

BI/ DW System Design Project with Professor Dr Turel, Spring 2017

**CREATING A PROPOSAL TO IMPLEMENT A SALE DATA BASED BI/DW
SOLUTION IN THE REAL ESTATE ORGANISATION 'HOMEARLY'**



ISDS 556 Project 2 by Team 8:

**Anush Pagadala, Ayush Upadhyay, Jerome
Meripo, Megha Nair, Prajakta Diwan,
Sanjana Raj**

SALES DATA BASED BI SOLUTION

PROPOSAL

to the

**Homearly Real Estate Group
3000 E Birch Street, Ste 102
Brea, CA, 92821**

Contents

OVERVIEW.....	3
HOMEARLY REAL ESTATE GROUP	3
Key existing systems:	3
High Level Enterprise Data Warehouse Bus matrix	4
FEATURES OF THE REQUIRED BI SYSTEM.....	5
Type of Reports and analyses:	5
DIMENSIONAL DATA MODEL.....	6
Fact Tables:	6
Dimension Tables:.....	7
DATA ANALYSIS	7
Available data.....	7
Data Source	8
CHALLENGES IN ETL PROCESS	8
Data Extraction:	8
Data Transformation:.....	9
RECOMMENDED ARCHETECTURE	10
PRELIMINARY BUDGET PLAN	10
SUMMARY, CONCLUSION, & RECOMMENDATION	11
REFERENCES	12

OVERVIEW

In today's digital economy, to make effective decisions, a real estate manager must be able to assess trends and view information to make effective and efficient decisions. All major real-estate companies aggregate massive amounts of data every day for potential customers, revenue, market potential and so on. Company executives also perform complex tasks to keep pace with the current dynamic market trends. Analyzing, and sharing information has become an inevitable task to achieve success. This proposal is to implement a sales data based business intelligence solution for 'Homearly'. The BI/DW solution proposed will provide an overview of the historical, current and predictive snapshot of the company's performance, enabling the company in making intelligent decisions aimed at improving bottom line results.



This proposal is to implement a sale based business intelligence solution for 'Homearly'. The BI/DW solution proposed will provide an overview of the historical, current and predictive snapshot of the company's performance, enabling the company in making intelligent decisions aimed at improving bottom line results.

A BI system will make sense of the immense amount of data that has been gathered by the company over past few years. It will help the company in making intelligent decisions aimed at improving bottom line results.

The BI/DW solution proposed by our team will help Homearly to better understand the trends in the real estate business. With a business intelligence solution in place, the company will be able to implement decisions to maximize profit and minimize risk.

HOMEARLY REAL ESTATE GROUP

Homearly, the real estate has been in the market for 10 years, helping people to buy and sell their homes with the best price available along with a dedicated trained staff. The company has been maintaining their website past few years to remain competent among its competitors in the real estate industry. The company understands the value of implementing a BI solution to excel in a highly competitive and volatile market. To stand out from its growing competition, the company needs to make efficient use of their limited resources.

Key existing systems

The company has a small-scale IT systems to store and retrieve basic property listings. A few types of reports are generated that are very general, such as period wise sales, pending inquiries etc. The company generates reports that displays data pertaining to home prices, sales, and local market listings to better inform the customer. But none of these reports leverage on modern BI capabilities. The company has systems/process for handling daily transactions and

report generation. However, most of it relies on spreadsheets/small scale apps to store and process information.

Over the years, the company has realized the need to add more capabilities to their current system. The top management is willing to invest for a good proposal and understands the need of educating its employees on this subject for a friendly acceptance. The requirement is to consolidate, clean, and prepare the existing operational data to extract knowledge and use it for decision making. BI reporting is required to carry out the following services:

- Monitor sales data and identify the type of property which is in demand.
- To identify the properties which sell out fast.
- To figure out the areas with stagnant real estate growth so that eventually the company could allocate its resources to those areas where it is more profitable.

High Level Enterprise Data Warehouse Bus matrix

	Time	Customer	Property	Agent	Finance	Marketing
Underwriting Transactions	×	×	×	×		
Agent's Commission	×		×	×		
Lead Generation (Sales History)	×	×				×
Revenue Generation (Sales)	×				×	

FEATURES OF THE REQUIRED BI SYSTEM

The process/KPI that will be in focus for this project will be Revenue. Revenue is a key component to monitor as it projects the growth of a company. Revenue has an inclusion of other KPIs like sale and customer satisfaction which reflects operational efficiency, marketing etc.

