Lección 3.1 Definición y relevancia:

Ya en el módulo anterior vimos que una de las partes más importantes de un sistema de BI son las fuentes de datos. En este módulo revisaremos en profundidad el concepto de fuente de datos y veremos cómo afecta el nuevo contexto de Big Data a la naturaleza de las mismas.

1. Definición

En primer lugar, ¿Qué es una fuente de datos?

- En BI, una fuente de datos es un punto de abastecimiento de datos con información potencialmente útil para el análisis de un proceso de negocio de nuestra organización
- Los datos pueden proporcionarse de distintas formas y en una gran variedad de formatos…(por ejemplo)
    - Conjuntos de datos (archivos, consultas a bases de datos,...) o generados de forma continua (datos procedentes de sensores)
    - Formatos: CSV, XML, JSON, texto, vídeo, audio, imagen, relacional (conjuntos de filas y columnas),...
- Cuando hablamos de fuentes de datos nos referimos a información digital o que es digitalizada para su procesamiento.

2. Fuentes de datos Big Data

- Una de las ventajas del uso de Big Data es que permite enriquecer la información interna disponible en una organización con información de fuentes externas
- De esta forma, en Big Data es frecuente disponer de:
    - Información interna
        - Data Warehouses o repositorios de tipo Big Data (ej. BD NoSQL) con información interna de la actividad de la organización recopilada en el tiempo
        - No solo tenemos información estructurada en forma de filas y columnas: texto, imagen, vídeo, XML, JSON...
    - Información externa
        - Proporcionada por otras empresas de la competencia mediante conexión privada (ej. en caso de sinergias)

- Proporcionada por otras organizaciones a través de internet, ya sea de forma pública o comercial.

3. Fuentes de datos Big Data

Precisamente, este último tipo de fuentes ha sido el que más aumentado en los últimos años y las responsables (en gran medida) del nuevo de contexto de Big Data (como ya vimos en el Módulo 1)

- Amplio abanico de fuentes externas disponibles a través de la red
  - Open Data, Redes Sociales, Internet of the Things (IoT)....
- En Big Data el objetivo es aprovechar este tipo de fuentes para complementar el resto de información de la que disponemos
- Dos de las más usadas para el desarrollo de aplicaciones Big Data para BI son:
  - Open Data
  - Redes Sociales

4. Fuentes externas Big Data. Open Data

Con una gran acogida en los últimos tiempos tenemos el….

- Fenómeno **Open Data**: Muchas instituciones y comunidades han decidido publicar y compartir en Internet la información que manejan.
  a. datos de PIB de países, encuestas, consumo energético, precios de energía, paradas de autobús de una ciudad y otros servicios, restaurantes y muchos más.
- Es habitual que la información...
  a. Se proporcione en distintos formatos (ej. Excel, XML, JSON)
  b. Sea accesible a través de API's que nos permiten consultar y recibir los datos en nuestras aplicaciones
  c. Soporte para lenguajes de consulta estándares como SPARQL
- Algunos ejemplos:
  a. Catálogo de Información Pública del Sector Público http://datos.gob.es/
  b. Santander Datos Abiertos http://datos.santander.es/

DATOS | VISTAS | VERSIONES

Mostrar elementos — Mostrando 1000 de 32545 registros (Limitado a 1000 ?)

Arrastre aquí una columna para agrupar por dicha columna

1 2 3 4 5 6 7 8 9 10 ...

id	ayto:fechaRegistro	ayto:fechaRegistroMes	ayto:asunto	ayto:fechaRegistroAnyo	ayto:destino	ayto:codAsunto	dc:modified	dc:i
0	2003-01-02	1	TRAMITACION ABREVIADA	2.003	OBRAS	20	2013-12-13T03:37:02.741Z	200.3
1	2003-01-02	1	LICENCIA OBRA MAYOR NUEVA PLANTA	2.003	OBRAS	34	2013-12-13T03:37:02.741Z	200.3
2	2003-01-02	1	TRAMITACION ABREVIADA	2.003	OBRAS	20	2013-12-13T03:37:02.741Z	200.3
3	2003-01-02	1	TRAMITACION ABREVIADA	2.003	OBRAS	20	2013-12-13T03:37:02.741Z	200.3
4	2003-01-02	1	TRAMITACION ABREVIADA	2.003	OBRAS	20	2013-12-13T03:37:02.741Z	200.3
5	2003-01-07	1	TRAMITACION ABREVIADA	2.003	OBRAS	20	2013-12-13T03:37:02.737Z	200.3
6	2003-01-07	1	TRAMITACION ABREVIADA	2.003	OBRAS	20	2013-12-13T03:37:02.737Z	200.3
7	2003-01-07	1	LICENCIA OBRA MAYOR	2.003	OBRAS	37	2013-12-13T03:37:02.737Z	200.3

Datos de Solicitud de Licencias de Obras (Fuente: http://datos.santander.es/)

5. Fuentes externas Big Data. Redes sociales

Otras de las fuentes Big Data más relevantes son las....

- Redes sociales: Millones de usuarios generan a diario una gran cantidad de información útil que podemos aprovechar
    a. Texto, imágenes, audio, vídeo...
    b. Opiniones, sentimientos, localizaciones...
- Facebook, Twitter, Instagram, Foursquare, Google + o Linkedin son algunos ejemplos.
- Acceso a los datos públicos y a los privados (de nuestra cuenta)
    a. API datos históricos, API streaming, SPARQL...
- Algunos ejemplos
    a. API's de Twitter: https://dev.twitter.com/
    b. Facebook Graph API: https://developers.facebook.com/docs/graph-api

6. Importancia en la selección de las fuentes

- La combinación de la información disponible en fuentes como las anteriores con la información interna de nuestra organización puede ser de gran utilidad para:
    - Estudios de mercado sobre un determinado producto
    - Marketing personalizado
    - Campañas políticas
    - Análisis y detección de comunidades sociales y sus flujos de movimiento
    - ...entre muchas otras posibles aplicaciones. (infinitas?)
- Por ello, es importante rastrear exhaustivamente la red en busca de fuentes de información que puedan ser útiles en el análisis de nuestro proceso de negocio objetivo.
- Sin embargo, no toda la información externa es útil o tiene la calidad suficiente para satisfacer nuestros objetivos de análisis y permitir la extracción de conocimiento fiable.
    - Más aún en los casos en los que la información proviene de proveedores ajenos a la actividad de nuestra organización.
- Si el conocimiento extraído no es fiable puede llevar a la toma de decisiones errónea respecto al proceso de negocio que pretendemos mejorar
    - Puede dar lugar a pérdidas económicas y al fracaso empresarial
- Por ello es importante verificar la **calidad** de las fuentes de datos externas usadas, así como la correlación con los datos internos con los que se combinan para el análisis.