BÁO CÁO ĐỒ ÁN THỰC HÀNH

HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH

GVHD:ThS. Hồ Thị Hoàng Vy

ThS. Tiết Gia Hồng

Nhóm: 05

THÔNG TIN NHÓM

STT	MSSV	Họ tên	Công việc	% Hoàn thành
1	21127246	Lê Minh Đức	Cài đặt ETL	100%
2	21127609	Nguyễn Đức Vĩnh Hòa	QC	100%
3	21127640	Huỳnh Hữu Lộc	Thiết kế cơ sở dữ liệu DDS	100%
4	21127641	Nguyễn Xuân Lộc	Viết báo cáo	100%

THIẾT KẾ CƠ SỞ DỮ LIỆU

1. Các chiều dữ liệu

1.1. DimDate

- **DateID (PK, SK):** ID cho mỗi ngày tháng.
- **Year:** Năm trong ngày tháng (chẳng hạn 2024).
- **Quarter:** Quý trong năm (1 tới 4).
- **Month:** Tháng trong ngày tháng (1 tới 12).
- **Day:** Ngày trong ngày tháng (1 tới 31).
- **FullDate (NK):** Ngày tháng đầy đủ (định dạng YYYY-MM-DD).
- **IsDayLightSaving:** Là "true" nếu ngày từ 12/3/2023 đến 5/11/2023. Là "false" khi khác.
- **createdDay:** Dấu thời gian ngày tạo dữ liệu.
- **updatedDay:** Dấu thời gian ngày cập nhật dữ liệu gần nhất.

1.2. DimGeography

- **GeographyID (PK, SK):** ID cho mỗi vị trí địa lý.
- **StateID (FK):** ID của bang.
- **CountyID (FK):** ID của quận.
- **createdDay:** Dấu thời gian ngày tạo dữ liệu.
- **updatedDay:** Dấu thời gian ngày cập nhật dữ liệu gần nhất.

1.3. DimState

- **StateID (PK, SK):** ID cho mỗi bang.
- **StateName:** Tên bang.
- **StateCode** (NK): Mã FIPS của bang (định dạng SS).
- **createdDay:** Dấu thời gian ngày tạo dữ liệu.
- **updatedDay:** Dấu thời gian ngày cập nhật dữ liệu gần nhất.

1.4. DimCounty

- **CountyID (PK, SK):** ID cho mỗi quận.
- **CountyName:** Tên quận.
- **CountyCode (NK):** Mã FIPS cho quận (định dạng CCC).
- **createdDay:** Dấu thời gian ngày tạo dữ liệu.
- **updatedDay:** Dấu thời gian ngày cập nhật dữ liệu gần nhất.

1.5. DimCategory

- **CategoryID (PK, SK):** ID cho mỗi phân loại.
- **CategoryName (NK):** Tên phân loại.
- **createdDay:** Dấu thời gian ngày tạo dữ liệu.
- **updatedDay:** Dấu thời gian ngày cập nhật dữ liệu gần nhất.

1.6. DimParameter

- **ParameterID (PK, SK):** ID cho mỗi yếu tố ảnh hưởng nhất tới chỉ số không khí.
- **ParameterName (NK):** Tên của yếu tố ảnh hưởng (chẳng hạn ozone, PM2.5).
- **createdDay:** Dấu thời gian ngày tạo dữ liệu.
- **updatedDay:** Dấu thời gian ngày cập nhật dữ liệu gần nhất.

1.7. AQIData

- **AQIDataID (PK, SK):** ID cho mỗi bản ghi AQI.
- **GeographyID(PK, SK):** Tham chiếu tới DimGeography(GeographyID).
- **DateID (FK):** Tham chiếu Date(DateID).
- **AQI:** Giá trị chất lượng không khí.
- **CategoryID (FK):** Tham chiếu Categories(CategoryID).
- **ParameterID (FK):** Tham chiếu DefiningParameters(ParameterID).
- **NumberOfSites:** Tổng số báo cáo từ các trạm.
- **createdDay:** Dấu thời gian ngày tạo dữ liệu.
- **updatedDay:** Dấu thời gian ngày cập nhật dữ liệu gần nhất.

