Demo Setup for Calling a Spark Machine Learning Model With IBM Db2 Big SQL

IBM NA Big Data Science Technical Team

December 2017

Contents

Intro Demo Setup for Calling a Spark Machine Learning Model with IBM Db2 Big SQL

IBM Db2 Big SQL (Big SQL) is a SQL engine on Hadoop that has been making strides with the fast-evolving open source ecosystem. The core capabilities of Big SQL focusses on federation, SQL compatibility, scalability, performance, and of course enterprise security, making it a desirable query engine to seek insights from disparate data sources including Hadoop.

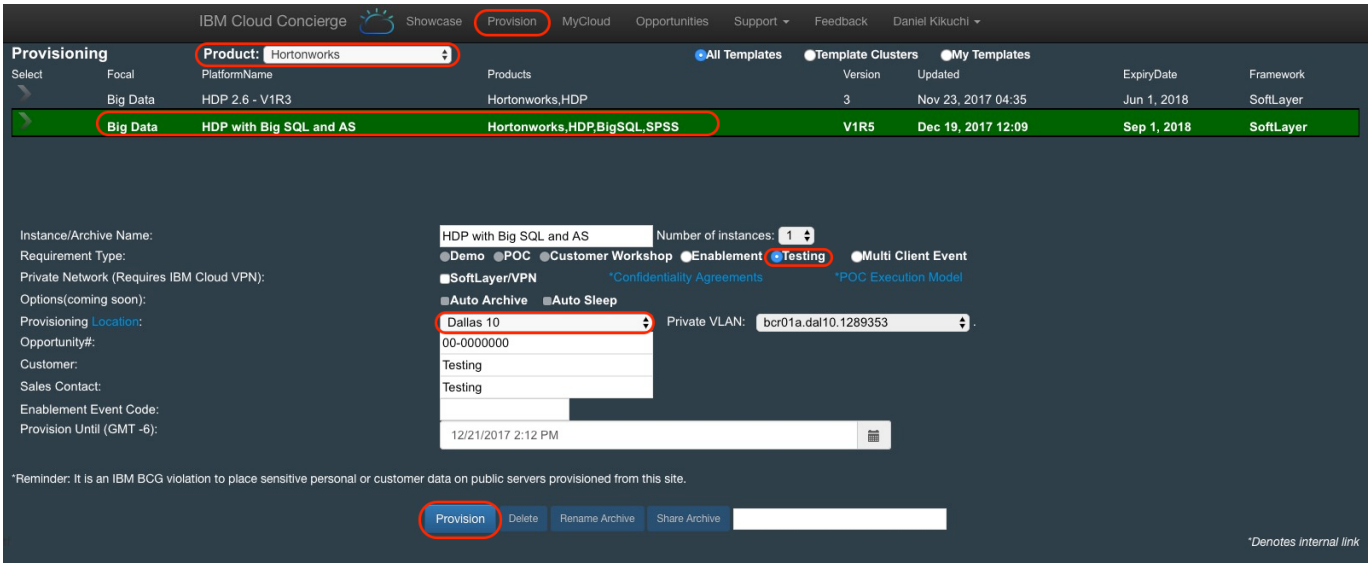
In addition, Big SQL provides tight integration with Apache Spark featuring efficient synthesis between Spark executors and Big SQL workers.

This tutorial will focus on the steps to demonstrate Big SQL calling a Spark machine learning model. This will be using the Cloud Concierge template for the Hortonworks Data Platform provided by Frank Ketelaars.

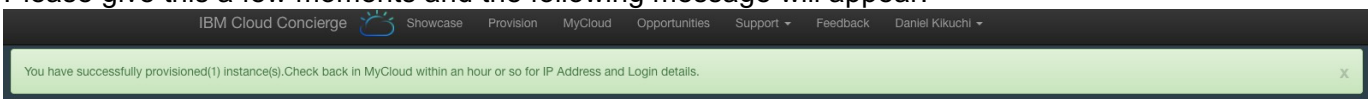
Part 1 Provision Cloud Concierge Instances

This part of the tutorial will provision two Cloud Concierge instances. One will be the Hortonworks Data Platform (HDP) with Big SQL instance and the other will be a base CentOS instance that will be used as a 'jump host'. This jump host will be allow us to ssh into the HDP image since the ssh port 22 for HDP is closed on the public interface.

