Name: Nikhil Prema chandra rao
Student ID: 02105149

Title: CSV Query Performance Benchmark - Documentation

---

1. Introduction

This project benchmarks the performance of SQL queries executed on a set of CSV files of varying sizes. The goal is to evaluate how query execution time varies with dataset size. The project involves loading CSV files into an SQLite database, executing a series of queries, and visualizing the execution time for each query.

Dataset Overview

The dataset used in this project is a simulated salary tracker, containing information about individuals, their job titles, earnings, and employment status. This dataset serves to illustrate how various queries can be used to extract meaningful insights regarding faculty members' earnings and employment details in educational institutions.

Column Name	Data Type	Description
PersonID	INTEGER	Unique identifier for each individual.
PersonName	TEXT	Full name of the individual.
JobTitle	TEXT	Title of the individual's job position.
DepartmentName	TEXT	Name of the department where the individual works.
SchoolName	TEXT	Name of the school where the individual is employed.
SchoolCampus	TEXT	Campus location of the school.
Earnings	INTEGER	Annual earnings of the individual.
BirthDate	DATE	Birthdate of the individual.
StillWorking	TEXT	Employment status ('yes' or 'no').

---

2. Queries Performed

Query	SQL Statement	Execution Time (1 MB)	Execution Time (10 MB)	Execution Time (100 MB)
Query 1	`SELECT PersonName FROM salary_tracker WHERE BirthDate < '1975-01-01' AND Earnings = (SELECT MAX(Earnings) FROM salary_tracker WHERE PersonID = salary_tracker.PersonID) AND Earnings > 130000;`	0.0009 seconds	0.0115 seconds	0.1135 seconds
Query 2	`sql SELECT PersonName, SchoolName FROM salary_tracker WHERE Earnings > 400000 AND StillWorking = 'no';`	0.0006 seconds	0.0072 seconds	0.0732 seconds
Query 3	`sql SELECT PersonName FROM salary_tracker WHERE JobTitle = 'Lecturer' AND SchoolName = 'University of Texas' AND StillWorking = 'no';`	0.0005 seconds	0.0060 seconds	0.0570 seconds
Query 4	`sql SELECT SchoolName, SchoolCampus, COUNT(*) AS ActiveFacultyCount FROM salary_tracker WHERE StillWorking = 'yes' GROUP BY SchoolName, SchoolCampus ORDER BY ActiveFacultyCount DESC LIMIT 1;`	0.0006 seconds	0.0059 seconds	0.0580 seconds
Query 5	`sql SELECT PersonName, JobTitle, DepartmentName, SchoolName, MAX(Earnings) AS MostRecentEarnings FROM salary_tracker WHERE PersonName = 'Nikhil Premachandra Rao' GROUP BY PersonName, JobTitle, DepartmentName, SchoolName;`	0.0004 seconds	0.0045 seconds	0.0417 seconds

Query 6	`sql SELECT DepartmentName, AVG(Earnings) AS AverageEarnings FROM salary_tracker GROUP BY DepartmentName ORDER BY AverageEarnings DESC LIMIT 1;`	0.0016 seconds	0.0213 seconds	0.2124 seconds

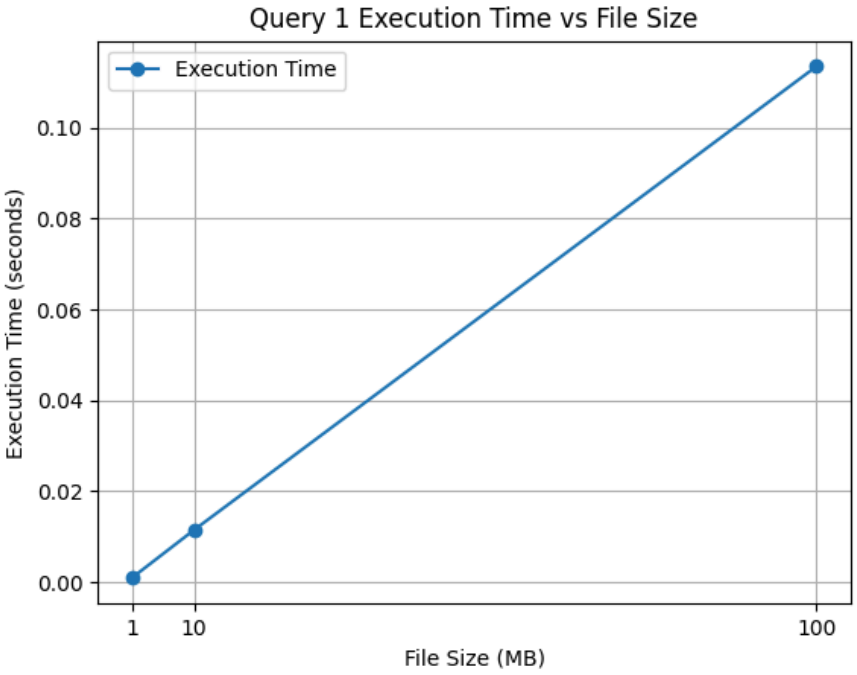
---

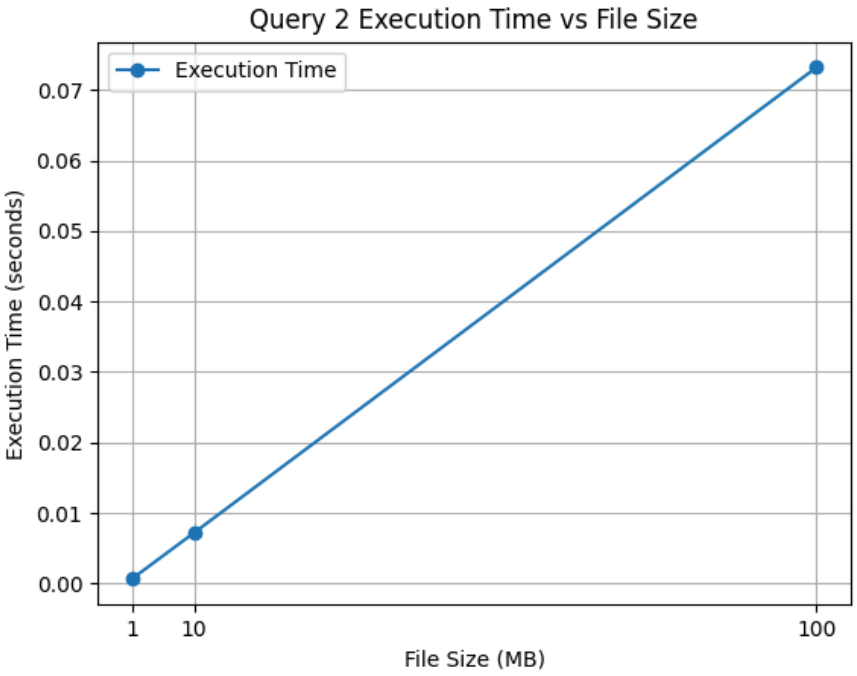
3. Time Measurement Methods

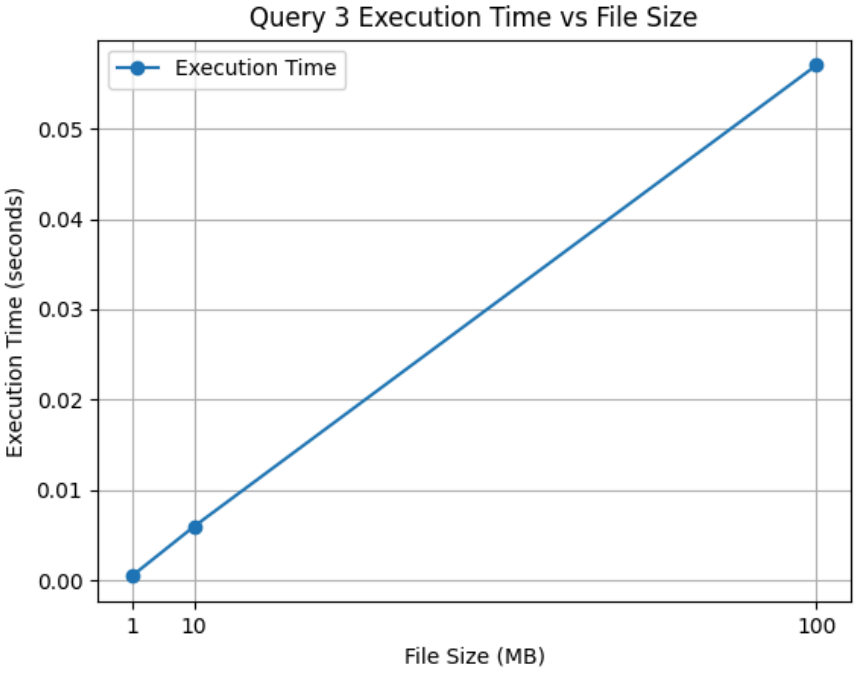
- **Execution Time Measurement**: Python's built-in timing routines were used to measure each query's execution time. Each query was timed both before and after it was executed, and the difference between the two was computed to get the overall execution time.
- **Data Loading Time**: The query execution timings did not include the time it took to load each dataset into the SQLite database, but it was measured anyway. An overview of the performance of loading bigger datasets is provided by this.

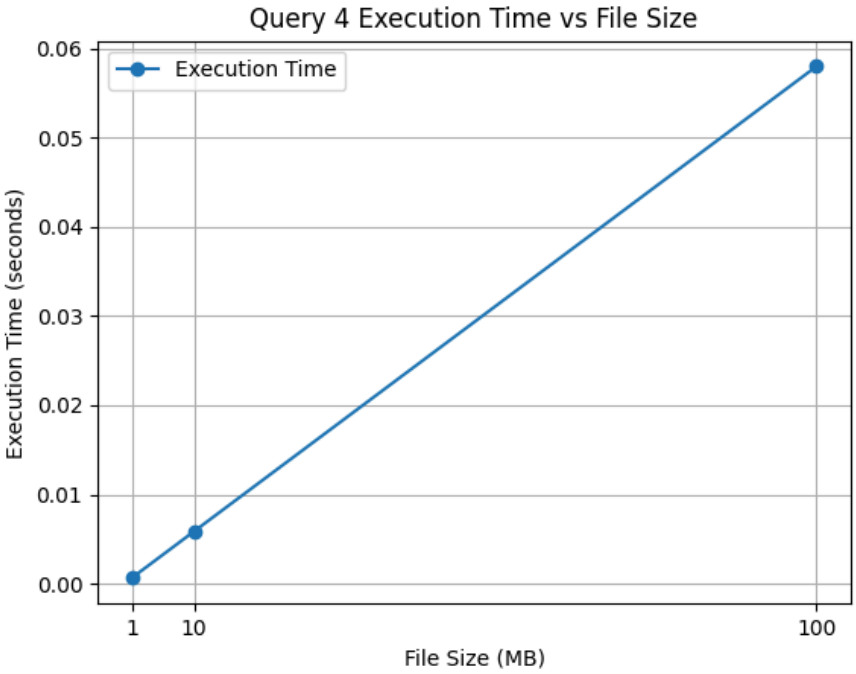
---

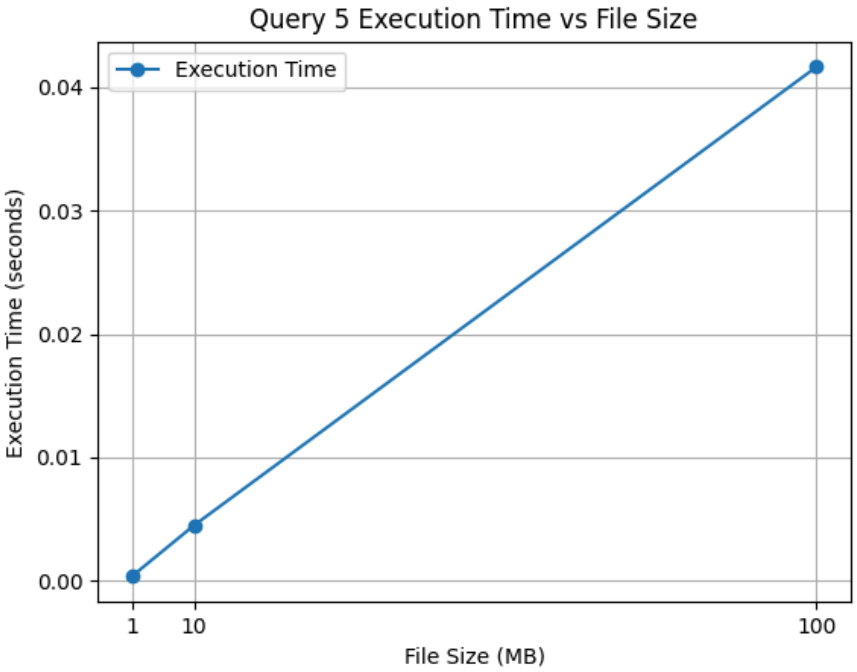
4. Graphs

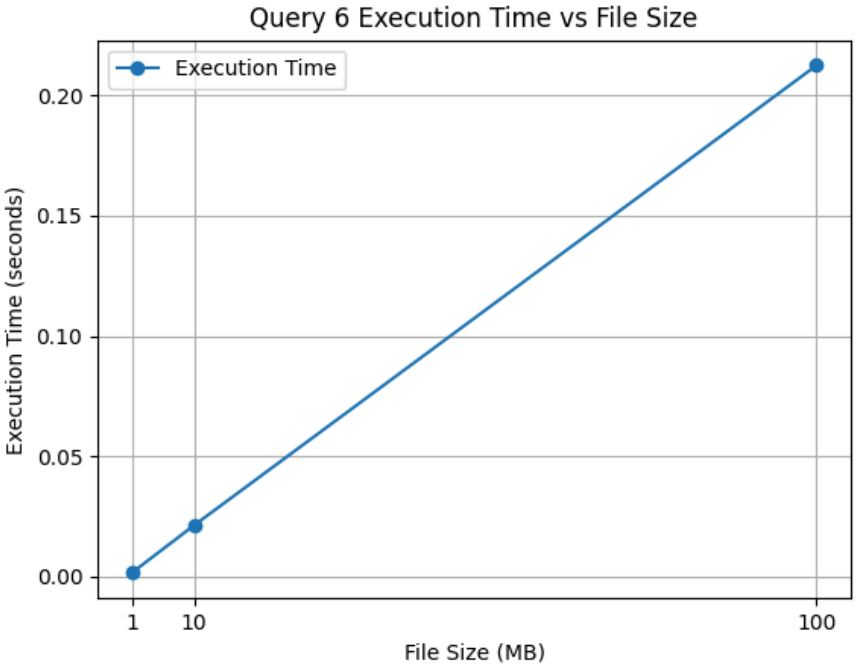

Query 1 Execution Time vs File Size
Execution Time
0.10
0.08
0.06
0.04
0.02
0.00
Execution Time (seconds)
1
10
100
File Size (MB)


Query 2 Execution Time vs File Size
Execution Time
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Execution Time (seconds)
1
10
100
File Size (MB)


Query 3 Execution Time vs File Size
Execution Time
0.05
0.04
0.03
0.02
0.01
0.00
Execution Time (seconds)
1
10
100
File Size (MB)


Query 4 Execution Time vs File Size
Execution Time
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Execution Time (seconds)
1
10
100
File Size (MB)


Query 5 Execution Time vs File Size
Execution Time
0.04
0.03
0.02
0.01
0.00
Execution Time (seconds)
1
10
100
File Size (MB)


Query 6 Execution Time vs File Size
Execution Time
0.20
0.15
0.10
0.05
0.00
Execution Time (seconds)
1
10
100
File Size (MB)

5. Conclusion

This documentation covers the queries and techniques used to calculate execution times. The results show a strong relationship between dataset size and query performance.