Data Management Pipeline

A close-up of a computer screen

Description automatically generated

Figure 1: 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).

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]

*Code sample to retrieve property history data*

