US EPA

Environmental database

Ravisha SK, Jeff Neilson, Nayan Gupta, Lavanya Basava Raju, Ryan Borowicz

MSBA6320 – Gold Cohort - Section 002

Team 14

December 16, 2016

Contents

[1.0 Introduction 2](#_Toc469631669)

[2.0 Requirements Definition 3](#_Toc469631670)

[3.0 Design & Implementation 3](#_Toc469631671)

[3.1. Data Analysis 3](#_Toc469631672)

[3.2. Source Database Design 7](#_Toc469631673)

[3.3. Target Data Warehouse Design 10](#_Toc469631674)

[3.4. Source to Target Mapping 12](#_Toc469631675)

[3.5. ETL Design & Development 13](#_Toc469631676)

[4.0 Analysis & Results 19](#_Toc469631677)

[4.1. Addressing Business Requirements 19](#_Toc469631678)

[5.0 Conclusion 21](#_Toc469631679)

[6.0 Bibliography 23](#_Toc469631680)

[7.0 Appendix 24](#_Toc469631681)

[7.1. Source Database Create Table Statements 24](#_Toc469631682)

[7.2. Target Data Warehouse Create Table Statements 24](#_Toc469631683)

[7.3. Source to Target Mapping Spreadsheet 24](#_Toc469631684)

[7.4. Analysis SQL Queries 24](#_Toc469631685)

1. Introduction

The harmful effects of pollution can be most readily observed through changes in the local ecosystem: water quality issues, soil contamination, and changes to the balance of plant and animal life in a particular region. When thinking about the dangers of pollution, these environmental factors are typically the first to come to mind. But pollution can also have an economic impact. According to the OECD, the economic impact of air pollution alone is estimated to have an annual cost of 1% of global GDP by 20601. In addition to the more obvious sources of economic impact, such as lower efficiency in the agriculture industry, this calculation also takes into account the cost of sick days and medical bills attributed to pollution.

It seems reasonable to conclude that elevated levels of fine particles may lead to higher rates of respiratory disease. However, in addition to more hospitalizations due to respiratory diseases, researchers from the Harvard School of Public Health have found that long term exposure to areas with higher long-term average fine particle levels lead to higher admission rates for heart attacks, strokes, and diabetes2. These are serious conditions that may lead to loss of life. The OECD has quantified this impact, reporting that air pollution is expected to cause between 6 to 9 million premature deaths by 20601.

Clearly, pollution is a serious problem that can burden a society in many ways. To address this problem, the U.S. government created the Environmental Protection Agency (EPA) in 1970 (citation). In short, the mission of the EPA is to “protect human health and the environment4.” The main role of the EPA is to both draft and enforce environmental regulations based on laws ratified by congress (citation).

In addition to the literature already produced by the EPA supporting links between air pollution and chronic diseases, the EPA is currently soliciting a funding opportunity to further the understanding of long-term exposure to air pollution and how it relates to the development of cardiovascular disease5. The EPA hopes gain valuable insights that will provide compelling evidence when lobbying for more effective environmental regulations. The results of the study will simultaneously build awareness for the issue and lay the groundwork necessary to solve it.

The potential for such a large, meaningful impact and the alignment with the current interests of the EPA serve as the motivation for our project. We will build a data warehouse that integrates chronic disease, air pollution, climate, weather, and census data. This data warehouse will be designed to support the EPA and related agencies in reporting. Combined with asking the right analysis questions, this data warehouse will aid in the discovery of meaningful patterns and associations between air pollution and chronic disease. These findings will be used to both monitor trends and strengthen the lobbying efforts of the EPA. Ultimately, the conclusions drawn through analysis of the information contained in the data warehouse will ideally lead to a reduction of the adverse effects of air pollution, saving ecosystems, the economy, and lives.

1. Requirements Definition

The EPA wants to spread awareness about the alarming increase in air pollution. This data warehouse will serve two primary sets of priorities. First, it will allow different governmental agencies to be able to report annually on different environmental measurements such as ozone, carbon dioxide, and emission levels to assist in regulatory and marketing efforts for their specific organization. Second, the information can be looked at together by data scientists to identify impacts of the different environmental factors on cancer rates across different US regions.

The list below provides a few of the requirements we are looking to address through the creation of this warehouse. The list was kept short for this document to give a high-level idea of the requirements. In a lengthier project, a full requirements traceability matrix would have been developed to document the requirements.

| Requirement | Business Area |
| --- | --- |
| Which states have carbon monoxide emissions higher than the acceptable carbon monoxide AQI for the latest year of 2016 | Pollution, Emissions |
| Is there any correlation trend between the annual lung cancer rate and the corresponding levels of Nitrogen dioxide for each of the states in USA for the year 2012? | Chronic Diseases, Pollution, Emissions |
| Which are the 5 states that account for highest incidence of lung cancer and bronchus cancer and what is the corresponding mortality rate? | Chronic Diseases |
| Is there any correlation trend between Sulphur Dioxide pollutant level and average temperature over a period of time (at a monthly level) for 5 states with the highest concentration of the pollutant | Pollution, Temperature |

1. Design & Implementation
   1. Data Analysis

Our datasets were chosen from Kaggle and from US government provided data. We chose datasets that capture pollution, chronic disease, carbon emissions, U.S. census data regarding population, and temperature information. These particular sources were selected because they are reliable and they contain all the information necessary to answer our analysis questions and fit our business requirements.

**3.1.1 Pollution**

The pollution data was obtained from https://www.kaggle.com/sogun3/uspollution. It originally contained 1,746,661 records and 29 variables, and we utilized the following fields:

| Field Name | Description | Data Type | Character Length | Null |
| --- | --- | --- | --- | --- |
| State\_Name | State of monitoring site | NVARCHAR | 25 |  |
| NO2\_Mean | The arithmetic mean of concentration of NO2 within a given day | DECIMAL | 10,2 | X |
| NO2\_1st\_Max\_Value | The maximum value obtained for NO2 concentration in a given day | DECIMAL | 10,2 | X |
| NO2\_1st\_Max\_Hour | The hour when the maximum NO2 concentration was recorded in a given day | DECIMAL | 10,2 | X |
| NO2\_AQI | The calculated air quality index of NO2 within a given day | DECIMAL | 10,2 | X |
| O3\_Mean | The arithmetic mean of concentration of O3 within a given day | DECIMAL | 10,2 | X |
| O3\_1st\_Max\_Value | The maximum value obtained for O3 concentration in a given day | DECIMAL | 10,2 | X |
| O3\_1st\_Max\_Hour | The hour when the maximum O3 concentration was recorded in a given day | DECIMAL | 10,2 | X |
| O3\_AQI | The calculated air quality index of O3 within a given day | DECIMAL | 10,2 | X |
| SO2\_Mean | The arithmetic mean of concentration of SO2 within a given day | DECIMAL | 10,2 | X |
| SO2\_1st\_Max\_Value | The maximum value obtained for SO2 concentration in a given day | DECIMAL | 10,2 | X |
| SO2\_1st\_Max\_Hour | The hour when the maximum SO2 concentration was recorded in a given day | DECIMAL | 10,2 | X |
| SO2\_AQI | The calculated air quality index of SO2 within a given day | DECIMAL | 10,2 | X |
| CO\_Mean | The arithmetic mean of concentration of CO within a given day | DECIMAL | 10,2 | X |
| CO\_1st\_Max\_Value | The maximum value obtained for CO concentration in a given day | DECIMAL | 10,2 | X |
| CO\_1st\_Max\_Hour | The hour when the maximum CO concentration was recorded in a given day | DECIMAL | 10,2 | X |
| CO\_AQI | The calculated air quality index of CO within a given day | DECIMAL | 10,2 | X |

**Notes regarding pollution:**

* County Code and Site Num from the original dataset were aggregated and the state level using R
* Year of Monitoring and Month of Monitoring from the original dataset were aggregated at the monthly level using R
* Agency Name data was added manually

**3.1.2 Chronic Disease Indicators (CDI)**

The chronic disease information was obtained from <https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi-e50c9>. It contains the following fields:

| Field Name | Description | Data Type | Character Length | Null |
| --- | --- | --- | --- | --- |
| YearStart | Start Year | NVARCHAR | 4 |  |
| YearEnd | End Year | NVARCHAR | 4 |  |
| LocationAbbr | State Abbreviation | NVARCHAR | 2 |  |
| LocationDesc | State Name | TEXT |  |  |
| DataSource | Name of monitoring agency | NVARCHAR | 50 |  |
| Alcohol\_Binge\_Freq\_Adults\_Mean | Binge drinking prevalence among adults | DECIMAL | 10,2 | X |
| Alcohol\_Liver\_Disease\_Deaths\_Avg\_No | Chronic liver disease mortality | DECIMAL | 10,2 | X |
| Alcohol\_Percap\_Cons\_Mean | Per capita alcohol consumption among persons aged >= 14 years | DECIMAL | 10,2 | X |
| Cancer\_Colon\_Avg\_No | Cancer of the colon and rectum (colorectal), incidence | DECIMAL | 10,2 | X |
| Cancer\_Colon\_Deaths\_Avg\_No | Cancer of the colon and rectum (colorectal), deaths | DECIMAL | 10,2 | X |
| Cancer\_Breast\_Deaths\_Avg\_No | cancer of the female breast, deaths | DECIMAL | 10,2 | X |
| Cancer\_Cervix\_Deaths\_Avg\_No | Cancer of the female cervix, mortality | DECIMAL | 10,2 | X |
| Cancer\_Lung\_Avg\_No | Cancer of the lung and bronchus, incidence | DECIMAL | 10,2 | X |
| Cancer\_Lung\_Deaths\_Avg\_No | Cancer of the lung and bronchus, mortality | DECIMAL | 10,2 | X |
| Cancer\_Oral\_Deaths\_Avg\_No | Cancer of the oral cavity and pharynx, mortality | DECIMAL | 10,2 | X |
| Cancer\_Prostate\_Deaths\_Avg\_No | Cancer of the prostate, mortality | DECIMAL | 10,2 | X |
| Cancer\_Invasive\_Avg\_No | Invasive cancer (all sites combined), incidence | DECIMAL | 10,2 | X |
| Cancer\_Invasive\_Deaths\_Avg\_No | Invasive cancer (all sites combined), mortality | DECIMAL | 10,2 | X |
| Cancer\_Cervix\_Invasive\_Avg\_No | Invasive cancer of the female cervix | DECIMAL | 10,2 | X |
| Cancer\_Breast\_Invasive\_Avg\_No | Invasive cancer of the female breast, incidence | DECMAL | 10,2 | X |
| Cancer\_Oral\_Invasive\_Avg\_No | Invasive cancer of the oral cavity or pharynx, incidence | DECIMAL | 10,2 | X |
| Cancer\_Prostate\_Invasive\_Avg\_No | Invasive cancer of the prostate, incidence | DECIMAL | 10,2 | X |
| Cancer\_Melanoma\_Invasive\_Avg\_No | Invasive melanoma, incidence | DECIMAL | 10,2 | X |
| Cancer\_Melanoma\_Deaths\_Avg\_No | Melanoma, mortality | DECIMAL | 10,2 | X |
| Diabetes\_Deaths\_Avg\_No | Mortality due to diabetes reported as any listed cause of death | DECIMAL | 10,2 | X |
| Tobacco\_Sale\_of\_Cig\_No | Current smoking among adults aged >= 18 years | DECIMAL | 10,2 | X |

**3.1.3 Carbon Emissions**

The carbon emissions information was obtained from <https://www.google.com/#q=kaggle+carbon+emissions>. It contains the following fields:

| Field Name | Description | Data Type | Character Length | Null |
| --- | --- | --- | --- | --- |
| MSN | Code tied to the description related to type of emissions data | VARCHAR | 7 |  |
| YYYYMM | Year and month of the measurement | VARCHAR | 6 |  |
| Value | Measurement value | DECIMAL | 10,2 | X |
| Description | Type of emissions data (i.e. Coal Electric Power Sector CO2 Emissions) | TEXT |  | X |
| Unit | Measurement unit for the emissions value (i.e. million metric tons of carbon dioxide) | TEXT |  | X |

**Notes:**

* Agency name was added manually

**3.1.4 US Census**

The U.S. Census information was obtained from https://www.census.gov/popest/data/counties/totals/2015/CO-EST2015-alldata.html. It contains the following fields:

| Field Name | Description | Data Type | Character Length | Null |
| --- | --- | --- | --- | --- |
| state\_name | State name | NVARCHAR | 25 |  |
| population\_Estimate | 7/1/xxxx resident total population estimate | NUMERIC |  | X |
| Births | Births in period 7/1/xxxx to 6/30/xxxx | NUMERIC |  |  |
| Deaths | Deaths in period 7/1/xxxx to 6/30/xxxx | NUMERIC |  |  |
| year\_num | Year | NVARCHAR | 4 |  |
| Agencyname | Name of monitoring agency | NVARCHAR | 50 |  |

**Notes:**

* population\_Estimate, Births, and Deaths were aggregated at the state level using R.
* year\_num was derived through transformation using R.
* agencyname was derived through the ETL process

**3.1.5 Temperature**

The temperature information was obtained from <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data>. It contains the following fields:

| Field Name | Description | Data Type | Character Length | Null |
| --- | --- | --- | --- | --- |
| Dt | Month/Year of temperature measurement | DATETIME |  |  |
| State | State of temperature measurement | NVARCHAR | 25 |  |
| Country | US | NVARCHAR | 25 |  |
| AverageTemperature | global average land temperature in Celsius | DECIMAL | 10,2 | X |
| AverageTemperatureUncertainty | the 95% confidence interval around the average | DECIMAL | 10,2 |  |

**Notes:**

* agencyname was added manually
  1. Source Database Design

**Technical Architecture Overview:**

For this project, we are utilizing Microsoft SQL Server as the database for storing our EPA source database and data warehouse. We are using the university provided server (msba6320.oit.umn.edu), and storing the results in our individual databases. We used one of our team member’s databases for the final product (db\_basav015), and this database will store both the source system tables as well as the data warehouse facts and dimensions. Security is enforced on the database at the student level.

After identifying the datasets needed to address the business requirements, we first wanted to build our database structure in a traditional relational (RDBMS) format, then use that database as the source to populate our data warehouse.

**Creation of Source Database:**

The tables that were created in Microsoft SQL Server are documented below. Refer to Appendix 7.1 for the create table statements related to the tables.

**Pollution Table:**

| Field Name | Data Type | Character Length | Null | Primary Key |
| --- | --- | --- | --- | --- |
| State\_name | NVARCHAR | 25 | NOT NULL | X |
| Year\_num | VARCHAR | 4 | NOT NULL | X |
| Month\_num | VARCHAR | 2 | NOT NULL | X |
| NO2\_Mean | DECIMAL | (10,2) | NULL |  |
| NO2\_1st\_Max\_Value | DECIMAL | (10,2) | NULL |  |
| NO2\_1st\_Max\_Hour | DECIMAL | (10,2) | NULL |  |
| NO2\_AQI | DECIMAL | (10,2) | NULL |  |
| O3\_Mean | DECIMAL | (10,2) | NULL |  |
| O3\_1st\_Max\_Value | DECIMAL | (10,2) | NULL |  |
| O3\_1st\_Max\_Hour | DECIMAL | (10,2) | NULL |  |
| O3\_AQI | DECIMAL | (10,2) | NULL |  |
| SO2\_Mean | DECIMAL | (10,2) | NULL |  |
| SO2\_1st\_Max\_Value | DECIMAL | (10,2) | NULL |  |
| SO2\_1st\_Max\_Hour | DECIMAL | (10,2) | NULL |  |
| SO2\_AQI | DECIMAL | (10,2) | NULL |  |
| CO\_Mean | DECIMAL | (10,2) | NULL |  |
| CO\_1st\_Max\_Value | DECIMAL | (10,2) | NULL |  |
| CO\_1st\_Max\_Hour | DECIMAL | (10,2) | NULL |  |
| CO\_AQI | DECIMAL | (10,2) | NULL |  |
| agencyname | NVARCHAR | 50 | NOT NULL |  |

**Chronic Disease Table:**

| Field Name | Data Type | Character Length | Null | Primary Key |
| --- | --- | --- | --- | --- |
| YearStart | NVARCHAR | 4 | NOT NULL | X |
| YearEnd | NVARCHAR | 4 | NOT NULL | X |
| LocationAbbr | NVARCHAR | 2 | NOT NULL | X |
| LocationDescr | TEXT | NA | NOT NULL |  |
| DataSource | NVARCHAR | 50 | NOT NULL | X |
| Alcohol\_Binge\_Freq\_Adults\_Mean | DECIMAL | (10,2) | NULL |  |
| Alcohol\_Liver\_Disease\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Alcohol\_Percap\_Cons\_Mean | DECIMAL | (10,2) | NULL |  |
| Cancer\_Colon\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Colon\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Breast\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Cervix\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Lung\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Lung\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Oral\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Prostate\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Invasive\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Invasive\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Cervix\_Invasive\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Breast\_Invasive\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Oral\_Invasive\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Prostate\_Invasive\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Melanoma\_Invasive\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Cancer\_Melanoma\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Diabetes\_Deaths\_Avg\_No | DECIMAL | (10,2) | NULL |  |
| Tobacco\_Sale\_of\_Cig\_No | DECIMAL | (10,2) | NULL |  |

**Carbon Emissions Table:**

| Field Name | Data Type | Character Length | Null | Primary Key |
| --- | --- | --- | --- | --- |
| MSN | VARCHAR | 7 | NOT NULL | X |
| YYYYMM | VARCHAR | 6 | NOT NULL | X |
| Value | DECIMAL | (10,2) | NULL |  |
| Column\_Order | VARCHAR | 1 | NULL |  |
| Description | TEXT | NA | NULL |  |
| Unit | TEXT | NA | NULL |  |
| agencyname | NVARCHAR | 50 | NOT NULL |  |

**US Census Table:**

| Field Name | Data Type | Character Length | Null | Primary Key |
| --- | --- | --- | --- | --- |
| State\_name | NVARCHAR | 25 | NOT NULL | X |
| Population\_Estimate | NUMERIC | NA | NULL |  |
| Births | NUMERIC | NA | NOT NULL |  |
| Deaths | NUMERIC | NA | NOT NULL |  |
| year\_num | NVARCHAR | 4 | NOT NULL | X |
| agencyname | NVARCHAR | 50 | NOT NULL |  |

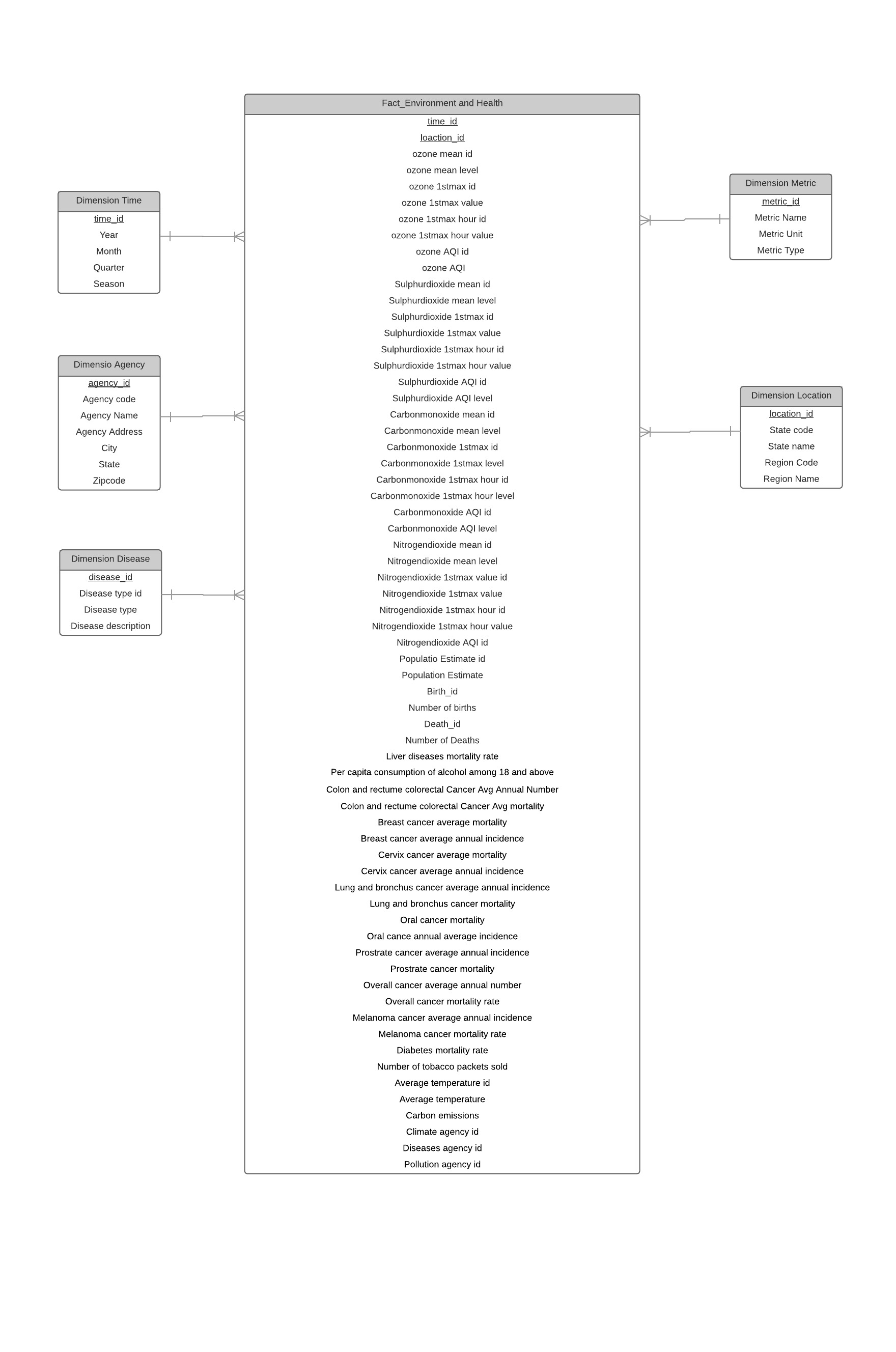
**Temperature Table:**

| Field Name | Data Type | Character Length | Null | Primary Key |
| --- | --- | --- | --- | --- |
| Dt | DATETIME | NA | NOT NULL | X |
| State | NVARCHAR | 25 | NOT NULL | X |
| Country | NVARCHAR | 25 | NOT NULL | X |
| AverageTemperature | DECIMAL | (10,2) | NULL |  |
| AverageTemp\_Uncert | DECIMAL | (10,2) | NULL |  |
| agencyname | NVARCHAR | 50 | NOT NULL |  |

* 1. Target Data Warehouse Design

The diagram below displays the star schema for our data warehouse:

**Data Warehouse Star Schema Diagram**



**Grain:**

The grain for our model is the time id and location id. Each measure in the fact table represents monthly amounts and averages for each state.

**Dimensions:**

The dimensions used to describe the environmental facts are time, location, agency, metric, and disease type. A description of each of the dimensions in provided below:

* **Time** - The Time dimension includes all year/month combinations relevant to the dataset
* **Agency** - The Agency dimension includes information about the governmental agency who manages the data
* **Location** – The Location dimension includes information about the state and region related to the measurement
* **Metric** – The Metric dimension includes information about the type of measurement being performed
* **Disease Type** - The disease type dimension includes information about the disease including the type

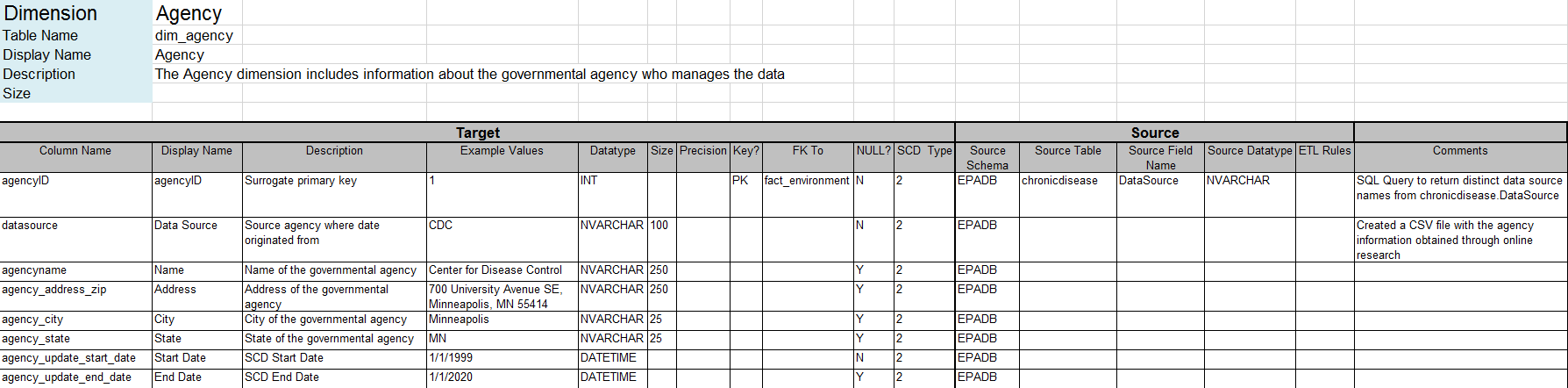
**Facts:**

One fact table was created to include the different environmental measures for pollution, temperature, chronic disease, and census data. The fact table is very long and narrow in our model. A description for each of these values can be found in the Data Dictionary tab on the Source-Target Mapping spreadsheet, which we will discuss in the next section.

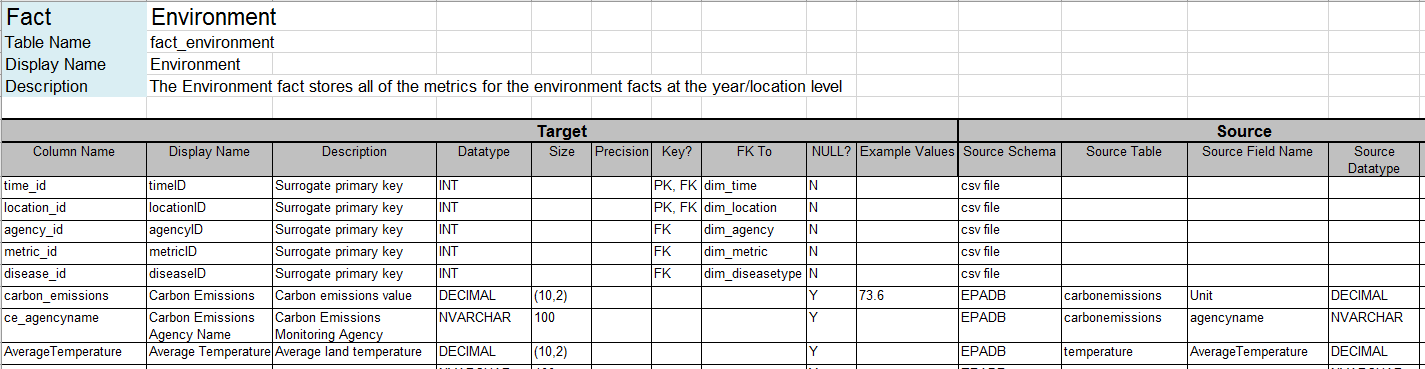
* 1. Source to Target Mapping

The Source to Target Mapping spreadsheet provides the details showing the links between source and target fields, as well as associated ETL processing rules and slowly changing dimension information. The spreadsheet includes individual tabs for each of the dimensions (Time, Agency, Location, Metric, Disease Type), as well as our fact table (Environment). Below we have provided a sample of the spreadsheet for our fact and one of the dimension tables. Please refer to appendix 7.3 for the complete spreadsheet values.

**Ex. Agency Dimension**



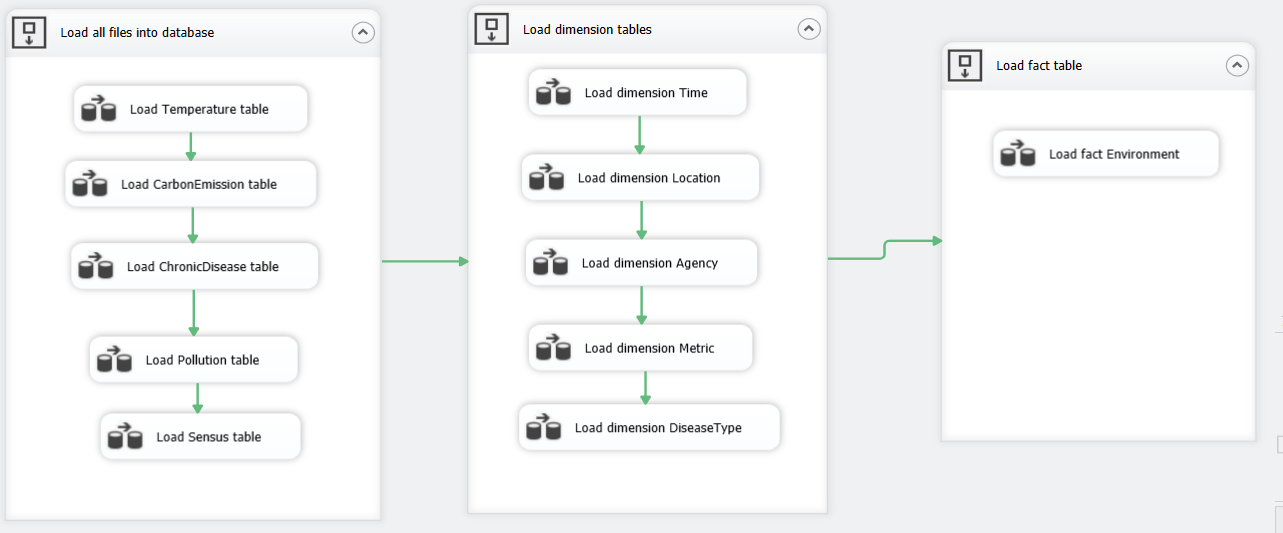
**Ex. Environment Fact**



* 1. ETL Design & Development

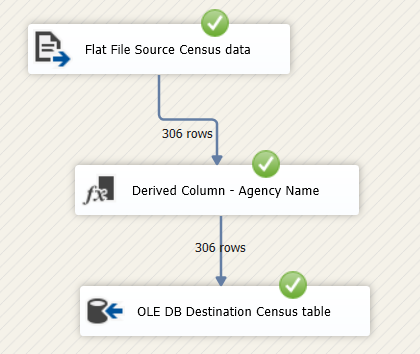
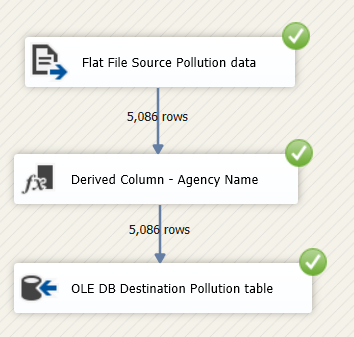
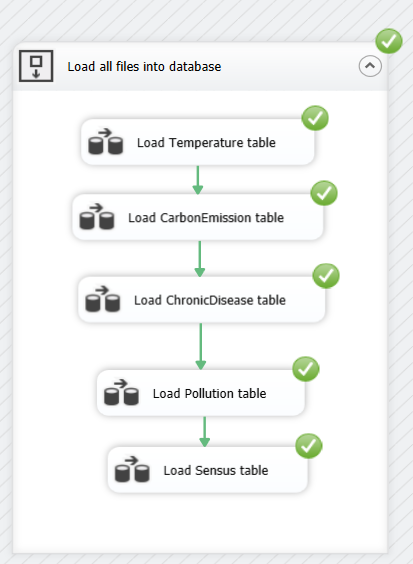
**SSL Package Overview:**

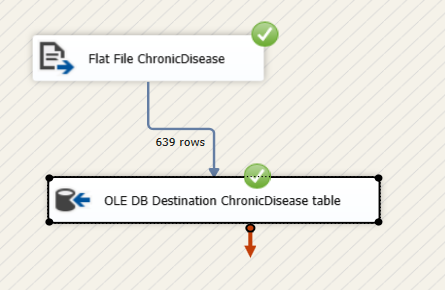
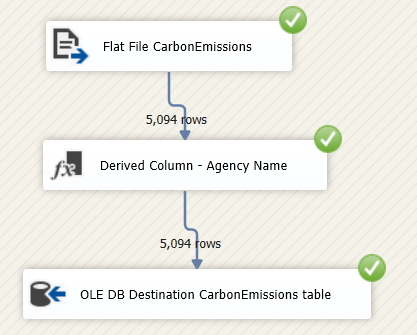
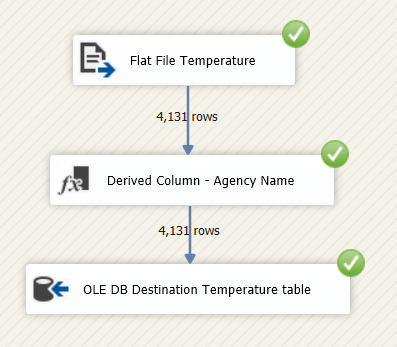
The diagram below shows the high-level sequence of our extract, transform, and load (ETL) process in SSIS. We first loaded the flat files into our EPA database, then used ETL processes to load the dimension and fact tables.



**Step 1: Load flat files into the database**

The first stage of our ETL process involved loading our database tables with the CSV flat file data files. The five source files included one each for pollution, temperature, carbon emissions, chronic diseases, and US census data. We used derived columns for the datasets that did not have an agency name in the original data to populate these values with a default agency name (i.e. for pollution “Environment Protection Agency”).



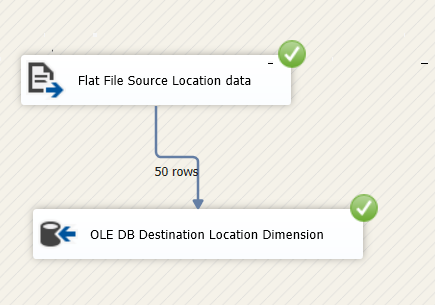
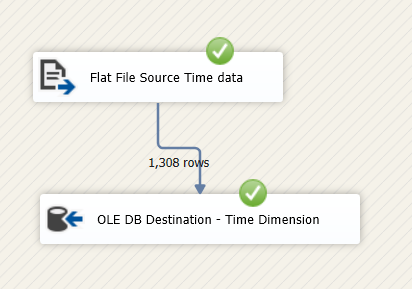
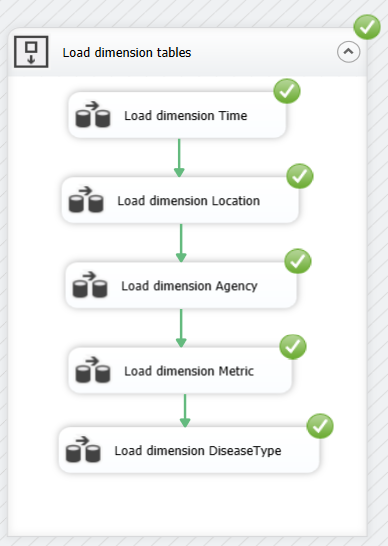


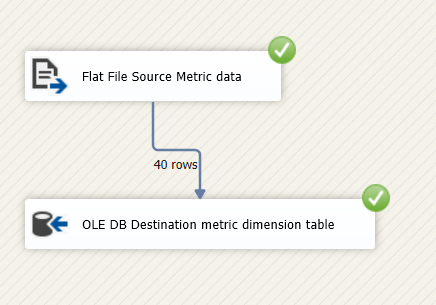
**Step 2: Load Dimension Tables**

Our data warehouse model includes five dimensions of time, location, agency, disease type, and metric. Three of these dimensions are static (time, location, metric), while two will be slowly changing dimensions (agency, diseasetype):

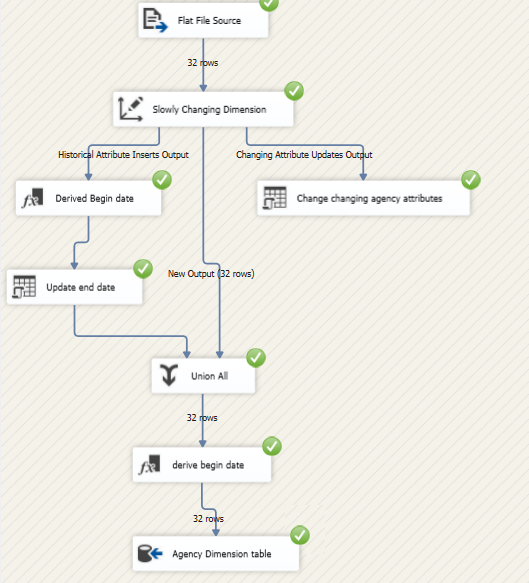
* **Static Dimensions** – Time, Location, Metric
* **Slowly Changing Dimensions (Type 2)** – Agency, Disease Type
  + These will both be handled as type 2 dimensions where we will track the history of the change through adding start and end date fields to our target dimension

The diagram below displays the sequence container with the load dimensions control flow tasks, as well as the individual data flow tasks to load each dimension from the source file/database.





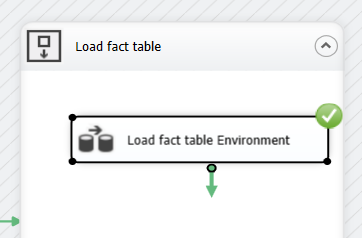
For slowly changing dimensions there is some additional setup required in SSIS. An example of the different steps is provided below for the Agency dimension.



Derived columns were used to capture the begin and end dates needed to store the history information for the slowly changing dimensions.

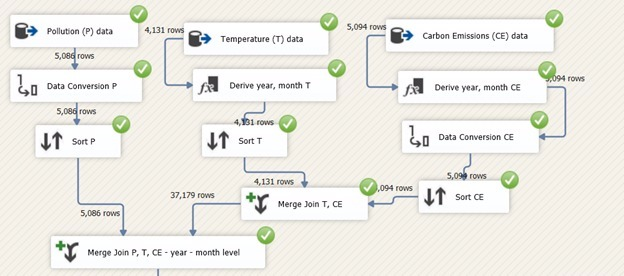
**Step 3: Load Fact table**

After loading the dimensions the fact\_environment table was populated.

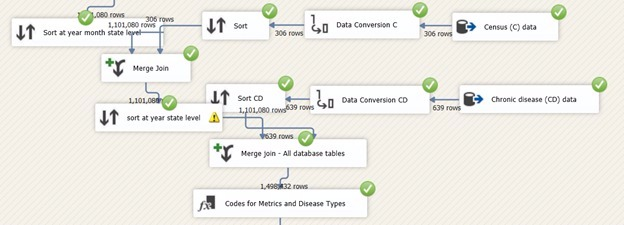


This is where most our ETL processing took place, and included a variety of sorting, merging, and data conversion steps as exhibited in the diagrams below. The end results were to merge the five different tables at the year/month/state level.

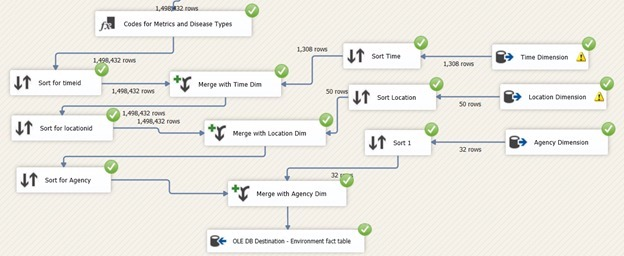
First, Temperature and Carbon emissions datasets were combined at the year month level as temperature data is at the daily level for each state, but carbon emissions data is at the monthly level. Next pollution, temperature, and carbon emissions datasets are combined at the year/month/state level.



The census data is at the year/state level, so this is joined to the previous merged table of (pollution, temperature, emissions) to get the population and birth and death rates. All of the months in the year for a particular state will have the same value for that year.



After merging these tables, ETL codes were created for every measure to map to the metric and disease type tables. Then, the primary keys of time id and location id were obtained from the dimension tables.



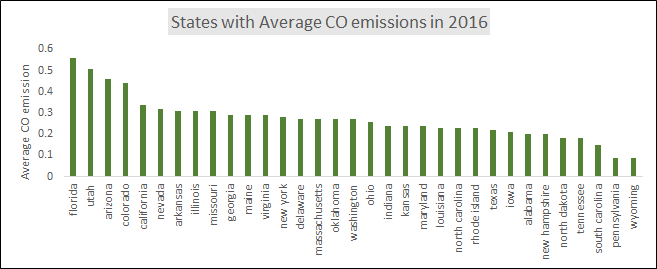
**Alternative Approaches:**

We made a bad choice in terms of project topics, as the datasets we were able to located were at varying levels of detail and not clean. Because of the fact that our data sources are at different levels redundancy was inevitable in the creation of the data warehouse. While we decided to load our source flat files into tables in the database before loading to the data warehouse, there was an option to just load directly from source to data warehouse. The advantages of our choice include the source data being available anytime if the organization wants to go back and modify the data warehouse build in a different format or get additional fields or levels of data detail. The disadvantage would be consuming storage space for potential minimal value add, so this trade-off would have to be considered.

1. Analysis & Results
   1. Addressing Business Requirements

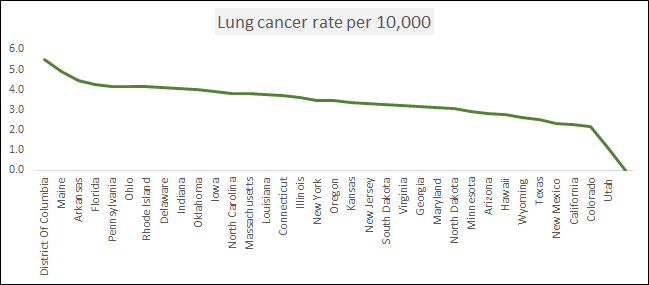
* **For the year 2016, which states had carbon monoxide emissions higher than the acceptable carbon monoxide AQI?**

We obtained 33 states in USA which had carbon monoxide emissions higher than the acceptable AQI. The CO AQI is 0 and we have listed all states which have emissions above 0. Florida, Utah and Arizona have the highest CO emissions for the year 2016.

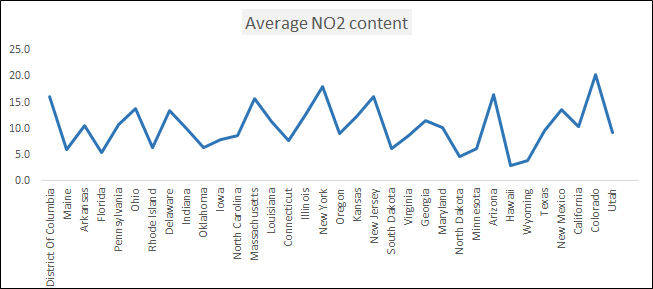


* **What was the annual lung cancer rate per 10,000 for each of the states in USA for the year 2012 and the corresponding levels of Nitrogen dioxide?**

Below the graphs obtained from the results.

****

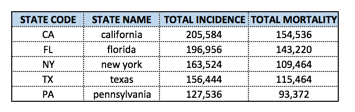
Here, we can see that District of Columbia, Maine, Florida and Pennsylvania were the states with highest lung cancer rate per 10,000. To attribute or derive a correlation between the lung cancer rate, we obtain the results which were plotted below



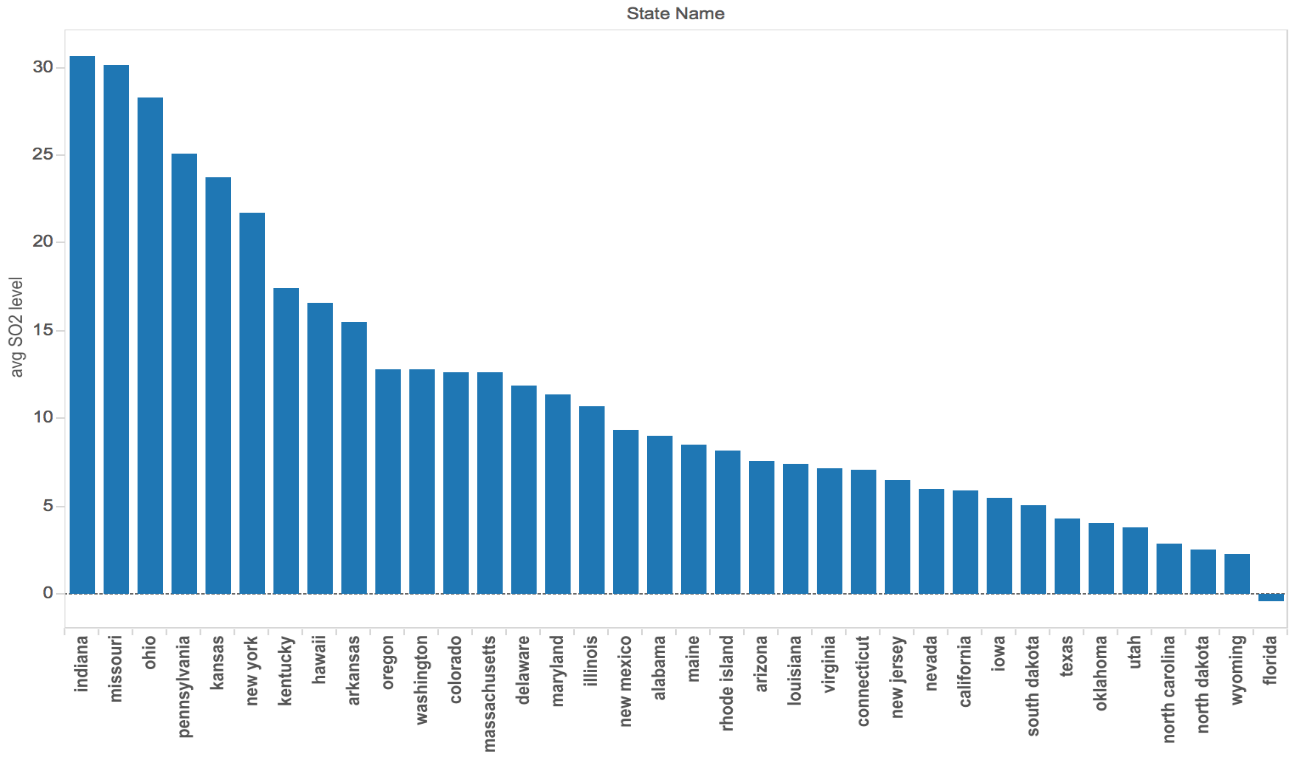
There is no obvious trend in the content of Nitrogen Dioxide pollutant level. The pollutant level is recorded in parts per billion. The highest level of Nitrogen Dioxide for year 2012 was found to be in Colorado, New York, Arizona and New Jersey. However, these states do not show a high lung cancer rate.

* **Which are the 5 states that account for highest incidence of lung cancer and bronchus cancer and what is the mortality?**

As can be seen from the result table below, california has the highest incidence of lung and bronchus cancer averaged over the time period 2008-2012.

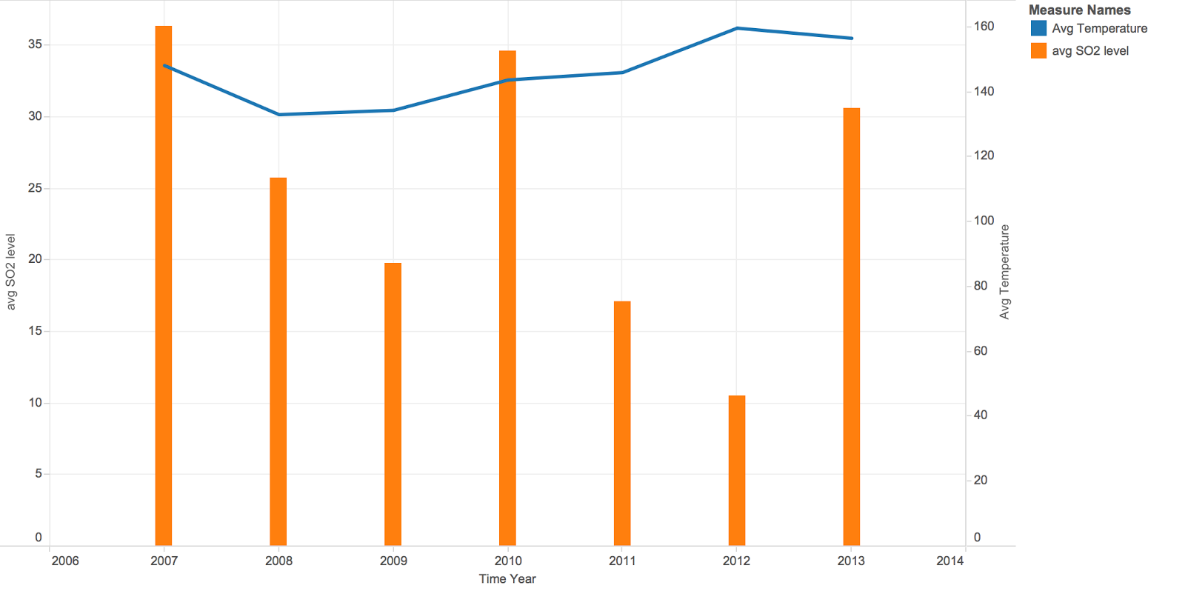
****

* **What is the trend of Sulphur Dioxide and average temperature over a period of time (at a monthly level)?**

****

As can be seen from the graph above, states of Indiana, Missouri and Ohio have unusually high levels SO2. The amount of SO2 in Florida is very low.

In order to understand the impact of SO2 on avg temperature, we selected a state “Indiana” and plotted its the temperature and SO2 trend over the years. As can be seen from the graph below, the average SO2 and average temperature varies greatly over the years. However, we did not see any correlation between the two trends.



1. Conclusion

The primary aim of this exercise was to solve business questions related to the relationship between different environmental factors and incidences of cancer. Due to the diverse data sources and complexity of the datasets, the solution to these questions would not have been possible without having a robust data warehouse structure, as laid out by our team.

Even though the analysis provided was not conclusive because of the varied interdependencies of the datasets, this data warehouse can be utilized by various government and environmental agencies to monitor the level of pollution in the air for different pollutants, observe the trend in average temperature over a period of time, and analyse the incidence/mortality of various diseases. Since this data warehouse brings all datasets and variables under one roof,  this can also be utilized to perform data mining tasks and identify statistically significant relationships between the variables.

The current design for the Data Warehouse has a lot of discrepancies. These primarily arise due to the diverse sources of the data and also lack consistency in the information from various sources. The chronic diseases data consists of data from 2008 to 2012 whereas the temperature data has data starting from year 2010 to 2013. For this reason, we were only able to identify the lung cancer incidence rate for year 2012 which is the latest year. Also, we were unable to find convincing correlations to justify our hypothesis which can be accounted to the lack of overlapping information between data sources.

Overall, this project sets the framework for agencies like the EPA that are working towards validating several hypotheses associated with cancer rates and rises in pollution levels.

1. Bibliography

Environmental Protection Agency. (2016, May 18). *Long-term Exposure to Air Pollution and Development of Cardiovascular Disease*. Retrieved from Research Grants: https://www.epa.gov/research-grants/long-term-exposure-air-pollution-and-development-cardiovascular-disease

Harvard School of Public Health. (2012, Apri 17). *Long-term exposure to air pollution may increase risk of hospitalization for lung, heart disease*. Retrieved from Press Releases: https://www.hsph.harvard.edu/news/press-releases/fine-particles-hospitalizations-lung-health-disease/

Kimball Group. (2011). *The Microsoft Data Warehouse Toolkit, 2nd Edition*. Retrieved from data-warehouse-business-intelligence-resources: http://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/microsoft-data-warehouse-dw-toolkit/

OECD. (2016, September 6). *Air pollution to cause 6-9 million premature deaths and cost 1% GDP by 2060*. Retrieved from Environment: https://www.oecd.org/environment/air-pollution-to-cause-6-9-million-premature-deaths-and-cost-1-gdp-by-2060.htm

1. Appendix
   1. Source Database Create Table Statements



* 1. Target Data Warehouse Create Table Statements



* 1. Source to Target Mapping Spreadsheet



* 1. Analysis SQL Queries

