Hiring Case - Data Engineer

Applicant: Gustavo Lagares

1) Considering you are a Data Engineer for a big brewery company and your manager asked you to join a meeting for new task of data ingestion. You join the meeting and with you, we have the stakeholder, solution architect and the tech lead, during the meeting the stakeholder asked wants the beer sell data from the orders made in the app. With this on mind, answer the question.

➔ Describe what information or question you need to have answer to do the task the question above.

To develop the immediate activities, these would be the questions needed to structure how the data will be made available:

1. About the origin of the data:

- Is the application data stored in specific tables?
- Where is this data currently stored (transactional database, API, other)?
- Is there any kind of treatment or standardization already applied to this data?
- Is the data already available in a data lake or data warehouse?
- Will it be necessary to request access or involve a DBA?

2. About data ingestion:

- Is there already an ingestion pipeline for this data or will it be something new?
- Will the ingestion be batch, near real-time or real-time?
- What tools and technologies are available or recommended for this ingestion?
- Is there an internal standard or framework already adopted for ingestion processes?
- Will the data be stored in an existing lake or will a new domain be structured?

3. Data processing and transformation (ETL/ELT):

- What business rules are involved in identifying and filtering the "beer" product?
- Does the data need to be enriched or merged with other sources?
- Is there a need to anonymize or mask sensitive data (such as customer data)?
- Which fact and dimension tables should be modeled?
- Is there a standard tool for orchestration and transformation?

4. On data delivery and consumption:

- Who are the consumers of this data (dashboards, APIs, other systems)?
- What business questions does this data need to answer?
- What metrics and aggregations should be made available?
- Is there a data model already defined for the consumption layer (Gold)?
- How often should the data be updated?

5. About the business scope:

- Will we only deal with beer sales or will we also deal with other products?
- How will it be identified that the item sold is a beer?
- Will it be necessary to consider historical data or only future orders?
- Does the data need to be segmented by region, customer, channel, etc.?

With this information, it's already possible to structure ideas for building a scalable pipeline in line with business needs.

➔ Describe how you would do this ingestion, if possible, do a blueprint of your solution

Considering that the application does not yet have the data available in the Lake, that the processing will be done in batch with D-1 data and the final data will be made available on a dashboard, the development of the data ingestion will be considered ingesting and transforming the D-1 data and using medallion architecture (Bronze, Silver, Gold), guaranteeing more organization, traceability and data quality in the processing.

In the ingestion stage:

- The data will be extracted from a relational database via Spark with JDBC, using the D-1 date as a filter. This ensures that only relevant data from the previous day is processed.
- This extraction will be orchestrated by Databricks Workflows, and the application will need access to the bank, which will be requested from the infrastructure team.
- The data will be stored in Delta format in Bronze, without prior processing, with partitioning by ingestion date, which facilitates future queries and audits.
- Alternatively, Azure Data Factory, AWS DMS or Airflow can be used, depending on the company's context.

In the processing and enrichment stage:

- With PySpark in Databricks, Bronze's data will be processed, including:
    - Deduplication;
    - Treatment of nulls;
    - Standardization of dates and values;
    - Enrichment with dimensions such as region, product category and device type.
- After these treatments, the data will be saved in Silver, already standardized and ready for analysis, with appropriate partitioning (e.g. date_ordered).

In the consolidation stage in Gold:

- From Silver, aggregations and metrics will be carried out to feed the dashboard with information such as:
    - Sales by region and device;
    - Seasonality of orders;
    - Identification of event-related peaks.

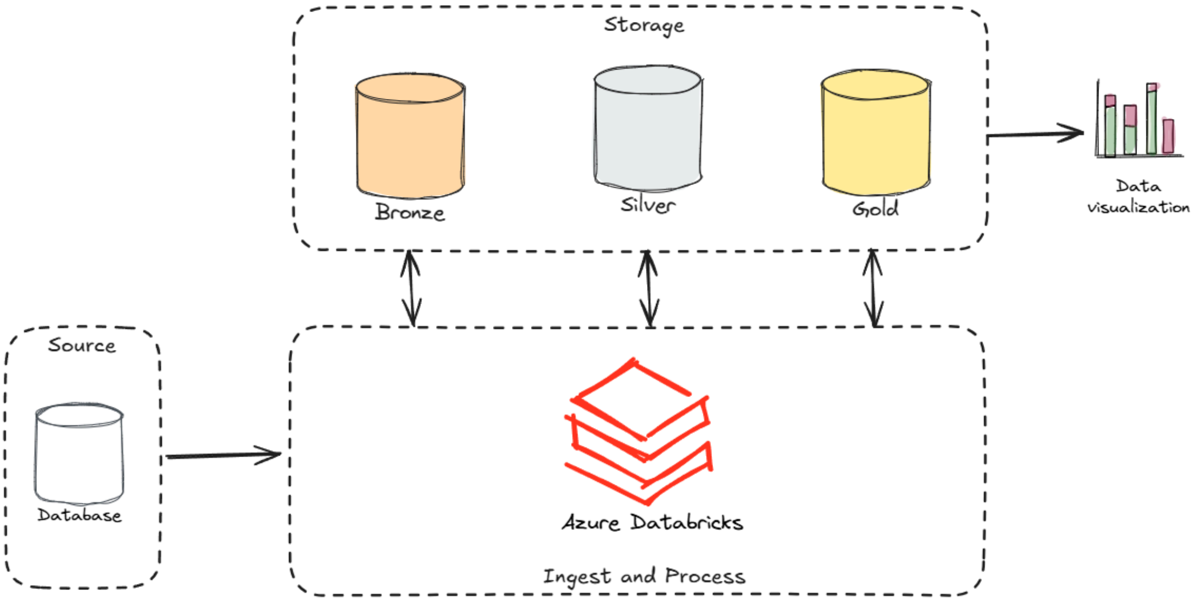
- The data will be made available in Delta tables in the Gold layer, ready for consumption by tools such as Power BI, Tableau or similar.

As for Orchestration and Monitoring:

- The flow will be controlled with Databricks Workflows, including:
    - Daily schedules;
    - Control of dependencies between layers;
    - Reruns in the event of failure;
    - Logging and alerts for operational monitoring.

With this structure, we can guarantee an efficient, reliable and scalable pipeline from D0, structuring the basis for value analysis and decision support.

Below is an illustration of the proposed solution.



2) As a data engineer, you are invited to join a data team to figure out some information's about the covid-19. Since it's a war against the virus, the data team didn't have time to make all the data in the same format, so you have 2 data sets with different formats, JSON and CSV. Using spark, they expecting to have some answers. (We suggest the use of Databricks community https://community.cloud.databricks.com/login.html)

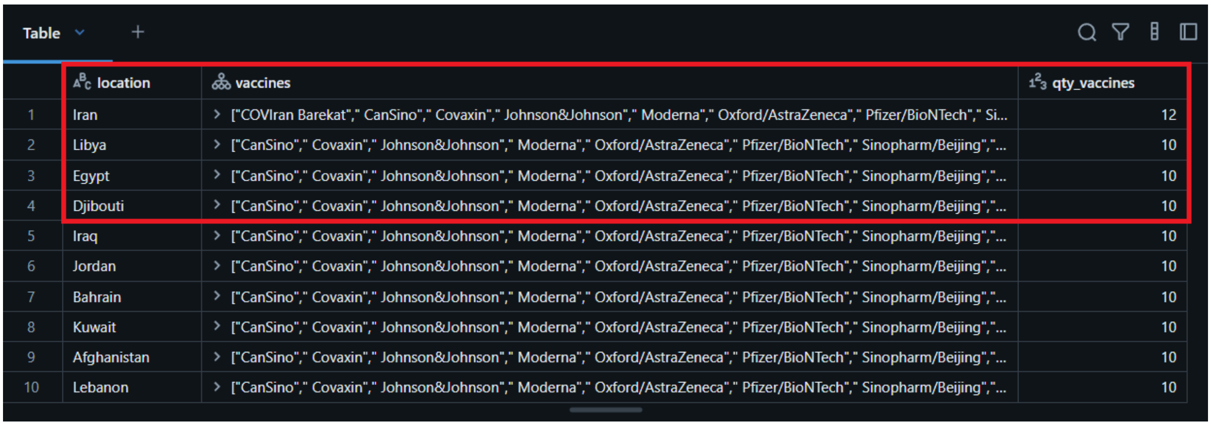
To solve this challenge, an application was developed with Spark, where the source data is ingested into the Bronze layer, in notebook *01.Stage*.

This is followed by processing and storage in the Silver layer, where we have the query to extract the first two questions, in notebook *02.Transformation*.

Finally, in notebook *03.Gold*, the tables are aggregated to extract the answer to the third question and stored in the Gold layer.

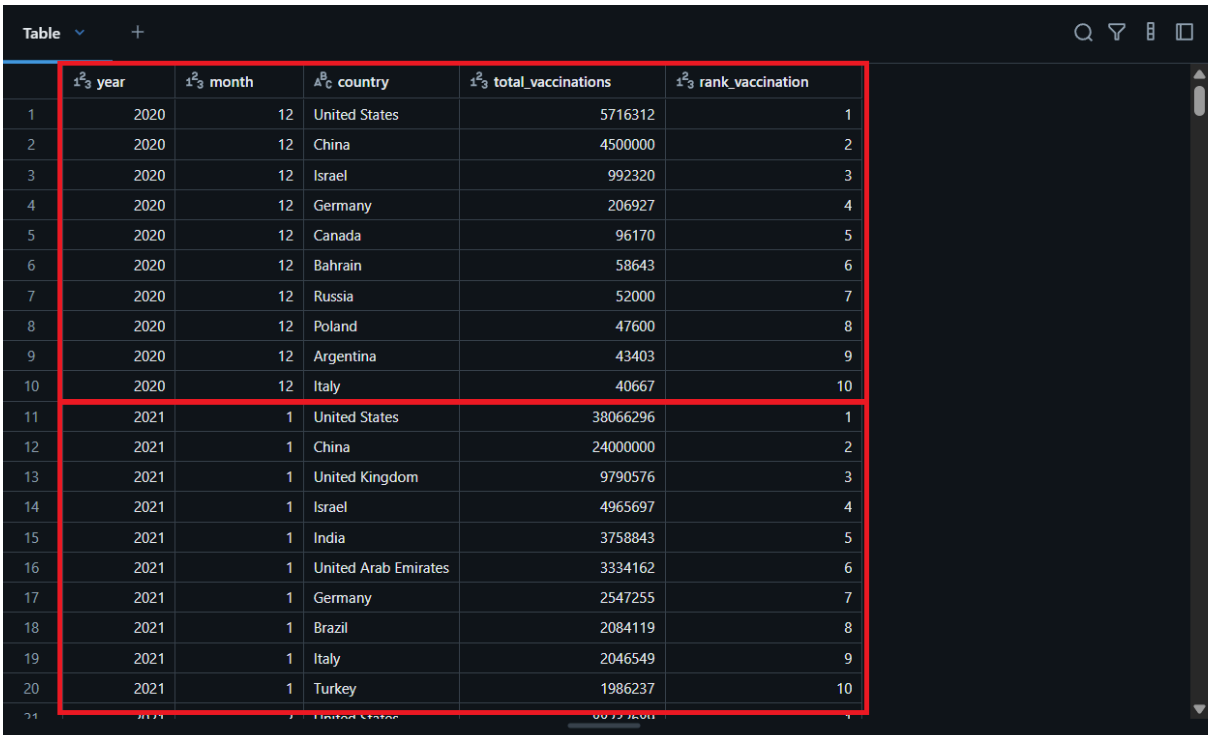
As a preview, here is a screenshot of the query with the data that answers the questions below.

➔ What country(s) use more kind of vaccines (use the file locations.csv)



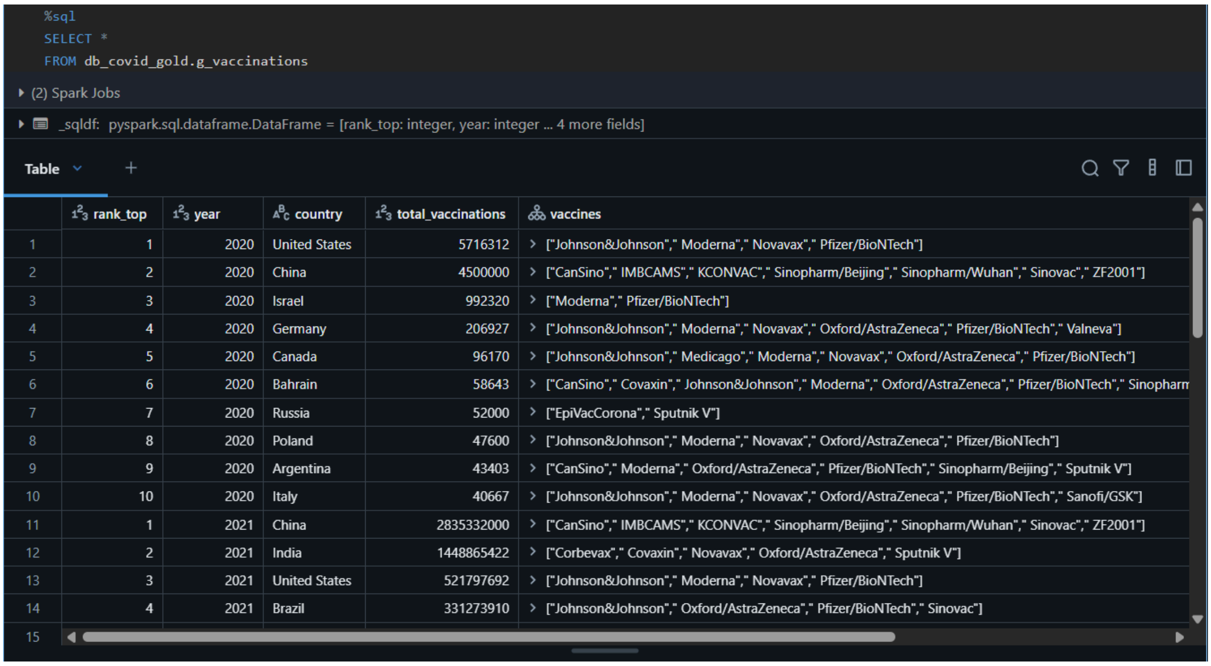
	location	vaccines	qty_vaccines
1	Iran	> ["COVIran Barekat"," CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Si...	12
2	Libya	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10
3	Egypt	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10
4	Djibouti	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10
5	Iraq	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10
6	Jordan	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10
7	Bahrain	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10
8	Kuwait	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10
9	Afghanistan	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10
10	Lebanon	> ["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing","...	10

