Cloudera Altus on Azure - Hands-on Workshop

Exercise Manual

Created by Cloudera

## 

[**Summary**](#_27zcunqi9vbg) **1**

[**Exercise 1: Logging in to Altus**](#_nh0q1u7u39vk) **1**

[**Exercise 2: Getting to know the Altus UI**](#_6dpwcp7krw6a) **4**

[**Exercise 3: How to create a cluster**](#_9oparr1bfgc8) **9**

[**Exercise 4: Preparing the object store**](#_c2y1eh3n5cov) **12**

[**Exercise 5: Running a Job.**](#_uf2a13jbrhj5) **17**

[**Exercise 6: Workload Analytics**](#_18zg4wyssknj) **20**

[**Exercise 7: Combine Structured and UnStructured data**](#_2l2i8go5f4b6) **25**

[**Conclusion: Visualizing the results.**](#_wf2siv4h9o8x) **28**

## Summary

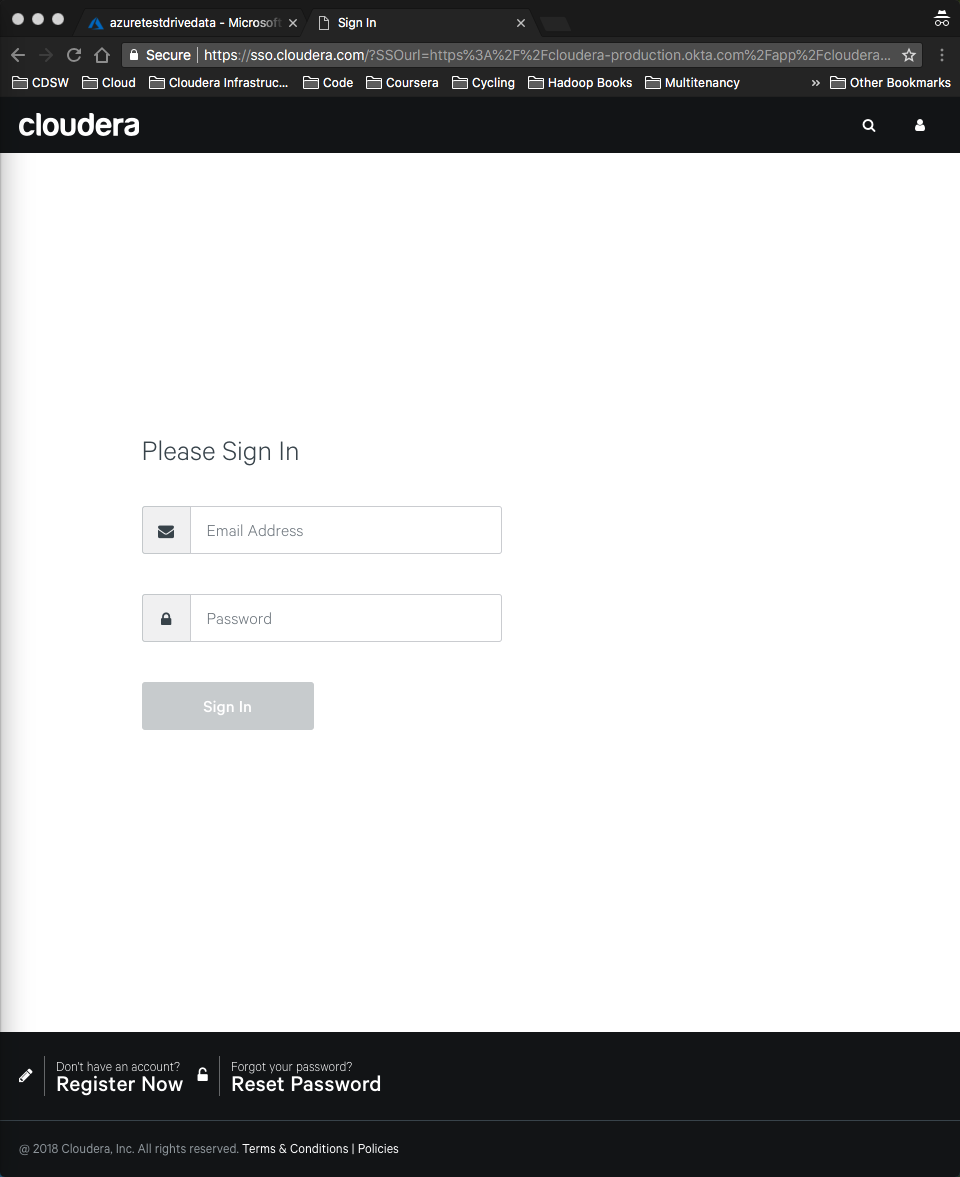
Welcome to the Altus on Azure Hands-on-Lab attendee guide. Below are a series of exercises designed to familiarize yourself with the Cloudera Altus platform offering. In addition to this, we will walk through a simple data engineering problem and how Altus empowers the end users to solve it and gain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

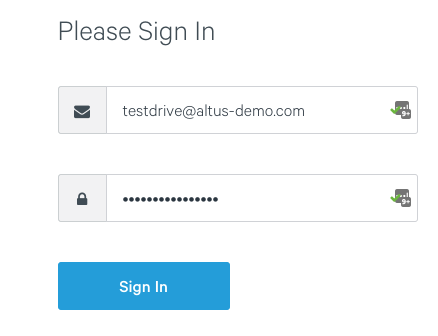
<http://altus.cloudera.com/>



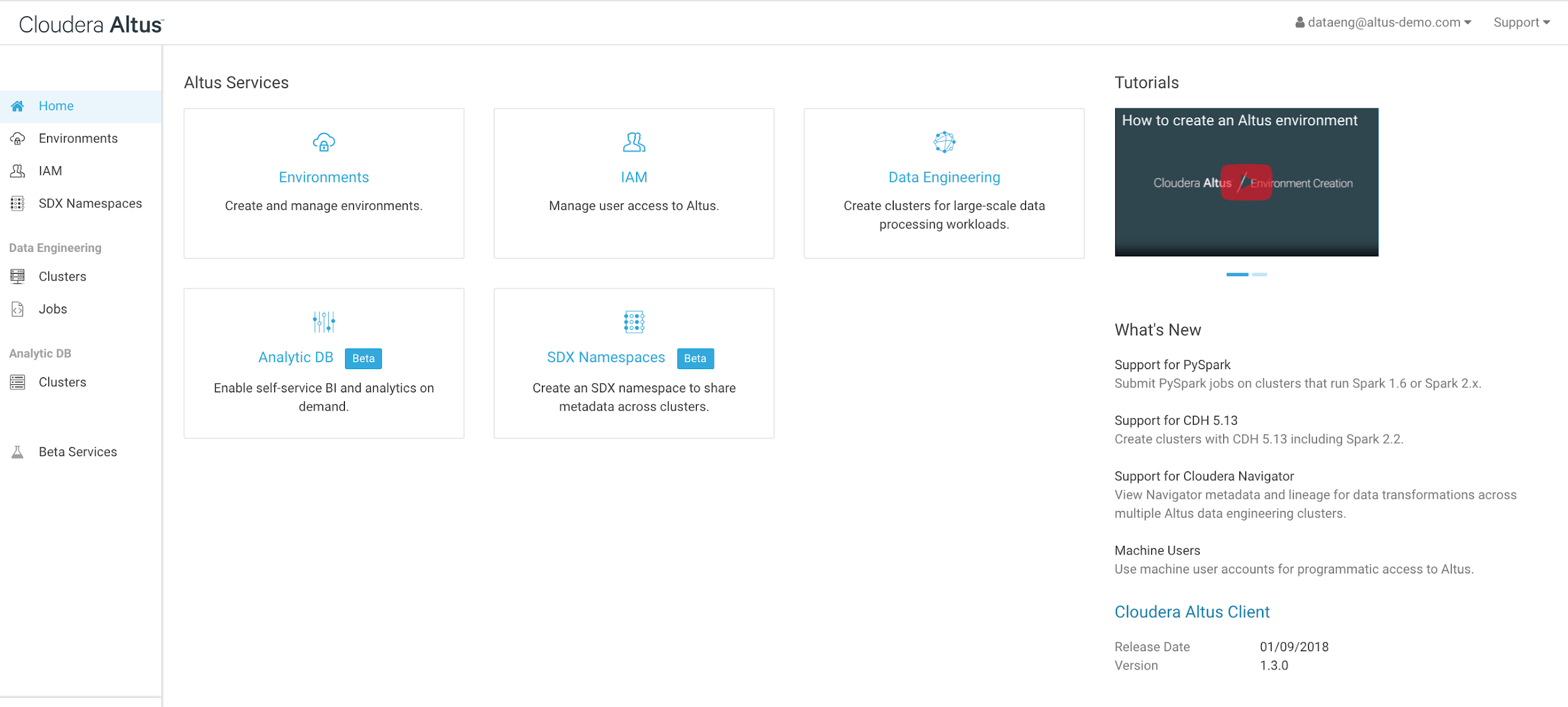
Login with your username and password. Please note that this is a shared login which will be used by all lab attendees.

Username: [testdrive@altus-demo.com](mailto:dataeng@altus-demo.com)

Password: In the email you received when you registered.



