Boston 311 ETL Pipeline
Data Warehousing & Orchestration

Introduction

"311 is an easy-to-remember telephone number that connects you with highly-trained constituent service representatives. They are ready to help you with requests for non-emergency City services and information."

This project is focused on preparing raw 311 requests data from **Boston** for advanced analytics. The goal is to build a robust data pipeline that extracts, transforms, and loads (ETL) data into a **structured data warehouse**, enabling seamless reporting, visualization, and analytics.

The pipeline processes **millions of records** and leverages **PySpark for distributed data processing**, **Airflow for orchestration**, and **MySQL as the Data Warehouse**. This ensures that historical and real-time service request data is available for analysis and business insights.

Links

Dataset: Boston 311 | Boston.gov

GitHub: amey379/311_Request_Analysis

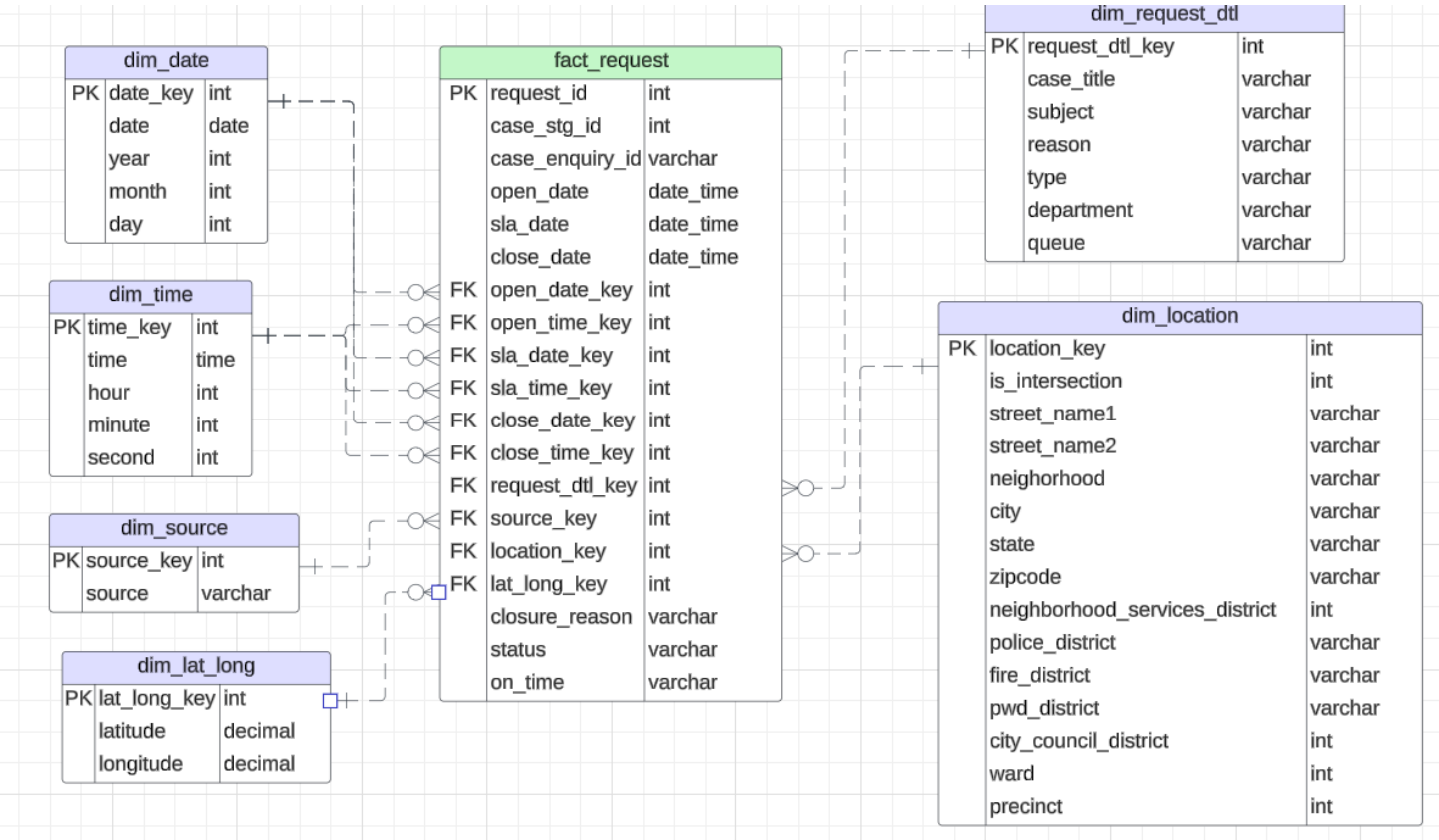
Tools

Databases & Data Processing: *MySQL – Data Warehouse for structured storage* *PySpark & Spark – Distributed data processing*	**Development & System:** *Python – Scripting for ETL processing* *Unix/Linux (WSL) – Environment setup & scripting*
Orchestration & Workflow Management: *Apache Airflow – Orchestrating ETL pipelines*	**Visualization & Reporting:** *Power BI – Dashboarding & data visualization*

Skills

- Dimensional Modeling & Data Warehousing
- ETL Development (Extract, Transform, Load)
- Distributed Data Processing with PySpark
- Workflow Automation using Airflow
- SQL Query Optimization & Performance Tuning
- Data Cleaning, Transformation & Enrichment
- Unix Shell Scripting
- Handling Large-Scale Datasets Efficiently

Dimensional Model



Fact Table:

- **Fact_311_Requests** – Stores key metrics like resolution time, case status, request type

Dimension Tables:

- **Dim_Date Dim_Time** – Date and time details for trend analysis
- **Dim_Location** – Geographic information of incidents
- **Dim_Request_Details** – Categories and types of 311 requests
- **Dim_Source** – Channel through which requests were received

This **dimensional model** enables efficient querying and analytical processing by minimizing redundancy while ensuring high performance.

Technical Implementation:

Data Ingestion & Storage

- Extracted raw 311 request data from CSV files stored in Unix filesystem (WSL)
- Stored data in MySQL tables for structured processing

ETL Pipeline with PySpark

- Data Cleaning & Transformation
  - Handled missing values, inconsistencies, and date conversions
  - Standardized categorical values (e.g., uppercase formatting)
  - Performed geo-coding & location mapping
- Optimized batch processing using Spark DataFrames

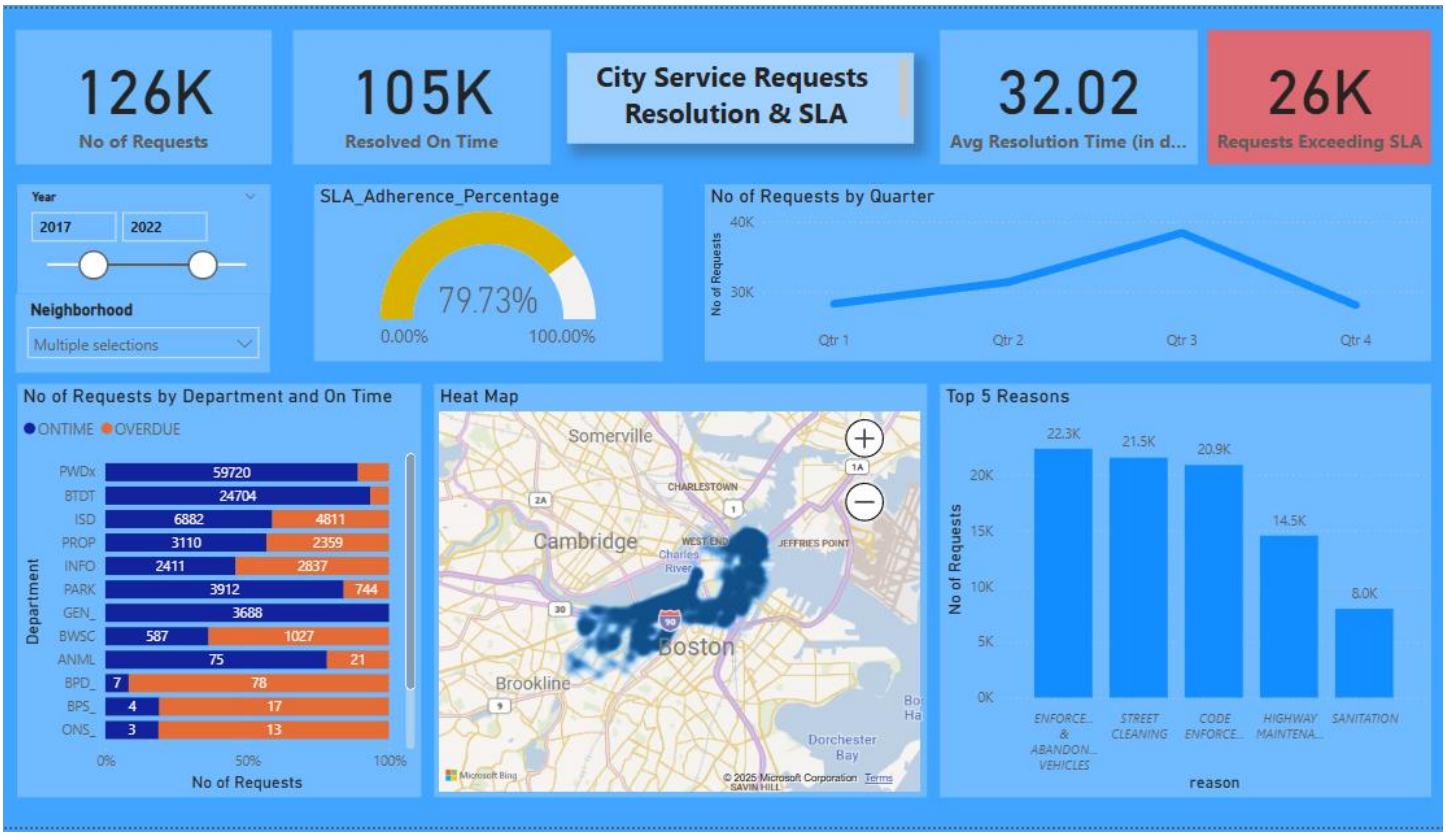
Orchestration with Apache Airflow

- Automated DAGs to schedule ETL tasks sequentially:
  - Stage Load – Raw ingestion to MySQL
  - Lookup Table Load – Enriching request data with reference tables
  - Dimensional Load – Populating dimension tables
  - Fact Load – Populating fact table for analytics
- Ensured task dependencies and error handling mechanisms

Performance Optimization

- Implemented batch processing for efficient large-scale data ingestion
- Used JDBC batch writes to MySQL for faster inserts
- Indexed key columns for query performance improvement

Data Visualization



SQL queries

Find the Top 5 Most Frequent Service Requests in the Last 5 Years

```sql
SELECT
    dr.reason AS request_reason,
    COUNT(f.request_id) AS total_requests
FROM fact_311_request f
JOIN dim_request_dtl dr ON f.request_dtl_key = dr.request_dtl_key
WHERE open_date >= DATE_SUB(CURDATE(), INTERVAL 5 YEAR)
GROUP BY dr.reason
ORDER BY total_requests DESC
LIMIT 5;
```

esult Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

request_reason	total_requests
ENFORCEMENT & ABANDONED VEHICLES	285691
STREET CLEANING	199236
SANITATION	176288
CODE ENFORCEMENT	138697
HIGHWAY MAINTENANCE	110571

Find Average Case Resolution Time by Request Type

```sql
SELECT
    dr.reason AS request_reason,
    ROUND(AVG(TIMESTAMPDIFF(HOUR, open_date, closed_date)), 2) AS avg_resolution_time
FROM fact_311_request f
JOIN dim_request_dtl dr ON f.request_dtl_key = dr.request_dtl_key
WHERE f.closed_date IS NOT NULL
GROUP BY dr.reason
ORDER BY avg_resolution_time_hours ASC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

request_reason	avg_resolution_time_hours
DISABILITY	0.72
MASSPORT	3.00
ADMINISTRATIVE	5.00
NEEDLE PROGRAM	17.39
CODE ENFORCEMENT	29.98
HEALTH	66.59
WEIGHTS AND MEASURES	66.94
CONSUMER AFFAIRS ISSUES	68.83
GENERAL REQUEST	86.01
CURRENT EVENTS	88.08

Track SLA Compliance Rate by Department

```sql
SELECT
    dr.department,
    COUNT(*) AS total_requests,
    SUM(CASE WHEN on_time = 'ONTIME' THEN 1 ELSE 0 END) AS on_time_requests,
    ROUND(SUM(CASE WHEN on_time = 'ONTIME' THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS sla_compliance_rate
FROM fact_311_request f
JOIN dim_request_dtl dr ON f.request_dtl_key = dr.request_dtl_key
GROUP BY dr.department
ORDER BY sla_compliance_rate DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

