**BIG DATA**

1. **Identificar elementos o características importantes de BIG DATA**

**Características:**

1. **Volumen:** El volumen se refiere a la gran cantidad de datos que se crean cada segundo en comparación con las fuentes de datos tradicionales.
2. **Variedad:** La variedad se refiere a que los datos provienen de diferentes fuentes y están siendo creados tanto por máquinas como por personas.
3. **Velocidad:** La velocidad se refiere a la velocidad de los datos que se generan extremadamente rápido; un proceso que nunca se detiene ni siquiera mientras dormimos.
4. **Veracidad:** La veracidad se refiere a Big Data que se obtiene de muchos lugares diferentes; como resultado, debe probar la veracidad / calidad de los datos.
5. [**Viabilidad**](https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/#viabilidad)**:** La inteligencia empresarial es un componente fundamental para la viabilidad de un proyecto y el éxito empresarial. Se trata de la capacidad que tienen las compañías en generar un uso eficaz del gran volumen de datos que manejan.
6. [**Visualización de los datos**](https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/#visualizacion-datos)**:** Cuando hablamos de visualización nos referimos al modo en el que los datos son presentados. Una vez que los datos son procesados (los datos están en tablas y hojas de cálculo), necesitamos representarlos visualmente de manera que sean legibles y accesibles, para encontrar patrones y claves ocultas en el tema a investigar.
7. **Valor:** Está muy bien tener acceso a Big Data, pero si no podemos generar valor a partir de los datos, se vuelve absolutamente inútil. Por lo tanto, es lo más importante en Big Data y debemos tener una comprensión clara de los costos y beneficios de Big Data antes de iniciar cualquier iniciativa relacionada en esta área.

**Elementos:**

* **Fuentes:** las más habituales son los registros históricos de la compañía, los almacenes de datos, los dispositivos inteligentes, los sistemas de gestión de datos, Internet y el Internet de las Cosas.
* **Capa de almacenamiento:** su función es la de recoger y transformar los datos sin perder de vista la normativa legal. Además, tiene que dar acceso a los datos independientemente de su formato, volumen, frecuencia u origen.
* **Capa de análisis:** se encarga de leer los datos almacenados. Mediante la utilización de los modelos, los algoritmos y las herramientas adecuadas, proporciona visibilidad sobre los datos para que puedan ser consultados en la capa de consumo.
* **Capa de consumo:** son muchos los proyectos y usuarios que se benefician del conocimiento extraído en todo este proceso. La forma de consumir los datos dependerá del destinatario, pero será habitual verlos en forma de *reporting* o visualización en tiempo real.

1. **Revisar la herramienta Apache Hadoop**

**Apache Hadoop**es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial. **Hadoop** hace posible a las organizaciones obtener conocimiento rápidamente a partir de cantidades masivas de datos, estructurados y no estructurados, posicionándolas al nivel de las exigencias actuales de los mercados en términos de dinamismo y capacidad.

El ecosistema Hadoop cuenta con soluciones de todo tipo para cubrir cualquier necesidad que pueda presentarse al negocio en materia de:

1. Gestión de datos.
2. Acceso a los datos.
3. Gobernabilidad e integración de datos.
4. Seguridad de la información.
5. Operaciones.
6. **Explicar el concepto de data warehouse**

**Data warehouse**

Es una colección de datos específicos de la empresa que se extraen de varias funciones organizativas y luego se hacen para ajustarse a una estructura estándar **de forma segura, fiable, fácil de recuperar y fácil de administrar**.

**La arquitectura de un data warehouse puede ser dividida en tres estructuras simplificadas:**

* **Con una estructura básica**, sistemas operativos y archivos planos proporcionan datos en bruto que se almacenan junto con [metadatos](https://www.powerdata.es/el-valor-de-la-gestion-de-datos/la-importancia-de-los-metadatos-en-el-gobierno-de-datos). Los usuarios finales pueden acceder a ellos para su análisis, generación de informes y minería.
* **Al añadir un área de ensayo** que se puede colocar entre las fuentes de datos y el almacén, ésta proporciona un lugar donde los datos se pueden limpiar antes de entrar en el almacén. Es posible personalizar la arquitectura del almacén para diferentes grupos dentro de la organización.
* **Se puede hacer agregando data marts**, que son sistemas diseñados para una línea de negocio en particular. Se pueden tener data marts separados para ventas, inventario y compras, por ejemplo, y los usuarios finales pueden acceder a datos de uno o de todos los data marts del departamento.

1. **Resumir el concepto de datalake**

Data Lake es un repositorio compartido que le permite adquirir y almacenar grandes cantidades de datos procedentes de sistemas heterogéneos en formato nativo, es decir, datos en bruto estructurados, semiestructurados y no estructurados.

|  | **DATA WAREHOUSE** | **DATA WAREHOUSE** |
| --- | --- | --- |
| **Estructura de los datos** | *Brutos (estructurados, semiestructurados y no estructurados)* | *Estructurados, procesados* |
| **Finalidad de los datos** | *Por definir, definida* ***Nota: Es posible que haya datos cuyo propósito no se haya definido (para uso futuro)*** | *Definida* |
| **Esquema** | *On Read* | *On Write* |
| **Usuarios** | *Data Scientists* | *Usuarios empresariales* |
| **Accessibilidad** | *Gran accesibilidad y fácil actualización* | *Acceso y actualizaciones más complicadas y costosas* |
| **Almacenamiento** | *Almacenamiento distribuido y costes limitados (potencialmente ampliable a la nube)* | *Costes y revisión de costosos procesos de ingesta* |

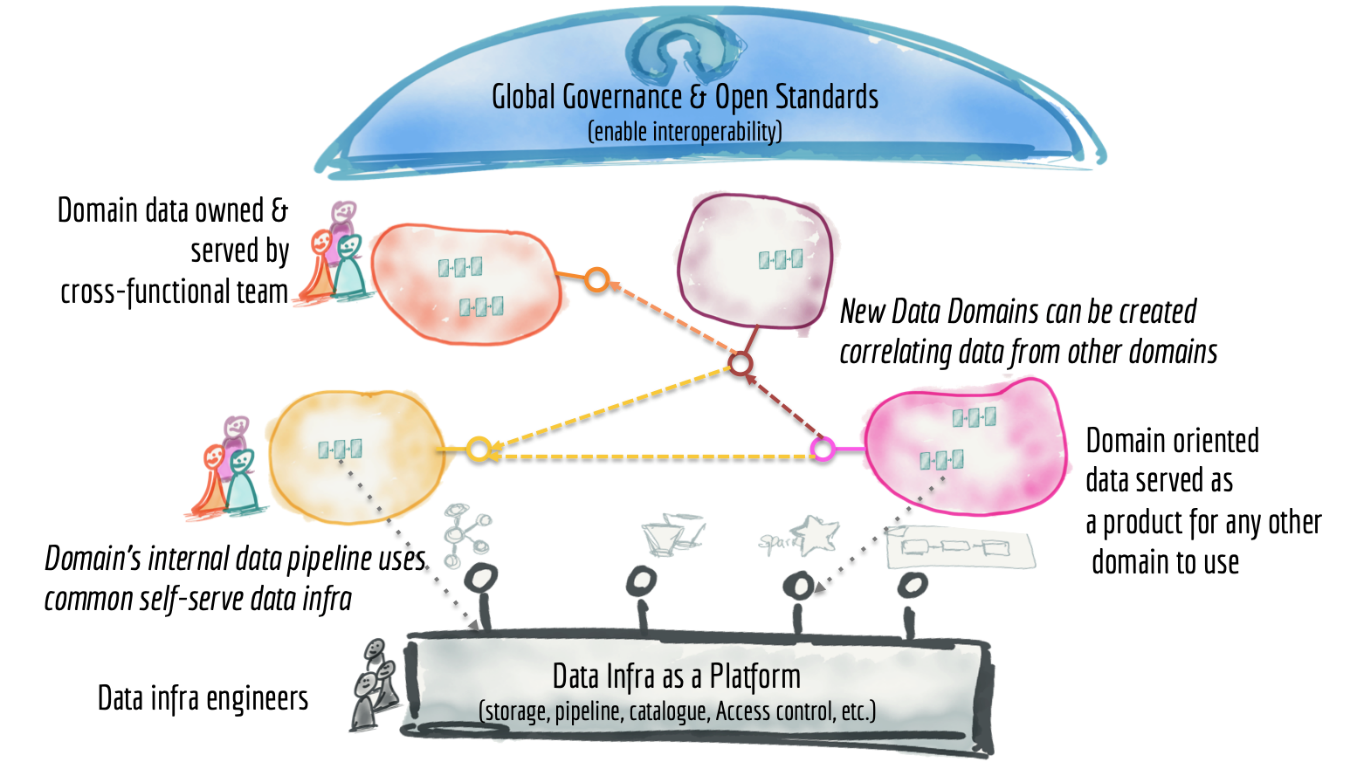
1. **Identificar la propuesta de "Distributed data mesh"**

**Data mesh**

Es un paradigma arquitectónico que desbloquea datos analíticos a escala; desbloquear rápidamente el acceso a un número cada vez mayor de conjuntos de datos de dominio distribuidos, para una proliferación de escenarios de consumo, como aprendizaje automático, análisis o aplicaciones de datos intensivos en toda la organización.

**Data mesh** aborda los modos de falla comunes del[**lago de datos**](https://martinfowler.com/bliki/DataLake.html) centralizado tradicional o arquitectura de plataforma de datos, con un cambio del paradigma centralizado de un lago, o su predecesor, el almacén de datos.

Como lo definió por primera vez Zhamak Dehghani, **Data mesh** es un tipo de arquitectura de plataforma de datos que abarca la ubicuidad de los datos en la empresa al aprovechar un diseño de autoservicio orientado al dominio.



**Casos de uso de Data Mesh**

Con una arquitectura de malla de datos, pueden:

* crear catálogos de datos virtuales a partir de una variedad de fuentes de datos.
* También puede crear lagos y almacenes de datos virtuales para análisis y capacitación en aprendizaje automático y, quizás lo más importante,
* podrá conectar aplicaciones en la nube a datos confidenciales que se encuentran en las instalaciones y / o en transmisión o datos en tiempo real de los dispositivos.
* Además, los desarrolladores de aplicaciones y los equipos de DevOps podrán consultar datos de una variedad de almacenes de datos sin tener que preocuparse por cómo acceden a estos datos.

**Conclusión:**

Una malla de datos permite a su organización escapar de los límites analíticos y de consumo de las arquitecturas de datos monolíticas y conecta los datos en silos para permitir el aprendizaje automático y el análisis automatizado a escala.

Con una malla de datos, su empresa estará realmente impulsada por los datos, renunciando a los problemas de los lagos y almacenes y reemplazándolos con el poder del acceso, el control y la conectividad de los datos.