Recommendation at scale

Simon Dollé

Berlin Buzzwords, June 2nd, 2015

Where is the need for tech ?

We buy

- Inventory ! (ad spaces)
- Billions of times a day
- All over the Internet
- For 95% of the population

We sell

- Clicks !
  - (that convert)
    - (that convert a lot)

We take the risk

You pay only for what you get

Traffic

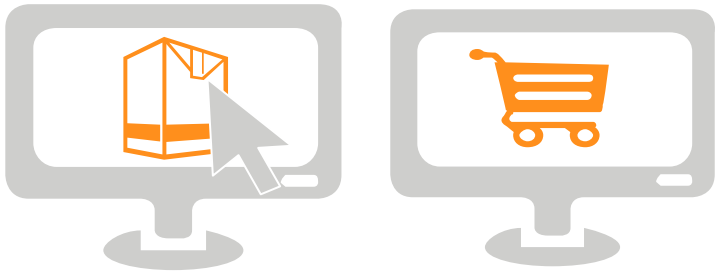
- 800 k HTTP requests / sec (peak activity)
- 29000 impressions /sec (peak activity)
- Less than 10 ms to process an RTB request

Physical infrastructure

- 6 Data centers on 3 continents operated and conceived in-house
- ~ 12000 servers, largest Hadoop cluster in Europe
- More than 35 PB of storage Big Data

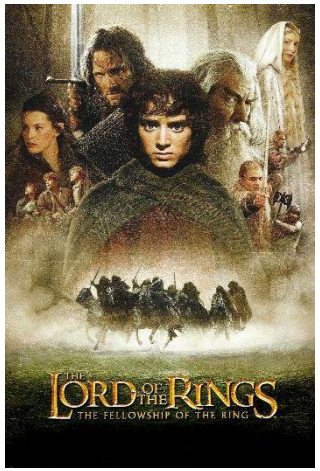
Data Sources

Catalog data

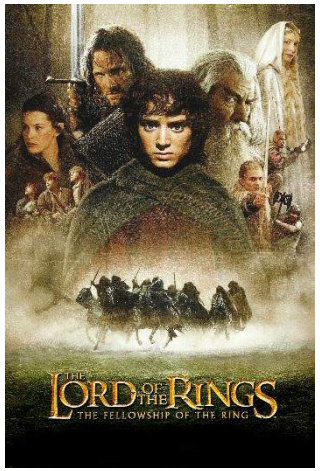


User behavior data

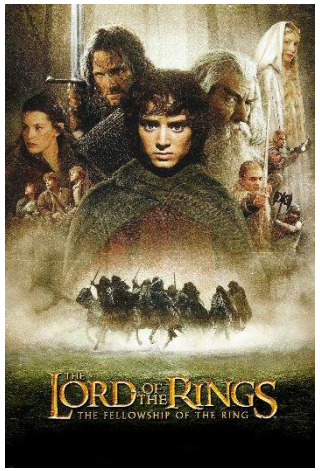
Ad display data



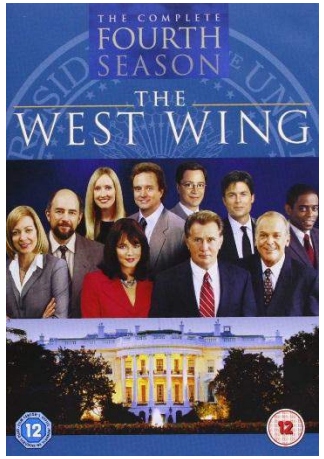
criteo

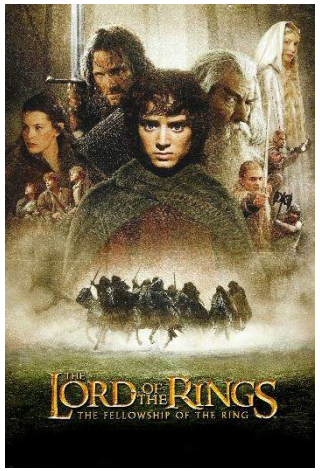


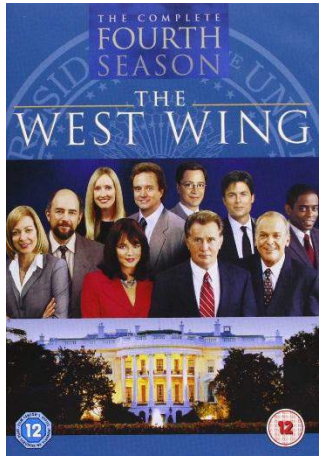
criteo

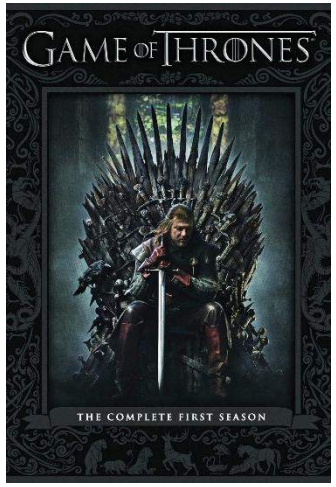


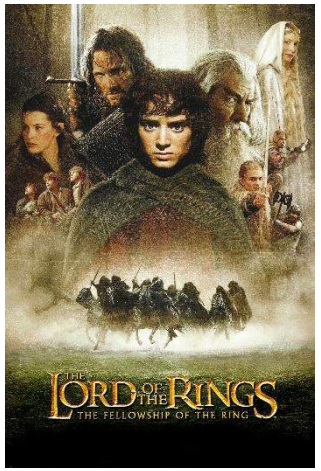
criteo

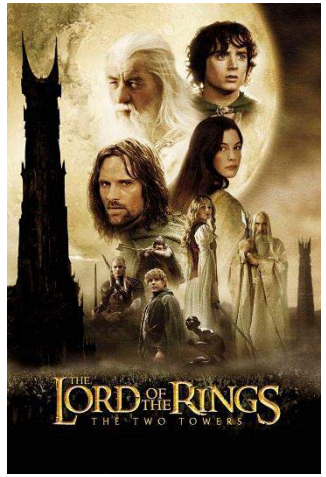


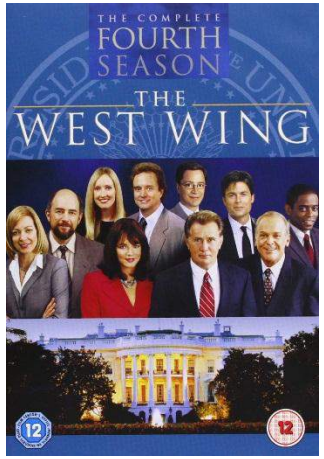


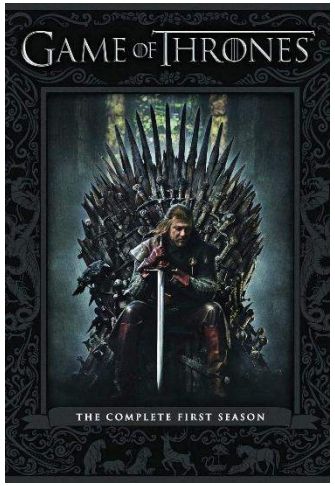


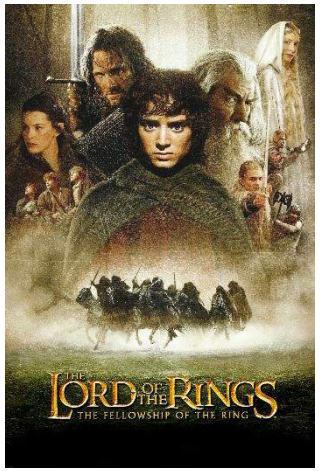


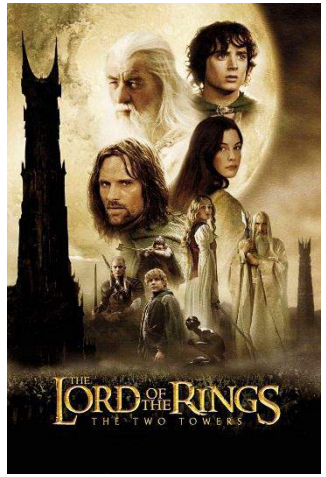


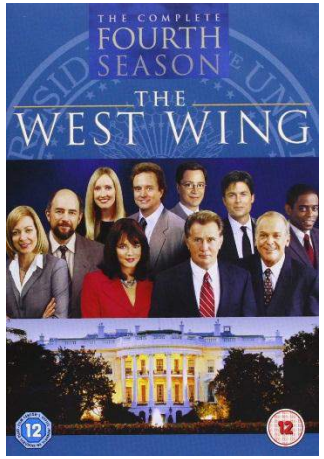


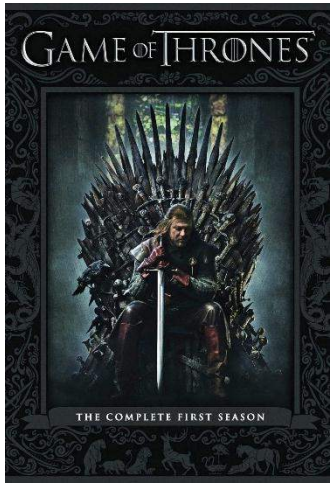










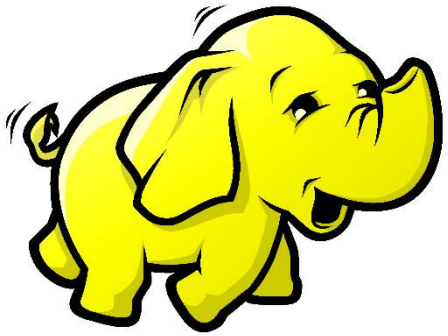


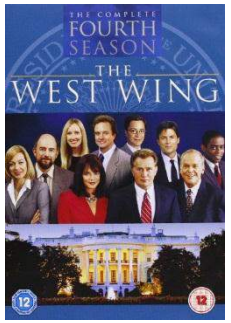
Offline

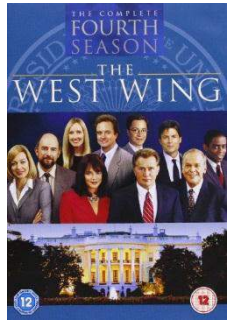
- Similarities computed on browsing data
- Based on coevents

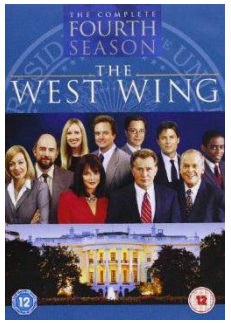
- Computed on Hadoop cluster
- Map reduce jobs, pig
- Takes around 12 hours

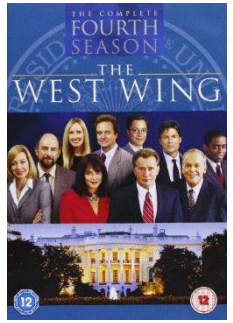
- Tradeoff performance, speed



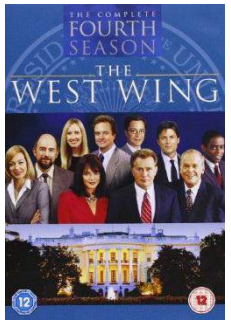








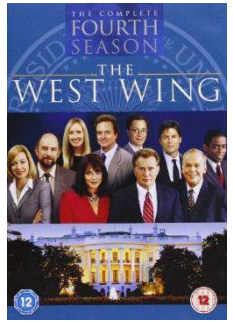
0.04 0.02 0.02 0.05



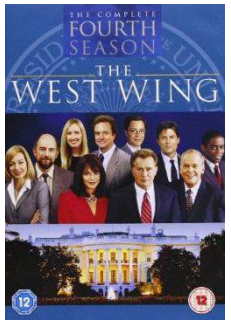
0.05

0.04

0.02



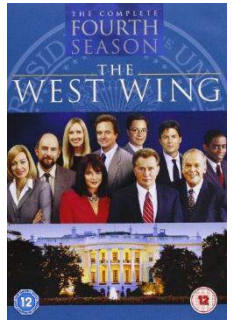
0.02



0.05

0.04

0.02



0.02

ML model

- Logistic regression models
- Features

Product-specific

User-specific

User-product interactions

Display-specific

Online optimizations

- Algorithmic
  - Use simpler ML model
  - Quickly discard candidates
- Technical
  - Memcache + local cache
  - Async I/O

criteo

Upcoming challenges

- Long(er)-term user profiles
- More and better product information (images, semantic, NLP)
- Instant-update of similarities
  - (because batch computation is soooo last year)

Fancy a try ?

On your own:

- Our 1st public dataset is online: http://bit.ly/1vgw2XC
  - 4GB display and click data, Kaggle challenge in 2014
- NEW : 1TB dataset released a few weeks ago: http://bit.ly/1PyH4Vq
  - Hosted on Microsoft Azure, just waiting for you

With us !

http://labs.criteo.com/jobs/

criteo

Questions?

criteo

Thank you !

s.dolle@criteo.com