# 1.3 Big Data – (video introductorio)

https://www.youtube.com/watch?v=6HVe54Yfa5A&feature=youtu.be

### 1.3.1 Introducción

Denominamos *Big Data* a la gestión y análisis de enormes volúmenes de datos que no pueden ser tratados de manera convencional, ya que superan los límites y capacidades de las herramientas de software habitualmente utilizadas para la captura, gestión y procesamiento de datos.

Dicho concepto engloba infraestructuras, tecnologías y servicios que han sido creados para dar solución al procesamiento de enormes conjuntos de datos estructurados, no estructurados o semi-estructurados (mensajes en redes sociales, señales de móvil, archivos de audio, sensores, imágenes digitales, datos de formularios, emails, datos de encuestas, logs etc,) que pueden provenir de sensores, micrófonos, cámaras, escáneres médicos, imágenes?


**Bases de datos relacionales y Big Data: ¿son compatibles?**

Tal y como hemos comentado anteriormente, en las bases de datos relacionales los datos se almacenan con una relación definida sobre una estructura basada normalmente en tablas que contienen filas y columnas.

Los principales problemas a los que se enfrentan las bases de datos relacionales con el Big Data son los siguientes:

- No son flexibles y por lo tanto no están diseñadas para posibles cambios.
- Tienen problemas para manejar datos heterogéneos.
- Su diseño no está optimizado para realizar labores operativas y de análisis, resultando ineficientes.
- No están preparadas para el desarrollo de aplicaciones modernas debido a que estas recurren a lenguajes de programación orientados a objetos.

Sin embargo, estos problemas a los que se enfrentan las bases de datos SQL (o bases de datos relacionales) no implica su desaparición, sin embargo, las Bases de Datos no relacionales se adaptan mejor al entorno Big Data, lo que significa que serán cada vez más utilizadas.

Además, gracias al avance en el uso de las bases de datos no relacionales será posible combinarlas con las bases de datos relacionales para determinadas aplicaciones.

Lo que es un hecho es que los próximos años van a estar marcados por la generación continua de datos, su manejo y el posterior análisis.

### 1.3.2 Objetivos del Bigdata

El objetivo de *Big Data*, al igual que los sistemas analíticos convencionales, es convertir el Dato en información que facilita la toma de decisiones, incluso en tiempo real.

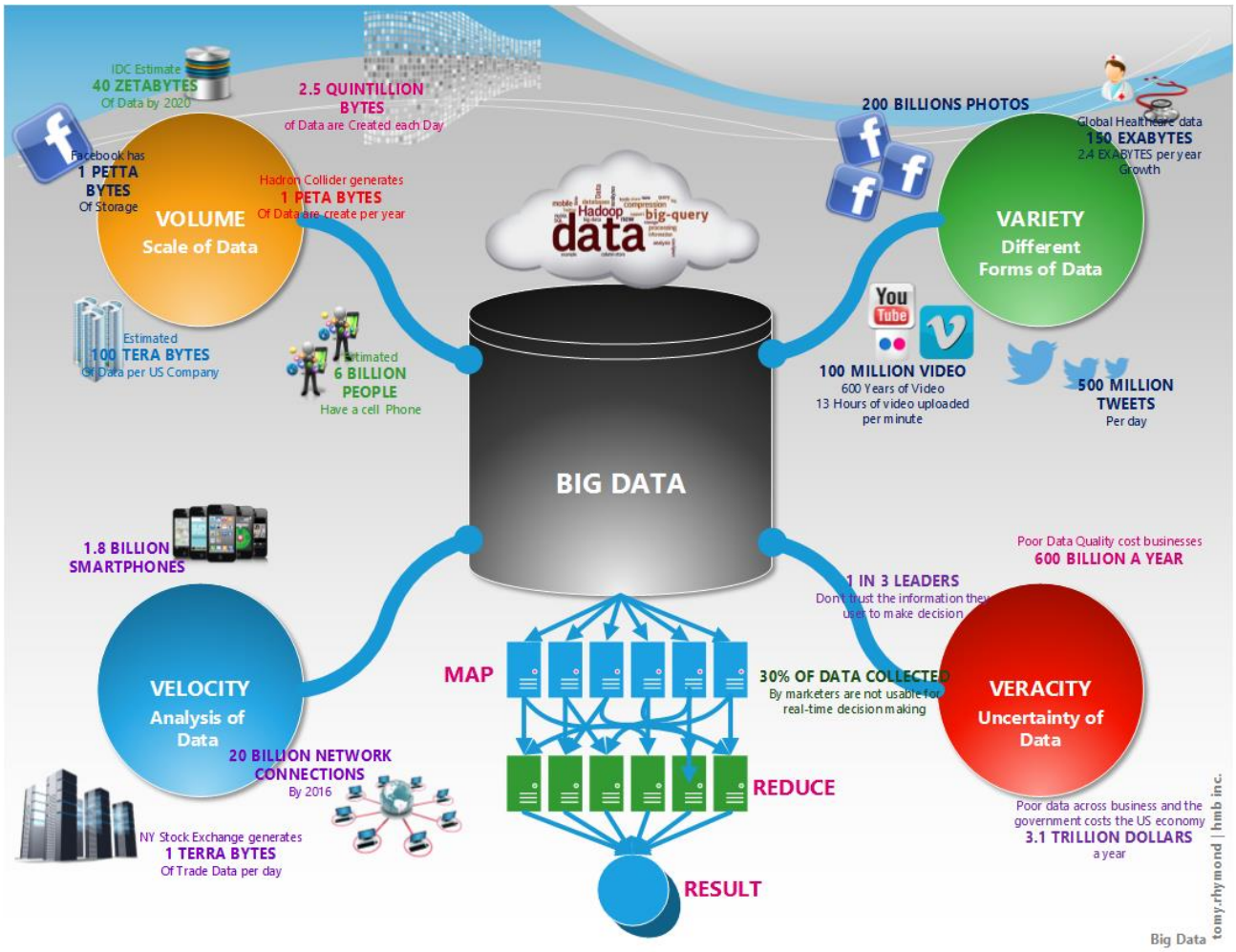
Más que una cuestión de tamaño de datos, es una oportunidad de negocio. Las empresas ya están utilizando *Big Data* para entender el perfil, las necesidades y el sentir de sus clientes respecto a los productos y/o servicios vendidos. Esto adquiere especial relevancia ya que permite adecuar la forma en la que interactúa la empresa con sus clientes y en cómo les prestan servicio.

Asociar *Big Data* con grandes volúmenes de datos no es nuevo. La gran mayoría de las empresas ya llevan mucho tiempo manejando grandes volúmenes de datos y han desarrollado DataWarehouses y potentes herramientas analíticas que les permiten tratar de forma adecuada esos grandes volúmenes. La evolución de la tecnología y los menores costes del almacenamiento han hecho que los volúmenes manejados por estas aplicaciones hayan aumentado de manera muy importante.

### 1.3.3 Las "Vondades" o las 5 'Vs' del *Big Data*

Inicialmente la diferencia entre las aplicaciones analíticas y de gestión y los nuevos conceptos de *Big Data*? se asocian a tres palabras, las tres 'Vs' del *Big Data*: Volumen, Variedad y Velocidad (3Vs). Debido al uso y evolución se ha ampliado la definición original, añadiendo nuevas características como son la Veracidad y Valor del dato dando lugar a las 5 V's:

- Volumen, Variedad, Velocidad, Veracidad y Valor del dato.




En *Big Data* los **<u>volúmenes</u>** superan la capacidad del software habitual para ser manejados y gestionados. Este concepto se encuentra en continuo movimiento porque los avances tecnológicos permiten tratamientos de volúmenes mayores. Cuando hablamos de grandes

volúmenes nos referimos a tratamientos de Terabytes o Petabytes. El concepto de volumen es muy variable y cada día que pasa eleva lo que podemos considerar grandes volúmenes de datos.

**Unidades de Medidas de Almacenamiento**

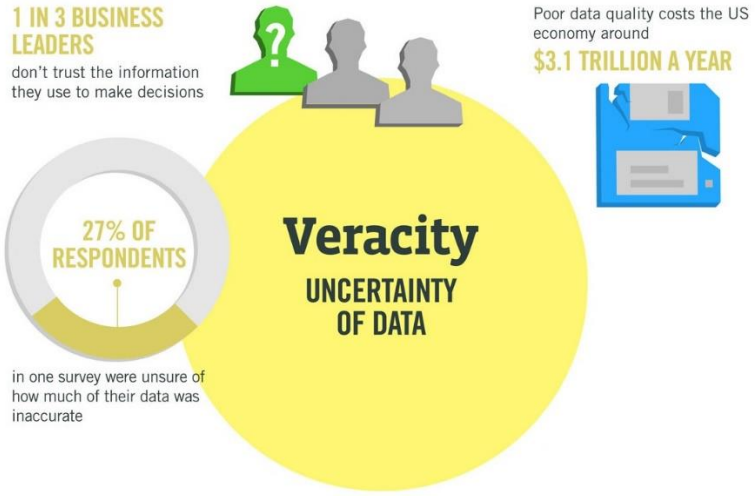
| Medida | Simbologia | Equivalencia | Equivalente en Bytes |
|---|---|---|---|
| byte | b | 8 bits | 1 byte |
| kilobyte | Kb | 1024 bytes | 1 024 bytes |
| megabyte | MB | 1024 KB | 1 048 576 bytes |
| gigabyte | GB | 1024 MB | 1 073 741 824 bytes |
| terabyte | TB | 1024 GB | 1 099 511 627 776 bytes |
| Petabyte | PB | 1024 TB | 1 125 899 906 842 624 bytes |
| Exabyte | EB | 1024 PB | 1 152 921 504 606 846 976 bytes |
| Zetabyte | ZB | 1024 EB | 1 180 591 620 717 411 303 424 bytes |
| Yottabyte | YB | 1024 ZB | 1 208 925 819 614 629 174 706 176 bytes |
| Brontobyte | BB | 1024 YB | 1 237 940 039 285 380 274 899 124 224 bytes |
| Geopbyte | GB | 1024 BB | 1 267 650 600 228 229 401 496 703 205 376 bytes |

En el concepto de **variedad** nos referimos a la inclusión de otros tipos de fuentes de datos diferentes a las que se utilizan de forma tradicional. Nos referimos a información obtenida en diferentes Redes Sociales, en el número cada vez mayor de dispositivos electrónicos conectados, la explotación de sensores que permiten conocer los movimientos y hábitos de vida, de información externa de diversas fuentes, etc.


La información que procesan los Datawarehose es información estructurada que ha pasado por numerosos filtros de calidad para poder garantizar que la información de salida tiene una precisión y una exactitud determinada. Sin embargo, cuando hablamos de *Big Data* nos referimos a información que puede estar semiestructurada o no tener ninguna estructuración. La gestión de esta información desestructurada precisa de una tecnología diferente y permite tomar decisiones basadas en información que tiene importantes grados de inexactitud. Muchos de estos algoritmos se relacionan con los tratamientos de sistemas avanzados de lógica difusa.

El concepto de **velocidad** se refiere a la rapidez con que los datos se reciben, se procesan y se toman decisiones a partir de ellos. A la mayoría de los sistemas tradicionales les es imposible analizar de forma inmediata los grandes volúmenes de datos que les llegan, sin embargo, incorporar el concepto de tiempo real es imprescindible para sistemas de detección del fraude o la realización de oferta personalizadas a los clientes.


Pero, no menos importante al barajar este concepto, es la **veracidad**, esto es, confianza de los datos, extraer datos de calidad eliminado la imprevisibilidad inherente de algunos, como el tiempo, la economía etc, para, de esta forma, llegar a una correcta toma de decisiones

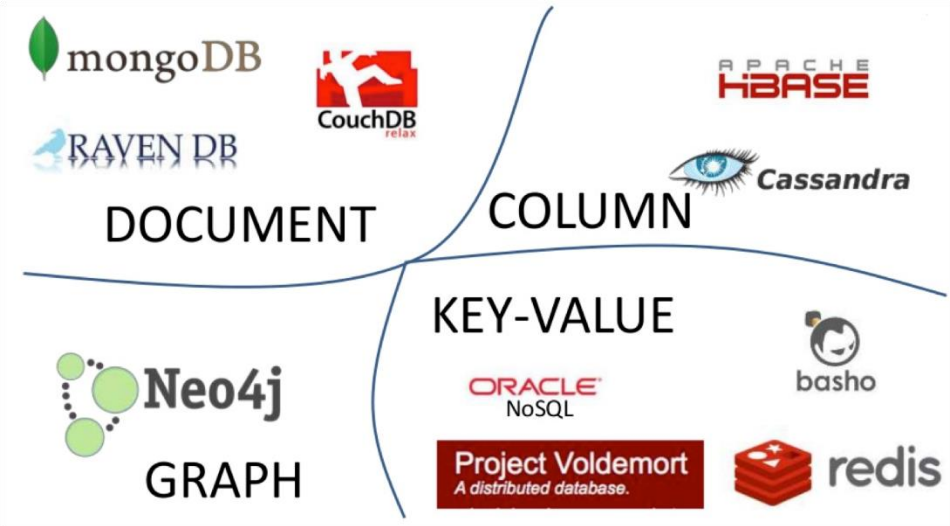



Finalmente, se añade el **valor del dato**. La importancia del dato para el negocio, saber que datos son los que se deben analizar, es fundamental. Tanto que ya se empieza a hablar del científico de datos, un profesional con perfil científico, tecnológico...y visión de negocio

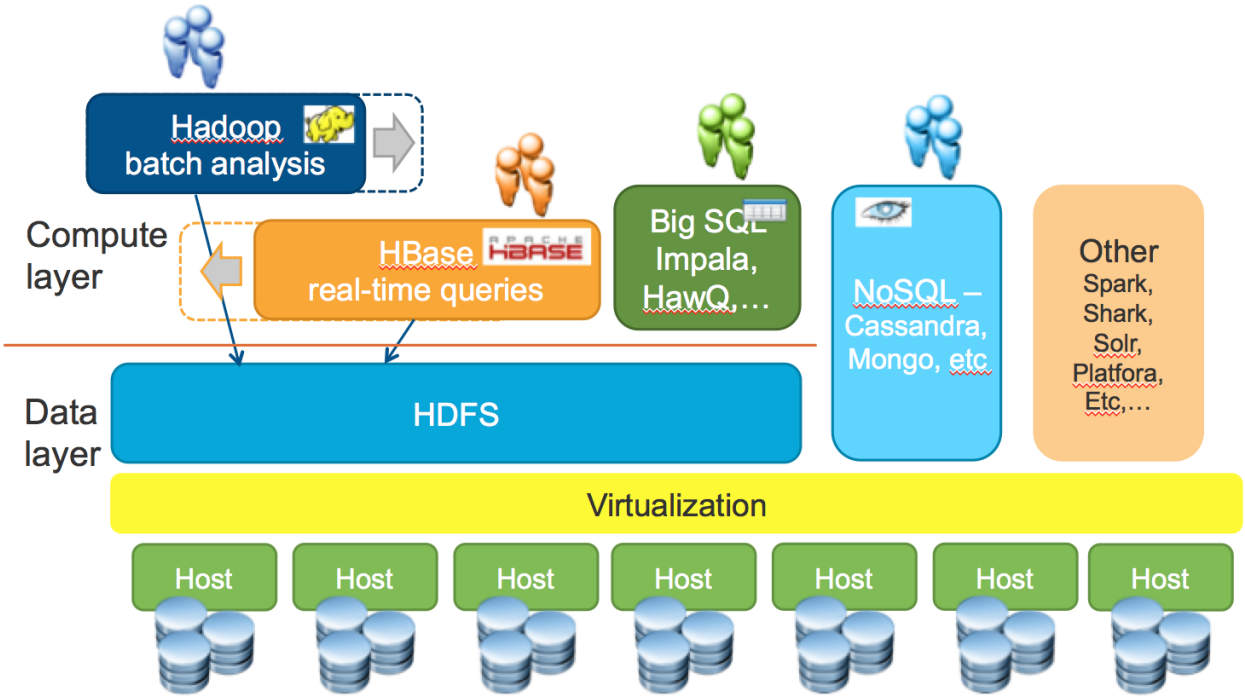
# 1.3.4 Tecnologías vinculadas con Big Data

- No Sql (Not Only Sql) – Video Explicativo (https://www.youtube.com/watch?v=EU8YpTYtZWk)

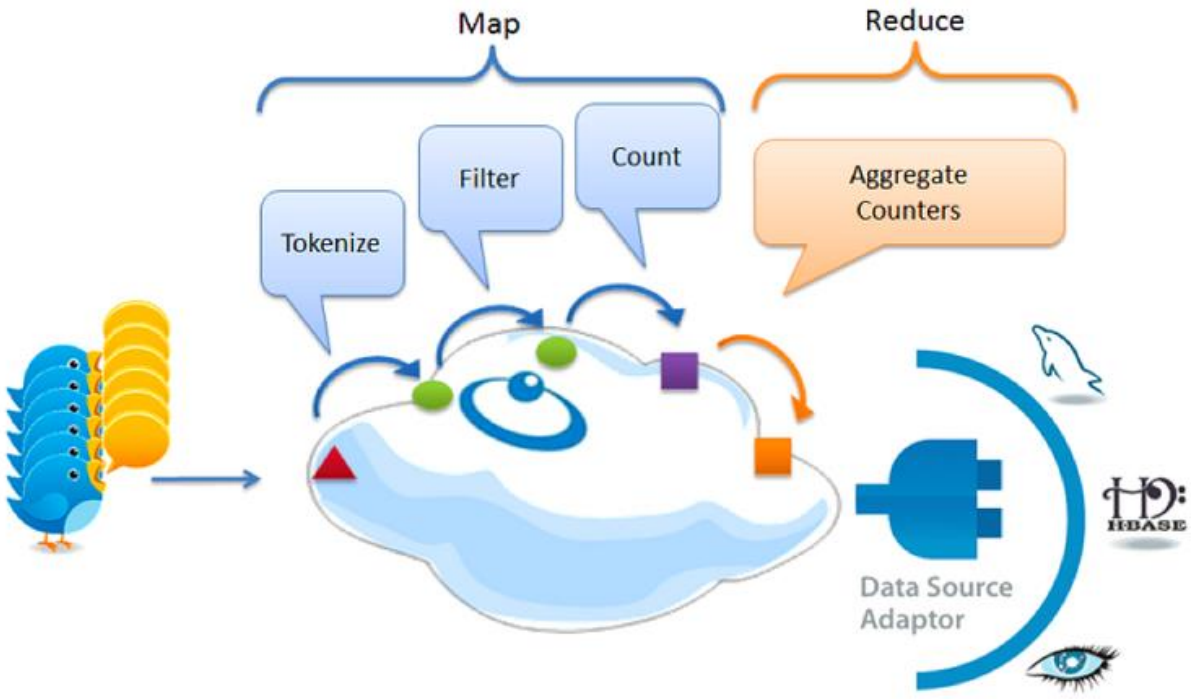

Clasificación de las bases de datos NoSql:

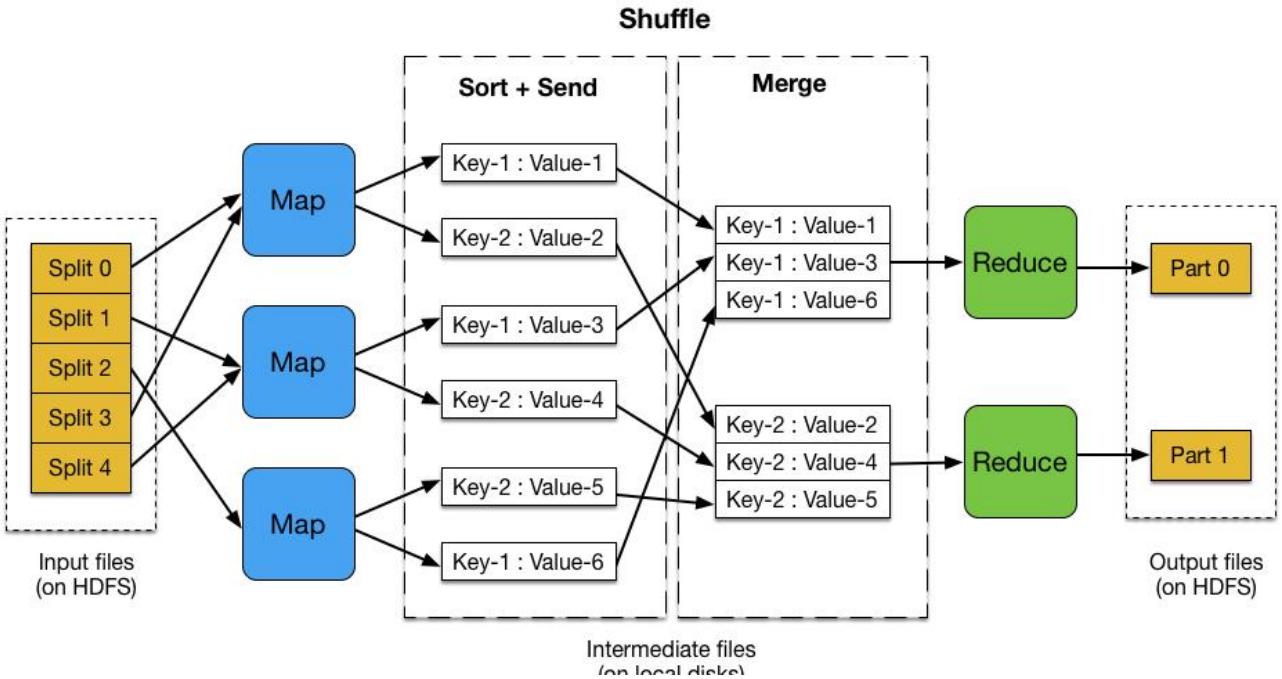


- Base de datos Hadoop como herramienta de Analítica

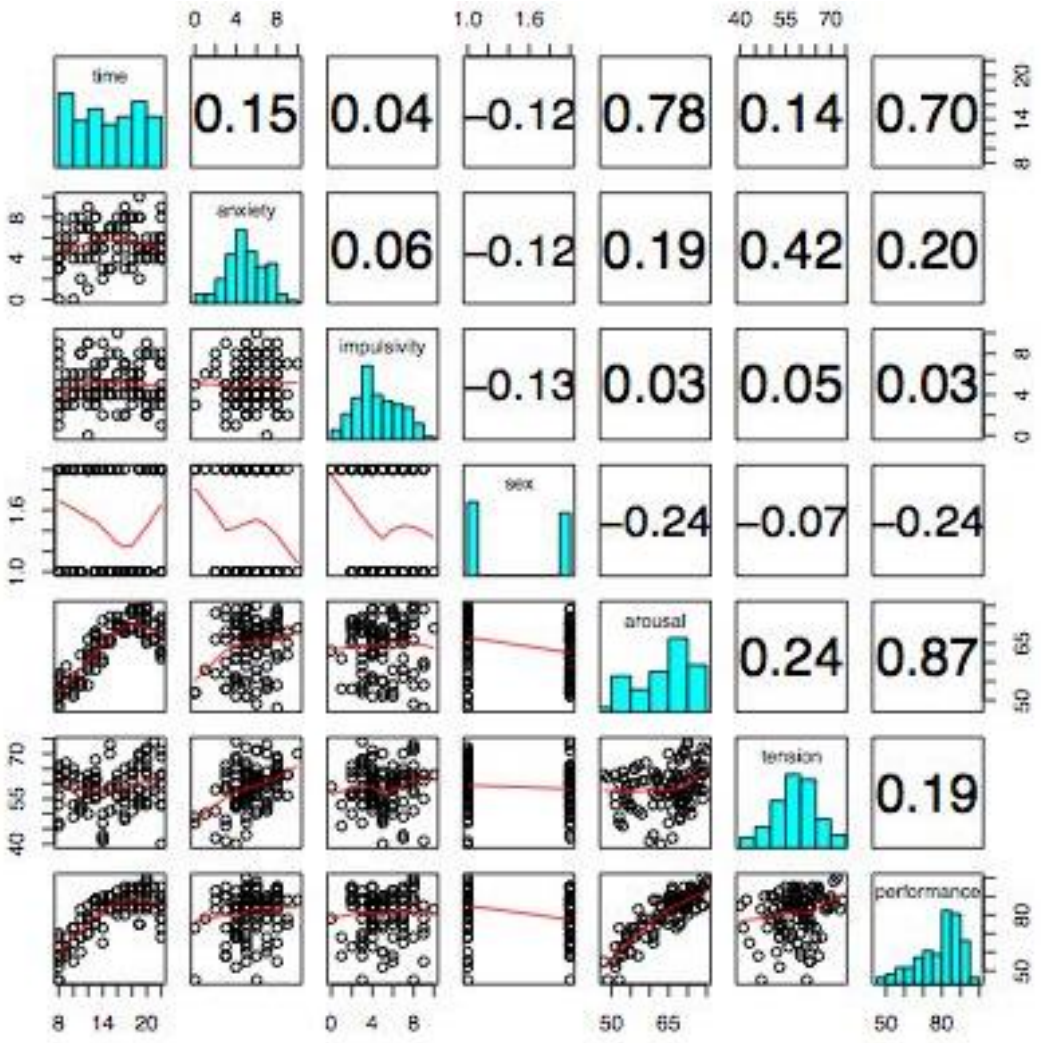


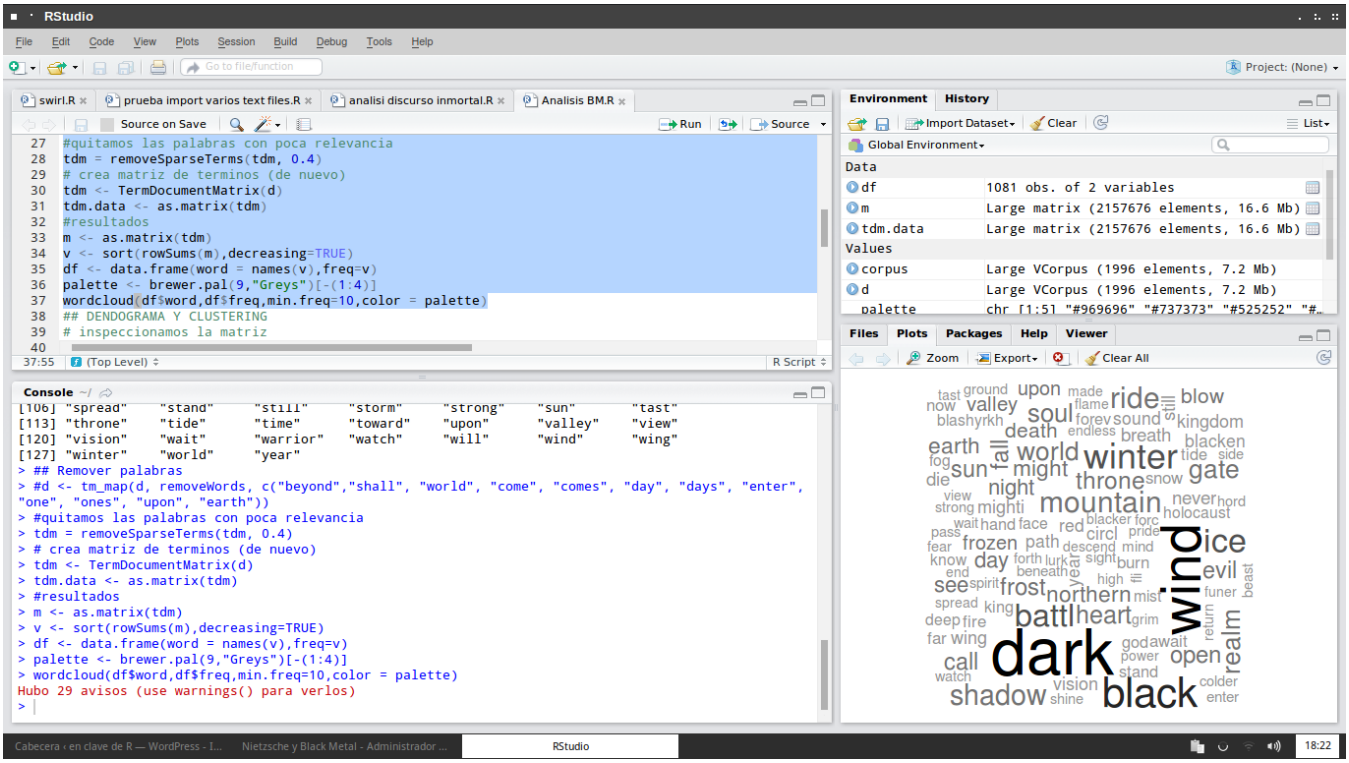

- Tecnología Map Reduce de Hadoop y otras…






- Lenguaje R (lenguaje de programación)






# *Aspectos de futuro de Big Data*

[Tendencias en análisis de datos 2020]

https://discoverthenew.ituser.es/predictive-analytics/2020/02/6-tendencias-que-los-analistas-de-datos-deben-saber-en-2020