# CLOUDERA 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”.

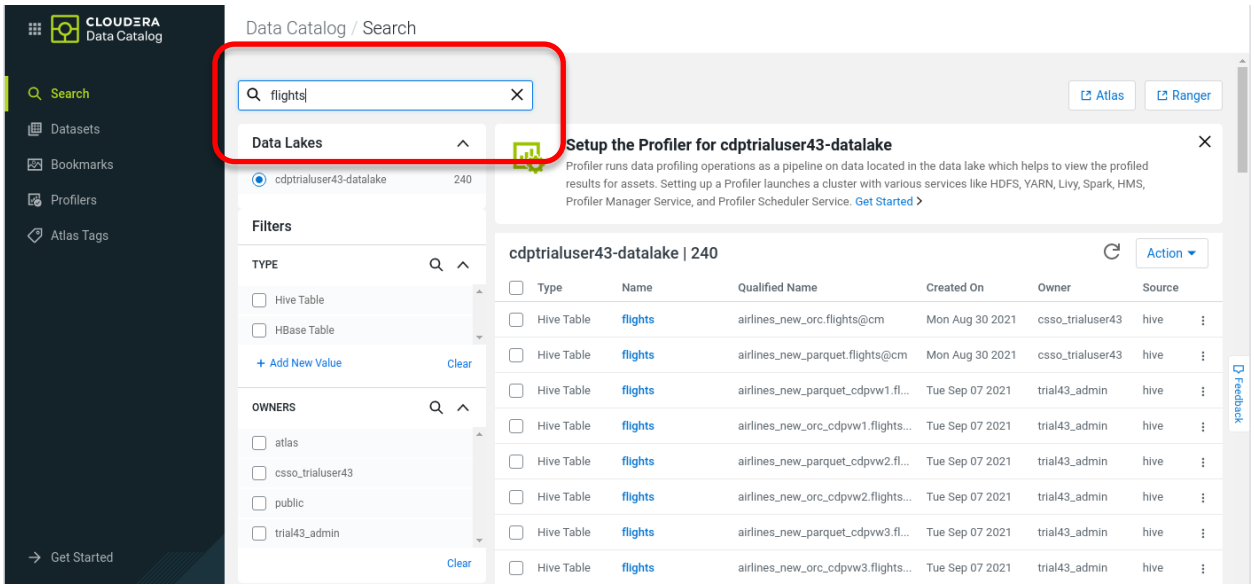
- Find what database(s) the table “flights” is located.
- Find out at least one year that the “flights” table was generated from.
- Find out how many columns the table “flights” contains.

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

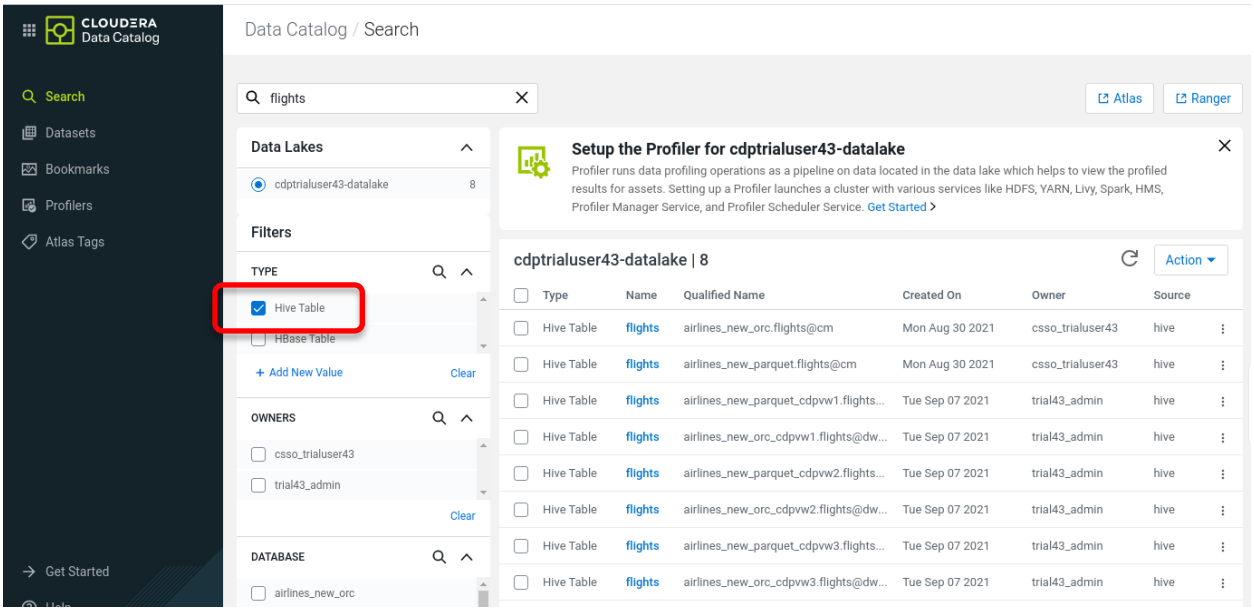
Your link should look something like (remember click the link in your email not the link below)
http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X
*xx represents the trial user #
*X represents the password

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

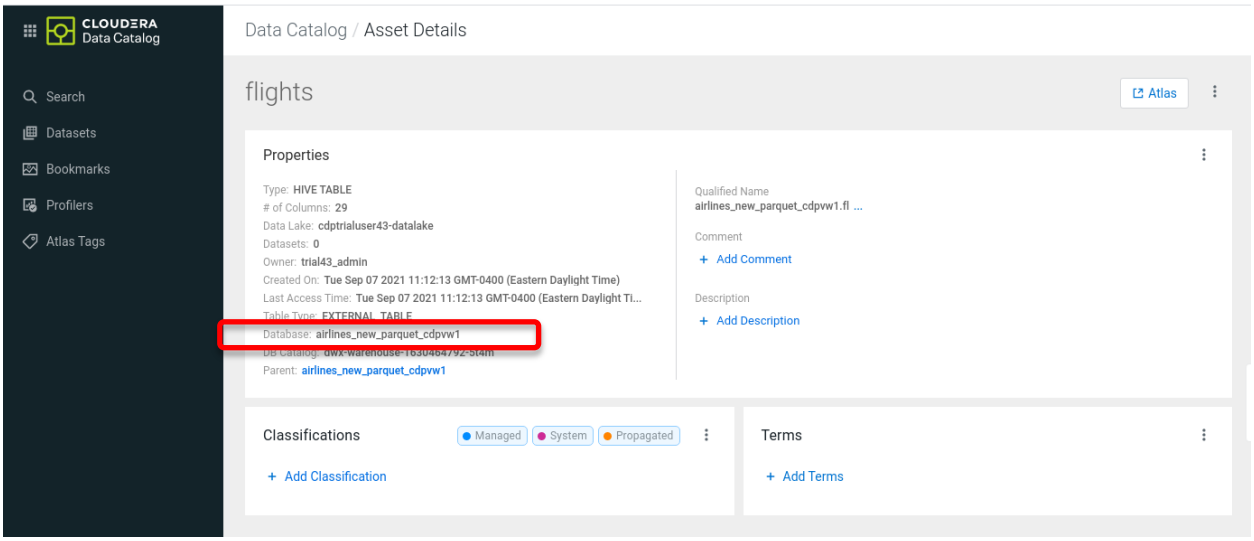

3) Type “flights” in the search box and press enter on your keyboard



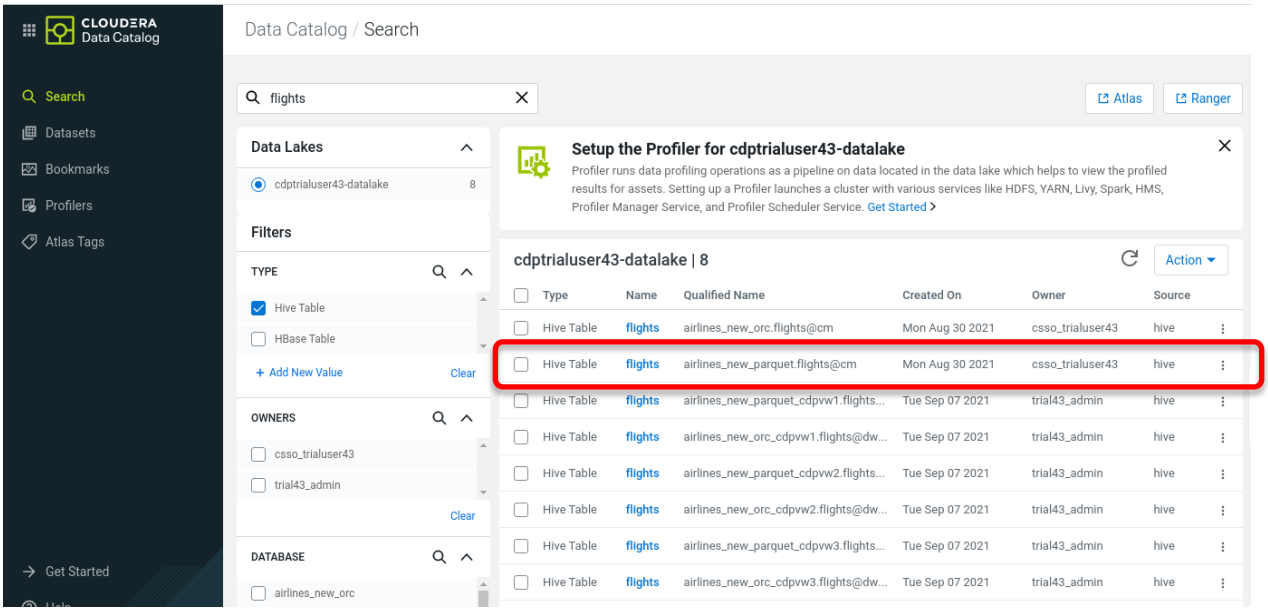

4) Click “Hive Table” under Filters on the left



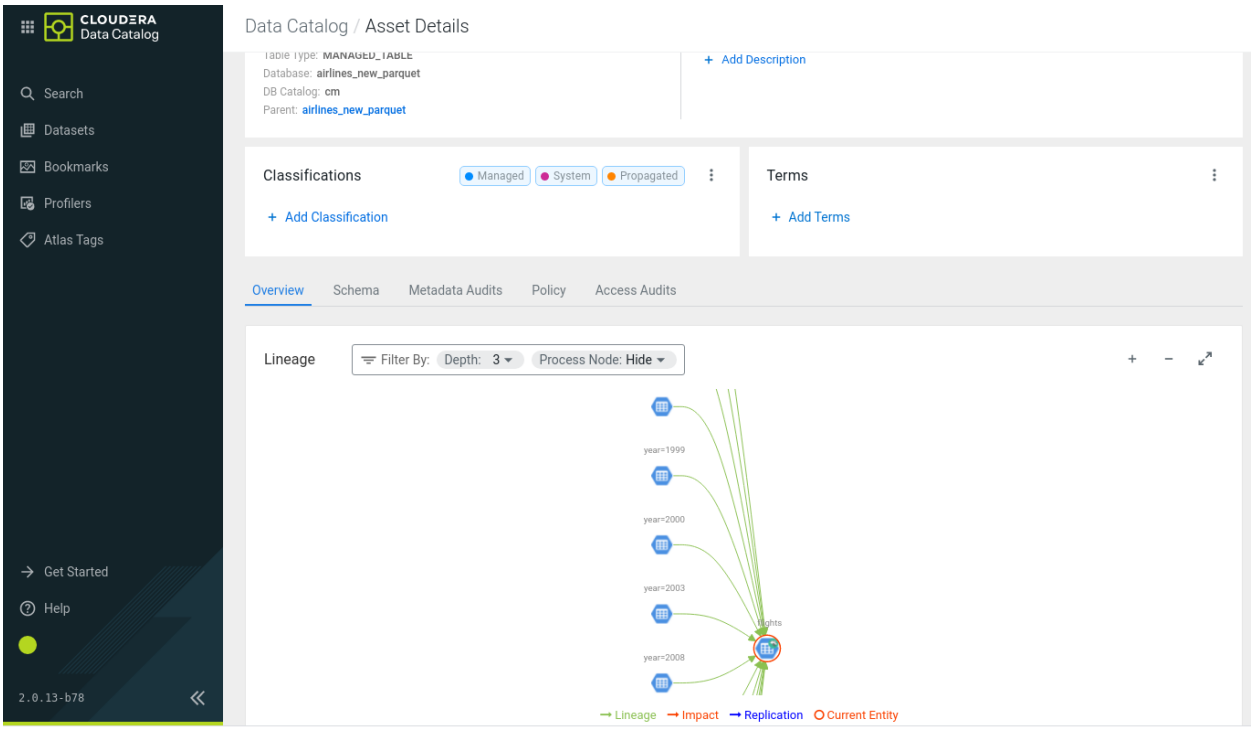

*Find what database(s) the table “flights” is located.




5) Click “flights” table for your breakout CDW




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




*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?

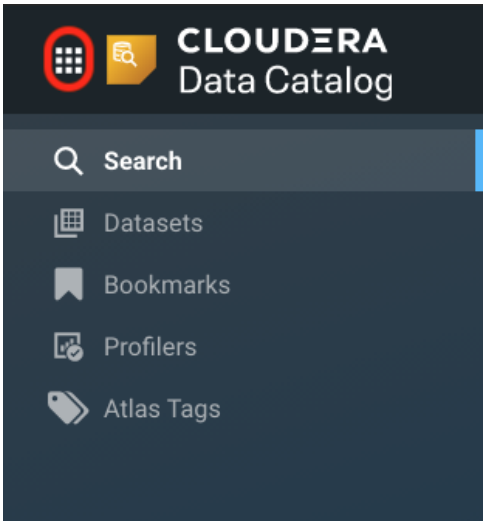
1) Open CDP, using the “admin” user within the Test Drive link.

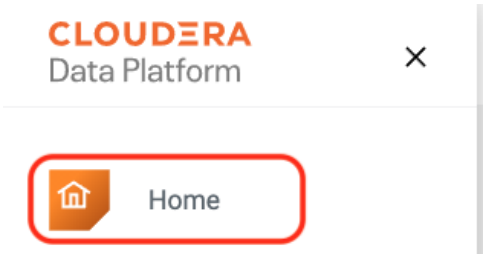
Your link should look something like (remember click the link in your email not the link below)
http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X
*xx represents the trial user #
*X represents the password

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


How do you get to the CDP Home Screen?

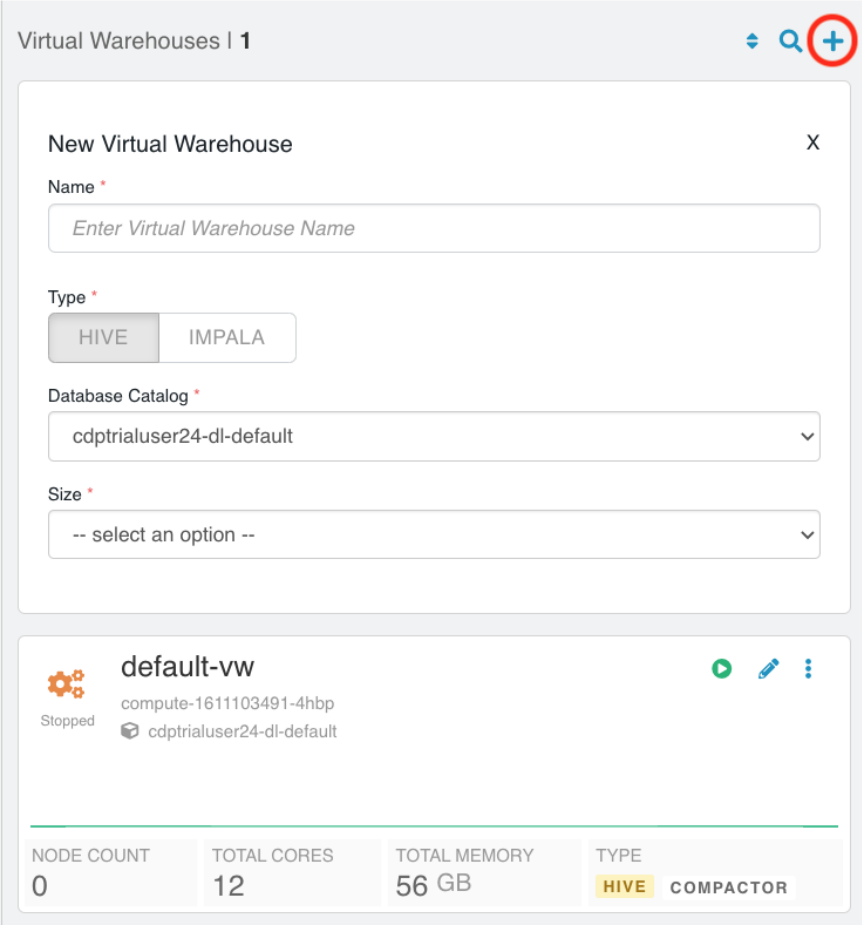
- From any experience such as “Data Catalog”, click the 9 square at the top left and then click “Home”



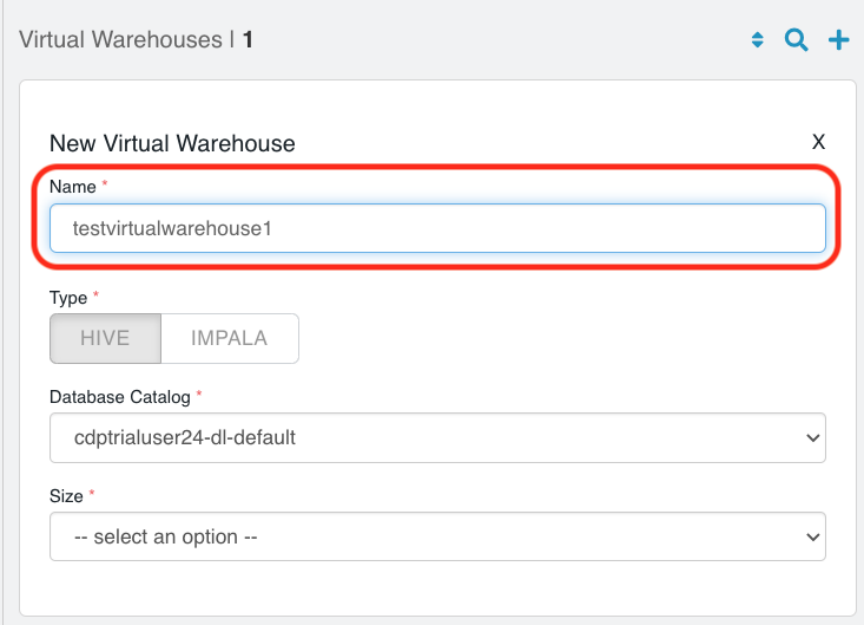

CLOUDERA
Data Platform
Home

# ***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”

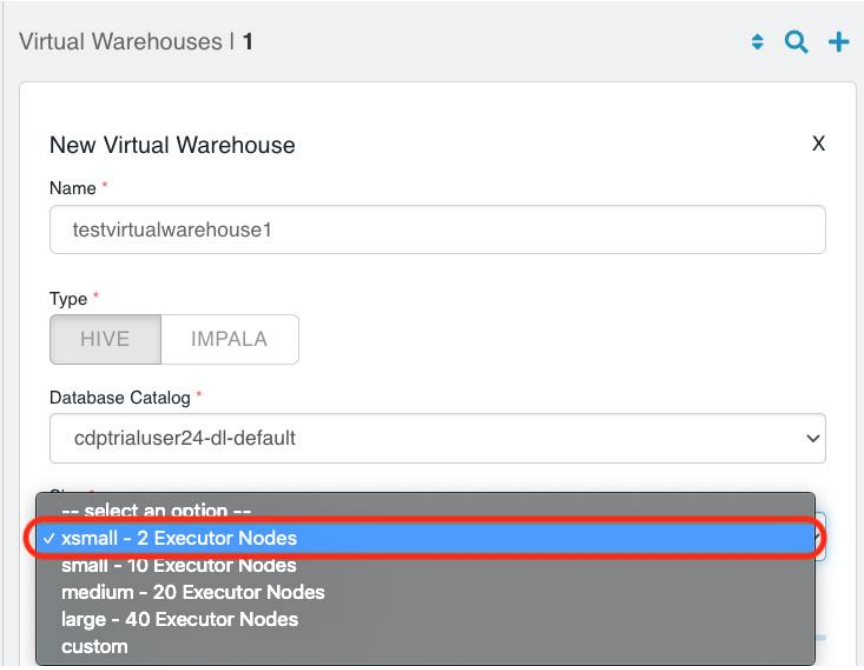



4) DO NOT  Enter a name for your New Virtual Warehouse



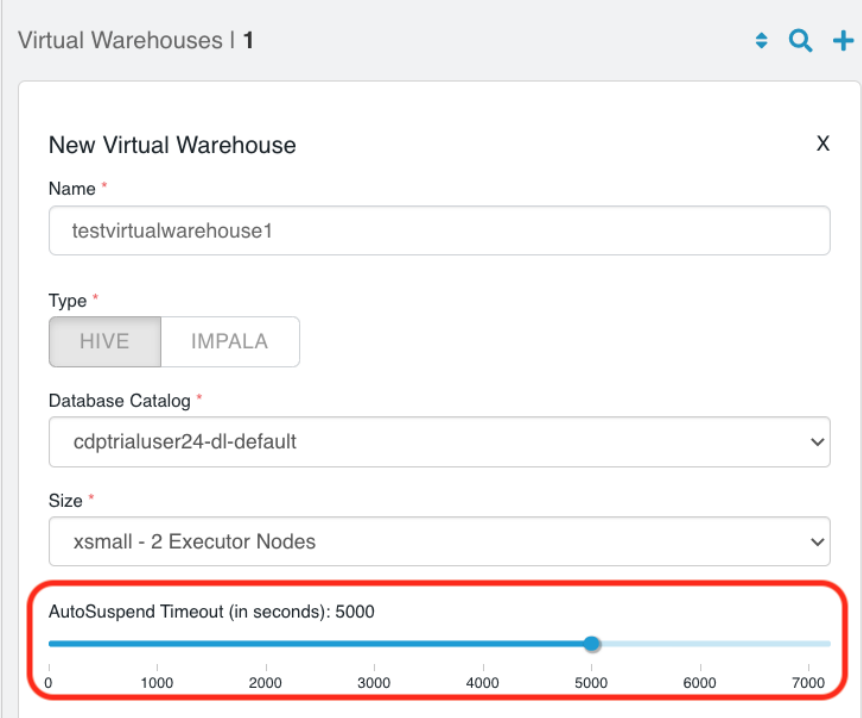

5) ) DO NOT  Select the Size of “xsmall - 2 Executor Nodes”

*How do I choose a size? Initial concurrent users



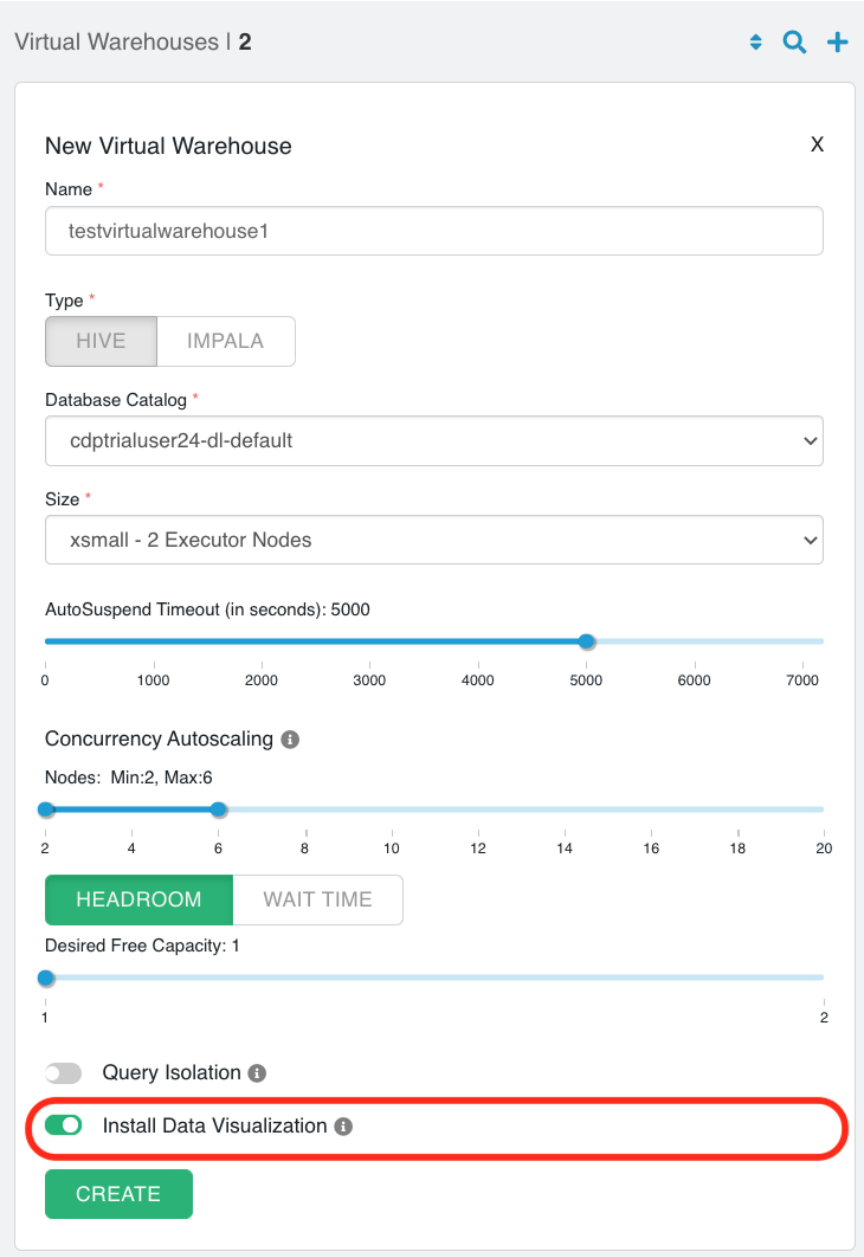

6) ) To save money you stop the instances you aren’t using.  Cloudera lets you define if you spin down to zero, if you have some Kubernetes pods running all the time, and how long these live when there is no workload.  DO NOT Set the AutoSuspend Timeout (in seconds) between 4500 and 5500:

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




7) DO NOT Choose “Install Data Visualization” to be on

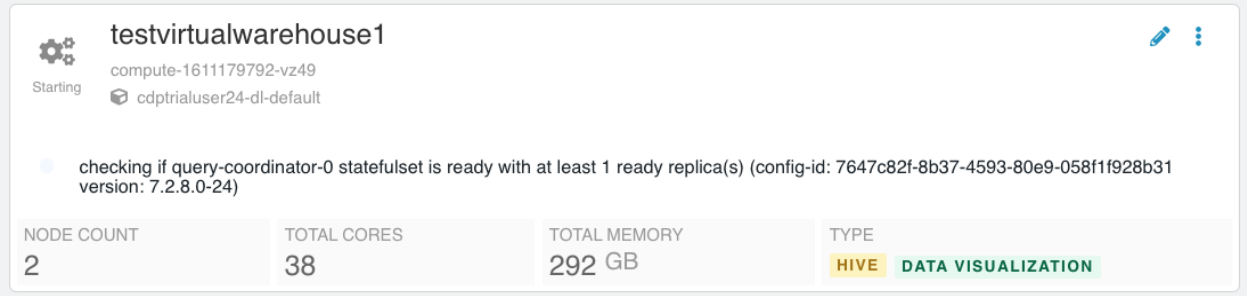
*Allowing for Data Visualizations in Part 3

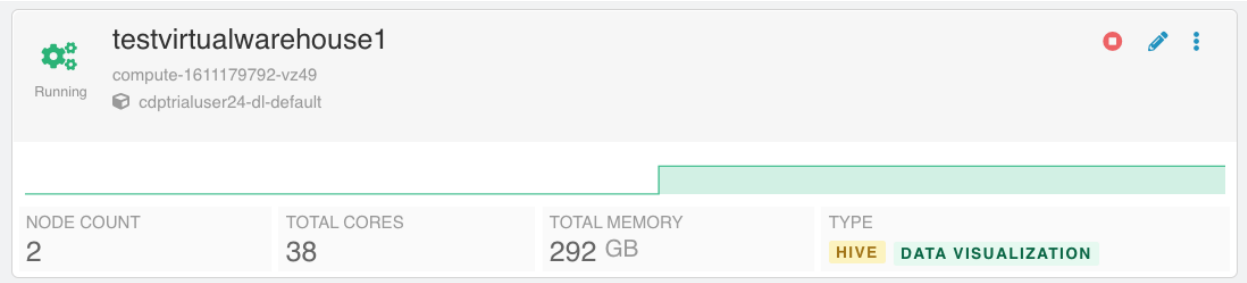



8) DO NOT , DO NOT , DO NOT, REALLY DO NOT: Click “Create” to create your Virtual Warehouse

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


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

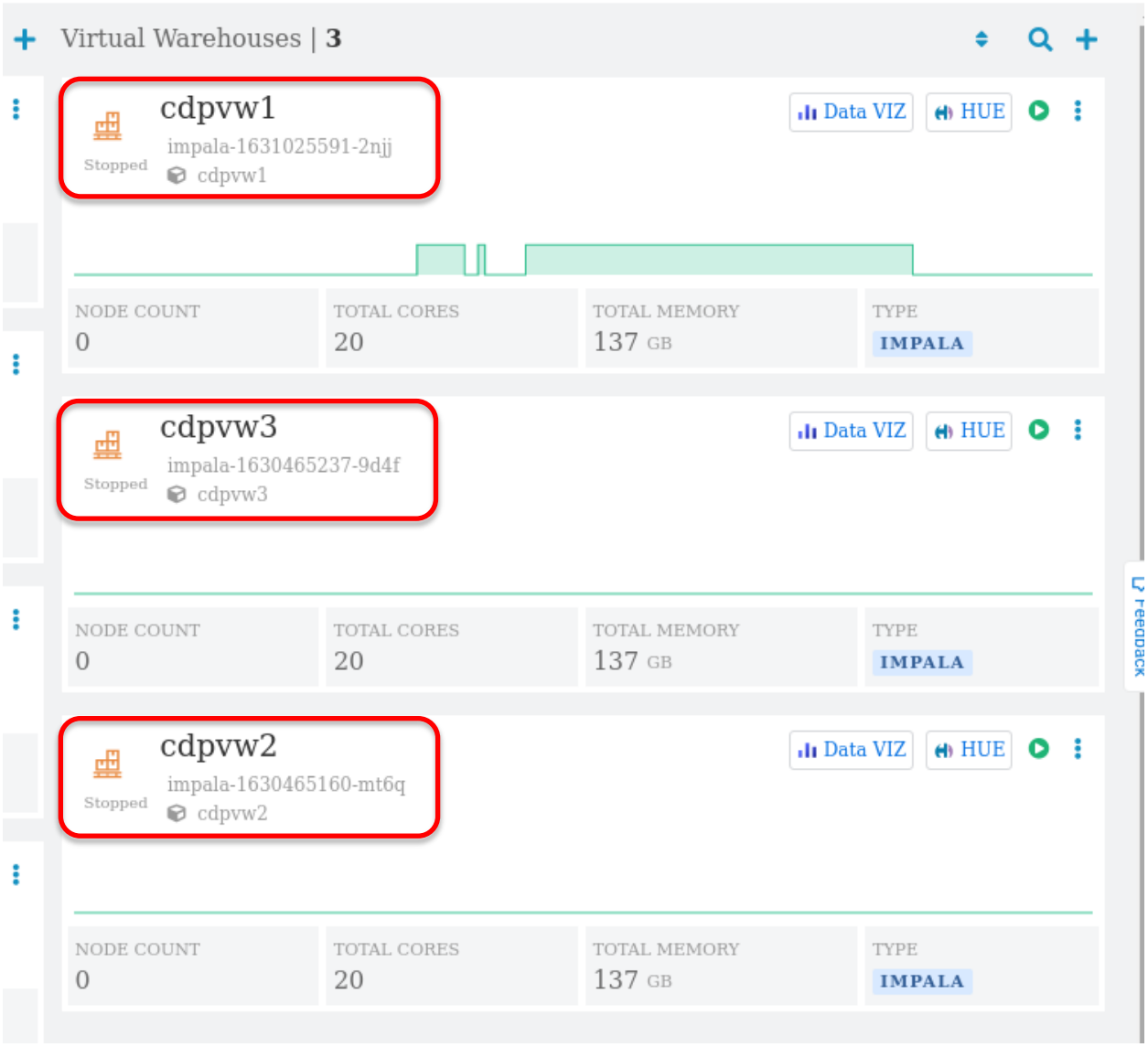


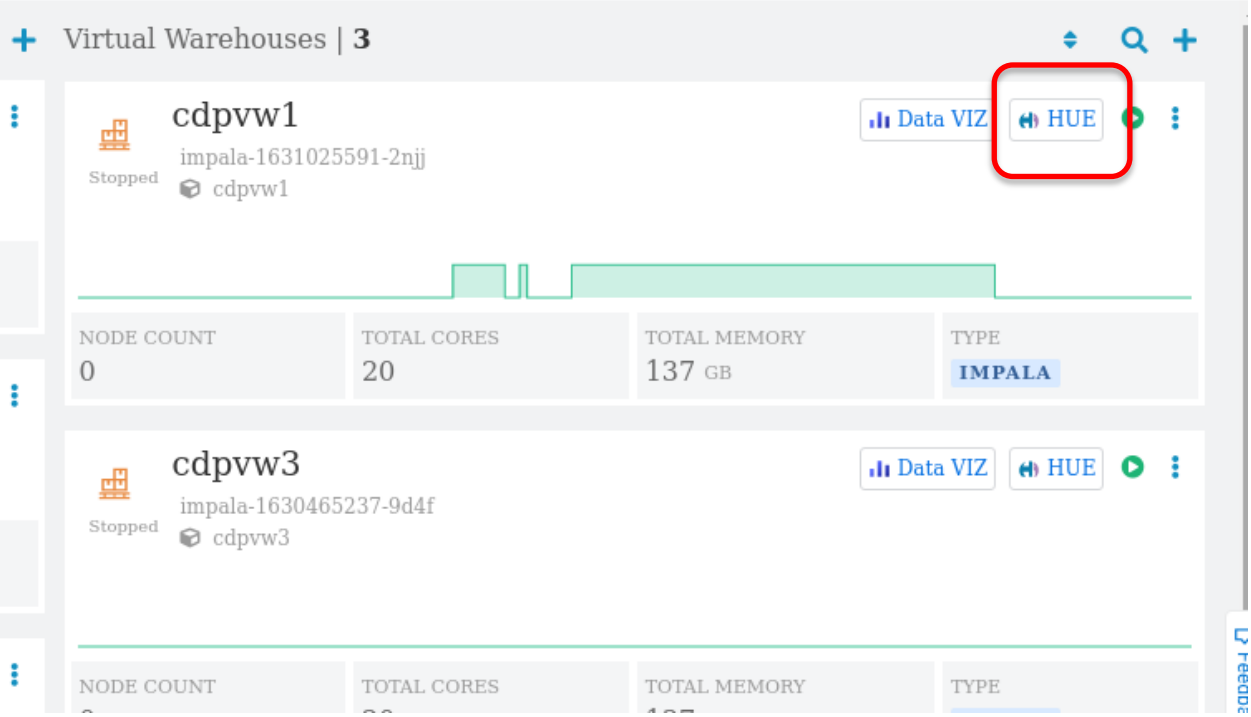

If you can read this it is the end of the background information and it is

# 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 and you have admin powers.***

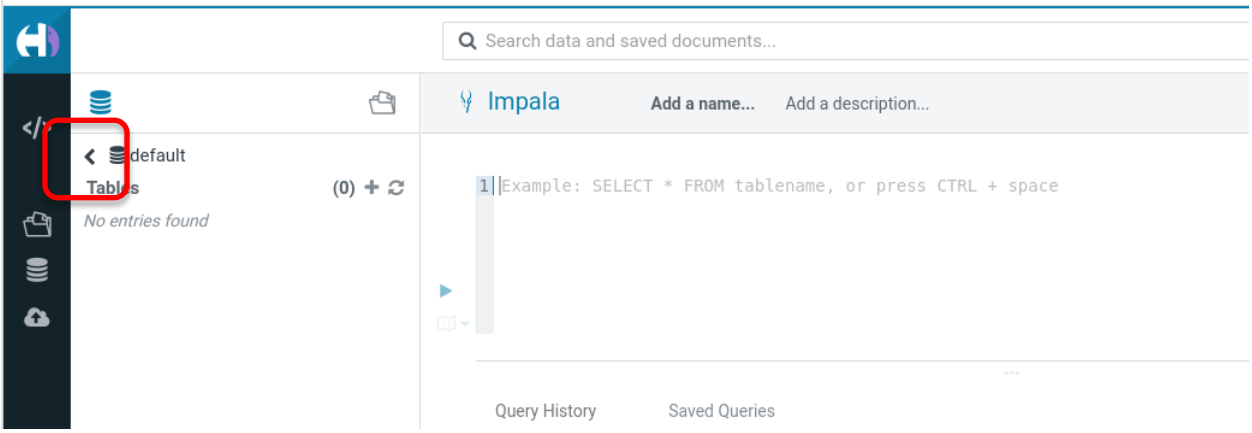
9) Notice there multiple Virtual Warehouses (VWs). You will be working on one of the VWs. If you are in Zoom Breakout Room 1 use cdpvw1, room2 uses cdpvw2 etc



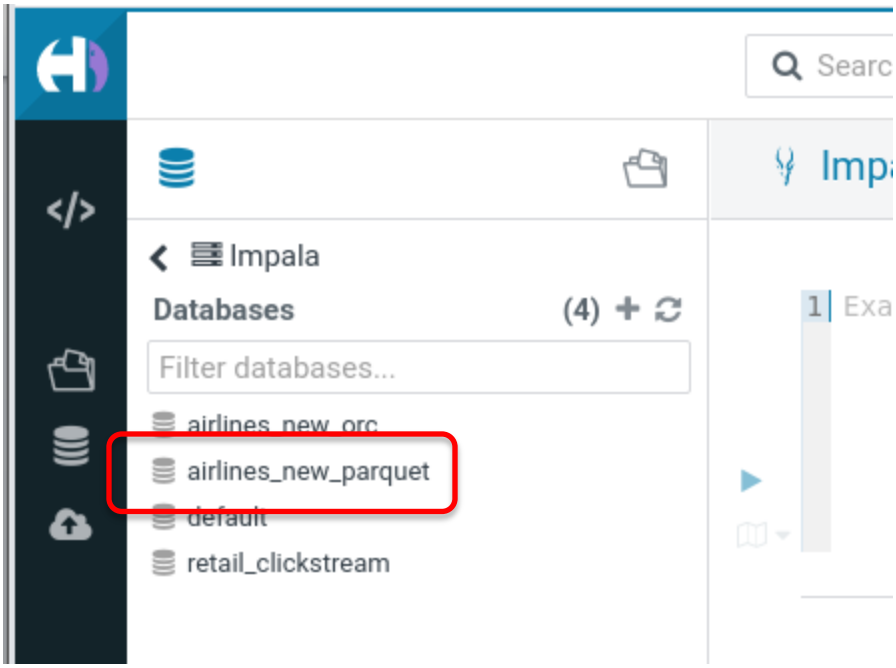

10) Click on HUE to enter the “Hadoop User Experience” in your designated room.




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.

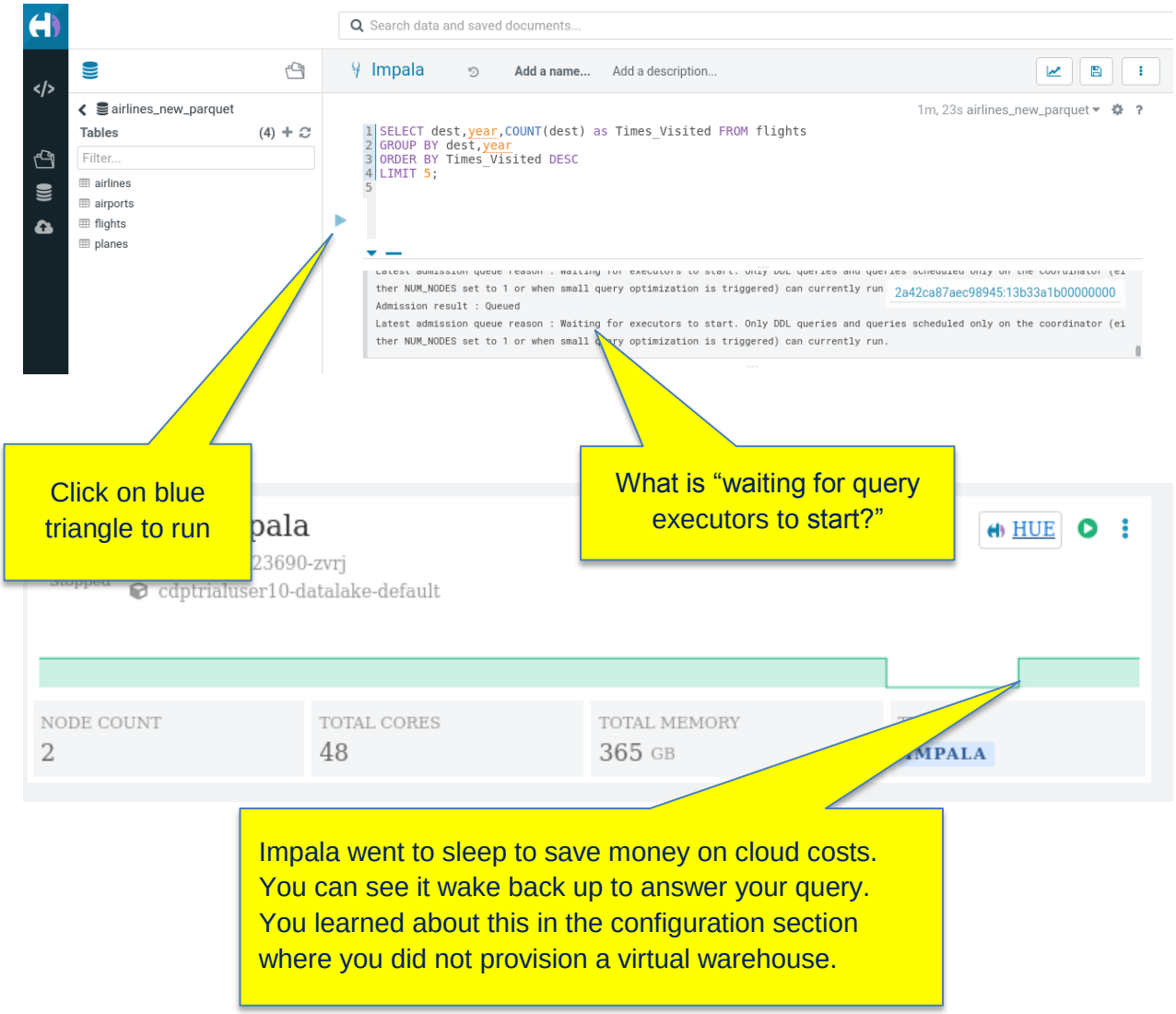



If you ever need to get back to this screen layout click on the “editor” shown here:


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;
```

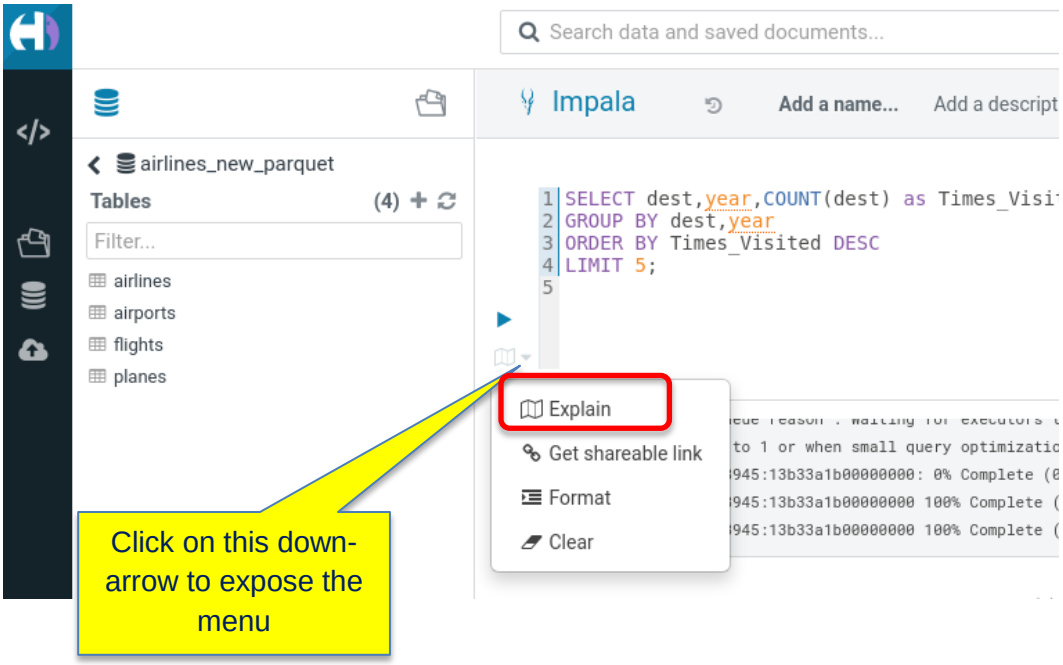
Why doesn’t it run right away? You promised it was fast ☺

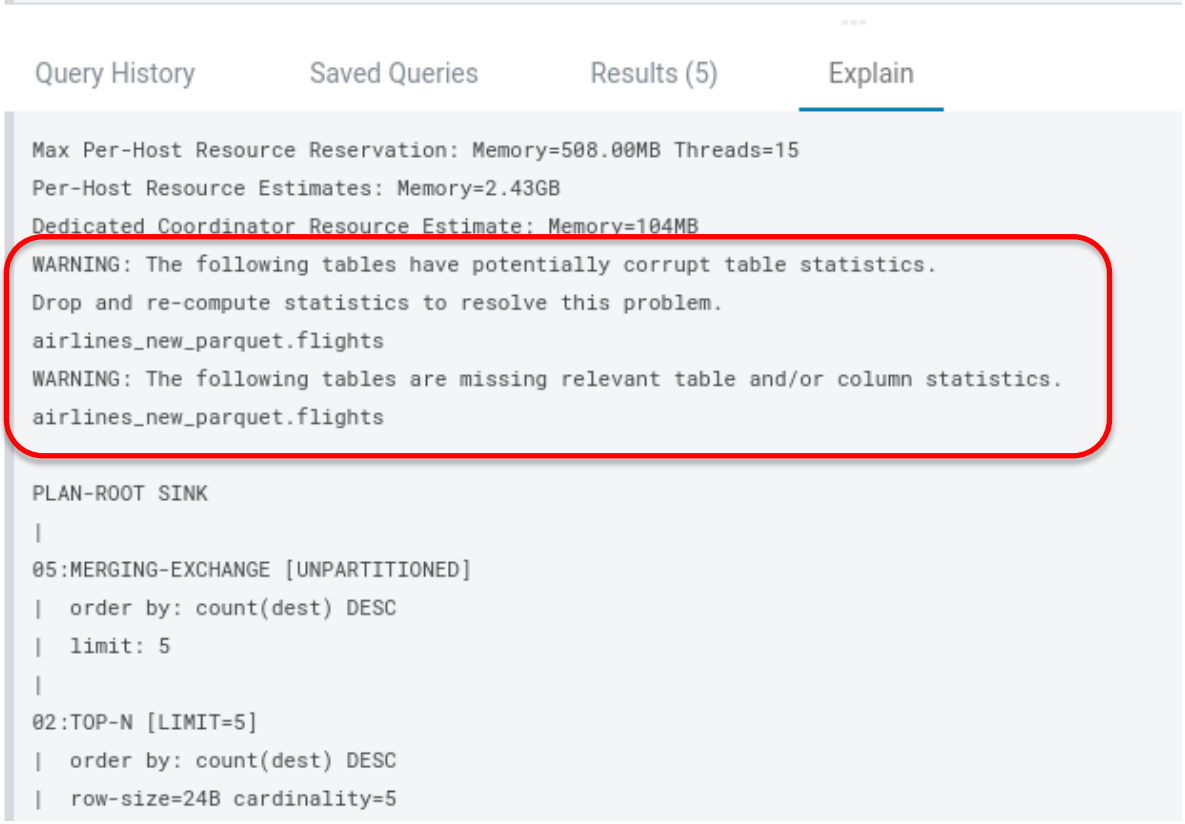



You may get the query result without having to wait for Impala to “wake up.” You get to decide how long an idle time you want to wait before you scale down to zero, or if you have the budget you can have Impala always available.

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



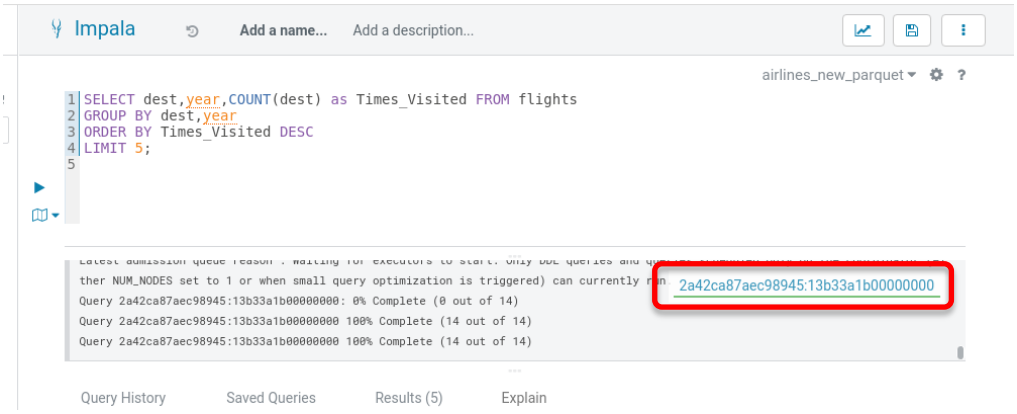




The “Explain Plan” shows you how the query will execute. It is asking you to update statistics for the tables in the query. This is a good idea for performance. After everyone reaches this point in the lab designate one person to run: **`compute stats airlines_new_parquet.flights;`**

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



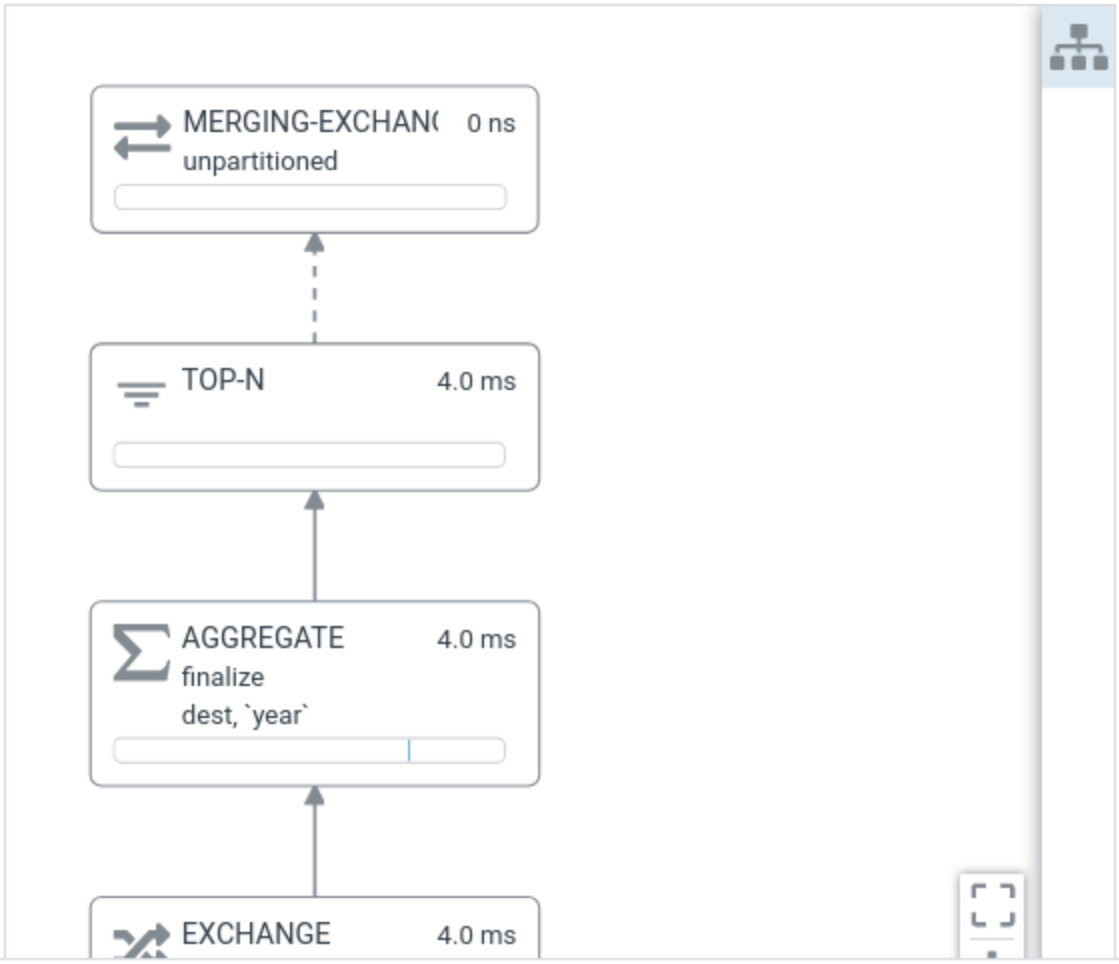

16) Explore the different tabs in the query pop-up. The Visual Plan shows how the query was executed. These get more interesting with multi table joins. The Summary show how much time was spent in each stage of the query, the memory used, and the rows produced. The Profile includes the summary and great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

Impala

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

ID
2a42ca87aec98945:13b33a1b00000000

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;
```

