

Data Lifecycle on CDP Public Cloud

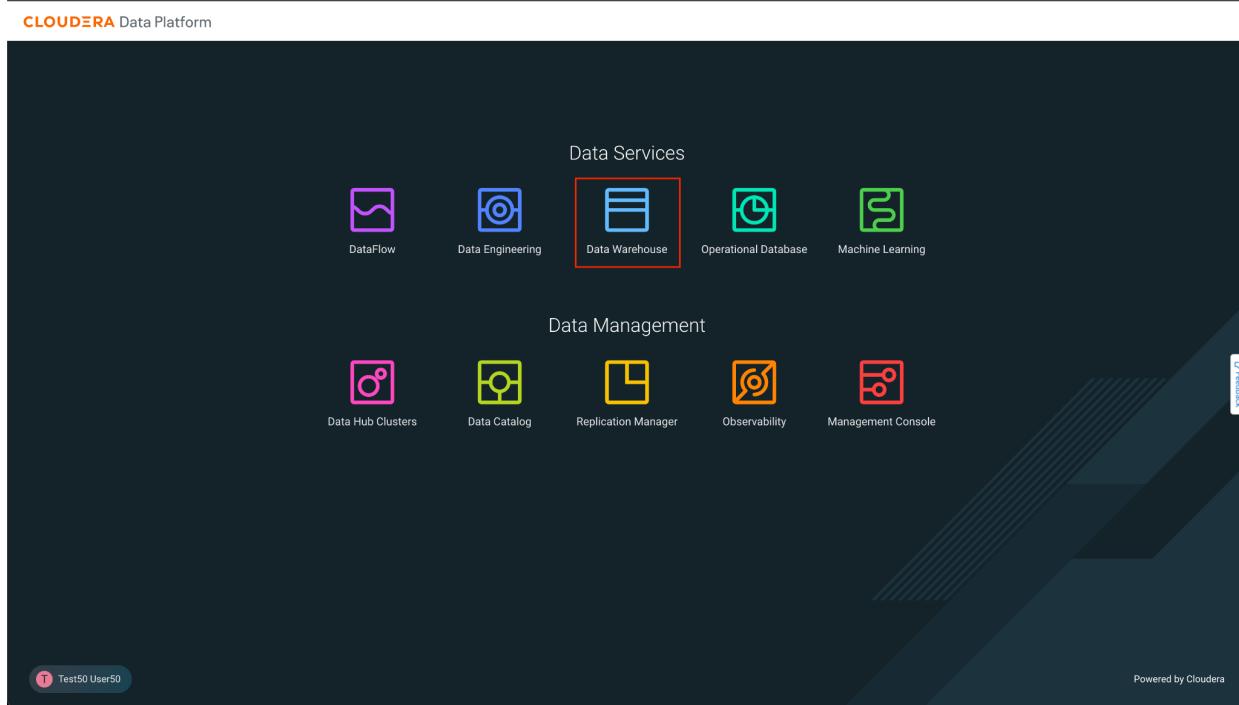
Data Warehouse Lab

Part 1: Dashboard development

Goals:

- Create a dataset pointing to the table
- Create a dashboard with metrics and dimensions

1. Click on Data Warehouse from CDP PC Home:



2. Data Warehouse welcome screen. Click on Data Visualization in the left menu.

3. In Data Visualization, click on the button **Data Viz** from which they were assigned.

NAME	DATA VISUALIZATION ID	Environment ID	VERSION	CPU	MEMORY	UPTIME	CREATED BY	
dataviz-0	viz-1685400615-2kkq	env-r9gpp	7.1.1-b30	2	8 GB	an hour	acampos	Data Viz

4. Once in Data Visualization, go to the Data option from the top menu, and then to the Connector **G1/G2/G3/G4/G5** from the left menu.

The screenshot shows the Datasets page in a data visualization tool. On the left, there's a sidebar with connection management options: '% NEW CONNECTION', 'All Connections', 'ImpalaConn' (which is highlighted with a red box), and 'samples'. The main area displays a table of datasets. The columns are: Title/Table, ID, Created, Last Updated, Modified By, and # Dashboards. The rows list various datasets such as 'Food Stores Inspection in NYC', 'Cereals', 'World Life Expectancy', 'Earthquake Data January 2019', 'US State Populations Over Time', 'US County Population', 'Global Information Security Threats', and 'Restaurant Inspection SF'. Each dataset row includes a preview icon, an edit icon, and a delete icon.

5. We have to create a new data source, for that, click on New Dataset and a window will appear to enter the information of the new data source.

The screenshot shows the 'NEW DATASET' creation window. The top bar has buttons for '% NEW CONNECTION', 'NEW DATASET' (which is highlighted with a red box), 'ADD DATA', and '...'. Below the bar is a header with 'Datasets' and 'Connection Explorer'. The main area is a table titled 'Title/Table' with columns: ID, Created, Last Updated, Modified By, and # Dashboards. A single row is present with the status 'No data'.

6. Enter the information for the new data source:

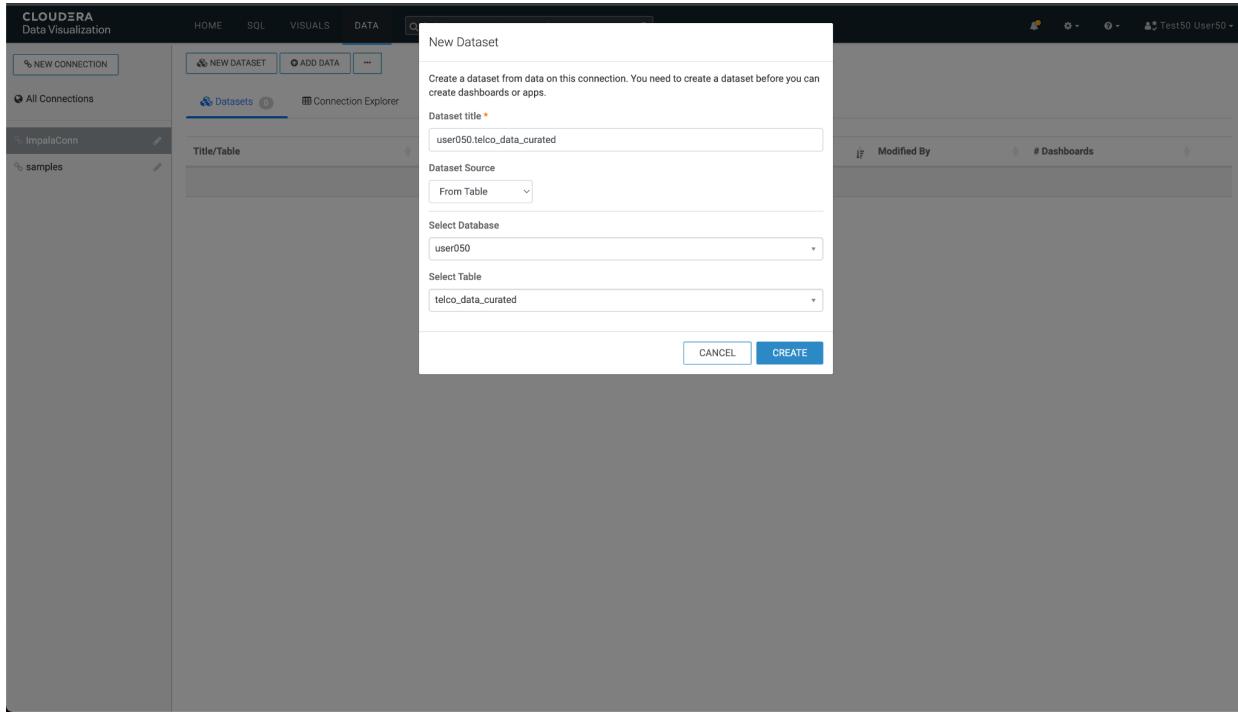
Dataset title: <assigned_user>.telco_curated_data

Dataset Source: From table

Select Database: <assigned_user>

Select Table: telco_data_curated

Click on Create to create the new Dataset.



7. The new Dataset should appear in the list. Click on the dataset that you just created.

Title/Table	ID	Created	Last Updated	Modified By	# Dashboards
user050 telco_data_curated	16	May 29, 2023	a few seconds ago	user050	0
user050.telco_data_curated					

8. Here you will see the details of the dataset.

The screenshot shows the 'Dataset Detail' page for the dataset 'user050.telco_data_curated'. The left sidebar contains navigation links: Dataset Detail, Related Dashboards, Fields, Data Model, Time Modeling, Segments (0), Filter Associations (0), and Permissions. The main content area displays dataset details: Dataset: user050.telco_data_curated, Table: user050.telco_data_curated, Connection Type: Impala, Data Connection: ImpalaConn, Description: (empty), Join Elimination: Enabled, Result Cache: From Connection, Incremental Results: Disabled. Below these are creation and update logs: ID: 16, Created on: May 29, 2023 06:15 PM, Created by: user050, Last updated: May 29, 2023 06:15 PM, Last updated by: user050. At the top right are 'CLONE DATASET' and 'NEW DASHBOARD' buttons.

9. Click on **Fields** (left menu) to see the fields automatically captured during the dataset creation process.

The screenshot shows the 'Fields' page for the dataset 'user050.telco_data_curated'. The left sidebar is identical to the previous screenshot. The main content area shows the 'Fields' section with a 'Dimensions' panel containing 19 items: multiplelines, paperlessbilling, gender, onlinesecurity, internetservice, techsupport, contract, churn, seniorcitizen, deviceprotection, streamingtv, streamingmovies, partner, customerid, dependents, onlinebackup, phoneservice, and paymentmethod. To the right is a 'Measures' panel with 3 items: totalcharges, monthlycharges, and tenure. At the top right are 'Edit Fields' and 'New Dashboard' buttons.

10. You can also preview the data from this screen. Click on **Data Model** (left menu) and then on the button **Show Data** that appears in the center.

The screenshot shows the Cloudera Data Visualization interface. The left sidebar has a 'Data Model' section selected. In the main area, there is a dataset named 'telco_data_curated'. A prominent blue button labeled 'SHOW DATA' is centered in the main content area, with a red box drawn around it. Below this button is a checkbox labeled 'Apply Display Format'. At the top right of the main area, there is a 'NEW DASHBOARD' button.

11. At this moment, a query to the Virtual Warehouse is executed to retrieve the data from the data set. Notice the columns and values. Click New Dashboard to create a new dashboard.

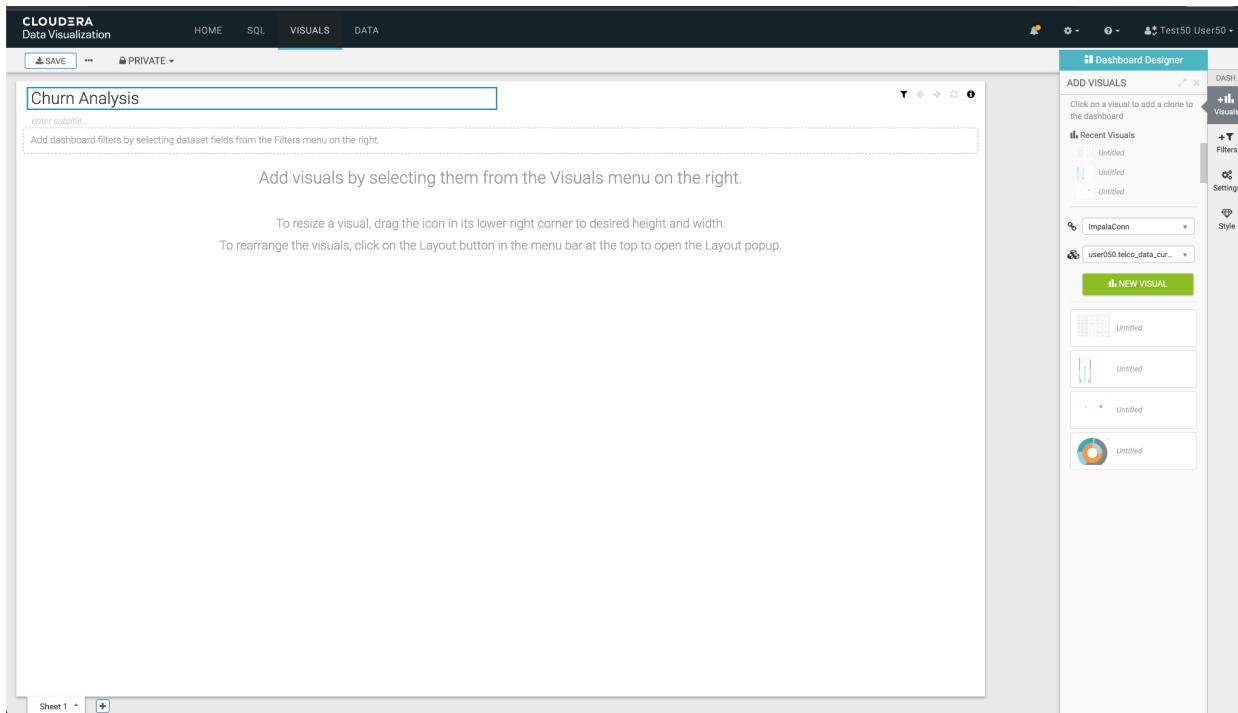
This screenshot shows the same Cloudera Data Visualization interface as the previous one, but the 'SHOW DATA' button has been clicked, revealing a detailed data preview. The preview is titled 'telco_data_curated' and displays a table with 19 columns. The columns are: multipelines, paperlessbilling, gender, onlinesecurity, internetservice, techsupport, contract, churn, seniorcitizen, deviceprotection, streamingtv, streamingmovies, totalcharges, partner, monthlycharges, customerid, and de. The data consists of 10 rows of information about telephone service users. A red box highlights the 'HIDE DATA' button at the top of the preview area. The 'NEW DASHBOARD' button is also visible at the top right.

multipelines	paperlessbilling	gender	onlinesecurity	internetservice	techsupport	contract	churn	seniorcitizen	deviceprotection	streamingtv	streamingmovies	totalcharges	partner	monthlycharges	customerid	de
No phone service	Yes	Female	No	DSL	No	Month-to-month	No	0	No	No	No	29.850000381469727	Yes	32.602622985839844	7590-VHVEG	Ni
No	No	Male	Yes	DSL	No	One year	No	0	Yes	No	No	1889.5	No	79.32872009277344	5575-GNVE	Ni
No	Yes	Male	Yes	DSL	No	Month-to-month	Yes	0	No	No	No	108.1500015258789	No	53.849998474121094	Q97B/KYBVK	Ni
No phone service	No	Male	Yes	DSL	Yes	One year	No	0	Yes	No	No	1840.75	No	39.008785247802734	7795-CFCW	Ni
No	Yes	Female	No	Fiber optic	No	Month-to-month	Yes	0	No	No	No	151.64999389648438	No	70.69999694824219	9237-HQITU	Ni
Yes	Yes	Female	No	Fiber optic	No	Month-to-month	Yes	0	Yes	Yes	Yes	820.5	No	99.6500015258789	9305-CDSKC	Ni
Yes	Yes	Male	No	Fiber optic	No	Month-to-month	No	0	No	Yes	No	1949.4000244140625	No	154.11448669433594	1452-KIOVK	Ye
No phone service	No	Female	Yes	DSL	No	Month-to-month	No	0	No	No	No	301.8999938964844	No	46.75687789916992	6713-OKOMC	Ni
Yes	Yes	Female	No	Fiber optic	Yes	Month-to-month	Yes	0	Yes	Yes	Yes	3046.050048828125	Yes	104.80000305175781	7892-POOKP	Ni

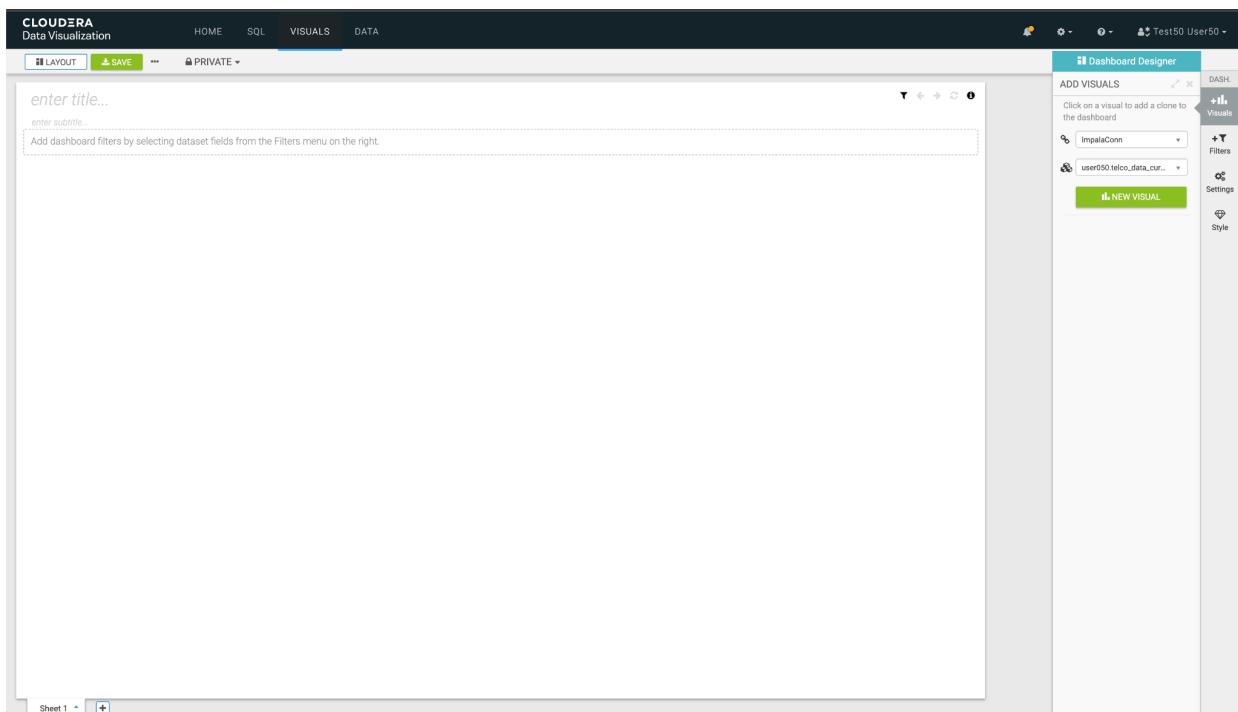
12. When opening the design canvas of a new panel, remove the element that is added by default, by clicking on the three dots (...) button at the top right of the element, and then clicking on the option **Delete Visual**

The screenshot shows the Cloudera Data Visualization interface. On the left, there's a dashboard canvas with a table visualization. A context menu is open over the table, with the 'Delete Visual' option highlighted. The interface includes a top navigation bar with tabs for HOME, SQL, VISUALS, and DATA, and a sidebar on the right with sections for DASH, VISUAL, and BUILD.

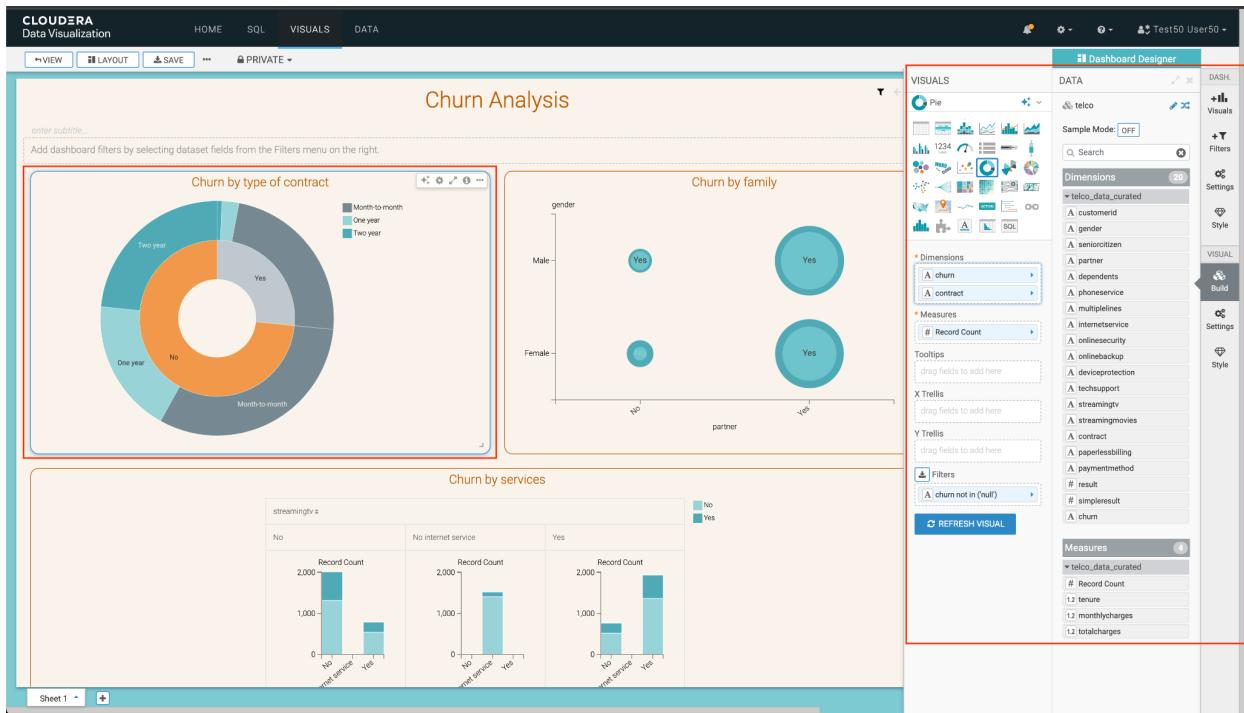
At the top of the canvas, in the enter title field, enter the name *Churn Analysis* to identify the dashboard.



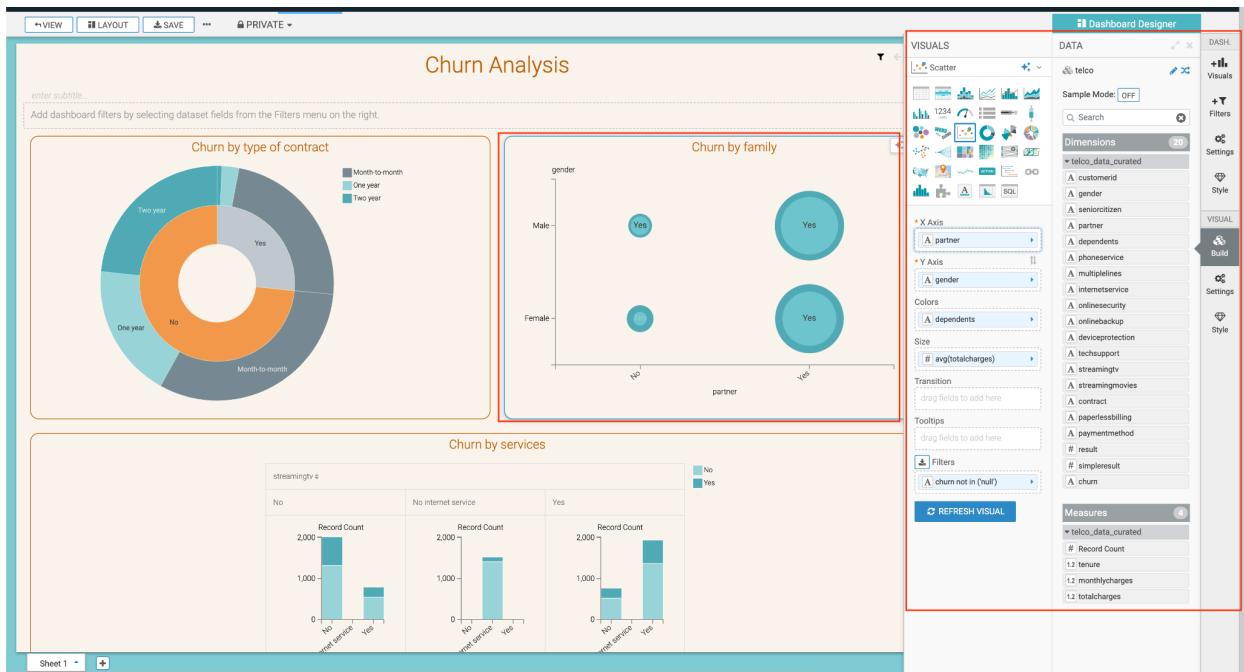
13. To add a new visual element, click on the button **Visuals** from the right menu, select the dataset that corresponds to them, and click on the button **New Visual**.



14. Add the first visual element, which is a pie chart with the dimensions **churn** and **contract**, with the metric of **Record count**. Once finished, click the button **Refresh Visual**.

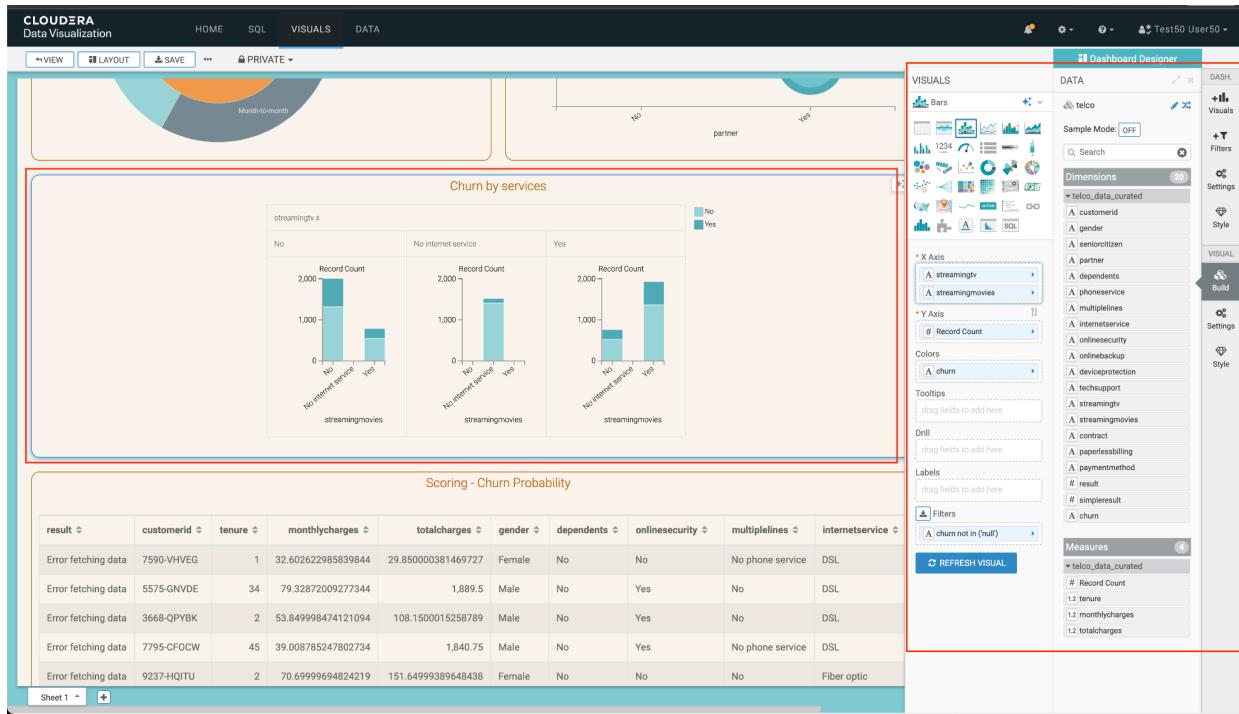


15. Add the second visual element, which is a scatter chart with the dimension **partner** like X Axis, **gender** how Y Axis, **dependents** as Colors and **avg (total charges)** as Size. Once finished, click the button **Refresh Visual**.



15. Add the third visual element, which is a bar chart with the dimensions **streamingtv** and **streamingmovies** like X Axis,

Record Count how Y Axis and **churn** how Colors. Once finished, click the button **Refresh Visual**.



16. Add the fourth and last visual element, which is a table with the dimensions and metrics of the dataset. Be sure to add all 17 dimensions and 3 metrics to the table. Once finished, click the button **Refresh Visual**.

The screenshot shows a Cloudera Data Visualization dashboard. At the top, there are three bar charts under the heading 'No internet service'. The first chart shows 'Record Count' for 'No' and 'Yes' categories. The second chart shows 'Record Count' for 'No' and 'Yes' categories. The third chart shows 'Record Count' for 'No' and 'Yes' categories. Below these charts is a table titled 'Scoring - Churn Probability'.

customerid	tenure	monthlycharges	totalcharges	gender	dependents	onlinesecurity	multiplelines	internetservice	seniorcitizen
7590-VHVEG	1	32.602622985839844	29.850000381469727	Female	No	No	No phone service	DSL	0
5575-GNVDE	34	79.32872009277344	1,889.5	Male	No	Yes	No	DSL	0
3668-QPYBK	2	53.849998474121094	108.1500015258789	Male	No	Yes	No	DSL	0
7795-CFOCW	45	39.008785247802734	1,840.75	Male	No	Yes	No phone service	DSL	0
9237-HQITU	2	70.69999694824219	151.64999389648438	Female	No	No	No	Fiber optic	0
9305-CDSKC	8	99.6500015258789	820.5	Female	No	No	Yes	Fiber optic	0
1452-KIOVK	22	154.11448669433594	1,949.4000244140625	Male	Yes	No	Yes	Fiber optic	0
6713-OKOMC	10	46.75687789916992	301.8999938964844	Female	No	Yes	No phone service	DSL	0

Save the dashboard by clicking the button **Save** from the top menu.

Part 2: Add new field

Goals:

- Add a new field that makes calls to the ML model
- Add the new field to the dashboard

1. Edit the previously created Dataset, in Data -> <user_assigned>.telco_data_curated.

The screenshot shows the Cloudera Data Visualization interface. On the left, there's a sidebar with connection management (New Connection, All Connections, ImpalaConn, samples). The main area is titled 'Datasets' and shows a table with one dataset entry:

Title/Table	ID	Created	Last Updated	Modified By	# Dashboards
user050.telco_data_curated	16	May 29, 2023	a few seconds ago	user050	0

2. Once in the Dataset, go to **Fields** in the left menu and then click on **Edit Field** to edit the fields of your dataset.

The screenshot shows the 'Fields' page for the 'telco_data_curated' dataset. The left sidebar has sections for Dataset Detail, Related Dashboards, Fields, Data Model, Time Modeling, Segments, Filter Associations, and Permissions. The main area is divided into 'Dimensions' and 'Measures' sections. The 'Dimensions' section lists fields like multiplelines, paperlessbilling, gender, onlinesecurity, internetservice, techsupport, contract, churn, seniorcitizen, deviceprotection, streamingtv, streamingmovies, partner, customerid, dependents, onlinebackup, phoneservice, and paymentmethod. The 'Measures' section lists totalcharges, monthlycharges, and tenure.

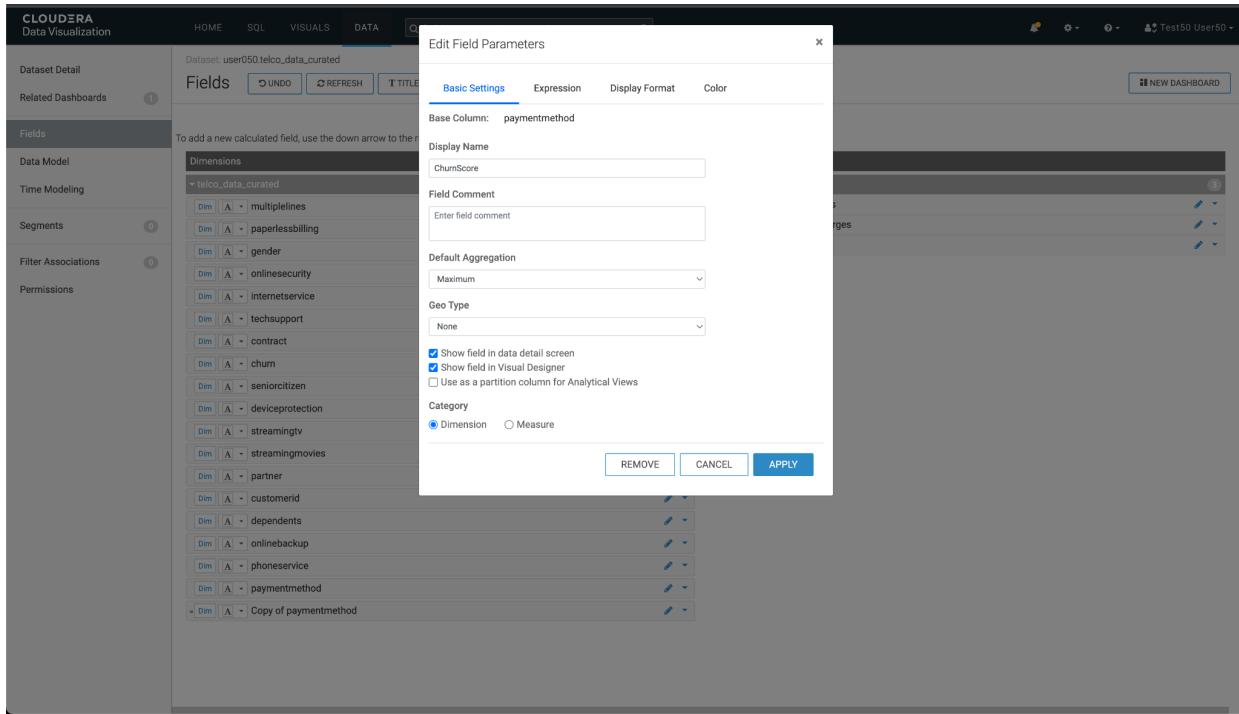
3. In the list of **Dimensions**, click the down arrow of the last field in the list, and select the option **Clone**.

The screenshot shows the Cloudera Data Visualization interface. On the left, a sidebar lists 'Dataset Detail', 'Related Dashboards' (with 1 item), 'Fields' (selected), 'Data Model', 'Time Modeling', 'Segments' (with 0 items), 'Filter Associations' (with 0 items), and 'Permissions'. The main area is titled 'Dataset: user050.telco_data_curated'. It has tabs for 'Fields' (selected), 'UNDOS', 'REFRESH', 'TITLE CASE', 'SAVE', and 'Show Comments'. A search bar at the top right says 'find titles, viz types, datasets, authors...'. On the right, there are 'NEW DASHBOARD' and user info ('Test50 User50'). The 'Fields' section contains two panels: 'Dimensions' (18 items) and 'Measures' (3 items). In the Dimensions panel, the last item is 'Dim [A] - paymentmethod'. Below these panels is a context menu with options: 'Clone' (highlighted), 'Hide', and 'Create Hierarchy'.

4. Once the field is cloned, click on the pencil next to the field to edit it.

This screenshot shows the same interface after cloning the 'paymentmethod' dimension. The 'Dimensions' panel now includes a new item: 'Dim [A] - Copy of paymentmethod'. To its right is a small 'Edit Field' button, which is highlighted with a red box. The rest of the interface remains the same, including the sidebar, dataset details, and the context menu for the cloned field.

5. In the popup window that appears, enter the name of the new field in **Display Name**. We suggest that you enter *ChurnScore*.



6. Go to the Expressions tab and enter the following value in the Expression field. This will allow you to call the REST API of the Model you have previously deployed.

```
cviz_rest('{"url":"<url_del_workspace>","accessKey":"<access_key>","colnames":["monthlycharges","totalcharges","tenure","gender","dependents","onlinesecurity","multiplelines","internetservice","seniorcitizen","techsupport","contract","streamingmovies","deviceprotection","paymentmethod","streamingtvtv","phoneservice","paperlessbilling","partner","onlinebackup"],"response_colname":"result"}')
```

The screenshot shows the Cloudera Data Visualization interface. A modal window titled 'Edit Field Parameters' is open, specifically on the 'Expression' tab. The expression input field contains a complex JSON object:

```

1 cviz_rest({"url": "curl_del_workspace", "access_key": "ec1cbe", "token": "Cmonthlycharges", "totalcharge": "1000", "gender": "dependents", "onlinesecurity": "multiplelines", "internetservice": "seniorcitizen", "techsupport": "contract", "streamingmovies": "deviceprotection", "paymentmethod": "streamingtv", "phoneservice": "paperlessbilling", "partner": "onlinebackup", "response_colname": "result"});
```

Below the expression input are several checkboxes: 'Expression contains an aggregation' (unchecked), 'Autocomplete on VALIDATE EXPRESSION' (checked), and 'Save expression only after validation succeeds' (checked). At the bottom of the dialog are 'REMOVE', 'CANCEL', and 'APPLY' buttons.

7. Being in CML in another tab of the web browser, go to the section of **Models** of your project, and click on the Model that begins with the name *Model/Viz*, followed by your assigned username.

The screenshot shows the Cloudera Machine Learning (CML) interface. On the left, a sidebar navigation includes 'All Projects', 'Overview' (selected), 'Sessions', 'Data', 'Experiments', 'Models' (selected), 'Jobs', 'Applications', 'Files', 'Collaborators', and 'Project Settings'. The main content area is titled 'user050 / user050-telco-churn'. It shows the 'Models' section with two entries:

Model	Source	Status	Replicas	CPU	Memory	Last Deployed	Actions
ModelViz_user050	13_mod...	Deployed	1 / 1	1	2.00 GiB	May 29, 2023, 03:54 PM	<button>Stop</button>
ModelOpsChurn_user050	11_best...	Deployed	1 / 1	1	2.00 GiB	May 29, 2023, 03:53 PM	<button>Stop</button>

Below the models is the 'Jobs' section:

Name	Runs / Failures	Duration	Status	Latest Run	Actions
deploy_best_model	0 / 0	00:00	Not Yet Run	-	<button>Run</button>
retrain	0 / 0	00:00	Not Yet Run	-	<button>Run</button>
avisoPerformance	0 / 0	00:00	Not Yet Run	-	<button>Run</button>
Check Model	0 / 0	00:00	Not Yet Run	-	<button>Run</button>

The 'Files' section shows a file tree:

- __pycache__
- flask
- images
- models
- raw
- O_bootstrap.py
- Ob_create_jobs.py

At the bottom, it says 'Workspace: ssa-cml-workspace' and 'Cloud Provider: aws (AWS)'.

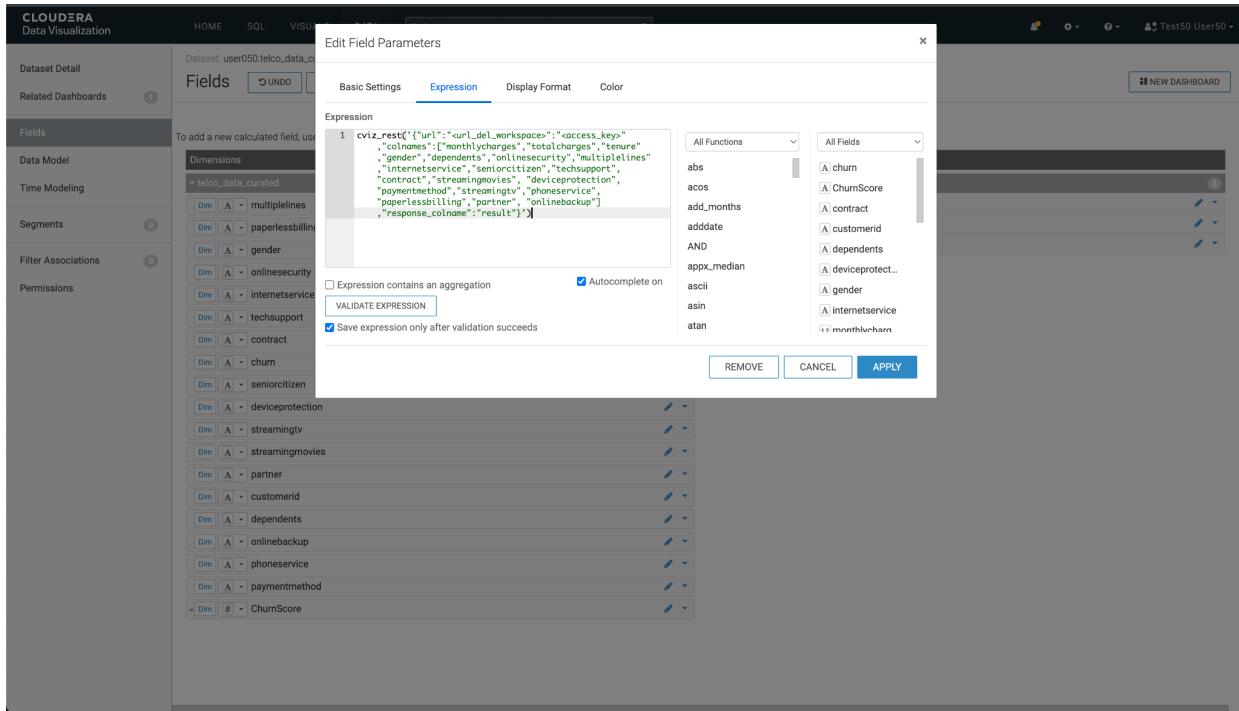
8. In the Overview tab, copy the URL that allows you to interact and call the workspace API.

Replace the copied value in the attribute <url_del_workspace> of the Expression field.

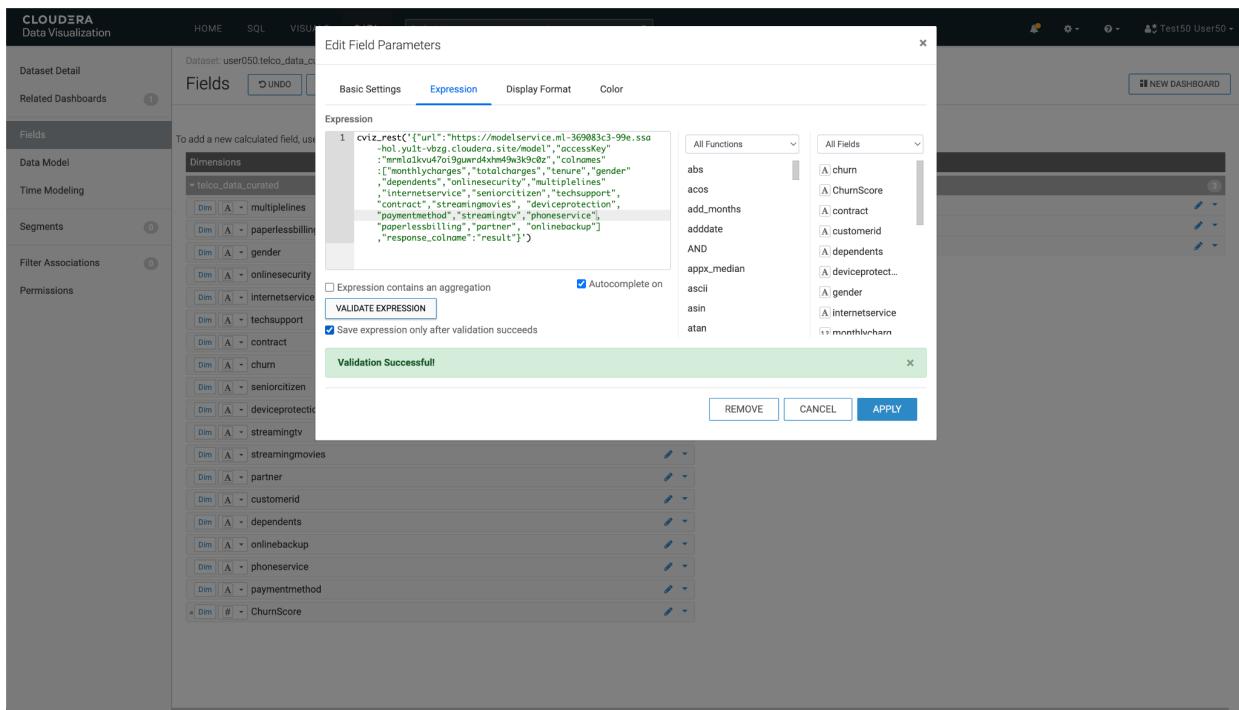
9. Returning to the CML, copy the accessKey of the model.

Replace the copied value in the attribute `<access_key>` of the Expression field. The format should be as follows, e.g.

```
cviz_rest('{"url":"https://modelservice.ml-b200bd6f-fb9.za-mtn-l.yu1t-vbzg.cloudera.site/model","accessKey":"mjy1fowabqiwpfjb19s9ht6xmuvy0f2j","colnames":["monthlycharges","totalcharges","tenure","gender","dependents","onlinesecurity","multiplelines","internetservice","seniorcitizen","techsupport", "contract", "streamingmovies", "deviceprotection", "paymentmethod", "streamingtvtv", "phoneservice", "paperlessbilling", "partner", "onlinebackup"],"response_colname":"result"})
```



10. Finish the process of copying the *url del workspace* and the *accessKey*, click the Validate Expression button at the top of the window. If the message appears in green *Validation Successful*, Click on **Apply** to save the settings made.



11. The new field should appear in the list of fields. Change the data type, selecting the type **Integer**, which is represented by the symbol #

The screenshot shows the Cloudera Data Visualization interface. The left sidebar has sections for Dataset Detail, Related Dashboards, Fields (selected), Data Model, Time Modeling, Segments, Filter Associations, and Permissions. The main area is titled 'Dataset: user050.telco_data_curated'. It shows two panels: 'Dimensions' on the left and 'Measures' on the right. In the Dimensions panel, there are many dimensions listed under 'telco_data_curated'. In the Measures panel, there are three measures: 'totalcharges', 'monthlycharges', and 'tenure'. At the bottom of the Dimensions list, there is a context menu with options: Boolean, Integer, Real, String, Timestamp, Remove CAST, and Integer. The 'Integer' option is highlighted with a blue background.

12. Finish the process by clicking on the green button with the legend **SAVE** in the top menu.

This screenshot shows the same interface as the previous one, but the 'ChurnScore' field is now listed in the 'Dimensions' panel under 'telco_data_curated'. The context menu is no longer open, and the 'SAVE' button in the top menu bar is visible but not highlighted.

13. Return to the dashboard, selecting the option **VISUALS** from the top menu, and clicking on the name of the dashboard that was previously created.

The screenshot shows the Cloudera Data Visualization interface. At the top, there's a navigation bar with 'HOME', 'SQL', 'VISUALS', and 'DATA' tabs. Below the navigation bar is a search bar and some action buttons: 'MOVE TO WORKSPACE', 'EXPORT', and 'DELETE'. On the left side, there's a sidebar with sections for 'All', 'My Favorites', 'WORKSPACES' (Public and Private), and a 'Sample Dashboards' section. The main area displays a grid of dashboard thumbnails. One specific dashboard, 'Churn Analysis', is highlighted with a red box. Other visible dashboard titles include 'Deficiency Details', 'State of NYC', 'Sample App', 'Store Details', 'Cereal Comparisons', 'Earthquakes Around the World', 'Life Expectancy Dashboard', 'World Population & GDP Trends', 'Animated world population - GDP vs life expectancy', 'US State Population Trends', 'Census Dashboard', 'Global Threats', 'Time & Industry Threat View', 'Inspector View', 'Consumer View', 'Iris species w/ images', and 'Taxi rides application'.

14. Once in the dashboard, click on the button **Edit** which is in the upper left.

The screenshot shows the 'streamingtv' dashboard in edit mode. At the top, there's a toolbar with 'EDIT' and 'PRIVATE' buttons. The main area contains three bar charts under the heading 'streamingtv'. Each chart has 'Record Count' on the y-axis and categories 'No internet service' and 'Yes' on the x-axis. The first chart is for 'streamingmovies', the second for 'streamingtv', and the third for 'partner'. Below the charts is a detailed data table with columns: totalcharges, monthlycharges, tenure, multiplelines, paperlessbilling, gender, onlinesecurity, internetservice, techsupport, contract, and chu. The table contains several rows of data. At the bottom right of the dashboard area, there are page navigation buttons (1, 2, 3, 4, 5, >).

15. Edit the lower table by clicking on it and then on the option **Build** from the right vertical menu. Add the new field, **ChurnScore**, at the beginning of the table, by clicking and dragging from the option **Dimensions** available.

The screenshot shows the Cloudera Data Visualization interface. At the top, there are navigation tabs: HOME, SQL, VISUALS, and DATA. Below the tabs, there are buttons for VIEW, LAYOUT, SAVE, and PRIVATE. On the right side, there is a vertical toolbar with various icons: DASH, Visuals, Filters, Settings, Build, and Style. The 'Build' icon is highlighted.

In the center, there are three stacked bar charts under the heading 'streamingtv'. Each chart has 'Record Count' on the y-axis (0 to 2,000) and 'No' and 'Yes' on the x-axis. The first chart is for 'No internet service', the second for 'No streamingmovies', and the third for 'streamingmovies'.

Below the charts is a table with the following data:

	totalcharges	monthlycharges	tenure	multiplelines	paperlessbilling	gender	onlinesecurity	internetservice	techsupport	contract
29.850000381469727	32.602622985839844	1	No phone service	Yes	Female	No	DSL	No	Month-to-month	
1,889.5	79.32872009277344	34	No	No	Male	Yes	DSL	No	One year	
108.1500015258789	53.8499998474121094	2	No	Yes	Male	Yes	DSL	No	Month-to-month	
1,840.75	39.008785247802734	45	No phone service	No	Male	Yes	DSL	Yes	One year	
151.64999389648438	70.69999694824219	2	No	Yes	Female	No	Fiber optic	No	Month-to-month	
820.5	99.6500015258789	8	Yes	Yes	Female	No	Fiber optic	No	Month-to-month	

At the bottom left, it says 'Sheet 1'. On the right side of the table, there are buttons for 'REFRESH VISUAL' and 'Limit: 100'.

16. Click on the Refresh Visual button to update the data. The new column should appear *ChurnScore* then at the beginning of the table, with a value of numeric type. Finish the process by clicking the button **SAVE** from the top left menu.

CLOUDERA Data Visualization

HOME SQL VISUALS DATA

VIEW LAYOUT SAVE PRIVATE

streamingtv \$

No internet service Yes

Record Count

streamingmovies

Record Count

streamingmovies

Record Count

streamingmovies

Dimensions

- # ChurnScore
- I2 totalcharges
- I2 monthlycharges
- I2 tenure
- A multiplelines
- A paperlessbilling
- A gender
- A onlinesecurity
- A internetservice
- A techsupport
- A contract
- A churn
- A seniorcitizen
- A deviceprotection
- A streamingtv
- A streamingmovies
- A partner
- A customerid
- A dependents
- A onlinebackup
- A phoneservice
- A paymentmethod
- # ChurnScore

Measures

- telco_data_curated
- # Record Count
- I2 totalcharges
- I2 monthlycharges
- I2 tenure

Filters

Search

DASH.

Visuals

Settings

Build

Style

Build

Style

DATA

Sample Mode: OFF

Search

Dimensions

- telco_data_curated
- A multiplelines
- A paperlessbilling
- A gender
- A onlinesecurity
- A internetservice
- A techsupport
- A contract
- A churn
- A seniorcitizen
- A deviceprotection
- A streamingtv
- A streamingmovies
- A partner
- A customerid
- A dependents
- A onlinebackup
- A phoneservice
- A paymentmethod
- # ChurnScore

Measures

- drag fields to add here

Tooltips

- drag fields to add here

Filters

- drag fields to add here

Limit: 100

REFRESH VISUAL

Sheet 1 +

ChurnScore	totalcharges	monthlycharges	tenure	multiplelines	paperlessbilling	gender	onlinesecurity	internetservice	techsupport
0	29.850000381469727	32.602622985839844	1	No phone service	Yes	Female	No	DSL	No
0	1,889.5	79.32872009277344	34	No	No	Male	Yes	DSL	No
0	108.1500015258789	53.849998474121094	2	No	Yes	Male	Yes	DSL	No
0	1,840.75	39.008785247802734	45	No phone service	No	Male	Yes	DSL	Yes
6	151.64999389648438	70.69999694824219	2	No	Yes	Female	No	Fiber optic	No
10	820.5	99.6500015258789	8	Yes	Yes	Female	No	Fiber optic	No