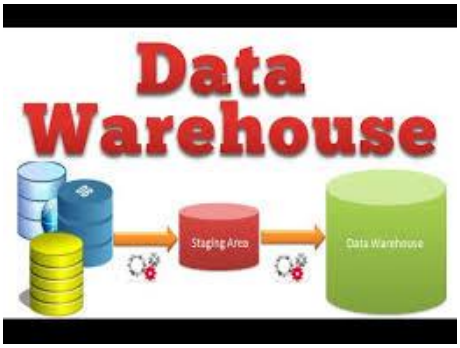
Group Project - Team 3

CIS 9440 - Section QTRA

12/19/19

Data Warehousing and Analytics - Airbnb



Mengxing Dong (mengxing.dong@baruchmail.cuny.edu)
Anthony Li (anthony.li@baruchmail.cuny.edu)
Shobhit Ratan (shobhit.ratan@baruchmail.cuny.edu)

Table of Contents

1. **Introduction** 2
2. **Problem and Opportunity** 2
3. **Management Information Needs** 3
4. **KPIs** 3
5. **Dimensional Model Diagram** 4
   - 5.1 **Dimensional Modelling** 4
   - 5.2 **Airbnb Data Warehousing - Dimensional Model** 8
6. **ETL Programming** 9
   - 6.1 **Introduction** 9
   - 6.2 **ETL Implementation** 9
     - 6.2.1 **Inventory** (Inventory_Dimension) 10
     - 6.2.2 **Time** (Date_Dimension) 12
     - 6.2.3 **Booking** (Booking_Dimension) 14
     - 6.2.4 **Geography** (Geography_Dimension) 16
     - 6.2.5 **Customer Reviews** (Reviews_Dimension) 18
     - 6.2.6 **Earnings** (Earnings_Fact) 20
     - 6.2.7 **Occupancy Rate** (Occupancy_Rate_Fact) 22
   - 6.3 **Data Migration From H2 Database to Oracle Database** 24
7. **Dashboard Programming** 25
   - 7.1 **Migration of Final Schema of Airbnb DW from Oracle to Tableau** 25
   - 7.2 **Dashboards** 27
     - 7.2.1 **Dashboard #1: Annual Earnings and Zillow Home Value Index Trend** 27
     - 7.2.2 **Dashboard #2: Top 10 Neighborhoods with the Highest Earnings** 28
     - 7.2.3 **Dashboard #3: Borough with the Highest Average Occupancy Rate(%)** 29
     - 7.2.4 **Dashboard #4: Room Types with the Highest Average Occupancy Rate** 30
8. **Narrative Conclusion: *Group's experience with the project*** 31
   - 8.1 **Most difficult Steps** 31
   - 8.2 **Easiest Steps** 31
   - 8.3 **Learnings that we did not imagine we would have?** 31
   - 8.4 **What would we do differently, If we had to do it all over again?** 32
   - 8.5 **Realization of proposed benefits to the new system** 32

1. Introduction

As a group, we have decided to build a data warehouse and analyze Airbnb's data. The site contains listings and reviews datasets for cities across the globe. We plan to use *Bottom Up Architecture* and build Data Marts based on geographic regions such as North America, Europe, Asia, Middle East etc. We have identified following three data sources from public domain. Two for the Airbnb(http://insideairbnb.com/get-the-data.html) and one external data set of Zillow (https://www.zillow.com/research/data/). The details of data set are as follows:

Airbnb Datasets

- Listing Dataset. *Listings dataset* contains Listing ID, Listing name, Host ID, Host name, Neighborhood, Geospatial data, Room type, Price, Minimum nights, Number of reviews, Last review, Reviews per month, Calculated host listings count and Availability 365.

- Reviews Dataset. *Reviews dataset* contains the Listing ID and the review dates for conducting time-based analytics.

Zillow Dataset

Zillow data sets contains Zillow Home value datasets for homes across the United States. We used the dataset for homes in the New York City area. Dataset contains Date, RegionID, RegionName, State, Metro, County, City, SizeRank, Zhvi(Zillow Home Value Index), MoM, QoQ, YoY, 5 Year, 10 Year, PeakMonth, PeakQuarter, PeakZHVI, PctFallFromPeak, LastTimeAtCurrentZHVI.

2. Problem and Opportunity

Salient problems faced by hospitality industry in general and specific to Airbnb are:

- Listing the right inventory,
- Unused inventory
- Poor Customer Review

With a growing middle class and increased disposable income the number of people with increased discretional spending on travel and tourism is consistently increasing. Hence

there is considerable opportunity for the hospitality industry. Brick and mortar travel and tour operators have been replaced by online booking portals. Today users to these sites can book the accommodation of their choice anytime, anywhere. Further success of Uberised business model has made people more open to use their personal property for earning revenue. This provides another opportunity to businesses like Airbnb to scale inventory of rooms without any investment in the capital assets.

3. Management Information Needs

Data is the new capital and management teams that best leverage the data analytics will have a competitive advantage over their competitors. This data warehouse is envisaged to provide Decision Support and Business Intelligence to Management. Salient facts which could be of interest for effective decision making are:

- Occupancy rates:
    - Based on Region
    - Based on Customer Reviews
    - Based on overall inventory
- Earnings:
    - Globally
    - By Region

- Global Footprint of Airbnb

Note: For our Group project, we have restricted our scope of Airbnb data warehousing to the New York City neighborhoods.

4. KPIs

KPIs for effective decision making and business intelligence are:

- What is the percentage of occupancy?
- Which area has maximum occupancy?
- Which room types are the most rented?
- In which price range are maximum number of rooms rented?

5. Dimensional Model Diagram

The list of envisaged Dimensions and Facts for our data warehousing project is as follows:

Dimensions

- Inventory
- Time
- Booking
- Geography
- Customer Reviews

Facts

- Earnings
- Occupancy Rate

5.1 Dimensional Modelling

Dimensional modelling for Dimensions and Facts is as follow:

Inventory_dim

FieldName	***Description***
Inventory_dim_id	Primary Key
listing_id	
listing_name	
host_id	
host_name	
room_type	

Date_dim

FieldName	*Description*
Date_dim_id	Primary Key
Date	
Week	
Month	
Quarter	
Year	

Booking_dim

FieldName	*Description*
Booking_dim_id	Primary Key
price	
Minimum nights	
Booked_nights	
Total_number_of_nights	

Geography_dim

FieldName	*Description*
Geography_dim_ID	Primary Key
Neighborhood	
Neighborhood Group	
City	
State	

Reviews_dim

FieldName	*Description*
Reviews_dim_id	Primary Key
Number of Reviews	
Last review	

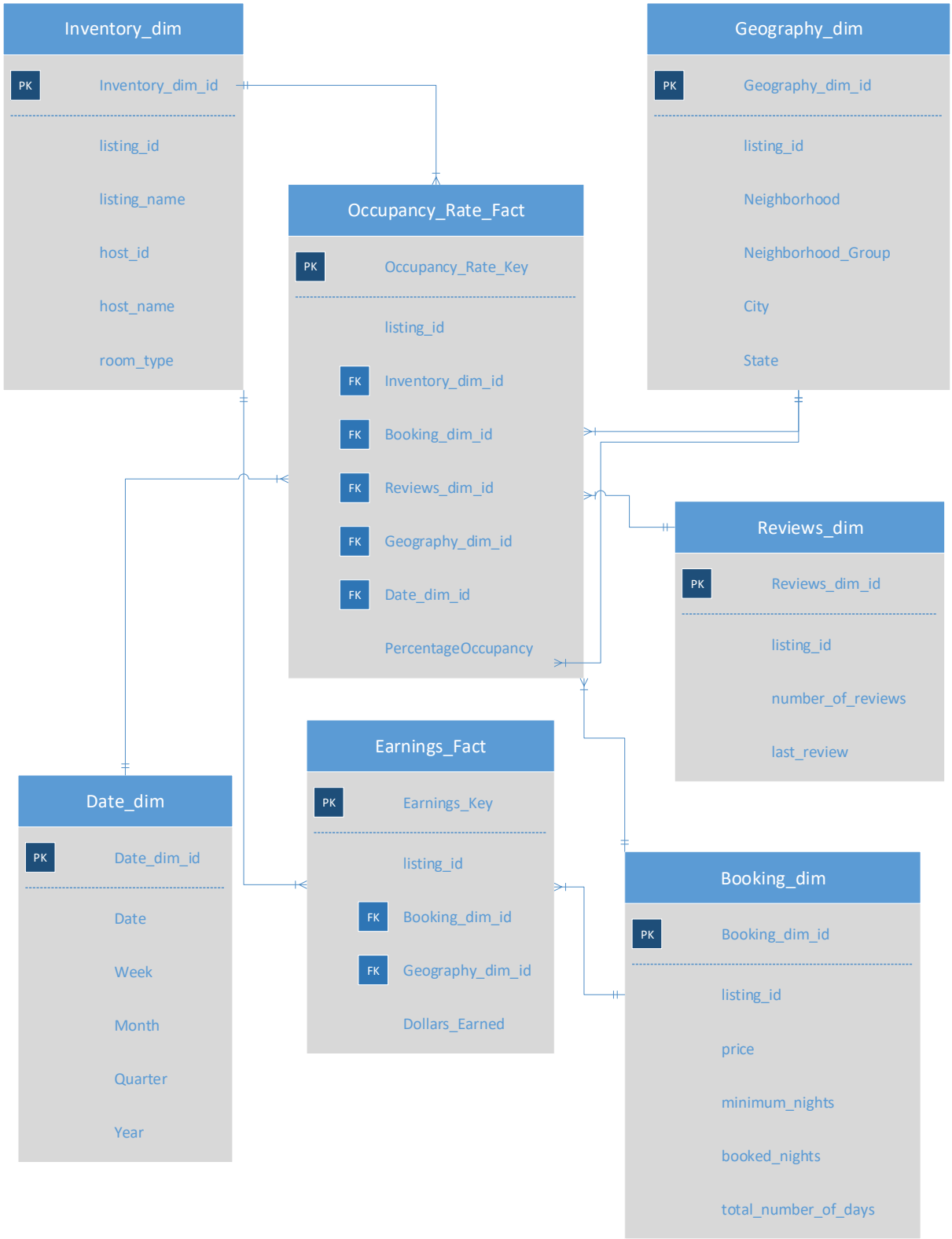
Earnings_Fact

FieldName	*Description*
Earnings_Key	Primary Key
Listing_id	Degenerate Dimension
Booking_dim_id	Foreign Key
Date_dim_id	Foreign Key
Geography_dim_id	Foreign Key
Dollars Earned	

Occupancy_Rate_Fact

FieldName	*Description*
Occupancy_Rate_Key	Primary Key
listing_id	Degenerate Dimension
Inventory_dim_id	Foreign Key
Booking_dim_id	Foreign Key
Reviews_dim_id	Foreign Key
Geography_dim_id	Foreign Key
Date_dim_id	Foreign Key
Percentage Occupancy	

5.2 Airbnb Data Warehousing - Dimensional Model



6. ETL Programming

6.1 Introduction

We have used Pentaho for our ETL programming. The various steps used to implement ETL programming are as follows:

Step 1: Import data csv files

Step 2: Transformation of the data imported in Step 1

Step 3: Load the transformed data into a target DBMS

Initially we used H2, as our target database, to validate working of our ETL programming. In the next phase, we used Oracle as our target database. We preferred Oracle because it is the leading Database application in the market and Oracle Database Architects are in high demand.

6.2 ETL Implementation

ETL programming has been done to implement following dimensions and facts:

Dimensions

- Inventory (*Inventory_Dimension)*
- Time (Date_Dimension)
- Booking (Booking_Dimension)
- Geography (Geography_Dimension)
- Customer Reviews (Reviews_Dimension)

Facts

- Earnings (Earnings_Fact)
- Occupancy Rate (Occupancy_Rate_Fact)

ETL implementation processes for our *Dimensions* and *Facts,* along with the associated execution steps, are as follows:

6.2.1 **Inventory** (Inventory_Dimension)

Inventory_Dimension Import

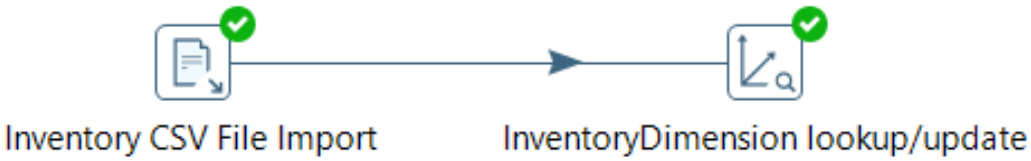


Figure 1: Inventory_dim Import Transformation Process

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Inventory CSV File Import	0	0	48377	48378	0	0	0	0	Finished	19.1s	2,538	-
2	InventoryDimension lookup/update	0	48377	48377	48377	48377	0	0	0	Finished	23.4s	2,066	-

Figure 2: Inventory_dim Import Step Metrics

Rows of step: Inventory_dim (100 rows)

#	INVENTORY_DIM_ID	VERSION	DATE_FROM	DATE_TO	LISTING_ID	LISTING_NAME	HOST_ID	HOST_NAME	ROOM_TYPE
1	0	1	<null>	<null>	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	3647.0	THE VILLAGE OF HARLEM....NEW YORK !	4632.0	Elisabeth	Private room
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	3831.0	Cozy Entire Floor of Brownstone	4869.0	LisaRoxanne	Entire home/apt
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5022.0	Entire Apt: Spacious Studio/Loft by central park	7192.0	Laura	Entire home/apt
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5099.0	Large Cozy 1 BR Apartment In Midtown East	7322.0	Chris	Entire home/apt
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5121.0	BlissArtsSpace!	7356.0	Garon	Private room
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5178.0	Large Furnished Room Near B'way	8967.0	Shunichi	Private room
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5203.0	Cozy Clean Guest Room - Family Apt	7490.0	MaryEllen	Private room
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5222.0	Best Hideaway	7516.0	Marilyn	Entire home/apt
1..	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5238.0	Cute & Cozy Lower East Side 1 bdrm	7549.0	Ben	Entire home/apt

Figure 3: Inventory_dim Sample Data

Execution Steps for *Inventory_Dimension* ETL Implementation: We imported and transformed the Inventory_dim table as a Type 2 SCD in *Spoon*. Steps used to load the sample data, as shown in *Figure 3 above,* are as follows:

1. Clicked on File -> New -> Transformation.

2. Copied the CSV file input option from the Input drop down menu.
3. Edited the properties of the file input and copied the Inventory-export.csv file to Pentaho.
4. Extracted and edited the datatypes into the CSV file input.
5. Copied the Dimension lookup / update from the Data Warehouse menu.
6. Added the database connection ‘Localhost_H2_Airbnb’ to the H2 database.
7. Added the target schema and target table.
8. Added the keys to the table: listing_id (Dimension) = listing_id (Field in Stream).
9. Copied the other data types to the dimension table with “Insert” (Type 2 SCD) listed in the type of dimension updates under Fields.
10. Added Inventory_dim_id to the technical key field.
11. Clicked on the SQL tab and executed the SQL code to create the dimension table.
12. Saved and ran the transformation (*Figure 1*) and looked at the Step Metrics (*Figure 2*).
13. Clicked on the tools menu -> Database -> Explore and selected the Localhost_H2_Airbnb option.
14. Clicked on the Inventory_dim under Schemas -> Public and previewed the sample data.

6.2.2 Time (Date_Dimension)

Date_Dimension Import

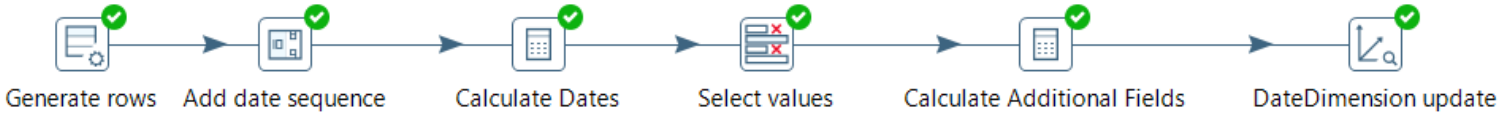


Figure 4: Date_Dim Transformation Process

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Generate rows	0	0	110000	0	0	0	0	0	Finished	20.1s	5,474	-
2	Add date sequence	0	110000	110000	0	0	0	0	0	Finished	22.9s	4,800	-
3	Calculate Dates	0	110000	110000	0	0	0	0	0	Finished	25.9s	4,253	-
4	Select values	0	110000	110000	0	0	0	0	0	Finished	28.8s	3,824	-
5	Calculate Additional Fields	0	110000	110000	0	0	0	0	0	Finished	31.7s	3,470	-
6	DateDimension update	0	110000	110000	110000	0	0	0	0	Finished	34.8s	3,164	-

Figure 5: Date_Dim Step Metrics

Rows of step: Date_dim (100 rows)

#	DATE_DIM_ID	VERSION	DATE_FROM	DATE_TO	ORDER_DATE	ORDER_DAY_OF_YEAR	ORDER_MONTH	ORDER_YEAR	ORDER_QUARTER	ORDER_MONTH_NAME
1	0	1	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2008/08/02 00:00:00.000000000	215.0	8.0	2008.0	3.0	August
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2008/08/03 00:00:00.000000000	216.0	8.0	2008.0	3.0	August
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2008/08/04 00:00:00.000000000	217.0	8.0	2008.0	3.0	August
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2008/08/05 00:00:00.000000000	218.0	8.0	2008.0	3.0	August
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2008/08/06 00:00:00.000000000	219.0	8.0	2008.0	3.0	August
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2008/08/07 00:00:00.000000000	220.0	8.0	2008.0	3.0	August
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2008/08/08 00:00:00.000000000	221.0	8.0	2008.0	3.0	August
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2008/08/09 00:00:00.000000000	222.0	8.0	2008.0	3.0	August

Figure 6: Date_Dim Sample Data

Execution Steps for *Date_Dimension* ETL Implementation: In this step, we generated and embellished the date dimension in *Spoon*. Steps used to generate the sample data, as shown in *Figure 6 above, are* as follows:

1. Clicked on File -> New -> Transform.
2. Copied the Generate Rows option from the Input drop down menu.
3. Added StartDate field starting from 8/1/2008 (Inception of Airbnb).
4. Previewed the data and saw the same date repeated many times.
5. Copied the Add date sequence option for incrementing the dates.
6. Copied the Calculator option for incrementing the dates to the StartDate.
7. Renamed the Calculator to Calculate Dates and created a new field Order_Date where we incremented the dates in StartDate.
8. Copied the Select values option for selecting the sales date so that we can calculate additional fields from Order_Date.
9. Copied the Calculator Option for adding additional fields to the date dimension.
10. Renamed the Calculator to Calculate Additional Fields to add a hierarchy for the dates from days to years.
11. Copied the Dimension lookup / update from the Data Warehouse menu.
12. Added the database connection to the H2 database.
13. Added the target schema and target table.
14. Added the keys to the table: Order_Date (Dimension) = Order_Date (Field in Stream).
15. Copied the other data types to the dimension table with “Update” (Correct error in last version) listed in the type of dimension update under Fields.
16. Added Date_dim_id to the technical key field.
17. Clicked on the SQL tab and executed the SQL code to create the dimension table.
18. Saved and ran the transformation (Figure 4) and looked at the Step Metrics (Figure 5).
19. Clicked on the tools menu -> Database -> Explore and selected the Localhost_H2_Airbnb option.
20. Clicked on the Date_dim under Schemas -> Public and previewed the sample data.

6.2.3 **Booking** (Booking_Dimension)

Booking_Dimension Import

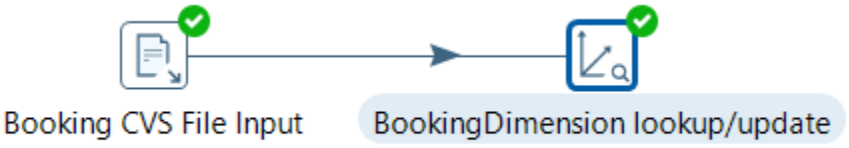


Figure 7: Booking_dim Import Transformation Process

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Booking CVS File Input	0	0	48377	48378	0	0	0	0	Finished	14.2s	3,409	-
2	BookingDimension lookup/update	0	48377	48377	48377	48377	0	0	0	Finished	17.6s	2,752	-

Figure 8: Booking_dim Step Metrics

#	BOOKING_DIM_ID	VERSION	DATE_FROM	DATE_TO	LISTING_ID	PRICE	MINIMUM_NIGHTS	BOOKED_NIGHTS	TOTAL_NUMBER_OF_DAYS
1	0	1	<null>	<null>	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	3647	150	3	0	365
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	3831	89	1	173	365
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5022	80	10	365	365
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5099	200	3	352	365
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5121	60	45	365	365
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5178	79	2	119	365
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5203	79	2	365	365
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5222	116	30	18	365
1..	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5238	150	1	365	365

Figure 9: Booking_dim Sample Data

Execution Steps for *Booking_Dimension* ETL Implementation: We imported and transformed the Booking_dim table as a Type 1 SCD in *Spoon*. Steps used to load the sample data, as shown in *Figure 9 above,* are as follows:

1. Clicked on File -> New -> Transformation.
2. Copied the CSV file input option from the Input drop down menu.

3. Edited the properties of the file input and copied the Booking-export.csv file to Pentaho.
4. Extracted and edited the datatypes into the CSV file input.
5. Copied the Dimension lookup / update from the Data Warehouse menu.
6. Added the database connection to the H2 database.
7. Added the target schema and target table.
8. Added the keys to the table: listing_id (Dimension) = listing_id (Field in Stream).
9. Copied the other data types to the dimension table with “Punch through” (Type 1 SCD) listed in the type of dimension update under Fields.
10. Added Booking_dim_id to the technical key field.
11. Clicked on the SQL tab and executed SQL code to create the dimension table.
12. Saved and ran the transformation (Figure 7) and looked at the Step Metrics (Figure 8).
13. Clicked on the tools menu -> Database -> Explore and selected the Localhost_H2_Airbnb option.
14. Clicked on the Booking_dim under Schemas -> Public and previewed the sample data.

6.2.4 Geography (Geography_Dimension)

Geography_Dimension Import

Figure 10: Geography_dim Import Transformation Process

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Geography CSV file input	0	0	48377	48378	0	0	0	0	Finished	21.4s	2,265	-
2	GeographyDimension lookup/update	0	48377	48377	48377	48377	0	0	0	Finished	26.5s	1,829	-

Figure 11: Geography_dim Import Step Metrics

Rows of step: Geography_dim (100 rows)

#	GEOGRAPHY_DIM_ID	VERSION	DATE_FROM	DATE_TO	LISTING_ID	NEIGHBOURHOOD_GROUP	NEIGHBOURHOOD	CITY	STATE	LATITUDE	LONGITUDE
1	0	1	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	3647.0	Manhattan	Harlem	New York City	NY	40.80902	-73.9419
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	3831.0	Brooklyn	Clinton Hill	New York City	NY	40.68514	-73.95976
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5022.0	Manhattan	East Harlem	New York City	NY	40.79851	-73.94399
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5099.0	Manhattan	Murray Hill	New York City	NY	40.74767	-73.975
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5121.0	Brooklyn	Bedford-Stuyvesant	New York City	NY	40.68688	-73.95596
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5178.0	Manhattan	Hell's Kitchen	New York City	NY	40.76489	-73.98493
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5203.0	Manhattan	Upper West Side	New York City	NY	40.80178	-73.96723
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5222.0	Manhattan	East Village	New York City	NY	40.72764	-73.97949
1..	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5238.0	Manhattan	Chinatown	New York City	NY	40.71344	-73.99037

Figure 12: Geography_dim Sample Data

Execution Steps for *Geography_Dimension* ETL Implementation: We imported and transformed the Geography_dim table as a Type 2 SCD in *Spoon*. Steps used to load the sample data, as shown in *Figure 12 above,* are as follows:

1. Clicked on File -> New -> Transformation.
2. Copied the CSV file input option from the Input drop down menu.

3. Edited the properties of the file input and copied the Geography-export.csv file to Pentaho.
4. Extracted and edited the datatypes into the CSV file input.
5. Copied the Dimension lookup / update from the Data Warehouse menu.
6. Added the database connection ‘Localhost_H2_Airbnb’ to the H2 database.
7. Added the target schema and target table.
8. Added the keys to the table: listing_id (Dimension) = listing_id (Field in Stream).
9. Copied the other data types to the dimension table with “Insert” (Type 2 SCD) listed in the type of dimension updates under Fields.
10. Added Geography_dim_id to the technical key field.
11. Clicked on the SQL tab and executed the SQL code to create the dimension table.
12. Saved and ran the transformation (*Figure 1*) and looked at the Step Metrics (*Figure 2*).
13. Clicked on the tools menu -> Database -> Explore and selected the Localhost_H2_Airbnb option.
14. Clicked on the Geography_dim under Schemas -> Public and previewed the sample data.

6.2.5 Customer Reviews (Reviews_Dimension)

Reviews_Dimension Import

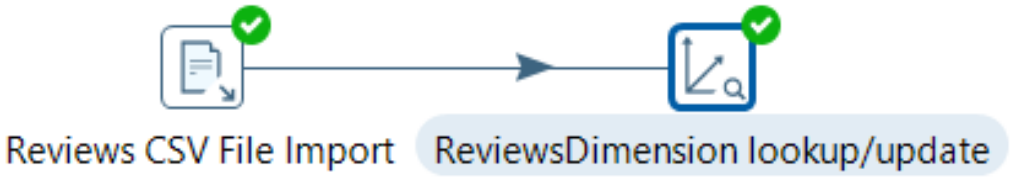


Figure 13: Reviews_dim Import Transformation Process

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Reviews CSV File Import	0	0	38726	38727	0	0	0	0	Finished	12.9s	2,997	-
2	ReviewsDimension lookup/update	0	38726	38726	38726	38726	0	0	0	Finished	17.6s	2,196	-

Figure 14: Reviews_dim Import Step Metrics

Rows of step: Reviews_dim (100 rows)

#	REVIEWS_DIM_ID	VERSION	DATE_FROM	DATE_TO	LISTING_ID	NUMBER_OF_REVIEWS	LAST_REVIEW
1	0	1	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	3831.0	279.0	2019/08/29 00:00:00.000000000
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5022.0	9.0	2018/11/19 00:00:00.000000000
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5099.0	75.0	2019/07/21 00:00:00.000000000
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5121.0	49.0	2017/10/05 00:00:00.000000000
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5178.0	443.0	2019/08/27 00:00:00.000000000
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5203.0	118.0	2017/07/21 00:00:00.000000000
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5222.0	94.0	2016/06/15 00:00:00.000000000
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5238.0	161.0	2019/07/29 00:00:00.000000000
1..	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	5295.0	54.0	2019/08/03 00:00:00.000000000

Figure 15: Reviews_dim Sample Data

Execution Steps for *Reviews_Dimension* ETL Implementation: We imported and transformed the Reviews_dim table as a Type 1 SCD in *Spoon*. Steps used to load the sample data, as shown in *Figure 15 above,* are as follows:

1. Clicked on File -> New -> Transformation.
2. Copied the CSV file input option from the Input drop down menu.
3. Edited the properties of the file input and copied the Reviews-export.csv file to Pentaho.
4. Extracted and edited the datatypes into the CSV file input.

5. Copied the Dimension lookup / update from the Data Warehouse menu.
6. Added the database connection to the H2 database.
7. Added the target schema and target table.
8. Added the keys to the table: listing_id (Dimension) = listing_id (Field in Stream).
9. Copied the other data types to the dimension table with “Punch through” (Type 1 SCD) listed in the type of dimension update under Fields.
10. Added Reviews_dim_id to the technical key field.
11. Clicked on the SQL tab and executed SQL code to create the dimension table.
12. Saved and ran the transformation (Figure 7) and looked at the Step Metrics (Figure 8).
13. Clicked on the tools menu -> Database -> Explore and selected the Localhost_H2_Airbnb option.
14. Clicked on the Reviews_dim under Schemas -> Public and previewed the sample data.

6.2.6 Earnings (Earnings_Fact)

Earnings_Fact Import

Figure 16: Earnings_Fact Import Transformation Process

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Earnings CSV file input	0	0	48377	48378	0	0	0	0	Finished	9.6s	5,064	-
2	Booking Dim Lookup	0	48377	48377	48377	0	0	0	0	Finished	11.9s	4,060	-
3	Geography Dim Lookup	0	48377	48377	48377	0	0	0	0	Finished	12.7s	3,819	-
4	Earnings Total	0	48377	48377	0	0	0	0	0	Finished	12.7s	3,818	-
5	Load Earnings Fact Table	0	48377	48377	0	48377	0	0	0	Finished	13.0s	3,735	-

Figure 17: Earnings_Fact Import Transformation Step Metrics

#	EARNINGS_KEY	LISTING_ID	BOOKING_DIM_ID	GEOGRAPHY_DIM_ID	Annual Earnings
1	1	3647	1	1	0.0
2	2	3831	2	2	15397.0
3	3	5022	3	3	29200.0
4	4	5099	4	4	70400.0
5	5	5121	5	5	21900.0
6	6	5178	6	6	9401.0
7	7	5203	7	7	28835.0
8	8	5222	8	8	2088.0
9	9	5238	9	9	54750.0
1..	10	5295	10	10	43875.0

Figure 18: Earnings_Fact Import Transformation Sample Data

Execution Steps for *Earnings_Fact* ETL Implementation: In this step, we developed the dimensional lookup transformations for Booking and Geography dimensions and loaded the Earnings_Fact Table in *Spoon*. Steps used to load the sample data, shown in Figure 18, are as follows:

1. Clicked on File -> New -> Transformation.
2. Copied the CSV file input option from the Input drop down menu.
3. Edited the properties of the file input and copied the Earnings_export.csv file to Pentaho.
4. Extracted and edited the datatypes into the CSV file input.
5. Copied the Dimension / lookup update from the Data Warehouse menu.
6. Added the database connection to the H2 database.
7. Added the target schema and target table.
8. Added the keys to the table: listing_id (Dimension) = listing_id (Field in Stream) for the Booking Dimension.
9. Repeat steps 5-8 for Geography Dimension.
10. Copied the calculator option for adding the Earnings to the Earnings fact table.
11. Added Annual Earnings using the price and booked_nights from the Booking table.
12. Copied a table output option from the Outputs dropdown menu.
13. Enabled the auto-generated key and named it Earnings_Key.
14. Imported data from other dimensions and removed extra data with only listing_id (degenerate dimension), Booking_dim_id, Geography_dim_id and Annual Earnings field remaining.
15. Clicked on the SQL tab and executed the SQL code to create the fact table.
16. Saved and ran the transformation (Figure 16) and looked at the Step Metrics (Figure 17).
17. Clicked on the tools menu -> Database -> Explore and selected the Localhost_H2_Airbnb option.
18. Clicked on the Earnings_Fact under Schemas -> Public and previewed the sample data.

6.2.7 **Occupancy Rate** (Occupancy_Rate_Fact)

Occupancy_Rate_Fact Import

Figure 19: Occupancy_Rate_Fact Import Transformation Process

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Occupancy Rate CSV file input	0	0	48377	48378	0	0	0	0	Finished	10.3s	4,705	-
2	Inventory Dim Lookup	0	48377	48377	48377	0	0	0	0	Finished	12.9s	3,742	-
3	Booking Dim Lookup	0	48377	48377	48377	0	0	0	0	Finished	13.0s	3,733	-
4	Reviews Dim Lookup	0	48377	48377	48377	0	0	0	0	Finished	13.0s	3,726	-
5	Geography Dim Lookup	0	48377	48377	48377	0	0	0	0	Finished	13.0s	3,718	-
6	Date Dim Lookup	0	48377	48377	48377	0	0	0	0	Finished	13.0s	3,712	-
7	Occupancy Rate Calculation	0	48377	48377	0	0	0	0	0	Finished	14.2s	3,412	-
8	Load OccupancyRate Table	0	48377	48377	0	48377	0	0	0	Finished	16.6s	2,918	-

Figure 20: Occupancy_Rate_Fact Import Transformation Step Metrics

#	OCCUPANCY_RATE_KEY	LISTING_ID	INVENTORY_DIM_ID	BOOKING_DIM_ID	REVIEWS_DIM_ID	GEOGRAPHY_DIM_ID	DATE_DIM_ID	PERCENTOCCUPANCY
1	1	3647	1	1	0	1	0	0
2	2	3831	2	2	1	2	4045	47
3	3	5022	3	3	2	3	3762	100
4	4	5099	4	4	3	4	4006	96
5	5	5121	5	5	4	5	3352	100
6	6	5178	6	6	5	6	4043	32
7	7	5203	7	7	6	7	3276	100
8	8	5222	8	8	7	8	2875	4
9	9	5238	9	9	8	9	4014	100
1..	10	5295	10	10	9	10	4019	89
1..	11	5441	11	11	10	11	4055	98

Figure 21: Occupancy_Rate_Fact Import Transformation Sample Data

Execution Steps for *Earnings_Fact* ETL Implementation: In this step, we developed the dimensional lookup transformations for all the dimensions and loaded the Occupancy_Rate_Fact Table in *Spoon*. Steps used to load the sample data, shown in Figure 21, are as follows:

1. Clicked on File -> New -> Transformation.
2. Copied the CSV file input option from the Input drop down menu.

3. Edited the properties of the file input and copied the Occupancy_export.csv file to Pentaho.
4. Extracted and edited the datatypes into the CSV file input.
5. Copied the Dimension / lookup update from the Data Warehouse menu.
6. Added the database connection to the H2 database.
7. Added the target schema and target table.
8. Added the keys to the table: listing_id (Dimension) = listing_id (Field in Stream) for the Inventory Dimension.
9. Repeat steps 5-8 for the remaining Dimensions.
10. Copied the calculator option for adding the Percent OccupancyRate to the Occupancy_Rate fact table.
11. Added PercentOccupancy using the booked_nights and total_number_of_days from the Booking table.
12. Copied a table output option from the Outputs dropdown menu.
13. Enabled the auto-generated key and named it Occupancy_Rate_Key.
14. Imported data from other dimensions and removed extra data with only listing_id (degenerate dimension), Inventory_dim_id, Booking_dim_id, Reviews_dim_id, Geography_dim_id, Date_dim_id and PercentOccupancy field remaining.
15. Clicked on the SQL tab and executed the SQL code to create the fact table.
16. Saved and ran the transformation (Figure 16) and looked at the Step Metrics (Figure 17).
17. Clicked on the tools menu -> Database -> Explore and selected the Localhost_H2_Airbnb option.
18. Clicked on the Occupancy_Rate_Fact under Schemas -> Public and previewed the sample data.

6.3 Data Migration From H2 Database to Oracle Database

We first tested the ETL programming in H2 database. Once the work was vetted, we migrated the data into Oracle database as it is scalable, high end commercially preferred DBMS. In order to do this, we changed the Target database connection on Pentaho from H2 to Oracle. The migration process for Booking Dimension from H2 to Oracle is shown in Fig 22 below. The same was undertaken for all other Dimensions and Facts.

Booking_Dim

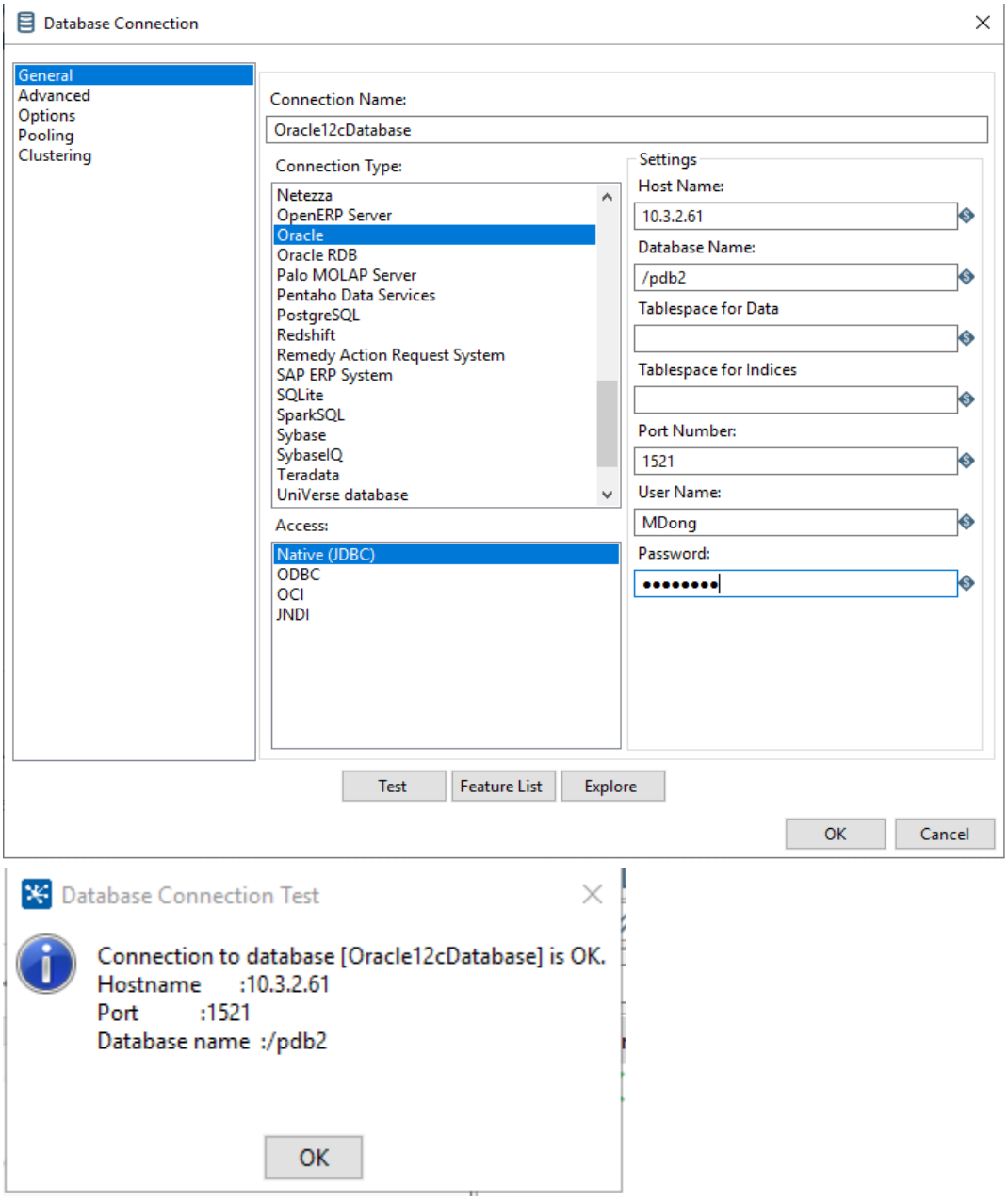


Figure 22: Migration of Booking_Dim from H2 to Oracle

7. Dashboard Programming

7.1 Migration of Final Schema of Airbnb DW from Oracle to Tableau

We needed to access Airbnb Data Warehouse data to undertake dashboard programming (for Business Analytics) on Tableau. Hence, we used Tableau to connect to Oracle server to retrieve our data, which was eventually used for the Dashboard programming. Screen shots showing *Connection to Oracle server* and *Migration of DW data into Tableau* are shown in Fig 23 and 24 respectively.

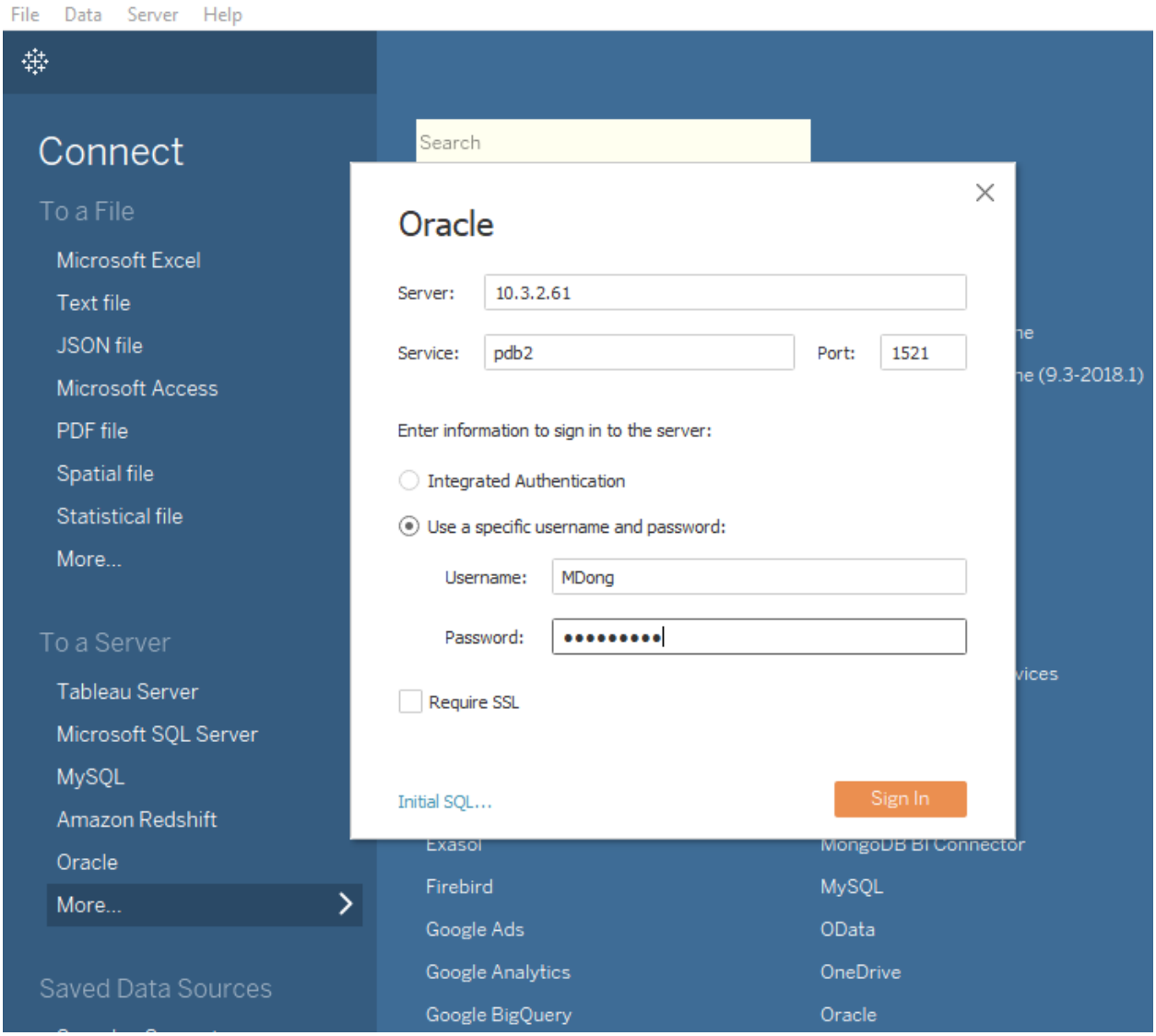


Figure 23: Connection to Oracle Server

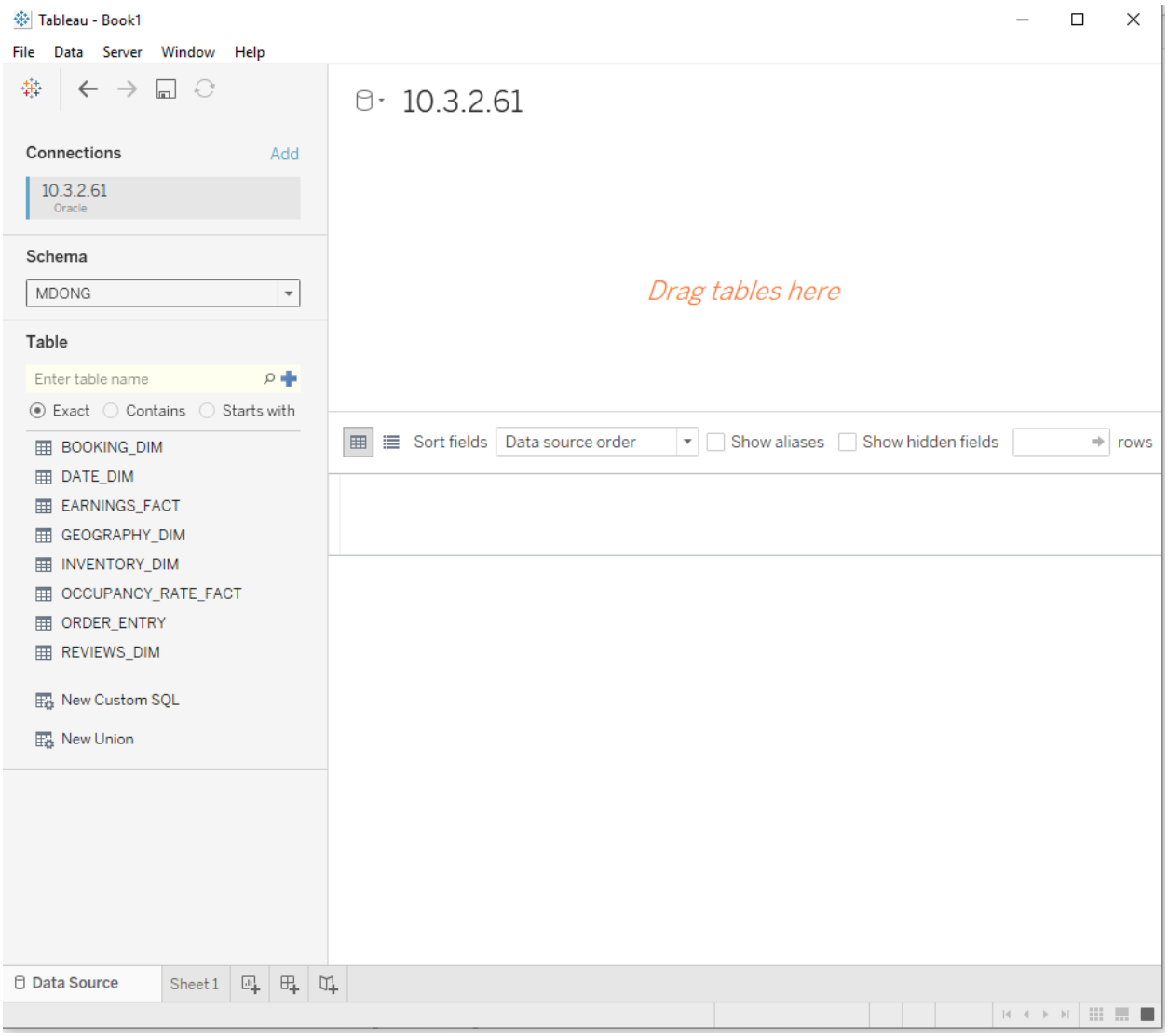


Figure 24: Migration of Airbnb DW data from Oracle Server into Tableau

7.2 Dashboards

7.2.1 Dashboard #1: Annual Earnings and Zillow Home Value Index Trend (NYC)

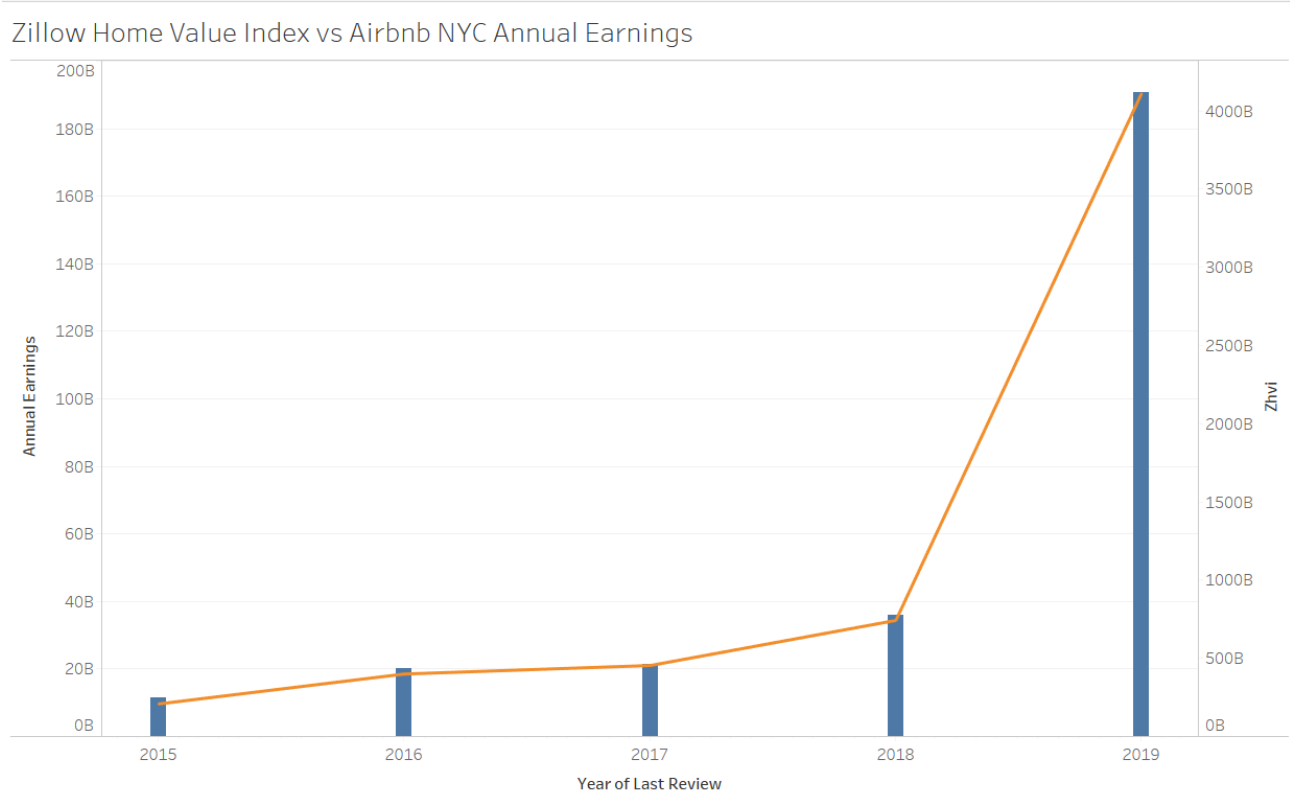


Figure 25: Visualization for Dashboard #1

Key Insights:

- There is a positive relationship between Zillow Home Value Index and Airbnb's Annual Earnings.
- Sharp increase in Airbnb Annual Earnings and Zillow Home Value Index, both measures quadrupled between 2018 and 2019.

7.2.2 Dashboard #2: Top 10 Neighborhoods with the Highest Earnings

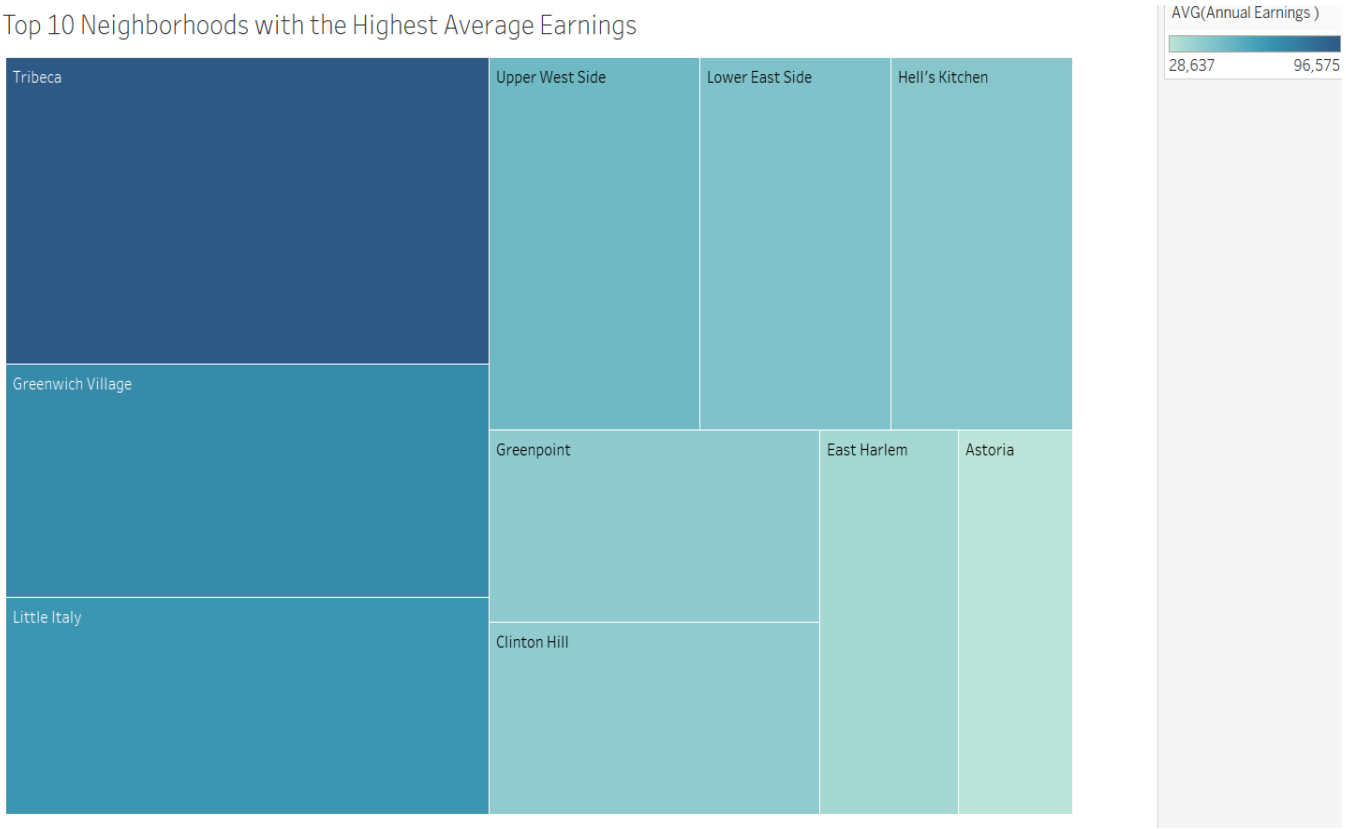


Figure 26: Visualization for Dashboard #2

Key Insights:

- Tribeca and Greenwich Village have the highest average Annual Earnings for Airbnb of $96,570 and $73,399, respectively.
- 7 of the top 10 neighborhoods are located in Manhattan.
- Astoria is the only neighborhood to feature from Queens.

7.2.3 Dashboard #3: Borough with the Highest Average Occupancy Rate(%)

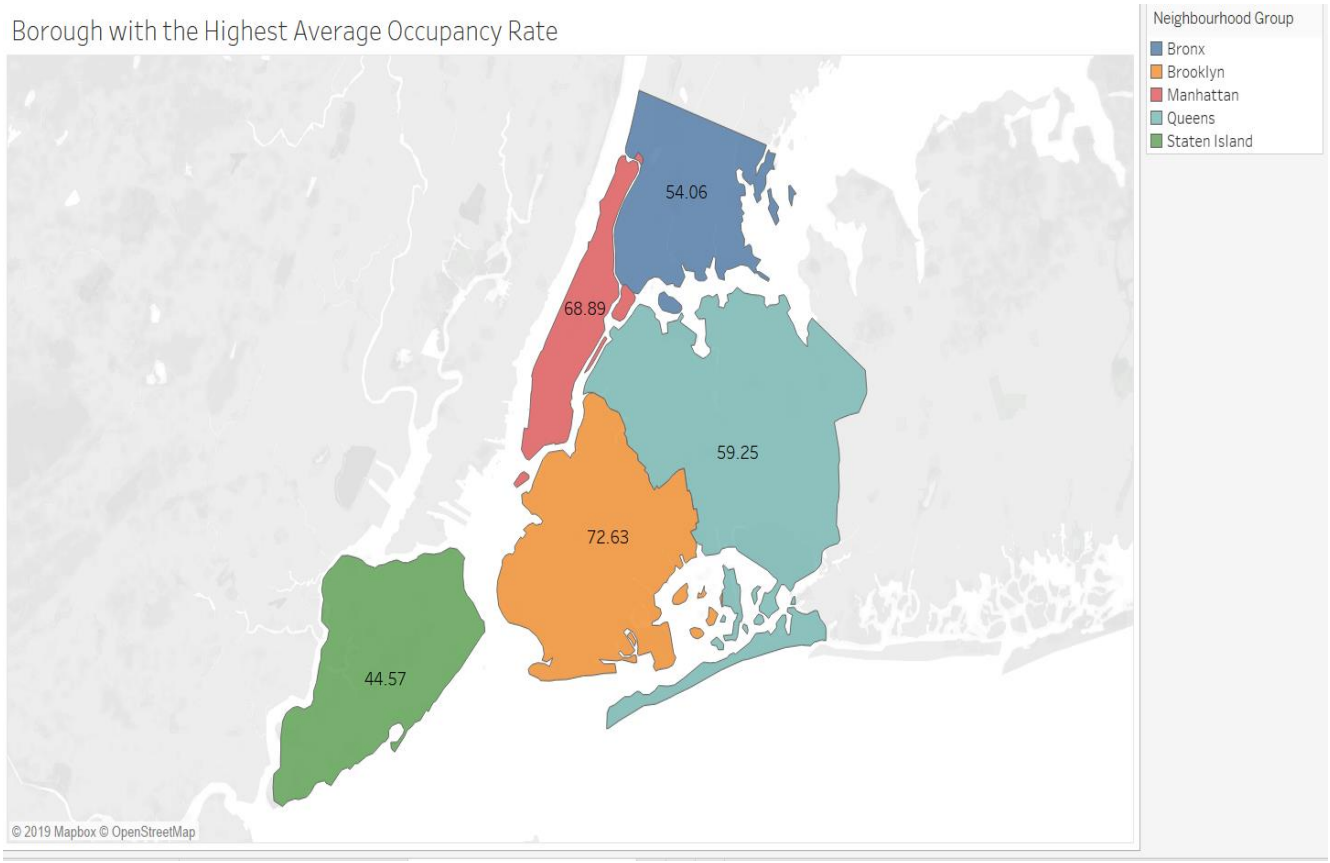


Figure 27: Visualization for Dashboard #3

Key Insights:

- Brooklyn has the highest Average Occupancy Rate and Staten Island has the lowest Average Occupancy Rate.
- Staten Island is the only Borough to have an Average Occupancy Rate < 50%.

7.2.4 Dashboard #4: Room Types with the Highest Average Occupancy Rate

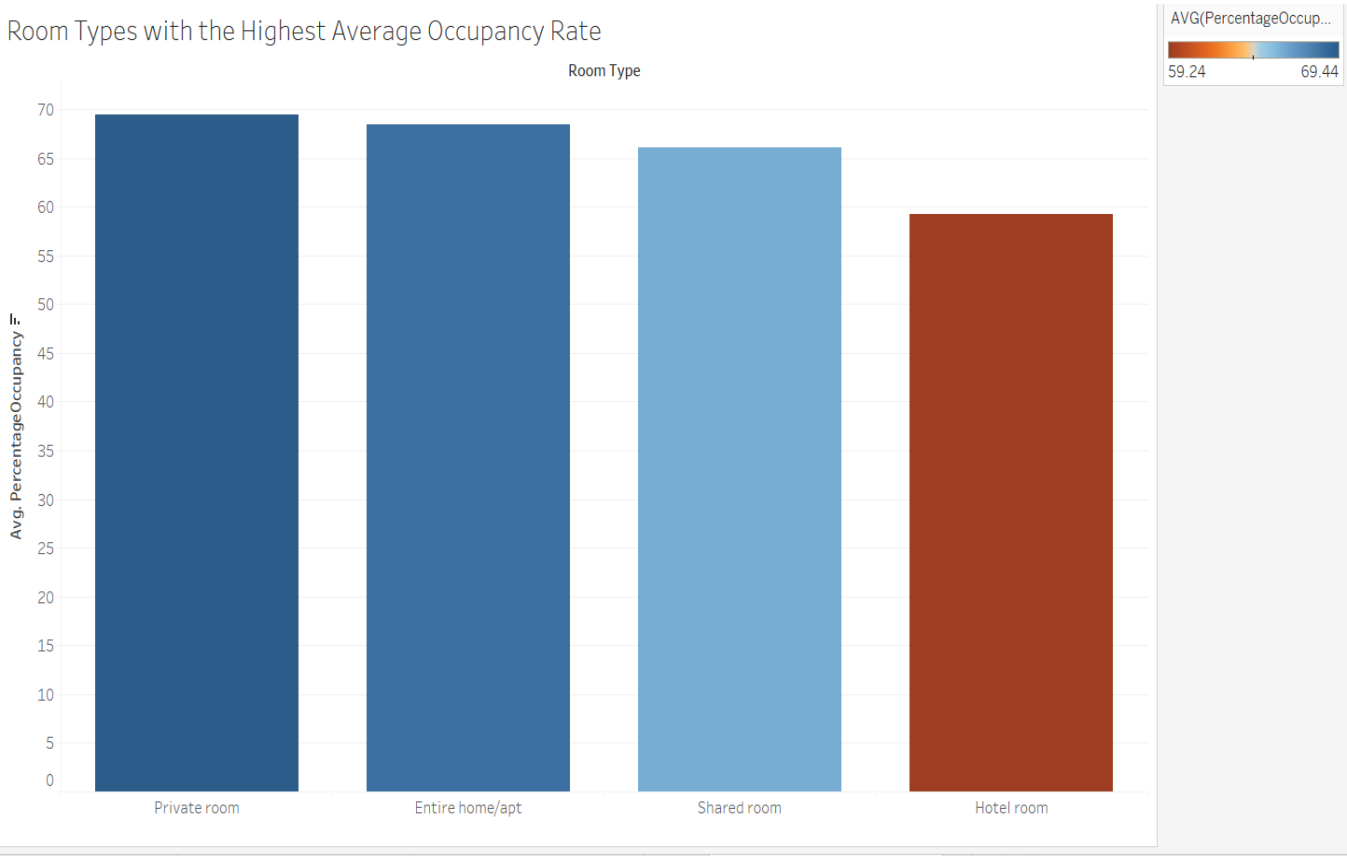


Figure 28: Visualization for Dashboard #4

Key Insights:

- Private rooms have the highest average Occupancy Rate and Hotel rooms have the lowest average Occupancy Rate.
- On average, Private rooms and Entire home / apartment have a higher chance of being rented as compared to Shared rooms and Hotel rooms.

8. Narrative Conclusion: *Group's experience with the project*

It was one of our best group projects. Requirements and milestone deliverables had a direct bearing to the course curriculum. The project looked quite abstract at the beginning and required a lot of reading / hands on practice to acquire the requisite skills to execute the work. The project gave a holistic understanding about Data Warehousing and Business Analytics. Our views on various aspects of the Group Project Tasks are as follows:

8.1 Most difficult Steps

- Finding two datasets that related to each other was one of the most difficult. was the most difficult, since it was hard to find two that could be linked through the same key.
- Creating the Dimensional Model and completing the ETL process was also quite challenging.

8.2 Easiest Steps

- Data Cleaning
- Dashboard Programming
- Playing with the data in Tableau. Making different visualizations so that we can see the possibilities of exactly what we can analyze with our data was interesting. This step was easy since there were not any issues importing the data into Tableau.

8.3 Learnings that we did not imagine we would have?

- We gained a lot of knowledge about the implementation of ETL process. Having the chance to do it from start to finish was challenging yet rewarding at the same time. We also learned a lot about Airbnb and how rooted its operations in NYC are.
- We learnt how to analyze the problem at a higher level, create the Dimensional Model Diagrams, import the data to a working Data Warehouse and export the data from Data Warehouse to a Data Visualization tool.
- Practical implementation of Kimble life cycle, which looked quite intricate theoretically

8.4 What would we do differently, If we had to do it all over again?

- *Alteryx self service data analytics platform* is considered a market leader in the field of Data Analytics and has larger corporate acceptance than Pentaho. Hence, if we had to do the project all over again, then we would prefer Alteryx for the ETL process and would host the Data Warehouse in a cloud-based service.

- We used Oracle to host the Data Warehouse and exported CSV files to Tableau to do Visualizations remotely. A cloud-based Data Warehouse would be preferred as in the present case we were restricted by limitation of using Oracle only on campus, at the Baruch's computer labs.

8.5 Realization of proposed benefits to the new system

- ***The proposed benefits would be realized by the new system as we were able to create visualizations demonstrating the KPIs that we set out to find.***