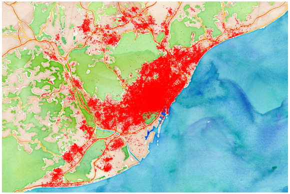
A GEO-stack for Big Data

Driving Spatial Analysis Beyond the Limits of Traditional Storage

Joana Simões [1]

[1]BDigital, CASA, CICS.NOVA

March 9, 2015



Big Data Word Cloud; source:http://olap.com/big-data/

Table of Contents

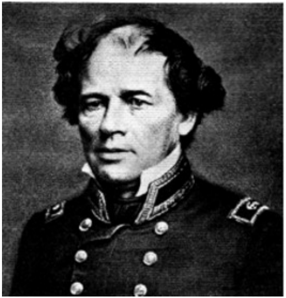
Warning

* This presentation may contain tech talk, such as: databases, clusters, NoSQL.
If you are susceptible to these concepts, you may want to leave the room **now**.
From this point beyond, you are at your own risk! *

The Value of Data

Not a New Idea!

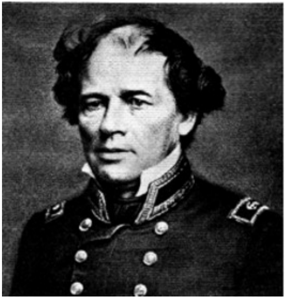
- Matthew Fontaine Maury (1806-1873).



The Value of Data

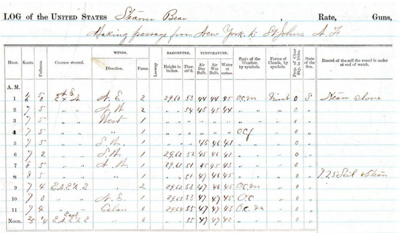
Not a New Idea!

- Matthew Fontaine Maury (1806-1873).
- Foresaw the hidden value on captain's ships logs, when analysed collectively.
- Used time series data to carry out analysis that would enable him to recomend optimal shipping routes.



The Value of Data

Data Mining, Open-source and Crowd-sourcing



LOG of the UNITED STATES Steamer Bear, Rate, Guns,

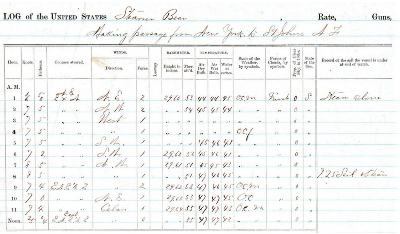
Making passage from New York to St Johns N.F.

Hour.	Knots.	Fathoms.	Courses steered.	Winds. Direction.	Winds. Force.	Leeway	Barometer. Height in inches.	Barometer. Ther. att'd.	Temperature. Air Dry Bulb.	Temperature. Air Wet Bulb.	Temperature. Water at surface.	State of the Weather, by symbols.	Forms of Clouds, by symbols.	Prop. of Clear Sky, in 10ths.	State of the Sea.	Record of the sail the vessel is under at end of watch.
A. M.																
1	6	5	E x N	N. E.	2		29.60	53	44	44	45	O.C.m	Nimb	0	S	Steam alone
2	7	5	" "	N. W.	2		" "	54	45	45	44	" "	"	0	"	
3	7	5	" "	West	1		" "	"	"	"	"	" "	"	0	"	
4	7	5	" "	"	1		" "	"	"	"	"	O.C.f	"	0	"	
5	7	5	" "	S. W.	1		" "	"	45	46	41	" "	"	0	"	
6	7	2	" "	S. W.	1		29.62	52	45	46	41	" "	"	0	"	
7	8	5	" "	N. W.	1		29.61	51	45	45	43	" "	"	0	"	
8	8	5	" "	"	1		" "	51	47	46	45	" "	"	0	"	7.25 Sail & Steam
9	7	4	E.N.E½E	"	2		29.62	53	47	46	45	O.C.m	"	0	"	
10	7	0	" "	N. E.	1		29.63	53	47	47	45	O.C	"	0	"	
11	7	4	" "	Calm	0		29.64	55	47	47	43	O.C.m	"	0	"	
Noon.	25	04	East E.N.E½E	"	0		" "	55	47	47	43	" "	"	0	"	

The Value of Data

Data Mining, Open-source and Crowd-sourcing

- In 1848, Captain Jackson was the the first person to try the route recommended by Maury, and as a result he was able to save 17 days on his outbond trip.



LOG of the UNITED STATES Steamer Bear , Rate, Guns,

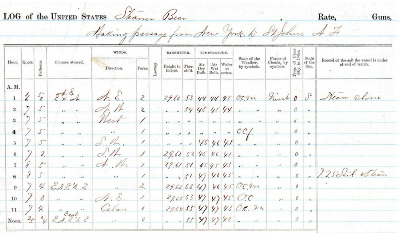
Making passage from New York to St Johns N.F.

Hour.	Knots.	Fathoms.	Courses steered.	WINDS. Direction.	Force.	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the Weather, by symbols.	Forms of Clouds, by symbols.	Prop. of Clear Sky, in 10ths.	State of the Sea.	Record of the sail the vessel is under at end of watch.
A. M.																
1	6	5	E x N	N.E.	2		29.60	53	44	44	45	O.C.m	Nimb	0	S	Steam alone
2	7	5	" "	N.W.	2		" "	54	45	45	44	" "	"	0	"	
3	7	5	" "	West	1		" "	"	"	"	"	" "	"	0	"	
4	7	5	" "	"	1		" "	"	"	"	"	O.C.f	"	0	"	
5	7	5	" "	S.W.	1		" "	"	45	46	41	" "	"	0	"	
6	7	2	" "	S.W.	1		29.62	52	45	46	41	" "	"	0	"	
7	8	5	" "	N.W.	1		29.61	51	45	45	43	" "	"	0	"	
8	8	5	" "	"	1		" "	51	47	46	45	" "	"	0	"	7.25 Sail & Steam
9	7	4	E.N.E ½ E	"	2		29.62	53	47	46	45	O.C.m	"	0	"	
10	7	0	" "	N.E.	1		29.63	53	47	47	45	O.C	"	0	"	
11	7	4	" "	Calm	0		29.64	55	47	47	43	O.C.m	"	0	"	
Noon.	25	04	East E.N.E ½ E	"	0		" "	55	47	47	43	" "	"	0	"	

The Value of Data

Data Mining, Open-source and Crowd-sourcing

- In 1848, Captain Jackson was the the first person to try the route recommended by Maury, and as a result he was able to save 17 days on his outbond trip.
- Apart from collecting existing logs, Maury encouraged the collection of more regular and systematic time series, by creating a template.



LOG of the UNITED STATES Steamer Bear, Rate, Guns,

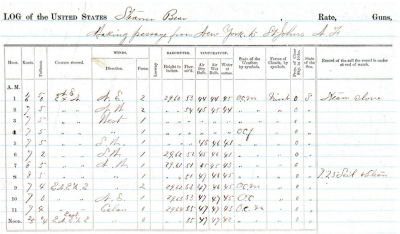
Making passage from New York to St Johns N.F.

Hour.	Knots.	Fathoms.	Courses steered.	WINDS. Direction.	WINDS. Force.	Leeway	BAROMETER. Height in inches.	BAROMETER. Ther. att'd.	TEMPERATURE. Air Dry Bulb.	TEMPERATURE. Air Wet Bulb.	TEMPERATURE. Water at surface.	State of the Weather, by symbols.	Forms of Clouds, by symbols.	Prop. of Clear Sky, in 10ths.	State of the Sea.	Record of the sail the vessel is under at end of watch.
A. M.																
1	6	5	E x N	N.E.	2		29.60	53	44	44	45	O.C.m	Nimb	0	S	Steam alone
2	7	5	" "	N.W.	2		" "	54	45	45	44	" "	"	0	"	
3	7	5	" "	West	1		" "	"	"	"	"	" "	"	0	"	
4	7	5	" "	"	1		" "	"	"	"	"	O.C.f	"	0	"	
5	7	5	" "	S.W.	1		" "	"	45	46	41	" "	"	0	"	
6	7	2	" "	S.W.	1		29.62	52	45	46	41	" "	"	0	"	
7	8	5	" "	N.W.	1		29.61	51	45	45	43	" "	"	0	"	
8	8	5	" "	"	1		" "	51	47	46	45	" "	"	0	"	7.25 Sail & Steam
9	7	4	E.N.E.½E	"	2		29.62	53	47	46	45	O.C.m	"	0	"	
10	7	0	" "	N.E.	1		29.63	53	47	47	45	O.C	"	0	"	
11	7	4	" "	Calm	0		29.64	55	47	47	43	O.C.m	"	0	"	
Noon.	25	4	East E.N.E.½E	"	0		" "	55	47	47	43	" "	"	0	"	

The Value of Data

Data Mining, Open-source and Crowd-sourcing

- In 1848, Captain Jackson was the the first person to try the route recommended by Maury, and as a result he was able to save 17 days on his outbond trip.
- Apart from collecting existing logs, Maury encouraged the collection of more regular and systematic time series, by creating a template.
- Collected Data: longitude, latitude, currents, magnetic variation, air and water temperature, general wind direction, etc.

