# DARREN CHEN

ADS 507 Data Engineering

Project Design Document: Illicit Drug Intervention Data

Spring 2025

**Table 1: Pipeline Resources:**

|  |  |
| --- | --- |
| Resource: | Location: |
| GitHub Repo | <https://github.com/darrencheninfo/data-engineering-pipeline/tree/main> |
| Database | mysqldchen.mysql.database.azure.com |
| Database Monitoring |  |
| Data Source CDC | <https://www.cdc.gov/yrbs/files/2021/XXH2021_YRBS_Data.dat> |
| Data Source ADSS | [https://www.icpsr.umich.edu/web/NAHDAP/studies/3088#](https://www.icpsr.umich.edu/web/NAHDAP/studies/3088) |
| Data Source DAWN | <https://www.icpsr.umich.edu/web/NAHDAP/studies/34565/versions/V3> |

# Data Pipeline Intent

The intent of this project is to develop a robust and secure data pipeline architecture capable of aggregating, analyzing, and visualizing critical substance abuse data from multiple, authoritative public health sources. The primary audience includes public health officials, policymakers, healthcare providers, and researchers aiming to understand, monitor, and mitigate the impact of substance abuse on different demographics. The project aims to deliver actionable insights through real-time analytics and visualizations, empowering decision-makers to proactively respond to trends and allocate resources effectively.

The pipeline integrates three distinct data sources, each contributing unique perspectives on substance abuse:

Drug Abuse Warning Network (DAWN) provides nationally representative hospital emergency department data on drug-related incidents, drug types, patient demographics, diagnoses, and visit outcomes, serving as an early warning for emerging drug abuse trends.

Youth Risk Behavior Surveillance System (YRBSS) offers survey-based insights into risky behaviors among adolescents and young adults, emphasizing substance abuse patterns and preventive strategies.

Alcohol and Drug Services Study (ADSS - ICPSR) delivers detailed information on drug treatment services, including facility characteristics, client profiles, treatment modalities, and effectiveness, to enhance substance abuse intervention planning

By securely integrating these diverse datasets into a centralized Azure MySQL database, utilizing SSL encryption, and automating transformations via Python scripts and stored procedures, the project ensures data integrity, confidentiality, and reliability. Finally, leveraging Power BI for intuitive visualizations makes complex datasets accessible, actionable, and impactful for informed policy-making and intervention planning.

Some potential shortcomings of this project include limited real-time ingestion capabilities, as the current pipeline primarily relies on periodic batch processing, which might hinder timely responsiveness to emerging drug abuse trends. Scalability might also become an issue with increased data volumes or additional sources, indicating a need for performance optimization such as indexing, partitioning, or migration to distributed data storage solutions. Adaptations for advanced data mining and predictive analytics could require further integration of machine learning frameworks or additional preprocessing to ensure dataset compatibility with such tools. Finally, the project currently lacks an API interface, which would be essential for securely serving the processed data to a broader research community, facilitating external collaborations, and supporting expanded analytical use cases.

**System Architecture** A diagram of a flowchart

AI-generated content may be incorrect.

Figure 1: System Architecture Overview

In the Data Tier, data from three sources—the Drug Abuse Warning Network (DAWN), Youth Risk Behavior Surveillance System (YRBSS), and Alcohol and Drug Services Study (ADSS)—are securely extracted and centralized into an Azure Database for MySQL Flexible Server using SSL certificates for encryption. The Application Logic Tier uses Python scripts developed in Visual Studio Code to automate the extraction, transformation, and loading (ETL) of data, while MySQL stored procedures facilitate additional data cleansing, transformation, and aggregation tasks. In the Presentation Tier, Power BI securely accesses the processed data via SSL-encrypted connections to provide interactive dashboards, visualizations, and reports, enabling stakeholders to effectively monitor and respond to substance abuse patterns in near real-time.

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

•

Data Sources:

**Drug Abuse Warning Network (DAWN)**: Provides nationally representative data collected from hospital emergency departments about drug-related incidents, capturing details such as drug types, patient demographics, diagnosis, and visit outcomes. It serves as a vital early warning system for tracking emerging drug abuse patterns, overdoses, adverse reactions, and drug-related suicides or accidental poisonings.

Focuses specifically on adolescents and young adults, providing extensive survey data regarding risky behaviors—including substance use and abuse, alcohol consumption, illicit drug use, and associated risk factors. This dataset offers critical insights into preventive strategies targeting younger populations.

**Alcohol and Drug Services Study (ADSS - ICPSR)**: Supplies comprehensive data related to drug treatment services, including facility characteristics, client profiles, types of treatments provided, and treatment outcomes. This source helps identify the effectiveness and accessibility of substance abuse treatment programs, aiding stakeholders in refining intervention strategies and treatment availability.

Table 2: 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. |  |
|  |
|  |

# 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 Source 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 3: 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.

Describe how data will be extracted from the MySQL database.

▪

Explain the transformations that will be applied using Pandas.

▪ TSV TO CSV USING THE TERMINAL

Describe how the transformed data will be loaded into the MYSQL 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.

◦

# Pipeline Output

* You must document the output of your pipeline.
* Why is the output useful?
* Architecture Diagram
* A Diagram of the final schema that was used must be present
* You should document the gaps in your system.
* Will the system scale as the dataset size grows?
* Is the system secure?
* Is the system extensible?

# Shortcomings of the current system.

◦

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 Azure MySQL Database

.

# 4. Running the ETL Pipeline

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

# DAWN PREPROCESSING

Is the dawn table indexed?

A screenshot of a computer

AI-generated content may be incorrect.

