Final Project

Clickpath Analytics

Satish Vittalam

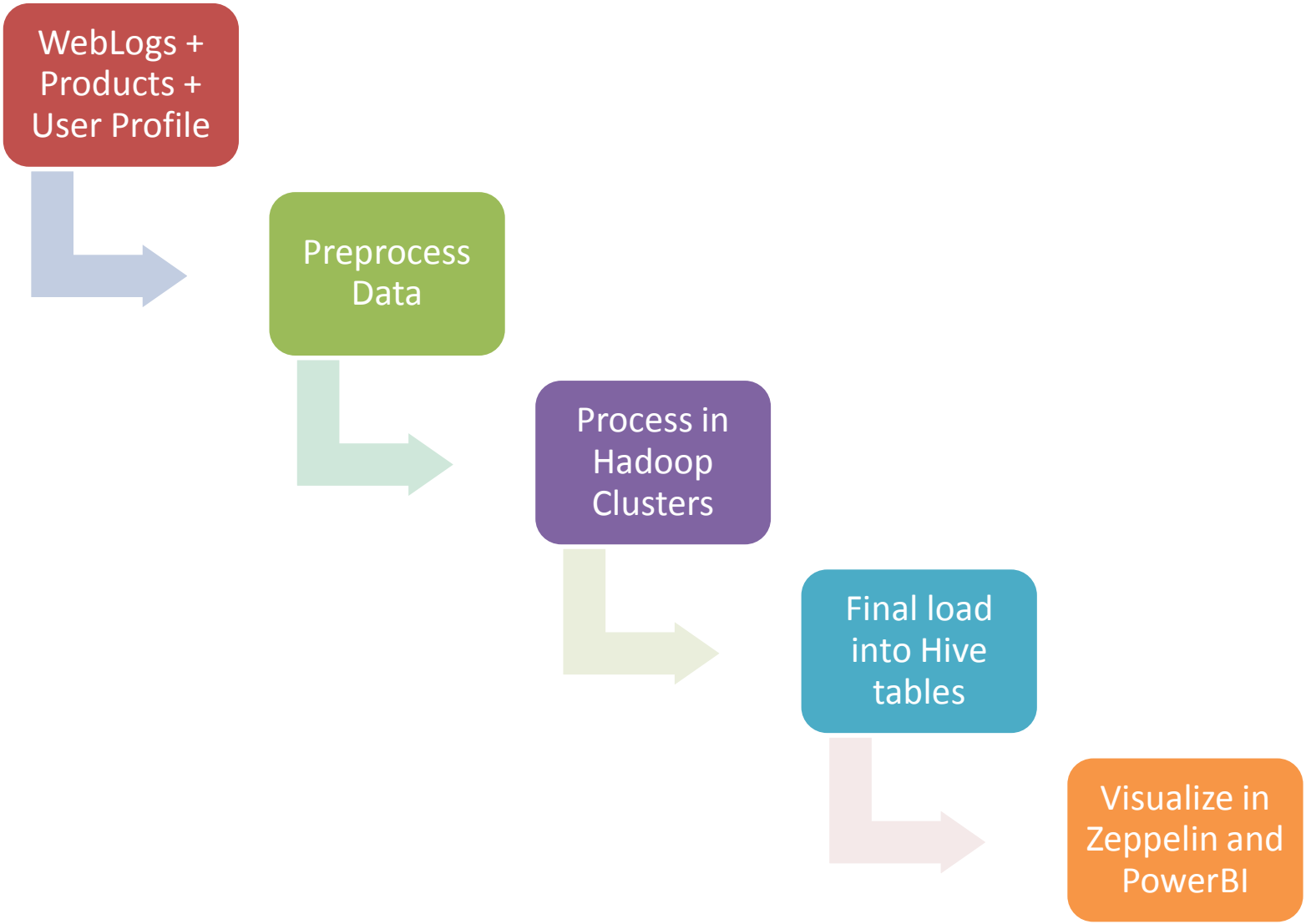
Problem Statement

- Ecommerce retailers and other companies who have an online presence are trying gather more details about the customers browsing or online shopping patterns, the products they buy, the products they may be interested in the future and also provide a better shopping experience. They perform Basket Analysis, Path optimization and even try to analyze the next product to buy. To achieve this, companies have to process massive amounts of data sets in terms of web server logs which is also referred to as Clickpath or Clickstream data. This information is captured by Webserver as customer navigates around the website.

