Google Pipeline

Suppose we are tasked to create an ETL Data Pipeline on Google Cloud with Airflow or Cloud Data Fusion, to extract some employee data for example from various sources, mask sensitive information within the data, and load it into BigQuery. It will require us to build and configure a data engineering infrastructure in google cloud platform.

Requirements

- **Data Extraction** from different sources such as databases, CSV files or APIs.
- **Data Masking (Transformation)** is done when sensitive information is identified
- **Data Loading into BigQuery** is after identified data is extracted and masked

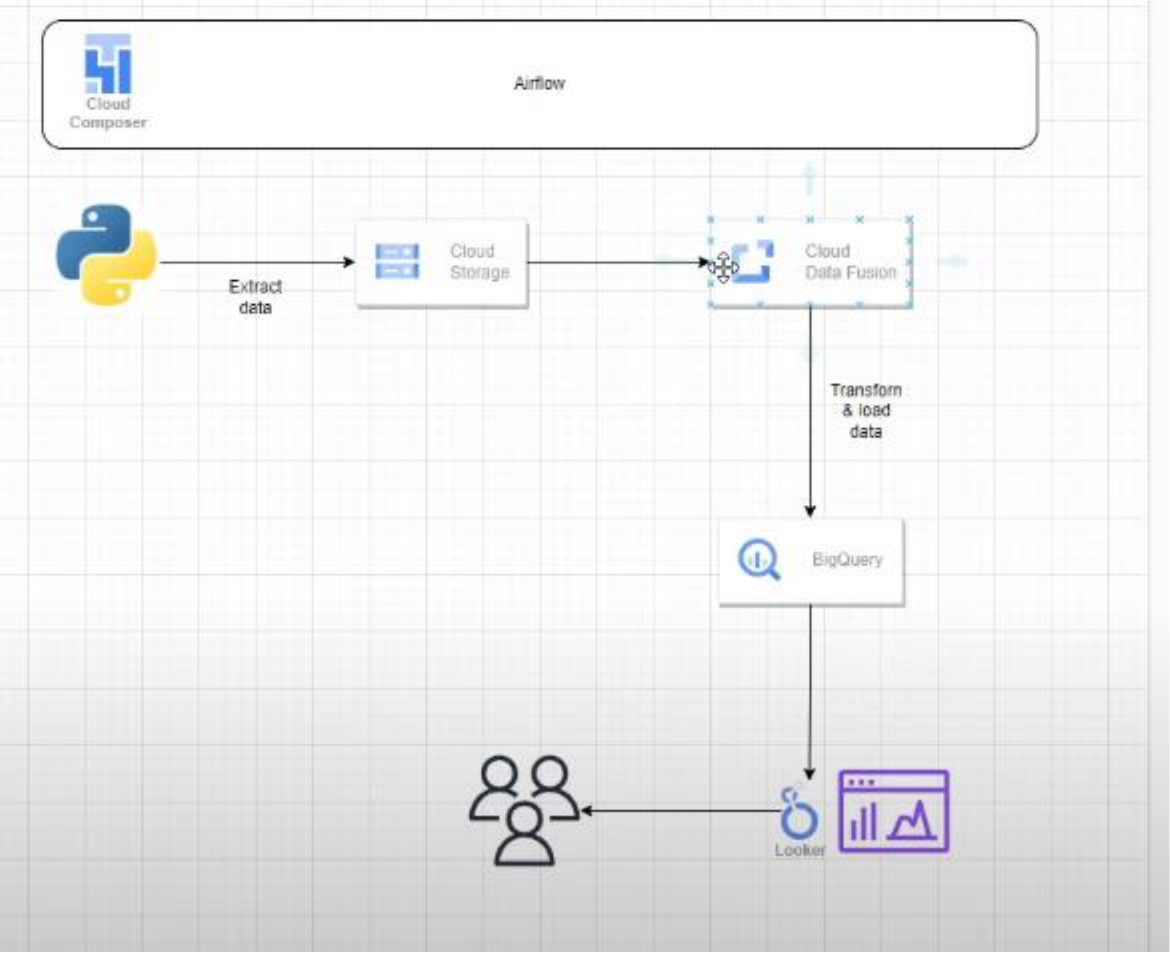
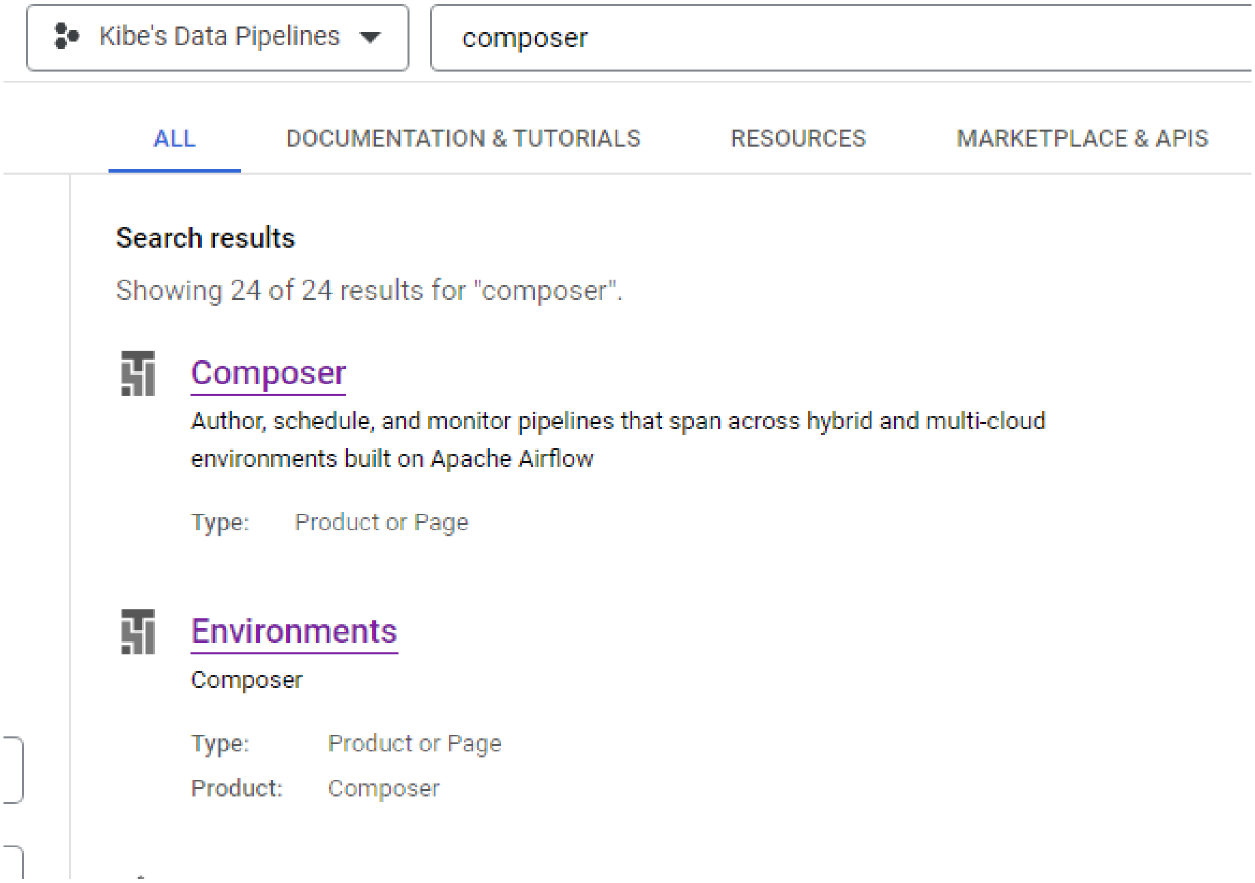


Figure 1: Data engineering infrastructure

This work flow is supposed to inform our Google pipeline project, following below steps

Step 1: Create composer environment



Search results

Showing 24 of 24 results for "composer".

Composer

Author, schedule, and monitor pipelines that span across hybrid and multi-cloud environments built on Apache Airflow

Type: Product or Page

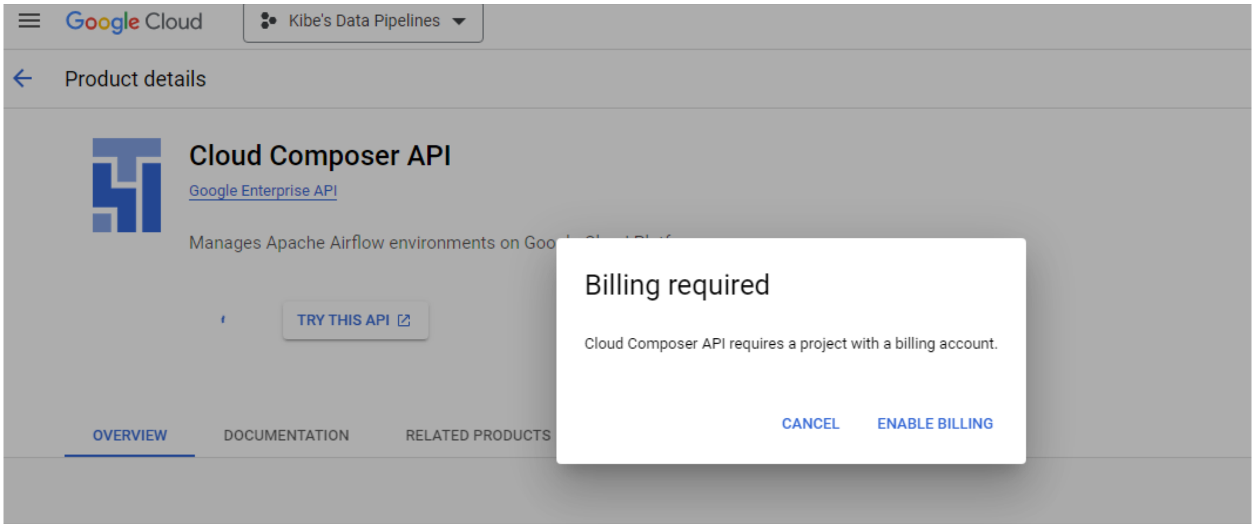
Environments

Composer

Type: Product or Page

Product: Composer

Error



Cloud Composer API

Google Enterprise API

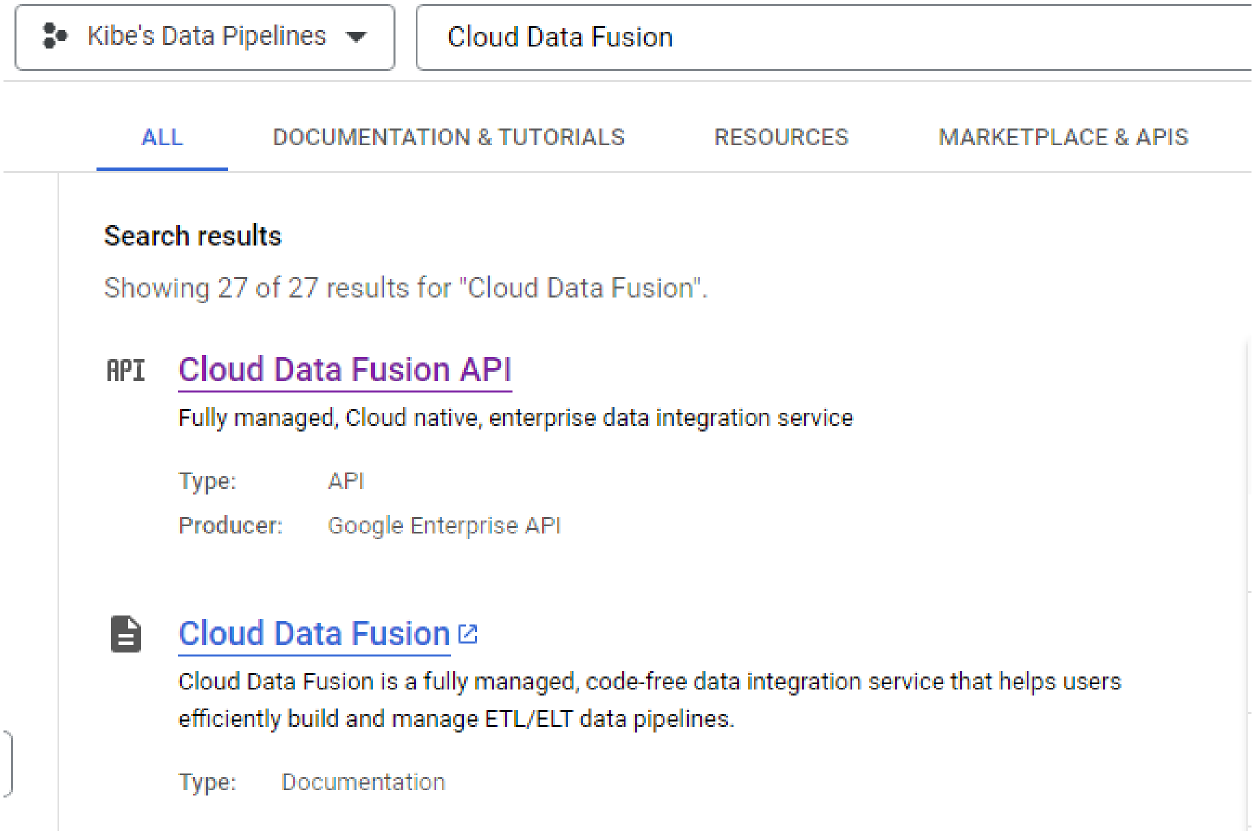
Manages Apache Airflow environments on Goo

Billing required

Cloud Composer API requires a project with a billing account.

We would have been able to create a composer environment which normally takes some 20-30 minutes.

Step 2: Create a data fusion environment



Search results

Showing 27 of 27 results for "Cloud Data Fusion".

API **Cloud Data Fusion API**

Fully managed, Cloud native, enterprise data integration service

Type: API

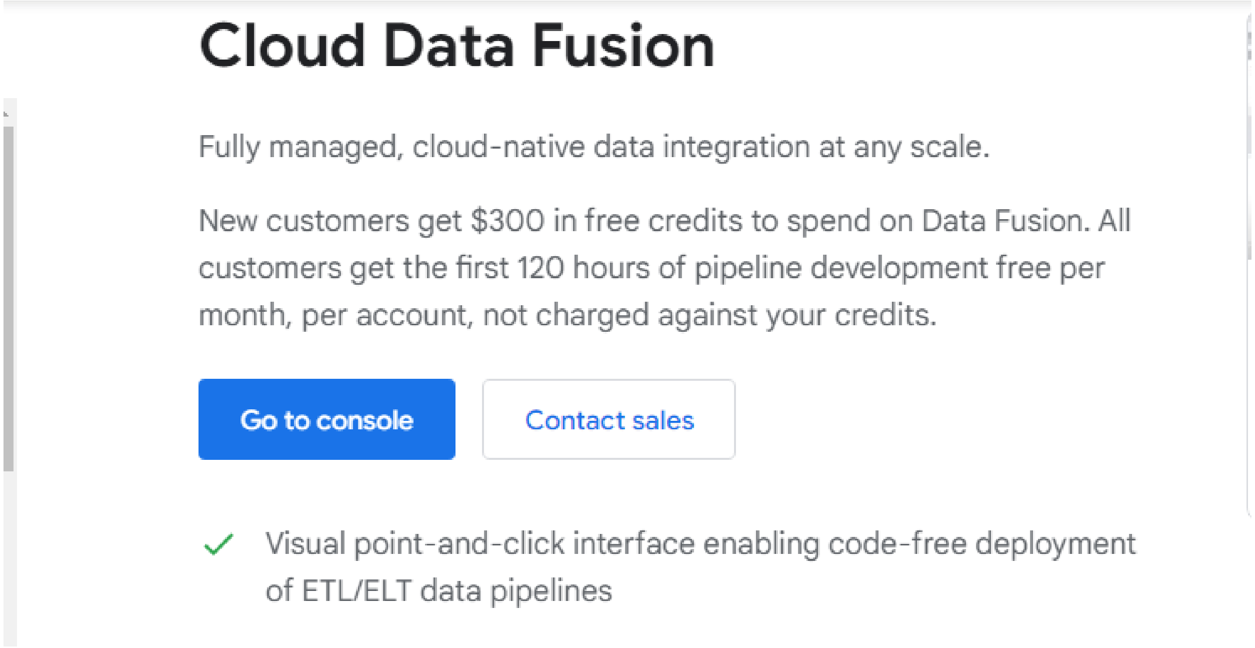
Producer: Google Enterprise API

Cloud Data Fusion

Cloud Data Fusion is a fully managed, code-free data integration service that helps users efficiently build and manage ETL/ELT data pipelines.

Type: Documentation

Option 1



Cloud Data Fusion

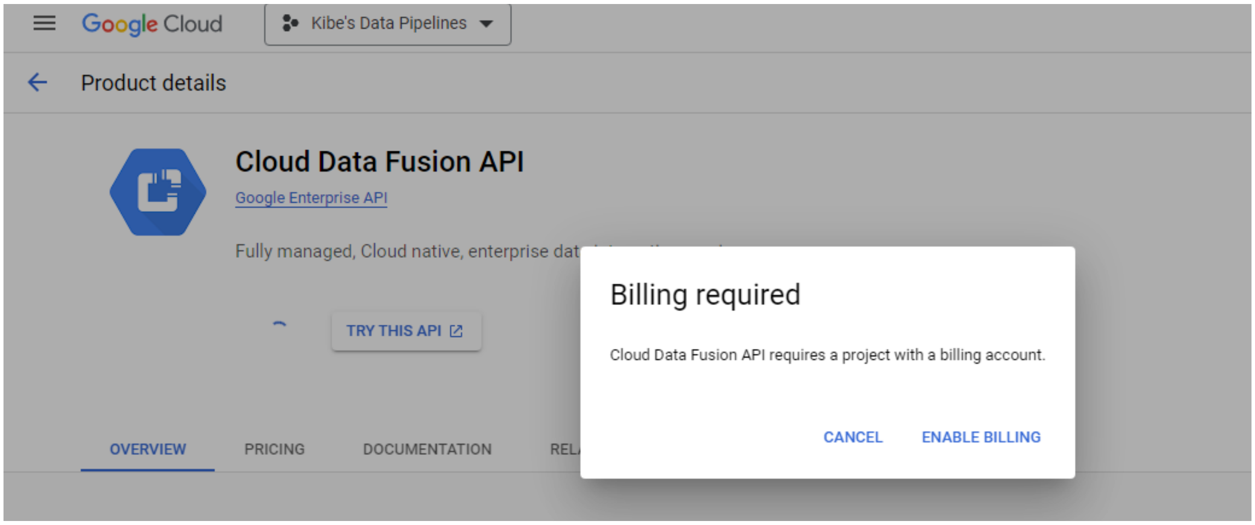
Fully managed, cloud-native data integration at any scale.

New customers get $300 in free credits to spend on Data Fusion. All customers get the first 120 hours of pipeline development free per month, per account, not charged against your credits.