Type of Reports and Analyses

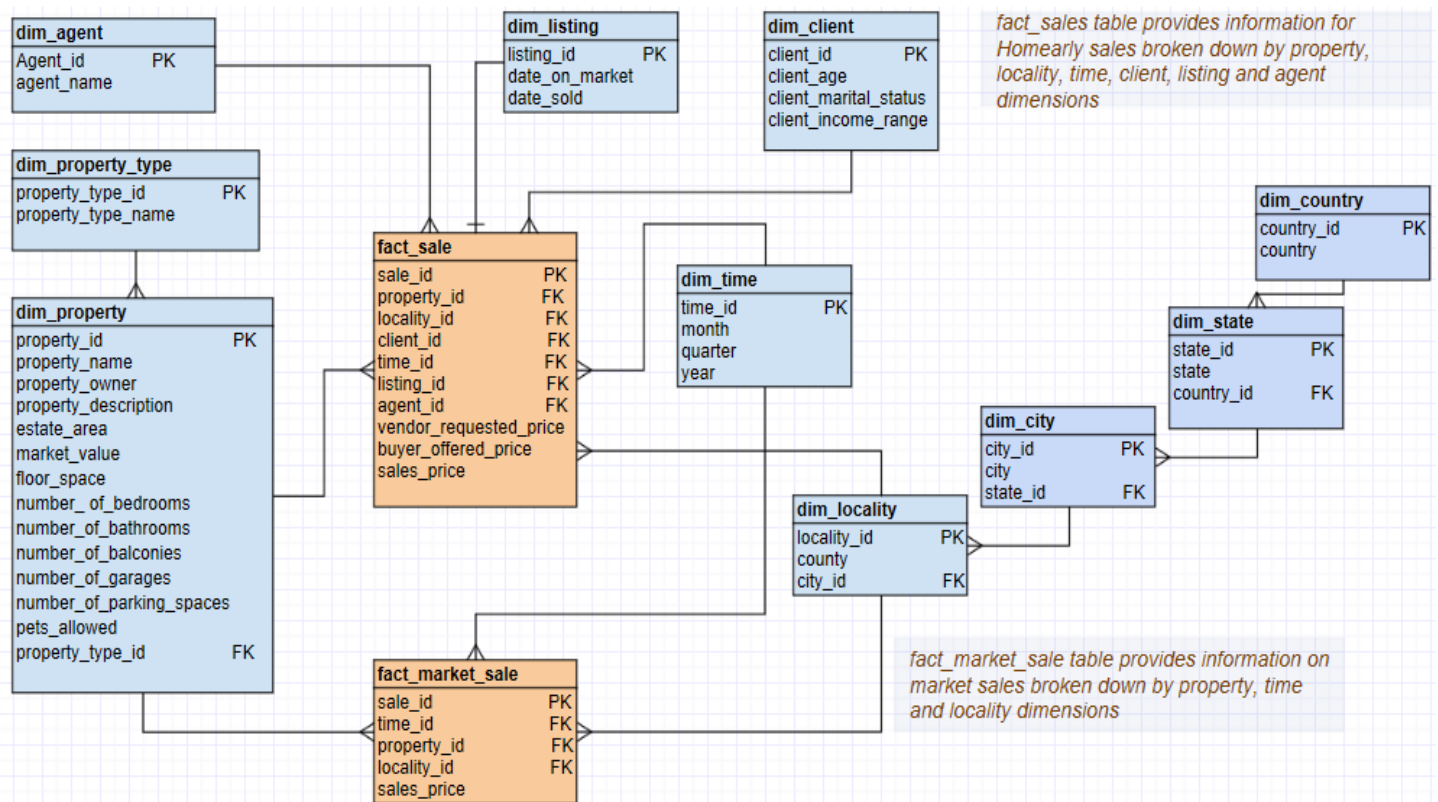
The project aims to undertake the following analysis and reports:

- **Generate reports on sales trend in market over a period** – Analysing market data to understand the movement of market and market challenges in real estate business.
- **Performance report of Homeearly generated by comparing sales trend in the market with Homeearly sales report.**
- **Generate report to identify risk and opportunities for company in properties and clients**, this report will be used by marketing team to formulate the marketing strategies. Generation of following kind of specific reports to identify risk and opportunities:
 - Detail report on the trend of type of properties which are more in demand in market over a period.
 - Sales pattern report by property types and customer categories (like age, marital status, income level etc.)

These reports can be used by top level executives, managers and associates to compare and understand the company's performance with respect to industry, to formulate future market strategies, to figure out stagnant growth areas, categorizing and targeting future clients etc. Reports can be generated on monthly, quarterly and yearly period.

Many different analytical tools can be used to extract data from the data warehouse for presentation and/or analyses such as SQL to query, Tableau for data visualization, R for data mining and other tools such as SPSS for statistical analysis.

DIMENSIONAL DATA MODEL



To ensure easy query retrieval and better performance the schema has some denormalized dimension tables (snow flaking). Snow flaking avoids data integrity problems and saves disk space but will not be used for all possible scenarios (like time dimension) as query performance will be impacted.

Fact Tables:

The design consists of two fact tables. Fact_sale table will hold information of each sale done by Homeearly. It consists of a primary key, an assigned surrogate key to uniquely identify each row and foreign keys to dimensional data. Granularity of a record per sales is chosen so that BI reports can drill down to that level of detail. Fact_market_sale table will hold information of historical sales in Homeearly's main operating geographical area.

- In the **Homeearly fact_sale table**, vendor requested price, buyer offered price and sale price are stored and grouped based on values in property, locality, time, client, listing and agent dimension tables.
- In the **fact_market_sale table**, sales price is stored and grouped based on property, locality, time, dimension tables.

Dimension Tables:

- **dim_property** and **dim_property_type**: These tables hold all the property data. This is a snowflake design as it will avoid data integrity issues and will save database space. The **dim_property** dimension table is split into two tables. The **dim_property_type** will hold property type information like Apartment/Real estate etc.
- **dim_locality** will hold sales property location in the granularity to county level. The normalization rules were applied on **dim_locality** dimension table to form snow flake schema. By this data is more structured, so data integrity problems are reduced.
- **dim_time** table hold time of sale transaction. Even though we have individual sale transaction recorded, time granularity is at month level as it will be more efficient to meet BI requirements. Model has a demoralized **dim_time** table. This reduces the number of query joins needed thus improves performance. Also, we won't lose a notable amount of the disk space, as all the table attributes and foreign key attributes share the integer type. Time dimension tables describes monthly as the lowest level of time granularity for which events are recorded in the star schema.
- The **dim_client** table holds client information. Basic client details required for analysis is included.
- The **dim_listing** table has property listing information. This can be used to identify if a property was slow moving in the market.
- **dim_agent** table stores basic agent details.

DATA ANALYSIS

Available data

To implement a sales data based business intelligence solution for Homearly, there are mainly two sources from where data will be extracted.



Historical market sales data is available from an external agency in excel spreadsheets. This consists of property sales in Orange county for a given period.

- The internal source of Homearly: Sales Data, listing data, customer data, employee data from the database of the **Homearly sales management system**.
- External source: Historical market sales data is available from an **external agency in the form of excel spread sheets**. Market sales data consist of property sales in

Orange county for a given period. Property data will also be sourced from external agency.

Data Source

For the analysis, data is collected from both Primary and Secondary sources.

- **Primary source** of data will be the sales management system's backend relational database.
- **Secondary source** of data will be the historical market sales information provided by external agency periodically in spread sheet format.

Data Cleansing

The data sourced from external agencies will be cleansed to remove corrupt and inappropriate values. When analysed by plotting a heat map, it was observed that the data requires imputation process and removal of records where imputation is not feasible. The irrelevant variables for the analysis will be deleted as part of cleansing.

CHALLENGES IN ETL PROCESS

The Extract, Transform & Load process is one of the crucial process in creating a data warehouse. Before the data can be loaded into the data warehouse, it should undergo three major processes, which are discussed in detail below:

Data Extraction:

The first step is to extract the data from the various relevant sources.

- From sales management system's normalized RDBMS:
Extraction of relevant data such as Sale, property listing, customer details and employee data will be done from the relational database of Homearely sales management system.
- Historical market sales data from spreadsheets:
Historical market sales data which consists of information on properties and sales over a period in the market will be collected from external agencies. This data will be provided periodically in the form of spreadsheets.

There are various extraction strategies which can be used for the extraction of the data. For the initial extraction, a Full Extraction strategy can be implemented in which all the data from the source files is extracted. From the second extraction, onwards a partial updating strategy or a partial extraction strategy can be implemented to extract only the data which has been modified, let it be updated, deleted or added. Updating of the data will be scheduled in such a way that it doesn't affect the regular usage of the database; i.e. the updating of the data can be scheduled every day during the night so to minimize the impact of database being locked.

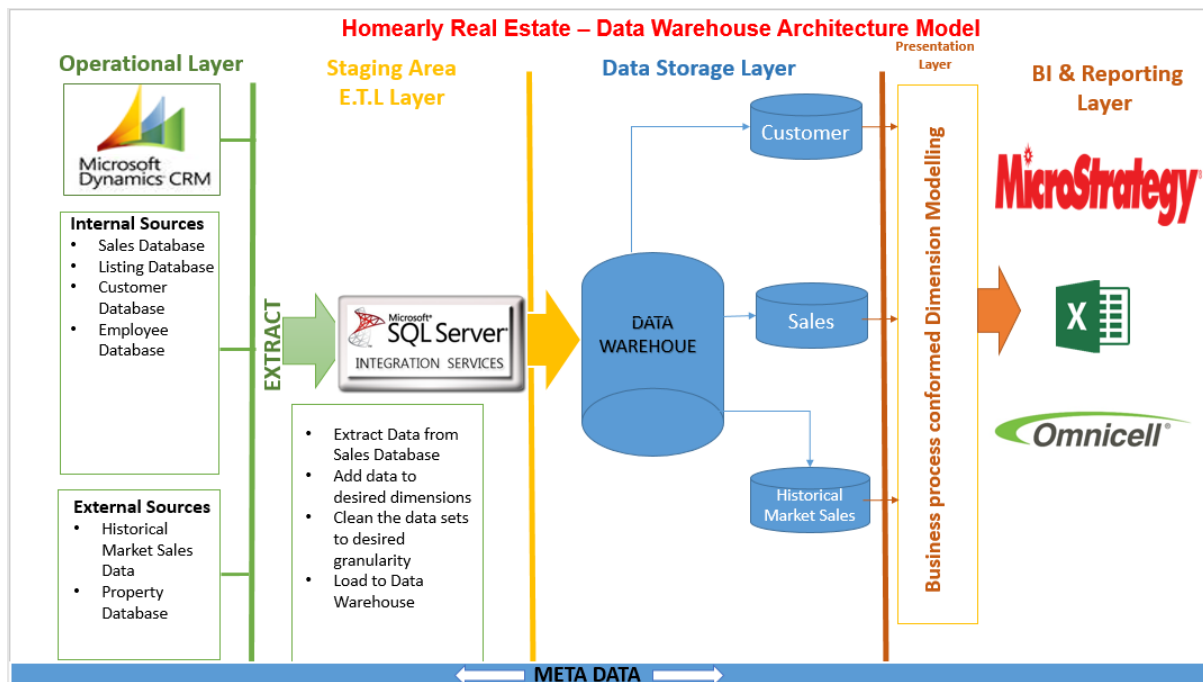
Data Transformation

The next step is to transform the extracted data. From all the data present within the source DB and files, only the relevant data from each source needs to be loaded. For example we are interested in working with the dimensions regarding Agent (which including Agent_id, & Agent_name), Listings (which Includes Listing_id, date_on_market, date_sold), Client (including Client_id, Client_age, Client_marital_status, & Client_income_range), Time (Including Time_id, Month, quarter, & year), Locality (Locality_id, County, City, State, & Country), Estate (Including Property_id, estate_area, & Market_value), House (including Property_id, floor_space, number_of_bedrooms, number_of_bathrooms, number_of_balconies, number_of_garages, number_of_parking_spaces, & pets_allowed), and Property (including Property_id, Property_name, Property_owner, & Property_description). As the source data are maintained by Homearely, the amount of irrelevant data is to a bare minimum from the beginning. We then match the data from different source files and then perform the cleaning process on the data, any null values may be replaced with terms such as “Unknown” or “Not Applicable”, data from different source file need to be changed into a similar format to ensure uniformity in data. Applying an internal search algorithm and using index or partitioning the table can help us reduce the query return time, we can partition the table as per the year and month for faster access. Several other transformation functions can be applied per our requirements such as deduplication where duplicate values can be omitted, summarization of the data, splitting or merging fields, etc. De-normalization should be used as we are using a database with a normalized architecture to shorten the query time.

Data Loading

The last step is to fetch the prepared or transformed data and load it into the database. The sales data extracted from the Homearely sales database is loaded to the fact table as one record per sale and time dimension in month level granularity. The historical market sales data is also loaded in other data mart in DW. Like the other processes, there are several types of loading function and an Initial Load function can be used to populate the data warehouse tables for the first time, and an Incremental Load function from the second time onwards to Load the ongoing changes as necessary in a periodic manner. The Loading process can be scheduled similarly to the extract phase, i.e. daily during the night time to prevent locking of the database while it is being updated. In between the extraction phase and the loading phase a buffer time should be present to ensure extraction is completed, only after the buffer time should the loading phase begin. We don't need to update our data in real time as the data we are dealing with isn't volatile and susceptible to frequent changes.

RECOMMENDED ARCHETECTURE



The operational layer will have data from sources such as sales, listing database, customer database which are internal sources of Homearely. The company also gets demographic regional data, property sales data from external data sources. The data from these sources will then be extracted, cleaned and loaded into the required dimensions using Microsoft SSIS. Depending on the type of reports needed the reporting layer will have applications supporting various forms of reports.

PRELIMINARY BUDGET PLAN

Implementing can be costly depending on how high you want to scale your data warehouse and what your business requirements are for the business intelligence project. We decided to take what we think is a conservative approach in setting up the necessary tools to implement a business intelligence solution. As the solution improves and as the importance of the solution is solidified with management, we can add more expensive components. Below is our preliminary budget plan that we believe will suffice the initial requirements for a business intelligence solution.

Product	# of Licenses	Price	Total
Microsoft Dynamics CRM	10	\$ 115.00	\$13,800.00
Microsoft SQL Server	-	\$ 3,717.00	\$ 7,434.00
Lenova All in One PC	10	\$ 799.99	\$ 7,999.90
Microsoft Office Professional	10	\$ 399.99	\$ 3,999.90
MicroStrategy	10	\$ 1,200.00	\$12,000.00
Design & Installation			\$10,000.00
Total Budget			\$55,233.80

SUMMARY, CONCLUSION, & RECOMMENDATION

Business intelligence allows businesses to make better decisions by accessing big data. Even small companies can derive better insights and benefits from better analysis. Homearly, the real estate company from Brea, California is a small-scale consulting group which helps the clients in buying and selling of houses. The company uses the traditional method to generate reports about sales, from its database. We have thus come up with the proposal for Homearly to implement the BI system, describing the KPI's, data analysis and the architecture.

The proposed solution is not just about gathering intelligence but about making sense of data in a way that can be useful to the management. Data which is now scattered across different systems/sources can be consolidated and made available to the management by the proposed BI solution. This is accomplished through processing the data and using visualization tools for generating informative reports. BI also offers dashboards, as well as performance scorecards. Key performance indicators and business metrics are much easier to understand when displayed as visualized data.

Adapting to the BI system will improve the business performance of the company. Some of the advantages include easier decision making, structured path for growing revenue, enhanced quality of customer service, edge over its competitors and improved efficiency of business operations. The company will have a better understanding of the sales with respect to the market value. It can make use of BI to predict the market value of houses in the coming years, based on the prices in the previous years and spot marketing opportunities. This will help generate better sales for the company.

Initially, only large companies would afford to adapt to business intelligence systems, but nowadays, technologies like Tableau, Saas (Software as a service), which are cost efficient business intelligence systems, are widespread in the market. We thus recommend Homearly to adapt to the BI system as explained in our report.

We recommend Homearly to review the proposal and suggest any changes required. Based on that, low level design can be done and resources should be procured to build the DW infrastructure.

REFERENCES

Drkušić, E.D. (2016). *The Snowflake Schema*. Retrieved 8 May, 2017, from <http://www.vertabelo.com/blog/technical-articles/data-warehouse-modeling-the-snowflake-schema>

Drkušić, E.D. (2016). *Star Schema vs Snowflake Schema*. Retrieved 8 May, 2017, from <http://www.vertabelo.com/blog/technical-articles/data-warehouse-modeling-star-schema-vs-snowflake-schema>

Foreman, E. (2017). Real Estate Business Intelligence - Steps to Success. Retrieved 8 May, 2017, from <https://www.realcomm.com/advisory/445/1/real-estate-business-intelligence-steps-to-success>

Wikipedia.org. (2017). Star schema. Retrieved 8 May, 2017, from https://en.wikipedia.org/wiki/Star_schema