StayMore

Data Warehouse Methodology

Business Intelligence & Analytics
Project Future 5
C1: Team 7

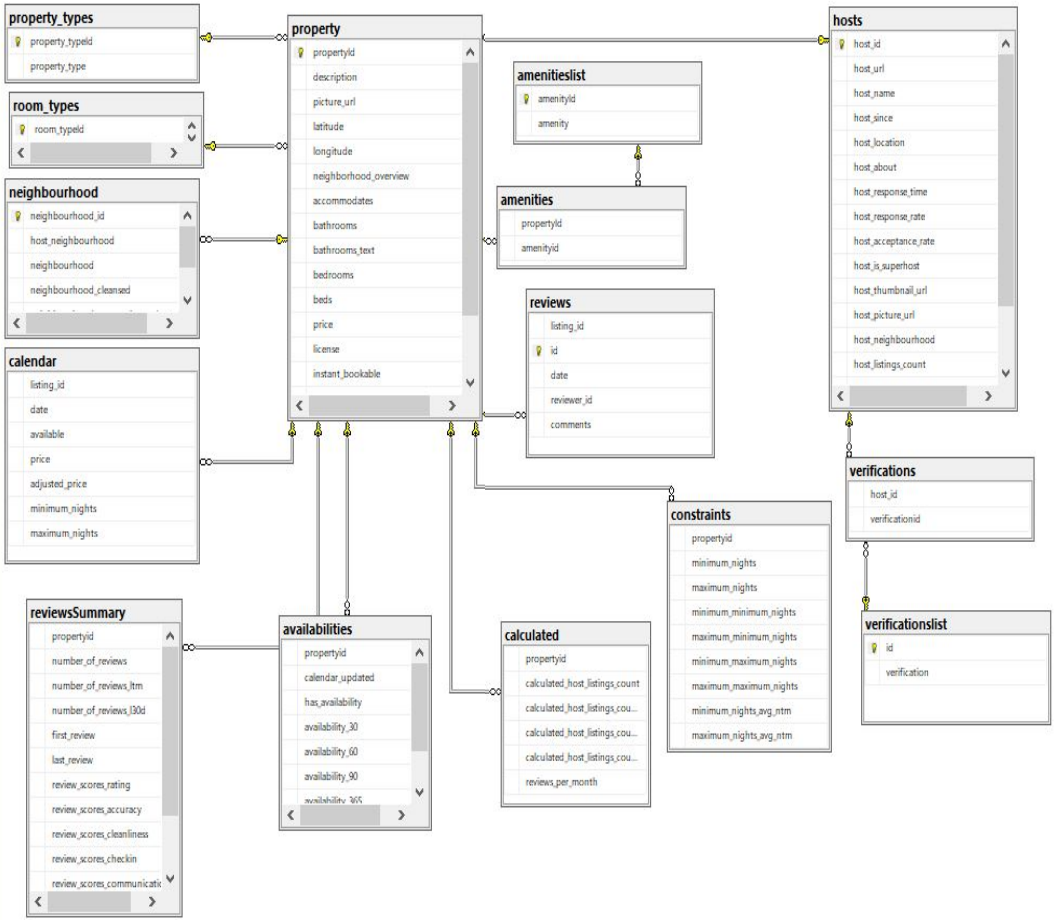
Business Needs

1. Provide **Booking** reports for:
    a. Price
    b. Reviews
    c. Hosts
    d. Properties (value and amenities)
    e. Dates

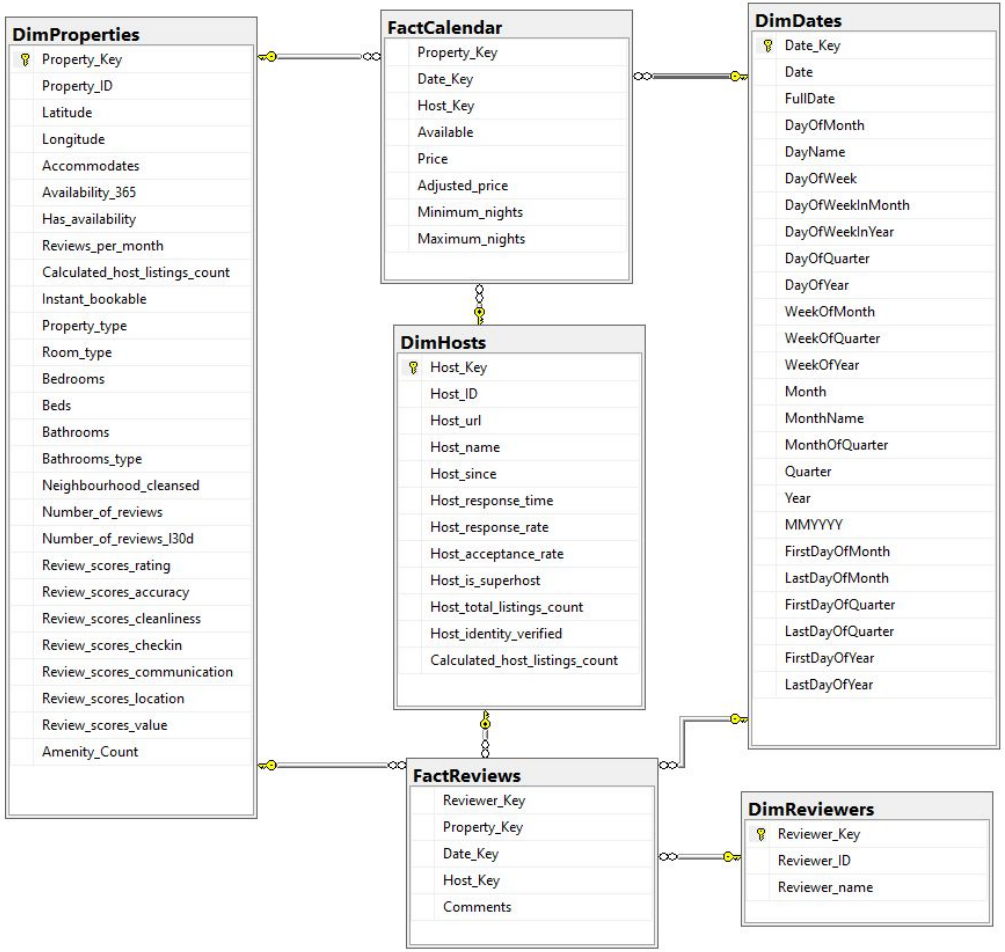
Introduction

- We begin with a given OLTP database and we have to build the Data Warehouse.
- To reach our goal, we created a staging environment.
- Included ETL using statements on MSSQL.
- This way we built StayMore Data Warehouse
- Data marts will be accessible by business users to query

OLTP Schema


property_types
property_typeId
property_type
room_types
room_typeId
neighbourhood
neighbourhood_id
host_neighbourhood
neighbourhood
neighbourhood_cleansed
calendar
listing_id
date
available
price
adjusted_price
minimum_nights
maximum_nights
reviewsSummary
propertyid
number_of_reviews
number_of_reviews_ltm
number_of_reviews_l30d
first_review
last_review
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communicatic
property
propertyId
description
picture_url
latitude
longitude
neighborhood_overview
accommodates
bathrooms
bathrooms_text
bedrooms
beds
price
license
instant_bookable
availabilities
propertyid
calendar_updated
has_availability
availability_30
availability_60
availability_90
amenitieslist
amenityId
amenity
amenities
propertyId
amenityid
reviews
listing_id
id
date
reviewer_id
comments
calculated
propertyid
calculated_host_listings_count
calculated_host_listings_cou...
calculated_host_listings_cou...
calculated_host_listings_cou...
reviews_per_month
constraints
propertyid
minimum_nights
maximum_nights
minimum_minimum_nights
maximum_minimum_nights
minimum_maximum_nights
maximum_maximum_nights
minimum_nights_avg_ntm
maximum_nights_avg_ntm
hosts
host_id
host_url
host_name
host_since
host_location
host_about
host_response_time
host_response_rate
host_acceptance_rate
host_is_superhost
host_thumbnail_url
host_picture_url
host_neighbourhood
host_listings_count
verifications
host_id
verificationid
verificationslist
id
verification

DW Schema


DimProperties
Property_Key
Property_ID
Latitude
Longitude
Accommodates
Availability_365
Has_availability
Reviews_per_month
Calculated_host_listings_count
Instant_bookable
Property_type
Room_type
Bedrooms
Beds
Bathrooms
Bathrooms_type
Neighbourhood_cleansed
Number_of_reviews
Number_of_reviews_l30d
Review_scores_rating
Review_scores_accuracy
Review_scores_cleanliness
Review_scores_checkin
Review_scores_communication
Review_scores_location
Review_scores_value
Amenity_Count
FactCalendar
Property_Key
Date_Key
Host_Key
Available
Price
Adjusted_price
Minimum_nights
Maximum_nights
DimDates
Date_Key
Date
FullDate
DayOfMonth
DayName
DayOfWeek
DayOfWeekInMonth
DayOfWeekInYear
DayOfQuarter
DayOfYear
WeekOfMonth
WeekOfQuarter
WeekOfYear
Month
MonthName
MonthOfQuarter
Quarter
Year
MMYYYY
FirstDayOfMonth
LastDayOfMonth
FirstDayOfQuarter
LastDayOfQuarter
FirstDayOfYear
LastDayOfYear
DimHosts
Host_Key
Host_ID
Host_url
Host_name
Host_since
Host_response_time
Host_response_rate
Host_acceptance_rate
Host_is_superhost
Host_total_listings_count
Host_identity_verified
Calculated_host_listings_count
FactReviews
Reviewer_Key
Property_Key
Date_Key
Host_Key
Comments
DimReviewers
Reviewer_Key
Reviewer_ID
Reviewer_name

Steps

- Creation of an OLTP database (air2)
- 1NF on the OLTP database
- Creation of the staging environment
- Data loading in the staging database (air2_staging)
- Creation of the Data Warehouse database (StayMore)
- Data loading from the staging database to the Data Warehouse
- Creation of Calendar and Reviews Data Marts

OLTP

Firstly, we need to create the air2 database, which includes the complete data source. This database needs to be created in the SQL Server Instance.

1NF

From the OLTP database, reviewers have been separated as an instance from reviews.

Additionally, they are connected with reviewer_id and id.

Script is provided.

Staging - Overview

- For the staging area, we have created a new database (air2_staging).
- All data, which are essential for our project development, will be transferred from the OLTP to staging area database air2_staging
- Specifically, air2_staging will include :
  - Reviews
  - Hosts
  - Properties
  - Reviewers
  - Calendar
- Moreover, we need a dimension named Date to keep track of time

Script is provided.

Staging - Details

In particular, three of these tables (Reviews, Reviewers and Calendar) will be formed as following:

Reviews

- listing_id
- date
- reviewer_id
- host_id
- comments

Reviewers

- reviewer_id
- fullname

Calendar

- listing_id
- date
- host_id
- available
- price
- adjusted_price
- minimum_nights
- maximum_nights

Staging - Details

The Hosts and Properties tables will be formed like this:

Hosts

- host_id
- host_url
- host_name
- host_since
- host_response_time
- host_response_rate
- host_acceptance_rate
- host_is_superhost
- host_total_listings_count
- host_identity_verified
- calculated_host_listings_count

Properties

- propertyId
- latitude
- longitude
- accommodates
- availability_365
- has_availability
- reviews_per_month
- calculated_host_listings_count
- instant_bookable
- property_type
- room_type
- bedrooms
- beds
- bathrooms_text
- neighbourhood_cleansed
- number_of_reviews
- number_of_reviews_l30d
- review_scores_rating
- review_scores_accuracy
- review_scores_cleanliness
- review_scores_checkin
- review_scores_communication
- review_scores_location
- review_scores_value
- AmenityCount

Staging - Details

Lastly, the new Date will also be included in the staging database.

Script is provided.

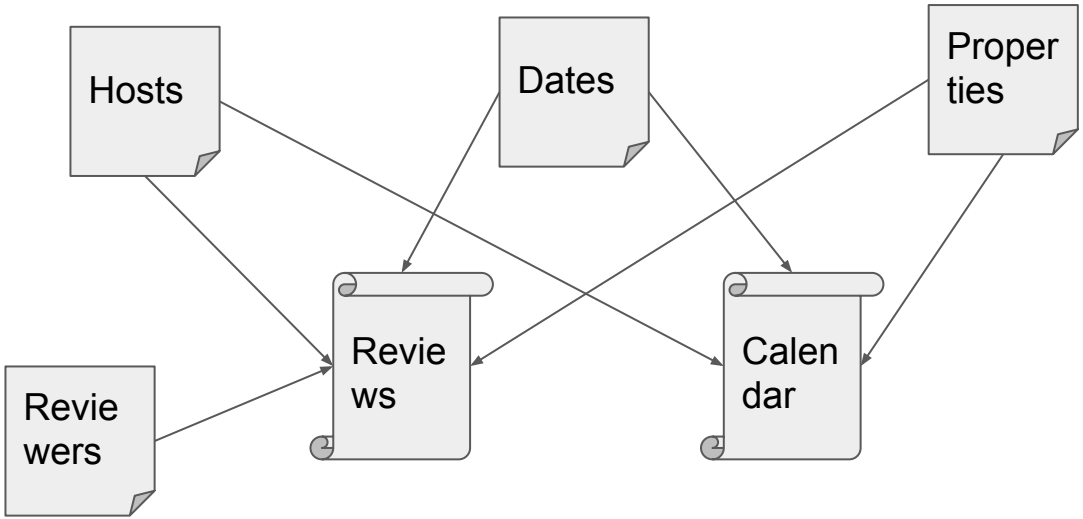
Inmon’s Corporate Information Factory

Staymore adopted the Inmon’s Corporate Information Factory architecture. Its business needs focus on:

- Subject-oriented, integrated, time-variant and nonvolatile repository or operational data
- Management and decision-making process

Data Warehouse Data Representation: Constellation Schema

In order to address the business needs, we built a constellation Schema for the Data Warehouse. This schema includes two Fact tables and four dimensional tables, as we can see below:



Data Organization

- Facts / Measures:
  - Calendar
  - Reviews
- Dimensions
  - Hosts
  - Properties
  - Dates
  - Reviewers

- Attributes
  - Full dates calendar
  - Number of reviews
  - Review_scores_rating
  - Room_type
  - Property_type
  - Price
  - Available
  - Reviewer name
  - etc.

Data Warehouse - Overview

- The next and most important step is the Data Warehouse creation.
- StayMore database will be created based on the Fact Constellation Schema.
- All relevant data will be loaded from air2_staging to StayMore.
- StayMore will include the following tables:
  - DimDates
  - DimHosts
  - DimProperties
  - FactCalendar
  - FactReviews

Data Warehouse - Details

DimReviewers dimension will include:

Reviewer_key	PK IDENTITY(1,1) NOT NULL PRIMARY KEY
Reviewer_ID	INT NOT NULL
Reviewer_name	NVARCHAR(50) NOT NULL

Data Warehouse - Details

DimProperties dimension will include:

Property_Key	PK IDENTITY(1,1) NOT NULL PRIMARY KEY	Instant_bookable	INT NOT NULL
Property_ID	INT NOT NULL	Property_type	NVARCHAR(40) NOT NULL
Latitude	FLOAT NOT NULL	Room_type	NVARCHAR(40) NOT NULL
Longitude	FLOAT NOT NULL	Bedrooms	INT NOT NULL
Accommodates	INT NOT NULL	Beds	INT NOT NULL
Availability_365	INT NOT NULL	Bathrooms	FLOAT NOT NULL
Has_availabitity	INT NOT NULL	Bathrooms_type	NVARCHAR(20) NOT NULL
Reviews_per_month	FLOAT NOT NULL	Neighbourhood_cleansed	NVARCHAR(40) NOT NULL
Calculated_host_listings_count	FLOAT NOT NULL	Review_scores_rating	INT NOT NULL

Data Warehouse - Details

DimProperties dimension will include:

Number_of_reviews	INT NOT NULL
Number_of_reviews_l30d	INT NOT NULL
Review_scores_rating	INT NOT NULL
Review_scores_accuracy	FLOAT NOT NULL
Review_scores_cleanliness	FLOAT NOT NULL
Review_scores_checkin	FLOAT NOT NULL
Review_scores_communication	FLOAT NOT NULL
Review_scores_location	FLOAT NOT NULL
Review_scores_value	FLOAT NOT NULL
Amenity_count	INT NOT NULL

Data Warehouse - Details

DimHosts dimension will include:

Host_Key	PK IDENTITY(1,1) NOT NULL PRIMARY KEY
Host_ID	INT NOT NULL
Host_url	NVARCHAR(100) NOT NULL
Host_name	NVARCHAR(100) NOT NULL
Host_since	NVARCHAR(20) NOT NULL
Host_response_time	NVARCHAR(25) NOT NULL
Host_response_rate	FLOAT NOT NULL
Host_acceptance_rate	FLOAT NOT NULL
Host_is_superhost	INT NOT NULL
Host_total_listings_count	INT NOT NULL
Host_identity_verified	INT NOT NULL
Calculated_host_listings_count	INT NOT NULL

Data Warehouse - Details

DimDates dimension will be similar to the Date table from the staging environment.

Data Warehouse - Fact Table

- StayMore data warehouse will include two fact tables named:
  - FactCalendar
  - FactReviews

Data Warehouse - Fact Table

FactCalendar dimension will include:

Property_Key	INT NOT NULL
Date_key	INT NOT NULL
Host_key	INT NOT NULL
Available	INT NOT NULL
Price	FLOAT NOT NULL
Adjusted_price	INT NOT NULL
Minimum_nights	INT NOT NULL
Maximum_nights	INT NOT NULL

Data Warehouse - Fact Table

FactReviews dimension will include:

Reviewer_Key	INT NOT NULL
Property_Key	INT NOT NULL
Date_Key	INT NOT NULL
Host_Key	INT NOT NULL
Comments	NVARCHAR(4000) NOT NULL

Data Warehouse - Fact Table

Additionally, StayMore will observe the following FK constraints:

FactCalendar_DimProperties_Property_Key_fk	Property_key → DimProperties(Property_Key)
FactCalendar_DimHosts_Host_key_fk	Host_Key → DimHosts(Host_key)
FactCalendar_DimDates_Date_ID_fk	Date_Key → DimDates(Date_key)
FactReviews_DimReviewers_Reviewer_Key_fk	Reviewer_Key → DimReviewers(Reviewer_Key)
FactReviews_DimProperties_Property_Key_fk	Property_Key → DimProperties(Property_Key)
FactReviews_DimDate_Date_ID_fk	Date_Key → DimDates(Date_key)
FactReviews_DimHosts_Host_Key_fk	Host_Key → DimHosts(Host_Key)

Data Integration

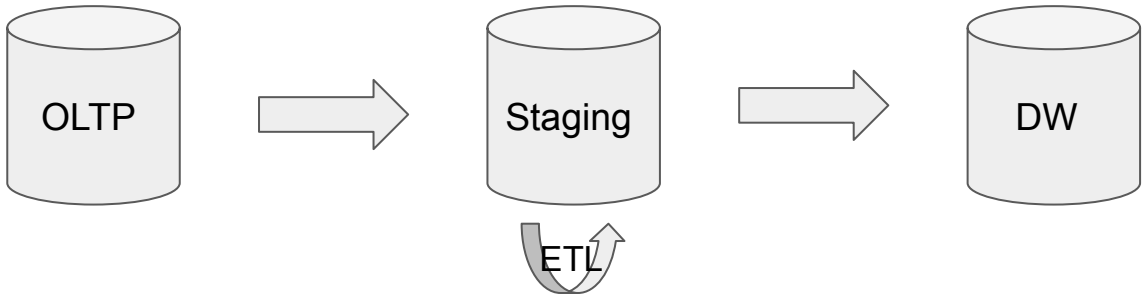
A key factor for an integrated Data Warehouse is Data Integration of all data sources that each business has in disposal. These data need to be cohesive and consolidated for the user.

The main functions/steps to achieve our goal are the following:

- Extraction
  - Retrieve data in a source system
  - Get these data as efficiently as possible
- Transformation
  - Inspect, clean and perform dimension conforming
  - Perform data calculations, if needed
- Loading
  - Load the data in the main Storage
  - Update the Warehouse

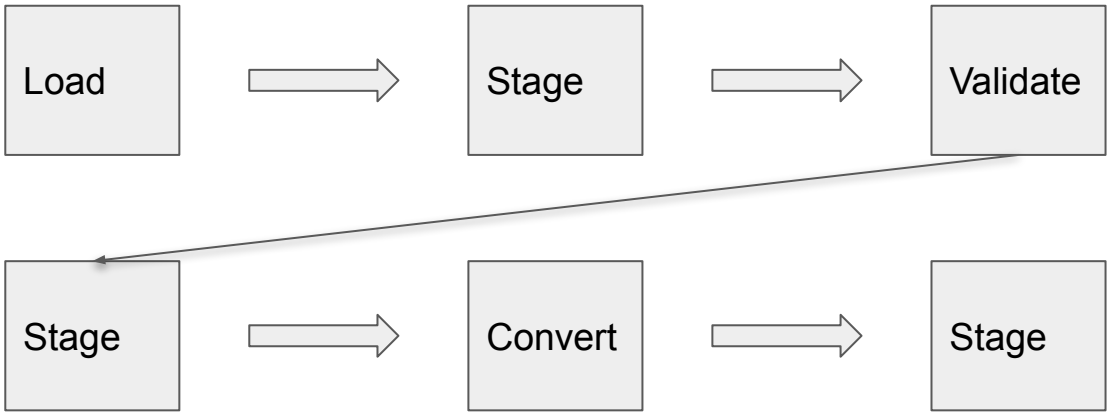
ETL Architecture

A Two-Stage process was used to create our system, as seen below. A given OLTP database was also used for the Staging environment creation and the DW was built under ETL processes.



Data Transformation

- Remove inconsistencies and redundancies.
- Conform to the Data Warehouse conceptual schema.



Details on Transform & Load (1)

- As we noted before, on tables' creation data will be loaded from the staging database to StayMore Data Warehouse.

- As a general rule, we have defined as NOT NULL all attributes in the Data Warehouse's tables.

  To ensure we don't load any NULL data values, we had to follow a set of rules for transforming NULL data before loading them in the Data Warehouse:

  - Float or int data with NULL values were transformed to 0
  - Categorical (eg nvarchar) data with NULL values were transformed to 'N/A'

Details on Transform & Load (2)

- Here are some examples of the transformations performed:
  - **Bathroom_text** has been splitted into two columns/attributes by utilizing the case SQL statement and using regular expressions:
    - Bathrooms (float not null)
    - Bathrooms_type (nvarchar(20) not null)
      - We have created 2 distinct categories “private bath(s)” and “shared bath(s)” (also any NULL values transformed to “N/A”)
  - Review table’s **Comments** attribute that surpass the maximum character number (4,000) set for the attribute in the DW are casted to nvarchar(4,000) and when they are NULL, then they should be set to ‘N/A’
  - **host_response_time** is a string and we transform it to ‘N/A’ when NULL.
  - **host_is_superhost** and **instant_bookable** attributes are transformed to int type. If the data contains a **t** value, then it’s transformed to 1, else 0.
  - All the transformations are visible in the sql file that creates and loads the data in the DW.

Data Marts

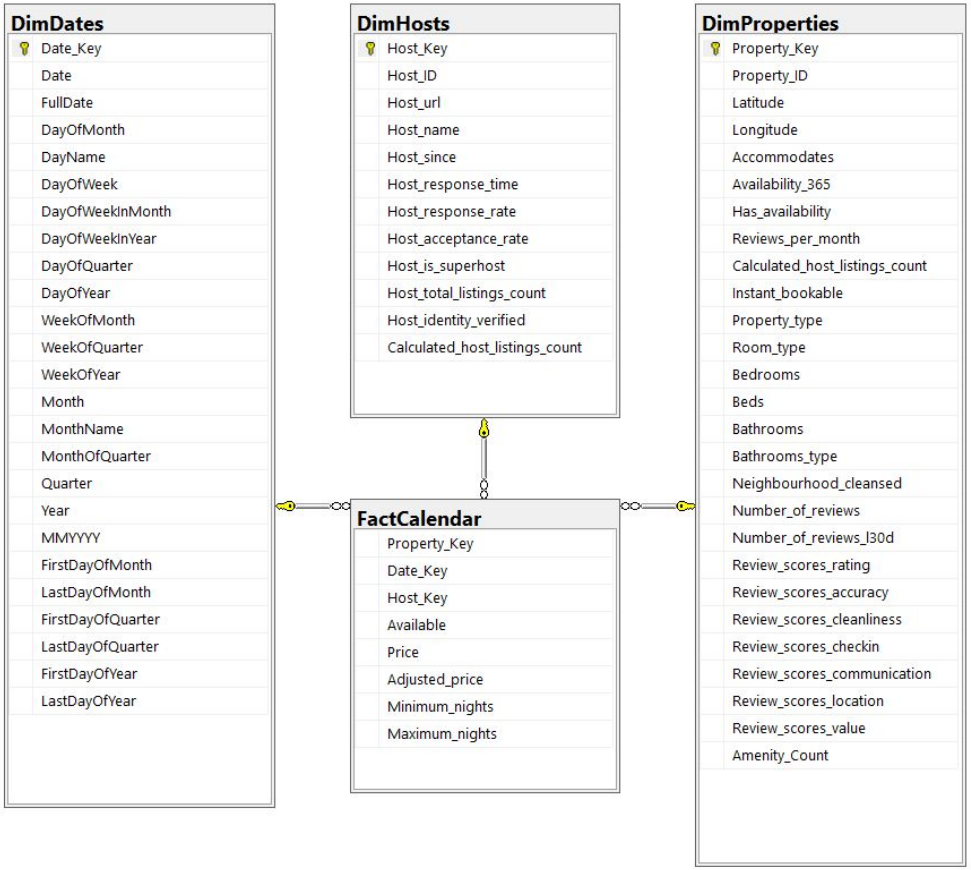
Data Marts consist a subset of our data warehouse with purpose to support the requirements of different business departments.

We can think Data marts as a small-scale Data Warehouse which stores specific data for a department. In addition, both ‘CalendarDW’ and ‘ReviewsDW’ Marts are able to support two different business needs.

Data Marts

In conclusion, a CalendarMart and a ReviewsMart view will be created from the Data Warehouse as in the next slide:

Calendar Schema



Reviews Schema