

Amazon Web Services

Data Engineering Immersion Day

Bridging Data for AI/ML using Amazon SageMaker

Jun 2019

Table of Contents

[Introduction 2](#_Toc10212787)

[Create Amazon SageMaker Notebook Instance 3](#_Toc10212788)

[Connect the SageMaker Jupyter notebook to Athena 8](#_Toc10212789)

[Working in Pandas 11](#_Toc10212790)

# Introduction

Amazon SageMaker is an end-to-end machine learning platform that lets you build, train, and deploy machine learning models in AWS. It is a highly modular service that lets you use each of these components independently of each other.

You will Learn:

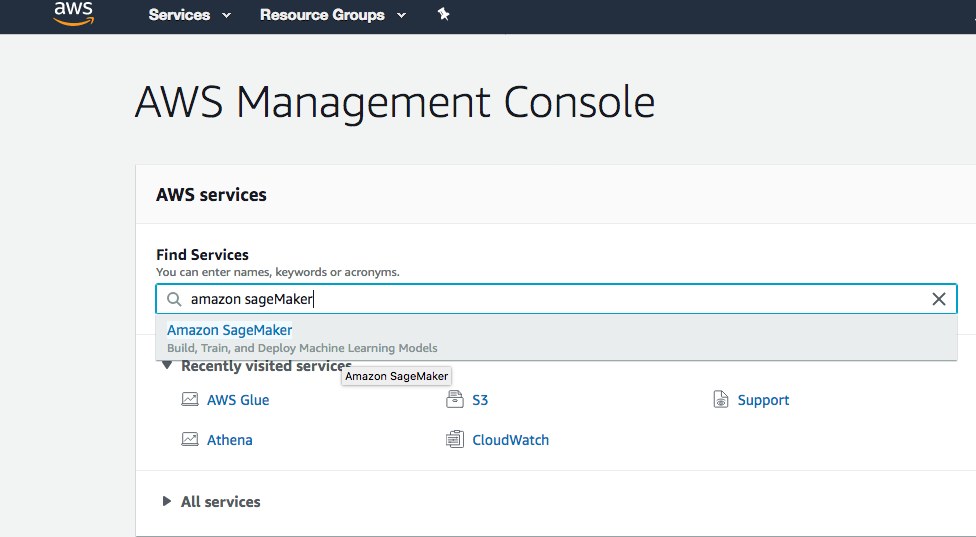
* How to use the Jupyter notebook component of Sagemaker to integrate with the data lake using Athena
* Populate data frames for data manipulation.

This process to prepare the data to satisfy the needs of ML algorithms is iterative. To prepare the data, we will make the table definitions in Athena available in a Jupyter notebook instance of SageMaker.

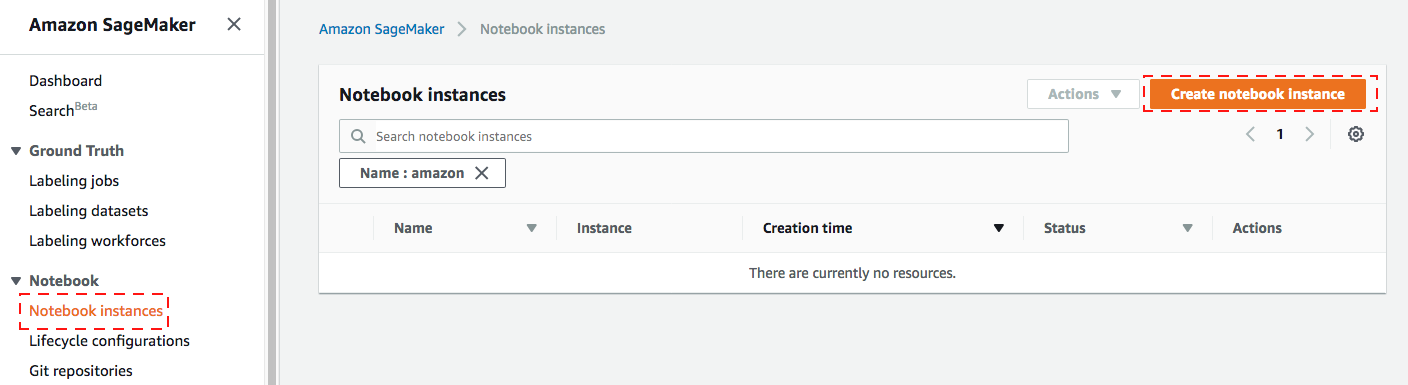
Jupyter notebooks are popular among data scientists and used to visualize data, perform statistical analysis, complete data manipulations, and make the data ready for machine learning work.

# Create Amazon SageMaker Notebook Instance

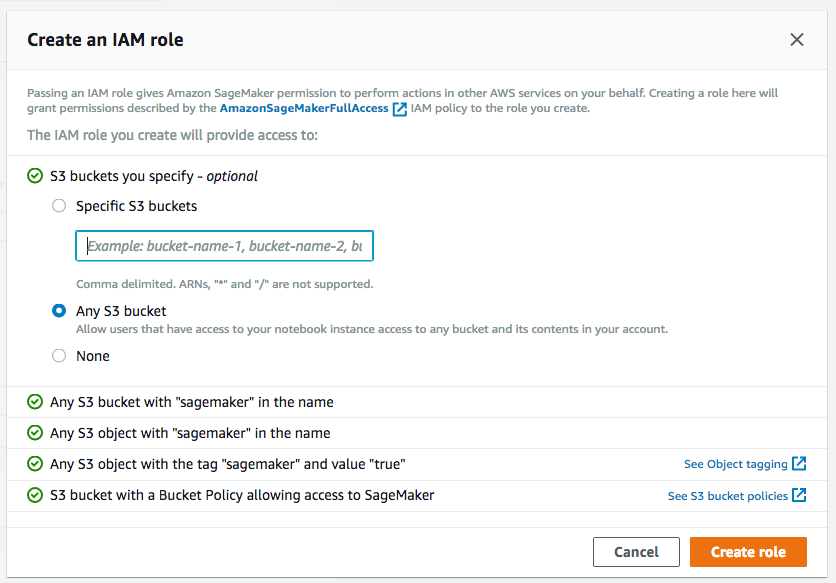
1. Go to the Amazon Sagemaker from AWS console



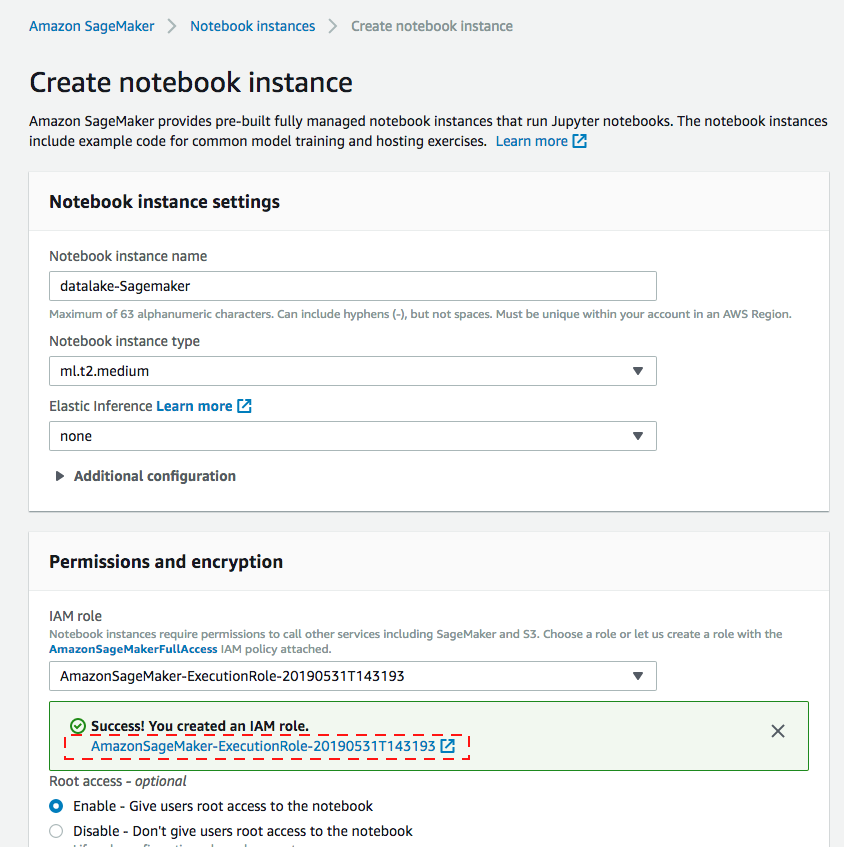
1. In the Amazon SageMaker navigation pane, click **Notebook instances a**ndClick **Create notebook instance**.



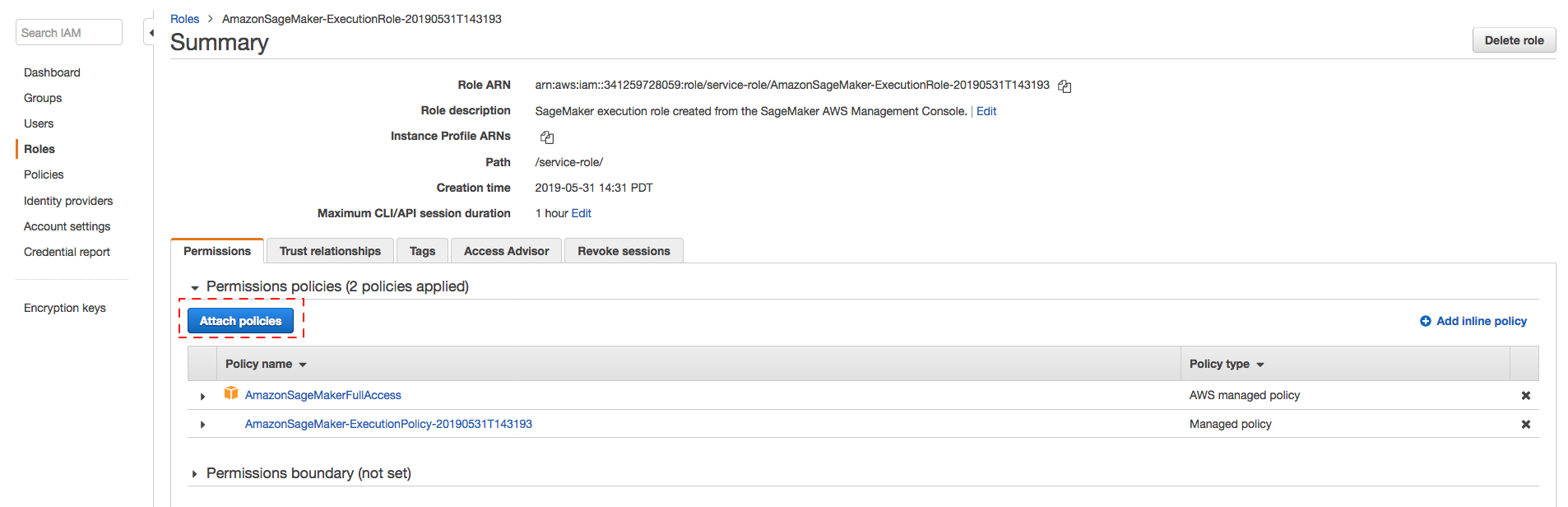
1. Put following values to create instance
   1. **Give your choice of name for Notebook instance name e.g. “datalake-Sagemaker”**
   2. You can leave **Notebook instance type** as default value “ml.t2.medium” for this lab.
   3. Leave Elastic **Inference** as null. This is to add extra resources.
   4. Choose a role for the notebook instances in Amazon SageMaker to interact with Amazon S3. As role doesn’t exist select the **Create new role** option.
   5. In **Create an IAM Role pop** up window, choose the specific S3 bucket you will grant access to. For this lab, choose **Any S3 bucket** as shown below and click **Create Role**:



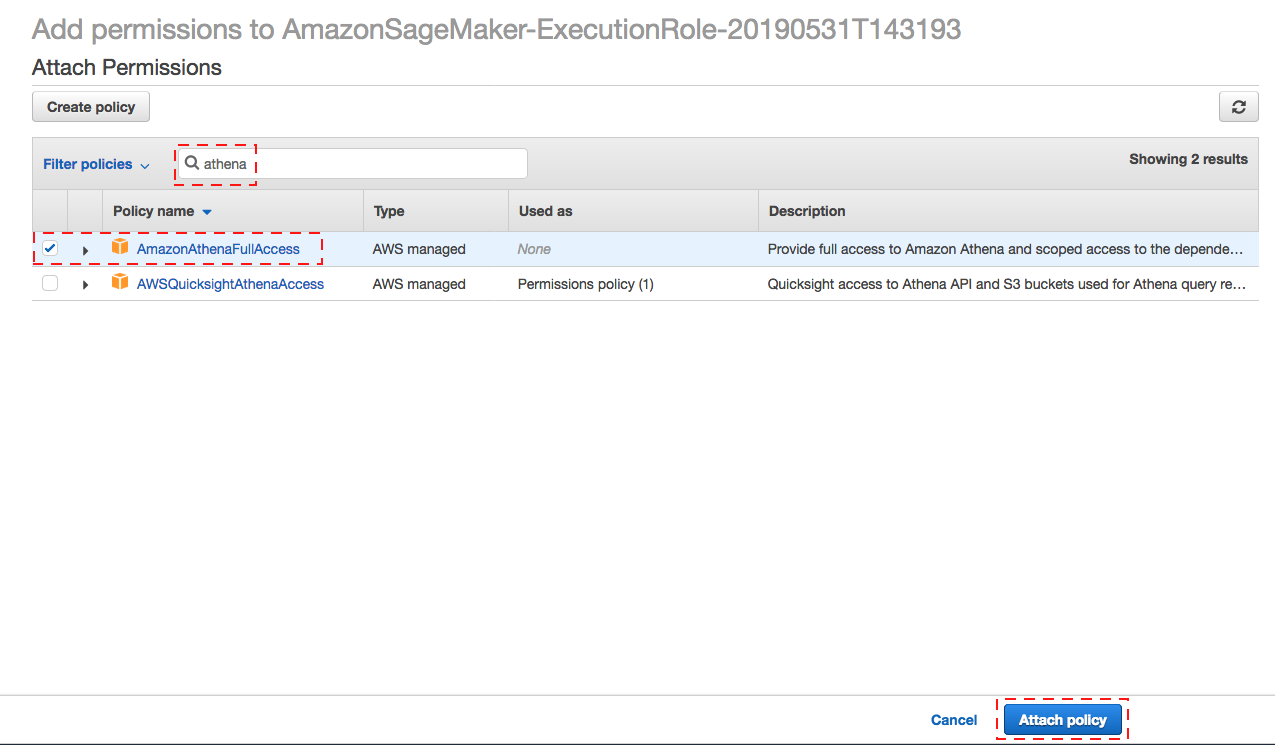
* 1. You will see a Role got create , As you are going to Access Athena from SageMaker, so SageMaker execution role must have the necessary permission to access Athena. To achieve that click the link for newly created IAM Role and in the new window



* 1. There is no Athena permission available in your SageMaker execution role. In this case, it is “AmazonSageMaker-ExecutionRole-20190531T143193”. Click **Attach Policies button** in **Permissions** tab.



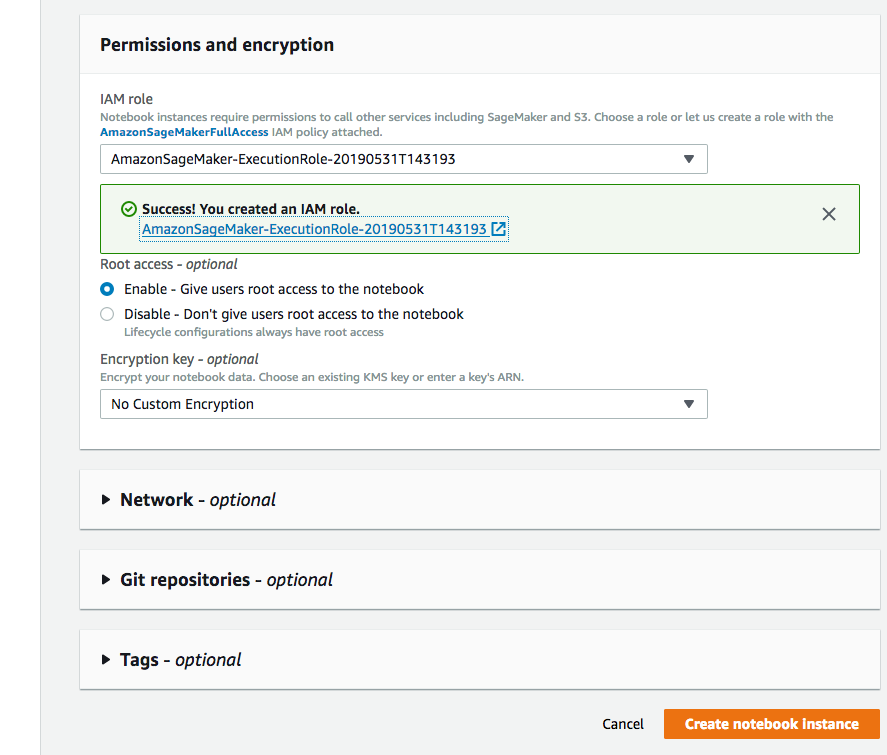
* 1. **Filter policies** by “Athena”, check **AmazonAthenaFullAccess** managed policy and click **Attach Policy** button at the bottom of the screen. Close the new window/tab.



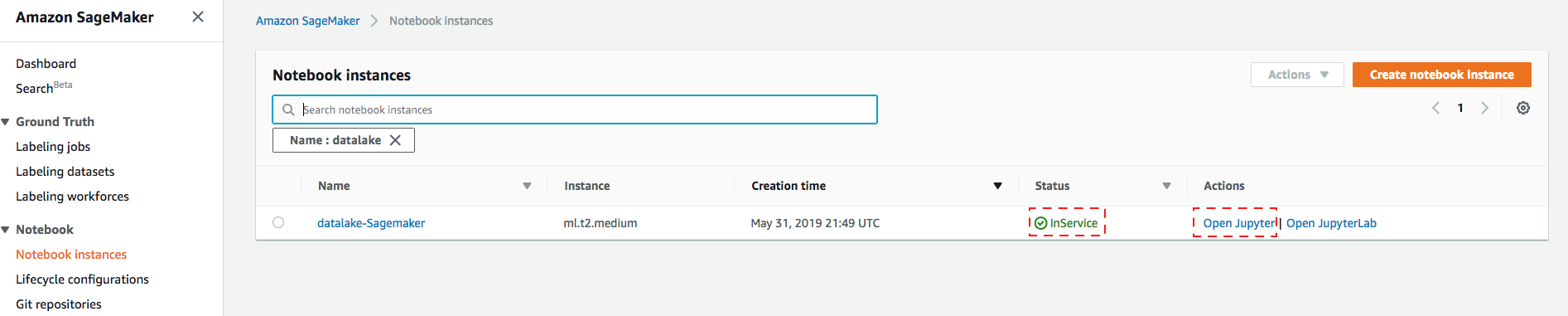
After Athena policy to IAM Role , you can close this window and go back

SageMaker browser window.

