

MIDS W205

Lab #	7	Lab Title	Introduction to Using Tableau With Hive
Related Module(s)	6	Goal	Introduction to Using Tableau With Hive
Last Updated	10/14/15	Expected Duration	30–40 minutes

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.

We will cover the following topics in this lab:

- Creating a Hive table and running a 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

Step 0: Running the Required Services

The following steps launch the required services and prepare your instance for the next steps.

- 1- Make sure port 10000 is open on your instance. To do this, edit your instance security group and add a TCP port 10000 accessible from anywhere in the Inbound section.
- 2- SSH to your instance and do not forget to mount /data
- 3- `cd /data`
- 4- Start Hadoop: `./start_hadoop.sh`
- 5- Start Postgres for hive meta data: `./start_postgres.sh`
- 6- Start Hive Metastore: `sudo -u w205 /data/start_metastore.sh`

Step 1: Creating Hive Table and Running a Sample Query on Hive

The following code creates a table—`web_session_log`—from the weblog data introduced in previous labs. While this lab uses Tableau to explore the weblog data, feel free to proceed with any table 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 into the Table:

From HDFS:

```

LOAD DATA INFILE '/mnt/weblog/weblog.csv'
INTO TABLE Web_Session_Log;

```

From Local File System:

```

LOAD DATA LOCAL INPATH '/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
GROUP BY REFERERURL;

```

Please see below for the 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 2: Starting a Hive Thrift Server for Remote Hive Access

HiveServer2 is a server interface that allows remote clients to execute queries against Hive. In Tableau, we will be able to extract data from our Web_Session_Log table by sending requests through HiveServer2.

How to start:

```

hive --service hiveserver2
or
hive --service hiveserver --help
hive --service hiveserver

```

(Note that as of Hive 1.0.0, HiveServer was removed in favor of HiveServer2. It is recommended you use 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: Installing Tableau and ODBC Driver for Connecting to Hive

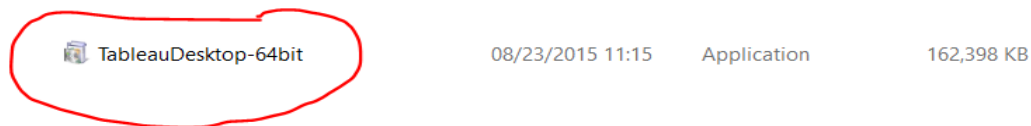
Hive is a data warehouse technology for working with data in your Hadoop cluster using a combination of traditional SQL expressions and advanced Hadoop-specific data analysis and transformation operations. Tableau works with Hadoop using Hive to provide a user experience that requires no programming.

In this lab, we will connect Tableau Desktop Pro with the HiveServer in order to access the Hive table `Web_Session_Log`, which we created earlier.

To install Tableau on your desktop (Tableau Desktop Pro Edition):

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

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



To install the ODBC driver for Hive:

For both HiveServer and HiveServer2, you must install the Cloudera, Hortonworks, MapR, or Amazon EMR ODBC driver from the Drivers page. Ensure that the version of the driver you download matches the bit version of your operating system.

