HHS Hospital Data Pipeline

Team Dancer
December 4, 2025

Engineering Team

Ryan Logue

Data Engineer

CMU MADS 2026

Jay Louissaint

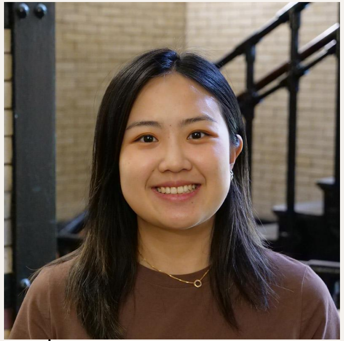
Data Engineer

CMU MADS 2026

Joanne Li

Data Engineer

CMU MADS 2026



Jasmine Kwok

Data Engineer

CMU MADS 2026

Agenda

1. Data Background

2. Database Design

3. Design Choices

4. Error Handling

5. Future Adjustments & Computation Time

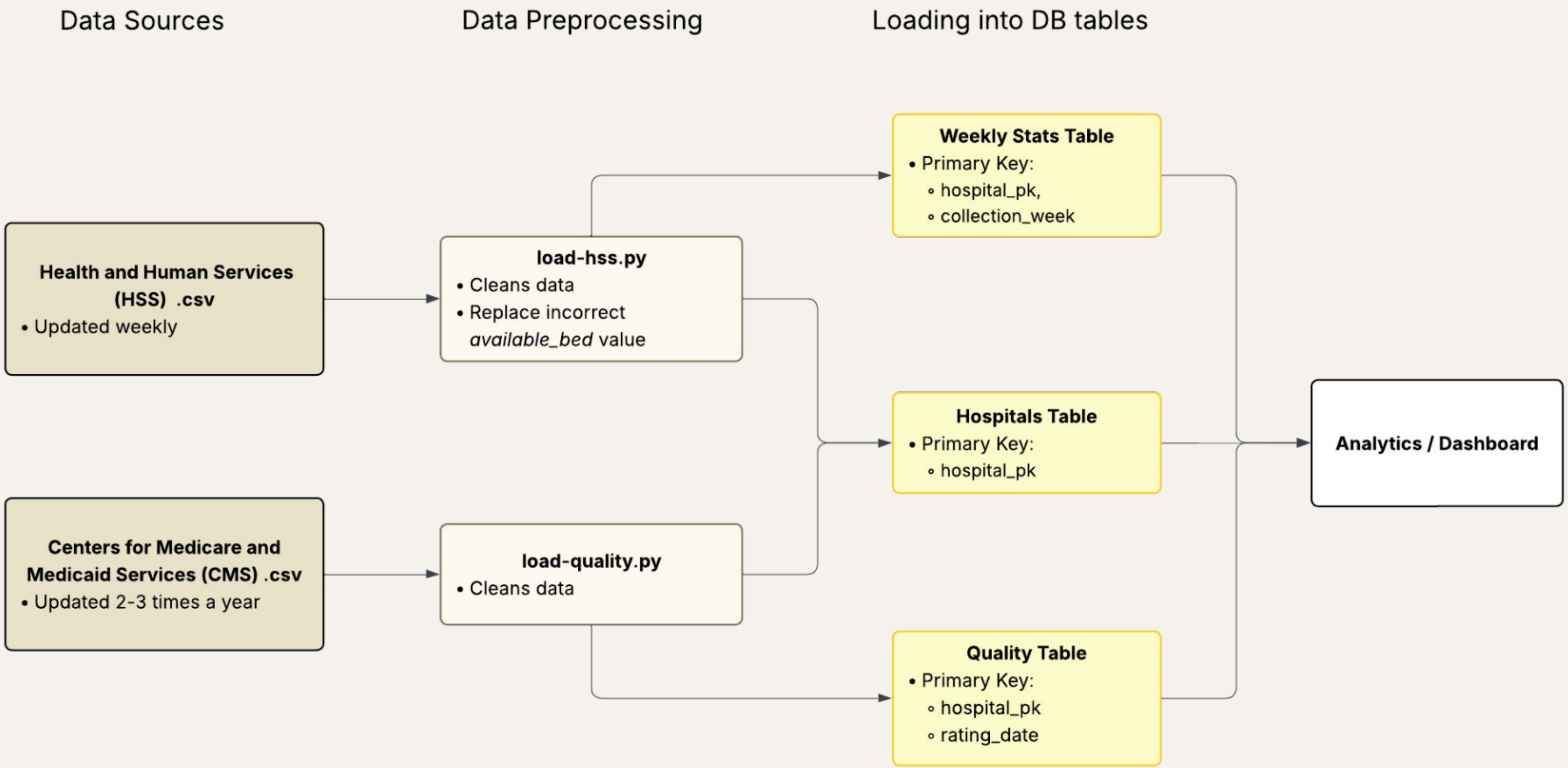
Executive Overview

Purpose	Automated Python → PostgreSQL pipeline for HHS weekly stats and CMS quality ratings
Core Components	Scripts: load-hhs.py (weekly data), load-quality.py (CMS ratings) Key features: cleaning, validation, rollback on error
Key Capabilities	Auto-insert/update hospital metadata; maintains historical ratings
Error Handling	Enforces constraints: bed counts, coordinates, rating ranges, state codes
Workflow	Initial load → weekly updates → periodic CMS rating updates
Outcome	Reliable, scalable, high-integrity hospital data pipeline

Project & Data Background

- Team Dasher → Team Dancer
- Two data sources:
  1. US Department of Health and Human Services (HHS)
     a. Contains weekly hospital-level information on location, bed availability and usage, and COVID-19 patient counts
     b. Recorded Sunday to Saturday – collection_week date is the Sunday it starts on
  2. Centers for Medicare and Medicaid Services (CMS)
     a. Contains hospital identifiers, location, ownership details and time-tracked quality ratings
     b. Data collection period varies, 2-3 times per year

Database Design and Workflow



Data Sources

Data Preprocessing

Loading into DB tables

Health and Human Services (HSS) .csv
- Updated weekly

Centers for Medicare and Medicaid Services (CMS) .csv
- Updated 2-3 times a year

load-hss.py
- Cleans data
- Replace incorrect *available_bed* value

load-quality.py
- Cleans data

Weekly Stats Table
- Primary Key:
  - hospital_pk,
  - collection_week

Hospitals Table
- Primary Key:
  - hospital_pk

Quality Table
- Primary Key:
  - hospital_pk
  - rating_date

