Plan your trip with Kayak

Construction and Management of
a Data Infrastructure

Context

- Create an application for Kayak customers
- Provide information about
  - Weather forecast
  - Hotels in the area
- Selection of 35 top locations to visit

  Mont St Michel, St Malo, Bayeaux, Le Havre, Rouen, ...

Tasks

- Get geographic coordinates of locations
- Get weather forecast for each location
- Get information about hotels in each location
- Archive the data in a data lake
- Store the data in a database
- Use the data to deliver a recommendation application

Pipeline architecture

Data collection

Geographic coord.

Nominatim

Weather forecast

Open-Meteo

Hotels info

Booking.com

Requests

API calls

Web scraping

Selenium 4

CSV

CSV

CSV

Data storage

Database

NEON

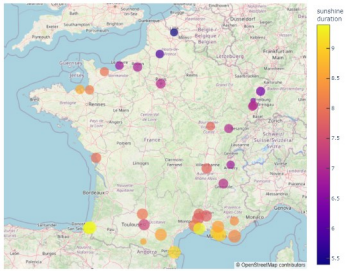
PostgreSQL

Amazon S3

Data lake

Request data

Application



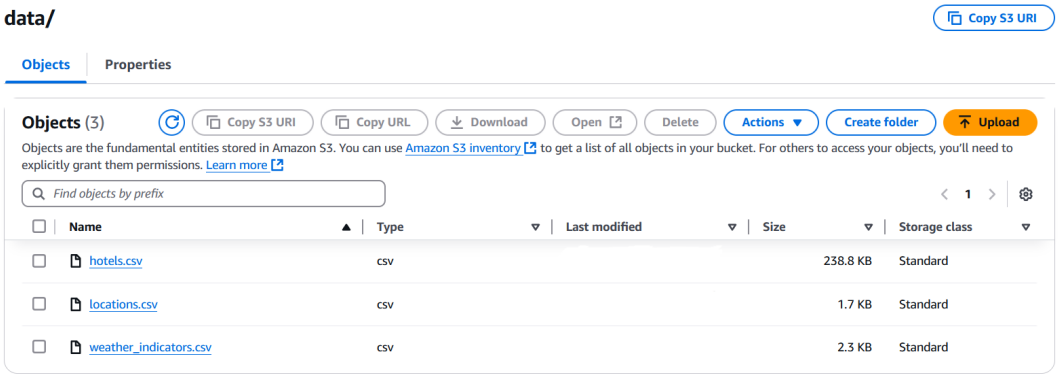
Data collection

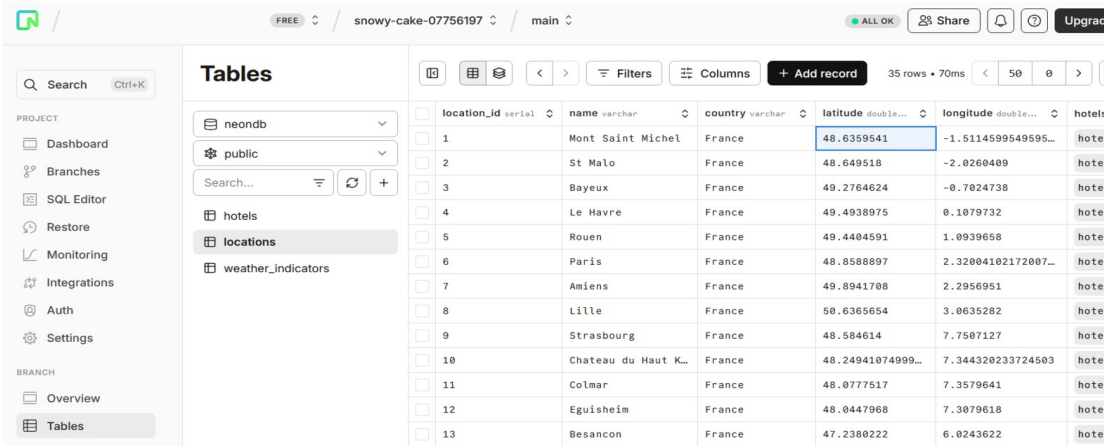
- Get geographic coordinates
  - Nominatim API
  - Rate-limiting: throttle requests
- Get weather forecast
  - Open-Meteo API
  - Use geographic coordinates
  - Average 14 days forecasts
- Get hotels information
  - Scrape booking.com
  - Use Selenium web-driver
    - Allows to execute JavaScript
    - Adpated to complex websites
    - Suited to **infinite scrolling**
  - Difficulties of scraping
    - Must avoid bot detection
    - Websites change over time
      (eg coords not available anymore)

Temp storage in csv format on local filesystem

Data storage

- Archiving in AWS S3 bucket
  - Transfer csv files
  - Programmatic interface: `boto3`
- Database storage
  - DBMS : PostgreSQL on Neon
  - Database structure implemented with `SQLAlchemy`
  - Programmatic interface
    - `psycopg` driver
    - SQLAlchemy.orm.Session

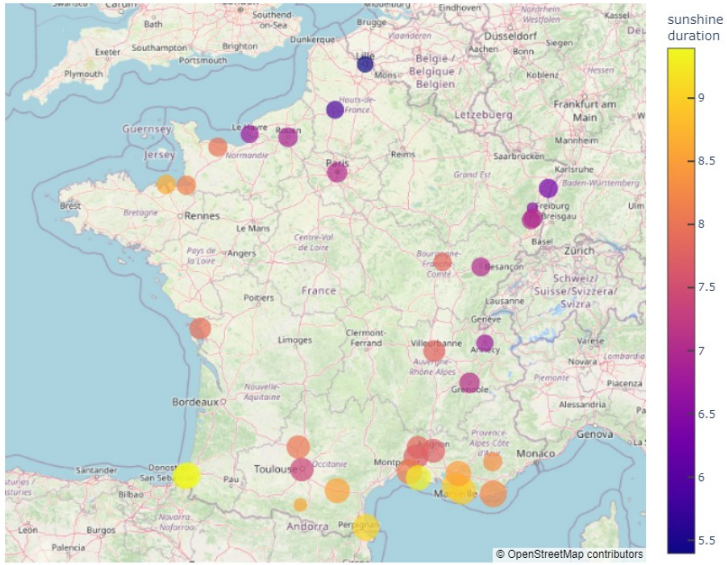




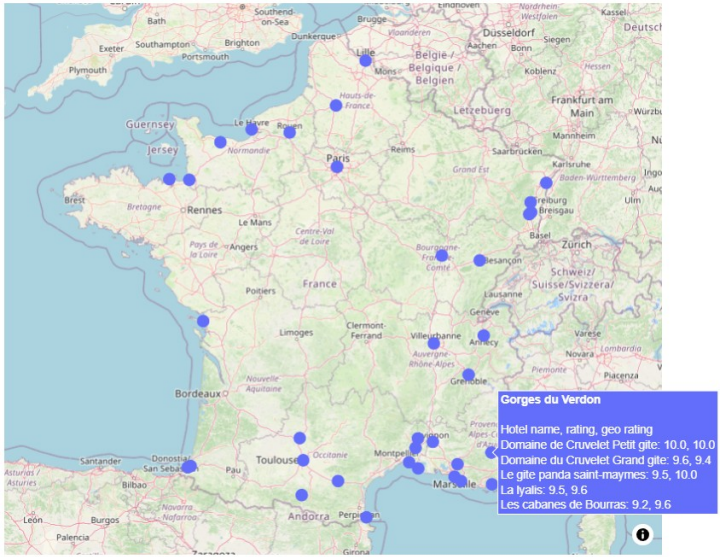
Running the application

- Fetch data from the database
- Render on maps with plotly

Weather forecast



Hotels information



Thanks!