2. Mối quan hệ giữa các bảng

2.1. DimDate ↔ AQIData (1:n)

- **Mối quan hệ:** Một ngày trong bảng DimDate có thể liên quan đến nhiều bản ghi chất lượng không khí trong bảng AQIData.
- **Chi tiết:** DateID trong DimDate là khoá chính (PK) và được dùng làm khoá ngoại (FK) trong AQIData.

2.2. DimGeography ↔ AQIData (1:n)

- **Mối quan hệ:** Một vị trí địa lý trong DimGeography có thể liên quan đến nhiều bản ghi trong AQIData.
- **Chi tiết:** GeographyID trong DimGeography làm khoá chính (PK), tham chiếu từ CountyID và StateID trong bảng AQIData.

2.3. DimState ↔ DimGeography (1:n)

- **Mối quan hệ:** Một bang trong DimState liên kết với nhiều vị trí địa lý trong DimGeography.

- **Chi tiết:** StateID trong DimState là khoá chính (PK), và là một phần của GeographyID trong DimGeography.

2.4. DimCounty ↔ DimGeography (1:n)

- **Mối quan hệ:** Một quận trong DimCounty liên kết với nhiều vị trí địa lý trong DimGeography.
- **Chi tiết:** CountyID trong DimCounty là khoá chính (PK), và là một phần của GeographyID trong DimGeography.

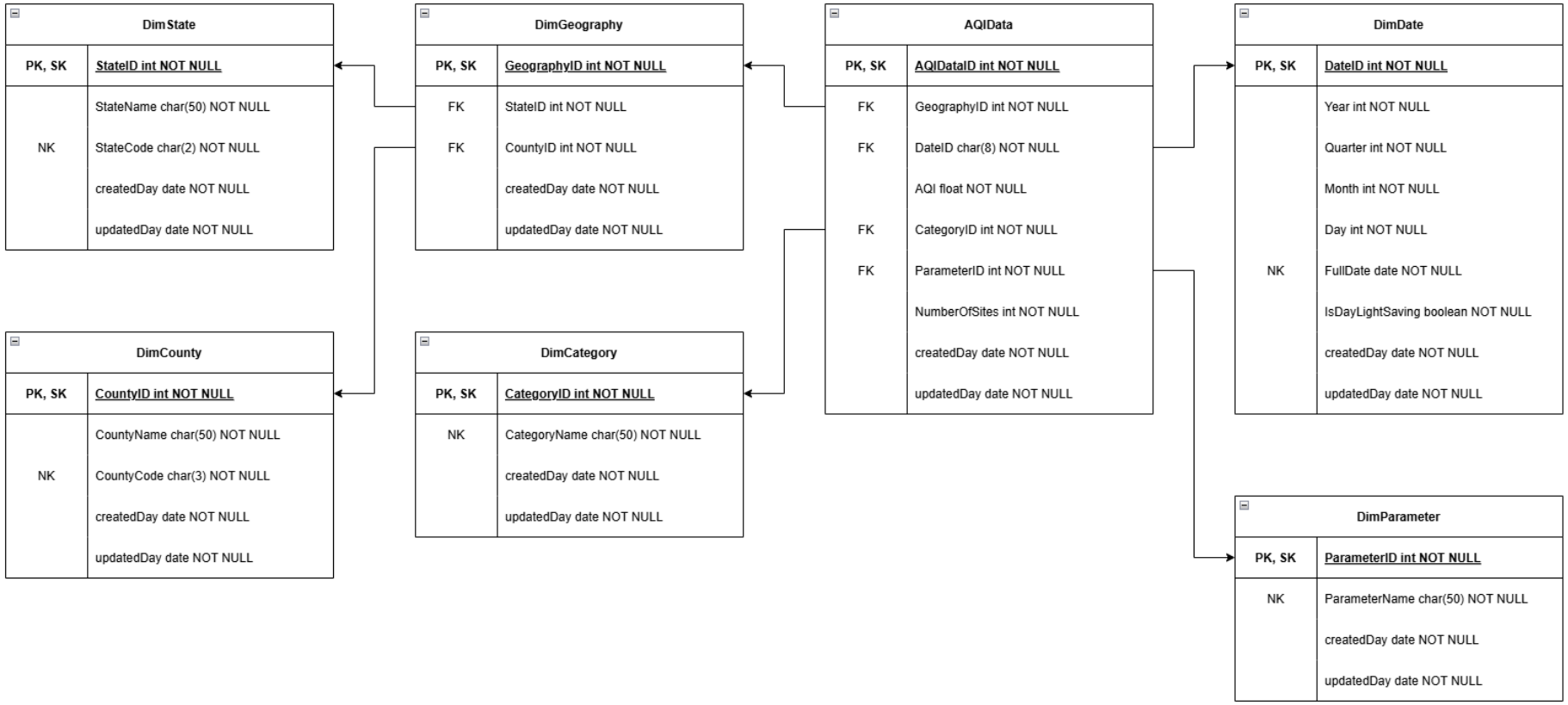
2.5. DimCategory ↔ AQIData (1:1)