Technology Overview

Azure HDInsight offers a cost-effective way to process massive amounts of data. Hadoops framework and its ecosystem helps to analyze this information easier, get better insights about the customer and help improve the effectiveness of the shopping. We use the tools and technologies provided to process large datasets of log files, get the required information from the logs and combine them with user profile and products data (these could be available from the OLTP application) to perform the required analytics. Hadoop offers multiple analytics tools for these big datasets. We will load, refine and visualize the log data.

Approach

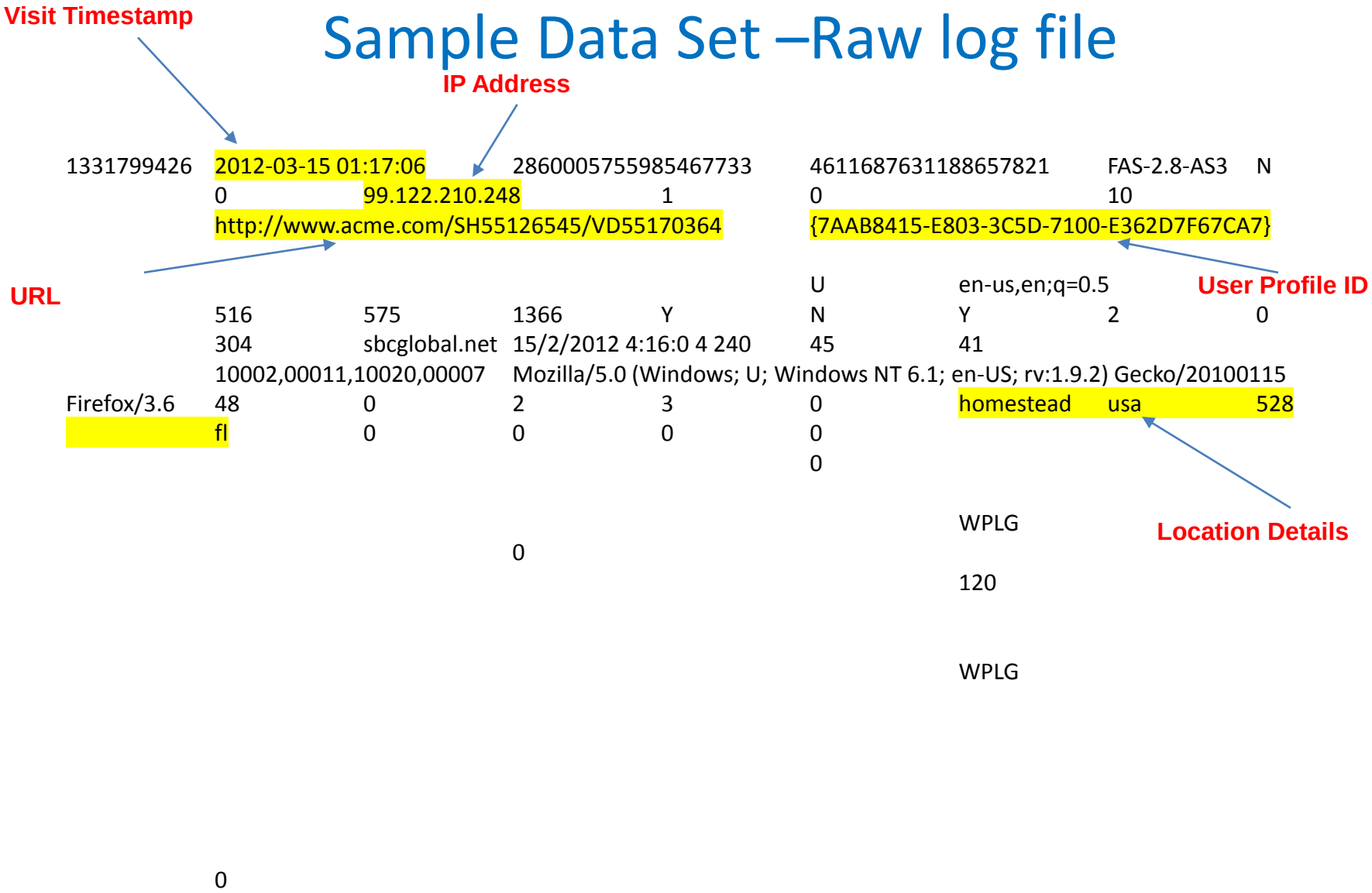

WebLogs + Products + User Profile
Preprocess Data
Process in Hadoop Clusters
Final load into Hive tables
Visualize in Zeppelin and PowerBI

Data Set

- **Data set obtained at** : https://s3.amazonaws.com/hw-sandbox/tutorial8/RefineDemoData.zip

- The following datasets are primarily used for this project:
  - Webserver log data
  - Product information data
  - User Profile data

Sample Data Set –Raw log file



Ref: https://hortonworks.com/blog/how-to-capitalize-on-clickstream-data-with-hadoop/

Sample Data Set –Raw log file

Sample Data for Products:

url	category
http://www.acme.com/	books
http://www.acme.com/SH55126545/VD55149415	movies
http://www.acme.com/SH55126545/VD55163347	games
http://www.acme.com/SH55126545/VD55165149	electronics

Sample data for Users:

SWID	BIRTH_DT	GENDER_CD
0001BDD9-EABF-4D0D-81BD-D9EABFCD0D7D	8-Apr-84	F
00071AA7-86D2-4EB9-871A-A786D27EB9BA	7-Feb-88	F
00071B7D-31AF-4D85-871B-7D31AFFD852E	22-Oct-64	F
0007967E-F188-4598-9C7C-E64390482CFB	1-Jun-66	M

