

**Agenda (Day 1):**

- a) CDP Overview (25 minutes)
- b) CDW Overview and Benefits (15 minutes)
- c) Hands-on (step-by-step below) (105 minutes[1 hour and 45 minutes] plus 5-10 minute break)
  - i) Part 1 - Data Catalog (20 minutes)
  - ii) Part 2 - Create a Virtual Warehouse and Run Queries (45 minutes)
  - iii) Part 3 - Data Visualization (25 minutes)
  - iv) Part 4 - Import a File into a Table (15 minutes)
- d) Q&A/Wrap-up (10-15 minutes)

## **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/realm/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx\\_admin@trycdp.com&p=X](http://login.trycdp.com/auth/realm/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 click “flights” under suggestions

Data Catalog / Search

The screenshot shows the Data Catalog / Search interface. At the top, there is a search bar with the query "flights". Below the search bar, the results are categorized into "Entities" and "Suggestions".

**Entities**

- actualelapsedtime (hive\_column)
- securitydelay (hive\_column)
- year (hive\_column)
- cancelled (hive\_column)
- weatherdelay (hive\_column)

**Suggestions**

- flights

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

The screenshot shows the Data Catalog / Search interface with filters applied. The "Filters" section on the left has "TYPE" selected, with "Hive Table" checked. This has narrowed the results to two entries in the main table.

Type	Name	Location
Hive Table	flights	/airlines_new_orc
Hive Table	flights	/airlines_new_parquet

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

5) Click “flights” where the Location = /airlines\_new\_orc

The screenshot shows a search interface for "flights". In the results table, there are two entries for "flights": one with a location of "/airlines\_new\_orc" and another with "/airlines\_new\_parquet". The entry with the location "/airlines\_new\_orc" is highlighted with a red box.

Type	Name	Location
Hive Table	<b>flights</b>	/airlines_new_orc
Hive Table	flights	/airlines_new_parquet

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

The screenshot shows the "Asset Details" page for the "flights" table. It includes sections for Overview, Schema, Metadata Audits, Policy, Access Audits, Asset Properties, Managed Classifications, and Lineage.

**Asset Properties:**

- Owner: csso\_trialuser31
- Qualified Name: airlines\_new\_orc.flights@cm
- Created On: Wed Jan 13 2021 01:10:57 GMT-0600 (Central Standard Time)
- Last Access Time: Wed Jan 13 2021 01:10:57 GMT-0600 (Central Standard Time)

**Managed Classifications:**

- Table Type: MANAGED\_TABLE
- Database: airlines\_new\_orc
- DB Catalog: cm
- Parent: airlines\_new\_orc

**Lineage:**

A lineage diagram shows the "flights" table connected to several source tables in the "/cdp-lake/data/airlines" directory. One specific node in this lineage is highlighted with a red box and has an "i" icon, indicating it can be clicked for more information.

**Managed Classification Detail:**

Details for the classification entry: /cdp-lake/data/airlines/airlines\_new\_orc.db/flights/year=1997

- Guid: be5a2094-f572-432e-9fa4-95498f7db550
- Type Name: aws\_s3\_pseudo\_dir
- Classifications(0): -
- Owner: -NA-
- Qualified Name: s3a://prod-cdptrialuser31-trycdp-com/cdp-lake/data/airlines/airlines\_new\_orc.db/flights/year=1997@cm
- Created On: -NA-
- Update Time: -NA-
- Created By: csso\_trialuser31
- Updated By: csso\_trialuser31

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

\*xx represents the trial user #

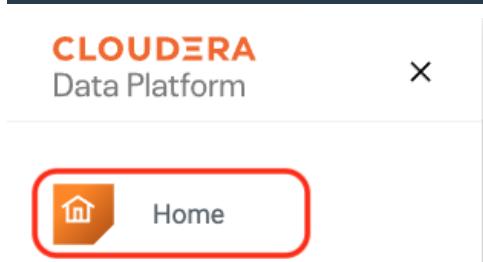
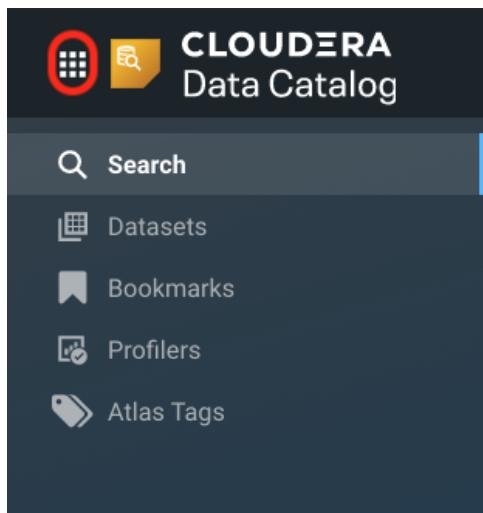
\*X represents the password

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 the “+” at the top right next to “Virtual Warehouses”

The screenshot shows the 'Virtual Warehouses' page with a count of 1. At the top right, there is a search icon and a red-circled '+' button. Below the header, a modal window titled 'New Virtual Warehouse' is open. It contains fields for 'Name' (with placeholder 'Enter Virtual Warehouse Name'), 'Type' (radio buttons for 'HIVE' and 'IMPALA' with 'HIVE' selected), 'Database Catalog' (dropdown set to 'cdptrialuser24-dl-default'), and 'Size' (dropdown set to '-- select an option --'). Below the modal, a list shows an existing virtual warehouse named 'default-vw' with status 'Stopped', compute node 'compute-1611103491-4hbp', and database catalog 'cdptrialuser24-dl-default'. A table at the bottom provides resource details: NODE COUNT 0, TOTAL CORES 12, TOTAL MEMORY 56 GB, and TYPE HIVE COMPACTOR.

4) Enter a name for your New Virtual Warehouse

The screenshot shows the 'Virtual Warehouses' page with a count of 1. The 'New Virtual Warehouse' modal is open, and the 'Name' field has been filled with 'testvirtualwarehouse1', which is highlighted with a red rectangle. The other fields in the modal are identical to the previous screenshot: Type (HIVE selected), Database Catalog (cdptrialuser24-dl-default), and Size (select an option).

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

\*How do I choose a size? Initial concurrent users

Virtual Warehouses | 1

New Virtual Warehouse

Name \*

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) 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

Name \*

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) Choose “Install Data Visualization” to be on