Table 4: DAWN Data Table Index

|  |  |
| --- | --- |
| **Index Feature** | **Meaning** |
| **Table** | The name of the table (er\_data). |
| **Non\_unique** | 0 means this index is **unique**, meaning values in the indexed column (caseid) **must be unique**. |
| **Key\_name** | PRIMARY means this is the **primary key index** (every row has a unique caseid). |
| **Seq\_in\_index** | 1 means caseid is the **first (and only) column** in this index. |
| **Column\_name** | caseid is the indexed column. |
| **Collation** | A (Ascending order)—how the index is stored internally. |
| **Cardinality** | 4334 is an **estimate of unique values** in caseid. Higher values = better index performance. |
| **Sub\_part** | NULL means the full column is indexed (not a prefix index). |
| **Packed** | NULL means the index **is not compressed**. |
| **Null** | Empty means caseid **cannot be NULL** (because it's a primary key). |
| **Index\_type** | BTREE means the index uses a **B-tree structure**, the default and most efficient for searching. |
| **Comment** | Empty—no special properties. |
| **Index\_comment** | Empty—no extra details. |
| **Visible** | YES means the index is **active and being used** by queries. |
| **Expression** | NULL means this is a normal index, not an expression-based index. |

NULL VALUES (-7)

A screenshot of a computer

AI-generated content may be incorrect.

-- Get\_Annual\_Drug\_Cases\_Per\_DrugCode

DELIMITER $$

CREATE PROCEDURE Get\_Annual\_Drug\_Cases\_Per\_DrugCode()

BEGIN

SELECT MIN(caseid) AS sample\_caseid,

MIN(metro) AS metro\_area,

MIN(agecat) AS age\_category,

MIN(sex) AS gender,

sdled\_3\_1 AS drug\_code,

COUNT(\*) AS total\_cases

FROM dawn.er\_data

Table 5: MySQL Output: Total Cases per Drug Code

A screenshot of a table

AI-generated content may be incorrect.

GROUP BY sdled\_3\_1

ORDER BY total\_cases DESC;

END $$

DELIMITER ;

CALL Get\_Annual\_Drug\_Cases\_Per\_DrugCode();

# 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.

# Database Performance Monitoring:

Azure Metrics: [link](https://portal.azure.com/#@darrenchenoutlook.onmicrosoft.com/blade/Microsoft_Azure_MonitoringMetrics/Metrics.ReactView/Referer/MetricsExplorer/ResourceId/%2Fsubscriptions%2F464a0109-8f64-4735-bee3-522cbdb76f9b%2FresourceGroups%2FADS507%2Fproviders%2FMicrosoft.DBforMySQL%2FflexibleServers%2Fmysqldchen/TimeContext/%7B%22relative%22%3A%7B%22duration%22%3A604800000%7D%2C%22showUTCTime%22%3Afalse%2C%22grain%22%3A1%7D/ChartDefinition/)A screenshot of a computer

AI-generated content may be incorrect.

MySQL Workbench – Administration / Performance ReportsA screenshot of a computer

AI-generated content may be incorrect.

Statement Statistics

Statement Analysis:

A screenshot of a computer

AI-generated content may be incorrect.

Linting:

Linting involves analyzing source code to flag programming errors, bugs, stylistic errors, and suspicious constructs. Linters are tools that perform this analysis, helping developers adhere to coding standards and identify potential issues early in the development process. In VS Code, linting is typically facilitated through extensions specific to the programming language in use. For this project, the VS Code Extension Pylint was integrated to provide linting capabilities. ​A screenshot of a computer

AI-generated content may be incorrect.

A black screen with white and blue lines

AI-generated content may be incorrect.

Figure 2 VS Code Extension: PyLint Report

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure : VS Code Extension - MarkdownLint

# Future Considerations & System Gaps

Potential improvements:

▪

**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.

Dashboard Deployment:

The current Tableau dashboard serves as a prototype. A production deployment would require either a Tableau Professional account or a

custom-built web application for greater flexibility and control.

Monitoring & Alerts:

The system could benefit from automated email notifications during the ELT process. This would enable faster response times to failures

and minimize dashboard downtime.

Pipeline Extensibility:

The current pipeline is easily adaptable to new data sources without major modifications. However, structural changes to the existing API

or new data fields would require minor adjustments at the transformation stage.

Security Considerations:

A public read-only access account allows Tableau to extract data and provides controlled access to certain users.

Write/modification access is secured via user authentication.

For large-scale deployment, SSL encryption and restricted IP access should be considered to enhance security.

Database Optimization:

The current MySQL database is semi-relational, primarily linking data through shared road attributes. Future iterations could benefit from

better normalization to strengthen relationships and improve query efficiency.

Scalability & Storage Costs:

The pipeline is currently designed to handle increasing data volume, but storage costs must be evaluated as historical data grows. Future

iterations may consider partitioning or archiving strategies to manage long-term data storage efficiently.

# Development Log

[ADS\_507\_Tracker.xlsx](https://1drv.ms/x/s!AtYkM_aIvmqSg9sS-7nmUcNLxz3b9w?e=kRqjZS&nav=MTVfezJGQ0MxRjAzLUYyMkMtNDEyRS05QkZELUUxNkM2ODhCMDZBN30)

# **References**

Beaulie, A. (2020). Learning SQL. O'Reilly Media.

Microsoft (n.d.) Azure. Retrieved from https://azure.microsoft.com/en-us/

Reis, J., & Housley, M. (2022). Fundamentals of data engineering. O'Reilly Media

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>

# 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

### ATTEMPTED AIRFLOW ON AZURE DATA FACTORY / MS FABRIC

AirFlow on Azure Data Factory

A screenshot of a computer

AI-generated content may be incorrect.

Airflow Environment migrated to MS Fabric, beta testers only,

Airflow on Azure Virtual Machine?

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

If I were to successfully run AIRFLOW on Azure or another host machine, I would use it to automate the