**Group One**

**Development Team Project: Project Report**

**September 18, 2023**

**A Design Report for Building a Relational Database for Manhattan Real Estate Listings**

**Module:** Deciphering Big Data

TABLE OF CONTENTS

[Introduction 1](#_Toc145790199)

[Objectives 2](#_Toc145790200)

[Data Management Pipeline 4](#_Toc145790201)

[Logical Design 5](#_Toc145790202)

[Conclusion 7](#_Toc145790203)

[References 8](#_Toc145790204)

**TABLE OF FIGURES**

[**Figure 1: Data Management Pipeline 4**](#_Toc145790332)

[**Figure 2: Entity Relationship Diagram 5**](#_Toc145790333)

**Introduction**

In 2016, the real estate sector claimed the top spot as the world's largest and most valuable asset (Andonov et al., 2013). The collective value of various real estate properties across the globe amounted to USD 217 trillion (Barnes et al., 2016). To put this into perspective, it surpassed the Gross Domestic Product (GDP) of the United States by a factor of 11 and China's GDP by 19. This revelation underscores the pivotal role that real estate plays not only as an individual sector but also in the global economy. It’s not just bricks and mortar, it’s an economic cornerstone.

In this digital era, real estate has witnessed profound transformation. Property listings, once confined to print publications and physical storefronts, now thrive in digital spaces (Bond T. et al., 2000). Online platforms have emerged as the epicenter of property discovery connecting buyers, renters, agents, and investors with an unparalleled array of real estate offerings. This technical capability has, in turn, ushered in an era of data abundance and proliferation.

The real estate industry now stands as a data-driven frontier, with big data emerging as the very lifeblood of its operations. This data deluge brought about by big data presents both an extraordinary opportunity and formidable challenge. It’s an opportunity because big data, when harnessed effectively, has the power to unlock new dimensions of understanding, facilitate informed decisions, and catalyze growth (Oluwunmi et al., 2019). Yet, it’s also a challenge because the sheer volume, velocity, and variety of data demand a strategic approach to capture, manage, store, and utilize it effectively (Munawar et al., 2020).

Using Zillow, a real-estate and rental marketplace, as the primary data source, this report delves into the design of a logical database that harmonises real estate for sale listings data from Manhattan, New York.

Objectives

The design and development of the logical database is driven by a set of clear and strategic objectives. These objectives are twofold: to capture the essence of real estate data in the era of big data and to construct a database that serves as a catalyst for informed decisions in the dynamic Manhattan real estate market. To achieve these objectives the following key goals have been outlined:

1. To maintain high data quality and integrity through data cleaning and normalization, ensuring that all information presented in the database is accurate. Exhaustive data cleaning will be applied to ensure the removal of inconsistencies and errors. Additionally, data normalization techniques will be employed to minimize data redundancy and enhance data integrity (Beaubouef et al., 2005).
2. To provide a comprehensive repository for real estate listings in Manhattan, encompassing diverse property types, amenities, and price ranges. The primary aim is to empower users with a rich source of property information by offering a wide spectrum of listings, from apartments to townhouses. This also includes detailed property descriptions, like square footage, the number of bedrooms and bathrooms, and available amenities, with the aim to provide a holistic view of each listing.
3. To offer comprehensive data on real estate agents and brokers associated with each listings. This is so as to enhance user-agent interactions, providing users with direct access to expert guidance.
4. To empower users with historical price data for each property listing, enabling the assess of market trends and property value changes. The price history entity in the database would allow users to track price trends over time. By accessing historical pricing information, users can make informed decisions about property investments and gain insights into market dynamics.

Data Management Pipeline

A close-up of a computer screen

Description automatically generated

Figure : Data Management Pipeline

Environmental Setting up:

**Python Installation:**

All the necessary libraries for data wrangling and scraping were installed together with Python (Python Software Foundation**)**. Python was our option for this project because it is simple to use and provides a wealth of modules for manipulating data with little code (Kazil & Jarmul, 2016)

**Setup of the MySQL database:**

A local MySQL instance was setup, and using MySQL Workbench, three data tables were first constructed to contain the scraped information, including the property listing, the property price history, and the agent details(Letkowski,2015).

