

WHITE PAPER

Data Fabric, Data Mesh o ambos: ¿cuál es mejor para tu empresa?

Mike Ferguson
Intelligent Business Strategies
April 2023

Investigación patrocinada
por:





Índice

Introducción.....	3
Data fabric: ¿qué es?	4
Implementación flexible y automatizada	4
Conectividad segura	5
Catálogo de datos.....	5
Desarrollo colaborativo, altamente productivo y mejorado por inteligencia artificial	5
Rendimiento y escalabilidad	6
Extensibilidad.....	6
Gobernanza unificada de datos	6
Data marketplace.....	7
¿Qué es un data mesh?	9
Productos de datos	10
Creación de nuevos productos a partir de otros productos de datos	10
Infraestructura self-service como plataforma	11
Gobernanza federada de productos de datos.....	11
¿Cómo ayuda el data fabric a crear un data mesh?	11
Cómo el data fabric de Denodo Platform ayuda a crear un data mesh	12
Denodo Platform: flexibilidad y opciones de implementación	13
Conectividad segura de Denodo Platform	14
Catálogo de datos de Denodo Platform	14
Integración de Denodo Platform con catálogo de datos de terceros	15
Desarrollo colaborativo, respaldado por inteligencia artificial, de Denodo Platform	16
Rendimiento y escalabilidad de Denodo Platform.....	17
Extensibilidad de Denodo Platform	18
Gestión unificada de datos de Denodo Platform.....	18
Denodo Platform data marketplace	19
Conclusiones	20

INTRODUCCIÓN

Los directivos consideran que los datos, el business intelligence y la inteligencia artificial son fundamentales para impulsar una estrategia empresarial con una mirada hacia el futuro.

Contar con datos de alta calidad, seguros y conformes a nivel normativo es fundamental para el éxito empresarial.

Los datos deben estar listos para usarse y compartirse para así conseguir un mejor “time-to-value”

La complejidad de los datos es un obstáculo enorme en este proceso.

Tanto el data fabric como el data mesh se utilizan para acelerar el desarrollo de una infraestructura de datos de calidad en una empresa data-driven.

Actualmente, la mayor parte de las empresas tanto de tamaño mediano como grandes tienen grandes expectativas respecto a la contribución de los datos y la analítica para ayudar al negocio a ser *data-driven*. De hecho, se podría afirmar que los datos y el uso de la inteligencia empresarial, el aprendizaje automático (machine learning) y la inteligencia artificial (IA) son ahora imprescindibles para dirigir la estrategia empresarial de cara al futuro.

Sin embargo, que esto se haga realidad depende de la capacidad de la organización para disponer de una infraestructura de datos de alta calidad, segura y conforme a nivel normativo sobre la que generar insights de valor para el negocio. La calidad de esos cimientos, de esos datos, tiene que ser impecable para que la empresa pueda mejorar el rendimiento de su negocio y ser una empresa más centrada en el cliente. Además, los datos deben respetar las obligaciones de cumplimiento normativo para poder considerarlos fiables para entrenar modelos de machine learning y tratar de sacar información y recomendaciones más innovadoras. Asimismo, necesitan estar listos para usarse y compartirse dentro y fuera de la organización, con el fin de conseguir acortar el “time-to-value”. Lo mismo puede decirse para cuando las empresas quieren reducir costes y mejorar la productividad mediante la automatización por inteligencia artificial.

Sin embargo, existen obstáculos en este proceso hacia el éxito empresarial. Uno de los principales desafíos es la complejidad de los datos: hay cada vez más fuentes de datos y almacenes de datos ubicados on-premises, en la nube, incluso en múltiples nubes, en infraestructuras de proveedores de aplicaciones SaaS (software como servicio) y en el edge. En la actualidad, muchas empresas tienen datos distribuidos en un entorno multinube verdaderamente híbrido.

Dada esta complejidad, existen algunos requisitos críticos que son necesarios para construir y acelerar el desarrollo de una arquitectura de datos sólida en una empresa que pretende ser “data-driven”. Entre ellos, se incluye la necesidad de un software de data fabric para conectarse y gestionar fuentes de datos distribuidas, así como la necesidad de crear un “data mesh” de productos de datos mediante la democratización de la ingeniería de datos para acortar el “time-to-value”.

Este documento analiza los conceptos tanto del data fabric como del data mesh para comprender qué son y cómo —o si— pueden ir de la mano para crear y acelerar el desarrollo de una infraestructura de datos que sean de alta calidad, seguros y conformes a nivel normativo para una empresa “data-driven”.



DATA FABRIC: ¿QUÉ ES?

La Figura 1 muestra el problema al que se enfrenta la mayoría de las empresas que desean construir una arquitectura de datos sólida como base de una empresa “data-driven”. El problema al que se enfrentan es que los datos se almacenan en bases de datos de aplicaciones SaaS; sistemas de procesamiento de transacciones locales, bases de datos relacionales y NoSQL; sistemas de archivos locales y sistemas de gestión de contenido; diferentes bases de datos relacionales y NoSQL en múltiples nubes diferentes; sistemas de archivos en la nube; almacenes de contenido; y datos almacenados en el edge o retransmitidos desde allí. Además, puede que haga falta recopilar datos externos de múltiples proveedores de datos, data clean rooms y/o marketplaces de datos públicos.

Muchas empresas ahora disponen de datos distribuidos en un entorno multinube híbrido.

Se recopilan datos externos de muchas fuentes de datos diferentes.

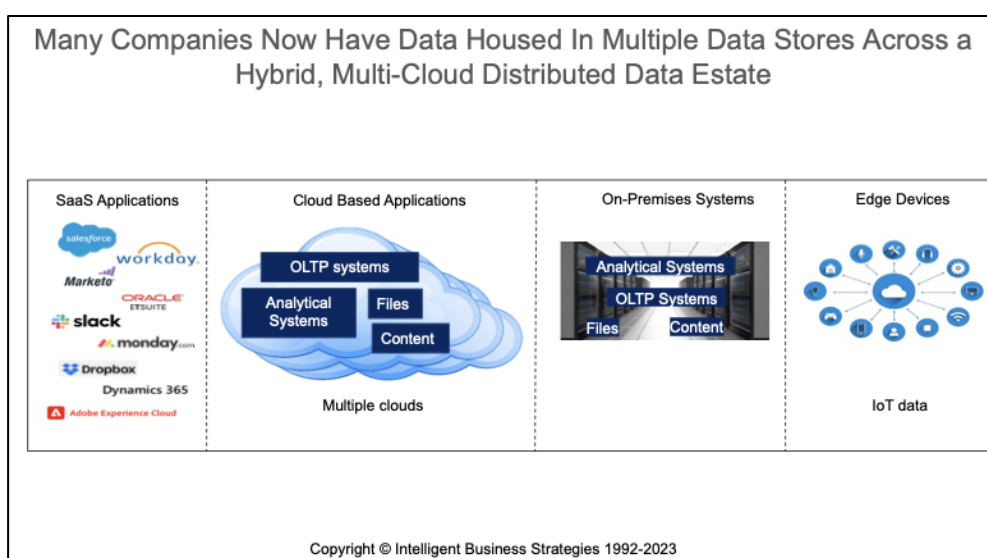


Figura 1

¿Y cómo abordar con esta realidad de datos tan compleja? La respuesta es con un software de data fabric.

