ENTERPRISE DATA ARCHITECTURE AND THE ANALYTICS PLAYGROUND

Overview and best practices.

January 2020

Internal Use Only

AGENDA

Welcome: Course Introduction

Module 1: E*TRADE Data Architecture Overview

Module 2: E*TRADE Analytics Playground Overview

Module 3: Hands-on Exercises - Playground Best Practices

Course Overview

This course provides an introduction and overview of E*TRADE's analytics playground, with an emphasis on E*TRADE's data architecture, metadata, data lineage, and using analytics best practices that lead to:

continuously discovering, developing and refining business processes in support of operational excellence.

surfacing marketplace insights that help provide a sustainable competitive advantage in the marketplace.

The course also includes hands on exercises around the prescribed analytics best practices.

This course is a prerequisite for gaining access to the analytics playground.

Course Objectives

- Introduce E*TRADE's *future state* data architecture and the Amazon Web Services (AWS) analytics playground.
- Review the capabilities, features, and functionalities of the analytics playground.
- Learn the core code of conduct allowed when working in the analytics playground.
- Review best practices for working in the analytics playground.

Training Prerequisites

The following are prerequisites for this training:

Environment access

- Prod: *447222439265*

AD group membership

- DataScientist

AWS workgroup membership

- DataScientist
- SageMaker

Application access

- Tableau
- Business Objects

Other

- Knowledge and familiarity with the related analytics applications and E*TRADE business processes.

*Training on specific tools or fundamental analytics topics will not be covered in this training. See Appendix A for reference material.

Course Modules

This course is comprised of the following three modules:

Module 1: E*TRADE Data Architecture Overview

- E*TRADE enterprise data architecture overview
- Detailed topics discussion:
  - Data lineage
  - Metadata
  - Tokenized data

Module 2: E*TRADE Analytics Playground Overview

- Analytics playground components
- Analytics playground roles
- Analytics playground best practices

Module 3: Hands-on Exercises - Playground Best Practices

- **Exercise 1**: Query the data lake with Athena and save the results.
- **Exercise 2:** Report against the data lake with Tableau and save the output.
- **Exercise 3:** Prepare and review a predictive model with SageMaker.

MODULE 1

E*TRADE Data Architecture Overview:

- E*TRADE enterprise data architecture review
- Detailed topics discussion:
  - Data lineage
  - Metadata
  - Tokenized data

Mission Statements

The mission statements of E*TRADE's enterprise data architecture and the analytics playground are as follows:

E*TRADE Enterprise Data Architecture

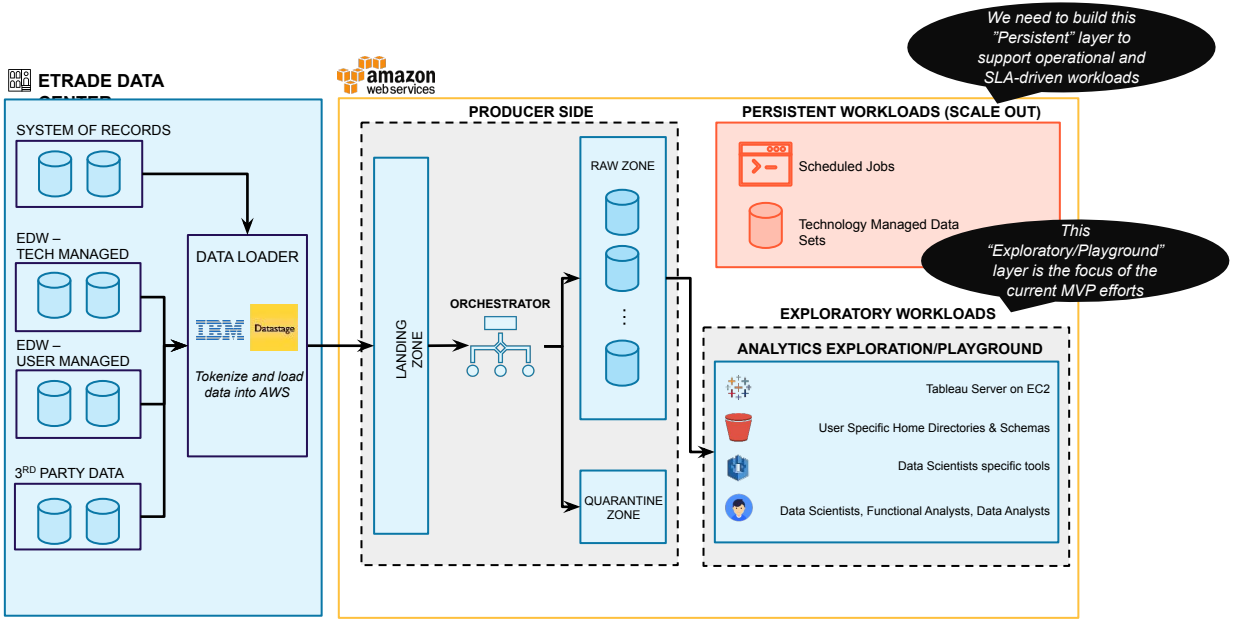
- to modernize and advance the analytics capabilities, both for retroactive reporting and (more strategically) for predictive algorithm development and refining, forecasting, and segmentation.

E*TRADE Analytics Playground

- to provide a common mechanism for presenting important data assets to knowledge workers in an ideal format for reporting and discovery.

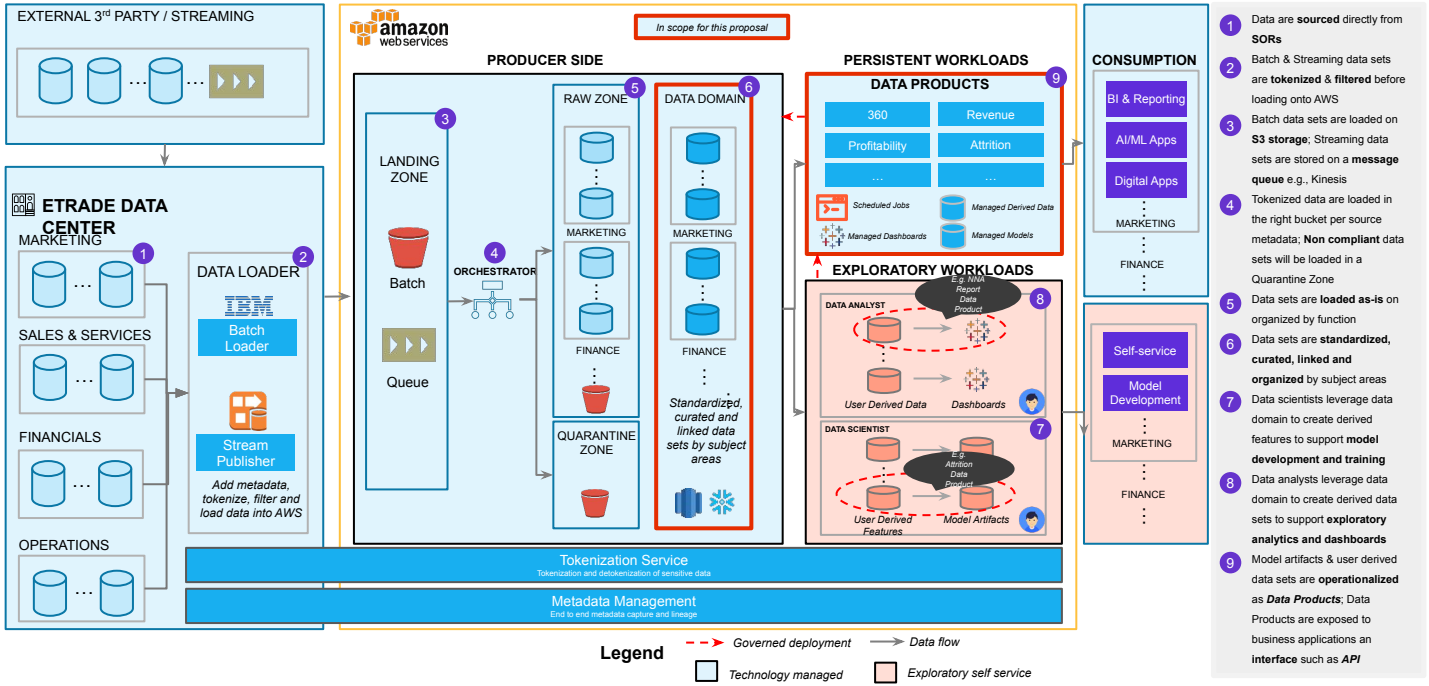
E*TRADE Enterprise *Current State* Data Architecture

Developing and operationalizing an enhanced Revenue data product requires scaling out the current architecture.



E*TRADE Enterprise *Future State* Data Architecture

We should continue to build towards the target state analytics architecture to scale out onboarding of additional teams and business analytical use cases.



Raw Zone & Data Lake

Data in the analytics playground originates as data sets from source systems that first land inside AWS in the *raw zone (S3 bucket)*.

Raw Zone

The data here appears “as is” from the E*TRADE data center source systems managed via a detailed *step function*.

Data Lake

Data validated by the spec file and trigger file in the raw zone is pushed to the data lake to be used for the types of analytics projects discussed in this course.

The data landing in the raw zone is validated using:

- Spec file
- Trigger file
- Data file
- If spec file doesn’t validate, data gets pushed to *quarantine zone S3 bucket*

Data lake access privileges are driven by:

- AWS account access
- AWS console login roles (users may have multiple roles)

Requesting AWS roles:
https://bmcsmrt.corp.etradegrp.com/ux/myitapp/#/catalog/home

Additional *Data Lake* info:
https://confluence.corp.etradegrp.com/display/IE/Analytics+Data+Lake

Analytics Playground

The AWS-based playground is a dynamic, self-purging environment (45-day lifecycle policy) that only allows for assets and projects that are progressing in a “meaningful” way (as defined in this training) to persist. Below are key structural components that define the playground.

Playground *user access*

User access is SSO to AWS console and privileges are defined by AD and AWS role(s) membership (as defined in a following slide).

Additional info:
https://confluence.corp.etradegrp.com/display/IE/SSO+Roles

Playground *environment*

The playground is on E*TRADE Production environment (**447222439265**).

Additional info:
https://confluence.corp.etradegrp.com/display/IE/Analytics+Data+Lake

Playground *data source*

The playground data sources are S3 buckets (e.g. the data lake is an S3 bucket.)

Users in the playground will also have an S3 bucket associated to their AWS account to store assets of active projects.

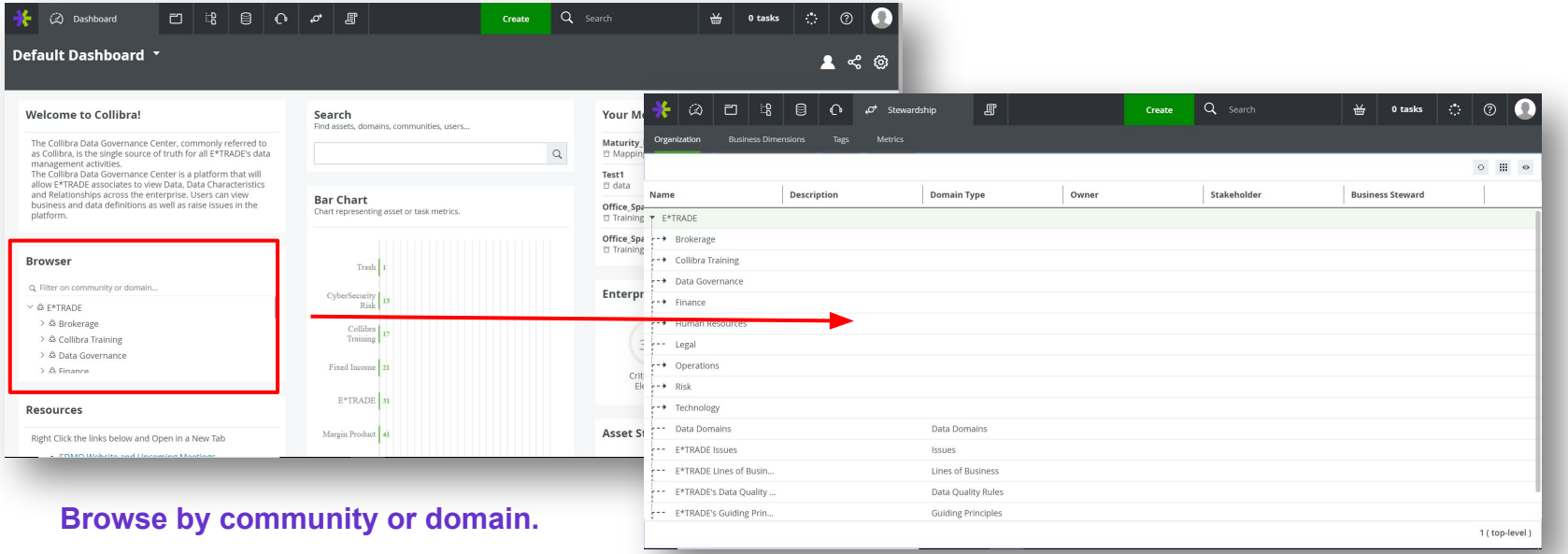
Available playground *tools*

E*TRADE provides a curated stack of AWS tools (Athena, SageMaker, Glue, etc.) and non-AWS tools (e.g. BusinessObjects, Tableau).

Additional info:
https://confluence.corp.etradegrp.com/display/BUSAN/Code+Corner+-+Tableau

Data Lineage

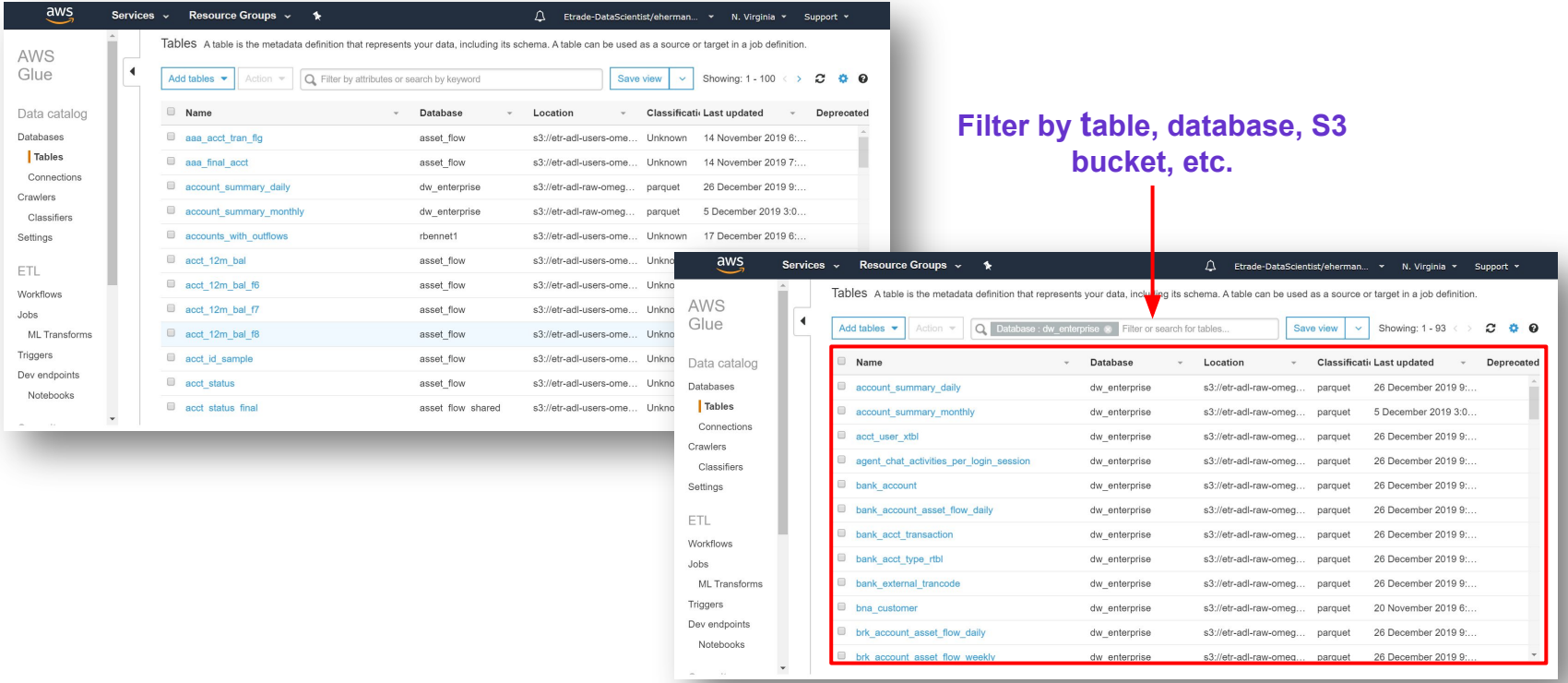
E*TRADE's data lineage can be reviewed using Collibra by going to data.etrade.com.



Browse by community or domain.

Metadata

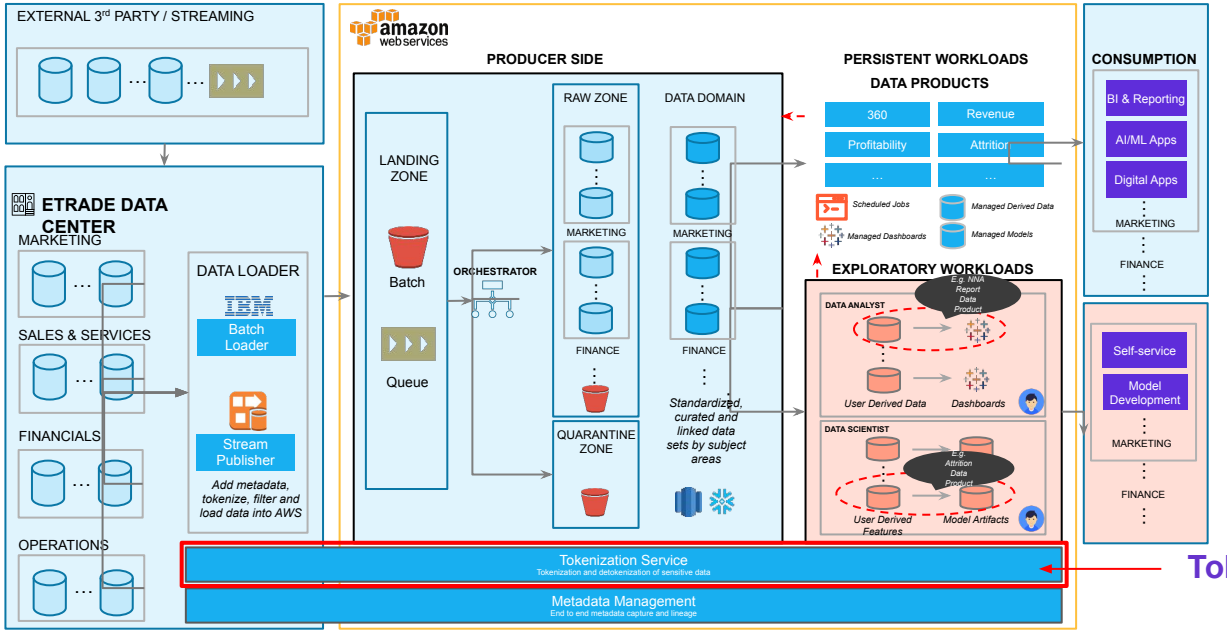
E*TRADE's data catalog can be searched using AWS Glue.



Tokenized Data

The *Tokenization Service* anonymizes detailed, atomic level operational data containing sensitive and/or personal information.

- The service provides both *tokenization* and *de-tokenization* of sensitive data.



MODULE 2

E*TRADE Analytics Playground Overview:

- Analytics playground components
- Analytics playground roles
- Analytics playground best practices

Analytics Playground Components

The playground is comprised of the following components:

- Exploratory, playground environment
- Role based access
  - Users may belong to multiple roles
- Self-purging environment
  - 45-day lifecycle policy

- All data starts as SORs/raw sets in the *raw zone* and validated prior to being pushed to the *data lake*
- Not all E*TRADE data available in the data lake

- AWS Athena
- AWS Glue
- AWS SageMaker
- Tableau, BO

Analytics Playground Roles

Access to the AWS console and specific AWS tools is controlled by membership to AD group roles. Users can claim multiple roles in the playground.

Additional info:
https://confluence.corp.etradegrp.com/display/IE/SSO+Roles

Functional analyst

Functional analysts work on initiatives dealing with large and complex trading systems.

F*unctional analyst* tools:
- Cloudwatch
- S3
- Athena
- Glue
- Redshift

Data analyst

Data analysts provide recurring and ad hoc C-level reporting and analytics.

D*ata analyst* tools:
- Cloudwatch
- S3
- Athena
- Glue
- Redshift

Data scientist

Data scientists build new & improve existing models & data products to power business decision making.

D*ata scientist* tools:
- Cloudwatch
- S3
- SageMaker
- Athena
- Glue
- Redshift

Data engineer

Data engineers research, evangelize, apply, lead, monitor backend systems and data sources.

D*ata engineer* tools:
- Cloudwatch
- S3
- SageMaker
- Athena
- Glue
- Redshift

Analytics Playground Best Practices

Users of the playground are expected to follow best practices related to:

- Utilization of data assets of the analytics playground.
- Querying the data lake zone and saving results to personal locations in analytics playground.
- Creating a report against data sources in data lake and/or data saved in personal location(s).
- Creating, preparing and deploying predictive analytic models to production.
- Documenting and registering a solution to be captured in knowledge base and protected from deletion by regular “laboratory cleanup” jobs.
- Uploading external data to playground for inclusion in above processes.

MODULE 3

Hands-on Exercises - Playground Best Practices:

- **Exercise 1:** Query the data lake with Athena and save the results.
- **Exercise 2:** Report against the data lake with Tableau and save the output.
- **Exercise 3:** Prepare and review a predictive model with SageMaker.

Exercise Objectives

The exercises in this module are intended to demonstrate accessing the playground, illustrate the features, functionalities, and capabilities of the playground, as well as review and discuss the best practice for working in the playground. The exercises are as follows:

Exercise 1: Query the data lake with Athena and save the results.

- Task 1 – Access the AWS console and Athena.
- Task 2 – Query the data lake using Athena.
- Task 3 – Save the results to an S3 bucket and as .csv file.

Exercise 2: Report against the data lake with Tableau and save the output**.**

- Task 1 – Login to Tableau.
- Task 2 – Create a workbook using the Analytics Playground training data set.
- Task 3 – Save the workbook.

Exercise 3: Prepare and review a predictive model with SageMaker.

- Task 1 – Access the AWS console and SageMaker.
- Task 2 – Create a notebook instance.
- Task 3 – Review a notebook instance.

Analytics Playground Training Data Set

The Analytics Playground Training Data Set is comprised of the following:

Trade Fact

- 500 generated rows of dummy trade facts.
- Each trade has an execution date, customer, stock, number of shares, price of share, etc.
- Customers and stocks are picked randomly by simple algorithm.
- Trade Type (Buy or Sell) randomly generated for trades.
- Execution dates randomly selected between account open date and today.
- Stock prices randomly selected between high and love values from the Stock dimension.

Customer Dimension:

- One row per (Disney) character, including date of birth and a randomized date they opened their E*TRADE account
- Date of birth used to calculate age at the time of trade
- Randomly generated account open date.

Stock Dimension:

- One row per (entertainment industry) stock.
- Includes an arbitrary high and low price/stock.

Static Copy of Trade Fact

- Copy of the above fact, “values” only. It should therefore be sta
  recalculates whenever anything changes.
- This tab is provided to give you something stable to export, if ne

Customer ID	Customer First Name	Customer Last Name	Customer Full Name	Customer Date of Birth	Customer Open Account Date
1	Mickey	Mouse	Mickey Mouse	1/1/1930	1/11/1986

Trade ID	Customer ID	Customer Full Name	Trade Execution Da	Customer Age at Trade	Stock	Stock Symbo	Stock Name	Trade Type	Trade Share Count	Trade Share Price	Trade Value Amount
1	8	Doc Dwarf	10/3/2008	57	4	VIA	Viacom, Inc.	Buy	508	$ 233	$ 118,364
2	7	Bashful Dwarf	4/8/2014	64	4	VIA	Viacom, Inc.	Buy	121	$ 206	$ 24,926
3	6	Louie Duck	7/17/2016	49	3	DIS	Walt Disney Company	Sell	1,182	$ 80	$ 94,560
4	9	Dopey Dwarf	10/18/2019	62	4	VIA	Viacom, Inc.	Buy	1,333	$ 200	$ 266,600
5	18	Dale Monk	6/3/2018	25	8	IMAX	IMAX Corp.	Buy	133	$ 74	$ 9,842
6	20	Blue Fairy	8/27/2016	45	9	CBS	CBS Corp.	Buy	710	$ 232	$ 164,720
7	16	Pluto Dog	1/10/2013	63	8	IMAX	IMAX Corp.	Sell	1,562	$ 131	$ 204,622
8	15	Goofy Dog	1/4/2010	65	7	AMC	AMC Entertainment Holdings Inc.	Sell	1,260	$ 110	$ 138,600
9	13	Sneezy Dwarf	5/21/2004	24	6	DISCA	Discovery Inc.	Sell	1,015	$ 237	$ 240,555
10	13	Sneezy Dwarf	4/14/2018	38	6	DISCA	Discovery Inc.	Sell	661	$ 273	$ 180,453
11	22	Dumbo Elephant	7/12/2018	37	10	CNK	Cinemark Holdings Inc.	Sell	279	$ 179	$ 49,941
12	17	Chip Monk	2/15/2018	27	8	IMAX	IMAX Corp.	Sell	434	$ 115	$ 49,910
13	1	Mickey Mouse	8/18/2016	86	1	T	AT&T	Sell	949	$ 48	$ 45,552
14	15	Goofy Dog	1/19/2019	74	7	AMC	AMC Entertainment Holdings Inc.	Sell	1,655	$ 196	$ 324,380
15	3	Donald Duck	3/17/1956	31	2	FOX	Twenty-First Century Fox	Sell	1,484	$ 53	$ 78,652
16	14	Snow White	4/23/2016	26	6	DISCA	Discovery Inc.	Buy	286	$ 294	$ 84,084
17	4	Huey Duck	6/12/2014	87	2	FOX	Twenty-First Century Fox	Sell	1,567	$ 69	$ 108,123
18	14	Snow White	8/2/2014	24	7	AMC	AMC Entertainment Holdings Inc.	Sell	119	$ 119	$ 14,161

Stock ID

E*TRADE

Analytics Playground Training Data Set (cont.)

