UNIVERSIDADE DA CORUÑA

Workshop on
Big Data and Statistics

LIDIA
Laboratorio de I+D en Inteligencia Artificial

INVESTIGAR EN BIG DATA LEARNING: ¿UNA PROFESIÓN CON FUTURO?

AMPARO ALONSO-BETANZOS, Junio 2015

¿Por qué Big Data?

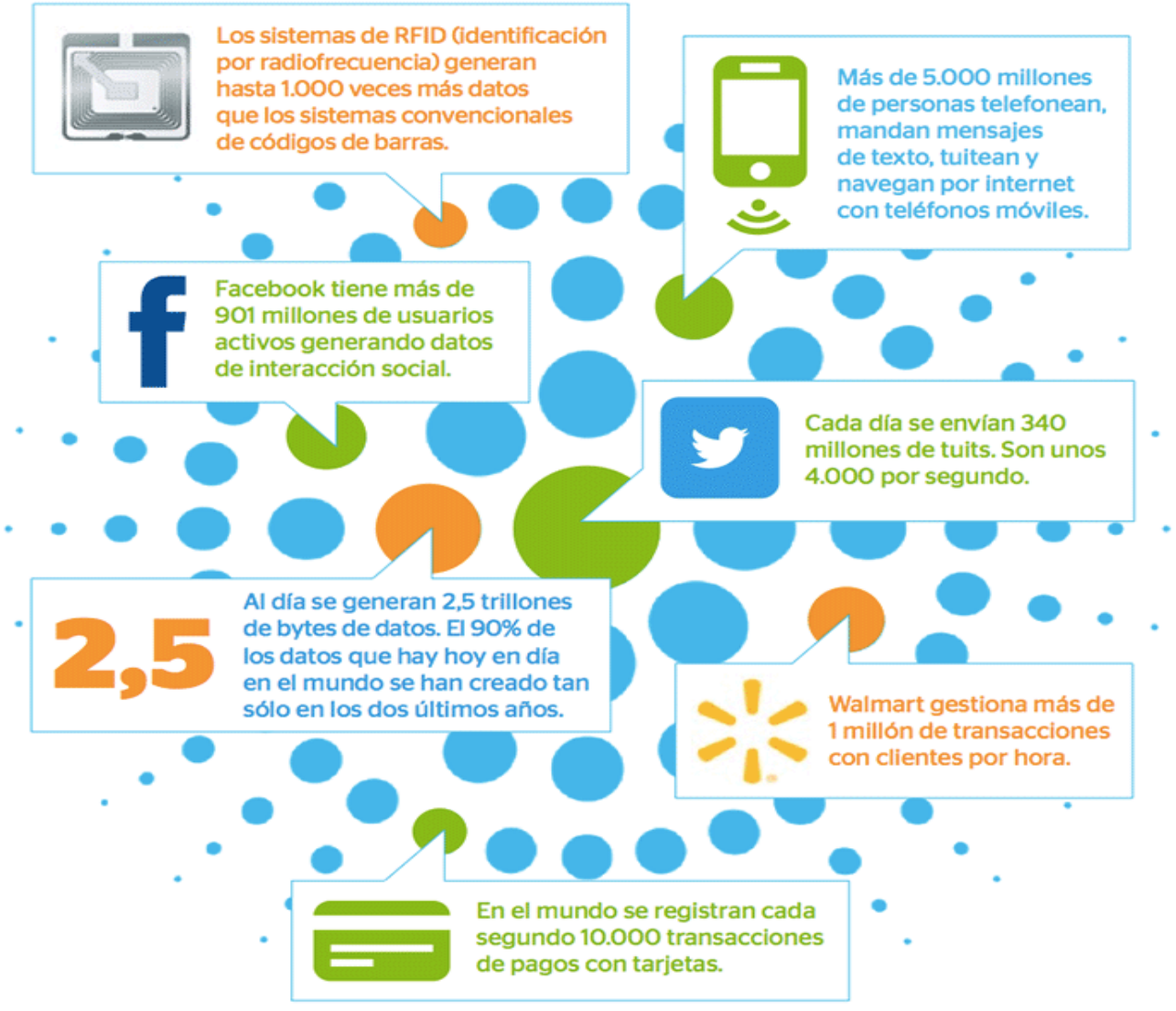
- 2,5 quintillones bytes/ día
- 90% de los datos que existen se han creado en los dos últimos años
- 5V: Volumen, Velocidad, Variedad, Valor y Veracidad

Herramientas software sin capacidad de actuación en tiempos admisibles

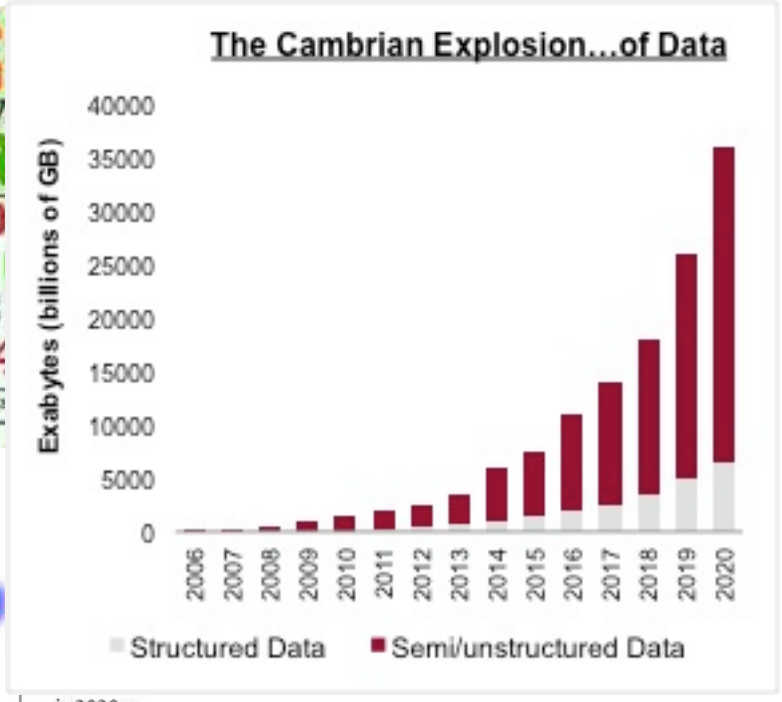
Científicos de datos: Una profesión nueva

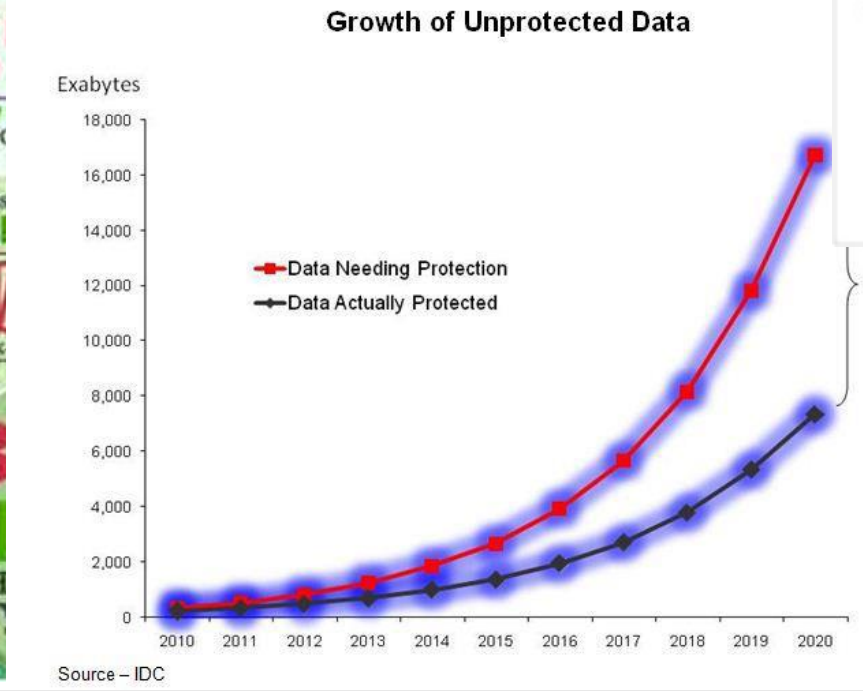
- Harvard Bussiness Review, October 2012

¿De dónde salen los datos?



Y no paran de crecer….en las 5V





Las 5V de Big Data

V de Volumen

Lo que antes se consideraba grande, ahora ya no lo es tanto.

- Unidad 'básica' de almacenamiento GB (10^9 bytes)
- Big Data habla en PB (10^{15} bytes)..
- … y pasando a Zettabytes (10^{21} bytes)

Algunas cifras….

2.5 exabytes de datos/día, duplicando cada 40 meses.

+ datos/s. que total almacenados hace 20 años.

Walmart

+ 2.5 PB datos de transacciones de clientes/h.

V de Big Data

V de Velocidad

- Gran volumen de datos sin variaciones rápidas → días, horas análisis.
- El montante de información crece por TB

tiempo de procesamiento de la información

V de Velocidad

V de Big Data

V de Variedad

- Big Data no versa en la mayoría de ocasiones sobre datos estructurados
- No es sencillo incorporar grandes volúmenes a una base de datos relacional.
- Infinidad de tipos de datos se aglutinan dispuestos a ser tratados y **aumenta el grado de complejidad** en almacenamiento y análisis.

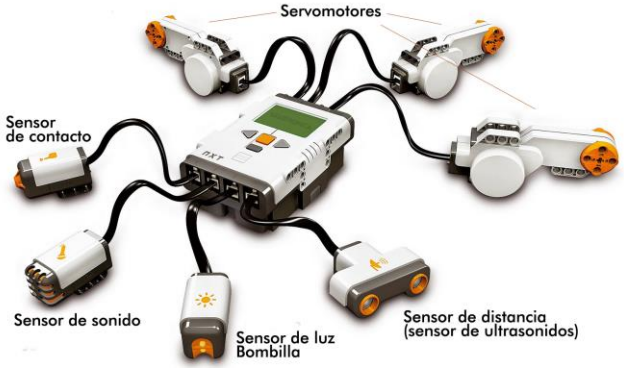
¿Variedad?

2006

2004

2007

2010



V de Big Data

V de Veracidad

Fuente: Celent

Veracidad

Balance tiempo/esfuerzo

V de Big Data

V de Valor

¿Y entonces qué?

Director de Investigación

Google

"We don't have better algorithms. We just have more data."

Las 10 top áreas de Big Data

1. Entender Clientes y centrarse en ellos

- Comportamientos y Preferencias
  - Datos tradicionales
  - Browser logs
  - Analítica de textos
  - Datos de sensores

Crear modelos predictivos

- Ejemplos:
  - US retailer Target, capaz de predecir cuando un cliente espera un bebé.
  - Compañías telefónicas pueden predecir cuando un cliente les abandona
  - Walmart puede predecir qué productos se venderán
  - Aseguradoras de coches pueden entender cómo conducen sus clientes
  - El papel de Big Data Analytics en la campaña a las presidenciales de Obama en 2012.

Las 10 top áreas de Big Data

2. Entender y optimizar procesos de negocio

- Optimización de stock basado en predicciones de datos de redes sociales, búsquedas de tendencias en la web, predicciones meteorológicas.
- Optimización de cadenas de suministros o rutas de entrega
- Procesos de búsqueda de Recursos Humanos

Las 10 top áreas de Big Data

3. Cuantificación personal y optimización de rendimiento

- Mejora personal basada en wearables

Jawbone

- Análisis de datos collectivos a partir de wearables personales
- Love-online

Las 10 top áreas de Big Data

4. Mejorar áreas de Salud y Salud pública

- Decodificar cadenas DNA, nuevos tratamientos, predicción de patrones de enfermedades
- Estudios clínicos con grandes muestras
- Monitorización de bebés en Unidades Prematuros y UCI neonatales y pediátricas
- Predecir el desarrollo de epidemias

Las 10 top áreas de Big Data

5. Mejorar rendimiento deportivo

- IBM SlamTracker para competiciones de tenis del Grand Slam
- Monitorizar el rendimiento de todos los jugadores de fútbol americano o baseball
- Tecnología sensórica en equipo deportivo (pelotas de baloncesto, golf, etc)
- Monitorizar atletas fuera de su entorno (hábitos nutrición, sueño, estado emocional, etc.)

Las 10 top áreas de Big Data

6. Mejorar la Ciencia y la Investigación

- El centro de datos del CERN tiene 65.000 procesadores para analizar sus 30 petabytes de datos.

- Usa la capacidad computacional de miles de computadores distribuidos en 150 centros de datos.

Las 10 top áreas de Big Data

7. Optimizar el rendimiento de las máquinas

- Maquinaria más inteligente y autónoma
  - Coche de Google
  - Toyota Prius
  - Ahorro energético

Las 10 top áreas de Big Data

8. Mejorar la seguridad y reforzar la ley

- NSA usa Big data para frustrar planes terroristas
- Detectar y prevenir ciberataques
- Predecir actividad criminal
- Detectar transacciones fraudulentas con tarjetas de crédito

Las 10 top áreas de Big Data

9. Mejorar y optimizar ciudades y países

- Optimizar flujos de tráfico usando información tráfico en tiempo real, redes sociales, datos climatológicos, etc.

- Proyectos Smart Cities

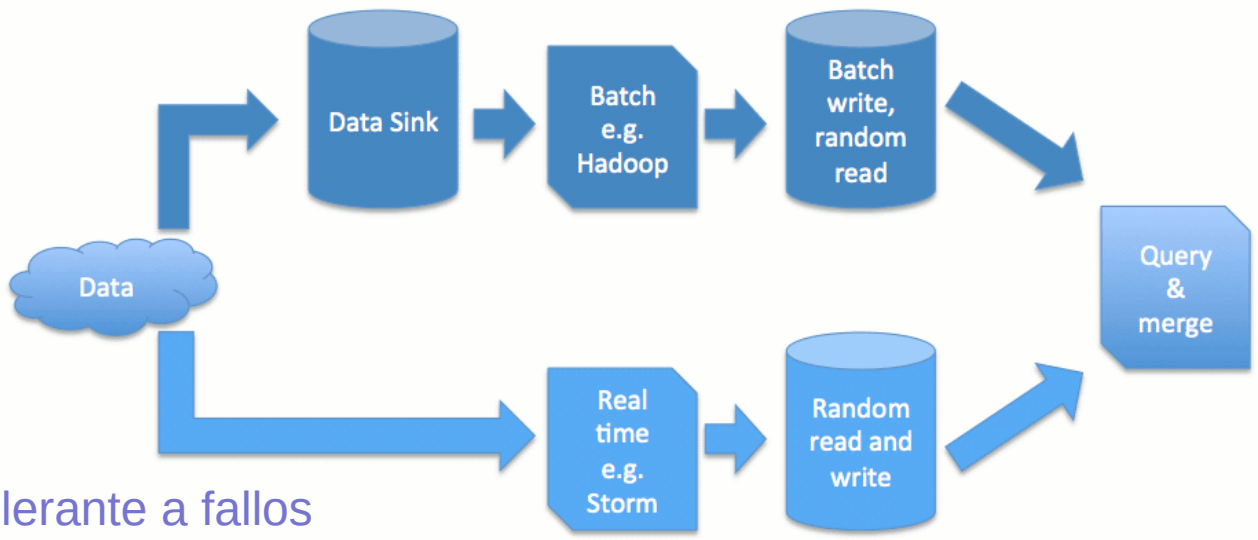
Las 10 top áreas de Big Data

10. Toma de decisiones en negocios. Tráfico financiero

Arquitecturas y casos de uso

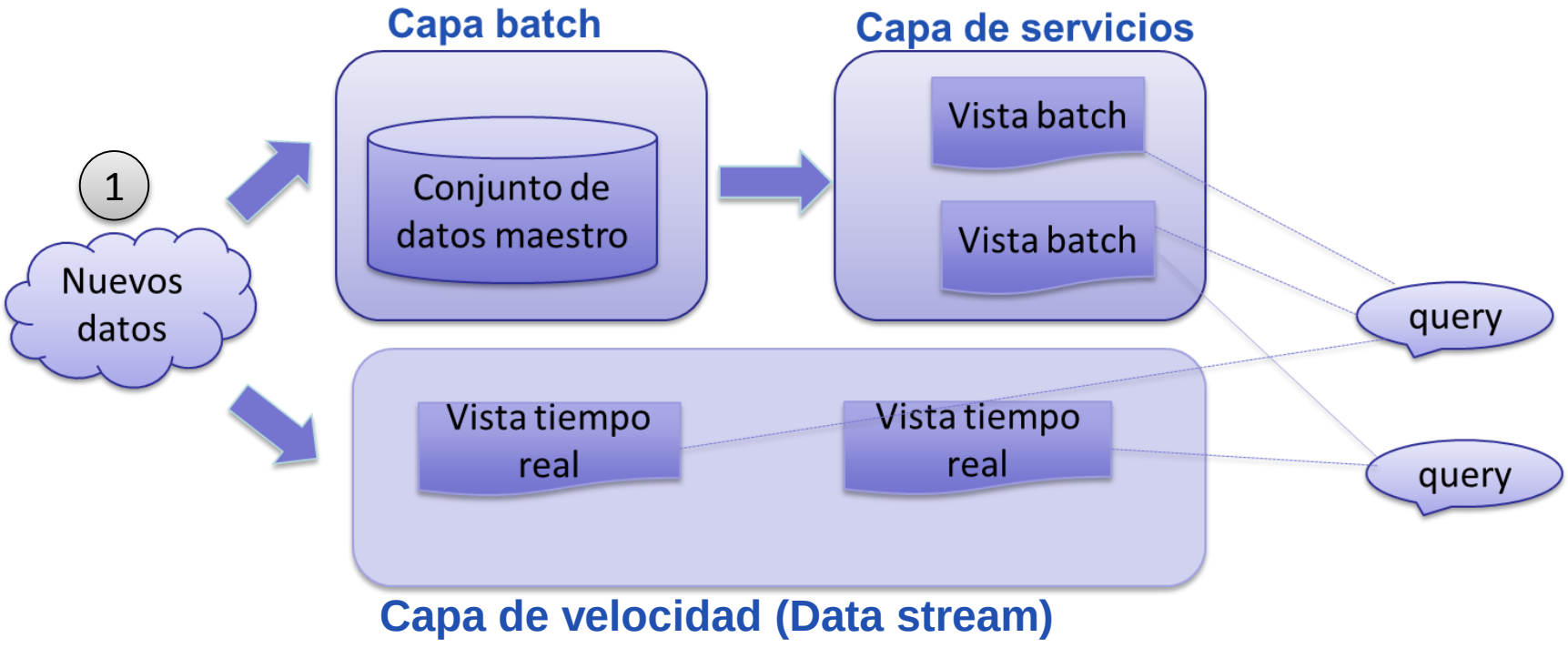
Arquitectura lambda

- Paradigmas de arquitecturas en Big Data:
  - Batch
    - Hadoop/MapReduce
  - Tiempo real
    - Storm
  - Solución híbrida
    - Arquitectura Lambda



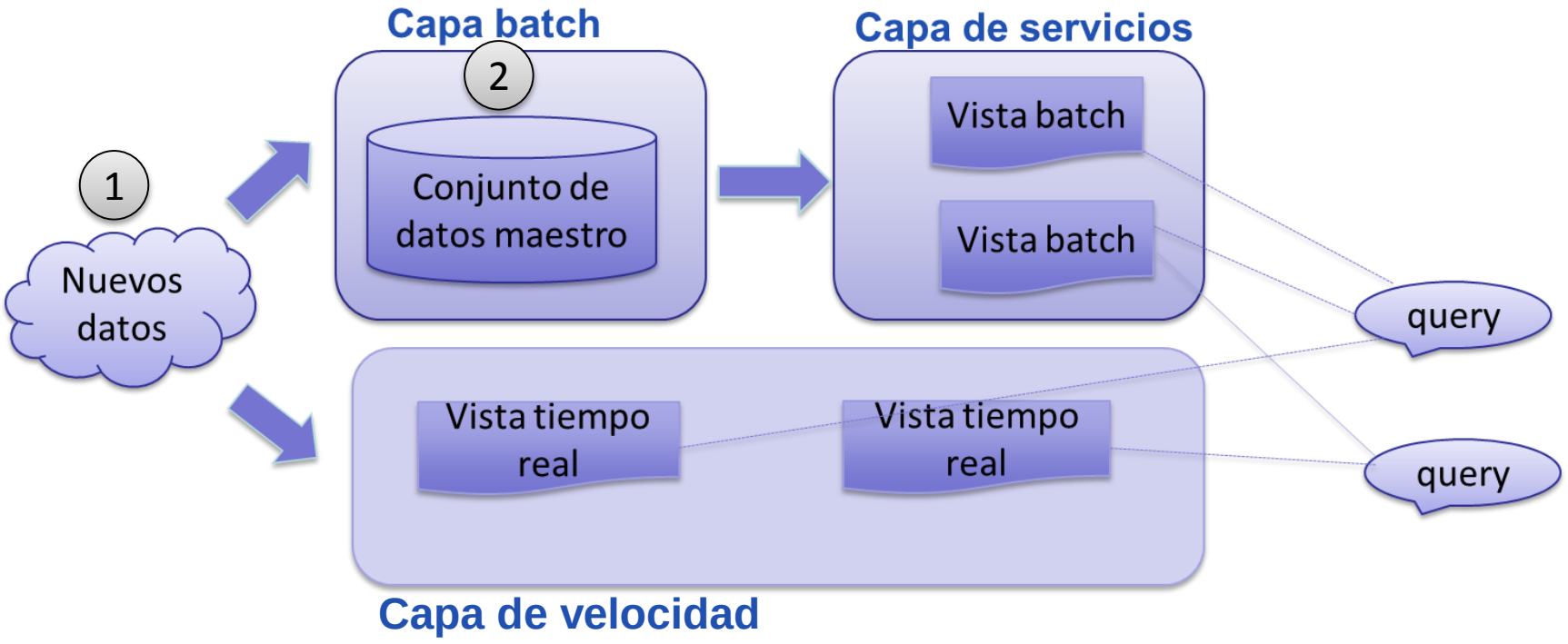
Genérica, escalable y tolerante a fallos

Arquitectura lambda



1 Todos los datos que entran se envían a ambas capas para su procesado

Arquitectura Lambda

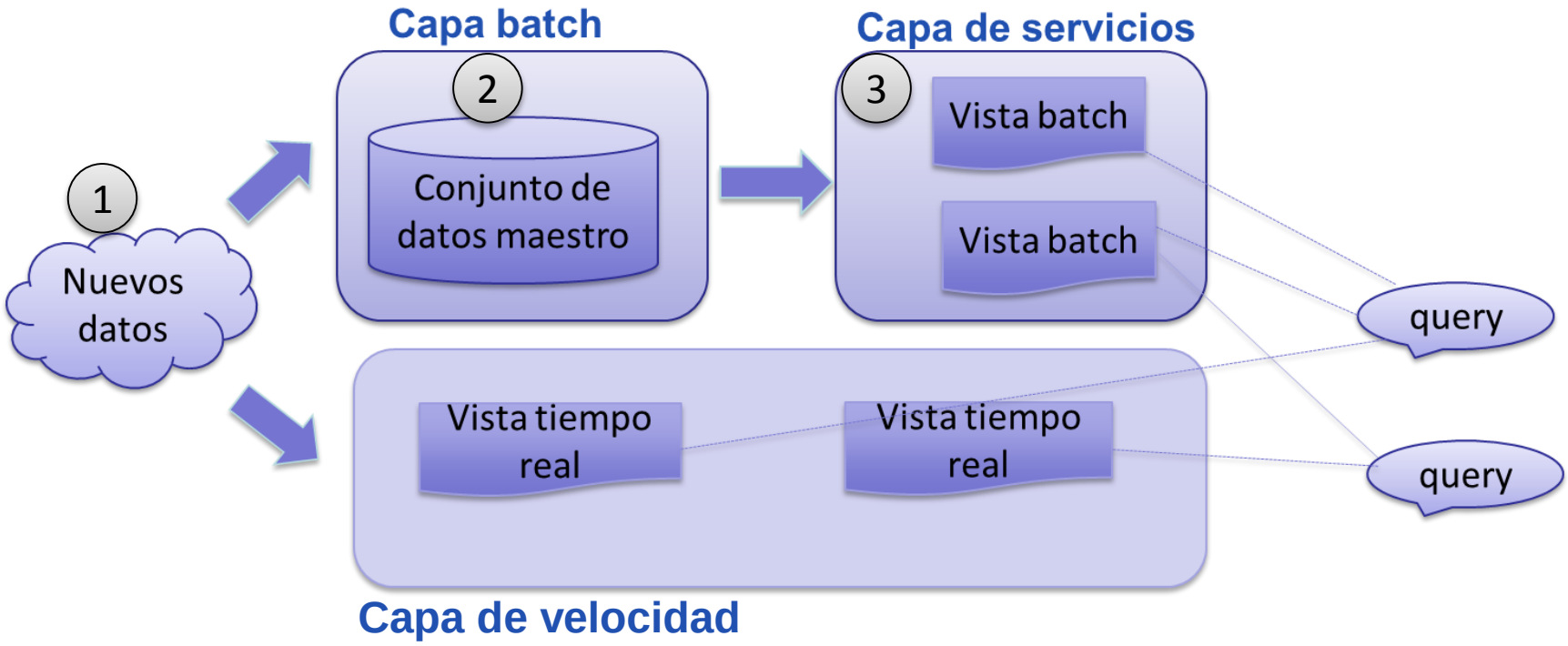


(1) gestionar el conjunto de datos maestro (immutable, append-only)

2

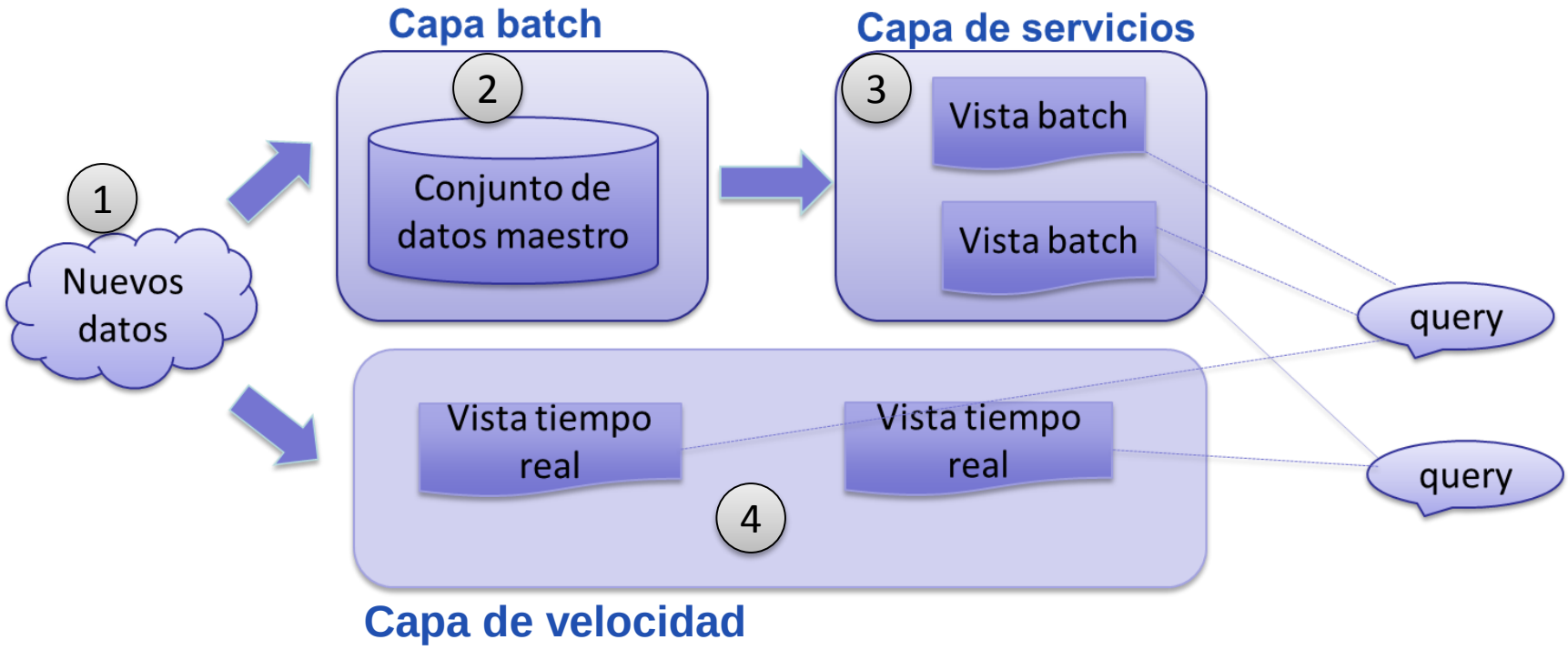
(2) Precomputar las vistas batch.

Arquitectura Lambda



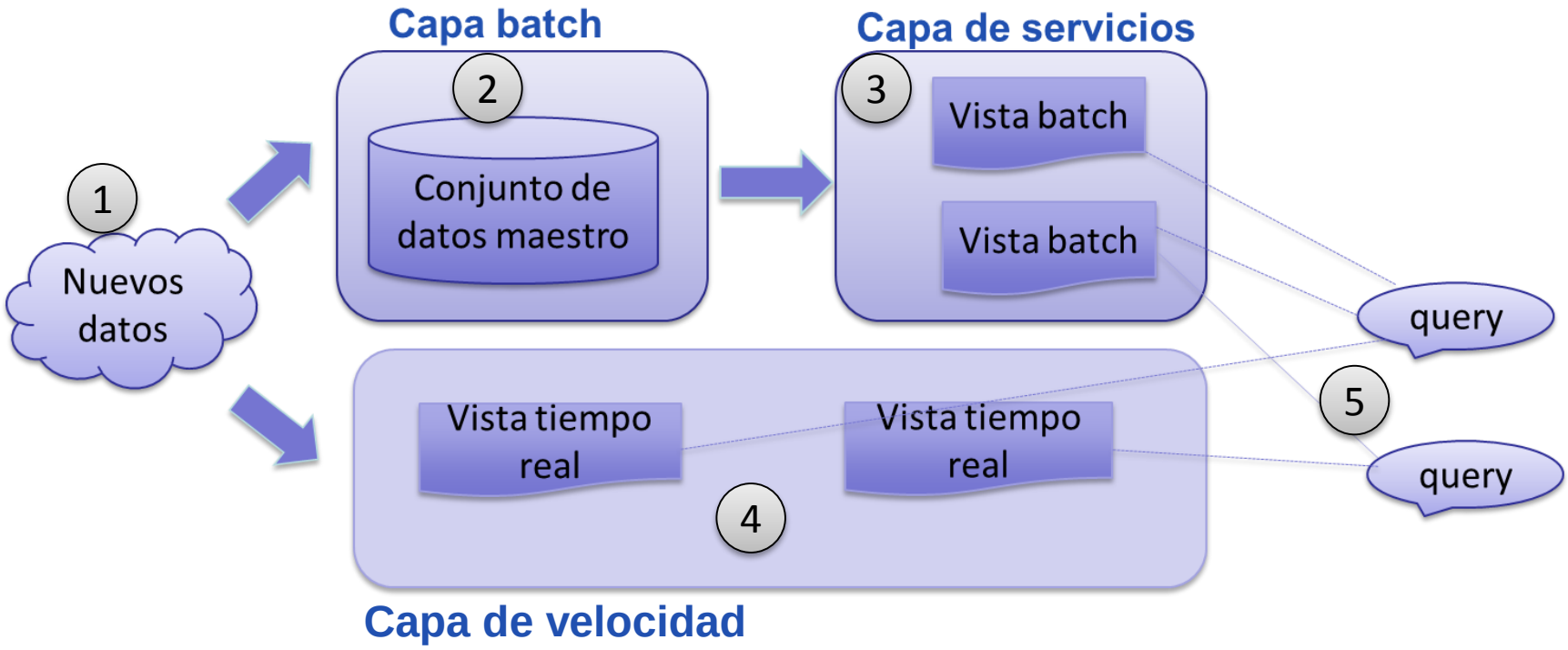
3 Indexa las vistas batch para queries ad-hoc y con baja latencia.