Analytics / Dashboard

Design Choices

- Separate files for loading HHS and CMS data
  - CMS Hospital Quality
    - Modular cleaning functionality: blanks / NA's → "None", invalid values → "None"
  - HHS Weekly Data
    - Modular cleaning functionality: empty/NA → "None", no special characters, -999 → "None"
- Panda tables for easier cleaning, renaming, and data insertion
- UPSERT is utilized by both programs
  - ensure there are no duplicate hospitals
- All or nothing data loads
  - avoids any discrepancies or duplications for time varying statistics
- STAR schema format
- Error logs for both loaders

Error Handling

- Errors are logged in .txt files for tracking purposes

  load_hss_error_log.txt
  load_quality_error_log.txt

Source-related Errors

→ Program stops

1. File not found:

   Check CSV file path
2. Invalid date format

   Ensure YYYY-MM-DD format

Database-related Errors

→ Rollback, pinpoint problematic row

1. Foreign key violation

   Ensure *hospitals* is up to date
2. Constraint violation

   Fix problematic row

Future Adjustments

- Separate functions: parse → clean → insert
  - Can easily extend to new sources
- If new columns are added or there are different naming conventions
  - Update processing logic and create new column
  - Reusable data cleaning functions
- If the dataset grows larger (more rows)
  - Minimal changes - current pipeline uses batch inserts
  - Current computation time is 3 seconds for around 5000 rows for both HHS and CMS
- If there is new data that we would like to track
  - New table needs to be created if data does not fit into the existing tables
  - Use foreign keys to link to existing tables

Thank You!

Questions

Appendix

- Resource Links :
  - Github - HHS ETL Pipeline - https://github.com/Vrajmp1/Dasher_DEDE
  - Centers for Medicare and Medicaid Services (CMS) Data - https://data.cms.gov/provider-data/dataset/xubh-q36u#data-dictionary
  - US Department of Health and Human Services (HHS) Data - https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/uqq2-txqb/about_data