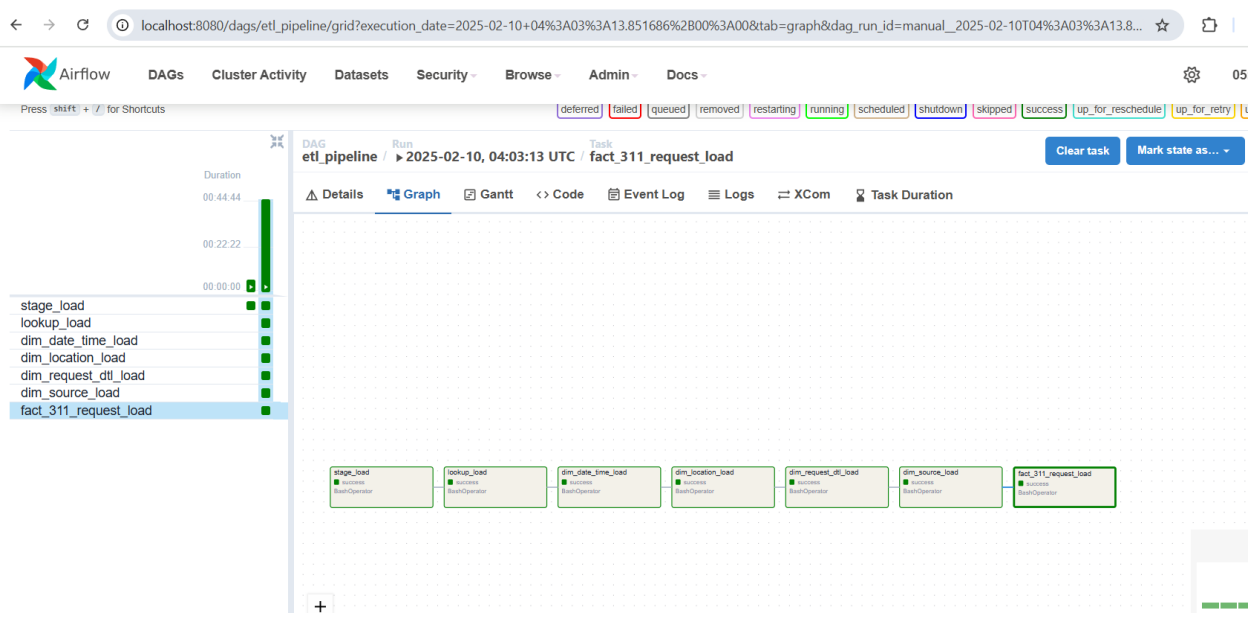
department	total_requests	on_time_requests	sla_compliance_rate
GRNi	12	4	33.33
No Q	13	9	69.23
BHA_	1022	107	10.47
DND_	368	132	35.87
ONS_	686	179	26.09
BPS_	1239	181	14.61
BPD_	2374	306	12.89
DISB	1610	1465	90.99
ANML	6194	5669	91.52
BWSC	32107	14633	45.58
PROP	34776	19286	55.46
GEN_	64881	64838	99.93
INFO	134606	67778	50.35

Identify the Top 3 Most Frequent Request Types in Each Neighborhood

```
SELECT neighborhood, request_reason, total_requests
FROM (
  SELECT
    dl.neighborhood,
    dr.reason AS request_reason,
    COUNT(f.request_id) AS total_requests,
    RANK() OVER (PARTITION BY dl.neighborhood ORDER BY COUNT(f.request_id) DESC) AS rank_order
  FROM fact_311_request f
  JOIN dim_location dl ON f.location_key = dl.location_key
  JOIN dim_request_dtl dr ON f.request_dtl_key = dr.request_dtl_key
  GROUP BY dl.neighborhood, dr.reason
) AS RankedRequests
WHERE rank_order <= 3
ORDER BY neighborhood, rank_order;
```

Output:

Orchestration



Stage_load

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content: | Fetch rows:

bos_case_id	case_enquiry_id	open_dt	sla_target_dt	closed_dt	on_time	case_status	closure_reason	case_title
1	101001438027	2015-07-28 00:03...	2015-08-11 04:30:00	2015-07-28 04:23:44	ONTIME	CLOSED	Case Closed. Closed date : 2015-07-28 08:23:4...	PRINTED
2	101001240753	2015-01-01 01:31...	2015-01-06 03:30:00	2015-01-01 05:49:51	ONTIME	CLOSED	Case Closed Case Resolved citywide cleaned up...	REQUESTS FOR
3	101001269105	2015-02-03 03:06...	2015-02-04 03:30:00	2015-02-04 02:13:33	ONTIME	CLOSED	Case Closed Case Resolved	TRAFFIC SIGNA
4	101001597643	2015-09-25 02:56...	2015-09-26 02:56:17	2015-09-30 10:43:34	OVERDUE	CLOSED	Case Closed. Closed date : 2015-09-30 14:43:3...	TRAFFIC SIGNA
5	101001541887	2015-08-27 06:19...	2015-09-15 06:19:26	2015-09-10 05:34:21	ONTIME	CLOSED	Case Closed. Closed date : 2015-09-10 09:34:2...	ABANDONED VI
6	101001375205	2015-05-12 03:03...	2015-11-08 02:03:04	2015-05-12 06:27:54	ONTIME	CLOSED	Case Closed. Closed date : 2015-05-12 10:27:5...	TRAFFIC SIGNA
7	101001412168	2015-06-18 04:30...	2015-06-23 04:30:48	2015-07-08 07:36:45	OVERDUE	CLOSED	Case Closed. Closed date : 2015-07-08 11:36:4...	BUILDING INSP
8	101001438028	2015-07-28 00:07...	2015-09-30 04:30:00	2015-08-20 02:55:20	ONTIME	CLOSED	Case Closed. Closed date : 2015-08-20 06:55:2...	GRAFFITI REM
9	101001541888	2015-08-27 06:20...	NULL	2015-09-04 09:25:10	ONTIME	CLOSED	Case Closed. Closed date : 2015-09-04 13:25:1...	SCHEDULE A BU
10	101001269106	2015-02-03 03:06...	NULL	NULL	ONTIME	OPEN		MRTA

Dim_date

date_key	date_key_str	date	year	month	day_of_month	day_of_week	db_created_datetime	db_modified_datetime	created_by	modified_by	process_id
20150101	20150101	2015-01-01	2015	1	1	5	2025-02-09 23:16:59	NULL	system	system	47075
20150102	20150102	2015-01-02	2015	1	2	6	2025-02-09 23:16:59	NULL	system	system	47075
20150103	20150103	2015-01-03	2015	1	3	7	2025-02-09 23:16:59	NULL	system	system	47075
20150104	20150104	2015-01-04	2015	1	4	1	2025-02-09 23:16:59	NULL	system	system	47075
20150105	20150105	2015-01-05	2015	1	5	2	2025-02-09 23:16:59	NULL	system	system	47075
20150106	20150106	2015-01-06	2015	1	6	3	2025-02-09 23:16:59	NULL	system	system	47075
20150107	20150107	2015-01-07	2015	1	7	4	2025-02-09 23:16:59	NULL	system	system	47075
20150108	20150108	2015-01-08	2015	1	8	5	2025-02-09 23:16:59	NULL	system	system	47075
20150109	20150109	2015-01-09	2015	1	9	6	2025-02-09 23:16:59	NULL	system	system	47075
20150110	20150110	2015-01-10	2015	1	10	7	2025-02-09 23:16:59	NULL	system	system	47075
20150111	20150111	2015-01-11	2015	1	11	1	2025-02-09 23:16:59	NULL	system	system	47075

dim_date 23 ×

Apply

Dim_time

time_key	time_key_str	time	hour	minute	second	db_created_datetime	db_modified_datetime	created_by	modified_by	process_
0	000000	00:00:00	0	0	0	2025-02-09 23:17:10	NULL	system	system	47075
1	000001	00:00:01	0	0	1	2025-02-09 23:17:10	NULL	system	system	47075
2	000002	00:00:02	0	0	2	2025-02-09 23:17:10	NULL	system	system	47075
3	000003	00:00:03	0	0	3	2025-02-09 23:17:10	NULL	system	system	47075
4	000004	00:00:04	0	0	4	2025-02-09 23:17:10	NULL	system	system	47075
5	000005	00:00:05	0	0	5	2025-02-09 23:17:10	NULL	system	system	47075
6	000006	00:00:06	0	0	6	2025-02-09 23:17:10	NULL	system	system	47075
7	000007	00:00:07	0	0	7	2025-02-09 23:17:10	NULL	system	system	47075
8	000008	00:00:08	0	0	8	2025-02-09 23:17:10	NULL	system	system	47075
9	000009	00:00:09	0	0	9	2025-02-09 23:17:10	NULL	system	system	47075
10	000010	00:00:10	0	0	10	2025-02-09 23:17:10	NULL	system	system	47075

