Mobility data storage and analysis

Antònia Tugores, Pere Colet

IFISC

CSIC

Universitat de les Illes Balears

Index

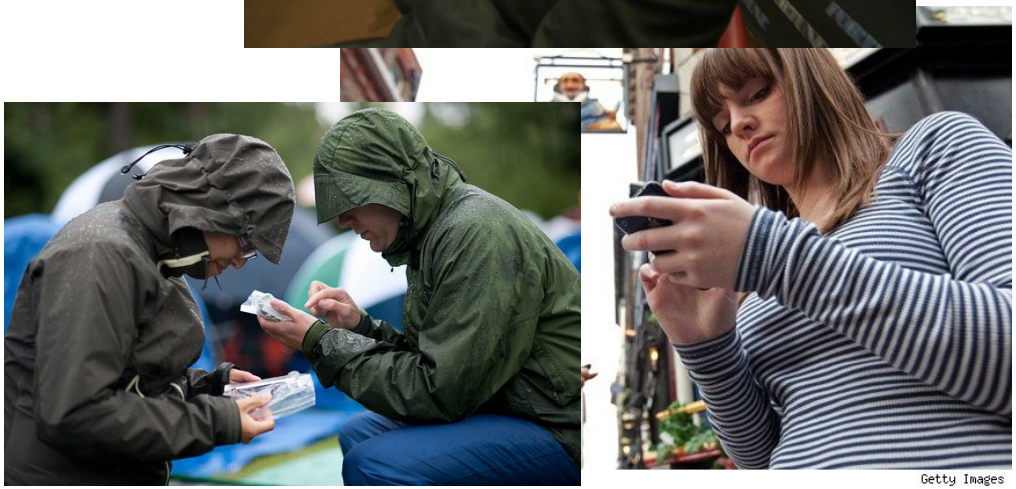
- Motivation
- Data storage
- Insert/Query performance
- Preliminary results and on going work

Motivation

Study urban mobility by using their tweets (specially the geolocated ones)

Compare with other on going studies
- Barcelona
- London
- Zurich

Not only tweets (in the future)



eunoia EVOLUTIVE USER-CENTRIC NETWORKS FOR INTRAURBAN ACCESSIBILITY

http://www.flickr.com/photos/mdgovpics/6756487295 http://www.flickr.com/photos/calgaryfolkfest/5966128182/

JSON document

JavaScript Object Notation
text-based open standard
designed for human-readable data interchange
language-independent
parsers available for many languages
alternative to XML
fields are in an arbitrary order


```
{
  "contributors": null,
  "truncated": false,
  "text": "Home, sweet home! I demà, a \"passar la tarda\" a Valencia. ^^
(@ Granvia Centre w/ @rubenralc) [pic]: http://t.co/7iA2XK52",
  "in_reply_to_status_id": null,
  "id": 277909591700942849,
  "retweeted": false,
  "coordinates":
    {
      "type": "Point",
      "coordinates": [2.13029439, 41.3665202]
    },
  "source": "foursquare",
  "in_reply_to_screen_name": null,
  "id_str": "277909591700942849",
  "retweet_count": 0,
  "in_reply_to_user_id": null,
  "favorited": false,
  "source_url": "http://foursquare.com",
  "user":
    {
  "geo":
    {
      "type": "Point",
      "coordinates": [41.3665202, 2.13029439]
    },
  "in_reply_to_user_id_str": null,
  "possibly_sensitive": false,
  "created_at": "2012-12-09T22:56:24",
  "in_reply_to_status_id_str": null,
  "place":
    {
      "country_code": "ES",
      "url": "http://api.twitter.com/1/geo/id/1aca5ffbc553e95b.json",
      "country": "Spain",
      "place_type": "city",
      "bounding_box":
        {
          "type": "Polygon",
          "coordinates": [[[2.086323, 41.336062], [2.137433,
41.336062], [2.137433, 41.380548], [2.086323, 41.380548]]]
        },
      "full_name": "Hospitalet de Llobregat, Barcelona",
      "attributes": {},
      "id": "1aca5ffbc553e95b",
      "name": "Hospitalet de Llobregat"
    },
  "_id": "5107d50e66946222fce81631"
}
```


GeoJSON

open format for encoding collections of simple feature objects
points
line strings
polygons
and multi-part collections of points, lines and polygons
with their non-spatial attributes
using JavaScript Object Notation

IFISC

Tweet

(+ RT info)

tweet content (140 characters)

id

geolocation (if enabled)

source

user information

creation time

place (if defined)

fields are in an arbitrary order
fields scheme can change


```
{
  "contributors": null,
  "truncated": false,
  "text": "Home, sweet home! I demà, a \"passar la tarda\" a Valencia. ^^
(@ Granvia Centre w/ @rubenralc) [pic]: http://t.co/7iA2XK52",
  "in_reply_to_status_id": null,
  "id": 277909591700942849,
  "retweeted": false,
  "coordinates":
    {
      "type": "Point",
      "coordinates": [2.13029439, 41.3665202]
    },
  "source": "foursquare",
  "in_reply_to_screen_name": null,
  "id_str": "277909591700942849",
  "retweet_count": 0,
  "in_reply_to_user_id": null,
  "favorited": false,
  "source_url": "http://foursquare.com",
  "user":
    {
  "geo":
    {
      "type": "Point",
      "coordinates": [41.3665202, 2.13029439]
    },
  "in_reply_to_user_id_str": null,
  "possibly_sensitive": false,
  "created_at": "2012-12-09T22:56:24",
  "in_reply_to_status_id_str": null,
  "place":
    {
      "country_code": "ES",
      "url": "http://api.twitter.com/1/geo/id/1aca5ffbc553e95b.json",
      "country": "Spain",
      "place_type": "city",
      "bounding_box":
        {
          "type": "Polygon",
          "coordinates": [[[2.086323, 41.336062], [2.137433,
41.336062], [2.137433, 41.380548], [2.086323, 41.380548]]]
        },
      "full_name": "Hospitalet de Llobregat, Barcelona",
      "attributes": {},
      "id": "1aca5ffbc553e95b",
      "name": "Hospitalet de Llobregat"
    },
  "_id": "5107d50e66946222fce81631"
}
```

