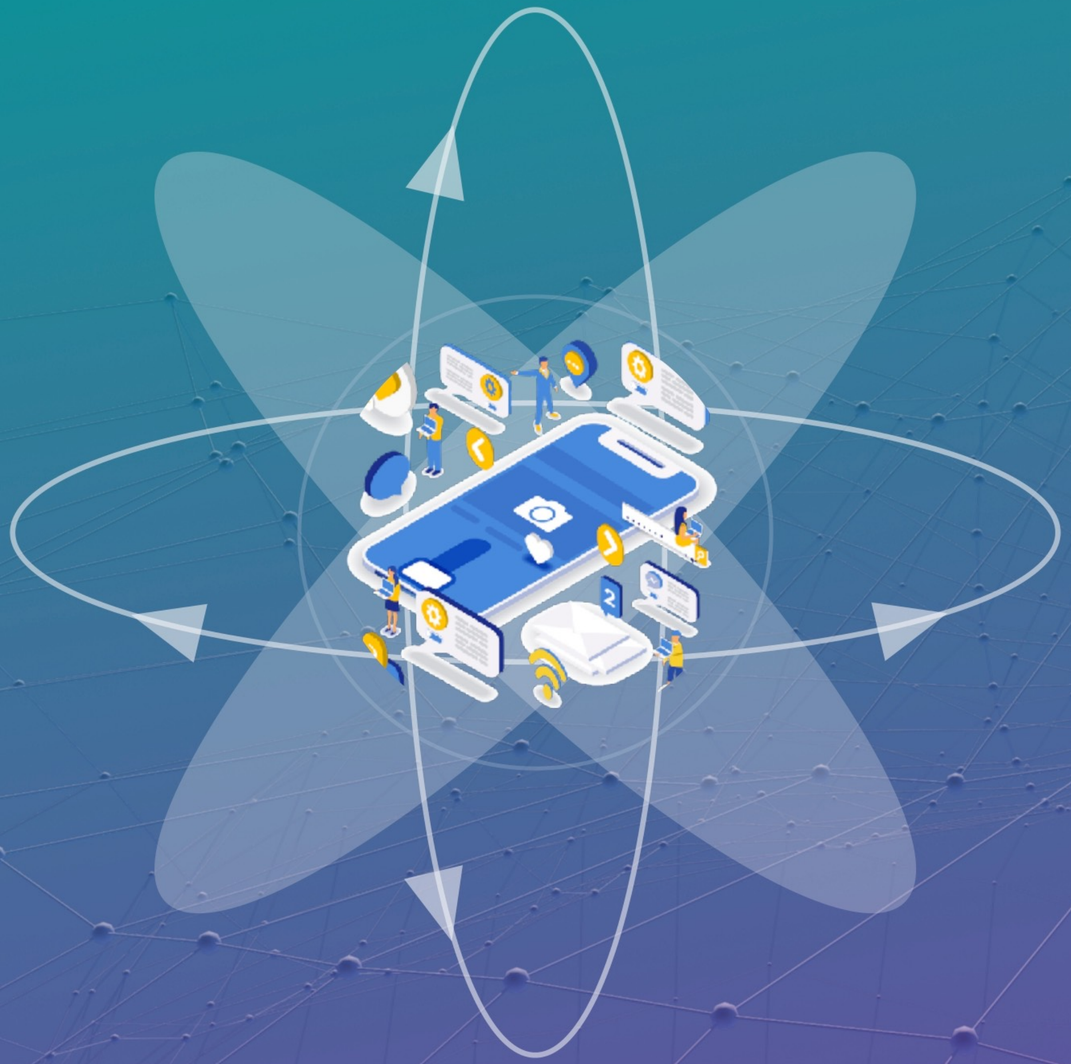


Big Data Platform-as-a-Service for Cross-Media Monitoring

LIANA NAPALKOVA

PhD, SENIOR DATA SCIENTIST

EURECAT

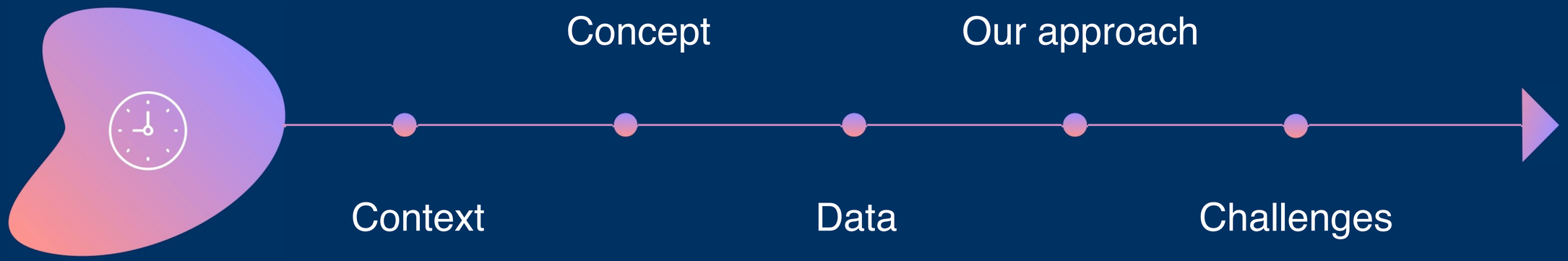
JUAN CARLOS CASTRO

DIGITAL PRODUCT MANAGER

EURECAT

DataEngConf - Barcelona, 2018

Agenda



Context

There are a lot of social media monitoring platforms.

BUT…

Classical Social Media Monitoring Platforms

1. **CONTENT ANALYSIS**
2. **TEXT MINING**
3. **NATURAL LANGUAGE PROCESSING**

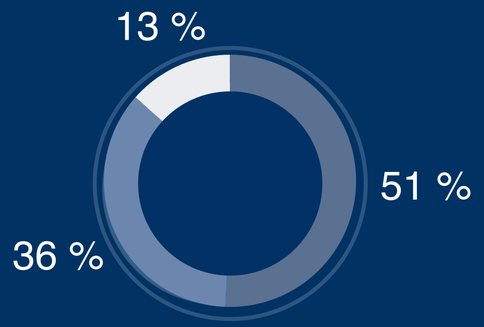
Sample questions answered:

- *What is the number and percent of users who mention my brand on Facebook?*
- *Which is users' attitude towards my products?*

Classical Social Media Monitoring Platforms

Example

NUMBER OF MENTIONS



51% (102K) Twitter
36% (72K) Facebook
13% (26K) Blogs

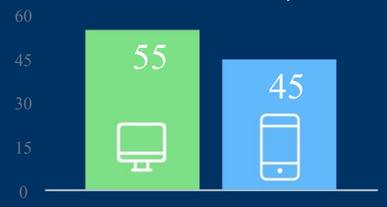
GENDER

48% Male
52% Female

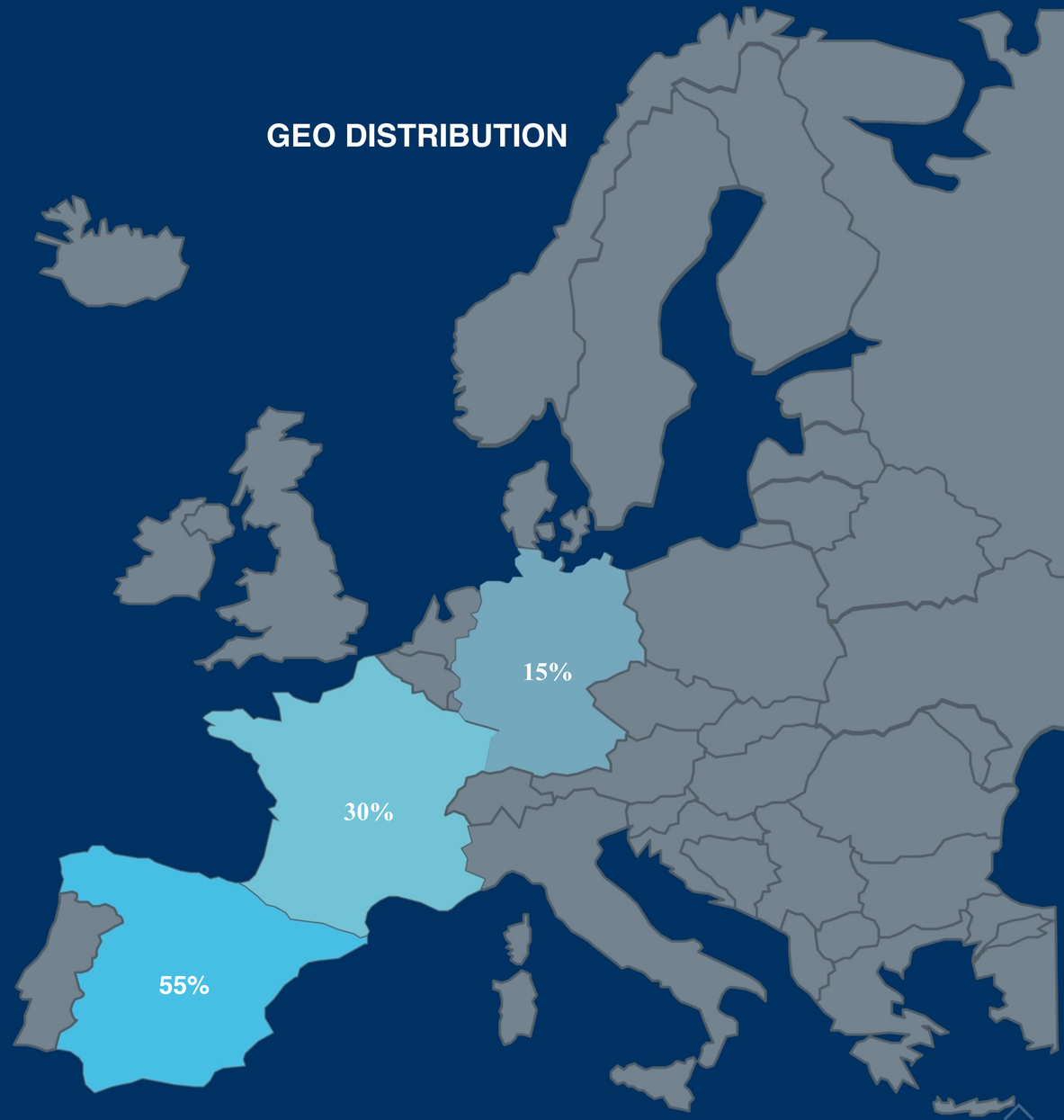
OS

45% Android
55% Linux

DEVICE TYPE, %



GEO DISTRIBUTION



BUT WHAT IF WE WANT TO DIVE DEEPER?

Let’s take a look at MWC'2015

TyN MAGAZINE SECCIONES VIDEO

PORTADA » EMPRESAS / DESTACADA 05/03/2015

Tras la tercera jornada, decrece el interés en Twitter sobre el MWC2015

Ya pasados los lanzamientos fuertes, la cantidad de posteos se redujo considerablemente: de 94.589 registrados en el segundo día hasta los 81.358 del tercero. Barcelona se mantuvo como la ciudad donde se originaron más tweets sobre el evento.

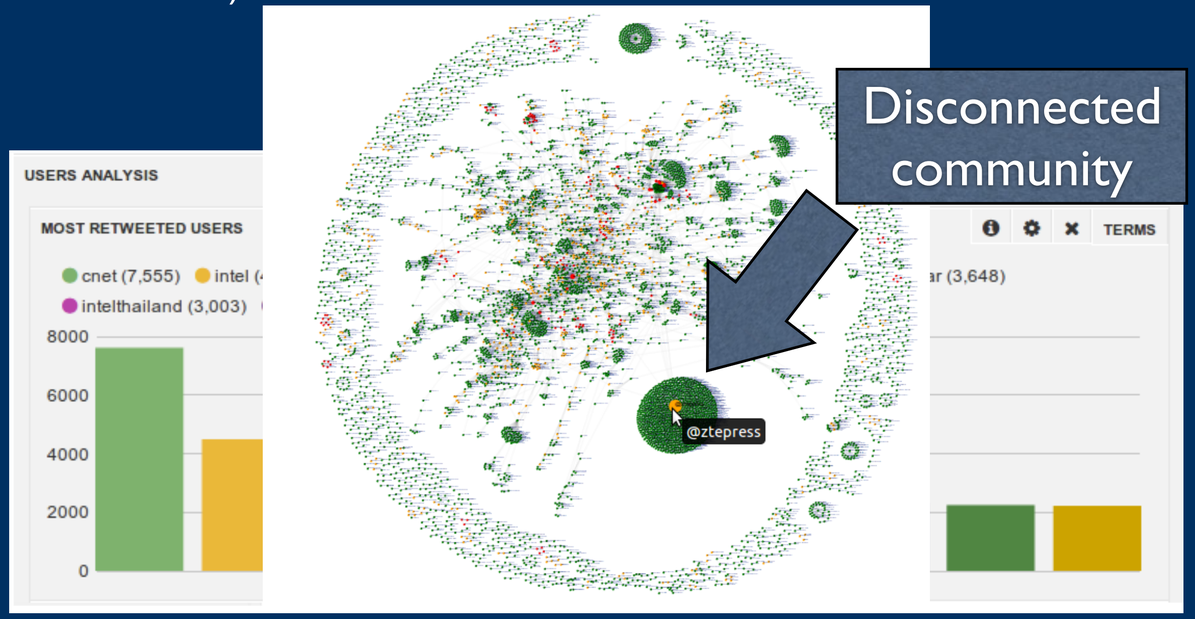
La tercera jornada del Mobile World Congress **mostró una reducción en el interés de los usuarios por las marcas en sí**: los tres primeros hashtags más importantes estuvieron relacionados con el evento: #MWC15 (27.526 impresiones), #MWC2015 (3.316) y #mwc15 (1,778). Los datos fueron proporcionados por la herramienta de monitoreo en redes sociales QSocialNow.

Recién en **el cuarto puesto apareció una marca: Intel, con 1.708 menciones**. La firma china **ZTE es la segunda que apareció, entre los puestos 6 y 8**, con los hashtags #ZTE (1.119), #ZTEOverseas (1.062) y #ZTELightUptheFuture (1.061). Por su parte, **la cuenta de twitter más leída fue @Mashable (propiedad del blog especializado del mismo nombre) que regitró 5.060.707 personas**. El podio lo completaron @Intel (4.044.603) y @el_pais (3.985.654).

Mobile World Congress 2015

ZTE Corporation was **one of the most relevant brands** on Twitter buzz…

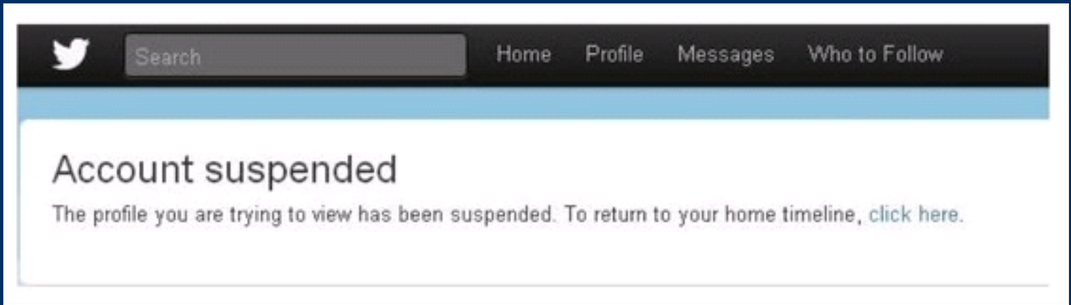
ZTE Corporation was mostly poked by a huge cluster of **users who did not belong to the giant component of the graph** (composed of the organisers of the event, mobile providers and mass-media).

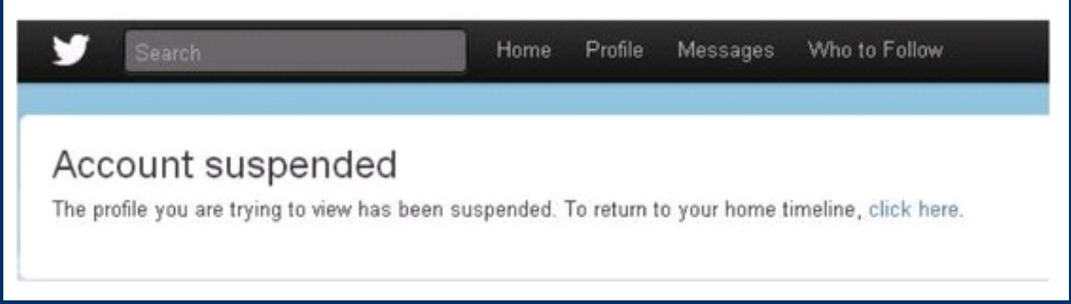


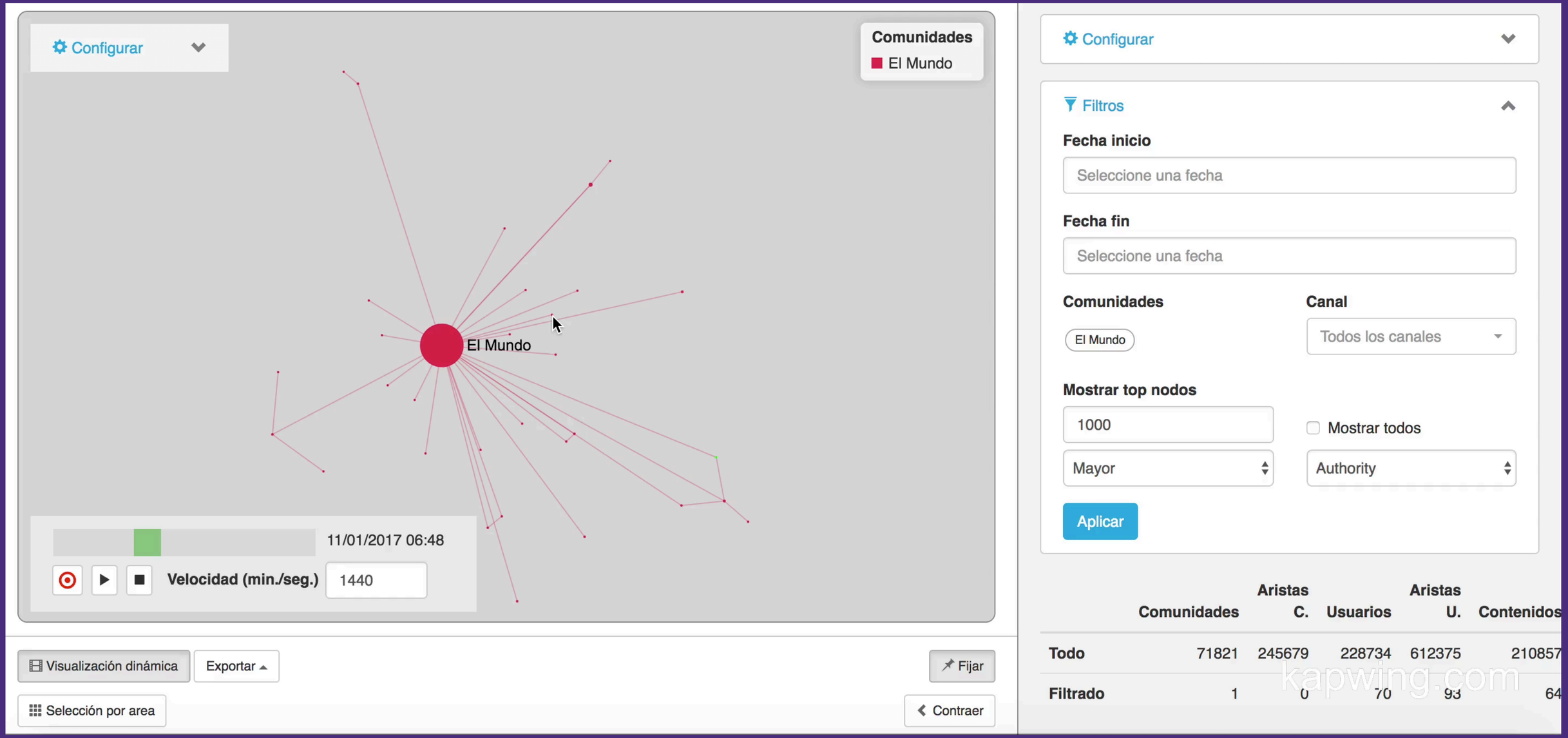
And the reason is…

The reason behind this unexpected scenario is that the ZTE retweet-graph community was formed by **BOTS**, a common practice hardly detected by most SMM industrial tools.

Each tweet posted by @ztepress was retweeted by 1000 users who were finally suspended.



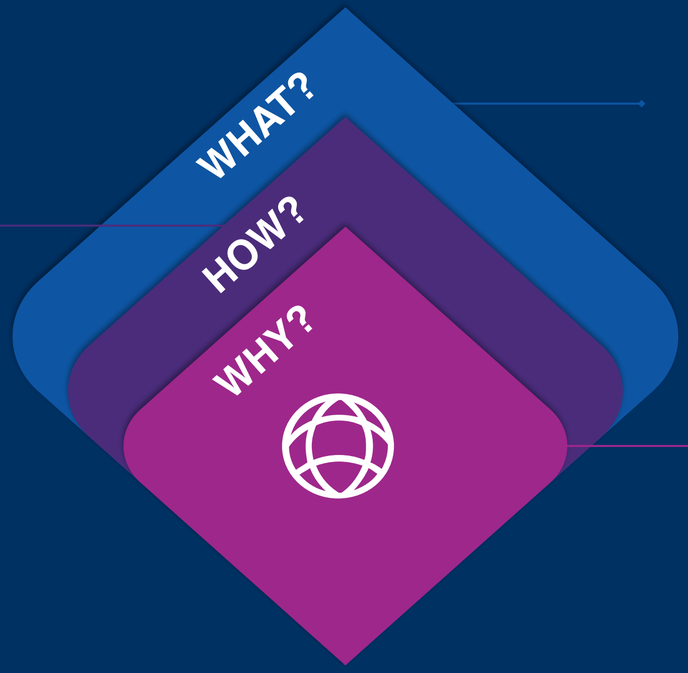




	Comunidades	Aristas C.	Usuarios	Aristas U.	Contenidos
Todo	71821	245679	228734	612375	210857
Filtrado	1	0	70	93	64

KALIUM: Let’s put it on the “Why” axis

How does the ecosystem behave over time?



What is the impact of the most frequent content?

What a certain community is talking about?

Why users of a certain community talk about particular topic?

KALIUM: Comparison to other tools

Gephi

Scales to large networks **but is limited** by the amount of memory allocated to it in JVM.

SNAP - Stanford Network Analysis Platform

Scales to massive networks with billions of edges **on a single big memory machine**.

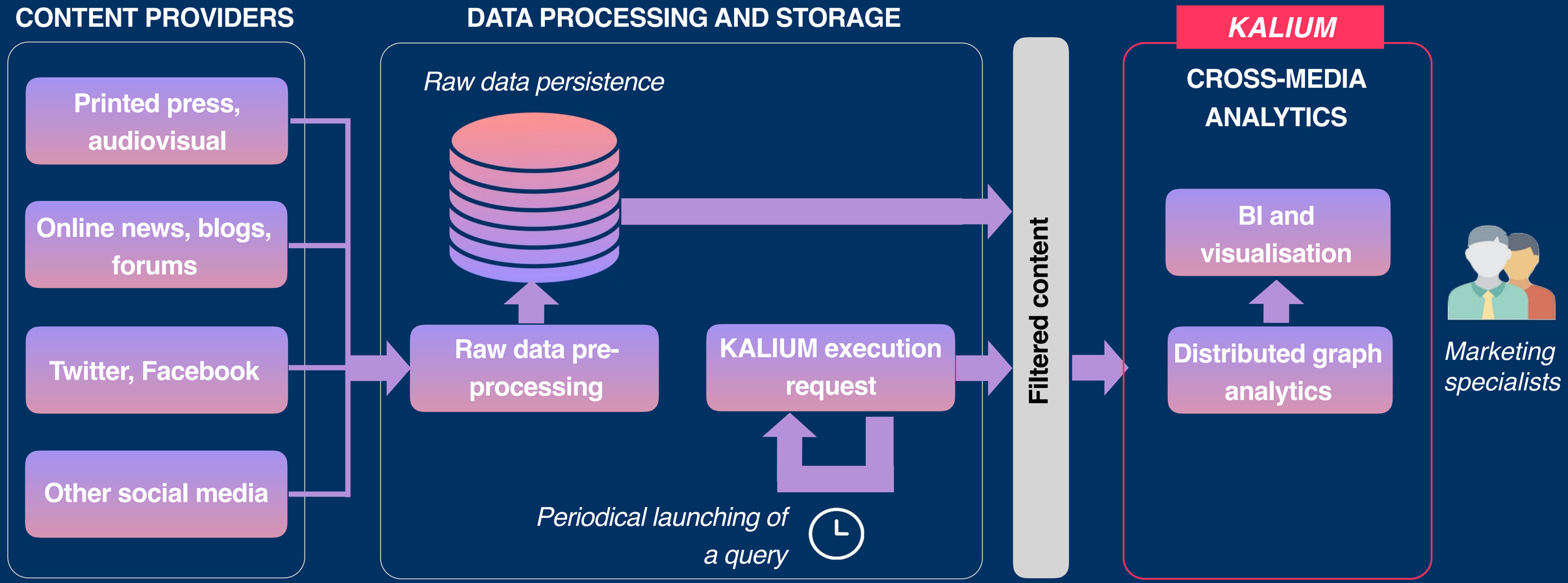
KALIUM

Scalability: distributed environment instead of a single machine

Visualisation of sub-graphs including video generation

Customisation of graph generation and algorithms

Concept


CONTENT PROVIDERS
Printed press, audiovisual
Online news, blogs, forums
Twitter, Facebook
Other social media
DATA PROCESSING AND STORAGE
Raw data persistence
Raw data pre-processing
KALIUM execution request
Periodical launching of a query
Filtered content
KALIUM
CROSS-MEDIA ANALYTICS
BI and visualisation
Distributed graph analytics
Marketing specialists

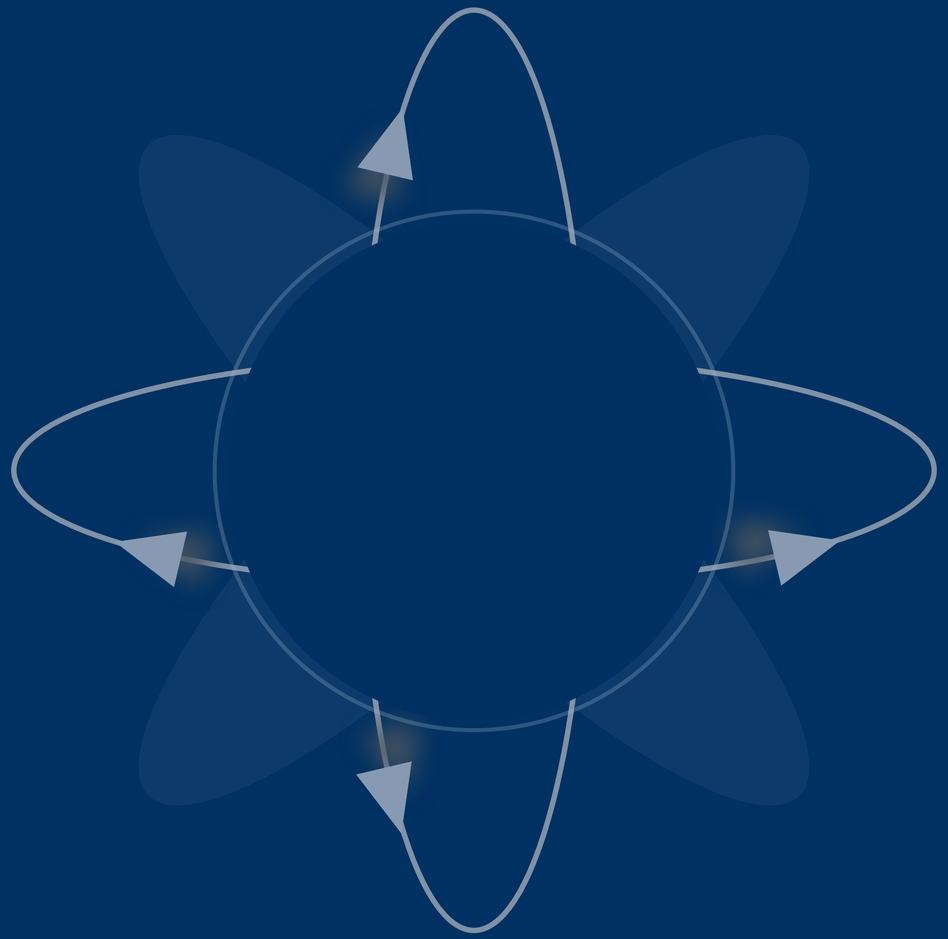
Input data

DATA VOLUME

KALIUM has been developed to deal with millions of records. The algorithms are scalable.

DATA FORMAT

Currently KALIUM works with raw data stored in Elasticsearch.



DATA DIVERSITY

Data may come from different sources which is defined the input query.

DATA VELOCITY

Our current assumption is that the graph generation queries arrive on a daily basis.

Challenges

CHALLENGE #1

How to flexibly create new graphs that better fit the goals of particular marketing analysis?

CHALLENGE #2

How to scale and optimise graph mining algorithms to handle large complex networks?

CHALLENGE #3

How to efficiently persist and query the graphs that represent large cross-media networks?

CHALLENGE #4

How to visualise large complex networks in an interpretable way?

Graph creation

CHALLENGE #1

How to flexibly create new graphs that better fit the goals of particular marketing analysis?

The query defines how to create a graph

INPUT QUERY

Example of a query


```
query_id":"59d9d4e8-05cf-11e8-ba89-0ed5f89f718b",
"elastic_query":{
  "elastic_uri": [
    {"host": "XXX.XX.XX.XX", "port": 9200}
  ],
  "elastic_index" : ["testindex/testitem"],
  "es_field_array_include": ["project","cluster","author","client","twitter_mentioned_user","hashtag","expan
},
"edges":{
  "attributes": {
    "edgeTimestampField": "publication_date",
    "audienceField": "audience",
    "contentUrlField": "url",
    "userUrlField": "source_url",
    "typologyField": "typology"
  },
  "implicit_rule":[
   {
      "nodeIdField": "remote_media_code",
      "nodeLabelField": "source_name",
      "nodeField": "",
      "nodeAttr":["typology", "media_type", " publishing_platform"],
      "reference": {
        "sourceField": "expanded_outbound_link",
        "targetField": ["url","source_url"],
        "referenceStrategy": "url"
      }
    }
  ],
  "explicit_rule":[
    {
      "sourceNodeIdField": "source_remote_media_code",
      "sourceNodeLabelField": "twitter_retweeted_user",
      "sourceNodeField": "twitter_retweeted_user_id",
      "nodeAttr":["typology", "media_type", " publishing_platform"],
      "targetNodeIdField": "remote_media_code",
      "targetNodeLabelField": "twitter_user",
      "targetNodeField": "twitter_user_id",
      "targetNodeAttr": ""
    }
  ],
  "cluster_rule": [
    {
      "nodeIdField": "remote_media_code",
```


Explicit rule

Implicit rule

Cluster rule

Graph creation

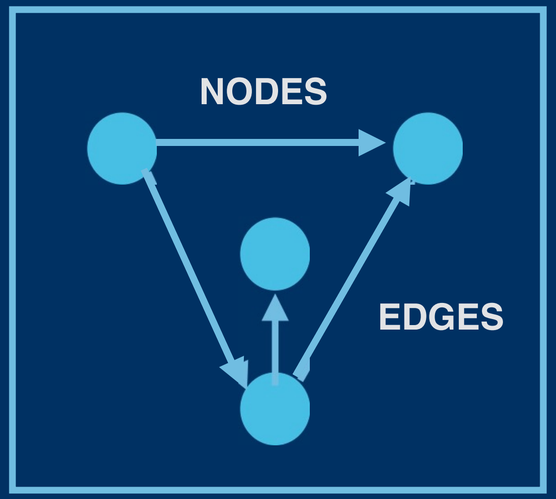
CHALLENGE #1

How to flexibly create new graphs that better fit the goals of particular marketing analysis?

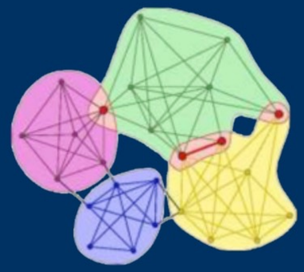
The query defines how to create a graph

INPUT QUERY

CUSTOM GRAPH CREATION



GRAPH ANALYSIS TOOLKIT



- **Social interaction types**
- **Hidden patterns in the network**
- **Information diffusion over time**

Scalability of graph algorithms

CHALLENGE #2

How to scale and optimise graph mining algorithms to handle large complex networks?

Detection of communities

Louvain algorithm

Ranking of nodes

HITS (authorities and hubs)

PageRank*

Roles of nodes

Within module degree

Participation coefficient

Graph storage

CHALLENGE #3

How to efficiently persist and query the graphs that represent large cross-media networks?

NEO4J

FEATURES

- NATIVE GRAPH STORAGE
- GPL V3 LICENSE
- NO SHARDING
- GRAPH QUERY

ELASTICSEARCH

FEATURES

- NON- NATIVE GRAPH STORAGE
- APACHE VERSION 2
- SHARDING
- GRAPH EXPLORE API

JANUS GRAPH

FEATURES

- NATIVE GRAPH STORAGE
- APACHE VERSION 2
- PARTITIONING DEPENDS ON STORAGE BACKEND
- GREMLIN

ORIENTDB

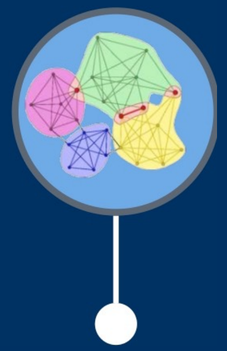
FEATURES

- MULTI-MODEL DATABASE
- APACHE VERSION 2
- SHARDING
- SUPPORTS SQL-LIKE QUERIES

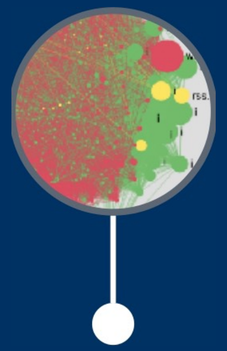
Graph visualisation

CHALLENGE #4

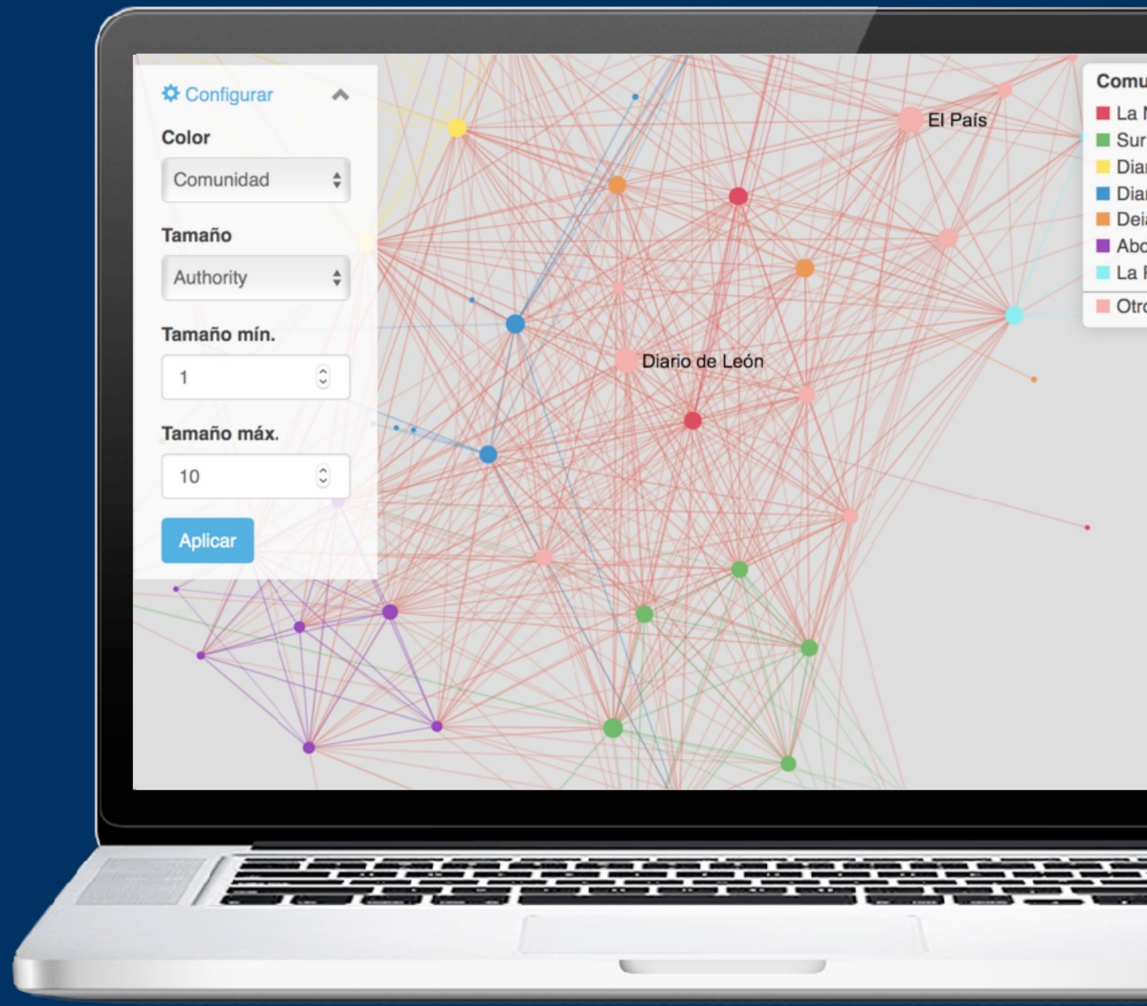
How to visualise large complex networks in an interpretable way?



Graph of communities



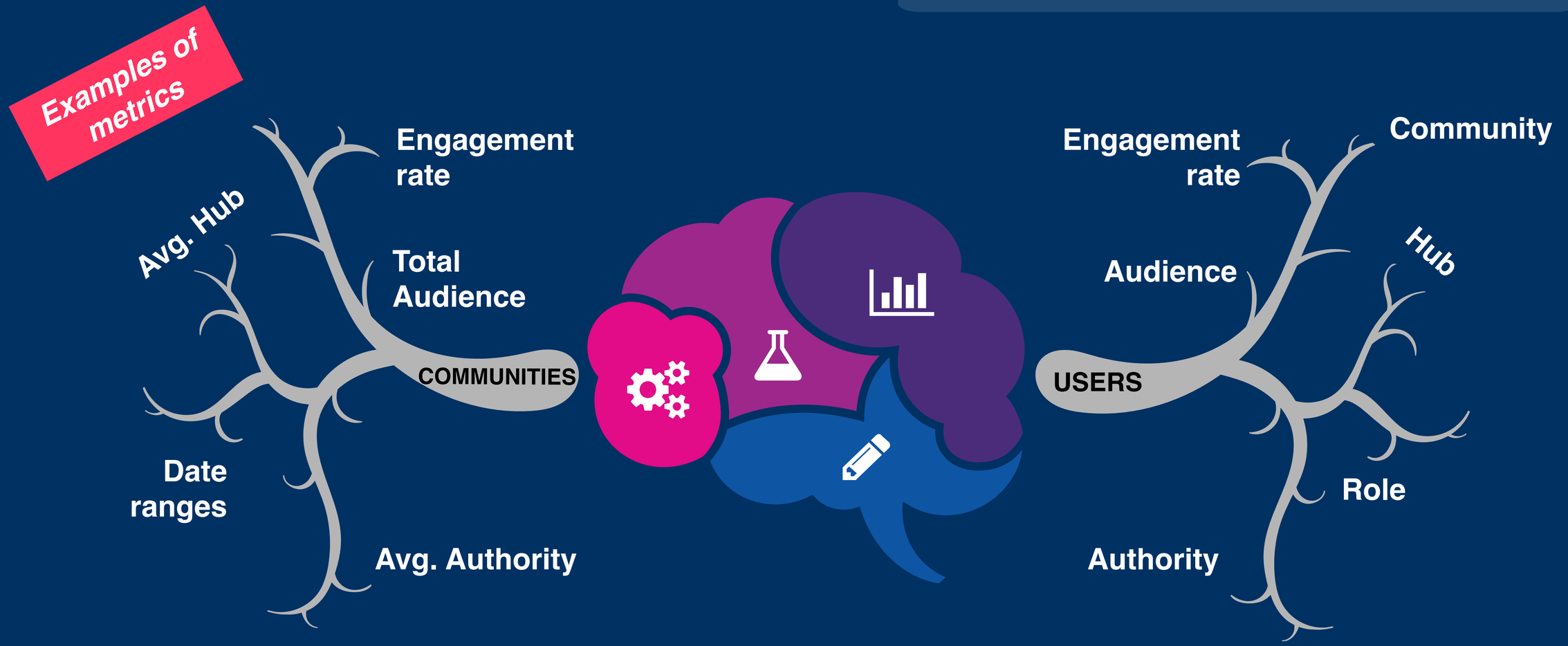
Graph of users



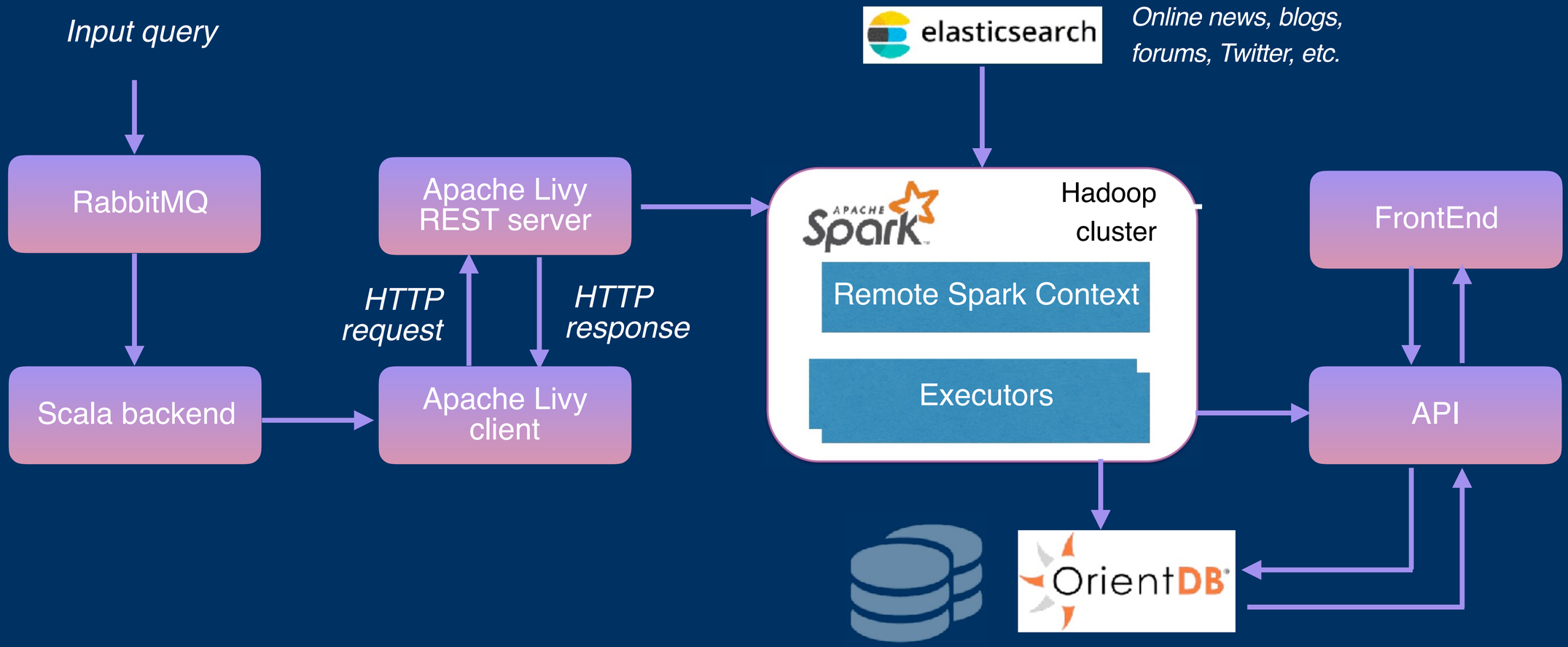
Graph visualisation

CHALLENGE #4

How to visualise large complex networks in an interpretable way?



Architecture



Cluster setup

SPARK CLUSTER (AMBARI)

3 INSTANCES

EACH INSTANCE:

- RAM: 16GB
- VCPU: 8
- DISK: 200GB

DISTRIBUTED DB (ORIENTDB)

2 INSTANCES

EACH INSTANCE:

- RAM: 16GB
- VCPU: 8
- DISK: 30GB

CONTROL INSTANCE

1 INSTANCE

INSTANCE SETUP:

- RAM: 4GB
- VCPU: 2
- DISK: 10GB

FRONT-END AND API

1 INSTANCE

INSTANCE SETUP:

- RAM: 4GB
- VCPU: 2
- DISK: 10GB

Our team

JUAN CARLOS CASTRO
DIGITAL PRODUCT MANAGER

PABLO ARAGÓN
SENIOR DATA SCIENTIST

JORDI RODA
M.Sc.
SOFTWARE ENGINEER

LIANA NAPALKOVA
SENIOR DATA SCIENTIST

EDUARDO RODRIGUEZ
SOFTWARE ENGINEER

Thanks for your attention!

Send your message at:
juancarlos.castro@eurecat.org
liana.napalkova@eurecat.org

Come visit us at the office in Barcelona:
Eurecat - Technology Center of Catalonia
Carrer de Bilbao 72
08005, Barcelona

Give us a call at:
+34 932 381 400

Follow us:
Eurecat_news
Eurecatorg