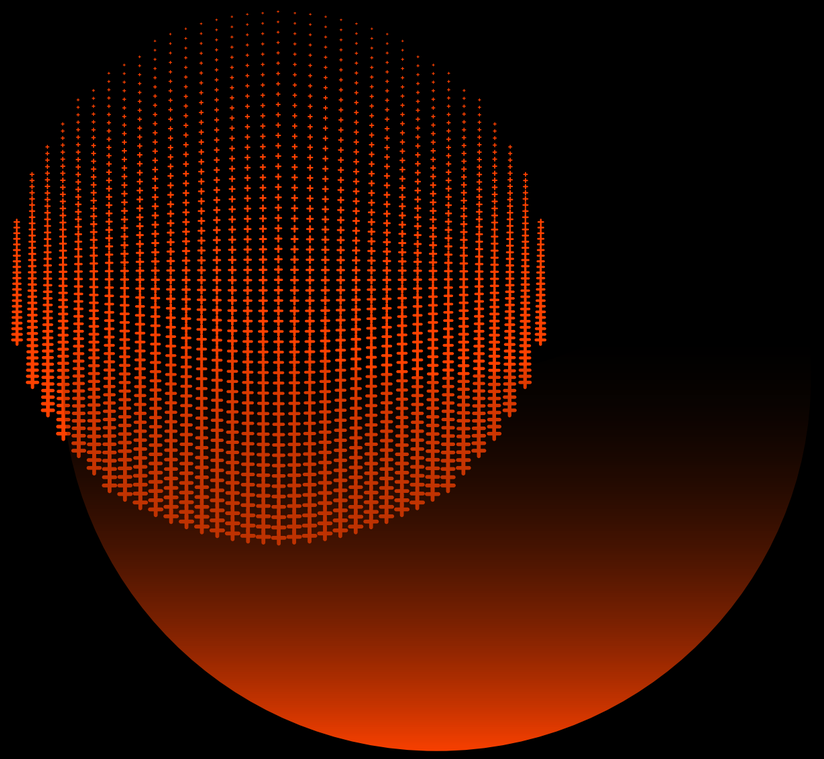
e-book

Cloudera Data Mesh

Lectura - 20min

índice



Cloudera Data Mesh, qué es, ventajas, casos de uso 1

¿Por qué apoyarse en el Data Mesh? 2

Casos de uso de Data Mesh 3

¿Qué no es un Data Mesh? 4

Implementación Cloudera para Data Mesh 5

1

Data Mesh: qué es, ventajas y casos de uso

El mundo de los datos evoluciona a ritmos muy altos con el objetivo de hacer los procesos de gestión más fáciles y certeros.

¿Qué es Data Mesh?

El Data Mesh divide datos almacenados en cualquier repositorio en muchos dominios de negocio que contarán con un equipo de profesionales que se van a encargar de crear productos de datos para hacérselos llegar a diversas áreas de la organización.
El Data Mesh es un paso más hacia la personalización de los productos o servicios que las empresas ofrecen a cada cliente, pues jerarquiza los niveles de datos y, a su vez, hace los procesos más rápidos para las necesidades digitales que demanda el mercado.
El usuario navega desde más de un dispositivo y las sincronizaciones entre ellos deben ser ágiles. El Data Mesh permite esto, que los productos de datos se unan entre dominios permitiendo el intercambio de datos sin almacén.

Principios del Data Mesh

Para entender mejor cómo puede el Data Mesh revolucionar el mundo de los datos, vamos a explicar los 4 principios que lo diferencian

Gestión de los datos de forma independiente

El hecho de que no haya un almacén ofrece una gestión mucho más cómoda y barata mediante dominios, sin almacén. Es una gestión de datos independiente.

Datos usables y sencillos como si fueran un producto

Los datos analíticos proporcionados por los dominios deben tratarse como un producto, y los consumidores de esos datos deben tratarse como clientes. Para que esto funcione, el propietario del producto de datos de dominio debe tener un conocimiento profundo de quiénes son los usuarios de esos datos, cómo los usan y cuáles son los métodos con los que se sienten cómodos para consumirlos. Este conocimiento de los usuarios permite el diseño de interfaces de productos de datos que satisfacen sus necesidades.

Principios del Data Mesh

Plataforma de autoservicio de datos

Los integrantes de cada dominio pueden acceder a ella para hacer uso de los datos,
requiere de alta tecnología si se quiere acceder a los productos de los datos correctamente.

Gobernanza federada

Los conjuntos de datos son independientes, pero para serlo de verdad, para no necesitar un almacén, deben ser capaces de interoperar entre sí. De esta forma, se crean uniones, se encuentran intersecciones, se realizan gráficos y operaciones a gran escala

2

¿Por qué apoyarse en el Data Mesh?

Las formas tradicionales de la analítica no están dando bien en cuanto al costo beneficio.
Las empresas invierten mucho tiempo y dinero en la integración de plataformas digitales; mientras, descuidan un factor crítico: sus arquitecturas tecnológicas monolíticas son engorrosas, costosas e inflexibles.

