PRESTO

Distributing queries across different data stores

What is Presto

- It's a lot like Drill
  - *It can connect to many different "big data" databases and data stores at once, and query across them*
  - *Familiar SQL syntax*
  - *Optimized for OLAP – analytical queries, data warehousing*
- Developed, and still partially maintained by Facebook
- Exposes JDBC, Command-Line, and Tableau interfaces

Why Presto

- Vs. Drill? Well, it has a Cassandra connector for one thing.
- If it's good enough for Facebook…
  - "Facebook uses Presto for interactive queries against several internal data stores, including their 300PB data warehouse. Over 1,000 Facebook employees use Presto daily to run more than 30,000 queries that in total scan over a petabyte each per day."
  - Also used by DropBox and AirBNB
- "A single Presto query can combine data from multiple sources, allowing for analytics across your entire organization."
- "Presto breaks the false choice between having fast analytics using an expensive commercial solution or using a slow "free" solution that requires excessive hardware."

What can Presto connect to?

- Cassandra (It's Facebook, after all)
- Hive
- MongoDB
- MySQL
- Local files
- And stuff we haven't talked about just yet:
  - *Kafka, JMX, PostgreSQL, Redis, Accumulo*

Let’s just dive in

- Set up Presto
- Query our Hive ratings table using Presto
- Spin Cassandra back up, and query our users table in Cassandra with Presto
- Execute a query that joins users in Cassandra with ratings in Hive!