Project - HR Analytics - Cloud Deployment

Purpose

In this project, you will practice deploying a pipeline to the cloud and estimating the cost needed for the cloud deployment.

Scenario

You are a data engineer in a HR agency. You have previously engaged in a proof of concept project to develop a data warehouse pipeline locally (refer to appendix A1 for the background). The stakeholders, talent acquisition specialists, are very satisfied with your work and would like to scale up your work. Your manager in the data engineering team now asks you to work on the cloud deployment of your pipeline, and would like to understand how much the cloud deployment can cost in order to decide whether to continue to the cloud deployment of the pipeline in the future. Work on the tasks below in respond to the request.

Mandatory tasks

Task 1 - Cloud deployment

Start with creating dlt, dbt, dagster and streamlit scripts for a data warehouse pipeline using duckdb. The dlt scripts should load job ads from JobTech API for three optional occupational fields. Data transformation should base on the dimensional model in appendix A2 (in case you want to deviate from this model, you need to motivate the changes). The streamlit scripts should produce at least four relevant visualizations. Then, containerize these scripts and deploy them to Azure with relevant resources.

You should keep your scripts centrally in your Github repo. All students should then use the same scripts in the repo for deployment in their own Azure account. Keep good documentation on the cloud deployment in the README file in the repo.

```
Refer to lecture(s): 05-09
```

Task 2 - Cost estimation

You manager would like to understand how much the cloud deployment can cost. This is to decide whether the company should continue to deploy the data pipeline on Azure. Your cost estimation should base on this plan:

- the duckdb data warehouse should be updated once per day
- the deployed dashboard should be always available to users

Document your cost estimation on the README file in the repo.

```
Refer to lecture(s): 04
```

Bonus tasks

Task 3 - IaC with Terraform

Your manager would like you to explore the use of Terraform to create resources used in the cloud deployment. Because the procedure of the cloud deployment can then be better documented and reproducible. Now, use Terraform to create resources that you used in task 1 above, and deploy the data pipeline again.

Track your Terraform scripts in the same Github repo.

`Refer to lecture(s): 12 and 13`

Task 4 - Cost comparison to cloud data warehouse

Your manager is also very interested to make use of cloud data warehouse instead of duckdb. Therefore, you are asked to illustrate the extra cost that will be incurred if duckdb is replaced by Snowflake. You DON'T need to work on Snowflake in this project. You can estimate the cost based on, for instance, hypothetical computing duration needed for this pipeline. Extend your documentation of cost estimation in task 2 by including extra cost if Snowflake is used as the data warehouse. Also, explain to your manager what are the pros and cons for using Snowflake.

`Refer to lecture(s): 04 and Snowflake cost documentation`

Submission and Grading

There are group and individual submissions for this project. For group submission, you can receive the grade of IG or G. For individual submissions, you can receive the grade of IG, G or VG. In order to get G in this course, both submissions should be at least G. For VG in this course, you need also to obtain VG for individual submission. Below are requirements for each submission:

Group

You should complete at least all manadatory tasks. Include a README.md for your repo to document your project which can guide others to understand your project.

You will submit/complete the followings together as a group:

- a public Github repo containing your shared code base with the project Kanban. The commit history of the code base and the Kanban board should show what each person has done during the project. One student in the group should send in the repo URL on Learnpoint before 31 Oct
- each group will present for 10 minutes on 31 Oct, including these points in the presentation:
    - illustrate the process of cloud deployment and show the deployed dashboard with the Azure account of one student
    - present your cost estimation of the cloud deployment
    - share your way of working in the group

Individual

Individually, you need to submit a video of around 10 minutes to demonstrate your understanding of the project.

More specifically, you need to individually replicate the team work in your individual Azure account. Using the resources created in your individual account to make a video to illustrate how the cloud deployment of the pipeline works. When demonstrating, you need to show that the cloud deployment works on your individual account.

In order to get VG, the deployment where your video demonstration bases on should cover the bonus tasks.

Appendix

A1 - HR Analytics Project - Proof of Concept

Imagine you are a data engineer for a HR agency. Here's an overview of the business model of this agency:

Talent acquisition specialists work with different occupation fields. According to the opening job ads on Arbetsförmedlingen, they will:

- search and contact potential candidates from LinkedIn
- contact and market those potential candidates to corresponding employers

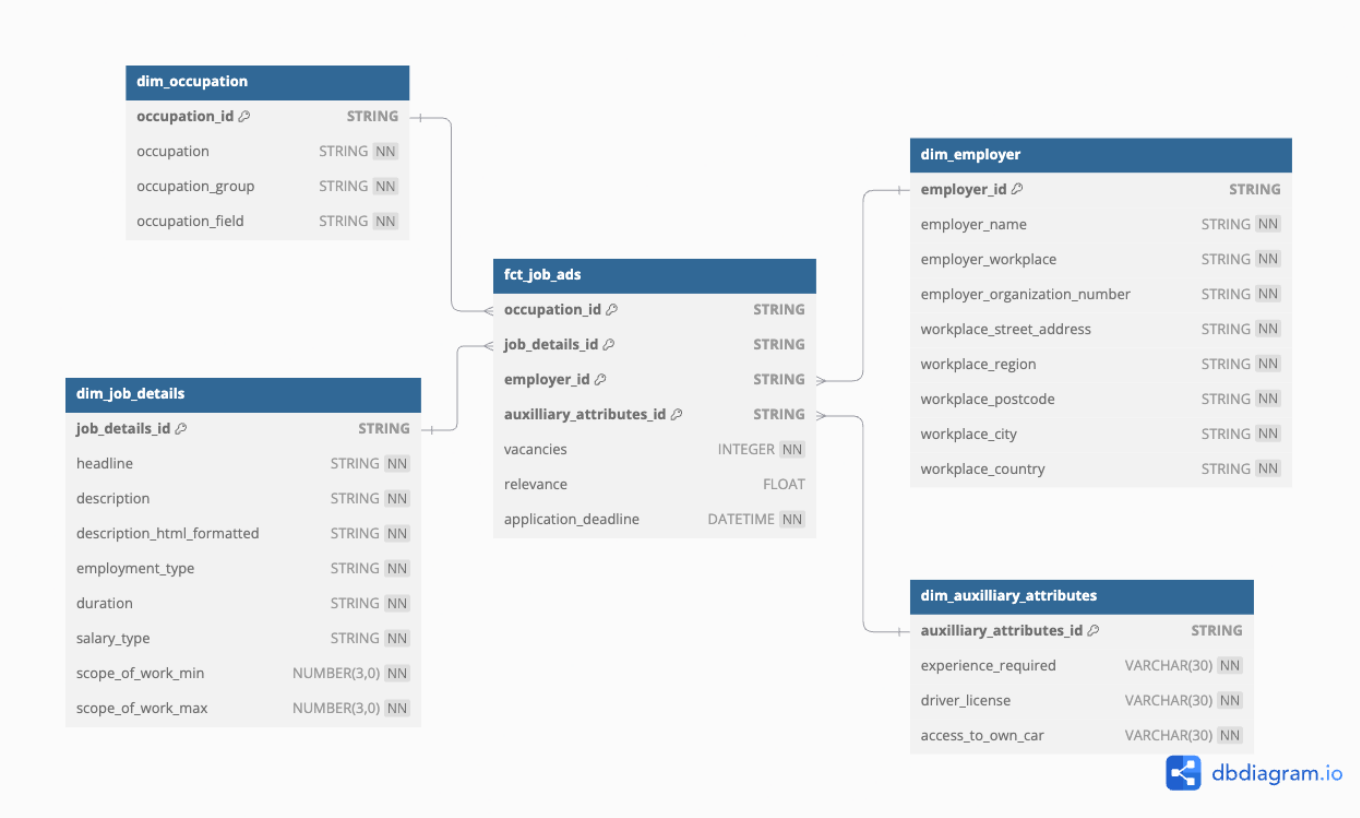
Therefore, they constantly analyze job ads in order to understand which types of candidates they should approach. Currently, every begining of the week, they manually browse the homepage of Arbetsförmedlingen and download a list of opening job ads to guide their work over the week. However, they are not able to draw insights from these job ads as:

- the information are messy
- they have spent too much time to manually collect and clean data so that they do not have much time to analyze the data, which is important to improve the efficiency of their work

Now, you are given a task to create a data pipeline for the team of talent acquisition specialists to:

- automate the data extraction from Jobtech API of Arbetsförmedlingen
- transform and structure data according to a dimensional model
- design a dashboard for talent acquisition specialists to analyse numbers of vacancies by city, by occupation and by employment types etc, for each of the occupation fields
- etc...

A2 - Data Model


dim_occupation
occupation_id STRING
occupation STRING NN
occupation_group STRING NN
occupation_field STRING NN
dim_job_details
job_details_id STRING
headline STRING NN
description STRING NN
description_html_formatted STRING NN
employment_type STRING NN
duration STRING NN
salary_type STRING NN
scope_of_work_min NUMBER(3,0) NN
scope_of_work_max NUMBER(3,0) NN
fct_job_ads
occupation_id STRING
job_details_id STRING
employer_id STRING
auxilliary_attributes_id STRING
vacancies INTEGER NN
relevance FLOAT
application_deadline DATETIME NN
dim_employer
employer_id STRING
employer_name STRING NN
employer_workplace STRING NN
employer_organization_number STRING NN
workplace_street_address STRING NN
workplace_region STRING NN
workplace_postcode STRING NN
workplace_city STRING NN
workplace_country STRING NN
dim_auxilliary_attributes
auxilliary_attributes_id STRING
experience_required VARCHAR(30) NN
driver_license VARCHAR(30) NN
access_to_own_car VARCHAR(30) NN
dbdiagram.io