Implementation & Deviations from Initial Plan

Big Data Management and Processing

Maria Christofi
University of Nicosia
christofi.m9@live.unic.ac.cy

Refinement of the Initial Proposal

As the project progressed, the initial dataset and conceptual proposal were refined and implemented through concrete technical steps. Bitcoin transaction dataset was sourced from Google BigQuery public datasets, which provide scalable access to large volume blockchain data. Instead of performing all preprocessing locally, data cleaning and transformation were executed directly within Google BigQuery, leveraging its distributed query engine to efficiently handle type conversions, filtering, and aggregation at scale. This approach reduced local computational overhead and ensured that the dataset was standardized before downstream analysis.

Once the dataset was cleaned and transformed into a daily-level structure suitable for time-series analysis, it was exported as a CSV file and ingested into MongoDB for NoSQL storage and further analytics using Python. This implementation remains consistent with the original proposal's objectives by combining cloud-based Big Data processing with NoSQL ingestion and analytical workflows.

In the processing stage, data cleaning and transformation were performed using Python and Pandas prior to database insertion. These steps included timestamp normalization, numeric type conversion, handling missing values, and ensuring a consistent schema suitable for analytical queries. Indexes were created in MongoDB to optimize time-series queries and anomaly detection workflows.

Batch analytics were then applied through MongoDB aggregation pipelines and Apache Spark, enabling grouped computations such as yearly transaction averages, fee aggregation, and detection of abnormal transaction spikes. These steps replaced purely conceptual descriptions with executable analytics pipelines.

Deviations from the Initial Project Plan

Some deviations from the original project proposal occurred during implementation, which is a natural outcome of exploratory data science and Big Data projects.

1. Data Granularity Adjustment
The original proposal planned to focus on minute-level transaction data. During exploration, it became evident that daily aggregation provided clearer insights into long-term network growth, trends, and anomalies while significantly reducing computational complexity. As a result, the analysis focused on daily transaction metrics rather than minute-level streaming data.

2. Emphasis on Transaction Activity Over Price Data
While the proposal initially referenced price and volatility analysis, the available dataset did not include bid/ask or direct market price information. Consequently, the analysis pivoted toward on-chain activity metrics such as transaction count and transaction fees, which still provide strong indicators of network usage, congestion, and market behavior.

3. Limited Use of Streaming Analytics
Although the proposal included streaming data processing as a possible extension, the final implementation focused primarily on batch analytics. This decision was made due to the historical nature of the dataset and the project's emphasis on trend detection and anomaly identification over long time horizons. Nevertheless, the batch analytics approach still aligns with Big Data processing principles and MapReduce style aggregation.

Reflection

These deviations ultimately strengthened the project by allowing the data itself to guide analytical decisions. Rather than strictly adhering to the original plan, the project evolved toward a more realistic, interpretable, and technically robust analysis. This adaptive process reflects a core principle of data science: meaningful insights often emerge through iterative exploration rather than rigid execution of an initial design.