

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', 'metrics.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.

Cloud Data Fusion runs pipelines using Dataproc clusters. Cloud Data Fusion automatically provisions ephemeral Dataproc clusters, runs pipelines on them, and then tears down the clusters after the pipeline run completes.

The running of pipeline can be automated .

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

API/Service Details

 [DISABLE API](#)

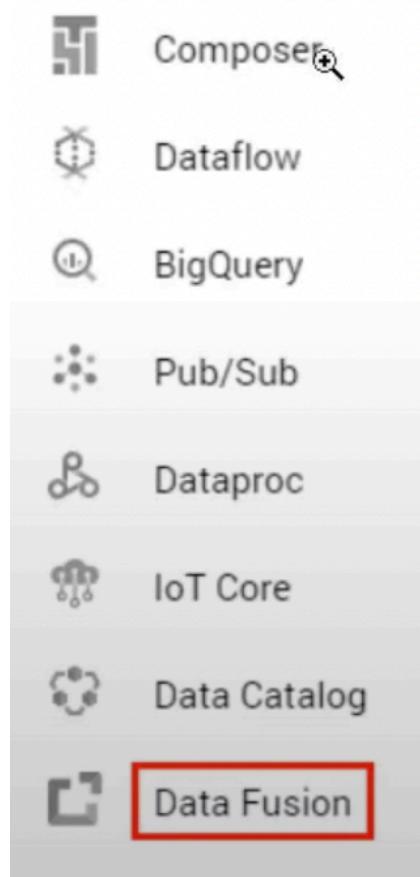


Cloud Data Fusion API

Cloud Data Fusion is a fully-managed, cloud native service for quickly building and managing data pipelines. It

Create Cloud Data Fusion Instance

BIG DATA



Data Fusion

Create an instance

Cloud Data Fusion - a fully-managed, cloud native, enterprise data integration service for quickly building and managing data pipelines.

[CREATE AN INSTANCE](#)

Edition - Basic

Advanced Options

Private IP

- Enable Private IP

Private IP connectivity requires additional AI
your organization's administrator for help en
cannot be disabled once it has been enabled

Logging and monitoring

- Enable Stackdriver logging service
- Enable Stackdriver monitoring service

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

In the Console navigate to the **IAM & admin > IAM**.

On the IAM Permissions page, click **Add**.

In the New Members field paste the service account.

Click into the Select a role field and start typing **Cloud Data Fusion API Service Agent**, then select it.

Add members, roles to "qwiklabs-gcp-01-ca64ad9602c6" project

Enter one or more members below. Then select a role for these members to grant them access to your resources. Multiple roles allowed. [Learn more](#)

New members

cloud-datafusion-management-sa@mc207729dc530ec03p-tp.iam.gserviceaccount.com

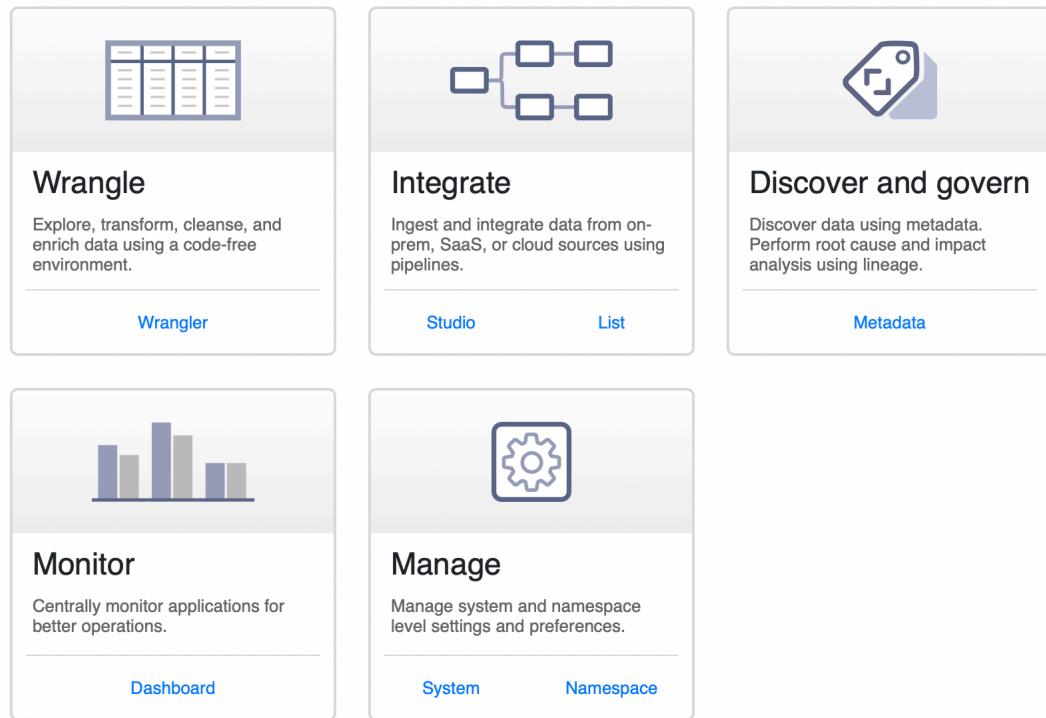
? X

<p>Role</p> <div style="border: 1px solid #ccc; padding: 2px; border-radius: 5px; width: 90%;"><p>Cloud Data Fusion API Se... ▾</p></div> <p>Gives Cloud Data Fusion service account access to Service Networking, Cloud Dataproc, Cloud Storage, BigQuery, Cloud Spanner, and Cloud Bigtable resources.</p> <p>+ ADD ANOTHER ROLE</p>	<p>Condition</p> <p>Add condition</p> <p>X</p>
--	---

SAVE CANCEL

Click on view instance

Instance ID	daily-covid-data-pipeline
Instance URL	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**

The Wrangler interface shows the following connections in the "default" project:

- Upload
- BigQuery(2)
- Database(0)
- GCS(2)
 - Cloud Storage Default**
 - Sample Buckets

The "Select Data" panel is open, showing the "Root" node with the following buckets listed under "Name":

- de-project-bucket
- df-6178416974983842983-ddbc5zvjeyi6zc6eaizbbqaaaa
- gcf-sources-463796539833-us-east1
- us.artifacts.de-project-mar22.appspot.com

We have to choose the data source from GCS bucket

The dataset opens in wrangler window

The screenshot shows the 'current-data.csv' dataset in the Wrangler interface. The top navigation bar includes 'Create a Pipeline' and 'More'. Below the header, there are tabs for 'Data' and 'Insights'. The main area displays the raw CSV data with four rows of sample data. The first row is labeled 'String' and contains the column 'body'. The second row starts with '0,2,US,AK,...'. The third row starts with '1,1,US,AL,...'. The fourth row starts with '2,5,US,AR,...'. On the right side, there are sections for 'Columns (1)', 'Transformation steps (0)', and a search bar. A progress bar at the bottom indicates 'Completion 100%'.

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.

The screenshot shows the 'Parse as CSV' dialog box overlaid on the Wrangler interface. The dialog title is 'Parse as CSV' and it says 'Please select the delimiter'. It lists several options: 'Comma' (selected), 'Tab', 'Space', 'Pipe', '^A', '^D', and 'Custom delimiter'. There is also a checked checkbox 'Set first row as header'. At the bottom are 'Apply' and 'Cancel' buttons.

	fips	country	state	county	level	lat	locationId	long	population	metrics	body
1	02	US	AK		state		iso1:us#iso2:us-ak		731545	0.068	
2	01	US	AL		state		iso1:us#iso2:us-al		4903185	0.032	
3	05	US	AR		state		iso1:us#iso2:us-ar		3017804	0.027	
4	04	US	AZ		state		iso1:us#iso2:us-az		7278717	0.034	

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

The column header with red band indicates max null values

body	unused7	ac
		149
		1564

List of transformation can be checked from transformation tab ,

Columns (39)	Transformation steps (14)
edRa	1 parse-as-csv :body ', true
	2 drop :body
	3 drop :country,:county,:level

4	drop :lat,:locationId,:long	x
5	drop :unused1	x
6	drop :unused2	x
7	drop :unused3	x
8	drop :unused4	x
9	drop :unused5	x
10	drop :unused6	x
11	drop :unused7	x
12	drop :unused8	x
13	drop :url	x
14	parse-as-simple-date :lastUpdatedDate yyyy-M...	x