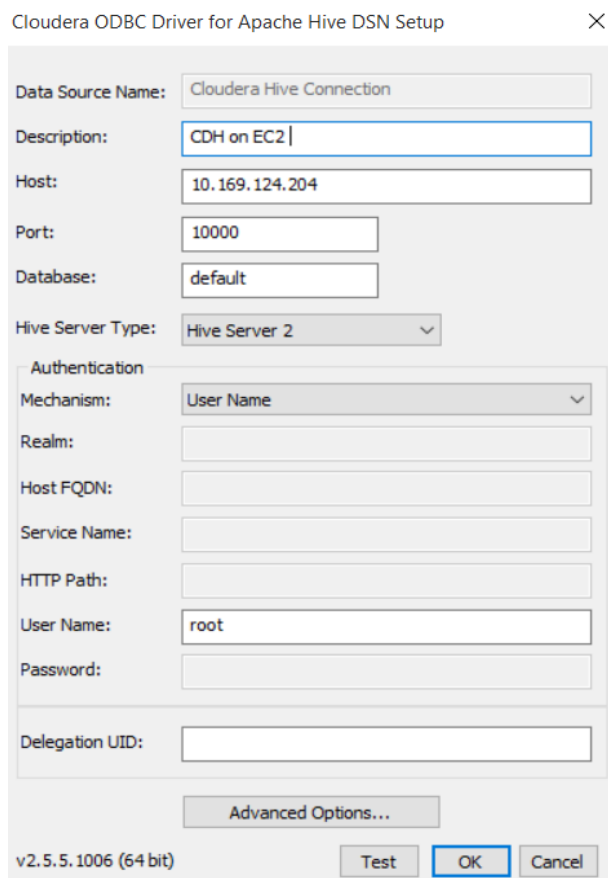
- **Cloudera (Hive):** Cloudera ODBC Driver for Apache Hive 2.5.x, 32-bit or 64-bit
 - For use with Tableau Server 8.0.0–8.0.7; for 8.1.0–8.1.3, use version 2.5.0.1000.
 - For use with Tableau Server 8.0.8 and later; for 8.1.4 and later, use [driver](#) version 2.5.0.1001 or later.
 - Cloudera drivers can be found [here](#).
- **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, uninstall that driver before installing the version provided on the Cloudera website.

Step 4: Configuring and Connecting to Hadoop Hive From Tableau Using ODBC Driver (Windows Only)

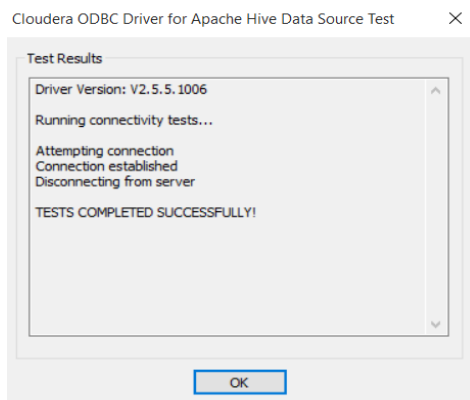
After installing the appropriate ODBC driver, you need to configure the ODBC System DSN on Windows. Download the driver for the appropriate version of Tableau Desktop (x86 or x64), and install it. Go to Control Panel -> System & Security -> Administrative Tools -> ODBC Data sources (32-bit or 64-bit).

Open the ODBC data source, go to the System DSN tab, and configure it by applying Host (Hadoop server name), Port (10000), Database(default), HiveServer Type (HiveServer2), and the authentication panel, then enter the username and password credentials (if you configured security settings on server authentication).



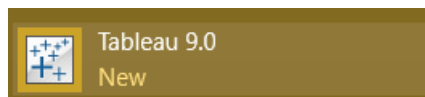
The screenshot shows the 'Cloudera ODBC Driver for Apache Hive DSN Setup' dialog box. The 'Data Source Name' is 'Cloudera Hive Connection'. The 'Description' is 'CDH on EC2'. The 'Host' is '10.169.124.204'. The 'Port' is '10000'. The 'Database' is 'default'. The 'Hive Server Type' is 'Hive Server 2'. The 'Authentication' section is expanded, showing 'Mechanism' as 'User Name', 'Realm' as an empty field, 'Host FQDN' as an empty field, 'Service Name' as an empty field, 'HTTP Path' as an empty field, 'User Name' as 'root', and 'Password' as an empty field. There is a 'Delegation UID' field at the bottom. An 'Advanced Options...' button is located below the 'Delegation UID' field. At the bottom of the dialog, there is a version string 'v2.5.5.1006 (64 bit)' and three buttons: 'Test', 'OK', and 'Cancel'.

Click **Test** to check the connectivity of Tableau to the Hadoop server. If the connection is successful, a success message is displayed. Otherwise, an error message is displayed.

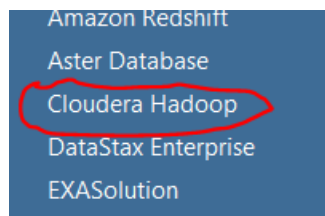


Step 5: Connect Tableau to HiveServer/HiveServer2 Using ODBC Driver

Open Tableau Desktop Pro from its quick-launch option.



Next, go to Data -> New Data Source tab, and click **Cloudera Hadoop** to connect to the Hadoop server.



