Social media data analysis & real world cases

Nikos Tsirakis

9/12/2017

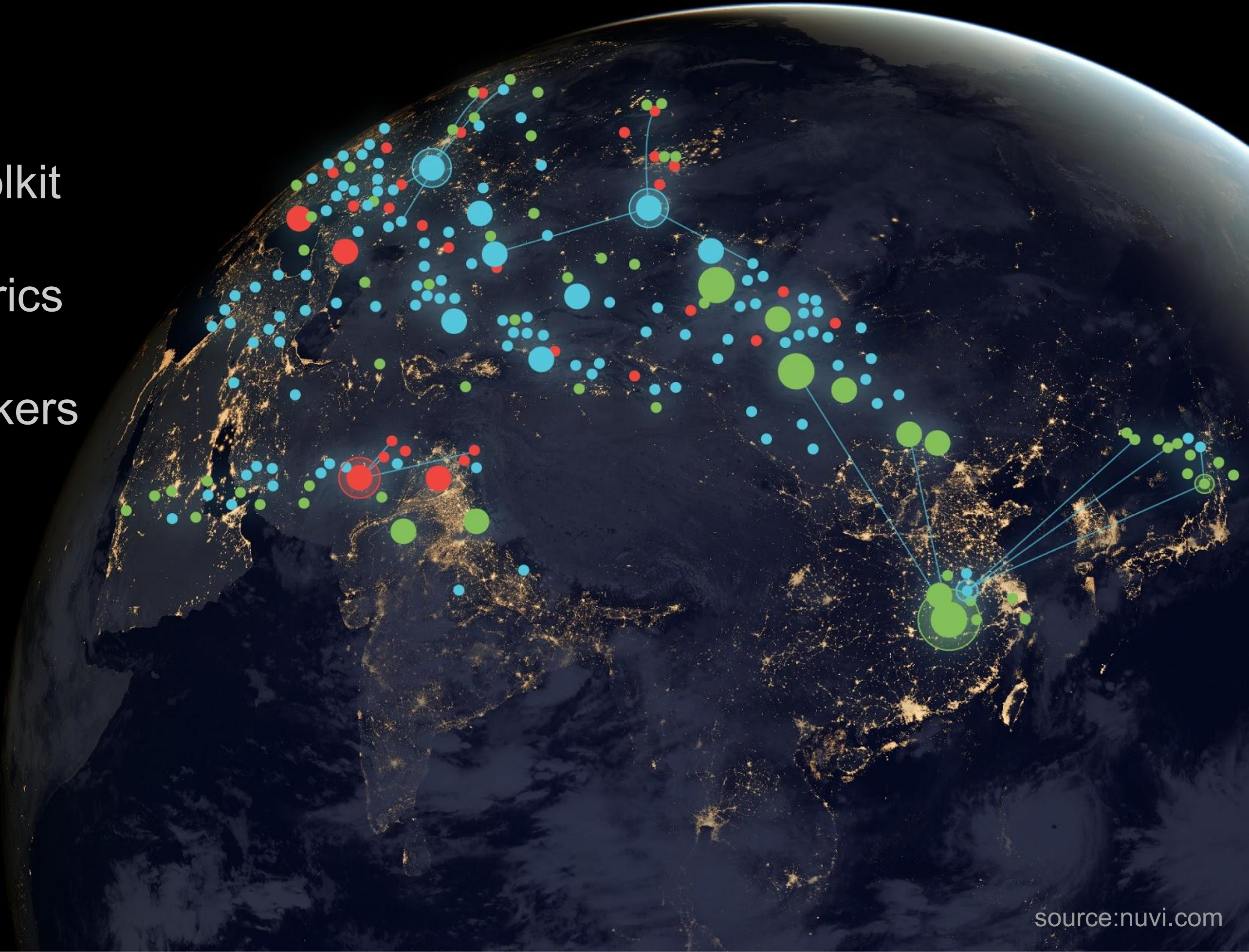
Agenda

1. Introduction to News & Social Media, Reputation Management Platforms
2. Data, Data, …. Big Data
3. Analyse, Transform, Enhance, Present
4. Challenges
5. Technologies
6. Case Study
7. Q & A

Introduction

Brandwatch
Mention
Sysomos
Nuvi
Talkwalker
Agorapulse
Lithium
Synthesio
Trackur
Keyhole
Buzzlogix
Mentionlytics

Mediatoolkit
Digimind
Ubermetrics
Brand24
Socialbakers
Sprinklr
...



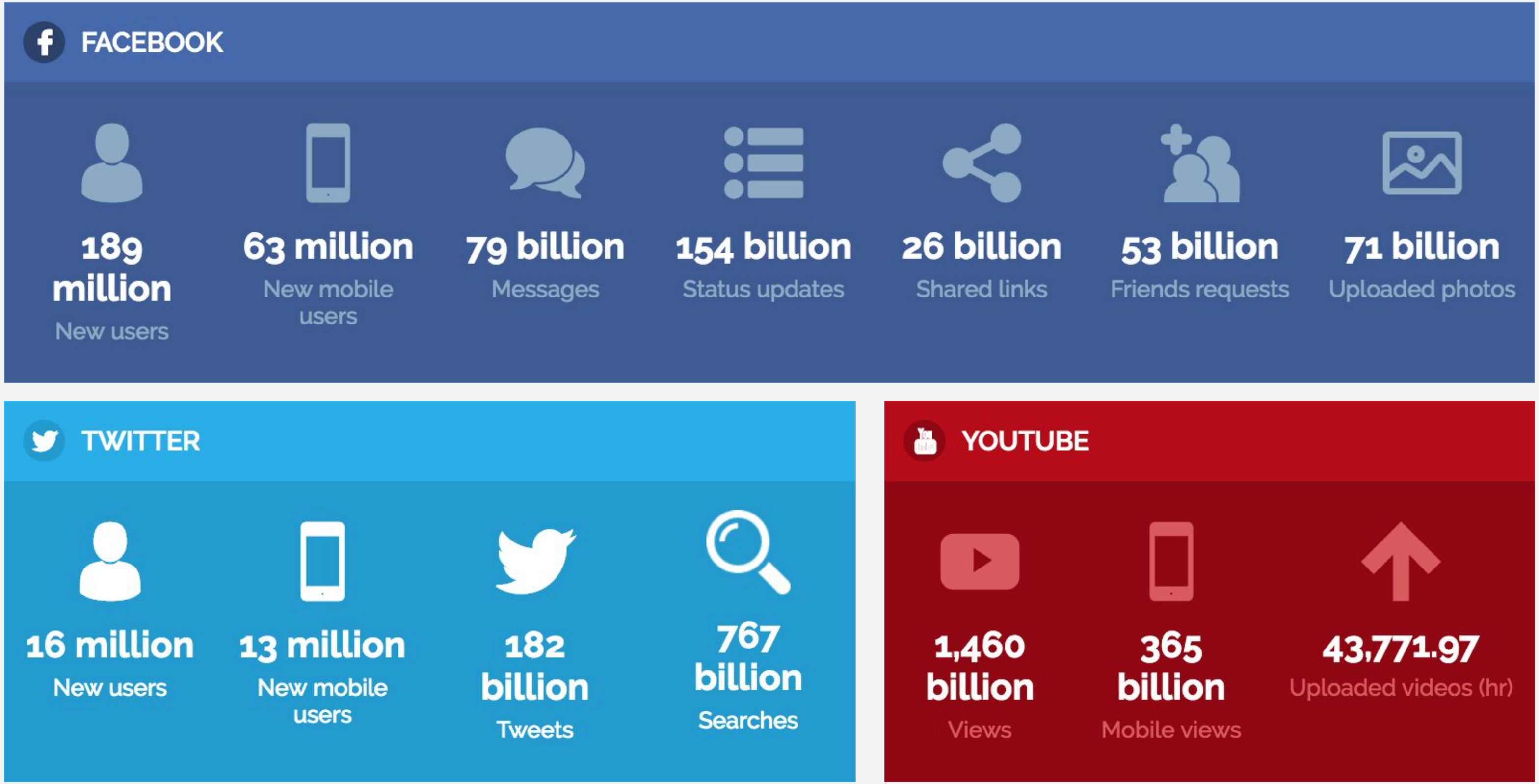
source:nuvi.com

Data & Precision in Numbers

- ~20k news sites & blogs
- +20 popular forums
- ~700k twitter users
- ~5m facebook users
- ~2.1m instagram users (mostly between 18-24)
- ~0.5m articles/d (news, blogs, twitter, facebook, youtube, forums)
- ~7.0m internet users

- ~1 million entities
- ~3 million entity terms
- Sentiment:
  - ~87% in sentence level
  - ~70% in entity level
- NER
  - ~70% with 60% recall

Social Media in numbers (1 year period)



FACEBOOK

189 million New users

63 million New mobile users

79 billion Messages

154 billion Status updates

26 billion Shared links

53 billion Friends requests

71 billion Uploaded photos

TWITTER

16 million New users

13 million New mobile users

182 billion Tweets

767 billion Searches

YOUTUBE

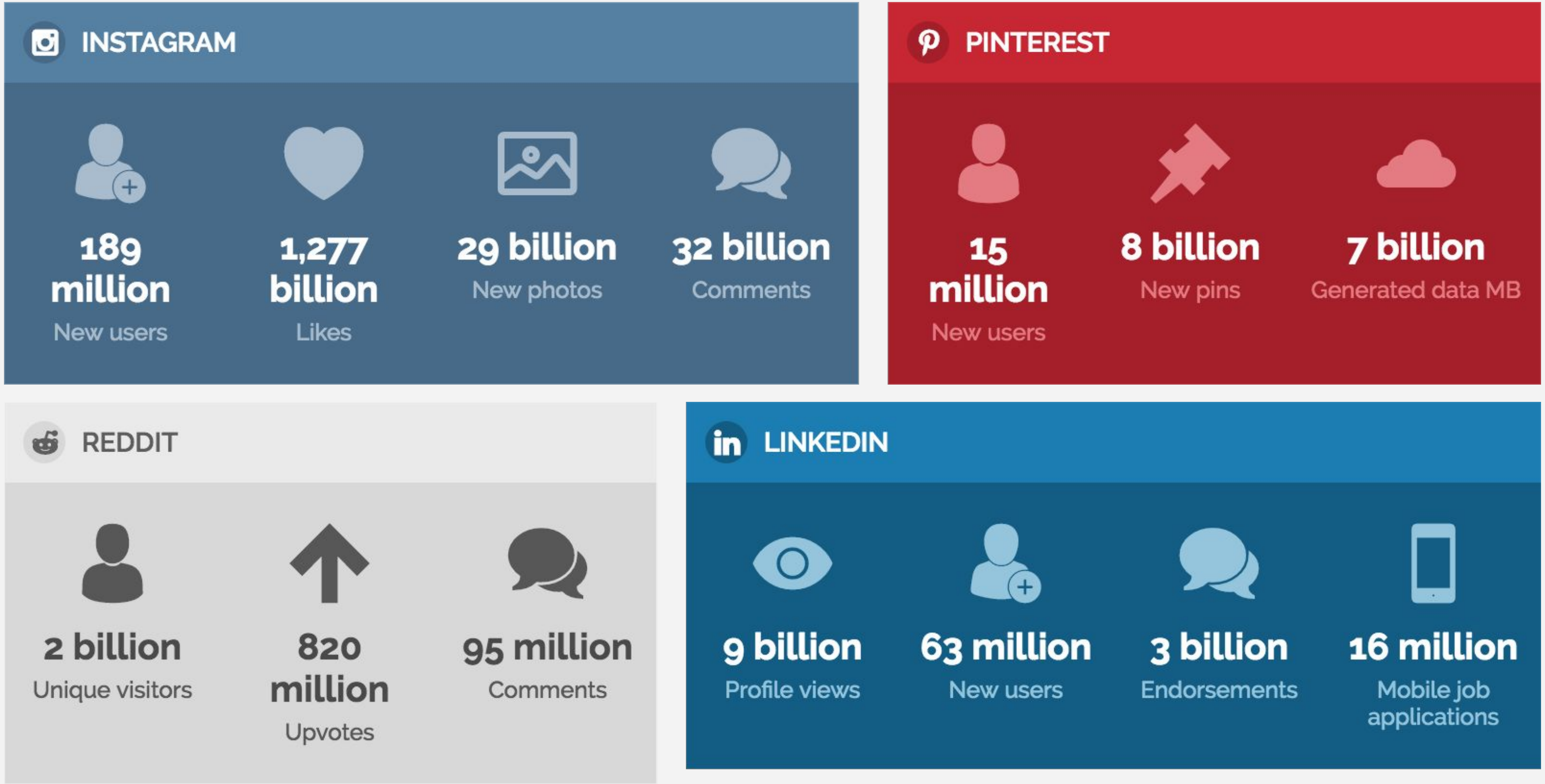
1,460 billion Views

365 billion Mobile views

43,771.97 Uploaded videos (hr)

source:coupofy.com

Social Media in numbers (1 year period)



INSTAGRAM

- 189 million New users
- 1,277 billion Likes
- 29 billion New photos
- 32 billion Comments

PINTEREST

- 15 million New users
- 8 billion New pins
- 7 billion Generated data MB

REDDIT

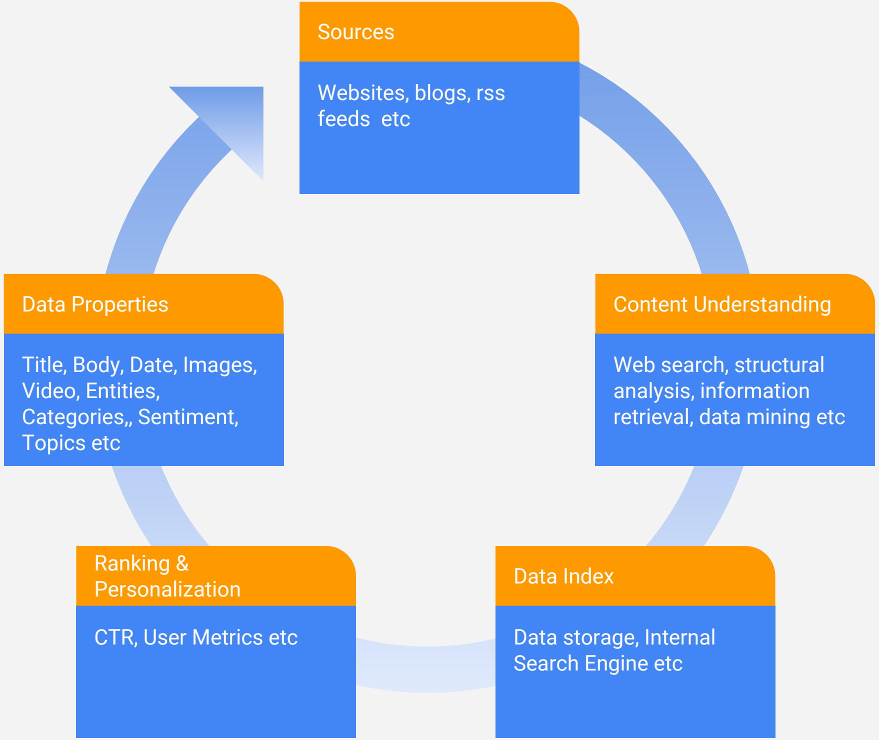
- 2 billion Unique visitors
- 820 million Upvotes
- 95 million Comments

LINKEDIN

- 9 billion Profile views
- 63 million New users
- 3 billion Endorsements
- 16 million Mobile job applications

source:coupofy.com

Content Lifecycle



Sources

Websites, blogs, rss feeds etc

Content Understanding

Web search, structural analysis, information retrieval, data mining etc

Data Index

Data storage, Internal Search Engine etc

Ranking & Personalization

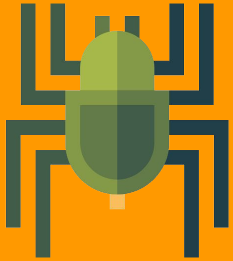
CTR, User Metrics etc

Data Properties

Title, Body, Date, Images, Video, Entities, Categories,, Sentiment, Topics etc

Data Collection

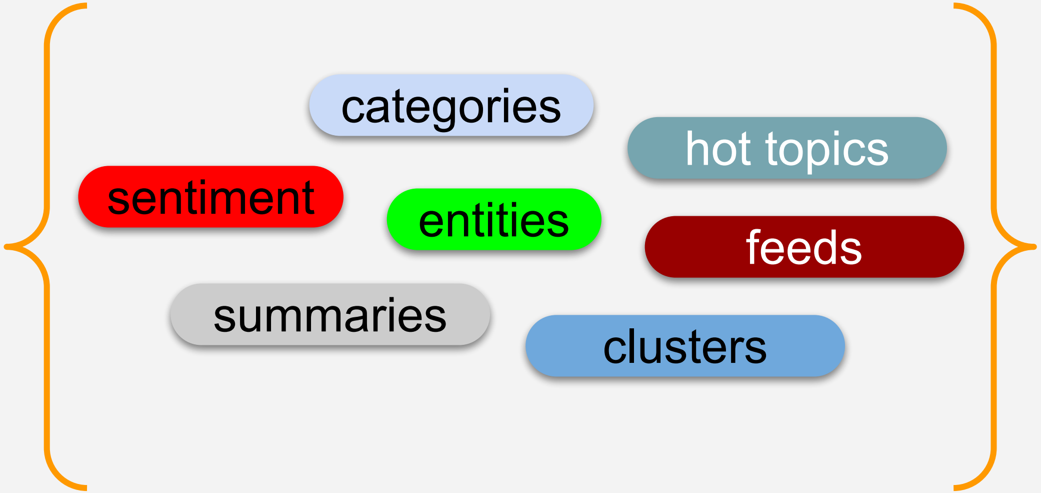
- Web-page Crawlers (xpath queries, regex)
  - ○ News Detection Content
  - ○ ...
- Blogs APIs
  - ○ Blogger
  - ○ Wordpress
  - ○ Medium
  - ○ ….
- Forum Crawlers
  - ○ PHPBB
  - ○ vBulletin
  - ○ ….



- Social Media Streams
  - ○ Twitter, Facebook ...
- Social Media APIs
  - ○ Twitter, Facebook, Youtube, Instagram ...
- Third Party Services
  - ○ Amazon Alexa
  - ○ Datasift
  - ○ …

- Linked Data
  - ○ DBpedia, Google Relation Extraction Corpus

News Content



categories

hot topics

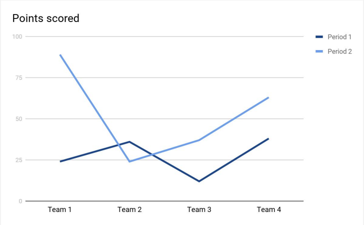
sentiment

entities

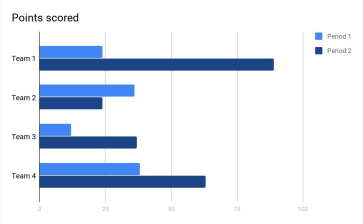
feeds

summaries

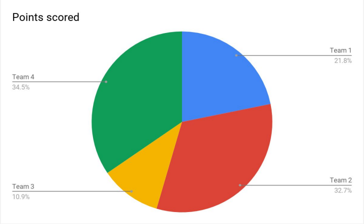
clusters



Points scored

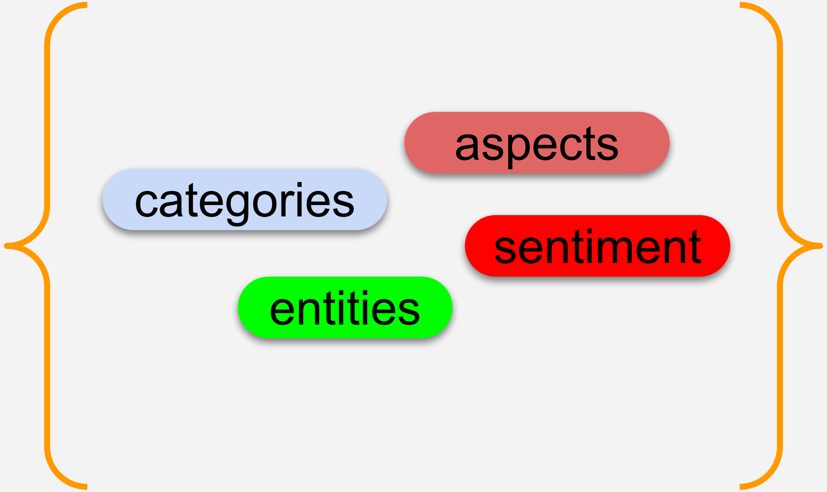


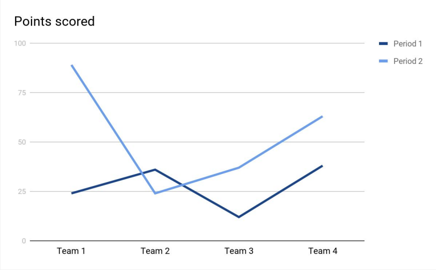
Points scored

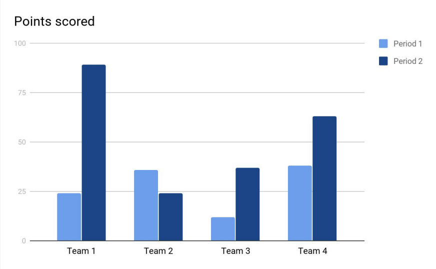


Points scored

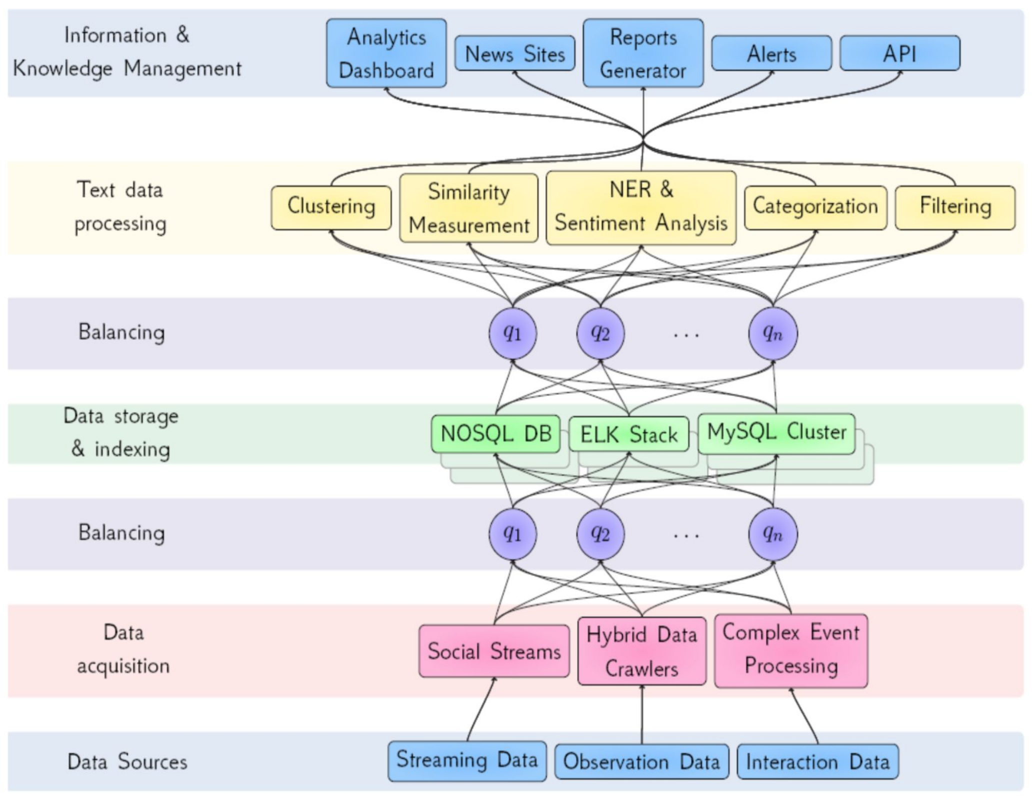
Social Media Content



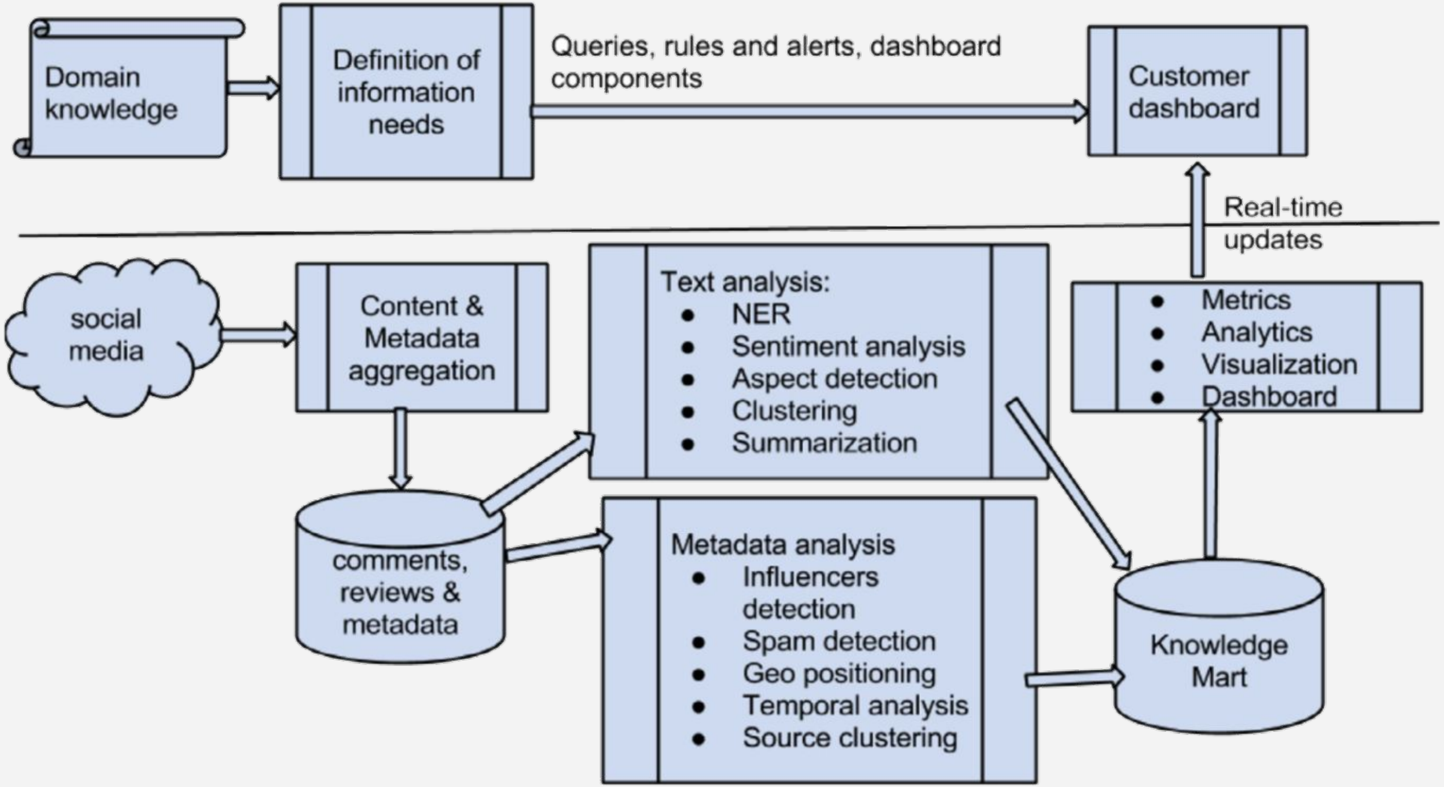




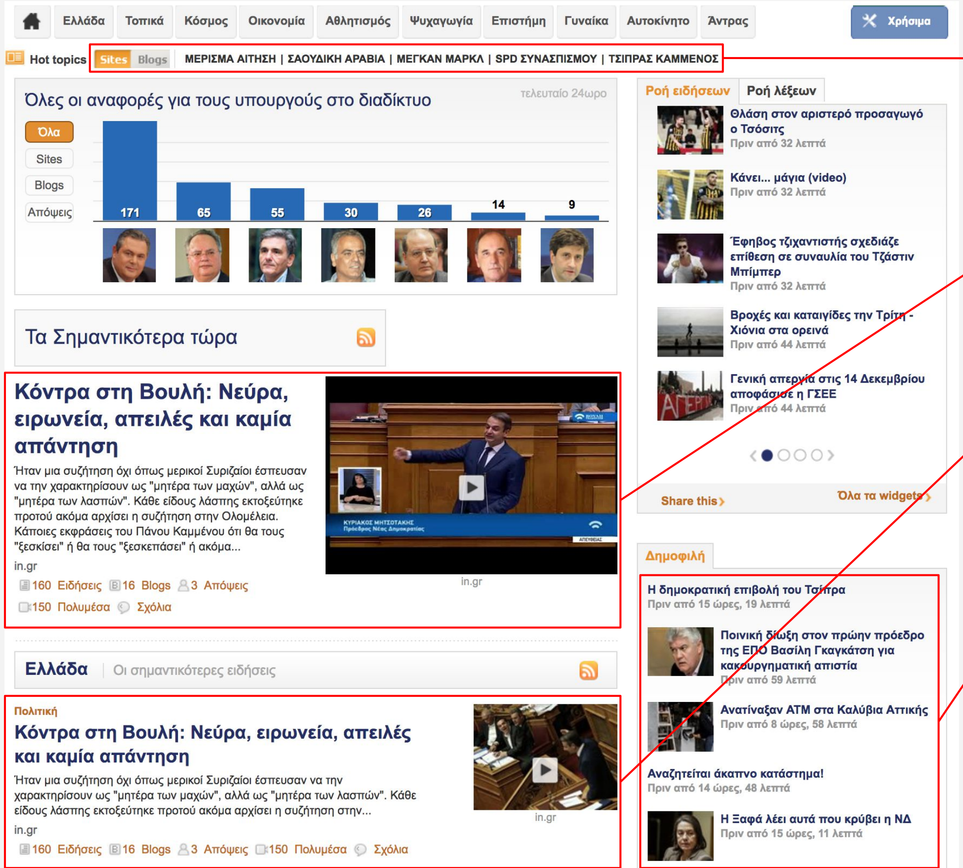
Architecture



Information Flow



Challenges (I)



Hot topics Sites Blogs ΜΕΡΙΣΜΑ ΑΙΤΗΣΗ | ΣΑΟΥΔΙΚΗ ΑΡΑΒΙΑ | ΜΕΓΚΑΝ ΜΑΡΚΛ | SPD ΣΥΝΑΣΠΙΣΜΟΥ | ΤΣΙΠΡΑΣ ΚΑΜΜΕΝΟΣ

Όλες οι αναφορές για τους υπουργούς στο διαδίκτυο

τελευταίο 24ωρο

Τα Σημαντικότερα τώρα

Κόντρα στη Βουλή: Νεύρα, ειρωνεία, απειλές και καμία απάντηση

Ήταν μια συζήτηση όχι όπως μερικοί Συριζαίοι έσπευσαν να την χαρακτηρίσουν ως "μητέρα των μαχών", αλλά ως "μητέρα των λασπών". Κάθε είδους λάσπης εκτοξεύτηκε προτού ακόμα αρχίσει η συζήτηση στην Ολομέλεια. Κάποιες εκφράσεις του Πάνου Καμμένου ότι θα τους "ξεσκίσει" ή θα τους "ξεσκεπάσει" ή ακόμα...

in.gr

160 Ειδήσεις 16 Blogs 3 Απόψεις 150 Πολυμέσα Σχόλια

in.gr

Ελλάδα | Οι σημαντικότερες ειδήσεις

Πολιτική

Κόντρα στη Βουλή: Νεύρα, ειρωνεία, απειλές και καμία απάντηση

Ήταν μια συζήτηση όχι όπως μερικοί Συριζαίοι έσπευσαν να την χαρακτηρίσουν ως "μητέρα των μαχών", αλλά ως "μητέρα των λασπών". Κάθε είδους λάσπης εκτοξεύτηκε προτού ακόμα αρχίσει η συζήτηση στην...

in.gr

160 Ειδήσεις 16 Blogs 3 Απόψεις 150 Πολυμέσα Σχόλια

in.gr

Ροή ειδήσεων | Ροή λέξεων

Θλάση στον αριστερό προσαγωγό ο Τσόσιτς
Πριν από 32 λεπτά

Κάνει... μάγια (video)
Πριν από 32 λεπτά

Έφηβος τζιχαντιστής σχεδιάζε επίθεση σε συναυλία του Τζάστιν Μπίμπερ
Πριν από 32 λεπτά

Βροχές και καταιγίδες την Τρίτη - Χιόνια στα ορεινά
Πριν από 44 λεπτά

Γενική απεργία στις 14 Δεκεμβρίου αποφάσισε η ΓΣΕΕ
Πριν από 44 λεπτά

Share this › Όλα τα widgets ›

Δημοφιλή

Η δημοκρατική επιβολή του Τσίπρα
Πριν από 15 ώρες, 19 λεπτά

Ποινική δίωξη στον πρώην πρόεδρο της ΕΠΟ Βασίλη Γκαγκάτση για κακουργηματική απιστία
Πριν από 59 λεπτά

Ανατίναξαν ΑΤΜ στα Καλύβια Αττικής
Πριν από 8 ώρες, 58 λεπτά

Αναζητείται άκαπνο κατάστημα!
Πριν από 14 ώρες, 48 λεπτά

Η Ξαφά λέει αυτά που κρύβει η ΝΔ
Πριν από 15 ώρες, 11 λεπτά

Hot topics

Top News / Breaking News

Clustering

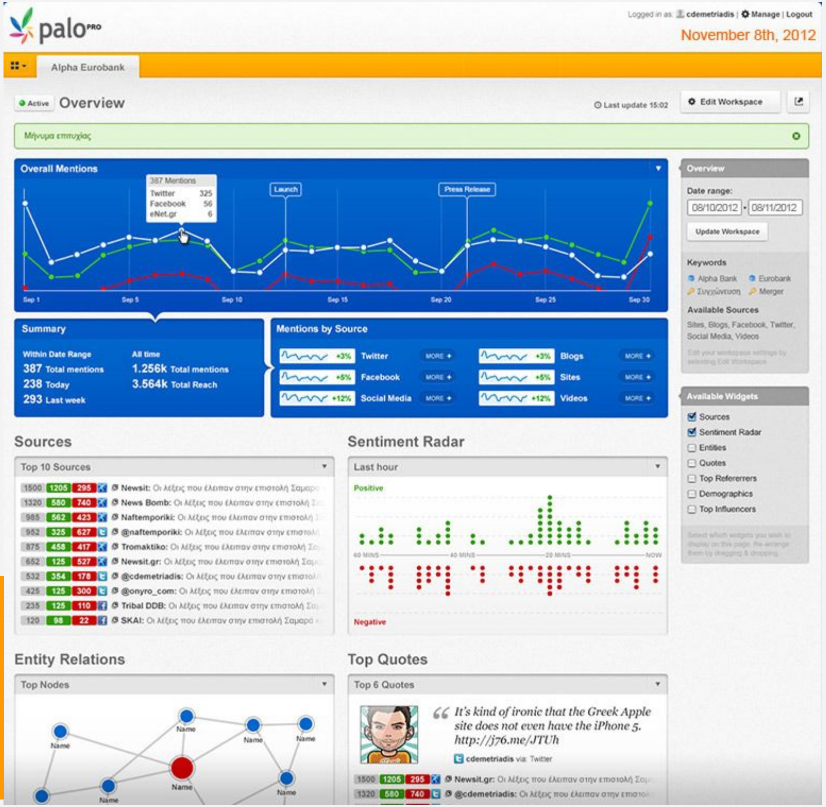
Personalized Popular

Challenges (II)

- Poor Content (outdated stories, missing media etc)
- Wrong Format (misspelling, wrong media etc)
- Legal Issues (copyright, plagiarism etc)
- Topic evolution (sport games, breaking news etc)
- Cluster Size (big news in short period etc)

The palopro.io example

Version 1.0



Logged in as: cdemetriadis | Manage | Logout

November 8th, 2012

Alpha Eurobank

● Active Overview

Last update 15:02

Edit Workspace

Μήνυμα επιτυχίας

Overview

Date range: 08/10/2012 - 08/11/2012

Update Workspace

Keywords

Alpha Bank, Eurobank, Συγχώνευση, Merger

Available Sources

Sites, Blogs, Facebook, Twitter, Social Media, Videos

Summary

Within Date Range

387 Total mentions
238 Today
293 Last week

All time

1.256k Total mentions
3.564k Total Reach

Mentions by Source

+3%	Twitter	+3%	Blogs
+5%	Facebook	+5%	Sites
+12%	Social Media	+12%	Videos

Available Widgets

☑ Sources
☑ Sentiment Radar
☐ Entities
☐ Quotes
☐ Top Referenrers
☐ Demographics
☐ Top Influencers

Sources

Top 10 Sources

1500	1205	295	Newsit: Οι λέξεις που έλειπαν στην επιστολή Σαμαρ
1320	580	740	News Bomb: Οι λέξεις που έλειπαν στην επιστολή Σ
985	562	423	Naftemporiki: Οι λέξεις που έλειπαν στην επιστολή
952	325	627	@naftemporiki: Οι λέξεις που έλειπαν στην επιστολ
875	458	417	Tromaktiko: Οι λέξεις που έλειπαν στην επιστολή Σα
652	125	527	Newsit.gr: Οι λέξεις που έλειπαν στην επιστολή Σαμ
532	354	178	@cdemetriadis: Οι λέξεις που έλειπαν στην επιστολ
425	125	300	@onyro_com: Οι λέξεις που έλειπαν στην επιστολή
235	125	110	Tribal DDB: Οι λέξεις που έλειπαν στην επιστολή Σα
120	98	22	SKAI: Οι λέξεις που έλειπαν στην επιστολή Σαμαρά

Sentiment Radar

Last hour

Entity Relations

Top Nodes

Top Quotes

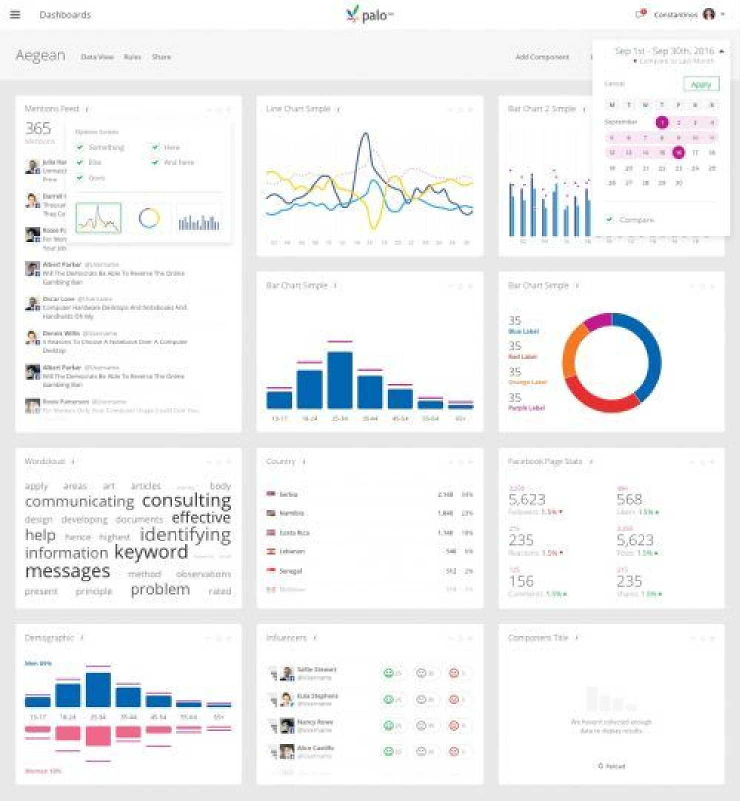
Top 6 Quotes

“It's kind of ironic that the Greek Apple site does not even have the iPhone 5. http://j76.me/JTUh

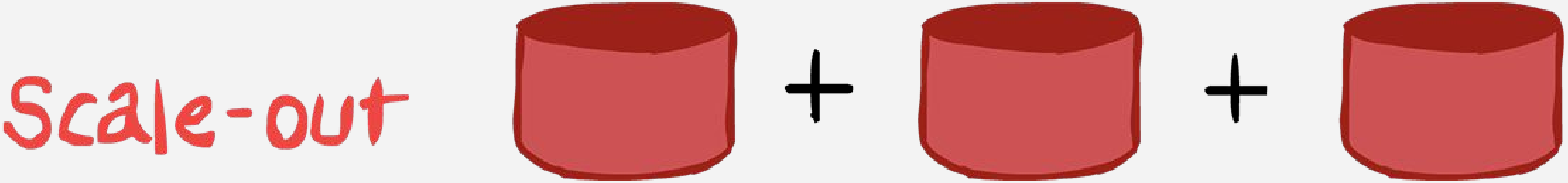
cdemetriadis via Twitter

1500	1205	295	Newsit.gr: Οι λέξεις που έλειπαν στην επιστολή Σα
1320	580	740	@cdemetriadis: Οι λέξεις που έλειπαν στην επιστολ