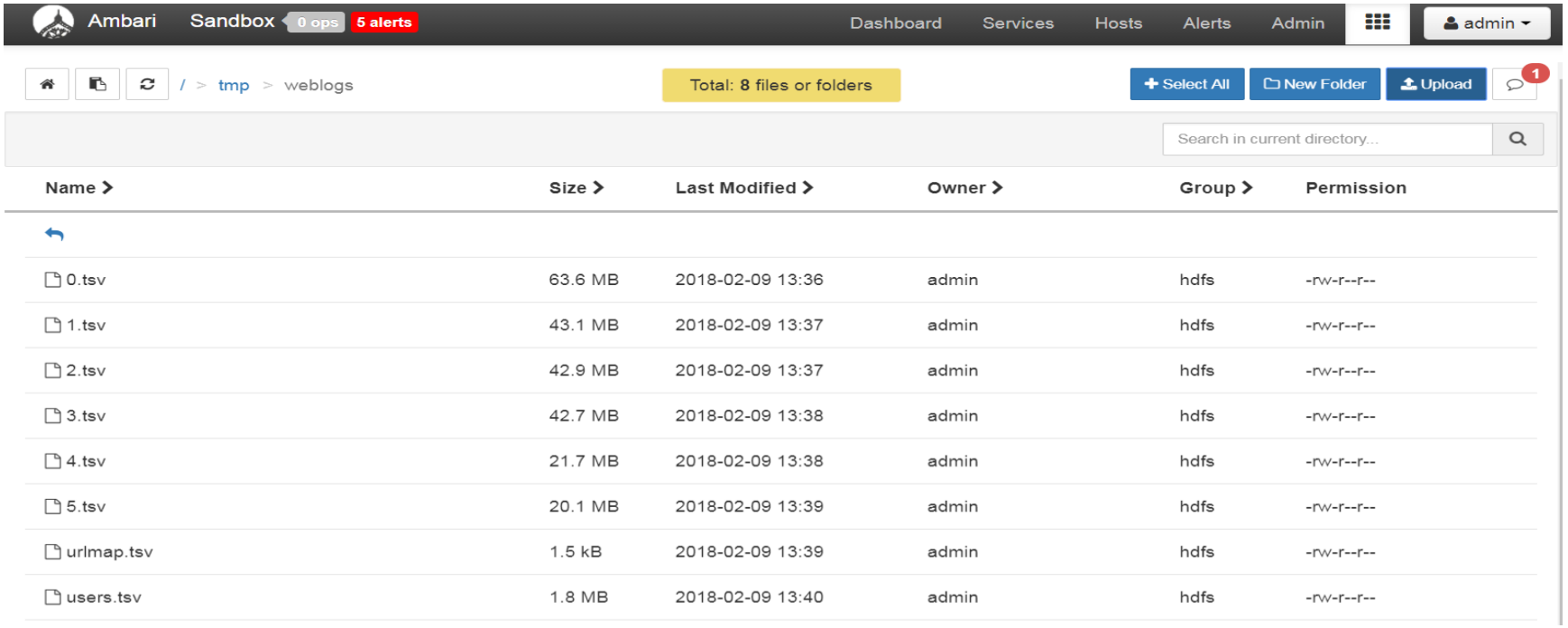
Process Overview

- Created a HortonWorks Cluster available on Azure and configure it for SSH
- Pre-process the Web logs sample data that is obtained
- Create a SQL database in Azure
- Move the sample data into Hadoop File system
- Use Pig latin script to combine the web server logs into one.
- Use Sqoop to move the data from traditional RDBMS (OLTP system) to Hive
- Create custom Hive tables that will store the final data that is created for visualization.
- Use Zepplin and Power BI for Analytics and visualization

Benefits

- These are the potential benefits of Clickpath Analysis:
  - What is the most efficient path for a site visitor to research a product, and then buy it?
  - What products do visitors tend to buy together, and what are they most likely to buy in the future?
  - Where should I spend resources on fixing or enhancing the user experience on my website?

Input Files uploaded




```
[root@sandbox tmp]# hadoop fs -ls /tmp/weblogs
Found 8 items
-rwxrwxrwx   3 admin hdfs   66685542 2018-02-09 19:36 /tmp/weblogs/0.tsv
-rwxrwxrwx   3 admin hdfs   45157110 2018-02-09 19:37 /tmp/weblogs/1.tsv
-rwxrwxrwx   3 admin hdfs   44952637 2018-02-09 19:37 /tmp/weblogs/2.tsv
-rwxrwxrwx   3 admin hdfs   44732597 2018-02-09 19:38 /tmp/weblogs/3.tsv
-rwxrwxrwx   3 admin hdfs   22784226 2018-02-09 19:38 /tmp/weblogs/4.tsv
-rwxrwxrwx   3 admin hdfs   21122289 2018-02-09 19:39 /tmp/weblogs/5.tsv
-rwxrwxrwx   3 admin hdfs       1522 2018-02-09 19:39 /tmp/weblogs/urlmap.tsv
-rwxrwxrwx   3 admin hdfs    1870304 2018-02-09 19:40 /tmp/weblogs/users.tsv
[root@sandbox tmp]#
```

Using Scoop to import into Hive from RDBMS

```
sqoop import --connect
"jdbc:sqlserver://svsqlserver.database.windows.net:1433;databaseName=svsqldemodb" \
 --username svittalam -P --table product \
 --target-dir /tmp/urlmap --fields-terminated-by "," \
 --hive-import --create-hive-table --hive-table default.urlmap \
 -m 1

sqoop import --connect
"jdbc:sqlserver://svsqlserver.database.windows.net:1433;databaseName=svsqldemodb" \
 --username svittalam -P --table user \
 --target-dir /tmp/user --fields-terminated-by "," \
 --hive-import --create-hive-table --hive-table default.user \
 -m 1
```

