San Francisco 311 Requests
Data Streaming ETL Pipeline

Project Overview

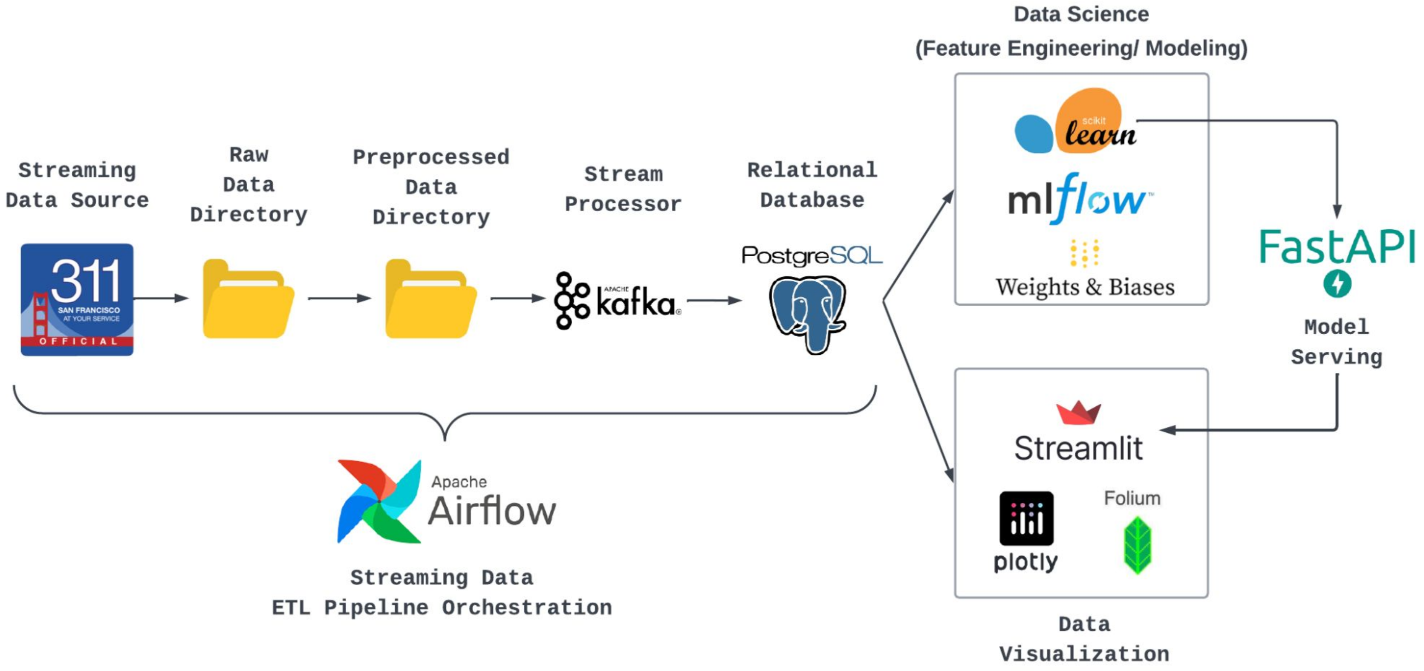
Summary

San Francisco faces the challenge of efficiently managing and addressing the diverse array of incidents reported through its **311 system**. With a continuous influx of cases, there is a pressing need to **streamline the allocation of resources** and **enhance response times to improve service quality**.

Objectives

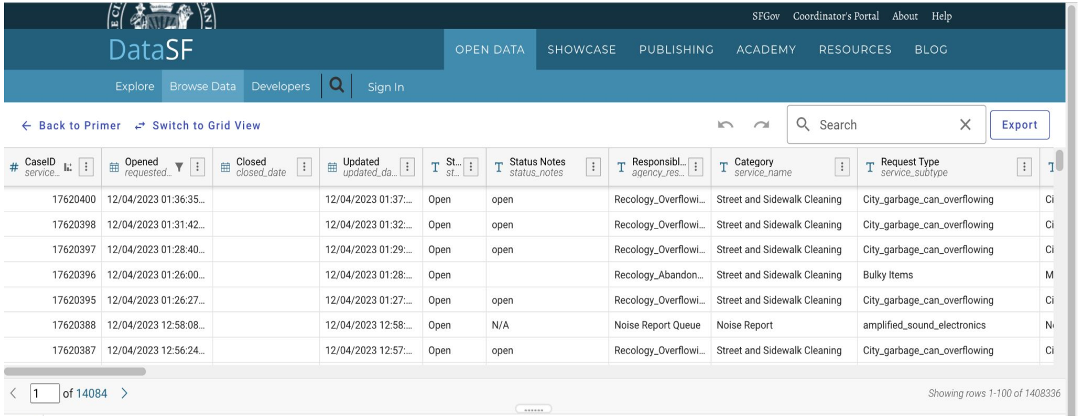
1. **Streaming Data ETL Pipeline**: Leveraging skills from Data Streaming course to build a real-time streaming data pipeline.
2. **Data Visualization:** Mapping 311 cases across the city to identify hotspots and patterns.
3. **Closure Time Prediction**: Using machine learning to predict case closure times, considering factors like case type and location.

Architecture Diagram



Dataset

- Open source dataset managed by San Francisco Government
- **Dataset Creation:** October, 2011
- **Update frequency:** Daily (multiple times per hour)
- **Dataset Size:** 2.2 GB
- **Number of Records:** 6.6 M
- **Features:** 48 (interpretable: 15)



CaseID service...	Opened requested...	Closed closed_date	Updated updated_da...	St... st...	Status Notes status_notes	Responsibl... agency_res...	Category service_name	Request Type service_subtype	
17620400	12/04/2023 01:36:35...		12/04/2023 01:37:...	Open	open	Recology_Overflowi...	Street and Sidewalk Cleaning	City_garbage_can_overflowing	Ci
17620398	12/04/2023 01:31:42...		12/04/2023 01:32:...	Open	open	Recology_Overflowi...	Street and Sidewalk Cleaning	City_garbage_can_overflowing	Ci
17620397	12/04/2023 01:28:40...		12/04/2023 01:29:...	Open	open	Recology_Overflowi...	Street and Sidewalk Cleaning	City_garbage_can_overflowing	Ci
17620396	12/04/2023 01:26:00...		12/04/2023 01:28:...	Open		Recology_Abandon...	Street and Sidewalk Cleaning	Bulky Items	M
17620395	12/04/2023 01:26:27...		12/04/2023 01:27:...	Open	open	Recology_Overflowi...	Street and Sidewalk Cleaning	City_garbage_can_overflowing	Ci
17620388	12/04/2023 12:58:08...		12/04/2023 12:58:...	Open	N/A	Noise Report Queue	Noise Report	amplified_sound_electronics	N
17620387	12/04/2023 12:56:24...		12/04/2023 12:57:...	Open	open	Recology_Overflowi...	Street and Sidewalk Cleaning	City_garbage_can_overflowing	Ci

1 of 14084

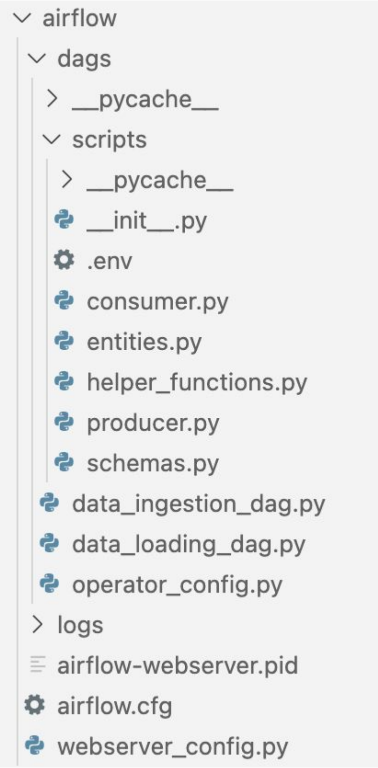
Showing rows 1-100 of 1408336

Source: https://data.sfgov.org/City-Infrastructure/311-Cases/vw6y-z8j6/explore

Tools & Packages

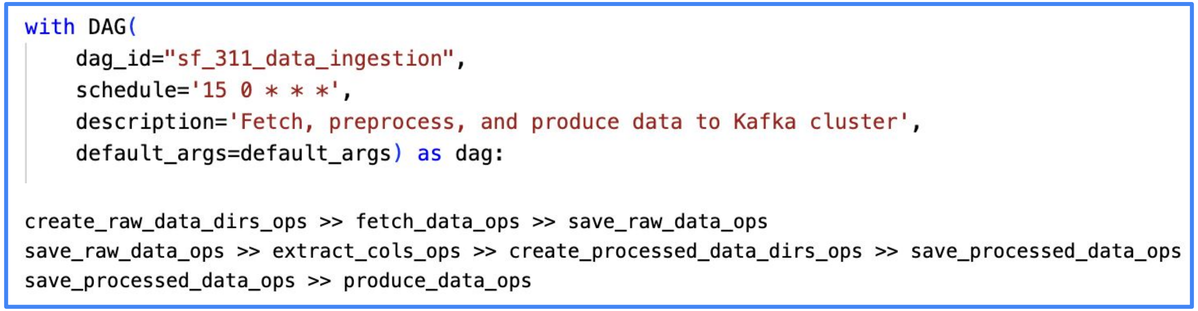
- **Data Orchestration:** apache-airflow, sodapy
- **Data Streaming:** avro, confluent_kafka, fastavro
- **Database:** pgAdmin, psycopg2, sqlalchemy
- **Modeling & Serving:** fastapi[all], mlflow, scikit-learn
- **Data Visualization:** geopandas, folium, numpy, pandas, plotly, streamlit
- **Version Control:** git, Github

