



Cloudera Altus on Azure - Hands-on Workshop

Exercise Manual

Created by Cloudera

Summary	1
Exercise 1: Logging in to Altus	1
Exercise 2: Getting to know the Altus UI	4
Exercise 3: How to create a cluster	9
Exercise 4: Preparing the object store	12
Exercise 5: Running a Job.	17
Exercise 6: Workload Analytics	20
Exercise 7: Combine Structured and UnStructured data	25
Conclusion: Visualizing the results.	32

Summary

Welcome to the Altus on Azure Hands-on-Lab attendee guide. Below are a series of exercises designed to familiarize yourself with the Cloudera Altus platform offering. In addition to this, we will walk through a simple data engineering problem and how Altus empowers the end users to solve it and gain further insights into the data.

The data engineering problem consists of correlating structured data with unstructured data. Let's say you work for a sports retail company and you'd like to know the top selling products by state. That's easy enough. Just query the data stored in your relational database and you have your answer. Let's take it one step further and ask the question: Are the highest selling products also the most popular products on our website? How do we answer this question? At the end of this lab we will find out!

Exercise 1: Logging in to Altus

Now that we have had an overview of Altus, let's get our hands dirty. The first thing we'll need to do is log into the system. Navigate to the following URL in your browser, click **Sign In** at the top right:

<http://altus.cloudera.com/>

azuretestdrivedata - Microsoft x Sign In x

Secure | <https://sso.cloudera.com/?SSOurl=https%3A%2F%2Fcloudera-production.okta.com%2Fapp%2Fcloudera...>

CDSW Cloud Cloudera Infrastruc... Code Coursera Cycling Hadoop Books Multitenancy >> Other Bookmarks

cloudera 🔍 👤

Please Sign In

✉

🔒

Sign In

✎

Don't have an account?
Register Now

🔒





Forgot your password?
Reset Password

Login with your username and password. Please note that this is a shared login which will be used by all lab attendees.

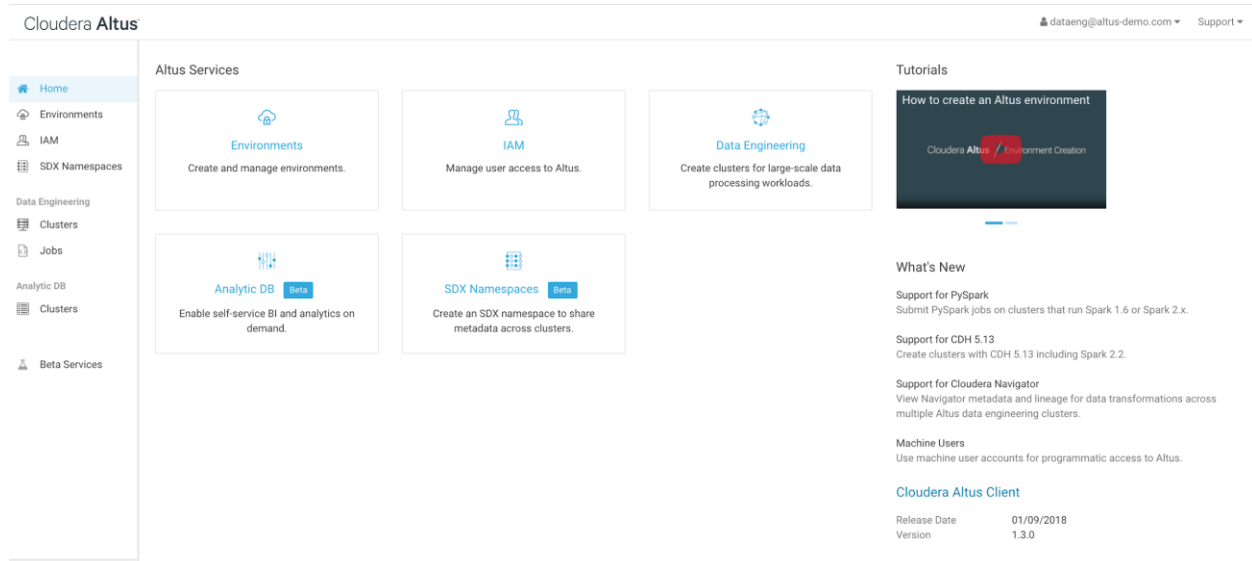
Username: testdrive@altus-demo.com

Password: In the email you received when you registered.

Please Sign In

	<input type="text" value="testdrive@altus-demo.com"/>	
	<input type="password" value="....."/>	
<input type="button" value="Sign In"/>		

Once you are logged in, we can get started. Your home page should look something similar to below.



The Altus organization you're currently logged in to is called the Altus-demo organization. This organization is used by Cloudera Engineers for demos as well as hands-on-labs, just like today's! Since this is a live demo environment, **please do NOT delete any clusters or resources.**

Exercise 2: Getting to know the Altus UI

Now that we have logged in, let's get familiar with the UI.

In the top-right hand corner of the page, click on Support -> Documentation

The screenshot shows the Cloudera Altus dashboard. On the left is a navigation sidebar with links to Home, Environments, IAM, SDX Namespaces, Data Engineering, Clusters, Jobs, Analytic DB, and Beta Services. The main content area is titled 'Altus Services' and contains five service cards: Environments (Create and manage environments), IAM (Manage user access to Altus), Data Engineering (Create clusters for large-scale data processing workloads), Analytic DB (Enable self-service BI and analytics on demand, marked as Beta), and SDX Namespaces (Create an SDX namespace to share metadata across clusters, marked as Beta). On the right, there is a 'Tutorials' section with a video thumbnail titled 'How to create an Altus environment' and a 'Documentation' dropdown menu with links to Documentation (Analytic DB), Documentation (SDX), Community, and Support Center. Below the tutorials is a 'What's New' section with updates on Microsoft Azure support, Custom Tags, CDH 5.14, and Secure Clusters. At the bottom right, there is a 'Cloudera Altus Client' section showing the release date (06/06/2018) and version (1.4.1).

You can visit these docs at any time in case you have further questions about Altus components.

On the right-side of the page, you'll see the What's New section which details the latest features of Altus. Below that is the Latest Cloudera Altus CLI download link. Everything we're doing here today can be done through the CLI in a more programmatic fashion.

What's New

Support for Microsoft Azure

Altus Data Engineering is now generally available.

Support for Custom Tags

Define tags to associate with your cluster instances.

Support for CDH 5.14

Create clusters with CDH 5.14 in addition to CDH 5.13 and CDH 5.12.

Support for Secure Clusters

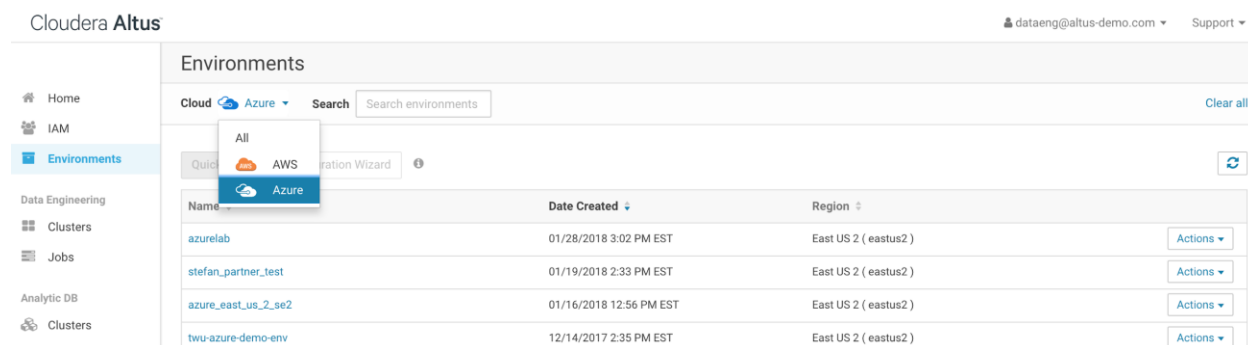
Create clusters with secure communication between services in the cluster and with encrypted data in the cluster.

Cloudera Altus Client

Release Date 06/06/2018

Version 1.4.1



Click on Environments in the left-hand pane. Then click on the Cloud button in the main pane and select Azure.





Cloudera Altus

dataeng@altus-demo.com Support

Environments

Cloud  Azure  Search Search environments Clear all

Quick  AWS  Azure Migration Wizard

Name	Date Created	Region	Actions
azurelab	01/28/2018 3:02 PM EST	East US 2 (eastus2)	Actions
stefan_partner_test	01/19/2018 2:33 PM EST	East US 2 (eastus2)	Actions
azure_east_us_2_se2	01/16/2018 12:56 PM EST	East US 2 (eastus2)	Actions
twu-azure-demo-env	12/14/2017 2:35 PM EST	East US 2 (eastus2)	Actions

An Altus environment: Defines the resources in your Azure subscription that are used by Cloudera Altus to create clusters and jobs. In a production environment, an administrator can set up and assign separate Altus environments to different users and groups. **Please do not create any new environments for this lab.**

Click on the environment called “**AzureTestDrive3**” and view the environment details:

The screenshot displays the Cloudera Altus web interface. The top navigation bar includes 'Home', 'Environments', 'SDX Namespaces', 'IAM', 'Users', 'Groups', 'Data Engineering', 'Clusters', 'Jobs', 'Data Warehouse', and 'Query Editor'. The 'Environments' section is active, showing a list of environments with 'AzureTestDrive2' selected. The 'Details' tab is open, displaying a green status bar indicating 'Environment permissions are up-to-date'. Below this, the 'Primary Actions' section contains three buttons: 'Submit Jobs', 'Create Data Engineering Cluster', and 'Create Data Warehouse Cluster'. The 'General Settings' section lists 'Azure Active Directory Tenant ID' (redacted), 'Workload Analytics' (Enabled), and 'Secure Clusters' (Disabled). The 'Network Settings' section lists 'Virtual Network Name' (azuretestdrive-vnet), 'Virtual Network Resource Group' (azuretestdrive-rg), 'Subnet Name' (default), 'Network Security Group Name' (azuretestdrive-nsg), and 'Network Security Group Resource Group' (azuretestdrive-rg). The 'Instance Settings' section lists 'Cluster Node Resource Group' (azuretestdrive-rg), 'User Assigned Managed Service Identity Name' (azuretestdrive-ua-msi), 'User Assigned Managed Service Identity Resource Group' (adi://azuretestdrivedata.azuredatastore.net/AltusLogs), and 'Data Lake Store Log Archive Path'.

As you can see, Altus leverages existing resources in the customer’s Azure subscription (resources such as vnets, network security groups etc. created by an Azure administrator). An Altus administrator then creates an Altus environment using those Azure resources.

Click on Clusters in the left-hand pane.

Clusters

Cloud All ▾ Environment All ▾ Status All ▾ Service Type All ▾ Search

Create Cluster

Cloud ▾	Name ▾	Status ▾	Type ▾	Workers ▾	Creation Time ▾
	Altus_DG_RetailData	✔ Created	Hive on Spark	3	01/25/2018 3:56 PM EST
	AltusDE-Spark-Demo	✔ Created	Spark 2.x	3	01/22/2018 6:59 PM EST
	engineering-training	✔ Created	Hive	5	01/09/2018 12:12 PM EST

Displaying 1 - 3

Here you will see a list of existing clusters you or other data engineers have created.

Click on Jobs in the left-hand pane.

Here you will see a list of previous jobs you have run. Since you are using a shared login, you will see past jobs that others have run as the testdrive@altus-demo.com user.

Exercise 3: How to create a cluster

In this exercise, we will walk through the steps to create a data engineering Hive on Spark cluster. However, in the interest of time, we have created some clusters for you to use in the lab. This exercise will show you the create process but **please DO NOT CLICK CREATE** at the end of the exercise.

Click on Clusters in the left pane then Create Cluster in the main pane.

Fill out the information for building a cluster.

General Information:

Cluster Name: **leave blank (since we are not going to create a cluster at the end)**








Service Type: **Hive on Spark**


CDH Version: **CDH 5.15**

SDX Namespace: **altus-test-drive-namespace**

Environment: **AzureTestDrive3**

General Information

Cluster Name *		<input type="text" value="Enter cluster name"/>
Service Type *		<div>Spark 2.x</div>
CDH Version *		<div>CDH 5.15 (Spark 2.2)</div>
SDX Namespace		<div>altus-test-drive-namespace</div>
Environment *		<div>  AzureTestDrive (Region: eastus2)</div>

 The environment you selected is configured for secure clusters. All clusters created using this environment will require authentication.

Node Configuration:

Worker: **Set the Number of Nodes to 3** (instead of 5).

Node Configuration

Node Type	Instance Type ?	Number of Nodes
Worker ?	STANDARD_DS12_V2 28 GiB 4vCPU	3
Master	STANDARD_DS12_V2 28 GiB 4vCPU	1
Cloudera Manager	STANDARD_DS12_V2 28 GiB 4vCPU	1

Credentials:

For the SSH Public Key, use any public key you have access to. If you don't currently have a public key, then click on **"Direct Input"** and copy and paste the below public key into the SSH Public Key box:

ssh-rsa

```
AAAAB3NzaC1yc2EAAAADAQABAAQDbmWtQ9V8S41bBV83/VI8m+3DQwTxJYigroT30bZv7UYYZ7Guk+zW9DaHEeHdIOeO3RDovPzU2j0C2Zw6nGE2iRZ8KbPUAFQxUlkkhYetMPb+ILV6mB4y9Rf1nx//r8npFbuTnYcocklUZggMS+NRoQY+vF03CATL1Lfr0R9CCWb4GfeAOaCpDHuPrthlf2s2QT//yxyf9++0S87ysqqC00x9csdmYMUgfPWWVzD4noOcu3nKySxxbNm0aAAH0d+H42+cnFjwxfgg/mr+8PZOTot1LOqSNj5YXPORab/V9/+3pIO+AFA+08JESX75uy5AozoDKYXGaBBOZXGJxcfr poweradmin@altus-demo.com
```

Credentials

SSH Public Key * ?

☐ File Upload

☒ Direct Input

```
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAQDbmWtQ9V8S41bBV83/VI8m+3DQwTxJYigroT3
0bZv7UYYZ7Guk+zW9DaHEeHdIOeO3RDovPzU2j0C2Zw6nGE2iRZ8KbPUAFQxUlkkhYetM
Pb+ILV6mB4y9Rf1nx
//r8npFbuTnYcocklUZggMS+NRoQY+vF03CATL1Lfr0R9CCWb4GfeAOaCpDHuPrthlf2s2Q
T
//yxyf9++0S87ysqqC00x9csdmYMUgfPWWVzD4noOcu3nKySxxbNm0aAAH0d+H42+cnFj
wxfgg/mr+8PZOTot1LOqSNj5YXPORab
/V9/+3pIO+AFA+08JESX75uy5AozoDKYXGaBBOZXGJxcfr poweradmin@altus-demo.com|
```

Also pick a username and password for Cloudera Manager. You can use **"cloudera"** for both the username and password.

Credentials

SSH Public Key * 

☒ File Upload ☐ Direct Input

[Choose File](#)

Cloudera Manager 

Username *

Cloudera Manager 

Password *



Confirm Cloudera Manager

Password *



Now click Cancel. **Please do NOT CLICK "Create Cluster".**

[Cancel](#)

[Create Cluster](#)

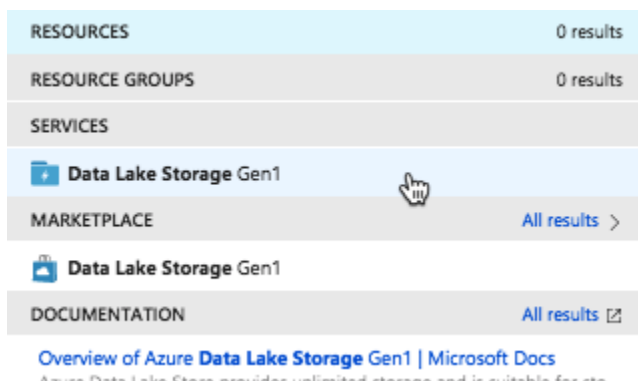
Exercise 4: Preparing the object store

In this exercise, you will be familiarizing yourself with the data stored in ADLS. You will also create a subfolder (named after yourself) to host the output from an Altus job you will run.

