MIDS W205

Lab #	9	Lab Title	Introduction to using Tableau with Hive
Related Module(s)	8	Goal	Introduction to using Tableau with Hive
Last Updated	9/29/15	Expected duration	30-40 min

BI Tableau Visualization Using Hive warehouse

Introduction

Tableau is one of the world's fastest-growing business intelligence companies. Tableau offers to quickly analyze, visualize and share information. As with most BI tools, Tableau can use Apache Hive (via ODBC connection) as the de facto standard for SQL access in Hadoop.

Here are a few topics, we will cover in this lab:

- Prerequisites and external resources
- Starting hive thrift server for remote hive access
- Creating Hive table and running sample query on hive
- Installing Tableau and ODBC Driver for connecting to Hive
- Configuring and connecting to Hadoop Hive from Tableau using ODBC driver
- Build visualization on Weblog Clickstream Analytics using Tableau

Let's go!

Step-1: Prerequisites and external resources:

For connecting to Hive Server, you must have one of the following distributions:

1. Cloudera distribution including Apache Hadoop CDH3u1 or later, which includes Hive 0.7.1 or later.
2. Hortonworks
3. MapR Enterprise Edition (M5)
4. Amazon EMR
5. Tableau Desktop Pro 8.x or 9 (recommended). Click here to Download.

Additionally, you must have the correct Hive ODBC driver installed on each machine running Tableau Desktop or Tableau Server. You can download it from Drivers page.

Step-2: Starting hive thrift server for remote hive access:

How to start:

```
hive –service hiveserver –help
```

hive –service hiveserver
or
hive –service hiveserver2

Here is a connection string for your reference:

with username and password:
jdbc:hive2://myhost.example.com:21050/test_db;user=fred;password=xyz123

without username and password:
jdbc:hive2://ec2-54-157-182-212.compute-1.amazonaws.com:10000/default

Step-3: Creating Hive table and running sample query on hive:

Here is the step to create the Web_Session_Log table that you have access to; Similarly, you can create any other table of you choice if you want.

```
create table Web_Session_Log
(DATETIME varchar(500), USERID varchar(500),
SESSIONID varchar(500),PRODUCTID varchar(500),
REFERERURL varchar(500))
row format delimited fields terminated by '\t'
stored as textfile
tblproperties("skip.header.line.count"="1");
```

Load data in to table:

HDFS:
LOAD DATA INFILE '/mnt/weblog/weblog.csv' INTO TABLE Web_Session_Log;
Local file system:
LOAD DATA LOCAL INFILE '/mnt/weblog/weblog.csv' INTO TABLE Web_Session_Log;

Sample Hive Query:
This is a query to find out which REFERERURL has been referred and for how many times.
select REFERERURL,count(*) from Web_Session_Log_1 GROUP BY REFERERURL;

Please refer below on exact code:

```
hive> create table Web_Session_Log
    > (DATETIME varchar(500),
    > USERID varchar(500),
    > SESSIONID varchar(500),
    > PRODUCTID varchar(500),
    > REFERERURL varchar(500))
    > row format delimited fields terminated by "\t"
    > stored as textfile;
OK
Time taken: 0.069 seconds
hive> LOAD DATA LOCAL INPATH "/mnt/weblogdata/" INTO TABLE Web_Session_Log;
Loading data to table default.web_session_log
Table default.web_session_log stats: [numFiles=1, totalSize=4513792]
OK
Time taken: 0.312 seconds
hive> select REFERERURL,count(*) from Web_Session_Log GROUP BY REFERERURL;
Query ID = root_20150903141515_a45a88a8-3a1d-4a4e-b0f3-1d78d9e35b1c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2015-09-03 14:15:16,239 Stage-1 map = 0%,  reduce = 0%
2015-09-03 14:15:17,251 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local504579210_0001
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
NULL    1
REFERERURL      1
http://www.abc.com      3303
http://www.amazon.com   3252
http://www.ebay.com     3263
```

Step 4: Installing Tableau and ODBC Driver for connecting to Hive

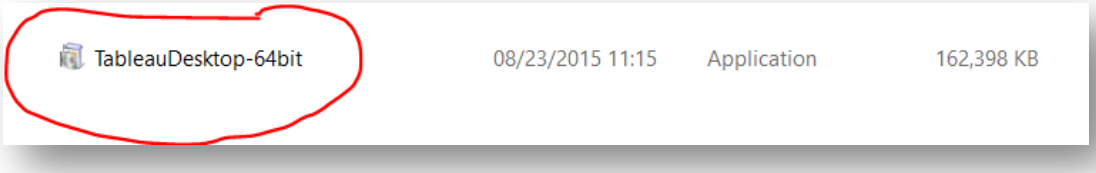
Hive is a data warehouse technology for working with data in your Hadoop cluster by using a combination of traditional SQL expressions and advanced Hadoop-specific data analysis and transformation operations. Tableau works with Hadoop using Hive to provide a user experience that requires no programming.

In this lab, we will connect the Tableau Desktop Pro with Hadoop server in order to access the hive table 'Web_Session_Log' which was created above.

Here is the information on how you can get Tableau on your desktop (Tableau Desktop Pro Edition):

You can download the Tableau Desktop Pro directly from Tableau's website (Windows or Mac) version based on your operating system(32 bit or 64 bit).

Once the product is downloaded, you need to install it just by double-clicking the installer package.



Here is the step to install the ODBC Driver for Hive:

For both Hive Server and Hive Server 2, you must install the Cloudera, Hortonworks, MapR, or Amazon EMR ODBC driver from the Drivers page. Ensure that the bit version of the driver you download matches the bit version of your operating system.

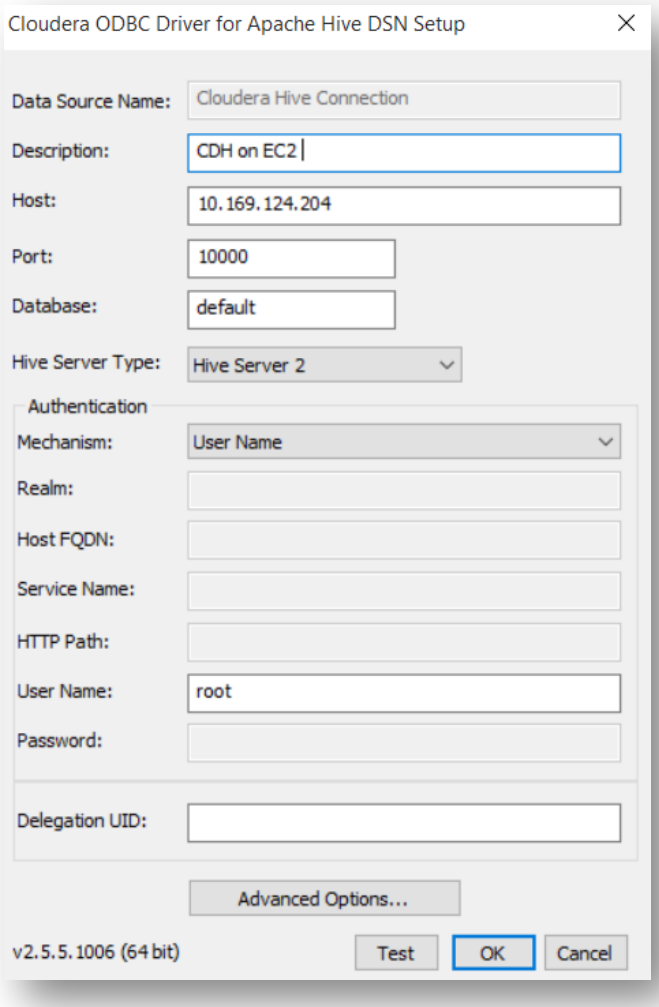
- **Cloudera (Hive):** Cloudera ODBC Driver for Apache Hive 2.5.x, 32-bit or 64-bit.
    - For use with Tableau Server 8.0.0-8.0.7, or 8.1.0-8.1.3, use version 2.5.0.1000.
    - For use with Tableau Server 8.0.8 and later, or 8.1.4 and later, use driver version 2.5.0.1001 or later.
- **Hortonworks**: Hortonworks Hive ODBC Driver 1.2.x (32-bit or 64-bit)
- **MapR**: MapR_odbc_2.1.0_x86.exe or later, or MapR_odbc_2.1.0_x64.exe or later
- **Amazon EMR**: HiveODBC.zip or ImpalaODBC.zip

Note: If you have a different version of the driver installed, first uninstall that driver before installing the version provided on the Drivers page.

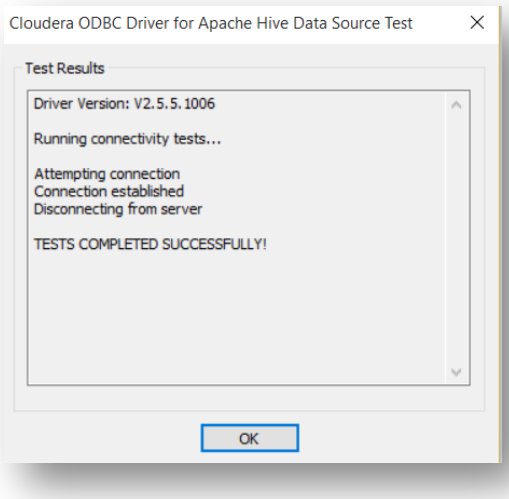
Step 5: Configuring and connecting to Hadoop hive from Tableau using ODBC driver

After installation of the appropriate ODBC driver, you need to configure the 'ODBC System DSN' on Windows. Based on the version of your Tableau Desktop (x86 or x64), download the driver & install it. Upon installation, open Control Panel -> System & Security -> Administrative Tools -> ODBC Data sources (32 bit or 64 bit).

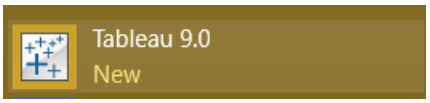
Open the ODBC data source & go to System DSN tab & configure it by applying Host (Hadoop server name), Port (10000), Database(default), Hive Server Type (Hive Server 2) & authentication panel , you can enter the Username & Password credentials (if you configured security settings on server authentication).



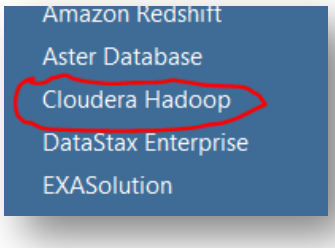
First you click on 'Test' button to check the connectivity of tableau to Hadoop server, if the connection is successful, it should show a success prompt otherwise, it would display a message with error flag.



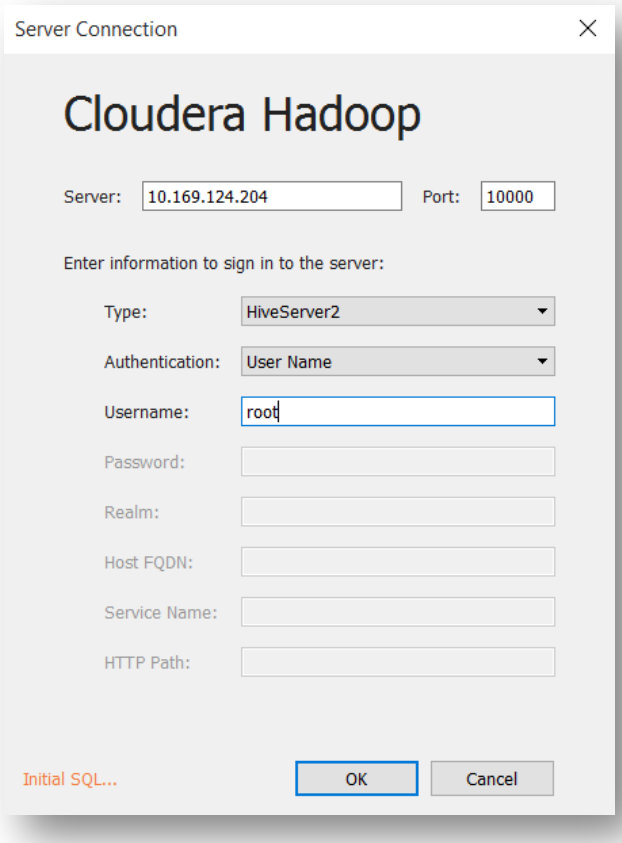
Now, Open Tableau Desktop Pro from quick launch option.



Next, select Data -> New Data Source tab, click on Cloudera Hadoop to connect to Hadoop server.



The Cloudera Hadoop connection pane should open where you need enter the Hadoop server credentials to connect. Make sure that to connect to hive tables the default port number should be '10000', type should be 'HiveServer2', and the username needs to be provided to connect to Hadoop server.

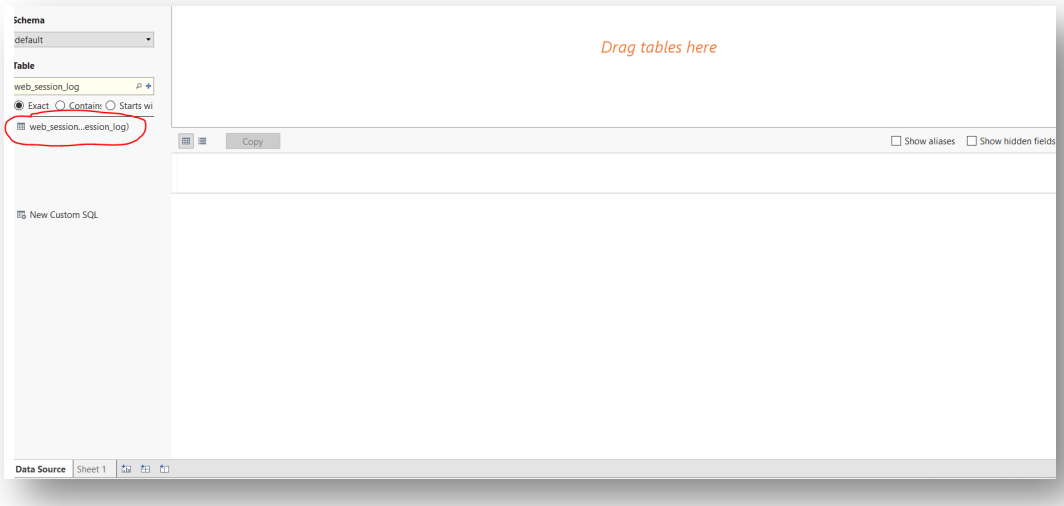

Server Connection
Cloudera Hadoop
Server:
10.169.124.204
Port:
10000
Enter information to sign in to the server:
Type:
HiveServer2
Authentication:
User Name
Username:
root
Password:
Realm:
Host FQDN:
Service Name:
HTTP Path:
Initial SQL...
OK
Cancel

Step 6: Build visualization on Weblog, Clickstream Analytics using Tableau

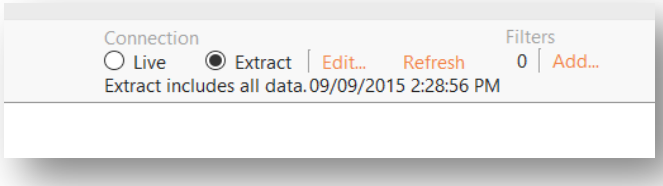
Once, the data source connection is made successfully, next you can connect to the hive table & click on 'Extract' to get data locally & store tableau in-memory in order to get rid of latency as hive is based on batch processing mechanism.

Switch back to Tableau, under data connection pane, enter the schema name, write 'default' & click on the 'search' icon on the right side of the textbox for schema.

