Burroughs

BRIDGING THE GAP BETWEEN KAFKA AND SQL

Ainesh Sootha, Andrew Riordan, Aneesh Chakravarthula, Erika Ergart, Jason Cao,
Mihira Krishnaswamy, Vandana Chari, Wyatt Klueber

PURDUE UNIVERSITY

Background

What is Apache Kafka?

Kafka is a popular platform for data transmission from many collection sources to many receiving consumers, but it has some problems:

- Data analysis must be done downstream from Kafka
- Need a consumption hook, a special program that moves the data off of Kafka
- Must wait a long time for this to happen

Recently, KSQLdb is a tool that has emerged to manipulate Kafka "streams" (real time collections of data) in similar ways to how one would treat data in a RDBMS.

However, it has some limitations:

- Specialized syntax
- Outputs query results onto another Kafka stream
- No closer to integrating with other tools

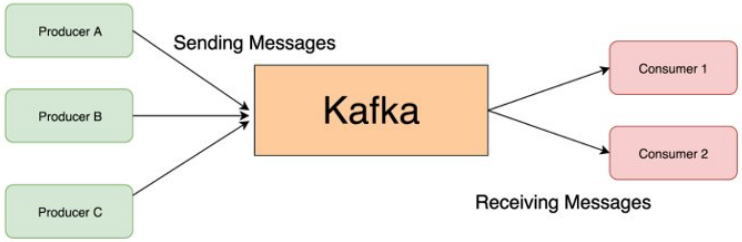
Project Statement / Burroughs Goals

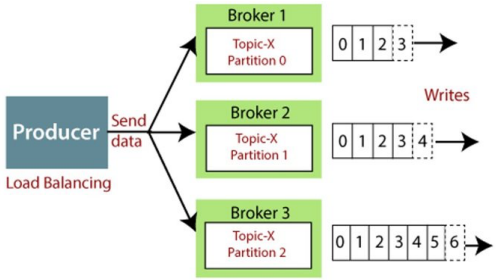
Performing data analysis on relational database is much easier than on the Kafka stream.

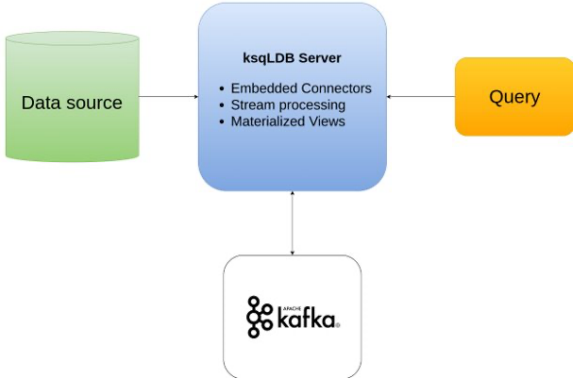
The goal of Burroughs is to bring real-time data from Kafka into the wheelhouse of the analyst/data scientist by leveraging KSQL.

Why Burroughs?

- Replace Consumption Hooks
  - Time consuming
  - Esoteric
  - Potentially redundant (blunt object)
- Mirroring into a database anticipates future use cases
- Leverages existing KsqlDB infrastructure







Research Methodology

Getting to know Kafka / Constructing a Realtime Data Streaming Pipeline

In anticipation of working with complex Confluent Platform, we first learned how to work with low-level Kafka technology to enable us to work on high-level Burroughs project; this step allowed us to accomplish cloning a KSQL table into RDBMS.

- Created a Kafka Producer
- Created a Kafka Consumer
- Created Streams using ksqlDB
- Ran simple queries, built tables, ran table aggregations
  - Identified limitations of ksqlDB

Burroughs & Query Translation

Keeping the limitations of ksqlDB in mind, we are continuing to work on the query translation aspect of Burroughs.

- Check for a valid SQL query
- Check for keywords that require translation
- Define queries that Burroughs can / cannot support
- Limitations of ksqlDB:
  - Not all SQL functions are supported
  - Cascading Aggregations
  - Distinct Keyword doesn't exist