DE Final Project

Shanistha Jagannath Shetty

Task at Hand

A large IT company is losing its employees at a particular location recently and management would like to closely monitor the HC and Attrition of this location.

We have been tasked to set up an automated dashboard for the management. HR data will be provided by HR Talent team on a regular basis. We would like to set up a workflow management to ensure that dashboard is automatically updated.

To achieve this, we have set up Airflow running on Dockers. Workflow will update the data in Google BigQuery table which in turn will be connected to a Looker Studio Report Dashboard providing overview of HC, Attrition and some Deep Dives into Attrition reasons.

Resources Required

Dockers, Python, Airflow, Google Cloud Storage, BigQuery and Looker Studio

Workflow Management in Airflow

Task 1 > BashOperator – Download csv file from a server

Task 2 > PythonOperator - Perform ETL and save the file

Task 3 > PythonOperator – Upload the file in Google Cloud Storage Bucket

Task 4 > PythonOperator – Check for table in BigQuery database. If does not exist then create it.

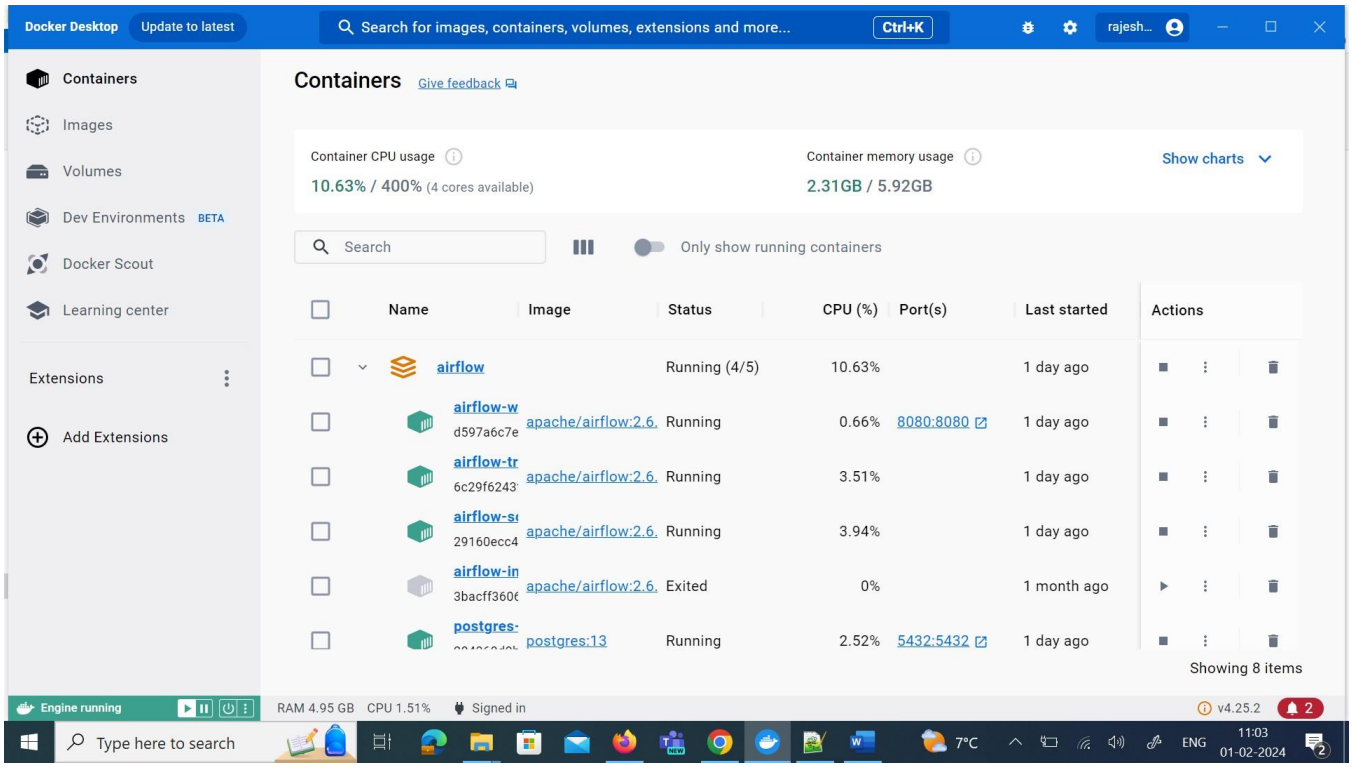
Task 5 > Load the data from Google Storage Bucket to BigQuery Table

