**SRI LANKA INSTITUE OF**

**INFORMATION TECHNOLOGY**



**ID No: IT20067342**

**Name: Jayasuriya J.A.D.A.S**

**Batch: DS weekday**

**Assignment: 01**

Table of Contents

[01).Data set selection 3](#_Toc103696779)

[ER Diagram 3](#_Toc103696780)

[02).Preparation of Data Sources 4](#_Toc103696781)

[03).Solution architecture 5](#_Toc103696782)

[04). Data warehouse design & development 7](#_Toc103696783)

[05).ETL Development 8](#_Toc103696784)

[Data Extraction 8](#_Toc103696785)

[5.1 Accident Data from Source to Staging 8](#_Toc103696786)

[5.2 Flight Data from Source to Staging 9](#_Toc103696787)

[5.3 Airport Data from Source to Staging 10](#_Toc103696788)

[Overall control flow 11](#_Toc103696789)

[Data Profiling 12](#_Toc103696790)

[b)Load Slowly changing Dimensions 13](#_Toc103696791)

[5.11 Customer Data from Staging to Data Warehouse 13](#_Toc103696792)

[5.12 Airport Data from Staging to Datawarehouse 14](#_Toc103696793)

[5.15 Creation of Date Dimension 15](#_Toc103696794)

[c)Load Fact Table 17](#_Toc103696795)

[6).ETL development – Accumulating fact tables 19](#_Toc103696796)

# 01).Data set selection

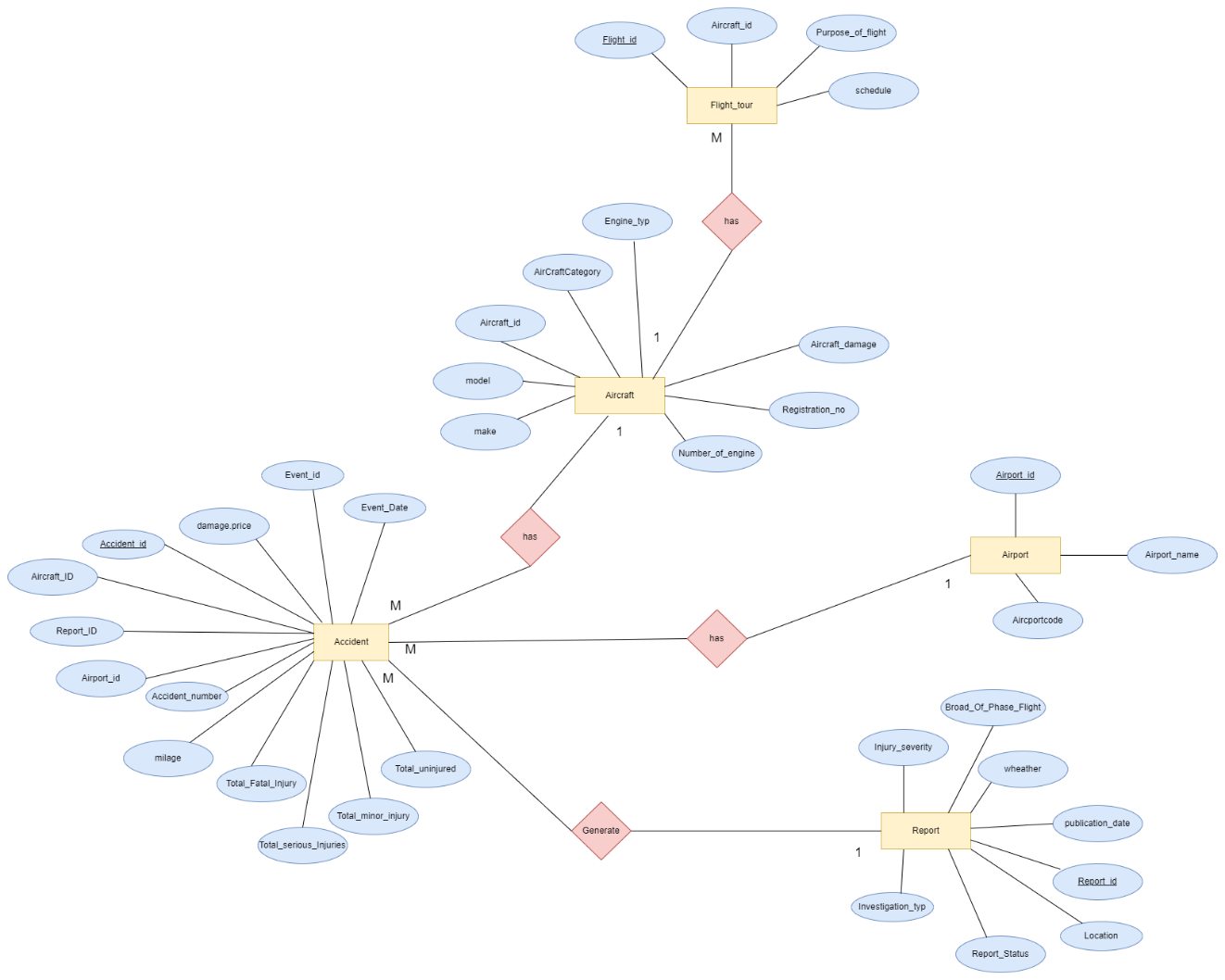
Data set : Aviation Accident Database & Synopses

Source : Kaggle

Link to the source: <https://www.kaggle.com/datasets/khsamaha/aviation-accident-database-synopses>

The dataset contains the aviation accident information .The data set consists of five files: three csv file, one excel file and one text files (Necessary modifications has been done in order to meet the requirements).

### ER Diagram



*Figure 1.0-ER diagram*

# 02).Preparation of Data Sources

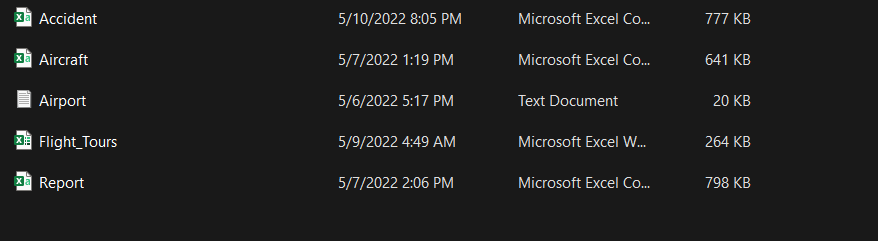
The dataset was originally in the form of one excel file. Data in the file has been seperated into 5 different files of type Excel,CSV and text.

|  |  |
| --- | --- |
| **Table** | **File type** |
| Accidents | Csv file(.csv) |
| Aircraft | Csv file(.csv) |
| Reports | Csv file(.csv) |
| Aircrafts | Excel file(.xlsx) |
| Flights | Text file(.txt) |
|  |  |
|  |  |

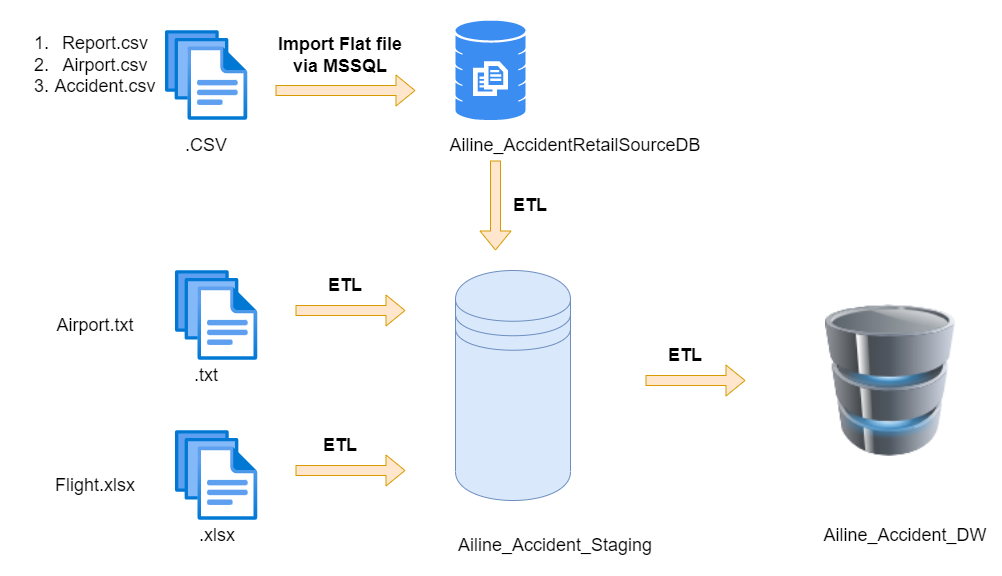
*Figure 1.1-sample source table creation*

Similarly other tables has also been created and then the tables has been exported in relevant file types.

Final set of Sources:

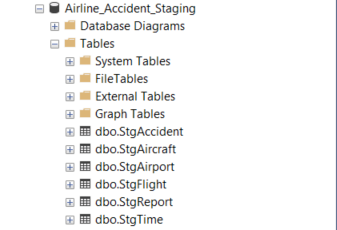


# 03).Solution architecture



*Figure 2.0-Solution architecture*

As can be seen in the figure three different resource types has been used to extract data to staging. Staging layer has been used to have all the tables in a single location as in the below figure.

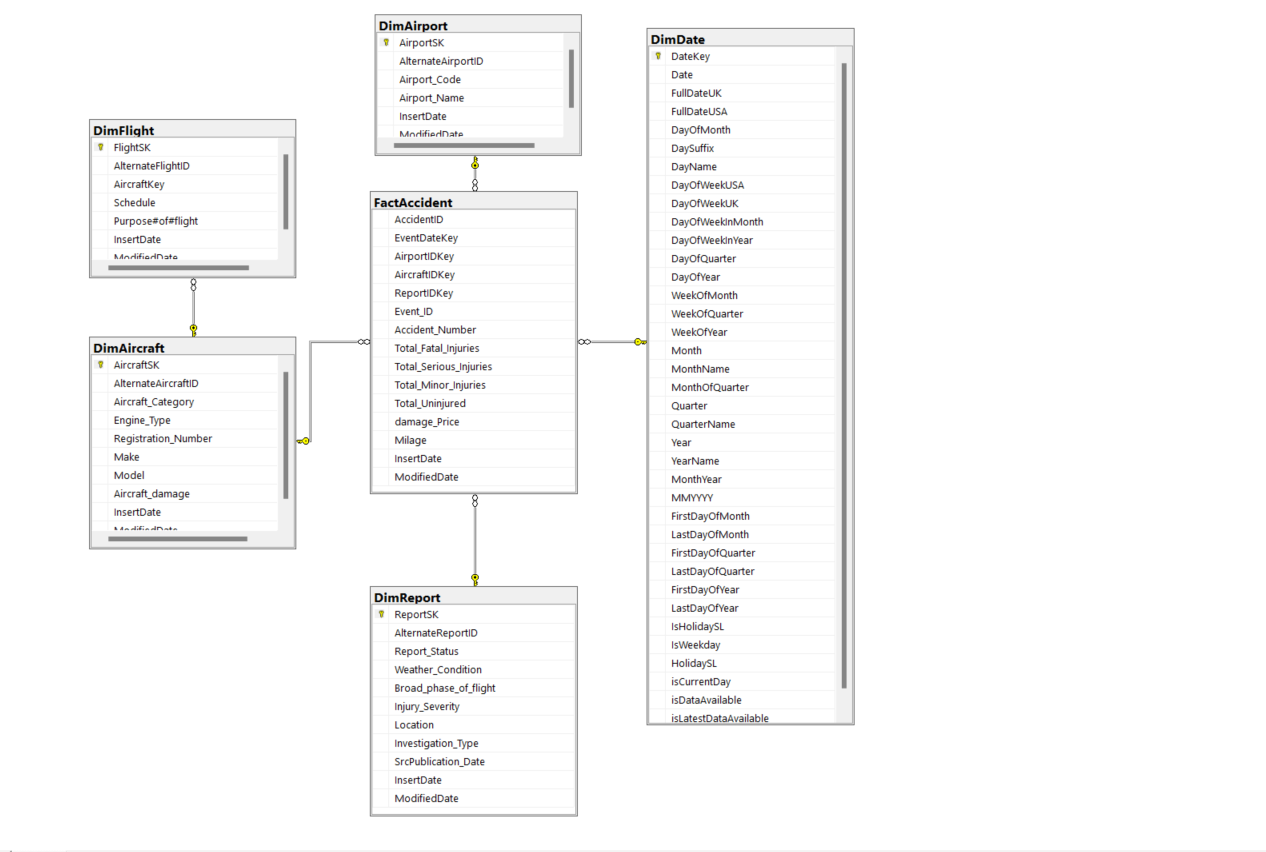


*Figure 3.0-Staging*

The tables at the staging are then profiled and after performing a rich set of ETL tasks, data is loaded to the data warehouse where from that several reporting tools and analysing tools can use data for reporting mining and analysing.

# 04). Data warehouse design & development

The datawarehouse is designed as a snow flake schema with one fact table and five dimension table.



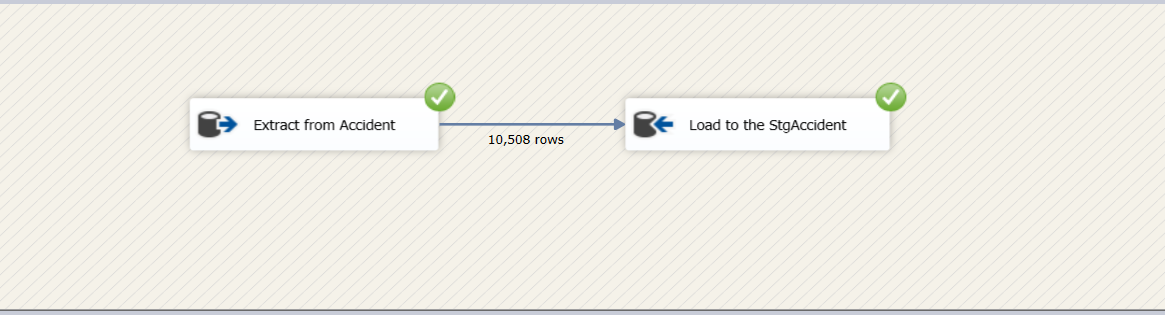
*Figure 4.0-Datawarehouse design*

# 05).ETL Development

As the first step data has been extracted from sources to staging area. Data flow task has been used for every extraction.

## Data Extraction

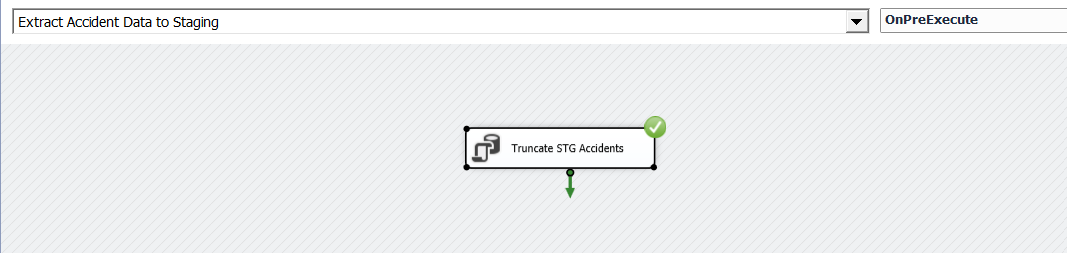
### 5.1 Accident Data from Source to Staging

5.1.1 Data Flow 

*Figure 5.1.1-Accident data flow*

5.1.2 Event handler

Before executing ‘extract Accident to staging’ existing data in the staging layer has been truncated.



*Figure 5.1.2.-Accident Event handler*

### 5.2 Flight Data from Source to Staging

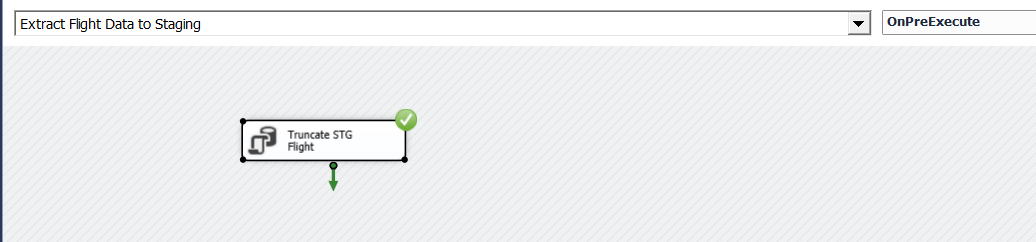
5.2.1 Data Flow

|  |  |
| --- | --- |
|  |  |

*Figure 5.2.1-Flight data flow*

5.2.2 Event handler

Before executing ‘extract flight to staging’ existing data in the staging layer has been truncated.



*Figure 5.2.2.-Flight Event handler*

### 5.3 Airport Data from Source to Staging

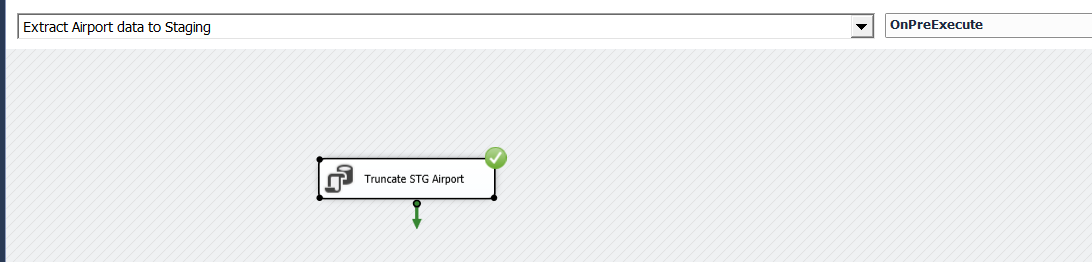
5.3.1 Data Flow

|  |  |
| --- | --- |
|  |  |

*Figure 5.3.-Airport data flow*

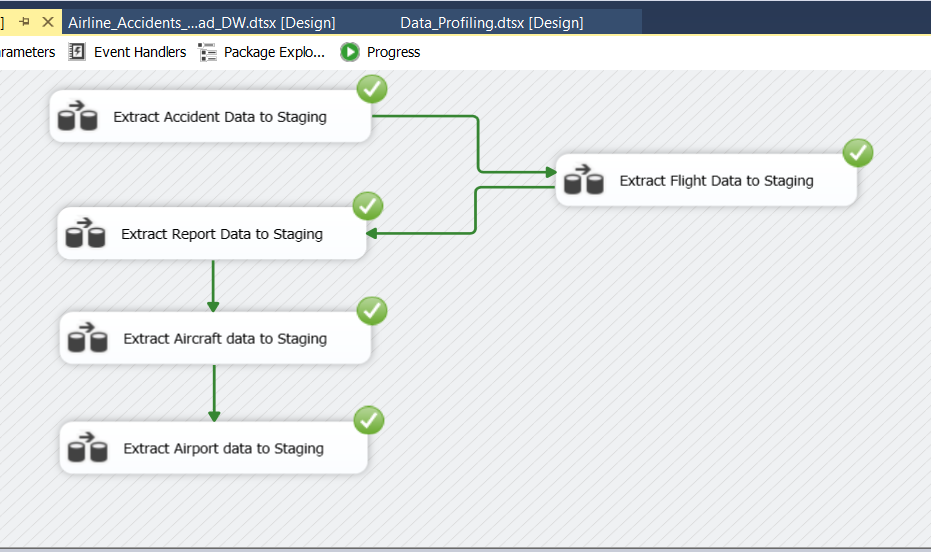
5.3.2 Event handler

Before executing ‘extract Airport to staging’ existing data in the staging layer has been truncated.



*Figure 5.3.1.-Airport Event handler*

### Overall control flow

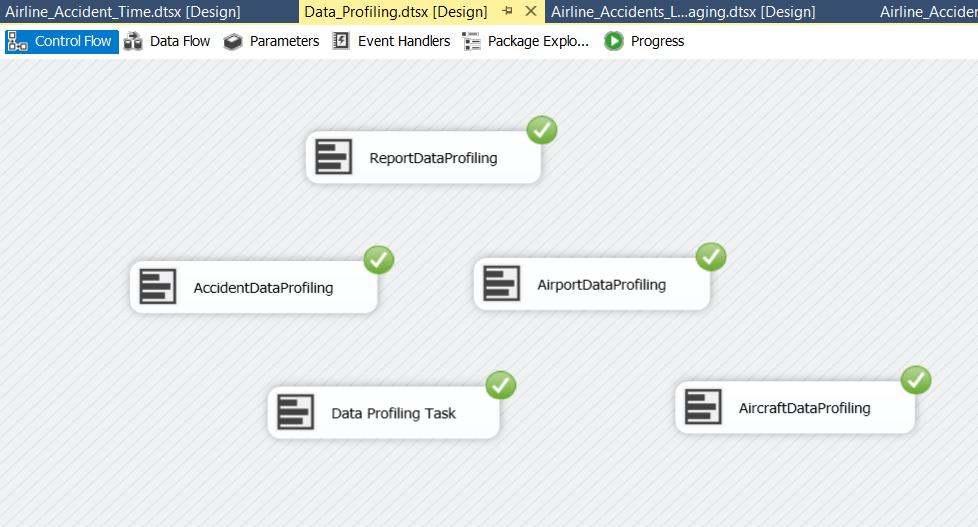
**

*Figure 5.8.-source to staging control flow*

### Data Profiling

Before Loading staging tables to the data warehouse data has to be enriched to obtain the most suitable data for analysing. Data profiling has been done in order to identify what need to be corrected in ETL process in order to meet this requirement.

Each and every table at staging is profiled and stored in a specific file location.

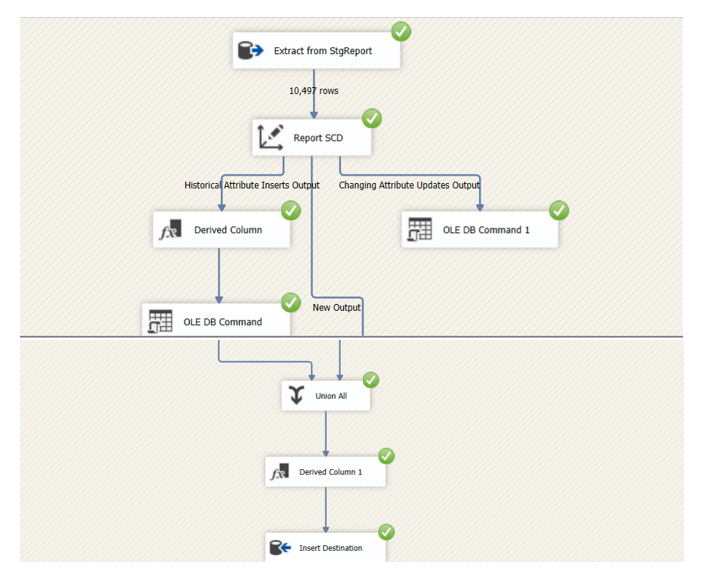


*Figure 5.9.1-profiling diagram*

Data Transforming and loading

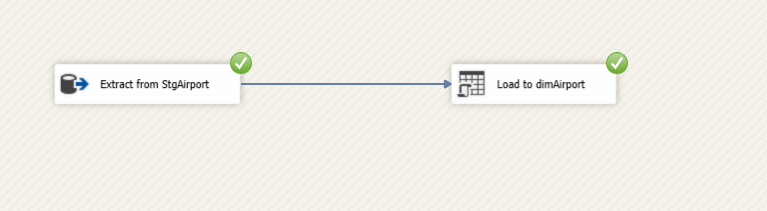
### b)Load Slowly changing Dimensions

### 5.11 Report Data from Staging to Data Warehouse



### 5.12 Airport Data from Staging to Datawarehouse

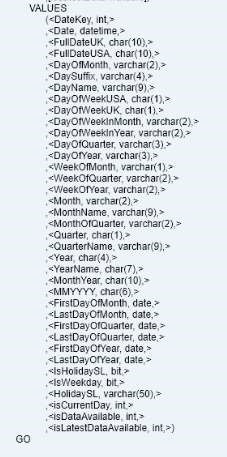
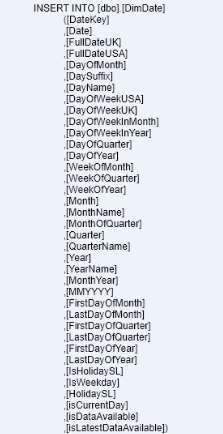
StgAirport data has been loaded to dimAirport



*Figure 5.12.1-load to DimAirport*

### 5.15 Creation of Date Dimension

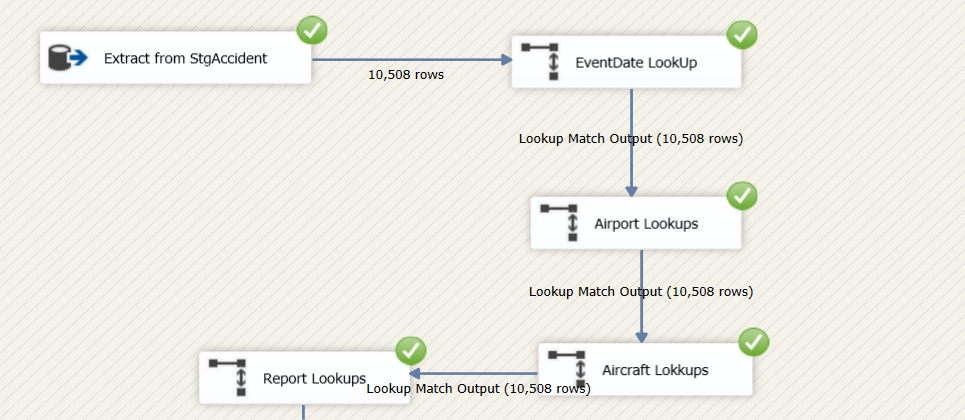
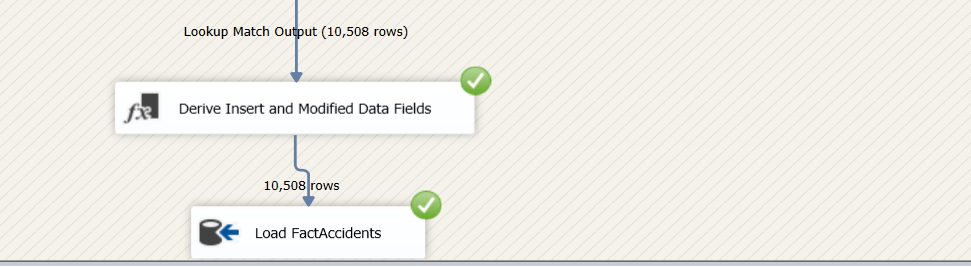
  *Figure 5.14.1-Load to DimDate*



Query used to create and load data to date dimension is listed above. Date dimension is assumed to tally with publication\_date in report table

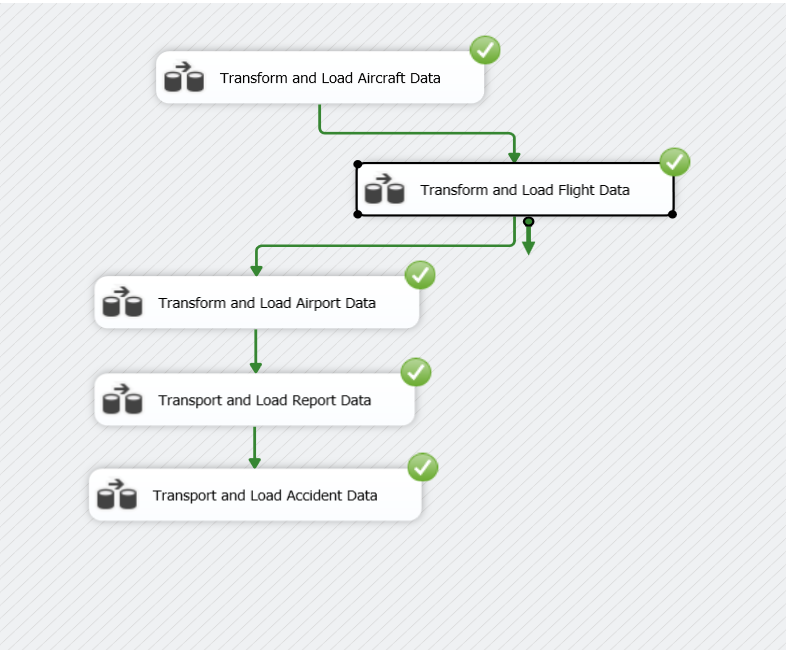
### c)Load Fact Table

StgAccident is loaded and merged to obtain Accident\_ID. All required surrogate keys has been loaded to data warehouse after a lookup through alternate keys in dimension tables.

*Figure 5.16.1-FactAccident ETL*

Overall ETL Transformation



*Figure 5.17.1-overall ETL to data warehouse*

# 6).ETL development – Accumulating fact tables

Fact table was extended by adding last three attributes as shown below.

Table

Description automatically generated

Then a separate data source (.txt) named Time.txt was created.

Table

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

**Extract Time data to Staging**

A picture containing chart

Description automatically generated

Time data from staging to datawarehouse

Graphical user interface, text, application, chat or text message

Description automatically generated

Time data source merge to the Fact table.

Data from Time Staging table and FactAccident fact table were extracted and merged. Merge has been performed by sorting both tables using the field fact\_table\_natural\_key(txt\_id)

Chart

Description automatically generated

Relevant column mapping is shown below.

Graphical user interface, Word

Description automatically generated

A derived column task has been used to derive the values for txt\_process\_time\_hours column by getting the date different of\_acc\_txt\_complete\_time and accn\_txt\_create\_time

Graphical user interface, text, application, Word

Description automatically generated

The following procedure is used to in order to load data.

A screenshot of FactAccident table after accumulating the (completing the step- 06) is given below

Table

Description automatically generated