Arquitectura Lambda



4 Compensa a la capa de servicios la alta latencia de las actualizaciones y solamente trata con datos recientes

Arquitectura Lambda



5 Cualquier query entrante se puede responder uniendo resultados de las vistas batch y tiempo real.

Ventajas

- Procesado de datos con alta precisión
- Compensación entre las dos capas
- Tolerancia a fallos
- Útil en varios workloads y para múltiples casos de uso
- Sistema final escalable linealmente

RAD - Outlier Detection on Big Data

Problema:

- Detección de anomalías en bases de datos crecientes
- Netflix tiene gran número de bases de datos de gran crecimiento
- Se necesita un proceso automático de detección de anomalías

ROBUST ANOMALY DETECTION (RAD)

Retos genericos:

- Alta cardinalidad en dimensiones
- Minimización de falsos positivos
- Estacionalidad
- Los datos no siguen siempre una distribución normal

RAD - Outlier Detection on Big Data

Algoritmo:

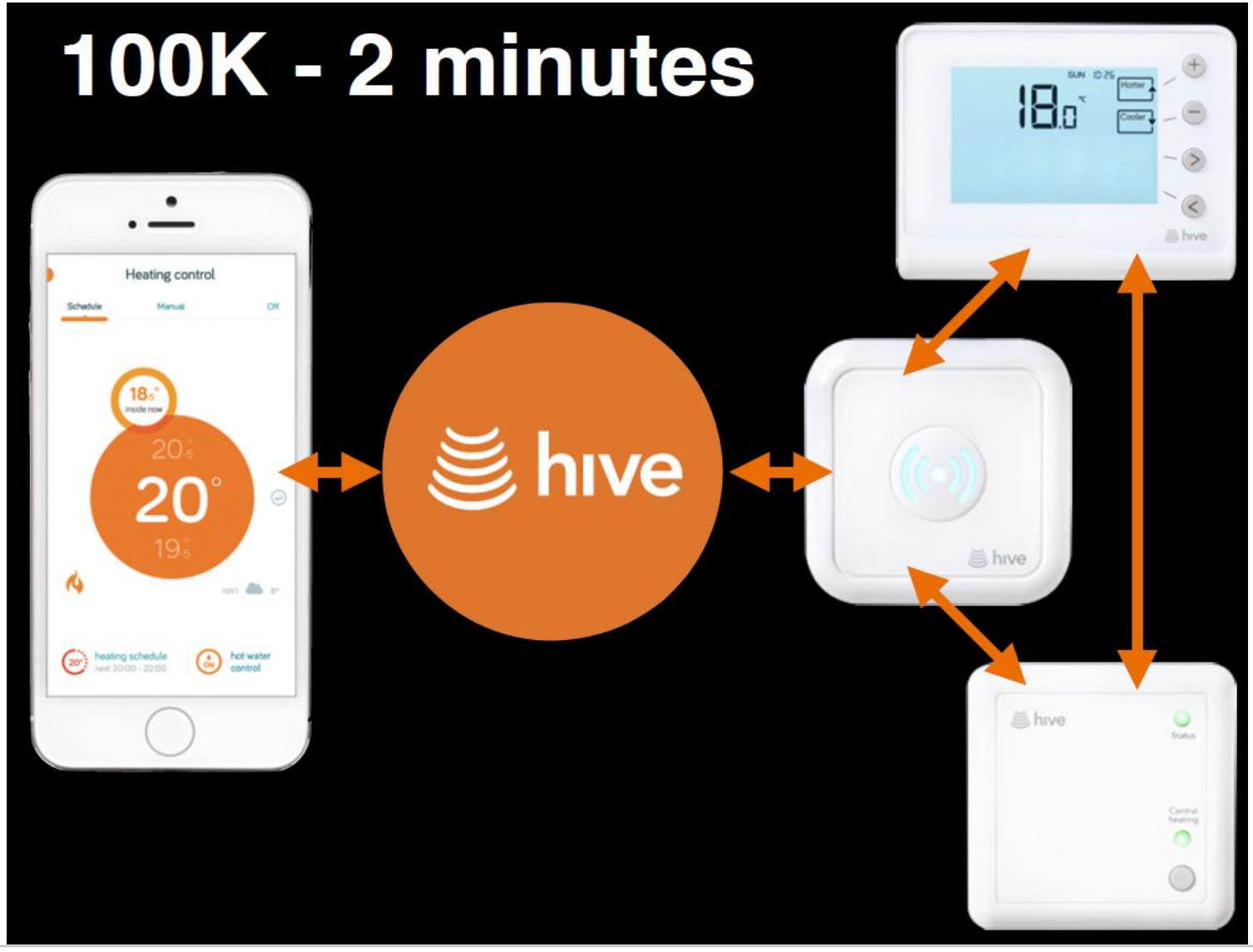
- Robust PCA http://netflix.github.io/demos/rsvd/
- ETL es Apache Pig

Procesos en Netflix:

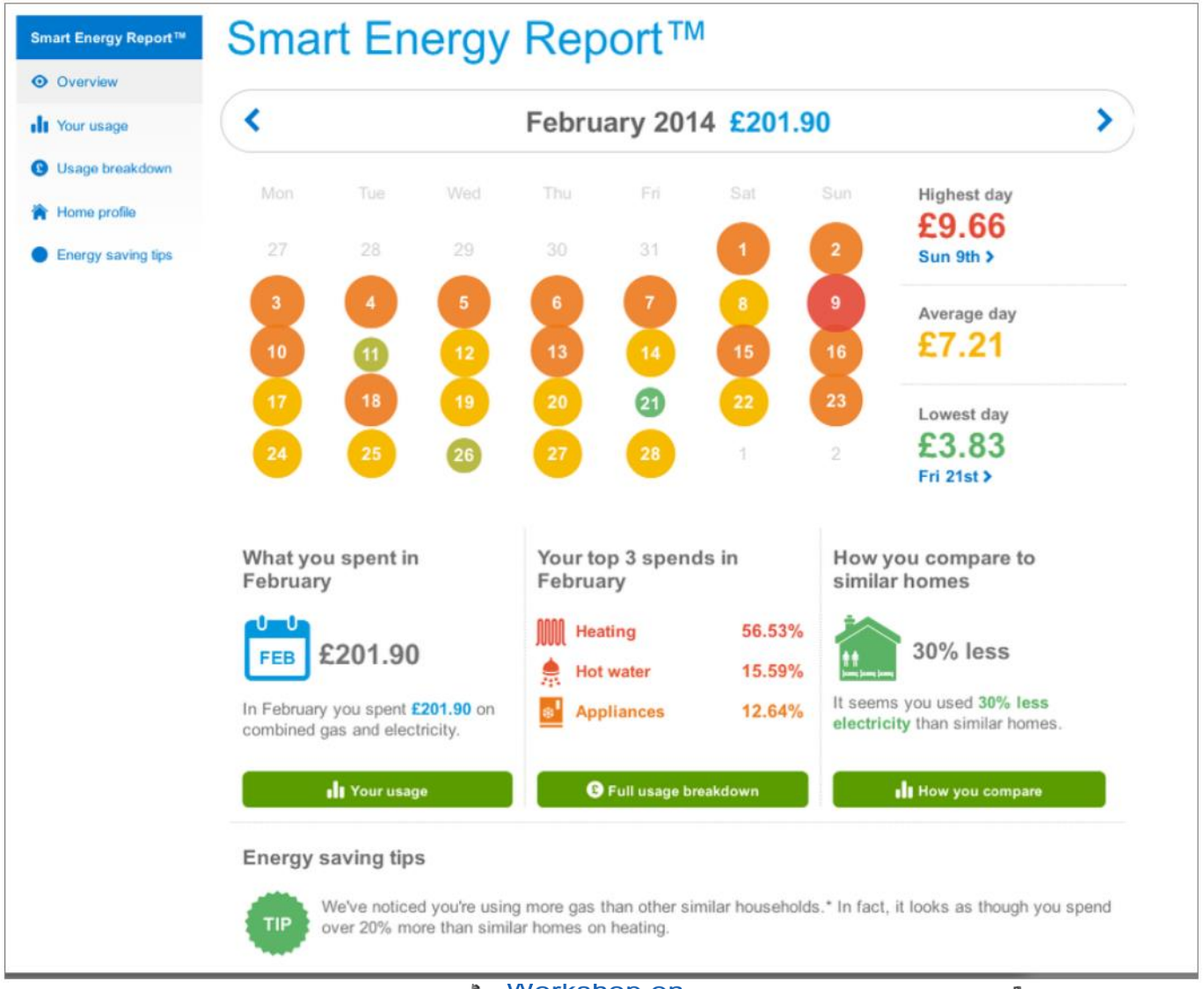
- Detección de anomalías en los fallos en la red de pagos al nivel del banco
  - Millones de transacciones diarias a través de decenas de miles de bancos en batch y tiempo real

- Identificar anomalías en el flujo de inscripciones a Netflix

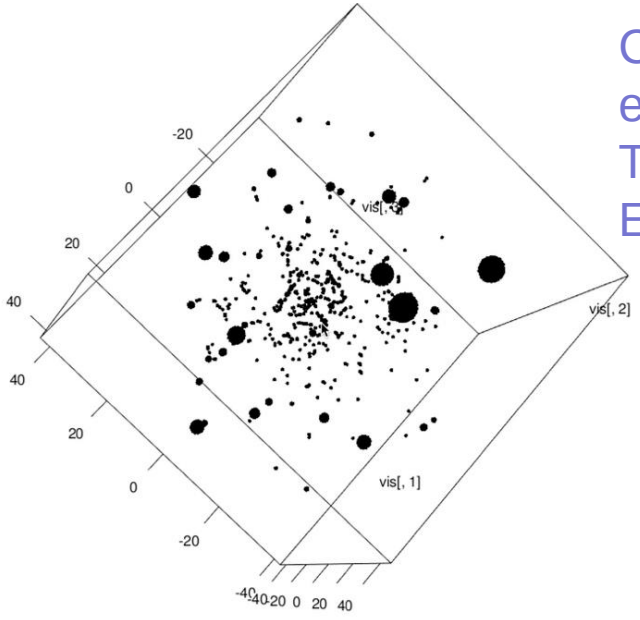
Ejemplos: Hive- Connected Homes in UK



Ejemplo: Hive



Técnicas en Hive



Conjunto de datos de 218 dimensiones expresado en 3D
T-distributed Stochastic Neighbour Embedding

Spark

RabbitMQ

R

C*

Scala

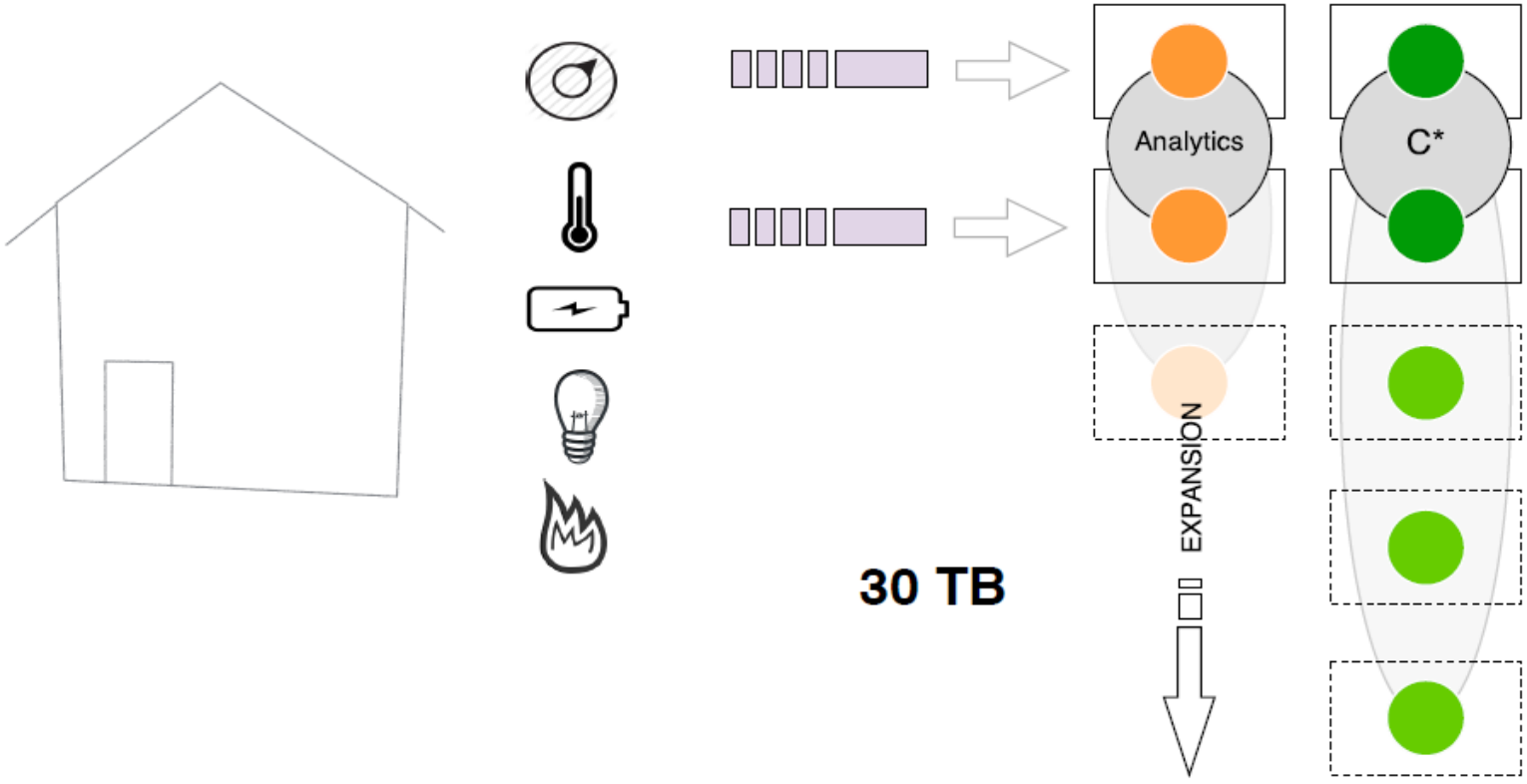
Python

Java

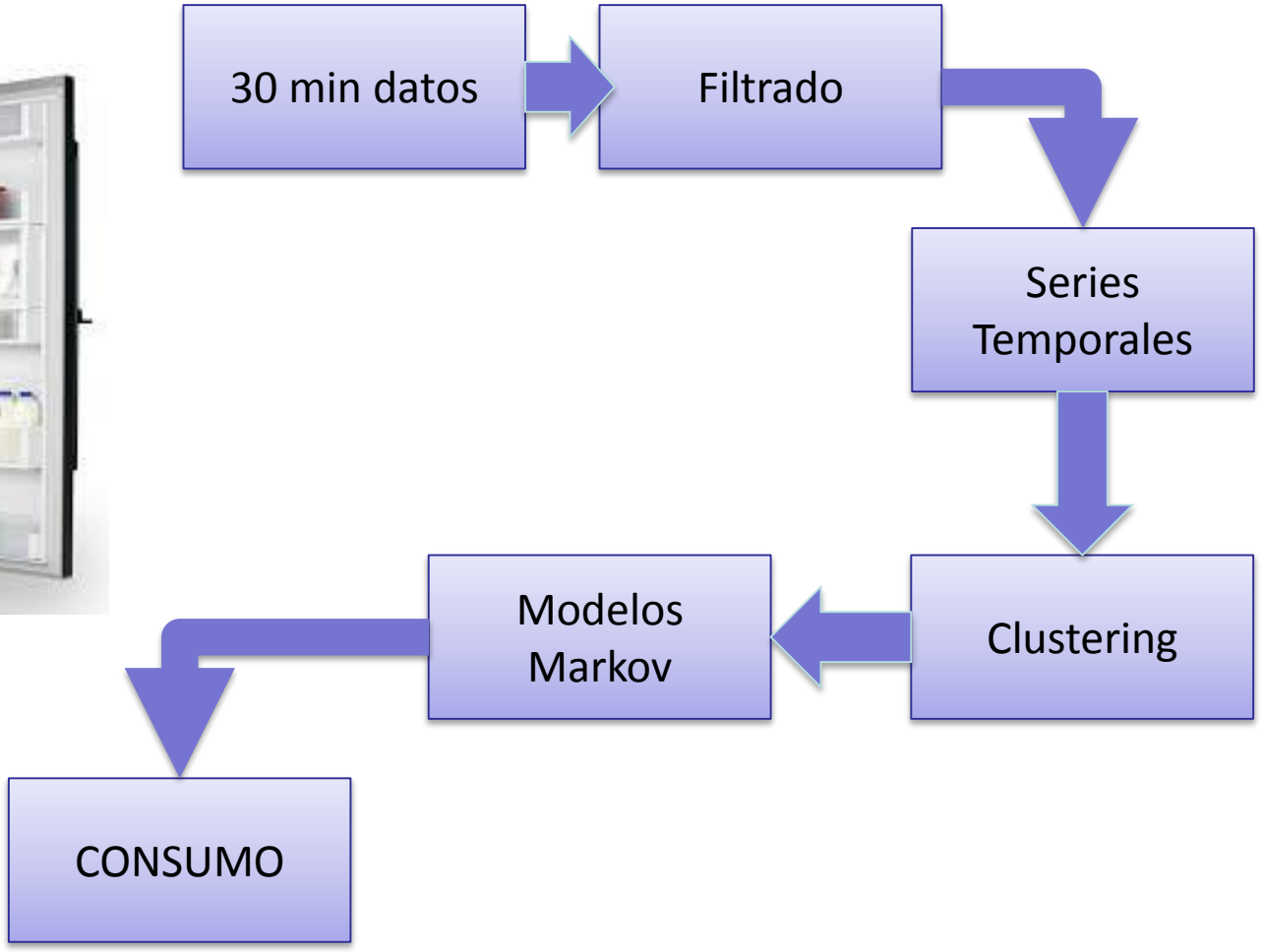
DataStax OpsCenter

Esquema general

Recogida de datos de sensores en el hogar → Procesado → Almacena miento



Lo excitante que puede ser una nevera



Recogida de datos

Spark streaming

Conector Spark-C*

Hive Home: 200k usuarios, 15000 mensajes/s
Calderas conectadas: 25k usuarios,
2500 mensajes/s
Energía: 50 k usuarios
8500 mensajes/s

Ciencia de Datos

R y C#→Java/Scala
Nevera- 1000 líneas de código→ 400 en Spark

Ejemplo: ONG International Alert (UK)

- Áreas en conflicto
  - Evaluación, impacto, medidas de cambio
    - Actores a involucrar
    - Temas a tratar

Descubrimiento de información

Análisis de sentimientos

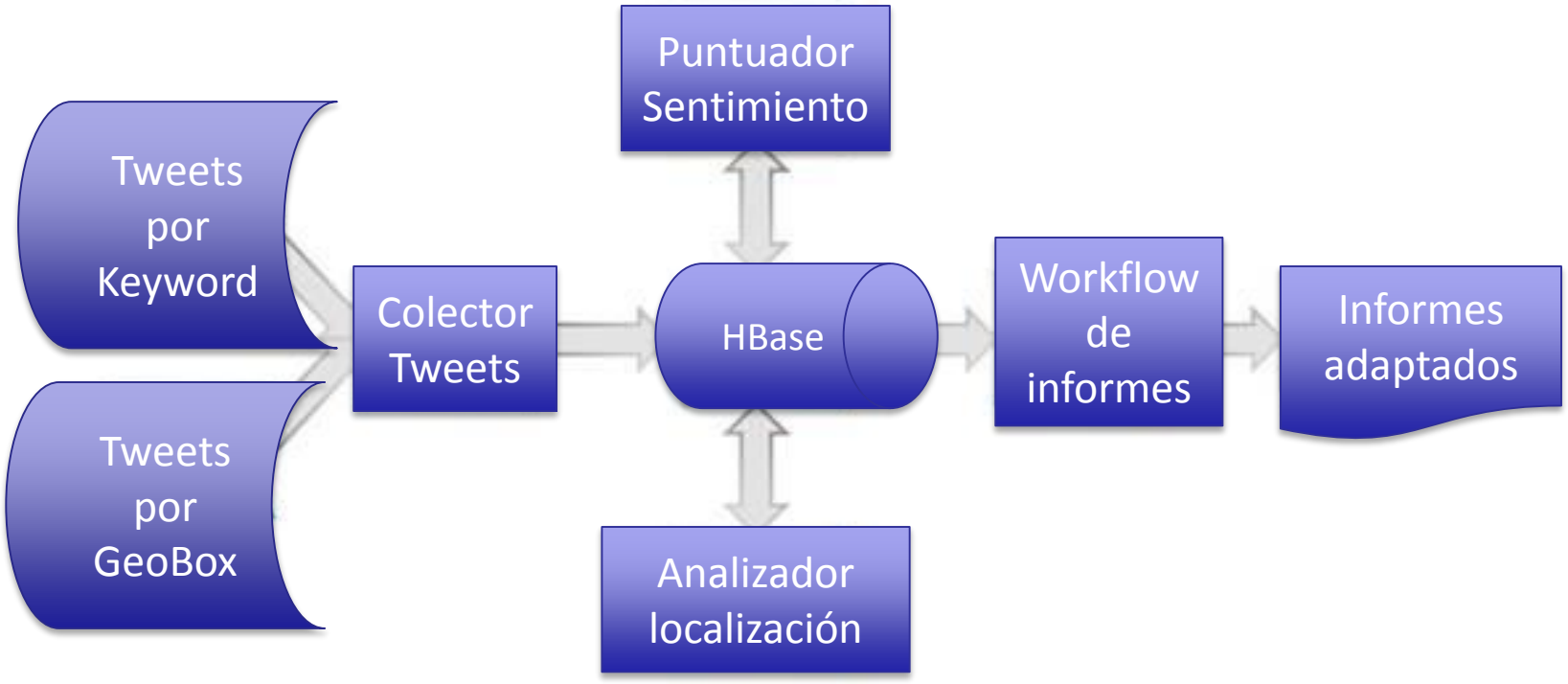
International Alert.

Understanding conflict. Building peace.

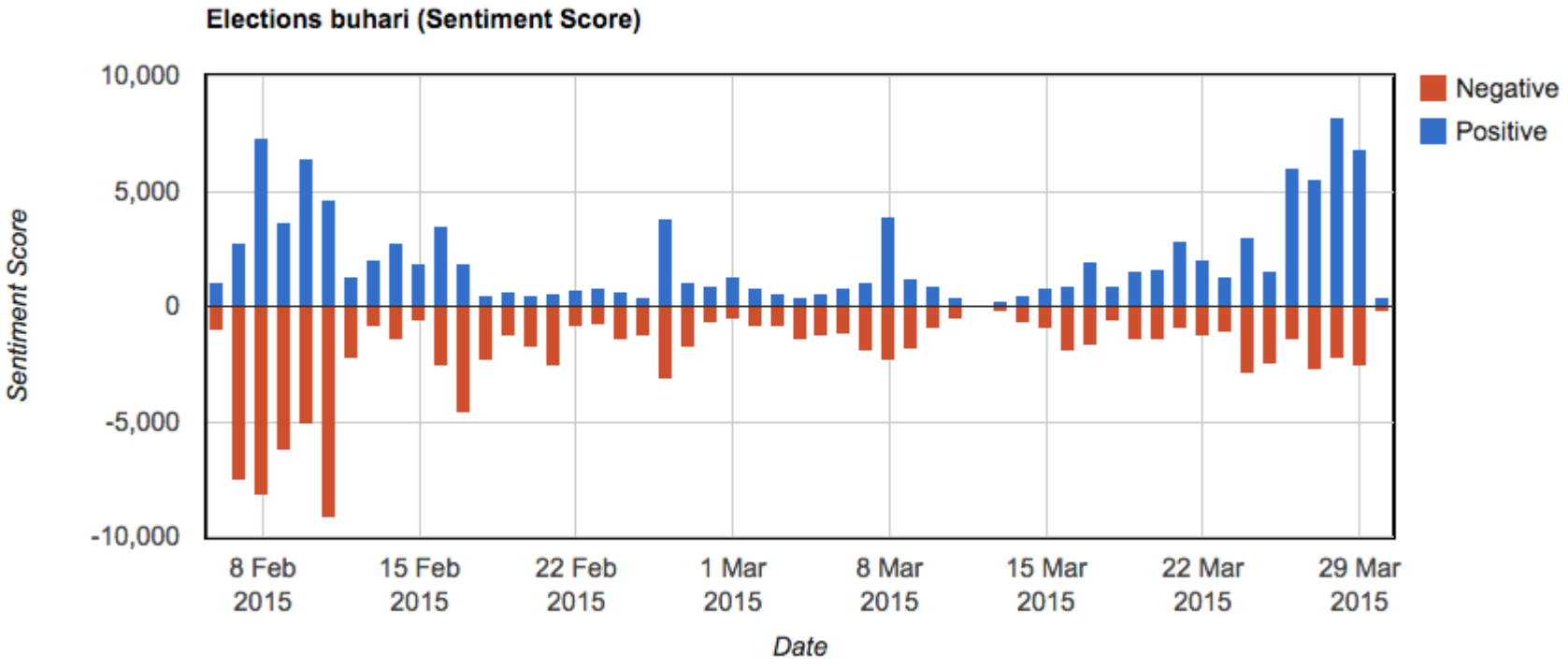
Descubrimiento de Información

Contenidos en internet

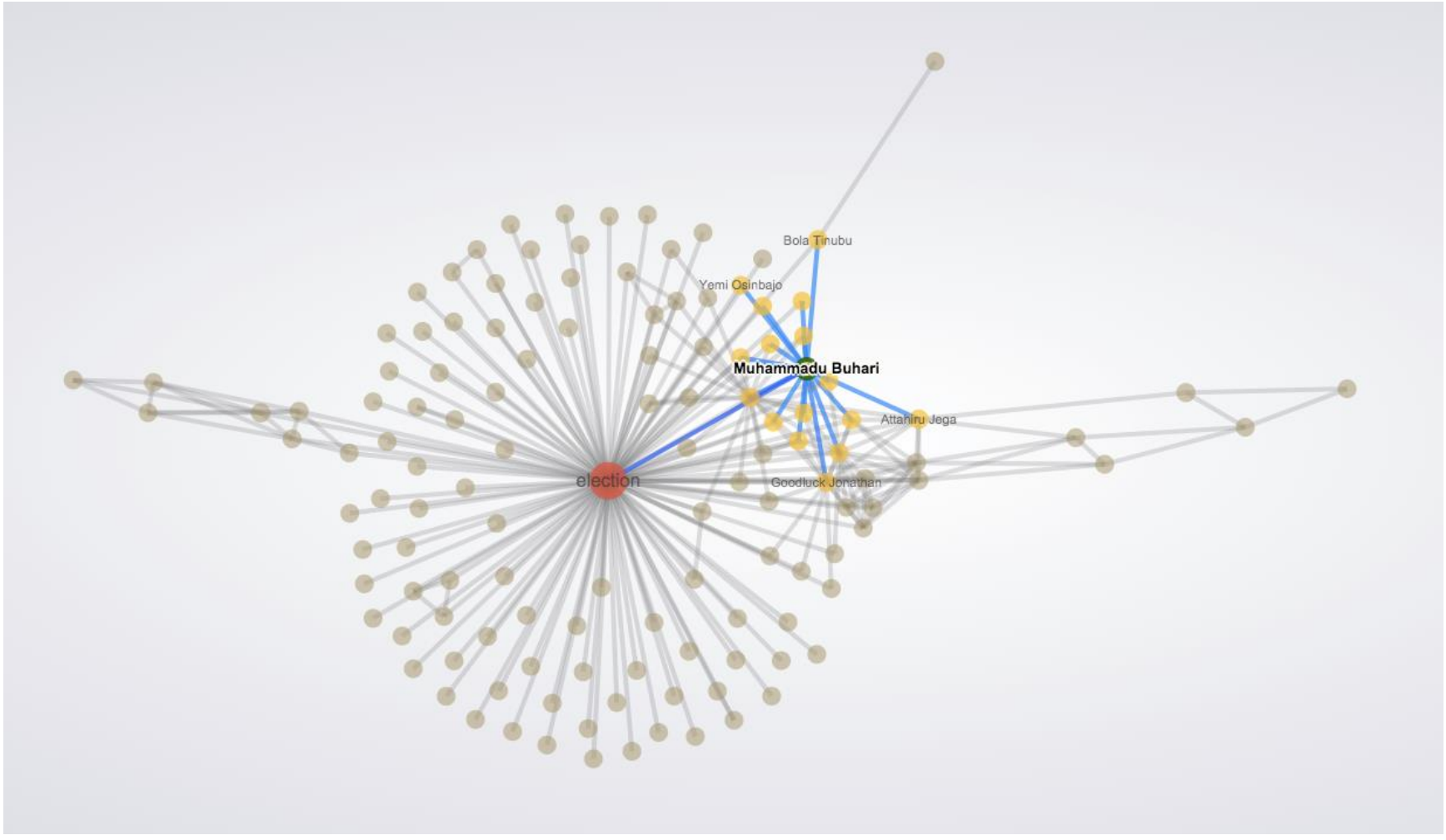
Análisis de sentimientos



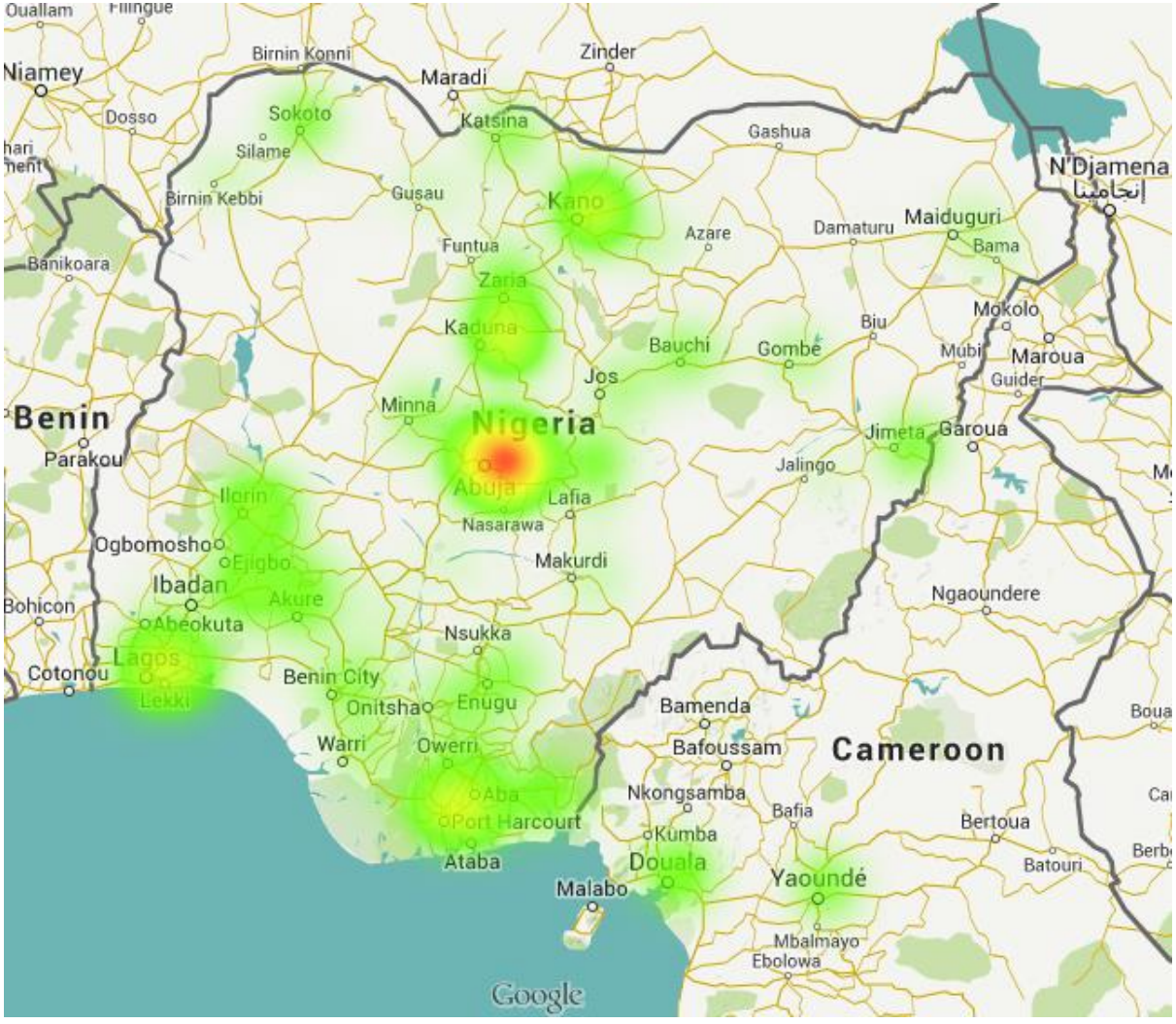
Análisis de sentimientos



Gráfica de actores



Heat map de Tweets



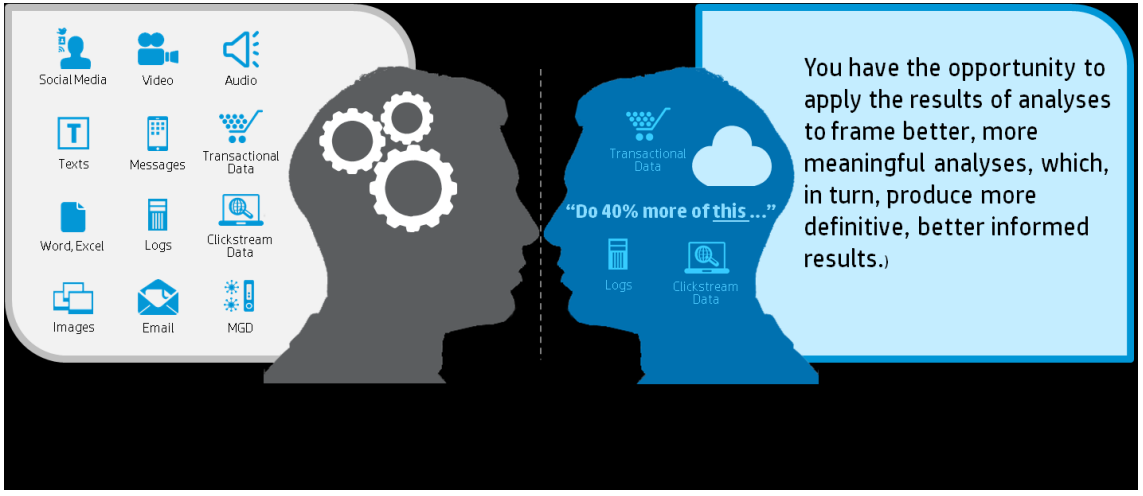
Motores y barreras

Motores


Fuentes de valor
Crear valor estratégico
Decisiones más rápidas
Decisiones mejores
Decisiones proactivas
Mejorar eficiencia
Mejora de capacidades
Aumento de automatización
Eliminar herramientas redundantes
Reestructuración procesos

Barreras: Acceso a datos

- Colocar en su sitio la tecnología y el talento apropiados
- Estructurar workflows e incentivos para optimizar el uso de Big Data. Datos transparentes y usables
- Facilitar el acceso a los datos: integrar información de múltiples fuentes



Barreras: Políticas de datos

- Privacidad
- Seguridad
- Propiedad intelectual
- Fiabilidad
- Responsabilidad legal

Sectores: Inversión, barreras, motores

Inversión
Motores
Barreras

Inversión
Motores
Barreras

Inversión
Motores
Barreras

Inversión Motores Barreras

Algunos modelos de negocio: diferente madurez y riesgo

Vino viejo en
Botella nueva

Mejorar servicios

Publicidad

Datos= Mejora negocio

Datos= Mejor publicidad

Usar los datos para entender y mejorar el negocio y los productos

Usar los datos para enfocar a usuarios con anuncios relevantes

amazon.com

ebay

NETFLIX

Google

Algunos modelos de negocio: diferente madurez y riesgo

Vino viejo en Botella nueva

Mejorar servicios

Datos= Mejora negocio

Usar los datos para entender y mejorar el negocio y los productos

amazon.com

ebay

NETFLIX

Publicidad

Datos= Mejor publicidad

Usar los datos para enfocar a usuarios con anuncios relevantes

Google

Big Data

Acceso a conocimiento

Datos= Negocio

El entendimiento ayuda a mejorar negocios y gobiernos

comScore

Experian
A world of insight

Indice de Rentabilidad

Guardas de datos personales

Datos= Riesgo=Negocio

Datos digitales
Privacidad
Oportunidad de negocio

reputation.com

PersonalCloud

Dinero, Dinero, Dinero….

21% MÁS DE GANANCIAS CORPORATIVAS GLOBALES

Conseguidas por empresas que aprovechan big data, según el Foro Económico Mundial.

45% DE ROI EN PROMEDIO GANADO POR SECTORES PRINCIPALES EN TODO EL MUNDO

De las telecomunicaciones a los seguros, pasando por la fabricación, según el Foro Económico Mundial.

83% DE MEJORA EN TOMA DE DECISIONES

Conseguida, en promedio y en todo el mundo, por empresas que utilizan big data, según TATA Consultancy.

Big data = Big Business?. Estudio EEUU

+300 billones $ en valor/ año
(2/3 en reducir gastos en un 8%)
0.7% anual de crecimiento en productividad

Más del doble del total anual en España

+ 60% aumento margen operaciones
0,5-1% anual de crecimiento en productividad

McKinsey&Company

Big Data = Big Business?
Estudio sectores (UE y global)

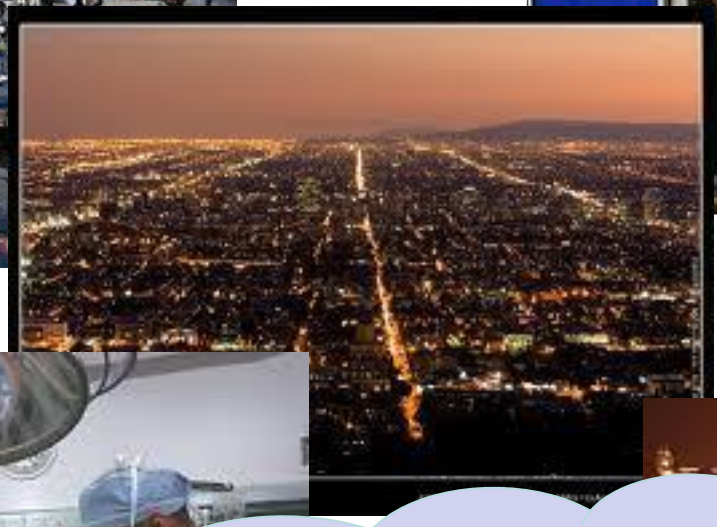
- 100 billones € en mejoras de eficiencia operacional
- Aparte Big Data para reducir fraude, errores y mejorar la recogida de impuestos
- 0,5% de crecimiento anual en productividad

Sólo temas de geo-localización podrían capturar unos $600 billones en superavit de consumidores