The Cloudera Hadoop connection pane opens. Enter the Hadoop server credentials to connect. To connect to the HiveServer, the default port number should be 10000, the Type should be HiveServer2, and the username needs to be provided.

Server Connection ✕

Cloudera Hadoop

Server: Port:

Enter information to sign in to the server:

Type:

Authentication:

Username:

Password:

Realm:

Host FQDN:

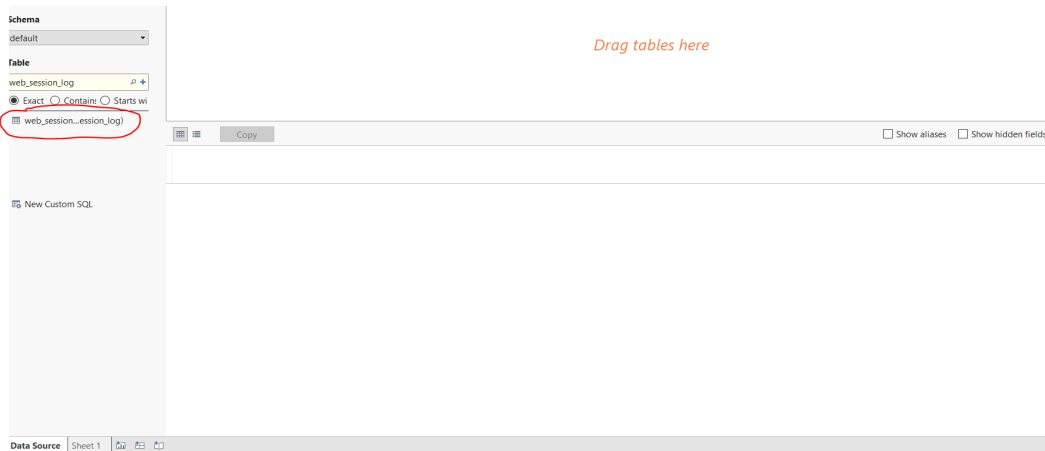
Service Name:

HTTP Path:

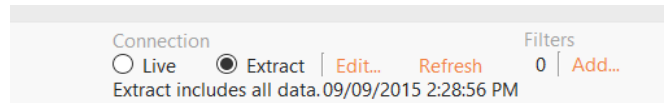
[Initial SQL...](#)

Step 6: Build Visualizations on Weblog, Clickstream Analytics Using Tableau

After the data source connection is made successfully, you can connect to the Hive table. Click **Extract** to get data locally and store Tableau in memory reduce latency because Hive is based on a batch-processing mechanism. Switch back to Tableau. Under the Data Connection pane, enter the schema name, enter **default**, and click the **Search** icon on the right side of the textbox. Next, enter the table name `Web_Session_Log` in the Table name textbox, click the **Search** icon on the right side of the textbox, and select and drag the table to the upper-right side of the Tableau window, as shown in the following screenshot.



Now you can extract the data locally on an in-memory Tableau dataset to avoid unnecessary server latency.

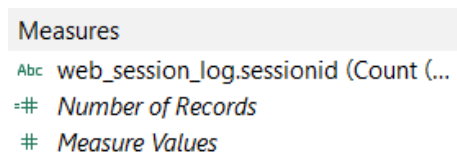


Click **Automatic update** to refresh the dataset. It should look like the following screen:

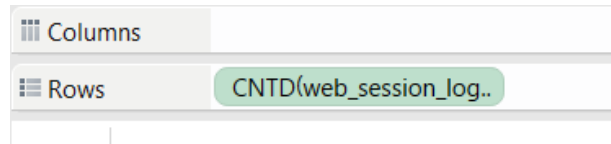
01/31/2008 3:54:25 PM	2ADB2	:+ ASPXAUTH=C31HDWD05KU00943S	/product/YJ29IOCVQ	http://www.abc.com
12/08/2005 2:36:30 AM	13233	:+ ASPXAUTH=H7HTS9Q9CC8ZXSERD	/product/MV19HHP8A	http://www.ebay.com
06/07/2015 11:27:58 PM	B322B	:+ ASPXAUTH=58SZL3FPGFUS8KLNA	/search/P5XKO3AC9	http://www.abc.com
03/12/2009 3:16:27 AM	1A1C2	:+ ASPXAUTH=V8WZJIR6CG85SOM3	/product/A13025WBT	http://www.shophealthy.com
07/23/2014 8:36:03 AM	2B1C2	:+ ASPXAUTH=VXBLEXUC177T4S7AA	/search/SP19XD6LZ	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/OJ2011BUN	http://www.homeshop18.com
01/26/2012 12:39:57 PM	DD1BC	:+ ASPXAUTH=SEWRRGG8HGP2G6H2J	/product/OA3QGXF1U	http://www.xyz.com
04/30/2008 2:01:34 AM	C3CDA	:+ ASPXAUTH=60B103SJY0RGI3UXM	/search/K1RBE1DU	http://www.abc.com

Click the **Worksheet** tab, and select **New Worksheet** to start building visualizations.

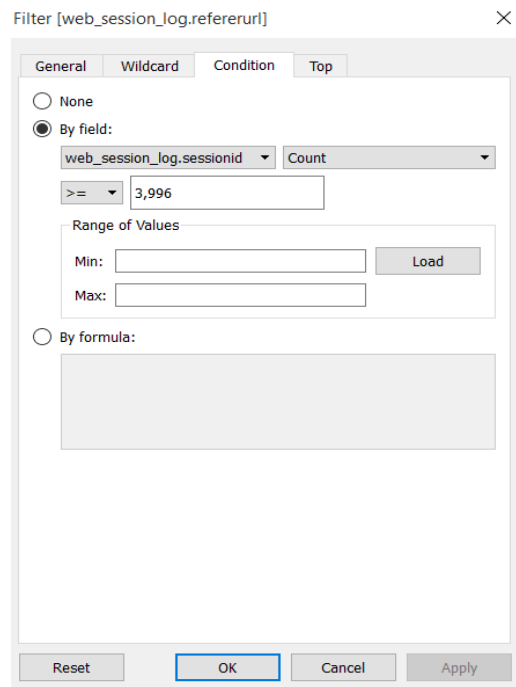
On the new worksheet tab, on the Data pane, drag the `web_session_log.sessionid` field to the Measures pane.



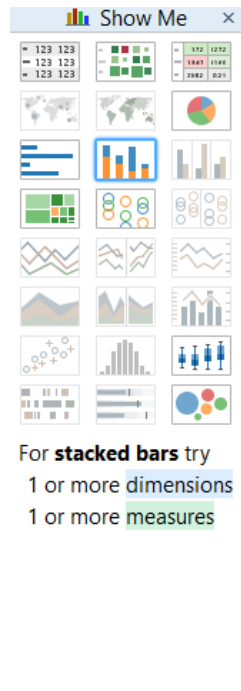
Next, drag the field to the **Rows** shelf. Hover over the field to expand the drop-down menu. Change the aggregation method (under Measure) from Sum to Count (Distinct).



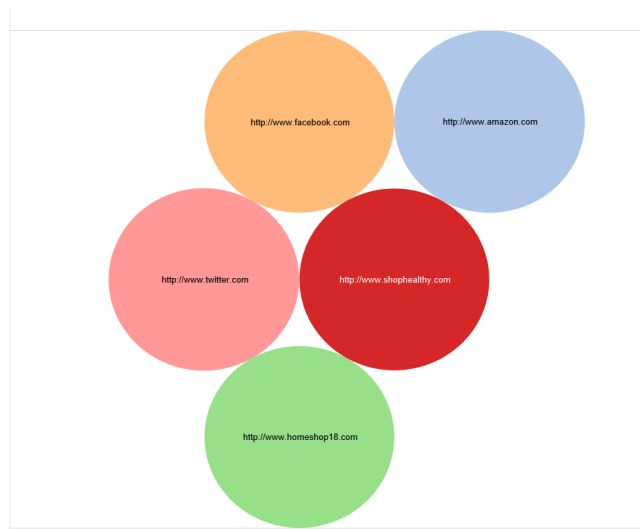
Drag the `web_session_log.referrerurl` field to the Column shelf, then right-click the value to create a filter to visualize the top five referring URLs. Select **Filter**, and click the **Condition** tab to get URLs based on `web_session_log.sessionids` beyond a certain value (e.g., `web_session_log.sessionid >= 3,996`) to drill down to the top five referring URLs.



After you define the settings, change the chart type in the Show Me pane on the right side of the window.

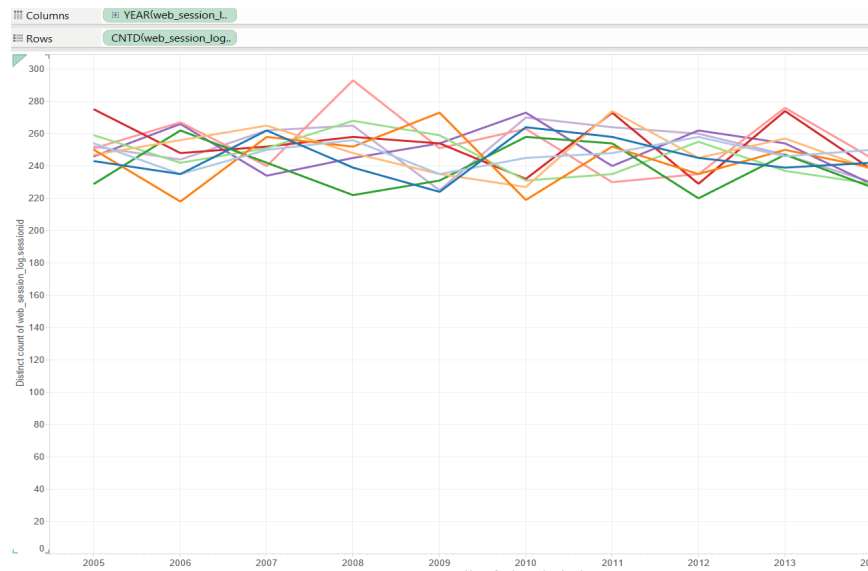


Select the bubbles chart type. You will see the graph, “Top 5 Referring URLs,” as follows:

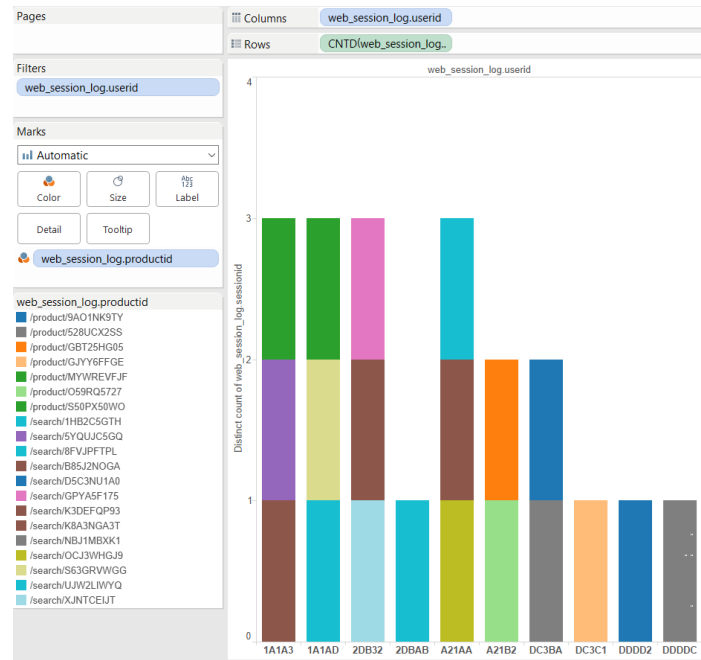


Next, create a new worksheet and name it **Top Referring URLs over last 10 years**. Drag `web_session_log.sessionid` to the Rows shelf. On the Data pane, right-click on the `datetime` field, and change its data type to Date and Time. Then, drag it to the Columns shelf. Drag the Referrer URL field to the Color section of the

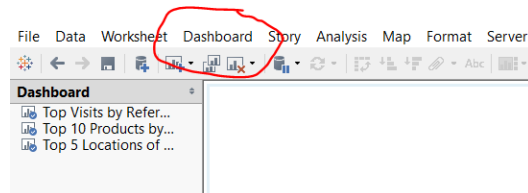
Marks pane. In the Show Me section, select the lines chart type.



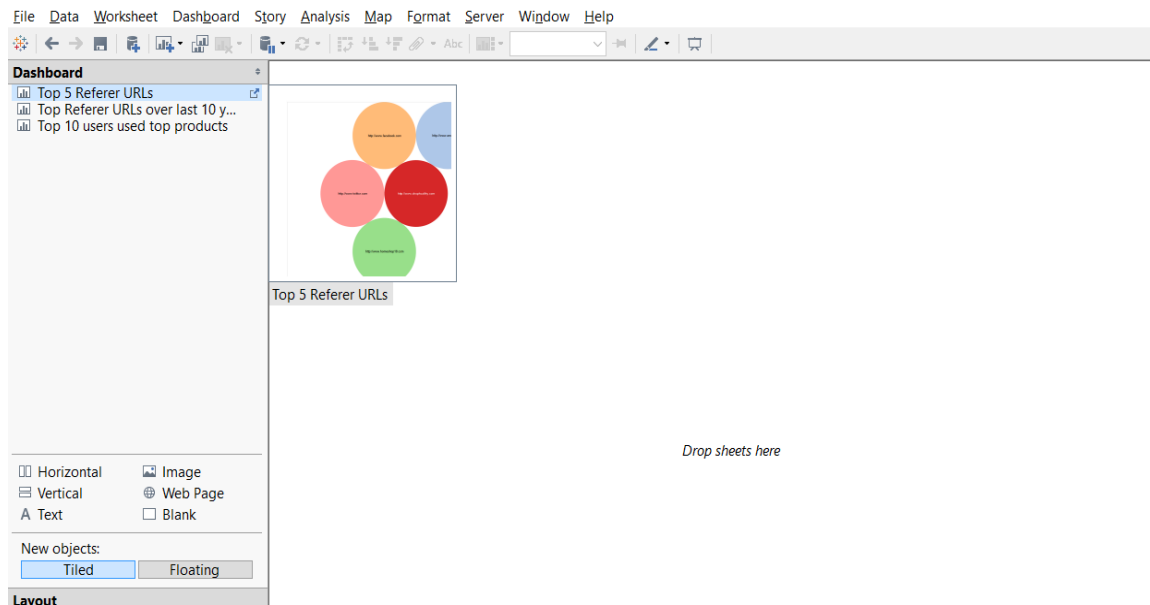
Go to a new worksheet tab, and name it **Top 10 users who used top 10 products**. On this worksheet tab, drag the `sessionid` field to the Rows shelf and the `userid` field to the Columns shelf. Filter the users based on the maximum session count. You need to filter the `userid` data based on condition of the `sessionid` value. Drag the `productID` field to the color field in the Marks pane to differentiate between products. You may select the simple stacked-bar chart type.



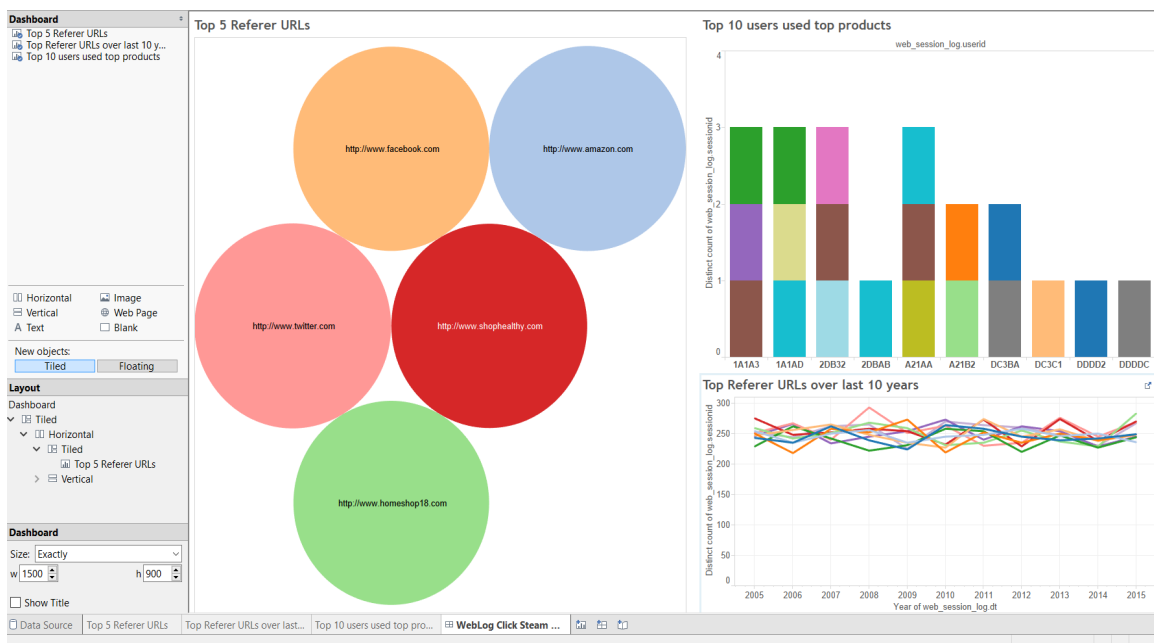
Once, you build such few graphs, click the **Dashboard** tab on top of the Tableau window



Select **New Dashboard** to start implementing the **weblog-clickstream analytics** dashboard. On the new dashboard page, drag the existing implemented worksheet graphs to the right pane.

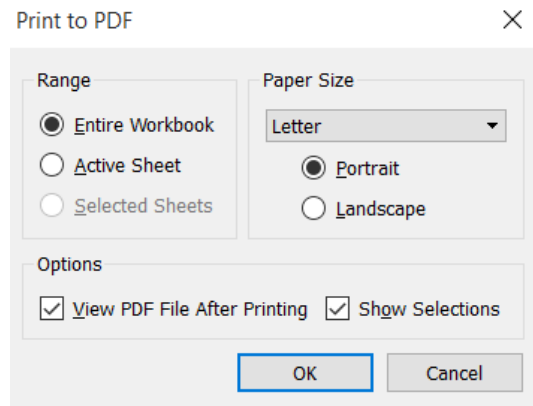


When you drag and drop the implemented graphs on the existing worksheet, you will see the Weblog Clickstream analytics dashboard on Tableau Desktop, as shown in the following screenshot:

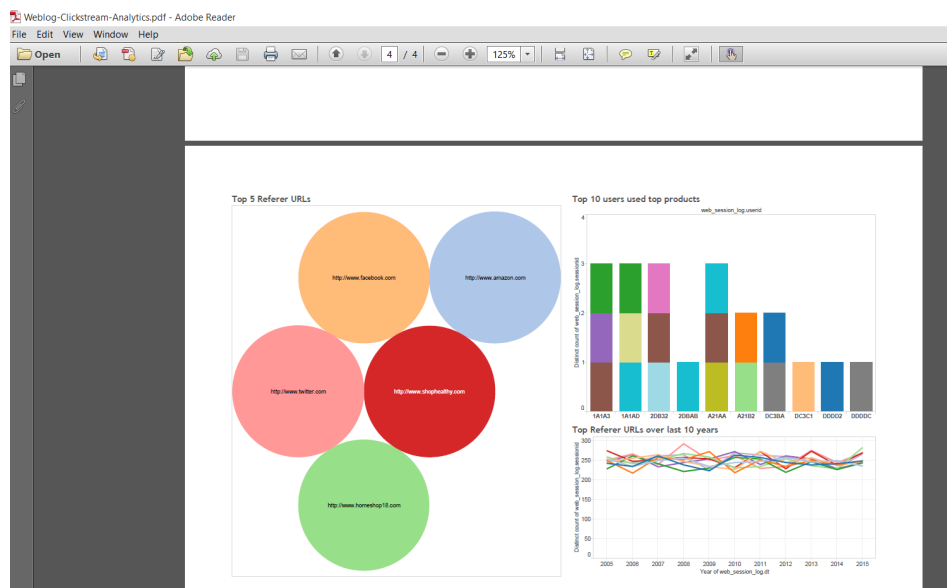


You can also publish the workbook on Tableau Server if you have those credentials. Always save the workbook on your local drive using the File -> Export as Packaged workbook option. You can

also print the entire workbook or selected worksheets in PDF format as required.



The Tableau ***weblog-hive-clickstream analytics workbook*** is printed in PDF format.



Task: Submit the printed PDF ***weblog-hive-clickstream analytics workbook***