

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

*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 a search interface with a search bar containing "flights". Below the search bar, there are two sections: "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 same search interface as above, but with filters applied. The "Filters" section on the left has "TYPE" selected, with "Hive Table" checked.

Data Lakes

NA	Type	Name	Location
cdptrialuser31-dl	Hive Table	flights	/airlines_new_orc
	Hive Table	flights	/airlines_new_parquet

Filters

TYPE

- Hive Table
- HBase Table

+ Add New Value

*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 as a Hive Table located at "/airlines_new_orc" and another as a Hive Table located at "/airlines_new_parquet". The entry at "/airlines_new_orc" is highlighted with a red box.

Type	Name	Location
Hive Table	flights	/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: 0

Lineage: A diagram showing the data flow from multiple sources (labeled "/cdp-lake/data/al...") into the "flights" table, which then has an arrow pointing to a specific location: "/cdp-lake/data/airlines/airlines_new_orc.db/flights/year=1997".

Detailed Information for Lineage Node:

- Table Type: MANAGED_TABLE
- Database: airlines_new_orc
- DB Catalog: cm
- Parent: airlines_new_orc

Details for Target Location:

- Guild: be5a2094-f572-432e-9fa4-95498f7db650
- 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

*xx represents the trial user #

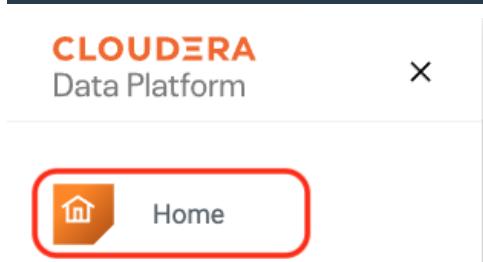
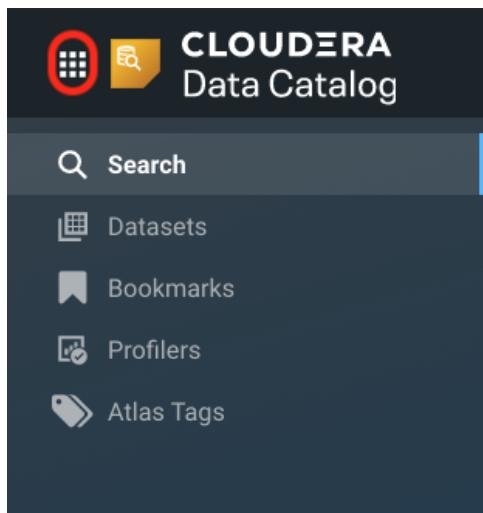
*X represents the password

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' creation interface. At the top right, there is a search icon and a red-circled '+' button. Below the header, the form is titled 'New Virtual Warehouse'. It contains fields for 'Name' (with placeholder 'Enter Virtual Warehouse Name'), 'Type' (radio buttons for 'HIVE' and 'IMPALA' with 'HIVE' selected), 'Database Catalog' (dropdown menu showing 'cdptrialuser24-dl-default'), 'Size' (dropdown menu showing '-- select an option --'), and a preview section for a virtual warehouse named 'default-vw' which is currently 'Stopped'. The preview shows node 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' creation interface. The 'Name' field is highlighted with a red border. The rest of the form is identical to the previous screenshot, including the 'Type' (HIVE selected), 'Database Catalog' (cdptrialuser24-dl-default), and 'Size' dropdown.

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

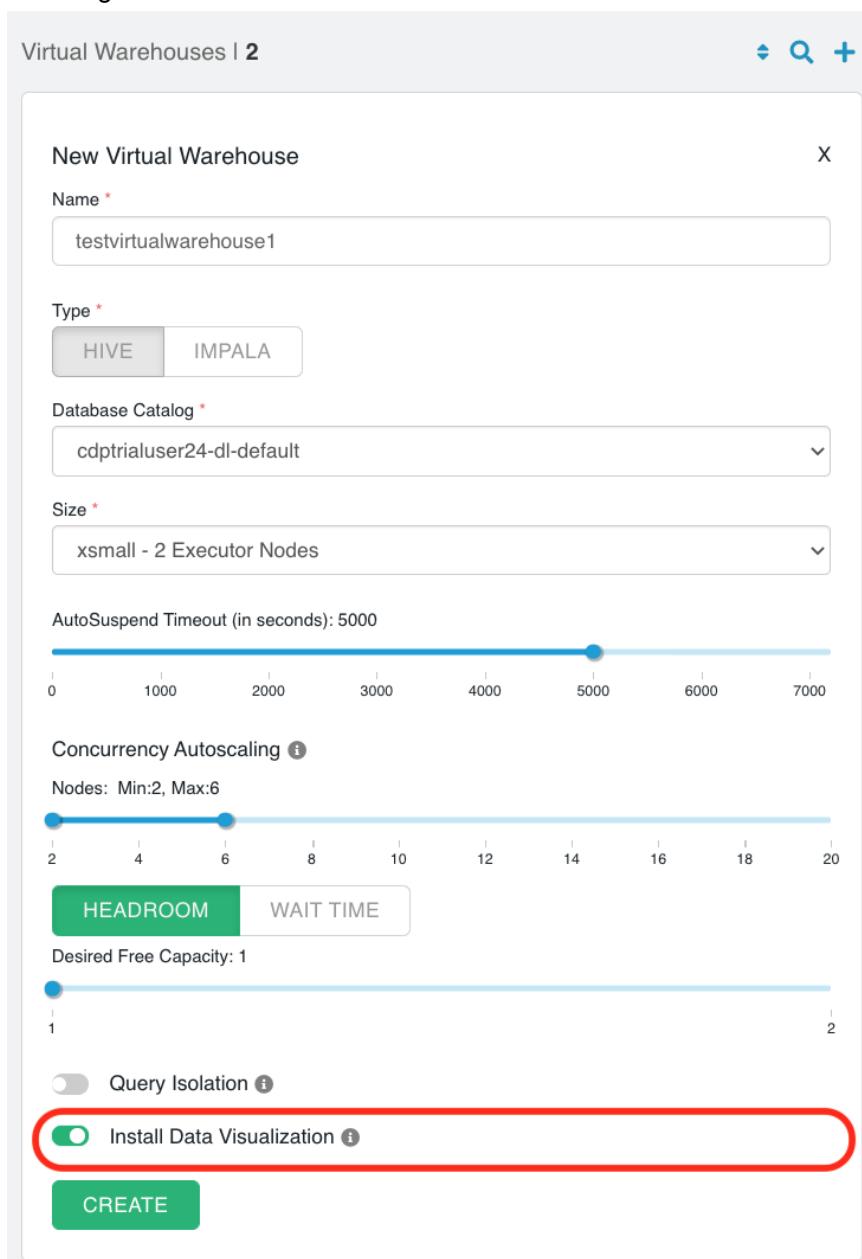
HEADROOM WAIT TIME

Desired Free Capacity: 1

Query Isolation ⓘ

Install Data Visualization ⓘ

CREATE



8) Click “Create” to create your Virtual Warehouse

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



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

testvirtualwarehouse1
compute-1611179792-vz49
cdptrialuser24-dl-default

Starting

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

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

testvirtualwarehouse1
compute-1611179792-vz49
cdptrialuser24-dl-default

Running

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

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

Virtual Warehouses | 3

testvirtualwarehouse1
compute-1611179792-vz49
cdptrialuser24-dl-default

Running

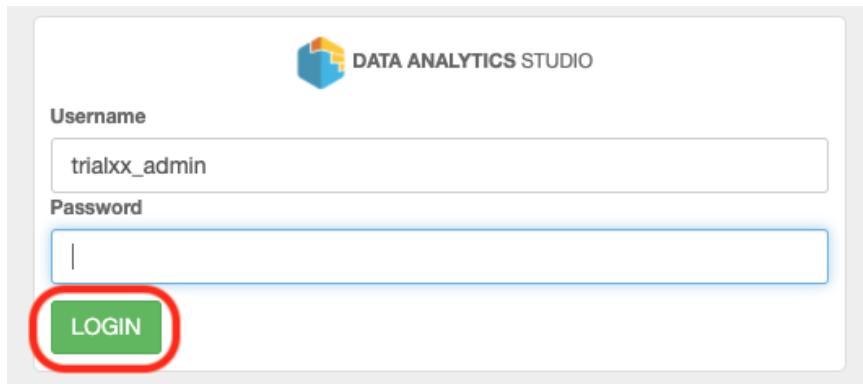
NODE COUNT 2 TOTAL CORES 38 TOTAL MEMORY 292 GB TYPE HIVE DATA VISUALIZATION

mschoeni-iso-1
compute-1611173596-db7v
cdptrialuser24-dl-default

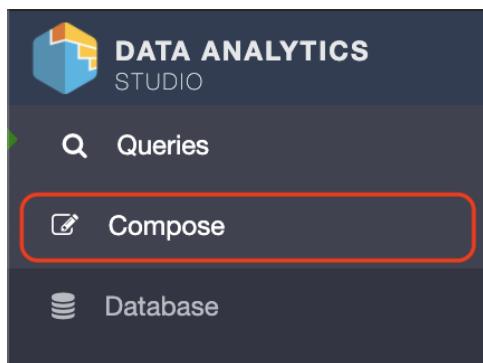
Stopped

Suspend
Clone
Edit
Delete
Upgrade
Copy JDBC URL
Download JDBC Jar
Open DAS
Open Data Visualization

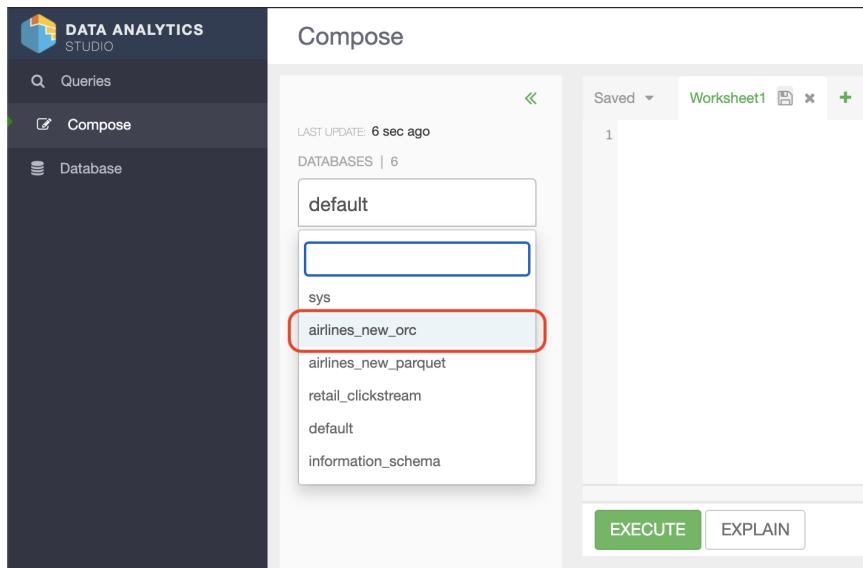
- 10) Enter the login information from step #1 above using the user, then click “LOGIN”
*Changing “trialxx_admin” to the trail user you’re using and password defined by “X” in #1 above



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

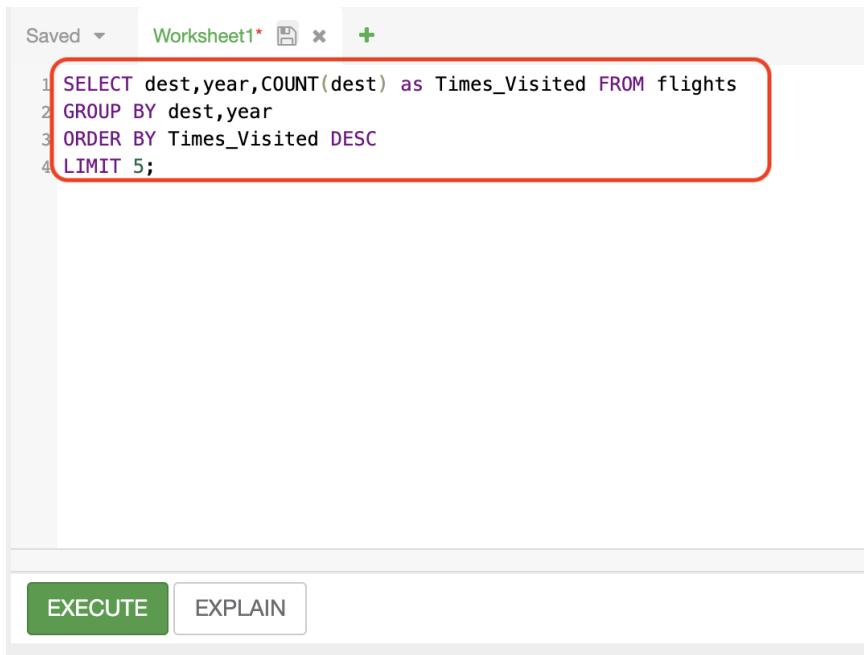


- 12) Choose the database “airlines_new_orc” that we found in Part 1 “Data Catalog”



- 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

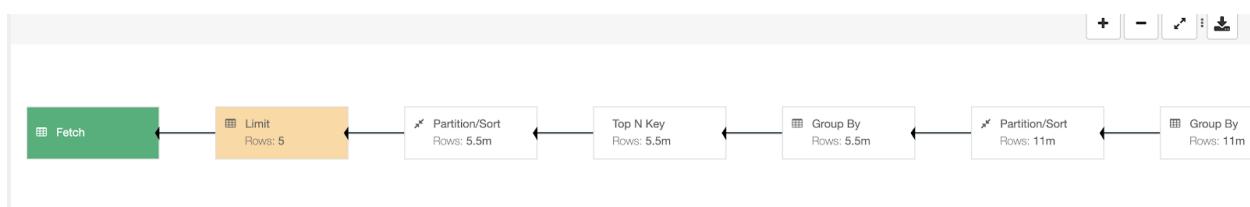
14) Click “EXPLAIN” to see the visual explain plan prior to running the query

*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 circle), a file icon, and a close button. 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”, 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?”

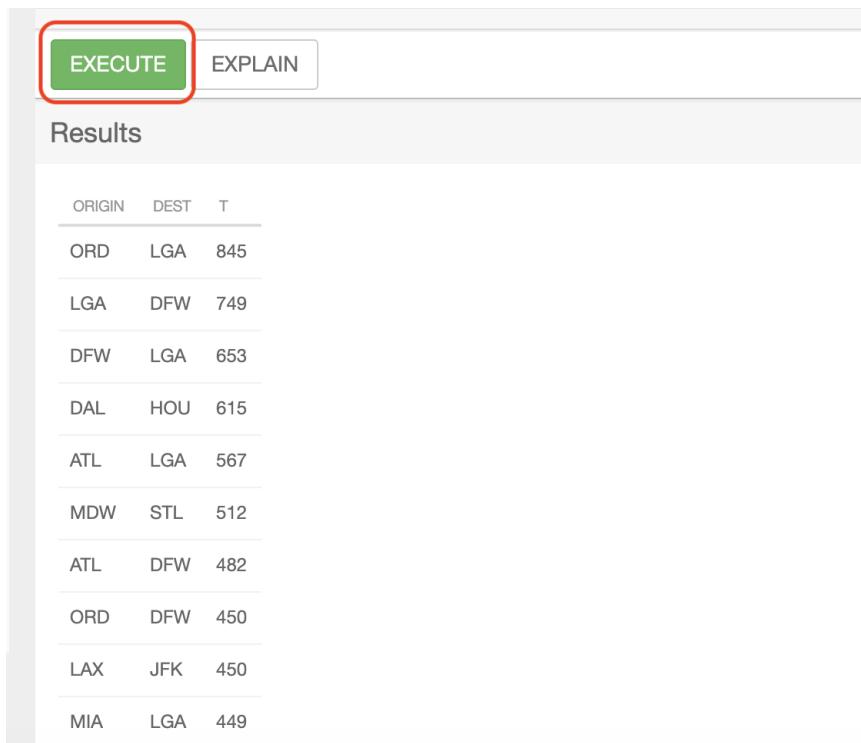
```
SELECT origin,dest,COUNT(Diverted) as t FROM flights
WHERE Diverted = 1
GROUP BY origin,dest
ORDER BY t DESC
LIMIT 10;
```

The screenshot shows a database worksheet interface with two tabs: "Worksheet1*" and "Worksheet2*". The "Worksheet2*" tab is active. The query window contains the following SQL code:

```
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;|
```

The code is highlighted with a red rectangle. Below the code are two buttons: "EXECUTE" and "EXPLAIN".

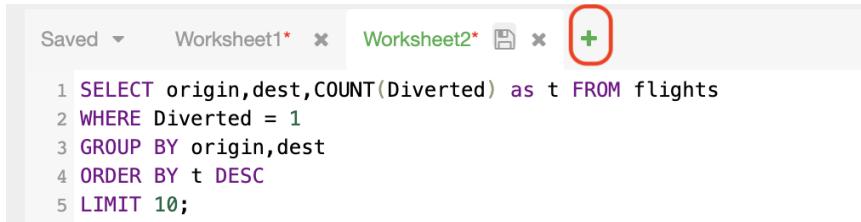
- 19) Click “EXECUTE” to execute the query, answering the question “What are the top 10 routes (origin and dest) that have seen maximum diversions?”



The screenshot shows a user interface for executing SQL queries. At the top, there are two buttons: "EXECUTE" (highlighted with a red box) and "EXPLAIN". Below them is a section titled "Results" containing a table with the following data:

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



The screenshot shows a workspace with multiple worksheets open. The tabs are labeled "Saved", "Worksheet1*", "Worksheet2*", and a new tab represented by a plus sign icon. Below the tabs, a query editor contains the following SQL code:

```
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”, 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?”

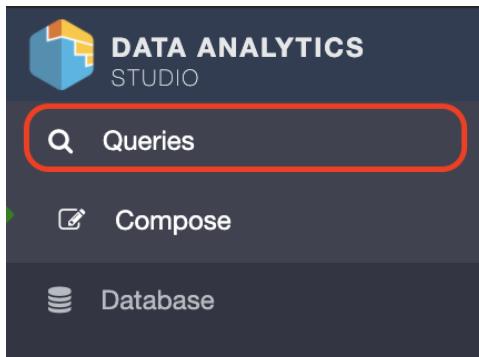
The screenshot shows a user interface for executing SQL queries. At the top, there are two buttons: "EXECUTE" (highlighted with a red box) and "EXPLAIN". Below them is a button labeled "Execute Query". The main area is titled "Results" and contains a table with the following data:

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)

The screenshot shows a table titled "QUERIES (159)" with the following columns: QUERY, STATUS, QUEUE, USER, TABLES READ, TABLES WRITTEN, START TIME, DURATION, DAG ID, and ACTIONS. The first five rows of data are:

QUERY	STATUS	QUEUE	USER	TABLES READ	TABLES WRITTEN	START TIME	DURATION	DAG ID	ACTIONS
SELECT month,COUNT(Cancelled) as ...	SUCCESS	None	trial24_admin	flights (airlines_new_o...)	Not Available!	8 seconds ago	00:00:00	Not Available!	
SELECT month,COUNT(Cancelled) as ...	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)

The screenshot shows a table titled "QUERIES (159)" with the same columns and data structure as the previous table. The second row of data is highlighted with a red box around its "Compare" icon.

QUERY	STATUS	QUEUE	USER	TABLES READ	TABLES WRITTEN	START TIME	DURATION	DAG ID	ACTIONS
SELECT month,COUNT(Cancelled) as ...	SUCCESS	None	trial24_admin	flights (airlines_new_o...)	Not Available!	4 minutes ago	00:00:00	Not Available!	
SELECT month,COUNT(Cancelled) as ...	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

The screenshot shows the CDW interface with two queries in the editor. The first query is "SELECT month,COUNT(Cancelled) as num_of_cancellations I" and the second is "SELECT month,COUNT(Cancelled) as num_of_cancellations I". A red box highlights the "COMPARE" button at the top right of the editor area. Below the editor, there is a "Compare two queries" link.

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

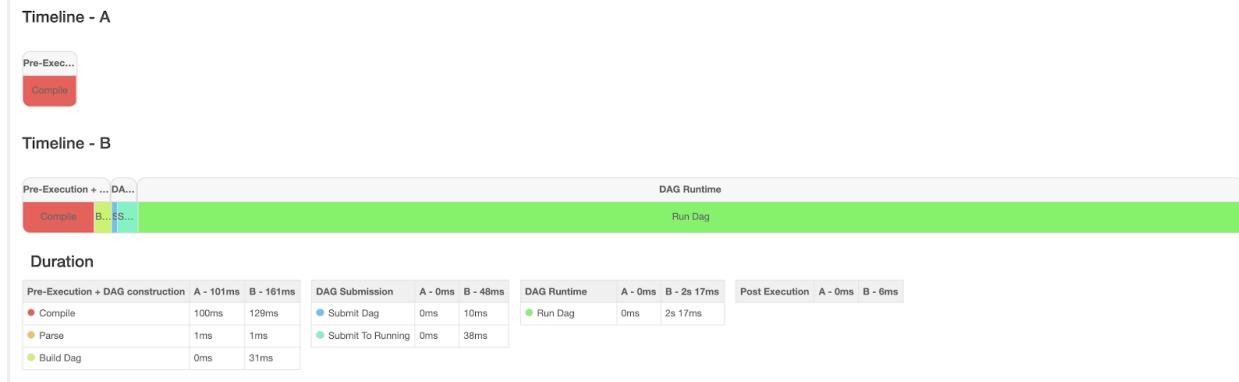
END TIME
21 Jan 2021 09:35:15

DURATION
2s 311ms

28) Click on “timeline” at the top

The screenshot shows the CDW interface with a navigation bar at the top. The tabs are "RECOMMENDATIONS", "QUERY DETAILS", "VISUAL EXPLAIN", "CONFIGS", "TIMELINE" (which is highlighted with a red box), and "DAG INFO".

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

*xx represents the trial user #

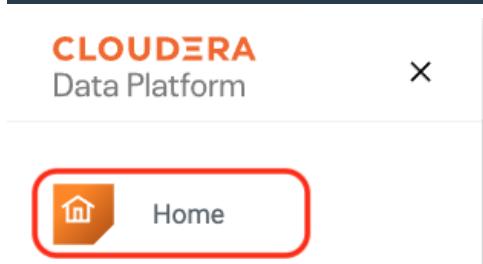
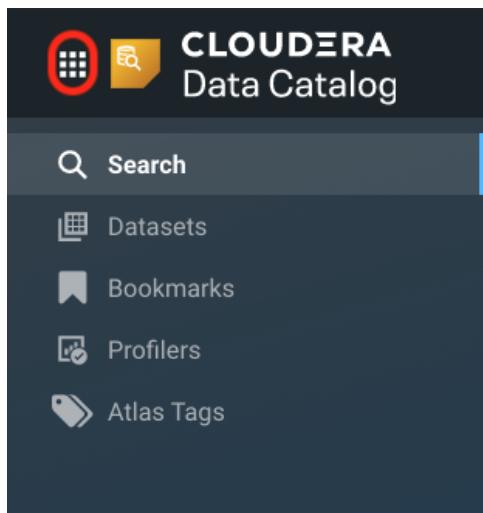
*X represents the password

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 shows a 'NEW DATASET' button, a 'Connection Explorer' link, and a table header for 'Title/Table' and 'Created'.

7) Click "NEW DATASET" to add our "flights" data

The screenshot shows the same interface again, with the 'NEW DATASET' button in the main toolbar highlighted by a red circle. The sidebar still shows 'Default Hive VW' selected. The main area includes a 'Connection Explorer' link and a table header.

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

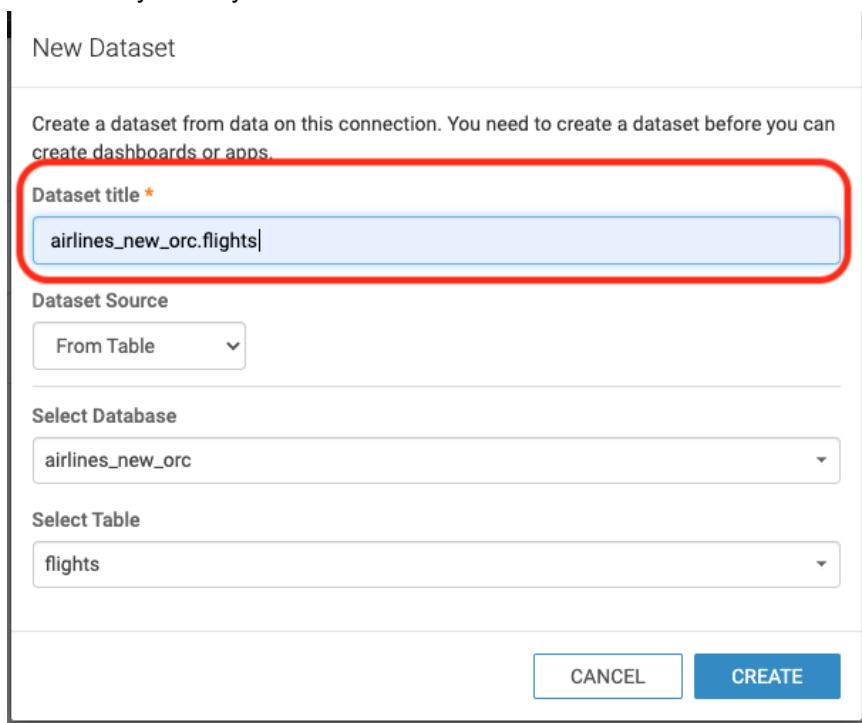
Select Database

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

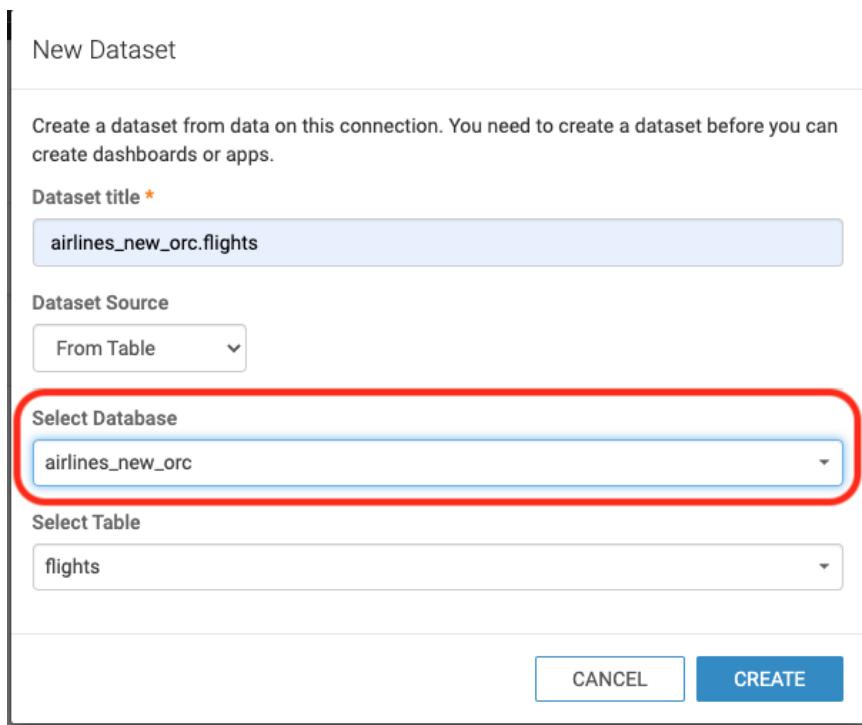
Select Database

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

 flights

CANCEL CREATE

11) Click “CREATE”

New Dataset

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

Dataset title *

Dataset Source

From Table

Select Database

Select Table

 flights

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'. The first item was created on Jan 21, 2021, and last updated a few seconds ago by triad24_admin with 0 visual.

13) Choose “Treemap” under “VISUALS”

The screenshot shows the 'Dashboard Designer' interface. On the left, the 'VISUALS' tab is selected, displaying various visualization icons. One icon, 'Treemap', is highlighted with a red circle. Below the icons are sections for 'Dimensions', 'Measures', 'Toolips', and 'Filters', each with a 'drag or click fields to add here' placeholder. A 'Limit' input field is set to 100. At the bottom is a blue 'REFRESH VISUAL' button. To the right, the 'DATA' tab shows a dataset named 'airlines_new.orc.flights' with a sample mode of 'OFF'. It lists dimensions like 'flights', 'uniquecarrier', 'tailnum', etc., and measures like '# Record Count', '# month', etc. On the far right, a sidebar shows navigation links for '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 Edit Delete
1234 LABEL	Sample Mode: OFF
	Search: <input type="text"/> X
Dimensions	Dimensions (6)
	▼ flights
	A uniquecarrier
	A tailnum
	A origin
	A dest
	A cancellationcode
	A diverted
Measures	Measures (24)
	▼ 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
REFRESH VISUAL	

The Dimensions section on the left contains two items highlighted with a red box: "dest" and "origin".

The right sidebar includes sections for Favorites (Airdrop, Recents, Application, Desktop), Filters, Settings, Style, and Tags (Red, Orange, Yellow, Green, Blue).

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. In the center, a 'FIELD PROPERTIES' pane is open for a visual containing the measure '# Record Count'. The 'Order and Top K' section is highlighted with a red box. Under 'Order and Top K', 'Descending' is checked, and 'Top K' and 'Bottom K' both have 'eg. 100' entered. A note below states 'Top K/Bottom K applies to granular dimensions'. On the right, the sidebar shows 'Build' is selected, along with 'Visuals', 'Filters', 'Settings', and 'Style'.

VISUALS

- Treemap
- 1234
- WORD
- ACTION
- SQL

* Dimensions

- dest
- origin

* Measure

- # Record Count

Tooltips

X Trellis

Y Trellis

Filters

REFRESH VISUAL

FIELD PROPERTIES

- Date/Time Functions
- Text Functions
- Analytic Functions
- Change Type
- Order and Top K

✓ Descending

Ascending

Top K: eg. 100

Bottom K: eg. 100

Top K/Bottom K applies to granular dimensions

[] Enter/Edit Expression

Display Format

Alias

Description

Duplicate

Save Expression

Remove

DASH.

Visuals

Filters

Settings

Style

Build

Visuals

Filters

Settings

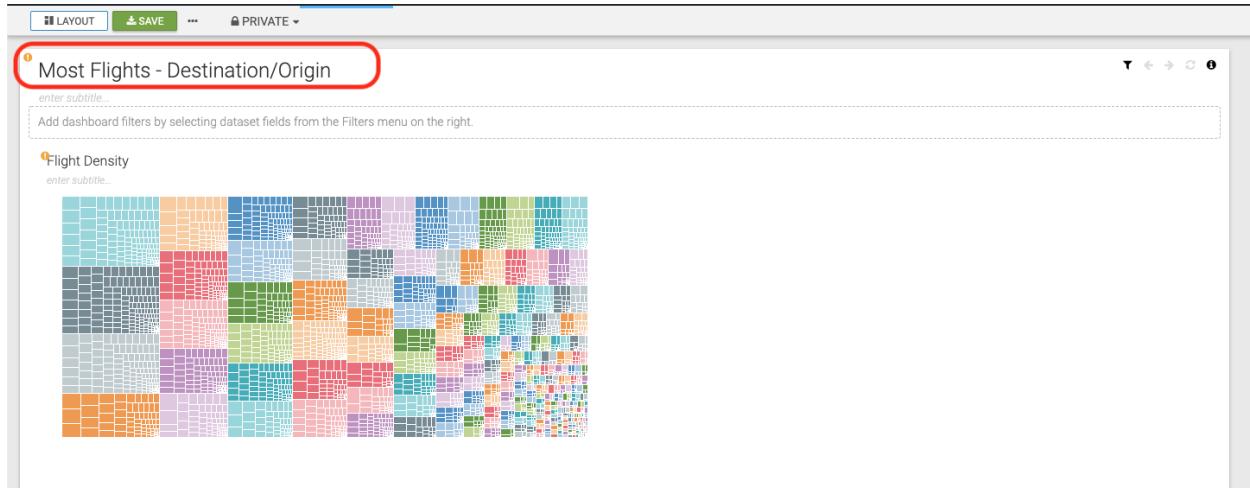
Style

17) Click "REFRESH VISUAL"

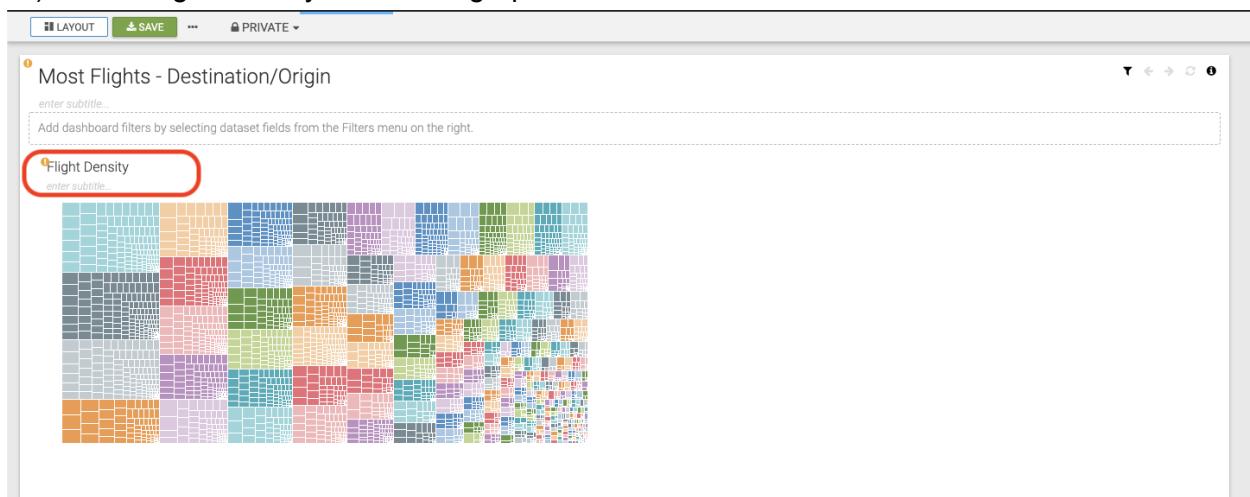
*Notice - you can have other Visuals chosen to be displayed with the Dimensions and Measure(s), then click REFRESH VISUALS