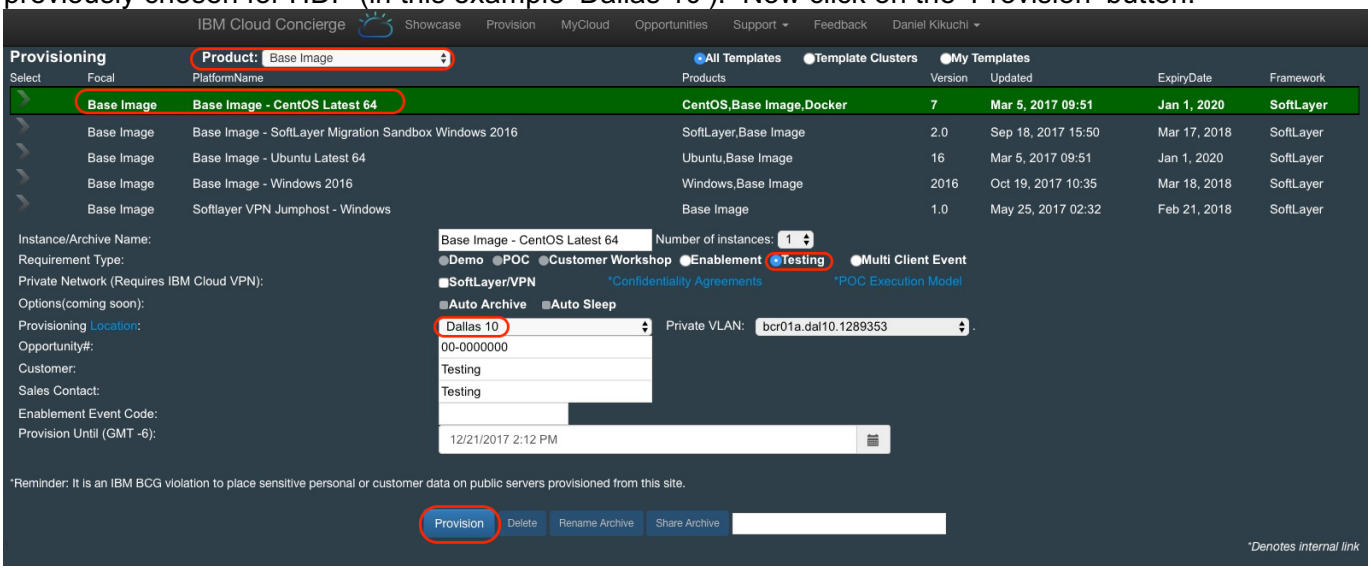
a) Open your browser and navigate to http://demo.ibmcloud.com to access IBM Cloud Concierge. Once logged in, click on 'Provision' from the top tab choices and then select 'Hortonworks' from the 'Product' pull down. Please pick the 'HDP with Big SQL and AS' template. Select the 'Requirement Type' that applies (in this example, 'Testing') and select a 'Provisioning Location' (in this example 'Dallas 10'). Please remember this 'Provisioning Location' for the next step. Now click on the 'Provision' button at the bottom.



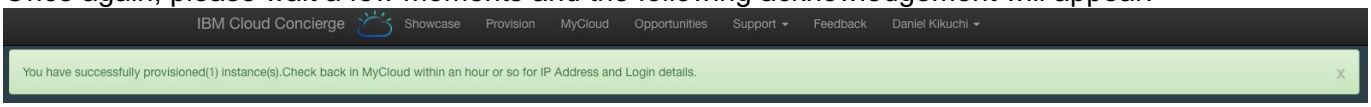
b) Please give this a few moments and the following message will appear.



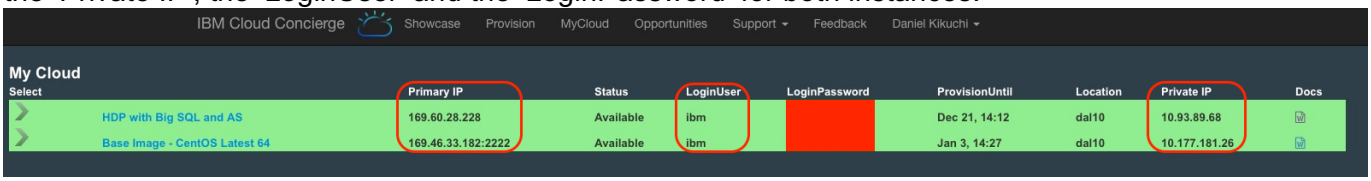
c) While that is being provisioned, let's also provision the jump host. Select 'Base Image' from the 'Product' pull down. Select the 'Base Image – CentOS Latest 64' template. Select the 'Requirement Type' that applies (in this example 'Testing'). Select the same 'Provisioning Location' that was previously chosen for HDP (in this example 'Dallas 10'). Now click on the 'Provision' button.



d) Once again, please wait a few moments and the following acknowledgement will appear.



e) As indicated, give this time to provision. When it is provisioned, you will see the following under the 'MyCloud' tab choices on the top of the page. Please note and remember the 'Primary IP' (public), the 'Private IP', the 'LoginUser' and the 'LoginPassword' for both instances.