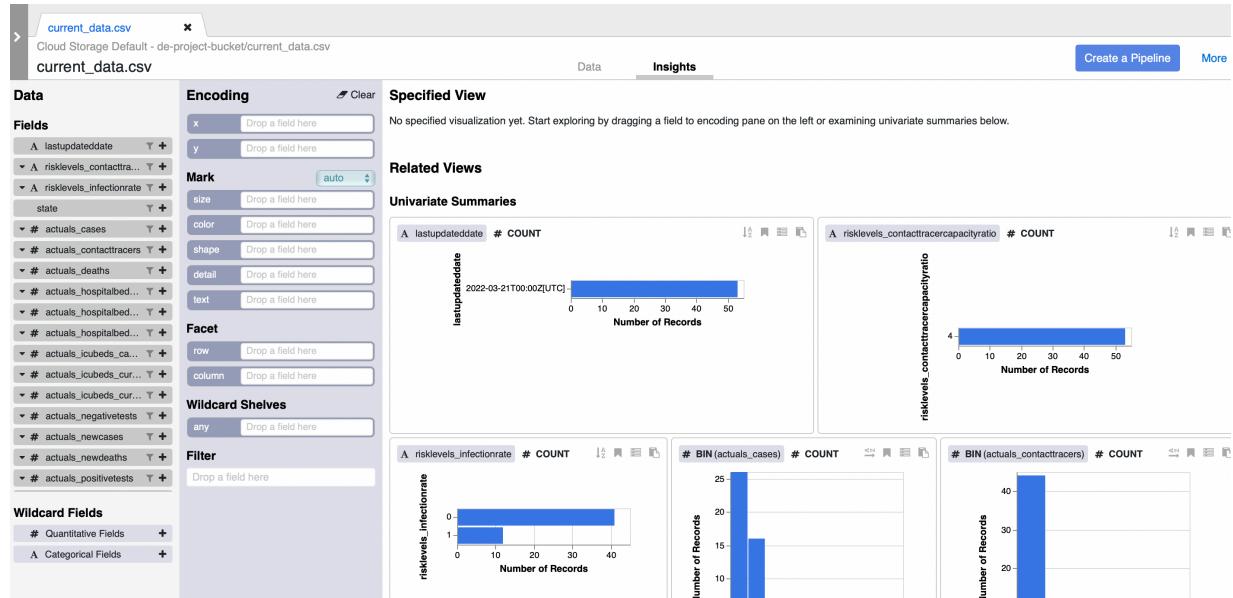
There are 37 columns after transformation in current data.

[Create a Pipeline](#)

[MORE](#)

	Str	Columns (37)	Transformation steps (18)
<input type="checkbox"/>	<input type="button" value="▼"/> me		
0.77		<input type="text" value="Search"/> Column names ▾	
0.85		<input checked="" type="checkbox"/> # Name	Completion
		<input type="checkbox"/> 1 fips	100%
		<input type="checkbox"/> 2 state	100%

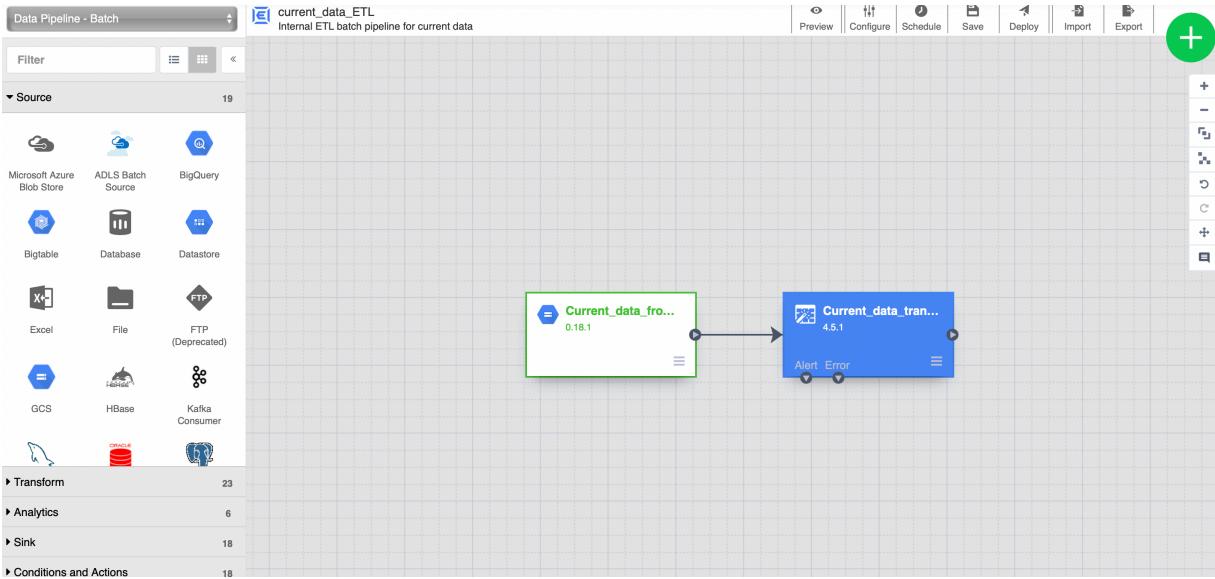
The insights tab gives visual appreciation of column



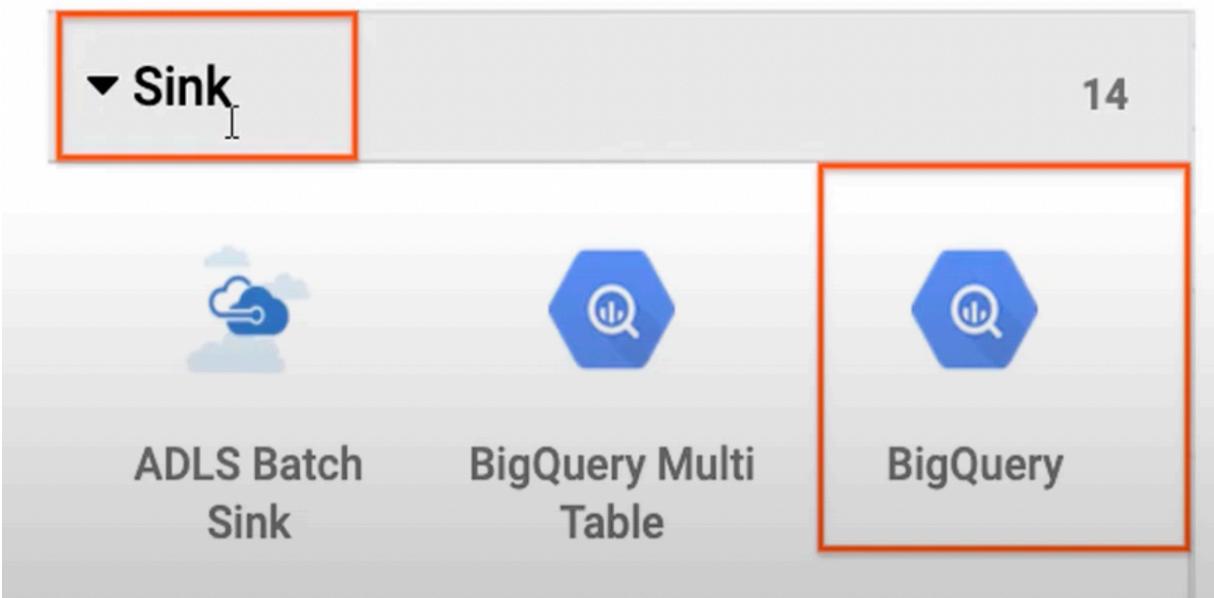
Hit on CREATE PIPELINE



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:

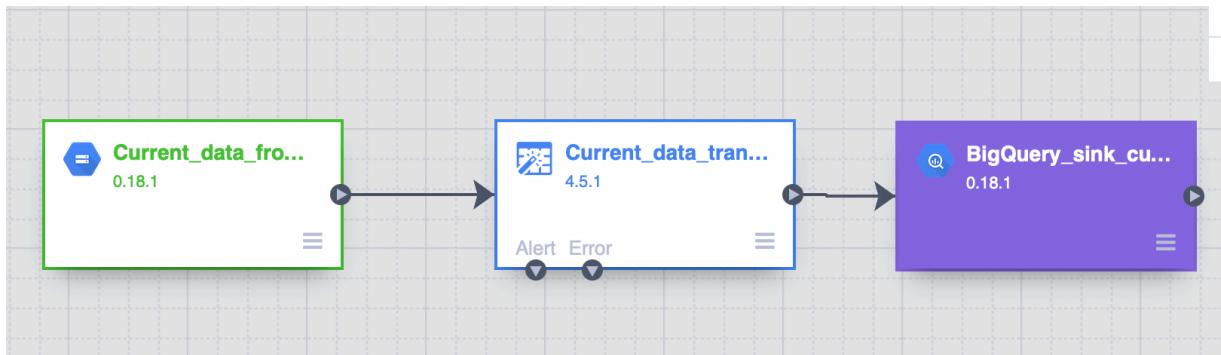
a. For **Reference Name**, enter `Titanic_BQ`

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 pipeline



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	Output Records					
	body		tips	state	population	metrics_testp...	me...	
1	fips,country,state,county,level,lat,locationId,long,population,metrics.testPositivityRatio,metrics.testPositivi...		1	02	AK	731545	0.068	28
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...	CO	2	01	AL	4903185	0.032	6.4
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...		3	05	AR	3017804	0.027	18

The preview for SINK

Properties Preview Documentation

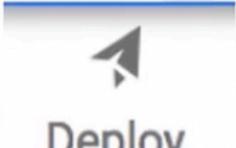
Input Records

	fips	state	population
1	02	AK	731545
2	01	AL	4903185

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.

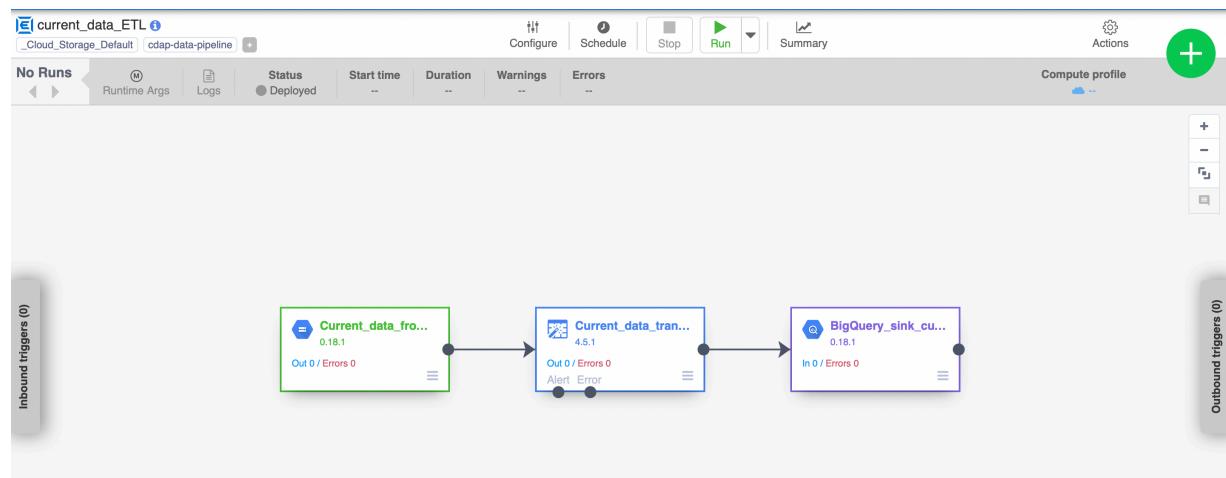
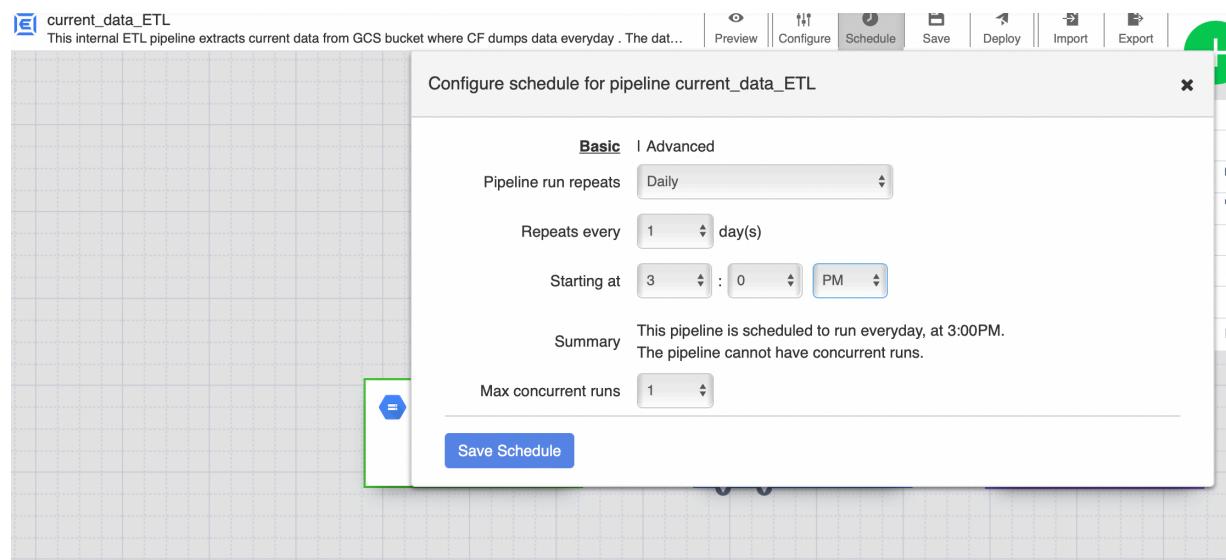
You will see a confirmation dialog that your pipeline is being deployed:

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

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

i. 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.



The screenshot shows the CDAP Data Pipeline interface. At the top, there's a header with the pipeline name 'current_data' and a status indicator. Below the header, a navigation bar includes 'Configure', 'Schedule', 'Stop', 'Run', and 'Summary' buttons. Underneath the navigation bar, a summary row displays 'Run 1 of 1', 'Runtime Args', 'Logs', 'Status Succeeded', 'Start time 03-21-2022 07:16:01 PM', 'Duration 7 mins 39 secs', 'Warnings 0', and 'Errors 0'.

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

The screenshot shows the BigQuery UI. On the left, there's a sidebar with icons for 'Explorer', 'Search' (highlighted), 'Pinned projects', 'Jobs', 'Tables', and 'Metrics'. The main area is titled 'FEATURES & INFO' and 'SHORTCUT'. It shows an 'Explorer' section with a search bar and a pinned project 'Viewing pinned projects' containing 'de-project-mar22' (with sub-table 'covid_data') and 'current_data'.

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

Processing location: US

Query results

Row	f0_
1	53

The ETL pipeline seems to run fine ..

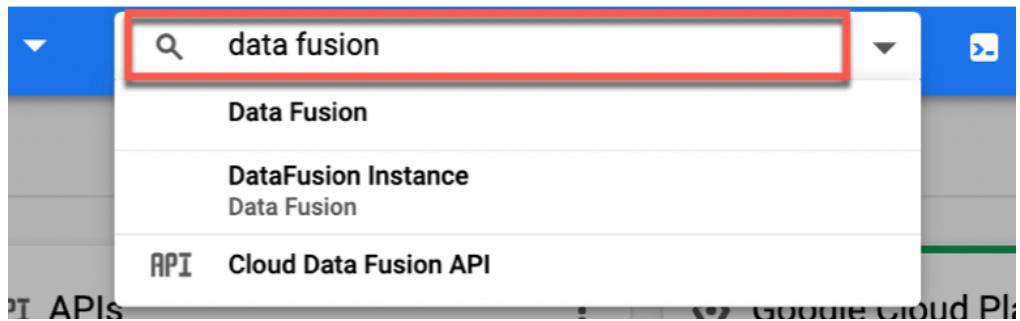
ETL PIPELINE FOR COMBINED(MERGED DATA) IN DATA FUSION

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 .

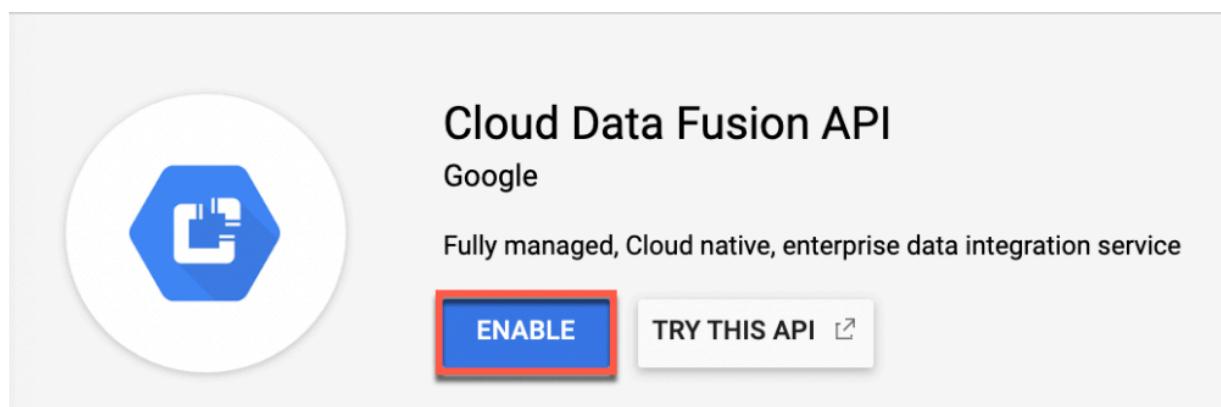
The screen shots indicating the steps (only samples included as extensive step wise process has been already indicated above) have been shown below

(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.



Instance name

flights-data-etl

1

Alphanumeric 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

Logging and monitoring

- Enable Stackdriver logging service
 Enable Stackdriver monitoring service

Labels

[+ ADD LABEL](#)

The instance creation will take approximately 20 minutes.

[CREATE](#)

[CANCEL](#)

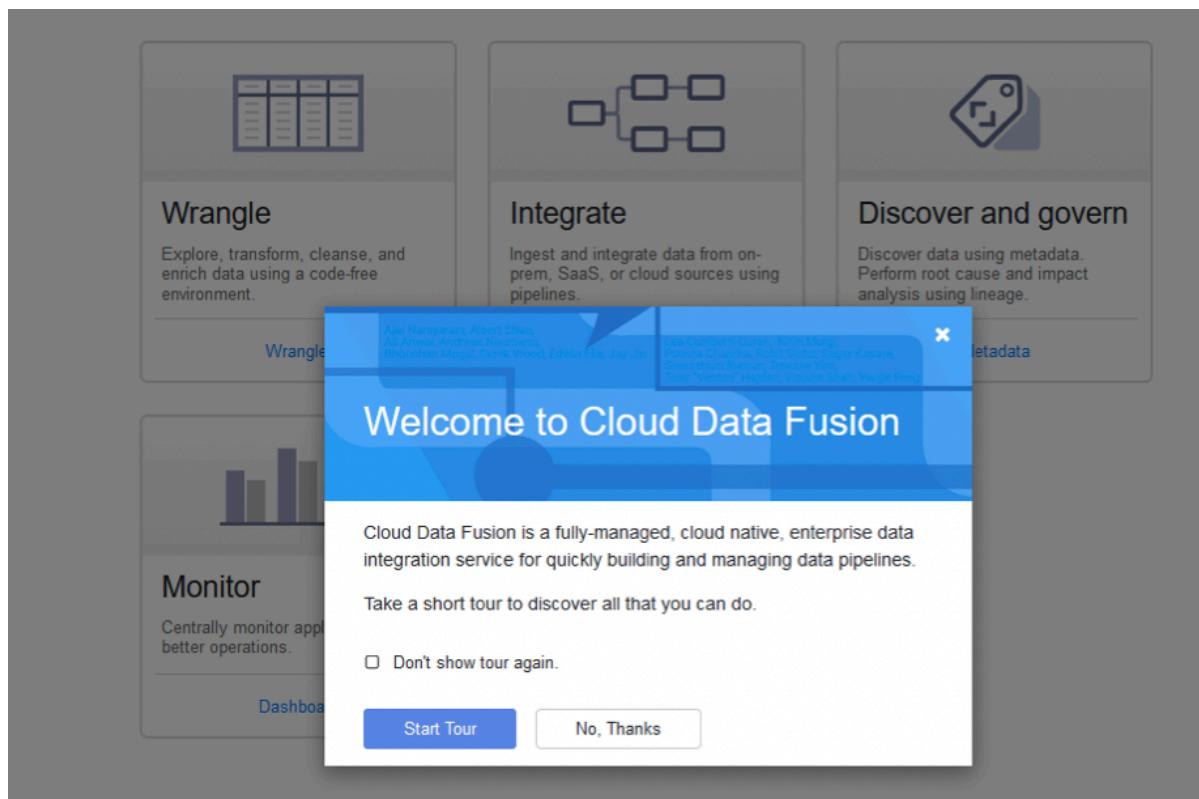
After the instance deploys successfully, a green check mark appears. Click *View Instance* to continue.



✓ flights-data-etl

Instance ID	flights-data-etl
Instance URL	View Instance
Description	--
Edition	BASIC
Region	northamerica-northeast1
Created	Nov 25, 2019, 2:12:12 PM
Last updated	Nov 25, 2019, 2:25:59 PM
Stackdriver logs	Disabled
Stackdriver monitoring	Disabled
Service Account	cloud-datafusion-management-sa@x8396b757c93a97b7-tp.iam.gserviceaccount.com
Version	6.1.0.3
Private IP	Disabled

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.

The screenshot shows the 'Cloud Storage Default' interface. On the left, there's a sidebar with 'Connections in "default"' containing 'Upload', 'BigQuery(2)', 'Database(0)', and 'GCS(2)'. Below this is a section titled 'Cloud Storage Default' with three dots. On the right, a 'Select Data' panel shows 'Root / de-project-bucket' with three files: 'current-data.csv', 'hist-data.csv', and 'merged-data.csv'.

The screenshot shows a 'Parse as CSV' dialog box over a main interface. The dialog title is 'Parse as CSV' and it says 'Please select the delimiter'. It has a radio button for 'Comma' (selected), 'Tab', 'Space', 'Pipe', '^A', '^D', and 'Custom delimiter'. There's also a checked checkbox for 'Set first row as header'. At the bottom are 'Apply' and 'Cancel' buttons.

Now we can carry out transformation on the data set.

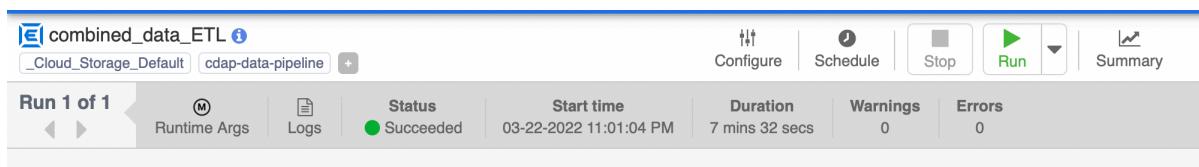
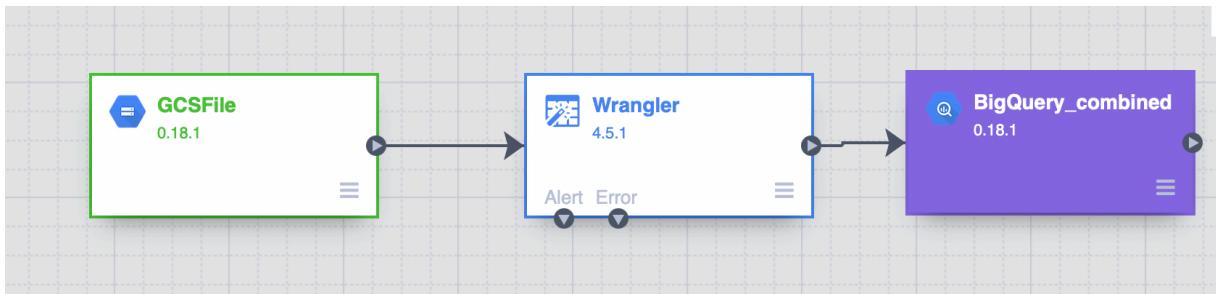
(a) Delete body column and other not relevant columns

(B) parse lastUpdatedDate 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.