Data Engineer Challenge
Globant

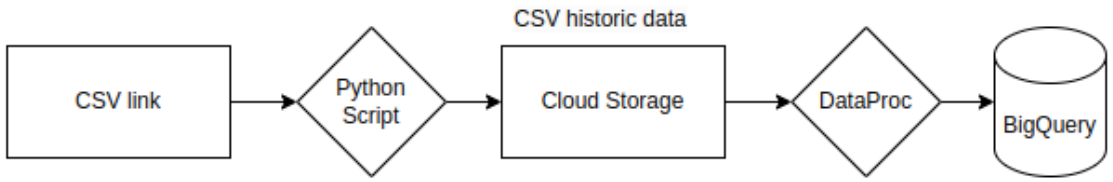
Felipe Ribeiro de Mello Ferreira

May 7th, 2023

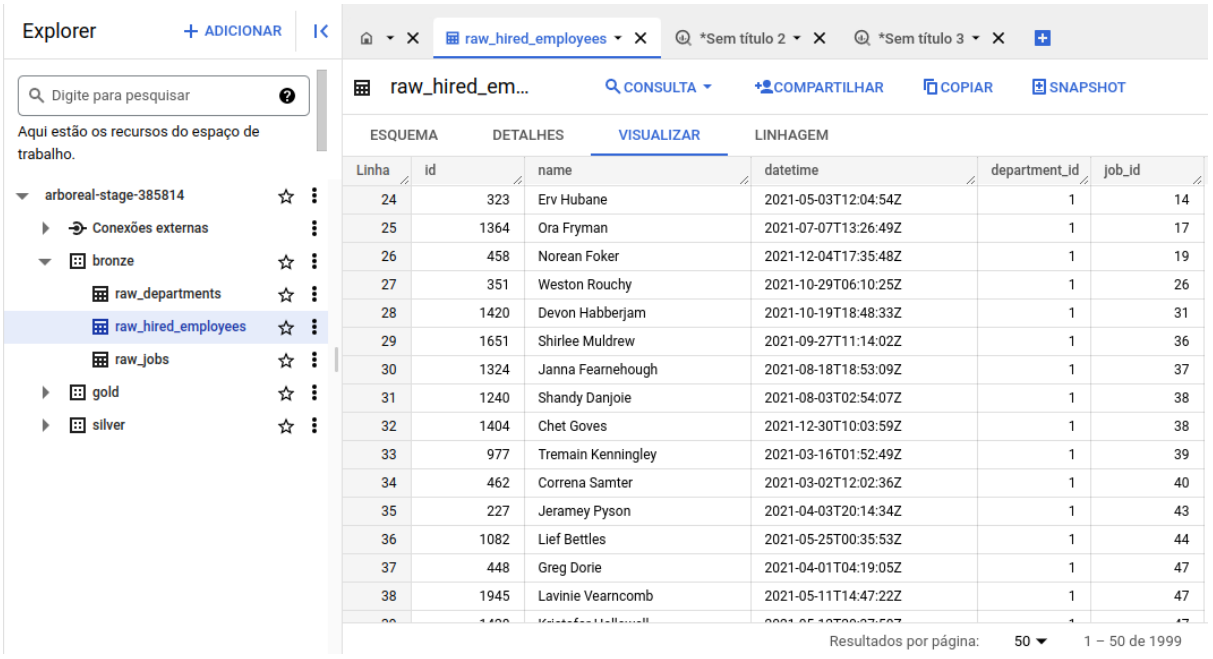
Challenge #1

To solve this challenge I chose to develop the solution in GCP.

Move historic data from files in CSV format to GCP Bucket. Then tables created in Google BigQuery as the following diagram:

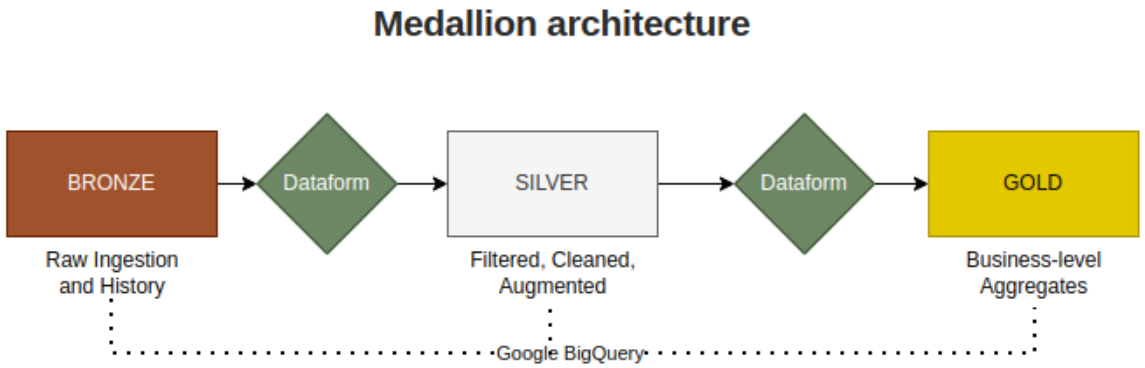


Tables created in Google Bigquery:



Data Architecture

When designing your architecture, the initial and foremost factor to consider is how your data platform will be utilized. Depending on whether you have a centralized and shared platform or a federated multi-platform structure that is used by multiple domains, your architecture will differ significantly. Data is organized with medallion architecture:



Additionally, the layering of your architecture will depend on whether you align your platform(s) with the source-system or consuming side. Generally, it is simpler to standardize the layering and structure of a source-system aligned platform as compared to a consumer-aligned platform, as there are more varied data usage characteristics on the consumption side.

Challenge #2

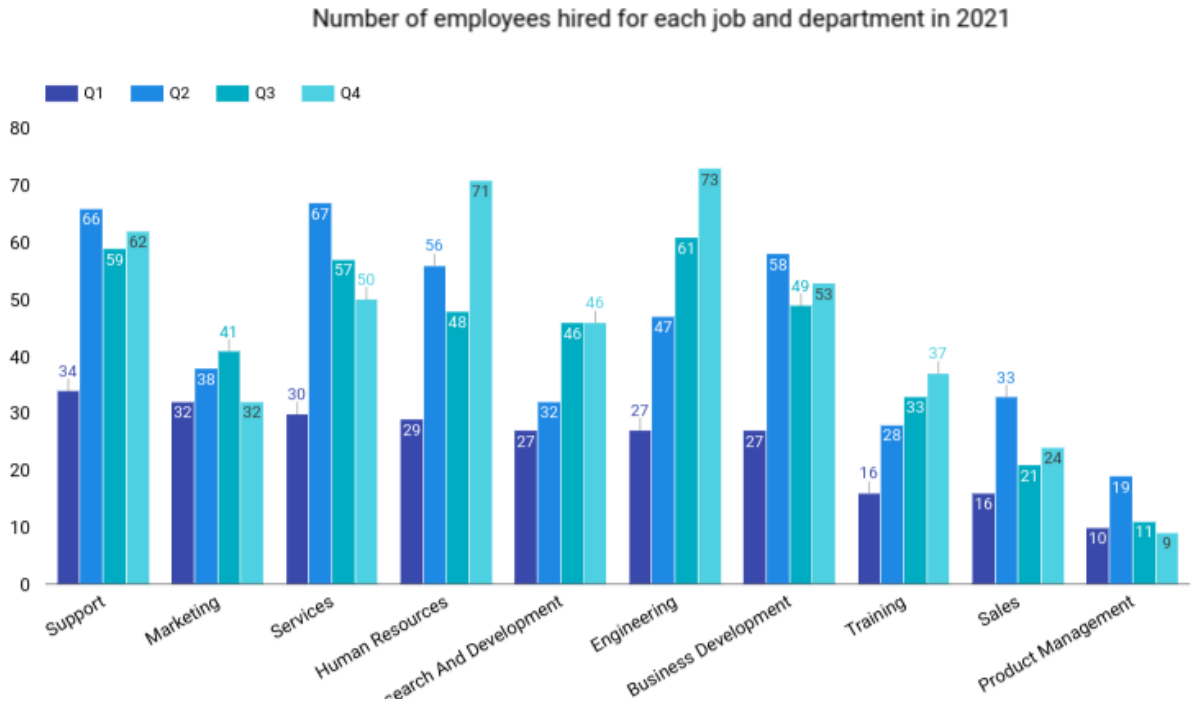
List of chosen tools:

- **Database and SQL Console:** Google BigQuery
- **BI - Visual Report:** Looker Studio: https://lookerstudio.google.com/s/ifLyY6vtqwQ

I - Number of employees hired for each job and department in 2021 divided by quarter. The table must be ordered alphabetically by department and job.

SQL Query:

```
SELECT
 INITCAP(d.department) AS department,
 INITCAP(j.job) AS job,
 COUNT(CASE WHEN EXTRACT(quarter FROM DATE(datetime)) = 1 THEN 1 END) AS Q1,
 COUNT(CASE WHEN EXTRACT(quarter FROM DATE(datetime)) = 2 THEN 1 END) AS Q2,
 COUNT(CASE WHEN EXTRACT(quarter FROM DATE(datetime)) = 3 THEN 1 END) AS Q3,
 COUNT(CASE WHEN EXTRACT(quarter FROM DATE(datetime)) = 4 THEN 1 END) AS Q4
FROM
 `arboreal-stage-385814.bronze.raw_hired_employees` e
 LEFT JOIN `arboreal-stage-385814.bronze.raw_departments` d
   ON e.department_id = d.id
 LEFT JOIN `arboreal-stage-385814.bronze.raw_jobs` j
   ON e.job_id = j.id
WHERE
 EXTRACT(year FROM DATE(datetime)) = 2021
GROUP BY
 department,
 job
ORDER BY
 department ASC,
 job ASC
```



II - List of ids, name and number of employees hired of each department that hired more employees than the mean of employees hired in 2021 for all the departments, ordered by the number of employees hired (descending).

SQL Query:

```
WITH
 department_hires AS (
   SELECT
     department_id,
     COUNT(*) AS hires_count,
     AVG(COUNT(*)) OVER () AS avg_hires_count
   FROM
     `arboreal-stage-385814.bronze.raw_hired_employees`
   WHERE
     EXTRACT(YEAR FROM DATE(datetime)) = 2021
   GROUP BY
     department_id
 )
SELECT
 dh.department_id,
 d.department,
 dh.hires_count
FROM
 department_hires dh
LEFT JOIN `arboreal-stage-385814.bronze.raw_departments` d
 ON dh.department_id = d.id
WHERE
 dh.hires_count > dh.avg_hires_count
ORDER BY
```

```
hires_count DESC
```

BigQuery Console Results

Linha	department_id	department	hires_count
1	8	Support	221
2	5	Engineering	208
3	6	Human Resources	204
4	7	Services	204
5	4	Business Development	187
6	3	Research and Development	151
7	9	Marketing	143

Visual Report