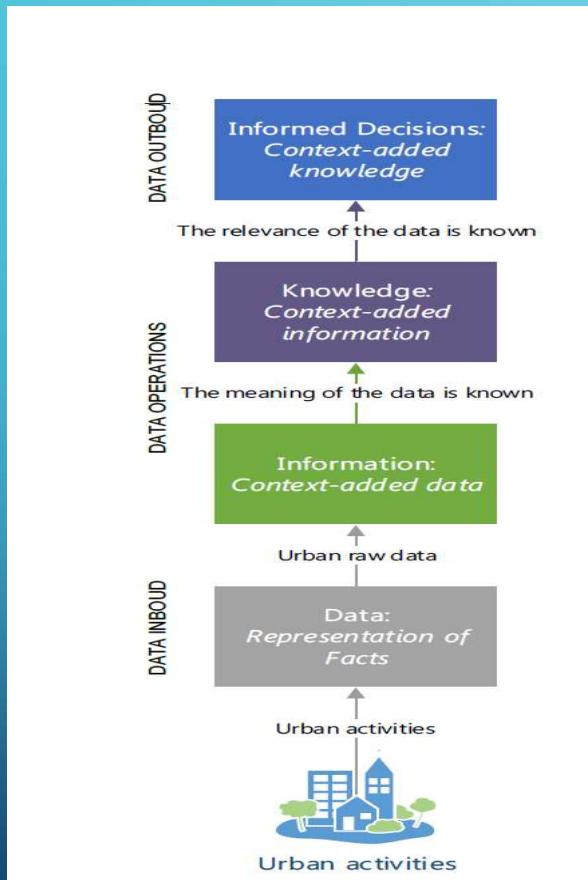


# **Módulo “Análisis de Datos Científicos y Geográficos”**

## **DIVERSAS APLICACIONES PARA CIUDADES INTELIGENTES**

**ITBA**

# CONTEXTO DE CIUDAD INTELIGENTE



Ciudad Inteligente no es sólo tecnología !

Infraestructura  
+  
Stakeholders  
+  
Análisis Profundo  
+  
Toma de Decisiones  
y Acciones

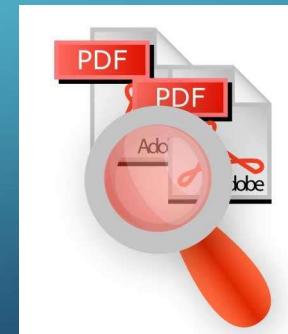
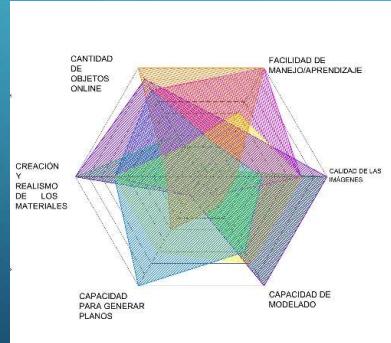
# INTRODUCCIÓN

Casos de Uso sobre investigaciones:

## Simulaciones y Aplicaciones Reales

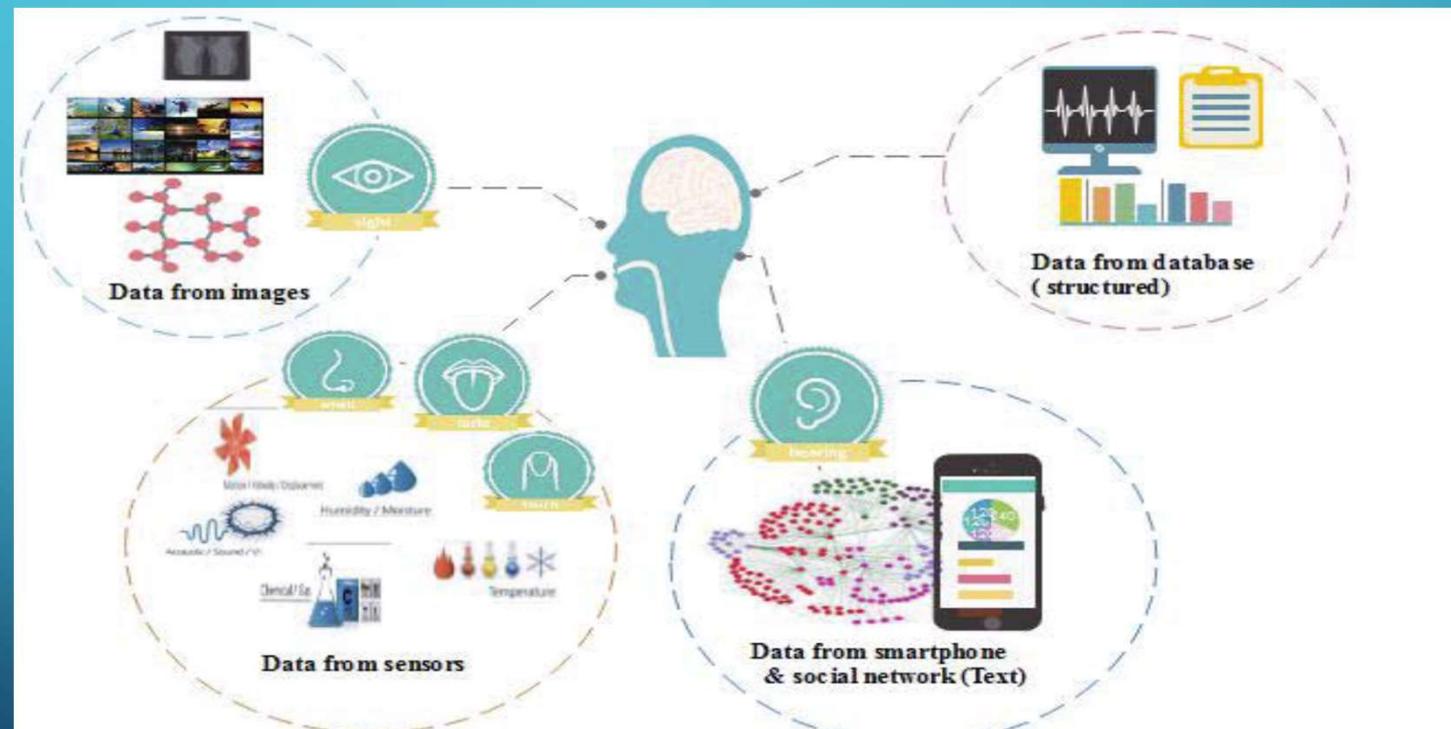
Qué es lo importante?

## El Modelado y el Análisis de los Resultados!



# BIG DATA + APRENDIZAJE INTELIGENTE

Nuevo diseño a partir del principio de cinco sentidos humanos





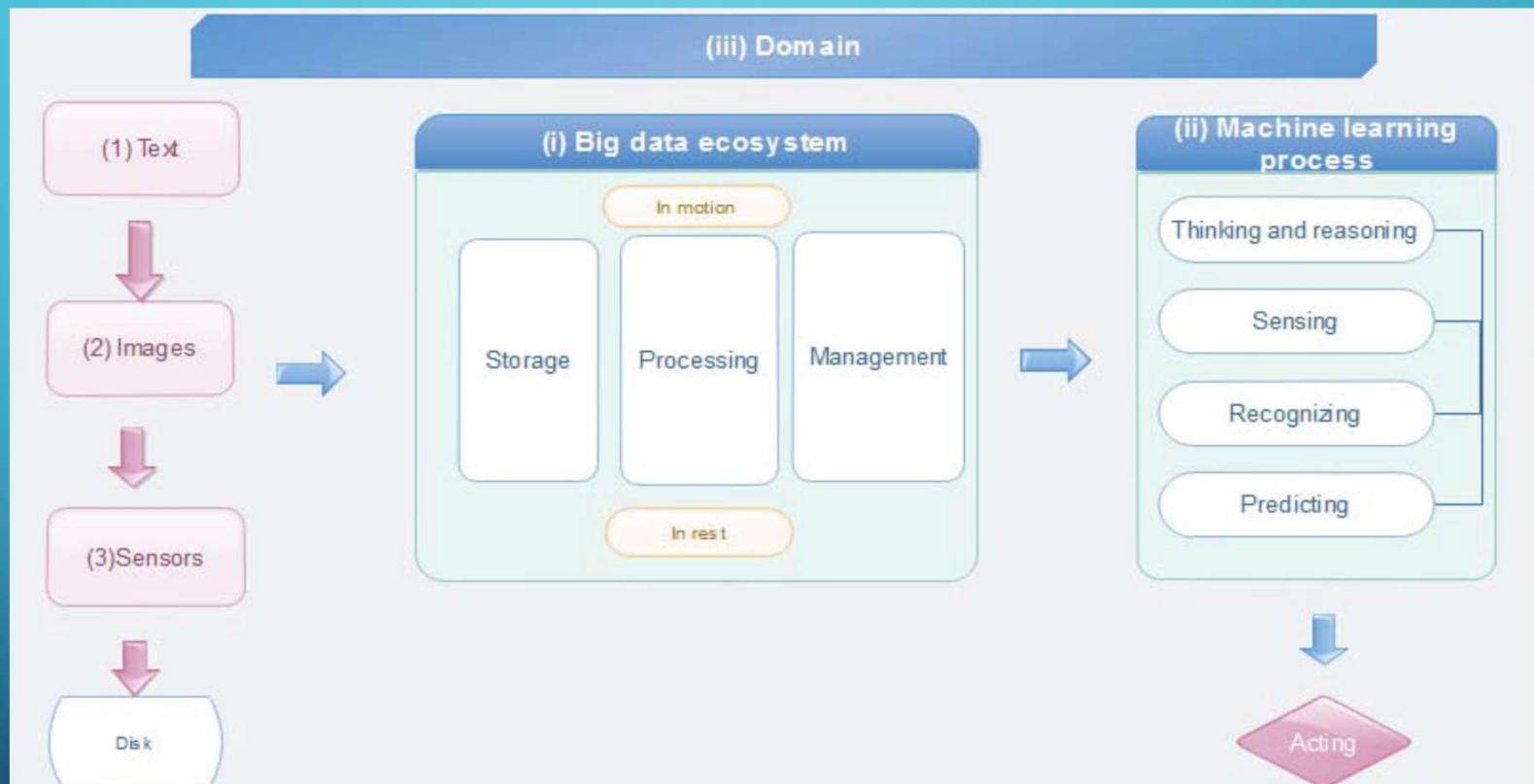
## **BIG DATA + APRENDIZAJE INTELIGENTE**

**El conocimiento del dominio facilita el descubrimiento de patrones ML**

**5V ofrece la posibilidad de información valiosa:**

**Supervisión del comportamiento humano, Sistemas Geográficos de Información, bioinformática, monitoreo de datos de tráfico, etc.**

# BIG DATA + APRENDIZAJE INTELIGENTE



**Big Data + Machine Learning + Domain Platform**

ESPECIALIZACIÓN EN CIENCIA DE DATOS - ANÁLISIS DE DATOS CIENTÍFICOS Y GEOGRÁFICOS

# BIG DATA + APRENDIZAJE INTELIGENTE

Una ciudad inteligente → Tecno + Colaborativo + Inteligencia Artificial (ML)

- Manufactura Inteligente
- Sociedad Inteligente
- Cuidado Inteligente
- Energía Inteligente
- Movilidad Inteligente,
- Trabajo Inteligente,
- Almacenamiento Inteligente
- Entorno Inteligente
- Hogar Inteligente
- Etc.





# UN FRAMEWORK PARA CIUDADES INTELIGENTES

**Dos enfoques para producir grandes datos en una ciudad inteligente:**

- **Técnico-céntrico**
- **Colaborativo**

Ciudad poblada por sensores que recopilan todos los datos para conducir todos los servicios urbanos

+

Monitoreo continuo

datos, servicios y aplicaciones que les permiten a los ciudadanos involucrarse en la administración diaria de la ciudad.



# UN FRAMEWORK PARA CIUDADES INTELIGENTES

**Dos enfoques para producir grandes datos en una ciudad inteligente:**

- Técnico-céntrico
- Colaborativo

**Correlacionar los datos provienen de sensores y detectores con los datos producido por las redes sociales y aplicaciones móviles para:**

- comprender mejor la información
- detectar rápidamente patrones de datos
- hacer predicciones más eficientes

# UN FRAMEWORK PARA CIUDADES INTELIGENTES

Dos enfoques para producir grandes datos en una ciudad inteligente:

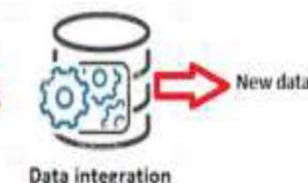
- Técnico-céntrico
- Colaborativo

Ejemplo de correlación entre Twitter y GPS

ID	IMAGE	X_GPS	Y_GPS	speed
12092	AA291A989898C090..	3.92387689900	-8219382392.4	63.4
12092	AA291A989898C090..	4.923829482001	-13893102392.5	66.0
12092	AA291A989898C090..	5.923828733001	2300001392.4	52.4
12092	3AF00000000000001..	3.923871520001	-5109302392.4	63.7
12092	A291A989898C090..	7.9231877120001	-7309302392.4	33.4
12092	AA291A989898C090..	8.923829442001	-2309302392.0	187.8
12092	AA291A989898C090..	8.923829482001	-2309302392.5	44.7

TWEET_ID	content	IMAGE	like
109787	Hello, i like .....	8218009998C090..	1491
382897	Hello World, i like .....	EA291A28000000000..	0385
109882	Phenomeno, .....	AA291A989898002111..	1247
644483	Hi my baby, i like .....	F324C41444441C090..	2223
328739	I guess, i like .....	AA291A989898C090..	492
122376	Information, i like .....	AA291A989898C090..	98300

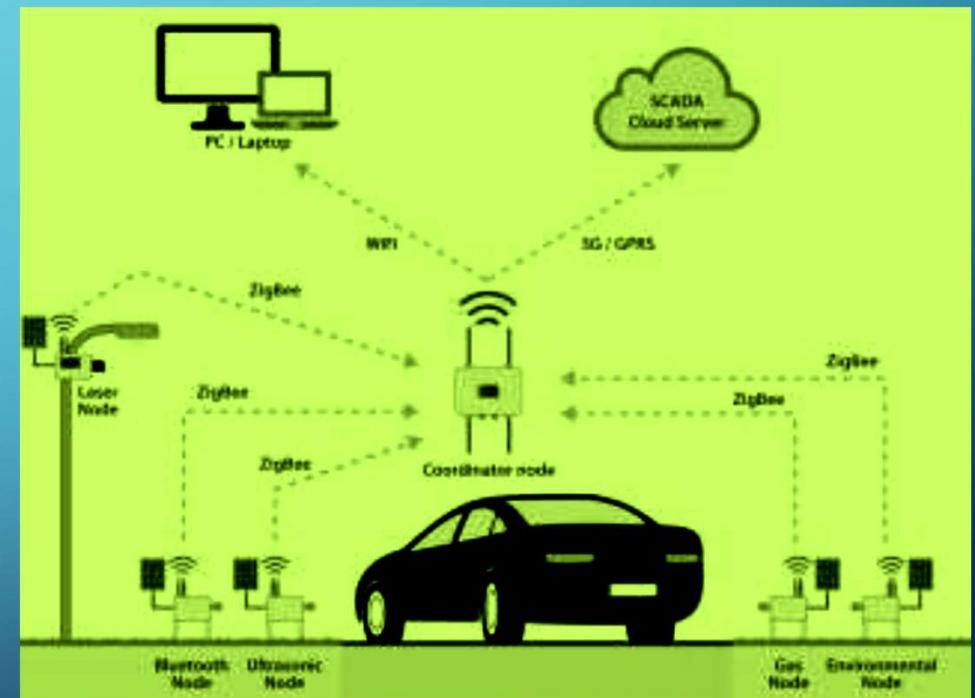


# UN FRAMEWORK PARA CIUDADES INTELIGENTES

La fuente de datos se diversifica:

- Sensores
- Detectores
- GPS
- Tarjetas con chip
- App móviles
- RRSS
- Etc.

**Problema:** Cada fuente usa una tecnología específica (RFID, WiFi, ZigBee, NB-IoT, LoRa, LTE-M) →  
**Desafío en la integración!**



# UN FRAMEWORK PARA CIUDADES INTELIGENTES

Area	Sub-area	Data sources
Mobility and transportation	Traffic	<ul style="list-style-type: none"><li>Connected vehicles</li><li>GPS vehicle location</li><li>RFID chips</li><li>Mobile application (Waze...)</li></ul>
	Parking	<ul style="list-style-type: none"><li>Parking place sensors</li></ul>
	Carpooling	<ul style="list-style-type: none"><li>Mobile applications</li><li>Web applications offer the carpooling services.</li></ul>
Environment and energy	Waste Management <i>(optimize waste collection routes)</i>	<ul style="list-style-type: none"><li>Chip card (<i>each household have a chip card to dispose their garbage</i>).</li><li>Sensors (<i>are placed inside the containers that measure all kinds of data</i>).</li><li>Usage trends data or historical data</li></ul>
	Electricity/Water/Gas consumption	<ul style="list-style-type: none"><li>IoT (An Intelligent Counter)</li><li>Sensors.</li></ul>
	Pollution	<ul style="list-style-type: none"><li>Pollution detectors</li></ul>
	Public lighting	<ul style="list-style-type: none"><li>High Quality Sensors</li></ul>
Security	People security	<ul style="list-style-type: none"><li>Video surveillance systems</li><li>Satellite imagery and data</li></ul>
	Criminal and terrorist activities	<ul style="list-style-type: none"><li>Social media</li><li>Video surveillance systems</li><li>Historic traveling</li></ul>
	Cyber attack	<ul style="list-style-type: none"><li>Transactional applications, logs files, etc.</li></ul>
Healthcare	Healthcare	<ul style="list-style-type: none"><li>Electronic Medical Records - EMR</li><li>Electronic Health Records - EHR</li><li>IoT (Smartphones, Wearable monitors and devices, Telemedicine devices...)</li><li>Health Information Exchanges HIE</li></ul>



# UN FRAMEWORK PARA CIUDADES INTELIGENTES

## Ejemplo en tráfico:

Las carreteras deben ser pobladas con varios sensores que generen datos (velocidad de vehículos, emisiones de CO<sub>2</sub>, contador de vehículos cruzando la calle, ...).

En paralelo, recibir información generada por aplicaciones móviles sobre accidentes, problemas en el camino, etc. (Waze, enfoque colaborativo, ...)



# UN FRAMEWORK PARA CIUDADES INTELIGENTES

Al encontrar correlaciones, se puede lanzar gran cantidad de información útil:

- Calculo dinámico de la duración de un viaje
- Propuestas para caminos alternativos con menos congestión
- Generación de tendencias de congestión vial basadas en datos históricos



La ganancia es enorme:

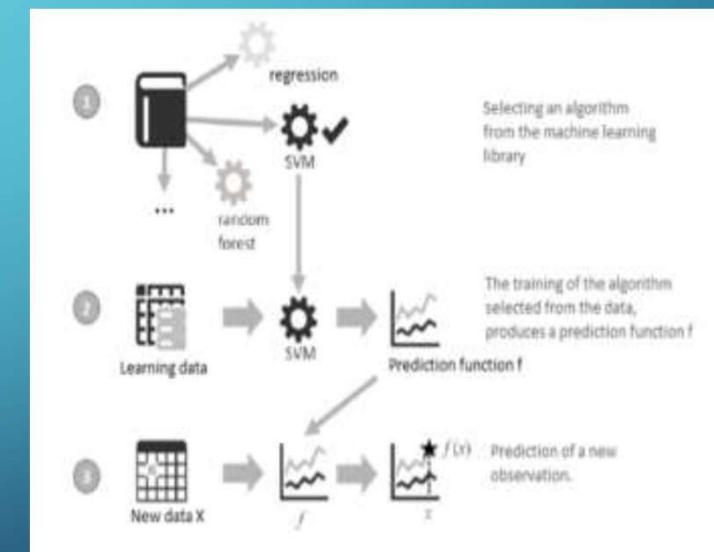
- Menos congestión (tráfico inteligente)
- Menos contaminación (entorno inteligente)
- Menos estrés (gente inteligente)

**Big Data Analytics**  
=>  
**Calidad de Vida**

# UN FRAMEWORK PARA CIUDADES INTELIGENTES

## Aprendizaje Automático:

- **Algoritmo supervisado como Regresión lineal, Maquina Soporte Vectorial(SVM): generalizar el relación entre las variables explicativas ( $X_1, X_2, X_3, \dots$ ) y las variables objetivo ( $y_1, y_2, y_3, \dots$ )**
- **Algoritmos no supervisados como la agrupación con k-means: inferir una función para describir estructuras ocultas de datos no etiquetados**

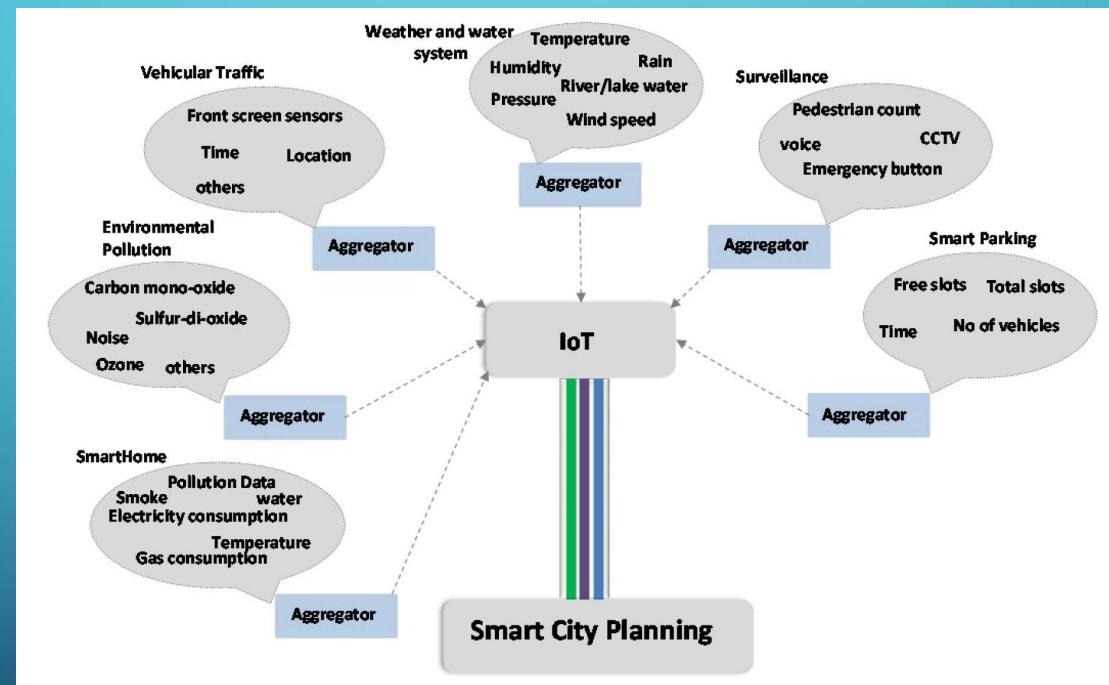


# UN FRAMEWORK PARA CIUDADES INTELIGENTES

<b>Area</b>	<b>Sub Area</b>	<b>Data Sources</b>	<b>Types of data</b>	<b>Real-time analytics (1), or batch processing (2)</b>	<b>Supervised/Unsupervised statistical machine learning algorithms</b>
<b>Mobility and transportation</b>	<b>Traffic</b>	<ul style="list-style-type: none"> <li>Connected vehicles</li> <li>GPS vehicle location</li> <li>RFID chips</li> <li>Mobile application (Waze,...)</li> </ul>	image and video; shipfile; structured relational data	1 and 2 (traffic patterns)	Unsupervised statistical machine learning algorithms
	<b>Parking</b>	<ul style="list-style-type: none"> <li>Anisotropic Magneto-resistive (AMR)</li> </ul>	Unstructured text	1 and 2	Unsupervised machine learning algorithms.
	<b>Carpooling</b>	<ul style="list-style-type: none"> <li>Mobile applications</li> <li>Web applications offer the carpooling services.</li> </ul>	Structured relational data	1 and 2	Supervised & Unsupervised machine learning algorithms.
<b>Environment and energy</b>	<b>Waste Management (optimize waste collection routes).</b>	<ul style="list-style-type: none"> <li>Chip card (each household have a chip card to dispose their garbage).</li> <li>Sensors are placed inside the containers that measure all kinds of data.</li> <li>Usage trends data for historical data</li> </ul>	Unstructured text Structured relational data XML	1 and 2	Unsupervised machine learning algorithms.
	<b>Electricity/water/gaz consumption</b>	<ul style="list-style-type: none"> <li>IoT (An Intelligent Counter)</li> <li>Sensors.</li> </ul>	Unstructured text Structured relational data XML	2	Unsupervised machine learning algorithms
	<b>Pollution</b>	<ul style="list-style-type: none"> <li>Pollutant detectors</li> </ul>	Relational data	1 and 2	Unsupervised machine learning algorithms
	<b>Public lighting</b>	<ul style="list-style-type: none"> <li>High Quality Sensors</li> </ul>	Relational data	1 and 2	Unsupervised machine learning algorithms.
<b>Security</b>	<b>People security</b>	<ul style="list-style-type: none"> <li>Video Surveillance Systems</li> <li>Satellite Imagery and data</li> </ul>	Image and video	1	Unsupervised machine learning algorithms.
	<b>Criminal and terrorist activities</b>	<ul style="list-style-type: none"> <li>Social media</li> <li>Video Surveillance Systems</li> <li>Historic traveling</li> </ul>	Unstructured text Image and video	1 and 2	Supervised and Unsupervised machine learning algorithms.
	<b>Cyber attack</b>	Transactional Applications, logs files,...	Unstructured text Structured relational data	1 and 2	Unsupervised machine learning algorithms.
<b>Healthcare</b>	<b>Healthcare</b>	<ul style="list-style-type: none"> <li>Electronic Medical Records (EMR)</li> <li>Electronic Health Records (EHR)</li> <li>IoT (Smartphones, Wearable monitors and devices, Telemedicine devices,...)</li> <li>Health Information Exchanges (HIE)</li> </ul>	XML Unstructured text	2	Supervised machine learning algorithms.

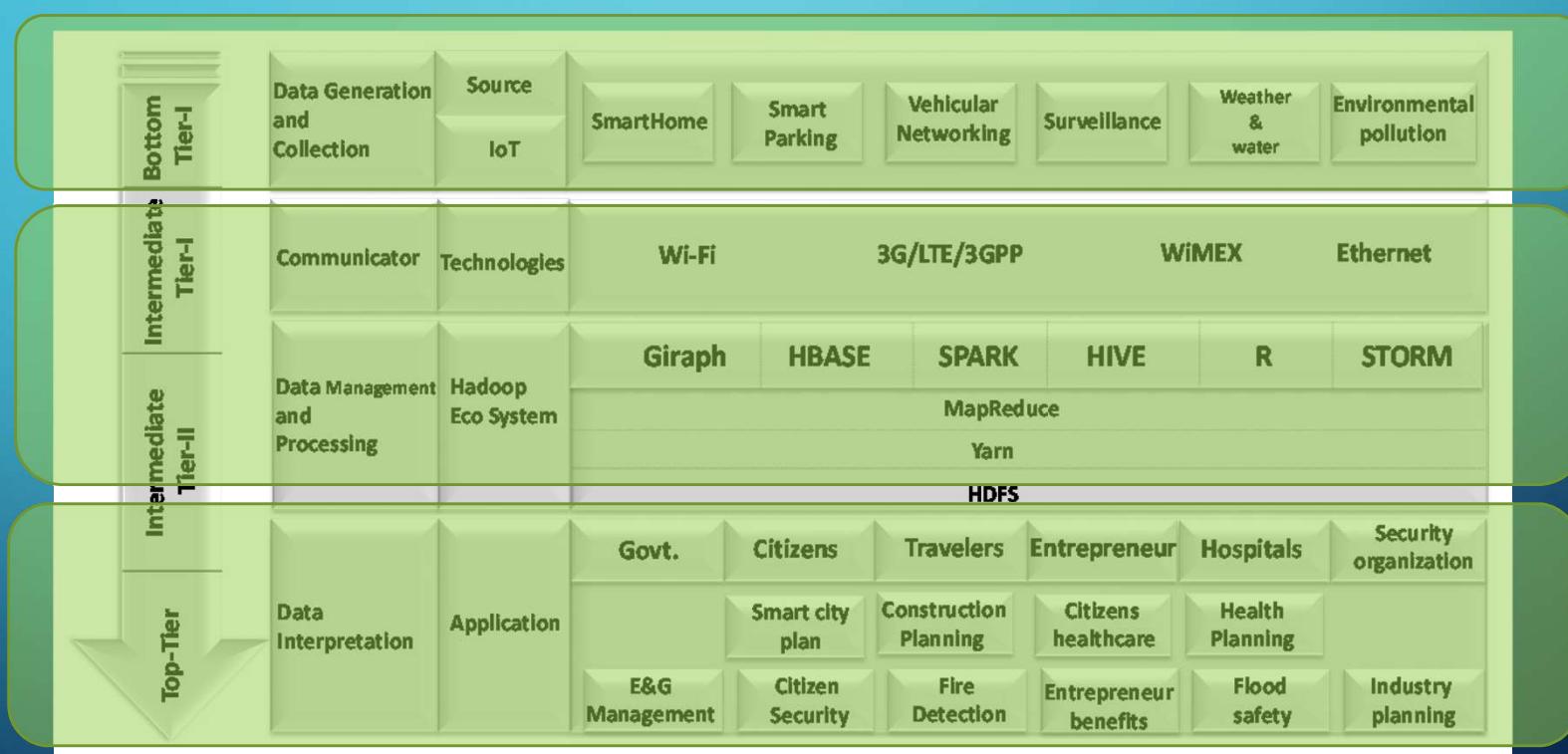
# BIG DATA AND IOT PARA PLANEAMIENTO URBANO

Sistema de generación de datos



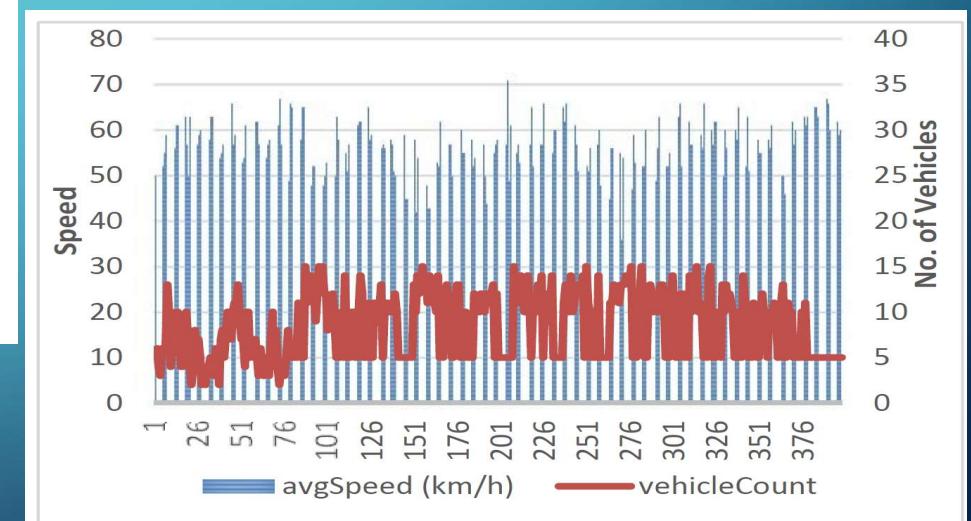
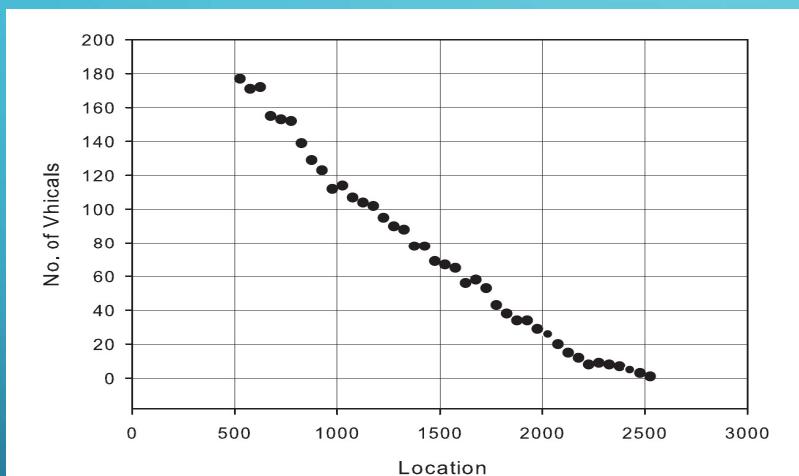
# BIG DATA AND IOT PARA PLANEAMIENTO URBANO

## TRES ÁREAS QUE NO PUEDEN FALTAR



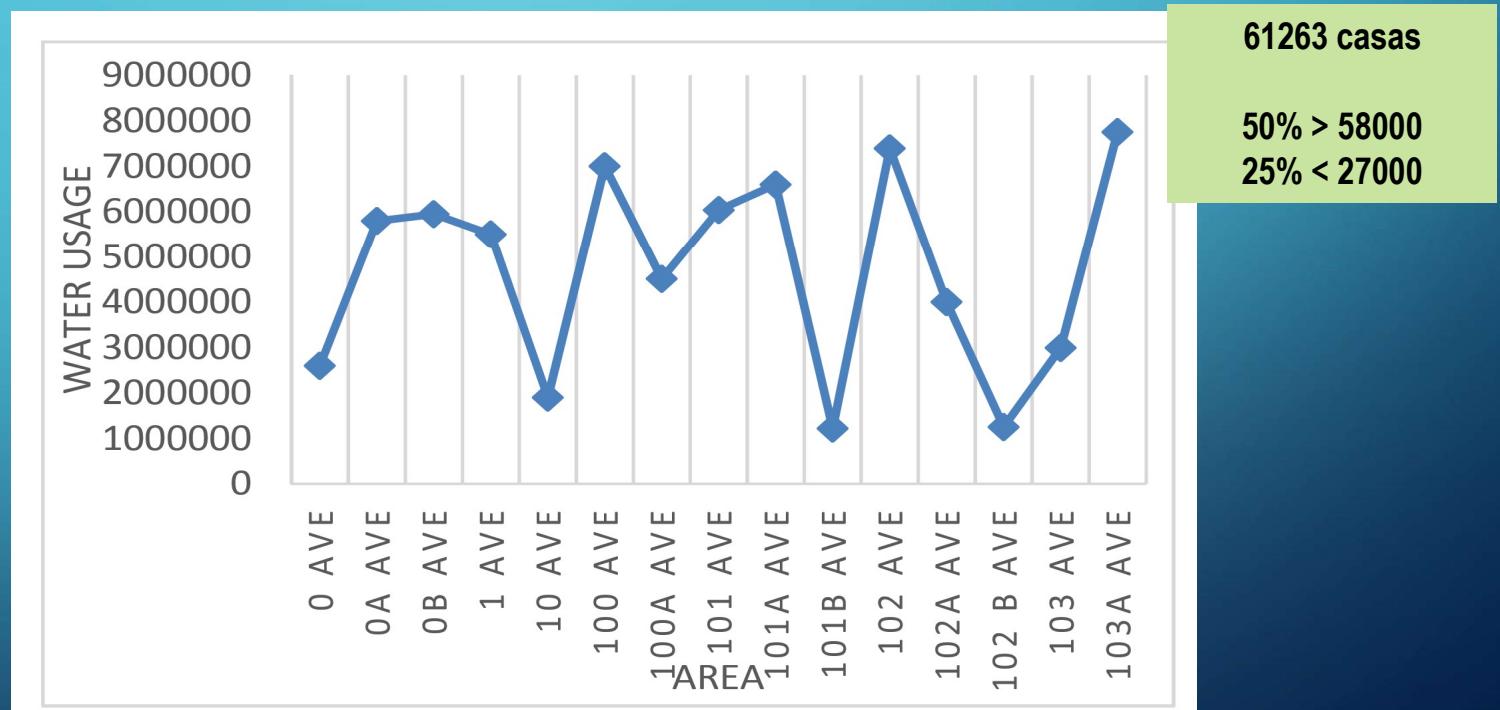
# BIG DATA AND IOT PARA PLANEAMIENTO URBANO

Ejemplo 1: Localización e intensidad de vehículos en Madrid



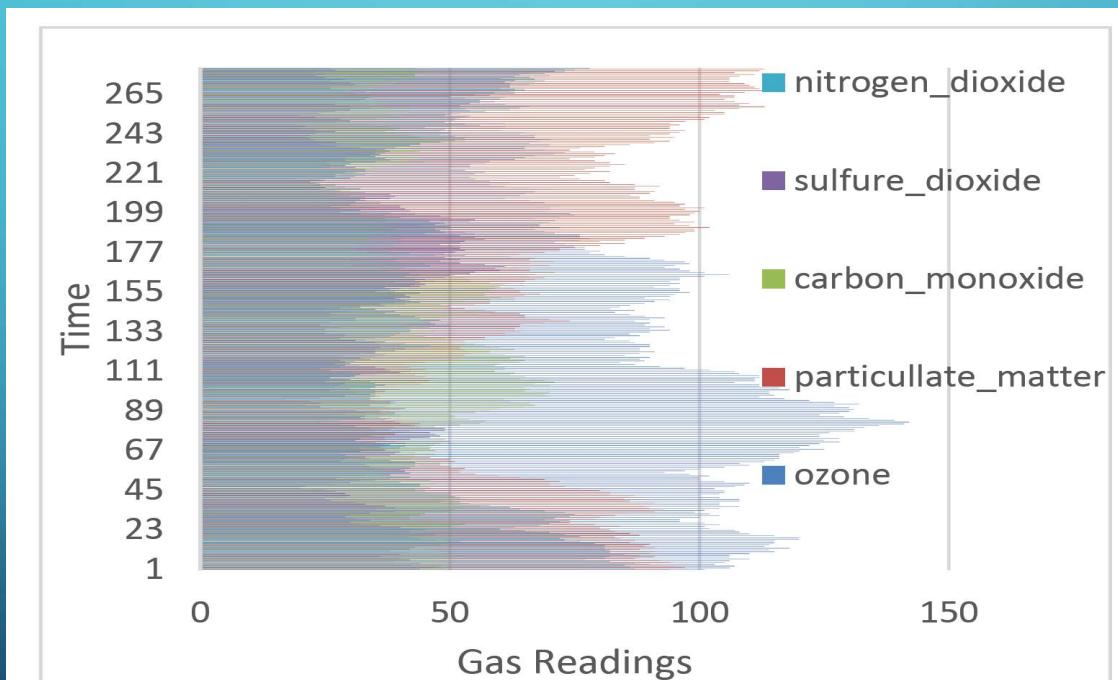
# BIG DATA AND IOT PARA PLANEAMIENTO URBANO

Ejemplo 2: Uso del Agua por Zona en Canada



# BIG DATA AND IOT PARA PLANEAMIENTO URBANO

Ejemplo 3: Nivel de polucion en horas del dia



Aumenta  
polucion por  
vehículos  
particulares  
...

Que acción  
se puede  
disparar?



## Módulo “Análisis de Datos Científicos y Geográficos”

### BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

(Ver paper completo en Campus→Lecturas)



## BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

Tradicionalmente, los sistemas de gestión de bases de datos relacionales (RDBMS) se utilizan para almacenar datos históricos, proporcionando diferentes perspectivas analíticas con respecto a varios procesos comerciales.

Con la corriente avances en las técnicas y tecnologías Big Data, el concepto de Big Data Warehouse (BDW) surge para superar varias limitaciones de DW tradicionales.

### **Base Transaccional vs Warehouse y OLAP**



## BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

Avances en TICs permite que las cosas y los humanos estén constantemente interconectados

En el contexto de Smart Cities, los datos se producen constantemente:

teléfonos inteligentes

medidores inteligentes

sensores de temperatura

sensores de ruido

sensores de ubicación

+

sistemas de bases de datos transaccionales

archivos geoespaciales

datos censales

datos proporcionados por empresas privadas responsables de servicios

IOT  
+  
BigData



## BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

Los DW relacionales se modelan de acuerdo a las técnicas rígidas de modelo de datos →→→ ineficaces / ineficientes para contextos de transmisión y para almacenar grandes cantidades de datos no estructurados o semiestructurados, incluyendo texto, video, imágenes, mapas.

Para grandes volúmenes no estructurados aparecieron nuevas tecnologías:

- Sistema de archivos distribuidos (Hadoop HDFS)
- Procesamiento de datos distribuidos (Hadoop MapReduce, Spark)
- Bases de datos NoSQL (Cassandra, HBase o MongoDB)



## BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

Proyecto SusCity es una pruebas de concepto de un Big Data Warehouse

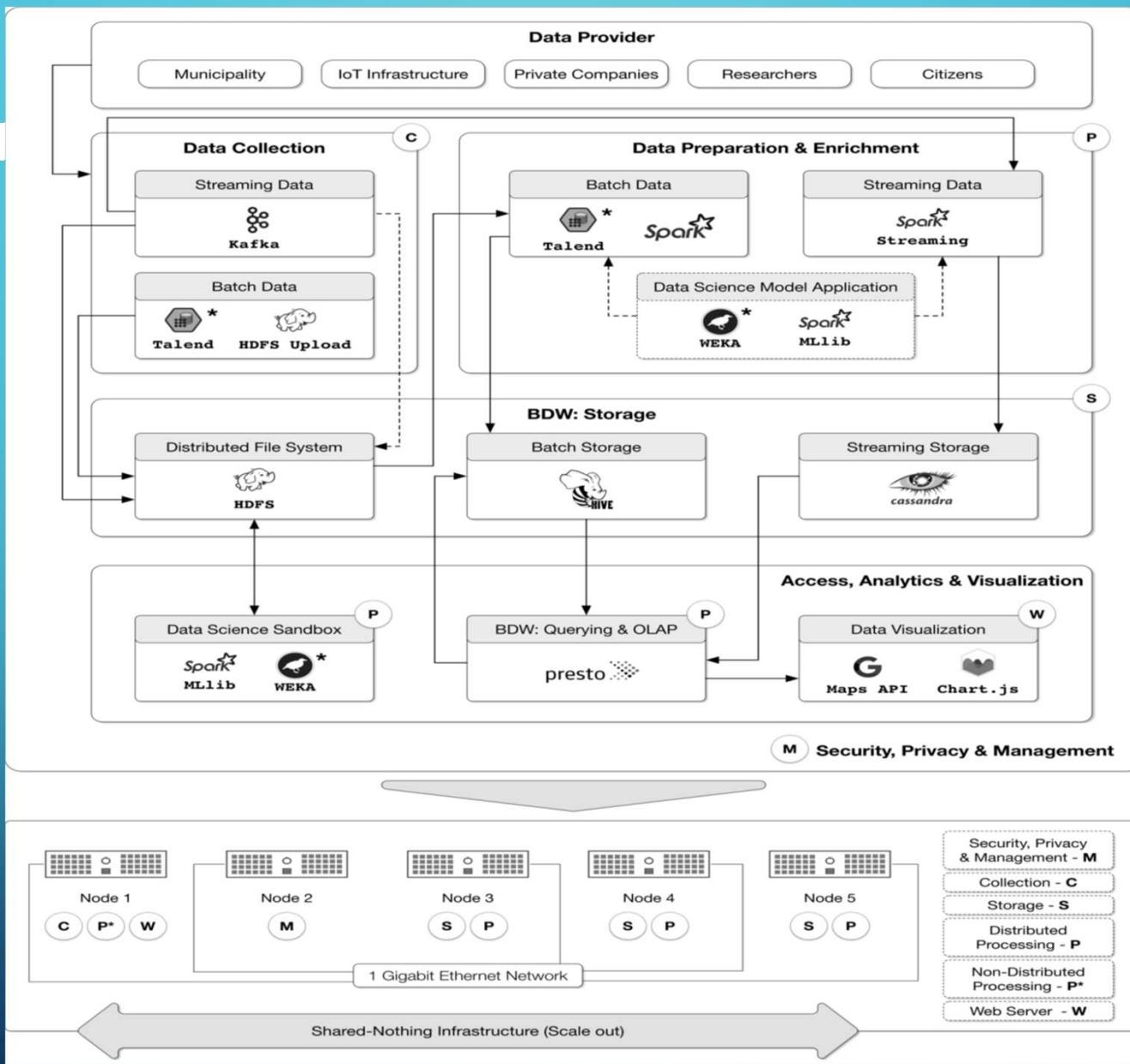
La capa lógica ayuda al investigador y a los profesionales, entendiendo los componentes lógicos del sistema y cómo los datos fluyen a través de estos componentes. Utiliza parte de la taxonomía de Big Data Reference Architecture de la Instituto Nacional de Estándares y Tecnología (NIST)

La capa física se centra en la tecnología utilizada para instanciar los componentes lógicos y la infraestructura en que estas tecnologías se implementan



BI

ELIGENTES





# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

## Proveedores de Datos

- **Municipio:** puede hacer disponible varias fuentes de datos relevantes para tareas analíticas, incluida la información de edificios o representaciones geoespaciales de las infraestructuras de la ciudad. Los sistemas transaccionales de la ciudad también pueden ser fuentes valiosas
- **Infraestructura IoT:** incluye diferentes tipos de sensores para informar el consumo de electricidad, la temperatura, el ruido y patrones de movilidad, por ejemplo. Esta información es significativamente relevante para comprender eventos y patrones en tiempo real en la ciudad



# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

## Proveedores de Datos (cont.)

- **Empresas privadas:** las infraestructuras de la ciudad no son siempre públicas y, por lo tanto, las interacciones con empresas privadas son de gran importancia en Smart Cities, con el fin de para recopilar datos históricos de consumo de energía, edificios certificados, consumo de agua, datos censales, entre muchos otras fuentes de datos
- **Investigadores y ciudadanos:** los proyectos de investigación llevados a cabo en la ciudad son una fuente de datos relevante para el BDW, incluidos los datos de simulación con respecto a diferentes fenómenos en la ciudad Además, los ciudadanos pueden proporcionar datos útiles

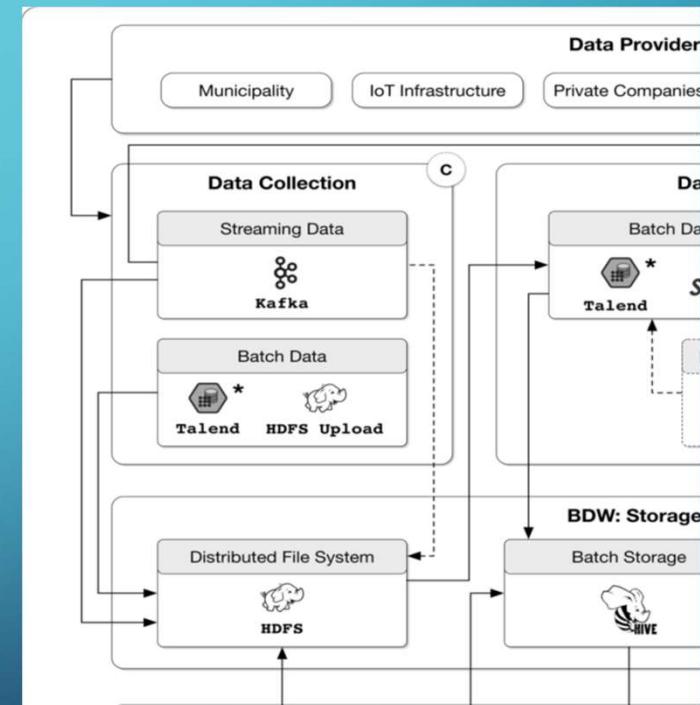
# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

## INPUT

Talend Open Studio para Big Data y un cliente HDFS

Kafka para streaming  
Periódicamente, se mueven a HDFS

Se suben los datos crudos a HDFS, antes de procesarlos.... Por que?



# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

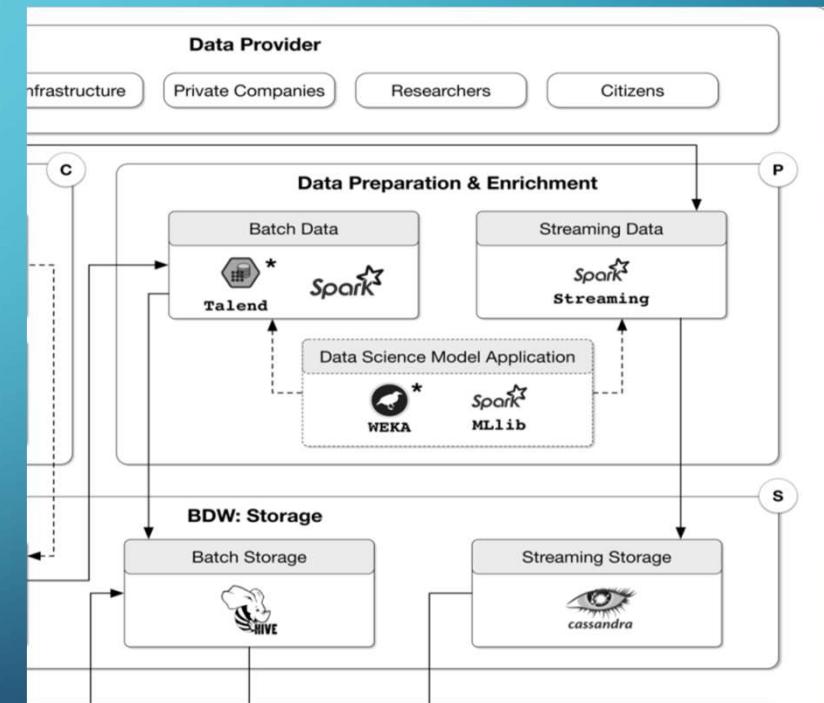
## PROCESAMIENTO (preparación y enriquecimiento)

Conjunto de datos no distribuido → Talen Open Studio para Big Data (filtrado, agregación, combinaciones) y una interfaz gráfica fácil de usar

Conjunto de datos de procesamiento distribuido → Spark

Para escenarios de transmisión → Spark Streaming (procesar datos a medida que llega al sistema BDW)

Datos previamente entrenados los modelos de minería también se pueden aplicar en esta fase,

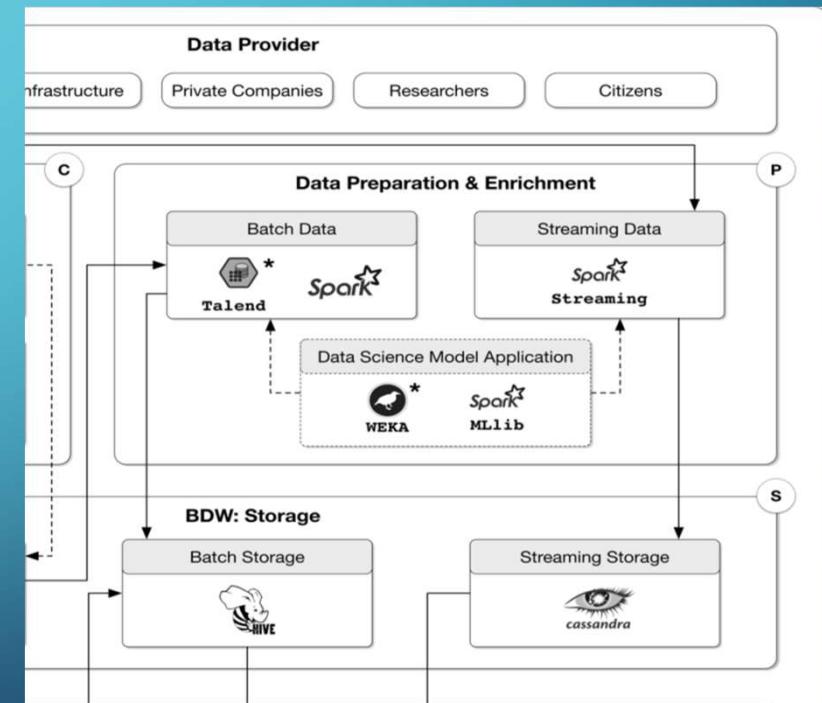


# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

## PROCESAMIENTO (preparación y enriquecimiento)

Para escenarios de transmisión → Spark Streaming (procesar datos a medida que llega al sistema BDW)

Datos previamente entrenados los modelos de minería también se pueden aplicar en esta fase, usando Para datos previamente entrenados WEKA para pequeña escala Spark MLlib para gran escala



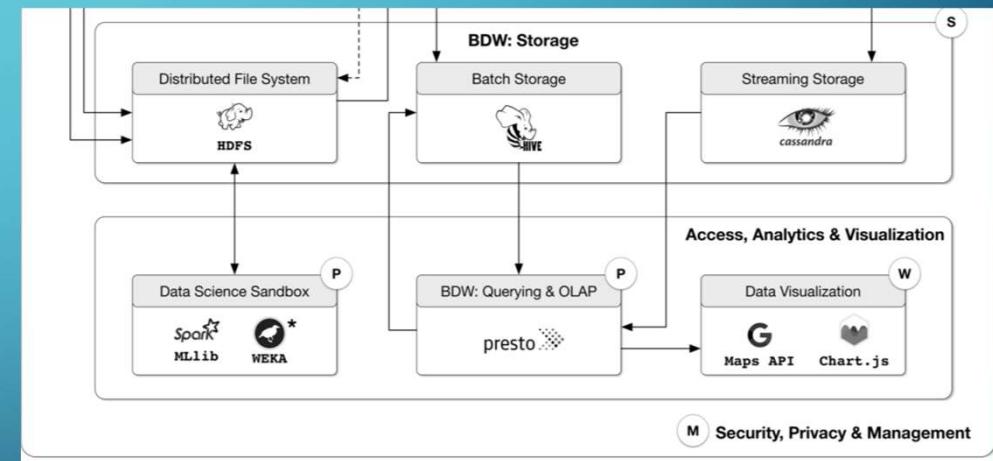
# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

## ALMACENAMIENTO

Batch con HDFS: lotes sin procesar  
(asegura flexibilidad capaz de manejar la variedad de datos de varias fuentes)

Almacenamiento distribuido estructurado con Hive

Almacenamiento streaming no estructurado con Cassandra

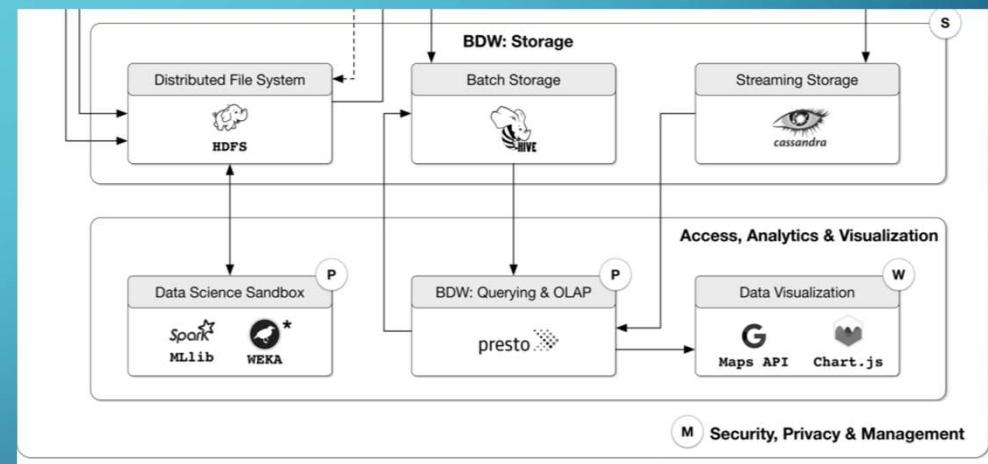


# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

## ANALISIS Y VISUALIZACIÓN

Consulta OLAP con Presto

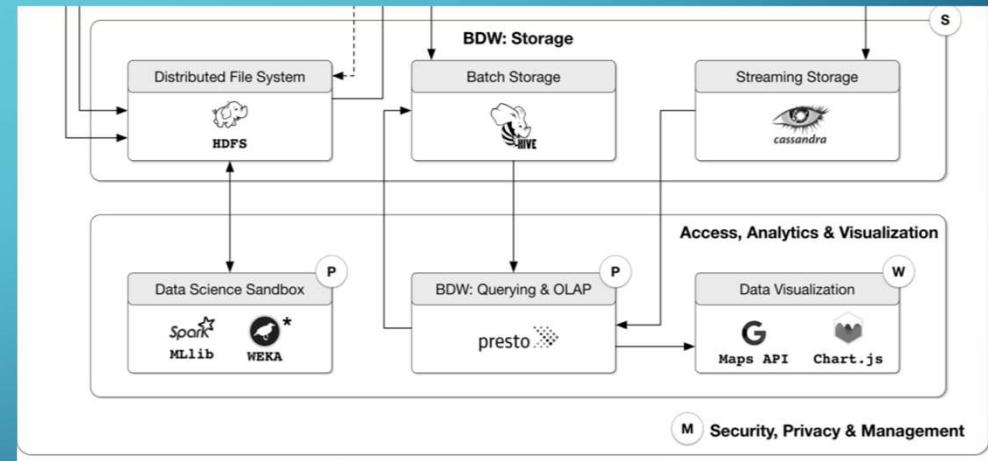
Como Cassandra es menos eficiente que Hive para un acceso secuencial rápido, se usa Presto para transferir datos entre Cassandra y Hive, evitando la acumulación de enormes cantidades de datos históricos en las familias de columnas de Cassandra.



# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

## ANALISIS Y VISUALIZACIÓN

Dado que los datos de lotes sin procesar y de transmisión pueden ser almacenados en HDFS, los científicos de datos pueden interactuar con estos datos para producir modelos capaces de extraer patrones y hacer predicciones cuando llegan nuevos datos a la preparación y componente de enriquecimiento WEKA y Spark son las fuerzas impulsoras para este propósitos



# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

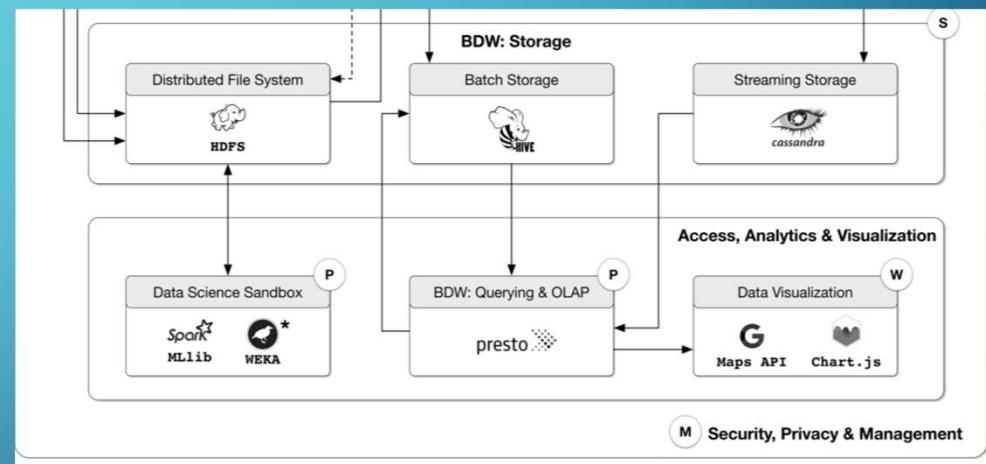
## SEGURIDAD

Kerberos es utilizado para autenticación segura en Hadoop.

Ranger se usa para desplegar rigurosas políticas de autorización, que definen a qué usuarios tienen acceso a ciertos archivos o tablas.

En Cassandra, se usa cifrado TLS / SSL para cliente-nodo o nodo-nodo.

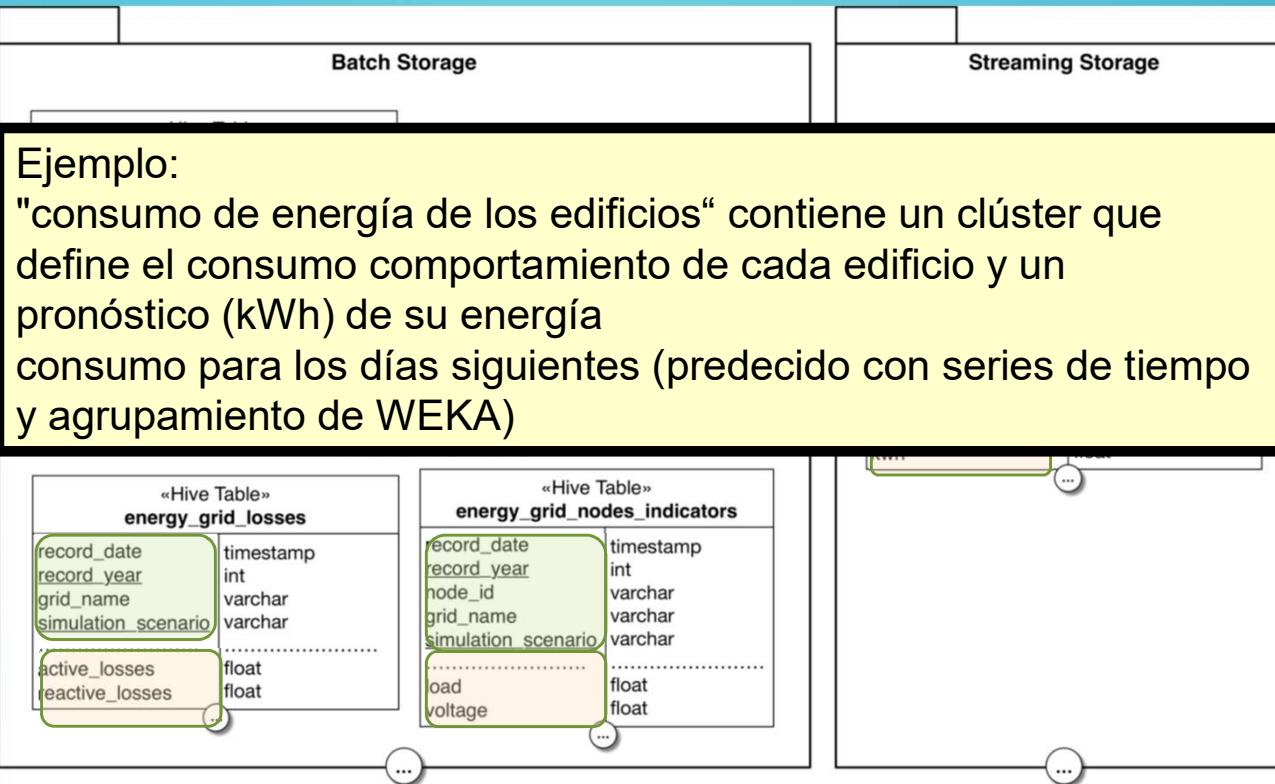
Ambari se usa para administrar / monitorear los componentes de Hadoop



En un contexto de Smart Cities, la privacidad de datos es una preocupación principal:  
Anonimización de datos confidenciales antes de almacenarlo en el BDW

# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

Atributos  
descriptivos

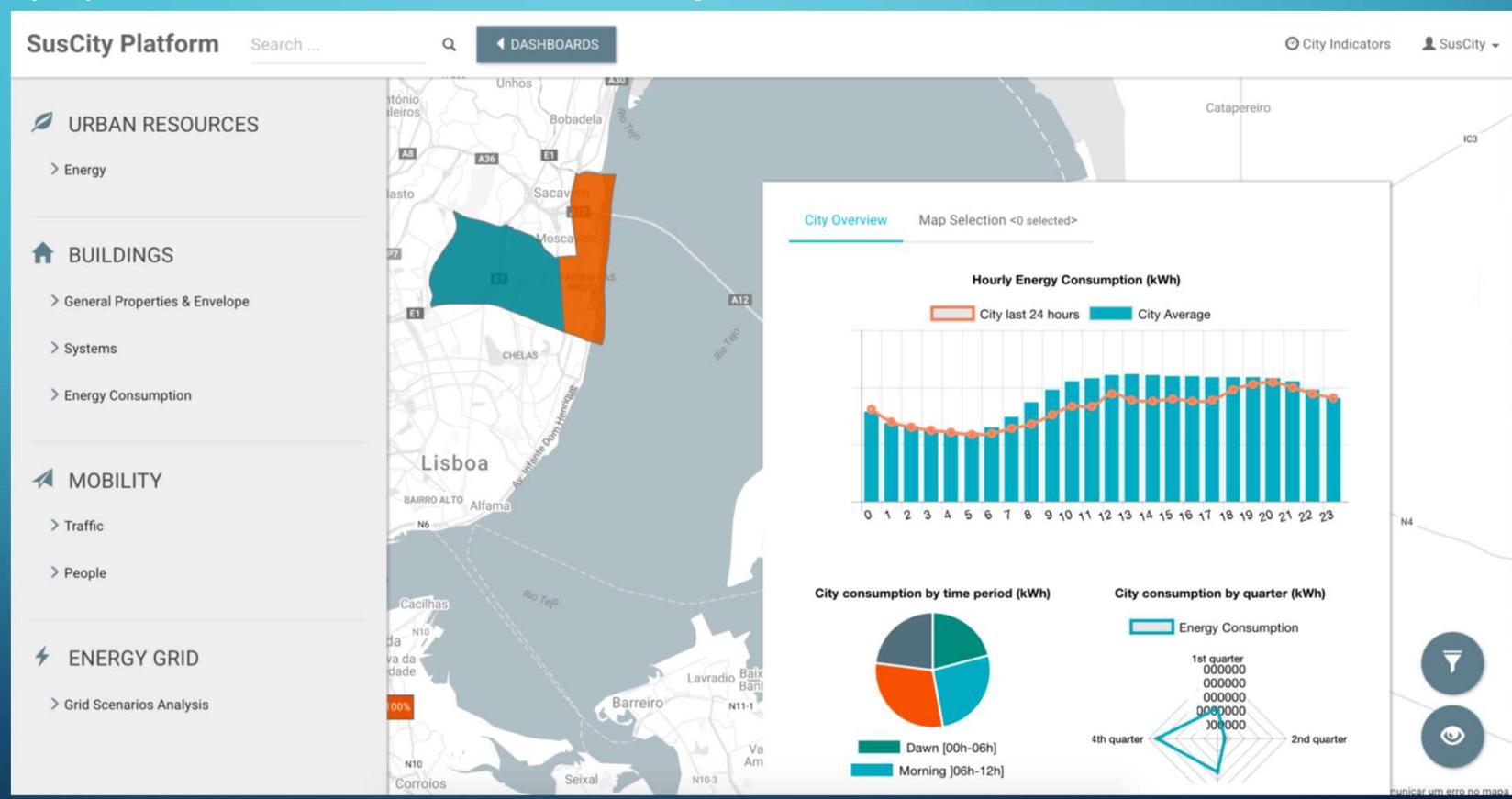


Atributos  
analíticos  
(indicadores  
históricos y  
predicciones)

tipos simples (por ejemplo, entero, flotante, varchar)  
o tipos complejos (por ejemplo, matrices, mapas y cadenas GeoJSON).

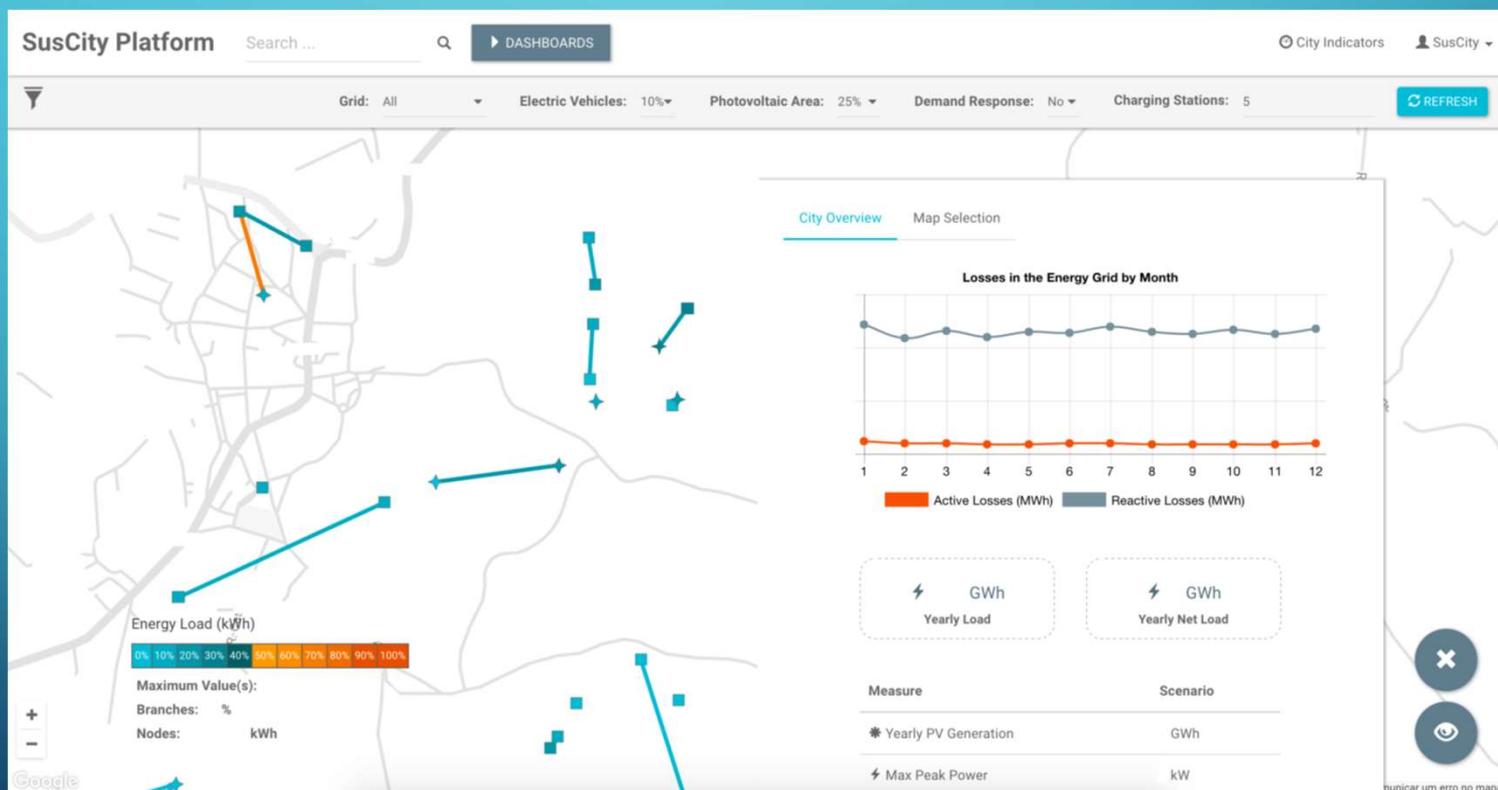
# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

Ejemplo de Visualización: Consumo de energía de cada barrio



# BIG DATA WAREHOUSE PARA CIUDADES INTELIGENTES

Ejemplo de Visualización: Tablero para la simulación de la red de energía





## Módulo “Análisis de Datos Científicos y Geográficos”

**UN MARCO DE INTEGRACIÓN DE DATOS URBANOS  
PARA EL ANÁLISIS DE LA MOVILIDAD EN  
CIUDADES INTELIGENTES**

(Ver paper completo en Campus→Lecturas)



## INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

Se presenta un enfoque novedoso para recopilar, integrar y realizar algunos análisis en movilidad dentro de ciudades inteligentes.

Se analizan patrones de movilidad basados en un grafo de múltiples aspectos (MAG).

Como caso de uso, se analiza la correlación spatio-temporal entre los datos reales recolectados desde dos fuentes diferentes en New York:

- Datos de redes sociales
- Datos de Yellow Taxi de la ciudad de Nueva York



## INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

Una forma de clasificar los datos generados en una ciudad es usar capas de detección.

Cada muestra de datos en la capa de detección incluye:

- un intervalo de tiempo (cuando se generaron los datos)
- una ubicación (donde se produjo)
- una especialidad de datos (mensaje en el tweet, una foto compartida, etc.)
- Un ID (para representar la entidad que creó los datos)

*La unificación de estos conjuntos de datos multicapa no es una tarea trivial*



## INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

- Algunas características que debe cumplir un framework para representar y analizar datos urbanos con éxito:
- ser capaz de integrar datos de fuentes heterogéneas: organizando los datos de una manera eficiente para la recuperación y la minería, manteniendo el consistencia de los datos para cada fuente independiente
- permitir el dominio cruzado para la combinación de datos
- Lograr manipular secuencias de coordenadas de ubicaciones con “sello de tiempo”



## INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

- Algunas características que debe cumplir un framework para representar y analizar datos urbanos con éxito:
- ser capaz de integrar datos de fuentes heterogéneas: organizando los datos de una manera eficiente para la recuperación y la minería, manteniendo el consistencia de los datos.
- permitir el dominio de los datos
- Lograr manipularlos “en tiempo”

A medida que los datos se vuelven cada vez más omnipresentes, generados en diferentes dispositivos y objetos, un aspecto importante para tratar con estos datos es hacer un seguimiento del ciclo de vida del mismo: retener información sobre cómo **se crean datos** y cómo **se transforma** a través de su uso.



## INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

Propuesta para representar datos de Centros urbanos:

### MAG - Gráfico de Múltiples Aspectos

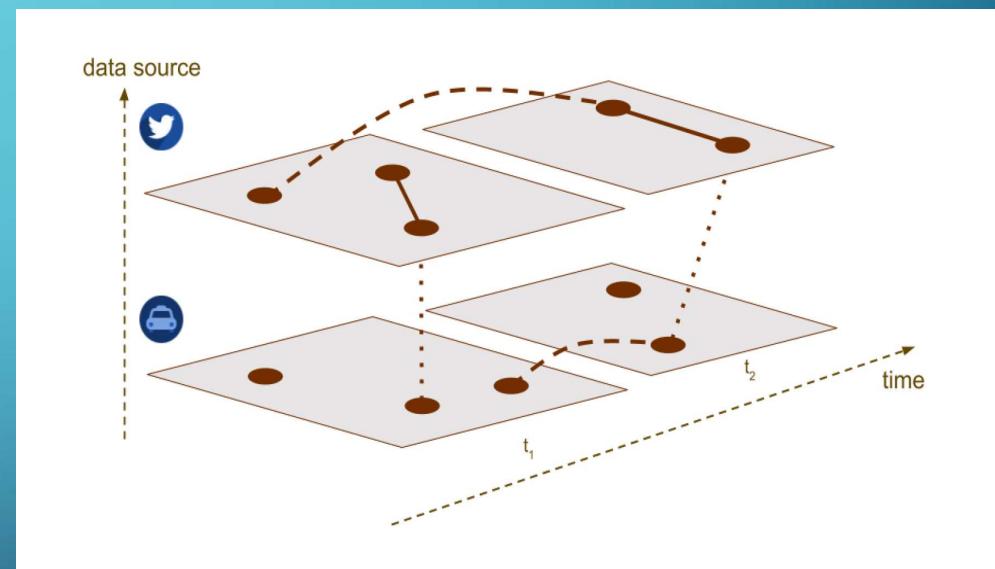
Permite la representación de diferentes características de datos mediante el uso de ASPECTOS.

Un aspecto es un eje donde se pueden crear capas

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

Ejemplo de MAG con dos aspectos:

- **Fuente de datos** → cada capa contiene datos de una fuente diferente
- **Tiempo** → discretizado, donde cada capa representa el momento de generación de los datos



# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

Ejemplo de MAG con dos aspectos:

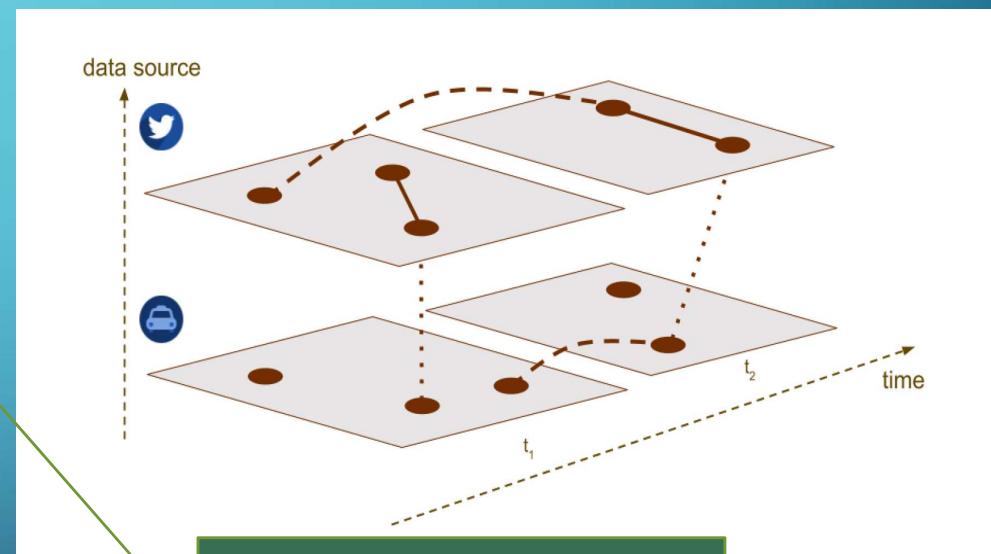
- **Fuente de datos** → cada capa contiene datos de una fuente diferente
- **Tiempo** → discretizado, donde cada capa representa el momento de generación de los datos



# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

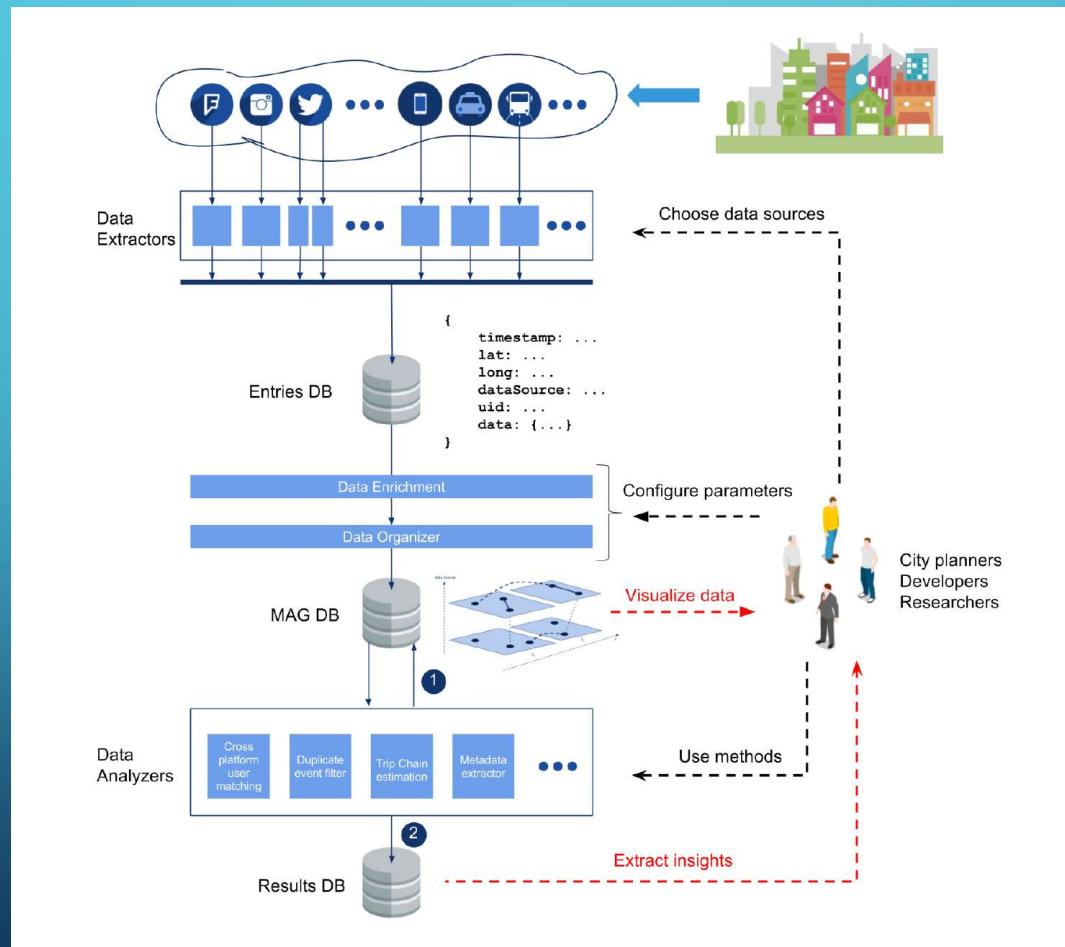
Hay tres tipos de relaciones:

- **Línea discontinua gruesa:** conecta nodos con el mismo ID
- **Línea punteada delgada:** representa nodos en diferentes capas, que fueron creado por la misma entidad
- **Línea continua:** representa contactos que suceden al instante (suficientemente rápido como para no usar más de uno paso de tiempo). Por ejemplo, un tweet enviado a otro usuario, que deja al primer usuario y llega al segundo en un segundo



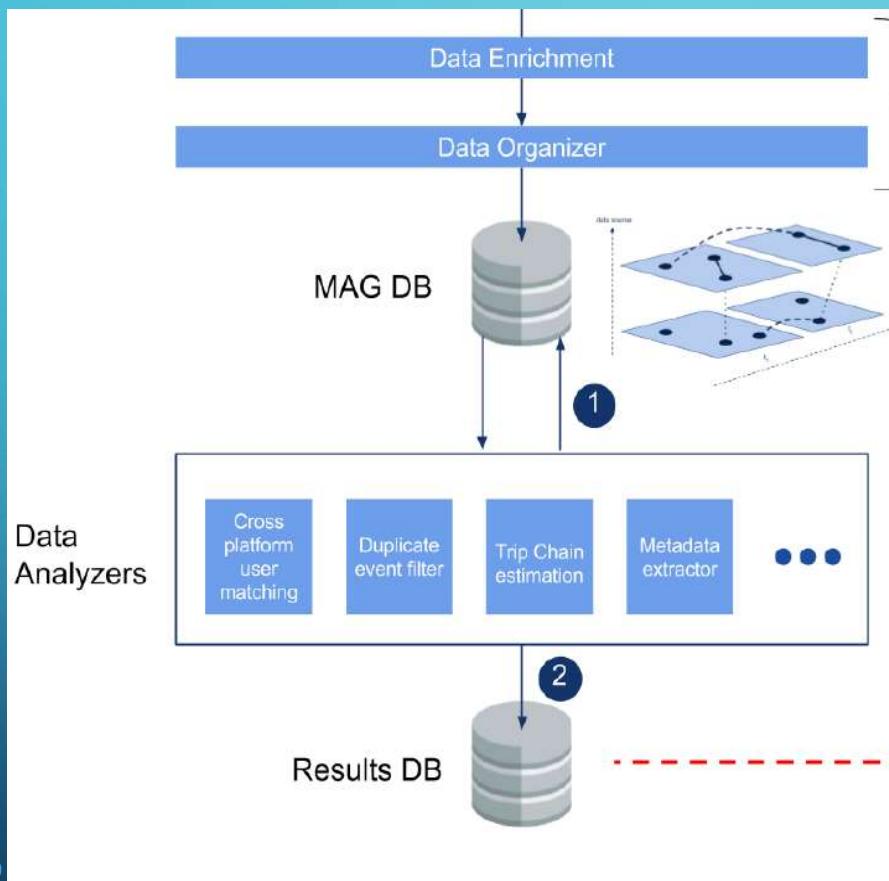
Como son de diferente ID,  
este tipo de conexión debe  
ser creado a través de  
tareas de análisis

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD



OGRÁFICOS

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD



*Los analizadores de datos pueden realizar dos tipos de tareas: pueden cambiar la estructura MAG para permitir la visualización de ciertos patrones, y pueden resumir resultados*

*Estos analizadores son parte clave para el análisis de resultados*



## INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

### Analizador **Fuzzy Matcher**

Identifica **coincidencias espaciotemporales** entre los nodos del MAG de diferentes capas y evalúa una puntuación temporal y espacial de las coincidencias

Los usuarios pueden especificar parámetros, como distancia, precisión de tiempo, y también depreciación de las funciones de distancia y tiempo

La forma en que las multitudes y otros flujos de movilidad se comportan en la ciudad se puede describir mejor mediante diferentes funciones de dispersión.



# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

## Analizador **Fuzzy Matcher**

Identifica **coincidencias espaciotemporales** entre los nodos del MAG de diferentes capas y evalúa una puntuación temporal y espacial de las coincidencias

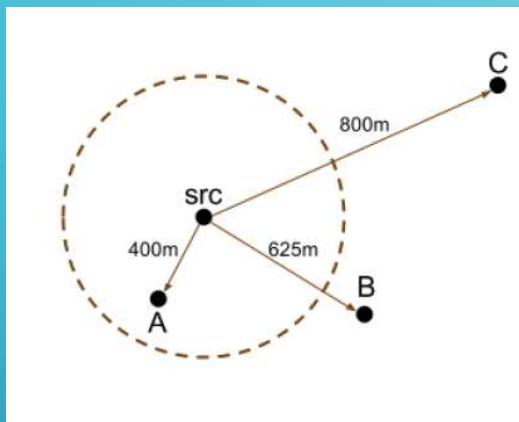
Los usuarios pueden especificar el tiempo, y también depreciar

La forma en que las multitudes en la ciudad se puede describir

*La forma en que las multitudes se mueven en una ciudad puede variar según un escenario particular.*

*Por ejemplo, una multitud en un desfile tiende a caminar distancias largas para hacer su causa más visible, mientras que una multitud que asiste a un concierto no se mueve mucho.*

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

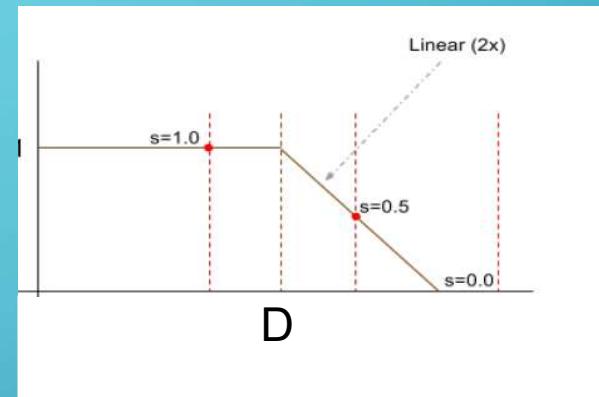


- Precisión de distancia D
- Precisión de tiempo T
- función de depreciación temporal  $td(t)$
- función de depreciación espacial  $sd(d)$   
(donde t y d son las distancias temporales y espaciales)

*el Fuzzy Matcher proporciona puntajes para representar qué tan fuerte / débil es una coincidencia, respecto de determinados umbrales*

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

Se crea una curva para evaluar el **Puntaje de la Coincidencia**



La puntuación espacial está dada  
 $SS(d) = f(d)/D$  donde  $f(d) = \{D \text{ if } d < D, s_d(d-D) \text{ if } d \geq D\}$

La puntuación espacial está dada  
 $TS(t) = f(t)/T$  donde  $f(t) = \{T \text{ if } t < T, s_t(t-T) \text{ if } t \geq T\}$

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

## Ejemplo

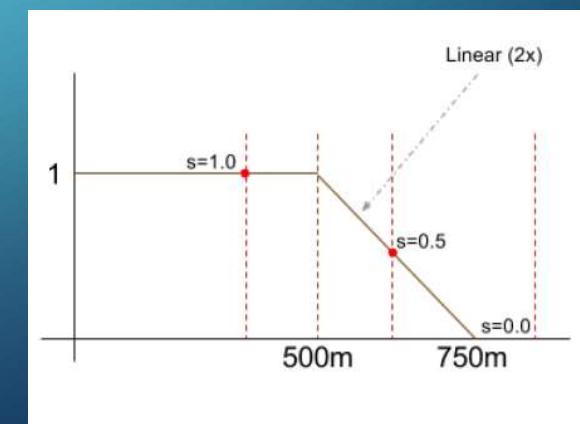
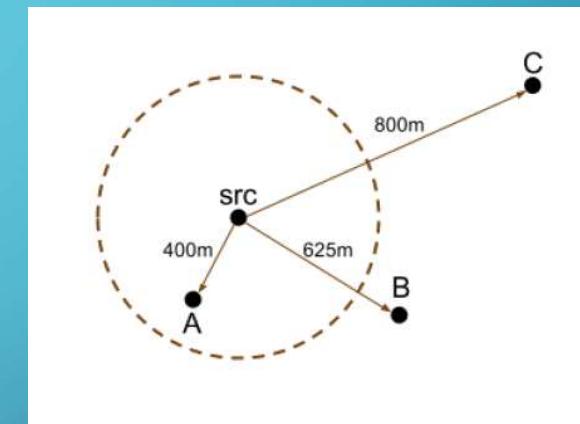
Con umbral  $D=500$

y depreciación espacial lineal  $sd(d) = \frac{500 - d}{250} + 1$

La coincidencia (SRC, A)  
tiene puntuación espacial  $SS(400) = 1$

La coincidencia (SRC, B)  
tiene puntuación espacial  $SS(625) = 0.5$

La coincidencia (SRC, C) tiene puntuación espacial  
 $SS(800) = 0 \rightarrow$  No hay coincidencia



