Big Data na Globo.com

Augusto Boranga e Matheus Pereira

6 times

- Analytics
- Cluster
- Pipeline
- Semântica
- Recomendação
- Growth

Alguns números

- 3 bilhões de eventos diários
- 2 milhões de conexões simultâneas
- 50 milhões de usuários únicos por mês
- 100 mil novos conteúdos por mês
- 100 mil recomendações por minuto
- +20 algoritmos de recomendação implementados
- +400 testes A/B

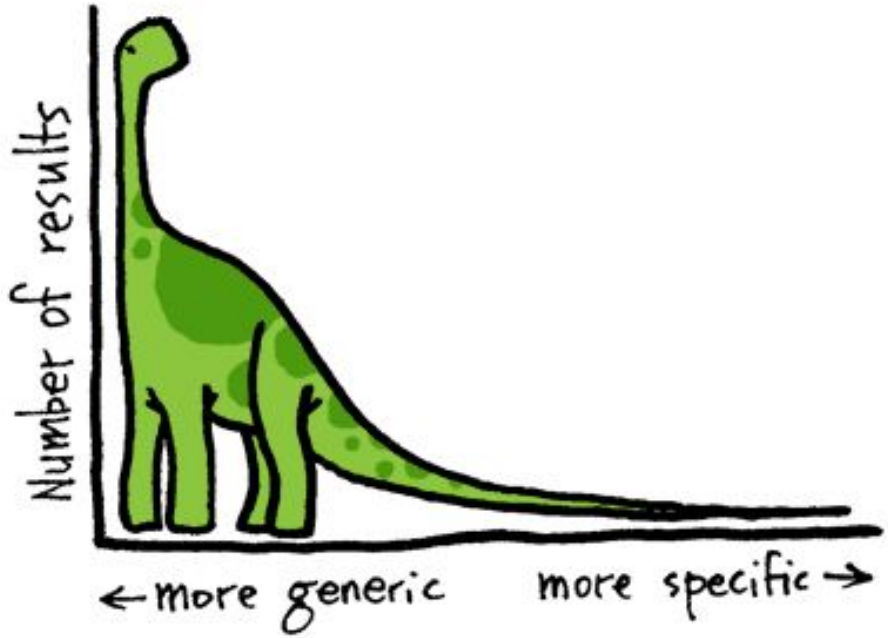
Sistemas de Recomendação

Para que servem:

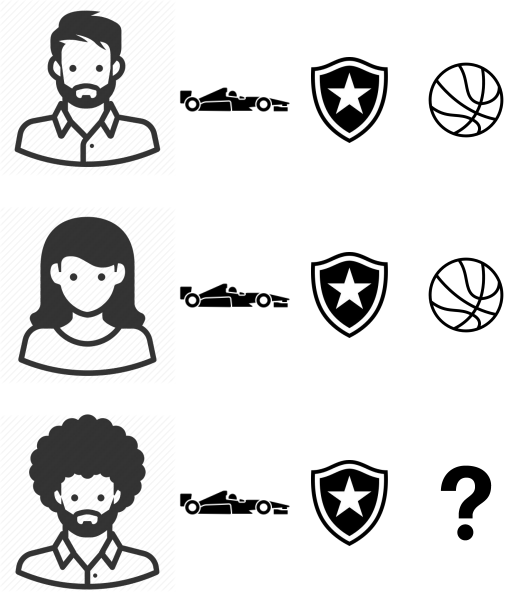
- Mostrar documentos que possam interessar o usuário, mas que ele não veria normalmente

Objetivo:

- Mais engajamento do usuário
- Ex: cliques, tempo online, vendas



COLLABORATIVE FILTERING



CONTENT BASED

TOP

1. Ministro do STF determina quebra de sigilo bancário e fiscal de Aécio Neves
2. Dólar sobe quase 2% e vai a R$ 3,28 com temor sobre a Previdência
3. Juiz pede prisão de Cristina Kirchner; Senado precisa dar aval

Collaborative Filtering

- O que usuários parecidos comigo viram que eu ainda não vi
- Maior taxa de conversão
- Atualização lenta
- É preciso conhecer o usuário
- Famoso pelo uso na Amazon

Content Based

- Documentos **similares** ao visualizado, com as mesmas palavras-chave
- Rápida atualização
- Usado no Spotify para a criação de playlists automáticas

- Abordagens usadas globo.com
  - TF-IDF: Automático para extração de palavras-chave
  - Semântica: Os editores indicam no momento da criação sobre quem é, onde ocorreu o fato etc

Top

- Conteúdos mais visitados
- Usado quando não sabemos nada sobre o usuário (aba anônima)
- Pode ser inteligente (top por região do usuário)

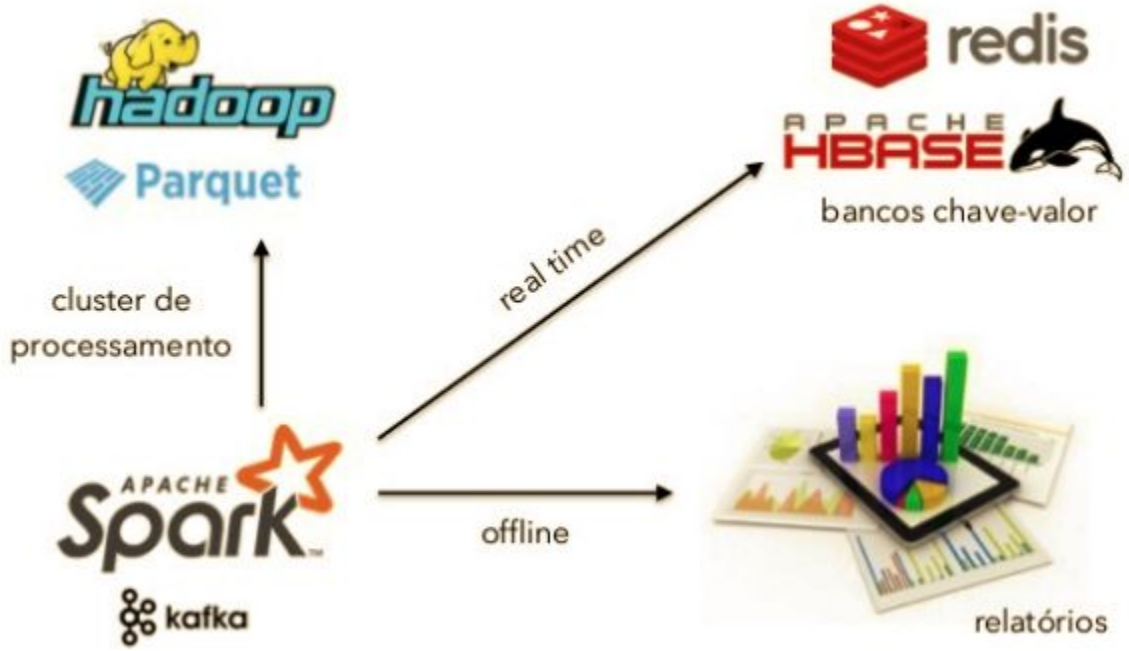
Testes A/B


50 % visitors
see variation A
Variation A
23%
conversion
50 % visitors
see variation B
Variation B
11%
conversion

Testes A/B

- Método de avaliação de desempenho entre grupos
- Usuário não tem conhecimento do teste
- Método científico

Pipeline


hadoop
Parquet
redis
APACHE HBASE
bancos chave-valor
cluster de processamento
real time
APACHE Spark
kafka
offline
relatórios

Data Science


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS