

ENTERPRISE DATA ARCHITECTURE AND THE ANALYTICS PLAYGROUND

Overview and best practices.

January 2020

Internal Use Only

© 2019 ETRADE Financial Corporation. All rights reserved.

This presentation contains confidential information and may not be disclosed without ETRADE Financial Corporation's written permission.

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

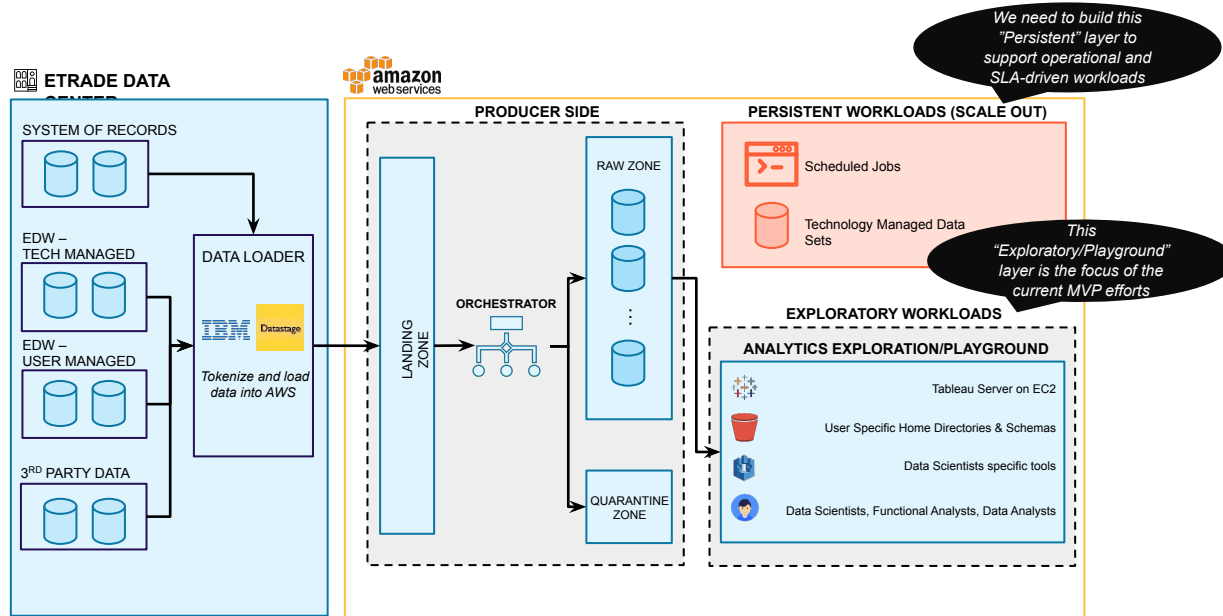
- to modernize and advance the analytics capabilities, both for retroactive reporting and (more strategically) for predictive algorithm development and refining, forecasting, and segmentation.

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.

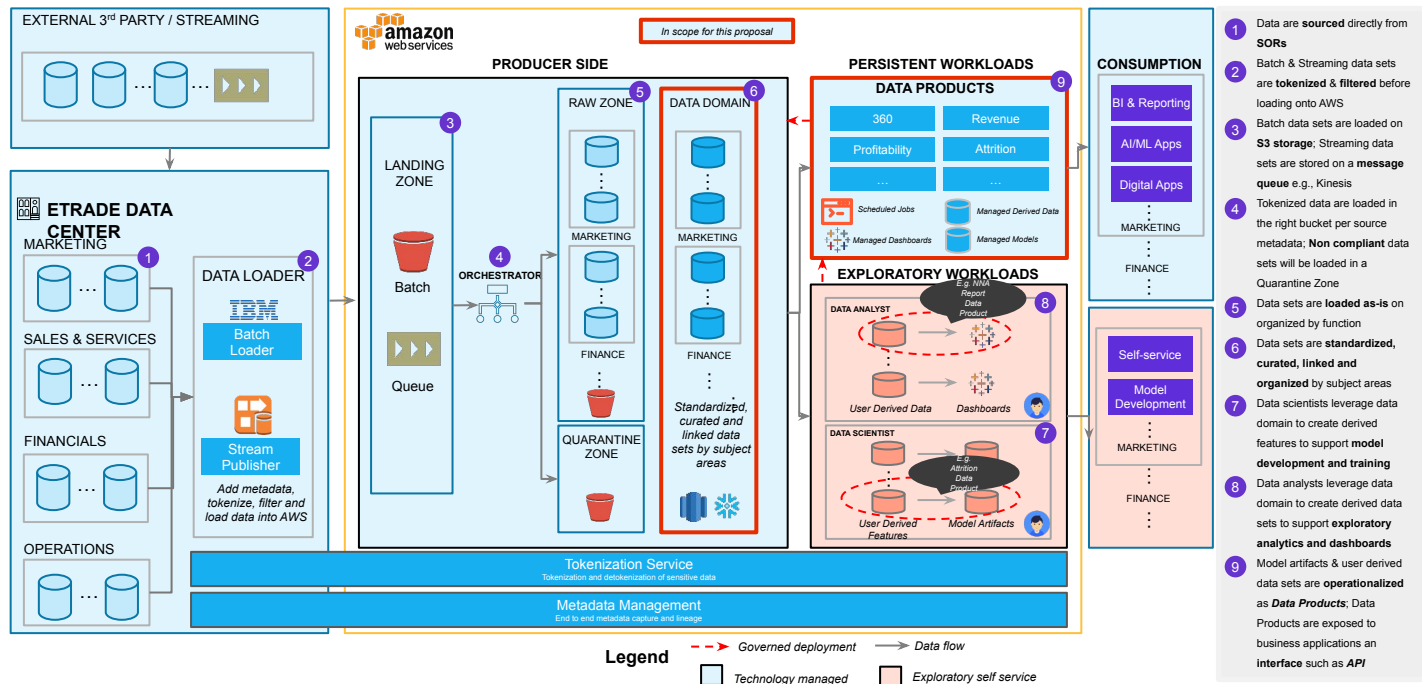
E*TRADE Enterprise *Current State* Data Architecture

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*.

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

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.

Data lake access privileges are driven by:

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

Requesting AWS roles:

<https://bmcsmrtpcorp.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.

The screenshot displays the Collibra user interface. On the left, the 'Browser' section is highlighted with a red box, showing a tree view of domains under 'E*TRADE'. A red arrow points from this section to a table in the main view. The table lists various data domains with columns for Name, Description, Domain Type, Owner, Stakeholder, and Business Steward.

Name	Description	Domain Type	Owner	Stakeholder	Business Steward
E*TRADE					
Brokerage					
Collibra Training					
Data Governance					
Finance					
Human Resources					
Legal					
Operations					
Risk					
Technology					
Data Domains		Data Domains			
E*TRADE Issues		Issues			
E*TRADE Lines of Busin...		Lines of Business			
E*TRADE's Data Quality ...		Data Quality Rules			
E*TRADE's Guiding Prin...		Guiding Principles			

Browse by community or domain.

Metadata

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

The screenshot shows the AWS Glue console interface. The left sidebar contains navigation options: Data catalog, Databases, Tables, Connections, Crawlers, Classifiers, Settings, ETL, Workflows, Jobs, ML Transforms, Triggers, Dev endpoints, and Notebooks. The main content area displays a table of metadata definitions. The search bar is empty, and the table shows various tables with columns for Name, Database, Location, Classification, Last updated, and Deprecated.

