Mi pregunta

Algunos autores consideran que los data lakes y los data warehouse ya no son la mejor solución. ¿Qué se propone como mejora para este modelo?

Respuesta

¡Has dado en el clavo con la pregunta más importante y actual en el mundo de la arquitectura de datos! Efectivamente, muchos autores y líderes de la industria (incluyendo a Piethein Strengholt en "Data Management at Scale") argumentan que los *data warehouses* y *data lakes* centralizados son un modelo del pasado que ya no escala.

La mejora que se propone no es simplemente una "versión 2.0" de lo mismo, sino un **cambio de paradigma fundamental** que se aleja de la centralización. La principal propuesta, y la que está generando más debate y adopción, es el **Data Mesh**.

Aquí te explico de forma sencilla el problema y la solución propuesta.

Primero, ¿Cuál es el Problema de los Data Warehouses y Data Lakes?

A pesar de sus diferencias, ambos comparten un pecado original: la **centralización**.

1. **El Cuello de Botella Central**: En ambos modelos, un **equipo central de datos** es el responsable de ingerir, limpiar, procesar y servir todos los datos de la empresa. Este equipo se convierte inevitablemente en un cuello de botella. Los equipos de negocio (Marketing, Ventas, etc.) tienen que hacer cola, esperando semanas o meses para obtener los datos que necesitan. Esto frena la innovación y la agilidad.

2. **La Pérdida de Contexto**: El equipo central no entiende el negocio tan bien como los equipos que generan los datos. Al intentar "limpiar" y "modelar" los datos de marketing, pueden malinterpretarlos o quitarles un contexto crucial.

3. **El Data Lake se Convierte en un Pantano (*Data Swamp*)**: La promesa del *data lake* era la flexibilidad ("tira todos tus datos aquí y ya veremos"). En la práctica, sin un gobierno claro y con un equipo central sobrecargado, el lago se llena de datos de mala calidad, sin documentación y duplicados, convirtiéndose en un pantano inútil.

La Mejora Propuesta: El Data Mesh

El **Data Mesh** es un cambio radical en la forma de pensar sobre la arquitectura y la organización de los datos. Fue propuesto por Zhamak Dehghani y se basa en una idea simple pero poderosa: **descentralizar la propiedad y la gestión de los datos**.

En lugar de un gran lago central, imagina una **red (una "malla") de productos de datos interconectados**, donde cada nodo es propiedad de un equipo de negocio.

El Data Mesh se sostiene sobre cuatro principios fundamentales:

1. Propiedad de los Datos Descentralizada y Orientada al Dominio

- **¿Qué significa?**: Se acaba el equipo central. La responsabilidad de los datos vuelve a los equipos que los generan y mejor los entienden: los **dominios de negocio**. El equipo de Marketing es ahora el dueño y responsable de los datos de marketing. El equipo de Logística es dueño de los datos de la cadena de suministro.

- **La Ventaja**: Los datos son gestionados por los expertos en la materia. La calidad y el contexto mejoran drásticamente. Se elimina el cuello de botella central.

2. Datos como un Producto (Data as a Product)

- **¿Qué significa?**: Este es el cambio de mentalidad más importante. Los dominios no deben tratar sus datos como un residuo de sus operaciones. Deben tratarlos como un **producto de primera clase** que ofrecen al resto de la empresa.

- **¿Qué implica un "producto de datos"?**: Al igual que un producto de software, un producto de datos debe ser:

  - **Descubrible**: Fácil de encontrar (ej: en un catálogo de datos central).

  - **Direccionable**: Con una ubicación única y permanente.

  - **Confiable y de Alta Calidad**: El equipo propietario garantiza su calidad.

  - **Autodescriptivo**: Bien documentado, con metadatos claros.

  - **Seguro**: Con políticas de acceso claras.

  - **Interoperable**: Servido en formatos estándar que todos puedan usar.

3. Plataforma de Datos de Autoservicio (Self-Serve Data Platform)

- **¿Qué significa?**: Forzar a cada equipo de marketing a convertirse en experto en infraestructura de *Big Data* sería un desastre. Para evitarlo, se crea un **equipo de**

plataforma central cuyo trabajo es proporcionar las herramientas y la infraestructura como un servicio.

- **La Analogía**: El equipo de plataforma no cocina la comida para todos (como hacía el equipo del *data warehouse*). En cambio, construye una **cocina industrial de última generación** con las mejores herramientas, para que cada chef (cada equipo de dominio) pueda cocinar sus propios platos (sus productos de datos) de forma fácil y eficiente.

4. Gobernanza Computacional Federada (Federated Computational Governance)

- **¿Qué significa?**: ¿Cómo evitamos el caos si cada equipo hace lo que quiere? La solución no es un gobierno central autoritario, sino una **federación**.

- **¿Cómo funciona?**: Se crea un "gremio" o un comité con representantes de cada dominio y del equipo de plataforma. Juntos, deciden y automatizan las **reglas globales** que todos deben seguir (ej: estándares de seguridad, políticas de interoperabilidad, formatos de metadatos). El cumplimiento de estas reglas se automatiza y se integra en la plataforma de autoservicio.

Otras Mejoras Relacionadas: Data Fabric y Lakehouse

Aunque el Data Mesh es el cambio de paradigma más grande, hay otros dos conceptos que también se proponen como mejoras:

- **Data Fabric**: Es un enfoque más **tecnológico** que organizacional. En lugar de cambiar quién es el dueño de los datos, propone crear una **capa de virtualización inteligente** que se conecta a todas las fuentes de datos existentes (warehouses, lakes, bases de datos operacionales). Usando metadatos activos e IA, el *data fabric* entiende qué datos hay, dónde están y cómo se relacionan, ofreciendo una vista unificada y simplificando el acceso sin mover físicamente los datos.

  - **Diferencia clave**: El Data Mesh descentraliza la arquitectura y la organización. El Data Fabric crea una capa inteligente **sobre** la arquitectura existente, que puede seguir siendo centralizada.

- **Data Lakehouse**: Esta es una **evolución**, no una revolución. Es la idea de combinar lo mejor de los *data warehouses* (estructura, gobernanza, transacciones ACID) con lo mejor de los *data lakes* (flexibilidad, bajo costo, datos no estructurados) en una única plataforma. Tecnologías como Delta Lake, Apache Iceberg y Apache Hudi permiten implementar esta visión.

En resumen, la principal mejora propuesta al modelo centralizado de *data warehouses* y *data lakes* es el Data Mesh, que aboga por una descentralización radical de la propiedad de los datos, tratándolos como productos gestionados por los propios dominios de negocio.