1. INTRODUCTION

Urbanization happens at a fast pace worldwide therefore metropolitan cities find protecting their air quality an absolute necessity. The heavily populated New York City confronts numerous air pollution problems that negatively affect both environmental conditions and community wellness. NYC air quality changes throughout time and geography will be analyzed through a dimensional data warehouse as the primary focus of this project.

The research aims to merge EPA and NYC Open Data sophisticated information into one interlinked system that enables both business intelligence investigations and clear revelations. A star schema data storage system enables comprehensive monitoring of PM2.5 together with second pollutant measurements for NO_2 and O_3 and health monitoring data to assess air quality impact on human health.

The developed data warehouse system remains a tool for stakeholders such as policymakers and healthcare organizations and public users to understand pollution patterns and find vulnerable areas while generating decisions to decrease environmental and health dangers. The project runs ETL procedures to maintain data quality along with consistency across datasets while complete data is achieved through this implementation. Additionally the application uses Tableau and Neo4j visualization tools to display vital trends and insights that lead to actionable solutions through intuitive interfaces.The initiative focuses on evaluating important research which enhances city sustainability and healthcare.

2. Business Vision and Stakeholders

Business Vision

Decision makers will receive high-quality integrated air quality trend insights in New York City through the primary business vision of the project. The health of urban populations depends heavily on air pollution because it affects both sustainable environments and public welfare while impacting the quality of city life. Respiratory disease and hospitalizations together with mortality rates in NYC act as a result of multiple air pollutants consisting of PM2.5 (fine particles), NO_2 (Nitrogen Dioxide), and O_3 (Ozone).

We design this dimensional data warehouse to deal with the present data problems of separated and inconsistent datasets stemming from EPA and NYC Open Data sources. NYC-Air-Quality-Analysis…. The warehouse integrates data through location-based and time-based dimensions to obtain complete knowledge of air quality variations across different seasons, neighborhoods and years.

The project delivers sophisticated analysis tools which facilitate strategic decision support on a large scale basis. Through this project authorities together with public health organizations gain the ability to pinpoint pollution hotspots and analyze seasonal patterns and predict health impact effects while determining policy effectiveness. Stakeholders use Tableau visualization tools to explore trends interactively and the Neo4j comparative database system helps conduct in-depth relationship-focused analysis on NYC-Air-Quality-Analysis..

Ultimately, the vision is to drive **evidence-based urban planning**, **targeted healthcare initiatives**, and **environmental policies** aimed at improving the overall quality of life for NYC residents.

Stakeholders

The key stakeholders for the NYC Air Quality Analysis project are:

1. **Environmental and Local Government Organisations**
    - Regional offices of the EPA and the NYC Department of Environmental Protection (DEP).
    - Interest: Identifying patterns in air pollution to develop new environmental laws and track adherence.

2. **Public Health Officials and Medical Researchers**
    - NYC Department of Health, local hospitals, research institutions.

- Public health campaign development and intervention strategies receive input from disease-respiratory and cardiovascular information showing air pollution associations.

3. **Urban Planners and Policy Makers**
    - Urban planners, transportation officials.
    - Users can develop cleaner more health-oriented urban areas through pollution pattern-based design while maximizing traffic flow and construction zones alongside industrial sites.

4. **Civil Society Organizations (CSOs) and Environmental Advocates**
    - Clean air initiatives, climate action groups.
    - Consumers should apply transparent data for environmental protection advocacy while raising public awareness about these issues.

5. **General Public and Residents**
    - NYC citizens, vulnerable populations (children, elderly).
    - The public needs trusted up-to-date air quality data which enables them to plan their outdoor time and healthcare needs.

Through integrated data insights enabled by data warehouses every stakeholder group will gain positive impacts which promote both accountability and faster policies while raising public awareness.

3. Data Warehouse Schema Design

Schema Design Overview

A data warehouse with star schema design supports efficient analytical requirements for air quality trend monitoring in New York City. The simplicity and high performance of a star schema along with its capability to organize large data efficiently made it the chosen design.A star schema design enables quick aggregation with single fact tables connected to many dimension tables thus facilitating effective work of Tableau and SSRSSchema NYC-Air-Quality-Analysis… applications.

Fact Table: AirQuality_Fact

AirQuality_Fact serves as the essential fact table where measurements for air quality are stored. A single record within the table shows the observed measurements of a particular pollutant indicator or aspect from both spatial and temporal dimensions.

- **Attributes:**
    - **Time_Key (FK)** – Dependent on Time Dimension
    - **Location_Key (FK)** – Dependent on Location Dimension
    - **Indicator_Key (FK)** – Dependent on Indicator Dimension
    - **Data_Value** – Measurement of pollutant concentration or health indicatorSchemaReport AirQuality_Fact

The data table contains numerical data for PM2.5 levels as well as NO_2 concentrations and O_3 levels and hospitalization rates that are associated with air pollution.

Dimension Tables

All facts in the schema use three dimensional tables that deliver contextual information for their facts.

1. **Time_Dimension**
    - The database system stores time-based information that includes periods, seasons and start date values.
    - Example attributes: Time_Key, Time_Period, Start_DateReport Time_Dimension.

2. **Location_Dimension**
    - The database contains a table for storing geographic data along with county and local district information.

- Example attributes: Location_Key, Geo_Type_Name, Geo_Join_ID, Geo_Place_NameReport Location_Dimension

3. **Indicator_Dimension**
    - The dimension outlines what health indicators and pollutants will be measured during analysis.
    - The example aspects include Indicator_Key, Indicator_ID, Name, Measure, and Measure_Info along with Indicator_Type

The dimensions are designed with complete normalization that lets users split data through different viewpoints (including time period, geographic areas and measurement types).

Justification for Schema Choice

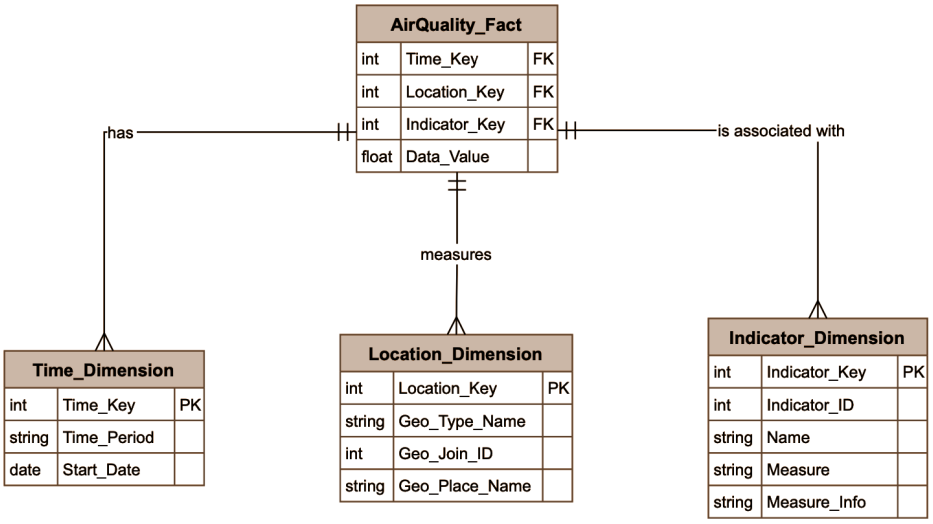
The company selected the star schema instead of alternate designs including snowflake schema because:

- Logical structures based on the star schema make it highly suitable for creating reports and queries because it benefits visual analysis tools such as Tableau.
- The schema enables quick aggregation performance through a reduced need of joins.
- Users with business expertise find it easy to navigate because all dimensions in the model have straightforward links to the fact table.
- Future expansions are made easy through the star schema since business users can incorporate new dimensions or facts like socioeconomic data and weather conditions.

The specifications regarding air quality data being multidimensional, time-series-oriented, and geographically distributed make the star schema the most suitable data model for this project NYC-Air-Quality-Analysis.

Schema Diagram

The following schema diagram visually represents the star schema architecture developed for this project:



4. ETL Process Implementation

Overview

Load and transform and extract procedures formed the main foundation to prepare NYC air quality data for analytical purposes. The diverse and inconsistent nature of data obtained from EPA and NYC Open Data portals demanded an advanced ETL pipeline which extracted and cleaned and transformed required information into an analytical structure for efficient loading into the NYC-Air-Quality-Analysis dimensional data warehouse.

Extract Phase

Multiple sources provided the data which was extracted for the study.

- EPA Air Quality Datasets: Detailed records of pollutant measurements (PM2.5, NO_2, O_3).
- The NYC Open Data platform contains additional health-related metrics besides those found in public data.

SSIS Microsoft SQL Server Integration Services functioned first to run automated processing of raw files for extracting both structured and semi-structured data.
NYC-Air-Quality-Analysis…. The extraction tasks focused on reading CSV files as well as Excel documents and database entries while maintaining metadata and data type integrity.

Transform Phase

The transformation phase demanded the most significant effort within ETL pipelines. Standardization and cleaning operations were conducted on the data through multiple procedures.

- For null and missing contaminant measurement values the ETL process followed business logic for imputation or removal.
- The data types across all fields underwent appropriate transformations which delivered standardized data consistency through this process.
- Standardization functions normalized textual elements in the data through a process which eradicated duplicate entries and enhanced linking between dimension keys.
- During this phase the project team established rational foreign key associations between the fact tables and dimension tables.

Record changes were developed through the use of SQL Server Data Tools (SSDT) and specific SQL scripts.

Load Phase

The data warehouse received the cleaned transformed information after the data delivery stage.

- **Fact Table:** AirQuality_Fact
- **Dimension Tables:** Time_Dimension, Location_Dimension, Indicator_Dimension

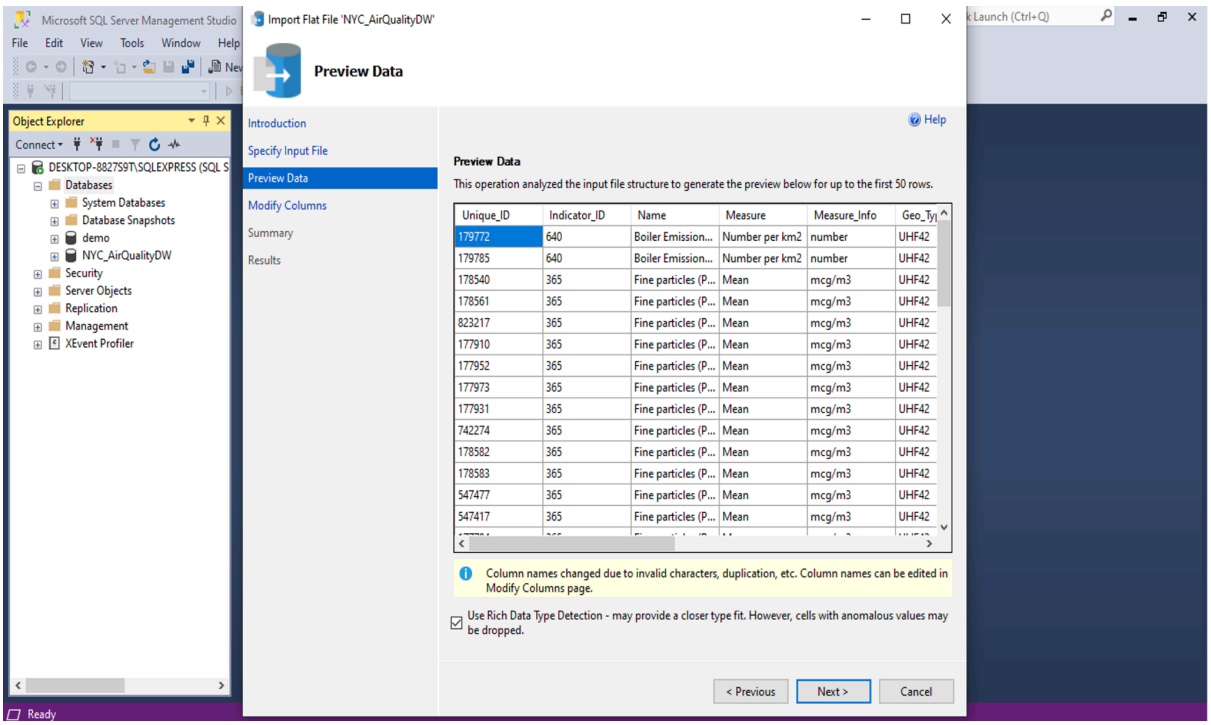
SSIS packages implemented control flow tasks for dependency management and transaction integrity during automated data loading operations. The system included error handling components which detected loading failures while automatically repairing issues to prevent interruptions of the whole process.

