Streamlining Redfin Housing Data: A Cloud-Based ETL and Analytics Solution

Jatan Sahu
MSc(Data Science)
DAIICT
Supervisor: Arpit Rana
Gandhinagar, Gujarat

Internship Details:
Data Engineer
GrowExx AI Solution LLP
Supervisor: Jenny Vasani
Ahmedabad, Gujarat

Duration: 6 months
Start Date: 22 January 2024
End Date: 22 July 2024
202218061@daiict.ac.in

***Abstract*—In the quickly developing focus of data engineering, the ability to build effective pipelines for data extraction, processing, and analysis is critical. This project aims to meet the requirement by developing an extensive data engineering pipeline established completely for the real estate domain, the use of the enormous dataset accessible on the Redfin platform. The pipeline, that makes use of Amazon Web Services (AWS) EC2 instances and Apache Airflow, allows for the extraction of relevant data from Redfin, transformation into a suitable format, and subsequent loading into a Snowflake data warehouse. The end product of this work is seamless integration with PowerBI, allowing stakeholders to gain important insights on city-related real estate trends.**

***Index Terms*—EC2, ETL Pipeline, Apache-Airflow, Snowflake, PowerBI**

I. Introduction

In today's decision-making based on data era, the real estate sector emerges as an essential arena that involves complex economic, social, and regional dynamics. The development of digital platforms such as Redfin has totally transformed access to real estate data, resulting in a new era of insights and opportunities. To effectively use the huge amount of information, complex data engineering pipelines for large-scale data extraction, transformation, and analysis are required. This project aims to close the gap between data availability and actionable insights in real estate by building a complete data engineering pipeline. At its foundation, the project aims to combine multiple technologies and approaches in order to realize the full potential of Redfin data. The pipeline speeds up the process from raw data to directed decision-making by using cloud computing, workflow automation, and data warehousing.

Beyond from technological execution, this effort represents an approach change toward data-driven techniques in real estate. Intuition-based decision-making is losing way to scientific data, statistical accuracy, and predictive modeling. Using data, individuals including purchasers to politicians may negotiate the complexity of the real estate market with greater clarity and confidence.

The project begins with the deployment of an Amazon Web Services (AWS) EC2 instance to serve as the foundation for the data engineering pipeline. This infrastructure option represents an essential decision to take advantage of cloud computing's scalability, reliability, and cost-effectiveness. Apache Airflow facilitates workflow orchestration by ensuring exact execution of data extraction, transformation, and loading (ETL). The extraction of data from Redfin, supported by Python modules such as Boto3, presents issues in balancing data accuracy and computational speed. The data is then turned into an analyzable format using techniques such as normalization and feature engineering. The process culminates with loading converted data into a Snowflake data warehouse, which allows for extensive analysis using tools such as PowerBI.

II. Background

Firstly, a grasp of foundational data engineering principles, encompassing data extraction, transformation, and loading (ETL) techniques, is essential. Additionally, familiarity with cloud computing concepts, particularly AWS services like EC2 and S3, is crucial, alongside knowledge of Apache Airflow for workflow orchestration. Understanding data warehousing fundamentals, including schema design principles like star and snowflake schemas, is pivotal for structuring the data effectively within Snowflake, the chosen data warehousing solution. Proficiency in SQL and relational database management systems (RDBMS) is necessary for data manipulation within Snowflake. Moreover, expertise in business intelligence and data visualization techniques, particularly with PowerBI, is vital for deriving insights and communicating findings effectively. Lastly, a foundational understanding of the real estate domain, encompassing market trends, pricing dynamics, and consumer behavior, provides context and enhances the interpretation of the data. By integrating these diverse areas of knowledge, stakeholders can successfully implement and leverage the data engineering pipeline to extract actionable insights from Redfin data, facilitating informed decision-making within the real estate sector.

III. Technologies

A. Cloud Computing

The term ”cloud computing” describes the pay-as-you-go offering of computer services through the internet (sometimes known as ”the cloud”), including servers, storage, databases, networking, software, and more. Through a cloud service provider, consumers can access computing resources remotely in place of purchasing hardware and maintaining infrastructure. Cloud computing plays a pivotal role in providing scalable and flexible infrastructure to support the data engineering pipeline. Leveraging Amazon Web Services (AWS), specifically EC2 instances and S3 storage, ensures reliable and cost-effective computing resources for tasks such as data extraction, transformation, and storage. The use of cloud services eliminates the need for on-premises hardware provisioning and maintenance, enabling rapid deployment and scalability as data processing demands fluctuate. Additionally, cloud computing offers inherent benefits such as high availability, fault tolerance, and elastic scalability, ensuring the robustness and reliability of the data pipeline. By harnessing the power of cloud computing, stakeholders can optimize resource utilization, reduce infrastructure costs, and focus on deriving insights from the vast datasets available through platforms like Redfin.

B. AWS - Amazon Web Services

The acronym AWS stands for Amazon Web Services, a full of features and widely used cloud computing platform provided by Amazon.com. Many services are offered by it, such as networking, databases, machine learning, analytics, storage choices, processing capacity, and more. Businesses can expand and innovate without having to make an upfront investment in physical infrastructure because to AWS's on-demand service access. Some of the key services offered by AWS include:

1) Amazon EC2 (Elastic Compute Cloud)
2) Amazon S3 (Simple Storage Service)
3) Amazon RDS (Relational Database Service)
4) Amazon VPC (Virtual Private Cloud)
5) Amazon SNS (Simple Notification Service)
6) Amazon Lambda

C. Amazon EC2 - Elastic Compute Cloud

Amazon EC2 (Elastic Compute Cloud) is one of the core services provided by Amazon Web Services (AWS). It offers scalable computing capacity in the cloud, allowing users to launch virtual servers, known as instances, on-demand. We utilized Amazon EC2 (Elastic Compute Cloud) in several key aspects:

- **Hosting Apache Airflow:** We launched an EC2 xlarge instance in AWS and installed Apache Airflow on it. This instance serves as the environment where Airflow is hosted and runs your ETL (Extract, Transform, Load) pipeline.
- **Data Transformation:** Within the EC2 instance, We performed the data transformation process. This involved extracting data from a URL, transforming it into the desired format (CSV), and then loading it into S3 bucket. This data transformation likely required computational resources, which were provided by the EC2 instance.
- **Executing Airflow DAGs:** The EC2 instance runs Airflow in standalone mode, allowing us to define and execute Directed Acyclic Graphs (DAGs). These DAGs orchestrate the entire ETL process, including tasks such as data extraction, transformation, and loading.
- **Processing and Storing Large CSV Files:** As part of our ETL pipeline, we processed large CSV files generated from the data transformation step. The EC2 instance's computing power would have been instrumental in handling these large datasets efficiently.

D. ETL Pipeline

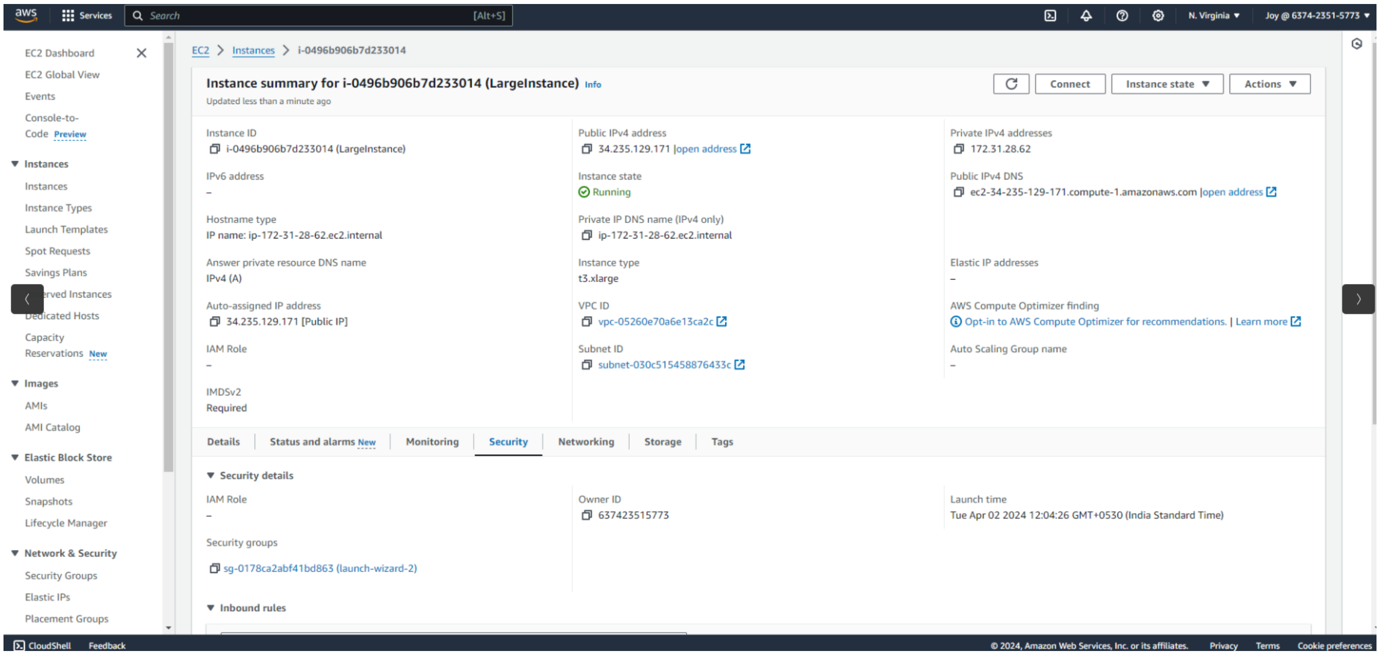
ETL stands for Extract, Transform, Load. An ETL pipeline is a set of processes used to extract data from various sources, transform it into a desired format, and load it into a target database or data warehouse. ETL pipelines are commonly used in data integration and data warehousing scenarios to consolidate data from multiple sources and make it available for analysis, reporting, and decision-making.

Our ETL pipeline involves several key steps, which can be broken down into Extract, Transform, and Load phases:

1) **Extract:** We start by launching an EC2 xlarge instance on AWS. Using Boto3, a Python library for AWS, you access the Redfin dataset stored in a URL. Apache Airflow, an orchestration tool, is deployed on the EC2 instance to manage the workflow. Airflow DAG (Directed Acyclic Graph) is created to automate the extraction process from the Redfin dataset URL.
2) **Transform:** After extraction, the data is transformed into CSV format within the EC2 instance. This transformation step may include data cleaning, normalization, or any necessary data manipulation to prepare it for analysis.
3) **Load:** Once transformed, the CSV file is loaded into an S3 bucket on AWS using Boto3. Additionally, a pipeline is established in Snowflake, a cloud-based data warehousing platform, which is triggered upon detecting the presence of the CSV file in the S3 bucket. Within Snowflake, We have likely defined a schema and created tables to accommodate the incoming data. A COPY command is used in Snowflake to load the CSV data from the S3 bucket into the corresponding Snowflake table.
4) **Analytics:** With the data now residing in Snowflake, PowerBI, a powerful business intelligence tool, is connected to Snowflake to perform analytics and generate insights from the Redfin data. PowerBI can execute various analyses such as visualizations, trend analysis, and statistical modeling to derive actionable insights from the dataset.

E. Apache-Airflow

Apache Airflow is an open-source platform used for orchestrating complex workflows and data pipelines. Airflow allows users to programmatically author, schedule, and monitor workflows as directed acyclic graphs (DAGs) of tasks. It provides a flexible and scalable framework for orchestrating data processing tasks and enables organizations



EC2 > Instances > i-0496b906b7d233014

Instance summary for i-0496b906b7d233014 (LargeInstance) Info
Updated less than a minute ago

Instance ID
i-0496b906b7d233014 (LargeInstance)

Public IPv4 address
34.235.129.171 |open address

Private IPv4 addresses
172.31.28.62

IPv6 address
–

Instance state
Running

Public IPv4 DNS
ec2-34-235-129-171.compute-1.amazonaws.com |open address

Hostname type
IP name: ip-172-31-28-62.ec2.internal

Private IP DNS name (IPv4 only)
ip-172-31-28-62.ec2.internal

Answer private resource DNS name
IPv4 (A)

Instance type
t3.xlarge

Elastic IP addresses
–

Auto-assigned IP address
34.235.129.171 [Public IP]

VPC ID
vpc-05260e70a6e13ca2c

AWS Compute Optimizer finding
Opt-in to AWS Compute Optimizer for recommendations. | Learn more

IAM Role
–

Subnet ID
subnet-030c515458876433c

Auto Scaling Group name
–

IMDSv2
Required

Security details

IAM Role
–

Owner ID
637423515773

Launch time
Tue Apr 02 2024 12:04:26 GMT+0530 (India Standard Time)

Security groups
sg-0178ca2abf41bd863 (launch-wizard-2)

Inbound rules

Fig. 1: EC2 Configuration

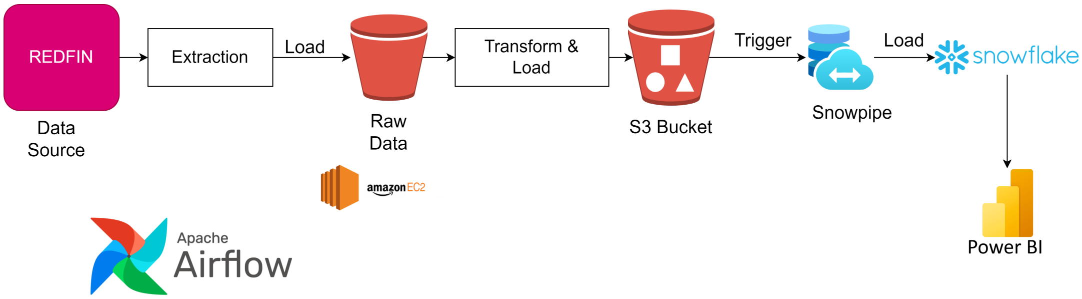
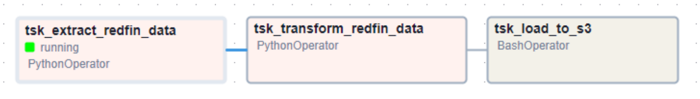


Fig. 2: Project Pipeline

to automate and streamline their data engineering workflows. **Airflow Standalone:** After installing Apache Airflow, you configured and ran it in standalone mode on your EC2 instance. This allows you to use Airflow for workflow management.

Creating DAG (Directed Acyclic Graph): You defined a DAG in Apache Airflow to orchestrate the ETL process. The DAG consists of tasks that represent different stages of the pipeline such as extracting data from the Redfin dataset, transforming it into CSV format, and loading it into your S3 bucket.


tsk_extract_redfin_data
running
PythonOperator

tsk_transform_redfin_data
PythonOperator

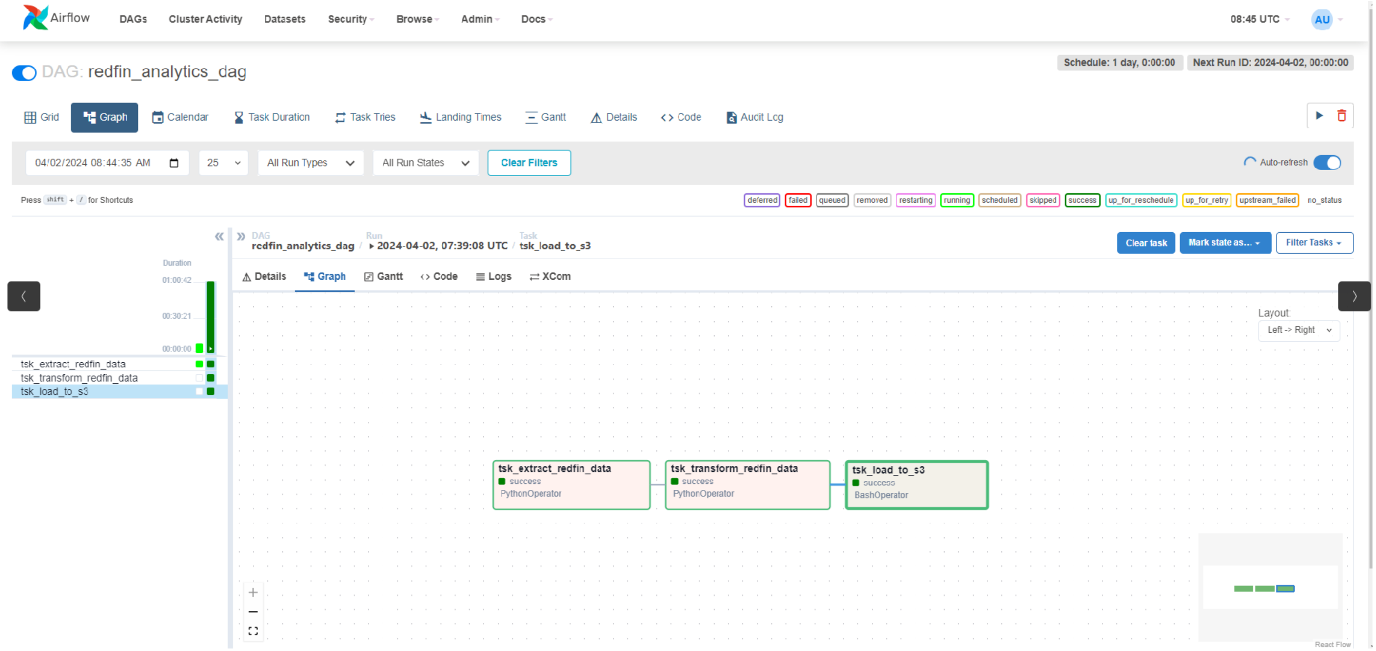
tsk_load_to_s3
BashOperator

Fig. 3: Airflow DAG

IV. Dataset

Redfin is a real estate brokerage, meaning they have direct access to data from local multiple listing services, as well as insight from that real estate agents across the country. That's why they are able to give the earliest and most reliable data on the state of the housing market. They publish existing industry data faster, and offer additional data on tours and offers that no one else has. Redfin gathers and analyzes housing market data from various sources, including local multiple listing services (MLS), county recorder's offices, and its own proprietary data.

- **Home Listings:** Redfin provides comprehensive listings of homes for sale, including details such as price, location, square footage, number of bedrooms and bathrooms, photos, and property descriptions. Users can search for homes based on their preferences and filter results accordingly.
- **Market Trends:** Redfin offers insights into housing



DAG: redfin_analytics_dag

Schedule: 1 day, 0:00:00

Next Run ID: 2024-04-02, 00:00:00

04/02/2024 08:44:35 AM

25

DAG redfin_analytics_dag / Run 2024-04-02, 07:39:08 UTC / Task tsk_load_to_s3

tsk_extract_redfin_data

tsk_transform_redfin_data

tsk_load_to_s3

tsk_extract_redfin_data
success
PythonOperator

tsk_transform_redfin_data
success
PythonOperator

tsk_load_to_s3
success
BashOperator

Fig. 4: Airflow ETL Pipeline

TABLE I: Redfin US House Market Dataset Schema

Column Name	Data Type	Description
period_begin	DATE	Start date of the period
period_end	DATE	End date of the period
period_duration	INT	Duration of the period (in days or months)
region_type	STRING	Type of region (e.g., city, state)
region_type_id	INT	Identifier for the region type
table_id	INT	Identifier for the table
is_seasonally_adjusted	STRING	Indicator for whether the data is seasonally adjusted
city	STRING	Name of the city
state	STRING	Name of the state
state_code	STRING	Code for the state
property_type	STRING	Type of property (e.g., single-family, condo)
property_type_id	INT	Identifier for the property type
median_sale_price	FLOAT	Median sale price of properties
median_list_price	FLOAT	Median list price of properties
median_ppsf	FLOAT	Median price per square foot
median_list_ppsf	FLOAT	Median list price per square foot
homes_sold	FLOAT	Number of homes sold
inventory	FLOAT	Inventory of available homes
months_of_supply	FLOAT	Number of months of supply
median_dom	FLOAT	Median days on market
avg_sale_to_list	FLOAT	Average sale-to-list ratio
sold_above_list	FLOAT	Percentage of homes sold above list price
parent_metro_region_metro_code	STRING	Code for the parent metro region
last_updated	DATETIME	Date and time of last update
period_begin_in_years	STRING	Start year of the period
period_end_in_years	STRING	End year of the period
period_begin_in_months	STRING	Start month of the period
period_end_in_months	STRING	End month of the period

market trends at both national and local levels. This includes data on median home prices, inventory levels, days on market, and sales activity. Users can track how the market is performing over time and make informed decisions about buying or selling property.

- **Home Value Estimates:** Redfin provides estimates of home values using its proprietary algorithm, which takes into account factors such as recent sales of similar properties in the area, market trends, and property characteristics. These estimates can be useful for homeowners who want to get an idea of their home's worth or for prospective buyers who are evaluating potential purchases.
- **Neighborhood Information:** Redfin offers information about neighborhoods, including demographics, school ratings, crime rates, amenities, and local attractions. This helps users evaluate the desirability of different areas and find the right neighborhood for their needs.
- **Market Reports:** Redfin periodically publishes market reports that provide in-depth analysis of housing market trends, including commentary from real estate experts and economists. These reports can be useful for investors, industry professionals, and anyone interested in understanding the dynamics of the real estate market.

V. Working Steps

The project described above involves node classification in the first step of semantic mapping. Here are the general steps that we followed to create this project:

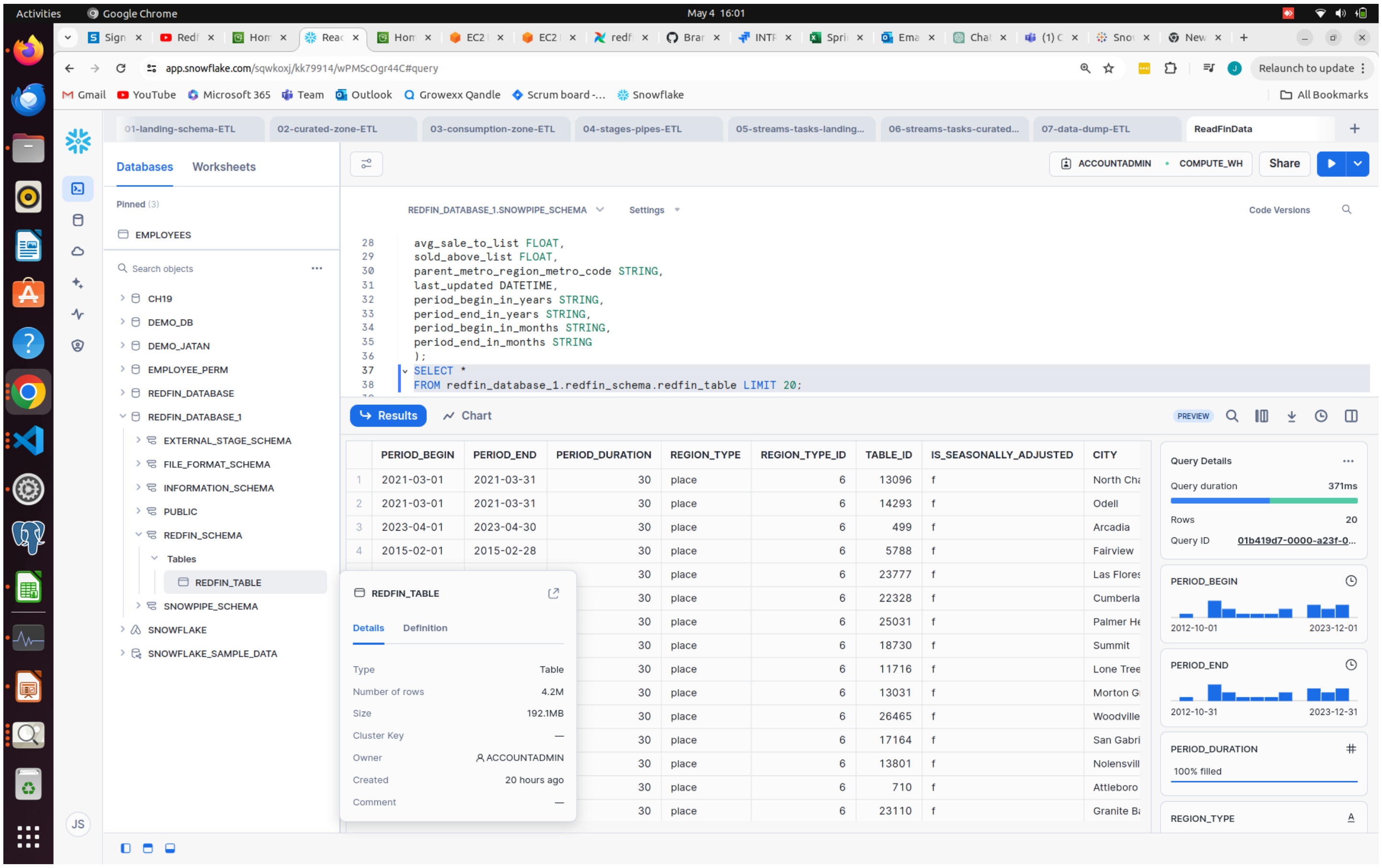
1) **Launch EC2 xlarge instance in AWS:** We begin by launching an EC2 xlarge instance on Amazon Web Services (AWS). This instance will serve as computing environment for running Apache Airflow and other necessary tools. We configure the instance settings such as network, security groups, and storage.
2) **Install Boto3 and Apache Airflow:** Once the EC2 instance is running, you install the required dependencies, including Boto3 (the AWS SDK for Python) and Apache Airflow, which is a platform to programmatically author, schedule, and monitor workflows.
3) **Run Airflow standalone:** We start the Apache Airflow service on the EC2 instance in standalone mode. This allows us to define and execute Directed Acyclic Graphs (DAGs) for data pipelines.
4) **Create Airflow DAG for Redfin data extraction:** You create a DAG in Apache Airflow to extract data from the Redfin dataset. This DAG defines the workflow for downloading the data from a specified URL and transforming it into CSV format.
5) **Data transformation and loading into S3:** Within the EC2 instance, the DAG executes tasks to transform the extracted data into CSV format. Once transformed, the data is loaded into an S3 bucket on AWS.
6) **Pipeline setup in Snowflake:** We set up a pipeline in Snowflake, a cloud-based data warehousing platform. This pipeline is triggered when a new CSV file is added to the S3 bucket.
7) **Pipeline setup in Snowflake:** We set up a pipeline in Snowflake, a cloud-based data warehousing platform. This pipeline is triggered when a new CSV file is added to the S3 bucket.
8) **Code for creating tables and schemas in Snowflake:** We write code (using SQL) within Snowflake's workbook to create tables and schemas that will store the data from the CSV files.
9) **Load CSV data into Snowflake:** The pipeline automatically loads the CSV data from the S3 bucket into Snowflake's data warehouse, using the tables and schemas we have defined.
10) **PowerBI access to Snowflake data:** Finally, we configure PowerBI, a business analytics tool, to access the data stored in Snowflake. This allows us to perform various analyses and create visualizations based on the Redfin dataset.
11) **Analytics with PowerBI:** With PowerBI connected to Snowflake, we can now perform analytics on the Redfin data. This may include creating dashboards, reports, and visualizations to gain insights into various aspects of the data, such as housing market trends, pricing analysis, geographical distribution, etc.

VI. Results

This data engineering approach resulted in an effective and scalable ETL pipeline that was smoothly managed using Apache Airflow on an EC2 xlarge machine. The project improved data extraction, transformation, and loading procedures by using technologies such as S3, Snowflake, and PowerBI, resulting in seamless data accessibility and integration across the pipeline. The PowerBI analytics provided stakeholders with actionable insights from the Redfin dataset, allowing them to make informed decisions based on trends, patterns, and anomalies found through interactive dashboards and reports. Extensive testing, monitoring, documenting, and maintenance methods assured the workflow's dependability, scalability, and performance, demonstrating the value of data-driven decision-making in delivering business growth and enabling strategic improvements in the real estate industry.

VII. Conclusion

In conclusion, this data engineering project exemplifies the transformative potential of leveraging advanced technologies and best practices to unlock valuable insights from complex datasets. Through the seamless integration of AWS, Apache Airflow, Snowflake, and PowerBI, the project successfully established an end-to-end ETL pipeline capable of efficiently processing, transforming, and analyzing the Redfin dataset. The project's outcomes, including actionable insights derived from PowerBI analytics and the establishment of robust monitoring and maintenance protocols, underscore the critical importance of data-driven decision-making in driving strategic advancements and operational efficiencies within the real estate domain and beyond.



```
avg_sale_to_list FLOAT,
sold_above_list FLOAT,
parent_metro_region_metro_code STRING,
last_updated DATETIME,
period_begin_in_years STRING,
period_end_in_years STRING,
period_begin_in_months STRING,
period_end_in_months STRING
);
SELECT *
FROM redfin_database_1.redfin_schema.redfin_table LIMIT 20;
```

	PERIOD_BEGIN	PERIOD_END	PERIOD_DURATION	REGION_TYPE	REGION_TYPE_ID	TABLE_ID	IS_SEASONALLY_ADJUSTED	CITY
1	2021-03-01	2021-03-31	30	place	6	13096	f	North Cha
2	2021-03-01	2021-03-31	30	place	6	14293	f	Odell
3	2023-04-01	2023-04-30	30	place	6	499	f	Arcadia
4	2015-02-01	2015-02-28	30	place	6	5788	f	Fairview
			30	place	6	23777	f	Las Flores
			30	place	6	22328	f	Cumberla
			30	place	6	25031	f	Palmer He
			30	place	6	18730	f	Summit
			30	place	6	11716	f	Lone Tree
			30	place	6	13031	f	Morton G
			30	place	6	26465	f	Woodville
			30	place	6	17164	f	San Gabri
			30	place	6	13801	f	Nolensvill
			30	place	6	710	f	Attleboro
			30	place	6	23110	f	Granite Ba

REDFIN_TABLE

Details | Definition

Type	Table
Number of rows	4.2M
Size	192.1MB
Cluster Key	—
Owner	ACCOUNTADMIN
Created	20 hours ago
Comment	—

Query Details

Query duration	371ms
Rows	20
Query ID	01b419d7-0000-a23f-0...

PERIOD_BEGIN 2012-10-01 2023-12-01

PERIOD_END 2012-10-31 2023-12-31

PERIOD_DURATION 100% filled

REGION_TYPE

Fig. 5: Snowflake Data

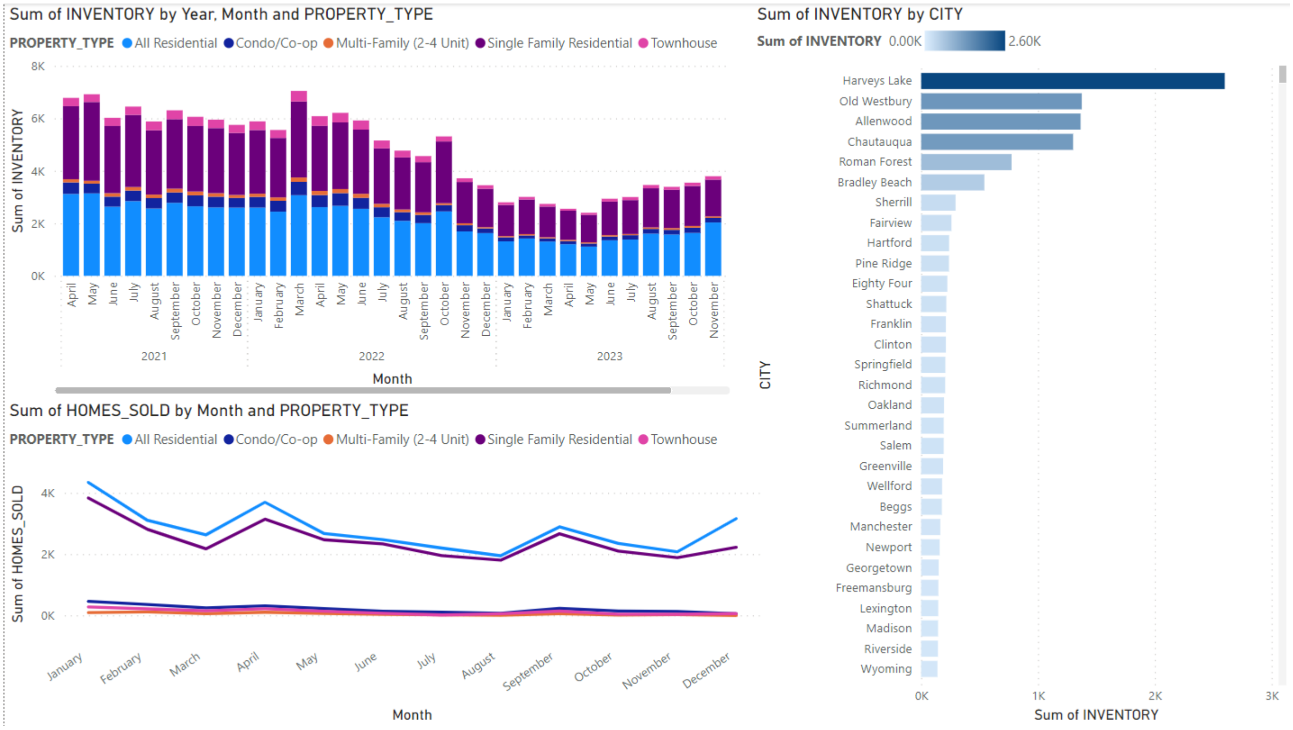


Fig. 6: Analysis in PowerBI

Looking ahead, the project sets a solid foundation for future enhancements and expansions, with opportunities to further optimize and scale the pipeline to accommodate evolving data requirements and analytical needs. By embracing a culture of continuous improvement and innovation, stakeholders can harness the power of data to drive informed decision-making, capitalize on emerging opportunities, and address complex challenges in the dynamic real estate landscape and other industries alike. This project stands as a testament to the transformative potential of data engineering in unlocking actionable insights and driving meaningful business outcomes in today's data-driven world.

VIII. Acknowledgement

I extend my heartfelt gratitude to all those who have played a part in my internship experience and the completion of this report. My sincere thanks go to **GrowExx Private Limited** for providing me with this valuable opportunity. I am also thankful for the guidance and support of my supervisor whose mentorship has been invaluable. I appreciate the collaboration and assistance of my colleagues at **Data Engineering Team**, as well as the guidance of **Prof.Arpit Rana** and the academic faculties. Lastly, I am grateful to my family and friends for their unwavering support throughout this journey. Thank you to everyone who has contributed to my growth and learning during this internship.

References

[1] *Redfin US House Market Tracker Data (City)*. https://www.redfin.com/news/data-center/

[2] *EC2 - Amazon Elastic Compute Cloud(Virtual Machine)*. https://docs.aws.amazon.com/ec2/

[3] *Apache Airflow*. https://airflow.apache.org/docs/

[4] *S3 - Amazon Simple Storage Service Documentation*. https://docs.aws.amazon.com/s3/

[5] *Snowflake*. https://docs.snowflake.com/

[6] *PowerBI*. https://learn.microsoft.com/en-us/power-bi/

[7] *DAGs - Directed Acyclic Graph Documentation*. https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/dags.html