DATA TRANSFORMATION IN CLOUD DATA FUSION

NOTE : THE GOOGLE CLOUD OBJECT NAMES INDICATED IN SCREEN SHOTS MAY NOT REFLECT ACTUAL NAMES USED IN THE PROJECT AND IS FOR GUIDANCE PURPOSES ONLY . PLEASE TAKE ADEQUATE PRECAUTIONS TO ENSURE YOU FOLLOW CORRECT OBJECT NAMES FOR CREATING THE PIPELINE.

The data ingestion part of the pipe line provides three csv files in the GCS bucket at 1400 hours every day . This data needs to be transformed/processed and loaded into Big Query for processing the dashboard. It is proposed to have two tiles of visualization in the dash board , one for current data and for aggregated combined data. The historic data in the GCS bucket is for reference purpose only and it not intended to be processed further.

INTERNAL ETL PROCESS

The data in the GCS bucket will be processed after due extraction by using **Google Data Fusion**

The source for the pipeline is **GCS bucket** and sink would be **Google Big Query**

DATA TRANSFORMATION - CURRENT DATA

The scrutiny of data frame for current data reveals following and irrelevant data can be removed ,

(a) The current data has 53 rows and 46 columns after removal of columns named unused

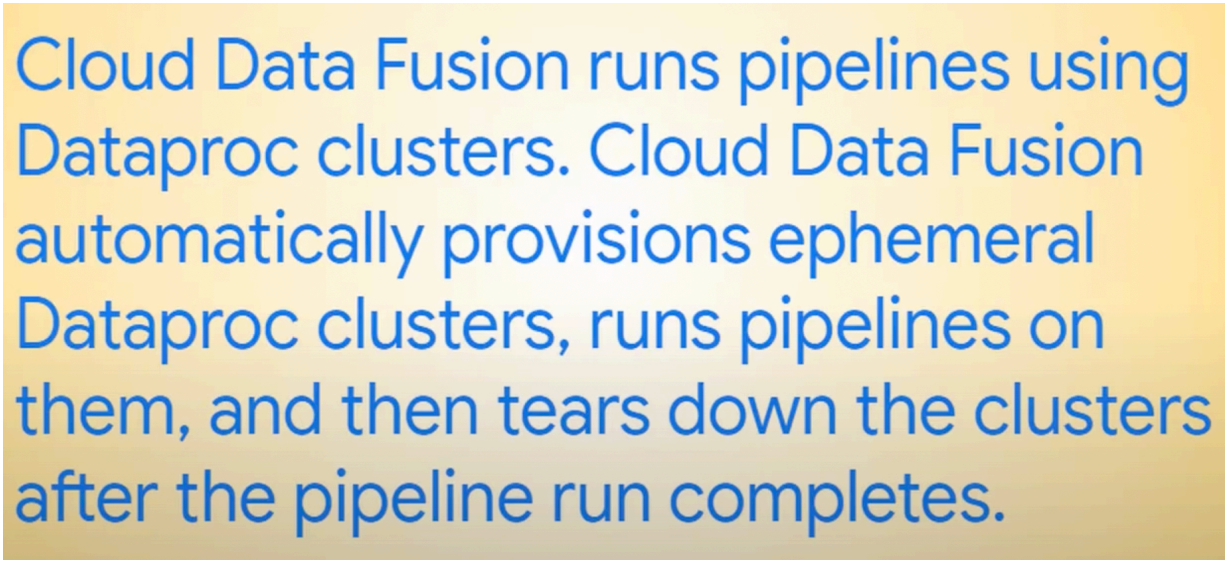
(b) Remove following columns (no relevant info or single value columns or NaN values) ,

```
'country','county', 'level', 'lat',
'locationId',

'long','metrics.testPositivityRatioDetails.source','met
rics.contactTracerCapacityRatio',
'url']
```

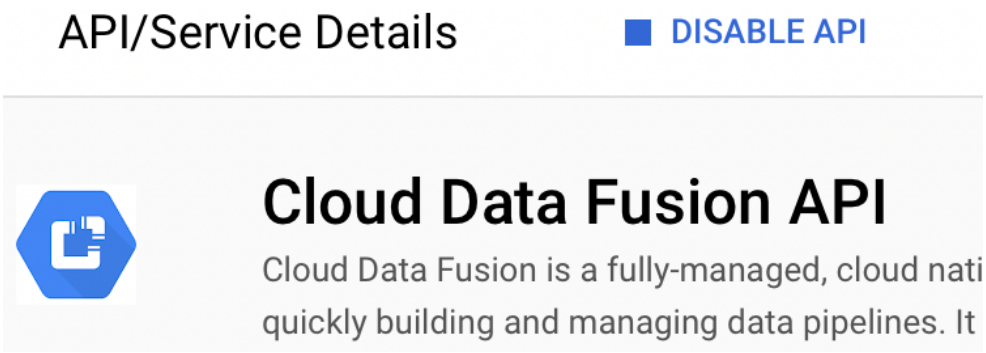
(C) convert lastUpdatedDate to datetime

Cloud Data Fusion is a fully managed, cloud-native, enterprise data integration service for quickly building and managing data pipelines. The Cloud Data Fusion web UI allows you to build scalable data integration solutions to clean, prepare, blend, transfer, and transform data, without having to manage the infrastructure.

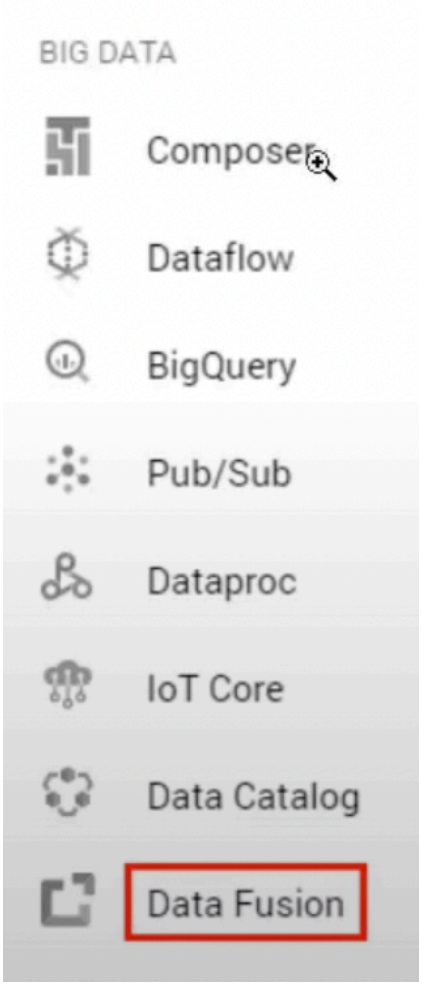


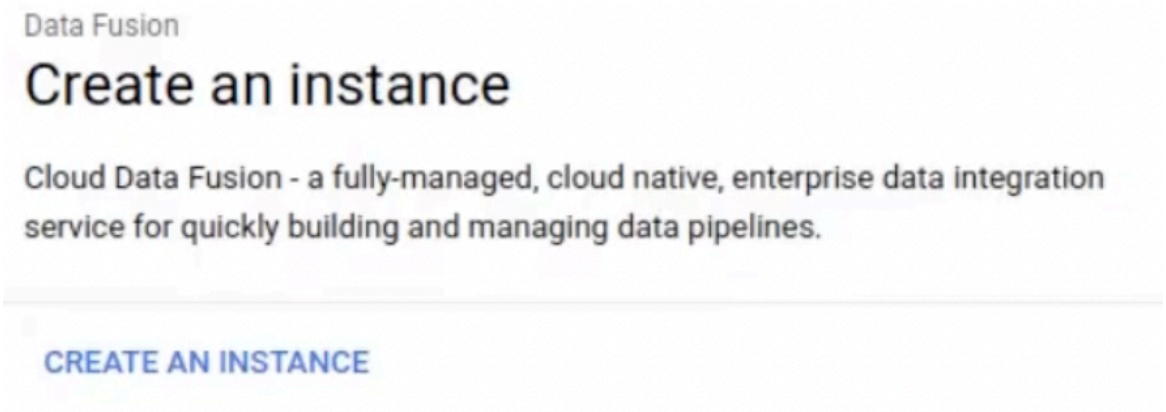
The running of pipeline can be automated .

At the start , Enable API **API Services >> Library**

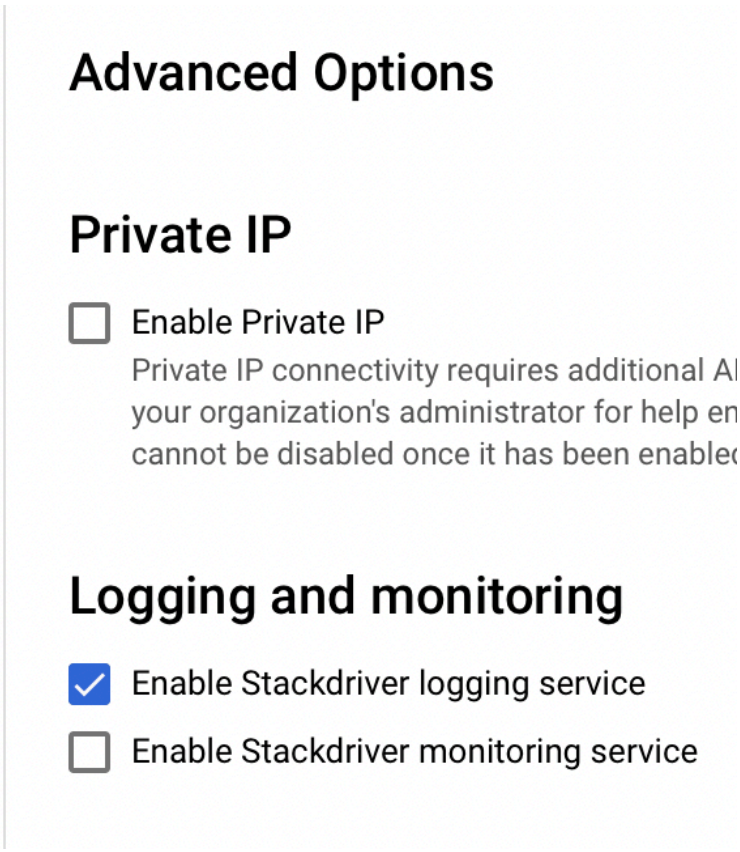


Create Cloud Data Fusion Instance





Edition - Basic



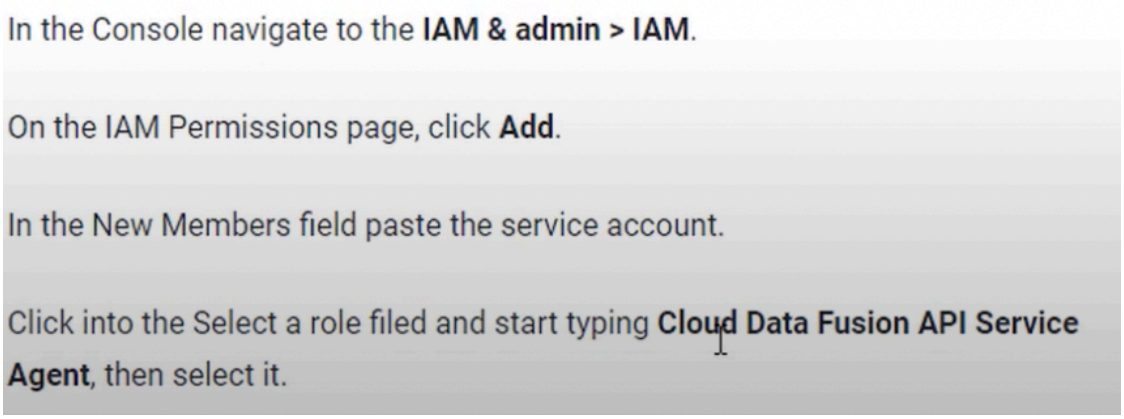
Creation of instance takes little time (may be around 10 minutes) . Also add following role to your service account

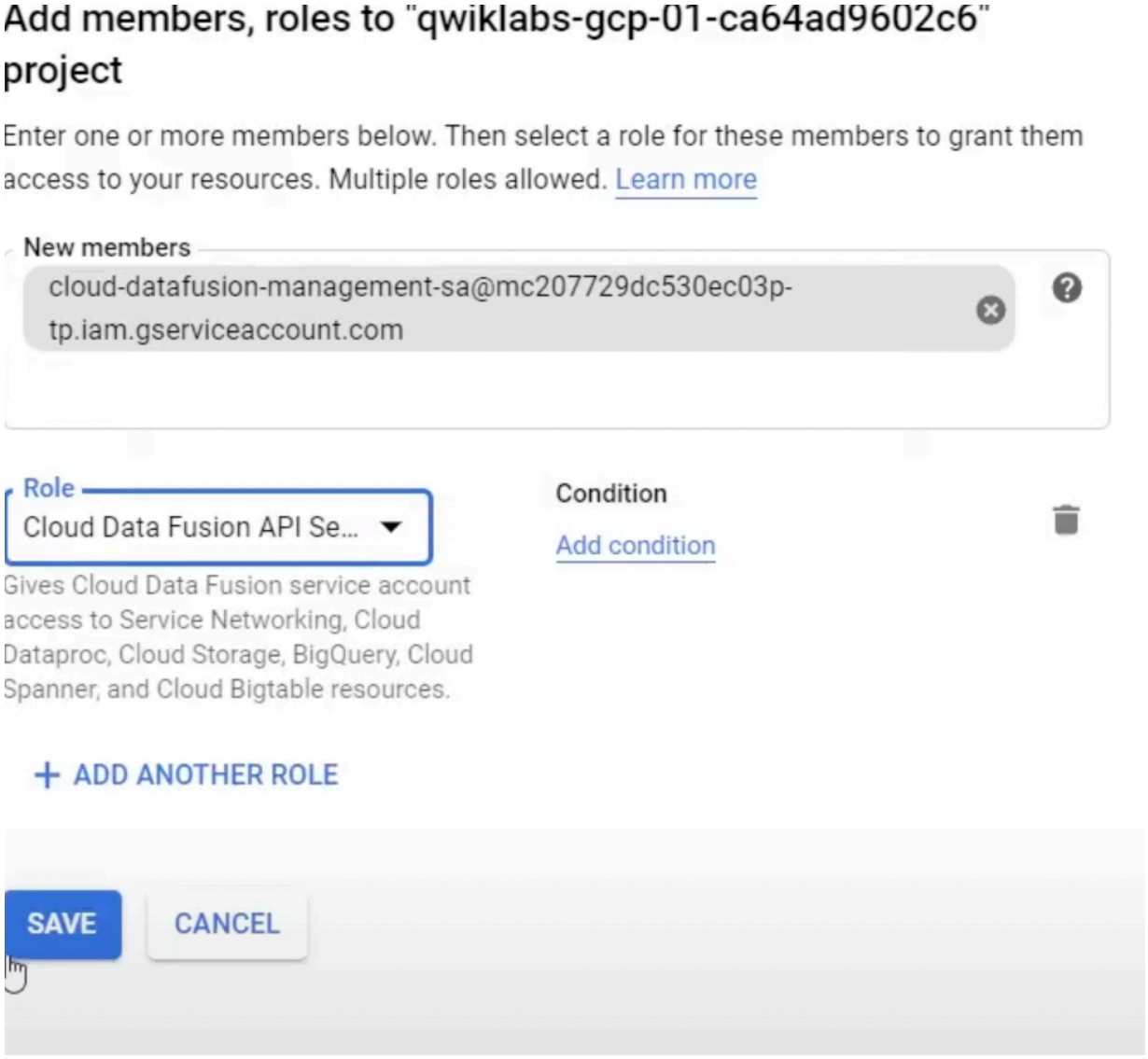
daily-covid-data-pipeline

Instance ID	daily-covid-data-pipeline
Instance URL	View Instance
Description	pipeline to process daily data
Edition	BASIC
Accelerators	ADD ACCELERATORS
Region	us-east1
Zone	--
Created	Mar 21, 2022, 10:43:00 AM
Last updated	Mar 21, 2022, 10:55:36 AM
Stackdriver logs	Enabled
Stackdriver monitoring	Disabled
Granular role based access control	ENABLE
Service Account	cloud-datafusion-management-sa@vdb24dc28d1d49f13-tp.iam.gserviceaccount.com
Version	6.5.1 (latest version)
Private IP	Disabled
Dataproc Service Account	463796539833-compute@developer.gserviceaccount.com
Encryption	Google-managed

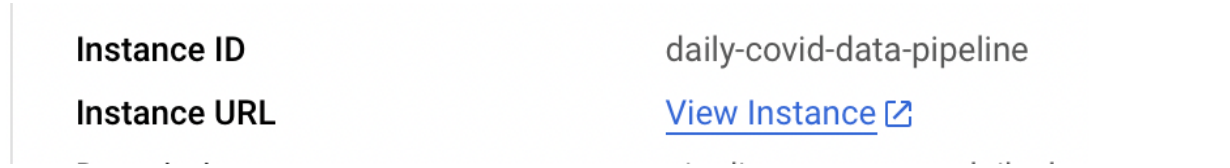
Click on the instance and copy service account .

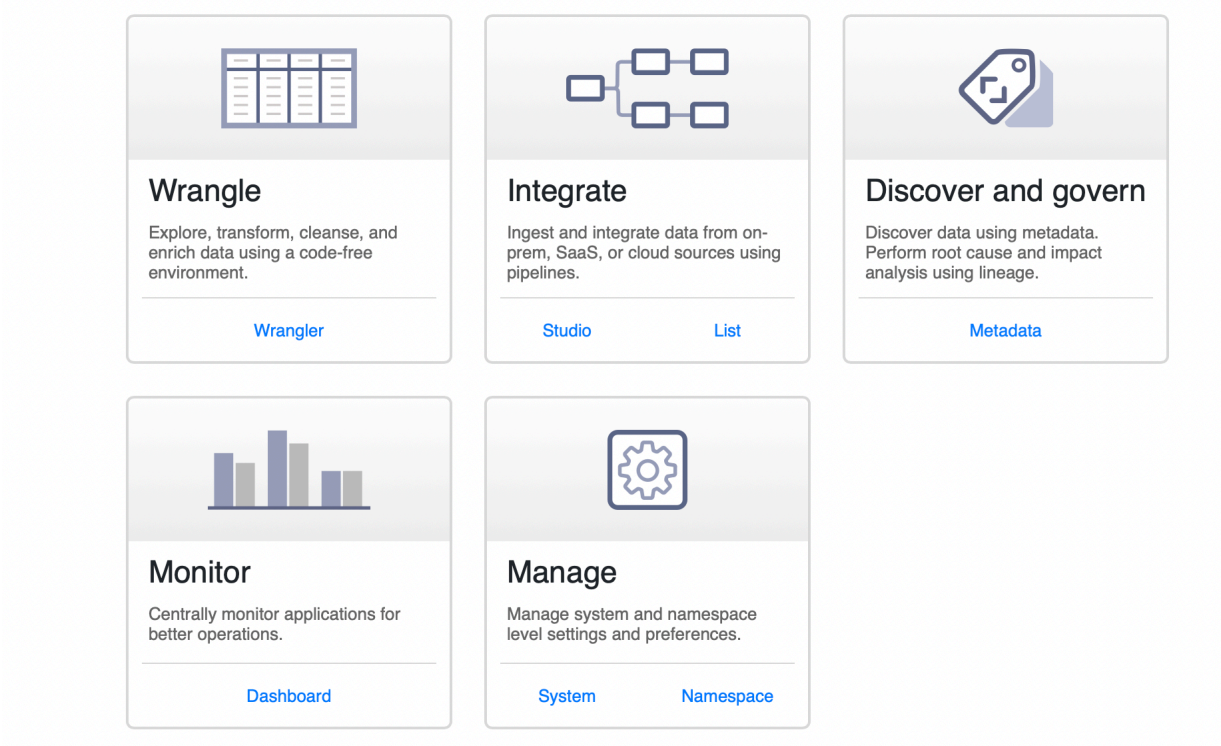
IAM & Admin >> IAM .>> ADD >> ADD PRINCIPALS (COPY THIS SA INTO THE FIELD) >> ROLE (DATA FUSION API SERVICE AGENT) >> SAVE





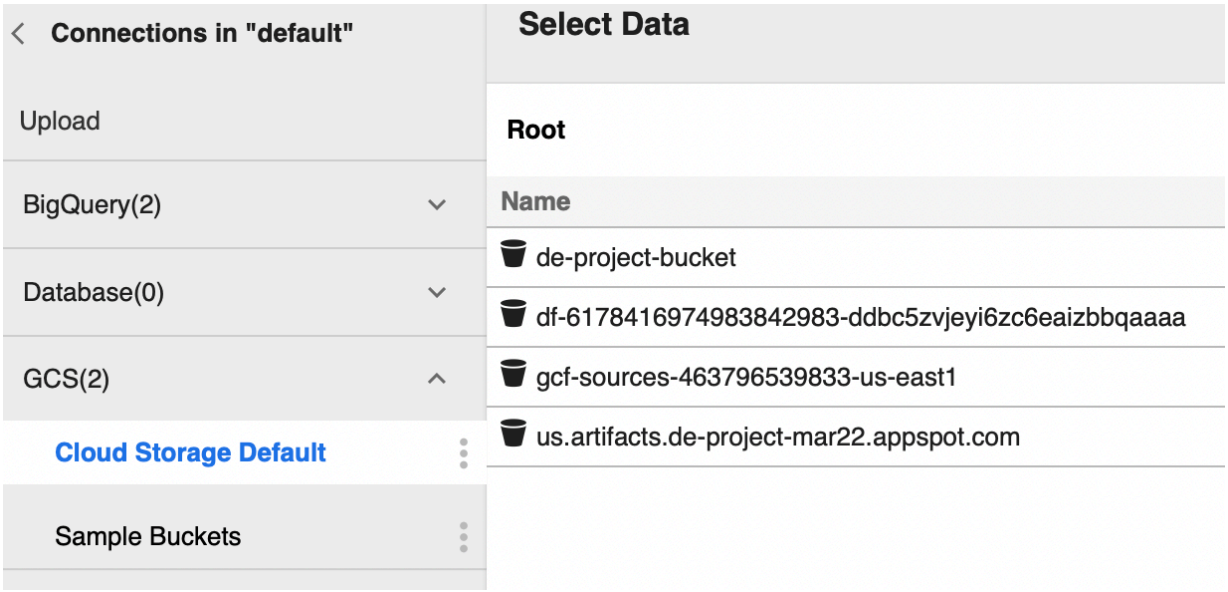
Click on view instance





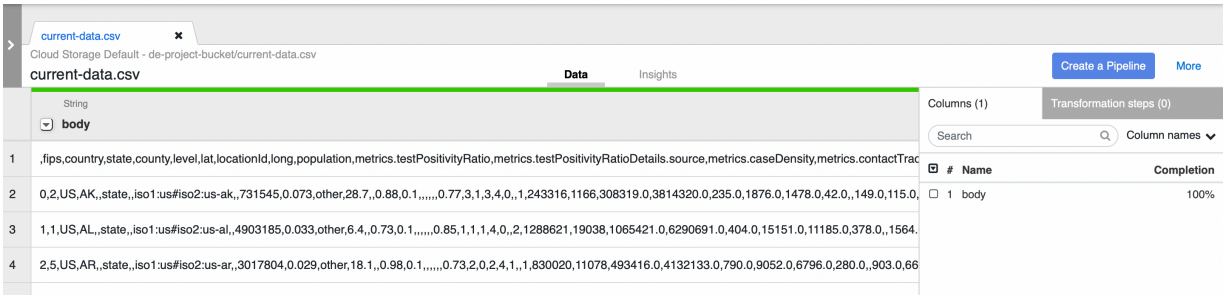
This opens main console for Data Fusion UI . In the cloud console , we can create or delete the Data Fusion instances . The Data Fusion UI is used for hanging the data pipeline.

Hit on **wrangler**



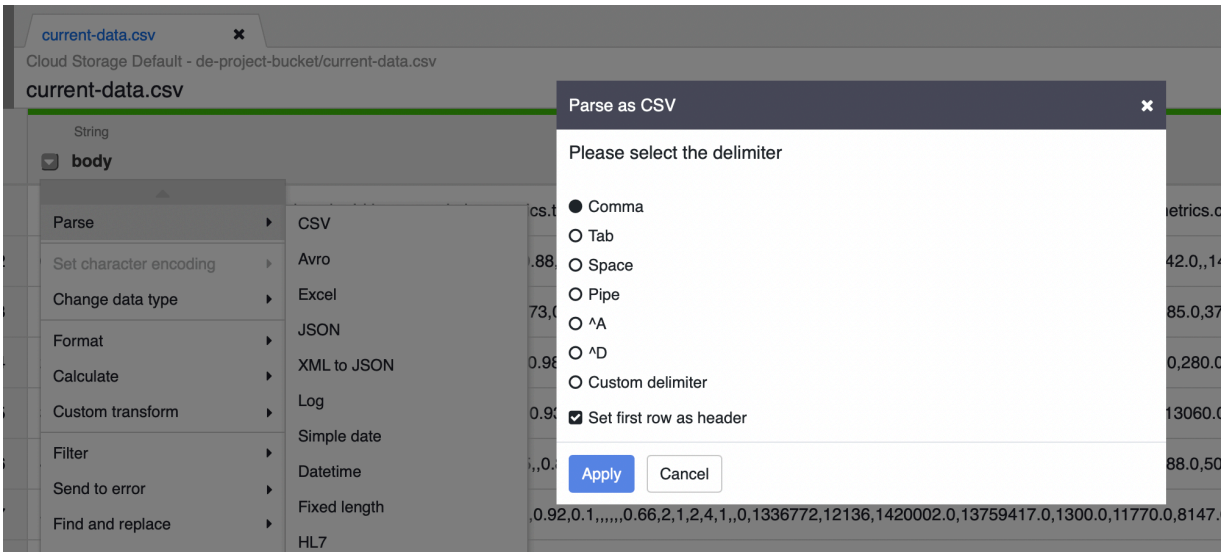
We have to choose the data source from GCS bucket

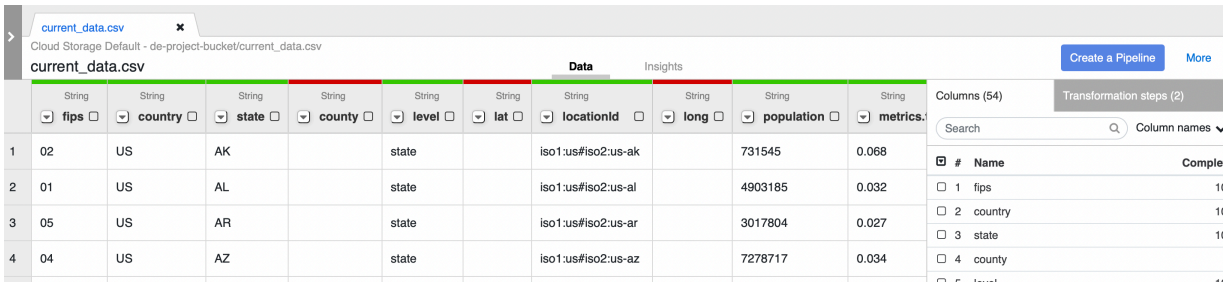
The dataset opens in wrangler window



We will use pipeline studio to all transformation on the data. The transformation can be done on CLI (right at the bottom) or using UI features. The data is loaded into
The wrangler screen

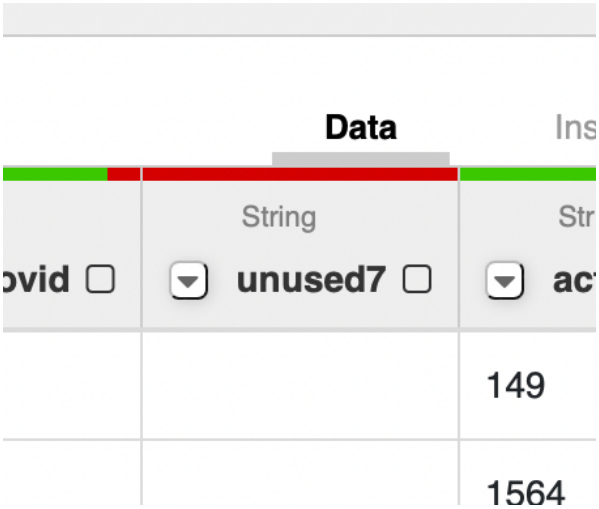
The first operation is to parse the raw CSV data into a tabular representation that is split into rows and columns. To do this, select the dropdown icon from the first column heading (body), and select the **Parse** menu item, and **CSV** from the submenu.



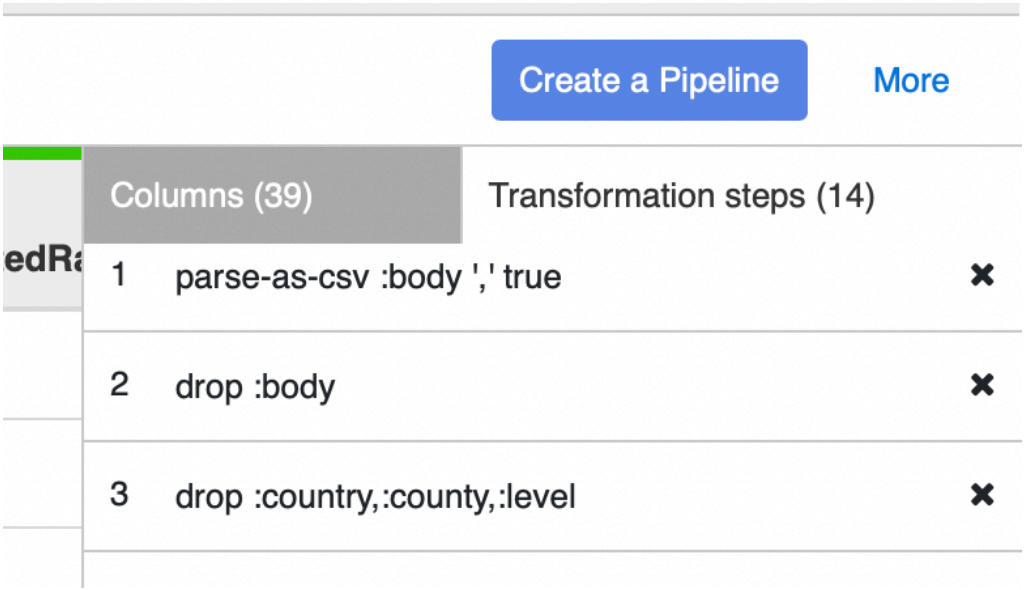


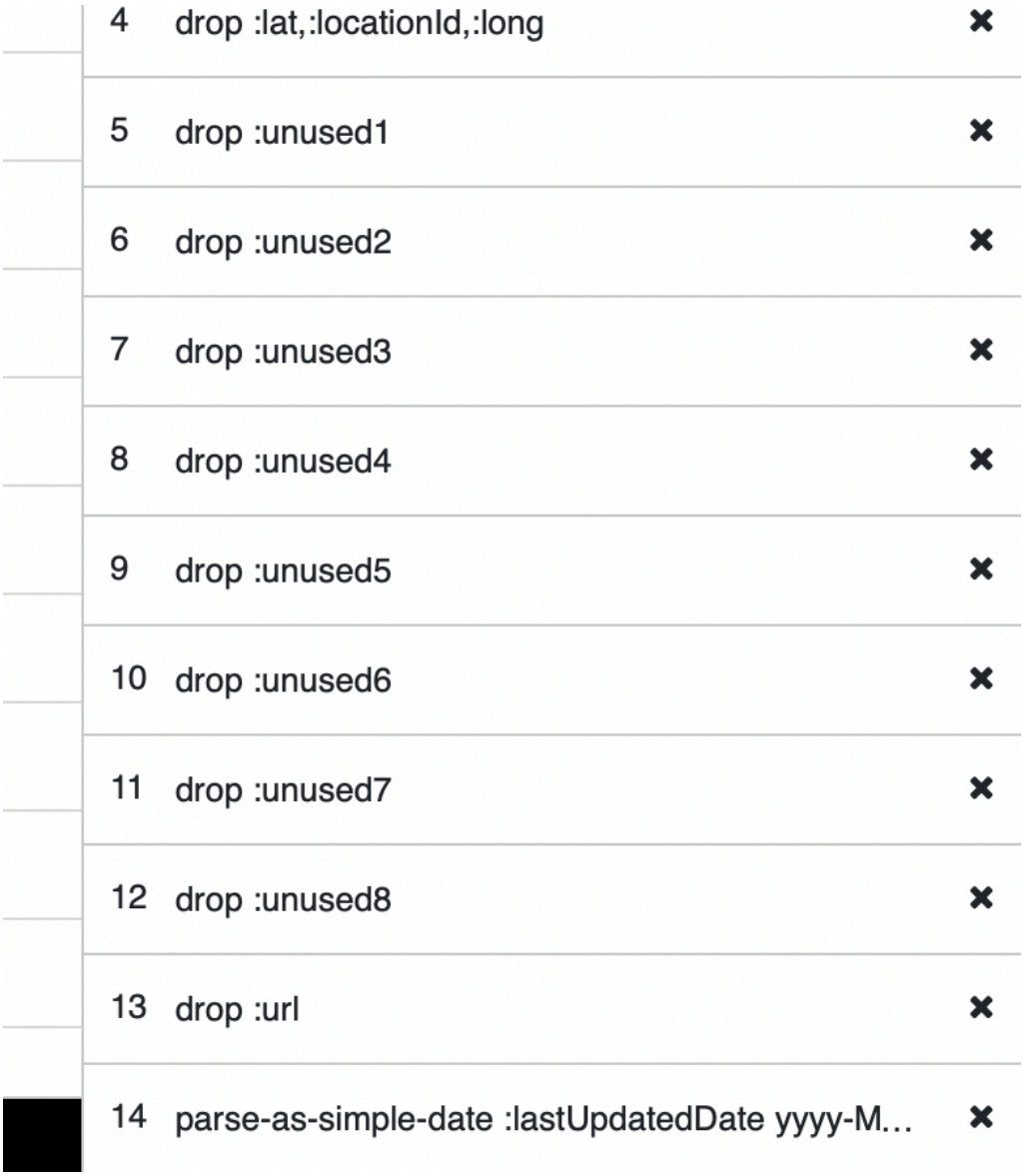
We can delete body column (drop down menu and hit delete column)