Airflow - ETL Pipeline Orchestration



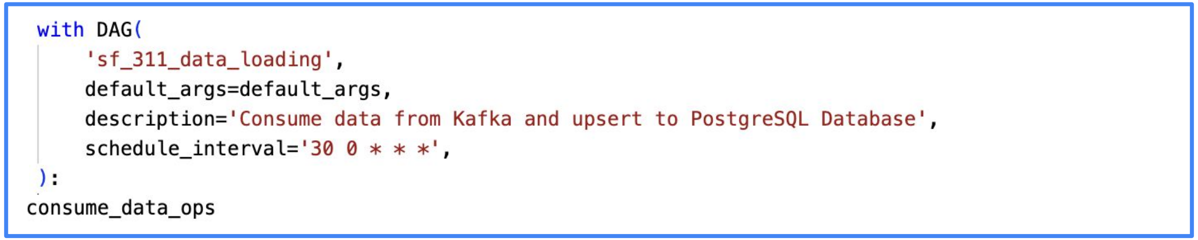
```
airflow
  dags
    __pycache__
    scripts
      __pycache__
      __init__.py
      .env
      consumer.py
      entities.py
      helper_functions.py
      producer.py
      schemas.py
    data_ingestion_dag.py
    data_loading_dag.py
    operator_config.py
  logs
  airflow-webserver.pid
  airflow.cfg
  webserver_config.py
```

DAG 1



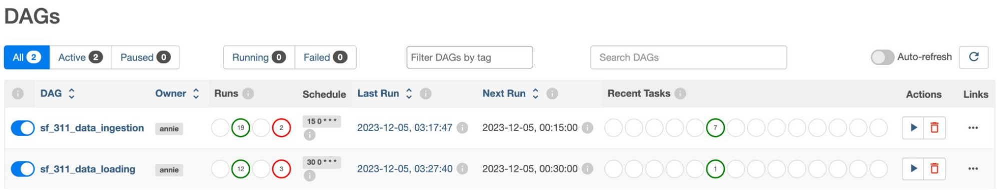
```
with DAG(
    dag_id="sf_311_data_ingestion",
    schedule='15 0 * * *',
    description='Fetch, preprocess, and produce data to Kafka cluster',
    default_args=default_args) as dag:

create_raw_data_dirs_ops >> fetch_data_ops >> save_raw_data_ops
save_raw_data_ops >> extract_cols_ops >> create_processed_data_dirs_ops >> save_processed_data_ops
save_processed_data_ops >> produce_data_ops
```

DAG 2



```
with DAG(
    'sf_311_data_loading',
    default_args=default_args,
    description='Consume data from Kafka and upsert to PostgreSQL Database',
    schedule_interval='30 0 * * *',
):
consume_data_ops
```

Airflow - ETL Pipeline Orchestration

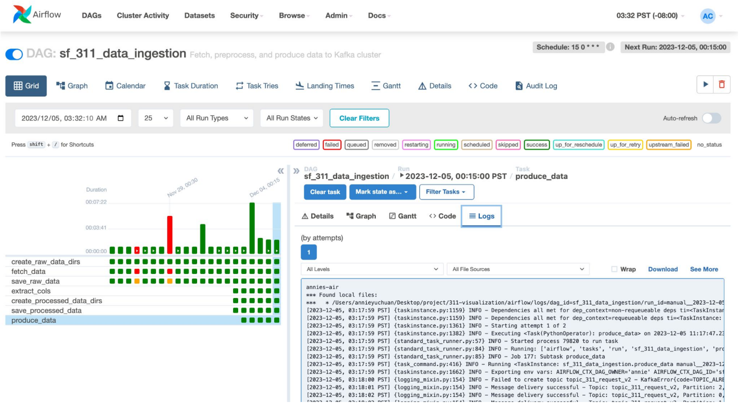


DAGs

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
sf_311_data_ingestion	annie	19 2	15 0 * * *	2023-12-05, 03:17:47	2023-12-05, 00:15:00	7		•••
sf_311_data_loading	annie	12 3	30 0 * * *	2023-12-05, 03:27:40	2023-12-05, 00:30:00	1		•••

DAG 1 - Data Ingestion, Preprocessing, Produce

Schedule: every night at 00:15 AM



DAG: sf_311_data_ingestion Fetch, preprocess, and produce data to Kafka cluster

Schedule: 15 0 * * *

Next Run: 2023-12-05, 00:15:00

2023/12/05, 03:32:10 AM | 25 | All Run Types | All Run States

create_raw_data_dirs
fetch_data
save_raw_data
extract_cols
create_processed_data_dirs
save_processed_data
produce_data

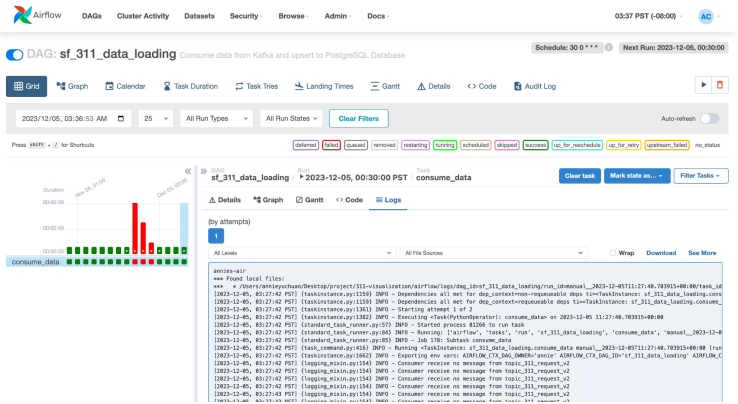
DAG sf_311_data_ingestion / Run 2023-12-05, 00:15:00 PST / Task produce_data

All Levels | All File Sources

```
annies-air
*** Found local files:
***   * /Users/annieyuchuan/Desktop/project/311-visualization/airflow/logs/dag_id=sf_311_data_ingestion/run_id=manual__2023-12-05
[2023-12-05, 03:17:59 PST] {taskinstance.py:1159} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstar
[2023-12-05, 03:17:59 PST] {taskinstance.py:1159} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance:
[2023-12-05, 03:17:59 PST] {taskinstance.py:1361} INFO - Starting attempt 1 of 2
[2023-12-05, 03:17:59 PST] {taskinstance.py:1382} INFO - Executing <Task(PythonOperator): produce_data> on 2023-12-05 11:17:47.23
[2023-12-05, 03:17:59 PST] {standard_task_runner.py:57} INFO - Started process 79820 to run task
[2023-12-05, 03:17:59 PST] {standard_task_runner.py:84} INFO - Running: ['airflow', 'tasks', 'run', 'sf_311_data_ingestion', 'pro
[2023-12-05, 03:17:59 PST] {standard_task_runner.py:85} INFO - Job 177: Subtask produce_data
[2023-12-05, 03:17:59 PST] {task_command.py:416} INFO - Running <TaskInstance: sf_311_data_ingestion.produce_data manual__2023-12
[2023-12-05, 03:17:59 PST] {taskinstance.py:1662} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='annie' AIRFLOW_CTX_DAG_ID='sf
[2023-12-05, 03:18:00 PST] {logging_mixin.py:154} INFO - Failed to create topic topic_311_request_v2 - KafkaError{code=TOPIC_ALRE
[2023-12-05, 03:18:01 PST] {logging_mixin.py:154} INFO - Message delivery successful - Topic: topic_311_request_v2, Partition: 2,
[2023-12-05, 03:18:02 PST] {logging_mixin.py:154} INFO - Message delivery successful - Topic: topic_311_request_v2, Partition: 0,
```

DAG 2 - Consume to Database

Schedule: every night at 00:30 AM



DAG: sf_311_data_loading Consume data from Kafka and upsert to PostgreSQL Database

Schedule: 30 0 * * *

Next Run: 2023-12-05, 00:30:00

2023/12/05, 03:36:53 AM | 25 | All Run Types | All Run States

consume_data

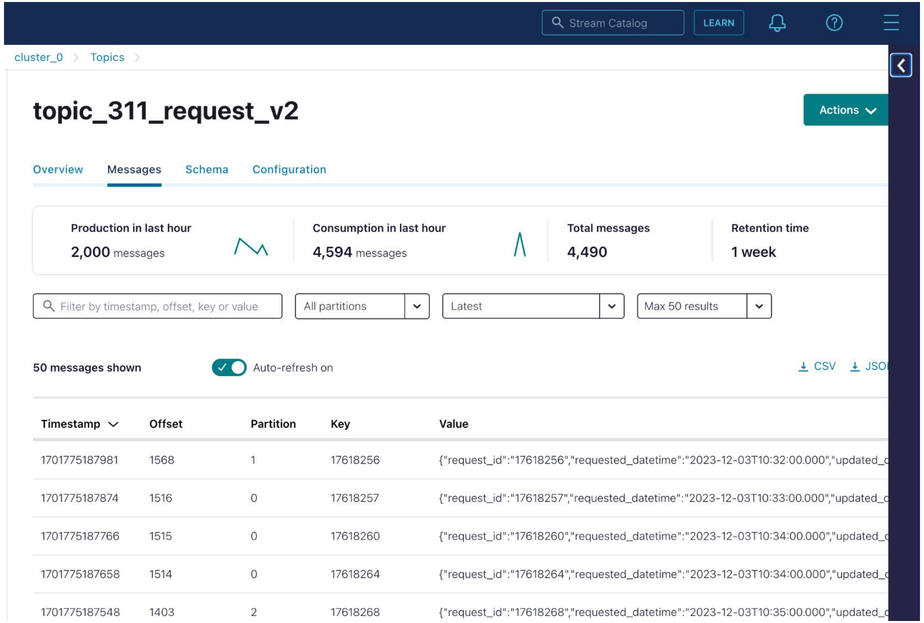
DAG sf_311_data_loading / Run 2023-12-05, 00:30:00 PST / Task consume_data

All Levels | All File Sources

```
annies-air
*** Found local files:
***   * /Users/annieyuchuan/Desktop/project/311-visualization/airflow/logs/dag_id=sf_311_data_loading/run_id=manual__2023-12-05T11:27:40.703915+00:00/task_id
[2023-12-05, 03:27:42 PST] {taskinstance.py:1159} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: sf_311_data_loading.cons
[2023-12-05, 03:27:42 PST] {taskinstance.py:1159} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: sf_311_data_loading.consume_
[2023-12-05, 03:27:42 PST] {taskinstance.py:1361} INFO - Starting attempt 1 of 2
[2023-12-05, 03:27:42 PST] {taskinstance.py:1382} INFO - Executing <Task(PythonOperator): consume_data> on 2023-12-05 11:27:40.703915+00:00
[2023-12-05, 03:27:42 PST] {standard_task_runner.py:57} INFO - Started process 81266 to run task
[2023-12-05, 03:27:42 PST] {standard_task_runner.py:84} INFO - Running: ['airflow', 'tasks', 'run', 'sf_311_data_loading', 'consume_data', 'manual__2023-12-0
[2023-12-05, 03:27:42 PST] {standard_task_runner.py:85} INFO - Job 178: Subtask consume_data
[2023-12-05, 03:27:42 PST] {task_command.py:416} INFO - Running <TaskInstance: sf_311_data_loading.consume_data manual__2023-12-05T11:27:40.703915+00:00 [run
[2023-12-05, 03:27:42 PST] {taskinstance.py:1662} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='annie' AIRFLOW_CTX_DAG_ID='sf_311_data_loading' AIRFLOW_C
[2023-12-05, 03:27:42 PST] {logging_mixin.py:154} INFO - Consumer receive no message from topic_311_request_v2
[2023-12-05, 03:27:42 PST] {logging_mixin.py:154} INFO - Consumer receive no message from topic_311_request_v2
[2023-12-05, 03:27:42 PST] {logging_mixin.py:154} INFO - Consumer receive no message from topic_311_request_v2
[2023-12-05, 03:27:42 PST] {logging_mixin.py:154} INFO - Consumer receive no message from topic_311_request_v2
[2023-12-05, 03:27:43 PST] {logging_mixin.py:154} INFO - Consumer receive no message from topic_311_request_v2
```

Kafka - Data Streaming

Kafka Topic



topic_311_request_v2

Production in last hour
2,000 messages

Consumption in last hour
4,594 messages

Total messages
4,490

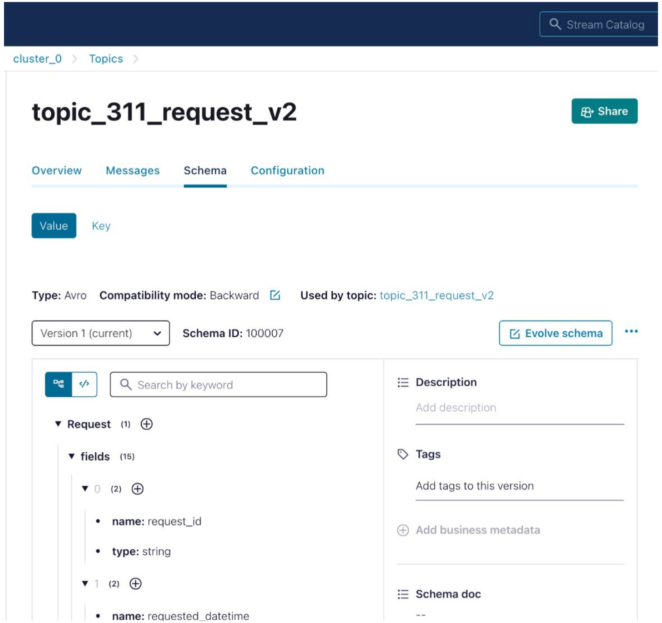
Retention time
1 week

50 messages shown

Auto-refresh on

Timestamp	Offset	Partition	Key	Value
1701775187981	1568	1	17618256	{"request_id":"17618256","requested_datetime":"2023-12-03T10:32:00.000","updated_
1701775187874	1516	0	17618257	{"request_id":"17618257","requested_datetime":"2023-12-03T10:33:00.000","updated_
1701775187766	1515	0	17618260	{"request_id":"17618260","requested_datetime":"2023-12-03T10:34:00.000","updated_
1701775187658	1514	0	17618264	{"request_id":"17618264","requested_datetime":"2023-12-03T10:34:00.000","updated_
1701775187548	1403	2	17618268	{"request_id":"17618268","requested_datetime":"2023-12-03T10:35:00.000","updated_

