

ECV-BD-310

Build recommendation engine with Amazon Machine Learning in Redshift

2016.11.03

Version 1.3

Agenda

Introduction	4
Overview	4
Topics covered	4
Redshift.....	4
What is Redshift?.....	4
Amazon Machine Learning.....	5
What is Amazon Machine Learning?.....	5
About this lab	6
Scenario	6
Architecture Diagram	6
Pre-request for this lab	7
Create a Security Group.....	7
Create a IAM for the credential	7
Add permission into Users.....	7
Create Redshift cluster	8
Create a Amazon Redshift Cluster	8
Setup EC2 environment for SQL WorkbenchJ	9
Setting EC2 environment.....	9
Connecting to your Amazon EC2 instances	10
RDP Connection for Linux and Mac Users:.....	10
RDP Connection for Windows Users:	10
Upload a file to S3 bucket	10
Establish a connection with Redshift cluster	11
Connect with SQLWorkbench.....	11
Create table import reference data into Redshift.....	12
Create a table in Redshift via SQLWorkbench	12

Import data in Redshift.....	12
Working with Amazon Machine Learning	13
Create Model via Amazon Machine Learning.....	13
Testing with Amazon Machine Learning	15
End your lab	15
Delete Redshift	15
Delete S3	15
Delete EC2	16
Delete Amazon Machine Learning	16

Introduction

Overview

In this lab, you will build a smart solution using Amazon Redshift and Amazon Machine Learning that predict rental bikes for Capital bikeshare system.

The dataset contains daily amount of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

You will learn how to use Redshift and predict using Machine Learning to create a model that will predict the rental bikes.

Topics covered

By the end of this lab, you will be able to:

- Create a Redshift cluster, build Redshift table and load data from Amazon S3.
- Create a Machine Learning Model
- Train the Machine Learning Model, using historic data about rental bikes.
- Predict the rental amount for the future sharebike system with Redshift and Amazon Machine Learning

Redshift

What is Redshift?

Amazon Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in [the cloud](#). Customers can start small for just \$0.25 per hour with no commitments or upfront costs and scale to a petabyte or more for \$1,000 per terabyte per year, less than a tenth of most other data warehousing solutions.

Traditional data warehouses require significant time and resource to administer, especially for large datasets. In addition, the financial cost associated with

building, maintaining, and growing self-managed, on-premise data warehouses is very high. Amazon Redshift not only significantly lowers the cost of a data warehouse, but also makes it easy to analyze large amounts of data very quickly. Amazon Redshift gives you fast querying capabilities over structured data using familiar SQL-based clients and business intelligence (BI) tools using standard ODBC and JDBC connections. Queries are distributed and parallelized across multiple physical resources. You can easily scale an Amazon Redshift data warehouse up or down with a few clicks in the AWS Management Console or with a single API call. Amazon Redshift automatically patches and backs up your data warehouse, storing the backups for a user-defined retention period. Amazon Redshift uses replication and continuous backups to enhance availability and improve data durability and can automatically recover from component and node failures. In addition, Amazon Redshift supports Amazon Virtual Private Cloud (Amazon VPC), SSL, AES-256 encryption and Hardware Security Modules (HSMs) to protect your data in transit and at rest. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use. Amazon Redshift lets you pay as you go.

Amazon Machine Learning

What is Amazon Machine Learning?

Amazon Machine Learning is a machine service that allows you to easily build predictive applications, including fraud detection, demand forecasting, and click prediction. Amazon Machine Learning uses powerful algorithms that can help you create machine learning models by finding patterns in existing data, and using these patterns to make predictions from new data as it becomes available. The AWS Management Console and API provide data and model visualization tools, as well as wizards to guide you through the process of creating machine learning models, measuring their quality and fine-tuning the predictions to

match your application requirements. Once the models are created, you can get predictions for your application by using the simple API, without having to implement custom prediction generation code or manage any infrastructure. Amazon Machine Learning is highly scalable and can generate billions of predictions, and serve those predictions in real-time and at high throughput. With Amazon Machine Learning there is no setup cost and you pay as you go, so you can start small and scale as your application grows.

