Tipos de datos en Big Data: clasificación por categoría y por origen

A la hora de crear proyectos Big Data que detecten, consuman, gestionen, organicen y presenten dichos datos de una manera optimizada y de forma que aporten algo a nuestro negocio generalmente nos enfrentamos a las siguientes preguntas:

- ¿De dónde obtengo los datos?
- ¿Qué datos aportan más información a mi negocio?
- ¿Qué datos hay disponibles fuera de mi organización que me pueden ayudar?
- ¿Qué volumen de datos tenemos que manejar?
- ¿Qué formato tienen?
- ¿Con qué frecuencia los utilizo?
- ¿Cómo integrarlos en nuestro sistema de gestión?

Aunque todas estas preguntas son importantes, la más importante de todas es:

- ¿Qué problema quiero resolver?

Si no tenemos claro el problema, no podemos plantearnos empezar a trabajar con datos para encontrar una solución.

Cuando hayamos localizado el problema que queremos resolver podremos plantearnos las preguntas iniciales y extraer información. El proceso de obtención de la misma a partir de los datos está reflejado en la famosa pirámide DIKW o pirámide del conocimiento, que relaciona cuatro componentes: Data, Information, Knowledge y Wisdow (Datos, Información, Conocimiento y Sabiduría).

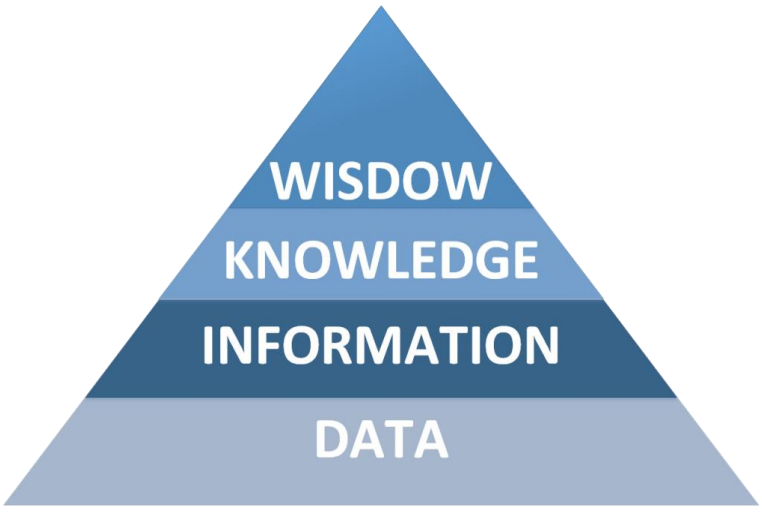

Ilustración 1: Pirámide DIKW

Tipos de datos de Big Data

La categorización de los datos es importante para cualquier proyecto, y en especial cuando vamos a trabajar con grandes volúmenes (Big Data).

Dos de las categorizaciones más utilizadas en Big Data suelen ser las que relacionan la estructura de los datos y las que dependen del origen de los mismos:

Tipos de datos por categorías

Los tipos de datos se suelen organizar en 2 categorías principales:

- Estructurados:
    - Creados: datos generados por nuestros sistemas de una manera predefinida (registros en tablas, ficheros XML asociados a un esquema)
    - Provocados: datos creados de manera indirecta a partir de una acción previa (valoraciones de restaurantes, películas, empresas (Yelp, TripAdvisor, …)
    - Dirigido por transacciones: datos que resultan al finalizar una acción previa de manera correcta (facturas autogeneradas al realizar una compra, recibo de un cajero automático al realizar una retirada de efectivo, …)
    - Compilados: resúmenes de datos de empresa, servicios públicos de interés grupal. Entre ellos nos encontramos con el censo electoral, vehículos matriculados, viviendas públicas, …)
    - Experimentales: datos generados como parte de pruebas o simulaciones que permitirán validar si existe una oportunidad de negocio.
- No estructurados:
    - Capturados: datos creados a partir del comportamiento de un usuario (información biométrica de pulseras de movimiento, aplicaciones de seguimiento de actividades (carrera, ciclismo, natación, …), posición GPS)
    - Generados por usuarios: datos que especifica un usuario (publicaciones en redes sociales, vídeos reproducidos en Youtube, búsquedas en Google, …)
- Multi-estructurados o híbridos:
    - Datos de mercados emergentes
    - E-commerce
    - Datos meteorológicos

Ilustración 2: Categorías

Tipos de datos por origen

Aunque no existe un criterio único para categorizar los tipos de datos lo más extendido es dividirlos en 5 grupos:

- Web y Redes Sociales
  - Información sobre clicks en vínculos y elementos
  - Búsquedas en Google
  - RRSS (fuentes de datos de Twitter, publicaciones en Facebook, otras RRSS)
  - Contenido Web (páginas, imágenes, enlaces, etc.)
- Comunicación entre máquinas
  - Lecturas RFID
  - Señales GPS
  - Otros sensores (parquímetros, máquinas expendedoras, cajeros, etc.)
- Transacciones
  - Registros de comunicaciones (llamadas, mensajería, VoIP, etc.)
  - Registros de facturación (pagos con tarjeta, pago online, etc.)
- Biométricos
  - Reconocimiento facial
  - Información genética (ADN)
- Generados por personas
  - Grabaciones a operadores de atención al cliente
  - E-mail
  - Registros médicos electrónicos

Ilustración 3: Orígenes

Conclusión

Una vez hayamos conseguido identificar nuestras fuentes de datos y hayamos sido capaces de categorizarlos convenientemente podremos pasar a la siguiente fase que consistirá en definir qué mecanismos vamos a utilizar para poder convertirlos en Información útil siguiendo la pirámide DIKW.