**Project Documentation**

# Purpose:

This document describes the data model, pipeline logic, data handling rules, and planning for future expansion. It serves as a guide for future team members or developers who may extend or maintain this system.

# Project Overview:

**Project Name:** UK User Activity Pipeline

**Author(s):** [Your Name]

**Environment(s):** GitHub Codespaces, Python 3.11, SQLite

**Objective:** Design a database schema and implement a data pipeline to ingest, clean, validate, and store user data from an external application from multiple countries.

# 2. Database Schema Design

## 2.1 ERD / Schema Diagram (Optional but recommended)

A simple diagram showing tables and relationships.

## 2.2 Tables and Fields

users

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Type | Description | Constraints |
| user\_id | INTEGER | Unique user identifier | Primary Key, AutoIncrement |
| first\_name | TEXT | First name | NOT NULL |
| Surname | TEXT | Surname | NOT NULL |
| middle\_initials | TEXT | Middle Initials | - |
| dob | DATE | Date of Birth | - |
| age\_last\_birthday | INTEGER | Age (at last birthday) | - |
| favourite\_colour | TEXT | Favourite colour | - |
| favourite\_animal | TEXT | Favourite animal | - |
| Favourite\_food | TEXT | Favourite food | - |
| Gender | TEXT | Gender | - |
| Password | TEXT | Password | NOT NULL |
| City | TEXT | City | - |
| County | TEXT | County | - |
| Postcode | TEXT | Postcode | - |
| Email | TEXT | Email | UNIQUE, NOT NULL |
| Phone | TEXT | Phone number | - |
| Mobile | TEXT | Mobile number | - |
| Education | TEXT | Highest level of education achieved for country | - |
| Rqf | TEXT | Highest level of education achieved | - |
| Salary | REAL | Salary | - |
| Currency | TEXT | Currency relating to the salary | CHECK(currency IN ('GBP', 'EUR')) NOT NULL |
| website\_visits\_last\_30\_days | INTEGER | Count of the number of visits to the website in the last 30 days | - |
| country\_code | TEXT | Country Code | CHECK(country\_code IN ('UK', 'FR')) NOT NULL |

logins

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Type | Description | Constraints |
| Login\_id | INTEGER | Login record ID | Primary Key, AutoIncrement |
| user\_id | INTEGER | Related user ID | Foreign Key → users.user\_id |
| login\_timestamp | TEXT | Time of login (ISO 8601) | NOT NULL |

# 3. Data Pipeline Logic

Two main tables are created: 'users' and 'logins'. Each includes a `country\_code` field to differentiate sources like 'UK' or 'FR'. Primary keys and foreign keys ensure referential integrity.

## 3.1 Input Sources

* UK User Data.csv (UK User metadata)
* UK-User-LoginTS.csv (UK Login timestamps)
* FR User Data.csv (FR User metadata)
* FR-User-LoginTS.csv (FR Login timestamps)

## 3.2 Processing Steps

* Load raw data from CSVs using pandas
* Clean column names (remove spaces, lowercase)
* Validate: Check for nulls, unique emails, timestamp consistency
* Transform: Rename columns, format timestamps
* Load into SQLite tables using .to\_sql()

## 3.3 Sample Validation Rules

* Emails must be unique and non-null
* Timestamps must follow ISO 8601 format (YYYY-MM-DD HH:MM:SS)

# 4. Test Data Handling

* Users: 10 fictional users from the UK; 10 from France
* Logins: login events per user
* Format Notes:
  + All fields expected to be clean and pre-formatted
  + Any anomalies (e.g., duplicate emails, bad timestamps) should be logged and skipped or fixed

# 5. Changelog

|  |  |  |
| --- | --- | --- |
| Version | Date | Comments |
| 1 | 12/06/2025 | Initial version with UK data integration |
| 2 | 19/06/2025 | Added support for French data and schema expansion |

# 6. Integration of French User Data

The pipeline and schema have been expanded to support French user data alongside UK data. This supports internationalisation and prepares the system for future expansion to other regions.

## 7.1 Schema Adjustments

* A new column `country\_code` has been added to both `users` and `logins` tables to differentiate source country.
* `country\_code` values are validated against a controlled list (e.g., 'UK', 'FR').
* The schema now supports translations of French column headers to English equivalents.

## 7.2 Pipeline Enhancements

* The data pipeline has been modularised to handle both UK and FR user datasets.
* Encoding differences are handled explicitly (e.g., `ISO-8859-1` for French files).
* French date-of-birth and salary formats are parsed using country-aware logic.
* French column names are mapped to the same unified schema as UK data.

## 7.3 Timezone Handling

* French login timestamps (CET/CEST) are converted to UTC for consistency.
* UK logins are also normalised to UTC using appropriate time zone conversions.

## 7.4 Input Files

* `FR User Data.csv` (user records with French headers)
* `FR-User-LoginTS.csv` (login timestamps in CET/CEST)

## 7.5 Validation and Merging

* User entries are merged using unique `user\_id` values.
* Login records are joined with their corresponding user ID.
* All records are stored in shared `users` and `logins` tables.

# 8. Recommendations:

* Modularise data cleaning functions for different locales
* Prepare for more country-specific data
* Prepare to support multiple ingestion formats (e.g., JSON, APIs)