

# ¿Qué es y qué no es BIG DATA?



- Iñaki Regidor Mendiola



# En las Redes Sociales



- [www.facebook.com/inaki.regidor](http://www.facebook.com/inaki.regidor)
- [www.pinterest.com/iregidor](http://www.pinterest.com/iregidor)
- [google.com/+IñakiRegidor](http://google.com/+IñakiRegidor)
- [es.linkedin.com/in/inakiregidor](http://es.linkedin.com/in/inakiregidor)
- [twitter.com/inaki\\_regidor](http://twitter.com/inaki_regidor) @Inaki\_Regidor



Denuncian las sanciones de hasta 120.000 euros impuesta por el comisionado por los contratos de publicidad que los propietarios de estas máquinas firman con las casas de tabaco. En la Asociación denuncian que "no se trata de acuerdos de publicidad sino de **venta de información**".

# ¿Qué es BIG DATA?

- Big data = datos masivos.
- Gestión y análisis de enormes volúmenes de datos.
- No pueden ser tratados de manera convencional.
- Superan los límites y capacidades de las herramientas de software habitualmente utilizadas para la captura, gestión y procesamiento de datos.
- Capacidades, competencias y retos relacionados con el almacenamiento y el análisis de estos grandes conjuntos de datos,
- Para apoyar un nivel de toma de decisiones más preciso y oportuno.

# Cifras del BIG DATA

- Facebook cuenta con más de 1.591 millones de usuarios, 1.300 desde móvil en 2015Q4
- En Twitter 320 millones de usuarios activos cada mes,
- Cada día se envían 340 millones de tweets (unos 4.000 por segundo),
- En 2016 habrá más de 6.300 millones de dispositivos conectados,
- En 2020 superará los 26.000 millones de dispositivos,
- Para 2020 cada individuo creará 1,7 megabytes de información nueva por segundo,
- realizamos 40.000 nuevas búsquedas en Google por segundo, lo que equivale a 3,5 millones de búsquedas por día y 1,2 billones por año,
- Se sube una media de 300 horas de vídeos a YouTube por minuto,
- En 2015 se tomaron 1 billón de fotografías, y miles de millones serán compartidas en la red. Para 2017 se estima que casi el 80% de las fotografías serán tomadas por smartphones,
- De momento solo se analiza y utiliza entre el 0,5% y el 3% de todos los datos que existen,
- En 2012 se estimaba que había **2.200 millones de usuarios de cuentas de correo electrónico**
- Se mandaron **144.000.000.000 mails al día. 52,56 Billones de correos al año!!**
- El **61% de los emails** fueron considerados **no esenciales** y casi el **70 % del tráfico de correo electrónico era Spam**
- **En 2014 Gmail** tenía entre 520-530 millones de cuentas de correo,
- Y a todo esto, China casi no cuenta, van a lo suyo.

# ¿Cuáles son las 5 características que definen al Big Data?

- **1.- Volumen:**

- El volumen es una de las principales características del Big Data.
- El volumen de datos nos puede indicar desde datos de identificación del cliente, el sexo, la edad, el origen geográfico, su dispositivo de navegación, qué páginas visita en nuestra web, el tiempo que permanece en ellas, con qué frecuencia visita nuestra página, los clicks que hace en ella, el histórico de su navegación, el idioma de navegación que usa... Los datos son ilimitados.

# ¿Cuáles son las 5 características que definen al Big Data?

- **2.- Variedad:**

- La información digital es tremendamente variada: texto, gráficos, imágenes, vídeos, presentaciones multimedias, juegos..., en diferentes formatos y tipos de archivos y provenientes, además, de variados dispositivos como smartphones, tabletas, portátiles, ordenadores de sobremesa...

# ¿Cuáles son las 5 características que definen al Big Data?

- **3.- Velocidad:**

- Una de las características intrínsecas del medio digital es que todo ocurre a mucha velocidad, e incluso en tiempo real: lo que hoy es trending topic en Twitter... mañana puede ya no serlo, un comentario negativo sobre una marca en Facebook, puede desencadenar un terremoto en cuestión de minutos...

# ¿Cuáles son las 5 características que definen al Big Data?

- **4.- Veracidad:**

- Es imprescindible poder medir y evaluar la veracidad y la autenticidad de los datos que obtenemos mediante el Big Data.
- La veracidad de los datos obtenidos es clave.

# ¿Cuáles son las 5 características que definen al Big Data?

- **5.- Valor:**

- Es imprescindible generar valor de negocio con los datos que nos proporciona el Big Data.
- Todas las empresas generan, trabajan y gestionan multitud de datos, pero la clave está en cómo obtener la mejor información, el mejor valor y conocimiento, para sacar la mayor rentabilidad posible:
- Saber que datos se deben de recolectar.
- Saber que soluciones analíticas usar.
- Saber cómo convertir los datos en conocimiento y el conocimiento en ¿dinero?.

# TECNOLOGÍAS





# Palabros

- IoT (Internet of things).
- Smart cities
- Cloud.
- Virtualización.
- Elasticidad.
- Amazon S3.
- BBDD noSQL. MongoDB
- 5 G
- IBM Watson
- Software libre
- Hadoop

# Perfiles profesionales. El Data Scientist

- Matemáticas, estadística, marketing y programación.
  - según la consultora **Gartner**, el analista Big Data debe tener una buena base de conocimientos informáticos –lenguajes como **Piton, Java o MapReduce**–, además de formación en estadística y matemáticas para ser capaz de manejar ingentes cantidades de datos desestructurados.
- Curiosidad extrema, propia de un científico y que lo empuja a buscar interpretaciones y tendencias más allá de la apariencia, dado que explora y examina datos de diferentes orígenes.
- Una fuerte visión para los negocios.
- Una gran capacidad de comunicación, fundamental para presentar y transmitir los descubrimientos.
- La demanda para este perfil es tan alta que solo se puede cubrir un 20% de esta. [McKinsey](#) calcula que, para 2018, EEUU podría enfrentarse a una escasez de unos 190 mil profesionales con habilidades de análisis, así como de unos 1,5 millones de gestores y analistas con el *know-how* para utilizar el análisis de grandes datos para tomar decisiones efectivas.

# Perfiles profesionales. El SEO

- **Es un buen investigador.**

- El SEO tiene la capacidad de investigar, analizar y evaluar información cualitativa y cuantitativa, que le permite deducir y detectar oportunidades y fortalezas para aplicar a sus estrategias.

- **Es Web master.**

- Aunque no se especializa en el tema, un SEO profesional conoce sobre servidores, funcionamiento web, navegadores, esta al tanto de la evolución en los lenguajes de programación y esta pendiente de los avances tecnológicos.

- **Sabe HTML5 y CSS3.**

- No es necesario que un SEO profesional tenga estudios superiores en lenguajes de programación o sea ingeniero de sistemas. Basta con conocer HTML y CSS básico para poder definir estructuras y funciones dentro del código, apoyándose en herramientas de desarrollo web.

- **Sabe sobre publicidad y marketing tradicional.**

- El SEO profesional debe saber y muchas veces apoyarse en marketing y publicidad tradicional, para lograr objetivos específicos dentro de su estrategia.

- **Puede desenvolverse como community manager.**

- El uso de redes sociales se hace presente en el día a día de un SEO profesional, apoyando al community manager en esta tarea y muchas veces supliendo su actividad a falta de presupuesto.

# El bigdantepasatus



Navajas y peines, bisturís y tijeras  
a mis órdenes todos están,  
tengo **recursos**, además de **oficio**,  
con la joven, con el caballero.  
¡Ah, que vida más hermosa,  
qué gran placer, qué gran placer,  
para un barbero de calidad, de calidad!

Todos me llaman, me **solicitan**,  
mujeres, muchachos, viejos y niñas.  
Trae la peluca...  
rápido, la barba...  
Trae las sanguijuelas, pronto, esta carta.  
¡Fígaro, Fígaro, Fígaro, Fígaro!  
¡Ay de mí, ay de mí! ¡Qué furia!  
¡Ay de mí! ¡Ay, qué multitud!  
¡De uno en uno, por caridad!

# Casos

- Panamá papers.
- DHL.
- Chicago Police Department → ENDECA  
Guardia Civil + Ertzaintza
- Ensamblaje coches GM Figueruelas.
- Triunfo de OBAMA 2008.
- Google y el euskera.
- Google now – Google ads. Gripe en USA

# Panamá papers.

**1,7 GB**  
Cablegate/Wikileaks (2010)

**260 GB**  
Offshore Leaks/ICIJ (2013)

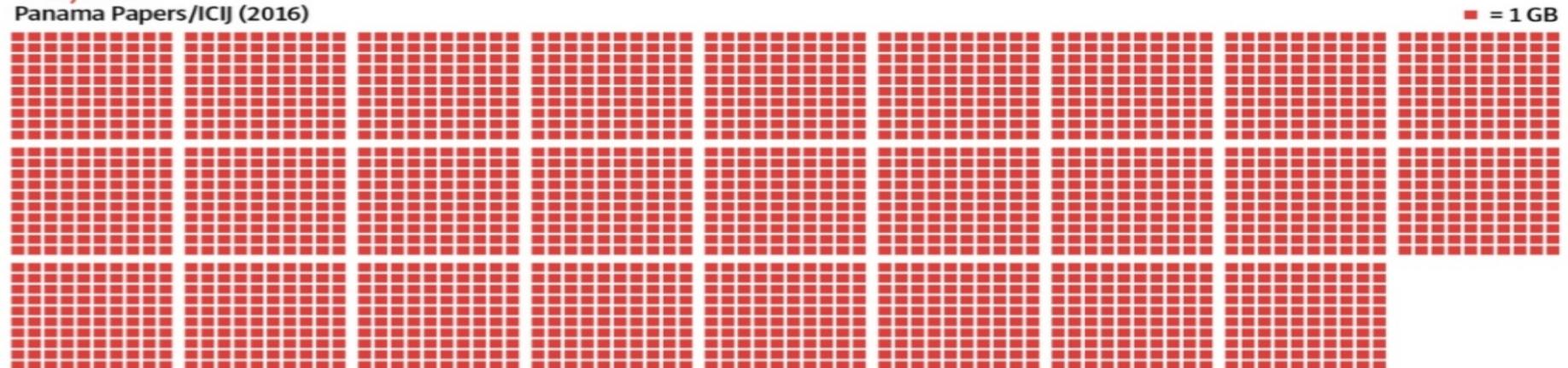
**4 GB**  
Luxemburg Leaks/ICIJ (2014)

**3,3 GB**  
Swiss Leaks/ICIJ (2015)

**≈ 2,6 TB**  
Panama Papers/ICIJ (2016)

## The scale of the leak

Volume of data compared to previous leaks



# Panamá papers.

- 2.6 Terabytes, 11,5 millones de documentos:
  - documentos de texto
  - .pst de correo electrónico
  - Imágenes
  - pdf, que deben ser pasados por un proceso de reconocimiento de caracteres OCR
- se dedicaron unos treinta servidores en paralelo,
- la navegación en la base de datos se desarrolló mediante [Nuix](#), un software de gestión documental habitualmente utilizado por despachos de abogados para el procesamiento de sumarios complejos o muy extensos,
- para su visualización posterior se utilizó [Linkurious](#),
- Para el proceso de autenticación y *login* se optó por [Google Authenticator](#) con verificación en dos pasos,
- Periodistas convertidos en expertos gestores de información, trabajando con herramientas muy alejadas de las que habitualmente componen la equipación de un periodista. Decididamente, periodismo del siglo XXI.

# DHL

- A través de una mejor predicción de la demanda, DHL ha observado cómo algunos de sus clientes han reducido entre el 20% y el 30% del inventario, dependiendo del sector en el que operen, al tiempo que la tasa media de utilización de la cadena de suministro aumentaba de 3 a 7 puntos porcentuales.

# Chicago Police Department → ENDECA

## Guardia Civil + Ertzaintza

- <http://www.kpipartners.com/webinar-Drive-Insight-From-Unstructured-Data-With-Endeca>
- [http://es.slideshare.net/sogeti\\_nl/oracle-wim-villiano-endeca-information-discovery](http://es.slideshare.net/sogeti_nl/oracle-wim-villiano-endeca-information-discovery)

### Increase Public Safety

- Multiple data sources: Police Incidents System, Social Media, News, Fora, etc.
- Research and Analyze security incidents
- Identify and Mitigate Risk and Threats



# Ensamblaje coches GM Figueruelas

- Conectaron robots ensamble a otros sistemas de información.
- Conclusiones.
  - Detectados los errores de ensamble.
  - Patrones de comportamiento humano, ligados a precariedad y no formación.

# Triunfo de OBAMA 2008.

- <http://www.infoworld.com/article/2613587/big-data/the-real-story-of-how-big-data-analytics-helped-obama-win.html>

# Google y el euskera.

- ¿Qué hicimos Gobierno Vasco y EITB?

# Google now – Google ads. Gripe en USA

- <http://www.whatsnew.com/2015/08/20/google-continua-trabajando-con-el-estudio-de-la-evolucion-de-la-gripe-analizando-las-busquedas/>

## Conoce la evolución de la gripe en todo el mundo

Hemos descubierto que ciertos términos de búsqueda sirven como buenos indicadores de la actividad de la gripe. Evolución de la gripe en Google utiliza los datos globales de las búsquedas en Google para hacer cálculos aproximados de la actividad de esta enfermedad. [Más información »](#)



# ¿Qué es y qué no es BIG DATA?



- **ESKERRIK ASKO!!**
- **Dudas, preguntas y comentarios.**