Dim_location

location_key	location_street_name	is_intersection	street_name1	street_name2	neighborhood	city	state	zipcode	ne
1	0 ALPHA RD	N	0 ALPHA RD	NULL	DORCHESTER	Boston	MA	02124	9
2	0 ADDISON ST	N	0 ADDISON ST	NULL	EAST BOSTON	Boston	MA	02128	1
3	0 ACADIA ST	N	0 ACADIA ST	NULL	SOUTH BOSTON / SOUTH BOSTON WATERFRONT	Boston	MA	02127	5
4	0 ADAMS ST	N	0 ADAMS ST	NULL	DORCHESTER	Boston	MA	02122	8
5	0 B ST	N	0 B ST	NULL	SOUTH BOSTON / SOUTH BOSTON WATERFRONT	Boston	MA	02210	5
6	0 BALDWIN PL	N	0 BALDWIN PL	NULL	ALLSTON / BRIGHTON	Boston	MA	02135	15
7	0 BOSTON UNIVERSITY BRG	N	0 BOSTON UNIVERSITY BRG	NULL	ALLSTON / BRIGHTON	Boston	MA	02215	14
8	0 A ST	N	0 A ST	NULL	SOUTH BOSTON / SOUTH BOSTON WATERFRONT	Boston	MA	02127	5
9	0 ARION ST	N	0 ARION ST	NULL	DORCHESTER	Boston	MA	02125	13
10	0 AMERICAN LEGION HWY	N	0 AMERICAN LEGION HWY	NULL	ROSLINDALE	Boston	MA	02131	10

Dim_source

source_key	source	db_created_datetime	db_modified_datetime	created_by	modified_by	process_id
1	CITIZENS CONNECT APP	2025-02-09 23:21:30	NULL	system	NULL	47744
2	CITY WORKER APP	2025-02-09 23:21:30	NULL	system	NULL	47744
3	CONSTITUENT CALL	2025-02-09 23:21:30	NULL	system	NULL	47744
4	EMPLOYEE GENERATED	2025-02-09 23:21:30	NULL	system	NULL	47744
5	MAXIMO INTEGRATION	2025-02-09 23:21:30	NULL	system	NULL	47744
6	SELF SERVICE	2025-02-09 23:21:30	NULL	system	NULL	47744
7	TWITTER	2025-02-09 23:21:30	NULL	system	NULL	47744

Dim_request_dtl

request_dtl_key	case_title	subject	reason	type	department	queue
1	"""ALL STREET LIGHTS OUT."" (CHECK OVERHE...	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	STREET LIGHT OUTAGES	PWDx	PWDx_Street Light_
2	"ABANDONED BICYCLE - ""CITY BIKE UNKNOW...	MAYOR'S 24 HOUR HOTLINE	ABANDONED BICYCLE	ABANDONED BICYCLE	BTDT	BTDT_Abandoned B
3	"ARM AND FIXTURE TRANSFER ASAP///ISSUED ...	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	STREET LIGHT OUTAGES	PWDx	PWDx_Street Light_
4	"CONTRACTOR LIGHT ""SENT TO MAVERICK"""	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	STREET LIGHT OUTAGES	PWDx	PWDx_Street Light_
5	"CONTRACTOR LIGHT ""SENT TOP MAVERICK"""	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	STREET LIGHT OUTAGES	PWDx	PWDx_Street Light_
6	"DCR STREET LIGHT OUTAGES ---REALLOCATE...	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	STREET LIGHT OUTAGES	INFO	INFO_Mass DCR
7	"GAS LIGHT OUTAGES //PRINTED// MD//""NO G...	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	STREET LIGHT OUTAGES	PWDx	PWDx_Street Light
8	"GAS LIGHT OUTAGES//PRINTED ""ON CALL"" //...	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	STREET LIGHT OUTAGES	PWDx	PWDx_Street Light
9	"GENERAL LIGHTING REQUEST - REPAIR BROK...	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	GENERAL LIGHTING REQUEST	PWDx	PWDx_Street Light_
10	"GENERAL LIGHTING REQUEST: TWO 19"" POLE...	PUBLIC WORKS DEPARTMENT	STREET LIGHTS	GENERAL LIGHTING REQUEST	PWDx	PWDx_Street Light

Future Scope

- Implement partitioning & indexing strategies for better query performance
- Add new business metrics and categorization for better insights
- Use NLP & AI models to analyze closed_reason text data
- Expand reports in Power BI or Tableau for deeper insights
- Deploy to AWS/GCP for scalability & automation
- Implement logging, alerting & retry mechanisms