The column header with red band indicates max null values

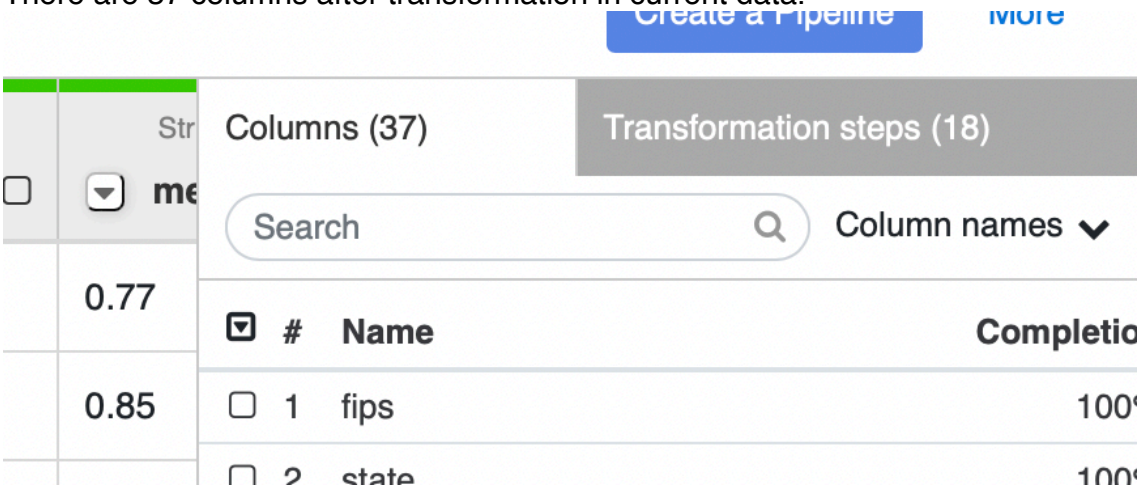


List of transformation can be checked from transformation tab ,

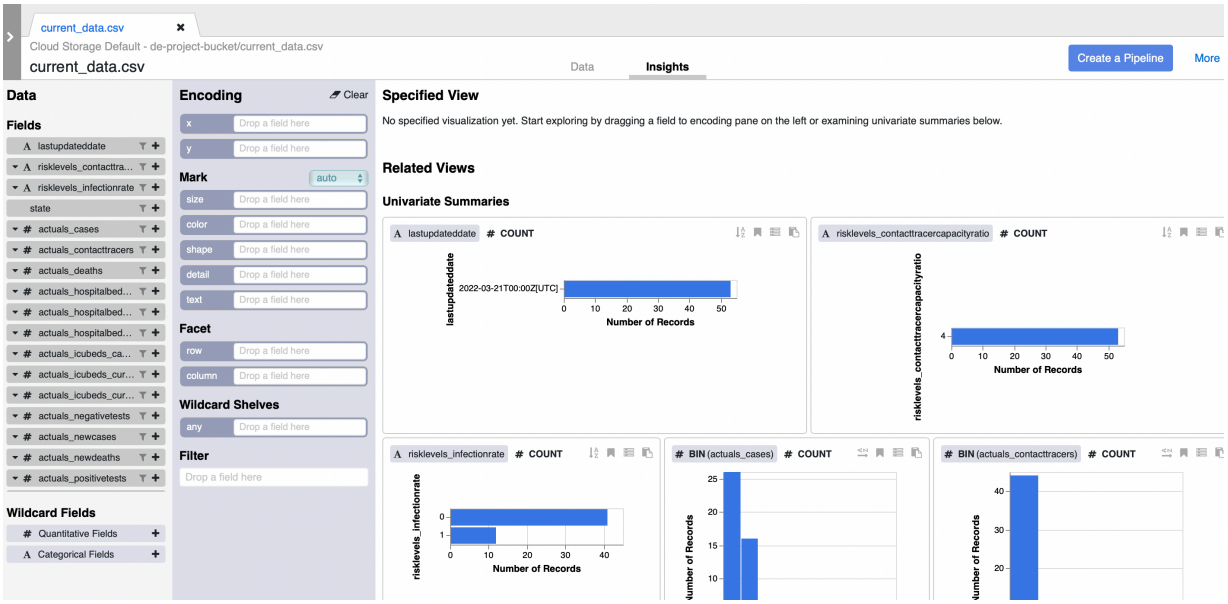




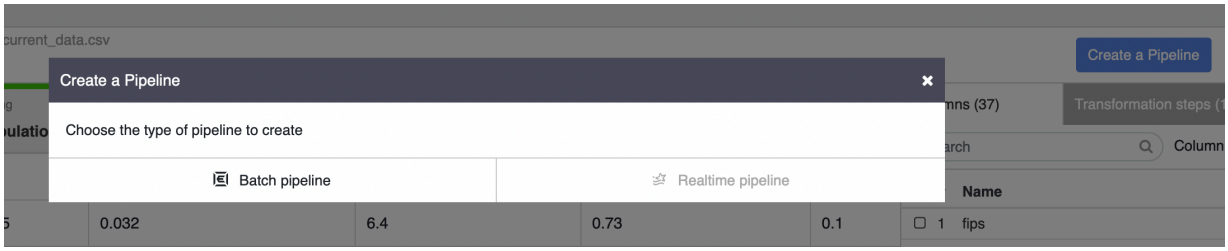
There are 37 columns after transformation in current data.



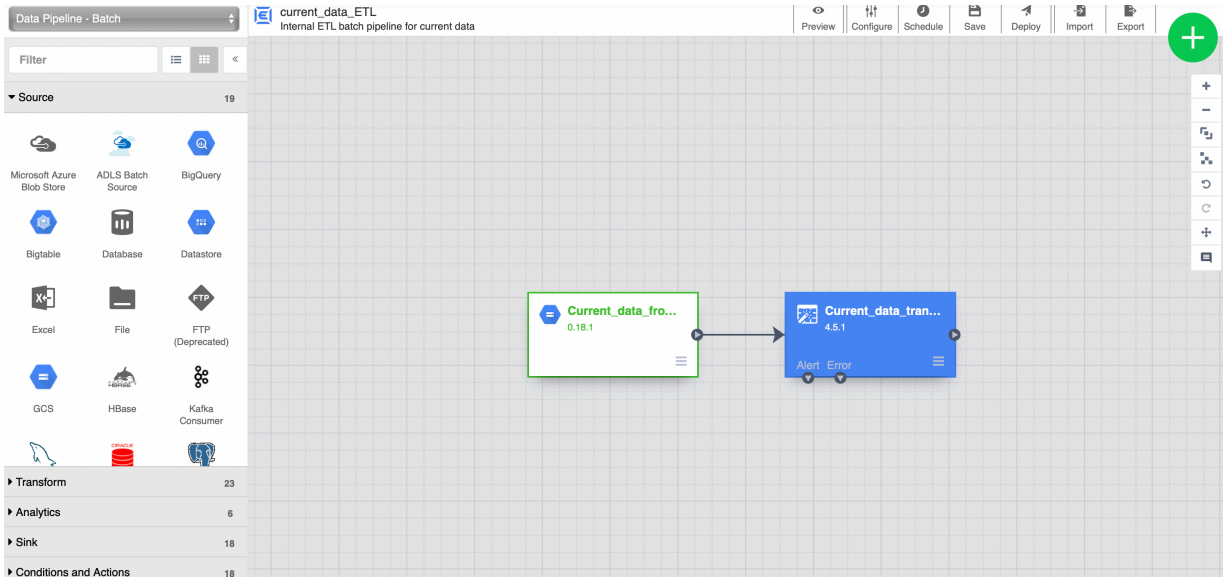
The insights tab gives visual appreciation of column



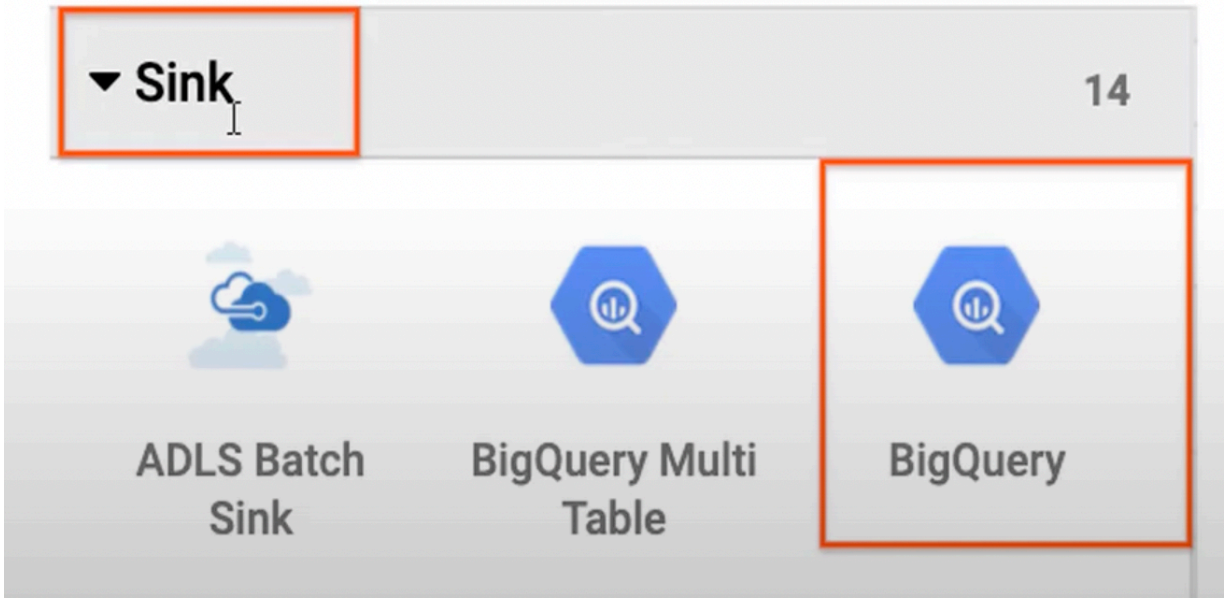
Hit on CREATE PIPELINE



The Data Fusion studio shows our pipeline consisting source and transformation part . We need to create our sink to BQ.



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



Hover your mouse over the **GCS** node and a **Properties** button will be displayed. Click on this button to open up the configuration settings.

The settings for config of BQ sink is as follows (names are indicative only !)

Hover your mouse over your BigQuery node, click on **Properties** and enter the following configuration settings:

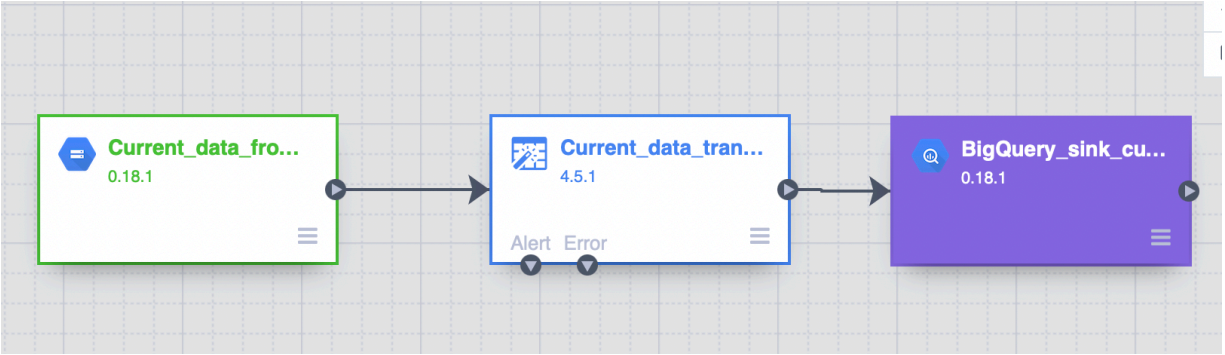
a. For **Reference Name**, enter `Titanic_BQ`

b. For **Dataset**, enter `demo`

c. For **Table**, enter `titanic`

d. Click on **X** on the top right of the Properties box to close it.

Final current data internal ETL pipelline



Now click **Save** from the upper right corner menu. You will be prompted you to give **Name** and add a **description** to the pipeline.

To test your pipeline click on the **Preview** icon. The button bar will now show a run icon that you can click to run the pipeline in preview mode.

Hit RUN to execute in PREVIEW mode.

Once PREVIEW RUN is completed , we can preview the output of each stage of pipeline.

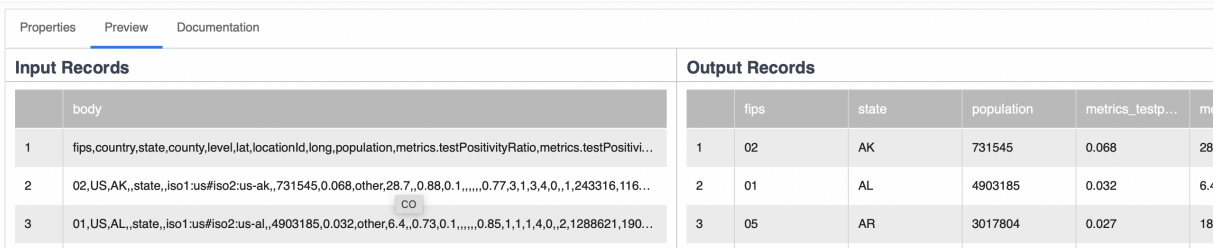
The preview for Extract part is as follows

Properties | Preview | Documentation

Output Records

	body
1	fips,country,state,county,level,lat,locationId,long,popula
2	02,US,AK,,state,,iso1:us#iso2:us-ak,,731545,0.068,oth
3	01,US,AL,,state,,iso1:us#iso2:us-al,,4903185,0.032,oth
4	05,US,AR,,state,,iso1:us#iso2:us-ar,,3017804,0.027,oth

The preview for TRANSFORM



Properties | Preview | Documentation

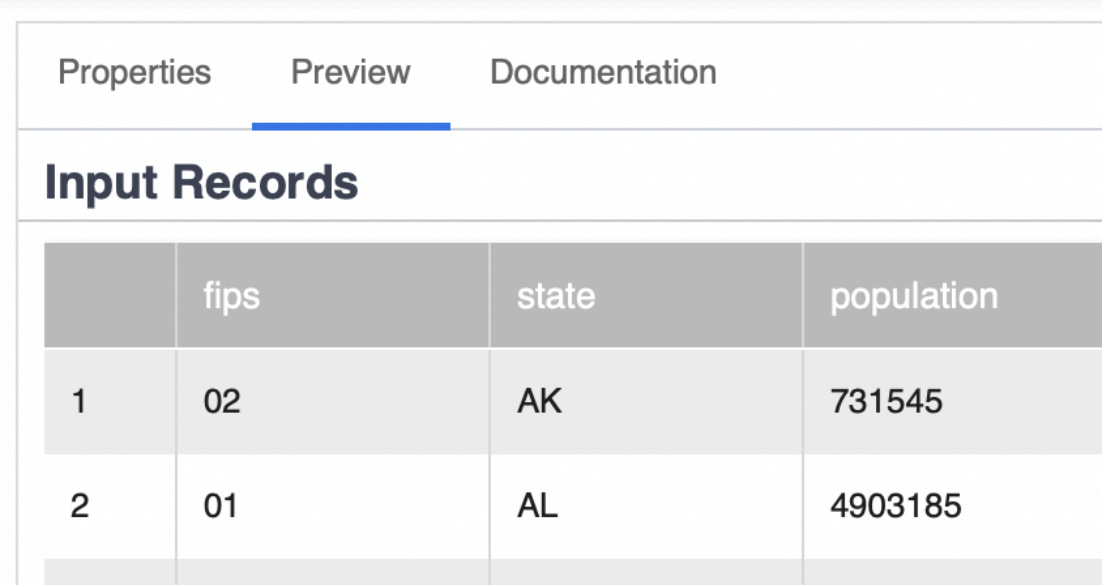
Input Records

	body
1	fips,country,state,county,level,lat,locationId,long,population,metrics.testPositivityRatio,metrics.testPositivi…
2	02,US,AK,,state,,iso1:us#iso2:us-ak,,731545,0.068,other,28.7,,0.88,0.1,,,,,,0.77,3,1,3,4,0,,1,243316,116…
3	01,US,AL,,state,,iso1:us#iso2:us-al,,4903185,0.032,other,6.4,,0.73,0.1,,,,,,0.85,1,1,1,4,0,,2,1288621,190…

Output Records

	fips	state	population	metrics_testp…	m
1	02	AK	731545	0.068	28
2	01	AL	4903185	0.032	6.4
3	05	AR	3017804	0.027	18

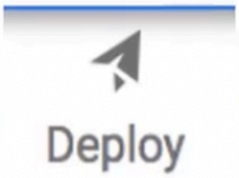
The preview for SINK



Once PREVIEW is satisfactory ,

Click on the **Preview** icon again, this time to toggle out of Preview mode.

If everything looks good so far, you can proceed to deploy the pipeline. Click on the

Deploy icon on the top right 

to deploy the pipeline.

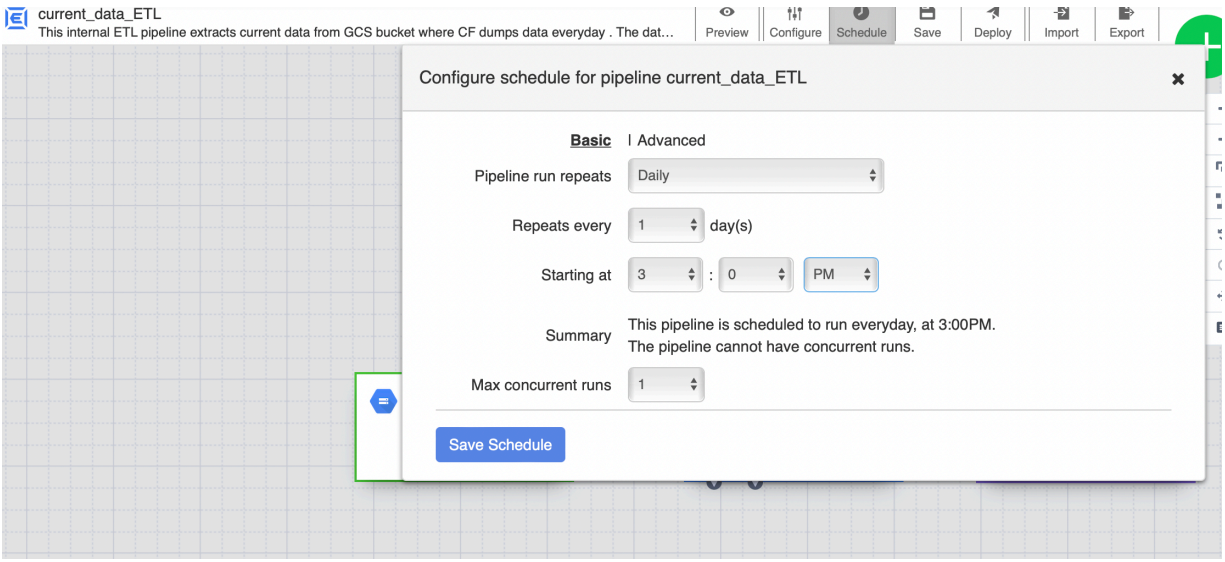
You will see a confirmation dialog that your pipeline is being deployed:

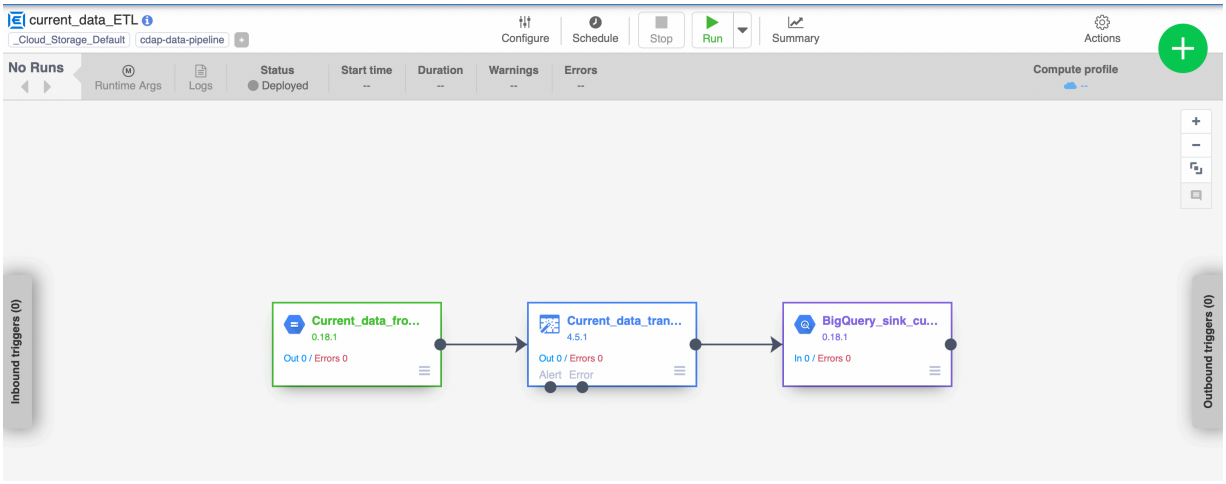
'. Once your pipeline has successfully deployed, you're now ready to run your ETL pipeline and load some data into BigQuery.

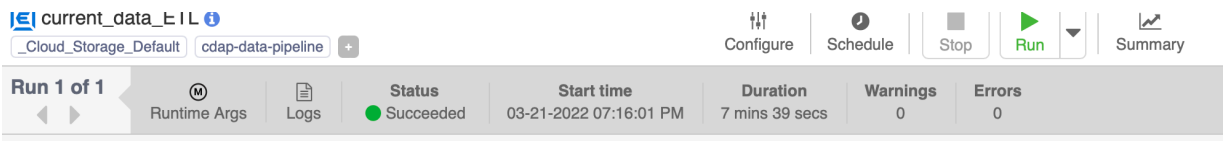
. Click on the **Run** icon to execute the ETL job.

. When done you should see pipeline status changes to **Succeeded** indicating that the pipeline ran successfully.

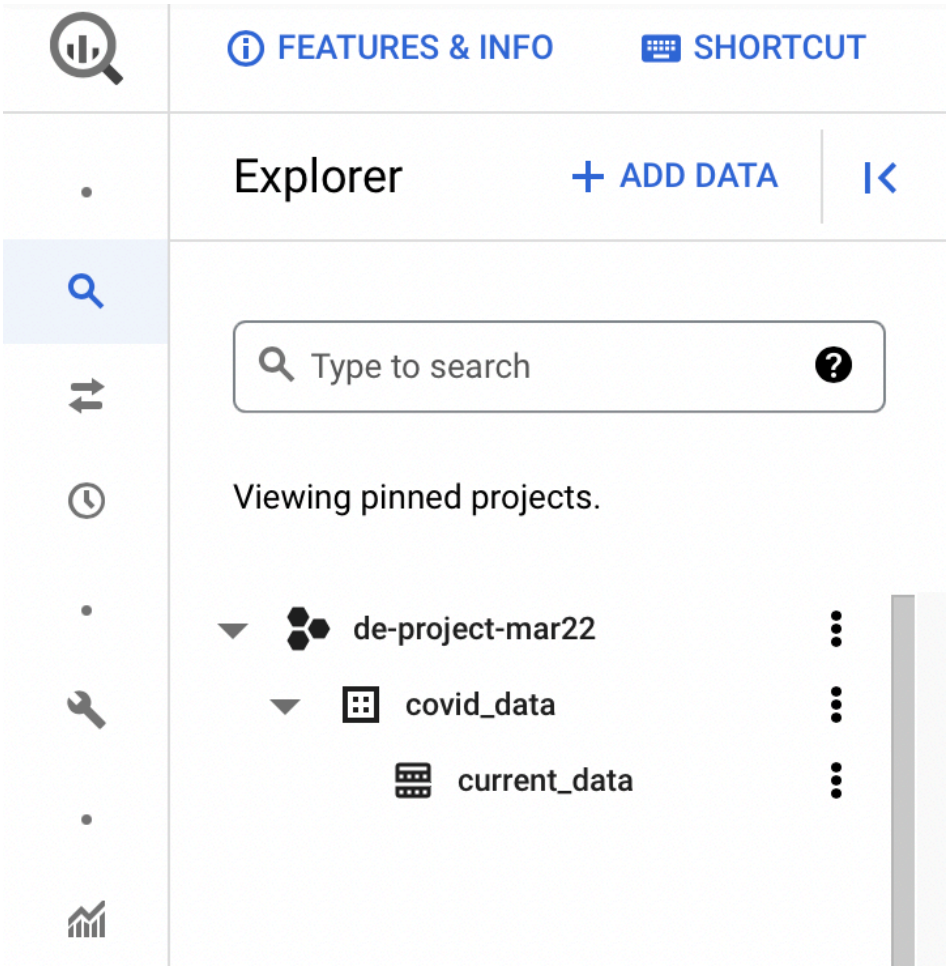
The pipeline can be scheduled to run at a particular time . It is to be remembered that the time indicated is UTC and not local time . As data ingestion takes place at 1400 hrs every day , we can run the current data pipeline at 1430hrs everyday.







We can head over to BQ to confirm our table has been uploaded

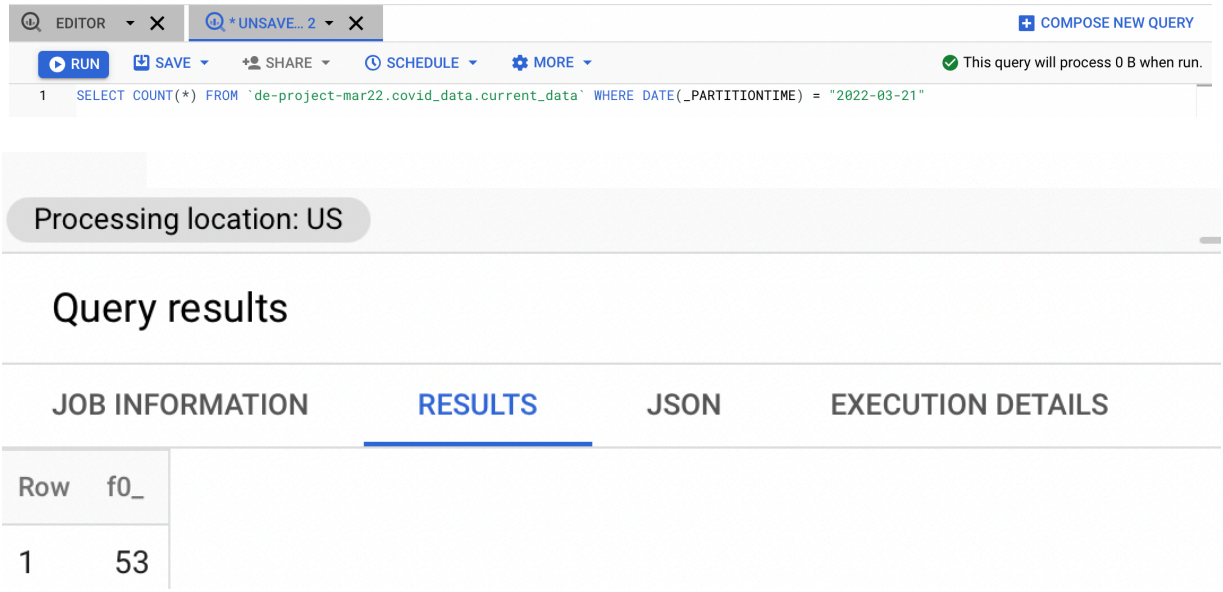


The preview of data in our BQ table is as shown

SCHEMA DETAILS PREVIEW

Row	fips	state	population	metrics_testpositivityratio	metrics_casedensity	metrics_infectionrate	metrics_infectionrateci90	metrics_icucapacityratio	risklevels_overall	riskle
1	20	KS	2913314	0.026	2.8	0.58	0.12	0.7	1	0
2	19	IA	3155070		3.5	0.67	0.1	0.66	1	4
3	46	SD	884659	0.032	3.9	0.65	0.2	0.56	1	1

We can run a query and confirm no of rows



The ETL pipeline seems to run fine ..

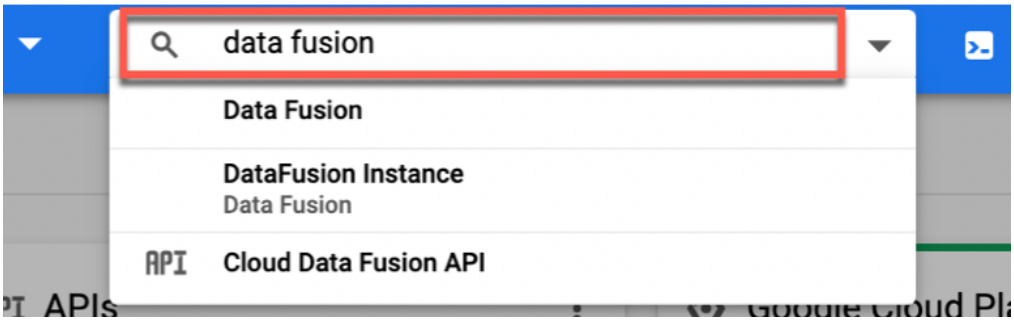
ETL PIPELINE FOR COMBINED(MERGED DATA) IN DATA FUSION

The sequence of activities for creating a pipeline for combined data to BQ table is exactly similar to the one indicated above . However, the list of transformations carried out may be slightly different .

