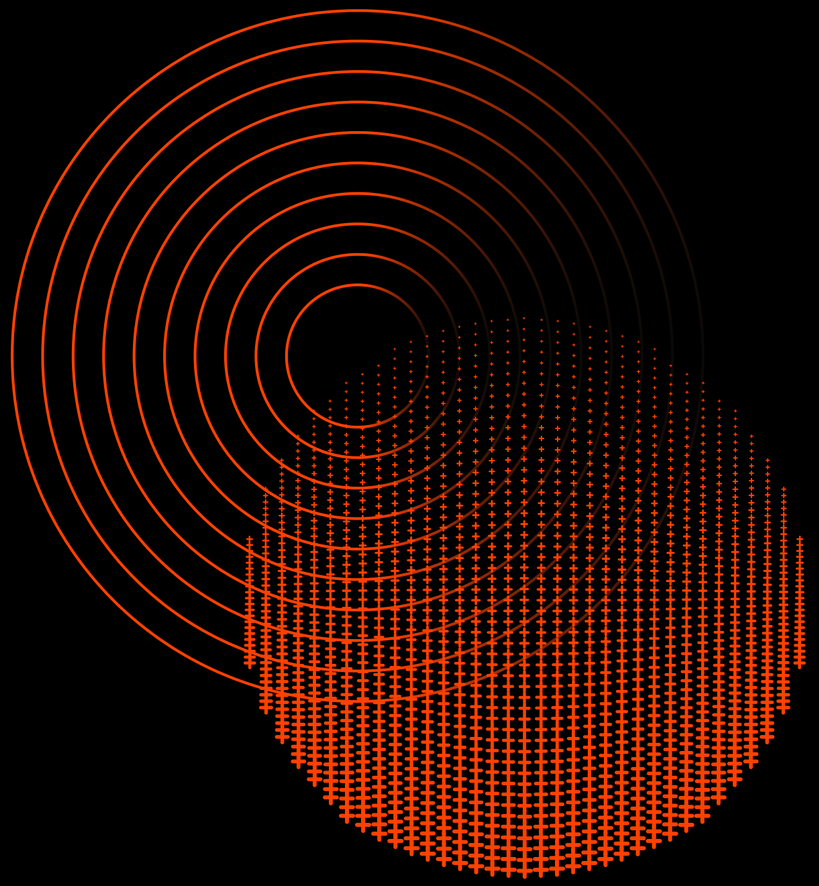
e-book

Cloudera Data Fabric

Lectura - 20min

índice



1

¿Qué es Data Fabric?

¿Qué es Data Fabric?

Data Fabric es un concepto que ha sonado entre las principales tendencias en tecnologías Cloud durante todo este 2021. En un mundo en el que cada vez se generan más datos, que son más diversos al proceder de muy distintas fuentes, gestionarlos y analizarlos se convierte en una prioridad de primer orden para las empresas.

Data Fabric es una capa de arquitectura que conecta los datos y los procesos analíticos. Es un entramado, una estructura o un tejido, y se utiliza para entender cómo se entretejen datos y procesos bajo este concepto. En otras palabras, hablamos de una arquitectura unificada con servicios (o tecnologías) corriendo sobre ella y que sirven de ayuda para las empresas en la tarea de la gestión de datos. Maximizar el valor de los datos y acelerar la transformación digital son dos de sus principales beneficios, mientras que la simplificación e integración de la gestión de datos en entornos Cloud (y también en las instalaciones) son otras ventajas claras.

¿Qué es Data Fabric?

El Data Fabric está diseñado para ayudar a las organizaciones a resolver problemas de datos complejos y casos de uso mediante la administración de sus datos, independientemente de los diversos tipos de aplicaciones, plataformas y ubicaciones donde se almacenan estos datos. El Data Fabric permite el acceso sin fricciones y el intercambio de datos en un entorno de datos distribuidos.

Es necesario tener la capacidad de asumir el constante «bombardeo» de datos, manipularlos y analizarlos para convertirlos en información útil, capaz de mejorar el proceso de toma de decisiones a todos los niveles. Los sistemas basados en inteligencia artificial nos aportan la potencia de cálculo y la velocidad necesaria para salir airosos y aprovechar al máximo los datos.

