****Logotipo

Descripción generada automáticamente**LICENCIATURA EN CIENCIA DE DATOS**

**Instituto Politécnico Nacional**

**Escuela Superior de Cómputo**

**UNIDAD DE APRENDIZAJE**

*BASES DE DATOS AVANZADAS*

**TÍTULO DEL TRABAJO**

Proyecto semestral

**ALUMNOS**

Aguilar Ramírez Carlos Francisco, Arista Romero Juan Ismael, Vázquez Martín Marlene Gabriela.

**PROFESOR (A)**

Andrés García Floriano

**GRUPO**

4AV1

**Fecha de entrega:** 6 de enero de 2025

**Introducción**

En este proyecto, se desarrollará un proceso ETL (Extracción, Transformación y Carga) enfocado en los datos de calidad del aire en India. Este proceso tiene como objetivo recopilar, transformar y cargar datos estructurados que permitan realizar análisis avanzados y generar reportes significativos sobre la calidad del aire en diferentes ciudades, estaciones del año y momentos del día.

El proyecto se divide en varias etapas clave que aseguran la integridad y usabilidad de los datos:

1. **Extracción de Datos**  
   Los datos serán descargados desde la fuente proporcionada y organizados en una carpeta denominada "landing-zone" dentro del entorno local.
2. **Perfilado y Reportes de Calidad de Datos**  
   Se analizarán los datos utilizando técnicas de perfilado para identificar inconsistencias y valores faltantes. Los reportes generados se almacenarán en la carpeta "data\_quality\_reports" dentro de la zona denominada "raw-zone".
3. **Transformación y Almacenamiento de Datos**  
   Se diseñarán reglas específicas para manejar valores nulos y se convertirán los archivos CSV de la "landing-zone" al formato Parquet, optimizando su almacenamiento y procesamiento.
4. **Carga y Refinamiento de Datos**  
   Los datos transformados se copiarán a una carpeta llamada "refined-zone" en formato Parquet, garantizando su accesibilidad y eficiencia para análisis posteriores.
5. **Integración con Base de Datos**  
   Los archivos procesados se cargarán en una base de datos SQL, permitiendo la consulta y recuperación estructurada de la información.
6. **Análisis y Reportes de Datos**  
   Se diseñarán consultas SQL que generen reportes sobre la calidad del aire segmentados por ciudad, estación del año y momento del día. Estas consultas incluirán agregaciones y filtros para obtener información relevante.
7. **Creación de Tableros de Visualización**  
   Se desarrollarán dashboards en Power BI para visualizar de manera intuitiva los insights obtenidos, facilitando la toma de decisiones informadas.

Este proyecto no solo reforzará los conceptos aprendidos a lo largo del curso, sino que también proporcionará un enfoque práctico para diseñar soluciones ETL robustas y escalables, con aplicaciones reales en el análisis de datos urbanos, específicamente los relacionados a cuestiones ambientales.

# Data Quality Reports

## Report for city\_day

Shape: (29531, 16)

Columns: City, Date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI\_Bucket

Null Values:

|  |  |
| --- | --- |
| Column | Null Count |
| City | 0 |
| Date | 0 |
| PM2.5 | 4598 |
| PM10 | 11140 |
| NO | 3582 |
| NO2 | 3585 |
| NOx | 4185 |
| NH3 | 10328 |
| CO | 2059 |
| SO2 | 3854 |
| O3 | 4022 |
| Benzene | 5623 |
| Toluene | 8041 |
| Xylene | 18109 |
| AQI | 4681 |
| AQI\_Bucket | 4681 |

Data Types:

- City: object

- Date: object

- PM2.5: float64

- PM10: float64

- NO: float64

- NO2: float64

- NOx: float64

- NH3: float64

- CO: float64

- SO2: float64

- O3: float64

- Benzene: float64

- Toluene: float64

- Xylene: float64

- AQI: float64

- AQI\_Bucket: object

Sample Data:

City Date PM2.5 PM10 NO NO2 NOx NH3 CO SO2 O3 Benzene Toluene Xylene AQI AQI\_Bucket  
0 Ahmedabad 2015-01-01 NaN NaN 0.92 18.22 17.15 NaN 0.92 27.64 133.36 0.00 0.02 0.00 NaN NaN  
1 Ahmedabad 2015-01-02 NaN NaN 0.97 15.69 16.46 NaN 0.97 24.55 34.06 3.68 5.50 3.77 NaN NaN  
2 Ahmedabad 2015-01-03 NaN NaN 17.40 19.30 29.70 NaN 17.40 29.07 30.70 6.80 16.40 2.25 NaN NaN

## Report for city\_hour

Shape: (707875, 16)

Columns: City, Datetime, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI\_Bucket

Null Values:

|  |  |
| --- | --- |
| Column | Null Count |
| City | 0 |
| Datetime | 0 |
| PM2.5 | 145088 |
| PM10 | 296737 |
| NO | 116632 |
| NO2 | 117122 |
| NOx | 123224 |
| NH3 | 272542 |
| CO | 86517 |
| SO2 | 130373 |
| O3 | 129208 |
| Benzene | 163646 |
| Toluene | 220607 |
| Xylene | 455829 |
| AQI | 129080 |
| AQI\_Bucket | 129080 |

Data Types:

- City: object

- Datetime: object

- PM2.5: float64

- PM10: float64

- NO: float64

- NO2: float64

- NOx: float64

- NH3: float64

- CO: float64

- SO2: float64

- O3: float64

- Benzene: float64

- Toluene: float64

- Xylene: float64

- AQI: float64

- AQI\_Bucket: object

Sample Data:

City Datetime PM2.5 PM10 NO NO2 NOx NH3 CO SO2 O3 Benzene Toluene Xylene AQI AQI\_Bucket  
0 Ahmedabad 2015-01-01 01:00:00 NaN NaN 1.00 40.01 36.37 NaN 1.00 122.07 NaN 0.0 0.0 0.0 NaN NaN  
1 Ahmedabad 2015-01-01 02:00:00 NaN NaN 0.02 27.75 19.73 NaN 0.02 85.90 NaN 0.0 0.0 0.0 NaN NaN  
2 Ahmedabad 2015-01-01 03:00:00 NaN NaN 0.08 19.32 11.08 NaN 0.08 52.83 NaN 0.0 0.0 0.0 NaN NaN

## Report for station\_day

Shape: (108035, 16)

Columns: StationId, Date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI\_Bucket

Null Values:

|  |  |
| --- | --- |
| Column | Null Count |
| StationId | 0 |
| Date | 0 |
| PM2.5 | 21625 |
| PM10 | 42706 |
| NO | 17106 |
| NO2 | 16547 |
| NOx | 15500 |
| NH3 | 48105 |
| CO | 12998 |
| SO2 | 25204 |
| O3 | 25568 |
| Benzene | 31455 |
| Toluene | 38702 |
| Xylene | 85137 |
| AQI | 21010 |
| AQI\_Bucket | 21010 |

Data Types:

- StationId: object

- Date: object

- PM2.5: float64

- PM10: float64

- NO: float64

- NO2: float64

- NOx: float64

- NH3: float64

- CO: float64

- SO2: float64

- O3: float64

- Benzene: float64

- Toluene: float64

- Xylene: float64

- AQI: float64

- AQI\_Bucket: object

Sample Data:

StationId Date PM2.5 PM10 NO NO2 NOx NH3 CO SO2 O3 Benzene Toluene Xylene AQI AQI\_Bucket  
0 AP001 2017-11-24 71.36 115.75 1.75 20.65 12.40 12.19 0.10 10.76 109.26 0.17 5.92 0.10 NaN NaN  
1 AP001 2017-11-25 81.40 124.50 1.44 20.50 12.08 10.72 0.12 15.24 127.09 0.20 6.50 0.06 184.0 Moderate  
2 AP001 2017-11-26 78.32 129.06 1.26 26.00 14.85 10.28 0.14 26.96 117.44 0.22 7.95 0.08 197.0 Moderate

## Report for station\_hour

Shape: (2589083, 16)

Columns: StationId, Datetime, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI\_Bucket

Null Values:

|  |  |
| --- | --- |
| Column | Null Count |
| StationId | 0 |
| Datetime | 0 |
| PM2.5 | 647689 |
| PM10 | 1119252 |
| NO | 553711 |
| NO2 | 528973 |
| NOx | 490808 |
| NH3 | 1236618 |
| CO | 499302 |
| SO2 | 742737 |
| O3 | 725973 |
| Benzene | 861579 |
| Toluene | 1042366 |
| Xylene | 2075104 |
| AQI | 570190 |
| AQI\_Bucket | 570190 |

Data Types:

- StationId: object

- Datetime: object

- PM2.5: float64

- PM10: float64

- NO: float64

- NO2: float64

- NOx: float64

- NH3: float64

- CO: float64

- SO2: float64

- O3: float64

- Benzene: float64

- Toluene: float64

- Xylene: float64

- AQI: float64

- AQI\_Bucket: object

Sample Data:

StationId Datetime PM2.5 PM10 NO NO2 NOx NH3 CO SO2 O3 Benzene Toluene Xylene AQI AQI\_Bucket  
0 AP001 2017-11-24 17:00:00 60.5 98.00 2.35 30.80 18.25 8.50 0.1 11.85 126.40 0.1 6.10 0.10 NaN NaN  
1 AP001 2017-11-24 18:00:00 65.5 111.25 2.70 24.20 15.07 9.77 0.1 13.17 117.12 0.1 6.25 0.15 NaN NaN  
2 AP001 2017-11-24 19:00:00 80.0 132.00 2.10 25.18 15.15 12.02 0.1 12.08 98.98 0.2 5.98 0.18 NaN NaN

## Report for stations

Shape: (230, 5)

Columns: StationId, StationName, City, State, Status

Null Values:

|  |  |
| --- | --- |
| Column | Null Count |
| StationId | 0 |
| StationName | 0 |
| City | 0 |
| State | 0 |
| Status | 97 |

Data Types:

- StationId: object

- StationName: object

- City: object

- State: object

- Status: object

Sample Data:

StationId StationName City State Status  
0 AP001 Secretariat, Amaravati - APPCB Amaravati Andhra Pradesh Active  
1 AP002 Anand Kala Kshetram, Rajamahendravaram - APPCB Rajamahendravaram Andhra Pradesh NaN  
2 AP003 Tirumala, Tirupati - APPCB Tirupati Andhra Pradesh NaN