

# Introducció al BigData

Ll. Gesa

Arquitectura i Tecnologies Software

UAB, 2018

## Professor

Lluís Gesa Boté.

Contacte a través del Campus Virtual o **Lluís.Gesa@uab.cat**

## Teoria i problemes

Intentarem donar una teoria amena i participativa entre tots amb exemples pràctics.

Cada setmana veurem algun exemple més concret i detallat.

BigData és Big en sí mateix, donarem breus conceptes per situar-nos.

- 13,20,27 Febrer, 6 Març : Teoria i Problemes
- Conferència sobre BigData
- 9 Març: Pràctica en Laboratori
- 17 Abril: Examen (Juntament amb MongoDB)

# Pràctica

Grups fins a 2 persones. La pràctica es divideix en 2 parts:

- Codificació, a realitzar a casa (75% nota)
- Estudi de dades. Al laboratori tancat (25% nota)

Grups fins a 2 persones. La pràctica es divideix en 2 parts:

- Codificació, a realitzar a casa (75% nota)
- Estudi de dades. Al laboratori tancat (25% nota)

## Codificació

S'actuarà com Data Engineer. Generar una aplicació que compti lletres d'arxiu aplicant l'algoritme Map-Reduce.

A realitzar amb qualsevol llenguatge entre : **Java<sup>tm</sup>**, **Python**, **C**, **C++**, **Perl**, **Ruby**

Enunciat sortirà el 15 Febrer i a entregar el 8 Març a les 23h59m.

El codi ha de ser compilable des de consola linux sense ajuda de IDEs ni entorns de desenvolupaments més enllà dels compiladors.

L'entrega ha d'anar acompanyat d'informe explicant codi, manual d'ús i report de resultats.

Grups fins a 2 persones. La pràctica es divideix en 2 parts:

- Codificació, a realitzar a casa (75% nota)
- Estudi de dades. Al laboratori tancat (25% nota)

## Codificació

S'actuarà com Data Engineer. Generar una aplicació que compti lletres d'arxiu aplicant l'algoritme Map-Reduce.

A realitzar amb qualsevol llenguatge entre : **Java<sup>tm</sup>**, **Python**, **C**, **C++**, **Perl**, **Ruby**

Enunciat sortirà el 15 Febrer i a entregar el 8 Març a les 23h59m.

El codi ha de ser compilable des de consola linux sense ajuda de IDEs ni entorns de desenvolupaments més enllà dels compiladors.

L'entrega ha d'anar acompanyat d'informe explicant codi, manual d'ús i report de resultats.

## Estudi de dades

S'actuarà com Data Scientist. Realitzar un estudi sobre un DataSet fent servir: [www.bigml.com](http://www.bigml.com)

24/48 Hores abans de la sessió és publicarà l'enunciat a fer i entregar durant la sessió.

Cal crear-se un perfil al entorn de BigML.com.

## Articles/Papers

Durant el curs es proporcionaran referències a articles, papers i urls que serviran per completar les explicacions donades. El seu contingut forma part de la 'teoria', per tan, son avaluable.

El problemes en línia amb les pràctiques:

- 13 Febrer: Classe de problemes de MapReduce
- 20 Febrer: Classe de problemes d'anàlisi de dades amb **R**
- 27 Febrer: Introducció al entorn BigML



El problemes en línia amb les pràctiques:

- 13 Febrer: Classe de problemes de MapReduce
- 20 Febrer: Classe de problemes d'anàlisi de dades amb **R**
- 27 Febrer: Introducció al entorn BigML

## Programació amb Llenguatge R

El llenguatge de programació R, és un llenguatge de programació i un entorn de desenvolupament de software per a l'obtenció de càlculs i gràfics estadístics.

De cara a la sessió del dia 20 de Febrer seria recomanable tenir instal·lat l'entorn en els laptops personal. <https://cran.r-project.org/>

Farem un problema d'anàlisi de dades de xarxes socials. Caldrà una compte twitter. 15 Febrer es publicarà un petit manual d'Instal·lació.