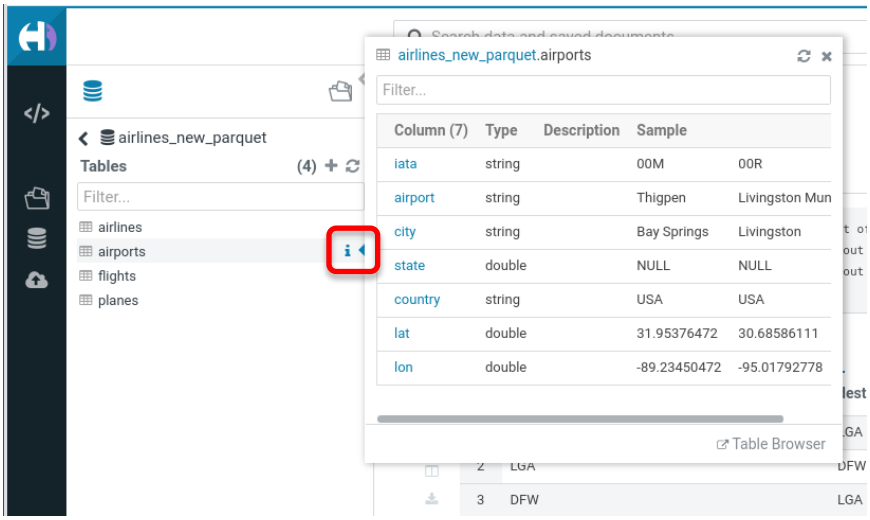
18) Notice the highlighting on the left edge. HUE is parsing based on the semi-colon and the execution arrow will run whatever is highlighted in blue, or whatever has been highlighted by the cursor. This way you can have multiple queries in the same canvas.


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?”

Query History  Saved Queries  Results (10)

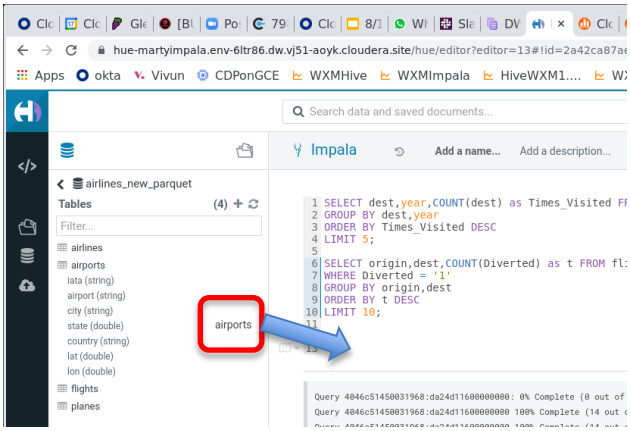
| | origin | dest | t |
|---|---|---|---|
| 1 | ORD | LGA | 845 |
| 2 | LGA | DFW | 749 |
| 3 | DFW | LGA | 653 |
| 4 | DAL | HOU | 615 |
| 5 | ATL | LGA | 567 |
| 6 | MDW | STL | 512 |
| 7 | ATL | DFW | 482 |
| 8 | ORD | DFW | 450 |

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.

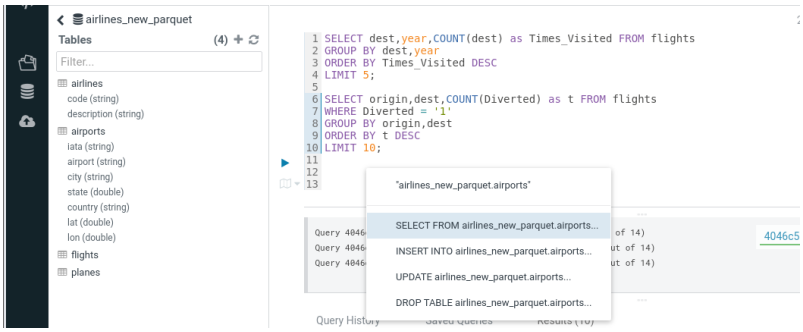



You can also click on the table name to expand all the columns. On the far right of the browser is help tooling.

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



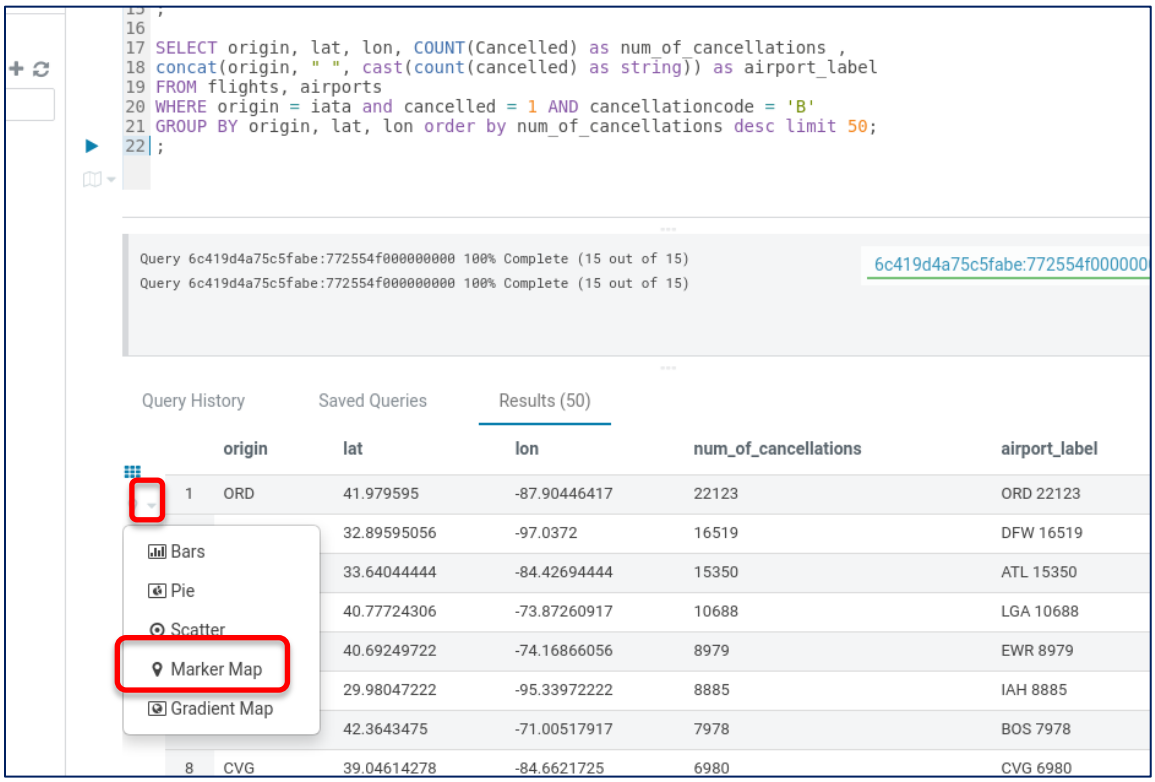
When you drop the table name you'll get to choose what SQL you want auto generated. Take the "select" option and run the query with the blue-triangle




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

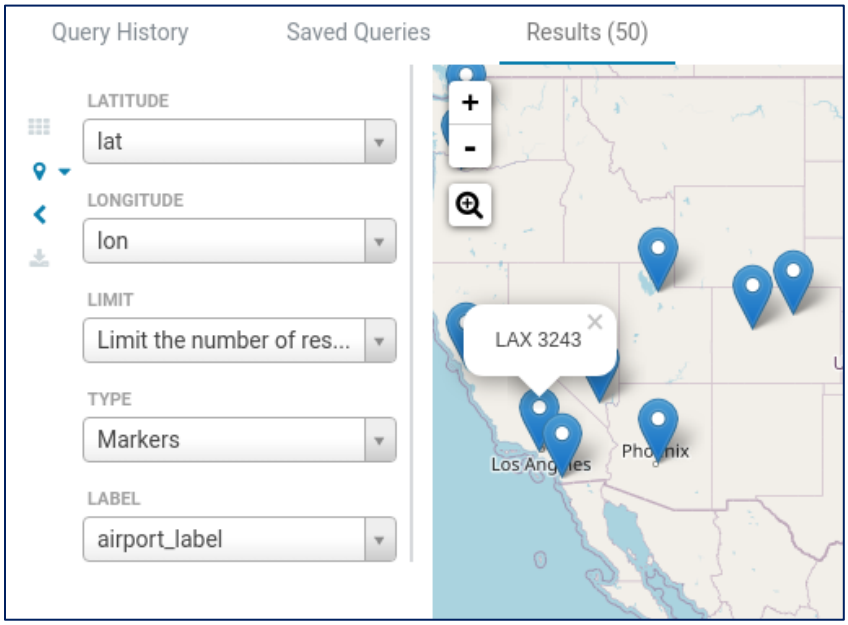
```
SELECT origin, lat, lon, COUNT(Cancelled) as num_of_cancellations ,
concat(origin, " ", cast(count(cancelled) as string)) as airport_label
FROM flights, airports
WHERE origin = iata and cancelled = 1 AND cancellationcode = 'B'
GROUP BY origin, lat, lon order by num_of_cancellations desc limit 50;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

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

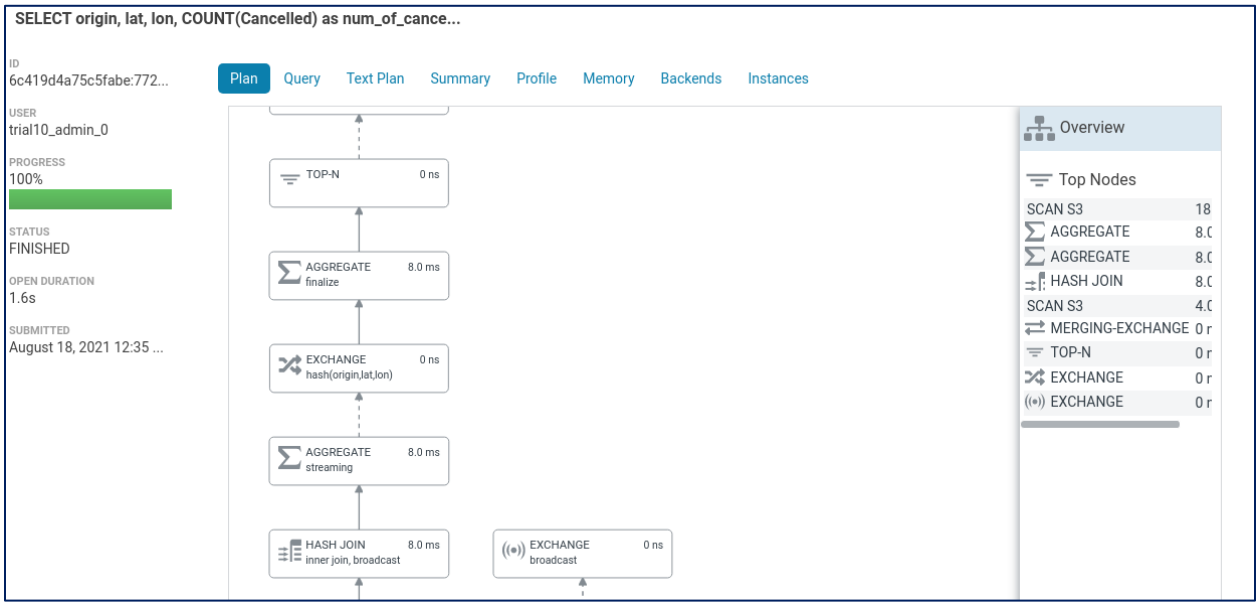



| | origin | lat | lon | num_of_cancellations | airport_label |
|---|---|---|---|---|---|
| 1 | ORD | 41.979595 | -87.90446417 | 22123 | ORD 22123 |
| | | 32.89595056 | -97.0372 | 16519 | DFW 16519 |
| | | 33.64044444 | -84.42694444 | 15350 | ATL 15350 |
| | | 40.77724306 | -73.87260917 | 10688 | LGA 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 |

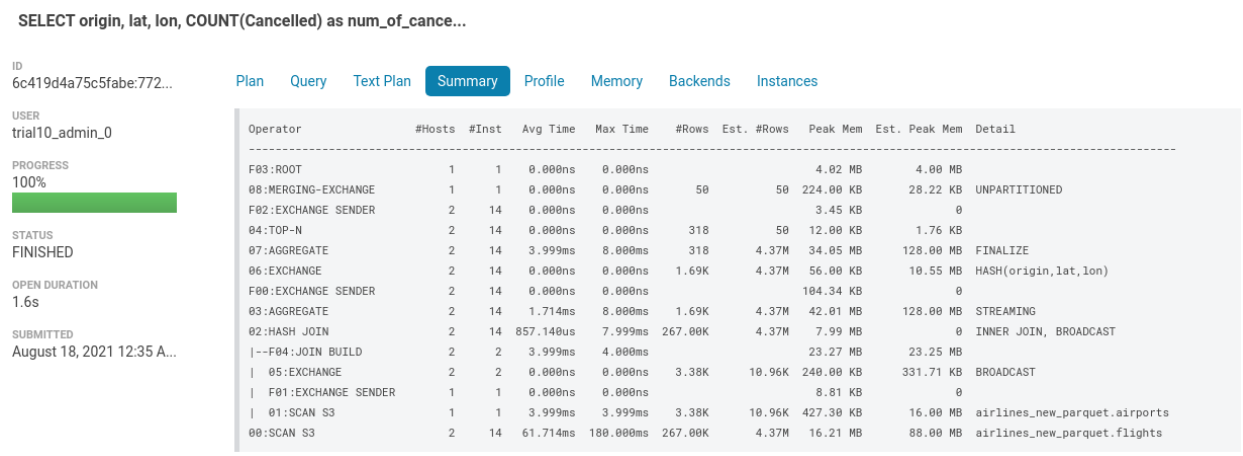
24) Configure the marker map per shown below. Clicking on one of the markers will pop up the value of the “ariport_label” column



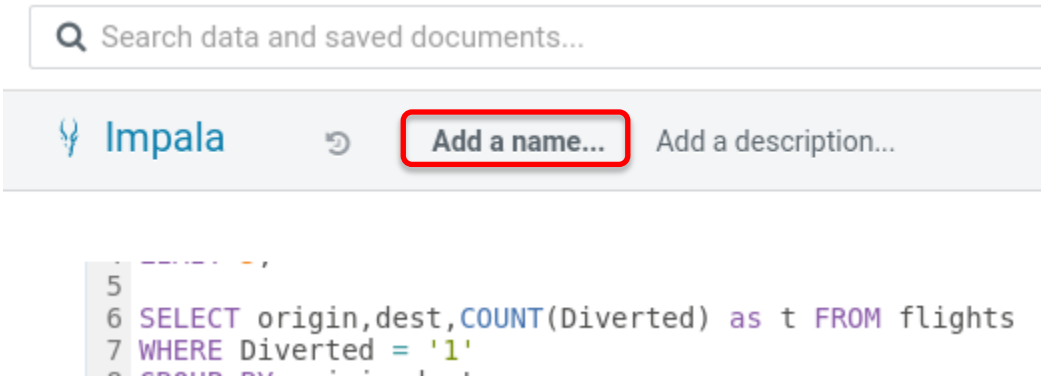

25) Look at the query plan – notice we now have a join in the tree



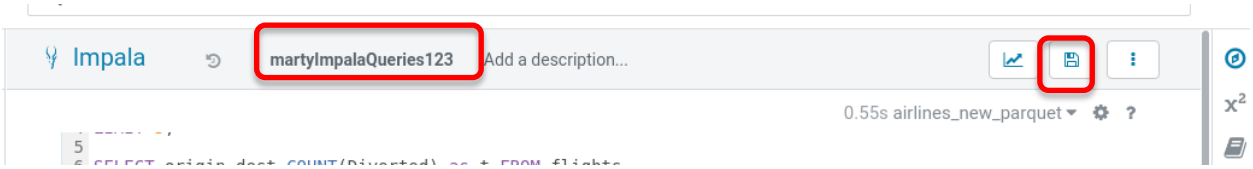

26) The summary shows details of all the stages in the join and their metrics

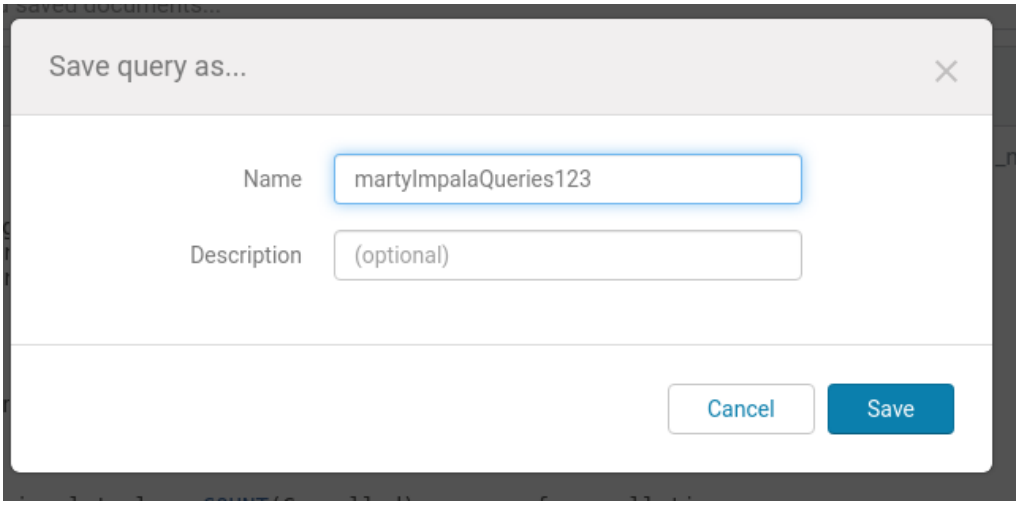



27) Time to save our work. Give your Impala SQL a name. For the lab use yourNameXXX where XXX is where you get to be creative. Use something unique, this is a multi user environment.




Then click “Save” and then “Save” in the popup






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

# 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

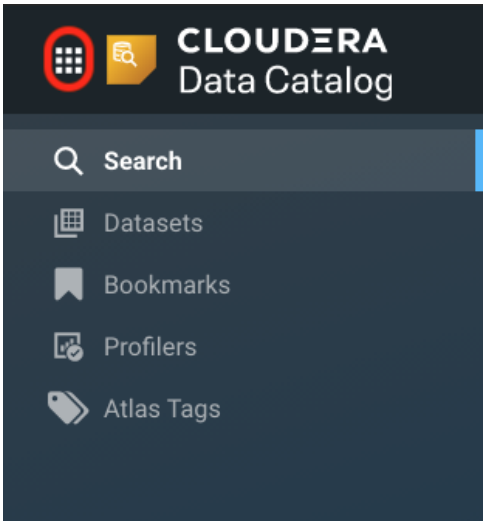
1) Open CDP, using the “admin” user within the Test Drive link.

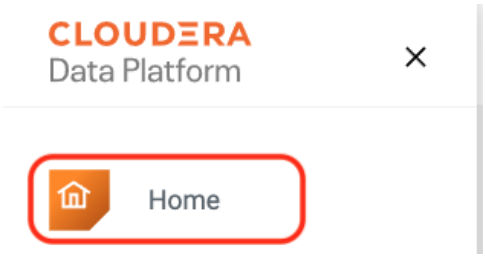
Your link should look something like (remember click the link in your email not the link below)
http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X
*xx represents the trial user #
*X represents the password

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

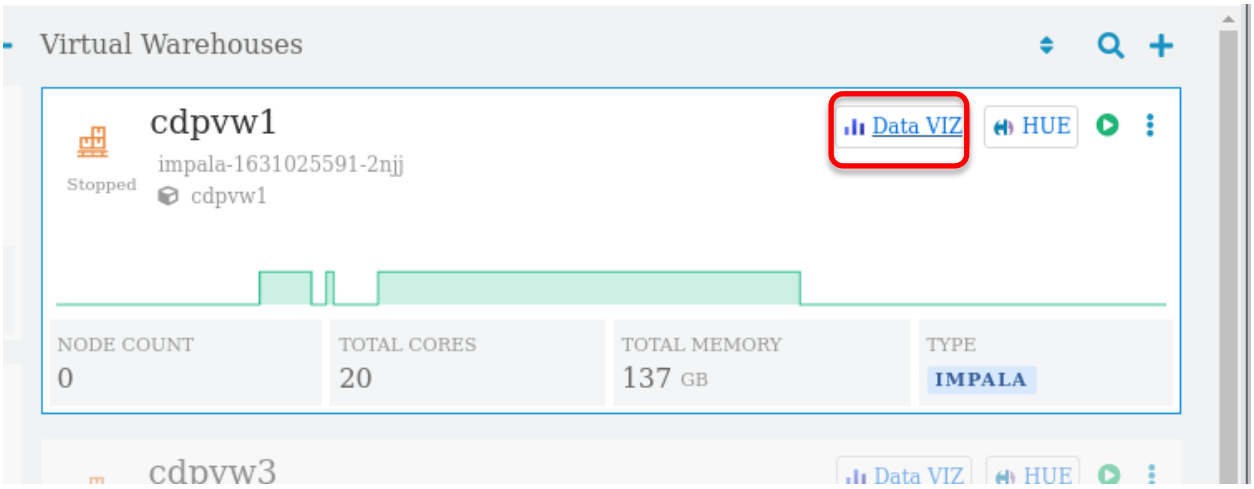

How do you get to the CDP Home Screen?

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



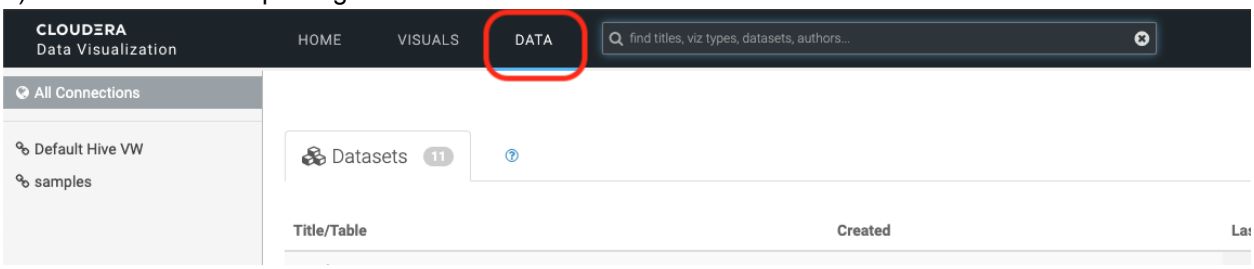



3) Click “Data VIZ” on your existing “Running” Virtual Warehouse.

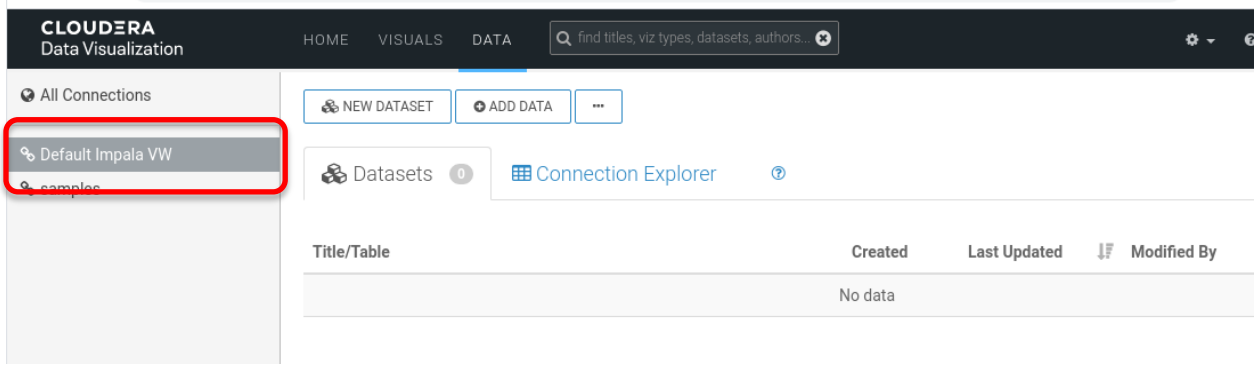



