

CLUDERA DATA PLATFORM DATA WAREHOUSE LAB

Step-by-step instructions:

If you can read this you found the PDF. If you are reading it in a Browser please download the PDF to your local computer. This will let you cut and paste from the PDF into the Cloudera Web User Interface.

Part 1 - Data Catalog [20 minutes]

Overview: What is Cloudera Data Catalog?

Data Catalog is a service that enables you to understand, manage, secure, and govern data assets across the enterprise. Data Catalog helps you understand data across multiple clusters and across multiple CDP environments. You can search to locate relevant data of interest based on various parameters. Using Data Catalog, you can understand how data is interpreted for use, how it is created and modified, and how data access is secured and protected.

Purpose: Search for a dataset (table) in Data Catalog, called “flights”.

- Find what database(s) the table “flights” is located.
- Find out at least one year that the “flights” table was generated from.
- Find out how many columns the table “flights” contains.

1) Open CDP, using the “admin” user within the Test Drive link.

Your link should look something like (remember click the link in your email not the link below)

http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X

*xx represents the trial user #

*X represents the password

2) Click the “Data Catalog” within the CDP Home Screen



3) Type “flights” in the search box and press enter on your keyboard

Data Catalog / Search

Search flights

Data Lakes

cdptrialuser43-datalake 240

Filters

TYPE

☐ Hive Table

☐ HBase Table

+ Add New Value

OWNERS

☐ atlas

☐ csso_trialuser43

☐ public

☐ trial43_admin

Setup the Profiler for cdptrialuser43-datalake

Profiler runs data profiling operations as a pipeline on data located in the data lake which helps to view the profiled results for assets. Setting up a Profiler launches a cluster with various services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. [Get Started >](#)

cdptrialuser43-datalake | 240

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	flights	airlines_new_orc.flights@cm	Mon Aug 30 2021	csso_trialuser43	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet.flights@cm	Mon Aug 30 2021	csso_trialuser43	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet_cdpw1.fl...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpw1.flights...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet_cdpw2.fl...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpw2.flights...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet_cdpw3.fl...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpw3.flights...	Tue Sep 07 2021	trial43_admin	hive

4) Click “Hive Table” under Filters on the left

CLUDERA
Data Catalog

Search

Datasets

Bookmarks

Profilers

Atlas Tags

Get Started

Data Catalog / Search

flights

Atlas Ranger

Setup the Profiler for cdptrialuser43-datalake

Profiler runs data profiling operations as a pipeline on data located in the data lake which helps to view the profiled results for assets. Setting up a Profiler launches a cluster with various services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. [Get Started >](#)

cdptrialuser43-datalake | 8

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	flights	airlines_new_orc.flights@cm	Mon Aug 30 2021	csso_trialuser43	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet.flights@cm	Mon Aug 30 2021	csso_trialuser43	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet_cdpww1.flights...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpww1.flights@dw...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet_cdpww2.flights...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpww2.flights@dw...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet_cdpww3.flights...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpww3.flights@dw...	Tue Sep 07 2021	trial43_admin	hive

*Find what database(s) the table “flights” is located.

CLUDERA
Data Catalog

Search

Datasets

Bookmarks

Profilers

Atlas Tags

Data Catalog / Asset Details

flights

Atlas

Properties

Type: HIVE TABLE

of Columns: 29

Data Lake: cdptrialuser43-datalake

Datasets: 0

Owner: trial43_admin

Created On: Tue Sep 07 2021 11:12:13 GMT-0400 (Eastern Daylight Time)

Last Access Time: Tue Sep 07 2021 11:12:13 GMT-0400 (Eastern Daylight Time)

Table Type: EXTERNAL TABLE

Database: **airlines_new_parquet_cdpww1**

De Catalog: dwx-warehouse-1630464792-30cm

Parent: airlines_new_parquet_cdpww1

Qualified Name

airlines_new_parquet_cdpww1.fl ...

Comment

+ Add Comment

Description

+ Add Description

Classifications

Managed System Propagated

+ Add Classification

Terms

+ Add Terms

5) Click “flights” table for your breakout CDW

CLUDERA
Data Catalog

Search

Datasets

Bookmarks

Profilers

Atlas Tags

Get Started

Data Catalog / Search

flights

Atlas Ranger

Setup the Profiler for cdptrialuser43-datalake

Profiler runs data profiling operations as a pipeline on data located in the data lake which helps to view the profiled results for assets. Setting up a Profiler launches a cluster with various services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. [Get Started](#)

cdptrialuser43-datalake | 8

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	flights	airlines_new_orc.flights@cm	Mon Aug 30 2021	csso_trialuser43	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet.flights@cm	Mon Aug 30 2021	csso_trialuser43	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_parquet_cdpvw1.flights...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpvw1.flights@dw...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpvw2.flights@dw...	Tue Sep 07 2021	trial43_admin	hive
<input type="checkbox"/> Hive Table	flights	airlines_new_orc_cdpvw3.flights@dw...	Tue Sep 07 2021	trial43_admin	hive

6) Zoom into the Lineage and scroll over one of the /cdp-lake/data, clicking the “i” for more information

CLUDERA
Data Catalog

Search

Datasets

Bookmarks

Profilers

Atlas Tags

Get Started

Help

2.0.13-b78

Data Catalog / Asset Details

Table type: MANAGED_TABLE
Database: airlines_new_parquet
DB Catalog: cm
Parent: airlines_new_parquet

+ Add Description

Classifications: Managed System Propagated

+ Add Classification

Terms: + Add Terms

Overview Schema Metadata Audits Policy Access Audits

Lineage

Filter By: Depth: 3 Process Node: Hide

→ Lineage → Impact → Replication ○ Current Entity

*Find out at least one year that the “flights” table was generated from.

*Find out how many columns the table “flights” contains.

Part 2 - Create a Virtual Warehouse and Run Queries [45 minutes]

Overview: What is Cloudera Data Warehouse?

We will explore features of Cloudera Data Warehouse (CDW) by performing some data exploration and create dashboards to share our results to a wider audience

We will be taking a look at a generated data set from a mock airline company containing flights information from its fleet of aircraft.

A virtual warehouse represents virtual compute resources to access data that is stored in a database catalog. This lets you create or destroy compute resources, auto-scale, or separate resources across different workloads, all running on the same underlying data.

CDW let's you choose from a set of default resources based on your predicted workload as well as give you fine grained control over autoscaling and timeout features so you can fine tune your system to be most cost effective.

Purpose: Create a virtual warehouse and run queries, answering the questions below:

- What are the top 5 visited destinations by year from (1995-2008)?
- What are the top 10 routes (origin and dest) that have seen maximum diversions?
- Which three months have seen the most number of cancellation due to bad weather?

1) Open CDP, using the “admin” user within the Test Drive link.

Your link should look something like (remember click the link in your email not the link below)

http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X

*xx represents the trial user #

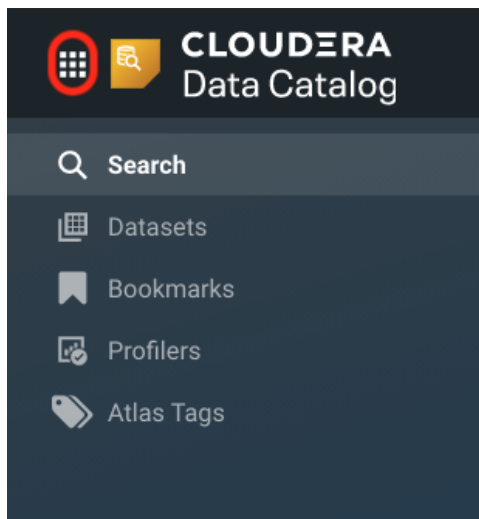
*X represents the password

2) Click the “Data Warehouse” within the CDP Home Screen



How do you get to the CDP Home Screen?

- From any experience such as “Data Catalog”, click the 9 square at the top left and then click “Home”



WE HAVE DONE THIS FOR YOU – DO NOT CREATE A NEW VIRUTAL WAREHOUSE – READ THROUGH THIS FOR BACKGROUND INFO ONLY...

3) DO NOT Click the “+” at the top right next to “Virtual Warehouses”

The screenshot shows the 'Virtual Warehouses' management interface. At the top, there's a header 'Virtual Warehouses | 1' with a search icon and a '+' button circled in red. Below this is a 'New Virtual Warehouse' form with the following fields:

- Name ***: A text input field with placeholder text 'Enter Virtual Warehouse Name'.
- Type ***: Two buttons, 'HIVE' (selected) and 'IMPALA'.
- Database Catalog ***: A dropdown menu with 'cdptrialuser24-dl-default' selected.
- Size ***: A dropdown menu with '-- select an option --'.

Below the form is a table listing existing virtual warehouses:

NAME	STATUS	COMPUTE	CATALOG
default-vw	Stopped	compute-1611103491-4hbp	cdptrialuser24-dl-default

At the bottom, there's a summary table for the 'default-vw' warehouse:

NODE COUNT	TOTAL CORES	TOTAL MEMORY	TYPE
0	12	56 GB	HIVE COMPACTOR

4) DO NOT Enter a name for your New Virtual Warehouse

Virtual Warehouses | 1

New Virtual Warehouse X

Name *

testvirtualwarehouse1

Type *

HIVE IMPALA

Database Catalog *

cdptrialuser24-dl-default

Size *

-- select an option --

5)) DO NOT Select the Size of “xsmall - 2 Executor Nodes”

*How do I choose a size? Initial concurrent users

Virtual Warehouses | 1

New Virtual Warehouse X

Name *

testvirtualwarehouse1

Type *

HIVE IMPALA

Database Catalog *

cdptrialuser24-dl-default

Size *

-- select an option --

- ✓ xsmall - 2 Executor Nodes
- small - 10 Executor Nodes
- medium - 20 Executor Nodes
- large - 40 Executor Nodes
- custom

6)) To save money you stop the instances you aren’t using. Cloudera lets you define if you spin down to zero, if you have some Kubernetes pods running all the time, and how long these live when there is no workload. DO NOT Set the AutoSuspend Timeout (in seconds) between 4500 and 5500:

*What is AutoSuspend Timeout? Automatically spin-down unused resources after timeout occurs.

Virtual Warehouses | 1

New Virtual Warehouse X

Name *

testvirtualwarehouse1

Type *

HIVE IMPALA

Database Catalog *

cdptrialuser24-dl-default

Size *

xsmall - 2 Executor Nodes

AutoSuspend Timeout (in seconds): 5000

0 1000 2000 3000 4000 5000 6000 7000

7) DO NOT Choose "Install Data Visualization" to be on
*Allowing for Data Visualizations in Part 3

New Virtual Warehouse X

Name *

Type *

HIVE IMPALA

Database Catalog *

▼

Size *

▼

AutoSuspend Timeout (in seconds): 5000

01000200030004000500060007000

Concurrency Autoscaling ⓘ

Nodes: Min:2, Max:6

2468101214161820

HEADROOM WAIT TIME

Desired Free Capacity: 1

12

☐ Query Isolation ⓘ

☒ Install Data Visualization ⓘ




CREATE

8) DO NOT , DO NOT , DO NOT, REALLY DO NOT: Click “Create” to create your Virtual Warehouse

*Allow for approximately 5 minutes for your Virtual Warehouse to become available for use

CREATE

When available for use, “Starting” will change to “Running” as shown below





**testvirtualwarehouse1**

compute-1611179792-vz49
cdptrialuser24-dl-default

Starting

checking if query-coordinator-0 statefulset is ready with at least 1 ready replica(s) (config-id: 7647c82f-8b37-4593-80e9-058f1f928b31 version: 7.2.8.0-24)

NODE COUNT	TOTAL CORES	TOTAL MEMORY	TYPE
2	38	292 GB	HIVE DATA VISUALIZATION

**testvirtualwarehouse1**

compute-1611179792-vz49
cdptrialuser24-dl-default

Running

NODE COUNT	TOTAL CORES	TOTAL MEMORY	TYPE
2	38	292 GB	HIVE DATA VISUALIZATION

If you can read this it is the end of the background information and it is

TIME FOR YOU TO JUMP BACK IN AND DO THE LAB. Please, Please, not drop any tables. Do not alter any tables. This is a shared environment. You all have the same userid.

9) Notice there multiple Virtual Warehouses (VWs). You will be working on one of the VWs. If you are in Zoom Breakout Room 1 use cdpvw1, room2 uses cdpvw2 etc

Virtual Warehouses | 3

cdpvw1
impala-1631025591-2nj
Stopped cdpvw1

[Data VIZ](#) [HUE](#) [▶](#) [⋮](#)

0 20 137 GB IMPALA

cdpvw3
impala-1630465237-9d4f
Stopped cdpvw3

[Data VIZ](#) [HUE](#) [▶](#) [⋮](#)

0 20 137 GB IMPALA

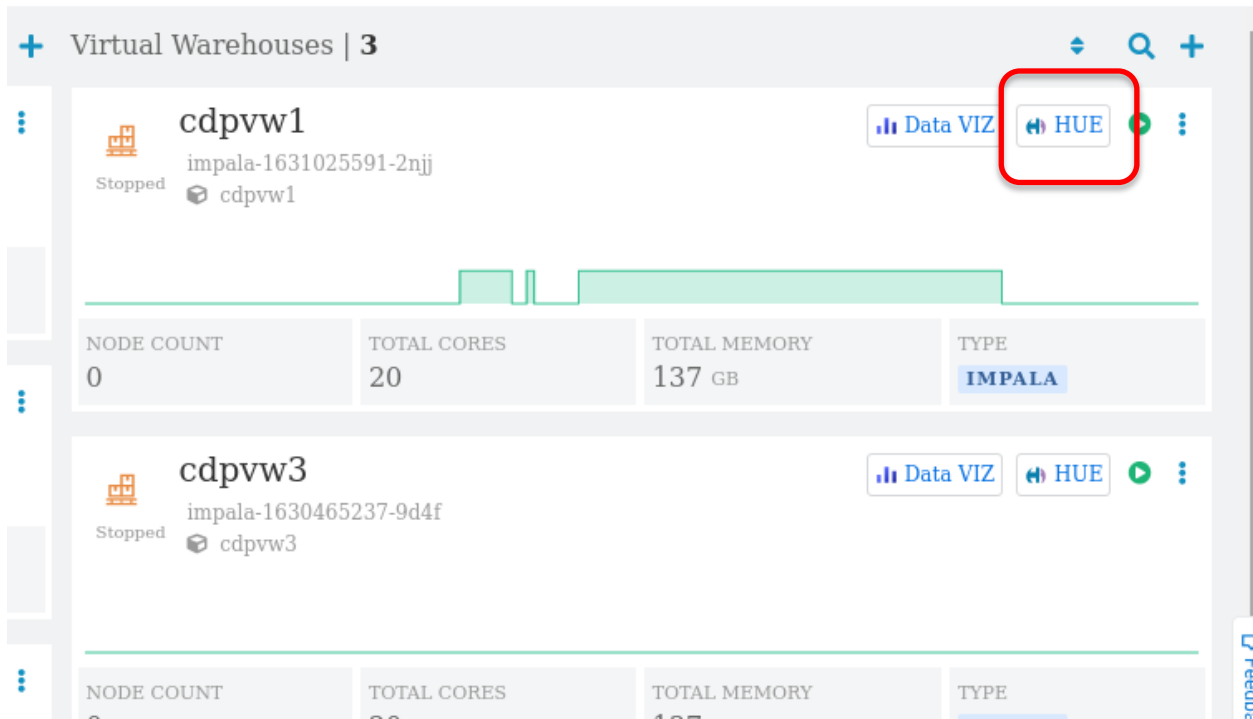
cdpvw2
impala-1630465160-mt6q
Stopped cdpvw2

[Data VIZ](#) [HUE](#) [▶](#) [⋮](#)

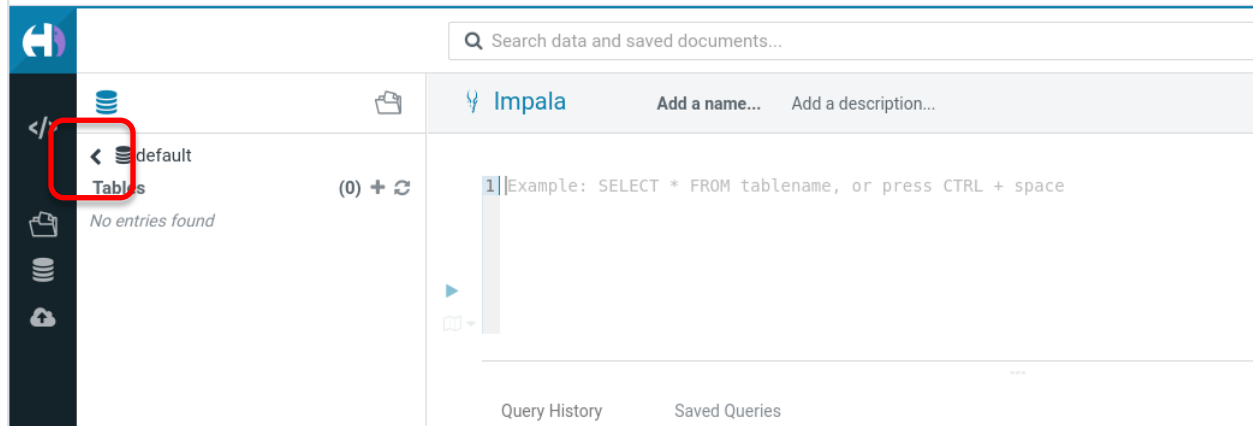
0 20 137 GB IMPALA

[Feedback](#)

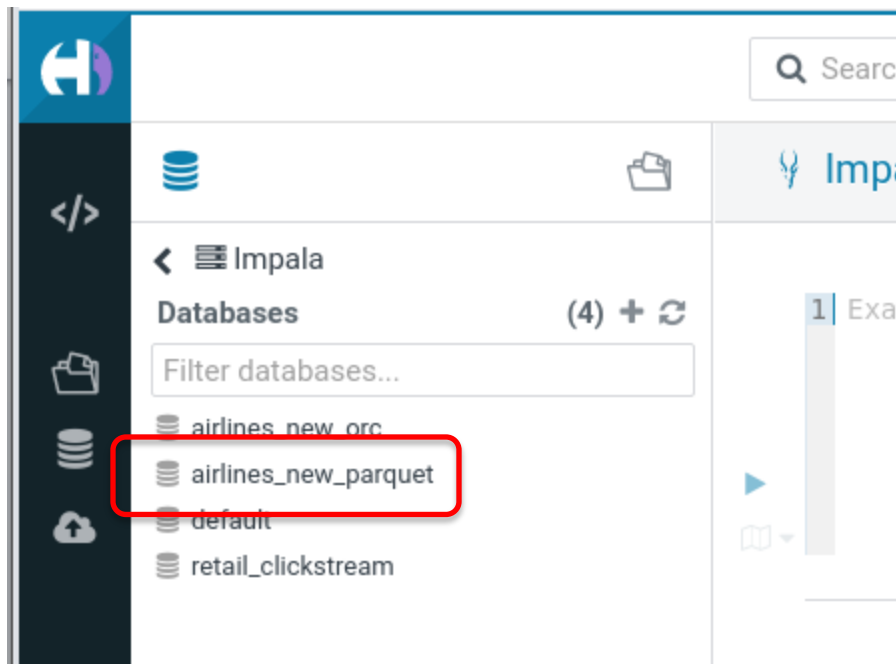
10) Click on HUE to enter the “Hadoop User Experience” in your designated room.



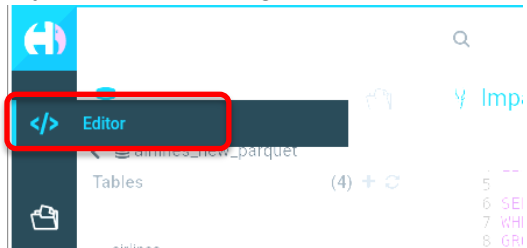
11) The landing page takes you to the “default” database. Click on the **<** to the left of the default database to select a different database



12) Click on the database “airlines_new_parquet” that we saw in Part 1 “Data Catalog.” Both Impala and Hive work with both Parquet and ORC files. As a rule of thumb if you’re mostly using Impala use Parquet format or Kudu. When working with Hive ORC is the preferred format.



If you ever need to get back to this screen layout click on the “editor” shown here:



13) Enter the following query, answering the question “show me the top 5 visited destination by year from (1995-2008)” Click on the blue triangle to run the query.

```
SELECT dest,year,COUNT(dest) as Times_Visited FROM flights
GROUP BY dest,year
ORDER BY Times_Visited DESC
LIMIT 5;
```

Why doesn't it run right away? You promised it was fast 😊

Click on blue triangle to run

What is “waiting for query executors to start?”

Impala went to sleep to save money on cloud costs. You can see it wake back up to answer your query. You learned about this in the configuration section where you did not provision a virtual warehouse.

You may get the query result without having to wait for Impala to “wake up.” You get to decide how long an idle time you want to wait before you scale down to zero, or if you have the budget you can have Impala always available.

14) Click “EXPLAIN” to see the explain plan prior to running the query

*Not required to execute the query - this gives us a plan on exactly what the query is doing

The screenshot shows the Impala web interface. On the left, a sidebar displays the database structure under 'airlines_new_parquet', listing tables: airlines, airports, flights, and planes. The main panel shows a SQL query: `1 SELECT dest, year, COUNT(dest) as Times_Visited
2 GROUP BY dest, year
3 ORDER BY Times_Visited DESC
4 LIMIT 5;
5`. A dropdown menu is open below the query, with options: Explain (highlighted with a red box), Get shareable link, Format, and Clear. A yellow callout box points to the dropdown arrow with the text: 'Click on this down-arrow to expose the menu'.

The screenshot shows the 'Explain' tab in the Impala interface. It displays resource estimates and warnings. A red box highlights the following warnings:

```
Max Per-Host Resource Reservation: Memory=508.00MB Threads=15
Per-Host Resource Estimates: Memory=2.43GB
Dedicated Coordinator Resource Estimate: Memory=104MB
WARNING: The following tables have potentially corrupt table statistics.
Drop and re-compute statistics to resolve this problem.
airlines_new_parquet.flights
WARNING: The following tables are missing relevant table and/or column statistics.
airlines_new_parquet.flights
```

Below the warnings, the execution plan is shown:

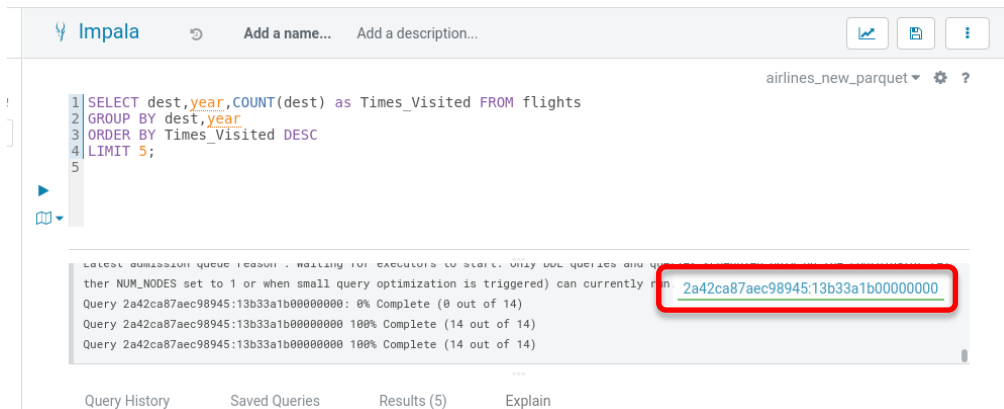
```
PLAN-ROOT SINK
|
05:MERGING-EXCHANGE [UNPARTITIONED]
| order by: count(dest) DESC
| limit: 5
|
02:TOP-N [LIMIT=5]
| order by: count(dest) DESC
| row-size=24B cardinality=5
```

The “Explain Plan” shows you how the query will execute. It is asking you to update statistics for the tables in the query. This is a good idea for performance. After everyone reaches this point in the lab designate one person to run: **`compute stats airlines_new_parquet.flights;`**

We are in a shared environment. If someone else has done “compute stats” you won’t see the message. In STEP 4, after the DataViz lab, you will have a chance to create your own tables and work on compute stats with your unique tables.

We can discuss optimizer paths and table statistic in more detail as part of the breakout room.

15) After you run the query you will have a link to lots of information about the query. Click on the link shown below



The screenshot shows the Impala web interface. At the top, there's a header with the Impala logo, a refresh button, and fields for "Add a name..." and "Add a description...". Below the header, there's a toolbar with icons for chart, table, and help. The main area displays a SQL query:

```
1 SELECT dest, year, COUNT(dest) as Times_Visited FROM flights
2 GROUP BY dest, year
3 ORDER BY Times_Visited DESC
4 LIMIT 5;
5
```

Below the query, there's a status bar showing the query's progress. A red box highlights a link: [2a42ca87aec98945:13b33a1b00000000](#). Below the link, there's a table showing the query's progress:

Query ID	Progress
Query 2a42ca87aec98945:13b33a1b00000000	0% Complete (0 out of 14)
Query 2a42ca87aec98945:13b33a1b00000000	100% Complete (14 out of 14)
Query 2a42ca87aec98945:13b33a1b00000000	100% Complete (14 out of 14)

At the bottom, there's a navigation bar with tabs: Query History, Saved Queries, Results (5), and Explain.

16) Explore the different tabs in the query pop-up. The Visual Plan shows how the query was executed. These get more interesting with multi table joins. The Summary show how much time was spent in each stage of the query, the memory used, and the rows produced. The Profile includes the summary and great detail of everything that happened on each node running the query.

SELECT dest,year,COUNT(dest) as Times_Visited FROM flights G...

ID

2a42ca87aec98945:13b33a1b000000000

Plan

Query

Text Plan

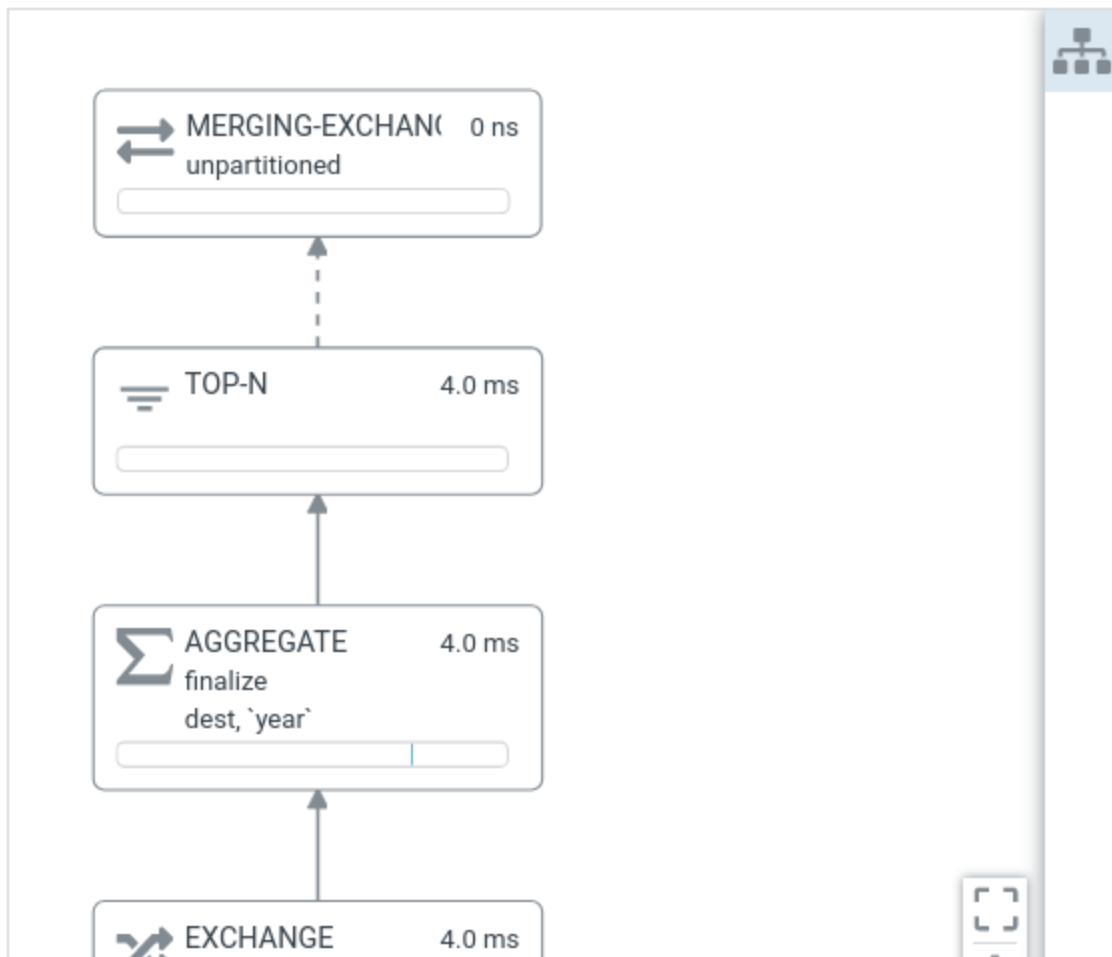
Summary

Profile

Memory

Backends

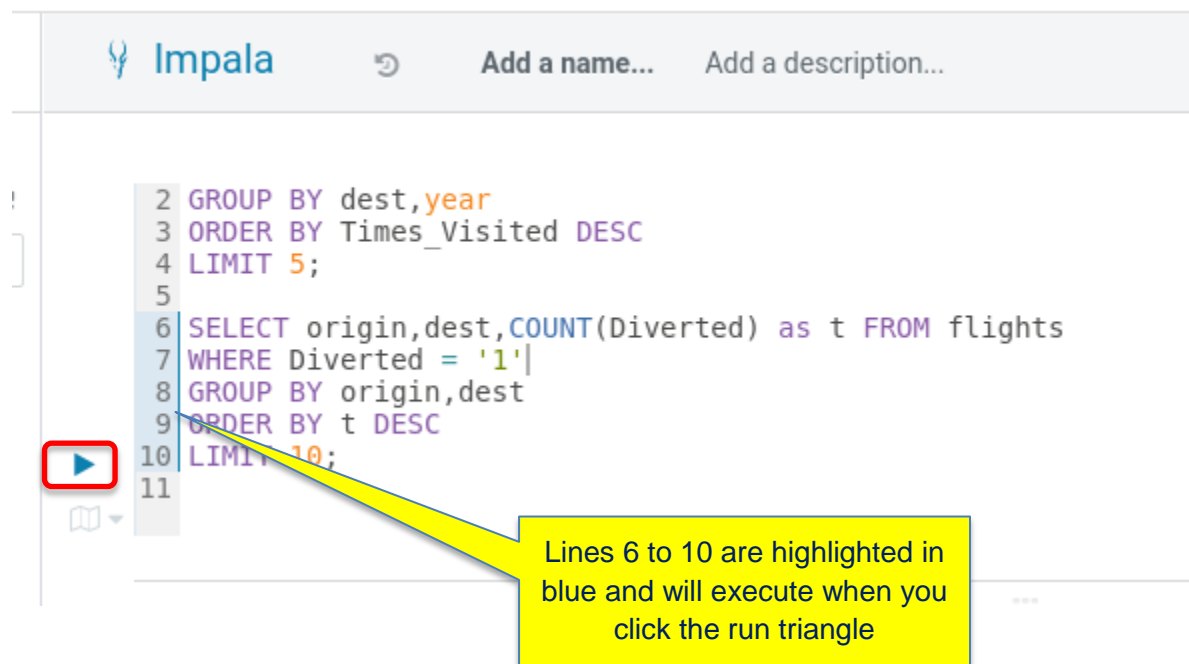
Instances



17) Add another query to the editor

```
SELECT origin,dest,COUNT(Diverted) as t FROM flights
WHERE Diverted = "1"
GROUP BY origin,dest
ORDER BY t DESC
LIMIT 10;
```

18) Notice the highlighting on the left edge. HUE is parsing based on the semi-colon and the execution arrow will run whatever is highlighted in blue, or whatever has been highlighted by the cursor. This way you can have multiple queries in the same canvas.



19) Click the blue triangle to execute the query, answering the question “What are the top 10 routes (origin and dest) that have seen maximum diversions?”

	origin	dest	t
1	ORD	LGA	845
2	LGA	DFW	749
3	DFW	LGA	653
4	DAL	HOU	615
5	ATL	LGA	567
6	MDW	STL	512
7	ATL	DFW	482
8	ORD	DFW	450

20) Hover over the disappearing **i** next to the airports table to see more information about the table. We're going to use the geo location of the airports to do a marker map in HUE.

Column (7)	Type	Description	Sample
iata	string	00M	00R
airport	string	Thigpen	Livingston Mun
city	string	Bay Springs	Livingston
state	double	NULL	NULL
country	string	USA	USA
lat	double	31.95376472	30.68586111
lon	double	-89.23450472	-95.01792778

You can also click on the table name to expand all the columns. On the far right of the browser is help tooling.

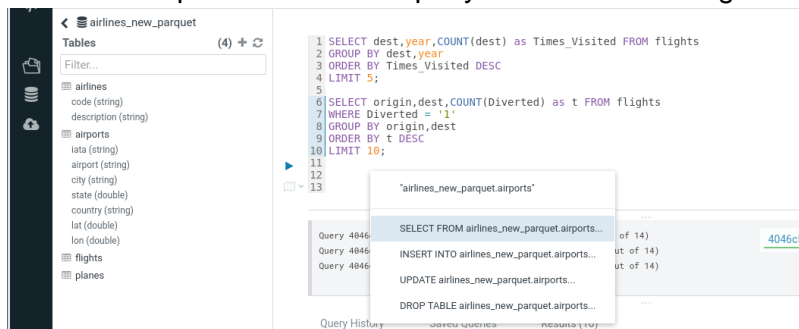
21) Try out the “drag and drop” option, drag and drop the table name “airports over to line 12

```

1 SELECT dest,year,COUNT(dest) as Times_Visited FI
2 GROUP BY dest,year
3 ORDER BY Times_Visited DESC
4 LIMIT 5;
5
6 SELECT origin,dest,COUNT(Diverted) as t FROM fl:
7 WHERE Diverted = '1'
8 GROUP BY origin,dest
9 ORDER BY t DESC
10 LIMIT 10;
11
12

```

When you drop the table name you'll get to choose what SQL you want auto generated. Take the "select" option and run the query with the blue-triangle



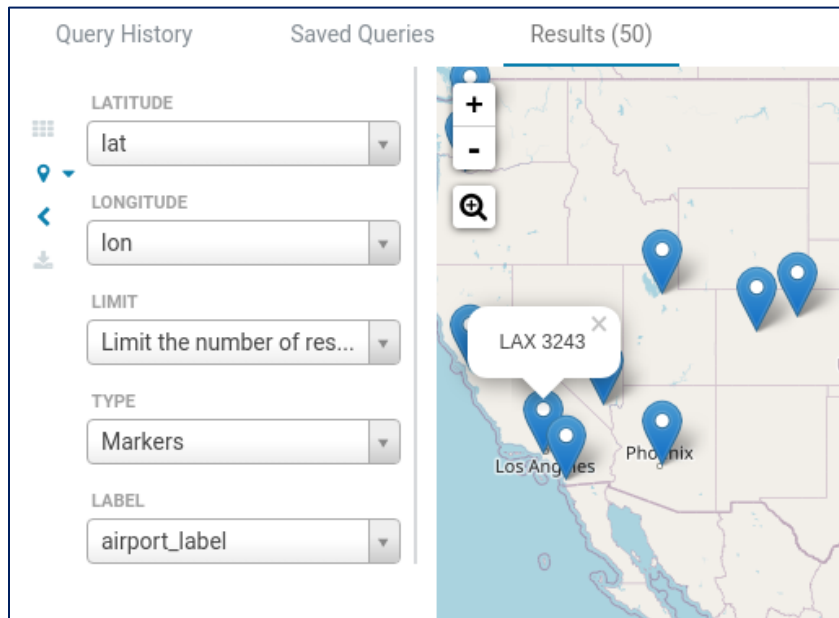
22) Let's now build a marker map of the airports with the most cancellations. This will correlate with the airports that have the most flights. Run the SQL shown below

```
SELECT origin, lat, lon, COUNT(Cancelled) as num_of_cancellations ,
concat(origin, " ", cast(count(cancelled) as string)) as airport_label
FROM flights, airports
WHERE origin = iata and cancelled = 1 AND cancellationcode = 'B'
GROUP BY origin, lat, lon order by num_of_cancellations desc limit 50;
```

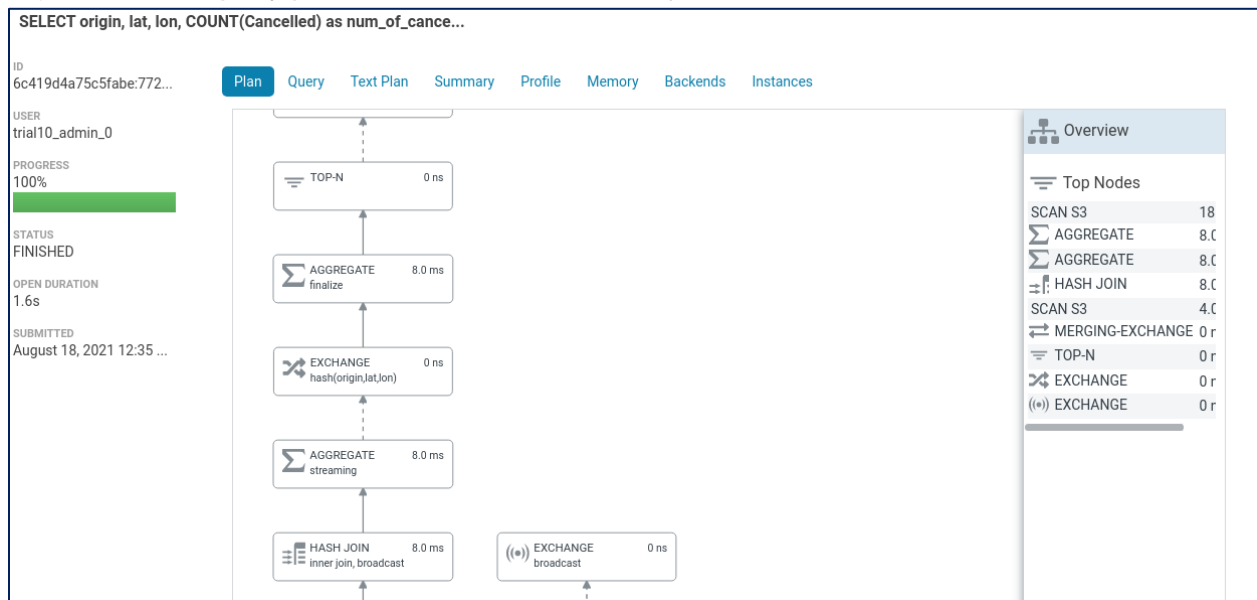
23) After you run the SQL use the down arrow to choose the type of output formatting. You've been using the data grid, we're now going to choose the marker-map

	origin	lat	lon	num_of_cancellations	airport_label
1	ORD	41.979595	-87.90446417	22123	ORD 22123
		32.89595056	-97.0372	16519	DFW 16519
		33.64044444	-84.42694444	15350	ATL 15350
		40.77724306	-73.87260917	10688	LGA 10688
		40.69249722	-74.16866056	8979	EWR 8979
		29.98047222	-95.33972222	8885	IAH 8885
		42.3643475	-71.00517917	7978	BOS 7978
8	CVG	39.04614278	-84.6621725	6980	CVG 6980

24) Configure the marker map per shown below. Clicking on one of the markers will pop up the value of the “airport_label” column



25) Look at the query plan – notice we now have a join in the tree



26) The summary shows details of all the stages in the join and their metrics

SELECT origin, lat, lon, COUNT(Cancelled) as num_of_cance...

ID
6c419d4a75c5fabe:772...

Plan Query Text Plan **Summary** Profile Memory Backends Instances

USER
trial10_admin_0

PROGRESS
100%

STATUS
FINISHED

OPEN DURATION
1.6s

SUBMITTED
August 18, 2021 12:35 A...

Operator	#Hosts	#Inst	Avg Time	Max Time	#Rows	Est. #Rows	Peak Mem	Est. Peak Mem	Detail
F03:ROOT	1	1	0.000ns	0.000ns			4.02 MB	4.00 MB	
08:MERCING-EXCHANGE	1	1	0.000ns	0.000ns	50	50	224.00 KB	28.22 KB	UNPARTITIONED
F02:EXCHANGE SENDER	2	14	0.000ns	0.000ns			3.45 KB	0	
04:TOP-N	2	14	0.000ns	0.000ns	318	50	12.00 KB	1.76 KB	
07:AGGREGATE	2	14	3.999ms	8.000ms	318	4.37M	34.05 MB	128.00 MB	FINALIZE
06:EXCHANGE	2	14	0.000ns	0.000ns	1.69K	4.37M	56.00 KB	10.55 MB	HASH(origin,lat,lon)
F00:EXCHANGE SENDER	2	14	0.000ns	0.000ns			104.34 KB	0	
03:AGGREGATE	2	14	1.714ms	8.000ms	1.69K	4.37M	42.01 MB	128.00 MB	STREAMING
02:HASH JOIN	2	14	857.140us	7.999ms	267.00K	4.37M	7.99 MB	0	INNER JOIN, BROADCAST
--F04:JOIN BUILD	2	2	3.999ms	4.000ms			23.27 MB	23.25 MB	
05:EXCHANGE	2	2	0.000ns	0.000ns	3.38K	10.96K	240.00 KB	331.71 KB	BROADCAST
F01:EXCHANGE SENDER	1	1	0.000ns	0.000ns			8.81 KB	0	
01:SCAN S3	1	1	3.999ms	3.999ms	3.38K	10.96K	427.30 KB	16.00 MB	airlines_new_parquet.airports
00:SCAN S3	2	14	61.714ms	180.000ms	267.00K	4.37M	16.21 MB	88.00 MB	airlines_new_parquet.flights

27) Time to save our work. Give your Impala SQL a name. For the lab use yourNameXXX where XXX is where you get to be creative. Use something unique, this is a multi user environment.

Search data and saved documents...

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

```

5
6 SELECT origin,dest,COUNT(Diverted) as t FROM flights
7 WHERE Diverted = '1'

```

Then click "Save" and then "Save" in the popup

Impala martyImpalaQueries123 Add a description...

0.55s airlines_new_parquet

```

5
6 SELECT origin,dest,COUNT(Diverted) as t FROM flights

```


Save query as...

Name

Description

Your saved queries will show up under the “Saved Queries” heading.

```
20 WHERE origin = iata and cancelled = 1 AND cancellationcode = 'B'
21 GROUP BY origin, lat, lon order by num_of_cancellations desc limit 50;
22;
```

Query 6c419d4a75c5fabe:772554f000000000 100% Complete (15 out of 15)

Query 6c419d4a75c5fabe:772554f000000000 100% Complete (15 out of 15)

6c419d4a75c5fabe:772554f000000000

Query History **Saved Queries** Results (50)

Name	Description	Owner	Last Modified
martyImpalaQueries123		trial10_admin_0	08/17/2021 1:55 PM -04:00

WELCOME TO THE DATA VISUALIZATION LAB

Part 3 - Data Visualization [25 minutes]

Overview: What is Data Visualization and how do we use it with our data?

Purpose: Create visualization using the flight information answering the question (visually with a density graph):

- What were the most number of flights from destination to origin between (1995-2008) - Route Density

1) Open CDP, using the “admin” user within the Test Drive link.

Your link should look something like (remember click the link in your email not the link below)

http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X

*xx represents the trial user #

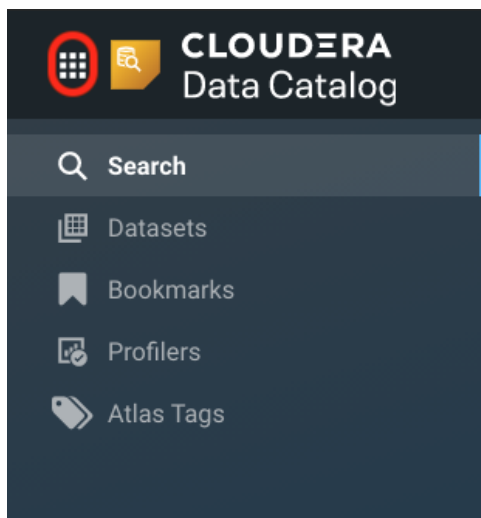
*X represents the password

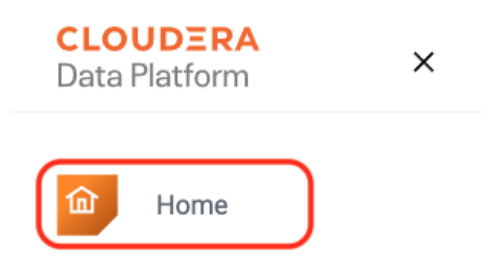
2) Click the “Data Warehouse” within the CDP Home Screen



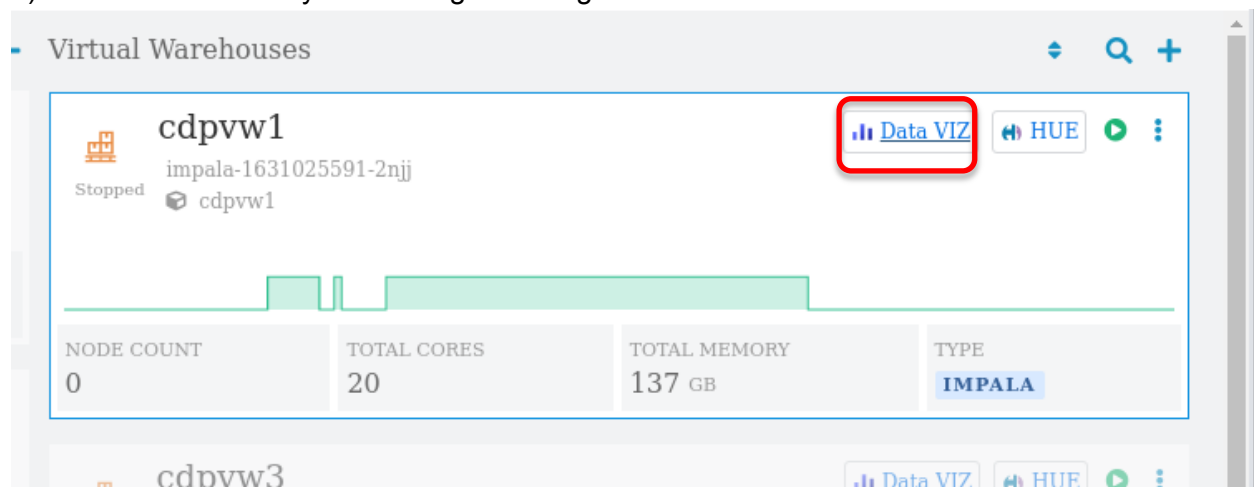
How do you get to the CDP Home Screen?

- From any experience such as “Data Catalog”, click the 9 square at the top left and then click “Home”



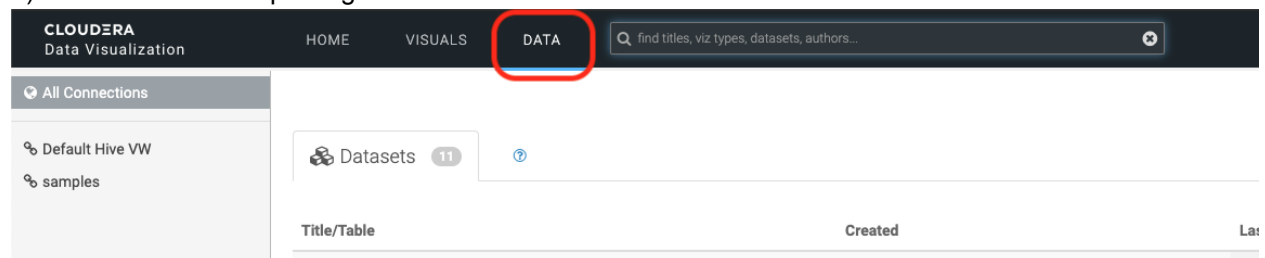


3) Click “Data VIZ” on your existing “Running” Virtual Warehouse.

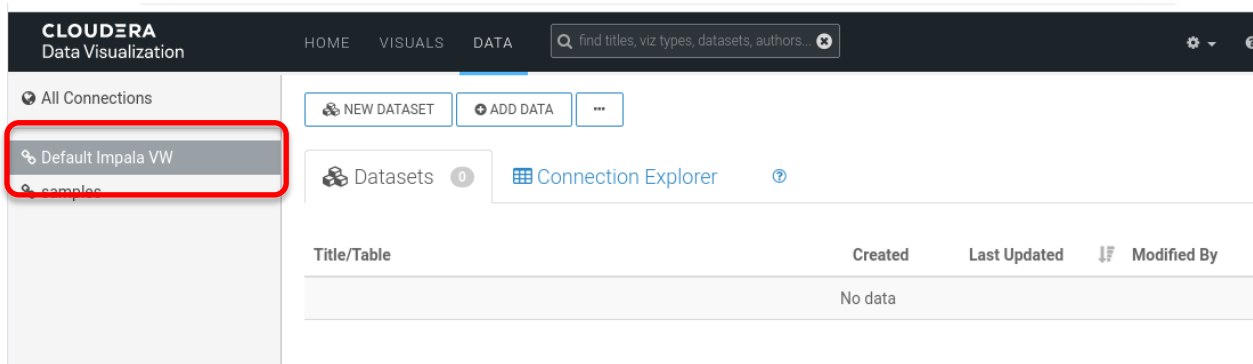


4) You are autologged into Data VIZ

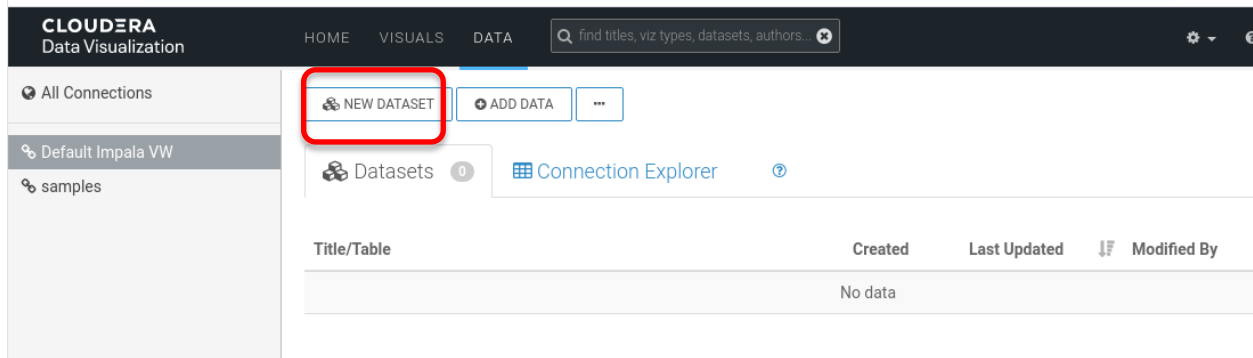
5) Click “DATA” the top navigation bar



6) Click “Default Impala VW” to add our dataset.



7) Click “NEW DATASET” to add our “flights” data



8) Enter a name for the Dataset title naming “mynamehere_airline_new_parquet_flights”
Recall Hive prefers ORC and Impala prefers Parquet. Our development team is making both engines like both file formats equally.

New Dataset

Create a dataset from data on this connection. You need to create a dataset before you can create dashboards or apps.

Dataset title *

martyflightsnewparquet

Dataset Source

From Table

Select Database

airlines_new_parquet_cdpw1

Select Table

flights

CANCEL

CREATE

9) Choose the database “airlines_new_parquet_YOURVIRTUALWAREHOUSENUMBER”

New Dataset

Create a dataset from data on this connection. You need to create a dataset before you can create dashboards or apps.

Dataset title *

martyflightsnewparquet

Dataset Source

From Table ▼

Select Database

airlines_new_parquet_cdpvw1 ▼

Select Table

flights ▼

CANCEL

CREATE

10) Choose the table “flights”

*Need to import multiple databases and tables? You’d use Dataset Source = SQL

New Dataset

Create a dataset from data on this connection. You need to create a dataset before you can create dashboards or apps.

Dataset title *

Dataset Source

From Table ▼

Select Database

airlines_new_parquet_cdpw1 ▼

Select Table

flights ▼

CANCEL

CREATE

11) Click "CREATE"

New Dataset

Create a dataset from data on this connection. You need to create a dataset before you can create dashboards or apps.

Dataset title *

Dataset Source

From Table ▼

Select Database

airlines_new_parquet_cdpw1 ▼

Select Table

flights ▼

CANCEL

CREATE

12) Click “+” to create a New Dashboard

The screenshot shows the Cloudera Data Visualization interface. The top navigation bar includes 'HOME', 'VISUALS', and 'DATA'. A search bar is present. The left sidebar shows 'All Connections' and 'Default Impala VW'. The main area displays 'NEW DATASET' and 'ADD DATA' buttons. Below these, there's a 'Datasets' section with a table listing datasets. The table has columns: 'Title/Table', 'Created', and 'Last Updated'. The first row is 'martyflightsnewparquet' with a red box highlighting a '+' icon next to it. The second row is 'airlines_new_parquet_cdpv1.flights'.

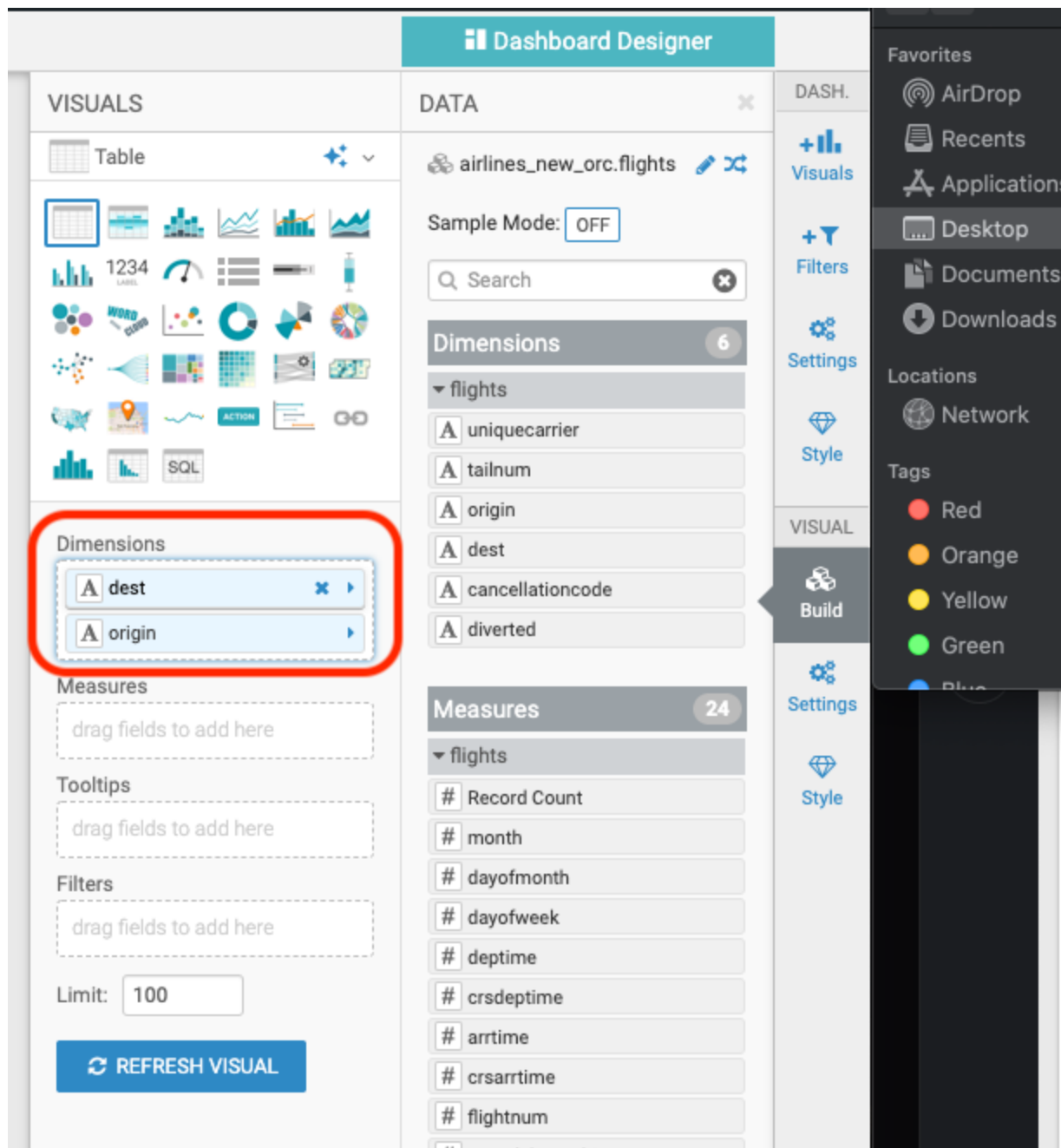
Title/Table	Created	Last Updated
martyflightsnewparquet	Sep 09, 2021	a few seconds ago
airlines_new_parquet_cdpv1.flights		

13) Choose “TreeMap” under “VISUALS” or explore with the Wizard

The screenshot shows the Cloudera Dashboard Designer interface. The 'VISUALS' panel is active, displaying various visualization options. A yellow callout box points to the 'TreeMap' visualization. The callout box contains the text: 'Explore many visualizations with this wizard this is much more fun and educational'. The 'DATA' panel shows the 'martyflightsnewparquet' dataset. The 'DASH.' panel shows the 'TreeMap' visualization.

Explore many visualizations with this wizard this is much more fun and educational

14) Drag-and-drop both “dest” and origin” from Dimensions->Flights into Dimensions under Visuals



15) Drag-and-drop "Record Count" from Measures->Flights into Measures under Visuals

Dashboard Designer

VISUALS

Table

1234

Dimensions

dest

origin

Measures **sum(1)**

Record Count

Tooltips

drag fields to add here

Filters

drag fields to add here

Limit: 100

REFRESH VISUAL

DATA

airlines_new_orc.flights

Sample Mode: OFF

Search

Dimensions 6

flights

uniquecarrier

tailnum

origin

dest

cancellationcode

diverted

Measures 24

flights

Record Count

month

dayofmonth

dayofweek

deptime

crsdeptime

arrtime

crsarrrtime

flightnum

actualelapsedtime

crselapsedtime

airtime

arrdelay

DASH.