IFISC

User Information

id

location

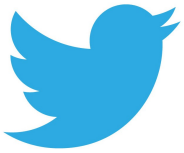
description

language

fields are in an arbitrary order


```
21    "user":
22      {
23        "follow_request_sent": false,
24        "profile_use_background_image": true,
25        "contributors_enabled": false,
26        "id": 260287521,
27        "verified": false,
28        "profile_image_url_https":
29    "https://si0.twimg.com/profile_images/1533280952/tw_12403594_1315422640_normal.
30    jpg",
31        "profile_sidebar_fill_color": "252429",
32        "profile_text_color": "666666",
33        "followers_count": 128,
34        "protected": false,
35        "id_str": "260287521",
36        "default_profile_image": false,
37        "location": "L'Hospitalet de Llobregat",
38        "utc_offset": 3600,
39        "statuses_count": 1747,
40        "description": "Necesito meterme algo dentro, un poco de café o
41    algo... Y luego, de alguna manera, el mundo será un poco mejor. (Sam Vimes,
42    Hombres de Armas, Terry Pratchett)",
43        "friends_count": 570,
44        "profile_link_color": "AA0000",
45        "profile_image_url": "http://a0.twimg.com/profile_images/1533280952/
46    tw_12403594_1315422640_normal. jpg",
47        "notifications": null,
48        "geo_enabled": true,
49        "profile_background_color": "000000",
50    "profile_background_image_url":"http://a0.twimg.com/profile_background_images/
51    600683524/n9duka667b9ju3n9wgw4. jpeg",
52        "screen_name": "jX09A",
53        "lang": "ca",
54        "following": false,
55        "profile_background_tile": false,
56        "favourites_count": 7,
57        "name": "Jordi Cant\u00f3",
58        "url": null,
59        "created_at": "2011-03-03T15:45:46",
60        "profile_background_image_url_https": "https://si0.twimg.com/
61    profile_background_images/600683524/n9duka667b9ju3n9wgw4. jpeg",
62        "time_zone": "Madrid",
63        "profile_sidebar_border_color": "181A1E",
64        "default_profile": false,
65        "is_translator": false,
66        "listed_count": 1
67      },
```

Data size estimation



2KB/tweet

Tweet: aprox. 20 key/value
User: aprox 30 key/value

- id
- content (140chars)
- geolocation
- creation time
- place
- user information
  - id
  - screen name
  - description
  - creation time
  - location
  - friends count
  - ….
- ...

Data size estimation

2KB/tweet x 15million tweets/day

30 GB/day

Tweet: aprox. 20 key/value
User: aprox 30 key/value

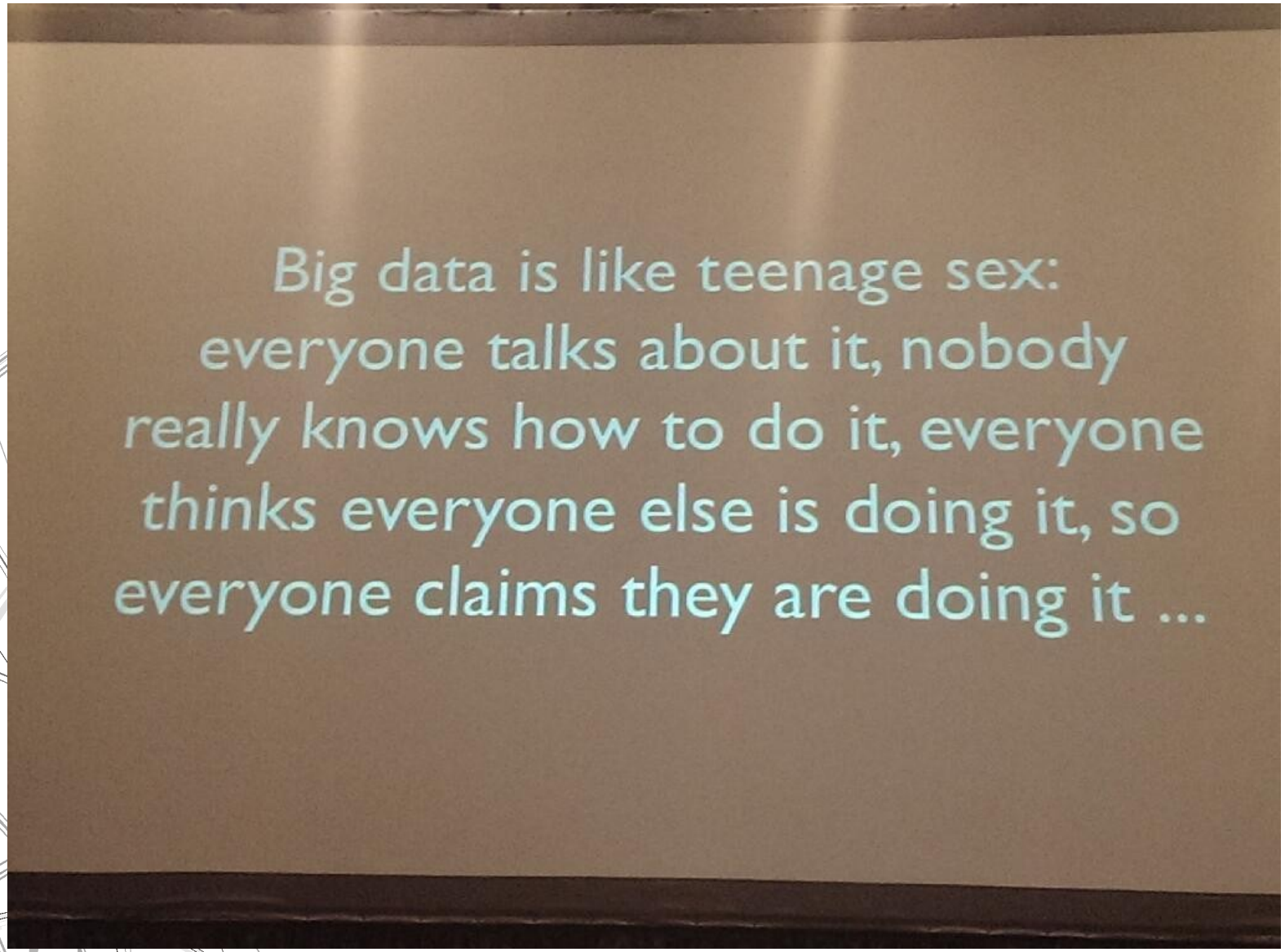
2500million tweets/year

6TB/year

- id
- content (140chars)
- geolocation
- creation time
- place
- user information
  - id
  - screen name
  - description
  - creation time
  - location
  - friends count
  - ….
- ...

