DOCUMENTATION REGRADING DASK -DATAFRAME QUERY PLANNING FEEDBACK

BY: Kritika and Emmanuel

Introduction:
This is an overview of a dask library data frame query planning feature implemented in the new version release 2024.3.0, which enabled query planning. Users are also required to use the latest dask-expr version, which supports the optimization of data set queries.

Project:
Firstly, we used the Dask and Pandas library to read three dataset containing housing sector prices. Secondly, we executed a query to get the “retailvalue” column from three datasets and calculated the time it took to query the data.

Objective:
The aim of this document is to demonstrate the efficiency of using the Dask library rather than pandas cause of the data frame query planning feature. It shows how querying data from the dataset is more time-efficient when using Dask than Pandas.

Coding Dataset: The project contains three house price prediction datasets which we are using for the implementation of our code with Dask framework and pandas package library.

Data that we used for both the libraries comparison:
Link to Kagle dataset used: Utrecht housing dataset (kaggle.com)
Platform Used: Jupyter Notebooks

Code:
Comparing Dask with Pandas Library: Given in the comment section

Comparing both Results: Query for “retail value” output in the dataset
From the results below, we observe that the time taken to query the result from the dataset using Pandas is more than the time taken by Dask.

Pandas:
Time taken by Pandas to query 'RetailValue' column from all datasets: 0.04046487808227539

Dask:
Time taken by Dask to query 'RetailValue' column from all datasets: 0.020566225051879883

Conclusion:
With the help of Dask framework query planning, the Dask library is more time efficient rather than pandas library when used to query data from datasets. Even for reading a dataset file, Dask was proven to be faster than Pandas for our project. While running the query dataframe the pandas take more time than Dask library whereas for query execution Dask takes way less amount of time than Pandas.