1.Introduction

The objective of this project is to design and optimize a relational database capable of storing and analyzing high-volume Futures & Options (F&O) market data from multiple Indian exchanges such as NSE, BSE, and MCX. The dataset represents time-series snapshots of derivative contracts and is intended for analytical use cases such as open interest analysis, volatility tracking, and cross-exchange comparisons.

Given the scale (2.5M+ rows) and financial nature of the data, the design prioritizes **normalization, scalability, and query performance**.

2. Domain Understanding and Data Characteristics

Each tradable contract is uniquely defined by:

- Underlying instrument (e.g., NIFTY, GOLD)
- Expiry date
- Strike price (for options)
- Option type (CE / PE)

Each contract generates **repeated time-series records** containing OHLC prices, volume, and open interest. This makes the dataset:

- Write-heavy
- Time-series oriented
- Highly repetitive without normalization

These characteristics strongly influence the database design choices.

3. Database Design and Normalization (3NF)

The schema is designed in **Third Normal Form (3NF)** to reduce redundancy and improve consistency.

Key Tables:

Exchanges

A separate exchanges table stores trading venues (NSE, BSE, MCX). Avoids repeating exchange names across millions of trade records and enables cross-exchange analytics.

Instruments

The instruments table represents the underlying asset or index (e.g., NIFTY, BANKNIFTY, GOLD).One instrument can have many derivative contracts across different expiries.

Expiries

The expiries table represents unique derivative contracts, defined by expiry date, strike price, and option type.Separating this entity prevents duplication of contract metadata in the trades table

Trades

The trades table is the fact table, storing prices, volume, and open interest for each contract over time.

4. Why Star Schema Was Avoided

A star schema was intentionally avoided for the following reasons:

- Contracts change frequently (weekly/monthly expiries)
- Strike prices and option types increase.
- Star schemas work better for static dimensions, not dynamic data
- Normalized schemas perform better for write-heavy, evolving datasets

Thus, a normalized relational model is more suitable for market data ingestion and analytics.

5. Scalability and Performance Optimization

Indexing Strategy (MySQL)-

Since MySQL does not support BRIN indexes, B-Tree indexes were used on frequently queried columns:

- expiry_dt → contract-centric filtering
- instrument_id → symbol-level aggregation
- exchange_id → cross-exchange analysis

  These indexes significantly improve query performance for analytical workloads.

Partitioning Strategy-

The trades table was partitioned by expiry date (YEAR(expiry_dt)) using RANGE partitioning.

- Derivative contracts naturally expire and become cold data
- Most analytical queries focus on active or recent expiries

6. Query Optimization and Validation

An optimized analytical query was executed to calculate Open Interest (OI) change across symbols and exchanges using expiry-based filtering.

To validate performance improvements, EXPLAIN ANALYZE was executed in MySQL Workbench.

Observed Improvements:

- Index range scans used instead of full table scans
- Only relevant expiry partitions were accessed
- Join operations used PRIMARY KEY lookups
- Overall query execution cost was significantly reduced

7. Analytical Use Cases Supported

The database design supports real-world trading analytics use cases:

- Open Interest buildup
- Option chain volume aggregation
- Volatility tracking using rolling statistics
- Cross-exchange futures price comparison
- Expiry-wise contract activity monitoring

8. Conclusion

This project demonstrates a production-ready database design for high-volume derivatives data. By combining 3NF normalization, expiry-based partitioning, and targeted indexing, the system achieves both scalability and analytical efficiency. The design is flexible enough to support additional exchanges, instruments, and higher data volumes (10M+ rows), making it suitable for real-world financial analytics environments.