✓ Visual point-and-click interface enabling code-free deployment of ETL/ELT data pipelines

Option 2 recommended

Error



Fully managed, Cloud native, enterprise dat

Billing required

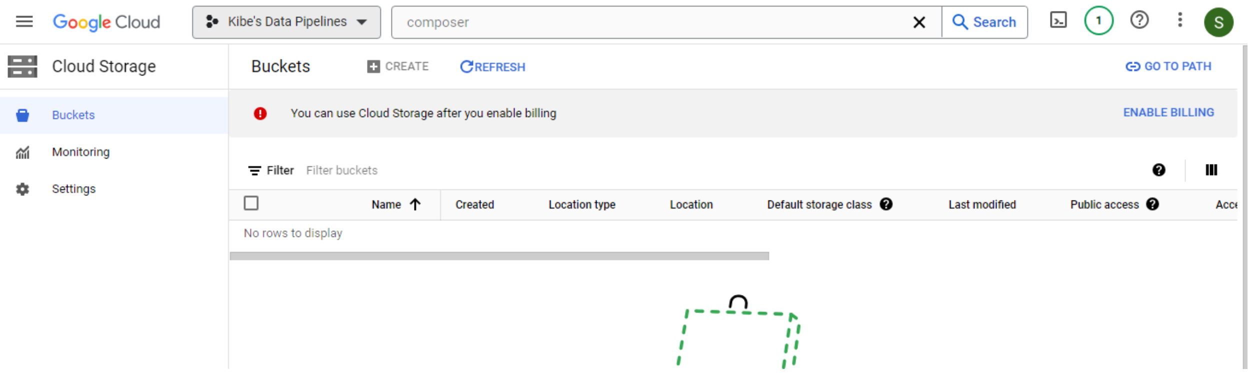
Cloud Data Fusion API requires a project with a billing account.

We would have been able to create a Data fusion Instance. Usually takes some time to create

Step 3: Creating dummy data then load it to cloud storage

Basically, a python script with faker library is used to create dummy data, which afterwards is then loaded into the cloud storage bucket, saved as a CSV file. This is easily created and configured through any IDE for example Visual code, which connects direct to a project in google platform using a Gmail account.

Error

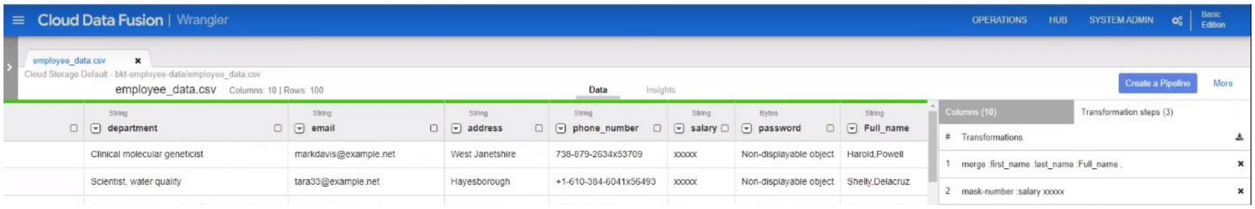


You can use Cloud Storage after you enable billing

No rows to display

Stem 4: Data pipeline

Our dummy data is now supposed to be transferred into BigQuery using a data pipeline customized in the data fusion environment. A lot of data transformation (masking) can be done in Data fusion without any coding.


Cloud Data Fusion | Wrangler
OPERATIONS
HUB
SYSTEM ADMIN
Basic Edition
employee_data.csv
Cloud Storage Default - bkt-employee-data/employee_data.csv
employee_data.csv Columns: 10 | Rows: 100
Data
Insights
Create a Pipeline
More
Columns (10)
Transformation steps (3)

String department	String email	String address	String phone_number	String salary	Bytes password	String Full_name
Clinical molecular geneticist	markdavis@example.net	West Janetshire	738-879-2634x53709	xxxxx	Non-displayable object	Harold,Powell
Scientist, water quality	tara33@example.net	Hayesborough	+1-610-384-6041x56493	xxxxx	Non-displayable object	Shelly,Delacruz

