

Project 1 - ELT pipeline

Ben - [rockerben](#)

Emily - [ZhiwenSong1](#)

Jessica - [JesSchattschneider](#)

Objective

Provide analytical datasets for new data engineers who want to know where to look for jobs and which cities have the potential of being less competitive based on their population size.

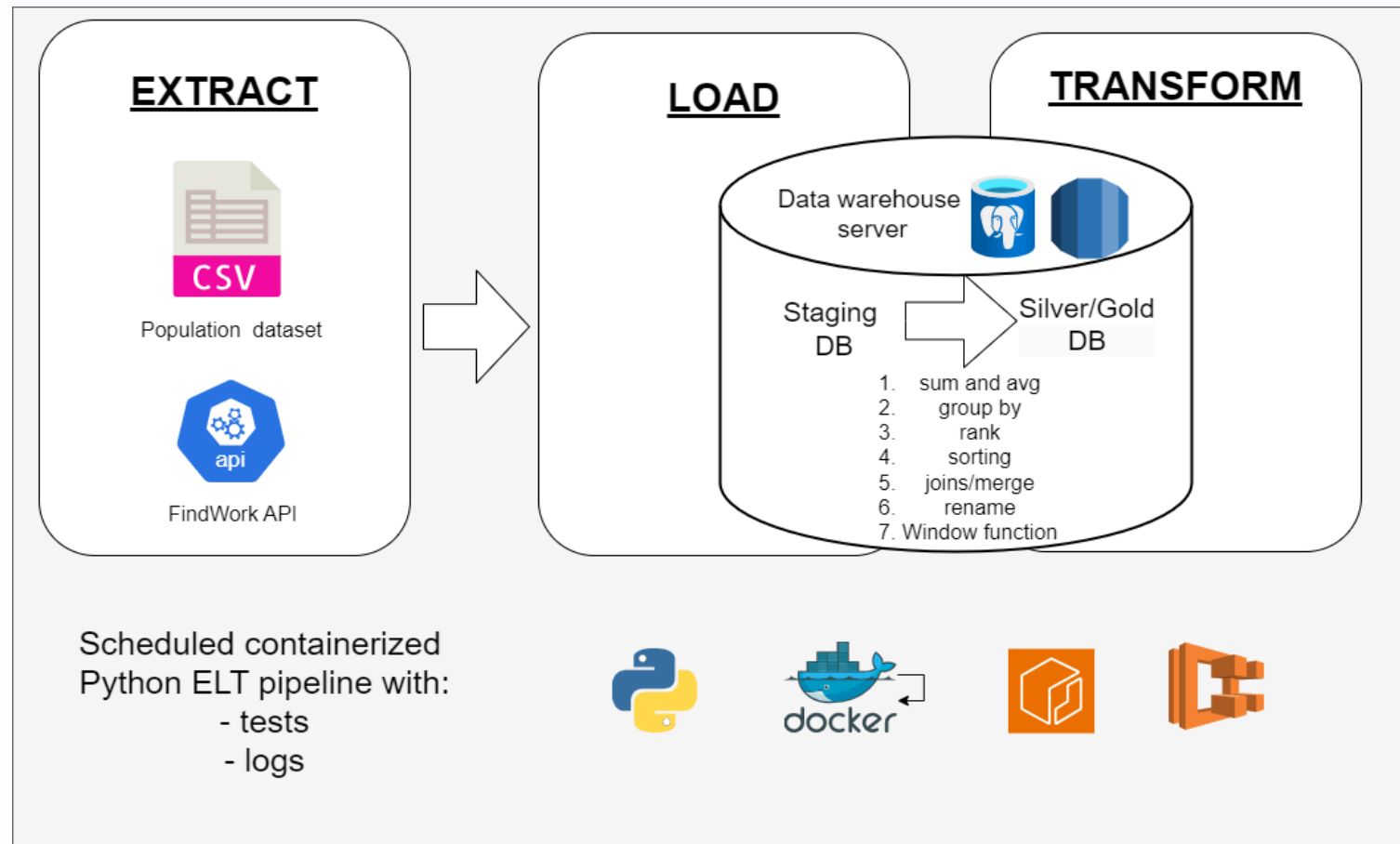
Questions

- Are there remote job opportunities available?
- What are the most common employment types?
- What is the population of the top 10 cities with more job opportunities?

Data sources

Source name	Source type	Source documentation
Population data	csv	TBC
findwork API	REST API	DOCS

Solution architecture



GIT

- GIT Repository
- Activity (9 branches, 17 pull requests and ~80 commits)



Reverse geocoding

Problem

Inconsistencies in how both datasets store their locations.

Example API - location attribute

Location

New York (NYC)

Remote (Europe, US)

Berlin, Germany

Canada

Amsterdam, Berlin, Ghent (EU) On-site/hybrid

Remote, US

Hamburg, Germany (or remote EU)

UK

Example CSV - city and country

City	Country
Tokyo	Japan
Shanghai	China
Dhaka	Bangladesh
Sao Paulo	Brazil

Solution

- Separate country and city from location attribute in the API response (transformation step - python approach)
- Use geopy to the reverse geocoding
- Apply the same approach to the population dataset - groom the original dataset before using it in the pipeline - 7 min process

Extract [Ben]

- CSV
- API - talk about the incremental step and the db checking to identify the number of pages to be extracted []


Load [Ben]

- Load the csv and API to a "bronze DB"

Transform

- Use a python based transformation function to "clean" the location name and load the groomed dataset to the "findwork_data_transformed" table. All the jinja based transformations dependend on the "findwork_data_transformed" table.
- jinja [EMILY]

Docker

 Dockerfile > ...

```
1 FROM python:3.9-slim-bookworm
2
3 WORKDIR /app
4
5 COPY requirements.txt .
6 RUN pip install -r requirements.txt
7
8 COPY src/ /app/src/
9
10 CMD ["python", "-m", "src.extract.pipelines.findwork"]
```

Production [TBD]

AWS

Results

- Are there remote job opportunities available?

	work_location text	num_of_job bigint
1	office work	207
2	remote work	435

- What are the most common employment types?

	employment_type character varying	max bigint
1	contract	10
2	full time	453
3	none	179

- What is the population of the top 10 cities with more job opportunities (**SQL needs review**)?

	city character varying 	num_of_job bigint 
1	berlin	441
2	new york	100
3	london	49
4	los angeles	9
5	dubai	1
6	lisbon	1
7	paris	1
8	san jose	1
9	amsterdam	1
10	tel-aviv	1
11	bengaluru	1
12	chicago	1
13	council of the city of sydney	1