

Sri Lanka Institute of Information Technology

**Data Warehousing and Business Intelligence**

**(IT3021)**

Assignment 1

**Hotel Reviews**

Submitted by :

**IT19374666**

**Dissanayake Adikaramge D.M.**

1. **Data Selection**

I selected a hotel review data set for the assignment which includes the ratings by the reviewers, number of reviews seen of a review by the reviewees, the commission paid to the reviewer, the relevant hotel details and the reviewer details.

This hotel review data set comprises of 10,000 records of hotel reviews over 16 years from 2002 – 2018.

The data set has hierarchies in hotel location such as country-> province -> city and in reviewer entity has user\_province -> user\_city hierarchy.

**Following is the ER Diagram for the chosen data set.**

Diagram

Description automatically generated

Data set was downloaded from the following link :

<https://www.kaggle.com/datafiniti/hotel-reviews>

1. **Preparation of data sources**

From the provided link above, I received a hotel review details data set in a csv file format.

The tables are review, review details, hotel , hotel location and reviewer.

Thus, I separated them into different tables in different source types.

They are as follows;

* Separated hotel location details mainly including address, city, postal code into a **text file**.
* Hotel details were separated into **csv file** including hotel name and primary category.
* Review table mainly including rating, commission, review dates was converted into a **excel file**.
* Reviewer details including username and review details mainly including review content were separated into **database tables**.

In each table I included a primary key . Furthermore, in review table I added foreign keys for review details ,hotel and reviewer tables and also in the hotel table added foreign key reference for the hotel location table.

1. **Solution Architecture** Diagram

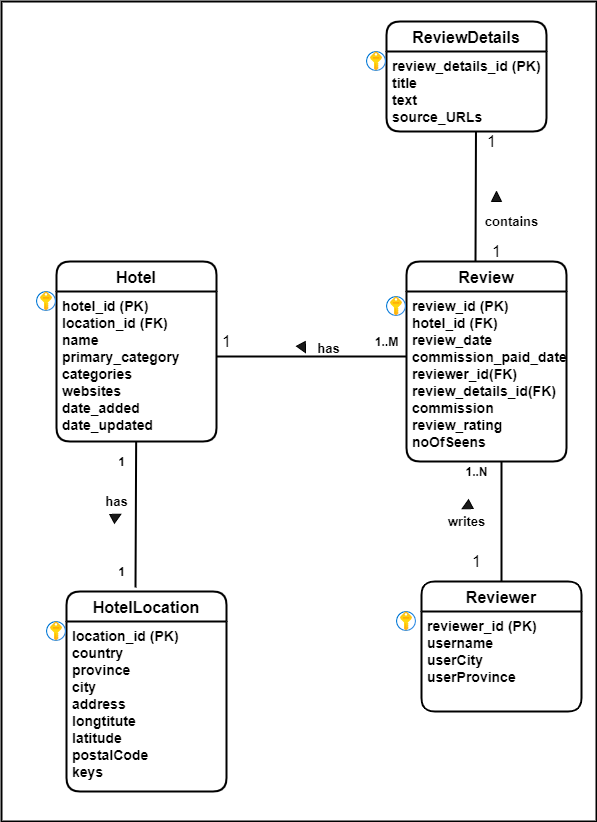
   Description automatically generated

First stage is represents the sources to the data warehouse which are in different format such as csv,excel, text and database files. Then I extracted the data from those sources and without performing any transactions load them into tables in staging database .

Then I extracted data from the staging tables and performed some transformations such as cleaning and adding derived columns and load the data to relevant dimensions and into a fact table (Fact Review) in data warehouse.

1. **Data warehouse design & development**

Below diagram is a **relational model** for hotel review data set;



**Dimensional Model**

**Diagram

Description automatically generated**

For the data warehouse of the review data set ,I implemented a star schema. DimHotel, DimReviewer , DimReviewDetails and DimDate are dimensions and Review is the fact table in the data warehouse.

I merged the hotel and hotel location tables as shown in the relational model above together to create a single dimension table named DimHotel as shown in the above dimensional model to develop it to a star schema.

Further I implemented the DimHotel dimension as a **slowly changing dimension**.

**Grain** : Details of a review on a specific hotel by a reviewer .

**Assumptions:**

I decided hotel dimension is a slowly changing dimension assuming hotel name, categories ,primary categories and location details can be changed over time.

1. **ETL development**
2. Extracting data from the source and load it to the staging. For each stage I truncate the staging tables to clean the database before loading to avoid duplication in the loading data.

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

Description automatically generated

1. Extracting hotel location details from a text file and loading.

Diagram

Description automatically generated

1. Extracting hotel details from a csv file and loading.

A screenshot of a graph

Description automatically generated with medium confidence

1. Extracting reviewer details from DB table and loading.

A screenshot of a computer

Description automatically generated with medium confidence

1. Extracting review details from DB table and loading.

A picture containing text, businesscard, stationary, envelope

Description automatically generated

1. Extracting review from an excel file and loading.

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

Description automatically generated

1. Perform a data profiling task to analyze source data including composite keys in the tables ,null value percentage and maximum and minimum length of a data in a column.

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

Description automatically generated

c)Extract .Transform and load data to data warehouse.

Diagram

Description automatically generated with low confidence

1. Extract,Transform and load Reviewer details to DimReviewer.

A picture containing text, businesscard

Description automatically generated

As the OLE DB command I executed the following stored procedure to insert data.

CREATE PROCEDURE dbo.UpdateDimReviewer

@reviewerID nvarchar(255),

@username nvarchar(255),

@userProvince nvarchar(255),

@userCity nvarchar(255)

AS

BEGIN

if not exists (select SK\_Reviewer from dbo.DimReviewer where [reviewer\_id] = @reviewerID )

BEGIN

insert into dbo.DimProduct ([reviewer\_id], [reviews#username],[reviews#userProvince] ,

[reviews#userCity])

values(@reviewerID, @username, @userProvince, @userCity)

END;

if exists (select SK\_Reviewer from dbo.DimReviewer where [reviewer\_id] = @reviewerID )

BEGIN

update dbo.DimReviewer

set [reviews#username] = @username,[reviews#userProvince] = @userProvince, [reviews#userCity]=@userCity

where [reviewer\_id] = @reviewerID

END;

END

1. Extract,transform and load Review Details to DimReviewDetails.

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

Description automatically generated

As the OLE DB command ,I executed the following stored procedure to insert data.

CREATE PROCEDURE dbo.UpdateDimReviewDetails

@sourceURLs nvarchar(255),

@text nvarchar(max),

@title nvarchar(255),

@detailId nvarchar(255)

AS

BEGIN

if not exists (select SK\_ReviewDetails from dbo.DimReviewDetails where [Detail\_id] = @detailId )

BEGIN

insert into dbo.DimReviewDetails ([reviews#sourceURLs], [reviews#text],[reviews#title],[Detail\_id])

values(@sourceURLs,@text,@title ,@detailId)

END;

if exists (select SK\_ReviewDetails from dbo.DimReviewDetails where [Detail\_id] = @detailId )

BEGIN

update dbo.DimReviewDetails

set [reviews#sourceURLs] =@sourceURLs,[reviews#text] =@text, [reviews#title]=@title

where [Detail\_id] = @detailId

END;

END

1. Extract,transform and load Hotel details to SCD DimHotel.

Diagram

Description automatically generated

First, I merged hotel and hotel location details after sorting the both tables by hotel\_location\_id.Then I used a derived column to convert null values in the postal Code to ‘NA’.Then I developed the whole Hotel dimension as a slowly changing dimension and load the data to DimHotel dimension.I maintained the slowly changing hotel dimension attributes under following types;

Type 1(Changing) - primary categories, categories, websites, keys

Type 2 (Historical) – Hotel name, address

1. Extract .transform and load Review to FactReview.

Timeline

Description automatically generated

**Error Handling**

* I got an error when using lookup for reviewer when loading data to FactReview as shown below and it was due to an unmatching output.

Diagram

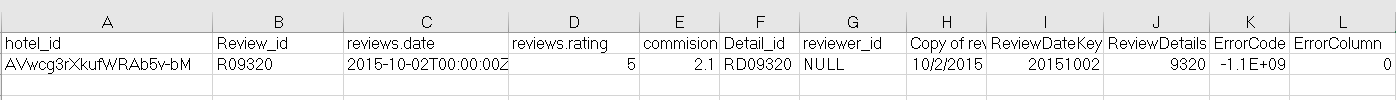
Description automatically generated

* Then I handled this using the error output rows redirecting option as shown below.

Timeline

Description automatically generated with medium confidence

* There was an unmatching output excluded in the excel destination as below.



* I found the reviewer\_id and the SK\_Reviewer from the following Query.

Graphical user interface, text, application, Word

Description automatically generated

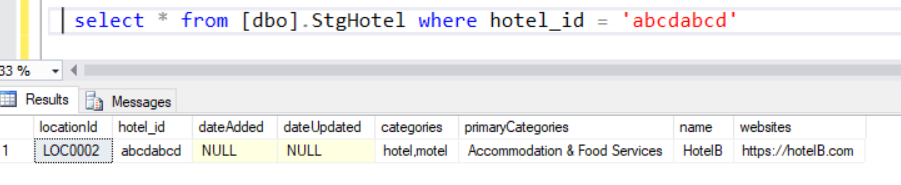
* Then updated record was inserted to the data warehouse FactReview.

Diagram

Description automatically generated

1. **Testing Methodology** 
   1. Testing the Slowly Changing Dimension Hotel – Type 2 attribute hotel name.

* Initially I inserted a new record into the Staging Hotel with the hotel\_id = ‘abcdabcd’ and load it to SCD Hotel.



* Then I updated the hotel name from ‘HotelB’ to ‘HotelB Updated’ in the StgHotel and inserted modified record into the Hotel Dimension.
* Selected all the recorded from the DimHotel where alternate\_hotel\_id = ‘abcdabcd’.

**Graphical user interface

Description automatically generated with low confidence**

* Status of the outdated record was set to ‘**Expired’** and new record was set to ‘**Current’.**
  1. Testing the Slowly Changing Dimension Hotel – Type 1 attribute hotel categories and primary categories.
* Change the above mentioned record from primecategory ‘Accomodation and food services’ to ‘Food Services’ and categoties from ‘hotel motel’ to ‘hotel’.

Graphical user interface, text, application

Description automatically generated

* Then inserted the record to DimHotel and check the values of the alternate\_hotel\_id = ‘abcdabcd’.

Graphical user interface, application

Description automatically generated

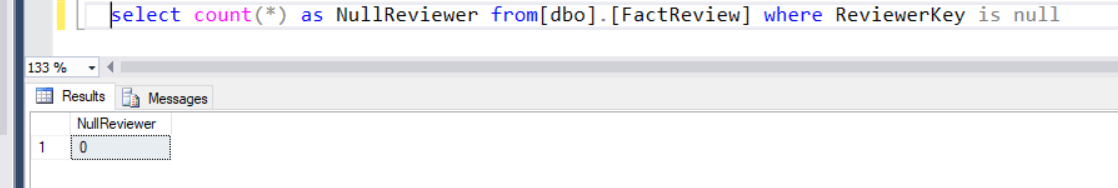
* Only the updated record was there under the current record.
  1. Testing null values in the fact table after look up.

|  |  |  |  |
| --- | --- | --- | --- |
| Test | Expected OutPut | Actual OutPut | Status |
| Test null Review Date Key columns. | 0 | 0  (Refer Attachment 6 1) | Pass |
| Test null Reviewer Key columns. | 0 | 0  (Refer Attachment 6 2) | Pass |
| Test null Hotel Key colunms. | 0 | 0  (Refer Attachment 6 3) | Pass |
| Test null Review Details Key colunms. | 0 | 0  (Refer Attachment 6 4) | Pass |
| Test null Commision Paid Date key colunms. | 0 | 0  (Refer Attachment 6 5) | Pass |

A picture containing timeline

Description automatically generated

Attachment 6 1



Attachment 6 2

Graphical user interface, text, application

Description automatically generated

Attachment 6 3

Text

Description automatically generated with medium confidence

Attachment 6 4

A picture containing table

Description automatically generated

Attachment 6 5

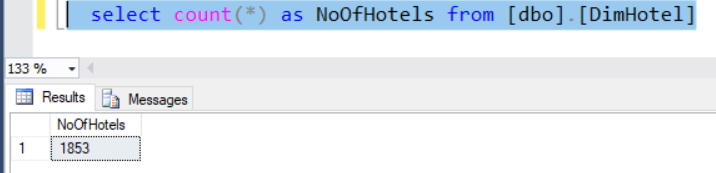
* 1. Test initial number of records in staging with relevant dimension and fact tables in data warehouse.

|  |  |  |  |
| --- | --- | --- | --- |
| Test | Expected OutPut | Actual OutPut | Status |
| Test number of records in DimReviewer | 7137 | 7137  (Refer Attachment 6 6) | Pass |
| Test number of records in DimHotel | 1853 | 1853  (Refer Attachment 6 7) | Pass |
| Test number of records in DimReviewDetails | 10000 | 10000  (Refer Attachment 6 8) | Pass |
| Test number of records in FactReview | 10000 | 10000  (Refer Attachment 6 9) | Pass |

Table

Description automatically generated with medium confidence

Attachment 6 6



Attachment 6 7

Table

Description automatically generated with medium confidence

Attachment 6 8

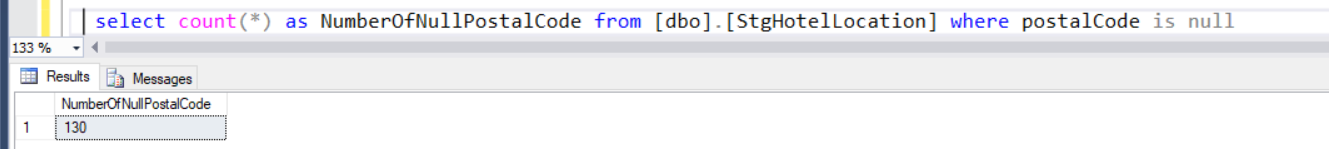
Table

Description automatically generated

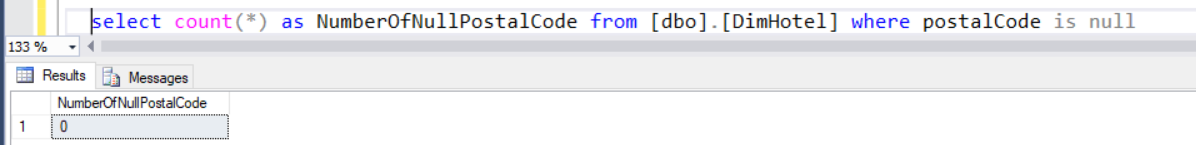
Attachment 6 9

* 1. Test the null values cleansing for the postal code.

Input :



Output :



* 1. Test the dimensions and fact table retrieving 10 records per table.
* Retrieve top 10 records from Fact Review

Graphical user interface, table

Description automatically generated

* Retrieve top 10 records from DimReviewer

Graphical user interface, table

Description automatically generated with medium confidence

* Retrieve top 10 records from DimReviewDetails.

Graphical user interface, text, application

Description automatically generated

* Retrieve top 10 records from DimHotel

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated