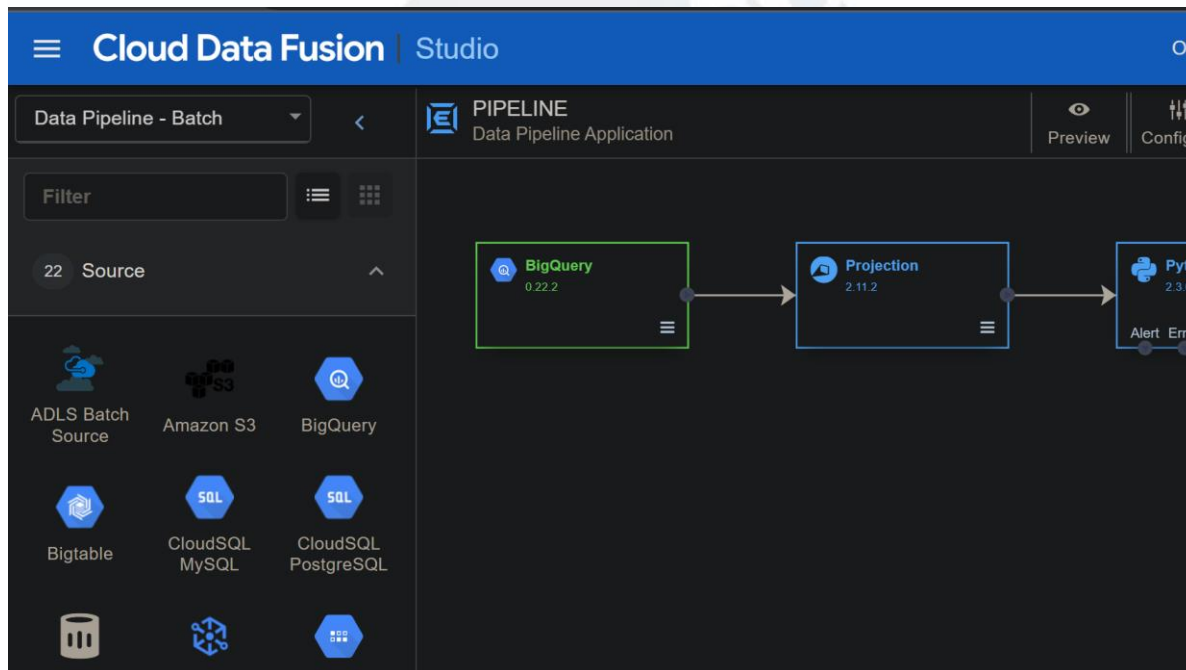

Big Data on GCP

Data Fusion como herramienta de desarrollo visual



Nos permite dibujar **flujos de procesamiento visual**, el flujo al ejecutarse distribuye su carga de trabajo sobre un clúster de SPARK