#	Transformations
1	merge :first_name :last_name :Full_name ,
2	mask-number :salary xxxxx

Figure 2: Sample Cloud Data fusion environment

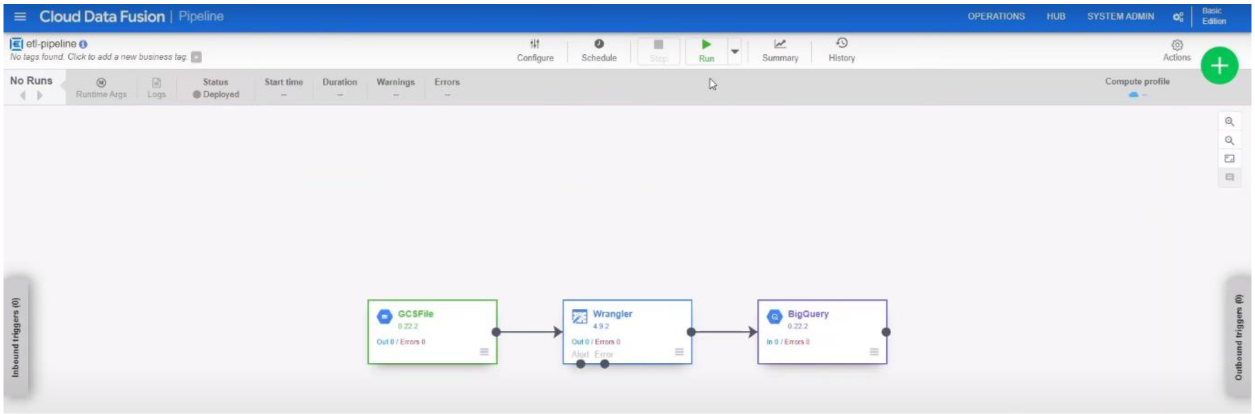
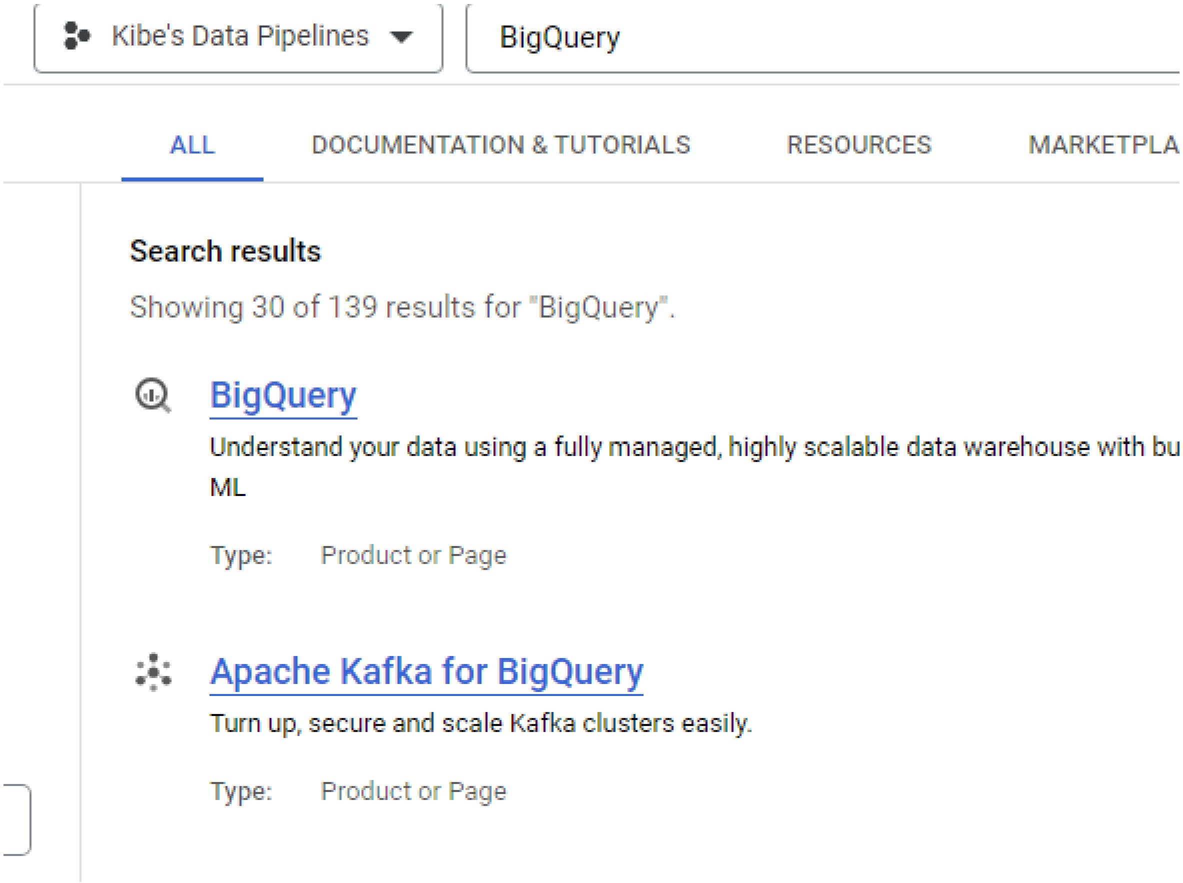


Figure 3: sample ETL-pipeline on Cloud Data Fusion

Step 4: Setup BigQuery Environment



Search results

Showing 30 of 139 results for "BigQuery".

BigQuery

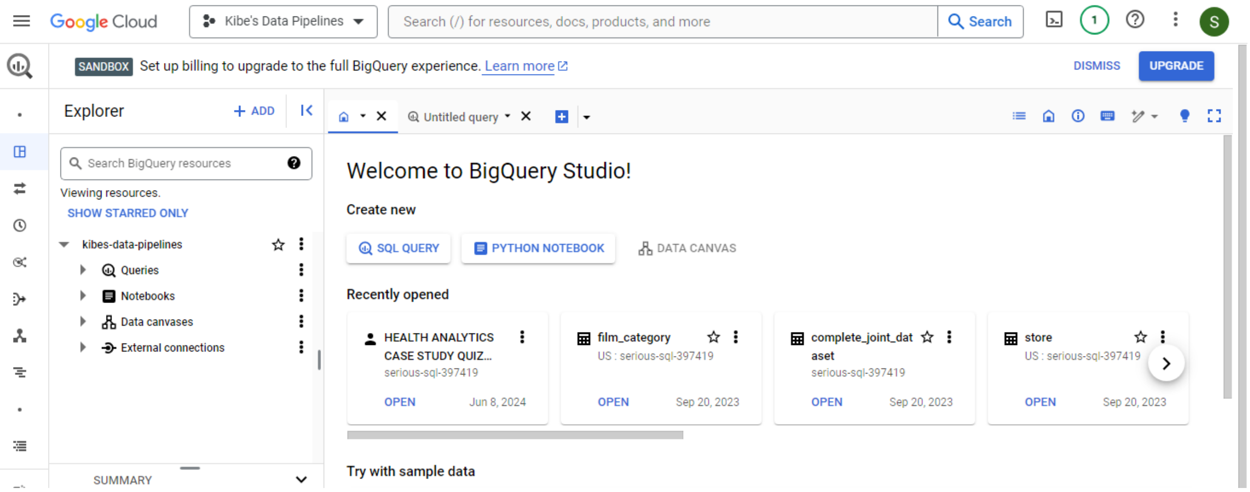
Understand your data using a fully managed, highly scalable data warehouse with bu ML

Type: Product or Page

Apache Kafka for BigQuery

Turn up, secure and scale Kafka clusters easily.

Type: Product or Page



Set up billing to upgrade to the full BigQuery experience. Learn more

Viewing resources.

Welcome to BigQuery Studio!

Create new

Recently opened

HEALTH ANALYTICS CASE STUDY QUIZ... serious-sql-397419 OPEN Jun 8, 2024	film_category US : serious-sql-397419 OPEN Sep 20, 2023	complete_joint_dataset serious-sql-397419 OPEN Sep 20, 2023	store US : serious-sql-397419 OPEN Sep 20, 2023

Try with sample data

Description

With these packages set up and running, we would now be able to perform ETL, the load the processed data into BigQuery using google pipelining as illustrated by our data flow on the first slide.