# DARREN CHEN

ADS 507 Data Engineering

Project Design Document: Illicit Drug Use Data Dashboard

•

GitHub Repository Link: <https://github.com/darrencheninfo/data-engineering-pipeline/tree/main>

•

System Architecture Overview:

◦

Describe the entire system, including all components and their interactions.

•

Source Data Description:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset Name** | **Source** | **Format** | **Records** | **Description** |
| **Drug Abuse Warning Network (DAWN)** | ICPSR / SAMHSA <https://www.icpsr.umich.edu/web/>  NAHDAP/studies/34565/versions/V3 | TSV | 229212 | Tracks emergency room visits related to drug misuse and abuse. |
|  |
|  |
| **Youth Risk Behavior Surveillance System (YRBSS)** | CDC <https://www.cdc.gov/yrbs/>  data/index.html | CSV | 16384 | Collects data on youth behaviors, including substance use, violence, and risky activities. |  |
|  |
| **National Addiction & HIV Data Archive Program (NAHDAP)** | ICPSR (ADSS) <https://www.icpsr.umich.edu/web/>  NAHDAP/studies/3088# | CSV | 5005 | Aggregates data related to addiction, drug use, and health issues. |  |
|  |
|  |

CDC. (2025, February 7). *Youth Risk Behavior Surveillance System (YRBSS) Data and Documentation*. Youth Risk Behavior Surveillance System (YRBSS). <https://www.cdc.gov/yrbs/data/index.html>

Quality, U. S. D. of H. and H. S. S. A. and M. H. S. A. C. for B. H. S. and. (2015). *Drug Abuse Warning Network (DAWN), 2011* [Dataset]. Inter-university Consortium for Political and Social Research. <https://doi.org/10.3886/ICPSR34565.v3>

Studies, U. S. D. of H. and H. S. S. A. and M. H. S. A. O. of A. (2009). *Alcohol and Drug Services Study (ADSS), 1996-1999: [United States]* [Dataset]. Inter-university Consortium for Political and Social Research. <https://doi.org/10.3886/ICPSR03088.v5>

Describe the source MySQL database, including its location (Azure), schema, tables, and data types.

Azure: <https://portal.azure.com/#@darrenchenoutlook.onmicrosoft.com/resource/subscriptions/464a0109-8f64-4735-bee3-522cbdb76f9b/resourceGroups/ADS507/providers/Microsoft.DBforMySQL/flexibleServers/mysqldchen/databases>

* Schema:
* Tables:
* Data Types:

Specify which tables and columns will be extracted.

•

Pipeline Specifics (ETL or ELT):

◦ write sql in python,

Create connection string /// (homework 4)

Create stored procedure

ETL: Detail the extraction, transformation, and loading processes.

▪use SQL to build complex DAGs (Directed Acyclic Graphs) using common table expressions, SQL scripts, or an orchestration tool.

Describe how data will be extracted from the MySQL database.

▪

Explain the transformations that will be applied using Pandas.

▪

Describe how the transformed data will be loaded into the PostgreSQL database.

◦

ELT: Detail the extraction, loading, and transformation processes.

▪

Describe how data will be extracted from the MySQL database.

▪

Explain how the data will be loaded into a staging area in the PostgreSQL database.

▪

Describe the transformations that will be applied using SQL queries within the PostgreSQL database.

•

Output Description:

◦

Describe the output, including the schema and tables in the destination PostgreSQL database.

◦

Explain how Power BI will be used to visualize the data.

•

Next Steps:

◦

Outline any shortcomings of the current system.

◦

Suggest potential improvements, such as:

▪

Incremental Loading: Implement incremental loading to process only new or updated data.

▪

Data Validation: Add data validation steps to ensure data quality.

▪

Error Handling: Implement more robust error handling and logging.

▪

Automation: Automate the pipeline using Apache Airflow or a similar tool.

2. Code Implementation (Python with Pandas in VSCode/Jupyter)

Here’s a sample code implementation that fulfills the requirements:

•

Install Libraries:

•

Python Script (etl\_pipeline.py):

3. Setting up PostgreSQL Database

•

Create a PostgreSQL database and user. This can be done via pgAdmin 4

.

# 4. Running the ETL Pipeline

•

Save the Python script (e.g., etl\_pipeline.py) and run it in VSCode or Jupyter:

# 5. Visualizing Data with Power BI

•

Open Power BI Desktop.

•

Connect to the PostgreSQL database using the PostgreSQL connector.

•

Select the destination table (your\_destination\_table).

•

Create visualizations as needed.

•

Publish the report to Power BI Service.

# 6. Key Components and Considerations

•

System Architecture: The system consists of a source MySQL database on Azure, a Python script for ETL, a destination PostgreSQL database, and Power BI for visualization.

•

Source Data: The source data resides in a MySQL database. The script extracts data from specified tables.

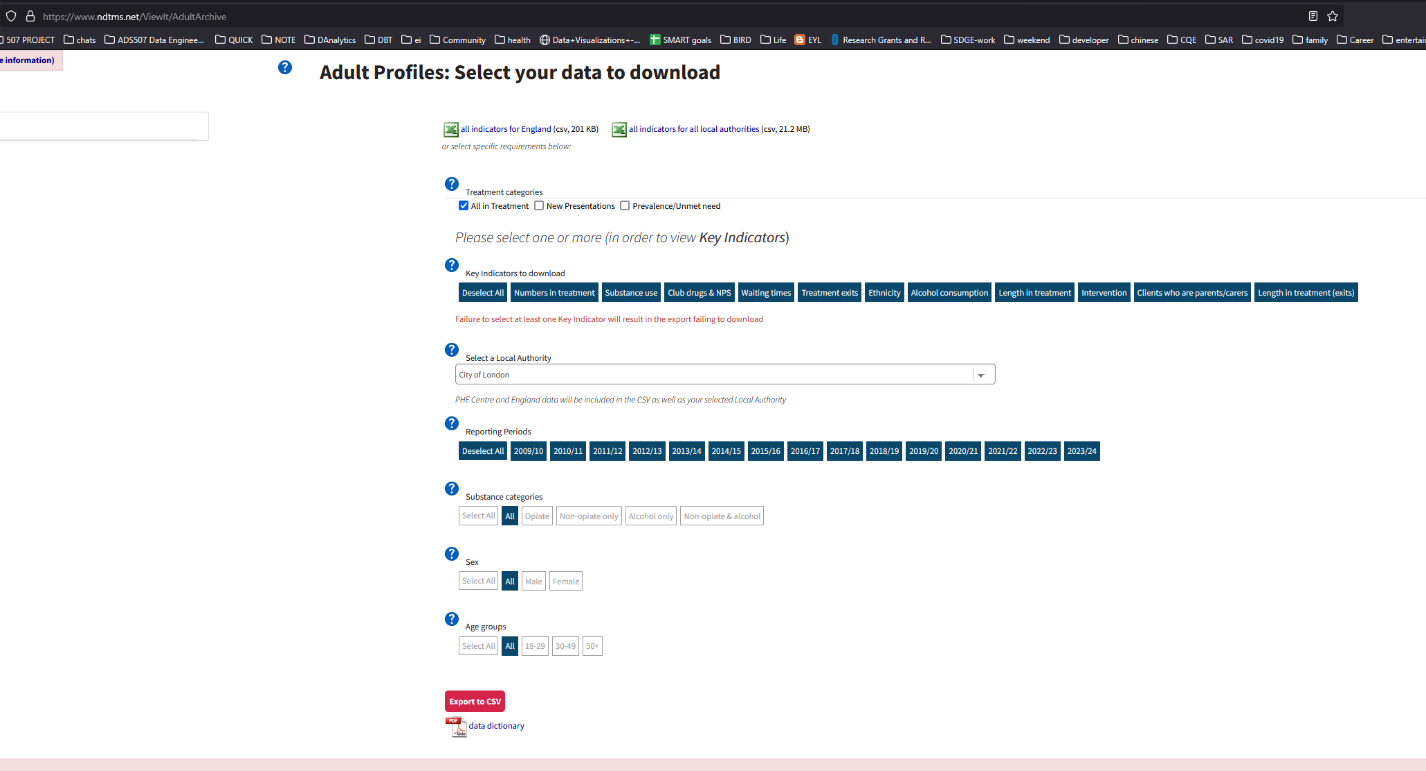
•

ETL Process: The Python script extracts data from MySQL, transforms it using Pandas, and loads it into PostgreSQL.

•

Output: The transformed data is loaded into a PostgreSQL database, which is then connected to Power BI for visualization.

# DATA SOURCES



# INSTRUCTIONS

CONNECT TO AZURE MYSQL SERVER:

# NAHDAP – RESEARCH : Alcohol and Drug Services Study (ADSS), 1996-1999: [United States] (ICPSR 3088)

[https://www.icpsr.umich.edu/web/NAHDAP/studies/3088#](https://www.icpsr.umich.edu/web/NAHDAP/studies/3088)

DELIMITED.

|  |  |  |
| --- | --- | --- |
| DS1 Phase I Facility Interview | 43 MB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS2 Phase II Administrator Interview | 6 MB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS3 Phase II Main/Incentive Abstract | 34 MB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS4 Phase II In-Treatment Methadone Abstract | 9 MB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS5 Phase II Early Dropout Abstract | 5 MB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS6 Phase III Main Study Follow-Up | 22 MB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS7 Phase III In-Treatment Methadone Follow-Up | 17 MB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS8 Phase III Early Dropout Follow-Up | 9 MB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS9 Phase I Finite Population Correction Factors | 566 KB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS10 Phase I Stratified Jackknife Factors | 565 KB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS11 Phase II/III Stratified Jackknife Factors | 564 KB |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
| DS12 Cost Study | 9 MB |

Set caseid and facid as a composite primary key

✅ Ensures uniqueness for caseid and facid

✅ Uses INSERT IGNORE to avoid duplicate primary key errors

✅ Dynamically maps CSV columns to MySQL table schema

# PROBLEMS:

creating this schema:

Error Code: 1059. Identifier name 'Interventions\_Total\_Individuals\_Any\_Struct\_Int\_with\_settings\_AllInTx' is too long