Composer como herramienta de orquestación

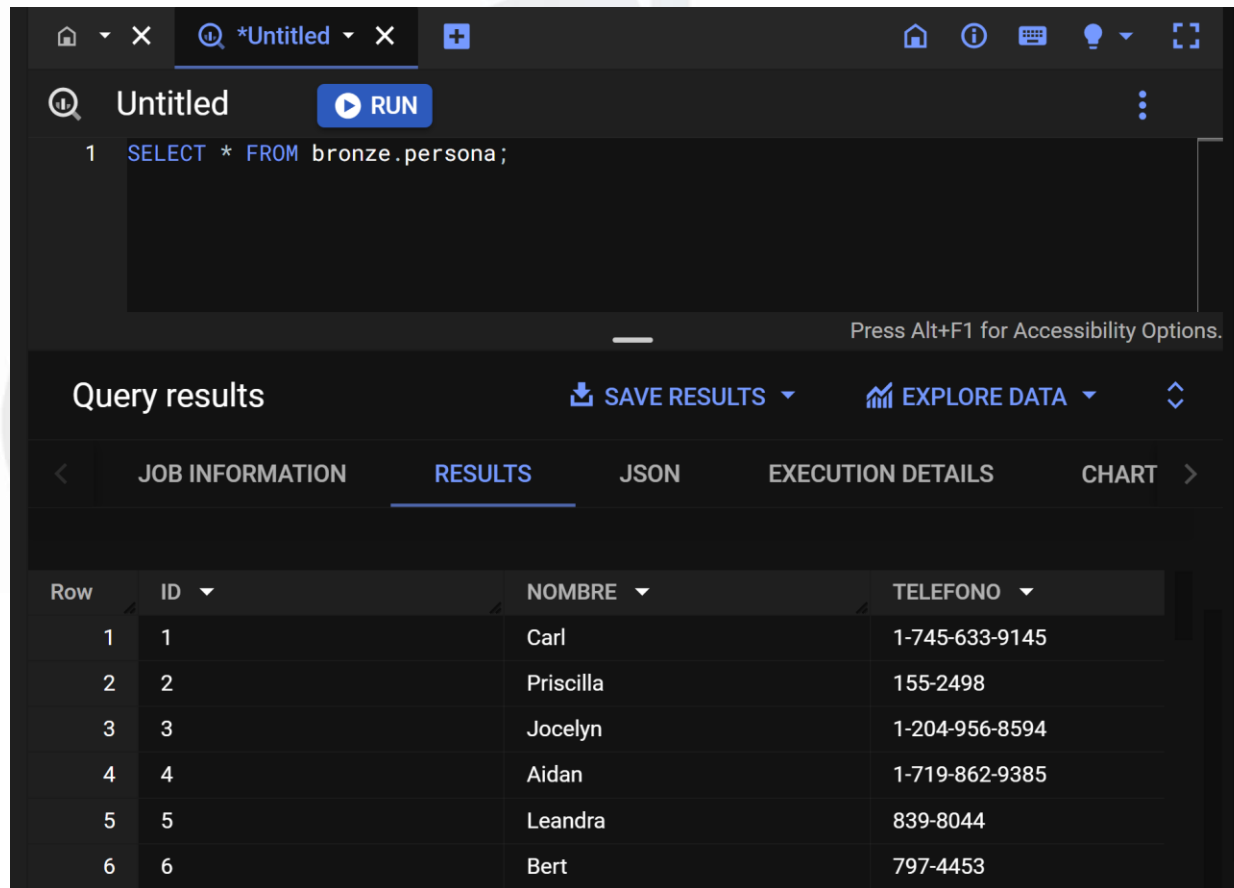
```
#Ejecución del preceso TO_SILVER
TO_SILVER = CloudDataFusionStartPipelineOperator(
    location = "us-west1", #Región donde vive la instancia
    instance_name = "serveret1", #Nombre de la instancia
    pipeline_name = "TO_SILVER", #Nombre del JOB de DATA FUSION
    task_id = "TO_SILVER" #Identificador del paso, general
)

# Segundo paso
TO_GOLD = CloudDataFusionStartPipelineOperator(
    location = "us-west1", #Región donde vive la instancia
    instance_name = "serveret1", #Nombre de la instancia
    pipeline_name = "TO_GOLD", #Nombre del JOB de DATA FUSION
    task_id = "TO_GOLD" #Identificador del paso, general
)

#Definimos el orden de ejecución
TO_SILVER >> TO_GOLD
```

Composer permite orquestar procesos. La orquestación se hace con código Python y con las librerías de Airflow, el cual es una herramienta open-source para orquestación

BigQuery como herramienta de tablas



The screenshot shows the BigQuery web interface. At the top, there's a toolbar with a home icon, a search icon, and a tab labeled '*Untitled'. Below the toolbar, the query editor shows a single line of SQL: `1 SELECT * FROM bronze.persona;`. To the right of the query is a blue 'RUN' button. Below the query editor, there's a section for 'Query results' with options to 'SAVE RESULTS' and 'EXPLORE DATA'. Below this, there are tabs for 'JOB INFORMATION', 'RESULTS' (which is selected), 'JSON', 'EXECUTION DETAILS', and 'CHART'. The 'RESULTS' tab displays a table with 6 rows and 4 columns: 'Row', 'ID', 'NOMBRE', and 'TELEFONO'. The data is as follows:

Row	ID	NOMBRE	TELEFONO
1	1	Carl	1-745-633-9145
2	2	Priscilla	155-2498
3	3	Jocelyn	1-204-956-8594
4	4	Aidan	1-719-862-9385
5	5	Leandra	839-8044
6	6	Bert	797-4453

BigQuery nos permite ver a los **archivos estructurados como tablas** de bases de datos

Dataproc para clústers de Spark

	Profile name	Provisioner
✓	Autoscaling Dataproc	Dataproc
★	Dataproc	Dataproc

Es el servicio que permite **crear clústers de Big Data** auto-escalables, es decir auto-define el tamaño en función de la volumetría