# [https://avatars2.githubusercontent.com/u/4156894?v=3&s=100](http://www.calstatela.edu/centers/hipic) CIS4560 Term Project Tutorial CIS4560 Term Project Tutorial

#### Authors: Adrian Soriano, Christian Momdjian, Joshua Ng, Louis Banegas, Marc Calvillo

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 11/16/2022

**Lab Tutorial (Team 2)**

Christian Momdjian ([cmomdji@calstatela.edu](mailto:cmomdji@calstatela.edu))

**Housing Costs in California**

**Objectives**

In this hands-on lab, you will learn how to:

1. Get data using *wget* and save it to the Oracle cluster
2. Upload the file to HDFS from Oracle
3. Create a database table of all the data
4. Create another table with filtered data to export to a new file
5. Visualize the exported data using Microsoft PowerBI

Step 1: Get data using WGET and save it to the Oracle cluster

This step details how you are going to get the dataset to work with for this tutorial. The original dataset is very large at around 49GB, which makes it hard to download the data straight from the source URL. So, to avoid issues with downloading it, we pulled 2GB worth of the data (which holds all our California records along with some other states) and stored it in a Dropbox location where it can be downloaded using the WGET tool:

1. Connect to the Oracle cluster using your username (Don't use cmomdji):

ssh **cmomdji**@144.24.14.145

1. Download the data from the Dropbox location we provided:

wget [https://www.dropbox.com/s/3xt3up4fve78me8/dolthub\_us-housing-](https://www.dropbox.com/s/3xt3up4fve78me8/dolthub_us-housing-prices_main_sales_REDUCED2.csv?dl=0) prices\_main\_sales\_REDUCED2.csv?dl=0

1. Rename the downloaded file:

mv dolthub\_us-housing-prices\_main\_sales\_REDUCED2.csv?dl=0

us-housing-prices.csv

1. Make sure the downloaded file is now in the Oracle cluster:

ls -al

Step 2: Upload the file to HDFS from Oracle

Now that we have the data, we need to upload it to HDFS so we can store it properly and avoid taking up too much space on the Oracle cluster.

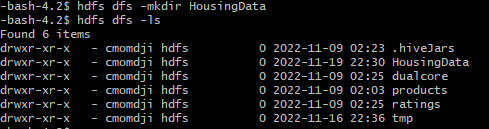
* + First start by creating a new directory to contain the data:

hdfs dfs -mkdir HousingData

* + Check to see that the directory was created in HDFS:

hdfs dfs -ls

\* Ignore the other directories. Just make sure HousingData appears like below



* + Next, we need to upload the data to HDFS:

hdfs dfs -put us-housing-prices.csv HousingData/

* + Make sure the file is uploaded:

hdfs dfs -ls HousingData/



* + Once the file is uploaded to HDFS, make sure to delete it from Oracle:

rm -rf us-housing-prices.csv

Step 3: Create a database table using the data with Hive (Beeline)

With our data loaded in HDFS, we can turn to Hive to create the database table and query the data.

1. Connect to Hive using the ‘beeline’ command:

beeline

1. Remember to use your database. Replace ‘cmomdji’ with your username:

use **cmomdji**;

1. Next, we’ll create the table with all the fields from the file. Again, make sure to replace **cmomdji** with your username:

DROP TABLE IF EXISTS housing\_data;

CREATE EXTERNAL TABLE housing\_data(state STRING, property\_zip5 STRING, property\_street\_address STRING, property\_city STRING, property\_county STRING, property\_id STRING, sale\_datetime STRING, property\_type STRING, sale\_price double, seller\_1\_name STRING, buyer\_1\_name STRING, building\_num\_units double, building\_year\_built double, source\_url STRING, book STRING, page STRING,transfer\_deed\_type STRING, property\_township STRING, property\_lat STRING, property\_lon STRING, sale\_id STRING, deed\_date STRING, building\_num\_stories double, building\_num\_beds double, building\_num\_baths double, building\_area\_sqft STRING, building\_assessed\_value double, building\_assessed\_date STRING, land\_area\_acres STRING, land\_area\_sqft STRING, land\_assessed\_value STRING, seller\_2\_name STRING, buyer\_2\_name STRING, land\_assessed\_date STRING, seller\_1\_state STRING, seller\_2\_state STRING, buyer\_1\_state STRING, buyer\_2\_state STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

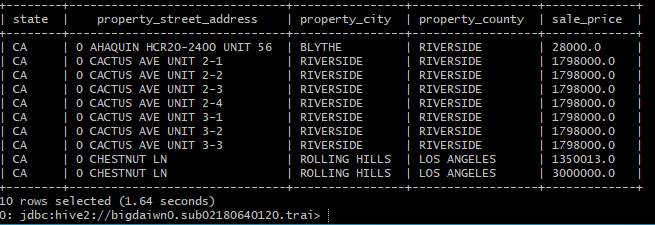
STORED AS TEXTFILE LOCATION '/user/**cmomdji**/HousingData/'

TBLPROPERTIES("skip.header.line.count"="1");

1. Make a test query to see if the table was created with all our data:

select state, property\_street\_address, property\_city, property\_county, sale\_price from housing\_data where state='CA' and sale\_price > 10000 limit 10;

**\* You should see a result similar to the one below**



Step 4: Create a new table with filtered data and save the data to a file

Now that our data is loaded into a database table in Hive, we can filter out the data and save it to a file for us to use later. **Remember to replace cmomdji with your username!**

* Create a new table that filters our data to only properties in California and sold in 2019 or later. We also want to make sure the sale price is over $100,000 to show realistic housing costs and have only residential properties such as single-family homes and condominiums, rather than, for example, vacant lots or office spaces:

DROP TABLE IF EXISTS california\_housing\_records;

CREATE TABLE california\_housing\_records

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE LOCATION '/user/**cmomdji**/HousingData/california\_housing\_records/' AS SELECT state, property\_street\_address, property\_city, property\_county, sale\_price, property\_type, sale\_datetime FROM housing\_data WHERE state='CA' AND sale\_datetime >= '2019-01-01 00:00:00' AND sale\_price > 100000 AND (property\_type LIKE '%RESIDENCE%' OR property\_type LIKE '%CONDO%');

* Check to see if the file was saved to HDFS (switch to the Oracle terminal):

hdfs dfs -ls HousingData/california\_housing\_records

* Download the file from HDFS to Oracle:

hdfs dfs –get /user/**cmomdji**/HousingData/california\_housing\_records/000000\_0 california\_housing\_records.csv

* Verify that the file is there:

ls –al

* Finally, open a new terminal and go to a directory where you would like to download the file. Once there, enter the following command to download the file to your computer (note the ‘.’ at the end):

scp [**cmomdji**@144.24.14.145:/home/**cmomdji**/california\_housing\_records.csv](mailto:cmomdji@144.24.14.145:/home/cmomdji/california_housing_records.csv) .

* You should now see the file in your current directory. This will be needed for the next section.

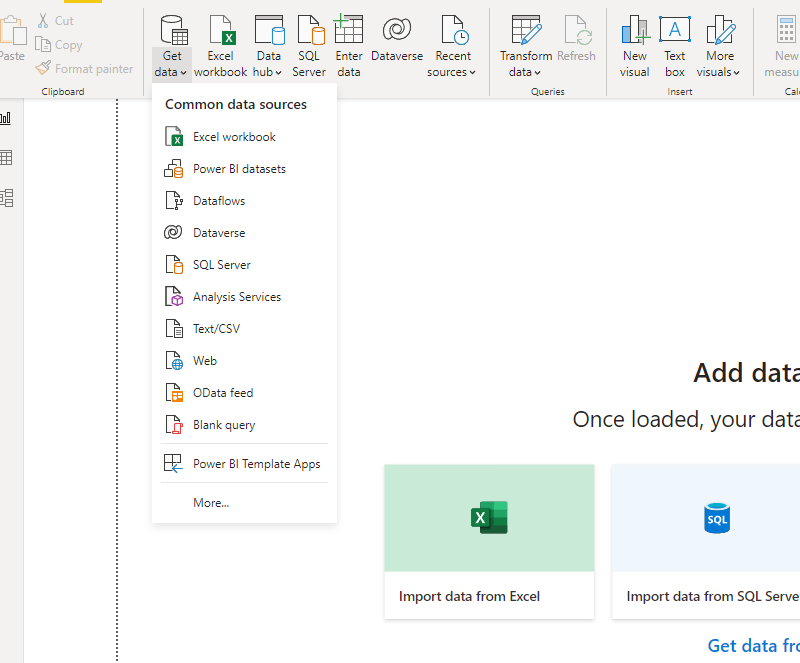
Step 5: Visualization

Now that we have our data filtered and downloaded to a new file on our computer, we can start to create useful visualizations to find helpful information about the data we have. The graphs we’ll make are going to be simple and straightforward, but they apply useful skills in gathering data, filtering through it and highlighting information about it that can potentially be helpful to us. This section requires Excel and PowerBI.

* Open the file you downloaded to your computer in Excel. Add a column at the top of the file above the first row and enter the following titles for each column:

**state**, **street\_address**, **city**, **county**, **sale\_price**, **property\_type**, **sale\_datetime**

* Download PowerBI: <https://powerbi.microsoft.com/en-us/downloads/>
* Once PowerBI is downloaded on your computer, open it and select the “Get data” dropdown on the top left corner. From the dropdown, select “Text/CSV”.



* Select the **california\_housing\_records.csv** file we downloaded earlier. Click “Open” and then click “Load”.

1. Click on the map icon to begin setting up our visualization

**Graphical user interface, application

Description automatically generated**

1. Click on the arrow on the left of the california\_housing\_records dataset to expand the menu and view the fields available

**Graphical user interface, application

Description automatically generated**

1. Drag the property\_city field to the Location area and the sale\_price field to the Bubble size area

**Graphical user interface, application, Word

Description automatically generated**

1. Use the drop-down menu under Bubble size to choose the **Average** of sale\_price. This will help us show the average property price in each city within the county of Los Angeles.

**Graphical user interface, application

Description automatically generated**

1. Drag the property\_city field to the Filters section and click **Select All**.Next, deselect both **LOS ANGELES** selections. This will keep the data focused on the surrounding cities of Los Angeles County and ignore downtown. This data will be explored separately.

**Graphical user interface, application, Word

Description automatically generated**

1. To add gradient color to our bubble visualization, click on the format button under Visualizations followed by the conditional formatting option under Bubbles and Colors.

**Graphical user interface, application, Word

Description automatically generated**

1. In the window that pops up, ensure that the bubbles are based on the **Average of sale\_price** and that the summarization is the **Average**.Next, set the color for **Minimum** to the lightest shade of red and set the color for **Maximum** to the darkest shade of red.

**Graphical user interface, application, Word

Description automatically generated**

1. Expand the map using its corners and enlarge it until all the space is used to get a better visual. Other cities around the world are marked, but these are simply cities with the same name. We will be focusing on the cities in the Los Angeles County area.

**Graphical user interface, application

Description automatically generated**

1. Zoom in using the CTRL + the middle click-wheel to get a clear view of Los Angeles County. The cities with the largest and darkest colored circles contain the most expensive properties currently on the market.

**Graphical user interface, application, map

Description automatically generated**

1. Click on File-> Export, then **Export to PDF** to save a copy of the data currently being viewed.

References

1. Dataset Source: <https://www.dolthub.com/repositories/dolthub/us-housing-prices/data/main>
2. Reduced Dataset: [https://www.dropbox.com/s/3xt3up4fve78me8/dolthub\_us-housing-](https://www.dropbox.com/s/3xt3up4fve78me8/dolthub_us-housing-prices_main_sales_REDUCED2.csv?dl=0)prices\_main\_sales\_REDUCED2.csv?dl=0
3. Github: <https://github.com/cmomdji/CIS-4560-Team-2>