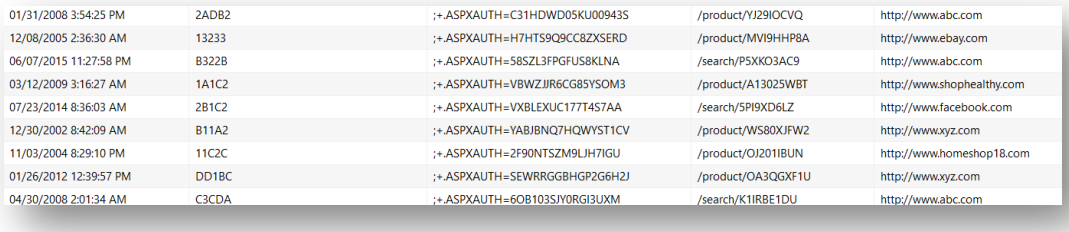
Next, click on 'Table' name textbox, enter the table name for the demo 'Web_Session_Log', click on the 'search' icon on the right side of the textbox & select it & drag it to the upper right side of the tableau window which is named as 'Drag tables here' like the following screenshot.



Now, once the table is dragged, you can extract the data locally on in-memory tableau dataset in order to avoid unnecessary server latency.



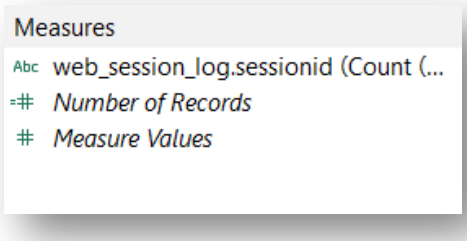
Click on ‘Automatic update’ to refresh the dataset. It should look like the following screen.



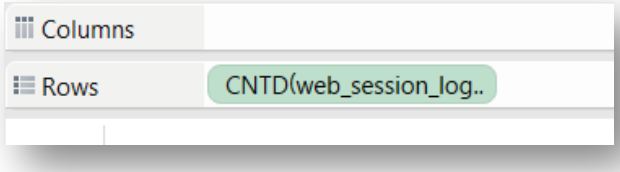
01/31/2008 3:54:25 PM	2ADB2	;+.ASPXAUTH=C31HDWD05KU00943S	/product/YJ29IOCVQ	http://www.abc.com
12/08/2005 2:36:30 AM	13233	;+.ASPXAUTH=H7HTS9Q9CC8ZXSERD	/product/MVI9HHP8A	http://www.ebay.com
06/07/2015 11:27:58 PM	B322B	;+.ASPXAUTH=58SZL3FPGFUS8KLNA	/search/P5XKO3AC9	http://www.abc.com
03/12/2009 3:16:27 AM	1A1C2	;+.ASPXAUTH=VBWZJJR6CG85YSOM3	/product/A13025WBT	http://www.shophealthy.com
07/23/2014 8:36:03 AM	2B1C2	;+.ASPXAUTH=VXBLEXUC177T4S7AA	/search/5PI9XD6LZ	http://www.facebook.com
12/30/2002 8:42:09 AM	B11A2	;+.ASPXAUTH=YABJBNQ7HQWYST1CV	/product/WS80XJFW2	http://www.xyz.com
11/03/2004 8:29:10 PM	11C2C	;+.ASPXAUTH=2F90NTSZM9LJH7IGU	/product/OJ201IBUN	http://www.homeshop18.com
01/26/2012 12:39:57 PM	DD1BC	;+.ASPXAUTH=SEWRRGGBHGP2G6H2J	/product/OA3QGXF1U	http://www.xyz.com
04/30/2008 2:01:34 AM	C3CDA	;+.ASPXAUTH=6OB103SJY0RGI3UXM	/search/K1IRBE1DU	http://www.abc.com

Click on ‘Worksheet’ tab on top left corner & select ‘New Worksheet’ to start building graphs.

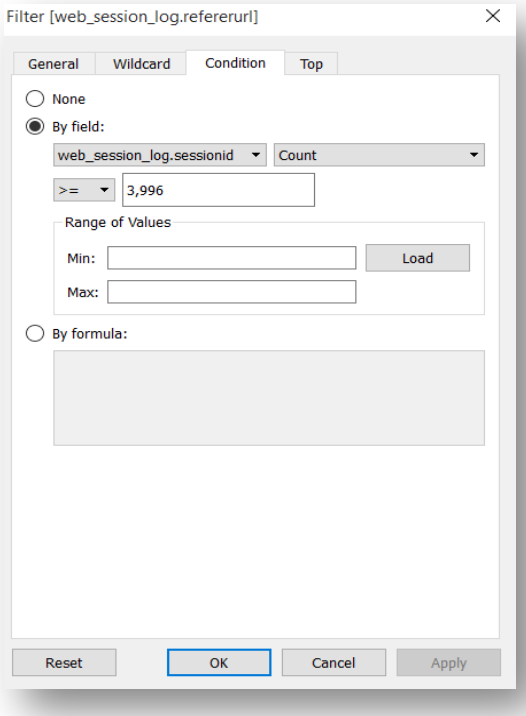
On the new worksheet tab, under ‘Data’ pane, drag the ‘web_session_log.sessionid’ field to the ‘Measures’ pane & add it to ‘Count the number of sessions’.



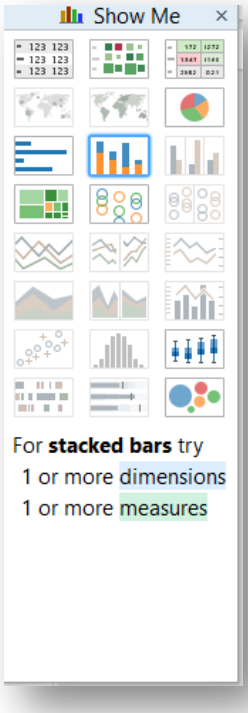
Next, drag it to the ‘Rows’ tab of graph.



Drag the field ‘web_session_log.refererurl’ to the Column tab & right click on the value to do a filter in order to visualize top 5 referring URLs. Select ‘Filter’ & click on ‘Condition’ tab in order to get URLs based on ‘Web_Session_log.sessionids’ beyond a certain value (e.g. web_session_log.sessionid >= 3,996) to drill down to top most 5 Referring URLs.

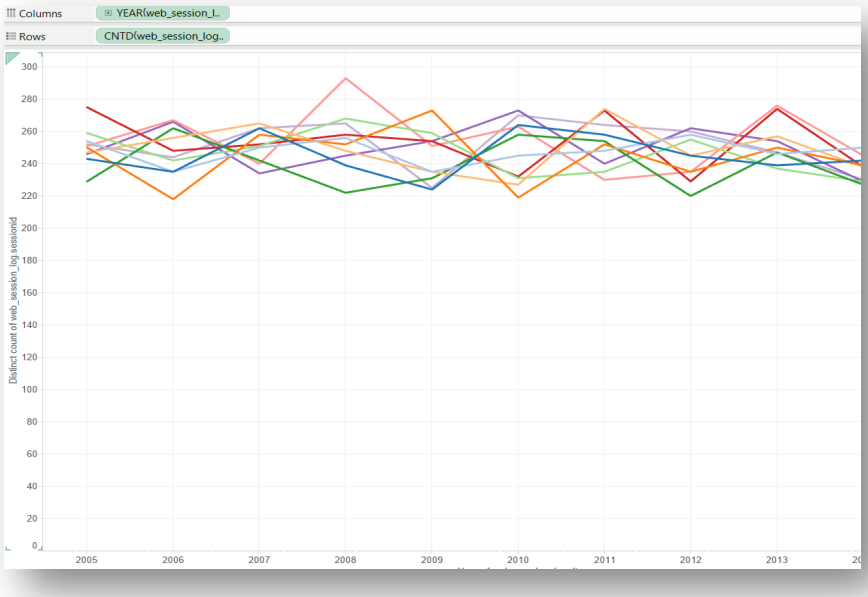


Once you finish the setting, select the chart type from ‘Show Me’ pane on right side of window.



Select the chart type as ‘Packed Bubbles’ & you will see the graph of ‘**Top 5 Referring URL**s’ like as the following graph.