Gracias a Data Fabric podemos procesar, administrar y almacenar los datos a medida que se mueven por la arquitectura. Además, los datos son accesibles por aplicaciones externas o internas. Por ejemplo, podemos conectarnos a cualquier fuente de datos a través de conectores y componentes previamente empaquetados, eliminando así la necesidad de codificación.
También, y es una de sus principales aplicaciones, soportar casos de uso por lotes, en tiempo real y de big data, o gestionar múltiples entornos cloud (híbrido, multicloud...), ya sea como fuentes o como consumidores de datos.

Este concepto pretende acelerar la inferencia de conocimiento a partir de los datos en bruto gracias a la automatización de procesos. Puede, incluso, en casos ideales, proporcionar conocimiento en tiempo real, gestionando a la vez el flujo de datos y la mejora de todas las fuentes de datos.

Data Fabric aglutina procesos como la integración de datos, el análisis o el dashboarding, todo en uno. Además sirve como solución de gestión, y permite acceso instantáneo a los datos para cualquier miembro de una organización, de nuevo, en tiempo real, permitiendo un acceso sin fricciones en un entorno distribuido.

2

¿Por qué usar un Data Fabric?

¿Por qué usar un Data Fabric?

Cualquier organización centrada en datos necesita un enfoque holístico que supere los obstáculos de tiempo, espacio, diferentes tipos de software y ubicaciones de datos. Los datos deben ser accesibles para los usuarios que los necesitan, no encerrados detrás de firewalls o ubicados poco a poco en una variedad de ubicaciones. Las empresas necesitan tener un entorno seguro, eficiente y unificado, y una solución de datos preparada para el futuro para poder prosperar. Un Data Fabric ofrece esto.

¿Por qué usar un Data Fabric?

La integración de datos tradicional ya no satisface las nuevas demandas comerciales de conectividad en tiempo real, autoservicio, automatización y transformaciones universales. Aunque la recopilación de datos de varias fuentes no suele ser el problema, muchas organizaciones no pueden integrar, procesar, seleccionar y transformar datos con otras fuentes. Esta parte crucial del proceso de administración de datos deberá darse para brindar una visión integral de los clientes, socios y productos, lo cual ofrecerá a las organizaciones una ventaja competitiva. De esta manera, podrán satisfacer mejor las demandas de los clientes, modernizar sus sistemas y aprovechar el poder de la informática en la nube.

El Data Fabric se puede visualizar como una tela, distribuida por todo el mundo, dondequiera que estén los usuarios de la organización. El usuario puede estar en cualquier lugar de este entramado de datos y seguir accediendo a los datos en cualquier otro lugar sin restricciones, en tiempo real.

Data Fabric es más que una simple red

Internet se creó para conectar a las personas de todo el mundo, brindándoles la capacidad de ignorar los obstáculos del tiempo y la distancia. Sin embargo, inicialmente solo conectaba personas y la transferencia de datos cuantificados era mínima. Hoy, la actividad en las plataformas digitales ha superado las previsiones iniciales y los datos se han convertido en un mundo en sí mismos. Cualquier actividad que sea cuantitativa, ya sea en línea o en la vida real, puede clasificarse como suministro de datos. Si bien estos datos crecen a pasos agigantados, es necesario establecer una infraestructura para administrarlos.

Anteriormente, el objetivo era administrar datos y, como beneficio adicional, extraer información de ellos. Con el paso del tiempo, el enfoque comenzó a pasar de solo administrar datos a poder extraer información de esos datos. Con un Data Fabric, el enfoque está cambiando de la simple administración de datos a la mejora de la calidad de los datos en sí, la disponibilidad de la información y los conocimientos automatizados derivados de ella.

En todo el mundo, el número de partes interesadas que ingresan al entorno en red está aumentando. Todo el mundo está conectado a internet y cada plataforma se ha convertido en una fuente de datos. Maximizar el valor de los datos se ha convertido en un problema complejo. Los desafíos de los datos actuales incluyen:

- Ubicación en múltiples ubicaciones locales y en la nube.
- Datos estructurados y no estructurados.
- Diferentes tipos de datos.
- Diferentes entornos de plataformas.
- Mantenido en diferentes sistemas de archivos, bases de datos y aplicaciones SaaS.