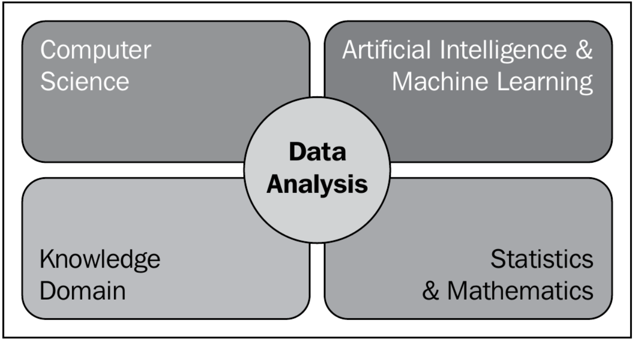

LOG of the United States Steamer Bear , Rate, Guns,

Making passage from New York to St Johns N. F.

Hour.	Knots.	Fathoms.	Courses steered.	Winds. Direction.	Winds. Force.	Leeway.	Barometer. Height in inches.	Barometer. Ther. att'd.	Temperature. Air Dry Bulb.	Temperature. Air Wet Bulb.	Temperature. Water at surface.	State of the Weather, by symbols.	Forms of Clouds, by symbols.	Prop. of Clear Sky, in 10ths.	State of the Sea.	Record of the sail the vessel is under at end of watch.
A. M.																
1	6	5	E x N	N. E.	2		29.60	53	44	44	45	O.C.m	[illegible]	0	S	Steam alone
2	7	5	" "	N. W.	2		"	54	45	45	44	" "	"	0	"	
3	7	5	" "	West	1		"	"	"	"	"	" "	"	0	"	
4	7	5	" "	"	1		"	"	"	"	"	O.C.f	"	0	"	
5	7	5	" "	S. W.	1		"	"	45	46	41	" "	"	0	"	
6	7	2	" "	S.W.	1		29.62	52	45	46	41	" "	"	0	"	
7	8	5	" "	N. W.	1		29.61	51	45	45	43	" "	"	0	"	
8	8	5	" "	"	1		"	51	47	46	45	" "	"	0	"	7.25 Sail & Steam
9	7	4	E.N.E ½ E	"	2		29.62	53	47	46	45	O.C.m	"	0	"	
10	7	0	" "	N. E.	1		29.63	53	47	47	45	O.C	"	0	"	
11	7	4	" "	Calm	0		29.64	55	47	47	43	O.C. [illegible]	"	0	"	
Noon.	25	4	E.N.E ½ E East	"	0		"	55	47	47	43	" "	"	0	"	

Data Analysis

A multidisciplinary field



Data Analysis

Traditional Stack

- Spreadsheets (e.g.: Excel, OpenOffice)
- RDBMS (e.g.: Oracle, PostgreSQL, MySQL)
- Statistical Packages (e.g.: R, Matlab)
- GIS Packages (e.g.: QGIS)
- Scripting and programming languages (e.g.: Python, Java)
- Libraries

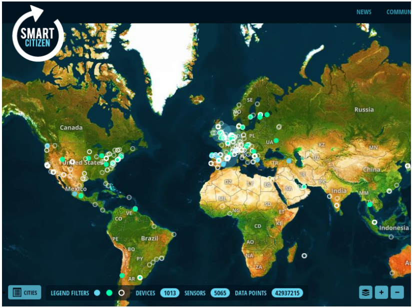
Big Data

What changed in recent years?

- Differences in the way global business and transportation are done have exploded the volume of traditional data sources.
- Widespread increase in data from sensors.

Big Data

Smart Citizen Project

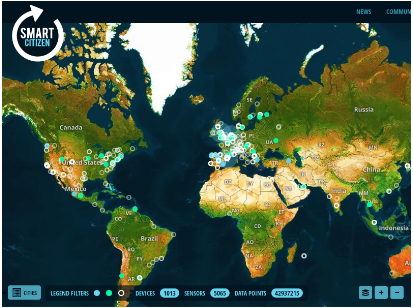


https://smartcitizen.me/

Big Data

Smart Citizen Project

- Platform to generate participatory processes of people in the cities, by connecting data, people and knowledge.

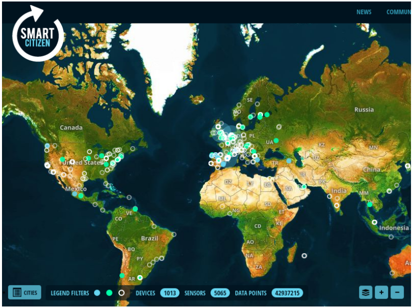


https://smartcitizen.me/

Big Data

Smart Citizen Project

- Platform to generate participatory processes of people in the cities, by connecting data, people and knowledge.
- Based on geolocation, Internet and free hardware and software for data collection and sharing.

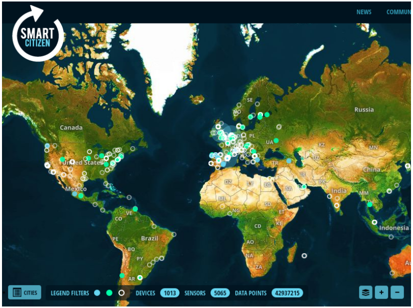


https://smartcitizen.me/

Big Data

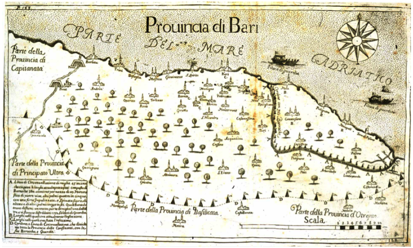
Smart Citizen Project

- Platform to generate participatory processes of people in the cities, by connecting data, people and knowledge.
- Based on geolocation, Internet and free hardware and software for data collection and sharing.
- Relies on the production of objects to connect people with their environment and their city.



`https://smartcitizen.me/`

Big Data

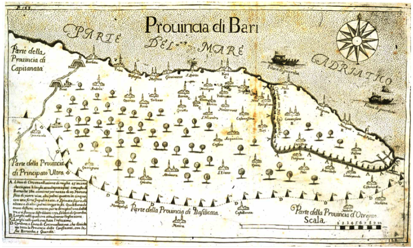


Based on reports, Fillipo Arrieta published maps describing a multi-stage containment plan designed to limit the plague in Bari (1864).

The iLab used the ushahidi platform to collect and display crowd-sorcing information about Ebola in Liberia (2014).

Big Data

- A great deal of this data is actually geo-located (e.g.: satellite navigate coordinates, ip addresses).

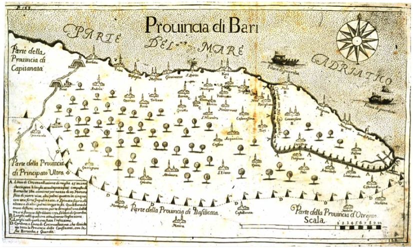


Based on reports, Fillipo Arrieta published maps describing a multi-stage containment plan designed to limit the plague in Bari (1864).

The iLab used the ushahidi platform to collect and display crowd-sorcing information about Ebola in Liberia (2014).

Big Data

- A great deal of this data is actually geo-located (e.g.: satellite navigate coordinates, ip addresses).
- Geography has finally the opportunity to switch from being based on guesses and samples, to become a truly data-driven science.

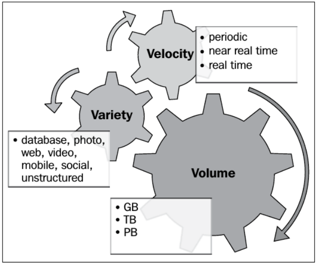


Based on reports, Fillipo Arrieta published maps describing a multi-stage containment plan designed to limit the plague in Bari (1864).

The iLab used the ushahidi platform to collect and display crowd-sorcing information about Ebola in Liberia (2014).

Big Data

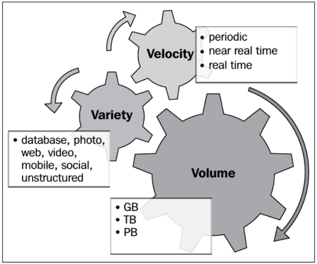
What characterizes Big Data?



Big Data

What characterizes Big Data?

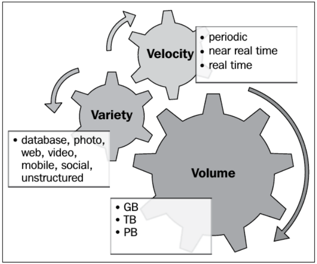
- Volume: Large amounts of data.



Big Data

What characterizes Big Data?

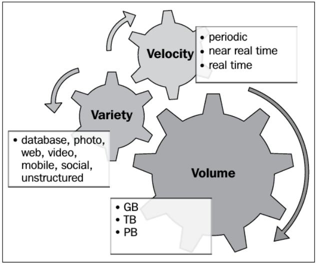
- Volume: Large amounts of data.
- Variety: Different types of structured, unstructured, and multi-structured data.



Big Data

What characterizes Big Data?

- Volume: Large amounts of data.
- Variety: Different types of structured, unstructured, and multi-structured data.
- Velocity: Needs to be analyzed quickly.



Technological Challenges

These characteristics map into challenges:

- Scalability
- Heterogeneity
- Low latency

When the traditional stack is no longer enough, a paradigm shift is required.

RDBMS vs NoSQL

RDBMS vs NoSQL

- NoSQL databases trade away some capabilities of relational databases (SQL), in order to improve scalability.

RDBMS vs NoSQL

- NoSQL databases trade away some capabilities of relational databases (SQL), in order to improve scalability.
- Advantages: NoSQL databases are simpler, can handle semi-structured and denormalized data and have an higher scalability.

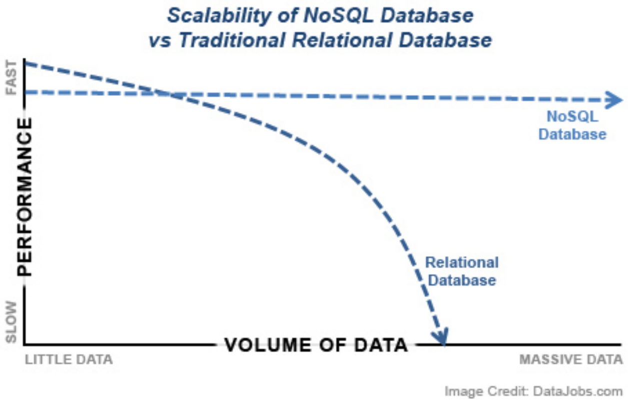
RDBMS vs NoSQL

- NoSQL databases trade away some capabilities of relational databases (SQL), in order to improve scalability.
- Advantages: NoSQL databases are simpler, can handle semi-structured and denormalized data and have an higher scalability.
- Disadvantages: loss of abstraction provided by the query optimizer, that increases the complexity of the applications.

RDBMS vs NoSQL

- NoSQL databases trade away some capabilities of relational databases (SQL), in order to improve scalability.
- Advantages: NoSQL databases are simpler, can handle semi-structured and denormalized data and have an higher scalability.
- Disadvantages: loss of abstraction provided by the query optimizer, that increases the complexity of the applications.
- Recently, tools were developed that bring back the full power of SQL language to the NoSQL ecosystem (e.g.: Apache Drill, Hive).

RDBMS vs NoSQL



Examples

cassandra
redis
CouchDB relax
riak
mongoDB
RAVENDB
OrientDB
symas
Neo4j
hadoop
HAZELCAST SOFTWARE
RethinkDB.
HYPERTABLE INC

MapReduce

Programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

- Map(): performs filtering and sorting.
- Reduce(): performs a summary operation.

MapReduce

Programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

- Map(): performs filtering and sorting.
- Reduce(): performs a summary operation.
- The framework coordinates the processing, by marshalling the distributed servers, running the tasks and parallel and managing all communications and data between the various parts of the system.

MapReduce

Programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

- Map(): performs filtering and sorting.
- Reduce(): performs a summary operation.
- The framework coordinates the processing, by marshalling the distributed servers, running the tasks and parallel and managing all communications and data between the various parts of the system.
- There are many libraries that implement MapReduce.

A Big Data Approach

Does this mean we have to throw away our traditional tools and methods?

A Big Data Approach

Does this mean we have to throw away our traditional tools and methods?

- First ensure that you **really** have, or will have Big Data at some point in the future.

A Big Data Approach

Does this mean we have to throw away our traditional tools and methods?

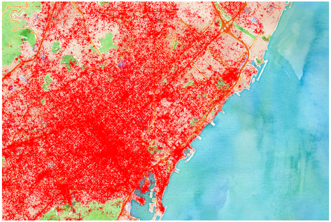
- First ensure that you **really** have, or will have Big Data at some point in the future.
- Then identify the stages of the workflow that are bottlenecks, in terms of the current technologies.

A Big Data Approach

Does this mean we have to throw away our traditional tools and methods?

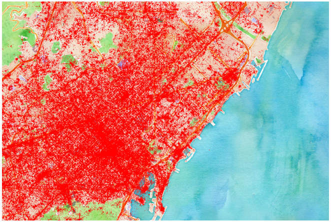
- First ensure that you **really** have, or will have Big Data at some point in the future.
- Then identify the stages of the workflow that are bottlenecks, in terms of the current technologies.
- It is possible to mix & and match.

Analysing geo-located Tweets



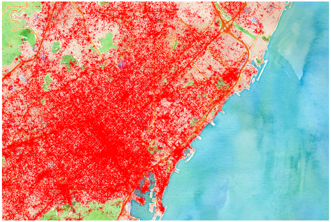
Analysing geo-located Tweets

- The purpose of this use case was to analyse the stream of geo-located Tweets, as a sensor for citizen presence.



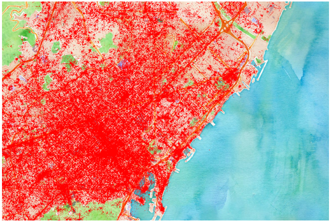
Analysing geo-located Tweets

- The purpose of this use case was to analyse the stream of geo-located Tweets, as a sensor for citizen presence.
- Number of tweets in Catalunya in around 3 months: +- 6 million.



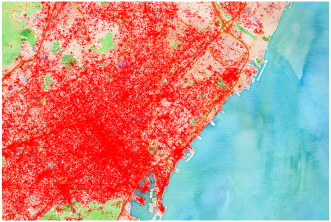
Analysing geo-located Tweets

- The purpose of this use case was to analyse the stream of geo-located Tweets, as a sensor for citizen presence.
- Number of tweets in Catalunya in around 3 months: +- 6 million.
- Continuous stream of data.



Analysing geo-located Tweets

- The purpose of this use case was to analyse the stream of geo-located Tweets, as a sensor for citizen presence.
- Number of tweets in Catalunya in around 3 months: +- 6 million.
- Continuous stream of data.
- This amount of data is not easily assimilated by the "human-eye", so we decided to create clusters of Tweets.



An Use Case

Clustering

An Use Case

Clustering

- It is a descriptive data mining technique, often used for dimensionality reduction.

An Use Case

Clustering

- It is a descriptive data mining technique, often used for dimensionality reduction.
- It groups a set of objects in such a way that objects in the same group are more similar to each other, than to object in other groups.

An Use Case

Clustering

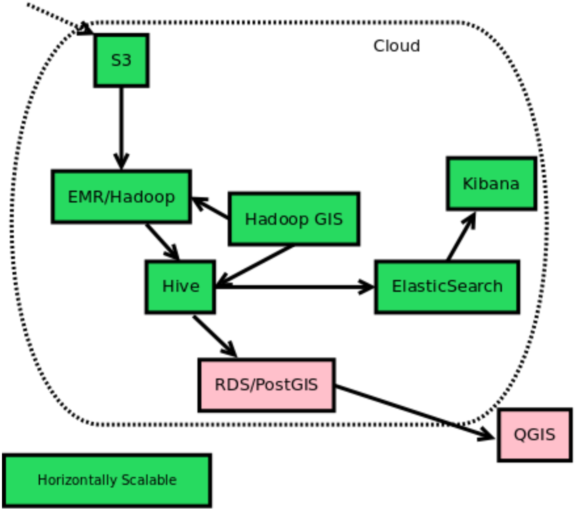
- It is a descriptive data mining technique, often used for dimensionality reduction.
- It groups a set of objects in such a way that objects in the same group are more similar to each other, than to object in other groups.
- Strictly, it corresponds to a family of unsupervised machine learning algorithms.

An Use Case

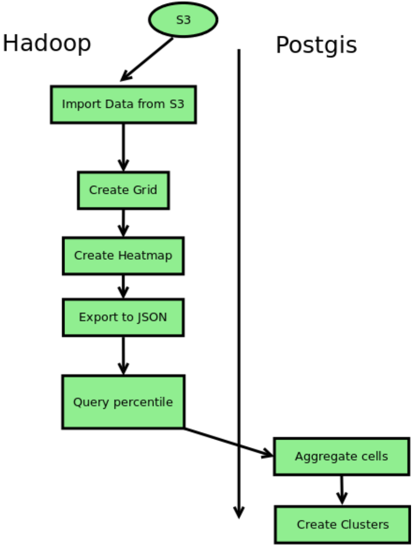
Clustering

- It is a descriptive data mining technique, often used for dimensionality reduction.
- It groups a set of objects in such a way that objects in the same group are more similar to each other, than to object in other groups.
- Strictly, it corresponds to a family of unsupervised machine learning algorithms.
- We wanted to implement this concept using only Hadoop, and apply it to spatial attributes.

Technological Stack

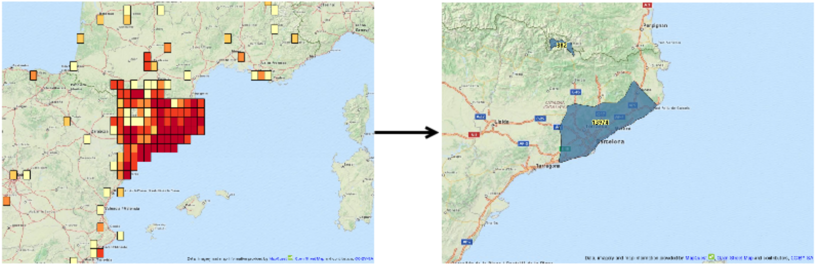


Workflow


S3
Hadoop
Postgis
Import Data from S3
Create Grid
Create Heatmap
Export to JSON
Query percentile
Aggregate cells
Create Clusters

Results

Using this workflow we were able to turn the original raw tweets, first into a density grid, and then into clusters.



As we identified and solved the bottlenecks with the relevant tools, in theory this algorithm is scalable to any Petabytes of data.

Final Remarks

Final Remarks

- It is ok to use non scalable tools, at certain points of the workflow.

Final Remarks

- It is ok to use non scalable tools, at certain points of the workflow.
- We are working on the edge of existing technologies; some functions were not implemented yet, and bugs are common.

Final Remarks

- It is ok to use non scalable tools, at certain points of the workflow.
- We are working on the edge of existing technologies; some functions were not implemented yet, and bugs are common.
- Since it is a niche, this is even more true for spatial technologies.

Final Remarks

- It is ok to use non scalable tools, at certain points of the workflow.
- We are working on the edge of existing technologies; some functions were not implemented yet, and bugs are common.
- Since it is a niche, this is even more true for spatial technologies.
- There are not ready-made solutions: the particular stack and workflow should be compiled for the specific case study.

Final Remarks

- It is ok to use non scalable tools, at certain points of the workflow.
- We are working on the edge of existing technologies; some functions were not implemented yet, and bugs are common.
- Since it is a niche, this is even more true for spatial technologies.
- There are not ready-made solutions: the particular stack and workflow should be compiled for the specific case study.
- this is one stack to solve this problem; it is not the only one, and it may not even be the "best" one;

Acknowledgements

I would like to thank Ellen Friedman (MapR, Apache Mahout, Apache Drill), for her inspiring work on Big Data and Machine Learning.

References

- Cuesta, H. "Practical Data Analysis". Packt Publishing (2013)
- Dunning, T. and Friedman, E. "Time Series Databases: New Ways to Store and Access Data". O'Reilly Media; 1 edition (October, 2014)
- Myatt, G and Johnson, W. "Making Sense Of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining". O'Reilly : 2 edition (2014).

Thank You!

This presentation is available at:
`http://tinyurl.com/pcmgzxp`

Next:

Thank You!

This presentation is available at:
`http://tinyurl.com/pcmgzxp`

Next:

- 9as Jornadas de SIG Libre - 26th-27th March 2015, Girona.