The Analytics Playground Training Data Set is comprised of the following:

AWS location of training data set:
Production S3 raw zone:
s3 - etr-adl-raw-omega-ue1

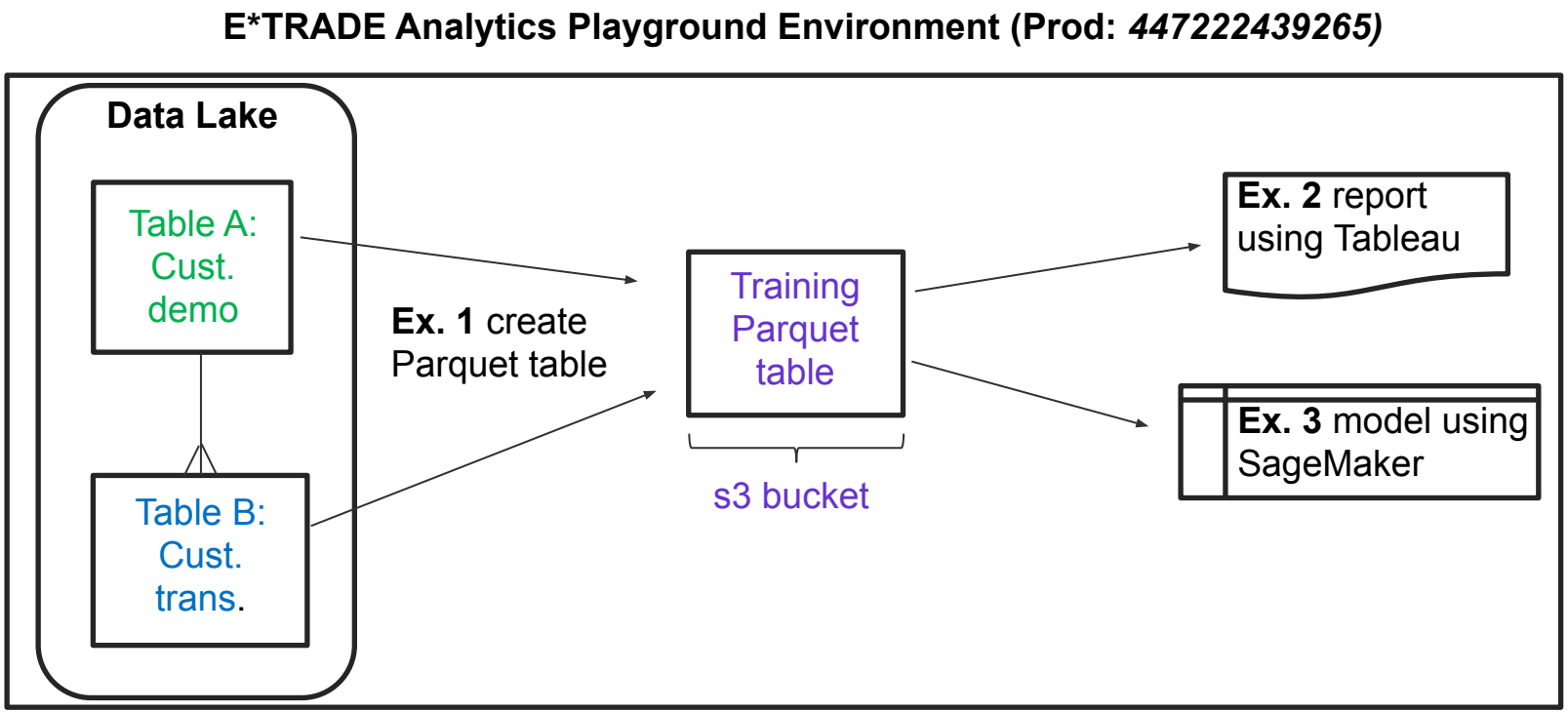
Directory structure:
s3://etr-adl-raw-omega-ue1/ANALYTICS_TRAINING/SAMPLE_TRAINING_DATA/CustDim.csv

s3://etr-adl-raw-omega-ue1/ANALYTICS_TRAINING/SAMPLE_TRAINING_DATA/StockDim.csv

s3://etr-adl-raw-omega-ue1/ANALYTICS_TRAINING/SAMPLE_TRAINING_DATA/TradeFact.csv

Hands-on Exercise Diagram

This diagram provides a high-level picture of the exercises and how they relate.



EXERCISE 1

Query the data lake with Athena and save the results.
Task 1 – Access the AWS console and Athena.
Task 2 – Query the data lake using Athena.
Task 3 – Save the results to an S3 bucket and as a .csv file.

Exercise 1: Query the data lake with Athena and save the results.

Exercise overview:

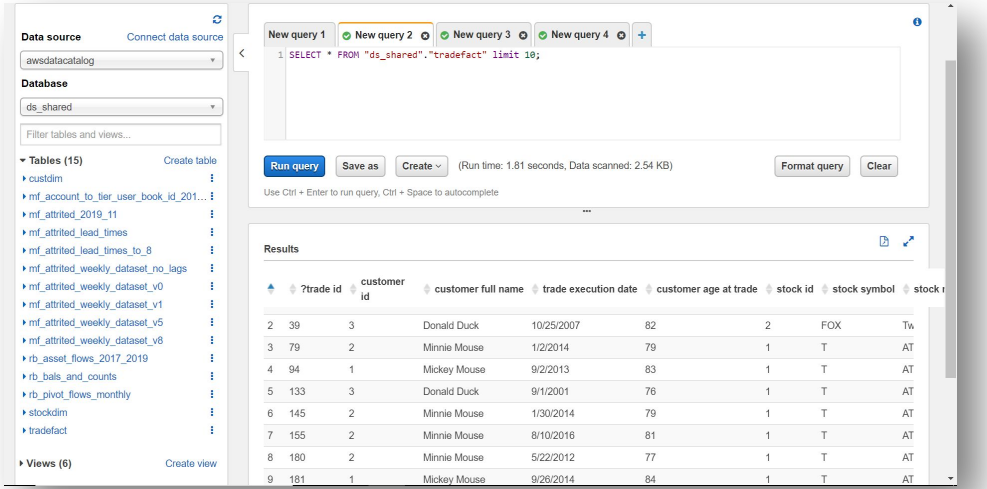
This exercise reviews best practices for querying the data lake using AWS Athena and saving the results.

Exercise objectives:

In this exercise, we will accomplish the following tasks:

Task 1 – Access the AWS console and Athena.
Task 2 – Query the data lake using Athena.
Task 3 – Save the results to an S3 bucket and as .csv

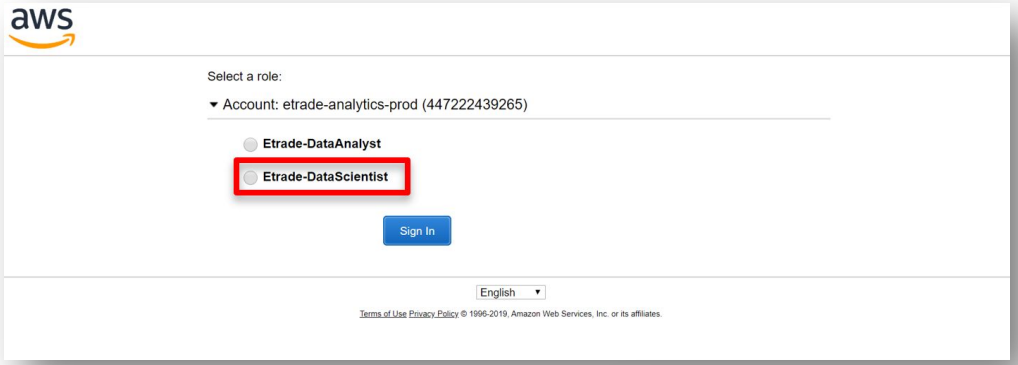
Amazon Athena is an interactive query service that makes it easy to analyze data directly in Amazon Simple Storage Service (Amazon S3) using standard SQL. With a few actions in the AWS Management Console, you can point Athena at your data stored in Amazon S3 and begin using standard SQL to run ad-hoc queries and get results in seconds.

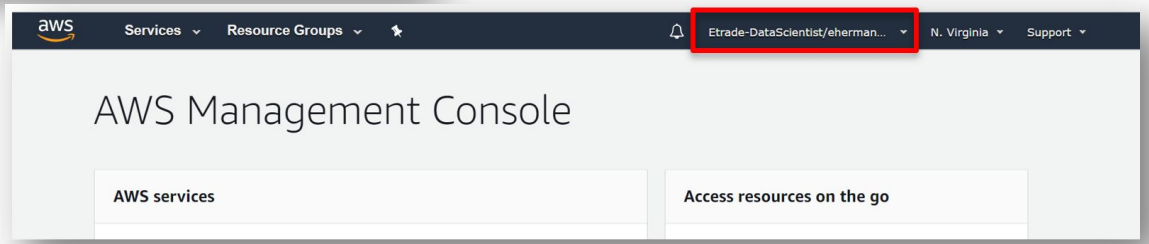


Exercise 1: Query the data lake and save the results (cont.)

Task 1 – Access the AWS Management Console and Athena.

1. Click **link provided by your administrator** to login to the **AWS management console**.
2. Log in as **Etrade-DataScientist**.

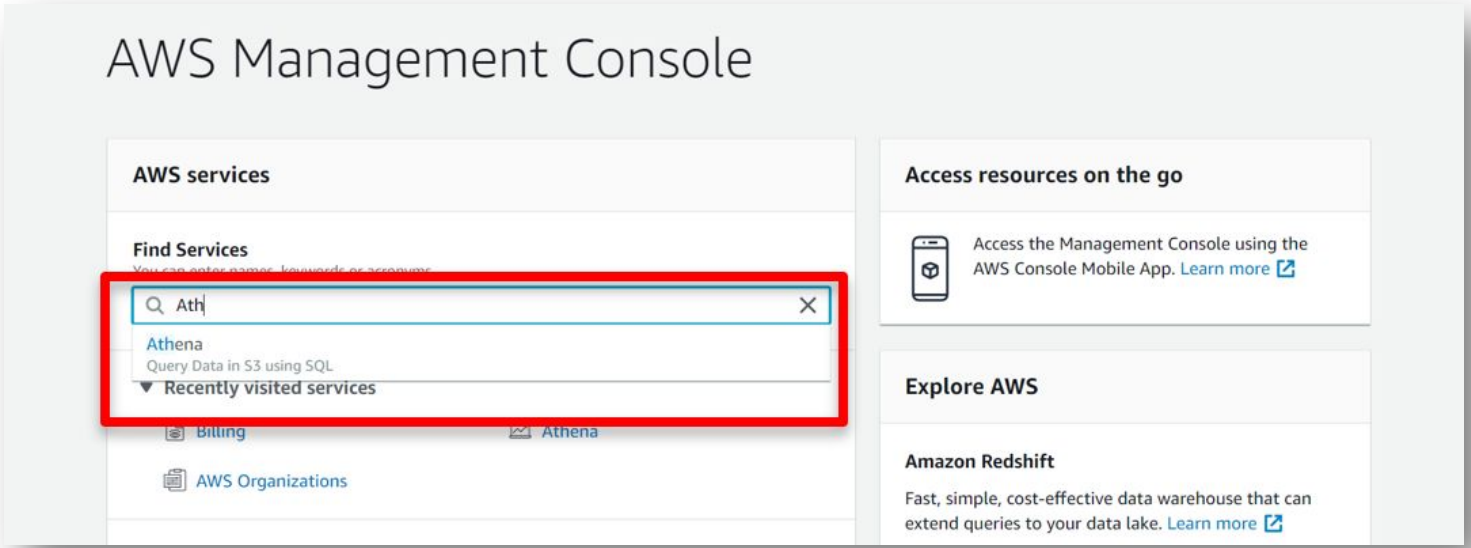




Exercise 1: Query the data lake and save the results (cont.)

Task 1 – Access the AWS Management Console and Athena (cont.)

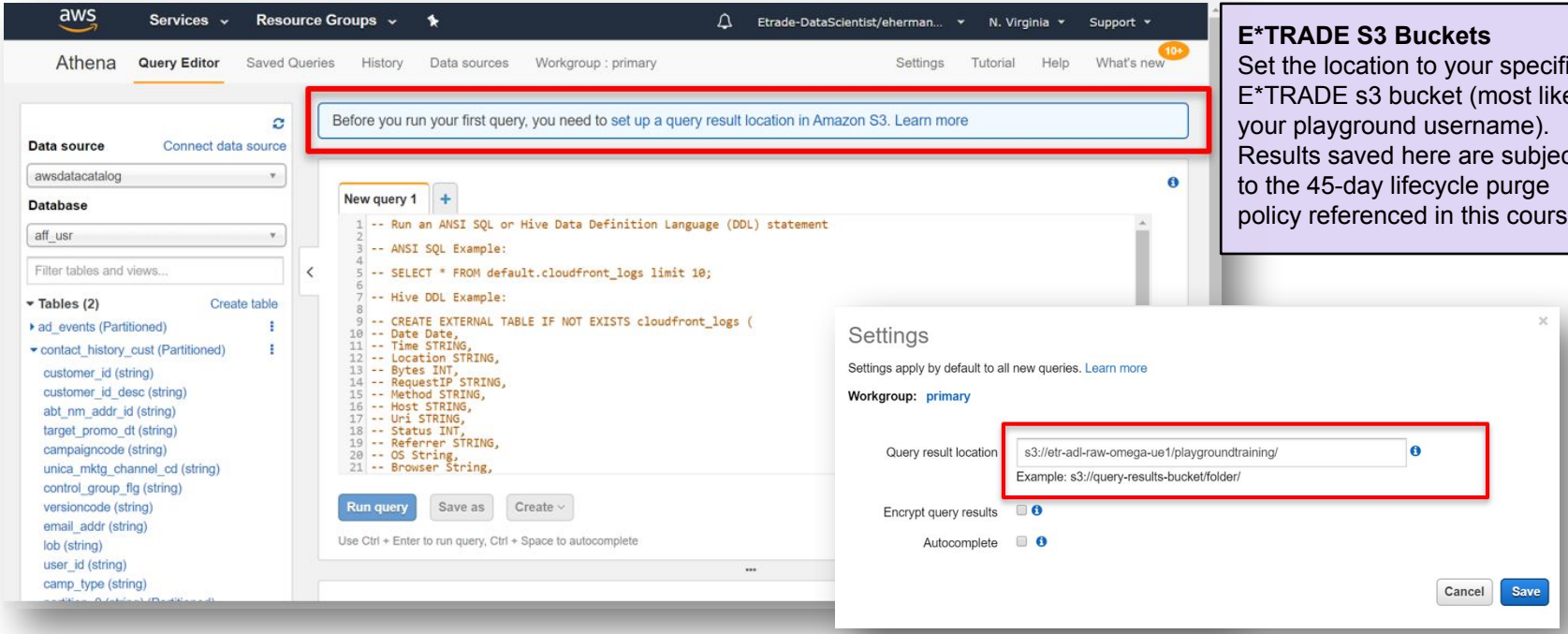
3. In the **AWS Management Console**, search for and select **Athena.**



Exercise 1: Query the data lake and save the results (cont.)

Task 1 – Access the AWS Management Console and Athena (cont.)