Un data fabric es un software, una plataforma de gestión de datos que es capaz de conectarse a una amplia gama de fuentes de datos en un entorno híbrido y multinube de datos distribuidos. Ofrece todas las funcionalidades para descubrimiento automático de datos, su clasificación, gestión, limpieza, transformación e integración, orquestación, automatización, uso compartido y aprovisionamiento seguro y conforme a nivel normativo.

El software de data fabric tiene una serie de funcionalidades, como:

IMPLEMENTACIÓN FLEXIBLE Y AUTOMATIZADA

- Despliegue de manera automática y ejecución del software de data fabric en cualquier lugar, por ejemplo, en Kubernetes, on-premises, como un servicio administrado SaaS o en diferentes nubes públicas como Amazon Web Services (AWS), Microsoft Azure o Google Cloud Platform (GCP) o incluso en cualquier combinación de los anteriores. Esto incluye la posibilidad de configurar y automatizar clústers para Kubernetes y despliegues en la nube.

El software de data fabric debería poder ejecutarse en cualquier lugar o estar disponible como servicio totalmente administrado.

- La ejecución en la nube o como un servicio administrado SaaS permitiría a los ingenieros de datos que trabajan de manera individual o en equipo a comenzar rápidamente con una suscripción o un servicio basado en el consumo.

CONECTIVIDAD SEGURA

- Conecta a una amplia gama de fuentes de datos y destinos en un entorno de datos distribuidos en un marco de conectividad segura, por ejemplo a:
 - sistemas de archivos locales, bases de datos relacionales, sistemas Hadoop, bases de datos NoSQL, aplicaciones personalizadas y empaquetadas, colas de mensajes y software de bus de servicios empresariales;
 - almacén de objetos en la nube (p. ej., archivos de Parquet en un data lake en la nube), sistemas de archivos en la nube, bases de datos relacionales en la nube y almacenes de bases de datos NoSQL en varios entornos en la nube diferentes;
 - servicios de transmisión de datos como AWS Kinesis, Azure Event Hubs, Confluent Kafka, Google Pub/Sub, BM Event Streams, Solace;
 - aplicaciones basadas en SaaS que se ejecutan en diferentes centros de datos de proveedores de aplicaciones SaaS con conectividad a datos mediante APIs o acceso seguro con credenciales al almacén de datos subyacente de la aplicación SaaS;
 - dispositivos, puertas de enlace y bases de datos en el edge.

El software de data fabric debería proporcionar una conectividad segura a una amplia gama de fuentes de datos.

CATÁLOGO DE DATOS

- Ofrecer un catálogo de datos, integración con catálogos de datos de terceros o las dos cosas para permitir:
 - detección automática de datos, esquemas y relaciones de datos dentro y entre múltiples almacenes de datos, así como flujos de datos del sistema de origen;
 - mapeo automático de nombres de datos físicos detectados en las distintas fuentes a términos de negocio comunes en un glosario empresarial dentro de un catálogo de datos para abstraer la complejidad de las fuentes de datos subyacentes y facilitar que los usuarios finales entiendan el significado de la información de origen;
 - clasificación automática de datos para detectar y etiquetar automáticamente los tipos de datos confidenciales en los sistemas de origen para poder gestionar y proteger esta información de acuerdo con las obligaciones de cumplimiento normativo;
 - proporcionar un linaje de extremo a extremo.

El software de data fabric debería incluir un catálogo de datos o integrarse con catálogos de terceros.

DESARROLLO COLABORATIVO, ALTAMENTE PRODUCTIVO Y MEJORADO POR INTELIGENCIA ARTIFICIAL

- Permite que múltiples equipos de productores de datos trabajen en la misma plataforma.

El software de data fabric debería soportar múltiples equipos de ingenieros de datos y utilizar la inteligencia artificial para acelerar la productividad de los desarrolladores.

- Utiliza las mejoras por inteligencia artificial para incrementar la productividad y acelerar el desarrollo de integración de datos. Algunos ejemplos de las funcionalidades son:
 - recomendar fuentes de datos relevantes;
 - inferir de manera automática el esquema de las fuentes de datos;
 - detectar de forma automática anomalías en las fuentes de datos;
 - recomendar oportunidades de estandarización de datos (p. ej., usar agrupación en clústeres de datos similares);
 - recomendar tareas de limpieza de datos;
 - recomendar transformaciones de datos;
 - recomendar asignaciones de integración de datos;
 - automatizar tareas (p. ej., eliminar duplicados).
- Orquesta los pipelines para controlar la secuencia de tareas de integración necesarias para generar datos compatibles de alta calidad;
- Comparte con facilidad metadatos entre ingenieros de datos en uno o más equipos, al proporcionar acceso a un catálogo de datos común para efectuar búsquedas desde el software de data fabric. Esto prevé compartir metadatos sobre términos del glosario empresarial, las fuentes de datos, sus relaciones, calidad, clasificación de información confidencial, linaje, componentes de integración, así como productos de datos;
- Ofrece DataOps para habilitar el desarrollo colaborativo y el control de versiones. Entre otras cosas, facilita el desarrollo continuo y la entrega continua (CI/CD, por sus siglas en inglés) para administrar el desarrollo, automatizar las pruebas, gestionar la producción y comparar de manera sencilla los entornos de desarrollo y de producción para detectar diferencias.

Se necesita orquestación para ensamblar los componentes de los pipelines necesarios para diseñar y gestionar los datos.

Se precisan metadatos compartidos y DataOps para facilitar y acelerar el desarrollo colaborativo.

RENDIMIENTO Y ESCALABILIDAD

El software de data fabric debería aprovechar la optimización por coste, el procesamiento paralelo y la inteligencia artificial para mejorar el rendimiento.

- Supervisa de forma automática el uso de datos en función de todos los usuarios que acceden y procesan datos usando el software de data fabric;
- Utiliza modelos de machine learning integrados para predecir y automatizar de manera dinámica qué datos deben almacenarse en caché para maximizar el rendimiento basándose en el procesamiento de datos y seguimiento de las consultas en directo;
- Utiliza modelos de machine learning integrados para automatizar las recomendaciones de consultas;
- Permite una escalabilidad dinámica (p. ej. mediante la configuración dinámica de clústeres y el procesamiento paralelo masivo [MPP]);
- Utiliza la optimización pushdown en los sistemas origen para maximizar el rendimiento y minimizar el tráfico de red durante la integración de datos;
- Utiliza un motor MPP para procesar datos en ficheros (e.g. parquet, iceberg) dentro del almacenamiento en la nube, ya que no es posible delegar el procesamiento de consultas a un sistema de archivos;
- Facilita la optimización basada en reglas, la optimización por coste y mediante MPP para acelerar la ejecución de la integración de datos;

También se necesita una escalabilidad automatizada.

EXTENSIBILIDAD

- Utiliza código personalizado para seleccionar de forma dinámica fuentes de datos alternativas, gestionar controles de calidad, llevar a cabo transformaciones complejas, etiquetar datos, etc.

GOBERNANZA UNIFICADA DE DATOS

El data fabric necesita emplear términos de negocio comunes en un glosario para crear una capa semántica común.

También debería poder controlar el acceso a los datos y garantizar que toda la información confidencial esté protegida.

- Crea una capa semántica común con los términos del glosario empresarial;
- Permite una gobernanza unificada del acceso a los datos y su confidencialidad:
 - integración con servicios empresariales de gestión de acceso de identidad;
 - detectar de forma automática tipos de datos confidenciales u obtenerlos de un catálogo de datos;
 - creación de políticas de control de acceso a datos y privacidad de datos asociadas con tipos de información confidencial y clasificaciones de confidencialidad de datos, incluidas políticas para:
 - enmascaramiento dinámico y cifrado de datos confidenciales;
 - exclusión de datos;
 - control de acceso basado en roles;
 - control de acceso basado en atributos.

DATA MARKETPLACE

El data fabric también necesita habilitar un data marketplace para publicar productos de datos reutilizables y ponerlos a disposición para su uso.

- Gobernar la publicación, el intercambio y el consumo de datos, así como la funcionalidad de:
 - contar con un proceso estándar para publicar datos empresariales listos para usar y vistas de datos que permitan:
 - definir las condiciones de uso de los datos empresariales listos para usar, que deben aceptarse antes de compartir dicha información;
 - definir políticas para controlar la seguridad de acceso, la privacidad, el intercambio y la retención de datos en el marketplace de datos;
 - permitir el control de versiones de conjuntos de datos;
 - buscar datos empresariales listos para usar;
 - solicitar al propietario el acceso a datos empresariales listos para usar;
 - definir un proceso estándar para que los propietarios de datos aprueben o rechacen el acceso a los datos;
 - permitir que los usuarios accedan a un glosario empresarial que describa el significado de los datos empresariales listos para usar y disponibles en un marketplace de datos;
 - permitir que los usuarios accedan al linaje completo de metadatos para entender cómo se generaron los datos empresariales listos para usar y disponibles en el marketplace de datos;

- auditar y llevar a cabo un seguimiento de los datos compartidos para monitorizar el uso y garantizar el cumplimiento de las obligaciones normativas y legislativas.

¿QUÉ ES UN DATA MESH?

Un data mesh es un enfoque descentralizado organizativamente para generar conjuntos de datos reutilizables y de alta calidad, conocidos como «productos de datos».

Un data mesh es un enfoque para democratizar y descentralizar la ingeniería de datos con el objetivo de conseguir acortar el “time-to-value”. Se basa en la idea de que los expertos de cada dominio en cuestión, que usan datos específicos de dicho dominio a diario, estén capacitados para crear pipelines de ingeniería de datos para generar productos de datos de alta calidad y listos para usar en la empresa a partir de fuentes con las que están familiarizados.

Los ingenieros de datos profesionales pueden integrarse en los diferentes dominios empresariales para ayudar a mejorar las competencias de los «citizen data engineers» y fomentar el empleo de buenas prácticas para lograrlo.

El objetivo es acelerar la ingeniería de datos y, al mismo tiempo, eliminar la dependencia de un equipo centralizado siempre limitado de informáticos y data engineers, que a menudo apenas pueden satisfacer la demanda y lidiar con un volumen cada vez mayor de nuevas solicitudes de datos por parte de negocio.

Los productos de datos generados por «citizen data engineers» expertos en la materia se pueden luego compartir y reutilizar en toda la empresa. La intención es contar con más ingenieros de datos e industrializar el desarrollo de productos de datos de alta calidad para que se usen y reutilicen en múltiples cargas de trabajo analíticas.

El objetivo es contar con más ingenieros de datos, eliminar cualquier retraso y acelerar el desarrollo de productos de datos.

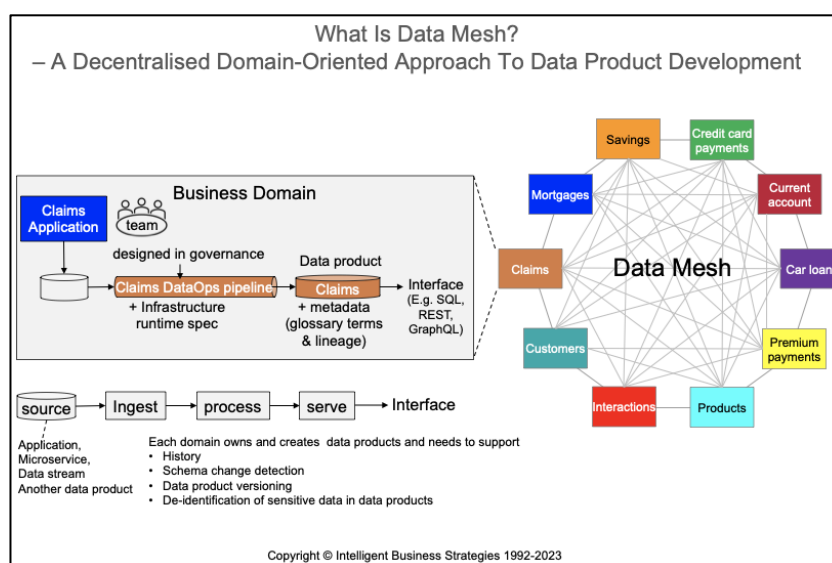


Figura 2

El concepto de data mesh¹ se basa en cuatro principios fundamentales:

- arquitectura y propiedad de datos descentralizadas y específicas del dominio de negocio;
- datos como producto;
- infraestructura de datos self-service como plataforma;
- gobernanza de datos computacionalmente federada.

Existen cuatro principios fundamentales relativos al data mesh.

¹ [How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh](#), Zhamak Dehghani, May 2019

PRODUCTOS DE DATOS

Los productos de datos son fundamentales para el planteamiento del data mesh. Existen conjuntos de datos reutilizables que se caracterizan por ser:

- explorables;
- direccionables;
- fiables;
- autodescriptivos;
- interoperables;
- seguros.

Cada producto de datos tiene un identificador empresarial único, nombrado en términos de negocio, metadatos para indicar cómo se generó y una interfaz para acceder a él.

Además de los datos en sí, un producto de datos tiene un identificador empresarial único e incluye nombrado en términos de negocio, reglas y asignaciones de metadatos para limpiar, transformar e integrar datos, la especificación necesaria para la ejecución del pipeline que genera el producto de datos y las APIs para acceder a dicho producto.

Pueden existir diferentes tipos de productos de datos en un data mesh:

- conjuntos de datos almacenados físicamente;
- vistas virtuales que integran datos bajo demanda a partir de una o más fuentes de datos subyacentes;
- secuencias de comandos y consultas SQL almacenadas que se pueden invocar bajo demanda para generar un producto de datos y hacer que sea operativo.

El producto de datos puede ser tanto físico como virtual.

Ejemplos de productos de datos incluyen entidades “core” de datos empresariales tales como “clientes”, “productos”, “pedidos”, “envíos”, “pagos”, “devoluciones”, “proveedores”, “materiales”, etc. Una buena manera de identificar productos de datos es crear un modelo conceptual de datos.

CREACIÓN DE NUEVOS PRODUCTOS A PARTIR DE OTROS PRODUCTOS DE DATOS

Uno de los principales beneficios del data mesh es que fomenta la reutilización, lo que ayuda a conseguir más resultados en menos tiempo. Una vez que se ha creado un producto de datos, es posible que los usuarios lo utilicen en otro pipeline de ingeniería de datos para integrarlo con información de otras fuentes y así generar un nuevo producto de datos, que puede añadirse al data mesh.

El “time-to-value” debería hacerse cada vez más corto a medida que se puedan reutilizar cada vez más productos de datos.

El enfoque consiste en aprovechar lo que los demás han hecho antes. En otras palabras, los nuevos proyectos no tienen por qué comenzar desde cero con datos sin procesar cada vez. Al contrario, se cuenta con la ventaja de poder seleccionar y reutilizar productos de datos empresariales relevantes y listos para usar, que han sido creados por otros en la organización. Si se gestiona bien, este planteamiento incremental para generar productos de datos debería traducirse en nuevos proyectos cada vez más rápidos a medida que cada vez más y más de los que se necesita para construir un nuevo producto de datos ya está disponible.

INFRAESTRUCTURA SELF-SERVICE COMO PLATAFORMA

Múltiples equipos de productores de datos utilizan una plataforma común para generar nuevos productos de datos.

Además de los citizen data engineers especializados en un determinado dominio que crean productos de datos, la idea detrás de un data mesh es que varios equipos que se dedican a ello en sus respectivos dominios empresariales utilicen un software de infraestructura self-service para crear pipelines de datos; estos últimos recopilan la información de las fuentes de datos que se usan en ese dominio y generan productos de datos disponibles en un data mesh.

GOBERNANZA FEDERADA DE PRODUCTOS DE DATOS

Es necesario definir políticas locales y globales para toda la empresa que regulen el acceso a los datos, su privacidad y el intercambio de productos de datos.

Se trata de definir y hacer cumplir las políticas locales y globales que afecten a toda la empresa para gobernar los datos en un data mesh. Las políticas de privacidad de datos se aplican a todos los productos de datos, mientras que las políticas de acceso a los mismos las pueden definir localmente en sus respectivos dominios los propietarios de los productos de datos. Las políticas de gobernanza de datos permiten gestionar la seguridad del acceso a los datos, su privacidad, intercambio, retención y calidad. Se pueden implementar tanto durante como después del desarrollo del producto de datos. Esto garantiza que todos los productos de datos generados cumplan con las normas y se puedan compartir de manera fiable.

¿CÓMO AYUDA EL DATA FABRIC A CREAR UN DATA MESH?

Una vez entendidos los conceptos de data fabric y data mesh, el siguiente paso es entender cómo encajan entre sí. Encontramos la respuesta en la Figura 3. Muestra que el software de data fabric proporciona una plataforma común de gestión de datos que permite que múltiples equipos de productores de datos, expertos en distintos dominios empresariales, puedan explorar, desarrollar y administrar datos y así como crear productos de datos de alta calidad y conformes a nivel normativo.

El software Denodo Platform ofrece una plataforma común para una ingeniería de datos y una gobernanza descentralizadas, y para compartir productos de datos en un data mesh.

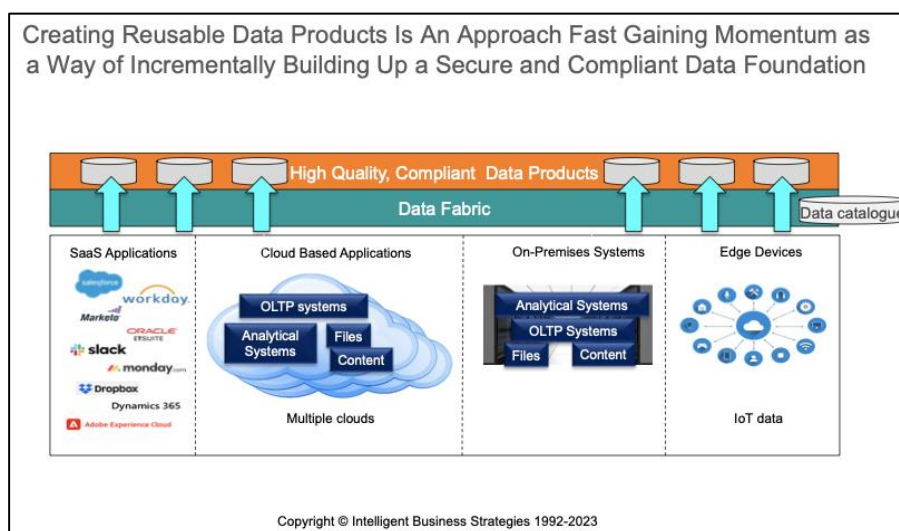


Figura 3

Un data fabric proporciona una plataforma común para una ingeniería de datos y gobernanza descentralizada de datos federados que cada equipo de productores de datos expertos en un ámbito empresarial pueden usar para:

- definir los términos de negocio comunes para cada producto de datos en un glosario empresarial del catálogo de datos asociado al data fabric;
- hacer un modelado del esquema de productos de datos o importarlo;
- conectar, explorar de forma automática y clasificar la información de origen necesaria para los productos de datos;
- crear pipelines de integración de datos para generar productos de datos almacenados y virtuales;
- enmascarar de forma dinámica datos confidenciales en productos de datos;
- definir políticas para gestionar el acceso a los productos de datos;
- publicar productos de datos en un data marketplace con términos empresariales en un glosario empresarial y el linaje completo de los metadatos;
- proporcionar seguridad de acceso universal a los productos de datos;
- aprovisionar datos dentro de productos de datos a través de una interfaz estándar como consultas SQL, REST o GraphQL API.

Definir, hacer un modelado, conectar, desarrollar, enmascarar, gestionar, publicar y aprovisionar productos de datos de una plataforma común de gestión de datos.

Aunque el data mesh no es para todos, es probable que cualquier organización que tenga dificultades con datos distribuidos necesite un software de data fabric.

CÓMO EL DATA FABRIC DE DENODO PLATFORM AYUDA A CREAR UN DATA MESH

Tras entender qué rol tiene el data fabric a la hora de gestionar, explorar e integrar datos para generar productos de datos administrados en un data mesh, explicamos en esta sección del documento cómo la tecnología de un proveedor de esta tecnología pretende conseguir estos objetivos. Ese proveedor es Denodo.

Denodo proporciona software de logical data fabric que se puede utilizar para desarrollar productos de datos en una data mesh.

La empresa Denodo, fundada en 1999, es líder en la gestión de datos y cuenta con muchos clientes en todo el mundo. Su producto estrella es Denodo Platform, una plataforma de gestión de datos que se puede utilizar como software de logical data fabric común para permitir a varios equipos de ingenieros de datos expertos en un ámbito empresarial acceder, explorar y organizar la ingeniería de datos para así integrar información, generar, usar y gestionar productos de datos en un data mesh.

La plataforma brinda una serie de funcionalidades de data fabric definidas anteriormente en este documento.

DENODO PLATFORM: FLEXIBILIDAD Y OPCIONES DE IMPLEMENTACIÓN

Denodo Platform está disponible con los siguientes tipos de suscripción:

Versión	Finalidad
Denodo Professional	Pequeños proyectos de departamentos
Denodo Standard	Múltiples casos de uso a nivel de departamento
Denodo Enterprise	en toda la empresa para múltiples casos de uso y grupos; incluye el catálogo de datos y la aceleración de consultas
Denodo Enterprise Plus	Mayor automatización mediante inteligencia artificial y machine learning, colaboración y seguridad avanzada para implementaciones en toda la empresa

Denodo Platform está disponible en distintas versiones y se puede configurar e implementar localmente, en múltiples nubes o como un servicio totalmente administrado.

La plataforma puede ejecutarse on-premises, en múltiples nubes (p. ej., AWS y Azure) y también como un servicio gestionado. La implementación en AWS y Azure se puede automatizar con el Denodo Platform Solution Manager, que permite configurar e iniciar rápidamente una configuración de clústeres según tus necesidades. También se ofrecen servicios de monitorización y auto-escalado. Una vez finalizado el despliegue, se conecta a la web del servicio de soporte de Denodo para conseguir actualizaciones de forma automática.

También puede ejecutarse en Kubernetes.

También es compatible con Kubernetes al poner a disposición imágenes de este contenedor oficiales en un registro público autenticado junto con gráficos de Helm para la configuración automática de:

- la seguridad de la capa de transporte (TLS), inicio de sesión único (SSO), red, JVM, registros, bases de datos de metadatos, etc.;
- el plano de control (gestión de sistemas, SSO, etc.);
- el plano de ejecución (DataPort Virtual o Data Catalog de Denodo, etc.);

Existen pods de Kubernetes para el clúster de servidores de Denodo Virtual DataPort, las herramientas de Denodo Platform (Design Studio, Data Catalog,

Solution Manager), así como el plano de control. El plano de control se ejecuta en la cuenta de Denodo en la nube, mientras que el plano de ejecución se ejecuta en la cuenta del cliente.

CONECTIVIDAD SEGURA DE DENODO PLATFORM

Denodo Platform brinda una conectividad segura a través de almacenes de credenciales (Key Vaults) y OAuth en bases de datos para:

- bases de datos relacionales a través de JDBC u ODBC, por ejemplo, Microsoft SQL Server, Oracle, IBM Db2, Snowflake, PostgreSQL, Exasol, SAP Hana y muchos más. Entre las funcionalidades que ofrece se hallan las de insertar, actualizar y eliminar datos en una fuente;
- bases de datos multidimensionales como Microsoft SQL Server Analysis Services, Oracle Essbase, Mondrian y SAP BW;
- bases de datos NoSQL como Neo4J, MongoDB, Cassandra, AWS DynamoDB;
- almacén de objetos en la nube como AWS S3, Azure Data Lake Storage (por API);
- Hadoop HDFS, Kudu, Hive e Impala;
- Delta en Databricks;
- aplicaciones SaaS;
- servicios web;
- servicios de mensajería JMS;
- temas de Kafka (se pueden usar como fuente y destino);
- datos semiestructurados XML y JSON;
- archivos delimitados como, por ejemplo, archivos de valores separados por comas (CSV);
- hojas de cálculo de Microsoft Excel;
- datos de páginas web como formularios web;
- índices de búsqueda —creados con Denodo Aracne— para consultar información no estructurada e integrar datos estructurados y semiestructurados;
- directorios LDAP;
- las BAPIs (interfaces de programación de aplicaciones empresariales) de SAP para extraer datos almacenados en aplicaciones SAP ERP.

Denodo Platform proporciona una conectividad segura a una amplia gama de fuentes de datos, incluyendo datos en streaming y aplicaciones SaaS.

Ofrece vistas predefinidas de datos en aplicaciones SaaS populares para simplificar y acelerar el acceso a los datos.

Con respecto a las aplicaciones SaaS, Denodo proporciona aceleradores SaaS que son vistas preconfiguradas en aplicaciones SaaS como Salesforce, ServiceNow, Marketo, Microsoft Dynamics 365, SharePoint y Google Apps.

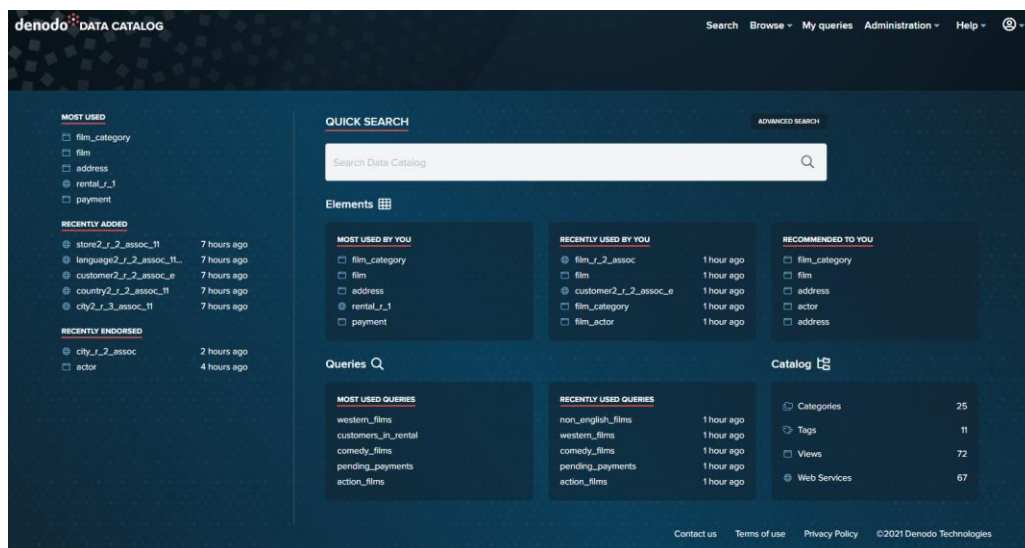
Además, Denodo dispone de una API para que los usuarios puedan crear sus propios conectores personalizados para extraer, insertar, actualizar o eliminar datos de fuentes específicas.

CATÁLOGO DE DATOS DE DENODO PLATFORM

Otro componente clave de Denodo Platform es su catálogo de datos. La Figura 4 muestra otra captura de pantalla al respecto.

También incluye un catálogo de datos mejorado por inteligencia artificial para ayudar a los usuarios a encontrar con rapidez datos relevantes.

El catálogo de datos de Denodo proporciona metadatos sobre los datos disponibles a los que se puede acceder mediante Denodo Platform. Cuenta con un cuadro de búsqueda y una funcionalidad de búsqueda avanzada para encontrar datos de forma más sencilla. También te permite guardar consultas de búsqueda para reutilizarlas en el futuro. Además, Denodo Data Catalog ofrece una serie de otras funcionalidades para ayudar a los usuarios a ser más productivos y encontrar lo que buscan con más rapidez.



Source: Denodo

Figura 4

Si nos fijamos en la Figura 4, se incluyen:

- los elementos de datos más utilizados por todos los usuarios de la organización, entre ellos las vistas virtuales y los servicios web;
- elementos de datos que se han añadido recientemente al catálogo;
- vistas y servicios web respaldados recientemente por una persona autorizada en tu empresa, como un propietario o administrador de datos;
- vistas virtuales y servicios web que el usuario ha utilizado o solicitado más a menudo recientemente;
- recomendaciones por inteligencia artificial que clasifican los elementos del catálogo que puedan interesar al usuario en función de su historial de interacciones con Data Catalog y estadísticas de uso recopiladas a partir de consultas anteriores efectuadas en el servidor de virtualización de datos Denodo Platform Virtual DataPort;
- consultas que el usuario lleva a cabo con más frecuencia;
- las consultas más recientes;
- el número de elementos gestionados en Denodo Data Catalog desglosado por tipo de elemento (p. ej. categorías, etiquetas, vistas, etc.).

INTEFRACIÓN DE DENODO PLATFORM CON CATÁLOGO DE DATOS DE TERCEROS

Además de su propio catálogo, Denodo Platform es capaz de integrarse con otros catálogos de datos de terceros como Alation, Apache Atlas, Collibra, IBM IGC, Informatica y data.world.

La integración con Collibra es bidireccional. Por ejemplo, los metadatos creados al explorar fuentes de datos en el catálogo de Collibra se pueden utilizar en Denodo, incluidos los datos clasificados como confidenciales por Collibra. Al importar estos metadatos, Denodo conocerá ya los tipos de datos confidenciales y pondrá en marcha políticas para gestionar el acceso a ellos a través de vistas virtuales. Además, los metadatos sobre las vistas virtuales de Denodo, las columnas de la vista y el linaje de datos de las dependencias de vistas se pueden importar dentro de Collibra y visualizar allí.

Muestra los datos y vistas más utilizados, añadidos y aprobados recientemente.

Recomienda información que pueda resultar interesante a los usuarios en función de su uso de datos.

Denodo Platform también puede usar y proporcionar metadatos a catálogos de datos y herramientas de business intelligence de terceros.

Denodo Data Catalog también es compatible con herramientas de business intelligence de terceros como Microsoft Power BI y Salesforce Tableau.

DESARROLLO COLABORATIVO, REPALDADO POR INTELIGENCIA ARTIFICIAL, DE DENODO PLATFORM

Denodo Platform permite una ingeniería descentralizada de productos de datos en un data mesh.

Las versiones Enterprise y Enterprise Plus de Denodo Platform están pensadas para los múltiples equipos de productores de datos que desarrollan productos de datos en un data mesh. Permite a informáticos y a citizen data engineers crear productos de datos mediante vistas virtuales que integran datos sin la necesidad de replicarlos. Los productos de datos que crea un equipo pueden utilizarlos otros equipos sin la necesidad de copiar datos.

Denodo Data Catalog puede compartir metadatos entre varios equipos para permitir el desarrollo colaborativo y el uso repetido.

Denodo Data Catalog fomenta la colaboración al compartir metadatos entre múltiples equipos descentralizados de productores y consumidores de datos. Entre ellos, comparte metadatos sobre términos empresariales comunes (glosario empresarial), fuentes de datos, recursos de datos disponibles en fuentes de datos, relaciones entre los datos, clasificaciones de datos, linaje de datos de fuentes a vistas y dependencias de vistas, categorías, vistas virtuales, servicios web de consulta y etiquetas.

Los miembros del equipo pueden añadir comentarios, propiedades personalizadas, etiquetas, recomendaciones y plantear preguntas. Asimismo, pueden ver mensajes de advertencia y avisos de desuso en vistas virtuales no utilizadas o usadas con poca frecuencia. Este tipo de colaboración suma de forma incremental el conocimiento de expertos, lo que se traduce en una mayor productividad. Este aspecto es relevante en un entorno de ingeniería de datos descentralizado que suele caracterizar las implementaciones de un data mesh.

La inteligencia artificial proporciona recomendaciones guiadas para ayudar a los usuarios a encontrar datos relevantes con más rapidez.

Además, Denodo Platform permite un desarrollo respaldado por inteligencia artificial, que hace recomendaciones basadas en los datos (Figura 4) para guiar y ayudar a los usuarios a encontrar datos relevantes con más rapidez. También puede inferir esquemas y relaciones de datos cuando se conecta a fuentes de datos.

Control de versiones, pruebas automatizadas, implementación rápida y reversión de cambios hacia y desde entornos de producción.

Para gestionar el desarrollo colaborativo, Denodo ofrece CI/CD. Esto implica, entre otras cosas, la posibilidad de comparar las diferencias de elementos y esquemas entre el origen y el destino y en los entornos de desarrollo, prueba y producción. También permite cambios rápidos de aprobación y reversión en miles de vistas virtuales en estos entornos. Todo esto ayuda con el control de versiones y la posibilidad de asociar versiones con solicitudes de cambio en toda la empresa.

Para más detalles sobre cómo desarrollar de forma incremental productos de datos en un data mesh con un planteamiento en capas en Denodo Platform, se puede consultar el artículo [El valor de la virtualización de datos en un data mesh.](#)

Los flujos de autorización controlan, administran y rastrean el estado de todas las solicitudes de acceso a datos y cambios relativos a los mismos.

Gracias a una nueva colaboración ampliada, Denodo Data Catalog ahora permite flujos de autorización para permitir la creación y gestión de solicitudes de acceso. Mediante los flujos de autorización, los usuarios pueden solicitar acceso a datos, informar un problema, solicitar un cambio o hacer preguntas. Los flujos de autorización de Denodo Data Catalog pueden integrarse con sistemas externos como ServiceNow o Jira donde las solicitudes y preguntas se pueden almacenar y administrar. Como alternativa, aquellos clientes que no dispongan de un sistema externo adecuado pueden gestionar dentro del Denodo Data Catalog flujos de autorización sencillos de acceso a datos,

informes de incidencias, etc. Además, Denodo Data Catalog puede obtener el estado de cualquier solicitud de cambio y acceso llevada a cabo desde herramientas externas para poder consultar todas las solicitudes y hacer un seguimiento de su estado.

RENDIMIENTO Y ESCALABILIDAD DE DENODO PLATFORM

Otros elementos importantes del data fabric son su rendimiento y escalabilidad. Denodo Platform ofrece ambos de varias maneras.

Ante todo, permite la delegación del procesamiento de consultas y su optimización para que los sistemas de origen de datos subyacentes hagan la mayor parte del trabajo de filtrado y recuperación de datos relevantes. De esta manera, los datos se procesan en el origen y se minimiza la cantidad de datos que se necesita transferir a Denodo Platform.

En segundo lugar, Denodo también analiza y hace un seguimiento de las consultas activas y del procesamiento de consultas de forma automática mediante inteligencia artificial:

- predice y decide qué datos almacenar en caché de forma dinámica para maximizar el rendimiento;
- recomienda los mejores resúmenes para aumentar el rendimiento según el historial de uso y las simulaciones de costes para generarlos.

Asimismo, puede analizar consultas anteriores recurriendo a la inteligencia artificial para detectar de manera automática patrones para clasificar consultas similares e identificar así conjuntos de resultados intermedios frecuentes y comunes para cada grupo de consultas similares (summaries). A continuación, puede precalcular los datos resultantes intermedios que maximizan el rendimiento en la carga de trabajo. El optimizador luego los usará durante la ejecución de consultas en tiempo de ejecución para agilizar el procesamiento de consultas que siguen patrones similares.

Un tercer elemento clave para el rendimiento es el motor de procesamiento paralelo masivo (MPP) basado en Presto e integrado en Denodo Platform. Su relevancia se debe al hecho de que muchas empresas ahora disponen de ingentes cantidades de datos almacenados en archivos Parquet en el almacenamiento de data lake, que suele ser o bien en la nube como Amazon S3 y Microsoft Azure Data Lake Storage (ADLS) o bien de tipo local en Hadoop Distributed File Systems (HDFS). El problema de esta solución es que los sistemas de almacenamiento no pueden optimizar las consultas. El motor MPP integrado en Denodo Platform resuelve este inconveniente al optimizar la parte de una consulta distribuida que necesita acceder de manera eficiente a los datos en el almacenamiento del data lake incluido en las vistas virtuales. El motor MPP integrado de Denodo proporciona funcionalidades integradas de seguridad, almacenamiento en caché de gestión, optimización basada en costes y procesamiento paralelo para agilizar las consultas. A continuación se muestra un ejemplo de una consulta que combina millones de filas de datos de archivos de Parquet de S3 y Oracle.

```
SELECT i_category, c_last_name, SUM(ss_net_paid_inc_tax) AS price
FROM tpcds_parquet.store_sales
  INNER JOIN tpcds_oracle.customer ON ss_customer_sk = c_customer_sk
  INNER JOIN tpcds_oracle.item ON i_item_sk = ss_item_sk
GROUP BY i_category, c_last_name
```

Fuente: Denodo

Delegación de consultas, optimización basada en costes, almacenamiento automatizado en caché impulsado por inteligencia artificial, conjuntos de resultados intermedios comunes y recomendaciones resumidas.

Procesamiento paralelo masivo de consultas en grandes volúmenes de datos almacenados en data lakes basados en almacenamiento de objetos.

Finalmente, Denodo Platform brinda un escalado automatizado a través de la configuración dinámica de clústeres.

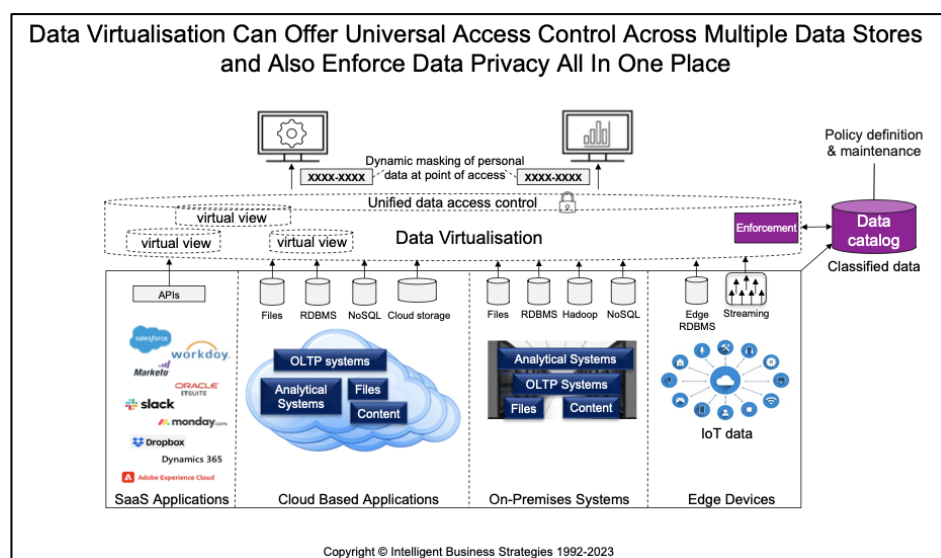
EXTENSIBILIDAD DE DENODO PLATFORM

Además de la funcionalidad out-of-the-box, Denodo Platform tiene un lenguaje similar a SQL llamado PL/VQL, con el que podrás ampliar las funcionalidades de la plataforma creando lógica basada en procedimientos. Estos procedimientos se pueden usar para seleccionar de forma dinámica fuentes de datos alternativas, gestionar transformaciones complejas de datos en pipelines, definir reglas más eficaces, automatizar el etiquetado de datos, mejorar la seguridad y mucho más.

Código personalizado para ampliar las funcionalidades de la plataforma.

GESTIÓN UNIFICADA DE DATOS DE DENODO PLATFORM

En términos de gestión de datos, Denodo Platform proporciona una capa semántica común que utiliza términos del glosario empresarial y con la que las organizaciones pueden cumplir de forma global con la seguridad de acceso a los datos y la privacidad de los mismos en múltiples almacenes de datos desde una única plataforma (ver Figura 5).



Control de acceso a datos basado en atributos y roles universales a datos en múltiples almacenes de datos desde una única plataforma.

Figura 5

Denodo Platform brinda un control de acceso universal a los datos en múltiples almacenes de datos y aplicaciones al restringir lo que los usuarios pueden hacer. Asimismo, permite la administración de roles, de privilegios basados en roles asociados con colecciones de catálogos de datos u objetos de catálogo únicos y control de acceso basado en roles. Se pueden otorgar privilegios para controlar quién puede administrar y/o crear fuentes de datos y vistas, así como ejecutar consultas. Es posible elegir entre un control de acceso basado en roles y uno basado en atributos que se detallarían en cada caso.

El enmascaramiento dinámico de datos y las vistas virtuales para excluir los tipos de datos confidenciales detectados de forma automática ayudan a garantizar el cumplimiento de la legislación relativa a la privacidad.

Con respecto a la privacidad de los datos, los metadatos de la información clasificada como confidencial se pueden gestionar dentro de Denodo Platform o se pueden importar a esta plataforma desde catálogos de datos de terceros como Collibra. A continuación, la plataforma puede hacer cumplir la privacidad de los datos de conformidad con las obligaciones legales, puesto que permite generar vistas virtuales que excluyan campos confidenciales para ocultar datos confidenciales. También es posible definir restricciones de fila con condicionales y enmascarar de forma dinámica los datos confidenciales al consultarlos para evitar que los vean usuarios no autorizados.

MARKETPLACE DE DATOS DE DENODO PLATFORM

Un data marketplace que utiliza Denodo Data Catalog para controlar la publicación, la búsqueda y el consumo de productos de datos en un data mesh.

Denodo Data Catalog también se puede utilizar para crear un marketplace de datos para publicar productos de datos virtuales listos para usar en ámbito empresarial en un data mesh. Estos productos de datos se pueden aprovisionar sin la necesidad de copiar datos.

Para simplificar su uso, el catálogo de datos se integra con herramientas populares de business intelligence para permitir a los analistas empresariales ver los productos de datos disponibles sin tener que abandonar la herramienta que están usando. Los asistentes para consultas de Denodo Data Catalog facilitan aún más la dinámica de uso, al crear consultas para acceder a estos datos. Los productos de datos se pueden usar mediante SQL y a través de las APIs.

CONCLUSIONES

Es necesario superar la complejidad de las fuentes de datos y los obstáculos relativos a la ingeniería de datos si las empresas quieren conseguir una arquitectura de datos sólida que permita ofrecer datos de alta calidad.

Las soluciones a estos problemas son un data fabric y data mesh.

El data mesh ayuda a aumentar la cantidad de data engineers y proporciona un planteamiento incremental para conseguir una arquitectura de datos sólida que ofrezca datos de alta calidad.

Un modelo operativo federado y una oficina encargada de gestionar el programa, administrada por un responsable de datos, también son importantes.

El software de data fabric de Denodo Platform puede ayudar a las empresas a implementar un data mesh y conseguir más resultados en menos tiempo.

There is also a bottleneck in data engineering and as demand for insights grows to improve business outcomes, so does the need to accelerate data engineering. No hay duda de que, para lograr el éxito de una empresa impulsada por los datos ("data-driven"), primero deben conseguir una arquitectura de datos que permita datos de alta calidad, seguros y conformes a nivel normativo, a partir de los cuales desarrollar información empresarial valiosa y competitiva. El principal desafío a la hora de hacerlo es la complejidad debida al número cada vez mayor de fuentes y almacenes de datos dispersos en entornos híbridos de datos distribuidos en múltiples nubes. Asimismo, existe un reto relativo a la ingeniería de datos y, a medida que crece la demanda de información para mejorar los resultados empresariales, también crece la necesidad de acelerar la ingeniería de datos.

Existen actualmente dos soluciones para superar estos problemas. La primera es el data mesh, que supone un planteamiento para democratizar la ingeniería de datos y así superar el cuello de botella. De hecho, permite que varios equipos de ingenieros de datos especializados en determinados ámbitos empresariales desarrollen de manera gradual un conjunto de productos de datos reutilizables y listos para usar con identificadores empresariales únicos que se pueden compartir y usar dentro de la organización para múltiples cargas de trabajo analíticas. Es una manera de iniciar a construir los cimientos de una arquitectura de datos y, a la vez, aumentar el número de ingenieros de datos.

Sin embargo, para que el data mesh tenga éxito, se precisa de una estructura organizativa federada con una oficina encargada de gestionar el programa, administrada por un responsable de datos (CDO), que coordine toda la actividad de ingeniería de datos, expertos en la materia y una reorganización del departamento informático para crear un centro de formación e incorporar ingenieros de datos profesionales en los dominios empresariales para mejorar sus competencias y ayudarles a adoptar buenas prácticas.

El segundo requisito para tener éxito es una plataforma de datos común para conectarse de forma segura a todas las diferentes fuentes de datos, almacenes de datos, flujos de datos y aplicaciones en un entorno híbrido multinube de datos distribuidos. De esta forma, es mucho más sencillo para los usuarios saber qué datos hay, gestionarlos de manera fácil para crear productos de datos en un data mesh y ponerlos a disposición para su uso. Esa plataforma común de administración de datos la proporciona el software de data fabric que puede implementarse on-premises, en la nube o como un servicio administrado e incluir un catálogo de datos para entender qué datos están disponibles. Asimismo, tiene que permitir el intercambio de metadatos entre varios equipos, la colaboración y el desarrollo de productos de datos respaldado por inteligencia artificial, ser escalable, proporcionar extensibilidad y un data marketplace para publicar productos de datos, a la vez que brinda un gobernanza de datos unificada. Se trata de una gran cantidad de funcionalidades que necesitan estar disponibles en una única plataforma de data fabric.

Denodo Platform ofrece una plataforma común para que varios equipos puedan encontrar, desarrollar, publicar y conectarse a productos de datos en un data mesh.

Aunque el data mesh no es para todos, es probable que cualquier organización que tenga dificultades con datos distribuidos precise de un software de data fabric.

El software de data fabric de Denodo Platform ofrece estas funcionalidades. Las empresas que buscan acelerar la ingeniería de datos para conseguir más resultados en menos tiempo al crear un data mesh requieren una plataforma común de gestión de datos. Necesitan software de data fabric y Denodo Platform es un fuerte candidato.



Sobre Intelligent Business Strategies

Intelligent Business Strategies es una empresa independiente de investigación, formación y consultoría cuyo objetivo es ayudar a las empresas a comprender y aprovechar al máximo los nuevos desarrollos en business intelligence, machine learning, análisis avanzado, gestión de datos, big data e integración empresarial. Juntas, estas tecnologías ayudan a las empresas a convertirse en negocios inteligentes.

Autor



Mike Ferguson es director general de Intelligent Business Strategies Ltd. Como analista y consultor independiente del sector informático, es especialista en business intelligence, análisis y gestión de datos. Con más de 40 años de experiencia en el campo informático, Mike ha asesorado a decenas de empresas sobre business intelligence, análisis, estrategia de datos, selección de tecnología, arquitectura empresarial y gestión de datos. Mike también es presidente de «Big Data LDN», la mayor conferencia de datos y análisis en Europa y es miembro del consejo ejecutivo y asesor EDM CDMC. Ha impartido ponencias en eventos en todo el mundo y es autor de numerosos artículos. Anteriormente, fue director y cofundador de Codd & Date Europe Ltd, los inventores del modelo relacional, *chief architect* en Teradata para sus sistemas de administración de bases de datos (DBMS), así como director general europeo de Database Associates. Imparte seminarios y ponencias muy populares sobre modernización de data warehouses, big data, gobernanza centralizada de entorno de datos distribuidos, recomendaciones prácticas para implementar un data mesh, machine learning y análisis avanzado, análisis integrado, aplicaciones inteligentes y automatización mediante inteligencia artificial.



Teléfono: (+44)1625 520700

Web: www.intelligentbusiness.biz

E-Mail: info@intelligentbusiness.biz

Data fabric, data mesh o los dos: ¿cuál le conviene a tu empresa?

Copyright © 2023, Intelligent Business Strategies

Todos los derechos reservados.

Todas las capturas de pantalla y diagramas facilitados por Denodo que aparecen en este documento siguen siendo propiedad intelectual de Denodo, que por lo tanto detiene los relativos derechos de autor.