**Proposal for ETL Project**

**Overview**

Our group, Thomas Law, Thomas Knies, and Grace Huefner were given three datasets to work with and organize. We were instructed that these three data sets should be pulled into a single format database and simpler to use and read by Saturday.

**Extract:**

We have extracted extracting the below Datasets from Kaggle.

<https://www.kaggle.com/epa/epa-historical-air-quality>

<https://www.kaggle.com/bappekim/air-pollution-in-seoul?>

<https://www.kaggle.com/rohanrao/air-quality-data-in-india?select=city_hour.csv>

**Transform:**

Each team member selected a specific city dataset.

* Thomas Laws: New Delhi, India
* Thomas Knies: Seoul, South Korea
* Grace Huefner: New York, USA

We each cleaned the data to make sure that categories and measurements were aligned. Then split them out by year and air pollutant.

We then put that cleaned data into a SQL database. From that SQL database we will join the data into three tables. Adding in the year of each and measurement of the pollutants per city.

**Load:**

Tables by pollutant by year (Pollutant columns, City rows) like the below into a SQL database:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| City | SO2 | NO2 | O3 | CO | PM10 | PM2.5 |
| New Delhi |  |  |  |  |  |  |
| New York |  |  |  |  |  |  |
| Seoul |  |  |  |  |  |  |