A esto se le suma que:

· El 70-80% de las transformaciones digitales fallan.
· El auge de las arquitecturas distribuidas, aplicaciones, datos y arquitecturas en la nube se está volviendo menos centralizado/monolítico.
· El bloqueo en la nube es real y puede volverse más costoso.
· Los lagos de datos rara vez tienen éxito y solo se centran en el análisis.
· Los silos organizacionales exacerban los problemas de intercambio de datos.
· El mundo se está acelerando: el ritmo de la innovación y la competencia.

El costo de las interrupciones de datos operativos está aumentando.

El Data Mesh no es un modelo que resuelve todos los problemas, pero sí provee un camino para iniciar la transformación y pasar la página de los proyectos analíticos fallidos.

El Data Mesh comparte los principios, las prácticas y las tecnologías que deben alinearse para resolver los desafíos principales de la modernización de las arquitecturas de datos, con el fin de responder a los problemas comerciales, a través de los datos.

3

Casos de uso de Data Mesh

Aplicaciones hay muchas y muy variados, tanto para la gestión operativa de la información como desde su punto de vista de explotación y analítica.

Modernización de aplicaciones

Las nuevas arquitecturas de microservicios requieren de un movimiento estructurado de datos, que independice el origen de los mismos y lo estructure por dominios y reciprocidad de la información.

Continuidad de negocio y disponibilidad del dato

En organizaciones globales es tremendamente complejo velar por la integridad del dato y su disponibilidad en diferentes regiones. Esta geodistribución del dato puede ser la solución para transacciones de baja latencia.

Modernización de aplicaciones

La integración de datos basada en eventos permite derribar las barreras de gestión de estos entre los diferentes contextos de cada grupo de microservicios.

Integración basada en eventos

Las grandes organizaciones tienen una mezcla de sistemas antiguos y nuevos, aplicaciones monolíticas y microservicios, almacenes de datos operativos y analíticos. Data Mesh puede ayudar a unificar estos recursos a través de diferentes dominios de negocio y datos.

Ingesta de datos continua para analítica

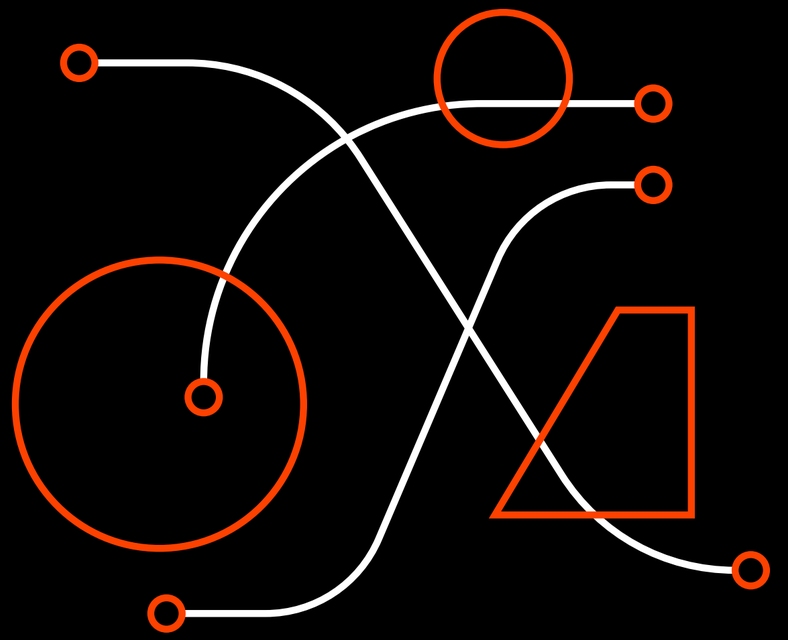
En general hay dos formas de cargar datos en los almacenes de datos analíticos: con procesos batch o mediante ingesta continua (streaming). Una aproximación Data Mesh puede ayudar a resolver el problema de gestionar diferentes tipos de eventos como son los orígenes de datos, las propias aplicaciones o dispositivos.

Canalización continua de datos o streaming data pipeline

Una malla de datos puede añadir un gobierno del dato independiente que conecte con los almacenes de datos analíticos, proporcionando y descubriendo los datos, gestión entre dominios, preparación y consistencia del dato.

Streaming analytics

Hay eventos que ocurren continuamente y pueden analizarse por flujos en tiempo real, algo determinante para procesos como la telemetría o en contextos tan complejos como un partido de futbol.



4

¿Qué no es un Data Mesh?

El Data Mesh es un modelo organizacional que combina las arquitecturas de datos descentralizadas, las acciones impulsadas por eventos y un estilo de diseño de microservicios centrado en la transmisión de información.

El Data Mesh no es

- Lago de datos en una sola nube, «dominios», catálogos y acceso SQL.
- Catálogo de datos: el Data Mesh necesita una implementación física.
- Productos específicos. Ningún proveedor tiene un producto único para Data Mesh.
- Fábrica de datos. Estas incluyen muchas arquitecturas de datos monolíticas.
- Analítica de autoservicio súper fácil de usar.

5

Implementación Cloudera para Data Mesh

La implementación Data Mesh de Cloudera es sin duda la mas completa del mercado, considerando su nacimiento como única empresa que implemento exitosamente Data Lake´s a costo beneficio razonable, sin duda pudo extender su visión sobre la gestión descentralizada de datos.

Cloudera propone la utilización de su plataforma Shared Data Experience (SDX) y Cloudera Data Platform (CDP) para ayudar a las organizaciones a resolver las limitaciones de las tradicionales plataformas monolíticas y estar listas para dar soluciones descentralizadas y orientadas al self service de datos.

CLOUDERA

Cloudera Data Platform (CDP) es una plataforma de datos híbrida diseñada para proporcionar autonomía a la hora de utilizar cualquier nube, cualquier análisis, cualquier dato.

CDP ofrece una gestión de datos más rápida y sencilla y análisis de datos sobre cualquier lugar, con rendimiento, escalabilidad y seguridad óptimos.

Con CDP obtiene todos los beneficios que ofrecen tanto la nube privada como la pública; esto permite maximizar la rentabilidad de la inversión ya realizada, como así también tener control centralizado de ambos mundos.

CDP brinda la libertad de mover aplicaciones, datos y usuarios de forma segura bidireccionalmente entre el centro de datos y múltiples nubes de datos, independientemente de dónde viven. Todo gracias a las modernas arquitecturas de datos:

- Un Data Fabric unificado que organiza de forma centralizada fuentes de datos dispares de forma inteligente y segura a través de múltiples nubes y on premise.
- Un Data Lakehouse abierto permite el análisis multifunción tanto de la transmisión como de los datos almacenados en un almacén de objetos nativo de la nube a través de múltiples nubes híbridas.
- Un Data Mesh escalable que ayuda a eliminar los silos de datos mediante la distribución de la propiedad a los equipos, manteniendo una infraestructura de datos común.

Common Control Plane

El plano de control común en CDP proporciona un servicio ubicuo que es consistente y abarca las instancias de implementación de una organización. En el diagrama se muestra cómo una instancia de nube pública comparte servicios como el de gobernanza con la instancia de nube privada.

El plano de control es un servicio federado que permite que los metadatos, la seguridad, el cifrado y la gobernanza sean gestionados como un servicio centralizado pero federado. Los bloques de construcción fundamentales se construyen sobre componentes de código abierto y tienen una API accesible que proporciona integración a un ecosistema más amplio de servicios y además admite estándares abiertos e interoperabilidad.

Data Catalog

El catálogo de datos de CDP se encuentra dentro del plano de control común. Aborda muchas de las cualidades fundamentales de la construcción hacia un enfoque de datos como producto.

Hace que los datos sean detectables a través de los nodos de la malla, proporciona acceso a los puntos finales para hacer los datos direccionables y captura metadatos seleccionados por el usuario para determinar la integridad de los productos de datos.

Además de los metadatos mejorados por el uso, los metadatos técnicos y comerciales pueden ser capturados de manera integral para auto describir los productos.
Finalmente, todos los datos están protegidos en reposo y en tránsito por encriptación de nivel FIP 140-2, y se almacena en formatos y estándares abiertos.
Esto permite el enfoque de datos de dominio como producto.

Shared Data Experience (SDX)

Cloudera SDX combina seguridad, gobierno y gestión centralizados de nivel empresarial capacidades con metadatos compartidos y un catálogo de datos. Proporciona una capa de gobernanza que abarca planos de control e instancias de implementación para asignar propiedad, capturar auditoría y aplicar políticas globales.

Cloudera SDX está federado y administrado como un servicio compartido en un entorno global.
Gobernanza federada para habilitar ecosistemas e interoperabilidad.

Data Hub

Data Hub permite a los usuarios implementar clústeres analíticos en todo el ciclo de vida de los datos mediante la implementación elástica de IaaS. Proporciona gran control sobre las configuraciones de clústeres, incluidos hardware y componentes de servicios. Su diseño nativo en la nube admite separación de cómputo y almacenamiento. Eso proporciona compatibilidad con el escalamiento automático de recursos en función del entorno.

Self-service Data Services

Los servicios de datos proporcionan aplicaciones analíticas informáticas en contenedores que escalan dinámicamente y se puede actualizar de forma independiente. Mediante el uso de contenedores desplegados en la nube gestionados por servicios de Kubernetes como Amazon EKS, Microsoft Azure AKS y Google GKE, los usuarios pueden implementar clústeres similares a lo que es posible en Data Hub, pero con la ventaja adicional de ser entregado a través de una experiencia PaaS.

Agnostic

www.agnosticit.com

Impulsando la
transformación digital
de las empresas