📘 PhonePe Transaction Insights End-to-End Data Analytics Project Documentation

Prepared by: Yamini S
Domain: Digital Payments / FinTech Analytics
Tools: Python, MySQL, SQL, Streamlit, Plotly

Executive Summary

India's rapid shift toward digital payments has generated massive volumes of transaction data. **PhonePe Pulse** is an open analytics initiative that exposes this data to understand how users transact, engage, and adopt new financial services.

This project builds a **data-driven analytics platform inspired by PhonePe Pulse**, transforming raw JSON data into **structured insights** through:

- Robust data ingestion pipelines
- Relational database design
- SQL-based analysis
- Interactive dashboards
- Business-oriented reporting

The outcome is a **scalable, explainable, and decision-ready analytics system**.

Problem Statement

With increasing reliance on digital payment platforms, stakeholders need answers to questions such as:

- Which regions contribute most to transaction value?
- How do transaction patterns vary by category and time?
- Which devices dominate user engagement?
- Where is insurance adoption growing or lagging?
- Which districts should be targeted for expansion?

Raw data alone cannot answer these questions.
This project addresses the problem by **structuring, analyzing, and visualizing PhonePe data** for actionable insights.

Objectives

The primary objectives of this project are:

1. Extract and transform raw PhonePe Pulse JSON data
2. Design a relational database optimized for analytics
3. Perform SQL-driven exploratory and business analysis
4. Build an interactive dashboard for insights exploration
5. Generate professional reports for decision-makers

Dataset Overview

Data Source:
PhonePe Pulse GitHub Repository (Open Data Initiative)

Nature of Data

- Semi-structured JSON files
- Hierarchical structure (country → state → district → quarter)
- Time-series data (year & quarter)

Data Categories

- Transactions
- Users
- Insurance

Each category is analyzed at **state-level and district-level granularity**.

Data Extraction & ETL Pipeline

Step-by-Step Process

1. **Data Extraction**
    - Cloned PhonePe Pulse GitHub repository
    - Read nested JSON files using Python

2. **Data Transformation**
    - Parsed hierarchical JSON structures
    - Normalized data into tabular format
    - Created separate DataFrames for each analytical use case
3. **Data Loading**
    - Exported cleaned DataFrames as CSV files
    - Loaded data into MySQL using Python scripts
    - Validated data integrity using sample queries

This ETL pipeline ensures **consistency, reusability, and scalability**.

Database Design & Schema

A **MySQL relational database** was designed with **9 analytics-ready tables**, grouped into three layers:

1) Aggregated Tables

- aggregated_transaction
- aggregated_user
- aggregated_insurance

Purpose:
High-level analysis by state, year, and quarter.

2) Map Tables

- map_transaction
- map_user
- map_insurance

Purpose:
District-level insights and regional comparisons.

3) Top Tables

- top_transaction
- top_user
- top_insurance

Purpose:
Ranking and benchmarking of top-performing regions.

This layered design avoids redundancy and enables **fast analytical queries**.

Business Case Analysis

☐ Case 1: Decoding Transaction Dynamics

Table Used: aggregated_transaction

Analysis Performed:

- Transaction type distribution
- Quarterly transaction trends

Insight:

- Merchant payments and peer-to-peer transfers dominate usage
- Digital payment adoption is consistently growing

☐ Case 2: Device Dominance & User Engagement

Table Used: aggregated_user

Analysis Performed:

- Device brand-wise user distribution

Insight:

- Android devices (Xiaomi, Samsung) account for the majority of users
- Platform optimization should prioritize Android ecosystem

☐ **Case 3: Insurance Penetration Analysis**

Table Used: aggregated_insurance

Analysis Performed:

- State-wise insurance transaction value

Insight:

- Insurance adoption is concentrated in a few states
- Significant untapped potential exists in other regions

☐ **Case 4: Market Expansion Strategy**

Table Used: map_transaction

Analysis Performed:

- District-wise transaction value ranking

Insight:

- Urban districts generate disproportionately high value
- Ideal targets for marketing and infrastructure expansion

- ☐ **Case 5: User Engagement & Growth Strategy**

Table Used: map_user

Analysis Performed:

- District-wise registered user ranking

Insight:

- High-user districts are optimal for feature rollouts and pilot programs

Dashboard & Visualization

An interactive **Streamlit dashboard** was developed with:

- India-level choropleth maps (Transactions, Users, Insurance)
- KPI cards (Total value, count, averages)
- Top-10 states and districts
- Business case analysis tabs
- Database documentation view
- Creator and About sections
- Dynamic PDF report generation

The dashboard enables **self-service analytics** for users.

Reporting & Insights Generation

The system generates **quarterly analytical reports** based on:

- Selected category (Transactions / Users / Insurance)
- Year and quarter filters

Reports include:

- Executive summary
- Key metrics
- Top-performing regions
- Business interpretation of trends

This bridges the gap between **technical analysis and business decision-making**.

Key Outcomes

- Built an end-to-end analytics pipeline
- Converted raw JSON into structured insights
- Designed scalable SQL schemas
- Developed a PhonePe Pulse–style analytics platform
- Strengthened data storytelling and analytical thinking skills

Limitations & Future Enhancements

☐ **Current Limitations**

- Static historical data
- No real-time streaming

☐ **Future Enhancements**

- Real-time data ingestion (Kafka / APIs)
- Predictive modeling for transaction growth
- Fraud detection analytics
- User segmentation models

Conclusion

This project demonstrates how **data engineering, SQL analytics, and visualization** can be combined to build a **real-world FinTech analytics solution**. It closely mirrors industry-grade platforms like PhonePe Pulse and highlights the power of data-driven decision-making.

Author

Yamini S, B. Sc (CT), (MCA)
Aspiring Data Scientist.