

# Building Transformations and Preparing Data with Wrangler in Cloud Data Fusion | Google Cloud Skills Boost

Qwiklabs : 17-21 minutes

## Overview

Data integration is all about your data. When working with data, it's handy to be able to see what the raw data looks like so that you can use it as a starting point for your transformation. With Wrangler, you can take a data-first approach to your data integration workflow.

The most common source of data for ETL (Extract-Transform-Load) applications is typically data stored in comma-separated value (CSV) format text files, as many database systems export and import data in this fashion. For the purposes of this lab you're using a CSV file, but the same techniques can be applied to database sources and any other data source that you have available in Cloud Data Fusion.

## Objectives

In this lab you learn how to perform the following tasks:

- Create a pipeline to ingest from a CSV file.
- Use Wrangler to apply transformations by using point-and-click and the CLI interfaces.

For most of this lab, you're working with Wrangler Transformation Steps which are used by the Wrangler plugin so that your transformations are encapsulated in one place and you can group transformation tasks into manageable blocks. This data-first approach will let you quickly visualize your transformations.

## Setup

For each lab, you get a new Google Cloud project and set of resources for a fixed time at no cost.

1. Sign in to Google Cloud Skills Boost using an **incognito window**.
2. Note the lab's access time (for example, 02:00:00), and make sure you can finish within that time. There is no pause feature. You can restart if needed, but you have to start at the beginning.
3. When ready, click **Start lab**.

**Note:** Once you click **Start lab**, it will take about **15 - 20 minutes** for the lab to provision necessary resources and create a Data Fusion instance. During that time, you can read through the steps below to get familiar with the goals of the lab. When you see lab credentials (**Username** and **Password**) in the left panel, the instance is created and you can continue logging into the console.

4. Note your lab credentials (**Username** and **Password**). You will use them to sign in to the Google Cloud console.
5. Click **Open Google console**.

6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.

If you use other credentials, you'll receive errors or **incur charges**.

7. Accept the terms and skip the recovery resource page.

**Note:** Do not click **End lab** unless you have finished the lab or want to restart it. This clears your work and removes the project.

## Log in to Google Cloud Console

1. Using the browser tab or window you are using for this lab session, copy the **Username** from the **Connection Details** panel and click the **Open Google Console** button.

**Note:** If you are asked to choose an account, click **Use another account**.

2. Paste in the **Username**, and then the **Password** as prompted.

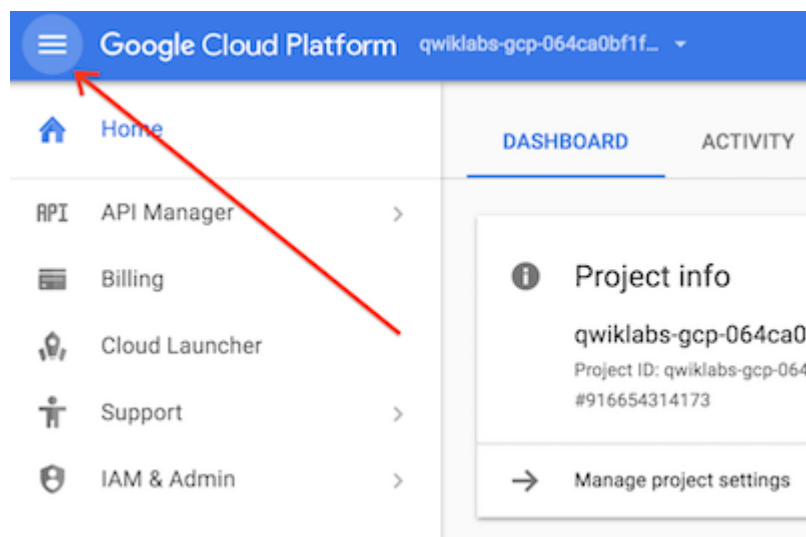
3. Click **Next**.

4. Accept the terms and conditions.

Since this is a temporary account, which will last only as long as this lab:

- Do not add recovery options
- Do not sign up for free trials

5. Once the console opens, view the list of services by clicking the **Navigation menu** (≡) at the top-left.



## Activate Cloud Shell

Cloud Shell is a virtual machine that contains development tools. It offers a persistent 5-GB home directory and runs on Google Cloud. Cloud Shell provides command-line access to your Google Cloud resources. `gcloud` is the command-line tool for Google Cloud. It comes pre-installed on Cloud Shell and supports tab completion.

1. Click the **Activate Cloud Shell** button (▶) at the top right of the console.

2. Click **Continue**.

It takes a few moments to provision and connect to the environment. When you are connected, you

are also authenticated, and the project is set to your *PROJECT\_ID*.

## Sample commands

- List the active account name:

```
gcloud auth list
```

(Output)

```
Credentialed accounts: - <myaccount>@<mydomain>.com (active)
```

(Example output)

```
Credentialed accounts: - google1623327_student@qwiklabs.net
```

- List the project ID:

```
gcloud config list project
```

(Output)


```
[core] project = <project_ID>
```

(Example output)

```
[core] project = qwiklabs-gcp-44776a13dea667a6 Note: Full documentation of gcloud is available in the gcloud CLI overview guide.
```

## Check project permissions

Before you begin working on Google Cloud, you must ensure that your project has the correct permissions within Identity and Access Management (IAM).

1. In the Google Cloud console, on the **Navigation menu** () , click **IAM & Admin > IAM**.
2. Confirm that the default compute Service Account `{project-number}-compute@developer.gserviceaccount.com` is present and has the editor role assigned. The account prefix is the project number, which you can find on **Navigation menu > Cloud overview**.

Google Cloud Platform qwiklabs-gcp-03-e30ac90a32e4 Search products and resources

**IAM & Admin**

**IAM** ADD REMOVE

**PERMISSIONS** **RECOMMENDATIONS HISTORY**

**Permissions for project "qwiklabs-gcp-03-e30ac90a32e4"**

These permissions affect this project and all of its resources. [Learn more](#)

View By: **PRINCIPALS** **ROLES**

**Filter** Enter property name or value

Type	Principal	Name	Role	Se
<input type="checkbox"/>	407543585891-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor	
<input type="checkbox"/>	407543585891@cloudbuild.gserviceaccount.com		Cloud Build Service Account	
<input type="checkbox"/>	407543585891@cloudservices.gserviceaccount.com	Google APIs Service Agent	Editor	
<input type="checkbox"/>	admiral@qwiklabs-services-prod.iam.gserviceaccount.com		Owner	
<input type="checkbox"/>	qwiklabs-gcp-03-e30ac90a32e4@qwiklabs-gcp-03-e30ac90a32e4.iam.gserviceaccount.com	Qwiklabs User Service Account	App Engine Admin BigQuery Admin	

If the account is not present in IAM or does not have the editor role, follow the steps below to assign the required role.

1. In the Google Cloud console, on the **Navigation menu**, click **Cloud overview**.
2. From the **Project info** card, copy the **Project number**.
3. On the **Navigation menu**, click **IAM & Admin > IAM**.
4. At the top of the **IAM** page, click **Add**.
5. For **New principals**, type:

{project-number}-compute@developer.gserviceaccount.com

Replace {project-number} with your project number.

6. For **Select a role**, select **Basic** (or Project) > **Editor**.
7. Click **Save**.

## Task 1. Add necessary permissions for your Cloud Data Fusion instance

1. In the Google Cloud console, from the **Navigation menu** select **Data Fusion > Instances**. You should see a Cloud Data Fusion instance already set up and ready for use.

Next, you will grant permissions to the service account associated with the instance, using the following steps.

2. Click the instance name. On the Instance details page copy, the **Service Account** to your clipboard.
3. From the Google Cloud console, navigate to the **IAM & Admin > IAM**.
4. On the IAM Permissions page, click **+Grant Access**.
5. In the New principals field paste the service account.
6. Click into the **Select a role field** and start typing "Cloud Data Fusion API Service Agent", then select it.
7. Click **Save**.

Click *Check my progress* to verify the objective. Add Cloud Data Fusion API Service Agent role to service account

## Grant service account user permission

1. In the console, on the **Navigation menu**, click **IAM & admin > IAM**.
2. Select the **Include Google-provided role grants** checkbox.
3. Scroll down the list to find the Google-managed Cloud Data Fusion service account that looks like `service-{project-number}@gcp-sa-datafusion.iam.gserviceaccount.com` and then copy the service account name to your clipboard.

service-690232861244@gcp-sa-datafusion.iam.gserviceaccount.com	Cloud Data Fusion Service Account	Cloud Data Fusion API Service Agent
--	-----------------------------------	-------------------------------------

4. Next, navigate to the **IAM & admin > Service Accounts**.
5. Click on the default compute engine account that looks like `{project-number}-compute@developer.gserviceaccount.com`, and select the **Permissions** tab on the top navigation.
6. Click on the **Grant Access** button.
7. In the **New Principals** field, paste the service account you copied earlier.
8. In the **Role** dropdown menu, select **Service Account User**.
9. Click **Save**.

## Task 2. Load the data

Next you will create a Cloud Storage bucket in your project so that you can load some sample data for Wrangling. Cloud Data Fusion will later read data out of this storage bucket

1. In Cloud Shell, execute the following commands to create a new bucket:

```
export BUCKET=$GOOGLE_CLOUD_PROJECT gcloud storage buckets create gs://$BUCKET
```

The created bucket name is your Project ID.

2. Run the command to copy the data file (a CSV file) into your bucket:

```
gcloud storage cp gs://cloud-training/OCBL163/titanic.csv gs://$BUCKET
```

Click *Check my progress* to verify the objective. Load the data

Now you're ready to proceed further.

## Task 3. Navigate the Cloud Data Fusion UI

In the Cloud Data Fusion UI you can use the various pages, such as **Pipeline Studio** or **Wrangler**, to use Cloud Data Fusion features.

To navigate the Cloud Data Fusion UI, follow these steps:

1. In the Console return to **Navigation menu > Data Fusion**.
2. Then click the **View Instance** link next to your Data Fusion instance.
3. Select your lab credentials to sign in.

If prompted to take a tour of the service click **No, Thanks**. You should now be in the Cloud Data Fusion UI.

The Cloud Data Fusion web UI comes with its own navigation panel (on the left) to navigate to the page you need.

4. In the Cloud Data UI, click the **Navigation menu** on the top left to expose the navigation panel.
5. Then choose **Wrangler**.

## Task 4. Working with Wrangler

**Wrangler** is an interactive, visual tool that lets you see the effects of transformations on a small subset of your data before dispatching large, parallel-processing jobs on the entire dataset.

1. When **Wrangler** loads, on the left side is a panel with the pre-configured connections to your data, including the Cloud Storage connection.
2. In the **GCS**, select **Cloud Storage Default**.
3. Click the bucket corresponding to your Project ID.
4. Click **titanic.csv**.
5. In the parsing options select **text** format from the drop-down.

# Parsing Options

Format \*

text

?

File encoding

UTF-8

?

IMPORT SCHEMA ?

CANCEL    CONFIRM

6. Click **Confirm**. The data is loaded into the Wrangler.

Cloud Data Fusion | Wrangler

titanic.csv

Google Cloud Storage - cloud\_storage\_default

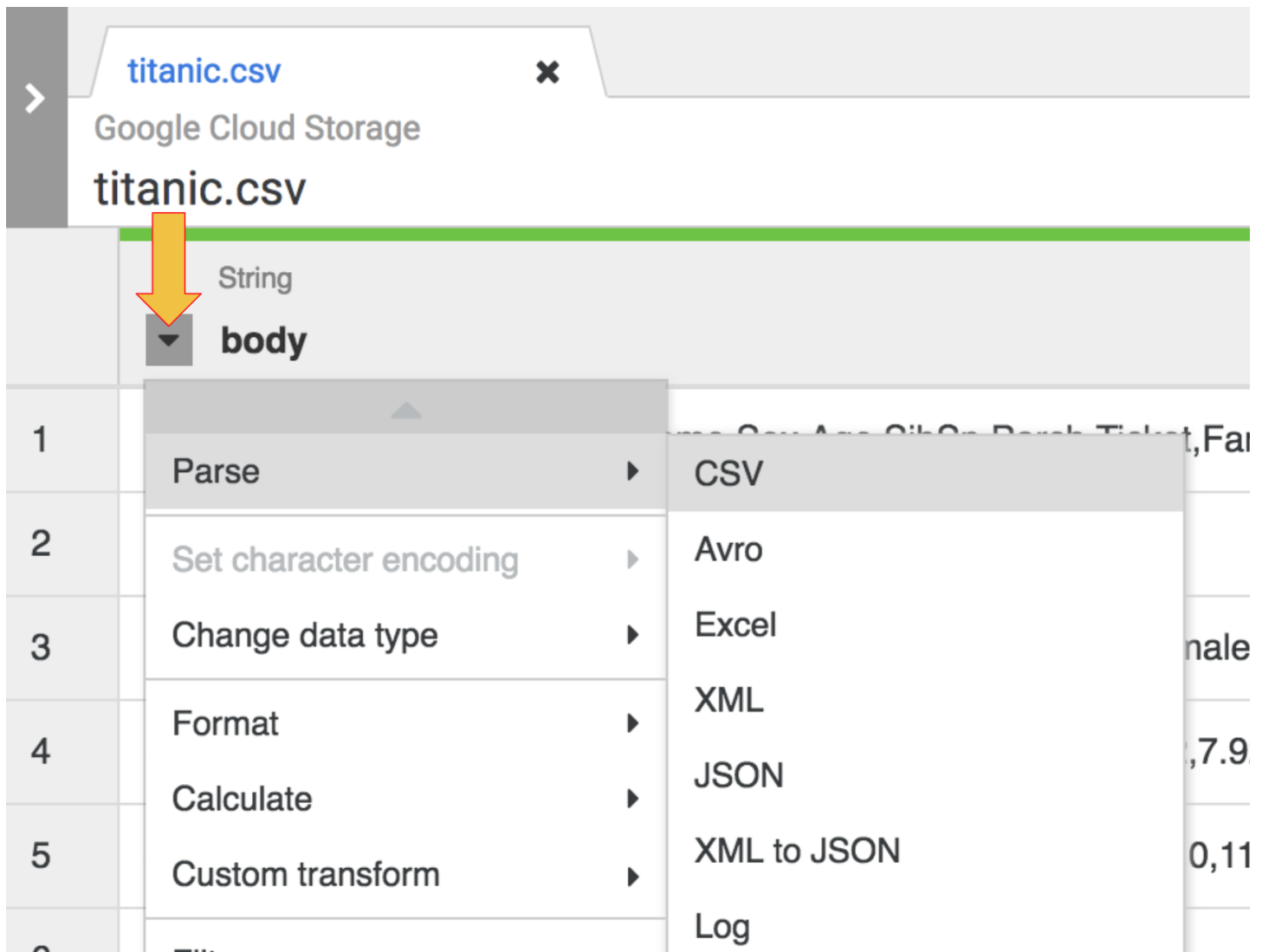
titanic.csv

Data

Insights

	String
	<div>body</div>
1	PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
2	1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
3	2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
4	3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
5	4,1,1,"Futelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
6	5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S

7. The first operation is to parse the raw CSV data into a tabular representation that is split into rows and columns. To do this you will select the drop-down icon from the first column heading, then select the **Parse** menu item, and **CSV** from the submenu.



8. In the raw data we can see that the first row consists of column headings, so you need to select the option to **Set first row as header** in the dialog box for **Parse as CSV** that is presented to you, then click **Apply**.
9. At this stage, the raw data is parsed and you can see the columns generated by this operation on the right of the *body* column
10. You no longer need the **body** column, so remove it by selecting the drop-down icon next to the **body** column heading, and select the **Delete column** menu item.

**Note:** To apply transformations, you can also use the command line interface (CLI). The CLI is the black bar at the bottom of the screen (with the green \$ prompt). As you start typing commands the autofill feature kicks in and presents you with a matching option. For example, to drop the body column, you could have alternatively used the directive: **drop: body**.



11	11,1,3,"Sandstrom, Miss. Marguerite
12	12,1,1,"Bonnell, Miss. Elizabeth",fen
13	13,0,3,"Saundercock, Mr. William He
14	14,0,3,"Andersson, Mr. Anders Joha

**drop**

Drop one or more columns.

**Usage:**

```
drop :column [,:column ]*
```

```
$ drop :body|
```

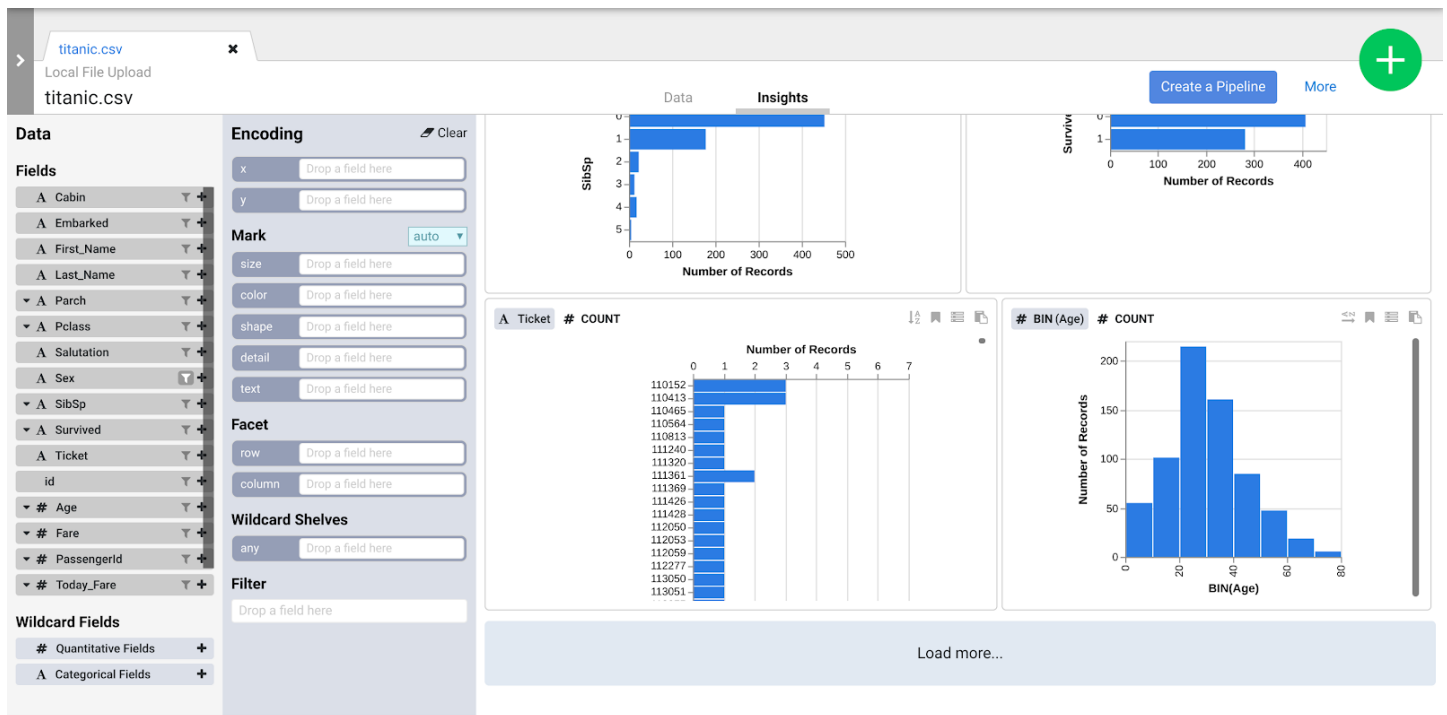
11. Click the **Transformation steps** tab on the far right of your Wrangler UI. You will see the two transformations you have applied so far.

**Note:** Both the menu selections and the CLI create directives that are visible on the **Transformation steps** tab on the right of the screen. Directives are individual transformations that are collectively referred to as a recipe.

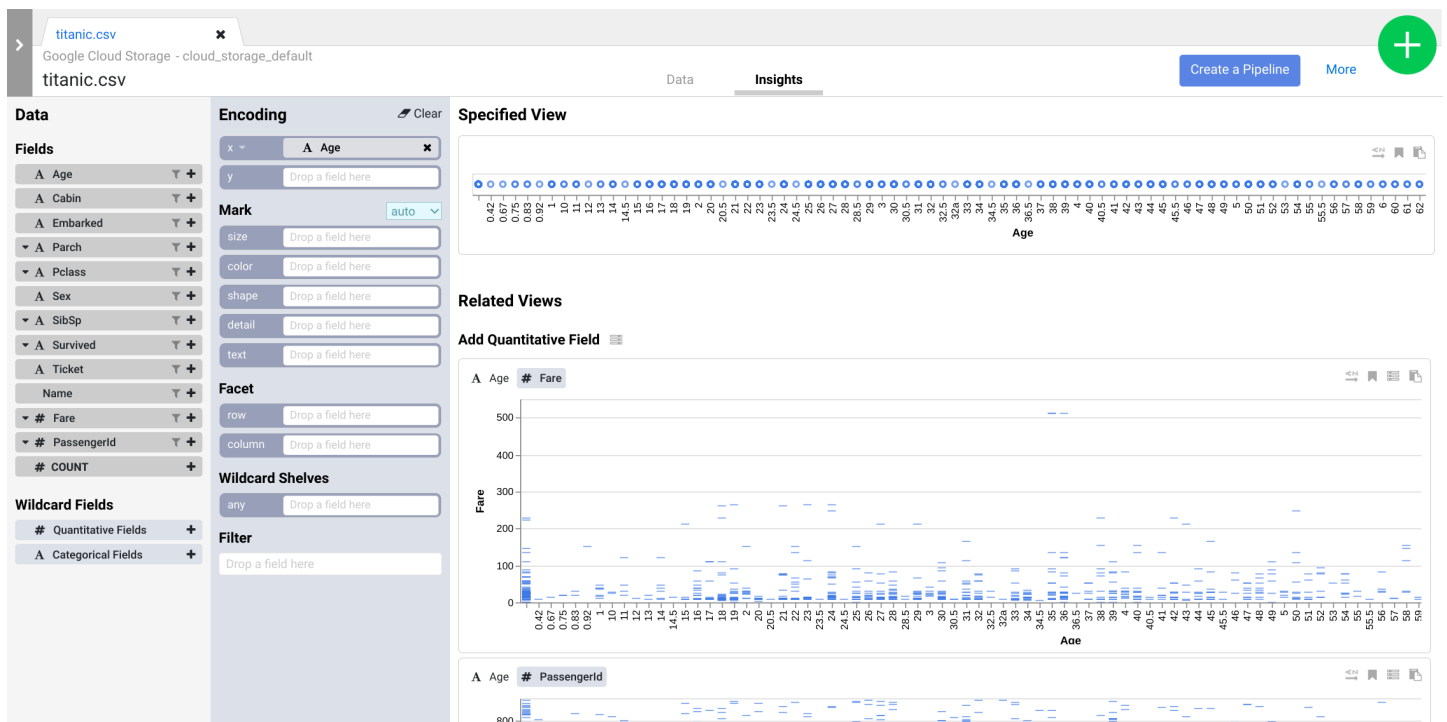
In a later part of the lab, you will add more transformation steps using the CLI.

As you apply Transformation Steps to your dataset, the transformations affect the sampled data and provide visual cues that can be explored through the *Insights* browser.

12. Click the **Insights** tab in the top middle area, to see how the data is distributed across the various columns.



13. Explore the interface to discover new ways of analyzing your data. Drag-and-drop the **Age** field to the **x** encoding to see how your data perspectives change.

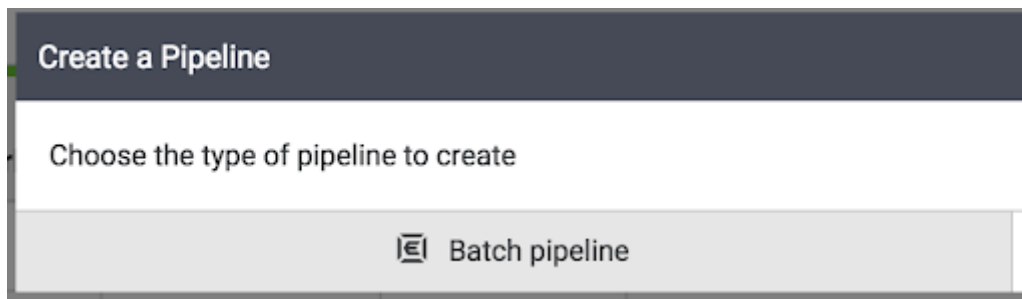


Instance id: qwiklabs-gcp-04-4dcc81d178f6/cdf

14. You can click the **Create Pipeline** button to transition to the pipeline development mode, where you can check the directives that you created within the Wrangler plugin.

Create a Pipeline

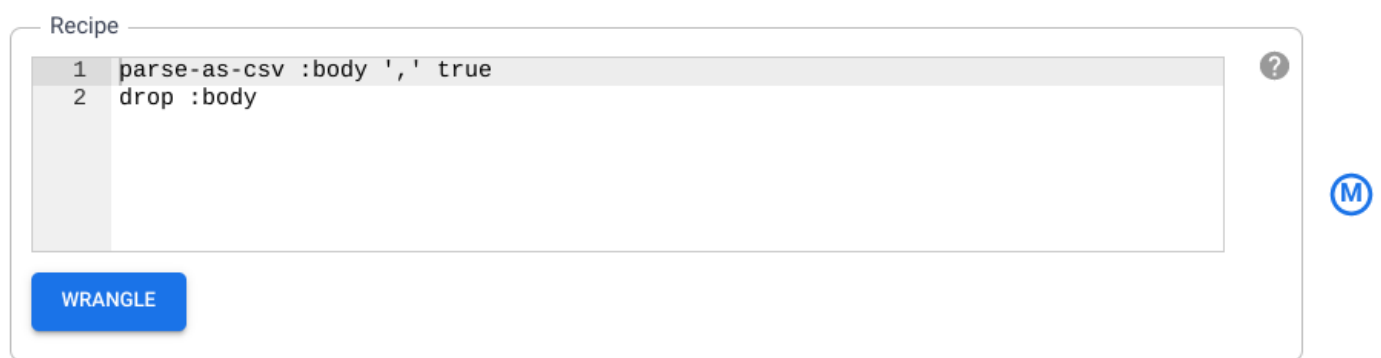
15. When presented with the next dialog select **Batch pipeline** to continue.



16. Once the *Pipeline Studio* opens, point to the **Wrangler** node and click **Properties**

17. Following **Directives** review the recipe of directives that were added by you earlier. In the next section, you will add more transformation steps using the CLI.

## Directives



## Task 5. Working with Transformation Steps

In this section, you will continue working in the Wrangler UI to explore the CSV dataset and apply transformations through CLI.

1. Click the **Wrangle** button under the **Directives** section of your Wrangler node's *Properties* box. You will be back in the Wrangler UI.
2. Click the **Transformation steps** on the far right of your Wrangler UI to expose the directives. Verify that you currently have two transformation steps.

You will now add more transformation steps using the CLI and see how they modify the data. The CLI is the black bar at the bottom of the screen (with the green \$ prompt).

3. Copy the directives, and paste them into your CLI at the \$ prompt. You will see the **Transformation Steps** on the right of your screen get updated.

```
fill-null-or-empty :Cabin 'none' send-to-error empty(Age) parse-as-csv :Name ',' false drop Name fill-null-or-empty :Name_2 'none' rename Name_1 Last_Name rename Name_2 First_Name set-type :PassengerId integer parse-as-csv :First_Name '.' false drop First_Name drop First_Name_3 rename First_Name_1 Salutation fill-null-or-empty :First_Name_2 'none' rename First_Name_2 First_Name send-to-error !dq:isNumber(Age) || !dq:isInteger(Age) || (Age == 0 || Age > 125) set-type :Age integer set-type :Fare double set-column Today_Fare (Fare * 23.4058)+1 generate-uuid id mask-shuffle First_Name
```

Following is an explanation of what the directives do to your data. **DO NOT** enter them again in the CLI as you have just already done so.


- a. `fill-null-or-empty :Cabin 'none'` fixes the **Cabin** column so that it's 100% complete.
- b. `send-to-error empty(Age)` fixes the **Age** column so there are no empty cells
- c. `parse-as-csv :Name ',' false` splits the **Name** columns into two separate columns containing the first name and last name
- d. `rename Name_1 Last_Name` and `rename Name_2 First_Name` rename the newly created columns, **Name\_1** and **Name\_2**, into **Last\_Name** and **First\_Name**
- e. `drop Name` removes the **Name** column as it's no longer needed
- f. `set-type :PassengerId integer` converts the **PassengerId** column into an integer
- g. The directives extract the salutation from the **First\_Name** column, delete the redundant column and rename the newly created columns accordingly:  
  
`parse-as-csv :First_Name '.' false drop First_Name drop First_Name_3 rename First_Name_1 Salutation`  
`fill-null-or-empty :First_Name_2 'none' rename First_Name_2 First_Name`
- h. the `send-to-error !dq:isNumber(Age) || !dq:isInteger(Age) || (Age == 0 || Age > 125)` directive performs data quality checks on the **Age** column while the `set-type :Age integer` sets it as an Integer column
- i. `set-type :Fare double` converts the **Fare** column to a Double so you can perform some arithmetic with the column values
- j. `set-column Today_Fare (Fare * 23.4058)+1` multiplies the **Fare** column by the inflation rate of the Dollar since 1912 to get the adjusted Dollar value
- k. `generate-uuid id` creates an identity column to uniquely identify each record
- l. `mask-shuffle First_Name` will mask the **Last\_Name** column to de-identify the person, i.e. PII

4. Click **More** link on the top right above your **Transformation steps**, and then click on **View Schema** to examine the schema that the transformations generated, and click the **download** icon to download it to your computer.

Schema			
Name	Type	Null	
PassengerId	int	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Survived	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Pclass	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Sex	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Age	int	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
SibSp	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Parch	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Ticket	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Fare	double	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Cabin	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Embarked	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Last_Name	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
Salutation	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>
First_Name	string	<input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>

5. Click **X** to close the Schema page.

6. You can click the **download** icon under **Transformation steps** to download the directives recipe to your computer to keep a copy of the transformation steps for future use.

Columns (16)		Transformation steps (20)	
#	Transformations		
1	parse-as-csv :body ", true		
			<input checked="" type="checkbox"/>

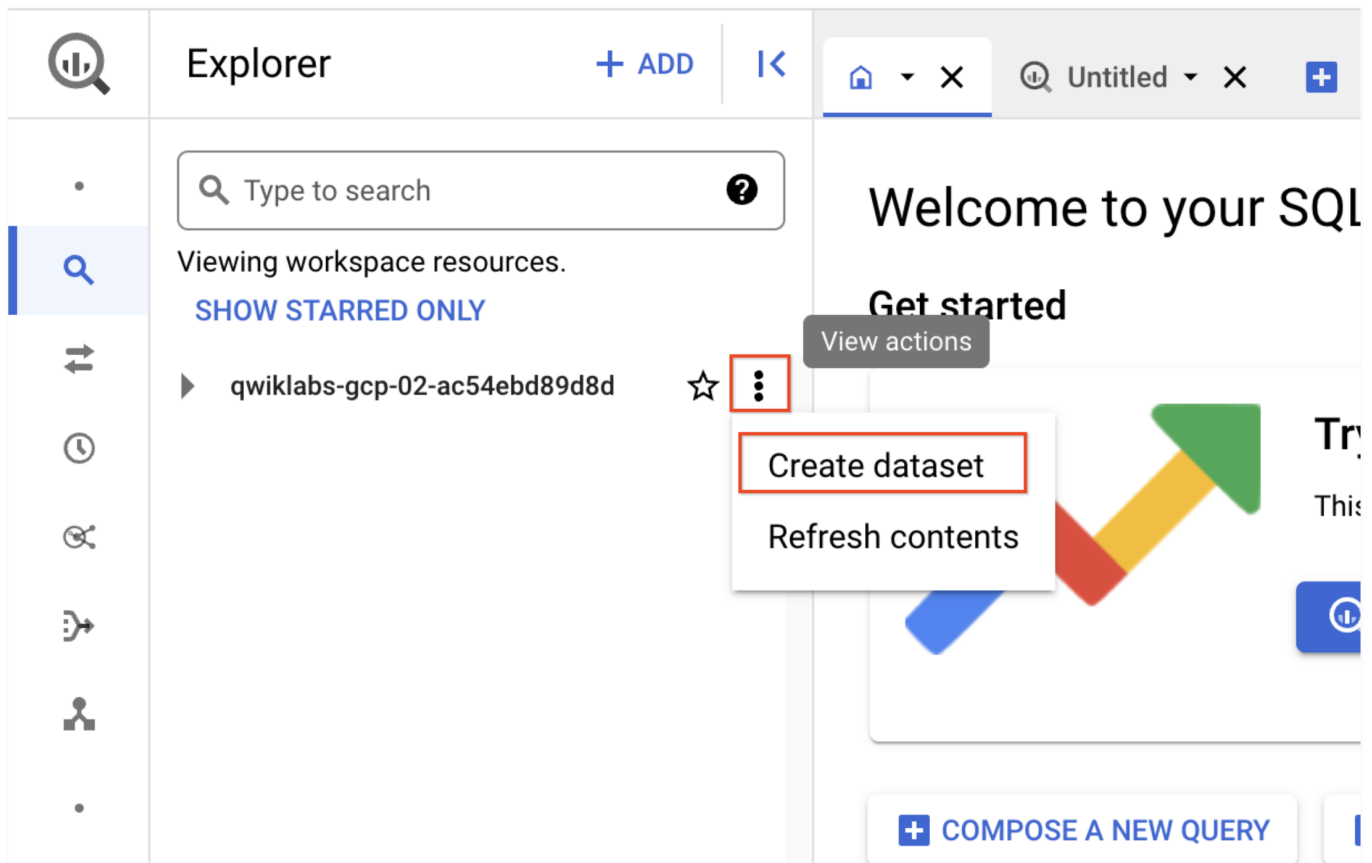
7. Click the **Apply** button on the top right to ensure all the newly entered transformation steps get added to the Wrangler node's configuration. You will then be redirected back to the properties box of the **Wrangler** node.

8. Click **X** to close it. You're back in the **Pipeline Studio**.

## Task 6. Ingestion into BigQuery

In order to ingest the data into BigQuery, create a dataset.

1. In a new tab, [open the BigQuery in the Google Cloud Console](#) or right-click on the Google Cloud console tab and select **Duplicate**, then use the **Navigation menu** to select **BigQuery**. If prompted click **Done**.
2. In the Explorer pane, click the **View actions** icon next to your Project ID (it will start with qwiklabs) and then select **Create dataset**.



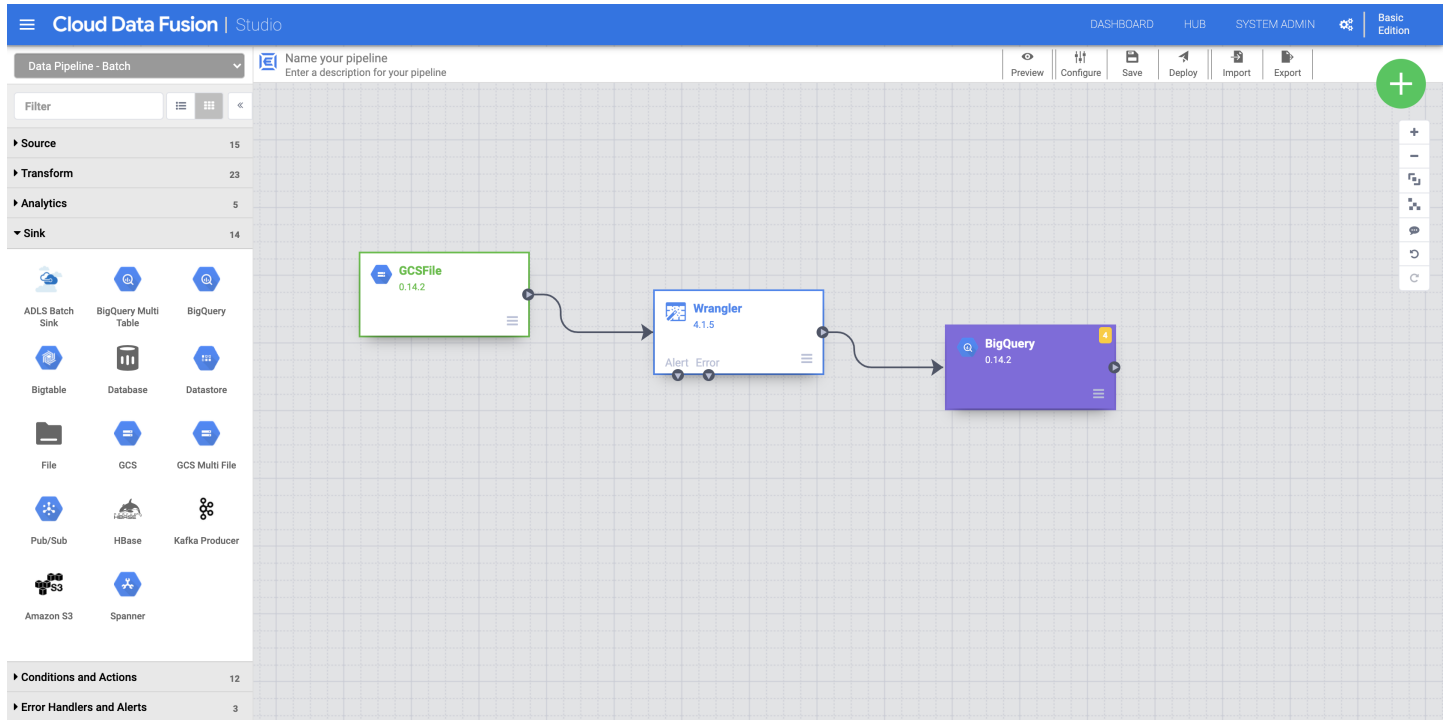
a. Dataset ID: demo\_cdf

b. Click **Create dataset**. Record the name to use later in the lab.

3. Navigate back to the Cloud Data Fusion UI tab

a. To add the BigQuery sink to the pipeline navigate to the **Sink** section on the left panel and click the **BigQuery** icon to place it on the canvas.

b. Once the BigQuery sink has been placed on the canvas, connect the Wrangler node with the BigQuery node. Do this by dragging the arrow from the Wrangler node to connect to the BigQuery node as illustrated.



c. Point your mouse over your BigQuery node, click **Properties** and enter the following configuration settings:

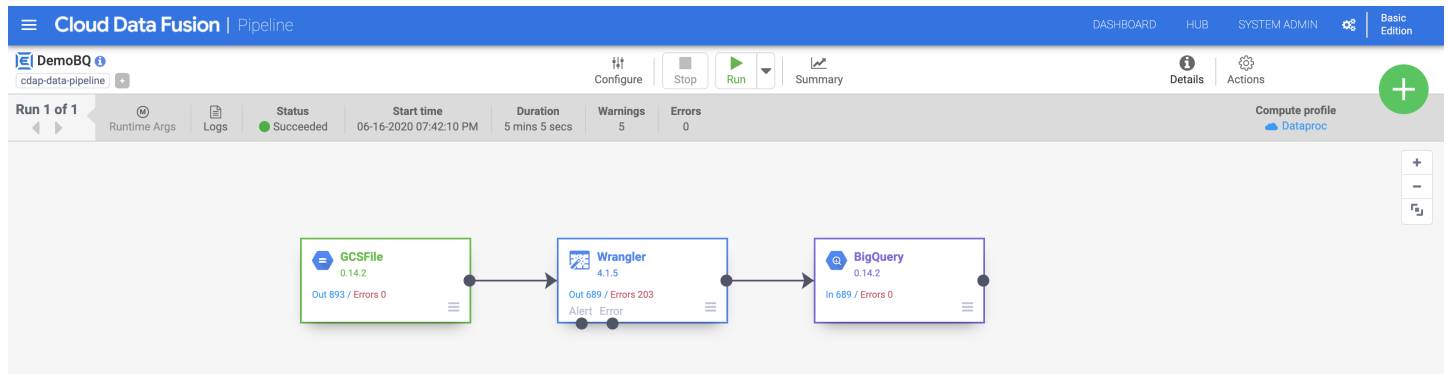
Field	Value
Reference Name	DemoSink
Dataset Project ID	Your Project ID
Dataset	demo_cdf(the dataset you created in the previous step)
Table	Enter an appropriate name (like titanic)

The table will be automatically created.

- d. Click the **Validate** button to check if everything is set up correctly.
- e. Click **X** to close it. And you're back in the Pipeline Studio.

4. Now you're ready to execute your pipeline.
- a. Give your pipeline a name (like DemoBQ)
- b. Click **Save** and then click **Deploy** on the top right to deploy the pipeline.
- c. Click **Run** to start the pipeline execution. You may click the **Summary** icon to look at some statistics.

After execution completes, the status changes to **Succeeded**. Navigate back to your BigQuery Console to query your results.



Click *Check my progress* to verify the objective. Ingestion into BigQuery

## Congratulations!

In this lab, you explored the Wrangler UI. You learned how to add transformation steps (directives) through the menu as well as using the CLI. Wrangler allows you to apply many powerful transformations to your data iteratively and you can use the Wrangler UI to view how it affects the schema of your data before you deploy and run your pipeline.

## End your lab

When you have completed your lab, click **End Lab**. Qwiklabs removes the resources you've used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click **Submit**.

The number of stars indicates the following:

- 1 star = Very dissatisfied
- 2 stars = Dissatisfied
- 3 stars = Neutral
- 4 stars = Satisfied
- 5 stars = Very satisfied

You can close the dialog box if you don't want to provide feedback.

For feedback, suggestions, or corrections, please use the **Support** tab.

Copyright 2022 Google LLC All rights reserved. Google and the Google logo are trademarks of Google LLC. All other company and product names may be trademarks of the respective companies with which they are associated.