Interview Challenge

Description

Galp is going to open a new store in New York City and our team needs to create a new pipeline to ingest all the generated transactions to our data lake. As the local currency (dollars) is different, we need to use a Rest API to do the exchange rate to euros. Airflow will be used for the store ingestion and NiFi will be used to get the dollar / euro exchange rate.

Airflow

We need to ingest all transactions from the new store in New York. The data is locally stored on a database (MySQL). Once a day, Airflow pipeline will read the data for the previous day and save it on Galp's data lake.

All transactions table data format:

Field Name	**Type**
timestamp	Timestamp
product_id	Varchar
product_name	Varchar
product_price	Double
cliente_id	Long

Think how you can build this pipeline in Airflow.

SQL

The table shown below store all the readings of a set of sensors in a refinery

Field Name	**Type**	**Comments**
name	string	Sensor name
ts	timestamp	Sensor reading timesatmp
value	double	Sensor reading value
status	Int	Sensor status
year	Int	Sensor reading year
month	Int	Sensor reading month
day	int	Sensor reading day

Write the SQL statements that allow you to answer the following questions

- Total number of rows;
- Number of distinct sensors present on the database;
- Number of rows for the sensor PPL340;
- The number of rows by year for the sensor PPL340;
- Average number of readings by year for the sensor PPL340;

- For PPL340, Identify the years in which the number of readings is less than the average;

Spark

Answer the questions from the previous exercise (SQL) using any Apache Spark API

Bash

Every six months, we need to run a script bash to delete all the files older than six months.

Think how you can create a new script bash to do the housekeeping.

Data Building Tool - dbt

What is dbt? how it works?

Python (Database extraction)

Imagine you need to extract all the data from a proprietary database through an ODBC driver on an hourly basis. The primary key is a field called “sensor_name” and there are more than 46000 sensors. The output of the extraction will be a Parquet file for all the sensors and to a one-hour window (example: 2020-01-01 00:00:00 to 2020-01-01 00:59:59).

Think how you would build your code to do this hourly extraction so that it will be ready to be consumed by other applications as soon as possible.

Fields and data format:

Field Name	**Type**
timestamp	Timestamp
sensor_name	Varchar
value	Double

Notes:

- It takes more than one hour to complete the query: “`SELECT * FROM table WHERE timestamp >= '2020-01-01 00:00:00' AND timestamp < '2020-01-01 01:00:00'`“.
- You cannot modify anything on the database side.

Data lake structure:

- Raw – where source data lands.
- Standardized – transform raw data to Parquet.
- Aggregated – aggregated data based in standardized data.

Note: For all exercises keep in mind the data lake structure.

Appendix

To create an Airflow cluster use Docker with the image “apache/airflow:1.10.15”.