Take home assessment

Instructions

You are a **senior data engineer**, responsible for overseeing data engineering delivery within a fictional company called ”jara.com”. The goal of this company is to rollout a new back-office data ingestion platform on our cloud backend (GCP or AWS). As part of the high-level requirements, the platform should be able to extract, transform and load data from diverse data sources

You have been given the following sample data file containing sales data. You have been tasked with developing a python based data pipeline that injects its row data from the provided data storage option i.e. google storage (current source is shared here). Your data pipeline should group the results by the **product** and **deal stage** fields in the csv file. The final result should be exported to a grouped by bar chart where the x-axis shows the product groupings and the y-axis, the corresponding number of deal stages found per product.

You are to answer both parts of this take-home exercise below taking into account the following considerations:

- Assume data storage location (google storage), can contain multiple csv files in theory at any point in time
- Performance and scalability should be accounted for in your code
- Assume code & architecture is meant to run on production.
- Assume graph only needs to show the top 5 sales agents
- Assume we care about deals only closed in 2016
- Any data record coming from a sales agent with with first names: “Darcel”, “Kami” or “Jonathan”
- The script you provide runs every 30 seconds

Once complete please send one email containing 2 links to the following:

- your gitlab repo
- solution architecture

Send to < adeola.ojo@stears.co, yewande.olagbaju@stears.co , bode@stears.co and oladele_abeeb@stearsng.com>. Kindly ensure your submissions require no additional permissions to access them 🙏. Please do not spend more than 6 hours on this, focusing only on key aspects for delivering production-ready software. Finally, do not attempt to use AI tooling to answer the questions below as this would show in your work.

Part 1: Solution Architecture

Design, document and share a solution architecture that demonstrates how your data pipeline would consume the data files from a location of your choice in a production environment.

Things your diagram should explain:

1. How it would possibly consume a large number of data files and data records (volume)
2. Considerations made to achieve speed of data ingestion and chart generation (velocity)
3. How the same pipeline can be made to ingest diverse data sources in the future (variety)
4. Data quality considerations (veracity)
5. Key infrastructure and tooling decisions to be captured in your architecture diagram using a level 2 C4 diagram

Part 2: Coding exercise

Create a simple but production-ready python project that meets the requirements above. We expect a decent level of test coverage.

Code Submission:

- Use Python to implement this data pipeline.
- Include a README file detailing how to test and run the pipeline on a local machine.
- Docker file for running the script on any environment (extra points)

Collaboration:

- Submit your code in a single **gitlab** repository and invite the following individuals to review your submission:
    - @foluso_ogunlana
    - @adeola.ojo1
    - @oladele_abeeb

Evaluation Criteria

Candidates will be assessed based on:

- Be as detailed as you can
- Clarity and comprehensiveness of their performance management strategy.
- Creativity and practicality in architectural design.
- Code quality, functionality, and adherence to continuous integration best practices.
- Communication of key technical concepts effectively through your architectural document.
- (Extra points) for providing code that connects to a mock cloud storage

What to ignore

- Do not attempt to add K8s manifests
- Do not attempt to store final table data in a database