Data Report

1. Question

Analysing the daily flow and storage of Gas in France.

2. Data Sources

Describe your data source: Why you have chosen them, where they are from, and what data they contain? What is the data structure and quality of your sources?

Data Source 1:

The data source is trustable and of governmental authority of data sharing, the data is accessed from https://www.data.gouv.fr/fr/datasets/stock-quotidien-dans-les-stockages-de-gaz-a-partir-de-novembre-2010/ but originally sourced from https://odre.opendatasoft.com/explore/dataset/stock-quotidien-stockages-gaz/information/?disjunctive.source&disjunctive.pits, This dataset presents the gas stock present in the Teréga and Storengy gas storages, by PITS and at the end of the gas day since November 1, 2010 (GWh PCS 0°C). It's a tabular structured data and the quality of the source is accurate, consistent and relevant as well.

Data Source 2:

Similarly this dataset is also accessed from the government website https://www.data.gouv.fr/fr/datasets/debit-quotidien-des-stockages-de-gaz-a-partir-de-novembre-2010/ and originally sourced from https://odre.opendatasoft.com/explore/dataset/debit-quotidien-stockages-gaz/, This dataset presents the accumulation of gas flow rates (injection/withdrawal) moving between the gas transport network and the gas storages of Teréga and Storengy, by PITS and during a gas day since November 1 , 2010 (GWh/d PCS 0°C).

The data is tabular structured and ensures the quality of data on the factors like accuracy, consistency, timeliness and relevancy.

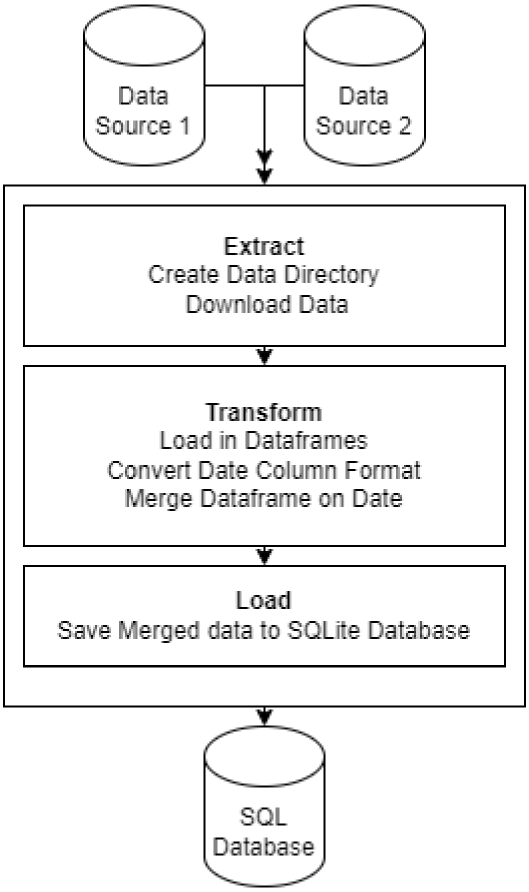
Describe the licenses of your data sources, why you are allowed to use the data and how you are planning to follow their obligations.

The data owns Licence Ouverte / Open Licence version 2.0 As part of the Government's policy in favour of the opening of public data ("Open Data"), Etalab has designed the "Open License". This license, developed in consultation with all stakeholders concerned, facilitates and encourages the reuse of public data made available free of charge.

3. Data Pipeline

Describe your data pipeline on a high level, which technology did you use to implement it, Which transformation or cleaning steps did you do and why?
What problems did you encounter and how did you solve them? Describe how your pipeline deals with errors or changing input data?

This pipeline begins by importing necessary libraries and setting up a directory to store data files. It then specifies the URLs of two datasets and downloads them into CSV files in the designated directory. The data from these files is read into pandas DataFrames, and the date columns are converted to datetime format to ensure accurate merging. The DataFrames are then merged on the date column, retaining only the intersecting dates. Finally, the merged data is saved into a SQLite database, effectively consolidating the information from both datasets into a single, accessible format. As the data is structured and the quality from source was essential to proceed So we didn't need to do cleaning part. The only problem in the available data was to find out a reliable data with intersecting dates of record so that the interpretation could be result conveying.



4. Result and Limitation

Describe the output data of your data pipeline? What is the data structure and quality of your result? What data format did you choose as the output of your pipeline and why? Critically reflect on your data and any potential issues you anticipate for your final report.

The output data of the pipeline is a merged dataset stored in a SQLite database. The merged data consists of records from two original CSV files, combined based on a common date column. Each row in the merged dataset represents a record that contains data from both source files, ensuring that only intersecting dates (common dates in both datasets) are included.
The quality of the data is dependent on the integrity and completeness of the original datasets. Assuming the source data is accurate and complete, the merged dataset should maintain high quality. Any data quality issues such as missing values, duplicate entries, or incorrect formats should be addressed during the data cleaning and preparation phase.

The chosen output format for the pipeline is a SQLite database. This format was selected for several reasons:

- **Efficiency**: SQLite databases are efficient for storing and querying structured data, making it easy to handle large datasets.
- **Portability**: The database file is a single file that can be easily transferred, shared, and integrated with other systems.
- **Flexibility**: SQLite supports various data types and allows for complex queries, facilitating further data analysis and manipulation.

By addressing following potential issues and ensuring a thorough data preparation process, the final merged dataset will be of high quality, suitable for further analysis and reporting. The choice of SQLite as the output format provides a robust and flexible foundation for storing and querying the data, facilitating deeper insights and decision-making

- **Data Interpretation**: Ensuring that the merged data is interpreted correctly is vital. Clear documentation of the data sources, merging process, and any transformations applied will aid in accurate analysis and reporting.
- **Validation**: Validating the merged dataset against known benchmarks or summary statistics from the original datasets can help ensure accuracy.