# Detecting events on the Web in real-time with **Java**, **Kafka** & **ZooKeeper**

Dr. James Stanier | brandwatch.com | jamess@brandwatch.com

# Coming Up/

- Me, Brandwatch and new problems
- Moving to Kafka
- Processing data
- Distributing work
- Finding meaning

Who?

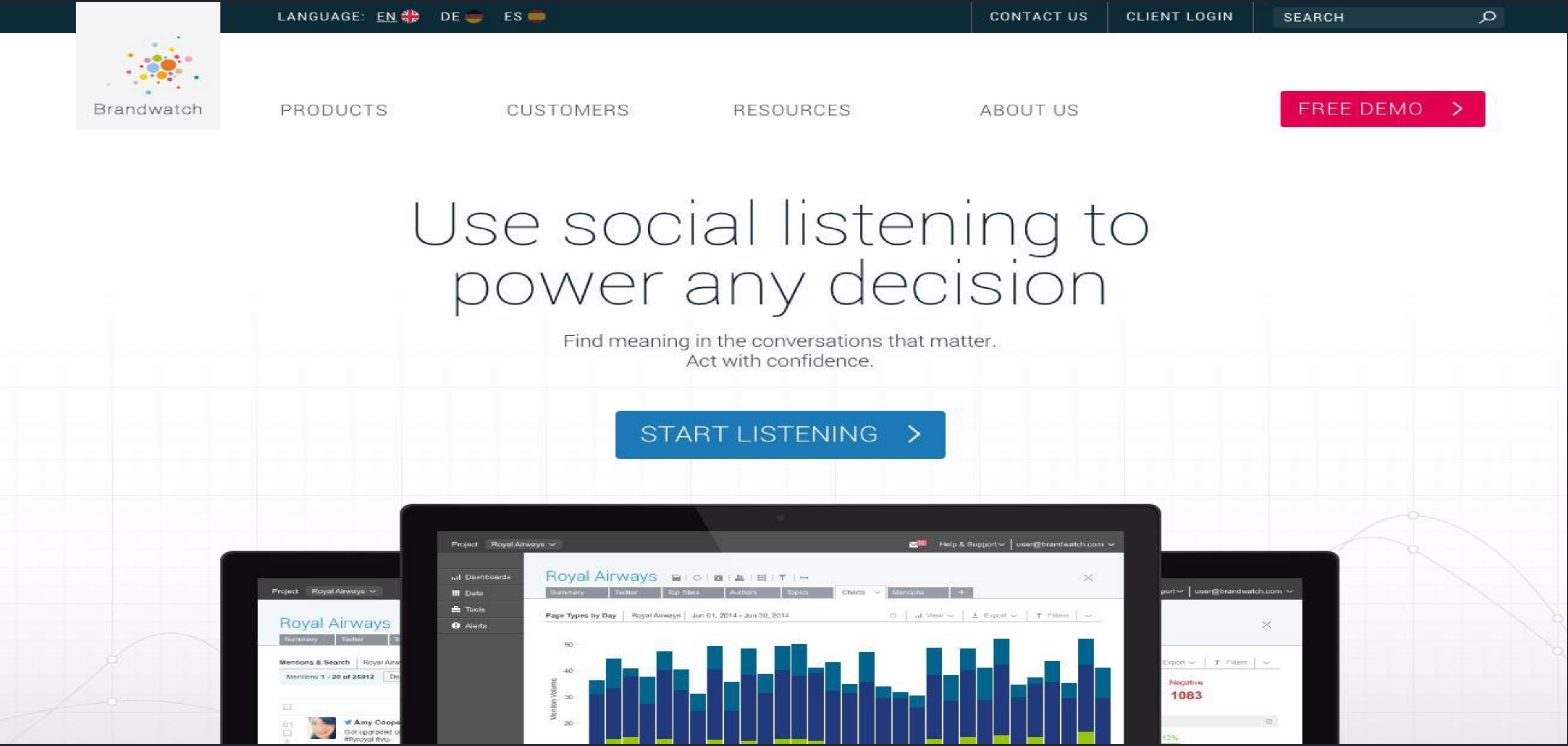
## Dr. James Stanier

VP Engineering,

Product (Backend)

Brandwatch

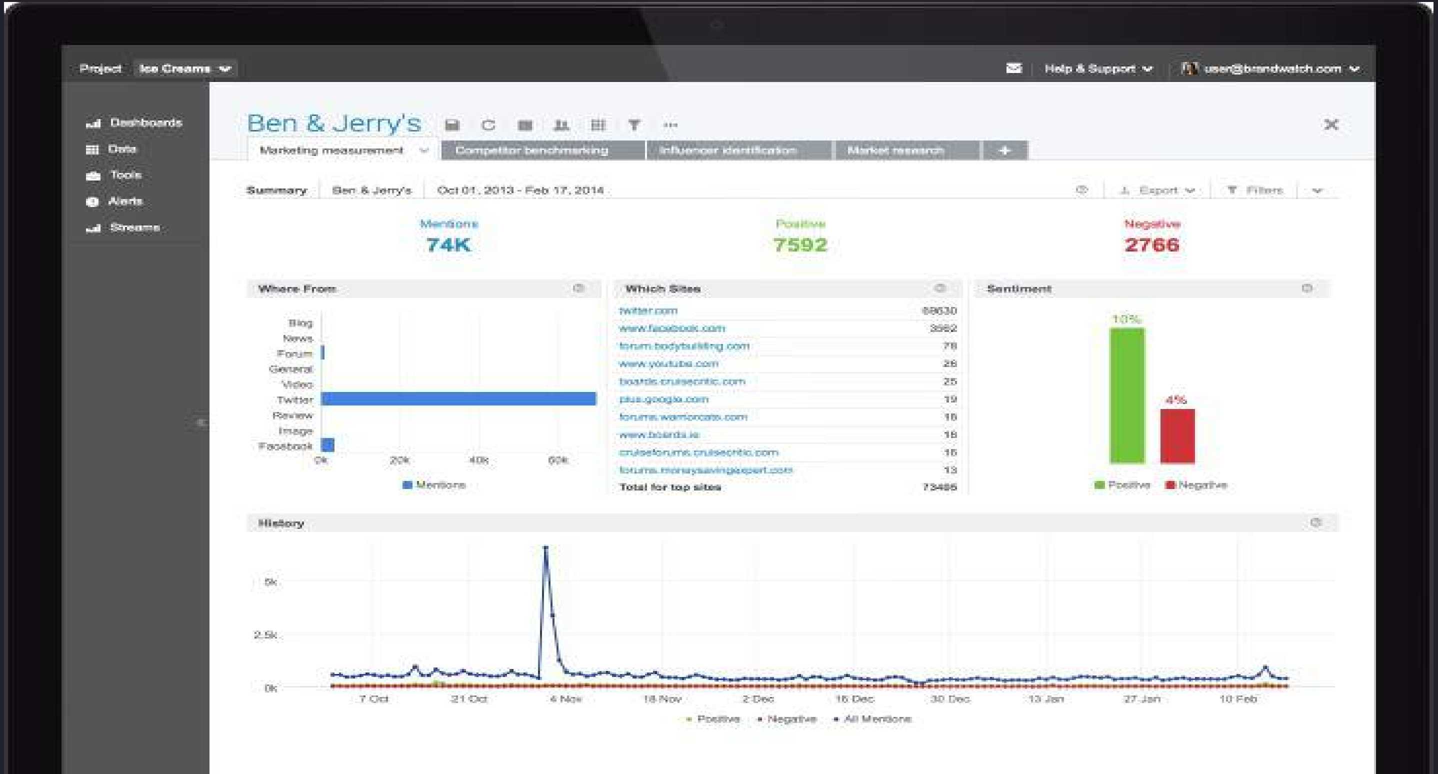
@jstanier | jamess@brandwatch.com



# Use social listening to power any decision

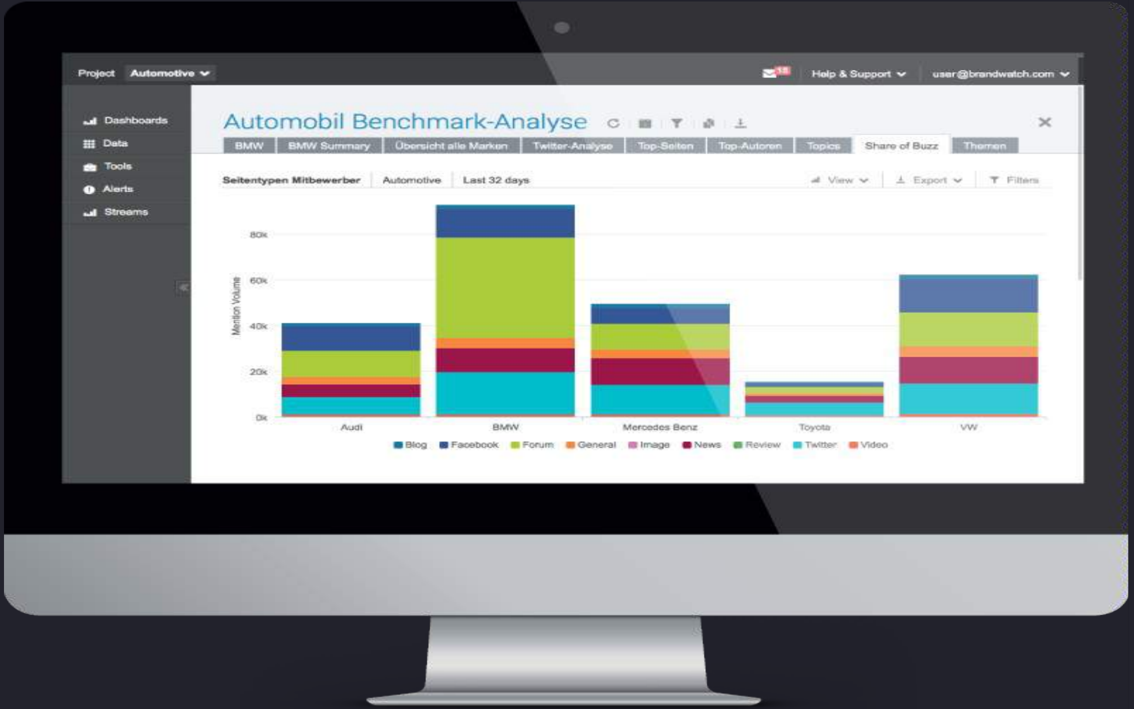
Find meaning in the conversations that matter.
Act with confidence.

# Data/ Presentation

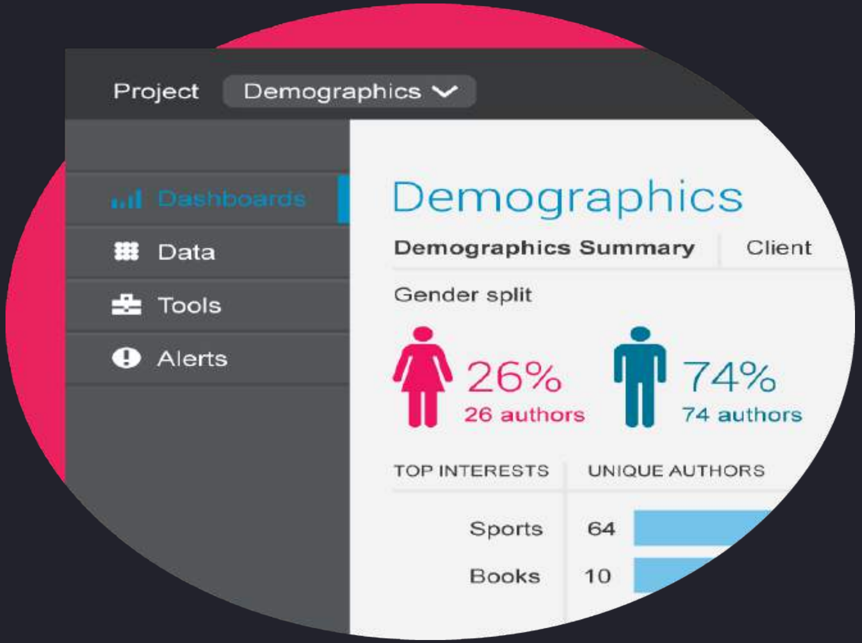


| Which Sites | |
|---|---|
| twitter.com | 69630 |
| www.facebook.com | 3562 |
| forum.bodybuilding.com | 78 |
| www.youtube.com | 26 |
| boards.cruisecritic.com | 25 |
| plus.google.com | 19 |
| forums.warriorcats.com | 16 |
| www.boards.ie | 16 |
| cruiseforums.cruisecritic.com | 15 |
| forums.moneysavingexpert.com | 13 |
| **Total for top sites** | 73485 |

