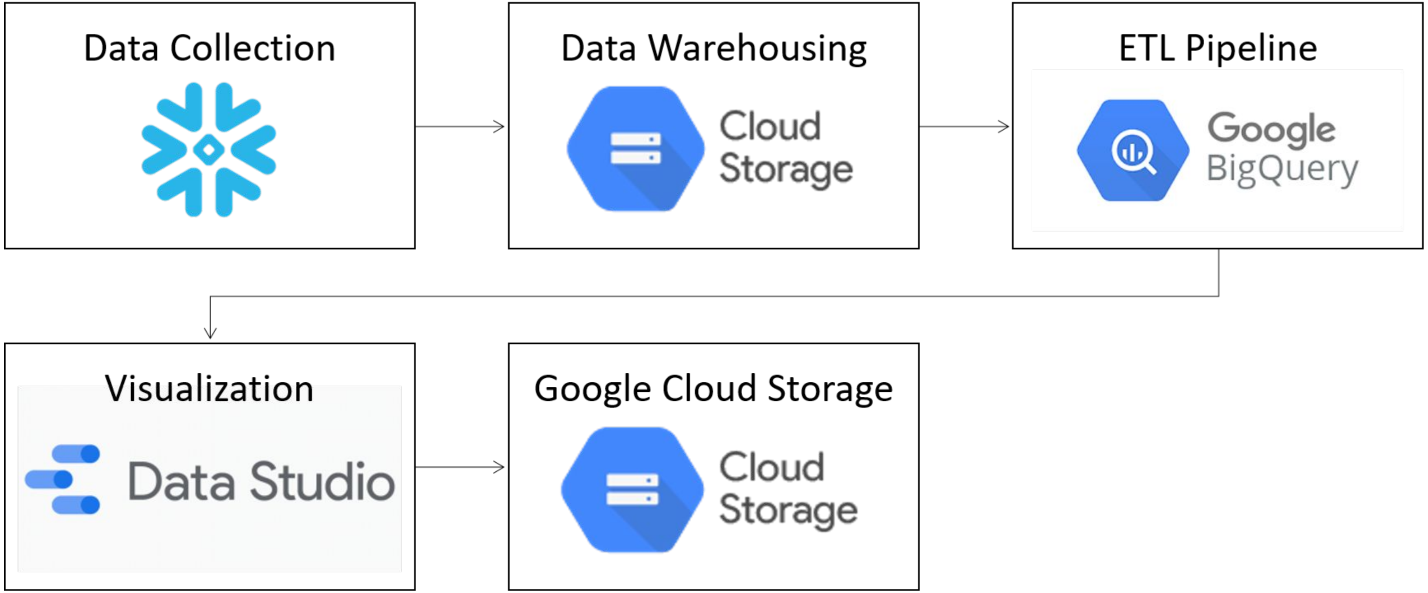
Egen Final Project

Zillow ETL Pipeline and visualizations

Architecture



Data Collection

Data Warehousing

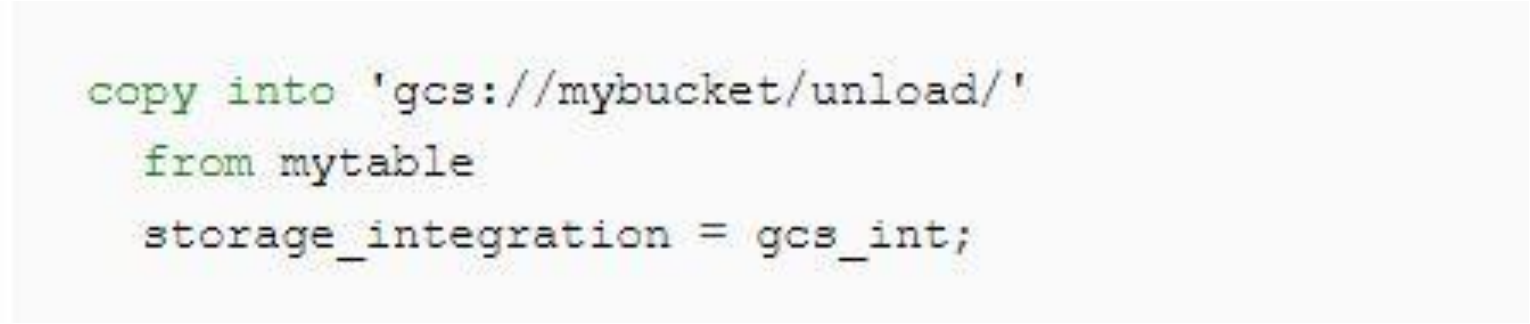
ETL Pipeline

Visualization

Google Cloud Storage

Data Collection

- Data was obtained from the Snowflake marketplace
- Then the data was transferred from Snowflake to Cloud Bucket using Snowflake API
-



```
copy into 'gcs://mybucket/unload/'
  from mytable
  storage_integration = gcs_int;
```

Data Warehousing

- We are gathering data from third party vendor, in this situation snowflake data warehouse.
- The data come in flat file with 18 columns and 6.9 Million rows.
- The data warehouse has to be in **dimensional** model

So we are using google data cloud bucket as object storage, imitating hadoop HDFS files systems. So when even, Third party vendor append a data then the file will in bucket will also get appended with conditional refresh.

Why google cloud as data warehouse?

Google data warehouse is object storage. So In future we decided to included images, and application data, then it will be on single storage rather than going to various storage.

ETL Pipeline

- Data clean up removing null
- Data transformations
- Calculated fields
- CTE (Common Table Expression)

So 18*6.9m data is not required for reporting. For our case scenario 1 where we want to find top 5 states in US in terms of values we need, only states and value columns with all the rows. Null excluded.

So we created temporary table in Bigquery with only 2 columns and 6.9 million rows.

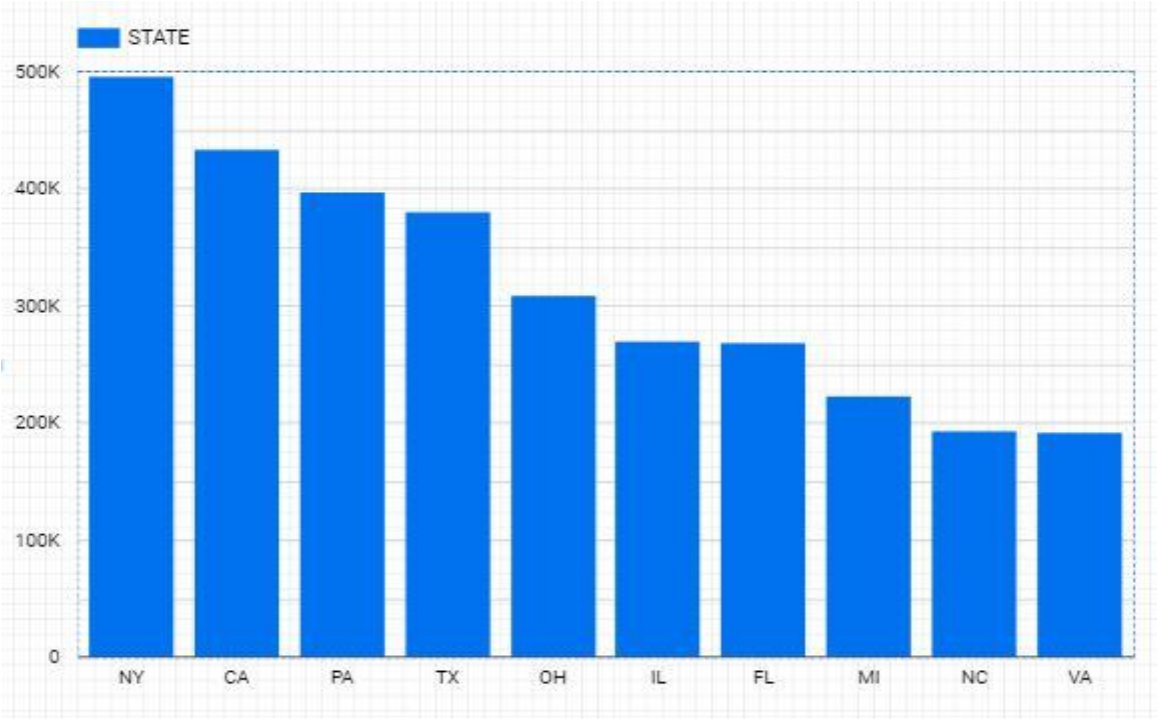
Why big query and not google clouds mysql ???

Mysql is optimized for relational databases. So if we need to scale up our app for 20 times with multiple tables joined. Will take day to compute report. Hence, We will go with Bigquery which uses columnar compression for query processing and is faster.

