`

KULLIYYAH OF INFORMATION & COMMUNICATION TECHNOLOGY

INFO 4311 DATA WAREHOUSING

SEMESTER 1, 2024/2025

SECTION 01, GROUP A

GROUP PROJECT:

DEVELOPMENT OF A STAR SCHEMA DATA WAREHOUSE

Aviation Accident (2021 - 2023)

PREPARED BY:

NO.	NAME	MATRIC NO
1	MUHAMMAD IKMAL HAKIMI BIN ROSLI	2210827
2	AHMAD IMAN BIN TURIN	2123349
3	MUHAMMAD MUSLIHUDDIN BIN MUSTAFFAR	2213263
4	ARMAN NURI ANUAR	2119279
5	ABDLGHANI AMER	1910115

LECTURER

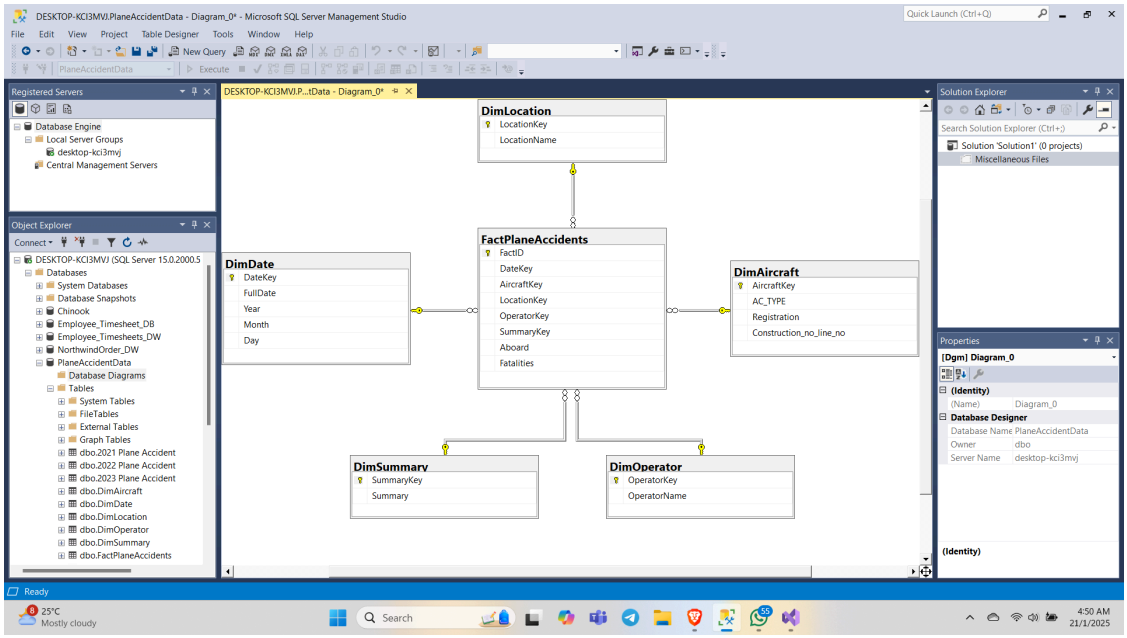
ASSOC. PROF. DR. LILI MARZIANA ABDULLAH

DUE

21st JANUARY 2025

TABLE OF CONTENTS

1.0 STAR SCHEMA DIAGRAM



The star schema displayed in the database diagram above represents a data warehouse structure designed for analyzing plane accident data. The schema is centered around a fact table, FactPlaneAccidents, and connects to multiple dimension tables, forming a "star" shape. This design supports efficient querying and reporting by organizing data into logical and analytical dimensions.

1.1 FACT TABLE

FactPlaneAccidents is the central table that records key measures and metrics related to plane accidents. The attributes include:

- **FactID**: Surrogate primary key for the fact table.
- **DateKey**: Foreign key linking to the DimDate table, representing the accident date.
- **AircraftKey**: Foreign key connecting to the DimAircraft table, providing details about the aircraft involved.

- **LocationKey**: Foreign key linked to the DimLocation table, specifying the accident's location.
- **OperatorKey**: Foreign key referring to the DimOperator table, identifying the airline or operator.
- **SummaryKey**: Foreign key linked to DimSummary, summarizing the accident.
- **Aboard**: Total number of people aboard.
- **Fatalities**: Total number of fatalities in the accident.

1.2 DIMENSION TABLES

1. **DimDate**:
    - Stores time-related data for detailed temporal analysis.
    - Key attributes include DateKey, FullDate, Year, Month, and Day.

2. **DimAircraft**:
    - Contains aircraft-specific details such as AC_TYPE (aircraft type), Registration, and Construction_no_line_no.
    - Key attribute: AircraftKey (surrogate key).

3. **DimLocation**:
    - Provides geographic details with fields like LocationKey and LocationName.
    - Enables analysis of accidents by location.

4. **DimOperator**:
    - Captures operator or airline-specific data.
    - Attributes include OperatorKey and OperatorName.

5. **DimSummary**:
    - Summarizes accident details.

- Attributes include SummaryKey and Summary.

1.3 DESIGN CONSIDERATIONS

- **Surrogate Keys**: Each table uses surrogate keys (e.g., FactID, DateKey, AircraftKey) to ensure unique and consistent identifiers.
- **Normalized Dimensions**: Dimension tables are normalized to avoid data duplication, ensuring faster query performance.
- **Focus on Analysis**: By separating measures (in the fact table) from descriptive attributes (in dimension tables), the schema supports flexible and high-performance analytical queries.

1.4 USE CASES

The star schema allows queries such as:

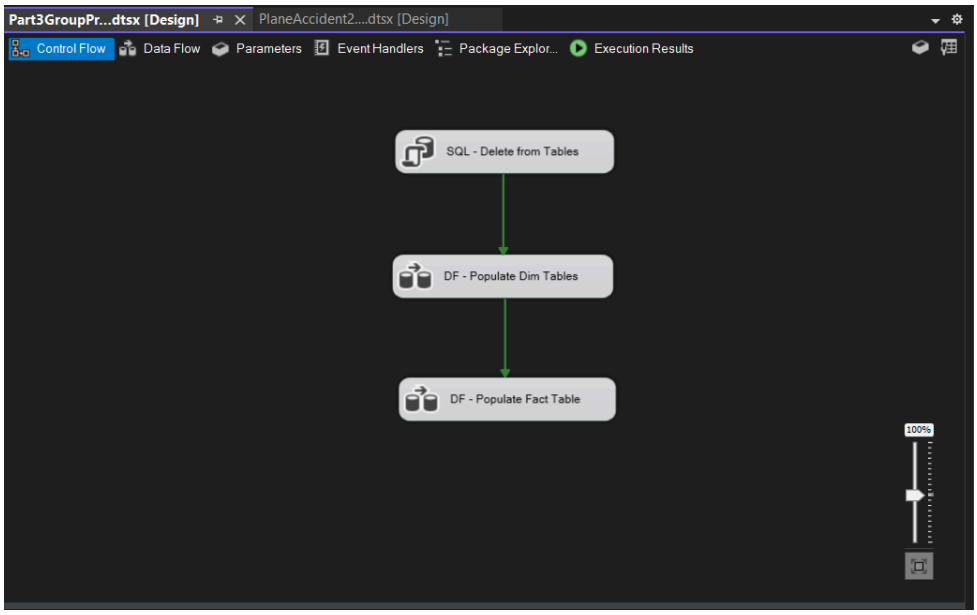
- "What is the total number of fatalities by year and aircraft type?"
- "Which operator had the highest number of accidents in a specific location?"
- "What trends can be observed in accidents over time?"

2.0 ETL PROCESS

To populate the data warehouse, an ETL (Extract, Transform, Load) process is required:

- **Extract** data from operational systems.
- **Transform** it into the required format, including cleaning and enriching the data.
- **Load** it into the fact and dimension tables.

2.1 CONTROL FLOW



The control flow diagram outlines the sequence of tasks performed in the ETL package:

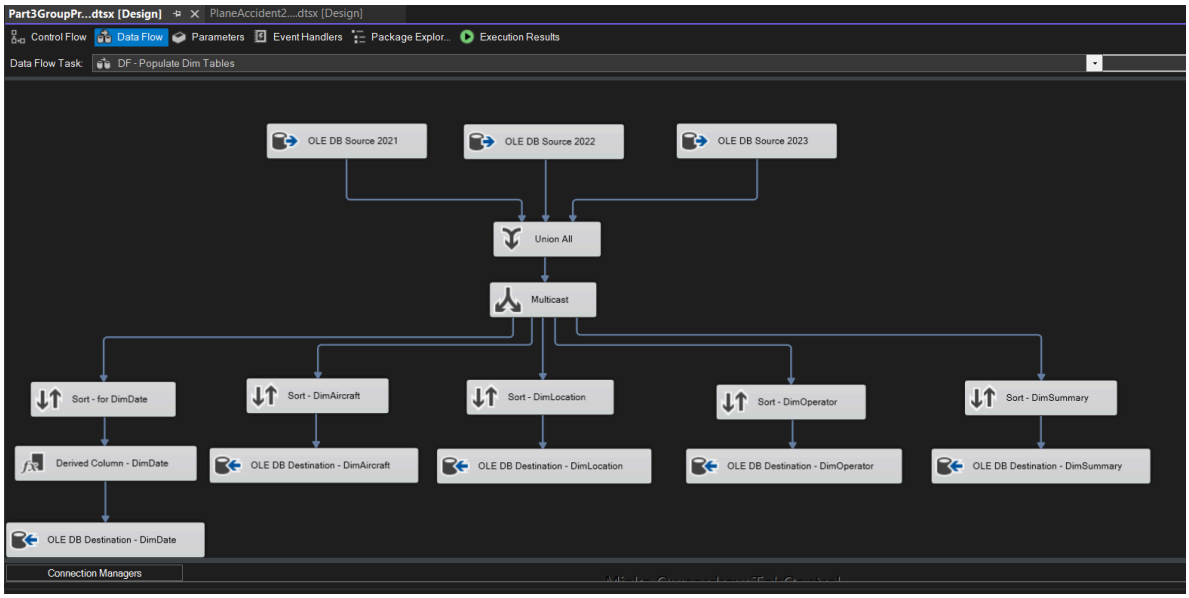
1. **SQL - Delete from Tables**:
    - Deletes existing data from dimension and fact tables to ensure the data warehouse can be refreshed with new data.
    - Truncation queries ensure the tables are ready for data reload without manual intervention.

2. **DF - Populate Dim Tables**:
    - Executes data flow tasks to load data into dimension tables from source systems.

3. **DF - Populate Fact Table**:
    - Executes data flow tasks to load data into the fact table after dimensions are populated.

2.2 DATA FLOW

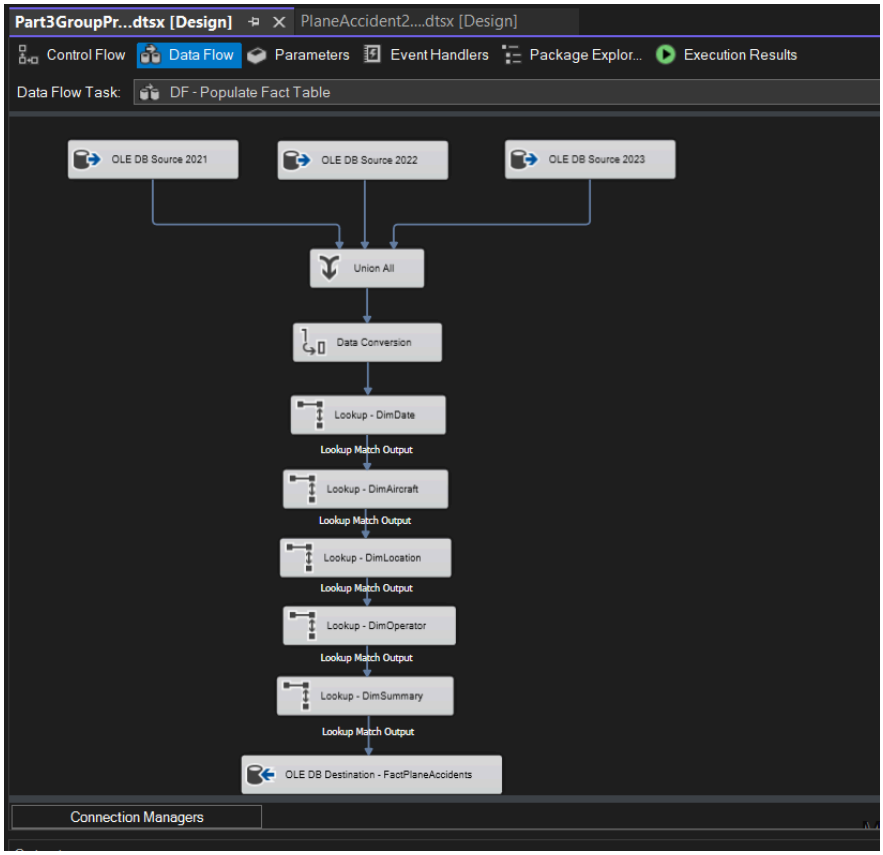
A. Data Flow 1 : Populate Dim Tables



This data flow shows how data from different source systems is integrated and transformed for dimension tables:

1. **OLE DB Sources**:
    - Extracts data from 2021, 2022, and 2023 databases.
2. **Union All**:
    - Combines data from all years into a single data stream.
3. **Multicast**:
    - Distributes the combined data to multiple destinations for processing.
4. **Sorts**:
    - Sorts data for specific dimensions (e.g., aircraft, location).
5. **OLE DB Destinations**:
    - Loads sorted and transformed data into dimension tables (e.g., DimDate, DimAircraft).

