Large Scale Data Management

Semester Exercise

Instructor: Dr. Lefteris Sidirourgos

Overview

You are requested to follow a series of steps to experience at first hand the process of setting up a large-scale DBMS and a distributed OLAP system, load data, query that data, and finally perform an analytical task.

"The exercise also contained setting up and operating a MonnetDB database, although because of its low use in the industry it was cut for this GitHub Repository."

Loading the Data

You will be given a data file, namely: zillow.csv

The origin of the data is from Zillow. It consists of a dataset including ~100K listings from the BOSTON, MA area in CSV format. You are requested to load these files in PySpark.

Analysis

Analysis consists of the following operations:

1) Extract number of bedrooms. You will implement a UDF that processes the `facts_and_features` column and extracts the number of bedrooms.
2) Extract number of bathrooms. You will implement a UDF that processes the `facts_and_features` column and extracts the number of bathrooms.
3) Extract sqft. You will implement a UDF that processes the `facts_and_features` column and extracts the sqft.
4) Extract type. You will implement a UDF that processes the `title` column and returns the type of the listing (e.g., condo, house, apartment)
5) Extract offer. You will implement a UDF that processes the `title` column and returns the type of offer. This can be `sale`, `rent`, `sold`, `foreclose`.
6) Filter out listings that are not for sale.
7) Extract price. You will implement a UDF that processes the `price` column and extract the price. Prices are stored as strings in the CSV. This UDF parses the string and returns the price as an integer.
8) Filter out listings with more than 10 bedrooms
9) Filter out listings with price greater than 20000000 and lower than 100000
10) Filter out listings that are not houses.
11) Calculate average price per sqft for houses for sale grouping them by the number of bedrooms.