

# Data Engineer Challenge

## Globant

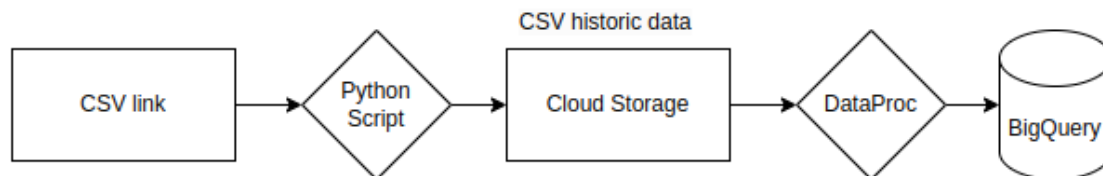
Felipe Ribeiro de Mello Ferreira

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:

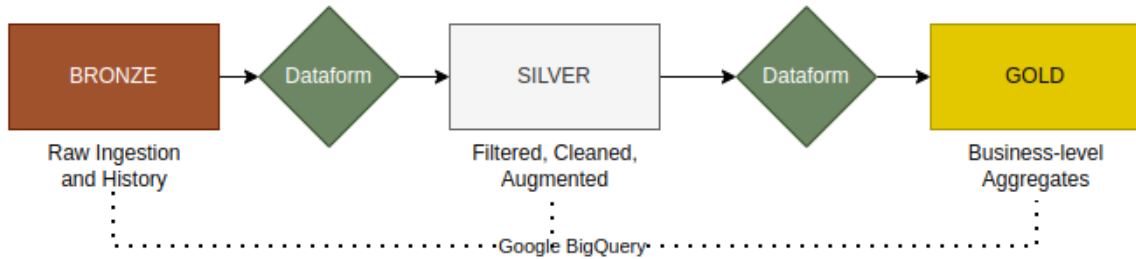
The screenshot shows the Google BigQuery Explorer interface. On the left, a tree view shows the project 'arboreal-stage-385814' with folders for 'Conexões externas', 'bronze', 'gold', and 'silver'. Under 'bronze', there are tables 'raw\_departments', 'raw\_hired\_employees' (selected), and 'raw\_jobs'. The main panel displays the 'raw\_hired\_employees' table with columns: 'id', 'name', 'datetime', 'department\_id', and 'job\_id'. The table contains 38 rows of data. The bottom of the interface shows 'Resultados por página: 50' and '1 - 50 de 1999'.

| Linha | id   | name               | datetime             | department_id | job_id |
|-------|------|--------------------|----------------------|---------------|--------|
| 24    | 323  | Erv Hubane         | 2021-05-03T12:04:54Z | 1             | 14     |
| 25    | 1364 | Ora Fryman         | 2021-07-07T13:26:49Z | 1             | 17     |
| 26    | 458  | Norean Foker       | 2021-12-04T17:35:48Z | 1             | 19     |
| 27    | 351  | Weston Rouchy      | 2021-10-29T06:10:25Z | 1             | 26     |
| 28    | 1420 | Devon Habberjam    | 2021-10-19T18:48:33Z | 1             | 31     |
| 29    | 1651 | Shirlee Muldrew    | 2021-09-27T11:14:02Z | 1             | 36     |
| 30    | 1324 | Janna Fearnemough  | 2021-08-18T18:53:09Z | 1             | 37     |
| 31    | 1240 | Shandy Danjole     | 2021-08-03T02:54:07Z | 1             | 38     |
| 32    | 1404 | Chet Goves         | 2021-12-30T10:03:59Z | 1             | 38     |
| 33    | 977  | Tremain Kenningley | 2021-03-16T01:52:49Z | 1             | 39     |
| 34    | 462  | Correna Samter     | 2021-03-02T12:02:36Z | 1             | 40     |
| 35    | 227  | Jeramey Pyson      | 2021-04-03T20:14:34Z | 1             | 43     |
| 36    | 1082 | Lief Bettles       | 2021-05-25T00:35:53Z | 1             | 44     |
| 37    | 448  | Greg Dorie         | 2021-04-01T04:19:05Z | 1             | 47     |
| 38    | 1945 | Lavinie Vearncomb  | 2021-05-11T14:47:22Z | 1             | 47     |

## 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:

## 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

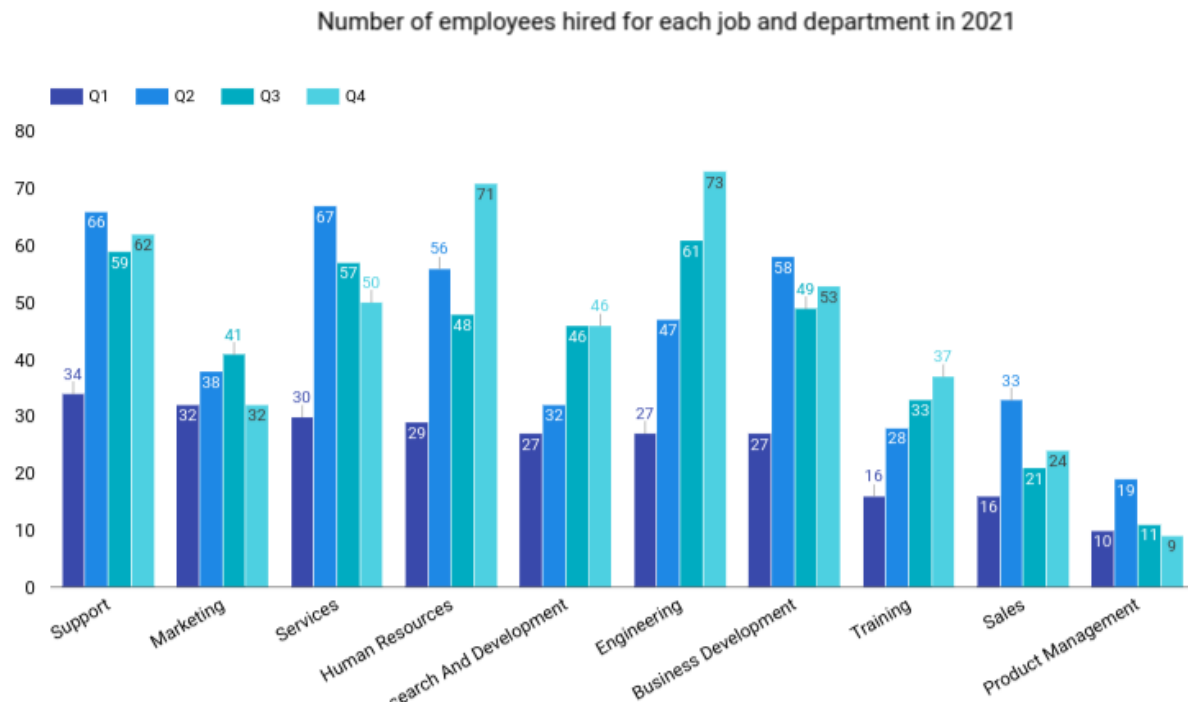
List of chosen tools:

- **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

