AquaInsight in Île de France over 10 last years

Group Members
Elhadi CHITER, Luka LAFAYE DE MICHEAUX, Marc KASPAR

Data Acquisition, Extraction and Storage - IASD Master

December 2024

Table of contents

Introduction

Introduction

Objective:

- Analyze water body quality trends in Île de France using data acquired from APIs.

Data Sources:

- HubEau API (5 APIs) and Sandre API (1 API).

Data Processing & Cleaning:

- Contructed CSV files and cleaned data using Python (Pandas).

Data Storage:

- Built a relational database in PostgreSQL.

APIs

Sandre API

- Provides data on French water bodies and their associated communes.
- Used command-line Bash scripts to retrieve all communes in France with water bodies.
- Cleaned and filtered the dataset using Python (Pandas library).
- Retained only the most relevant columns for analysis.
- Focused on communes in Île-de-France.
- Easy and Straightforward.
- Retrieved dataset: **communes_IDF.csv**

HubEau API - Qualité des cours d'eau

- Provides physico-chemical water quality measurements for rivers and water bodies.
- Used two API endpoints:
  - **stations_pc:** Retrieved measurement stations using command-line Bash scripts.
  - **analyses_pc:** Retrieved measurement results using Python scripts (easier for pagination handling).

HubEau API - Qualité des cours d'eau

stations_pc endpoint

- Retrieved dataset: **stations_pc_idf.csv**
- Handling NaN values on the dataset: Used group by and aggregation.

analyses_pc endpoint

- Retrieved dataset: **analyses_pc.csv**
- Exhaustive list of physico-chemical parameters (628).
  - Decided to focus only on 20 most important parameters.
- Retrieved data from $01/01/2014$.
- Handling NaN values on the dataset: Represented only 1% of it, so dropped them.

HubEau API - Ecoulement des cours d'eau

- Provides data on the flow of small and medium-sized watercourses in mainland France.
- Used Bash and Python scripts to:
  - Retrieve and clean measurement station data.
  - Retrieve flow observations via API.
- Cleaned and processed the resulting datasets using Python (Pandas library).
- Retrieved dataset: **stations_ecoul_idf.csv** & **observations_ecoul.csv**

HubEau API - Qualité de l'eau potable

- Provides results of sanitary inspections of water quality distributed by municipality in France.
- Used Python to:
  - Retrieve measurements via API.
  - Clean and process the dataset using Python (Pandas library).
- Focused on 20 key parameters.
- Handled NaN values also.
- Retrieved dataset: **analyses_eau_potable.csv**

HubEau API - Qualité des nappes

- Provides physico-chemical water quality measurements for underground aquifers.
- Used Python to:
  - Retrieve measurement stations.
  - Retrieve water quality measurement results via API.
- Cleaned and processed the resulting dataset using Python (Pandas library).
- Focused only on 20 parameters as well.
- Parameters measurements per year.
- Retrieved datasets: **stations_qualite_nappes.csv** & **resultats_qualite.csv**.

HubEau API - Piezométrie

- Provides data on the depth and water levels of underground aquifers.
- Used Python to:
  - Retrieve measurement stations.
  - Retrieve water level measurements via API.
- Cleaned and processed the dataset using Python (Pandas library).
- Retrieved datasets: **stations_piezo.csv** & **piezometrie.csv**

Relational Database: PostgreSQL

Relational Database: PostgreSQL

- We chose a **relational database** for our datasets because it efficiently handles large amounts of data across all communes.
- It's easier to **store** and **organize** data, making it simpler to understand and analyze.
- Allows us to easily **conduct experiments** and manage relationships between data points.

Relational Database Schema

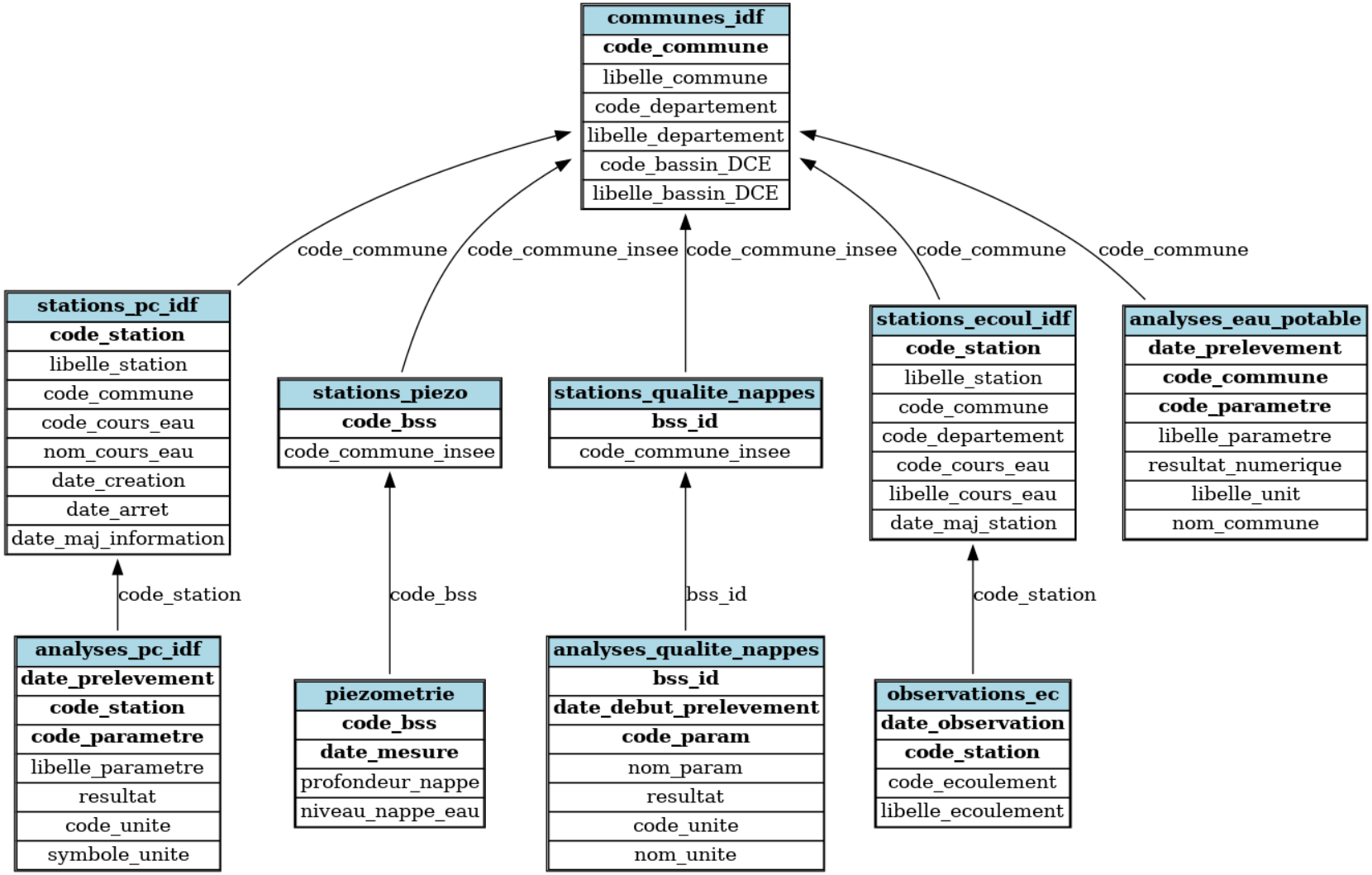

Figure: Relational Database Schema

Use Cases

Use Cases

The dataset we constructed can be used in several applications:

- **Visualizations:** Create plots to analyze water quality metrics and trends.
- **Comparisons**
  - Compare water quality across different regions or types of water bodies.
  - Analyze trends across different periods (e.g., pre-COVID, during COVID, and post-COVID).
- **Machine Learning:** Develop models to forecast and predict water quality trends.

Thank you!