Project Proposal

Real-Time Big Data Streaming with Kafka

Abstract

In today's data-driven world, data streaming has emerged as an essential element for businesses worldwide.

Edris Safari

DSC680 Project2, Milestone1

Topic

This project proposes to demonstrate the ability to manage live streaming data. By manage, we mean to send data from a source or producer to one or more destinations or consumers.

Business Problem

Studies have shown that data-driven organizations are 23 times more likely to acquire customers, six times as likely to retain customers, and 19 times as likely to be profitable (reference 1). For businesses to be competitive, turning to data must be considered as essential to the future of those businesses.

Datasets

To demonstrate live streaming data, we chose to obtain stock market data from yfinance. Other viable options are to get air traffic data, weather report, or any other time-series data that can be analyzed or graphed on the receiving end.

Methods

We will use Apache Kafka as the live streaming application. To use it, there are two options. Install it locally, or use Amazon Web Services, or google confluent. We propose to use the 1st method. The later method is costly; however, it must be considered when the scale requires it.

We will stream 1 year of stock data augmented with the current stock of several companies (i.e. Amazon, Google, Microsoft, etc.) every 10 seconds. On the receiving end, we will calculate the stock's technical indicators and display them on the screen of the consumer. We will create two consumers, one will display one stock, and the other a different stock. This is to demonstrate the multi-consumer scenario.

Ethical Considerations

These are the ethical considerations that are found common.

Data privacy: Ensuring that the data collected is obtained in a legal and ethical manner and that personal information is not misused or disclosed without consent.

Bias and fairness: Ensuring that the algorithms and models used in the project do not perpetuate or amplify biases and discrimination against any group.

Transparency: Ensuring that the data sources, analysis methods, and findings are transparent and easily understandable to all stakeholders.

Security: Ensuring that the data is secure and protected against unauthorized access or theft.

Informed consent: Ensuring that individuals whose data is being used in the project are fully informed about the project's purpose, risks, and benefits and have given their informed consent to participate.

Impact on society: Ensuring that the project's results do not have a negative impact on society or vulnerable populations.

Challenges/Issues

Challenges and issues that we could face in this project must be considered and addressed. This is part of risk assessment that all projects must go through. The table below lists the risks and their mitigation.

Risk	Mitigation
Data Quality	Ensure data quality by performing a preliminary analysis
Data Security	Ensure data is secure both incoming and outgoing. Enable/utilize security measures.
Data Availability	Ensure data is available without interruption or delay that may affect the performance of the analysis
Technical Challenges	Enforce fault tolerance, redundancy, connection integrity
Ethical Violations	Ensure procedures are put in place that will reduce and remove risk of ethical violations by all parties involved.

References

1. https://blog.logrocket.com/apache-kafka-real-time-data-streaming-app/
2. https://kafka.apache.org/intro
3. https://analyticshut.com/kafka-producer-and-consumer-in-python/
4. https://www.youtube.com/watch?v=tFlYoEJsT2k
5. https://www.youtube.com/watch?v=LjjPjT6R9Bg
6. https://medium.com/@alipazaga07/setting-up-your-first-apache-kafka-development-environment-in-google-cloud-in-15-minutes-ffdb1623e125
7. https://www.learningjournal.guru/courses/kafka/kafka-foundation-training/kafka-in-gcp/
8. https://www.youtube.com/watch?v=RYC-7wECMds&list=PLa7VYi0yPIH14oEOfwbcE9_gM5lOZ4ICN
9. https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html
10. https://phoenixnap.com/kb/install-spark-on-ubuntu
11. https://www.makeuseof.com/pyscript-python-visualizations-web/
12. François Chollet. Deep Learning with Python (Kindle Locations 1504-1508). Manning Publications Co.. Kindle Edition.