Task in Google Cloud

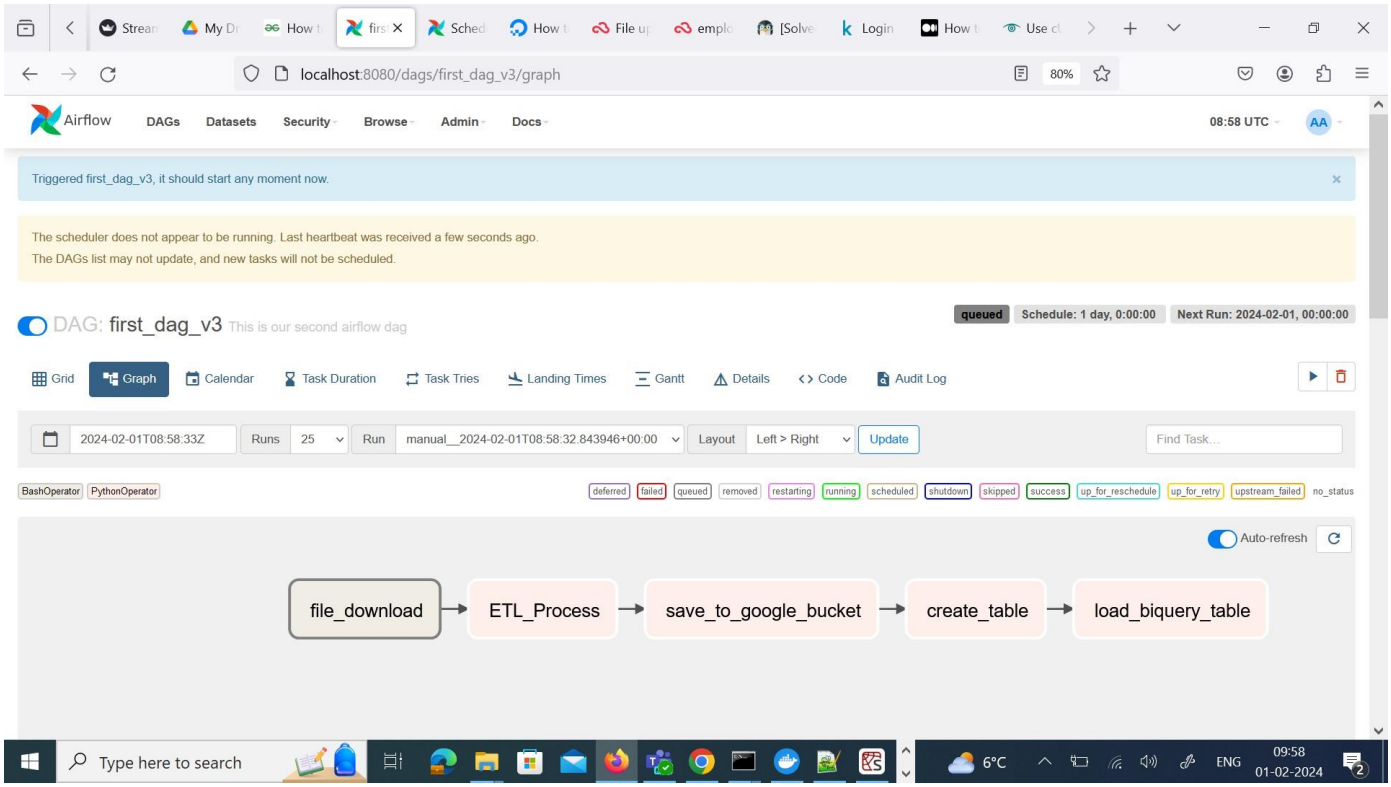
Set up Terraform in Google Cloud Terminal to create Storage Bucket

Create Looker Studio Report based on the requirements of the management to show HC and Attrition Analysis

Airflow set up and running on Dockers



Airflow running – Dag queued



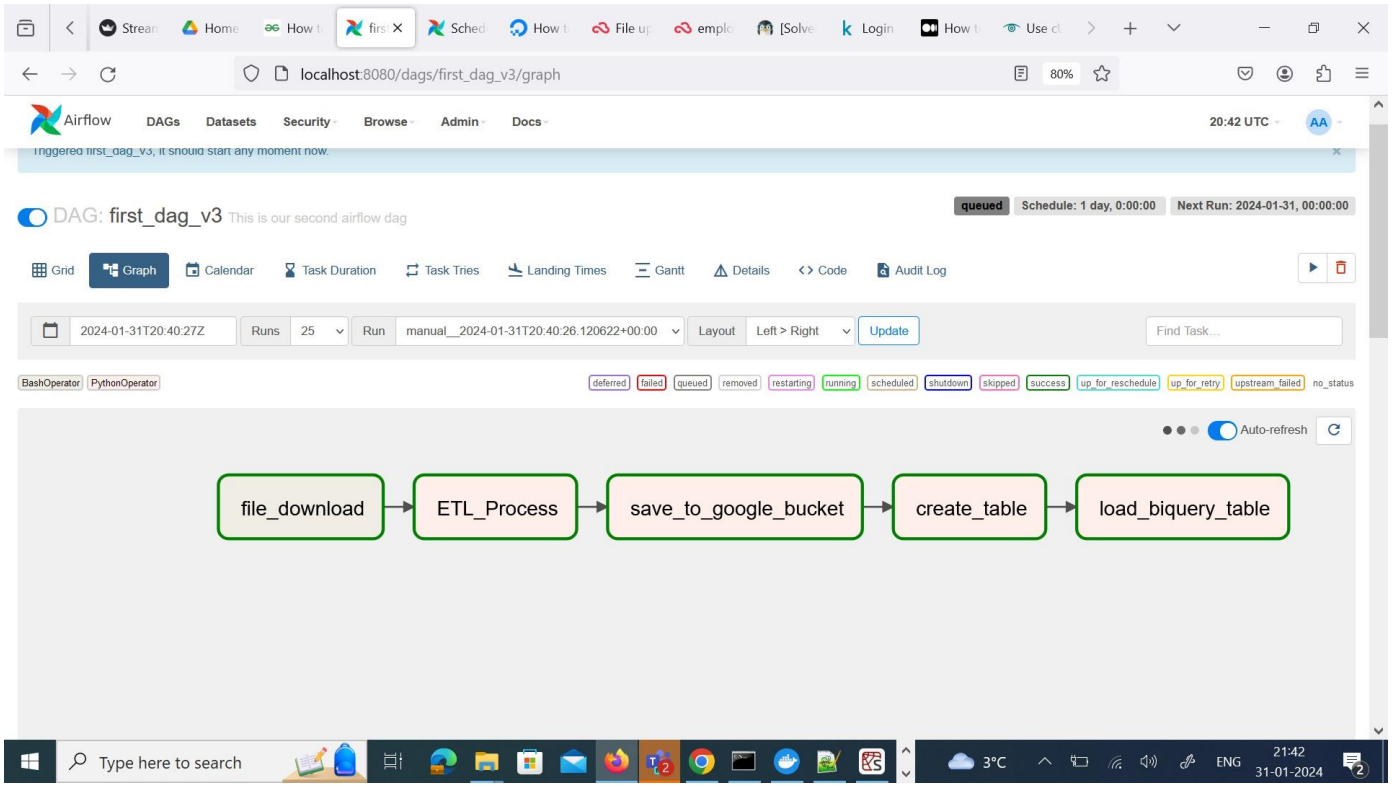
First task is to download the hr data csv file hosted on an intranet file server and rename it to current date. For example

20240201.csv.

ETL is performed on this file in next step, one field is changed from categorical to numerical required for analysis. Extra columns which are not required are dropped and a new file is saved.

For example employee_hr_data20240201.csv.

Subsequently the next task continues.



Bucket Created on Google Cloud using Terraform. This file was saved main.tf was run on Google Cloud Terminal

```
provider "google"{
credentials=file("access-keys.json")
project=var.project_id
}
variable "bucket_name"{
type=string
description="Bucket created using terraform"
}
variable "bucket_location"{
type=string
default="us-east1"
}
variable "project_id"{
type=string
}
variable "storage_class"{
```

type=string

}

resource "google_storage_bucket" "default"{

name=var.bucket_name

storage_class=var.storage_class

location=var.bucket_location

}

Airflow runs and uploads the new created csv file after the ETL process and it is then uploaded in the designated Bucket

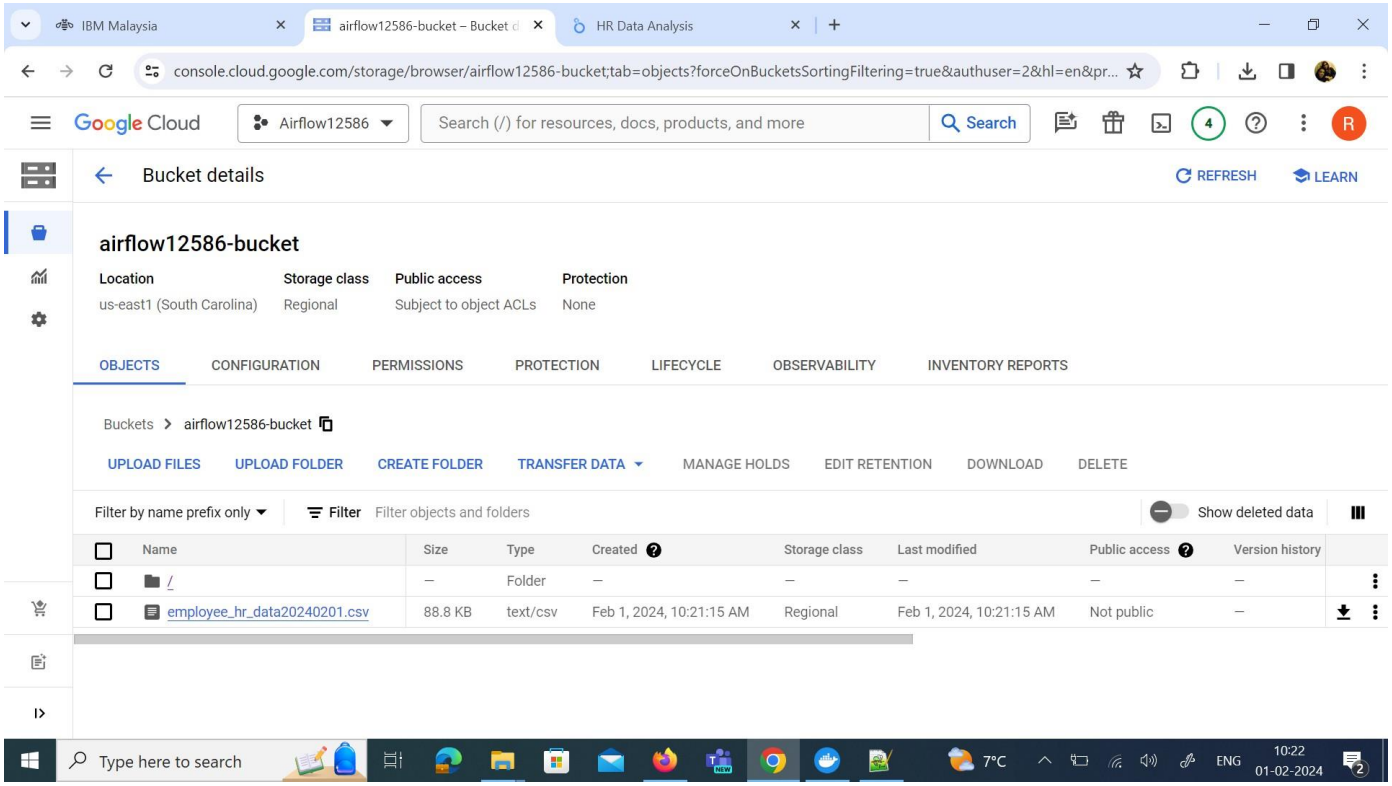
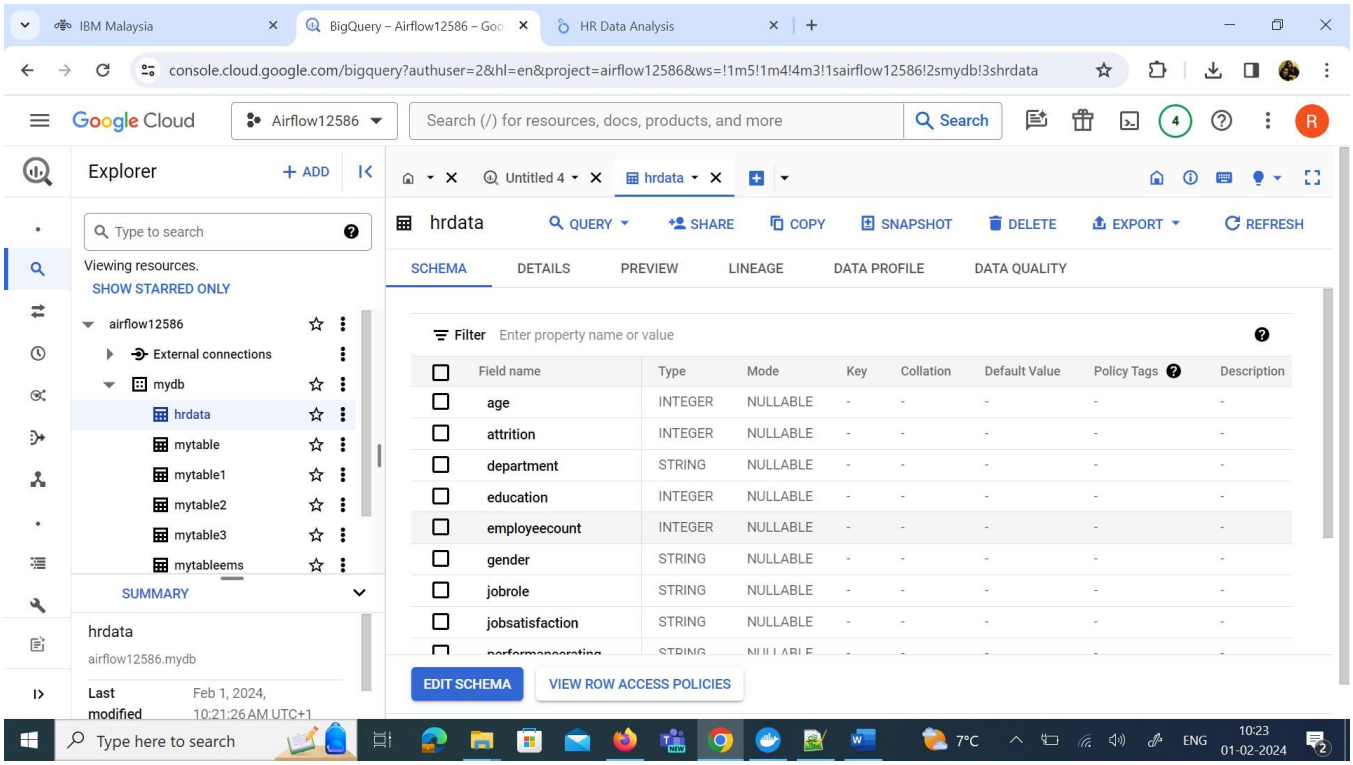
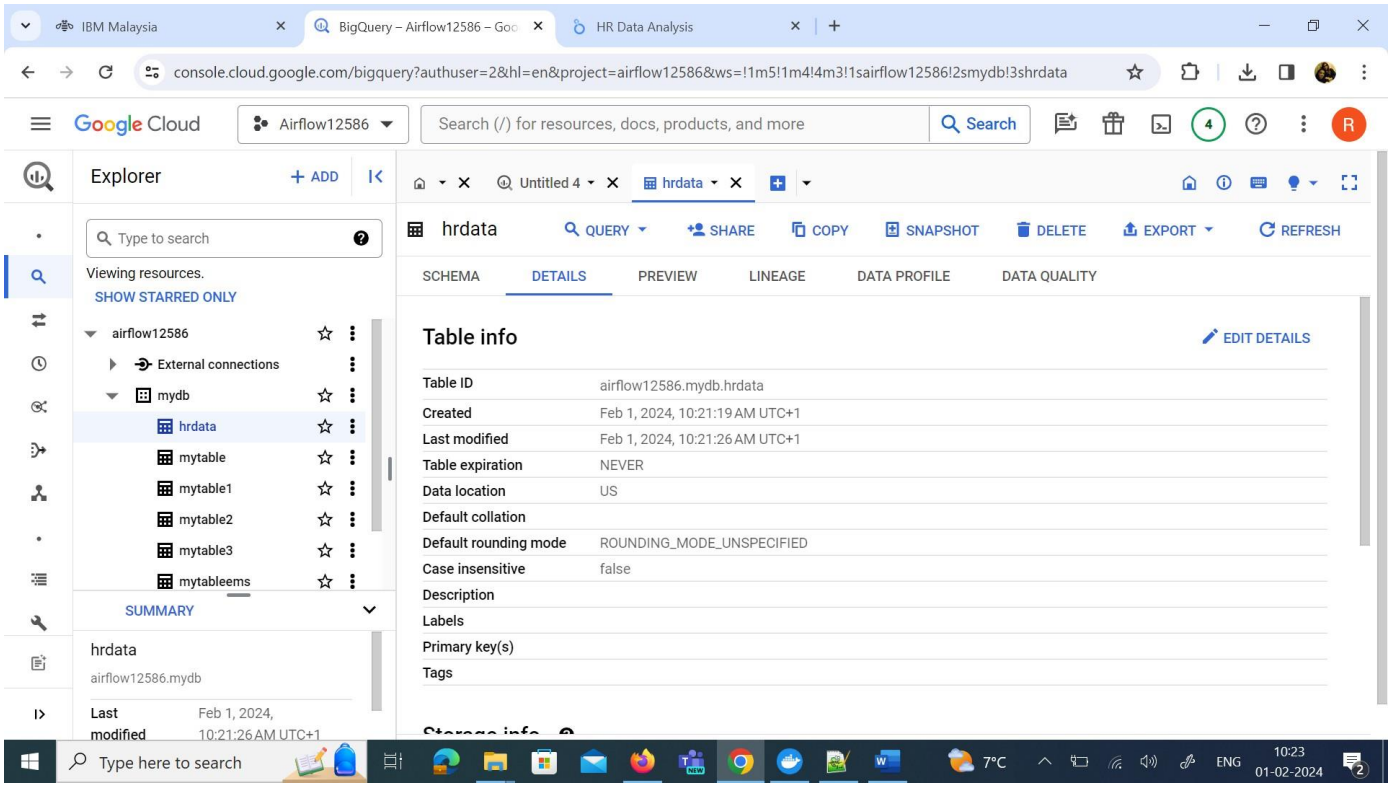
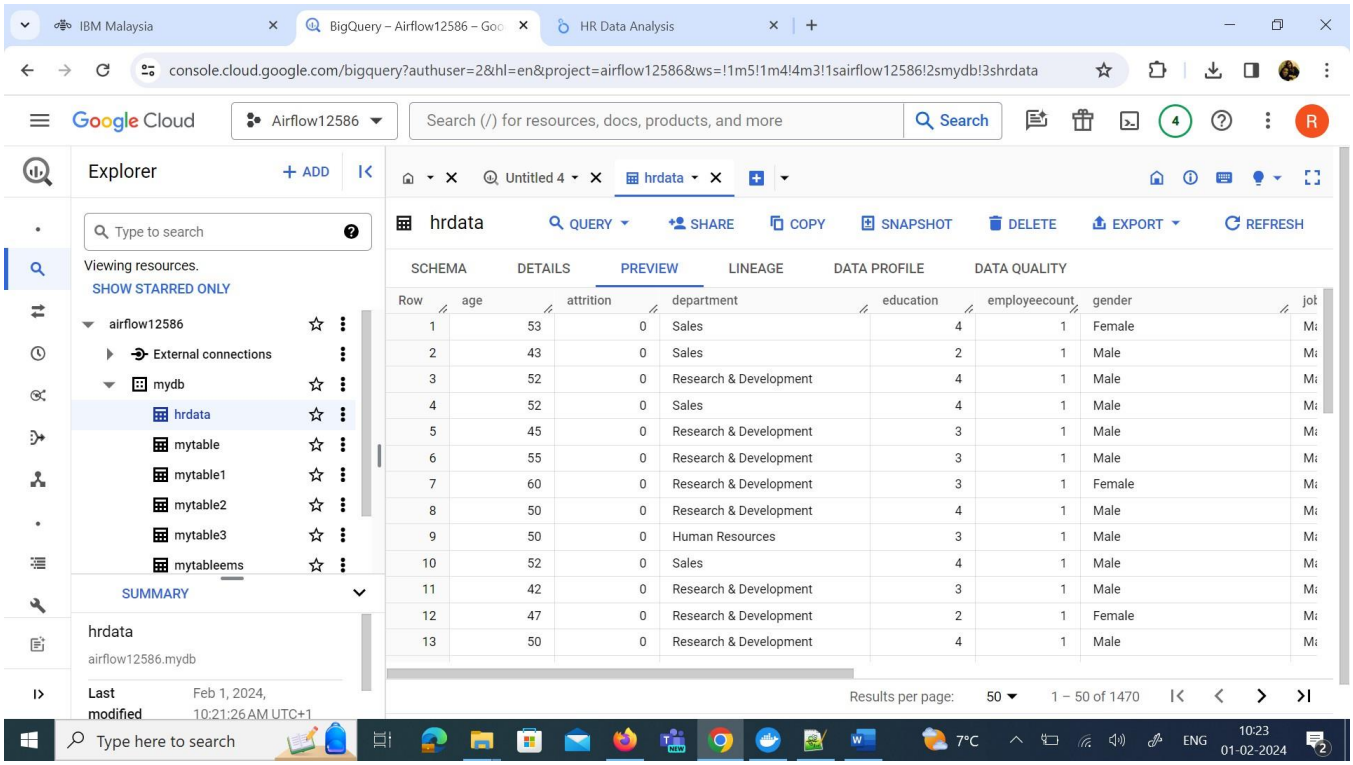


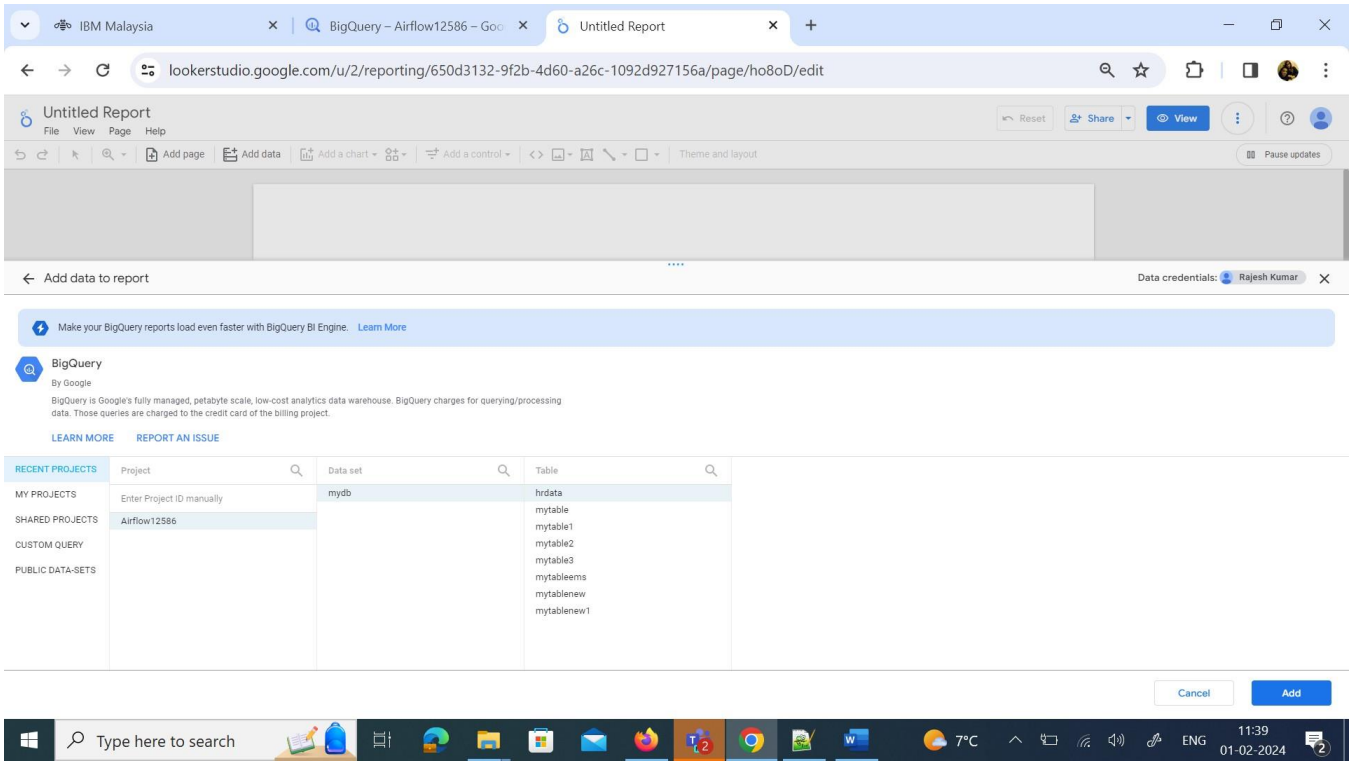
Table is created from Airflow into the Bigquery Database

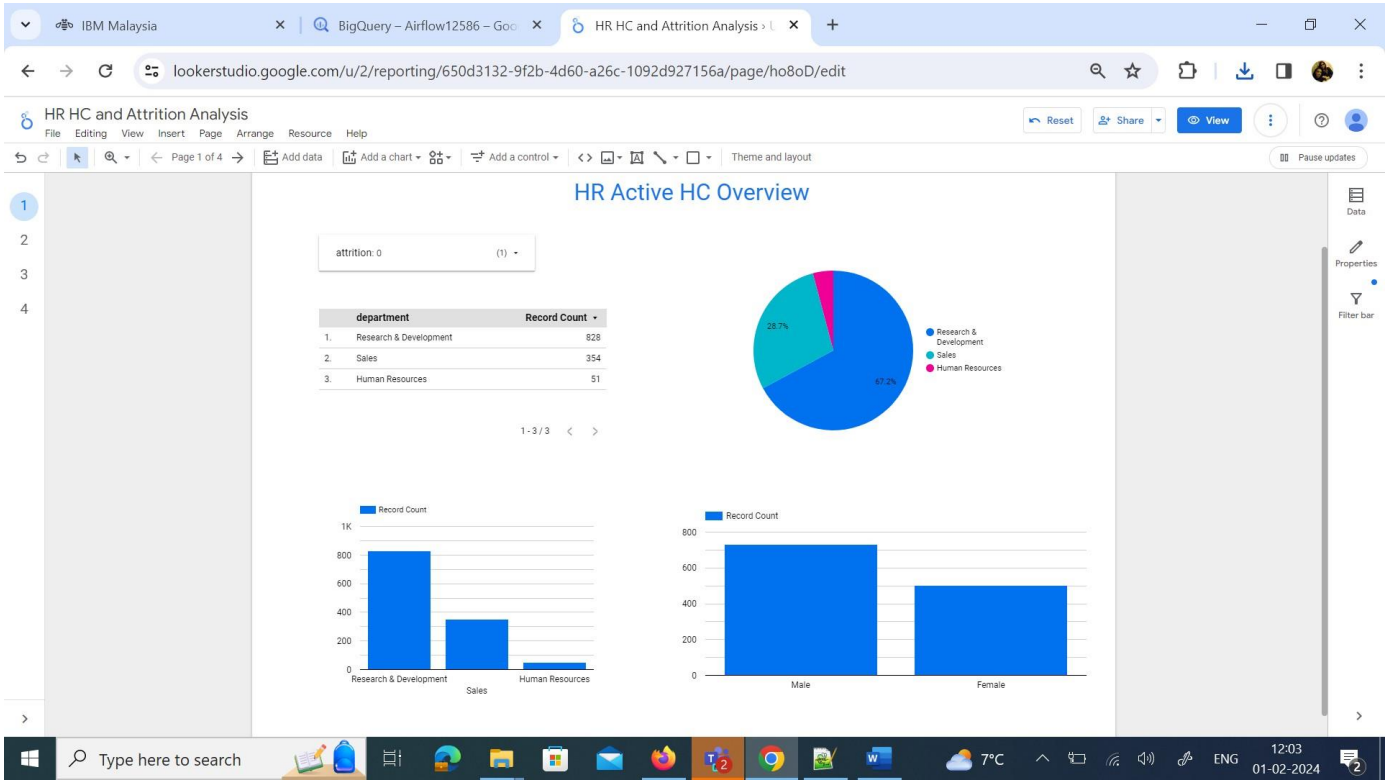


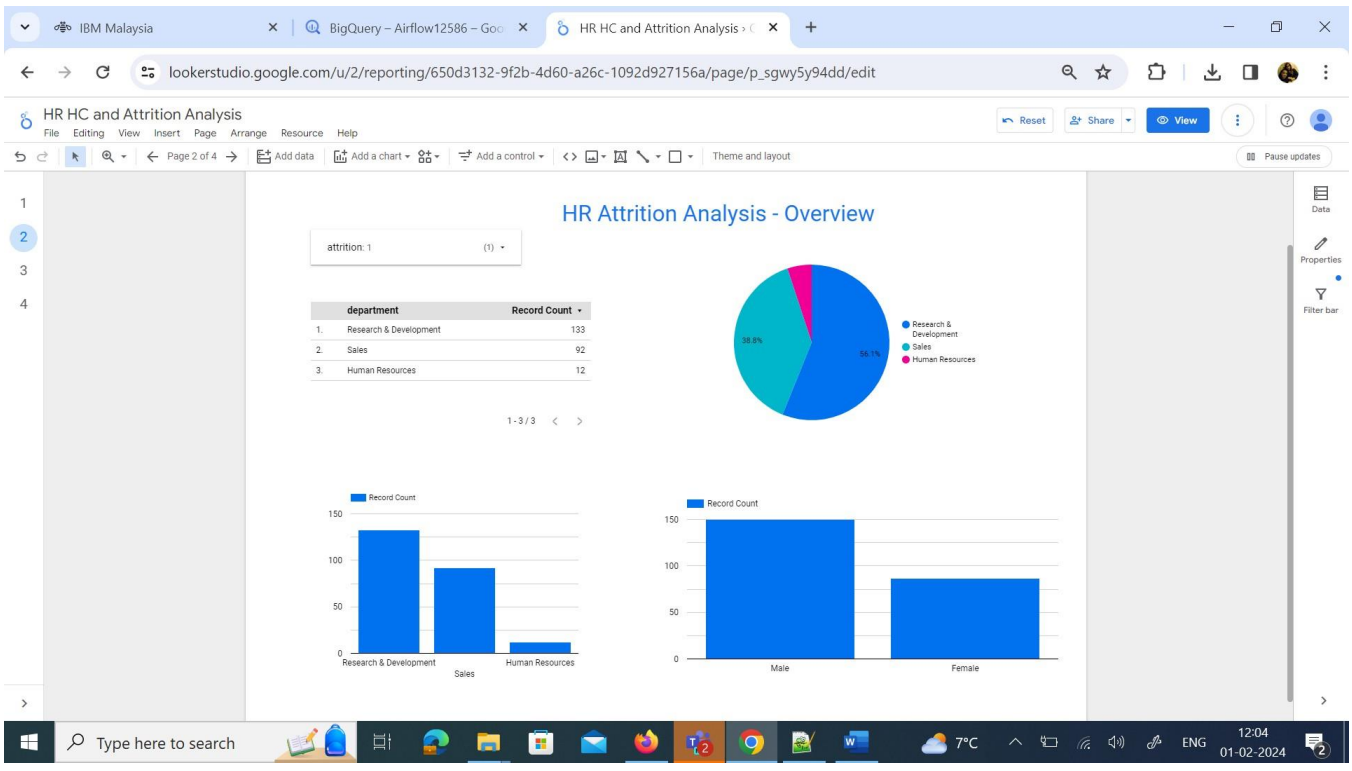


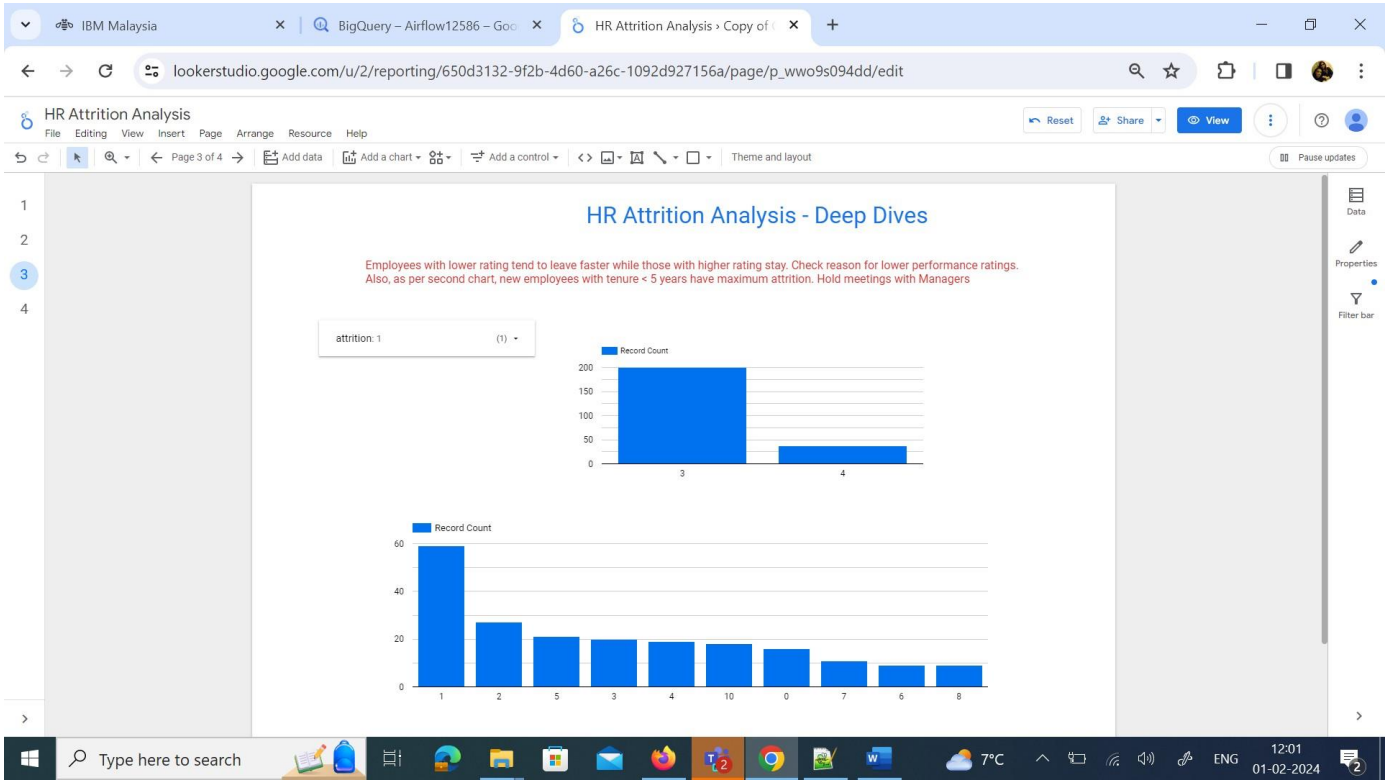
Data loaded from the csv file in the Bucket into the the BigQuery table.