Los datos están creciendo exponencialmente, por lo que estos problemas se están multiplicando. Estos problemas y variaciones hacen que sea complejo acceder o usar fácilmente los datos. Y, si las organizaciones desean producir u operar AI y ML, necesitan que sus datos se recopilen, transformen y procesen.

En la actualidad, la mayoría de las organizaciones tienden a lidiar con el problema en silos, creando muchas formas diferentes de administrar los datos en una organización. Aunque esta solución hace que los datos estén disponibles para grupos particulares, el acceso a ellos en toda la empresa se vuelve casi imposible, frecuentemente relegando los datos a permanecer inactivos y sin usar.

La falta de acceso y uso de datos completos da como resultado un bajo retorno de la inversión en la infraestructura, la falta de disponibilidad de datos para producir predicciones útiles y una menor productividad. Bajo tales condiciones que el Data Fabric viene al rescate.

Data Fabric vs Status quo

Actualmente, muchas organizaciones utilizan lagos de datos y almacenes de datos para administrar datos. Sin embargo, en una inspección más cercana, estos enfoques son intensivos en tecnología en lugar de centrados en datos. Con los lagos de datos y los almacenes de datos, el énfasis está en recopilar o extraer los datos sin procesar, almacenarlos y usarlos cuando se obtienen conocimientos.

Estas soluciones no se diseñaron teniendo en cuenta los problemas actuales y hacen que sea difícil obtener una vista unificada de los datos. Sin embargo, estas técnicas frecuentemente conducen a retrasos y costos crecientes. Con la creciente cantidad de datos y las limitaciones de tiempo con las que trabajan los responsables de la toma de decisiones de una organización, no son deseables los retrasos en el acceso y el procesamiento de los datos. En tales escenarios, el Data Fabric ofrece la ventaja de almacenar, extraer y procesar datos en el punto de origen en tiempo real, lo que permite a los responsables de la toma de decisiones obtener información sobre la marcha.

Data Fabric vs. Virtualización de datos

El Data Fabric frecuentemente se confunde con la virtualización de datos. La virtualización de datos crea una capa de abstracción de datos y, a menudo, se confía en ella cuando necesita integrar datos rápidamente. Conecta, recopila y transforma datos de muchas fuentes diferentes, ya sea en las instalaciones o en la nube, para obtener información ágil, de autoservicio y en tiempo real. Por otro lado, el Data Fabric se refiere a una arquitectura de administración de datos integral y global que se utiliza para casos de uso más amplios, como el análisis de preferencias de clientes y análisis de IoT, incluido un conjunto más grande de componentes de pila. Los analistas recomiendan usar la virtualización de datos como una herramienta que contribuye a su arquitectura de Data Fabric. A medida que utiliza más y más herramientas de integración de datos, puede hacer crecer su solución en un Data Fabric que sea específica para los objetivos de su organización.

Implementación de Data Fabric

El Data Fabric comienza con conceptos de procesamiento de transacciones en línea (OLTP), el cual se inserta, actualiza y carga en una base de datos información detallada sobre cada transacción. Los datos se estructuran, limpian y almacenan en silos en un centro para su uso posterior. Cualquier usuario de los datos, en cualquier punto de la estructura, puede tomar los datos sin procesar y usarlos para derivar múltiples hallazgos, lo que ayudará a las organizaciones a aprovechar sus datos para crecer, adaptarse y mejorar.

Implementación de Data Fabric

Una implementación exitosa del Data Fabric requiere:

· Aplicación y servicios: donde se construye la infraestructura necesaria para la adquisición de datos. Esto incluye el desarrollo de aplicaciones e interfaces gráficas de usuario (GUIs) para que el cliente interactúe con la organización.

· Desarrollo e integración de ecosistemas: creación del ecosistema necesario para recopilar, gestionar y almacenar los datos. Los datos del cliente deben transferirse al administrador de datos y a los sistemas de almacenamiento de manera que se evite la pérdida de datos.

· Seguridad: los datos recopilados de todas las fuentes deben administrarse con la seguridad adecuada.

· Gestión de almacenamiento: los datos se almacenan de manera accesible y eficiente, con posibilidad de escalar cuando sea necesario.

· Transporte: construcción de la infraestructura necesaria para acceder a los datos desde cualquier punto de las ubicaciones geográficas de la organización.

· Terminales: desarrollo de la infraestructura definida por software en los puntos de almacenamiento y acceso para permitir información en tiempo real.

3

¿Cómo funciona la inteligencia artificial o el Machine Learning con Data Fabric?

En las fases iniciales del almacenamiento de datos, los ingenieros de datos y los científicos de datos intentaron conectar los puntos en los datos para encontrar patrones. Descubrieron que con las técnicas tradicionales de integración de datos, pasaban la mayor parte del tiempo en la logística de datos en lugar de aprender sobre los datos. No es un enfoque sostenible si deseamos obtener información más rápidamente.

¿Cómo funciona la inteligencia artificial o el Machine Learning con Data Fabric?

Un Data Fabric es esencialmente una capa operativa de datos que no solo reúne todos los datos, sino que los transforma y procesa mediante Machine Learning para descubrir patrones y conocimientos. Sin un Data Fabric, todo esto tiene que suceder en cada aplicación individual, lo cual no es una solución sostenible.

Un Data Fabric puede preparar datos para satisfacer las necesidades de AI y ML automáticamente y en niveles sostenibles. El Machine Learning puede proporcionar los datos y los conocimientos de forma proactiva, lo cual ayuda a los responsables de la toma de decisiones a tener mejores conocimientos e información más oportuna. Los resultados deseables radican en descubrir hechos ocultos de los datos sin que se busquen o soliciten específicamente, mientras se encuentran soluciones para problemas o conocimientos comerciales.

4

Riesgos con Data Fabric y Beneficios de Data Fabric

Una preocupación creciente para las organizaciones es la amenaza a la seguridad de los datos cuando se transportan de un punto a otro en el Data Fabric. Es obligatorio que la infraestructura para el transporte de datos incorpore firewalls y protocolos de seguridad para garantizar la seguridad frente a violaciones de seguridad. Con un número cada vez mayor de ataques cibernéticos que afectan a las organizaciones, la seguridad de los datos en todos los puntos del ciclo de datos es primordial.

Beneficios de Data Fabric

El Data Fabric es ideal para organizaciones con base en diferentes partes del mundo, tienen múltiples fuentes de datos y enfrentan problemas de datos complejos o casos de uso.

Con los continuos avances en las capacidades del hardware, la globalización se está expandiendo a regiones que antes no estaban conectadas. Con las velocidades de conectividad aceleradas, las organizaciones pueden verse abrumadas por los datos de los dispositivos y servicios. Si bien los datos se han utilizado durante bastante tiempo para obtener información, el Data Fabric proporciona una solución que ofrece las siguientes ventajas:

· Cuenta con un modelo ágil que permite cambios en los sistemas, se adapta y ajusta según sea necesario y funciona en todos los sistemas operativos y de almacenamiento.
· Es escalable con mínima interferencia, sin inversión en hardware enormemente costoso o personal altamente capacitado y costoso.
· Proporciona la máxima integridad y cumple con las regulaciones, manteniendo la accesibilidad y el flujo de información en tiempo real.

Las cantidades masivas de datos a las que las empresas pueden acceder deben explotarse para obtener información única. Las áreas que incluyen pronósticos, ventas y optimización de la cadena de suministro, marketing y comportamiento del consumidor brindan a la organización una ventaja competitiva y liderazgo en datos en su campo. La derivación de información en tiempo real puede hacer que la organización se destaque del resto.

5

Implementación Cloudera para Data Fabric

Cloudera Data Platform (CDP) se ha creado desde cero para admitir entornos híbridos y multinube, gestión de datos en apoyo de una arquitectura Data Fabric. En esta sección proporcionamos una introducción a CDP, con un enfoque en las capacidades de gestión de datos que permiten el Data Fabric.

Cloudera Data Platform (CDP) es una plataforma de datos híbrida diseñada para brindar la libertad para elegir cualquier nube, cualquier análisis, cualquier dato. CDP ofrece una gestión de datos más rápida y sencilla y análisis para datos en cualquier lugar, con un rendimiento, escalabilidad y seguridad óptimos.
Con CDP obtiene el valor de CDP Private Cloud y CDP Public Cloud para rapidamente valorar el mayor control de TI.

- **Cloudera Data Platform** 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 recida el dato. Como resultado, la plataforma está en condiciones para implementar arquitecturas de datos modernas:

- Un Data Fabric unificado que organiza centralmente fuentes de datos dispares de manera inteligente y de forma segura a través de múltiples nubes y también on premise.

- Un Data Lakehouse abierto que permite análisis multifunción tanto en transmisión como almacenado datos en un almacén de objetos nativos de la nube a través de múltiples nubes híbridas.

- Una Data Mesh escalable que ayuda a eliminar los silos de datos mediante la distribución de la propiedad a los equipos multifuncionales mientras se mantiene una infraestructura de datos comú

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.

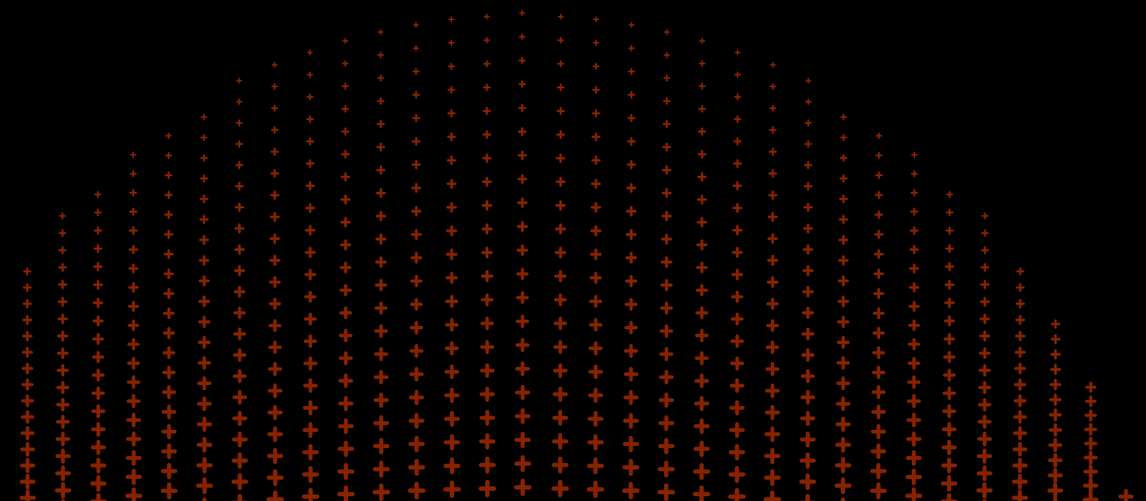
Este catálogo global proporciona un inventario de búsqueda de todos los activos que forman parte de Data Fabric, lo que hace que los activos de datos sean fácilmente accesibles.

- Integral: soporte para todas las entidades que conforman el ecosistema de nube híbrida: Hive tablas, tópicos de Kafka, flujos de Nifi, tablas de HBase, modelos de aprendizaje automático, etc. Cada activo podrá mostrarse junto con sus metadatos contextuales, como esquema, políticas de seguridad, etiquetas y clasificaciones, perfil, reglas de gobierno y anotaciones comerciales.
- Visibilidad: ubicación única para descubrir y buscar datos de todos los nodos de Fabric.
- Gobernanza: creación de perfiles integrada para brindar información sobre la calidad y la sensibilidad de los datos, motor de clasificación que asigna atributos relacionados con la seguridad, el cumplimiento y las políticas, como PII.
- Linaje: la captura automática de información de linaje ayuda a comprender de dónde provienen los datos, de cómo se está utilizando, qué impacto tendrían los cambios. Se puede extender más para propagar políticas de seguridad en todo el Data Fabric, lo que hace que sea más seguro y fácil compartir datos.
- Política: las políticas de seguridad, cumplimiento y gobernanza se pueden asignar a cualquier recurso de datos directamente desde el Catálogo.
- Seguridad: registro de auditoría completo de todos los accesos y modificaciones realizadas en los conjuntos de datos y en cualquier parte del Data Fabric.
- Colaboración: admite anotaciones comerciales y metadatos, conservación y colaboración

Esto aborda los requisitos de la capa de gestión de datos de Data Fabric, cuando se implementa junto con Shared Data Experience (SDX

Shared Data Experience (SDX)

Cloudera SDX combina capacidades de gestión, gobernanza y seguridad de nivel empresarial con metadatos compartidos que se implementan localmente en cada nodo del Data Fabric y se federan a través del Plano de Control. Proporciona una capa de gobernanza que es verdaderamente global: control de expansión de planos e instancias de implementación para asignar propiedad, capturar para auditoría y aplicar políticas globales a través de implementaciones locales y nubes públicas.