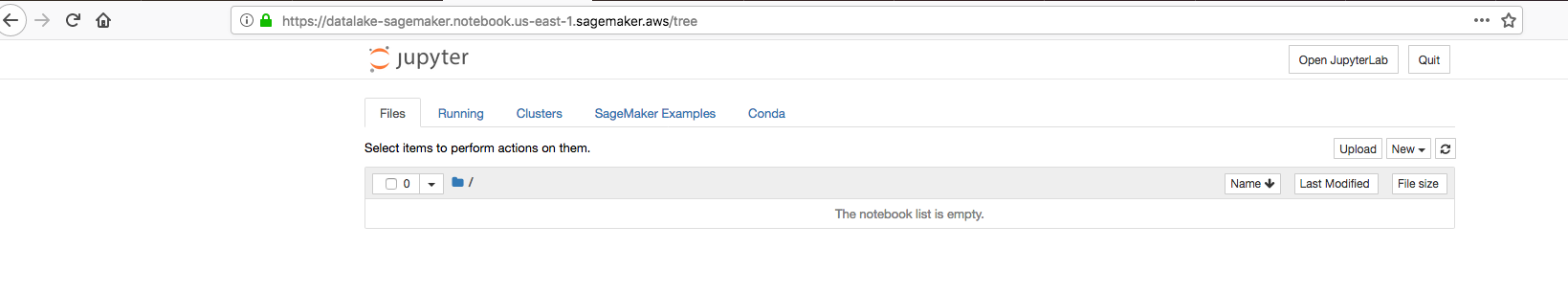
* 1. Leave all other options default. Click **Create notebook instance.**



1. Wait for the notebook instances to be created and the Status to change to **Inservice** and Click **Open Jupyter** in from **Actions column.**

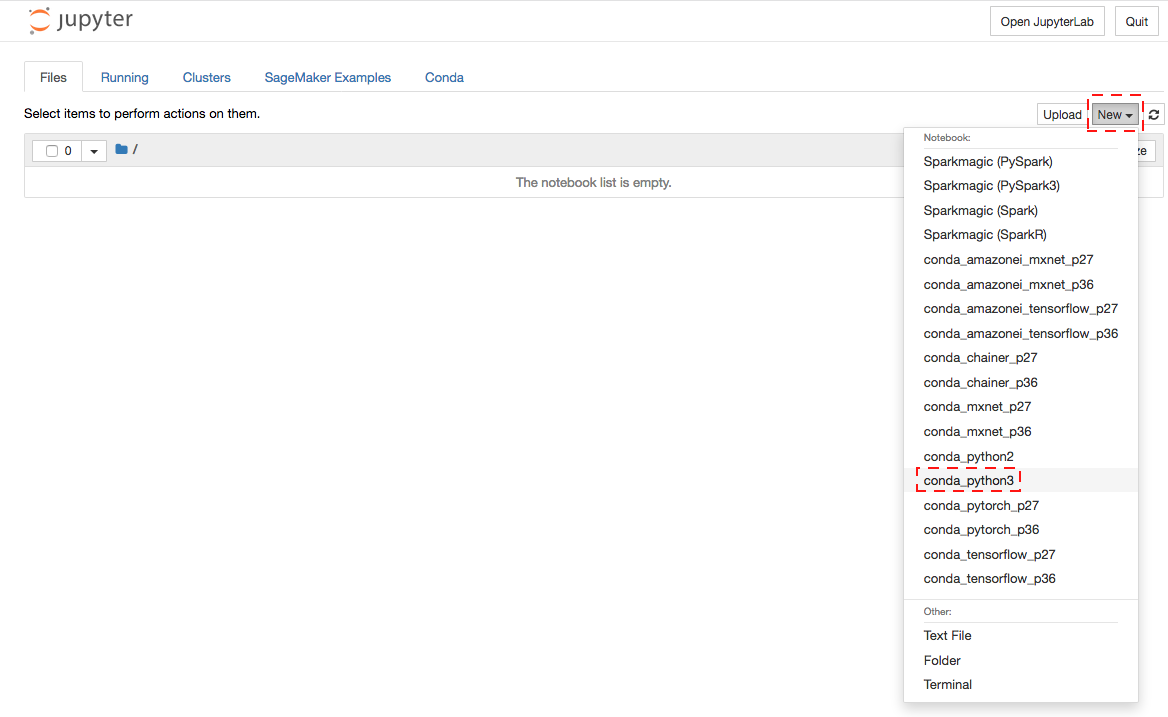


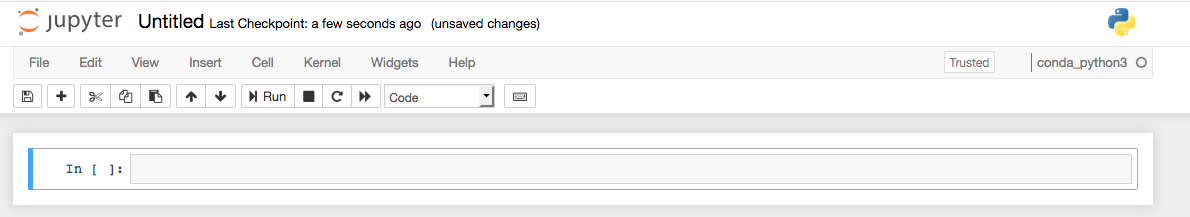
1. The notebook interface opens in a separate browser window.



# Connect the SageMaker Jupyter notebook to Athena

1. In the Jupyter notebook interface, click **New**. For the kernel, choose **conda\_python3**.

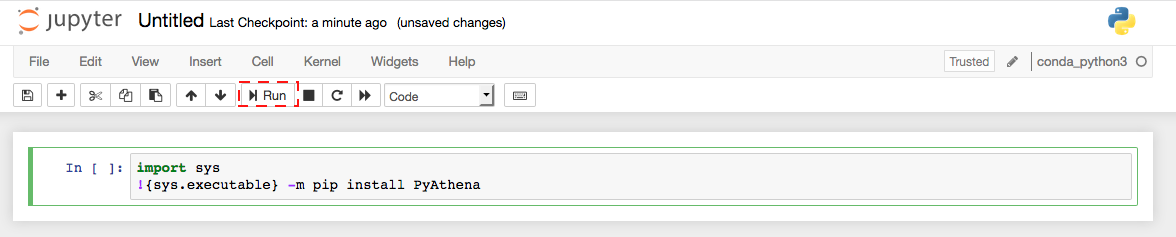




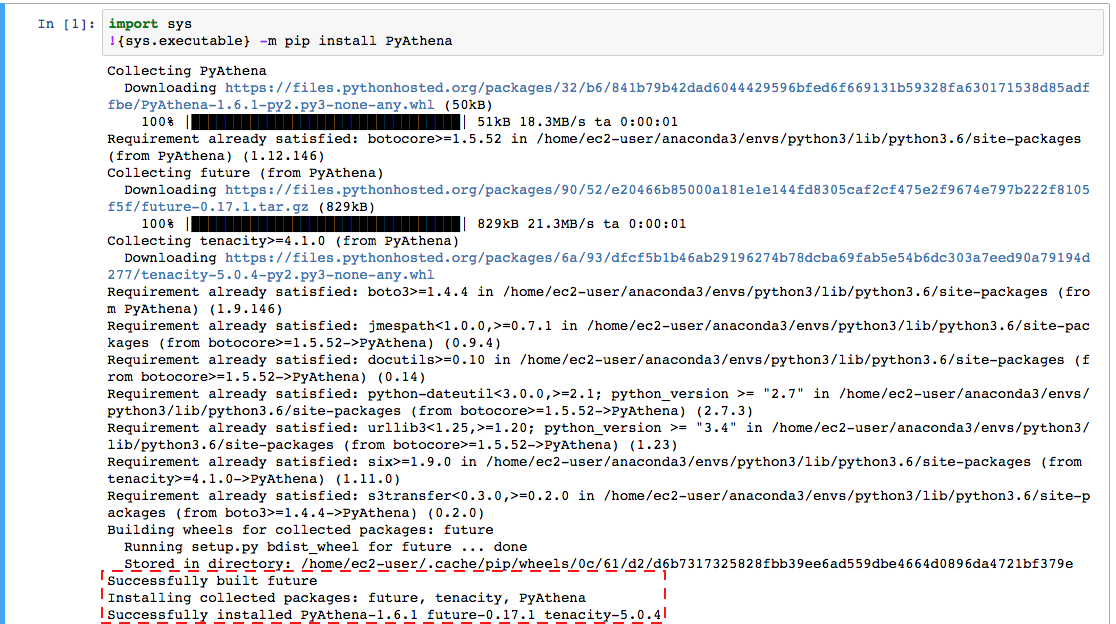
**Note:** Amazon SageMaker provides several kernels for Jupyter, including support for Python 2 and Python 3, MXNet, TensorFlow, and PySpark. This exercise uses Python because it includes the Pandas library.

1. Within the notebook, execute the following commands to install the Athena JDBC driver and In the top toolbar, click **Run**. (PyAthena is a Python [DB API 2.0 (PEP 249)](https://www.python.org/dev/peps/pep-0249/) compliant client for the [Amazon Athena JDBC driver](http://docs.aws.amazon.com/athena/latest/ug/connect-with-jdbc.html).)

Enter -> import sys  
Enter -> !{sys.executable} -m pip install PyAthena

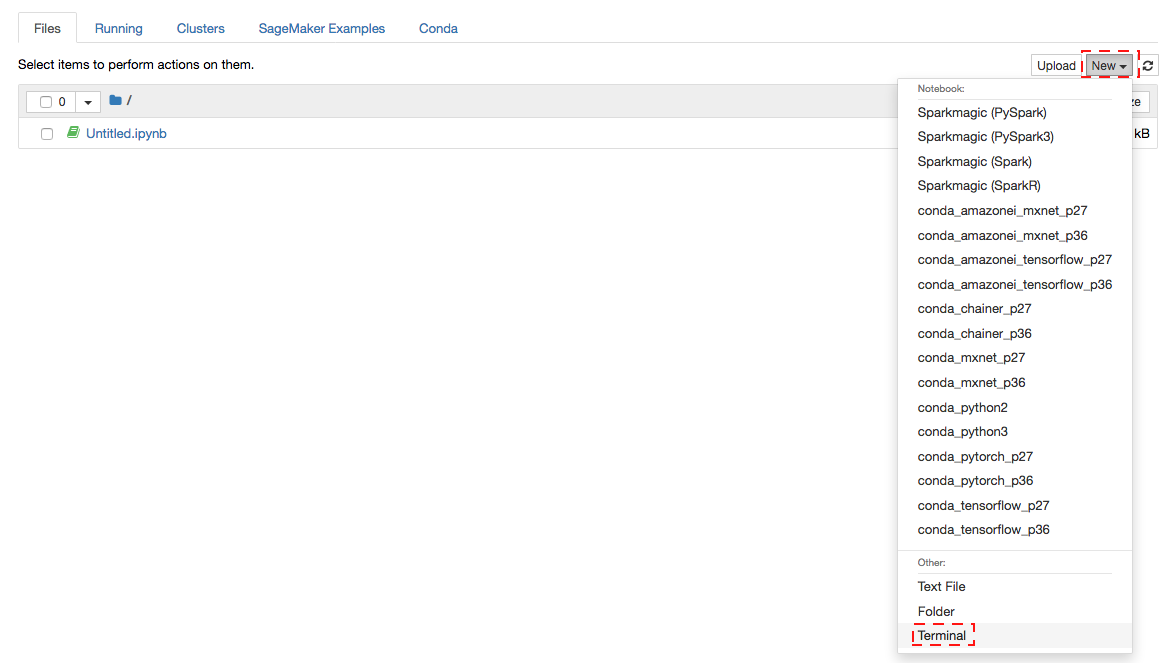


Observe success message in log:



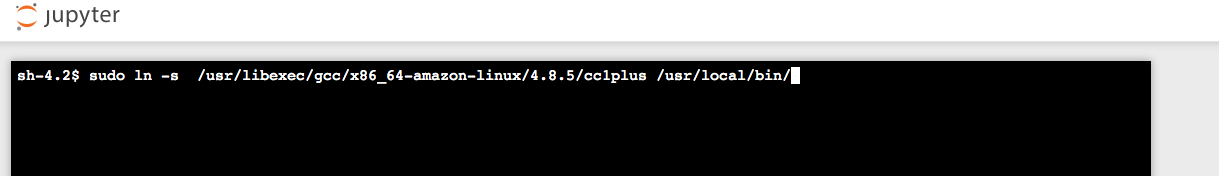
**Note:** If an error occurs, you may be required to set a file system path for an application that helps build the driver software. If this is the case, follow these steps.

* 1. In your home notebook window, select **New > Terminal**.



* 1. Copy following command and press **Enter**.

sudo ln -s /usr/libexec/gcc/x86\_64-amazon-linux/4.8.5/cc1plus /usr/local/bin/



* 1. Run the original paragraph code from Step 2 again to build the AthenaJDBC driver .

# Working in Pandas

After the Athena driver is installed, you can use the JDBC connection to connect to Athena and populate the Pandas data frames. For data scientists, working with data is typically divided into multiple stages: ingesting and cleaning data, analyzing and modeling data, then organizing the results of the analysis into a form suitable for plotting or tabular display. Pandas is the ideal tool for all of these tasks.

1. You can load Athena table data from Data lake to Pandas data frame and apply machine learning. Copy following code in your notebook and replace following:
   1. Account number : Open this link <https://console.aws.amazon.com/billing/home?#/account> and copy account ID.
   2. Your region : Throughout this lab you are working in **N. Virginia** region which us-east-1. Refer link here to get region name mapping - <https://docs.aws.amazon.com/general/latest/gr/rande.html>
   3. Your Athena Database name, this is the Glue Database name which you created during previous lab e.g. “ticketdata”

from pyathena import connect

import pandas as pd

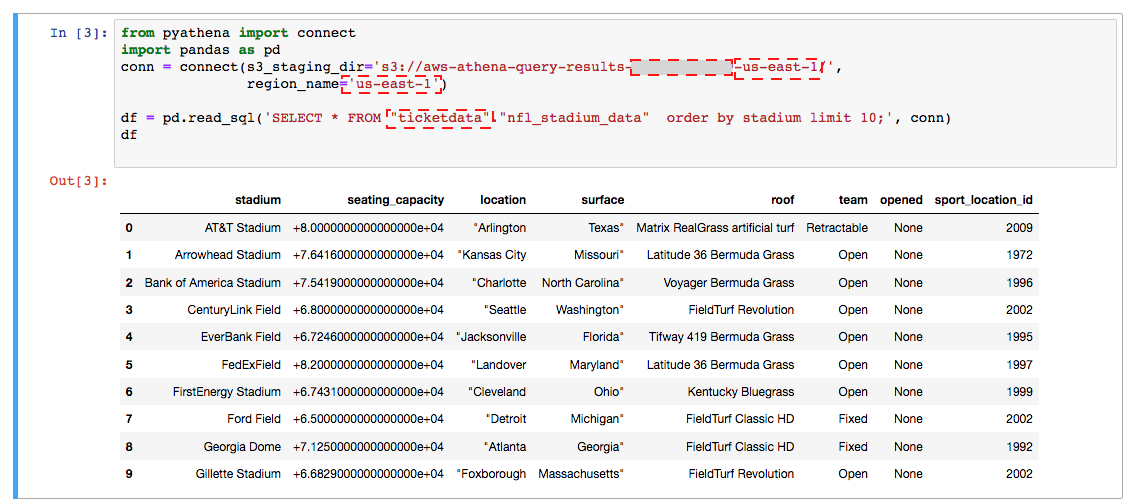
conn = connect(s3\_staging\_dir='s3://aws-athena-query-results-<youraccountnumber>-<yourregion>/',

region\_name='<yourregion>')

df = pd.read\_sql('SELECT \* FROM "<yourathenadatabase>"."nfl\_stadium\_data" order by stadium limit 10;', conn)

df

1. Click **Run** and data frame will display query output.



In this query , you are loading all nfl statdium information to panda dataframe from table nfl\_stadium\_data.

**Note**:

if you get a SageMaker does not have Athena execution permissions error issue. You need to add Athena Access to the Sagemaker role as steps provide in previous section.

1. You can use apply different ML algorithm in populated Pandas data frames. For example, draw a plot. Copy following code in your notebook and replace Your Athena Database name, this is the Glue Database name which you created during previous lab e.g. “ticketdata”.

In this query , you are loading all even ticket information to panda dataframe from table sporting\_event\_ticket\_info.

df = pd.read\_sql('SELECT sport, \

event\_date\_time, \

home\_team,away\_team, \

city, \

count(\*) as tickets, \

sum(ticket\_price) as total\_tickets\_amt, \

avg(ticket\_price) as avg\_ticket\_price, \

max(ticket\_price) as max\_ticket\_price, \

min(ticket\_price) as min\_ticket\_price \

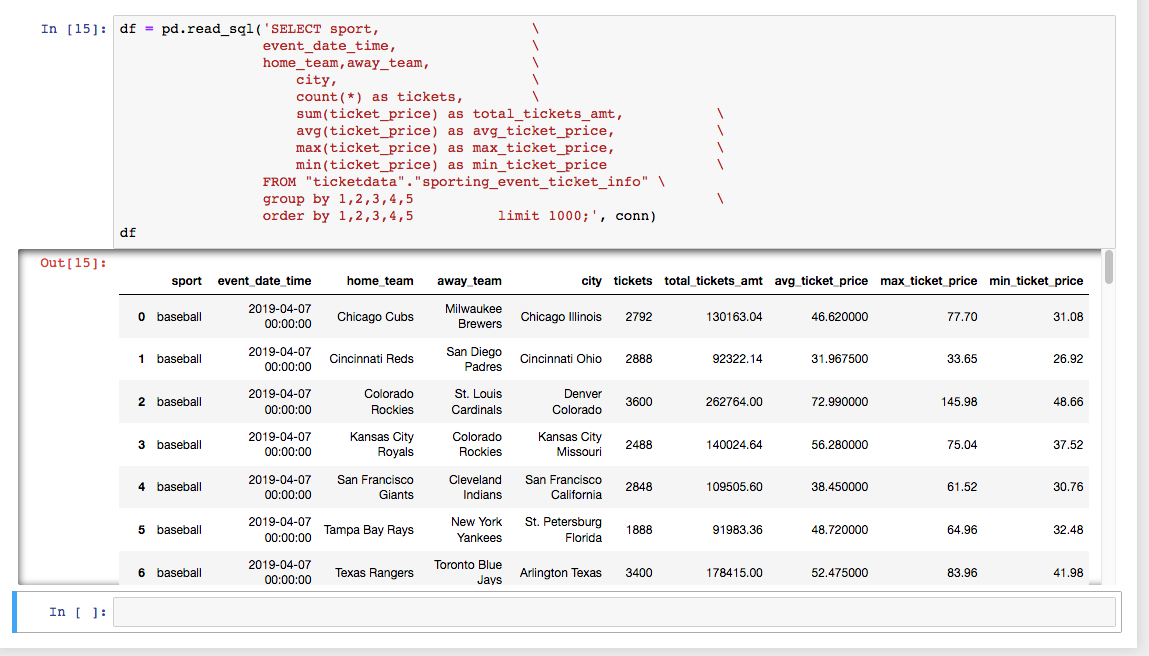
FROM "<yourathenadatabase>"."sporting\_event\_ticket\_info" \

group by 1,2,3,4,5 \

order by 1,2,3,4,5 limit 1000;', conn)

df

1. Click **Run** and data frame will display query output



1. In new execution line copy following code

import matplotlib.pyplot as plt

df.plot(x='event\_date\_time',y='avg\_ticket\_price')

1. Click Run and you will see data plot which got draw using matplotlib library.