# Data/ Aggregation



# Data/ Classification



Demographics

Demographics Summary | Client

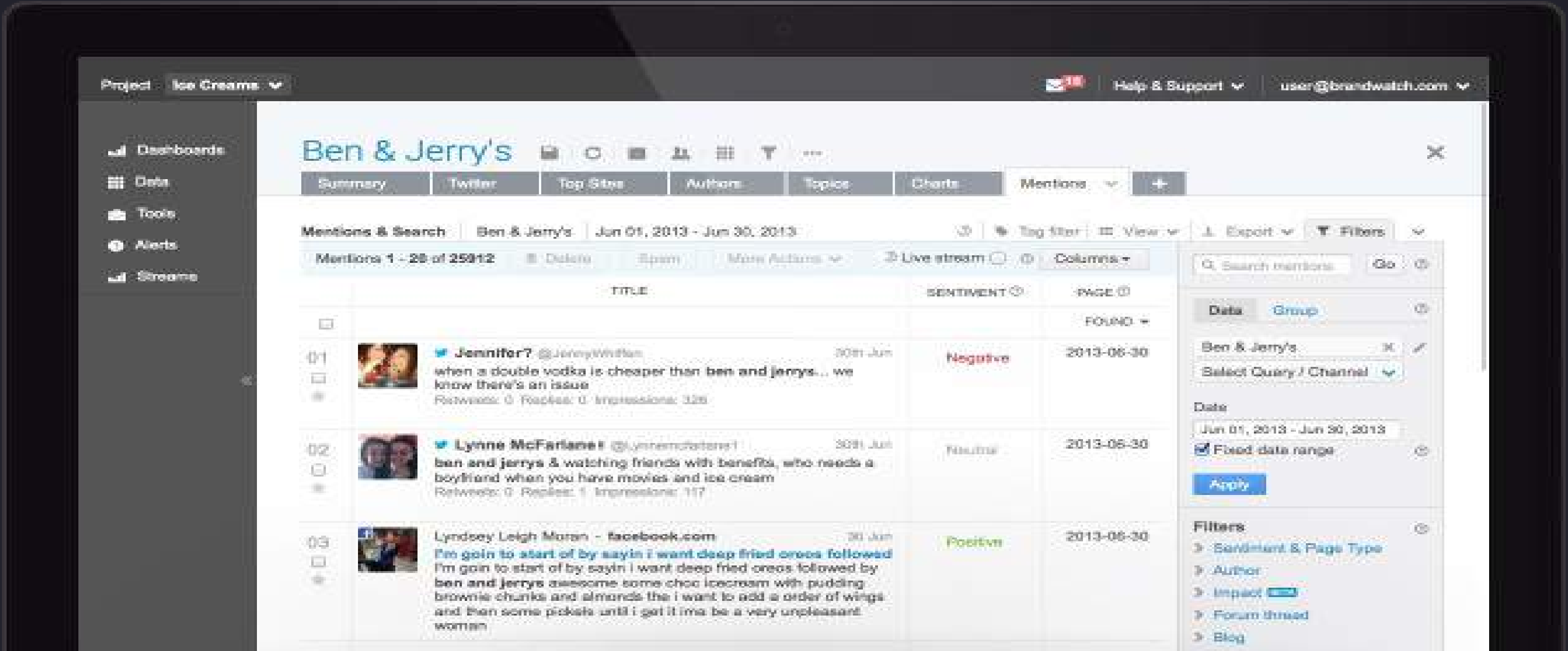
Gender split

26%
26 authors

74%
74 authors

| TOP INTERESTS | UNIQUE AUTHORS |
|---|---|
| Sports | 64 |
| Books | 10 |

# Data/ Not just top level metrics



Mentions & Search | Ben & Jerry's | Jun 01, 2013 - Jun 30, 2013

Mentions 1 - 20 of 25912

| | TITLE | SENTIMENT | PAGE FOUND |
|---|---|---|---|
| 01 | Jennifer? @JennyWhitfan 30th Jun<br>when a double vodka is cheaper than **ben and jerrys**... we know there's an issue<br>Retweets: 0 Replies: 0 Impressions: 326 | Negative | 2013-06-30 |
| 02 | Lynne McFarlane• @lynnemcfarlane1 30th Jun<br>**ben and jerrys** & watching friends with benefits, who needs a boyfriend when you have movies and ice cream<br>Retweets: 0 Replies: 1 Impressions: 117 | Neutral | 2013-06-30 |
| 03 | Lyndsey Leigh Moran - **facebook.com** 30 Jun<br>I'm goin to start of by sayin i want deep fried oreos followed<br>I'm goin to start of by sayin i want deep fried oreos followed by **ben and jerrys** awesome some choc icecream with pudding brownie chunks and almonds the i want to add a order of wings and then some pickels until i get it ima be a very unpleasant woman | Positive | 2013-06-30 |

Ben & Jerry's

Select Query / Channel

Date

Jun 01, 2013 - Jun 30, 2013

☑ Fixed date range

**Filters**

- Sentiment & Page Type
- Author
- Impact
- Forum thread
- Blog

# Data/ The numbers

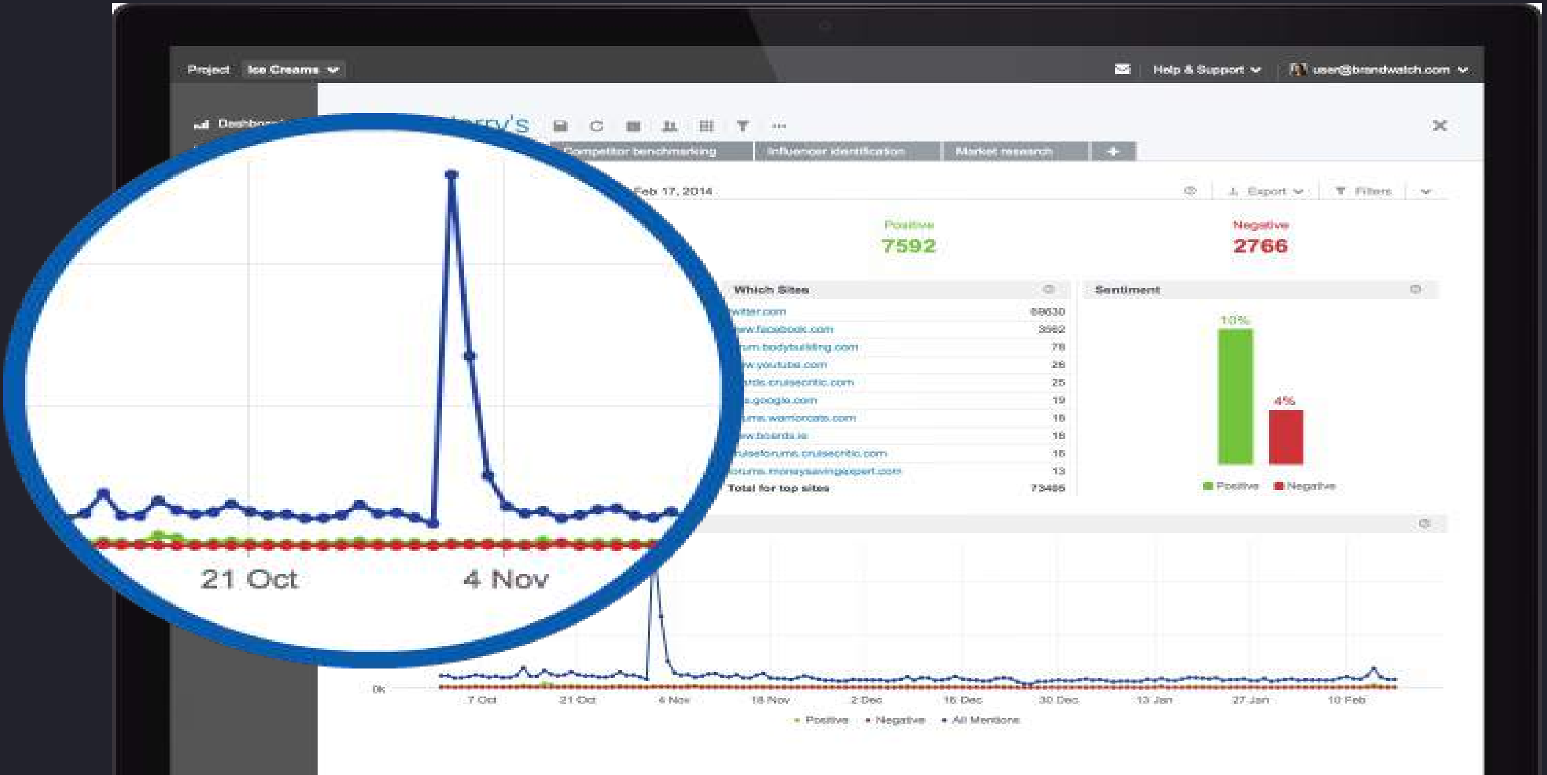
- 50+ Java Web Crawlers
- 10+ Historical crawlers for new queries
- Twitter via GNIP (now Twitter), Weibo, Disqus and more
- 80M+ query matches per day

My Brand Sentiment
Competitor 1
Competitor 2
741
7708
400
My Brand Mentions
Competitor 1 Mentions
Competitor 2 Mentions
21830
774
3222
My brand History
Competitor 1 History
Competitor 2 History
My Brand Topics
Competitor 2 Topics
01 Sep
Contact

Globe
September 29 1:56pm
TWITTER Mentions
Global
Apple Watch
Sony Smartwatch 3
Motorola Moto 360
8133
7271
164
220
Demo project
Display 1
Display 2

# A new challenge

# The challenge/ The signal from the noise



Positive
7592

Negative
2766

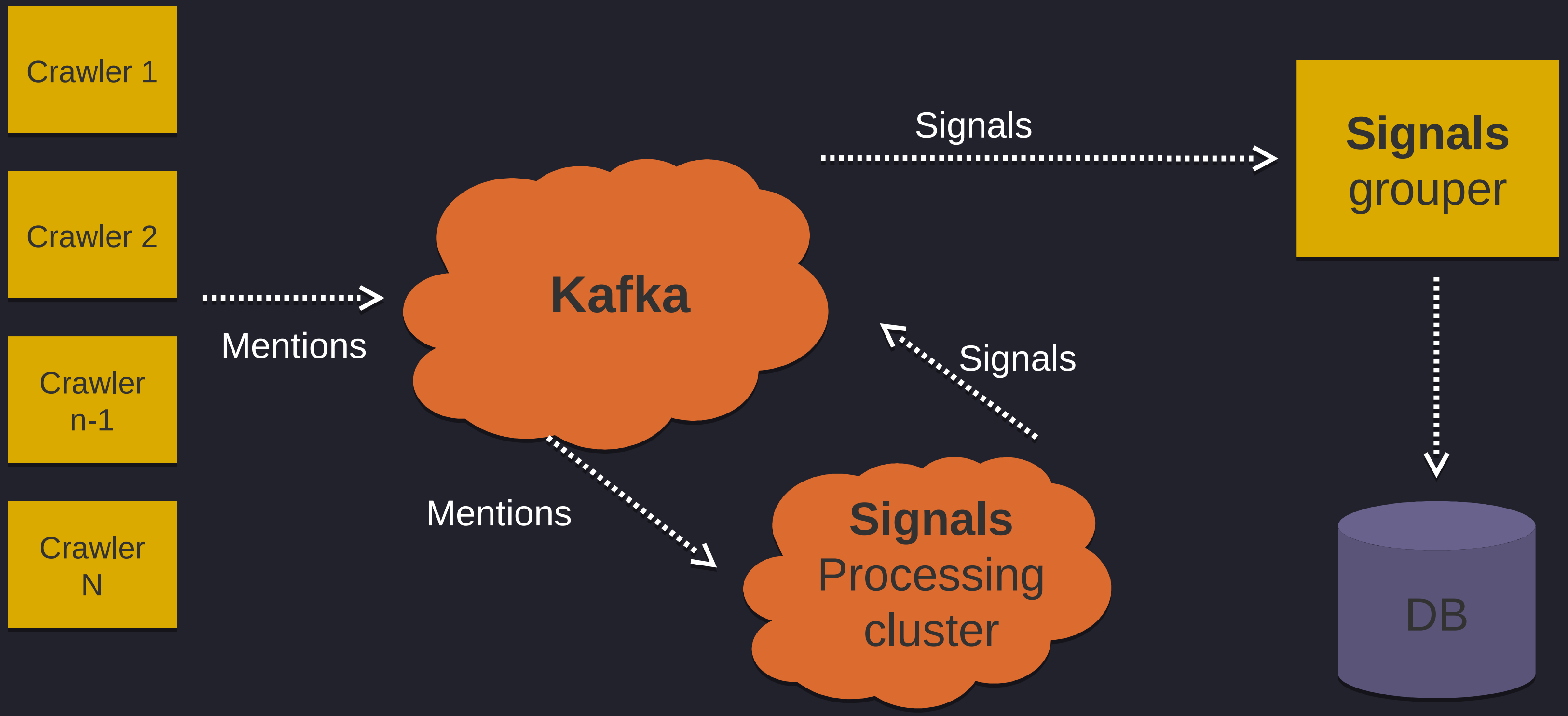
| Which Sites | |
|---|---|
| witter.com | 69630 |
| ww.facebook.com | 3562 |
| rum.bodybuilding.com | 78 |
| w.youtube.com | 26 |
| ards.cruisecritic.com | 25 |
| s.google.com | 19 |
| ums.warriorcats.com | 16 |
| ww.boards.ie | 16 |
| ruiseforums.cruisecritic.com | 16 |
| forums.moneysavingexpert.com | 13 |
| Total for top sites | 73486 |

Sentiment

# The challenge/ at scale

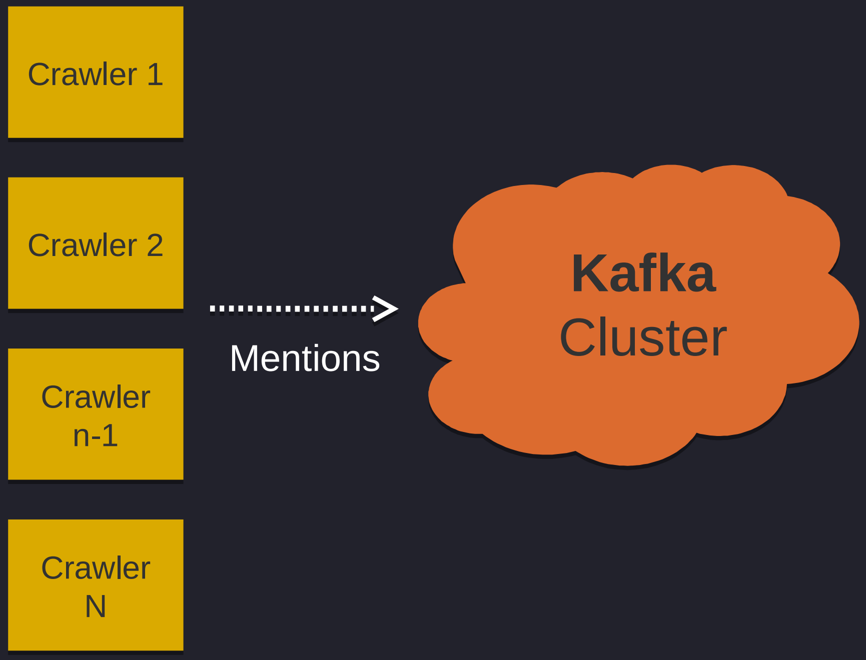
- 130K+ user queries
- 80M+ mentions per day
- Polling the data stores for mentions for all queries takes 8hrs for one pass

# The Problem/ How we handled it…



Kafka

# Step 1/ Kafka



Crawler 1

Crawler 2

Crawler n-1

Crawler N

Mentions

**Kafka** Cluster

© 2014 Brandwatch | www.brandwatch.com

# Kafka/ What is it?

- Apache Kafka is a publish-subscribe messaging system rethought as a distributed commit log
- Apache top level project November 2013
- Started at LinkedIn

# Kafka/ is…

- Fast: hundreds of MBs read/write per second from thousands of clients
- Scalable: clustered, partitioned over many machines, expanded without downtime
- Durable: messages persisted to disk and replicated in cluster

# Kafka/ Written to disk?

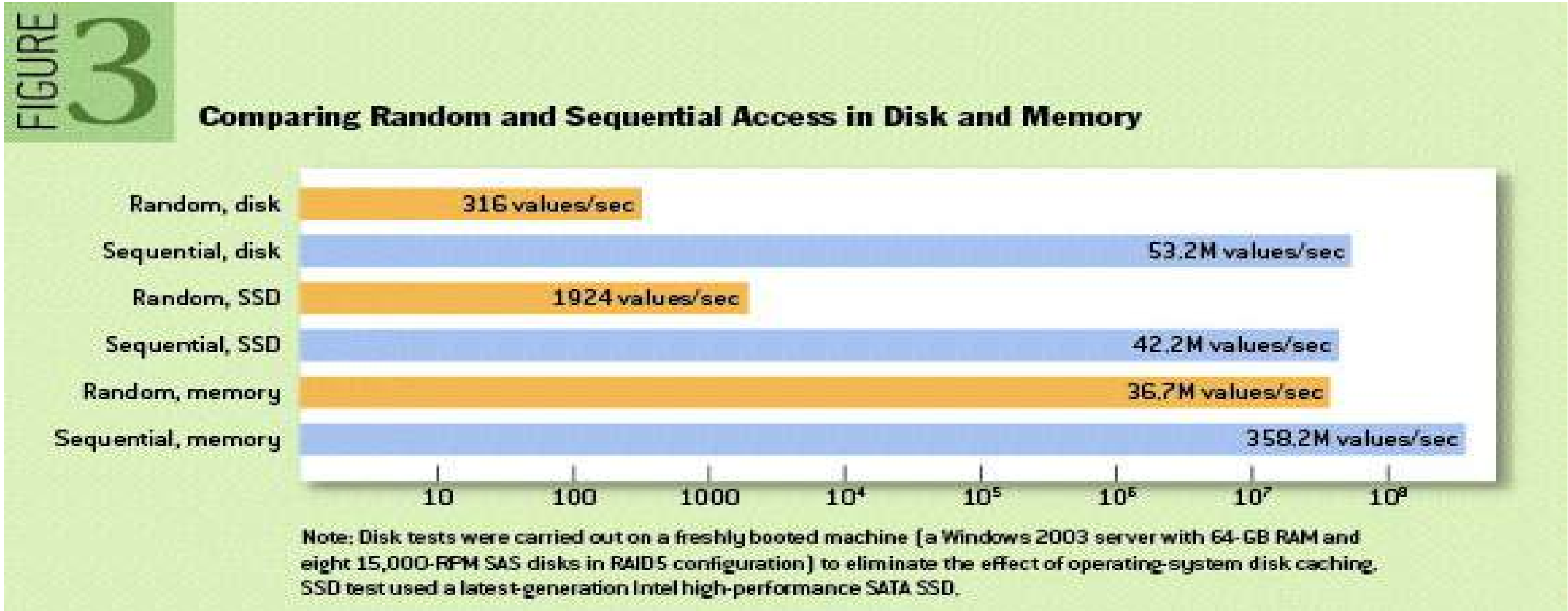


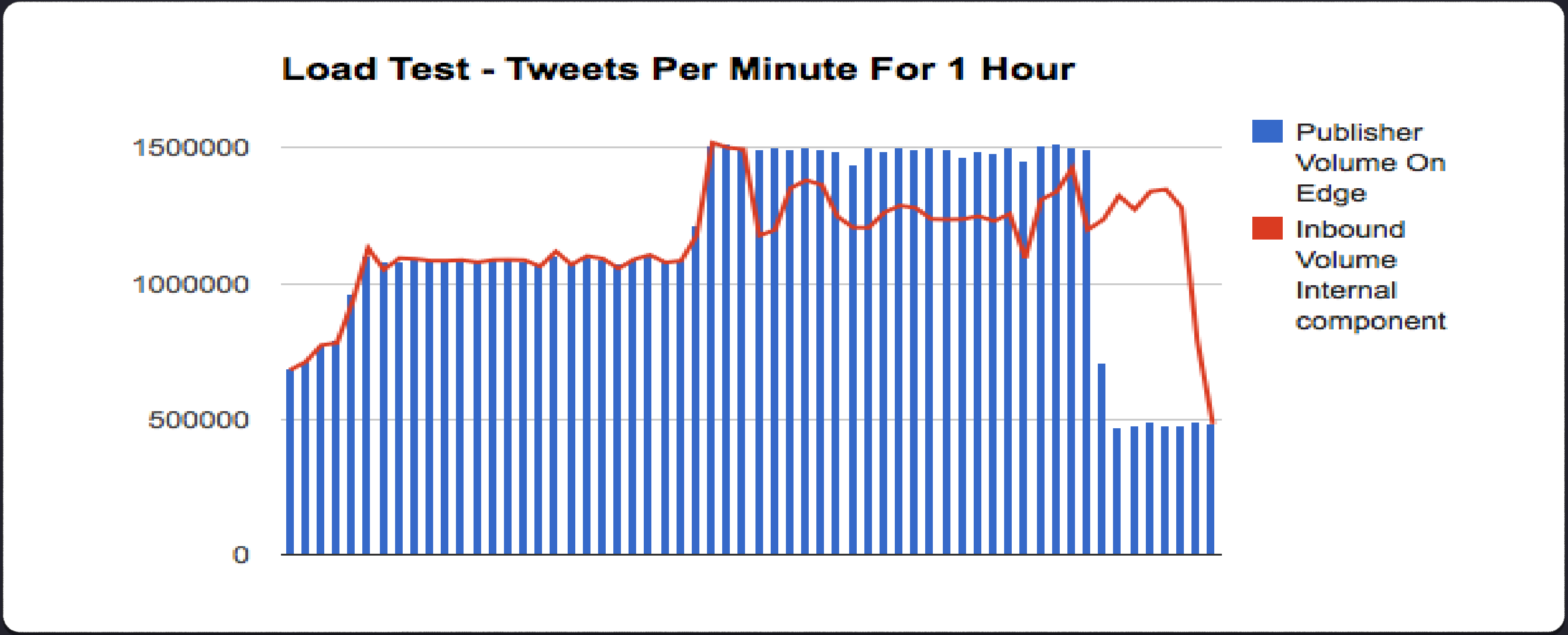
FIGURE 3

Comparing Random and Sequential Access in Disk and Memory

Note: Disk tests were carried out on a freshly booted machine (a Windows 2003 server with 64-GB RAM and eight 15,000-RPM SAS disks in RAID5 configuration) to eliminate the effect of operating-system disk caching. SSD test used a latest-generation Intel high-performance SATA SSD.

http://q.acm.org/detail.cfm?id=1563874

# Kafka/ Bending, not breaking



http://engineering.gnip.com/tag/kafka/

# Kafka/ Sending from the crawlers

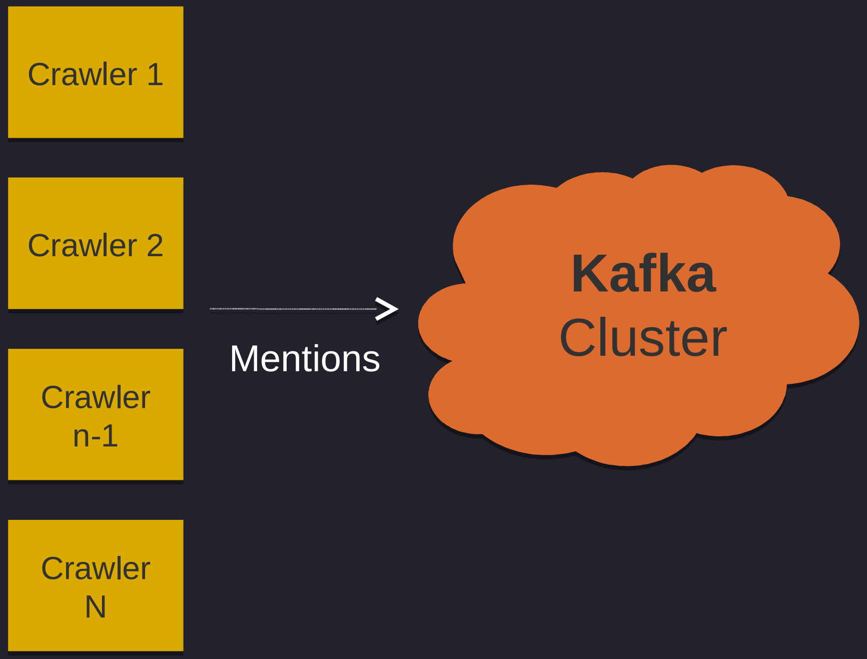
```
String message = toJson(...);

KeyedMessage<String, String> message = new

    KeyedMessage<String, String>("query.mentions",
queryId, message);

producer.send(message);
```

# Step 1/ Done



Crawler 1

Crawler 2

Crawler n-1

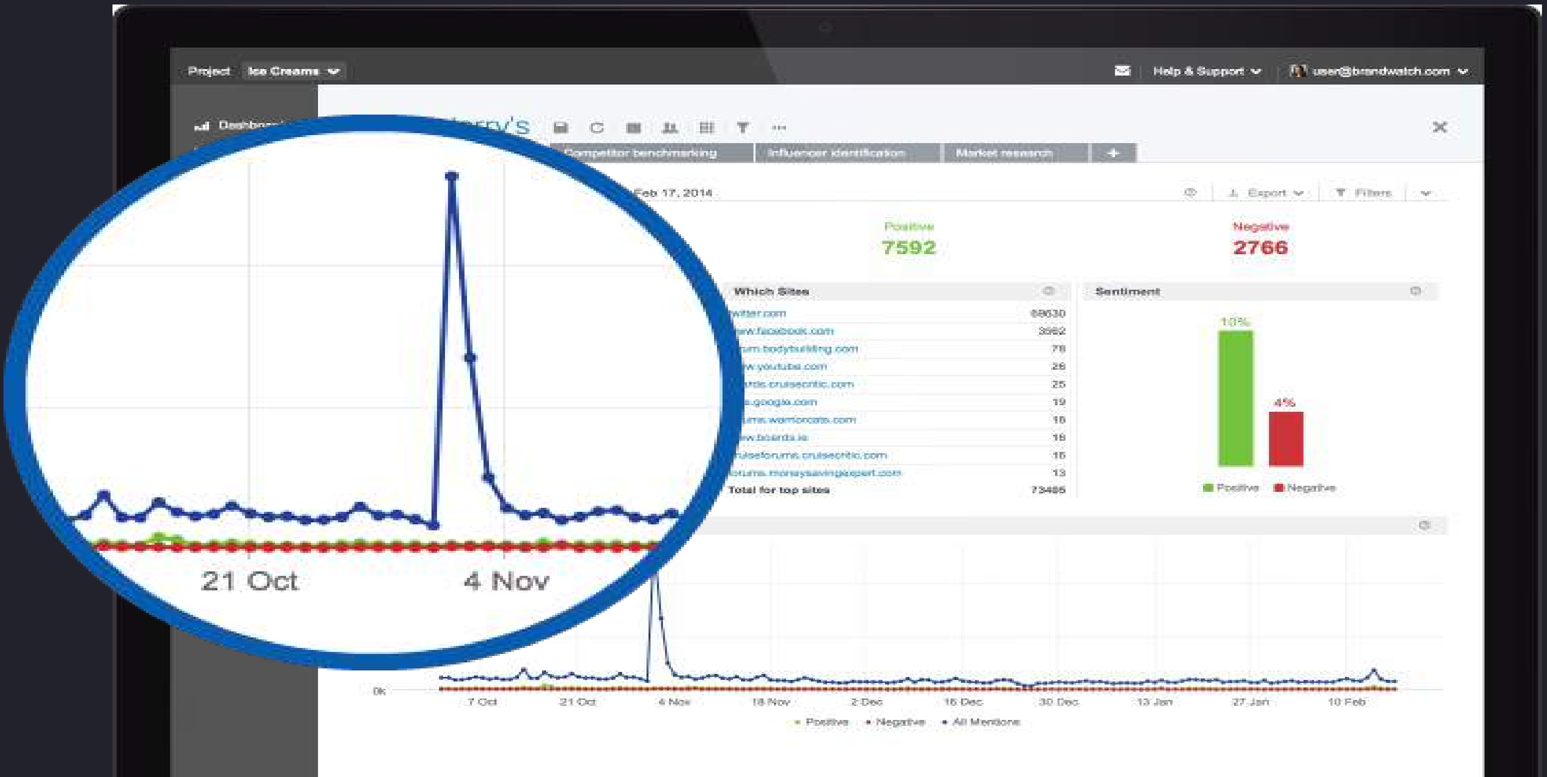
Crawler N

Mentions

**Kafka** Cluster

# Processing

# Processing/ What’s happening now?



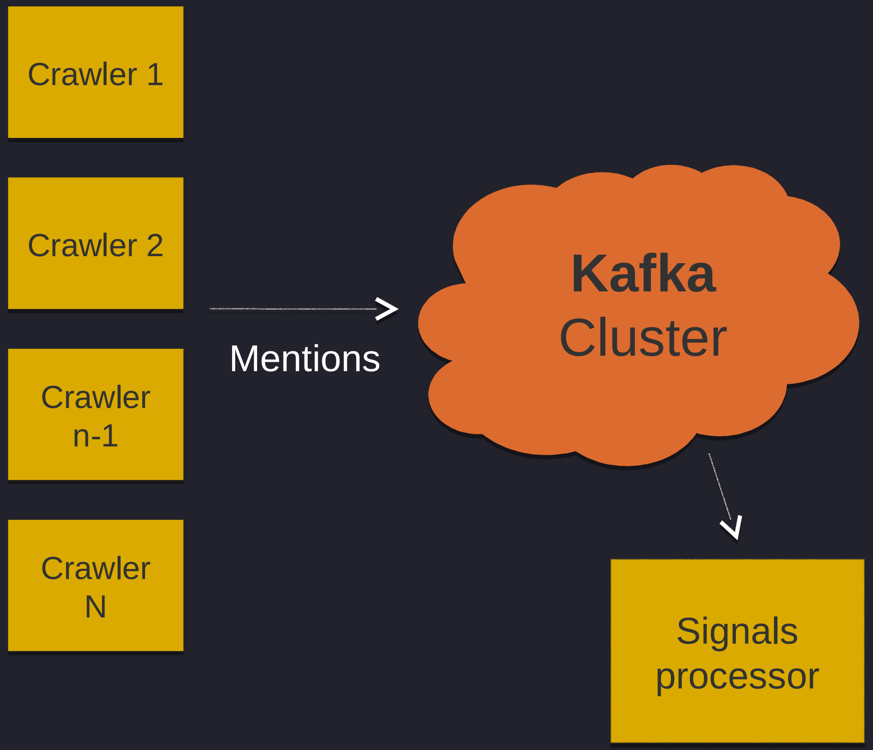
Feb 17, 2014

Positive
7592

Negative
2766

| Which Sites | |
|---|---|
| witter.com | 69630 |
| ww.facebook.com | 3562 |
| um.bodybuilding.com | 78 |
| w.youtube.com | 26 |
| ards.cruisecritic.com | 25 |
| s.google.com | 19 |
| ums.warriorcats.com | 18 |
| w.boards.ie | 18 |
| ruiseforums.cruisecritic.com | 16 |
| forums.moneysavingexpert.com | 13 |
| **Total for top sites** | 73485 |

# Step 2.1/ One processing JVM



Crawler 1

Crawler 2

Crawler n-1

Crawler N

Mentions

**Kafka** Cluster

Signals processor

# Processing/ A wild tweet appears!

Mention

date: 01/06/2015 16:05

pageType: twitter

author: @berlinperson

hashtags: [#berlinbuzzwords, #amazingtalk, #greatshoes]

mentionedTweeters: [@jstanier]

text: “@jstanier is at #berlinbuzzwords #amazingtalk #greatshoes”

# Processing/ Storing hashtags

```
Map<Date, Multiset<String>>
```

Initialise with the last 24 hours

# Processing/ Storing hashtags

```
Map<Date, Multiset<String>>
```

Mention

date: 01/06/2015 4:10PM

hashtags: [#berlinbuzzwords, #amazingtalk, #greatshoes]

# Processing/ Storing hashtags

```
Map<Date, Multiset<String>>
```

Mention

date: 01/06/2015 4:10PM

hashtags: [#berlinbuzzwords, #amazingtalk, #greatshoes]

```
add(“#berlinbuzzwords”)
add(“#amazingtalk”)
add(“#greatshoes”)
```

# Processing/ Cycling the buckets

```
@Scheduled(cron = "0 0 * * * *")
public void cycleBuckets() {
     Date oldest = buckets.lastKey();
     removeBucket(oldest);
     DateTime newest = new
       DateTime(buckets.firstKey());
     addBucket(newest.plusHours(1).toDate());
 }
```

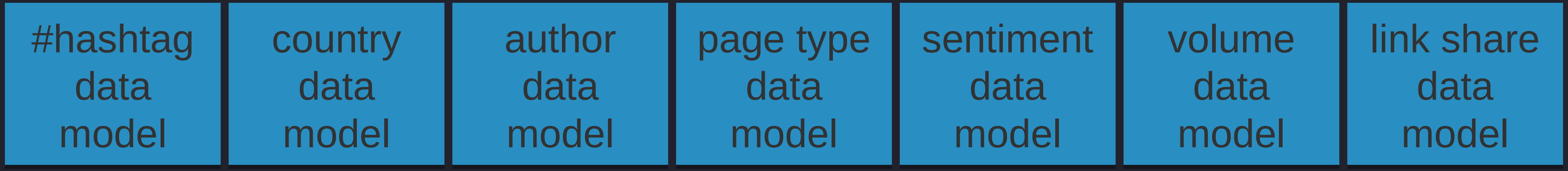
# Processing/ Detecting spikes

- At scheduled intervals
- For each #
  - Convert to a timeseries [5, …. 1002, 5499]
  - Compare previous hour to history
  - Give a score to it
- If score > threshold, it’s interesting
- Send it on a new Kafka topic

# Processing/ What we just did

#hashtag data model

# Processing/ But we also track…



#hashtag data model

country data model

author data model

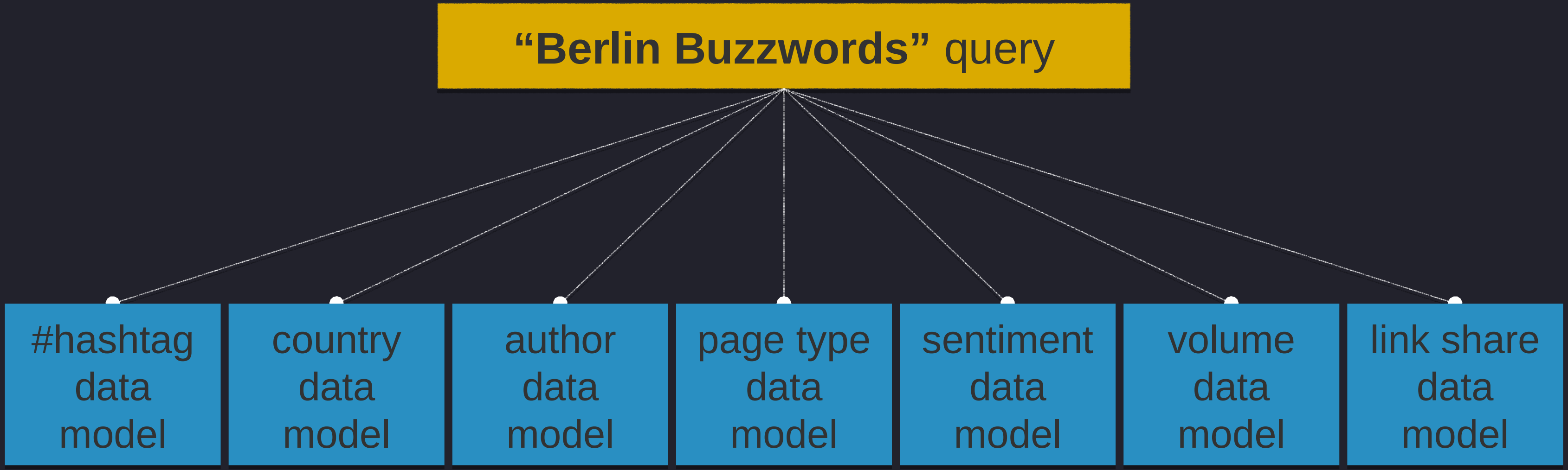
page type data model

sentiment data model

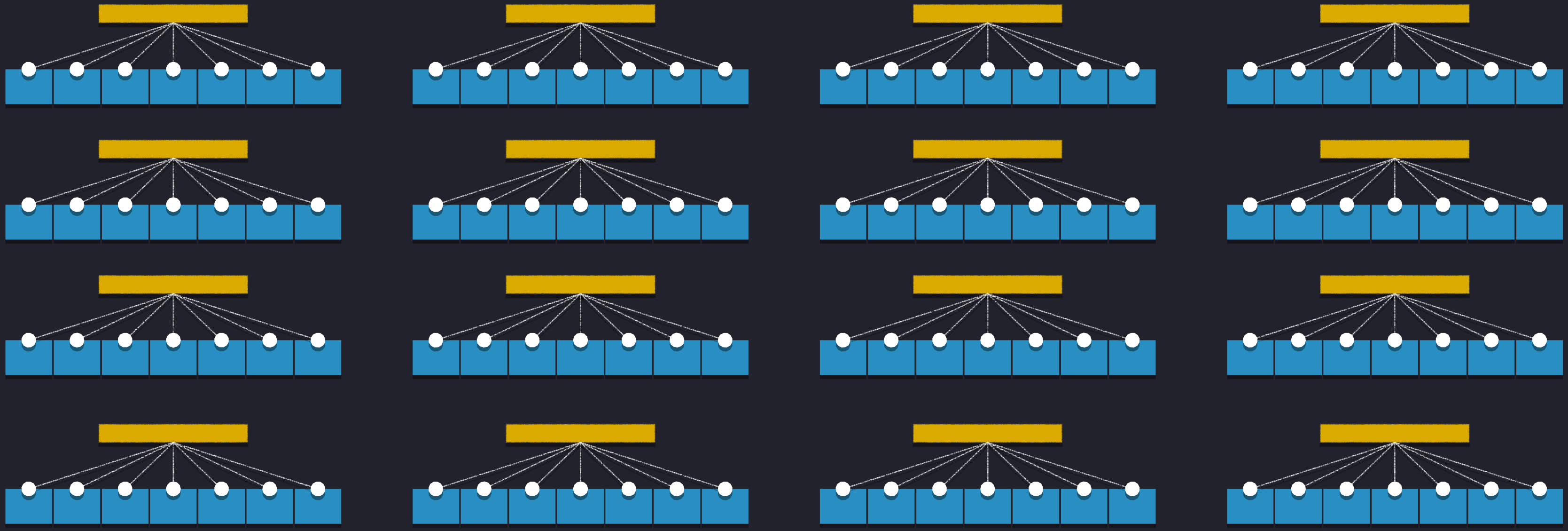
volume data model

link share data model

# Processing/ …for one query



# Processing/ 100K+ queries and rising

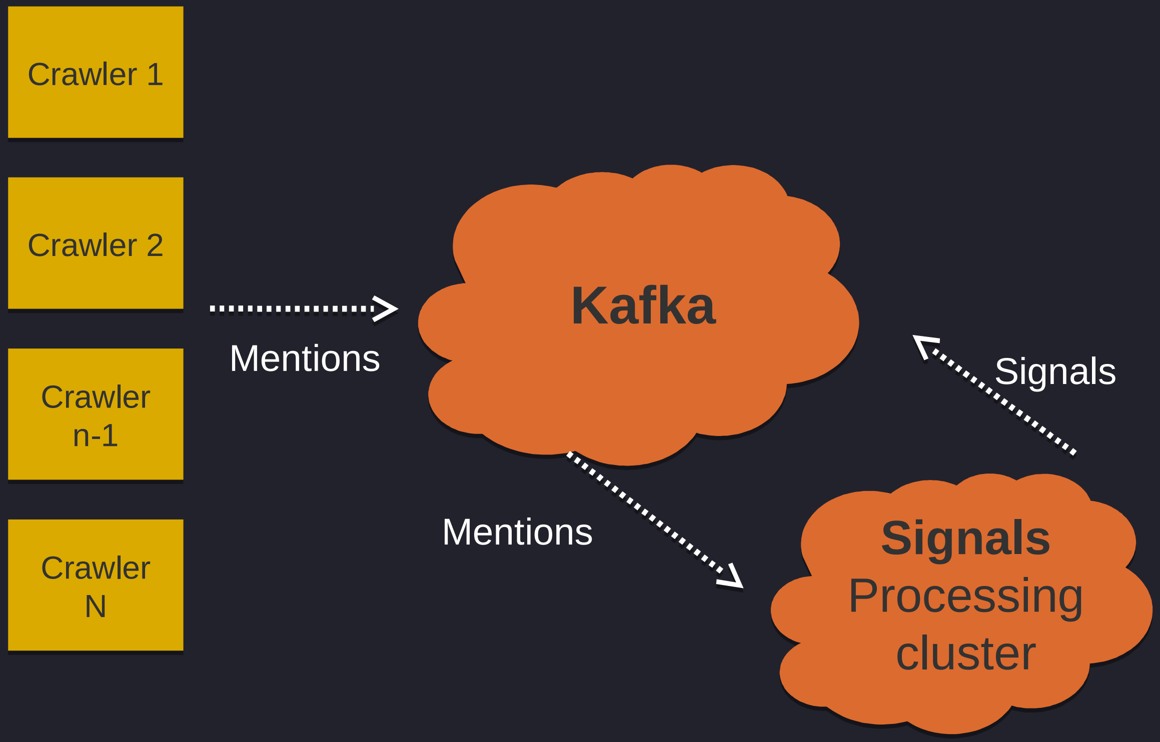


# Processing/ We need more JVMs

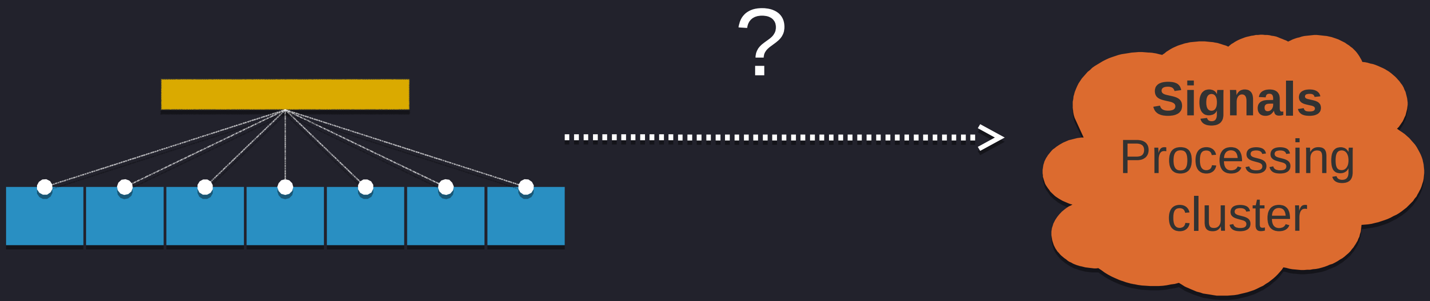
But how do we share the workload?

# Distribution of work

# Step 2.2/ A cluster of processing JVMs



# Distribution/ An atomic unit of work



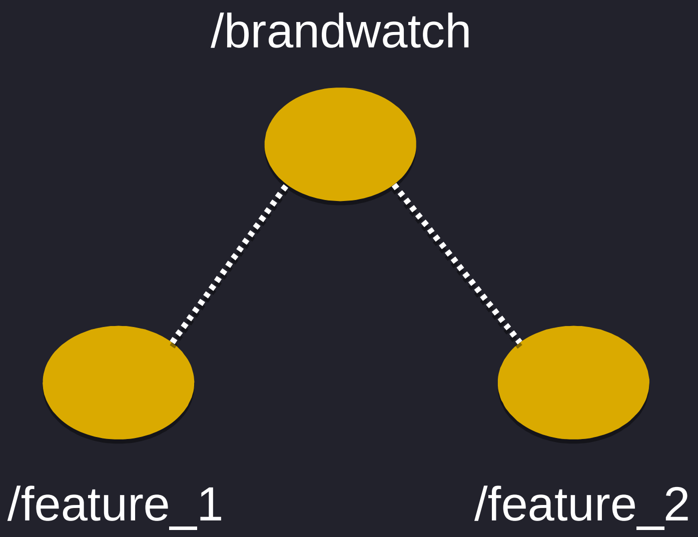
# Distribution/ Leader election

A way of deciding who is the leader for a task in a group of distributed nodes

# Distribution/ Zookeeper

A way of coordinating and managing distributed applications

# Zookeeper/ It’s like a file system

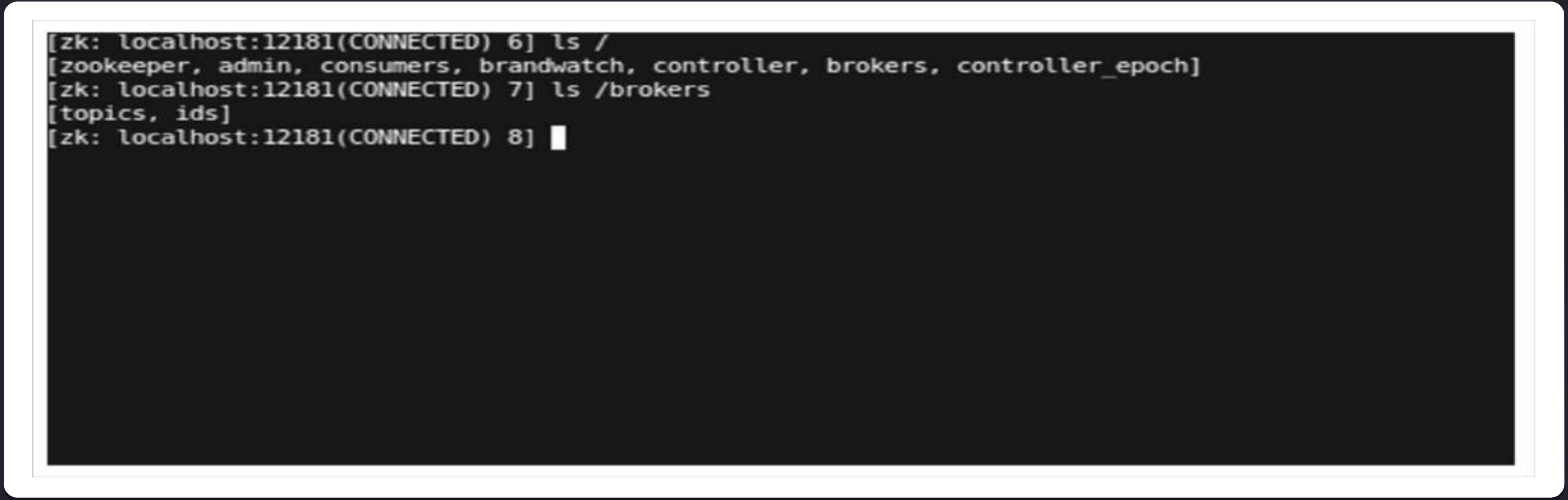


/brandwatch

/feature_1

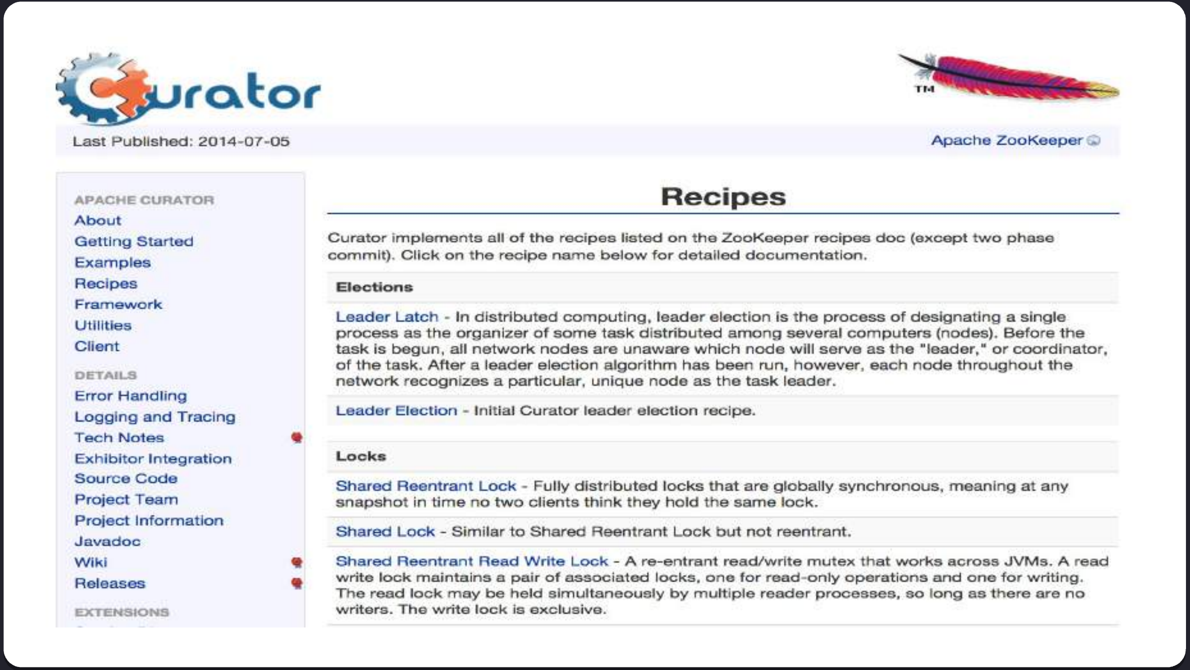
/feature_2

# Zookeeper/ At the command line



```
[zk: localhost:12181(CONNECTED) 6] ls /
[zookeeper, admin, consumers, brandwatch, controller, brokers, controller_epoch]
[zk: localhost:12181(CONNECTED) 7] ls /brokers
[topics, ids]
[zk: localhost:12181(CONNECTED) 8] 
```

# Distribution/ Recipes



## Recipes

Curator implements all of the recipes listed on the ZooKeeper recipes doc (except two phase commit). Click on the recipe name below for detailed documentation.

### Elections

Leader Latch - In distributed computing, leader election is the process of designating a single process as the organizer of some task distributed among several computers (nodes). Before the task is begun, all network nodes are unaware which node will serve as the "leader," or coordinator, of the task. After a leader election algorithm has been run, however, each node throughout the network recognizes a particular, unique node as the task leader.

Leader Election - Initial Curator leader election recipe.

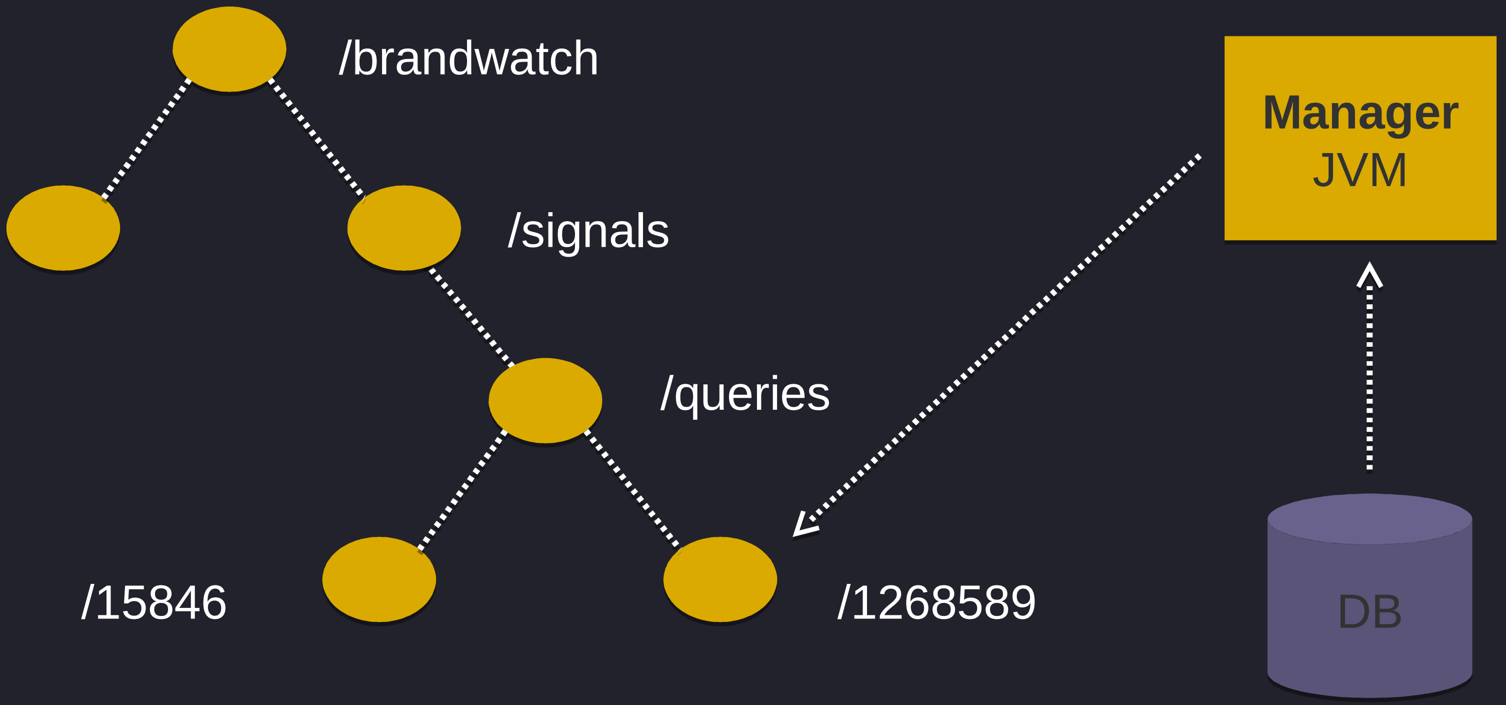
### Locks

Shared Reentrant Lock - Fully distributed locks that are globally synchronous, meaning at any snapshot in time no two clients think they hold the same lock.

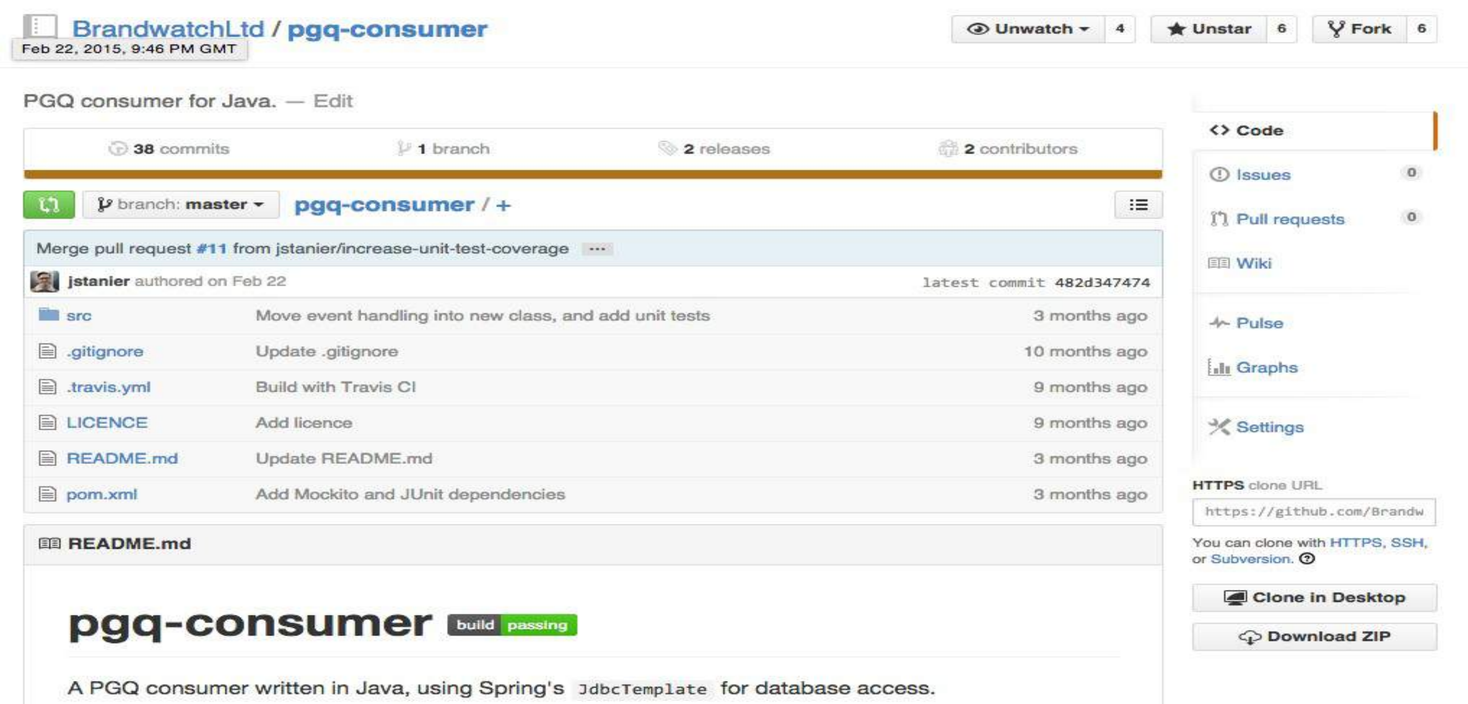
Shared Lock - Similar to Shared Reentrant Lock but not reentrant.

Shared Reentrant Read Write Lock - A re-entrant read/write mutex that works across JVMs. A read write lock maintains a pair of associated locks, one for read-only operations and one for writing. The read lock may be held simultaneously by multiple reader processes, so long as there are no writers. The write lock is exclusive.

# Distribution/ Offering jobs



# Distribution/ PGQ



BrandwatchLtd / pgq-consumer

Feb 22, 2015, 9:46 PM GMT

PGQ consumer for Java. — Edit

38 commits 1 branch 2 releases 2 contributors

branch: master pgq-consumer / +

Merge pull request #11 from jstanier/increase-unit-test-coverage …

jstanier authored on Feb 22 latest commit 482d347474

| | | |
|---|---|---|
| src | Move event handling into new class, and add unit tests | 3 months ago |
| .gitignore | Update .gitignore | 10 months ago |
| .travis.yml | Build with Travis CI | 9 months ago |
| LICENCE | Add licence | 9 months ago |
| README.md | Update README.md | 3 months ago |
| pom.xml | Add Mockito and JUnit dependencies | 3 months ago |

HTTPS clone URL

https://github.com/Brandw

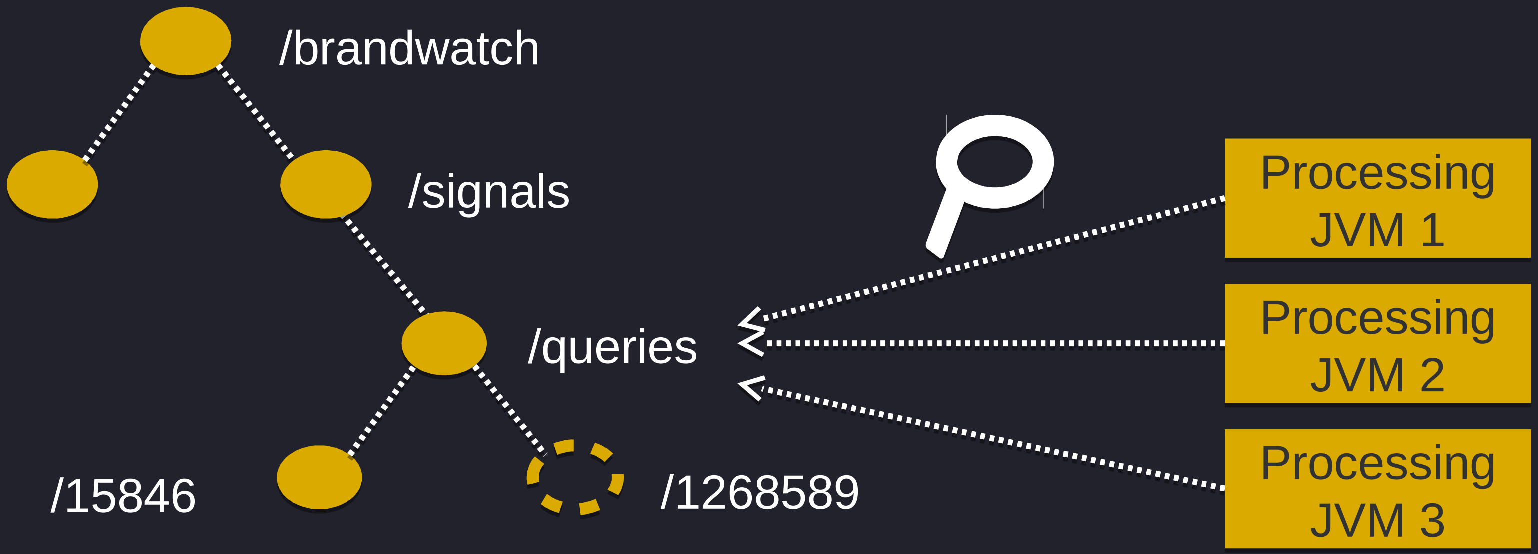
You can clone with HTTPS, SSH, or Subversion.

README.md

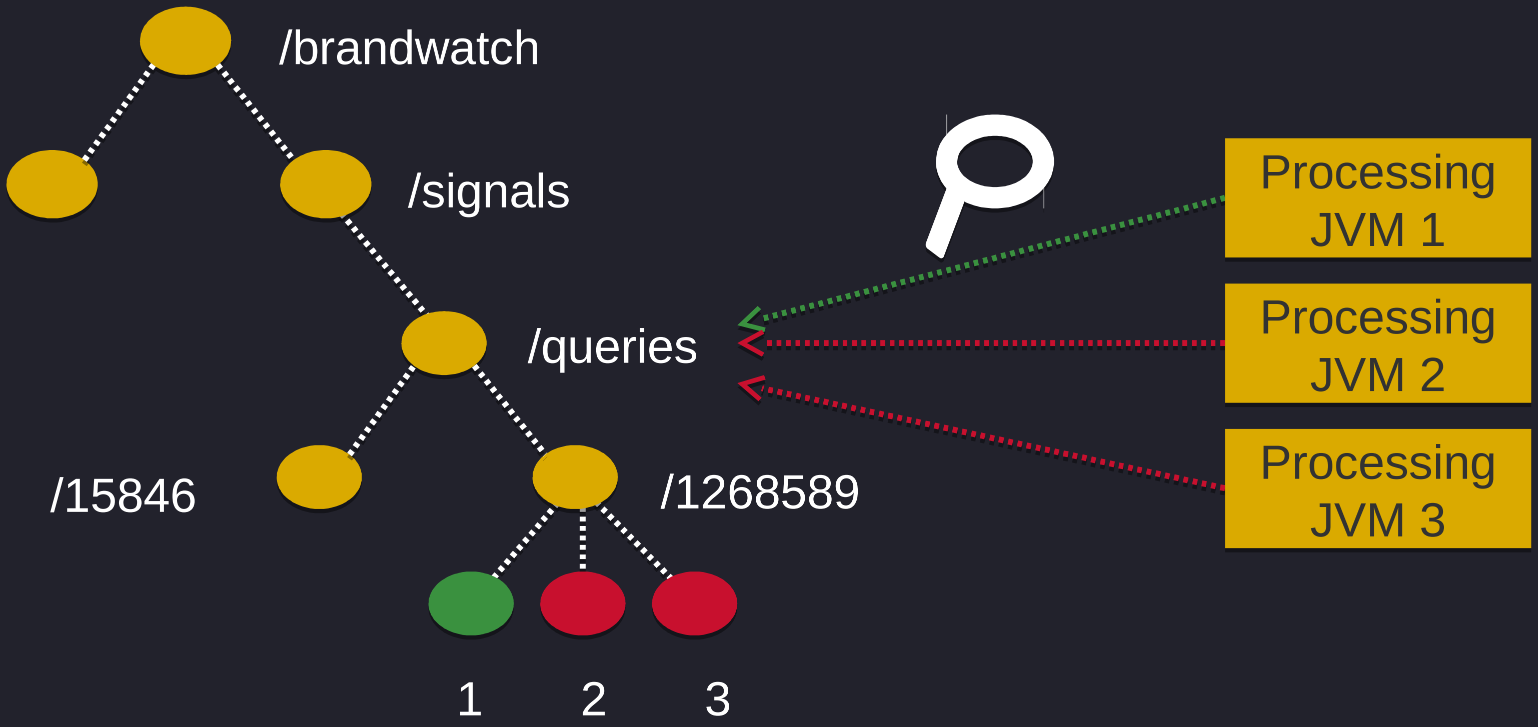
## pgq-consumer

A PGQ consumer written in Java, using Spring's `JdbcTemplate` for database access.

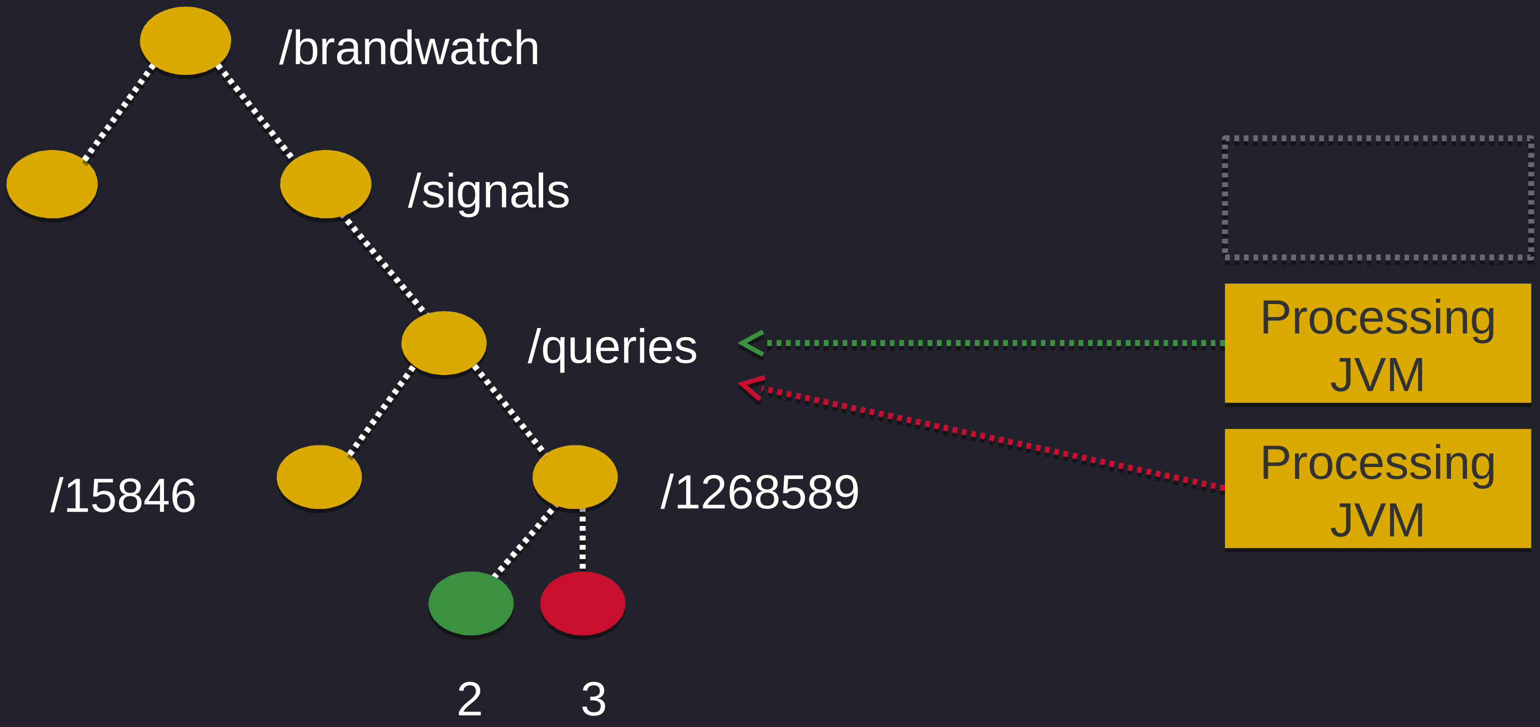
# Distribution/ Leader election 101



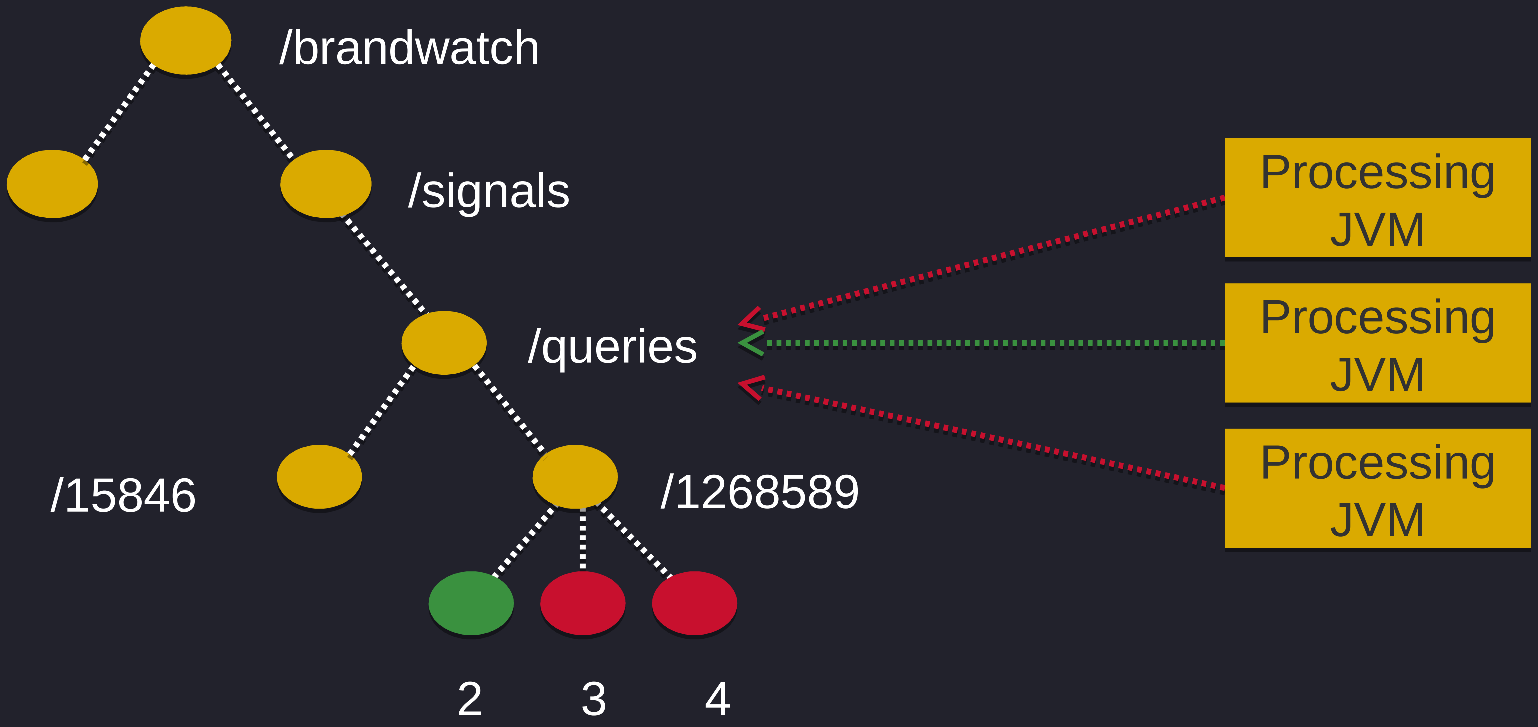
# Distribution/ Leader election 101



# Distribution/ The leader dies



# Distribution/ The dead rises again



# Distribution/ That’s how we do it

Each time someone turns on the feature, we leader elect for processing

# Distribution/ Almost there?

We are processing long running jobs

What about workers getting overloaded?

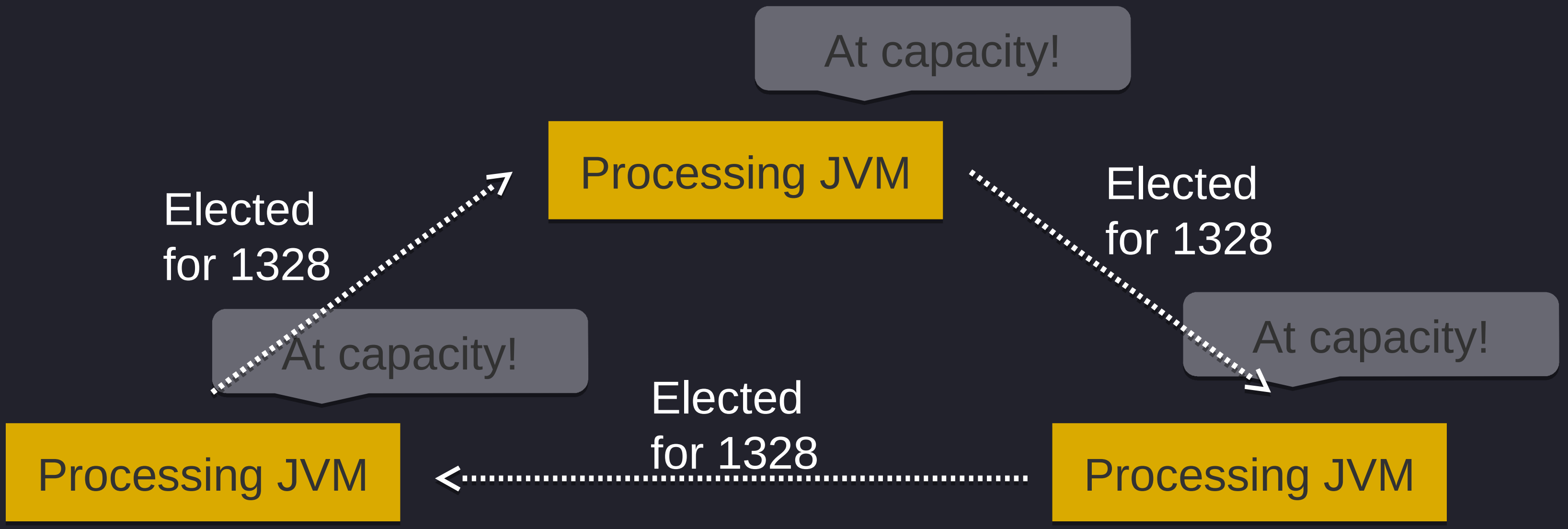
# Distribution/ After leader election

1. Take leadership
2. Hit max queries?
   a. No – go to 3
   b. Yes – give up leadership, try again
3. Start working

# Distribution/ Now we’re almost there?

Actually, no…

# Distribution/ Infinite election



At capacity!

Processing JVM

Elected for 1328

Elected for 1328

At capacity!

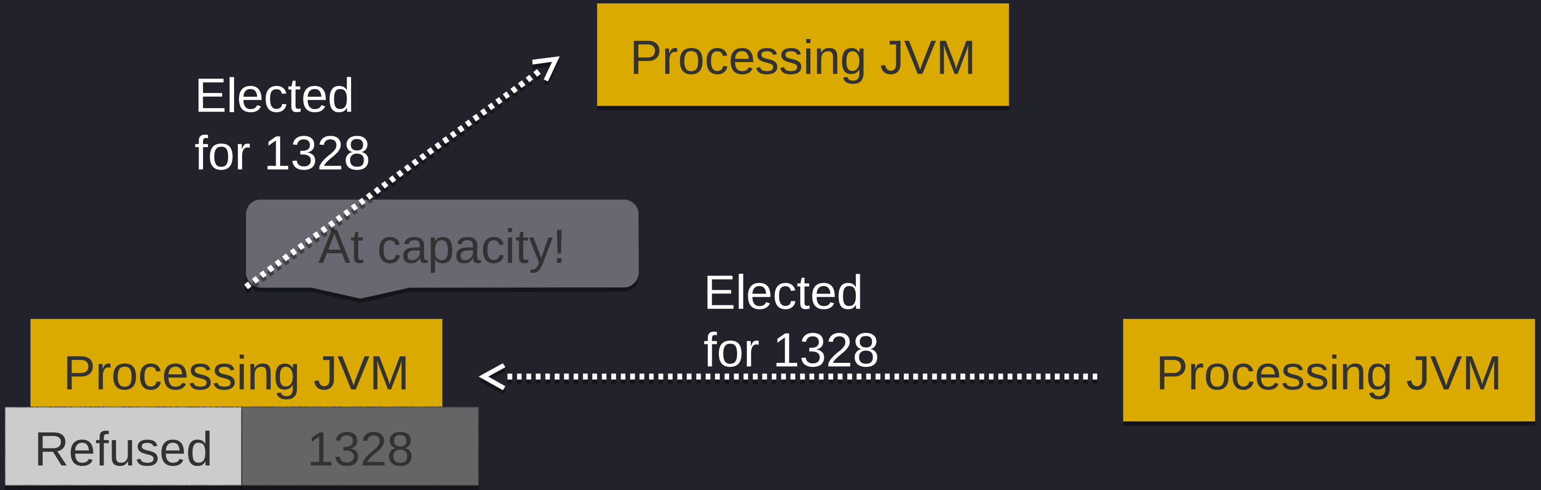
At capacity!

Elected for 1328

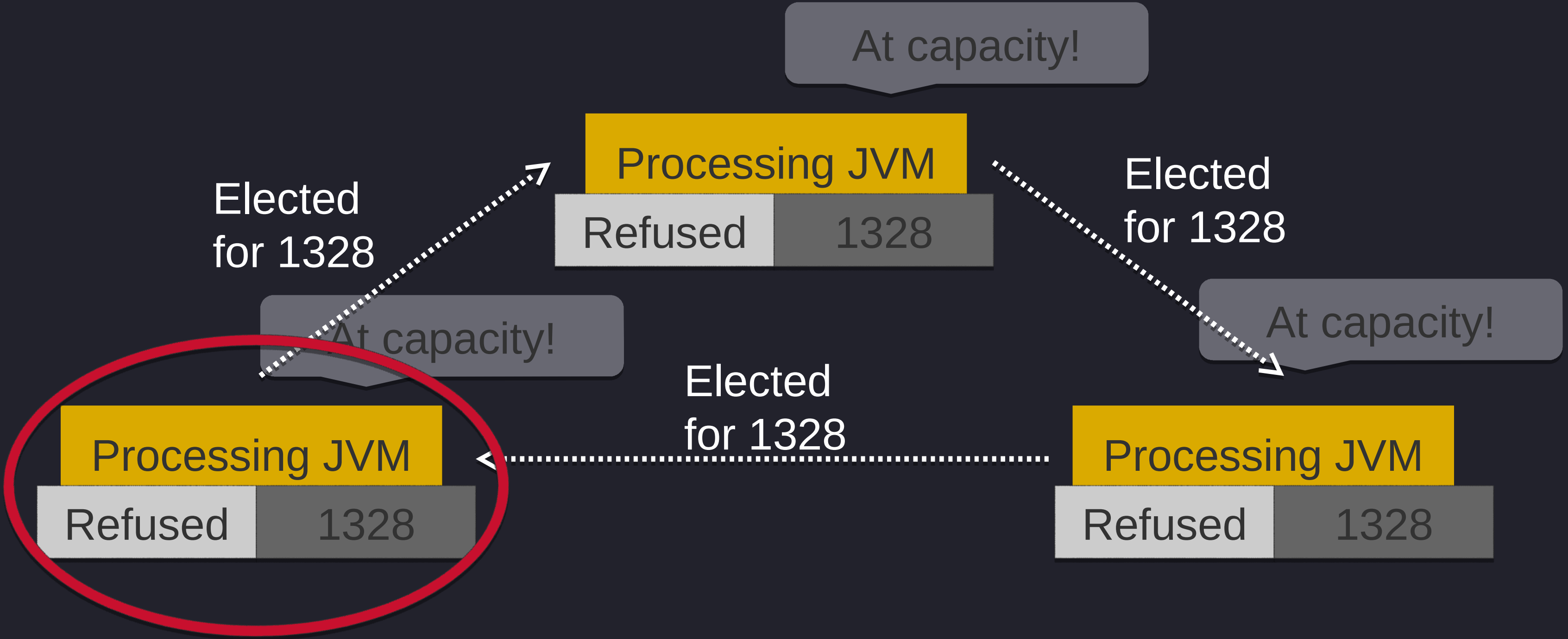
Processing JVM

Processing JVM

# Distribution/ Solution



# Distribution/ Solution



At capacity!

Processing JVM

| Refused | 1328 |
|---|---|

Elected for 1328

Elected for 1328

At capacity!

At capacity!

Processing JVM

| Refused | 1328 |
|---|---|

Elected for 1328

Processing JVM

| Refused | 1328 |
|---|---|

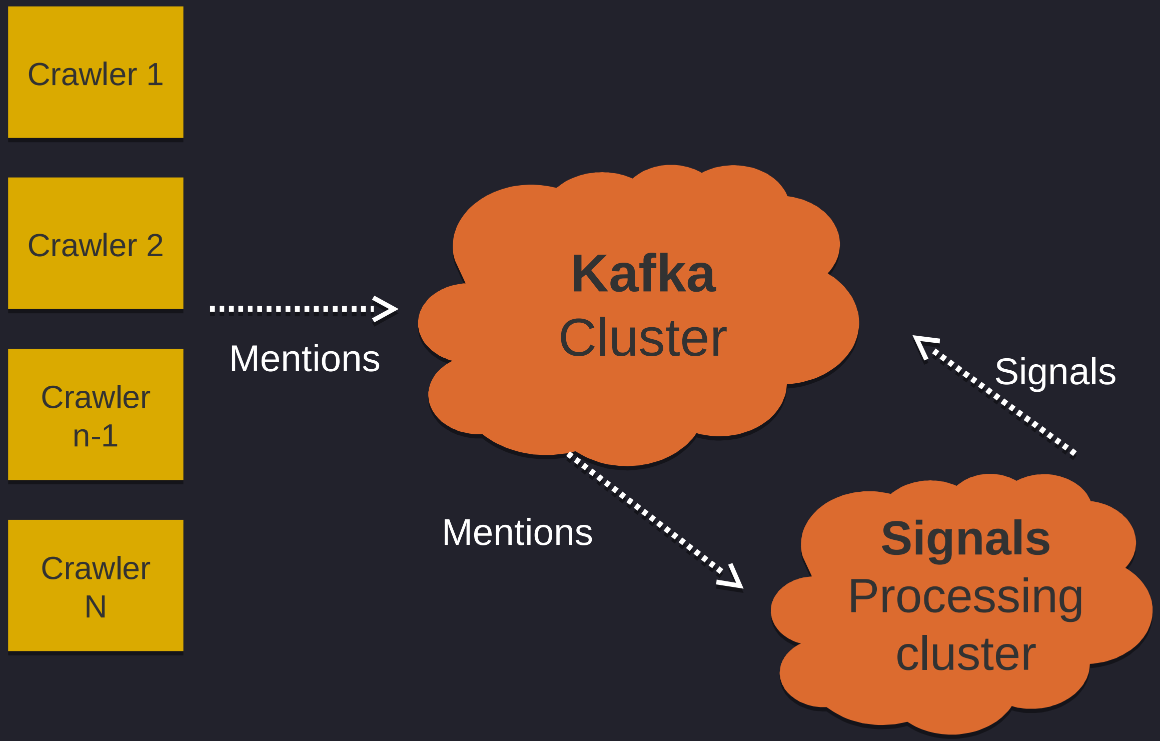
State

# State/ Snapshotting of worker data

If one worker dies, we want the other to pick up where it left off

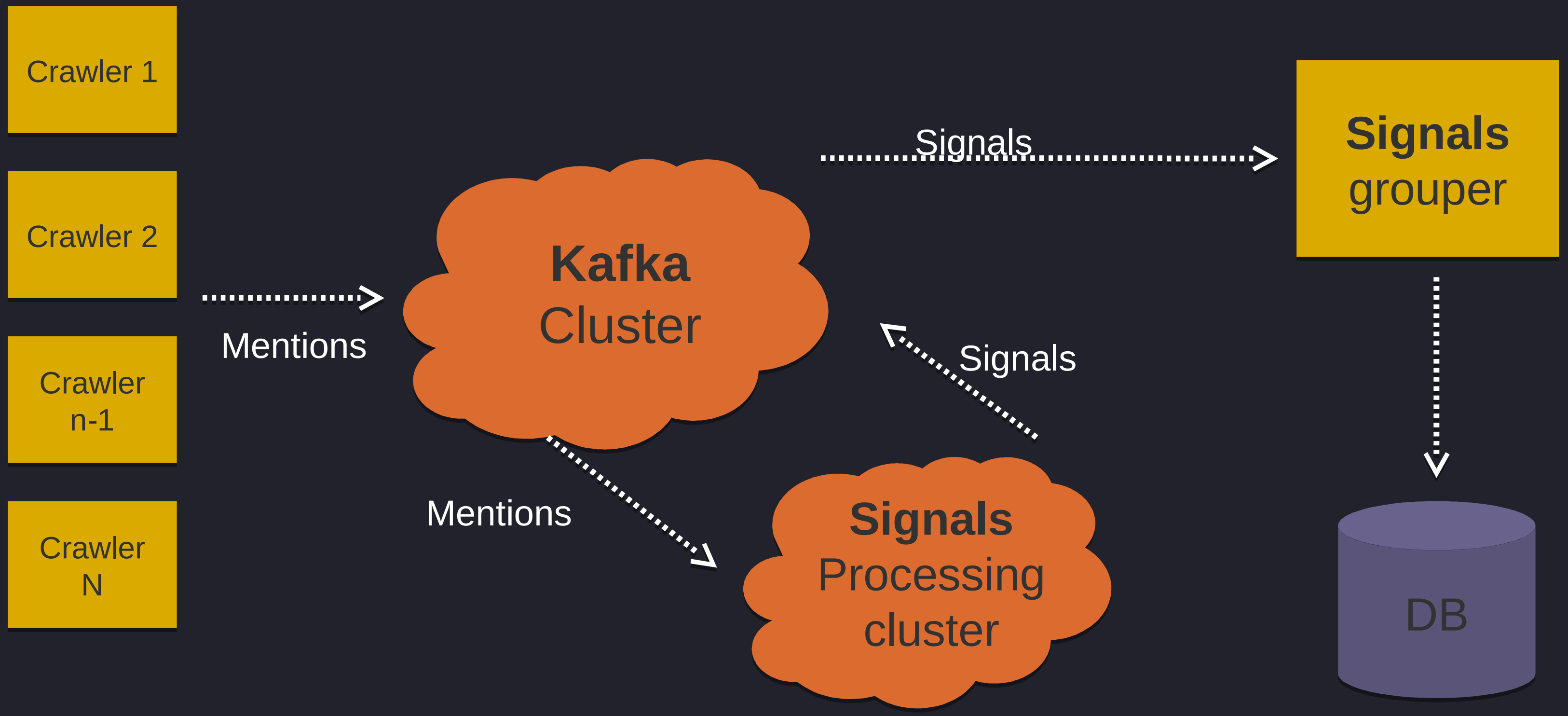
Regular snapshotting to HBase

# Step 2.2/ Done!

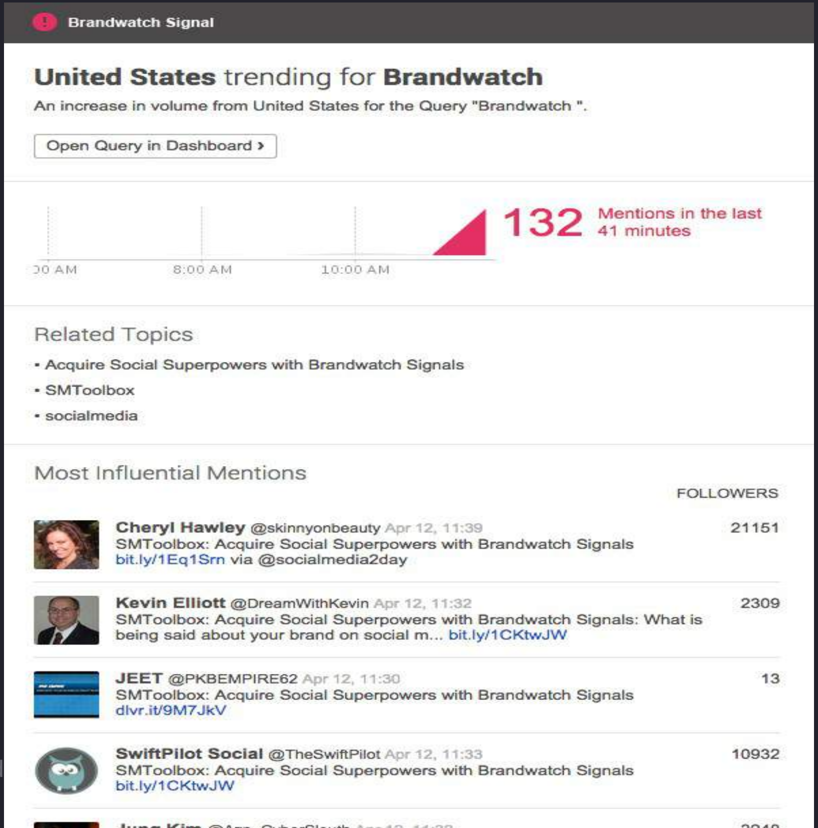

Crawler 1
Crawler 2
Crawler n-1
Crawler N
Mentions
Kafka Cluster
Signals
Mentions
Signals Processing cluster

# Finding meaning

# Step 3/ Meaning



# Meaning/ Desired outcome



Brandwatch Signal

**United States** trending for **Brandwatch**

An increase in volume from United States for the Query "Brandwatch ".

Open Query in Dashboard ›

132 Mentions in the last 41 minutes

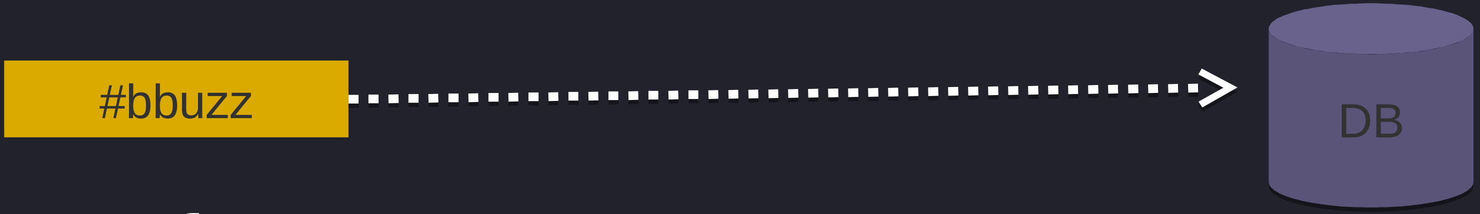
Related Topics

- Acquire Social Superpowers with Brandwatch Signals
- SMToolbox
- socialmedia

Most Influential Mentions

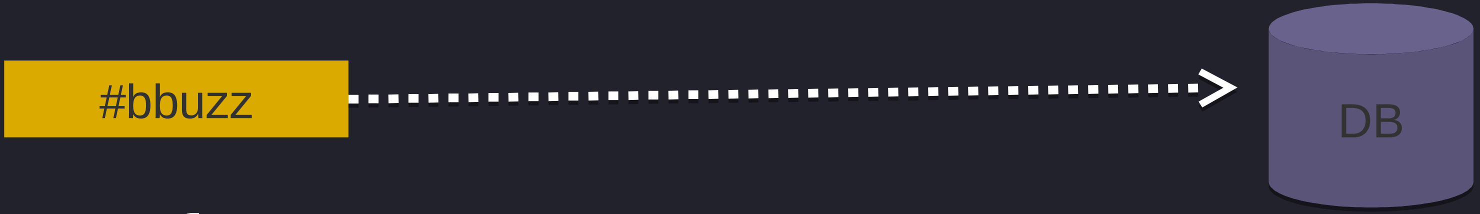
| | FOLLOWERS |
|---|---|
| **Cheryl Hawley** @skinnyonbeauty Apr 12, 11:39<br>SMToolbox: Acquire Social Superpowers with Brandwatch Signals bit.ly/1Eq1Srn via @socialmedia2day | 21151 |
| **Kevin Elliott** @DreamWithKevin Apr 12, 11:32<br>SMToolbox: Acquire Social Superpowers with Brandwatch Signals: What is being said about your brand on social m... bit.ly/1CKtwJW | 2309 |
| **JEET** @PKBEMPIRE62 Apr 12, 11:30<br>SMToolbox: Acquire Social Superpowers with Brandwatch Signals dlvr.it/9M7JkV | 13 |
| **SwiftPilot Social** @TheSwiftPilot Apr 12, 11:33<br>SMToolbox: Acquire Social Superpowers with Brandwatch Signals bit.ly/1CKtwJW | 10932 |

# Meaning/ 1. Topics



{ “Hey @jstanier Your talk sucks! #bbuzz”,
“I love #bbuzz”,
“Where’s the Club Mate? #bbuzz”,
… ,
“I have consumed ALL THE FREE COFFEE!
#bbuzz” }

# Meaning/ 1. Topics

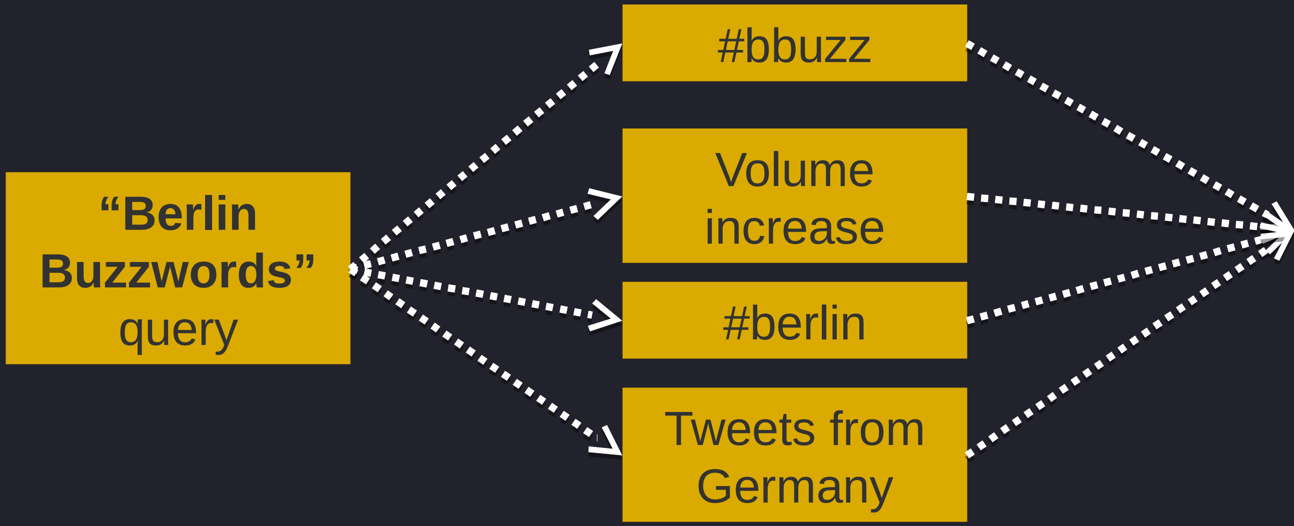


{

Berlin Buzzwords
ElasticSearch
Scaling

}

# Meaning/ 2. Grouping events



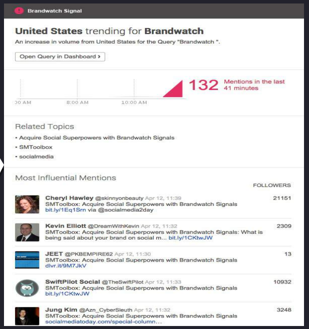
**“Berlin Buzzwords”** query

#bbuzz

Volume increase

#berlin

Tweets from Germany



Brandwatch Signal

**United States** trending for **Brandwatch**

An increase in volume from United States for the Query "Brandwatch ".

Open Query in Dashboard ›

132 Mentions in the last 41 minutes

Related Topics

- Acquire Social Superpowers with Brandwatch Signals
- SMToolbox
- socialmedia

Most Influential Mentions

FOLLOWERS

**Cheryl Hawley** @skinnyonbeauty Apr 12, 11:39 21151
SMToolbox: Acquire Social Superpowers with Brandwatch Signals bit.ly/1Eq1Srn via @socialmedia2day

**Kevin Elliott** @DreamWithKevin Apr 12, 11:32 2309
SMToolbox: Acquire Social Superpowers with Brandwatch Signals: What is being said about your brand on social m... bit.ly/1CKtwJW

**JEET** @PKBEMPIRE62 Apr 12, 11:30 13
SMToolbox: Acquire Social Superpowers with Brandwatch Signals dlvr.it/9M7JkV

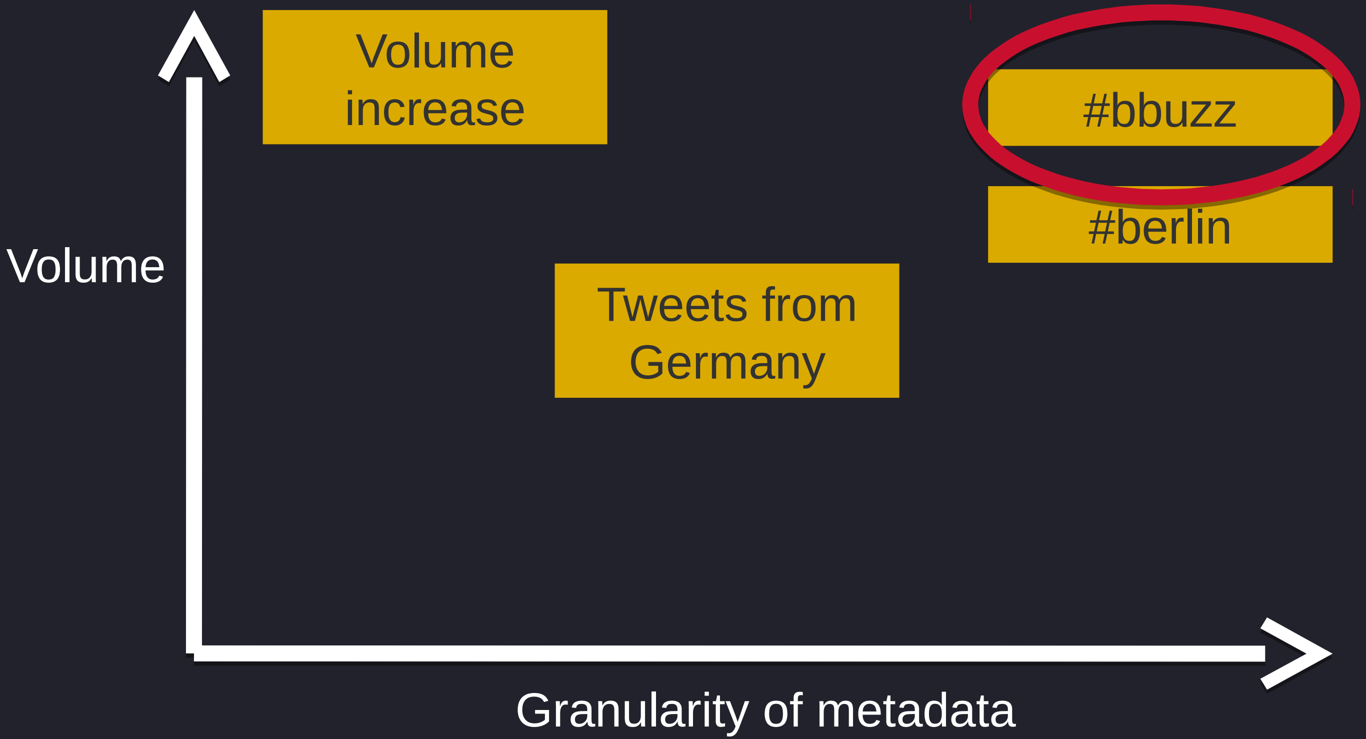
**SwiftPilot Social** @TheSwiftPilot Apr 12, 11:33 10932
SMToolbox: Acquire Social Superpowers with Brandwatch Signals bit.ly/1CKtwJW

**Jung Kim** @Azn_CyberSleuth Apr 12, 11:32 3248
SMToolbox: Acquire Social Superpowers with Brandwatch Signals socialmediatoday.com/special-column…

# Meaning/ 2. Grouping events

- Granularity
- Text similarity
- Shape of volume sparkline

# Meaning/ 2. Grouping events



Volume increase

#bbuzz

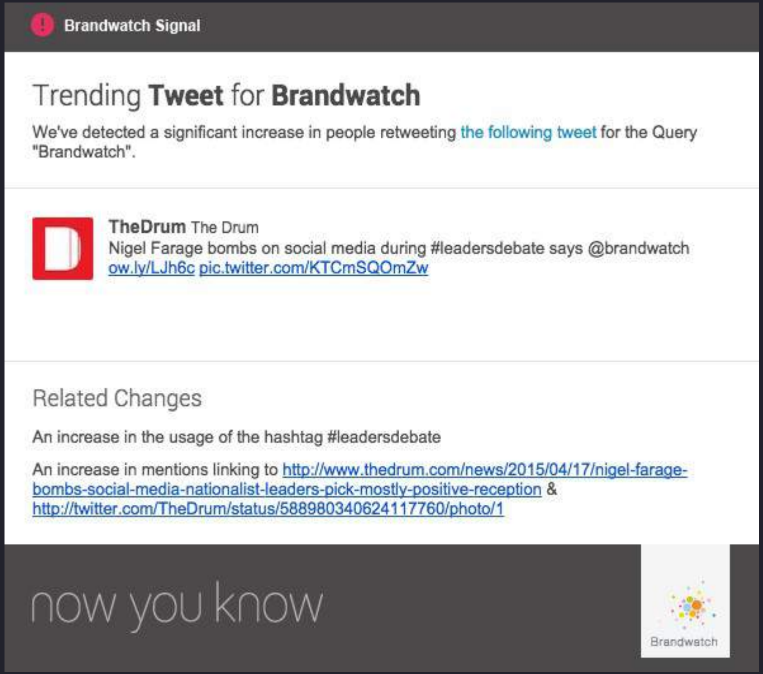
#berlin

Tweets from Germany

Volume

Granularity of metadata

# Meaning/ An example



Brandwatch Signal

Trending **Tweet** for **Brandwatch**

We've detected a significant increase in people retweeting the following tweet for the Query "Brandwatch".

**TheDrum** The Drum
Nigel Farage bombs on social media during #leadersdebate says @brandwatch ow.ly/LJh6c pic.twitter.com/KTCmSQOmZw

Related Changes

An increase in the usage of the hashtag #leadersdebate

An increase in mentions linking to http://www.thedrum.com/news/2015/04/17/nigel-farage-bombs-social-media-nationalist-leaders-pick-mostly-positive-reception & http://twitter.com/TheDrum/status/588980340624117760/photo/1

# Closing remarks

# Say hello/

@jstanier

# Q&A