Installing and Configuring the Demo containers

The MapR PACC Docker image is available from Docker Hub, and the containers and associated scripts for this demo are available on GitHub as shown below.

https://github.com/maprpartners/lenovo-demo
https://hub.docker.com/r/maprpartners/

These are the steps to follow to get the demo up and running:

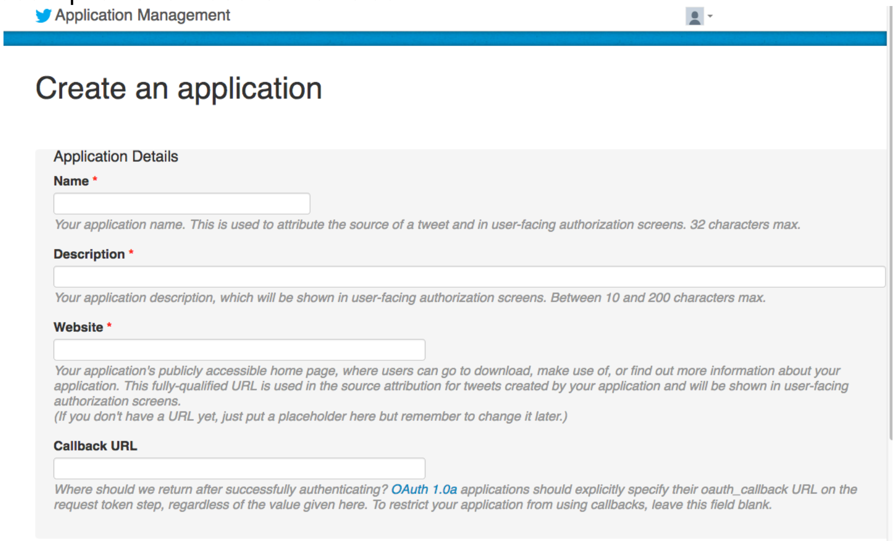
- Install Docker on an edge node
  Login to the edge node as root, issue below command to install Docker:

  ```
  yum –y install docker
  systcmctl enable docker
  systemctl start docker
  ```

  Now docker should be running on the edge node, you can issue this command to verify

  ```
  docker images
  ```

- Create your Twitter application and get the credentials for accessing the Twitter API
  You can go to https://twitter.com and create an account, or you can use your existing account. Once you have your account, go to https://apps.twitter.com and create an application by filling in the name, description and website info fields.



Create an application

Application Details

Name *

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

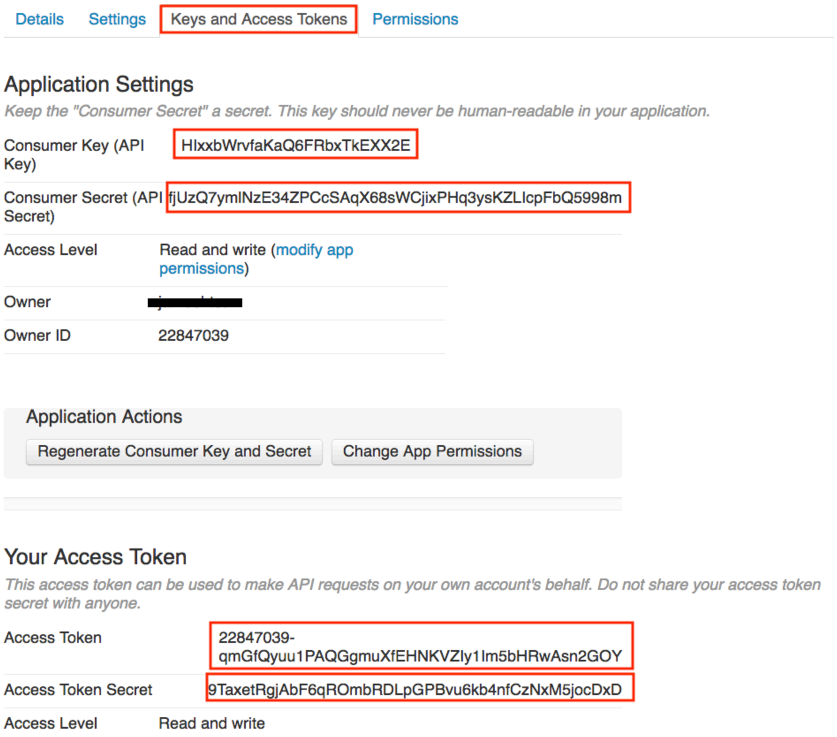
Website *

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Next, you will click the “Keys and Access Tokens” tab. You should then have the consumer key, consumer secret, access token and access token secret. Copy and save them somewhere, you will need them later.



Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	HlxxbWrvfaKaQ6FRbxTkEXX2E
Consumer Secret (API Secret)	fjUzQ7ymlNzE34ZPCcSAqX68sWCjixPHq3ysKZLlcpFbQ5998m
Access Level	Read and write (modify app permissions)
Owner	[redacted]
Owner ID	22847039

Application Actions

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	22847039-qmGfQyuu1PAQGgmuXfEHNKVZly1Im5bHRwAsn2GOY
Access Token Secret	9TaxetRgjAbF6qROmbRDLpGPBvu6kb4nfCzNxM5jocDxD
Access Level	Read and write

- Create a MapR stream and a MapR-DB table on the MapR cluster
  Login to the MapR head node as root and run the following commands:

  ```
  su mapr -c "echo <cluster user mapr’s password> | maprlogin password"
  su mapr -c "maprcli stream delete -path /tweets"
  su mapr -c "maprcli stream create -path /tweets"
  su mapr -c "maprcli table delete -path /tmp/tweets”
  su mapr -c "maprcli table create -path /tmp/tweets -tabletype json"
  ```

- Start the pre-built producer and consumer containers. The producer will get the tweets through the Twitter API and publish them to MapR-ES (MapR stream). The consumer will subscribe to MapR stream, get the tweets and save them into a MapR-DB table.

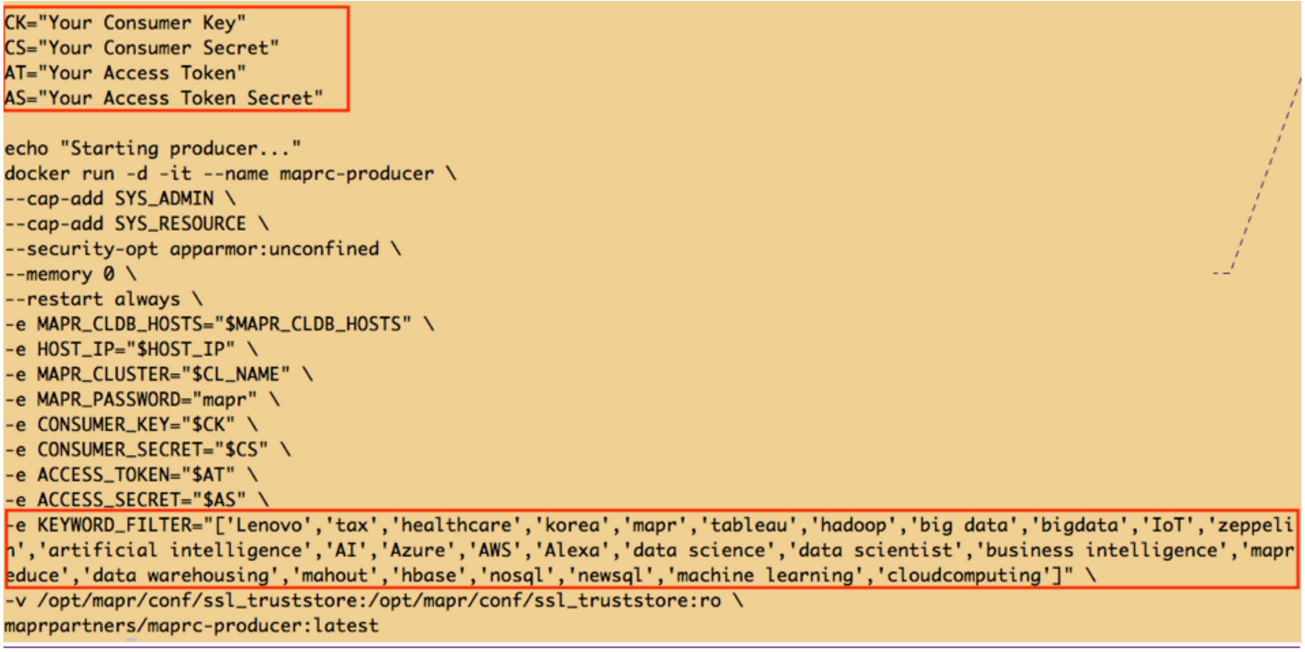
  Login to the edge node as root and run the following command to set the required environment variables:

  ```
  export MAPR_CLDB_HOSTS="MapR's CLDB IP addresses, i.e. 10.0.0.1,10.0.0.2,10.0.0.3"
  export CL_NAME="MapR cluster name, i.e. my.cluster.com"
  export HOST_IP="Your Edge node IP address in the same subnet as the MapR cluster's, i.e. 10.0.0.10"
  ```

  Now grab the scripts from Github

  ```
  wget https://raw.githubusercontent.com/maprpartners/lenovo-demo/master/1-run.producer
  wget https://raw.githubusercontent.com/maprpartners/lenovo-demo/master/2-run.consumer
  ```

  Edit 1-run.producer, fill in the Twitter tokens from earlier and the tweet keywords you are most interested in.



```
CK="Your Consumer Key"
CS="Your Consumer Secret"
AT="Your Access Token"
AS="Your Access Token Secret"

echo "Starting producer..."
docker run -d -it --name maprc-producer \
--cap-add SYS_ADMIN \
--cap-add SYS_RESOURCE \
--security-opt apparmor:unconfined \
--memory 0 \
--restart always \
-e MAPR_CLDB_HOSTS="$MAPR_CLDB_HOSTS" \
-e HOST_IP="$HOST_IP" \
-e MAPR_CLUSTER="$CL_NAME" \
-e MAPR_PASSWORD="mapr" \
-e CONSUMER_KEY="$CK" \
-e CONSUMER_SECRET="$CS" \
-e ACCESS_TOKEN="$AT" \
-e ACCESS_SECRET="$AS" \
-e KEYWORD_FILTER="['Lenovo','tax','healthcare','korea','mapr','tableau','hadoop','big data','bigdata','IoT','zeppelin','artificial intelligence','AI','Azure','AWS','Alexa','data science','data scientist','business intelligence','mapreduce','data warehousing','mahout','hbase','nosql','newsql','machine learning','cloudcomputing']" \
-v /opt/mapr/conf/ssl_truststore:/opt/mapr/conf/ssl_truststore:ro \
maprpartners/maprc-producer:latest
```

Before we can launch the containers, we need to copy the MapR ticket from the cluster to edge node by issuing the follow two commands:

```
scp <IP of one of the CLDB node>:/opt/mapr/conf/maprusterticket /tmp
chown mapr.mapr /tmp/maprusterticket
```

Now issues these commands to start the producer and consumer containers:

```
sh 1-run.producer
sh 2-run.consumer
```

- Start the Data Science Refinery container that will launch Zeppelin notebook for analytics visualization

  Download the script from Github

  wget https://raw.githubusercontent.com/maprpartners/lenovo-demo/master/3-run.dsr

  Start the container:

  sh 3-run.dsr

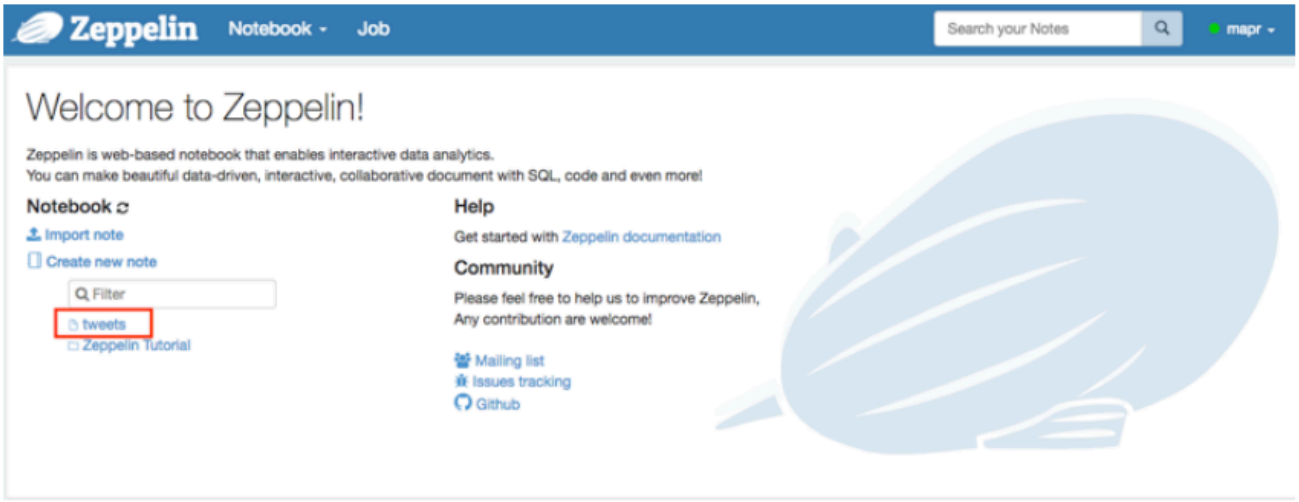
- Configure Zeppelin
  Issue these commands to configure Zeppelin

  wget wget https://raw.githubusercontent.com/maprpartners/lenovo-demo/master/4-config_zeppelin
  sh 4-config_zeppelin