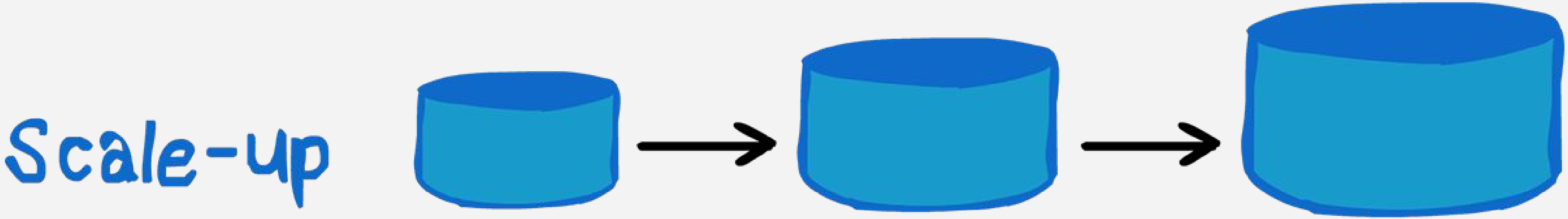
Version 2.0



Technologies



VS



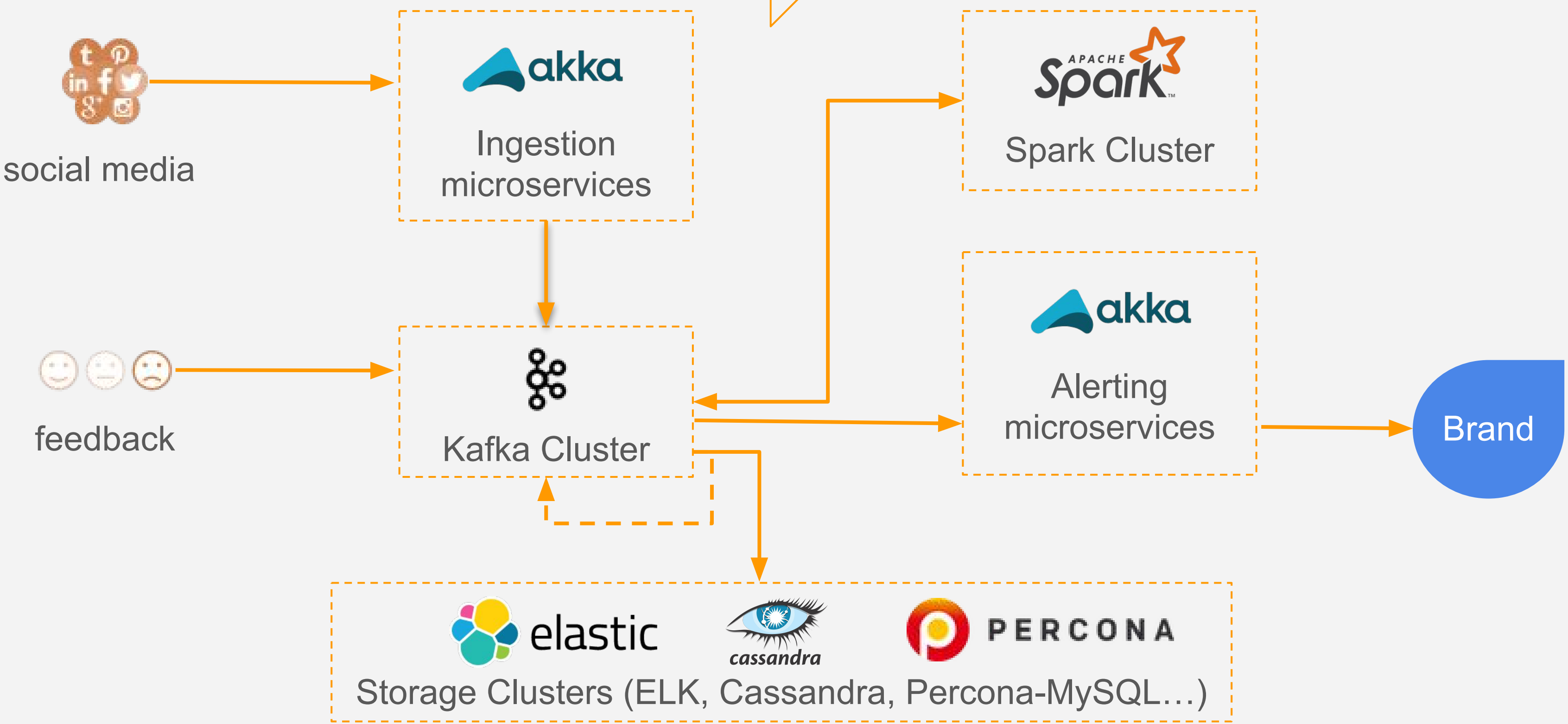
Technologies to scale up

Java & Python threaded services

RDBMS (MySQL)

Search (Solr)

Technologies to scale out



Case Study - Hands On

twitter

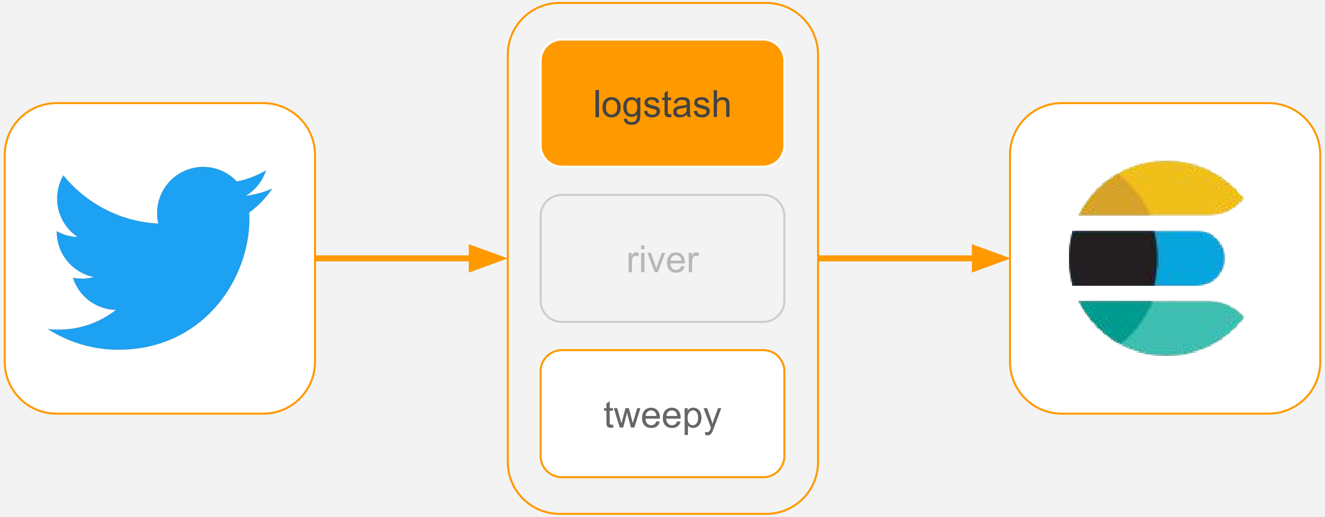
logstash

elasticsearch

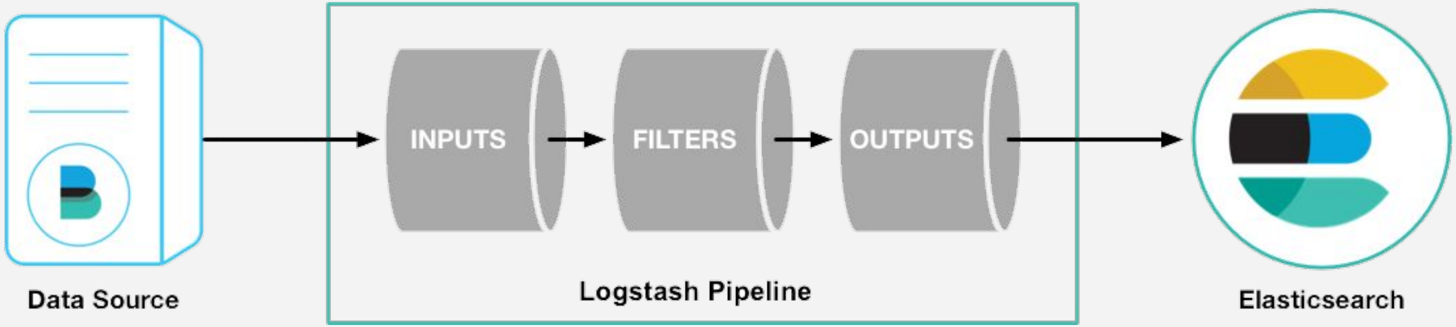
kibana

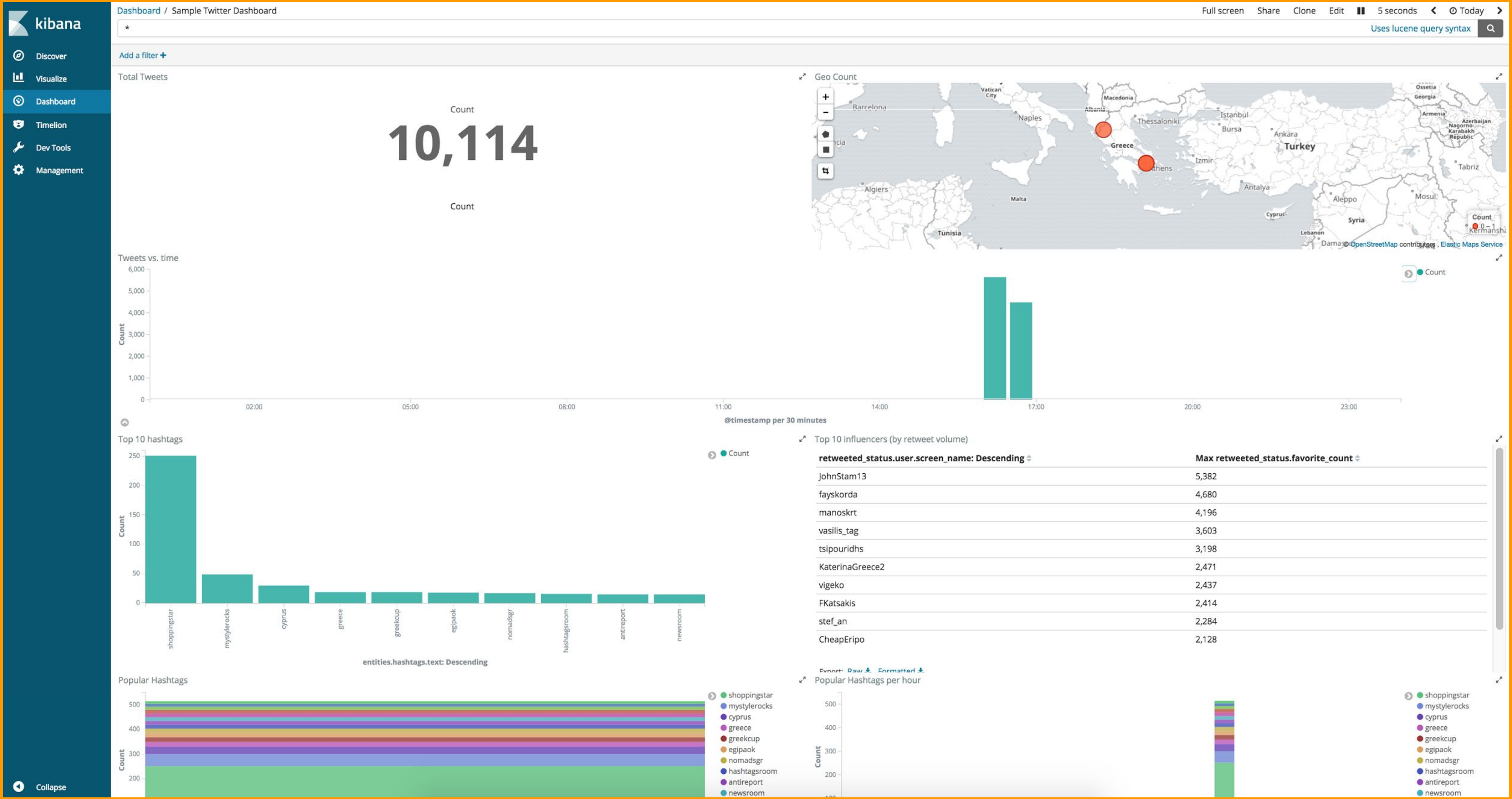
Ways to get data

- Logstash
- ~~Elasticsearch-twitter-river~~ (March 2015 - deprecated due to cluster stability)
- Tweepy



The Logstash Pipeline





Dashboard / Sample Twitter Dashboard

*

Total Tweets

Count

10,114

Count

Top 10 influencers (by retweet volume)

retweeted_status.user.screen_name: Descending	Max retweeted_status.favorite_count
JohnStam13	5,382
fayskorda	4,680
manoskrt	4,196
vasilis_tag	3,603
tsipouridhs	3,198
KaterinaGreece2	2,471
vigeko	2,437
FKatsakis	2,414
stef_an	2,284
CheapEripo	2,128

Export: Raw Formatted

Questions?

@tsirakis