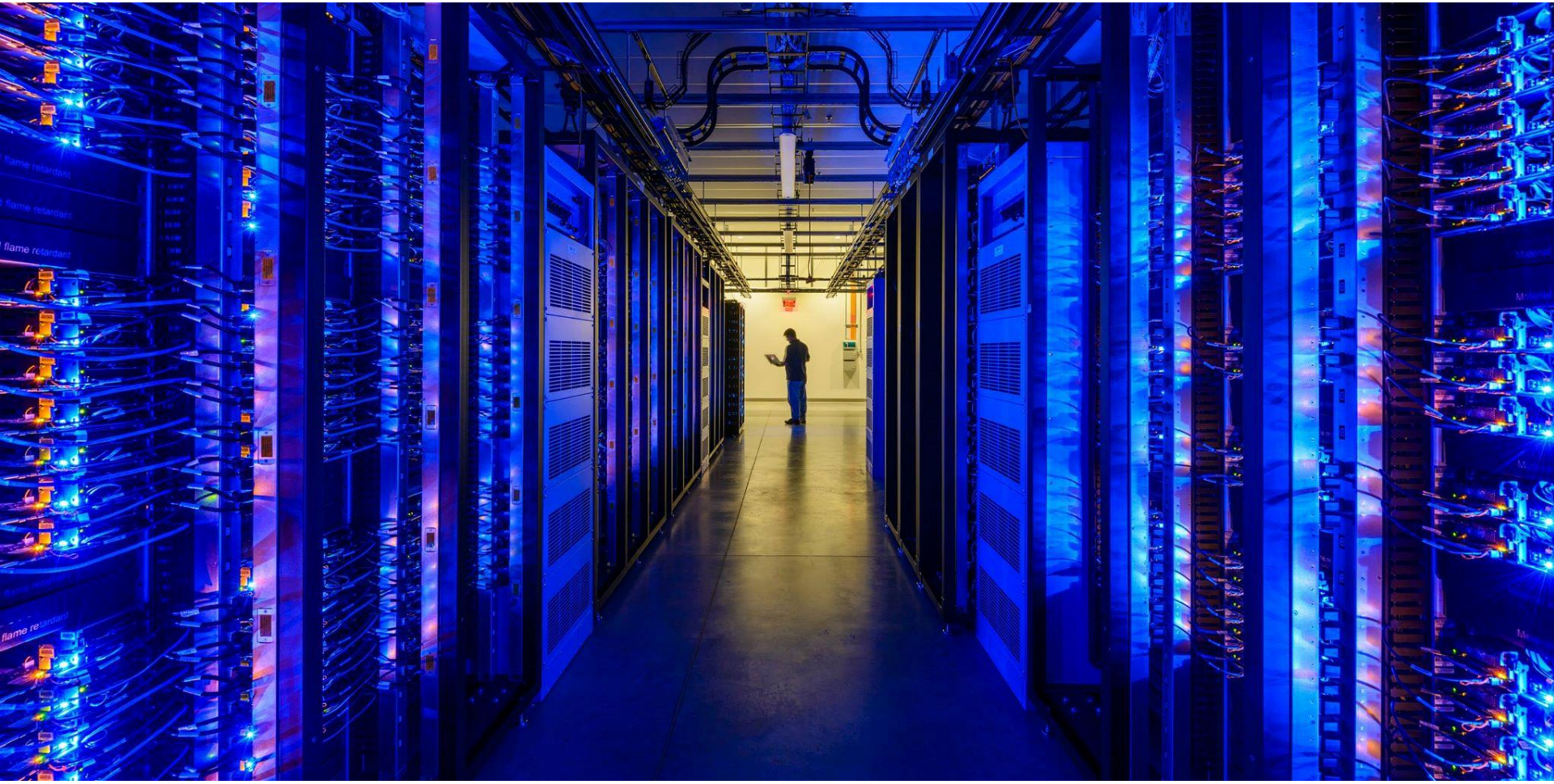


Tratamiento de datos masivos

Rafael Caballero Roldán – rafacr@ucm.es

Clases, evaluación, consultas

Estos apuntes de la asignatura Gestión de Datos Masivos no pretender sustituir las clases en aula, sino recoger ideas que sirvan de apoyo al curso

¿Dónde y cuándo?

Lunes en el laboratorio 9 de la Facultad de Informática, 19-21

Jueves, aula 1210 edif. Multiusos, 19-20j

¿Quién?

Rafael Caballero Roldán

Despacho 345

Facultad de Informática

UCM

Evaluación

Prácticas en el laboratorio: 80%

Presentación de un tema en clase 20%

Documentación

A través del campus virtual

https://cv.ucm.es

Tratamiento de datos masivos

Empezamos con un repaso histórico del papel que han jugado los datos desde los primeros ordenadores hasta el presente

Introducción a Big Data

Bases de datos NoSQL

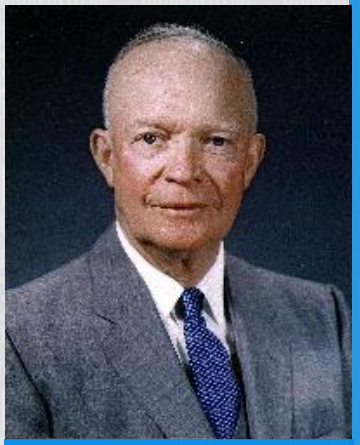
Arquitecturas Big Data

Análisis Científico de datos

Aprendizaje automático

Primeros Ordenadores

UNIVAC I fue adquirido por la oficina del Censo de los Estados Unidos en 1951, y lo siguió usando hasta 1963.
Por cierto que ni ENIAC ni UNIVAC son el primer “ordenador”; la primera máquina “Turing Completo” fue el Z3, de Konrad Zuse en 1941.

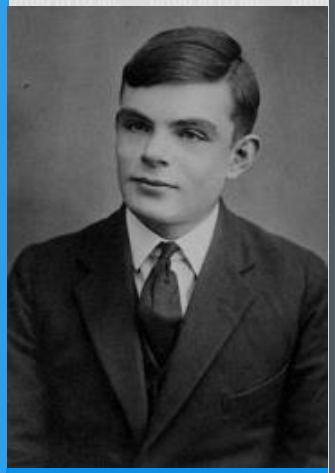


Dwight D. Eisenhower

Aplicaciones Civiles

Utilizados por su capacidad de gestionar datos

UNIVAC
Adquirido por la oficina del Censo de Estados Unidos



Alan Turing

Aplicaciones Militares

Utilizados por su capacidad de cómputo

- Cifrado: Colossus
- Balística: ENIAC

Los datos como sujeto pasivo

Durante los siguientes 40 años las bases de datos fueron ganando en popularidad, pero siempre considerándolas “recipientes” en los que guardar información, no posibles fuentes de nueva información

1. **¡Cero Papel!**
   Ideales para reemplazar grandes archivos en papel. Menos espacio y mayor rapidez de acceso

2. **Gestión de la empresa**
   Nóminas, contabilidad y todo tipo de tareas tediosas

3. Resúmenes de ventas, gastos, etc. Un anticipo de lo que estaba por venir

Business Intelligence

¿Qué es Big Data ?

Término más ambiguo cuanto más de moda, que habla de grandes cantidades de datos.

¿Cómo de grandes? ¿Basta con tener muchos datos para hablar de Big Data?

Introducción

Evolución de la Gestión de Datos

Definición de Big Data

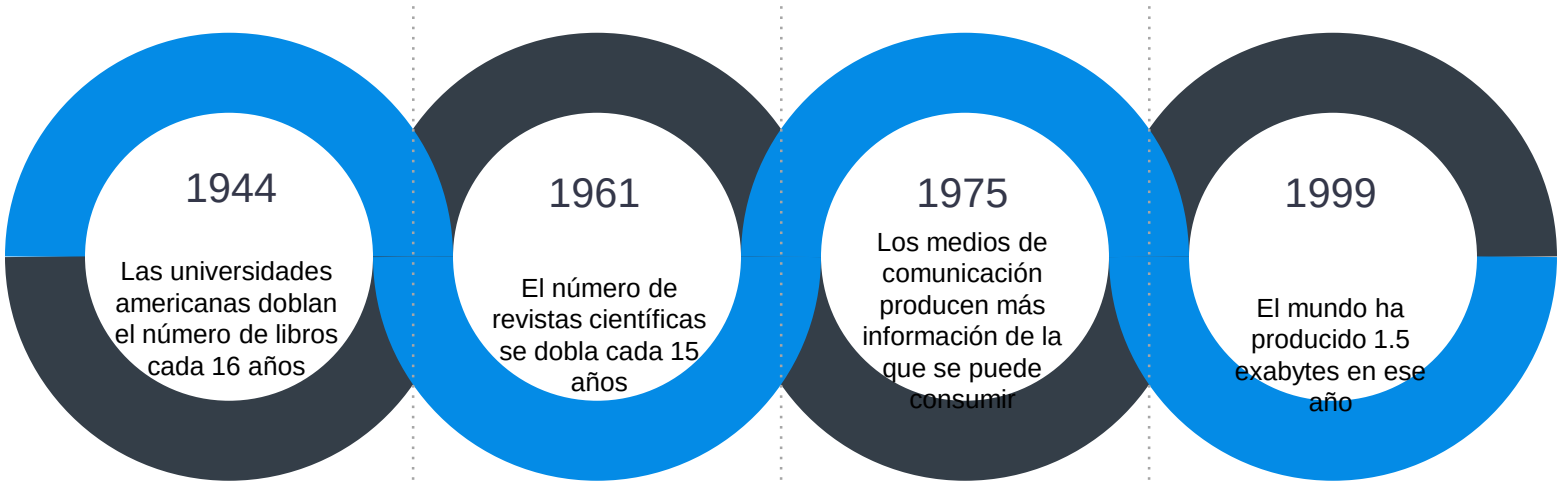
Arquitectura de Gestión

Rendimiento

Analítica Tradicional y Avanzada

¿Es realmente nuevo esto de Big Data?

El término Big Data es relativamente reciente, pero la preocupación por los datos viene de muy atrás



Freemont Rider bibliotecario

Derek Price

Informe del ministerio de correos y telecomunicaciones (Japón)

Peter Lyman y Hal R. Varian en "How Much Information?"

2017

1,259,413,690

Websites online right now

http://www.internetlivestats.com/total-number-of-websites/

Evolución en la gestión de los datos

Tiempos heroicos

Información almacenada en ficheros; lectura/escritura en bajo nivel. Bases de datos "artesanales"

Soporte: cintas, discos

Problemas de redundancia y calidad de datos; informes inconsistentes

Bases de datos jerárquicas (forma de árbol) → poco flexibles

Bases de datos relacionales

Datos vistos como tablas que se relacionan entre sí

Vista lógica y la vista física

Se evita la redundancia (normalización)

Bases de datos ACID (Atomicity, Consistency, Isolation, Durability)

Eficiencia: índices

Problema de impedancia;

Papel activo de los datos: almacenes de datos, OLAP,…

NoSQL

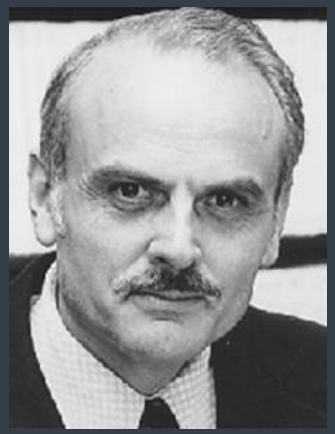
Dados sin estructura o semi-estructurados

Importancia de la distribución en clúster; vuelve a pensarse en la capa física

Muchos modelos alternativos

Orientados a Big Data

Se renuncia a ACID en favor de BASE (Basic Availability, Soft State, Eventual consistency)



Dr. Edgar F. Codd

1923-2003

Edgard F. Codd estudió Matemáticas y Químicas en Oxford. Piloto aéreo con la RAF durante la segunda guerra mundial, en 1948 marchó a Estados Unidos para ingresar en IBM, aunque marchó a Canadá en 1953 en desacuerdo con la “caza de brujas” del senador McCarthy. Regresó 10 años más tardes, doctorándose por la Universidad de Michigan con una tesis sobre autómatas celulares. Tras su regreso a IBM propuso las bases de lo que hoy es el modelo relacional. Sin embargo sus ideas no tuvieron mucho éxito. Las ideas expuestas en su artículo fueron rechazadas inicialmente tanto como publicación científica (hoy tiene casi 10000 citas) como por su empresa. Al final, un poco a regañadientes, IBM terminó creando un lenguaje inspirado en las ideas de Codd al que llamaron SEQUEL, y que fue pronto mejorado por Larry Ellison el creador de Oracle dando lugar a SQL

