Sri Lanka Institute of Information Technology

Data Warehousing and Business Intelligence

(IT3021)

Assignment 1

Hotel Reviews

Submitted by :

IT19374666

Dissanayake Adikaramge D.M.

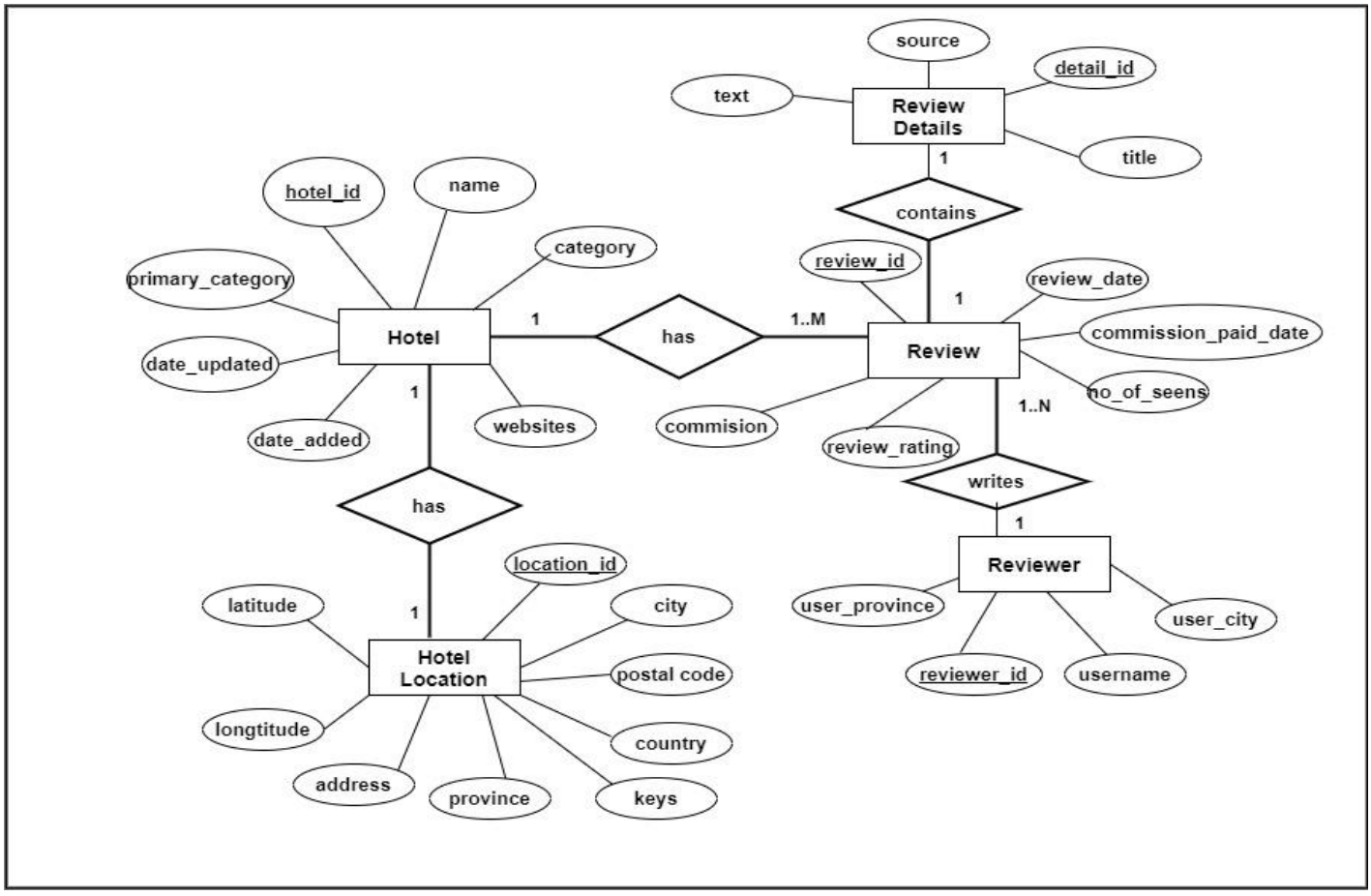
1. Data Selection

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.



Data set was downloaded from the following link :

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

2. 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.

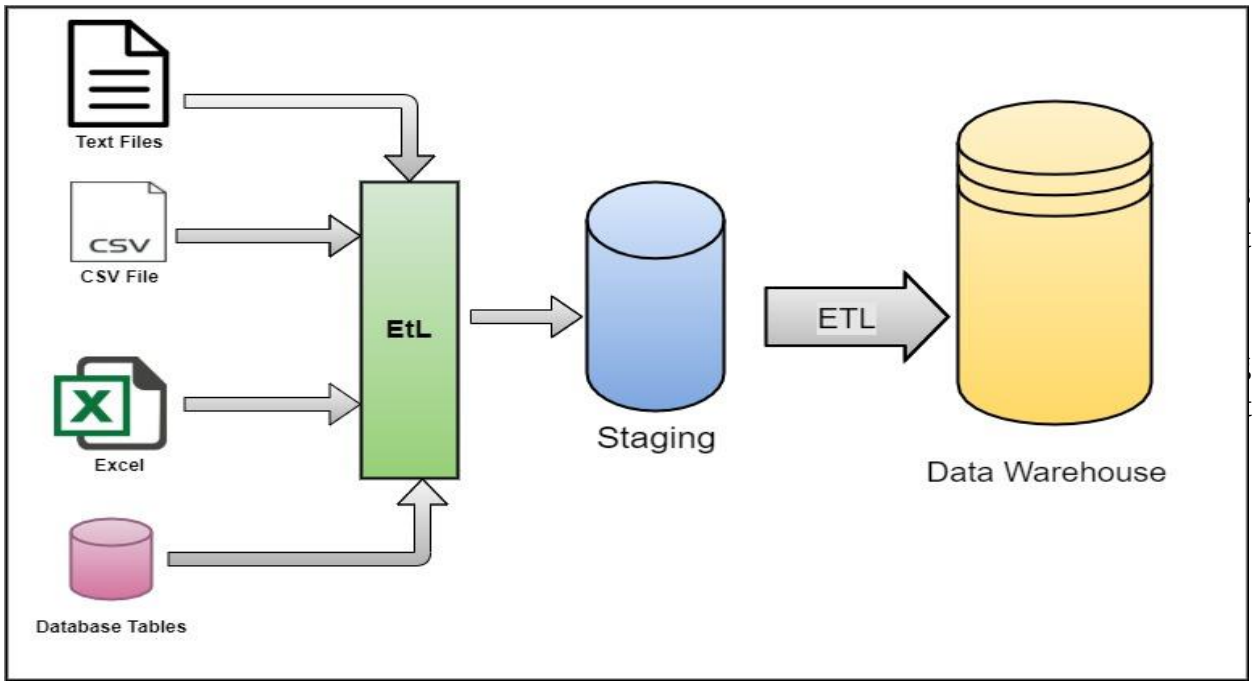
Thus, I separated them into different tables in different source types.

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**.

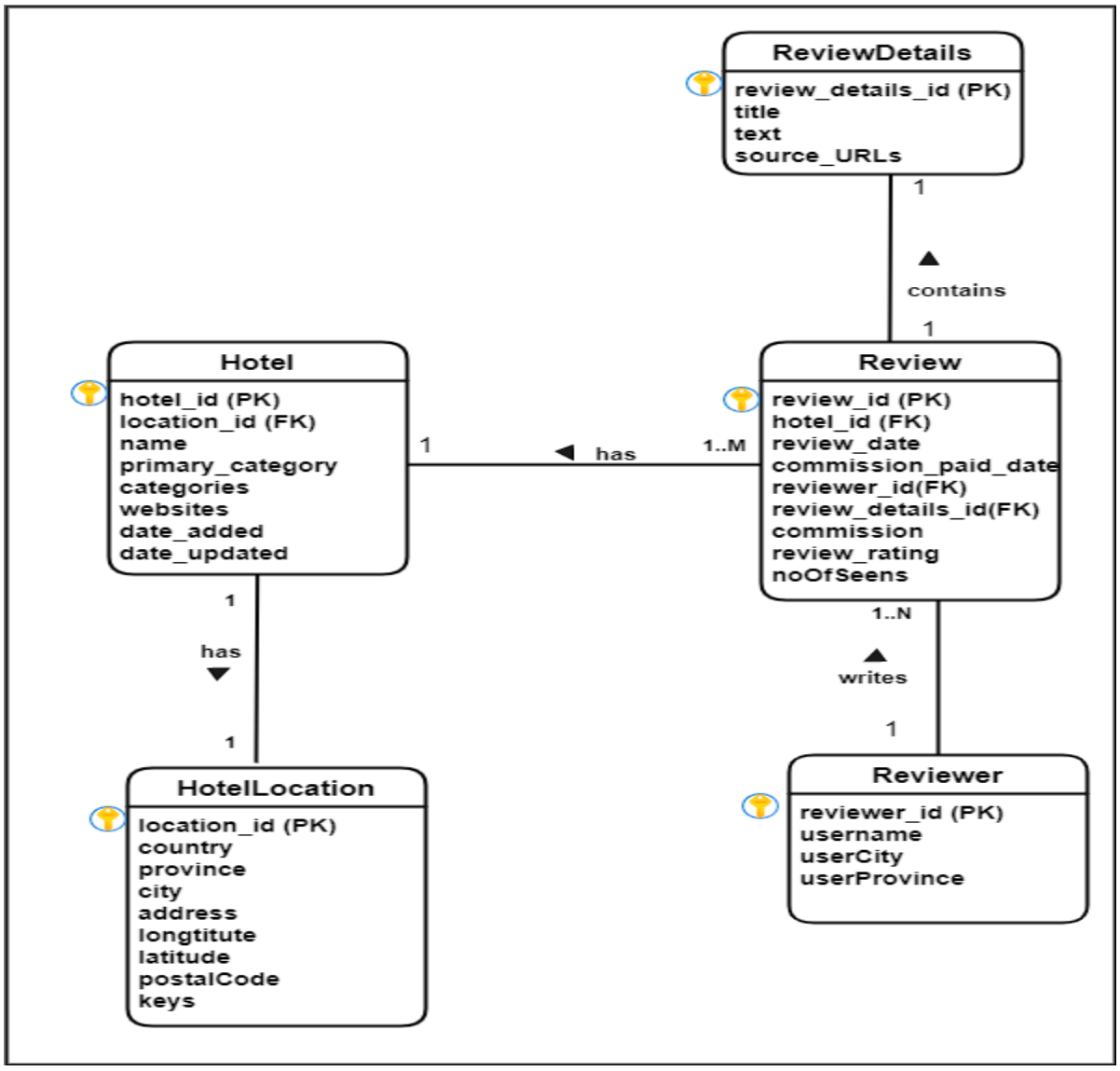
In each table I included a primary key . Furthermore, in review table I added foreign keys for review details ,hotel and reviewer tables and also in the hotel table added foreign key reference for the hotel location table.

3. Solution Architecture

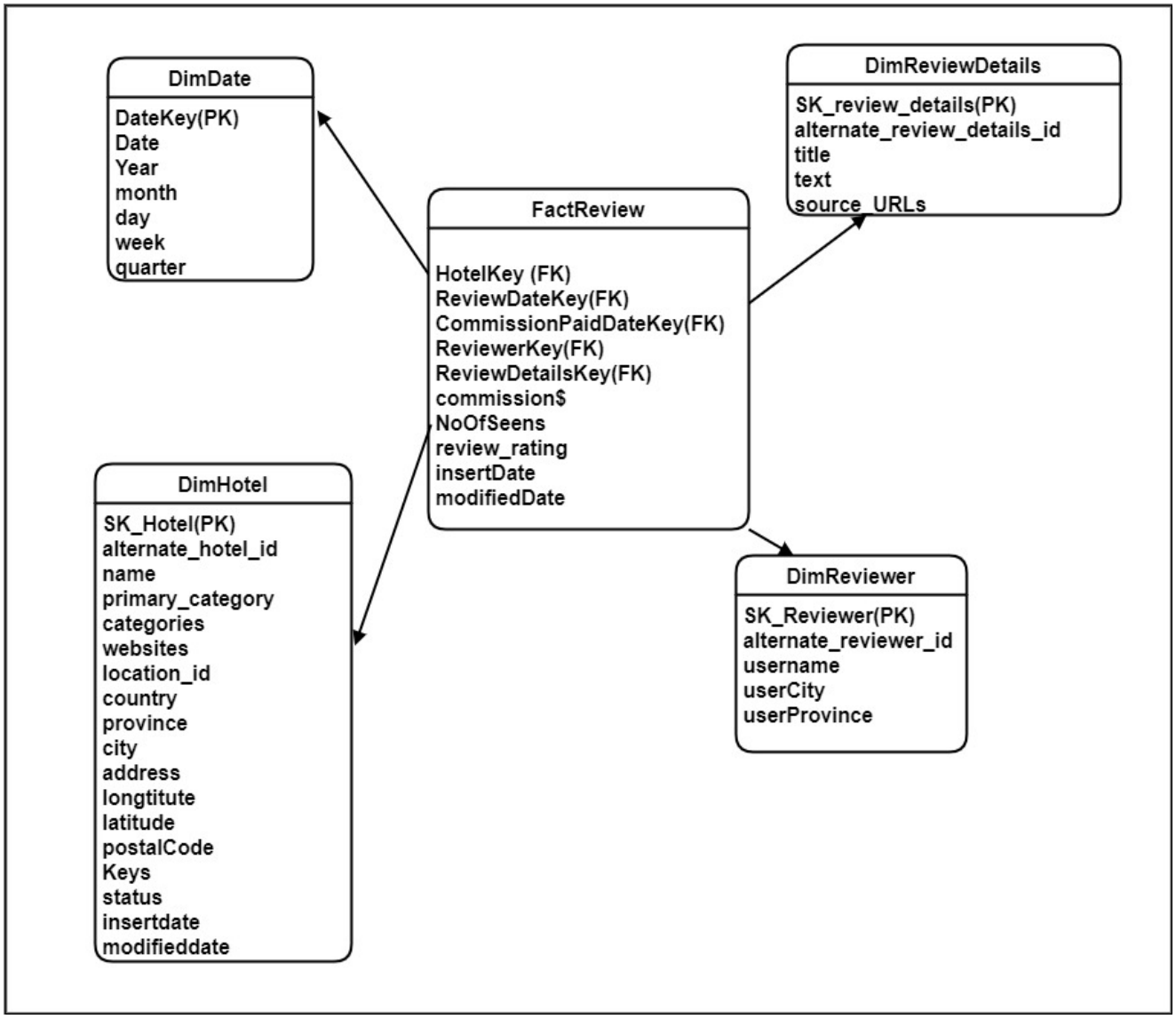


4. Data warehouse design & development

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



Dimensional Model



For the data warehouse of the review data set ,I implemented a 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**.

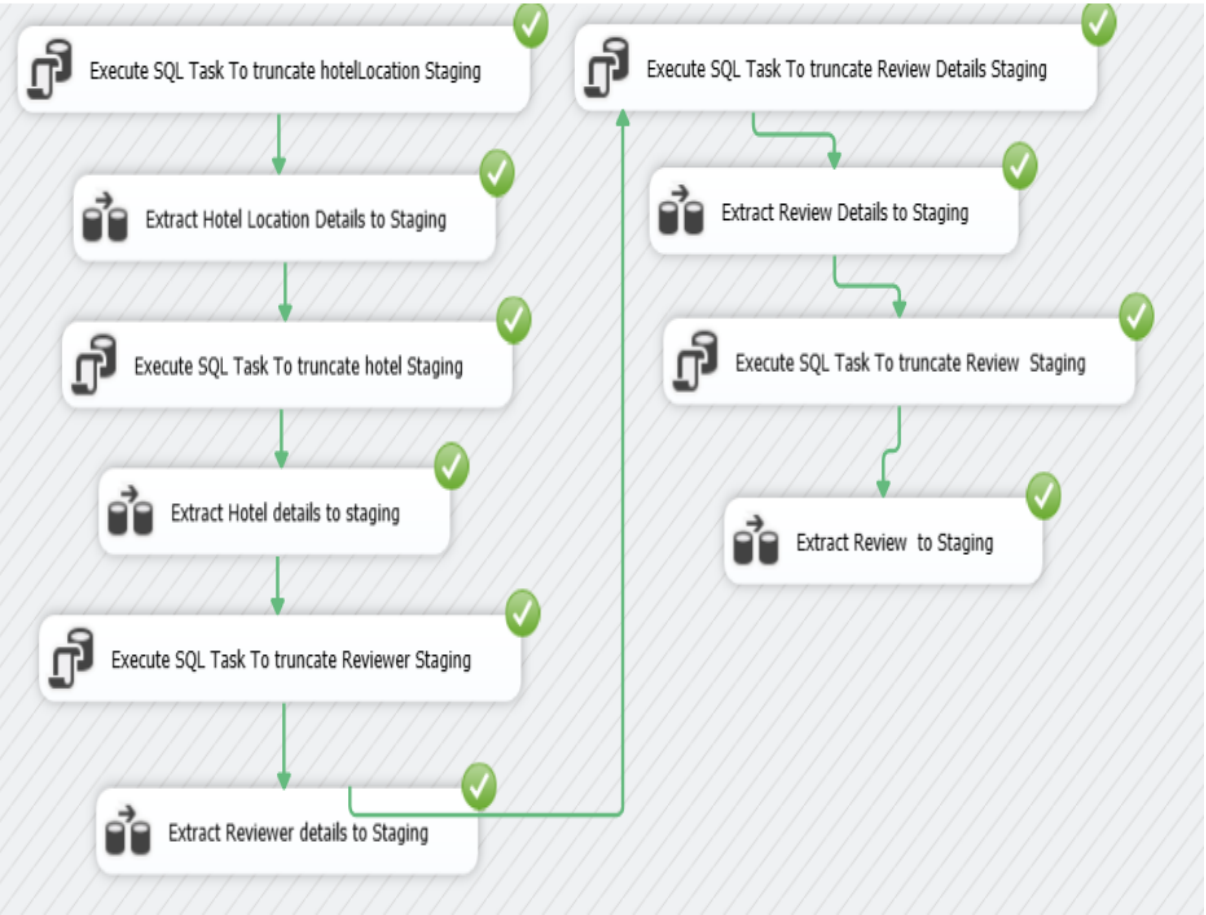
Grain : Details of a review on a specific hotel by a reviewer .

Assumptions:

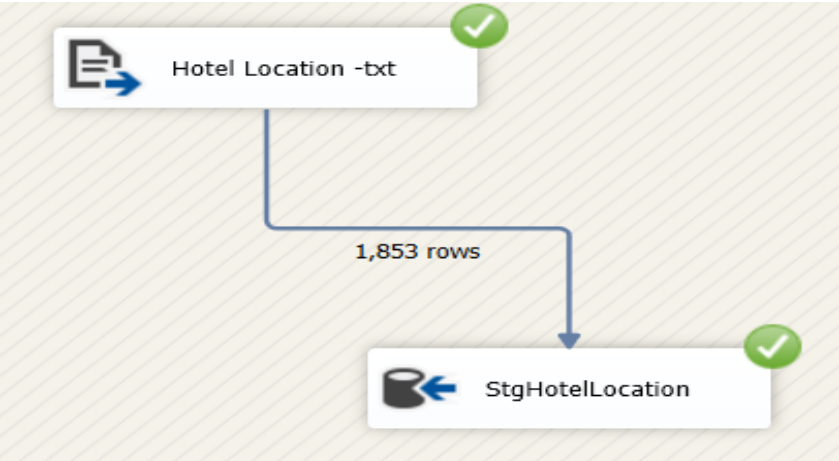
I decided hotel dimension is a slowly changing dimension assuming hotel name, categories ,primary categories and location details can be changed over time.

5. ETL development

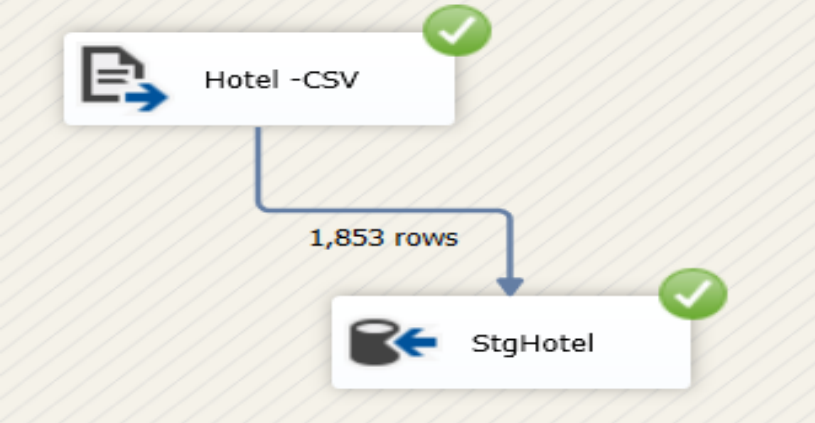
a) Extracting data from the source and load it to the staging. For each stage I truncate the staging tables to clean the database before loading to avoid duplication in the loading data.



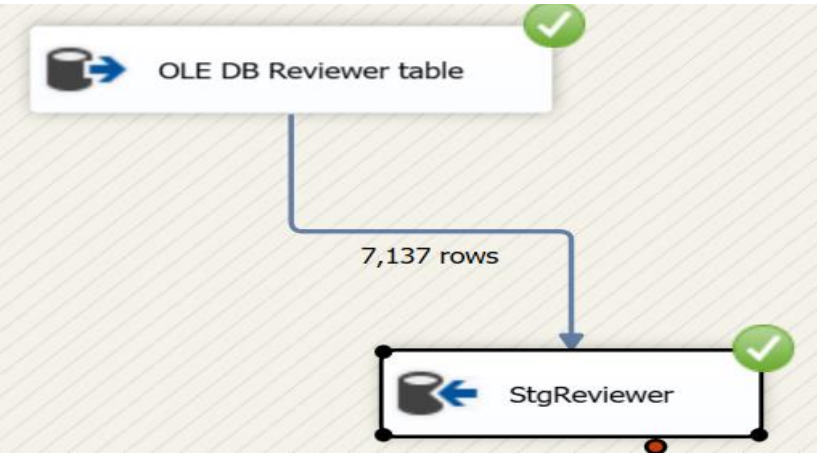
i. Extracting hotel location details from a text file and loading.



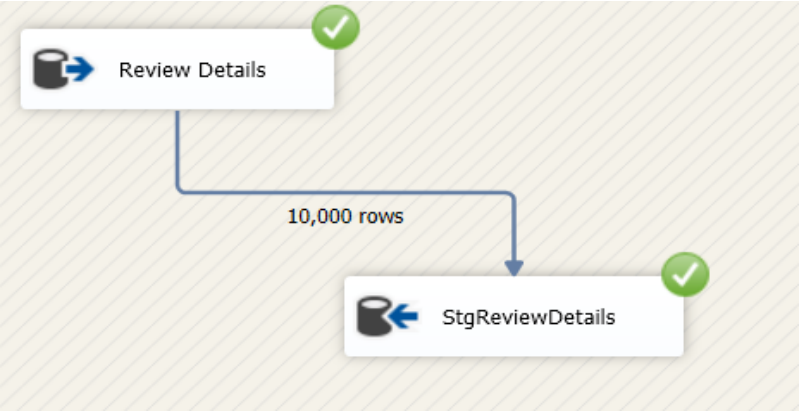
ii. Extracting hotel details from a csv file and loading.



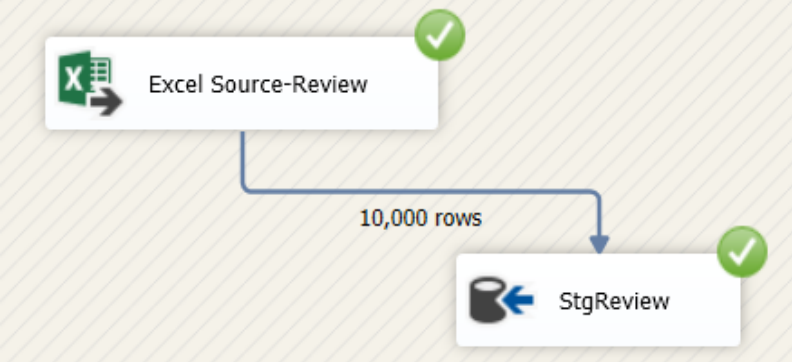
iii. Extracting reviewer details from DB table and loading.



iv. Extracting review details from DB table and loading.



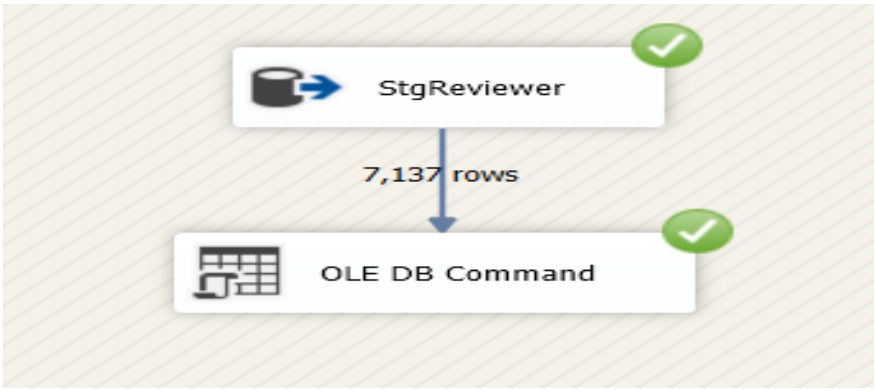
v. Extracting review from an excel file and loading.



b) 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.

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

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

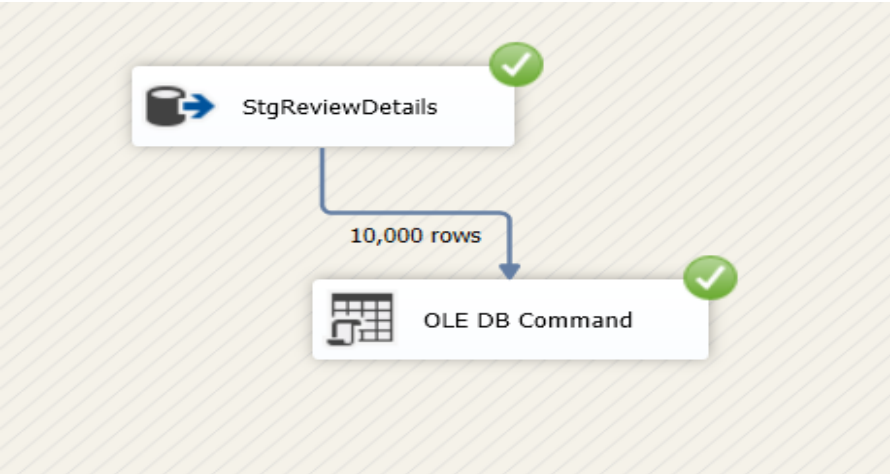


As the OLE DB command I executed the following stored procedure to insert data.

```
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
```

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

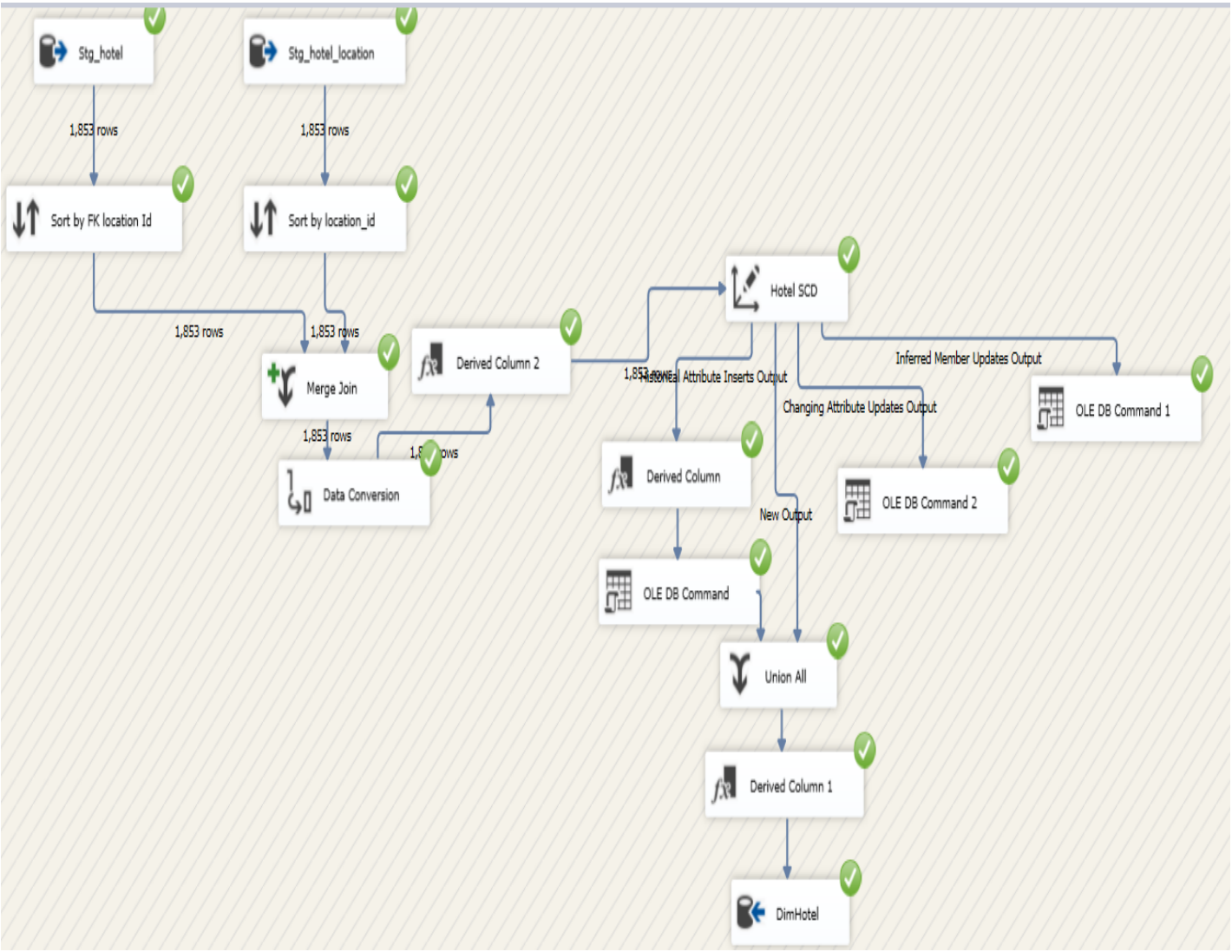


As the OLE DB command ,I executed the following stored procedure to insert data.

```
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
```

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

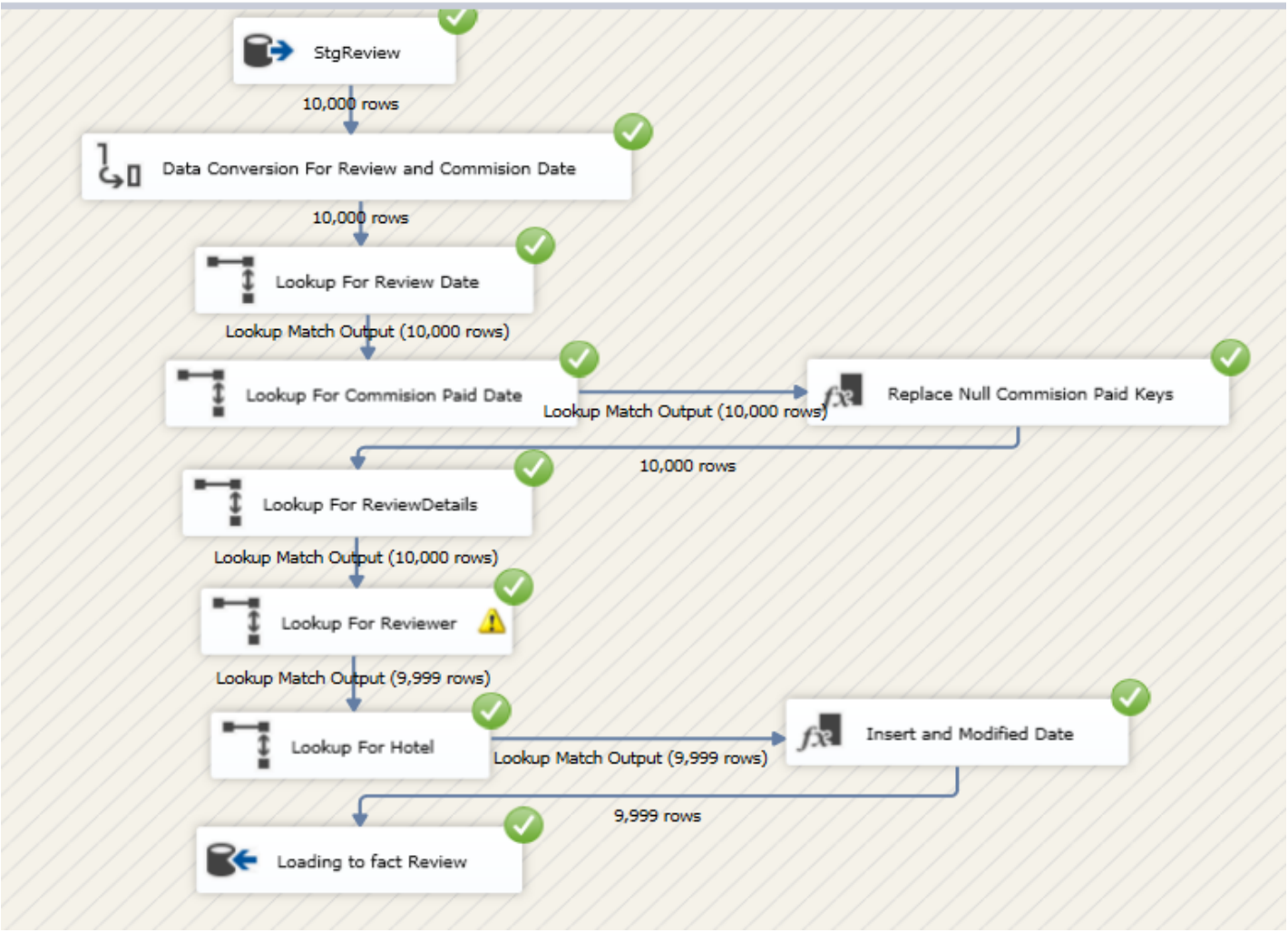


First, I merged hotel and hotel location details after sorting the both tables by hotel_location_id.Then I used a derived column to convert null values in the postal Code to 'NA'.Then I developed the whole Hotel dimension as a slowly changing dimension and load the data to DimHotel dimension.I maintained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

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

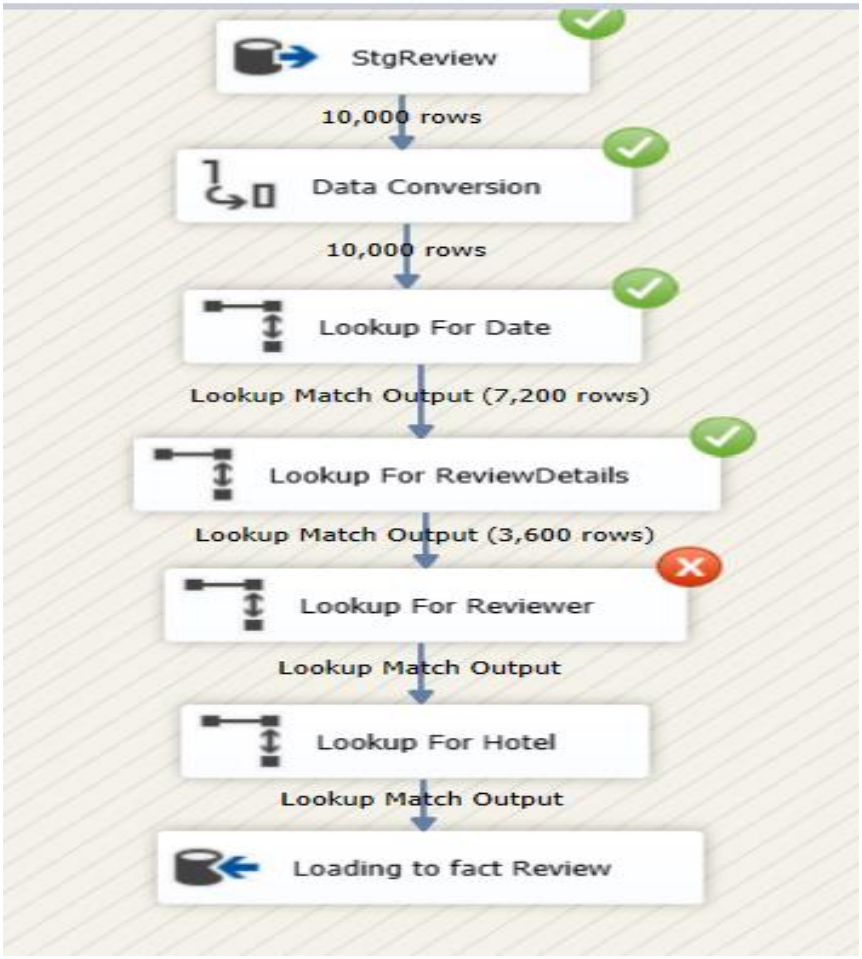
Type 2 (Historical) – Hotel name, address

iv. Extract .transform and load Review to FactReview.

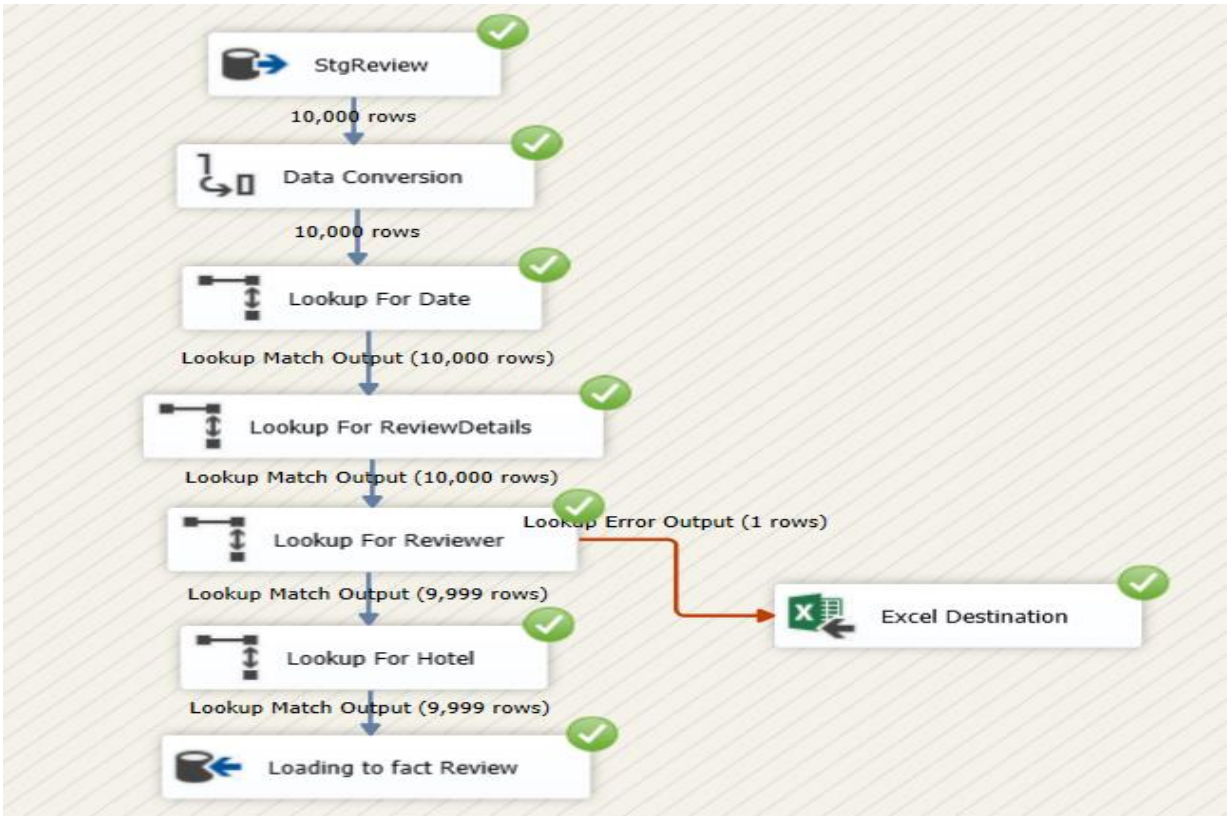


Error Handling

- 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.



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



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

A	B	C	D	E	F	G	H	I	J	K	L
hotel_id	Review_id	reviews.date	reviews.rating	commision	Detail_id	reviewer_id	Copy of rev	ReviewDateKey	ReviewDetails	ErrorCode	ErrorColumn
AVwcg3rXkufWRAb5v-bM	R09320	2015-10-02T00:00:00Z	5	2.1	RD09320	NULL	10/2/2015	20151002	9320	-1.1E+09	0

- I found the reviewer_id and the SK_Reviewer from the following Query.

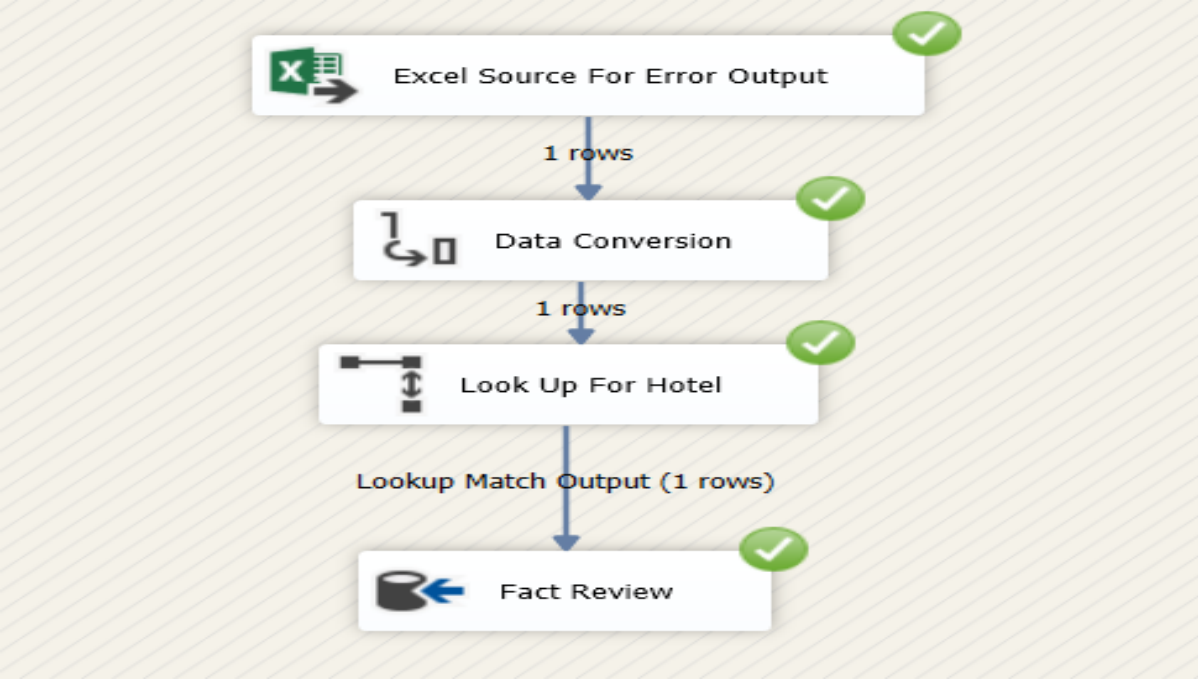
```
select d.reviewer_id,d.SK_Reviewer
from [dbo].[DimReviewer] d left outer join
[dbo].[FactReview] f on d.SK_Reviewer = f.ReviewerKey where f.ReviewerKey is null
```

133 %

Results | Messages

	reviewer_id	SK_Reviewer
1	RWR2665	2665

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



6. Testing Methodology

I. 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.

```
select * from [dbo].StgHotel where hotel_id = 'abcdabcd'
```

33 %

Results | Messages

	locationId	hotel_id	dateAdded	dateUpdated	categories	primaryCategories	name	websites
1	LOC0002	abcdabcd	NULL	NULL	hotel,motel	Accommodation & Food Services	HotelB	https://hotelB.com

- 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’.

```
select * from [DataWarehouse_Hotel_Reviews].[dbo].[DimHotel] where alternate_hotel_id = 'abcdabcd'
```

133 %

Results | Messages

	SK_hotel	alternate_hotel...	cat...	prima...	name	web...	city	provin...	address	postalCode	keys	longitude	latitude	status	modifieddate	insertdate
1	2	abcdabcd	ho...	Acc...	hotelB	http...	Albany	NY	3 Watervliet A...	12206	us/ny/albany/3water...	-73.78336	42.68045	Expired	2021-05-08	2021-05-08
2	1855	abcdabcd	ho...	Acc...	HotelB Updated	http...	Albany	NY	3 Watervliet A...	12206	us/ny/albany/3water...	-73.78336	42.68045	Current	2021-05-08	2021-05-08

- Status of the outdated record was set to ‘**Expired’** and new record was set to ‘**Current’.**

II. 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’.

```
update StgHotel set  categories = 'hotel' , primaryCategories='Food Services' where hotel_id = 'abcdabcd'
select * from [dbo].StgHotel where hotel_id = 'abcdabcd'
```

33 % ▾

Results | Messages

	locationId	hotel_id	dateAdded	dateUpdated	categories	primaryCategories	name	websites
1	LOC0002	abcdabcd	NULL	NULL	hotel	Food Services	HotelB Updated	https://hotelB.com

- Then inserted the record to DimHotel and check the values of the alternate_hotel_id = 'abcdabcd'.

```
select * from [dbo].[DimHotel] where alternate_hotel_id = 'abcdabcd'
```

133 % ▾

Results | Messages

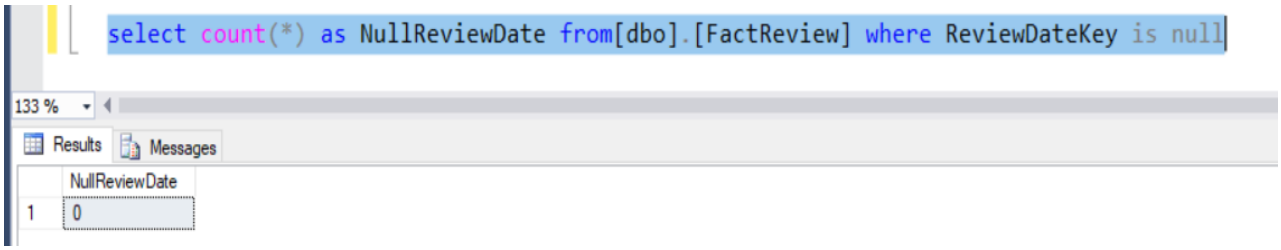
	SK_hotel	alternate_hotel_id	categories	primaryCategories	name	w...	lo...	co...	city	pr...	address	postalCode	keys	longi...	l	status	modifiedd...	insertdate
1	2	abcdabcd	hotel,motel	Accommodation & Food Services	hotelB	h...	L...	US	A...	NY	3 Water...	12206	us/...	-73....		Expired	2021-05-...	2021-05-08
2	1855	abcdabcd	hotel	Food Services	HotelB Updated	h...	L...	US	A...	NY	3 Water...	12206	us/...	-73....		Current	2021-05-...	2021-05-08

- Only the updated record was there under the current record.

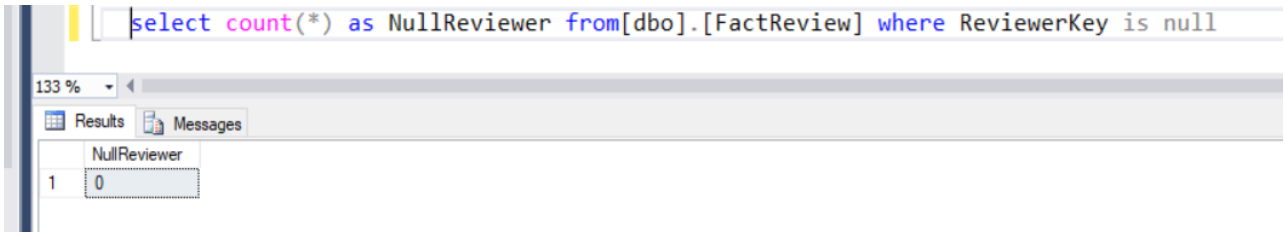
III. Testing null values in the fact table after look up.

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

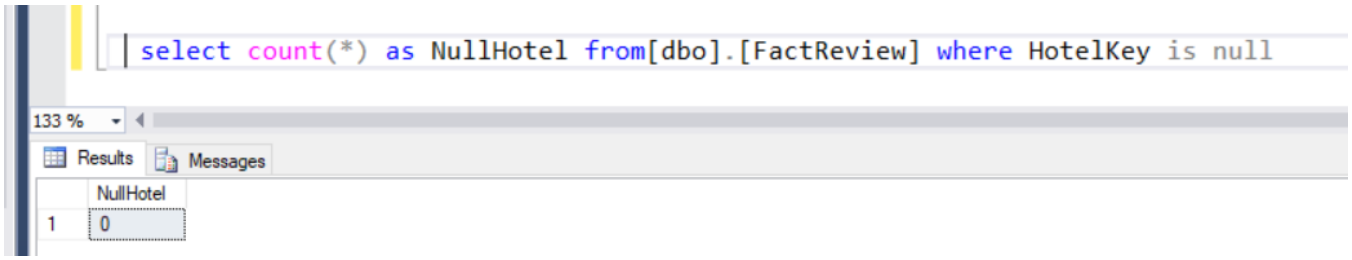
Test null Commision Paid Date key columns.	0	0 *(Refer Attachment 6 5)*	Pass



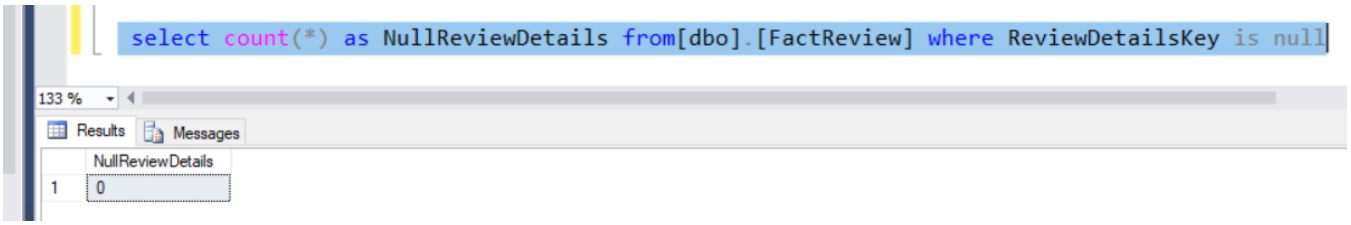
Attachment 6 1



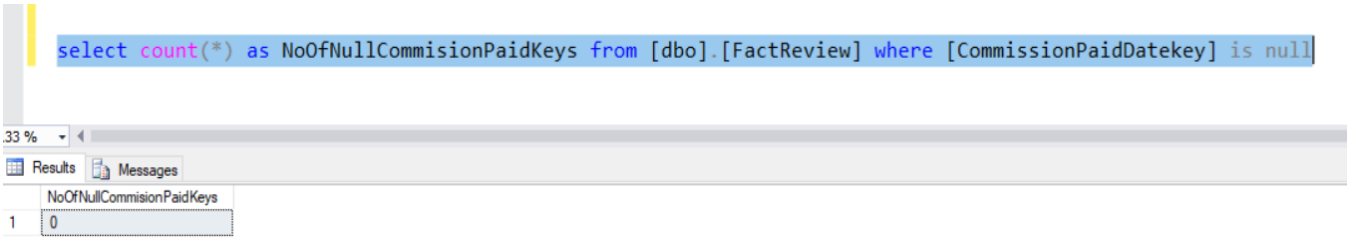
Attachment 6 2



Attachment 6 3



Attachment 6 4



Attachment 6 5

IV. 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

```
select count(*) as NoOfReviewers from  [dbo].[DimReviewer]
```

133 %

Results | Messages

	NoOfReviewers
1	7137

Attachment 6 6

```
select count(*) as NoOfHotels from [dbo].[DimHotel]
```

133 %

Results | Messages

	NoOfHotels
1	1853

Attachment 6 7

```
select count(*) as NoOfReviewDetails from [dbo].[DimReviewDetails]
```

133 %

Results | Messages

	NoOfReviewDetails
1	10000

Attachment 6 8

```
select count(*) as NoOfReviews from[dbo].[FactReview]
```

133 %

Results | Messages

	NoOfReviews
1	10000

Attachment 6 9

V. Test the null values cleansing for the postal code.

Input :

```
select count(*) as NumberOfNullPostalCode from [dbo].[StgHotelLocation] where postalCode is null
```

133 %

Results | Messages

	NumberOfNullPostalCode
1	130

Output :

```
select count(*) as NumberOfNullPostalCode from [dbo].[DimHotel] where postalCode is null
```

133 %

Results | Messages

	NumberOfNullPostalCode
1	0

VI. Test the dimensions and fact table retrieving 10 records per table.

- Retrieve top 10 records from Fact Review

```
select TOP 10 * from [dbo].[FactReview]
```

133 %

Results | Messages

	ReviewDateKey	CommissionPaidDatekey	ReviewDetailsKey	ReviewerKey	HotelKey	insertDate	modifiedDate	reviews#rating	commision#$	noOfSeens
1	20140629	20140706	8898	5382	833	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	5	2.1	3
2	20160716	20160723	8899	3759	1419	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	5	2.7	2
3	20170608	20170615	8900	1271	1006	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	4	2.1	9
4	20160909	-1	8901	5381	1420	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	2	0	0
5	20160323	20160330	8902	7100	1482	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	5	0.8	13
6	20160311	20160318	8903	2633	622	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	5	2	2
7	20150309	20150316	8904	1390	178	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	4.8	2	9
8	20180309	20180316	8905	544	1671	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	5	2.1	9
9	20160723	20160729	8906	5894	572	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	3	2	4
10	20150908	20150915	8907	4828	1740	2021-05-10 17:11:32.160	2021-05-10 17:11:32.160	4	1.7	11

- Retrieve top 10 records from DimReviewer

```
select TOP 10 * from [dbo].[DimReviewer]
```

133 %

Results | Messages

	SK_Reviewer	reviewer_id	reviews#username	reviews#userProvince	reviews#userCity
1	1	RWR0001	112traveler47	NewYork	Jamestown
2	2	RWR0002	146maryw2016	GA	Kingsland
3	3	RWR0003	155hannahp	NotAvailable	Milwaukee
4	4	RWR0004	216Cheezer	NotAvailable	NotAvailable
5	5	RWR0005	2seniorcitizens	NotAvailable	Sayre
6	6	RWR0006	307CatherineW	Texas	Houston
7	7	RWR0007	324kathleenh	FL	Pinecrest
8	8	RWR0008	490aracelic	FL	Fort Myers
9	9	RWR0009	505jll	VA	Richmond
10	10	RWR0010	5168pp	OR	Fairview

- Retrieve top 10 records from DimReviewDetails.

```
select TOP 10 * from [dbo].[DimReviewDetails]
```

133 %

Results | Messages

	SK_ReviewDetails	reviews#sourceURLs	reviews#text	reviews#title	Detail_id
1	1	https://www.hotels.com/hotel/112919/reviews%20/	Very clean, comfortable, friendly hotel. Nicely mainta...	Best Hotel for the value in the area	RD00001
2	2	https://www.tripadvisor.com/Hotel_Review-g60898-d...	I was going to be in downtown Atlanta for a conven...	Truly unprofessional, did not match their high r...	RD00002
3	3	http://www.tripadvisor.com/Hotel_Review-g32332-d5...	We were very pleased with how clean this Super8 ...	Great New Place to Stay	RD00003
4	4	https://www.tripadvisor.com/Hotel_Review-g34296-d...	This hotel is in a great location. It is close to many re...	Nice, Quaint Hotel	RD00004
5	5	https://www.tripadvisor.com/Hotel_Review-g34141-d...	The sea captain has it all, it is the friendliest place I ...	Family holiday	RD00005
6	6	https://www.tripadvisor.com/Hotel_Review-g45964-d...	I love this hotel. I took a chance on it about two yea...	Girls Weekend	RD00006
7	7	http://www.tripadvisor.com/Hotel_Review-g30196-d1...	The staff was friendly, the room was great and spaci...	A Must Stay!!!	RD00007
8	8	http://www.tripadvisor.com/Hotel_Review-g45438-d2...	It's a shame that I didn't do more research as far as t...	Loud people all hours of the night ,classifed as...	RD00008
9	9	http://www.expedia.com/Hotels.h8842065-p43.Hotel-...	This was a great hotel. Everything was brand new a...	Business Stay. One night. Perfect.	RD00009
10	10	http://www.tripadvisor.com/Hotel_Review-g34583-d1...	We don't give many 5 stars, but this hotel deserves i...	All hotels should take a lesson from this one!	RD00010

- Retrieve top 10 records from DimHotel

```
select top 10 * from [dbo].[DimHotel]
```

133 %

Results | Messages

	SK_hotel	alternate_...	categori...	primar...	name	websites	locationId	c...	city	province	address	postalCode	keys	insertdate	longitude	latitude	modifiedd...	status
1	1	AVwcpbn...	Hotels ...	Acco...	Wester...	http://www.thewe...	LOC0001	U.	Abilene	TX	3201 S 1st St	79605	us/tx/a...	2021-05-11 12:...	-99.7629	32.4505	2021-05-...	Current
2	2	AVwdbrLr...	Hotel,...	Acco...	Ramad...	https://www.wynd...	LOC0002	U.	Albany	NY	3 Watervliet Ave. Ext.	12206	us/ny/a...	2021-05-11 12:...	-73.78336	42.68045	2021-05-...	Current
3	3	AVwcw3a...	Hotels,...	Acco...	Americ...	https://www.redlio...	LOC0003	U.	Albert Lea	MN	2306 E Main St	56007	us/mn/...	2021-05-11 12:...	-93.33518	43.65524	2021-05-...	Current
4	4	AVwd7mj...	Hotels,...	Acco...	Algoma...	http://www.algom...	LOC0004	U.	Algoma	WI	1500 Lake St	54201	us/wi/a...	2021-05-11 12:...	-87.44031	44.59898	2021-05-...	Current
5	5	AVweSsy-...	Hotels ...	Acco...	Wingat...	http://www.wingat...	LOC0005	U.	Alpharetta	GA	1005 Kingswood Pl	30009	us/ga/...	2021-05-11 12:...	-84.32345	34.03955	2021-05-...	Current
6	6	AVwdfXN...	Hotel,...	Acco...	Countr...	https://www.count...	LOC0006	U.	Ames	IA	2605 SE 16th St	50010	us/ia/a...	2021-05-11 12:...	-93.57829	42.00878	2021-05-...	Current
7	7	AVwcg3r...	Hotels,...	Acco...	Eden ...	http://www.edenr...	LOC0007	U.	Anaheim	CA	1830 S West St	NA	us/ca/...	2021-05-11 12:...	-117.9236	33.80212	2021-05-...	Current
8	8	AVweT72...	Resort...	Acco...	Holiday...	https://www.ihg.c...	LOC0008	U.	Anaheim	CA	1915 S Manchester Ave	NA	us/ca/...	2021-05-11 12:...	-117.9016	33.80036	2021-05-...	Current
9	9	AVwcw6l...	Hotels,...	Acco...	Quality ...	https://www.choic...	LOC0009	U.	Ashland	VA	107 N Carter Rd	23005	us/va/...	2021-05-11 12:...	-77.46111	37.76192	2021-05-...	Current
10	10	AVweiyLa...	Hotels,...	Acco...	Ranch...	http://ranchoteem...	LOC0010	U.	Atascadero	CA	6895 El Camino Real	93422	us/ca/...	2021-05-11 12:...	-120.6663	35.48774	2021-05-...	Current