

BIG DATA for BUSINESS

1.INTRODUCCIÓN & NEGOCIO

1.1.

Introducción a Big Data

Conecta Empleo

Contenido desarrollado por
Synergic Partners



PROGRAMA

1. INTRODUCCIÓN A BIG DATA

1.1 Introducción a Big Data

1.2 Compañías Data Driven

1.3 Casos de uso

1.4 Metodologías ágiles

1.5 New Trends in Data



1.1 Introducción a Big Data

1. INTRODUCCIÓN AL BIG DATA

DATOS

ANALÍTICA

TECNOLOGÍA



1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

TECNOLOGÍA

Monolítico



- Primeras menciones del término “Big Data”: 1997

Este término se empleó por primera vez en un artículo de los investigadores de la NASA Michael Cox y David Ellsworth, donde ambos afirmaron que el ritmo de crecimiento de los datos empezaba a ser un problema para los sistemas informáticos actuales. Esto se denominó el “problema del Big Data”

- Superordenador más potente de la actualidad: Sunway TaihuLight (Dios del Lago) - *China*
 - 40.960 procesadores con 10.649.600 núcleos = **100.000 billones de cálculos por segundo**
 - Memoria de 1.310 Tb = 1,3 Petabytes
 - Utilizado para prospección de petróleo, ciencias de la vida, el tiempo, el diseño industrial y la investigación de fármacos.



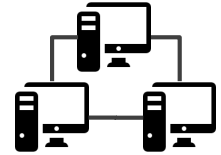
Precio:
\$ 273.000.000

1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

TECNOLOGÍA

Distribuido



Inicialmente existía un **Sistema Monolítico**:

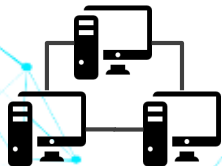


Monolítico

- Una **única máquina**
- Esa máquina individual procesaba la información
- **Limitación de cantidad** de datos a procesar
- **Limitación de velocidad** de dicho procesamiento



Ante esas limitaciones surgen los **Sistemas Distribuidos**:



Sistema Distribuido

- Un **conjunto unificado de máquinas**
- Permite procesar grandes cantidades de datos
- **Limitación de velocidad** porque la programación era muy costosa
- Fallo de una máquina podría generar **pérdida de información**

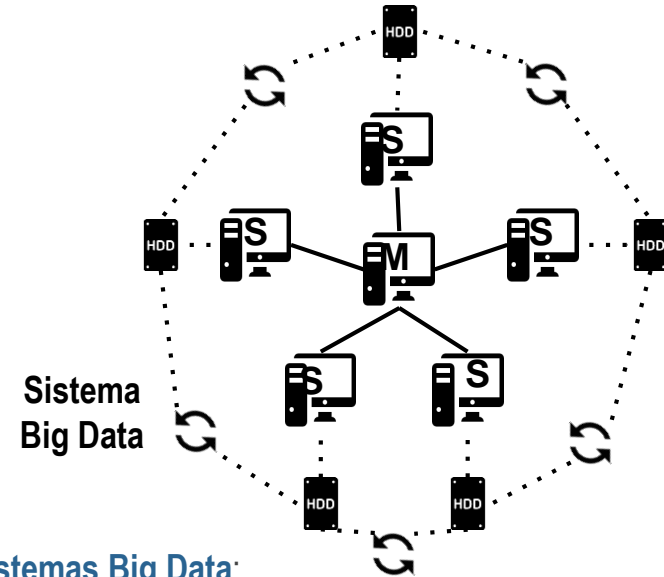
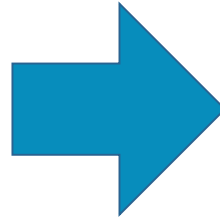
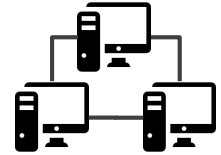
**Sistemas Distribuidos
son efectivos, pero no
siempre eficientes...**

1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

TECNOLOGÍA

Distribuido



Para corregir esas limitaciones existentes surgen los **Sistemas Big Data**:

- **Cambios de la distribución de los elementos físicos:** un **ordenador maestro** gestiona el resto de ordenadores que son los **trabajadores**, procurando que la información a procesar esté bien repartida entre los mismos (cada “worker” trabaja en los datos que tiene almacenados en su propio disco duro)
- **Implementación de una capa de software:** se encarga de la gestión del grupo de ordenadores y abstrae al usuario

1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

TIPOS DE DATOS en la analítica tradicional



Datos
operacionales



Tendencias
de mercado



Información
demográfica del cliente

1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

TIPOS DE DATOS



Datos operacionales



Información demográfica del cliente



Tendencias de mercado



TRADICIONALMENTE

ACTUALIDAD

Datos Estructurados

- Datos con un formato de dato establecido y estructura.
- Ejemplo: Datos transaccionales.

Datos Semi-Estructurados

- Datos de texto con un patrón reconocible, el cual es apto para ser parseado (troceado).
- Ejemplo: Archivos XML que son definidos por un esquema XSD

Datos Quasi-Estructurados

- Datos de texto con un patrón de datos difícil de identificar. Pueden ser formateados con esfuerzo, tiempo y herramientas específicas.
- Ejemplo: Registros de eventos o acciones en una web, logs.

Datos No Estructurados

- Datos que no tienen ninguna coherencia ni patrón y usualmente están almacenados en distintos tipos de archivos
- Ejemplo: Archivos de texto, PDFs, Imágenes, Videos..

1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

EJEMPLOS TIPOS DE DATOS

Datos Estructurados

| | nombre | color | edad | altura | peso | puntuacion |
|----|---------|----------|------|--------|------|------------|
| 1: | Paco | Rojo | 24 | 182 | 74.8 | 83 |
| 2: | Juan | Green | 30 | 170 | 70.1 | 500 |
| 3: | Andres | Amarillo | 41 | 169 | 60.0 | 20 |
| 4: | Natalia | Green | 22 | 183 | 75.0 | 865 |
| 5: | Vanesa | Verde | 31 | 178 | 83.9 | 221 |
| 6: | Miriam | Rojo | 35 | 172 | 76.2 | 413 |
| 7: | Juan | Amarillo | 22 | 164 | 68.0 | 902 |

Datos Semi-Estructurados

```
{  "marcadores": [
    {
      "latitude": 40.416875,
      "longitude": -3.703308,
      "city": "Madrid",
      "description": "Puerta del Sol"
    },
    {
      "latitude": 40.417438,
      "longitude": -3.693363,
      "city": "Madrid",
      "description": "Paseo del Prado"
    }
  ],
}
```

Datos No Estructurados

CAPÍTULO PRIMERO

Que trata de la condición y ejercicio del famoso hidalgo D. Quijote de la Mancha

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lentejas los viernes, algún palomino de añadidura los domingos, consumían las tres partes de su hacienda. El resto della concluían sayo de velarte, calzas de



1.1 INTRODUCCIÓN AL BIG DATA

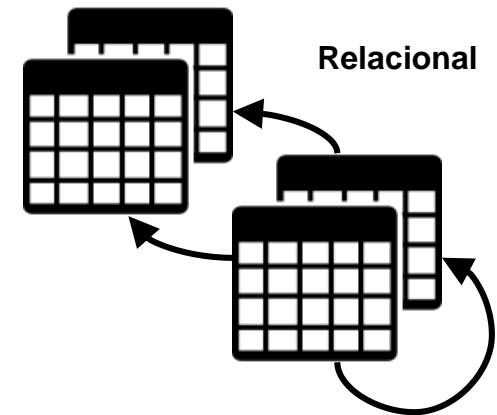
Evolución analítica empresarial

BBDD SQL

Las bases de datos **SQL o relacionales** se caracterizan por estar **formadas por tablas**.

Cada una de las tablas contiene una o varias columnas del mismo o distintos tipos que permiten almacenar la información en filas.

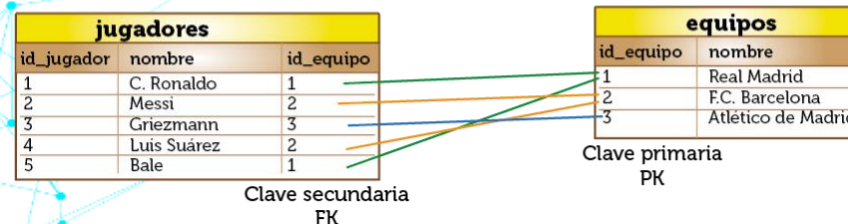
Cada una de las tablas se puede relacionar con ninguna o con varias tablas para formar el modelo relacional que soporta de manera lógica la realidad del negocio.



Modelo relacional



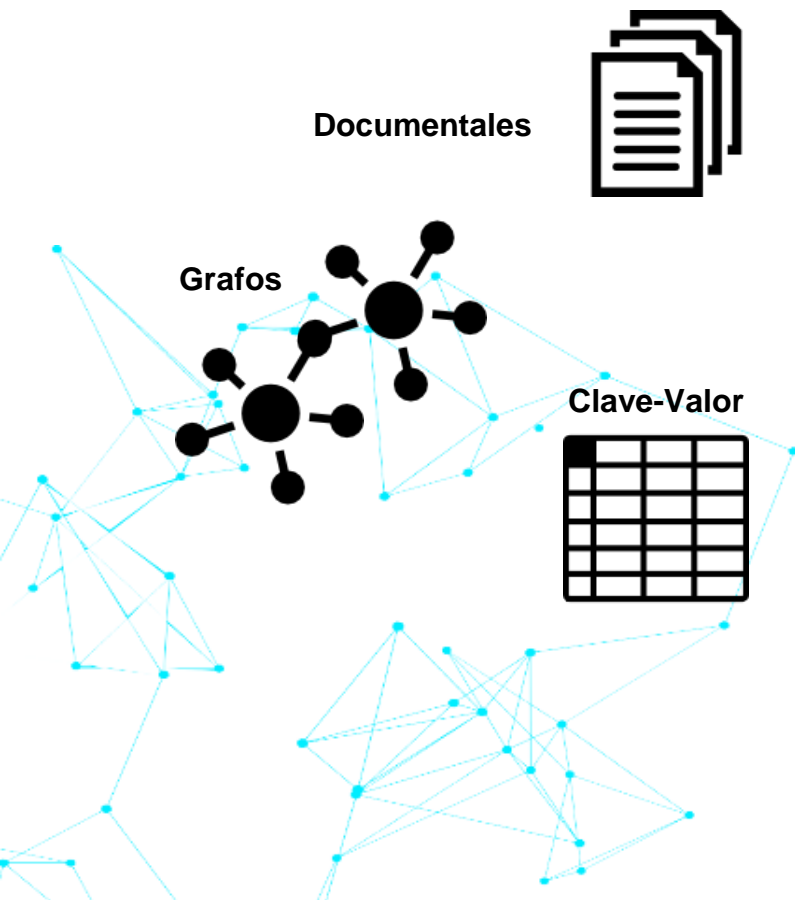
Ejemplo de datos



1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

BBDD DATOS NOSQL



Las bases de datos **NoSQL o no relacionales** abarcan a todos los tipos de bases de datos que no son SQL.

Todas tienen la filosofía de un esquema flexible, que permite adaptarlas de una manera rápida a cambios en las necesidades del negocio.

Entre las más extendidas destacan las siguientes:

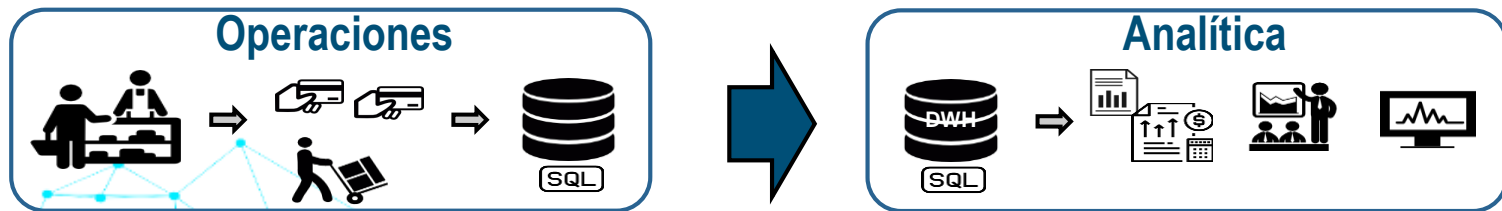
- **Clave-Valor**: Formadas por una tabla con una columna que es la clave y otra en la que se almacena toda la información.
- **Documentales**: Los datos se almacenan en documentos con estructura (XML, JSON, etc).
- **Grafos**: La red de grafos está formada por nodos y relaciones que almacenan la información.

1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

ANALÍTICA TRADICIONAL (BUSINESS INTELLIGENCE)

Permite conseguir los objetivos empresariales a nivel productos/servicios a partir de un análisis de datos.



FOCO DE ANÁLISIS: Informes, KPIs, tendencias

ANÁLISIS: Retrospectivo y Descriptivo

PROCESO DE ANÁLISIS: Comparativo

1.1 INTRODUCCIÓN AL BIG DATA

Evolución analítica empresarial

LIMITACIONES BUSINESS INTELLIGENCE

TECNOLOGÍA

- **Procesamientos lentos:** procesar cantidades requiere de procesos muy pesados y lentos
- **Alto riesgo en posible fallo de las máquinas:** el fallo de una máquina puede generar una pérdida de la información en caso de que no se gestione correctamente.
- **Almacenamiento centralizado de los datos:** único punto de acceso.

ANALÍTICA

- **Silos de información,** falta de compartición de la información
- Análisis por áreas, no globales.
- Analítica de hechos pasados, no predictivo.

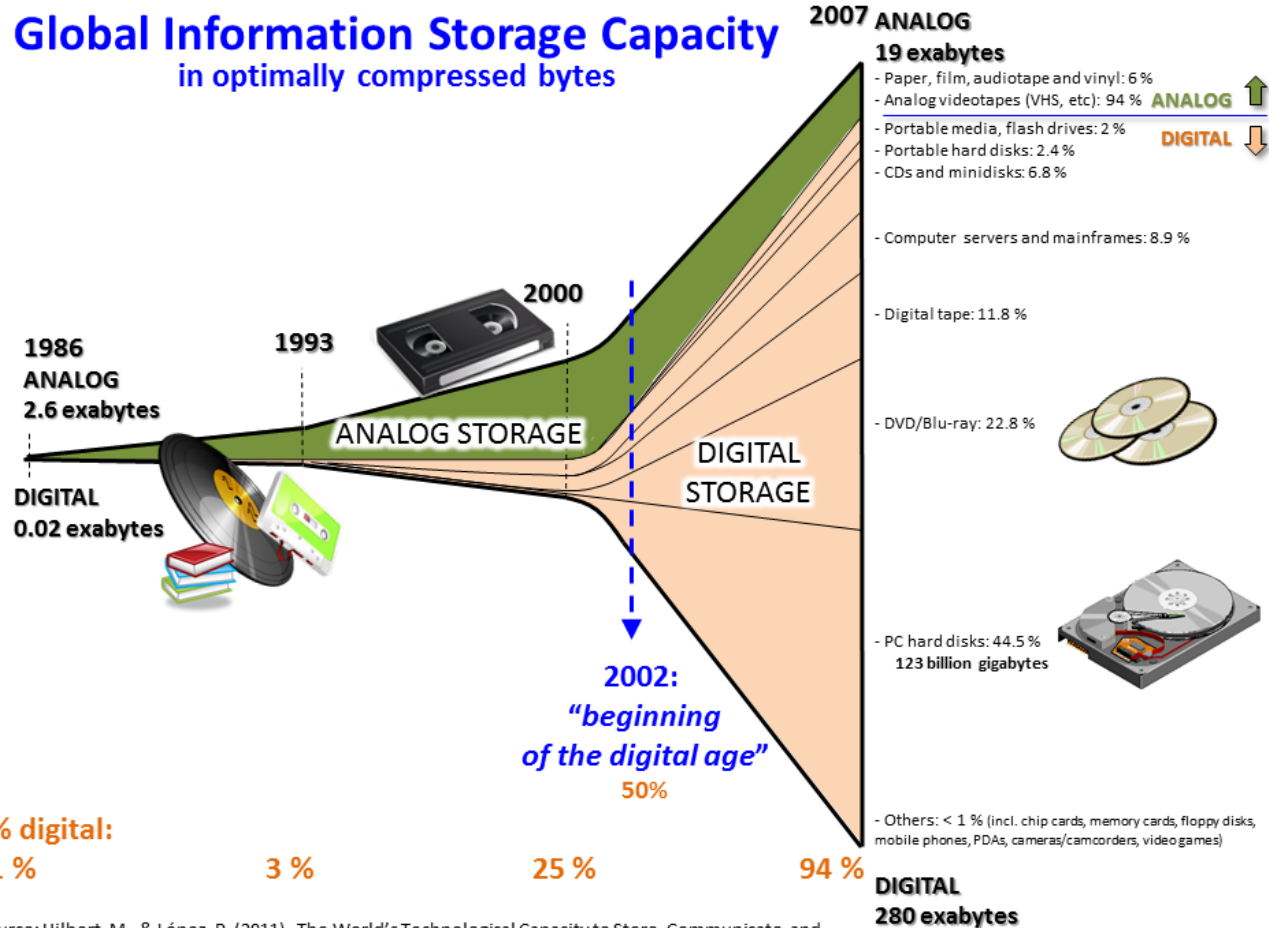
DATOS

- Uso de **datos internos** (y a veces enriquecidos con estudios de mercados muy genéricos).
- **Falta de capacidad** para almacenar todos los datos
- **Infra-explotación** de los datos disponibles (debido a las limitaciones en términos de procesamiento y tiempo).