Using Pig Latin to merge datasets

```
grunt>
grunt>
grunt> File0 = LOAD '/tmp/weblogs/0.tsv' USING PigStorage();
grunt> File1 = LOAD '/tmp/weblogs/1.tsv' USING PigStorage();
grunt> File2 = LOAD '/tmp/weblogs/2.tsv' USING PigStorage();
grunt> File3 = LOAD '/tmp/weblogs/3.tsv' USING PigStorage();
grunt> File4 = LOAD '/tmp/weblogs/4.tsv' USING PigStorage();
grunt> File5 = LOAD '/tmp/weblogs/5.tsv' USING PigStorage();
grunt> Final_file = UNION File0, File1, File2, File3, File4, File5;
grunt> STORE Final_file INTO '/tmp/Final_pigOutput' USING PigStorage();
2018-02-09 20:42:41,773 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNION
2018-02-09 20:42:41,865 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.7.3.2.5.0.0-1245      0.16.0.2.5.0.0-1245     root    2018-02-09 20:42:42     2018-02-09 20:43:47     UNION

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1518207558664_0002  6       0       39      34      37      38      0       0       0       0       File0,File1,File2,File3,File4,File5,Final_file  MAP_ONLY        /tmp/Final_pigOutput,

Input(s):
Successfully read 36270 records from: "/tmp/weblogs/5.tsv"
Successfully read 77529 records from: "/tmp/weblogs/1.tsv"
Successfully read 77137 records from: "/tmp/weblogs/2.tsv"
Successfully read 76782 records from: "/tmp/weblogs/3.tsv"
Successfully read 39078 records from: "/tmp/weblogs/4.tsv"
Successfully read 114470 records from: "/tmp/weblogs/0.tsv"

Output(s):
Successfully stored 421266 records (245434401 bytes) in: "/tmp/Final_pigOutput"

Counters:
Total records written : 421266
Total bytes written : 245434401
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1518207558664_0002


2018-02-09 20:43:47,954 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
2018-02-09 20:43:47,954 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox.hortonworks.com/172.17.0.2:8050
2018-02-09 20:43:47,956 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at sandbox.hortonworks.com/172.17.0.2:10200
2018-02-09 20:43:47,989 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-02-09 20:43:48,321 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
2018-02-09 20:43:48,321 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox.hortonworks.com/172.17.0.2:8050
2018-02-09 20:43:48,322 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at sandbox.hortonworks.com/172.17.0.2:10200
2018-02-09 20:43:48,373 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-02-09 20:43:48,661 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
2018-02-09 20:43:48,673 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox.hortonworks.com/172.17.0.2:8050
2018-02-09 20:43:48,673 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at sandbox.hortonworks.com/172.17.0.2:10200
2018-02-09 20:43:48,720 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-02-09 20:43:48,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Creating tables in Hive and loading data

```
[root@sandbox ~]# hive

Logging initialized using configuration in file:/etc/hive/2.5.0.0-1245/0/hive-log4j.properties
hive>
    > ;
hive>
    > use default;
OK
Time taken: 4.668 seconds
hive>
    >
    > CREATE EXTERNAL TABLE users
    > (swid string, birth_dt string, gender_cd string)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE
    > LOCATION "/tmp/weblogs/users";
OK
Time taken: 1.154 seconds
hive> CREATE EXTERNAL TABLE urlmap
    > (url string, category string)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE
    > LOCATION "/tmp/data/urlmap";
OK
Time taken: 0.541 seconds
hive>
```

