

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

The screenshot displays the Docker Desktop application window. The left sidebar contains navigation options: Containers, Images, Volumes, Dev Environments (marked BETA), Docker Scout, and Learning center. Below these is an Extensions section with an 'Add Extensions' button. The main panel is titled 'Containers' and shows system metrics: Container CPU usage at 10.63% / 400% (4 cores available) and Container memory usage at 2.31GB / 5.92GB. A search bar and a toggle for 'Only show running containers' are present. A table lists the containers:

Name	Image	Status	CPU (%)	Port(s)	Last started	Actions
airflow		Running (4/5)	10.63%		1 day ago	[Stop] [Refresh] [Delete]
airflow-w d597a6c7e	apache/airflow:2.6	Running	0.66%	8080:8080	1 day ago	[Stop] [Refresh] [Delete]
airflow-tr 6c29f6243	apache/airflow:2.6	Running	3.51%		1 day ago	[Stop] [Refresh] [Delete]
airflow-si 29160ecc4	apache/airflow:2.6	Running	3.94%		1 day ago	[Stop] [Refresh] [Delete]
airflow-in 3bacff360f	apache/airflow:2.6	Exited	0%		1 month ago	[Restart] [Refresh] [Delete]
postgres- c0c0c0c0c	postgres:13	Running	2.52%	5432:5432	1 day ago	[Stop] [Refresh] [Delete]

At the bottom, a status bar shows 'Engine running', system resources (RAM 4.95 GB, CPU 1.51%), and a 'Signed in' status. The Windows taskbar at the very bottom shows the search bar, taskbar icons, and system tray with the date 01-02-2024 and time 11:03.

Airflow running – Dag queued

The screenshot displays the Apache Airflow web interface in a web browser. The browser's address bar shows the URL `localhost:8080/dags/first_dag_v3/graph`. The Airflow interface includes a top navigation bar with links for DAGs, Datasets, Security, Browse, Admin, and Docs. A status bar at the top right indicates the time as 08:58 UTC. A blue notification banner at the top states: "Triggered first_dag_v3, it should start any moment now." Below this, a yellow warning banner reads: "The scheduler does not appear to be running. Last heartbeat was received a few seconds ago. The DAGs list may not update, and new tasks will not be scheduled." The main content area shows the DAG "first_dag_v3" with a status of "queued". It includes a "Schedule: 1 day, 0:00:00" and a "Next Run: 2024-02-01, 00:00:00". Below the DAG name, there are tabs for Grid, Graph, Calendar, Task Duration, Task Times, Landing Times, Gantt, Details, Code, and Audit Log. The "Graph" tab is selected, showing a workflow with five tasks: "file_download", "ETL_Process", "save_to_google_bucket", "create_table", and "load_biquery_table". A status bar at the bottom of the DAG graph lists various task states: deferred, failed, queued, removed, restarting, running, scheduled, shutdown, skipped, success, up_for_reschedule, up_for_retry, upstream_failed, and no_status. The "queued" state is highlighted. An "Auto-refresh" toggle is visible on the right side of the DAG graph. The Windows taskbar at the bottom shows the search bar, taskbar icons, and system tray information including the date and time (09:58, 01-02-2024).

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.

The screenshot shows the Apache Airflow web interface at localhost:8080/dags/first_dag_v3/graph. The DAG is named 'first_dag_v3' and is currently in a 'queued' state. The interface displays a flow graph with five tasks: 'file_download', 'ETL_Process', 'save_to_google_bucket', 'create_table', and 'load_biquery_table'. The tasks are connected in a linear sequence. The interface also shows a top navigation bar with links to DAGs, Datasets, Security, Browse, Admin, and Docs. A status bar at the top indicates the DAG is 'queued' with a schedule of '1 day, 0:00:00' and a next run of '2024-01-31, 00:00:00'. Below the graph, there is a task list with columns for Task ID, Task Name, Task Type, Task Status, Task Duration, Task Tries, Landing Times, Gantt, Details, Code, and Audit Log. The task list shows the following tasks: 'file_download' (BashOperator), 'ETL_Process' (PythonOperator), 'save_to_google_bucket' (BashOperator), 'create_table' (PythonOperator), and 'load_biquery_table' (PythonOperator). The task list also includes a 'Find Task...' search bar and a 'Layout' dropdown menu.

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

The screenshot displays the Google Cloud Storage console for the bucket 'airflow12586-bucket'. The bucket's location is 'us-east1 (South Carolina)' and its storage class is 'Regional'. The public access is set to 'Subject to object ACLs' and protection is 'None'. The 'OBJECTS' tab is selected, showing a list of objects. A single CSV file, 'employee_hr_data20240201.csv', is present, with a size of 88.8 KB and a creation time of Feb 1, 2024, at 10:21:15 AM. The file is stored in the 'Regional' storage class and is not publicly accessible.

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history
employee_hr_data20240201.csv	88.8 KB	text/csv	Feb 1, 2024, 10:21:15 AM	Regional	Feb 1, 2024, 10:21:15 AM	Not public	—

Table is created from Airflow into the Bigquery Database

IBM Malaysia | BigQuery - Airflow12586 - Google Cloud | HR Data Analysis

console.cloud.google.com/bigquery?authuser=2&hl=en&project=airflow12586&ws=!1m5!1m4!4m3!1sairflow12586!2smysql!3shrdata

Google Cloud | Airflow12586 | Search (/) for resources, docs, products, and more

Explorer | Type to search | Viewing resources. SHOW STARRED ONLY

- airflow12586
 - External connections
 - mysql
 - hrdata
 - mytable
 - mytable1
 - mytable2
 - mytable3
 - mytablems

hrdata | airflow12586.mysql | Last modified Feb 1, 2024, 10:21:26 AM UTC+1

hrdata | QUERY | SHARE | COPY | SNAPSHOT | DELETE | EXPORT | REFRESH

SCHEMA | DETAILS | PREVIEW | LINEAGE | DATA PROFILE | DATA QUALITY

Filter Enter property name or value

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
age	INTEGER	NULLABLE	-	-	-	-	-
attrition	INTEGER	NULLABLE	-	-	-	-	-
department	STRING	NULLABLE	-	-	-	-	-
education	INTEGER	NULLABLE	-	-	-	-	-
employeecount	INTEGER	NULLABLE	-	-	-	-	-
gender	STRING	NULLABLE	-	-	-	-	-
jobrole	STRING	NULLABLE	-	-	-	-	-
jobsatisfaction	STRING	NULLABLE	-	-	-	-	-
performance	STRING	NULLABLE	-	-	-	-	-

EDIT SCHEMA | VIEW ROW ACCESS POLICIES

IBM Malaysia | BigQuery - Airflow12586 - Google Cloud | HR Data Analysis

console.cloud.google.com/bigquery?authuser=2&hl=en&project=airflow12586&ws=!1m5!1m4!4m3!1sairflow12586!2smysql!3shrdata

Google Cloud | Airflow12586 | Search (/) for resources, docs, products, and more

Explorer | Type to search | Viewing resources. SHOW STARRED ONLY

- airflow12586
 - External connections
 - mysql
 - hrdata
 - mytable
 - mytable1
 - mytable2
 - mytable3
 - mytablems

hrdata | airflow12586.mysql | Last modified Feb 1, 2024, 10:21:26 AM UTC+1

hrdata | QUERY | SHARE | COPY | SNAPSHOT | DELETE | EXPORT | REFRESH

SCHEMA | DETAILS | PREVIEW | LINEAGE | DATA PROFILE | DATA QUALITY

Table info | EDIT DETAILS

Table ID	airflow12586.mysql.hrdata
Created	Feb 1, 2024, 10:21:19 AM UTC+1
Last modified	Feb 1, 2024, 10:21:26 AM UTC+1
Table expiration	NEVER
Data location	US
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Case insensitive	false
Description	
Labels	
Primary key(s)	
Tags	

Storage info

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

IBM Malaysia

BigQuery – Airflow12586 – Go

HR Data Analysis

console.cloud.google.com/bigquery?authuser=2&hl=en&project=airflow12586&ws=!1m5!1m4!4m3!1sairflow12586!2smydb!3shrdatabigquery

Google CloudAirflow12586Search (/) for resources, docs, products, and moreSearch

Explorer

Type to search

Viewing resources.
SHOW STARRED ONLY

airflow12586

External connections

mydb

hrdata

mytable

mytable1

mytable2

mytable3

mytablems

SUMMARY

hrdata

airflow12586.mydb

Last modified Feb 1, 2024, 10:21:26 AM UTC+1

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	Mr
2	43	0	Sales	2	1	Male	Mr
3	52	0	Research & Development	4	1	Male	Mr
4	52	0	Sales	4	1	Male	Mr
5	45	0	Research & Development	3	1	Male	Mr
6	55	0	Research & Development	3	1	Male	Mr
7	60	0	Research & Development	3	1	Female	Mr
8	50	0	Research & Development	4	1	Male	Mr
9	50	0	Human Resources	3	1	Male	Mr
10	52	0	Sales	4	1	Male	Mr
11	42	0	Research & Development	3	1	Male	Mr
12	47	0	Research & Development	2	1	Female	Mr
13	50	0	Research & Development	4	1	Male	Mr

Results per page: 501 – 50 of 1470

Type here to search

IBM Malaysia

BigQuery – Airflow12586 – Go

Untitled Report

lookerstudio.google.com/u/2/reporting/650d3132-9f2b-4d60-a26c-1092d927156a/page/ho8oD/edit

Untitled Report

File View Page Help

ResetShareView

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

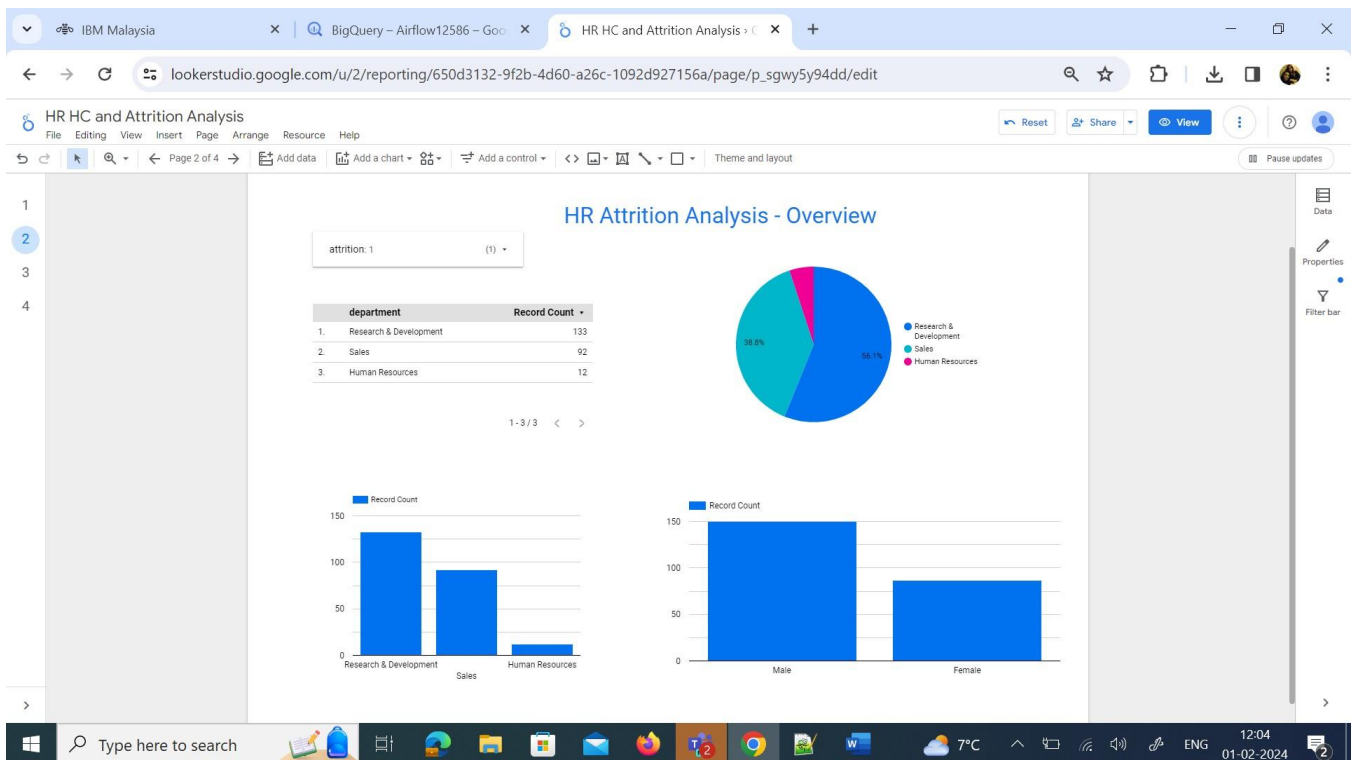
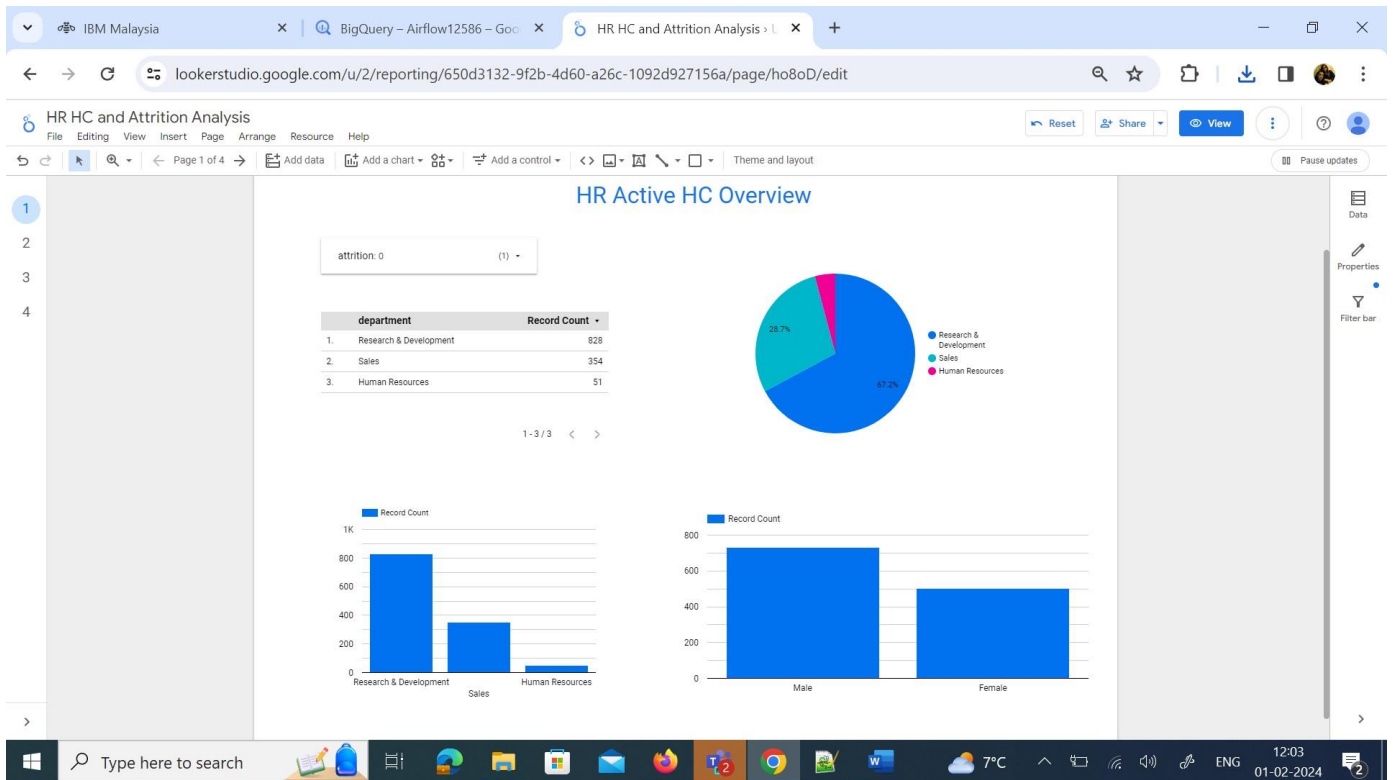
BigQuery is Google's fully managed, petabyte scale, low-cost analytics data warehouse. BigQuery charges for querying/processing data. Those queries are charged to the credit card of the billing project.

LEARN MOREREPORT AN ISSUE

RECENT PROJECTS	Project	Data set	Table
MY PROJECTS	Enter Project ID manually	mydb	hrdata
SHARED PROJECTS	Airflow12586		mytable
CUSTOM QUERY			mytable1
PUBLIC DATA-SETS			mytable2
			mytable3
			mytablems
			mytablenew
			mytablenew1

CancelAdd

Type here to search



IBM Malaysia

BigQuery – Airflow12586 – Go

HR Attrition Analysis › Copy of

lookerstudio.google.com/u/2/reporting/650d3132-9f2b-4d60-a26c-1092d927156a/page/p_ww0s094dd/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

1

2

3

4

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

3	4
200	50

Record Count

1	2	5	3	4	10	0	7	6	8
60	25	20	20	20	15	10	10	10	10

Data

Properties

Filter bar

Type here to search

7°C

12:01

01-02-2024