# Avaluació

## Problemes

Hi haurà un problema lliure a realitzar INDIVIDUALMENT amb llenguatge R. En cas de tenir més de 4.5 a teoria, la realització del problema (opcional) ajudaria a tenir com a màxim 0.5pt més.

## Pràctiques

La part de codificació puntua 75% sobre la nota de pràctiques. La sessió d'anàlisi de dades 25%.

Hi haurà un exercici opcional de la part d'anàlisi a entregar abans d'examen que pot pujar 1pt la nota en cas de tenir més de 4.

Pràctica a validar el dia del exàmen.

## Teoria

L'examen incorporarà preguntes de la Part teòrica com de Problemes.

## Participació

Sempre ajuda...

## Big Data

Ordenar ofertas por:

☐ Fecha de publicación

Palabra clave

Tipo de oferta

☐ Sólo Executive (6)

Fecha

☐ Cualquier fecha

☐ Últimas 24 horas

☒ Últimos 7 días

☐ Últimos 15 días

Provincia

☐ Madrid (126)

☐ Barcelona (75)

☐ Illes Balears (4)

☐ Girona (3)

[Mostrar todas](#)

241 ofertas de trabajo de **big data** encontradas

**Técnico Sistemas Linux - Big Data**

[Data Architecture And technology](#)

Pozuelo De Alarcon | Hace 14h Nueva

Desarrollamos una nueva arquitectura de referencia y aplicaciones de negocio que aceleran y posibilitan una mejor experiencia de cliente, adaptando los productos del banco a la gente rea...

Contrato no especificado | Jornada completa | 24.000€ - 36.000€ Bruto/año


**Big Data Project Manager - everis Data Innovation**

[everis Ofertas de empleo Profesionales](#)


Madrid | Hace 34m Nueva


Would you like to join us ? We offer you a career plan, to be involved in innovative projects with top customers in EMEA, a training and certification plan, access to entrepreneur networks and ...

Contrato indefinido | Jornada completa | Salario no especificado




an NTT DATA Company





ITALIAN FASHION



Data Engineer (SQL / Python / Spark / Big Data)

## PostGrau

```
http://www.uab.cat/web/postgrau/  
diplomatura-de-postgrau-en-processament-big-data-per-a-ciencies-de-la-vida/  
informacio-general-1203328491238.html/param1-3695_ca/param2-2008/
```

- 1 Introducció BigData
- 2 Estructures i Plataformes BigData
  - Exemple: Hadoop
  - Exemple: Spark
  - Mes enllà de Hadoop i Spark
- 3 Enginyeria de software
  - Algoritmes / Patrons
    - MapReduce, cloud dataflow
  - Models / Estructures
  - Seguretat: Robustesa, coherencia, integritat, accesibilitat
- 4 IoT
- 5 Data mining / Machine Learning / Deep Learning / Predictive analytics
  - Exemple: BigML
  - Data scientists

- 1 Introducció BigData
- 2 Estructures i Plataformes BigData
  - Exemple: Hadoop
  - Exemple: Spark
  - Mes enllà de Hadoop i Spark
- 3 Enginyeria de software
  - Algoritmes / Patrons
    - MapReduce, cloud dataflow
  - Models / Estructures
  - Seguretat: Robustesa, coherencia, integritat, accesibilitat
- 4 IoT
- 5 Data mining / Machine Learning / Deep Learning / Predictive analytics
  - Exemple: BigML
  - Data scientists

# Que és Bigdata?

# Que és Bigdata?

## Wikipèdia

“Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate”



# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?

# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes

# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes
  - 2006 – > 160 Petabytes ( $10^{15}$ )

# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes
  - 2006 – > 160 Petabytes ( $10^{15}$ )
  - 2012 – > 4.5 Exabytes ( $10^{18}$ )

# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes
  - 2006 – > 160 Petabytes ( $10^{15}$ )
  - 2012 – > 4.5 Exabytes ( $10^{18}$ )
  - 2020 – > 44 Zettabytes ( $10^{21}$ )

# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes
  - 2006 – > 160 Petabytes ( $10^{15}$ )
  - 2012 – > 4.5 Exabytes ( $10^{18}$ )
  - 2020 – > 44 Zettabytes ( $10^{21}$ )
- Quant és un zettabyte?

# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes
  - 2006 – > 160 Petabytes ( $10^{15}$ )
  - 2012 – > 4.5 Exabytes ( $10^{18}$ )
  - 2020 – > 44 Zettabytes ( $10^{21}$ )
- Quant és un zettabyte?
  - 1,000,000,000,000,000,000,000 bytes

# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes
  - 2006 – > 160 Petabytes ( $10^{15}$ )
  - 2012 – > 4.5 Exabytes ( $10^{18}$ )
  - 2020 – > 44 Zettabytes ( $10^{21}$ )
- Quant és un zettabyte?
  - 1,000,000,000,000,000,000,000 bytes
  - Una filera de 1TB hard disks de 25,400 km llarg



# Que és Bigdata?

## Algunes xifres Un oceà de bytes!

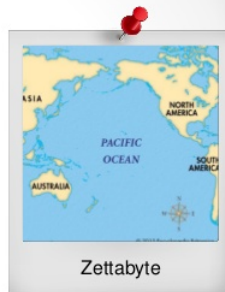
- Quantes

- 200
- 200
- 201
- 202

- Quant é

- 1,00
- Un

Byte : one grain of rice  
Kilobyte : cup of rice  
Megabyte : 8 bags of rice  
Gigabyte : 3 Semi trucks  
Terabyte : 2 Container Ships  
Petabyte : Blankets Manhattan  
Exabyte : Blankets west coast states  
Zettabyte : Fills the Pacific Ocean



# Que és Bigdata?

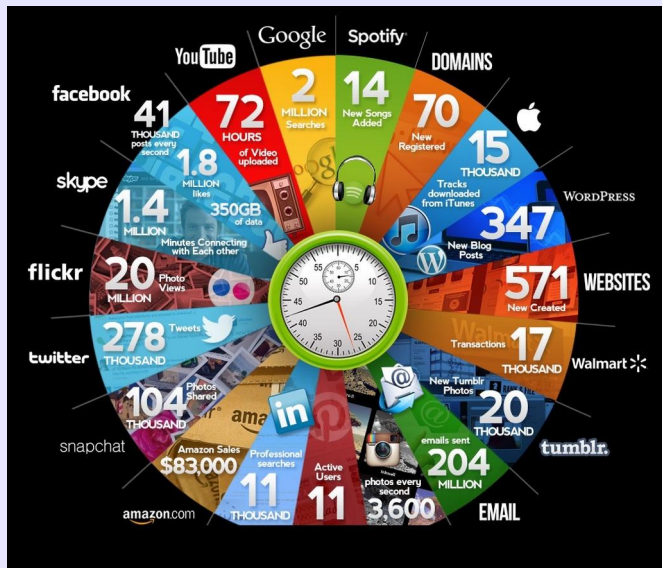
## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes
  - 2006 – > 160 Petabytes ( $10^{15}$ )
  - 2012 – > 4.5 Exabytes ( $10^{18}$ )
  - 2020 – > 44 Zettabytes ( $10^{21}$ )
- Quant és un zettabyte?
  - 1,000,000,000,000,000,000,000 bytes
  - Una filera de 1TB hard disks de 25,400 km llarg
- Quantes dades es generen en un dia?
  - Twitter 7TB
  - Facebook 10TB

# Que és Bigdata Un minut d'activitat

## Algunes xifres

- Quantes dades
  - 2000 — >
  - 2006 — >
  - 2012 — >
  - 2020 — >
- Quant és un z
  - 1,000,000
  - Una filera
- Quantes dades
  - Twitter 7
  - Facebook

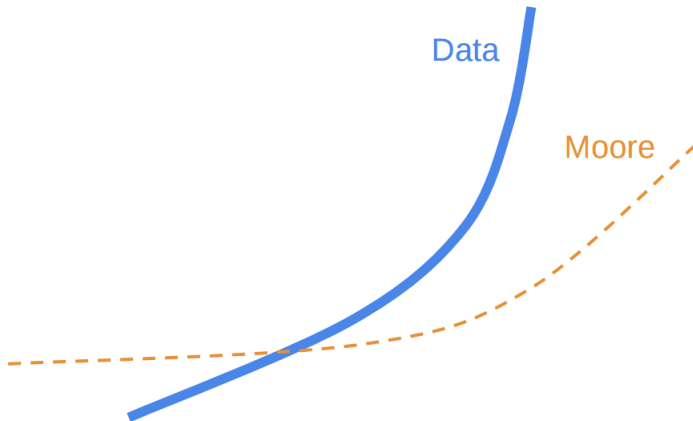


# Que és Bigdata?

## Algunes xifres

- Quantes dades hi ha al mon?
  - 2000 – > 800 Terabytes
  - 2006 – > 160 Petabytes ( $10^{15}$ )
  - 2012 – > 4.5 Exabytes ( $10^{18}$ )
  - 2020 – > 44 Zettabytes ( $10^{21}$ )
- Quant és un zettabyte?
  - 1,000,000,000,000,000,000,000 bytes
  - Una filera de 1TB hard disks de 25,400 km llarg
- Quantes dades es generen en un dia?
  - Twitter 7TB
  - Facebook 10TB
- El 90% de les dades generades en els 2 últims anys!

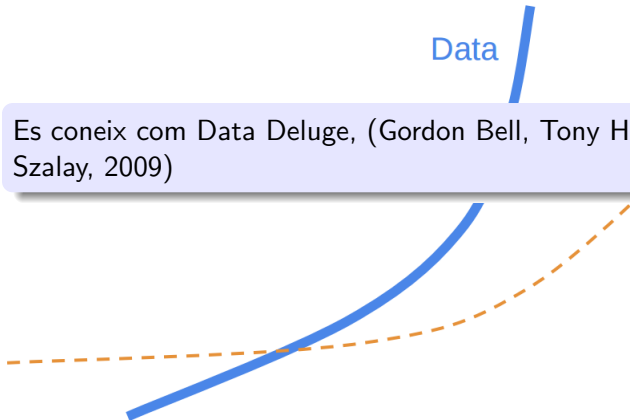
## Data vs Moore's Law



## Data vs Moore's Law

Data

Es coneix com Data Deluge, (Gordon Bell, Tony Hey, Alex Szalay, 2009)



Quan va començar tot

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.



## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria (III a.C). Ptolomeu II amb Zenodoto. 900mil papirs

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria (III a.C). Ptolomeu II amb Zenodoto. 900mil papirs
- 1450 la impremta de Gutenberg.

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria (III a.C). Ptolomeu II amb Zenodoto. 900mil papirs
- 1450 la impremta de Gutenberg.
- 1800-1940. Targetes punxades, dades binaries. 1911 (IBM)

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria (III a.C). Ptolomeu II amb Zenodoto. 900mil papirs
- 1450 la impremta de Gutenberg.
- 1800-1940. Targetes punxades, dades binaries. 1911 (IBM)
- 1940-1970. Cintes magnètiques, base de dades NDBMS

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexar NDBMS (Network Database Management System) 50mil papirs
- 1450 la impremta
- 1800-1940. Targetes punxades, dades binaries. 1911 (IBM)
- 1940-1970. Cintes magnètiques, base de dades NDBMS

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria (III a.C). Ptolomeu II amb Zenodoto. 900mil papirs
- 1450 la impremta de Gutenberg.
- 1800-1940. Targetes punxades, dades binaries. 1911 (IBM)
- 1940-1970. Cintes magnètiques, base de dades NDBMS
- 1980. Base de dades Relacionals. ACID. Client/Server i computació paral.lel.

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria ACID (Atomicity, Consistency, Isolation, Durability) 50mil papirs
- 1450 la impremta
- 1800-1940. Targetes punxades, dades binaries. 1911 (IBM)
- 1940-1970. Cintes magnètiques, base de dades NDBMS
- 1980. Base de dades Relacionals. ACID. Client/Server i computació paral.lel.

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria (III a.C). Ptolomeu II amb Zenodoto. 900mil papirs
- 1450 la impremta de Gutenberg.
- 1800-1940. Targetes punxades, dades binaries. 1911 (IBM)
- 1940-1970. Cintes magnètiques, base de dades NDBMS
- 1980. Base de dades Relacionals. ACID. Client/Server i computació paral.lel.
- 1990-2000: Internet



## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria (III a.C). Ptolomeu II amb Zenodoto. 900mil papirs
- 1450 la impremta de Gutenberg.
- 1800-1940. Targetes punxades, dades binaries. 1911 (IBM)
- 1940-1970. Cintes magnètiques, base de dades NDBMS
- 1980. Base de dades Relacionals. ACID. Client/Server i computació paral.lel.
- 1990-2000: Internet
- 1996 Sergey Brin y Lawrence Page (The Anatomy of a Large-Scale Hypertextual Web Search Engine)

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria 700 mil papirs
- 1450 la impremta
- 1800-1940. Targetes perforades
- 1940-1970. Cinta magnètica
- 1980. Base de dades Relacionals. ACID. Client/Server i computació paral·lela.
- 1990-2000: Internet
- 1996 Sergey Brin y Lawrence Page (The Anatomy of a Large-Scale Hypertextual Web Search Engine)

### Introdueixen 3 conceptes

- PageRank
- Google File System (2003)
- MapReduce (2004)

## Quan va començar tot

- 4000 B.C . Apareix l'escriptura, comença l'era d'emmagatzemar informació.
- Biblioteca Alexandria (III a.C). Ptolomeu II amb Zenodoto. 900mil papirs
- 1450 la impremta de Gutenberg.
- 1800-1940. Targetes punxades, dades binaries. 1911 (IBM)
- 1940-1970. Cintes magnètiques, base de dades NDBMS
- 1980. Base de dades Relacionals. ACID. Client/Server i computació paral.lel.
- 1990-2000: Internet
- 1996 Sergey Brin y Lawrence Page (The Anatomy of a Large-Scale Hypertextual Web Search Engine)
- 2005 Apache Hadoop.

En l'expressió Big Data que és important?

- El Big

En l'expressió Big Data que és important?

- El Big
- La Data

## En l'expressió Big Data que és important?

- El Big
- La Data
- Els 2 conceptes

## En l'expressió Big Data que és important?

- El Big
- La Data
- Els 2 conceptes
- Cap dels 2?

## En l'expressió Big Data que és important?

- El Big
- La Data
- Els 2 conceptes
- Cap dels 2?

Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom. - Clifford Stoll



- Volum

## Les 4 V del emmagatzematge de dades

- Volum
- Velocitat

## Les 4 V del emmagatzematge de dades

- Volum
- Velocitat
- Varietat

## Les 4 V del emmagatzematge de dades

- Volum
- Velocitat
- Varietat
- Veracitat

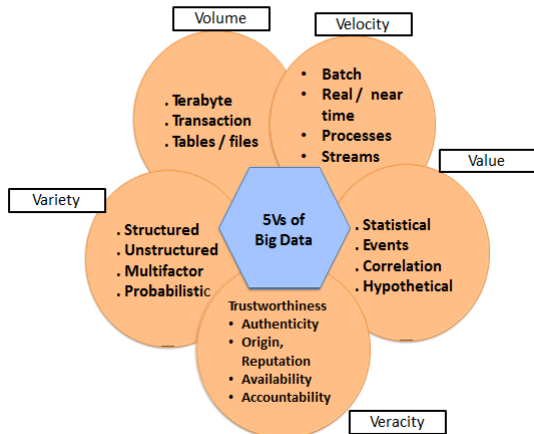
## Les 4 V del emmagatzematge de dades

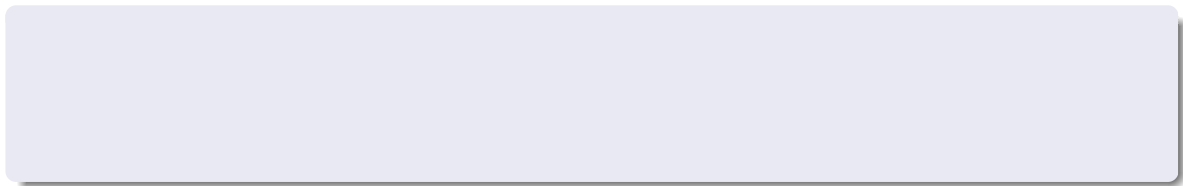
- Volum
- Velocitat
- Varietat
- Veracitat
- **Valor**

- Volum
- Velocitat
- Varietat
- Veracitat
- **Valor**

## 5vs

### 5 'V's of Big Data





- Internet dels continguts



- Internet dels continguts
- Internet de les Persones

- Internet dels continguts
- Internet de les Persones
- Internet de les coses (IoT)

## Libelium Smart World

### Air Pollution

Control of CO<sub>2</sub> emissions of factories, pollution emitted by cars and toxic gases generated in farms.

### Forest Fire Detection

Monitoring of combustion gases and preemptive fire conditions to define alert zones.

### Wine Quality Enhancing

Monitoring soil moisture and trunk diameter in vineyards to control the amount of sugar in grapes and grapevine health.

### Offspring Care

Control of growing conditions of the offspring in animal farms to ensure its survival and health.

### Sportsmen Care

Vital signs monitoring in high performance centers and fields.

### Structural Health

Monitoring of vibrations and material conditions in buildings, bridges and historical monuments.

### Quality of Shipment Conditions

Monitoring of vibrations, strokes, container openings or cold chain maintenance for insurance purposes.

### Smartphones Detection

Detect iPhone and Android devices and in general any device which works with Wifi or Bluetooth interfaces.

### Perimeter Access Control

Access control to restricted areas and detection of people in non-authorized areas.

### Radiation Levels

Distributed measurement of radiation levels in nuclear power stations surroundings to generate leakage alerts.

### Electromagnetic Levels

Measurement of the energy radiated by cell stations and WiFi routers.

### Traffic Congestion

Monitoring of vehicles and pedestrian affluence to optimize driving and walking routes.

### Smart Roads

Warning messages and diversions according to climate conditions and unexpected events like accidents or traffic jams.

### Smart Lighting

Intelligent and weather adaptive lighting in street lights.

### Intelligent Shopping

Getting advices in the point of sale according to customer habits, preferences, presence of allergic components for them or expiring dates.

### Noise Urban Maps

Sound monitoring in bar areas and centric zones in real time.

### Water Leakages

Detection of liquid presence outside tanks and pressure variations along pipes.

### Vehicle Auto-diagnosis

Information collection from CanBus to send real time alarms to emergencies or provide advice to drivers.

### Item Location

Search of individual items in big surfaces like warehouses or harbours.

### Waste Management

Detection of rubbish levels in containers to optimize the trash collection routes.

### Smart Parking

Monitoring of parking spaces availability in the city.

### Golf Courses

Selective irrigation in dry zones to reduce the water resources required in the green.

### Water Quality

Study of water suitability in rivers and the sea for fauna and eligibility for drinkable use.

**libelium**  
www.libelium.com

- Internet dels continguts
- Internet de les Persones
- Internet de les coses (IoT)
- Internet dels llocs

## Conclusions