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

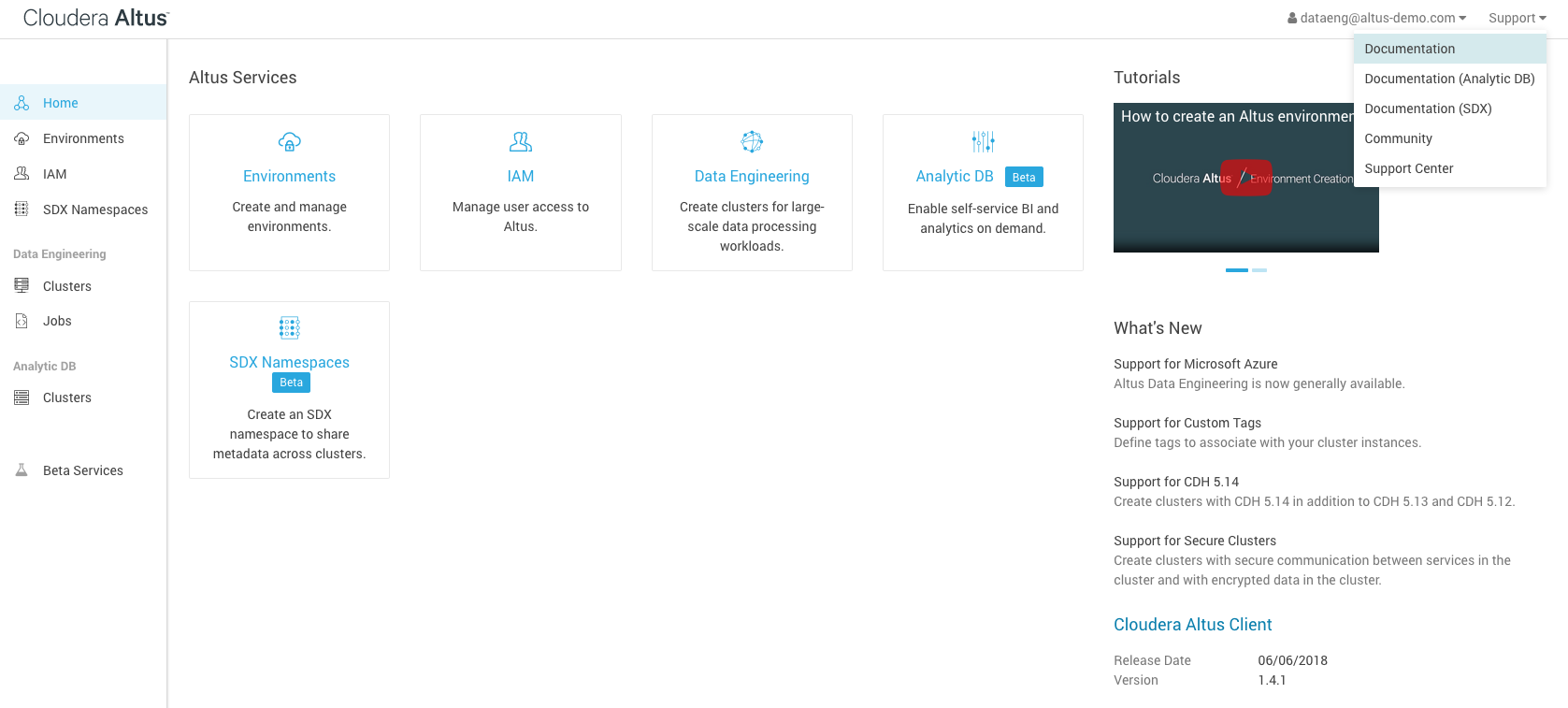


The Altus organization you’re currently logged in to is called the Altus-demo organization. This organization is used by Cloudera Engineers for demos as well as hands-on-labs, just like today’s! Since this is a live demo environment, **please do NOT delete any clusters or resources.**

## 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

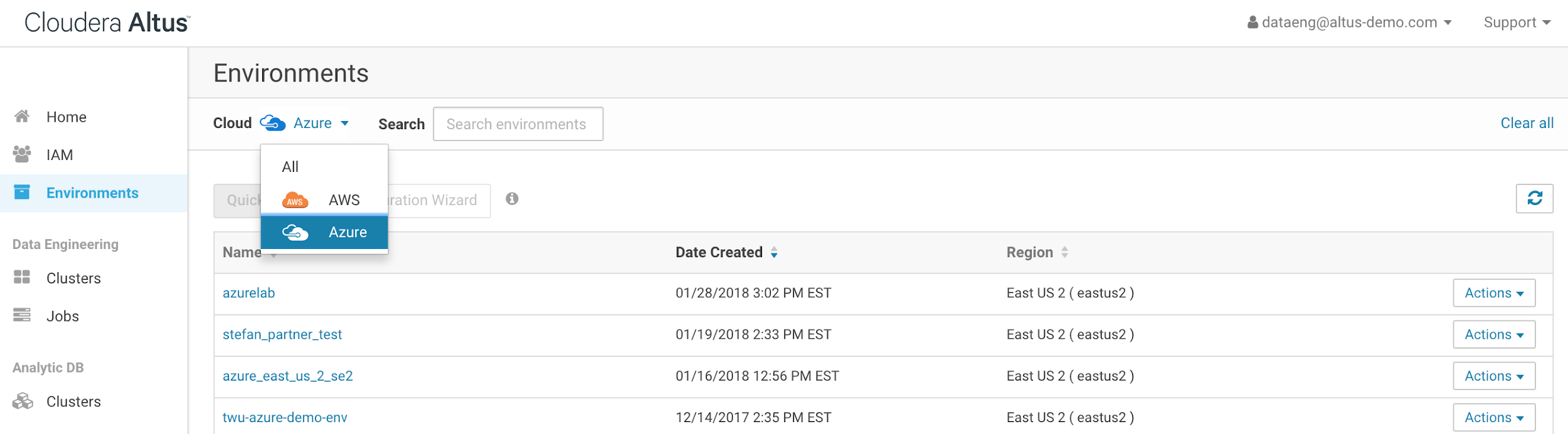


You can visit these docs at any time in case you have further questions about Altus components.

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.

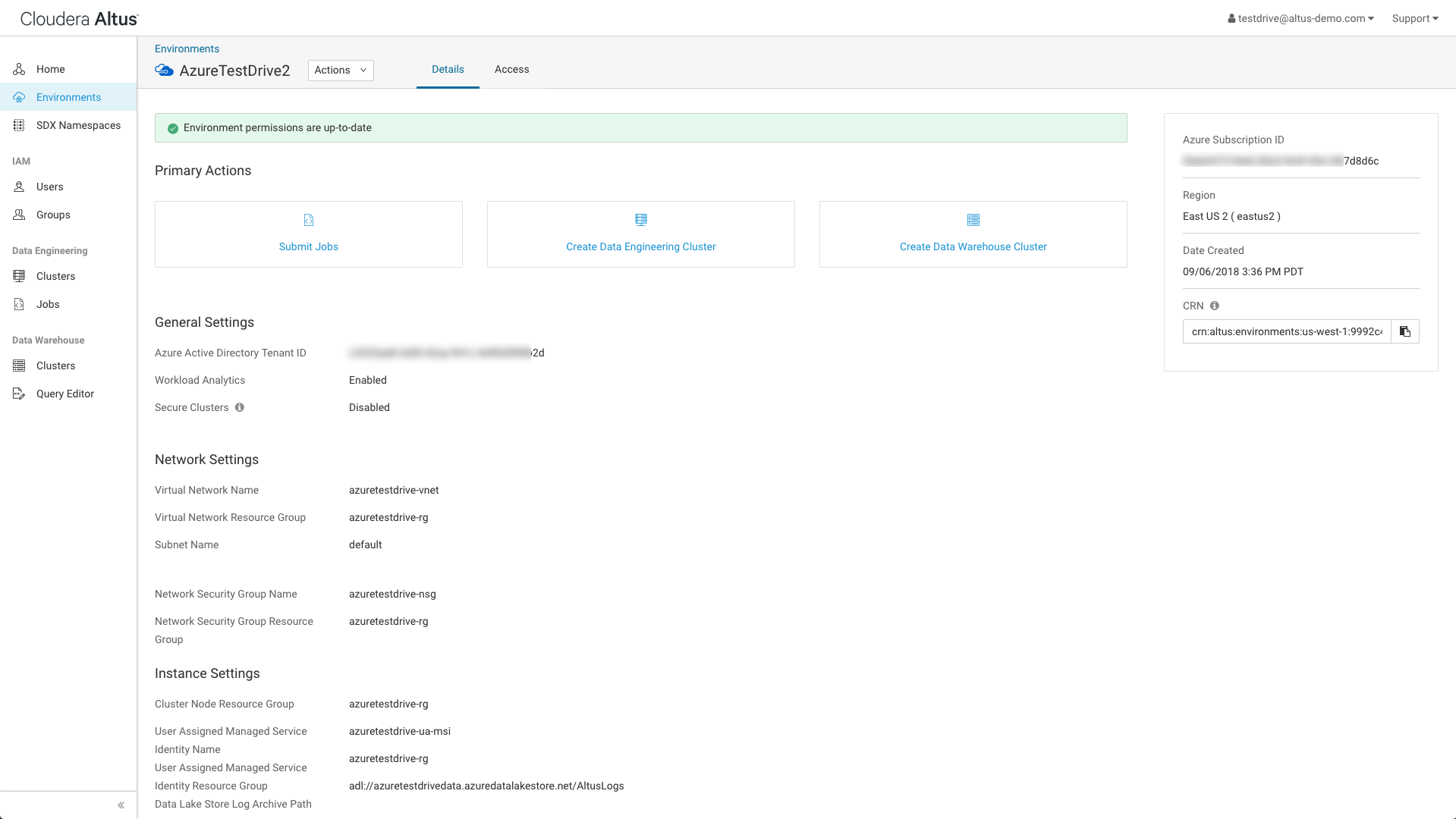


Click on Environments in the left-hand pane. Then click on the Cloud button in the main pane and select Azure.



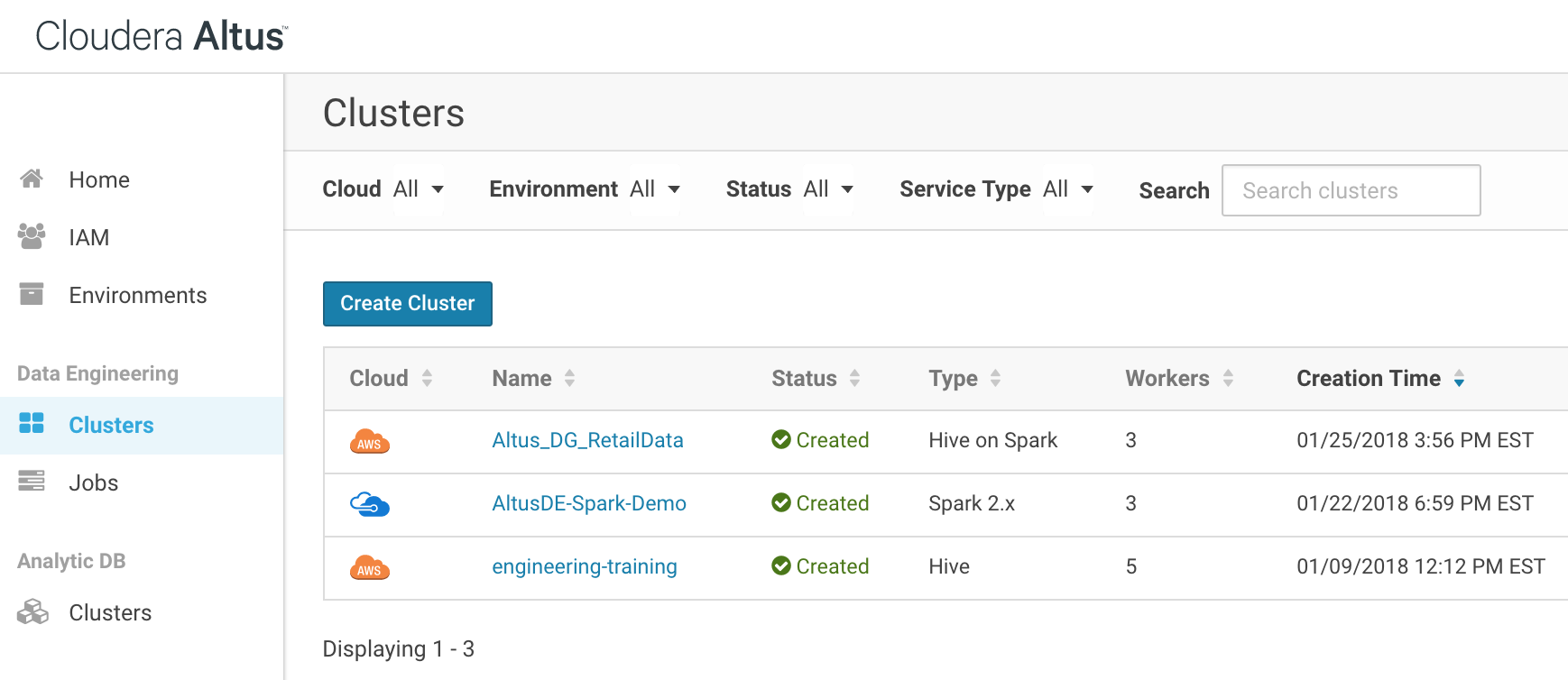
**An Altus environment**: Defines the resources in your Azure subscription that are used by Cloudera Altus to create clusters and jobs. In a production environment, an administrator can set up and assign separate Altus environments to different users and groups. **Please do not create any new environments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

Click on the environment called “**AzureTestDrive2**” and view the environment details:



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.



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](mailto:dataeng@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:**

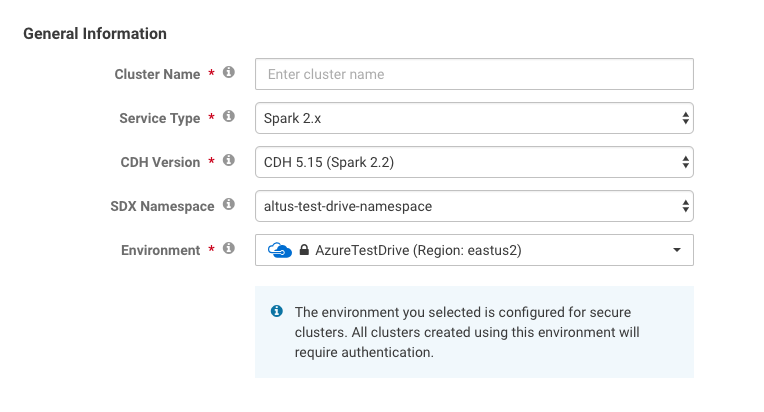
Cluster Name: **leave blank (since we are not going to create a cluster at the end)**

Service Type: **Hive on Spark**

CDH Version: **CDH 5.14**

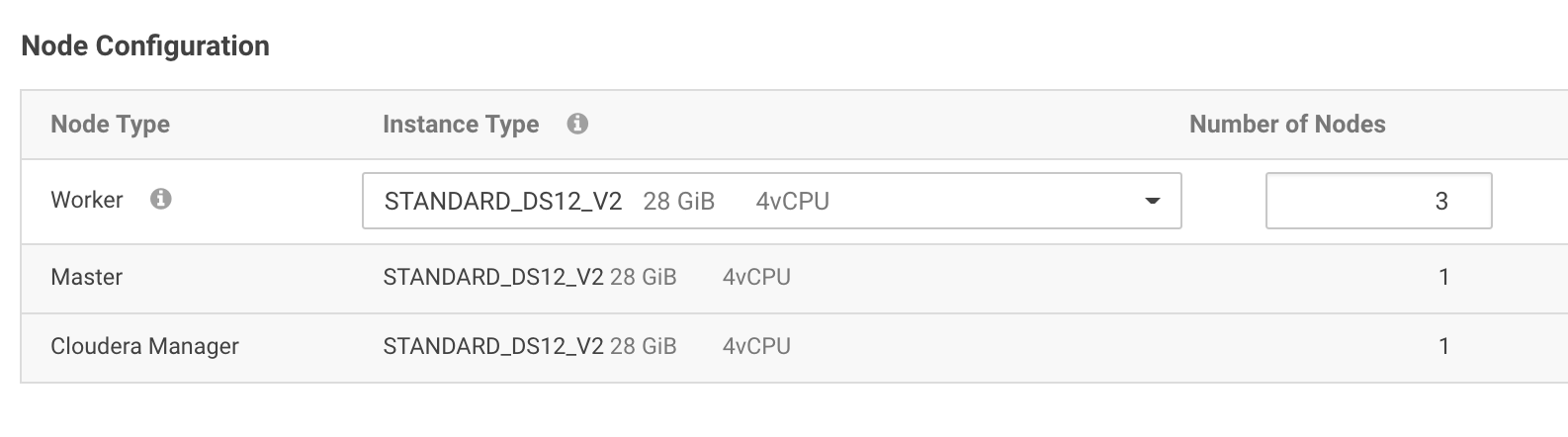
SDX Namespace: **altus-test-drive-namespace**

Environment: **AzureTestDrive2**



**Node Configuration:**

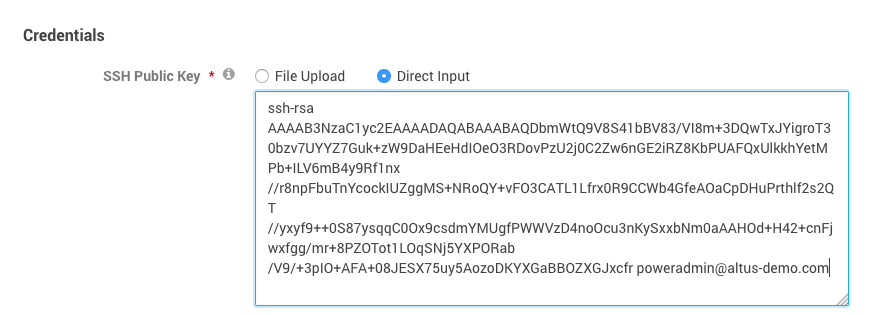
Worker: **Set the Number of Nodes to 3** (instead of 5).



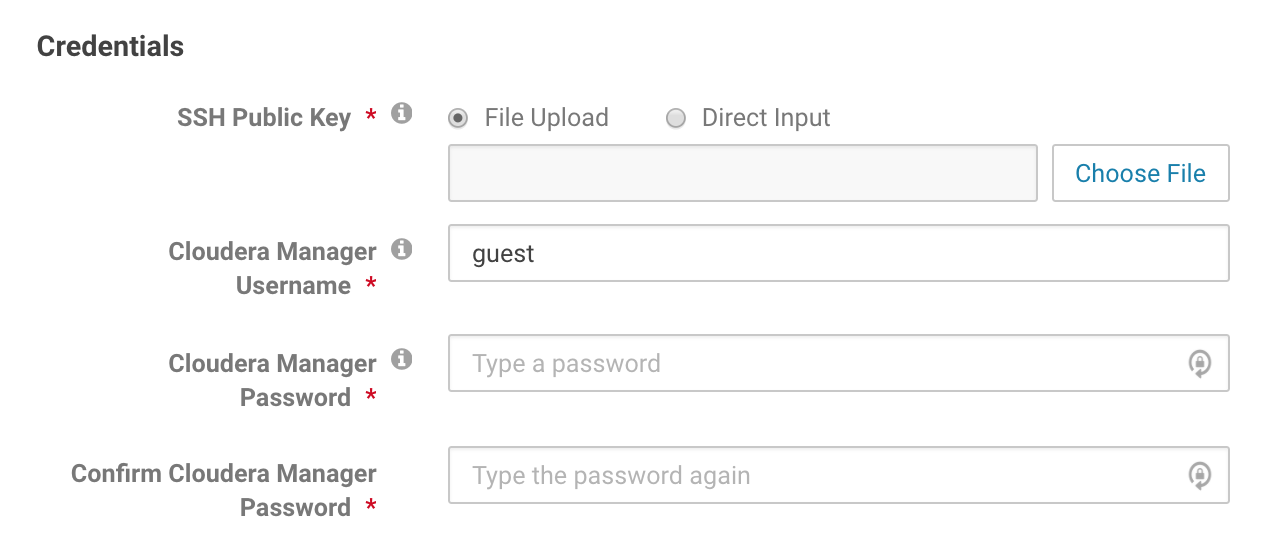
**Credentials:**

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:

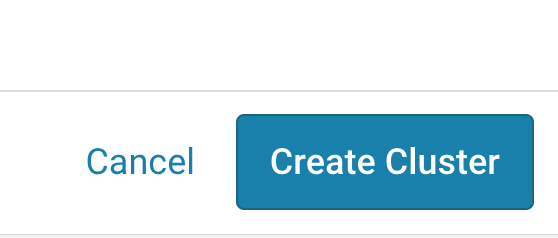
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQDbmWtQ9V8S41bBV83/VI8m+3DQwTxJYigroT30bzv7UYYZ7Guk+zW9DaHEeHdIOeO3RDovPzU2j0C2Zw6nGE2iRZ8KbPUAFQxUlkkhYetMPb+ILV6mB4y9Rf1nx//r8npFbuTnYcockIUZggMS+NRoQY+vFO3CATL1Lfrx0R9CCWb4GfeAOaCpDHuPrthlf2s2QT//yxyf9++0S87ysqqC0Ox9csdmYMUgfPWWVzD4noOcu3nKySxxbNm0aAAHOd+H42+cnFjwxfgg/mr+8PZOTot1LOqSNj5YXPORab/V9/+3pIO+AFA+08JESX75uy5AozoDKYXGaBBOZXGJxcfr poweradmin@altus-demo.com



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



Now click Cancel**. Please do NOT CLICK** “**Create Cluster”**.



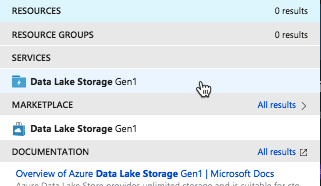
## Exercise 4: Preparing the object store

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

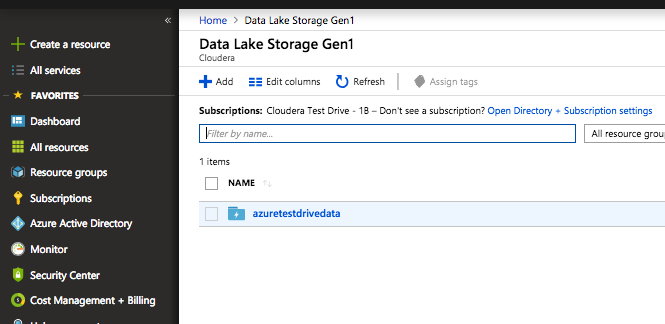
Once you’ve logged in to the Azure portal, go to the search bar at the top of the screen and type in “data lake storage”



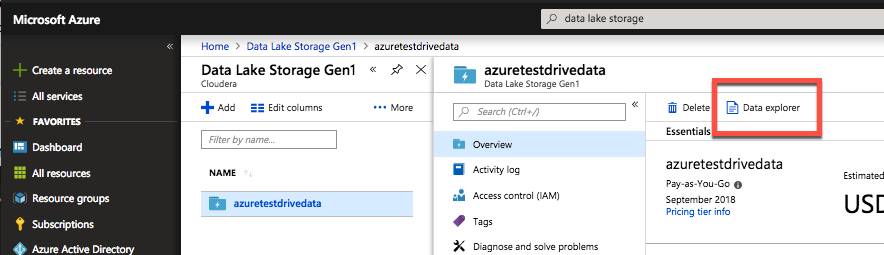
Click on “Data Lake Storage Gen 1” under Services



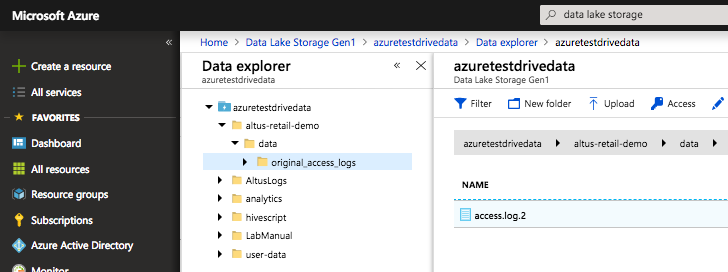
Click on **azuretestdrivedata**



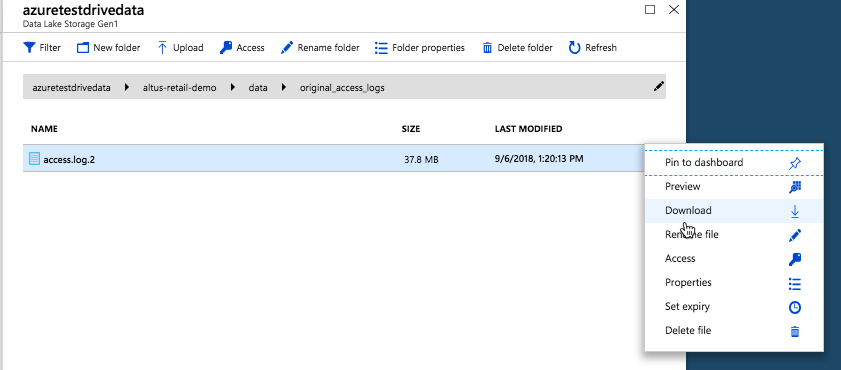
Click on “**Data Explorer**” which is in the right-most pane at the top:



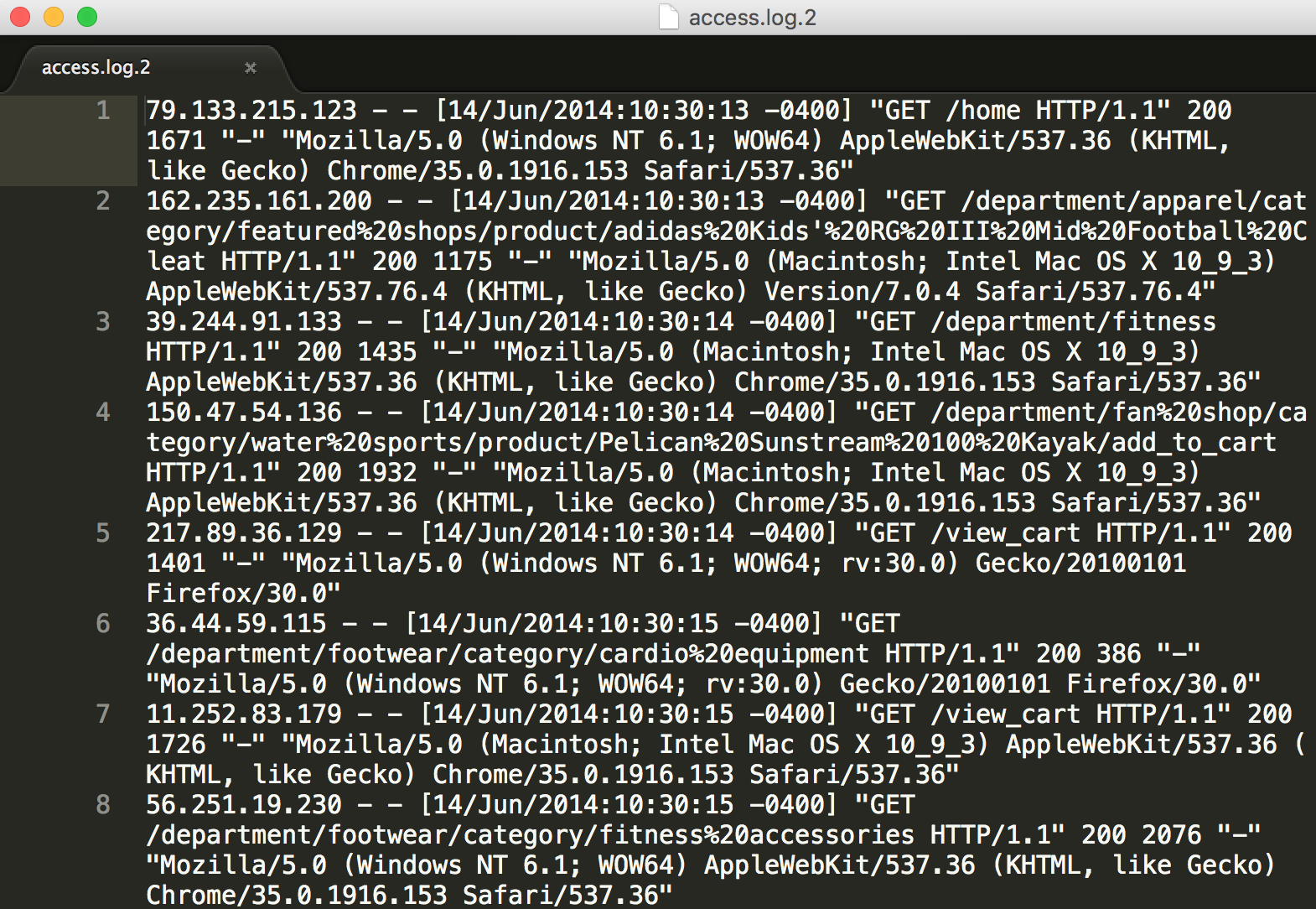
Navigate to the **azuretestdrivedata** ADLS folder. Go into the **altus-retail-demo/data/original\_access\_logs** folder.



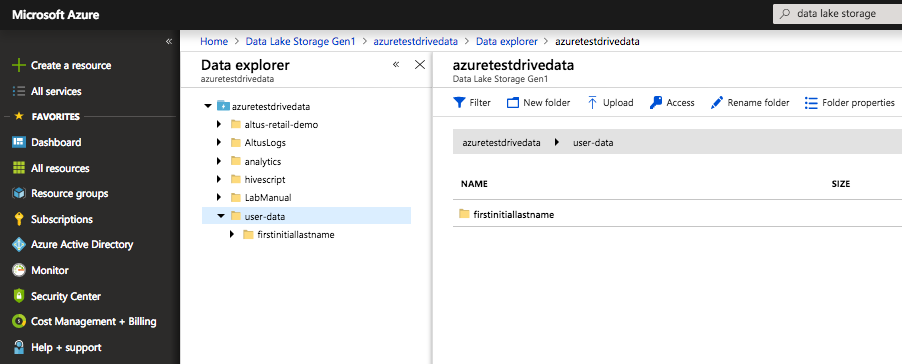
Inside the **original\_access\_logs** folder is **the access.log.2** file which contains weblog data. Download the file (by clicking on the three dots next to the filename) and open it up with your favourite text editor (notepad or text editor).



This file looks similar to below:



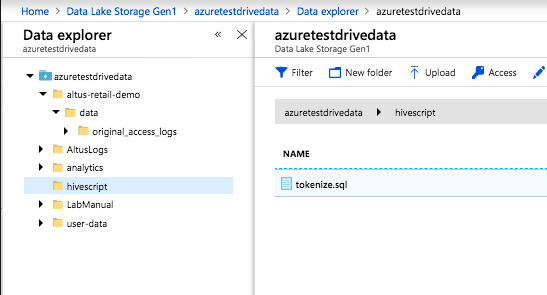
Navigate back to the top-level **azuretestdrivedata** folder and go into **user-data** folder.



**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.



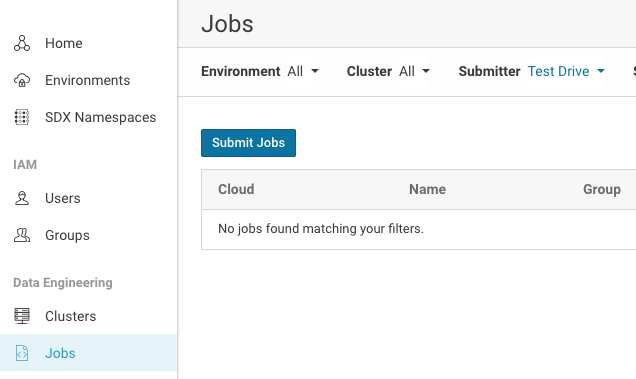
Open and review the tokenize.sql file. This is the Data Engineering query you are going to run.

|  |
| --- |
| -- 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$$s %2$$s %3$$s %4$$s %5$$s %6$$s %7$$s %8$$s %9$$s")  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';  add jar /opt/cloudera/parcels/CDH/lib/hive/lib/hive-contrib.jar;  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**.



**Submission**: Single Job

**Job type**: Hive

**Job Name**: use your name

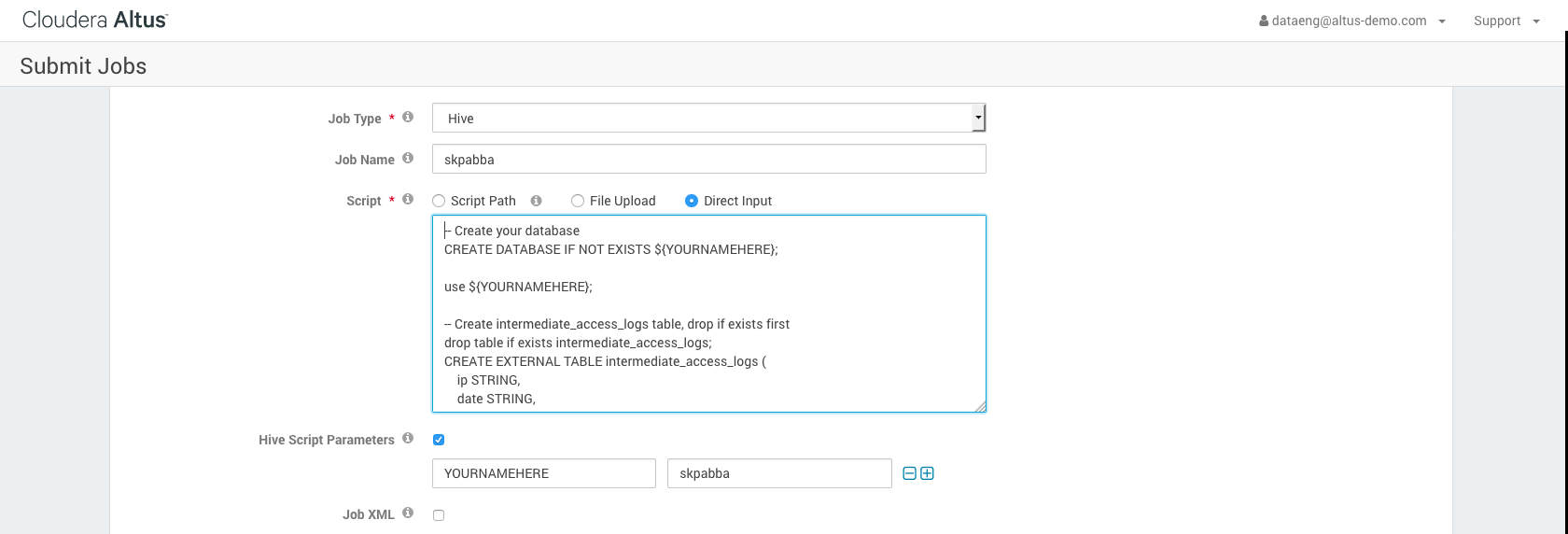
**Script**: Simply select ‘direct input’ and copy/pasted the tokenize.sql code

**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.

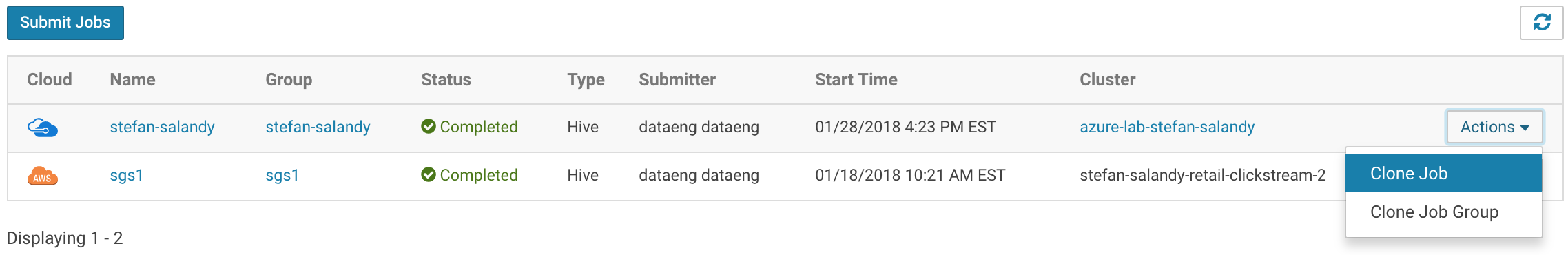
**Job XML**: None

**Action on Failure**: None

**Cluster**: Use the cluster name you got in the email.



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.

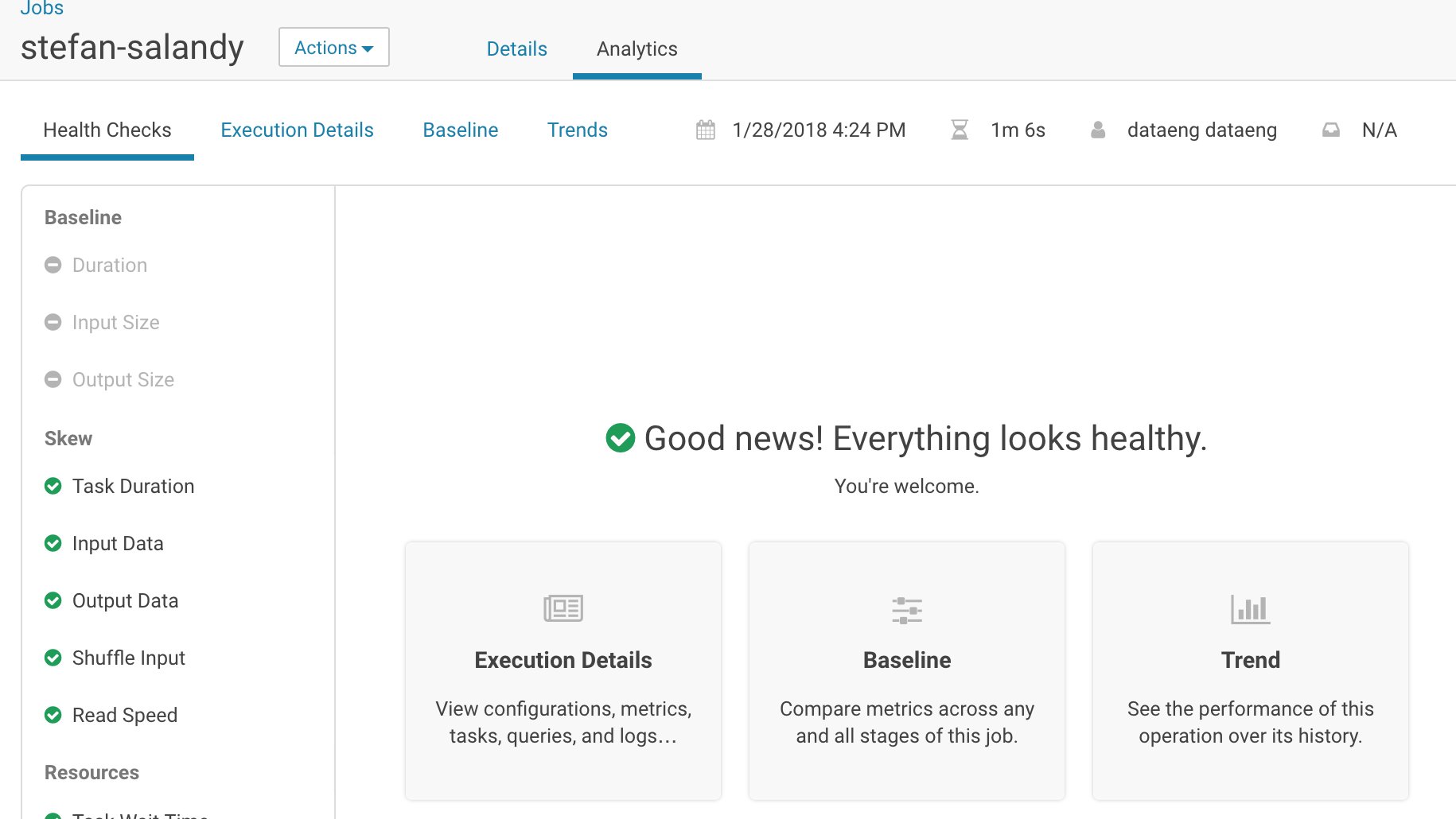


## Exercise 6: Workload Analytics

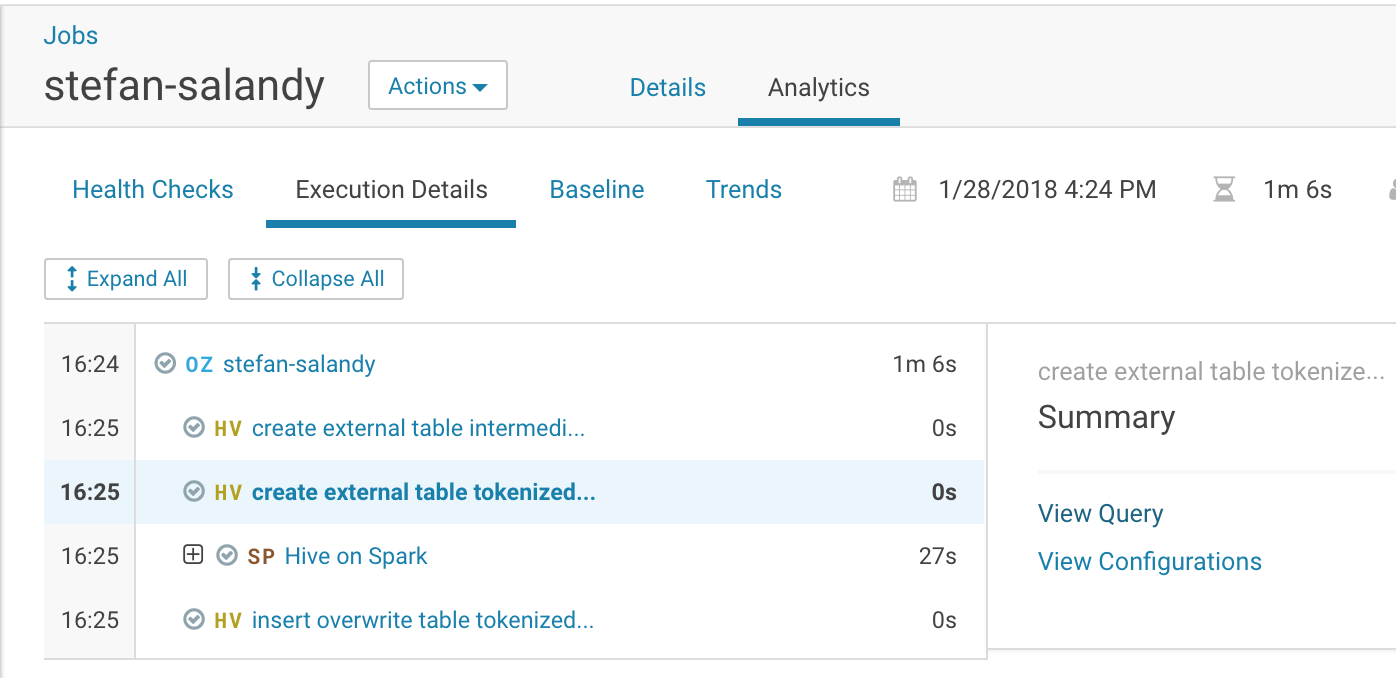
Click on your completed job. Here you can view details about the job you just ran.



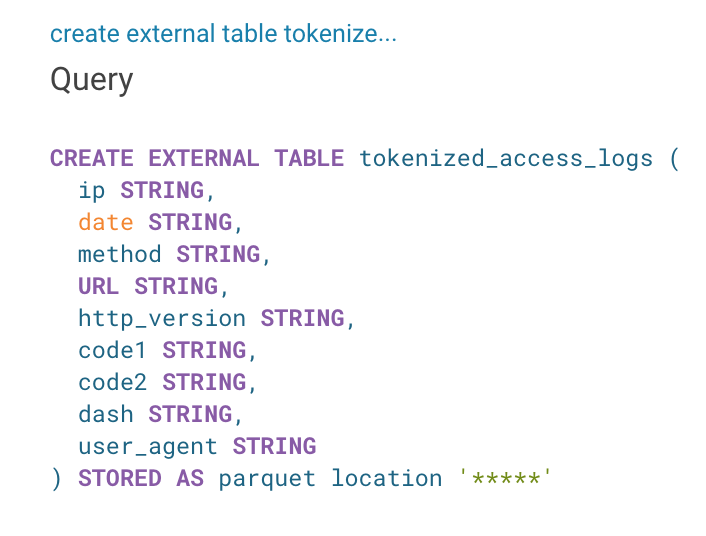
Click on **Analytics**. It might take a few minutes to appear.



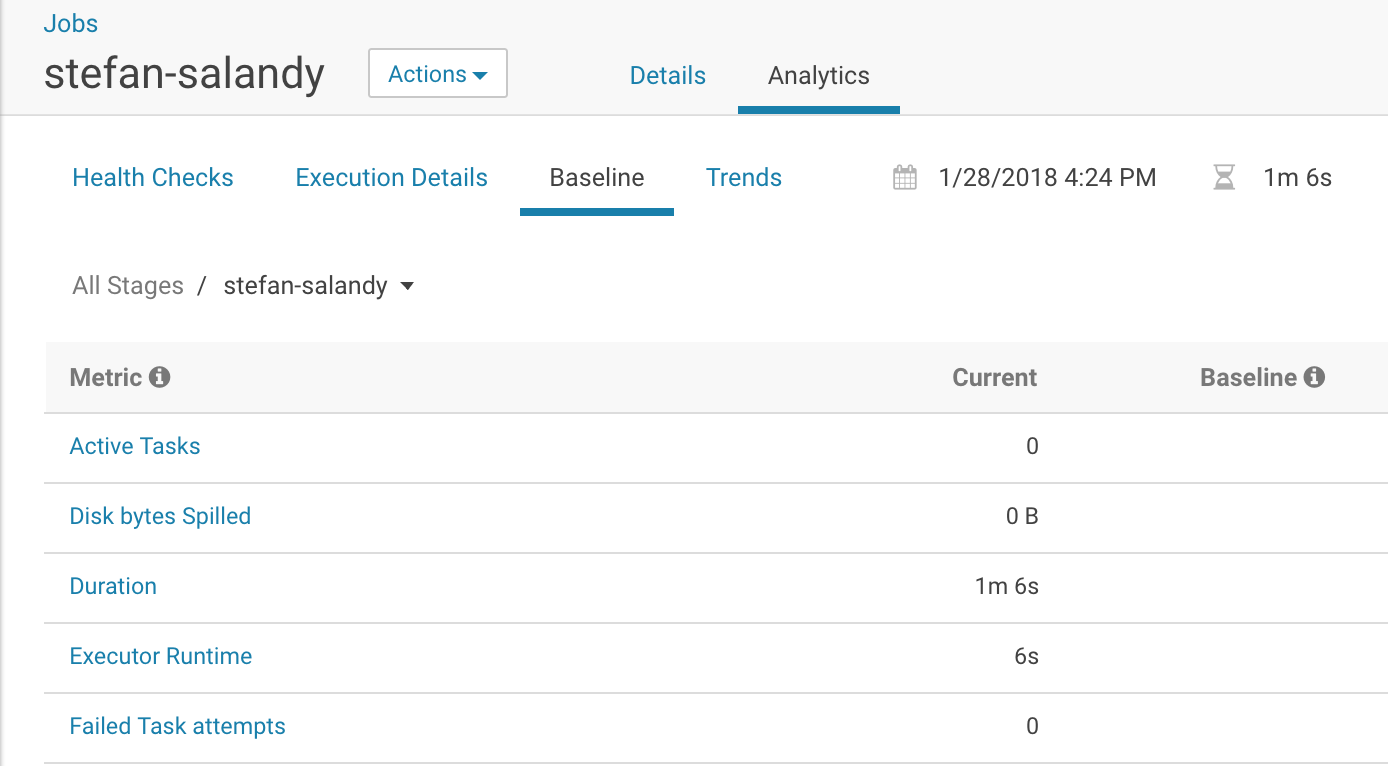
Click on **Execution Details** and then click on any one of the SQL queries. Then click on **View Query**.



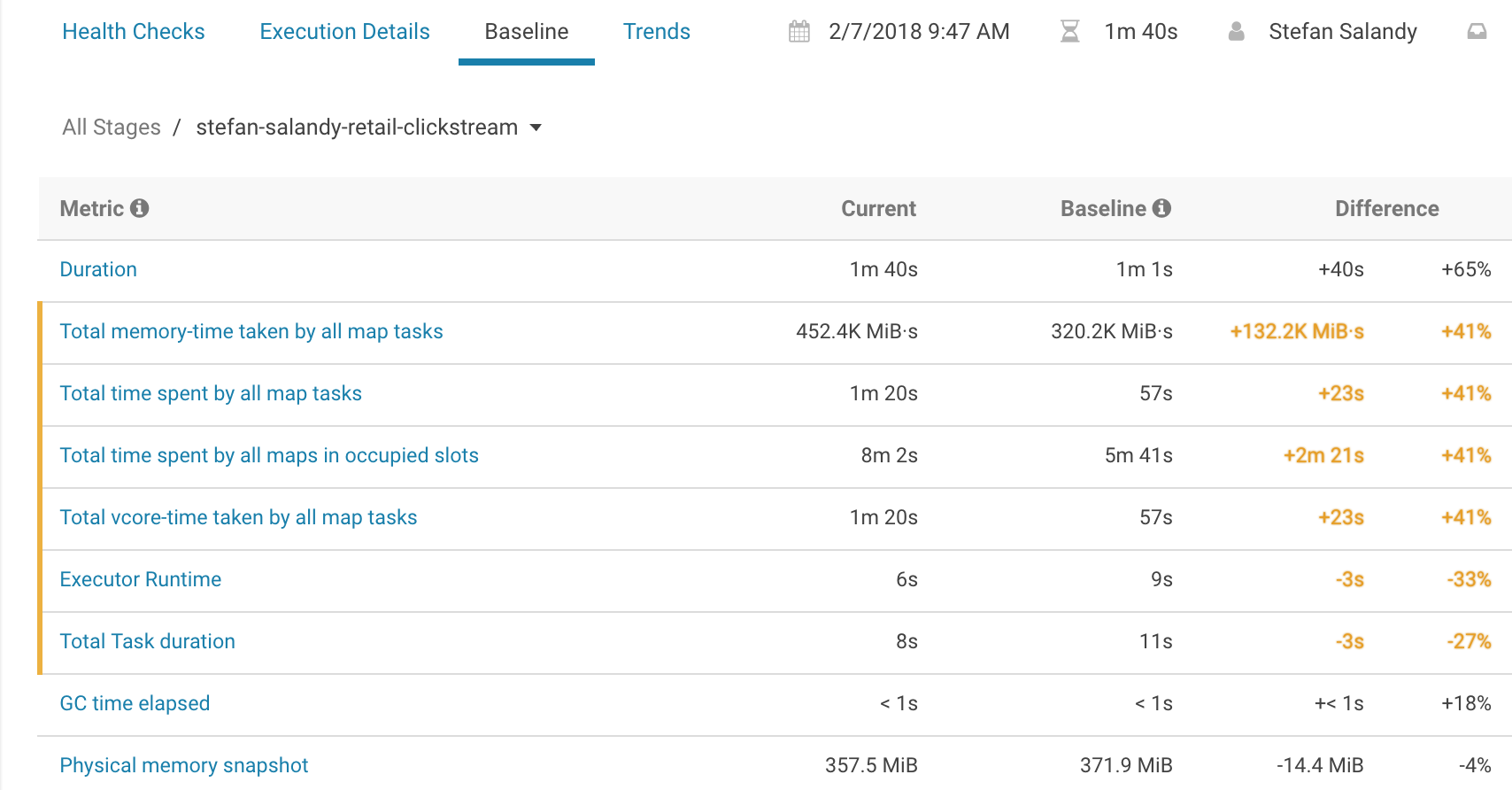
Looks familiar? It’s part of the SQL script that you ran.

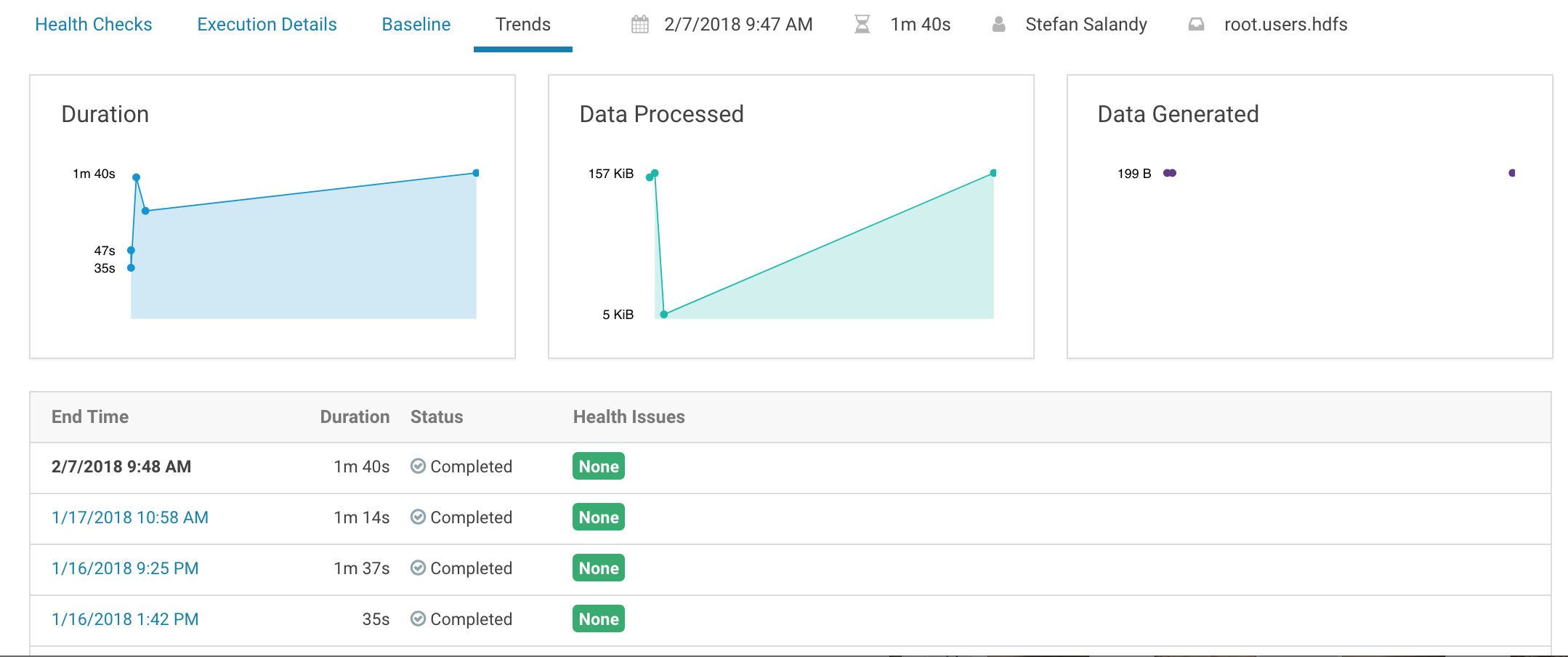


Click on Baseline to view more info about your job. Note that since this is your first time running a job, Altus Workload Analytics does not have enough info from previous runs to create the baseline.

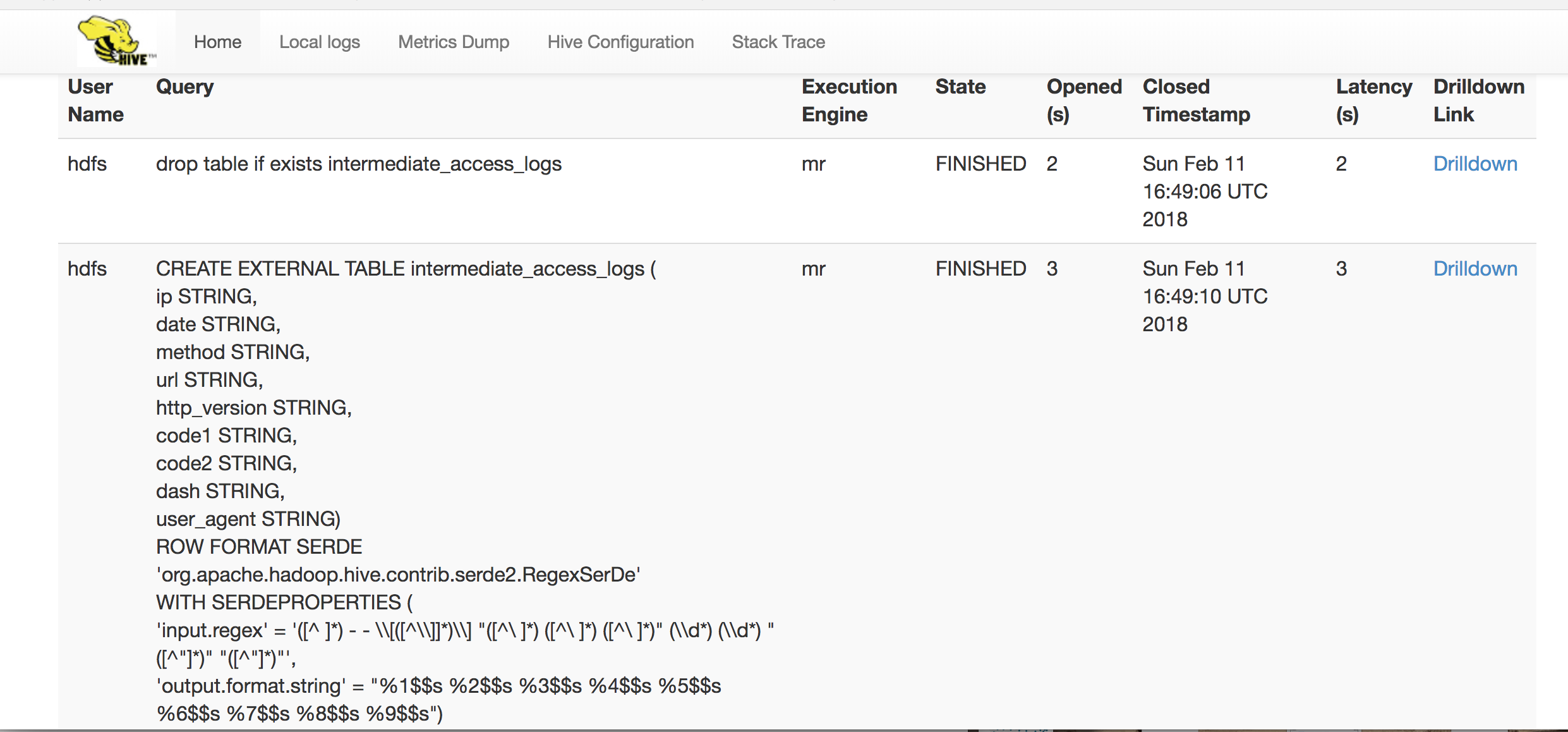


For example purposes we have included screenshots of previously run jobs. These jobs have Baseline and Trend analysis:



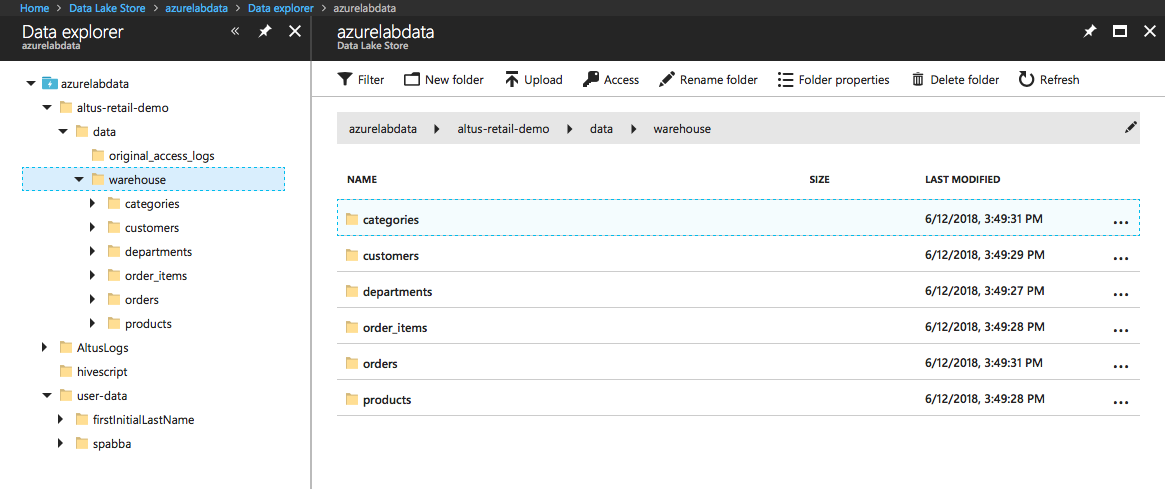


Finally, if the Altus cluster has not yet been terminated, you can navigate to Cloudera Manager and view the queries that were run from the Hive server web UI. **For the purposes of this lab, we have not configured the network to allow access to Cloudera Manager for the attendees.** Instead, please review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

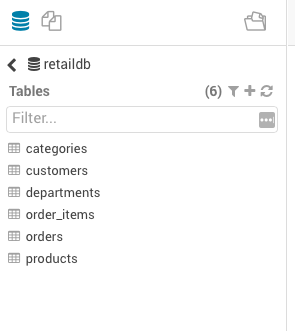


## Exercise 7: Combine Structured and UnStructured data

Altus Data Warehouse (DW) clusters can be used to run Analytics. The Structured warehouse data is stored in the ADLS account and mapped to tables,



These tables are in the SDX namespace, altus-test-drive-namespace, under retaildb already.

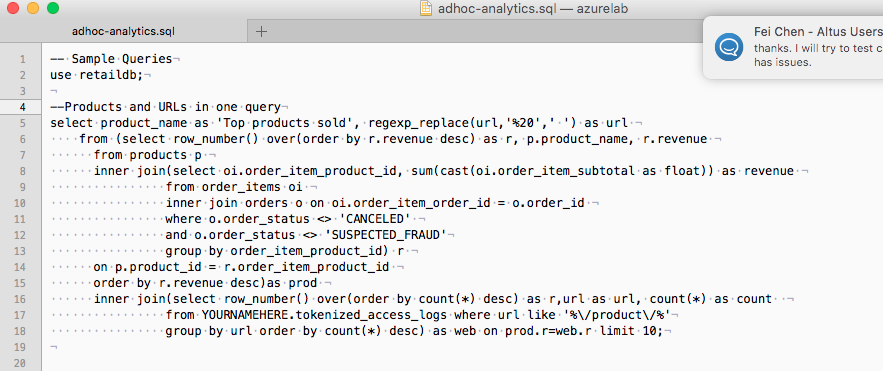


The DE job you ran in Exercise 5, converted unstructured log files into tokenized logs table. This table is also in the same namespace. Using Hue on an DW cluster you will be able to run queries where you can correlate structured Sales Data and unstructured Website Access Logs.

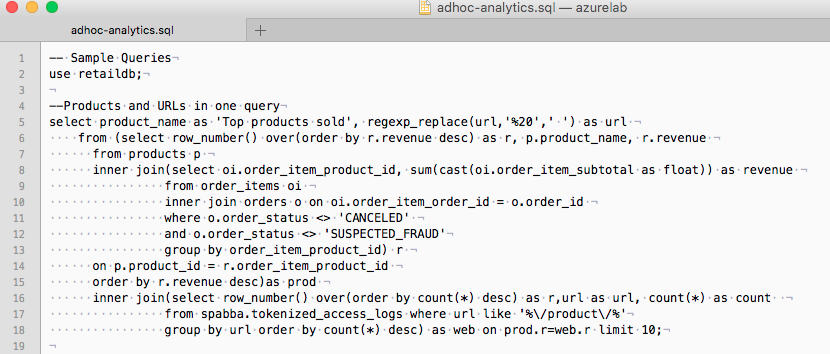
Download the adhoc-analytics SQL from the **azuretestdrivedata** ADLS account, **analytics** folder.



Open the adhoc-analytics.sql in your favorite text editor and search for the text “YOURNAMEHERE”.

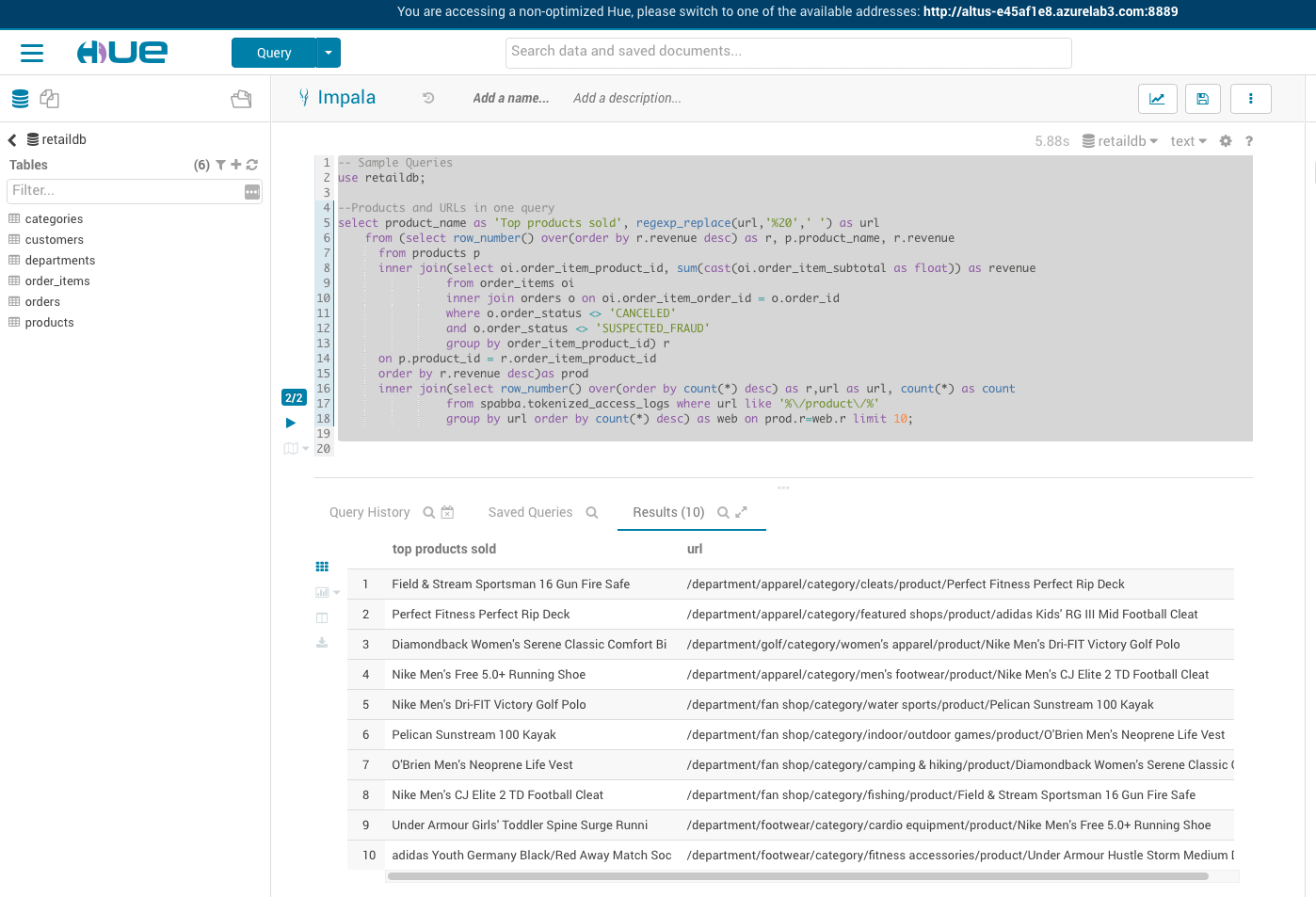


Replace all instances of “YOURNAMEHERE” with the exact name of the sub-folder you created. Save the file.



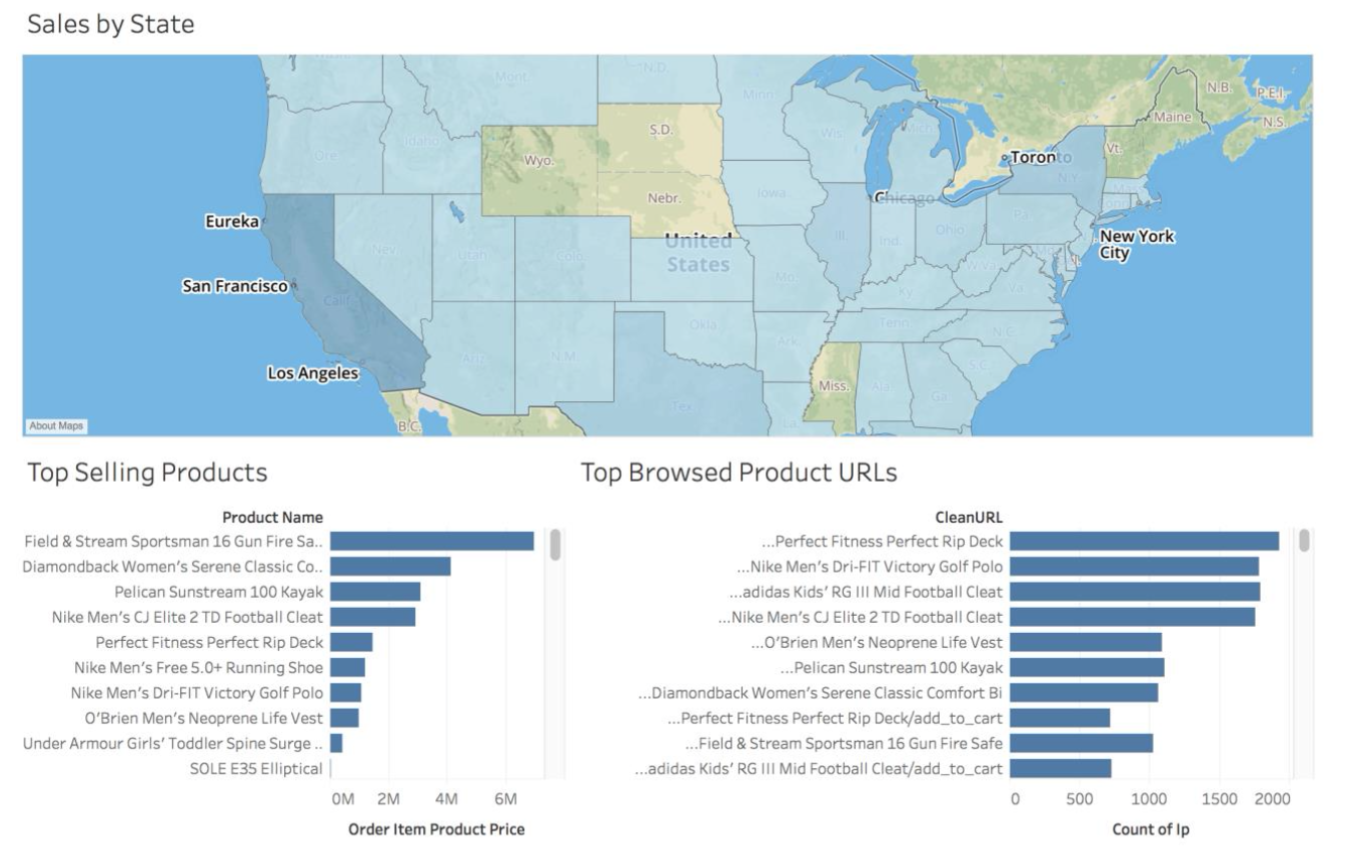
Instructor will give you URL for Hue and access credentials.

Login to Hue and run the SQL query as shown below. You should see the results from both Structured and UnStructured data.



## Conclusion: Visualizing the results.

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



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!