The raw tweets should have started flooding in at this point if everything goes smoothly. Issue this command to see the raw tweets:

docker logs –f maprc-producer

- Using Zeppelin notebook to analyze/visualize the tweets
  Point your browser to the edge node's IP address at port 9995; e.g. https://10.0.0.10:9995; login as user mapr; password as mapr. The "tweets" notebook is shown as below and can be selected.



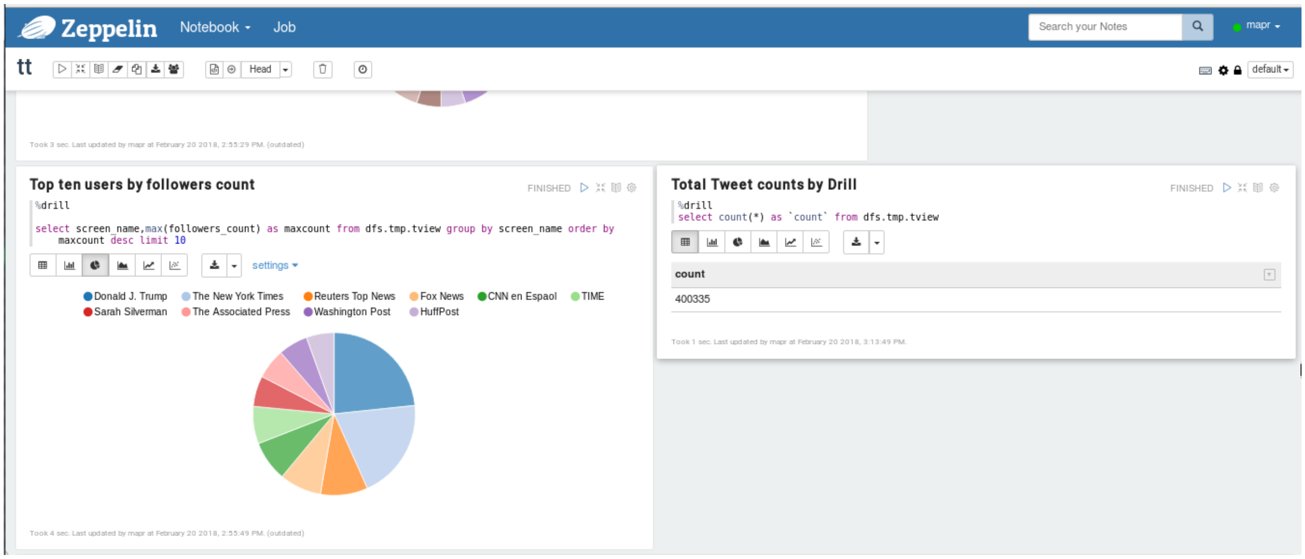
Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

Get started with Zeppelin documentation

Please feel free to help us to improve Zeppelin,
Any contribution are welcome!

In the tweets notebook, the following charts with the corresponding Data Science Refinery and Drill commands show that over 400,000 tweets were ingested and stored in MapR-DB for analysis. The Top Ten twitter users measured by their follower count (followers_count) is analyzed from the stored data and presented as a pie chart.



Took 3 sec. Last updated by mapr at February 20 2018, 2:55:29 PM. (outdated)

Top ten users by followers count

```
%drill

select screen_name,max(followers_count) as maxcount from dfs.tmp.tview group by screen_name order by
    maxcount desc limit 10
```

Took 4 sec. Last updated by mapr at February 20 2018, 2:55:49 PM. (outdated)

Total Tweet counts by Drill

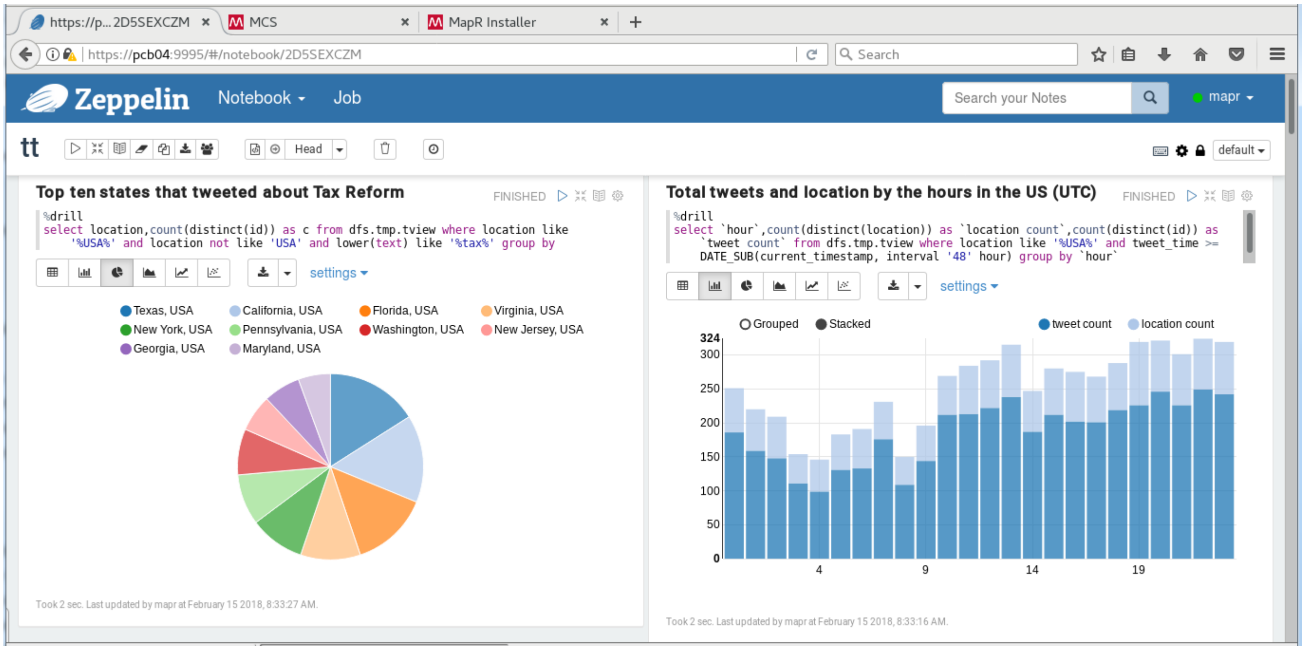
```
%drill
select count(*) as `count` from dfs.tmp.tview
```

count
400335

Took 1 sec. Last updated by mapr at February 20 2018, 3:13:49 PM.

Figure 1: Data Science Refinary Drill analysis and total tweet count

In this example analysis, tweets from top ten states with subject regarding tax reform were queried and presented in a pie chart format. In addition, the total tweet count by hour for this query is presented in the hourly bar chart.



Top ten states that tweeted about Tax Reform

```
%drill
select location,count(distinct(id)) as c from dfs.tmp.tview where location like
    '%USA%' and location not like 'USA' and lower(text) like '%tax%' group by
```

Took 2 sec. Last updated by mapr at February 15 2018, 8:33:27 AM.

Total tweets and location by the hours in the US (UTC)

```
%drill
select `hour`,count(distinct(location)) as `location count`,count(distinct(id)) as
    `tweet count` from dfs.tmp.tview where location like '%USA%' and tweet_time >=
    DATE_SUB(current_timestamp, interval '48' hour) group by `hour`
```

Took 2 sec. Last updated by mapr at February 15 2018, 8:33:16 AM.

Figure 2: Data Science Refinery - Drill query and count by hour

To capture additional tweets with different sets of keywords, the 1-run.producer script can be modified to launch several producers at the same time for a multi-stream ingest and higher data ingestion rate. It is very easy to scale up and down the container deployment to fit the analytical requirements.

The Zeppelin notebook is a primary tool for teams of data scientists to share and use various tools (i.e. Drill, Spark, and Hive) to extract data from a MapR cluster, then perform and visualize the data analysis with Data Science Refinery. For more information about MapR Data Science Refinery, refer to this URL: https://mapr.com/products/data-science-refinery/

If you are also interested in deploying this demo in the cloud, please refer to this MapR blog: https://mapr.com/blog/real-time-twitter-analytics-with-mapr-data-science-refinery-clouds/