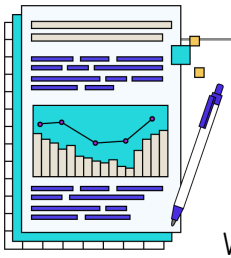
ETUDE DE CAS

PREPARATION AUX TESTS TECHNIQUES DATA ENGINEER 29/11/2024 - 11H30

DataScientest - Data Engineer, Technical interview



Dear [Name],

We would like to invite you to complete a technical task as part of our hiring process for the Data Engineer role.

The task is to collect and analyze data related to jobs in France using one publicly available API.

To complete the task, you are expected to implement ETL processes to extract, transform, and load the data into Snowflake, and use dbt to query the data and compute the desired metrics.

I. Data Ingestion from the Adzuna API

Q1 - Data Collection : Using the website:

- https://developer.adzuna.com/

Collect a list of Data Engineering jobs available in France.

Q2 - Expected Data : The API response contains fields such as job_id, title, company, location, salary_min, salary_max, description, and created. Use these fields to create a structured dataset that can be loaded into a Snowflake staging table.

Q3 - Scheduling : Describe how you would schedule a daily extraction of this data using an orchestration tool such as Apache Airflow.

ETUDE DE CAS

PREPARATION AUX TESTS TECHNIQUES DATA ENGINEER 29/11/2024 - 11H30

II. Loading Data into Snowflake

Q1 - Schema Design : Create a Snowflake schema to store the job postings data. Define appropriate data types for each field, ensuring that the schema can accommodate daily data loads.

Q2 - Data Loading : Describe the steps to load data from the staging area (e.g., a CSV file or temporary table) into the main Snowflake tables. Explain how you would handle incremental data loads to keep the main tables up-to-date with new or modified job postings.

Q3 - Performance Optimization : Describe any partitioning, clustering, or indexing strategies you would use to improve query performance on these tables.

III. Data Transformation with dbt

Using dbt, create transformation models on the Snowflake data to achieve the following:

Q1 - Job Category Summary : A table that summarizes the average salary range (salary_min and salary_max) and total number of jobs for each job category.

Q2 - Location-Based Analysis : A table that aggregates the total number of jobs, average salary range, and unique companies hiring in each location.

Q3 - Monthly Job Trends : A table showing the total number of job postings per month over the past year

Provide a high-level overview of how you would set up dbt models and configure your dbt project to connect to Snowflake. Briefly describe how you would handle development and production environments within dbt.

We would like to request that you provide a written explanation of your implementation decisions, including any assumptions or simplifications made, and a detailed explanation of any problems or issues encountered during the implementation.

Please let us know if you have any questions or concerns about the task or the requirements.