```
hive> LOAD DATA INPATH '/tmp/weblogs1/weblogs/urlmap/urlmap.tsv' INTO TABLE URLMAP;
Loading data to table default.urlmap
Table default.urlmap stats: [numFiles=2, totalSize=3044]
OK
Time taken: 1.244 seconds
hive> select count(1) from urlmap;
Query ID = root_20180209233924_e933b8fb-b40a-4b7d-b2e8-b108f7130518
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1518207558664_0010)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.66 s
--------------------------------------------------------------------------------
OK
64
Time taken: 9.552 seconds, Fetched: 1 row(s)
```

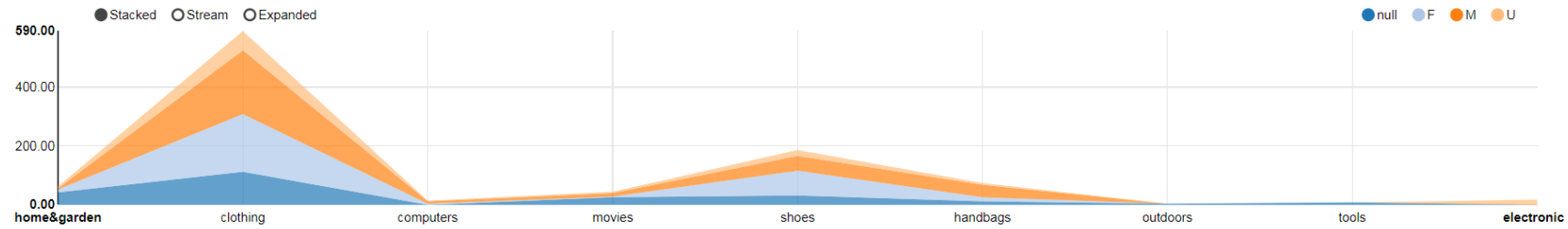
Creating dataset for analysis

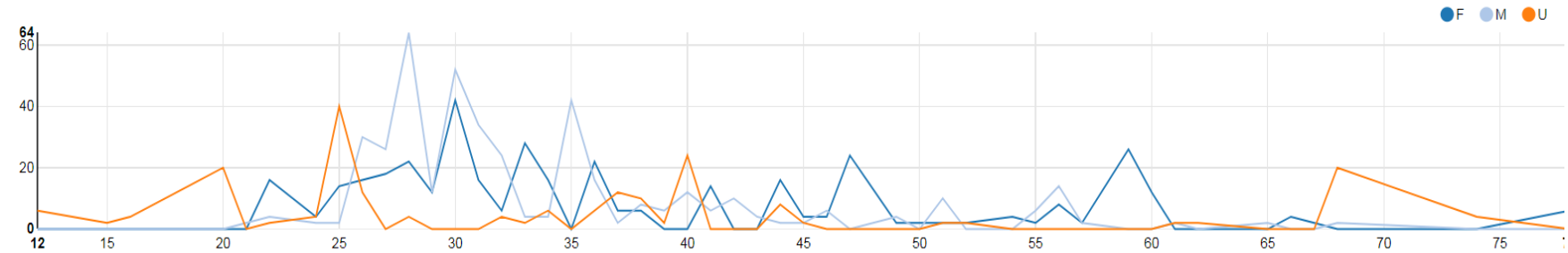
```
Logging initialized using configuration in file:/etc/hive/2.5.0.0-1245/0/hive-log4j.properties
hive>
    >
    >
    > create table clickpathanalytics as
    > select
    >         to_date(o.ts) logdate,
    >         o.url,
    >         o.ip,
    >         o.city,
    >         upper(o.`state`) `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
    >         webserverlogview o
    >         left outer join urlmap p on o.url = p.url
    >         left outer join users d on o.swid = concat('{', d.swid , '}');
Query ID = root_20180209235418_7f45bd95-38ff-48b7-878f-4ec1212d264c
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1518207558664_0011)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 2 ..........   SUCCEEDED      1          1        0        0       0       0
Map 3 ..........   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 42.55 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/clickpathanalytics
Table default.clickpathanalytics stats: [numFiles=1, numRows=842512, totalSize=81845684, rawDataSize=81003172]
OK
Time taken: 52.439 seconds
```

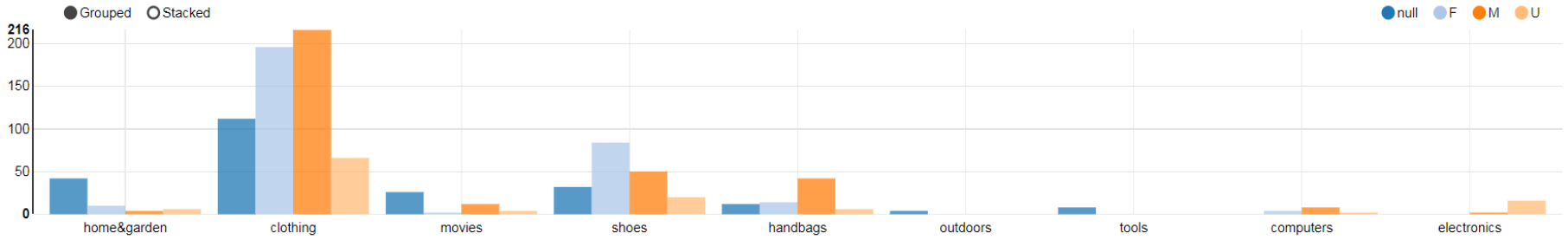
Final dataset for analysis

logdate	url	ip	city	state	country	category	age	gender
3/15/2012	http://www.acme.com/SH55126545/VD55170364	99.122.210.248	homestead	FL	usa	home&garden		
3/15/2012	http://www.acme.com/SH55126545/VD55170364	99.122.210.248	homestead	FL	usa	home&garden		
3/15/2012	http://www.acme.com/SH55126545/VD55177927	69.76.12.213	coeur d alene	ID	usa	clothing	36	F
3/15/2012	http://www.acme.com/SH55126545/VD55177927	69.76.12.213	coeur d alene	ID	usa	clothing	36	F
3/15/2012	http://www.acme.com/SH55126545/VD55166807	67.240.15.94	queensbury	NY	usa	computers	35	M
3/15/2012	http://www.acme.com/SH55126545/VD55166807	67.240.15.94	queensbury	NY	usa	computers	35	M
3/15/2012	http://www.acme.com/SH55126545/VD55149415	67.240.15.94	queensbury	NY	usa	movies	35	M
3/15/2012	http://www.acme.com/SH55126545/VD55149415	67.240.15.94	queensbury	NY	usa	movies	35	M
3/15/2012	http://www.acme.com/SH55126545/VD55179433	98.234.107.75	sunnyvale	CA	usa	shoes	22	M
3/15/2012	http://www.acme.com/SH55126545/VD55179433	98.234.107.75	sunnyvale	CA	usa	shoes	22	M
