Problemática que plantea Big Data

Módulo 1 – Lección 4

ibigdata.es

Universitat d'Alacant
Universidad de Alicante

Agenda

- Introducción a la problemática del Big Data
- Problemas
  - Volumen
  - Integración de las fuentes
  - Verificación de la calidad
  - Gestión de datos Streaming
  - Otros problemas
- Resumen del módulo 1

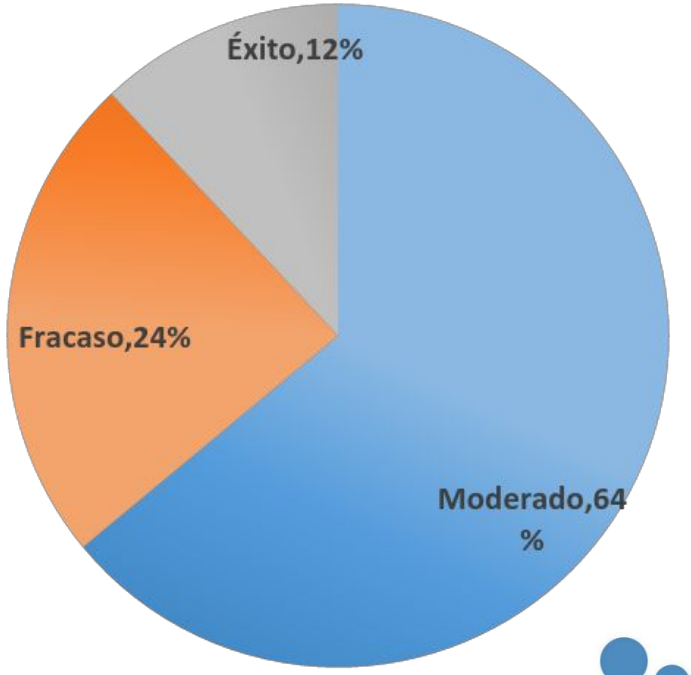
Introducción a la problemática del Big Data

Big Data proporciona un gran abanico de posibilidades a las organizaciones pero ...

Las 5v’s son un problema en sí mismas

Introducción a la problemática del Big Data

Encuesta reciente a empresas que usan Big Data muestra que:



- Solo un 12% afirman tener gran éxito en su uso
- Frente a un 64% de éxito moderado y un 24% de fracaso

Fuente: tdwi THE DATA WAREHOUSING INSTITUTE

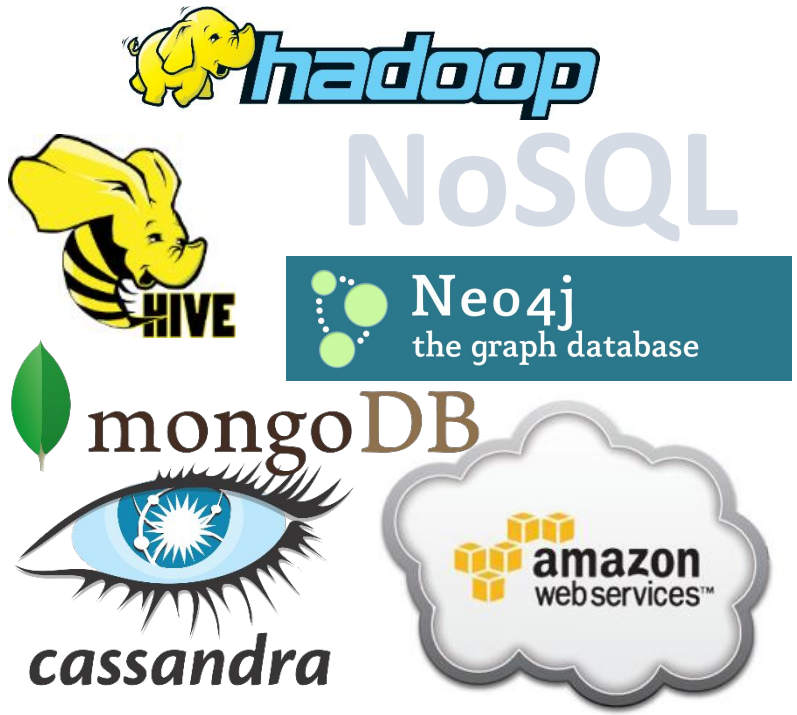
Introducción a la problemática del Big Data

El fracaso se debe a problemas como:

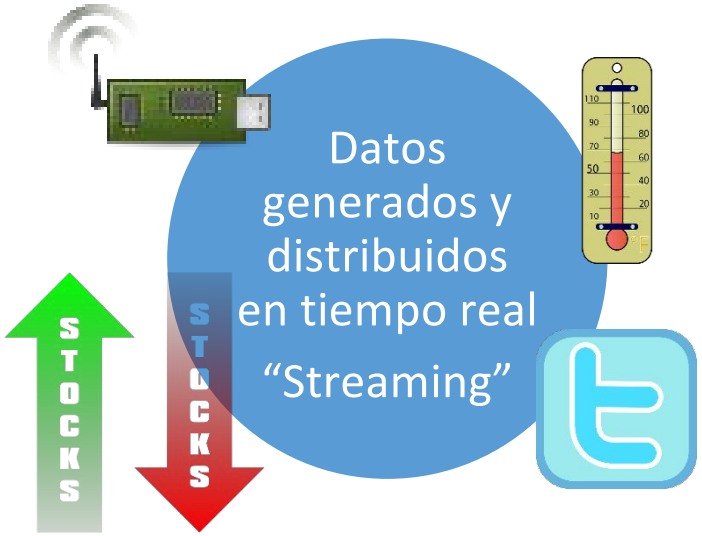
- Complejidad de la **integración** de las fuentes de datos
- Pobre **calidad** de los datos
- Gestión de datos generados en **tiempo real**
- Falta de **personal** con las habilidades adecuadas
- Elección de la **arquitectura** incorrecta

Problemas. Volumen

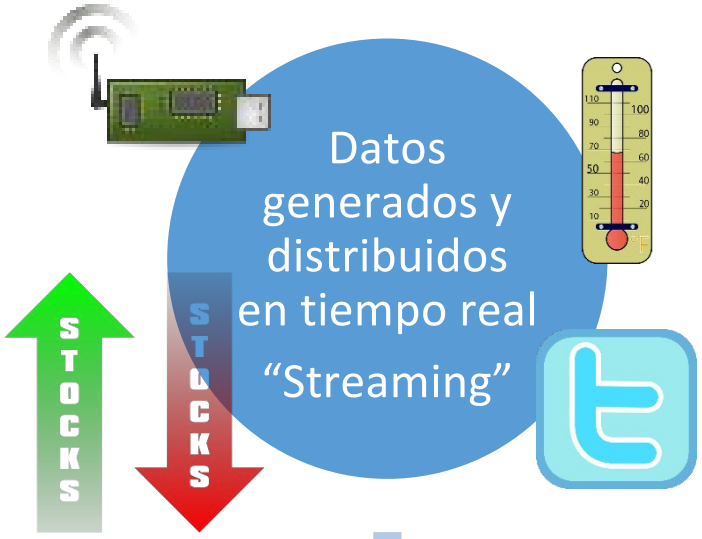
- El procesamiento y análisis de los enormes **volúmenes** es uno de los problemas más evidentes y antiguos.
- Sin embargo, la tecnología actual aporta soluciones como Apache Hadoop y las bases de datos NoSQL
  - De **bajo coste y escalables** en volumen
  - Procesamiento y análisis de **terabytes de datos en minutos o segundos** sobre hardware comercial



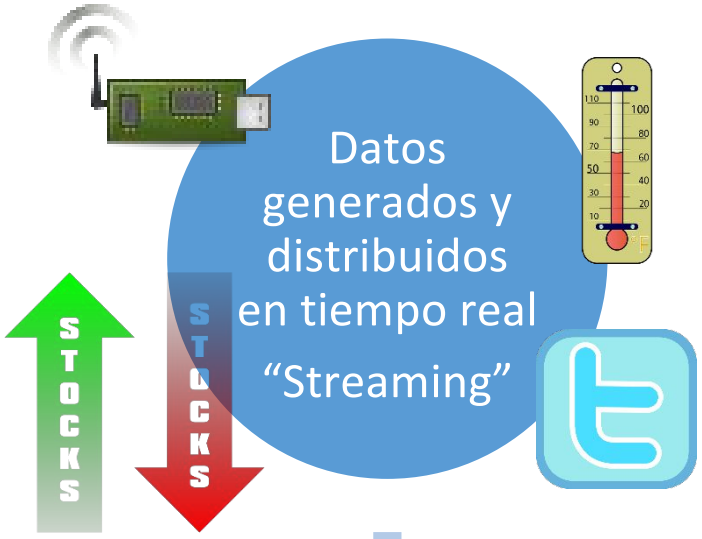
Problemas. Gestión de datos Streaming



Problemas. Gestión de datos Streaming



Problemas. Gestión de datos Streaming



Datos generados y distribuidos en tiempo real

"Streaming"

Oportunidad de Análisis. "Real Time"

Retos

i) Integración

ii) Verificación de la calidad

* En tiempo real

Problemas. Integración de las fuentes

- Combinación o **integración** de fuentes de datos, internas y externas
  - Una de las **formas de añadir valor** a los datos originales y prepararlos para el análisis

Problemas. Integración de las fuentes

- Combinación o **integración** de fuentes de datos, internas y externas
  - Una de las **formas de añadir valor** a los datos originales y prepararlos para el análisis
- La **Variedad** o heterogeneidad de las fuentes, hace que esta tarea requiera un **gran esfuerzo humano**
  - Sobre todo cuando se maneja un gran número de fuentes
  - A tener en cuenta: Distintos modelos de datos, distinto formato, falta o inexistencia de metadatos...

Problemas. Verificación de la calidad

- Es muy difícil comprobar la **Veracidad o precisión de los datos** contenidos en las fuentes externas
  - Su generación no depende de nosotros
  - **Falta de datos**, ruido, alteraciones…

Indicador	1995	1996	1997	1998
Licencias ambientales únicas vigentes	119	12	2	
Árboles plantados	2010177	122789	1842636	3747566
Superficie reforestada	523	709	695	4154
Licencias ambientales únicas vigentes		9	2	
Árboles plantados	492082	21749		
Superficie reforestada	75	42		
Licencias ambientales únicas vigentes		0	0	
Árboles plantados	250418	40032		
Superficie reforestada	17	60		

Problemas. Verificación de la calidad

- Es muy difícil comprobar la **Veracidad o precisión de los datos** contenidos en las fuentes externas
  - Su generación no depende de nosotros
  - **Falta de datos**, ruido, alteraciones…

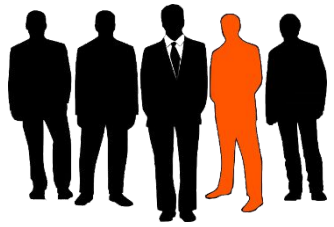
Indicador	1995	1996	1997	1998
Licencias ambientales únicas vigentes	119	12	2	
Árboles plantados	2010177	122789	1842636	3747566
Superficie reforestada	523	709	695	4154
Licencias ambientales únicas vigentes		9	2	
Árboles plantados	492082	21749		
Superficie reforestada	75	42		
Licencias ambientales únicas vigentes		0	0	
Árboles plantados	250418	40032		
Superficie reforestada	17	60		

- Es necesario garantizar la calidad de los datos antes y después de los procesos de integración

Otros problemas

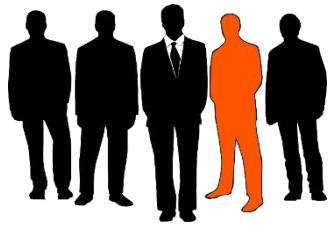
- La falta de personal con las **habilidades** adecuadas

Otros problemas



- La falta de personal con las **habilidades** adecuadas
- Selección de la **arquitectura** idónea
  - ¿Qué base datos NoSQL es la más adecuada?
  - ¿Clúster local o uso de servicios y almacenamiento en la nube?

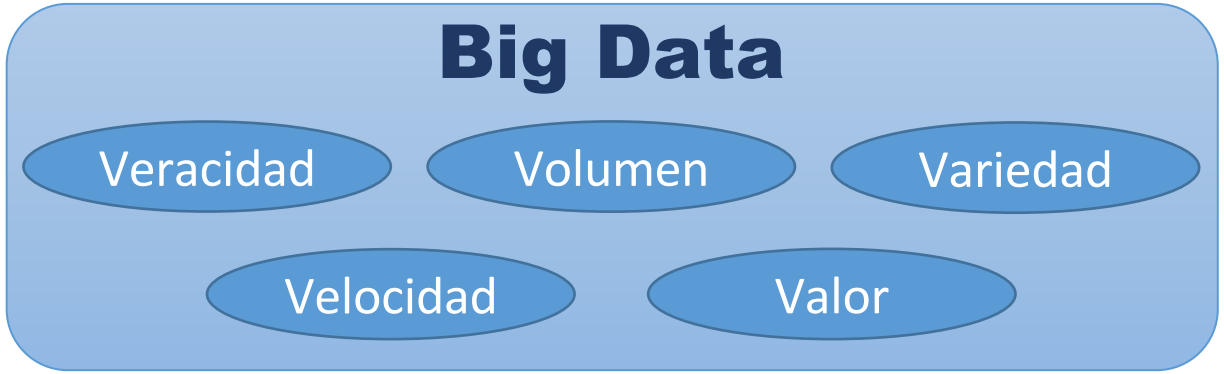
Otros problemas



- La falta de personal con las **habilidades** adecuadas
- Selección de la **arquitectura** idónea
  - ¿Qué base datos NoSQL es la más adecuada?
  - ¿Clúster local o uso de servicios y almacenamiento en la nube?
- Otros problemas
  - Coexistencia con Almacenes de Datos (Data Warehouses)
  - Coste de implementación y mantenimiento
  - Pobre integración entre herramientas Big Data existentes

Resumen del módulo 1

- Lección 1.1: Contexto donde surge Big Data: Dispositivos y aplicaciones



- Lección 1.2: Definición de Big Data
- Lección 1.3: Aplicaciones analíticas
- **Lección 1.4: Principales problemas de las 5v's**


problem
analysis
solution

Contacto

- Dudas, comentarios y noticias acerca del curso
  - Email: lucentialab@gmail.es
  - Twitter: @lucentialab
  - Página web del curso: http://ibigdata.es
- Otras webs
  - Grupo de investigación Lucentia http://www.lucentia.es/
  - Canal de YouTube https://www.youtube.com/user/lucentialab
  - Plataforma UniMOOC http://unimooc.com/