

Weather Data ETL Project

Diego Iván Perea Montealegre diego.perea@uao.edu.co
Facultad de Ingeniería, Universidad Autónoma de Occidente
Cali, Valle del Cauca

Introduction

The analysis of meteorological data is essential for understanding climate patterns, predicting weather conditions, and studying the impact of climate on various areas such as agriculture, transportation, and disaster management. This synthetic dataset provides a valuable opportunity to research and develop models without the restrictions of accessing real data, facilitating experimentation and the improvement of analytical techniques.

The dataset **Weather Data** of kaggle contains synthetic weather data generated for ten different locations, including New York, Los Angeles, Chicago, Houston, Phoenix, Philadelphia, San Antonio, San Diego, Dallas, and San Jose. The data includes information about temperature, humidity, precipitation, and wind speed, with 1 million data points generated for each parameter.

Features:

- **Location:** The city where the weather data was simulated.
- **Date_Time:** The date and time when the weather data was recorded.
- **Temperature_C:** The temperature in Celsius at the given location and time.
- **Humidity_pct:** The humidity in percentage at the given location and time.
- **Precipitation_mm:** The precipitation in millimeters at the given location and time.
- **Wind_Speed_kmh:** The wind speed in kilometers per hour at the given location and time.

Additional Information:

- **Variability and Complexity:** The dataset incorporates variability and complexity to simulate realistic weather patterns. For example, adjustments have been made to temperature and precipitation based on seasonal variations observed in certain locations. In New York, higher temperatures and precipitation are simulated during the summer months, while in Phoenix, lower temperatures and increased precipitation are simulated during the winter months.

Importance and Applications

1. **Development of Weather Prediction Models:** With parameter-based data, this dataset allows researchers and data scientists to train weather prediction models that can improve accuracy in estimating temperature, humidity, precipitation, and wind speed across different cities.
2. **Climate Studies and Trend Analysis:** By providing simulated weather data from various locations, this dataset enables the analysis of specific climate patterns in each region, the identification of long-term trends, and the evaluation of seasonal variations.
3. **Educational and Academic Use:** This dataset serves as an excellent tool for teaching data analysis, visualization, and modeling in meteorology. Students and educators can use it in courses related to data science, artificial intelligence, and applied statistics.
4. **Optimization in Climate-Dependent Sectors:** Industries such as agriculture, aviation, and logistics can utilize this data to simulate scenarios and develop adaptation strategies for

climate variations, contributing to better planning and decision-making.

5. **Evaluation of Machine Learning Algorithms:** Thanks to the dataset's quantity and diversity, different machine learning approaches can be tested and compared for weather prediction, allowing for model validation before applying them to real-world data.

This synthetic dataset represents a versatile and valuable tool for research, education, and the practical application of climate analysis and prediction techniques. Its availability facilitates the development of innovative solutions without relying on access restrictions to real meteorological data.

Objective

The ETL (Extract, Transform, Load) process for visualizing this synthetic weather dataset would primarily aim to efficiently prepare the data to generate charts and visual analyses that help identify climate patterns, trends, and correlations between variables.

Specific ETL Objectives:

Extraction (Extract):

- Retrieve data from the original Kaggle dataset, ensuring data integrity.
- Load the data into an analysis environment such as Python (using Pandas), SQL, or a visualization tool like Power BI or Tableau.

Transformation (Transform):

- Data Cleaning:
 - Remove null or outlier values that could affect visualization accuracy.
 - Verify the consistency of measurement units (temperature in °C, wind speed in km/h, etc.).
- Format Conversion:
 - Convert the Date_Time column into the appropriate date and time format for temporal analysis.
- Data Aggregation:
 - Group data by city to calculate daily, weekly, or monthly average values for temperature, humidity, precipitation, and wind speed.
- Generation of New Metrics:
 - Calculate moving averages or variation rates to observe climate trends.

Final Project Delivery Steps

1. Data Sources: Select one or more data sources (e.g., CSVs, APIs, databases).
2. Data Extraction: Use Python to extract data from the selected source and store it in a relational database.
3. Exploratory Data Analysis (EDA): Analyze the raw data to understand its structure and quality.
4. Read Raw Data: Retrieve the raw data from the staging area database using Python.
5. Data Transformation: Perform necessary transformations to create value and address the problem.
6. Merge Data (if needed): Combine different data sources through a merge task.
7. Load Processed Data: Store the transformed dataset back into the database.
8. Dashboard Creation: Retrieve data from the ETL pipeline database and create a

dashboard using your preferred tool (e.g., Power BI, Looker Studio)

Steps Project

- **Phase1** : Identification of the data problem or objective and dataset selection ,Data extraction or collection
- **Phase2** : Data transformation , Data pre-analysis and visualization (EDA)
- **Phase3**: Data load in a SQL database , Presentation and Data story telling

Phase 1

The first phase involves extracting data from a source, in this case a .csv file, using Python and uploading the information to a database, which in this case is Supabase.

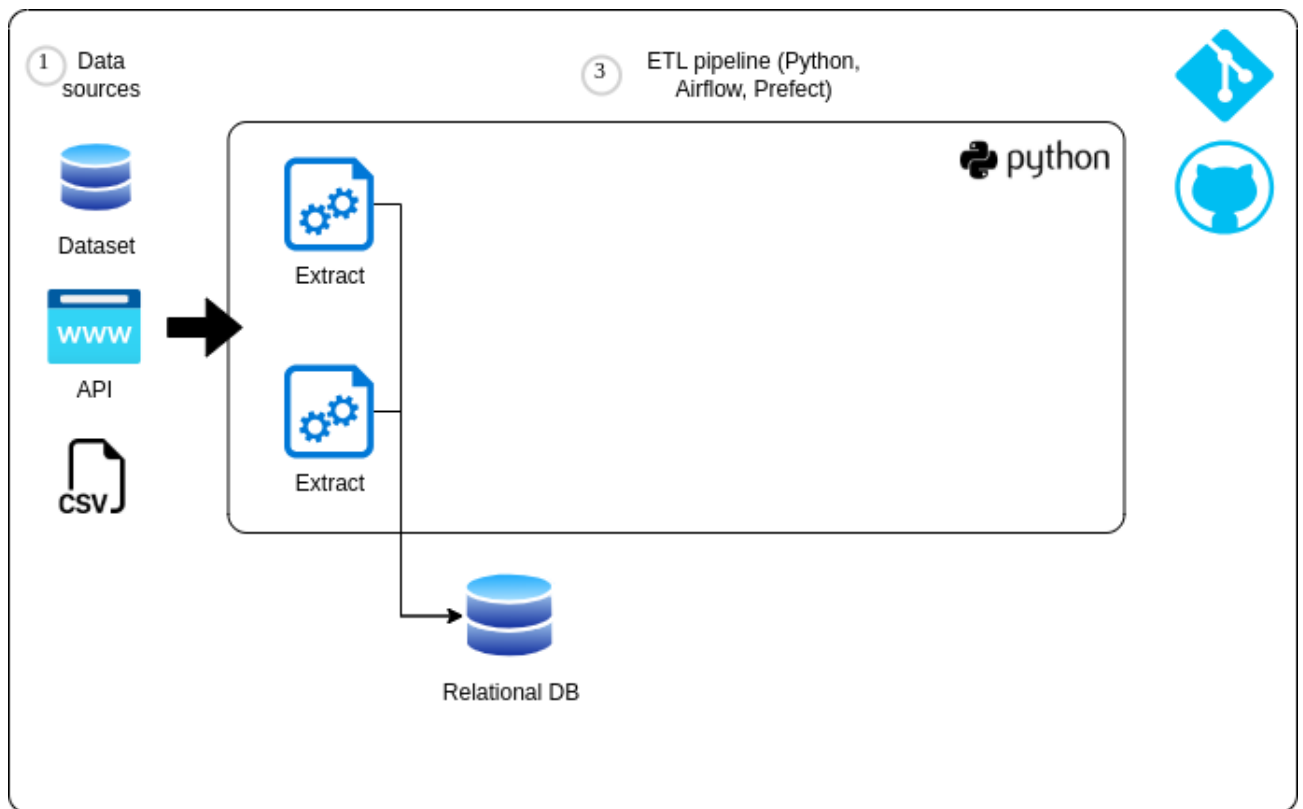
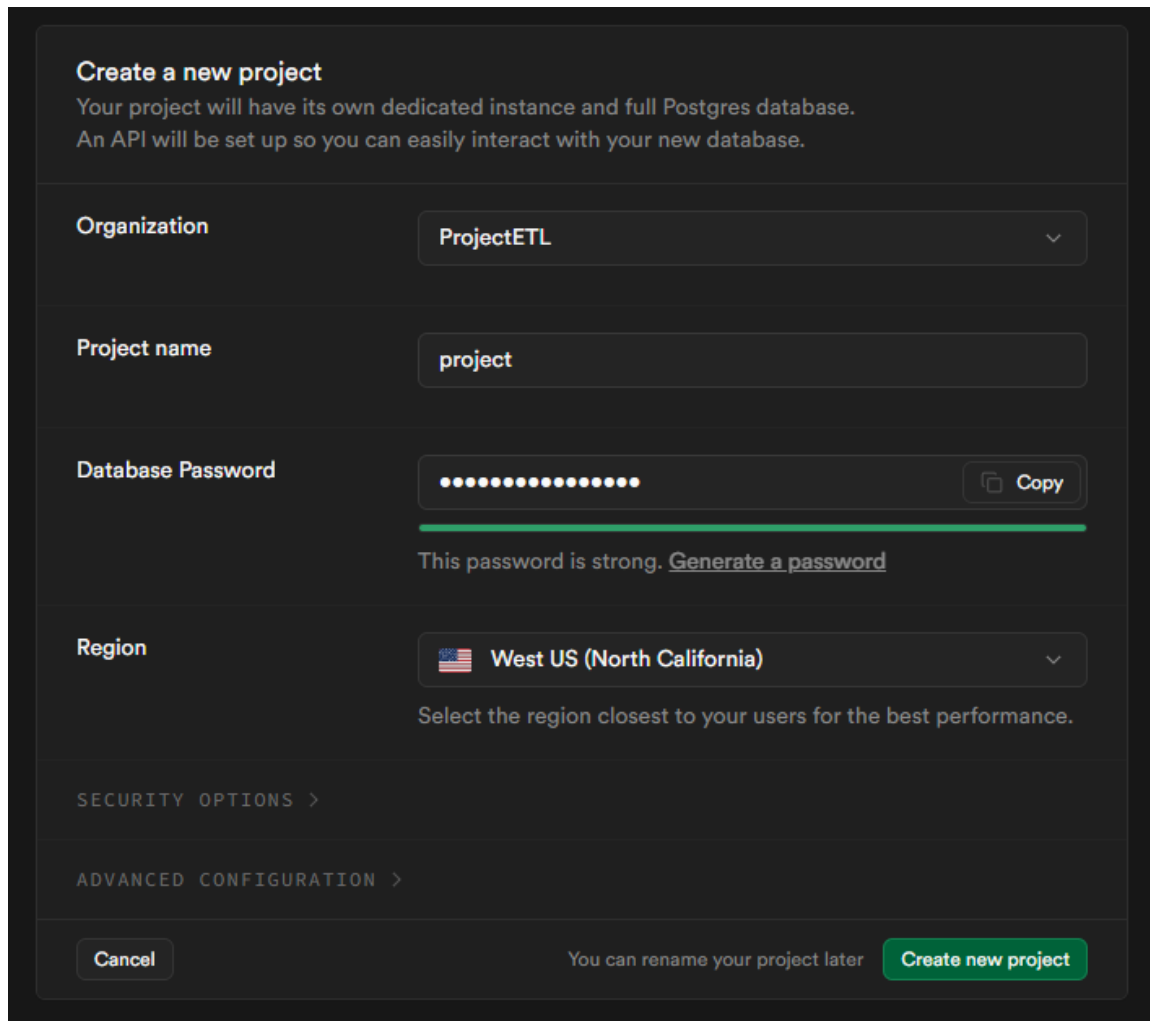