Hasta 50% reducción en desarrollo de producto y costes de ensamblado

McKinsey&Company

Big data: Beneficios sociales

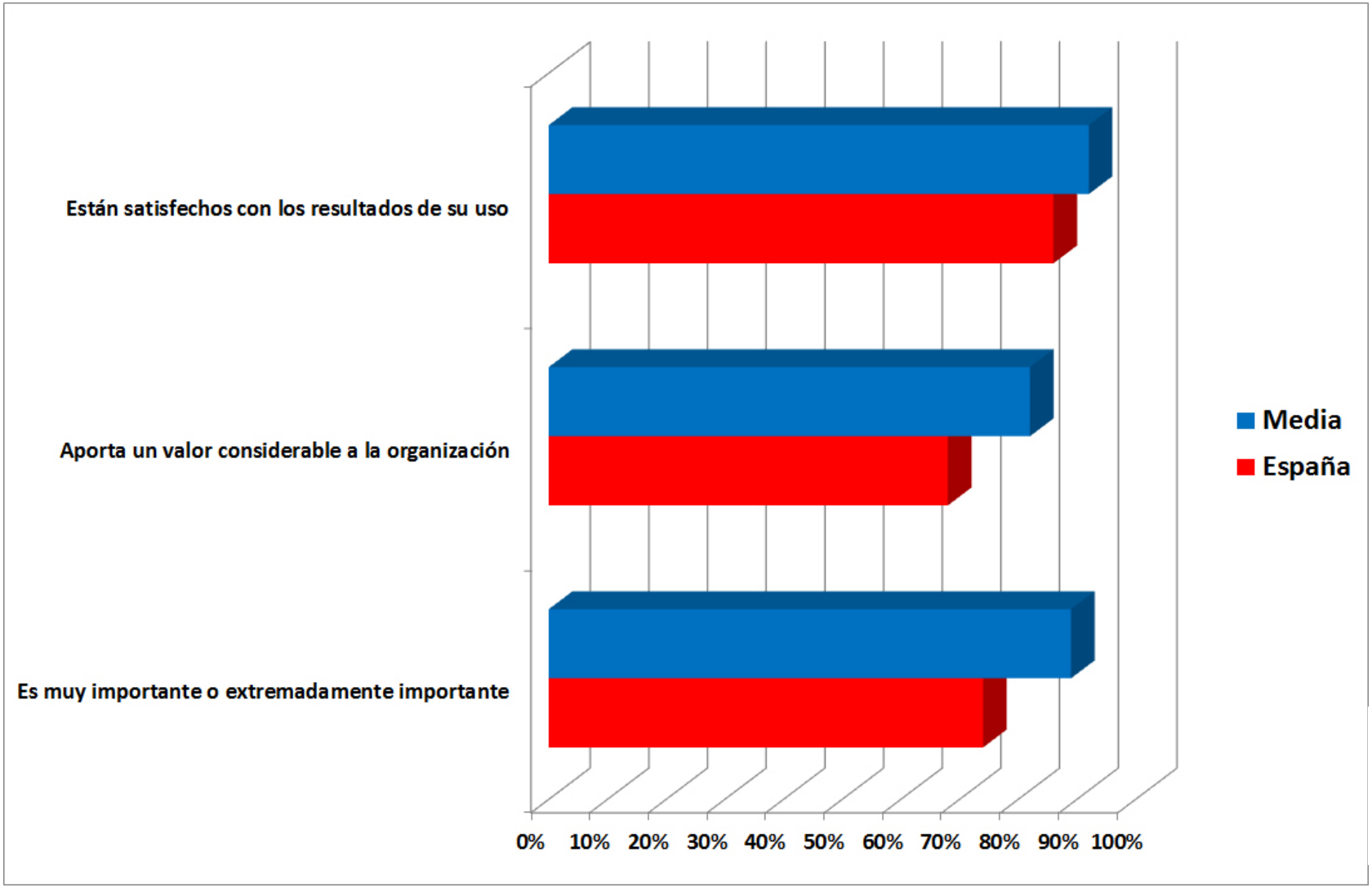


Gestión del tráfico y rutas
Transporte público
Iluminación inteligente
Medicina y Salud
Ciudades Inteligentes

Modelos de negocio emergentes

BBVA
INNOVATION CENTER
www.centrodeinnovacionbbva.com

Radiografía de Big data en Europa, ¿por qué?

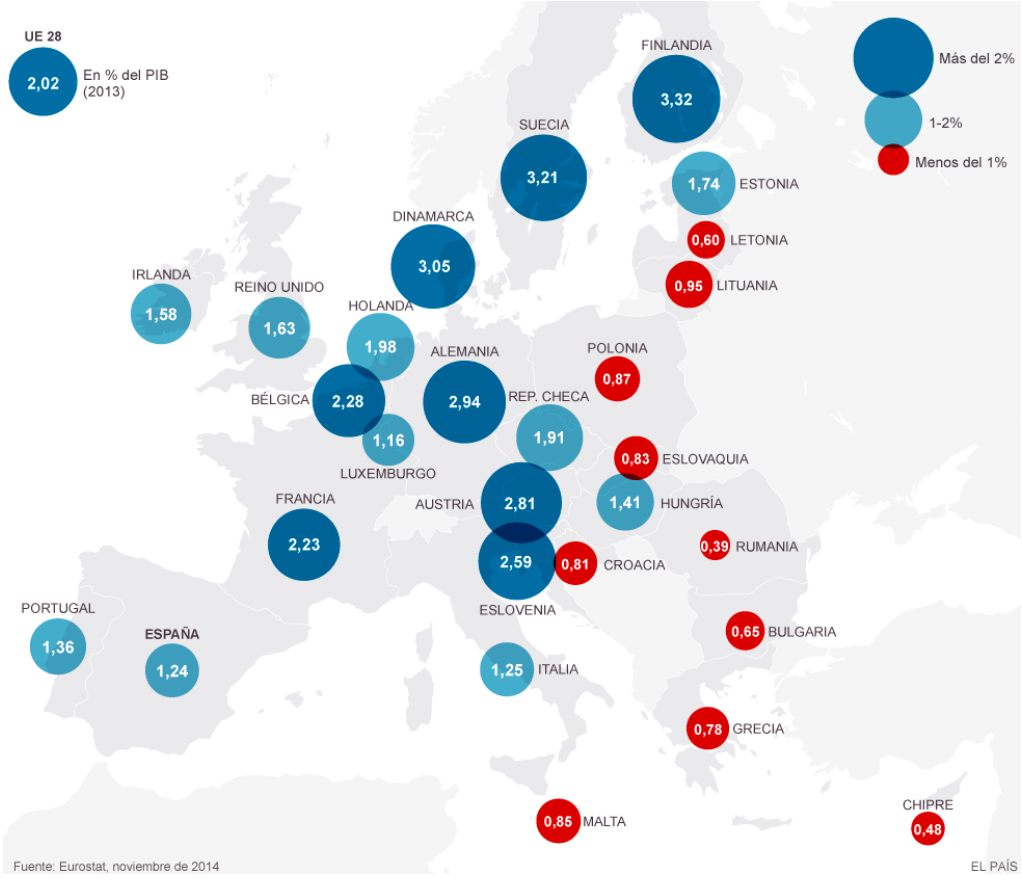


Big Data en el mundo

¿Por qué?

Consejo de Lisboa "Convertir UE en la economía basada en conocimiento más competitiva y dinámica del mundo"

- Lento avance hacia una inversión media del 3% en I+D
  - Crisis económica
  - Baja inversión empresas



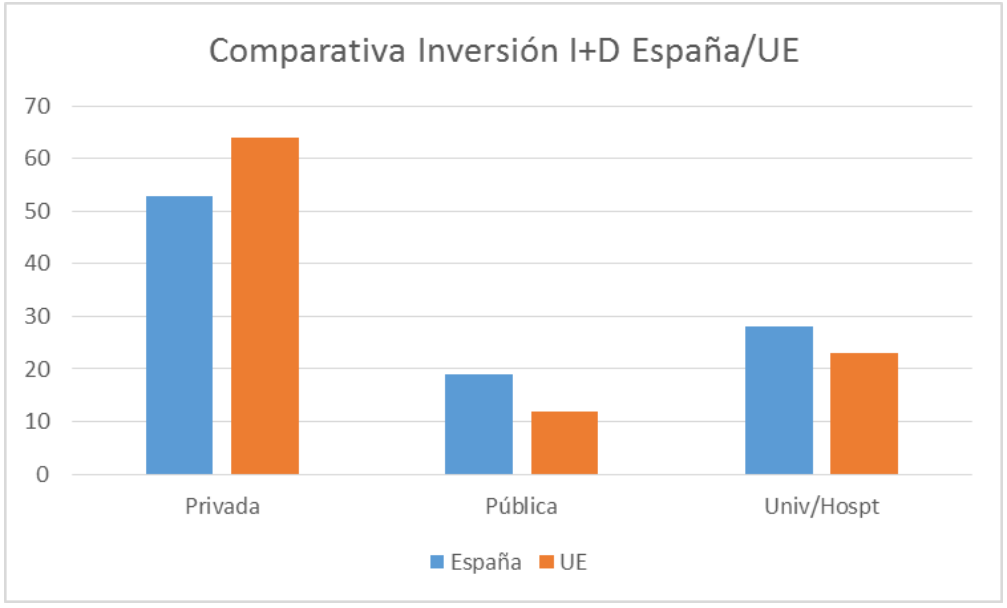
Algunas medidas para la UE

- EU es un jugador top en excelencia científica y producción de conocimiento
  - Pierde fuelle en resultados de investigación
  - Bajo nivel de transferencia
  - Mejorar participación público-privada

- Necesario aumento en cooperación internacional

- Las PYMES son innovadoras, pero no crecen suficiente

Comparativa inversión I+D



Privadas sin ánimo lucro

- 1% media UE

- Reino Unido 2%
- Italia 3%
- Portugal 9%
- España ----

Algunas noticias

- **UE:**
  - Los países que más invierten en I+D son los que mejor salen de la crisis (Alemania, Finlandia, Suecia)

- **España:**
  - CAV, Navarra, Madrid

- **Empresas:**
  - Volkswagen record de beneficios en 2011
  - Esfuerzo I+D+i en 2010 ≈ ∑ 20 empresas españolas más innovadoras (Telefónica, Acciona, Iberdrola, Indra, ..)

Programas de investigación

- Neelie Kroes, Vice-Presidenta EC responsable de la Agenda Digital

  "But this data revolution, moving to a data-driven economy, won't happen by itself. It needs a helping hand, and the right framework. Data needs to be freely available for use and re-use. It needs to be easy to transport and inter-operate — without different rules and standards for every country and dataset. And it needs the framework that safeguards privacy and builds trust"

"demand for data experts is booming: it has grown twelve-fold over twenty years. Jobs like "data scientist" – a concept that hardly existed a few years ago – are now in high demand.
Let's make sure we are responding. Let's ensure our education and training systems, and our industry, are preparing the ground for tomorrow's digital job market. That's exactly what we're doing with our grand coalition for digital jobs"

Survey de Big data en España

- BigDataHispano.org (1500 personas)

Big Data architect	15	20%
Data engineer	12	16%
Data scientist	11	14%
Business analyst	11	14%
Big Data analyst	9	12%
Database administrator	6	8%
Business intelligence expert	4	5%
Data miner	3	4%
Database manager	2	3%
Business intelligence systems manager	2	3%
Data warehouse manager	1	1%

25-40.000 €/año

crece 12%

crecen un 52%

Sexy… con talento, y …con ganas de ver mundo

82.000€

Sin exp. 71.000€
3-9 años 133.000€
Resp. equipo 205.000€

4.4M de puestos de trabajo

140–190.000 expertos en Big Data

1.5Millones de gestores y analistas con el know-how para usar Big Data Analytics en toma de decisiones efectiva.

UNIVERSIDADE DA CORUÑA

Workshop on
Big Data and Statistics

19 06 2015

LIDIA
Laboratorio de I+D en Inteligencia Artificial

MUCHAS GRACIAS

AMPARO ALONSO-BETANZOS