Raw Data

Census Business Data EPA Data Asthma Data

ETL

Census Data - extract with API, transform, store EPA Data - extract CA county info with API, transform Asthma Data -Download csv file of CA data, save to DataLake, Transform Link EPA CA data to Asthma CA Data and store in two SQL tables to use for ML training EPA Data - extract US county in 2021, transform, store

Streaming Data
Producer -- Pull EPA US County data from storage, randomly select, feed into Kafka

Storage

Data Lake:

Asthma Data

SQL Tables: EPA CA data by county, CA Asthma rates by county Process
For Producer - Pull in
2021 air quality data
for random county,
feed to Kafka
For Consumer - Take
data from Kafka and
process to prepare for
sending to
PowerBI/ML
prediction

Enrich

Train ML model w/ Asthma + Air Quality Data

Linear Regression

Store ML Model in Data Lake

PowerBI Dashboard

Import ML model Import data from Kafka with selected county and Air Quality date Pull in matching census data for that county Run ML prediction with AQ data from Kafka to show predictions of Asthma-related ED visits for that county

Serve

Users

Local Health Professionals

People w/ Asthma can assess daily risk for their county

- 1. Extract CSV files and store in data lake
- 2. Create databrick and establish mountpoint to clean and transform CSV data
- 3. Connect to APIs, read relevant data into dataframe, clean and transform data
- 4. Create databricks for Kafka producer and consumer, produce + consume Kafka messages to stream EPA API data
- 5. Construct SQL tables and read data from Azure blobs to SQL database
- 6. Set up data factory to trigger producer + consumer activity / create pipeline to update data in SQL
- 7. Create databrick and develop linear regression ML model using asthma + historic air quality data
- 8. Save ML model and dump into PowerBI / implement on EPA API data to get daily predictions
- 9. Create PowerBI dashboard incorporating above ML model + historic data visuals