Visuals

Filters

Settings

Style

VISUAL

Build

Settings

Style

Favorites

AirDro

Recen

Applic

Deskt

Docur

Downl

Locations

Netwo

Tags

Red

Orang

Yellow

Green

Blue

16) Click the right arrow next to Record Count and select “Descending” under Order and Top K

*Notice - you can have other Visuals chosen to be displayed with the Dimensions and Measure(s), then click REFRESH VISUALS

Dashboard Designer

VISUALS

Treemap

1234 LABEL

WORD CLOUD

ACTION

SQL

Dimensions

dest

origin

Measure

Record Count

Tooltips

drag fields to add here

X Trellis

drag fields to add here

Y Trellis

drag fields to add here

Filters

drag fields to add here

REFRESH VISUAL

DATA

airlines_new_orc.flights

Sample Mode: OFF

Search

Dimensions

6

flights

uniquecarrier

tailnum

origin

dest

cancellationcode

diverted

Measures

24

flights

Record Count

month

dayofmonth

dayofweek

deptime

crsdeptime

arrtime

crsarrrtime

flightnum

actualelapsedtime

crselapsedtime

airtime

arrdelay

DASH.

Visuals

Filters

Settings

Style

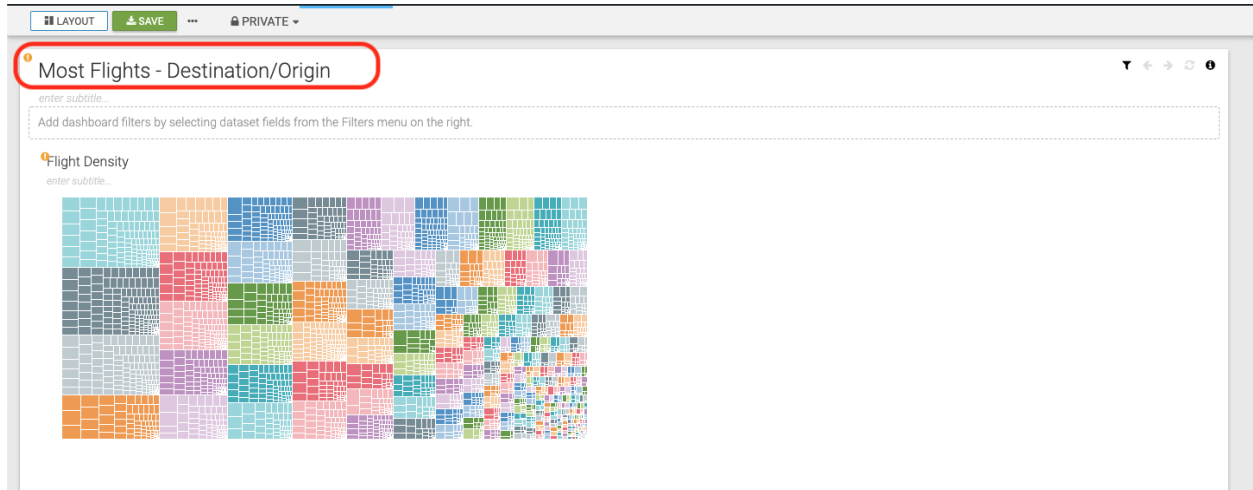
VISUAL

Build

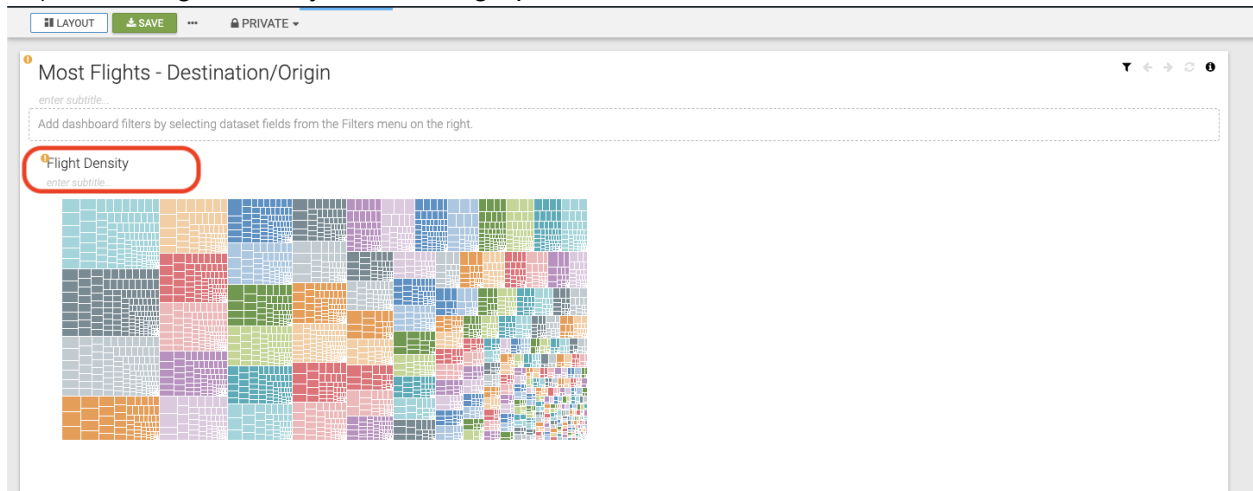
Settings

Style

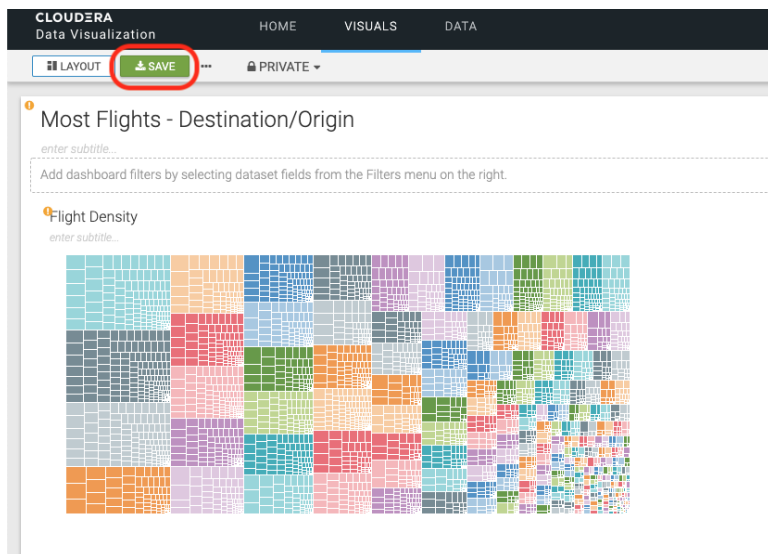
18) Enter a title “Most Flights - Destination/Origin”



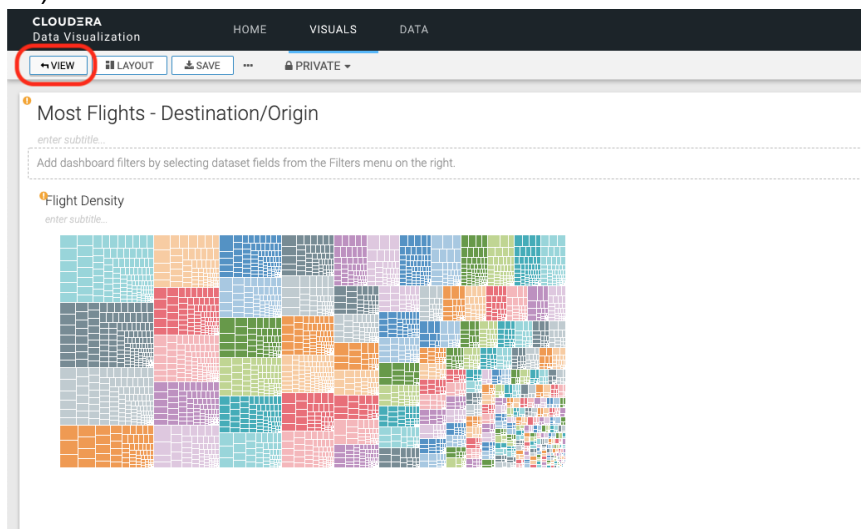
19) Enter “Flight Density” under the graph’s title



20) Click “SAVE”



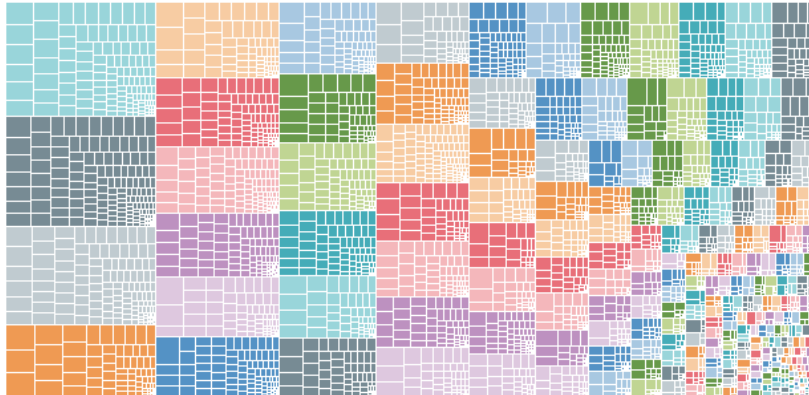
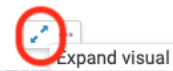
21) Click “VIEW”



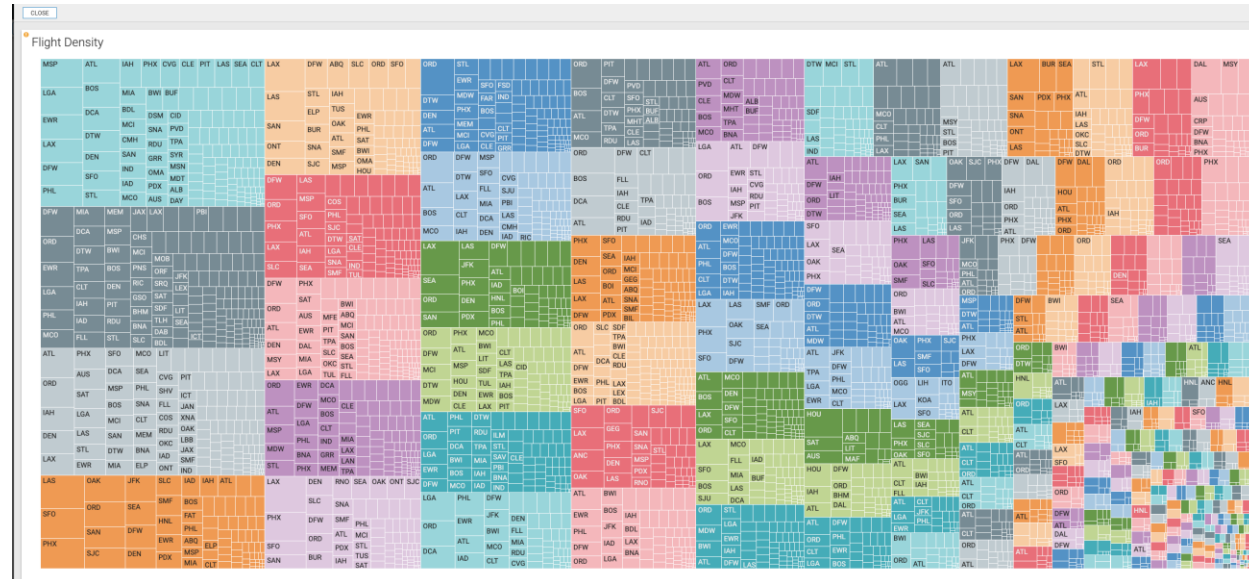
22) Scroll over the graph and click “Expand Visual”