4. First time users are prompted to **set up a query result location in Amazon S3** by specifying an S3 bucket.

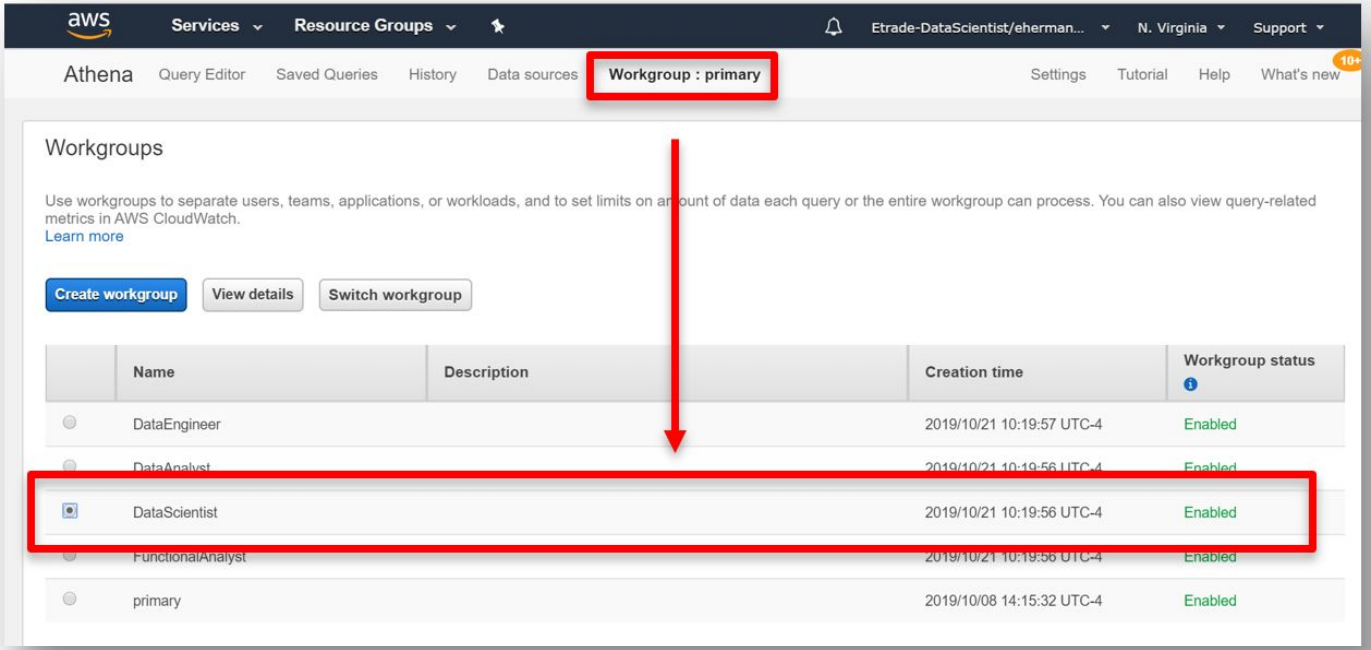


E*TRADE S3 Buckets
Set the location to your specific E*TRADE s3 bucket (most likely your playground username). Results saved here are subject to the 45-day lifecycle purge policy referenced in this course.

Exercise 1: Query the data lake and save the results (cont.)

Task 1 – Access the AWS Management Console and Athena (cont.)

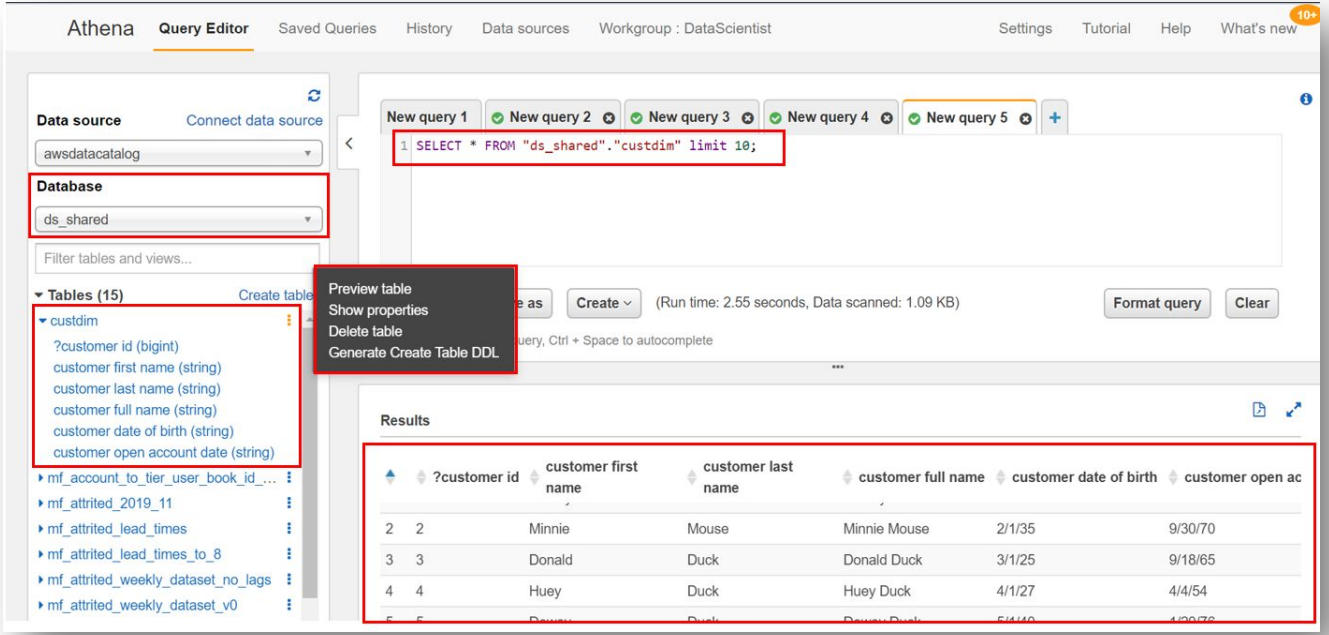
5. Under **Workgroup : primary**, ensure the **DataScientist** group is selected.



Exercise 1: Query the data lake and save the results (cont.)

Task 2 – Query the data lake using Athena.

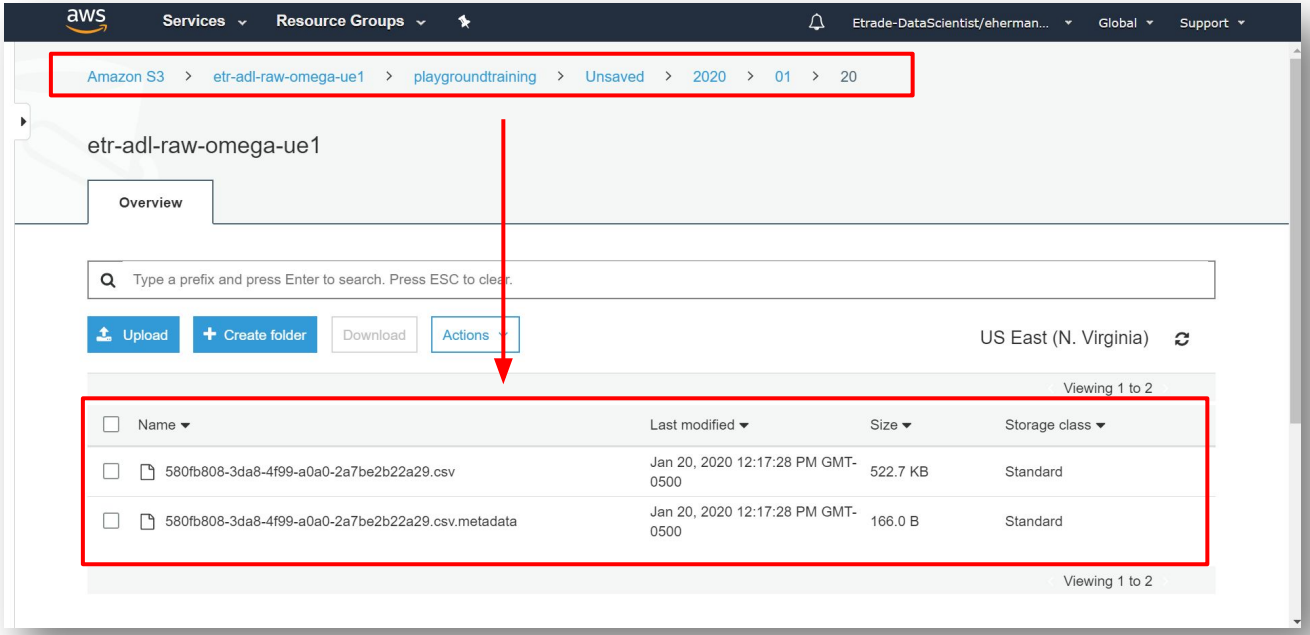
1. Under **Database**, choose **ds_Shared**, expand **custdim** and review the data items, then click the ellipsis and choose **Preview table** and review the **Results**. (Repeat the process to view the **stockdim** and **tradefact** tables.)



Exercise 1: Query the data lake and save the results (cont.)

Task 3 – Save the results to an S3 bucket.

1. From the **AWS console**, navigate to your **S3 bucket** defined in *task 1*. By default, all run queries are saved here.

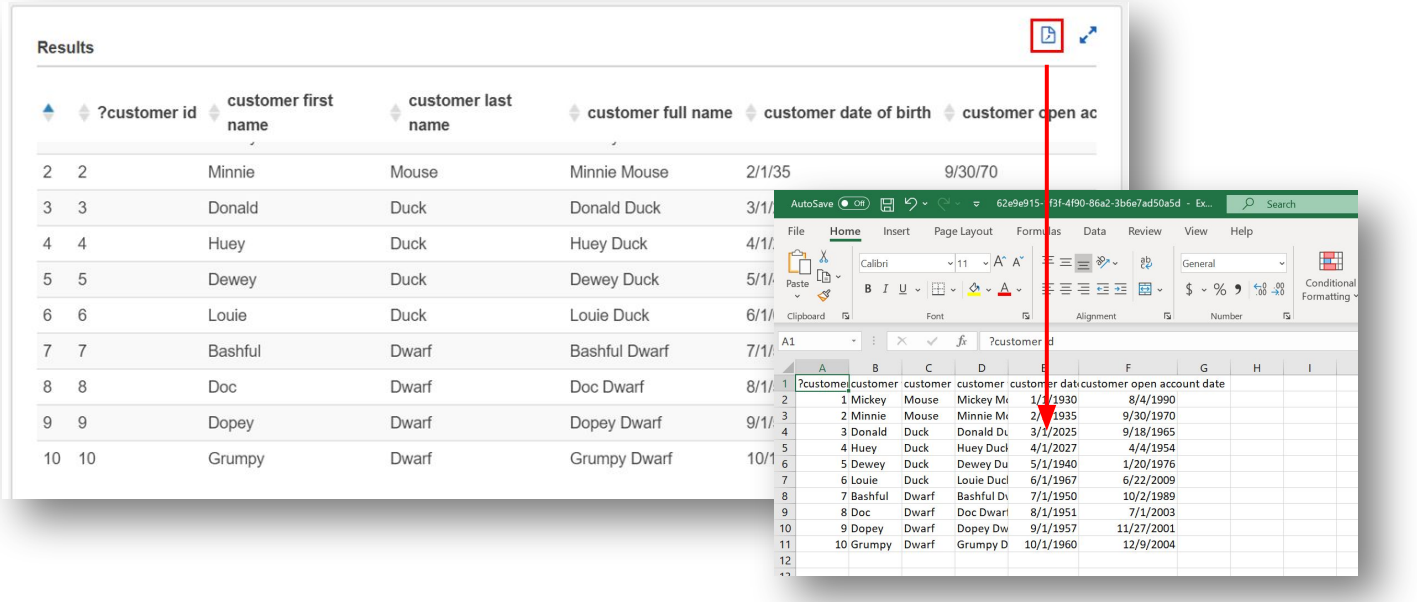


Note: Query results are purged from the playground after 45 days.

Exercise 1: Query the data lake and save the results (cont.)

Task 3 – Save the results to an S3 bucket (cont.)

2. Back in **Athena,** under **Results**, click the **Save** icon in the upper right corner to save the results locally as a .csv file.



THIS COMPLETES EXERCISE 1.

Continue to **Exercise 2: Report against the data lake with Tableau and save the output**.

EXERCISE 2

Report against the data lake with Tableau and save the output.
Task 1 – Log in to Tableau.
Task 2 – Create a workbook using the Analytics Playground Training data set.
Task 3 – Save the workbook.

Exercise 2: Report against the data lake with Tableau.

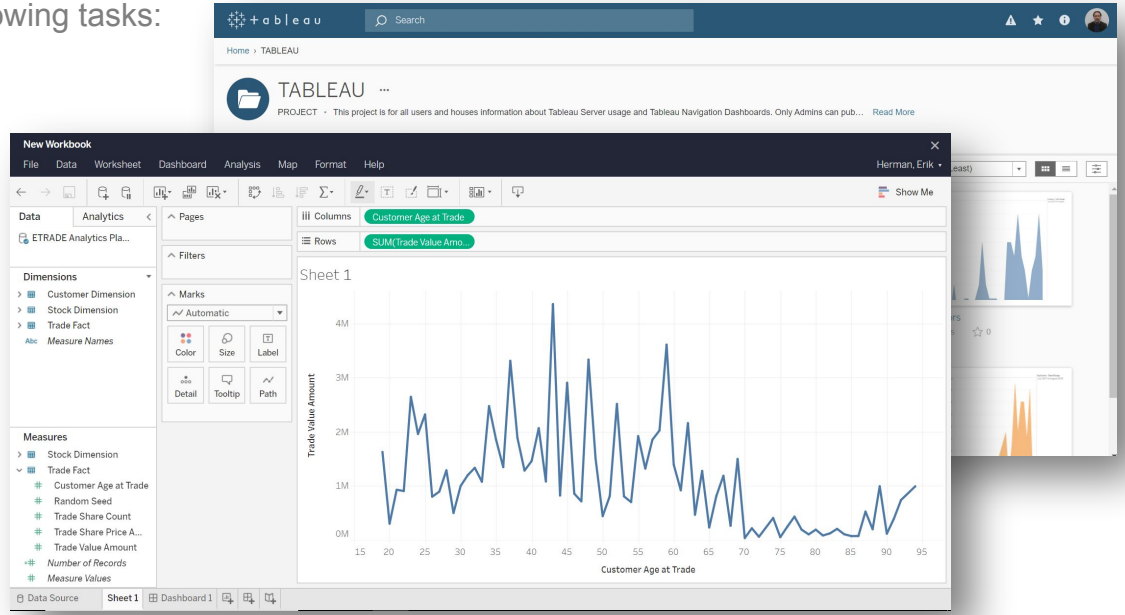
Exercise overview:

This exercise reviews best practices for reporting against the data lake using Tableau and saving the results.

Exercise objectives:

In this exercise, we will accomplish the following tasks:

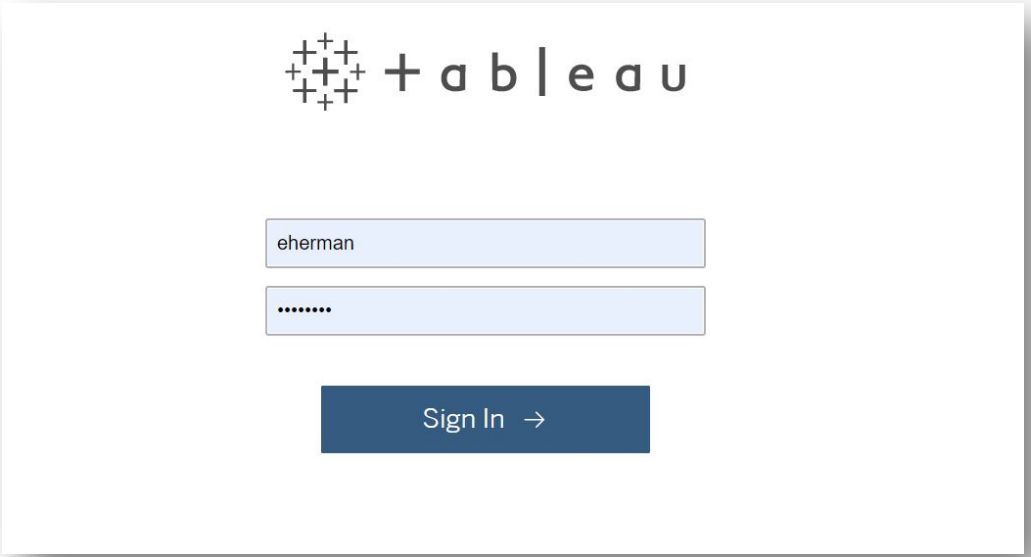
Task 1 – Log in to Tableau.
Task 2 – Create a workbook.
Task 3 – Save the workbook.



Exercise 2: Report against the data lake (cont.)

Task 1 – Log in to Tableau.

1. Log in to **Tableau** using the *link provided by your administrator*.



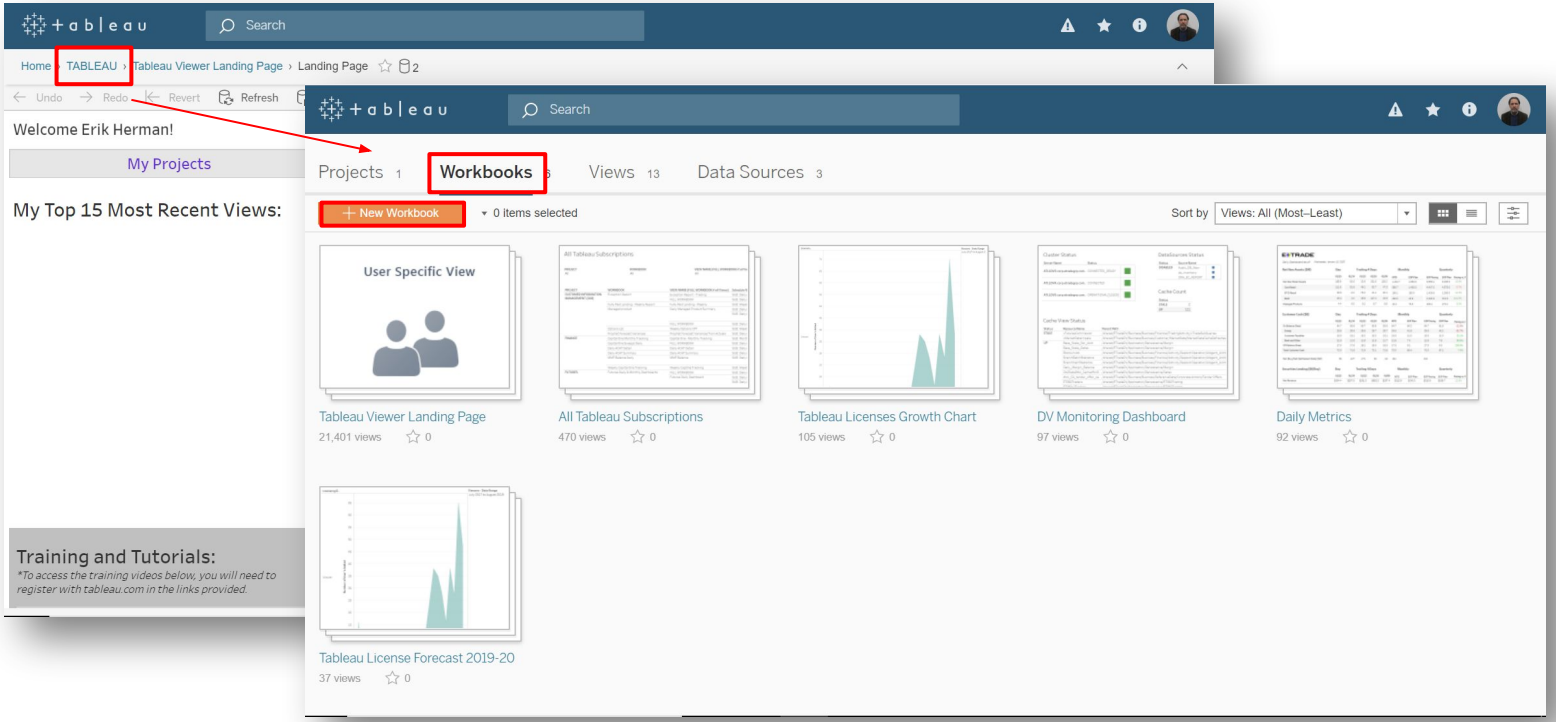
Using Tableau

- Tableau is a drag and drop visualization and dashboarding tool designed to be user friendly, as well as also feature rich and powerful. This course does not teach how to user Tableau. Additional Tableau training materials are available in the Appendix and available for bulk purchase through Safari Bookshelf.
- In this example, we create two simple visualizations to get familiar with the data we'll use in the following predictive model.

Exercise 2: Report against the data lake (cont.)

Task 1 – Log in to Tableau (cont.)

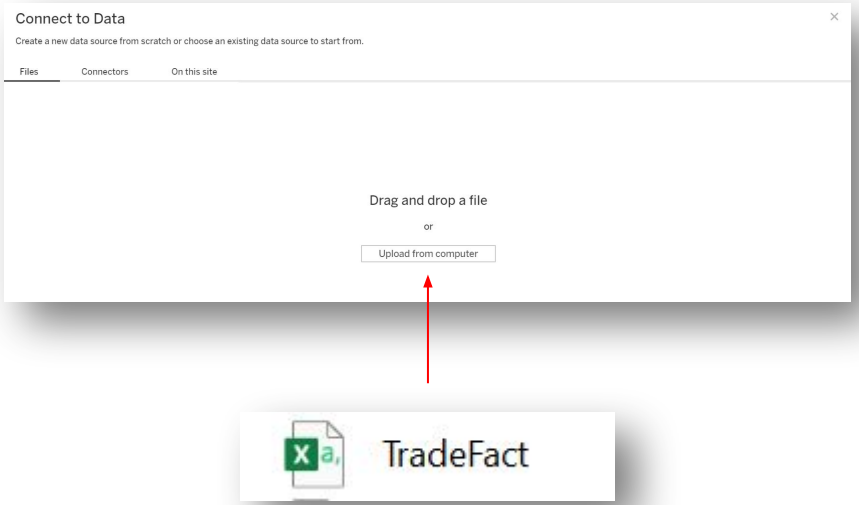
2. From the welcome screen, click **Tableau** and then under **Workbooks** click **New Workbook**.



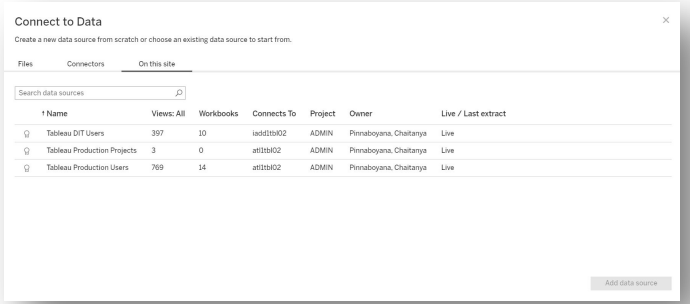
Exercise 2: Report against the data lake (cont.)

Task 1 – Log in to Tableau (cont.)

3. Drag and drop the TradeFact.csv file created and exported in the previous example .



Note: Users may connect to the trade fact table by directly accessing it via the s3 location or by uploading the exported csv file (as shown in this example).

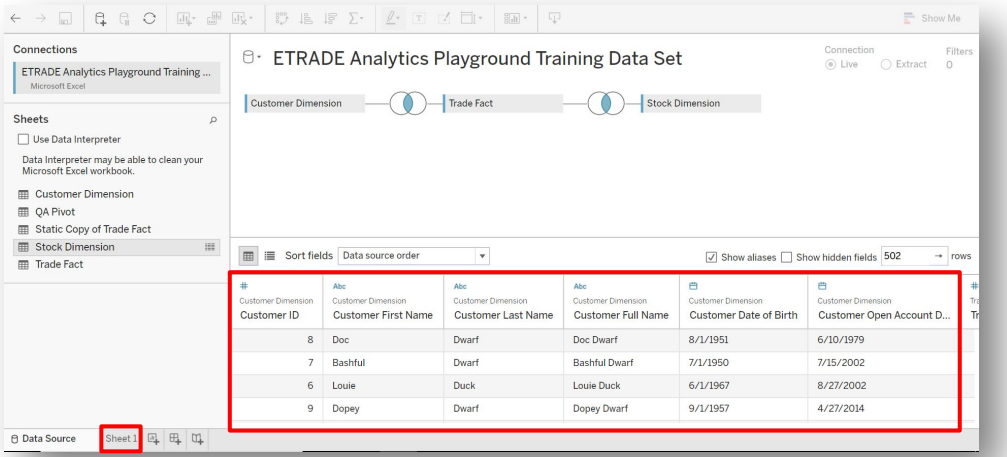


E*TRADE

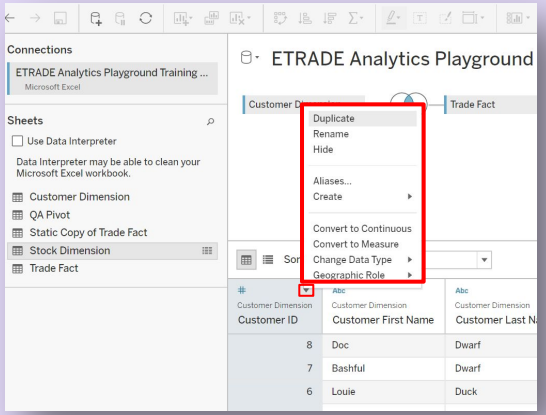
Exercise 2: Report against the data lake (cont.)

Task 1 – Log in to Tableau (cont.)

4. Review the **Data Source** tab and then click on **Sheet 1** to create a visualization.



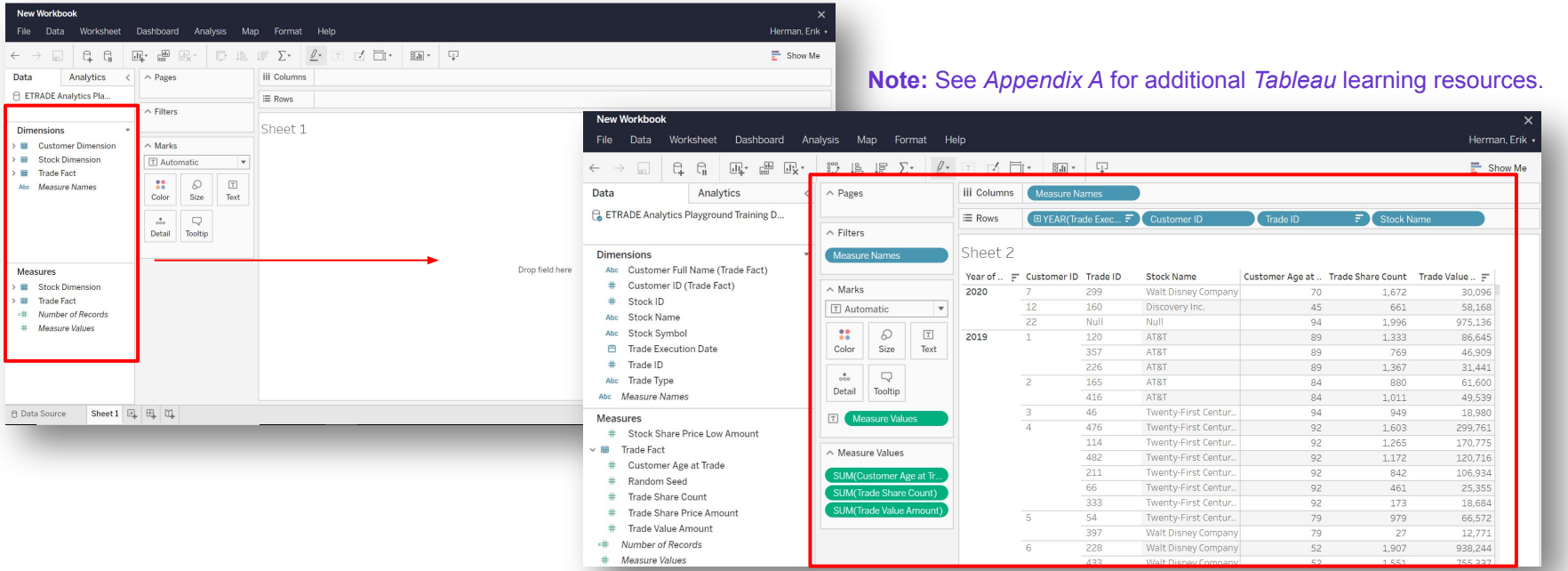
Note: Light data transformation can be handled on the *Data Source* tab, including changing data types, table/column names, create calculations, etc.



Exercise 2: Report against the data lake (cont.)

Task 2 – Create a workbook using the Analytics Playground Training data set.

1. Populate **Sheet 1** with items from the **Data** tab as shown below. Note, these are data items used in the next exercise when we create a linear regression model analyzing the relationships between customer age and trading behavior.



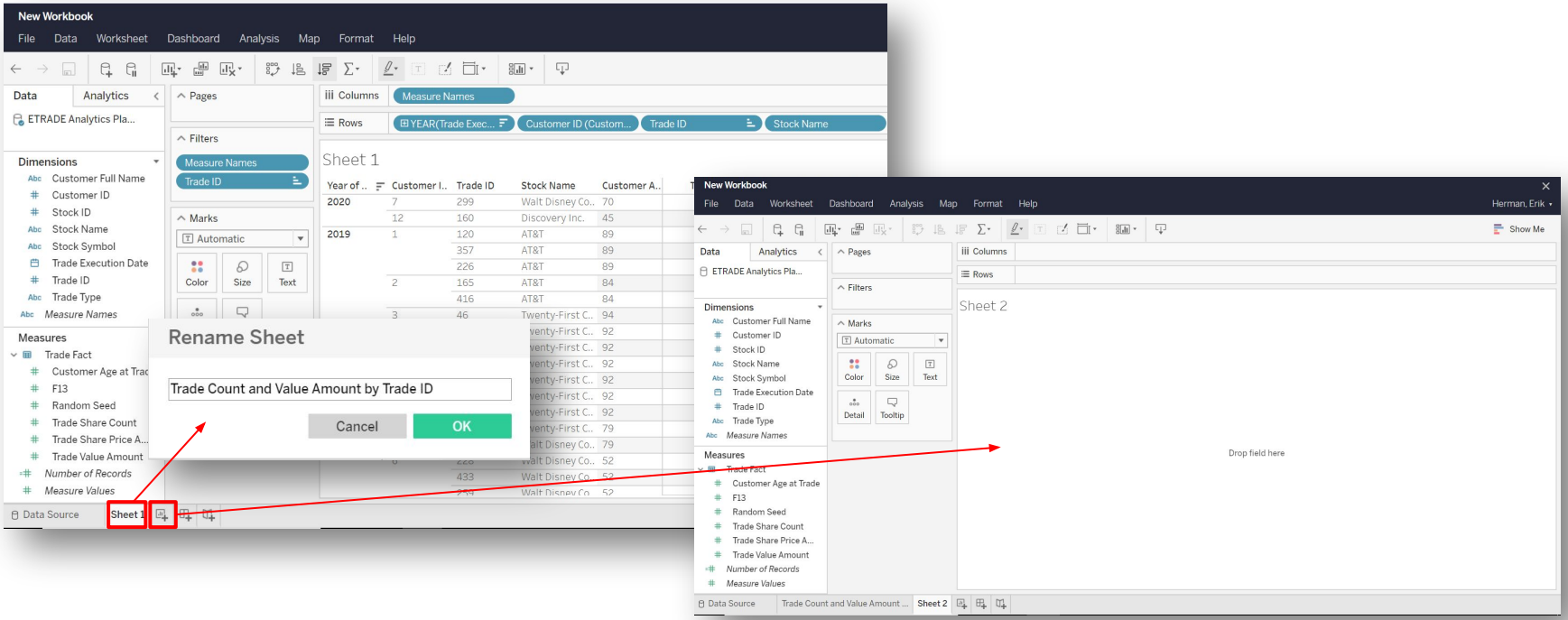
Note: See *Appendix A* for additional *Tableau* learning resources.

E*TRADE

Exercise 2: Report against the data lake (cont.)

Task 2 – Create a workbook using the Analytics Playground Training data set (cont.)

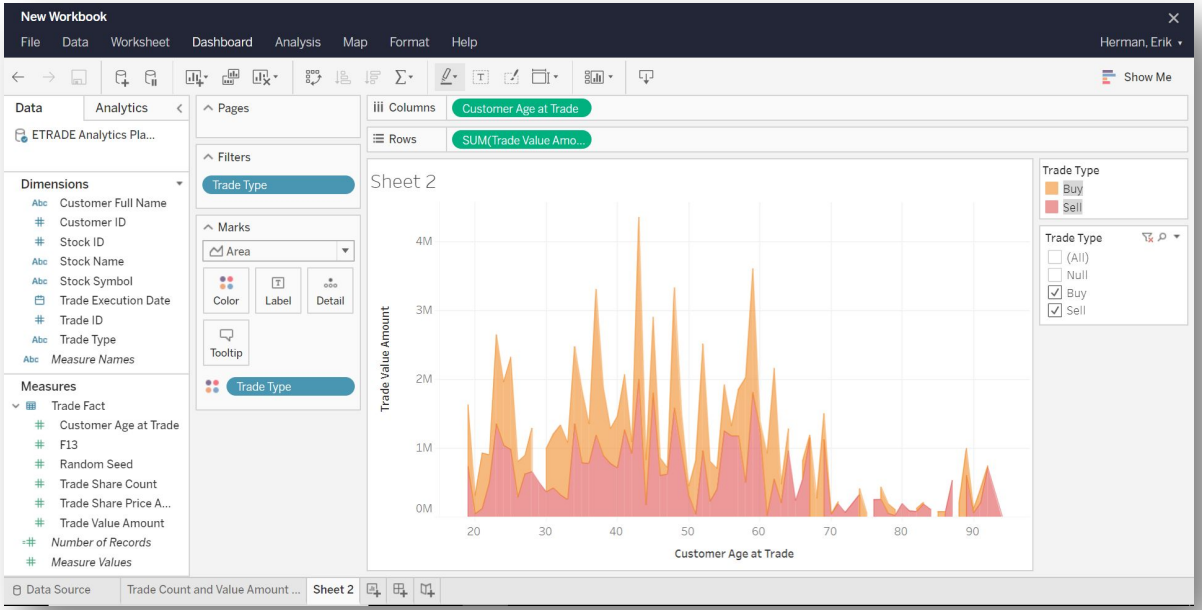
2. At the bottom of the workbook, double click **Sheet 1** and rename it **Trade Count and Amount by Trade ID**, then click **New Worksheet**.



Exercise 2: Report against the data lake (cont.)

Task 2 – Create a workbook using the Analytics Playground Training data set (cont.)

2. On the new worksheet, create a chart showing **Customer Age** by **Trade Type** and **Trade Value Amount** (as shown below).

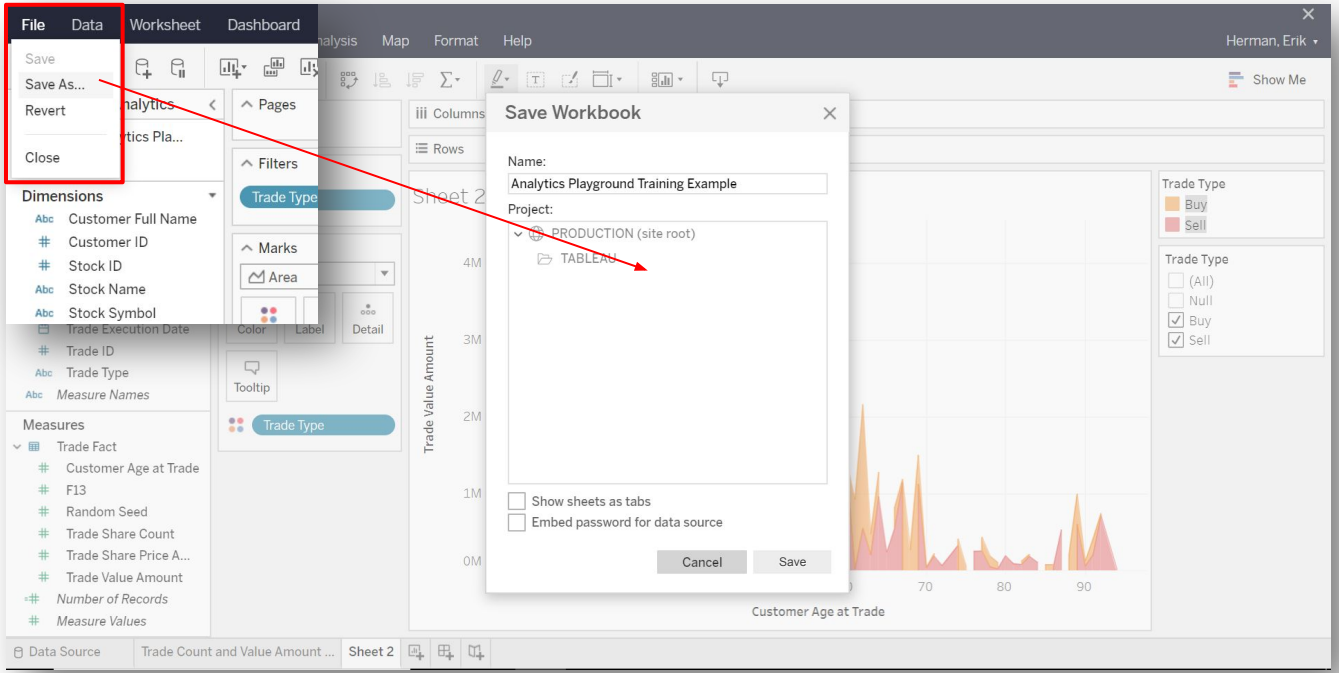


Note: Looking at this visualization we can already see patterns of buy/sell behavior based on customer age. A linear regression model will allow us to look at those patterns at a more granular level, as well as create predictions for future behavior.

Exercise 2: Report against the data lake (cont.)

Task 3 – Save the report.

1. From the **File** menu, choose **Save** or **Save as** to save the workbook.



Saving Tableau Workbooks
Add a note here about ETRADE's policy, where things get saved.

THIS COMPLETES EXERCISE 2.

Continue to **Exercise 3: Prepare and review a predictive model with SageMaker.**

EXERCISE 3

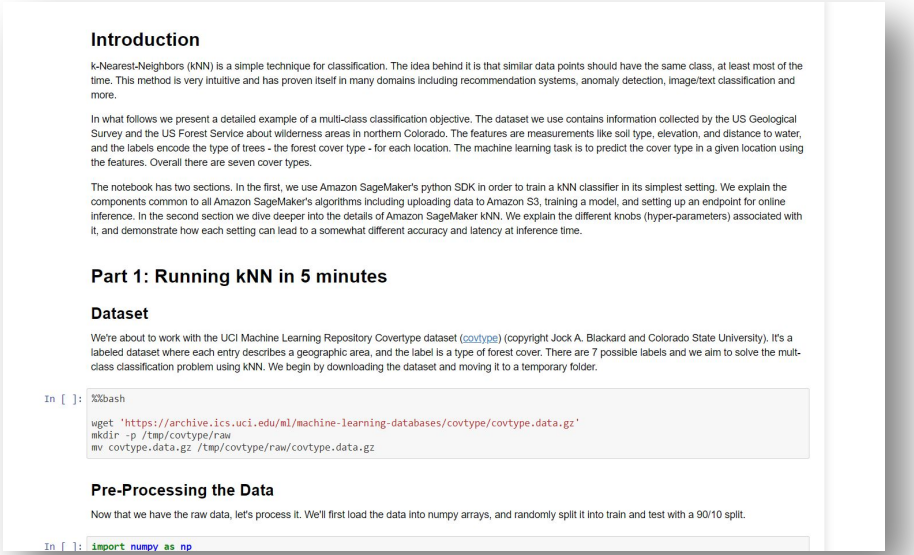
Prepare and review a predictive model with SageMaker.
Task 1 – Access the AWS console and SageMaker.
Task 2 – Create a notebook instance.
Task 3 – Review a notebook instance.

Prepare and review a predictive model with SageMaker.

Exercise overview:

This exercise reviews best practices for creating a predictive model against the data lake using SageMaker.

Task 1 – Access the AWS console and SageMaker.
Task 2 – Create a notebook instance.
Task 3 – Review a notebook instance.



Introduction

k-Nearest-Neighbors (kNN) is a simple technique for classification. The idea behind it is that similar data points should have the same class, at least most of the time. This method is very intuitive and has proven itself in many domains including recommendation systems, anomaly detection, image/text classification and more.

In what follows we present a detailed example of a multi-class classification objective. The dataset we use contains information collected by the US Geological Survey and the US Forest Service about wilderness areas in northern Colorado. The features are measurements like soil type, elevation, and distance to water, and the labels encode the type of trees - the forest cover type - for each location. The machine learning task is to predict the cover type in a given location using the features. Overall there are seven cover types.

The notebook has two sections. In the first, we use Amazon SageMaker's python SDK in order to train a kNN classifier in its simplest setting. We explain the components common to all Amazon SageMaker's algorithms including uploading data to Amazon S3, training a model, and setting up an endpoint for online inference. In the second section we dive deeper into the details of Amazon SageMaker kNN. We explain the different knobs (hyper-parameters) associated with it, and demonstrate how each setting can lead to a somewhat different accuracy and latency at inference time.

Part 1: Running kNN in 5 minutes

Dataset

We're about to work with the UCI Machine Learning Repository Covertype dataset (covtype) (copyright Jock A. Blackard and Colorado State University). It's a labeled dataset where each entry describes a geographic area, and the label is a type of forest cover. There are 7 possible labels and we aim to solve the multi-class classification problem using kNN. We begin by downloading the dataset and moving it to a temporary folder.

In [ ]:
```
%%bash

wget 'https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.data.gz'
mkdir -p /tmp/covtype/raw
mv covtype.data.gz /tmp/covtype/raw/covtype.data.gz
```

Pre-Processing the Data

Now that we have the raw data, let's process it. We'll first load the data into numpy arrays, and randomly split it into train and test with a 90/10 split.

In [ ]:
```
import numpy as np
```

Exercise 3: Prepare and review a predictive model (cont.)

The following are *notebook conditions* E*TRADE applies to all SageMaker notebooks:

Notebook Instance Settings

- region: you must be in the N. Virginia (us-east-1) region
- root-access: must be disabled
- instance-type: currently only allows ml.t2.medium

Permissions and encryption

- custom-role-arn: arn:aws:iam::<account>:role/service-role/SageMaker-<username>
  - ex: arn:aws:iam::447222439265:role/service-role/SageMaker-bcolema2
- encryption key: you must choose etr-t3-alias

Network

- vpc: you must select a vpc
- subnet: for sandbox 13 choose either available subnet:
  - subnet-etrade-Private-app1
  - subnet-etrade-Private-app2
- direct internet access: must be disabled
- security group: you must choose sagemaker-notebooks-sg

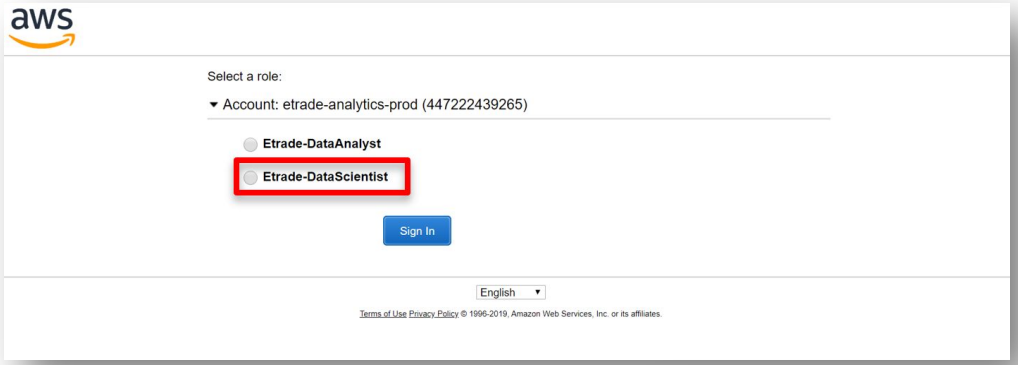
Tags

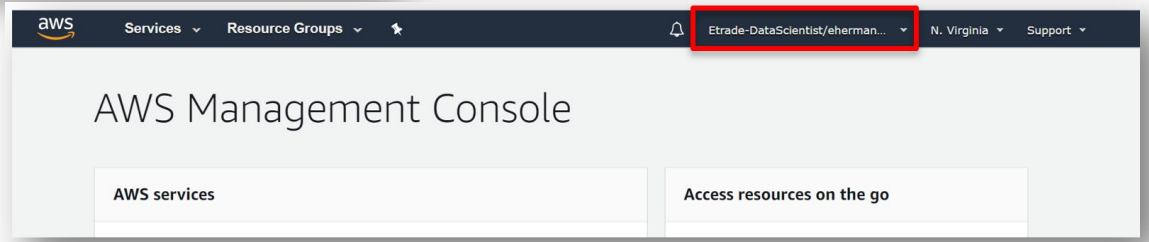
- You must create the following tag:
  - key = username
  - value = your AD username

Exercise 3: Prepare and review a predictive model (cont.)

Task 1 – Access the AWS Management Console and SageMaker

1. Click **link provided by your administrator** to login to the **AWS Management Console**.
2. Log in as **Etrade-DataScientist**.

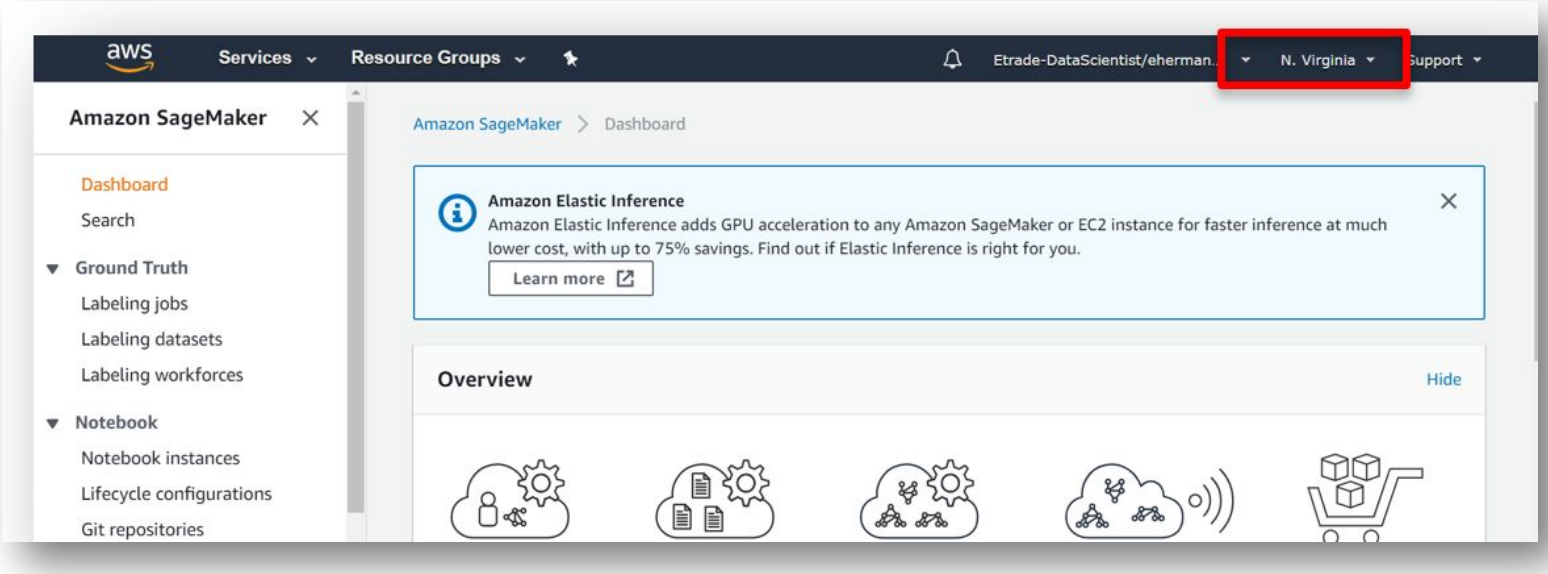




Exercise 3: Prepare and review a predictive model (cont.)

Task 1 – Access the AWS Management Console and SageMaker (cont.)

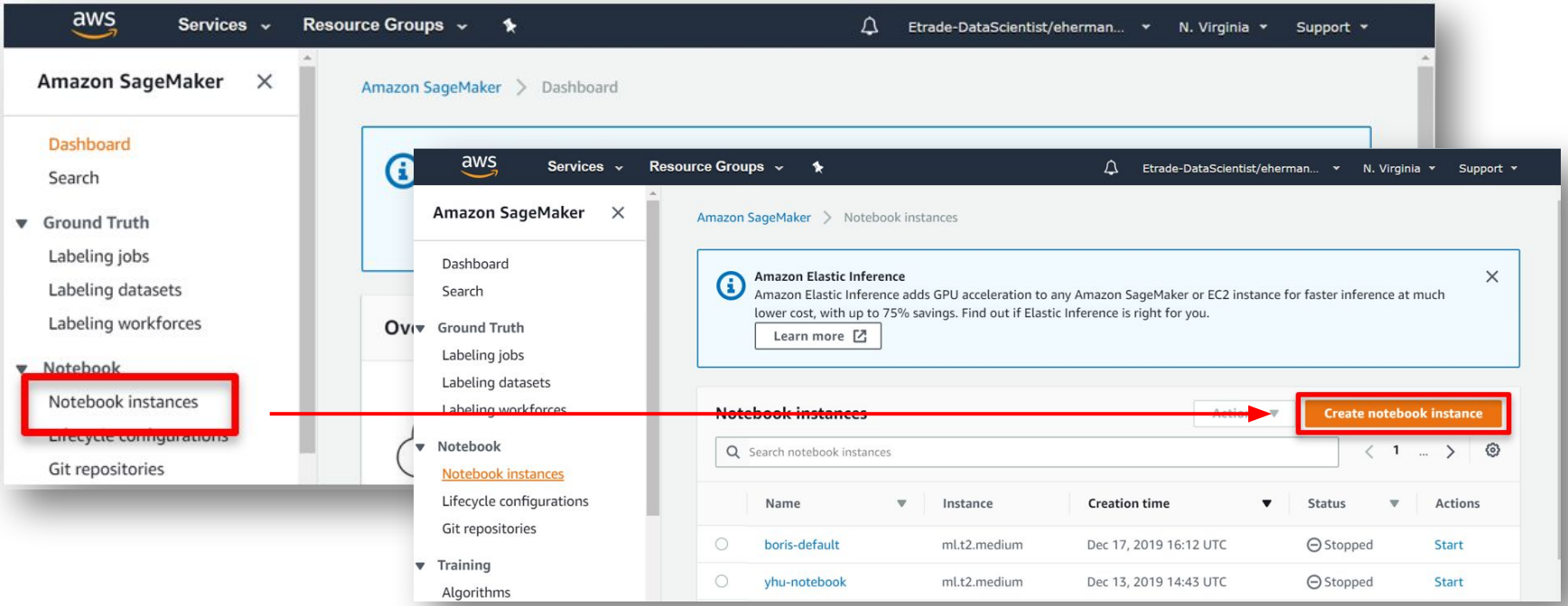
3. On the **SageMaker page**, make sure you are in **N. Virginia** region.



Exercise 3: Prepare and review a predictive model (cont.)

Task 2 – Create a Jupyter notebook with SageMaker

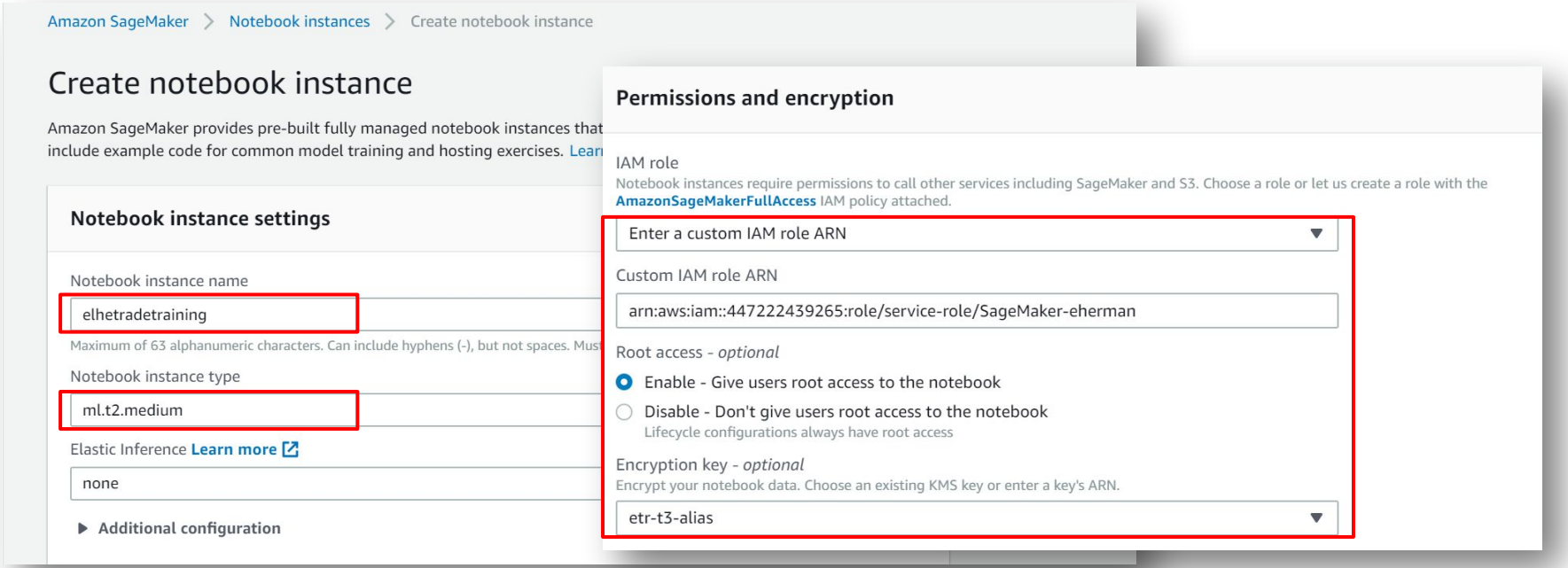
1. Click on **Notebook** instance from the menu on the left and the click **Create note instance** (on the right).



Exercise 3: Prepare and review a predictive model (cont.)

Task 2 – Create a Jupyter notebook with SageMaker (cont.)

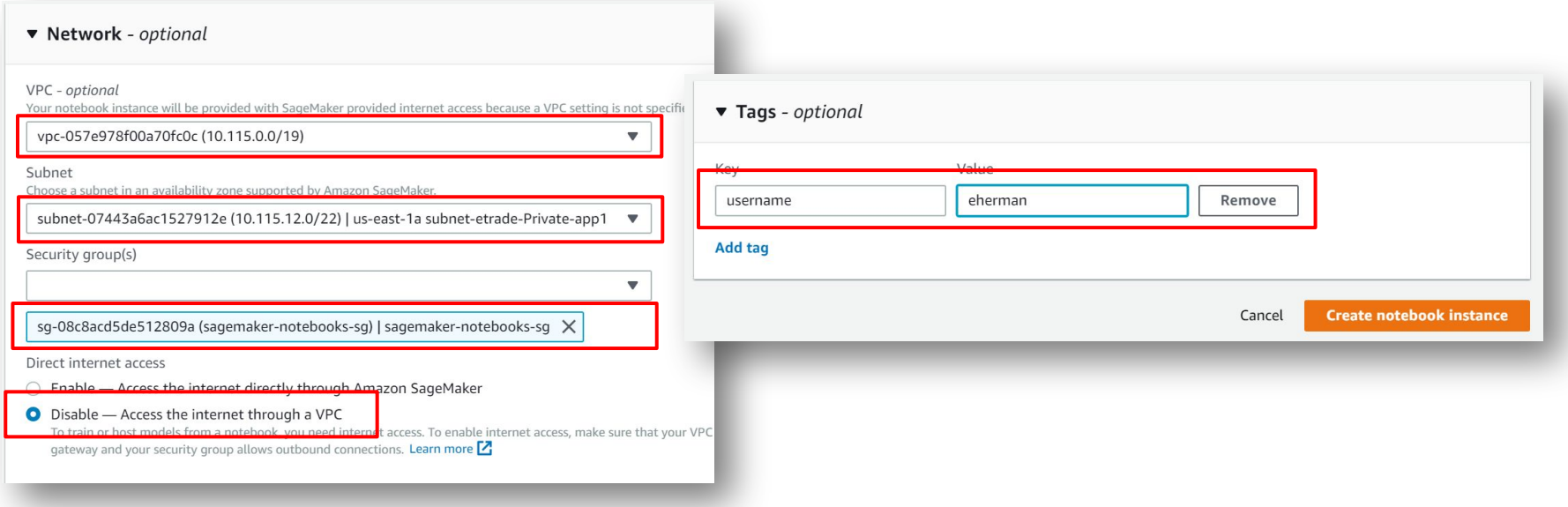
2. Configure your notebook per the *E*TRADE AWS Notebook Conditions* at the beginning of this exercise (as shown below).



Exercise 3: Prepare and review a predictive model (cont.)

Task 2 – Create a Jupyter notebook with SageMaker (cont.)

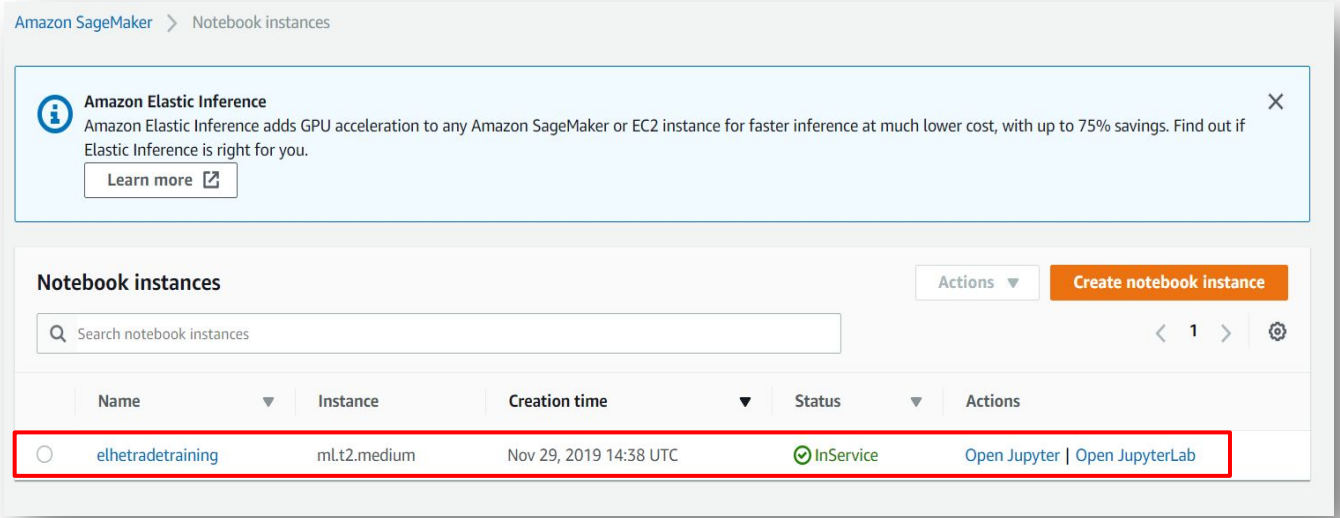
3. Configure your notebook per the E*TRADE AWS Notebook Conditions at the beginning of this exercise (as shown below) and click **Create notebook instance**.



Exercise 3: Prepare and review a predictive model (cont.)

Task 2 – Create a Jupyter notebook with SageMaker (cont.)

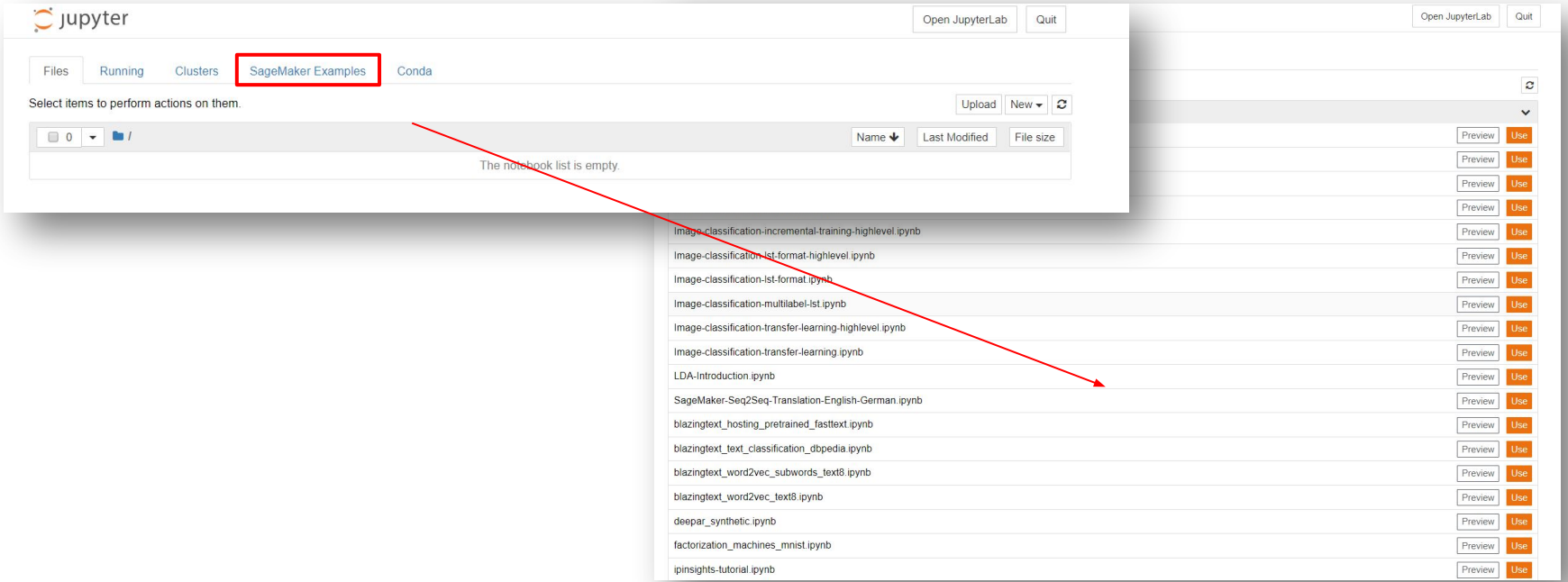
4. Once the **Status** turns to **InService**, click **Open Jupyter**.



Exercise 3: Prepare and review a predictive model (cont.)

Task 2 – Create a Jupyter notebook with SageMaker (cont.)

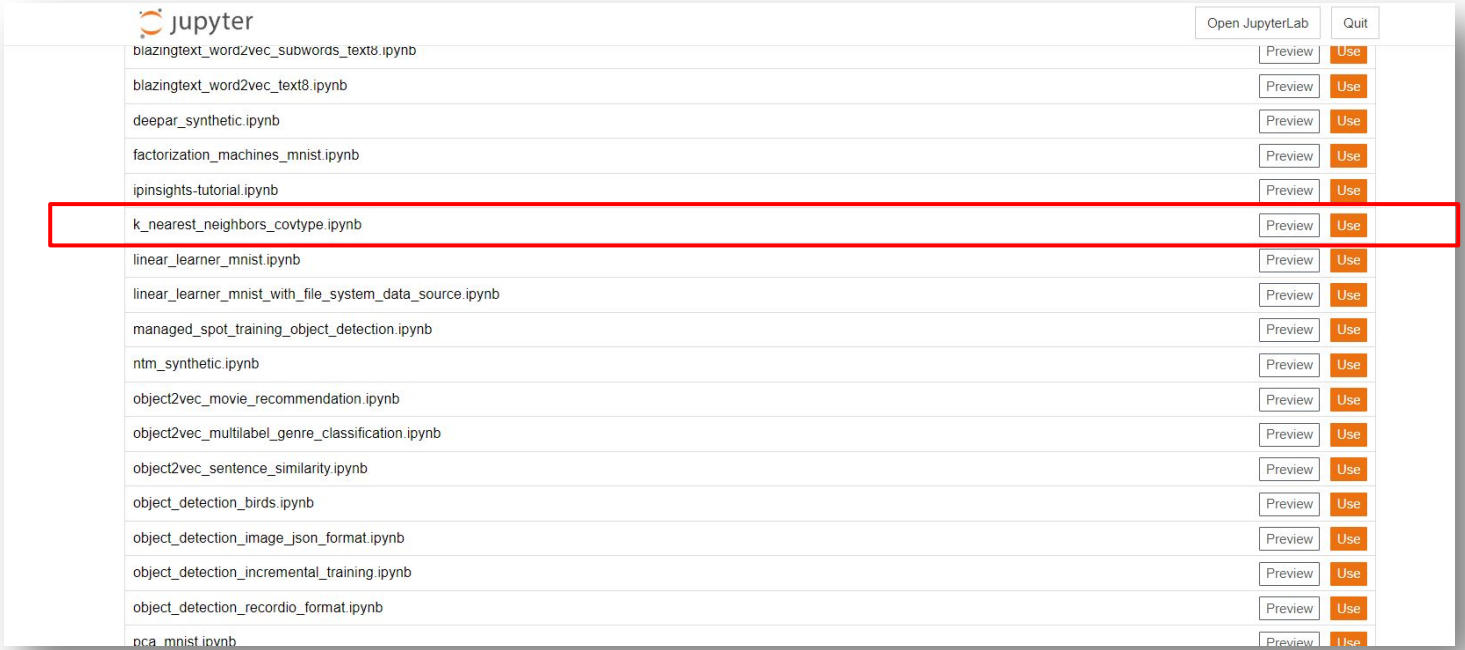
5. Once the **Jupyter notebook** opens, click the **SageMaker Examples** tab.



Exercise 3: Prepare and review a predictive model (cont.)

Task 2 – Create a Jupyter notebook with SageMaker (cont.)

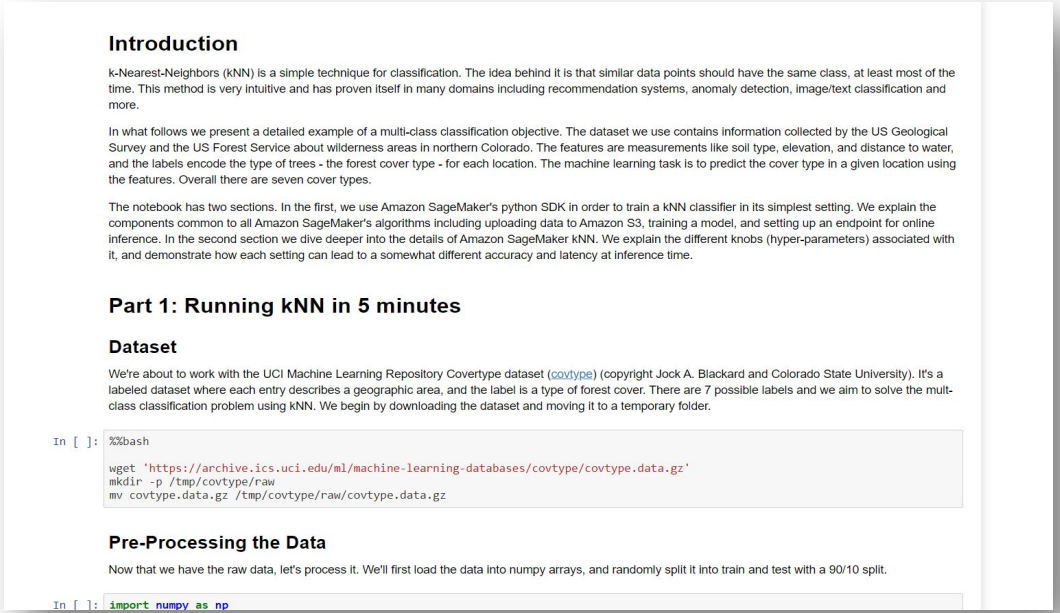
6. From the available examples, find **k_nearest_neighbors_covtype.ipynb** and click **Preview**.



Exercise 3: Prepare and review a predictive model (cont.)

Task 2 – Create a Jupyter notebook with SageMaker (cont.)

2. The examples provided by SageMaker are very valuable. Work through this document, revising it to look at the analytics playground training data set.



Introduction

k-Nearest-Neighbors (kNN) is a simple technique for classification. The idea behind it is that similar data points should have the same class, at least most of the time. This method is very intuitive and has proven itself in many domains including recommendation systems, anomaly detection, image/text classification and more.

In what follows we present a detailed example of a multi-class classification objective. The dataset we use contains information collected by the US Geological Survey and the US Forest Service about wilderness areas in northern Colorado. The features are measurements like soil type, elevation, and distance to water, and the labels encode the type of trees - the forest cover type - for each location. The machine learning task is to predict the cover type in a given location using the features. Overall there are seven cover types.

The notebook has two sections. In the first, we use Amazon SageMaker's python SDK in order to train a kNN classifier in its simplest setting. We explain the components common to all Amazon SageMaker's algorithms including uploading data to Amazon S3, training a model, and setting up an endpoint for online inference. In the second section we dive deeper into the details of Amazon SageMaker kNN. We explain the different knobs (hyper-parameters) associated with it, and demonstrate how each setting can lead to a somewhat different accuracy and latency at inference time.

Part 1: Running kNN in 5 minutes

Dataset

We're about to work with the UCI Machine Learning Repository Covertype dataset (covtype) (copyright Jock A. Blackard and Colorado State University). It's a labeled dataset where each entry describes a geographic area, and the label is a type of forest cover. There are 7 possible labels and we aim to solve the mult-class classification problem using kNN. We begin by downloading the dataset and moving it to a temporary folder.

In [ ]:
```
%%bash

wget 'https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.data.gz'
mkdir -p /tmp/covtype/raw
mv covtype.data.gz /tmp/covtype/raw/covtype.data.gz
```

Pre-Processing the Data

Now that we have the raw data, let's process it. We'll first load the data into numpy arrays, and randomly split it into train and test with a 90/10 split.

In [ ]:
```
import numpy as np
```

THIS COMPLETES EXERCISE 3.

This also completes the course **Enterprise Data Architecture and the Analytics Playground Overview and Best Practices.**

APPENDIX A

Analytics Fundamentals Resources

Appendix A - Analytics Fundamentals Resources

Note: Training materials for tools referenced in this training may be bulk purchased via Safari Bookshelf.

Amazon Athena
https://aws.amazon.com/athena/
https://docs.aws.amazon.com/athena/latest/ug/getting-started.html
AWS
https://en.wikipedia.org/wiki/Amazon_Web_Services
https://aws.amazon.com/
https://aws.amazon.com/training/
https://confluence.corp.etradegrp.com/display/IE/AWS+Training+and+Certification
AWS SageMaker
https://en.wikipedia.org/wiki/Amazon_SageMaker
https://aws.amazon.com/sagemaker
https://www.youtube.com/playlist?list=PLhr1KZpdzukcOr_6j_zmSrvYnLUtgqsZz&sc_icampaign=YT_deep-dive&sc_icontent=awssm-2747&sc_iplace=console-sagemaker-learning
Collibra
https://www.collibra.com
https://university.collibra.com

Appendix A - Analytics Fundamentals Resources (cont.)

Note: Training materials for tools referenced in this training may be bulk purchased via Safari Bookshelf.

Metadata modeling
https://en.wikipedia.org/wiki/Metadata_modeling
SQL
https://en.wikipedia.org/wiki/SQL
Machine learning
https://en.wikipedia.org/wiki/Machine_learning
https://towardsdatascience.com/machine-learning-basics-part-1-a36d38c7916
https://archive.ics.uci.edu/ml/index.php
Predictive modeling
https://en.wikipedia.org/wiki/Predictive_modelling
Python
https://en.wikipedia.org/wiki/Python_(programming_language)
https://www.python.org/about/gettingstarted/
R
https://en.wikipedia.org/wiki/R_(programming_language)
https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

THANK YOU!

This concludes the training.