
KULLIYAH 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

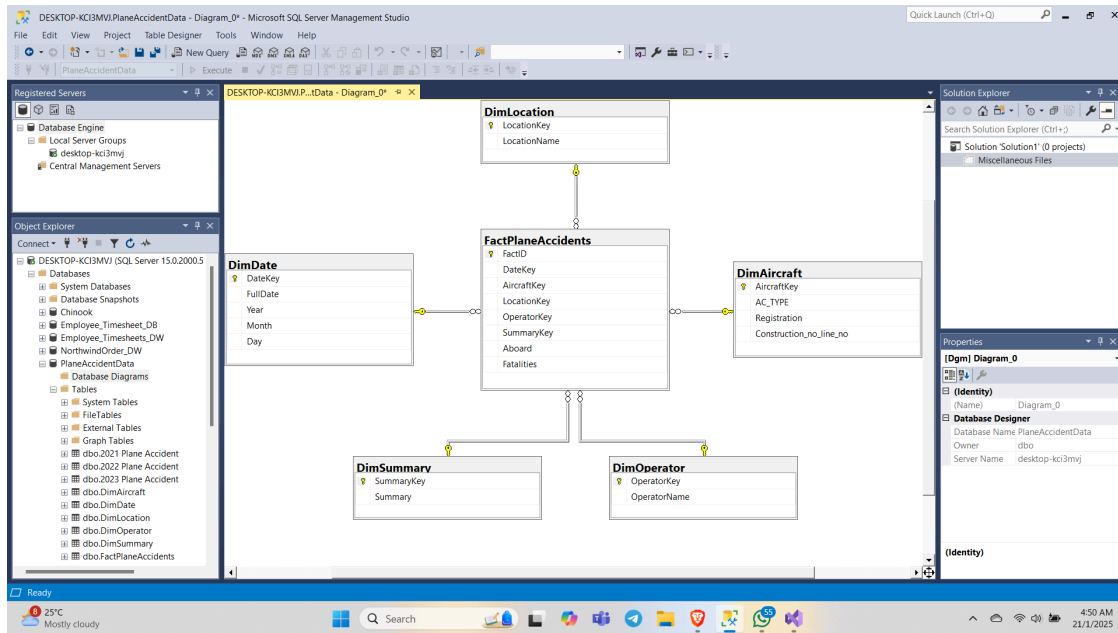
DUE

21st JANUARY 2025

TABLE OF CONTENTS

1.0 STAR SCHEMA DIAGRAM.....	2
1.1 FACT TABLE.....	2
1.2 DIMENSION TABLES.....	3
1.3 DESIGN CONSIDERATIONS.....	4
1.4 USE CASES.....	4
2.0 ETL PROCESS.....	4
2.1 CONTROL FLOW.....	5
2.2 DATA FLOW.....	6
2.3 EXECUTE SQL TASK.....	8
2.4 EXECUTION FLOW.....	8
2.5 MAXIMIZED OUTPUT.....	11
2.6 OUTPUT FOR EACH TABLE.....	14

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:

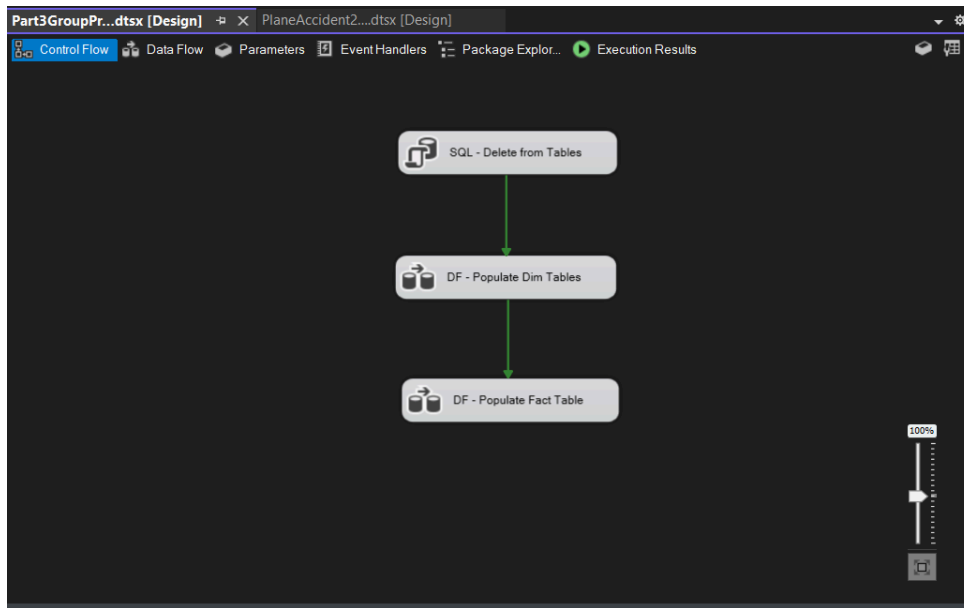
- "What is the total number of fatalities by year and aircraft type?"
- "Which operator had the highest number of accidents in a specific location?"
- "What trends can be observed in accidents over time?"

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:**

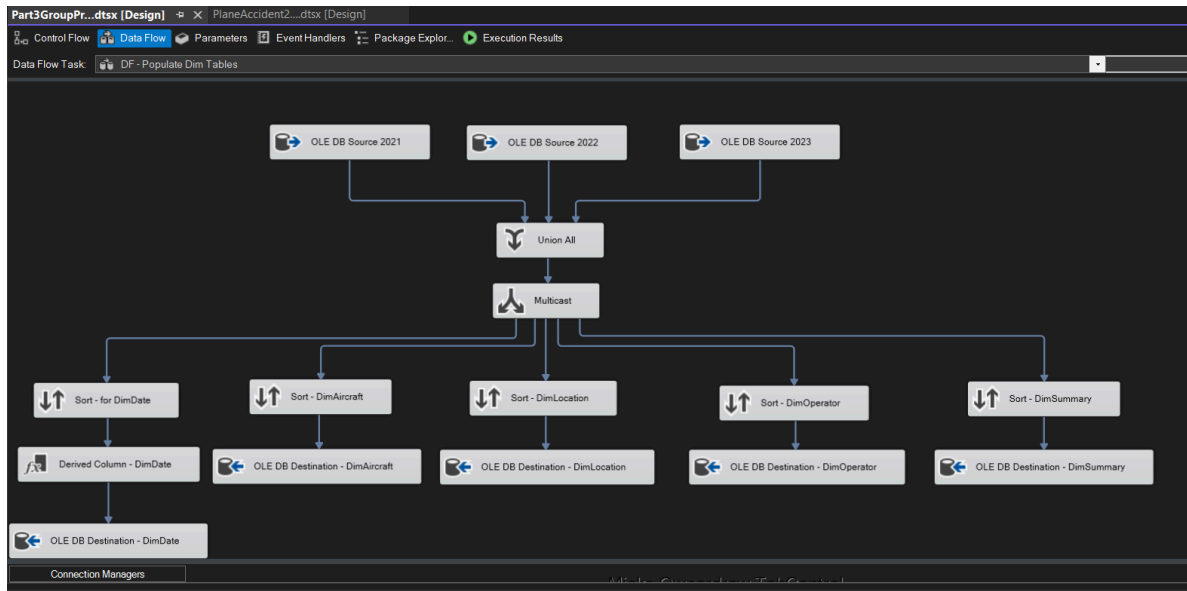
- Executes data flow tasks to load data into dimension tables from source systems.

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.

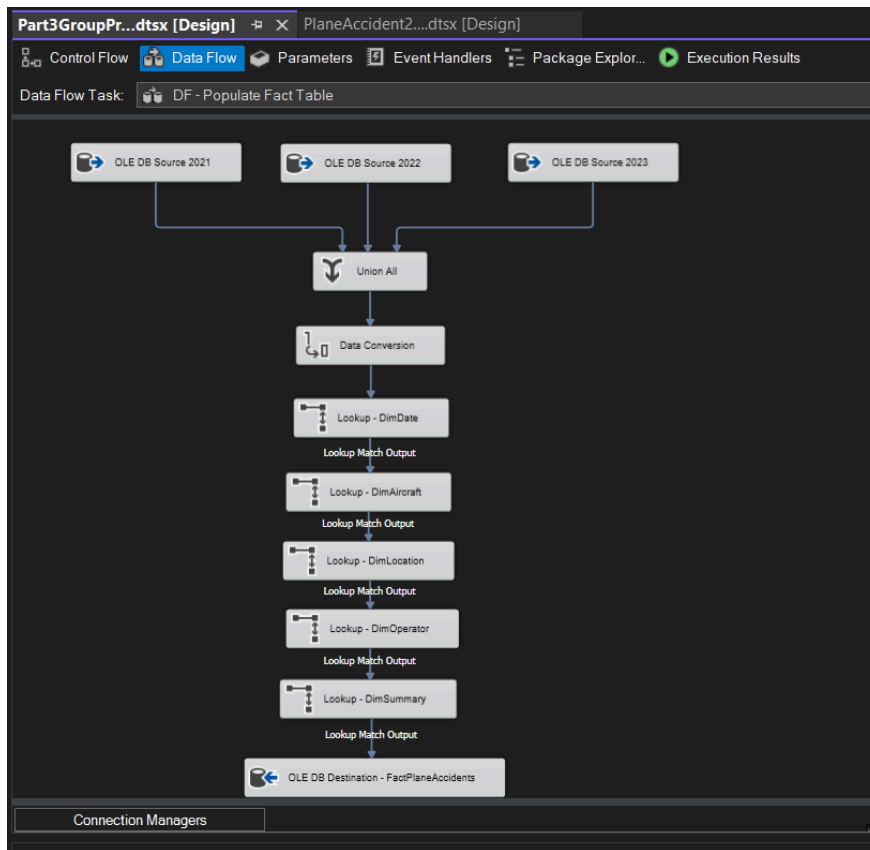
4. **Sorts:**

- 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

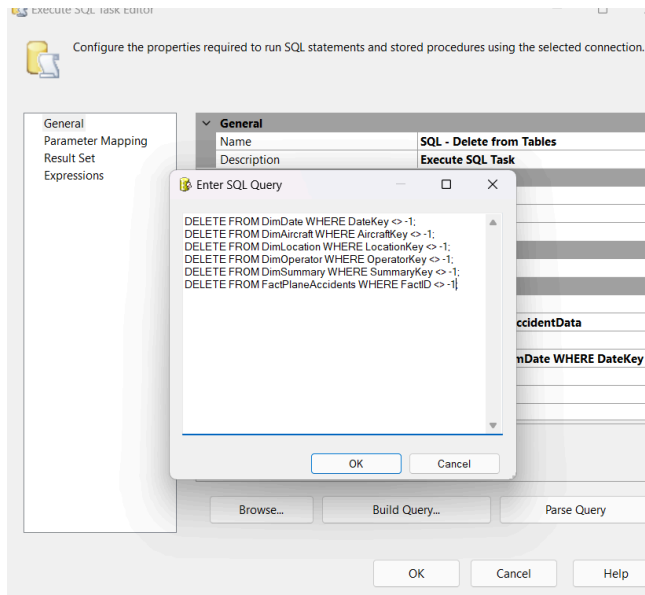


This data flow outlines the steps to load the fact table:

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



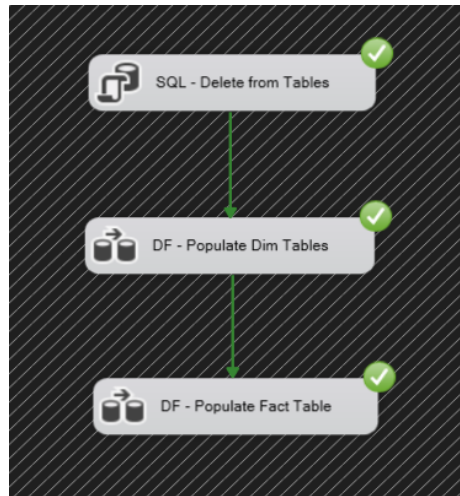
This SQL task deletes existing data from all tables to ensure the integrity and readiness of the tables for data reload.

2.4 EXECUTION FLOW

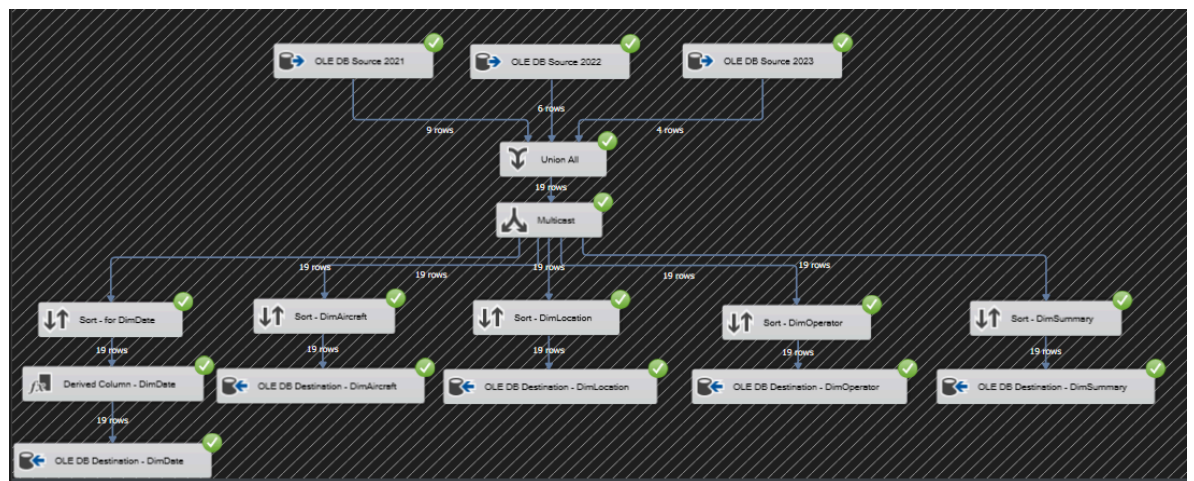
The execution flow confirms the successful completion of ETL tasks:

- 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



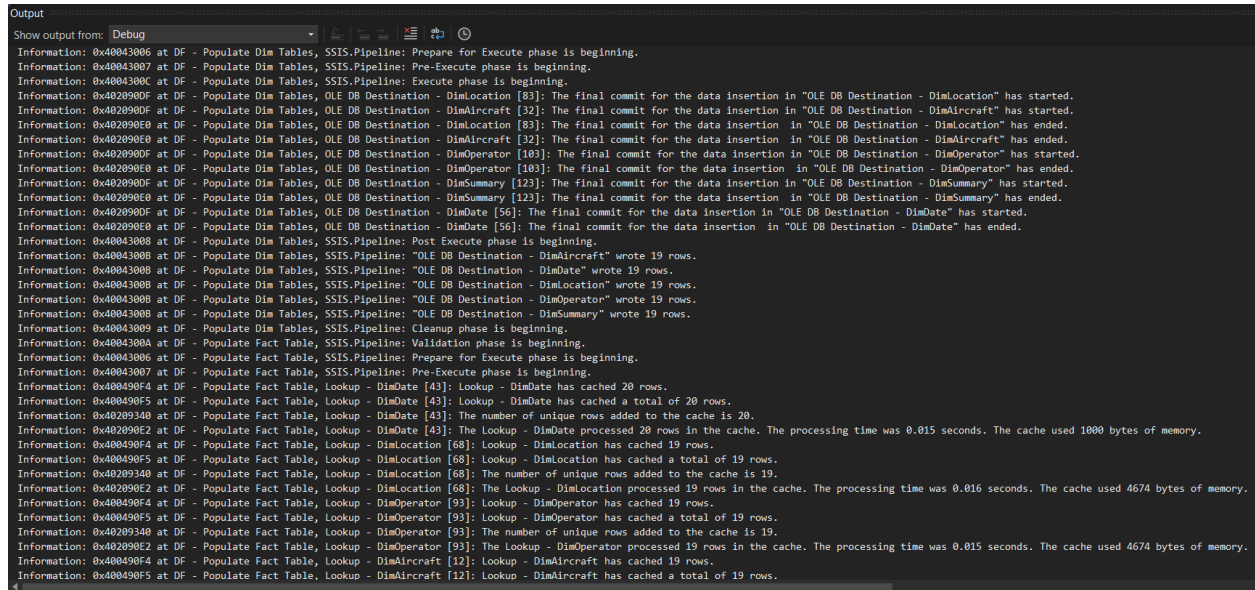
B. Data Flow 1 - Populate Dim Tables



C. Data Flow 2 - Populate Fact Table



2.5 MAXIMIZED OUTPUT



```
Output
Show output from: Debug
Information: 0x40043006 at DF - Populate Dim Tables, SSIS.Pipeline: Prepare for Execute phase is beginning.
Information: 0x40043007 at DF - Populate Dim Tables, SSIS.Pipeline: Pre-Execute phase is beginning.
Information: 0x40043008 at DF - Populate Dim Tables, SSIS.Pipeline: Execute phase is beginning.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimLocation [83]: The final commit for the data insertion in "OLE DB Destination - DimLocation" has started.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimAircraft [32]: The final commit for the data insertion in "OLE DB Destination - DimAircraft" has started.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimLocation [83]: The final commit for the data insertion in "OLE DB Destination - DimLocation" has ended.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimAircraft [32]: The final commit for the data insertion in "OLE DB Destination - DimAircraft" has ended.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimOperator [183]: The final commit for the data insertion in "OLE DB Destination - DimOperator" has started.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimSummary [123]: The final commit for the data insertion in "OLE DB Destination - DimSummary" has started.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimSummary [123]: The final commit for the data insertion in "OLE DB Destination - DimSummary" has ended.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimDate [56]: The final commit for the data insertion in "OLE DB Destination - DimDate" has started.
Information: 0x4020900F at DF - Populate Dim Tables, OLE DB Destination - DimDate [56]: The final commit for the data insertion in "OLE DB Destination - DimDate" has ended.
Information: 0x40043008 at DF - Populate Dim Tables, SSIS.Pipeline: Post-Execute phase is beginning.
Information: 0x40043008 at DF - Populate Dim Tables, SSIS.Pipeline: "OLE DB Destination - DimAircraft" wrote 19 rows.
Information: 0x40043008 at DF - Populate Dim Tables, SSIS.Pipeline: "OLE DB Destination - DimDate" wrote 19 rows.
Information: 0x40043008 at DF - Populate Dim Tables, SSIS.Pipeline: "OLE DB Destination - DimLocation" wrote 19 rows.
Information: 0x40043008 at DF - Populate Dim Tables, SSIS.Pipeline: "OLE DB Destination - DimOperator" wrote 19 rows.
Information: 0x40043008 at DF - Populate Dim Tables, SSIS.Pipeline: "OLE DB Destination - DimSummary" wrote 19 rows.
Information: 0x40043009 at DF - Populate Dim Tables, SSIS.Pipeline: Cleanup phase is beginning.
Information: 0x4004300A at DF - Populate Fact Table, SSIS.Pipeline: Validation phase is beginning.
Information: 0x40043006 at DF - Populate Fact Table, SSIS.Pipeline: Prepare for Execute phase is beginning.
Information: 0x40043007 at DF - Populate Fact Table, SSIS.Pipeline: Pre-Execute phase is beginning.
Information: 0x400490F4 at DF - Populate Fact Table, Lookup - DimDate [43]: Lookup - DimDate has cached 20 rows.
Information: 0x400490F5 at DF - Populate Fact Table, Lookup - DimDate [43]: Lookup - DimDate has cached a total of 20 rows.
Information: 0x40209340 at DF - Populate Fact Table, Lookup - DimDate [43]: The number of unique rows added to the cache is 20.
Information: 0x402090E2 at DF - Populate Fact Table, Lookup - DimDate [43]: The Lookup - DimDate processed 20 rows in the cache. The processing time was 0.015 seconds. The cache used 1000 bytes of memory.
Information: 0x400490F4 at DF - Populate Fact Table, Lookup - DimLocation [68]: Lookup - DimLocation has cached 19 rows.
Information: 0x40209340 at DF - Populate Fact Table, Lookup - DimLocation [68]: The number of unique rows added to the cache is 19.
Information: 0x402090E2 at DF - Populate Fact Table, Lookup - DimLocation [68]: The Lookup - DimLocation processed 19 rows in the cache. The processing time was 0.016 seconds. The cache used 4674 bytes of memory.
Information: 0x400490F4 at DF - Populate Fact Table, Lookup - DimOperator [93]: Lookup - DimOperator has cached 19 rows.
Information: 0x400490F5 at DF - Populate Fact Table, Lookup - DimOperator [93]: Lookup - DimOperator has cached a total of 19 rows.
Information: 0x40209340 at DF - Populate Fact Table, Lookup - DimOperator [93]: The number of unique rows added to the cache is 19.
Information: 0x402090E2 at DF - Populate Fact Table, Lookup - DimOperator [93]: The Lookup - DimOperator processed 19 rows in the cache. The processing time was 0.015 seconds. The cache used 4674 bytes of memory.
Information: 0x400490F4 at DF - Populate Fact Table, Lookup - DimAircraft [12]: Lookup - DimAircraft has cached 19 rows.
Information: 0x400490F5 at DF - Populate Fact Table, Lookup - DimAircraft [12]: Lookup - DimAircraft has cached a total of 19 rows.
```

A. KEY PHASES OF EXECUTION

a. Pre-Execute Phase:

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

b. Execute Phase:

- This phase represents the actual execution of data processing:
- Dimension Tables:** Data is loaded into the dimension tables (DimLocation, DimAircraft, DimOperator, DimSummary, DimDate).
- Each dimension table insertion is logged with details such as start and end times, the number of rows written, and final commit messages.
- 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:

- 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

The screenshot shows the Microsoft SQL Server Management Studio interface. The query editor contains the following SQL script:

```
/****** Script for SelectTopRows command from SSMS *****/
SELECT TOP (1000) [DateKey]
               [FullDate]
               [Year]
               [Month]
               [Day]
FROM [PlaneAccidentData].[dbo].[DimDate]
```

The Results pane displays the output of the query, showing columns: DateKey, FullDate, Year, Month, and Day. The data includes dates from 1900-01-01 to 2022-11-21.

The Properties pane on the right shows the current connection parameters for the connection named 'DESKTOP-KC33MVJ'.

2. DimAircraft

The screenshot shows the Microsoft SQL Server Management Studio interface. The query editor contains the following SQL script:

```
/****** Script for SelectTopRows command from SSMS *****/
SELECT TOP (1000) [AircraftKey]
               [AC_TYPE]
               [Registration]
               [Construction_no_line_no]
FROM [PlaneAccidentData].[dbo].[DimAircraft]
```

The Results pane displays the output of the query, showing columns: AircraftKey, AC_TYPE, Registration, and Construction_no_line_no. The data includes aircraft types such as Airbus 320-271N, Antonov An-26B-100, and Boeing 737-800.

The Properties pane on the right shows the current connection parameters for the connection named 'DESKTOP-KC33MVJ'.

3. DimLocation

The screenshot shows the Microsoft SQL Server Management Studio interface. The query editor contains the following SQL script:

```
/****** Script for SelectTopRows command from SSMS *****/
SELECT TOP (1000) [LocationKey]
FROM [PlaneAccidentData].[dbo].[DimLocation]
```

The Results pane displays the following data:

LocationKey	LocationName
77	Baretois, Brazil
78	Bukoba, Tanzania
79	El Cajon, California
80	Eidhousopolis, Greece
81	Kazakhstanskoye, Russia
82	Lima, Peru
83	Medellin, Colombia
84	Near Sable, Alaska
85	Near Jakarta, Indonesia
86	Near Kaduna, Nigeria
87	Near Late Pass
88	Near Pyaw Oo Lwin, Myanmar
89	Palana, Russia
90	Patikul, Bulu, Philippines
91	Pieri, Sudan
92	Pokhara, Nepal
93	...

The status bar indicates the query was executed successfully, returning 19 rows.

4. DimOperator

The screenshot shows the Microsoft SQL Server Management Studio interface. The query editor contains the following SQL script:

```
/****** Script for SelectTopRows command from SSMS *****/
SELECT TOP (1000) [OperatorKey]
FROM [PlaneAccidentData].[dbo].[DimOperator]
```

The Results pane displays the following data:

OperatorKey	OperatorName
39	AerPosta SAS
40	Aerosevica/SILA
41	ART Taxi Aero
42	China Eastern Airlines
43	Jet Rescue Air Ambulance
44	Kamchaltla Aviation Enterprise
45	LATAM
46	Manaus Aerotaxi
47	Med Jet
48	Meridien
49	Military - Myanmar Air Force
50	Military - Nigerian Air Force
51	Military - Philippine Air Force
52	Precision Air
53	Soloy Helicopters
54	South Sudan Supreme Airlines
...	...

The status bar indicates the query was executed successfully, returning 19 rows.

5. DimSummary

The screenshot shows the SQL Server Enterprise Manager interface. The left pane displays the database structure, including the PlaneAccidentData database and its tables. The right pane shows the Properties window for the DimSummary table, indicating it is a fact table with 19 rows. The bottom pane displays the results of a query executed successfully, showing 19 rows of data from the DimSummary table.

SummaryKey	Summary
1	The crew while in initial climb, encountered engine...
2	A business jet operating as an air ambulance, over...
3	Before approach, the pilot requested a change of r...
4	Flying at 29,000 feet, the jetliner started descendin...
5	One of the engines on the aircraft failed 10 minute...
6	Snwoyae Air flight 102 was climbing through 10,90...
7	The air ambulance flight was completing a turn for...
8	The air taxi crashed in heavy rain while attempting...
9	The Airbus A320 collided with a fire truck while it w...
10	The aircraft was on it's second approach in heavy f...
11	The cargo plane carrying eight crew members and...
12	The commuter plane crashed while on a flight to a...
13	The passenger plane crashed into the top of a cliff...
14	The plane was carrying military personnel and mil...
15	The plane was chartered to carry a team of six peo...
16	The sightseeing helicopter crashed after missing L...
17	
18	
19	

6. FactPlaneAccidents

The screenshot shows the SQL Server Enterprise Manager interface. The left pane displays the database structure, including the PlaneAccidentData database and its tables. The right pane shows the Properties window for the FactPlaneAccidents table, indicating it is a fact table with 19 rows. The bottom pane displays the results of a query executed successfully, showing 19 rows of data from the FactPlaneAccidents table.

FactID	DateKey	AircraftKey	LocationKey	OperatorKey	SummaryKey	Aboard	Fatalities
1	20210109	65	85	55	25	62	62
2	20210302	73	91	54	24	10	10
3	20210328	70	64	53	35	6	5
4	20210521	64	86	50	38	11	11
5	20210610	63	88	49	33	14	12
6	20210704	75	90	51	36	96	50
7	20210706	59	89	44	32	28	28
8	20210912	74	81	40	29	16	4
9	20211227	71	79	47	26	4	4
10	2020322	66	95	42	23	132	132
11	2020929	68	87	56	31	22	22
12	2020716	60	80	48	30	8	8
13	20221106	61	78	52	37	43	19
14	20221118	58	82	45	28	108	0
15	2022121	76	83	39	34	8	8
16	2020115	62	62	57	22	72	72
17	2020916	68	77	46	27	14	14