**Real Estate Insights: Web Data ETL Pipeline & Tableau Dashboard**

**Project Objective: -**

The project aims to develop a comprehensive ETL (Extract, Transform, Load) pipeline and an interactive Tableau dashboard to provide detailed real estate insights at the state, county, and city levels. The objective is to automate the extraction, cleaning, transformation, and storage of real estate data, and then visualize key metrics and trends to support strategic decision-making in the real estate sector.

Data Extraction

In the Real Estate Insights project, the data extraction phase involved automating the download of real-estate data files for state, county, and zip-code levels from a public repository. To achieve this, I wrote a Python function using the boto library. This function automatically retrieved the required flat files and stored them in an AWS S3 bucket. This automated approach ensured that the data extraction process was efficient and scalable, reducing manual intervention and the risk of errors.

The use of AWS S3 as a storage solution provided a centralized and secure location for the data. It also facilitated easy access and management of the files, which was crucial for subsequent data processing steps. By automating the download process, I ensured that the data was up-to-date and consistently available for analysis, laying a solid foundation for the next stages of the ETL (Extract, Transform, Load) pipeline.

Data Transformation

During the data transformation stage, I extracted the files from the AWS S3 bucket and loaded them into pandas data frames. Each file contained around 2-3 million rows and 48 columns. From these, I selected 7-8 relevant columns such as median listing price, average listing price, and active and total listing counts for further analysis. The transformation process included cleaning the data by removing null values and converting string dates into datetime format. Additionally, I extracted the year and month from the date for more granular analysis.

To enrich the dataset, I calculated Month-over-Month (MoM) and Year-over-Year (YoY) values using pandas’ shift function, which mimics SQL’s lead and lag functions. This allowed for the calculation of growth rates and trends. Furthermore, I standardized state names by replacing abbreviated versions with their full names and filled in missing zip codes in the city\_zipcode file using USPS zip code data. These steps ensured the data was accurate, consistent, and ready for in-depth analysis.

Data Load

After completing the data cleaning and transformation processes, I wrote the transformed data into the Oracle Database. This step ensured that the cleaned and processed data was stored securely and was easily accessible for further use. Writing the data to Oracle Database maintained the data's integrity and facilitated smooth integration with other tools and platforms used in the analysis phase.

Storing the transformed data in Oracle Database also enabled seamless connectivity with data visualization tools such as Tableau. By maintaining a structured and organized dataset in Oracle Database, I ensured that the data was ready for efficient retrieval and analysis, supporting the project's goal of providing comprehensive real estate insights.

Tableau Dashboard

For the Tableau dashboard, I utilized Tableau's built-in connector to establish a connection with the Oracle Database. Using Tableau's Relationship concept, I logically joined the state, county, and city\_zip\_code files on common fields such as state names, county names, dates, and city names. The dashboard was designed to provide a comprehensive analysis across three levels: state, county, and city with zip code. Users could seamlessly navigate between these levels using Tableau's navigator feature.

The state-level analysis included filters for year, month, and state, and displayed the top 10 and bottom 10 states based on the selected criteria. A geographical map showcased the median listing prices, while line charts depicted overall median and average listings, and active and total listing counts over time. A bar chart showed the trend of median listing prices from the beginning to the current month, with a 12-month forecast using Tableau's forecasting feature. At the county level, similar analyses were conducted, with filters automatically adjusting to show relevant counties within the selected state. The city level offered parametric sorting options to display high-to-low or low-to-high rankings. Trend lines at the city\_zip\_code level illustrated the performance of median listing prices over time, providing a detailed and interactive view of real estate trends across different geographical scales.

Visualization: -

1. Geographical Map represents the color encoding based on median listing price.
2. Tables that show the Top 10 and Bottom 10 based on median listing price.
3. Line charts show the 12-month trend of Median listing price, Average Listing Price and Active & Total Listing count.
4. Bar-chart and Trend-line shows the median listing price from Oct 2016 to present and represents the next 12 months forecast for the median listing price.
5. Display four key metrics: median listing price, average listing price, median listing price per square foot, and total & active listing count, showing the prices for the selected state, county, and city for the current month and year along with MoM & YoY.

Architecture Diagram: -

Data Source

State Data

AWS S3

Python

County Data

City Data

Python

AWS Data Bricks

Convert into appropriate data types

Oracle Database

Tableau

Aggerate the data

Handling Missing & Duplicate Data

ER Diagram:-

