# 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** A diagram of a flowchart

AI-generated content may be incorrect.

Figure : System Architecture Overview

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

•

Table 1: Source Data Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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. |  |
|  |
| Alcohol and Drug Services Study (ADSS) | ICPSR (ADSS)  <https://www.icpsr.umich.edu/web/>  NAHDAP/studies/3088# | CSV Delimited | 5005 | A National study of substance abuse treatment facilities and clients. |  |
|  |
|  |

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>

# Data Pipeline

**Database**: Azure SQL DB

**ETL**: Azure Data Factory + Python

**Web Form:** Azure Static Web Apps + Functions

**Authentication**: Azure AD

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:

## Data Extraction Details:

The CDC Youth Risk Behavior Surveillance System (YRBSS), established in 1990, monitors key youth health risks linked to major causes of death and social issues in the U.S. It tracks demographics, health behaviors, substance use (tobacco, alcohol, drugs), and student experiences (e.g., bullying, violence). Data from 1991 to 2021 includes over 5 million high school students across 2,200+ surveys. This specific dataset extracted from the CDC includes 16,384 student records.

**Data Source:** [CDC YRBSS Data](https://www.cdc.gov/yrbs/data/index.html)

Table 2: YRBSS Extracted Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Actual CSV Column Name | Short Description | Coded Answers |
| RecordID 🔑 | record **🔑** | Unique identifier for each survey record | *(Primary Key, Unique numeric/text identifier)* |
| Age | q1 | Age of the student | A= ≤12, B=13, C=14, D=15, E=16, F=17, G=18 or older |
| Sex | q2 | Sex of the student | A=Female, B=Male |
| Grade | q3 | Grade level of the student | A=9th, B=10th, C=11th, D=12th, E=Ungraded or other |
| Height | q6 | Height of student | Continuous numeric value (height in meters calculated from feet/inches) |
| Weight | q7 | Weight of student | Continuous numeric value (weight in kilograms calculated from pounds) |
| Ever\_Alcohol\_Use | q40 | Ever had at least one drink of alcohol | A=Yes, B=No |
| Current\_Alcohol\_Use | q42 | Currently drinks alcohol | A=0 days, B=1-2 days, C=3-5 days, D=6-9 days, E=10-19 days, F=20-29 days, G=All 30 days |
| Binge\_Drinking | q43 | Had 4 or more drinks of alcohol in a row (binge drinking) | A=0 days, B=1 day, C=2 days, D=3-5 days, E=6-9 days, F=10-19 days, G= ≥20 days |
| Ever\_Marijuana\_Use | q45 | Ever used marijuana | A=Yes, B=No |
| Marijuana\_Use\_Before\_13 | q46 | Used marijuana before age 13 | A=Never, B= ≤8 years, C=9-10 years, D=11-12 years, E=13-14 years, F=15-16 years, G= ≥17 years |
| Current\_Marijuana\_Use | q47 | Currently uses marijuana | A=0 times, B=1-2 times, C=3-9 times, D=10-19 times, E=20-39 times, F= ≥40 times |

•

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

▪

Describe the transformations that will be applied using SQL queries.

•

Output Description:

◦

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

◦

## Visualization

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

◦

# 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.

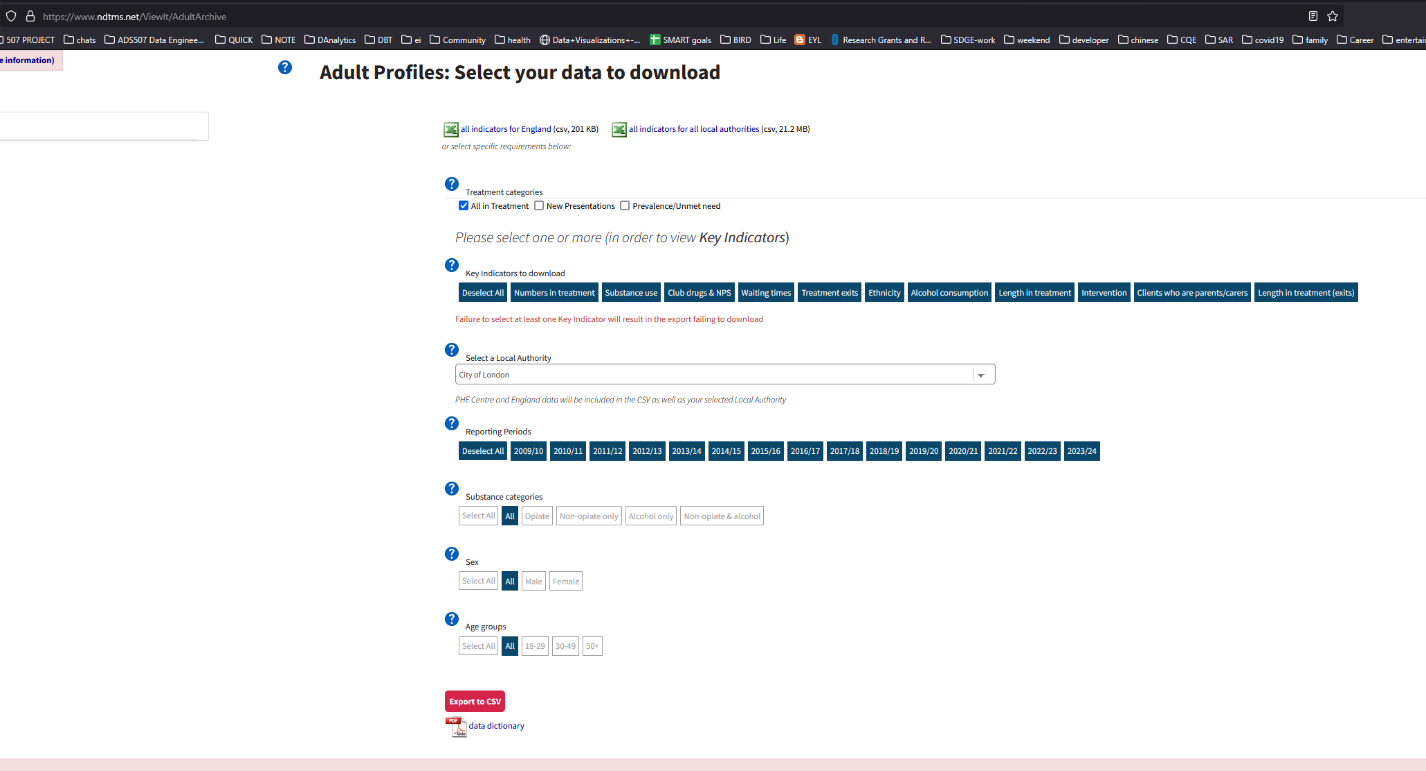
•

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