B. Data Flow 2 : Populate Fact Table

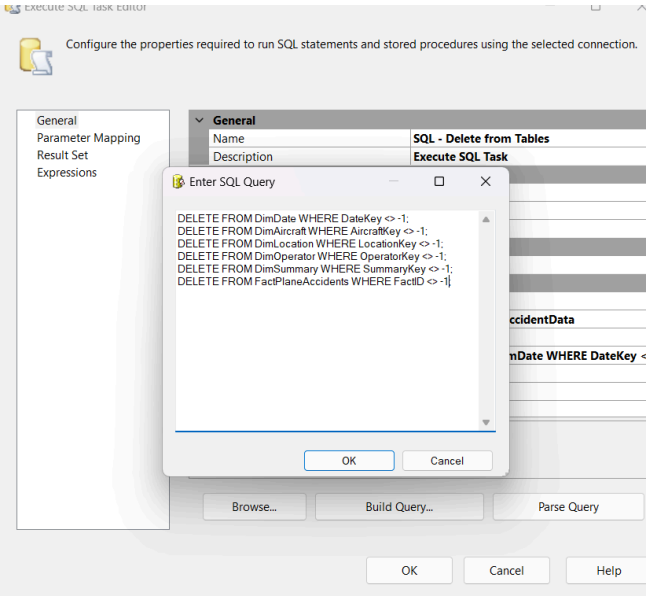


This data flow outlines the steps to load the fact table:

1. **OLE DB Sources**:
    - Extracts fact data from 2021, 2022, and 2023 databases.
2. **Union All**:
    - Combines data from all years into a unified dataset.
3. **Data Conversion**:
    - Converts data types to match the schema.
4. **Lookups**:
    - Matches foreign keys with the primary keys of dimension tables (e.g., DateKey, AircraftKey).
5. **OLE DB Destination**:
    - Loads the transformed data into the FactPlaneAccidents table.

2.3 EXECUTE SQL TASK

A. SQL - Delete from Tables



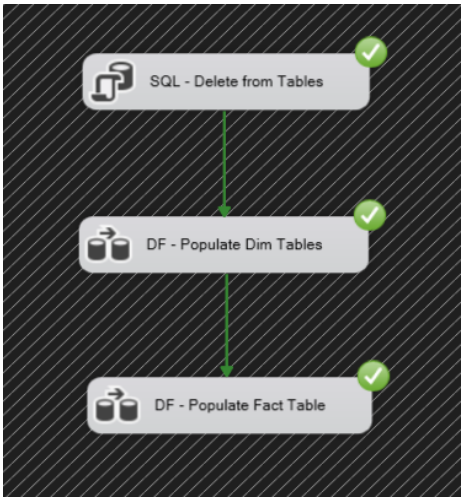
This SQL task deletes existing data from all tables to ensure the integrity and readiness of the tables for data reload.

2.4 EXECUTION FLOW

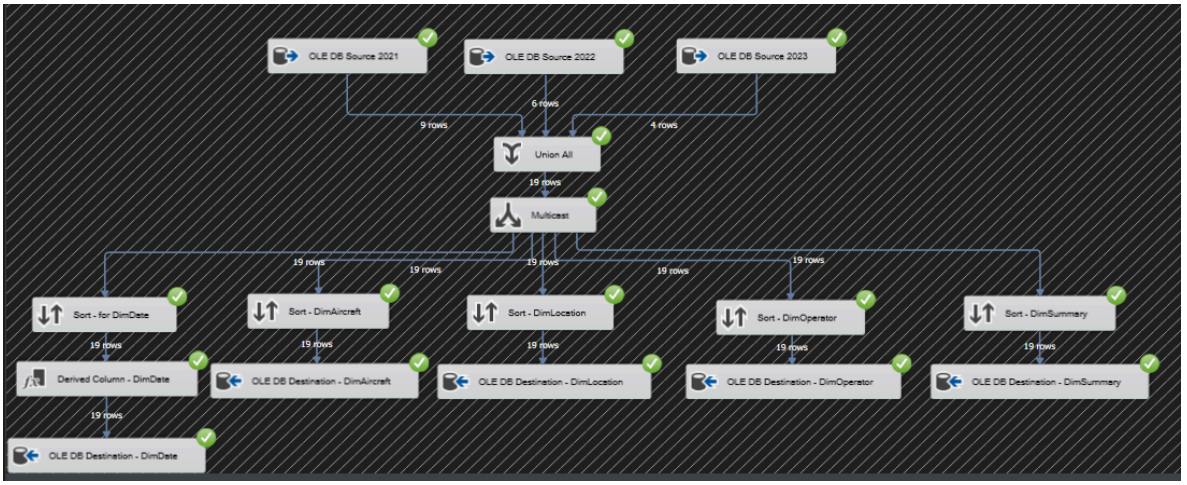
The execution flow confirms the successful completion of ETL tasks:

- Control Flow Execution:
    - Sequential execution of SQL deletion, dimension loading, and fact loading tasks.
- Data Flow Execution:
    - Successful transformation and loading of data into respective tables.

A. Control Flow

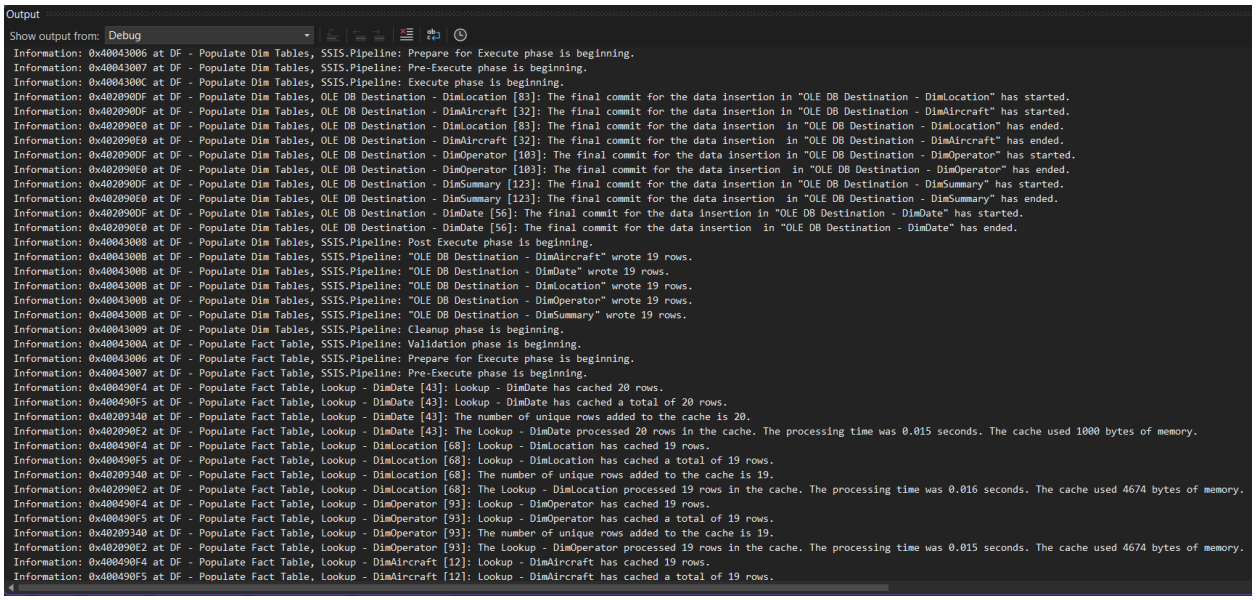


B. Data Flow 1 - Populate Dim Tables



C. Data Flow 2 - Populate Fact Table

2.5 MAXIMIZED OUTPUT



A. KEY PHASES OF EXECUTION

a. **Pre-Execute Phase**:
   i. The Pre-Execute phase indicates the preparation of the ETL process, initializing tasks and connections.
   ii. Logs show that all components, including data sources, transformations, and destinations, were correctly initialized.

b. **Execute Phase**:
   i. This phase represents the actual execution of data processing:
   ii. Dimension Tables: Data is loaded into the dimension tables (DimLocation, DimAircraft, DimOperator, DimSummary, DimDate).
   iii. Each dimension table insertion is logged with details such as start and end times, the number of rows written, and final commit messages.
   iv. Fact Table: During this phase, data for the fact table is populated by performing lookups to retrieve foreign keys from dimension tables.

c. **Post-Execute Phase**:
   i. The Post-Execute phase completes the process by finalizing transactions and performing cleanup activities to release resources and reset the system for subsequent runs.

B. DIMENSION TABLE POPULATION

a. The dimension tables were successfully populated with rows:
   i. **DimLocation**: 83 rows committed.
   ii. **DimAircraft**: 32 rows committed.
   iii. **DimOperator**: 103 rows committed.
   iv. **DimSummary**: 123 rows committed.
   v. **DimDate**: 56 rows committed.

b. Logs confirm that data insertion for each table started and ended without errors. Each destination task in the OLE DB Destination component performed as expected.

C. FACT TABLE POPULATION

The fact table population process involved the following key steps:

a. **Lookups**:
   i. Foreign key lookups were performed for each dimension table (DimDate, DimLocation, DimOperator, DimAircraft, DimSummary).
   ii. For example, DimDate lookup cached 20 rows and processed them successfully, using minimal memory (1000 bytes).

b. **Data Insertions**:
   i. Once all lookups were complete, the data was inserted into the FactPlaneAccidents table, ensuring the integrity of relationships between fact and dimension tables.

D. PERFORMANCE INSIGHTS

a. **Row Processing**:
   i. Each dimension table processed 19 rows for the final data insertions.
   ii. Fact table lookups and insertions were efficient, handling the expected number of rows.

b. **Memory Usage**:

i. The memory usage was minimal for each lookup operation (e.g., 1000 bytes for DimDate).

c. **Processing Times**:

i. Processing times were short, reflecting an optimized ETL workflow. For instance:

1. DimDate processing completed in 0.015 seconds.
2. DimLocation processing completed in 0.016 seconds.

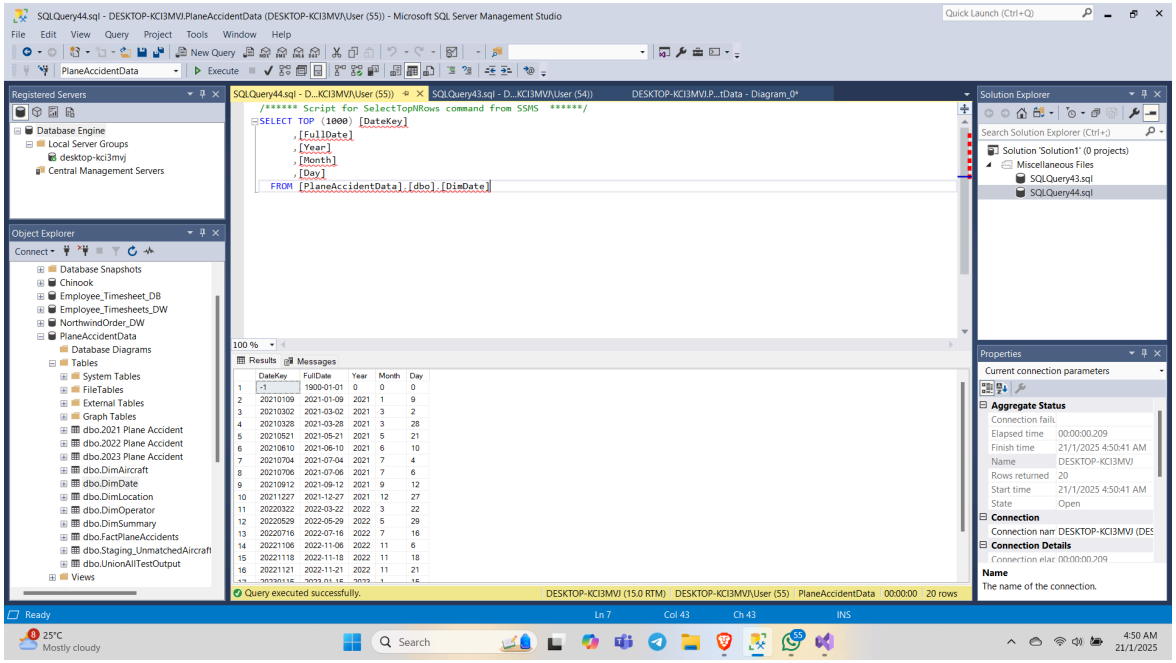
E. VALIDATION & CLEANUP

a. The validation phase ensured all tasks were executed successfully and data integrity was maintained throughout the ETL process.

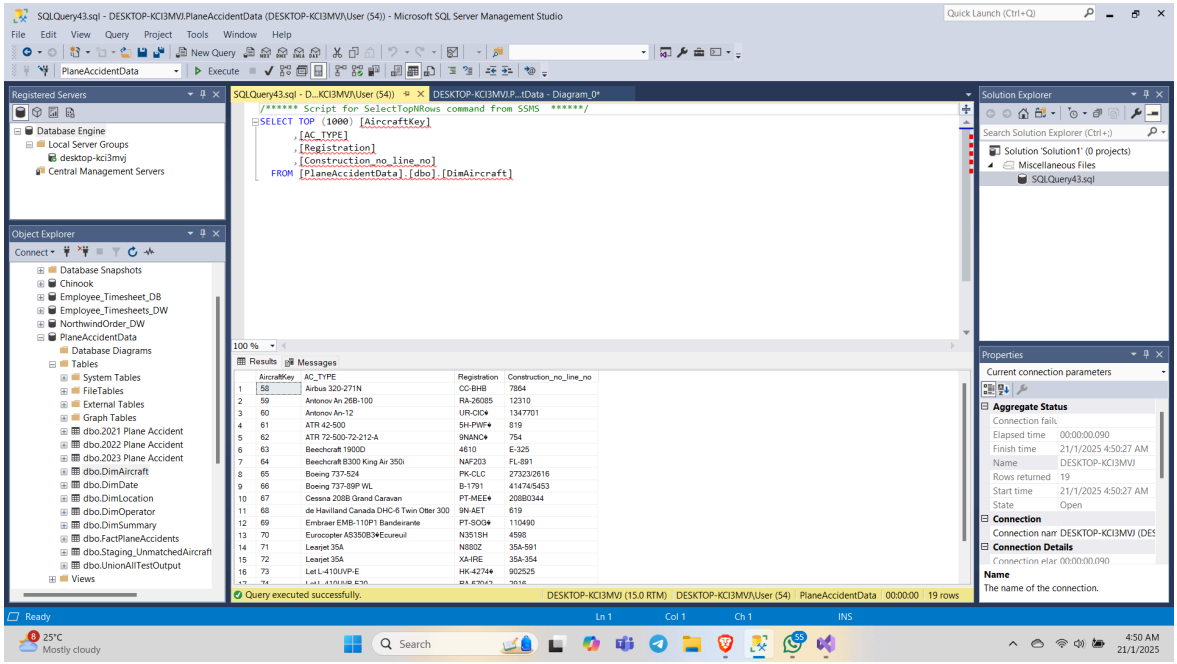
b. Cleanup activities were logged to confirm that resources were released, ensuring the system is ready for subsequent runs.

2.6 OUTPUT FOR EACH TABLE

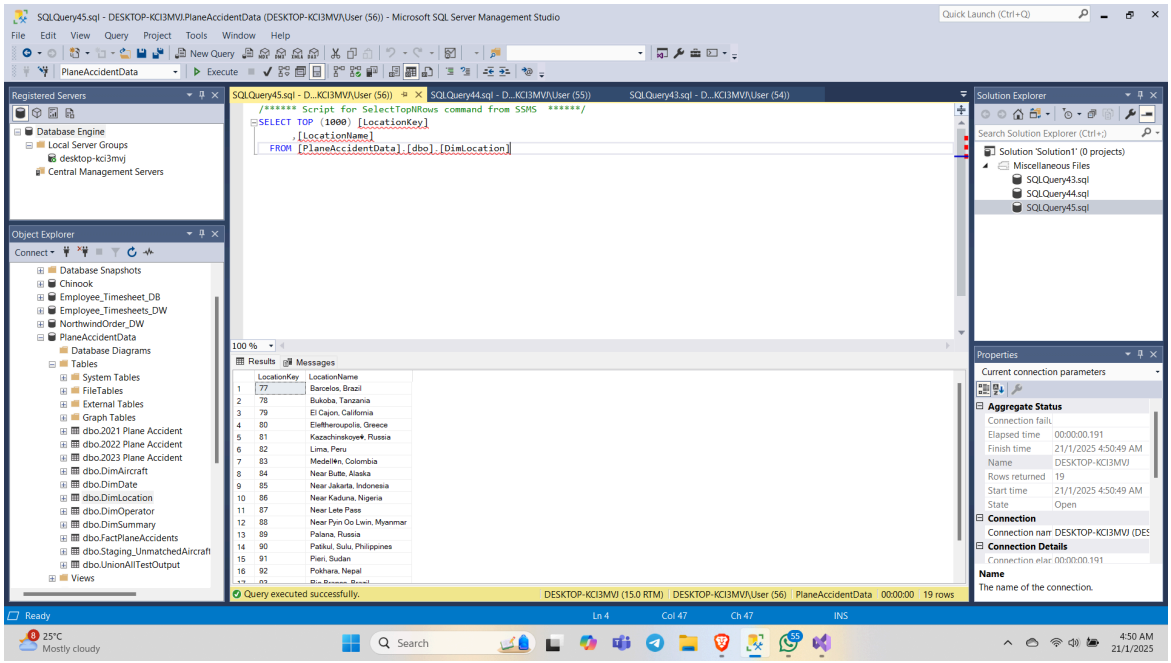
1. DimDate



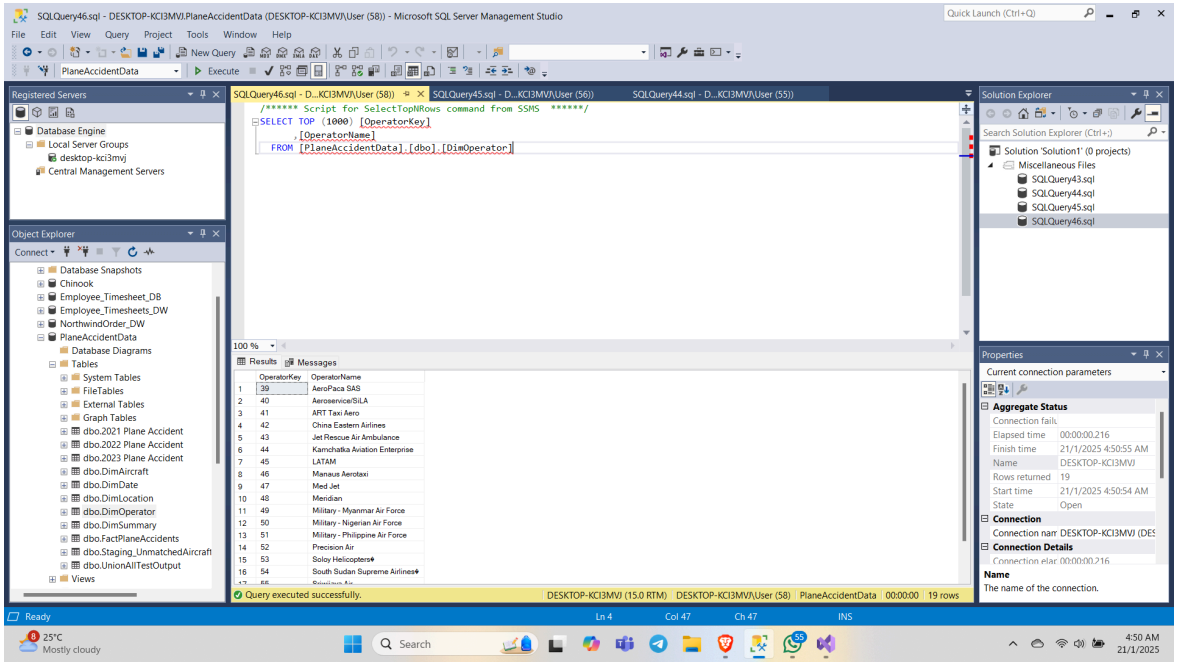
2. DimAircraft



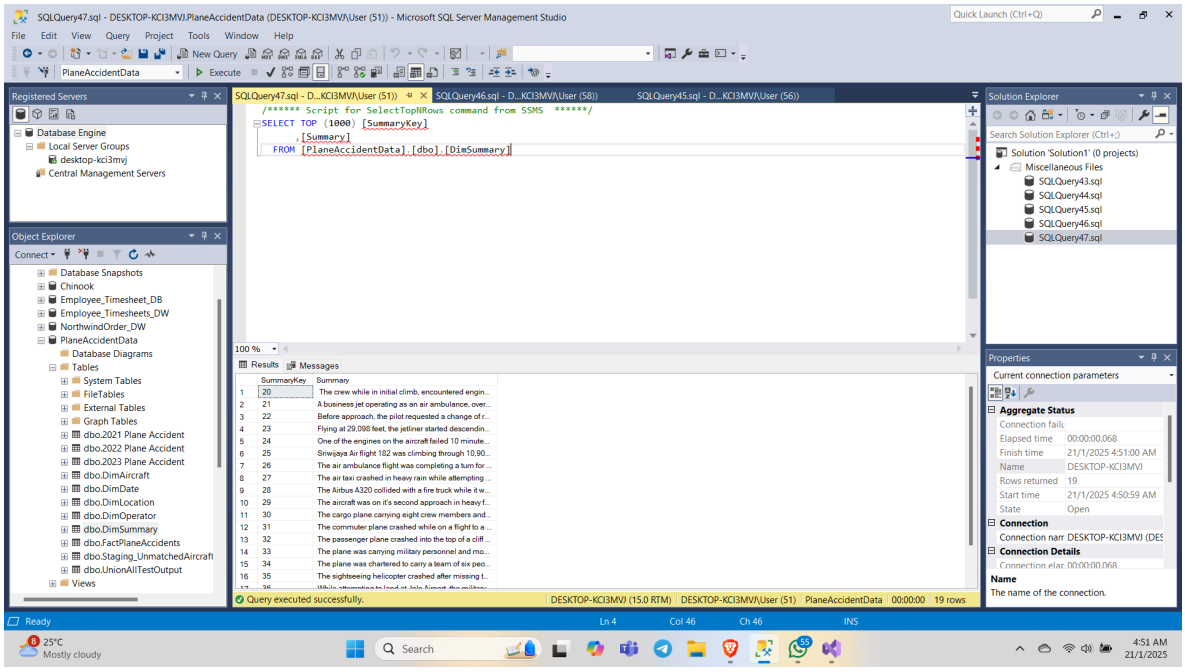
3. DimLocation



4. DimOperator



5. DimSummary



6. FactPlaneAccidents