Visualization

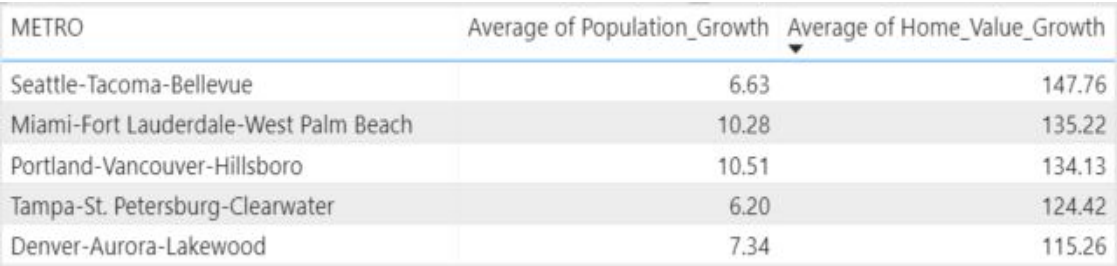
- As previously explained we are finding top states with highest values of home prices.
- So from 18*6.9 million rows has come to 10 * 2 rows. So the processing optimization has been achieved.
- Now, With that said. Aggregation on the basis summation is going are discrete value and state are categorical variables. Hence, Bar chart will be the best visualization.

Bar graph of the states with highest home values

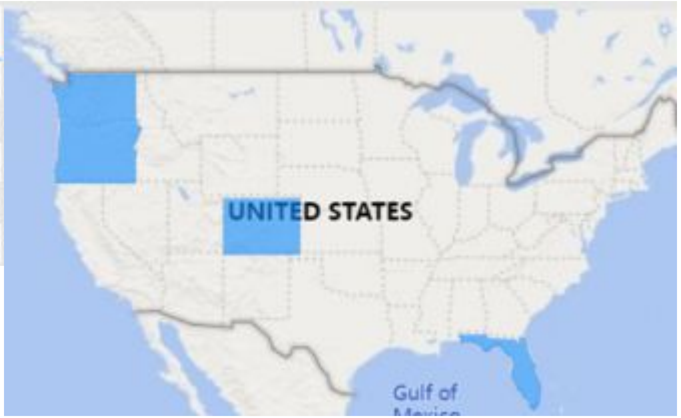


Some other visualisations and conclusions

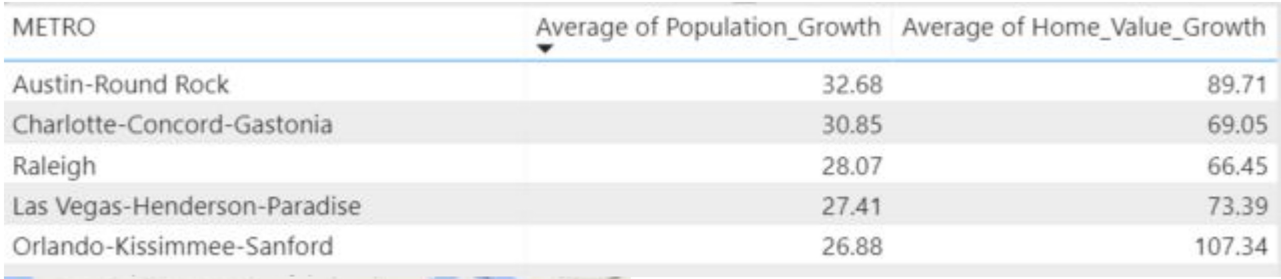
- Top 5 metros by average home value growth



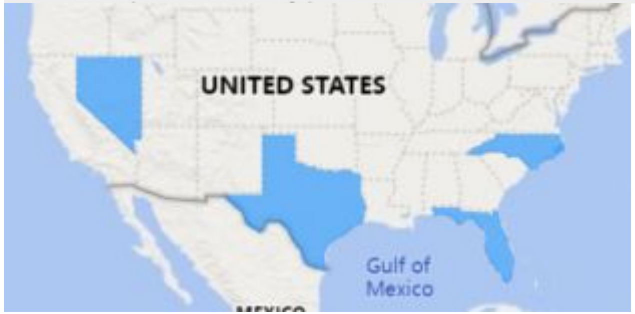
METRO	Average of Population_Growth	Average of Home_Value_Growth
Seattle-Tacoma-Bellevue	6.63	147.76
Miami-Fort Lauderdale-West Palm Beach	10.28	135.22
Portland-Vancouver-Hillsboro	10.51	134.13
Tampa-St. Petersburg-Clearwater	6.20	124.42
Denver-Aurora-Lakewood	7.34	115.26



- Top 5 metros by population growth



METRO	Average of Population_Growth	Average of Home_Value_Growth
Austin-Round Rock	32.68	89.71
Charlotte-Concord-Gastonia	30.85	69.05
Raleigh	28.07	66.45
Las Vegas-Henderson-Paradise	27.41	73.39
Orlando-Kissimmee-Sanford	26.88	107.34


UNITED STATES
Gulf of Mexico
MEXICO