Conclusión:
Con el **BI** no basta...

1.1 INTRODUCCIÓN AL BIG DATA

Cambio de paradigma



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

1.1 INTRODUCCIÓN AL BIG DATA

Cambio de paradigma



FredCavazza.net

1.1 INTRODUCCIÓN AL BIG DATA

Cambio de paradigma

2018 *This Is What Happens In An Internet Minute*



BIG DATA



1.1 INTRODUCCIÓN AL BIG DATA

Cambio de paradigma



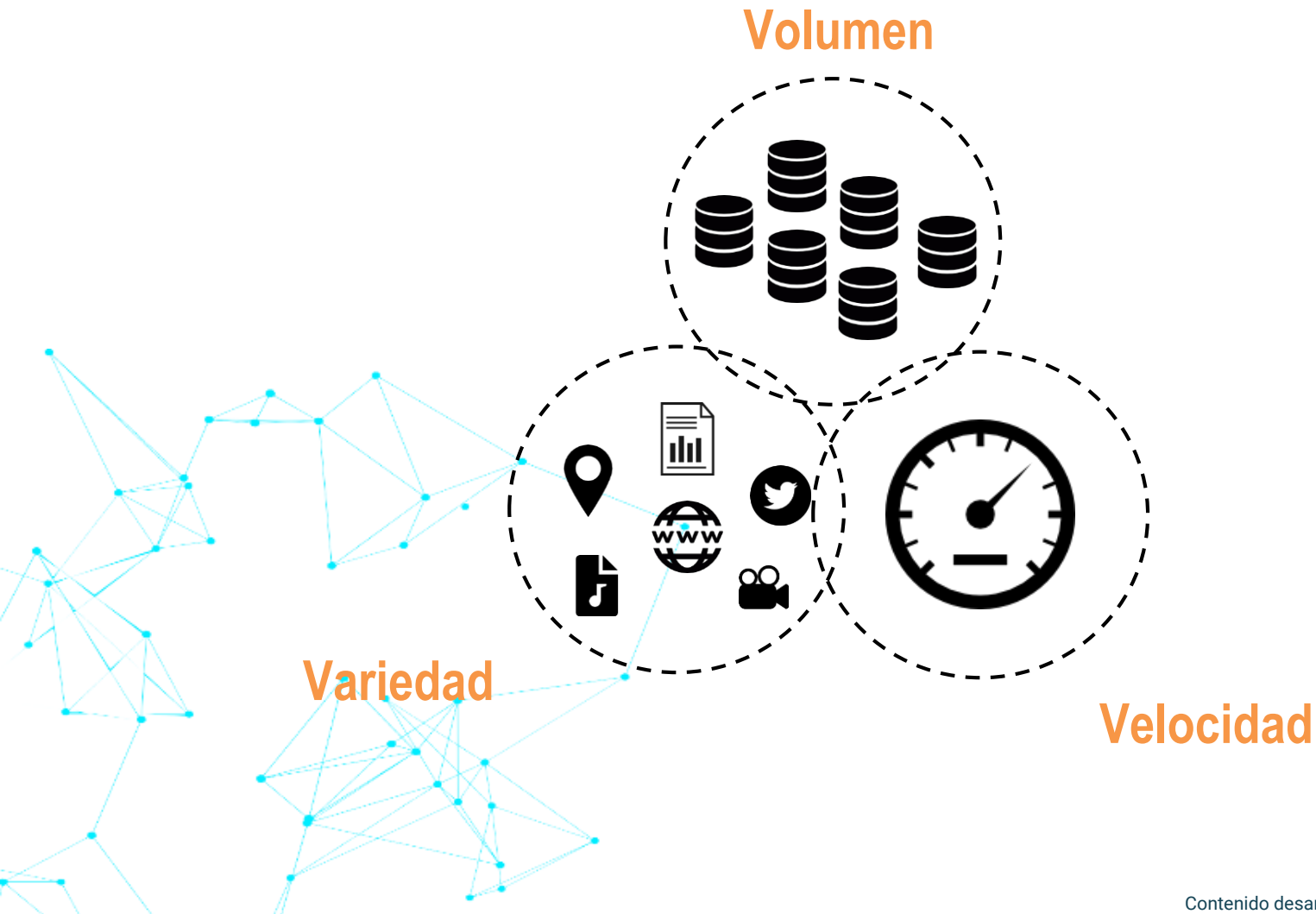
¿Qué es Big Data?

“**Big Data** es un concepto que hace referencia a grandes **conjuntos de datos tan grandes** que aplicaciones informáticas tradicionales del procesamiento de datos no son suficientes para tratar con ellos y a los **procedimientos usados** para encontrar patrones repetitivos dentro de esos datos”

“**Big Data** es un concepto que hace referencia a un **conjunto de procesos, tecnologías y modelos** basados en el almacenamiento **masivo de datos, procesamiento y transformación** de los mismos en **conocimiento**, para **anticipar** lo que sucederá en un mundo complejo con muchas interacciones.”