Name	Database	Location	Classification	Last updated	Deprecated
aaa_acct_tran_flg	asset_flow	s3://etr-adi-users-ome...	Unknown	14 November 2019 6:...	
aaa_final_acct	asset_flow	s3://etr-adi-users-ome...	Unknown	14 November 2019 7:...	
account_summary_daily	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
account_summary_monthly	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	5 December 2019 3:0...	
accounts_with_outflows	rbennet1	s3://etr-adi-users-ome...	Unknown	17 December 2019 6:...	
acct_12m_bal	asset_flow	s3://etr-adi-users-ome...	Unknown		
acct_12m_bal_f6	asset_flow	s3://etr-adi-users-ome...	Unknown		
acct_12m_bal_f7	asset_flow	s3://etr-adi-users-ome...	Unknown		
acct_12m_bal_f8	asset_flow	s3://etr-adi-users-ome...	Unknown		
acct_id_sample	asset_flow	s3://etr-adi-users-ome...	Unknown		
acct_status	asset_flow	s3://etr-adi-users-ome...	Unknown		
acct status final	asset flow shared	s3://etr-adi-users-ome...	Unknown		

Filter by table, database, S3 bucket, etc.

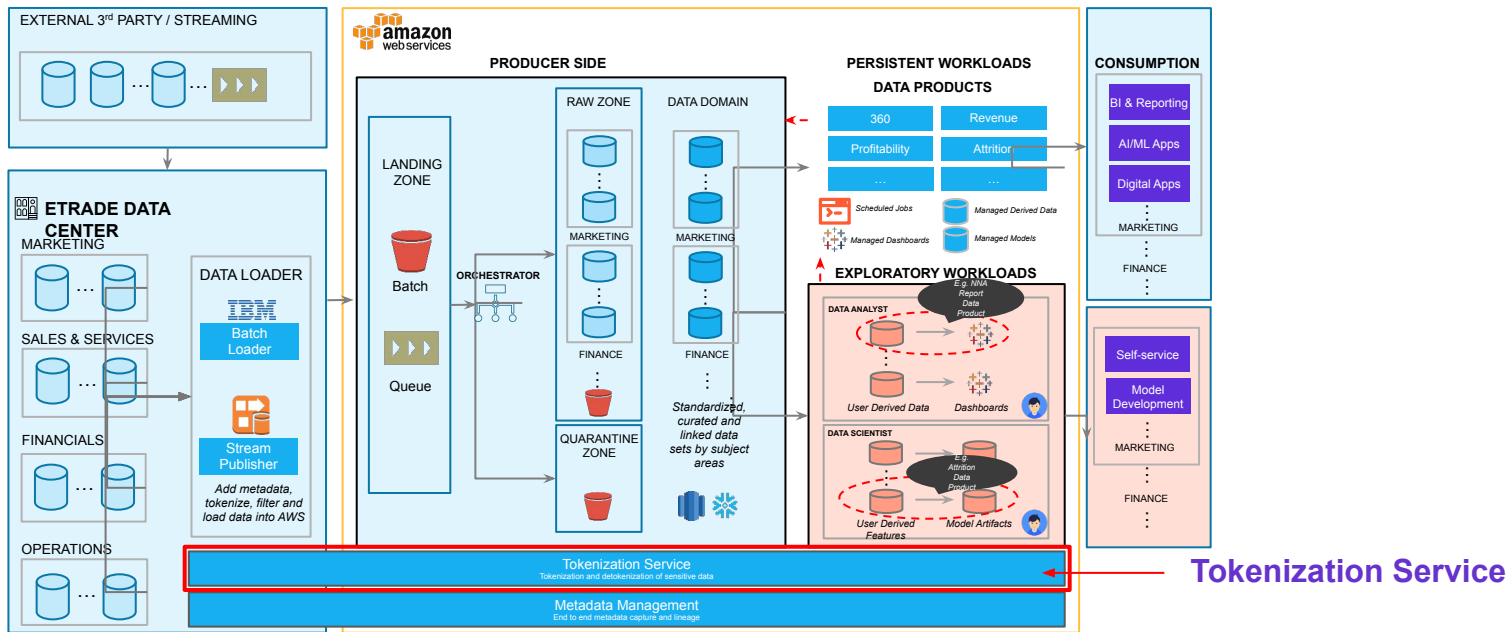
The screenshot shows the AWS Glue console with the search filter set to "Database: dw_enterprise". The search results are displayed in a table, which is highlighted with a red border. The table lists various tables from the "dw_enterprise" database, including their names, locations, classifications, and last updated dates.

Name	Database	Location	Classification	Last updated	Deprecated
account_summary_daily	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
account_summary_monthly	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	5 December 2019 3:0...	
acct_user_xtbl	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
agent_chat_activities_per_login_session	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
bank_account	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
bank_account_asset_flow_daily	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
bank_acct_transaction	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
bank_acct_type_rtbl	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
bank_external_trancode	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
bna_customer	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	20 November 2019 6:...	
brk_account_asset_flow_daily	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	
brk account asset flow weekly	dw_enterprise	s3://etr-adi-raw-omeg...	parquet	26 December 2019 9:...	

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:

AWS console

Data Lake (S3 bucket)

Available analytics tools

- 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.

Functional analyst tools:

- Cloudwatch
- S3
- Athena
- Glue
- Redshift

Data analyst

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

Data analyst tools:

- Cloudwatch
- S3
- Athena
- Glue
- Redshift

Data scientist

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

Data scientist tools:

- Cloudwatch
- S3
- SageMaker
- Athena
- Glue
- Redshift

Data engineer

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

Data 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 low 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 static and not recalculate whenever anything changes.
- This tab is provided to give you something stable to export, if ne

Trade ID	Customer ID	Customer First Name	Customer Last Name	Customer Full Name	Customer Date of Birth	Customer Open Account Date	Trade ID	Customer	Customer Full	Customer Age at Trade	Stock	Stock	Stock	Trade	Trade	Trade	
								ID	Name	Trade Execution Date	Trade	Symbol	Stock Name	Type	Share Count	Share Price	Trade Value
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	Ginemark 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						

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

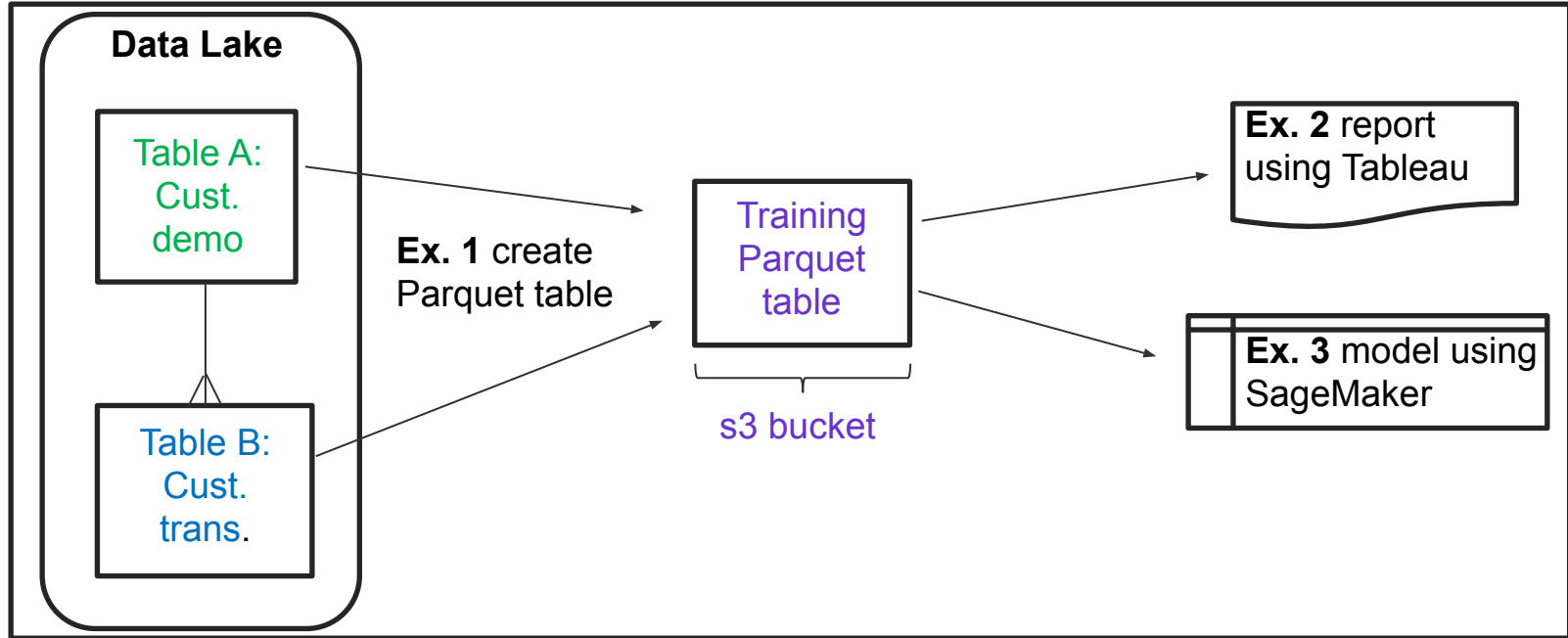
s3://etr-adl-raw-omega-ue1/ANALYTICS_TRAINING/SAMPLE_TRAINING_DATA/StockDim.csv

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.

E*TRADE Analytics Playground Environment (Prod: 447222439265)



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:

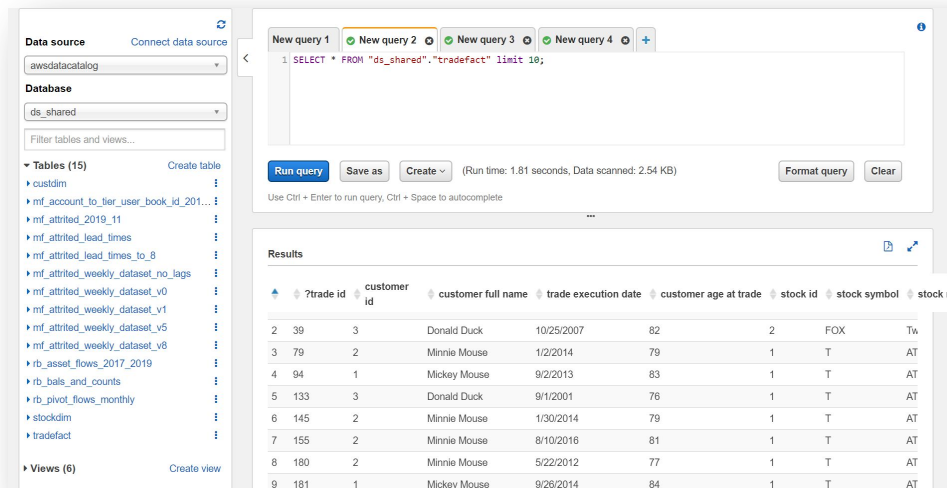
In this exercise, we will accomplish the following tasks:

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.



The screenshot displays the AWS Athena console interface. On the left, the 'Data source' is set to 'awsdatacatalog' and the 'Database' is 'ds_shared'. A list of tables is visible, including 'custdim', 'mf_account_to_tier_user_book_id_201...', 'mf_atrtrid_2019_11', 'mf_atrtrid_lead_times', 'mf_atrtrid_lead_times_to_8', 'mf_atrtrid_weekly_dataset_no_lags', 'mf_atrtrid_weekly_dataset_v0', 'mf_atrtrid_weekly_dataset_v1', 'mf_atrtrid_weekly_dataset_v5', 'mf_atrtrid_weekly_dataset_v8', 'rb_asset_flows_2017_2019', 'rb_bals_and_counts', 'rb_pivot_flows_monthly', 'stockdim', and 'tradefact'. The 'Views' section shows 'stockdim' and 'tradefact'. The main area shows a query editor with the following SQL query:

```
1 SELECT * FROM "ds_shared"."tradefact" limit 10;
```

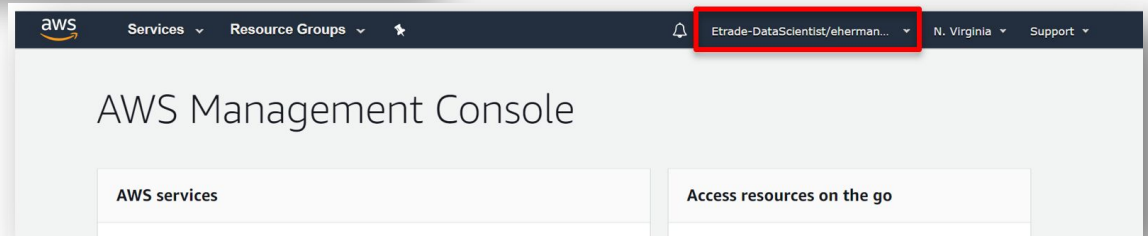
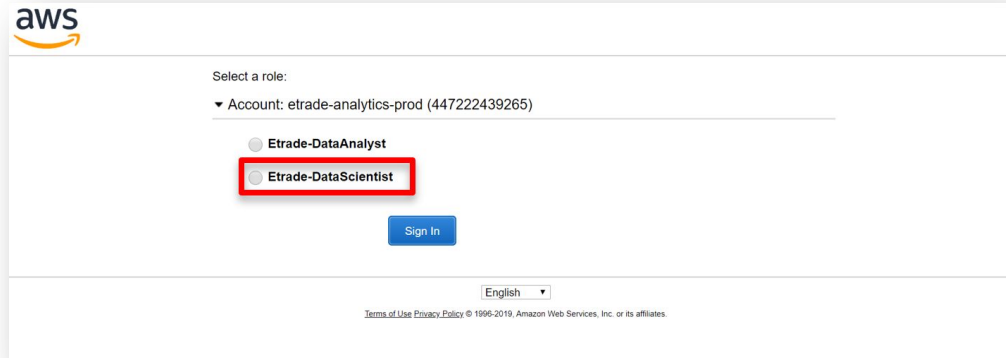
Below the query editor, there are buttons for 'Run query', 'Save as', 'Create', 'Format query', and 'Clear'. The 'Run time' is 1.81 seconds and 'Data scanned' is 2.54 KB. The 'Results' section shows a table with the following columns: 'trade id', 'customer id', 'customer full name', 'trade execution date', 'customer age at trade', 'stock id', 'stock symbol', and 'stock i'.

	trade id	customer id	customer full name	trade execution date	customer age at trade	stock id	stock symbol	stock i
2	39	3	Donald Duck	10/25/2007	82	2	FOX	Tw
3	79	2	Minnie Mouse	1/2/2014	79	1	T	AT
4	94	1	Mickey Mouse	9/2/2013	83	1	T	AT
5	133	3	Donald Duck	9/1/2001	76	1	T	AT
6	145	2	Minnie Mouse	1/30/2014	79	1	T	AT
7	155	2	Minnie Mouse	8/10/2016	81	1	T	AT
8	180	2	Minnie Mouse	5/22/2012	77	1	T	AT
9	181	1	Mickey Mouse	9/26/2014	84	1	T	AT

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

Task 1 – Access the AWS Management Console and Athena.

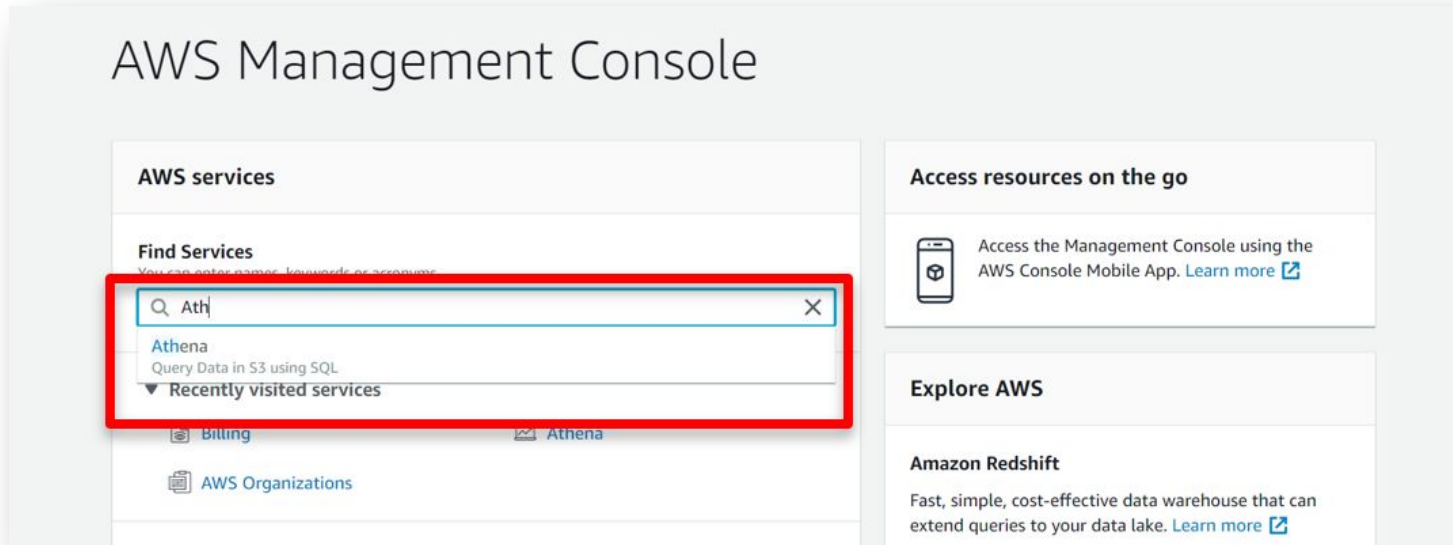
1. Click link provided by your administrator to login to the **AWS** management console.
2. Log in as **Etrade-DataScientist**.



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.

The screenshot shows the AWS Athena Query Editor interface. A notification banner at the top states: "Before you run your first query, you need to set up a query result location in Amazon S3. Learn more". The interface includes a left sidebar with "Data source" (awsdatacatalog) and "Database" (aff_usr). The main area shows a SQL query editor with a "New query 1" tab. Below the editor are buttons for "Run query", "Save as", and "Create". A "Settings" dialog box is open, showing "Settings apply by default to all new queries. Learn more" and "Workgroup: primary". The "Query result location" field is highlighted with a red box and contains the text "s3://etr-adr-raw-omega-ue1/playgroundtraining/". Below this field is an example: "Example: s3://query-results-bucket/folder/". There are also checkboxes for "Encrypt query results" and "Autocomplete".

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.

The screenshot shows the AWS Athena console interface. At the top, the navigation bar includes 'Services', 'Resource Groups', and a user profile 'Etrade-DataScientist/eherman...'. The main header shows 'Athena' and several tabs: 'Query Editor', 'Saved Queries', 'History', 'Data sources', and 'Workgroup : primary' (which is highlighted with a red box). Below the header, there are buttons for 'Create workgroup', 'View details', and 'Switch workgroup'. A table lists the workgroups with columns for Name, Description, Creation time, and Workgroup status. The 'DataScientist' row is highlighted with a red box, and a red arrow points from the 'Workgroup : primary' tab to this row.

	Name	Description	Creation time	Workgroup status
<input type="radio"/>	DataEngineer		2019/10/21 10:19:57 UTC-4	Enabled
<input type="radio"/>	DataAnalyst		2019/10/21 10:19:56 UTC-4	Enabled
<input checked="" type="radio"/>	DataScientist		2019/10/21 10:19:56 UTC-4	Enabled
<input type="radio"/>	FunctionalAnalyst		2019/10/21 10:19:56 UTC-4	Enabled
<input type="radio"/>	primary		2019/10/08 14:15:32 UTC-4	Enabled

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.)

The screenshot displays the Athena Query Editor interface. The top navigation bar includes 'Athena', 'Query Editor', 'Saved Queries', 'History', 'Data sources', 'Workgroup : DataScientist', 'Settings', 'Tutorial', 'Help', and 'What's new'. The 'Data source' is set to 'awsdatacatalog'. The 'Database' dropdown is set to 'ds_shared'. The 'Tables (15)' list is expanded to show 'custdim', with its schema details visible: '?customer id (bigint)', 'customer first name (string)', 'customer last name (string)', 'customer full name (string)', 'customer date of birth (string)', and 'customer open account date (string)'. A context menu is open over the 'custdim' table, with 'Preview table' selected. The query editor contains the SQL: `SELECT * FROM "ds_shared"."custdim" limit 10;`. The 'Results' section shows a table with 10 rows of customer data.

	?customer id	customer first name	customer last name	customer full name	customer date of birth	customer open ac
2	2	Minnie	Mouse	Minnie Mouse	2/1/35	9/30/70
3	3	Donald	Duck	Donald Duck	3/1/25	9/18/65
4	4	Huey	Duck	Huey Duck	4/1/27	4/4/54
5	5	Dewey	Duck	Dewey Duck	5/1/26	4/20/76

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.

The screenshot shows the AWS S3 console interface. At the top, the breadcrumb navigation path is highlighted with a red box: Amazon S3 > etr-adl-raw-omega-ue1 > playgroundtraining > Unsaved > 2020 > 01 > 20. Below this, the bucket name 'etr-adl-raw-omega-ue1' is displayed. A search bar and action buttons (Upload, Create folder, Download, Actions) are visible. A red arrow points from the 'Actions' button to a table of objects. The table is also highlighted with a red box and contains the following data:

Name	Last modified	Size	Storage class
580fb808-3da8-4f99-a0a0-2a7be2b22a29.csv	Jan 20, 2020 12:17:28 PM GMT-0500	522.7 KB	Standard
580fb808-3da8-4f99-a0a0-2a7be2b22a29.csv.metadata	Jan 20, 2020 12:17:28 PM GMT-0500	166.0 B	Standard

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.)

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

The screenshot shows the Athena Results console with a table of customer data. A red box highlights the 'Save' icon in the top right corner. An Excel spreadsheet is overlaid on the bottom right, showing the data from the table in a grid format.

customer id	customer first name	customer last name	customer full name	customer date of birth	customer open ac
2	Minnie	Mouse	Minnie Mouse	2/1/35	9/30/70
3	Donald	Duck	Donald Duck	3/1/2025	9/18/1965
4	Huey	Duck	Huey Duck	4/1/2027	4/4/1954
5	Dewey	Duck	Dewey Duck	5/1/1940	1/20/1976
6	Louie	Duck	Louie Duck	6/1/1967	6/22/2009
7	Bashful	Dwarf	Bashful Dwarf	7/1/1950	10/2/1989
8	Doc	Dwarf	Doc Dwarf	8/1/1951	7/1/2003
9	Dopey	Dwarf	Dopey Dwarf	9/1/1957	11/27/2001
10	Grumpy	Dwarf	Grumpy Dwarf	10/1/1960	12/9/2004

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.

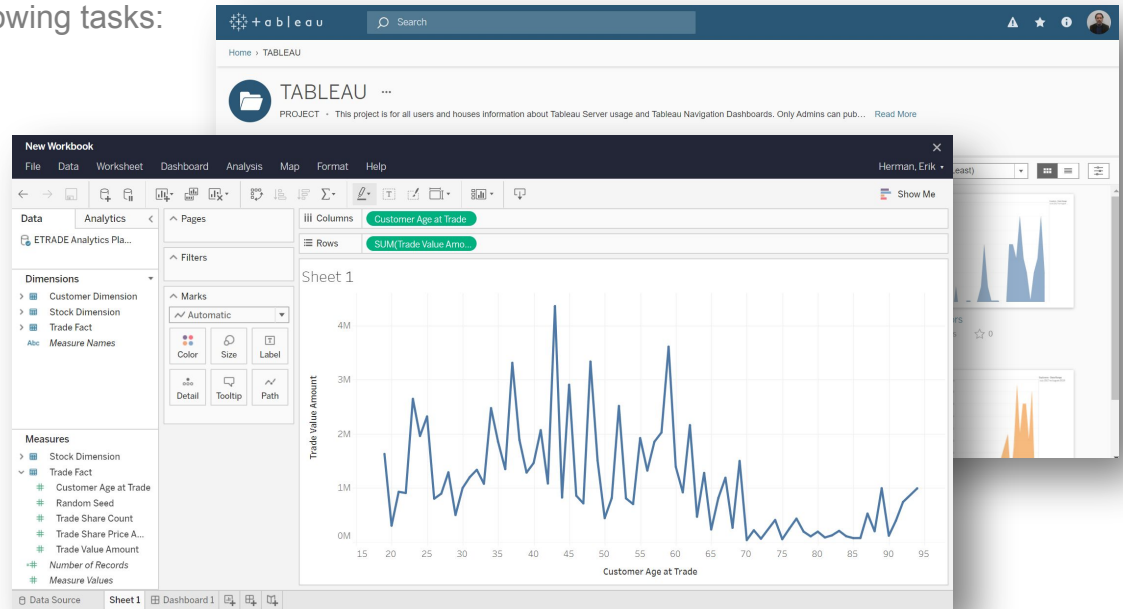
Exercise objectives:

In this exercise, we will accomplish the following tasks:

Task 1 – Log in to Tableau.

Task 2 – Create a workbook.

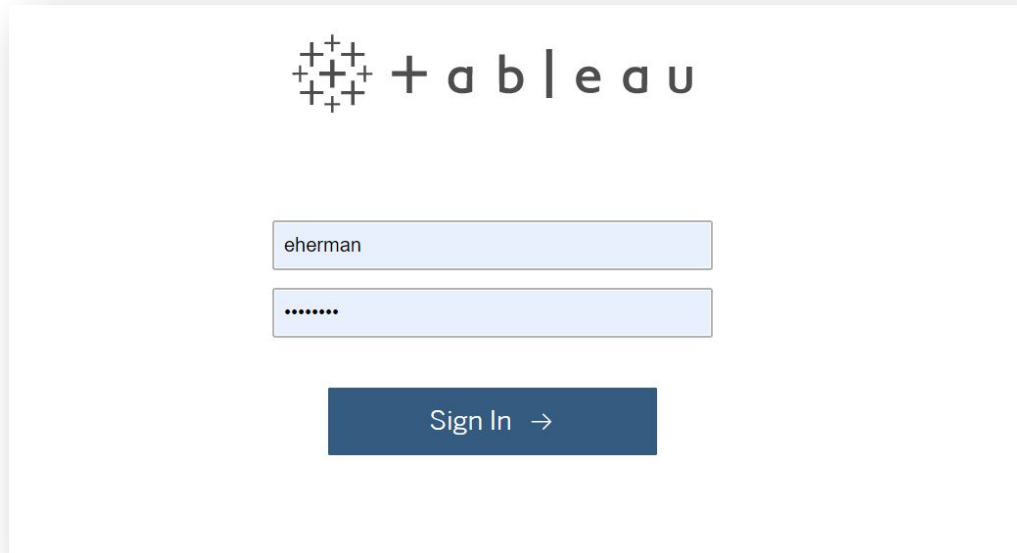
Task 3 – Save the workbook.



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*.



tableau

eherman

.....

Sign In →

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 use 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.)

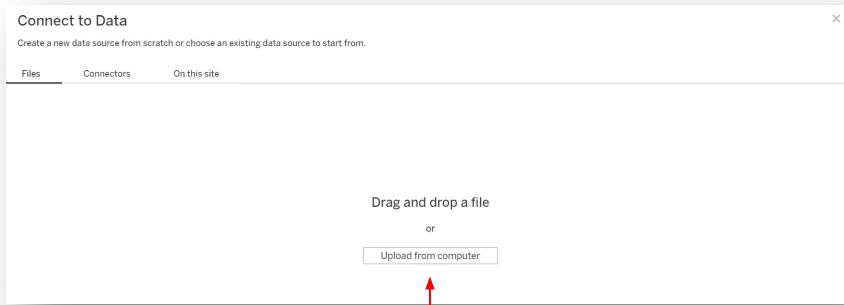
2. From the welcome screen, click **Tableau** and then under **Workbooks** click **New Workbook**.

The screenshot displays the Tableau web interface. At the top, the navigation bar includes the Tableau logo, a search bar, and user profile information. Below the navigation bar, the breadcrumb trail shows 'Home > TABLEAU > Tableau Viewer Landing Page > Landing Page'. The main content area is titled 'Welcome Erik Herman!' and features a 'My Projects' section. A red box highlights the 'TABLEAU' link in the breadcrumb trail, with a red arrow pointing to the 'Workbooks' tab in the navigation bar. Another red box highlights the 'Workbooks' tab, and a third red box highlights the '+ New Workbook' button. The 'Workbooks' tab shows a list of recent views, including 'User Specific View', 'Tableau Viewer Landing Page', 'All Tableau Subscriptions', 'Tableau Licenses Growth Chart', 'DV Monitoring Dashboard', 'Daily Metrics', and 'Tableau License Forecast 2019-20'. Each view card displays a thumbnail, title, and view count. A 'Training and Tutorials' section is visible in the bottom left corner, with a note: '*To access the training videos below, you will need to register with tableau.com in the links provided.'

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).

Connect to Data

Create a new data source from scratch or choose an existing data source to start from.

Files Connectors On this site

Search data sources

Name	Views: All	Workbooks	Connects To	Project	Owner	Live / Last extract
Tableau DIT Users	307	10	isd11to02	ADMIN	Pinnaboyana, Chaitanya	Live
Tableau Production Projects	3	0	atl11to02	ADMIN	Pinnaboyana, Chaitanya	Live
Tableau Production Users	769	14	atl11to02	ADMIN	Pinnaboyana, Chaitanya	Live

Add data source

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

Task 1 – Log in to Tableau (cont.)

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

The screenshot shows the Tableau interface with the 'Data Source' tab selected. A data table is displayed with the following columns and rows:

#	Customer Dimension	Customer Dimension	Customer Dimension	Customer Dimension	Customer Dimension	Customer Dimension
	Customer ID	Customer First Name	Customer Last Name	Customer Full Name	Customer Date of Birth	Customer Open Account D...
8	Doc	Dwarf	Doc Dwarf	8/1/1951	6/10/1979	
7	Bashful	Dwarf	Bashful Dwarf	7/1/1950	7/15/2002	
6	Louie	Duck	Louie Duck	6/1/1967	8/27/2002	
9	Dopey	Dwarf	Dopey Dwarf	9/1/1957	4/27/2014	

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

The screenshot shows the Tableau interface with the 'Data Source' tab selected. A context menu is open over the 'Customer Dimension' field, showing the following options:

- Duplicate
- Rename
- Hide
- Aliases... Create
- Convert to Continuous
- Convert to Measure
- Change Data Type
- Geographic Role

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.

The image displays two screenshots of the Tableau Desktop interface. The left screenshot shows the 'New Workbook' window with the 'Data' tab selected. The 'Dimensions' pane on the left is highlighted with a red box, showing 'Customer Dimension', 'Stock Dimension', 'Trade Fact', and 'Measure Names'. The 'Measures' pane below it is also highlighted with a red box, showing 'Stock Dimension', 'Trade Fact', 'Number of Records', and 'Measure Values'. A red arrow points from the 'Measure Values' item in the Measures pane to the 'Measure Values' item in the 'Marks' card of the right screenshot. The right screenshot shows the same 'New Workbook' window with the 'Analytics' tab selected. The 'Columns' shelf contains 'Measure Names'. The 'Rows' shelf contains 'YEAR(Trade Execution Date)', 'Customer ID', 'Trade ID', and 'Stock Name'. The 'Marks' card is set to 'Automatic'. The 'Measure Values' section of the Marks card is highlighted with a red box, showing 'SUM(Customer Age at Trade)', 'SUM(Trade Share Count)', and 'SUM(Trade Value Amount)'. The main view shows a table with columns for Year of Trade, Customer ID, Trade ID, Stock Name, Customer Age at Trade, Trade Share Count, and Trade Value Amount. The table data is as follows:

Year of Trade	Customer ID	Trade ID	Stock Name	Customer Age at Trade	Trade Share Count	Trade Value Amount
2020	7	299	Walt Disney Company	70	1,672	30,096
	12	160	Discovery Inc.	45	661	58,168
	22	Null	Null	94	1,996	975,136
2019	1	120	AT&T	89	1,333	86,645
		357	AT&T	89	769	46,909
		226	AT&T	89	1,367	31,441
	2	165	AT&T	84	880	61,600
		416	AT&T	84	1,011	49,539
	3	46	Twenty-First Century	94	949	18,980
	4	476	Twenty-First Century	92	1,603	299,761
		114	Twenty-First Century	92	1,265	170,775
		482	Twenty-First Century	92	1,172	120,716
		211	Twenty-First Century	92	842	106,934
		66	Twenty-First Century	92	461	25,355
		333	Twenty-First Century	92	173	18,684
5	54	Twenty-First Century	79	979	66,572	
	397	Walt Disney Company	79	27	12,771	
6	228	Walt Disney Company	52	1,907	938,244	
	433	Walt Disney Company	82	1,561	756,327	

Note: See Appendix A for additional Tableau learning resources.

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

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

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

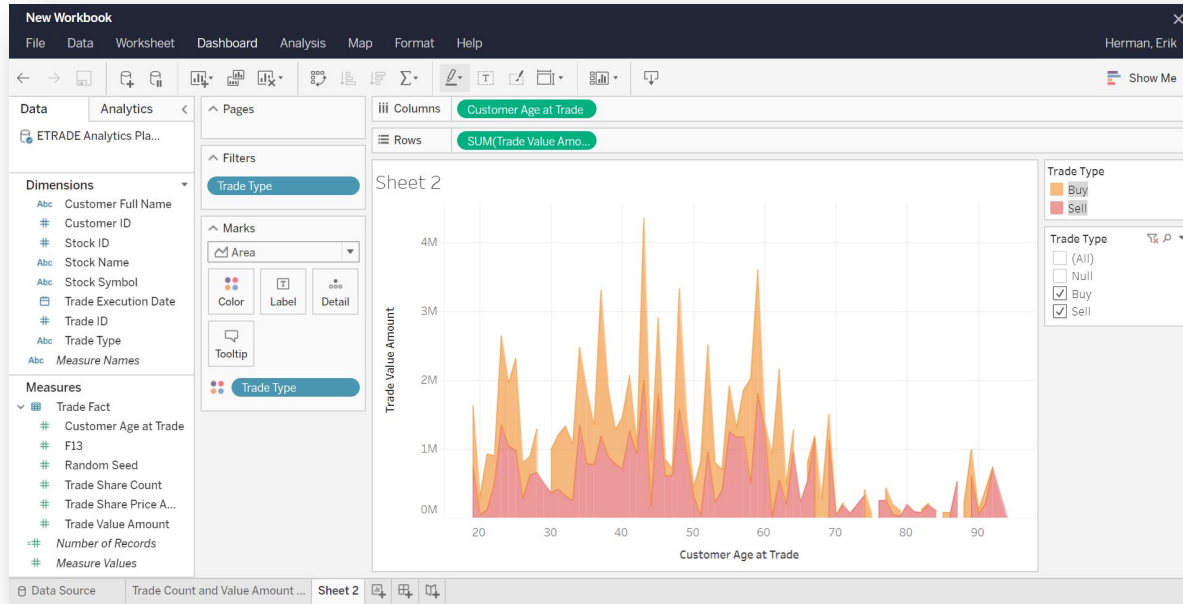
The image consists of two screenshots of the Tableau Desktop interface. The left screenshot shows a 'New Workbook' with a data source 'ETRADE Analytics Pla...'. The 'Columns' shelf contains 'YEAR(Trade Exec...)', 'Customer ID(Customi...', 'Trade ID', and 'Stock Name'. The 'Rows' shelf is empty. A 'Rename Sheet' dialog box is open, with the text 'Trade Count and Value Amount by Trade ID' entered. A red arrow points from the 'Sheet 1' tab at the bottom to the dialog box. The right screenshot shows the same workbook with 'Sheet 2' selected. A red arrow points from the 'New Worksheet' button at the bottom to the 'Sheet 2' tab.

Year of ..	Customer I..	Trade ID	Stock Name	Customer A..
2020	7	299	Walt Disney Co..	70
	12	160	Discovery Inc..	45
2019	1	120	AT&T	89
		357	AT&T	89
		226	AT&T	89
	2	165	AT&T	84
		416	AT&T	84
	3	46	Twenty-First C..	94
			enty-First C..	92
			enty-First C..	92
			enty-First C..	92
			enty-First C..	92
			enty-First C..	92
			enty-First C..	92
			enty-First C..	79
			Walt Disney Co..	79
			Walt Disney Co..	52
			Walt Disney Co..	52
			Walt Disney Co..	52

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).

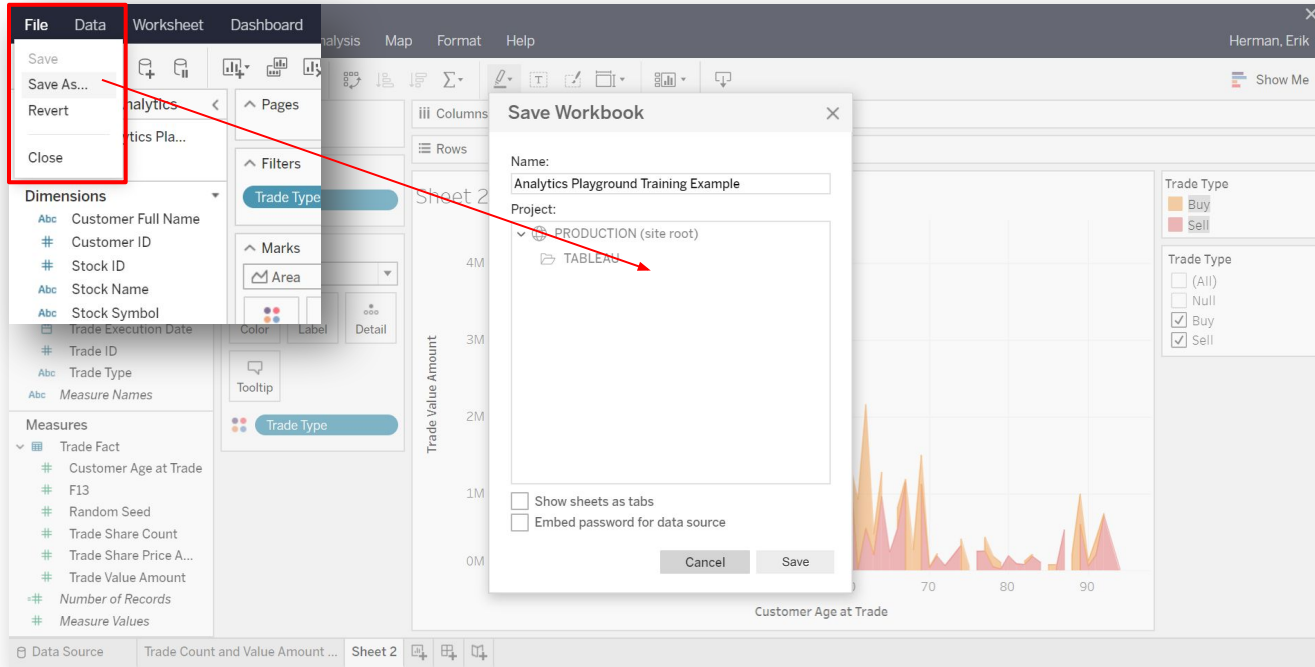


Note: Looking at this visualization we can already see patterns of buy/sell behavior based on customer age. A linear regression model will allow us to look at those patterns at a more granular level, as well as create predictions for future behavior.

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

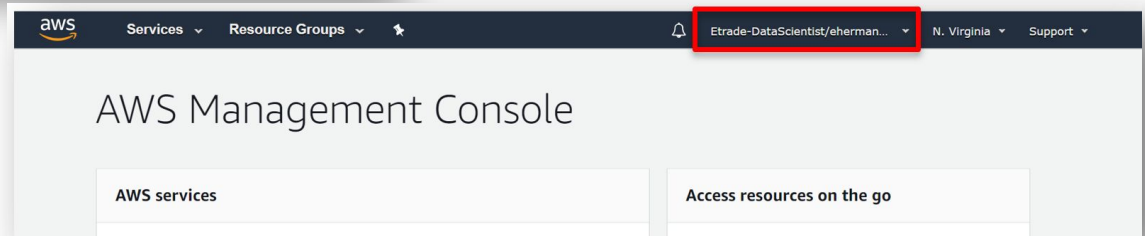
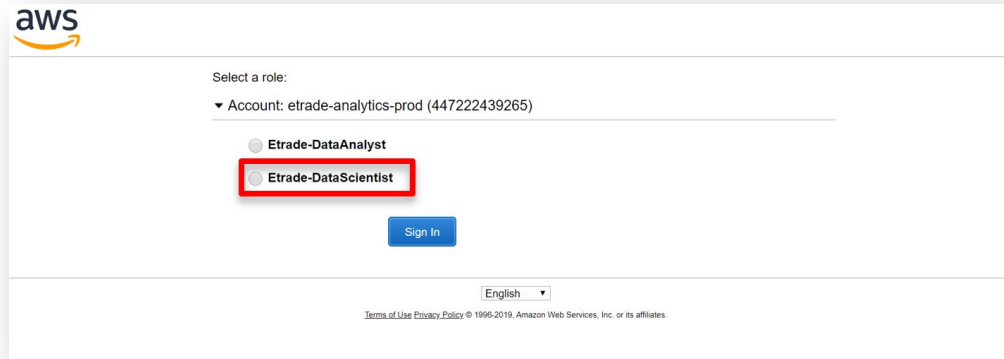
Tags

- 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

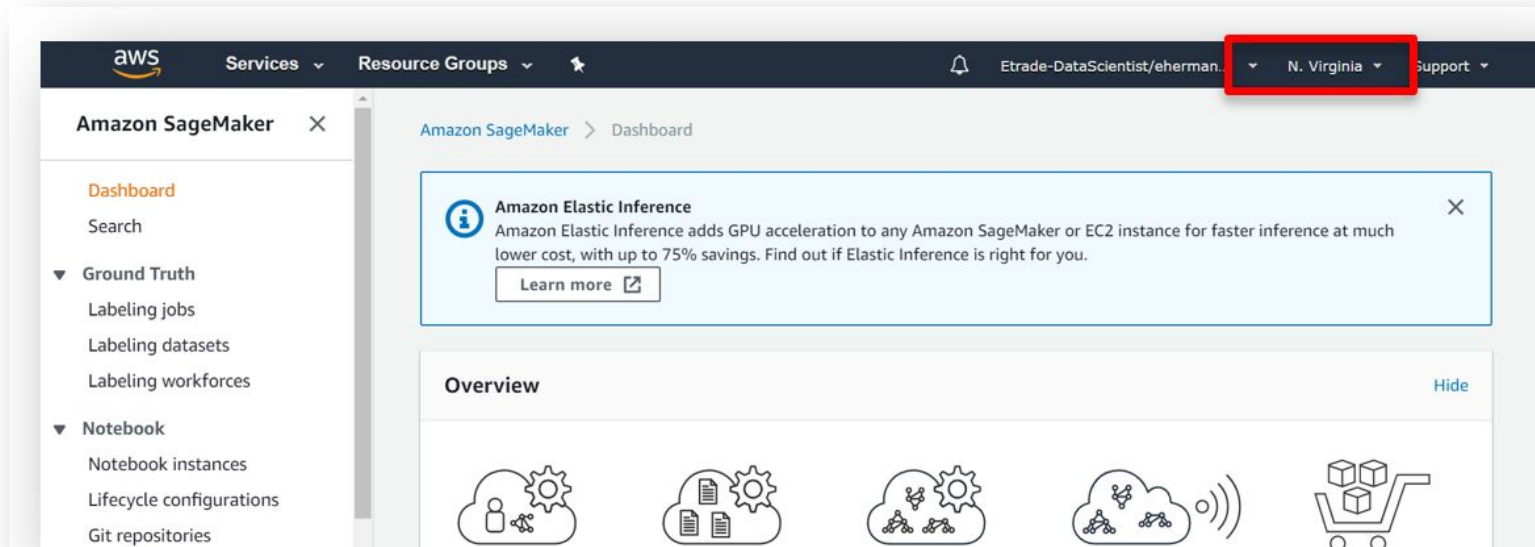
1. Click link provided by your administrator to login to the **AWS Management Console**.
2. Log in as **Etrade-DataScientist**.



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).

The screenshot displays the AWS SageMaker console interface. On the left, the navigation menu is visible, with 'Notebook instances' highlighted by a red rectangular box. A red arrow originates from this box and points to the 'Create notebook instance' button, which is also highlighted with a red rectangular box. The main content area shows the 'Notebook instances' page, featuring a table with columns for Name, Instance, Creation time, Status, and Actions. Two notebook instances are listed: 'boris-default' and 'yhu-notebook', both with a status of 'Stopped'. An 'Amazon Elastic Inference' notification is present at the top of the page, indicating GPU acceleration options.

Name	Instance	Creation time	Status	Actions
boris-default	ml.t2.medium	Dec 17, 2019 16:12 UTC	⊖ Stopped	Start
yhu-notebook	ml.t2.medium	Dec 13, 2019 14:43 UTC	⊖ Stopped	Start

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).

Amazon SageMaker > Notebook instances > Create notebook instance

Create notebook instance

Amazon SageMaker provides pre-built fully managed notebook instances that include example code for common model training and hosting exercises. [Learn](#)

Notebook instance settings

Notebook instance name:

Notebook instance type:

Elastic Inference [Learn more](#)

▶ Additional configuration

Permissions and encryption

IAM role
Notebook instances require permissions to call other services including SageMaker and S3. Choose a role or let us create a role with the [AmazonSageMakerFullAccess](#) IAM policy attached.

Enter a custom IAM role ARN

Custom IAM role ARN

Root access - optional

Enable - Give users root access to the notebook

Disable - Don't give users root access to the notebook
Lifecycle configurations always have root access

Encryption key - optional
Encrypt your notebook data. Choose an existing KMS key or enter a key's ARN.

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**.

The screenshot displays the configuration options for a SageMaker notebook instance, divided into two main sections: Network and Tags.

Network - optional

- VPC - optional**: Your notebook instance will be provided with SageMaker provided internet access because a VPC setting is not specified. The selected VPC is `vpc-057e978f00a70fc0c (10.115.0.0/19)`.
- Subnet**: Choose a subnet in an availability zone supported by Amazon SageMaker. The selected subnet is `subnet-07443a6ac1527912e (10.115.12.0/22) | us-east-1a subnet-ettrade-Private-app1`.
- Security group(s)**: The selected security group is `sg-08c8acd5de512809a (sagemaker-notebooks-sg) | sagemaker-notebooks-sg`.
- Direct internet access**: The **Disable** option is selected, indicating that internet access will be provided through a VPC. A note states: "To train or host models from a notebook, you need internet access. To enable internet access, make sure that your VPC gateway and your security group allows outbound connections. [Learn more](#)".

Tags - optional

Key	Value	
username	eherman	Remove

Buttons: **Cancel** and **Create notebook instance**.

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

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

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

Amazon SageMaker > Notebook instances

Amazon Elastic Inference ×

Amazon Elastic Inference adds GPU acceleration to any Amazon SageMaker or EC2 instance for faster inference at much lower cost, with up to 75% savings. Find out if Elastic Inference is right for you.

[Learn more](#)

Notebook instances Actions ▾ Create notebook instance

Search notebook instances < 1 > ⚙

Name	Instance	Creation time	Status	Actions
<input type="radio"/> elhetradetraining	ml.t2.medium	Nov 29, 2019 14:38 UTC	✔ InService	Open Jupyter Open JupyterLab

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

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

- Once the Jupyter notebook opens, click the SageMaker Examples tab.

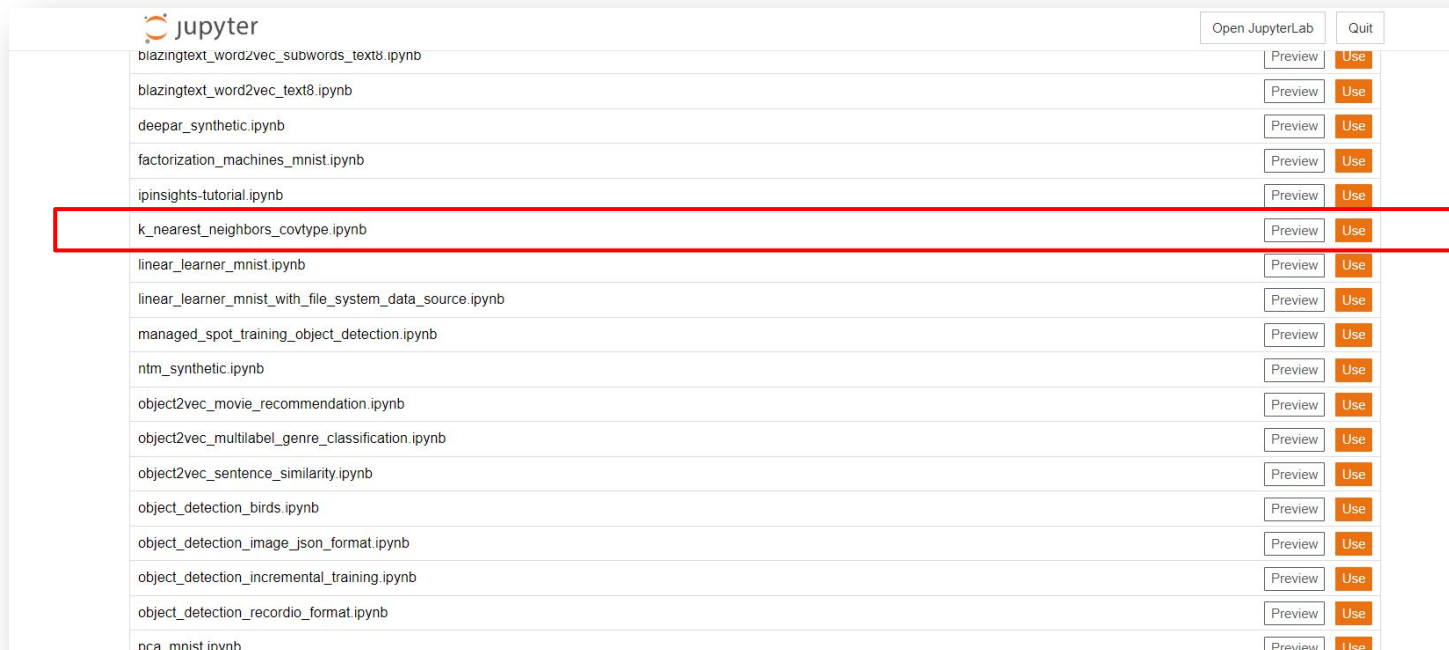
The screenshot shows the JupyterLab interface. At the top, there are tabs for 'Files', 'Running', 'Clusters', 'SageMaker Examples', and 'Conda'. The 'SageMaker Examples' tab is highlighted with a red box. Below the tabs, there is a message 'Select items to perform actions on them.' and a file browser area showing '0' items. A red arrow points from the 'SageMaker Examples' tab to a list of example notebooks. The list includes:

Name	Preview	Use
Image-classification-incremental-training-highlevel.ipynb	Preview	Use
Image-classification-1st-format-highlevel.ipynb	Preview	Use
Image-classification-1st-format.ipynb	Preview	Use
Image-classification-multilabel-1st.ipynb	Preview	Use
Image-classification-transfer-learning-highlevel.ipynb	Preview	Use
Image-classification-transfer-learning.ipynb	Preview	Use
LDA-introduction.ipynb	Preview	Use
SageMaker-Seq2Seq-Translation-English-German.ipynb	Preview	Use
blazingtext_hosting_pretrained_fasttext.ipynb	Preview	Use
blazingtext_text_classification_dbpedia.ipynb	Preview	Use
blazingtext_word2vec_subwords_text8.ipynb	Preview	Use
blazingtext_word2vec_text8.ipynb	Preview	Use
deepar_synthetic.ipynb	Preview	Use
factorization_machines_mnist.ipynb	Preview	Use
ipinsights-tutorial.ipynb	Preview	Use

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

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

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



The screenshot shows the JupyterLab interface with a list of notebooks. The notebook `k_nearest_neighbors_covtype.ipynb` is highlighted with a red box. The interface includes a header with the Jupyter logo, the text "jupyter", and buttons for "Open JupyterLab" and "Quit". The list of notebooks is as follows:

Notebook Name	Preview	Use
blazingtext_word2vec_subwords_text8.ipynb	Preview	Use
blazingtext_word2vec_text8.ipynb	Preview	Use
deepar_synthetic.ipynb	Preview	Use
factorization_machines_mnist.ipynb	Preview	Use
ipinsights-tutorial.ipynb	Preview	Use
k_nearest_neighbors_covtype.ipynb	Preview	Use
linear_learner_mnist.ipynb	Preview	Use
linear_learner_mnist_with_file_system_data_source.ipynb	Preview	Use
managed_spot_training_object_detection.ipynb	Preview	Use
ntm_synthetic.ipynb	Preview	Use
object2vec_movie_recommendation.ipynb	Preview	Use
object2vec_multilabel_genre_classification.ipynb	Preview	Use
object2vec_sentence_similarity.ipynb	Preview	Use
object_detection_birds.ipynb	Preview	Use
object_detection_image_json_format.ipynb	Preview	Use
object_detection_incremental_training.ipynb	Preview	Use
object_detection_recordio_format.ipynb	Preview	Use
pca_mnist.ipynb	Preview	Use

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

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

- 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](https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.data.gz)) (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
```

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_zmSrvYnLUTggsZz&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://en.wikipedia.org/wiki/Python_(programming_language))

<https://www.python.org/about/gettingstarted/>

R

[https://en.wikipedia.org/wiki/R_\(programming_language\)](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.