Real-Time Stock Price Streaming Pipeline Documentation

# 1. Overview

This document describes the architecture, setup, and operational flow of a real-time data pipeline that fetches stock price data using the Finnhub API, sends it to a Kafka topic, processes it with Spark Structured Streaming, stores it in PostgreSQL, and visualizes it using BI tools.

# 2. Components

## 2.1 Kafka Producer (finhub\_producer.py)

This Python script fetches stock quotes for a list of predefined symbols from the Finnhub API and sends the JSON responses to a Kafka topic called 'stock\_prices'.

Main fields:

• c: Current price  
• h: High price of the day  
• l: Low price of the day  
• o: Open price of the day  
• pc: Previous close price  
• t: Timestamp  
• symbol: Stock symbol

## 2.2 Kafka & Zookeeper (docker-compose.yml)

Docker Compose is used to bring up Kafka and Zookeeper services. The Kafka container listens on port 9092.

## 2.3 Spark Streaming (spark\_kafka\_alerts.py)

Spark reads the data stream from Kafka, parses JSON messages, and writes the data into a PostgreSQL database table using the JDBC connector.

## 2.4 PostgreSQL Database

A local PostgreSQL database stores the streaming data. The table `stock\_prices` includes columns: c, h, l, o, pc, t, symbol.

## 2.5 ETL to BigQuery (ETL\_schedule.py)

Periodically transfers data from PostgreSQL into Google BigQuery for integration with Looker or other BI tools.

# 3. Setup Instructions

## 3.1 Start Docker Services

Run: docker-compose up -d

## 3.2 Run the Kafka Producer

Run: python finhub\_producer.py

## 3.3 Run Spark Streaming

Run: spark-submit --jars postgresql-<version>.jar spark\_kafka\_alerts.py

## 3.4 PostgreSQL Access

Check tables using pgAdmin or CLI: \dt or SELECT \* FROM stock\_prices

## 3.5 Export to BigQuery

Run: python ETL\_schedule.py after setting GOOGLE\_APPLICATION\_CREDENTIALS and pyarrow

# 4. BI Integration

You can visualize the data using tools like Looker, Power BI, or Metabase. BigQuery acts as the bridge between PostgreSQL and the BI tool.

Document generated on: 2025-06-11 00:49:17