\*Allowing for Data Visualizations in Part 3

Virtual Warehouses | 2

New Virtual Warehouse

Name \*

Type \*

HIVE IMPALA

Database Catalog \*

cdptrialuser24-dl-default

Size \*

xsmall - 2 Executor Nodes

AutoSuspend Timeout (in seconds): 5000

Concurrency Autoscaling ⓘ

Nodes: Min:2, Max:6

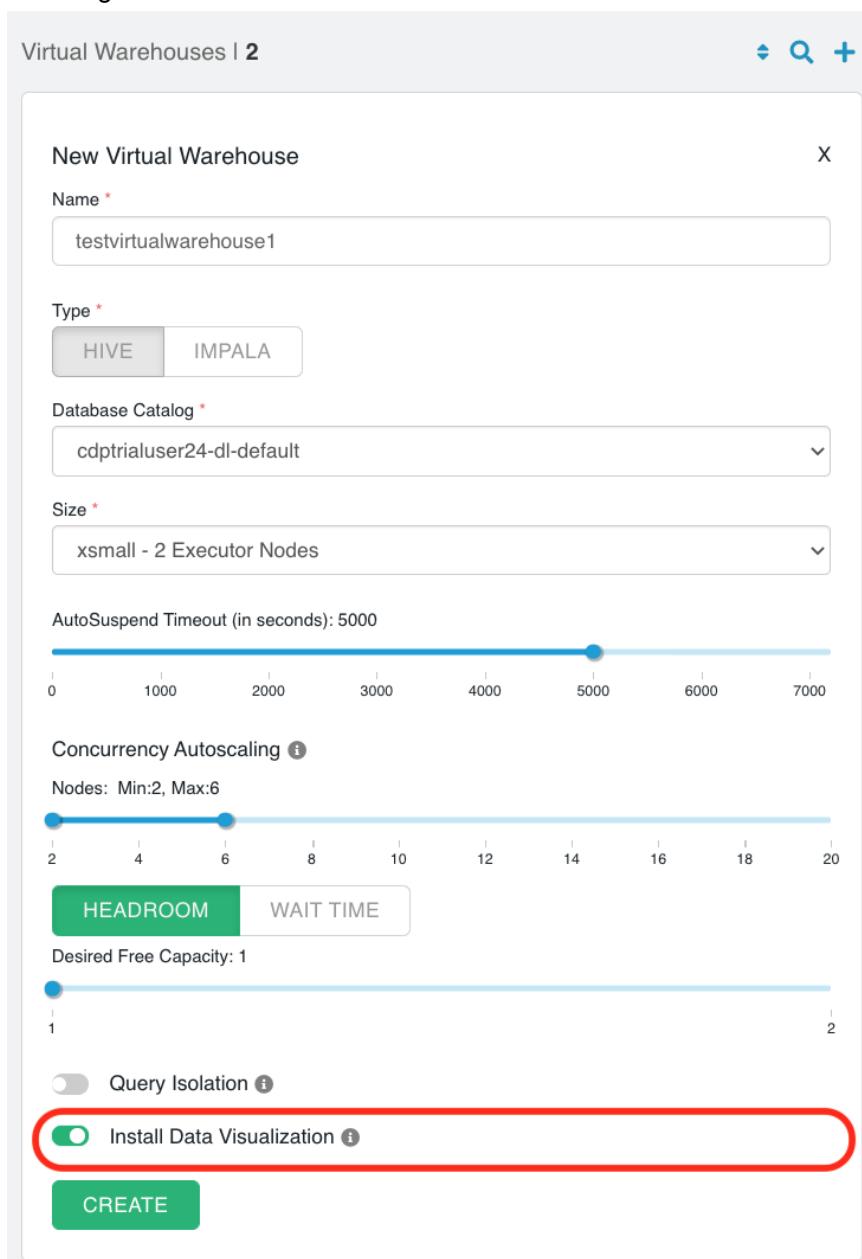
HEADROOM WAIT TIME

Desired Free Capacity: 1

Query Isolation ⓘ

Install Data Visualization ⓘ

**CREATE**



8) Click “Create” to create your Virtual Warehouse

\*Allow for 2.5 to 3 minutes for your Virtual Warehouse to become available for use



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

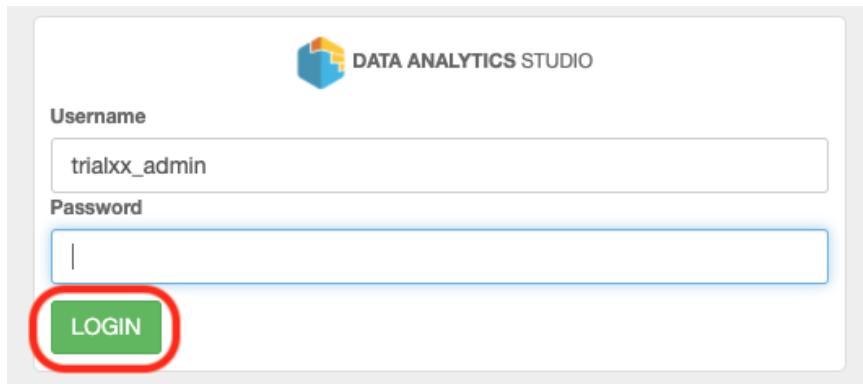
The screenshot shows the details of a virtual warehouse named "testvirtualwarehouse1". The status is currently "Starting". The configuration includes a single node labeled "compute-1611179792-vz49" and a database named "cdptrialuser24-dl-default". A progress bar at the bottom indicates the status. Below the progress bar, the warehouse's specifications are listed: NODE COUNT (2), TOTAL CORES (38), TOTAL MEMORY (292 GB), and TYPE (HIVE, DATA VISUALIZATION).

The screenshot shows the same virtual warehouse "testvirtualwarehouse1" but with the status now set to "Running". The configuration and specifications remain the same as in the previous screenshot. The progress bar is now fully green, indicating the warehouse is fully operational.

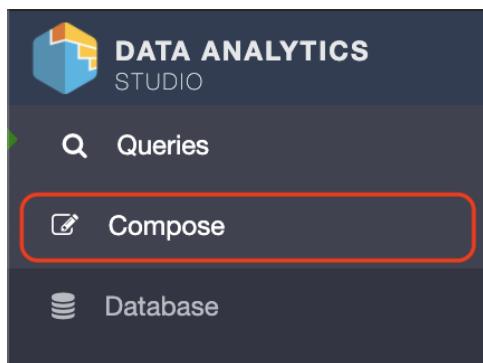
9) Once your Virtual Warehouse is “Running”, click the line in the top right and then click “Open DAS”

The screenshot displays a list of virtual warehouses. The first entry is "testvirtualwarehouse1", which is currently "Running". The second entry is "mschoeni-iso-1", which is "Stopped". To the right of the "testvirtualwarehouse1" row, a context menu is open, showing various options like Suspend, Clone, Edit, Delete, Upgrade, Copy JDBC URL, Download JDBC Jar, and Open DAS. The "Open DAS" option is highlighted with a red box.

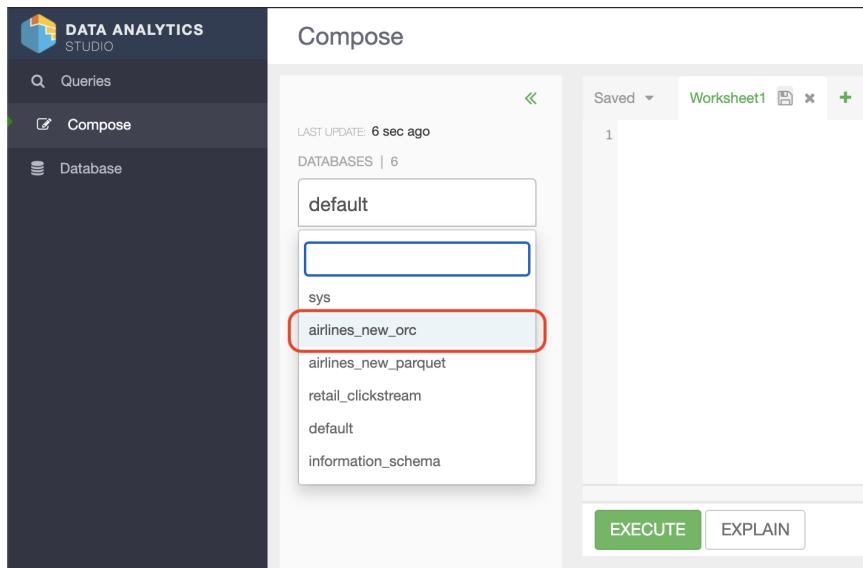
- 10) Enter the login information from step #1 above using the user, then click “LOGIN”  
\*Changing “trialxx\_admin” to the trail user you’re using and password defined by “X” in #1 above



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

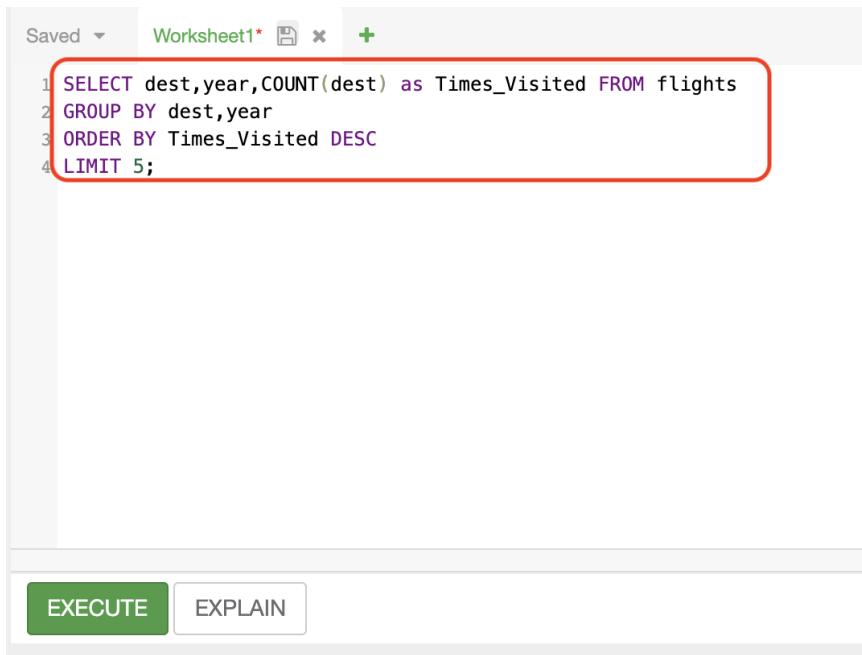


- 12) Choose the database “airlines\_new\_orc” that we found in Part 1 “Data Catalog”



- 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;
```



The screenshot shows a database worksheet window titled "Worksheet1". The query code is displayed in a text area, with the entire code block highlighted by a red rectangular box. Below the text area are two buttons: "EXECUTE" (in green) and "EXPLAIN" (in white).