Definición de Big Data: Las 3Vs, 4Vs, 5Vs…

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. Doug Laney (Gartner)

Gran Volumen

Por grande queremos decir tan grandes como para necesitar de varios ordenadores (clúster). La idea de Big Data suele ir asociada a la de información distribuida

Gran Velocidad

En teoría la velocidad se refiere tanto a la escritura como a la lectura (consultas), pero en la práctica se piensa sobre todo en la llegada de datos

Gran variedad

Se piensa en datos heterogéneos, semi-estructurados, que no pueden ser tabulados con facilidad

Veracidad y/o Valor

A menudo se añade la “veracidad” indicando la “falta de veracidad”, es decir que algunos de los datos pueden ser poco fiables. El Valor (la 5ª V) se refiere a que debe tratarse de datos valiosos

Si estas definiciones no nos gustan podemos ver estas 12 definiciones en Forbes

¿Cuáles son los puntos débiles del modelo relacional?

Dificultad para manejar enormes cantidades de datos (Big Data)

Dificultad para manejar datos heterogéneos originados en Internet, IoT

Escalado Vertical y Horizontal

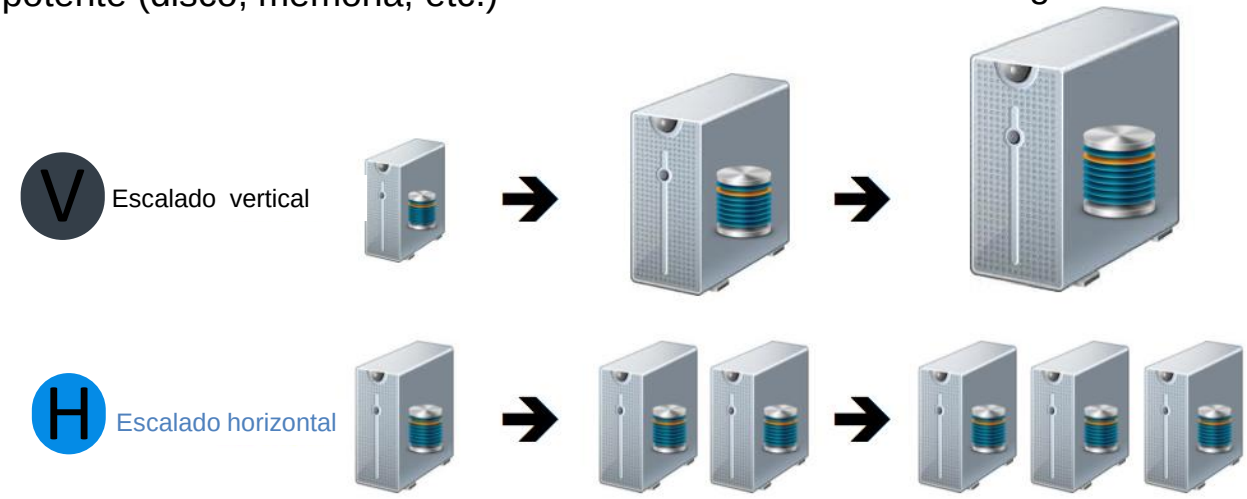
Cuando superan las posibilidades del sistema tenemos dos posibilidades