Figure 1. Diagram of Phase 1

Steps for develop Phase 1:

1. Create a database in Supabase. Supabase was selected because is an excellent choice for database management due to its ease of use, seamless setup, and cost-effectiveness. Unlike traditional databases that require complex installations and configurations, Supabase provides a fully managed, serverless PostgreSQL solution that can be accessed directly from the browser or through APIs. This eliminates the need for intricate setup processes, making it ideal for developers who want a hassle-free experience. Additionally, Supabase offers a generous free tier, allowing users to build and deploy applications without upfront costs. Its intuitive interface, real-time capabilities, and built-in authentication features further enhance its appeal, making it a powerful yet accessible choice for modern web and mobile applications.



Create a new project
Your project will have its own dedicated instance and full Postgres database.
An API will be set up so you can easily interact with your new database.

Organization ProjectETL

Project name project

Database Password Copy

This password is strong. [Generate a password](#)

Region West US (North California)

Select the region closest to your users for the best performance.

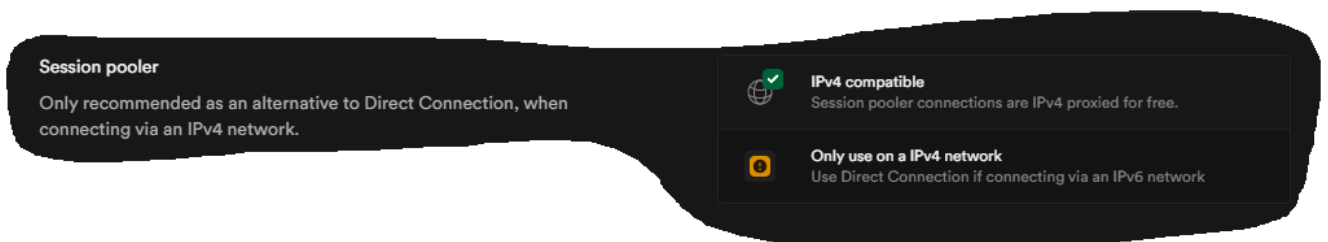
[SECURITY OPTIONS >](#)

[ADVANCED CONFIGURATION >](#)

Cancel You can rename your project later Create new project

Figure 2. SQL Database Creation in Supabase

2. Connect the Supabase database and choose the connection, in this case, "Session pooler"

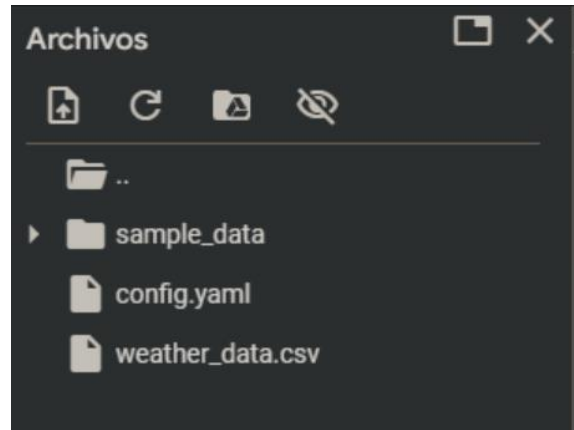


Session pooler
Only recommended as an alternative to Direct Connection, when connecting via an IPv4 network.

- ☒ **IPv4 compatible**
Session pooler connections are IPv4 proxied for free.
- ☐ **Only use on a IPv4 network**
Use Direct Connection if connecting via an IPv6 network

Figure 3. Connection with Session pooler

3. Upload data from the .csv to the database by following the file Phase1Load\001_loadPhase1.ipynb
Download the **weather_data.csv** file [1], create the **config.yaml** file following the steps, and upload it to Google Colab **.ipynb**.



4. View the uploaded data in the Supabase database. The next image shows evidence of the successful upload of the 11000 records handled in the previous step.

| id | Location | Date_Time | Temperature_C | Humidity_pct | Precipitation_mm | Wind_Speed_kmh |
|----|--------------|---------------------|-------------------|------------------|-------------------|------------------|
| 1 | San Diego | 2024-01-14 21:12:46 | 10.6830010947154 | 41.1967535669445 | 4.0201871570867 | 8.23354024687302 |
| 2 | San Diego | 2024-05-17 15:22:10 | 8.7541397823536 | 58.3191073955202 | 9.1162344822938 | 27.7161612568925 |
| 3 | San Diego | 2024-05-11 09:30:59 | 11.6304363129309 | 38.820752691595 | 4.60751137714603 | 28.7329512882362 |
| 4 | Philadelphia | 2024-02-26 17:32:39 | -6.62897589569391 | 54.0744739759617 | 3.18371974780765 | 26.3673026725366 |
| 5 | San Antonio | 2024-04-29 13:23:51 | 39.8082129746316 | 72.8999079529431 | 9.59828213674966 | 29.8986216692961 |
| 6 | San Diego | 2024-01-21 08:54:56 | 27.341054869194 | 49.0332360683476 | 9.16654330273274 | 27.4738960845188 |
| 7 | San Jose | 2024-01-13 02:10:54 | 1.88868336796818 | 65.7423245369102 | 0.22170904319721 | 1.073117812286 |
| 8 | New York | 2024-01-25 19:04:34 | -6.89476554095456 | 30.8048940082328 | 6.02762351418163 | 16.8483368651636 |
| 9 | New York | 2024-03-29 05:20:30 | 0.963544894354438 | 38.8191676409565 | 3.64012916874015 | 7.98902397469938 |
| 10 | San Jose | 2024-05-18 09:14:02 | -1.60708803842582 | 82.1987013132337 | 4.10149297151851 | 25.6472820704365 |
| 11 | New York | 2024-03-04 13:47:15 | 35.145559065071 | 64.7528656584597 | 8.34919496614864 | 25.4303100195307 |
| 12 | Houston | 2024-03-07 22:03:36 | 15.8167635681006 | 80.1199020345908 | 3.76000418615803 | 16.7821322062365 |
| 13 | Dallas | 2024-02-27 21:07:10 | 32.0168976079419 | 63.1943706432686 | 3.58267105442589 | 3.05019638566425 |
| 14 | Houston | 2024-05-09 00:53:10 | 38.6410269264086 | 85.957725559403 | 0.470781842603626 | 20.7792640109 |
| 15 | Houston | 2024-05-12 15:47:55 | 39.666779833673 | 72.7470258115083 | 1.2637215889168 | 6.47949209737064 |
| 16 | Philadelphia | 2024-03-09 01:51:24 | 28.290146882928 | 35.2391696910979 | 9.34720473250742 | 14.0667646401653 |
| 17 | San Antonio | 2024-02-10 15:05:28 | 16.3497895003687 | 65.8126073101361 | 0.109090217726223 | 6.59703303956028 |
| 18 | Chicago | 2024-01-06 02:59:46 | 26.7368110913922 | 31.5136144229176 | 0.496024001203168 | 22.9800953046045 |
| 19 | San Antonio | 2024-05-08 16:20:53 | 35.0795476379057 | 35.0830709963318 | 9.59729410606991 | 4.50786270579149 |
| 20 | San Diego | 2024-01-31 06:38:46 | 14.6058192276471 | 66.6422331815403 | 1.518637132483 | 29.4318902008544 |
| 21 | San Diego | 2024-01-25 12:59:32 | 33.0233907964717 | 62.6074849485407 | 0.21242687171479 | 16.7333348913416 |
| 22 | New York | 2024-02-19 12:26:07 | -7.38381096899866 | 54.0899734429906 | 1.90573059680887 | 6.63706435992247 |

Figure 4. Uploaded Data in supabase database

References

[1] Prasad22. (n.d.). *Weather data* [Dataset]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/prasad22/weather-data>