- **Mối quan hệ:** Mỗi bản ghi AQI trong AQIData liên kết với một loại phân loại trong DimCategory.
- **Chi tiết:** CategoryID trong DimCategory là khoá chính (PK) và được sử dụng làm khoá ngoại (FK) trong AQIData.

2.6. DimParameter ↔ AQIData (1:1)

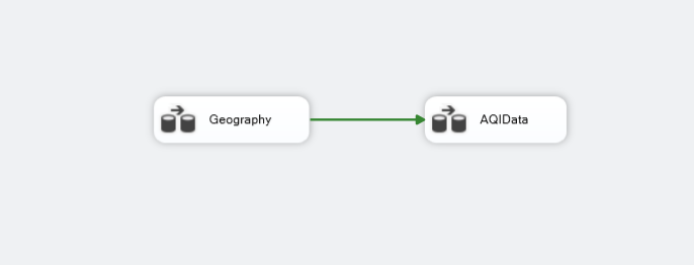
- **Mối quan hệ:** Mỗi bản ghi AQI trong AQIData liên quan đến một yếu tố ảnh hưởng (pollutant) trong DimParameter.
- **Chi tiết:** ParameterID trong DimParameter là khoá chính (PK) và được dùng làm khoá ngoại (FK) trong AQIData.

3. Mô hình ERD


DimState
PK, SK
StateID int NOT NULL
StateName char(50) NOT NULL
NK
StateCode char(2) NOT NULL
createdDay date NOT NULL
updatedDay date NOT NULL
DimGeography
PK, SK
GeographyID int NOT NULL
FK
StateID int NOT NULL
FK
CountyID int NOT NULL
createdDay date NOT NULL
updatedDay date NOT NULL
AQIData
PK, SK
AQIDataID int NOT NULL
FK
GeographyID int NOT NULL
FK
DateID char(8) NOT NULL
AQI float NOT NULL
FK
CategoryID int NOT NULL
FK
ParameterID int NOT NULL
NumberOfSites int NOT NULL
createdDay date NOT NULL
updatedDay date NOT NULL
DimDate
PK, SK
DateID int NOT NULL
Year int NOT NULL
Quarter int NOT NULL
Month int NOT NULL
Day int NOT NULL
NK
FullDate date NOT NULL
IsDayLightSaving boolean NOT NULL
createdDay date NOT NULL
updatedDay date NOT NULL
DimCounty
PK, SK
CountyID int NOT NULL
CountyName char(50) NOT NULL
NK
CountyCode char(3) NOT NULL
createdDay date NOT NULL
updatedDay date NOT NULL
DimCategory
PK, SK
CategoryID int NOT NULL
NK
CategoryName char(50) NOT NULL
createdDay date NOT NULL
updatedDay date NOT NULL
DimParameter
PK, SK
ParameterID int NOT NULL
NK
ParameterName char(50) NOT NULL
createdDay date NOT NULL
updatedDay date NOT NULL

XÂY DỰNG ETL PROCESS

4. Cấu trúc dự án: 2425.BI.DATH#2



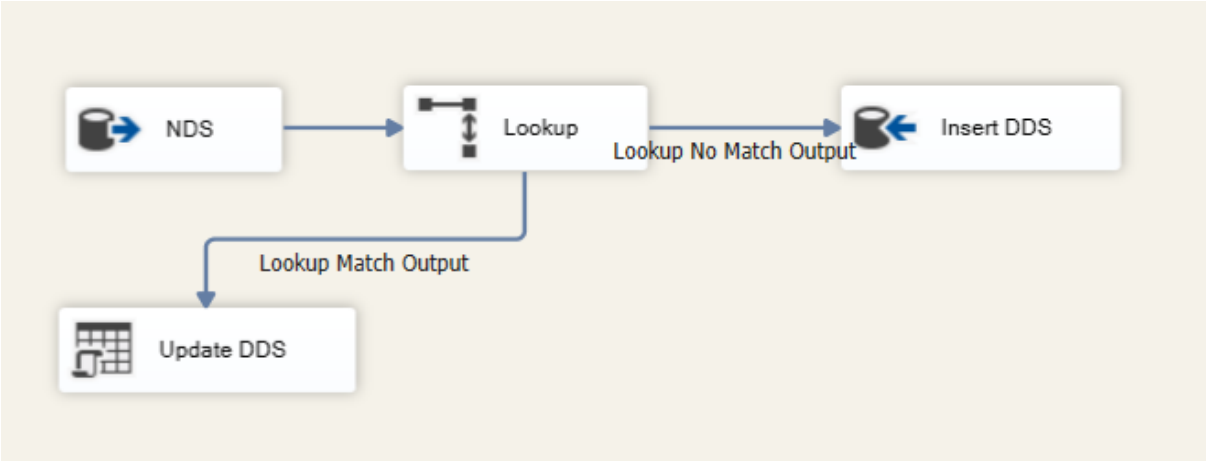
Package.dtsx *Package1.dtsx*

- **Package.dtsx:** Lấy dữ liệu từ NDS sang các bảng chiều tại DDS không có tham chiếu.
- **Package1.dtsx:** Lấy dữ liệu từ NDS sang các bảng chứa tham chiếu tại DDS. Fact sẽ được đổ sau khi Dim cuối cùng được đổ dữ liệu thành công.

5. Cấu trúc các dataflow

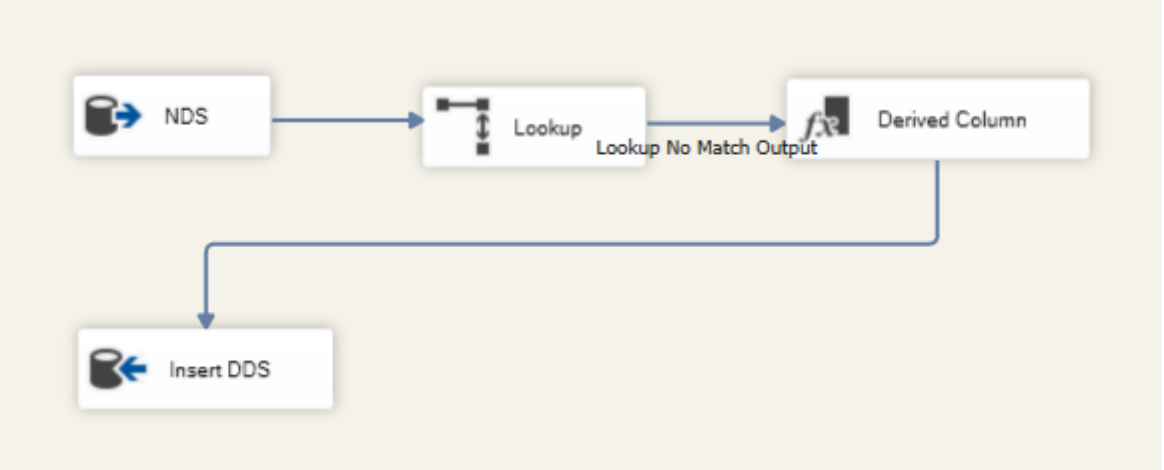
5.1. Package

County, State, Parameter, Category



- *Bước 1:* Truy vấn lấy dữ liệu tại NDS.
- *Bước 2*: Lookup. Phân nhánh tập bản ghi tìm thấy tại DDS: f và tập bản ghi chưa ghi nhận tại DDS: u.
- *Bước 3:* Map và ghi dữ liệu vào DDS với tập f. Cập nhật dữ liệu tại DDS với map của tập u.

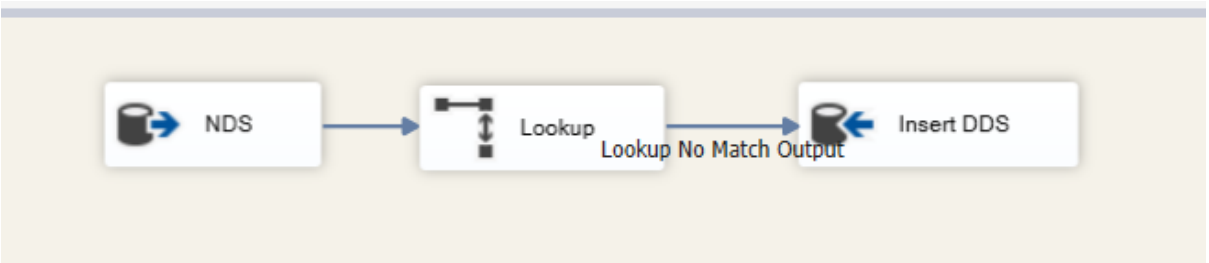
Date



- *Bước 1:* Truy vấn lấy dữ liệu từ NDS.
- *Bước 2:* Lookup những bản ghi về Date chưa tồn tại trong DDS.
- *Bước 3:* Thực hiện suy diễn, gán IsDayLightSaving cho các bản ghi mới. IsDayLightSaving = True chỉ khi thuộc đoạn [12-03-2023, 05-11-2023].
- *Bước 4:* Map tập các bản ghi sau khi suy diễn, ghi vào DDS.

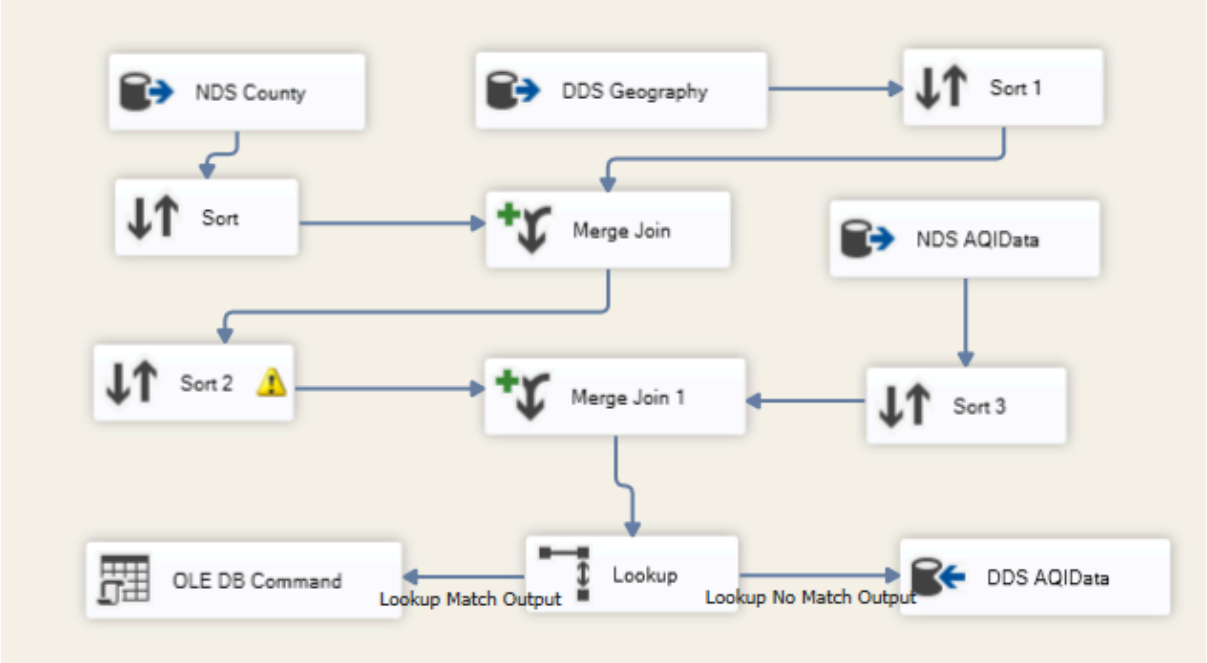
5.2. Package1:

Geography:



- *Bước 1:* Truy vấn lấy dữ liệu từ NDS, County.
- *Bước 2:* Lookup lấy các cặp CountySK, StateSK chưa tồn tại trong DDS.
- *Bước 3:* Map và ghi dữ liệu vào DDS.

AQIData:



- *Bước 1:* Sắp xếp và join 2 bản NDS County và DDS Geography để lấy về GeographySK thay vì dùng CountySK.
- *Bước 2:* Sắp xếp và join tiếp tục tập kết quả với NDS AQIData để gắn GeographySK vào tập các bản ghi thu được từ NDS.
- *Bước 3:* Lookup. Phân nhánh tập bản ghi tìm thấy tại DDS: f và tập bản ghi chưa ghi nhận tại DDS: u.
- *Bước 4:* Map và ghi dữ liệu vào DDS với tập f. Cập nhật dữ liệu tại DDS với map của tập u.

6. Kết quả

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [AQIDataSK]
      ,[GeographySK]
      ,[DateSK]
      ,[AQI]
      ,[CategorySK]
      ,[ParameterSK]
      ,[NumberOfSites]
      ,[CreatedDate]
      ,[UpdatedDate]
  FROM [2425.BI.DATH_DDS].[FACT].[AQIDATA]
```

100 %

Results | Messages

	AQIDataSK	GeographySK	DateSK	AQI	CategorySK	ParameterSK	NumberOfSites	CreatedDate	UpdatedDate
1	1	197	272	68	3	4	9	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
2	2	197	273	84	3	3	9	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
3	3	197	274	158	4	5	10	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
4	4	197	275	173	4	5	9	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
5	5	197	276	223	6	5	9	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
6	6	197	277	176	4	5	10	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
7	7	197	278	164	4	5	9	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
8	8	197	279	77	3	5	10	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
9	9	197	280	155	4	5	11	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
10	10	197	281	38	1	5	10	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
11	11	197	282	40	1	3	10	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
12	12	197	283	64	3	5	11	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
13	13	197	284	163	4	4	10	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
14	14	197	285	38	1	4	10	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
15	15	197	286	58	3	5	11	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973
16	16	197	287	59	3	5	10	2024-12-07 10:11:35.973	2024-12-07 10:11:35.973

Query executed successfully. | PC (16.0 RTM) | PC\phamg (75) | 2425.BI.DATH_DDS | 00:00:01 | 187,933 rows

FACT.AQIDATA

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [StateSK]
      ,[StateID]
      ,[StateName]
      ,[CreatedDate]
      ,[UpdatedDate]
  FROM [2425.BI.DATH_DDS].[DIM].[STATE]
```

