Movie Recommendations to a website

Working backwards

* Users want to discover movies they haven’t yet seen that they might enjoy
* Their own behaviour (ratings, purchases, views) are probably the best predictors
* As before, availability and partition-tolerance are important, Consistency not so much.

So maybe Cassandra

* But any NoSQL approach would do these days

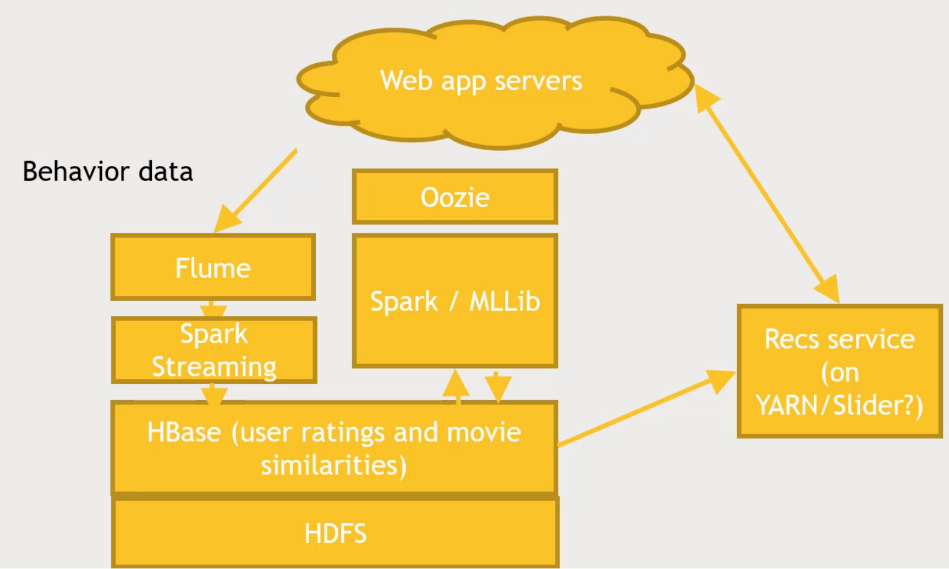
How do movie recommendations get into Cassandra?

* We need to do machine learning
  + Spark MLLib
  + Flink could also be an alternative
* Timeliness requirements need to be thought out
  + Real-time ML is a tall order - do you really need recommendations based on the ratings you just left?

Creative Thinking

* Pre-computing recommendations for every user isn’t timely and waste resources
* item-base collaborative filtering
  + Store movies similar to the other movies
  + At runtime, recommend movies similar to ones you’ve liked (based on real-time behavior data)
* So we need something that can quickly look up movies similar to ones you’ve liked at scale
  + Could reside within web app, but probably want your own service for this
* We also need to quickly get at your past ratings/views/etc.

Requirements

* Some web service to create recommendation on damand
* It will talk to a Fast NoSQL data store with movie similarities data
* it also need your past ratings/purchases/ etc.
* Movie Similarities (which are expensive) can be updated infrequently, based on log data with views/ratings/etc.
* Recs Service custom service, Docker based system? (on YARN/Slider)

Exercise: Design Web Analytics

* Work for Big website
* Some manager wants a graph of total number of sessions per day
* And for some reason they don’t want to use an existing service.

Requirements.

* Run daily based on previous day’s activity
* Sessions are defined as traffic same IP address within a sliding one hour window
  + Hint: Spark Streaming etc. can handle ‘stateful’ data like this
* Let assume your existing web logs do not have session data in them
* Data is only used for analytic purposes, internally

Consider

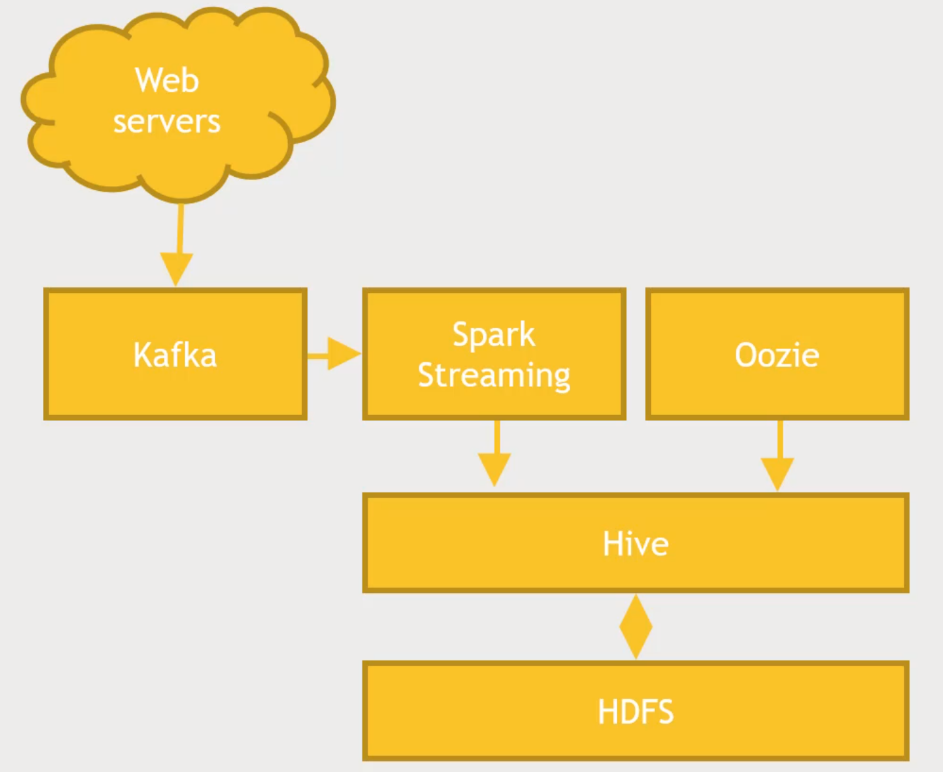
* A daily SQL query run automatically is all you need
* But this query needs some table that contains session data
  + need to be built up throughout the day

My answer?

Web server logs is to be logged 1 hour, check the IP addresses for duplicates.

Storage and Processing: MongoDB or HBase. Apache Drill to access the logs?

Flume or Spark Streaming for getting the data

Solution: