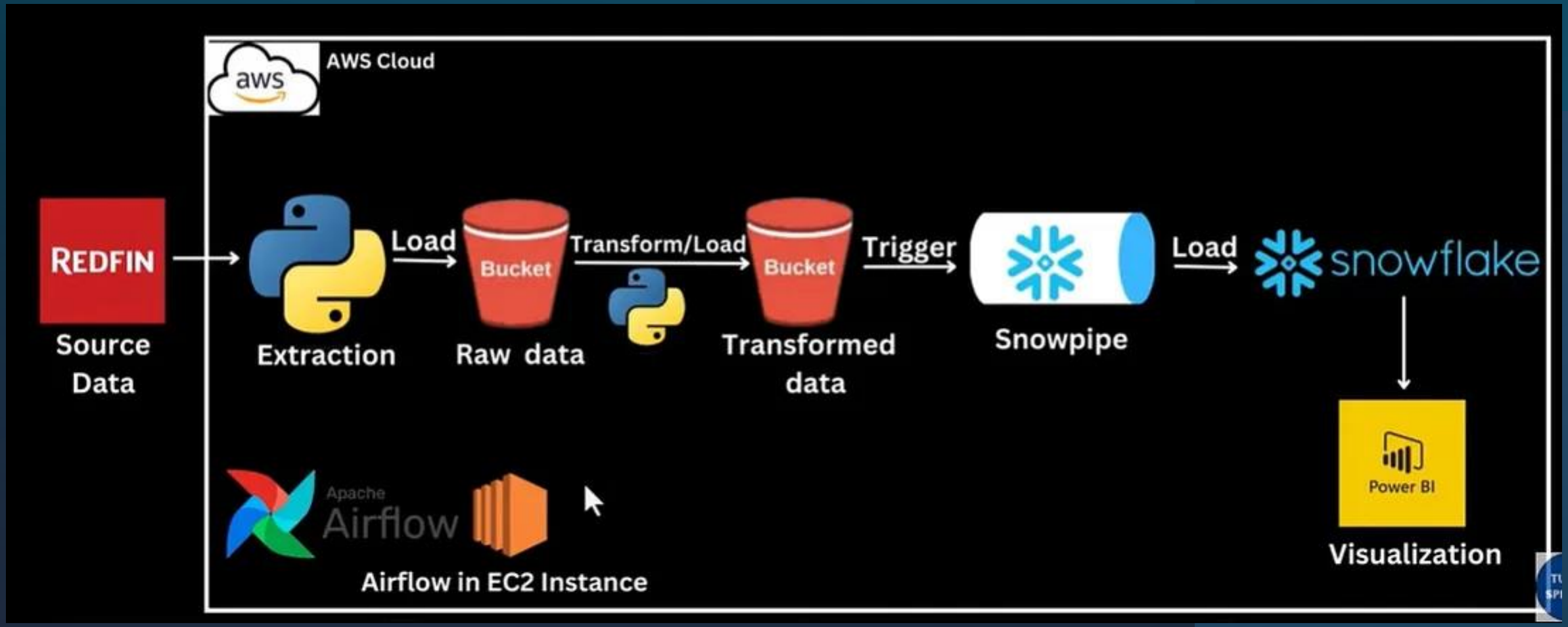
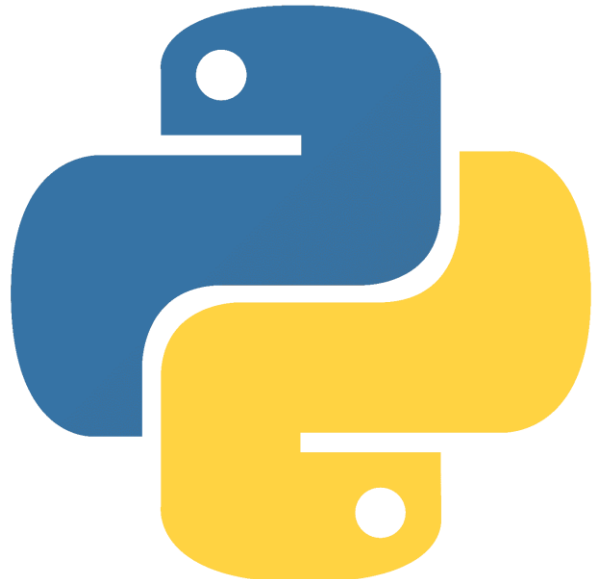

AWS Cloud
REDFIN
Source Data
Extraction
Load
Bucket
Raw data
Transform/Load
Bucket
Transformed data
Trigger
Snowpipe
Load
snowflake
Power BI
Visualization
Apache Airflow
Airflow in EC2 Instance



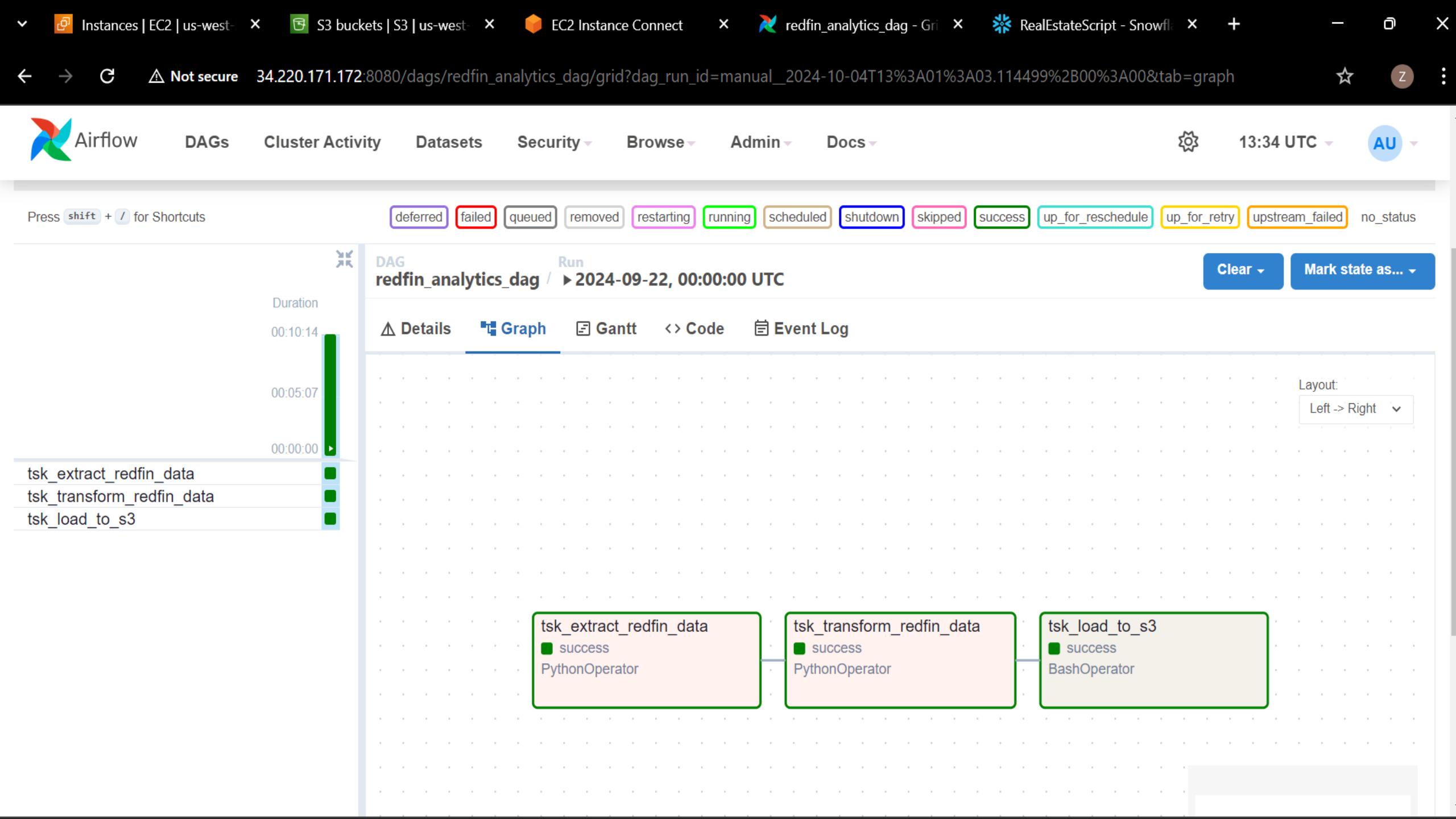
python™

redfin_analytics.py

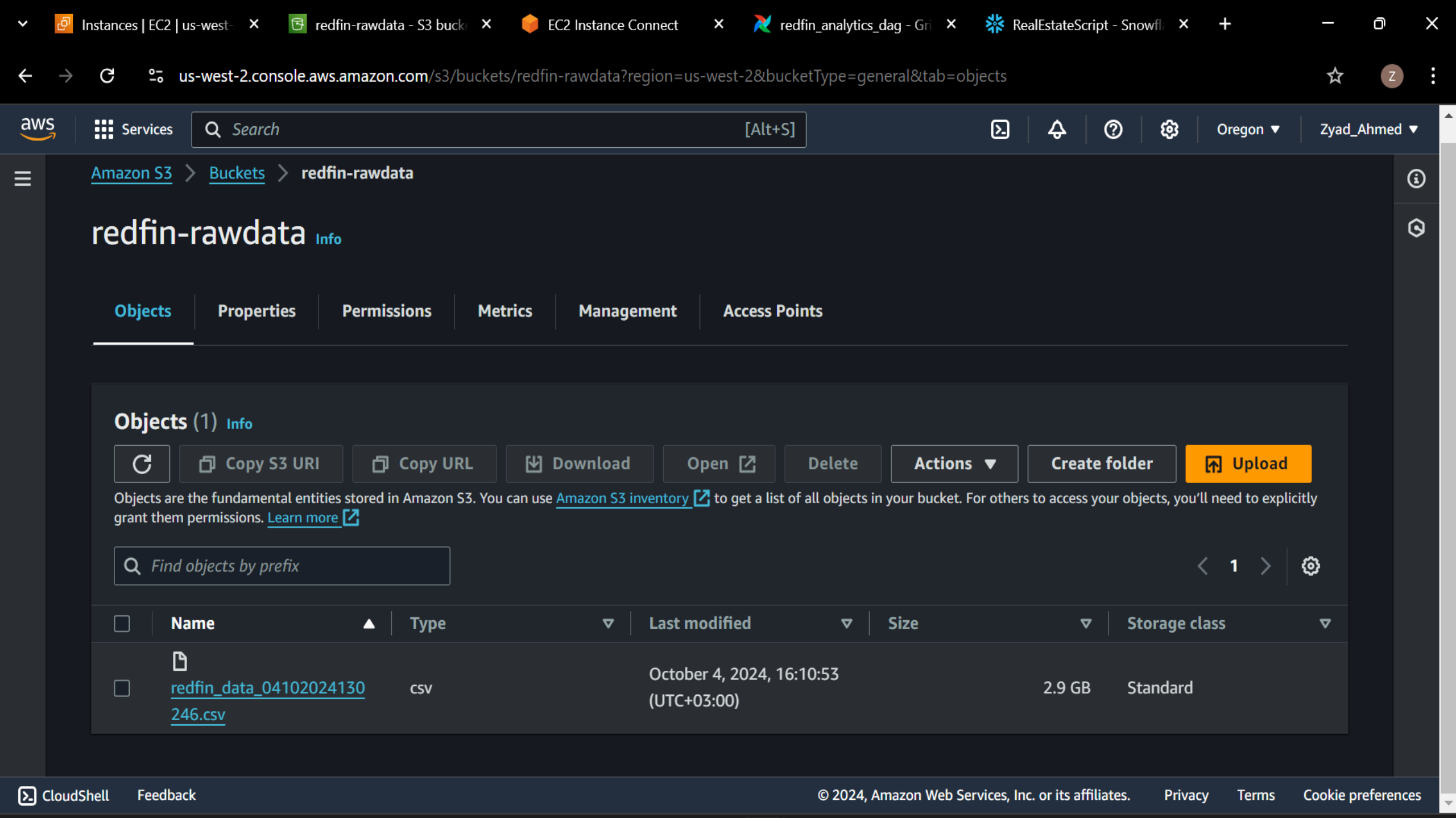
C: > Users > DELL > Desktop > Un_Uploaded Projects > Redfin_Snowpipe > redfin_analytics.py > transform_data

```python
# s3 buckets
target_bucket_name = 'redfin-transformed-data'

url_by_city = 'https://redfin-public-data.s3.us-west-2.amazonaws.com/redfin_market_tracker/city_market_tracker.tsv000.gz'

def extract_data(**kwargs):
    url = kwargs['url']
    df = pd.read_csv(url, compression='gzip', sep='\t')
    now = datetime.now()
    date_now_string = now.strftime("%d%m%Y%H%M%S")
    file_str = 'redfin_data_' + date_now_string
    df.to_csv(f"{file_str}.csv", index=False)
    output_file_path = f"/home/ubuntu/{file_str}.csv"
    output_list = [output_file_path, file_str]
    return output_list

def transform_data(task_instance):
    data = task_instance.xcom_pull(task_ids="tsk_extract_redfin_data")[0]
    object_key = task_instance.xcom_pull(task_ids="tsk_extract_redfin_data")[1]
    df = pd.read_csv(data)

    # Remove commas from the 'city' column
    df['city'] = df['city'].str.replace(',', '')
    cols = ['period_begin','period_end','period_duration', 'region_type', 'region_type_id', 'table_id',
    'is_seasonally_adjusted', 'city', 'state', 'state_code', 'property_type', 'property_type_id',
    'median_sale_price', 'median_list_price', 'median_ppsf', 'median_list_ppsf', 'homes_sold',
    'inventory', 'months_of_supply', 'median_dom', 'avg_sale_to_list', 'sold_above_list', 'parent_metro_region_metro_code', 'last_updated']
    df = df[cols]
    df = df.dropna()
```

Apache
Airflow


Instances | EC2 | us-west-
S3 buckets | S3 | us-west-
EC2 Instance Connect
redfin_analytics_dag - Gri
RealEstateScript - Snowfl
Not secure
34.220.171.172:8080/dags/redfin_analytics_dag/grid?dag_run_id=manual__2024-10-04T13%3A01%3A03.114499%2B00%3A00&tab=graph
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
13:34 UTC
AU
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:10:14
00:05:07
00:00:00
tsk_extract_redfin_data
tsk_transform_redfin_data
tsk_load_to_s3
DAG
redfin_analytics_dag
Run
2024-09-22, 00:00:00 UTC
Clear
Mark state as...
Details
Graph
Gantt
Code
Event Log
Layout:
Left -> Right
tsk_extract_redfin_data
success
PythonOperator
tsk_transform_redfin_data
success
PythonOperator
tsk_load_to_s3
success
BashOperator

Amazon
S3


Instances | EC2 | us-west-
redfin-rawdata - S3 buck
EC2 Instance Connect
redfin_analytics_dag - Gri
RealEstateScript - Snowfl
us-west-2.console.aws.amazon.com/s3/buckets/redfin-rawdata?region=us-west-2&bucketType=general&tab=objects
aws
Services
Search
[Alt+S]
Oregon
Zyad_Ahmed
Amazon S3 > Buckets > redfin-rawdata
redfin-rawdata Info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (1) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
redfin_data_04102024130246.csv
csv
October 4, 2024, 16:10:53 (UTC+03:00)
2.9 GB
Standard
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

redfin-transformed-data Info

Objects | Properties | Permissions | Metrics | Management | Access Points

Objects (1) Info

Copy S3 URI | Copy URL | Download | Open | Delete | Actions | Create folder | Upload

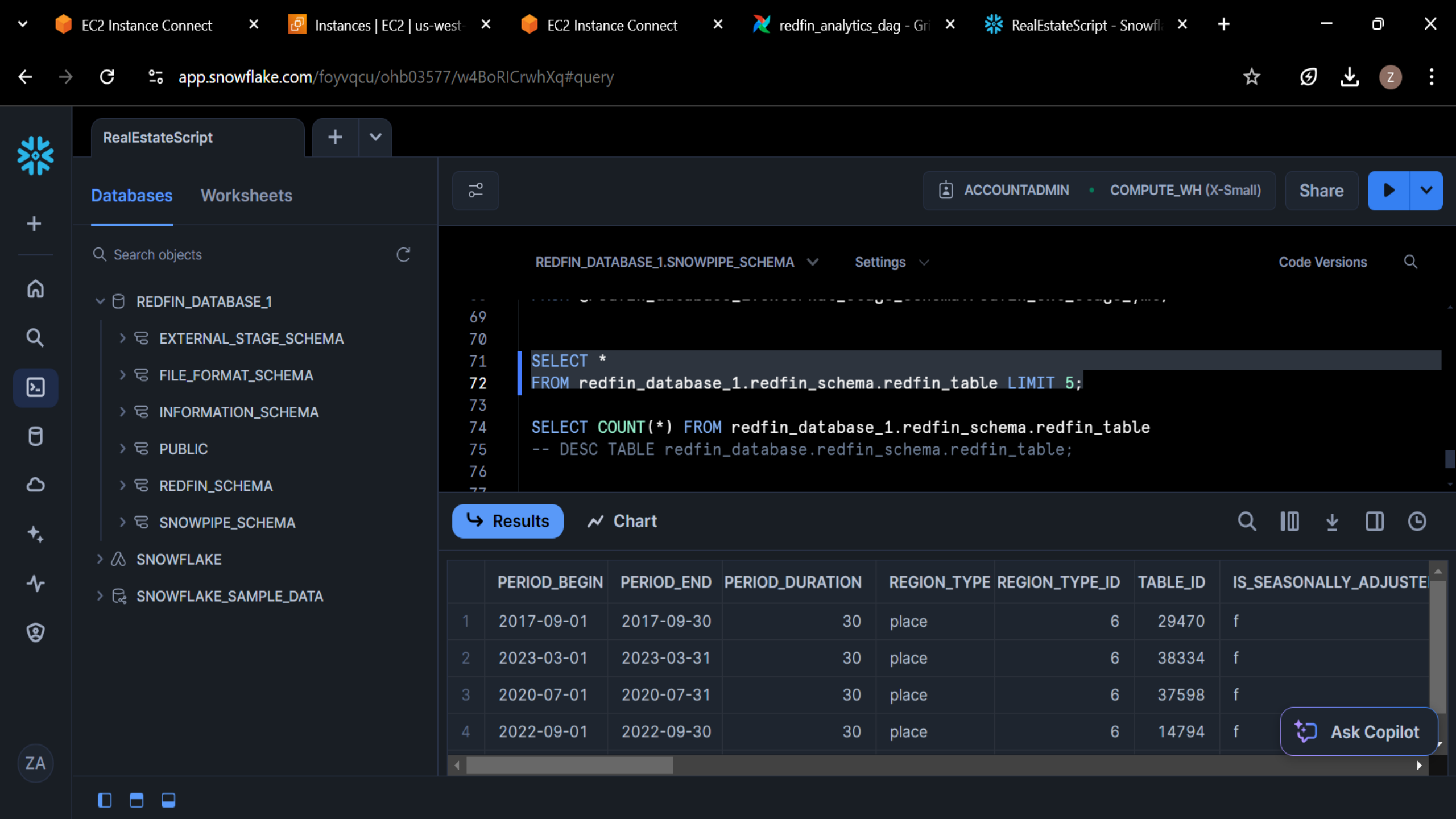
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Find objects by prefix

1

Name	Type	Last modified	Size	Storage class
redfin_data_04102024130246.csv	csv	October 4, 2024, 16:10:40 (UTC+03:00)	953.9 MB	Standard

snowflake®


EC2 Instance Connect
Instances | EC2 | us-west-
EC2 Instance Connect
redfin_analytics_dag - Gri
RealEstateScript - Snowfl
app.snowflake.com/foyvqcu/ohb03577/w4BoRICrwhXq#query
RealEstateScript
Databases
Worksheets
Search objects
REDFIN_DATABASE_1
EXTERNAL_STAGE_SCHEMA
FILE_FORMAT_SCHEMA
INFORMATION_SCHEMA
PUBLIC
REDFIN_SCHEMA
SNOWPIPE_SCHEMA
SNOWFLAKE
SNOWFLAKE_SAMPLE_DATA
ACCOUNTADMIN
COMPUTE_WH (X-Small)
Share
REDFIN_DATABASE_1.SNOWPIPE_SCHEMA
Settings
Code Versions
69
70
71 SELECT *
72 FROM redfin_database_1.redfin_schema.redfin_table LIMIT 5;
73
74 SELECT COUNT(*) FROM redfin_database_1.redfin_schema.redfin_table
75 -- DESC TABLE redfin_database.redfin_schema.redfin_table;
76
Results
Chart
PERIOD_BEGIN PERIOD_END PERIOD_DURATION REGION_TYPE REGION_TYPE_ID TABLE_ID IS_SEASONALLY_ADJUSTE
1 2017-09-01 2017-09-30 30 place 6 29470 f
2 2023-03-01 2023-03-31 30 place 6 38334 f
3 2020-07-01 2020-07-31 30 place 6 37598 f
4 2022-09-01 2022-09-30 30 place 6 14794 f
Ask Copilot
ZA

EC2 Instance Connect | Instances | EC2 | us-west- | EC2 Instance Connect | redfin_analytics_dag - Gri | RealEstateScript - Snowfl

app.snowflake.com/foyvqcu/ohb03577/w4BoRICrwhXq#query

RealEstateScript

Databases | Worksheets

Search objects

- REDFIN_DATABASE_1
  - EXTERNAL_STAGE_SCHEMA
  - FILE_FORMAT_SCHEMA
  - INFORMATION_SCHEMA
  - PUBLIC
  - REDFIN_SCHEMA
  - SNOWPIPE_SCHEMA
- SNOWFLAKE
- SNOWFLAKE_SAMPLE_DATA

ACCOUNTADMIN • COMPUTE_WH (X-Small) | Share

REDFIN_DATABASE_1.SNOWPIPE_SCHEMA | Settings | Code Versions

```
69
70
71  SELECT *
72  FROM redfin_database_1.redfin_schema.redfin_table LIMIT 5;
73
74  SELECT COUNT(*) FROM redfin_database_1.redfin_schema.redfin_table
75  -- DESC TABLE redfin_database.redfin_schema.redfin_table;
76
```

Results | Chart

	COUNT(*)
1	4331055

Query Details

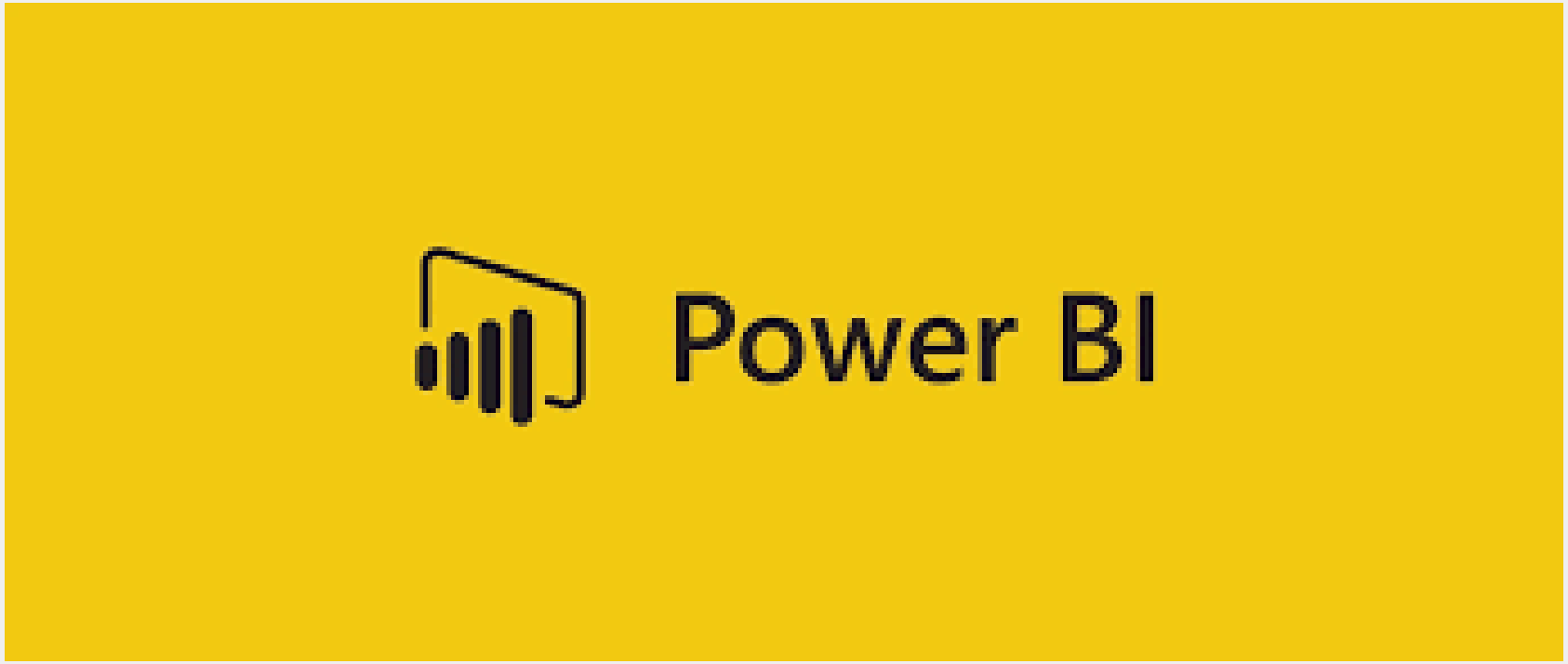
Query duration 76ms

Rows 1

Query ID 01b7778a-0003-45e8-0...

Show more | Ask Copilot

ZA


Power BI

Redfin_visualization • Last saved: Today at 7:04 PM

Search

zyad ahmed

File Home Insert Modeling View Optimize Help

Share

Paste Cut Copy Format painter — Clipboard

Get data Excel workbook OneLake data hub SQL Server Enter data Dataverse Recent sources — Data

Transform data Refresh — Queries

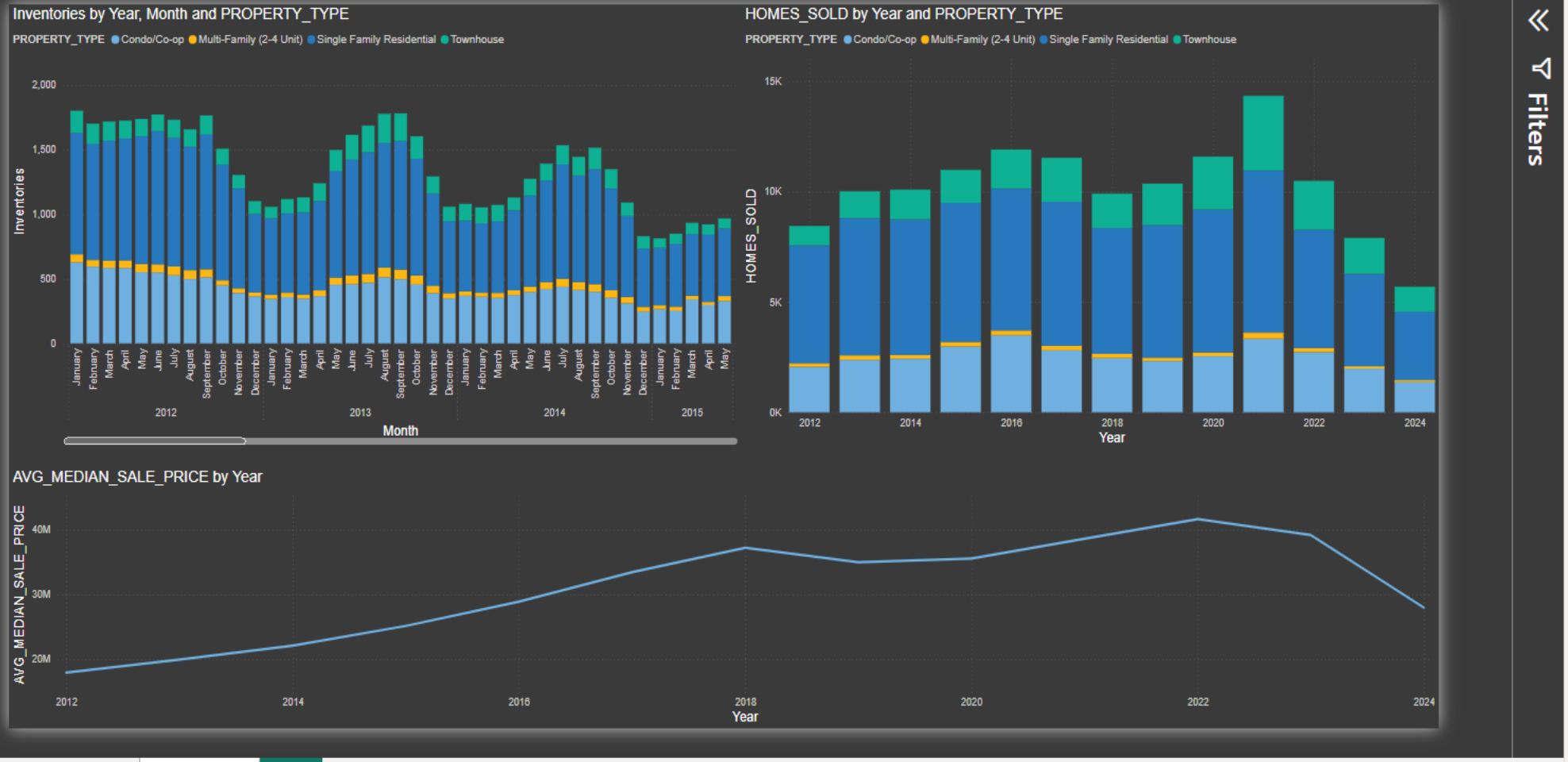
New visual Text box More visuals — Insert

New visual calculation New measure Quick measure — Calculations

Sensitivity — Sensitivity

Publish — Share

Copilot — Copilot



Filters

Visualizations

Data

Search

REDFIN_TABLE

Page 1

Page 1 of 1

50%