📝

Technical test - Data Engineer

Introduction

You are a Data Engineer at Side, and you have been assigned a project: to help product and operational teams match candidates (which we call Siders) and missions (which we call tasks) so that recruitment process is efficient and we can fill as many tasks as possible as fast as possible.

Goal

You are given four data sources:

```
https://s3-us-west-2.amazonaws.com/secure.notion-static.com/2805a453-ea67-4115-8778-76c61c7d905d/data-engineer-technical-test.zip
```

- Tasks
- Task locations
- Siders
- Sider availabilities

After brainstorming with your colleagues, you decide to go as follows:

- Read and understand the data you’ve been given.
- Create automatic pipelines to integrate your data sources to a central data storage system.
  - Questions around the distance between a sider and task will come up, find a way to compare text addresses and geographic coordinates. You may find this kind of resources useful.

- Answer some analysis questions raised during the definition of the mission:
    1. Compute and order by the city with the most siders first:
        a. the total number of Siders per city
        b. the total number of Siders per city that have CACES 5
        c. the percentage of Siders per city that have CACES 5
    2. Compute the number of available Siders per task based on geographical location and on as-the-crows-flies (straight line) distance (Note: transport preference max distance indicate the maximum distance a Sider is ready to travel to go to work).
    3. *BONUS*: Write a SQL query that returns the number of Siders available per task based on all criteria that you find relevant? Here are some relevant explanations of the columns you'll find:

status	Indicates if the Sider can be recruited on a mission or not.
blacklisted	Indicates that the Sider is longer able to work with our company.
caces 5	Indicates if the Sider has a license to drive specific logistic vehicle and until when the license is valid.
driving license	Indicates if the Sider has a driving license to drive a certain category of vehicle.
transport preference: max distance	Indicates the maximum distance a Sider is ready to travel to go to work.
availability: type options	Values are day or night. Indicates if the Sider is ready to accomplish missions during the day, the night, or both. In our case: night hours are from 22:00 / 10PM to 06:00 / 6AM.
availability: start and end date	Indicates the dates between which the Sider is available. If both fields are empty, infer the Sider is unavailable. If start date is filled, consider the sider available from start date to the end of times or until further notification from the Sider. If end date is filled, consider

	the Sider available from the beginning of times to the end date or until further notification from the Sider.
task start date and end date	Indicates the timeframe a mission is happening. Extracting time date from start date and end date will provide you with the start time and end time of each day.

4. *BONUS*: Knowing that missions are worked only during business days, can you compute the Sider’s salary for doing the whole mission?

Report

While doing your project, you will write a documentation that will help your colleagues understand what you did. They should be able to access your project and launch it locally on their machine. You can help them do so by describing how to install and run your project.

To avoid any debate, it is highly encouraged to justify your technical and technological choices so that your teammates understand why you went in this direction.

Otherwise, you are free to do anything you deem necessary: software, programming language, data storage system, etc… and the way you choose to deliver analytics: API, ad-hoc script, dashboard, etc…

Every additional skill shown throughout this exercise will be welcomed especially if documented and adapted to the situation.