



# **A Case Study on Innovative Data Integration Strategies for Real Estate: Building ETL Pipelines with Airflow**

By Dr. Ehibhahiem Nelson Ughele

# Executive Summary for Real-yo Realtor



- Real-yo Realtor, a market leader in the real estate industry, recognizes the pressing need for technological innovation to stay ahead in a highly competitive market. The Executive Summary for Project AIR-REAL encapsulates the essence of the initiative, emphasizing the strategic shift towards leveraging data integration for improved operational efficiency and informed decision-making.
- By adopting innovative ETL pipelines with Apache Airflow, Python, and AWS services, Real-yo Realtor aims to revolutionize its workflows, setting the stage for a more responsive and data-driven approach to real estate management.

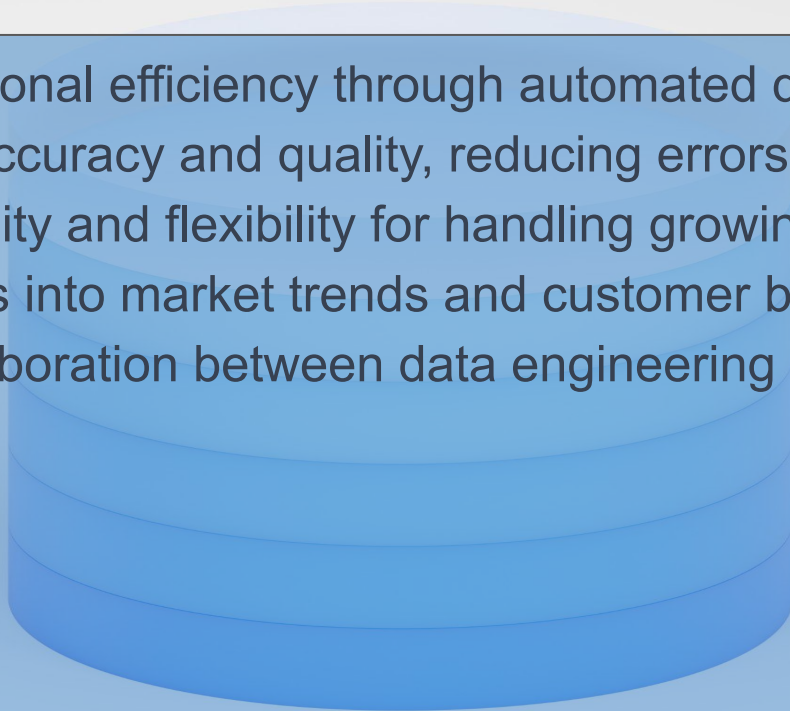
- The current state of Real-yo Realtor's data infrastructure poses significant challenges to the organization's agility and responsiveness. Manual data handling processes result in time-consuming workflows, leading to delays in decision-making. Disparate data sources contribute to inconsistencies and inaccuracies, hindering the reliability of information crucial for strategic planning. The absence of a robust ETL solution further exacerbates these issues, limiting the organization's ability to harness the full potential of its data assets.
- Real-yo Realtor's business problem statement can be summarized as follows:
  - Manual Workflow Bottlenecks: Current processes heavily reliant on manual data handling, leading to inefficiencies and delays.
  - Disparate Data Sources: Data spread across various platforms and databases, causing challenges in consolidating and harmonizing information.
  - Decision-Making Impacts: Inaccuracies and delays in obtaining critical data impacting the ability to make timely and informed decisions.
  - Lack of Automation: Absence of an automated ETL solution hindering scalability and adaptability to changing data volumes and sources.

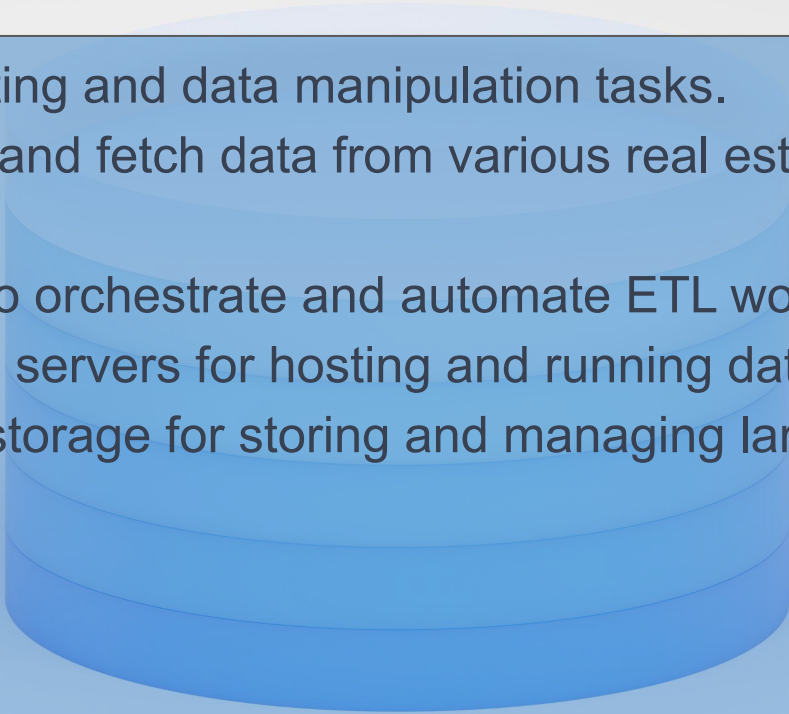
# Objectives (For Data Engineers)

- Develop and implement ETL pipelines using Airflow to automate data workflows.
- Integrate data from diverse sources such as property listings, market trends, and customer interactions.
- Ensure real-time or near-real-time data processing for timely decision-making.
- Improve data quality, accuracy, and reliability through transformation processes.
- Establish a scalable and maintainable data integration architecture.

# Benefits (For Data Engineers)

- Increased operational efficiency through automated data workflows.
- Enhanced data accuracy and quality, reducing errors in decision-making.
- Improved scalability and flexibility for handling growing data volumes.
- Real-time insights into market trends and customer behavior.
- Streamlined collaboration between data engineering and other business units.



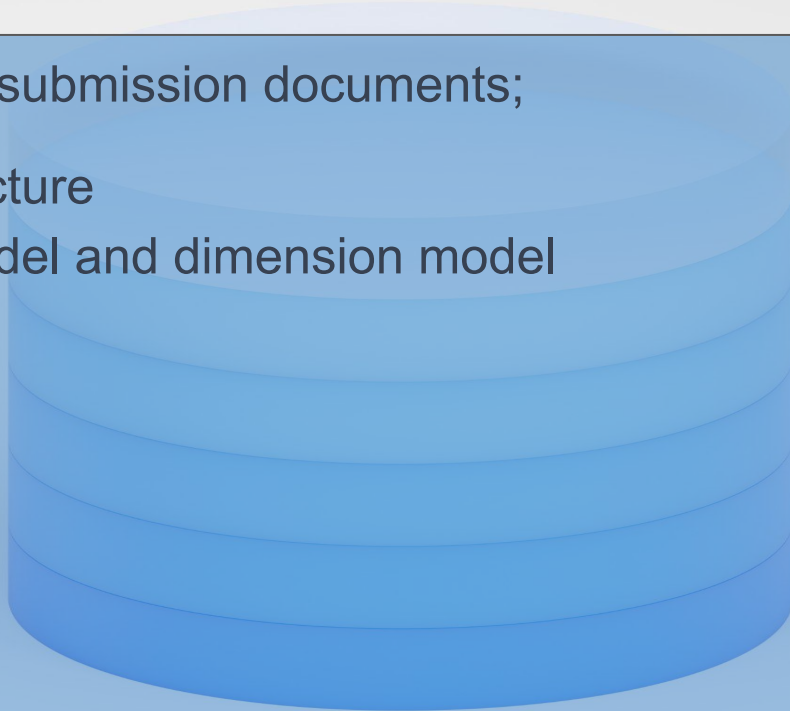
- 
- Python: For scripting and data manipulation tasks.
  - APIs: To connect and fetch data from various real estate platforms and databases.
  - Apache Airflow: To orchestrate and automate ETL workflows.
  - AWS EC2: Virtual servers for hosting and running data processing tasks.
  - AWS S3: Object storage for storing and managing large volumes of data.

The scope of Project AIR-REAL includes:

- Designing and implementing ETL pipelines for real-time data integration.
- Connecting to and extracting data from diverse real estate APIs and sources.
- Transforming and cleaning data to ensure accuracy and consistency.
- Loading processed data into a centralized data warehouse on AWS S3.
- Developing monitoring and logging mechanisms for pipeline visibility and troubleshooting.

Present the following submission documents;

- The Data architecture
- ERD for Data model and dimension model
- Python scripts





Present the following submission documents;

- The Data architecture
- ERD for Data model and dimension model
- Python scripts
- Data source (<https://rapidapi.com/realty-mole/api/realty-mole-property-api>)