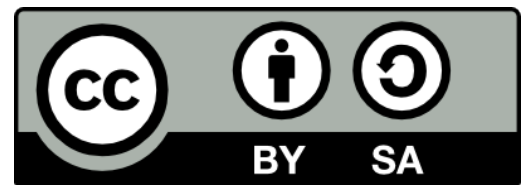
Big Data I:

Ingeniería de datos

Felipe Ortega
Dpto. de Estadística e Investigación Operativa
Universidad Rey Juan Carlos

March 3, 2015

Universidad
Rey Juan Carlos



Big data: problemas y soluciones tecnológicas

Mejor definición de big data hasta la fecha...

Dan Ariely

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

6 de enero de 2013 a la(s) 8:02 ·

Dimensiones big data

- Término controvertido, incluso para los propios profesionales.
- Consenso: definido por las 3 “Vs” [2-3].
  - **Volumen** (tamaño, procesamiento).
  - **Velocidad** (adquisición, procesamiento).
  - **Variedad** (dimensiones).
- A veces, se añade más factores (V’s):
  - *veracidad* (integridad de datos, corrección...)
  - *valor* (el valor añadido que aporta big data para el negocio o dominio de aplicación).
  - *variabilidad*, *visualización*, etc.
- Lo importante: no es sólo una cuestión de tamaño.

¿Cuántos son “muchos datos”?

- Típicamente, más de los que podamos procesar en un sólo computador (incluso en un servidor muy potente).
  - Por necesitar demasiada memoria.
  - Por requerir demasiado espacio de almacenamiento.
  - Porque no podemos almacenar el flujo de datos que nos llega de forma permanente (procesado *streaming* vs. *batch*).
  - Porque necesitamos resultados con gran rapidez para tomar decisiones operativas.

- A continuación, presentamos algunos ejemplos [4].

Algunos números sobre big data

- **Walmart**.
  - Fortune 500 Global.
  - Mayor empleador privado del mundo (+2 millones empleados).
  - Mayor distribuidor minorista del mundo.

- Sus servidores procesan más de un millón de transacciones de clientes cada hora.

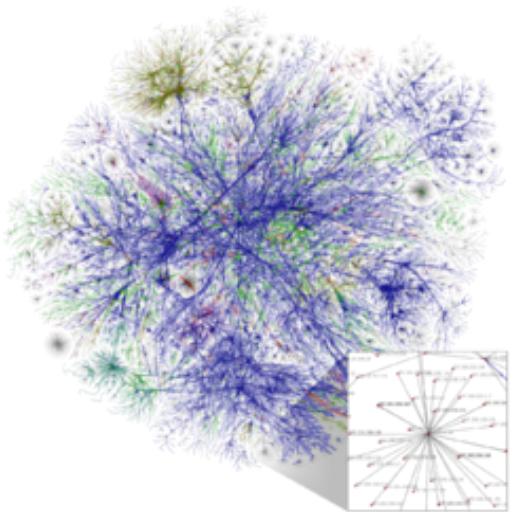
- Sus bases de datos almacenan más de 2,5 Petabytes (1 Petabyte = 1024 Terabytes).

Algunos números sobre big data

- **LHC (CERN)**.
  - Mayor y más potente colisionador de partículas del mundo.
  - Una de las mayores fuentes de datos de experimentos científicos del mundo.
- Se estima que genera unos 15 Petabytes de información anualmente.
- Se analizan en un sistema computacional distribuído y tolerante a fallos (grid computing):
  - 170 centros de computación,
  - 36 países participantes.
  - Red global de comunicación.

Algunos números sobre big data

- **Datos en la Web**.
- **Facebook** opera sobre 500 Terabytes de información de registro de actividad de sus usuarios, y sobre cientos de Terabytes de imágenes.
- Cada minuto se cargan 100 horas de vídeo en **Youtube**, y más de 135.000 horas de vídeo son vistas.
- **Twitter** sirve a casi 600 millones de usuarios que generan 9.100 tweets cada segundo.
- Los sistemas de **eBay** procesan más de 100 Petabytes de información al día.



Algunos números sobre big data

- **Sector aeronáutico**.
- Un avión comercial de Boeing puede generar alrededor de 10 Terabytes de información operacional cada 30 minutos de funcionamiento.
- Por tanto, en un vuelo transatlántico se pueden llegar a generar varios cientos de Terabytes de información.
- Se realizan alrededor de 22.000 vuelos diarios en todo el mundo.
- Esto nos ofrece una idea de la ingente cantidad de datos generada por máquinas y redes de sensores de manera regular.

Necesidades computacionales

- Precisamos potencia y capacidad de computación para ingeniería de datos.
- Problema: el tráfico de datos crece a mayor velocidad que nuestra capacidad de computación.
  - (2002-2009): volúmen global del tráfico de datos se multiplicó por 56; potencia de computación se multiplicó sólo por 16.
  - (1998-2005): centros de datos crecieron en tamaño un 173% anual [4], mientras que la eficiencia en consumo energético no mejoró a la par.
  - Esto generará una enorme *huella de consumo energético* para análisis de datos.
  - 50% de los centros de cómputo de datos (aprox.) solo funcionan al 50% de su rendimiento máximo.

Tipos de datos según su estructura

- **Datos estructurados**: Tienen una serie de campos con significado predefinido. Cada campo está asociado a un tipo de datos (numérico, textual, doble precisión, objeto serializado...). Ejemplo: RDBMS.

- **Datos semi-estructurados**: Se representan mediante un formato de codificación que aporta cierta estructura e información sobre los datos (metadatos). Sin embargo, su contenido (número de campos, formato de cada campo, etc.) puede ser muy variado. Ej: documentos XML.

- **Datos no estructurados**: El formato de los datos no está claramente definido de forma previa. Pueden aparecer mezclados datos numéricos, textuales o multimedia, y en un orden imprevisible.

Soluciones para datos no estructurados

- Necesitamos tecnologías y métodos flexibles para gestionar este tipo de fuentes de datos (necesidades dinámicas e imprevisibles). Ejemplo: **tecnologías NoSQL**.
  - Ejemplo: Esquemas **clave-valor**.
  - Almacenan duplas (*clave-valor*), donde las claves asociadas a cada valor son únicas (para acelerar las búsquedas) y los valores pueden ser también objetos complejos (tales como listas, tablas hash, etc).
  - Ejemplo: Bases de datos **documentales**.
  - Almacenan documentos representados en cierto formato de condificación (tales como XML, JSON o YAML).
  - También siguen un esquema de almacenamiento *clave-valor*, pero el contenido de los documentos es arbitrario, y además se ofrecen mecanismos para realizar búsquedas basadas en dichos contenidos (utilizando los metadatos del sistema de condificación).

Tipos de procesamiento de datos

- Clasificación de procesamiento de datos según requisitos de interacción:

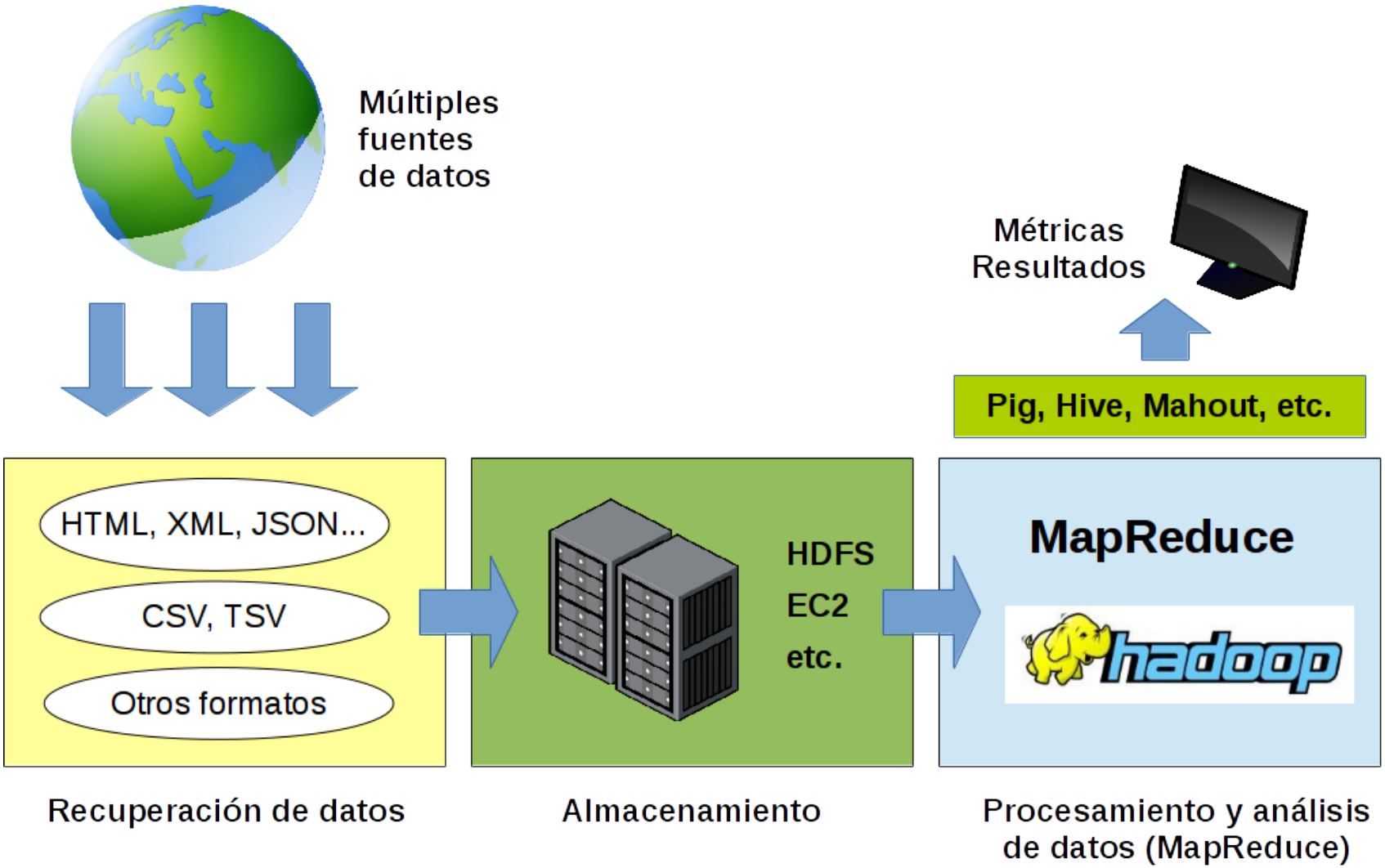
1. Procesamiento **batch** (también llamado *offline*): No existen requisitos estrictos en cuanto al tiempo que podemos emplear en la preparación, transformación y computación de los datos almacenados. Ejemplo: MapReduce (Hadoop).
2. Procesamiento de **flujos de datos** (*streaming*, también llamado *online*): Existen requisitos estrictos sobre el tiempo máximo que podemos emplear para preparar, trasformar y procesar los datos. Puede deberse a varias razones:
   - Análisis interactivo.
   - Interacción con usuarios finales (servicios, dashboards, etc.).
   - Excesiva velocidad o volumen de datos (no podemos almacenar localmente).

Tipos de procesamiento de datos

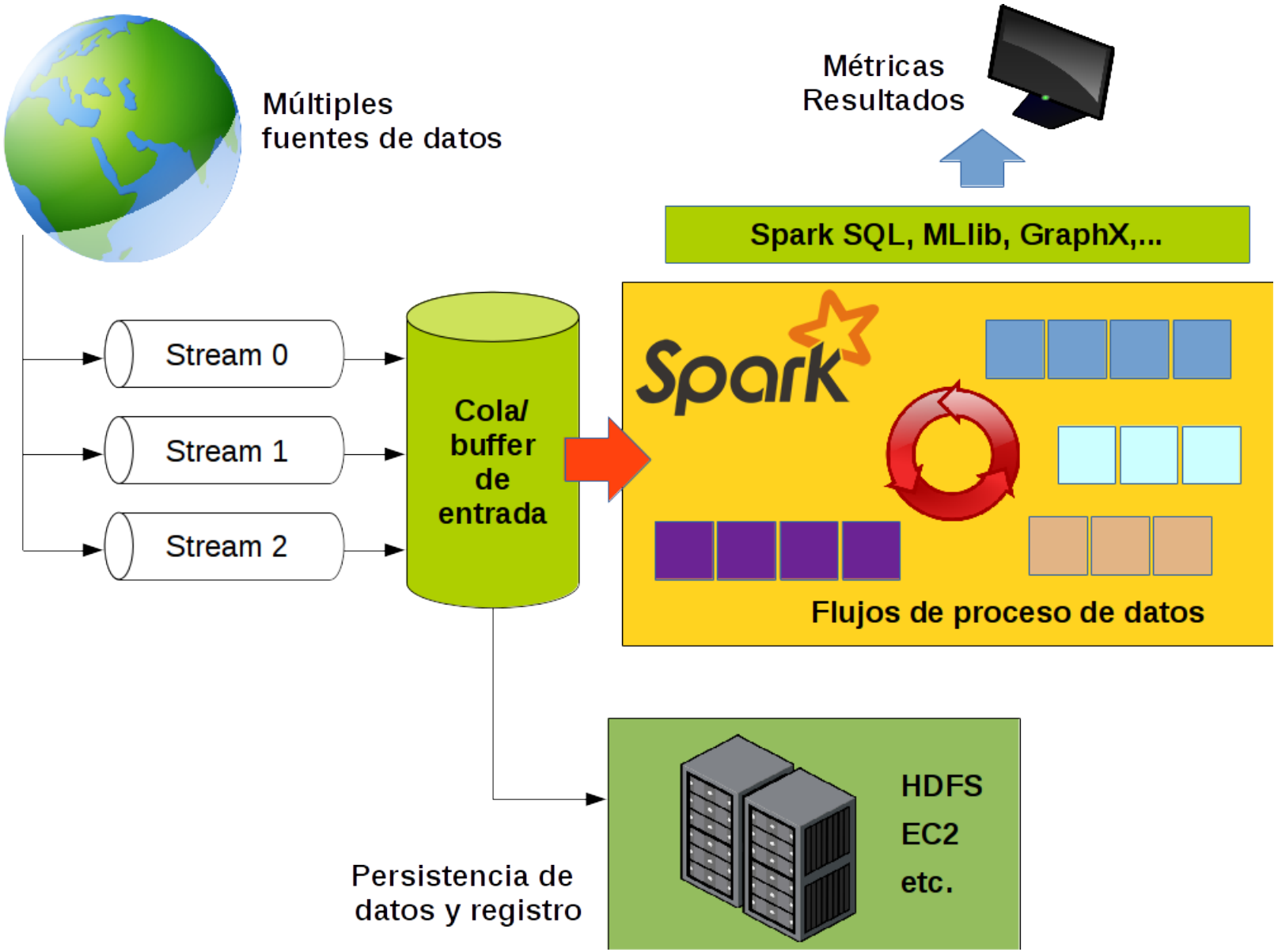
Fig. – Procesamiento *streaming*

Fig. – Procesamiento *batch*

Esquema procesamiento *batch*



Esquema procesamiento *streaming*



Tendencias procesamiento de datos

- El procesamiento **streaming se está imponiendo** rápidamente.
- MapReduce no es suficientemente flexible ni rápido para muchos problemas de análisis de datos.
  - Junio 2014: Google declara que dejaron de usar MapReduce hace años.
- MapReduce no es adecuado para muchos modelos de análisis de datos, que incluyen operaciones iterativas.
  - Exigen un importante esfuerzo para programar estos procesos de modo que se reduzca el número de pasadas sobre los datos (cada iteración es muy cara).
- Por contra, los sistemas de procesado *streaming* se pueden adaptar a muchos más tipos de análisis, han sido concecibos para ser rápidos y escalables.

Tendencias procesamiento de datos

- En procesado *streaming* se crean flujos de **datos inmutables**, que se procesan o transforman para generar nuevos flujos de datos. Se puede añadir cierta *persistencia*.
- También es posible combinar *streaming* con procesado *batch*, (la llamada **arquitectura lambda**) pero cuidado con la duplicidad de trabajo.
- Pero exige ciertos requisitos adicionales:
  - Sistemas de colas de mensajes / buffer de entrada que almacenen temporalmente los datos hasta que entren al flujo de procesado (idealmente sin péridas).
  - Incluir sistemas automáticos de distribución de carga y tolerancia ante fallos de nodos de procesamiento.

Bibliografía

1. Provost, F., Fawcett, T. Data Science for Business. O'Reilly Media Inc. Julio 2013.
2. Cathy O'Neil, Rachel Schutt. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media Inc. Octubre 2013.
3. Doug Laney. 3d Data management: controlling data volume, velocity and variety. Appl. Delivery Strategies Meta Group (949)(2001).
4. Kambatla, K. et al. Trends in big data analytics. Journal of Parallel and Distributed Computing (in press). Elsevier. Enero 2014.

Créditos

1. Imagen Walmart-exterior.jpg por see. CC-BY-SA-3.0, via Wikimedia Commons.
2. Imagen inside-CERN-LHC por Juhanson. CC-BY-SA-3.0, via Wikimedia Commons.
3. Imagen Internet map por The Opte Project. CC-BY-2.5 , via Wikimedia Commons.
4. Imagen Boeing Emirates por Faisal Akram desde Dhaka, Bangladesh. CC-BY-SA-2.0, via Wikimedia Commons
5. Imágenes clipart obtenidas de Openclipart, todas ellas disponibles en dominio público.
6. Todos los logos de proyectos y/o empresas son marcas registradas, utilizados simplemente con fines ilustrativos.

Contacto

e-mail: felipe.ortega@urjc.es

Twitter: @jfelipe