100 %

Results | Messages

	StateSK	StateID	StateName	CreatedDate	UpdatedDate
1	1	AK	Alaska	2024-11-23 16:59:16.400	2024-11-23 20:14:34.790
2	2	AL	Alabama	2024-11-23 16:59:16.400	2024-11-23 20:14:34.793
3	3	AR	Arkansas	2024-11-23 16:59:16.400	2024-11-23 20:14:34.793
4	4	AZ	Arizona	2024-11-23 16:59:16.400	2024-11-23 20:14:34.793
5	5	CA	California	2024-11-23 16:59:16.400	2024-11-23 20:14:34.793
6	6	CO	Colorado	2024-11-23 16:59:16.400	2024-11-23 20:14:34.797
7	7	CT	Connecticut	2024-11-23 16:59:16.400	2024-11-23 20:14:34.797
8	8	DC	District of Columbia	2024-11-23 16:59:16.400	2024-11-23 20:14:34.797
9	9	DE	Delaware	2024-11-23 16:59:16.400	2024-11-23 20:14:34.797
10	10	FL	Florida	2024-11-23 16:59:16.400	2024-11-23 20:14:34.797
11	11	GA	Georgia	2024-11-23 16:59:16.400	2024-11-23 20:14:34.800
12	12	HI	Hawaii	2024-11-23 16:59:16.400	2024-11-23 20:14:34.800
13	13	IA	Iowa	2024-11-23 16:59:16.400	2024-11-23 20:14:34.800
14	14	ID	Idaho	2024-11-23 16:59:16.400	2024-11-23 20:14:34.800
15	15	IL	Illinois	2024-11-23 16:59:16.400	2024-11-23 20:14:34.800
16	16	IN	Indiana	2024-11-23 16:59:16.400	2024-11-23 20:14:34.800

Query executed successfully. PC (16.0 RTM) | PC\phamg (74) | 2425.BI.DATH_DDS | 00:00:00 | 51 rows

DIM.STATE

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [ParameterSK]
      ,[ParameterName]
      ,[CreatedDate]
      ,[UpdatedDate]
  FROM [2425.BI.DATH_DDS].[DIM].[PARAMETER]
```

100 %

Results | Messages

	ParameterSK	ParameterName	CreatedDate	UpdatedDate
1	1	CO	2024-11-23 17:31:51.613	2024-11-23 17:31:51.613
2	2	NO2	2024-11-23 17:31:51.613	2024-11-23 17:31:51.613
3	3	Ozone	2024-11-23 17:31:51.613	2024-11-23 17:31:51.613
4	4	PM10	2024-11-23 17:31:51.613	2024-11-23 17:31:51.613
5	5	PM2.5	2024-11-23 17:31:51.613	2024-11-23 17:31:51.613

Query executed successfully. PC (16.0 RTM) | PC\phamg (69) | 2425.BI.DATH_DDS | 00:00:00 | 5 rows

DIM.PARAMETER

```sql
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [GeographySK]
      ,[CountySK]
      ,[StateSK]
      ,[CreatedDate]
      ,[UpdatedDate]
  FROM [2425.BI.DATH_DDS].[DIM].[GEOGRAPHY]
```

100 %

Results | Messages

	GeographySK	CountySK	StateSK	CreatedDate	UpdatedDate
1	1	1	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
2	2	2	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
3	3	3	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
4	4	4	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
5	5	5	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
6	6	6	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
7	7	7	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
8	8	8	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
9	9	9	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
10	10	10	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
11	11	11	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
12	12	12	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
13	13	13	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
14	14	14	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
15	15	15	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443
16	16	16	2	2024-12-07 10:11:34.443	2024-12-07 10:11:34.443

Query executed successfully. PC (16.0 RTM) | PC\phamg (72) | 2425.BI.DATH_DDS | 00:00:00 | 3,144 rows

DIM.GEOGRAPHY

```sql
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [CountySK]
      ,[CountyFips]
      ,[CountyName]
      ,[CreatedDate]
      ,[UpdatedDate]
  FROM [2425.BI.DATH_DDS].[DIM].[COUNTY]
```

100 %

Results | Messages

	CountySK	CountyFips	CountyName	CreatedDate	UpdatedDate
1	1	01001	Autauga	2024-11-23 18:00:51.233	2024-11-23 20:16:49.420
2	2	01003	Baldwin	2024-11-23 18:00:51.233	2024-11-23 20:16:49.427
3	3	01005	Barbour	2024-11-23 18:00:51.233	2024-11-23 20:16:49.427
4	4	01007	Bibb	2024-11-23 18:00:51.233	2024-11-23 20:16:49.430
5	5	01009	Blount	2024-11-23 18:00:51.233	2024-11-23 20:16:49.433
6	6	01011	Bullock	2024-11-23 18:00:51.233	2024-11-23 20:16:49.437
7	7	01013	Butler	2024-11-23 18:00:51.233	2024-11-23 20:16:49.440
8	8	01015	Calhoun	2024-11-23 18:00:51.233	2024-11-23 20:16:49.440
9	9	01017	Chambers	2024-11-23 18:00:51.233	2024-11-23 20:16:49.443
10	10	01019	Cherokee	2024-11-23 18:00:51.233	2024-11-23 20:16:49.443
11	11	01021	Chilton	2024-11-23 18:00:51.233	2024-11-23 20:16:49.447
12	12	01023	Choctaw	2024-11-23 18:00:51.233	2024-11-23 20:16:49.450
13	13	01025	Clarke	2024-11-23 18:00:51.233	2024-11-23 20:16:49.450
14	14	01027	Clay	2024-11-23 18:00:51.233	2024-11-23 20:16:49.453
15	15	01029	Cleburne	2024-11-23 18:00:51.233	2024-11-23 20:16:49.457
16	16	01031	Coffee	2024-11-23 18:00:51.233	2024-11-23 20:16:49.460

Query executed successfully. PC (16.0 RTM) | PC\phamg (68) | 2425.BI.DATH_DDS | 00:00:00 | 3,144 rows

DIM.COUNTY

```sql
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [DateSK]
      ,[Year]
      ,[Quarter]
      ,[Month]
      ,[Day]
      ,[FullDate]
      ,[IsDayLightSaving]
      ,[CreatedDate]
      ,[UpdatedDate]
  FROM [2425.BI.DATH_DDS].[DIM].[DATE]
```

100 %

Results Messages

	DateSK	Year	Quarter	Month	Day	FullDate	IsDayLightSaving	CreatedDate	UpdatedDate
1	1	2021	1	1	1	2021-01-01	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
2	2	2021	1	1	2	2021-01-02	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
3	3	2021	1	1	3	2021-01-03	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
4	4	2021	1	1	4	2021-01-04	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
5	5	2021	1	1	5	2021-01-05	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
6	6	2021	1	1	6	2021-01-06	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
7	7	2021	1	1	7	2021-01-07	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
8	8	2021	1	1	8	2021-01-08	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
9	9	2021	1	1	9	2021-01-09	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
10	10	2021	1	1	10	2021-01-10	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
11	11	2021	1	1	11	2021-01-11	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
12	12	2021	1	1	12	2021-01-12	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
13	13	2021	1	1	13	2021-01-13	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
14	14	2021	1	1	14	2021-01-14	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
15	15	2021	1	1	15	2021-01-15	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303
16	16	2021	1	1	16	2021-01-16	0	2024-11-23 17:48:36.303	2024-11-23 17:48:36.303

Query executed successfully. PC (16.0 RTM) | PC\phamg (75) | 2425.BI.DATH_DDS | 00:00:00 | 1,095 rows

DIM.DATE

```sql
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [CategorySK]
      ,[CategoryName]
      ,[CreatedDate]
      ,[UpdatedDate]
  FROM [2425.BI.DATH_DDS].[DIM].[CATEGORY]
```

100 %

Results Messages

	CategorySK	CategoryName	CreatedDate	UpdatedDate
1	1	Good	2024-11-23 17:26:12.167	2024-11-23 17:26:12.167
2	2	Hazardous	2024-11-23 17:26:12.167	2024-11-23 17:26:12.167
3	3	Moderate	2024-11-23 17:26:12.167	2024-11-23 17:26:12.167
4	4	Unhealthy	2024-11-23 17:26:12.167	2024-11-23 17:26:12.167
5	5	Unhealthy for Sensitive Groups	2024-11-23 17:26:12.167	2024-11-23 17:26:12.167
6	6	Very Unhealthy	2024-11-23 17:26:12.167	2024-11-23 17:26:12.167

Query executed successfully. PC (16.0 RTM) | PC\phamg (67) | 2425.BI.DATH_DDS | 00:00:00 | 6 rows

DIM.CATEGORY