The workshop's region will be in N. Virginia

About this lab

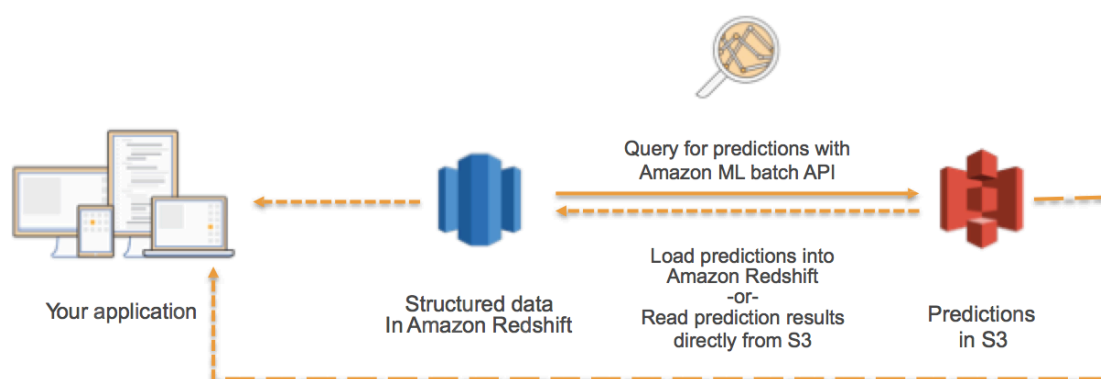
Scenario

The dataset contains daily amount of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

All we want to know is how much bikes we should prepare for the next week. To avoid the situation when the supply could not meet the demand.

Architecture Diagram

We used ec2 for import data into Redshift. Then, we used Amazon Machine Learning for training model and prediction. All of the output will be stored into S3.



Prerequisite for this lab

Create a Security Group

On the service menu, click 'EC2'

On the left panel, click 'Security Group'

Click 'Create Security Group'

For Security group name, type 'workshop-sg'

For Description, type 'workshop -sg'

For VPC, choose 'default'

For Inbound, click 'Add Rule'

For Inbound type, choose 'All traffic'

For Inbound Source, choose 'Anywhere'

Click 'Create'

Create a IAM for the credential

On the service menu, click 'IAM'

On the left panel, click 'Users'

Click 'Create New Users'

For Enter user name, type 'workshop-user'

Select 'Generate an access key for each user'

Click 'Create'

Click 'Download Credential'

When you downloaded the credential, then click 'Close'

Add permission into Users

On the left panel, click 'Users'

Click the user you created 'workshop-user'

For permissions, click 'Attach Policy'

For policy, select 'AdministratorAccess'

Click 'Attach policy'

Create Redshift cluster

Create a Amazon Redshift Cluster

On the service menu, click 'Redshift'

Click 'Launch Cluster'

IN CLUSTER DETAILS PART

For Cluster identifier, type 'redshiftdemo'

For Database name, type 'redshiftdemo'

For Database port, type '5439'

For Master User Name, type 'root'

For Master User Password, type 'Redshift123'

Confirm the password

Click 'Continue'

IN NODE CONFIGURATION PART

For Node Type, select 'dc1.large'

For Cluster Type, select 'Single Node'

For Number of Compute Nodes, type '1'

Click 'Continue'

IN ADDITIONAL CONFIGURATION PART

For Choose a VPC, select 'default VPC'

For VPC Security Groups, choose 'workshop-sg' which you created before

Click 'Continue'

IN REVIEW PART

Click 'Launch Cluster'

For now, you have created a Redshift cluster.

Continue the following guide, you will know how to establish a connection with

Redshift cluster.

Setup EC2 environment for SQL WorkbenchJ

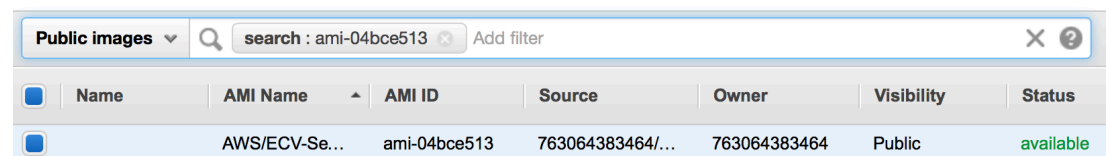
Setting EC2 environment

On the service menu, click 'EC2'

On the left panel, click 'AMIs'

For the filter, choose 'Public images'

For the filter, type 'ami-f4ca93e3', then click 'search'



	Name	AMI Name	AMI ID	Source	Owner	Visibility	Status
<input checked="" type="checkbox"/>		AWS/ECV-Se...	ami-04bce513	763064383464/...	763064383464	Public	available

Choose images, then click 'Launch'

For the instance type, choose 't2.small'

Click 'Next: Configure Instance Detail'

For network, keep the setting as default

Click 'Next: Add storage'

For storage, keep the setting as default

Click 'Next: Tag Instance'

For tag instance, type values as 'Workshop'

Click 'Next: Configure Security Group'

For configure Security Group, Select an existing security group: workshop-sg

Click 'Review and Launch'

For selecting an existing key pair part, select 'proceed without key'

Select yes about 'I acknowledge that I have access to the selected private key file (virginia-testing-ruoen.pem), and that without this file, I won't be able to log into my instance.'

Click 'Launch'

Connecting to your Amazon EC2 instances

RDP Connection for Linux and Mac Users:

If you are running Mac OS X, you can download the CoRD RDS client:

<http://cord.sourceforge.net/>

Open your RDP Client

For Server, paste in the Public DNS of EC2 in the blank

Click 'connect'

For Username, type 'Administrator'

For Password, type '5m@;xE3c9iB'

RDP Connection for Windows Users:

Open your RDP client

For Server, paste in the Public DNS you copied with EC2

Click Connect

For Username, type 'Administrator'

For Password, type '5m@;xE3c9iB'

Click OK

Click Yes to connect when you may see a security connection warning

After you login into EC2 which generated from AMI, we can use SQL Workbench as a connection tool in EC2. Continue the lab you will learn how to use SQL Workbench to connect to Redshift and import data into Redshift..

Before connect with SQL Workbench, we need to upload dataset into s3 bucket. Please following the guide to continue the lab.

Upload a file to S3 bucket

On the service menu, click 'S3'

Click 'Create Bucket'

For Bucket Name, type 'Unique Name'

For Region, choose 'US Standard

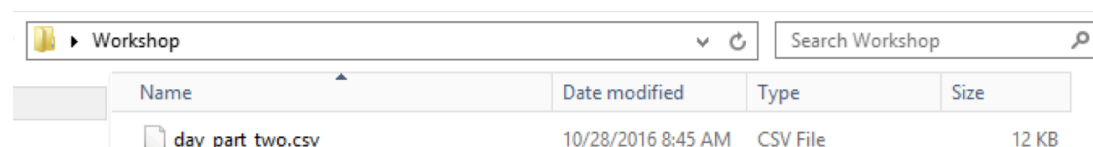
Click 'Create'

Select the bucket which you created before

Click 'Upload'

Click 'Add files'

Click the csv files which in the 'workshop folder'



Name	Date modified	Type	Size
day_part_two.csv	10/28/2016 8:45 AM	CSV File	12 KB

*workshop folder is in the ec2's desktop

Click 'Start Upload'

Establish a connection with Redshift cluster

Connect with SQLWorkbench

Click SQLWorkbenchJ which in the 'workshop folder'



Name	Date modified	Type	Size
SQLWorkbench64 - Shortcut	10/28/2016 8:33 AM	Shortcut	1 KB

About 'select connection profile':

For URL, type 'jdbc:redshift://REDSHIFT_ENDPOINT:5439/redshiftdemo'

*Redshoft endpoint/Database name showed on AWS redshift console

For Username, type the username which you created in the AWS console

For Password, type the password which you created in the AWS console

Select 'AutoCommit'

Click 'OK'

Congrats! You have established a connection with Redshift cluster.

Continue the following guide. You will create a table and import reference data into Redshift

Create table import reference data into Redshift

Create a table in Redshift via SQLWorkbench

Create a table in Redshift, type the following command into SQL WorkBench's statement

**Please do not copy the command in word, please copy the command in 'sql txt file' which is stored into workshop folder.*

```
create table bike(  
season varchar(10),  
mnth varchar(10),  
weekday varchar(10),  
workingday varchar(10),  
weathersit varchar(10),  
cnt int  
);
```

Then you will see 'Table bike created' in SQL Workbench's messages.

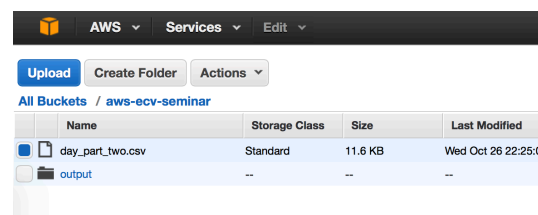
Import data in Redshift

Import bike data in Redshift's table, type the following command into SQLWorkBench's statement

**Please do not copy the command in word, please copy the command in 'sql txt file' which is stored into workshop folder.*

```
copy bike FROM 's3://bucketname/day_part_two.csv' credentials  
'aws_access_key_id=YOUR-ACCESS-KEY-ID;aws_secret_access_key=YOUR-SECRET-ACCESS-KEY' csv IGNOREHEADER 1;
```

**Please noticed that bucket name should be copy from your own AWS account's s3 bucket. For example my bucket name is 'aws-ecv-seminar':*



After submit the SQL commend, you will see 'Warnings: Load into table 'bike' completed, 731 record(s) loaded successfully. 0 rows affected. COPY executed successfully' in SQL Workbench's message.

Congrats! You have completed the Redshift part as a data source. Let's move to next section to learn how to use Amazon Machine Learning.

For next section, you will learn how to use Amazon Machine Learning (AML) as a tool for building a model and use AML's model to predict the business decision.

Working with Amazon Machine Learning

Create Model via Amazon Machine Learning

On the service menu, click 'Machine Learning'.

Click 'Get started'.

Click 'Launch'.

For 'where is your data', choose 'Redshift'.

For Cluster identifier, choose the redshift which you created.

For Database name, type database name which you created.

For Database user name, type user name which you created.

For Database password, type password which you created.

For IAM role, select 'Test Access'.

*You will see 'IAM role created. Amazon ML can now access Amazon Redshift.'.

For SQL Query, type 'select * from bike'.

For Amazon S3 staging location, type 'aws-ecv-seminar/output'.

*Find the bucket which you created before, then create a folder which named output.

For Datasource name, type 'aml-ver1'.

Click 'Verify'.

*You will see 'The validation is successful. To go to the next step, choose 'Continue'.

Click 'Continue'.

In Schema part

For Datatype, choose season/mnth/weekday/workingday/weathersit as Catagorical.

For Datatype, choose cnt as Numetric.

Click 'Continue'

In Target part

For target, choose 'cnt' as target for prediction.

Click 'Continue'.

In Row ID part

Click 'Review'.

In Review part

Click 'Finish'.

In ML model settings part

Click 'Review'.

In Review part

Click 'Create ML Model'.

For this moment, you will see the message said ' status: Pending', you can test

this machine learning until the status go to 'completed'.

Testing with Amazon Machine Learning

For Dashboard, click 'ML-model' which AML created.

For the left panel, click 'Try real-time predictions'.

For season, you can type 1 to 4.

For mnth, you can type 1 to 12.

For weekday, you can type 1 to 7.

For workingday, you can type 1 to 2.

For weathersit, you can type 1 to 4.

Then, click 'create prediction', you will see the predict value in the right panel.

End your lab

Delete Redshift

On the service menu, click 'Redshift'.

On the left panel, click 'Clusters'.

Select the cluster which you created before.

Select Cluster button, choose 'delete'.

For create snapshot, choose 'no'.

Select 'I acknowledge that when I delete this cluster, data changes since the most recent manual snapshot will be lost.'

Delete S3

On the service menu, click 'S3'.

Select the bucket which you created before.

Select Action, then click 'Delete Bucket'.

Confirm the action, then delete the bucket.

Delete EC2

On the service menu, click 'EC2'.

On the left panel, click 'Instances'.

Select the instance which you created before.

Select Action button, choose 'Instance state, then select 'Terminate'.

Select 'Yes, Terminate'.

Delete Amazon Machine Learning

On the service menu, click 'Machine Learning'.

Select the items which you created before.

Select 'Actions', then click 'Delete'.

Click 'click'.