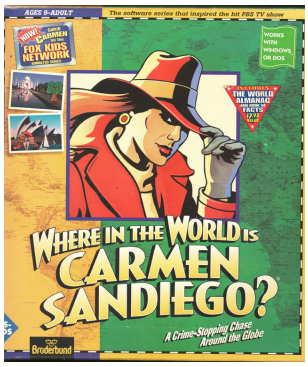
Skill Assessment: *dbt* Data Engineer

Background

👏A sincere congratulations on progressing your candidacy for the Data Engineer role at Cascade Debt! This role is a critical part of our ELT pipeline team, and will largely address transformation from client raw data models into frontend mart models. As such, we'd love for you to show off your dbt skills to us! 😎

For this final assessment, please build a dbt project that addresses the problem listed below. Ensure that you have it cloned to a git-based sharable repository, and that you have a README.md file in the repository's root discussing your analysis for the problem. You may use whichever SQL database flavor you prefer (we use PostgreSQL), just state which was chosen in your README.md.

Problem: *Where in the World is Carmen Sandiego?*



Some may know this popular game franchise in which players answer questions to solve the titular riddle. Your task today is an homage to this game! You've recently come aboard Interpol's team as a data engineer. Their dedicated data team has collected, parsed, and assembled several agent field reports over the years, but in Excel only. Their final collection is provided in the next subsection - your task is to engineer this data to provide analytical answers to the Interpol team!

Data Sources & Common Model Development

The data is contained in the attached Excel workbook *carmen_sightings_20220629061307.xlsx*. Note the sheets are organized by eight (nearly continential) **regions** - there is an Interpol agency HQ in a city of each region to which the agents report. Each agency HQ uses their own language or dialect to compile their regional reports, but those reports are in first normal form (1NF).

> 1. *The first step of your task is to extract data from Excel workbook, treating as initial sources.*
>
>   **HINT:** *CSV exports into* `seeds` *- whether by Excel or pandas - is a great way to start...* 👀

As seen from the data, agencies are free to name report columns according to their custom - but let's call each "yablaka" an apple! 🍎
Each source ought follow a *common data dictionary* of

Column	Description	Type
date_witness	Date of witness sighting	date
witness	Name of witness sighting the perpetrator	string
agent	Name of field agent filing the report	string
date_agent	Date of field agent filing the report	date
city_agent	HQ city where field agent files the report	string
country	Country of sighting	string
city	City of sighting	string
latitude	Latitude of sighting	float
longitude	Longitude of sighting	float
has_weapon	Was the perpetrator observed to be armed?	boolean
has_hat	Was the perpetrator wearing a hat?	boolean
has_jacket	Was the perpetrator wearing a jacket?	boolean
behavior	Short description of perpetrator behavior	string

> 2. *The second step of your task is to create view models that columnarly maps these eight different sources, each into this common data dictionary.*
> **HINT:** *You can do this as you wish - via CTE stages, macros, go wild! The end result however must be a view model for each source.*

Now that you have eight models, all with the same columns - join them together, but with a caveat:

> 3. *The third step of your task is to join the six different views into ONE schema that goes beyond 1NF ([2-6]NF, BCNF). You have a great deal of design freedom here, so get creative! Just persist final resulting schema as tables into a new schema - please present your design's entity-relation-diagram in your README.md.*

Analytics

- This new schema includes the >1NF model you've just developed, as tables. From this model, it ought be fairly straightforward for you to create analytical view(s) to answer the following questions:

  a. For each month, which agency region is Carmen Sandiego most likely to be found?

  b. Also for each month, what is the probability that Ms. Sandiego is armed **AND** wearing a jacket, but **NOT** a hat? What general observations about Ms. Sandiego can you make from this?

  c. What are the three most occuring behaviors of Ms. Sandiego?

d. For each month, what is the probability Ms. Sandiego exhibits one of her three most occurring behaviors?

> 4. *Create analytical views in your new schema to answer the four above questions. Document your steps and logic in your README.md.*
> **HINT:** `dbt docs` *(and its screenshots) are a great resource!*

Submission

> **NOTE:** Throughout, I've referenced `README.md`. If you are unfamiliar with GitHub Markdown, feel free to use your most convenient method: Word document, Google Doc, textfile (with image attachments), html - however you can best communicate your thoughts and ideas!

- **Ensure that your project runs fully BEFORE submission!**
- Verify that you have the f analytical questions answered in your README.md, and you are confident with your presentation.
- Push and merge into your git repository's main branch.
- Submit your project's git repository URL to either the Rippling comms channel or via email to `joshua@cascadedebt.com`.