**Web Scraping:**

A real estate listings website called Zillow was used to gather information for this project. In order to control data queries and extractions from the website to the local system, the zillow web scraping API from Scrapeak was utilized. In JavaScript Object Notation (JSON) format, Scrapeak provides systematic and reliable data retrieval (Glez-Pea et al., 2014).

Using Python libraries, the JSON data from the data request and response is then converted to DataFrames. Python DataFrames are tabular and straightforward which makes the process of wrangling data easier (Databricks, 2023).

In order to obtain price history data, we created a Python script that iterates over every record in the property listing to retrieve historical data on property prices i.e. a request to the API is processed for each zpid in order to retrieve the relevant pricing history data.

**Data Storage:**

The MySQL Database Management System was linked with the Python environment using the mysql-connector-python module to store the scraped real estate data(Vishal,2021).

***Methodology as discussed in today’s meeting Oct, 10th 2023***

**Data collection**

* **Web Scraping:**
* **API**
* **Zillow website**
* **Python**

**Data cleaning**

* **JSON transformation**
* **Python Dataframes**
* **Data type & conversion e.g. zpid**
* **Dropping unnecessary columns**
* **Dropping missing data**
* **Data transformation e.g. columns with dictionary like entries**

**Data integration**

* **Data storage**
* **MySQL DBMS**

**Database design (logical and Physical)**

* **ERD (MySQL workbench)**
* **Normalizations**

**Data Analysis**

* **TBD**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The data management pipeline for the database commences with the acquisition of data from the real estate listings website: Zillow, facilitated by Scrapeak, a web scraping API. The utilisation of Scrapeak offers structured and consistent data retrieval in JavaScript Object Notation (JSON) format (Glez-Peña et al., 2014). Subsequently, data transformation is undertaken to convert the JSON response to a DataFrame. This is done to ease the data cleaning process because of the tabular and simple structure (Databricks, 2023).

The data cleaning stage is closely integrated with data validation, which is executed to verify the accuracy and integrity of the acquired data. This dual process ensures that the data within the database is not only cleaned but also validated for reliability, thus upholding data quality standards.

Facilitated by a database connector in Python, the data is dumped into an initial database that contains a single table encompassing all the acquired information. However, this initial structure is a precursor to the subsequent vital phase of normalisation.

Normalisation involves breaking down data into smaller, related tables to optimize data organization by reducing data redundancy, inconsistency, and maintaining of the atomicity within a database relation (Kumar & Kumar Azad, 2017). The normalisation stage is imperative due to the choice of the relational database model, with MySQL as the database management system. The relational database model is suited for modeling real-world relationships between the entities, such as properties linked to agents, agencies, and price history (ElDahshan et al., 2022). To achieve this relational structure, normalisation is undertaken using the Entity Relationship Diagram (ERD) as the blueprint.

Logical Design

The logical design of the database is illustrated by an ERD that visually represents the entities, attributes, relationships, and data types.

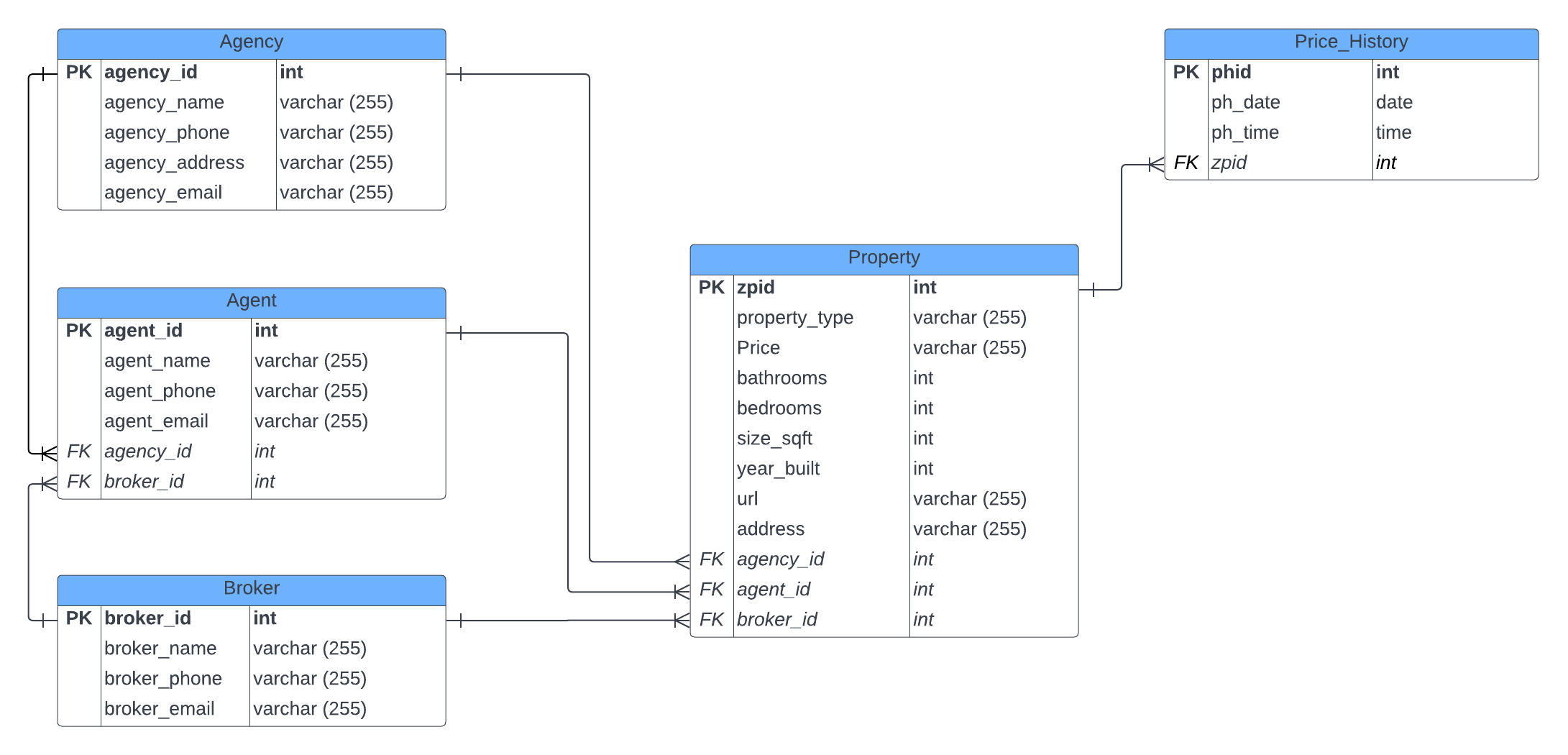


Figure : Entity Relationship Diagram (Harrington, 2016)

The “Property” entity serves as the central hub for the real estate listings database. This entity is linked to several others, including "Agent," "Agency," "Broker," and "Price History." Through these relationships, the association between properties, the professionals who manage them, and the historical pricing information is captured.

Each entity is equipped with specific attributes that define and describe them. For instance, "Property" includes attributes like zpid (Zillow Property ID), address, property type, size, bedrooms, bathrooms, and price. Similarly, "Agent" feature attributes such as agent ID, name, and contact details.

To ensure data accuracy and integrity, the appropriate data types and formats for each attribute are defined (Toerey et al., 2010). Numeric fields, such as property square footage and price data, employ suitable numerical formats. Textual data, including names and addresses, adhere to predefined character limits. Dates are stored in standardized date formats to facilitate accurate temporal analysis.

Furthermore, primary keys are established within each entity to uniquely identify records, ensuring data integrity by preventing duplication (Bopp et al., 2019). These primary keys serve two purposes, to not only uphold data integrity but also form the foundation for connecting data across different entities. This connection is achieved by creating references to the primary keys through foreign keys, establishing relationships that enhance the overall database structure.

Conclusion

This report sets the stage for the development of the real estate for sale listings database. It outlines objectives, data management pipeline, and logical design. It emphasizes data accuracy and integrity. This is in order to offer rich, reliable information for users, enhancing their understanding of Manhattan's real estate market and enabling informed decisions.

References

Andonov, A., Kok, N., & Eichholtz, P. (2013). A Global Perspective on Pension Fund Investments in Real Estate. The Journal of Portfolio Management, 39(5), 32–42. Available from: <https://doi.org/10.3905/jpm.2013.39.5.032> [Accessed 4 September 2023].

Barnes, Y., Tostevin, P., & Tikhnenko, V. (2016). Around the world in dollars and cents.

Beaubouef, T., Petry, F. E., & Ladner, R. (2005). Normalization in a Rough Relational Database (pp. 275–282). Available from: <https://doi.org/10.1007/11548669_29> [Accessed 14 September 2023].

Bond T., M., Seiler J., M., Seiler J., V., & Blake, B. (2000). Uses of Websites for Effective Real Estate Marketing. The Journal of Real Estate Portfolio Management, 6(2), 203–210.

Bopp, C., Benjamin, L. M., & Voida, A. (2019). The Coerciveness of the Primary Key. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–26. Available from: <https://doi.org/10.1145/3359153> [Accessed 16 September 2023].

Databricks. (2023, June 8). Tutorial: Work with PySpark DataFrames on Databricks | Databricks on AWS. Available from: <https://docs.databricks.com/en/getting-started/dataframes-python.html> [Accessed 15 September 2023].

ElDahshan, K., Selim, E., Ismail Ebada, A., Abouhawwash, M., Nam, Y., & Behery, G. (2022). Handling Big Data in Relational Database Management Systems. Computers, Materials & Continua, 72(3), 5149–5164. Available from: <https://doi.org/10.32604/cmc.2022.028326> [Accessed 6 September 2023].

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. Briefings in Bioinformatics, 15(5), 788–797. Available from: <https://doi.org/10.1093/bib/bbt026> [Accessed 12 September 2023].

Harrington, J. L. (2016). Entities and Relationships. In Relational Database Design and Implementation (4th ed.). Morgan Kaufmann.

Kumar, K., & Kumar Azad, S. (2017). Relational Database Normalization under Tabular Approach: A Design Methodology. International Journal of Advanced Research in Computer Science, 8(5), 2160–2164. [www.ijarcs.info](http://www.ijarcs.info)

Munawar, H. S., Qayyum, S., Ullah, F., & Sepasgozar, S. (2020). Big Data and Its Applications in Smart Real Estate and the Disaster Management Life Cycle: A Systematic Analysis. Big Data and Cognitive Computing, 4(2), 4. Accessed from: <https://doi.org/10.3390/bdcc4020004> [Accessed 8 September 2023].

Oluwunmi, A. O., Role, B. A., Akinwale, O. M., Oladayo, O. P., & Afolabi, T. O. (2019). Big Data And Real Estate: A Review Of Literature. Journal of Physics: Conference Series, 1378(3), 032015. Available from: <https://doi.org/10.1088/1742-6596/1378/3/032015> [Accessed 8 September 2023].

Toerey, J. T., Lightstone, S. S., Nadegu, T., & Jagadish, H. V. (2010). Database Modelling and Design (4th ed.). Morgan Kaufmann.

**Python (2023). Python Software Foundation. available at:** <https://www.python.org/> **(Accessed: 8 October 2023).**

Kazil, J. and Jarmul, K. (2016). Data Wrangling with Python. O'Reilly Media

Letkowski, J. (2015). Doing database design with MySQL. Journal of Technology Research. Volume 6. ResearchGate

Python Programming

Vishal, H. 2021. Python MySQL Database Connection using MySQL Connector. PYnative. Available from: https://pynative.com/python-mysql-database-connection/ [Accessed on 08 October 2023]