**21CSA523A Data Engineering for AI**

**Mid Review Report**

**City Air Quality Data Extraction, Transformation and Loading**

**Prepared by**

**Sachin Madhukar Garajkar**

**AA.SC.P2MCA2107413**

**MCA AI Sem-3**

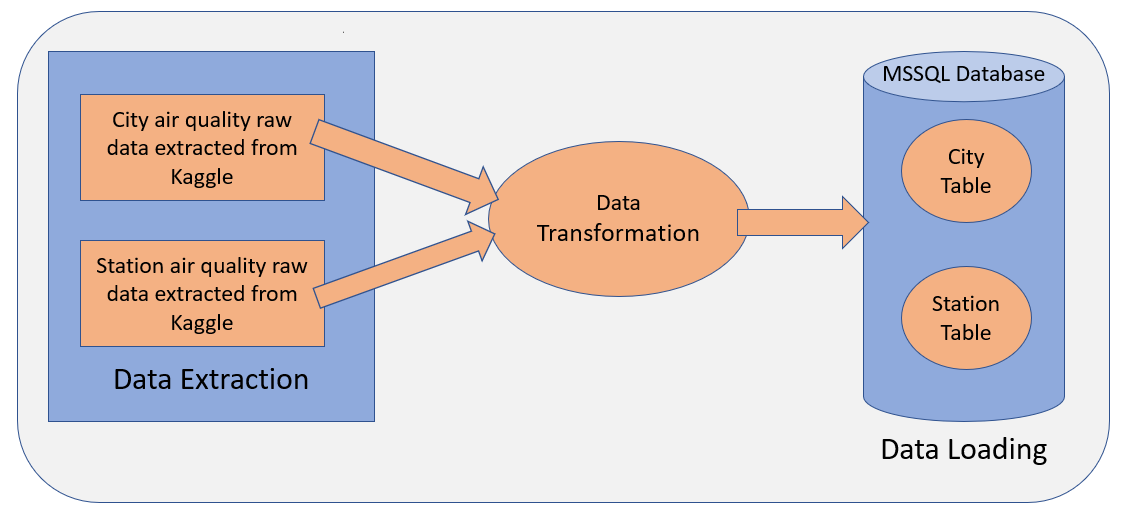
**Email:** [**garajkarsm\_mca2107413@ahead.students.amrita.edu**](mailto:garajkarsm_mca2107413@ahead.students.amrita.edu)

**Objective:**

Air is what keeps humans alive. Monitoring it and understanding its quality is of immense importance to our well-being. The Objective of this mini project is to collect major Indian cities air quality data, transform it and load it in the educational institute database for students use. I will analyse raw data and create ETL pipeline of major Indian Cities Air Quality data.I have taken raw dataset from Kaggle (<https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>) . The dataset contains air quality data and AQI (Air Quality Index) at hourly and daily level of various stations across multiple cities in India. Once data is transformed and loaded into the database than this historical data can be used for air quality studies by the students. As the loading database is belong to educational institute.

The Dataset have 26 different cities and various stations air quality and AQI data.

**Block Diagram:**

****

**Tools Used:**

1. Pandas, Numpy
2. Sqlalchemy, pyodbc libraries
3. EDA Techniques
4. MS SQL Server & Database

**Dataset details:**

Below 2 Datasets used from Kaggle ([https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-indiaa](https://www.kaggle.com/code/anjanirawat/air-quality-data-in-india))

1. city\_hour.csv

This Dataset is the citywise air quality data of 26 major Indian cities, which includes 25 different air quality parameters.

1. Station\_hour.csv

This Dataset has air quality data of different stations from different cities in India.

**Plan to Execute:**

1. Firstly I will extract raw data from available sources using Pandas library.
2. I will transform this data in the required format so that one can do data analysis to reach to the final conclusion.
3. I will then load this data to the dataset so that it can be used for Data analysis.