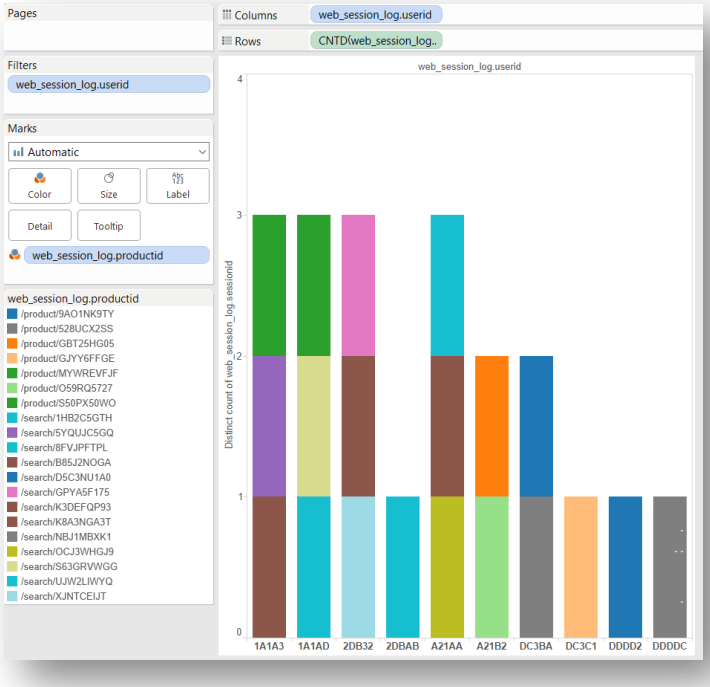
Next, select another worksheet to build out the ‘**Top Referring URLs over last 10 years’**. Drag the ‘web_session_log.sessionid’ to ‘Rows’ as ‘average count of sessions’ & ‘datetime’ to the ‘Columns’, check the chart type as ‘lines’.



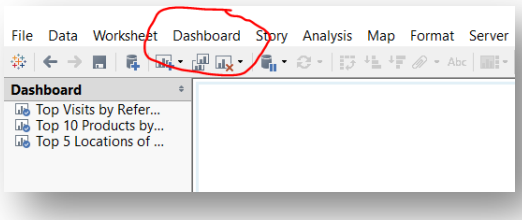
Go to Worksheet tab, start build a graph on '**Top referrer URLs based on user session count over last 5 yrs.**'.

On the next worksheet tab, drag the 'sessionid' field to 'Rows' & 'Userid' to the 'Columns' tab to build the graph on '**Top 10 users who used top 10 products'** by filtering the 'Users' based on maximum count of sessions. You need to filter the 'userid' data based on condition of 'sessionid' value.

You may select the chart type as simple 'stacked bar' chart.

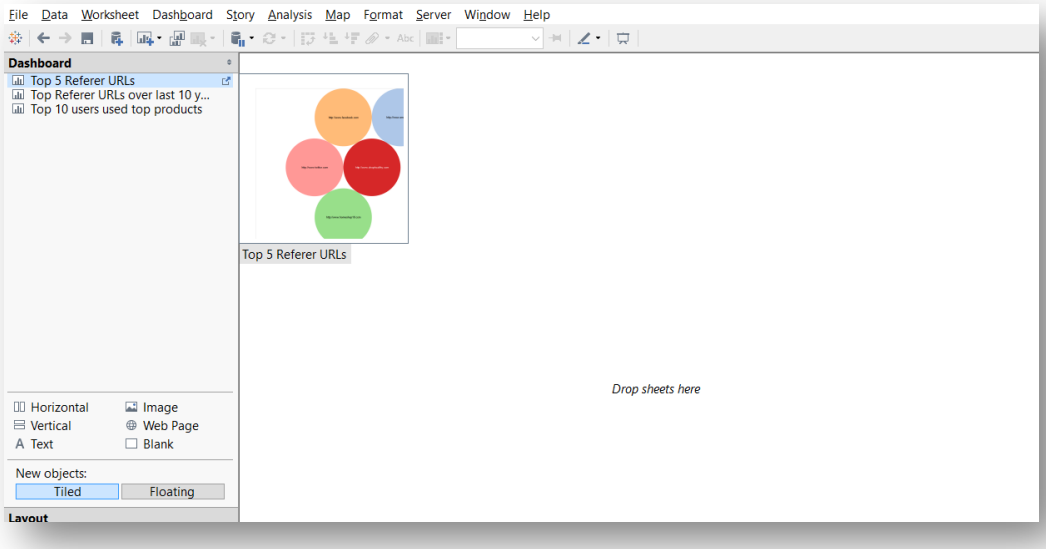


Once, you build such few graphs, click to 'Dashboard' tab on top of Tableau window



Select 'New Dashboard' in order to start implementing **weblog- clickstream analytics** dashboard.

