# Week 22 Challenge – Big data

<img src="ReadMe Pics/Cover Pic 1.png" width="817" height="456">

# Background

In this challenge, we were tasked to use SparkSQL to determine key metrics about home sales data. To accomplish this goal, we needed to use Spark to create temporary views, partition data, and cache and uncache data.

# Preprocessing

To start, home\_sales\_revised.csv was downloaded from an S3 bucket and read into a DataFrame. The datatypes were then viewed to assess if they needed to be updated.

<br>

<img src="ReadMe Pics/Pic 1.png" width="919" height="452">

<img src="ReadMe Pics/Pic 2.png" width="349" height="225">

The data was then cast to the appropriate datatypes, and a ‘year\_sold’ column was established and populated.

<br>

<img src="ReadMe Pics/Pic 4.png" width="1025" height="449">

<img src="ReadMe Pics/Pic 3.png" width="372" height="249">

Finally, a temporary view of the DataFrame was established to prepare the data to be queried.

<br>

# Quiring the data using SparkSQL

The data was queried to determine key metrics about home sales data. The completed queries are as follows:

<br>

#### Query # 1 - What is the average price for a four-bedroom house sold for each year?

<img src="ReadMe Pics/Pic 5.png" width="215" height="145">

#### Query # 2 - What is the average price of a home for each year the home was built, that has three bedrooms and three bathrooms?

<img src="ReadMe Pics/Pic 6.png" width="214" height="220">

#### Query # 3 - What is the average price of a home for each year the home was built, that has three bedrooms, three bathrooms, two floors, and is greater than or equal to 2,000 square feet?

<img src="ReadMe Pics/Pic 7.png" width="218" height="215">

#### Query # 4 - What is the average price of a home per "view" rating having an average home price greater than or equal to $350,000? Determine the run time for this query.

<img src="ReadMe Pics/Pic 8.png" width="272" height="476">

## The data was cached, and Query #4 was re-run to check for processing time improvements

<img src="ReadMe Pics/Pic 9.png" width="286" height="480">

## The cached data was partition by the ‘date\_built’ field and Query #4 was re-run to check for processing time improvements

<img src="ReadMe Pics/Pic 10.png" width="285" height="481">

# Summary

When comparing the processing time results for Query #4 query between the original dataset, the cached data, and the partition/cached data, increased processing times were observed. The cached data showed a 55.61% performance improvement, and the partition/cached data showed a 62.66% performance improvement over the processing time of the original query. The runtimes between the three queries can be seen below.

<br>

<img src="ReadMe Pics/Pic 11.png" width="366" height="274">

<br>

<img src="ReadMe Pics/Cover Pic 4.png" width="827" height="312">