web logs generated by Adobe Omniture and other hosting providers, I have downloaded sample Adobe Omniture web logs and other sample web logs directly from online. All logs are in the .tsv format. We have 3 types of web logs in this project:

   - ***Adobe Omniture:*** This sample Adobe Omniture has 178 different fields of user data such as ip address, url, timestamps, state, country, browser info, etc. It has around 115,000 number of records in it. This is the biggest web log file we're using. File name used is 0.tsv

   - ***Registered user data:*** This log has 3 fields "registered user's ID", "birth date" and "gender code". This log has around 38500 records. File name is regusers.tsv.

   - ***Url map:*** In this web log file categories are mapped with corresponding url. It has only 2 fields category name and url of the category. File name is urlmap.tsv

   (See appendix for more details about the web log files link)

3. **Uploading web logs using Ambari:**
   In this step, we'll be using Apache Ambari to upload these 3 types of web log files to HDFS via Files View to make things simpler and quick. Each web log is copied in it's own folder. Location of web logs is /tmp/CSP554/. This is here each separate folder is created for each log file type.

4. **Data selection using Hive:**
   This is the most important step in this project. We will create database named mydb first. Then using LOAD parameter in HiveQL queries, data from each web log is taken and an EXTERNAL table is created for each type of web log. So, we'll have 3 tables named omniturelogs, users and products. Now after this, a view named "omniture" is created from omniturelogs table for eliminating non-usable fields and keeping required fields for this project. These 7 columns we'll be having in view are ip, timestamps, url, city, country, state, swid.

   Now in the next step, we will join this view with users and products table to create new 9 column table called "**webanalytics**". This is the final table we'll be using for analysing data on Excel. Following are the columns contained by **webanalytics**:
   Timestamp, url, ip, city, state, country, category, age and gender.

   (See appendix for more details about the HiveQL queries)

5. **Setting up ODBC to connect to Hive database:**
   Now we will setup ODBC connection to connect to Hive database and tables in it. Using ODBC Data sources configuration in Windows, we will setup the connection by using following parameters.

(See appendix for more details about the configuration)

6. **Loading data in Microsoft Excel:**
   In this step, we will load external data source which we created in ODBC named “Hortonworks Hive Driver” to connect to our database through the option of inserting data through external resources in Microsoft Excel.

   (See appendix for more details)

7. **Throwing queries using Pivot Charts and Pivot Table in Excel:**
   In this final step, we select different parameters from the tables such as categories, url, age, etc and will visualise results using bar graphs, pie graphs and maps.

Expected results:

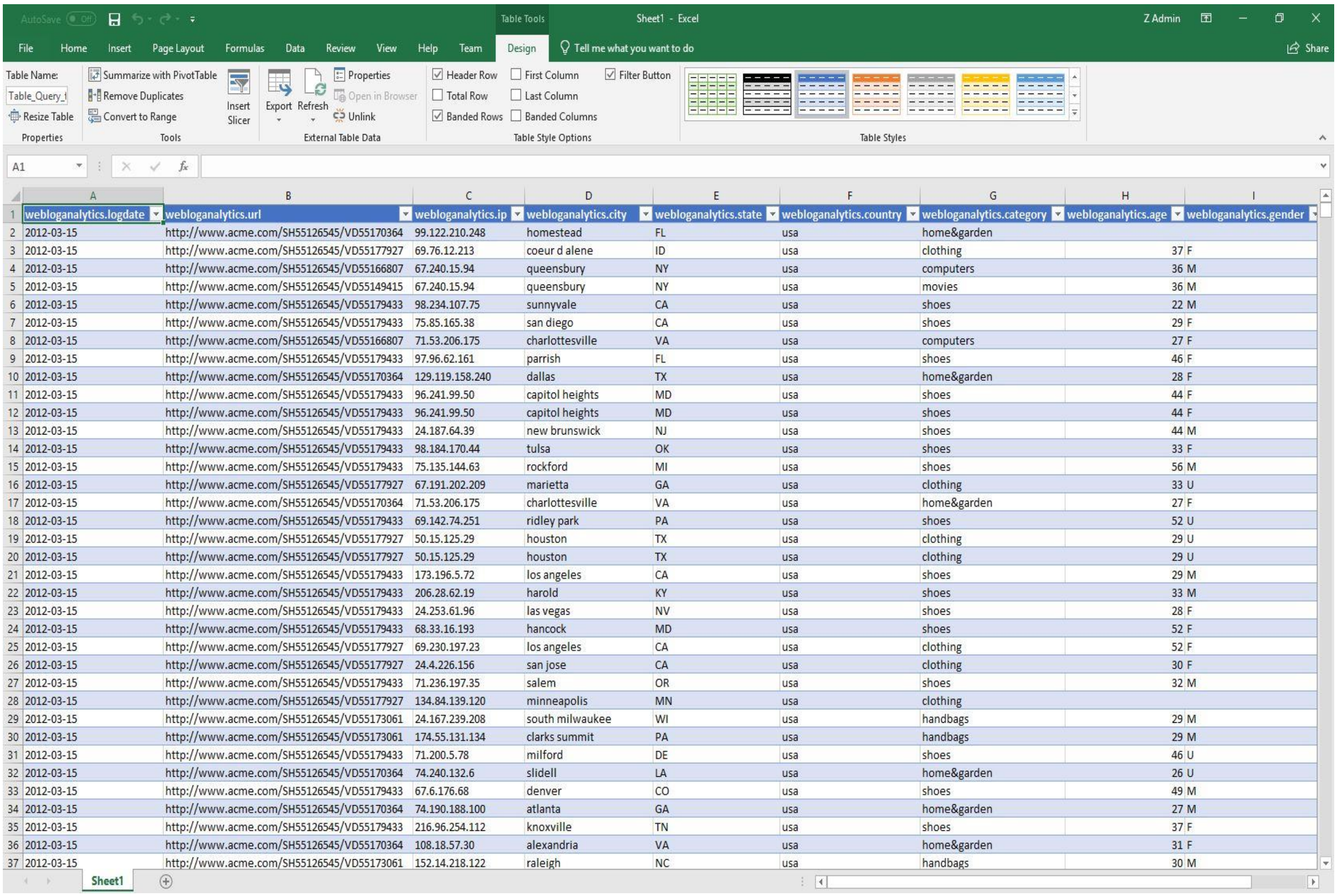
By selecting different parameters, we will be answering following questions about website:

- What is the average age for each category on the website?
- Which is the most popular category in every state/country?
- What is the average age and most popular category according to gender?
- What is the average visits of particular url?

RESULTS AND ANALYSIS

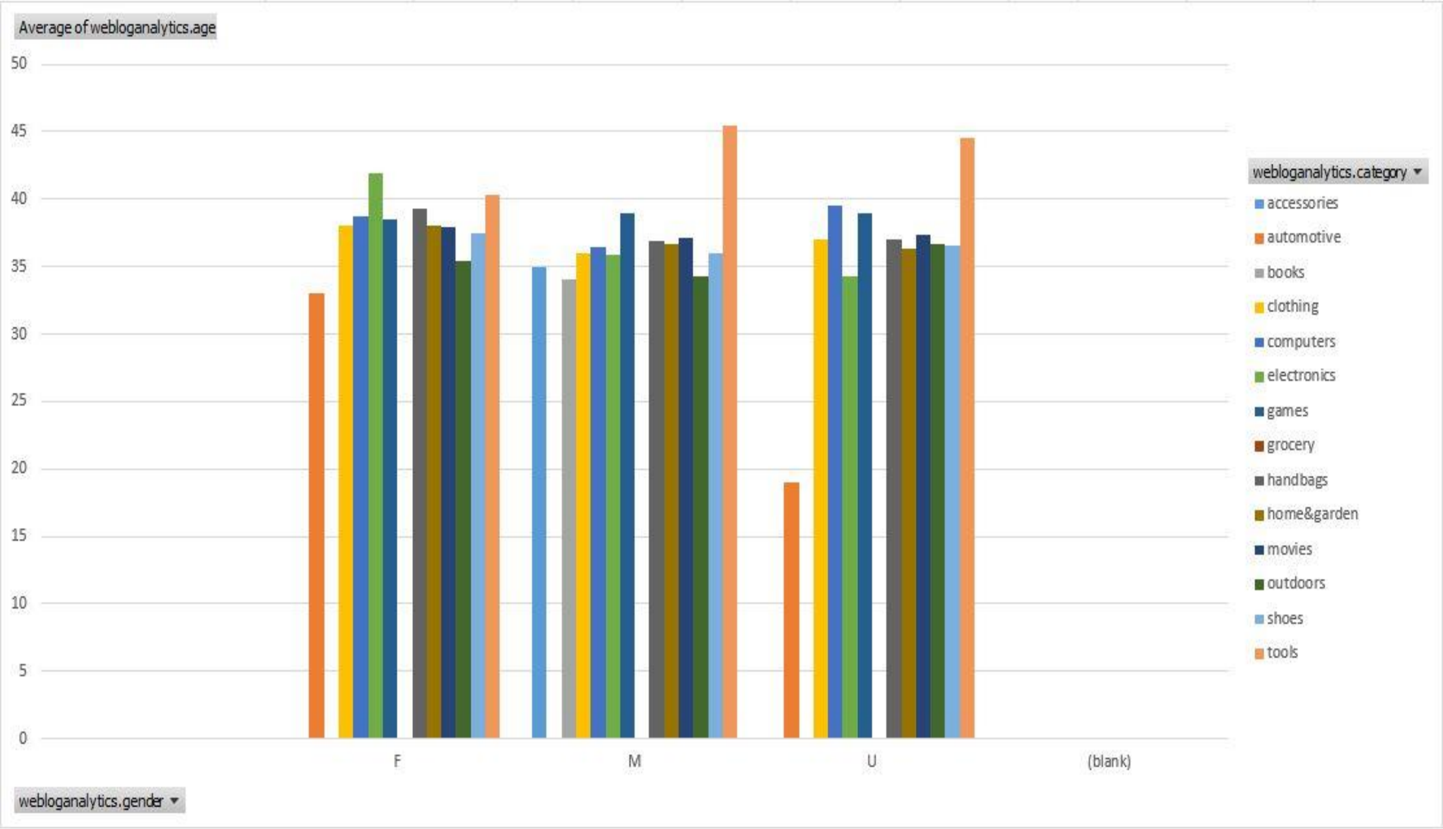
Following results are generated from the demo we performed

Table containing 9 fields Timestamp, url, ip, city, state, country, category, age and gender with around 115,000 records

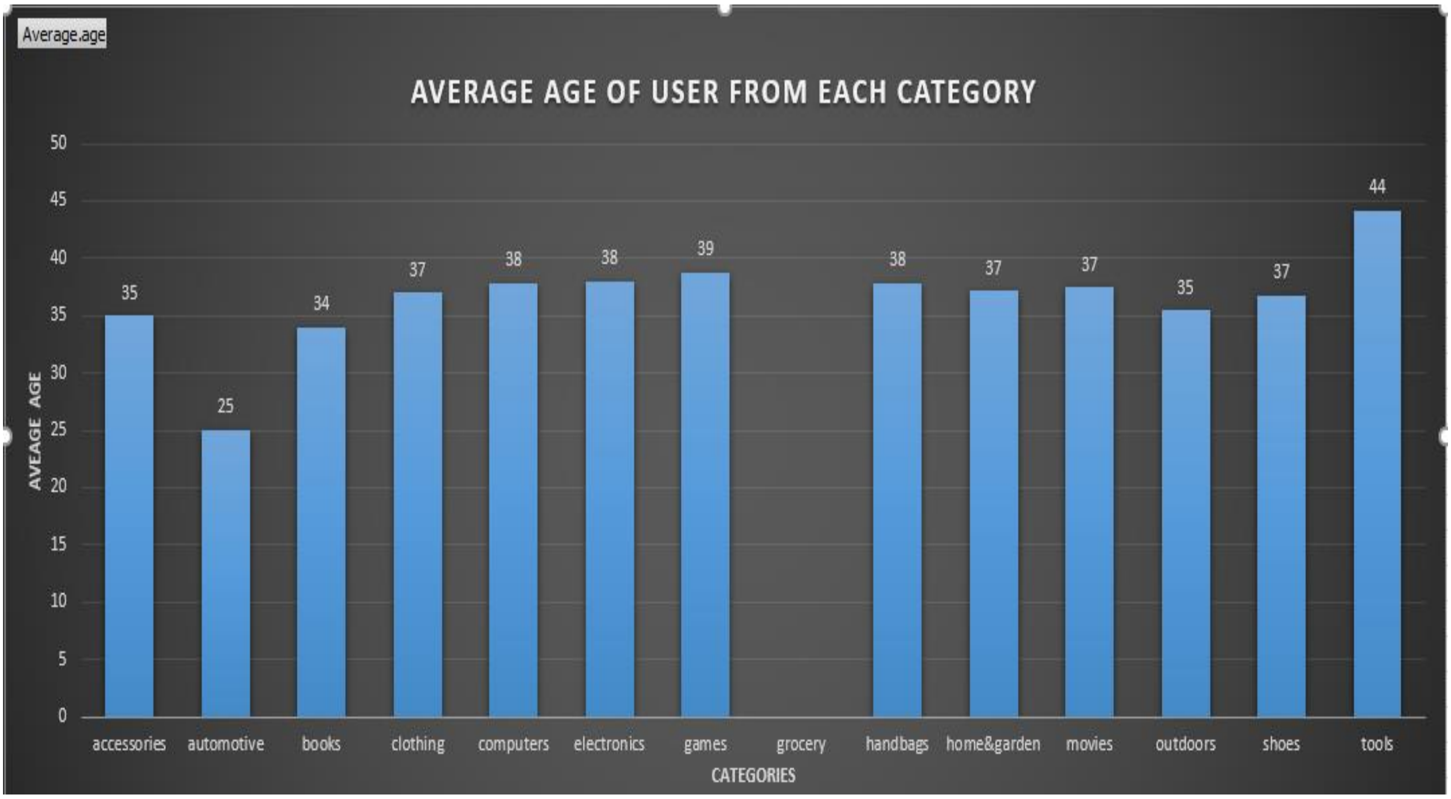


1	webloganalytics.logdate	webloganalytics.url	webloganalytics.ip	webloganalytics.city	webloganalytics.state	webloganalytics.country	webloganalytics.category	webloganalytics.age	webloganalytics.gender
2	2012-03-15	http://www.acme.com/SH55126545/VD55170364	99.122.210.248	homestead	FL	usa	home&garden		
3	2012-03-15	http://www.acme.com/SH55126545/VD55177927	69.76.12.213	coeur d alene	ID	usa	clothing	37	F
4	2012-03-15	http://www.acme.com/SH55126545/VD55166807	67.240.15.94	queensbury	NY	usa	computers	36	M
5	2012-03-15	http://www.acme.com/SH55126545/VD55149415	67.240.15.94	queensbury	NY	usa	movies	36	M
6	2012-03-15	http://www.acme.com/SH55126545/VD55179433	98.234.107.75	sunnyvale	CA	usa	shoes	22	M
7	2012-03-15	http://www.acme.com/SH55126545/VD55179433	75.85.165.38	san diego	CA	usa	shoes	29	F
8	2012-03-15	http://www.acme.com/SH55126545/VD55166807	71.53.206.175	charlottesville	VA	usa	computers	27	F
9	2012-03-15	http://www.acme.com/SH55126545/VD55179433	97.96.62.161	parrish	FL	usa	shoes	46	F
10	2012-03-15	http://www.acme.com/SH55126545/VD55170364	129.119.158.240	dallas	TX	usa	home&garden	28	F
11	2012-03-15	http://www.acme.com/SH55126545/VD55179433	96.241.99.50	capitol heights	MD	usa	shoes	44	F
12	2012-03-15	http://www.acme.com/SH55126545/VD55179433	96.241.99.50	capitol heights	MD	usa	shoes	44	F
13	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.187.64.39	new brunswick	NJ	usa	shoes	44	M
14	2012-03-15	http://www.acme.com/SH55126545/VD55179433	98.184.170.44	tulsa	OK	usa	shoes	33	F
15	2012-03-15	http://www.acme.com/SH55126545/VD55179433	75.135.144.63	rockford	MI	usa	shoes	56	M
16	2012-03-15	http://www.acme.com/SH55126545/VD55177927	67.191.202.209	marietta	GA	usa	clothing	33	U
17	2012-03-15	http://www.acme.com/SH55126545/VD55170364	71.53.206.175	charlottesville	VA	usa	home&garden	27	F
18	2012-03-15	http://www.acme.com/SH55126545/VD55179433	69.142.74.251	ridley park	PA	usa	shoes	52	U
19	2012-03-15	http://www.acme.com/SH55126545/VD55177927	50.15.125.29	houston	TX	usa	clothing	29	U
20	2012-03-15	http://www.acme.com/SH55126545/VD55177927	50.15.125.29	houston	TX	usa	clothing	29	U
21	2012-03-15	http://www.acme.com/SH55126545/VD55179433	173.196.5.72	los angeles	CA	usa	shoes	29	M
22	2012-03-15	http://www.acme.com/SH55126545/VD55179433	206.28.62.19	harold	KY	usa	shoes	33	M
23	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.253.61.96	las vegas	NV	usa	shoes	28	F
24	2012-03-15	http://www.acme.com/SH55126545/VD55179433	68.33.16.193	hancock	MD	usa	shoes	52	F
25	2012-03-15	http://www.acme.com/SH55126545/VD55177927	69.230.197.23	los angeles	CA	usa	clothing	52	F
26	2012-03-15	http://www.acme.com/SH55126545/VD55177927	24.4.226.156	san jose	CA	usa	clothing	30	F
27	2012-03-15	http://www.acme.com/SH55126545/VD55179433	71.236.197.35	salem	OR	usa	shoes	32	M
28	2012-03-15	http://www.acme.com/SH55126545/VD55177927	134.84.139.120	minneapolis	MN	usa	clothing		
29	2012-03-15	http://www.acme.com/SH55126545/VD55173061	24.167.239.208	south milwaukee	WI	usa	handbags	29	M
30	2012-03-15	http://www.acme.com/SH55126545/VD55173061	174.55.131.134	clarks summit	PA	usa	handbags	29	M
31	2012-03-15	http://www.acme.com/SH55126545/VD55179433	71.200.5.78	milford	DE	usa	shoes	46	U
32	2012-03-15	http://www.acme.com/SH55126545/VD55170364	74.240.132.6	slidell	LA	usa	home&garden	26	U
33	2012-03-15	http://www.acme.com/SH55126545/VD55179433	67.6.176.68	denver	CO	usa	shoes	49	M
34	2012-03-15	http://www.acme.com/SH55126545/VD55170364	74.190.188.100	atlanta	GA	usa	home&garden	27	M
35	2012-03-15	http://www.acme.com/SH55126545/VD55179433	216.96.254.112	knoxville	TN	usa	shoes	37	F
36	2012-03-15	http://www.acme.com/SH55126545/VD55170364	108.18.57.30	alexandria	VA	usa	home&garden	31	F
37	2012-03-15	http://www.acme.com/SH55126545/VD55173061	152.14.218.122	raleigh	NC	usa	handbags	30	M

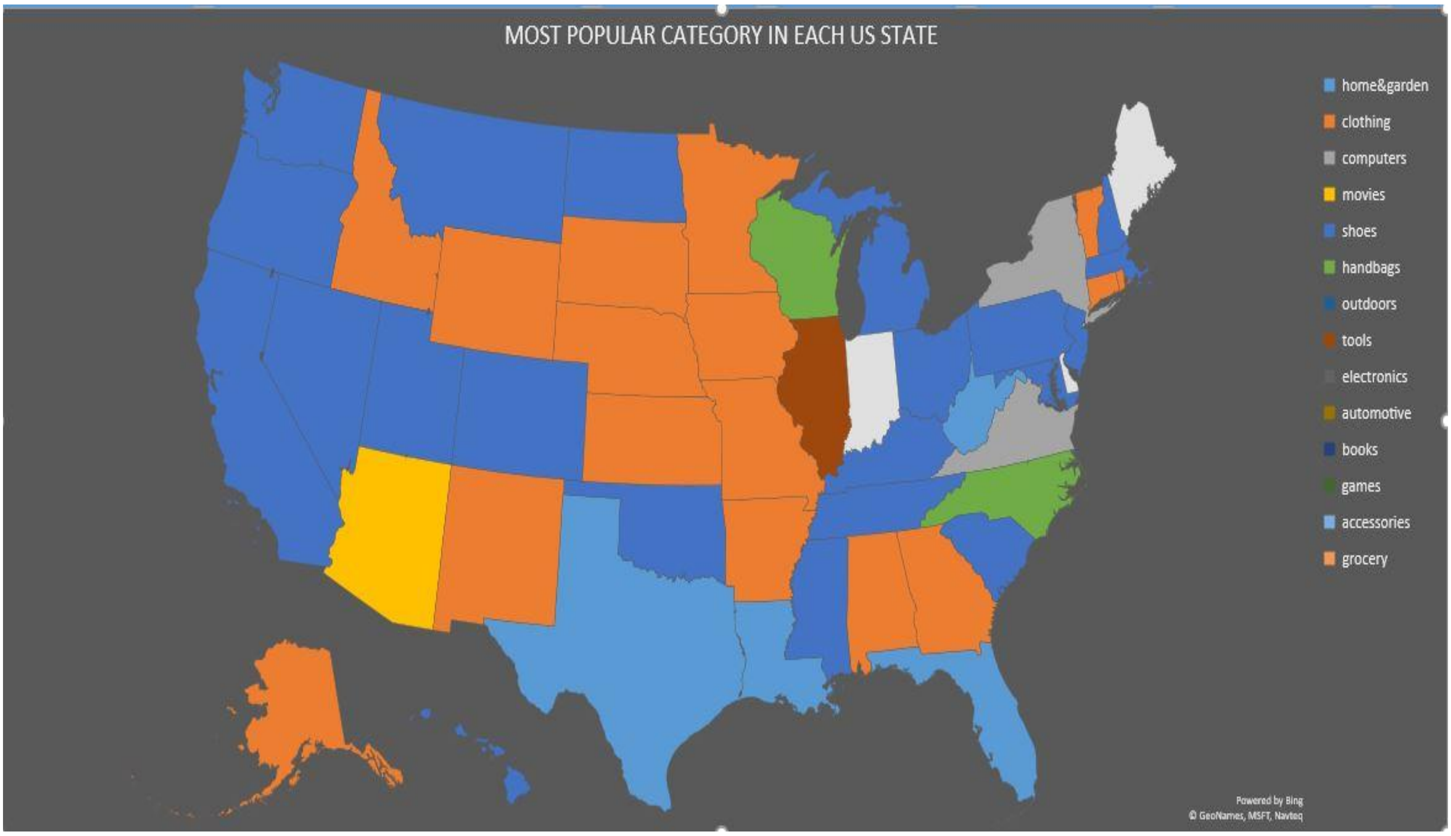
Clustered graph for most popular category and average age for each gender



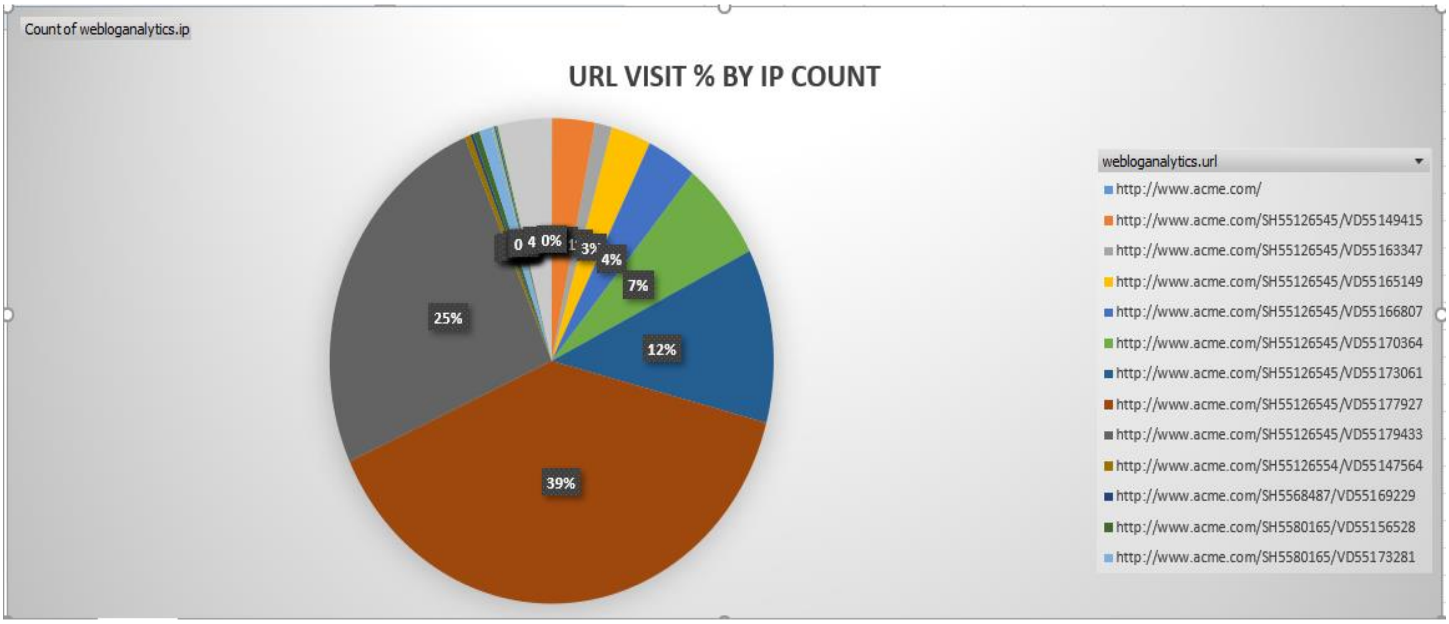
Bar graph showing average age of the user from each category on the website



Map showing most popular category in each US state



Pie chart showing URL visit percentage by IP count



Future Scope

In this project we have selected 9 fields for visualising the data. Adobe Omniture logs more than 100 fields. We can use other fields like language, browser info, flags, etc. to show results like most/least popular browser, most/least preferred language.

Also, by throwing different geolocation queries, selecting and filtering different parameters like age, gender, browser, ip, url, user id, etc we can find out more about the web visitor’s data.

So by adding new fields, throwing different queries and filtering different parameters which we find useful, we can visualise website visitor’s data and pattern from many different perspective.

Conclusion

By analyzation through the visualization of web visitors’ data generated via web logs of eCommerce website, companies study online market trends and website visitors’ pattern. This will help them to improve the overall online business by focusing on providing visitors with high quality website experience.

References

- **The Importance of Log Analysis** ***Written by Dave Collins, SoftwarePromotions Ltd.*** [1]

  https://www.davetalks.com/articles/importance-of-log-analysis/

- **An Open Source Approach to Log Analytics with Big Data In the Trenches with Big Data & Search** [2][3]

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

- **Web Server Log Processing using Hadoop** - International Journal for Research in Engineering Application & Management (IJREAM) Vol-01, Issue 10, Jan 2016. [4]

  https://www.ijream.org/papers/INJRV01I10001.pdf

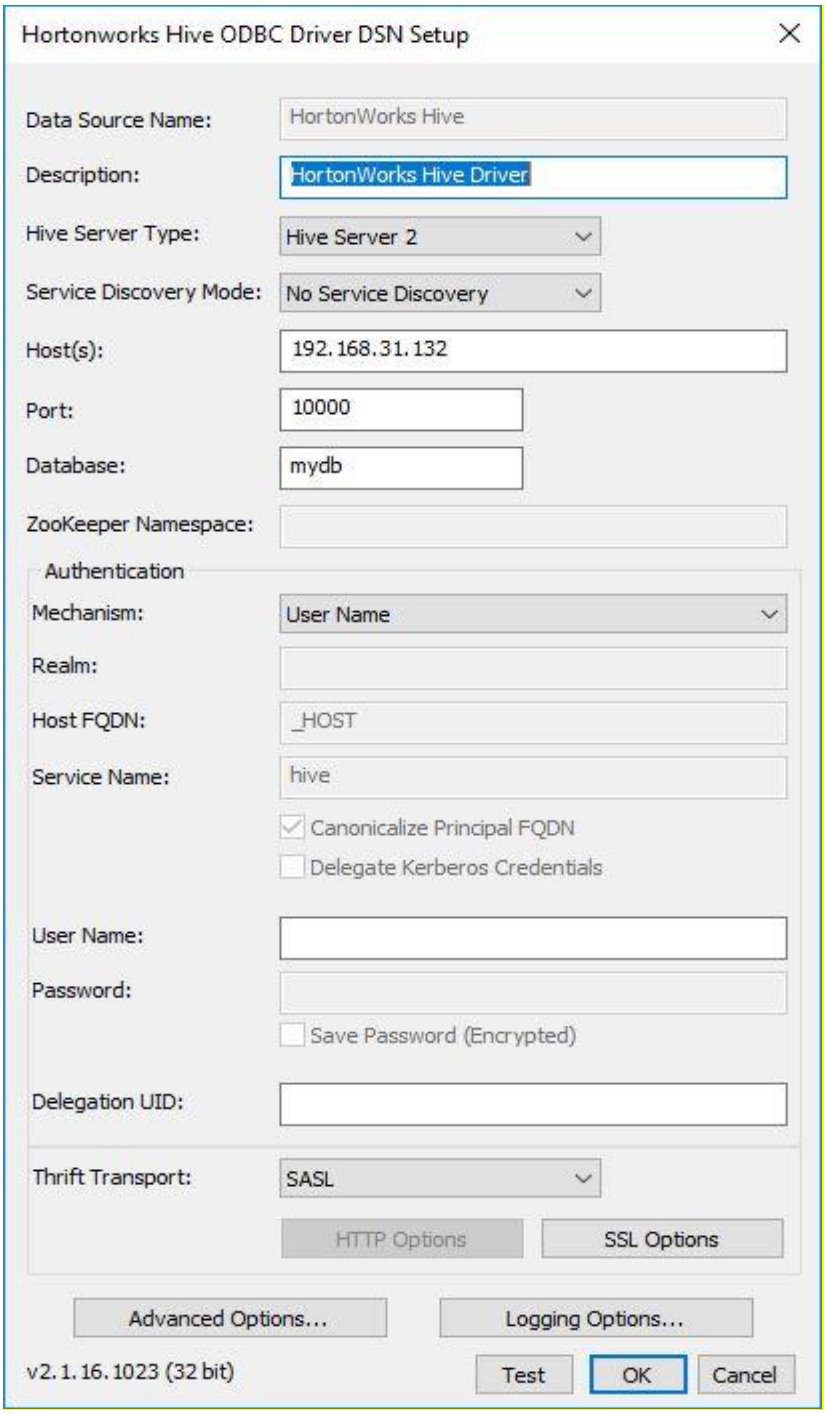
APPENDIX

Log Files:

HiveQL Queries:

```
CREATE DATABASE MyDb;

USE MyDb;


CREATE EXTERNAL TABLE users

(swid string, birth_dt string, gender_cd string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE

LOCATION "/tmp/CSP554/users"

tblproperties ("skip.header.line.count"="1");


CREATE EXTERNAL TABLE products

(url string, category string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE

LOCATION "/tmp/CSP554/products"

tblproperties ("skip.header.line.count"="1");


CREATE EXTERNAL TABLE omniturelogs

(

col_1 string, col_2 string, col_3 string, col_4 string, col_5 string, col_6 string, col_7 string, col_8 string, col_9 string, col_10
string, col_11 string, col_12 string, col_13 string, col_14 string, col_15 string, col_16 string, col_17 string, col_18 string, col_19
string, col_20 string, col_21 string, col_22 string, col_23 string, col_24 string, col_25 string, col_26 string, col_27 string, col_28
string, col_29 string, col_30 string, col_31 string, col_32 string, col_33 string, col_34 string, col_35 string, col_36 string, col_37
string, col_38 string, col_39 string, col_40 string, col_41 string, col_42 string, col_43 string, col_44 string, col_45 string, col_46
string, col_47 string, col_48 string, col_49 string, col_50 string, col_51 string, col_52 string, col_53 string, col_54 string, col_55
string, col_56 string, col_57 string, col_58 string, col_59 string, col_60 string, col_61 string, col_62 string, col_63 string, col_64
string, col_65 string, col_66 string, col_67 string, col_68 string, col_69 string, col_70 string, col_71 string, col_72 string, col_73
string, col_74 string, col_75 string, col_76 string, col_77 string, col_78 string, col_79 string, col_80 string, col_81 string, col_82
string, col_83 string, col_84 string, col_85 string, col_86 string, col_87 string, col_88 string, col_89 string, col_90 string, col_91
string, col_92 string, col_93 string, col_94 string, col_95 string, col_96 string, col_97 string, col_98 string, col_99 string,
col_100 string, col_101 string, col_102 string, col_103 string, col_104 string, col_105 string, col_106 string, col_107 string,
col_108 string, col_109 string, col_110 string, col_111 string, col_112 string, col_113 string, col_114 string, col_115 string,
col_116 string, col_117 string, col_118 string, col_119 string, col_120 string, col_121 string, col_122 string, col_123 string,
col_124 string, col_125 string, col_126 string, col_127 string, col_128 string, col_129 string, col_130 string, col_131 string,
col_132 string, col_133 string, col_134 string, col_135 string, col_136 string, col_137 string, col_138 string, col_139 string,
col_140 string, col_141 string, col_142 string, col_143 string, col_144 string, col_145 string, col_146 string, col_147 string,
col_148 string, col_149 string, col_150 string, col_151 string, col_152 string, col_153 string, col_154 string, col_155 string,
col_156 string, col_157 string, col_158 string, col_159 string, col_160 string, col_161 string, col_162 string, col_163 string,
col_164 string, col_165 string, col_166 string, col_167 string, col_168 string, col_169 string, col_170 string, col_171 string,
col_172 string, col_173 string, col_174 string, col_175 string, col_176 string, col_177 string, col_178 string
)
-- PARTITIONED BY (id string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE
LOCATION "/tmp/CSP554/omniturelogs";

CREATE VIEW omniture AS
SELECT
col_2 ts,
col_8 ip,
col_13 url,
col_14 swid,
col_50 city,
col_51 country,
col_53 `state`
from omniturelogs;

create table webloganalytics as
select
    to_date(o.ts) logdate,
    o.url,
    o.ip,
    o.city,
    upper(o.`state`) `state`,
    o.country,
    p.category,
    CAST(datediff(
    from_unixtime( unix_timestamp() ),
        from_unixtime( unix_timestamp(d.birth_dt, 'dd-MMM-yy'))) / 365  AS INT) age,
    d.gender_cd gender
from
    omniture o
    left outer join products p on o.url = p.url
    left outer join users d on o.swid = concat('{', d.swid , '}');
```

ODBC Configuration:



Hortonworks Hive ODBC Driver DSN Setup

Data Source Name: HortonWorks Hive

Description: HortonWorks Hive Driver

Hive Server Type: Hive Server 2

Service Discovery Mode: No Service Discovery

Host(s): 192.168.31.132

Port: 10000

Database: mydb

ZooKeeper Namespace:

Authentication

Mechanism: User Name

Realm:

Host FQDN: _HOST

Service Name: hive

☑ Canonicalize Principal FQDN

□ Delegate Kerberos Credentials

User Name:

Password:

□ Save Password (Encrypted)

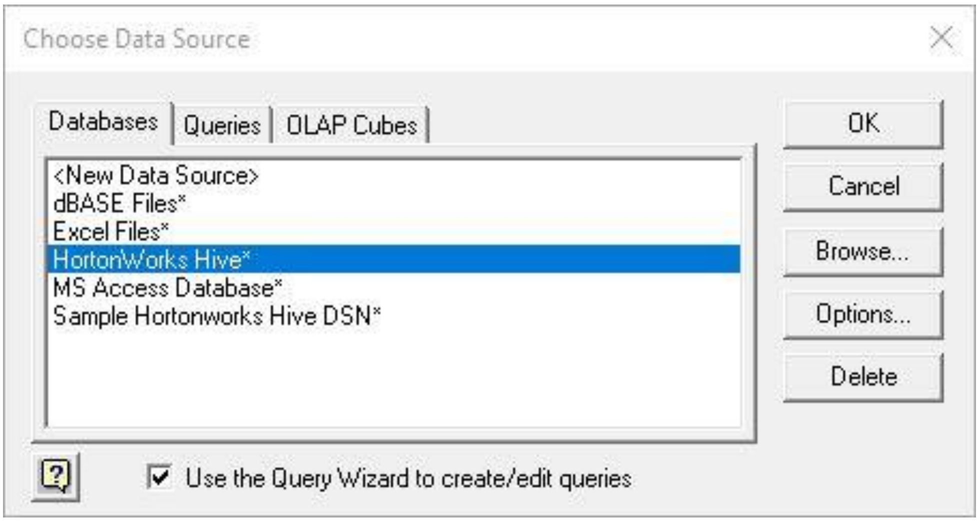
Delegation UID:

Thrift Transport: SASL

v2.1.16.1023 (32 bit)

Note: Host IP is IP of the Hortonworks VM. Database name is the name of the database you want to connect.

Microsoft Excel data source selection



<New Data Source>
dBASE Files*
Excel Files*
HortonWorks Hive*
MS Access Database*
Sample Hortonworks Hive DSN*

☑ Use the Query Wizard to create/edit queries