

Introducción a Big Data

Agenda

- **Introducción a Big Data**
- Datos estructurados y no estructurados
- Hadoop
- Ecosistema Hadoop
- Casos de Estudio

Definiciones de Big Data

- Un conjunto de tecnologías relacionadas y no relacionadas para analítica a gran escala
- Gran volumen, alta velocidad y gran variedad de información que demanda un procesamiento poco costoso para obtener conocimiento y tomar decisiones.

Las V's de Big Data

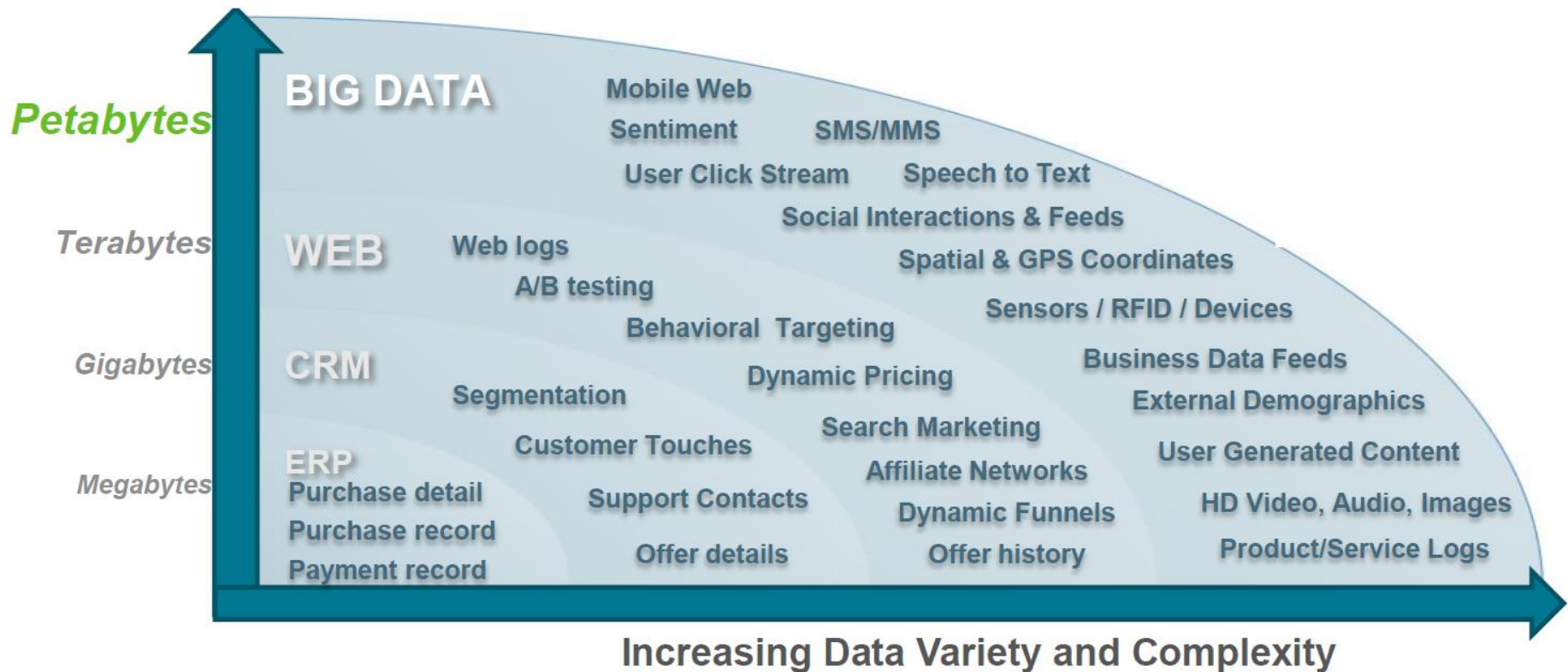
- **Volumen:** Terabytes, Petabytes, Exabytes
- **Velocidad:** hora, segundos, milisegundos
- **Variedad:** 5 formatos, 10 formatos, 20+ formatos
- **Variabilidad:** formatos cambian en el tiempo

No todas las V's tienen que estar presentes

ROI probado

- Google AdWords: Predicción de click through rates (CTR)
- Netflix: 75% del streaming de video viene de recomendaciones
- Amazon: 35% de las ventas de producto vienen de recomendaciones de producto

Nada nuevo, excepto las V's

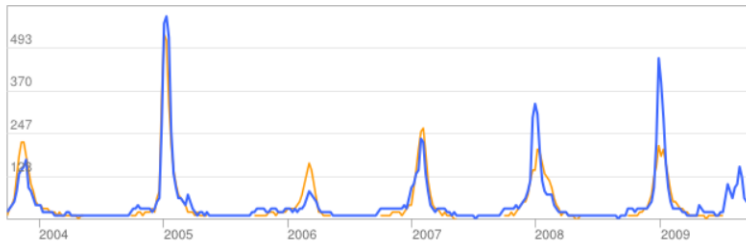


Correlación de Búsquedas

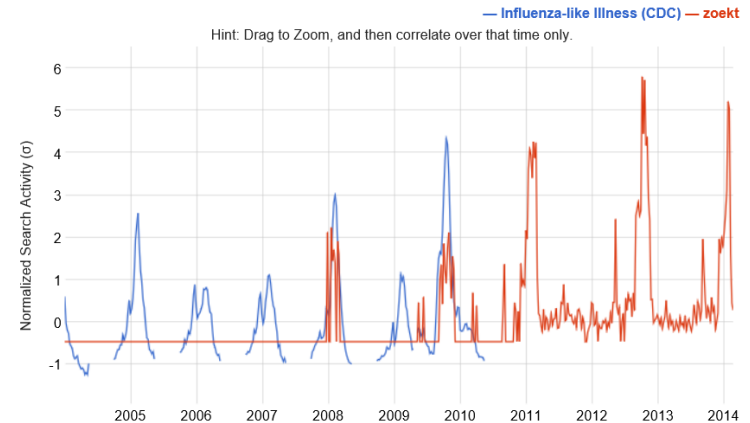
Actividad de la gripe en España

Estimaciones de la gripe

● Estimaciones de Evolución de la gripe
● Datos de España

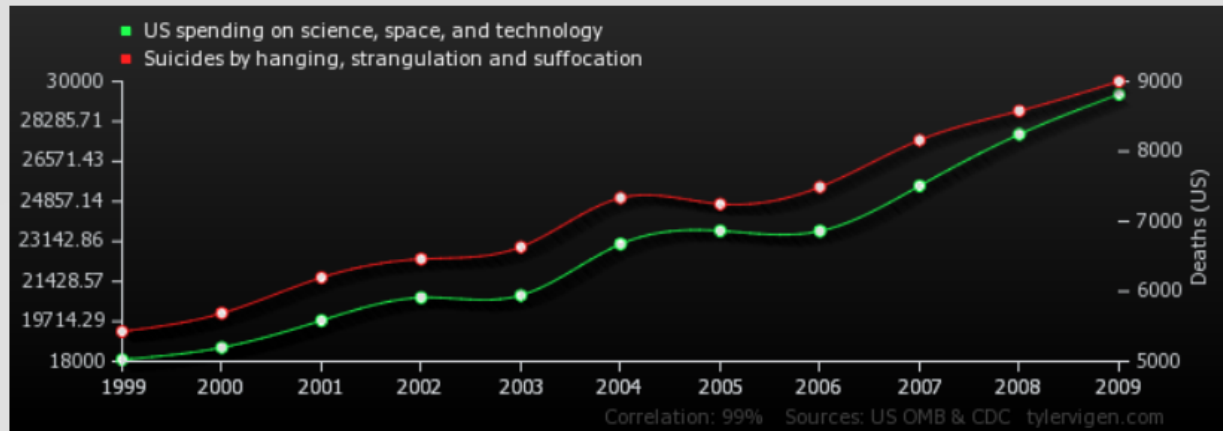


España: datos sobre enfermedades de tipo gripal publicados por la [Red Europea para la Vigilancia de la Gripe](#) del Centro



¿Sólo correlaciones?

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology Millions of todays dollars (US OMB)	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation Deaths (US) (CDC)	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

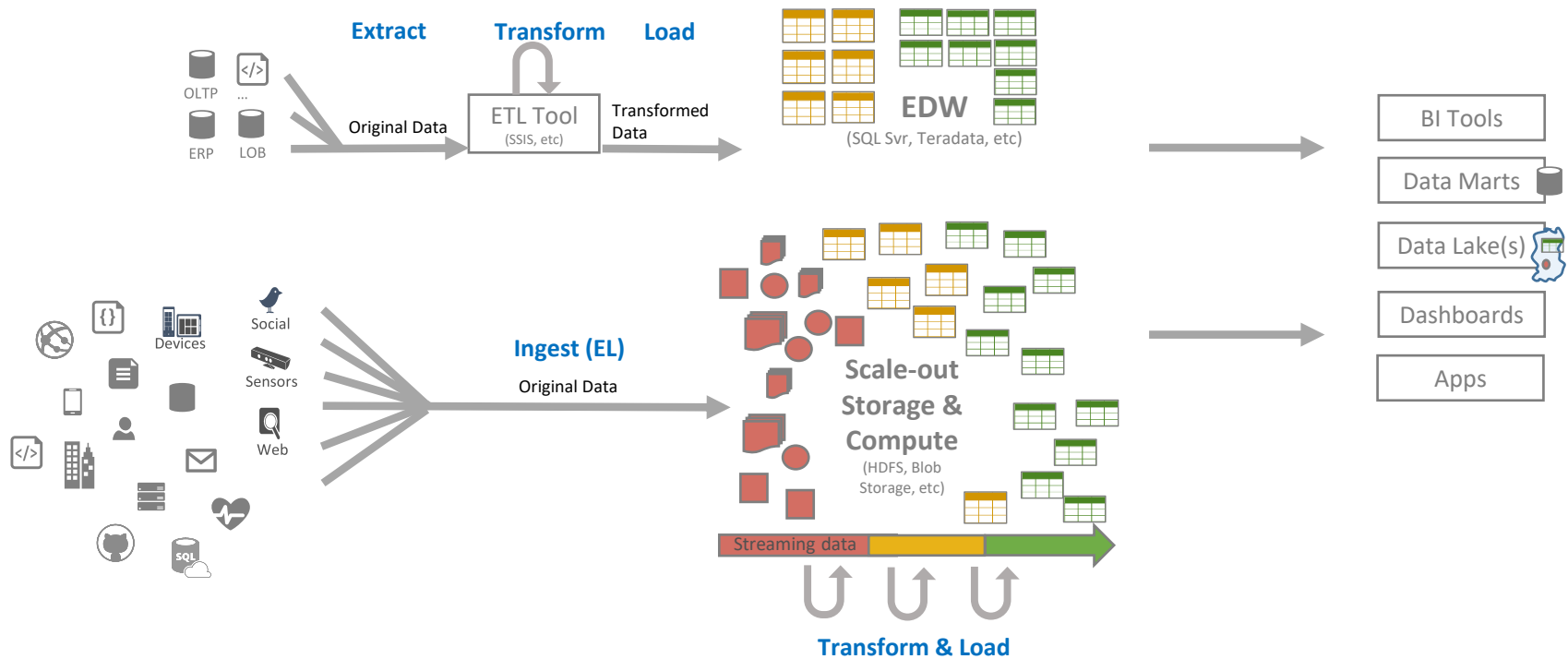
Correlation: 0.992082

<http://www.tylervigen.com/>

Big Data \neq BI Tradicional con más datos

- Big Data está redefiniendo los procesos de gestión de datos maestros, calidad de datos y gestión del ciclo de vida de la información
- Big Data NO reemplaza EDW y OLAP, suplementa esas inversiones
- El ecosistema Big Data incluye una gran variedad de tecnologías analíticas
- Bases de datos columnares, JSON y almacenes de ficheros no estructurados

Enfoque evolutivo



Cambios en patrones DW

El almacenamiento Big Data (hoy en día Data Lake) se caracteriza por tres atributos clave:

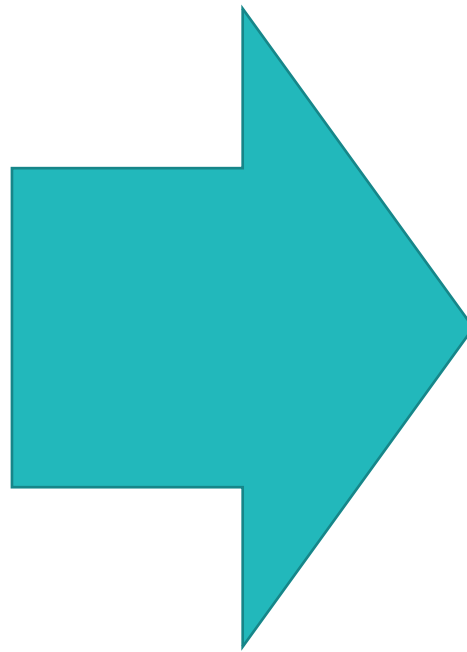
- Recoge todo
- Explótalo desde cualquier sitio
- Acceso Flexible

Cambios en patrones DW

Cambiar de Esquema primero a Esquema más tarde

1. Llegan los datos
2. Se deriva el esquema
3. Limpieza de Datos
4. Transformación
5. Carga en EDW
6. Análisis

Valor de
los datos
LENTO

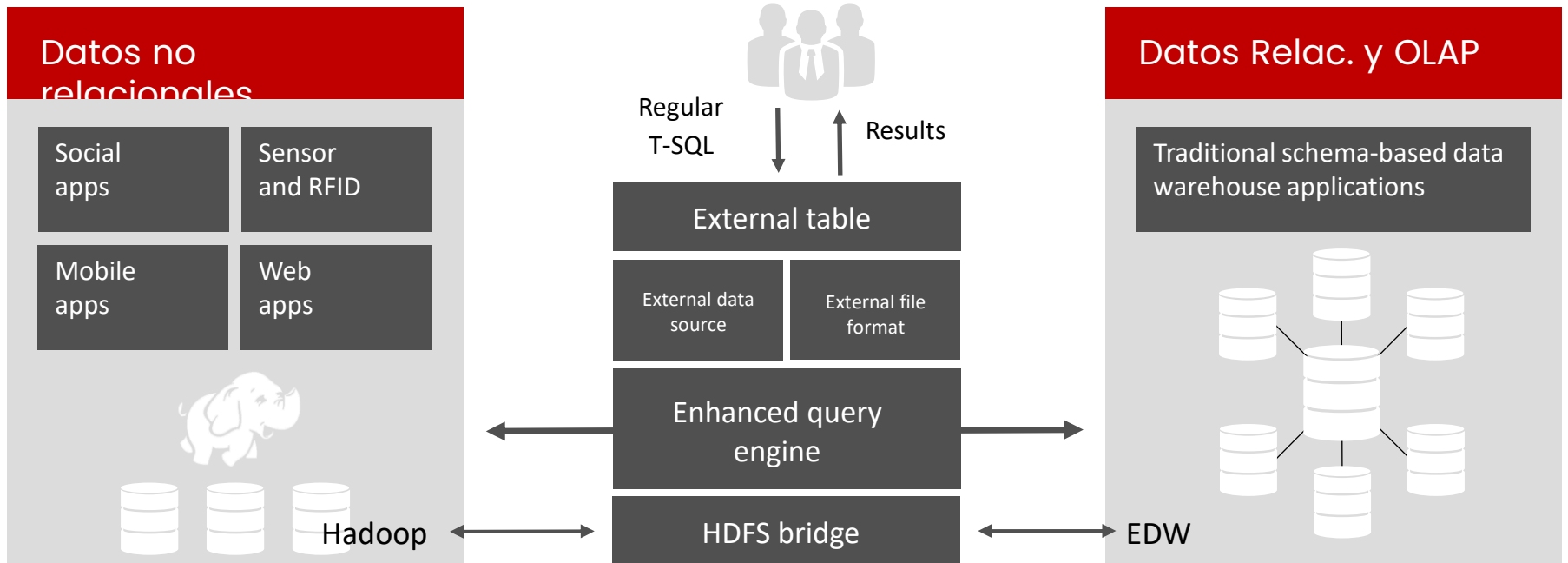


1. Llegan los datos
2. Se cargan en Hadoop
3. Análisis
4. Subconjuntos cargados en EDW

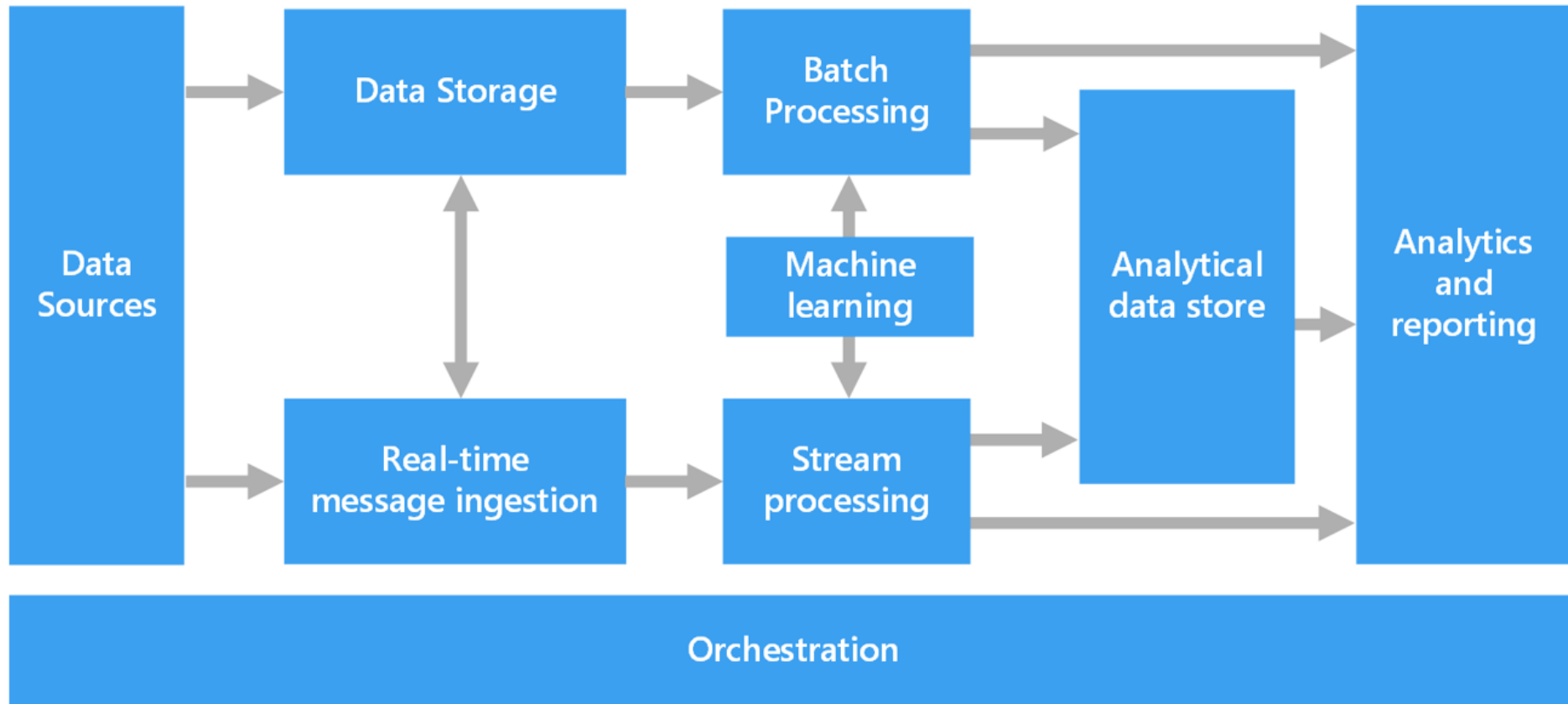
Rápido valor
de los datos

Cambios en Patrones

Básicamente construir un “puente” hacia Big Data



Componentes de una Arquitectura Big Data



Tipos de Carga de Trabajo

- Procesado batch de grandes orígenes de datos
- Procesado de grandes cantidades de datos en tiempo real
- Exploración interactiva de grandes cantidades de datos
- Analítica predictiva y Machine Learning
- Considerar big data cuando:
 - Se almacenan y procesan grandes cantidades de datos demasiado grandes para una base de datos tradicional
 - Transformar datos no estructurados para análisis y Reporting
 - Capturar, procesar y analizar grandes cantidades de datos en streaming en tiempo real o con muy baja latencia

Agenda

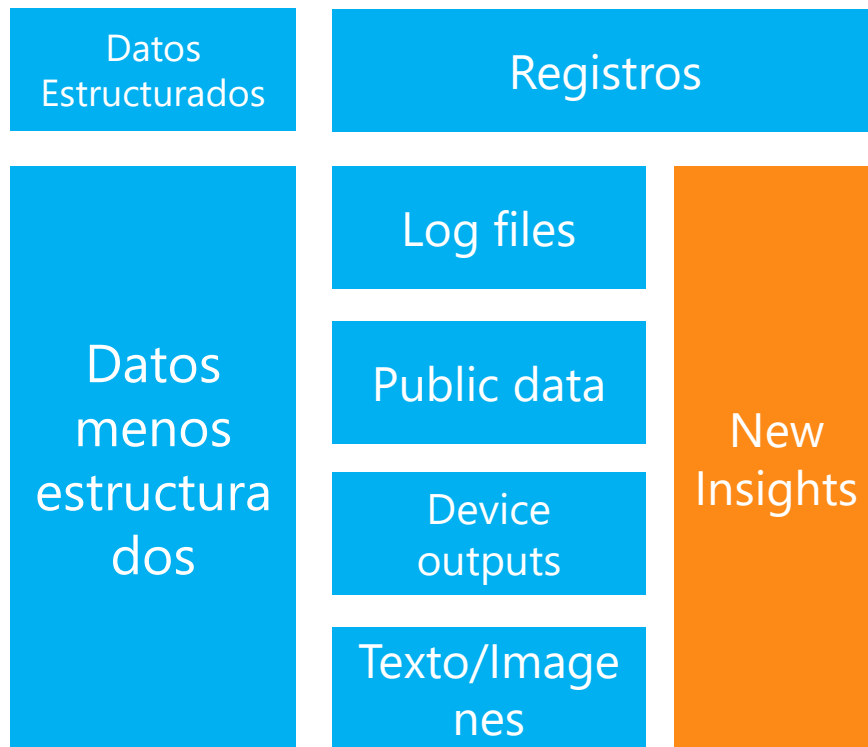
- Introducción a Big Data
- **Datos estructurados y no estructurados**
- Hadoop
- Ecosistema Hadoop
- Casos de Estudio

Datos no estructurados no se están analizando



- Datos estructurados
 - BBDD Relacionales
 - BBDD Analíticas

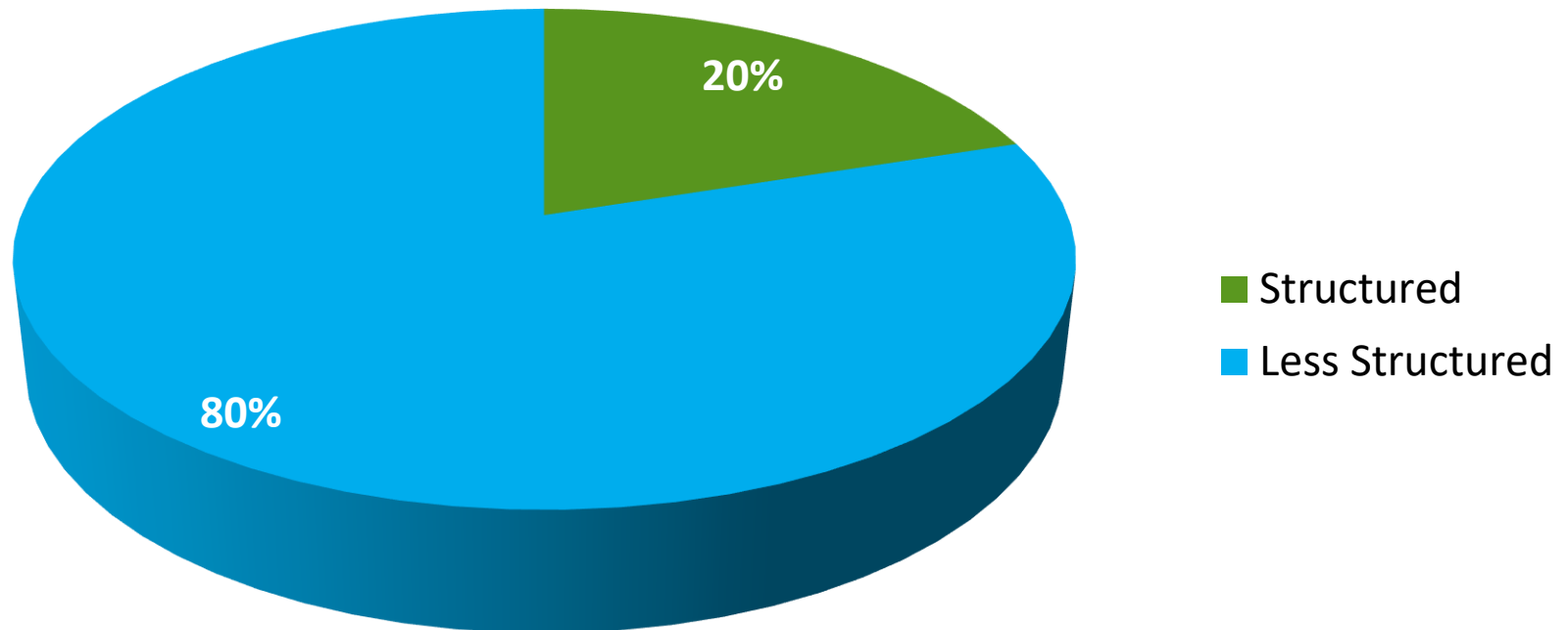
Datos no estructurados no se están analizando



- Datos estructurados
 - BBDD Relacionales
 - BBDD Analíticas
- Datos Menos estructurados
 - Crear ETL para transformar en Relacional
 - Mucho tiempo desarrollo
 - Susceptible cambio estructura
 - Archivado o borrado
 - Acceso caro

Datos en las organizaciones

Tipos de datos



Ejemplos de datos no estructurados

facebook

12 Tb
day

21 Pb
Hadoop
cluster

bing

7 Pb
Month
(search
queries info)

twitter

1 Tb
tweets/day

7 Tb
data/day

K KLOUT

75
Million
scores/day

4 Billion
Graph
edg/day

FINANCIAL TIMES

USA TODAY

THE WALL STREET JOURNAL.

B B C

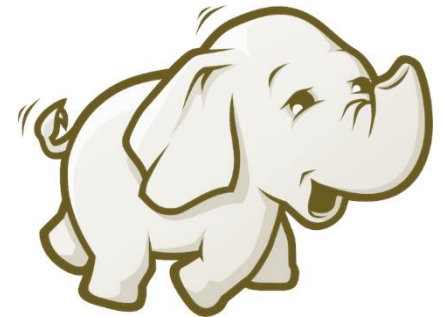
Millions of
opinions

Agenda

- Introducción a Big Data
- Datos estructurados y no estructurados
- **Hadoop**
- Ecosistema Hadoop
- Casos de Estudio

Hadoop

- Open Source ☺
- Plataforma de almacenamiento y procesado para **Big Data**
- Optimizado para manejar
 - Datos de forma masiva utilizando paralelismo
 - Variedad de datos (Estructurado, No estructurado, Menos estructurado)
 - Uso de hardware barato
- No para OLTP / OLAP
- Mover el cómputo hacia el dato



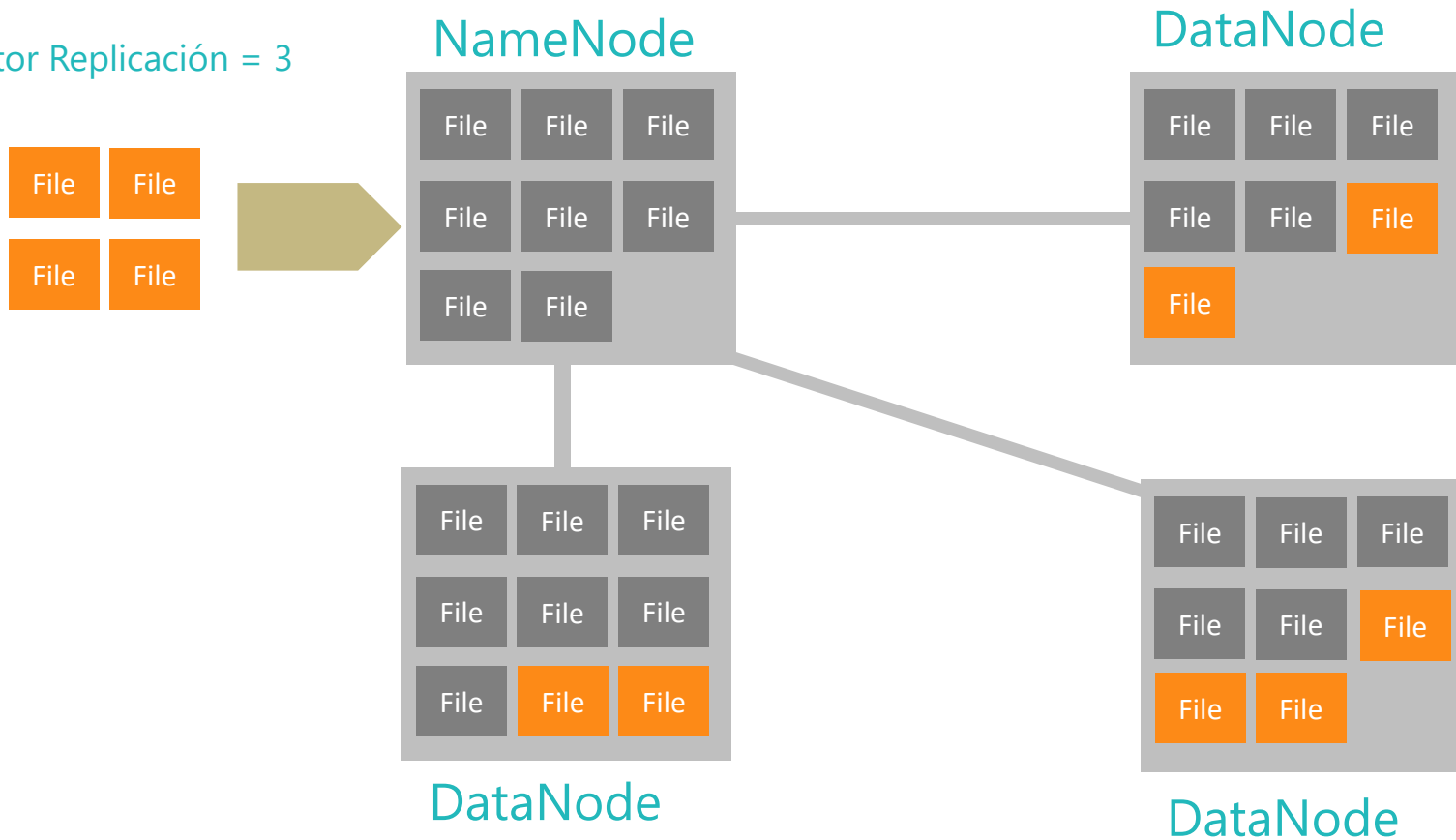
Hadoop

- Componentes core de Hadoop: HDFS & MapReduce
- Hadoop Distribution File System
 - Distribuido, tolerante a fallos, redundante, autorecuperable
- Map Reduce
 - Procesamiento distribuido, tolerante a fallos, procesa donde está el dato. Lectura y procesado distribuido.

Un cliente escribiendo datos en HDFS

Tamaño de Bloque = 64 Mb

Factor Replicación = 3



Hadoop

- Escalable
 - Escala linealmente en capacidad de almacenamiento y procesamiento.
- Tolerante a Fallos
 - Matrimonio entre un Sistema de ficheros distribuido y un framework tolerante a fallos utilizado para leer datos
- Procesamiento distribuido
 - Sigue la estrategia de divide y vencerás.

RDBMS vs Hadoop

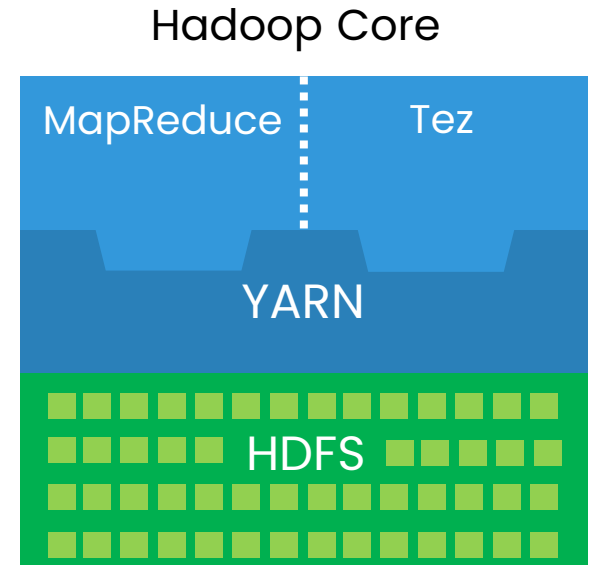
Característica	RDBMS	Hadoop
Tamaño de Datos	Gigabytes (Terabytes)	Petabytes (Hexabytes)
Acceso	Interactivo y Batch	Batch – NO Interactivo
Actualizaciones	Leer/ Escribir varias veces	Escribir una vez, leer varias veces
Estructura	Esquema estático	Esquema dinámico
Integridad	Alta (ACID)	Baja
Escalado	No lineal	Lineal
Tiempo de respuesta consultas	Puede ser casi inmediato	Tiene latencia (debido a procesamiento batch)

Historia Hadoop

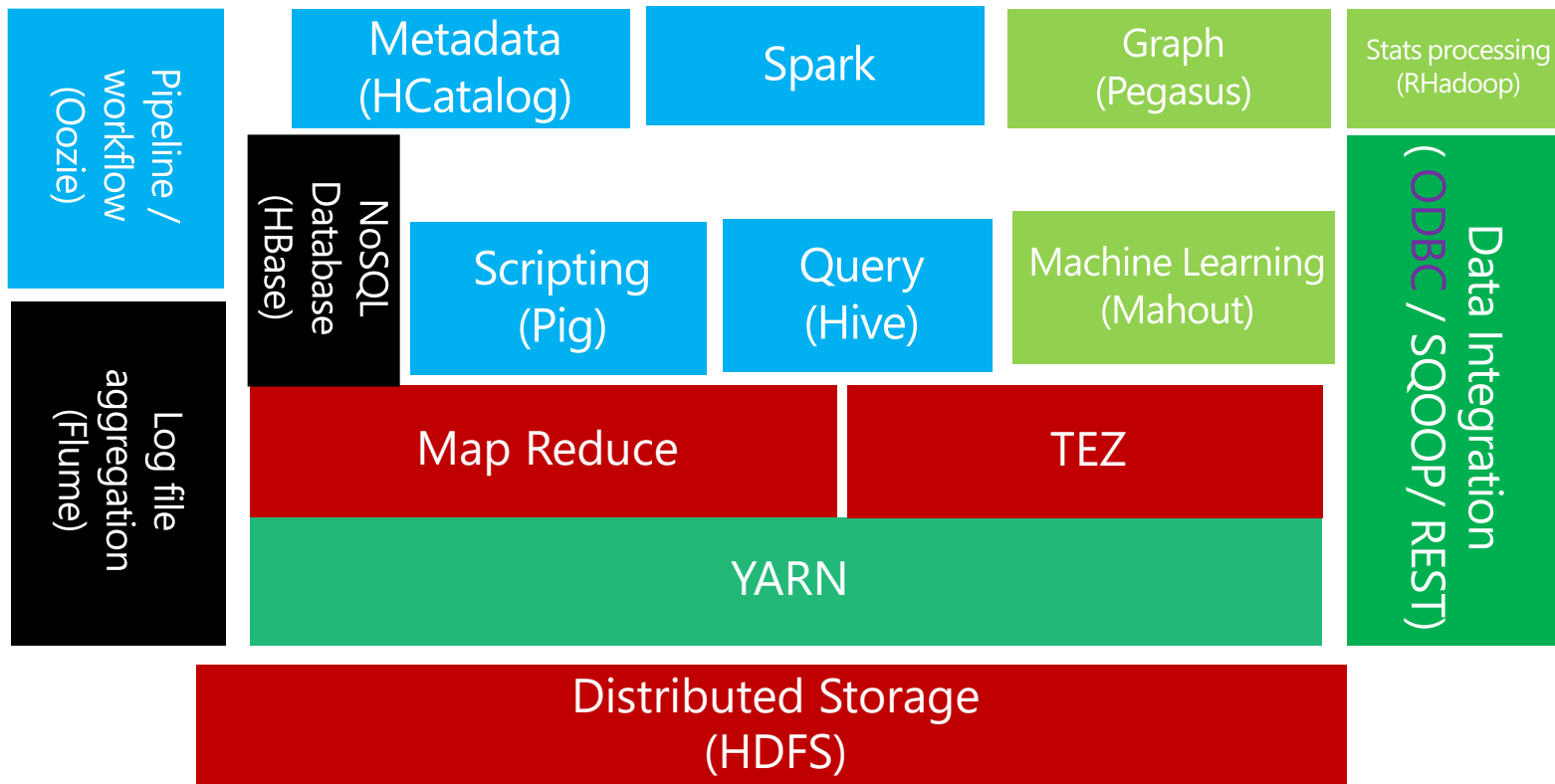
- 2002: **Nutch** open source motor de búsqueda por **Doug Cutting**
- 2003: Google publica un documento sobre **GFS** (Google Distribute File System)
- 2004: **Nutch** Distributed Files System (**NDFS**)
- 2004: Google publica un documento sobre **MapReduce**
- 2005: **MapReduce** se implementa en **NDFS**
- 2006: Doug Cutting se une a Yahoo! & inician Apache Hadoop Subproject
- 2008: Hadoop se convierte en un Proyecto top de Apache
 - El índice de Yahoo's se ejecuta en un cluster de 10.000 nodos
 - Hadoop rompe el record de 1TB en ordenación: 209s en 910 nodos
 - New York Times convierte 4TB de archivos en PDF en 24h en 100 nodos
- 2011: Yahoo crea **HortonWorks**, una compañía dedicada a Hadoop
- 2011: **HortonWorks** y **Microsoft** anuncian un acuerdo
- 2011: Microsoft libera la primera preview de **Isotop/HDInsight**
- 2018: **Cloudera** compra **Hortonworks**

El Zoo de Big Data

- El objetivo de Hadoop es crear un framework unificado para procesar big data
- Tres requisitos principales:
 - Escalabilidad
 - Eficiencia



Ecosistema Hadoop



Hive

- Sistema Data Warehouse para Hadoop
- Facilita las sumalizaciones de datos
- Consultas Ad-hoc
- Lenguaje consulta similar SQL: **HiveQL**
- Análisis de grandes conjuntos de datos almacenados en Hadoop
- Por detrás ejecuta
 - Trabajos **MapReduce**
 - **Stinger** / **Tez**
 - **LLAP** (Long Live and Process)

Pig

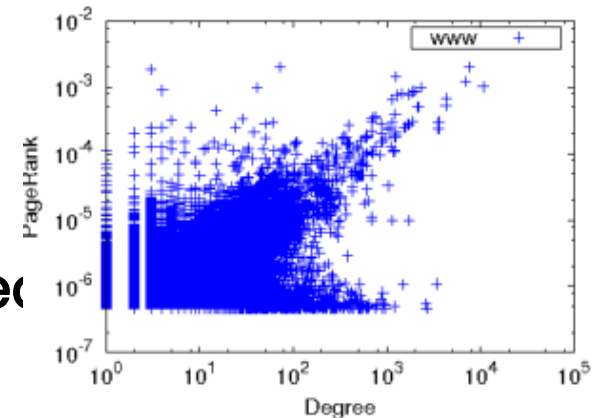
- Lenguaje scripting de Alto nivel
- Capa de procesamiento de Alto Nivel que se ejecuta en Hadoop
 - Usa ambos HDFS y Map Reduce
- Facilidad de programación
 - El Usuario se enfoca en semántica en lugar de eficiencia. Map Reduce es como lenguaje de ensamblador
- Extensibilidad

Flume & Sqoop

- Flume
 - Recolectar y mover grandes cantidades de datos
 - Ejecución distribuida
- Sqoop
 - Import y Export: RDBMS \leftrightarrow HDFS, Hive..
 - SQL Server, MySQL, Oracle
 - Ejecución distribuida

Mahout & Pegasus

- Mahout
 - Machine learning y data mining a gran escala
 - clusterización, recomendaciones, clasificación, y más.
- Pegasus
 - Page Rank y Graph Mining
 - Network Analysis.
- Por detrás se ejecutan Trabajos MapReduce



Agenda

- Introducción a Big Data
- Datos estructurados y no estructurados
- Hadoop
- Ecosistema Hadoop
- **Casos de Estudio**

Almacena ahora, averigua más tarde

- Hadoop facilita almacenamiento y procesado
- Fácil de desarrollar programas que obtienen conocimiento de datos no estructurados
- Framework para almacenar y procesar subconjunto de los datos a demanda

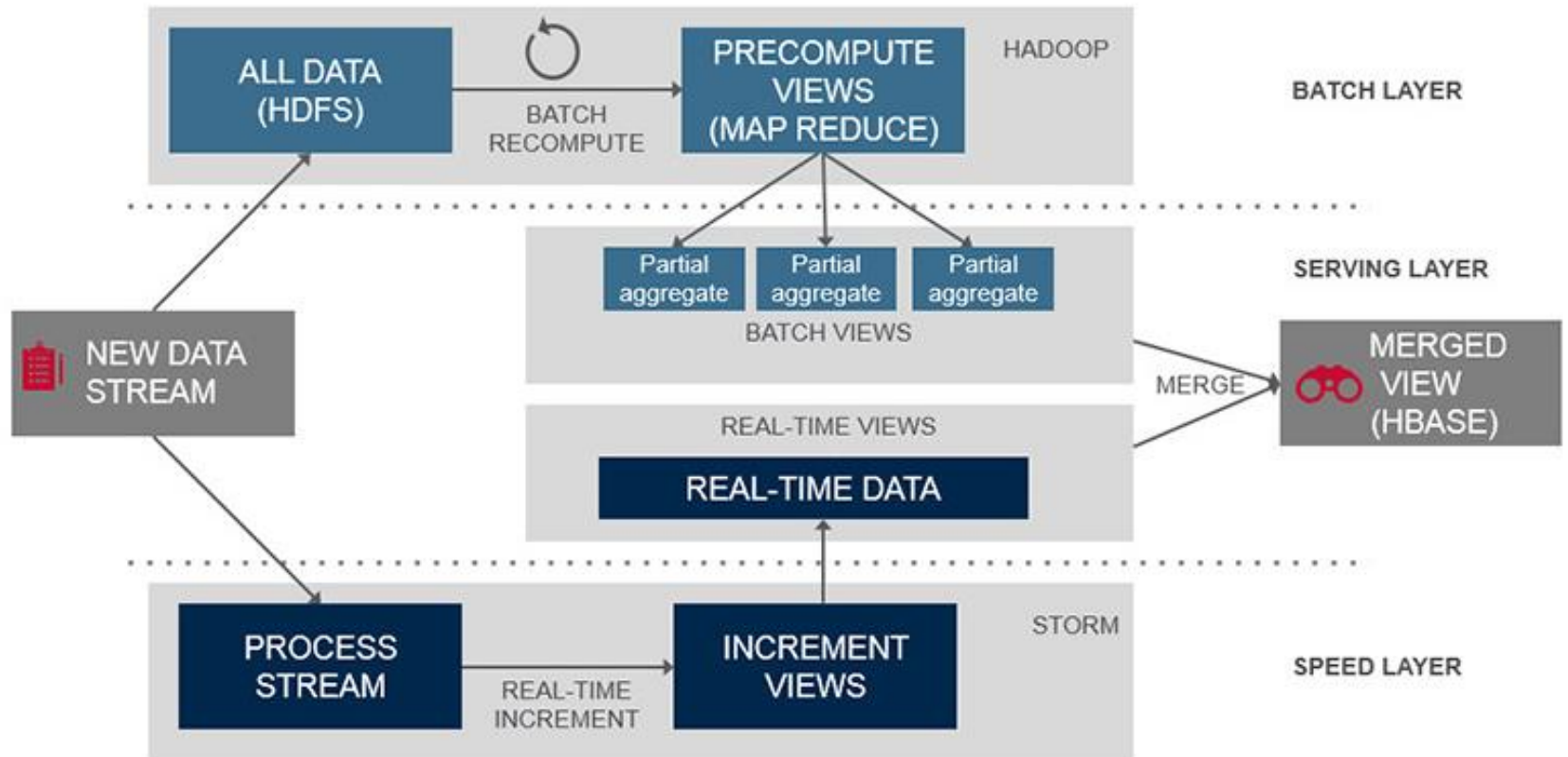
Datos en Tiempo Real

- Utilizar almacenes de datos operacionales en tiempo real (RT ODSS)
- Utilizar DW en Tiempo real
- Implementar CDC
- Presentar datos en tiempo real y datos históricos
- Definir umbrales aceptables y reglas de negocio para todas las entidades del tiempo real

Datos en Tiempo Real

- Flujo de datos continuo
- Manejar el stream como si fuese una cola
- Ventanas de tiempo
- Arquitectura de datos Hadoop y Lambda
- Enriquecer datos de streaming con datos de la organización
- Almacenar los datos de stream para construir la historia

Arquitectura Lambda



Casos de estudio Hadoop (I)

- Modelado de Riesgos
 - Banca y seguros,
- Análisis de rotación
 - Consultar logs y datos complejos desde múltiples orígenes
- Motor de recomendaciones

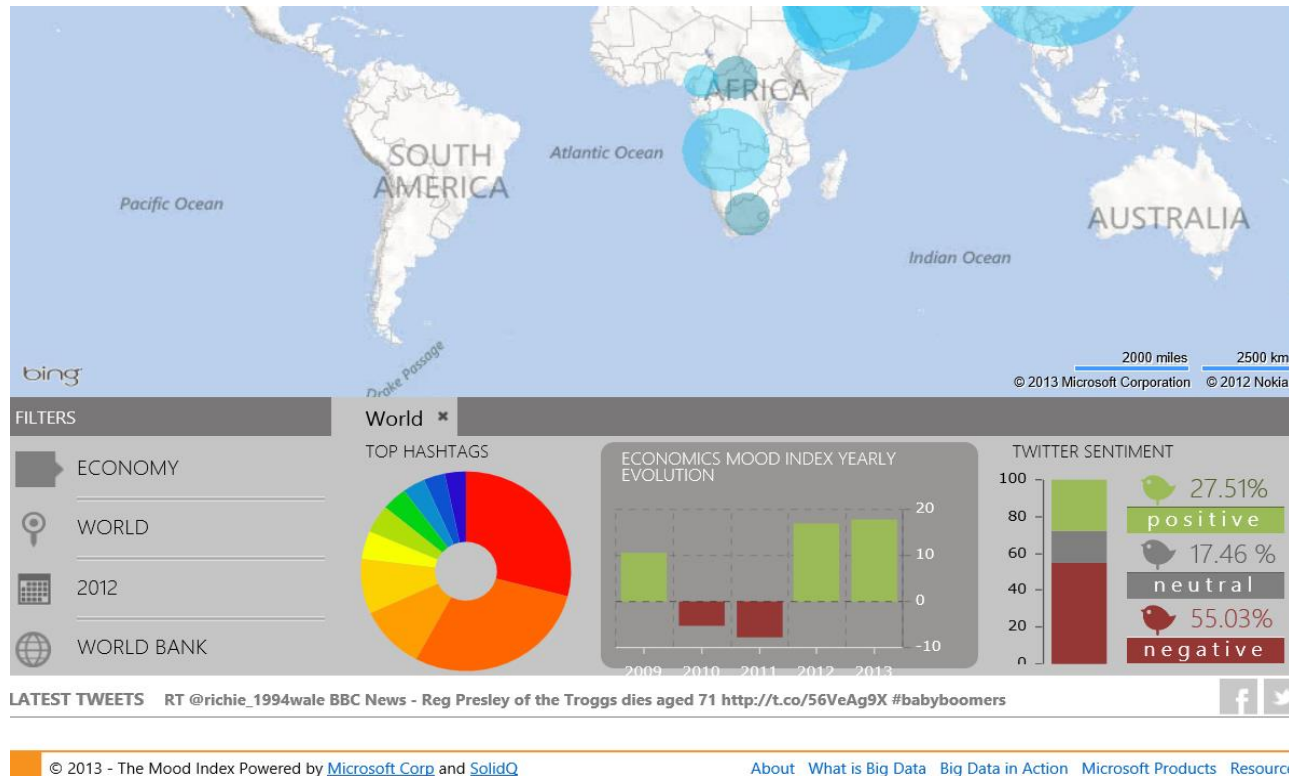
Casos de Estudio Hadoop (II)

- Ad Targeting
 - CTR, placement, auction
- Análisis de transacción en punto de venta
 - Análisis cesta de la compra, mejora de márgenes
- Datos de Redes
 - Predicción de fallos, ratios de transmission, protocolos de transmisión

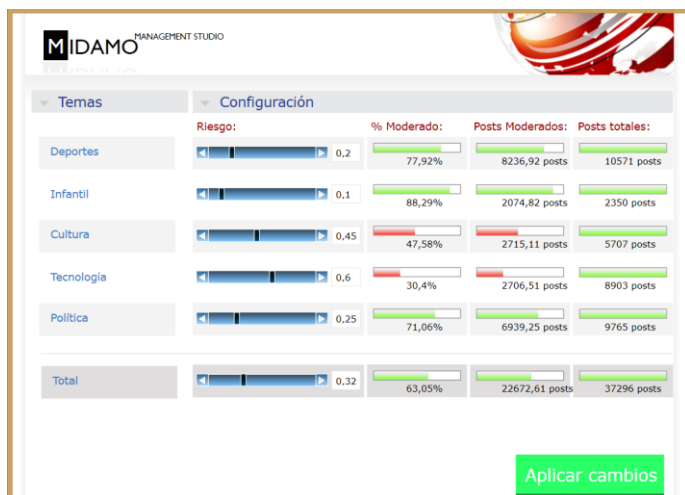
Casos de Estudio Hadoop (III)

- Detección de fraude
 - En transacciones
- Calidad búsquedas
 - Resultados relevantes, utilidad de búsquedas
- Data sandbox
 - Almacenado ahora y analizado después
- Análisis de Sentimiento
 - Twitter

Caso de Éxito: Mood Index



MIDAMO



SU NOMBRE FUE TAN COREADO COMO EL DE KAKÁ

Florentino sí que es un galáctico

Las más de 40.000 personas que se dieron cita en el coliseo de la Castellana tenían tantas ganas de ver a Kaká de blanco como de corear el nombre de la persona que ya trajo a Figo, Zidane, Ronaldo y Beckham.

Jose Antonio 12:30 29/7

A uno le cuesta imaginar cómo será la presentación de Cristiano Ronaldo, actual Balón de Oro y el traspaso más caro en la historia del fútbol, el próximo 6 de julio, pero desde ya les digo que el portugués tendrá muy difícil superar el espectáculo vivido la tarde de este martes en el Santiago Bernabéu.

Porque la presentación de hoy ha sido doble. La afición merengue ha tenido al fin la posibilidad de vitorear a su nuevo héroe y, de paso, al hombre que lo ha hecho posible, Florentino Pérez. La primera toma de contacto del presidente madridista con su afición en esta segunda etapa demostró hasta qué punto era deseado su regreso.



Comentarios:

Más comentarios: 85 ...

#85 **javi86**

Atención, última hora: el virus de Madriditis, según acaba de comentar la ministra de sanidad, alcanza el grado de Pandemia en Cataluña. Asimismo ha declarado que los principales afectados son tanto varones como mujeres. simpatizantes del Barsa. HALA

#84 **madridl**

I belong to Jesús!! I belong to Kaká!! simplemente de pie, callado, sin decir nada, solo sonriendo..... es elegante. Solo podías ir a un sitio y has elegido el mejor. HALA MADRID!!!

Bienvenido jose233, añade comentari

Me parece genial que un equipo como el Madrid fiche a un jugador de estas características
Como Madridista que soy estoy orgulloso de ello, nos va a hacer muy grandes!!
HALA MADRID!!

Publicar

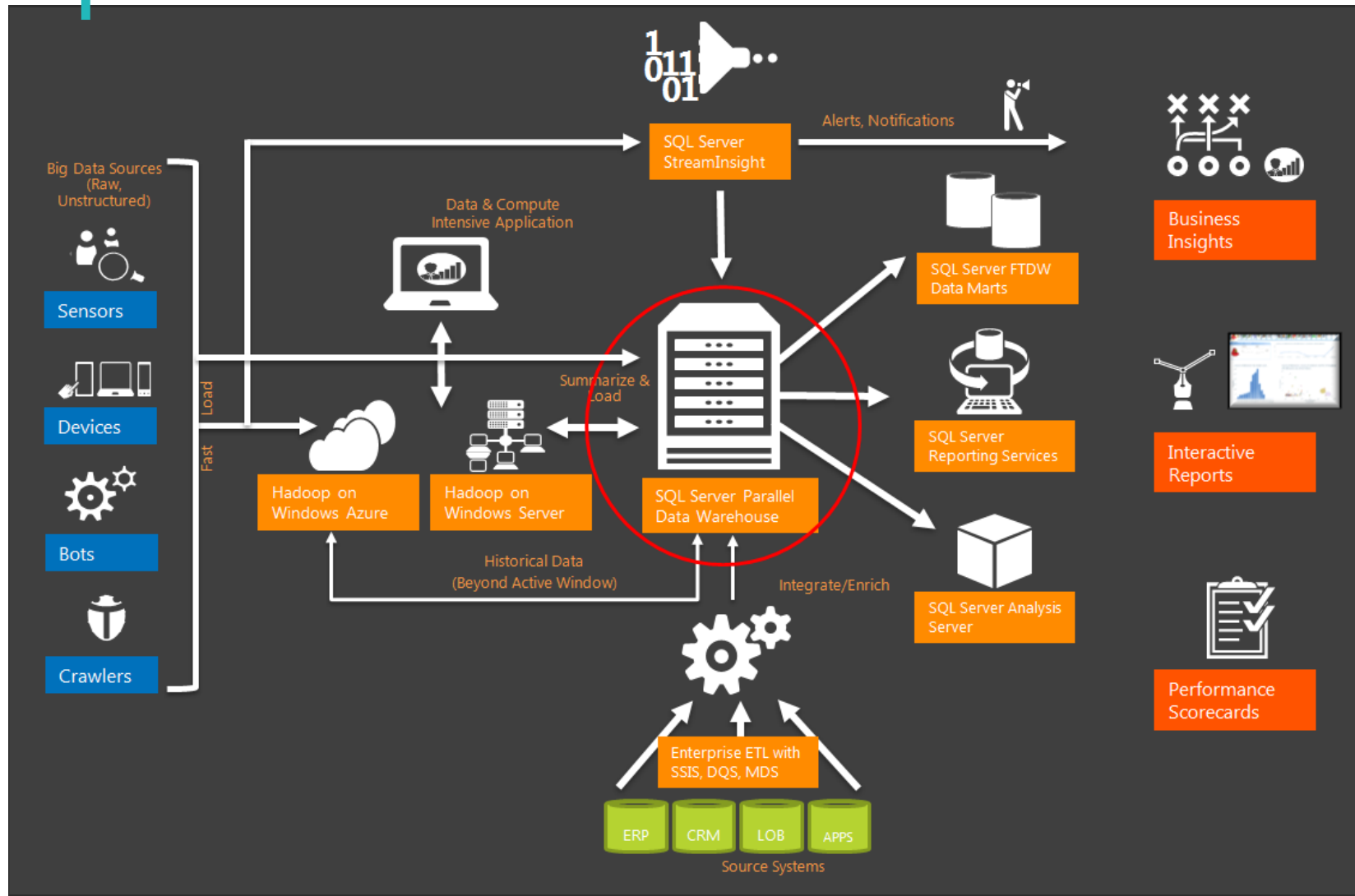
Normas:

Por favor, escribe correctamente, sin mayúsculas ni abreviaturas.

Recuerda que el tono del mensaje debe ser respetuoso. No se admitirán insultos ni faltas de respeto.

La redacción se reserva el derecho a eliminar

Opciones



Datalakehouse

