- **Análisis de *streaming*.** Pueden utilizarse para el análisis en tiempo real de datos de *streaming*, lo que permite la detección de patrones y la toma de decisiones en tiempo real.

Los ***data lakes*** son fundamentales en el ecosistema de datos moderno, ya que permiten a las organizaciones gestionar y analizar grandes volúmenes de datos de manera eficiente y flexible. Proporcionan una infraestructura escalable y económica para el almacenamiento de datos, así como capacidades avanzadas de análisis y exploración de datos, lo que les permite aprovechar al máximo su información y tomar decisiones basadas en datos de manera informada y estratégica. Seguidamente, examinemos las ventajas y las desventajas de un *data lake*.

Por el lado de las **ventajas** tenemos:

- **Almacenamiento sin procesar.** Permite almacenar datos en su forma original, sin procesar ni estructurar previamente, lo que proporciona una mayor flexibilidad y diversidad de datos.
- **Escalabilidad.** Los *data lake*s pueden manejar grandes volúmenes de datos, desde terabytes hasta petabytes o más, escalando horizontalmente según sea necesario.
- **Diversidad de datos.** Pueden almacenar una amplia variedad de tipos de datos, incluyendo datos estructurados, semiestructurados y no estructurados, así como datos en tiempo real y datos históricos.
- **Bajo costo.** Utilizan un enfoque de almacenamiento económico, ya que no requieren una estructura predefinida y pueden utilizar almacenamiento en la nube u otras soluciones de almacenamiento de bajo costo.
- **Facilidad de acceso.** Proporcionan capacidades de búsqueda y acceso flexible a los datos, lo que permite a los usuarios explorar y analizar datos de manera ágil y sin restricciones.

- **Análisis avanzado.** Facilitan el análisis avanzado de datos a través de herramientas y plataformas de análisis de datos, lo que permite a las organizaciones extraer información valiosa y descubrir patrones ocultos en los datos.

Como **desventajas** tenemos:

- **Gestión de la calidad de los datos.** Debido a la diversidad y volumen de datos almacenados en un *data lake*, la gestión de la calidad de los datos puede ser un desafío, ya que es necesario garantizar la integridad, precisión y consistencia de los datos.

- **Seguridad y privacidad.** Los *data lake*s pueden contener datos sensibles o confidenciales, lo que plantea preocupaciones sobre la seguridad y la privacidad de los datos. Es importante implementar medidas de seguridad robustas, como cifrado de datos, control de acceso y monitorización de actividades, para proteger los datos almacenados en el *data lake*.

- **Complejidad de la gestión.** La gestión de un *data lake* puede ser compleja y requerir habilidades técnicas especializadas, e incluir la integración de datos, la optimización del rendimiento y la administración de la infraestructura de almacenamiento.

- **Riesgo de convertirse en un *data swamp*.** Si no se gestionan adecuadamente, los *data lake*s pueden correr el riesgo de convertirse en un *data swamp*, donde los datos se acumulan sin un propósito claro o sin ser utilizados de manera efectiva. Es importante establecer políticas y procedimientos para la gestión y gobernanza de datos con el fin de evitar este escenario.

En resumen, aunque los ***data lakes*** ofrecen muchas ventajas en términos de flexibilidad, escalabilidad y análisis avanzado de datos, también presentan desafíos en términos de gestión de la calidad de los datos, seguridad y privacidad, complejidad de la gestión y riesgo de convertirse en un *data swamp*. Es importante abordar estos desafíos de manera proactiva para maximizar el valor de un *data lake* y garantizar su éxito en la organización.

Data warehouse

Un ***data warehouse*** es una base de datos diseñada específicamente para análisis y *reporting*, donde los datos son organizados y estructurados de manera que facilitan la consulta y el análisis. Utiliza un modelo dimensional o en estrella, que consiste en hechos (medidas cuantificables) que están rodeados por dimensiones (atributos descriptivos). Esto permite a los usuarios realizar consultas analíticas complejas y generar informes detallados sobre el rendimiento y las tendencias de negocio.

El **concepto** de *data warehouse* surgió en la década de 1980 y fue impulsado por las necesidades de las organizaciones para analizar grandes volúmenes de datos dispersos en diferentes sistemas operativos y bases de datos. La idea era crear un repositorio centralizado de datos que permitiera a las empresas consolidar y analizar información de múltiples fuentes para obtener una visión holística de sus operaciones y rendimiento.

William H. Inmon y **Ralph Kimball** son dos figuras clave en el desarrollo del concepto de *data warehouse* y en su evolución a lo largo de los años. Ambos han tenido un impacto significativo en la forma en que se diseñan, implementan y utilizan los *data warehouse*s en la industria.

William H. Inmon, a menudo considerado como el padre del *data warehousing*, ha sido una figura influyente en el campo de la gestión de datos durante décadas. En la década de 1980, Inmon fue pionero en el desarrollo del concepto de *data warehouse*

como una solución centralizada para almacenar y gestionar grandes volúmenes de datos empresariales.

Es conocido por su **enfoque de arquitectura centrado en los datos,** que enfatiza la importancia de la integración y la consistencia de los datos en un único repositorio centralizado. Inmon acuñó el término *data warehouse* y definió su arquitectura en su libro *Building the data warehouse*, publicado en 1992. En este libro, describió que un enfoque de *data warehouse* que prioriza la calidad y la integridad de los datos para garantizar la confiabilidad de la información analítica.

Ralph Kimball es otro influyente experto en el campo del *data warehousing*, quién es conocido por su enfoque de arquitectura dimensional. A diferencia de Inmon, Kimball aboga por un enfoque pragmático y orientado a usuarios para el diseño e implementación de *data warehouse*s. Kimball es conocido por popularizar el **modelo dimensional,** que utiliza un diseño basado en estrellas o copos de nieve para representar los datos.

En contraste con el enfoque de Inmon, que se centra en la integración de datos y la normalización, el **enfoque de Kimball** prioriza la simplicidad y la facilidad de uso, lo que permite a los usuarios acceder y analizar datos de manera rápida y eficiente. Kimball ha escrito varios libros sobre *data warehousing*, incluyendo *The data warehouse Toolkit*, que proporciona orientación práctica sobre cómo diseñar y construir *data warehouse*s utilizando su enfoque dimensional.

Las **contribuciones** de Inmon y Kimball al campo del *data warehousing* han sido significativas y complementarias. Mientras que Inmon estableció los fundamentos del *data warehouse* como una solución centralizada para la gestión de datos empresariales, Kimball introdujo enfoques prácticos y orientados a usuarios para su diseño e implementación. Ambos enfoques tienen sus ventajas y desventajas, y la elección entre ellos depende de los requisitos específicos y las prioridades de una organización. En la práctica, muchos proyectos de *data warehousing* combinan

elementos de ambos enfoques para aprovechar sus fortalezas y mitigar sus debilidades.

Entre sus principales **características,** podemos destacar:

- **Orientado a temas.** Los datos se organizan en torno a temas de negocio específicos, lo que facilita la comprensión y el análisis de estos.
- **Integración de datos.** Los datos son extraídos, transformados y cargados desde diferentes fuentes de datos operativos para crear un repositorio centralizado y coherente de información.
- **Almacena datos estructurados y preprocesados** optimizados para consultas analíticas.
- **Utiliza un esquema predefinido y una estructura organizada** para garantizar la consistencia y la integridad de los datos.
- **Modelo dimensional.** Utiliza un modelo dimensional o en estrella para representar los datos, con hechos y dimensiones que permiten realizar consultas analíticas complejas.
- **Historización de datos.** Los *data warehouse*s suelen mantener un historial de cambios en los datos a lo largo del tiempo, lo que permite realizar análisis comparativos y de tendencias.
- **Generación de informes.** Proporciona herramientas y funcionalidades para generar informes detallados y análisis de datos para apoyar la toma de decisiones empresariales.
- **Proporciona un rendimiento optimizado** para consultas complejas y análisis *ad hoc*.

En resumen, un ***data warehouse*** es un componente clave en la infraestructura de datos de una organización, proporcionando una plataforma centralizada y

estructurada para el análisis y la generación de informes sobre el rendimiento y las operaciones de negocio. Surgió en la década de 1980 como respuesta a la necesidad de las organizaciones de gestionar y analizar grandes volúmenes de datos dispersos en diferentes sistemas y fuentes.

Los *data warehouses* son fundamentales para una variedad de aplicaciones empresariales y análisis de datos. Aquí tienes una descripción de los principales **usos** de un *data warehouse*:

- **Análisis de negocios.** Permite la generación de informes. Un *data warehouse* proporciona una plataforma centralizada para generar informes y paneles de control que muestran métricas clave de negocio, como ventas, ingresos, inventario y rendimiento financiero.

- **Análisis de tendencias y patrones históricos.** Permite analizar datos históricos y tendencias a lo largo del tiempo para identificar patrones, oportunidades y áreas de mejora en las operaciones comerciales.

- **Análisis de clientes.** Facilita el análisis del comportamiento del cliente, la segmentación de clientes y la evaluación del valor del cliente para mejorar la retención y la satisfacción del cliente.

- **Planificación empresarial.** Proporciona datos y análisis que respaldan la planificación estratégica y la toma de decisiones a largo plazo, lo que incluye la expansión de mercado, las fusiones y adquisiciones, y el desarrollo de productos.

- **Gestión del rendimiento.** Ayuda a monitorear y evaluar el rendimiento empresarial en relación con los objetivos y metas estratégicas, e identifica las áreas de mejora y las oportunidades para optimizar la eficiencia operativa.

- **Segmentación de mercado.** Permite segmentar el mercado en función de diferentes criterios demográficos, geográficos y de comportamiento, para orientar las estrategias de *marketing* y ventas de manera más efectiva.

- **Análisis de canales.** Facilita el análisis del rendimiento de diferentes canales de *marketing* y ventas, como publicidad en línea, redes sociales, correo electrónico y ventas directas, para optimizar el retorno de la inversión (ROI).

- **Planificación de la demanda.** Ayuda a predecir la demanda de productos y servicios utilizando datos históricos y análisis predictivo, lo que permite una planificación más precisa de la producción y el inventario.

- **Optimización de inventarios.** Permite monitorear y optimizar los niveles de inventario en toda la cadena de suministro para reducir los costos de almacenamiento y mejorar la disponibilidad de productos.

- **Auditoría y cumplimiento.** Proporciona un registro centralizado y detallado de datos para fines de auditoría y cumplimiento normativo, ayudando a garantizar la precisión, la integridad y la seguridad de los datos.

- **Informes regulatorios.** Facilita la generación de informes y documentación necesaria para cumplir con regulaciones específicas de la industria, como Sarbanes-Oxley (SOX), GDPR y HIPAA.

Seguidamente, examinemos las ventajas y las desventajas de un *data warehouse*. Por el lado de las **ventajas** tenemos:

- **Centralización de datos.** Un *data warehouse* centraliza datos de múltiples fuentes en un único repositorio, lo que facilita el acceso y la gestión de la información empresarial.

- **Datos estructurados para análisis.** Los datos en un *data warehouse* están estructurados y optimizados para análisis, lo que permite a los usuarios realizar consultas complejas y generar informes detallados sobre el rendimiento y las tendencias del negocio.

- **Mejora la toma de decisiones.** Proporciona a los usuarios acceso a información actualizada y coherente, lo que facilita la toma de decisiones basada en datos y ayuda a identificar oportunidades de mejora y optimización.

- **Soporte para análisis avanzado.** Un *data warehouse* proporciona una plataforma robusta para análisis avanzado de datos, que incluye el análisis predictivo, la minería de datos y el modelado estadístico.

- **Histórico de datos.** Permite mantener un historial de cambios en los datos a lo largo del tiempo, lo que facilita el análisis comparativo y la identificación de tendencias a lo largo del tiempo.

Como **desventajas** tenemos:

- **Costo inicial y complejidad.** La implementación y mantenimiento de un *data warehouse* puede ser costosa y compleja, ya que requiere infraestructura de *hardware* y *software* especializada, así como personal técnico capacitado.

- **Tiempo de implementación.** El proceso de diseño, desarrollo e implementación de un *data warehouse* puede llevar mucho tiempo, lo que puede retrasar la disponibilidad de información para la toma de decisiones.

- **Dependencia de la calidad de los datos.** La calidad de los datos en un *data warehouse* es fundamental para la precisión y la confiabilidad de los análisis. Si los datos de entrada no son precisos o completos, los resultados del análisis pueden ser incorrectos o sesgados.

- **Rigidez en los esquemas de datos.** Los esquemas de datos en un *data warehouse* suelen ser rígidos y difíciles de cambiar una vez que están establecidos, lo que puede limitar la capacidad de adaptarse a cambios en los requisitos empresariales o en los datos de entrada.

- **Volumen de datos limitado.** Aunque un *data warehouse* puede manejar grandes volúmenes de datos, puede haber limitaciones en términos de escalabilidad y rendimiento a medida que los volúmenes de datos continúan creciendo.

En resumen, un ***data warehouse*** ofrece numerosas ventajas en términos de centralización de datos, soporte para análisis avanzado y mejora de la toma de decisiones, pero también presenta desafíos en términos de costos, complejidad, calidad de los datos y rigidez en los esquemas de datos. Es importante evaluar cuidadosamente estas ventajas y desventajas al considerar la implementación de un *data warehouse* en una organización.

Data lakehouse

El concepto de ***data lakehouse*** surge como una combinación de las ventajas del *data lake* y del *data warehouse* que intenta abordar las limitaciones individuales de cada uno. Veamos más en detalle, un *data lakehouse* es un enfoque arquitectónico que combina las características de un *data lake* y un *data warehouse* en una única plataforma unificada. Esta combinación busca aprovechar la capacidad de almacenamiento ilimitada y la diversidad de datos del *data lake*, junto con la estructura y el rendimiento optimizado para el análisis del *data warehouse*.

Sus principales **características** son:

- **Almacenamiento flexible.** Al igual que un *data lake*, un *data lakehouse* permite el almacenamiento de datos en su forma original, sin procesar ni estructurar previamente, lo que proporciona una gran flexibilidad y diversidad de datos.

- **Estructura para el análisis.** A diferencia de un *data lake* puro, un *data lakehouse* proporciona una estructura optimizada para el análisis de datos, similar a un *data warehouse*, lo que facilita consultas analíticas complejas y generación de informes.

- **Integración de datos.** Permite la integración de datos de múltiples fuentes en un único repositorio centralizado, lo que facilita el acceso y la gestión de la información empresarial.
- **Escalabilidad y rendimiento.** Un *data lakehouse* combina la escalabilidad del *data lake* con el rendimiento optimizado del *data warehouse*, lo que permite manejar grandes volúmenes de datos y realizar análisis avanzados de manera eficiente.
- **Análisis avanzado.** Proporciona capacidades avanzadas de análisis de datos, incluyendo análisis predictivo, ML y análisis de *big data*, que pueden aprovechar la diversidad y la profundidad de los datos almacenados en el *data lakehouse*.

Sus **ventajas** son:

- **Integración de capacidades.** Un *data lakehouse* combina las ventajas del *data lake* y del *data warehouse* en una única plataforma unificada, proporcionando flexibilidad, escalabilidad y capacidades avanzadas de análisis de datos.
- **Flexibilidad en el almacenamiento de datos.** Permite almacenar una amplia variedad de tipos de datos, desde datos estructurados hasta no estructurados, y utilizarlos para una variedad de aplicaciones y análisis.
- **Escalabilidad.** Puede manejar grandes volúmenes de datos y escalar horizontalmente según sea necesario para satisfacer las demandas cambiantes de almacenamiento y análisis.
- **Rendimiento optimizado para análisis de datos.** Proporciona un rendimiento optimizado para análisis de datos, lo que permite realizar consultas analíticas complejas de manera eficiente.
- **Integración de datos de múltiples fuentes.** Permite la integración de datos de múltiples fuentes en un único repositorio centralizado, lo que facilita el acceso y la gestión de la información empresarial.

- **Capacidades avanzadas de análisis de datos.** Proporciona capacidades avanzadas de análisis de datos, incluyendo análisis predictivo, ML y análisis de *big data*, que pueden aprovechar la diversidad y la profundidad de los datos almacenados en el *data lakehouse*.

- **Facilidad de acceso y exploración de datos.** Facilita el acceso y la exploración de datos para usuarios de negocio y analistas, ya que permite realizar consultas *ad hoc* y descubrir *insights* de manera ágil.

- **Costos potencialmente reducidos.** Aunque la implementación y el mantenimiento de un *data lakehouse* pueden ser costosos, puede resultar en una reducción de costos a largo plazo al consolidar y optimizar la infraestructura de datos de la organización.

En resumen, un *data lakehouse* ofrece una serie de **ventajas** significativas en términos de flexibilidad, escalabilidad, rendimiento y capacidades avanzadas de análisis de datos, lo que lo convierte en una opción atractiva para las organizaciones que buscan aprovechar al máximo su información empresarial.

Aunque un *data lakehouse* ofrece una serie de ventajas en términos de flexibilidad, escalabilidad y capacidades avanzadas de análisis de datos, también presenta algunas desventajas y desafíos. Aquí hay algunas **desventajas** potenciales de un *data lakehouse*:

- **Complejidad de implementación y mantenimiento.** Integrar las capacidades de un *data lake* y un *data warehouse* en una única plataforma unificada puede ser complejo y requerir una planificación cuidadosa, así como habilidades técnicas especializadas para la implementación y el mantenimiento.

- **Costes.** La implementación y el mantenimiento de un *data lakehouse* pueden ser costosos, ya que puede requerir infraestructura de *hardware* y *software* especializada, así como recursos humanos calificados para gestionar y mantener la plataforma.

- **Gestión de la calidad de los datos.** Mantener la calidad y la integridad de los datos en un entorno tan diverso puede ser un desafío, especialmente cuando se combinan datos estructurados y no estructurados de múltiples fuentes.

- **Seguridad y privacidad de los datos.** Almacenar una gran cantidad de datos en un único repositorio puede plantear preocupaciones sobre la seguridad y la privacidad de los datos, lo que requiere medidas de seguridad robustas para proteger los datos contra accesos no autorizados y filtraciones de datos.

- **Gestión del ciclo de vida de los datos.** Gestionar el ciclo de vida completo de los datos, incluyendo la ingestión, transformación, almacenamiento y eliminación de datos, puede ser complejo y requerir una planificación cuidadosa para garantizar la eficiencia y la conformidad con los requisitos regulatorios.

- **Riesgo de convertirse en un *data swamp*.** Si no se gestionan adecuadamente, un *data lakehouse* corre el riesgo de convertirse en un *data swamp*, donde los datos se acumulan sin un propósito claro o sin ser utilizados de manera efectiva. Es importante establecer políticas y procedimientos para la gestión y la gobernanza de datos con el fin de evitar este escenario.

En resumen, aunque un ***data lakehouse*** ofrece muchas ventajas en términos de flexibilidad, escalabilidad y capacidades avanzadas de análisis de datos, también presenta desafíos en términos de complejidad de implementación y mantenimiento, costos, gestión de la calidad de los datos, seguridad y privacidad de los datos, gestión del ciclo de vida de los datos y riesgo de convertirse en un *data swamp*. Es importante abordar estos desafíos de manera proactiva para maximizar el valor de un *data lakehouse* y garantizar su éxito en la organización.

Sus principales **desafíos** son los siguientes:

- **Gestión de la calidad de los datos.** Mantener la calidad y la integridad de los datos en un entorno tan diverso puede ser un desafío, especialmente cuando se combinan datos estructurados y no estructurados.
- **Seguridad y privacidad.** Almacenar una gran cantidad de datos en un único repositorio puede plantear preocupaciones sobre la seguridad y la privacidad de los datos, lo que requiere medidas de seguridad robustas.
- **Gestión del ciclo de vida de los datos.** Gestionar el ciclo de vida de los datos, incluyendo la ingestión, transformación, almacenamiento y eliminación de datos, puede ser complejo y requerir una planificación cuidadosa.

En resumen, un ***data lakehouse*** combina las ventajas del *data lake* y del *data warehouse* en una única plataforma unificada, proporcionando flexibilidad, escalabilidad, rendimiento y capacidades avanzadas de análisis de datos. Aunque presenta desafíos en términos de gestión de la calidad de los datos, seguridad y privacidad, y gestión del ciclo de vida de los datos, ofrece un enfoque poderoso para el almacenamiento y el análisis de datos en entornos empresariales modernos.

Enfoques de gestión de arquitectura de datos

Data mesh

Si tenemos que definir el concepto de diseño ***data mesh*** en forma simple, deberíamos decir que es la construcción de una infraestructura de autoservicio que permite a las diferentes unidades de negocio dentro de una empresa de utilizar recursos y herramientas bajo demanda, para acceder a los datos correctos, procesarlos, prepararlos y analizarlos.

Data mesh nace con el claro **propósito** de la descentralización del dato. Dota de la capacidad a toda organización de deconstruir los silos de información aislada tan existentes en las empresas y eliminar la complejidad que conlleva la centralización.

De esta manera, permite impulsar nuevos planteamientos estratégicos, lograr una gestión eficiente del negocio y caminar hacia una cultura *data-driven*. **Zhamak Dehghani** es la principal fundadora e impulsora de este paradigma.

¿Como definiríamos *data mesh*? Como una **técnica organizativa** de la descentralización del dominio de datos, la transformación y la entrega de los datos. Surge como una solución a arquitecturas centralizadas en las que su crecimiento se ve limitado por sus dependencias y su complejidad.

A medida que crece el número de fuentes y consumidores de datos, también lo hace el número de **canalizaciones de datos** necesarios para conectarlos. Esto hizo imprescindible contar con equipos especialistas en la carga de datos y habilidades para desarrollar tecnologías y administrar toda esa información. Esto fue generando paulatinamente un distanciamiento de quienes necesitan utilizar esos datos y aquellos que los gestionan.

Por eso, durante los últimos años, la **arquitectura** de *data mesh* surgió para resolver el desafío de las barreras de datos y así lograr una democratización de la información. A continuación, profundizaremos sobre sus características esenciales.

La **evolución** de las arquitecturas centralizadas tradicionales hacia una arquitectura descentralizada *data mesh* se pude ver en la Figura 7.

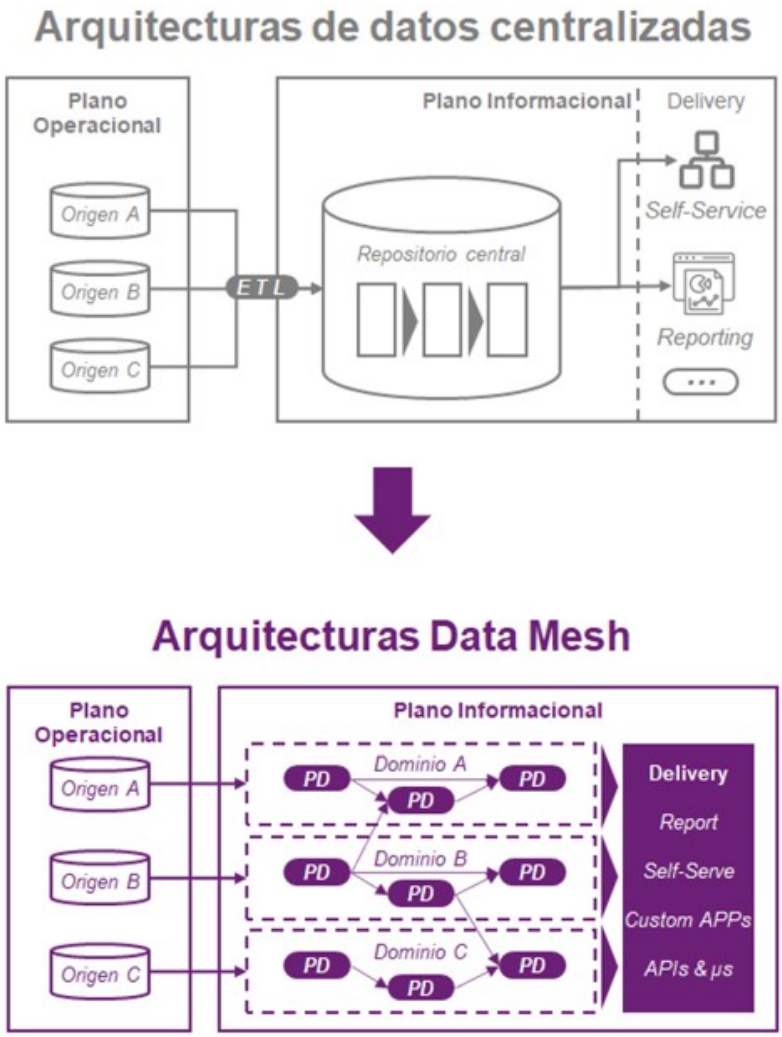

Figura 7. Evolución de las arquitecturas. Fuente: Rigoberto M., 2022.

Los **principios** del *data mesh* son los fundamentos que guían la implementación de este enfoque en la gestión de datos dentro de una organización. Teniendo en cuenta que la propiedad fundamental de una arquitectura *data mesh* es la descentralización de la información, sus principales principios de diseño serían los siguientes:

- **Descentralización.** Promueve la descentralización de la gestión de datos, donde cada dominio de negocio es responsable de sus propios datos y procesos de datos.

- **Autonomía de los equipos.** Fomenta la formación de equipos multidisciplinarios y autónomos en cada dominio de negocio, que son responsables de definir, gestionar y operar sus propios datos.
- **Orientación a servicios.** Se centra en la exposición de datos a través de servicios estandarizados y bien definidos, que facilitan la integración y el intercambio de datos entre diferentes dominios de negocio.
- **Gobernanza distribuida.** Establece políticas y prácticas de gobernanza de datos en cada dominio de negocio, que son aplicadas y gestionadas localmente por los equipos responsables de esos datos.
- **Arquitectura distribuida.** Propone una arquitectura distribuida y descentralizada para manejar la complejidad de los sistemas de datos, donde los datos se distribuyen y almacenan en cada dominio de negocio en lugar de centralizarse en un único repositorio de datos.
- **Evita el modelo de ingestión-centralización-transformación-distribución**. Se aleja del enfoque tradicional de ingestión centralizada, transformación y distribución de datos, y propone un modelo donde los datos se gestionan y transforman en el origen, antes de ser expuestos a través de servicios.
- **Cultura de datos como producto.** Fomenta una cultura centrada en los datos, donde los datos se tratan como productos que se desarrollan, gestionan y mejoran continuamente para satisfacer las necesidades cambiantes del negocio.
- **Transparencia y visibilidad.** Promueve la transparencia y la visibilidad de los datos en toda la organización, lo que permite que los usuarios entiendan y accedan fácilmente a los datos disponibles en la empresa.

- **Colaboración y comunidad.** Fomenta la colaboración y la construcción de una comunidad en torno a la gestión de datos, donde los equipos comparten conocimientos, herramientas y mejores prácticas para mejorar la eficacia y la calidad de los datos en toda la organización.

Estos son algunos de los principios comunes asociados con el *data mesh*, que proporcionan una **guía para la implementación** de este enfoque en la gestión de datos dentro de una organización.

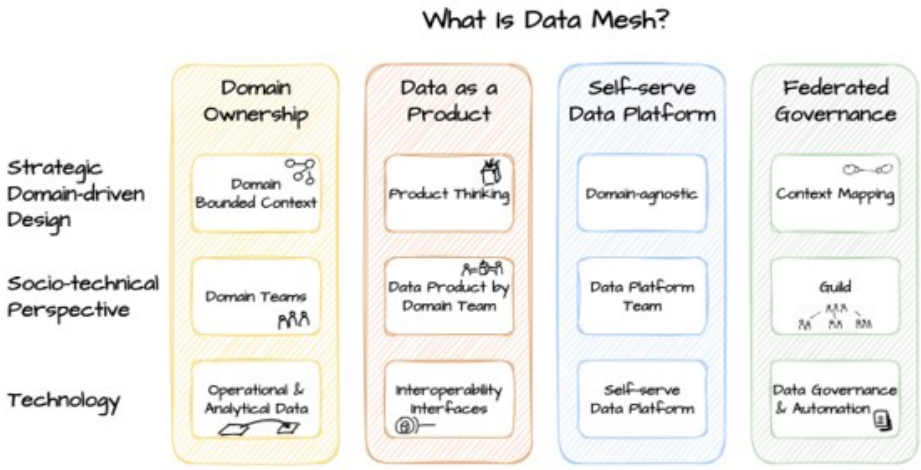

Figura 8. *Data mesh*. Fuente: Kumar, 2023.

Una **arquitectura** de *data mesh* se compone de varios elementos fundamentales que trabajan en conjunto para facilitar la gestión descentralizada y distribuida de los datos en una organización. Aquí están los elementos principales:

- **Dominios de datos.** Los dominios de datos representan áreas de negocio o funciones dentro de la organización, cada uno con su propio conjunto de datos y procesos de datos. Estos dominios son propietarios de sus datos y responsables de su gestión y operación.

- **Equipos autónomos.** Cada dominio de datos está asociado con un equipo multidisciplinario y autónomo que es responsable de definir, gestionar y operar los

datos en ese dominio. Estos equipos tienen autoridad para tomar decisiones sobre la gestión de datos en su dominio, incluyendo la definición de modelos de datos, la implementación de políticas de gobernanza y la exposición de datos a través de servicios.

- **Una plataforma de autoservicio** puede tener múltiples planos, cada uno de los cuales sirve a un perfil diferente de usuarios. En el siguiente ejemplo, se enumeran tres planos de plataforma de datos diferentes:

  - **Plano de aprovisionamiento de infraestructura de datos.** Admite el aprovisionamiento de la infraestructura subyacente, necesaria para ejecutar los componentes de un producto de datos y la malla de productos. Esto incluye el aprovisionamiento de un almacenamiento de archivos distribuido, cuentas de almacenamiento, sistema de administración de control de acceso, la orquestación para ejecutar el código interno de los productos de datos, el aprovisionamiento de un motor de consulta distribuido en un gráfico de productos de datos, etc. Solo los desarrolladores de productos de datos avanzados usan esta interfaz directamente. Este es un plano de gestión del ciclo de vida de la infraestructura de datos de nivel bastante bajo.

  - **Plano de experiencia del desarrollador de productos de datos.** Es la interfaz principal que utiliza un desarrollador de productos de datos típico. Esta interfaz abstrae muchas de las complejidades de lo que implica respaldar el flujo de trabajo de un desarrollador de productos de datos. Proporciona un mayor nivel de abstracción que el plano de aprovisionamiento. Utiliza interfaces declarativas simples para administrar el ciclo de vida de un producto de datos. Implementa automáticamente las preocupaciones transversales que se definen como un conjunto de estándares y convenciones globales, que se aplican a todos los productos de datos y sus interfaces.

  - **Plano de supervisión de malla de datos.** Hay un conjunto de capacidades que se proporcionan mejor a nivel de malla, un gráfico de productos de datos conectados, globalmente. Si bien la implementación de cada una de estas interfaces puede

depender de capacidades de productos de datos individuales, es más conveniente proporcionar estas capacidades al nivel de la malla. Por ejemplo, la capacidad de descubrir productos de datos para un caso de uso particular se proporciona mejor mediante la búsqueda o la exploración de la red de productos de datos; o la correlación de múltiples productos de datos para crear una perspectiva de mayor orden, se proporciona mejor a través de la ejecución de una consulta semántica de datos que puede operar en múltiples productos de datos en la malla.

En la Figura 9 podemos ver una ilustración de estos tres planos.

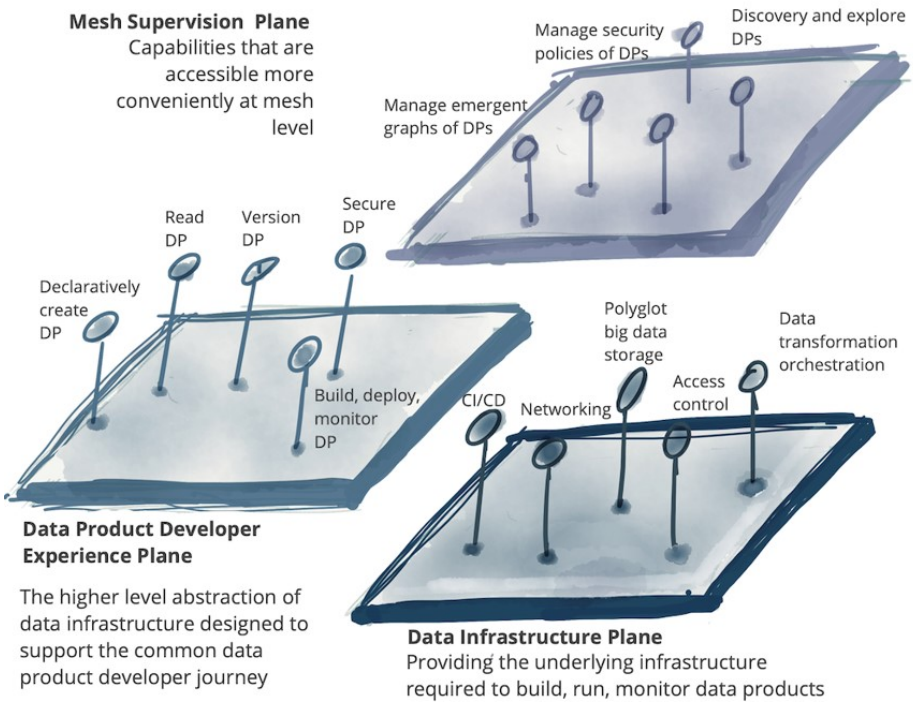

Figura 9. Planos de una plataforma de autoservicio. Fuente: Dehghani, 2020.

- **Catálogo de datos.** Proporciona un catálogo centralizado de datos que permite a los usuarios descubrir, entender y acceder a los datos disponibles en la organización. Este catálogo incluye metadatos sobre los datos, como su origen, estructura, calidad y uso.

Algunas de las **ventajas** que detectamos son las siguientes:

- **Descentralización y autonomía.** La arquitectura de *data mesh* permite a cada dominio de negocio ser responsable de sus propios datos y procesos de datos, lo que fomenta la autonomía y la toma de decisiones localizada.
- **Flexibilidad y agilidad.** Al distribuir la gestión de datos en equipos autónomos, la arquitectura de *data mesh* permite una mayor flexibilidad y agilidad para adaptarse a las necesidades y a los cambios en el negocio.
- **Mejora de la calidad de los datos.** Al fomentar la propiedad y la responsabilidad de los datos en cada dominio de negocio, la arquitectura de *data mesh* puede mejorar la calidad y la integridad de los datos en toda la organización.
- **Escalabilidad.** La arquitectura de *data mesh* puede escalar horizontalmente según sea necesario para manejar grandes volúmenes de datos y demandas de procesamiento, ya que cada dominio de negocio puede escalar de forma independiente.
- **Fomenta la innovación y la experimentación.** Al permitir que los equipos autónomos experimenten e iteren con sus propios datos y procesos de datos, la arquitectura de *data mesh* fomenta la innovación y la mejora continua en la gestión de datos.

Sin embargo, también supone **retos** a resolver:

- **Complejidad organizativa y cultural.** La implementación de una arquitectura de *data mesh* puede requerir cambios significativos en la estructura organizativa y en la cultura empresarial, lo que puede ser difícil de lograr en organizaciones grandes y establecidas.

- **Gestión de la complejidad técnica.** La distribución y la descentralización de los datos y los procesos de datos pueden introducir una mayor complejidad técnica en el manejo de datos, lo que puede requerir nuevas herramientas y habilidades para gestionar eficazmente esta complejidad.

- **Coordinación y alineación.** Aunque la descentralización puede fomentar la agilidad y la autonomía en los equipos de datos, también puede plantear desafíos en términos de coordinación y alineación entre diferentes dominios de negocio, especialmente en organizaciones grandes y complejas.

- **Riesgo de duplicación de esfuerzos.** La descentralización de la gestión de datos puede llevar a la duplicación de esfuerzos y recursos en diferentes dominios de negocio, lo que puede resultar en redundancias y falta de eficiencia en la gestión de datos.

- **Gestión de la seguridad y privacidad.** La distribución de los datos en diferentes dominios de negocio puede plantear desafíos en términos de seguridad y privacidad de los datos, lo que requiere medidas de seguridad robustas para proteger los datos contra accesos no autorizados y filtraciones de datos.

En resumen, una **arquitectura de *data mesh*** ofrece una serie de ventajas en términos de descentralización, autonomía, flexibilidad y escalabilidad, pero también presenta desafíos en términos de complejidad organizativa y técnica, coordinación y alineación, riesgo de duplicación de esfuerzos y gestión de la seguridad y privacidad. Es importante evaluar cuidadosamente estas ventajas y desventajas al considerar la implementación de una arquitectura *data mesh*.

El **paradigma *data mesh*** es reciente y difícil de implementar debido a su naturaleza descentralizada, pero se convierte en un requisito indispensable para resolver problemas de escalabilidad a los que muchas compañías se enfrentan hoy en día. Las arquitecturas centralizadas sirven mejor a compañías pequeñas o medianas que no son *data-driven* o, dicho de otra manera, que no basan sus decisiones estratégicas en el análisis de datos.

Entonces, ¿cómo saber si una compañía necesita migrar a un *data mesh*? La respuesta a esta pregunta depende de las **resistencias** que tenga actualmente una organización. Si encuentra problemas de escalabilidad, dependencias complejas entre equipos, problemas con la calidad del dato, cuellos de botella en flujos de datos o problemas de gobernanza y seguridad, probablemente necesite plantear una posible migración al paradigma *data mesh*.

Hay muchos factores, como los mencionados anteriormente, que se deben analizar antes de dar este paso. Dicho esto, cuando se pretenda responder a esta pregunta se deberán tener en cuenta principalmente tres **factores:**

- El número de **fuentes de datos** de la compañía.
- El tamaño del **equipo** dedicado a los datos.
- El número de **dominios de negocio** que tiene la compañía.

Figura 10. Factores de necesidad en el *data mesh*. Fuente: Grande y Labella, 2022.

Pero ¿cuándo tiene sentido aplicar este enfoque de arquitectura de datos? Existen muchos ejemplos y escenarios tanto desde el punto de vista para eficientizar la operativa de la información como desde el punto de vista de explotación y analítica. Algunos ejemplos de **casos de uso** serían:

- **Análisis de datos distribuidos.** Una empresa con múltiples unidades de negocio puede utilizar una arquitectura de *data mesh* para integrar datos dispersos en diferentes sistemas y ubicaciones, lo que les permite a los analistas acceder y analizar datos de manera centralizada para obtener una visión unificada del rendimiento empresarial.

- **Integración de datos en la nube.** Una empresa que migra sus sistemas a la nube puede utilizar una arquitectura de *data mesh* para integrar datos dispersos en diferentes plataformas en la nube, lo que les permite a los usuarios acceder y analizar datos de manera transparente desde cualquier ubicación.

- **Gestión de datos de IoT.** Una empresa que recopila datos de sensores y dispositivos de IoT puede utilizar una arquitectura de *data mesh* para integrar y procesar datos de manera distribuida, lo que permite el análisis en tiempo real y la detección de patrones y anomalías en los datos.

- **Análisis de datos en tiempo real.** Una empresa que necesita analizar grandes volúmenes de datos en tiempo real puede utilizar una arquitectura de *data mesh* para distribuir el procesamiento de datos en diferentes nodos, lo que permite análisis en tiempo real y la toma de decisiones basada en datos en tiempo real.

- **Gestión de datos de clientes.** Una empresa que recopila datos de clientes de diferentes fuentes, como transacciones, interacciones en redes sociales y comentarios en línea, puede utilizar una arquitectura de *data mesh* para integrar y analizar estos datos de manera centralizada, lo que permite una comprensión más profunda del comportamiento del cliente y la personalización de la experiencia del cliente.

- **Análisis de riesgos financieros.** Una institución financiera puede utilizar una arquitectura de *data mesh* para integrar y analizar datos dispersos en diferentes sistemas y ubicaciones, lo que permite la identificación y la mitigación de riesgos financieros de manera más eficiente y efectiva.

Estos son solo algunos ejemplos de casos de uso de una arquitectura de *data mesh*, que demuestran su versatilidad y aplicabilidad en una amplia variedad de entornos empresariales y verticales de la industria.

Data fabric

Data fabric es un enfoque arquitectónico que aborda los desafíos de la gestión de datos en entornos empresariales cada vez más complejos y distribuidos. Se trata de una estrategia integral que busca proporcionar una capa unificada y coherente sobre los datos dispersos en diferentes sistemas y ubicaciones, lo que facilita su acceso, gestión y análisis de manera eficiente y efectiva.

En un mundo donde los datos están distribuidos en una amplia variedad de sistemas y ubicaciones, como bases de datos locales y en la nube, sistemas de archivos, dispositivos móviles y dispositivos de IoT, entre otros, gestionar y operar estos datos de manera eficiente puede ser un desafío significativo para las organizaciones. *Data fabric* aborda este desafío proporcionando una **capa unificada** de virtualización, integración, gestión de metadatos, seguridad y cumplimiento sobre los datos dispersos, lo que les permite a los usuarios acceder y analizar los datos de manera fluida y transparente, independientemente de su ubicación o formato.

La clave de *data fabric* es su capacidad para crear una **vista unificada** de los datos dispersos en diferentes sistemas y ubicaciones, lo que permite a los usuarios acceder a ellos de manera transparente sin necesidad de conocer su ubicación física o la complejidad subyacente de la infraestructura de datos. Esto simplifica la gestión y operación de los datos en la organización, mejora la flexibilidad y la agilidad para adaptarse a las necesidades y cambios en el negocio, y acelera el desarrollo y la entrega de soluciones basadas en datos.

En resumen, *data fabric* es un **enfoque arquitectónico integral** diseñado para abordar los desafíos de la gestión de datos en entornos empresariales complejos y distribuidos, proporcionando una capa unificada y coherente sobre los datos dispersos en diferentes sistemas y ubicaciones, lo que facilita su acceso, gestión y análisis de manera eficiente y efectiva.

Sus **características** principales son las siguientes:

- **Virtualización de datos.** *Data fabric* utiliza técnicas de virtualización para proporcionar una vista unificada de los datos, lo que les permite a los usuarios acceder a ellos sin necesidad de conocer su ubicación física o la complejidad subyacente de la infraestructura de datos.
- **Integración de datos.** Permite la integración de datos de múltiples fuentes, incluyendo sistemas locales y en la nube, bases de datos relacionales y no relacionales, sistemas de archivos y servicios web, entre otros.
- **Gestión de metadatos.** *Data fabric* incluye una capa de metadatos que proporciona información sobre los datos disponibles en la organización, incluyendo su origen, estructura, calidad y uso. Esto facilita la búsqueda, descubrimiento y comprensión de los datos por parte de los usuarios.

- **Seguridad y cumplimiento.** Proporciona mecanismos de seguridad robustos para proteger los datos contra accesos no autorizados y cumplir con los requisitos regulatorios y de cumplimiento, independientemente de su ubicación o formato.
- **Escalabilidad y rendimiento.** *Data fabric* está diseñado para escalar horizontalmente según sea necesario para manejar grandes volúmenes de datos y demandas de procesamiento, proporcionando un rendimiento óptimo para el acceso y análisis de datos.

Los **componentes** de una arquitectura *data fabric* son:

- **Capa de virtualización de datos.** Proporciona una vista unificada de los datos dispersos en diferentes sistemas y ubicaciones, lo que les permite a los usuarios acceder a ellos de manera transparente.
- **Capa de integración de datos.** Facilita la integración de datos de múltiples fuentes, transformando y combinando los datos según sea necesario para satisfacer las necesidades de los usuarios.
- **Capa de gestión de metadatos.** Almacena y gestiona los metadatos sobre los datos disponibles en la organización, lo que proporciona información sobre su origen, estructura, calidad y uso.
- **Capa de seguridad y cumplimiento.** Proporciona mecanismos de seguridad y cumplimiento para proteger los datos contra accesos no autorizados y garantizar el cumplimiento de los requisitos regulatorios y de cumplimiento.

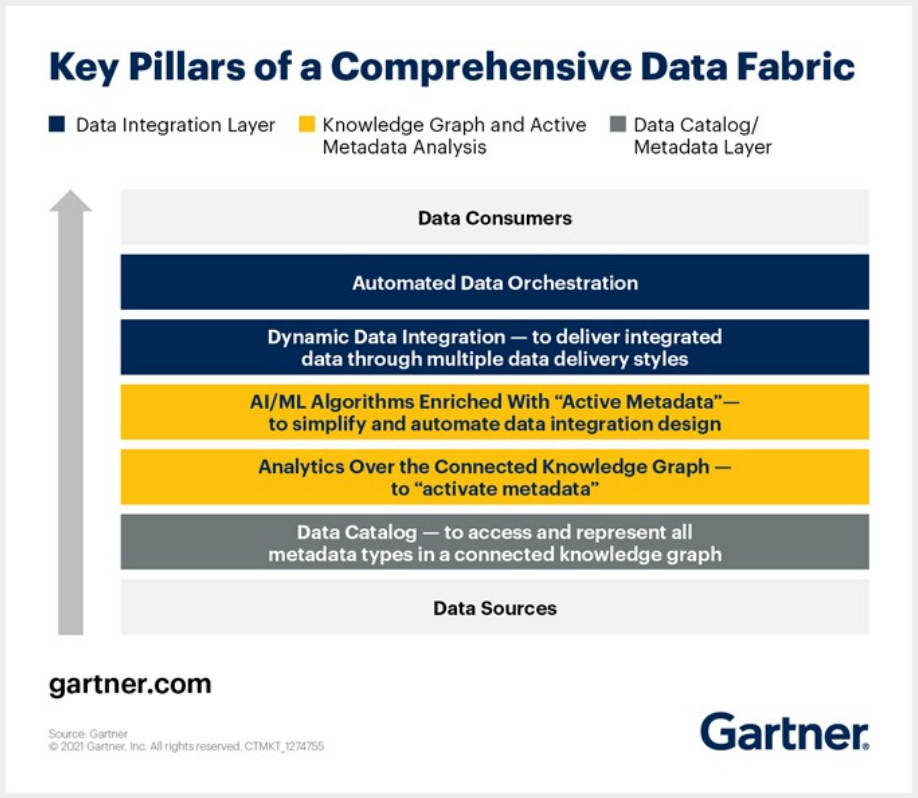

Figura 11. Pilares del *data fabric*. Fuente: Gartner.

Sus principales **ventajas** son:

- **Simplificación de la gestión de datos.** Proporciona una capa unificada y coherente sobre los datos dispersos en diferentes sistemas y ubicaciones, lo que simplifica la gestión y operación de los datos en la organización.
- **Flexibilidad y agilidad.** Permite a los usuarios acceder y analizar datos de manera fluida y eficiente, independientemente de su ubicación o formato, lo que facilita la toma de decisiones basada en datos y la innovación en la organización.
- **Mejora del tiempo de comercialización.** Al facilitar la integración y el acceso a los datos, *data fabric* puede acelerar el desarrollo y la entrega de soluciones basadas en datos, lo que reduce el tiempo de comercialización y mejora la competitividad de la empresa.

Entre los principales **desafíos de implementación** de una arquitectura *data fabric*, podemos destacar:

- **Complejidad técnica.** Implementar y gestionar una arquitectura de *data fabric* puede ser complejo, ya que requiere integrar múltiples tecnologías y sistemas en una única plataforma unificada.
- **Gestión del cambio organizativo.** Adoptar un enfoque de *data fabric* puede requerir cambios significativos en la cultura y la estructura organizativa, así como en las prácticas de gestión de datos, lo que puede ser difícil de lograr en organizaciones grandes y establecidas.

En resumen, ***data fabric*** es una arquitectura que proporciona una capa unificada y coherente sobre los datos dispersos en diferentes sistemas y ubicaciones, lo que facilita el acceso, la gestión y el análisis de datos en la organización. Si bien ofrece numerosas ventajas en términos de simplificación de la gestión de datos, flexibilidad y mejora del tiempo de comercialización, también presenta desafíos en términos de complejidad técnica y gestión del cambio organizativo.

Una vez que las organizaciones resuelven estos problemas, pueden comenzar a explorar nuevos **casos de uso** de estructuras de datos como los siguientes.

- **Mejorar la aplicación de IA sobre los datos.** Una enfoque de arquitectura *data fabric* puede proporcionar a los *data science* el acceso integral a los datos que les permita aplicar técnicas avanzadas de IA para mejorar la toma de decisiones. Por ejemplo, el mantenimiento predictivo requiere un acceso optimizado a los datos en tiempo real que proporciona una estructura de datos.
- **Mejorar la seguridad.** Una estructura de datos también puede mejorar las aplicaciones de seguridad al vincular datos y aplicaciones de todos los sistemas físicos y de TI. Por ejemplo, un equipo podría mejorar la seguridad al vincular la información de los lectores de llaves que se utilizan para abrir puertas, lo que podría correlacionarse con los datos de eventos de los sistemas informáticos a los que se

accede desde dentro de la instalación. Esto permitiría realizar un análisis más sofisticado del comportamiento típico y anómalo para activar alertas de seguridad en tiempo real cuando sea necesario.

- **Crear una visión holística del cliente.** Las organizaciones también pueden usar una arquitectura *data fabric* para entrelazar datos de las actividades de un cliente junto con los diversos roles que interactúan con ellos para obtener una visión más holística. Esto podría incorporar datos en tiempo real de diversas actividades de ventas, realización de ingresos potenciales, tiempo de incorporación de clientes o métricas de satisfacción del cliente.

- **Mejorar la comprensión empresarial.** Las empresas también pueden usar la estructura de datos para crear una visión más holística del negocio en todas las actividades y departamentos.

- **Facilitar la productivización de algoritmos.** *Data fabric* se pueden usar para entrenar, configurar e implementar algoritmos de predicción simples y desencadenar acciones que se ejecutan en varios puntos finales de aplicaciones empresariales. Estos tipos de casos de uso abarcan todo, desde la trazabilidad de la seguridad hasta el cumplimiento de auditorías y los eventos que generan ingresos, como la acción de abandono del carrito, la optimización de anuncios, la retención de clientes, el *marketing* e incluso la venta orquestada.

- **Implementación de *market places* de datos.** Las empresas que implementan una arquitectura de estructura de datos también pueden configurar un mercado de datos más accesible que facilite a los desarrolladores ciudadanos entretejer fuentes de datos dispares en nuevos modelos. Un mercado de datos permite a los ingenieros de datos configurar una infraestructura que se puede usar en múltiples casos de uso en lugar de crear una nueva infraestructura para cada caso de uso individualmente.

Diferencias data mesh vs. data fabric

Tanto *data mesh* como *data fabric* son enfoques arquitectónicos destinados a abordar los desafíos de la gestión de datos en entornos empresariales cada vez más complejos y distribuidos. Aunque comparten algunos conceptos y objetivos comunes, también presentan diferencias significativas en términos de enfoque, implementación y alcance. Aquí están las principales **diferencias** entre *data mesh* y *data fabric.*

Enfoque

- ***Data mesh.*** *Data mesh* se centra en descentralizar la gestión de datos al promover la autonomía de los equipos en cada dominio de negocio para definir, gestionar y operar sus propios datos. Proporciona una arquitectura distribuida y descentralizada para manejar la complejidad de los sistemas de datos, donde los datos se distribuyen y almacenan en cada dominio de negocio en lugar de centralizarse en un único repositorio de datos.

- ***Data fabric.*** *Data fabric* se centra en proporcionar una capa unificada y coherente sobre los datos dispersos en diferentes sistemas y ubicaciones. Utiliza técnicas de virtualización de datos, integración de datos, gestión de metadatos, seguridad y cumplimiento para crear una vista unificada de los datos, lo que facilita su acceso, gestión y análisis de manera eficiente y efectiva.

Alcance

- ***Data mesh.*** *Data mesh* se centra en la gestión descentralizada de datos en cada dominio de negocio, lo que promueve la autonomía y la responsabilidad de los equipos en la gestión de sus propios datos. Proporciona un marco para integrar y operar datos de manera distribuida, lo que permite a los equipos acceder y analizar datos de manera fluida y transparente.