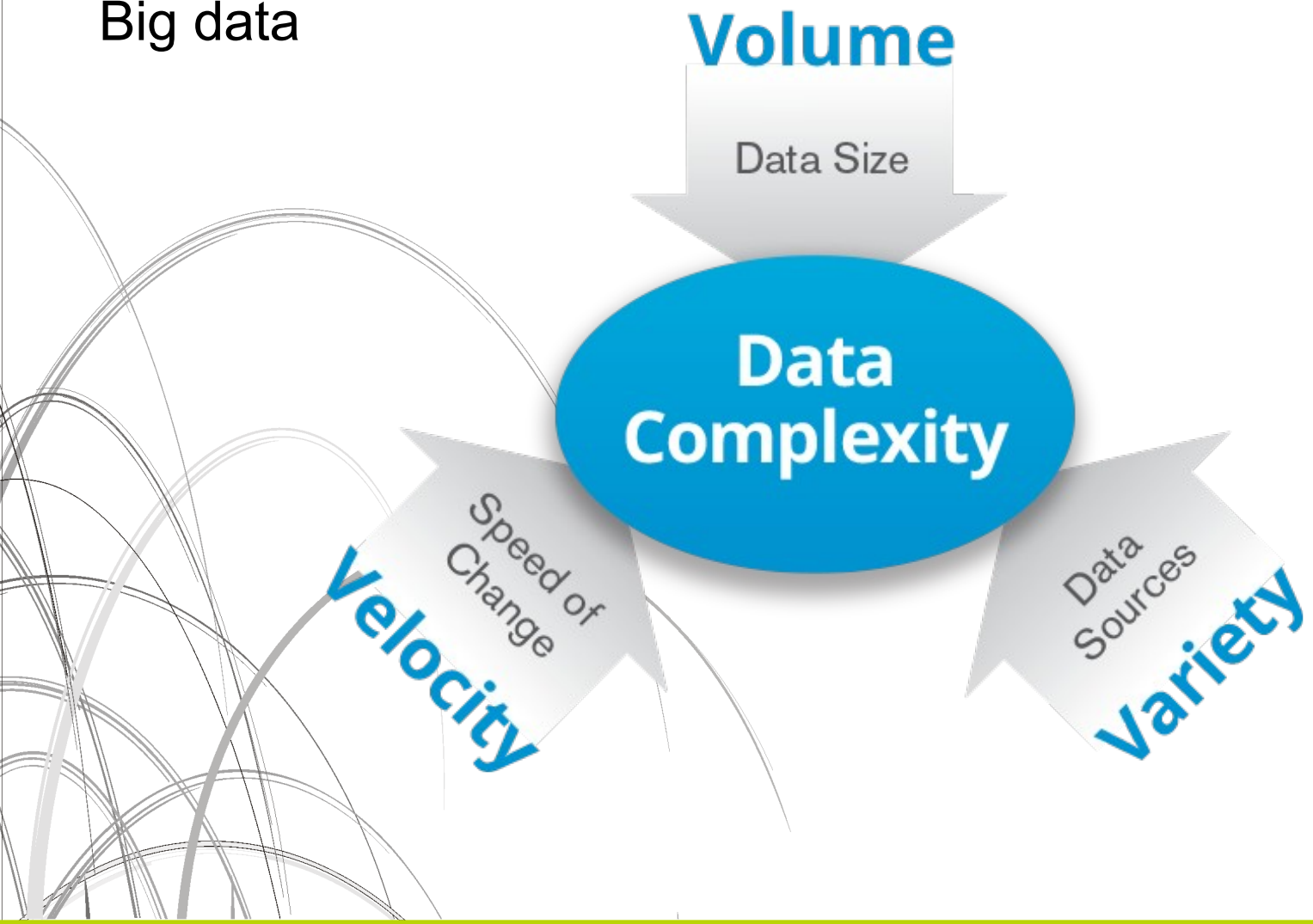
http://library.uoregon.edu/node/3475
http://www.barclayandtodds.com/pages/24-7_Service.cfm


big data is...



https://twitter.com/toniborrallo/status/391526056852996096

Big data


Volume
Data Size
Data Complexity
Speed of Change
Velocity
Data Sources
Variety

We need to...

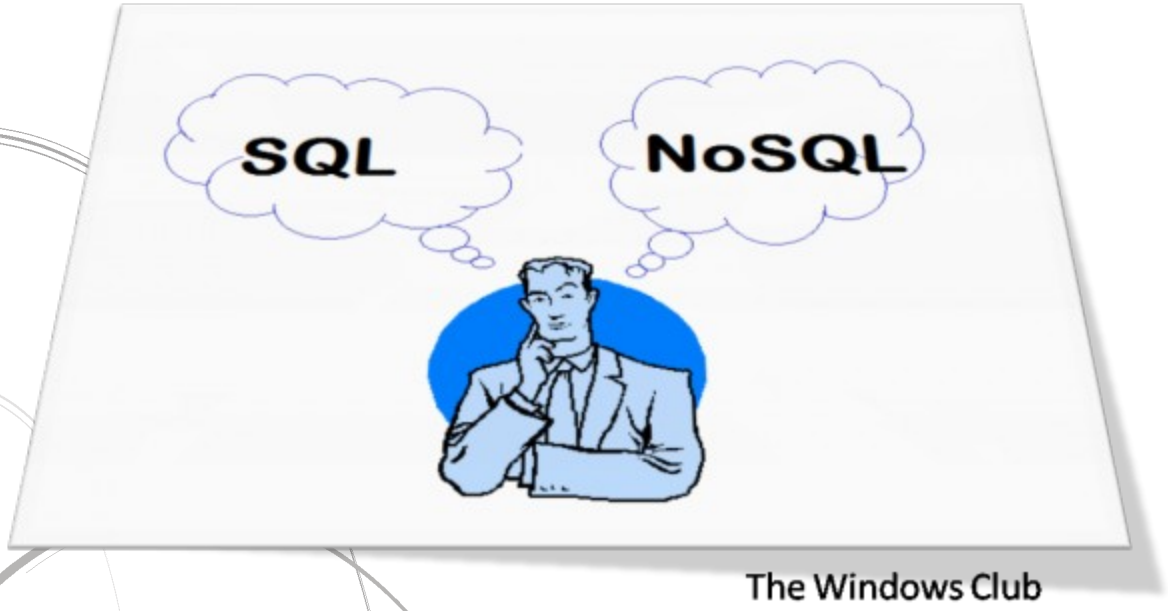
Efficiently store data
Efficiently manage data (select subsets)
Efficiently analyse data

http://bitzermobile.com/wp-content/uploads/2012/07/puzzle-piece.png
http://library.uoregon.edu/sites/default/files/node3393/data_management.jpg
http://t2.gstatic.com/images?q=tbn:ANd9GcTlGtxf7mKEl6IU0ebDQhPHgq1zTJzQssuDQMEv5Kw7C7N4GGgw

Requirements to store and manage the data

* Capability to store **billions** of documents
* Fast storage rate
* Scalability
* High availability
* High search performance
* Adaptative format (twitter can change the data format)

Databases



http://blog.izenda.com/wp-content/uploads/2013/05/SQL-600x309.png

Databases

Relational (SQL) databases

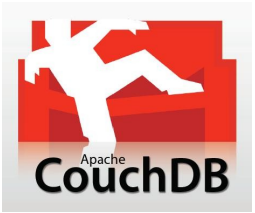
has a collection of tables of data items all of which is formally described and organized according to the relational model

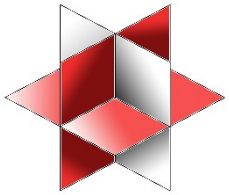
Examples
- Oracle
- PostgreSQL
- MySQL

Non relational databases

A NoSQL database provides a mechanism for storage and retrieval of data that uses looser consistency models than traditional relational databases.

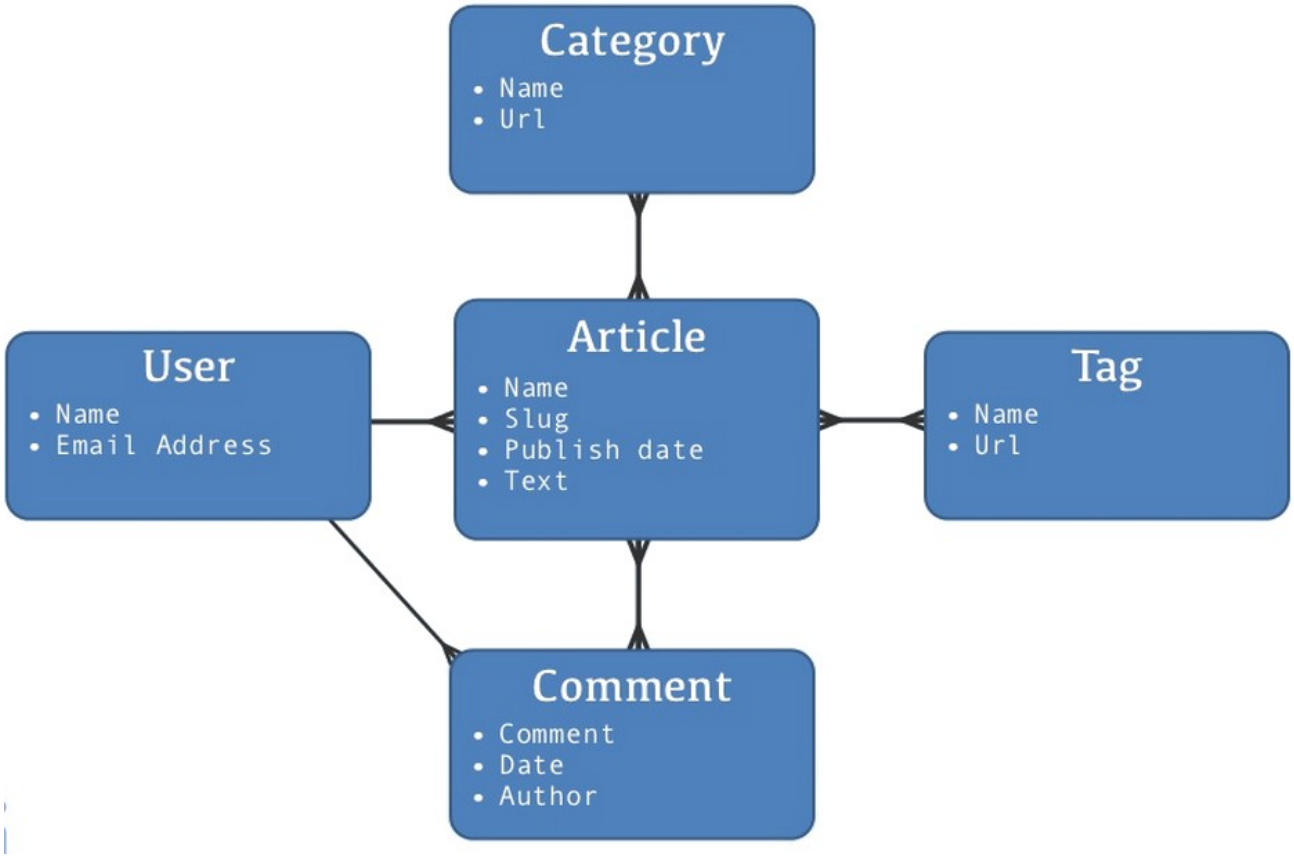
Examples
- MongoDB
- CouchDB
- HyperDex





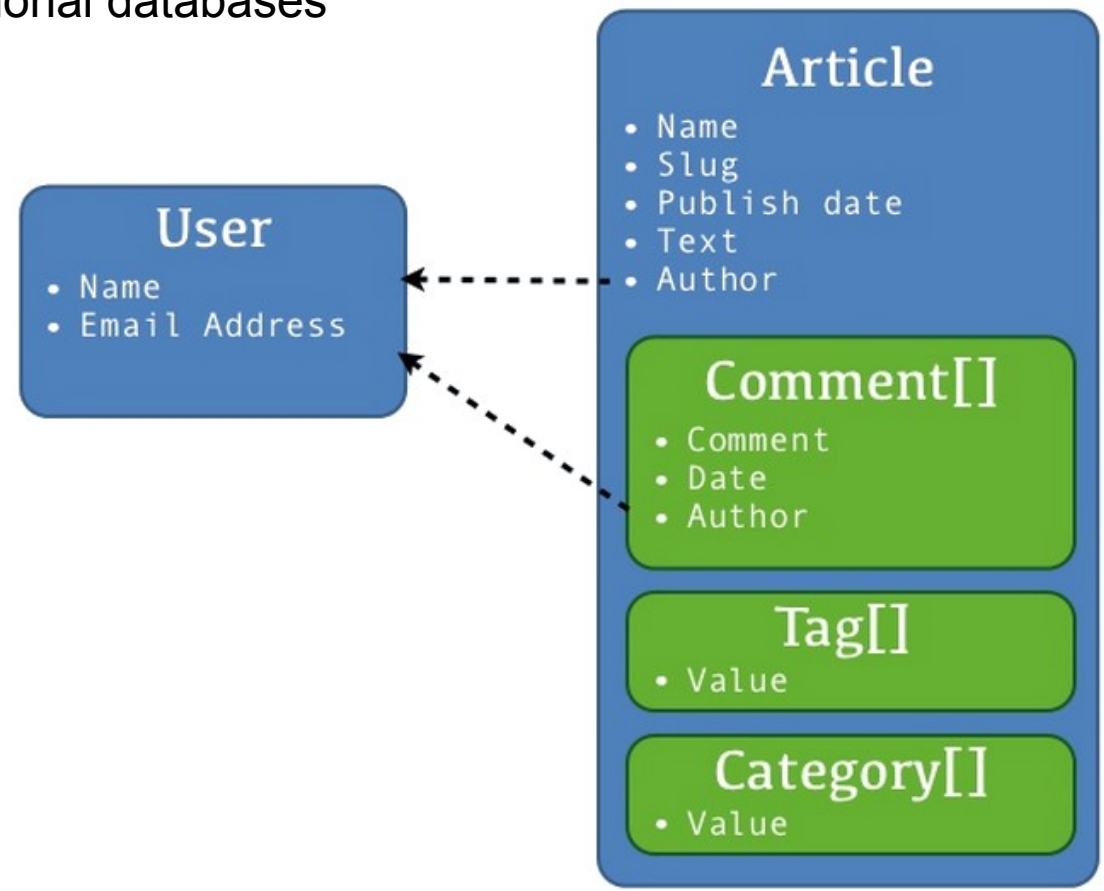
Databases

Relational (SQL) databases



Databases

Non Relational databases

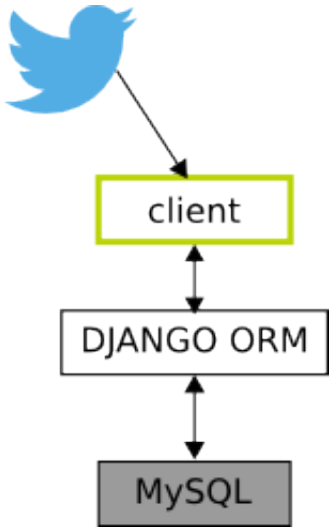


DB comparison

	SQL	noSQL
High performance	Indexes	Indexes
High availability	Automatic master failover and recovery	Replica sets with automatic master failover and recovery
Easy scalability	Hardware: $$$$$$$ Soft: MySQL or $$$$$ = $$$$$$$$$	Hardware: Add ANY computer Soft: MongoDB = Automatic sharding

Empiric comparison: SQL vs noSQL

MySQL configuration



Physical computer:

- 16GB RAM
- 8 cores (2x Xeon L5520 @ 2.27GHz)
- 2TB, 7200 rpm

Indexing:

- id
- user.id
- coordinates.coordinates
- created_at

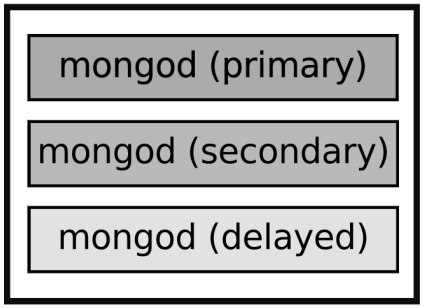
MongoDB minimal configuration

MongoDB “Data”

Holds the data

It can be

- A single server with the data
- A replica set with at least 3 servers with data replicated among them for security.
  One of the elements of the replica set acts as primary and the others are secondaries.
  Some of the secondary one can replicate the data with a delay time
  so that they can be used a back up.



Client / Application server

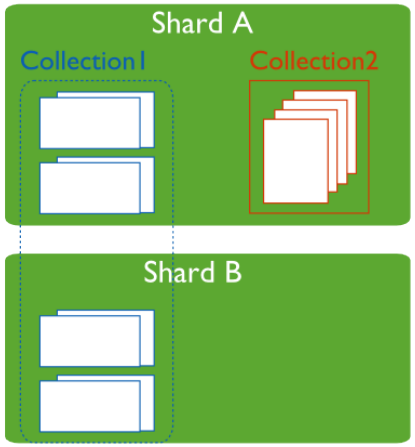
routes the reads and writes from applications to the data

MongoDB sharding

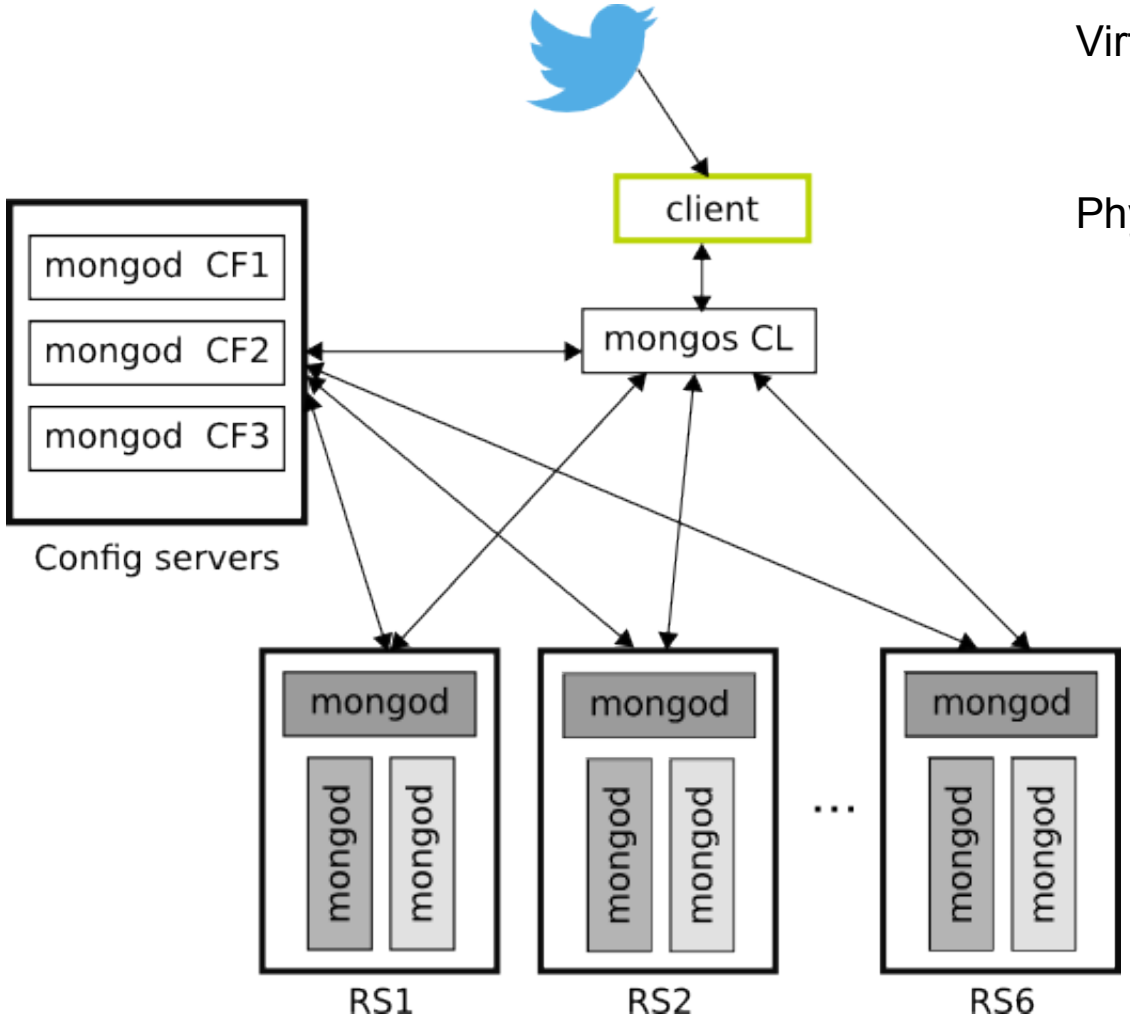
Several MongoDB shards (each formed by a replica set with at least three servers)
Data is distributed among the instances (sharding).

MongoDB configuration servers
distribute the data among the MongoDB instances (chunks)
hold the metadata of the cluster (relations between shard and chunks)

Client / Application server
routes the reads and writes from applications to the shards



MongoDB configuration



Virtual computers:
- 1-2GB RAM
- 1-2 cores

Physical computers:
- 16GB RAM
- 8 cores (2x Xeon L5520 @ 2.27GHz)
- 2TB, 7200 rpm

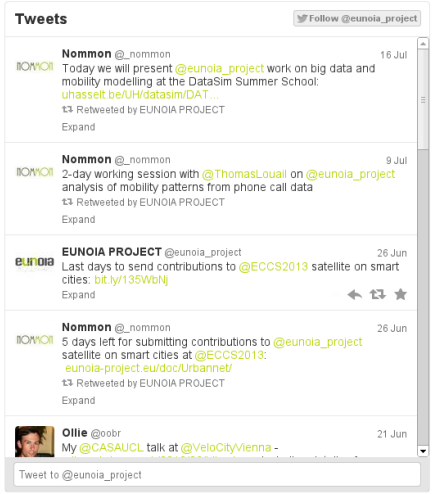
Indexing:
- id
- user.id
- coordinates.coordinates (2d)
- created_at

Data collection

Twitter APIs.