Part 2 Upload Data into HDFS and Create Big SQL Tables

This part of the tutorial will upload the data to HDP HDFS and create Big SQL tables.

a) ssh login to the HDP instance using the jump host. To do this, the following command will first ssh login into the jump host followed by an ssh login to HDP. Please accept the authenticity fingerprints and enter the passwords when prompted.

ssh ibm@<your_base_centos_public_ip> -p 2222 -t ssh ibm@<your_HDP_private_ip>

In this example, 'ssh ibm@169.46.33.182 -p 2222 -t ssh ibm@10.93.89.68' was used. You are now logged into HDP as user 'ibm'.

```
Daniels-MacBook-Pro:~ dkikuchi$ ssh ibm@169.46.33.182 -p 2222 -t ssh ibm@10.93.89.68
The authenticity of host '[169.46.33.182]:2222 ([169.46.33.182]:2222)' can't be established.
ECDSA key fingerprint is SHA256:VUjRR//lxuenq9yzta0Q3twVri/OGgAE43d4Q3we8Wk.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '[169.46.33.182]:2222' (ECDSA) to the list of known hosts.
ibm@169.46.33.182's password:
The authenticity of host '10.93.89.68 (10.93.89.68)' can't be established.
ECDSA key fingerprint is SHA256:yqe+f+MIpIJD8g1j0sqAwRE7kPiTFPsrJ8Aq+OHIiEg.
ECDSA key fingerprint is MD5:9b:35:a6:71:50:22:b0:74:e7:a9:b4:32:8a:6c:e0:19.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '10.93.89.68' (ECDSA) to the list of known hosts.
ibm@10.93.89.68's password:
Last login: Tue Dec 19 08:40:34 2017 from 10.3.220.74
[ibm@hdp ~]$
```

b) Switch to the user 'bigsql' by issuing the following command.

su - bigsql

Use the same password as before.
Issue the following commands to retrieve the 2 data files.

wget https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/GoSales_Tx_20.csv
wget https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/GoSales_Tx_80.csv

```
[ibm@hdp ~]$ su - bigsql
Password:
Last login: Wed Dec 20 15:02:36 CST 2017
[bigsql@hdp ~]$ wget https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/GoSales_Tx_20.csv
--2017-12-20 15:05:51--  https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/GoSales_Tx_20.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.48.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.48.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 501580 (490K) [text/plain]
Saving to: 'GoSales_Tx_20.csv'

100%[======================================================================================================================================================>] 501,580     --.-K/s   in 0.02s

2017-12-20 15:05:52 (26.7 MB/s) - 'GoSales_Tx_20.csv' saved [501580/501580]

[bigsql@hdp ~]$ wget https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/GoSales_Tx_80.csv
--2017-12-20 15:05:54--  https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/GoSales_Tx_80.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.48.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.48.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2009012 (1.9M) [text/plain]
Saving to: 'GoSales_Tx_80.csv'

100%[======================================================================================================================================================>] 2,009,012   --.-K/s   in 0.05s

2017-12-20 15:05:54 (35.3 MB/s) - 'GoSales_Tx_80.csv' saved [2009012/2009012]

[bigsql@hdp ~]$
```

c) Create folders on HDP HDFS for each of these files by issuing the following commands.

hdfs dfs -mkdir GoSales20
hdfs dfs -mkdir GoSales80

And copy the data files to these folders by issuing the following commands.

hdfs dfs -put GoSales_Tx_20.csv GoSales20
hdfs dfs -put GoSales_Tx_80.csv GoSales80

```
[bigsql@hdp ~]$ hdfs dfs -mkdir GoSales20
[bigsql@hdp ~]$ hdfs dfs -mkdir GoSales80
[bigsql@hdp ~]$ hdfs dfs -put GoSales_Tx_20.csv GoSales20
[bigsql@hdp ~]$ hdfs dfs -put GoSales_Tx_80.csv GoSales80
[bigsql@hdp ~]$ hdfs dfs -ls GoSales20
Found 1 items
-rw-r--r--   1 bigsql hdfs     501580 2017-12-20 15:13 GoSales20/GoSales_Tx_20.csv
[bigsql@hdp ~]$ hdfs dfs -ls GoSales80
Found 1 items
-rw-r--r--   1 bigsql hdfs    2009012 2017-12-20 15:13 GoSales80/GoSales_Tx_80.csv
[bigsql@hdp ~]$
```

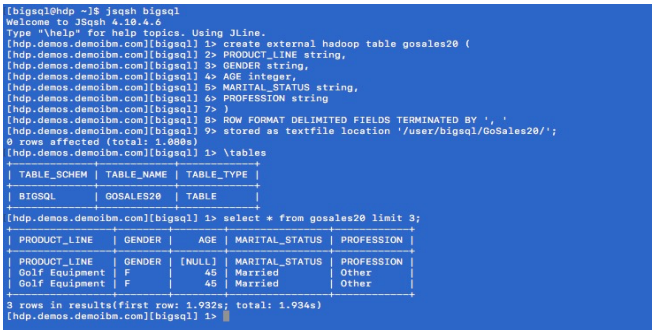
d) Start ‘jsqsh’ which is the Java SQL Shell command interface by issuing the command

```
jsqsh bigsql
```

Note that HDP and Big SQL may not be entirely initialized. To see the current status, issue ‘bigsql status’ and if needed, please wait until Big SQL is available and running.

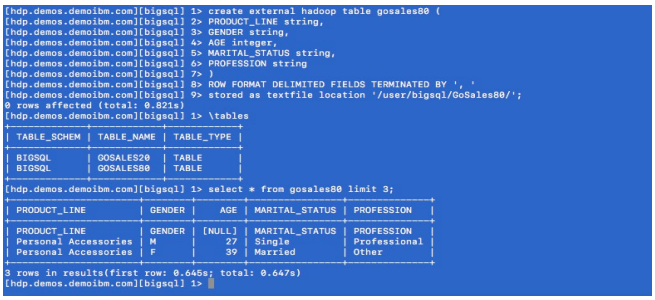
e) Create the ‘gosales20’ table by issuing the following command in the ‘jsqsh’ shell.

```
create external hadoop table gosales20 (
PRODUCT_LINE string,
GENDER string,
AGE integer,
MARITAL_STATUS string,
PROFESSION string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ', '
stored as textfile location '/user/bigsql/GoSales20/';
```



f) Do the same for the ‘gosales80’ table.

```
create external hadoop table gosales80 (
PRODUCT_LINE string,
GENDER string,
AGE integer,
MARITAL_STATUS string,
PROFESSION string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ', '
stored as textfile location '/user/bigsql/GoSales80/';
```



g) The data has been uploaded, inserted into HDP HDFS and Big SQL tables have been created.

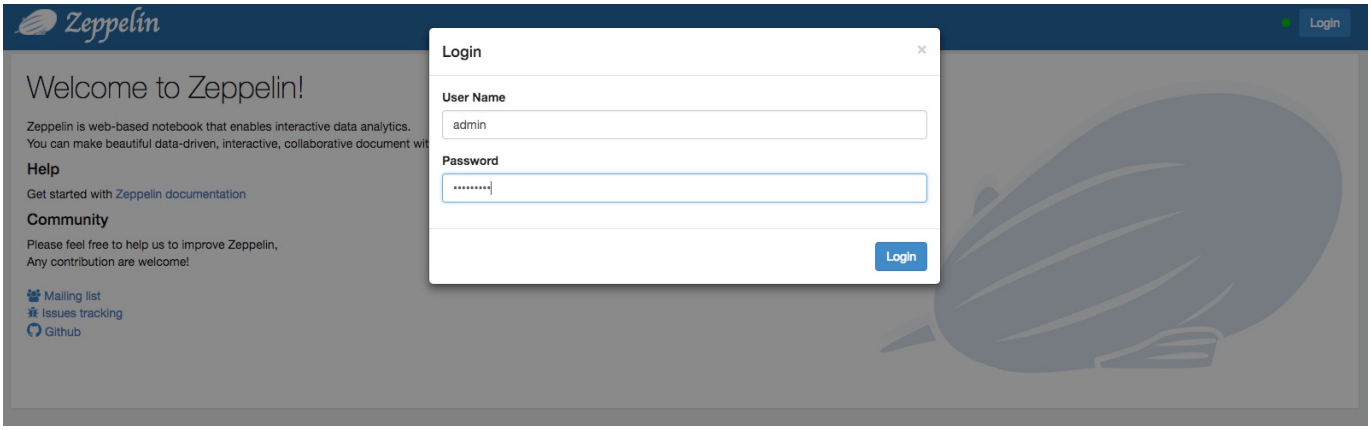
Part 3 Train and Save a Spark Machine Learning Model

This part of the tutorial will use a Zeppelin notebook to train and save a Spark Machine Learning model using the data in Part 2.

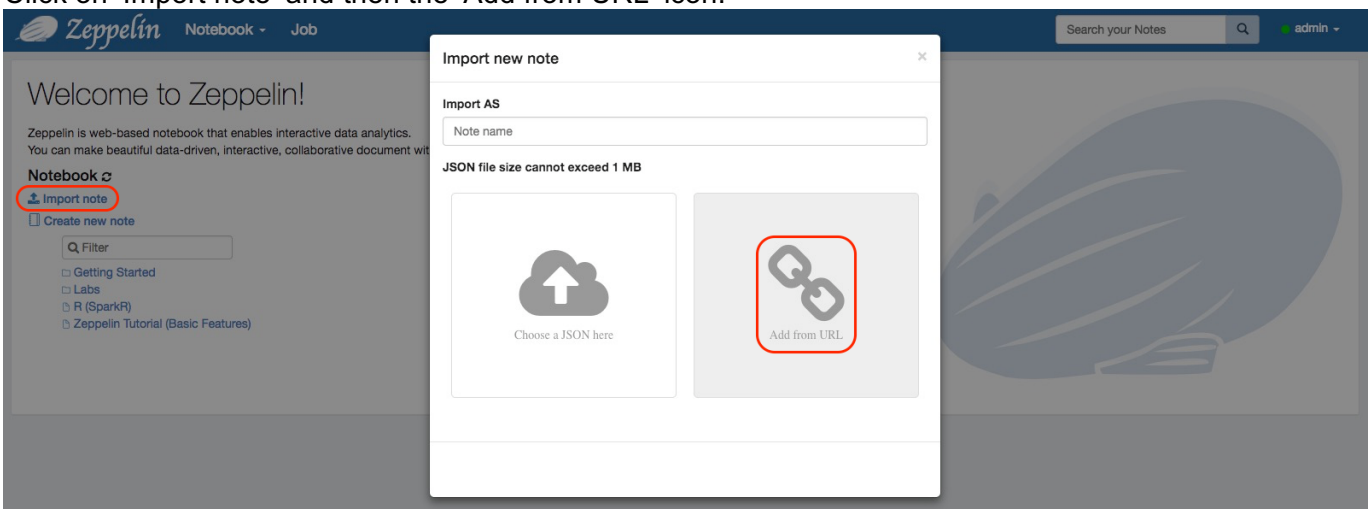
a) Login to Zeppelin on HDP by going to a browser and entering the following url.

https://<your_HDP_public_ip>:9995

In this example, https://169.60.28.228:9995 was used.
Please enter a 'User Name' of 'admin' and the same password as before.

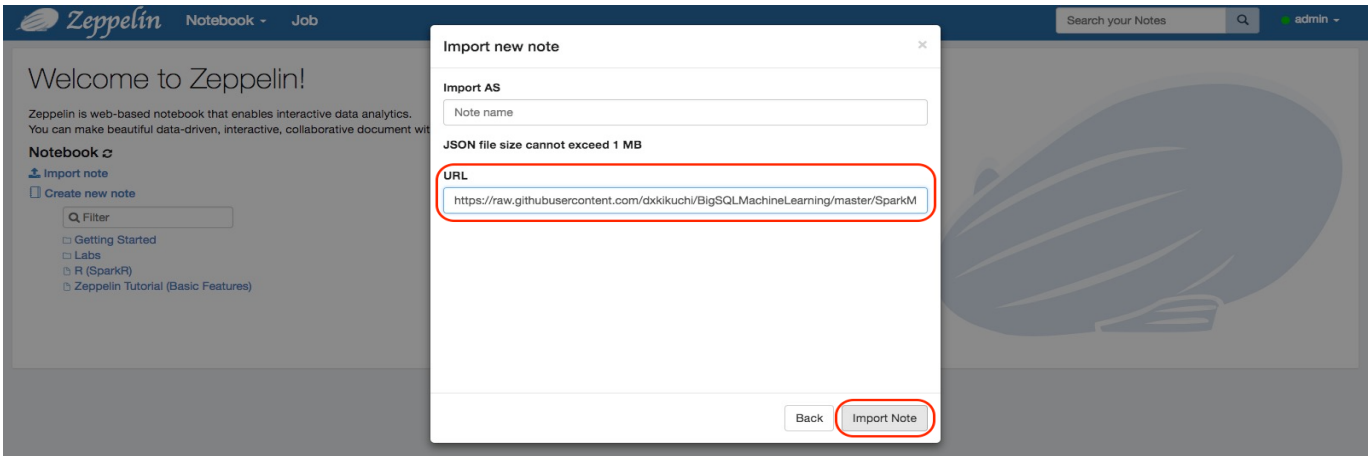


b) Click on 'Import note' and then the 'Add from URL' icon.

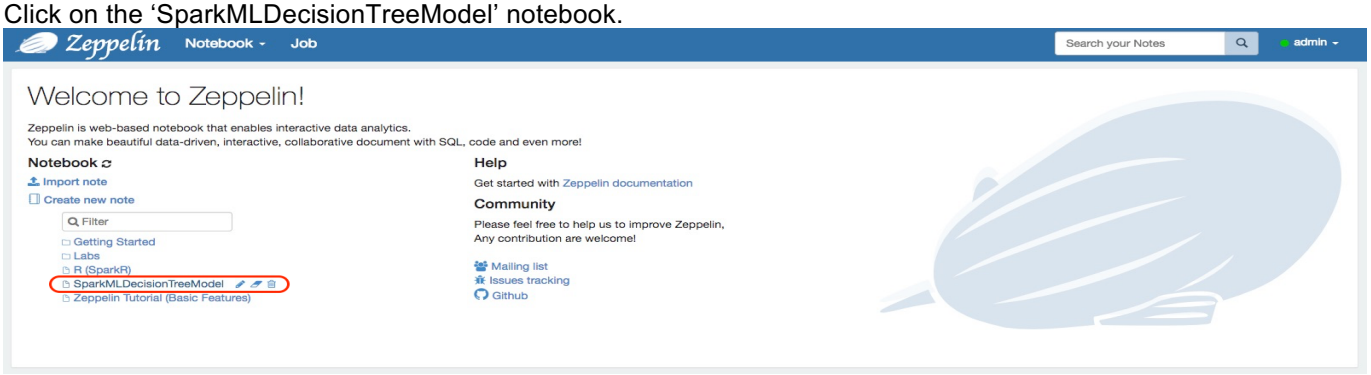


c) Enter the following for ‘URL’ and click on ‘Import Note’.

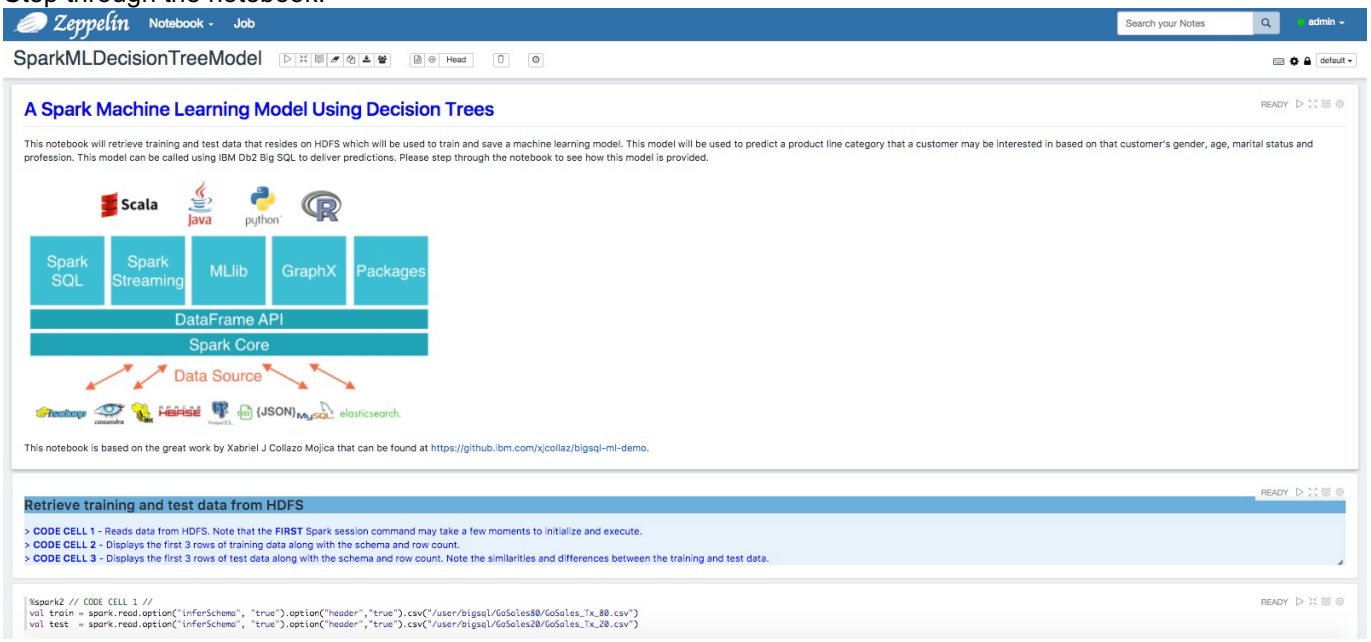
https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/SparkMLDecisionTreeModel.json



d) Click on the ‘SparkMLDecisionTreeModel’ notebook.



e) Step through the notebook.



f) This notebook will retrieve the training and test data that resides on HDP HDFS which will be used to train and save a machine learning model. This model will be used to predict a product line category that a customer may be interested in base on that customer's gender, age, marital status and profession. This model has been saved in HDP HDFS and will be called upon by IBM Db2 Big SQL to deliver predictions.

Part 4 Installing a PTF and Starting the Big SQL Spark Gateway

This part of the tutorial will install a PTF (Polymorphic Table Function) and start the Big SQL Spark Gateway. More detailed information on this can be found at https://www.ibm.com/support/knowledgecenter/SSCRJT_5.0.1/com.ibm.swg.im.bigsql.dev.doc/doc/biga_spark_int.html

a) From the HDP operating system prompt, retrieve the PTF using the following command.
wget https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/QueryMLModelPTF.jar

```
[bigsql@hdp ~]$ wget https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/QueryMLModelPTF.jar
--2017-12-20 16:05:35--  https://raw.githubusercontent.com/dxkikuchi/BigSQLMachineLearning/master/QueryMLModelPTF.jar
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.48.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.48.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 26224 (26K) [application/octet-stream]
Saving to: 'QueryMLModelPTF.jar'

100%[==========================================================================================================================================================>] 26,224      --.-K/s   in 0.002s

2017-12-20 16:05:36 (12.4 MB/s) - 'QueryMLModelPTF.jar' saved [26224/26224]

[bigsql@hdp ~]$
```

b) Access jsqsh using the following command.

jsqsh bigsql

c) Issue the following command in the jsqsh shell to install the PTF.

\install-jar --path=/home/bigsql/QueryMLModelPTF.jar --id myptfs

```
[bigsql@hdp ~]$ jsqsh bigsql
Welcome to JSqsh 4.10.4.6
Type "\help" for help topics. Using JLine.
[hdp.demos.demoibm.com][bigsql] 1> \install-jar --path=/home/bigsql/QueryMLModelPTF.jar --id myptfs
[hdp.demos.demoibm.com][bigsql] 1>
```

d) Exit out of jsqsh. Copy the bigsql-spark.conf template and add the PTF jar by issuing the following commands.

cp /usr/ibmpacks/bigsql/5.0.2.0/bigsql/conf/templates/bigsql-spark.conf /usr/ibmpacks/bigsql/5.0.2.0/bigsql/conf/bigsql-spark.conf
vi /usr/ibmpacks/bigsql/5.0.2.0/bigsql/conf/bigsql-spark.conf

Add /home/bigsql/QueryMLModelPTF.jar to the 'spark.driver.extraClassPath' and 'spark.executor.extraClassPath'. The file should look like the following. Save the file.

```
spark.executor.instances 2
spark.executor.memory 2g
spark.driver.memory 2g
spark.executor.cores 4
spark.driver.extraClassPath /usr/ibmpacks/current/bigsql/bigsql/userlib/ptf-examples.jar:/home/bigsql/QueryMLModelPTF.jar
spark.executor.extraClassPath /usr/ibmpacks/current/bigsql/bigsql/userlib/ptf-examples.jar:/home/bigsql/QueryMLModelPTF.jar
bigsql.spark.jars.regex .*
bigsql.spark.table.input.engine bigsql
bigsql.spark.check.uri.access true
bigsql.spark.default.package com.ibm.biginsights.bigsql.examples
bigsql.spark.string.size 1024
```

e) Stop and restart bigsql and bigsql spark.

```
bigsql stop -spark
bigsql stop
bigsql start
bigsql start -spark
```

```
[bigsql@hdp ~]$ bigsql stop -spark
Stopping Big SQL Spark gateway : OK
[bigsql@hdp ~]$ bigsql stop
Stopping Big SQL               : OK
Stopping Big SQL Scheduler     : OK
[bigsql@hdp ~]$ bigsql start
Global config update           : OK
Starting Big SQL Scheduler     : OK
Starting Big SQL               : OK
[bigsql@hdp ~]$ bigsql start -spark
Starting Big SQL Spark gateway : OK
[bigsql@hdp ~]$
```

f) The PTF has been installed and the Big SQL Spark Gateway has been started.

Part 5 Big SQL Calling Spark Machine Learning Model

This part of the tutorial will use Big SQL to call the model and provide predictions using 5 rows of the gosales20 Big SQL table.

In the jsqsh shell, enter the following command.

```
SELECT model.* FROM TABLE(SYSHADOOP.EXECSPARK(
  class => 'com.ibm.biginsights.bigsql.examples.QueryMLModel',
  model_uri => CAST('/user/zeppelin/CrossValidatedDecisionTree.sparkmodel' as SYSHADOOP.URI),
  feature_names => 'PRODUCT_LINE,GENDER,AGE,MARITAL_STATUS,PROFESSION',
  feature_values => TABLE(SELECT * FROM gosales20 where age is not null) as in,
  label_name => 'predictedLabel',
  label_type => 'string')) AS model limit 5;
```

```
[bigsql@hdp conf]$ jsqsh bigsql
Welcome to JSqsh 4.10.4.6
Type "\help" for help topics. Using JLine.
[hdp.demos.demoibm.com][bigsql] 1> SELECT model.* FROM TABLE(SYSHADOOP.EXECSPARK(
[hdp.demos.demoibm.com][bigsql] 2>    class => 'com.ibm.biginsights.bigsql.examples.QueryMLModel',
[hdp.demos.demoibm.com][bigsql] 3>    model_uri => CAST('/user/zeppelin/CrossValidatedDecisionTree.sparkmodel' as SYSHADOOP.URI),
[hdp.demos.demoibm.com][bigsql] 4>    feature_names => 'PRODUCT_LINE,GENDER,AGE,MARITAL_STATUS,PROFESSION',
[hdp.demos.demoibm.com][bigsql] 5>    feature_values => TABLE(SELECT * FROM gosales20 where age is not null) as in,
[hdp.demos.demoibm.com][bigsql] 6>    label_name => 'predictedLabel',
[hdp.demos.demoibm.com][bigsql] 7>    label_type => 'string')) AS model limit 5;
+-------------------------+--------+-----+----------------+--------------+----------------------+--------------------------------------------------------------------------------------------------------+
| PRODUCT_LINE            | GENDER | AGE | MARITAL_STATUS | PROFESSION   | PREDICTEDLABEL       | PROBABILITY                                                                                            |
+-------------------------+--------+-----+----------------+--------------+----------------------+--------------------------------------------------------------------------------------------------------+
| Golf Equipment          | F      |  45 | Married        | Other        | Camping Equipment    | [0.35398230088495575,0.3392330383480826,0.02654867256637168,0.20648967551622419,0.07374631268436578]   |
| Golf Equipment          | F      |  45 | Married        | Other        | Camping Equipment    | [0.35398230088495575,0.3392330383480826,0.02654867256637168,0.20648967551622419,0.07374631268436578]   |
| Personal Accessories    | F      |  32 | Single         | Professional | Personal Accessories | [0.027237354085603113,0.9455252918287937,0.019455252918287938,0.007782101167315175,0.0]                |
| Personal Accessories    | F      |  33 | Single         | Other        | Personal Accessories | [0.031413612565445025,0.9319371727748691,0.015706806282722512,0.005235602094240838,0.015706806282722512] |
| Mountaineering Equipment | M     |  23 | Single         | Professional | Camping Equipment    | [0.5662164886111671,0.1505744809514211,0.2541826244708728,0.01431163071961298,0.014714775246926022]     |
+-------------------------+--------+-----+----------------+--------------+----------------------+--------------------------------------------------------------------------------------------------------+
5 rows in results(first row: 4.093s; total: 4.099s)
[hdp.demos.demoibm.com][bigsql] 1>
```

Congratulations, you have completed this exercise.

Great work and congratulations, you have completed this exercise.

Screen captures in this document may vary slightly from yours.

NOTES

Please Recycle