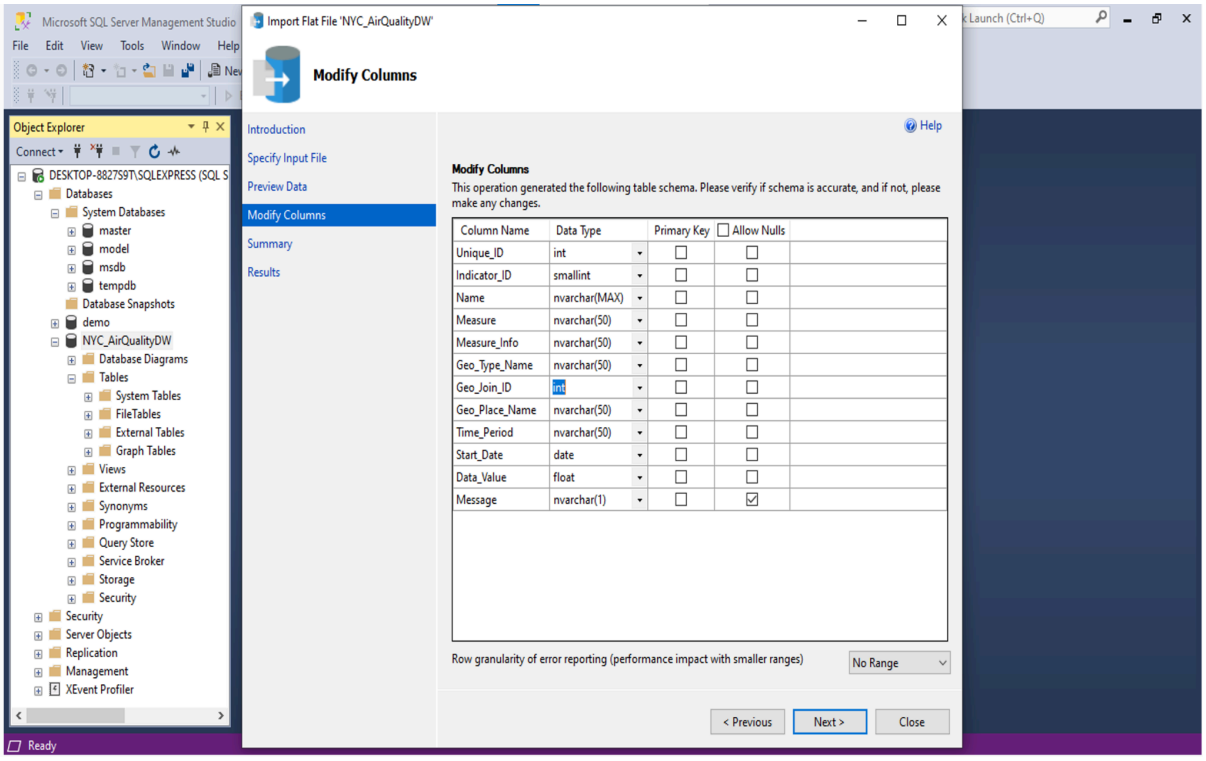
ETL Implementation Screenshots

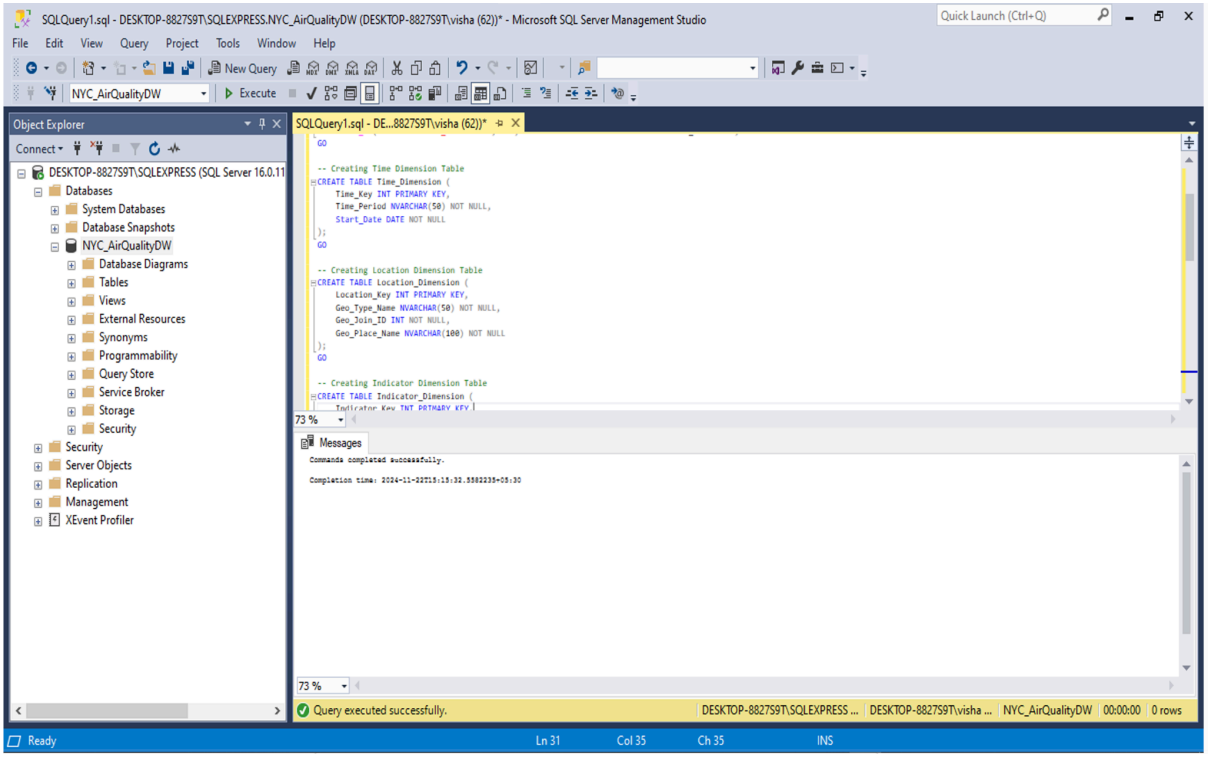
The following screenshots (attached) show:

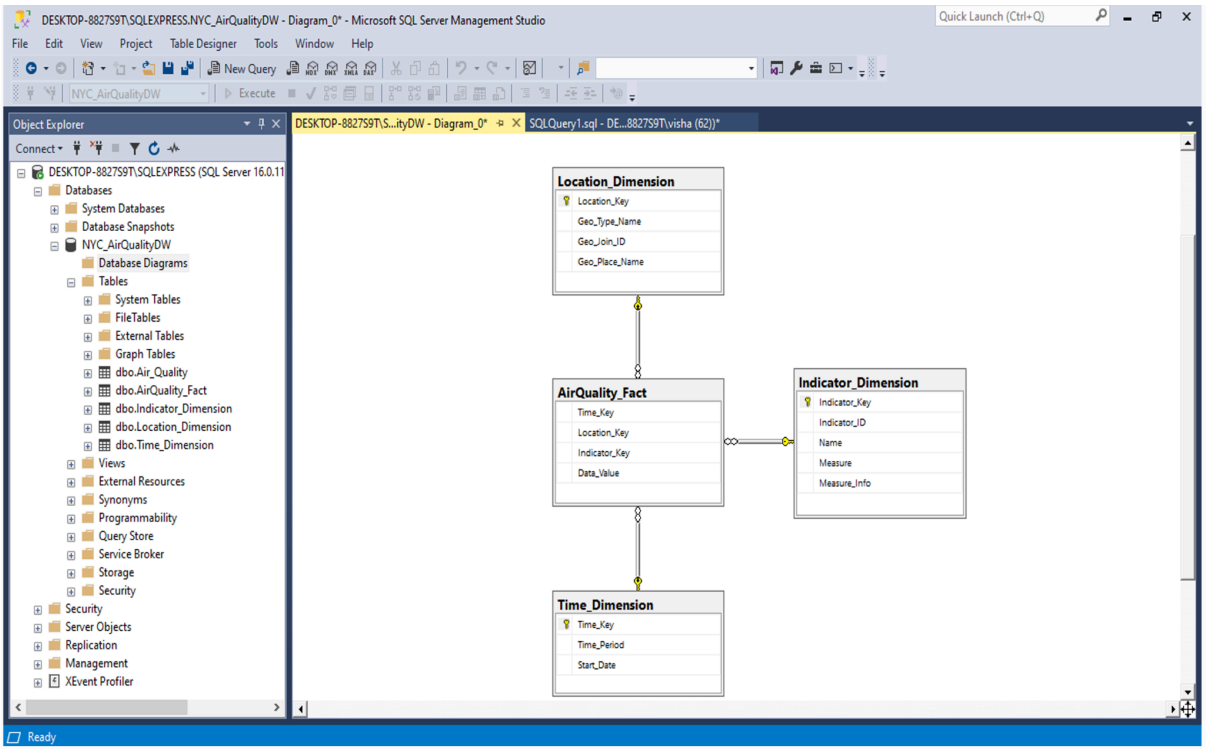
- SSIS package designer setup
- Data flow configurations
- The SSIS package makes use of transformation tasks which include Derived Column, Lookup and Data Transformation.
- Final loading processes

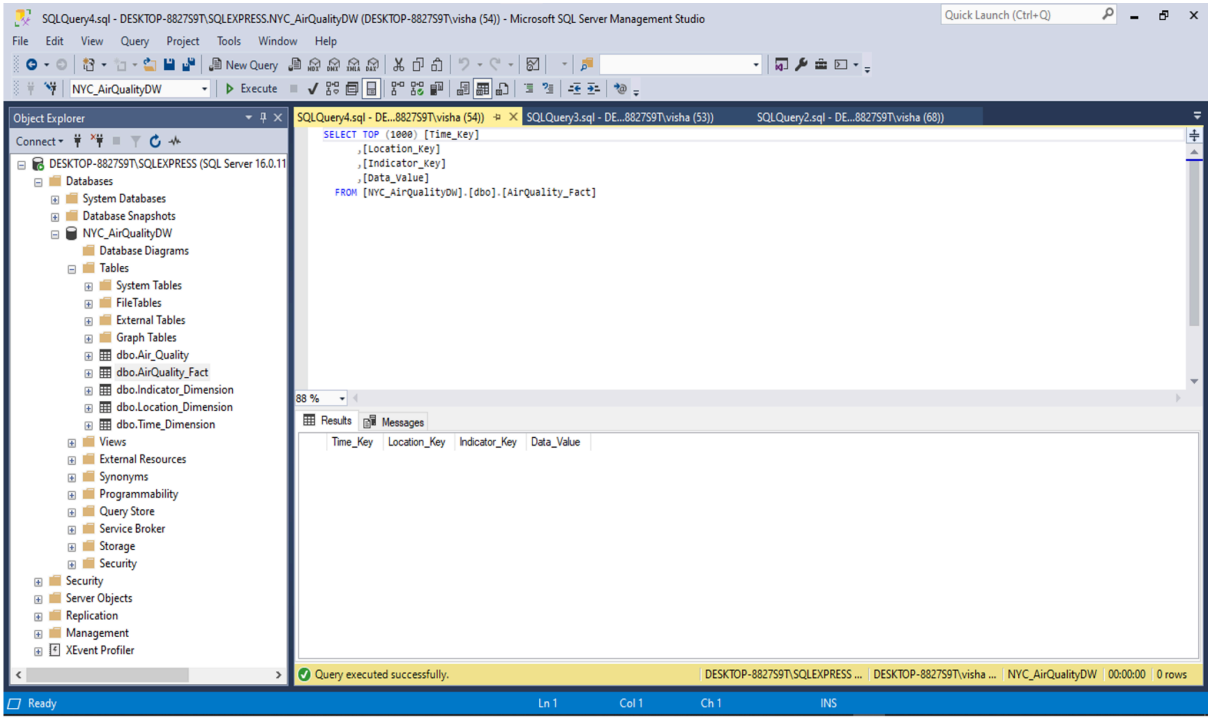
Visual indicators depict the entire data Moving journey starting from operational sources and finally ending with a dimensional model structure suitable for analytical purposes.

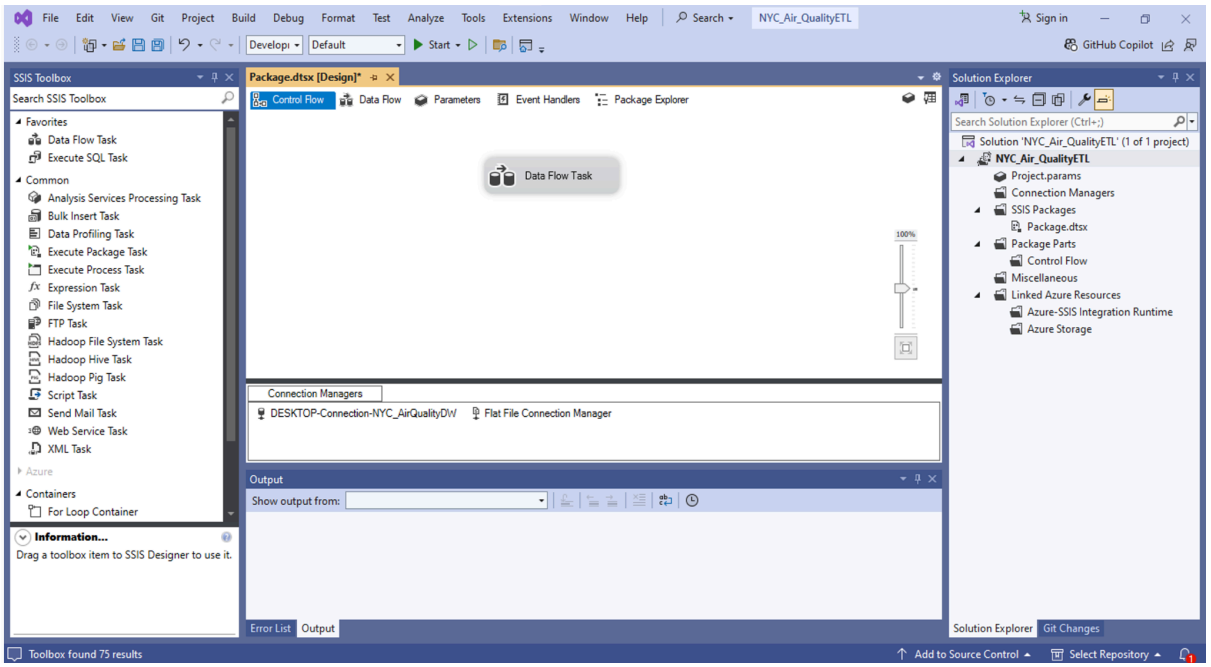

Microsoft SQL Server Management Studio
File Edit View Tools Window Help
Object Explorer
Connect
DESKTOP-8827S9T\SQLEXPRESS (SQL S
Databases
System Databases
Database Snapshots
demo
NYC_AirQualityDW
Security
Server Objects
Replication
Management
XEvent Profiler
Ready
Import Flat File 'NYC_AirQualityDW'
Preview Data
Introduction
Specify Input File
Preview Data
Modify Columns
Summary
Results
Help
Preview Data
This operation analyzed the input file structure to generate the preview below for up to the first 50 rows.
Unique_ID | Indicator_ID | Name | Measure | Measure_Info | Geo_Ty
179772 | 640 | Boiler Emission... | Number per km2 | number | UHF42
179785 | 640 | Boiler Emission... | Number per km2 | number | UHF42
178540 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
178561 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
823217 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
177910 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
177952 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
177973 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
177931 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
742274 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
178582 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
178583 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
547477 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
547417 | 365 | Fine particles (P... | Mean | mcg/m3 | UHF42
Column names changed due to invalid characters, duplication, etc. Column names can be edited in Modify Columns page.
Use Rich Data Type Detection - may provide a closer type fit. However, cells with anomalous values may be dropped.
< Previous
Next >
Cancel

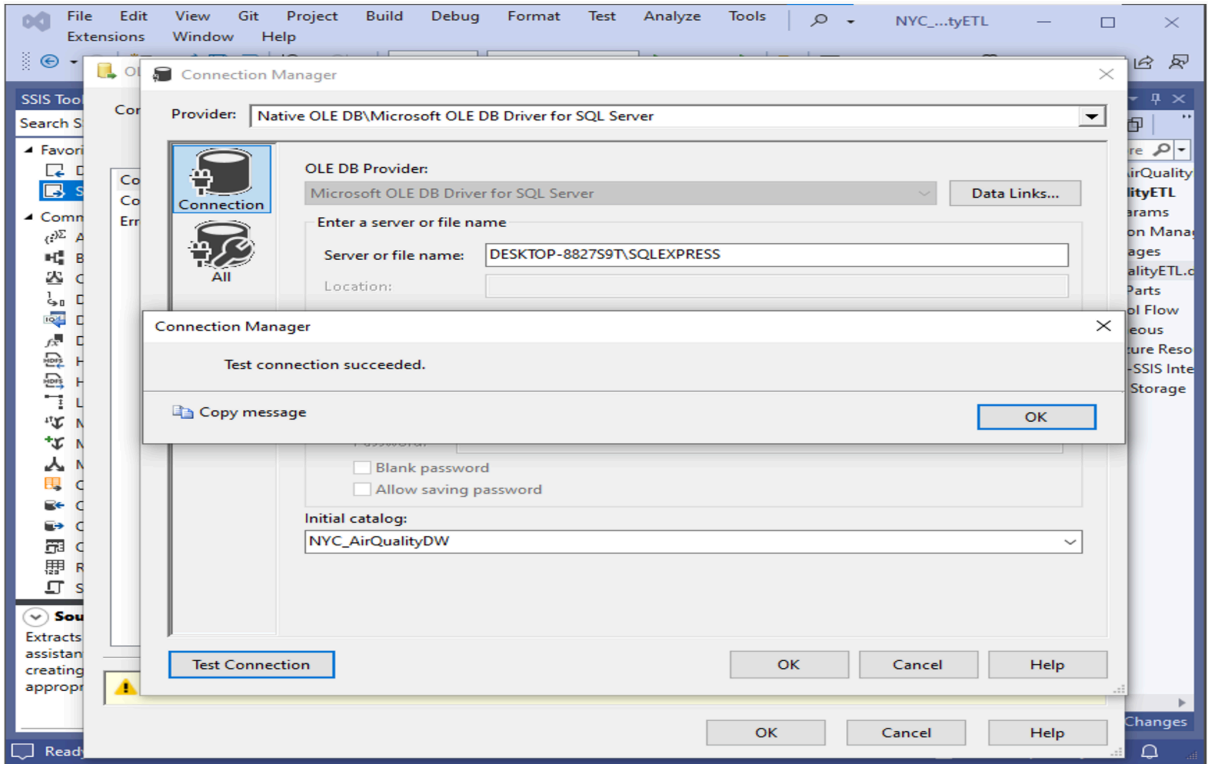

Microsoft SQL Server Management Studio
File Edit View Tools Window Help
Object Explorer
Connect
DESKTOP-8827S9T\SQLEXPRESS (SQL S
Databases
System Databases
master
model
msdb
tempdb
Database Snapshots
demo
NYC_AirQualityDW
Database Diagrams
Tables
System Tables
FileTables
External Tables
Graph Tables
Views
External Resources
Synonyms
Programmability
Query Store
Service Broker
Storage
Security
Security
Server Objects
Replication
Management
XEvent Profiler
Ready
Import Flat File 'NYC_AirQualityDW'
Modify Columns
Introduction
Specify Input File
Preview Data
Modify Columns
Summary
Results
Help
Modify Columns
This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.
Column Name | Data Type | Primary Key | Allow Nulls
Unique_ID | int
Indicator_ID | smallint
Name | nvarchar(MAX)
Measure | nvarchar(50)
Measure_Info | nvarchar(50)
Geo_Type_Name | nvarchar(50)
Geo_Join_ID | int
Geo_Place_Name | nvarchar(50)
Time_Period | nvarchar(50)
Start_Date | date
Data_Value | float
Message | nvarchar(1)
Row granularity of error reporting (performance impact with smaller ranges)
No Range
< Previous
Next >
Close

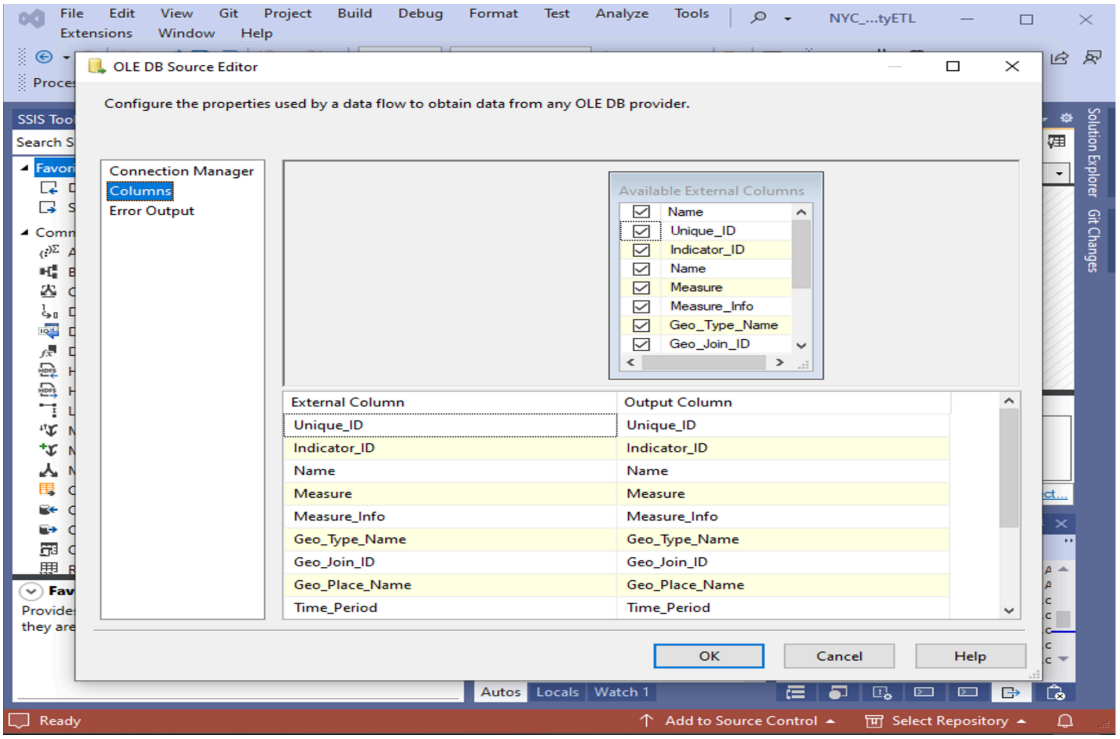

SQLQuery1.sql - DESKTOP-8827S9T\SQLEXPRESS.NYC_AirQualityDW (DESKTOP-8827S9T\visha (62))* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Query Project Tools Window Help
New Query
Execute
NYC_AirQualityDW
Object Explorer
Connect
DESKTOP-8827S9T\SQLEXPRESS (SQL Server 16.0.11
Databases
System Databases
Database Snapshots
NYC_AirQualityDW
Database Diagrams
Tables
Views
External Resources
Synonyms
Programmability
Query Store
Service Broker
Storage
Security
Security
Server Objects
Replication
Management
XEvent Profiler
SQLQuery1.sql - DE...8827S9T\visha (62))*
GO
-- Creating Time Dimension Table
CREATE TABLE Time_Dimension (
Time_Key INT PRIMARY KEY,
Time_Period NVARCHAR(50) NOT NULL,
Start_Date DATE NOT NULL
);
GO
-- Creating Location Dimension Table
CREATE TABLE Location_Dimension (
Location_Key INT PRIMARY KEY,
Geo_Type_Name NVARCHAR(50) NOT NULL,
Geo_Join_ID INT NOT NULL,
Geo_Place_Name NVARCHAR(100) NOT NULL
);
GO
-- Creating Indicator Dimension Table
CREATE TABLE Indicator_Dimension (
73 %
Messages
Commands completed successfully.
Completion time: 2024-11-22T15:15:32.5582235+05:30
73 %
Query executed successfully.
DESKTOP-8827S9T\SQLEXPRESS ...
DESKTOP-8827S9T\visha ...
NYC_AirQualityDW
00:00:00
0 rows
Ready
Ln 31
Col 35
Ch 35
INS

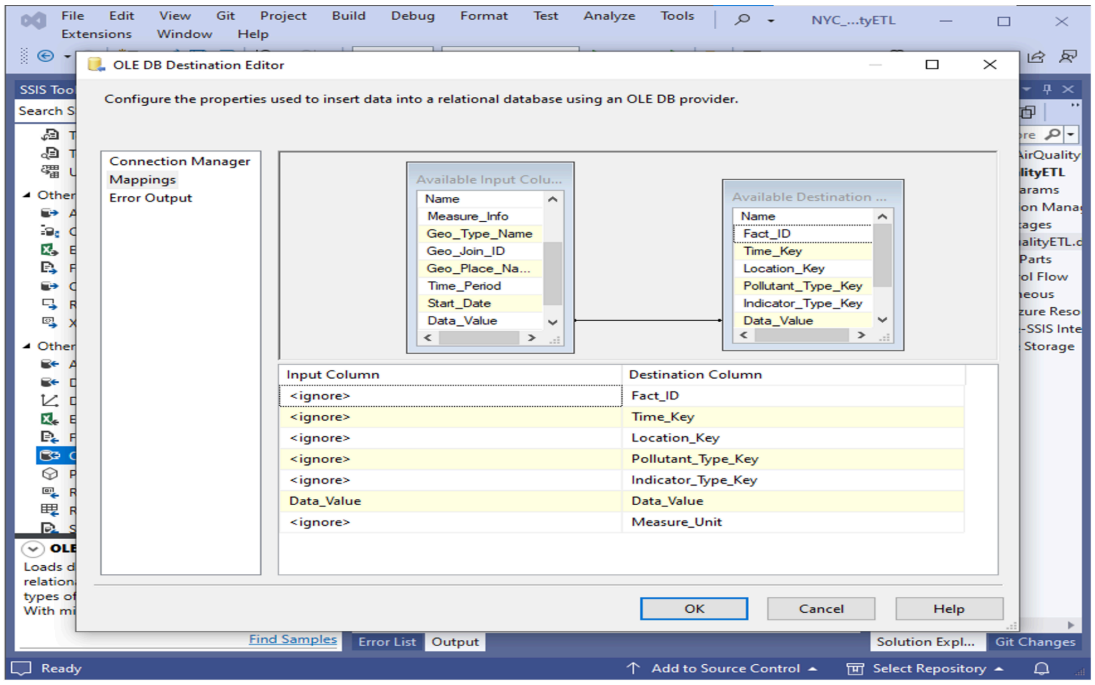

DESKTOP-8827S9T\SQLEXPRESS.NYC_AirQualityDW - Diagram_0* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Project Table Designer Tools Window Help
New Query
Execute
NYC_AirQualityDW
Object Explorer
Connect
DESKTOP-8827S9T\SQLEXPRESS (SQL Server 16.0.11
Databases
System Databases
Database Snapshots
NYC_AirQualityDW
Database Diagrams
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.Air_Quality
dbo.AirQuality_Fact
dbo.Indicator_Dimension
dbo.Location_Dimension
dbo.Time_Dimension
Views
External Resources
Synonyms
Programmability
Query Store
Service Broker
Storage
Security
Security
Server Objects
Replication
Management
XEvent Profiler
DESKTOP-8827S9T\S...ityDW - Diagram_0*
SQLQuery1.sql - DE...8827S9T\visha (62))*
Location_Dimension
Location_Key
Geo_Type_Name
Geo_Join_ID
Geo_Place_Name
AirQuality_Fact
Time_Key
Location_Key
Indicator_Key
Data_Value
Indicator_Dimension
Indicator_Key
Indicator_ID
Name
Measure
Measure_Info
Time_Dimension
Time_Key
Time_Period
Start_Date
Ready

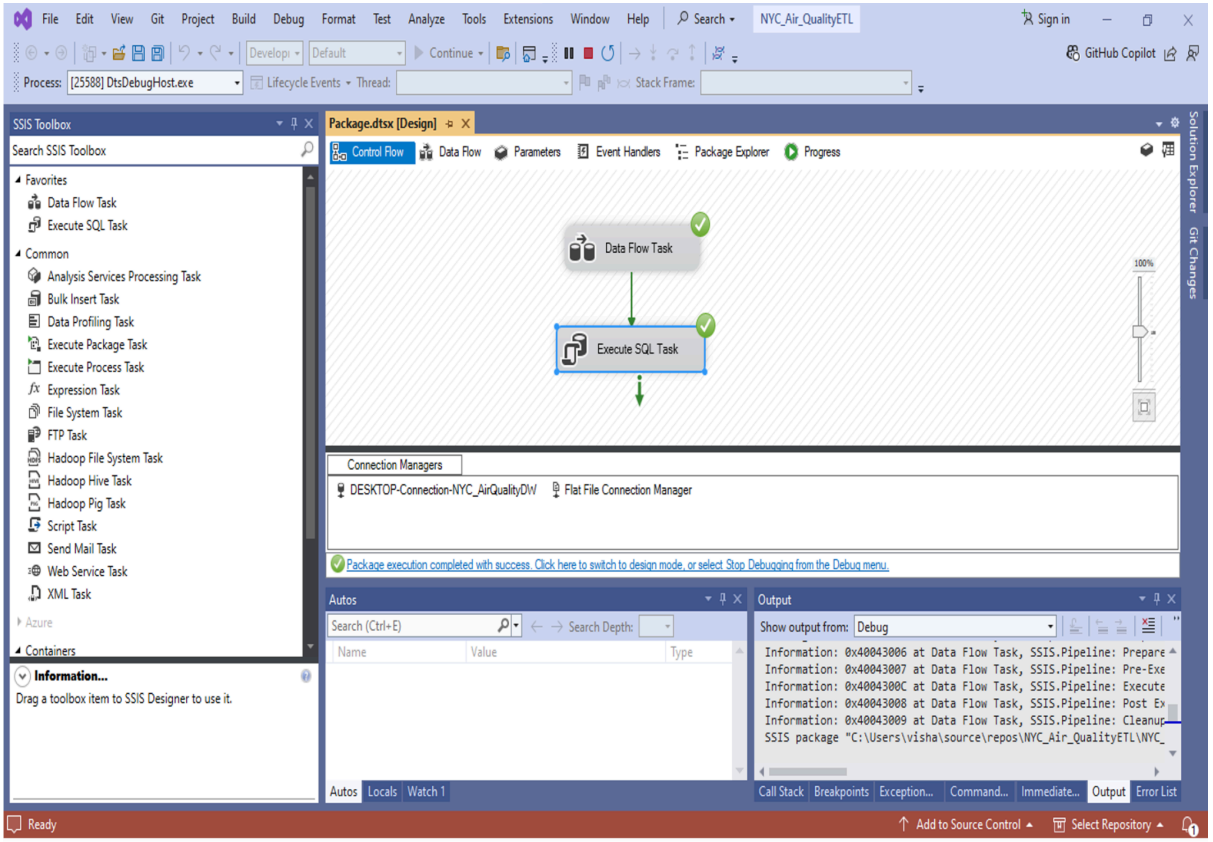

SQLQuery4.sql - DESKTOP-8827S9T\SQLEXPRESS.NYC_AirQualityDW (DESKTOP-8827S9T\visha (54)) - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Query Project Tools Window Help
New Query
NYC_AirQualityDW
Execute
Object Explorer
Connect
DESKTOP-8827S9T\SQLEXPRESS (SQL Server 16.0.11
Databases
System Databases
Database Snapshots
NYC_AirQualityDW
Database Diagrams
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.Air_Quality
dbo.AirQuality_Fact
dbo.Indicator_Dimension
dbo.Location_Dimension
dbo.Time_Dimension
Views
External Resources
Synonyms
Programmability
Query Store
Service Broker
Storage
Security
Security
Server Objects
Replication
Management
XEvent Profiler
SQLQuery4.sql - DE...8827S9T\visha (54))
SQLQuery3.sql - DE...8827S9T\visha (53))
SQLQuery2.sql - DE...8827S9T\visha (68))
SELECT TOP (1000) [Time_Key]
,[Location_Key]
,[Indicator_Key]
,[Data_Value]
FROM [NYC_AirQualityDW].[dbo].[AirQuality_Fact]
88 %
Results
Messages
Time_Key
Location_Key
Indicator_Key
Data_Value
Query executed successfully.
DESKTOP-8827S9T\SQLEXPRESS ...
DESKTOP-8827S9T\visha ...
NYC_AirQualityDW
00:00:00
0 rows
Ready
Ln 1
Col 1
Ch 1
INS

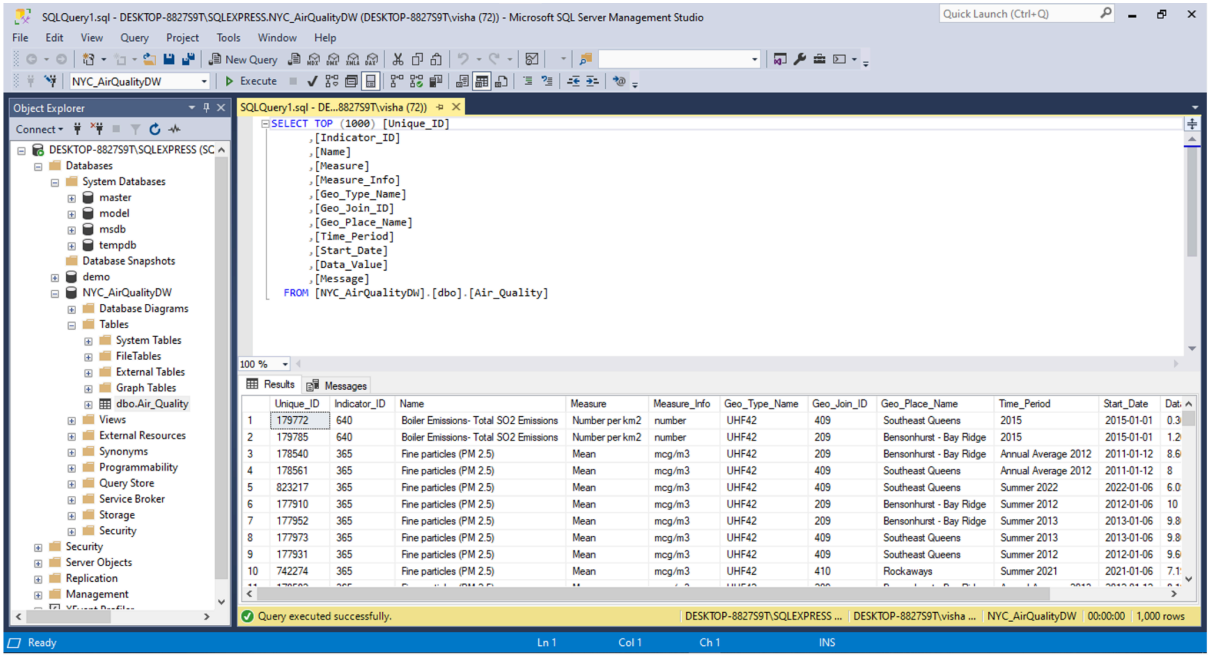

File Edit View Git Project Build Debug Format Test Analyze Tools Extensions Window Help
Search
NYC_Air_QualityETL
Sign in
GitHub Copilot
Start
SSIS Toolbox
Search SSIS Toolbox
Favorites
Data Flow Task
Execute SQL Task
Common
Analysis Services Processing Task
Bulk Insert Task
Data Profiling Task
Execute Package Task
Execute Process Task
Expression Task
File System Task
FTP Task
Hadoop File System Task
Hadoop Hive Task
Hadoop Pig Task
Script Task
Send Mail Task
Web Service Task
XML Task
Azure
Containers
For Loop Container
Information...
Drag a toolbox item to SSIS Designer to use it.
Package.dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Data Flow Task
Connection Managers
DESKTOP-Connection-NYC_AirQualityDW
Flat File Connection Manager
Output
Show output from:
Error List
Output
Solution Explorer
Search Solution Explorer (Ctrl+;)
Solution 'NYC_Air_QualityETL' (1 of 1 project)
NYC_Air_QualityETL
Project.params
Connection Managers
SSIS Packages
Package.dtsx
Package Parts
Control Flow
Miscellaneous
Linked Azure Resources
Azure-SSIS Integration Runtime
Azure Storage
Solution Explorer
Git Changes
Toolbox found 75 results
Add to Source Control
Select Repository


File Edit View Git Project Build Debug Format Test Analyze Tools
Extensions Window Help
Connection Manager
Provider:
Native OLE DB\Microsoft OLE DB Driver for SQL Server
Connection
All
OLE DB Provider:
Microsoft OLE DB Driver for SQL Server
Data Links...
Enter a server or file name
Server or file name:
DESKTOP-8827S9T\SQLEXPRESS
Location:
Connection Manager
Test connection succeeded.
Copy message
OK
Blank password
Allow saving password
Initial catalog:
NYC_AirQualityDW
Test Connection
OK
Cancel
Help
OK
Cancel
Help


OLE DB Source Editor
Configure the properties used by a data flow to obtain data from any OLE DB provider.
Connection Manager
Columns
Error Output
Available External Columns
Name
Unique_ID
Indicator_ID
Name
Measure
Measure_Info
Geo_Type_Name
Geo_Join_ID
External Column
Output Column
Unique_ID
Indicator_ID
Name
Measure
Measure_Info
Geo_Type_Name
Geo_Join_ID
Geo_Place_Name
Time_Period
OK
Cancel
Help


OLE DB Destination Editor
Configure the properties used to insert data into a relational database using an OLE DB provider.
Connection Manager
Mappings
Error Output
Available Input Colu...
Name
Measure_Info
Geo_Type_Name
Geo_Join_ID
Geo_Place_Na...
Time_Period
Start_Date
Data_Value
Available Destination ...
Name
Fact_ID
Time_Key
Location_Key
Pollutant_Type_Key
Indicator_Type_Key
Data_Value
Input Column
Destination Column
<ignore>
Fact_ID
<ignore>
Time_Key
<ignore>
Location_Key
<ignore>
Pollutant_Type_Key
<ignore>
Indicator_Type_Key
Data_Value
Data_Value
<ignore>
Measure_Unit
OK
Cancel
Help





Validation and Testing

Validation queries were written to:

- Rows ers should verify that the same number of rows exists between the source and target systems
- The system should check the integrity of foreign key relationships between its tables.
- Perform an examination for absent or peculiar data values after data load concludes

5. Reports and Visualizations

Overview

The NYC air quality data warehouse received its actionable insights through visualization and analytical reporting tools based on Tableau and SQL Server Reporting Services (SSRS). The created visual representations help stakeholders examine data trends through time-based and location-based and indicator-based analyses.

The monitoring reports combine pollutant measurement data with details of healthcare issues linked to polluted air environments. By studying problems through multiple perspectives this approach delivers full knowledge about environmental as well as social outcomes.

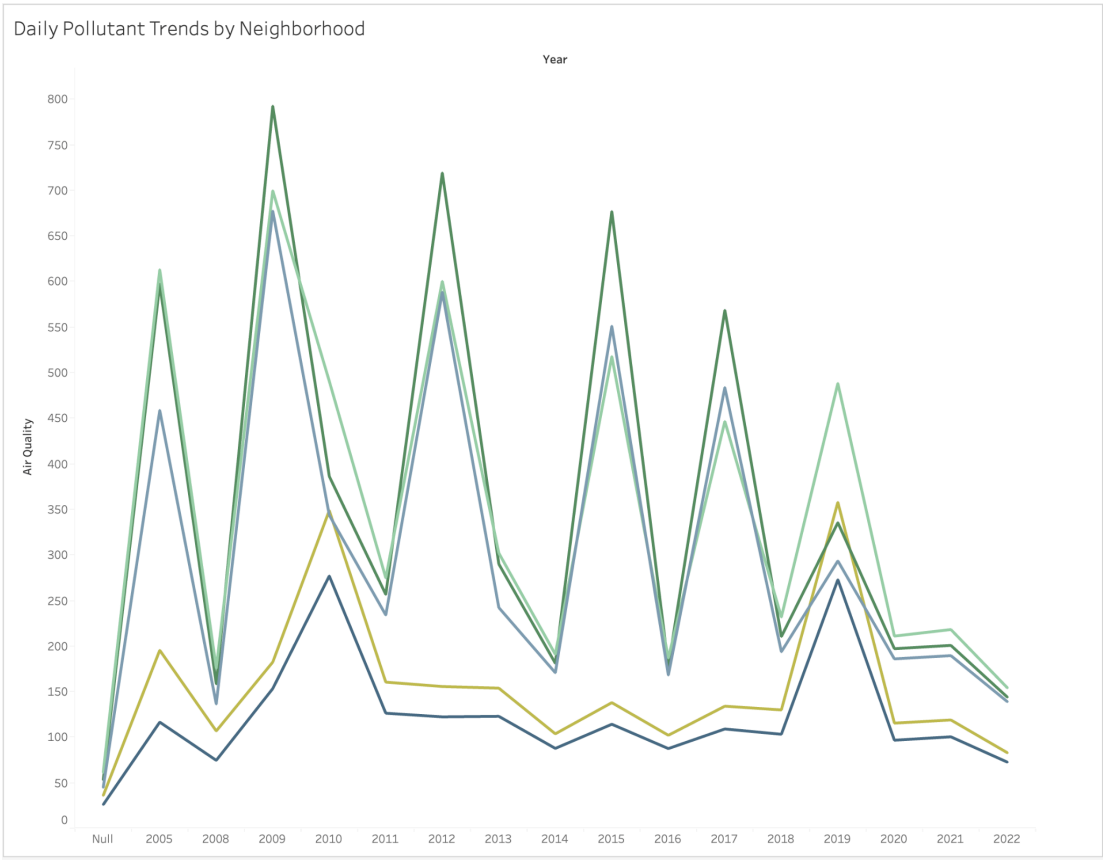
Tableau Visualizations

Fourth the key visualizations were improved across Tableau platforms and will further optimize data analysis.

The following four major Tableau visualizations were developed:

1. **Daily Pollutant Trends by Neighborhood.**
    - **What it is:** A line chart showing air quality index, or AQI, data daily in various New York City neighborhoods from 2005 to 2021. The colored lines show pollution levels in each district over time.

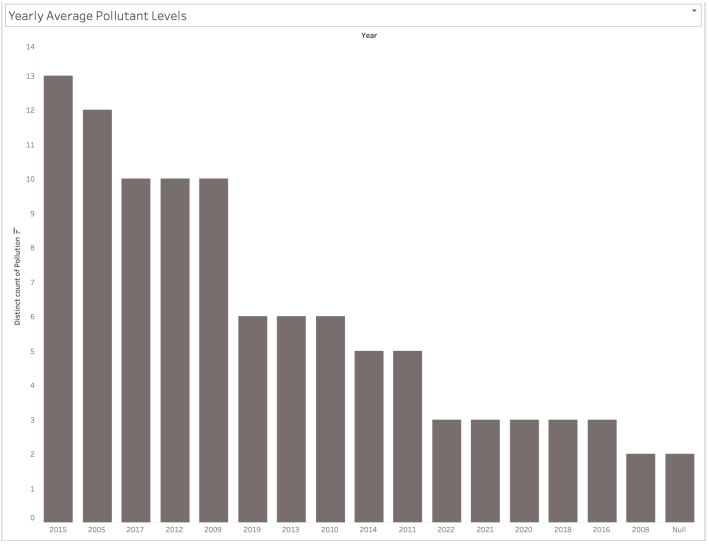
    - **Takeaway:** Pollution levels saw a dramatic seasonal increase each winter but had declined slowly since about 2010.



2. Yearly Average Pollutant Levels

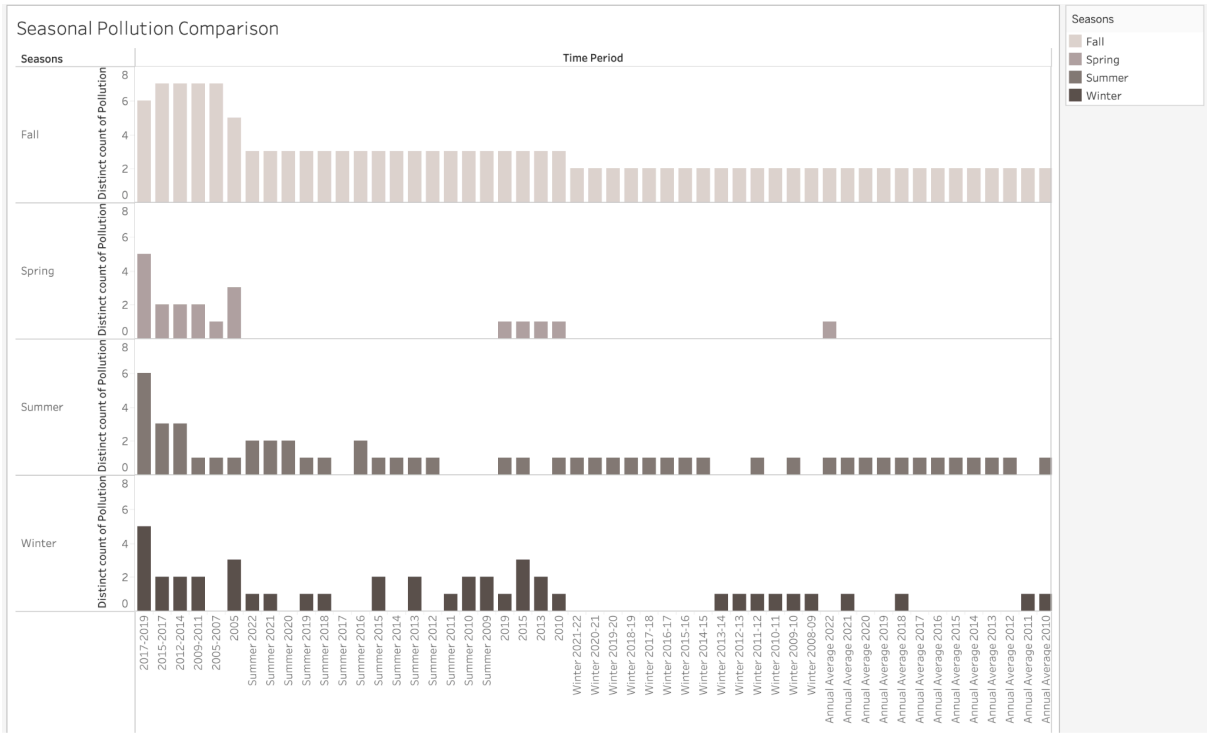
- Bar chart comparing the average annual number of unique pollution events, in many ways signaling improved air quality over time.

- **Bottom line:** Since 2005, pollution counts have decreased, indicating that awareness of and policies around environmental health are having an impact.

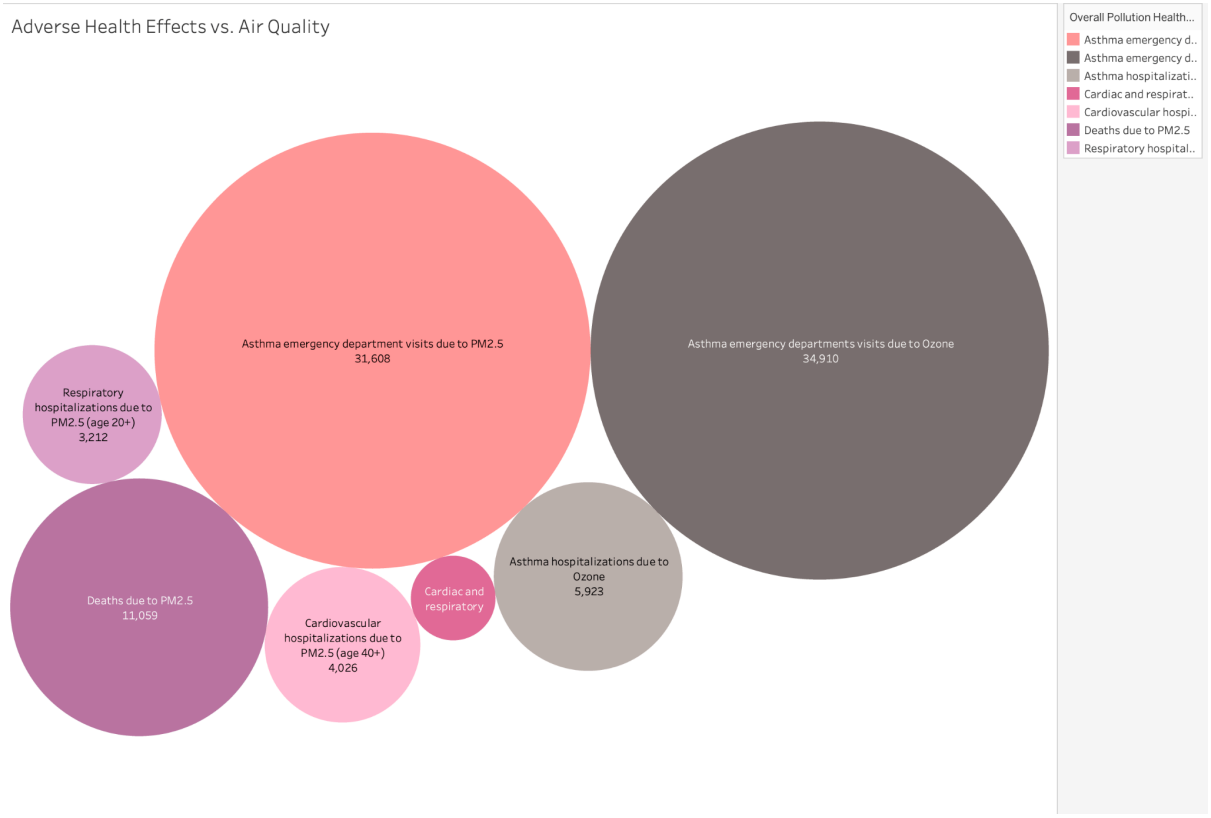


3. **Seasonal Pollution Comparison**
    - **Description:** A multi-year set of group bars showing how many pollutants were counted in each season—Fall, Spring, Summer, Winter—to reveal the seasonality of air quality.

    - **Bottom line:** Wintertime brings the highest levels of pollutants, likely because of heating sources and less air movement.



4. **Adverse Health Effects vs. Air Quality**
    - **Description:**
      A bubble chart correlating adverse health outcomes (e.g., asthma emergency visits, respiratory hospitalizations, deaths) with pollution exposure levels.
    - **Key Insight:**
      PM2.5 and O_3 pollution are strongly linked to spikes in asthma emergencies and cardiovascular issues, emphasizing the critical impact of air quality on public health.



SSRS Reports

Additional to Tableau visualizations SSRS reports were developed with four main sections that included:

- Daily pollutant trends by neighborhood.
- Yearly average pollutant levels.
- Seasonal pollution comparison.
- People experience adverse public health effects as a result of air quality modifications.

The SSRS system developed comprehensive tabular documents which official divisions transmitted to both public health departments as well as environmental agencies.

	Time_Key	Location_Key	Indicator_Key	Data_Value
1	55	5	9	19.6000003814697
2	55	5	10	1
3	12	5	2	21.6000003814697
4	34	5	2	11.8999996185303
5	35	5	16	7.5
6	14	5	1	11.8000001907349
7	34	5	1	5.90000009536743
8	29	5	11	0.6000000238418…
9	36	5	9	21
10	9	5	9	21.2000007629395
11	6	5	16	7.5
12	25	5	11	1.20000004768372
13	10	5	5	0.1000000014901…
14	12	5	1	11.8000001907349
15	33	5	2	18.7000007629395
16	10	5	7	2.70000004768372
17	10	5	6	0
18	24	5	3	35.2000007629395
19	8	5	3	34.2000007629395
20	51	5	1	7.80000019073486

Report_Indication_facts

	Indicator_Key	Indicator_ID	Name	Measure
1	1	365	Fine particles (PM 2.5)	Mean
2	2	375	Nitrogen dioxide (NO2)	Mean
3	3	386	Ozone (O3)	Mean
4	4	639	Deaths due to PM2.5	Estimated annual rate (age 3…
5	5	640	Boiler Emissions- Total SO2 Emissi…	Number per km2
6	6	641	Boiler Emissions- Total PM2.5 Emis…	Number per km2
7	7	642	Boiler Emissions- Total NOx Emissi…	Number per km2
8	8	643	Annual vehicle miles traveled	Million miles
9	9	644	Annual vehicle miles traveled (cars)	Million miles
10	10	645	Annual vehicle miles traveled (trucks)	Million miles
11	11	646	Outdoor Air Toxics - Benzene	Annual average concentration
12	12	647	Outdoor Air Toxics - Formaldehyde	Annual average concentration
13	13	648	Asthma emergency department visit…	Estimated annual rate (under…
14	14	650	Respiratory hospitalizations due to P…	Estimated annual rate
15	15	651	Cardiovascular hospitalizations due…	Estimated annual rate
16	16	652	Cardiac and respiratory deaths due t…	Estimated annual rate
17	17	653	Asthma emergency departments visi…	Estimated annual rate (under…
18	18	655	Asthma hospitalizations due to Ozone	Estimated annual rate (under…
19	19	657	Asthma emergency department visit…	Estimated annual rate (age 1…
20	20	659	Asthma emergency departments visi…	Estimated annual rate (age 1…
21	21	661	Asthma hospitalizations due to Ozone	Estimated annual rate (age 1…

Report Indicator_Type_Dimension

	Location_Key	Geo_Type_Name	Geo_Join_ID	Geo_Place_Name
1	1	Borough	1	Bronx
2	2	Borough	2	Brooklyn
3	3	Borough	3	Manhattan
4	4	Borough	4	Queens
5	5	Borough	5	Staten Island
6	6	Borough	1198269	NA
7	7	Borough	1198479	NA
8	8	Borough	1198689	NA
9	9	Borough	1198899	NA
10	10	Borough	1199109	NA
11	11	Borough	1199319	NA
12	12	Borough	1199528	NA
13	13	Borough	1199738	NA
14	14	Borough	1199948	NA
15	15	CD	101	Financial District...
16	16	CD	102	Greenwich Villa...
17	17	CD	103	Lower East Side...
18	18	CD	104	Clinton and Chel...
19	19	CD	105	Midtown (CD5)
20	20	CD	106	Stuyvesant Tow...
21	21	CD	107	Upper West Sid...
22	22	CD	108	Upper East Side...
23	23	CD	109	Morningside Hei...
24	24	CD	110	Central Harlem (...
25	25	CD	111	East Harlem (C...

Report Location_Dimension

	Time_Key	Time_Period	Start_Date
1	1	Summer 2018	2018-01-...
2	2	Summer 2013	2013-01-...
3	3	Annual Average 20...	2009-01-...
4	4	Winter 2019-20	2019-01-...
5	5	2015	2015-01-...
6	6	2012-2014	2012-02-...
7	7	Winter 2021-22	2021-01-...
8	8	Summer 2011	2011-01-...
9	9	2010	2010-01-...
10	10	2013	2013-01-...
11	11	Winter 2010-11	2010-01-...
12	12	Winter 2008-09	2008-01-...
13	13	Annual Average 20...	2018-01-...
14	14	Winter 2013-14	2013-01-...
15	15	Winter 2009-10	2009-01-...
16	16	Annual Average 20...	2011-01-...
17	17	Winter 2016-17	2016-01-...
18	18	Summer 2015	2015-01-...
19	19	Winter 2015-16	2015-01-...
20	20	2017-2019	2017-01-...
21	21	Winter 2012-13	2012-01-...
22	22	Annual Average 20...	2013-01-...
23	23	Summer 2022	2022-01-...
24	24	Summer 2010	2010-01-...
25	25	2011	2011-01-...

Report Time_Dimension

Dashboard Integration

To improve access to, and interaction with, the air quality data warehouse, all Tableau visualizations have been combined into a single interactive dashboard. Out of this came an interactive dashboard for stakeholders to view, in a simple and centralized way, the data that had been analyzed.

Key Features

Filtering Capabilities:
They can filter by county, by season or year, and by kinds of pollutants, like PM2.5 or NO_2 or O_3.

Detail Functionality:
By location and time period, the dashboard allows for more granular datamining to spot pollution hot spots.

Export Options:
Users can export filtered views and visual summaries to support reporting and documentation requirements in PDF or image file formats.

Business Impact

- It's designed as a self-serve analytic tool for policymakers, environmental scientists and the public to mine the data without having to navigate the database.

- It helps demystify models of pollution and health impact for data-driven decision making.

- By allowing a spatial view of time series health correlation data on one screen, dashboards give stakeholders rich insight with little effort.

6: Database Comparison: Relational vs. Graph Databases

Overview

The project involved a comparative study of two different database technologies: relational databases (SQL Server) and graph databases (Neo4j). The goal was to evaluate the efficiency, performance, and flexibility of each approach for storing and querying complex data sets using seven sample queries applied to both systems.
In a relational database, information is stored in tables with fixed schemas and primary and foreign key constraints; graph databases, by contrast, organize data into nodes, which are entities, and relationships, which are the connections between entities; this more closely mirrors how it's tracked in the real world and is a nimble and intuitive way to approach data in relationship to other data.

Implementation Approach

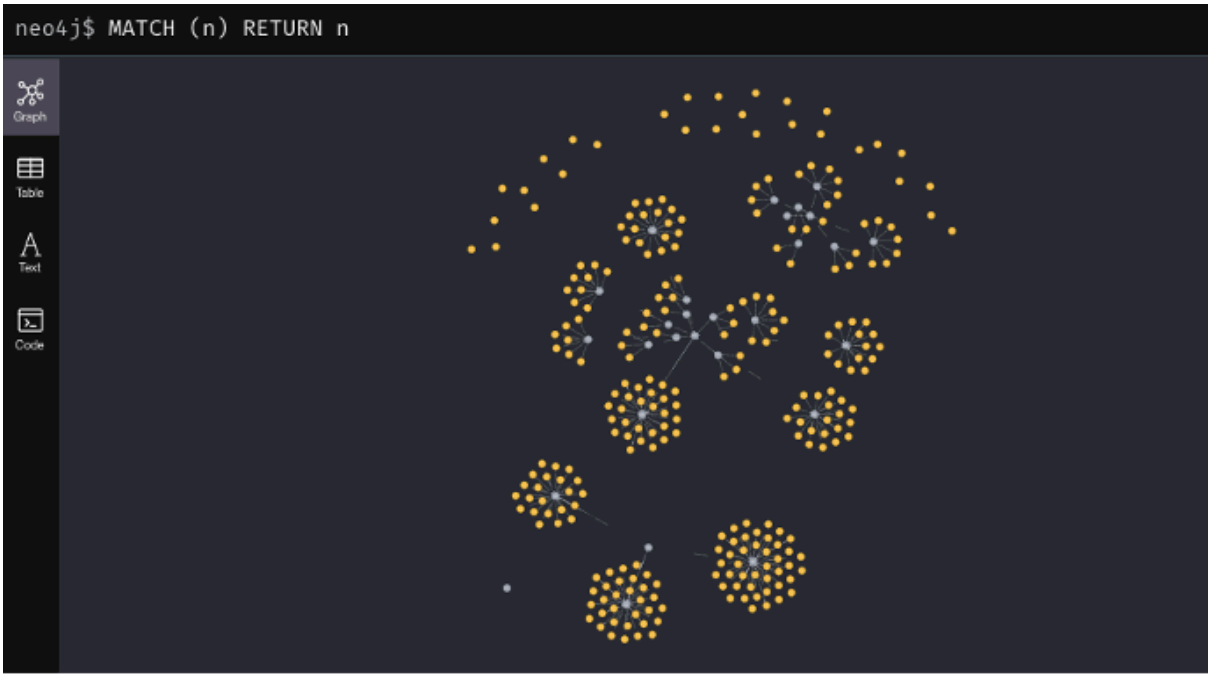
The relational database model linked Product to ProductCategory to ProductSubCategory to Vendor to ProductVendor by using SQL queries based on foreign key relationships. The same entities within Neo4j existed as nodes that formed connections through relationships including "BELONGS_TO_CATEGORY" and "SUPPLIES". The Neo4j import process used CSV file loading to create nodes while establishing relationships which followed the original relational design.

SQL vs CQL Queries and Results

The project required implementation of seven queries against both SQL Server using T-SQL and Neo4j using Cypher Query Language - CQL.

Neo4j Graph Relationships with its SQL equivalent using Adventure works dataset

Overall graph of dataset with nodes and relationships consisting of five tables viz: Product, ProductCategory, ProductSubcategory, Vendor, and ProductVendor, and sets of Relationship types and seven (7) sample relationships between two or more of them for both Neo4j and SQL.



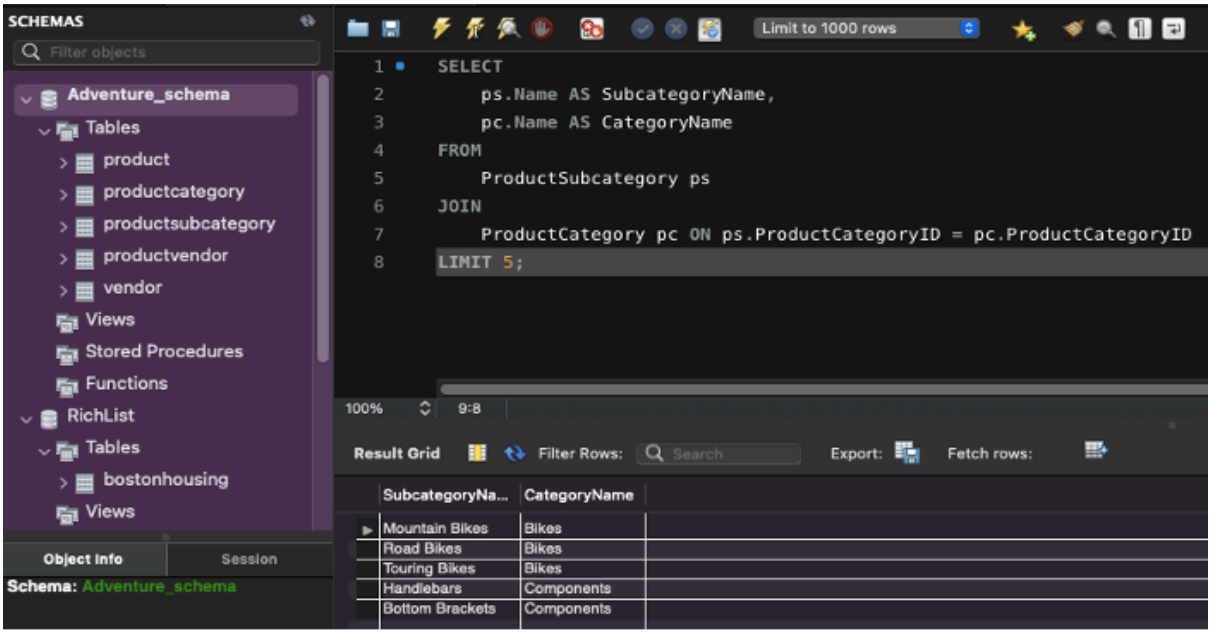
1. A Relationship between Subcategories and the (broader) Categories they belong to.

```
1 MATCH (sub:ProductSubcategory)-[:BELONGS_TO_CATEGORY]→(cat:ProductCategory)
2 RETURN sub.Name AS SubcategoryName, cat.Name AS CategoryName
3 LIMIT 5;
```

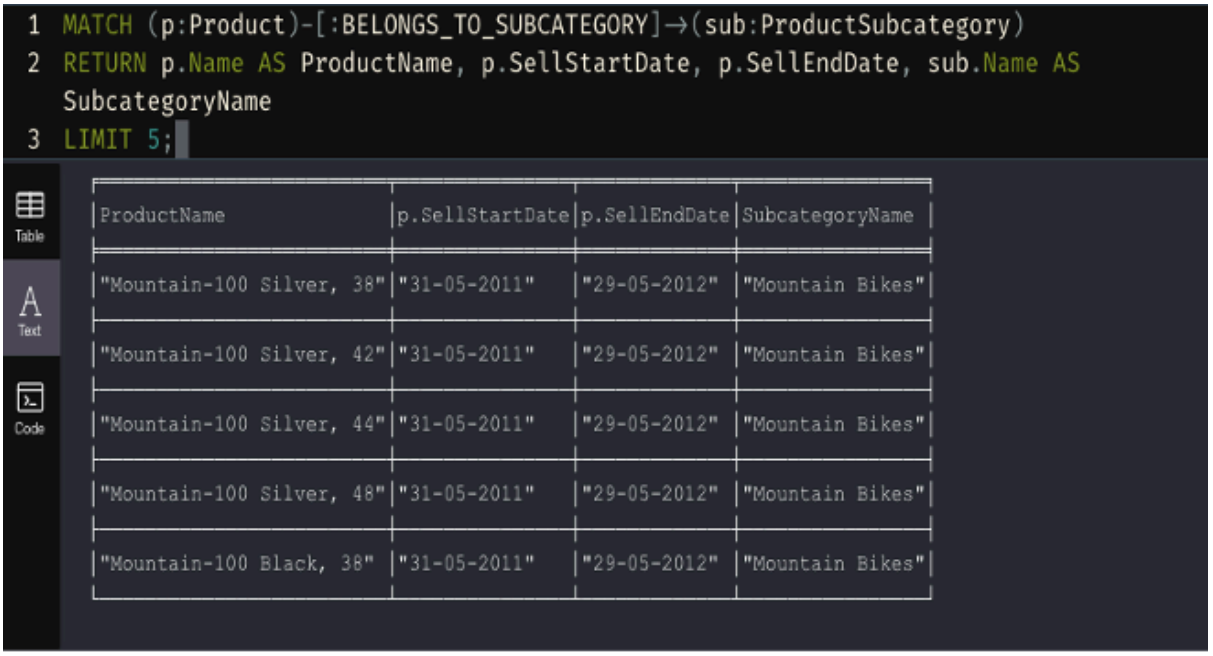
Table | Text | Code

	SubcategoryName	CategoryName
1	"Mountain Bikes"	"Bikes"
2	"Road Bikes"	"Bikes"
3	"Touring Bikes"	"Bikes"
4	"Handlebars"	"Components"
5	"Bottom Brackets"	"Components"

Started streaming 5 records after 19 ms and completed after 21 ms.



2. A relationship between Product, Sales starting and end dates, and the subcategories they belong to.



```
1  SELECT
2      p.Name AS ProductName,
3      p.SellStartDate,
4      p.SellEndDate,
5      ps.Name AS SubcategoryName
6  FROM
7      Product p
8  JOIN
9      ProductSubcategory ps ON p.ProductSubcategoryID = ps.ProductSubcategoryID
10 LIMIT 5;
```

100% 6:8

Result Grid | Filter Rows: Search | Export: | Fetch rows:

ProductName	SellStartDate	SellEndDate	SubcategoryNa...
HL Road Frame - Black, 58	30-04-2008		Road Frames
HL Road Frame - Red, 58	30-04-2008		Road Frames
Sport-100 Helmet, Red	31-05-2011		Helmets
Sport-100 Helmet, Black	31-05-2011		Helmets
Mountain Bike Socks, M	31-05-2011	29-05-2012	Socks

3. A relationship between some Products and the Subcategories they belong to.

```
1 MATCH (p:Product)-[:BELONGS_TO_SUBCATEGORY]→(sub:ProductSubcategory)
2 WHERE p.Name IS NOT NULL AND sub.Name IS NOT NULL // Filter out potential nulls if
  needed
3 RETURN p.Name AS Product, sub.Name AS Subcategory
4 LIMIT 5;
```

Table | Text | Code

Product	Subcategory
"Mountain-100 Silver, 38"	"Mountain Bikes"
"Mountain-100 Silver, 42"	"Mountain Bikes"
"Mountain-100 Silver, 44"	"Mountain Bikes"
"Mountain-100 Silver, 48"	"Mountain Bikes"
"Mountain-100 Black, 38"	"Mountain Bikes"

MAX COLUMN WIDTH:

```
1  SELECT
2      p.Name AS ProductName,
3      ps.Name AS SubcategoryName
4  FROM
5      Product p
6  JOIN
7      ProductSubcategory ps ON p.ProductSubcategoryID = ps.ProductSubcategoryID
8  LIMIT 5;
```

100% 9:8

Result Grid | Filter Rows: Search | Export: | Fetch rows:

ProductName	SubcategoryNa...
HL Road Frame - Black, 58	Road Frames
HL Road Frame - Red, 58	Road Frames
Sport-100 Helmet, Red	Helmets
Sport-100 Helmet, Black	Helmets
Mountain Bike Socks, M	Socks

4. Relationship with Vendor supplying Product with product Price

```
1  MATCH (v:Vendor)-[r:SUPPLIES]→(p:Product)
2  WHERE r.StandardPrice IS NOT NULL AND v.Name IS NOT NULL AND p.Name IS NOT NULL
3  RETURN v.Name AS Vendor, p.Name AS Product, r.StandardPrice AS Price
4  LIMIT 5;
```

Table | Text | Code

Vendor	Product	Price
"Australia Bike Retailer"	"Thin-Jam Lock Nut 9"	45.26
"Australia Bike Retailer"	"Thin-Jam Lock Nut 10"	43.26
"Australia Bike Retailer"	"Thin-Jam Lock Nut 1"	47.33
"Australia Bike Retailer"	"Thin-Jam Lock Nut 2"	43.26
"Australia Bike Retailer"	"Thin-Jam Lock Nut 15"	41.26

MAX COLUMN WIDTH:

```sql
SELECT
    v.Name AS Vendor,
    p.Name AS Product,
    pv.StandardPrice AS Price
FROM
    Vendor v
JOIN
    ProductVendor pv ON v.BusinessEntityID = pv.BusinessEntityID
JOIN
    Product p ON pv.ProductID = p.ProductID
WHERE
    pv.StandardPrice IS NOT NULL
-- WHERE clauses for v.Name and p.Name being NOT NULL are usually implicit
-- with an INNER JOIN, but can be added explicitly if needed.
LIMIT 5;
```

100% 9:15

Result Grid Filter Rows: Search Export: Fetch rows:

Vendor	Product	Price
Litware, Inc.	Adjustable Race	? 47.87
Wood Fitness	Bearing Ball	? 39.92
American Bicycles and Wheels	Headset Ball Bearings	? 54.31
Proseware, Inc.	LL Crankarm	? 25.77
Vision Cycles, Inc.	LL Crankarm	? 28.17

5. A Relationship between Vendors, the products they supply (by name) and the Lead Time

```cypher
MATCH (p:Product)<-[r:SUPPLIES]-(v:Vendor)
WHERE r.AverageLeadTime IS NOT NULL AND p.Name IS NOT NULL AND v.Name IS NOT NULL
RETURN p.Name AS Product, v.Name AS Vendor, r.AverageLeadTime AS LeadTime
LIMIT 5;
```

Table Text Code

Product	Vendor	LeadTime
"Thin-Jam Lock Nut 9"	"Australia Bike Retailer"	17
"Thin-Jam Lock Nut 10"	"Australia Bike Retailer"	17
"Thin-Jam Lock Nut 1"	"Australia Bike Retailer"	17
"Thin-Jam Lock Nut 2"	"Australia Bike Retailer"	17
"Thin-Jam Lock Nut 15"	"Australia Bike Retailer"	17

MAX COLUMN WIDTH:

```sql
SELECT
    p.Name AS Product,
    v.Name AS Vendor,
    pv.AverageLeadTime AS LeadTime
FROM
    Product p
JOIN
    ProductVendor pv ON p.ProductID = pv.ProductID
JOIN
    Vendor v ON pv.BusinessEntityID = v.BusinessEntityID
WHERE
    pv.AverageLeadTime IS NOT NULL
-- WHERE clauses for p.Name and v.Name being NOT NULL are usually implicit
-- with an INNER JOIN, but can be added explicitly if needed.
LIMIT 5;
```

100% 9:15

Result Grid | Filter Rows: Search | Export: | Fetch rows:

Product	Vendor	LeadTime
Adjustable Race	Litware, Inc.	17
Bearing Ball	Wood Fitness	19
Headset Ball Bearings	American Bicycles and Wheels	17
LL Crankarm	Proseware, Inc.	17
LL Crankarm	Vision Cycles, Inc.	19

6. Double relationship between Products belonging to a Product-Subcategory which belongs to a Product-Category

```cypher
MATCH (p:Product)-[:BELONGS_TO_SUBCATEGORY]→(:ProductSubcategory)-
[:BELONGS_TO_CATEGORY]→(cat:ProductCategory)
WHERE p.Name IS NOT NULL AND cat.Name IS NOT NULL
RETURN p.Name AS Product, cat.Name AS Category
LIMIT 5;
```

Table | Text | Code

	Product	Category
1	"Mountain-100 Silver, 38"	"Bikes"
2	"Mountain-100 Silver, 42"	"Bikes"
3	"Mountain-100 Silver, 44"	"Bikes"
4	"Mountain-100 Silver, 48"	"Bikes"
5	"Mountain-100 Black, 38"	"Bikes"

Started streaming 5 records after 19 ms and completed after 21 ms.

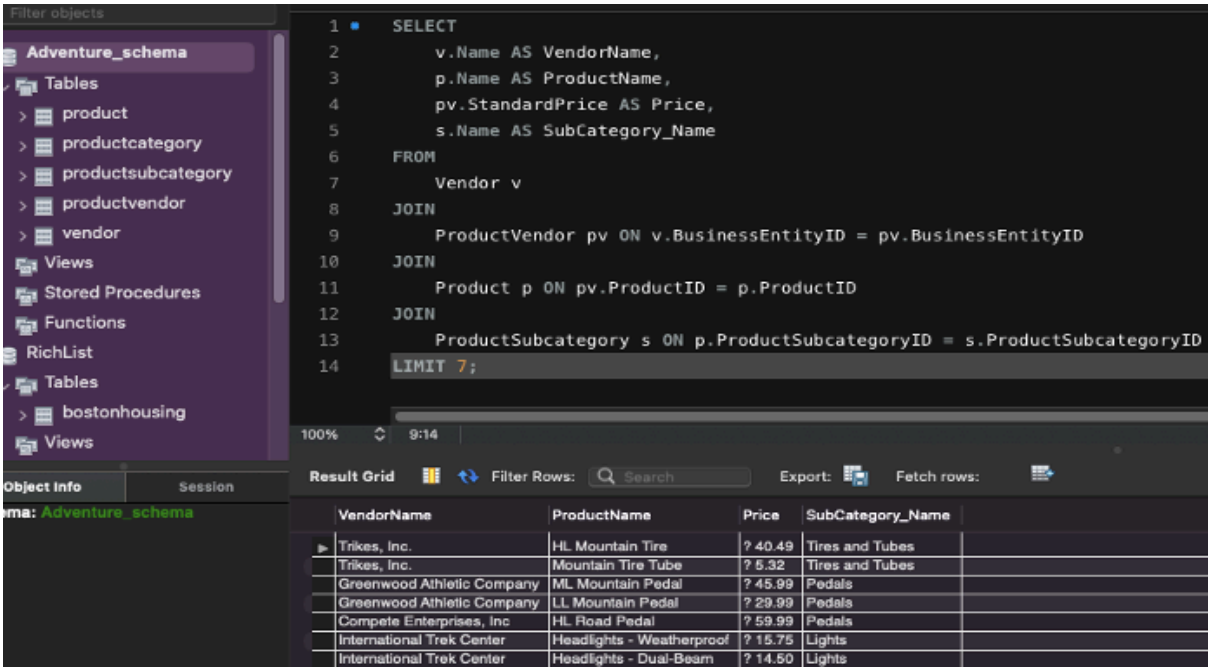
```sql
SELECT
    p.Name AS Product,
    pc.Name AS Category
FROM
    Product p
JOIN -- Join Product to its Subcategory
    ProductSubcategory ps ON p.ProductSubcategoryID = ps.ProductSubcategoryID
JOIN -- Join the Subcategory to its Category
    ProductCategory pc ON ps.ProductCategoryID = pc.ProductCategoryID
WHERE
    p.Name IS NOT NULL AND pc.Name IS NOT NULL -- Apply the filters
LIMIT 5;
```

100% 9:12

Result Grid | Filter Rows: Search | Export: | Fetch rows:

Product	Category
HL Road Frame - Black, 58	Components
HL Road Frame - Red, 58	Components
Sport-100 Helmet, Red	Accessories
Sport-100 Helmet, Black	Accessories
Mountain Bike Socks, M	Clothing

7. A Relationship between Vendors, name of Products they supply, the Prices and the Sub-categories those products belong to.



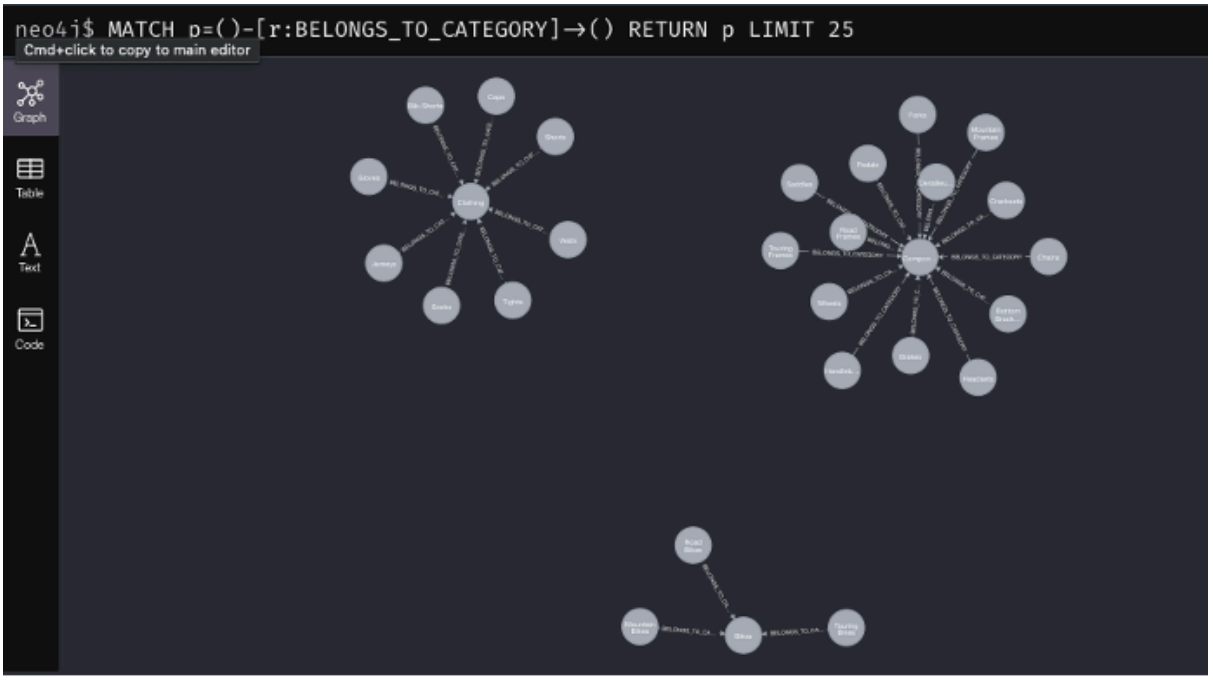
```
1 MATCH (v:Vendor)-[r:SUPPLIES]→(p:Product)-[:BELONGS_TO_SUBCATEGORY]→
  (s:ProductSubcategory)
2 RETURN v.Name AS VendorName, p.Name AS ProductName, r.StandardPrice AS Price,s.Name
  AS SubCategory_Name
3 LIMIT 7;
```

Table | Text | Code

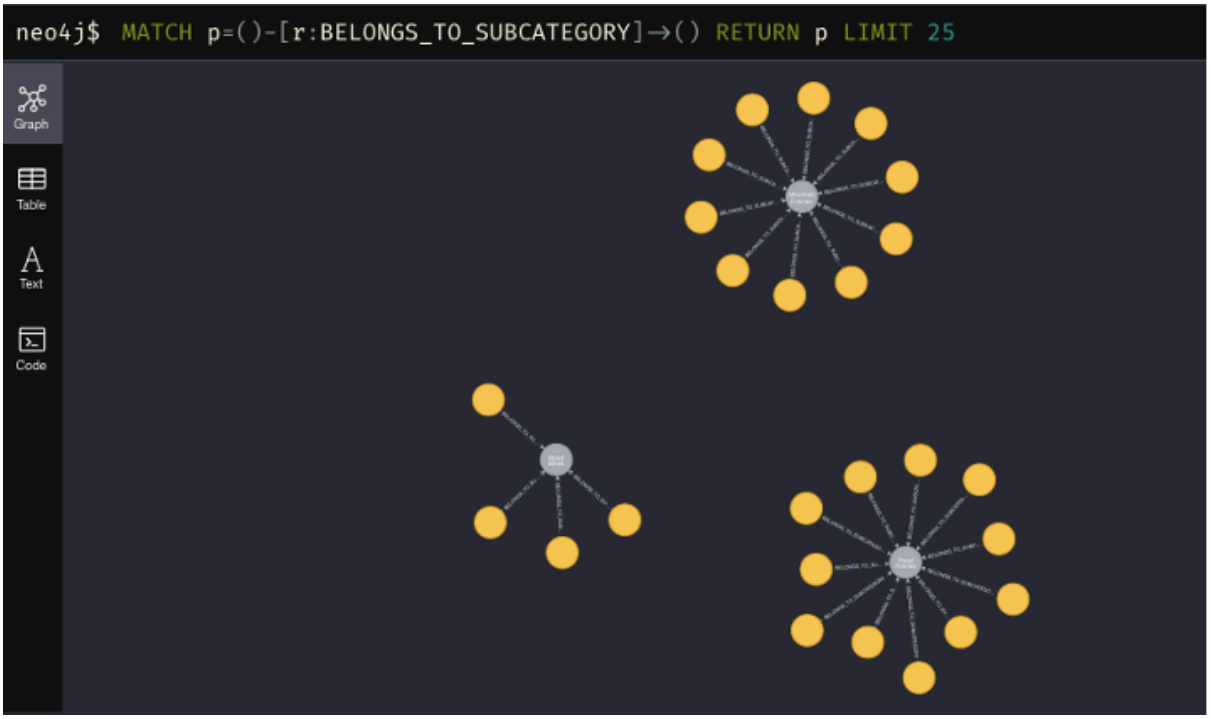
VendorName	ProductName	Price	SubCategory_Name
"Superior Bicycles"	"Rear Brakes"	78.89	"Brakes"
"Superior Bicycles"	"Front Brakes"	78.89	"Brakes"
"Varsity Sport Co."	"Chain"	14.99	"Chains"
"Greenwood Athletic Company"	"LL Mountain Pedal"	29.99	"Pedals"
"Crowley Sport"	"LL Mountain Pedal"	29.99	"Pedals"
"Greenwood Athletic Company"	"ML Mountain Pedal"	45.99	"Pedals"
"Crowley Sport"	"ML Mountain Pedal"	45.99	"Pedals"

MAX COLUMN WIDTH:

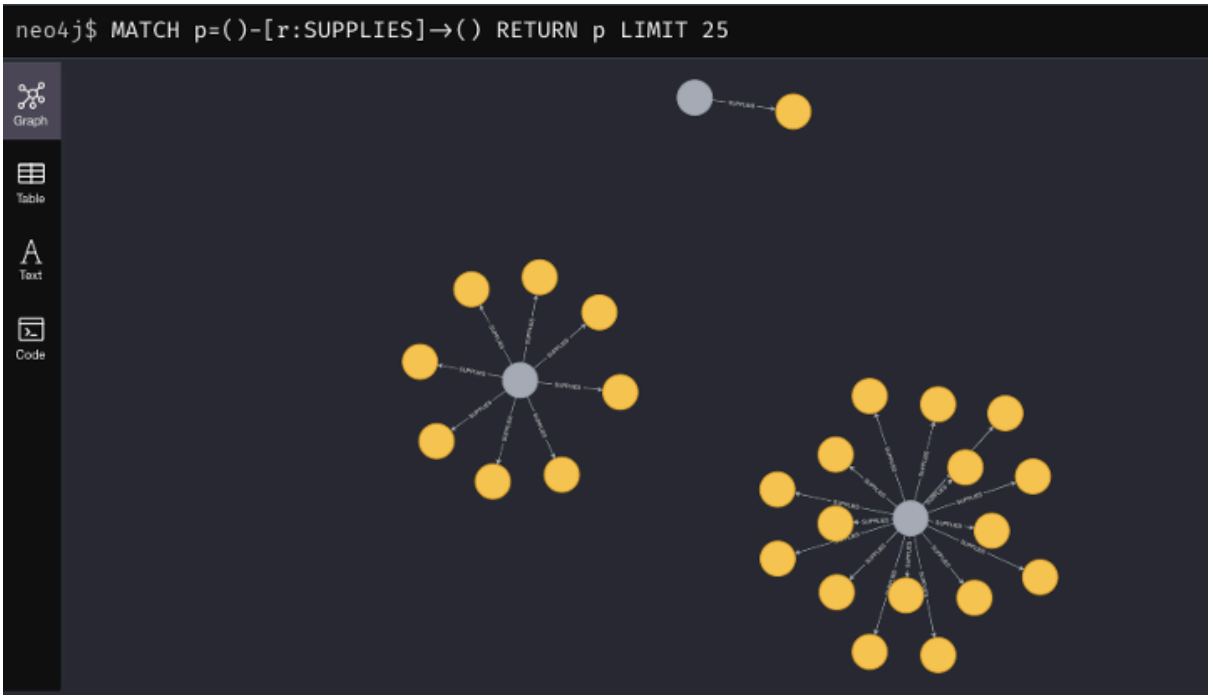
Relationship type 1 [BELONGS_TO_CATEGORY]:



Relationship type 2 [BELONGS_TO_SUBCATEGORY]:



Relationship type 3 [SUPPLIES]:



7. Overall Discussion and Reflection

Project Achievements

Utilizing an effectively structured dimensional data warehouse the NYC Air Quality Analysis project achieved transformation of complex fragmented datasets into useful business information.Different source data enabled us to develop a database based on the star schema design format which facilitated comprehensive analysis of pollution patterns and health results and geographical patterns in New York City.

The ETL process standardized data through cleaning methods which also repaired faulty input and validated complete records within both fact and dimension tables. As part of our work we designed the ETL workflows while performing data quality inspections. With the help of Tableau and SSRS complicated designs became simple to understand, entitle stakeholders such as policymakers and public health employees with interactive insights.

Additionally, the study which is comparative between relational(SQL Server) and Neo4j graph databases shows the adaptability of different storage models based on query complexity and relationship depth.

Challenges Encountered

Several challenges were encountered throughout the project:

- Data Inconsistencies and Missing Values: The raw data had missing fields and inconsistent formatting. Robust transformation rules and validation scripts had to be developed to address these issues.
- Schema Normalization and Key Mapping: Designing efficient foreign key relationships without losing analytical flexibility was complex and required multiple iterations.
- Managing High Data Volume: The data warehouse had to manage large datasets spanning multiple contaminants and locations without performance degradation during querying and visualization.

Skills and Knowledge Gained

The project strengthened my capabilities to perform technical and analytical work through several key elements.

- The design of star schemas together with fact/dimension table connectivity stands as the fundamental focus of dimensional modeling as well as analytics design concepts.

- Advancement in SQL Scripting and Data Validation includes writing robust scripts for extraction and transformation and testingSingle-Handed Project Execution.Individual Contribution….

ETL Automation with SSIS: Building robust, fault-tolerant data pipelines.
Comparative Database Analysis: Understanding the practical differences between relational and graph databases for different types of business queries.

References

chugugrace (2024) *SSIS How to Create an ETL Package - SQL Server Integration Services (SSIS)*. Microsoft.com. Available at: https://learn.microsoft.com/en-us/sql/integration-services/ssis-how-to-create-an-etl-package?view=sql-server-ver16 (Accessed: 10 April 2025).

erinstellato-ms (2024) *Download SQL Server Management Studio (SSMS)*. Microsoft.com. Available at: https://learn.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver16(Accessed: 10 April 2025).

GeeksforGeeks (2018) *Star Schema in Data Warehouse Modeling*. Available at: https://www.geeksforgeeks.org/star-schema-in-data-warehouse-modeling/ (Accessed: 10 April 2025).

Murphy, S.A. (2013) *Data Visualization and Rapid Analytics: Applying Tableau Desktop to Support Library Decision-Making*. Journal of Web Librarianship, 7(4), pp.465–476. Available at: https://doi.org/10.1080/19322909.2013.825148 (Accessed: 10 April 2025).

Neo4j Documentation (2024) *Neo4j Graph Data Platform*. Available at: https://neo4j.com/docs/ (Accessed: 10 April 2025).

Tableau (2023) *Tableau: Business Intelligence and Analytics Software*. Available at: https://www.tableau.com/(Accessed: 10 April 2025).

NYC Open Data (2025) *NYC Environmental & Health Data*. Available at: https://opendata.cityofnewyork.us/(Accessed: 10 April 2025).

Environmental Protection Agency (EPA) (2025) *Air Quality Data*. Available at: https://www.epa.gov/outdoor-air-quality-data (Accessed: 10 April 2025).