➔ Top 10 country that had more vaccinations per month and year (use the file vaccinations.json)



	year	month	country	total_vaccinations	rank_vaccination
1	2020	12	United States	5716312	1
2	2020	12	China	4500000	2
3	2020	12	Israel	992320	3
4	2020	12	Germany	206927	4
5	2020	12	Canada	96170	5
6	2020	12	Bahrain	58643	6
7	2020	12	Russia	52000	7
8	2020	12	Poland	47600	8
9	2020	12	Argentina	43403	9
10	2020	12	Italy	40667	10
11	2021	1	United States	38066296	1
12	2021	1	China	24000000	2
13	2021	1	United Kingdom	9790576	3
14	2021	1	Israel	4965697	4
15	2021	1	India	3758843	5
16	2021	1	United Arab Emirates	3334162	6
17	2021	1	Germany	2547255	7
18	2021	1	Brazil	2084119	8
19	2021	1	Italy	2046549	9
20	2021	1	Turkey	1986237	10

➔ Include in the top 10 country vaccinations per year, all the vaccine used during the fight against covid, ordering the top 10, first by most vaccine used and most vaccinated in the year. (use both files)



```
%sql
SELECT *
FROM db_covid_gold.g_vaccinations
```

(2) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [rank_top: integer, year: integer ... 4 more fields]

	rank_top	year	country	total_vaccinations	vaccines
1	1	2020	United States	5716312	["Johnson&Johnson"," Moderna"," Novavax"," Pfizer/BioNTech"]
2	2	2020	China	4500000	["CanSino"," IMBCAMS"," KCONVAC"," Sinopharm/Beijing"," Sinopharm/Wuhan"," Sinovac"," ZF2001"]
3	3	2020	Israel	992320	["Moderna"," Pfizer/BioNTech"]
4	4	2020	Germany	206927	["Johnson&Johnson"," Moderna"," Novavax"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Valneva"]
5	5	2020	Canada	96170	["Johnson&Johnson"," Medicago"," Moderna"," Novavax"," Oxford/AstraZeneca"," Pfizer/BioNTech"]
6	6	2020	Bahrain	58643	["CanSino"," Covaxin"," Johnson&Johnson"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm
7	7	2020	Russia	52000	["EpiVacCorona"," Sputnik V"]
8	8	2020	Poland	47600	["Johnson&Johnson"," Moderna"," Novavax"," Oxford/AstraZeneca"," Pfizer/BioNTech"]
9	9	2020	Argentina	43403	["CanSino"," Moderna"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinopharm/Beijing"," Sputnik V"]
10	10	2020	Italy	40667	["Johnson&Johnson"," Moderna"," Novavax"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sanofi/GSK"]
11	1	2021	China	2835332000	["CanSino"," IMBCAMS"," KCONVAC"," Sinopharm/Beijing"," Sinopharm/Wuhan"," Sinovac"," ZF2001"]
12	2	2021	India	1448865422	["Corbevax"," Covaxin"," Novavax"," Oxford/AstraZeneca"," Sputnik V"]
13	3	2021	United States	521797692	["Johnson&Johnson"," Moderna"," Novavax"," Pfizer/BioNTech"]
14	4	2021	Brazil	331273910	["Johnson&Johnson"," Oxford/AstraZeneca"," Pfizer/BioNTech"," Sinovac"]