3/15/2012	http://www.acme.com/SH55126545/VD55179433	75.85.165.38	san diego	CA	usa	shoes	28	F
3/15/2012	http://www.acme.com/SH55126545/VD55179433	75.85.165.38	san diego	CA	usa	shoes	28	F
3/15/2012	http://www.acme.com/SH55126545/VD55166807	71.53.206.175	charlottesville	VA	usa	computers	26	F
3/15/2012	http://www.acme.com/SH55126545/VD55166807	71.53.206.175	charlottesville	VA	usa	computers	26	F
3/15/2012	http://www.acme.com/SH55126545/VD55179433	97.96.62.161	parrish	FL	usa	shoes	45	F
3/15/2012	http://www.acme.com/SH55126545/VD55179433	97.96.62.161	parrish	FL	usa	shoes	45	F
3/15/2012	http://www.acme.com/SH55126545/VD55170364	129.119.158.240	dallas	TX	usa	home&garden	28	F
3/15/2012	http://www.acme.com/SH55126545/VD55170364	129.119.158.240	dallas	TX	usa	home&garden	28	F

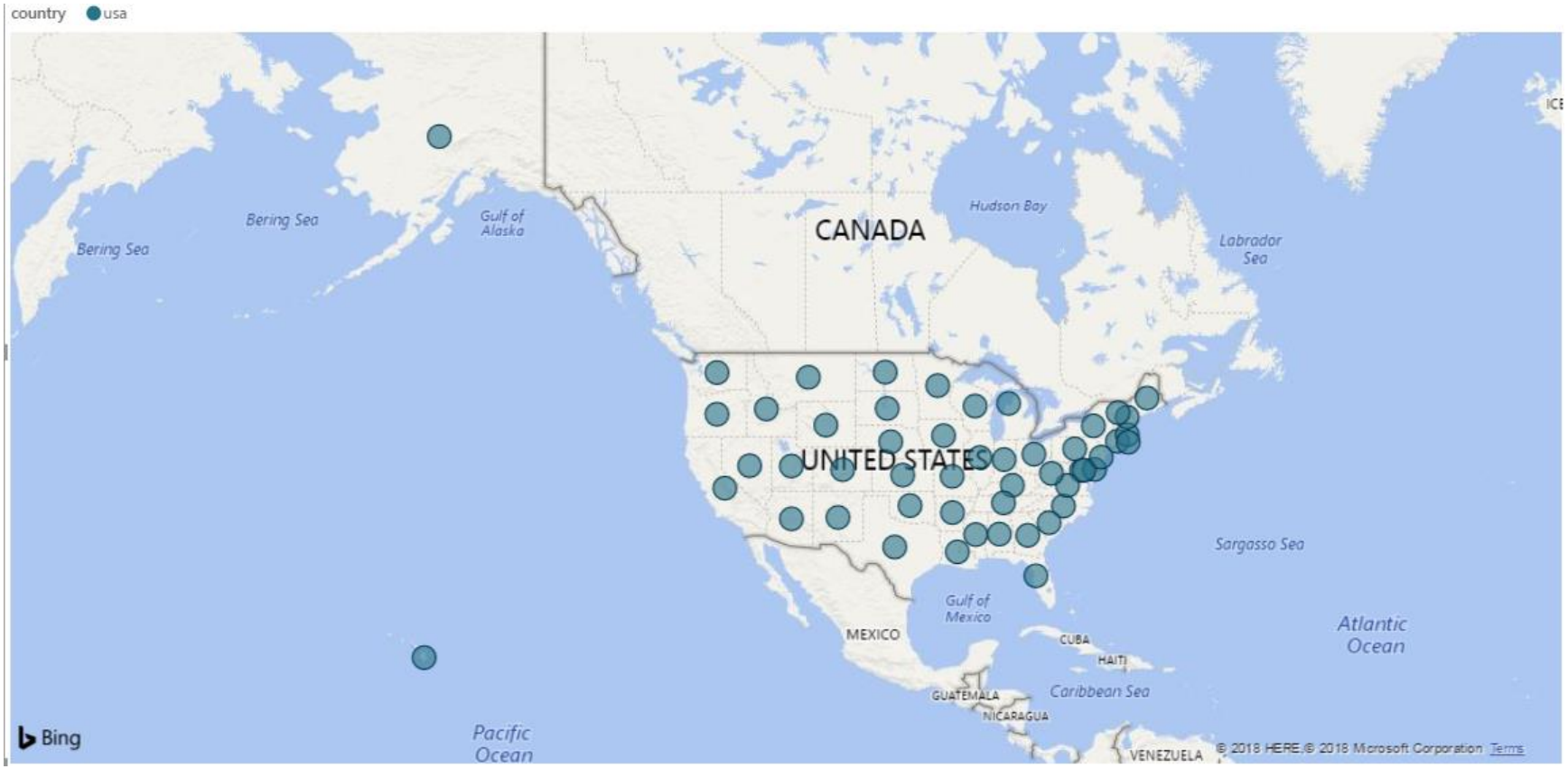
Visualization in Zepplin







Visualization in PowerBI



Lessons Learned

- When I used the HDInsight cluster and logged into the Ambari UI, I could not see all the utilities. For eg., I could not find the link for Zepplin.

YouTube URLs, GitHub URL, Last Page

- Two minute (short): https://youtu.be/eaO8d-c93G8
- 15 minutes (long): https://youtu.be/IYv9PkPdCjk
- GitHub Repository with all artifacts: https://github.com/satishvittalam/Azureproject

- https://hortonworks.com/tutorial/getting-started-with-apache-zeppelin/
- https://community.hortonworks.com/content/repo/56765/zeppelin-notebook-for-analysing-web-server-logs.html