On the new dashboard page, just drag the existing implemented worksheet graphs to the right side pane ‘Drag sheets option’.



Drag & Drop the implemented graphs on exiting worksheet & you will see the ‘**Weblog Clickstream analytics’** dashboard on Tableau Desktop like as following screenshot.

You can also publish the workbook on tableau server if you have those credentials for future reference & always save the workbook on your local drive from File -> Export as Packaged workbook' option.

You may also print the entire workbook or selected worksheet as pdf format as required.

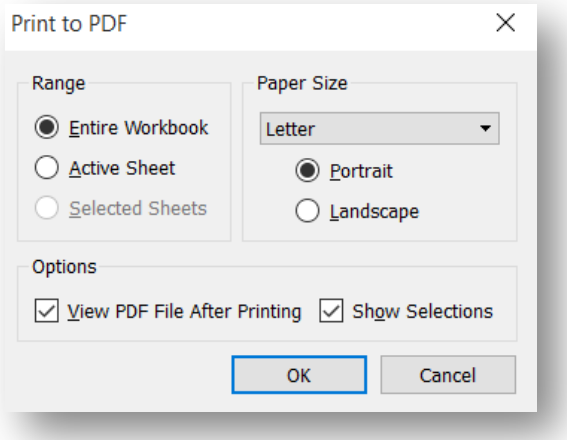
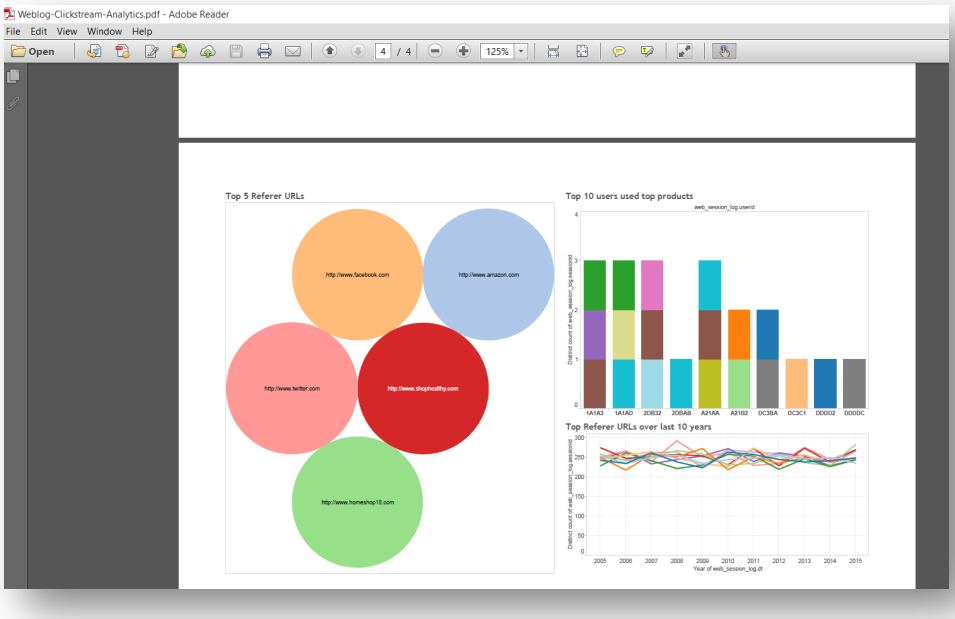


Tableau 'weblog-hive-clickstream analytics' workbook is printed on .pdf format.



Questions:

1. Which version of Tableau did you use?
2. What are the various backend databases you can connect Tableau to for reporting?
3. List most used features of Tableau in the Industry
4. What is the advantage of in memory feature of Tableau

Note. Please refer to Tableau official website for further reading and learnings.