Real Time Streaming and Analytics Pipeline using MongoDB, Kafka and Power BI

Lambton College

Agenda

- Introduction & Architecture of the project
- Real-time Streaming
- Dataset
- Producer and Consumer for the stream
- Project Demonstration
- Analytics
- Visualization

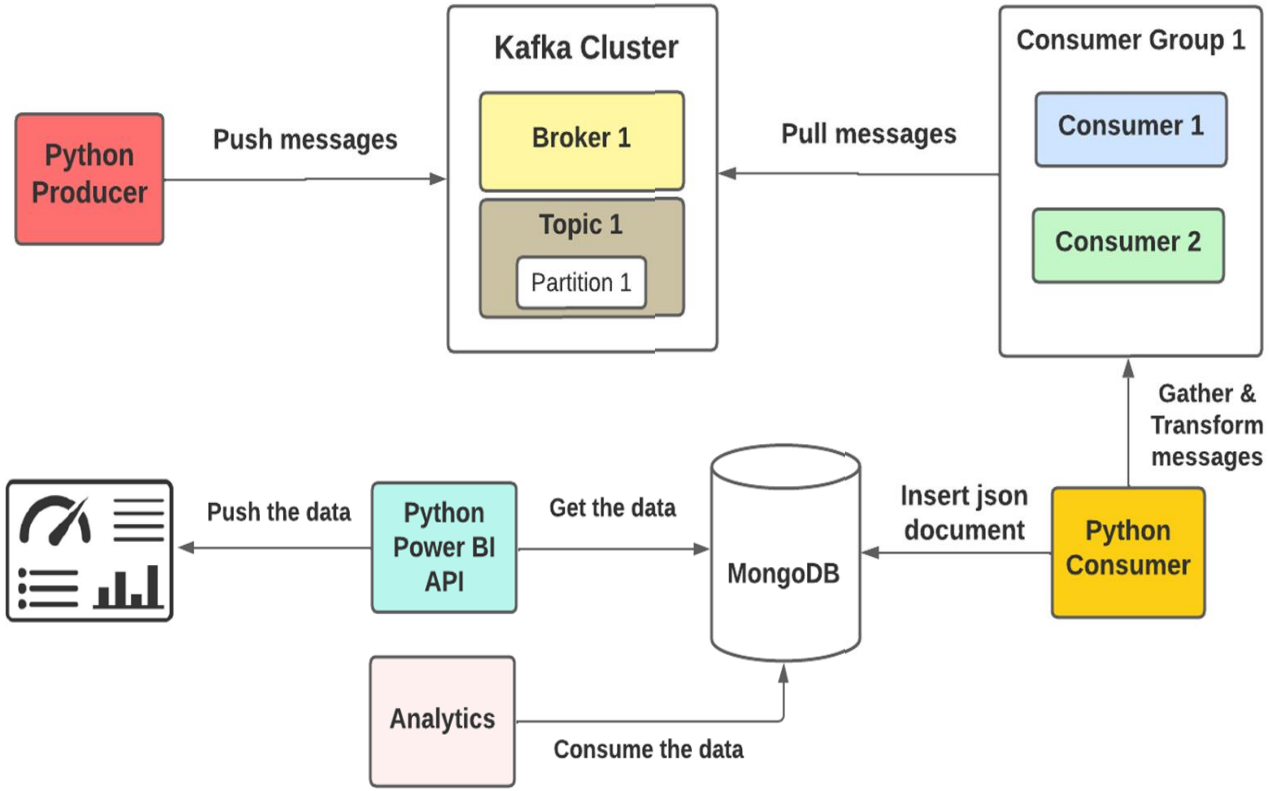
Lambton College

Introduction

- Project goals
- Use case
- Pipeline architecture
- Components
- Persisting data with MongoDB
- Visualization

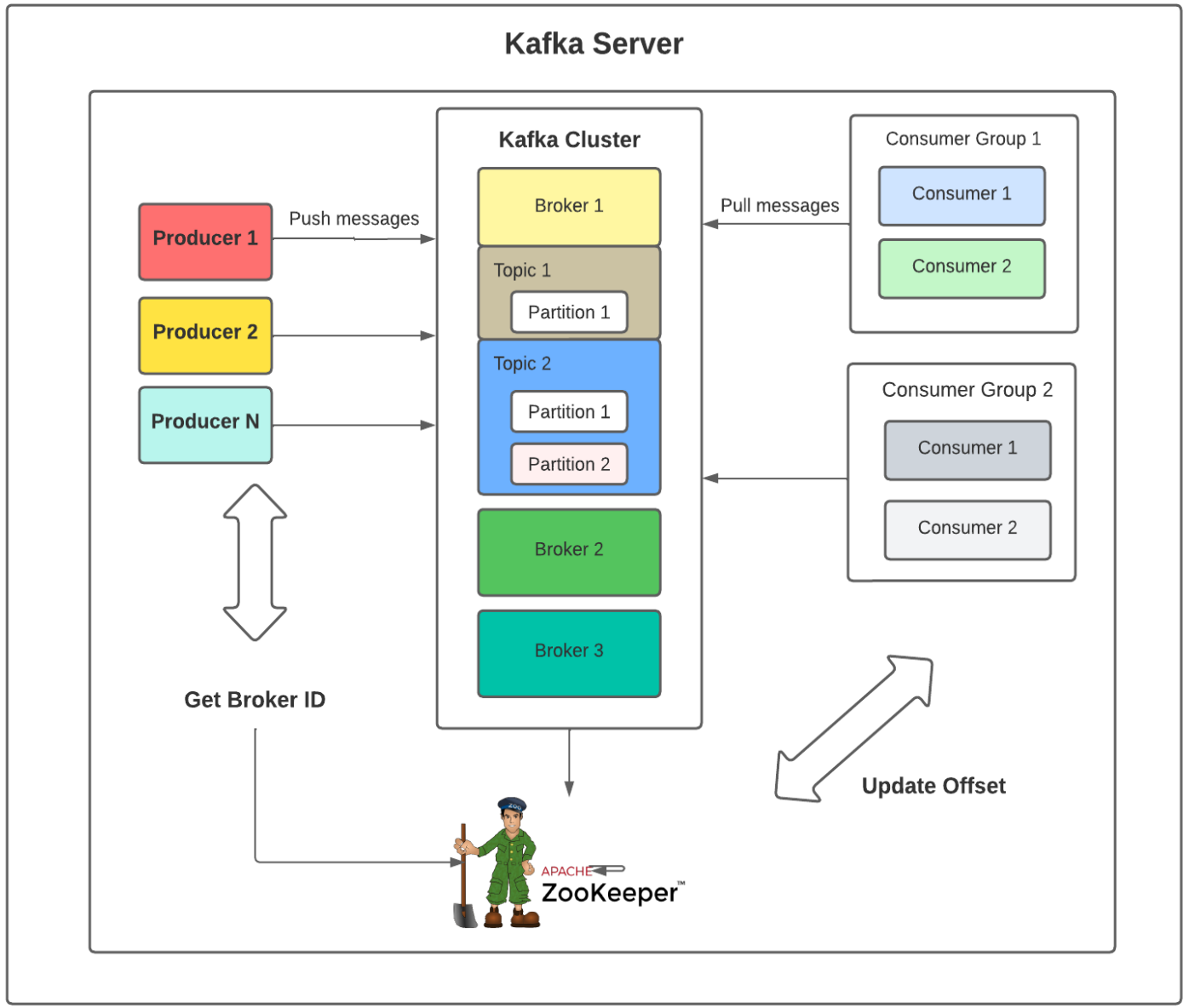
Project Architecture

- **The producer** is getting the data from **Yahoo Finance** , transforming the message topic contents to a json document and pushing to Kafka cluster every second.
- The consumer is responsible to collect messages every second
- The consumer allows the python application on local Server for gathering data from kafka cluster and loads the data into MongoDB
- Python Power BI API and Analytics get the data from MongoDB and pushes that to Power BI stream



Kafka Architecture

- Kafka Server provides a platform for distribution real-time streams among integrated systems of networks
- Streams could be messages or data, such as payments transactions, geolocation updates, and some else
- These consist of events organized in Topics
- Producer client connects with the Kafka brokers
- Kakfa Cluster (Brokers, Topics and Partitions)
- Consumer Groups reads the messages produced by brokers)
- ZooKeeper provides "Broker ID" and manages "Offset"



Real-time Streaming

- A distributed event streaming platform
- The producer pushes new record (message) each second to Kafka Cluster
- ZooKeeper service provides “Broker ID” to Producers push the data to Kafka Cluster on a Topic Server
- ZooKeeper service manages “Offset” to identify the position regarding each consumer in a partition
- A Broker contains a Topic storing data partitions from previous configuration
- The consumer collects the data (message) each second from a topic assigned

Dataset

- Stock price data from yahoo finance
- Ticker : AAPL (Apple INC)
- Last 60 days stock price data
- Variables : Datetime, Open, Close, Adjusted Close, Low, High and Volume

Lambton College

Python Producer

- The application was developed based on Python
- Using Yahoo Finance API to push the data (message) for every second to Kafka Cluster
- It also gets stock price data of Yahoo Finance publicly market data

Lambton College

Python Consumer

- The consumer was developed based on Python
- Using Kafka consumer library to pull the data (message) for every second from Kafka Cluster and pushing it to MongoDB
- It also gather and transform stock price data of publicly traded companies to a json document

Demonstration of The Pipeline

Lambton College

Analytics

- Connect using Pymongo to the database
- Query and insights
- Statistical features
- Visualizations in the notebook

Visualization

- Dashboards generated using PowerBI Web
- Connection between MongoDB and Power BI using API with Python script
- Python gets the data from MongoDB and pushes the data to Power BI
- The schema has to be pre specified in Power BI
- Refreshed manually

Lambton College

References

- Apache Kafka. https://kafka.apache.org/
- Yahoo!Finance. http://finance.yahoo.com

Lambton College

Thanks For Your Attention

Lambton College