One Page Summary
Stock Data Processing in Delta Lake

STATEMENT

Development of an ACID-compliant storage architecture and streaming processing of stock market data based on Databricks Delta Lake.

PROBLEM DESCRIPTION

Designing the architecture and building storage are important steps in Big Data Solutions engineering. The central issue of these tasks is the choice of persistent storage. Even though Data Lake and Data Warehouse are powerful tools, the classic solutions based on them have a high complexity of application architecture. Just as difficult is the implementation of CRUD scripts in storage. Architectures based on Delta Lake are designed to eliminate these drawbacks. This paper presents the implementation of Delta Lake architecture for stock market data. The stock market is a suitable data model for data streaming, moreover, there is a wide scope for Business Intelligence.

HIGH LEVEL OVERVIEW OF STEPS

- Data stream initialisation;
- Delta Lake pipeline setup (Bronze -> Silver -> Gold);
- ACID Jobs on Gold layer;
- Timetravel on Gold layer;
- Visualisation.

BIG DATASET

S&P 500
Historical stock data for all current S&P 500 companies

SOFTWARE

- macOS Big Sur Version 11.3.1 Darwin Kernel Version 20.4.0
- Apache Spark: spark-3.1.1-bin-hadoop3.2
- Python 3.9.2, Clang 12.0.0 (clang-1200.0.32.29) on darwin

Python packages:
- deltalake 0.4.7
- pandas 1.2.4
- plotly 4.14.3

SUMMARY

Delta Lake is a new chapter in the development of Big Data solutions. Combination of convenient DataFrame interface, storage organisation transparency, ACID-compatibility, batch-stream processing, metadata as DataFrame opens up opportunities for building extremely large, scalable and durable storages and Big Data processing systems in financial and other fields.

LINKS

YouTube Brief Demo Link | YouTube Full Demo Link

Github Link