1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato



1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

Volumen



+



1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

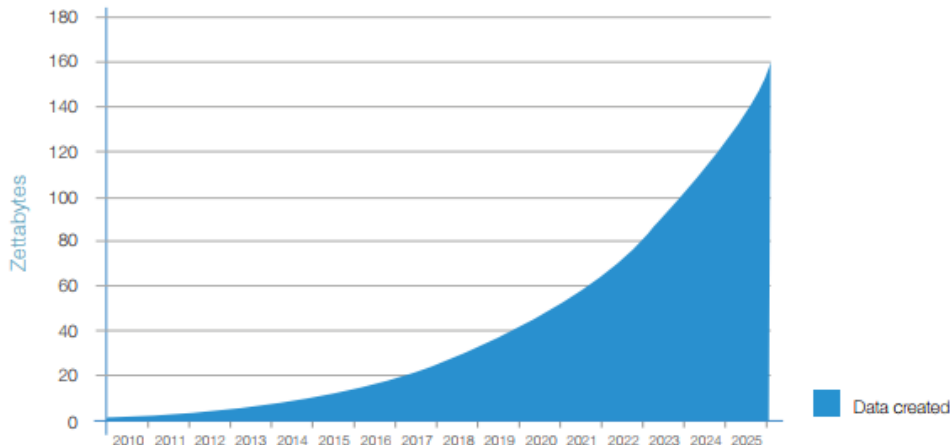
Volumen



Volumen de datos mundiales:

| | |
|------|----------------------------|
| 2009 | 0,8 Zb |
| 2010 | 1 Zb |
| 2011 | 1,8 Zb |
| 2018 | <i>estimado 35 Zb</i> |
| 2025 | <i>estimado 163 Zb (*)</i> |

Figure 2. Annual Size of the Global Datasphere



| Unidad | Abreviatura | Equivalencia |
|-------------|-------------|--------------|
| Byte/Octeto | B | 8 bits |
| Kilobyte | KB | 1024 bytes |
| Megabyte | MB | 1024 KB |
| Gigabyte | GB | 1024 MB |
| Terabyte | TB | 1024 GB |
| Petabyte | PB | 1024 TB |
| Exabyte | EB | 1024 PB |
| Zettabyte | ZB | 1024 EB |
| Yottabyte | YB | 1024 ZB |
| Brontobyte | BB | 1024 YB |
| Geopbyte | GeB | 1024 BB |
| ... | ... | ... |

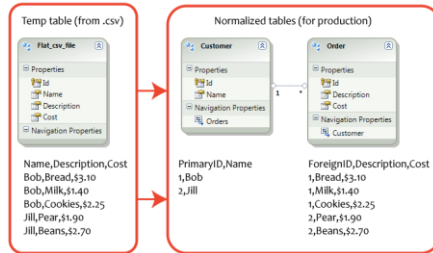
(*) Estudio "Data Age 2025" realizado por IDC y patrocinado por Seagate

1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

Volumen

Estructurados

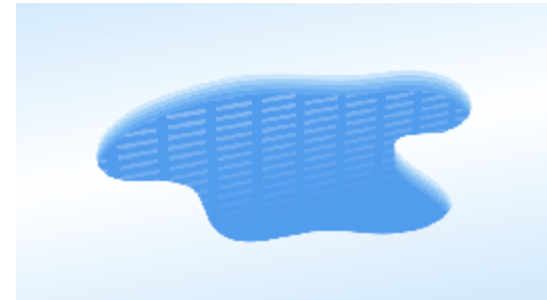


No estructurados

Social Media Landscape



Data Lake

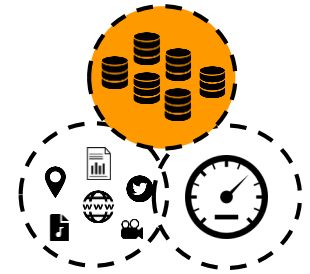


1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

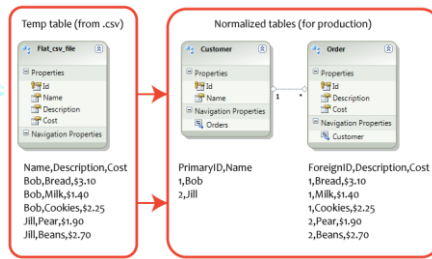


Volumen



¿De qué tipo de datos está compuesto el gran volumen de datos actuales?

Estructurados



No estructurados

Social Media Landscape



En torno al **90%** de los datos actuales son datos no estructurados

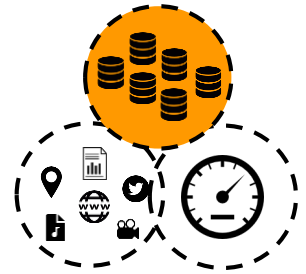
1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

Volumen

Lago de datos (Data Lake)

- Un Lago de Datos facilita datos a una organización para realizar análisis de todo tipo
- Es posible introducir el análisis en el Lago de Datos para generar conocimiento adicional de los datos cargados
- Un Lago de Datos gestiona repositorios compartidos de información para analizarla
- Cada repositorio del Lago de Datos se optimiza para un procesamiento particular
- Los datos pueden replicarse en múltiples repositorios en el Lago de Datos y tener distintos significados/usos



Lago de Datos = Gestión eficiente, Gobernanza, protección y acceso

Data Warehouse VS Data Lake

- Un Data Lake es un repositorio de almacenamiento que contiene gran cantidad de datos en bruto (estructurados, semi-estructurados y desestructurados) mientras que el Data Warehouse solo recoge datos procesados y/o estructurados. Es decir, en el DW encontramos datos a los que ya se les ha definido un uso concreto, mientras que en los DL se cargan todos los datos sin excepción.
- Los usuarios de un DW son profesionales dedicados a los negocios, mientras que en el DL están más enfocados a data scientist, etc.

1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

Velocidad



Los datos se generan muy rápido y necesitan ser procesados a una gran velocidad

Tipos de procesamiento de la información:

- **Procesamiento Batch:** los datos se van acumulando y procesando periódicamente, requiere de una arquitectura escalable y con gran capacidad de almacenamiento
- **Procesamiento Streaming:** los datos se procesan de forma inmediata y requieren una arquitectura de baja latencia
- **Procesamiento híbrido:** Batch + Streaming, han de cumplir una arquitectura con capacidad para ambos procesos.

Decisiones tomadas tarde conllevan a oportunidades perdidas

1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

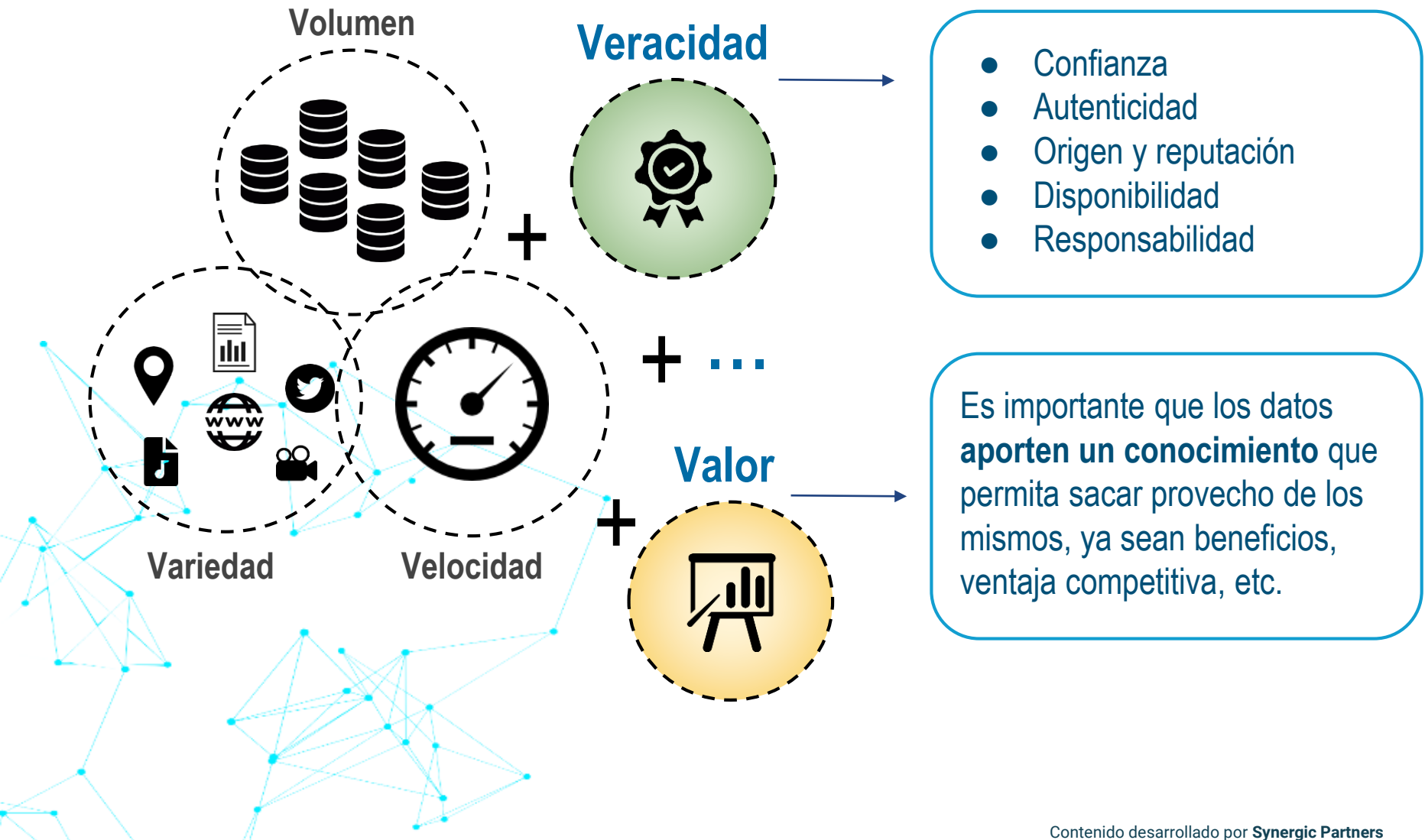
Variedad

- Diferentes formatos, tipos y estructuras
- Texto, números, imágenes, audio, video, secuencias, series temporales, datos de redes sociales, etc.
- Datos estáticos vs datos en tiempo real
- Una sencilla aplicación puede generar y almacenar muchos tipos diferentes de datos



1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato



1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

Valor & Veracidad

- Mejorar las estimaciones del riesgo de crédito
- Aumentar cartera de clientes
- Aminorar la pérdida de clientes
- Aumentar la satisfacción de los clientes
- Incrementar la eficiencia de los programas de Marketing
- Aumentar las ventas mediante análisis predictivos



Smart Data

1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

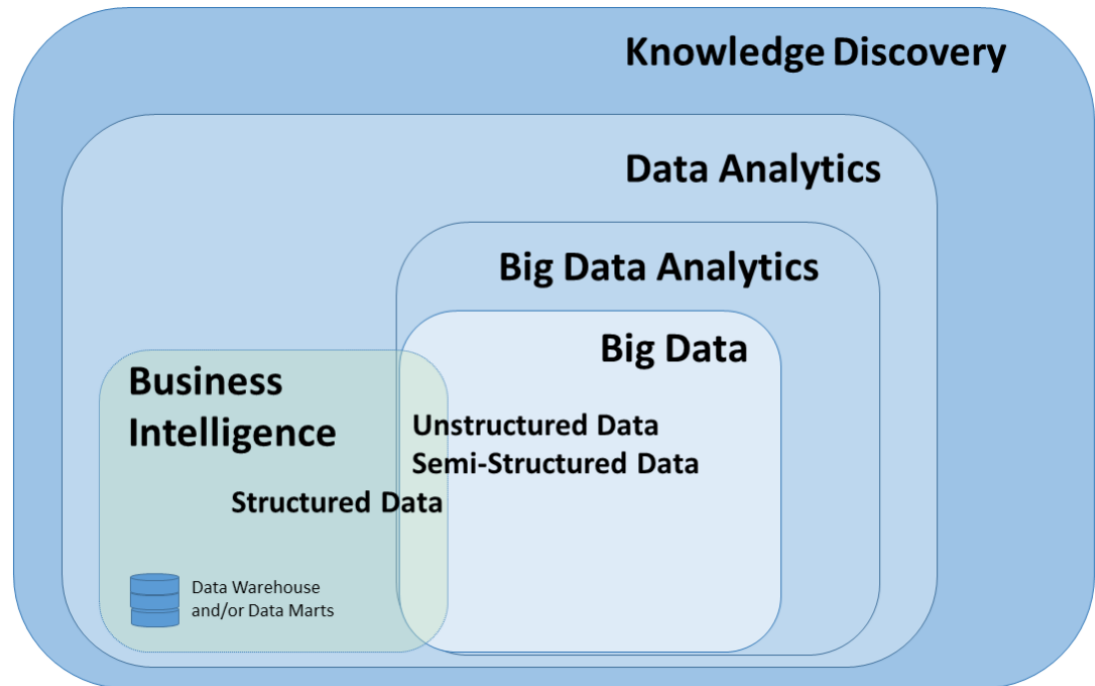
Business Intelligence & Big Data

Business Intelligence:

Enfoque analítico referido a todos los datos acumulados (**presente, pasado**) con datos estructurados y que permiten responder a preguntas objetivas sobre como es el estado actual o pasado.

Big Data:

Aporta nuevas herramientas para trabajar que permiten otras escalas de volumen, velocidad y flexibilidad para adaptarse a las nuevas tecnologías de datos. Se enfoca en hacer **predicciones**, modelos, previsiones etc. para un **futuro** a partir del análisis del pasado.



Business Intelligence y Big Data **NO son excluyentes**. Ambas forman parte del mismo ecosistema de soluciones analíticas, por lo que pueden complementarse e integrarse

1.1 INTRODUCCIÓN AL BIG DATA

Las V's del Big Data y la Ciencia del Dato

Resumen:

BUSINESS INTELLIGENCE

BIG DATA

| | | |
|--------------|--|---|
| FOCO | Informes, KPIs, tendencias | Patrones, correlaciones, modelos |
| PROCESO | Estático, comparativo | Exploración, experimentación, visualización |
| DATOS | Planificado, estructurados, crece lentamente | 'On the fly', según necesidad, no estructurados, crece por segundos |
| ANÁLISIS | Retrospectivo, descriptivo | Predictivo, prescriptivo |
| ARQUITECTURA | Centralizada | Distribuida |

Cuestiones a las que dan respuesta

Descriptivo:
QUIÉN?

Comportamental:
QUÉ?

Predicción:
QUÉ HARÁN?

Actitudinal:
POR QUÉ?

Interacción:
CÓMO?

1.1 INTRODUCCIÓN AL BIG DATA

Ejemplo



¿Qué productos descubrió Walmart que estaban entre los más vendidos cuando se aproximaba un huracán?

- Cerveza
- Linternas y velas
- Agua embotellada
- Pop-tarts (barritas dulces para el desayuno)
- Suministros de primeros auxilios
- Baterías y generadores



TODOS !!

- La cerveza era el producto que más aumentaba sus ventas
- La venta de pop-tarts se multiplicaba por 7



1.1 INTRODUCCIÓN AL BIG DATA

Aspectos legales y éticos a tener en cuenta

Es completamente **imprescindible cumplir la legalidad** en torno a los datos puesto que es una información muy sensible



1.1 INTRODUCCIÓN AL BIG DATA

Aspectos legales y éticos a tener en cuenta



TECNOLOGÍA

VS

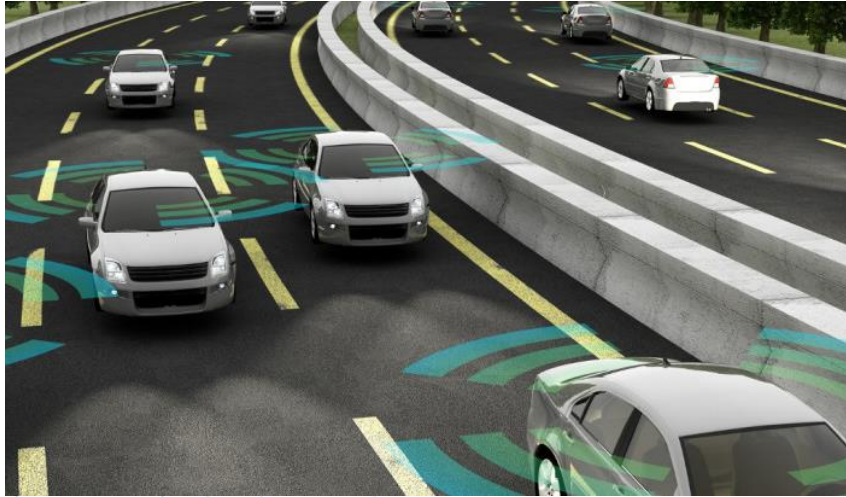


LEGISLACIÓN

1.1 INTRODUCCIÓN AL BIG DATA

Aspectos legales y éticos a tener en cuenta

DILEMA MORAL



Google™



Ante una situación de riesgo de atropello de peatones ¿el coche debe salvar la vida de los pasajeros o la de los peatones?



1.1 INTRODUCCIÓN AL BIG DATA

Aspectos legales y éticos a tener en cuenta

¿Qué es la protección de datos?

Es un **derecho fundamental** de todas las personas que se traduce en la potestad de **control sobre** el uso que se hace de sus **datos personales**.

Este control permite evitar que, a través del tratamiento de nuestros datos, se pueda llegar a disponer de información sobre nosotros que afecte a nuestra intimidad y demás derechos fundamentales y libertades públicas.

artículo 18 Constitución Española

En España la ley que regula este derecho es la Ley de Protección de Datos (LOPD), la cual ha sufrido modificaciones con la entrada en vigor el RGPD

1.1 INTRODUCCIÓN AL BIG DATA

Aspectos legales y éticos a tener en cuenta

Reglamento General de Protección de Datos (RGPD)



Objetivos del nuevo reglamento europeo:

- **Devolver a los ciudadanos el control sobre sus datos personales** en un mundo dominado por los teléfonos inteligentes y las redes sociales y donde la banca por Internet y el comercio online están a la orden del día.
- **Mejorar la seguridad en cuanto a la legislación para las empresas que operan dentro de la Unión Europea**, con el fin de impulsar la innovación y el desarrollo futuro del mercado único digital, y también se aplicará a las empresas fuera de Europa que se dirijan a los consumidores de la Unión Europea.

Entrada en vigor definitivo el 25 de Mayo de 2018

1.1 INTRODUCCIÓN AL BIG DATA

Principales diferencias entre la LOPD y la RGPD

SANCIONES

LOPD

Entre **900€** y
600.000€



RGPD


Hasta **20 MM de €**
o
4% de la facturación
global anual

1.1 INTRODUCCIÓN AL BIG DATA

La importancia de los datos

Taller:

*Identificar fuentes de datos
generadas por el negocio en otros
sectores*

An abstract network diagram consisting of numerous small blue dots connected by thin, light blue lines. The dots are scattered across the lower-left and bottom-center portions of the slide, with lines forming a complex, interconnected web that suggests data flow or relationships between nodes.

1.1 INTRODUCCIÓN AL BIG DATA

La importancia de los datos

FUENTES DATOS INTERNAS

- Customer Relationship Management (CRM):
 - Demográficos
 - DNI
 - Nacionalidad
 - Información financiera-nivel de consumo
- Interactive voice response (IVR)
- Datos generados en las plataformas online
- ERP (Enterprise Resource Planning)
- Módulos ERP (Finanzas, RRHH, Logística, Stock Management)

FUENTES DATOS EXTERNAS

- Open Data
- Redes Sociales
- B2B
- Datos de meteorología
- Social Listening
- Open Street Maps
- CIRBE (Central de Información de Riesgos)

1. INTRODUCCIÓN AL BIG DATA

La importancia de los datos

Nombre del caso de uso

IDENTIFICACIÓN DE FUENTES DE DATOS

SECTOR BANCARIO/ TELECOMUNICACIONES

1. INTRODUCCIÓN AL BIG DATA

La importancia de los datos

TELECOMUNICACIONES

- **Deep Packet Inspection (DPI)/ WebLogs:** Recogen detalles generados en el uso de datos de tráfico, en concreto, por en la navegación web o el uso de aplicaciones.
 - Duración de la conexión
 - Tipo de conexión: Navegación web, correo, juegos, contenido multimedia, compras, navegación en webs de compañías competidoras...
 - Consumo por conexión
 - User Agent: Tipo de dispositivo, navegador, sistema operativo.
 - Antena a la que se conectó
- **Call Detail Record (CDR):** Datos generados por llamadas telefónicas que documentan los detalles de una llamada u otra transacción, como pueden ser mensajes de texto.
 - El número de la parte que origina la llamada
 - El número que recibe la llamada
 - La hora de inicio de la llamada
 - La duración de la llamada
 - La factura generada por la llamada
 - Un número único secuencial generado por la llamada
 - Los resultados de la llamada, indicando, por ejemplo, si llegó a haber conexión
 - Cualquier fallo que pudiera resultar de la llamada
 - A qué antena se conectó el dispositivo que generador y el receptor de la llamada
 - Número de conexiones a las antenas
- **Red:** Recogen todos los detalles de los eventos que suceden en la red
 - Movimientos de usuarios,
 - Zonas hay más usuarios,
 - Tiempo de conexión en la antena
 - Hora de conexión
 - Datos de Roaming
- **Decodificadores de televisión:** Recogen los datos generados en su uso
 - Pulsaciones del mando
 - Canales utilizados
 - Tiempo de visualización
 - Contenido on demand
- **CRM y otras telco:** Datos relativos a sus clientes:
 - Parque Prepago
 - Parque Postpago
 - Altas/Bajas
 - Promociones
 - Campañas
 - Planes / Tarifas
 - Reclamos
 - Quejas / incidencias
 - Consultas de saldo
 - Intentos de consulta / recarga

1. INTRODUCCIÓN AL BIG DATA

La importancia de los datos

BANCARIOS

- **Datos Cajeros:** Cantidad de movimientos realizados en cajeros. Detalles relativos a los cajeros de donde se ha sacado dinero, lugar, hora/día
- **Cheques:** Número de cheques ingresados, pagados, identificando la cantidad o el importe.
- **Datos por cada uno de los productos bancarios contratados:** Información relativa a productos contratados, como por ejemplo cuenta nómina, cuenta ahorro, depósitos, derivados etc. En concreto:
 - Número de cuentas activas, identificando el número de meses que llevan activas, cuentas más antiguas, más nuevas
 - Fecha de alta
 - Saldo mensual en las cuentas (mínimo, media y máximo disponible)
 - N° meses desde el vencimiento más antiguo en créditos
- **Datos por cada uno de los productos financieros y fondos de inversión:**
 - Número de operaciones de compra, de venta de productos financieros
 - Número de traspasos de entrada y salida de Fondos (Renta Fija y Renta Variable), Número de reembolsos
- **Movimientos en cuenta:**
 - N° operaciones de cajeros al debe y al haber
 - N° transferencias
 - Importe operaciones
- **Riesgo:**
 - Scoring clientes
 - Operacional
 - De mercado
 - De crédito
- **Compliance/Prevención del Fraude:**
 - Propiedades de la persona
 - Documentación justificativa de las operaciones realizadas con la entidad
- **Sociodemográficos:**
 - Fecha de nacimiento
 - estado civil
 - sexo, edad
 - Ocupación (según clasificación) y Profesión
 - Situación de riesgo de la persona
 - Código postal
 - Importe de renta (a nivel familiar)
 - Saldo en inversión
 - Saldo en patrimonio
- **Incidencias / Call Center:** Datos relativos sobre el número de incidencias en relación a todos los productos y operaciones

PROGRAMA

1. INTRODUCCIÓN A BIG DATA

1.1 Introducción a Big Data

1.2 Compañías Data Driven

1.3 Casos de uso

1.4 Metodologías ágiles

1.5 New Trends in Data

Tareas para el TEC

Realizar el tipo test disponible en la plataforma

Participación en el foro



Conecta Empleo

