

Agenda :

Hands-on (step-by-step below) (30 minutes)

- i) Part 1 - Create a Virtual Warehouse and Run Queries (15 minutes)
- ii) Part 2 - Data Visualization (10 minutes)
- iii) Part 3 - Import a File into a Table (5 minutes)

Step-by-step instructions:

Part 1 - Create a Virtual Warehouse and Run Queries [25 minutes]

Commented [1]: Check time...

Overview: What is Cloudera Data Warehouse?

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

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

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

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

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

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

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

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

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

*xx represents the trial user #

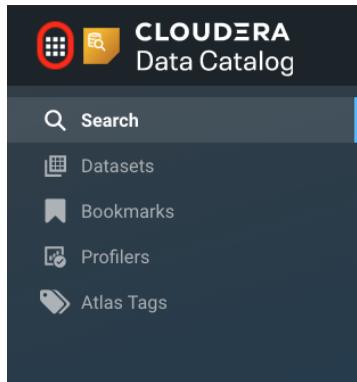
*X represents the password

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



How do you get to the CDP Home Screen?

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



CLOUDERA
Data Platform

X



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

Virtual Warehouses | 1

♦

New Virtual Warehouse

Name *

Type * HIVE IMPALA

Database Catalog *

Size *

default-vw
compute-1611103491-4hbp
Stopped cdptrialuser24-dl-default

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

4) Enter a name for your New Virtual Warehouse

Virtual Warehouses | 1

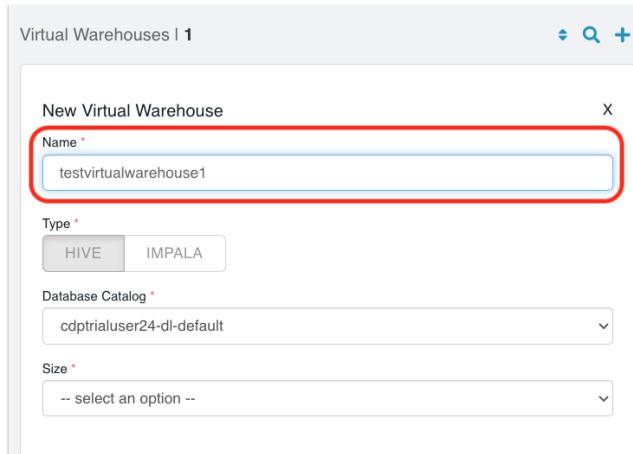
New Virtual Warehouse

Name * testvirtualwarehouse1

Type * HIVE IMPALA

Database Catalog * cdptrialuser24-dl-default

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

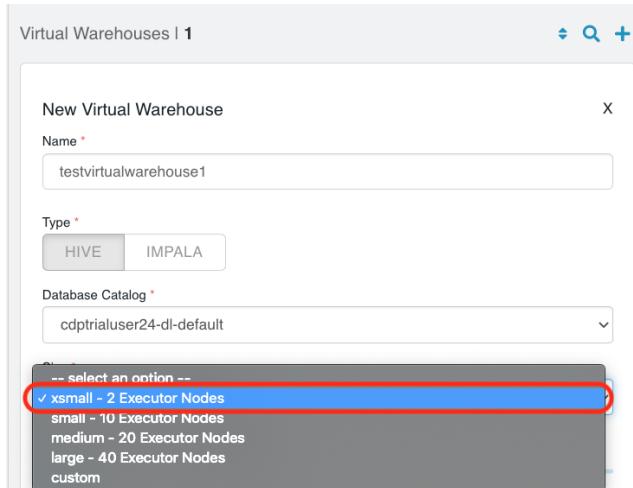
Name * testvirtualwarehouse1

Type * HIVE IMPALA

Database Catalog * cdptrialuser24-dl-default

-- 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

A slider for "AutoSuspend Timeout (in seconds)" ranging from 0 to 7000. The current value is set to 5000. The input field above the slider shows "5000". The entire input field and slider area are highlighted with a red rounded rectangle.

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 *

Size *

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

Starting cdptrialuser24-dl-default

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

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

testvirtualwarehouse1

compute-1611179792-vz49

Running cdptrialuser24-dl-default

NODE COUNT	TOTAL CORES	TOTAL MEMORY	TYPE
2	38	292 GB	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	Running cdptrialuser24-dl-default
2	38	292 GB

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

Suspend

Clone

Edit

Delete

Upgrade

Copy JDBC URL

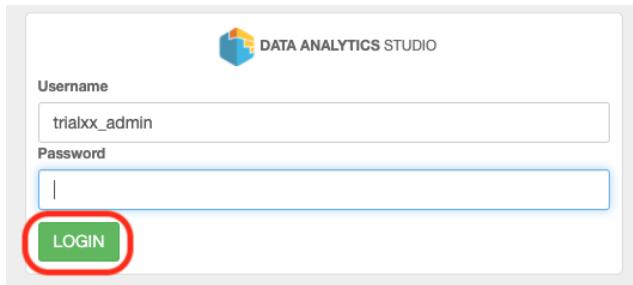
Download JDBC Jar

Open DAS

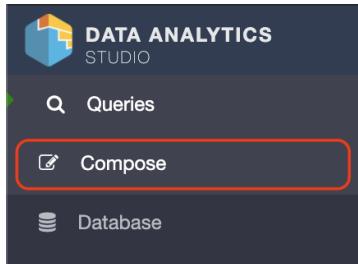
Open Data Visualization

10) Enter the login information from step #1 above using the user, then click "LOGIN"

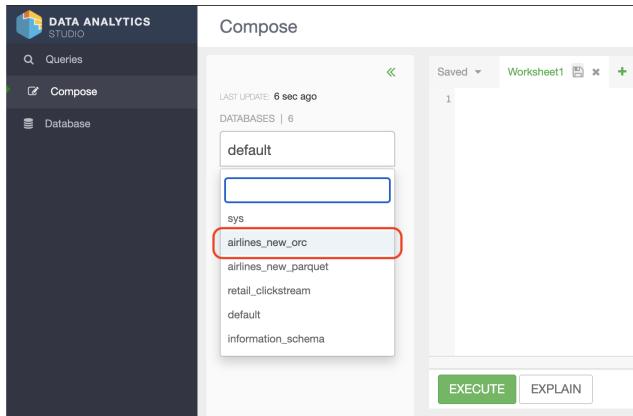
*Changing "trialxx_admin" to the trail user you're using and password defined by "X" in #1 above



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



12) Choose the database "airlines_new_orc".



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

```
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;
```

Commented [2]: Had to run twice ??

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

- 14) 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". The query entered is:

```
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;
```

Below the query, there are two buttons: "EXECUTE" (highlighted with a red box) and "EXPLAIN".

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

The screenshot shows a results table titled "Results" with the following data:

DEST	YEAR	TIMES_VISITED
ATL	2003	4296000
ATL	2004	418669
ATL	2008	414521
ATL	2007	413805
ATL	2006	404629

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

The screenshot shows a database worksheet titled "Worksheet1". The query entered is:

```
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;
```

Below the query, there are two buttons: "EXECUTE" and "EXPLAIN". Above the query, there is a "+" button (highlighted with a red circle).

- 17) 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?"

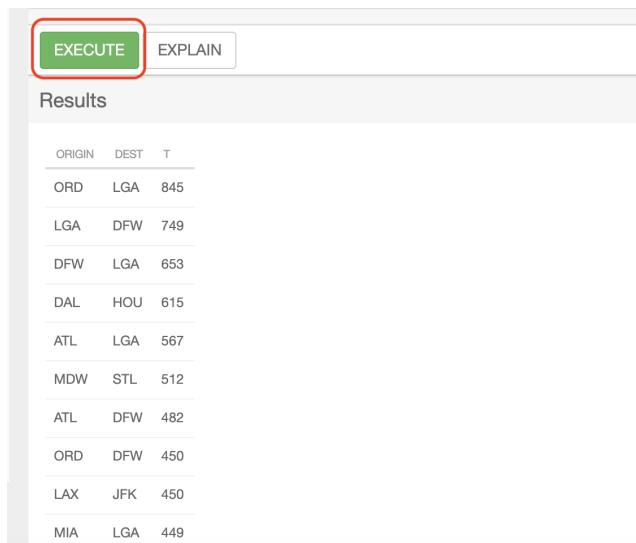
```
SELECT origin,dest,COUNT(Diverted) as t FROM flights  
WHERE Diverted = 1  
GROUP BY origin,dest  
ORDER BY t DESC  
LIMIT 10;
```

The screenshot shows a database interface with two worksheets open: 'Worksheet1' and 'Worksheet2'. The code in 'Worksheet2' is highlighted with a red box. At the bottom of the interface are 'EXECUTE' and 'EXPLAIN' buttons.

```
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;|
```

EXECUTE EXPLAIN

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

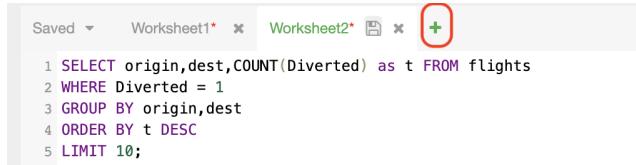


EXECUTE EXPLAIN

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

- 19) Going back to "Worksheet 2", click the "+" to add another Worksheet for the final query



```

Saved ▾ Worksheet1* × Worksheet2* + 
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;

```

- 20) 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;

```

21) 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

The screenshot shows the Data Analytics Studio interface. On the left, there is a navigation bar with several items: "Queries" (highlighted with a red box), "Compose", and "Database". The main area is dark-themed.

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

The screenshot shows a table of query runs. The first row, which is the most recent, has a "Compare" icon (two overlapping circles) highlighted with a red box in the "ACTIONS" column.

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!	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!	Not Available!	4 minutes ago	00:00:02	dag_161123648	
SELECT origin,dest,COUNT(diverted) as...	SUCCESS	None	trial24_admin	flights (airlines_new_o... Not Available!	Not Available!	8 minutes ago	00:00:03	dag_161123648	
SELECT dest,year,COUNT(dest) as Time...	SUCCESS	None	trial24_admin	flights (airlines_new_o... Not Available!	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!	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 of query runs. The second row from the top has a "Compare" icon (two overlapping circles) highlighted with a red box in the "ACTIONS" column.

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!	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!	Not Available!	8 minutes ago	00:00:02	dag_161123648	
SELECT origin,dest,COUNT(diverted) as...	SUCCESS	None	trial24_admin	flights (airlines_new_o... Not Available!	Not Available!	13 minutes ago	00:00:03	dag_161123648	
SELECT dest,year,COUNT(dest) as Time...	SUCCESS	None	trial24_admin	flights (airlines_new_o... Not Available!	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!	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 FROM ...". The second query is: "SELECT month,COUNT(Cancelled) as num_of_cancellations FROM ...". Below the queries is a red box highlighting the "COMPARE" button, which has a tooltip "Compare two queries".

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

The screenshot shows two "Query Details" sections. Section A (left) shows a duration of 118ms. Section B (right) shows a duration of 2s 311ms. Both sections include details like query ID, user, status, start time, end time, and a "DURATION" field.

Query Details - A	Query Details - B
QUERY ID hive_20210121153926_e3a56b9d-71f2-45dc-b23e-2c2e1146d61e	QUERY ID hive_20210121153513_37aefec1-0284-4897-bfb9-bf9bb5797252
USER trial24_admin	USER trial24_admin
STATUS ✓ SUCCESS	STATUS ✓ SUCCESS
START TIME 21 Jan 2021 09:39:26	START TIME 21 Jan 2021 09:35:13
END TIME 21 Jan 2021 09:39:26	END TIME 21 Jan 2021 09:35:15
DURATION 118ms	DURATION 2s 311ms

28) Click on “timeline” at the top

The screenshot shows the top navigation bar with tabs: RECOMMENDATIONS, QUERY DETAILS, VISUAL EXPLAIN, CONFIGS, and TIMELINE (which is highlighted with a red box). Below the timeline tab, there is a "Duration" section with a table comparing execution times for various stages between Query A and Query B.

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 2 - 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)

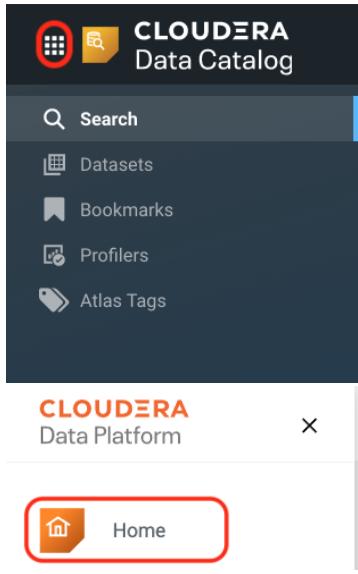
http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X
*xx represents the trial user #
*X represents the password

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

Virtual Warehouses

	NODE COUNT	TOTAL CORES	TOTAL MEMORY	TYPE
testvirtualwarehouse1	2	38	292 GB	HIVE
mschoeni-iso-1	NODE COUNT	TOTAL CORES	TOTAL MEMORY	
Stopped	compute-1611173596-dbvt	cdptrialuser24-dl-default		

Context menu for testvirtualwarehouse1:

- Suspend
- Clone
- Edit
- Delete
- Upgrade
- Copy JDBC URL
- Download JDBC Jar
- Open DAS
- Open Data Visualization** (highlighted with a red box)

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

CLOUDERA
Data Visualization

LOGIN

Username

Password

Invalid login

[Forgot your password?](#)

Remember me on this computer

LOGIN

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. Below the tabs is a search bar with placeholder text "find titles, viz types, datasets, authors...". On the left, there's a sidebar with "All Connections" and two items: "% Default Hive VW" and "% samples". The main content area shows a "Datasets" section with a count of 11, a "Connection Explorer" section, and a table header for "Title/Table" and "Created".

6) Click "Default Hive VW" to add our dataset

This screenshot is similar to the previous one, but the "% Default Hive VW" item in the sidebar is highlighted with a red circle. The main content area remains the same, showing the Datasets and Connection Explorer sections.

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

This screenshot shows the same interface as the previous ones, but the "NEW DATASET" button in the top right of the main content area is highlighted with a red circle. The sidebar still shows the "% Default Hive VW" connection selected.

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 *

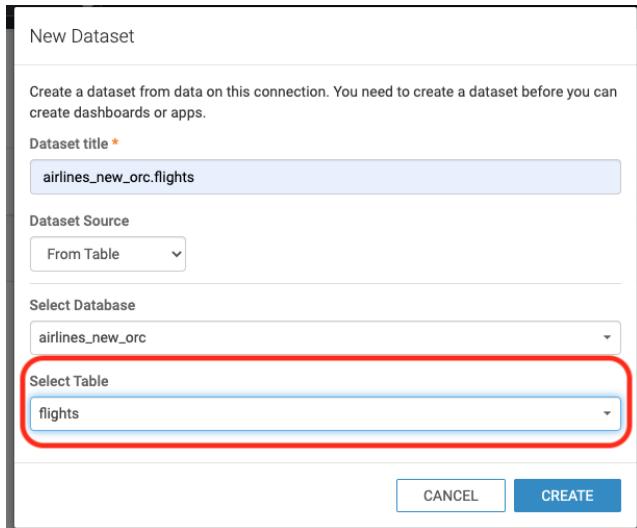
Dataset Source

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 *

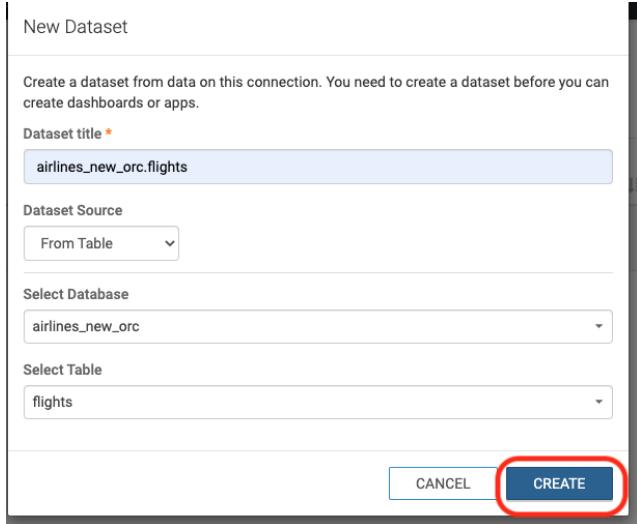
Dataset Source

From Table

Select Database

Select Table

CANCEL CREATE



12) Click "+" to create a New Dashboard

13) Choose "Treemap" under "VISUALS"

Dashboard Designer

VISUALS

- Table
- Treemap (highlighted with a red circle)
- 1234
- SQL

Dimensions
drag or click fields to add here

Measures
drag fields to add here

Toolips
drag fields to add here

Filters
drag fields to add here

Limit: 100

REFRESH VISUAL

DATA

airlines_new_orc.flights

Sample Mode: OFF

Search

Dimensions (6)

- flights
- A uniquecarrier
- A tailnum
- A origin
- A dest
- A cancellationcode
- A diverted

Measures (24)

- flights
- # Record Count
- # month
- # dayofmonth
- # dayofweek
- # deptime
- # crsdeptime
- # arrtime
- # crsarrrtime
- # flightnum
- # actualelapsedtime

DASH.

Visuals

Filters

Settings

Style

VISUAL

Build

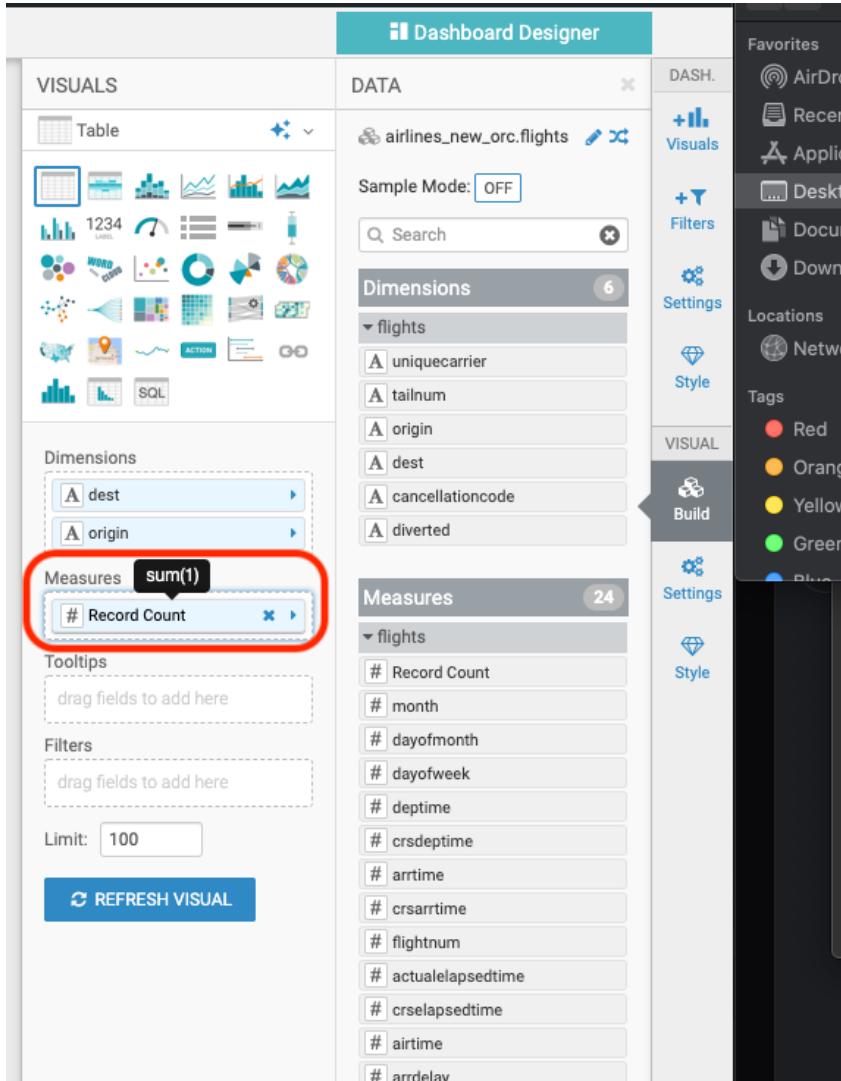
Settings

Style

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

The screenshot shows the Tableau Dashboard Designer interface. On the left, the **VISUALS** pane displays various visualization options like Table, Map, and Line charts. In the center, the **DATA** pane shows a connection to **airlines_new_orc.flights**. The **Dimensions** section lists **dest** and **origin**, which are highlighted with a red box. The **Measures** section lists **Record Count**, **month**, **dayofmonth**, **dayofweek**, **deptime**, **crsdeptime**, **artime**, **crsartime**, and **flightnum**. On the right, a sidebar shows **Favorites** (Airdrop, Recents, Application, Desktop), **Locations** (Network), and **Tags** (Red, Orange, Yellow, Green, Blue).

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



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

Dashboard Designer

VISUALS

- Treemap
- 1234
- LABEL
- Text
- Bar
- Line
- Area
- Stacked Bar
- Stacked Line
- Donut
- Treemap
- Map
- Card
- Table
- Matrix
- Image
- Action
- SQL

Dimensions

- dest
- origin

Measure

- Record Count

Tooltips

drag fields to add here

X Trellis

drag fields to add here

Y Trellis

drag fields to add here

Filters

drag fields to add here

FIELD PROPERTIES

DASH.

Visuals

Filters

Settings

Style

BUILD

Enter/Edit Expression

Display Format

Alias

Description

Duplicate

Save Expression

Remove

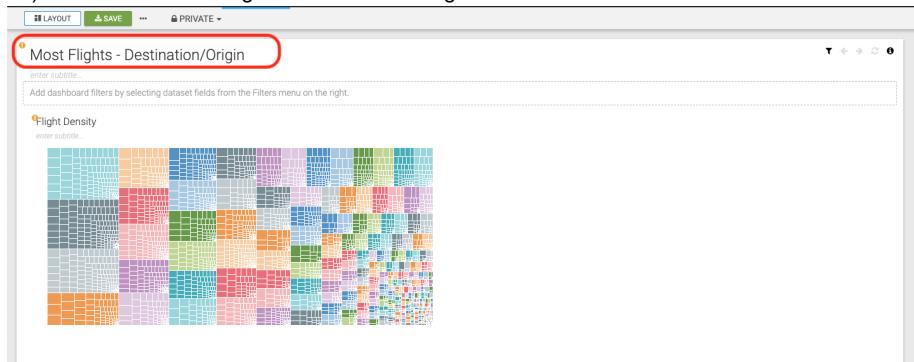
REFRESH VISUAL

The 'FIELD PROPERTIES' panel is open, showing various options for field manipulation. The 'Order and Top K' section is highlighted with a red box. It contains two radio buttons: 'Descending' (selected) and 'Ascending'. Below these are input fields for 'Top K:' (eg. 100) and 'Bottom K:' (eg. 100). A note states 'Top K/Bottom K applies to granular dimensions'. Other visible sections include 'Enter/Edit Expression', 'Display Format', 'Alias', 'Description', 'Duplicate', 'Save Expression', and 'Remove'.

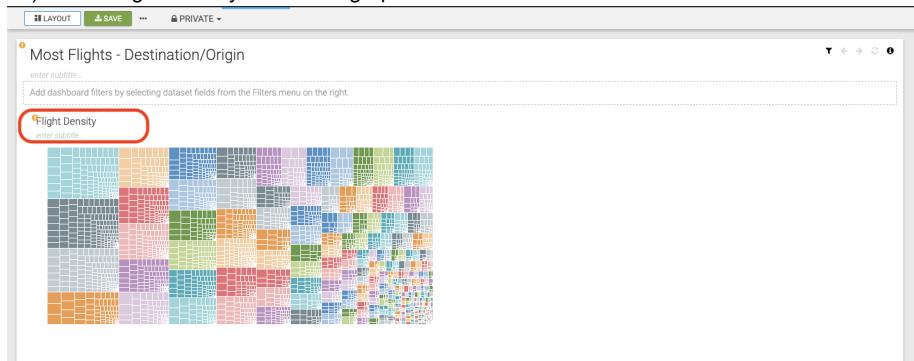
17) Click "REFRESH VISUAL"

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

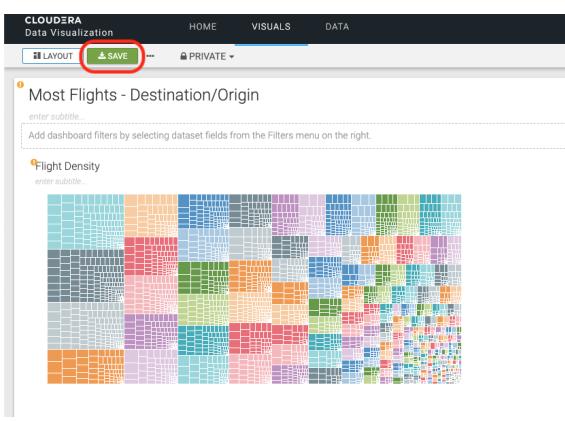
18) Enter a title "Most Flights - Destination/Origin"



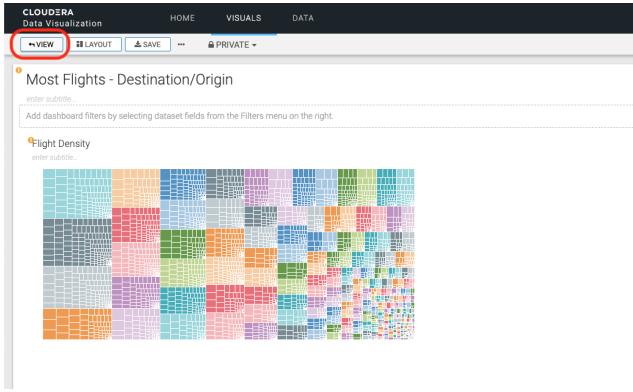
19) Enter "Flight Density" under the graph's title



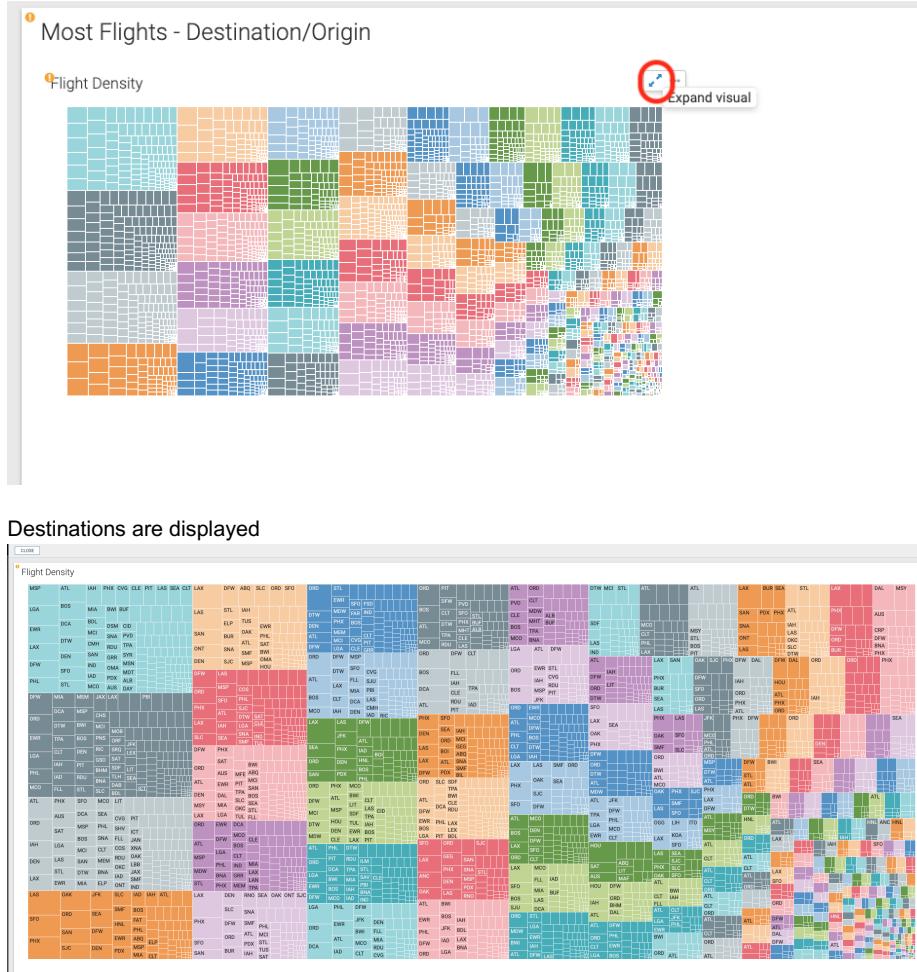
20) Click "SAVE"



21) Click "VIEW"



22) Scroll over the graph and click "Expand Visual"



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.

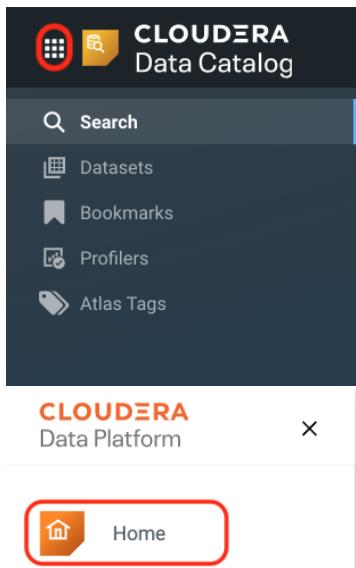
Your link should look something like (remember click the link in your email not the link below)
http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X
*xx represents the trial user #
*X represents the password

- 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

Virtual Warehouses

	testvirtualwarehouse1
Running	compute-1611179792-vz49 cdptrialuser24-dl-default
NODE COUNT	2
TOTAL CORES	38
TOTAL MEMORY	292 GB

	mschoeni-iso-1
Stopped	compute-1611173596-dbty cdptrialuser24-dl-default
NODE COUNT	0
TOTAL CORES	12
TOTAL MEMORY	56 GB

	default-vw
Stopped	compute-1611103491-4hbp
NODE COUNT	0
TOTAL CORES	0
TOTAL MEMORY	0 GB

Suspend Clone Edit Delete Upgrade

Copy JDBC URL Download JDBC Jar Open DAS Open Data Visualization Set Compactor Run AutoScaling Demo 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 trial user you're using and password defined by "X" in #1 above

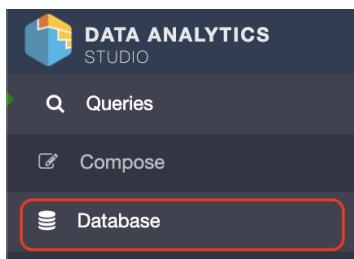
DATA ANALYTICS STUDIO

Username
trialxx_admin

Password

LOGIN

5) Click on Database on the left navigation bar



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

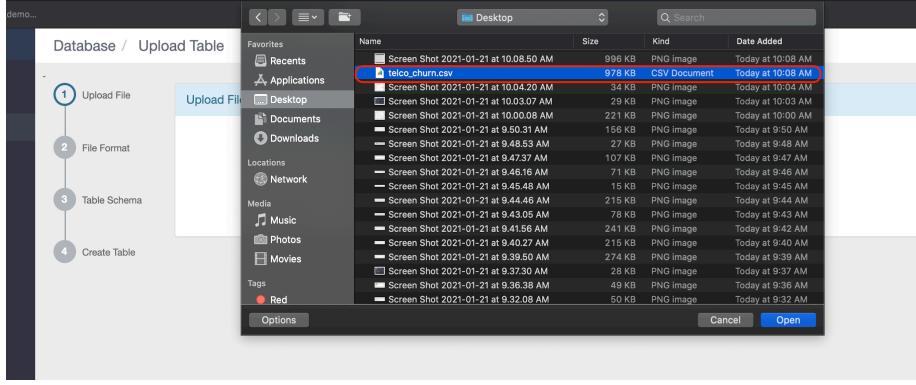
A screenshot of the Database Explorer. At the top, it says "LAST UPDATE: 7 sec ago DATABASES | 6". Below that is a section for the "default" database, which shows "TABLES | 0". To the right of this section are three icons: a refresh symbol, a plus sign, and an upward arrow (the "Upload Table" button).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

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

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	onthlyCharges	totalCharges	Churn						
7590-UNIVG	Female	0	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No					
5571-GHDE	Male	0	No	Yes	Yes	No	No	No	One year	No	Mailed check	56.91	1389.5	No					
3468-OPYKX	Male	0	No	2	Yes	No	DSL	Yes	Yes	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes			
7795-CFOQK	Male	0	No	45	Yes	No	phone service	DSL	Yes	No	One year	No	Bank transfer (automatic)	42.3	1840.75	No			
9237-HOI7U	Female	0	No	2	Yes	No	Fiber optic	No	No	No	Month-to-month	Yes	Electronic check	70.7	151.65	Yes			
9305-CDSKC	Female	0	No	8	Yes	Yes	Fiber optic	No	No	Yes	Month-to-month	Yes	Electronic check	99.65	820.5	Yes			
1452-KZOKV	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	Month-to-month	Yes	Credit card (automatic)	89.1	1949.4	No		
6713-OKOMC	Female	0	No	10	No	No	phone service	DSL	Yes	No	No	No	Month-to-month	No	Mailed check	29.75	301.9	No	
7892-POOKP	Female	0	Yes	No	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	62	Yes	No	DSL	Yes	Yes	No	One year	No	Bank transfer (automatic)	56.15	3487.95	Yes			
9763-GRSDK	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	Yes	No	No	Month-to-month	Yes	Mailed check	49.95	587.45	No	
7469-LKBCI	Male	0	No	16	Yes	No	No	internet service	No	internet service	No	internet service	No	internet service	No	internet service	No		
8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	Yes	No	Yes	One year	No	Credit card (automatic)	100.35	5681.1	No	
0280-XJGEK	Male	0	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	Month-to-month	Yes	Bank transfer (automatic)	103.7	5036.3	:			
5129-JLPIS	Male	0	No	25	Yes	No	Fiber optic	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	105.5	2686.05	No			
3655-SNOYZ	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	52	Yes	Yes	No	No	No	No	Month-to-month	No	Credit card (automatic)	106.7	7382.25	No			
9595-WOPFJ	Female	0	Yes	Yes	71	Yes	Yes	Fiber optic	Yes	Yes	No	Yes	Two year	No	Bank transfer (automatic)	106.7	7382.25	No	
4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	Yes	Yes	No	No	Month-to-month	No	Credit card (automatic)	55.2	528.35	Yes	
4183-MYFRB	Female	0	No	21	Yes	No	Fiber optic	No	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	90.05	1862.9	No		
8779-QRDMV	Male	1	No	1	Yes	No	phone service	DSL	No	No	No	No	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		
1066-JKSGK	Male	0	No	1	Yes	No	No	No	No	No	internet service	No	internet service	No	internet service	No			
3638-WEAWB	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	No	Two year	Yes	Credit card (automatic)	59.9	3505.1	No	
6322-HRPFA	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	No	Month-to-month	No	Credit card (automatic)	59.6	2970.3	No	
6865-JZJKO	Female	0	No	30	Yes	No	DSL	Yes	Yes	No	No	No	Month-to-month	Yes	Bank transfer (automatic)	55.3	1530.6	No	
6467-CHFWZ	Male	0	Yes	No	47	Yes	Yes	Fiber optic	No	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	99.35	4749.15	Yes	
8665-UTDHZ	Male	0	Yes	Yes	1	Yes	No	phone service	DSL	No	Yes	No	No	Month-to-month	No	Electronic check	30.2	30.2	Yes
5248-YGLJN	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	17	Yes	No	DSL	No	No	No	Yes	Yes	Month-to-month	Yes	Mailed check	64.7	1093.1	Yes	
3841-NPECX	Female	1	No	71	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	No	Two year	Yes	Credit card (automatic)	96.35	6766.95	No	
4929-XTHVW	Male	1	Yes	No	2	Yes	No	Fiber optic	No	Yes	No	Yes	Month-to-month	Yes	Credit card (automatic)	95.5	181.65	No	
6827-IEAUQ	Female	0	Yes	No	27	Yes	No	DSL	Yes	Yes	No	No	One year	No	Mailed check	66.15	1874.45	No	
7310-EGVHZ	Male	0	No	1	Yes	No	No	internet service	No	internet service	No	internet service	No	internet service	No	internet service	No		
3413-BMNZE	Male	1	No	1	Yes	No	DSL	Yes	Yes	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	Yes	Yes	Two year	No	Bank transfer (automatic)	99.9	7251.7	No	
6047-YHPVI	Male	0	No	5	Yes	No	Fiber optic	No	Yes	No	No	No	Month-to-month	Yes	Electronic check	69.7	316.9	Yes	
6572-ADKRS	Female	0	No	46	Yes	Yes	Fiber optic	No	Yes	No	No	No	Month-to-month	Yes	Credit card (automatic)	74.8	3548.3	No	
5380-WFKOV	Male	0	No	34	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Month-to-month	Yes	Electronic check	106.35	3549.25	Yes	
8168-UQWPF	Female	0	No	11	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Month-to-month	Yes	Bank transfer (automatic)	97.85	1105.4	No	
8865-TNNMX	Male	0	Yes	Yes	10	Yes	No	DSL	No	Yes	No	No	One year	No	Mailed check	49.55	475.7	No	
9489-DEVDP	Female	0	Yes	Yes	70	Yes	Yes	DSL	Yes	Yes	No	No	Two year	Yes	Credit card (automatic)	69.2	4872.35	No	
9867-JCZSP	Female	0	Yes	Yes	17	Yes	No	No	No	internet service	No	internet service	No	internet service	No	internet service	No		
4671-VJLCL	Female	0	No	63	Yes	Yes	DSL	Yes	Yes	Yes	No	Two year	Yes	Credit card (automatic)	79.45	4861.45	No		
4080-FLARD	Female	0	Yes	No	13	Yes	Yes	DSL	Yes	Yes	Yes	No	Month-to-month	Yes	Electronic check	76.2	981.45	No	
3714-NTNPO	Female	0	No	49	Yes	Yes	Fiber optic	No	Yes	No	No	No	Month-to-month	Yes	Electronic check	84.5	3906.7	No	
5948-USZLF	Male	0	No	5	Yes	No	DSL	No	Yes	No	No	No	Month-to-month	No	Mailed check	49.25	97.9	No	
7760-OPYD	Female	0	No	52	Yes	Yes	No	Fiber optic	No	Yes	No	No	Month-to-month	Yes	Electronic check	80.65	144.15	Yes	
7639-MLRQW	Male	0	No	52	Yes	Yes	DSL	Yes	Yes	No	No	No	One year	No	Credit card (automatic)	79.75	4317.8	No	
2954-PBKQ	Female	0	Yes	Yes	69	Yes	Yes	DSL	Yes	Yes	No	No	Two year	Yes	Credit card (automatic)	64.15	4254.1	No	
8012-SOUDQ	Female	1	No	43	Yes	Yes	Fiber optic	No	Yes	No	No	No	Month-to-month	Yes	Electronic check	90.25	3838.75	No	
9420-LOJRX	Female	0	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	Yes	No	No	Month-to-month	Yes	Credit card (automatic)	69.5	1752.65	No	
7495-OKKPV	Female	1	Yes	No	8	Yes	Yes	Fiber optic	No	Yes	No	No	Month-to-month	Yes	Credit card (automatic)	80.65	613.3	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

Select File Format

File type	CSV	<input type="button" value="Clear"/>
Field Delimiter	,	<input type="button" value="Clear"/>
Escape Character	\	<input type="button" value="Clear"/>
Quote Character	"	<input type="button" value="Clear"/>

Is first row header?

Contains endlines?

PREVIEW

10) Click “PREVIEW” prior to creating the table

PREVIEW

Table Preview

CUSTOMERID	GENDER	SENIORCITIZEN	PARTNER	DEPENDENTS	TENURE	PHONESERVICE	MULTIPLELINES	INTERNETSERVICE	ONLINESECURITY	ONLINESBACKUP	DEVICEPROTECTION	TECHSUPPORT	STREAMINGTV	STREAMINGMOVIES	CONTRV
7590-VHWEZ	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-month
5575-GNVDIE	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** **CANCEL**

11) Click "NEXT"

PREVIEW

Table Preview

CUSTOMERID	GENDER	SENIORCITIZEN	PARTNER	DEPENDENTS	TENURE	PHONESERVICE	MULTIPLELINES	INTERNETSERVICE	ONLINESECURITY	ONLINESBACKUP	DEVICEPROTECTION	TECHSUPPORT	STREAMINGTV	STREAMINGMOVIES	CONTRV
7590-VHWEZ	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-month
5575-GNVDIE	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** **CANCEL**

12) Enter 'telco_churn' as the Table Name. Click "CREATE"

Database / Upload Table

Table Name: telco_churn

COLUMNS

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

+ CREATE

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

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

Compose

TELCO_CHURN.CUSTOMERID	TELCO_CHURN.GENDER	TELCO_CHURN.SENIORCITIZEN	TELCO_CHURN.PARTNER	TELCO_CHURN.DEPENDENTS	TELCO_CHURN.TENURE	TELCO_CHURN.PHONESERVICE	TELCO_CHURN.MULTIPLINES	TELCO_CHURN
7590-WHVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVED	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL

Parking lot items

- Show Impala/Hue - for CDH customers