Introducción a Big Data

Agenda

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

Dataging
Managing through data

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 procesado 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

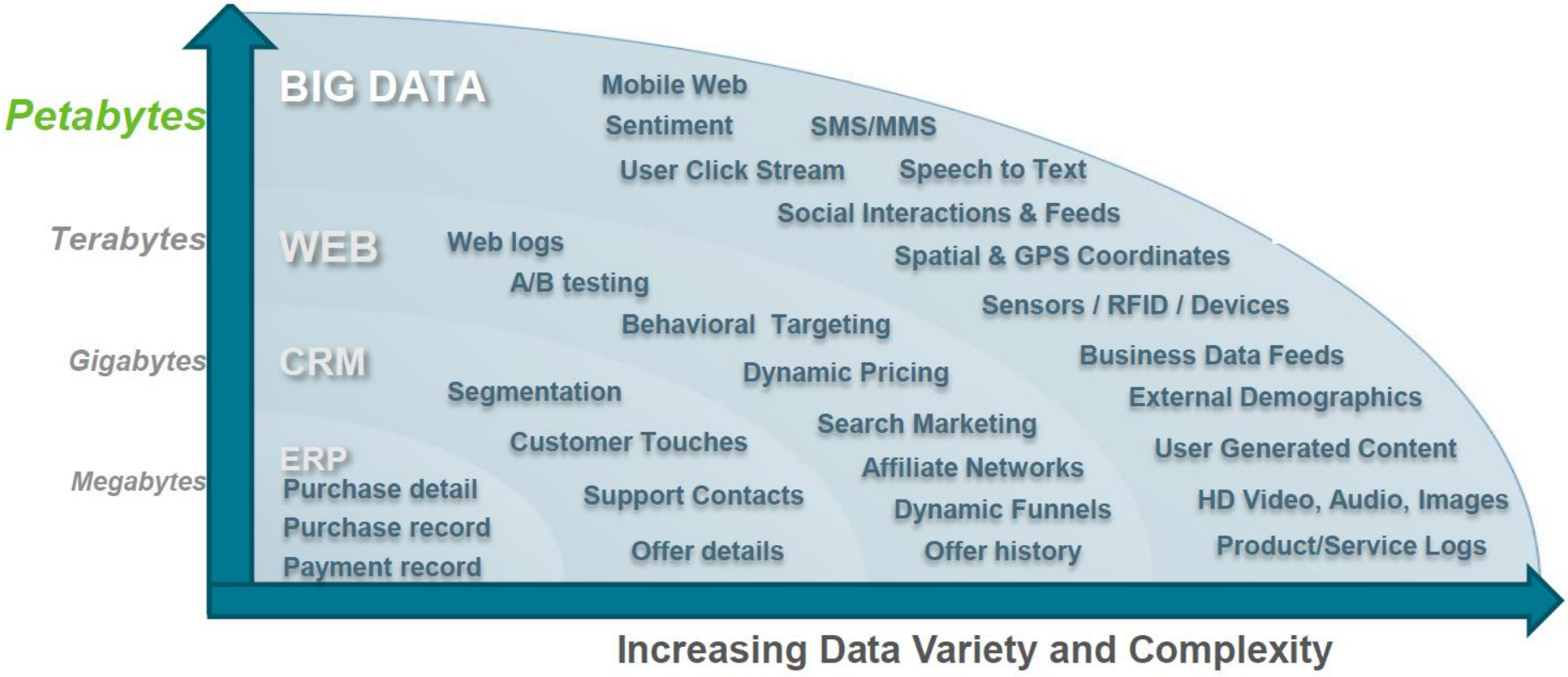
No todas las V's tienen que estar presentes

Dataging
Managing through data

ROI probado

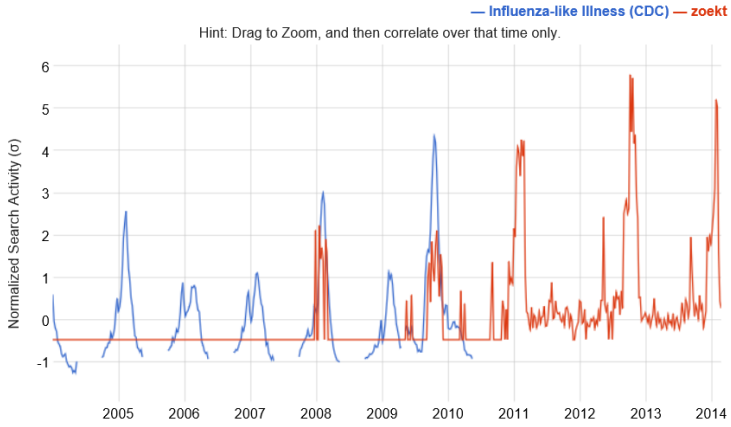
- Google AdWords: Predicción de click through rates (CTR)
- Netflix: 75% del streaming de video viene de recomendaciones
- Amazon: 35% de las ventas de producto vienen de recomendaciones de producto

Nada nuevo, excepto las V's



Dataging
Managing through data

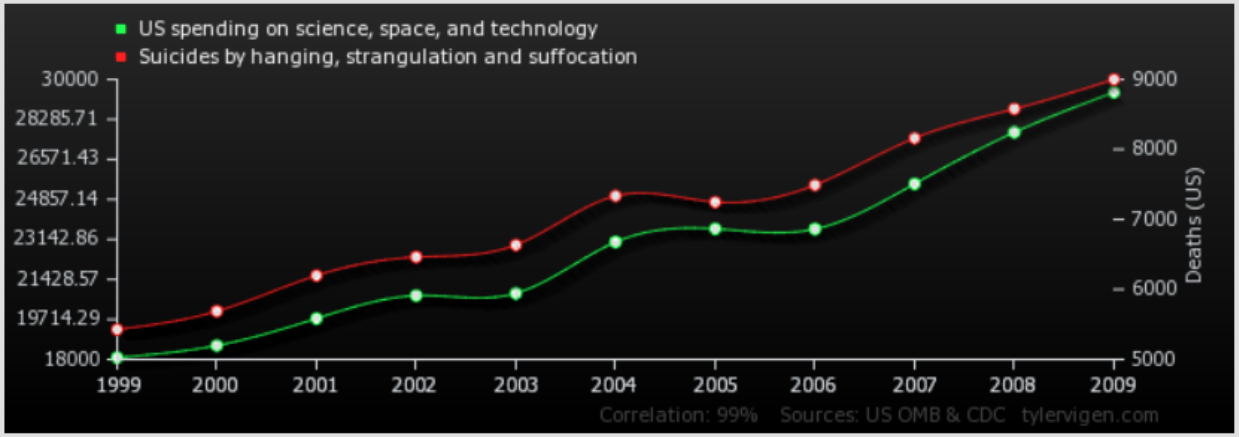
Correlación de Búsquedas



Dataging
Managing through data

¿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

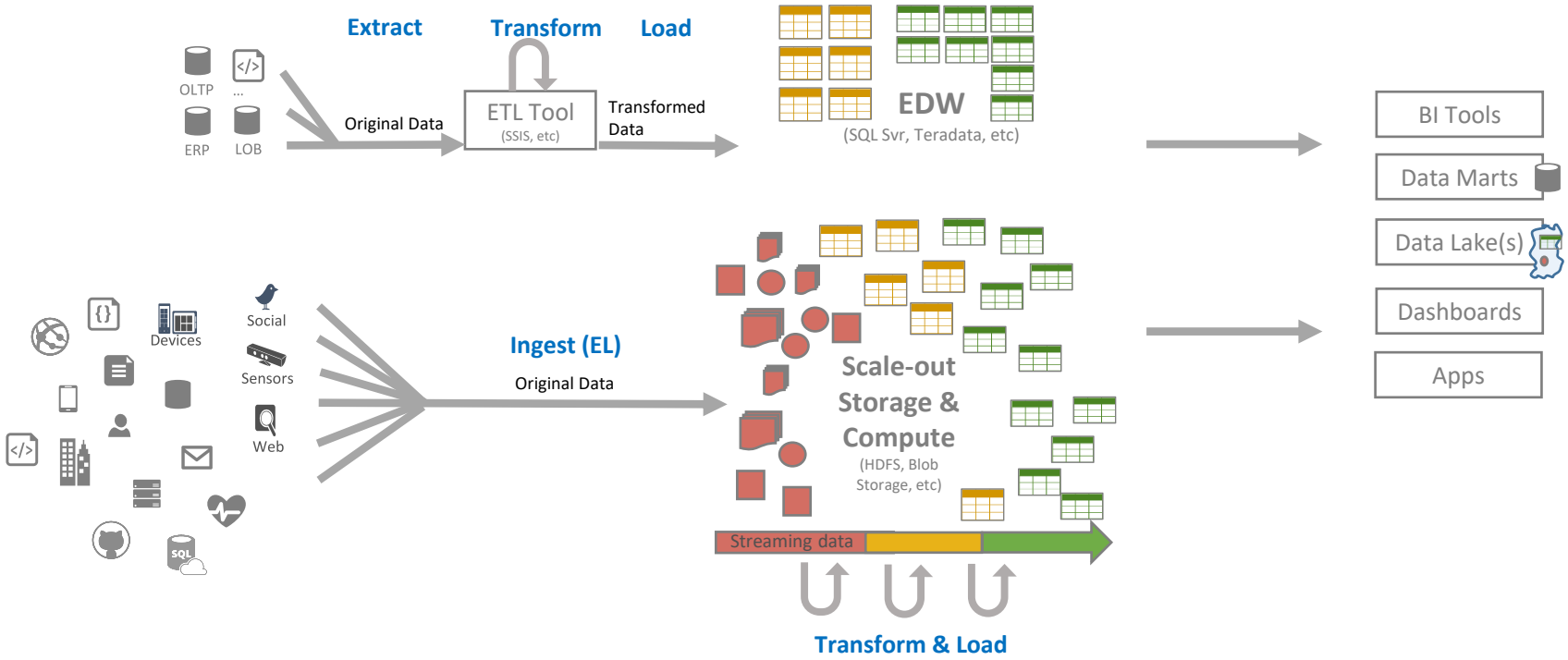
http://www.tylervigen.com/

Dataging
Managing through data

Big Data ≠ 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



Dataging
Managing through data

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

Dataging
Managing through data

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

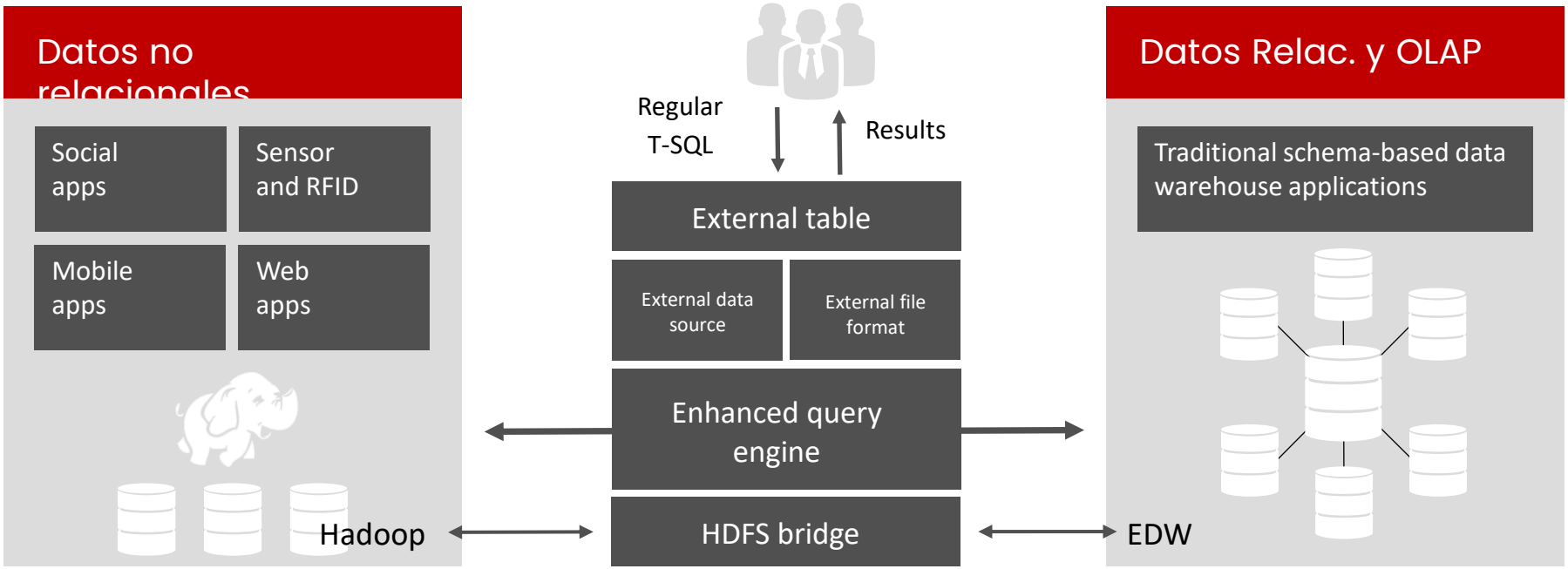
1. Llegan los datos
2. Se cargan en Hadoop
3. Análisis
4. Subconjuntos cargados en EDW

Rápido valor de los datos

Dataging
Managing through data

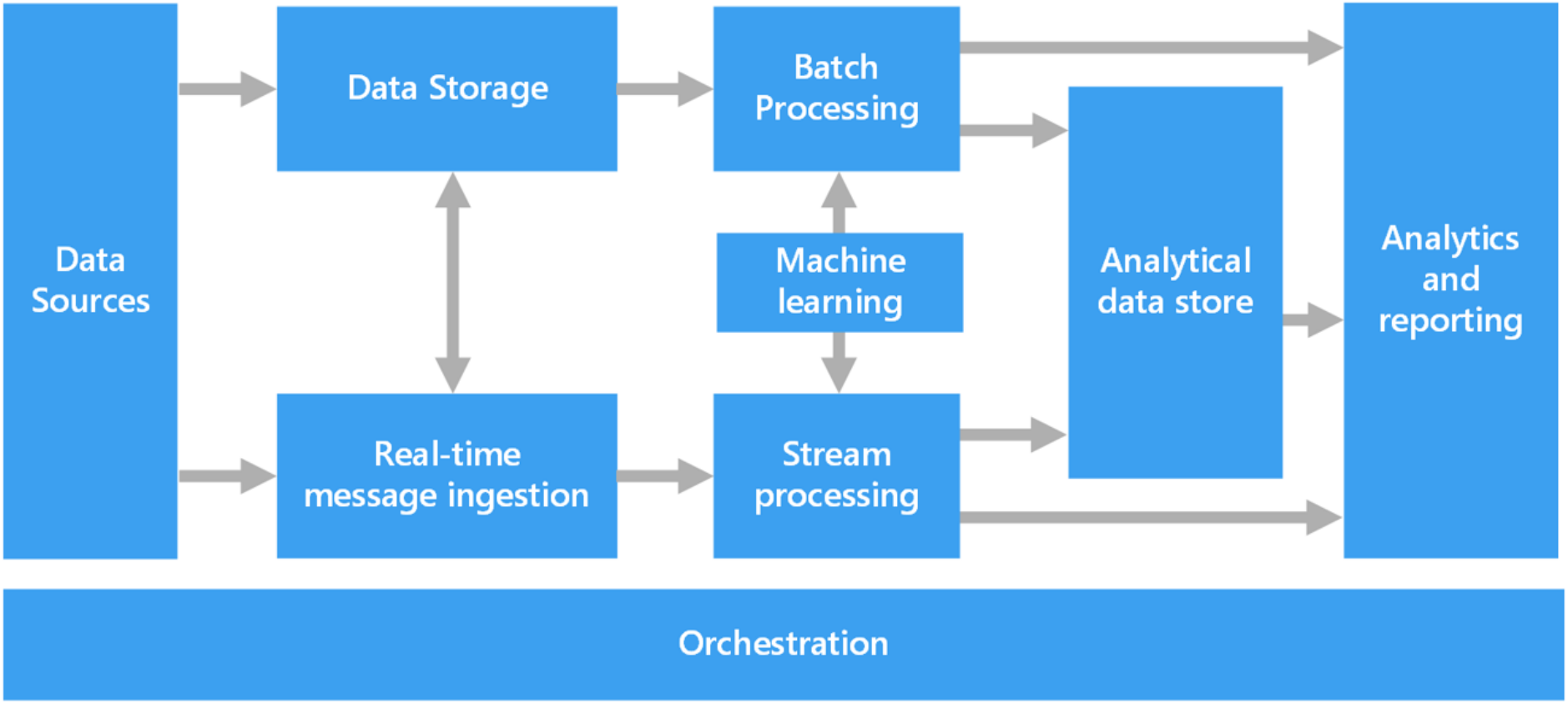
Cambios en Patrones

Básicamente construir un “puente” hacia Big Data



Dataging
Managing through data

Componentes de una Arquitectura Big Data



Dataging
Managing through 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

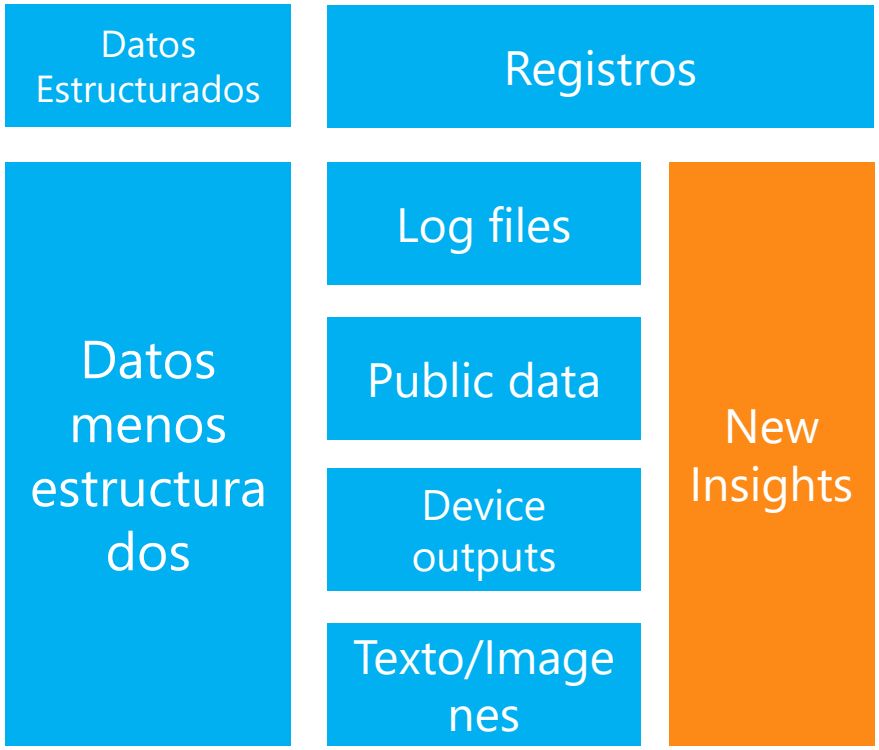
Dataging
Managing through data

Datos no estructurados no se están analizando

- Datos estructurados
  - BBDD Relacionales
  - BBDD Analíticas

Dataging
Managing through data

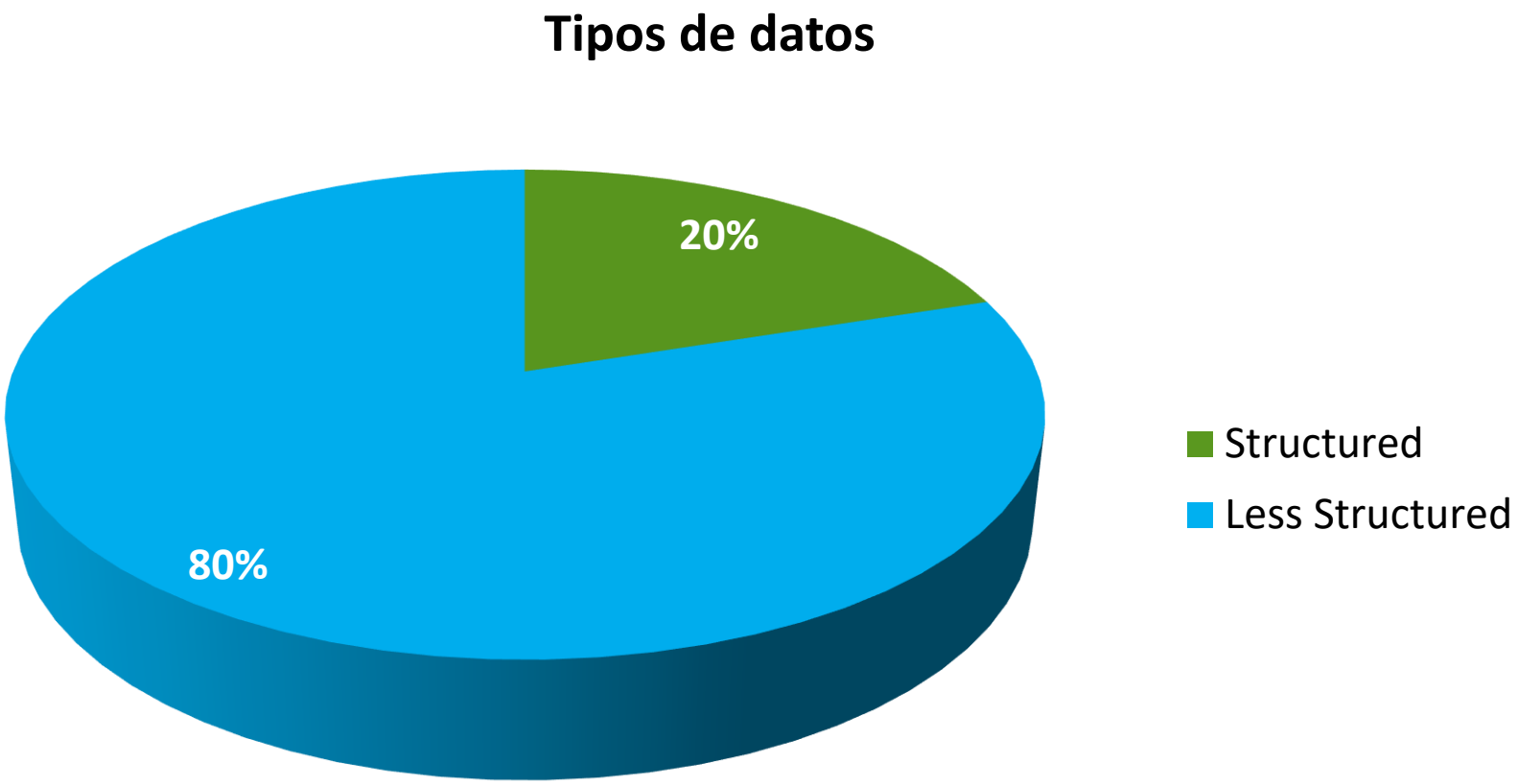
Datos no estructurados no se están analizando



- 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

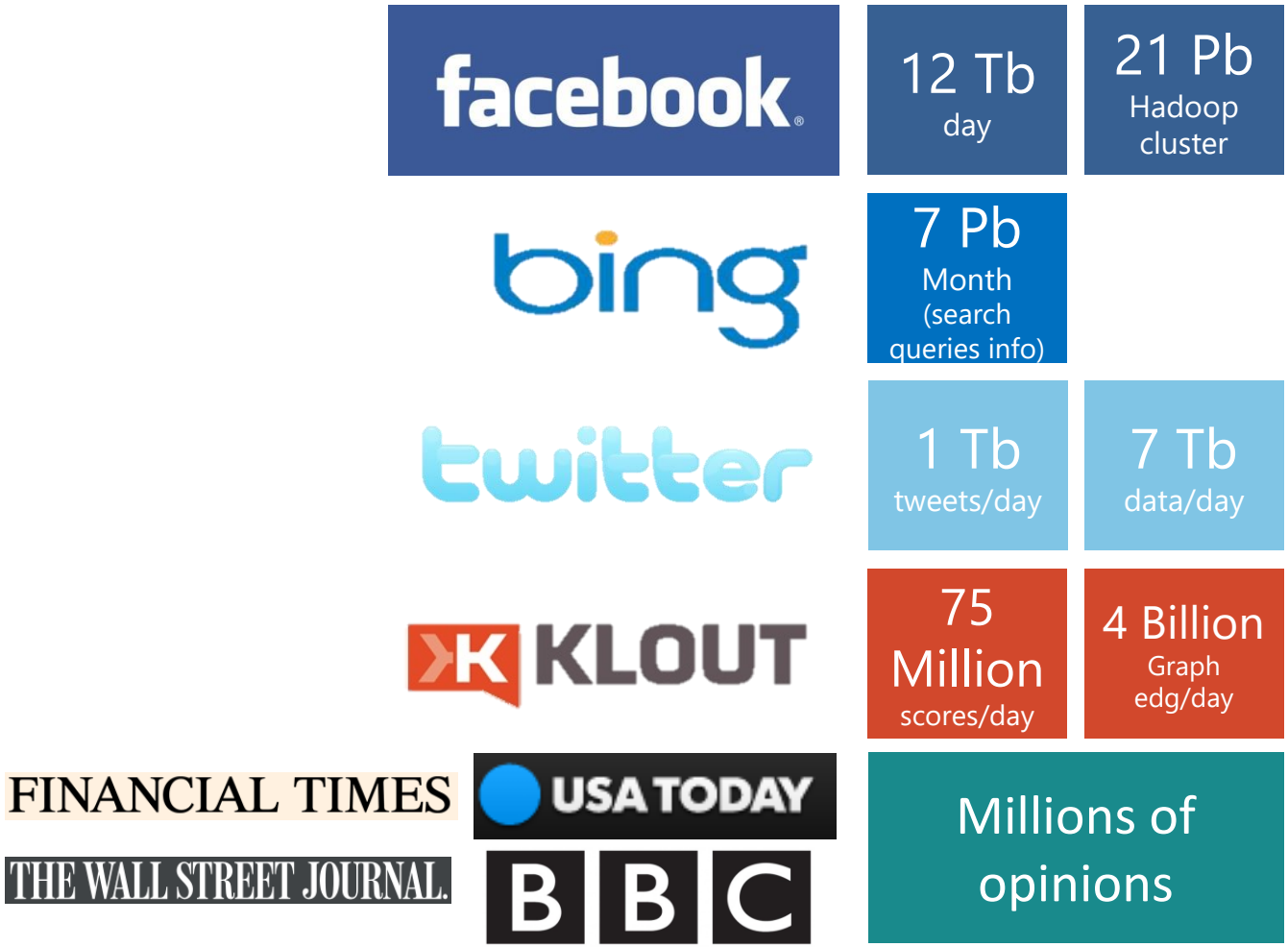
Dataging
Managing through data

Datos en las organizaciones



Dataging
Managing through data

Ejemplos de datos no estructurados



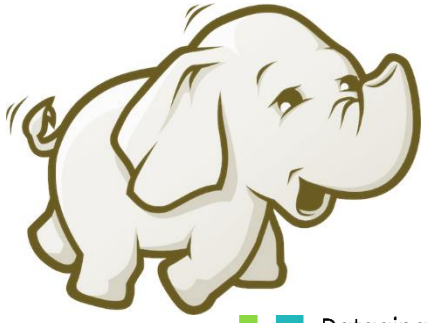
Dataging
Managing through data

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



Dataging
Managing through data

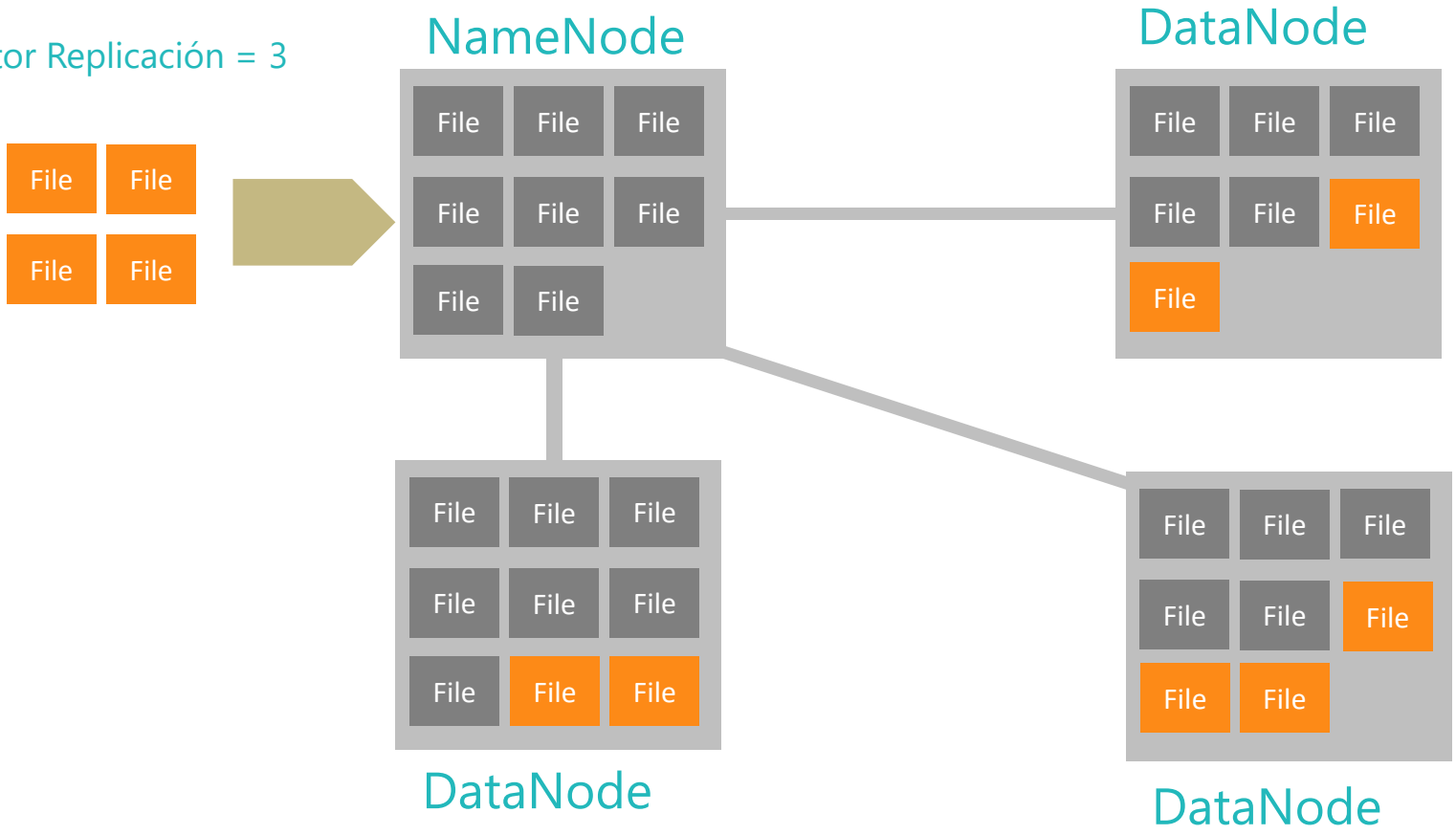
Hadoop

- 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



Dataging
Managing through data

Hadoop

- Escalable
  - Escala linealmente en capacidad de almacenamiento y procesado.
- 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)

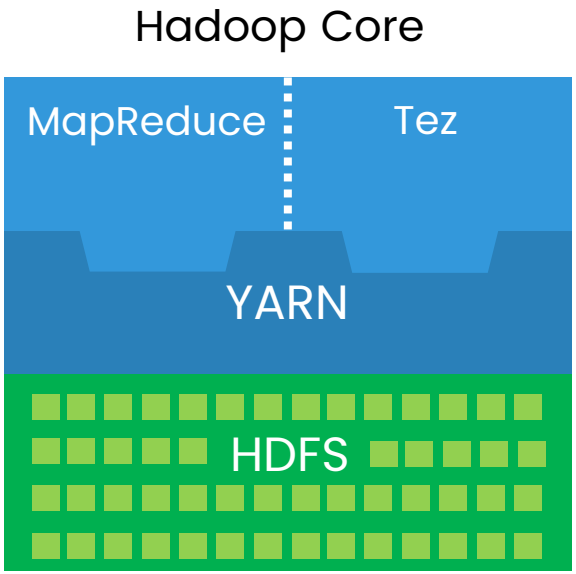
Historia Hadoop

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 Horthonworks

El Zoo de Big Data

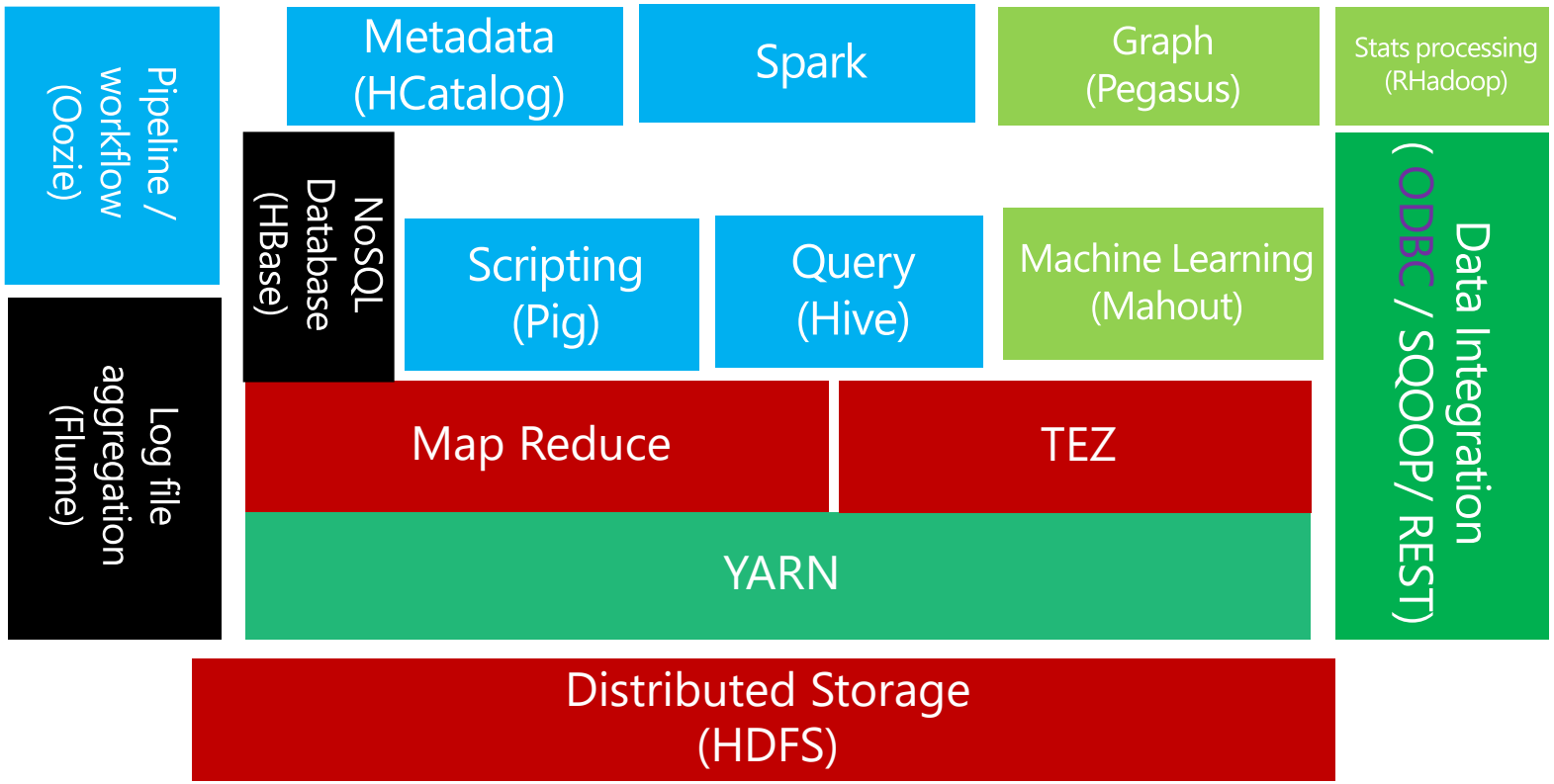
- El objetivo de Hadoop es crear un framework unificado para procesar big data

- Tres requisitos principales:
  - Escalabilidad
  - Eficiencia



Dataging
Managing through data

Ecosistema Hadoop



Dataging
Managing through data

Hive

- Sistema Data Warehouse para Hadoop
- Facilita las sumarizaciones 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

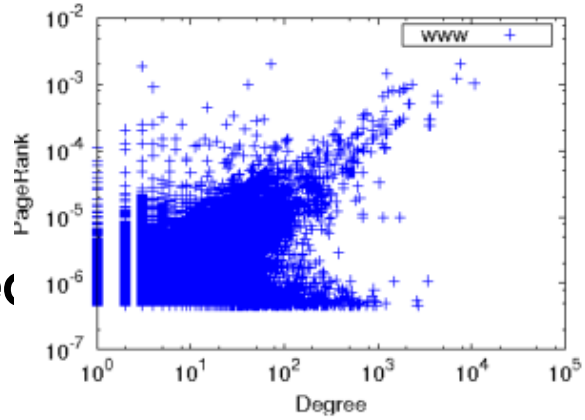
- Flume
  - Recolectar y mover grandes cantidades de datos
  - Ejecución distribuida

- Sqoop
  - Import y Export: RDBMS ←→ 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 **MapRe**



Dataging
Managing through data

Agenda

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

Dataging
Managing through data

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

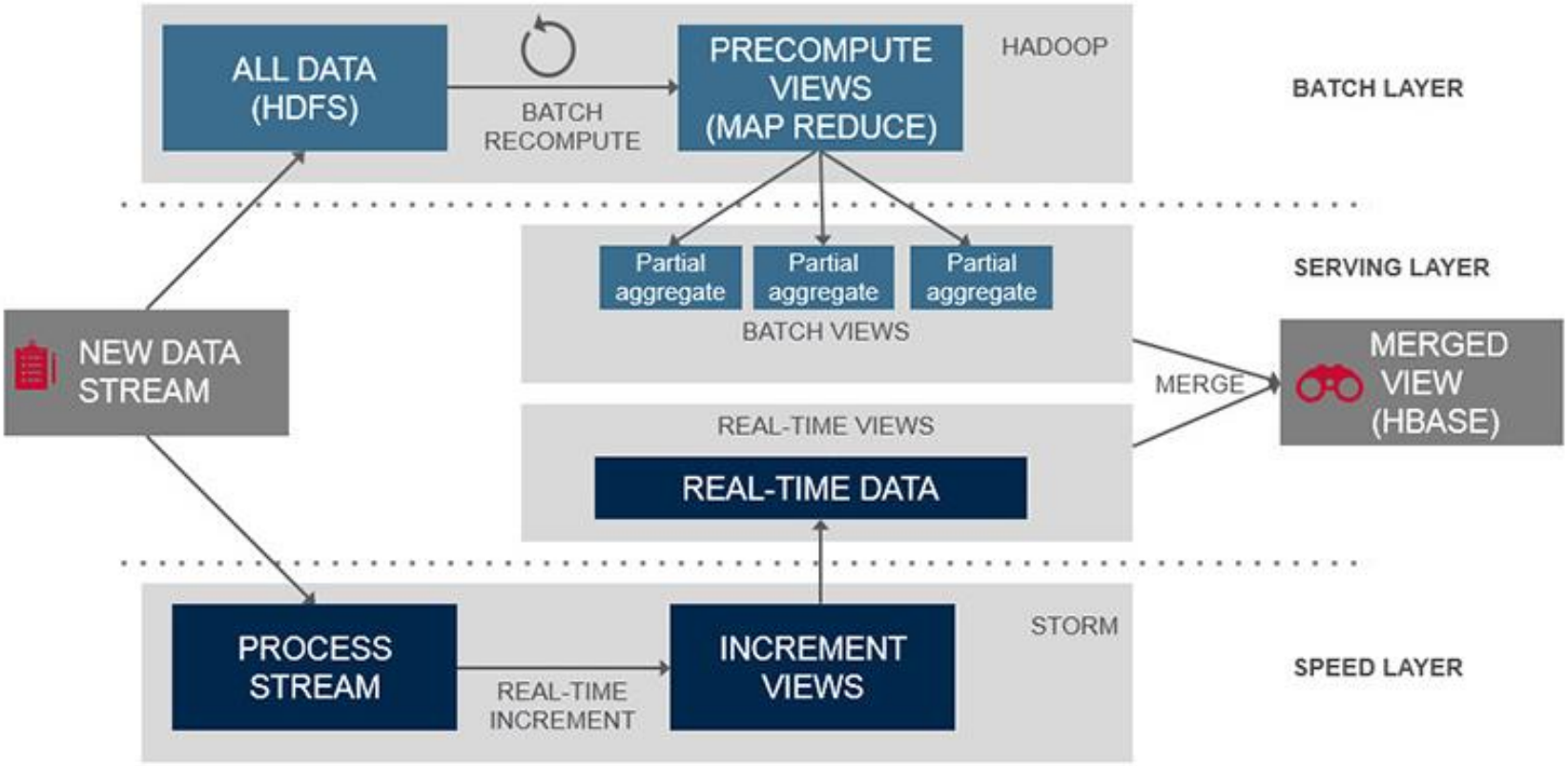
Datos en Tiempo Real

- 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



Dataging
Managing through data

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

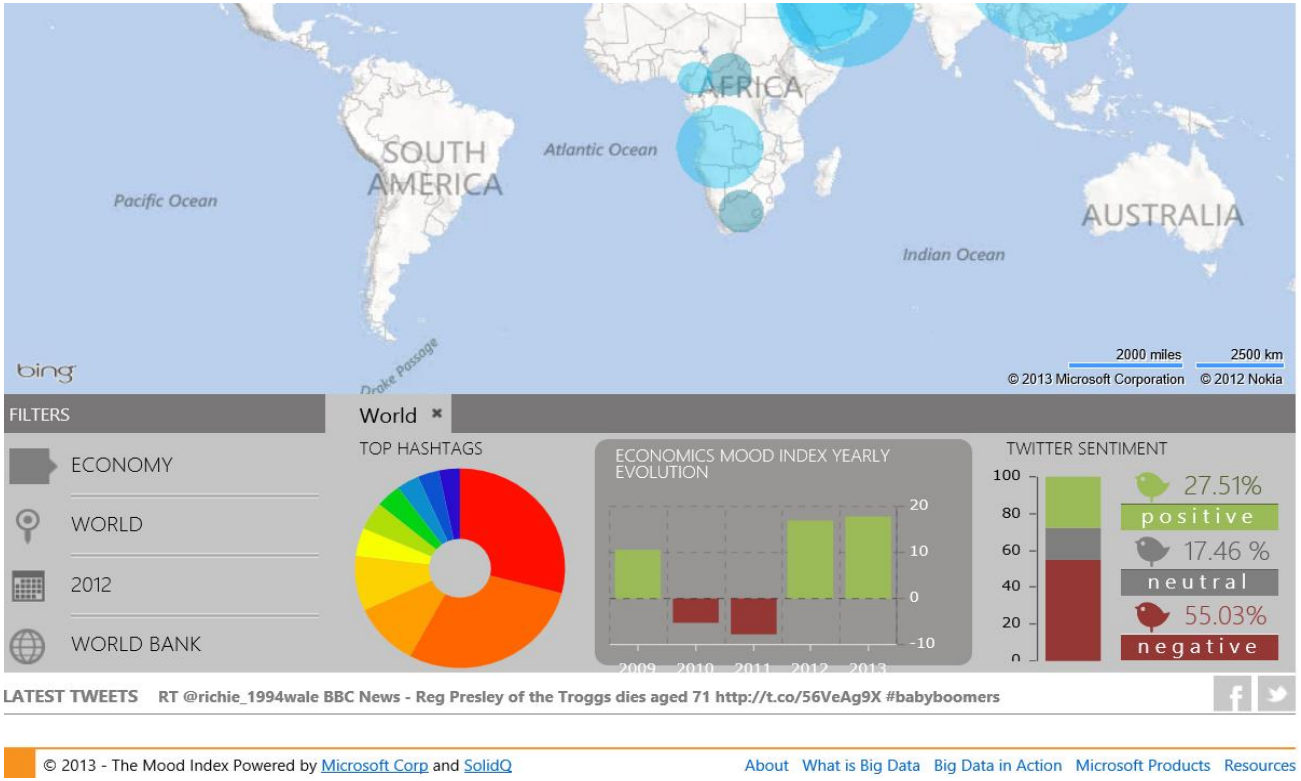
Casos de Estudio Hadoop (II)

- 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

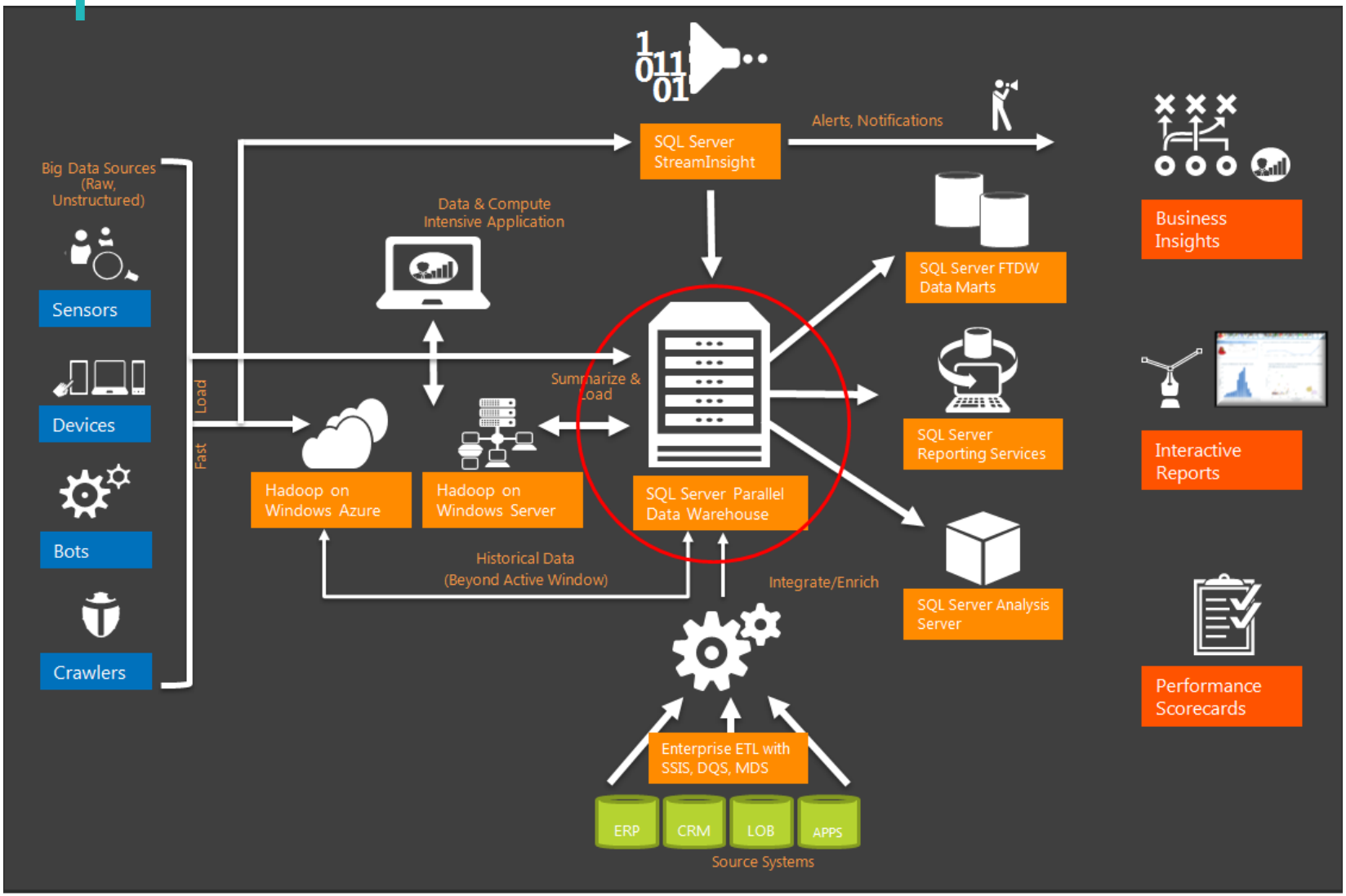


Dataging
Managing through data

MIDAMO

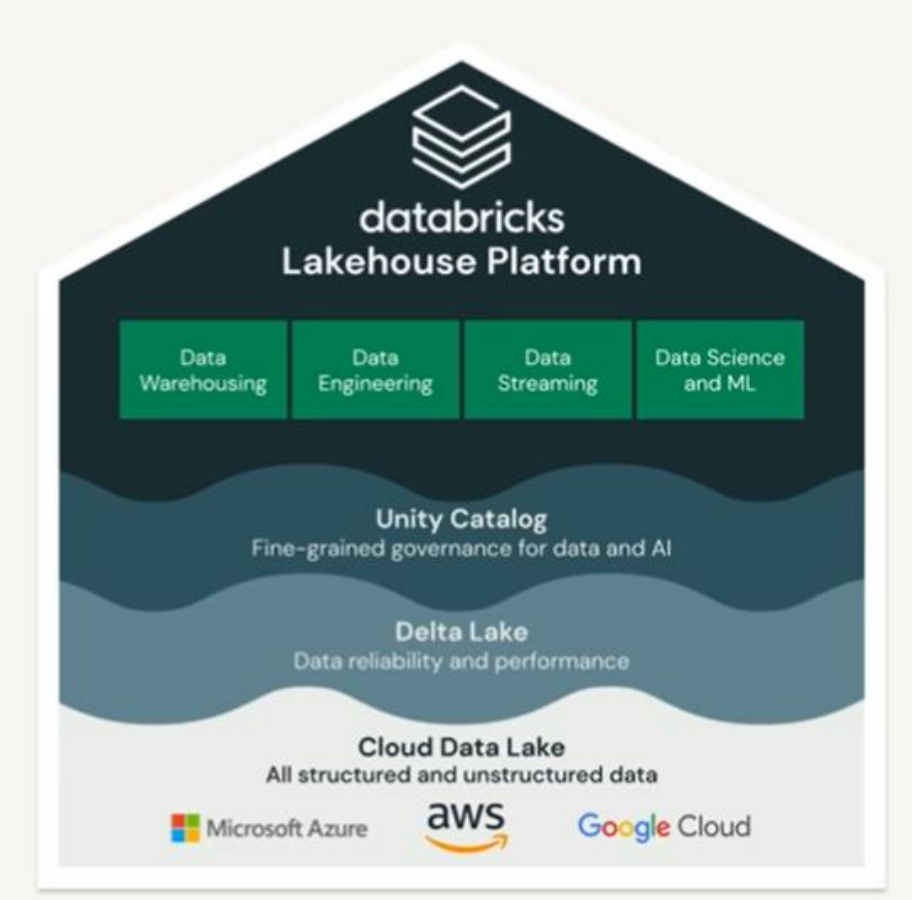
Dataging
Managing through data

Opciones



Dataging
Managing through data

Datalakehouse



Dataging
Managing through data