IBM Malaysia
BigQuery – Airflow12586 – Goo
HR Data Analysis
console.cloud.google.com/bigquery?authuser=2&hl=en&project=airflow12586&ws=!1m5!1m4!4m3!1sairflow12586!2smydb!3shrdata
Google Cloud
Airflow12586
Search (/) for resources, docs, products, and more
Search
Explorer
+ ADD
Type to search
Viewing resources.
SHOW STARRED ONLY
airflow12586
External connections
mydb
hrdata
mytable
mytable1
mytable2
mytable3
mytableems
SUMMARY
hrdata
airflow12586.mydb
Last modified
Feb 1, 2024, 10:21:26 AM UTC+1
Untitled 4
hrdata
QUERY
SHARE
COPY
SNAPSHOT
DELETE
EXPORT
REFRESH
SCHEMA
DETAILS
PREVIEW
LINEAGE
DATA PROFILE
DATA QUALITY
Row age attrition department education employeecount gender job
1 53 0 Sales 4 1 Female
2 43 0 Sales 2 1 Male
3 52 0 Research & Development 4 1 Male
4 52 0 Sales 4 1 Male
5 45 0 Research & Development 3 1 Male
6 55 0 Research & Development 3 1 Male
7 60 0 Research & Development 3 1 Female
8 50 0 Research & Development 4 1 Male
9 50 0 Human Resources 3 1 Male
10 52 0 Sales 4 1 Male
11 42 0 Research & Development 3 1 Male
12 47 0 Research & Development 2 1 Female
13 50 0 Research & Development 4 1 Male
Results per page: 50
1 – 50 of 1470
Type here to search
7°C
ENG
10:23
01-02-2024


IBM Malaysia
BigQuery – Airflow12586 – Goo
Untitled Report
lookerstudio.google.com/u/2/reporting/650d3132-9f2b-4d60-a26c-1092d927156a/page/ho8oD/edit
Untitled Report
File View Page Help
Reset
Share
View
Add page
Add data
Add a chart
Add a control
Theme and layout
Pause updates
Add data to report
Data credentials: Rajesh Kumar
Make your BigQuery reports load even faster with BigQuery BI Engine. Learn More
BigQuery
By Google
BigQuery is Google's fully managed, petabyte scale, low-cost analytics data warehouse. BigQuery charges for querying/processing data. Those queries are charged to the credit card of the billing project.
LEARN MORE
REPORT AN ISSUE
RECENT PROJECTS
MY PROJECTS
SHARED PROJECTS
CUSTOM QUERY
PUBLIC DATA-SETS
Project
Enter Project ID manually
Airflow12586
Data set
mydb
Table
hrdata
mytable
mytable1
mytable2
mytable3
mytableems
mytablenew
mytablenew1
Cancel
Add
Type here to search
7°C
ENG
11:39
01-02-2024


IBM Malaysia
BigQuery – Airflow12586 – Goo
HR HC and Attrition Analysis
lookerstudio.google.com/u/2/reporting/650d3132-9f2b-4d60-a26c-1092d927156a/page/ho8oD/edit
HR HC and Attrition Analysis
File Editing View Insert Page Arrange Resource Help
Reset
Share
View
Page 1 of 4
Add data
Add a chart
Add a control
Theme and layout
Pause updates
HR Active HC Overview
attrition: 0
(1)
department
Record Count
1. Research & Development 828
2. Sales 354
3. Human Resources 51
1 - 3 / 3
28.7%
67.2%
Research & Development
Sales
Human Resources
Record Count
1K
800
600
400
200
0
Research & Development
Sales
Human Resources
Record Count
800
600
400
200
0
Male
Female
Data
Properties
Filter bar
Type here to search
7°C
ENG
12:03
01-02-2024


IBM Malaysia
BigQuery – Airflow12586 – Goo
HR HC and Attrition Analysis
lookerstudio.google.com/u/2/reporting/650d3132-9f2b-4d60-a26c-1092d927156a/page/p_sgwy5y94dd/edit
HR HC and Attrition Analysis
File Editing View Insert Page Arrange Resource Help
Reset
Share
View
Page 2 of 4
Add data
Add a chart
Add a control
Theme and layout
Pause updates
HR Attrition Analysis - Overview
attrition: 1
(1)
department
Record Count
1. Research & Development 133
2. Sales 92
3. Human Resources 12
1 - 3 / 3
38.8%
55.1%
Research & Development
Sales
Human Resources
Record Count
150
100
50
0
Research & Development
Sales
Human Resources
Record Count
150
100
50
0
Male
Female
Data
Properties
Filter bar
Type here to search
7°C
ENG
12:04
01-02-2024


IBM Malaysia
BigQuery – Airflow12586 – Goo
HR Attrition Analysis › Copy of
lookerstudio.google.com/u/2/reporting/650d3132-9f2b-4d60-a26c-1092d927156a/page/p_wwo9s094dd/edit
HR Attrition Analysis
File Editing View Insert Page Arrange Resource Help
Reset
Share
View
Page 3 of 4
Add data
Add a chart
Add a control
Theme and layout
Pause updates
HR Attrition Analysis - Deep Dives
Employees with lower rating tend to leave faster while those with higher rating stay. Check reason for lower performance ratings.
Also, as per second chart, new employees with tenure < 5 years have maximum attrition. Hold meetings with Managers
attrition: 1
(1)
Record Count
200
150
100
50
0
3
4
Record Count
60
40
20
0
1
2
5
3
4
10
0
7
6
8
Data
Properties
Filter bar
Type here to search
7°C
ENG
12:01
01-02-2024