Assignment-2

Tingyan Lu s4226992

1 Introduction

In this assignment, I did a experiment with the NYC Cab dataset.I downloaded the NYC Cab dataset into Pandas, SQLite, and DuckDB. When I ran the sql, I compared the run time for each. Finally, I ran a machine learning algorithm in SQLite and DuckDB.

2 Loading The Data

In this part, we created Pandas Dataframe from Parquet, load data into SQLite from Pandas Dataframe, then created table and load data from Parquet into DuckDB.

Figure 1: Loading the data for DuckDB

3 Queries

3.1 Query Q1

In this part, I ran the code that provided by teacher.This code aimed to count the number of unique values in each table of the dataset.For SQLite, It operated at the database level, using COUNT(DISTINCT) to count the number of unique values in each column. For Pandas, I also used the unique() function to directly evaluate the unique values in each column of the datasets.

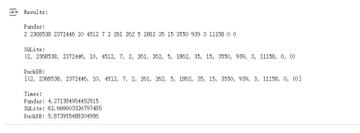

Figure 2: time and results for query Q1

3.2 Query Q2

In this part, the aim of this code was to execute queries using SQLite, DuckDB, and Pandas to count the number of taxi orders each day and hour and return the minimum, average, and maximum number of orders per day and hour.

For SQLite, I used the strftime() function to extract date and time like strftime(\%H',...) extracts the "day" from the date, while strftime (\%H',...) extract the hour.The SQL with statement defined a temporary table called events, which is used in subsequent queries.

For DuckDB, I used the EXTRACT() function to retrieve date and time.A standard SQL syntax of EXTRACT(DAY FROM) and EXTRACT(HOUR FROM) are different phrase. It also define a temporary table to restore the statement.

Figure 3: time and results for Q2

3.3 Comparison between SQLite, DuckDB and Pandas

MysQl is based in hash approach or an index scan to evaluate the unique values.In this part, I did not use a index to ran SQL so that they performed too much execution times.It was low efficiency.

Pandas used memory function when they load data. It is their advantage. They use RAM to operate in Datasets directly. They are the second optimized system in Three datasets.They can faster when they work.

DuckDb is highest optimized datasets for two queries. They are the least time when compared with others datasets. It use vectorized execution that can speed operations time.

datasets	q1 time	q2 time
Pandas	4.271	1.348
SQLite	62.666	21.118
DuckDb	5.873	0.467

Table 1: Execution time

4 Machine learning: Fare Estimation

4.1 For SQLite

For SQLite, I wrote a query that randomly select 1,000 trip records to ensure that both fare amounts and trip distances are greater,Then

I fit the query results retrieved from the SQLite database into a Pandas DataFrame format to show easily.

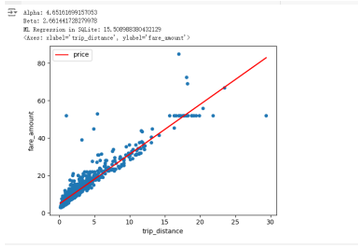

Figure 4: image of SQLite

4.2 For DUCKDB

For DuckDB, the first step is to computer the average fare and distance in yellow-trip-data.

Then, I created a SQL to choose a random sample of 1000 records in this datasets that aim is to visualise $\text{trip}_d istanceandfare - amount$.

finally, I used the pandas library to change the SQL query to fit it that is benefit to show manipulate and visualize.

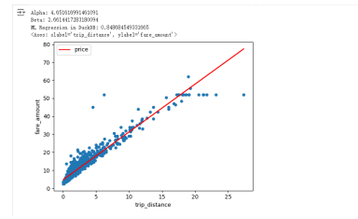

Figure 5: image of DUKCDB

4.3 Difference between SQLite and DuckDB

For replacement part, DuckDB takes advantages of their function that is columnar storage. This means it can process the involved columns.it can avoid that they ran irrelevant data which an speed up their runtime, more efficiency. When I used the SQLite, they sow a slowly time. because it use row storage so that it should read and write all data When I change one column.So when we apply it in largest dataset, they use unnecessary data to access and modify.