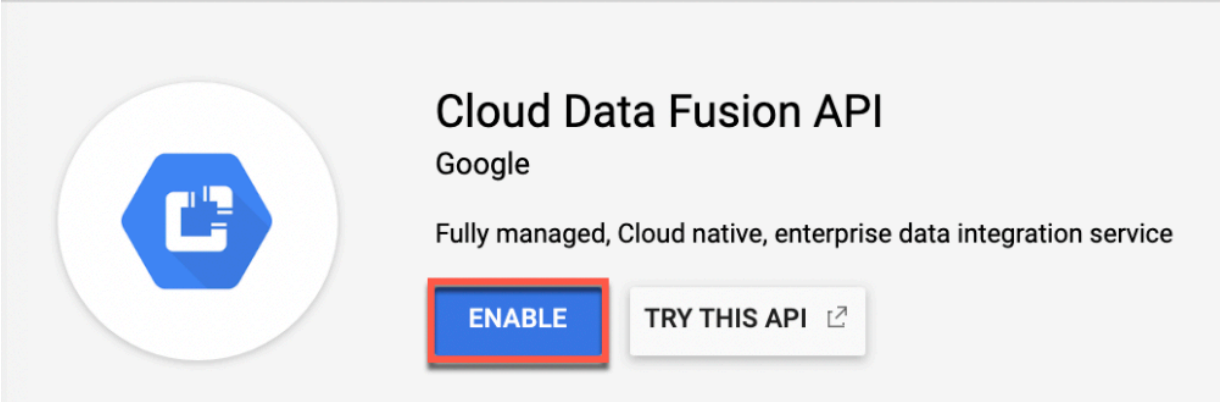
The screen shots indicating the steps (only samples included as extensive step wise process has been already indicated above) have been shown below

(A) **CREATE A DATA FUSION INSTANCE**

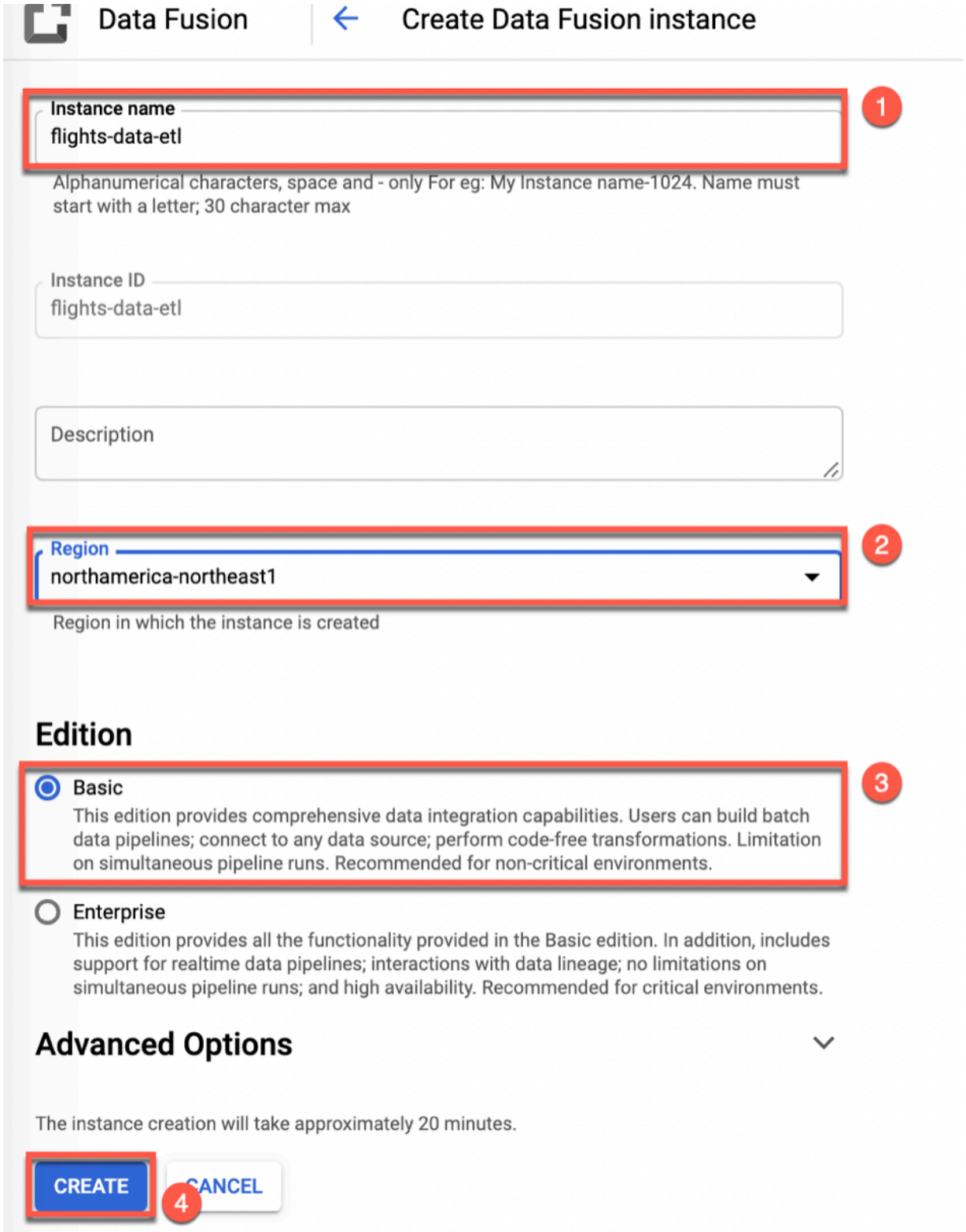
Go to the Cloud Data Fusion page. You can do this by typing “data fusion” in the resources and products search field and then selecting *Data Fusion*.

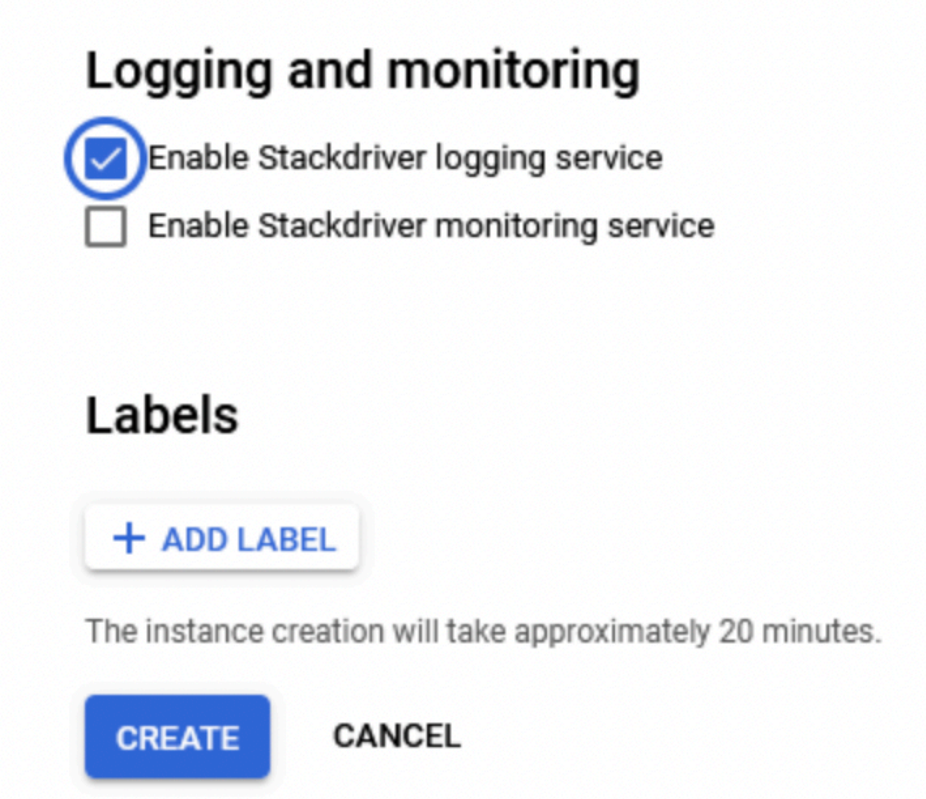


If the Cloud Data Fusion API is not already enabled, activate it by clicking *Enable*. This might take a few moments to complete.

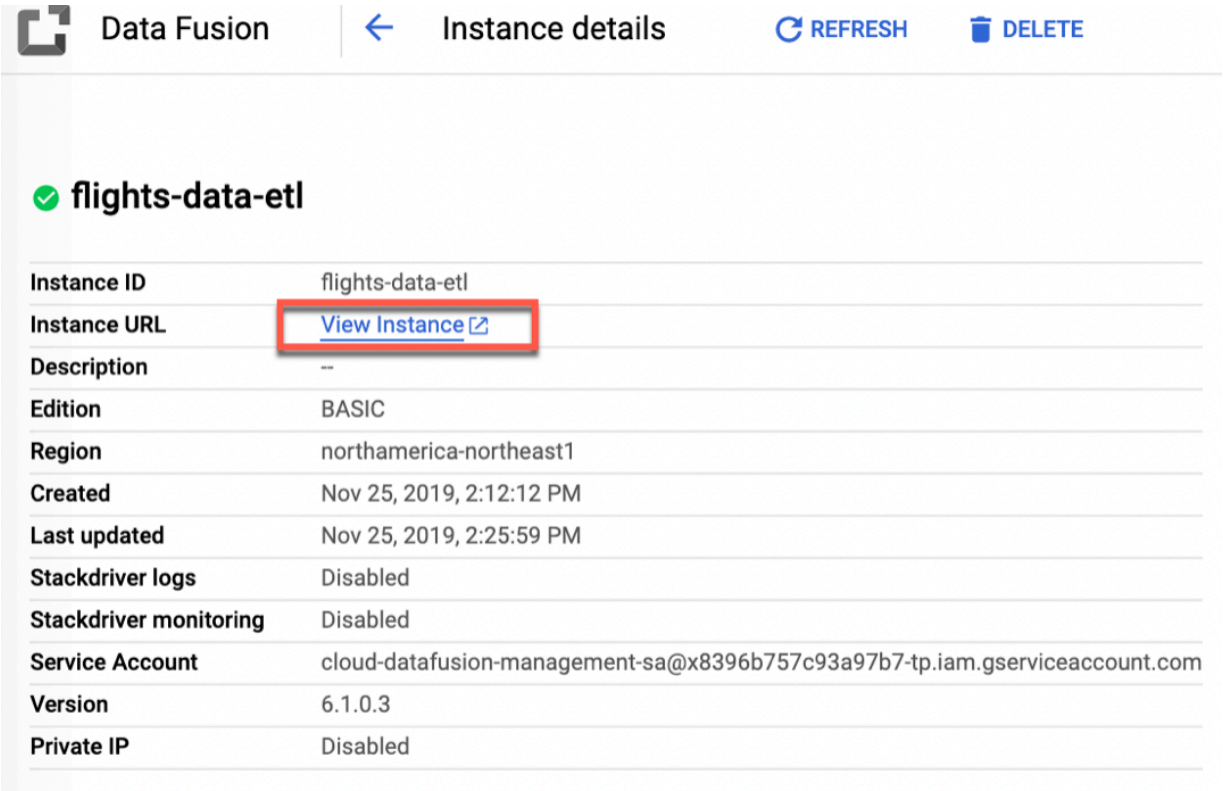


Navigate back to the Data Fusion page. Create a Data Fusion instance. Click *Create An Instance*. Enter an *Instance name* **(1)**, select your *Region* **(2)**, select *Basic* for *Edition* **(3)**, and click *Create* **(4)** to deploy your instance.

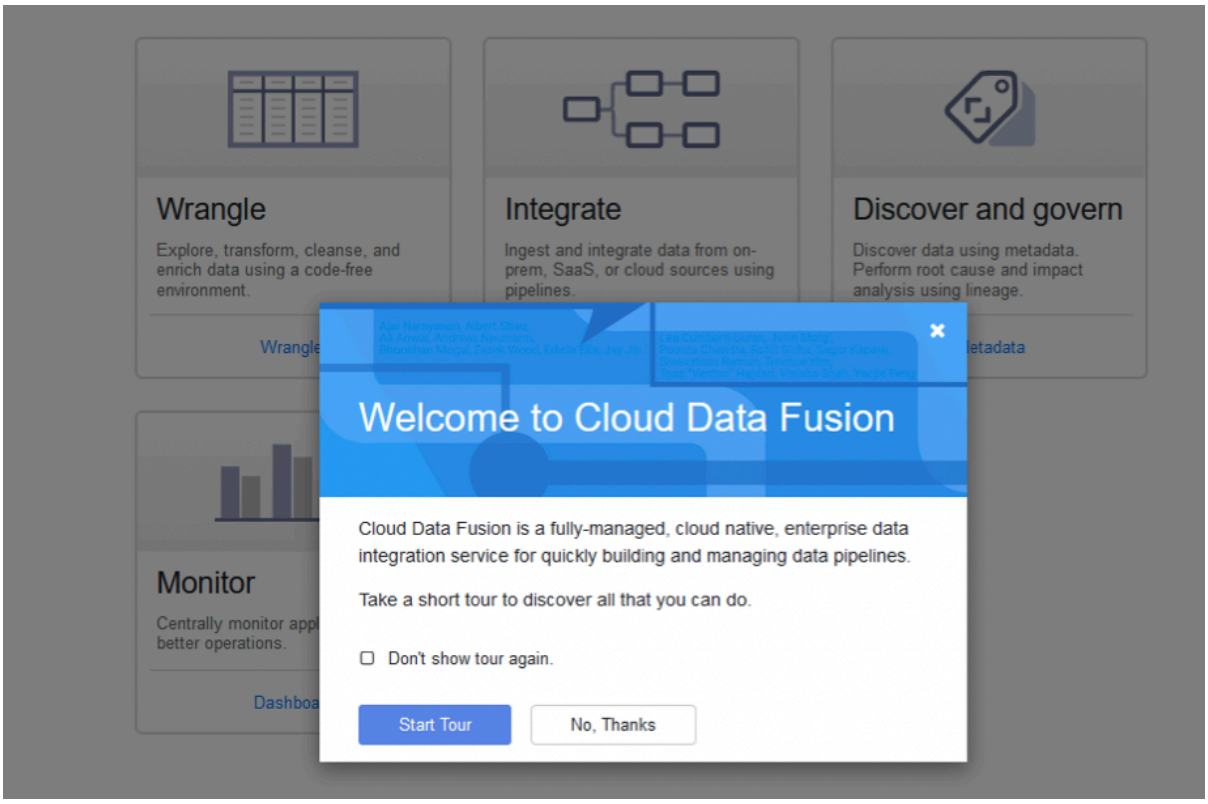

Data Fusion
Create Data Fusion instance
Instance name
flights-data-etl
1
Alphanumerical characters, space and - only For eg: My Instance name-1024. Name must start with a letter; 30 character max
Instance ID
flights-data-etl
Description
Region
northamerica-northeast1
2
Region in which the instance is created
Edition
Basic
This edition provides comprehensive data integration capabilities. Users can build batch data pipelines; connect to any data source; perform code-free transformations. Limitation on simultaneous pipeline runs. Recommended for non-critical environments.
3
Enterprise
This edition provides all the functionality provided in the Basic edition. In addition, includes support for realtime data pipelines; interactions with data lineage; no limitations on simultaneous pipeline runs; and high availability. Recommended for critical environments.
Advanced Options
The instance creation will take approximately 20 minutes.
CREATE
CANCEL
4



After the instance deploys successfully, a green check mark appears. Click *View Instance* to continue.

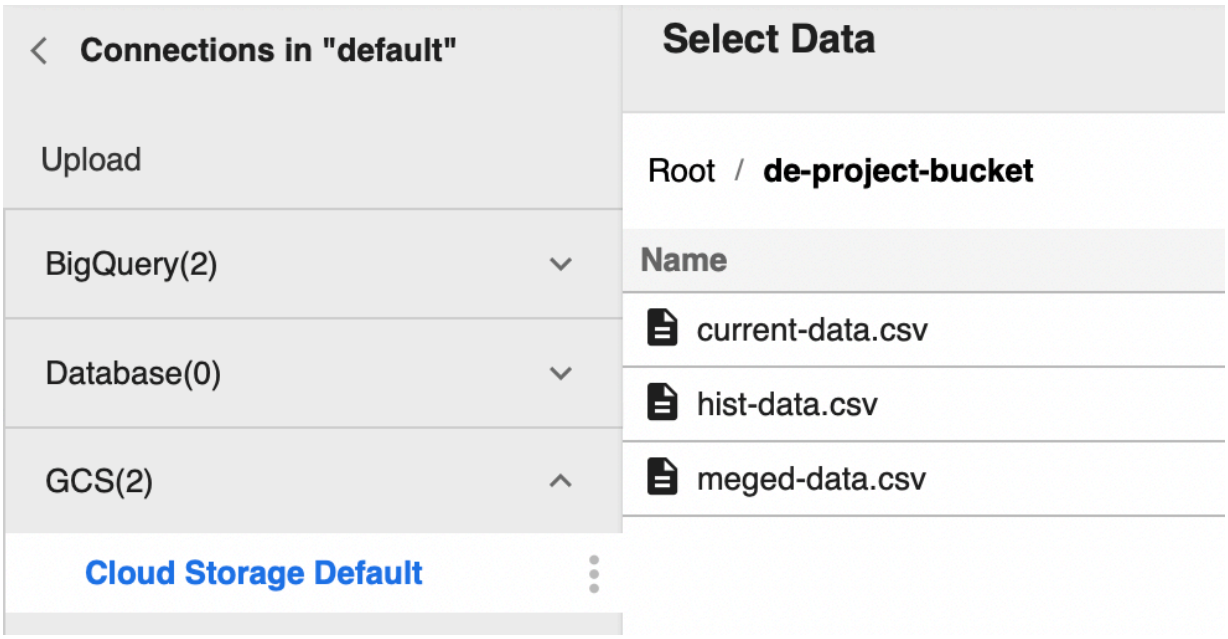


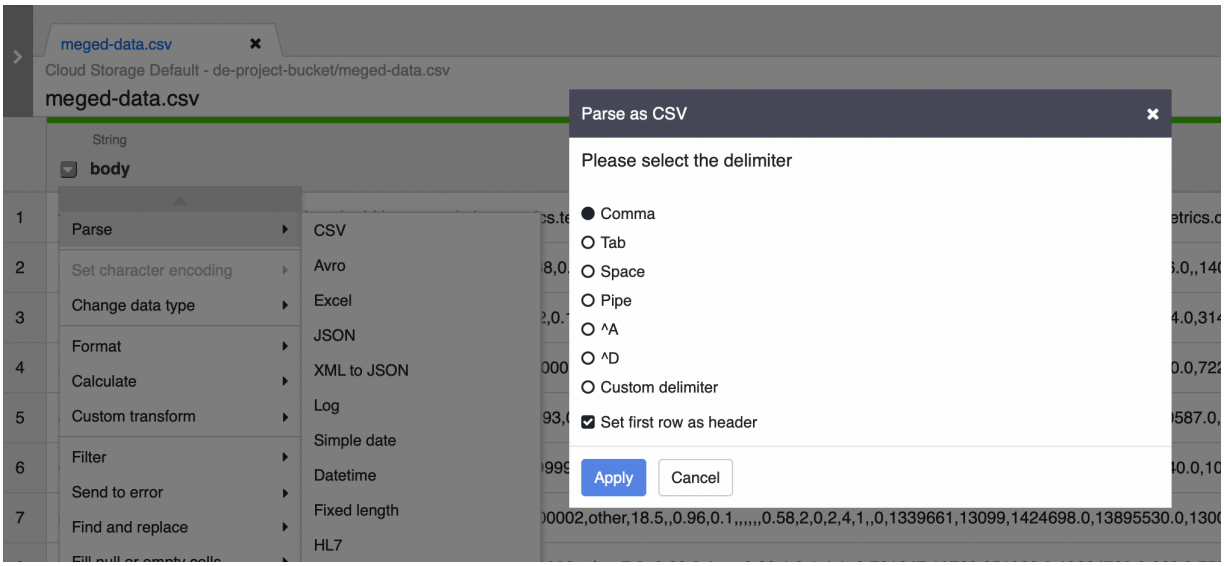
Once the Data Fusion instance is created, copy the Service Account Data Fusion is using and grant it the **“Cloud Data Fusion API Service Agent”** role by navigating to IAM and clicking the +ADD button, with this role assigned to the Data Fusion Service Account, Data Fusion can access data from/to other services such as Cloud Storage, BigQuery and Dataproc.



You may click on the Start Tour button if you're not familiar with Data Fusion to take a look and get familiar with it. For the purposes of this exercise we will skip it for now, click in the **No Thanks** button.

Click on WRANGLER . Navigate to our GCS bucket and locate merged-data.csv.





Now we can carry out transformation on the data set.

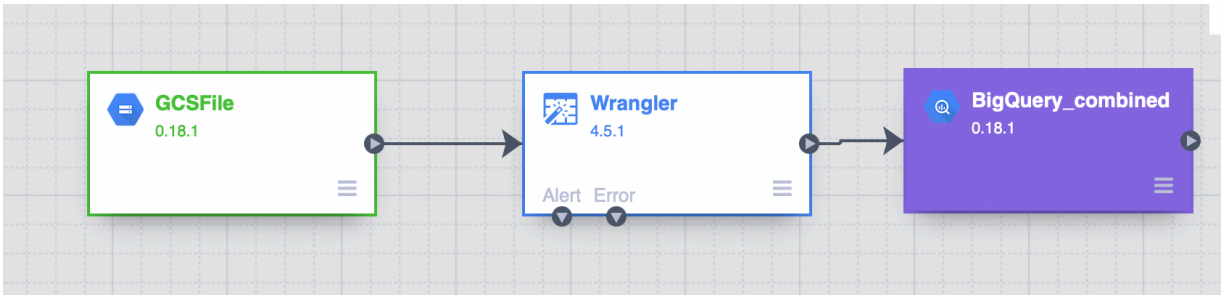
(a) Delete body column and other not relevant columns

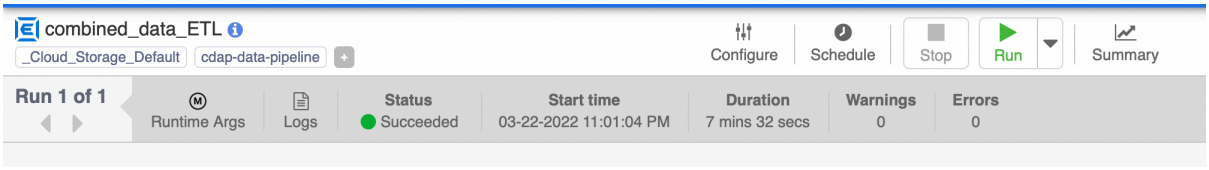
(B) parse lasUpdatedDate as simple date

The red color indicates missing values in the column (% of missing values).

Then take BQ as sink and fill in dataset and table details

By default, and as this is a CSV file, Data Fusion treats all the inputs as Strings, we need to change that to insert the data properly.We will use the Change Data Type option in the arrow menu on the columns to set the data types





We will schedule the pipeline to run everyday at 1500 hrs local (1900 UTC) . We can confirm the BQ table being populated with combined table.