How Machine Learning, Google Cloud and Hadoop make a difference in mining insights from Social data

Presenter: Mr Minh Truong
Tech Lead – R&D Department
YouNet Group

Date: May 28, 2019

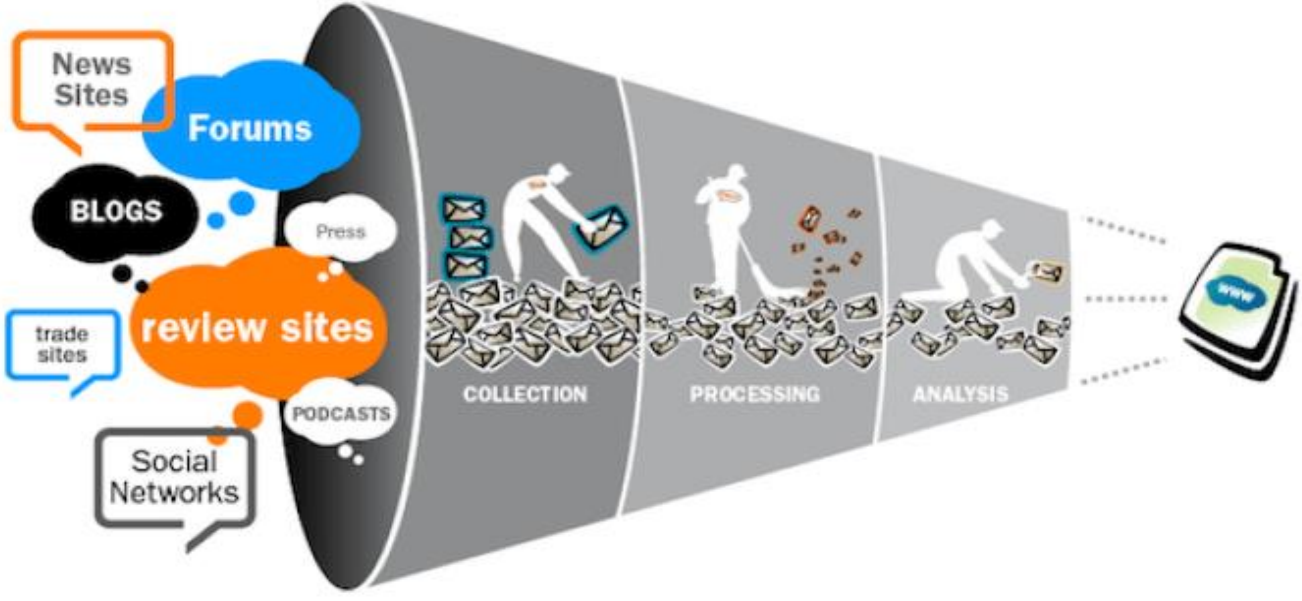
AGENDA

- Our Applications & Challenges
- Storing, retrieving and analyzing billion of records on Hadoop
- Filtering noises and extract advanced insights using Machine Learning
- Scaling on demand with Google Cloud
- The future

Our Applications & Challenges

SocialHeat

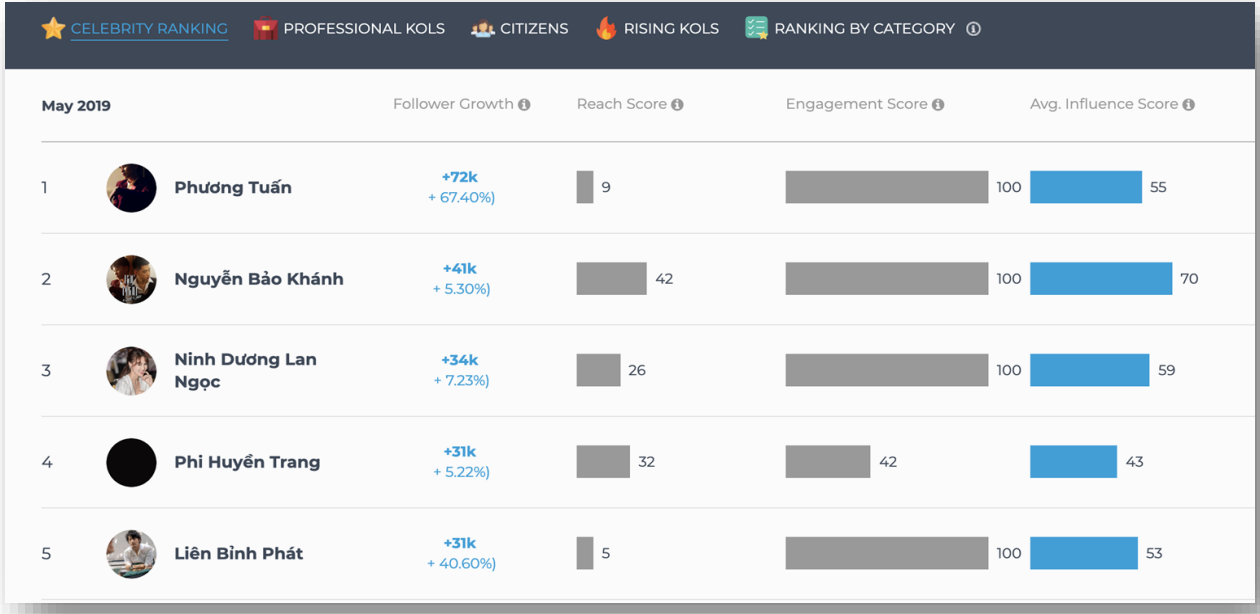
No 1 Social Listening Platform in Vietnam



SociaLift

Largest Influencer Database in Vietnam

socialift



Challenge 1

Storing, retrieving and analyzing billions of records every month

~ 30,000,000 / day

~ 1,000,000,000 / month

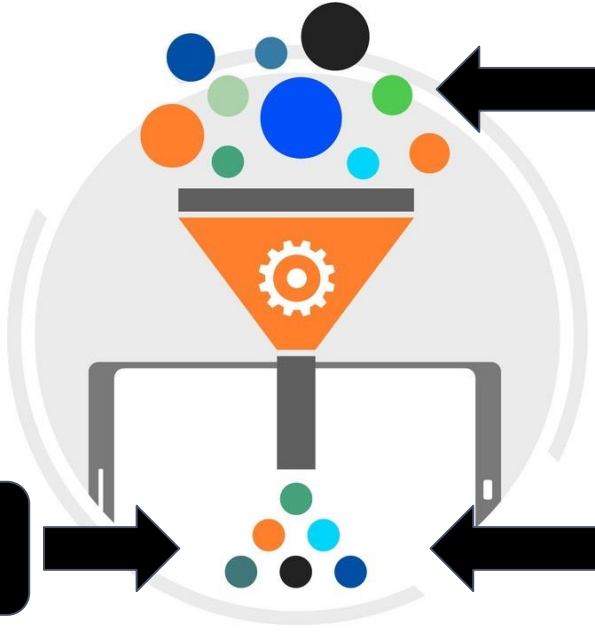
~ 12,000,000,000 / year

~ 100,000,000,000 / 10 years

Big Data!!

Challenge 2

Filtering noises and extract advanced insights



80% is noise

Category Classification

Sentiment Analysis

Challenge 3

Scaling on demand

10X traffic for just few hours

CHALLENGE 1

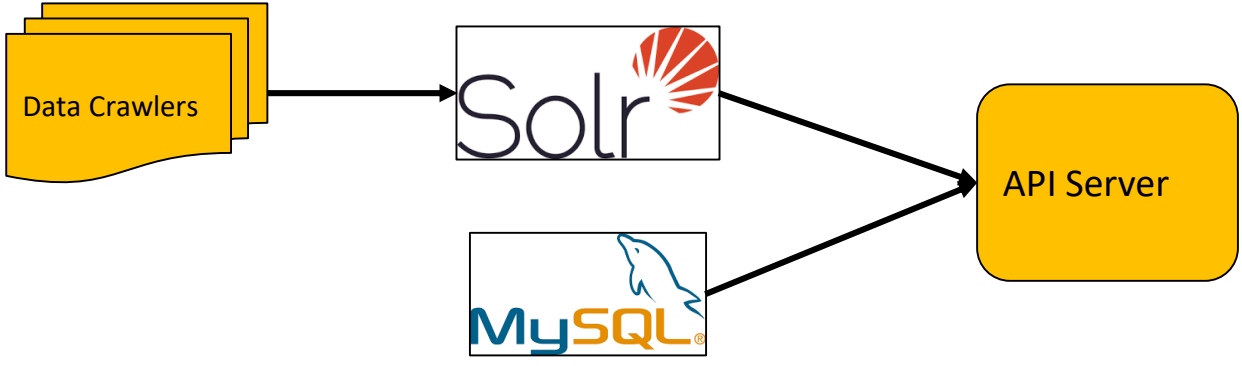
Storing, retrieving and analyzing billions

of records every month

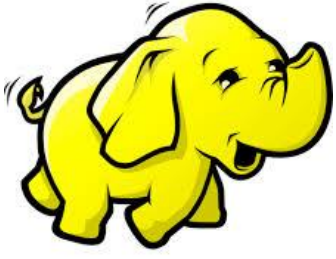
Store, index & search

- Open source
- Highly scalable (NoSQL)
- Very fast text search
- Fault tolerant
- Strong community

Original Architecture

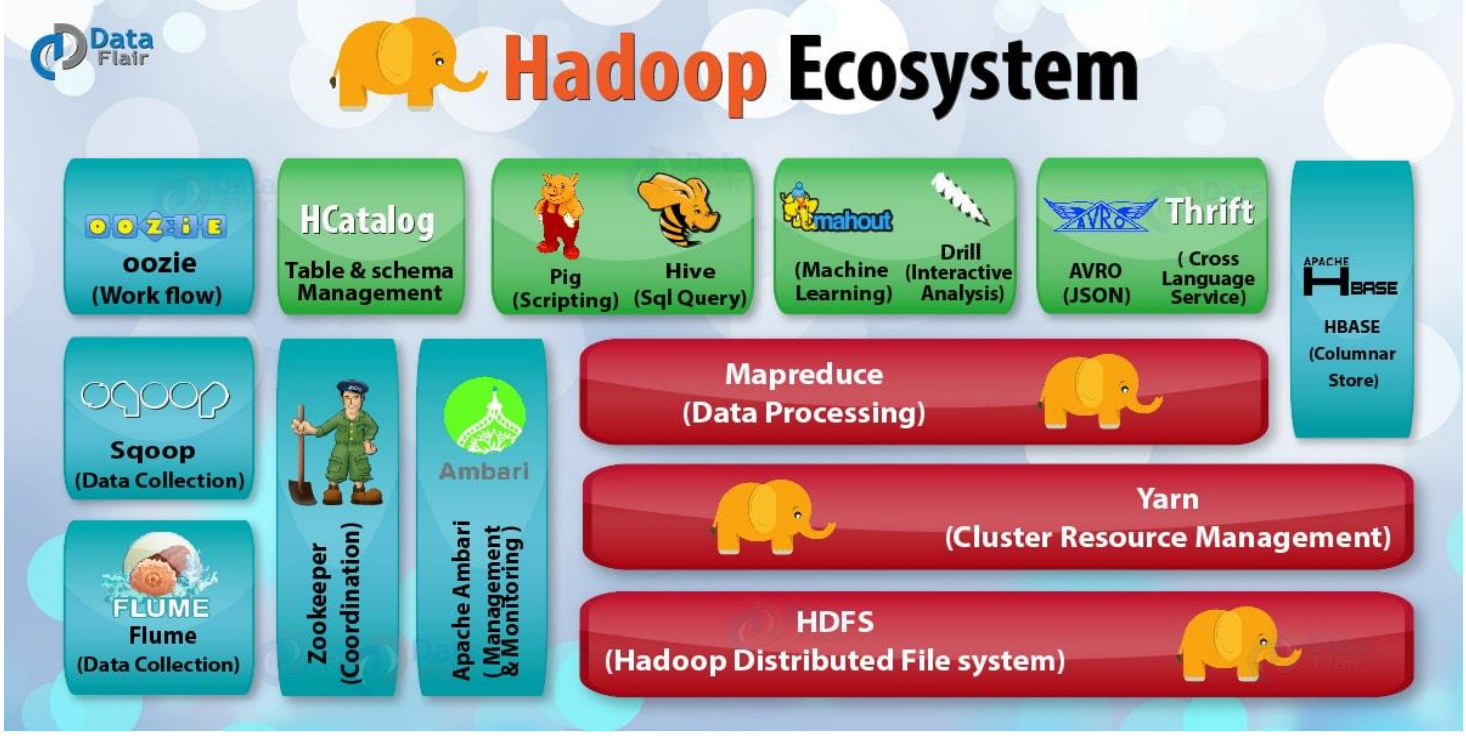


Advance Analytics

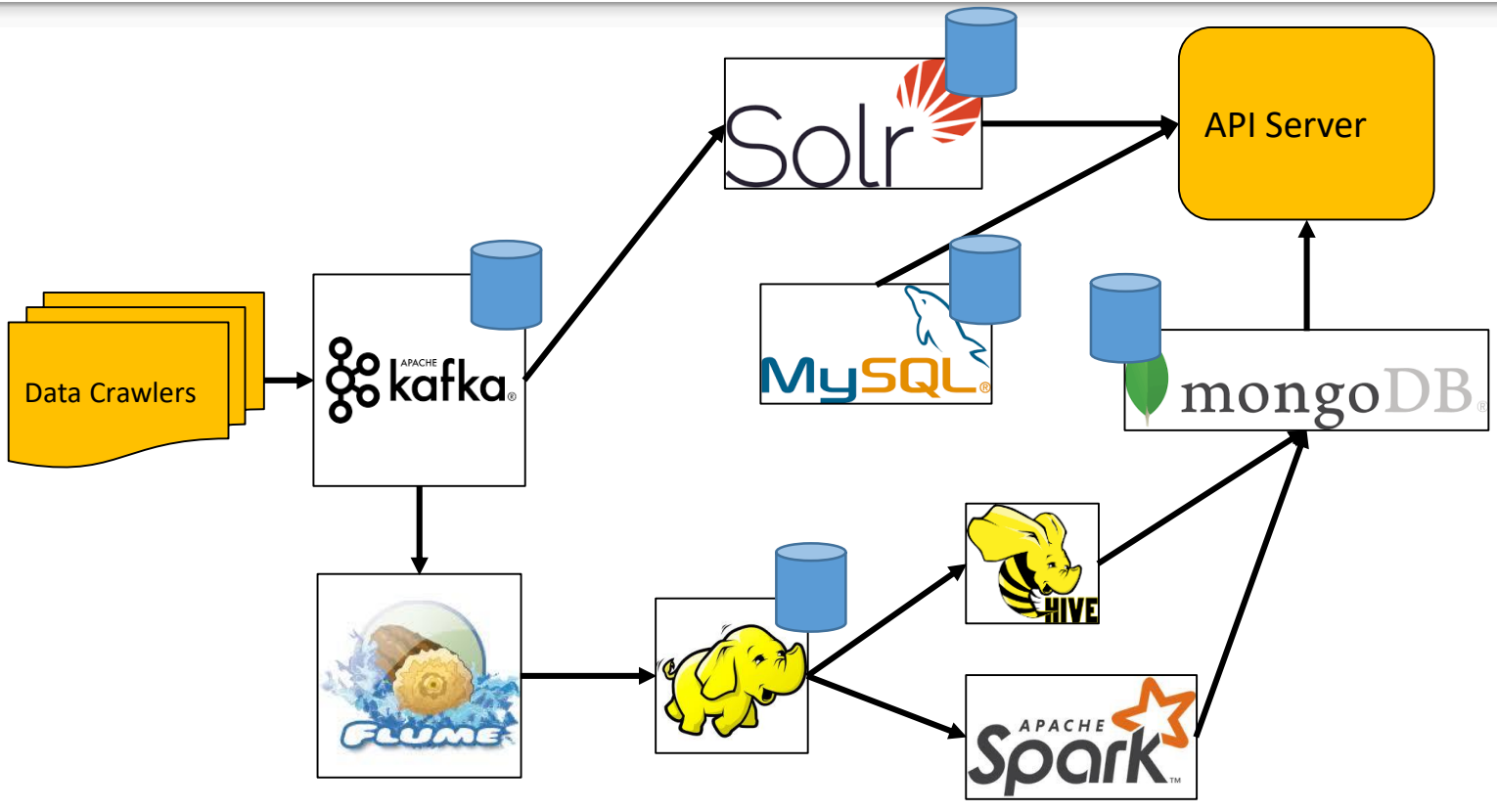


- What is the age range of people who liked posts on YouNet fan page about Social Listening from the West of Vietnam?
- How many people who liked YouNet Recruitment fan page, at age from 18-24 and from top VN universities ?

Brief Introduction to Hadoop



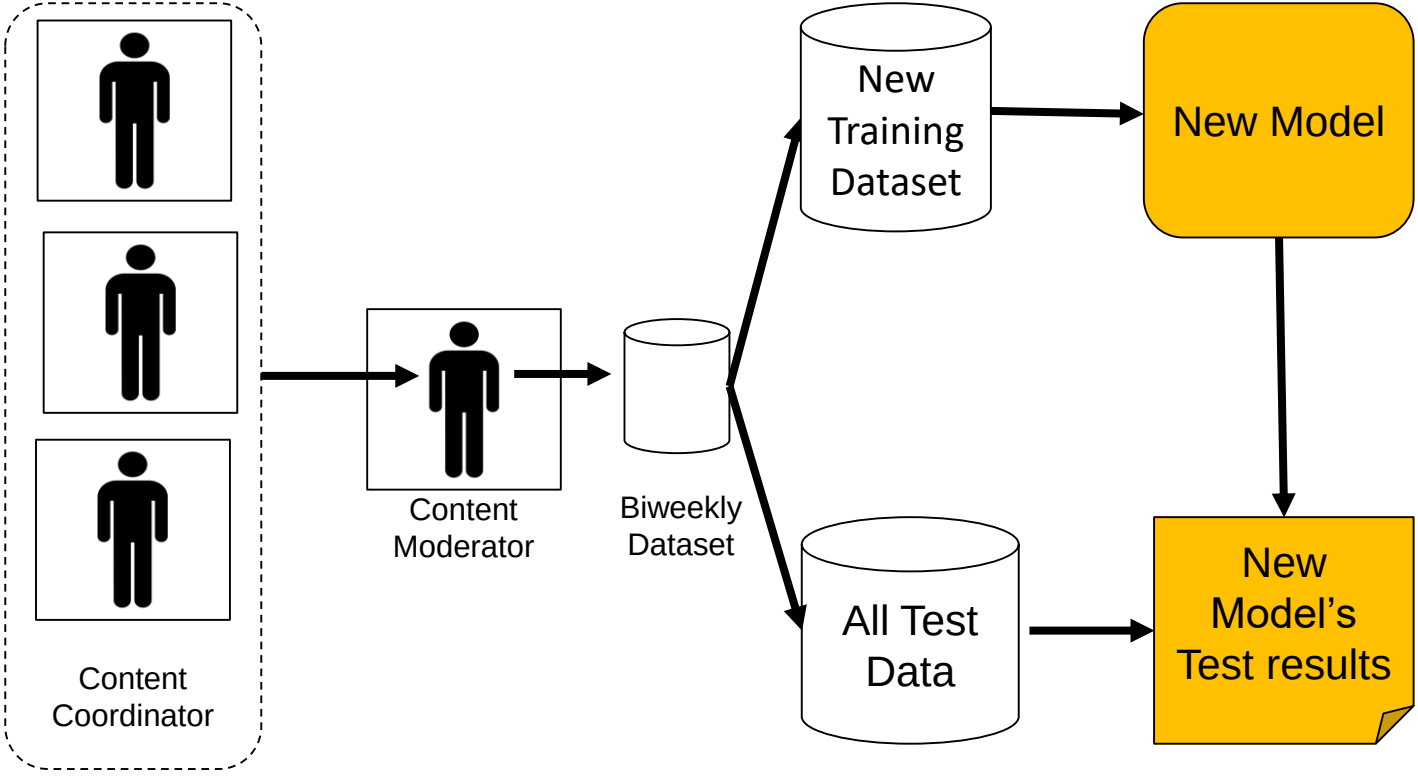
New Architecture with Hadoop



Filtering noises and

extract advanced insights with ML

Training Data Quality Management

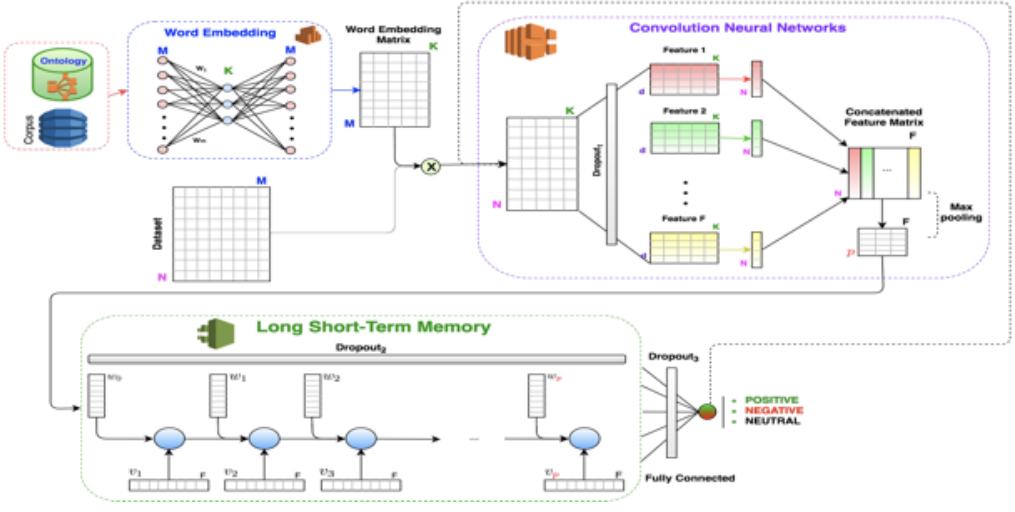


Spam Detection

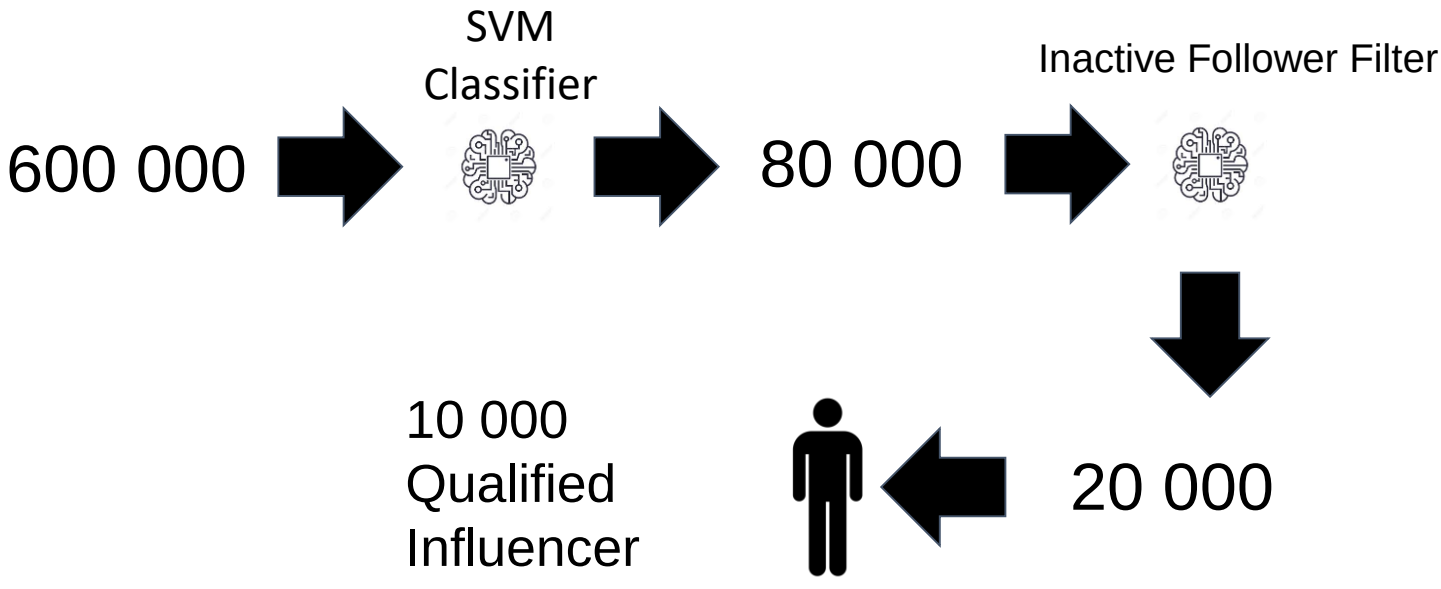
https://github.com/facebookresearch/fastText

Sentiment analysis

State of the art deep learning model using Tensorflow, CNN, RNN and word embedding with current F1-score: 0.81 on live data



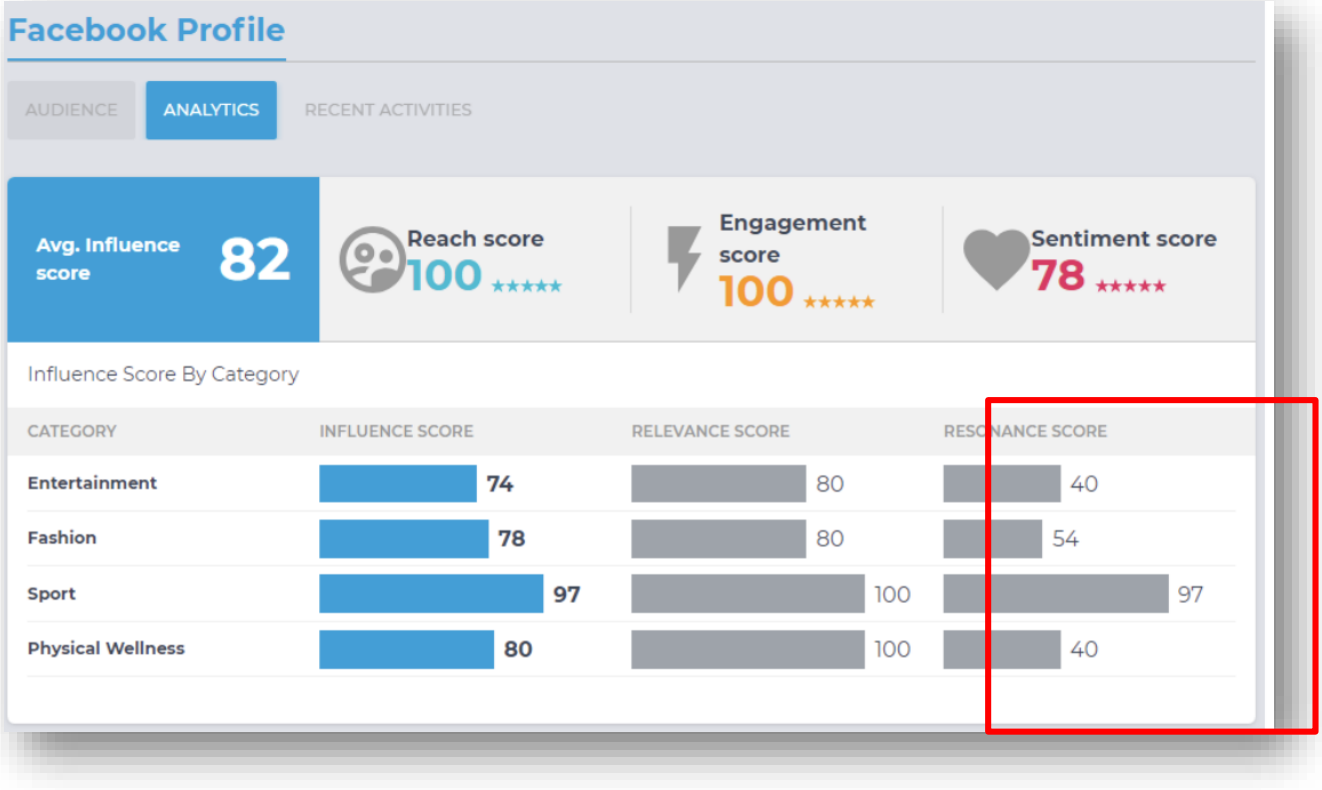
SociaLift's Influencer Filter



Influencer Score

Resonance Score

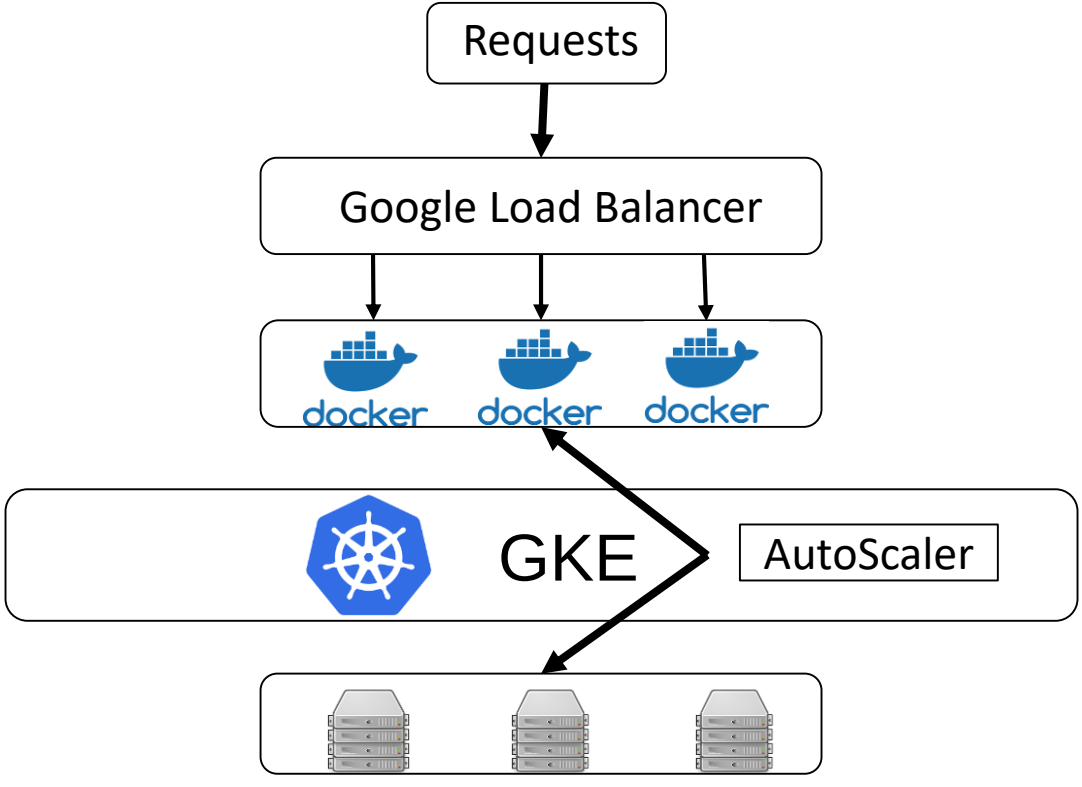
(Are comments relevant to post content?)



Scaling on demand with Google Cloud

Scale from 1 to 10X traffic

Docker & Kubernetes on GKE with preemptible instance



Lessons learned

- Starts with simple solution and evolves to more complex one as needed
- There is no single database that can fit all
- Data quality is the most important thing in ML
- Try out with hybrid model to utilize the flexibility and elasticity of public cloud

THE FUTURE

- Apply computer vision at scale to analyze images and videos
- Chatbot & automated customer care
- Classify violent, policy-violated content in real-time
- Creative assistant for marketing content

Thank you!

YouNet
social intelligence

www.younetgroup.com

/younetvietnam

(+84) 8 626 464 88

info@younetgroup.com

US OFFICE
9741 Bolsa Avenue, Suite 201, Westminster, CA 92683

VIETNAM OFFICE
2nd Floor, Lu Gia Plaza, 70 Lu Gia St., Dist. 11, HCMC