

Data Lifecycle en CDP Public Cloud

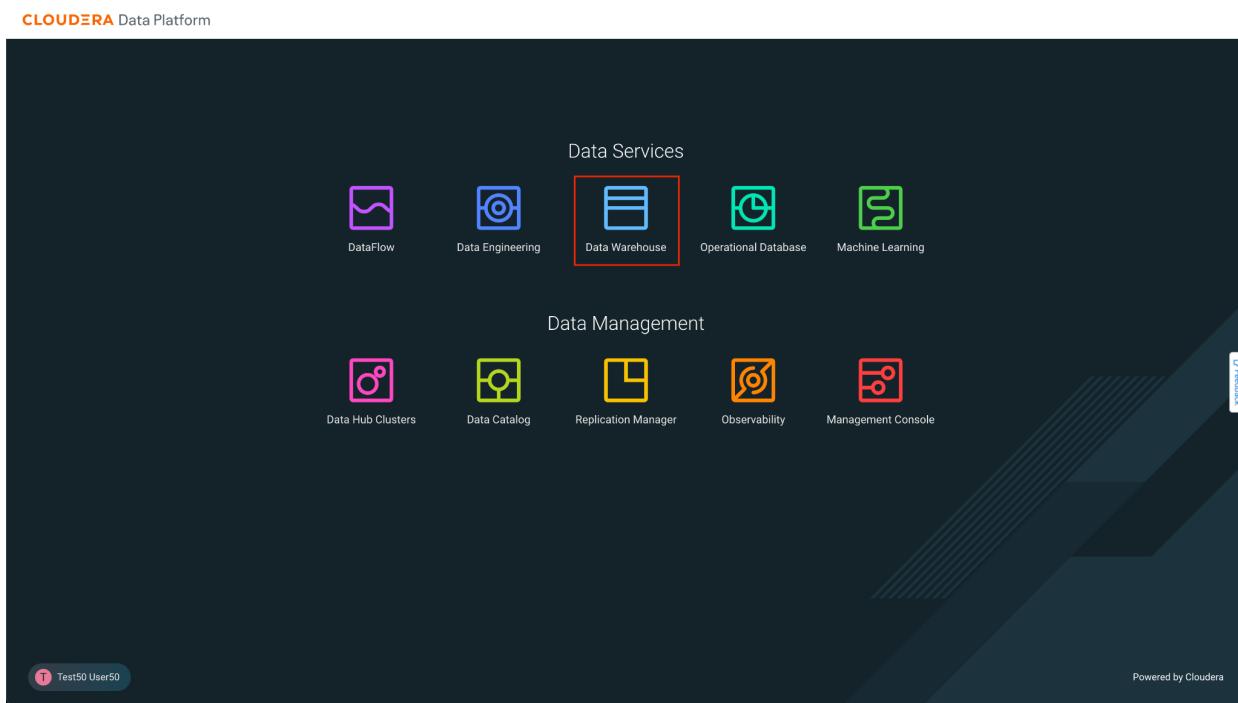
Data Warehouse Lab

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-rggpp	7.1.1-b30	2	8 GB	an hour	acampos

4. Once in Data Visualization, go to the Data option from the top menu, and then to the Connector **ImpalaConn** 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 with the following columns: Title/Table, ID, Created, Last Updated, Modified By, and # Dashboards. The datasets listed are:

Title/Table	ID	Created	Last Updated	Modified By	# Dashboards
Food Stores Inspection in NYC main.retail_food_store.inspections_current_critical_vio...	12	May 29, 2023	a few seconds ago	vizapps_admin	3
Cereals main.cereals	11	May 29, 2023	a few seconds ago	vizapps_admin	1
World Life Expectancy main.world_life.expectancy	9	May 29, 2023	a few seconds ago	vizapps_admin	1
Earthquake Data January 2019 main.earthquake_data2019	10	May 29, 2023	a few seconds ago	vizapps_admin	1
US State Populations Over Time main.census_pop	7	May 29, 2023	a few seconds ago	vizapps_admin	1
US County Population main.us_counties	8	May 29, 2023	a few seconds ago	vizapps_admin	1
Global Information Security Threats main.infosec_1559	6	May 29, 2023	a few seconds ago	vizapps_admin	1
Restaurant Inspection SF main.restaurant_scores_lives_standard	5	May 29, 2023	a few seconds ago	vizapps_admin	1

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. At the top, there are three buttons: '% NEW CONNECTION', 'NEW DATASET' (which is highlighted with a red box), 'ADD DATA', and '...'. Below the buttons, there's a 'Datasets' section with a 'Connection Explorer' tab. The main area is a table titled 'Title/Table' with columns: ID, Created, Last Updated, Modified By, and # Dashboards. The table currently displays the message 'No data'.

6. Enter the information for the new data source:

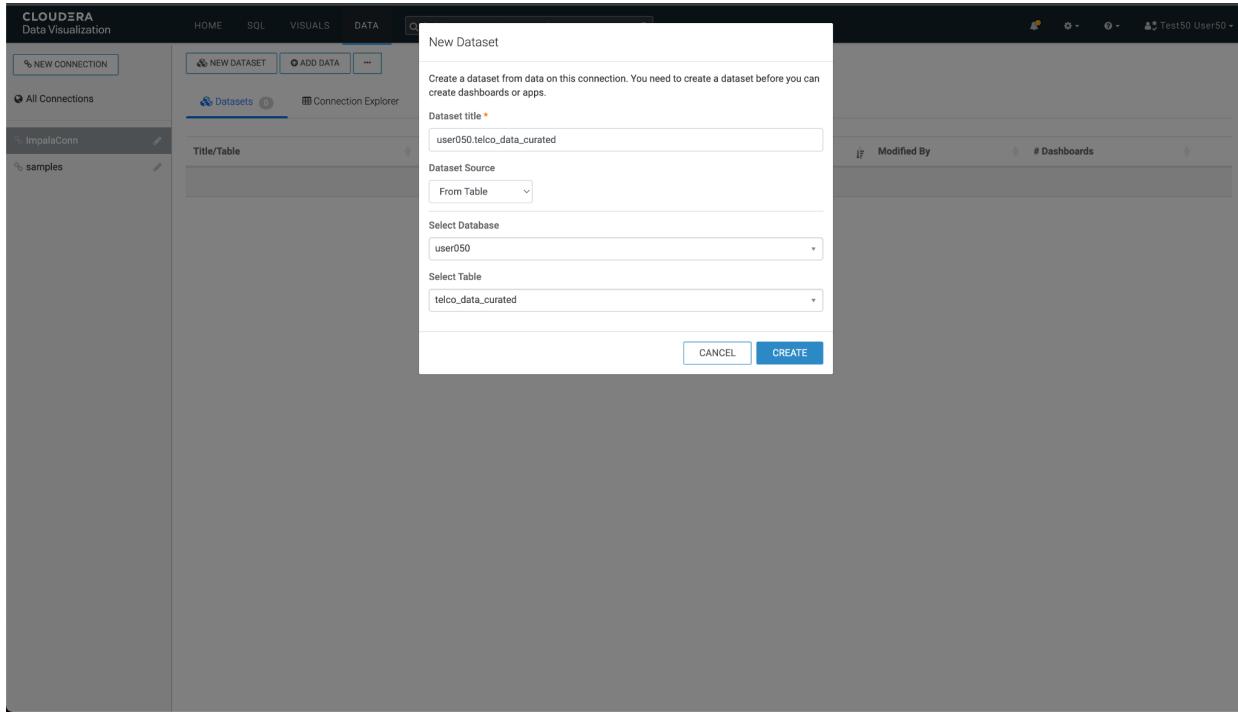
Dataset title: <assigned_user>.telco_curated_data

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 has sections for Dataset Detail, Related Dashboards, Fields, Data Model, Time Modeling, Segments (0), Filter Associations (0), and Permissions. The main panel 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. It also shows creation details: 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. Buttons at the top right include 'CLONE DATASET' and 'NEW DASHBOARD'.

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 has sections for Dataset Detail, Related Dashboards, Fields, Data Model, Time Modeling, Segments (0), Filter Associations (0), and Permissions. The main panel shows the 'Fields' section with tabs for 'Dimensions' and 'Measures'. Under Dimensions, there is a list of fields: telco_data_curated (19), including multipelines, paperlessbilling, gender, onlinesecurity, internetservice, techsupport, contract, churn, seniorcitizen, deviceprotection, streamingtv, streamingmovies, partner, customerid, dependents, onlinebackup, phoneservice, and paymentmethod. Under Measures, there is a list of fields: telco_data_curated (3), including totalcharges, monthlycharges, and tenure. A 'EDIT FIELDS' button is visible above the Dimensions list, and a 'NEW DASHBOARD' button is at the top right.

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, with a red box drawn around it. Below the button is a checked checkbox for 'Apply Display Format'. At the top right, 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.

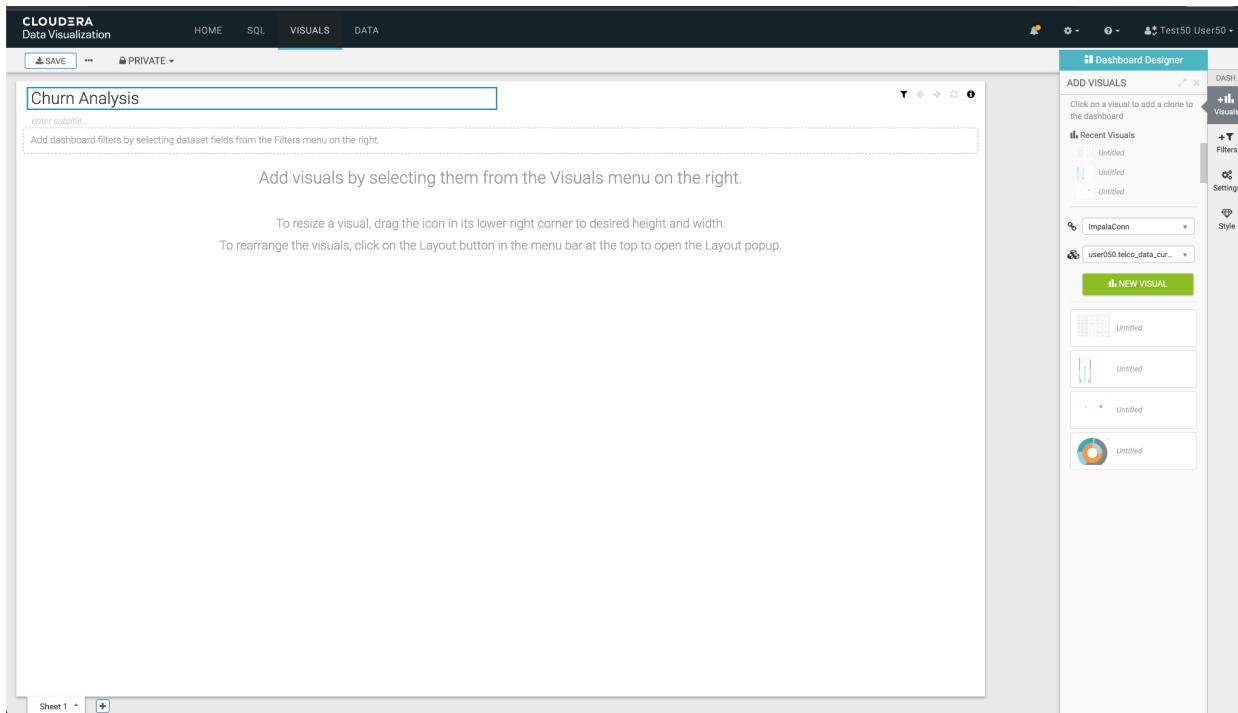
The screenshot shows the same Cloudera Data Visualization interface, but the 'SHOW DATA' button is now grayed out and labeled 'HIDE DATA'. The main area displays a table titled 'telco_data_curated' with various columns: multiplelines, paperlessbilling, gender, onlinesecurity, internetservice, techsupport, contract, churn, seniorcitizen, deviceprotection, streamingtv, streamingmovies, totalcharges, partner, monthlycharges, customerid, and de. The table contains several rows of data, such as 'No phone service' and 'Yes' for multiplelines, and 'Female' for gender.

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**

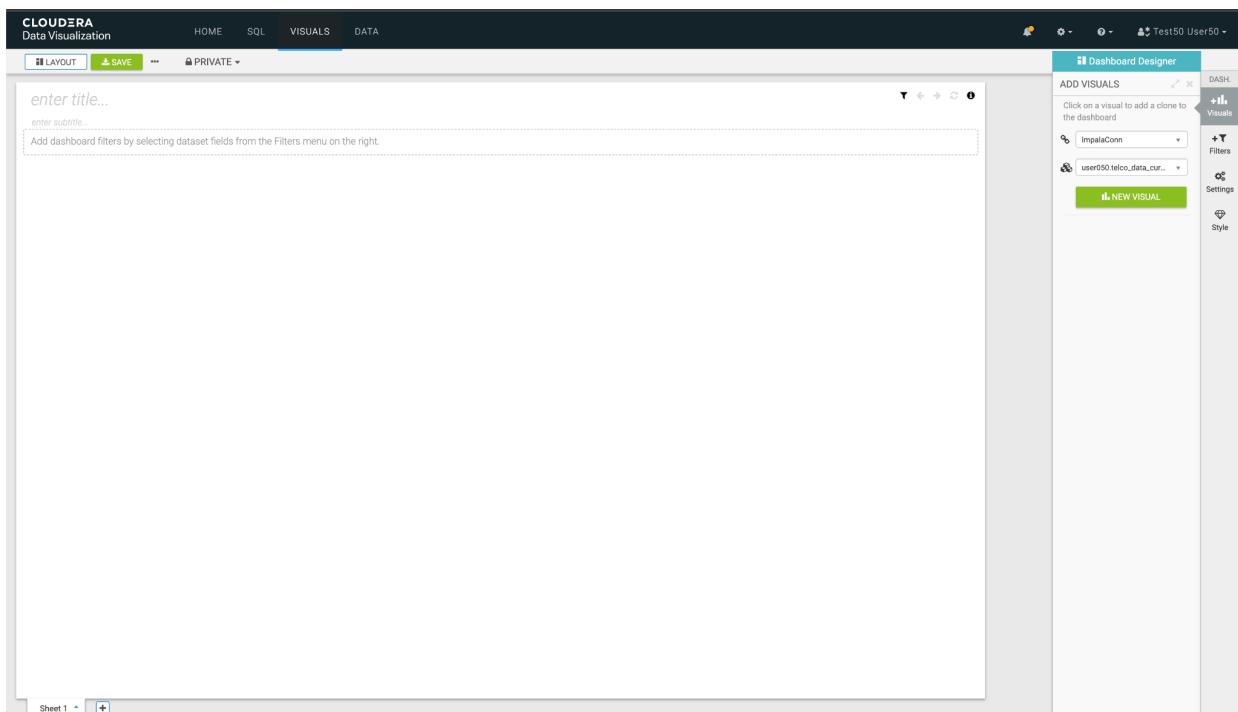
The screenshot shows the Cloudera Data Visualization interface. On the left, there's a dashboard canvas with a table visual. A context menu is open over the table, with the 'Delete Visual' option highlighted. The interface includes a top navigation bar with tabs like HOME, SQL, VISUALS, and DATA, and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

multiplelines	paperlessbilling	gender	onlinesecurity
No phone service	Yes	Female	No
No	No	Male	Yes
No	Yes	Male	Yes
No phone service	No	Male	Yes
No	Yes	Female	No
Yes	Yes	Female	No

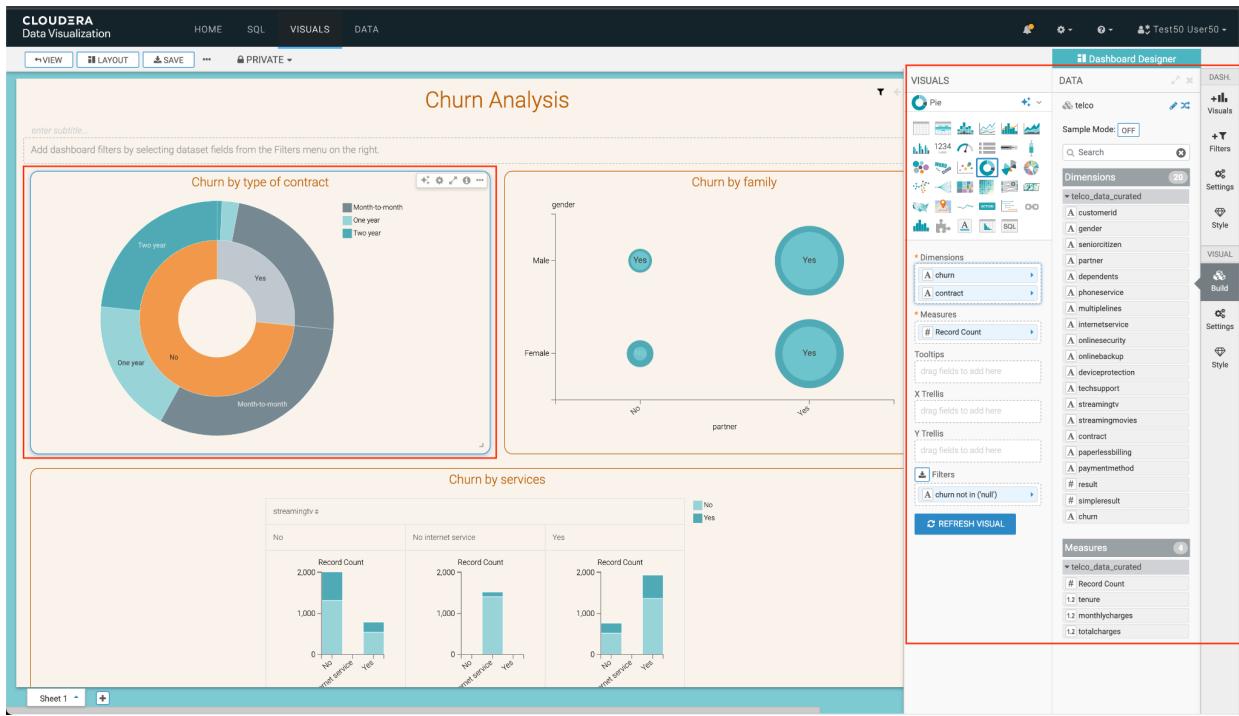
At the top of the canvas, in the enter title field, enter the name *Churn Analysis* to identify the dashboard.



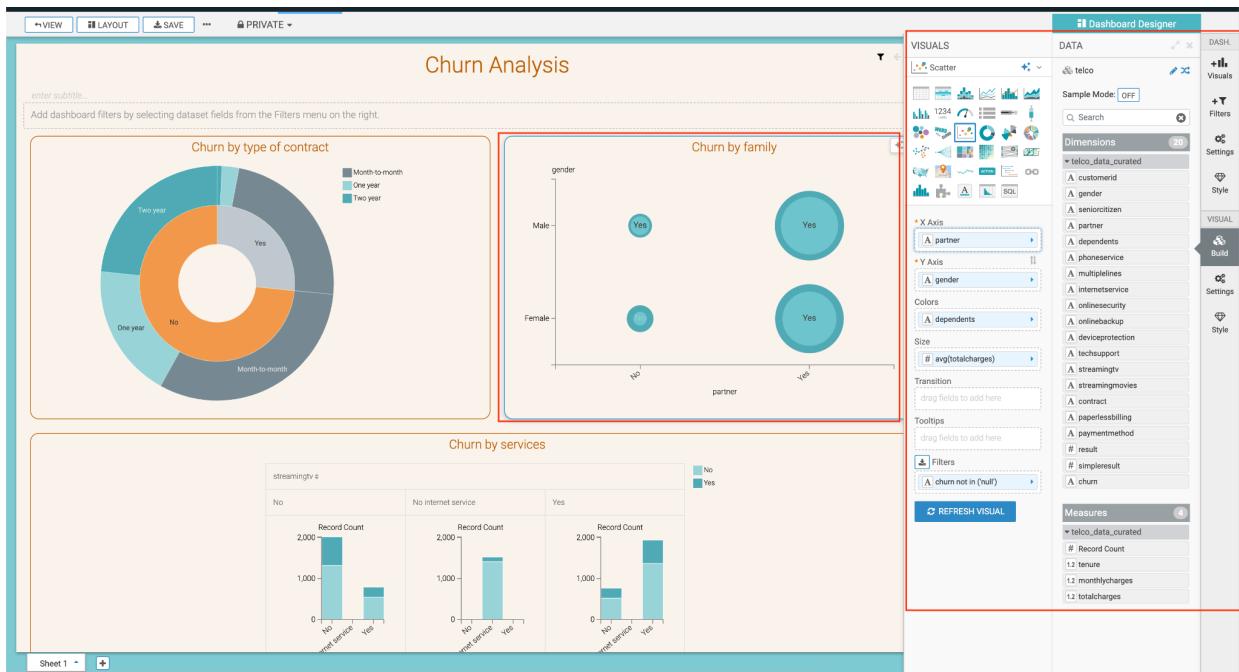
13. To add a new visual element, click on the button **Visuals** from the right menu, select the dataset that corresponds to them, and click on the button **New Visual**.



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**.

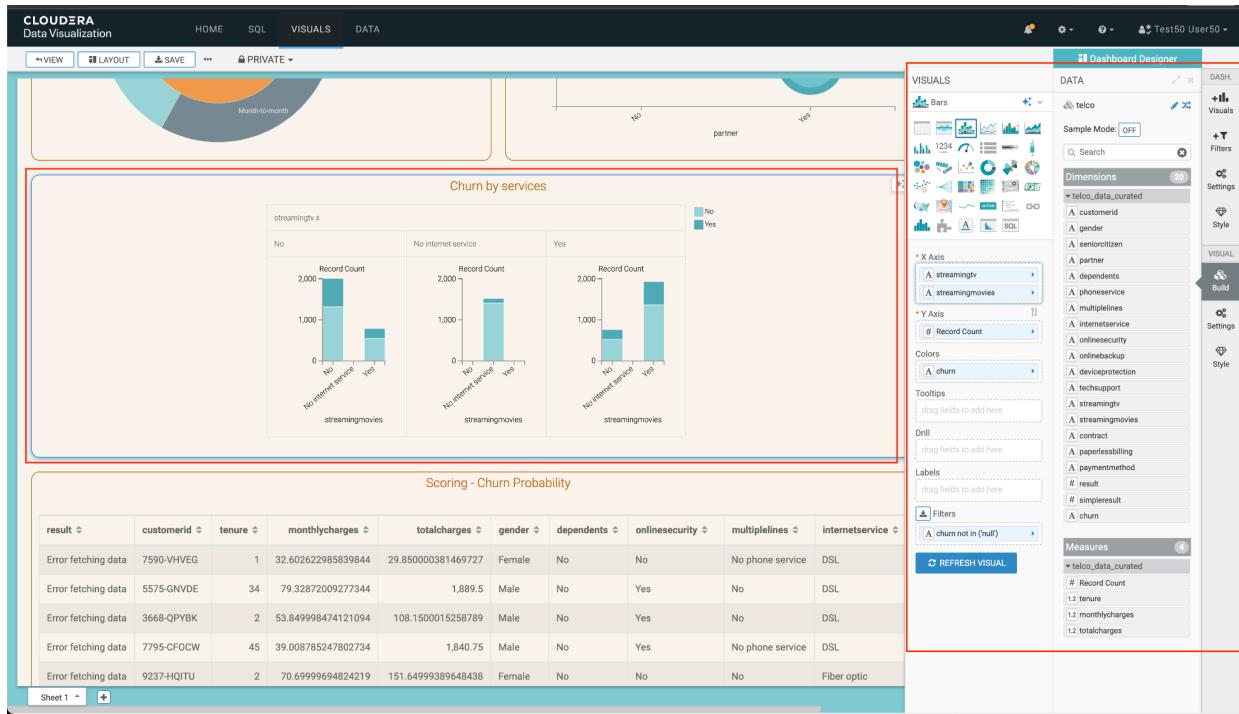


15. Add the second visual element, which is a scatter chart with the dimension **partner** like X Axis, **gender** how Y Axis, **dependents** as Colors and **avg (total charges)** as Size. Once finished, click the button **Refresh Visual**.

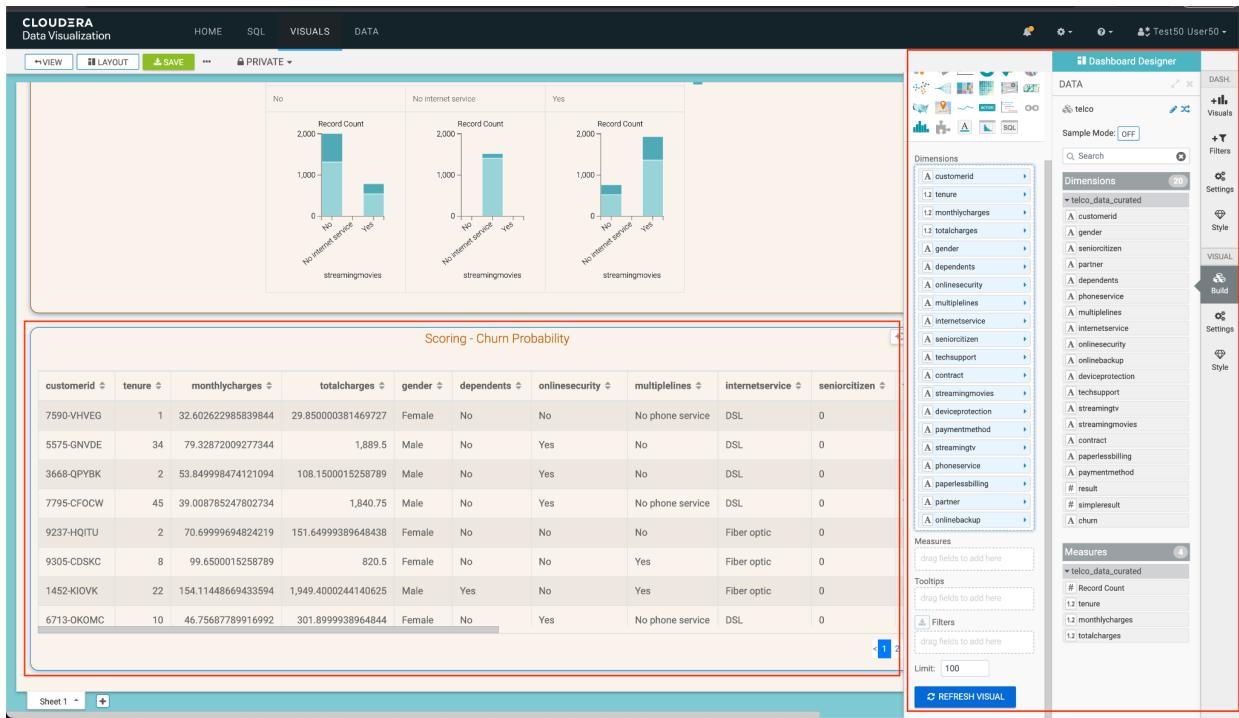


15. Add the third visual element, which is a bar chart with the dimensions **streamingtv** and **streamingmovies** like X Axis,

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**.



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. In the Dimensions section, there is a list of fields under 'telco_data_curated':

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

In the Measures section, there is a list of fields under 'telco_data_curated':

- totalcharges
- monthlycharges
- tenure

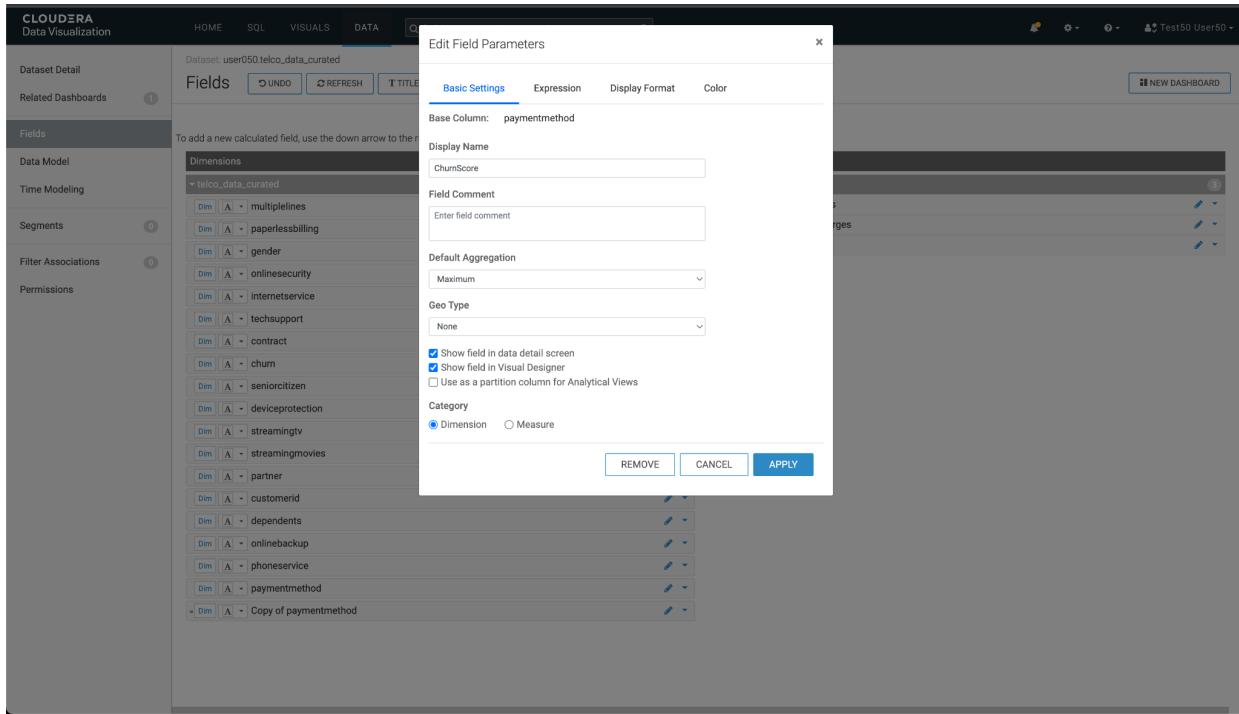
3. In the list of **Dimensions**, click the down arrow of the last field in the list, and select the option **Clone**.

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 'NEW DASHBOARD' button is in the top right. The 'Fields' section contains two panels: 'Dimensions' (left) and 'Measures' (right). The 'Dimensions' panel lists fields like 'multiplelines', 'paperlessbilling', 'gender', etc., each with a pencil icon for editing. The 'Measures' panel lists 'totalcharges', 'monthlycharges', and 'tenure'. A tooltip at the bottom right of the Dimensions panel says 'To add a new calculated field, use the down arrow to the right of a field to clone it, and then edit the expression of the cloned field.' A context menu is open over the 'paymentmethod' dimension field, with options 'Clone', '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 field. The 'Edit Field' button is highlighted on the 'Copy of paymentmethod' field in the Dimensions panel. The context menu from the previous screenshot is still visible at the bottom right of the Dimensions panel.

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"}')
```

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.

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>

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>

Name	Size	Last Modified
__pycache__	-	15 hours ago
flask	-	15 hours ago
images	-	15 hours ago
models	-	15 hours ago
raw	-	15 hours ago
0_bootstrap.py	1.95 kB	15 hours ago
0b_create_jobs.py	5.60 kB	15 hours ago

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.

The screenshot shows the Cloudera Data Visualization interface with a modal dialog titled "Edit Field Parameters". The dialog has tabs for "Basic Settings", "Expression" (which is selected), "Display Format", and "Color".

Expression Tab:

- Text input field: `1 cviz.rest('{"url":"curl_dsl_workspace","access_key": "colnames":["monthlycharges","totalcharges","tenure","internet","dependents","onlineservice","multiplelines","contract","streamingtv","streamingmovies","deviceprotection","paperlessbilling","partner","onlinebackup","response_columnname":"result"}')`
- Checkboxes:
 - Expression contains an aggregation
 - Autocomplete on VALIDATE EXPRESSION
 - Save expression only after validation succeeds
- Buttons: REMOVE, CANCEL, APPLY

Sidebar (Fields Section):

- Dataset Detail: Dataset: user050.telco_data.curated
- Related Dashboards: 1
- Fields: Fields D UNDO
- Dimensions: Dimensions
- telco_data_curated:
 - Dim A + multiplelines
 - Dim A + paperlessbilling
 - Dim A + gender
 - Dim A + onlinenetsecurity
 - Dim A + internetservice
 - Dim A + techsupport
 - Dim A + contract
 - Dim A + churn
 - Dim A + seniorcitizen
 - Dim A + deviceprotection
 - Dim A + streamingtv
 - Dim A + streamingmovies
 - Dim A + partner
 - Dim A + customerid
 - Dim A + dependents
 - Dim A + onlinebackup
 - Dim A + phoneservice
 - Dim A + paymentmethod
 - Dim # + ChurnScore

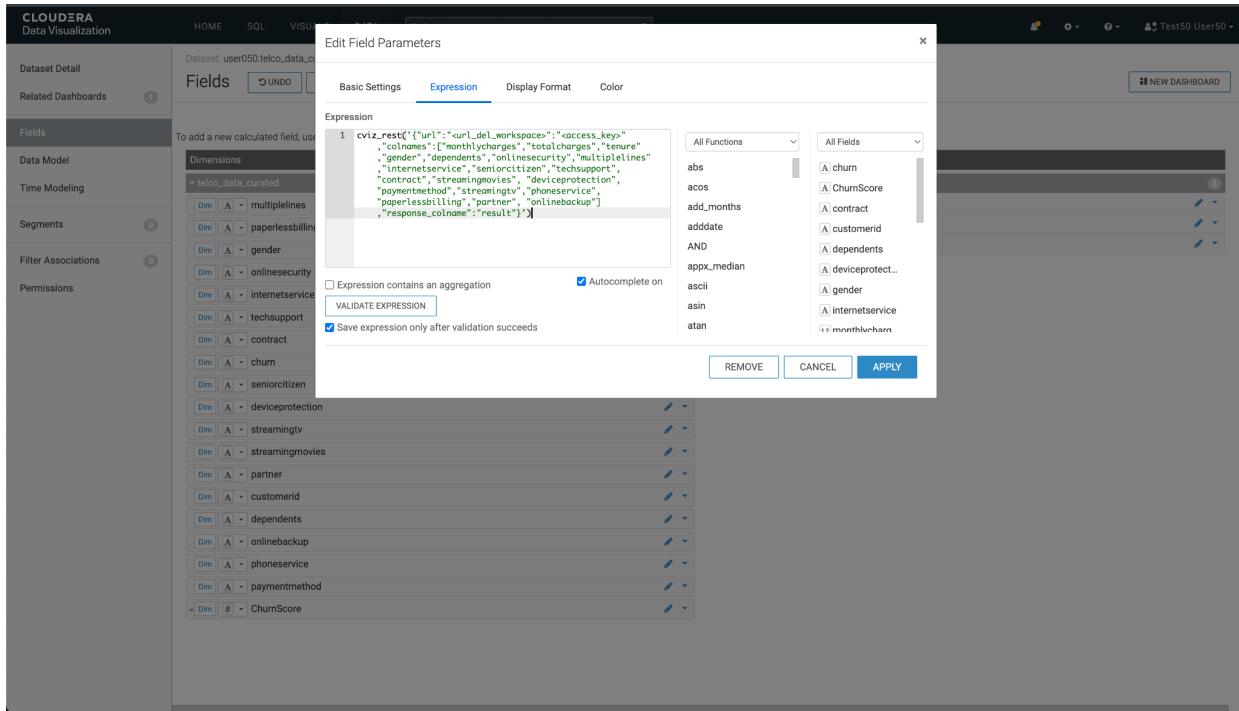
Top Right:

- NEW DASHBOARD

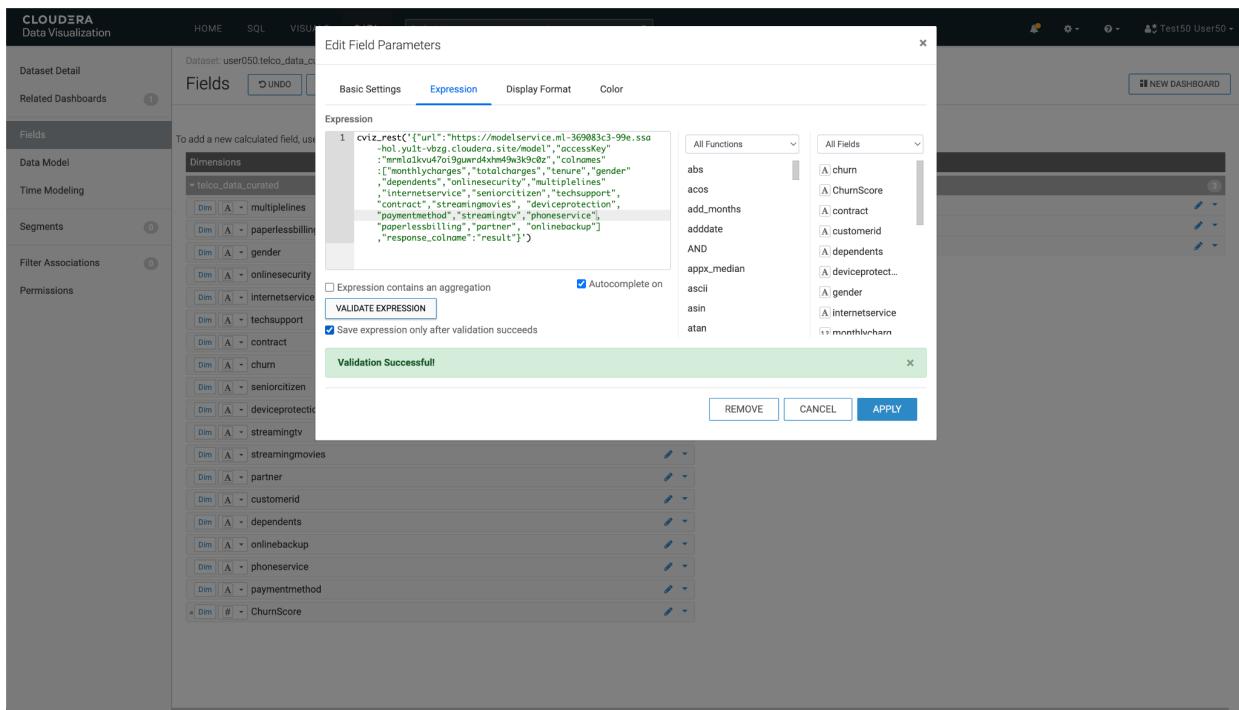
9. Returning to the CML, copy the accessKey of the model.

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.



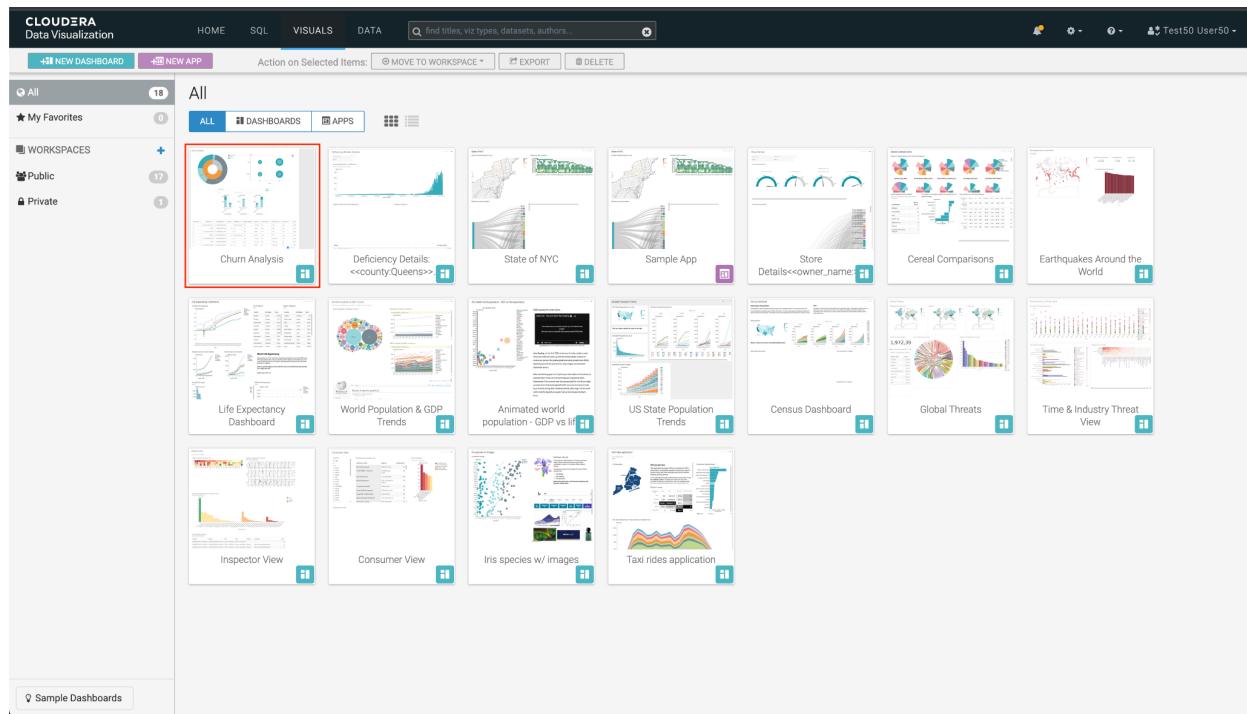
11. The new field should appear in the list of fields. Change the data type, selecting the type **Integer**, which is represented by the symbol #

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 shows the 'Fields' tab for the dataset 'user050.telco_data_curated'. It has two panels: 'Dimensions' on the left and 'Measures' on the right. In the Dimensions panel, there is a list of dimensions like 'multipleinlines', 'paperlessbilling', 'gender', etc. A context menu is open over one of these dimensions, showing options: Boolean, Integer, Real, String, Timestamp, Remove CAST, and Integer again (highlighted). The 'Measures' panel shows three measures: 'totalcharges', 'monthlycharges', and 'tenure'. At the top of the Fields tab, there are buttons for UNDO, REFRESH, TITLE CASE, SAVE (green), and Show Comments. A 'NEW DASHBOARD' button is also present.

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

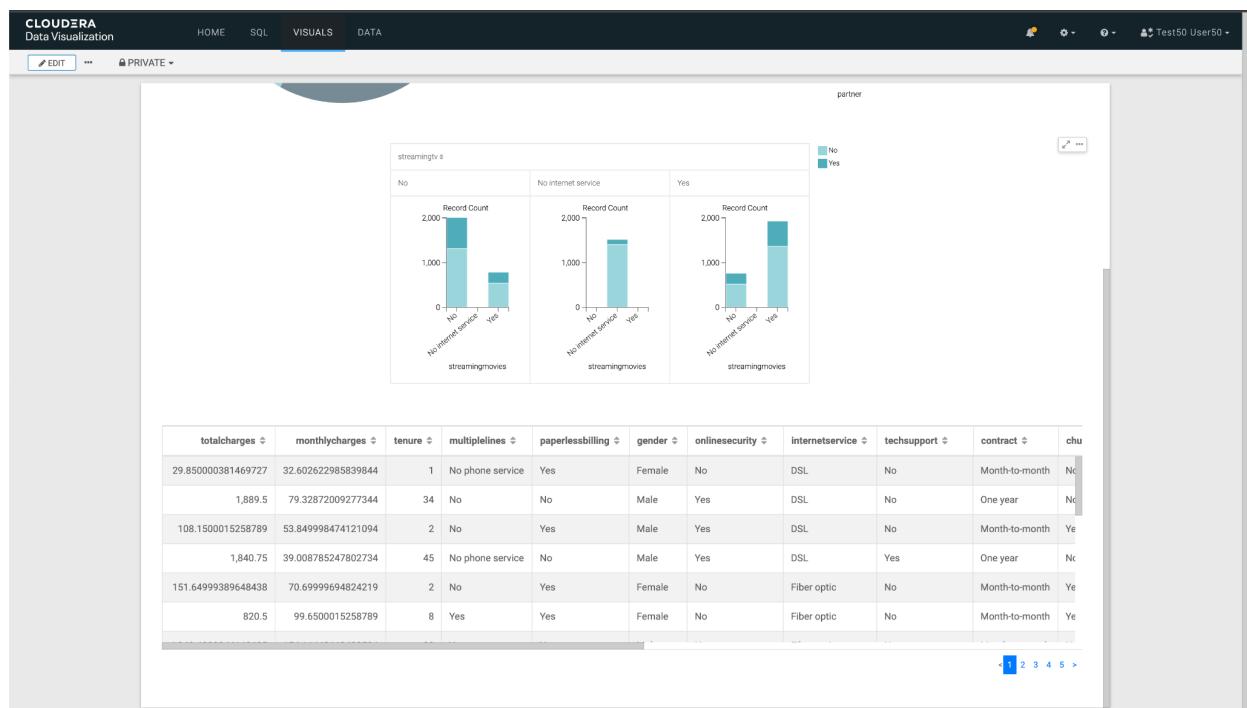
This screenshot shows the same interface after the changes have been saved. The 'Dimensions' panel now includes the newly added dimension 'ChurnScore' at the bottom of the list. The other dimensions like 'multipleinlines', 'paperlessbilling', 'gender', etc., remain in their original positions. The 'Measures' panel and the top menu remain the same as in the previous screenshot.

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 'All', 'My Favorites', 'WORKSPACES' (Public and Private), and a 'Sample Dashboards' section. The main area displays a grid of dashboard thumbnails. One dashboard, titled '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 dashboard itself contains three bar charts. The first chart, titled 'streamingtv', compares 'Record Count' for 'No' and 'Yes' categories. The second chart, also titled 'streamingtv', compares 'Record Count' for 'No internet service' and 'Yes'. The third chart, titled 'streamingmovies', compares 'Record Count' for 'No' and 'Yes'. Below the charts is a detailed 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, 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. The main area contains three stacked bar charts under the heading 'streamingtv'. The first chart is for 'No internet service', the second for 'No streamingmovies', and the third for 'Yes streamingmovies'. Each chart has two bars: 'No' (light blue) and 'Yes' (dark blue). 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	53.849998474121094	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, there is a 'Sheet 1' button and a '+' button. On the right side of the interface, there are sections for Dimensions and Measures. The 'Dimensions' section has a red box around the '# ChurnScore' item. The 'Measures' section lists '# Record Count', '# totalcharges', '# monthlycharges', and '# tenure'. There is also a 'Filters' section with a 'drag fields to add here' placeholder. A 'Refresh Visual' button is located at the bottom right of the interface.

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