V Escalado vertical

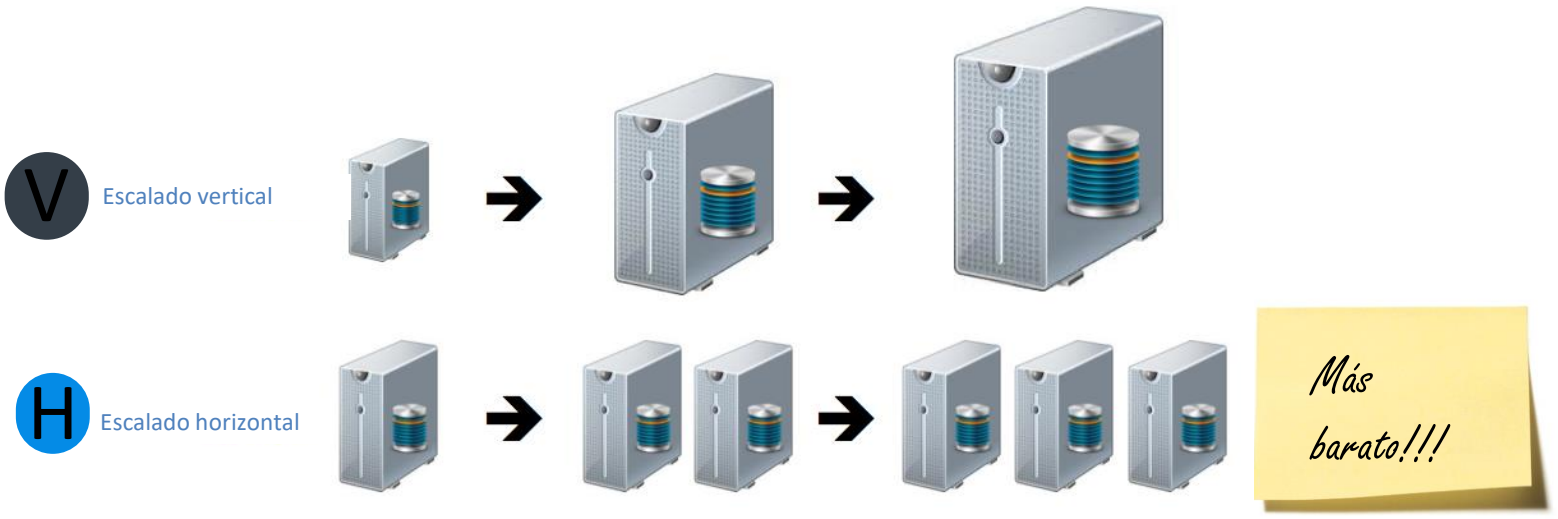
Se sustituye el equipo por otro mayor, más potente (disco, memoria, etc.)

H Escalado horizontal

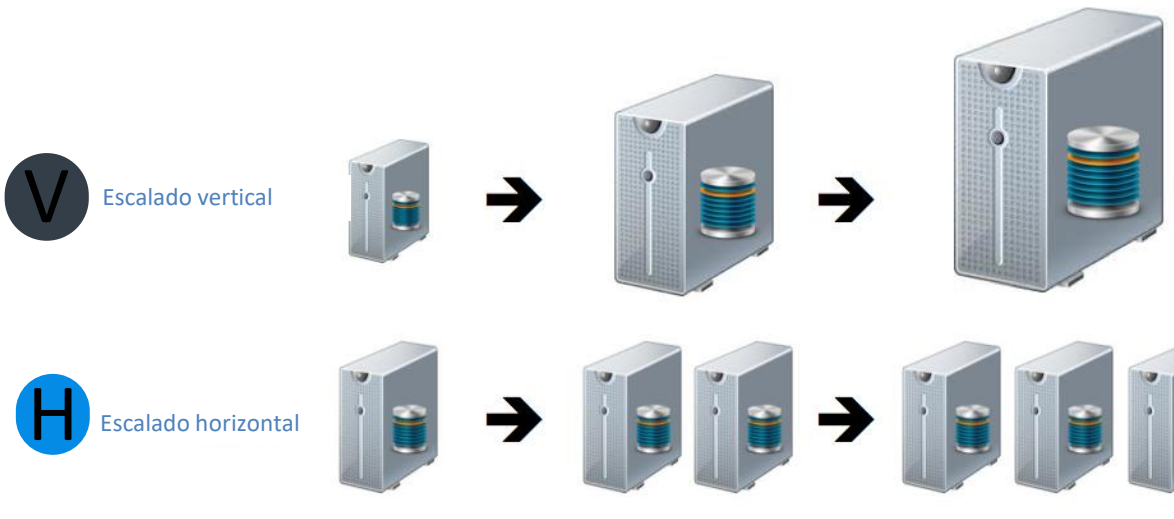
Se crea un clúster de equipos de gama baja, añadiendo más según demanda



Fin del reinado relacional: Escalado Vertical y Horizontal

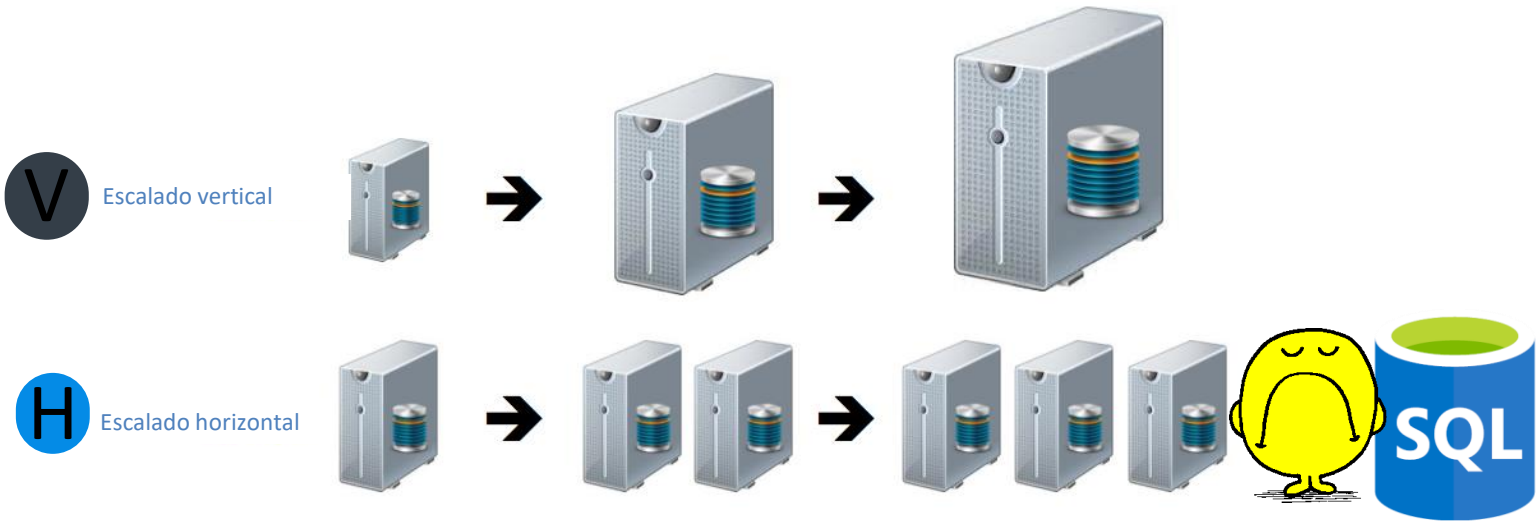


Fin del reinado relacional: Escalado Vertical y Horizontal



...pero poco adecuado para BBDD relacionales

Fin del reinado relacional: Escalado Vertical y Horizontal



Escalado Horizontal: pros y contras

PROS

1. Más económico
   Adquirir 10 servidores de tamaño medio resulta más barato que adquirir un servidor de grandes prestaciones
2. Sharding
   Si se organizan bien los datos se pueden hacer consultas solo a los servidores que tienen la información buscada (partición/sharding)
3. Paralelización
   Se puede enviar la misma consulta a todos los servidores a la vez y ahorrar tiempo (map/reduce)

CONTRAS

1. Réplicas; no tan económico
   Al tener muchos servidores aumenta la probabilidad de fallos (caídas). Para evitar la pérdida de datos se necesita crear servidores que mantienen réplicas de los datos
2. Joins
   Las consultas complejas y en particular las que implican joins pueden ser más lentas debido al tráfico de datos entre servidores. Esto se soluciona parcialmente mediante la desnormalización típica de los entornos semiestructurados
3. Inconsistencia
   El uso de réplicas en los clúster dificulta lograr la consistencia habitual en el modelo relacional. La consistencia se puede lograr a base de incrementar el tiempo de lectura/escritura.
   También se complica el concepto de atomicidad de las transacciones

Escalando los análisis

El cambio de arquitectura impulsado por Big Data conlleva cambios en los análisis de datos

Paralelización	Nuevos algoritmos	Cerca de los datos	Nuevos lenguajes
+	+	+	+
Muchos algoritmos necesitan ser refactorizados para tener en cuenta las nuevas arquitecturas y no perder eficiencia	Big Data genera nuevas posibilidades y requiere nuevos algoritmos	Con estas cantidades de datos los almacenes de datos pierden valor como entidades separadas	Lenguajes como Python incluyen ahora gran cantidad de librerías para tratar con Big Data
¿Cómo hacer regresión en Hadoop?	¿Cómo analizar los datos de los sensores de un coche?	¿Cómo hace análisis Google? ¿Se lleva los datos a un almacén?	¿Big Data en Python?

Datos heterogéneos

La diversidad en las fuentes de datos y los nuevos tipos de análisis generan nuevas posibilidades y demandan nuevas formas de tratar los datos

Dispositivos

IoT: datos constantes que desde su inicio deben ser pensados en relación con Big Data

Redes sociales

Acceso a información que permite detectar el gusto de los consumidores.

Más aún, el poder agrupar gente con gustos comunes permite análisis hasta ahora impensables

Predicción de stock

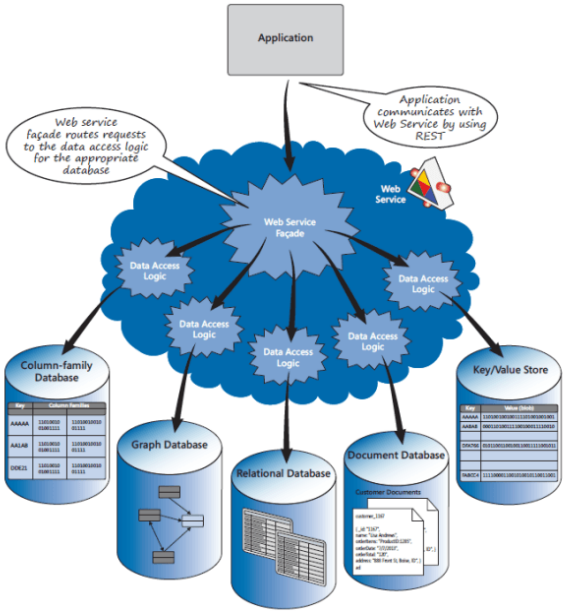
La gestión de almacenes de una gran cadena se reduce al mínimo si se es capaz de predecir en detalle qué necesidades tendrá cada tienda en cada momento

Detección de fraude

Los defraudadores, tanto bancarios, como de seguros o simples timadores, tienden a seguir patrones, que pueden ser descubiertos mediante bases de datos orientadas a grafos

Persistencia políglota

Distintos problemas exigen herramientas diversas; es importante saber qué se planea hacer con los datos y estar preparado para disponer de varios sistemas complementarios



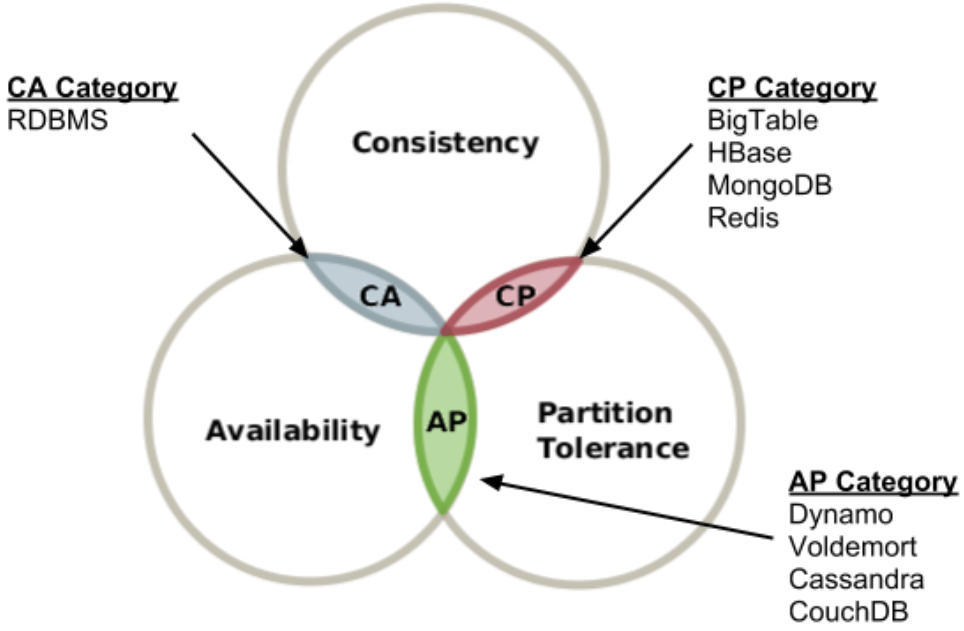
http://www.jamesserra.com/archive/2015/07/what-is-polyglot-persistence/


Lugares de Madrid más visitados
por turistas chinos
08:45
Llegada de turismo extranjero al aeropuerto de Madrid
y su impacto en la ciudad

Teorema CAP

Enunciado como conjetura por Brewer en el año 2000. Dice que en un sistema solo se pueden dar en un momento dado 2 de las siguientes 3 propiedades

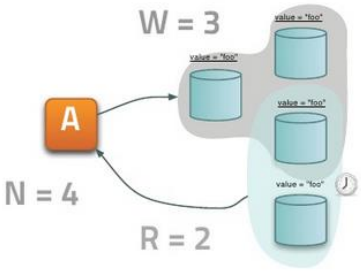
- Consistencia (*Consistency*): Información consistente en todos los nodos al acabar una escritura
- Disponibilidad (*Availability*): toda la información del sistema está siempre disponible (no hay “vuelva Vd. mañana”)
- Tolerancia de las Particiones (*Partition Tolerance*): las diferentes partes del sistema distribuido (los nodos) continuarán funcionando normalmente aunque la comunicación entre ellos se vea interrumpida



Escalado Horizontal: consistencia

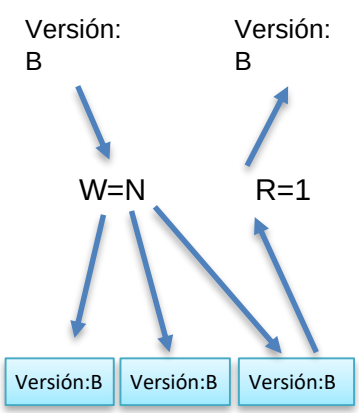
Los nodos del clúster son en realidad conjuntos de réplica con N servidores cada uno.

- N servidores tienen la misma información (salvo latencia)
- Llamamos W al número de servidores que deben contestar para asegurar una escritura
- Llamamos R al número de servidores que deben contestar para asegurar una lectura



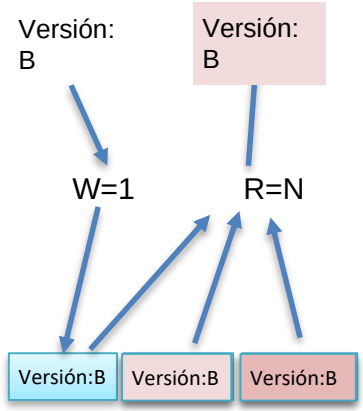
Consistencia fuerte
W+R>N

Se asegura que se lee lo mismo que se ha escrito, una vez que se ha confirmado la lectura



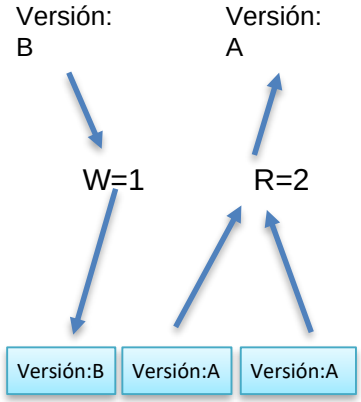
Consistencia fuerte por escritura W=N, R=1

Variante en el que lo que se retrasa es la escritura



Consistencia fuerte por lectura W=1,R=N

Escritura rápida, pero consulta lenta



Consistencia Eventual
W+R<N

Lectura y escritura rápida, pero se pierde la consistencia

Escalando los análisis

El cambio de arquitectura impulsado por Big Data conlleva cambios en los análisis de datos

Paralelización

+

Muchos algoritmos necesitan ser refactorizados para tener en cuenta las nuevas arquitecturas y no perder eficiencia

¿Cómo hacer regresión en Hadoop?

Nuevos algoritmos

+

Big Data genera nuevas posibilidades y requiere nuevos algoritmos

¿Cómo analizar los datos de los sensores de un coche?

Cerca de los datos

+

Con estas cantidades de datos los almacenes de datos pierden valor como entidades separadas

¿Cómo hace análisis Google?

¿Se lleva los datos a un almacén?

Nuevos lenguajes

+

Lenguajes como Python incluyen ahora gran cantidad de librerías para tratar con Big Data

¿Big Data en Python?