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).

The price history data was obtained using a Python script that iterates over every property id to fetch the corresponding price history data on property prices i.e. a data request is made to the property listing API(scrapeak) for every id.

**Data Transformation.**

Every data request made to the API gets returned with a JSON result. This data is transformed into a dataframe using the Pandas module in Python. Data wrangling is quick and easy thanks to the extensive abstraction provided by the Pandas library during data processing (Wes McKinney,2013).

**Data cleaning**

After conversion of JSON to Dataframes, inconsistencies and errors were identified in the property listing and price history data. The following steps were undertaken to correct the errors in the data.

* Handling Missing Values

In order to enforce data integrity, missing values especially for zpid were excluded from the dataset.

* Removing Duplicates
* Data Formatting

This was done to ensure consistency in data formats, such as date formats, numbers and strings data.

* Text Data Cleaning

Text data with inconsistent data was trimmed accordingly to bring about consistency

**Data Storage:**

The cleaned dataset was stored in MySQL, a relational data management system using sqlalchemy module and MySQL connector.

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

* **Findings**

Web scraping is one of the most effective ways to mine data in e-commerce, and in this study, we were able to gather a sizable amount of real estate data and store it in a MySQL database for analysis (Henrys, 2021).

Scraping Zillow property listing data has provided valuable insights into the real estate market trends, pricing, and property characteristics (Khder,2021).

With the integration of Python, JSON, Pandas, MySQL, and scrapeak API, we were able to collect, process, and analyze a lot of real estate data.

Database use cases:

* + - Price Analysis

Using historical data from the price history table in MySQL, you can analyze the trend of property prices over time. Visualize price changes for different types of properties (houses, apartments) and in different areas.

To achieve this, the database stores property historical data in MySQL, and uses Pandas for data manipulation and visualization.

* + - Property Features Analysis:

By examining detailed property descriptions data in the property table, investors in real estate can identify the most common features of properties (e.g. the number of bedrooms, baths, and amenities) on their impact on home prices.

* + - Location-based Insights:

The database stores real estate Geolocation data were users can analyze property prices based on different neighborhoods or proximity to amenities (schools, parks, public transport). Identify areas with the highest and lowest property values.

* + - Predictive Modeling:

Using historical property prices, detailed property features and location data. It is possible to build predictive models to forecast property prices based on historical data, features, and location. Property investors can evaluate the accuracy of the models to make informed predictions.

* Conclusion

We easily integrated real estate data from the Zillow website using web scraping, Scrapeak API, Python, Pandas module, and MySQL, a Relational Database Management System (RDMS). These solutions have shown to be strong and efficient ways to collect, transform and deploy a comprehensive database of property listings from Manhattan (Jarmul & Lawson, 2017).

By leveraging the synergy of these technologies, we have been able to develop scripts that can automate the process of gathering a wide variety of property listing information as well as evaluate and derive valuable insights that are essential for making well-informed real estate investment decisions (El Asikri et al.,2020)

Key Achievements:

Data Aggregation and Integration:

We gathered a variety of up-to-date data for Manhattan region from the Zillow website using web scraping with Python and Zillow Scraper. A comprehensive perspective of the Manhattan real estate market that takes into account property prices, features, locations, trends, and demand dynamics has been made possible by this database.

Data Quality and Reliability:

We ensured the data's correctness and consistency by using Python for data cleaning and processing. We handled missing values, outliers, and inconsistencies using Pandas cleaning techniques, which enhanced the database's consistency. (McKinney,2012).

Data Storage and Management:

Utilizing MySQL as our RDMS allowed for effective property data management, storage, and retrieval. The database's relational structure supported sophisticated analytical processes by ensuring data integrity and facilitating easy querying (Myers & Copeland, 2015).

Data Analysis and Insights:

Pandas played a pivotal role in exploring and analyzing the data. Various statistical analyses, trend evaluations, and predictive modeling were performed to derive actionable insights. These insights, ranging from market trends to rental yield calculations, provided invaluable information for decision-making processes.

* Recommendations

Scalability and Future Prospects:

With tables connected together, relational database design techniques have been used as the basis for database architectural designs that make it simple to integrate new features and data from new sources. It will be worthwhile to investigate adopting the NoSQL database in the future to ensure scalability as it can provide superior scalability and Data retrieval (Rautmare & Bhalerao,2016).

However, the safety of the data that is being stored must be guaranteed. To protect sensitive real estate information, encryption, access limits, and frequent security audits are required. It is imperative to give security careful thought before implementing a non-relational database system. Some NoSQL systems lack important security protections like encryption and weak authentication processes as compared to MySQL (Okman et al., 2011).

Ethical and Legal Compliance:

It is essential to follow moral guidelines and legal standards while obtaining data from websites and APIs. To ensure data confidentially and to comply with the General Data Protection Regulations, some data elements in the database used for property listings, such as phone numbers and names, have been de-identified. Online scraping is not technically prohibited by law, but it is important to follow data usage guidelines and regularly review the terms of service for the Zillow API in this context (Krotov & Silva,2018).

User Experience and Interface:

The creation of user-friendly tools and interfaces that enable seamless interaction between stakeholders and the database should be taken into consideration. It is possible to connect visualization libraries to provide data insights in a clear manner. Python offers a wide variety of frameworks, including Django, Flask, Pylon, and others. Python is frequently employed in web development (Saabith et al.,2019)

In summary, a robust, reliable. and informative property listing database for the Manhattan area has been created thanks to the collaboration of real estate data aggregation techniques, Python programming, MySQL database management, and Pandas data analysis. This database benefits researchers, investors, and real estate professionals alike by providing an invaluable tool for market analysis and providing the basis for informed decisions (Jarmul & Lawson, 2017). The database has the ability to transform how real estate investors view and interact with the dynamic world of real estate business through constant updates.

Review concepts underlying database modelling and

Critically evaluate the strengths and weaknesses of the relevant data models used in your design and build

Your recommendations should take into account the legal and compliance requirements applicable to the Organization.

You will be submitting a report that informs your client of the outcomes of your analysis and evaluation of your choice of data model highlighting strengths and limitations.

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]

Khder, M.A., 2021. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing & Its Applications*, *13*(3).

Henrys, K., 2021. Importance of web scraping in e-commerce and e-marketing. *Available at SSRN 3769593*.

Jarmul, K. and Lawson, R., 2017. *Python Web Scraping*. Packt Publishing Ltd.

McKinney, W., 2012. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".

El Asikri, M.S., Knit, S. and Chaib, H., 2020. Using web scraping in a knowledge environment to build ontologies using python and scrapy. *European Journal of Molecular & Clinical Medicine*, *7*(03), p.2020.

McKinney, W., 2012. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".

Myers, J. and Copeland, R., 2015. *Essential SQLAlchemy: Mapping Python to Databases*. " O'Reilly Media, Inc.".

Rautmare, S. and Bhalerao, D.M., 2016, October. MySQL and NoSQL database comparison for IoT application. In *2016 IEEE international conference on advances in computer applications (ICACA)* (pp. 235-238). IEEE.

Okman, L., Gal-Oz, N., Gonen, Y., Gudes, E. and Abramov, J., 2011, November. Security issues in nosql databases. In *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 541-547). IEEE.

Krotov, V. and Silva, L., 2018. Legality and ethics of web scraping. Available at <https://www.researchgate.net/profile/Vlad-Krotov/publication/324907302_Legality_and_Ethics_of_Web_Scraping/links/5aea622345851588dd8287dc/Legality-and-Ethics-of-Web-Scraping.pdf> [Accessed October 20, 2023]

Saabith, A.S., Fareez, M.M.M. and Vinothraj, T., 2019. Python current trend applications-an overview. *International Journal of Advance Engineering and Research Development*, *6*(10).

Available at <https://www.academia.edu/download/61390005/IJAERDV06I108548120191201-27336-1cr1jto.pdf> [Accessed October 20, 2023]

*Code sample to retrieve property history data*