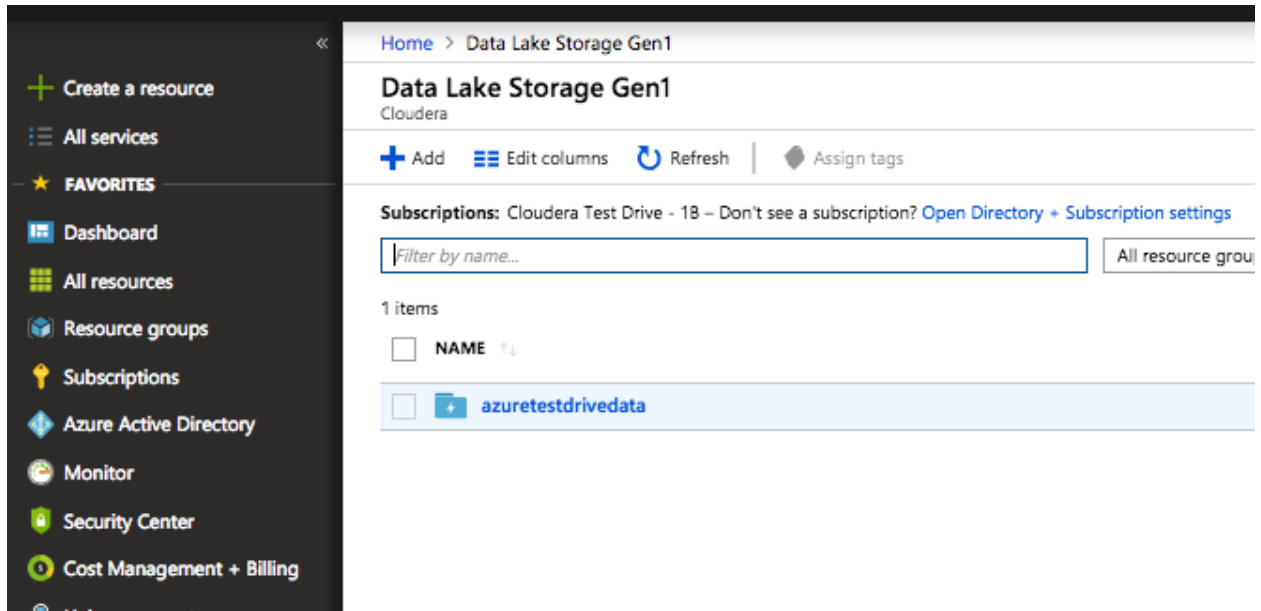
Once you've logged in to the Azure portal, go to the search bar at the top of the screen and type in "data lake storage"



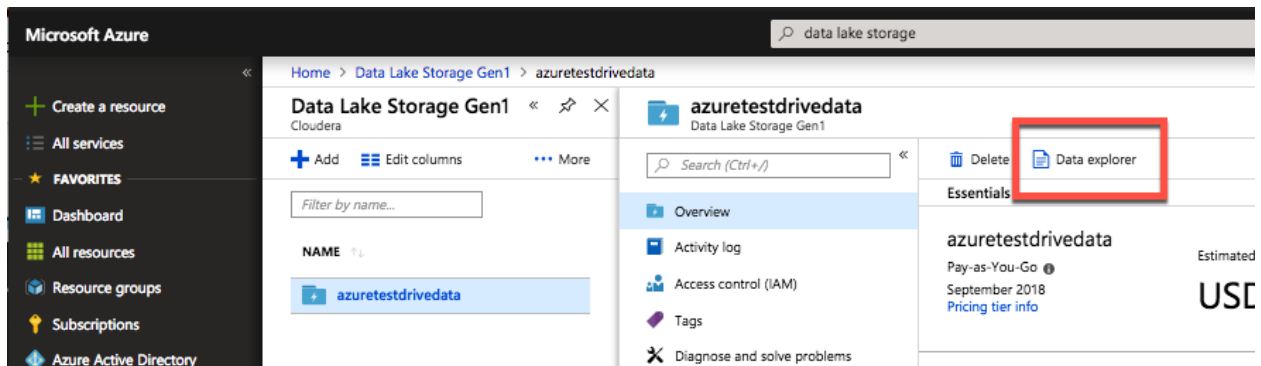
Click on "Data Lake Storage Gen 1" under Services



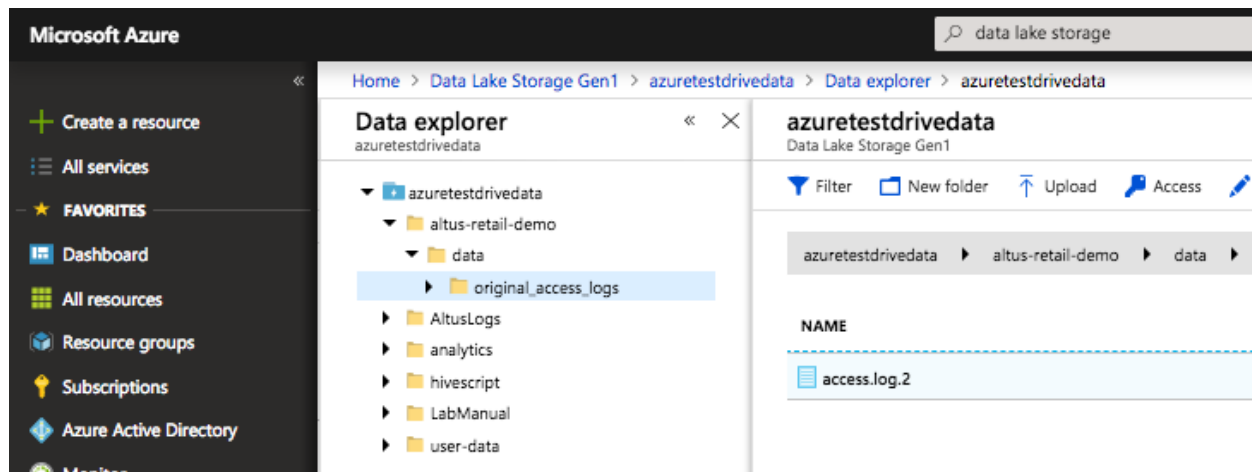
Click on **azuretestdrivedata**



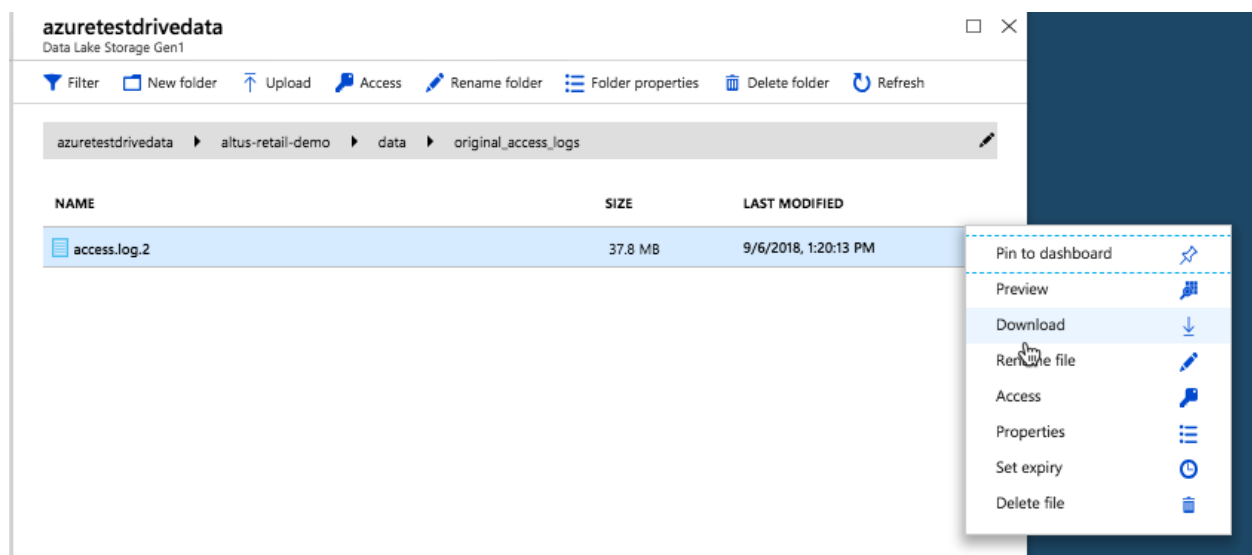
Click on “**Data Explorer**” which is in the right-most pane at the top:



Navigate to the **azuretestdrivedata** ADLS folder. Go into the **altus-retail-demo/data/original_access_logs** folder.



Inside the **original_access_logs** folder is the **access.log.2** file which contains weblog data. Download the file (by clicking on the three dots next to the filename) and open it up with your favourite text editor (notepad or text editor).



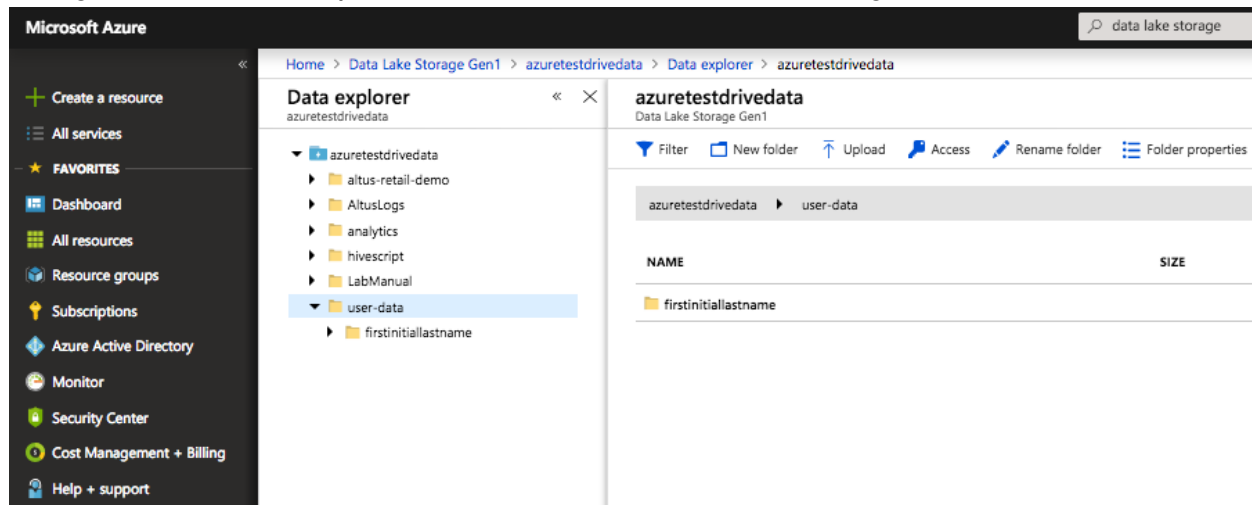
This file looks similar to below:

```

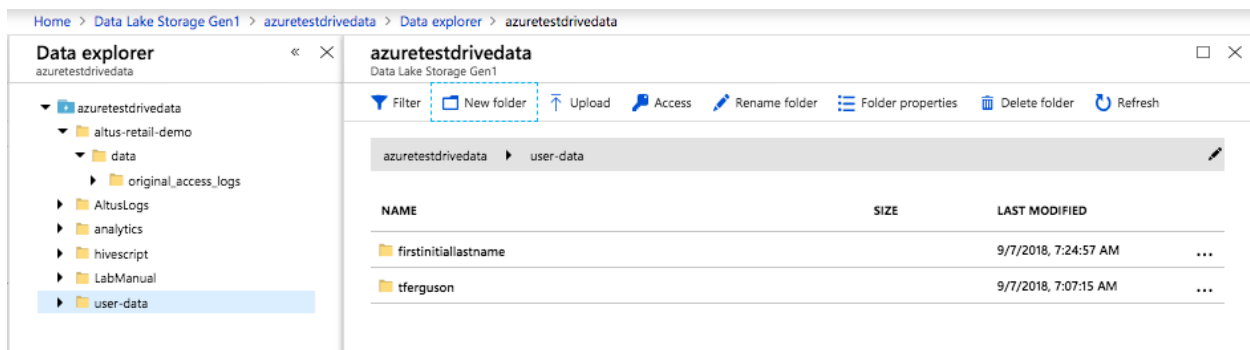
access.log.2
1 79.133.215.123 - - [14/Jun/2014:10:30:13 -0400] "GET /home HTTP/1.1" 200
1671 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/35.0.1916.153 Safari/537.36"
2 162.235.161.200 - - [14/Jun/2014:10:30:13 -0400] "GET /department/apparel/cat
egory/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20C
leat HTTP/1.1" 200 1175 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_3)
AppleWebKit/537.76.4 (KHTML, like Gecko) Version/7.0.4 Safari/537.76.4"
3 39.244.91.133 - - [14/Jun/2014:10:30:14 -0400] "GET /department/fitness
HTTP/1.1" 200 1435 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_3)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
4 150.47.54.136 - - [14/Jun/2014:10:30:14 -0400] "GET /department/fan%20shop/ca
tegory/water%20sports/product/Pelican%20Sunstream%20100%20Kayak/add_to_cart
HTTP/1.1" 200 1932 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_3)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
5 217.89.36.129 - - [14/Jun/2014:10:30:14 -0400] "GET /view_cart HTTP/1.1" 200
1401 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:30.0) Gecko/20100101
Firefox/30.0"
6 36.44.59.115 - - [14/Jun/2014:10:30:15 -0400] "GET
/department/footwear/category/cardio%20equipment HTTP/1.1" 200 386 "-"
"Mozilla/5.0 (Windows NT 6.1; WOW64; rv:30.0) Gecko/20100101 Firefox/30.0"
7 11.252.83.179 - - [14/Jun/2014:10:30:15 -0400] "GET /view_cart HTTP/1.1" 200
1726 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_3) AppleWebKit/537.36 (
KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
8 56.251.19.230 - - [14/Jun/2014:10:30:15 -0400] "GET
/department/footwear/category/fitness%20accessories HTTP/1.1" 200 2076 "-"
"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/35.0.1916.153 Safari/537.36"

```

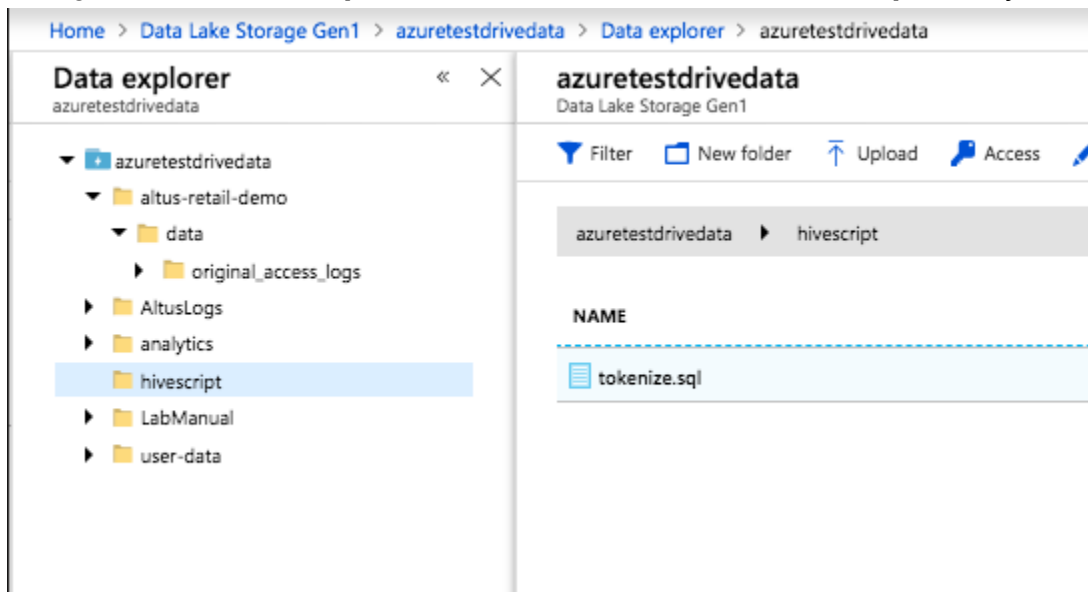
Navigate back to the top-level **azuretestdrivedata** folder and go into **user-data** folder.



Create a sub-folder with your name. Click on “New Folder” and enter your short name, first Initial last name (all lowercase and no spaces). For example tferguson.



Navigate to the **hivescript** folder and download the **tokenize.sql** file to your laptop.



Open and review the **tokenize.sql** file. This is the Data Engineering query you are going to run.

```
-- Create your database
CREATE DATABASE IF NOT EXISTS ${YOURNAMEHERE};

use ${YOURNAMEHERE};

-- Create intermediate_access_logs table, drop if exists first
drop table if exists intermediate_access_logs;
CREATE EXTERNAL TABLE intermediate_access_logs (
  ip STRING,
  date STRING,
  method STRING,
  url STRING,
  http_version STRING,
```

```

code1 STRING,
code2 STRING,
dash STRING,
user_agent STRING)
ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
  'input.regex' = '([\ ]*) - - \\\\[([^\ ]*)\\\[ "([\ ]*) ([^\ ]*) ([^\ ]*)" (\\d*) (\\d*)
"([\ ]*)" "([\ ]*)"',
  'output.format.string' = "%1$$$ %2$$$ %3$$$ %4$$$ %5$$$ %6$$$ %7$$$ %8$$$ %9$$$")
LOCATION 'adl://azuretestdrivedata.azuredatalakestore.net/altus-retail-
demo/data/original_access_logs';

-- Create tokenized_access_logs table, drop if exists first
drop table if exists tokenized_access_logs;
CREATE EXTERNAL TABLE tokenized_access_logs (
  ip STRING,
  date STRING,
  method STRING,
  url STRING,
  http_version STRING,
  code1 STRING,
  code2 STRING,
  dash STRING,
  user_agent STRING)
stored as parquet
LOCATION 'adl://azuretestdrivedata.azuredatalakestore.net/user-
data/${YOURNAMEHERE}/tokenized_access_logs';

INSERT OVERWRITE TABLE tokenized_access_logs SELECT * FROM intermediate_access_logs;

```

Exercise 5: Running a Job.

In this exercise, you will use pre-created clusters to run your job. Ask the instructor for guidelines on which clusters to use. The job you are about to run will take semi-structured web log data and will transform it into a structured format for SQL querying.

Click on **Jobs**, then **Submit Jobs**.

Home
 Environments
 SDX Namespaces

 IAM

 Users

 Groups

 Data Engineering

 Clusters

Jobs

Jobs

Environment All ▾
 Cluster All ▾
 Submitter Test Drive ▾

Submit Jobs

Cloud	Name	Group
No jobs found matching your filters.		

Submission: Single Job

Job type: Hive

Job Name: use your name

Script: Simply select 'direct input' and copy/pasted the tokenize.sql code


Hive Script Parameters: Add one parameter with variable name - YOURNAMEHERE and value with sub-folder name you previously created in ADLS. This name will also be a database name.


Job XML: None



Action on Failure: None

Cluster: Select "Use existing" and choose the cluster name you got in the email.

Submit Jobs

Job Type  Hive


Job Name  skpabba



Script  ☐ Script Path  ☐ File Upload ☒ Direct Input


```
-- Create your database
CREATE DATABASE IF NOT EXISTS ${YOURNAMEHERE};

use ${YOURNAMEHERE};

-- Create intermediate_access_logs table, drop if exists first
drop table if exists intermediate_access_logs;
CREATE EXTERNAL TABLE intermediate_access_logs (
  ip STRING,
  date STRING,
```

Hive Script Parameters  ☒



 

Job XML  ☐

Click on Submit Jobs. The job will take ~3 mins to complete. If your job fails, you can try cloning the job and rerunning it or contact the lab instructor if you need help debugging the problem.

Submit Jobs



Cloud	Name	Group	Status	Type	Submitter	Start Time	Cluster	Actions
	stefan-salandy	stefan-salandy	Completed	Hive	dataeng dataeng	01/28/2018 4:23 PM EST	azure-lab-stefan-salandy	Clone Job
	sgs1	sgs1	Completed	Hive	dataeng dataeng	01/18/2018 10:21 AM EST	stefan-salandy-retail-clickstream-2	Clone Job Group

Displaying 1 - 2

Exercise 6: Workload Analytics

Click on your completed job. Here you can view details about the job you just ran.

The screenshot shows the Cloudera Altus web interface. The top navigation bar includes the Cloudera Altus logo, a user profile for 'dataeng@altus-demo.com', and a 'Support' link. The left sidebar contains a navigation menu with options: Home, IAM, Environments, Data Engineering, Clusters, Jobs (highlighted), Analytic DB, and Clusters. The main content area is titled 'Jobs / stefan-salandy' and 'stefan-salandy'. It features a 'Completed' status indicator, a 'Job Settings' section with a script editor containing SQL code, and a 'Job Details' sidebar on the right. The script editor shows a Hive script to create an external table. The job details sidebar includes a 'Job Type' of 'Hive', a 'Timeline (EST)' with 'Created', 'Start', and 'End' times, 'Queue Time', and 'Execution Time'. It also lists the 'Submitter' as 'dataeng dataeng', the 'Cluster' as 'azure-lab-stefan-salandy', the 'ID' as 'a02b0ed9-0666-43a7-90ff-fe5bdc85c12', and the 'CRN' as 'crm.altus:dataeng-us-west-1:9992c4f'.

Cloudera Altus

dataeng@altus-demo.com Support

Jobs / stefan-salandy

stefan-salandy Actions Details Analytics

Completed

Job Settings

Script

```
drop table if exists intermediate_access_logs;  
CREATE EXTERNAL TABLE intermediate_access_logs (  
  ip STRING,  
  date STRING,  
  method STRING,  
  url STRING,  
  http_version STRING,  
  code1 STRING,  
  code2 STRING,  
  dash STRING,
```

Hive Script Parameters Unspecified

Job XML Unspecified

Action on Failure None

Job Type
Hive

Timeline (EST)

Created	01/28/2018 4:23 PM
Start	01/28/2018 4:24 PM
End	01/28/2018 4:25 PM
Queue Time	3s
Execution Time	1m 55s

Submitter
dataeng dataeng

Cluster
azure-lab-stefan-salandy

ID
a02b0ed9-0666-43a7-90ff-fe5bdc85c12

CRN
crm.altus:dataeng-us-west-1:9992c4f

Click on **Analytics**. It might take a few minutes to appear.

Jobs

stefan-salandy Actions ▾ Details Analytics

Health Checks Execution Details Baseline Trends 1/28/2018 4:24 PM 1m 6s dataeng dataeng N/A

Baseline

- Duration
- Input Size
- Output Size
- Skew
- Task Duration
- Input Data
- Output Data
- Shuffle Input
- Read Speed
- Resources

Good news! Everything looks healthy.
You're welcome.

Execution Details
View configurations, metrics, tasks, queries, and logs...

Baseline
Compare metrics across any and all stages of this job.

Trend
See the performance of this operation over its history.

Click on **Execution Details** and then click on any one of the SQL queries. Then click on **Query**.

Jobs

stefan-salandy Actions ▾ Details Analytics

Health Checks Execution Details Baseline Trends 1/28/2018 4:24 PM 1m 6s

Expand All Collapse All

16:24	✓ OZ stefan-salandy	1m 6s	create external table tokenize...
16:25	✓ HV create external table intermedi...	0s	Summary
16:25	✓ HV create external table tokenized...	0s	View Query View Configurations
16:25	⊞ ✓ SP Hive on Spark	27s	
16:25	✓ HV insert overwrite table tokenized...	0s	

Looks familiar? It's part of the SQL script that you ran.

create external table tokenize...

Query

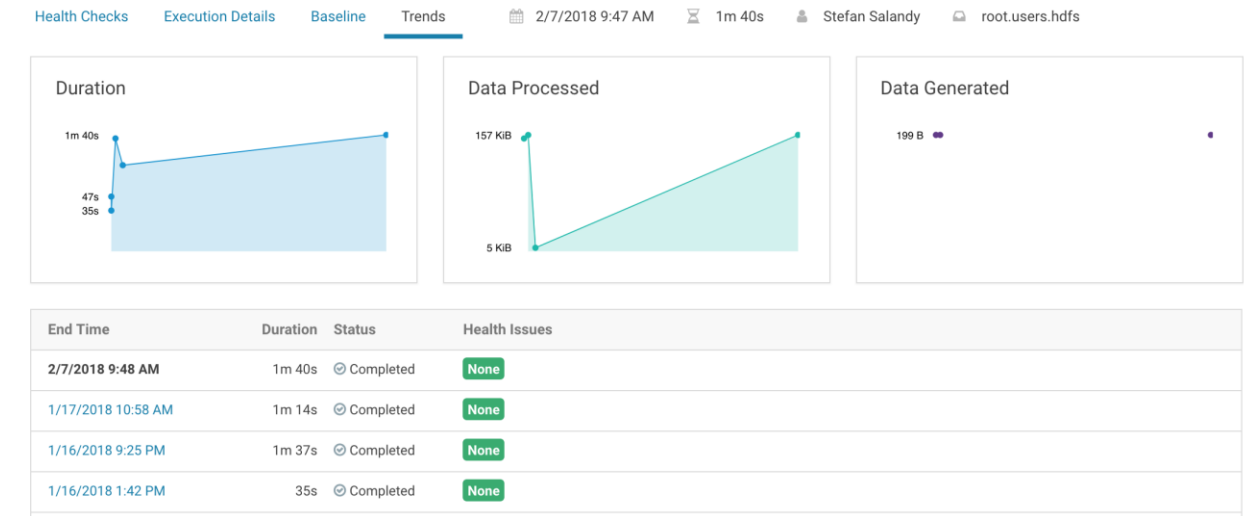
```
CREATE EXTERNAL TABLE tokenized_access_logs (  
  ip STRING,  
  date STRING,  
  method STRING,  
  URL STRING,  
  http_version STRING,  
  code1 STRING,  
  code2 STRING,  
  dash STRING,  
  user_agent STRING  
) STORED AS parquet location '*****'
```

Click on Baseline to view more info about your job. Note that since this is your first time running a job, Altus Workload Analytics does not have enough info from previous runs to create the baseline.

Jobs			
stefan-salandy		Actions ▾	Details Analytics
Health Checks	Execution Details	Baseline	Trends
All Stages / stefan-salandy ▾		1/28/2018 4:24 PM	1m 6s
Metric ⓘ	Current	Baseline ⓘ	
Active Tasks	0		
Disk bytes Spilled	0 B		
Duration	1m 6s		
Executor Runtime	6s		
Failed Task attempts	0		

For example purposes we have included screenshots of previously run jobs. These jobs have Baseline and Trend analysis:

Health Checks		Execution Details	Baseline	Trends	2/7/2018 9:47 AM	1m 40s	Stefan Salandy	
All Stages / stefan-salandy-retail-clickstream ▾								
Metric ⓘ	Current	Baseline ⓘ	Difference					
Duration	1m 40s	1m 1s	+40s	+65%				
Total memory-time taken by all map tasks	452.4K MiB-s	320.2K MiB-s	+132.2K MiB-s	+41%				
Total time spent by all map tasks	1m 20s	57s	+23s	+41%				
Total time spent by all maps in occupied slots	8m 2s	5m 41s	+2m 21s	+41%				
Total vcore-time taken by all map tasks	1m 20s	57s	+23s	+41%				
Executor Runtime	6s	9s	-3s	-33%				
Total Task duration	8s	11s	-3s	-27%				
GC time elapsed	< 1s	< 1s	+< 1s	+18%				
Physical memory snapshot	357.5 MiB	371.9 MiB	-14.4 MiB	-4%				



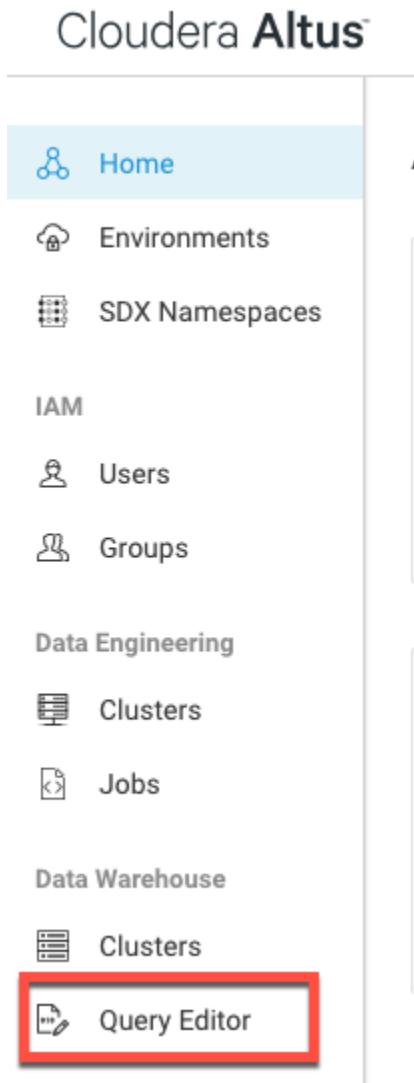
Finally, if the Altus cluster has not yet been terminated, you can navigate to Cloudera Manager and view the queries that were run from the Hive server web UI. **For the purposes of this lab, we have not configured the network to allow access to Cloudera Manager for the attendees.** Instead, please review the below screenshot showing the query from the Hive server UI.

<div> Home Local logs Metrics Dump Hive Configuration Stack Trace </div>								
User Name	Query	Execution Engine	State	Opened (s)	Closed Timestamp	Latency (s)	Drilldown Link	
hdfs	drop table if exists intermediate_access_logs	mr	FINISHED	2	Sun Feb 11 16:49:06 UTC 2018	2	Drilldown	
hdfs	CREATE EXTERNAL TABLE intermediate_access_logs (ip STRING, date STRING, method STRING, url STRING, http_version STRING, code1 STRING, code2 STRING, dash STRING, user_agent STRING) ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe' WITH SERDEPROPERTIES ('input.regex' = '([^\]*) - - \\[([^\]*)\] \[([^\]*) ([^\]*)" (\d*) (\d*)" ([^\]*)" ([^\]*)"', 'output.format.string' = "%1\$s %2\$s %3\$s %4\$s %5\$s %6\$s %7\$s %8\$s %9\$s")	mr	FINISHED	3	Sun Feb 11 16:49:10 UTC 2018	3	Drilldown	

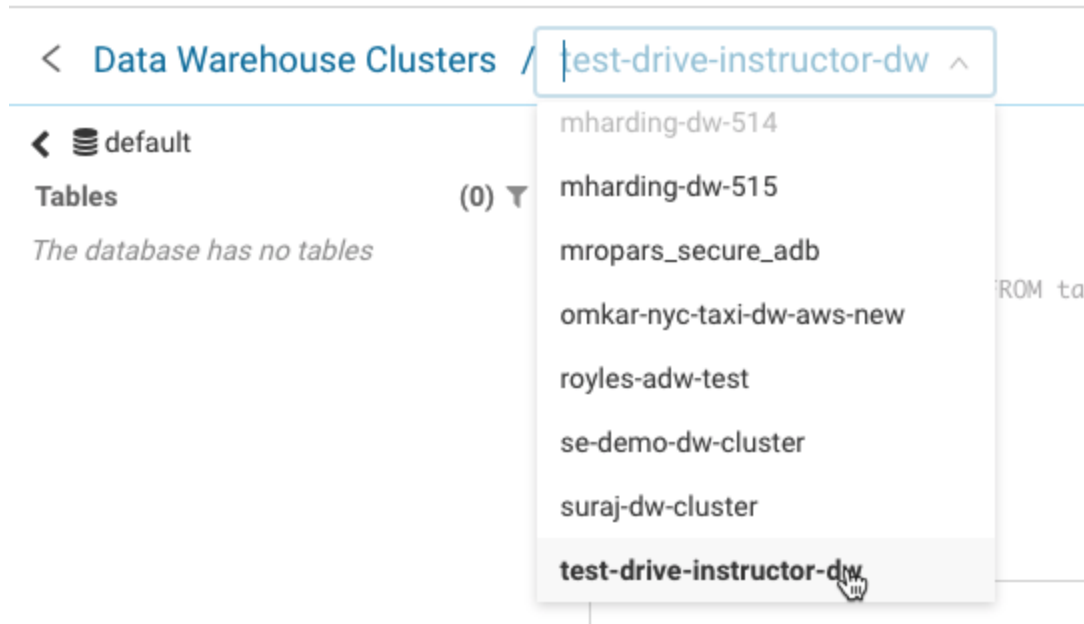
Exercise 7: Combine Structured and UnStructured data

Altus Data Warehouse (DW) clusters are used to run Analytics. A SQL Editor is built in and that allows you to run queries directly, or you can plug in a BI tool and use that if you prefer.