Kafka Schema Registry - Avro Serializer



topic_311_request_v2

Type: Avro Compatibility mode: Backward Used by topic: topic_311_request_v2

Version 1 (current) Schema ID: 100007

- Request (1)
  - fields (15)
    - 0 (2)
      - name: request_id
      - type: string
    - 1 (2)
      - name: requested_datetime

Description

Add description

Tags

Add tags to this version

Add business metadata

Schema doc

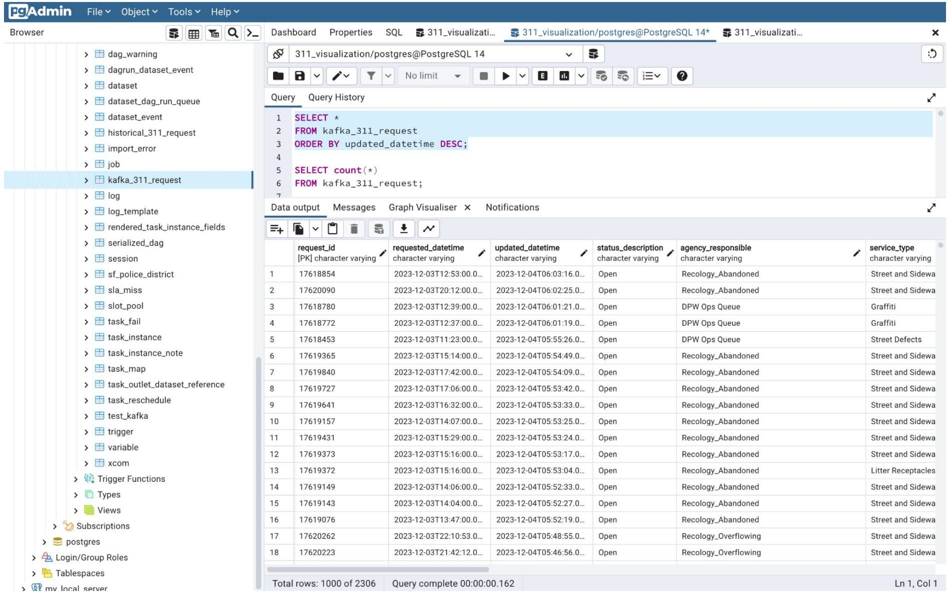
--

PostgreSQL Database

SQL Table Creation

```sql
CREATE TABLE IF NOT EXISTS historical_311_request
(
    request_id character varying PRIMARY KEY NOT NULL,
    requested_datetime character varying NOT NULL,
    updated_datetime character varying NOT NULL,
    status_description character varying,
    agency_responsible character varying,
    service_type character varying,
    service_subtype character varying,
    address character varying,
    street character varying,
    supervisor_district character varying,
    neighborhood character varying,
    police_district character varying,
    latitude double precision,
    longitude double precision,
    source character varying
);
```

pgAdmin - PostgreSQL Tool



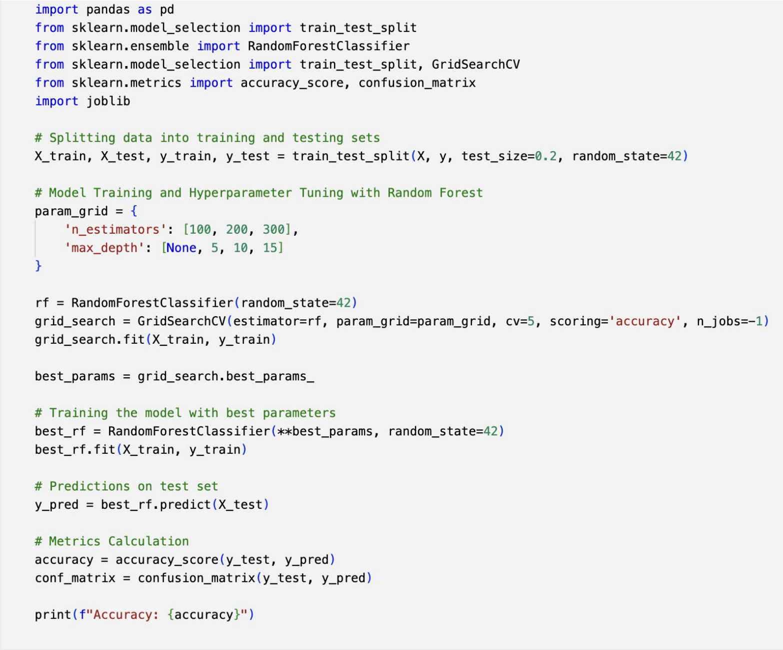
```sql
SELECT *
FROM kafka_311_request
ORDER BY updated_datetime DESC;

SELECT count(*)
FROM kafka_311_request;
```

	request_id [PK] character varying	requested_datetime character varying	updated_datetime character varying	status_description character varying	agency_responsible character varying	service_type character varying
1	17618854	2023-12-03T12:53:00.0...	2023-12-04T06:03:16.0...	Open	Recology_Abandoned	Street and Sidewa
2	17620090	2023-12-03T20:12:00.0...	2023-12-04T06:02:25.0...	Open	Recology_Abandoned	Street and Sidewa
3	17618780	2023-12-03T12:39:00.0...	2023-12-04T06:01:21.0...	Open	DPW Ops Queue	Graffiti
4	17618772	2023-12-03T12:37:00.0...	2023-12-04T06:01:19.0...	Open	DPW Ops Queue	Graffiti
5	17618453	2023-12-03T11:23:00.0...	2023-12-04T05:55:26.0...	Open	DPW Ops Queue	Street Defects
6	17619365	2023-12-03T15:14:00.0...	2023-12-04T05:54:49.0...	Open	Recology_Abandoned	Street and Sidewa
7	17619840	2023-12-03T17:42:00.0...	2023-12-04T05:54:09.0...	Open	Recology_Abandoned	Street and Sidewa
8	17619727	2023-12-03T17:06:00.0...	2023-12-04T05:53:42.0...	Open	Recology_Abandoned	Street and Sidewa
9	17619641	2023-12-03T16:32:00.0...	2023-12-04T05:53:33.0...	Open	Recology_Abandoned	Street and Sidewa
10	17619157	2023-12-03T14:07:00.0...	2023-12-04T05:53:25.0...	Open	Recology_Abandoned	Street and Sidewa
11	17619431	2023-12-03T15:29:00.0...	2023-12-04T05:53:24.0...	Open	Recology_Abandoned	Street and Sidewa
12	17619373	2023-12-03T15:16:00.0...	2023-12-04T05:53:17.0...	Open	Recology_Abandoned	Street and Sidewa
13	17619372	2023-12-03T15:16:00.0...	2023-12-04T05:53:04.0...	Open	Recology_Abandoned	Litter Receptacles
14	17619149	2023-12-03T14:06:00.0...	2023-12-04T05:52:33.0...	Open	Recology_Abandoned	Street and Sidewa
15	17619143	2023-12-03T14:04:00.0...	2023-12-04T05:52:27.0...	Open	Recology_Abandoned	Street and Sidewa
16	17619076	2023-12-03T13:47:00.0...	2023-12-04T05:52:19.0...	Open	Recology_Abandoned	Street and Sidewa
17	17620262	2023-12-03T22:10:53.0...	2023-12-04T05:48:55.0...	Open	Recology_Overflowing	Street and Sidewa
18	17620223	2023-12-03T21:42:12.0...	2023-12-04T05:46:56.0...	Open	Recology_Overflowing	Street and Sidewa

Total rows: 1000 of 2306 Query complete 00:00:00.162 Ln 1, Col 1

Scikit-Learn/ FastAPI - Modeling & Serving

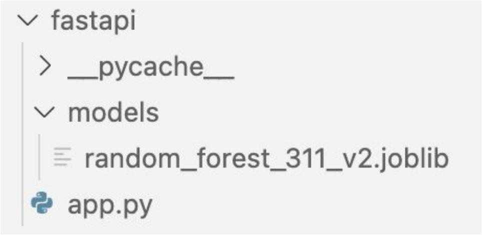
Scikit learn



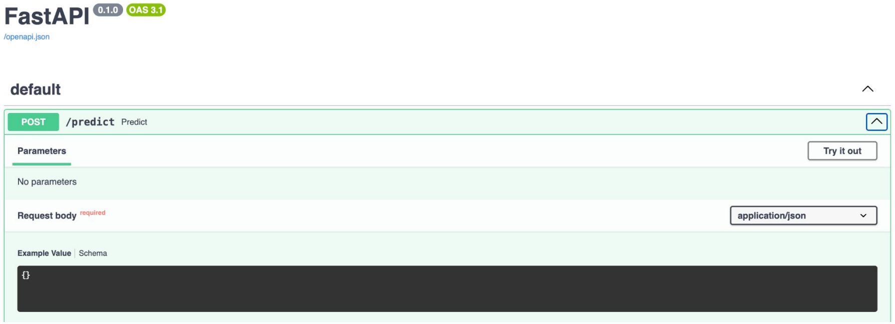
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import accuracy_score, confusion_matrix
import joblib

# Splitting data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Model Training and Hyperparameter Tuning with Random Forest
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 5, 10, 15]
}

rf = RandomForestClassifier(random_state=42)
grid_search = GridSearchCV(estimator=rf, param_grid=param_grid, cv=5, scoring='accuracy', n_jobs=-1)
grid_search.fit(X_train, y_train)

best_params = grid_search.best_params_

# Training the model with best parameters
best_rf = RandomForestClassifier(**best_params, random_state=42)
best_rf.fit(X_train, y_train)

# Predictions on test set
y_pred = best_rf.predict(X_test)

# Metrics Calculation
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.586056549770947
```

FastAPI



```
fastapi
  __pycache__
  models
    random_forest_311_v2.joblib
  app.py
```



No parameters

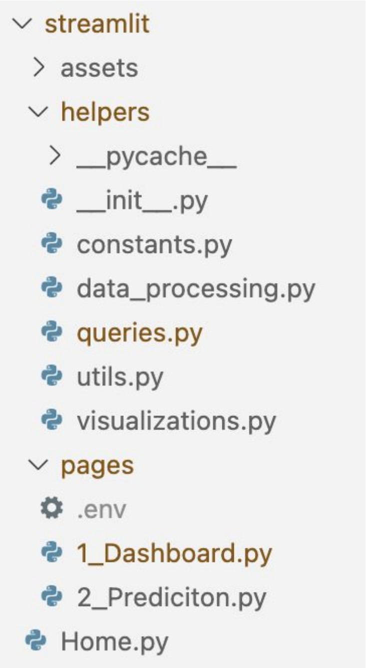
Request body required

application/json

Example Value | Schema

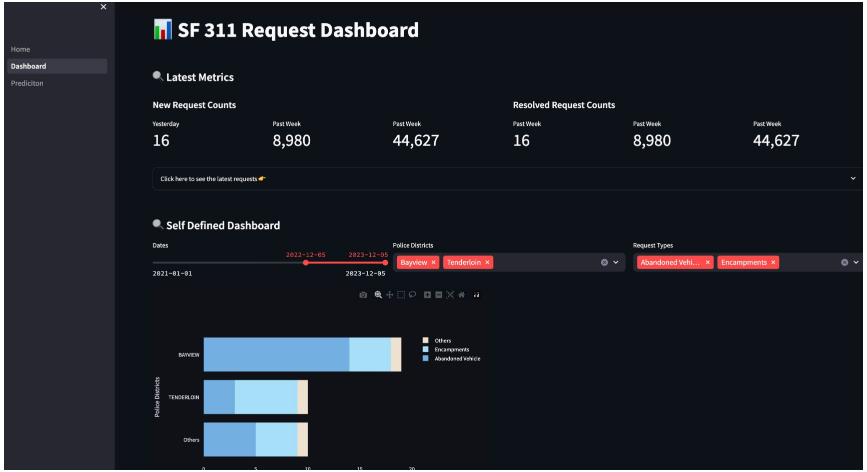
```
{}
```

Streamlit/ Plotly/ Folium - Visualization



```
streamlit
  assets
  helpers
    __pycache__
    __init__.py
    constants.py
    data_processing.py
    queries.py
    utils.py
    visualizations.py
  pages
    .env
    1_Dashboard.py
    2_Prediciton.py
  Home.py
```

Page 1 - Real Time Data Visualization Dashboard



📊 SF 311 Request Dashboard

🔍 Latest Metrics

New Request Counts

Yesterday 16

Past Week 8,980

Past Week 44,627

Resolved Request Counts

Past Week 16

Past Week 8,980

Past Week 44,627

Click here to see the latest requests 👉

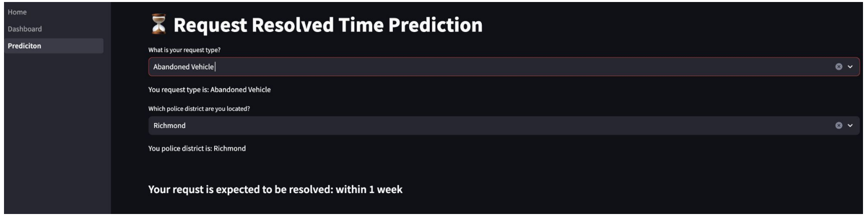
🔍 Self Defined Dashboard

Dates: 2022-12-05 – 2023-12-05 (2021-01-01 – 2023-12-05)

Police Districts: Bayview, Tenderloin

Request Types: Abandoned Vehi…, Encampments

Page 2 - Request Time Prediction



⏳ Request Resolved Time Prediction

What is your request type?

Abandoned Vehicle

You request type is: Abandoned Vehicle

Which police district are you located?

Richmond

You police district is: Richmond

Your requst is expected to be resolved: within 1 week

Demo