CLOUDERA DATA PLATFORM DATA WAREHOUSE LAB – Experimental, you are welcome to try this lab or use the prior version

Step-by-step instructions:

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”.

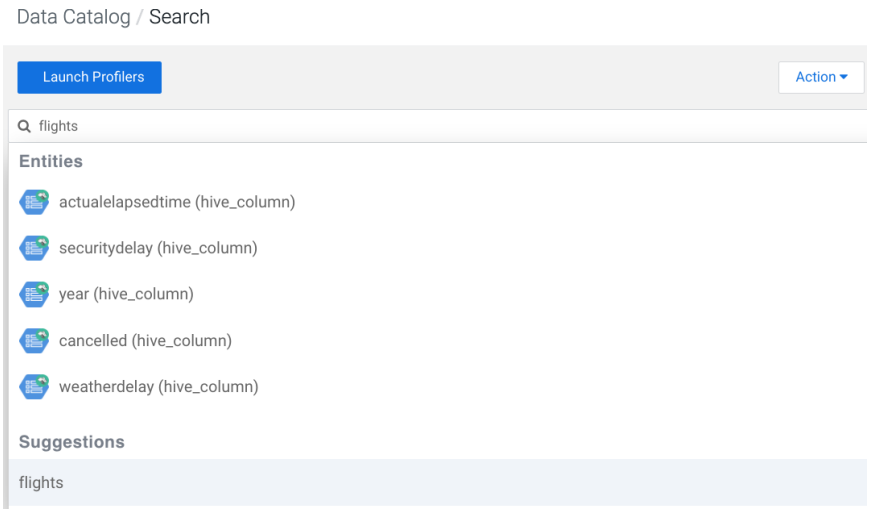
- Find what database(s) the table “flights” is located.
- Find out at least one year that the “flights” table was generated from.
- Find out how many columns the table “flights” contains.

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

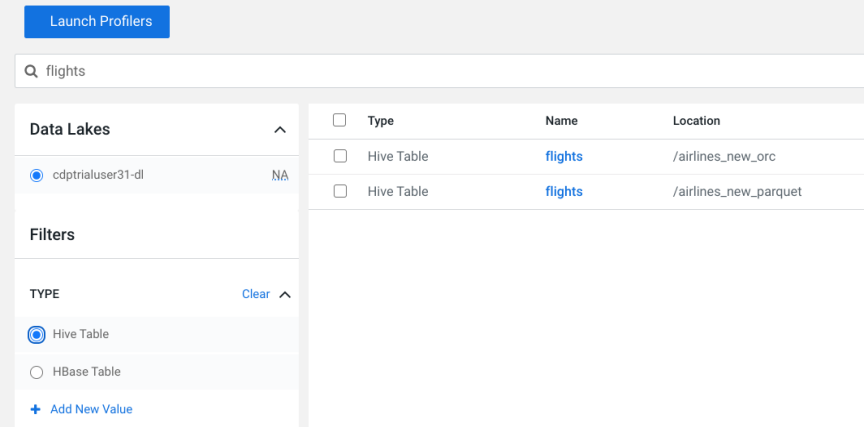
Your link should look something like (remember click the link in your email not the link below)
http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X
*xx represents the trial user #
*X represents the password

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

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

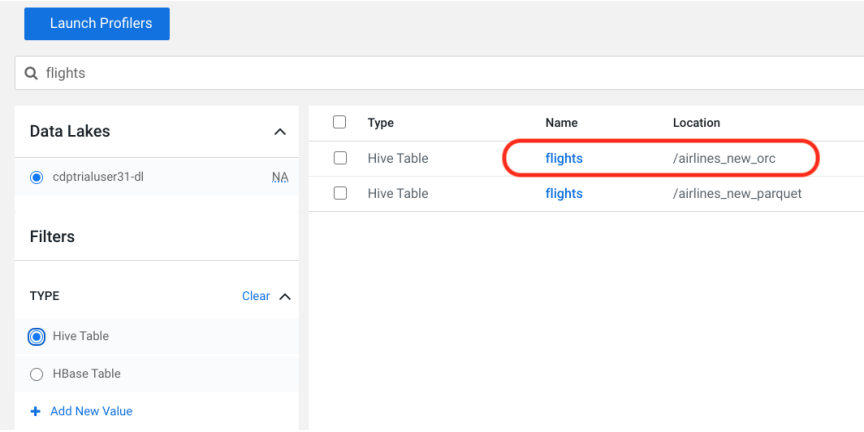


4) Click “Hive Table” under Filters on the left

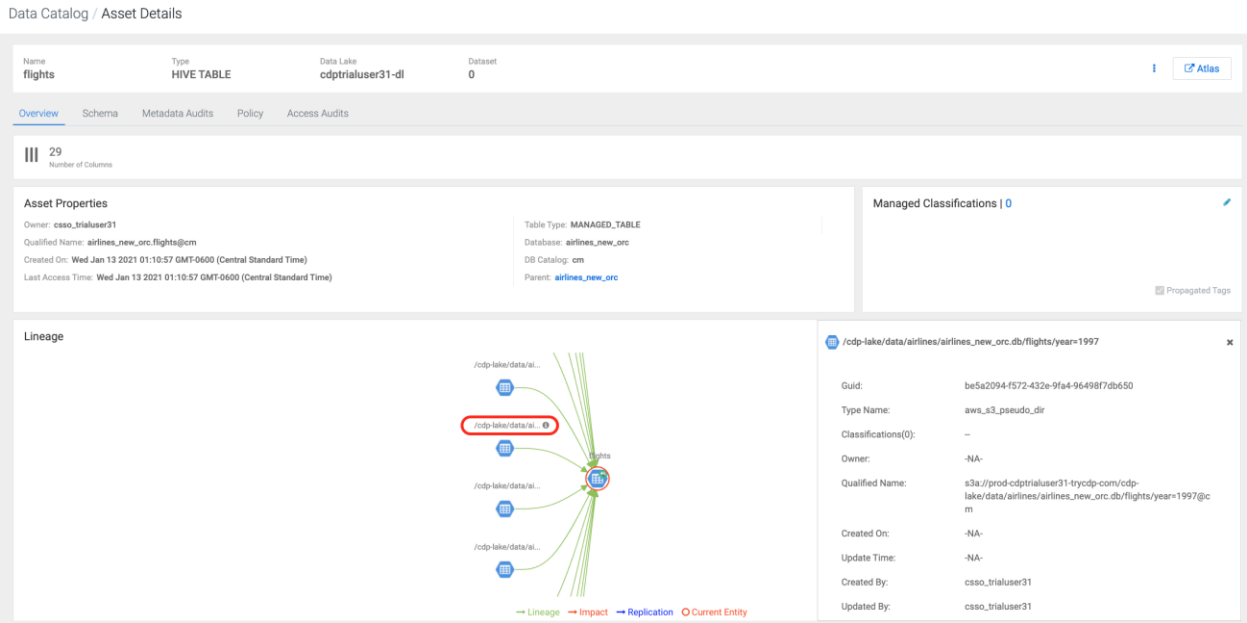


*Find what database(s) the table “flights” is located.

5) Click “flights” where the Location = /airlines_new_orc



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?

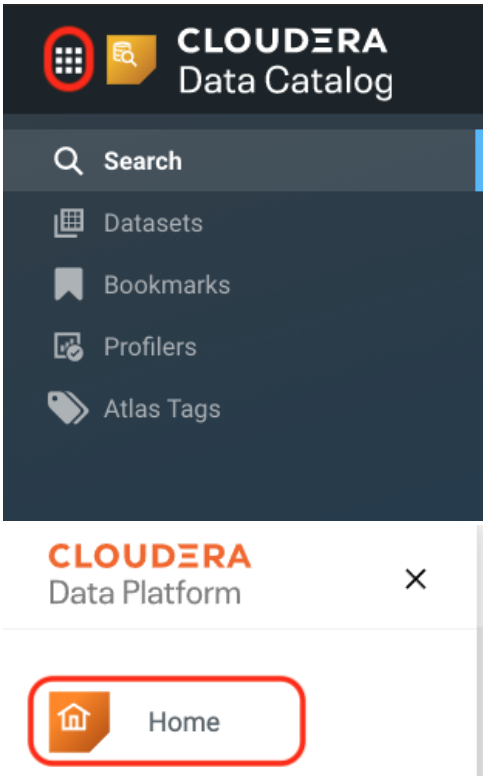
1) Open CDP, using the “admin” user within the Test Drive link.

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

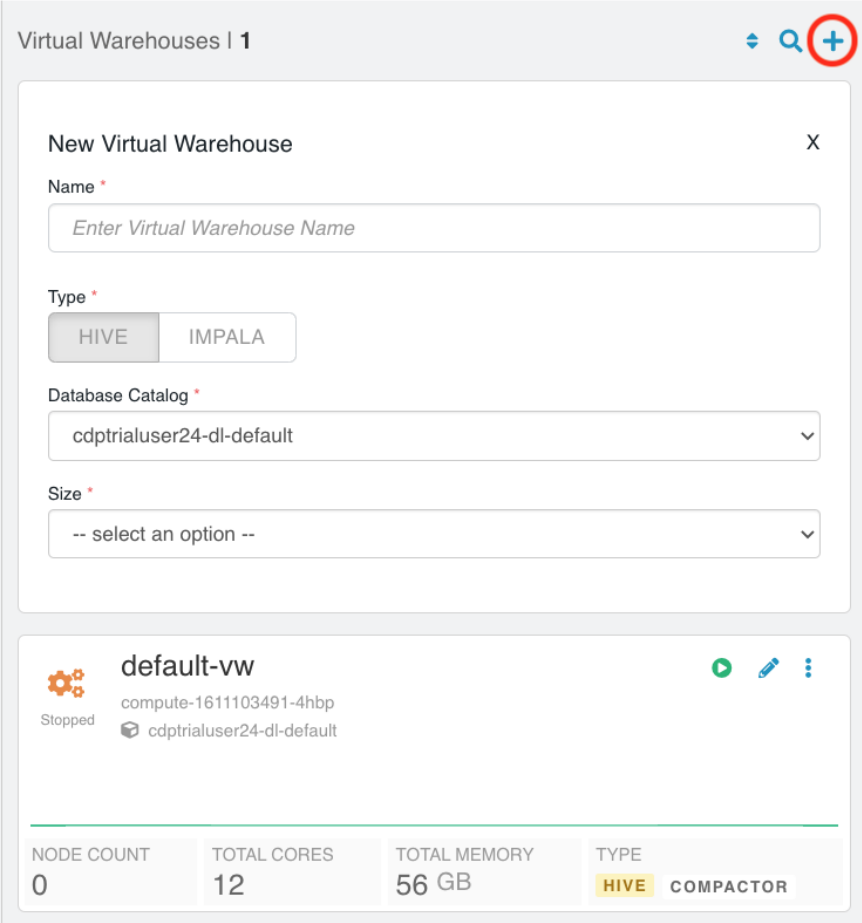
How do you get to the CDP Home Screen?

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

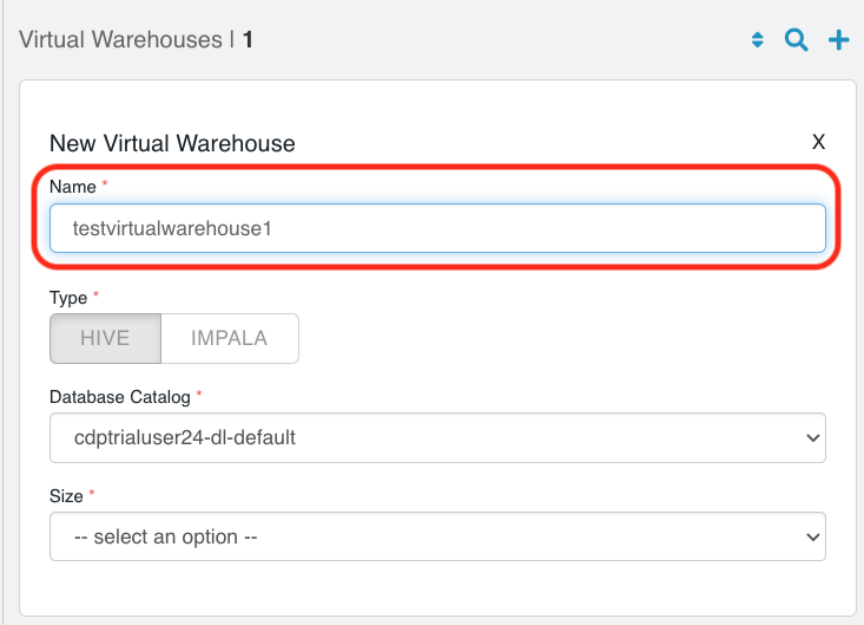


WE HAVE DONE THIS FOR YOU – DO NOT CREATE A NEW VIRUTAL WAREHOUSE – READ THROUGH THIS FOR BACKGROUND INFO…

3) Click the “+” at the top right next to “Virtual Warehouses”

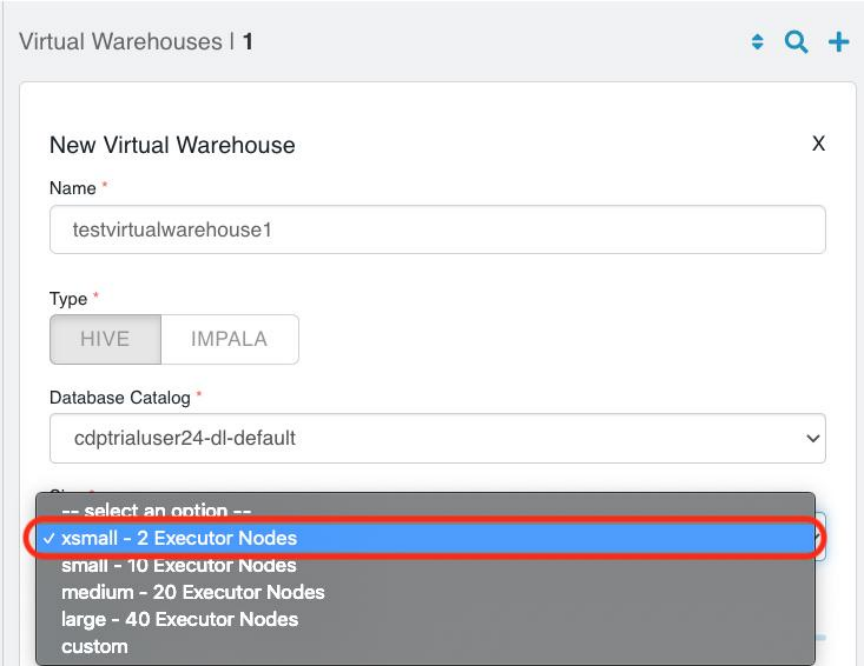


4) Enter a name for your New Virtual Warehouse



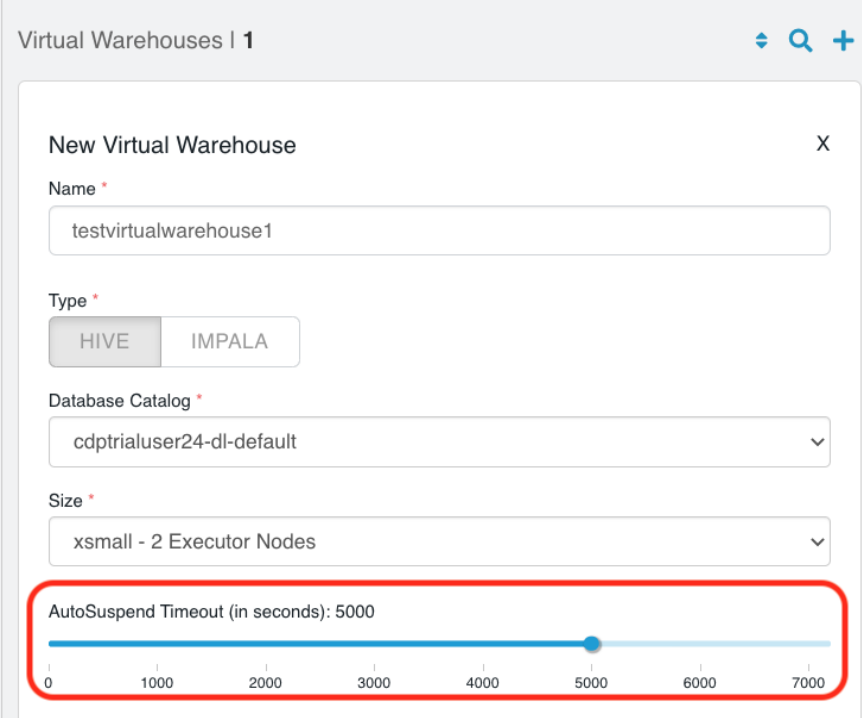
5) Select the Size of “xsmall - 2 Executor Nodes”

*How do I choose a size? Initial concurrent users



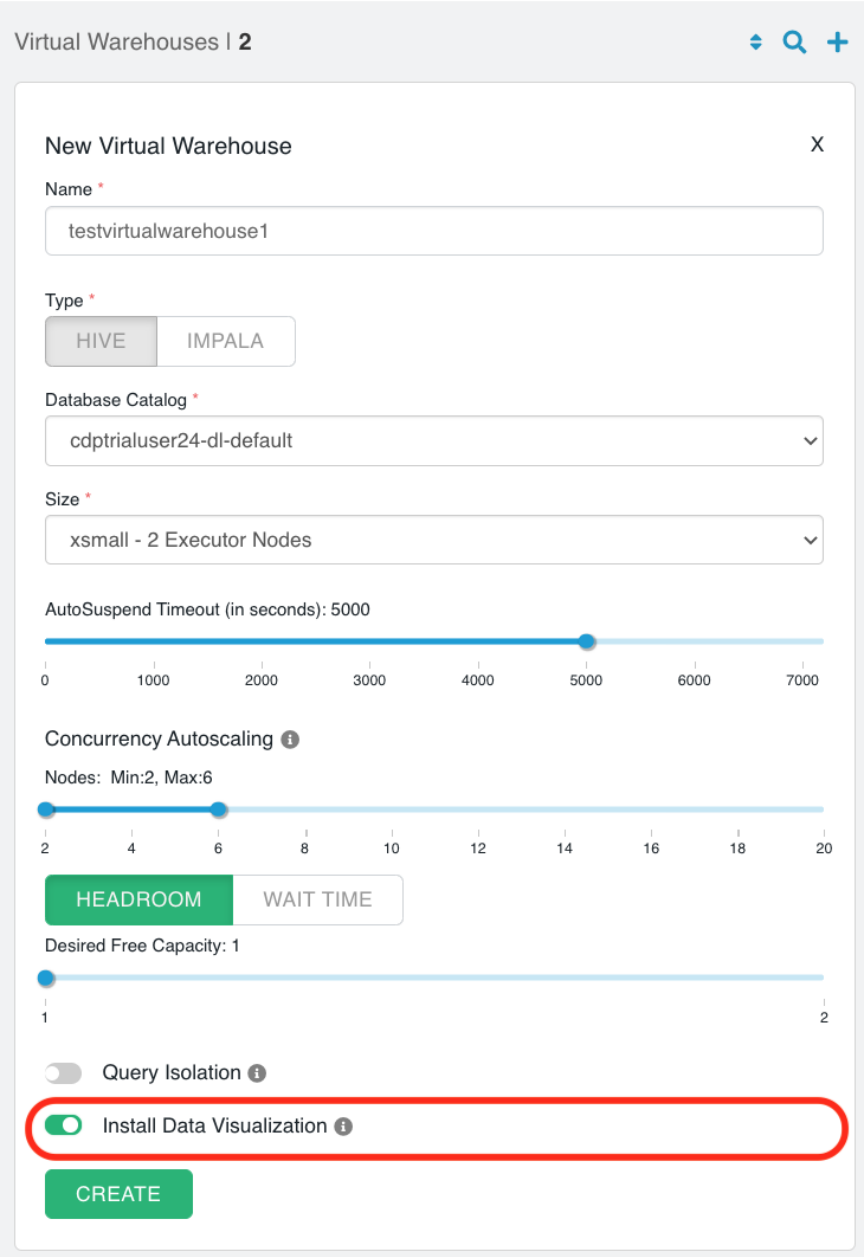
6) Set the AutoSuspend Timeout (in seconds) between 4500 and 5500:

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



7) Choose “Install Data Visualization” to be on

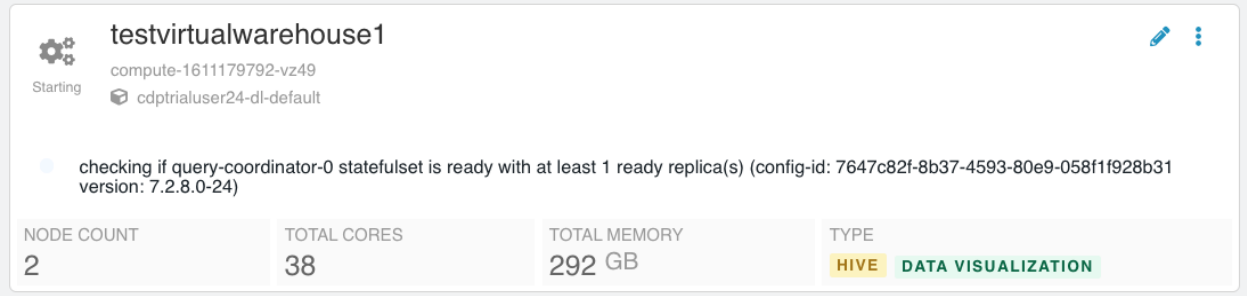
*Allowing for Data Visualizations in Part 3

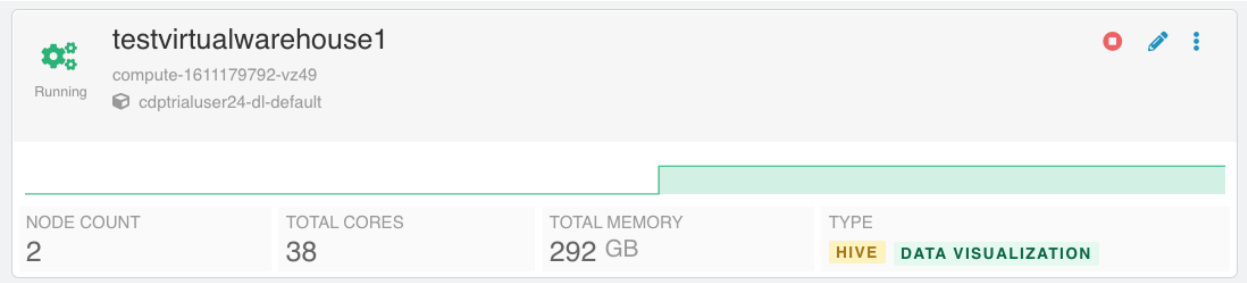


8) Click “Create” to create your Virtual Warehouse

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

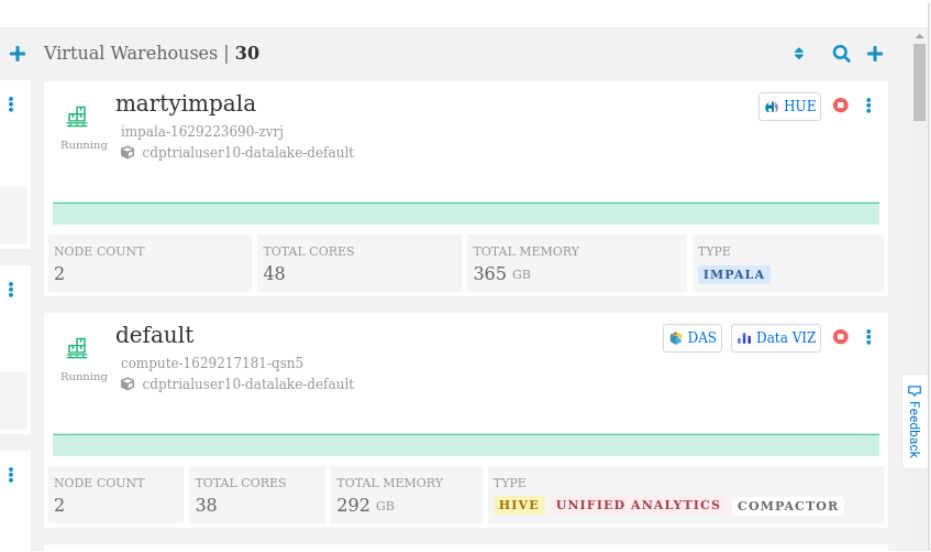
When available for use, “Starting” will change to “Running” as shown below



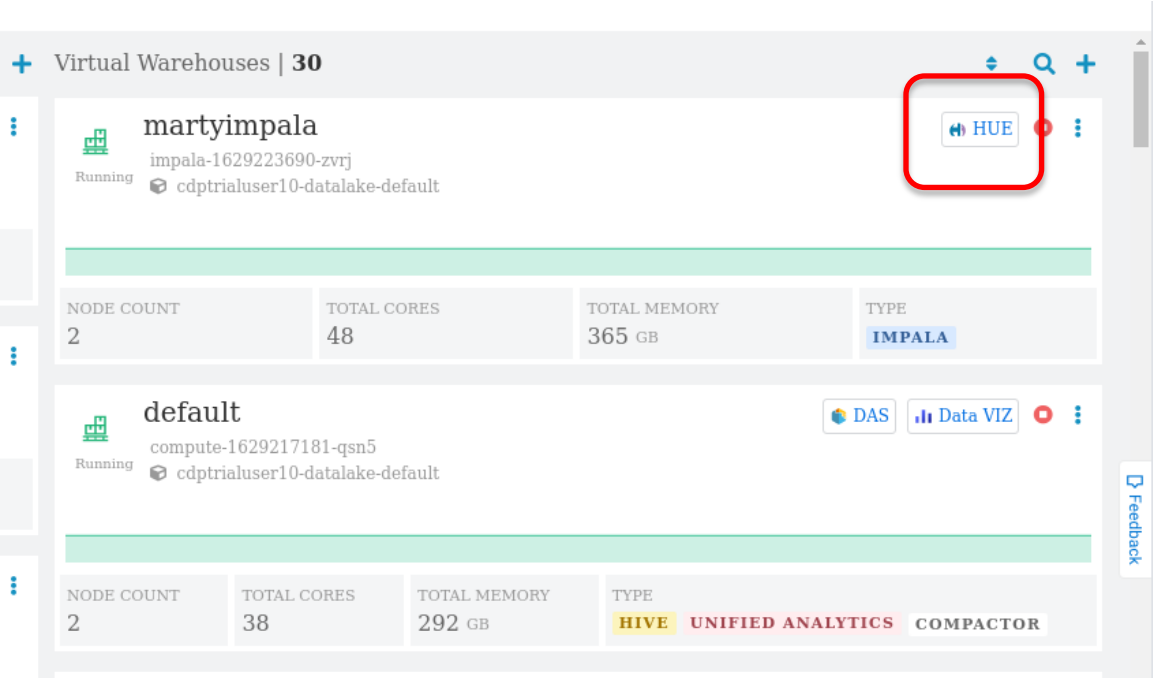


OK, TIME FOR YOU TO JUMP BACK IN. THE DATA ANALYTICS STUDIO (DAS) IS DEPRICATED. IT IS REPLACED WITH HUE. THE INSTRUCTIONS BELOW SHOW YOU HUE WITH IMPALA. IF YOU WANT TO USE DAS AND HIVE THOSE INSTRUCTIONS ARE AFTER THE IMPALA LABS

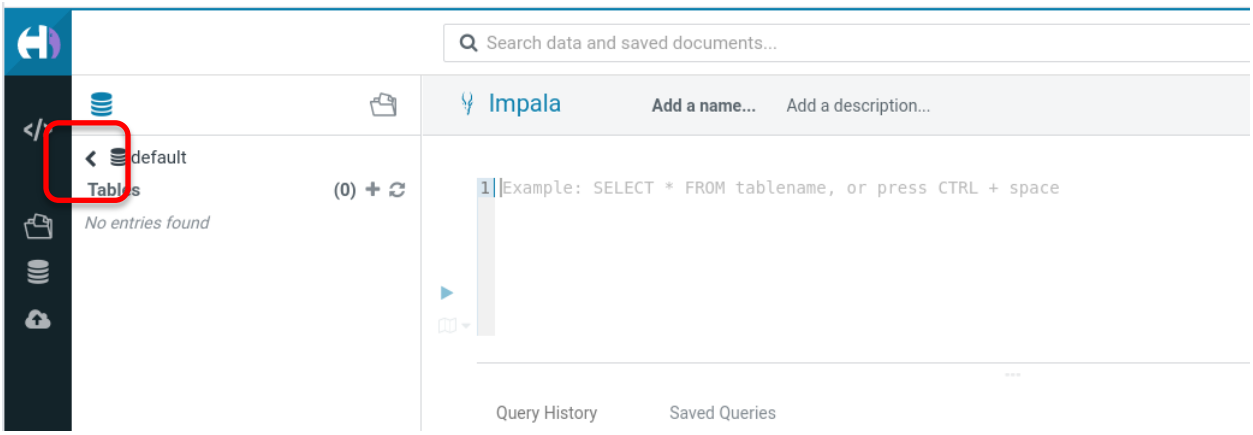
9) Notice there are two virtual warehouses built on the cdptrial data catalog. Only the “default” virtual warehouse is enabled for Data VIZ



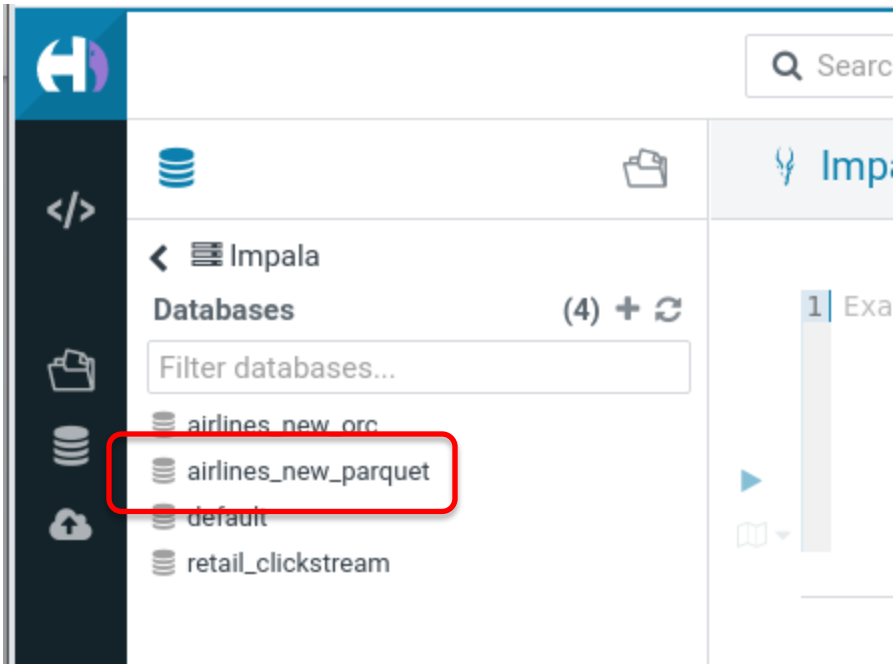
10) Click on HUE to enter the “Hadoop User Experience”



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.

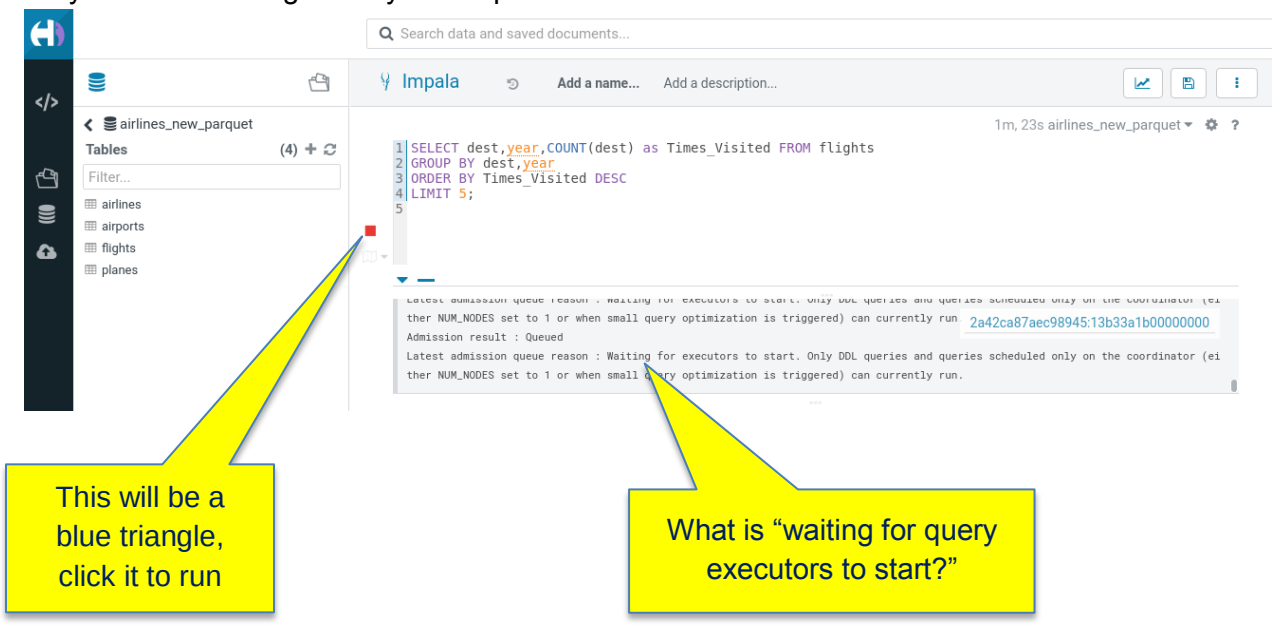


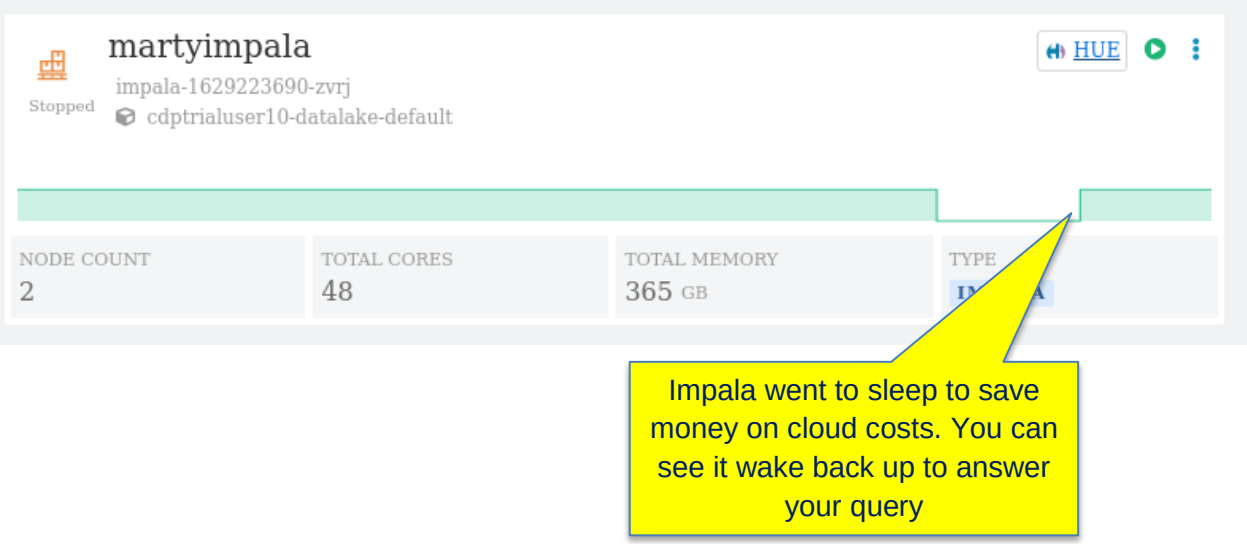
If you ever need to get back to this screen layout click on the “editor” shown here:

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 ☺

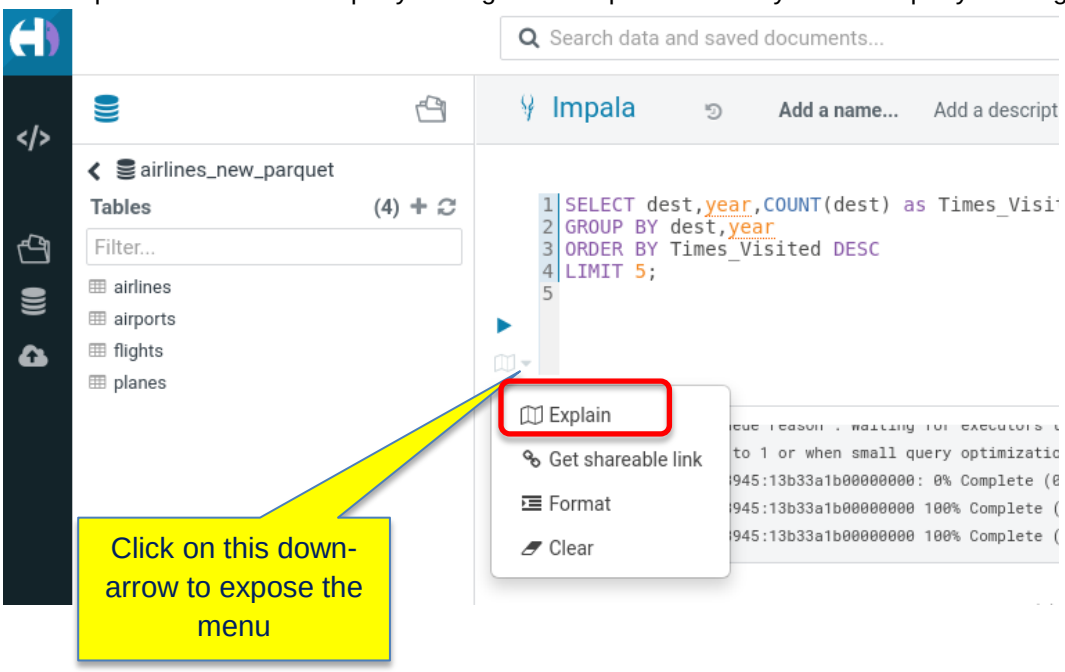


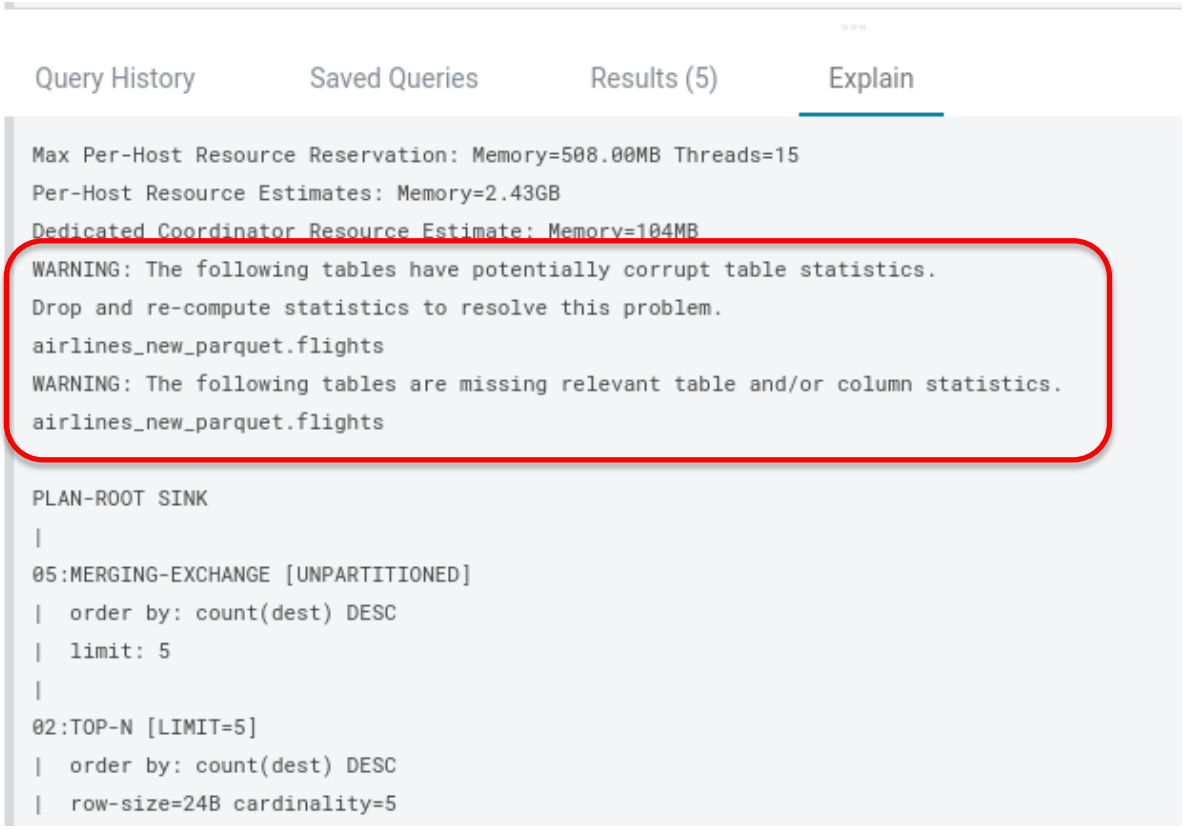


You may get the query result without having to wait for Impala to "wake up." You get to decide how long an idle time you want to wait before you scale down to zero, or if you have the budget you can have Impala always available.

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

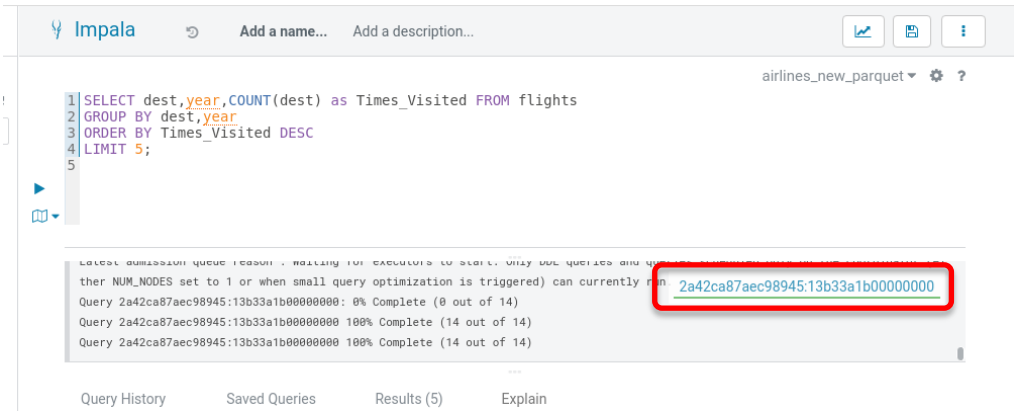
*Not required to execute the query - this gives us a plan on exactly what the query is doing



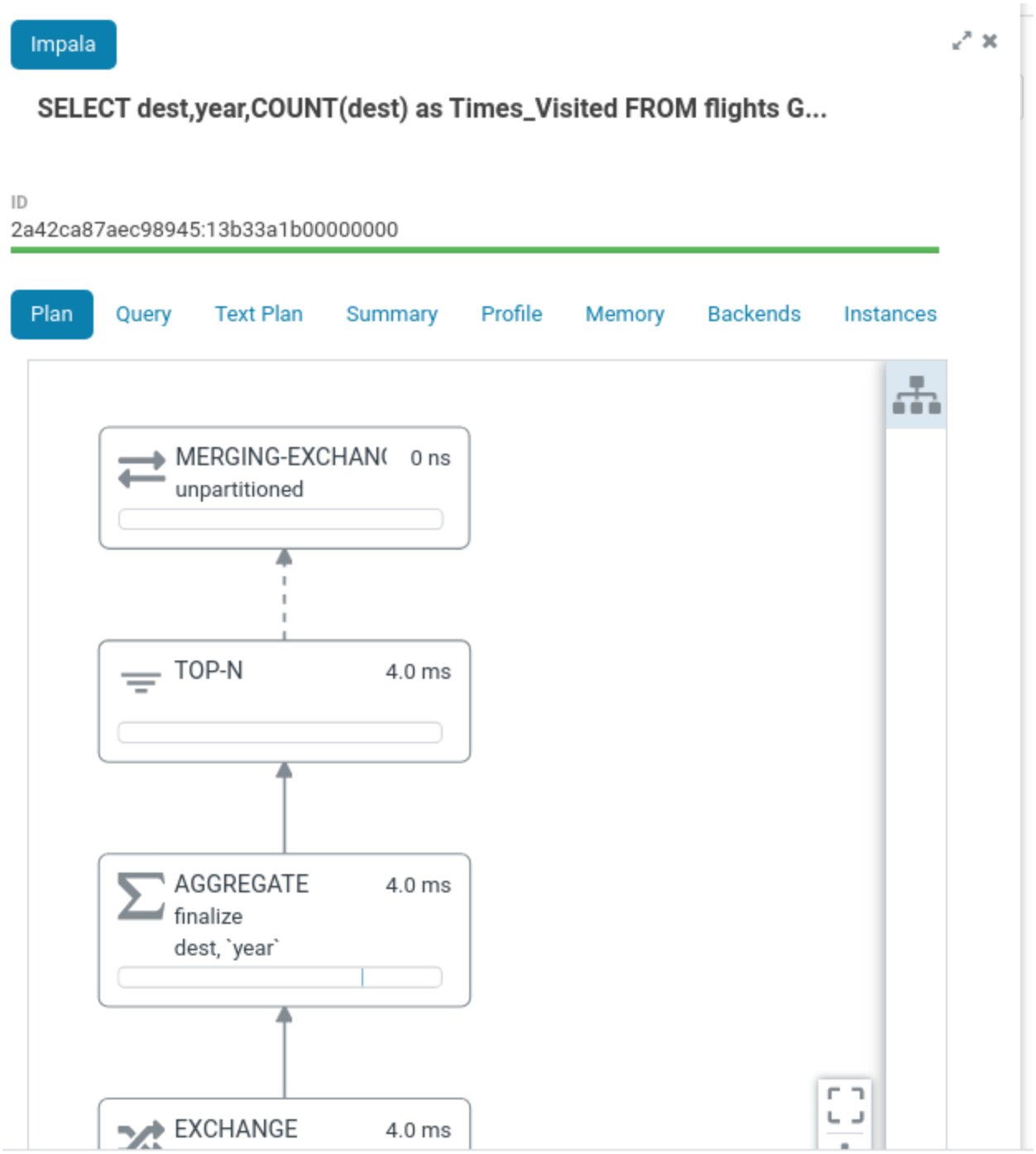


The “Explain Plan” shows you how the query will execute. It is asking you to update statistics for the tables in the query. This is a good idea for performance. We can discuss this in more detail.

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.



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;
```

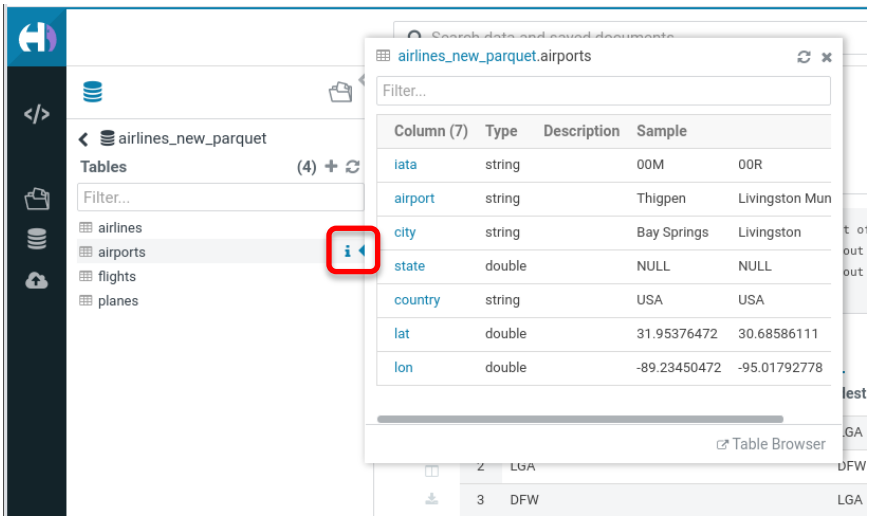
18) Notice the highlighting on the left edge. HUE is parsing based on the semi-colon and the execution arrow will run whatever is highlighted in blue, or whatever has been highlighted by the cursor. This way you can have multiple queries in the same canvas.

19) Click the blue arrow 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)

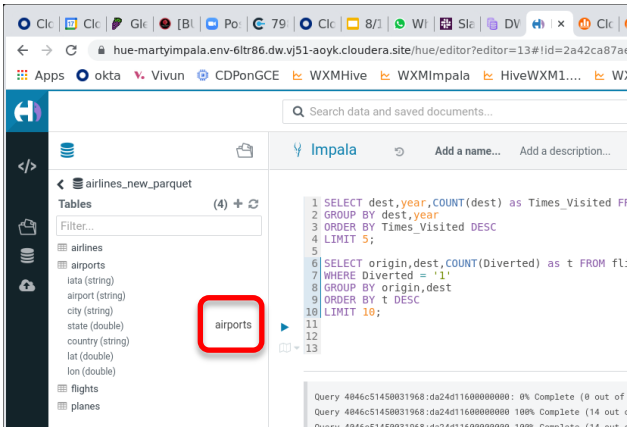
	origin	dest	t
1	ORD	LGA	845
2	LGA	DFW	749
3	DFW	LGA	653
4	DAL	HOU	615
5	ATL	LGA	567
6	MDW	STL	512
7	ATL	DFW	482
8	ORD	DFW	450

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.

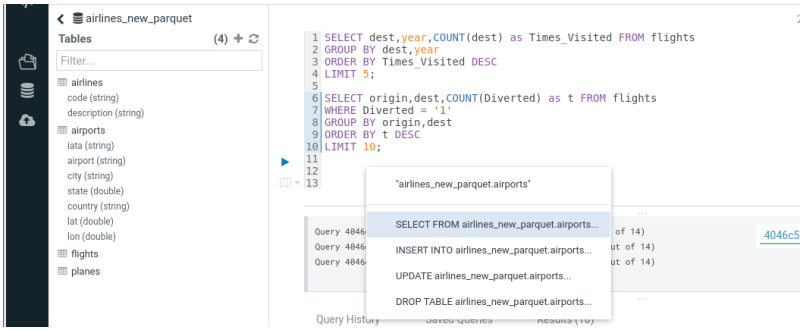


You can also click on the table name to expand all the columns. On the far right of the browser is help tooling.

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



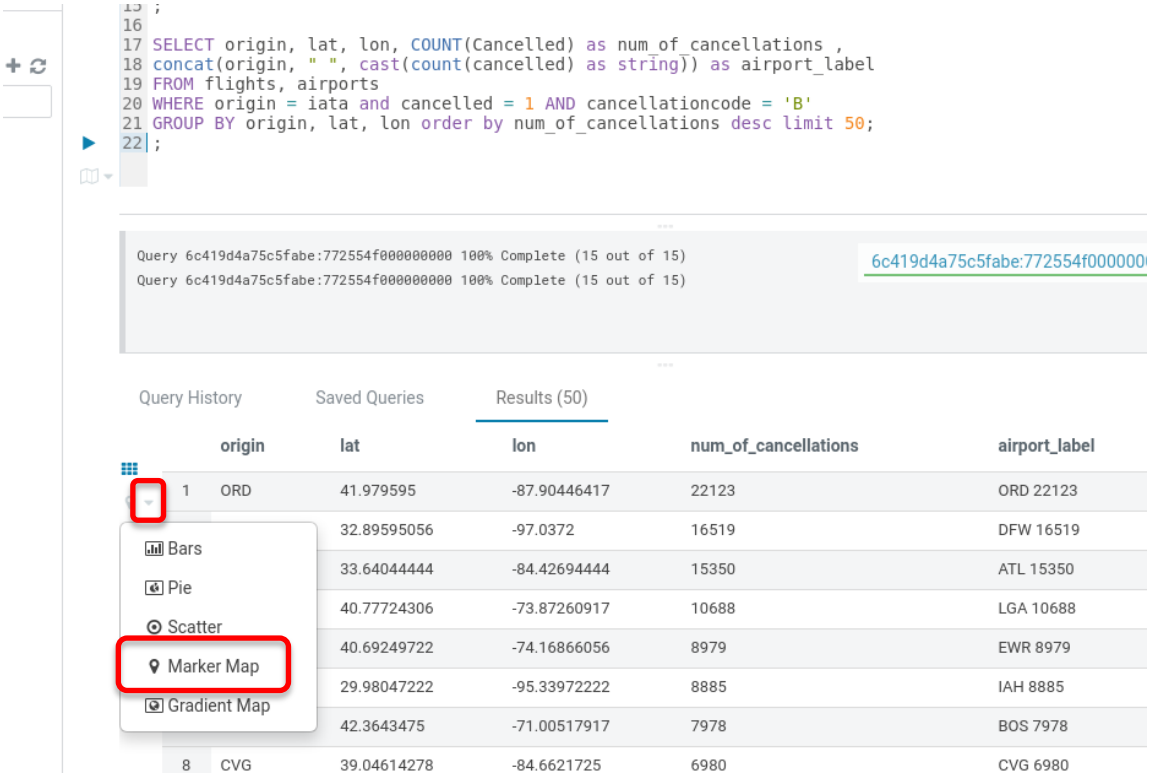
When you drop the table name you’ll get to choose what SQL you want auto generated. Take the “select” option and run the query with the blue-triangle



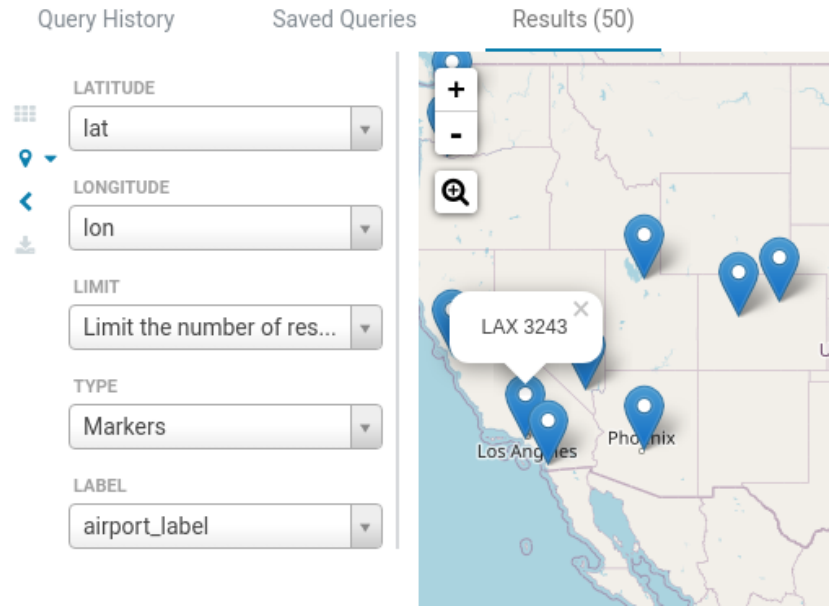
22) Let’s now build a marker map of the airports with the most cancellations. This will correlate with the airports that have the most flights. Run the SQL shown below

```
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;
```

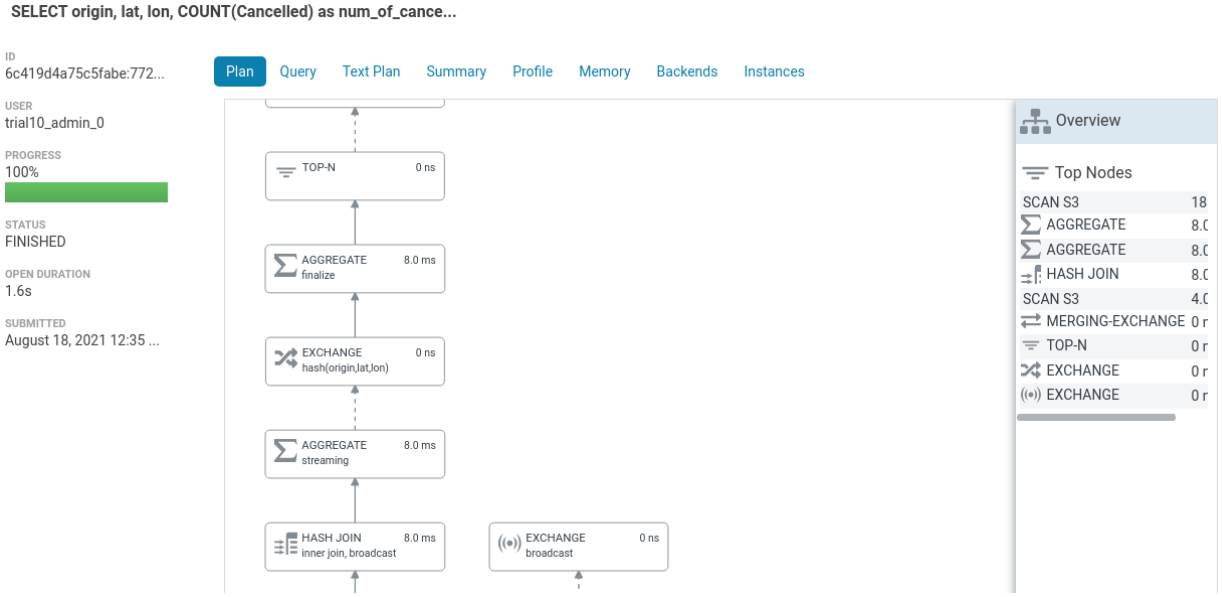
23) After you run the SQL use the down arrow to choose the type of output formatting. You've been using the data grid, we're now going to choose the marker-map



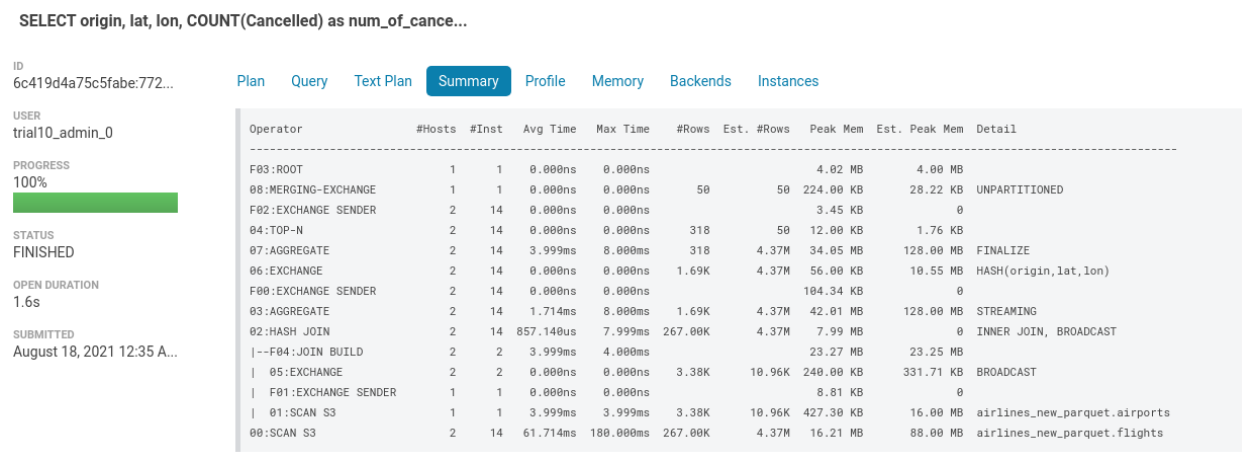
24) Configure the marker map per shown below. Clicking on one of the markers will pop up the value of the “ariport_label” column



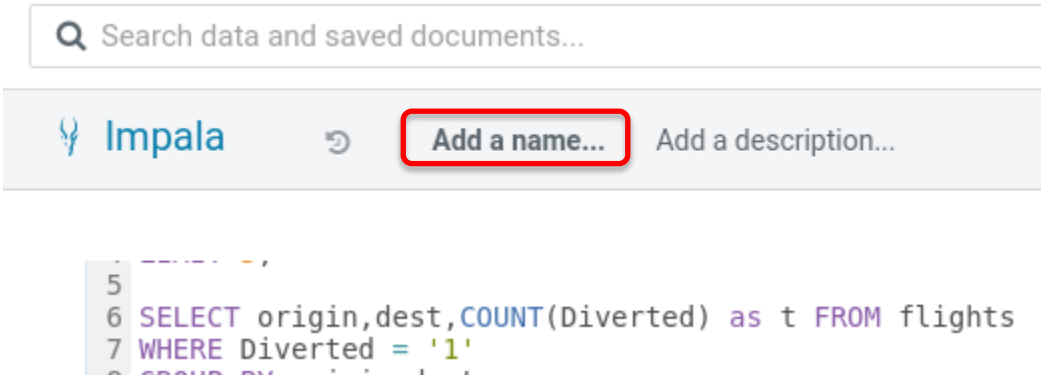
25) Look at the query plan – notice we now have a join in the tree



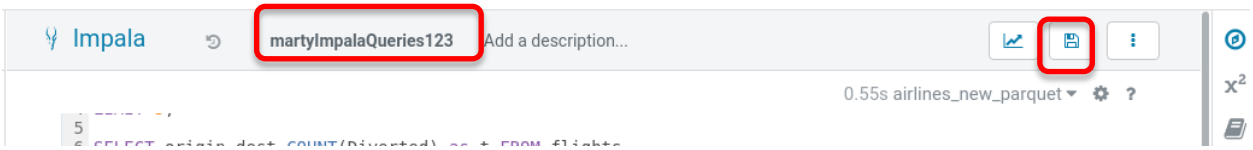
26) The summary shows details of all the stages in the join and their metrics

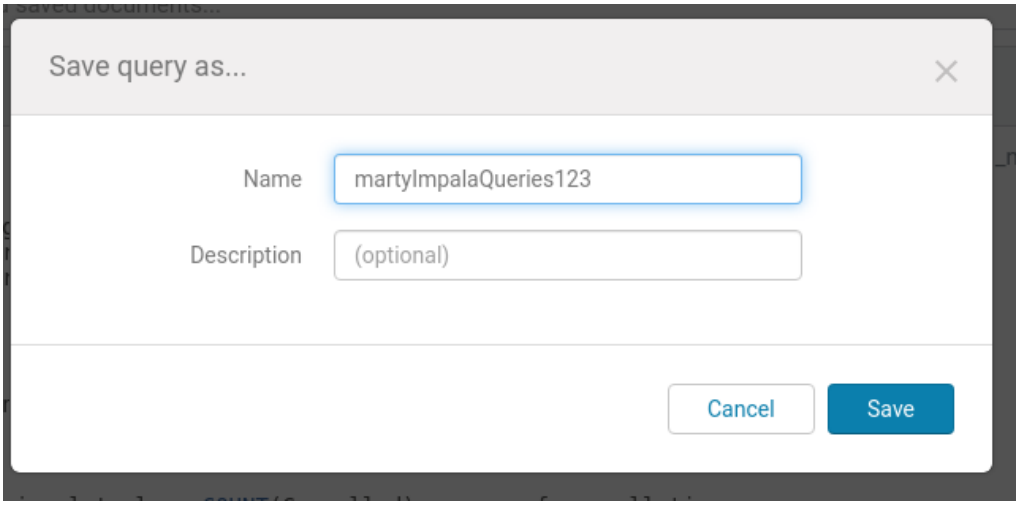


27) Time to save our work. Give your Impala session a name. Use something unique as this is a multi-user environment.



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

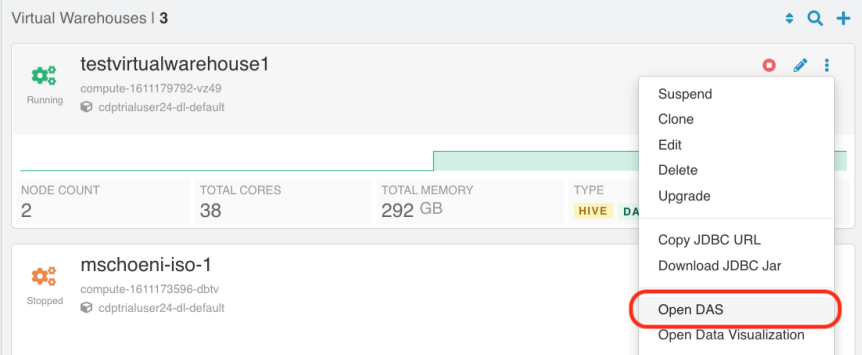




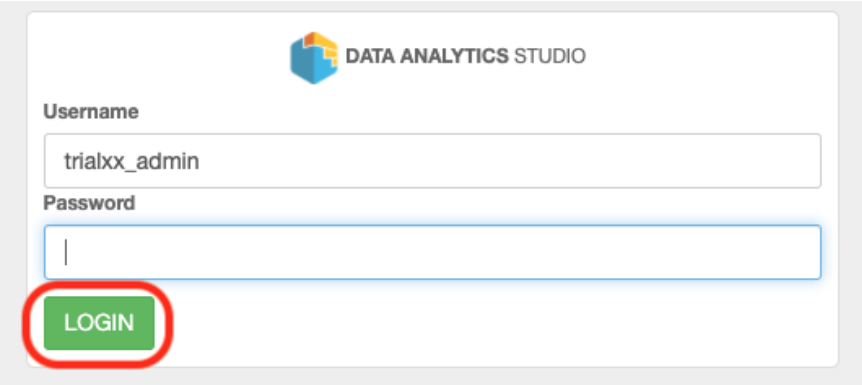
Your saved queries will show up under the “Saved Queries” heading.

SKIP TO THE DATA VISUALIZATION LAB. DAS INSTRUCTIONS BELOW. DAS IS DEPRECATED SO I URGE YOU TO JUMP TO THE DATA VISUALIZATION… so I don’t really have strong feelings about this…

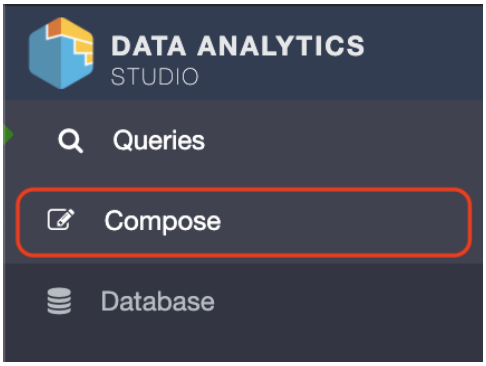
9) Once your Virtual Warehouse is “Running”, click the line in the top right and then click “Open DAS”



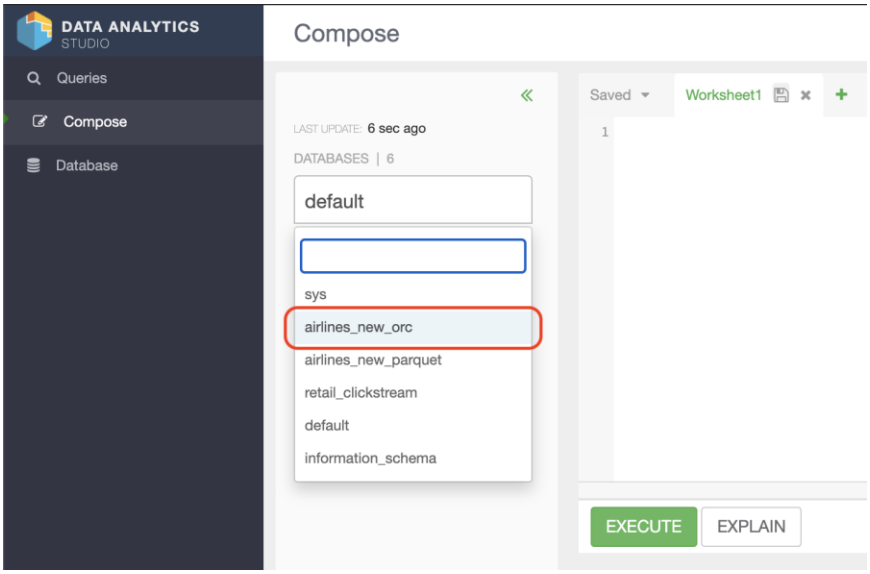
10) Enter the login information from step #1 above using the user, then click “LOGIN”
*Changing “trialxx_admin” to the trail user you’re using and password defined by “X” in #1 above



11) Click on “Compose”, to write the queries below to answer questions on the table “flights”

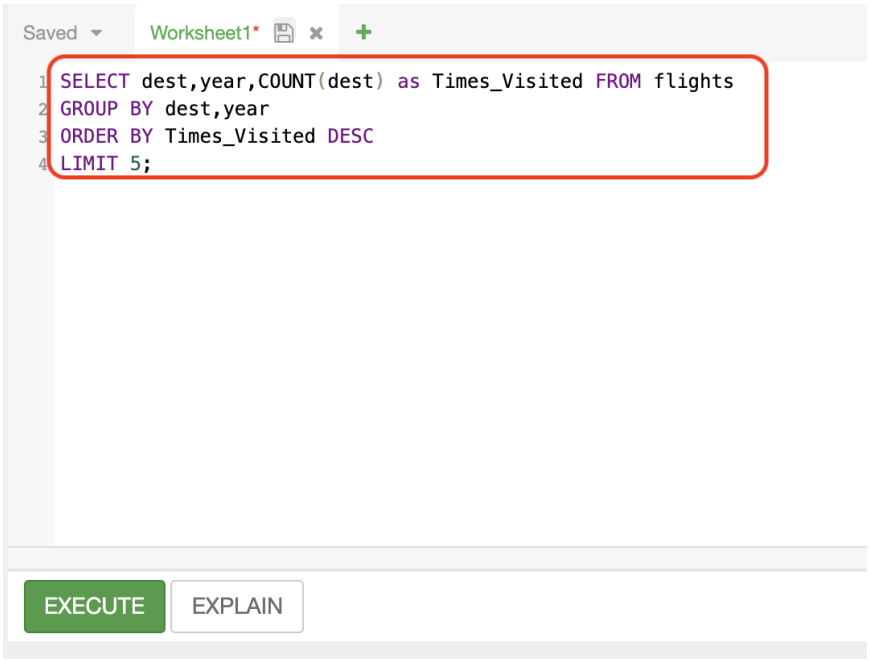


12) Choose the database “airlines_new_orc” that we found in Part 1 “Data Catalog”



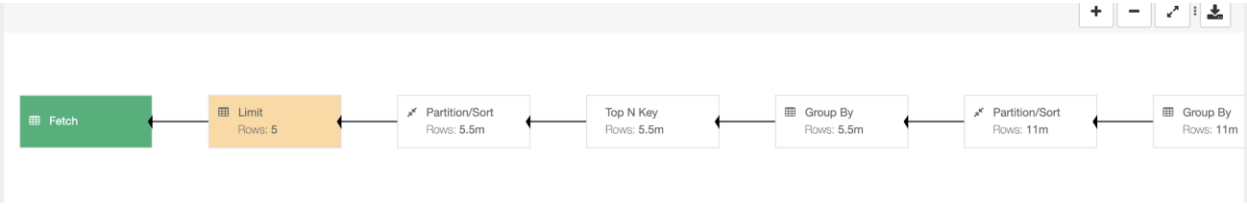
13) Enter the following query in Worksheet1, answering the question “show me the top 5 visited destination by year from (1995-2008)”

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

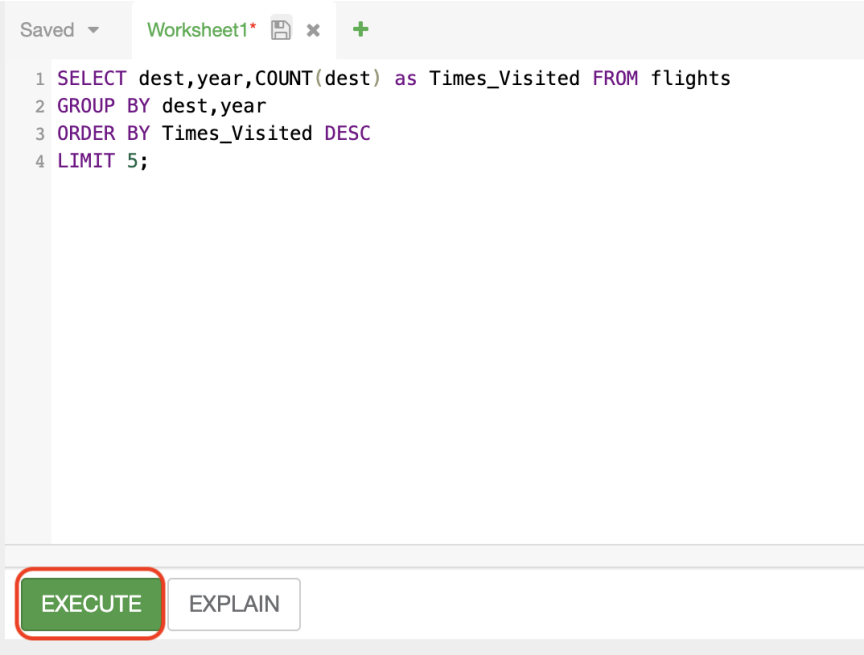


14) Click “EXPLAIN” to see the visual explain plan prior to running the query

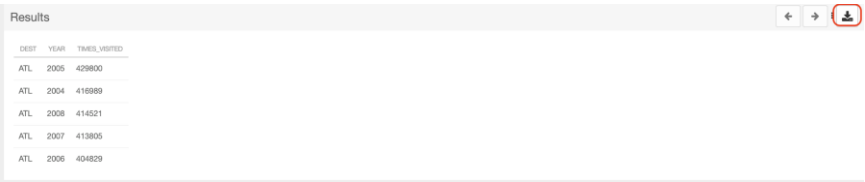
*Not required to execute the query - this gives us a plan on exactly what the query is doing



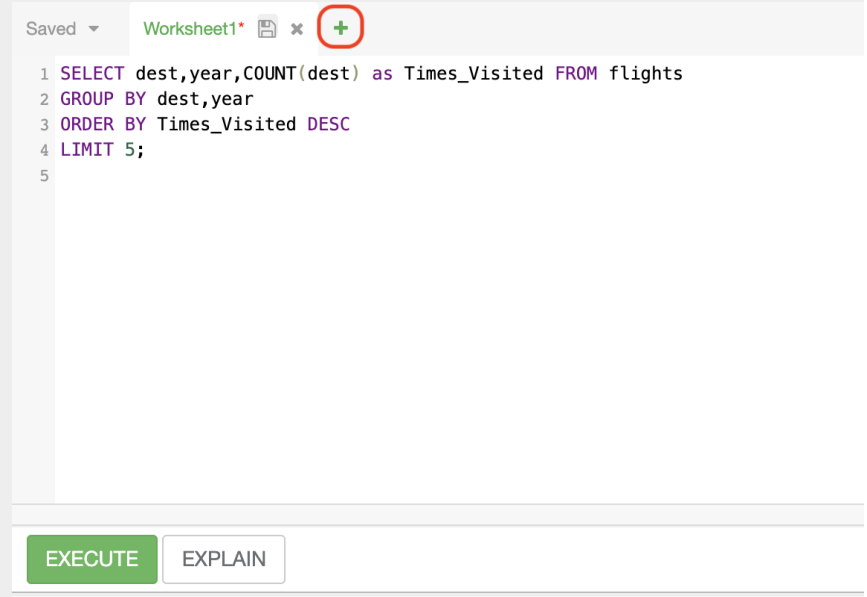
15) Click “EXECUTE” to execute the query, answering the question “show me the top 5 visited destination by year from (1995-2008)”



16) Click the download button on the top right, to download the results as a CSV file

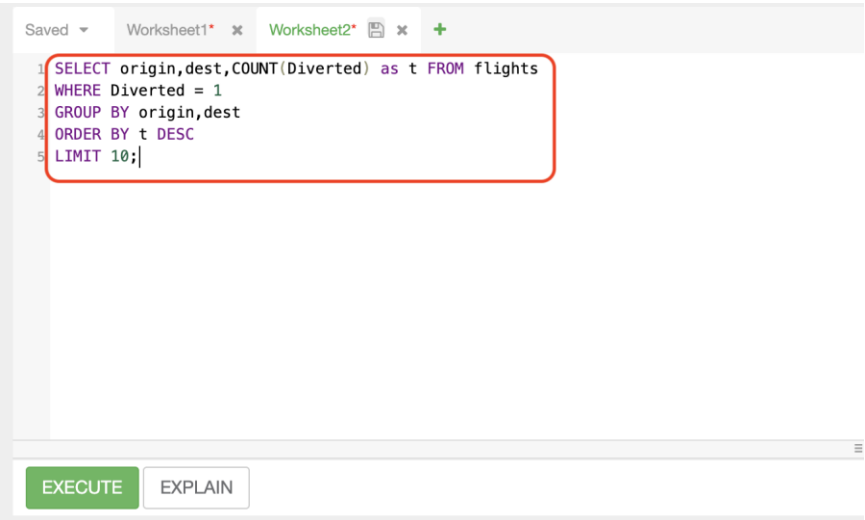


17) Going back to “Worksheet 1”, click the “+” to add another Worksheet for the next query

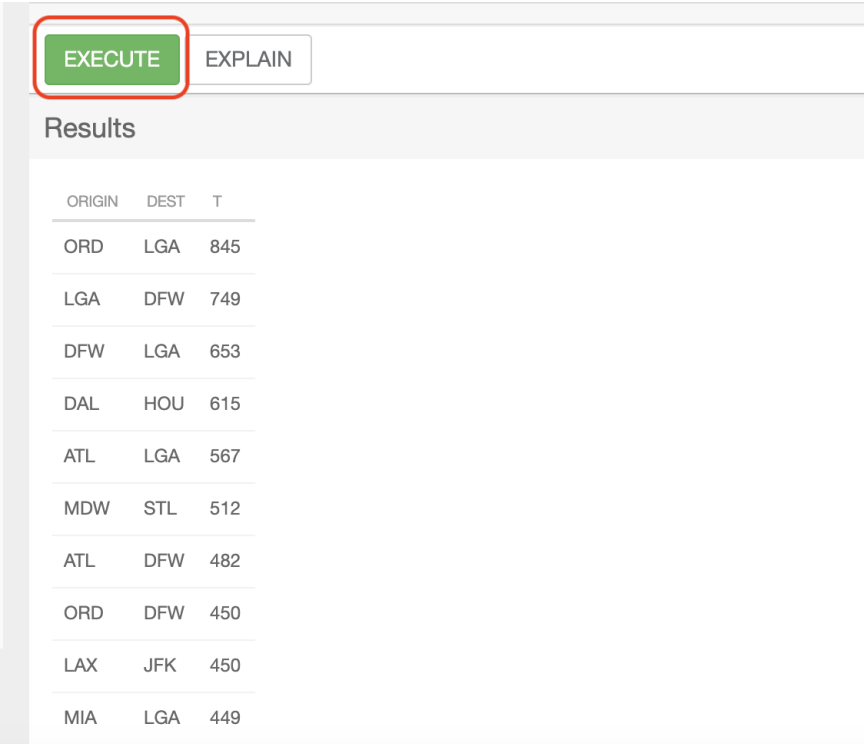


18) In “Worksheet 2”, Choose the database “airlines_new_orc” then copy-and-paste the following query, answering the question “What are the top 10 routes (origin and dest) that have seen maximum diversions?”

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

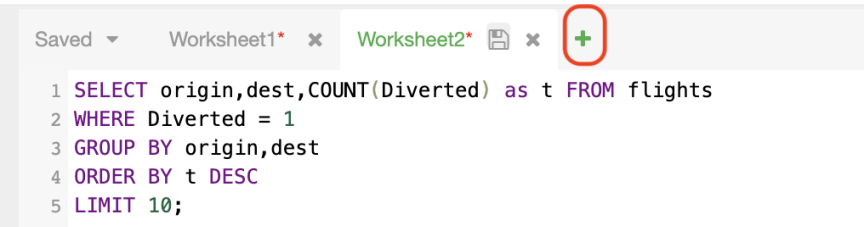


19) Click “EXECUTE” to execute the query, answering the question “What are the top 10 routes (origin and dest) that have seen maximum diversions?”



ORIGIN	DEST	T
ORD	LGA	845
LGA	DFW	749
DFW	LGA	653
DAL	HOU	615
ATL	LGA	567
MDW	STL	512
ATL	DFW	482
ORD	DFW	450
LAX	JFK	450
MIA	LGA	449

20) Going back to “Worksheet 2”, click the “+” to add another Worksheet for the final query

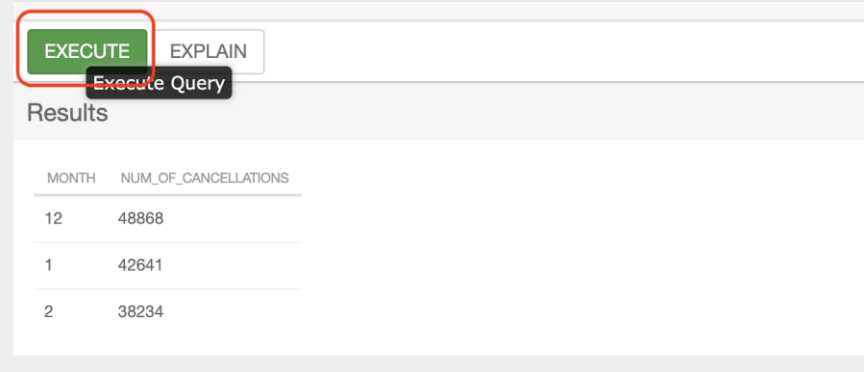



```
SELECT origin,dest,COUNT(Diverted) as t FROM flights
WHERE Diverted = 1
GROUP BY origin,dest
ORDER BY t DESC
LIMIT 10;
```

21) In “Worksheet 3”, Choose the database “airlines_new_orc” then copy-and-paste the following query, answering the question “Which three months have seen the most number of cancellation due to bad weather?”

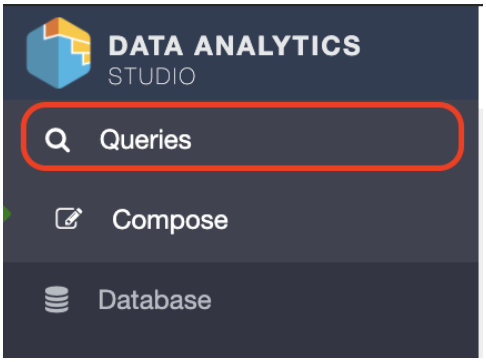
```
SELECT month,COUNT(Cancelled) as num_of_cancellations FROM flights
WHERE Cancelled = 1 AND CancellationCode = 'B'
GROUP BY month
ORDER BY num_of_cancellations DESC
LIMIT 3;
```

22) Click “EXECUTE” to execute the query, answering the question “Which three months have seen the most number of cancellation due to bad weather?”

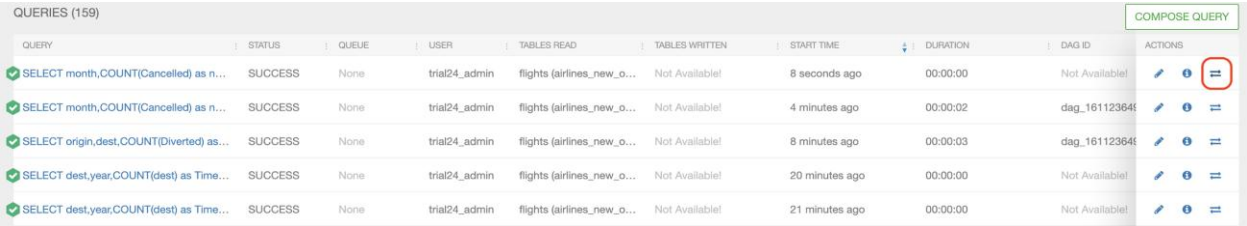


22) Click “EXECUTE” a second time - this will lead us to our last portion of Part 2

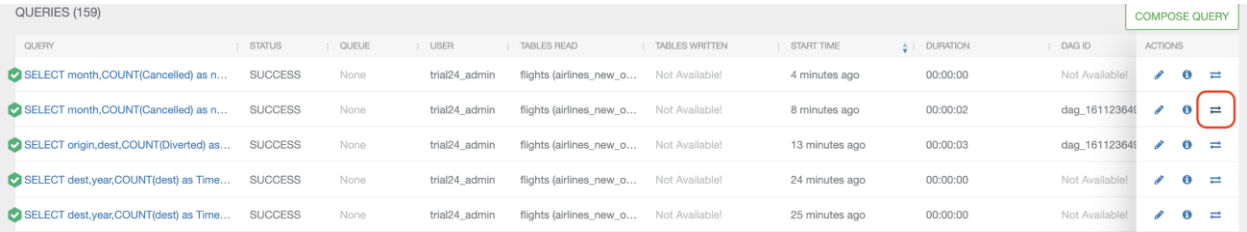
23) Click on “Queries” on the top left navigation bar
*We’ll look at our query history



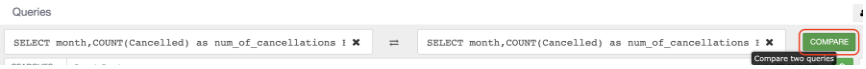
24) Click the “Compare” on the right of your last query run (query at the top)



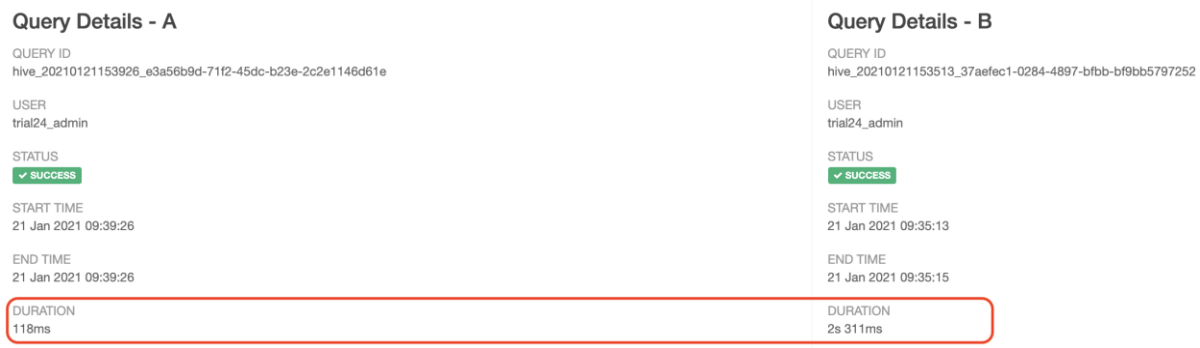
25) Click the “Compare” on the right of the query (second to the top)



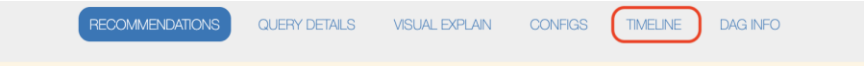
26) Click on the “COMPARE” button to compare the two queries



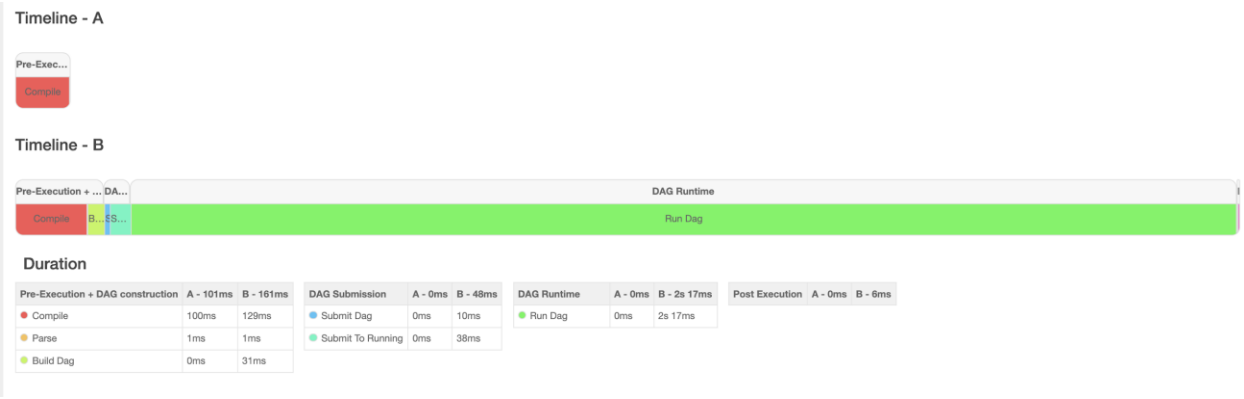
27) Notice the run duration is different between the two, let’s find out why



28) Click on “timeline” at the top



As shown, the faster query only did “compile and parse”, while the slower query did “compile, parse, build dag, submit dag, submit to running, run dag”. Why? Because if you run the same exact query twice, the results are cached (if the data didn’t change). CDW knows if the data changed.



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

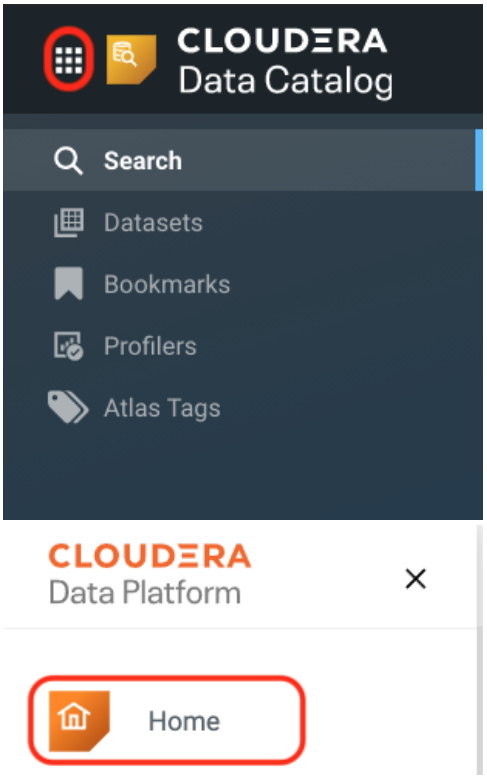
1) Open CDP, using the “admin” user within the Test Drive link.

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

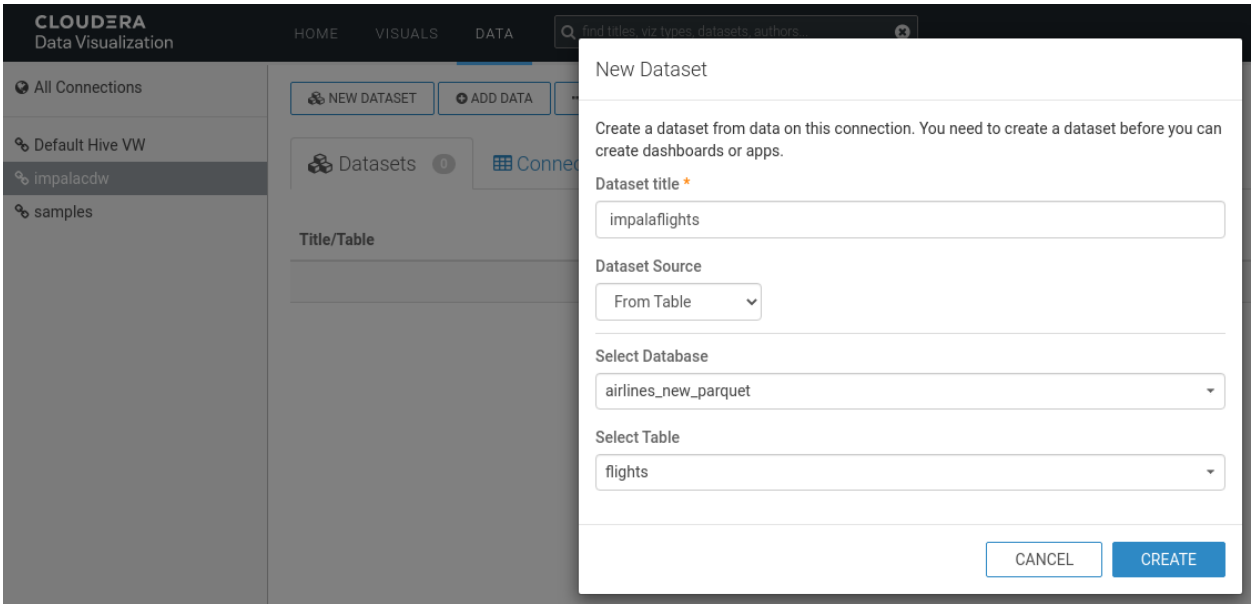
2) Click the “Data Warehouse” within the CDP Home Screen

How do you get to the CDP Home Screen?

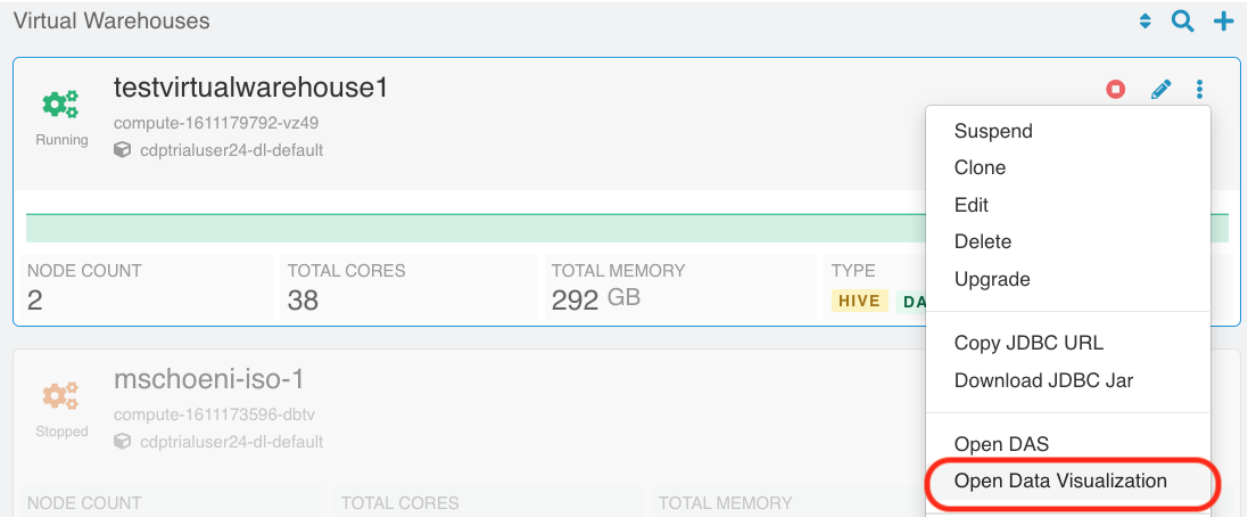
- From any experience such as “Data Catalog”, click the 9 square at the top left and then click “Home”



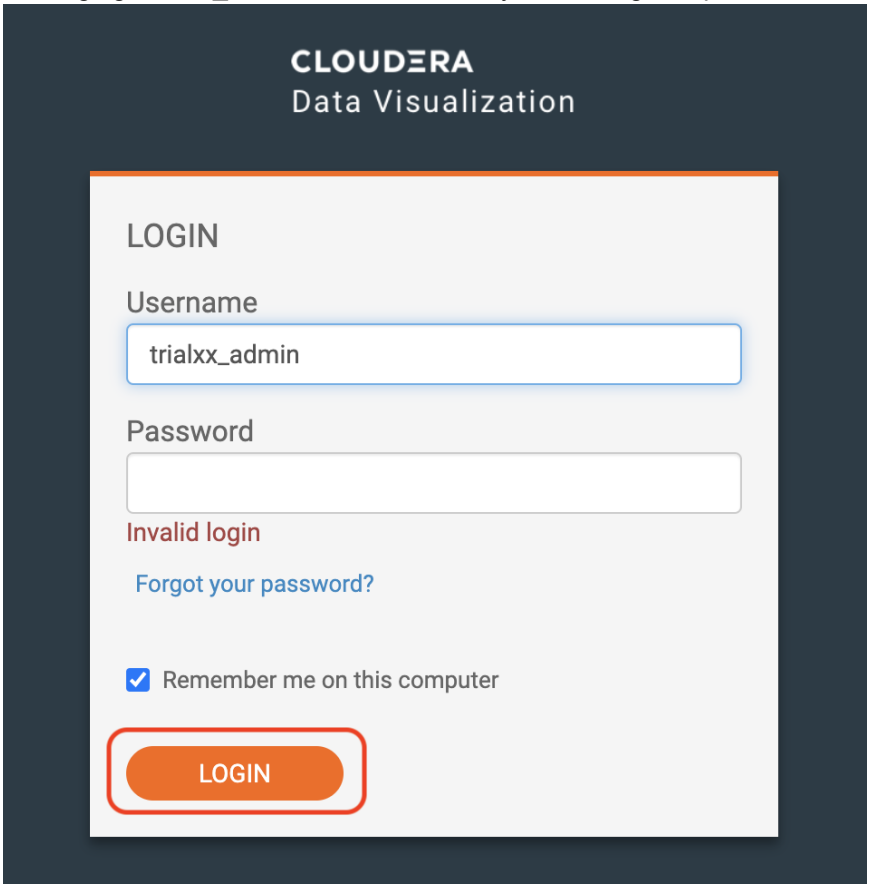
NOTES FOR NEW LAB: You will autologin, and when you get to create new dataset use the impalacdw and corresponding setting shown here:



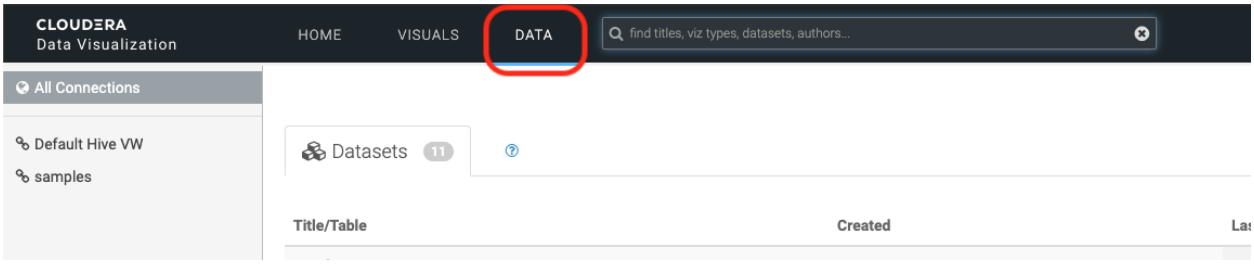
3) Click “Open Data Visualization” on your existing “Running” Virtual Warehouse



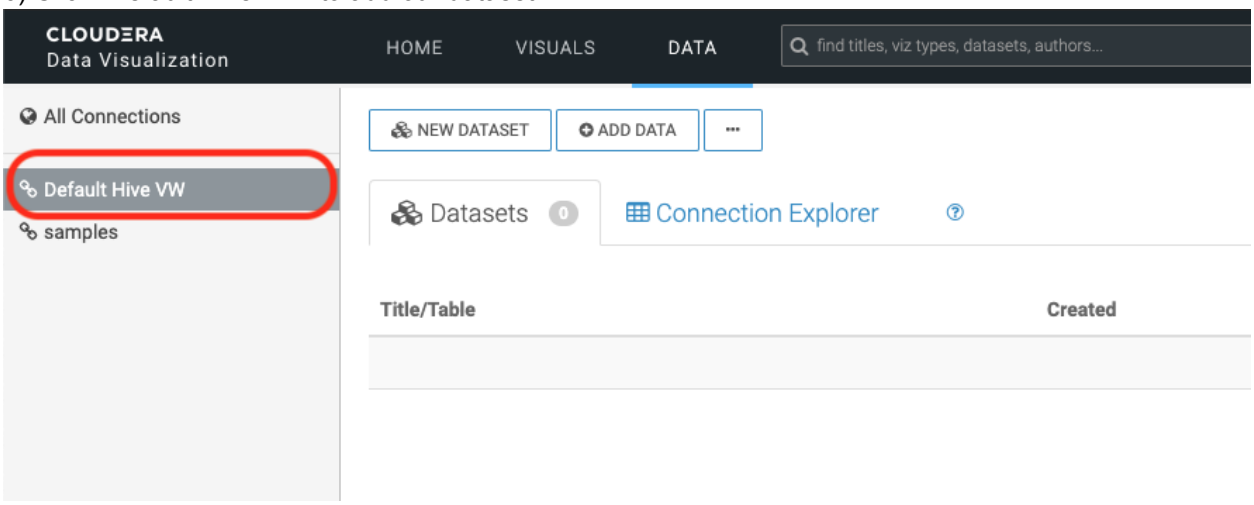
4) Enter the login information from step #1 above using the user, then click “LOGIN”
*Changing “trialxx_admin” to the trail user you’re using and password defined by “X” in #1 above



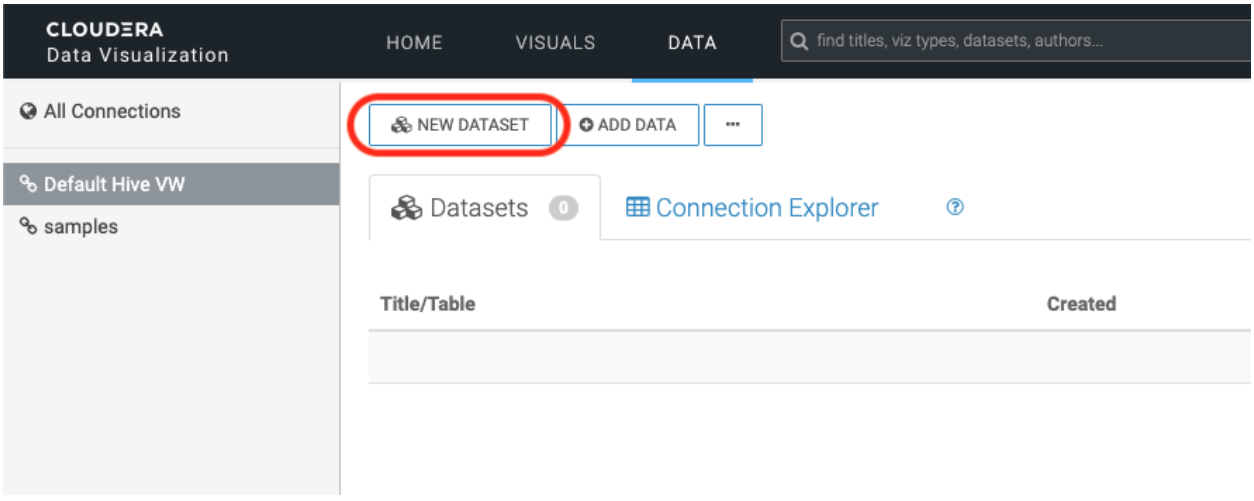
5) Click “DATA” the top navigation bar



6) Click “Default Hive VW” to add our dataset

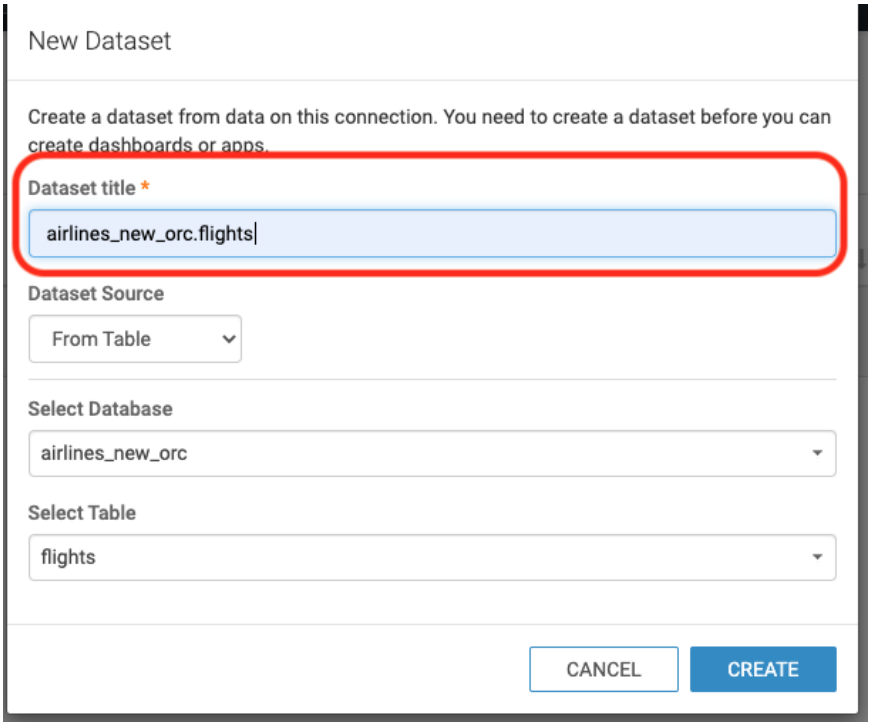


7) Click “NEW DATASET” to add our “flights” data

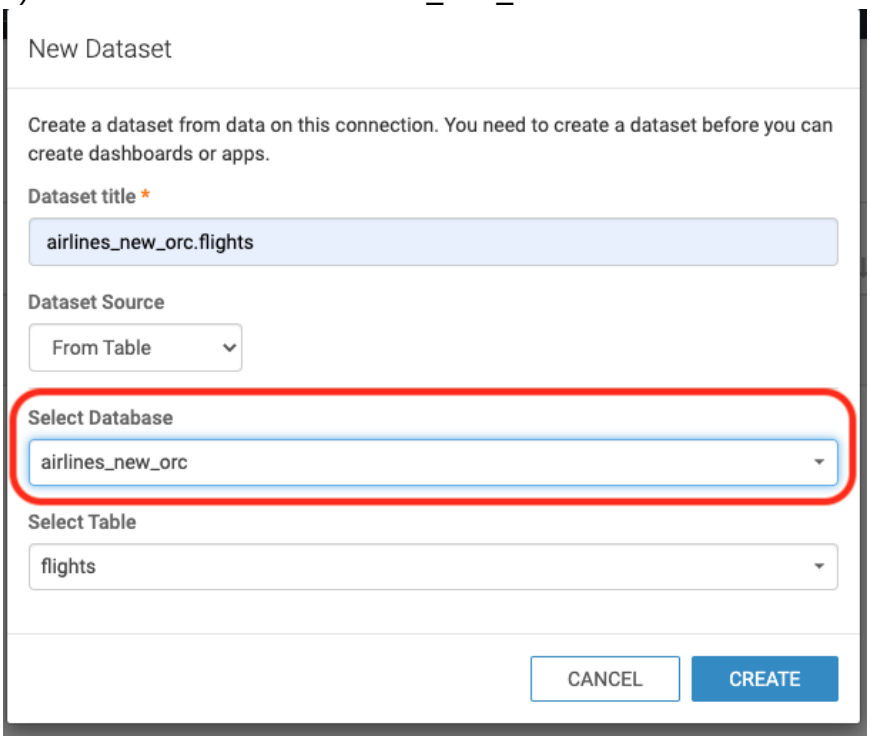


8) Enter a name for the Dataset title naming “airline_new_orc.flights”
*Can be any name you choose

Ls

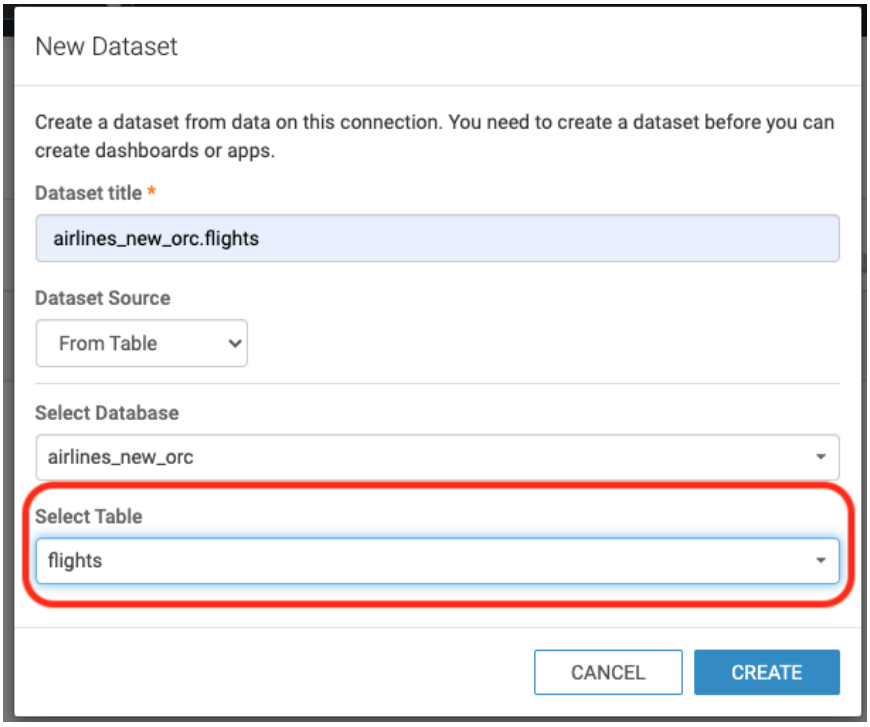


9) Choose the database “airlines_new_orc”

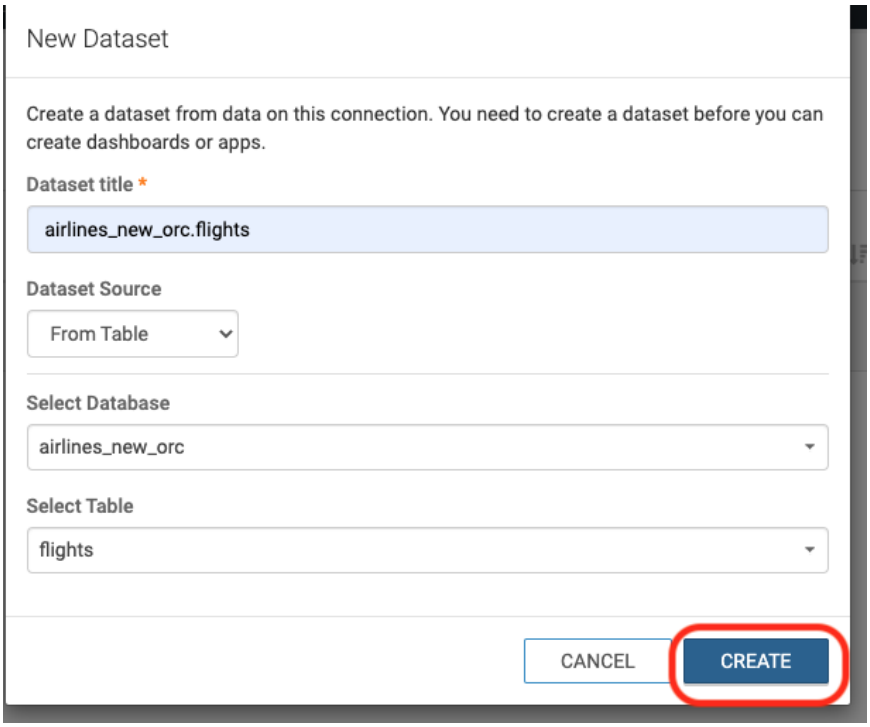


10) Choose the table “flights”

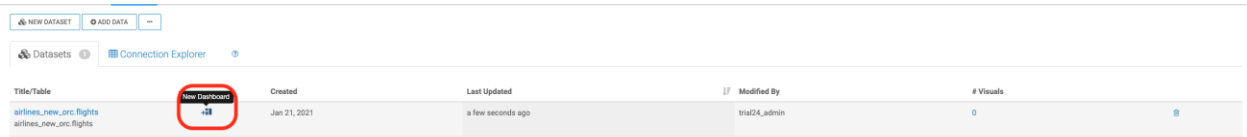
*Need to import multiple databases and tables? You’d use Dataset Source = SQL



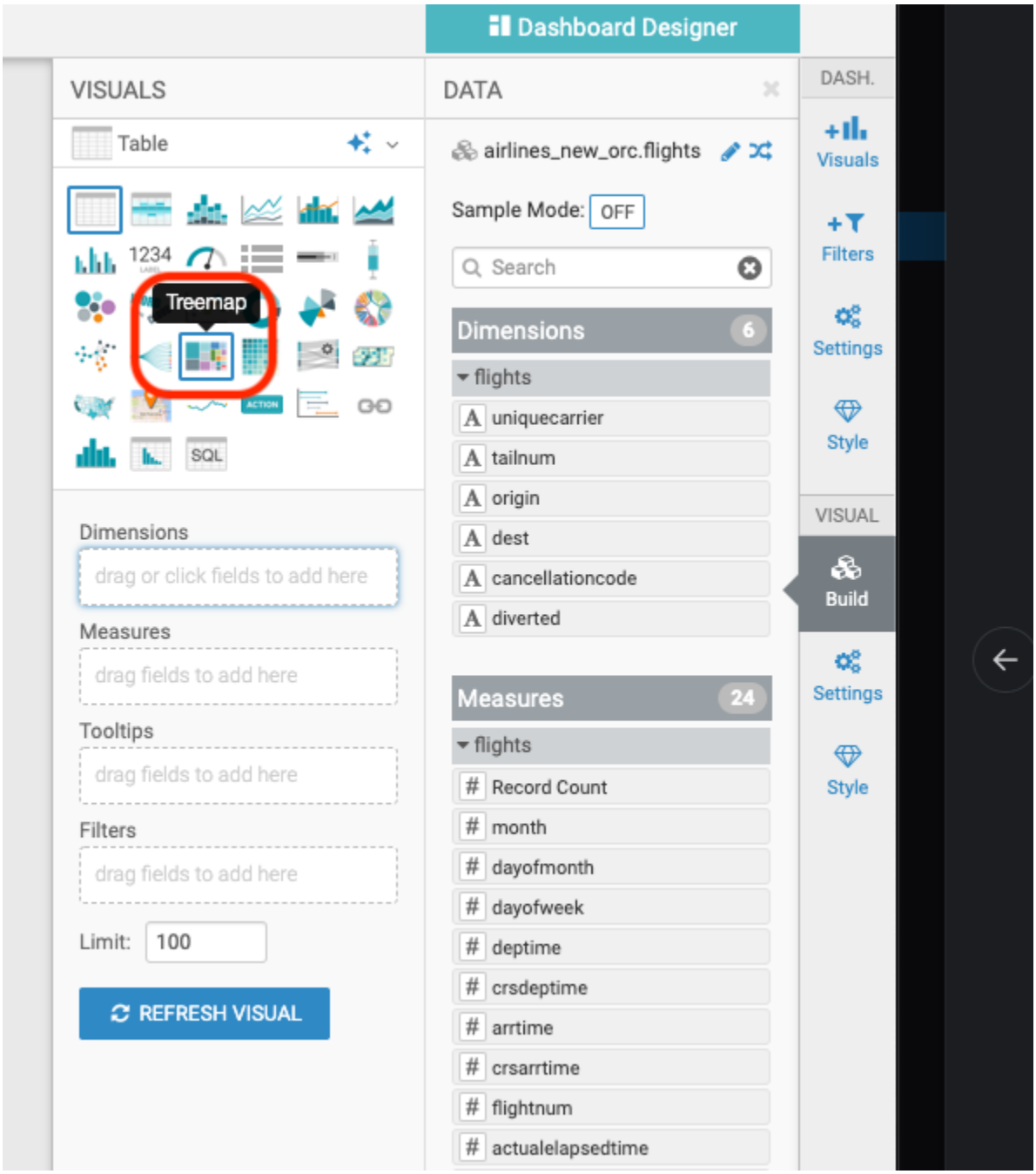
11) Click “CREATE”



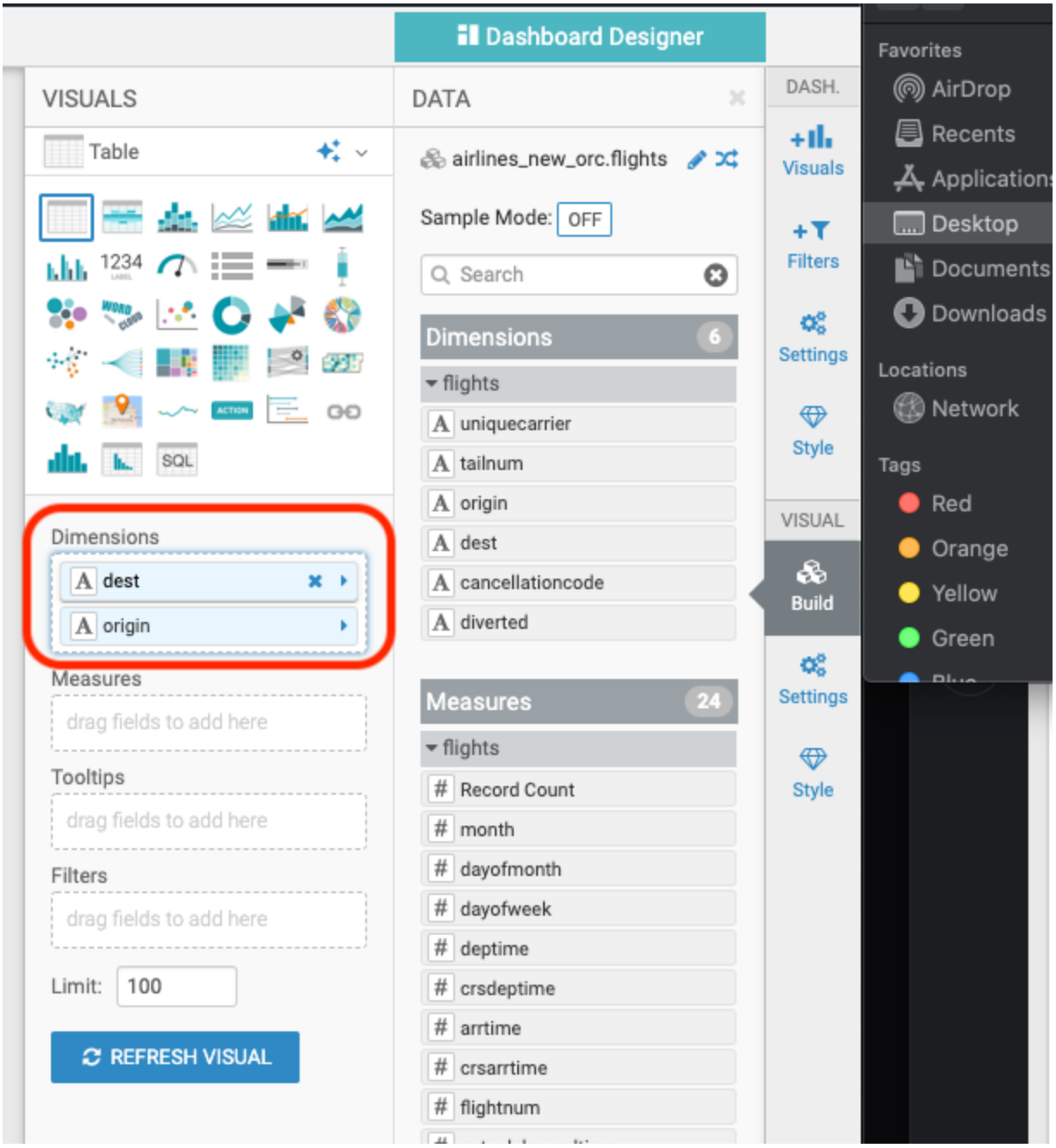
12) Click “+” to create a New Dashboard



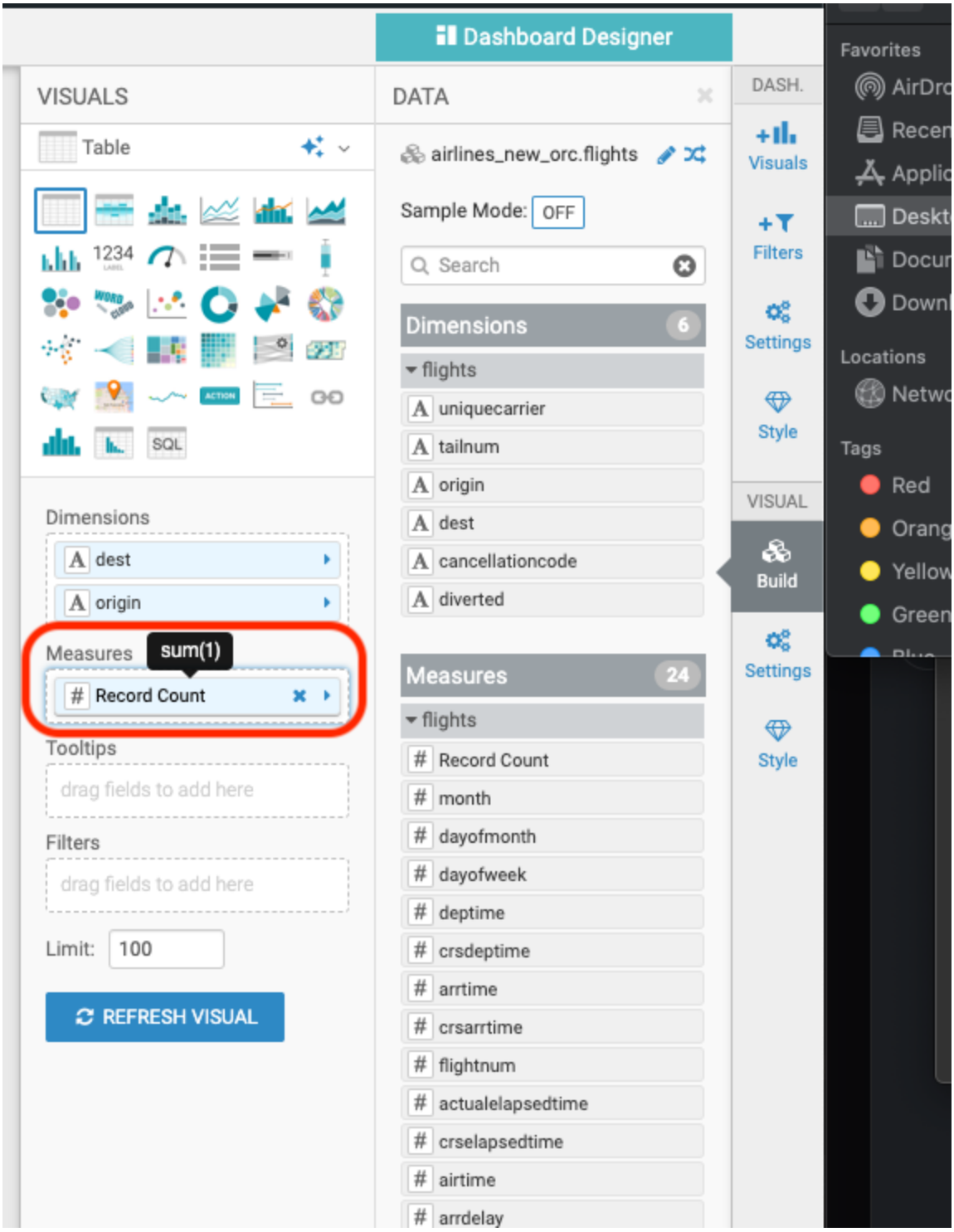
13) Choose “TreeMap” under “VISUALS”



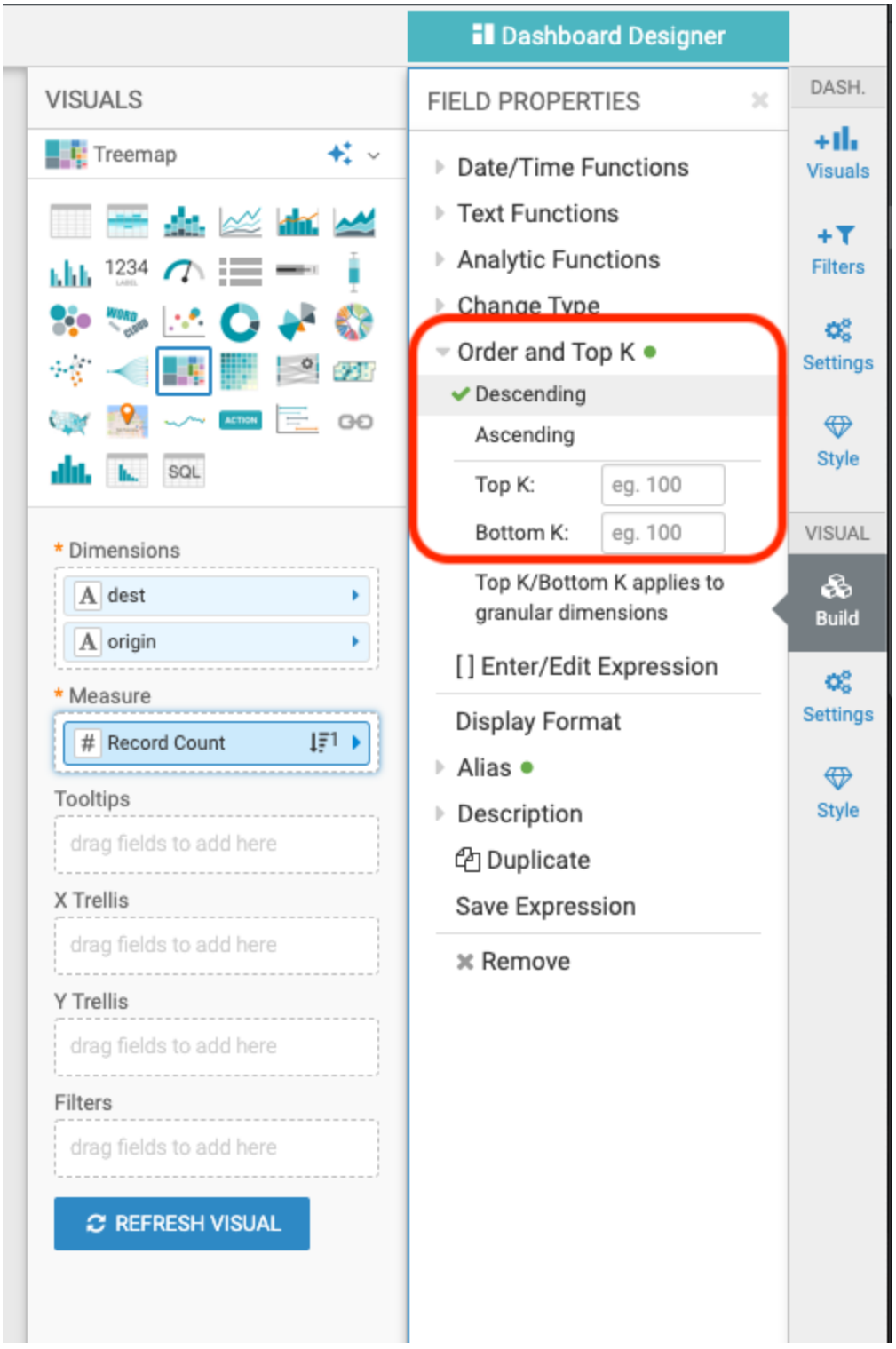
14) Drag-and-drop both “dest” and origin” from Dimensions->Flights into Dimensions under Visuals



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

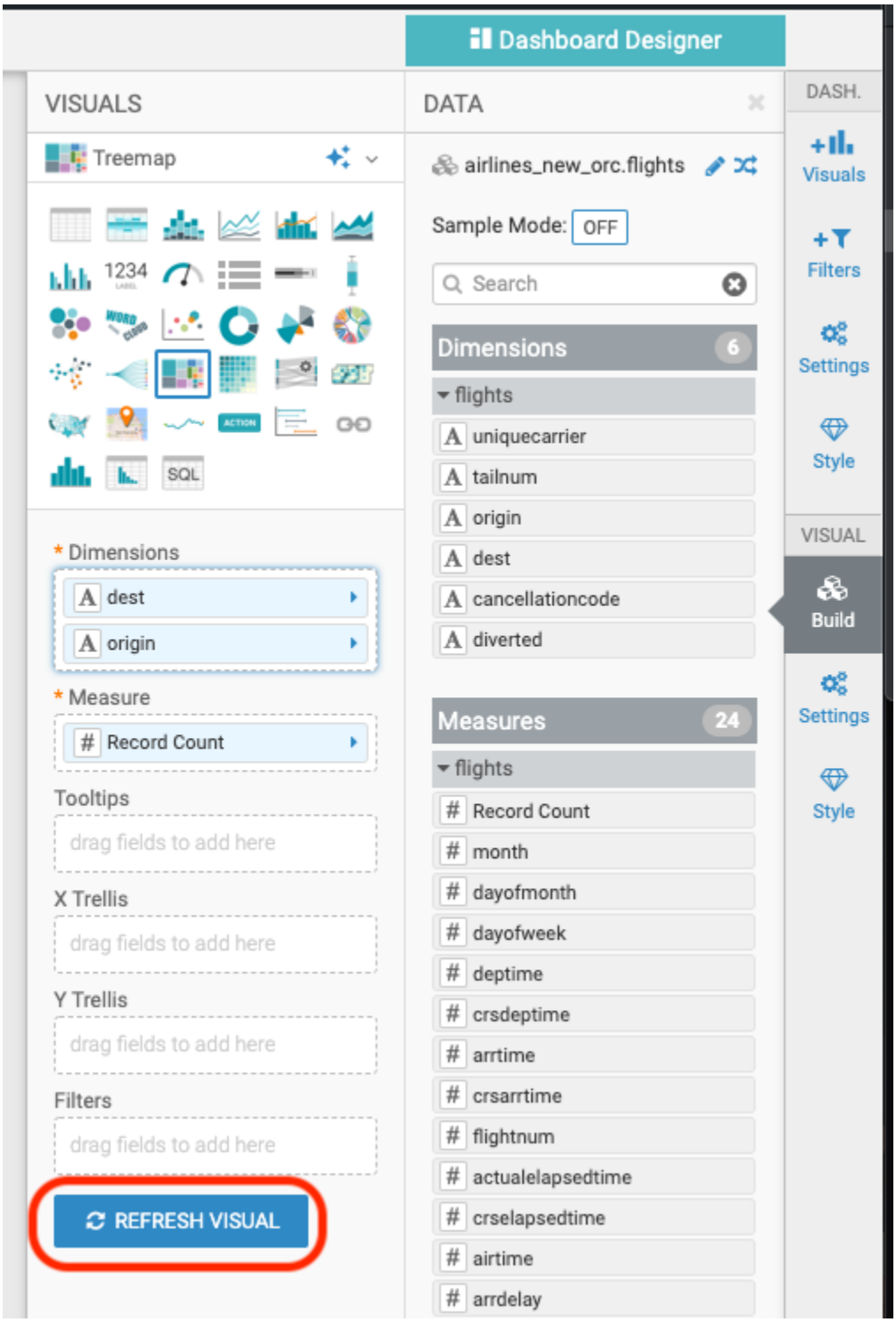


16) Click the right arrow next to Record Count and select “Descending” under Order and Top K

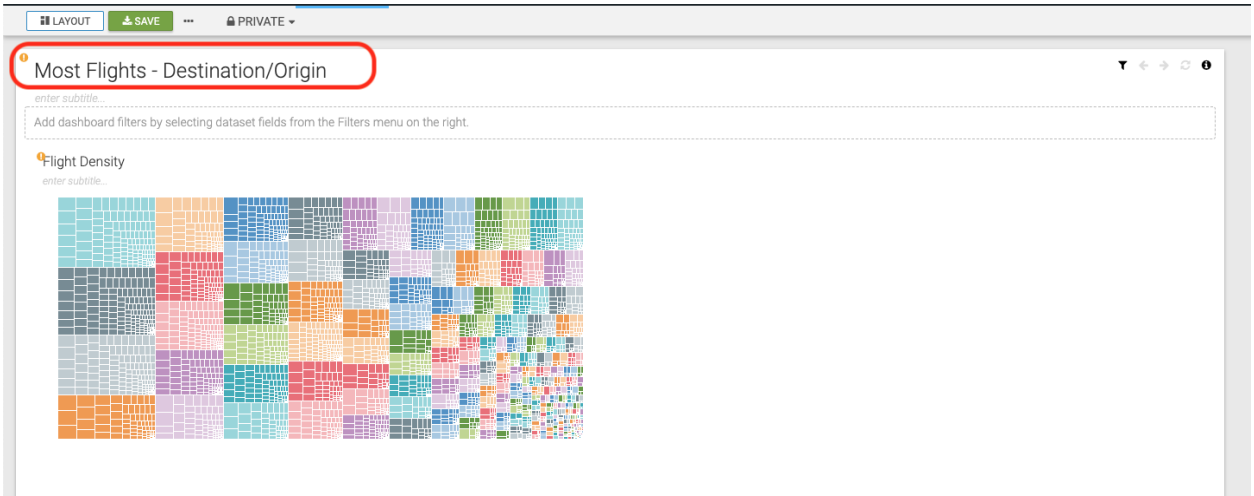


17) Click “REFRESH VISUAL”

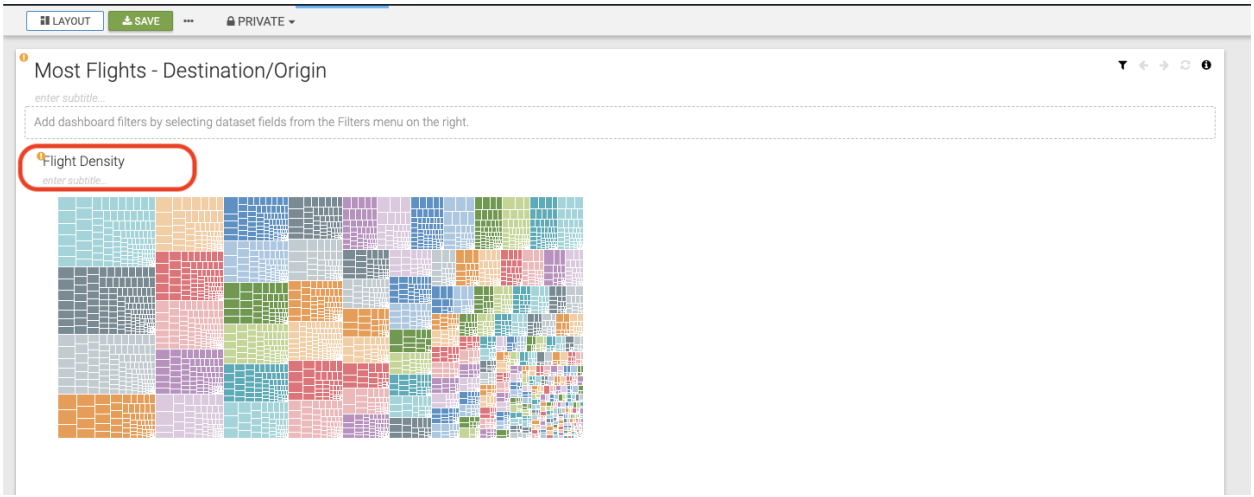
*Notice - you can have other Visuals chosen to be displayed with the Dimensions and Measure(s), then click REFRESH VISUALS



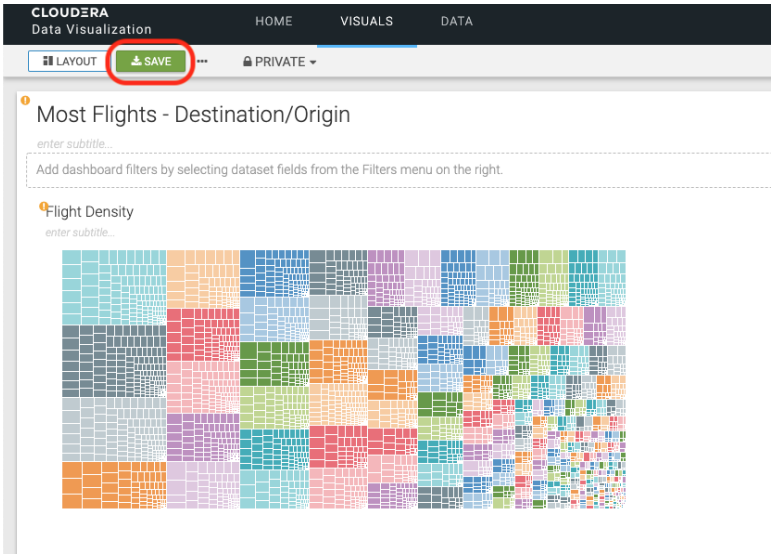
18) Enter a title “Most Flights - Destination/Origin”



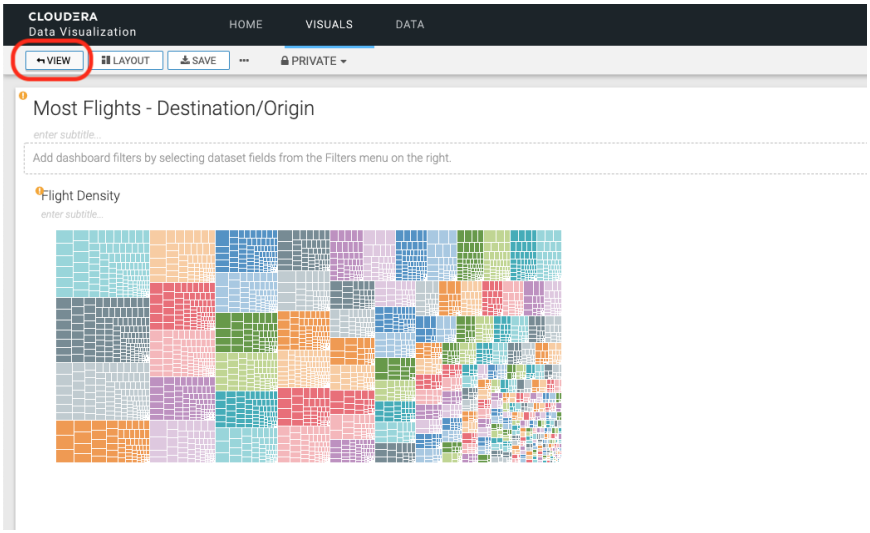
19) Enter “Flight Density” under the graph’s title



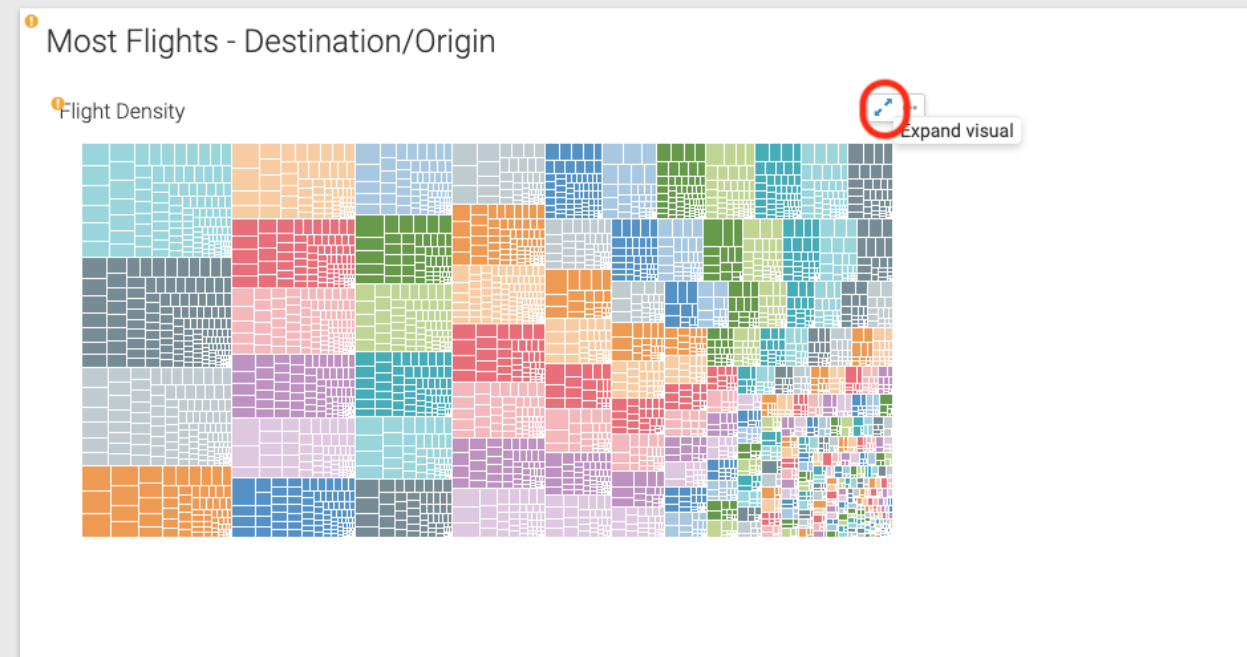
20) Click “SAVE”



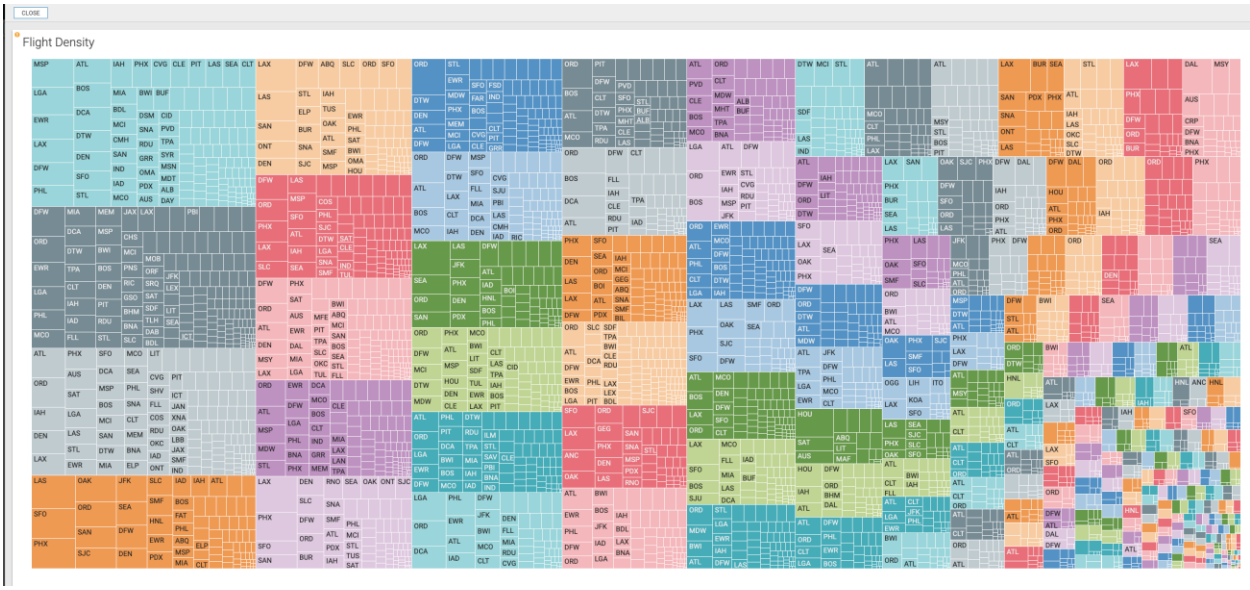
21) Click “VIEW”



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



Destinations are displayed



OPTIONAL DATA IMPORT LAB USING DAS. YOU WILL DO THIS DATA IMPORT IN SPARK TOMORROW USING THE CLOUDERA MACHINE LEARNING.

Part 4 - Import a File into a Table [15 minutes]

Overview: How do we import data (csv file), creating a table?

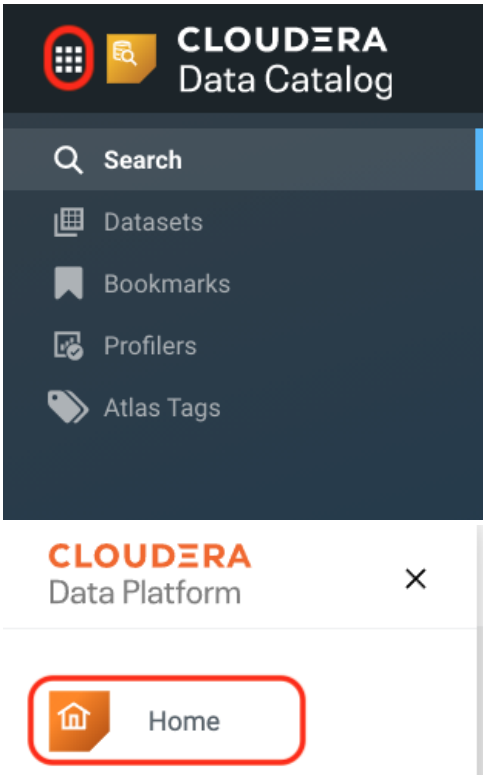
1) Open CDP, using the “admin” user within the Test Drive link.

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

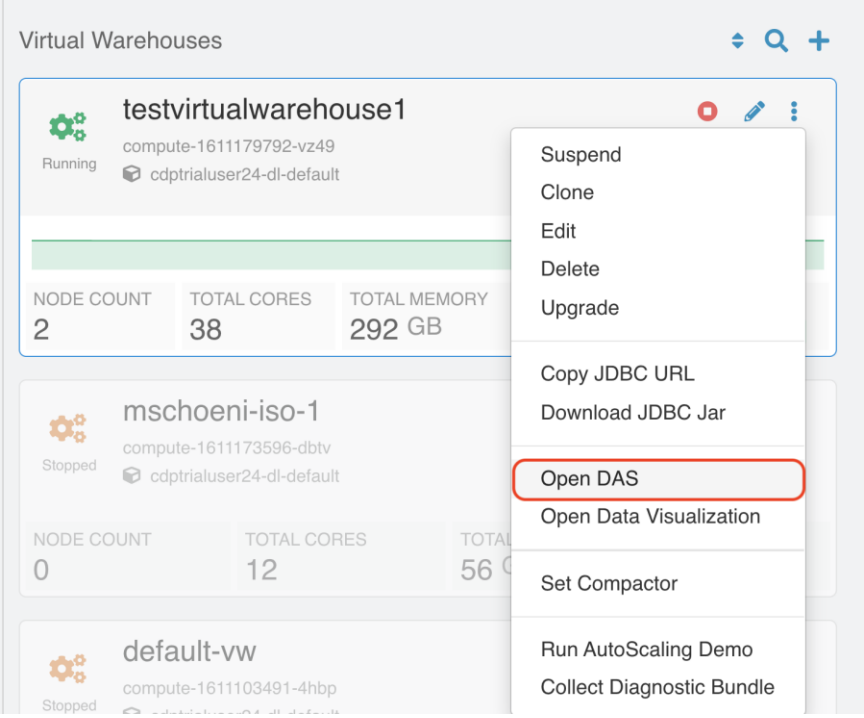
How do you get to the CDP Home Screen?

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



3) Click “Open DAS” on your existing “Running” Virtual Warehouse

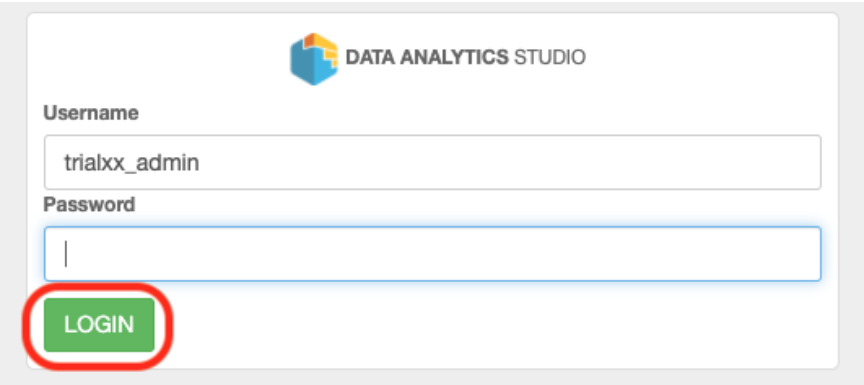
*The same steps you did in Part 2 to Open DAS



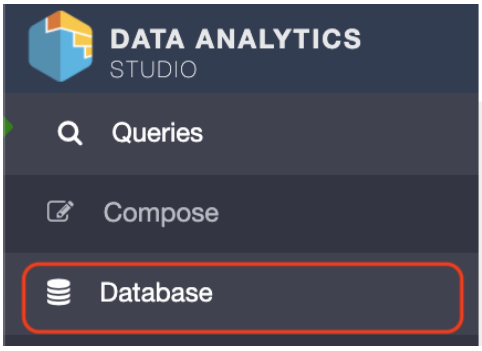
4) Enter the login information from step #1 above using the user, then click “LOGIN”

*You’ll likely already be authenticated from Part 2, you may not need to enter credentials

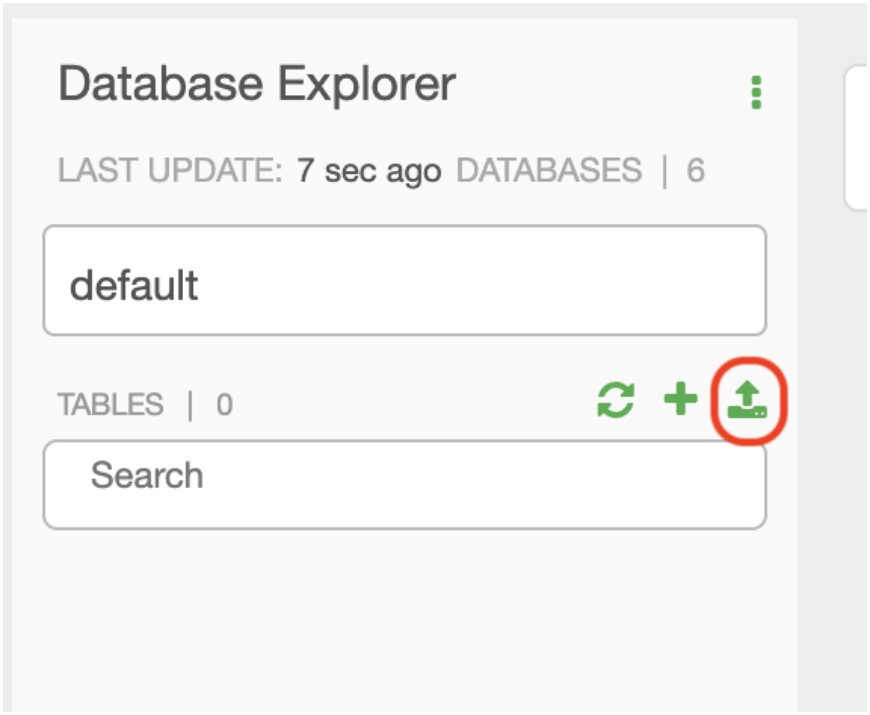
*Changing “trialxx_admin” to the trail user you’re using and password defined by “X” in #1 above



5) Click on Database on the left navigation bar

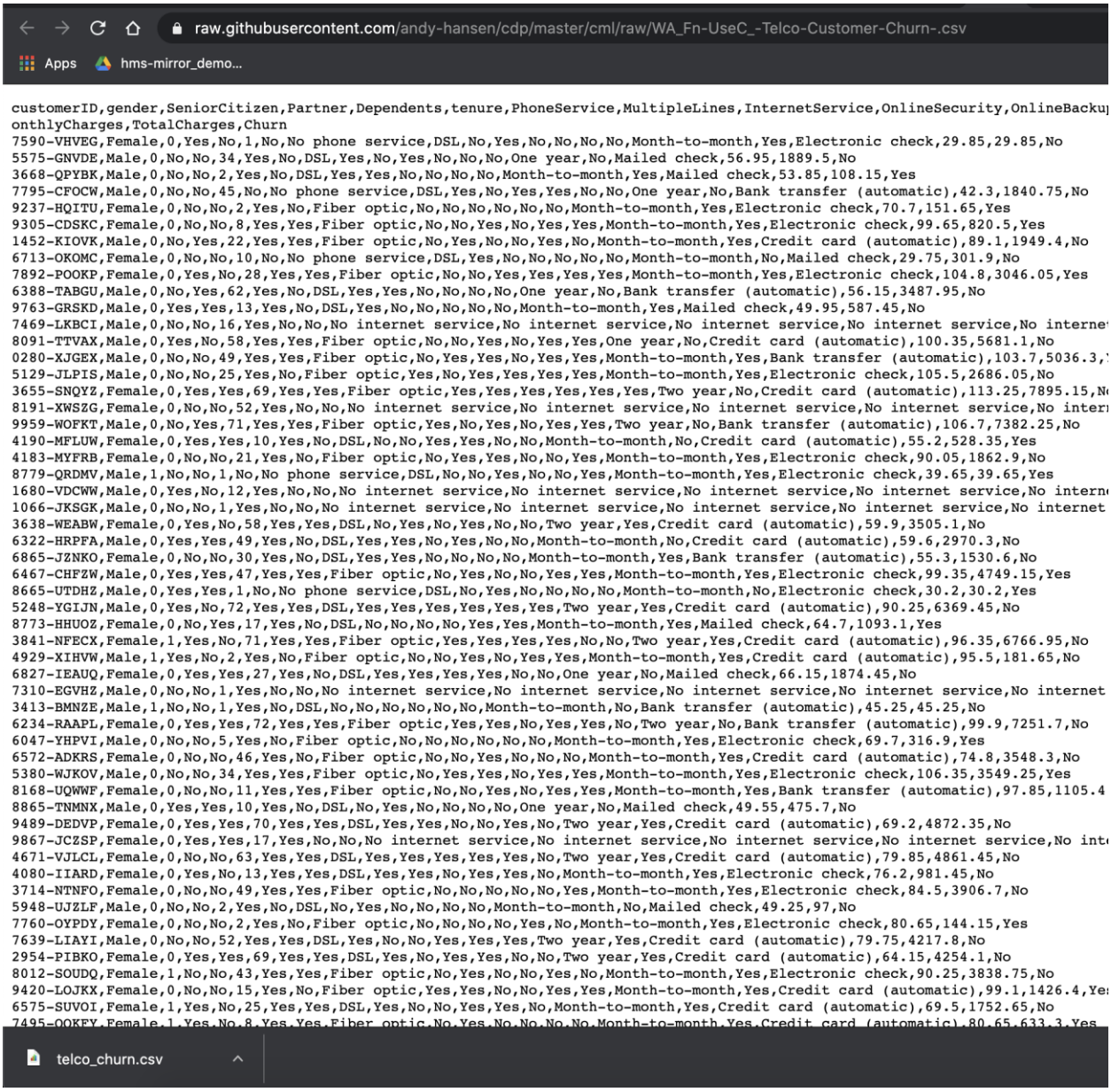


6) Click on “Upload Table”, using the “default” database

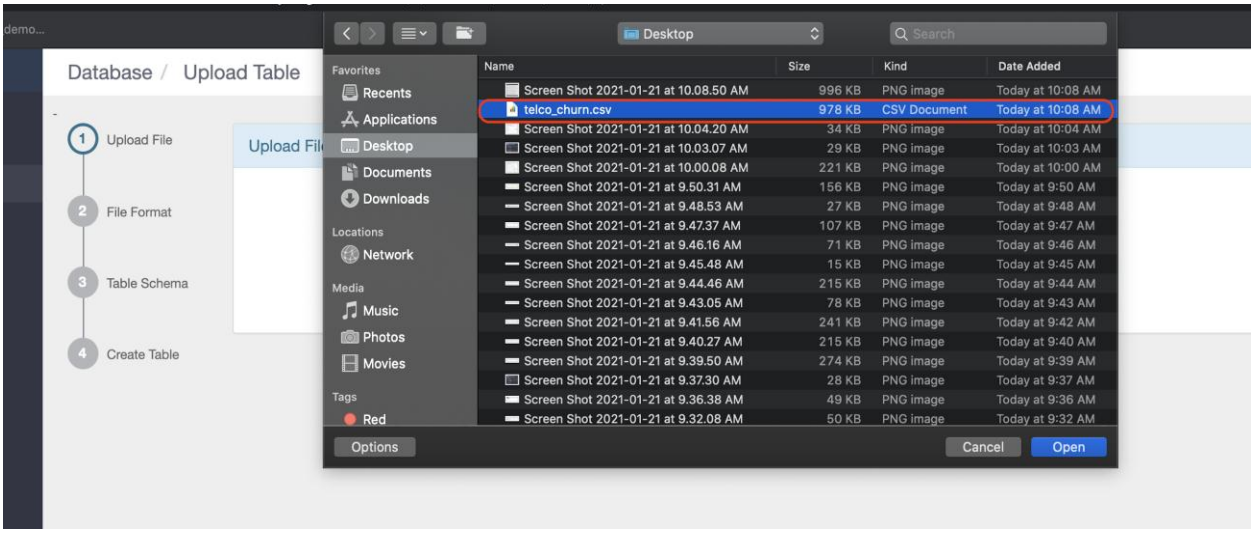


7) In a new browser window or tab, download the CSV file, saving to your desktop as “telco_churn.csv”

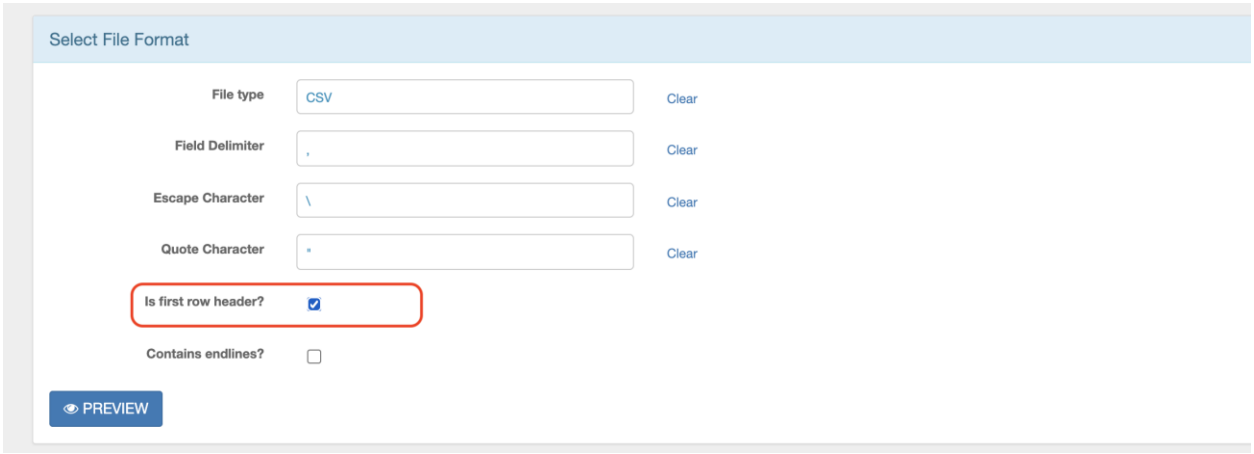
https://raw.githubusercontent.com/andy-hansen/cdp/master/cml/raw/WA_Fn-UseC_-Telco-Customer-Churn-.csv



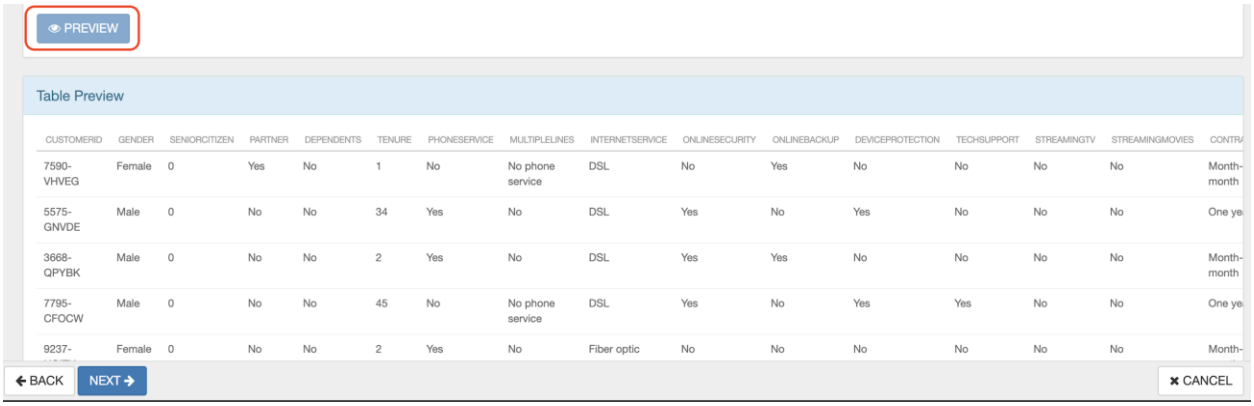
8) Going back to the window from step 6 above, upload the file “telco_churn.csv”



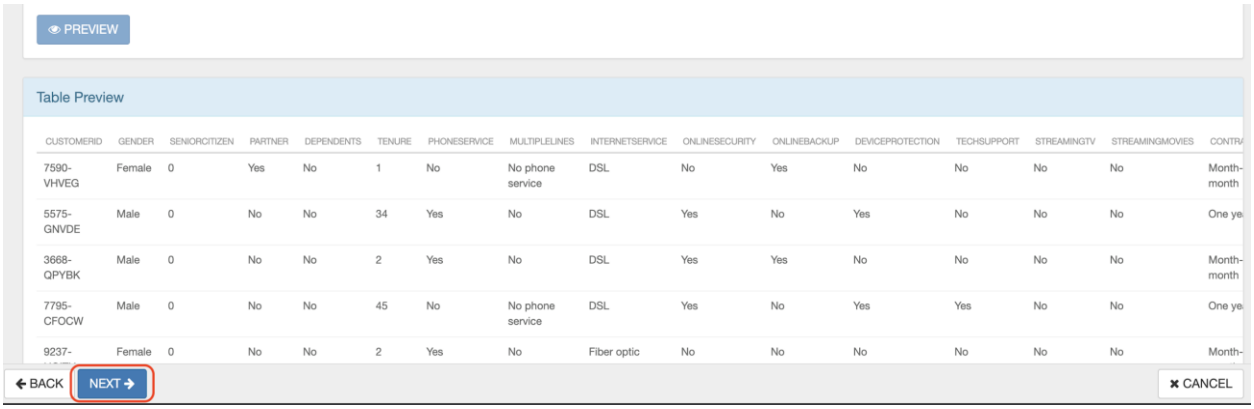
9) Click the “Is first row header?”, since the first row is a header



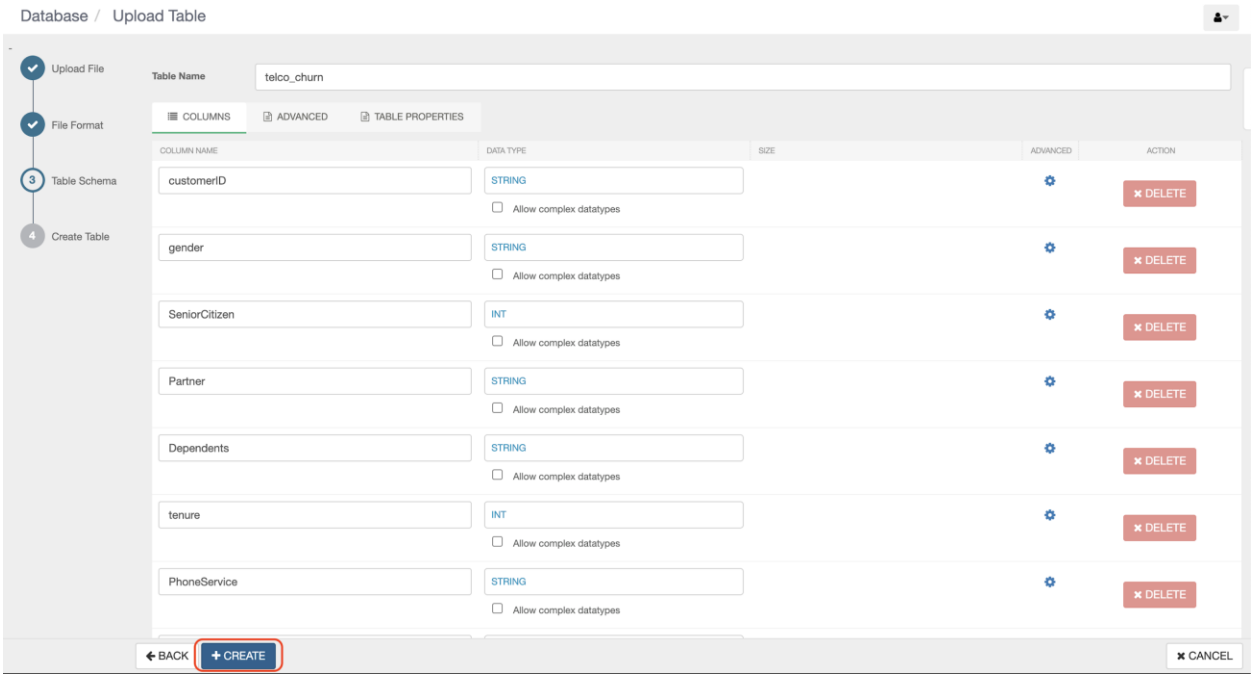
10) Click “PREVIEW” prior to creating the table



11) Click “NEXT”

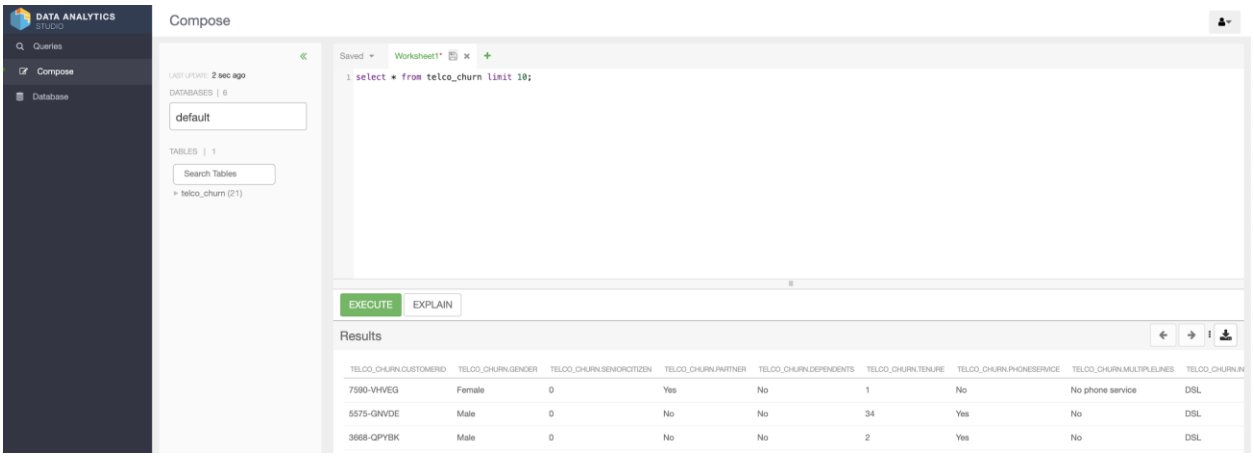


12) Enter ‘telco_churn’ as the Table Name. Click “CREATE”



13) Wait for about 2 minutes then Go-to “Compose” and within “Worksheet 1” run the following query on the new table

```
select * from telco_churn limit 10;
```



Parking lot items

- Show Impala/Hue - for CDH customers
- Add - Result Cache explanation
- Import CML github csv file