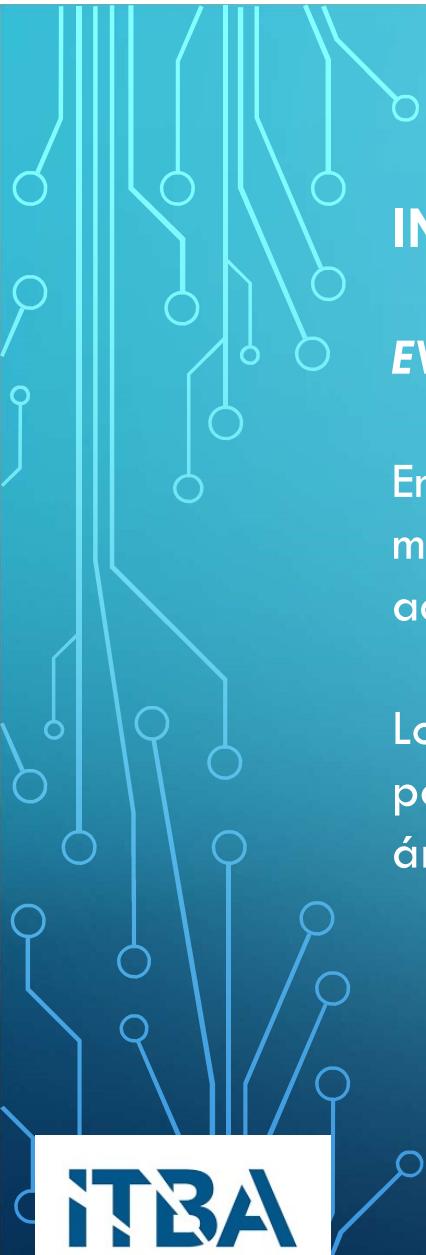


# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

## EVALUACIÓN 1

Utilizamos Fuzzy Matcher para analizar los datos recopilados de los dos diferentes fuentes de datos en la ciudad de Nueva York, definiendo un área de interés que cubre la región de Manhattan y algunos vecindarios cercanos:

- Conjunto de datos de 399.024 tweets geo-anotados recopilados durante el mes de enero 2016.
- Conjunto de datos disponible en el portal NYC Open Data con información de 10.580.378 viajes válidos sobre los viajes en Yellow Taxi de enero 2016

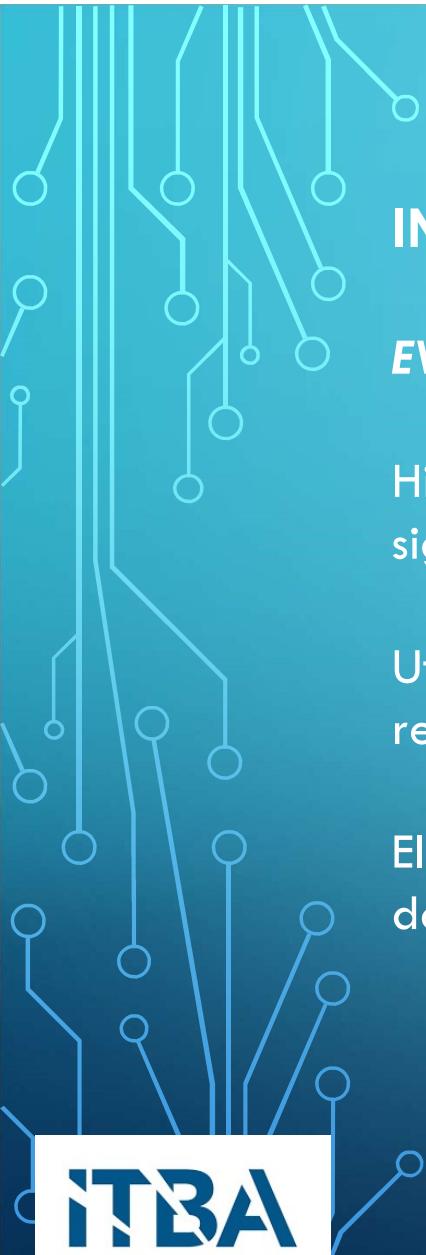


# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

## EVALUACIÓN 1

En este experimento, analizamos si podemos usar los datos de Twitter para una mejor posición de taxis dentro de la ciudad, para que estos taxis sean más accesibles para los ciudadanos

Los taxis amarillos en Nueva York sólo brindan sus servicios a través de la calle, por lo tanto, deben estar bien posicionados para tener una mejor cobertura del área de la ciudad, lo que resulta en un transporte más inteligente.



# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

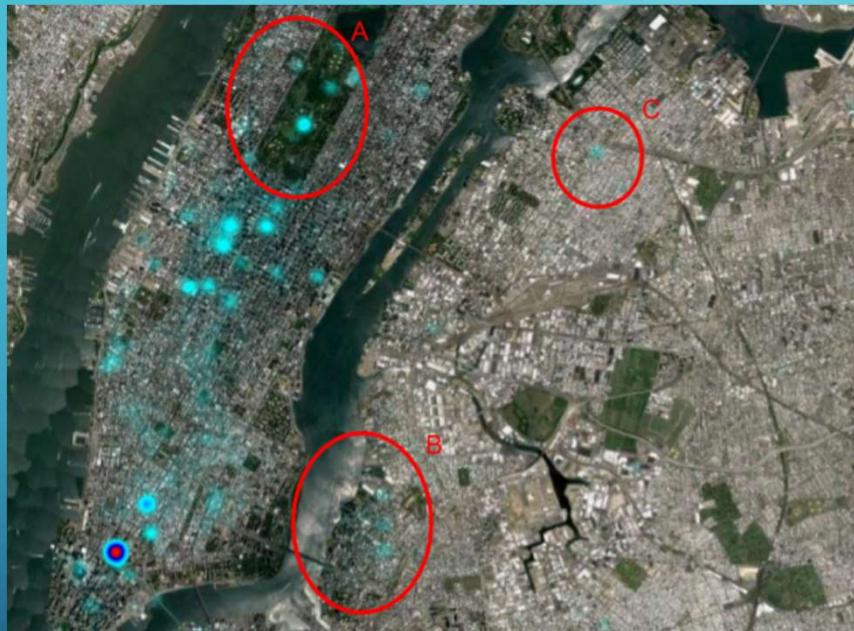
## EVALUACIÓN 1

Hipótesis: una región con un número relevante de tweets pueden indicar una cantidad significativa de posibles usuarios de taxis.

Utilizamos el algoritmo Fuzzy Matcher para analizar la correlación de muestras recolectadas de Twitter y las capas sensadas de Yellow Taxi.

El algoritmo identifica coincidencias espaciotemporales con una precisión espacial de 100 metros y una precisión temporal de 2 horas.

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD



Mapa de calor de la distribución de datos Twitter



Mapa de distribución de coincidencias

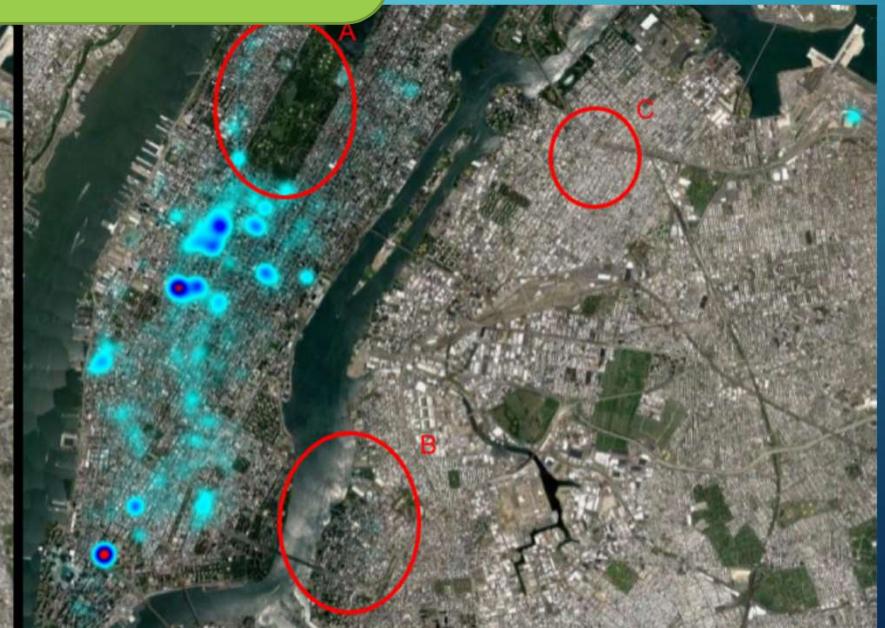
## INTEGRACIÓN DE

## ANÁLISIS DE MOVILIDAD

*Regiones con mayores volúmenes de tweets dieron lugar a regiones con mayor volúmenes de coincidencias*



Mapa de calor de la distribución de datos Twitter



Mapa de distribución de coincidencias



# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

## EVALUACIÓN 2

Encontrar viajes de taxis que tienen coincidencias borrosas, con capa de Twitter, al principio y al final del viaje (viajes que fueron realizados por el propietario de una cuenta de Twitter determinada).

Necesitamos un escenario donde los usuarios realmente twittearon, tomaron un taxi y luego twittearon nuevamente. En este escenario, 500 usuarios de Twitter enviaron 20,000 tweets en un día. Además, se generaron 5.000 viajes en taxi.

Usaron Matcher y analizaron los ID de usuario contenido en los partidos en ambas capas.



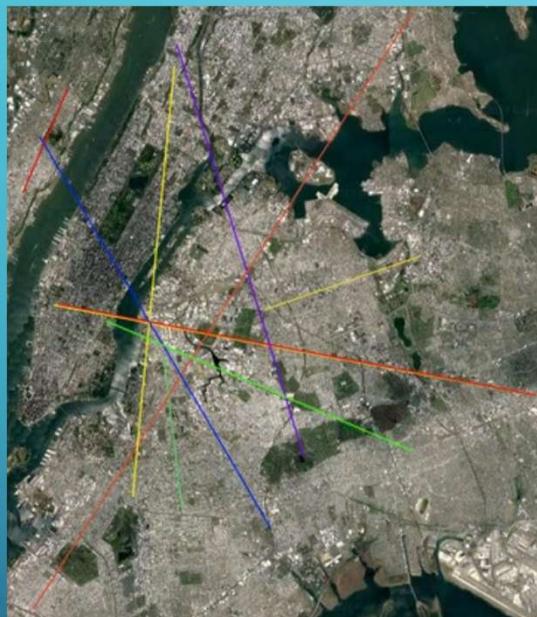
# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

## EVALUACIÓN 2

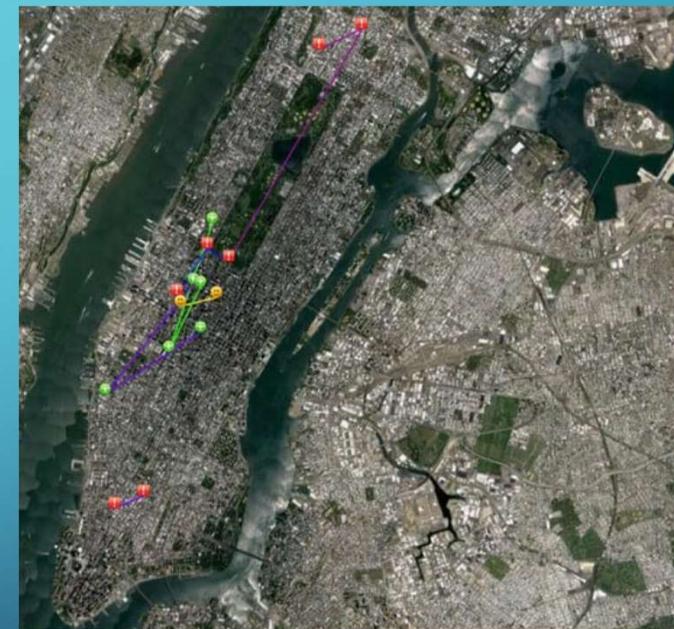
El algoritmo Fuzzy Matcher permite buscar coincidencias de un mismo nodo de capa Twitter con muchos nodos YellowTaxis.

Para considerar coincidencia, la distancia debe ser de al menos 500 metros, y un límite de velocidad máxima entre coincidencias de 100 km/h

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD

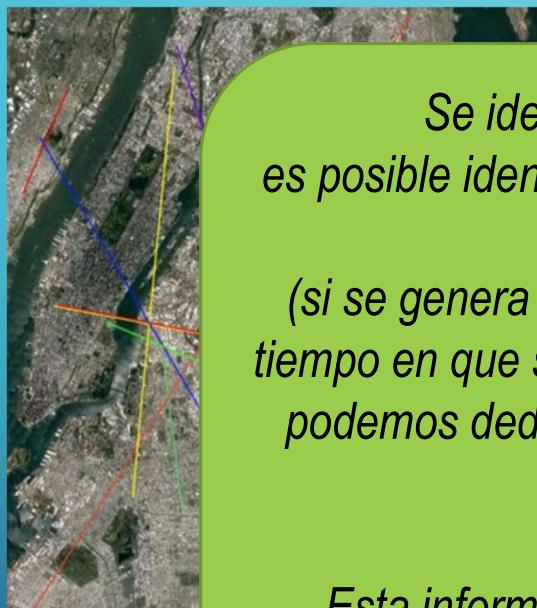


Escenario simulado



Escenario real

# INTEGRACIÓN DE DATOS URBANOS PARA ANÁLISIS DE MOVILIDAD



Escena



*Se identificaron 20.368 coincidencias persistentes: es posible identificar qué transporte fue utilizado por los usuarios en un momento establecido*

*(si se genera otra interacción del mismo usuario en la ventana de tiempo en que se estaba llevando a cabo la coincidencia persistente, podemos deducir que el usuario todavía estaba dentro del taxi en esa ubicación).*

*Esta información ayuda a mejorar la comprensión del uso del sistema de transporte.*



## Módulo “Análisis de Datos Científicos y Geográficos”

**UN MARCO INNOVADOR PARA SOPORTE EFECTIVO Y  
EFICIENTEMENTE DE BIG DATA ANALYTICS  
SOBRE GEO-LOCATED MOBILE  
MEDIOS DE COMUNICACIÓN SOCIAL**

(Ver paper completo en Campus→Lecturas)

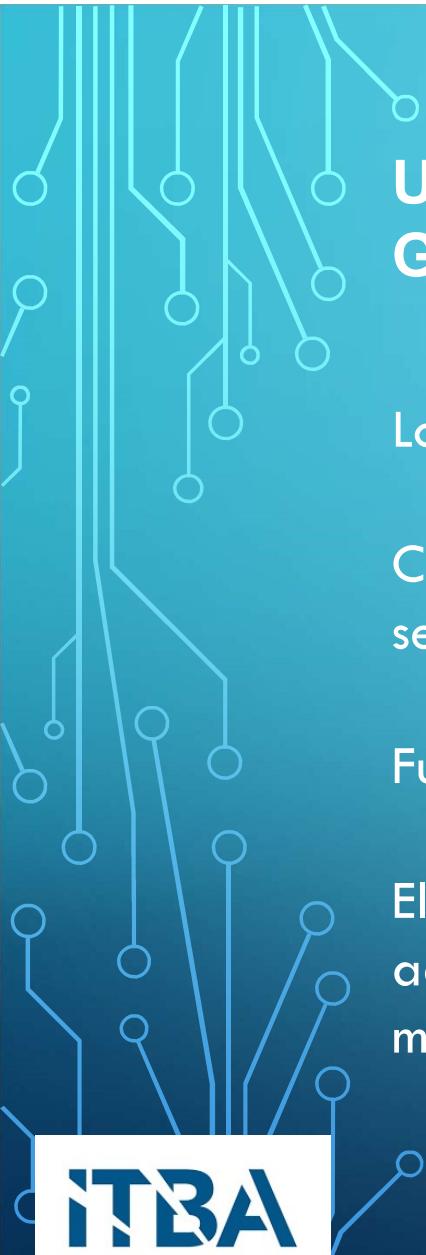


## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

Las redes sociales móviles (Twitter, Instagram, etc.) está produciendo grandes cantidades de datos que representan una fuente de conocimiento muy rica para análisis predictivo

Análisis de datos sobre redes sociales móviles geolocalizadas →  
Mayor potencia de análisis

Ejemplo: Plataforma FollowMe centrándose en el evento de la EXPO 2015 en Milán, Italia



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

Los teléfonos inteligentes modernos → Mobile Social Computación

Conjunto de metodologías, infraestructuras computacionales relacionadas y servicios para hacer frente a los problemas sociales

Fundamental: geo-localización de los mensajes

El conocimiento obtenido de mensajes geo-localizados es difícilmente adquirible por medio de métodos de encuesta tradicionales



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

### PROYECTO FOLLOWME

La Computacion social móvil puede ayudar a comprender dónde van los viajeros, qué sitios visitan, en que lugares pasan sus noches.

Mediante la recopilación de publicaciones geolocalizadas que envían durante su viaje, es posible reconstruir sus viajes

Pero para realizar este poderoso análisis de datos multimediales, necesitamos Big Data



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

### PROYECTO FOLLOWME

El aeropuerto ubicado en Bérgamo (norte de Italia) ha ido creciendo desde que una conocida compañía de bajo costo lo eligió como su Centro italiano.

En unos pocos años, se ha convertido en el cuarto aeropuerto italiano (Enero 2014)

El aeropuerto sirve un rango de 100 km con 9 millones de habitantes y 800.000 empresas que producen el 21% del PIB italiano.

La consecuencia es una visibilidad internacional aumentada y un incremento de turistas extranjeros que visitan la ciudad y su centro medieval.



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

### Objetivo de FOLLOWME

El objetivo del proyecto FollowMe es construir técnicas y herramientas permitiendo el seguimiento de los movimientos de turistas que visitan una región determinada, con el fin de construir análisis adecuados de big data sobre ellos.



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

Algunas definiciones

**Post (T)**

Una publicación T es una tupla  $\langle id, ID \text{ de usuario}, \text{fecha}, \text{hora}, \text{texto}, \text{lat}, \text{lon} \rangle$   
donde:

**id** identifica de manera única la publicación

**ID** es el identificador del usuario que envió la publicación  
fecha en que se envió la publicación;

hora en que se envió la publicación; (v) el texto es el texto en la publicación;  
Latitud, longitud de la posición donde el usuario envió la publicación a través de un dispositivo móvil



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

### **Hang Post (HT)**

Es una publicación geolocalizada enviada en un área del aeropuerto. permite descubrir el usuario de los usuarios que viajan.

Respecto de T, una publicación HT tiene un campo adicional llamado origen, que es el nombre del área del aeropuerto donde se colocó la publicación

### **Tracket Post (TT)**

Es una geolocalización publicación enviada por un usuario viajero, siempre que el usuario haya enviado un HT no más de 9 días antes de que se publique TT.



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

### Trip

Es una tupla trip: = <ID de usuario, fecha, origen, tSeq: (tt1 ;:::; ttN)>  
donde:

ID es el identificador del usuario que publicó el seguimiento  
publicaciones que componen el viaje  
fecha en que se publicó el seguimiento expedido  
origen es el aeropuerto de donde proviene el viaje  
tSeq es la secuencia real de publicaciones rastreadas que componen los viajes  
(las publicaciones se ordenan por fecha y hora).



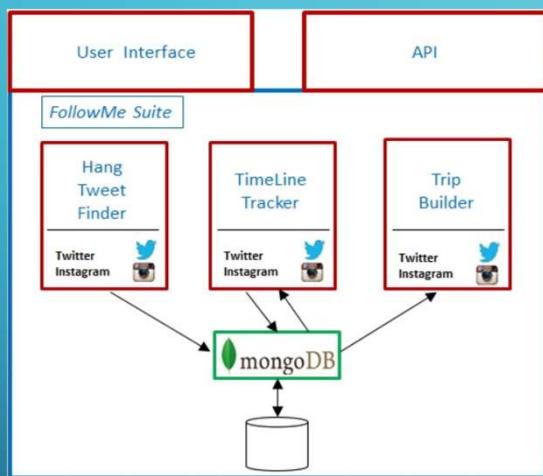
## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

Para descubrir viajeros y recuperar suficiente información sobre su procedencia, los mensajes deben buscarse en el aeropuerto de donde provienen los viajes de estos viajeros

De esta manera, es posible conocer el ID de usuario y la pista sus viajes al consultar su línea de tiempo, es decir, la lista revertida de publicaciones Twitter o Instagram para un usuario determinado.

El límite temporal de 9 días es razonable para considerar a los viajeros como turistas a través de los aeropuertos.

# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION



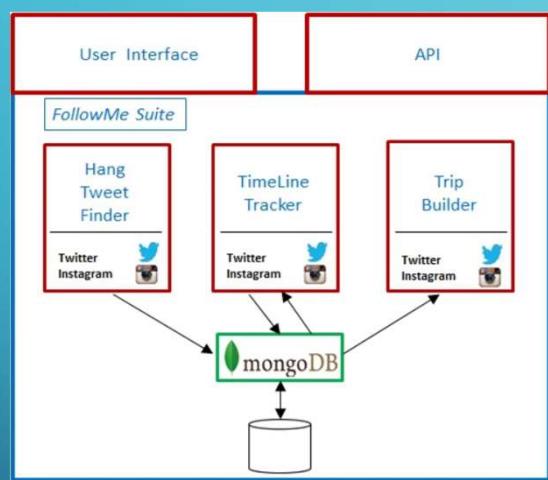
## MongoDB.

Un DBMS sin SQL, diseñado para tratar con colecciones de documentos donde cada documento se representa como un objeto JSON. La ventaja principal es la capacidad de administrar documentos con diferentes estructuras dentro de las mismas colecciones, superando el concepto de esquema en tablas.

## Hang Post Finder

Componente es responsable de consultar la API de Twitter y la API de Instagram para descubrir HT. Las API proporcionan la capacidad de buscar publicaciones ubicadas geográficamente, dadas las coordenadas del centro y el radio de un área de interés. Debido a la escalabilidad y la caducidad de las publicaciones en la API, el Hang Post Finder se lanza todos los días para encontrar publicaciones colgadas durante el día anterior.

# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION



## TimeLine Tracker

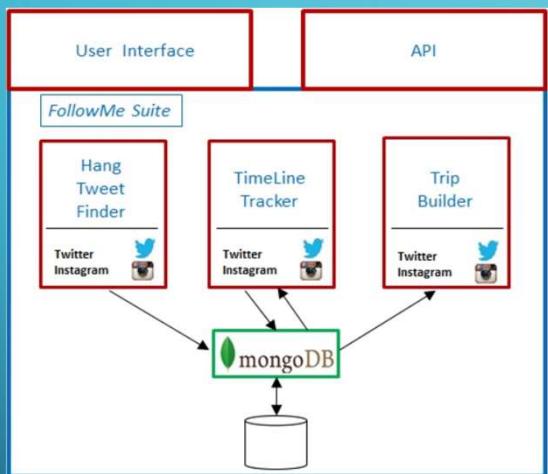
Componente que sigue el historial de publicaciones publicadas por cada Usuario

Debido a razones de escalabilidad, se lanza el Timeline Tracker todos los días, siguiendo los tiempos un día hacia atrás, considerando solo las publicaciones suspendidas no mayores a 9 días

## Trip Builder

Componente que, dado un cuadro delimitador BB, extrae todas las secuencias de publicaciones rastreadas que se publican en el área delimitada

# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION



El constructor de Trip produce un archivo en formato CSV, que tiene a la comodidad de ser procesable por herramientas externas, tales como MatLab, Excel, etc., para inspeccionar y analizar viajes.

Es importante visualizar viajes en mapas, para permitir el análisis visual de viajes descubiertos. Por esta razón, el FollowMe suite proporciona convertidores a varias representaciones KML de los viajes (formato aceptado por Google Earth y Google Maps API)

Para las tareas de análisis, Google Earth se convierte en una muy herramienta poderosa, ya que permite seleccionar información sobre los elementos que se van a mostrar en los mapas.

Los archivos KML permiten dividir información basados en una propiedad específica. Por ejemplo, un analista podría estar interesado en particionar viajes basados en el aeropuerto de origen.



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

### Post Alineación

Para hacer un análisis de ruta efectivo y habilitar agregaciones intermedias, cada publicación en un viaje está alineada la base de la distancia entre su fecha y la fecha del comienzo publicación del viaje.

$$t.td = (t.date - h.date) + 1$$

donde  $h$  modela el hang post del viaje (la publicación enviada en aeropuerto de origen)



# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

## Partición del horario

Para habilitar el análisis de dimensión basado en el horario, cada publicación se extiende con los siguientes intervalos de tiempo:

TS1: 10: 00pm - 05: 59am, Noche;  
TS2: 06: 00am - 11: 59am, Mañana;  
TS3: 00: 00pm - 5: 59pm, tarde;  
TS4: 6: 00 p.m. - 9:59 p.m., Noche.

*TS1 proporciona información sobre lugares donde los viajeros duermen  
TS4 proporciona información sobre lugares donde los viajeros cenan  
TS2 y TS3 brindan información sobre las actividades diarias*



## UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

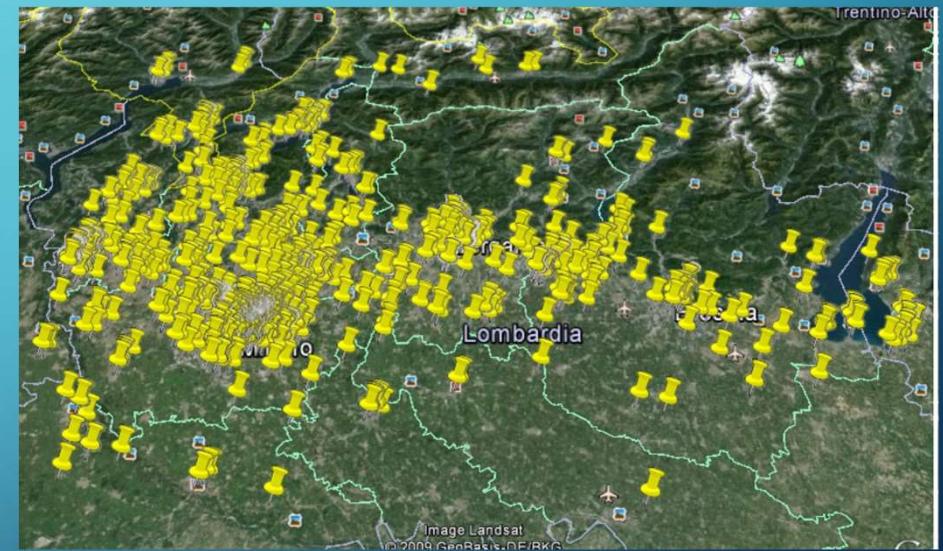
Qué tipo de análisis se puede realizar en los viajes?

- Camino
- Aeropuertos origen
- Ranuras de tiempo
- Días de semana

# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACIÓN

## ESTUDIO DE CASO: LA EXPO 2015 EN MILÁN

Para ilustrar la efectividad de nuestro análisis de big data marco, construimos un estudio de caso centrado en la conocida EXPO 2015 evento en Milán, Italia, y sobre la base de un conjunto de geo-ubicado publicaciones recopiladas por la suite FollowMe.



Distribución de publicaciones en la región de Lombardía

# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION



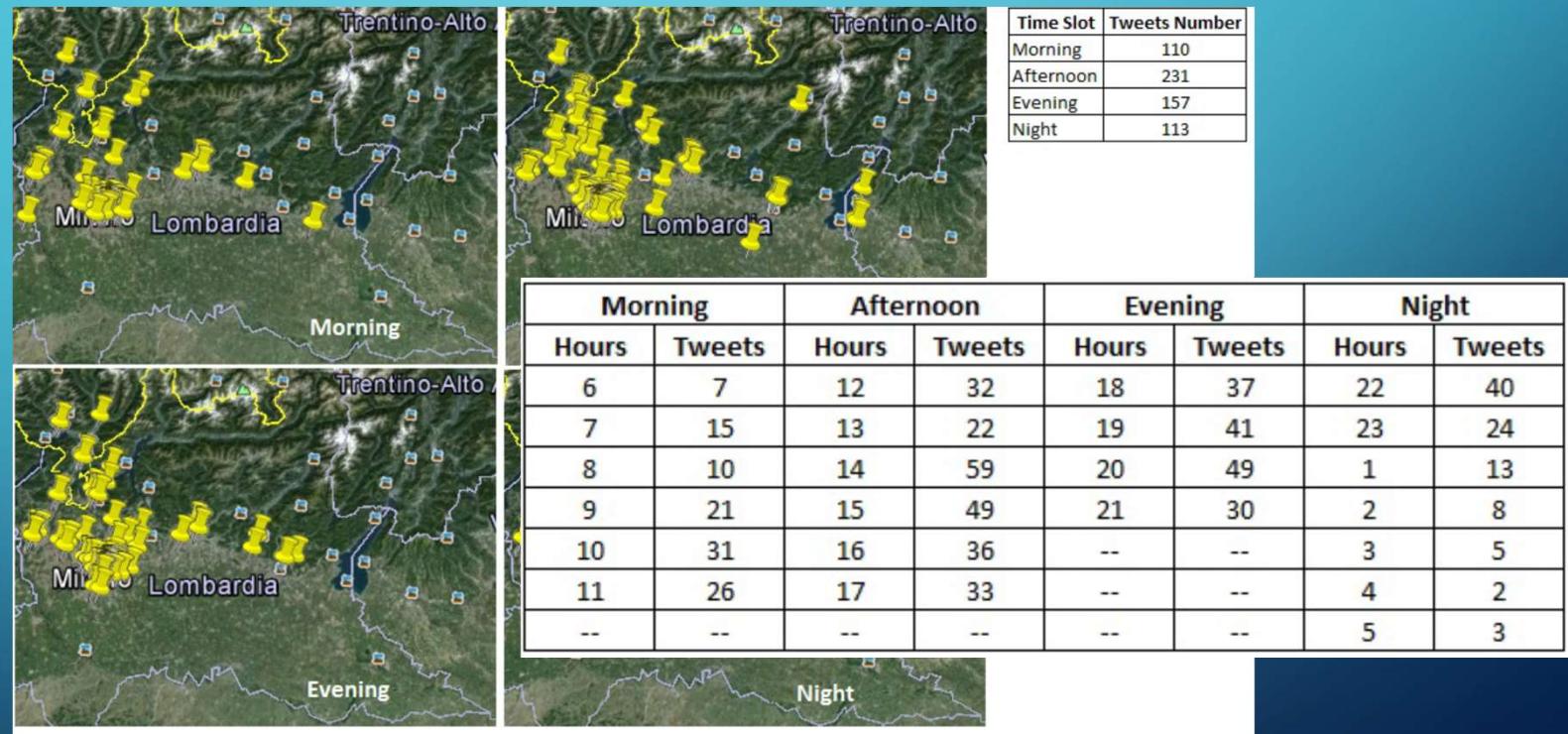
Distribucion de publicaciones con  
respecto a los viajeros provienen de  
Barcelona



Mensaje de un viajero español

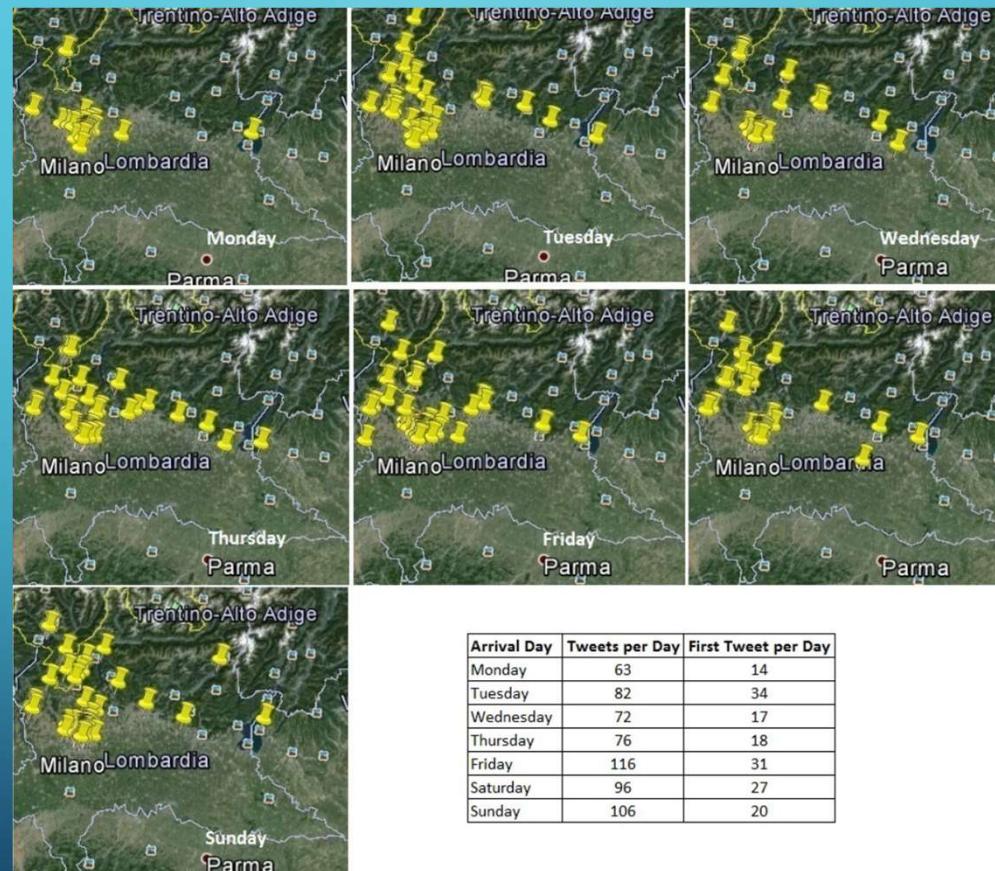
# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

Distribución de publicaciones en Time Slots



# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACIÓN

Páginas por dia de la semana

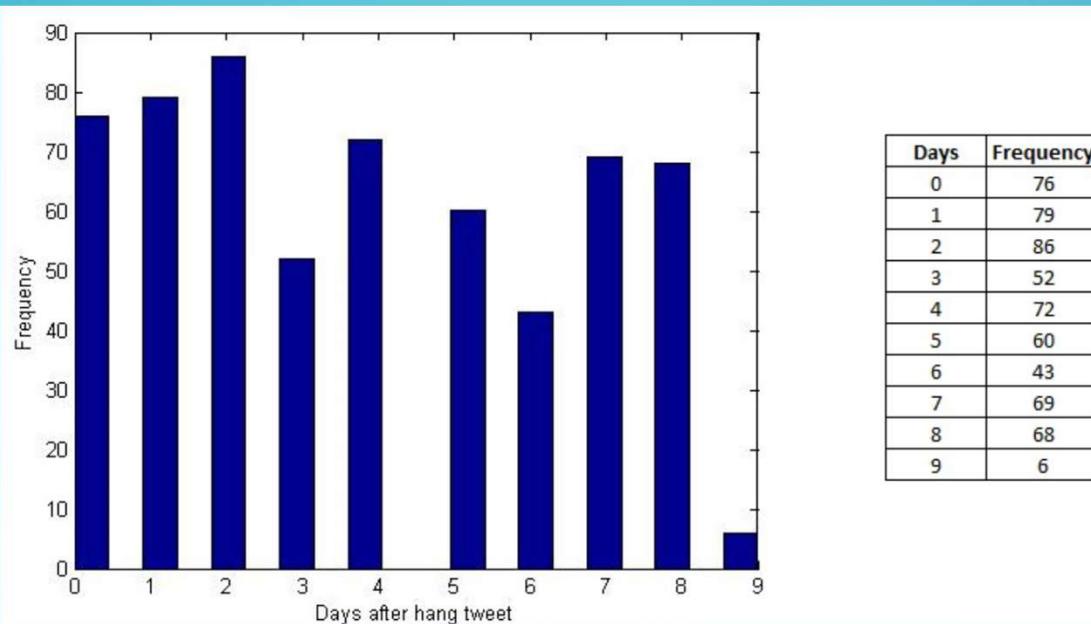


En sábado y domingo, se postea el 50% del total de publicaciones

# UN MARCO INNOVADOR PARA BIG DATA CON GEOLOCALIZACION

Las aplicaciones de Big Data son muy importantes para la economía de las regiones que visitan, y sus patrones pueden ayudar a mejorar aspectos de la administración públicas

Días transcurridos después de la primera publicación



Notar cómo un viajero publica su primera publicación principalmente dentro de los tres días de haber estado en el aeropuerto



## Módulo “Análisis de Datos Científicos y Geográficos”

### PLANIFICACION URBANA USANDO ANALISIS BIG DATA BASADO EN IOT

(Ver paper completo en Campus→Lecturas)



# PLANIFICACION URBANA CON IOT Y BIG DATA

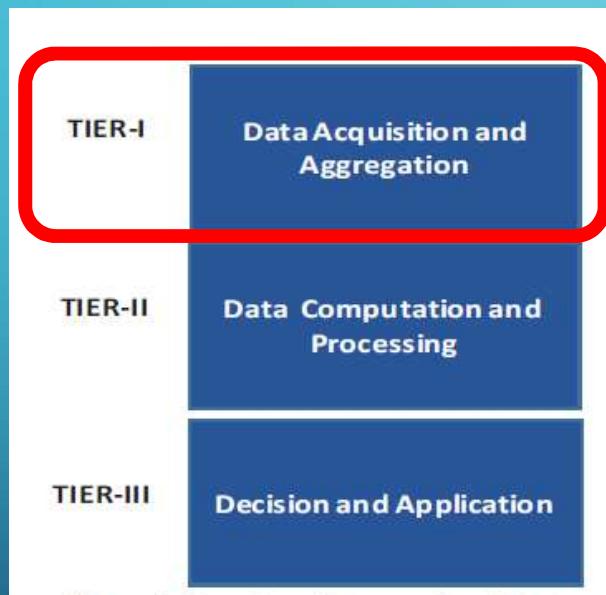
**Smart Urban Planning using Big Data Analytics based Internet of Things**

Muhammad Babar, Fahim Arif

Proponen una arquitectura práctica para ciudades inteligentes, incorporando análisis Big Data a IoT para la toma de decisiones inteligente en tiempo real.

# PLANIFICACION URBANA CON IOT Y BIG DATA

## Arquitectura



Fuentes de datos (con sensores): tráfico, clima, salud, hogar, desechos, uso de agua, etc. → ZigBee, Wi-Fi, Bluetooth y redes celulares 3G / 4G.

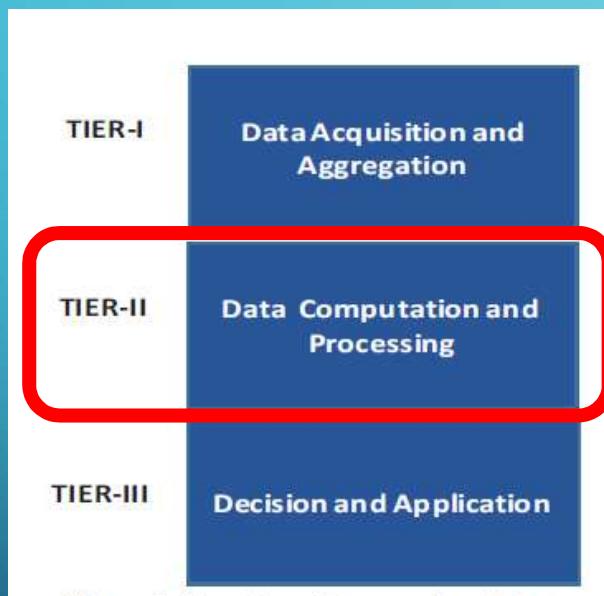
Agregación: divide & conquer

Preprocesamiento: eliminar datos fuera de rango, poco prácticos y valores perdidos.

Min-Max para normalización y filtro Kalman para eliminar el ruido

# PLANIFICACION URBANA CON IOT Y BIG DATA

## Arquitectura

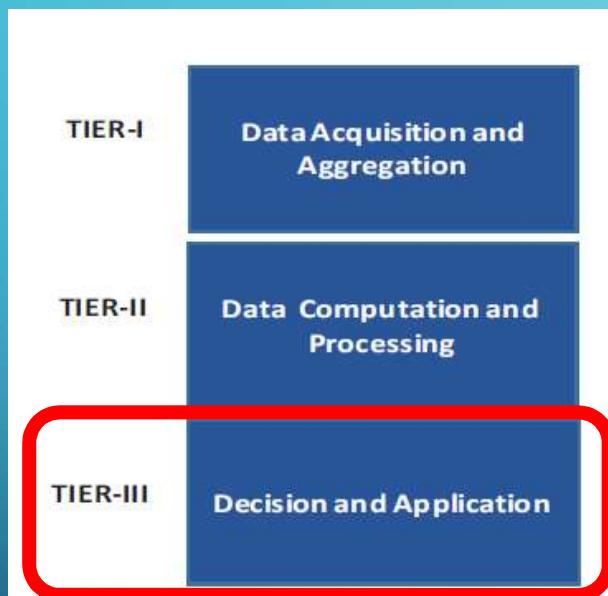


Procesamiento: Hadoop, LST (Least Slack Time) y balanceo RR (Round Robin).

Los datos procesados son separados por subservicio

# PLANIFICACION URBANA CON IOT Y BIG DATA

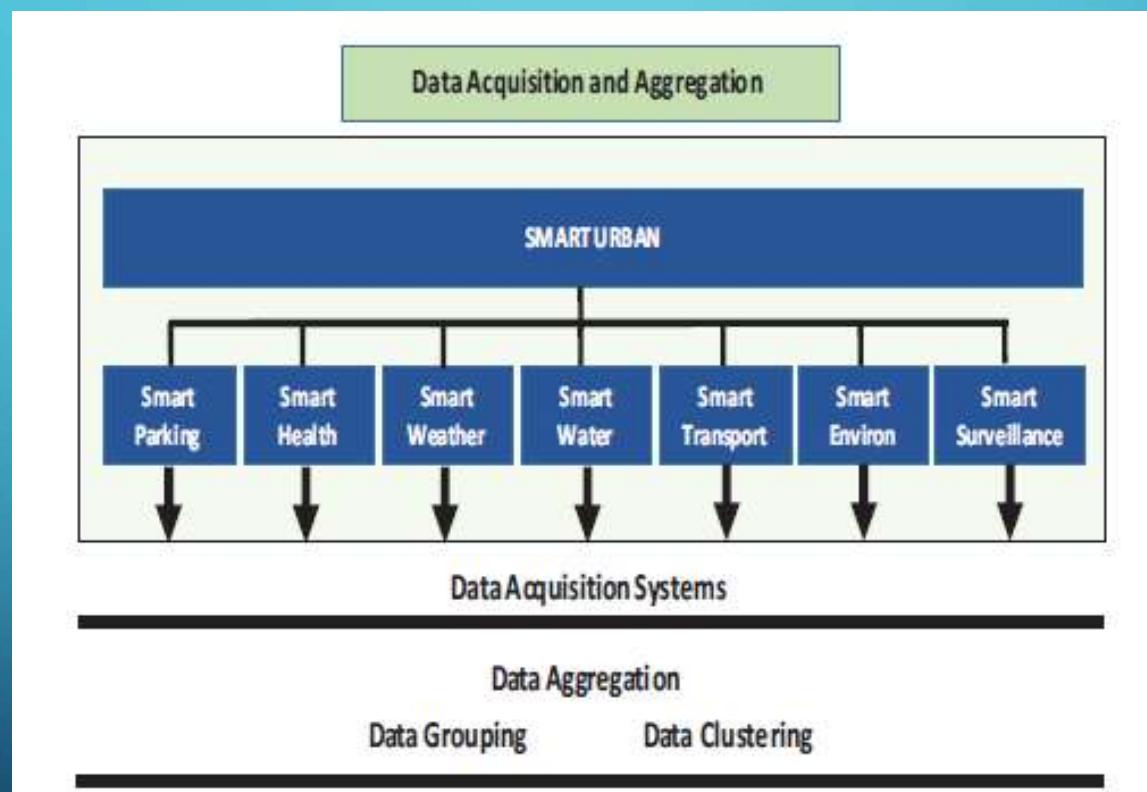
## Arquitectura



Toma de decisiones y la planificación: gestión propia de eventos

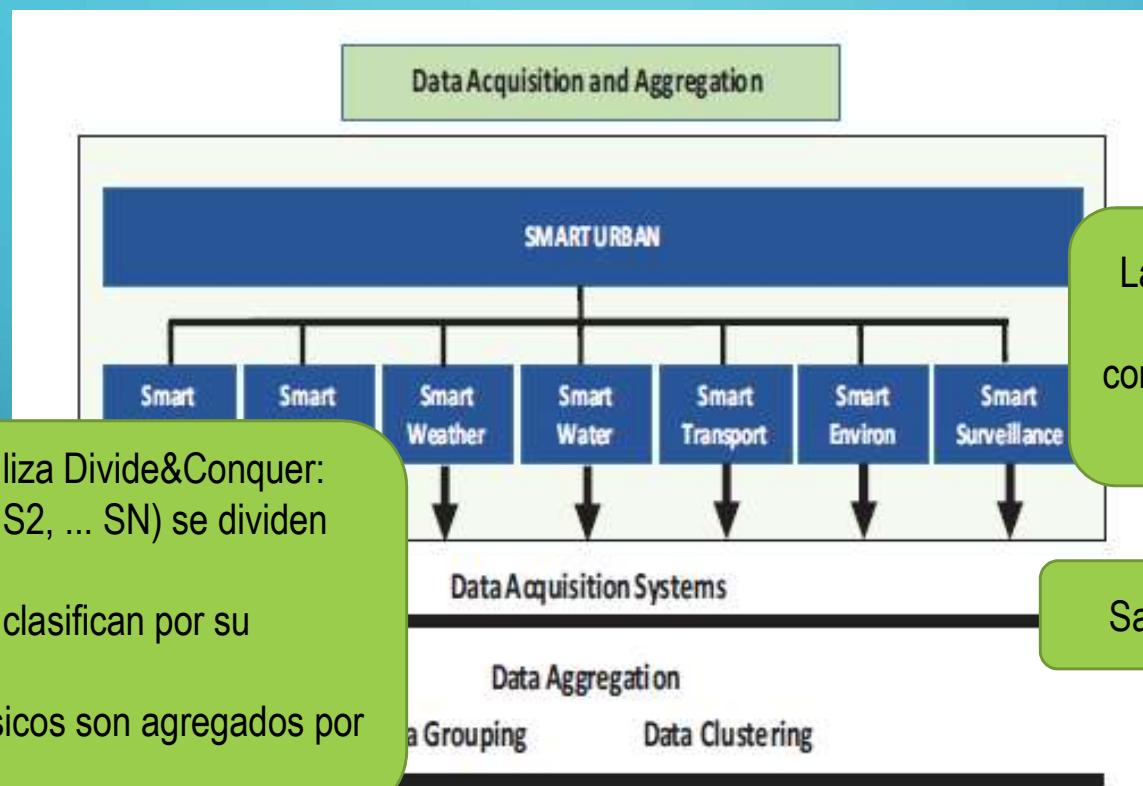
# PLANIFICACION URBANA CON IOT Y BIG DATA

## Capa 1: Adquisición y Agregación de Datos



# PLANIFICACION URBANA CON IOT Y BIG DATA

## Capa 1: Adquisición y Agregación de Datos





# PLANIFICACION URBANA CON IOT Y BIG DATA

## Capa 2: Procesamiento

Este nivel juega un papel vital en esta arquitectura, ya que es el módulo de procesamiento central

Este nivel está compuesto por

- 1) preprocesador de datos**
- 2) filtro de datos**
- 3) sistema Hadoop**
- 4) almacenamiento**



# PLANIFICACION URBANA CON IOT Y BIG DATA

## Capa 2: Procesamiento

Min-Max Normalization

Filtro de Kalman (KF) para acelerar los datos

Hadoop Ecosystem (MapReduce)

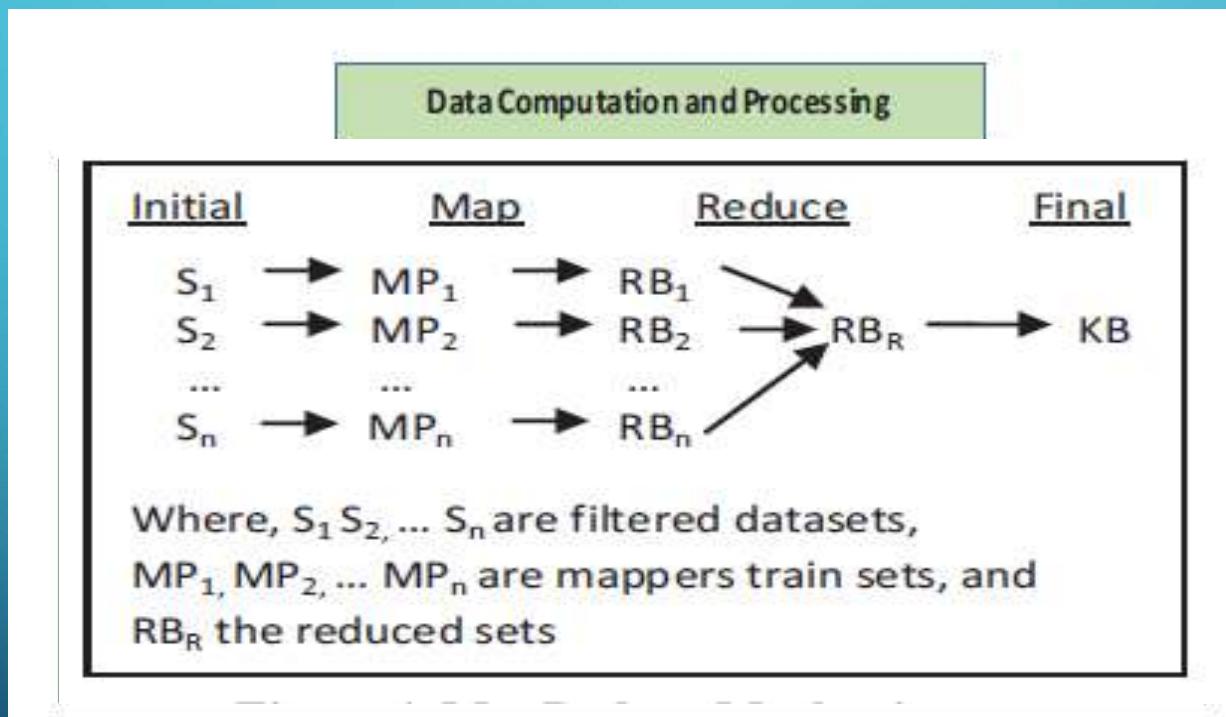
LST (least slack time) - selecciona aquellas tareas que tienen el menor "tiempo de holgura"

(holgura = tiempo deadline – tiempo actual – tiempo remanente de ejecución)

Almacenamiento: HDFS

# PLANIFICACION URBANA CON IOT Y BIG DATA

## Capa 2: Procesamiento





# PLANIFICACION URBANA CON IOT Y BIG DATA

## Capa 3: Toma de decisiones y nivel de aplicación

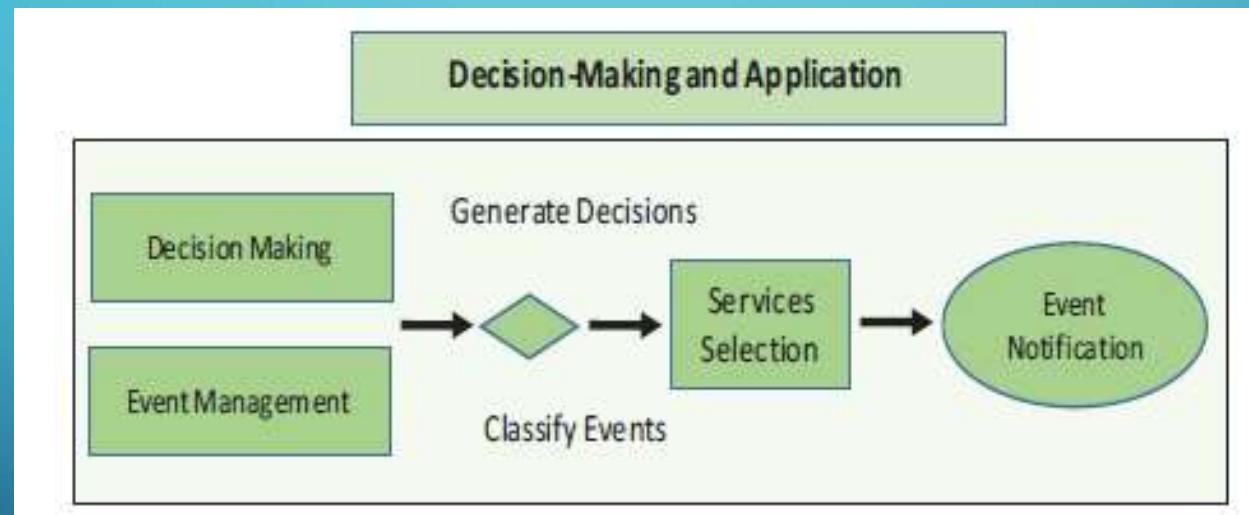
El servidor de toma de decisiones describe la decisión según una ontología que se usa para unificar los eventos.

La sociedad correspondiente realiza el servicio de selección para distinguir los eventos altos y bajos.

La unidad de selección de servicio genera el evento respectivo lo difunde a los niveles departamentales, de servicio o subservicio.

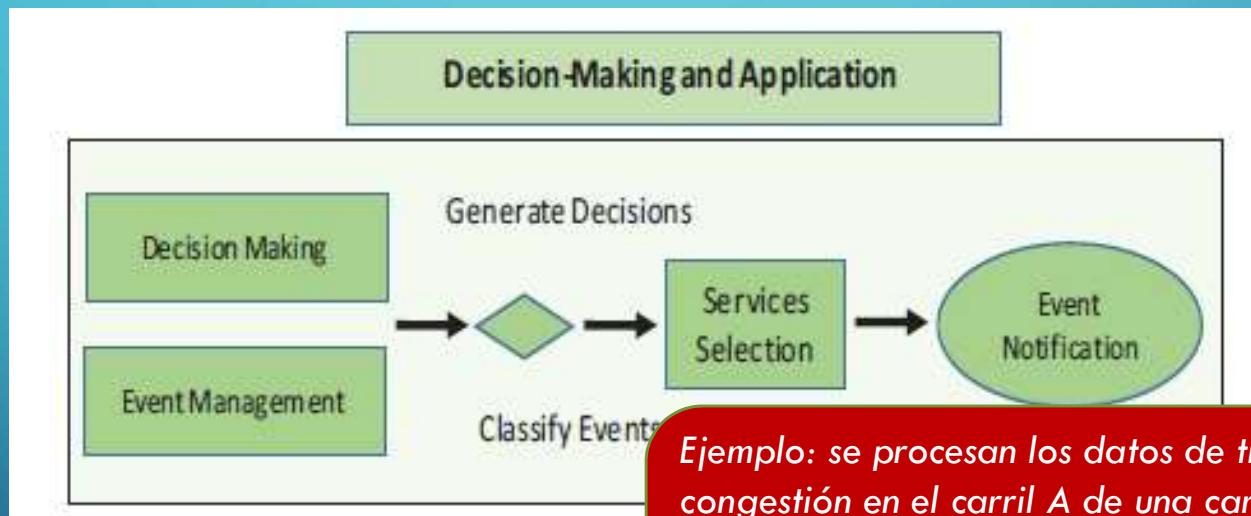
# PLANIFICACION URBANA CON IOT Y BIG DATA

## Capa 3: Toma de decisiones y nivel de aplicación



# PLANIFICACION URBANA CON IOT Y BIG DATA

## Capa 3: Toma de decisiones y nivel de aplicación



*Ejemplo: se procesan los datos de tráfico y se detecta una congestión en el carril A de una carretera. Se toma una decisión y el evento es enviado al departamento de la sociedad de control de tráfico inteligente, quien notificara a los usuarios involucrados para optar por otro carril, para evitar y controlar la congestión.*



# PLANIFICACION URBANA CON IOT Y BIG DATA

## Casos de Uso

Fuentes:

<http://data.surrey.ca/dataset>

El centro de análisis es evaluar toma de decisiones basados en ciertos umbrales.

# PLANIFICACION URBANA CON IOT Y BIG DATA

## Casos de Uso

Fuentes:

<http://data.surrey.ca/datasets>

El centro de análisis e  
basados en ciertos un

The screenshot shows the 'Datasets' page for the City of Surrey. The top navigation bar includes links for Datasets, Showcases, Suggest a Dataset, Subscribe, Contact, Disclaimer, and About. A search bar at the top right contains the placeholder 'Search datasets...' and a magnifying glass icon. Below the search bar, a message states '365 datasets found' and 'Order by: Relevance'. The main content area displays three sections: 'Fraser Health Restaurant Inspection Reports', '2015 Orthophoto', and '2016 Orthophoto'. Each section includes a brief description, a list of available formats (e.g., CSV, API, SID), and download links. The 'Fraser Health Restaurant Inspection Reports' section also includes a note about inspection frequency and regulatory requirements.

# PLANIFICACION URBANA CON IOT Y BIG DATA

## Casos de Uso

Fuentes:

<http://data.surrey.ca/datasets>

El centro de análisis e  
basados en ciertos un

The screenshot shows two versions of the City of Surrey's data portal side-by-side. Both versions feature a header with the City of Surrey logo and navigation links for Datasets, Showcases, Suggest a Dataset, Subscribe, Contact, Disclaimer, and About. The left version has a light blue background and displays a sidebar with filters for Formats (CSV, JSON, API, KML, FGDB, DWG, TIFF, XLSX, SID, PDF) and Categories (Transportation, Local Government, Land, Infrastructure, Environment, Business). The right version has a light orange background and also displays a sidebar with similar filters. Below the sidebar, a search bar says "Search datasets..." with a magnifying glass icon. A message indicates "8 datasets found". Under "Categories", "Transportation" is selected. Under "Tags", "traffic" is selected. The main content area shows three dataset cards: "Traffic Signals", "Traffic Calming", and "Traffic Counts 2013". Each card includes a brief description, a list of available formats (CSV, JSON, KML, FGDB, DWG, API), and a "CSV" button.

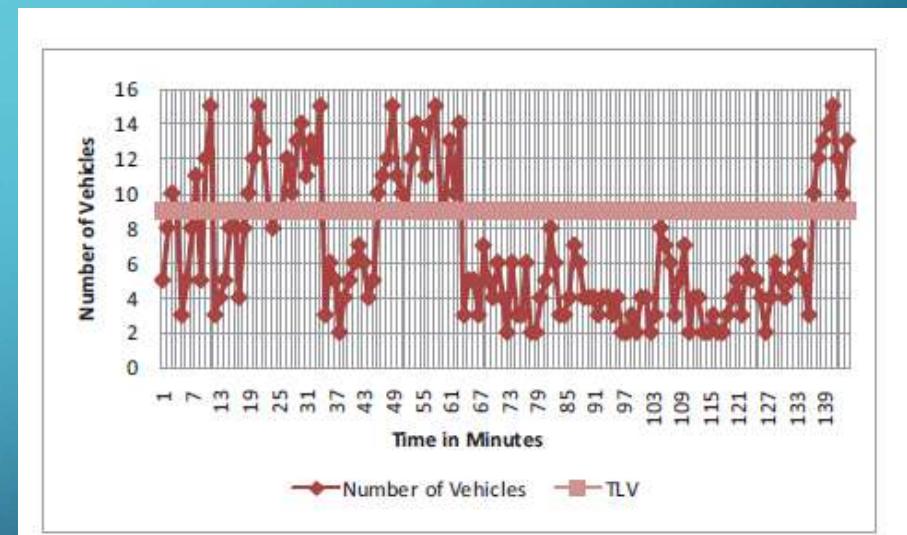
# PLANIFICACION URBANA CON IOT Y BIG DATA

## Caso 1: Trafico

Se mide el número de vehículos en diferentes carreteras en la ciudad de Aarhus, Dinamarca.

El umbral para la congestión del tráfico es de 9 vehículos por carril.

Cada vez, la cantidad de datos en un momento específico va más allá del umbral, se inicia un evento particular para el Departamento.



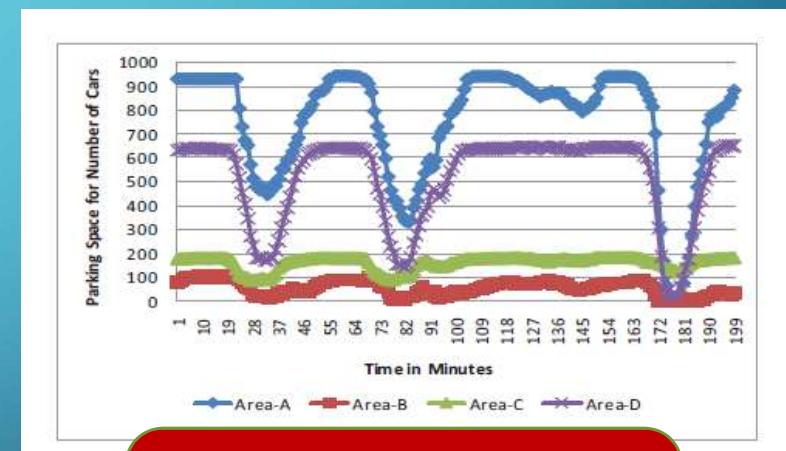
# PLANIFICACION URBANA CON IOT Y BIG DATA

## Caso 2: Parking

Se mide el número de espacios libres en los estacionamientos IoT en la ciudad de Aarhus, Dinamarca.

Los centros comerciales, las organizaciones y las tiendas ofrecen los datos de estacionamiento a través del propio sistema para dar el servicio de estacionamiento inteligente.

Los ciudadanos solo pueden reservar el espacio en los estacionamientos IoT antes de llegar al lugar.



parking inteligente →  
mejora el transito y baja la  
emisión de CO<sub>2</sub>.

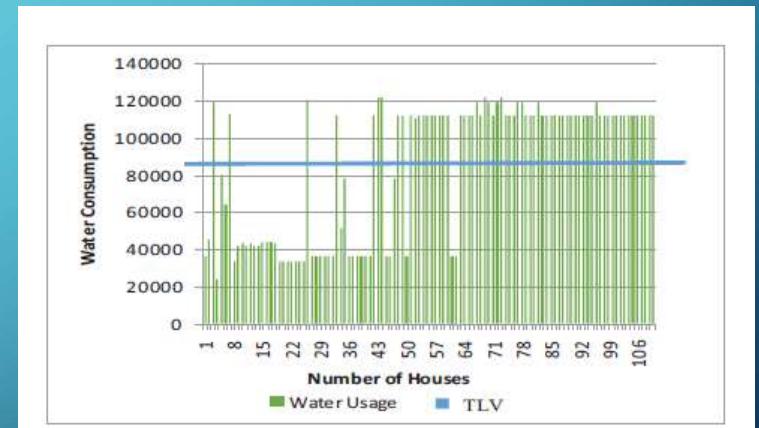
# PLANIFICACION URBANA CON IOT Y BIG DATA

## Caso 3: Consumo de Agua

Consumo de agua de hogares inteligentes de Surrey, Canadá (lecturas del medidor de 61263 casas)

El umbral para el consumo de agua es de 82000 litros por mes.

La cantidad de consumo de agua superior al umbral genera un evento para el departamento de gestión del agua.





## Módulo “Análisis de Datos Científicos y Geográficos”

### DESARROLLO DE SERVICIOS EN SMART CITY CON ANALISIS DE TEXTO EN RRSS

(Ver paper completo en Campus→Lecturas)



## BIG DATA Y RRSS PARA SMART CITY

**Developing Smart Cities Services through Semantic Analysis of Social Streams**

Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, Pasquale Lops

Desarrollo de una Herramienta dominio-agnóstica para el procesamiento inteligente de flujos de texto en RRSS, con aplicación a desarrollo en Smart Cities.



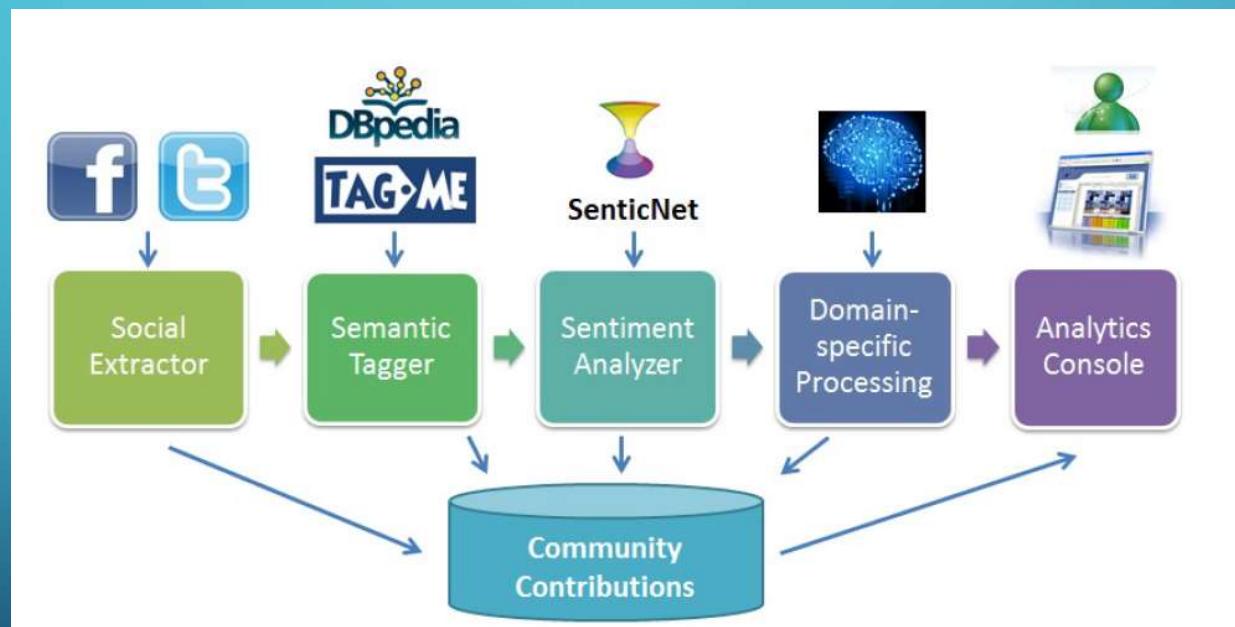
## **BIG DATA Y RRSS PARA SMART CITY**

### **Objetivo de la herramienta**

Extraer contenido de microblogs despierta cada vez mas atención, ya que hay mucha información latente sobre sentimientos, pensamientos, preferencias y opiniones de las personas que se puede extraer automáticamente de las secuencias de texto, allanando el camino para el desarrollo de nuevos servicios innovadores e inteligentes que dependen del análisis de contenido de estos datos.

# BIG DATA Y RRSS PARA SMART CITY

## Arquitectura de la Herramienta



Extrae, analiza y agrega datos para producir algunos análisis valiosos para los usuarios, de forma totalmente independiente del dominio → puede agregar y extraer cualquier tipo de contenido que el usuario quiere analizar



## **BIG DATA Y RRSS PARA SMART CITY**

### **Extraccion: componente principal**

Dadas algunas heurísticas de extracción, la herramienta se conecte a una red social y extrae algún contenido que coincida con la heurística.

Extrae contenido de Facebook y Twitter explotando las APIs.



# BIG DATA Y RRSS PARA SMART CITY

## Extraccion: componente principal

Heurísticas planteadas:

- Contenido: extrae todos los tweets que coincidan con un término especificado
- Usuario: extrae todos los tweets publicados por un usuario específico
- Geo: extrae todos los Tweets geolocalizados, dada la latitud, longitud y radio
- Contenido + Geo: extrae todos los Tweets geolocalizados que coincide con un término específico
- Página: extrae todas las publicaciones de Facebook provenientes de una página específica (la publicación principal y las respuestas)
- Grupo: extrae todas las publicaciones de Facebook que provienen de un grupo específico (los mensajes principales y las respuestas)



## BIG DATA Y RRSS PARA SMART CITY

### Extraccion: componente principal

Solo contenido público: extracción masiva a gran escala de contenido, sin necesidad de una autorización explícita de los usuarios.

Toda la información extraída es luego anonimizada y almacenada en una base MongoDB en forma continua durante todo el intervalo de tiempo del análisis deseado.

La base obtenido se llamará **Contenido Social** (sin importar las fuentes involucradas)

# BIG DATA Y RRSS PARA SMART CITY

## Análisis Semántico

Dado que el proceso de extracción se basa en simples coincidencias sobre palabras clave, se extrae un gran cantidad de contenido irrelevantes para el estudio de un caso.

*Ejemplo:*

*Supongamos que el término "L'Aquila" se usa para extraer todo*

*Tweets donde la gente habla sobre la ciudad afectada por el terremoto en 2009.*

*El problema es que L'Aquila es un término polisémico, que en italiano significa “águila”.*



Albero Vagabondo @AlberoVagabondo · 2 feb 2013

Lupo, aquila, lontra, cicogna nera. Sono animali rari, preziosi che abitano l'irpinia e sono minacciati d... [fb.me/1TOLb1QPj](http://fb.me/1TOLb1QPj)



PrimaDaNoi.it @PrimaDaNoi · 13 ott

Cialente lancia Sos all'Europa: «L'Aquila sta morendo» [ift.tt/1wsNVlm](http://ift.tt/1wsNVlm)

# BIG DATA Y RRSS PARA SMART CITY

## Análisis Semántico

Técnicas de PNL: + Semántica y – Ruido

**Un extra: encontrar algunos conceptos de alto nivel que pueden proporcionar una visión más amplia y abstracta de los datos → Algoritmos de enlace de entidades DBpedia Spotlight, Wikipedia Miner and Tag.me.**

Al mapear a Wikipedia, podemos enriquecer aún más el contenido mediante la introducción de las categorías de ancestros más relevantes de esa página.



Democrats Politics,  
L'Aquila Mayors

# BIG DATA Y RRSS PARA SMART CITY

## Análisis de Sentimientos

Se enriquece el contenido mediante el análisis de la opinión transmitida en cada contenido social con Puntaje Categórico (positivo, negativo, neutral) o Numérico (score)

Se utiliza SenticNet: Ofrece puntajes de sentimiento (en un rango entre -1 y 1) para 14,000 conceptos de sentido común.

*Ejemplo: contiene solo un término con una clara polaridad (morir), lo que influye negativamente*



Puntuación  
SenticNet  
-0.235



## BIG DATA Y RRSS PARA SMART CITY

### Procesamiento de Dominio Específico

Algunos pasos adicionales de procesamiento específicos del dominio. Podrían variar desde la aplicación de Técnicas basadas en el aprendizaje automático (como clase de texto o minería redes sociales) a las aplicaciones simples de heurísticas para dispersar o enriquecer los datos previamente extraídos.

*Se verán estrategias puntuales en cada caso de uso*



# BIG DATA Y RRSS PARA SMART CITY

## Consola de Análisis

Permite que el usuario visualice e interactúe con los resultados agregados del análisis.

Tres tipos de visualización:

- Mapas
- Nube de Etiquetas
- Gráficos

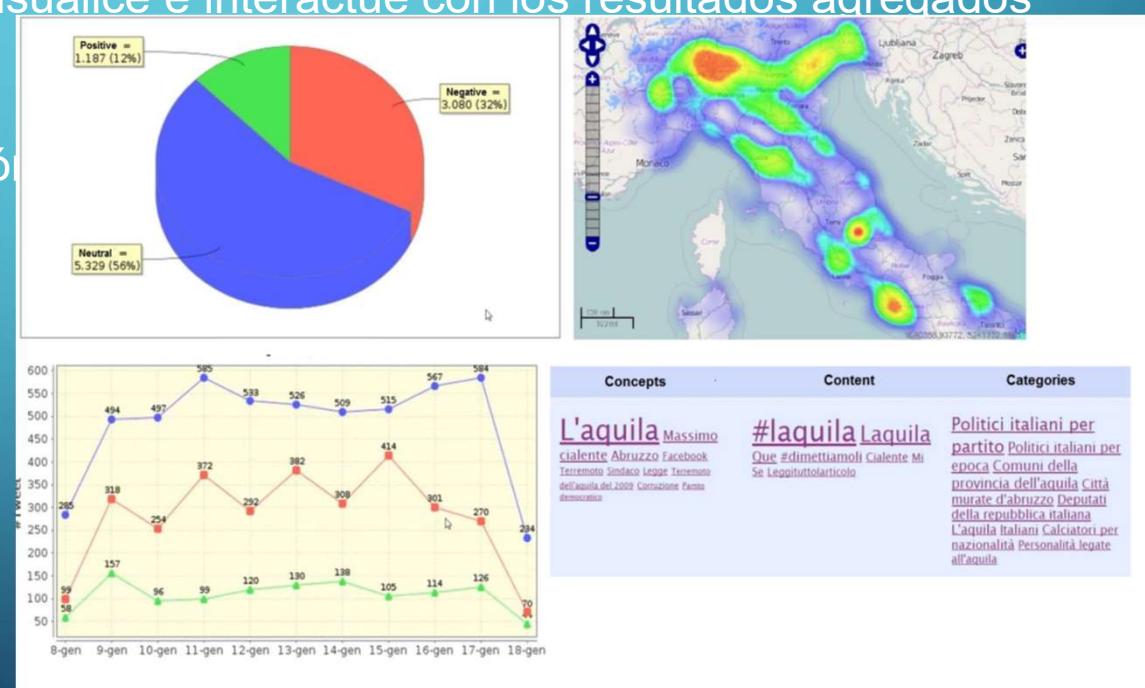
# BIG DATA Y RRSS PARA SMART CITY

## Consola de Análisis

Permite que el usuario visualice e interactúe con los resultados agregados del análisis.

Tres tipos de visualización:

- Mapas
- Nube de Etiquetas
- Gráficos



# BIG DATA Y RRSS PARA SMART CITY

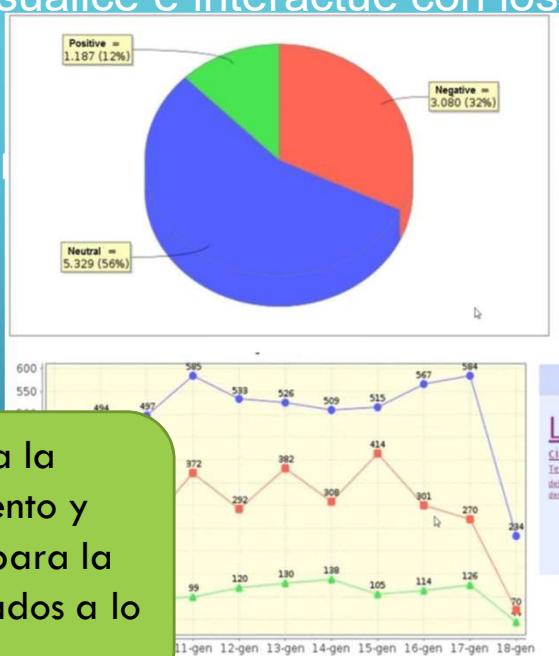
## Consola de Análisis

Permite que el usuario visualice e interactúe con los resultados del análisis.

Tres tipos de visualización:

- Mapas
- Nube de Etiquetas
- Gráficos

Diagrama pastel para la distribución del Sentimiento y gráfico de líneas se usa para la cantidad de tweets publicados a lo largo del tiempo

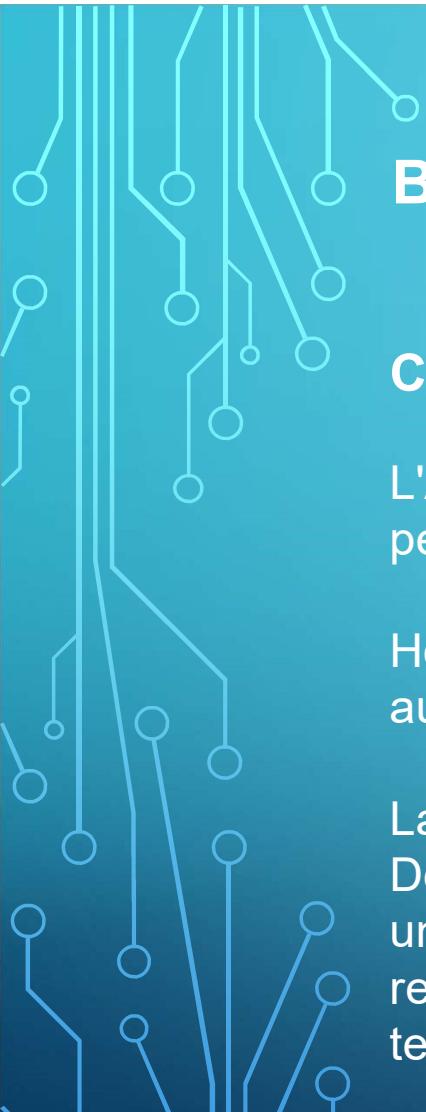


Ejemplo de utilidad: verificar el opinión de los ciudadanos sobre medidas administrativas recientes diferentes áreas de la ciudad, o para comprobar cuán popular es un tema en un área particular



Concepts	Content	Categories
L'aquila Massimo caliente Abruzzo Facebook Terremoto Jiloca Legge Intemperie dell'aquila del 2009 Corruzione Punto denuncia	#laquila Laquila Que #edimettiamoli Caliente Mi Se Legittimato articolo	Politici italiani per partito Politici italiani per epoca Comuni della provincia dell'aquila Città

Diferenciadas por Conceptos, Contenido, Categorías (es dinámico)



## BIG DATA Y RRSS PARA SMART CITY

### Caso de Uso 1: Red Social Urbana de L'Aquila

L'Aquila sufrió un tremendo terremoto en abril de 2009 que mató a 297 personas.

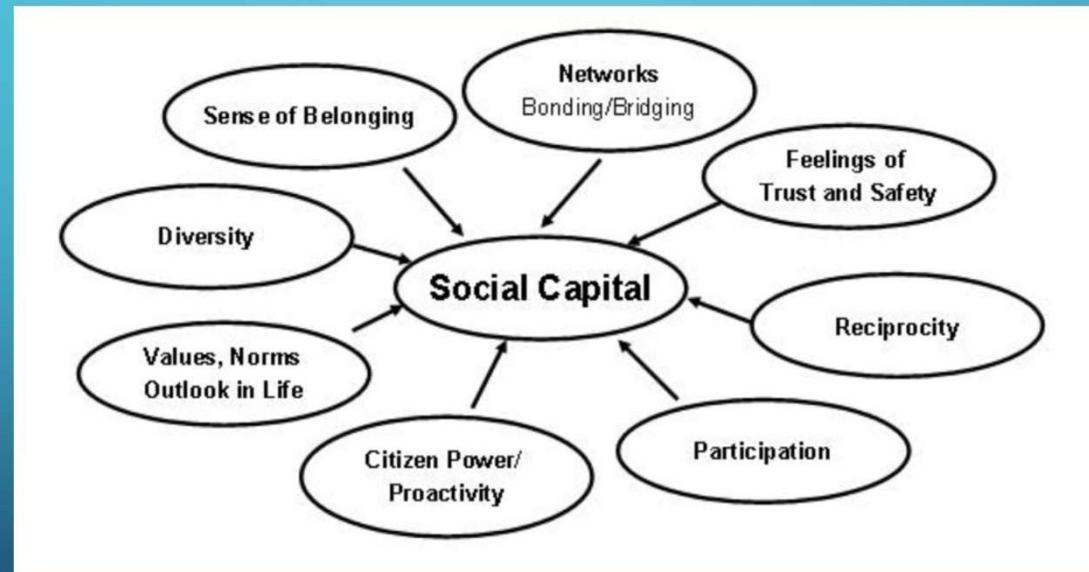
Hoy en día, el trauma severo a las estructuras físicas y psicosociales está aún en la fase de recuperación.

La ENEA (Agencia Nacional Italiana para Nuevas Tecnologías, Energía y Desarrollo Económico Sustentable) propuso el SUN (Social Urban Network), un proyecto relacionado con ciudades inteligentes, destinado a potenciar y revitalizar el patrimonio urbano y el capital social de la ciudad después del terrible terremoto.

# BIG DATA Y RRSS PARA SMART CITY

## Caso de Uso 1: Red Social Urbana de L'Aquila

Se definieron los siguientes 8 indicadores para SUN:





# **BIG DATA Y RRSS PARA SMART CITY**

## **Caso de Uso 1: Red Social Urbana de L'Aquila**

Social Extractor:

Para Facebook → páginas y grupos específicos gestionados por ciudadanos de L'Aquila

Para Twitter → heurísticas GEO (radio 50 km desde centro L'Aquila) y USER (tweets sobre principales periódicos locales y sus re-tweets)



# **BIG DATA Y RRSS PARA SMART CITY**

## **Caso de Uso 1: Red Social Urbana de L'Aquila**

Procesamiento de Dominio Específico:

Se explotó un conjunto de etiquetas ejemplo para aprender un modelo de clasificación multiclas. Luego, el modelo fue capaz de asociar cada nuevo texto al indicador social correspondiente.

Analisis de Sentimiento

Para proporcionar a cada indicador social un puntaje, cada contenido extraído ha sido procesado a través del Analizador de Sentimiento y el puntaje general de cada indicador social se ha obtenido sumando el puntaje de sentimiento de cada contenido refiriéndose a ese indicador

# BIG DATA Y RRSS PARA SMART CITY

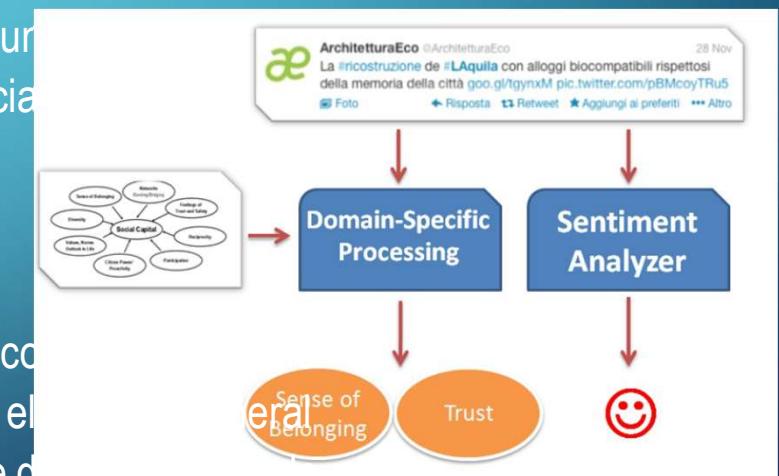
## Caso de Uso 1: Red Social Urbana de L'Aquila

Procesamiento de Dominio Específico:

Se explotó un conjunto de etiquetas ejemplo para aprender una clasificación multiclas. Luego, el modelo fue capaz de asociar nuevo texto al indicador social correspondiente.

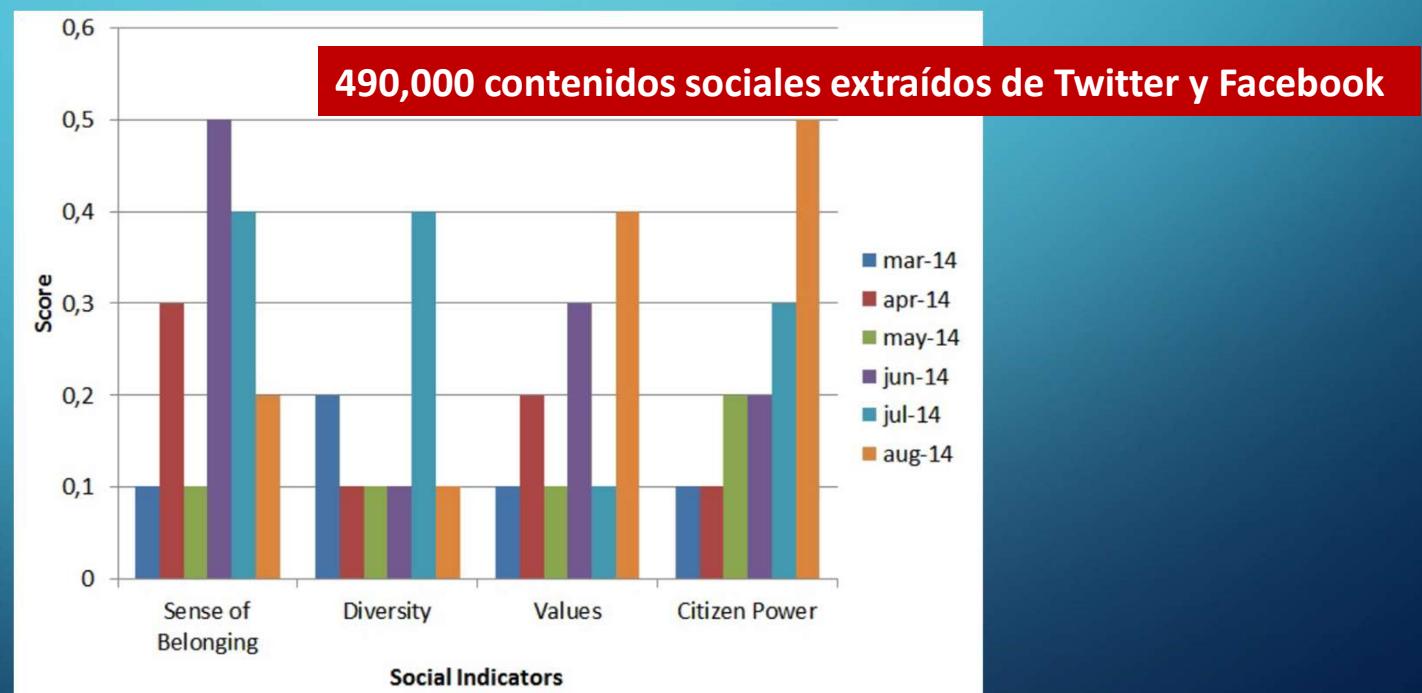
Análisis de Sentimiento

Para proporcionar a cada indicador social un puntaje, cada contenido ha sido procesado a través del Analizador de Sentimiento y el puntaje de cada indicador social se ha obtenido sumando el puntaje de todos los contenidos que se refieren a ese indicador.



# BIG DATA Y RRSS PARA SMART CITY

## Caso de Uso 1: Red Social Urbana de L'Aquila





## **BIG DATA Y RRSS PARA SMART CITY**

### **Caso de Uso 2: El Mapa del Odio Italiano**

El objetivo del proyecto fue analizar el contenido producido en las redes sociales para medir el nivel de intolerancia del país italiano y para orientar la definición de intervenciones específicas (recuperación y prevención) en el territorio.

El análisis fue realizado para varias facetas: homofobia, racismo, violencia contra las mujeres, antisemitismo y discapacidad



## BIG DATA Y RRSS PARA SMART CITY

### Caso de Uso 2: El Mapa del Odio Italiano

Extractor Social

Definición de un conjunto de términos razonables realizado por psicólogos con experiencia específica en este dominio → 47 términos

Solo se usó Twitter, ya que debido a las políticas de Facebook, no hay grupos o páginas con un claro propósito homofóbico o racista está disponible en la plataforma.



# BIG DATA Y RRSS PARA SMART CITY

## Caso de Uso 2: El Mapa del Odio Italiano

Procesamiento Semántico

Se eliminaron tweets ambiguos y no intolerantes.

Análisis de Sentimiento

Se eliminaron los tweets con una puntuación de sentimiento neutral o positiva

Se utilizó técnica GEO (utilizando geolocalización de un usuario, aun utilizándola desde tweets no intolerantes para marcar uno que si lo era)

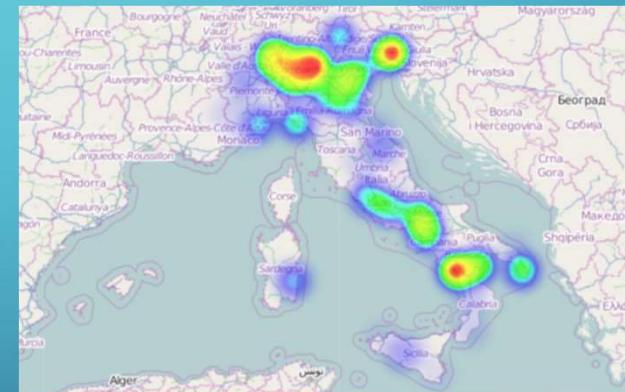
Dimension	#Tweets	#Geo	%Geo
Homophobia	110,774	8,501	7,66%
Racism	154,170	1,940	1,24%
Violence	1,102,494	28,886	2,62%
Disability	479,654	3,410	0,75%
Anti-Semitism	6,000	1,150	18,03%

# BIG DATA Y RRSS PARA SMART CITY

## Caso de Uso 2: El Mapa del Odio Italiano



Homofobia



Racismo

Ventana de 7 días