GHC 2019 Session Notes

Session: DS520: Real-time Analytics Platform Based on Lambda Architecture
Speaker: Yang Zhou, Software Engineer, *Intuit*
Note Taker: Sarvenaz Myslicki

Introduction of Quickbooks Example

- The speaker's story starts with quickbooks online, where analysts noticed subscription numbers were down. The discovery timeline to fix the issue was around 9 days. Two main problems:
    a. They didn't have end-to-end technical monitoring, and everyone's individual dashboards showed things were ok. The problem was occuring between teams.
    b. Analytics were not available in real time.
- Historically, transactional and analytics systems have been kept separate. That way analytics won't impact transactions. However, this means you can't make quick data driven decisions.

Lambda Architecture

- Lambda architecture is an excellent bridge when moving from an existing batch-processing ecosystem toward real-time. All that is needed to build is the streaming layer.
- Lambda is used to denote this architecture because of the split of data between the batch layer and speed layer.
- The batch layer is an immutable master dataset. Map reduce is used to compute batch views. Each iteration could take hours.
- The speed layer makes up for the multi-hour batch processing needs. It creates increment views.
- The final layer is the serving layer, which merges batch and real time views and allows users to create ad hoc queries.

Case Study: Intuit's Real Time Platform

- Intuit's real-time platform uses Apache Cassandra (distributed NoSQL data store), Apache Kafka (message queue for real time data), Apache Spark Streaming (consumes events from Kafka)
- When a user come to quickbooks and clicks around, JSON events go to Kafka, Spark Streaming, and then Cassandra.
- Internal teams can subscribe to a specific events that interest them.
- Once in Cassandra, there are dashboards on wavefront (automate alerting and monitoring) and tableau (slicing and dicing) as the serving layer.
- Now the subscription team can monitor with the new real time system. A threshold alert is set to go out to them within 5 seconds of an anomaly!
- Deal with roughly 4.5 million customers and 10-15 million events per day.