Inteligencia de Negocio y Gestión de Procesos (INGP)

BIG DATA

Tema 9

Profesores:

Juan C. Trujillo, Alejandro Maté

LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

1

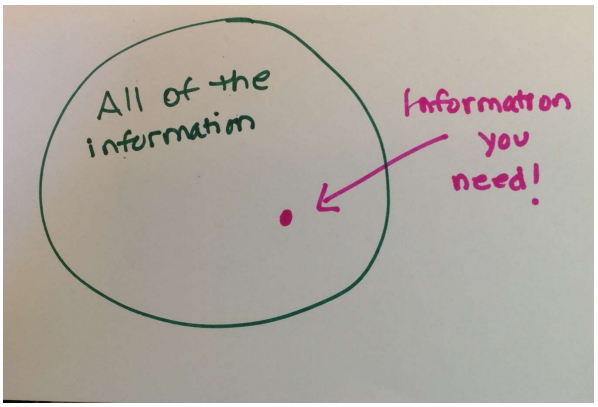
Índice

2

Indice

3

- Introducción de Big Data



INGP. 2021

3

Introducción Big Data

4

- Entendiendo Big Data

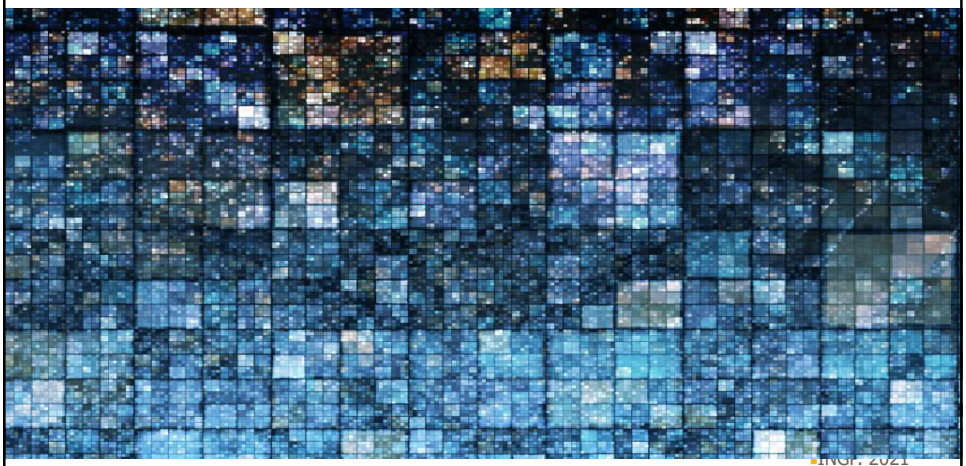
INGP. 2021

4

Introducción Big Data

5

- ¿Qué tal ?



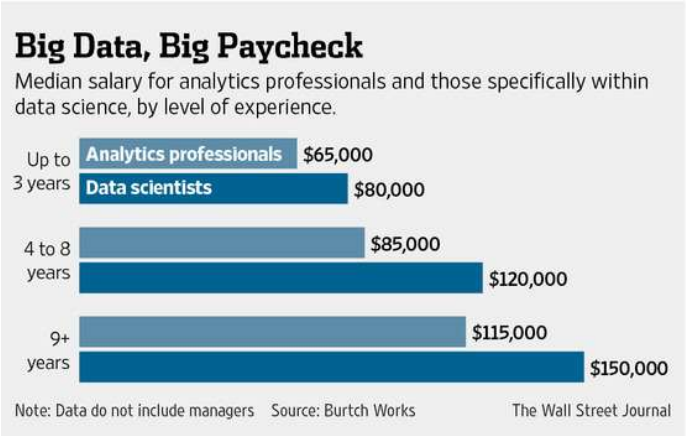
INGP. 2021

5

Introducción Big Data

6

- Seguro… que quiero trabajar en big data?



INGP. 2021

6

Introducción Big Data

7

- BIG DATA projects aren't one man thing
  - Servidores
  - Arquitectura
  - Programación
  - Diseño
  - Análisis
  - Dirección
    - DevOps, Backend, Frontend, Data scientist…

7

Introducción Big Data

8

- Data scientist
  - …a data scientist is 1) a data analyst in California or 2) a statistician under 35
    - Gartner blog post by analyst Svetlana Sicular
    - Estadística
    - R, Matlab, SAS, SPSS
    - Minería de datos
    - Procesamiento de lenguaje natural
    - Machine Learning
    - Map/Reduce, Hadoop, Hive, etc
    - Python
  - The notion of a Data Scientist is a little mad but then so is Big Data. Removing the buzzwords just leaves you with….*Data*.

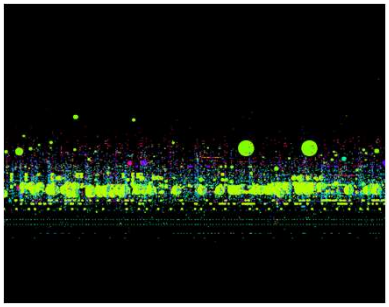
8

Introducción Big Data

9

No todo son analíticas

- creative coders, data designers and artists
  http://eyeofestival.com/



- http://content.stamen.com/visualizing_a_day_of_financial_transactions_on_nasdaq_part_2
- http://content.stamen.com/facebook_mapping_how_viral_photos_spread

9

Introducción Big Data

10

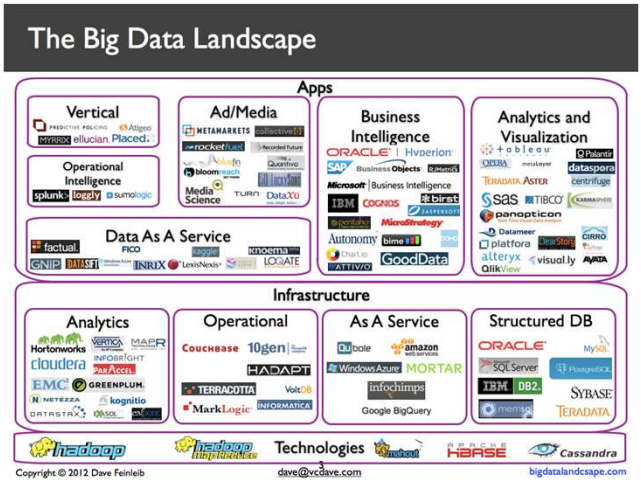
- BIG DATA para salvar el mundo
  - Siempre hemos tenido mucha información
  - Pero ahora gracias a nuevas herramientas se pueden analizar e interpretar
  - También se pueden almacenar más cantidad de información
    - Genoma Humano
    - Datos de Enfermedades
    - LHC

10

Introducción Big Data

11

- Ya hay muchos jugadores



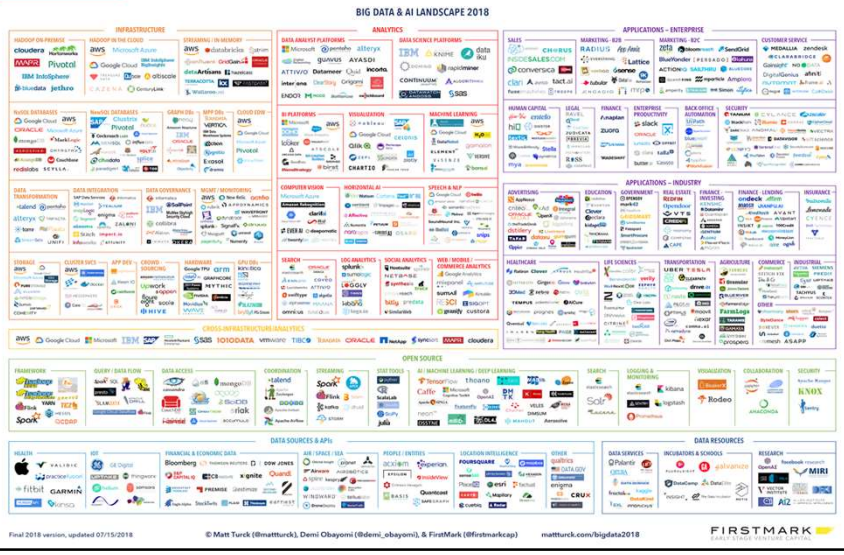
- INGP. 2021

11

Introducción Big Data

12

2018

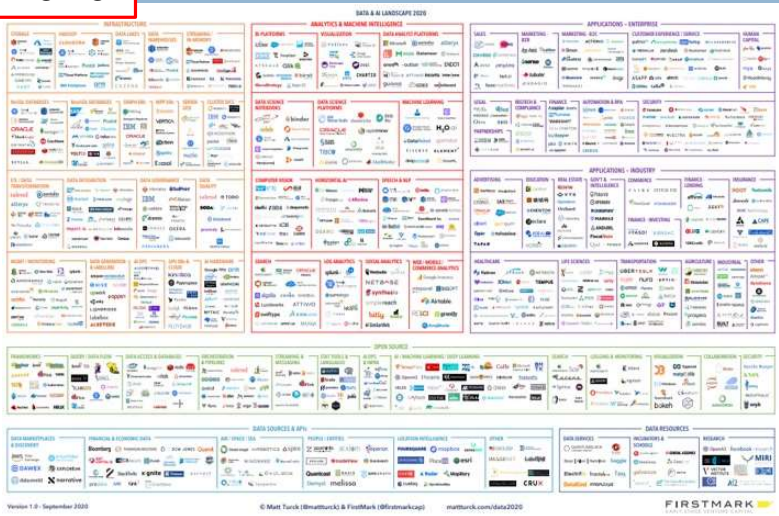


12

Introducción Big Data

13

2020



13

Introducción Big Data

14

- Dentro de Big Data se engloba o tiene que ver:
  - Smart City
  - Sensores
  - Seguridad
  - Privacidad
  - Inteligencia Artificial
    - …
  - Tendencias
  - Marketing
  - Psicología
    - …..

14

Índice

15

- Introducción Big Data 1° semana
- Introducción a Big Data y Business Intelligence
- Problemas
- Herramientas
- Casos de éxito
- Introducción NOSQL

15

Introducción a Big Data y Business Intelligence

16

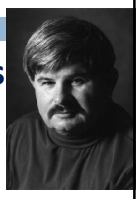
- Business Intelligence se basa en la **explotación de los recursos de información** de una organización, internos y externos
- Apoyo a la toma de decisiones estratégicas
- Respuestas a preguntas del tipo:
  - ¿Qué especialidad jurídica es la más demandada, en qué lugar y entre que segmento de población?
  - ¿Qué horario laboral me permite racionalizar el consumo energético de mi empresa?

16

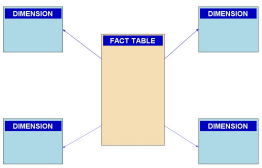
16

Introducción a Big Data y Business Intelligence

17



- La información se guarda en **Almacenes de Datos**
  - Desde finales de los 80
  - "Una colección de datos orientados por tema, integrados, variables en el tiempo y no volátiles que se emplea como apoyo a la toma de decisiones estratégicas" (Bill Inmon)
  - Características
    - **Datos estructurados**
    - Almacenados en SGBDR
    - **Volumen**→ Terabytes - Petabytes

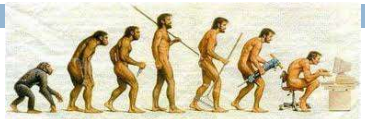


17

17

Introducción a Big Data y Business Intelligence

18



- Aumento de volumen y cambios en las características de los datos
  - Grandes empresas 1 TB / Hora
  - Facebook 10 TB / Hora
  - Datos semi-estructurados o no estructurados
    - Texto, imágenes, JSON, XML, RSS, …
  - Aumento velocidad de generación de los datos
- Las técnicas de Almacenes de Datos no son adecuadas para este análisis
- La solución a estos problemas es el enfoque **Big Data**

18

18

Big Data

19

- Big Data
  "Forma de afrontar el procesamiento o análisis de grandes volúmenes de información que por su naturaleza desestructurada no pueden ser analizados, y en un tiempo aceptable, usando los procesos y herramientas tradicionales de BI" (IBM)
- Características (5v's)

Veracidad | Velocidad | Valor | Volumen | Variedad

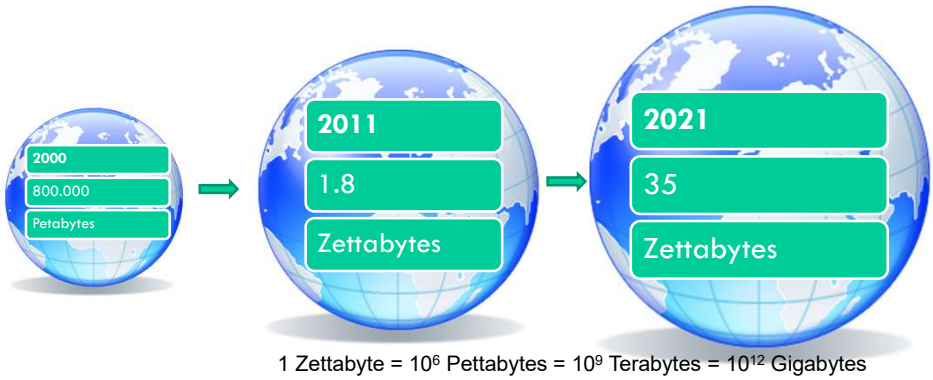
19

19

Big Data – Volumen

20

- Volumen: capacidad para procesar **grandes volúmenes de datos**



1 Zettabyte = 10^6 Pettabytes = 10^9 Terabytes = 10^{12} Gigabytes

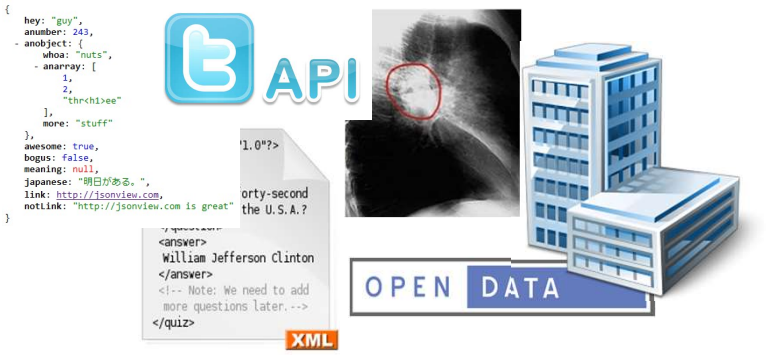
20

20

Big Data – Variedad

21

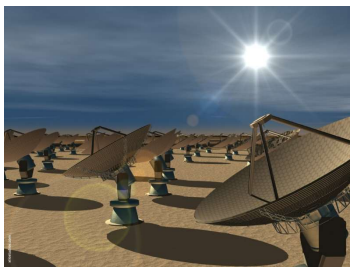
- Variedad: capacidad para soportar el aumento en la **heterogeneidad** de las **fuentes** a procesar.



Big Data – Velocidad

22

- Velocidad: velocidad a la que fluye la información.



Telescopio SKA
10 petabytes / hora

Twitter
100.000 tweets / min

Big Data – Veracidad

23

- Veracidad: incertidumbre datos = incertidumbre conocimiento extraído.
  - 1 de cada 3 ejecutivos **desconfía** de los datos que usan para tomar decisiones
  - ¿Encuestas precisas?
  - Uso de datos incorrectos supone **grandes perdidas** (varios billones de euros al año)

23

23

Big Data – Valor

24

- ¿Por qué queremos implementar esta tecnología?
- ¿Supone alguna ventaja para nuestra empresa?

24

24

Índice

25

- Introducción Big Data 1° semana
- Introducción a Big Data y Business Intelligence
- Problemas
- Herramientas
- Casos de éxito
- Introducción NOSQL

Problemas – Integración

26

- El uso de distintas fuentes de datos da lugar a problemas de **incoherencia**
- Distintas formas de representar los mismos datos
  - Descripción:
    - J.A. Rodríguez ←→ José A. Rodríguez
  - Unidades:
    - Estatura: 1,70 mts ←→ 170 cm
- Su resolución puede requerir la aplicación de procesos que tienen un alto coste temporal (**ETL's**)

Problemas – API's

27

- Depender de servicios de datos libres proporcionados por empresas externas
  - Cambios en el formato…
  - Cambios en las condiciones de servicio…
  - Cambios en las API's de obtención de datos…
  - Averías
  - Cierre del servicio

27

27

Problemas – Aspectos Legales

28

- Usar datos proporcionados a través de **terceros**
  - ¿De quién es la **propiedad** de los datos obtenidos tras el procesamiento y análisis?
  - ¿Es **lícito** usarlos para la creación de nuestras aplicaciones?

28

28

Problemas – Aspectos Legales

29

- La mayoría de la población **desconoce**:
  - Clausulas de privacidad
    - **Redes Sociales** : Geo localización activada por defecto, clausulas difíciles de comprender, complicadas opciones de privacidad...
  - Posibles usos de los datos:
    - **Correos electrónicos**: Usados por los proveedores del servicio para marketing…

29

29

Problemas – Aspectos Legales

30

- ¿Es **ético** analizar a una persona por los datos de las redes sociales? ¿Es **legal**?

  - **¿Sí? :** Ausentismo y rendimiento laboral, fraude al seguro, criminales, revueltas, epidemias….
  - **¿No? :** amistades, relaciones sentimentales, ideologías, pensamientos…
- En cualquier caso, hemos de estar muy seguros de la **veracidad de los datos y resultados obtenidos**

30

30

Índice

31

- Introducción Big Data 1° semana
- Introducción a Big Data y Business Intelligence
- Problemas
- Herramientas
- Casos de éxito
- Introducción NOSQL

Herramientas

32

- Que características tiene una herramienta para Big Data:
  - **Escalable** para que soporte fácilmente petabytes
  - **Distribuido** (en varios procesadores, diferentes lugares y características)
  - Guardar los datos en el **formato original,** pudiendo hacer querys sin convertir el formato o moverlo
  - Capacidad de poder realizar User-defined functions (UDFs)

Herramientas

33

- Ejecutar **UDFs** en petabyte data en minutos
- Permitir **guardar muchos formatos,** desde imágenes audio, datos jerarquizados, pares nombre-valor…
- Cargar datos de multiples fuentes al menos **GB/segundo**
- Cargar los datos en BD **antes de declarar o descubrir su estructura**
- ….

- 2 Soluciones **RDBMSs** y **MapReduce/Hadoop**

Herramientas

34

Arquitectura BigData típica

Características:

- Almacenamiento de diferentes tipos de datos
  - Semi-estructurados (Marketing/ campañas/ móbil/ web logs)
  - Estructurados
  - Ficheros de log
- Carga de datos desde diferentes bases de datos (MySQL, Oracle, PostgreSQL, MongoDB, etc)
- Minería de datos
- Analíticas
- Almacenes de datos para reporting
- Análisis por lotes (Hadoop)
- Web caching
- Search

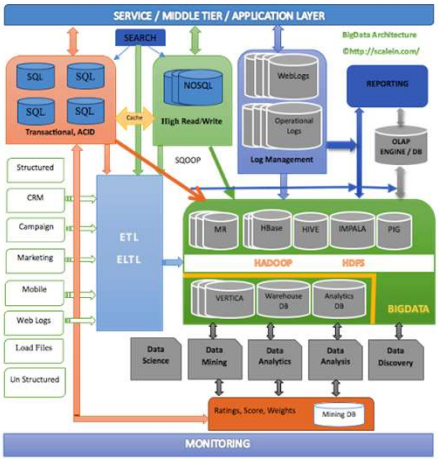


Imagen via (http://scalein.com/)

MapReduce/Hadoop

35

- Open source top-level Apache
- Desarrollado por Google 2000s
- MapReduce es un framework que ejecuta UDF
- Muchas Bases de datos están implementando interfaces para permitir que Hadoop Jobs, de forma distribuida en sus instancias de bases de datos.

35

MapReduce/Hadoop

36

Extended Relational DBMS	MapReduce/Hadoop
Proprietary, mostly	Open source
Expensive	Less expensive
Data must be structured	Data does not require structuring
Great for speedy indexed lookups	Great for massive full data scans
Deep support for relational semantics	Indirect support for relational semantics, e.g., Hive
Indirect support for complex data structures	Deep support for complex data structures
Indirect support for iteration, complex branching	Deep support for iteration, complex branching
Deep support for transaction processing	Little or no support for transaction processing

Figure 21-2: Comparison of relational DBMS and MapReduce/Hadoop architectures.

36

Hadoop
MapReduce

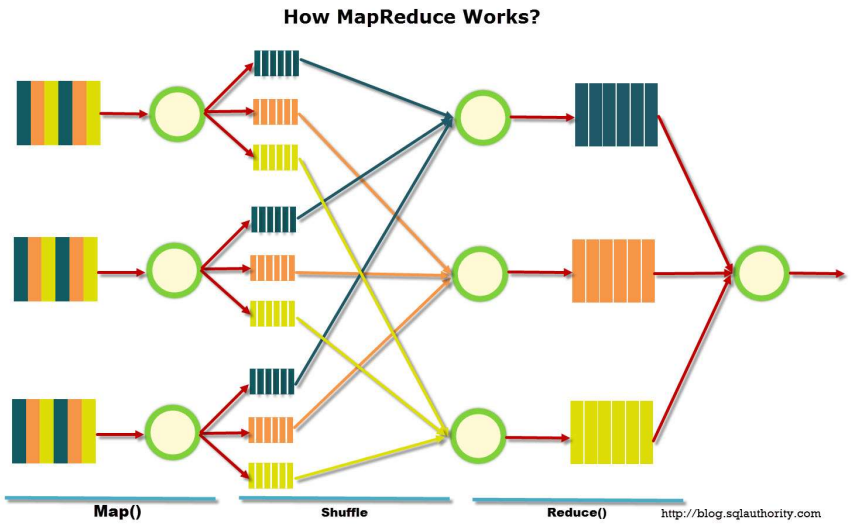
37

- ***MapReduce*** - software framework. Permite escribir programas para procesar grandes cantidades de datos no estructurados en clusters distribuidos de procesos.
  - 2 Fases
    - Map (Se realiza en paralelo para cada entrada):
      - 1- Entrada (clave, valor) y devuelve una lista de pares (clave2,valor2).
      - 2- Junta todos los pares con la misma clave de todas las listas y los agrupa. Creando un grupo por cada una de las diferentes claves generadas
      - Map(k_1,v_1) -> list(k_2,v_2)
    - Reduce
      - Entrada lista de valores, salida colección de valores
      - Reduce(k2, list (v2)) -> list(v3)

37

Hadoop
MapReduce

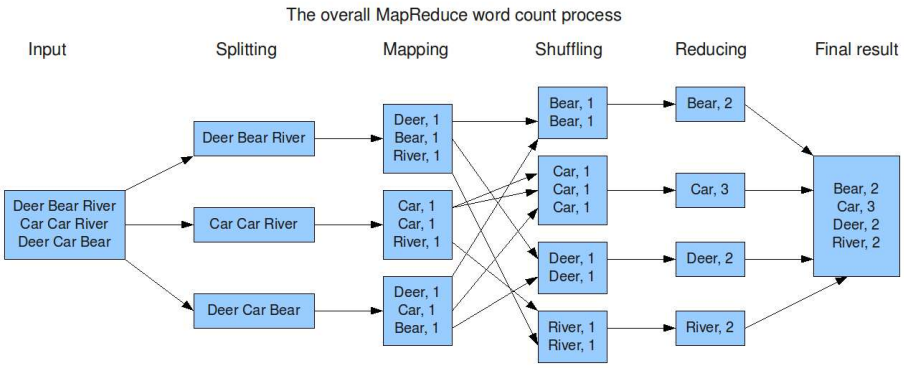
38



38

Hadoop MapReduce

39



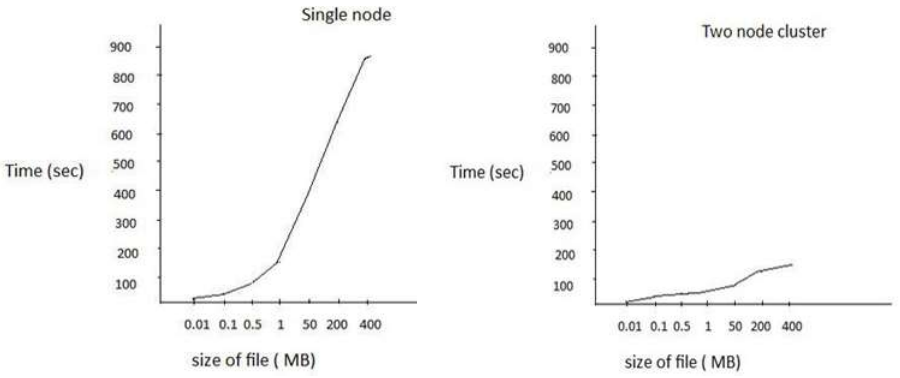
INGP. 2021

39

Hadoop MapReduce

40

- Fuente: http://www.ibm.com/developerworks/ssa/cloud/library/cl-mapreduce/



INGP. 2021

40

Hadoop

41

- ***Hive*** – Permite utilizar un lenguaje similar al estandar SQL. Hive Query Language (HQL). Hive re-escribe las consultas a operaciones de MapReduce para utilizarlas en clusters de Hadoop.
- ***NoSQL*** – Final del tema
- ***Hadoop Distributed File System (HDFS)*** – Sistema de ficheros, distribuido, escalable y portable en Java.

41

Hadoop

42

- ***Sqoop*** – Es una herramienta que ha sido diseñada para el volcado eficiente de datos entre una distribución Hadoop de Apache y Bases de datos relacionales. (SQL to Hadoop = Sqoop)
- • ***Pig*** – Es una plataforma de programción para escribir programas de MapReduce con Scripts de PIG.
- • ***Oozie*** –Es un planificador de flujos de trabajo para organizar Hadoops Jobs.
  - Hadoop jobs =Java map-reduce, Streaming map-reduce, Pig, Hive, and Sqoop.

42

Casos de éxito

43

- Casos de éxito
  - Recomendación Amazon
  - Elecciones OBAMA

43

43

Casos de éxito - Amazon

44

- Amazon usa un **sistema de recomendación** de productos a posibles compradores.
- Proporciona a cada visitante de Amazon.com una página web personalizada
  - Nos ofrece de forma **automática** los productos que el sistema determina que podríamos querer adquirir

44

44

Casos de éxito - Amazon

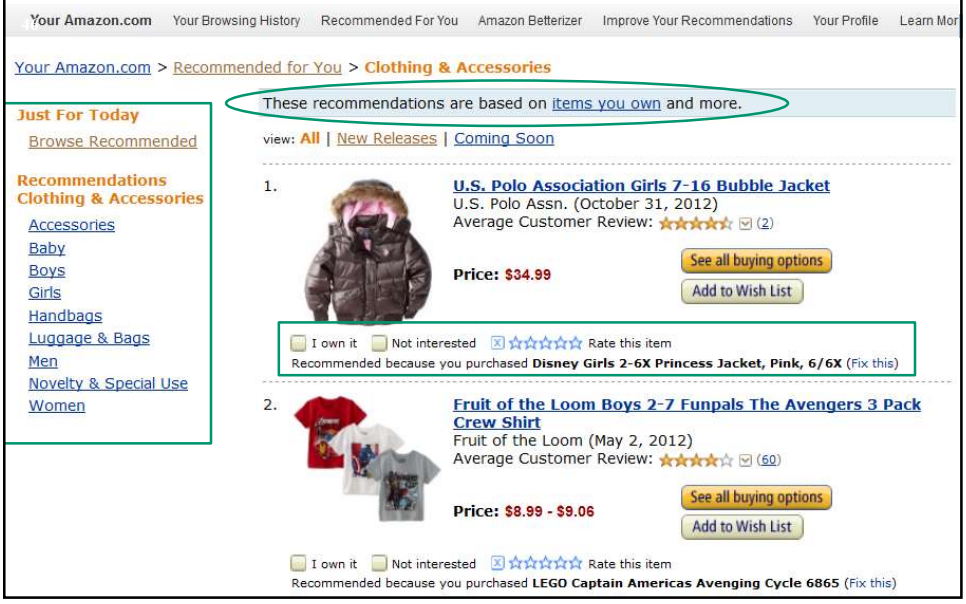
45

- Amazon implementa un **enfoque híbrido**
  - **"ítem to-ítem collaborative filtering":** historial de compras, artículos en el carrito de la compra, puntuaciones y "likes" sobre artículos, lo que han visto y comprado otros usuarios con perfiles similares…

45

45

Casos de éxito - Amazon



46

Casos de éxito – Campaña Obama

47

- Aplicando con éxito tecnología Big Data en sus campañas electorales desde 2008
  - Predicción de resultados electorales
  - Retroalimentación de la estrategia de campaña electoral
- Para las elecciones de 2012 conto con un equipo de 50 analistas y 50 ingenieros

Equipo Mitt Romney

Equipo B. Obama

47

47

Casos de éxito – Campaña Obama

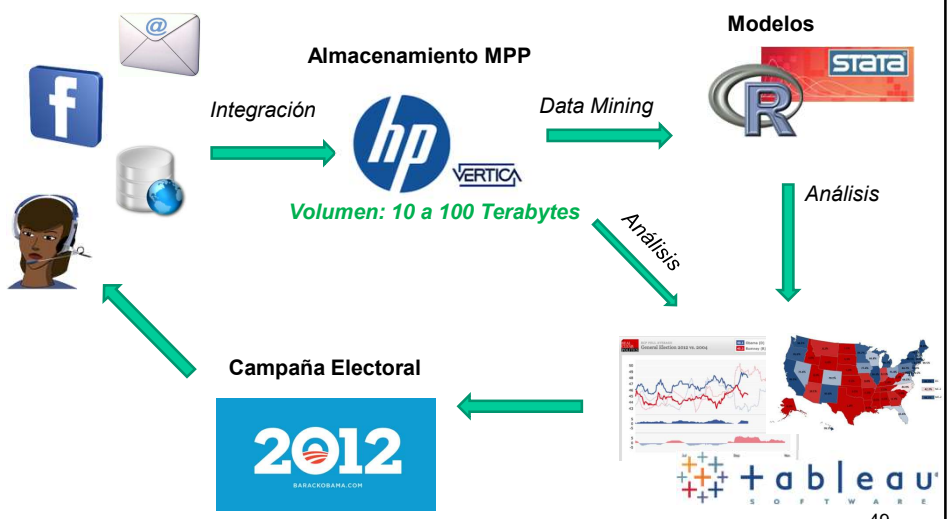
48

Medio	Aplicación	Datos / Conocimiento
Teléfono, Email (Comunicación Directa)	Encuestas individuales sobre las actividades y preferencias del votante	Sistema de puntuación que describe a los votantes de forma individual (+50 Variables)
Social Media	Facebook / Twitter	Búsqueda en páginas de apoyo a Obama de posibles simpatizantes / +50.000 Cuentas de Twitter asociadas a la política
Smartphones	Aplicación móvil	Agentes electorales – Encuestas intención de voto
Web	"Dashboard"	Sistema de recogida de opiniones de los ciudadanos
Otros	Bases de datos ya existentes	Datos de 180 millones de votantes, afiliados, voluntarios, donaciones, webs apoyo a Obama…

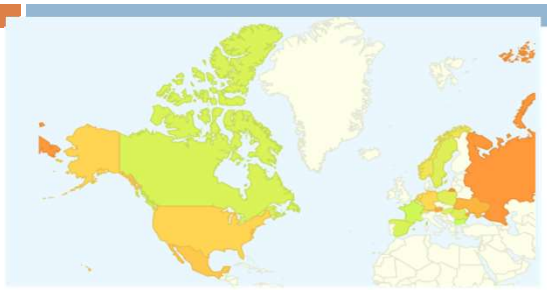
48

48

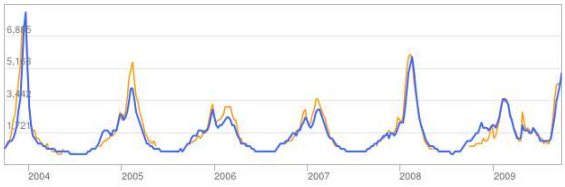
Casos de éxito – Campaña Obama



Predictive analytics: flu trends



- United states
- Flu Activity
- **United States**



Data

- **Google Flu Trends estimate**

Índice

51

- Introducción Big Data 1° semana
- Introducción a Big Data y Business Intelligence
- Problemas
- Herramientas
- Casos de éxito
- Introducción NOSQL

51

Introducción NoSQL

52

Conceptos básicos NoSQL "Not only SQL"

- Actualmente la mayoría de la información generada desde internet es:
  - NO ESTRUCTURADA.
    - El esfuerzo de estructurarla es demasiado grande.
  - Gran volumen de información

- Aparecen los nuevos sistemas de gestión de datos para **información no estructurada** y **distribuidas**.
  - Algunas soluciones:
    - Clave-valor (Amazon DynamoDB http://aws.amazon.com/es/dynamodb/)
    - Columnas (Cassanda http://cassandra.apache.org/)
    - Orientados a documentos (MongoDB http://www.mongodb.org/)
    - Grafos (Neo4j http://www.neo4j.org/)

52

Introducción NoSQL

53

Las BD NoSQL:

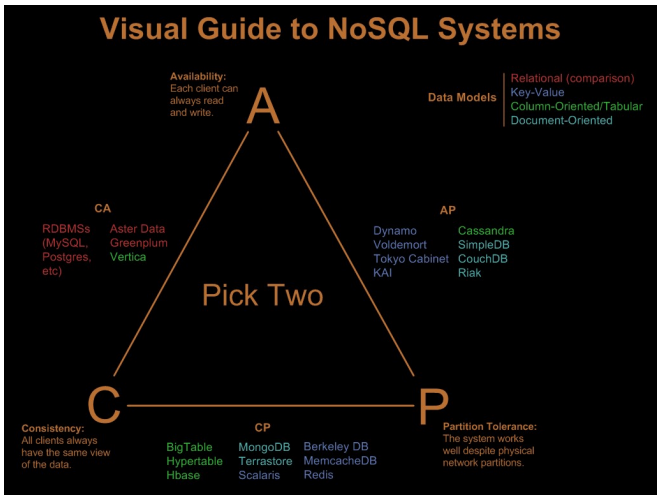
No tienen Schemas, no permiten Joins y escalan horizontalmente.

- Por ejemplo: Guardar historial de pacientes.
  - Almacén de datos:
    - Diseñar el esquema estrella identificando hechos y dimensiones.
  - NoSQL BD:
    - No hace falta diseñar el esquema de datos, solamente introducir los datos.
      - EJ: MongoDB (Orientado a documentos) insertando los JSON con la información, sería suficiente aunque fueran diferentes unos de otros.

Introducción NoSQL

54

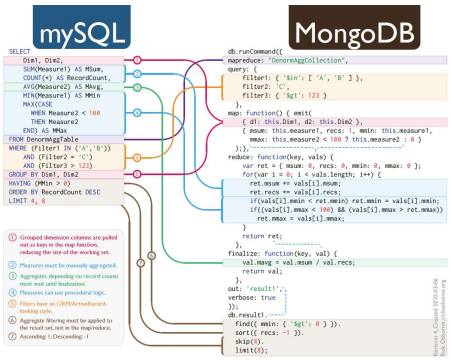
- Cual elegir?: (Fuente: http://blog.nahurst.com/visual-guide-to-nosql-systems)



Introducción NoSQL

55

- Cual elegir?:
  - Cuando elegir NoSQL sobre BD estructuradas
    - Cuando se necesita :
      - Alta disponibilidad
      - Alta escalabilidad
      - No se tienen claro el esquema de datos
      - Se necesita analizar y operar grandes cantidades de datos.
- Como regla general: Si la información es estructurada y la escalabilidad no es un punto crítico, mejor utilizar BD relacionales.
  - Las querys permiten más expresividad.
  - Principio ACID



55

Visualización de Big Data

56

Visualizar no es complicado, lo complicado es transmitir conocimiento a partir de la visualización.

Los métodos de visualización más usados:

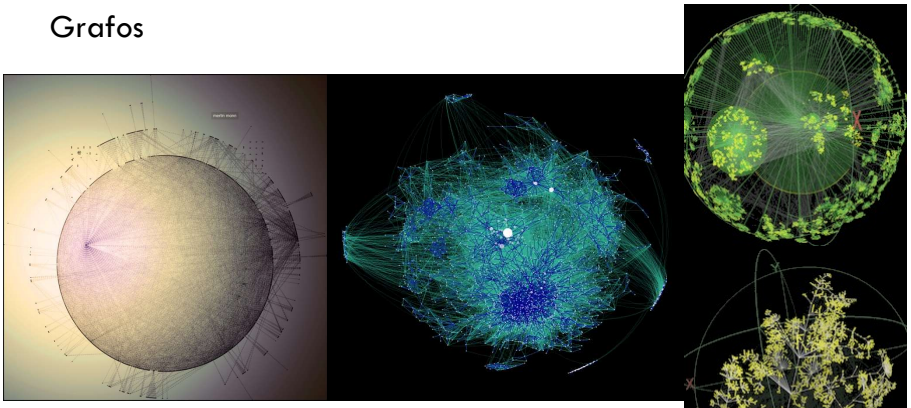
- Grafos
- Mapas

56

Visualización de Big Data

57

Grafos



Fuente:http://kunalanand.com/delicious/

http://www.caida.org/tools/visualization/walrus/gallery1/

http://datamining.typepad.com/gallery/blog-map-gallery.html

57

Visualización de Big Data

58

Mapas

http://www.akamai.com/html/technology/dataviz1.html

http://demographics.coopercenter.org/DotMap/index.html

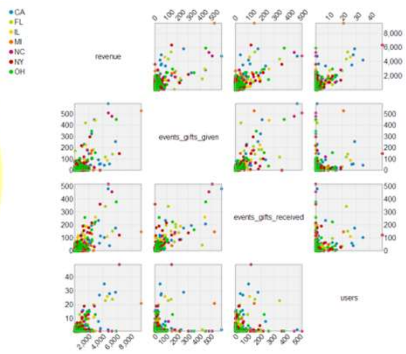
58

Visualización de Big Data

59

Gráficas de Pentaho para BigData

Sunburst

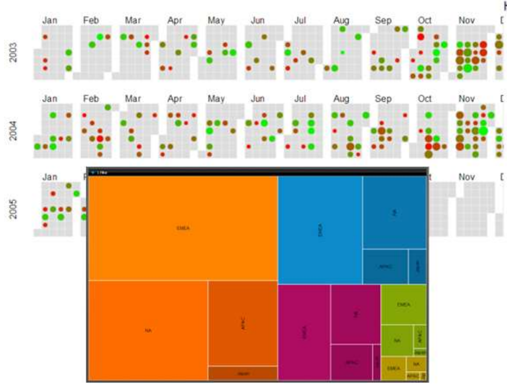


Trellis

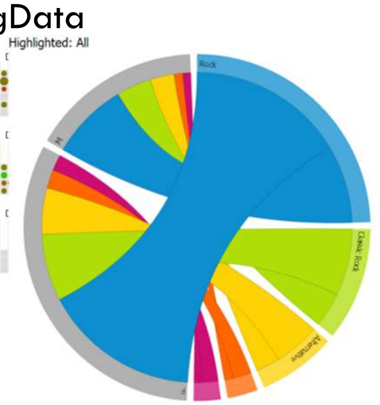
Visualización de Big Data

60

Gráficas de Pentaho para BigData



Treemap

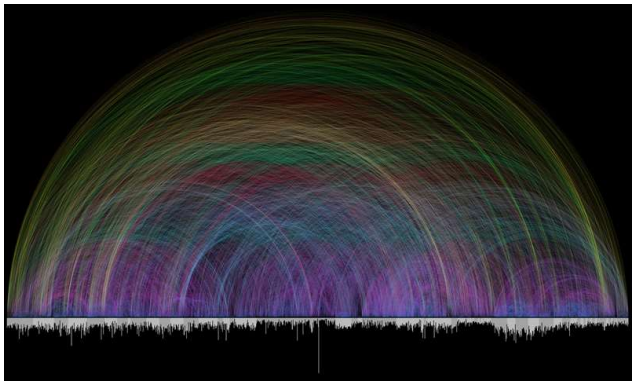


Chord

61

Otras visualizaciones

http://www.nytimes.com/interactive/science/space/keplers-tally-of-planets.html?_r=1&



61

Inteligencia de Negocio y Gestión de Procesos (INGP)

BIG DATA INTRODUCCIÓN

Tema 9

Profesores:

Juan C. Trujillo, Alejandro Maté

LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

62