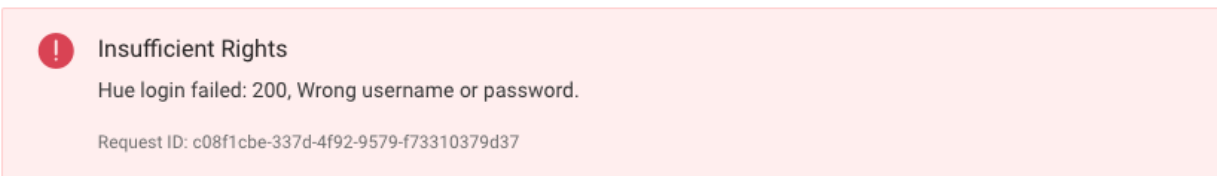
In this lab we'll go to the SQL Editor:



and choose the DW cluster we've created **test-drive-instructor-dw**:



If you see this:



then just refresh your screen and you'll be into the SQL Editor.

So now we're ready to run a query. But what are we querying? If we go back to Azure you can see that the structured warehouse data (i.e. the data about customers, orders etc) is stored in the ADLS account and mapped to tables:

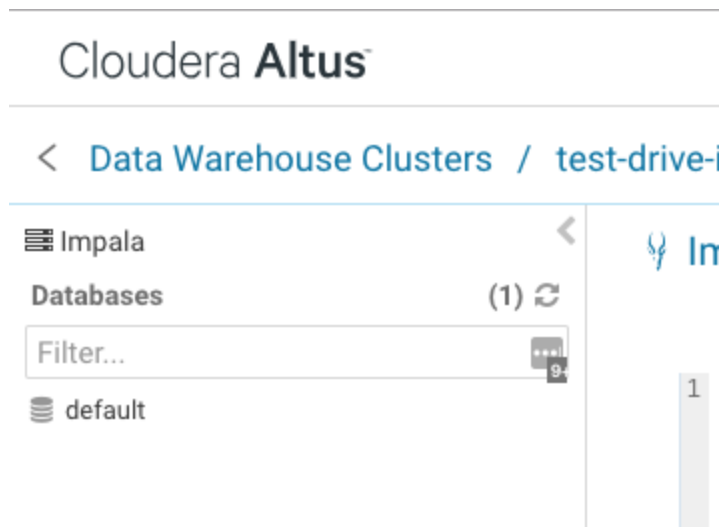
The screenshot shows the Azure Data Explorer interface. On the left, a folder tree under 'azurelabdata' shows 'altus-retail-demo' > 'data' > 'warehouse' selected. The main pane displays a table of the 'warehouse' contents:

NAME	SIZE	LAST MODIFIED
categories		6/12/2018, 3:49:31 PM
customers		6/12/2018, 3:49:29 PM
departments		6/12/2018, 3:49:27 PM
order_items		6/12/2018, 3:49:28 PM
orders		6/12/2018, 3:49:31 PM
products		6/12/2018, 3:49:28 PM

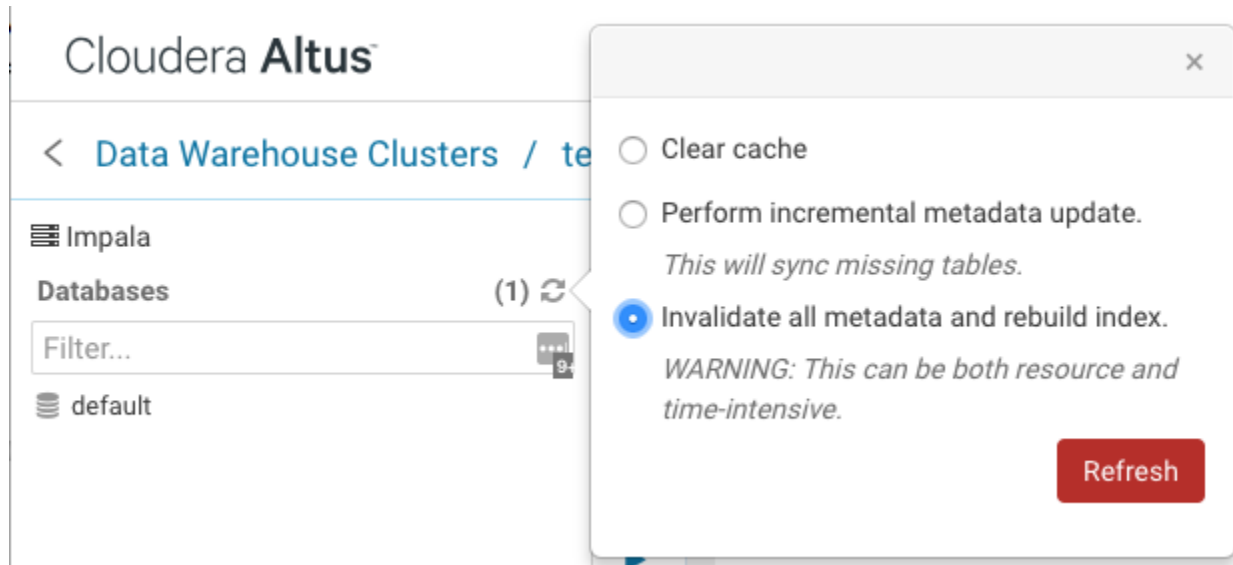
These tables are in the SDX namespace, altus-test-drive-namespace, under the **retaildb** already. You might have to invalidate the impala metadata to see the **retaildb** and the tables. (If you can see the **retaildb** and its tables you can skip this bit!):
Click the arrow to the right of **default**:

The screenshot shows the Cloudera Altus interface. The breadcrumb navigation shows 'Data Warehouse Clusters / test-drive-instruc'. On the left, 'default' is selected under 'Tables'. The main area displays the message: 'The database has no tables'. On the right, the 'Impala' logo is visible, and a small '1 Example' link is shown.

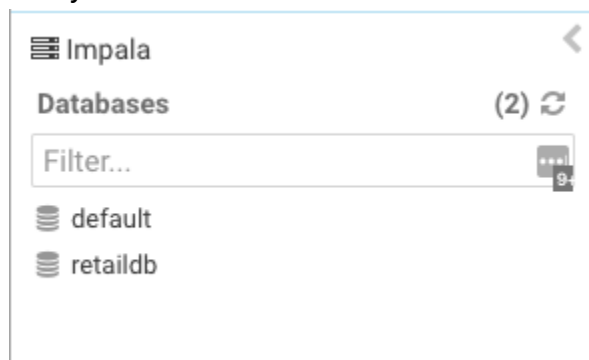
and you'll get to the 'top' of the database server, looking at all the possible databases:



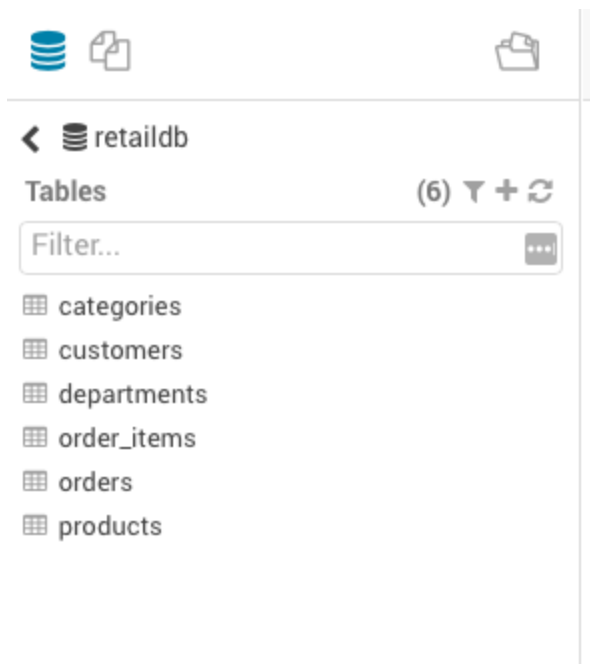
Right now you only see **default**. Select the little circle and choose 'invalidate metadata':



and you should now see the **retaildb**:



Click on **retaildb** and you should see the tables:



Copy the following SQL query into the SQL Editor:

```
-- Refresh the metadata
invalidate metadata;

-- Sample Queries
use retaildb;

--Products and URLs in one query
select product_name as 'Top products sold', regexp_replace(url,'%20',' ') as url
  from (select row_number() over(order by r.revenue desc) as r, p.product_name, r.revenue
        from products p
        inner join(select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float))
as revenue
                  from order_items oi
                  inner join orders o on oi.order_item_order_id = o.order_id
                  where o.order_status <> 'CANCELED'
                  and o.order_status <> 'SUSPECTED_FRAUD'
                  group by order_item_product_id) r
        on p.product_id = r.order_item_product_id
        order by r.revenue desc)as prod
  inner join(select row_number() over(order by count(*) desc) as r,url as url, count(*)
as count
            from ${YOURNAMEHERE}.tokenized_access_logs where url like '%/product/%'
            group by url order by count(*) desc) as web on prod.r=web.r limit 10;
```

Before you run it, enter your first initial/last name in the box at the bottom of the editor next to “YOURNAMEHERE”, just like you did earlier:

3/3



YOURNAMEHERE mkohs

Once you've done that, highlight ALL the text (Ctl-A on a Windows machine; Command A on a Mac, or drag with your mouse), and then click the little blue arrow to execute everything:

Impala

```

1 -- Refresh the metadata
2 invalidate metadata;
3
4 -- Sample Queries
5 use retaildb;
6
7 --Products and URLs in one query
8 select product_name as 'Top products sold', regexp_replace(url,'%20',' ') as url
9     from (select row_number() over(order by r.revenue desc) as r, p.product_name, r.revenue
10          from products p
11          inner join(select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as revenue
12                   from order_items oi
13                   inner join orders o on oi.order_item_order_id = o.order_id
14                   where o.order_status <> 'CANCELED'
15                   and o.order_status <> 'SUSPECTED_FRAUD'
16                   group by order_item_product_id) r
17          on p.product_id = r.order_item_product_id
18          order by r.revenue desc)as prod
19     inner join(select row_number() over(order by count(*) desc) as r,url as url, count(*) as count
20              from tferguson.tokenized_access_logs where url like '%\product\%'
21              group by url order by count(*) desc) as web on prod.r=web.r limit 10;

```

and you should see the results from querying the structured and unstructured data:

You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://altus-e45af1e8.azurelab3.com:8889>

HUE Query

Impala Add a name... Add a description...

5.88s retaildb text ?

retaildb (6)

Tables

Filter...

- categories
- customers
- departments
- order_items
- orders
- products

```

1 -- Sample Queries
2 use retaildb;
3
4 --Products and URLs in one query
5 select product_name as 'Top products sold', regexp_replace(url,'%20',' ') as url
6   from (select row_number() over(order by r.revenue desc) as r, p.product_name, r.revenue
7        from products p
8        inner join(select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as revenue
9                from order_items oi
10               inner join orders o on oi.order_item_order_id = o.order_id
11                where o.order_status <> 'CANCELED'
12                and o.order_status <> 'SUSPECTED_FRAUD'
13                group by order_item_product_id) r
14        on p.product_id = r.order_item_product_id
15        order by r.revenue desc)as prod
16   inner join(select row_number() over(order by count(*) desc) as r,url as url, count(*) as count
17        from spabba.tokenized_access_logs where url like '%\product/%'
18        group by url order by count(*) desc) as web on prod.r=web.r limit 10;
19
20

```

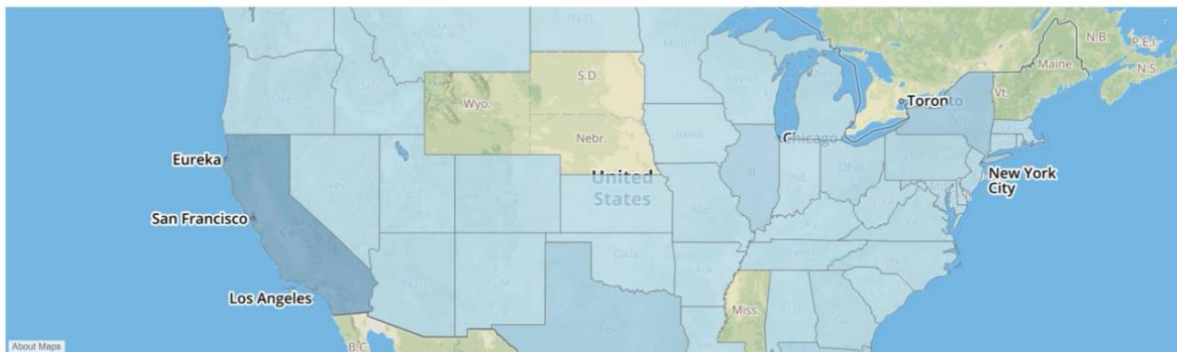
Query History Saved Queries Results (10)

	top products sold	url
1	Field & Stream Sportsman 16 Gun Fire Safe	/department/apparel/category/cleats/product/Perfect Fitness Perfect Rip Deck
2	Perfect Fitness Perfect Rip Deck	/department/apparel/category/featured shops/product/adidas Kids' RG III Mid Football Cleat
3	Diamondback Women's Serene Classic Comfort Bi	/department/golf/category/women's apparel/product/Nike Men's Dri-FIT Victory Golf Polo
4	Nike Men's Free 5.0+ Running Shoe	/department/apparel/category/men's footwear/product/Nike Men's CJ Elite 2 TD Football Cleat
5	Nike Men's Dri-FIT Victory Golf Polo	/department/fan shop/category/water sports/product/Pelican Sunstream 100 Kayak
6	Pelican Sunstream 100 Kayak	/department/fan shop/category/indoor/outdoor games/product/O'Brien Men's Neoprene Life Vest
7	O'Brien Men's Neoprene Life Vest	/department/fan shop/category/camping & hiking/product/Diamondback Women's Serene Classic
8	Nike Men's CJ Elite 2 TD Football Cleat	/department/fan shop/category/fishing/product/Field & Stream Sportsman 16 Gun Fire Safe
9	Under Armour Girls' Toddler Spine Surge Runni	/department/footwear/category/cardio equipment/product/Nike Men's Free 5.0+ Running Shoe
10	adidas Youth Germany Black/Red Away Match Soc	/department/footwear/category/fitness accessories/product/Under Armour Hustle Storm Medium

Conclusion: Visualizing the results.

Your instructor will show you a visualization of the results of your job using a BI Tool.

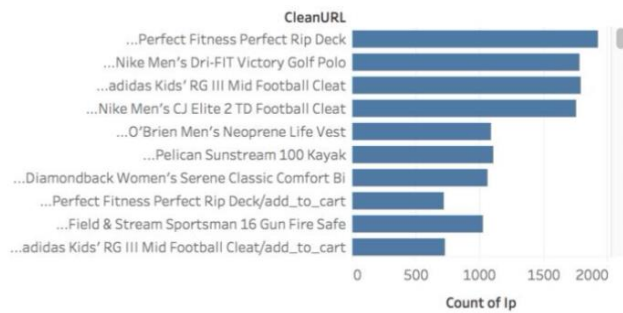
Sales by State



Top Selling Products



Top Browsed Product URLs



This brings us to the end of the lab. We've discussed the value Cloudera Altus on Azure brings to data engineers and business intelligence users. Cloudera Altus allows end users to be self-sufficient and easily correlate disparate data sources for faster insights. Thank you for your time and patience!