Proof of Concept

"Performance Evaluation of Big Data Queries: A Comparative Analysis of Execution Times Across Multiple Databases"

Aim: The aim of this study is to evaluate and compare the performance of fundamental CRUD operations—Insert, Update, Delete, and Select—across different databases for datasets with different volumes. Specifically, we contrast the performance of SQL Server, a relational database model, with Cassandra and MongoDB, two NoSQL models, to determine their effectiveness in handling massive and heterogeneous data generated across various sectors. The study also seeks to assess how these databases address challenges related to data volume, variety, and horizontal scaling.

Dataset used: Massive Bank Dataset

The dataset is described as follows:

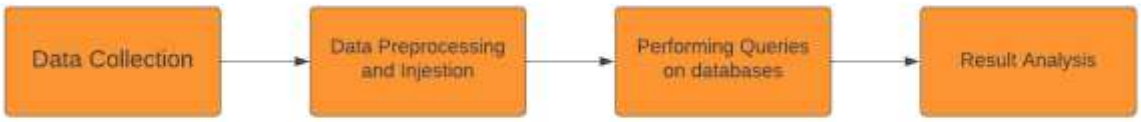
- Date - The date on which the transaction took place.
- Domain - Where or which type of Business entity made the transaction.
- Location - Where the data is collected from
- Value - Total value of transaction
- Count of transaction.

For example, in the very first row, the data can be read as:

" On the first of January 2022, 1932 transactions of summing up to INR 365554 from Bhuj were reported "

	A	B	C	D	E
1	Date	Domain	Location	Value	Transaction_count
2	01-01-2022	RESTRAUNT	Bhuj	365554	1932
3	01-01-2022	INVESTMENTS	Ludhiana	847444	1721
4	01-01-2022	RETAIL	Goa	786941	1573
5	01-01-2022	INTERNATIONAL	Mathura	368610	2049
6	01-01-2022	RESTRAUNT	Madurai	615681	1519
7	01-01-2022	INTERNATIONAL	Daman	1191092	1813
8	01-01-2022	INTERNATIONAL	Buxar	968883	2098
9	01-01-2022	PUBLIC	Trichy	1030297	606
10	01-01-2022	RESTRAUNT	Kullu	688655	1463

Data Flow:



From Massive Bank Dataset to Databases: Data from CSV files is imported into MySQL, Cassandra, and MongoDB databases using their respective import tools or scripts.

Query Execution: Queries (Insert, Update, Delete, Select) are executed on each database using their dedicated queries.

Evaluation and Analysis: Performance metrics (execution times) for CRUD operations across databases are recorded and analysed to compare their efficiency in handling different volumes and types of data.

Databases and Technologies:

- MySQL, Cassandra, MongoDB: The primary databases used for data storage and management.
- MySQL Server Management Studio: Utilized for executing and analyzing queries within MySQL.
- CQLSH (Cassandra Query Language Shell): Used for running and evaluating queries in Cassandra.
- MongoDB Compass: A tool for executing and analyzing queries in MongoDB.