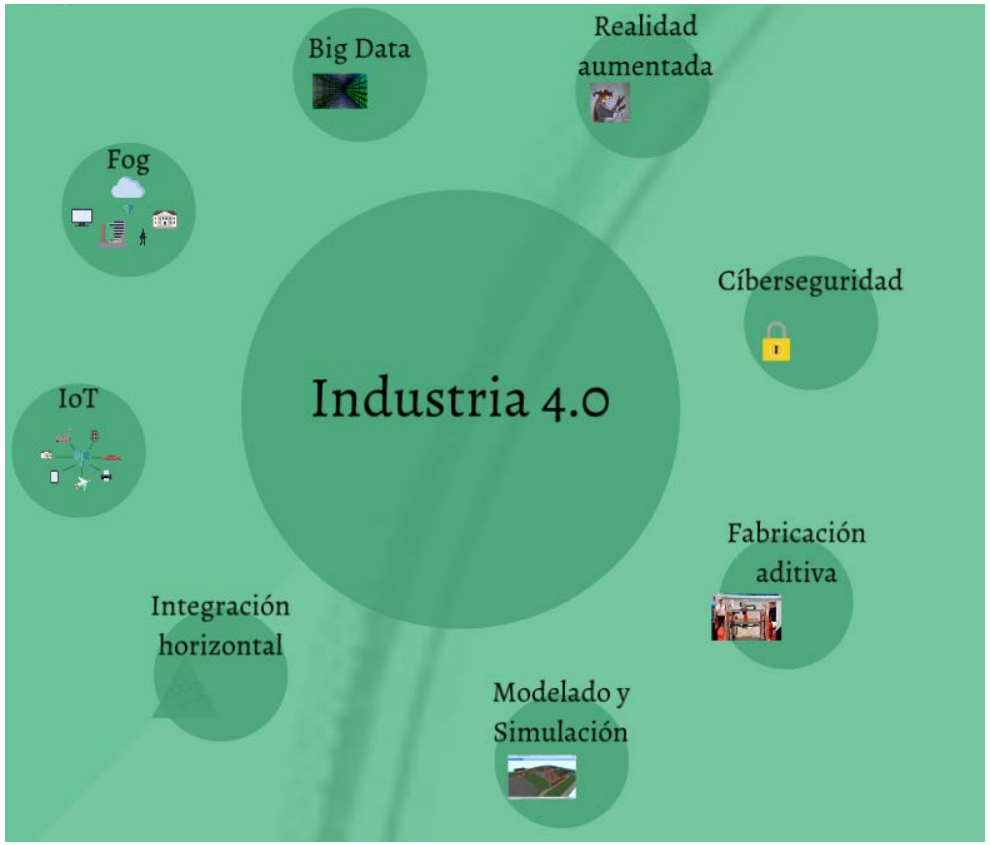
Big Data Analitics

Introducción básica para la asignatura de Informática Industrial

Rogelio Mazaeda
Eusebio de la Fuente

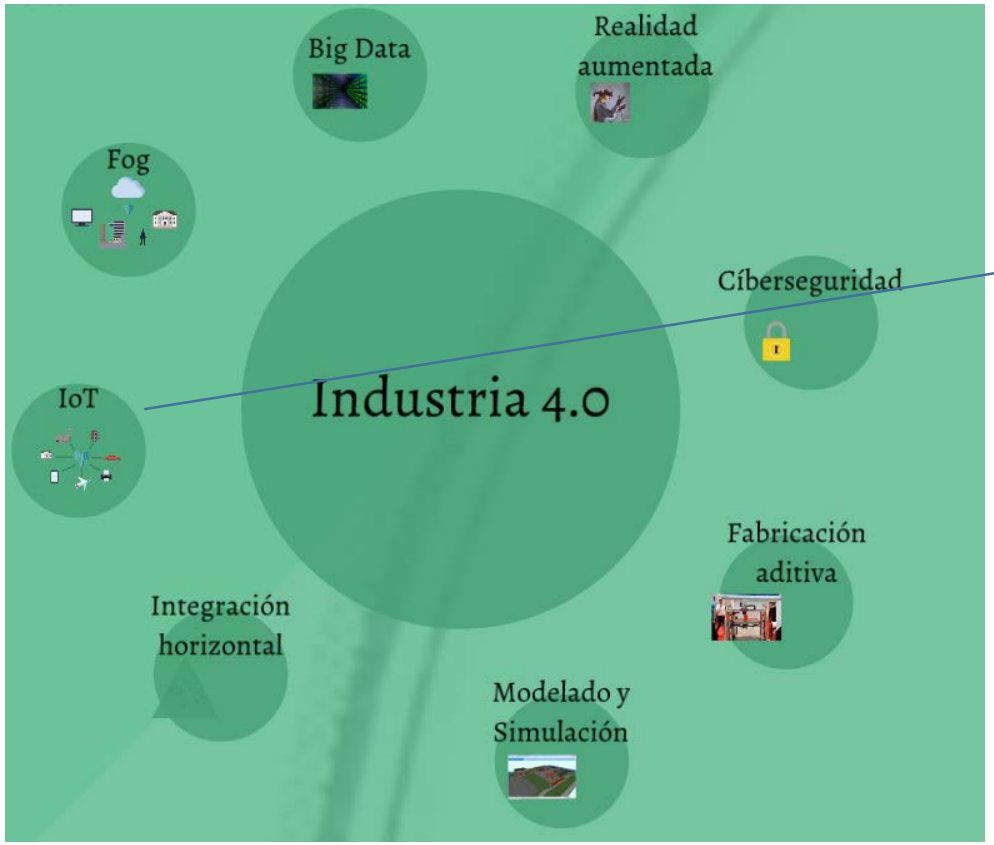
Introducción



Bajo los términos de **sistemas cíber-físicos**, **digitalización**, *smart factories* entre otros, se trata de agrupar la irrupción desde varios años, y de forma acelerada de una serie de tecnologías y paradigmas que prometen modificar de forma decisiva la vida económica y social.
Por supuesto la industria no es ajena a este proceso y se anuncia la llegada de la cuarta revolución industrial o **Industria 4.0**.

Elementos importantes de este proceso es la existencia de los avances tecnológicos que permiten la adquisición y procesamiento eficaz de ingentes cantidades de datos

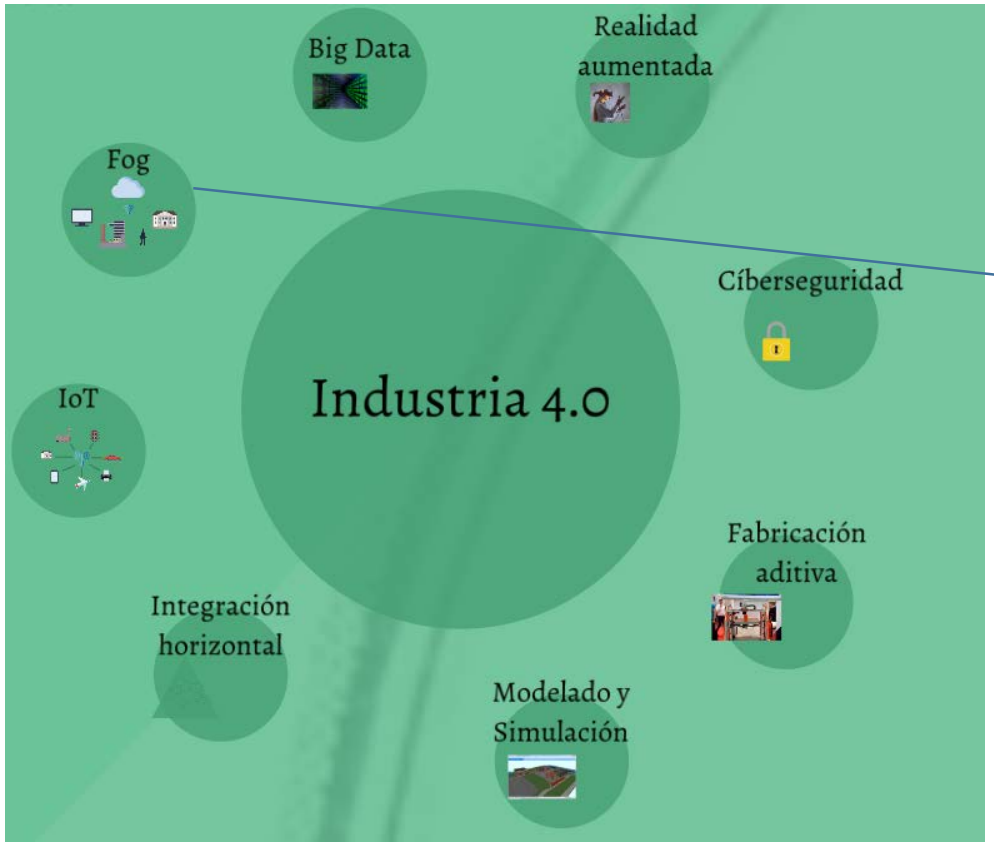
Introducción. Industria 4.0



Un elemento importante de este proceso es la existencia de los avances tecnológicos que permiten la adquisición y procesamiento eficaz de ingentes cantidades de datos:

- ***IoT: Internet de las cosas***. El abaratamiento de los elementos técnicos que permiten la adquisición de datos de forma masiva en lugares y condiciones que antes resultaban prohibitivas y la capacidad de poder ser transmitidas de forma inalámbrica (Wifi, IPV6) a los centros donde puedan ser procesados, de forma segura (cíber-seguridad).

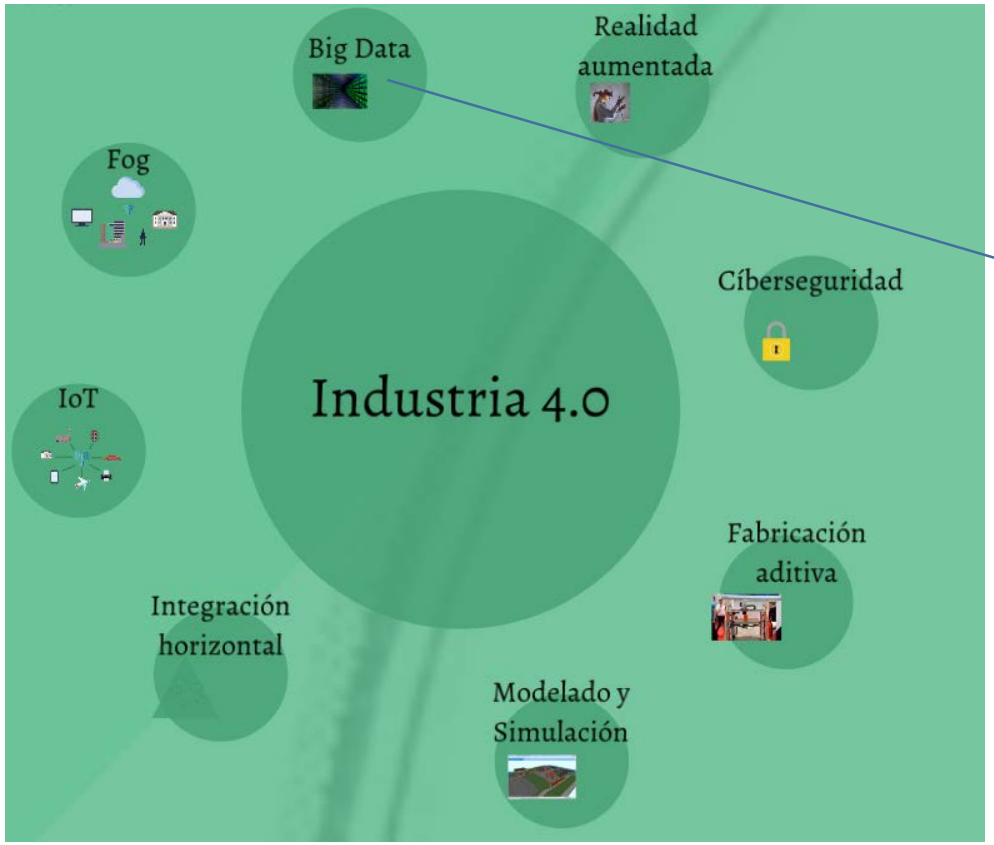
Introducción. Industria 4.0



Un elemento importante de este proceso es la existencia de los avances tecnológicos que permiten la adquisición y procesamiento eficaz de ingentes cantidades de datos:

- ***Procesamiento en la nube (cloud or fog computing): La creación de nuevos sistemas de negocio consistentes en brindar a las medianas y pequeñas empresas la posibilidad de satisfacer sus de procesamiento de datos a empresas especializadas de manera que resulte económicamente factible la utilización, a demanda, de los ingentes recursos de hardware y software requeridos.***

Introducción. Industria 4.0



Un elemento importante de este proceso es la existencia de los avances tecnológicos que permiten la adquisición y procesamiento eficaz de ingentes cantidades de datos:

- ***Big Data: Las emergencia de técnicas eficaces de almacenamiento y procesamiento de enormes cantidades de datos, obtenidos de diferentes fuentes y de muy distintos formatos en tiempo real.***
- ***La posibilidad de extraer información útil para la ayuda a la decisión de estos datos.***
- ***El desarrollo y relanzamiento de métodos basados en datos (machine learning)***

Big Data. Contexto.

El crecimiento de los datos disponibles electrónicamente se prevé explosivo. Se hacen predicciones que los datos de las empresas llegará el 2020 a 46 zettabytes (10^{21}) y se espera que crezca 40% anualmente. Claramente, el manejo eficiente de esos volúmenes de datos y la extracción de información útil de los mismos requiere de cambios muy profundos en los paradigmas, las tecnologías, y los algoritmos utilizados.

Fuentes del Big Data:

- IoT: el más importante desde el punto de vista industrial
- Medios Sociales (e-correo, facebook, etc).
- Información multimedia.

Desarrollo histórico

Aplicaciones en solitario	Aplicaciones en redes locales	Aplicaciones web	Aplicaciones complejas en el cliente	Big Data
Datos y algoritmos en ordenadores Individuales. Software de contabilidad	Datos disponibles en Servidores conectados Por redes locales. Aplicaciones de bases de datos relacionales (SQL)	Irrupción de Internet Y aplicaciones clientes-servidor web. Bases de Datos NoSQL.	Clientes más potentes. Utilizables en línea y Fuera de línea. Cloud computing	Datos de diversas fuentes. Haddoop, Map-reduce

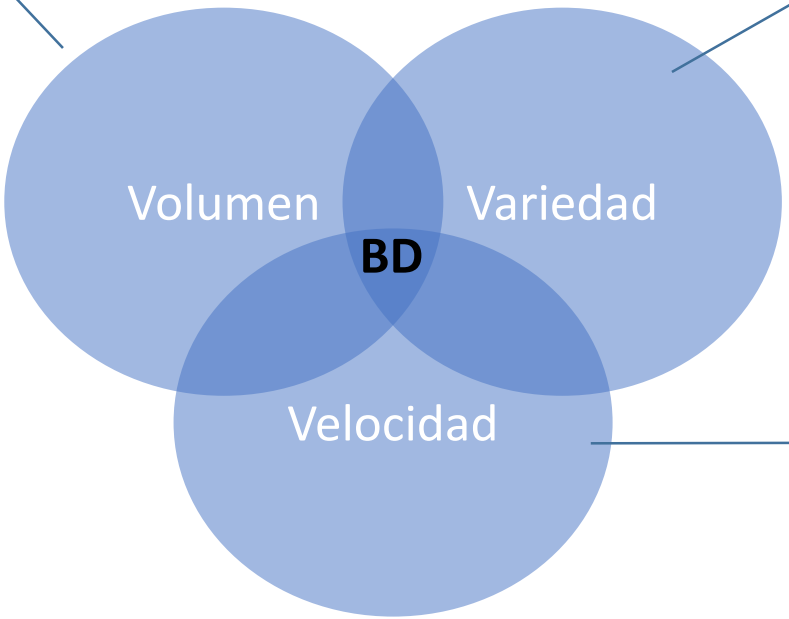
I. Yaqoob *et al.*, “Big data: From beginning to future,” *Int. J. Inf. Manage.*, vol. 36, no. 6, pp. 1231–1247, 2016.

Big Data. Definición

El Big Data actualmente se define a partir de sus atributos:

Definición de 3 V's de Big Data (BD)

Se refiere a la enorme cantidad de datos que llegan por unidad de tiempo así como a los datos que se acumulan en los medios de almacenamiento



Diversidad de formatos y estructuras diferentes y en principio incompatibles

La velocidad a la que se reciben los datos y la que se requiere al interactuar con ellos en la ayuda del proceso de decisión

R. Buyya, R. N. Calheiros, and A. V. Dastjerdi, *Big Data: Principles and Paradigms*. 2016.

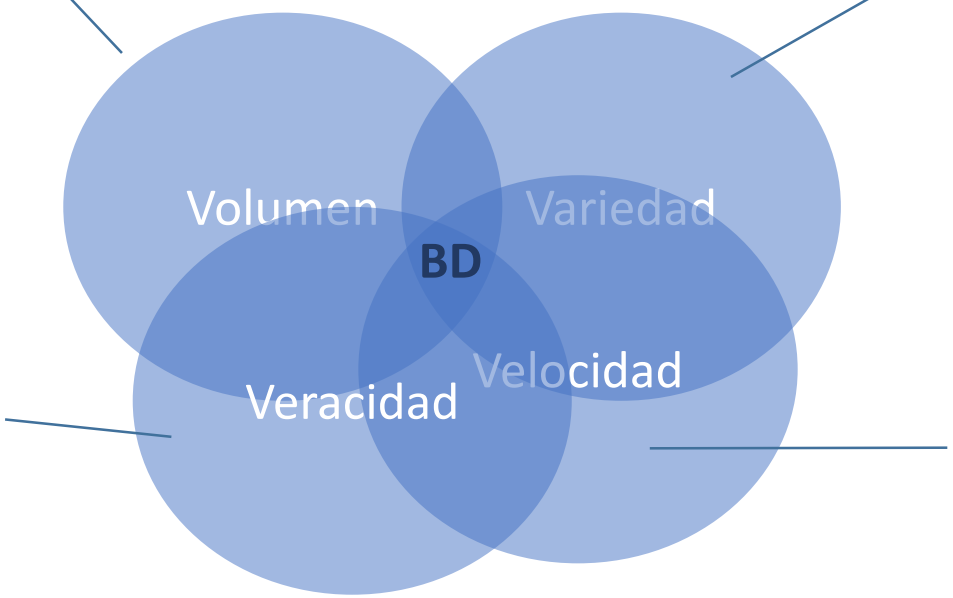
Big Data. Definición

Continuan añadiendose atributos (manteniendo las V's). Variabilidad (se refiere a la complejidad de los datos más que a la diversidad de formatos), Visibilidad (necesidad de mostrar una visión significativa de los datos que ayude a la decisión)

Definición de 4 V's de Big Data (BD)

Se refiere a la enorme cantidad de datos que llegan por unidad de tiempo así como a los datos que se acumulan en los medios de almacenamiento

Diversidad de formatos y estructuras diferentes y en principio incompatibles



La veracidad de los datos es cuestionable. Los datos arriban con gran velocidad y de diferentes fuentes, frecuentemente con alto nivel de redundancia pero plagado de incertidumbres

La velocidad a la que se reciben los datos y la que se requiere al interactuar con ellos en la ayuda del proceso de decisión

R. Buyya, R. N. Calheiros, and A. V. Dastjerdi, *Big Data: Principles and Paradigms*. 2016.

Big Data (BD) vs. Big Data Analitics (BDA)

BD hace referencia a lo que se entiende por el concepto, mientras que **BDA** hace énfasis en lo que se puede lograr, desde el punto de vista pragmático con esos datos al convertirlos en información sobre la que se pueda actuar. Se trata de encontrar patrones de "significado" en grandes volúmenes de datos. Otros términos relacionados: Minería de datos. Relación directa con la estadística.

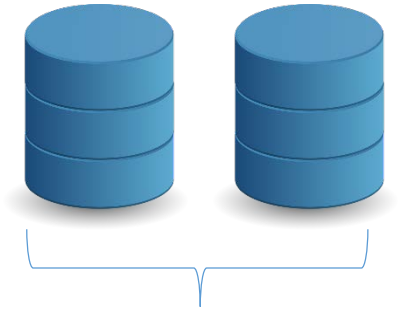
Elementos de BDA:

- **Herramientas de Extracción Transformación y Carga** (ETL: ExtractTransformLoad) con cantidades masivas de datos.
- ***Machine Learning***: Definida como la capacidad de dotar a los ordenadores de la capacidad de "aprender" sin haber sido explícitamente programados para hacerlo en ningún campo específico. Se trata de extraer "información" a partir de los "datos".
- ***Cloud computing***: Modelo de negocio que permite a las pequeñas empreses externalizar, a un coste asumible, los gastos de los dos elementos anteriores, que de lo contrario serían inasumibles.

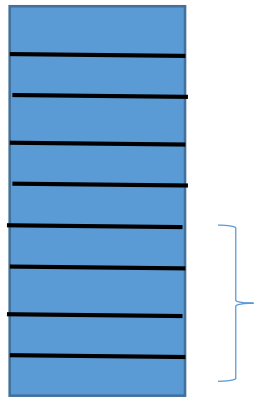
Se trata, en todo caso, de aprovechar la oportunidad y, al mismo tiempo, superar los retos que implican la existencia de enormes volúmenes de datos de diversas fuentes (internas y externas a la organización), diferentes formatos, grado de fiabilidad que crecen a una ritmo mucho mayor que el de las capacidades de almacenamiento y de velocidad de procedimiento de los ordenadores, obteniendo información útil de estos datos de una forma que sea costeable para la empresa.

Big Data. Tecnologías

Big Data



Modo *Batch*:
Acceso a todos los
Datos simultáneamente
Ej: Apache Hadoop



Modo *stream*:
Procesa datos en tiempo
Real, en la medida en que van
Llegando.
Ej: Storm

Big Data. Métodos de procesamiento

¿Cómo sacar información valiosa de forma rápida de grandes cantidades de datos?

- **Indexado**: Localizar información de forma eficiente en grandes volúmenes de datos.
- **Tablas hash**: Una búsqueda por índice puede ser ineficiente puesto que necesita la consideración de la base de datos completa. El uso de la técnica hash, una función que pone en correspondencia datos con una clave, para distribuirlos de forma eficiente en varios reservorios (buckets) pre-especificados, constituye una alternativa a considerar.
- **Filtro Bloom**: Determina si un elemento pertenece a un conjunto. Puede dar falsos positivos pero nunca falsos negativos.
- **Computación en paralelo**: El tratamiento de la información se puede acelerar, distribuyendo la tarea en varios segmentos y asignando estos a ordenadores separados.

Big Data Analitics (BDA). Técnicas de análisis de datos

Si la manipulación de los datos es una tarea exigente, aún lo es más el extraer información útil de esos datos. Algunas técnicas utilizadas.

- **Minería de Datos**: Regresión, clasificación, análisis de clusters, regla de asociación del aprendizaje, etc. Utiliza métodos de *machine learning* o de la estadística tradicional para extraer información de los datos. Los métodos tradicionales deben ser constantemente revisados y mejorados para adaptarlos a las demandas de volúmenes y velocidades de procesamiento cada vez mayores.
- **Minería de Web**: Descubre patrones en repositorios Web.
- **Métodos de Visualización**: Es un reto el cómo representar datos de muchas dimensiones de forma gráfica con el objetivo de poder extraer conclusiones acertadas de dicha representación.
- **Machine Learning**: Algoritmo supervisados, no supervisado o de aprendizaje reforzado para extraer información de los datos existentes. Mucho por avanzar a la hora de aplicarlo a BD.
- **Métodos de Optimización**: Optimización numérica matemática.

Retos de aplicación de BD

Retos en la aplicación del BD han sido identificados*:

- **Manipulación de BD:** Hay que hacerlo con eficiencia y es la clave de todo lo demás. Incluye procesos conocidos (limpieza, agregación, codificación, almacenamiento y acceso) desde siempre, complicado en BD por las 3(4) V's y la necesidad de acudir a sistemas distribuidos, con muchas aplicaciones de diferente naturaleza. Ej:
  - Limpieza de BD: ¿Cómo garantizar la fiabilidad de datos masivos? ¿Cómo detectar ruido, errores, incosistencias?
  - Agregación: Sincronizar datos internos y externos de la empresa, en aplicaciones distribuidas con multitud de formatos. Ej: datos de producción disponibles de sensores, con datos externos de predicción del tiempo o de demanda de mercado con satisfacción de los consumidores, etc.
- **BDA:** Se necesitan algoritmos nuevos, eficientes para extraer patrones y descubrir relaciones "ocultas" en los datos, que pueda brindar nuevas oportunidades o descubrir a tiempo potenciales peligros. Los retos: gran cantidad de información heterogénea versus necesidad de rápida respuesta:
  - Machine Learning (ML): Incorporar ML en un escenario de Data Stream (tiempo real): con limitaciones de almacenamiento y necesidad de responder en tiempo adecuado (nótese que las necesidades de tiempo real se relajan con respecto a lo comúnmente aceptado. Uso de Aprendizaje Profundo (DL): ha mostrado su superioridad sobre las tradicionales redes neuronales (ANN) sin embargo requieren el uso de muchos datos para el entrenamiento, ajuste de miles de millones de parámetros. Uso del aprendizaje incremental (data streams).

A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *J. King Saud Univ. - Comput. Inf. Sci.*, 2017.

El sistema Apache Hadoop

Apache Hadoop is una reconocida tecnología de BD*:

- **Sistema de ficheros ditribuido (HDFS):** No trae a memoria los datos distribuidos en otro ordenador sino que los procesa in situ liberando capacidad de tx. en la red.
- **Fiable en presencia de distribución de datos**: Tolerante a fallos en sistemas distribuidos permitiendo la replicación de los datos.
- **Basado en modelo de programación Map-Reduce**: Permite tratamiento efectivo de BD haciendo posible procesamiento masivo en paralelo.
- **Sistema abierto y modular**: Las contribuciones de los usuarios lo convierte en un ecosistema vivo y dinámico.

A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” *J. King Saud Univ. - Comput. Inf. Sci.*, 2017.

El sistema Apache Hadoop: Map-Reduce

Modelo programación Map-reduce*:

- **Map**:
  - Divide una tarea de gran tamaño en un conjunto de sub-tareas, creando una participación que se representa como {llave, valor}.
  - Se envía {llave, valor} a ser procesado de forma individual en los procesos Mapper distribuidos en varios ordenadores. Cada ordenador recibe una partición, realiza el procesamiento indicado y devuelve uno o más pares {llave, valor} intermedios.
  - Se recolectan y se ordenan por llave los anteriores pares.
- **Reduce**: Por cada llave, la función de reducción, agrega los valores asociados a esa llave según un procedimiento predefinido (ej: resumir, ordenar, promedio, etc). Produce uno o más pares {llave, valor}

A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” *J. King Saud Univ. - Comput. Inf. Sci.*, 2017.