4) You are autologged into Data VIZ

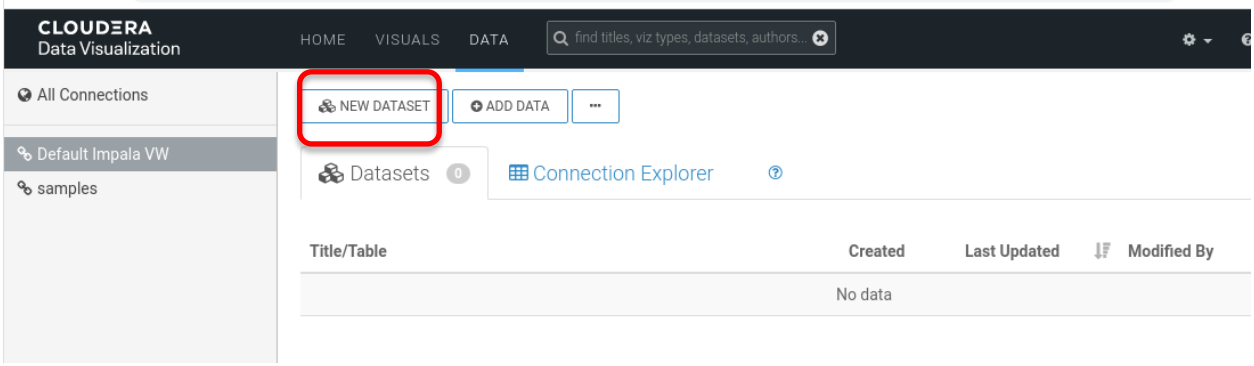
5) Click “DATA” the top navigation bar




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

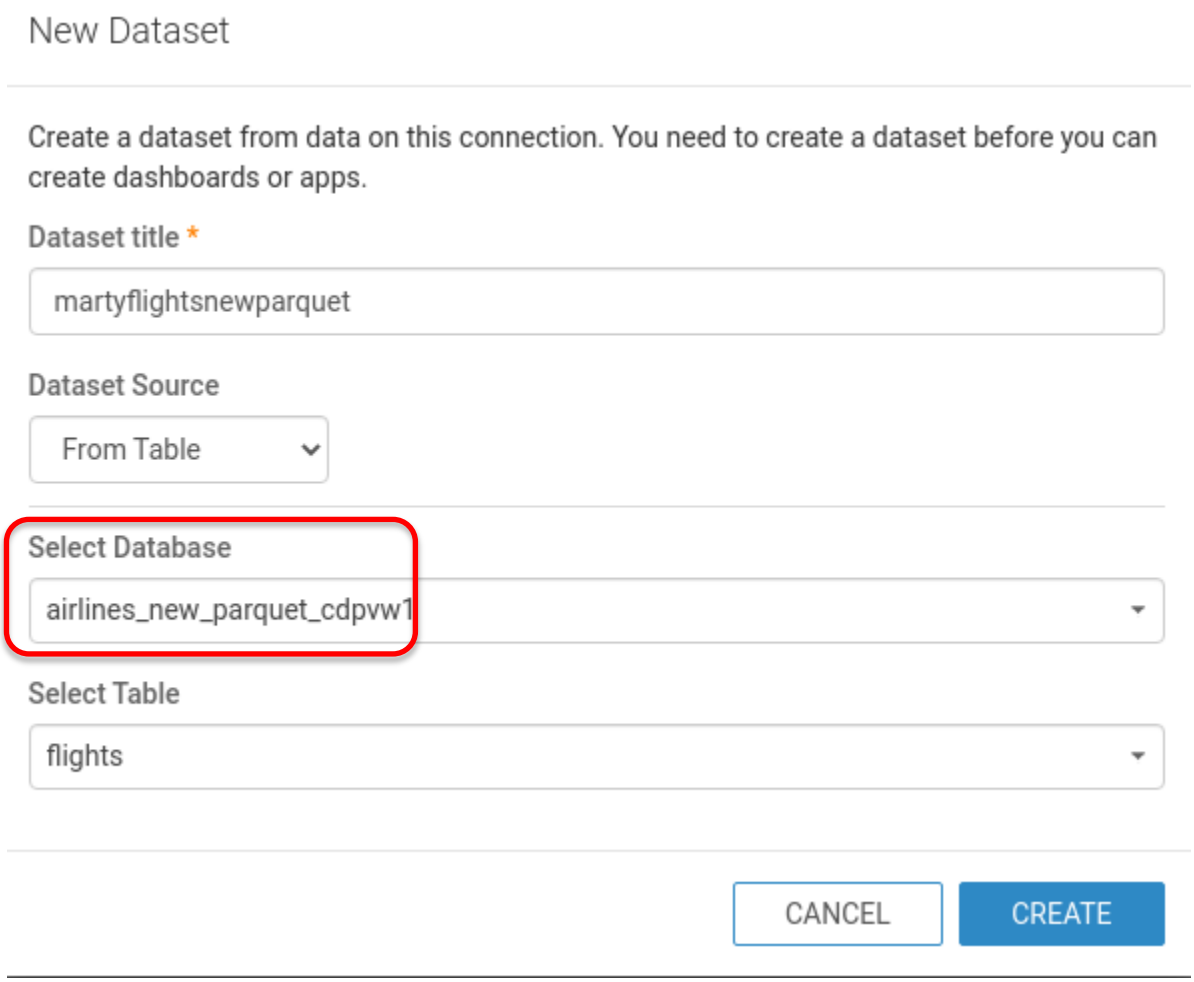
airlines_new_parquet_cdpvw1

Select Table

flights

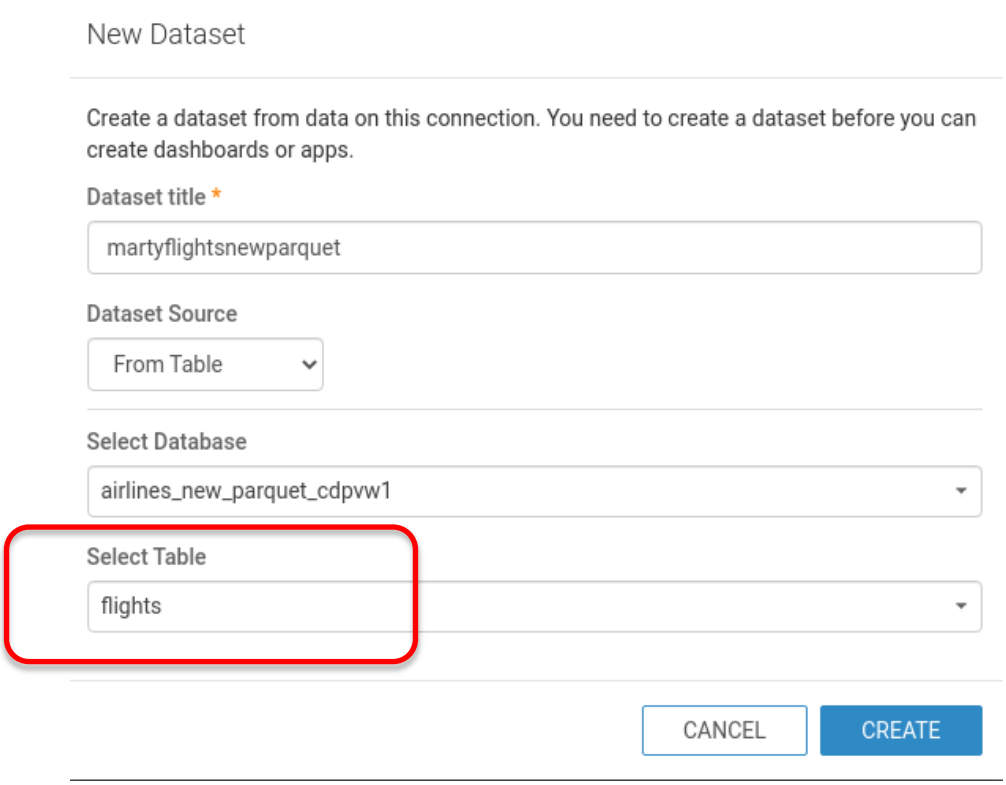
CANCEL CREATE

9) Choose the database “airlines_new_parquet_YOURVIRTUALWAREHOUSENUMBER”

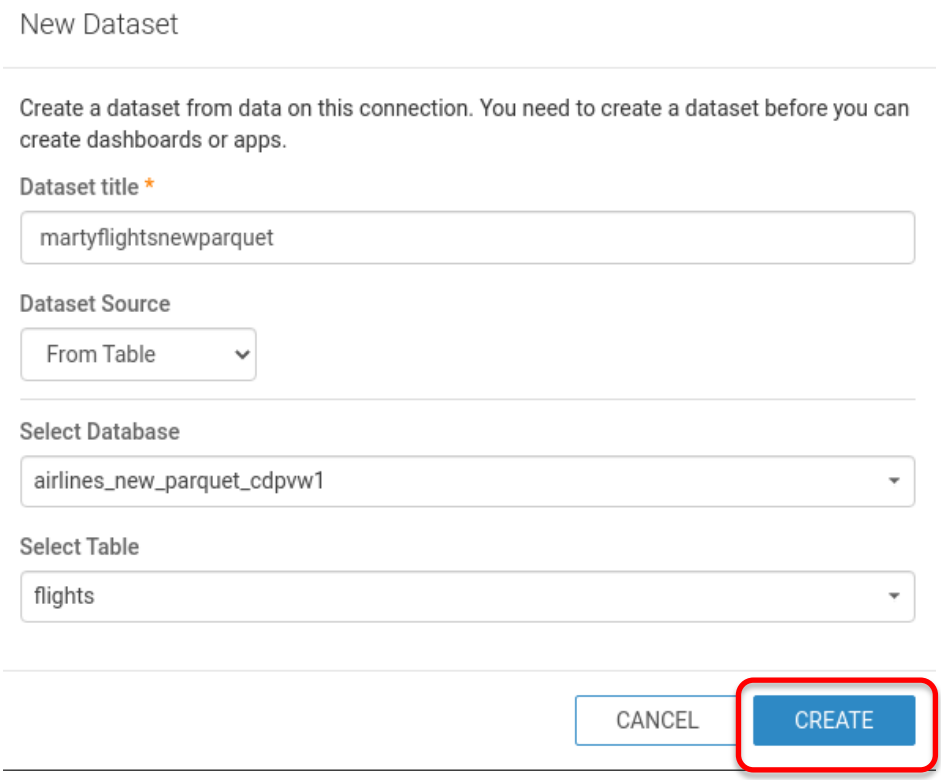



10) Choose the table “flights”

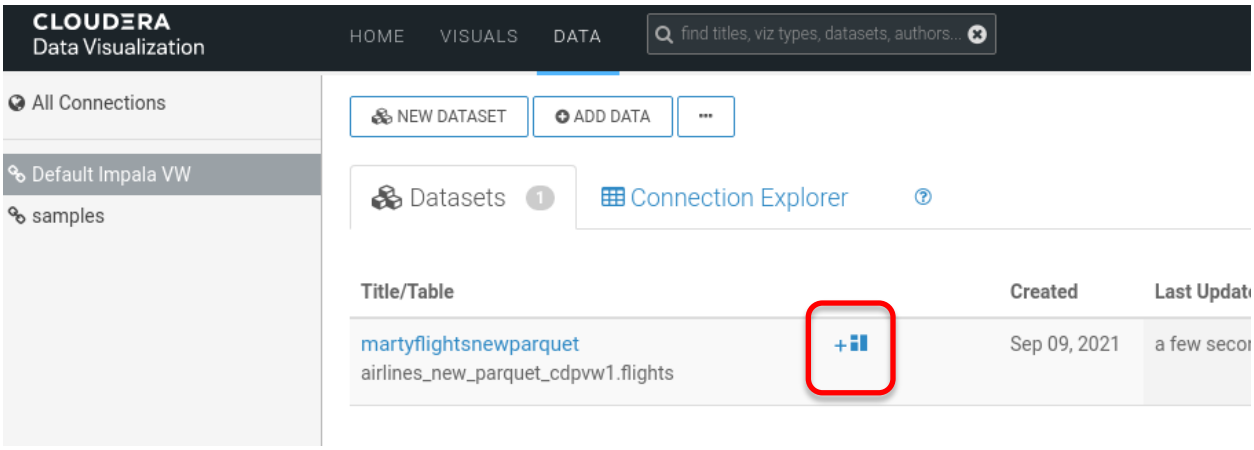
*Need to import multiple databases and tables? You’d use Dataset Source = SQL



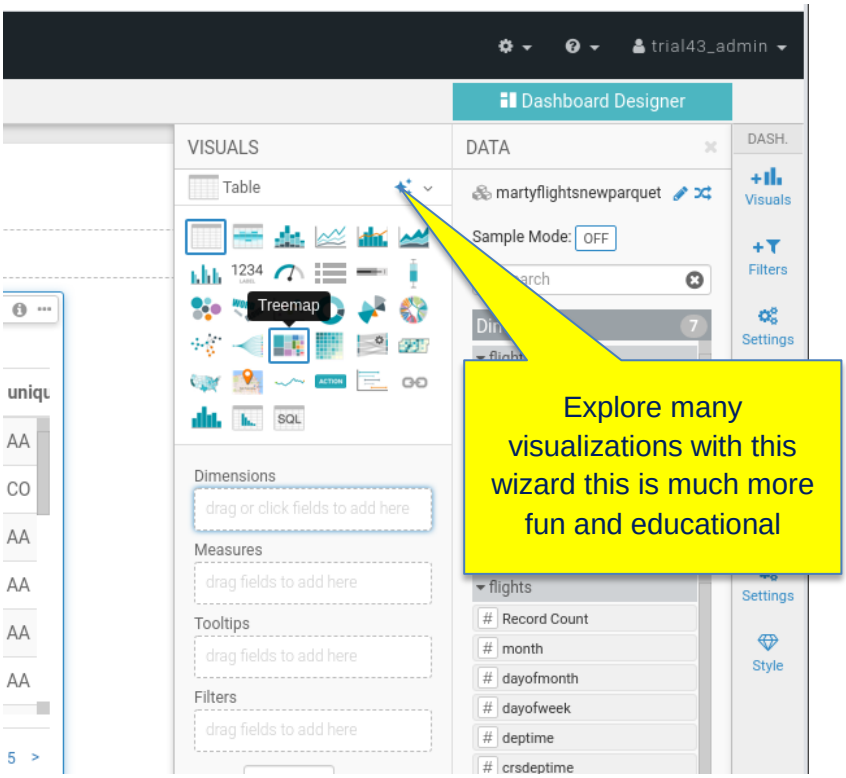

11) Click “CREATE”



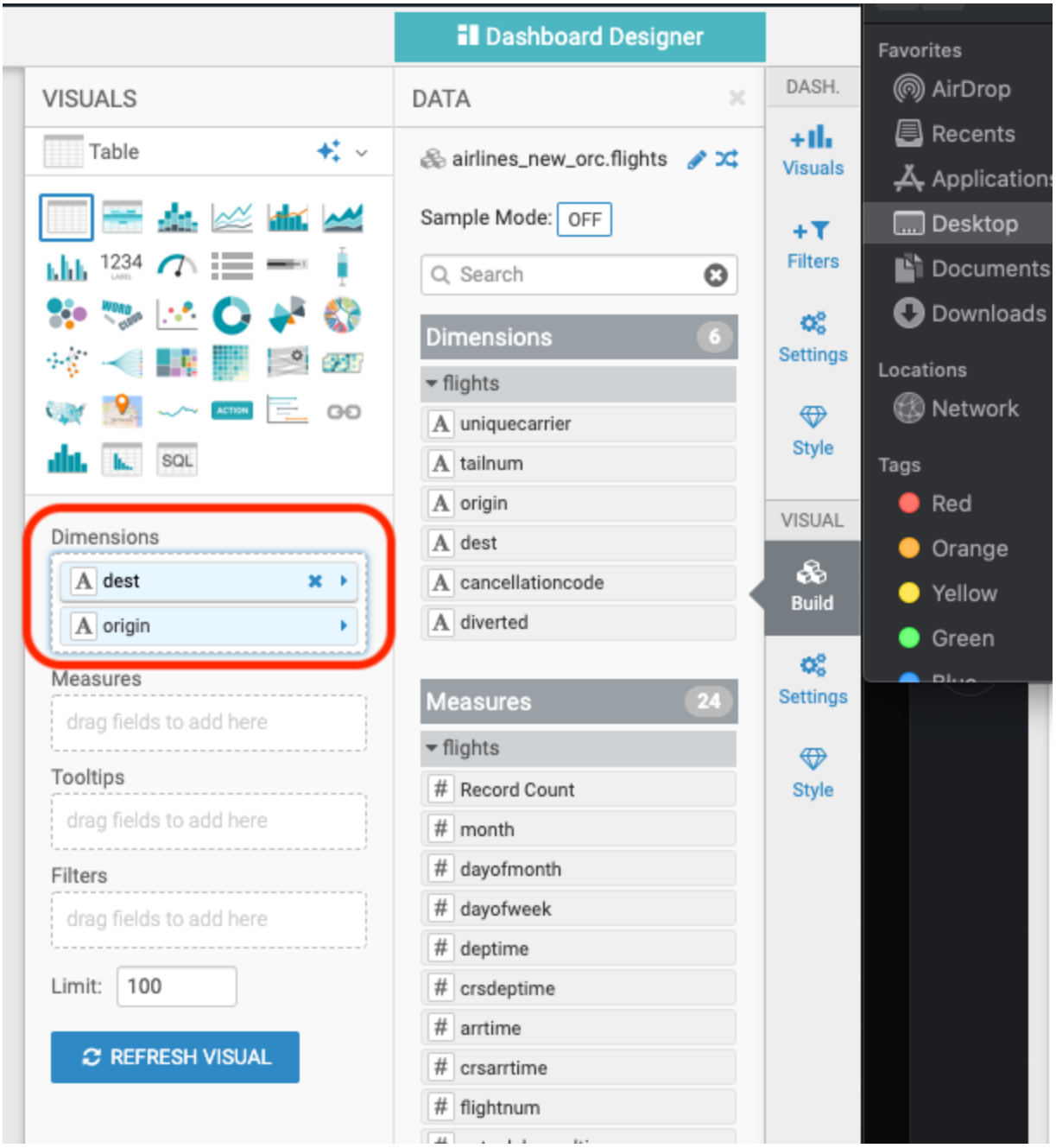
12) Click “+” to create a New Dashboard



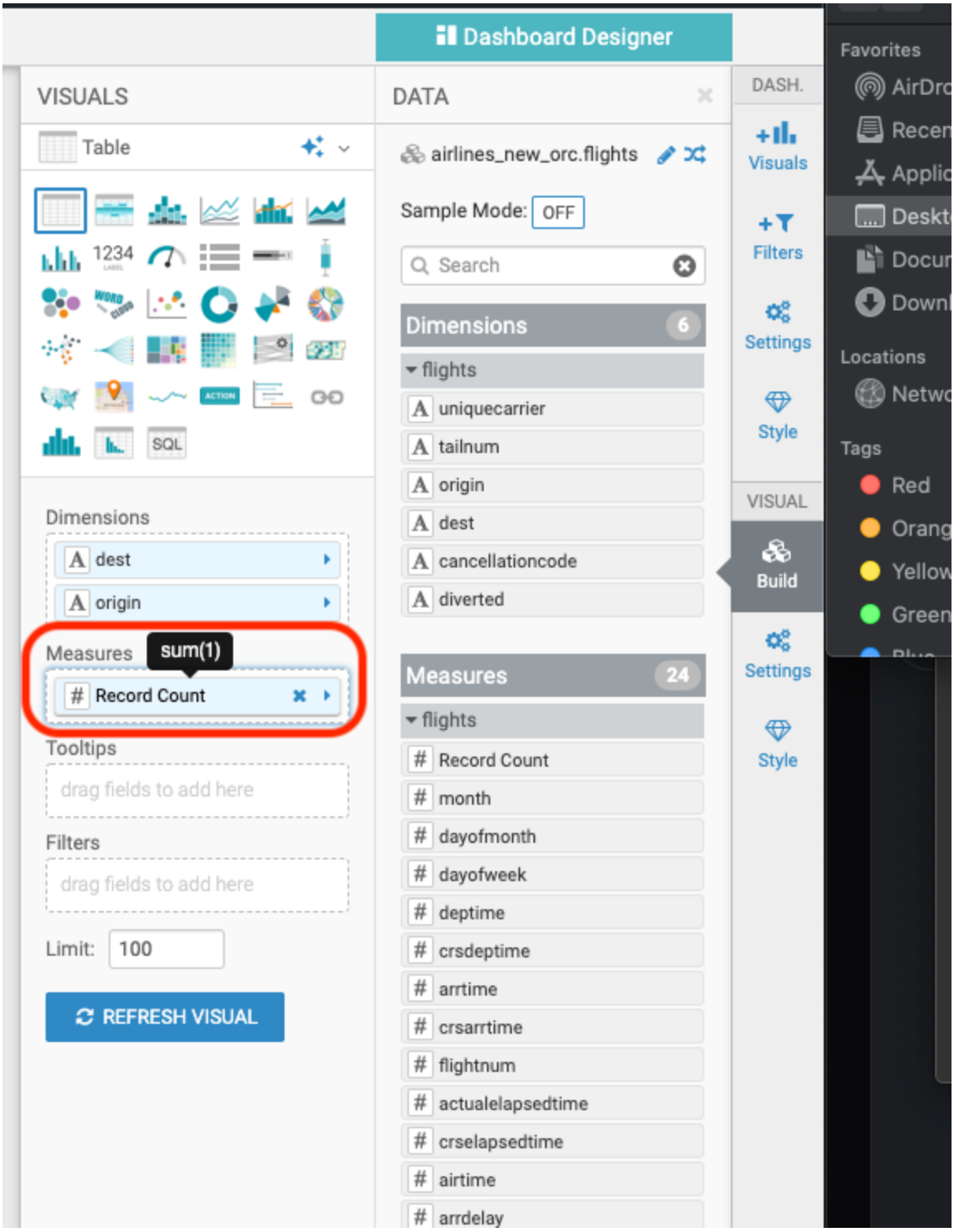

13) Choose “TreeMap” under “VISUALS” or explore with the Wizard



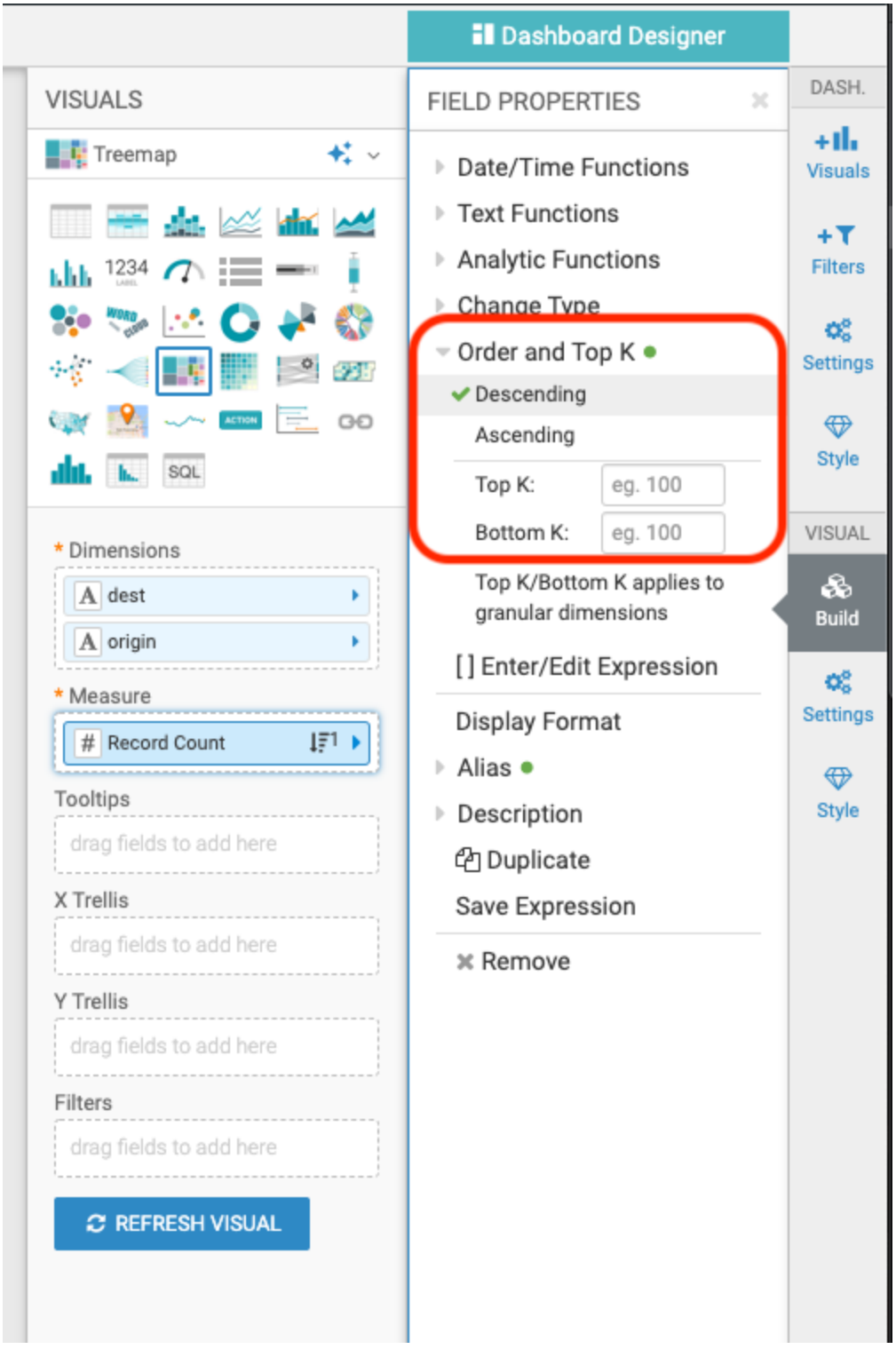

14) Drag-and-drop both “dest” and origin” from Dimensions->Flights into Dimensions under Visuals




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

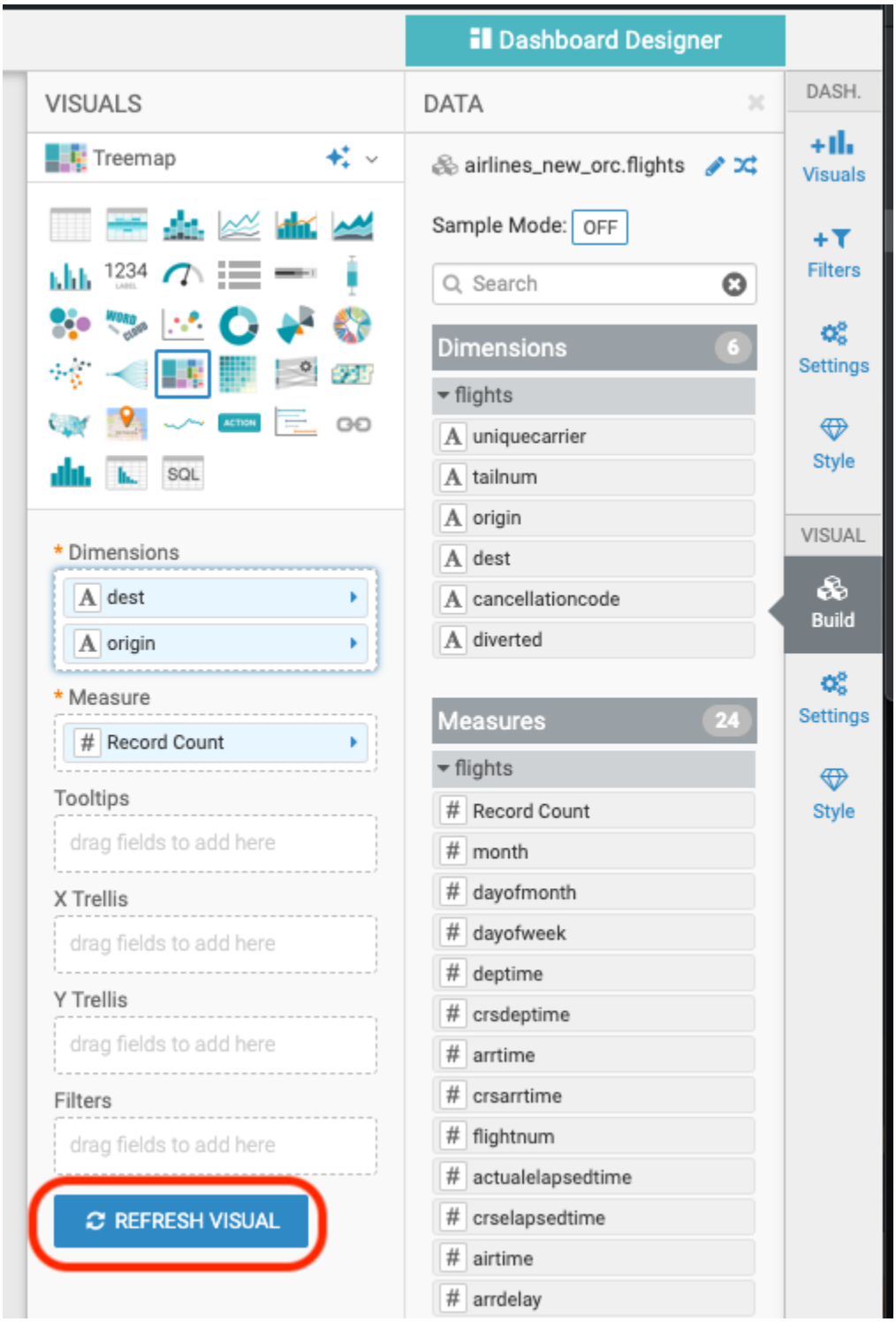



16) Click the right arrow next to Record Count and select “Descending” under Order and Top K

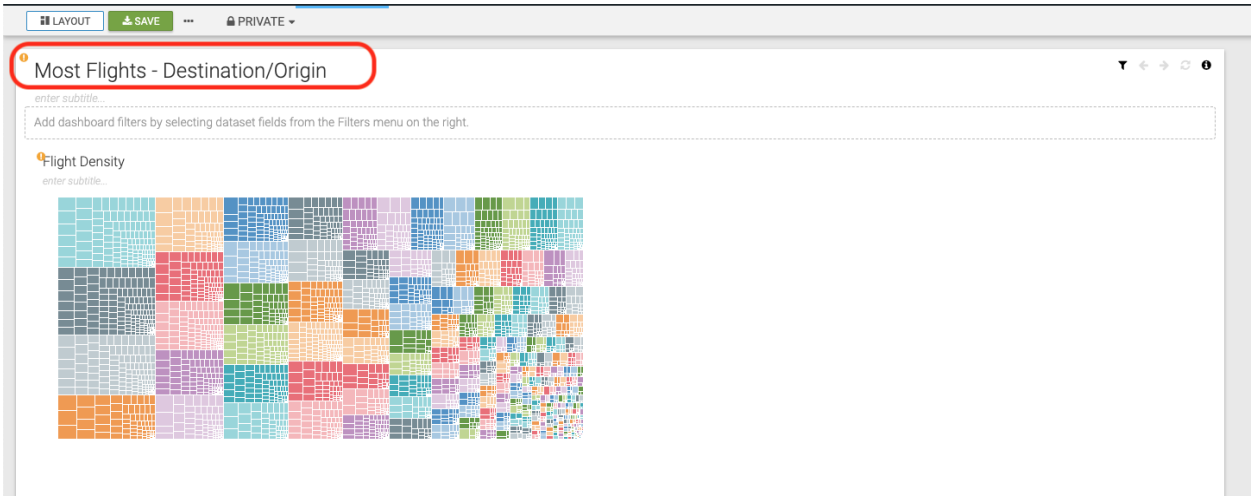



17) Click “REFRESH VISUAL”

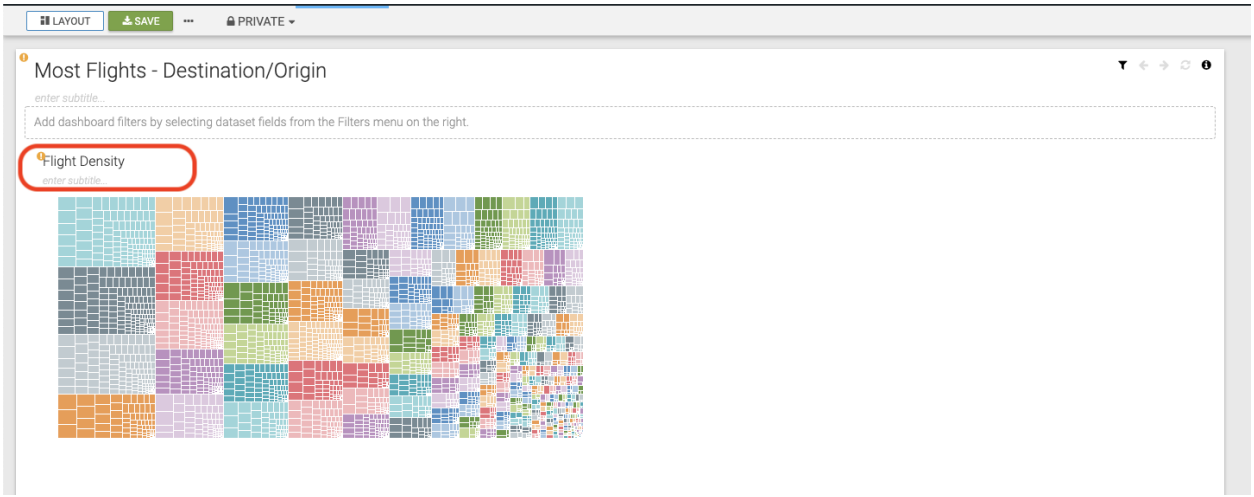
*Notice - you can have other Visuals chosen to be displayed with the Dimensions and Measure(s), then click REFRESH VISUALS



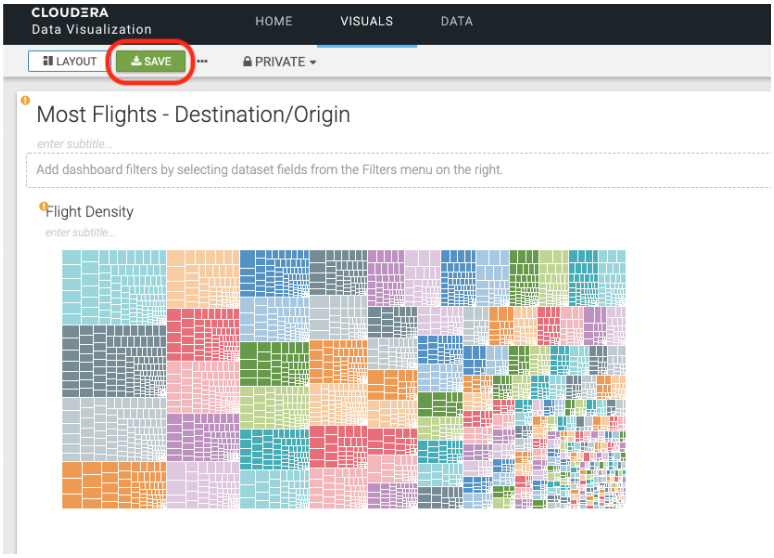
18) Enter a title “Most Flights - Destination/Origin”



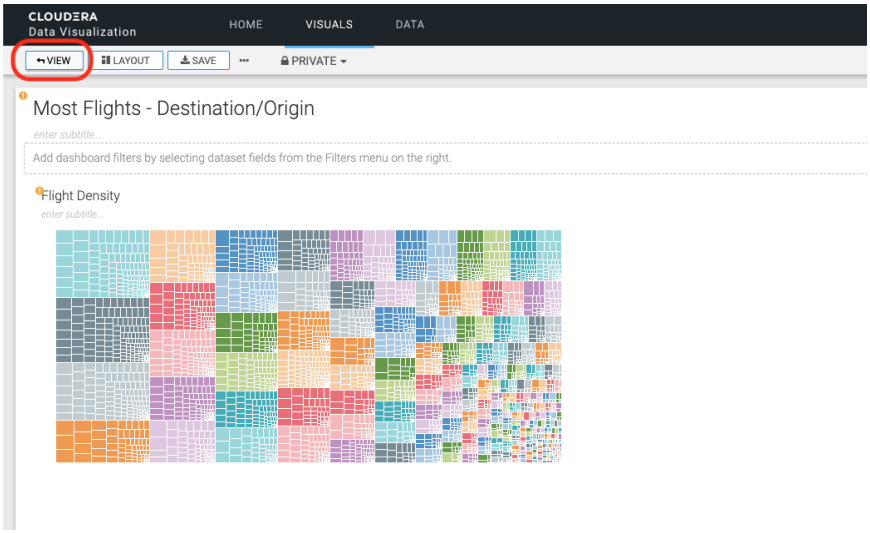

19) Enter “Flight Density” under the graph’s title



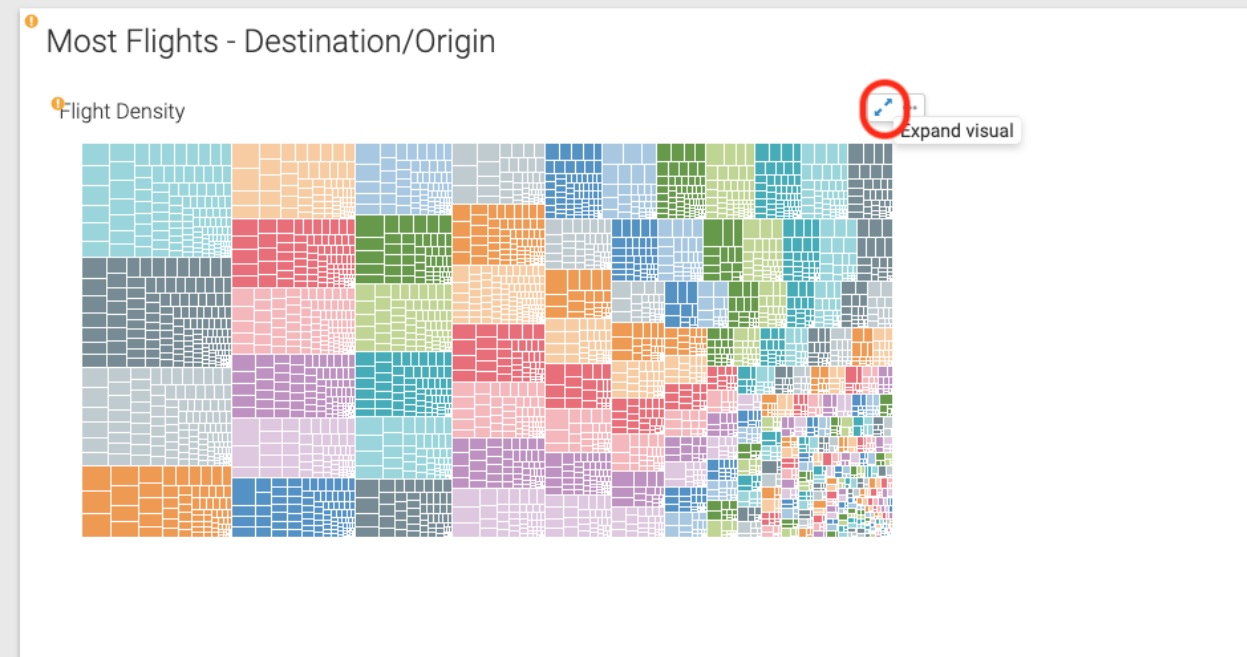

20) Click “SAVE”



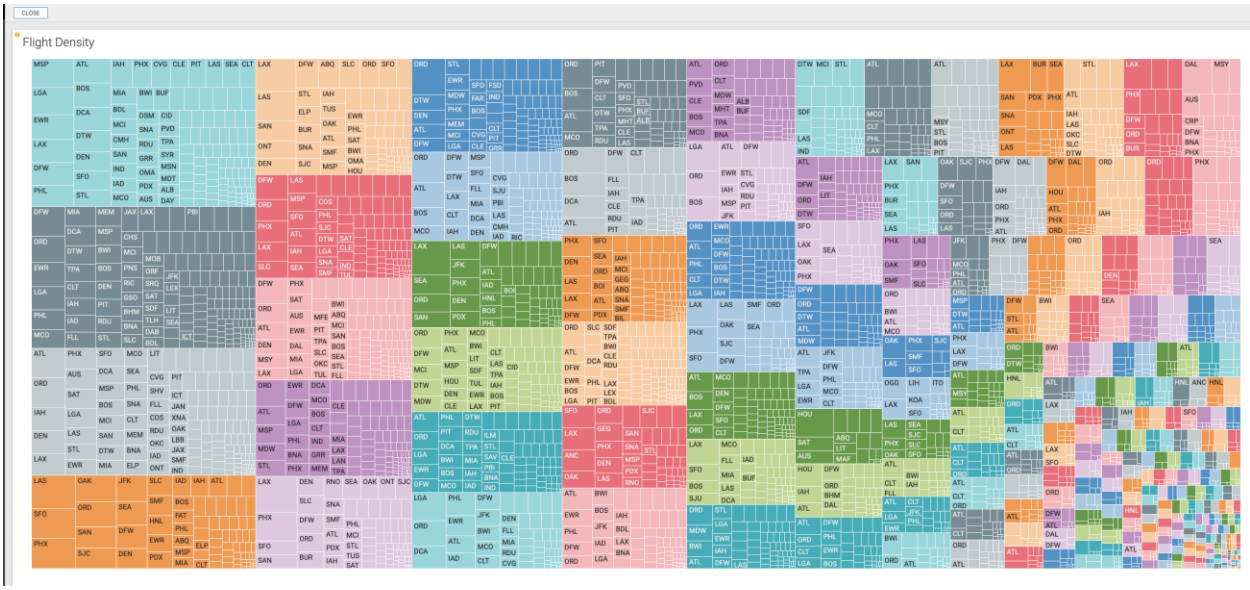

21) Click “VIEW”



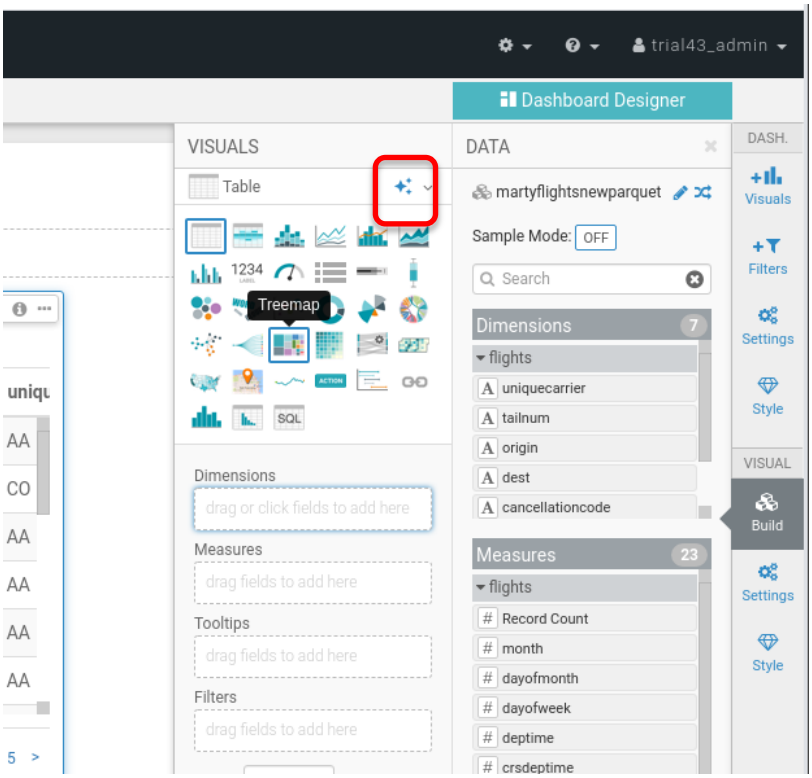

22) Scroll over the graph and click “Expand Visual”



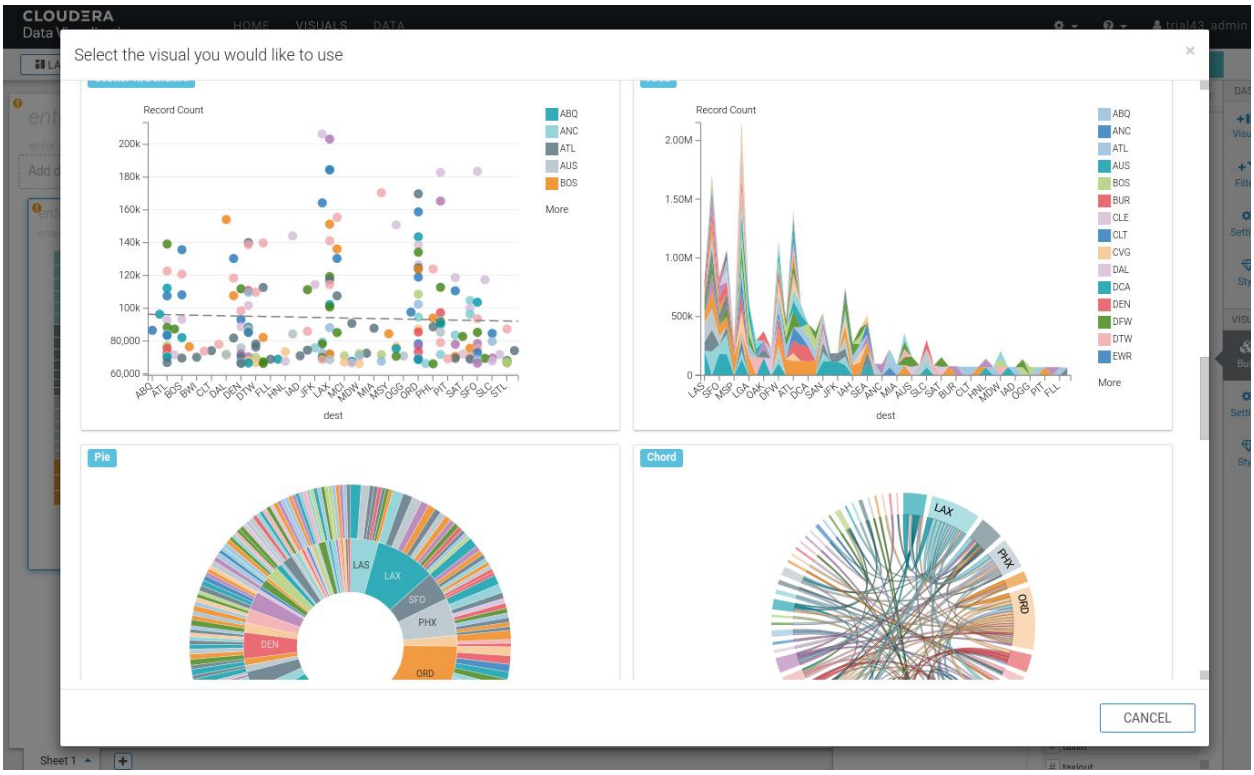

Destinations are displayed




23) Time to explore the other visualization options.




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



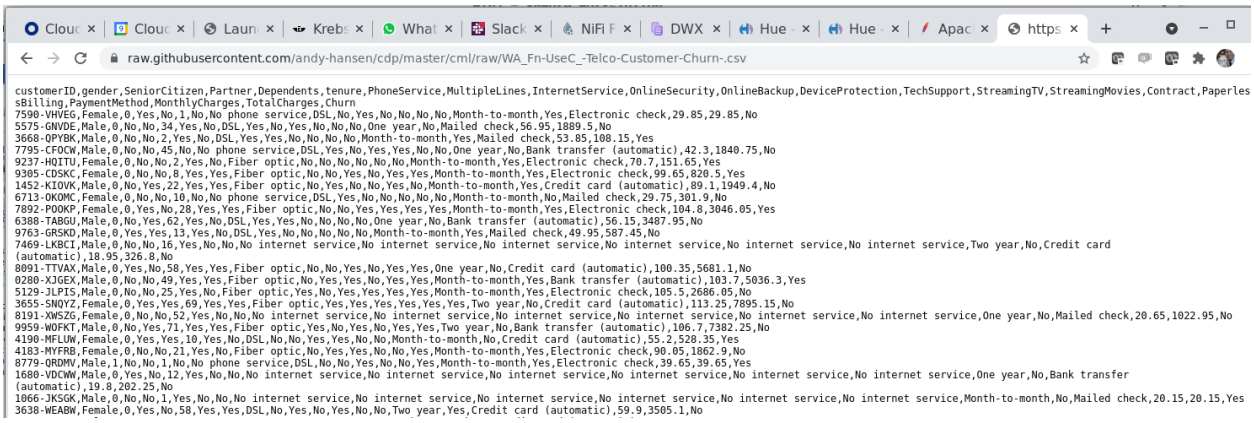
# Part 4 – How to create a table using data stored in S3

**Create table on S3 file [15 minutes]**

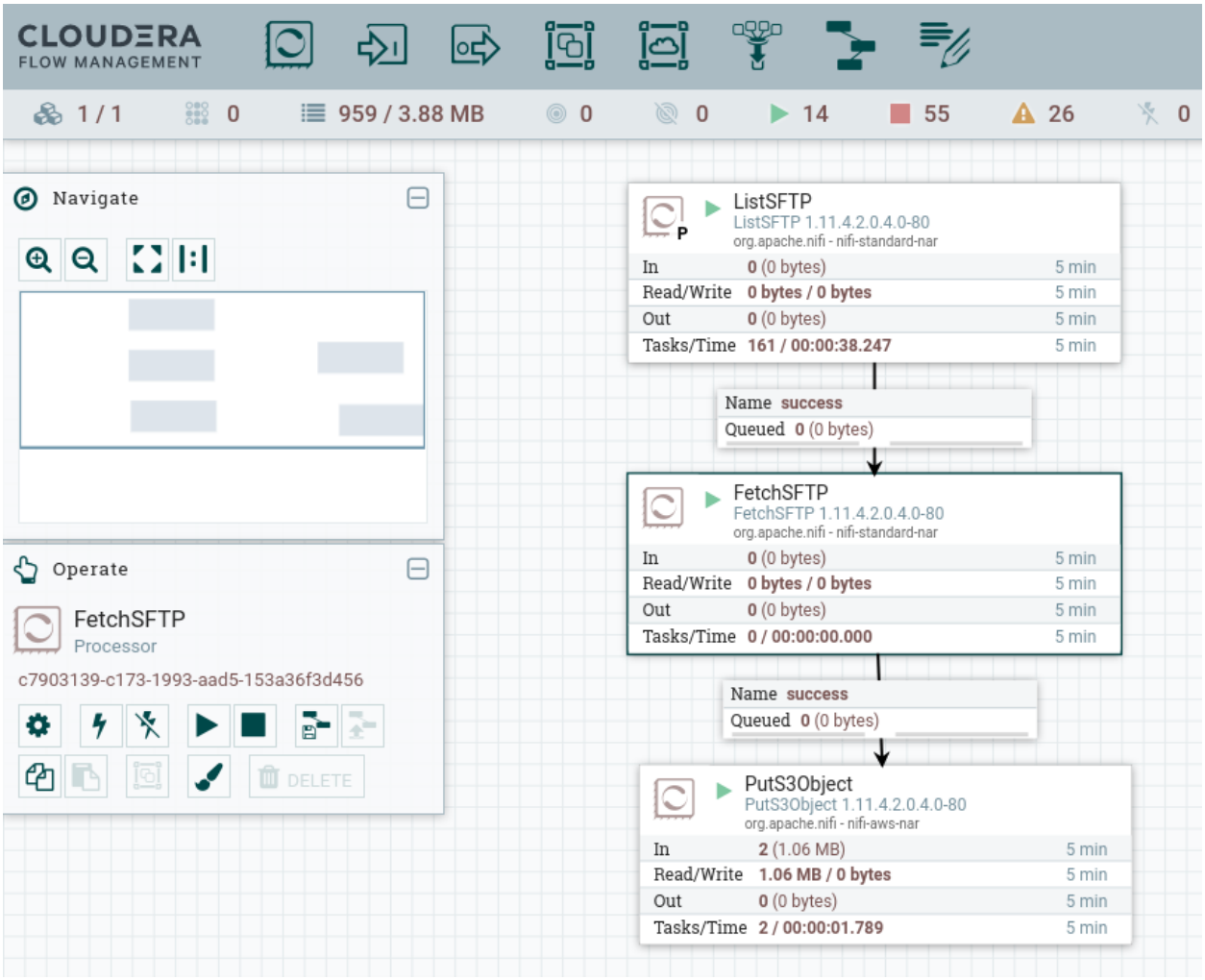
**Options:**
**Nifi**
**Command line**
**Create file in S3 bucket, run sql to create the table**

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




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
```

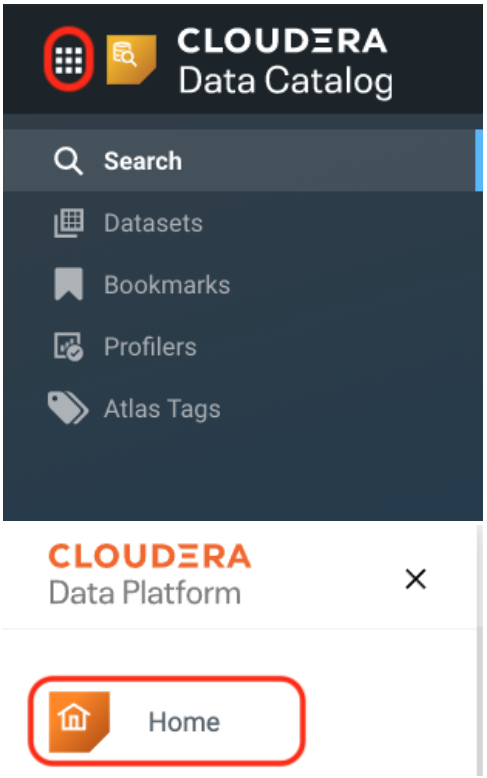
1) Open CDP, using the “admin” user within the Test Drive link.

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

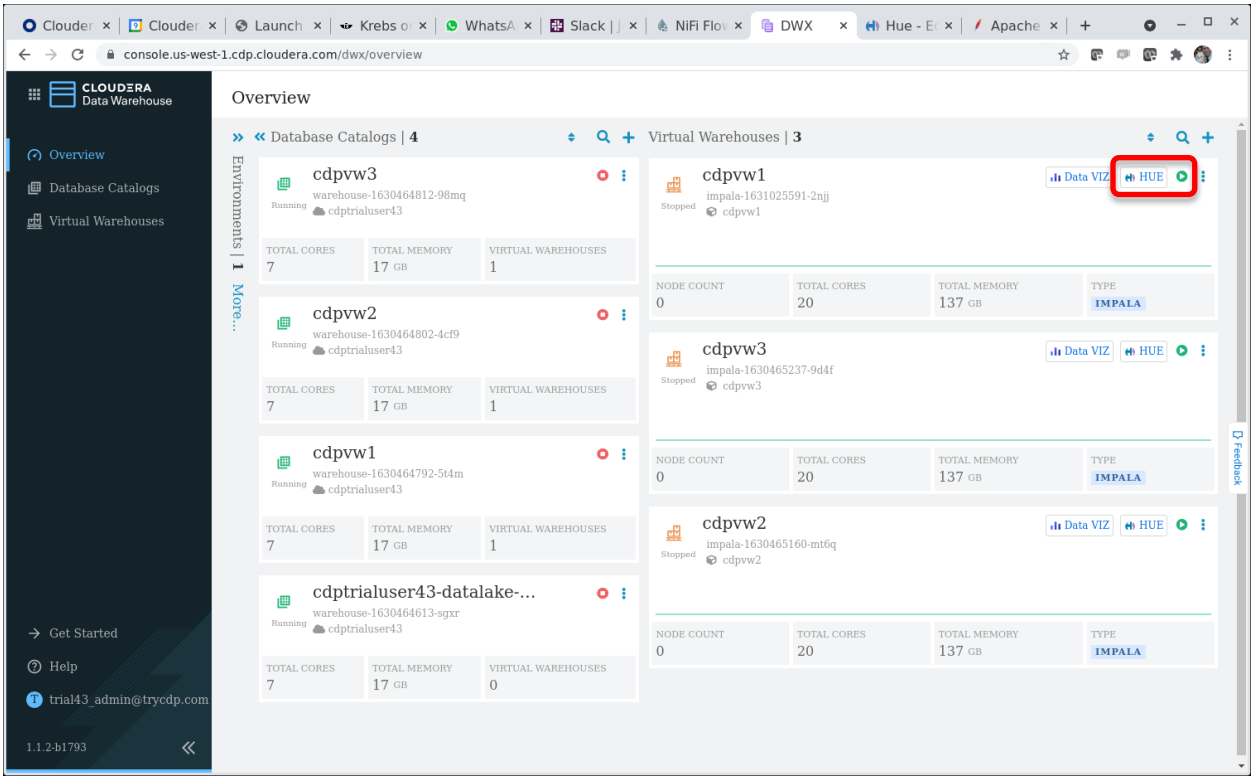
2) Click the “Data Warehouse” within the CDP Home Screen


How do you get to the CDP Home Screen?

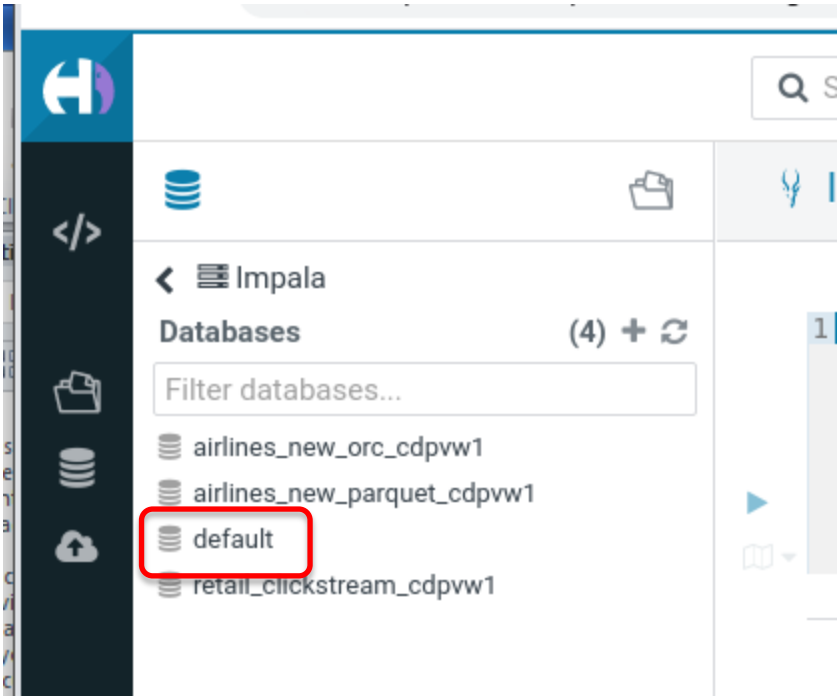
- From any experience such as “Data Catalog”, click the 9 square at the top left and then click “Home”



3) Open HUE, be sure you're going the the virtual warehouse
*The same steps you did in Part 2 to Open HUE




4) Navigate to the "Default" database


Impala
Databases
(4)
Filter databases...
airlines_new_orc_cdpvw1
airlines_new_parquet_cdpvw1
default
retail_clickstream_cdpvw1

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%";
```

## Auto-Scaling Demo

As time permits we will show you how Cloudera Data Warehouse will add nodes to the Kubernetes cluster to meet additional workload. The test harness is Apache Jmeter. The image below shows scaling from zero nodes to two to four and then back down to two nodes running AWS EKS.