Sri Lanka Institute of Information Technology

Data Warehousing and Business Intelligence
IT3021
Assignment 1

Submitted by:
Full name: - H.A.M Senadheera
Registration No.: - IT20125998
Batch: - Y3.S1.DS.WE.04

Table of Contents

Data Set Selection

Data Set name: - Boston Airbnb open data

Source: - Kaggle.com

Link to the source: - https://www.kaggle.com/datasets/airbnb/boston

Introduction to the dataset

This is a dataset of Airbnb bookings in the city of Boston in United States of America in the year 2016. The initial dataset is adjusted and arranged to meet the requirements of the assignment and suit the scenario

Content of the dataset	
Hosts	Contains the data of individuals who have listed their properties in Airbnb
Listings	Contains the data of properties listed in the Airbnb
Customers	Contains the data of individuals who have done bookings/reservations for listed properties
CustomerAddress	Contains addresses of the customers.
Bookings	Contains the details of bookings that customers have made to listed properties.

Preparation of the data set

Dataset was obtained as a single large csv containing listing and hosts.

While preparing the dataset listings and host were broken into two csv files.

Three more csv files were created by generating some sample data named as customer, booking and accumulated complete date csv.

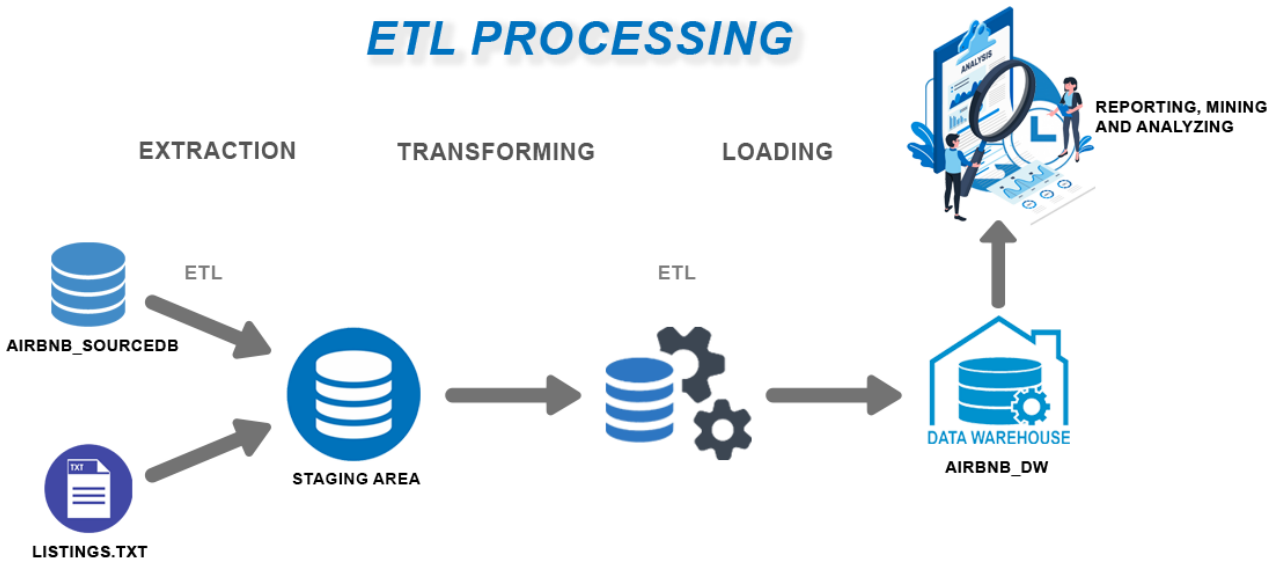
A text file was created by adding customer address information.

- Listing.csv
- Booking.csv
- Host.csv
- Customer.csv
- CustomerAddress.txt
- AccumulatedCompleteDate.csv

Using the csv files, a database named Airbnb_SourceDb was created. The sourceDB, csv and text files are as follows.

- Airbnb_SourceDB (source database)
  - dbo.Host
  - dbo.Listing
  - dbo.Customer
  - dbo.Booking
- CustomerAddress.txt
- AccumulatedCompleteDate.csv

Architecture



A data source is an initial location where information is obtained for a given scenario. The primary data source in this case is a database and secondary data source is a flat file.

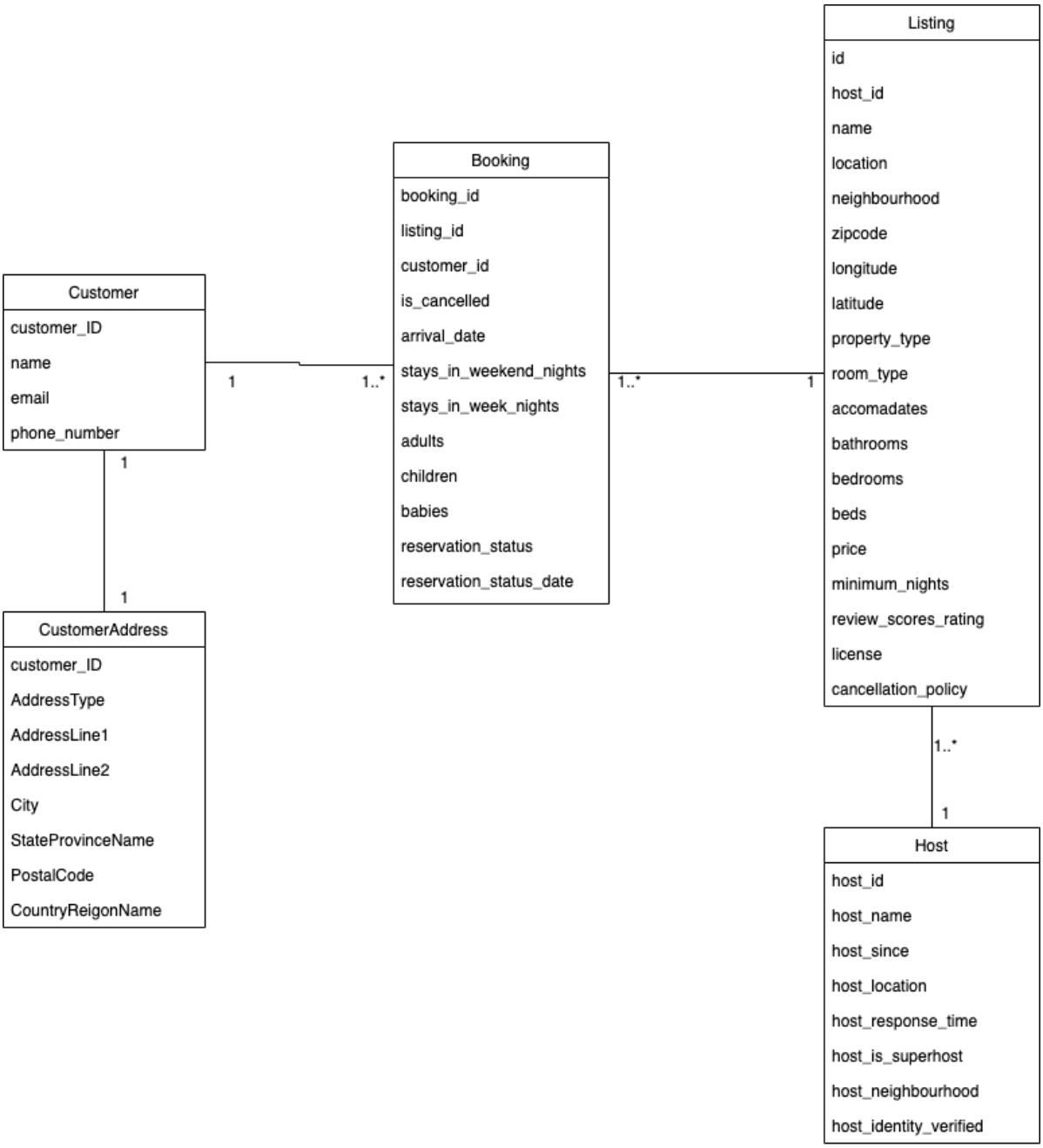
ETL, which stands for extract, transform and load, is a data integration process that extracts and combines data from multiple data sources into a single, consistent data store that is loaded into a Data Warehouse system or other target systems.

- Airbnb_Staging (staging database)
    - dbo.StgHost
    - dbo.StgListing
    - dbo.StgCustomer
    - dbo.StgBooking

The data warehouse (DWH) is a central repository where an organization electronically stores data by extracting it from operational systems and making it available for data analysis and scheduled reporting. In contrast, the process of building a data warehouse entail designing a data model that can quickly generate insights.

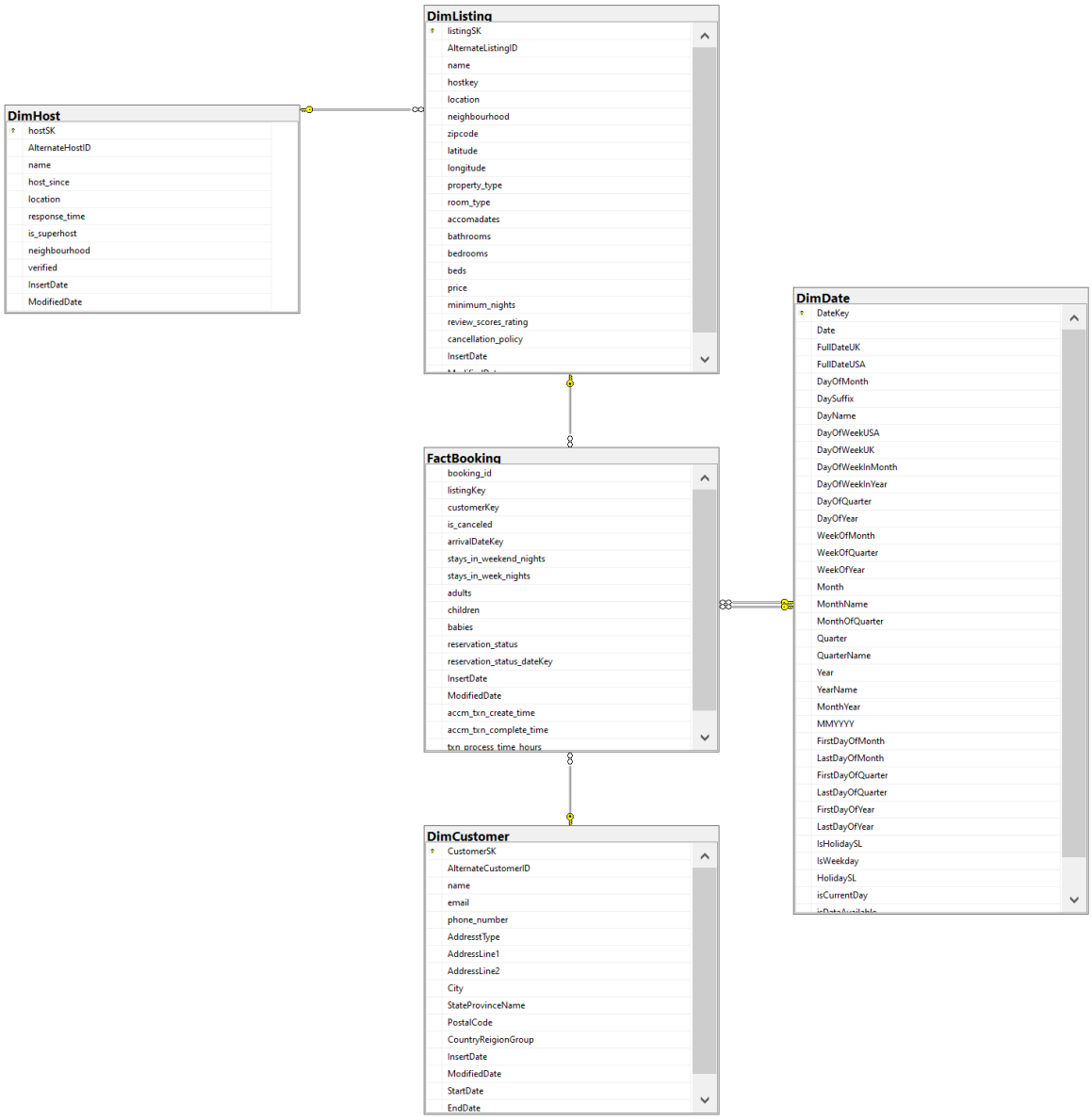
- Airbnb_DW (data warehousing database)
    - dbo.DimHost
    - dbo.DimListing
    - dbo.DimCustomer
    - dbo.DimDate
    - dbo.FactBooking

Class diagram using the sources tables



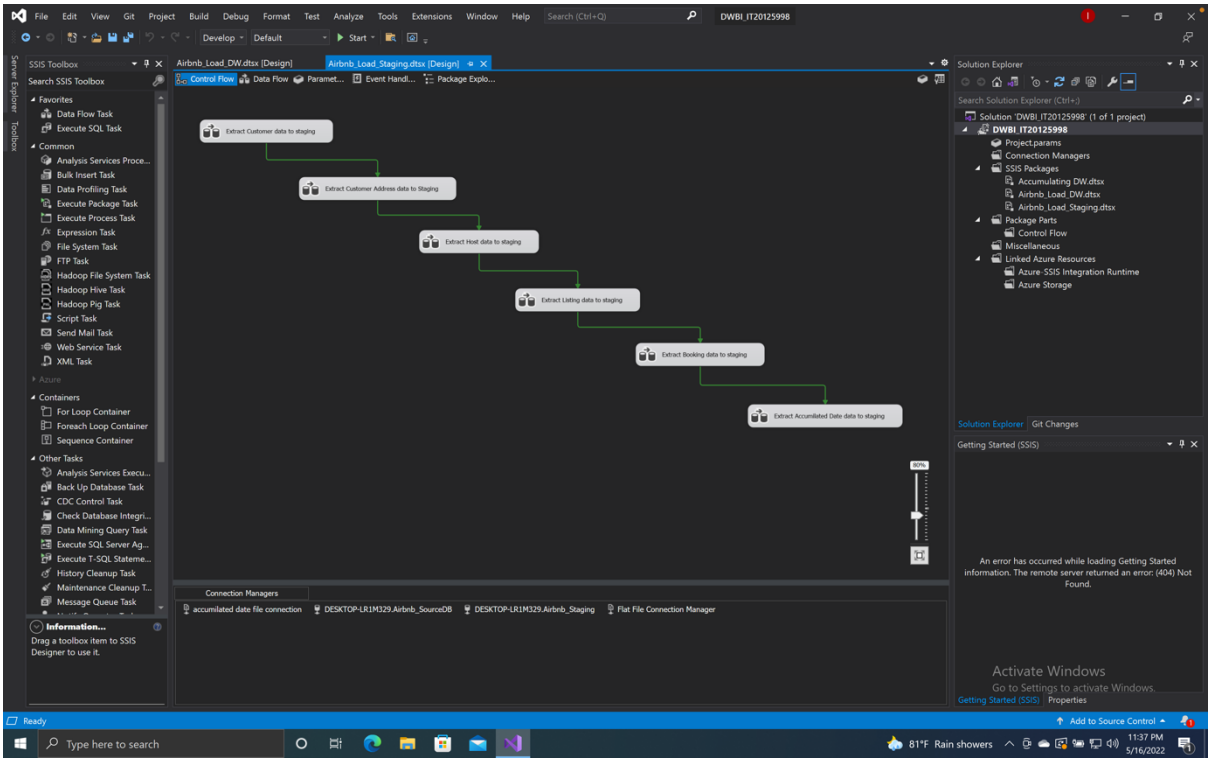
Snowflake Schema

DimCustomer is a slowly changing dimension



ETL Process Screenshots

Extraction from Source database and flat files to staging area.



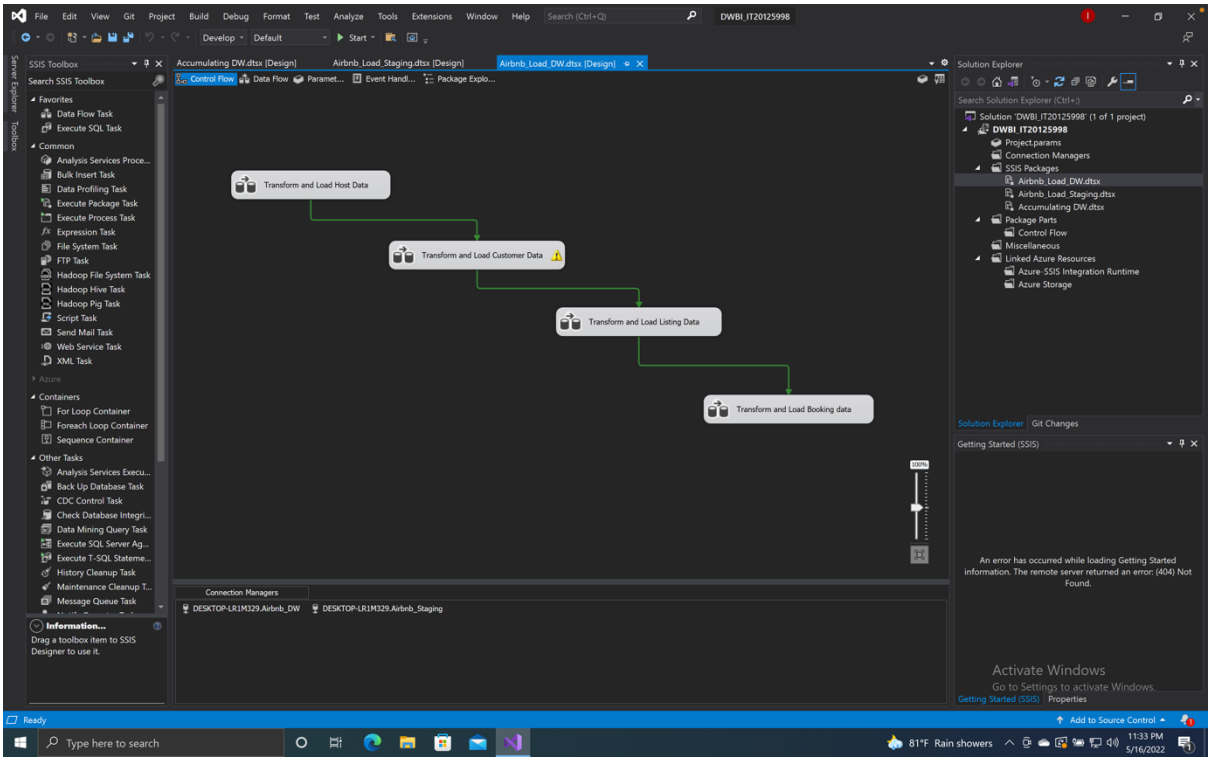
Procedures for updating DimHost table

```
CREATE PROCEDURE dbo.UpdateDimHost
@HostID int,
@HostName varchar(MAX),
@HostSince datetime,
@Location varchar(Max),
@ResponseTime varchar(Max),
@Superhost bit,
@Neighbourhood varchar(Max),
@Verified bit
AS
BEGIN
if not exists (select hostSK from dbo.DimHost where AlternateHostID = @HostID)
BEGIN
insert into dbo.DimHost
(AlternateHostID, name, host_since, location, response_time, is_superhost, neighbourhood, verified, InsertDate, ModifiedDate) values
(@HostID, @HostName, @HostSince, @Location, @ResponseTime, @Superhost, @Neighbourhood, @Verified, GETDATE(), GETDATE())
END;
if exists (select hostSK from dbo.DimHost where AlternateHostID = @HostID)
BEGIN
update dbo.DimHost
set name = @HostName,
    host_since = @HostSince,
    location = @Location,
    response_time = @ResponseTime,
    is_superhost = @Superhost,
    neighbourhood = @Neighbourhood,
    verified = @Verified,
    ModifiedDate = GETDATE()
where AlternateHostID = @HostID
END;
END;
```

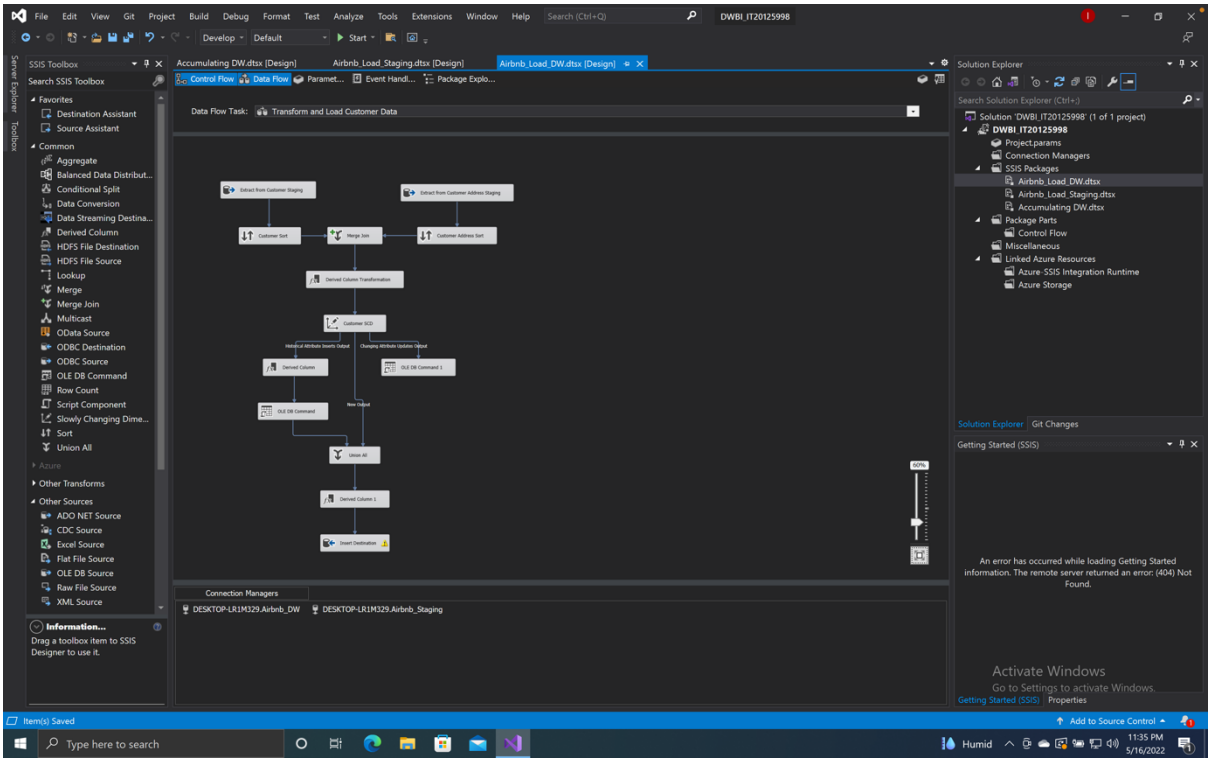
Procedure for updating DimListing table

```sql
CREATE PROCEDURE dbo.UpdateDimListing
@ListingID int,
@Name varchar(100),
@HostKey int,
@Location varchar(100),
@Neighbourhood varchar(50),
@Zipcode varchar(50),
@Latitude varchar(15),
@Longitude varchar(50),
@PropertyType varchar(50),
@RoomType varchar(50),
@Accomadates int,
@Bathrooms float,
@Bedrooms int,
@Beds int,
@Price money,
@MinimumNights int,
@Review int,
@Cancellation varchar(50)
AS
BEGIN
if not exists (select listingSK from dbo.DimListing where AlternateListingID = @ListingID)
BEGIN
insert into dbo.DimListing (AlternateListingID, [name], hostkey, [location], neighbourhood, zipcode, latitude, longitude, property_type, room_type,
                            accomadates, bathrooms, bedrooms, beds, price, minimum_nights, review_scores_rating, cancellation_policy, InsertDate, ModifiedDate)
values(@ListingID, @Name, @HostKey, @Location, @Neighbourhood, @Zipcode, @Latitude, @Longitude, @PropertyType, @RoomType,
       @Accomadates, @Bathrooms, @Bedrooms, @Beds, @Price, @MinimumNights, @Review, @Cancellation, GETDATE(), GETDATE())
END;
if exists (select listingSK from dbo.DimListing where AlternateListingID = @ListingID)
BEGIN
update dbo.DimListing
set [name] = @Name,
    hostkey = @HostKey,
    [location] = @Location,
    neighbourhood = @Neighbourhood,
    zipcode = @Zipcode,
    latitude = @Latitude,
    longitude = @Longitude,
    property_type = @PropertyType,
    room_type = @RoomType,
    accomadates = @Accomadates,
    bathrooms = @Bathrooms,
    bedrooms = @Bedrooms,
    beds = @Beds,
    price = @Price,
    minimum_nights = @MinimumNights,
    review_scores_rating = @Review,
    cancellation_policy = @Cancellation,
    ModifiedDate = GETDATE()
where AlternateListingID = @ListingID
END;
END;
```

Transform and load data to Data Warehousing tables.



Transform and load slowly changing dimension data to data warehouse.



Transform and load accumulating timestamp data to data warehouse.