



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

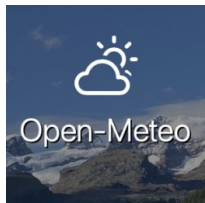


Pipeline architecture

Data collection

Geographic coord.
Nominatim

Weather forecast



Hotels info
Booking.com



API
calls



Web
scrapping



 Selenium⁴

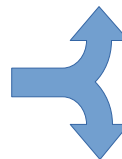


Data storage

Database

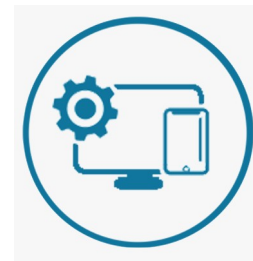


PostgreSQL



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
 - Adapted 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

data/

Copy S3 URI

Objects

Properties

Objects (3)



Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	hotels.csv	csv		238.8 KB	Standard
<input type="checkbox"/>	locations.csv	csv		1.7 KB	Standard
<input type="checkbox"/>	weather_indicators.csv	csv		2.3 KB	Standard

FREE / snowy-cake-07756197 / main

ALL OK Share

Tables

neondb public

Search... +

hotels locations weather_indicators

location_id	name	country	latitude	longitude	hotels
1	Mont Saint Michel	France	48.6359541	-1.5114599549595...	hote
2	St Malo	France	48.649518	-2.0260409	hote
3	Bayeux	France	49.2764624	-0.7024738	hote
4	Le Havre	France	49.4938975	0.1079732	hote
5	Rouen	France	49.4404591	1.0939658	hote
6	Paris	France	48.8588897	2.32004102172007...	hote
7	Amiens	France	49.8941708	2.2956951	hote
8	Lille	France	50.6365654	3.0635282	hote
9	Strasbourg	France	48.584614	7.7507127	hote
10	Chateau du Haut K...	France	48.24941074999...	7.344320233724503	hote
11	Colmar	France	48.0777517	7.3579641	hote
12	Eguisheim	France	48.0447968	7.3079618	hote
13	Besancon	France	47.2380222	6.0243622	hote

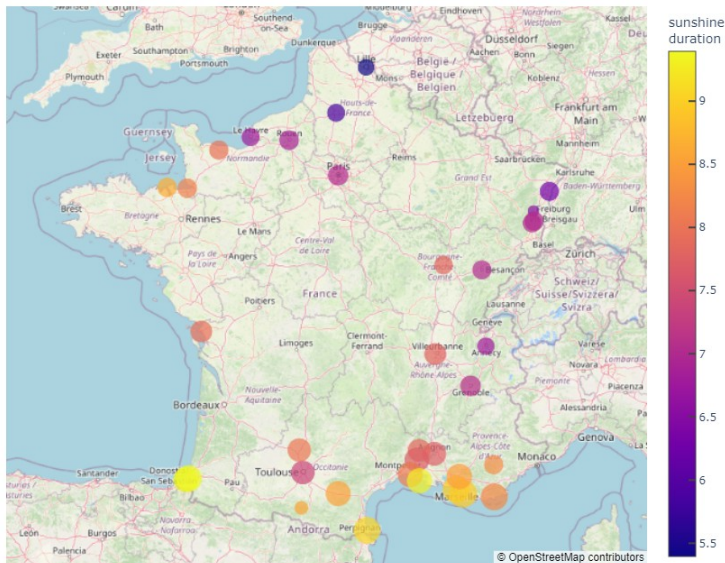


Running the application

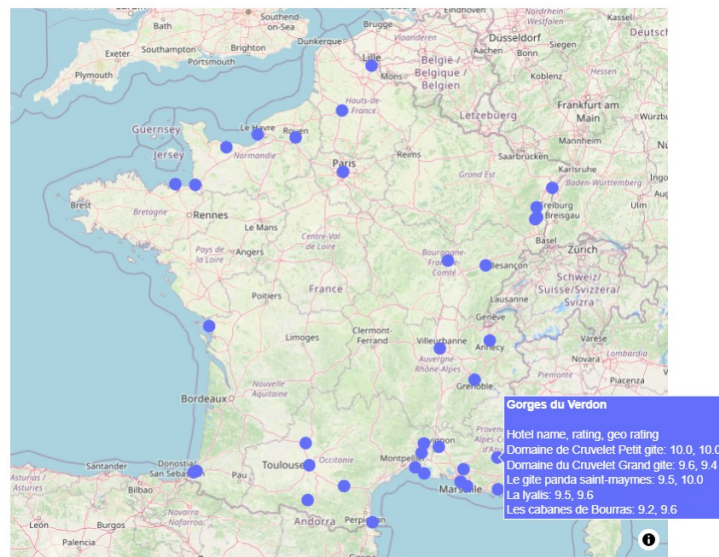
Fetch data from the database

Render on maps with plotly

Weather forecast



Hotels information





Thanks!