- StreamAPI
  - Client receives a small sample of all public statuses (aprox 1%)
  - Only 12% of them are geolocalized
    - and of those, only a small fraction are in the cities of interest.
- REST API
  - Allows to get a users' timeline.
  - We use that for the most active users located in the cities of interest.
  - After one year of data collection, in London there are 240K users and 6M tweets

We use **tweepy** to interface with the Twitter APIs
simplejson to convert from/to text/JSON
pymongo to insert the data to MongoDB database
django-orm to insert the data to MySQL database



Data collection

Stream

```
from tweepy import Stream, OAuthHandler
from tweepy.streaming import StreamListener


class StdOutListener(StreamListener):
    def on_data(self, data):
        print data
        return True


auth = OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
auth.set_access_token(ACCESS_KEY, ACCESS_SECRET)

listen = StdOutListener()
stream = Stream(auth, listen, gzip=True)

stream.sample()
```

Data collection

Users timeline

continuously:
    get user ids with tweets geolocated in the cities of interest
        (sorted by number of geolocated tweets)
    for the first 43200 uids (more active)
        get timeline (from last stored tweet until now)

Users network

continuously:
    get user ids with tweets geolocated in the cities of interest
        (sorted by number of geolocated tweets)
    for all of them
        get current list of following and followers (uids)

Data collection

Write to DB (MySQL with Django-ORM)

```
class Tweet(Model):
    twid = BigIntegerField(primary_key=True,db_index=True)
    place = ForeignKey(Place, null=True)
    text = CharField(max_length=2048, blank=True)
    retweet_count = IntegerField(null=True)
    parent_id = BigIntegerField(null=True)
    source = CharField(max_length=2048)
    coordinates = ForeignKey(BoundingBox, null=True)
    contributors = CharField(max_length=2048, null=True)
    retweeted = BooleanField()
    truncated = BooleanField()
    created_at = DateTimeField(null=True)
    user = ForeignKey(User)
    entities = ForeignKey(Entities, null=True)
    in_reply_to_status_id = BigIntegerField(null=True)
    in_reply_to_user_id = BigIntegerField(null=True)
    in_reply_to_screen_id = BigIntegerField(null=True)
    deleted = BooleanField()

    class Meta:
        app_label = 'twitter'
```

```
for line in tweets_file:
        tweet = fillTweet(line)
        tweet.save()
```

Idem for user, hashtag, coordinates, url, …

Data collection

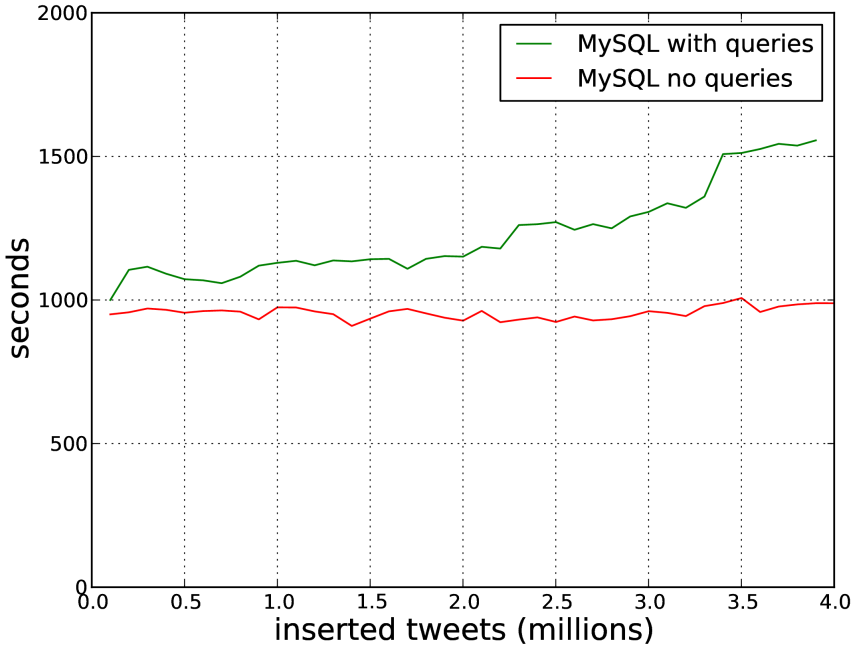
Write to DB (MongoDB)

```
mongoserver_uri = "mongodb://" + user + ":" + pwd + "@" +
                            host + ":" + port + "/" + dbname
conection = MongoClient(host=mongoserver_uri)
db = conection[dbname]
collection = db[collname]

for line in tweetsfile:
    tweet = simplejson.loads(line.encode('utf8'))
    collection.insert(tweet)
```

Insertion performance (time to insert 100K tweets)

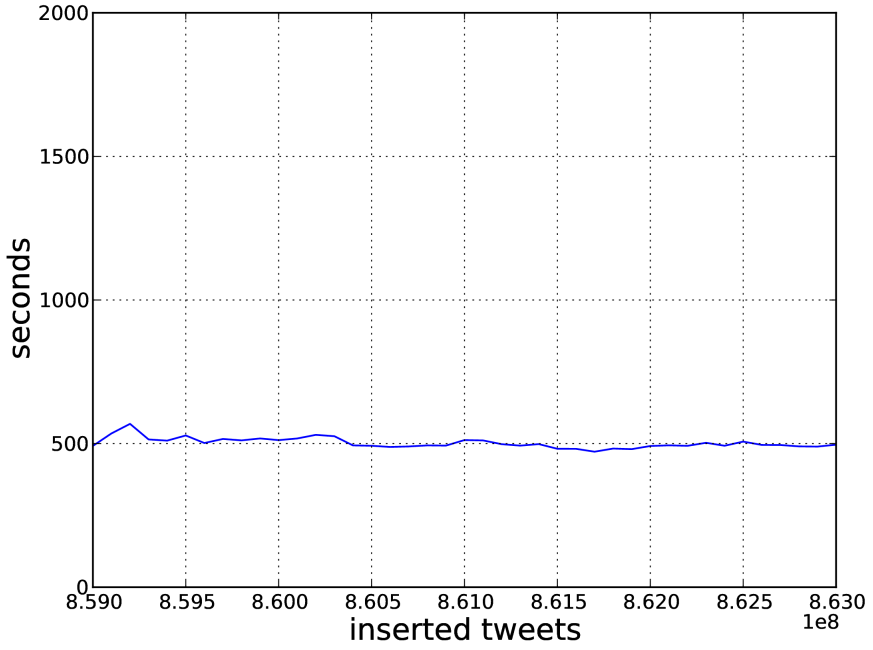
MySQL



empty DB

Django ORM
(+ relations)

MongoDB



850 million tweets
3 replicasets

JSON insertion + pymongo
(200 tweets/second)

Query performance (MongoDB)

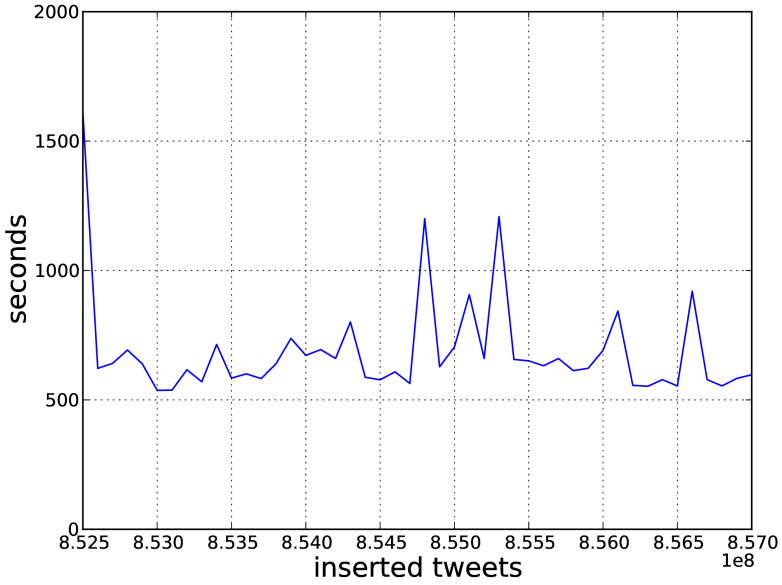
database with 1 thousand million tweets

geonear: get the closest documents to a given point with a maximum distance

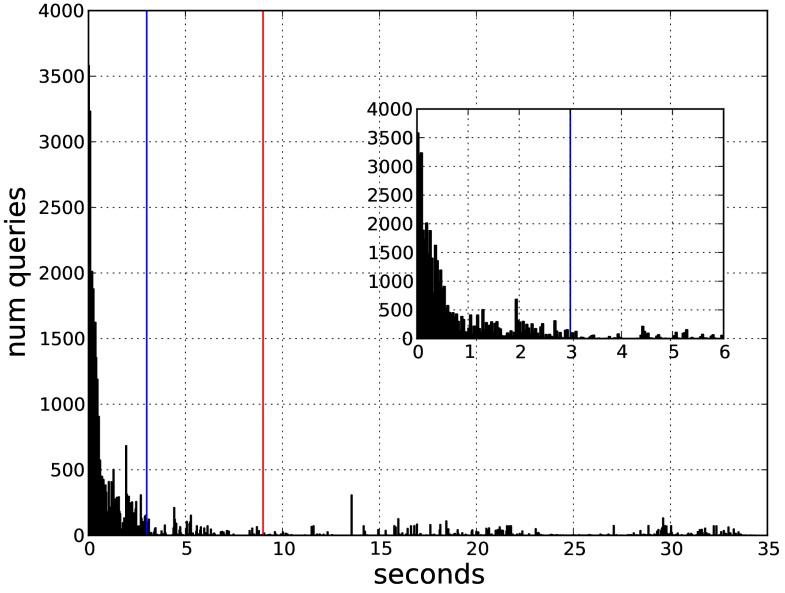
tessellate the city maps with cells of aprox 1km² (16MB size)

run a query per cell

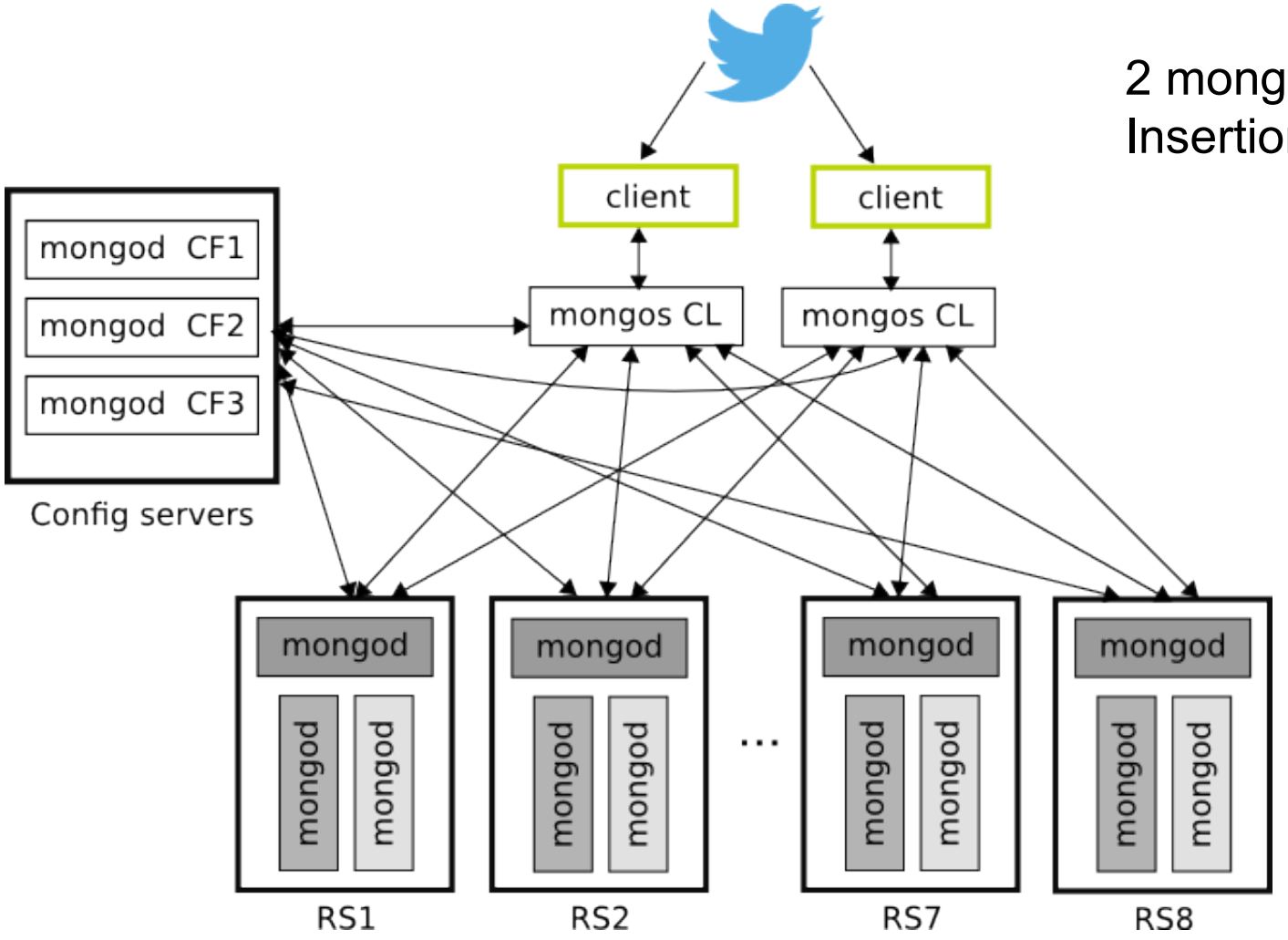
Insertion speed when querying the primary node. (not production)



Queries timing histogram for Barcelona metropolitan area. Blue line shows the median and red line the 70th percentile.



MongoDB configuration improvements



2 mongos CL
Insertion in bunches

MongoDB issues

- Documents (JSON) and Javascript for queries
- Authorisation/authentication → 2.4
- Default write concern
- Synchronization between primary and secondaries
- Geoindexes: 2d and 2dsphere
- “Big” data backups are not easy → use delayed members in RS

Twitter (and other social network) data issues

- Bias: more tweets in cities
- Signal: people not represented
- Scale: sometimes, small data is better than big data
- Correlations between different subsets
- Big bad data (data quality analysis)
- ….

Preliminary results and on going work

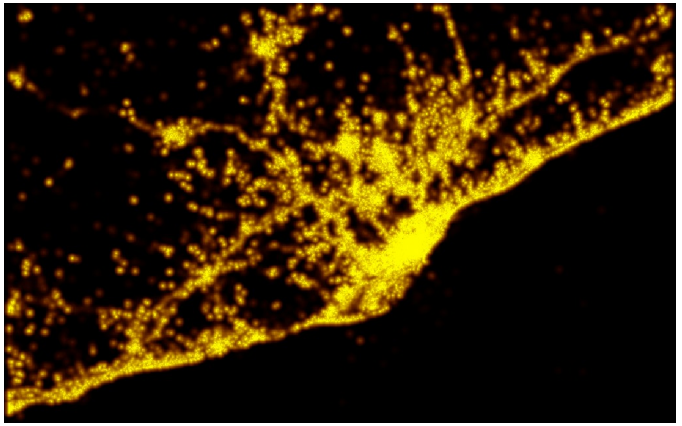
2 billion tweets stored in the database.
250 million geolocated (aprox 13%)

5.9 millions located in London
1.5 millions located in Barcelona
0.2 millions located in Zurich

Users with at least one tweet in the city
260 thousand users in London
25 thousand users in Barcelona
3 thousand users in Zurich

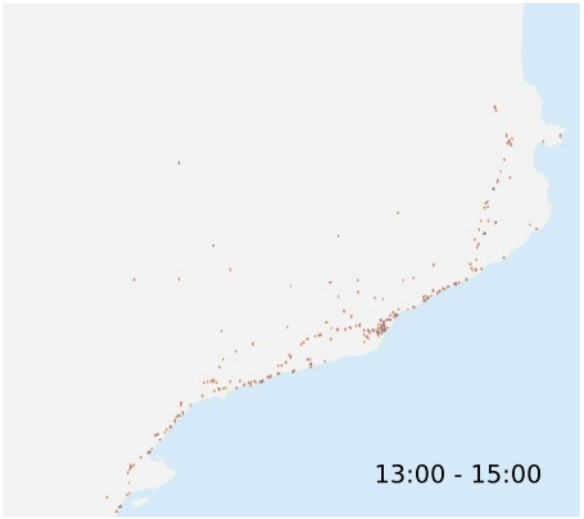
Preliminary data shows a good agreement with population distribution and the transportation network of the cities

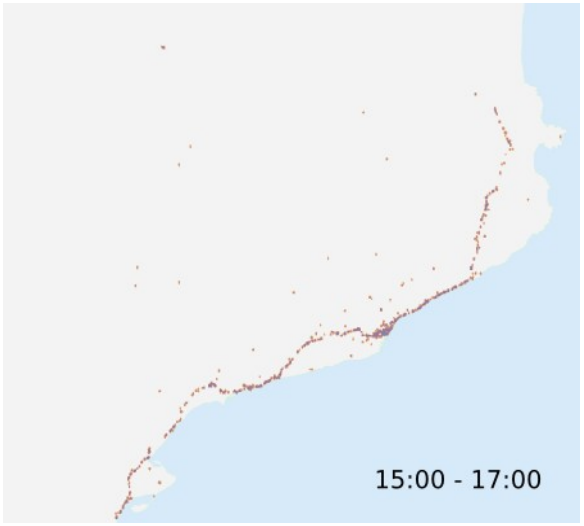
On going work: more detailed comparison with traffic data of the transportation networks with traditional datasets within the European project EUNOIA

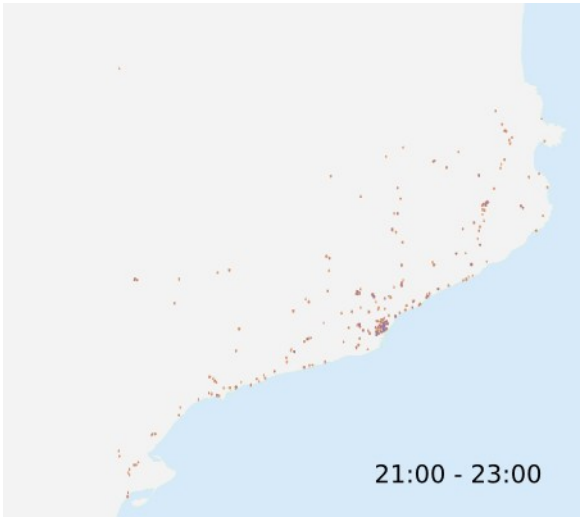


#ViaCatalana

11/09/2013: demonstration in the center of Barcelona and a human chain crossing Catalonia from North to South along the Mediterranean coast.

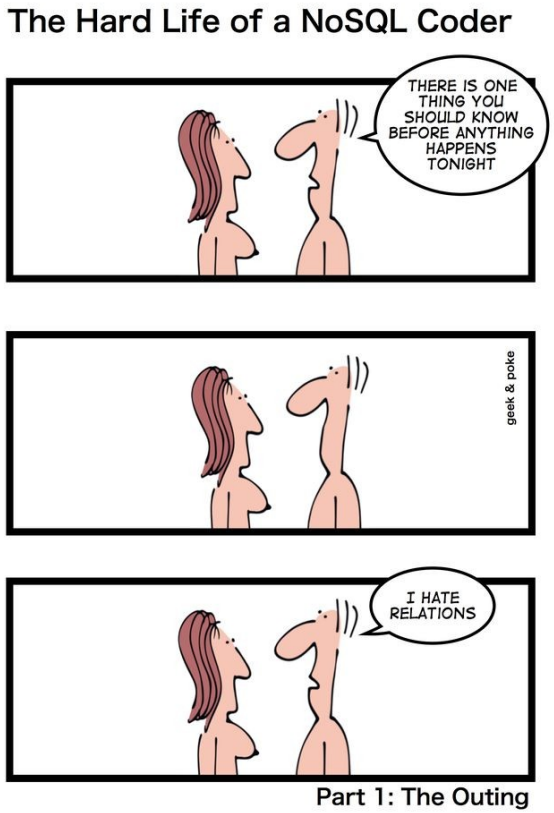






http://ifisc.uib-csic.es/humanmobility

aprox 175000 tweets
4600 geolocated



Thank you

Questions?