```
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;
```

EXECUTE EXPLAIN

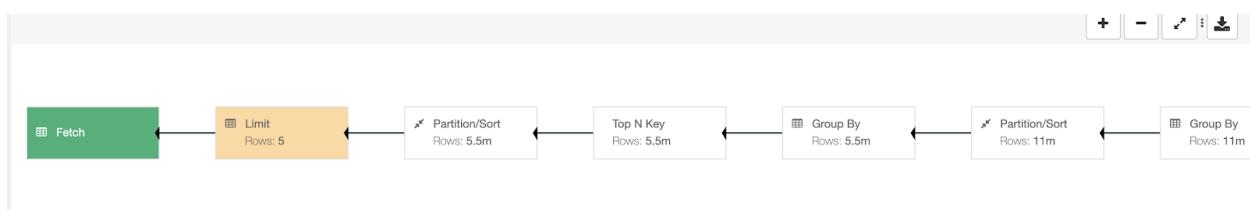
14) Click “EXPLAIN” to see the visual explain plan prior to running the query

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

Saved ▾ Worksheet1\*

```
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;
```

**EXECUTE** **EXPLAIN**



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

The screenshot shows a database worksheet titled "Worksheet1". At the top, there are tabs for "Saved", "Worksheet1\*", a file icon, and a close button. Below the tabs is a code editor containing the following 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;
```

At the bottom of the worksheet, there are two buttons: "EXECUTE" (highlighted with a red box) and "EXPLAIN".

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

The screenshot shows a results table titled "Results". The table has three columns: DEST, YEAR, and TIMES\_VISITED. The data is as follows:

DEST	YEAR	TIMES_VISITED
ATL	2005	429800
ATL	2004	416989
ATL	2008	414521
ATL	2007	413805
ATL	2006	404829

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

The screenshot shows a database worksheet titled "Worksheet1". At the top, there are tabs for "Saved", "Worksheet1\*" (highlighted with a red box), a file icon, and a close button. A new tab labeled "+" is also visible. Below the tabs is a code editor containing the same SQL query as in step 15:

```
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;
```

At the bottom of the worksheet, there are two buttons: "EXECUTE" and "EXPLAIN".

- 18) In “Worksheet 2”, 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;
```

The screenshot shows a database interface with two tabs: 'Worksheet1\*' and 'Worksheet2\*'. The 'Worksheet2\*' tab is active and contains the SQL query from the previous step. The 'EXECUTE' button is visible at the bottom left of the query editor.

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

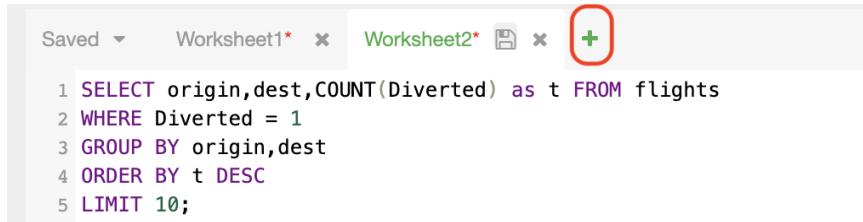
- 19) Click “EXECUTE” to execute the query, answering the question “What are the top 10 routes (origin and dest) that have seen maximum diversions?”

\*Change the database to use “airlines\_new\_orc” prior to executing the query

The screenshot shows the results of the executed query. The 'EXECUTE' button is highlighted with a red box. The results are displayed in a table titled 'Results'.

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



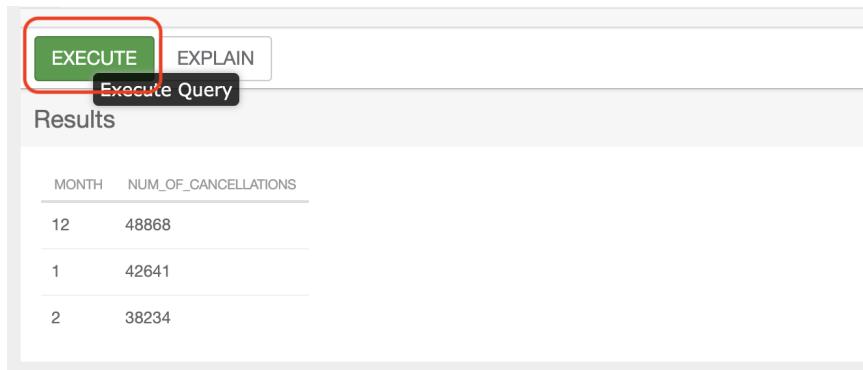
```
1 SELECT origin,dest,COUNT(Diverted) as t FROM flights
2 WHERE Diverted = 1
3 GROUP BY origin,dest
4 ORDER BY t DESC
5 LIMIT 10;
```

21) In “Worksheet 3”, 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?”

\*Change the database to use “airlines\_new\_orc” prior to executing the query

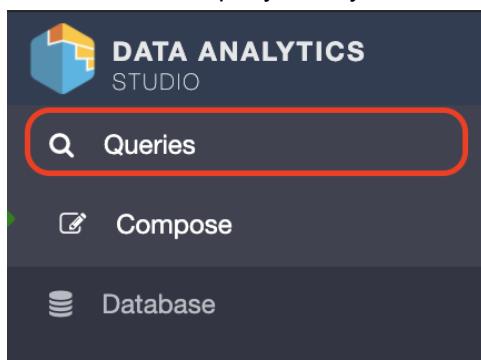


MONTH	NUM_OF_CANCELLATIONS
12	48868
1	42641
2	38234

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)

QUERIES (159)										COMPOSE QUERY
QUERY	STATUS	QUEUE	USER	TABLES READ	TABLES WRITTEN	START TIME	DURATION	DAG ID	ACTIONS	
✓ SELECT month,COUNT(Cancelled) as n...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	8 seconds ago	00:00:00	Not Available!		
✓ SELECT month,COUNT(Cancelled) as n...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	4 minutes ago	00:00:02	dag_161123649		
✓ SELECT origin,dest,COUNT(Diverted) as...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	8 minutes ago	00:00:03	dag_161123649		
✓ SELECT dest,year,COUNT(dest) as Time...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	20 minutes ago	00:00:00	Not Available!		
✓ SELECT dest,year,COUNT(dest) as Time...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	21 minutes ago	00:00:00	Not Available!		

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

QUERIES (159)										COMPOSE QUERY
QUERY	STATUS	QUEUE	USER	TABLES READ	TABLES WRITTEN	START TIME	DURATION	DAG ID	ACTIONS	
✓ SELECT month,COUNT(Cancelled) as n...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	4 minutes ago	00:00:00	Not Available!		
✓ SELECT month,COUNT(Cancelled) as n...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	8 minutes ago	00:00:02	dag_161123649		
✓ SELECT origin,dest,COUNT(Diverted) as...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	13 minutes ago	00:00:03	dag_161123649		
✓ SELECT dest,year,COUNT(dest) as Time...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	24 minutes ago	00:00:00	Not Available!		
✓ SELECT dest,year,COUNT(dest) as Time...	SUCCESS	None	trial24_admin	flights (airlines_new_o...	Not Available!	25 minutes ago	00:00:00	Not Available!		

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

Queries

```
SELECT month,COUNT(Cancelled) as num_of_cancellations
```

```
SELECT month,COUNT(Cancelled) as num_of_cancellations
```

COMPARE
  
Compare two queries

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

#### Query Details - A

QUERY ID  
hive\_20210121153926\_e3a56b9d-71f2-45dc-b23e-2c2e1146d61e

USER  
trial24\_admin

STATUS  
 **SUCCESS**

START TIME  
21 Jan 2021 09:39:26

END TIME  
21 Jan 2021 09:39:26

DURATION  
118ms

#### Query Details - B

QUERY ID  
hive\_20210121153513\_37aefec1-0284-4897-bfbb-bf9bb5797252

USER  
trial24\_admin

STATUS  
 **SUCCESS**

START TIME  
21 Jan 2021 09:35:13

END TIME  
21 Jan 2021 09:35:15

DURATION  
2s 311ms

## 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.



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

\*xx represents the trial user #

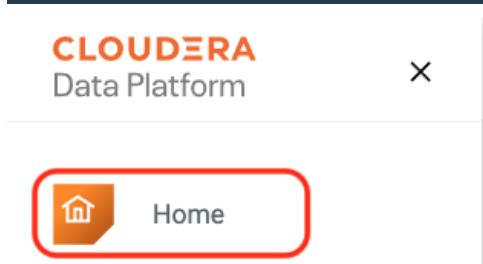
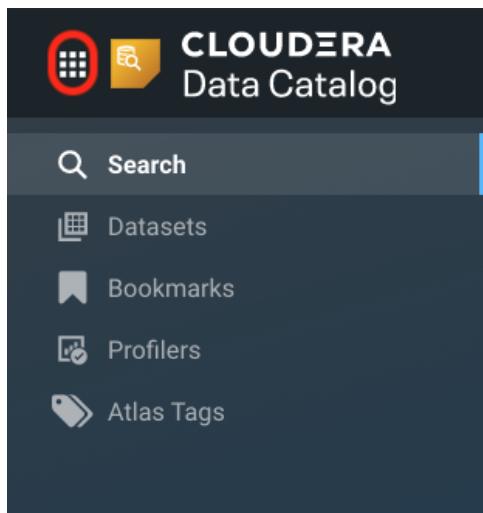
\*X represents the password

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 Data Visualization” on your existing “Running” Virtual Warehouse

The screenshot shows the Cloudera Manager interface for managing virtual warehouses. At the top, there's a search bar and a '+' button. Below it, a table lists virtual warehouses:

	Name	Status	Compute Nodes	Total Cores	Total Memory	Type
	testvirtualwarehouse1	Running	compute-1611179792-vz49	2	38	292 GB HIVE DA
	mschoeni-iso-1	Stopped	compute-1611173596-dbtv	2	38	292 GB DA

A context menu is open over the first row (testvirtualwarehouse1), listing options: Suspend, Clone, Edit, Delete, Upgrade, Copy JDBC URL, Download JDBC Jar, Open DAS, and Open Data Visualization. The 'Open Data Visualization' option is highlighted with a red circle.

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

The screenshot shows the Cloudera Data Visualization login interface. It features a dark header with the text "CLOUDERA Data Visualization". Below it is a light-colored login form with the following fields and elements:

- LOGIN** (Section title)
- Username**: Input field containing "trialxx\_admin".
- Password**: Input field (empty).
- Invalid login**: Error message displayed below the password field.
- Forgot your password?**: Link to password recovery.
- Remember me on this computer**: Checkbox.
- LOGIN** (Large orange button at the bottom, highlighted with a red circle).

5) Click “DATA” the top navigation bar

The screenshot shows the Cloudera Data Visualization interface. The top navigation bar has tabs for HOME, VISUALS, and DATA, with the DATA tab highlighted by a red circle. On the left, there's a sidebar with 'All Connections' (Default Hive VW, samples), a 'Datasets' section (1 item), and a search bar. The main area is titled 'Title/Table' and 'Created'.

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

The screenshot shows the same interface as above, but the 'Default Hive VW' connection in the sidebar is highlighted with a red circle. The main area still shows the 'Title/Table' and 'Created' columns.

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

The screenshot shows the interface again, with the 'Default Hive VW' connection selected. The 'NEW DATASET' button in the top right is highlighted with a red circle. The main area remains the same with 'Title/Table' and 'Created' columns.

