**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.

# 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 | - |
| Rqf | TEXT | Highest level of education achieved | - |
| Salary | REAL | Salary | - |
| website\_visits\_last\_30\_days | INTEGER | Count of the number of visits to the website in the last 30 days | - |

logins

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Type | Description | Constraints |
| Login\_id | INTEGER | Login record ID | Primary Key, AutoIncrement |
| Username | Text | Username of person logging in | - |
| login\_timestamp | TEXT | Time of login (ISO 8601) | - |

# Data Pipeline Logic

## 3.1 Input Sources

* UK User Data.csv (User metadata)
* UK-User-LoginTS.csv (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
* Logins: 1–3 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. Future Expansion Plan

* Different timestamp formats (e.g., US vs EU formats)
* Country-specific fields (e.g., postcode vs ZIP)
* New tables (e.g., session activity, location logs)
* Schema versioning and compatibility checks

# 6. Recommendations:

* Add a country\_code field to the users table
* Modularise data cleaning functions for different locales
* Prepare to support multiple ingestion formats (e.g., JSON, APIs)