- Metadatos: establece activos de información para aumentar la usabilidad, la confianza y el valor aprovechando todos los metadatos (estructurales, operativos, empresariales y sociales).
- Seguridad: políticas de seguridad granulares, dinámicas, basadas en roles y atributos. Prevenir y auditar el acceso no autorizado a datos confidenciales o restringidos a través de la plataforma.
- Cifrado: criptografía sólida para datos en movimiento y en reposo, autenticación centralizada con inicio de sesión único.
- Control: mueva los datos y las cargas de trabajo entre implementaciones para obtener un rendimiento y un costo óptimos y resiliencia, satisfaciendo las necesidades comerciales en constante cambio.
- Gobernanza: capacidades de auditoría, linaje y gobernanza de nivel empresarial aplicadas en la plataforma con amplia extensibilidad para integraciones de socios.

Replication Manager

El administrador de replicación está diseñado para atender una serie de casos de uso en torno a datos entre estructuras orquestación y replicación: migración de cargas de trabajo, sobre cargas de la nube, copias de seguridad, recuperación por desastres y replicación en apoyo de sistemas de desarrollo y prueba. Soporta replicación full e incremental para todos los tipos de almacenamiento de datos disponibles en la estructura.

Un principio clave de Unified Data Fabric es tener controles de gobierno y seguridad consistentes en todos los puntos finales de la estructura. Estrechamente integrado con SDX, el administrador de replicas admite que funcione moviendo políticas con los datos, replicando todos los metadatos asociados, clasificación de etiquetas, políticas de seguridad, reglas de cumplimiento e información de linaj

Seguridad global unificada con SDX

Seguridad global unificada con SDX

SDX admite políticas basadas en atributos mediante el uso de etiquetas, como "PII", que se pueden ser asignados a cualquier activo de datos, incluidas las columnas individuales de una tabla. La política de acceso a datos para los datos PII pueden ser especificados por un equipo centralizado responsable de las reglas de toda la empresa, mientras que la etiqueta en sí puede ser asignada por el creador del conjunto de datos, ya sea manualmente o de forma automática según clasificación. La captura automática de información de linaje a través de la canalización de datos permite tener herencia de etiquetas y, como tal, propagación de las políticas relevantes dentro de un nodo local en el tejido de forma autónoma.

El administrador de replicación es consciente de estas etiquetas. A medida que los datos se mueven entre entornos, las etiquetas de clasificación también se propagan y asignan automáticamente a los datos, esto hace cumplir las políticas apropiadas globalmente a través de los nodos de la estructura, y proporciona políticas unificadas de gestión y cumplimiento en todos los entornos de la organización, al tiempo que permite el acceso de autoservicio de los usuarios comerciales a datos confiables.

Data Services

Las capas 2 a 6 del Data Fabric se abordan mediante Cloudera Data Services

- Ingestión y transmisión de datos: proporcionados por Cloudera Data Flow (CDF) y Ingeniería de datos de Cloudera (CDE)
- Procesamiento y persistencia de datos: proporcionado por Cloudera Data Hub (CDH), Cloudera Data Warehouse (CDW) y Cloudera Operational Database (COD)
- Orquestación de datos: proporcionada por componentes integrados en Cloudera Data Engineering (CDE) y Cloudera Streaming Analytics (CSA)
- Descubrimiento de datos: proporcionado por Cloudera Data Visualization (CDV) y Cloudera Data Catalog
- Acceso a datos globales: proporcionado por Cloudera Data Warehouse (CDW), Cloudera Operational Database (COD) y Cloudera Machine Learning (CML)

Agnostic

www.agnosticit.com

Impulsando la
transformación digital
de las empresas