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 like flights, uniquecarrier, tailnum, origin, dest, cancellationcode, diverted, and measures like Record Count, month, dayofmonth, dayofweek, deptime, crsdeptime, arftime, crsarftime, flightnum, actualelapsedtime, crselapsedtime, airtime, and arrdelay. On the right, the **BUILD** pane contains settings and style options. A vertical sidebar on the far right provides navigation links for Visuals, Filters, Settings, Style, and Build.

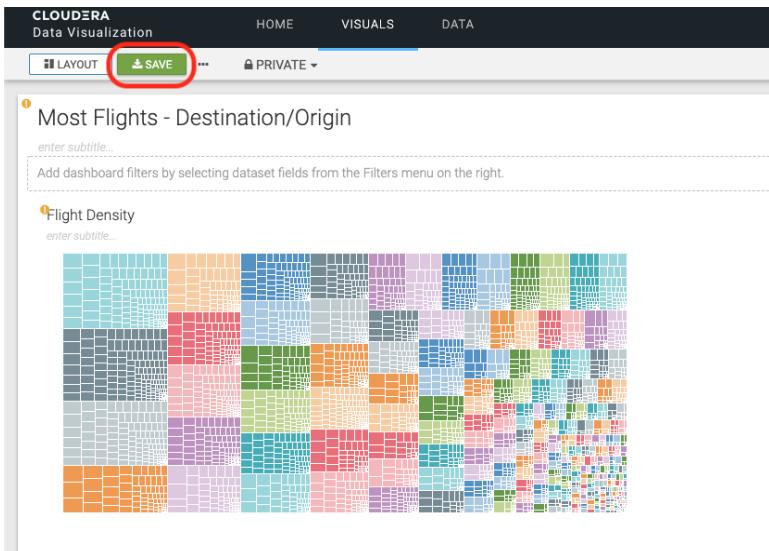
18) Enter a title “Most Flights - Destination/Origin”



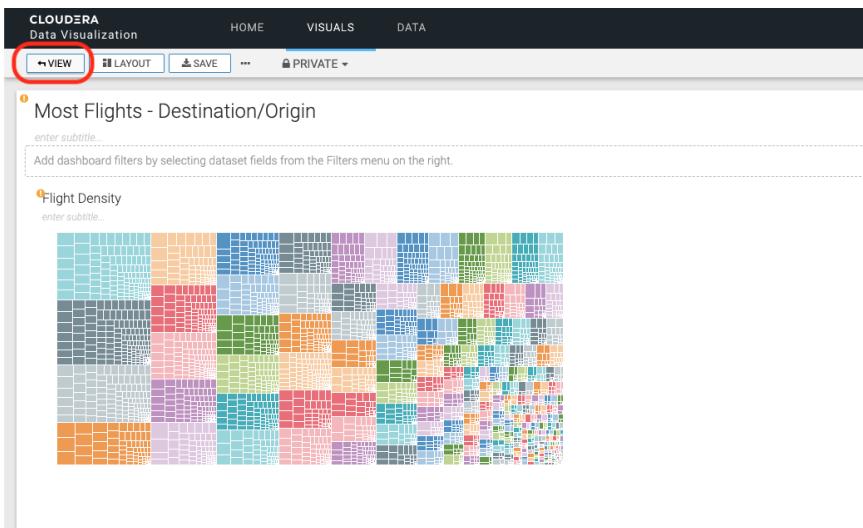
19) Enter “Flight Density” under the graph’s title



20) Click "SAVE"

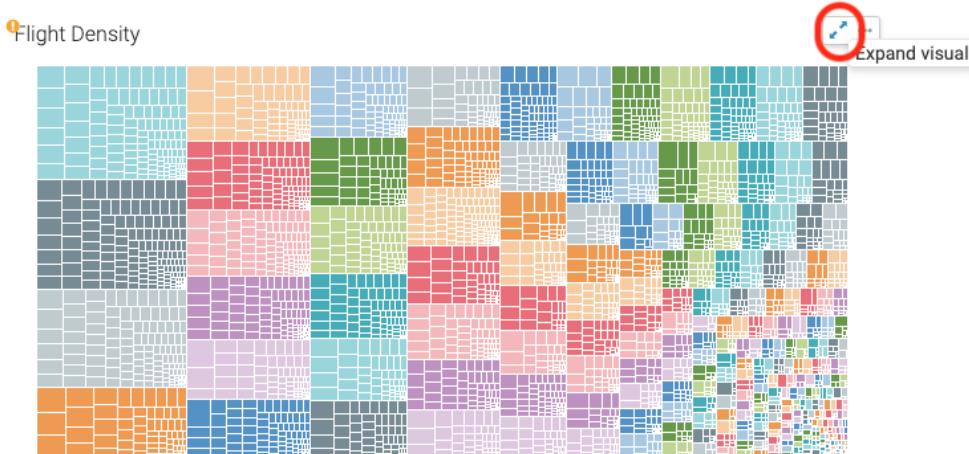


21) Click "VIEW"



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

Most Flights - Destination/Origin



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.

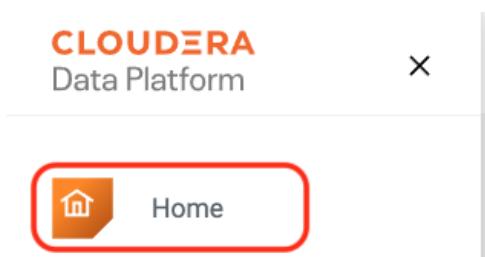
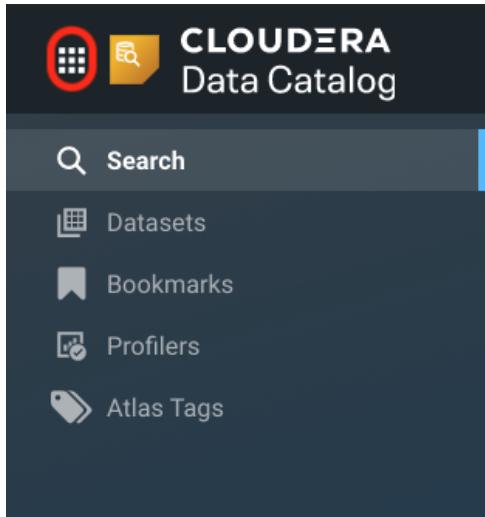
Your link should look something like (remember click the link in your email not the link below)
http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X
*xx represents the trial user #
*X represents the password

- 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.

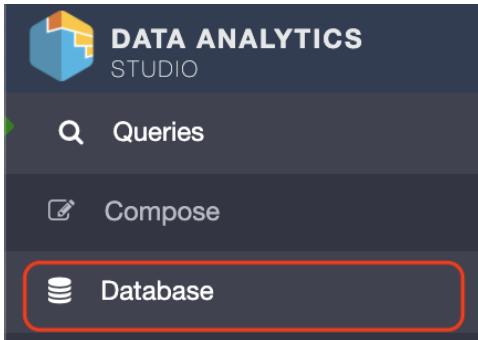
4) Enter the login information from step #1 above using the user, then click “LOGIN”

*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

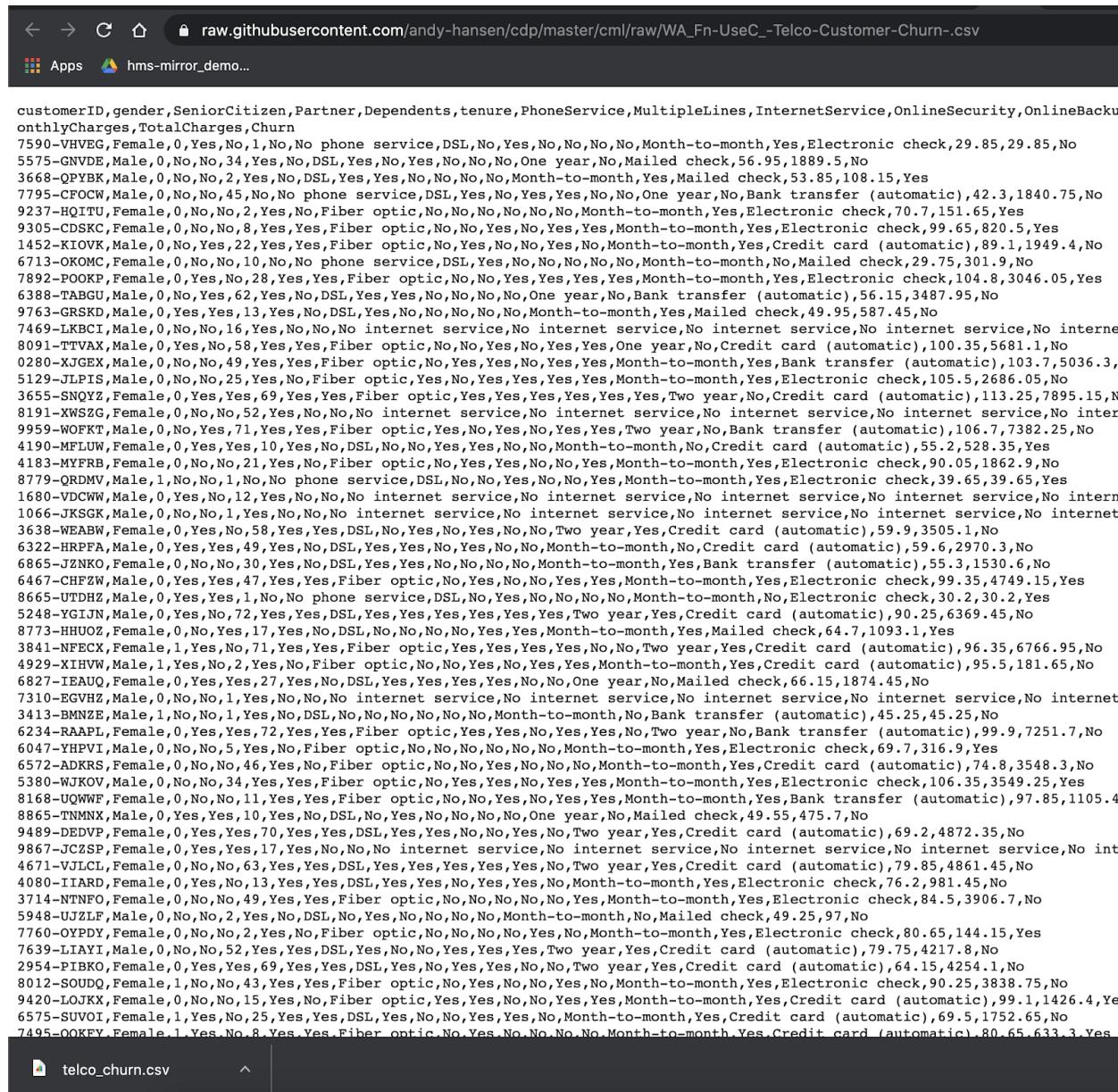


6) Click on “Upload Table”, using the “default” database

A screenshot of the Database Explorer interface. At the top, it says 'Database Explorer' and 'LAST UPDATE: 7 sec ago DATABASES | 6'. Below that, it shows a section for the 'default' database, which has 'TABLES | 0'. To the right of this section are three icons: a green circular refresh icon, a green plus sign icon, and a green icon with an upward arrow (the 'Upload Table' button). The green icon with the upward arrow is circled in red.

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

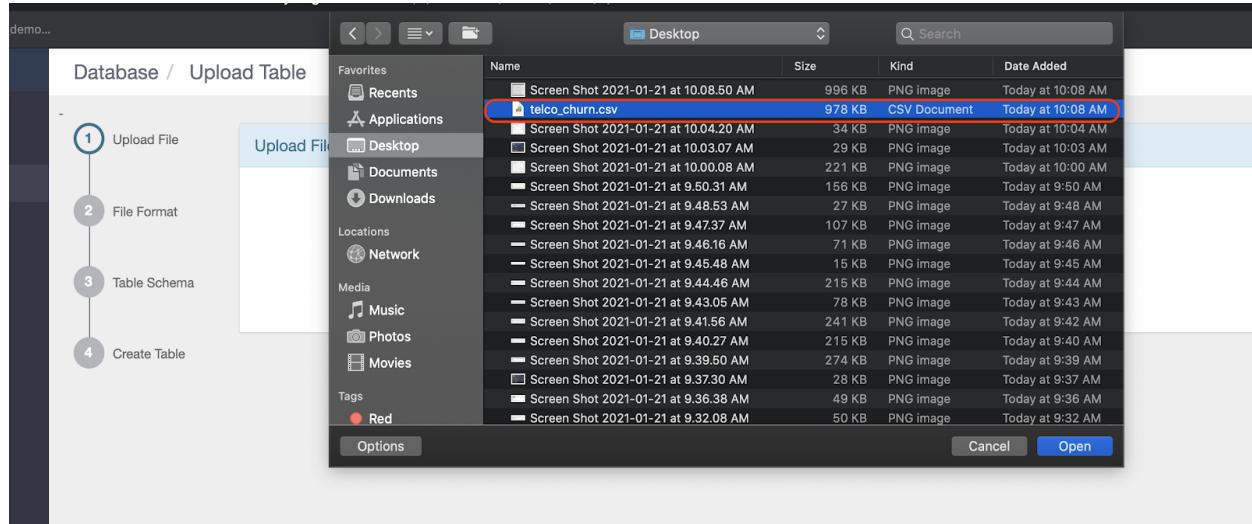
https://raw.githubusercontent.com/andy-hansen/cdp/master/cml/raw/WA_Fn-UseC_-Telco-Customer-Churn-.csv



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. The page displays a large amount of CSV data. At the bottom left, there is a download link labeled "telco_churn.csv".

```
customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,MonthlyCharges,TelcoChurn
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,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,No,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,Yes,Two year,No,Credit card (automatic),113.25,7895.15,No
8191-XWSZG,Female,0,No,No,52,Yes,No,No,internet service,No,internet service,No,internet service,No,internet service,No,internet service
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,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,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,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,Yes,27,Yes,No,DSL,Yes,Yes,Yes,No,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,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,Yes,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,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 comma), "Escape Character" (set to backslash), and "Quote Character" (set to double quote). 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	CONTR
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 ye
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 ye
9237-	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-
< BACK	NEXT >														x CANCEL

11) Click “NEXT”

Table Preview

CUSTOMERID	GENDER	SENIORCITIZEN	PARTNER	DEPENDENTS	TENURE	PHONESERVICE	MULTIPLELINES	INTERNETSERVICE	ONLINESECURITY	ONLINEBACKUP	DEVICEPROTECTION	TECHSUPPORT	STREAMINGTV	STREAMINGMOVIES	CONTR
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 ye
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 ye
9237-	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-
< BACK	NEXT >														x CANCEL

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

The screenshot shows a step-by-step process for creating a table. Step 3, 'Table Schema', is highlighted. The table name is 'telco_churn'. The schema defines seven columns:

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

At the bottom, there are buttons for 'BACK', '+ CREATE' (which is highlighted with a red box), and 'CANCEL'.

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

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

The screenshot shows the 'Compose' interface with a saved worksheet named 'Worksheet1'. The query 'select * from telco_churn limit 10;' is executed. The results are displayed in a table:

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