End-to-end Analytics Data Pipeline Design for the Real Estate Market of Armenia

Areen Sipan

Contents

Abstract

As the volume of data provided by commercial real estate companies continues to expand, manual consumption and analysis is becoming increasingly complex. Integrating such data to the Business Intelligence environment is becoming crucial during this age of technology. The real estate market of Armenia is no exception to this. The aim of the project is to create an efficient data warehouse for storing the real estate listings in Armenia, which would make it accessible for the real estate agencies to analyze the data, generate reports, and create easily interpretable dashboards. For bringing this project to life, data was gathered by scraping two Armenian real estate websites and an appropriate database architecture was designed for storing the data. Furthermore, a sample dashboard was created to showcase the efficiency and accessibility of the aforementioned database. The idea is to provide you a genuine picture of the market. The goal is to develop this product into a workable concept that demonstrates how to make access to information about real estate easier, quicker, and, most importantly, efficient.

Keywords: Business Intelligence, real estate, database design, data warehouse

Introduction

Integrating Business Intelligence tools to a company's workflow has already proven to be beneficial in a variety of areas, ranging from healthcare to banking. The commercial real estate industry can also utilize such tools to their benefits in order to find and capitalize on new prospects. As the data generated by commercial real estate agencies continues to grow to immense volumes, it is becoming extremely difficult to consume and analyze it manually. This issue can be addressed by processing, storing, and integrating these data streams into actionable insights by using Business Intelligence and other programming tools. This topic is quite appropriate for the Armenian real estate market, as throughout the recent years it has been growing steadily, and, as of Grant Thornton's report, the real estate transactions reached their all-time maximum in 2018, passing the threshold of 160,000 (2019). Although the situation changed, due to the ongoing pandemic, the Artsakh war, and the political crisis, resulting in a 15% decrease in transactions, nevertheless, growth is anticipated, as in December of 2020, transactions increased by 75.7% compared to the previous month (Grand Thornton Armenia, 2020). The main aim of this capstone project is to create an efficient data warehouse, where the real estate listings of Armenia can be stored and to integrate it to the Business Intelligence environment. The project is going to be created by having the real estate agencies of Armenia in mind, that is to say, it is going to make it much more accessible for them to work with data, create easily interpretable descriptive and analytical dashboards, analyze them, and use in decision-making.

Approach

The project would consist of three main phases - data collection, data warehouse creation, and dashboard. Data collection was done via preiodical scraping the listings from the real estate websites of Armenia. The websites used were myrealty.am and bars.am. These were chosen because they are some of the biggest ones with the most amount of data. Moreover, they cooperate with other small real estate agencies and independent agents and have their database lited on their website as well. All the relevant attributes of the listings were gathered, in order to have a well-rounded picture of the data that needs to be worked on, such attributes include price, area, floor, number of views, location, available facilities, etc. The next step was cleaning the data, which after some filtering was brought to an appropriate structure for analysis. Here, it is extremely important to adhere to the well-known 5 Cs of the data to make sure it is clean, consistent, conformed, current, and comprehensive (Sherman, 2015). This stage is the one responsible for delivering a clean, and comprehensive dataset to the database. The issues of consistency, conformity and currency will be solved in the next phase, while designing an appropriate database architecture. Furthermore, this is the stage when the outliers should be identified. Python was the tool used via Jupyter Notebook for both web scraping and data filtering. The second phase of the project was data engineering; this entailed coming up with an appropriate data architecture for building the database. This meant understanding what dimensional tables is it necessary to have, what attributes should be included in the fact table and what method is the most suitable for dealing with slowly changing dimensions. In order to answer these questions, all the extracted attributes were taken into account to come up with a design that would provide the most efficient way of storing the data. It is noteworthy, that the design is scalable and can accommodate data collected from different websites. The scraped data was then pushed to the data warehouse created by using dimensional modelling. The main tools utilized for this step were Microsoft SQL Server Management Studio and Azure Data Studio, along with Python, for creating automated functions that would push the extracted data to the warehouse. The third phase was creating the dashboard. The dashboard includes descriptive visuals that help to better interpret the data, accompanied by analytical visual, that could be useful to the real estate agencies for decision-making. The famous corporate tool - Microsoft Power BI Desktop was used for this phase.

Solution/Deliverables

The data manipulation and data scraping were somewhat combined to one another, to make sure that the extracted data was as clean as possible. This meant making sure the scraped format was correct, and the types were transformed. For the data manipulation part, the unique values were considered to see if there were any unreasonable values and if the data was scraped properly. Moreover, in some cases the quantiles were calculated to see if there were extreme outliers and be able to deal with them in the future. This step made sure the data was as comprehensive as possible.The latter was somewhat challenging, as there were a lot of unreasonable attribute values (e. g. houses with 25 storeys, apartments with 1 sqm area, etc.), which had to be dealt with individually for differentiating the ourliers from the errors. Although there were some doubts in the beginning about combining transactional and dimensional modelling, in the end it was chosen to use only dimensional. That is because “BI applications and BI-related data stores, [. . .] are better served with dimensional data models. BI tools will give the best performance with star or snowflakes in querying the relational database” (Sherman, 2015). Furthermore, dimensional models make it intelligible to generate clear and effective reports with easily interpretable dashboards. As for the database architecture, it was decided to use the snowflake schema. Its main characteristic is that it the fact table is surrounded by normalized hierarchies within a particular dimension, where “each level in the dimensional hierarchy becomes its own dimensional table with parent keys created to link the hierarchical structure together” (Sherman, 2015). One of the main reasons why this particular schema was chosen was because the data has dimensions with a large number of rows and with deep hierarchies, which are relatively static. The dimensions in their turn, were constructed by taking into account the accessibility of filtering and analyzing the data based on the attributes we have. In the cases of storing the locations of the listing and the dates, we have balanced dimensions. A separate date dimension was also constructed since our fact table has a date column. As for the fact table, it is a transaction non-additive table. From the former, it can be deduced that we have high granularity, as well. Since most of the data is relatively static, the most common SCD type in this project is type 0 – that is storing the original values. This is because most attributes of the listings, such as the address, the area, or the floor it is located on, do not change. However, there are some features that may change during time, for example the condition of the building or the number of rooms after renovations. A similar project was done about developing a data model for the city of Lisbon (Pereira, 2021). In these project SCD type 2 was decided to utilize for storing all the historical data, however, this would greatly slow down the query process and would require extra efforts for dealing with listings that cease to exist after some time. For this project, it was decided that SCD type 7 (SCD type 4 for some sources) provides the best format for storing them – by efficiently combining SCD type 1 and 2. This structure gives the opportunity to both “track historical data in type 2 processing and get type 1 query performance” (Sherman, 2015). Aside from providing accessibility, this particular structure helps deal with expired listings. If a listing exists in the historical table but is absent in the ongoing one, then it can be said that it is expired and was taken out of the website. After coming up with the design, an initial visualization was done via dbdiagram.io, which helped to get a better picture of the data warehouse to be created. The dimensional and fact tables were then constructed accordingly. Aside from these two, a staging table was also constructed, to have an intermediary area for storing the data before actually integrating it to the warehouse. In order to smoothly push the data from the staging area to the dimensional tables, several merge operations were done. The final step was testing the created architecture by pushing the filtered and clean data to the staging table and then to the warehouse. The third phase of the project was to build sample and quite simple dashboard. The dashboard mainly highlights information about the prices and popularities of listings. The goal for creating this dashboard was to showcase how and why this database can be used by the real estate agencies.

Further Improvements

It is important to note that only two websites were used to scrape, so for further improvements more websites could be included to create a bigger database, in order to show the full picture of the real estate market of Armenia. Moreover, some analytical components could be added by fitting appropriate machine learning models on the existing data for predicting prices and popularity (from the number of views) of the listings. The dashboard could also be improved by adding analytical visualization from those models, along with some informative geographic ones, since the coordinates of the listings are also stored in the database. It can then be further modified based on the business requirements of each of the real estate agencies. Another aspect to improve this project would be deploying the database, which was note done due to time restrictions. There is an intermediary company in Armenia called Dignisi, which cooperates with different real estate agencies, in order to combine all real estate listings of Armenia and create a database sponsored by teh cadastre of Armenia. After a meet up with the head of the company, I realized that the project could help them by providing an end-to-end data pipeline design for proper storing the data they have. This capstone project is simply a smaller simulation of a big tool that could help revolutionize the real estate market of Armenia and has huge prospects if enough time and effort is put into it.

Git-Hub Repo link: https://github.com/AreenSipan/Capstone.git

References

Grant Thornton Armenia. (2019). Real Estate Market in Armenia. granttphonton.am. Retrieved February 2, 2022, from https://amcham.am/wp-content/uploads/2019/07/GT-_-real-estate-market-analysis.pdf

Grant Thornton Armenia. (2020). Real Estate Market in Armenia 2019-2020. granttphonton.am.Retrieved February 2, 2022, from https://www.grantthornton.pr/globalassets/1.-member-firms/puerto-rico/publications/global-transparency-report-2018.pdf

Pereira, A. S. S. R. (2021). Real Estate: Developing A Data Model For The City Of Lisbon To Improve Information Sharing (dissertation). NOVA Information Management School Instituto Superior de Estatística e Gestão de Informação, Lisbon.

Sherman, R. (2015). Business intelligence guidebook: From Data Integration to Analytics. Elsevier, Morgan Kaufmann is an imprint of Elsevier.