Most Flights - Destination/Origin

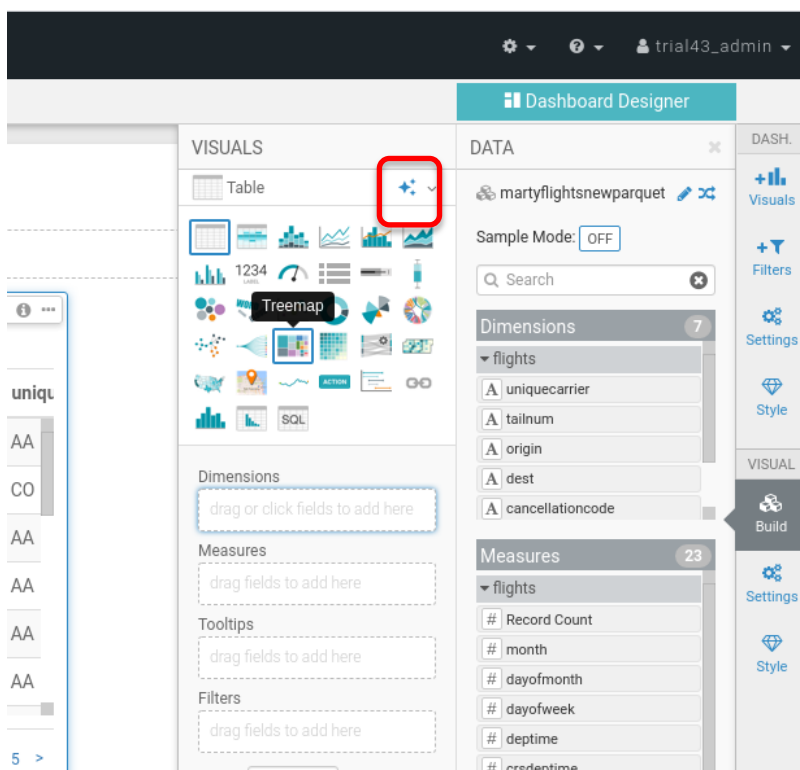
Flight Density



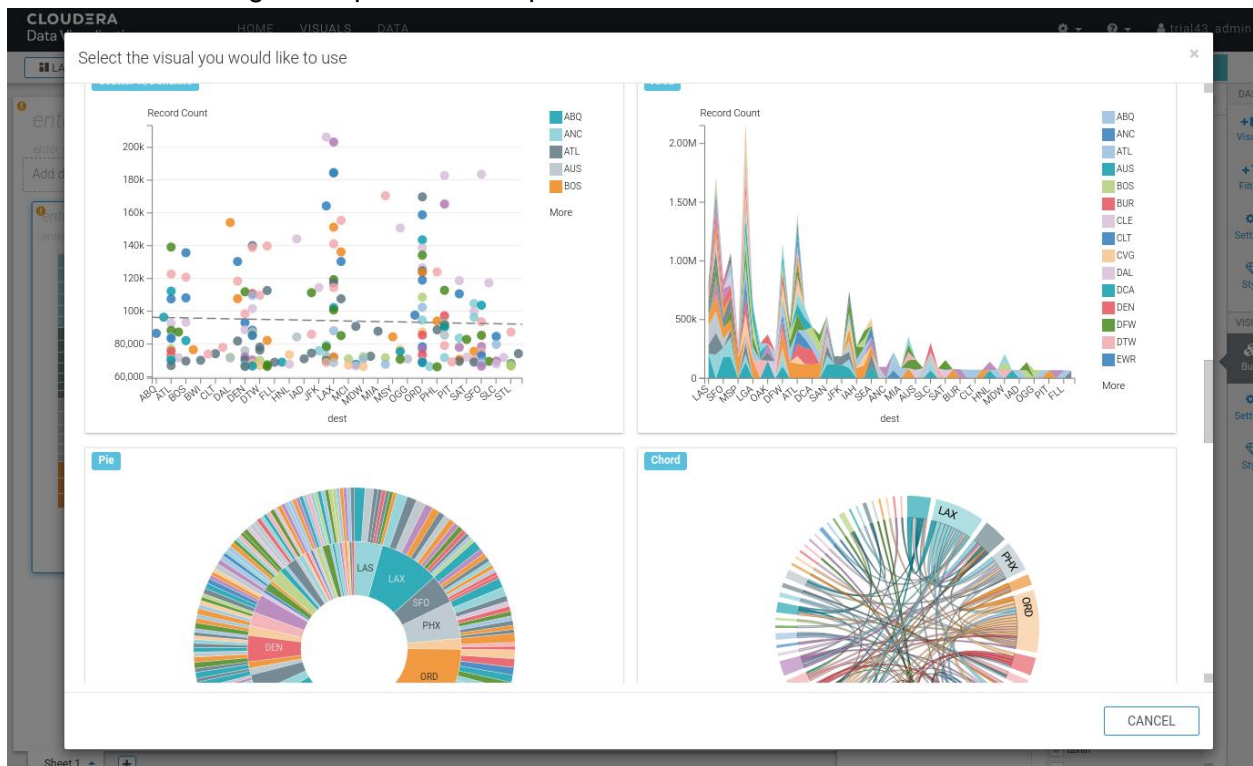
Destinations are displayed



23) Time to explore the other visualization options.



Please scroll through the options and experiment with some of the additional visuals.



Part 4 – How to create a table using data stored in S3

Create table on S3 file [15 minutes]

Options:

Nifi

Command line

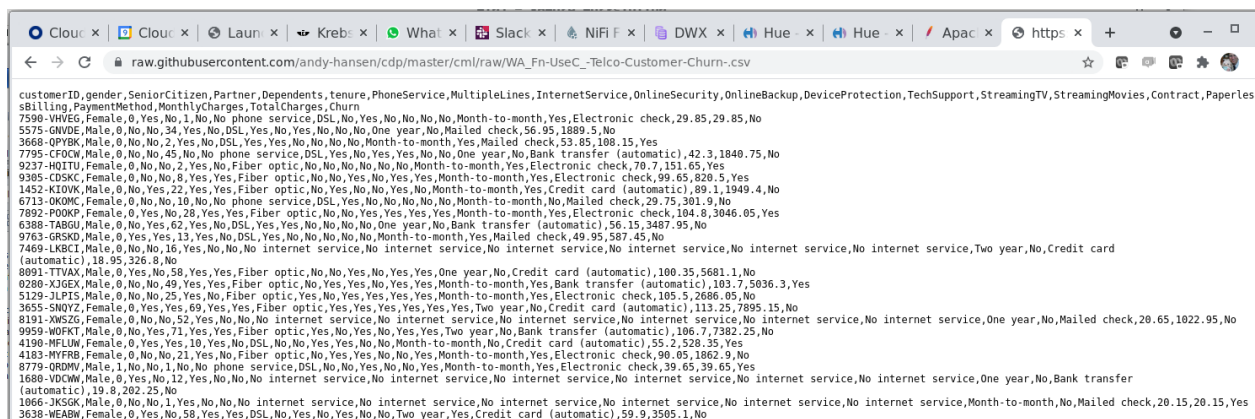
Create file in S3 bucket, run sql to create the table

Overview: What is an Impala table? What is a Hive table?

An Impala/Hive table is a combination of stored data and an entry in the Hive MetaStore. In the early days the stored data lived only in HDFS. Today storage options include HDFS, Kudu, Ozone, S3, ADLS, GoogleObjects, and more. In this lab we'll work with an existing S3 file and create a table on top of the existing data.

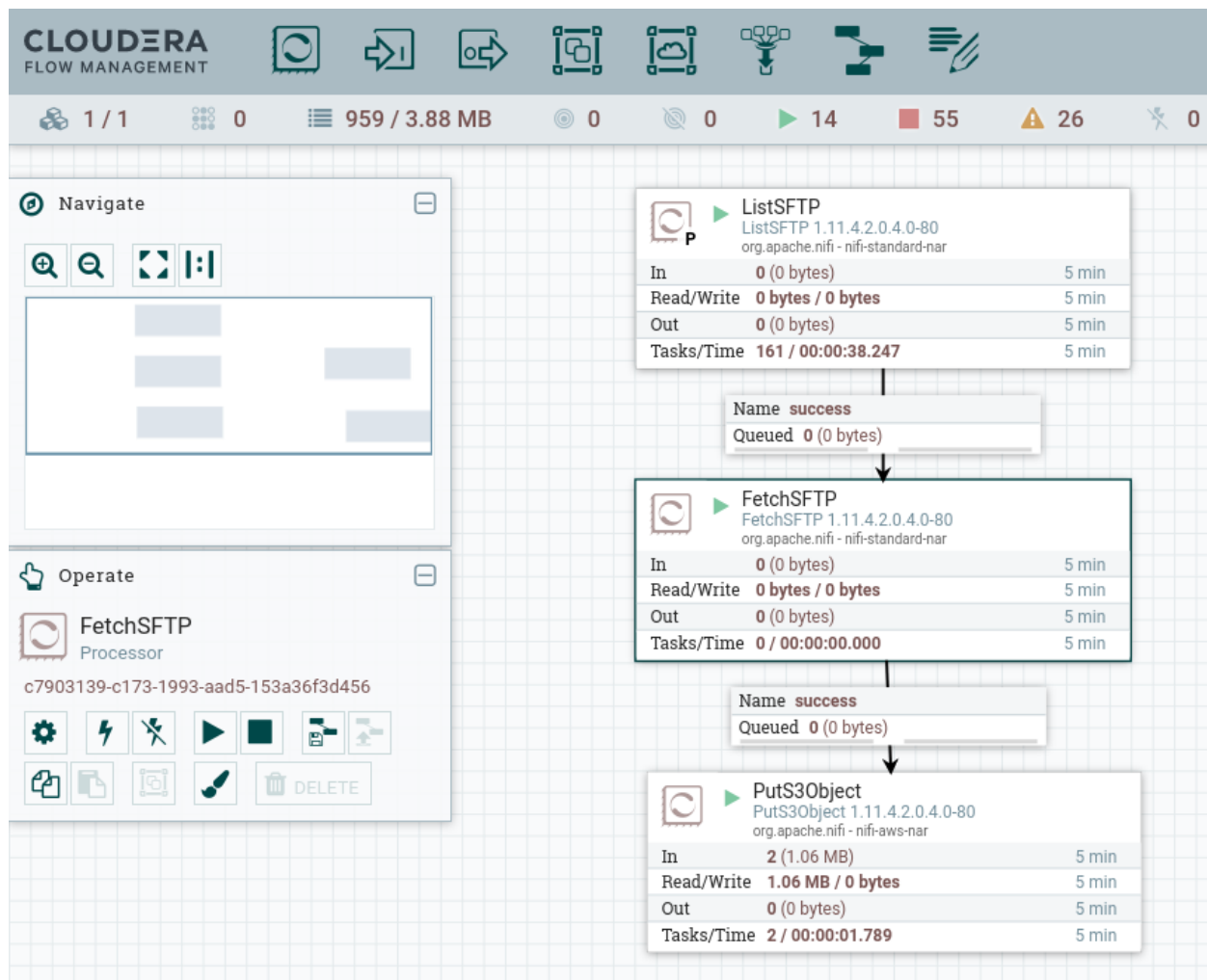
You can look at the data we'll be using at this link:

https://raw.githubusercontent.com/andy-hansen/cdp/master/cml/raw/WA_Fn-UseC_-Telco-Customer-Churn.csv



```
customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,DeviceProtection,TechSupport,StreamingTV,StreamingMovies,Contract,PaperlessBilling,PaymentMethod,MonthlyCharges,TotalCharges,Churn
7590-VHVEG,Female,0,Yes,No,1,No,No,phone service,DSL,No,Yes,No,No,No,Month-to-month,Yes,Electronic check,29.85,29.85,No
5575-QWDEI,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,No,No,No,One year,No,Mailed check,56.95,1889.5,No
3668-QPYBK,Male,0,No,No,2,Yes,No,DSL,Yes,Yes,No,No,No,Month-to-month,Yes,Mailed check,53.85,108.15,Yes
7795-CFOCW,Male,0,No,45,No,No,phone service,DSL,Yes,No,Yes,Yes,No,No,One year,No,Bank transfer (automatic),42.3,1849.75,No
9237-HQITU,Female,0,No,No,2,Yes,No,Fiber optic,No,No,No,No,No,Month-to-month,Yes,Electronic check,70.7,151.65,Yes
9385-CDKXK,Female,0,No,No,8,Yes,Yes,Fiber optic,No,No,Yes,No,Yes,Yes,Month-to-month,Yes,Electronic check,99.65,820.5,Yes
1452-KLOUK,Male,0,No,Yes,22,Yes,Yes,Fiber optic,No,Yes,No,No,Yes,No,Month-to-month,Yes,Credit card (automatic),89.1,1949.4,No
6713-QKMKC,Female,0,No,No,18,No,No,phone service,DSL,Yes,No,No,No,No,Month-to-month,No,Mailed check,29.75,301.9,No
7892-PQOKP,Female,0,Yes,No,28,Yes,Yes,Fiber optic,No,No,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,104.8,3046.05,Yes
6388-TABGU,Male,0,No,Yes,62,Yes,No,DSL,Yes,Yes,No,No,No,One year,No,Bank transfer (automatic),56.15,3487.95,No
9763-QRSKD,Male,0,Yes,Yes,13,Yes,No,DSL,Yes,No,No,No,No,Month-to-month,Yes,Mailed check,49.95,587.45,No
7469-LKICI,Male,0,No,No,16,Yes,No,No,No,Internet service,No,Internet service,No,Internet service,No,Internet service,No,Internet service,Two year,No,Credit card (automatic),18.95,326.8,No
8091-TTVAX,Male,0,Yes,No,58,Yes,Yes,Fiber optic,No,No,Yes,No,Yes,Yes,One year,No,Credit card (automatic),100.35,5681.1,No
0280-XJGEX,Male,0,No,No,49,Yes,Yes,Fiber optic,No,Yes,Yes,No,Yes,Yes,Month-to-month,Yes,Bank transfer (automatic),103.7,5036.3,Yes
5129-JLPTS,Male,0,No,No,25,Yes,No,Fiber optic,Yes,No,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,105.5,2686.05,No
3655-SNOYZ,Female,0,Yes,Yes,69,Yes,Yes,Fiber optic,Yes,Yes,Yes,Yes,Yes,Two year,No,Credit card (automatic),113.25,7895.15,No
8191-XWSZG,Female,0,No,No,52,Yes,No,No,Internet service,No,Internet service,No,Internet service,No,Internet service,No,Internet service,One year,No,Mailed check,20.65,1022.95,No
9950-WQKFT,Male,0,No,Yes,71,Yes,Yes,Fiber optic,Yes,No,Yes,No,Yes,Yes,Two year,No,Bank transfer (automatic),106.7,7382.25,No
4190-MFLIW,Female,0,Yes,Yes,10,Yes,No,DSL,No,No,Yes,Yes,No,No,Month-to-month,No,Credit card (automatic),55.2,528.35,Yes
4183-MYFRB,Female,0,No,No,21,Yes,No,Fiber optic,No,Yes,Yes,No,No,Yes,Month-to-month,Yes,Electronic check,90.05,1862.9,No
8779-QRDWV,Male,1,No,No,1,No,No,phone service,DSL,No,No,Yes,No,No,Yes,Month-to-month,Yes,Electronic check,39.65,39.65,Yes
1689-VDCWV,Male,0,Yes,No,12,Yes,No,No,No,Internet service,No,Internet service,No,Internet service,No,Internet service,No,Internet service,One year,No,Bank transfer (automatic),19.8,202.25,No
1866-JK5GK,Male,0,No,No,1,Yes,No,No,No,Internet service,No,Internet service,No,Internet service,No,Internet service,No,Internet service,Month-to-month,No,Mailed check,20.15,20.15,Yes
3638-WEABW,Female,0,Yes,No,58,Yes,Yes,DSL,No,Yes,No,Yes,No,No,Two year,Yes,Credit card (automatic),59.9,3505.1,No
```

How did the file get into S3? There are many options to populate S3, ADLS, Google Object store etc. Shown below is a sample Nifi flow that does a remote sftp and puts files into S3. We hope you can join us for a Nifi webinar in the future.



Your lab team will tell you the location of the file in S3. Use that location in the SQL below.

For reference, here are the commands used in a prior lab to stage the datafile in S3

```
hadoop fs -mkdir s3a://prod-cdptrialuser43-trycdp-com/cdp-lake/data/cdpvw1/telcochurn
hadoop fs -mkdir s3a://prod-cdptrialuser43-trycdp-com/cdp-lake/data/cdpvw2/telcochurn
hadoop fs -mkdir s3a://prod-cdptrialuser43-trycdp-com/cdp-lake/data/cdpvw3/telcochurn
hadoop fs -put WA_Fn-UseC_-Telco-Customer-Churn.csv s3a://prod-cdptrialuser43-trycdp-com/cdp-lake/data/cdpvw1/telcochurn
hadoop fs -put WA_Fn-UseC_-Telco-Customer-Churn.csv s3a://prod-cdptrialuser43-trycdp-com/cdp-lake/data/cdpvw2/telcochurn
hadoop fs -put WA_Fn-UseC_-Telco-Customer-Churn.csv s3a://prod-cdptrialuser43-trycdp-com/cdp-lake/data/cdpvw3/telcochurn
hadoop fs -ls s3a://prod-cdptrialuser43-trycdp-com/cdp-lake/data/cdpvw3/telcochurn
```

1) Open CDP, using the “admin” user within the Test Drive link.

Your link should look something like (remember click the link in your email not the link below)

http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X

*xx represents the trial user #

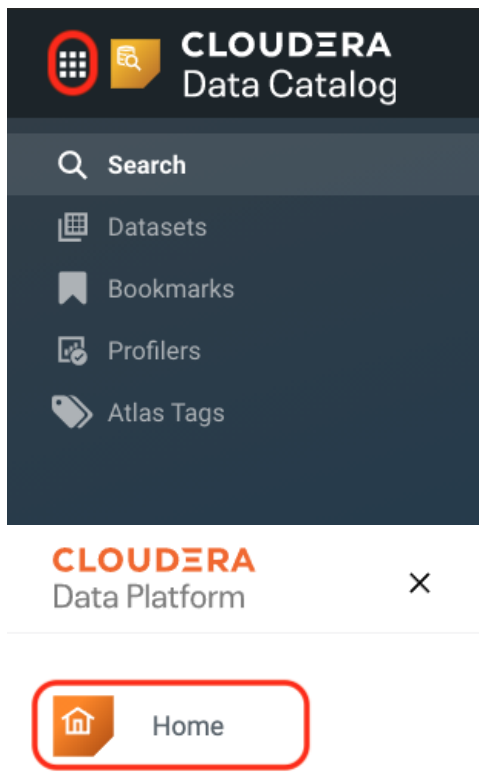
*X represents the password

2) Click the “Data Warehouse” within the CDP Home Screen



How do you get to the CDP Home Screen?

- From any experience such as “Data Catalog”, click the 9 square at the top left and then click “Home”

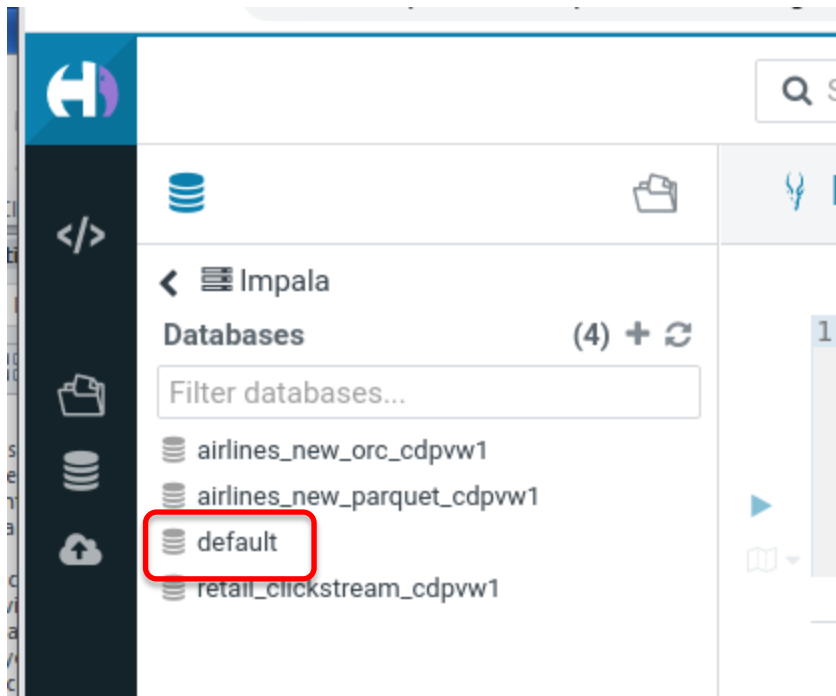


3) Open HUE, be sure you're going to the virtual warehouse

*The same steps you did in Part 2 to Open HUE

The screenshot shows the Cloudera Data Warehouse (CDW) Overview page. The left sidebar contains navigation links for Overview, Database Catalogs, and Virtual Warehouses. The main content area is divided into two sections: Database Catalogs (4) and Virtual Warehouses (3). The Database Catalogs section lists four catalogs: cdpww3, cdpww2, cdpww1, and cdptrialuser43-datalake-... Each catalog has a status (Running or Stopped), a warehouse ID, and a table showing Total Cores (7), Total Memory (17 GB), and Virtual Warehouses (1 or 0). The Virtual Warehouses section lists three warehouses: cdpww1, cdpww3, and cdpww2. Each warehouse has a status (Stopped), an Impala ID, and a table showing Node Count (0), Total Cores (20), Total Memory (137 GB), and Type (IMPALA). A red box highlights the 'HUE' button next to the 'cdpww1' virtual warehouse.

4) Navigate to the "Default" database



5) The following SQL will create a table on top of the file that lives in S3. Remember, a “hive table” or an “impala table” is really a data file and an entry in the HiveMetastore relational database.

```
create external table if not exists {YOURNAMEHERE}_telcochurn (  
  customerid string,  
  gender string,  
  seniorcitizen string,  
  partner string,  
  dependents string,  
  tenure int,  
  phoneservice string,  
  multiplelines string,  
  internetservice string,  
  onlinesecurity string,  
  onlinebackup string,  
  deviceprotection string,  
  techsupport string,  
  streamingtv string,  
  streamingmovies string,  
  contract string,  
  paperlessbilling string,  
  paymentmethod string,  
  monthlycharges decimal(10,2),  
  totalcharges decimal(10,2),  
  churn int  
)  
row format delimited fields terminated by ','  
stored as textfile  
location 's3a://USETHEPATHNAMEPROVIDEDBYYOURLABTEAM/telcochurn'  
;  
-- for example 's3a://prod-cdptrialuser43-trycdp-com/cdp-  
lake/data/cdpvw2/telcochurn'
```

5) Trust but verify

```
select * from yournamehere_telco_churn limit 10;  
select count(*) from yournamehere_telco_churn;  
select sum(totalcharges) from yournamehere_telcochurn;
```

Did you remember optimizer statistics? Please do a compute stats when you create a table or when a table changes by approx 15%.

```
compute stats yournamehere_telco_churn;
```

6) Multiple schemas on one file? Time to experiment with the power of “schema on read”

We aren't limited to only one schema for each file. The SQL below will create another table, but this time the entire row will appear as a single column. This is very useful when using SQL as a super-search in log files or other text files.

```
create external table if not exists {YOURNAMEHERE}_telcochurn_onecol (
myonecolumn string
)
row format delimited fields terminated by '^'
stored as textfile
location 's3a://USETHEPATHNAMEPROVIDEDBYYOURLABTEAM/telcochurn'
;
-- for example 's3a://prod-cdptrialuser43-trycdp-com/cdp-
lake/data/cdpvw2/telcochurn'
```

Now we have some interesting capabilities to search, treating every row as a single string:

```
select * from yournamehere_telcochurn_onecol;

select * from yournamehere_telcochurn_onecol
where myonecolumn like "%Mailed%";
```