8) Enter a name for the Dataset title naming “airline\_new\_orc.flights”

\*Can be any name you choose

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 \*

 airlines\_new\_orc.flights

Dataset Source

From Table

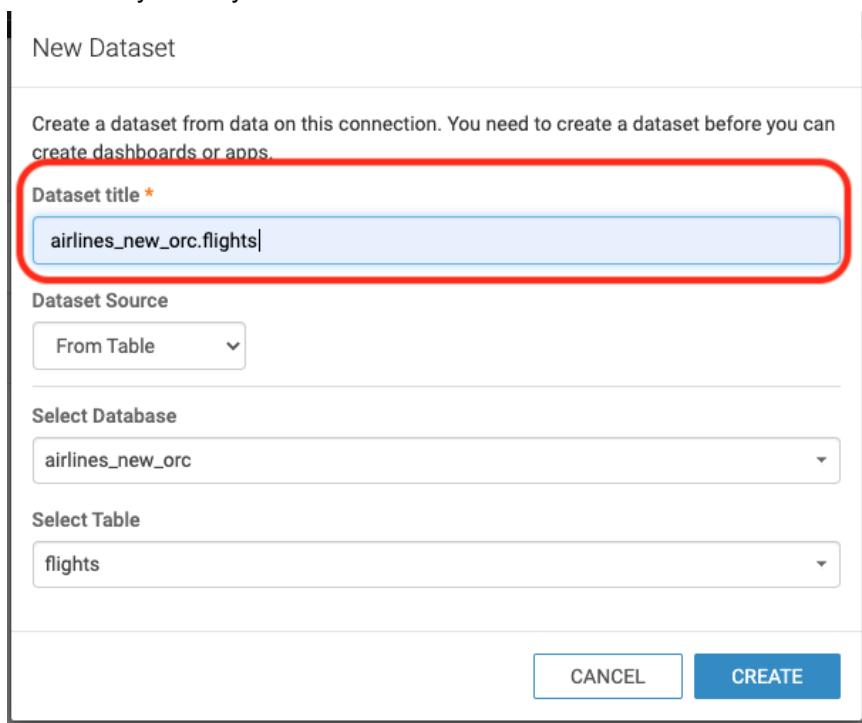
Select Database

airlines\_new\_orc

Select Table

flights

CANCEL CREATE



9) Choose the database “airlines\_new\_orc”

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 \*

 airlines\_new\_orc.flights

Dataset Source

From Table

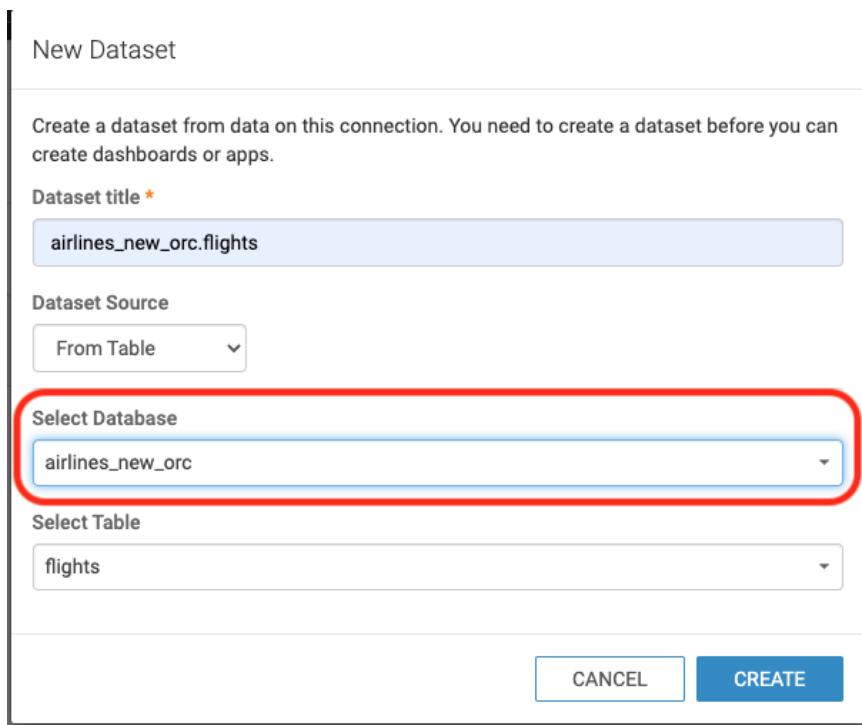
Select Database

airlines\_new\_orc

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

Select Table

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

Select Table

CANCEL

CREATE

12) Click “+” to create a New Dashboard

The screenshot shows a list of dashboards. At the top, there are buttons for 'NEW DATASET', 'ADD DATA', and a dropdown. Below this is a search bar and a filter for 'Datasets'. A 'New Dashboard' button is highlighted with a red circle. The list includes two items: 'airlines\_new.orc.flights' and 'airlines\_new.orc.flights'. Each item has columns for 'Created', 'Last Updated', 'Modified By', and '# Visuals'.

Title/Table	Created	Last Updated	Modified By	# Visuals
airlines_new.orc.flights	Jan 21, 2021	a few seconds ago	trial24_admin	0
airlines_new.orc.flights				

13) Choose “Treemap” under “VISUALS”

The screenshot shows the 'Dashboard Designer' interface. On the left, the 'VISUALS' section contains a 'Table' icon and a grid of visual icons, with 'Treemap' highlighted with a red circle. Below this are sections for 'Dimensions', 'Measures', 'Toolips', 'Filters', and a 'Limit' input field set to 100. At the bottom is a 'REFRESH VISUAL' button. On the right, the 'DATA' section lists datasets ('airlines\_new.orc.flights') and dimensions ('flights', 'uniquecarrier', 'tailnum', 'origin', 'dest', 'cancellationcode', 'diverted'). Below this is the 'Measures' section with 24 items including 'Record Count', 'month', 'dayofmonth', etc. A sidebar on the right lists 'DASH.', 'Visuals', 'Filters', 'Settings', 'Style', 'VISUAL', 'Build', 'Settings', and 'Style'.

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

**Dashboard Designer**

VISUALS	DATA
Table	airlines_new_orc.flights <span style="color: blue;">edit</span> <span style="color: red;">x</span>
1234 LABEL	Sample Mode: OFF
	Search <span style="color: red;">x</span>
Dimensions	<b>Dimensions</b> 6
dest	flights
origin	A uniquecarrier
	A tailnum
	A origin
	A dest
	A cancellationcode
	A diverted
Measures	<b>Measures</b> 24
drag fields to add here	flights
	# Record Count
	# month
	# dayofmonth
	# dayofweek
	# deptime
	# crsdeptime
	# arrtime
	# crsarrtime
	# flightnum
Tooltips	
drag fields to add here	
Filters	
drag fields to add here	
Limit: 100	
<span style="background-color: #0078D4; color: white; padding: 5px 10px;">REFRESH VISUAL</span>	

**DASH.**

- Visuals
- Filters
- Settings

**Favorites**

- AirDrop
- Recents
- Application
- Desktop
- Documents
- Downloads

**Locations**

- Network

**Tags**

- Red
- Orange
- Yellow
- Green
- Blue

**VISUAL**

- Build
- Settings
- Style

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

The screenshot shows the Tableau Dashboard Designer interface. On the left, the 'VISUALS' pane contains a 'Table' visualization. Below it, the 'Dimensions' section lists 'dest' and 'origin'. The 'Measures' section is highlighted with a red box and shows 'sum(1)' above '# Record Count'. To the right, the 'DATA' pane displays a connection to 'airlines\_new\_orc.flights' with 'Sample Mode: OFF'. A search bar and a 'Dimensions' section (containing 'flights' and several flight-related fields) are also visible. The 'Measures' section in the DATA pane lists 24 measures, including '# Record Count' at the top. The right side of the screen features a sidebar with 'Favorites', 'Filters', 'Settings', 'Style', 'Build' (which is selected), and 'Tags' (with color-coded entries for Red, Orange, Yellow, Green, and Blue).

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

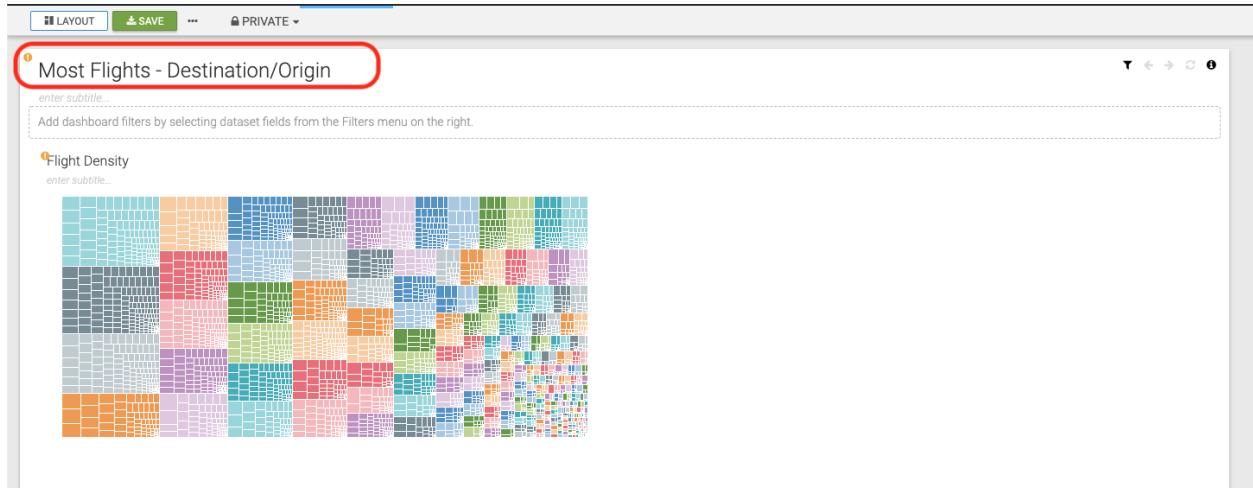
The screenshot shows the Tableau Dashboard Designer interface. On the left, the Visuals shelf has 'Treemap' selected. Below it, under Dimensions, are 'dest' and 'origin'. Under Measure, 'Record Count' is selected and has a blue border. To the right is the 'FIELD PROPERTIES' pane, which is open to the 'Order and Top K' section. This section contains two options: 'Descending' (selected, indicated by a green checkmark) and 'Ascending'. Below these are input fields for 'Top K:' (set to 'eg. 100') and 'Bottom K:' (set to 'eg. 100'). A red box highlights the 'Order and Top K' section. The right sidebar shows navigation links: DASH., Visuals, Filters, Settings, Style, and a 'Build' tab which is currently active. The 'Build' tab also includes 'Settings' and 'Style' sub-links.

17) Click "REFRESH VISUAL"

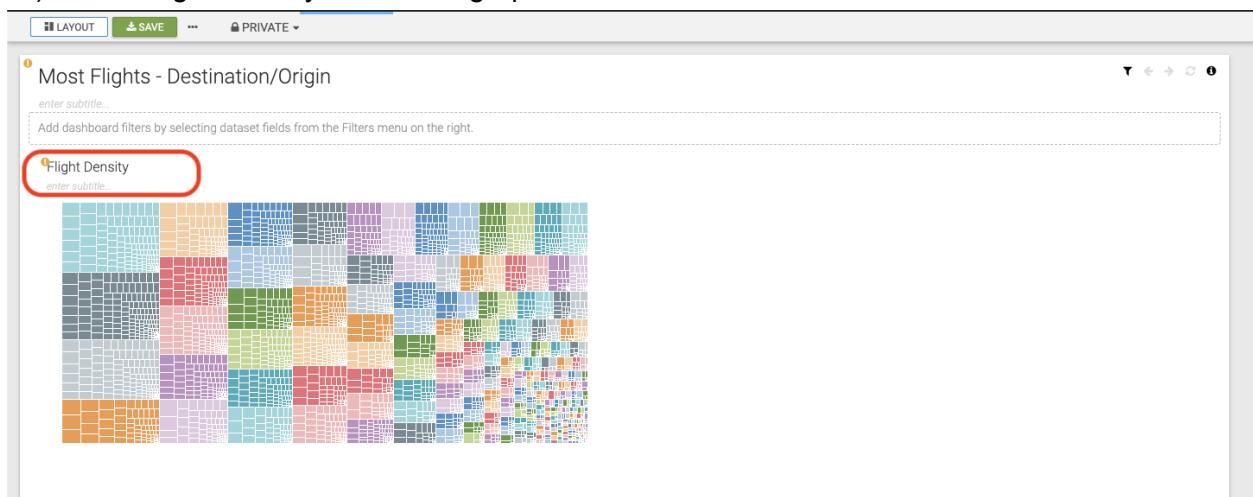
\*Notice - you can have other Visuals chosen to be displayed with the Dimensions and Measure(s), then click REFRESH VISUALS

The screenshot shows the Tableau Dashboard Designer interface. On the left, the **VISUALS** pane displays various visualization icons, with a Treemap icon selected. Below it, sections for **Dimensions** (dest, origin), **Measure** (# Record Count), **Tooltips**, **X Trellis**, **Y Trellis**, and **Filters** are shown. A red oval highlights the **REFRESH VISUAL** button at the bottom of this pane. In the center, the **DATA** pane shows a connection to **airlines\_new\_orc.flights** with **Sample Mode: OFF**. It lists **Dimensions** (flights, uniquecarrier, tailnum, origin, dest, cancellationcode, diverted) and **Measures** (flights, Record Count, month, dayofmonth, dayofweek, deptime, crsdeptime, arftime, crsarftime, flightnum, actualelapsedtime, crselapsedtime, airtime, arrdelay). On the right, the **BUILD** pane contains **Visuals**, **Filters**, **Settings**, **Style**, and **Build** sections. The **Build** section is currently active.

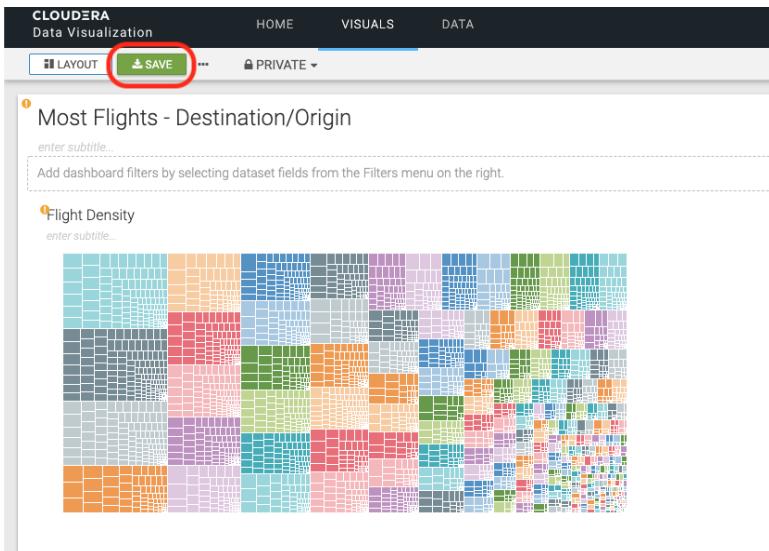
18) Enter a title “Most Flights - Destination/Origin”



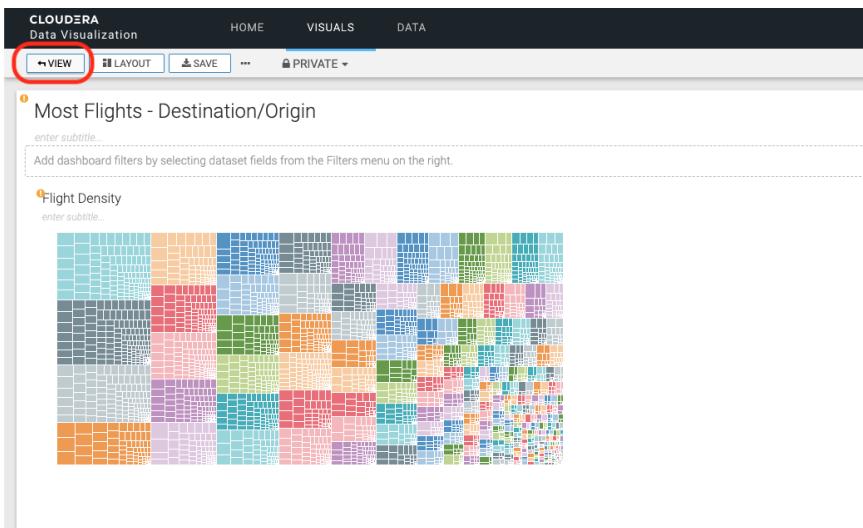
19) Enter “Flight Density” under the graph’s title



20) Click "SAVE"



21) Click "VIEW"

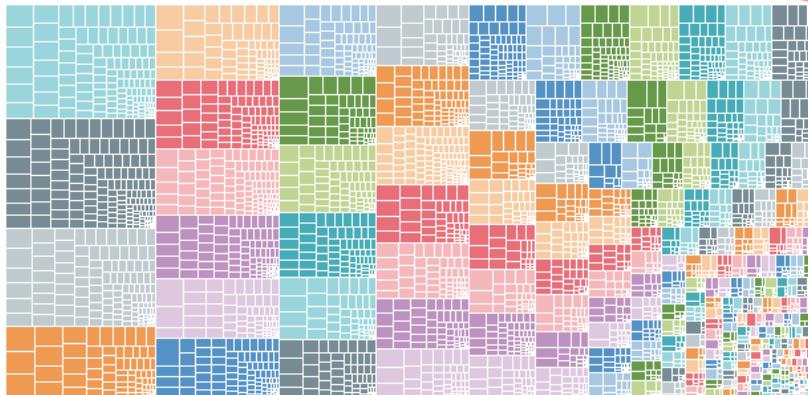


22) Scroll over the graph and click “Expand Visual”

### 1 Most Flights - Destination/Origin

#### Flight Density

Expand visual



Destinations are displayed



## 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.

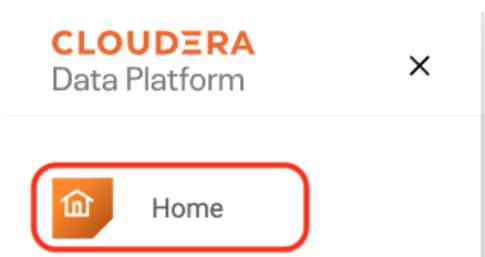
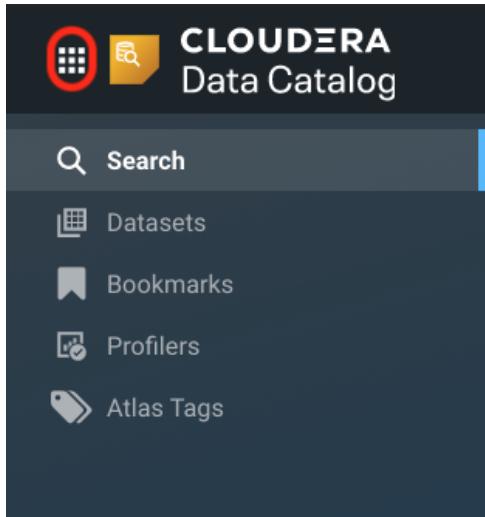
Your link should look something like (remember click the link in your email not the link below)  
[http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx\\_admin@trycdp.com&p=X](http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X)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

The screenshot shows the Cloudera Data Platform (CDP) interface for managing virtual warehouses. At the top, there's a search bar and a '+' button. Below it, a table lists three virtual warehouses:

	Name	Status	Compute Cluster	Database
	testvirtualwarehouse1	Running	compute-1611179792-vz49	cdptrialuser24-dl-default
	mschoeni-iso-1	Stopped	compute-1611173596-dbtv	cdptrialuser24-dl-default
	default-vw	Stopped	compute-1611103491-4hbp	

Below the table, there are columns for NODE COUNT, TOTAL CORES, and TOTAL MEMORY. A context menu is open over the first row (testvirtualwarehouse1), listing options: Suspend, Clone, Edit, Delete, Upgrade, Copy JDBC URL, Download JDBC Jar, Open DAS (which is highlighted with a red box), Open Data Visualization, Set Compactor, Run AutoScaling Demo, and Collect Diagnostic Bundle.

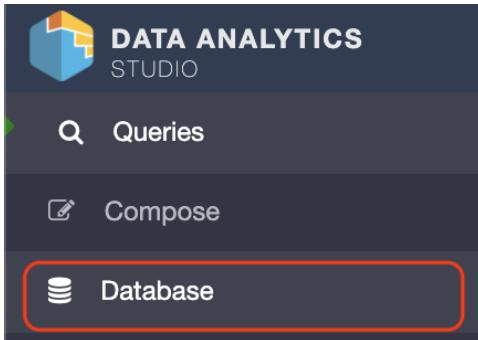
4) Enter the login information from step #1 above using the user, then click “LOGIN”

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

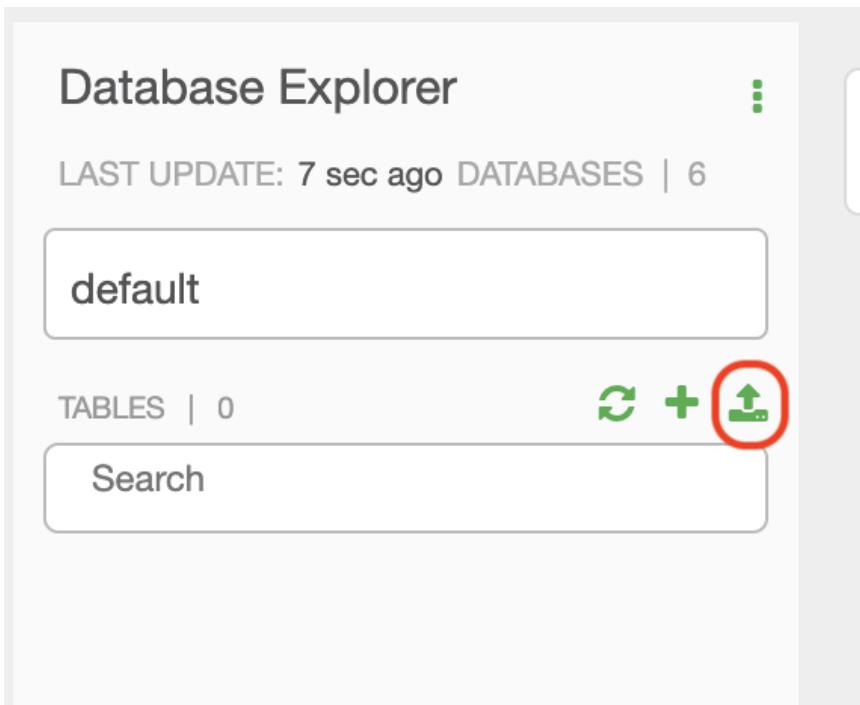
\*Changing “trialxx\_admin” to the trail user you’re using and password defined by “X” in #1 above

The screenshot shows the Data Analytics Studio login interface. It features a logo and the text "DATA ANALYTICS STUDIO". Below that are two input fields: "Username" containing "trialxx\_admin" and "Password" which is currently empty. At the bottom is a green "LOGIN" button, which is circled in red to indicate it should be clicked.

5) Click on Database on the left navigation bar

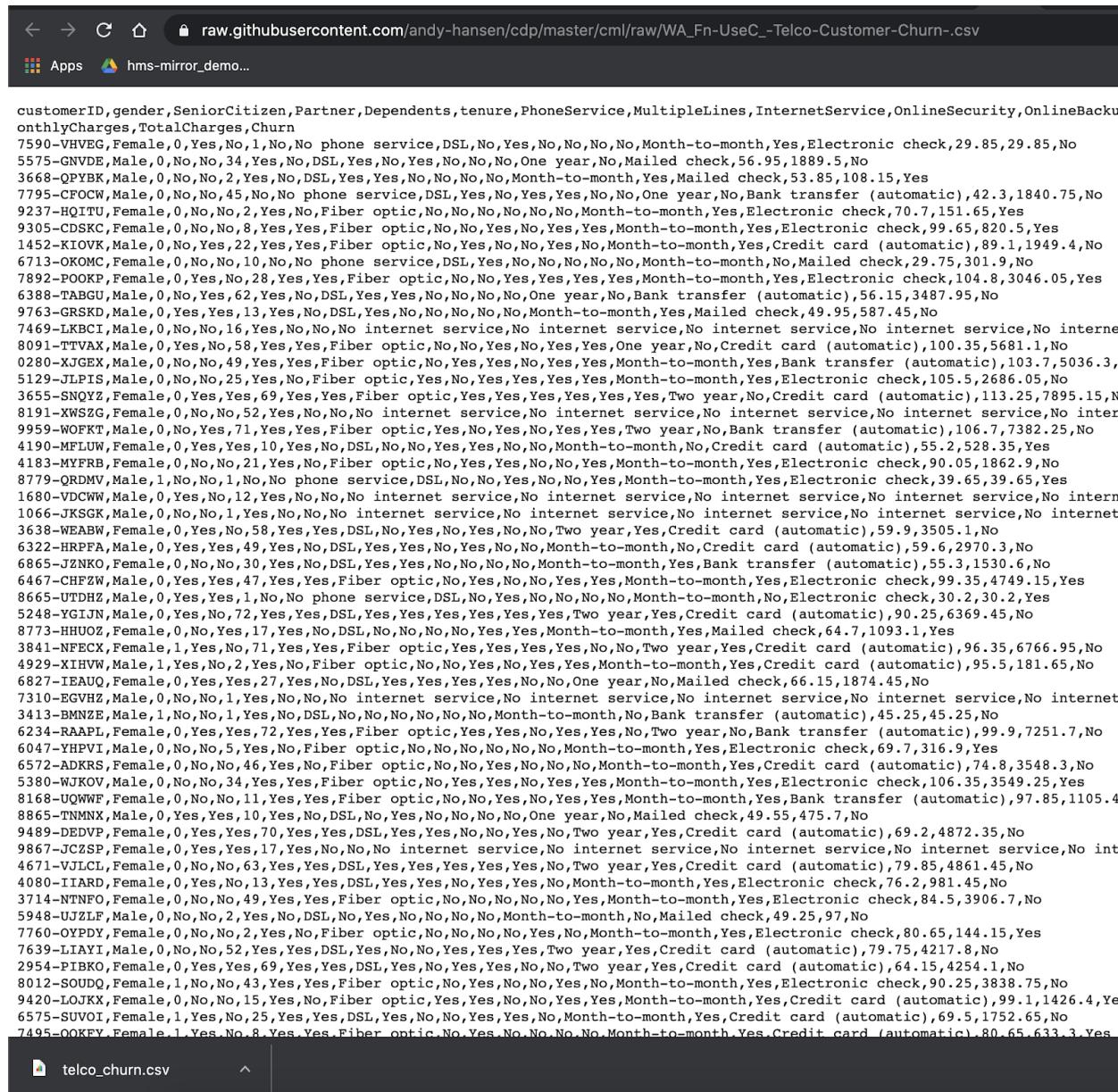


6) Click on “Upload Table”, using the “default” database



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



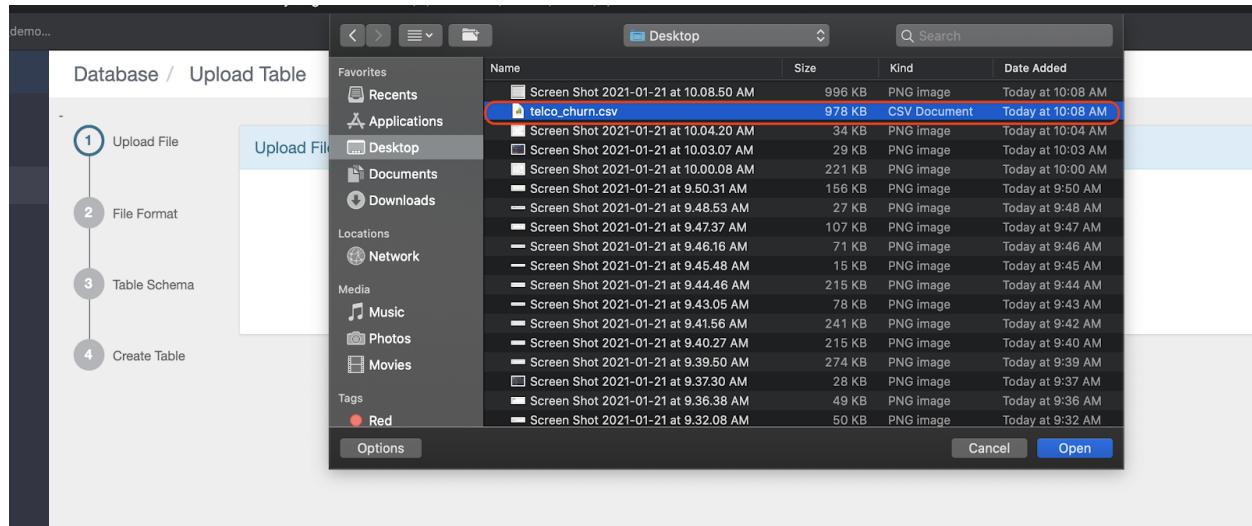
The screenshot shows a browser window with the URL [https://raw.githubusercontent.com/andy-hansen/cdp/master/cml/raw/WA\\_Fn-UseC\\_-Telco-Customer-Churn-.csv](https://raw.githubusercontent.com/andy-hansen/cdp/master/cml/raw/WA_Fn-UseC_-Telco-Customer-Churn-.csv). The page displays a large amount of comma-separated data, which is the CSV file content. The data includes various columns such as customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup,onthlyCharges, TotalCharges, Churn, and numerous other service-related variables. The browser interface shows standard navigation buttons (back, forward, home), a search bar, and a tab labeled "hms-mirror\_demo". At the bottom of the browser window, there is a file download dialog with the filename "telco\_churn.csv".

```

customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,onthlyCharges,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-GNVDE,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,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,No,45,No,No phone service,DSL,Yes,No,Yes,Yes,No,No,One year,No,Bank transfer (automatic),42.3,1840.75,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
9305-CDKCI,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-KIOVK,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-OOKMC,Female,0,No,No,10,No,No phone service,DSL,Yes,No,No,No,No,Month-to-month,No,Mailed check,29.75,301.9,No
7892-POOKP,Female,0,Yes,28,Yes,Yes,Fiber optic,No,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-GRSKD,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-LKBCI,Male,0,No,No,16,Yes,No,No,internet service,No,internet service,No,internet service,No,internet service,No,internet service
8091-TTVAX,Male,0,Yes,No,58,Yes,Yes,Fiber optic,No,No,Yes,No,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,
5129-JLPIS,Male,0,No,No,25,Yes,Yes,Fiber optic,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,105.5,2686.05,No
3655-SNQYZ,Female,0,Yes,69,Yes,Yes,Fiber optic,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,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
9959-WOFKT,Male,0,No,Yes,71,Yes,Yes,Fiber optic,Yes,No,Yes,No,Yes,Two year,No,Bank transfer (automatic),106.7,7382.25,No
4190-MFLUW,Female,0,Yes,Yes,10,Yes,No,DSL,No,No,Yes,Yes,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-QRDMV,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
1680-VDCWW,Male,0,Yes,No,12,Yes,No,No,No,internet service,No,internet service,No,internet service,No,internet service,No,internet service
1066-JKSGK,Male,0,No,No,1,Yes,No,No,internet service,No,internet service,No,internet service,No,internet service,No,internet service
3638-WEABW,Female,0,Yes,No,58,Yes,Yes,DSL,No,Yes,No,Yes,Two year,Yes,Credit card (automatic),59.9,3505.1,No
6322-HRFPA,Male,0,Yes,49,Yes,No,DSL,Yes,Yes,No,Yes,No,Month-to-month,No,Credit card (automatic),59.6,2970.3,No
6865-JZNKO,Female,0,No,No,30,Yes,No,DSL,Yes,Yes,No,No,No,Month-to-month,Yes,Bank transfer (automatic),55.3,1530.6,No
6467-CHFZW,Male,0,Yes,47,Yes,Yes,Fiber optic,No,Yes,No,Yes,Month-to-month,Yes,Electronic check,99.35,4749.15,Yes
8665-UTDHZ,Male,0,Yes,Yes,1,No,no phone service,DSL,No,Yes,No,No,No,Month-to-month,No,Electronic check,30.2,30.2,Yes
5248-YGJIN,Male,0,Yes,No,72,Yes,Yes,DSL,Yes,Yes,Yes,Yes,Two year,Yes,Credit card (automatic),90.25,6369.45,No
8773-HHUOZ,Female,0,No,Yes,17,Yes,No,DSL,No,No,No,Yes,Month-to-month,Yes,Mailed check,64.7,1093.1,Yes
3841-NFECX,Female,1,Yes,No,71,Yes,Yes,Fiber optic,Yes,Yes,Yes,Yes,No,Two year,Yes,Credit card (automatic),96.35,6766.95,No
4929-XIHFW,Male,1,Yes,No,2,Yes,No,Fiber optic,No,No,Yes,No,Yes,Month-to-month,Yes,Credit card (automatic),95.5,181.65,No
6827-IEAUQ,Female,0,Yes,27,Yes,No,DSL,Yes,Yes,Yes,No,One year,No,Mailed check,66.15,1874.45,No
7310-EGVHZ,Male,0,No,No,1,Yes,No,No,internet service,No,internet service,No,internet service,No,internet service,No,internet service
3413-BMNZE,Male,1,No,No,1,Yes,No,DSL,No,No,No,No,Month-to-month,No,Bank transfer (automatic),45.25,45.25,No
6234-RAAPL,Female,0,Yes,72,Yes,Yes,Fiber optic,Yes,Yes,No,Yes,Two year,No,Bank transfer (automatic),99.9,7251.7,No
6047-YHPVI,Male,0,No,5,Yes,No,Fiber optic,No,No,No,No,Month-to-month,Yes,Electronic check,69.7,316.9,Yes
6572-ADKRS,Female,0,No,No,46,Yes,No,Fiber optic,No,No,Yes,No,No,Month-to-month,Yes,Credit card (automatic),74.8,3548.3,No
5380-WJKOV,Male,0,No,No,34,Yes,Yes,Fiber optic,No,Yes,Yes,No,Yes,Month-to-month,Yes,Electronic check,106.35,3549.25,Yes
8168-UQWWE,Female,0,No,No,11,Yes,Yes,Fiber optic,No,Yes,No,Yes,Month-to-month,Yes,Bank transfer (automatic),97.85,1105.4
8865-TNNMX,Male,0,Yes,Yes,10,Yes,No,DSL,No,Yes,No,No,No,One year,No,Mailed check,49.55,475.7,No
9489-DEDVP,Female,0,Yes,Yes,70,Yes,Yes,DSL,Yes,Yes,No,Yes,No,Two year,Yes,Credit card (automatic),69.2,4872.35,No
9867-JC2SP,Female,0,Yes,Yes,17,Yes,No,No,internet service,No,internet service,No,internet service,No,internet service
4671-VJLCL,Female,0,No,No,63,Yes,Yes,DSL,Yes,Yes,Yes,Yes,No,Two year,Yes,Credit card (automatic),79.85,4861.45,No
4080-IIARD,Female,0,Yes,No,13,Yes,Yes,DSL,Yes,Yes,Yes,No,Yes,Month-to-month,Yes,Electronic check,76.2,981.45,No
3714-NTNFO,Female,0,No,No,49,Yes,Yes,Fiber optic,No,No,No,No,Yes,Month-to-month,Yes,Electronic check,84.5,3906.7,No
5948-UJZLF,Male,0,No,No,2,Yes,No,DSL,No,Yes,No,No,No,Month-to-month,No,Mailed check,49.25,97,No
7760-OYPDY,Female,0,No,No,2,Yes,No,Fiber optic,No,No,No,Yes,Month-to-month,Yes,Electronic check,80.65,144.15,Yes
7639-LIAIYI,Male,0,No,No,52,Yes,Yes,DSL,Yes,No,Yes,Yes,Two year,Yes,Credit card (automatic),79.75,4217.8,No
2954-PIBKO,Female,0,Yes,Yes,69,Yes,Yes,DSL,Yes,No,Yes,Yes,No,Two year,Yes,Credit card (automatic),64.15,4254.1,No
8012-SOUDQ,Female,1,No,No,43,Yes,Yes,Fiber optic,No,Yes,No,No,Yes,No,Month-to-month,Yes,Electronic check,90.25,3838.75,No
9420-LOJXK,Female,0,No,No,15,Yes,No,Fiber optic,Yes,Yes,No,Yes,Yes,Month-to-month,Yes,Credit card (automatic),99.1,1426.4,Yes
6575-SUVOI,Female,1,Yes,No,25,Yes,Yes,DSL,Yes,No,No,Yes,Yes,No,Month-to-month,Yes,Credit card (automatic),69.5,1752.65,No
7495-OOKFY,Female,1,Yes,No,8,Yes,Yes,Fiber optic,No,Yes,No,No,No,Month-to-month,Yes,Credit card (automatic),80.65,633.3,Yes

```

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

A screenshot of the "Select File Format" dialog. It has fields for "File type" (set to "CSV"), "Field Delimiter" (set to ","), "Escape Character" (set to "\"), and "Quote Character" (set to "\""). Below these is a section labeled "Is first row header?" which contains a checked checkbox. Underneath is another section labeled "Contains endlines?" with an unchecked checkbox. At the bottom is a "PREVIEW" button.

10) Click “PREVIEW” prior to creating the table

Table Preview															
CUSTOMERID	GENDER	SENIORCITIZEN	PARTNER	DEPENDENTS	TENURE	PHONESERVICE	MULTIPLELINES	INTERNETSERVICE	ONLINESECURITY	ONLINEBACKUP	DEVICEPROTECTION	TECHSUPPORT	STREAMINGTV	STREAMINGMOVIES	CONTRA
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-month
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-month
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year
9237-	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-

← BACK    **NEXT →**    CANCEL

11) Click “NEXT”

Table Preview															
CUSTOMERID	GENDER	SENIORCITIZEN	PARTNER	DEPENDENTS	TENURE	PHONESERVICE	MULTIPLELINES	INTERNETSERVICE	ONLINESECURITY	ONLINEBACKUP	DEVICEPROTECTION	TECHSUPPORT	STREAMINGTV	STREAMINGMOVIES	CONTRA
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-month
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-month
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year
9237-	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-

← BACK    **NEXT →**    CANCEL

## 12) Click “CREATE”

Database / Upload Table

Table Name: telco\_churn

COLUMNS ADVANCED TABLE PROPERTIES

COLUMN NAME	DATA TYPE	SIZE	ADVANCED	ACTION
customerID	STRING		<input type="checkbox"/> Allow complex datatypes	<span>DELETE</span>
gender	STRING		<input type="checkbox"/> Allow complex datatypes	<span>DELETE</span>
SeniorCitizen	INT		<input type="checkbox"/> Allow complex datatypes	<span>DELETE</span>
Partner	STRING		<input type="checkbox"/> Allow complex datatypes	<span>DELETE</span>
Dependents	STRING		<input type="checkbox"/> Allow complex datatypes	<span>DELETE</span>
tenure	INT		<input type="checkbox"/> Allow complex datatypes	<span>DELETE</span>
PhoneService	STRING		<input type="checkbox"/> Allow complex datatypes	<span>DELETE</span>

BACK + CREATE CANCEL

## 13) Go-to “Compose” and within “Worksheet 1” run the following query on the new table

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

DATA ANALYTICS STUDIO

Compose

Saved | Worksheet1 | +

DATABASES | 6

default

TABLES | 1

Search Tables

telco\_churn (21)

EXECUTE EXPLAIN

Results

TELCO_CHURN.CUSTOMERID	TELCO_CHURN.GENDER	TELCO_CHURN.SENIORCITIZEN	TELCO_CHURN.PARTNER	TELCO_CHURN.DEPENDENTS	TELCO_CHURN.TENURE	TELCO_CHURN.PHONESERVICE	TELCO_CHURN.MULTIPLELINES	TELCO_CHURN.IN
7590-VWVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL