Informatica™

Conversión de un data lake en un mercado de datos

Contenidos

Consejo: haga clic aquí para ir directamente a cualquier sección.

Introducción

Introducción

No hay duda de que los data lakes representan una gran oportunidad para ofrecer nuevas perspectivas a partir de enormes cantidades de datos procedentes de fuentes antiguas y nuevas.

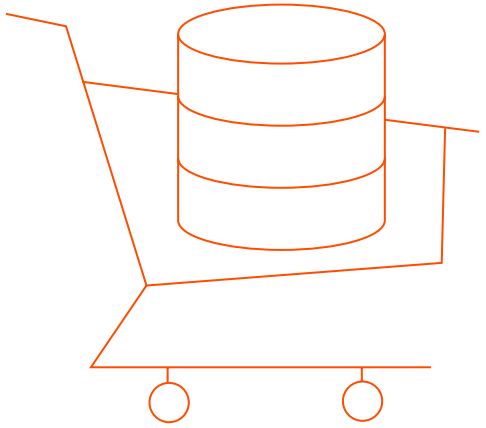
Sin embargo, la creación, el mantenimiento y el uso eficaz de los entornos de data lake presentan dificultades para las organizaciones. Como resultado, no pueden aprovechar la información basada en datos y, posiblemente, pierdan nuevas oportunidades. Mientras tanto, los data lakes corren el riesgo de convertirse en espacios pasivos para almacenar datos en lugar de ser espacios activos para distribuir datos a los consumidores de datos.

Un nuevo conjunto de capacidades tecnológicas y prácticas organizativas está empezando a formar la base para convertir los data lakes en mercados de datos. Para ello, son esenciales los principios del diseño orientado a la agilidad, la creación de una cadena de suministro de datos y la planificación de un éxito más rápido y colaborativo.

¿Qué es un mercado de datos?

Un mercado de datos es un nuevo tipo de arquitectura de gestión de la información que amplía el concepto tradicional de "data lake" para combinar un proceso estandarizado e industrializado que permite conservar activos de datos sin procesar y convertirlos en información de confianza, con un modo de interacción con los usuarios finales basados en la colaboración y el autoservicio, de modo que los consumidores de datos pueden adquirir rápida y fácilmente los datos que necesitan.

Este cuaderno tiene la finalidad de compartir los consejos y las mejores prácticas necesarios para maximizar el valor de los entornos de data lake de su organización y aprovechar el potencial de la disrupción inteligente basada en datos a través mediante un mercado de datos.



Primera parte

Diseño orientado a la agilidad

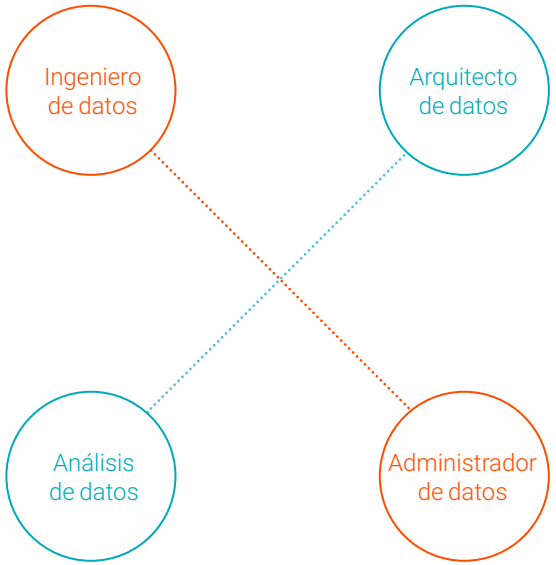
Obstáculos para la implementación eficaz del mercado de datos

Los entornos de mercado de datos prometen agilizar la detección de nueva información. Sin embargo, las organizaciones que emprenden estas iniciativas suelen verse incapaces de extraer el máximo valor de ellas por distintos motivos:

- **Los procesos de gestión de datos anticuados a menudo dificultan la velocidad, la flexibilidad y la colaboración.** Los procesos de recopilación de requisitos complejos y los ciclos de desarrollo prolongados causan retrasos en la obtención de la información que necesitan las líneas de negocio para demostrar valor y generar el impulso necesario.
- **Exceso de controles de TI.** Un exceso de controles puede ralentizar los proyectos, ya que la participación del personal de TI en las operaciones suele ser innecesaria.
- **Falta de herramientas eficaces para la colaboración.** Sin estas, los equipos no pueden obtener los beneficios del trabajo que otros equipos ya han creado.

Una vez que se salvan estos obstáculos, la importancia de los equipos transversales, formados por ingenieros y arquitectos de datos de TI, así como de los usuarios de líneas de negocios de los equipos de análisis y administración que trabajan para lograr un programa de negocio común es primordial. Estos miembros del equipo están facultados para representar las necesidades de su función a medida que el grupo ejecuta colectivamente el alcance de un proyecto desde el inicio hasta el final.

El principal beneficio de crear equipos transversales es la capacidad de integrar conocimientos exhaustivos funcionales provenientes de múltiples fuentes. Los proyectos de data lake requieren los conocimientos en implementación de la ingeniería de datos y el contexto de negocio de los administradores de datos, así como conocimientos analíticos de los científicos y analistas de datos. Tener múltiples perspectivas fomenta el desarrollo oportuno de información de negocio precisa y coherente, y garantiza que todo el mundo se ajuste a una comprensión común de los datos disponibles.



Preguntas que puede formularse:

¿Son los procesos de gestión de datos tan eficientes como deberían ser? Si no, ¿cómo se pueden optimizar para garantizar que no afecten al plazo de amortización?

¿Qué grado de accesibilidad y oportunidad tienen los datos? ¿Son los controles que está utilizando demasiado estrictos?

¿Se ha creado un equipo transversal con los usuarios pertinentes de todo el negocio, que pueda ofrecer diversas perspectivas y garantizar la obtención de los resultados adecuados del proyecto de mercado de datos?

¿Se han definido roles y responsabilidades, y se han proporcionado herramientas de apoyo para asegurar una colaboración eficaz?

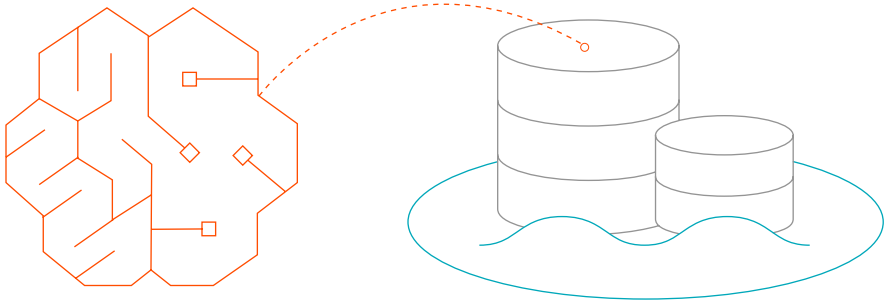
Permita a los especialistas en datos acceder a los datos que necesitan para ayudar en la preparación de datos

Las herramientas de visualización de datos de autoservicio, como Tableau, Qlik y Zoomdata, se han vuelto muy populares con el paso de los años para proporcionar a los analistas de negocio acceso directo a los datos. Las iniciativas de autoservicio son uno de los principios fundamentales de la creación de un mercado de datos que saca a la luz los datos de un almacén para ponerlos a disposición de los consumidores de una organización. Habilitar a los usuarios de las líneas de negocios para que adquieran directamente los datos del mercado que se ajustan a su finalidad les da la posibilidad de participar en el proceso de preparación de datos en activos de confianza.

Pero, con frecuencia, el problema de estas iniciativas de autoservicio es que los usuarios de negocio deben esperar a que los equipos de TI les proporcionen los datos que necesitan, o bien tienen que adoptar procesos manuales para conservar y limpiar los datos para obtenerlos de la forma que requieren (muchas veces en hojas de cálculo).

Aquí es donde la preparación de datos de autoservicio entra en escena. Esta permite que los usuarios analíticos de negocio con conocimientos fusionen, transformen y limpien los datos relevantes en formularios certificados y más fiables antes del análisis.

Herramientas sofisticadas permiten a los usuarios publicar sus conjuntos de datos preparados en espacios de trabajo colaborativos, de modo que varios usuarios de negocio pueden acceder juntos los datos. Además, las técnicas de inteligencia artificial y aprendizaje automatizado dentro de las herramientas pueden proporcionar una experiencia automatizada y guiada a los analistas de negocio a medida que exploran y detectan los datos del data lake.



Uso de la aportación colectiva y el etiquetado para gobernar activos de datos

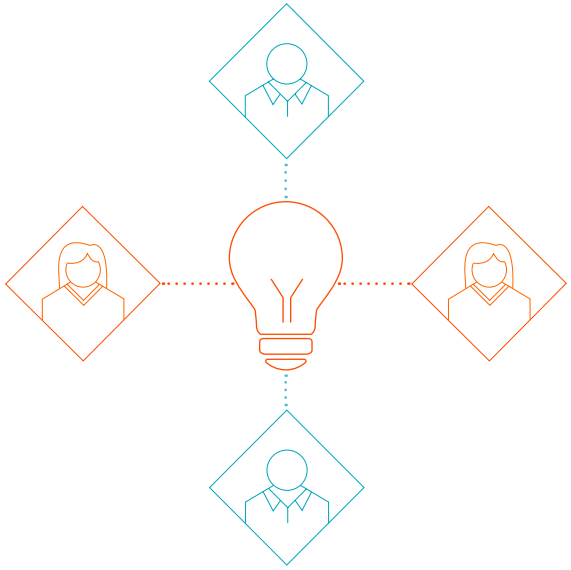
Con frecuencia, se cree que los data lakes pueden estar sin gobierno. Este es un peligroso mito: la adopción de data lakes por parte de las organizaciones para el procesamiento de los datos sensibles, como datos de pacientes o consumidores, necesita métodos eficaces de gobierno de datos. Sin embargo, los métodos de gobierno lentos y centralizados pueden invalidar los beneficios de agilidad que promete un data lake.

Los mercados minoristas online han aprovechado la sabiduría popular para que los consumidores puedan compartir comentarios y opiniones, de forma que los futuros consumidores puedan beneficiarse de una experiencia anterior. Como tal, este tipo de filtrado de colaboración y aportación colectiva de sabiduría es otro principio fundamental de un mercado de datos.

El gobierno de datos está destinado a ser una función de valor añadido que aumenta la calidad de los datos y garantiza el cumplimiento de los estándares y la protección de los datos sensibles. Como tales, los analistas de negocio y otros consumidores de datos tienen tanto interés en el gobierno de datos como sus compañeros de administración. Ahí es donde el concepto de gobierno de datos con aportación colectiva entra en escena.

La aportación colectiva es la capacidad de aprovechar la sabiduría de los analistas de negocio para obtener el conocimiento y la experiencia que mejoran en conjunto la calidad de los datos. En un entorno de autoservicio, todos los usuarios tienen la capacidad de aplicar sus conocimientos en la materia para mejorar la calidad y el contexto de los datos.

Los analistas de negocio deben aportar sus conocimientos, a través de etiquetas y otras clasificaciones, de modo que la calidad de los activos de datos aumente. La colaboración se convierte entonces en un mecanismo para garantizar la flexibilidad, ya que los analistas de negocio se ayudan unos a otros a mejorar la calidad de los activos de datos.



Preguntas que puede formularse:

¿Se ha facilitado en la medida de lo posible a los usuarios de la empresa la preparación de datos del data lake sin necesidad de involucrar al equipo de TI, o la preparación de datos sigue siendo un engorroso proceso manual de conservación y limpieza?

¿Se han implementado políticas de gobierno de datos suficientes para asegurar la protección de los datos sensibles y garantizar que la calidad de los datos es la adecuada para su uso en las decisiones clave?

¿Comprometen los procesos manuales de aplicación de gobierno la agilidad del data lake?

¿Tienen los analistas de negocio capacidad para aportar eficazmente sus conocimientos de contexto de datos al data lake mediante el uso de etiquetas u otras clasificaciones?

Segunda parte

Creación de un motor de cadena de suministro de datos

Problemas de los procesos manuales y especializados

La detección rápida de información de negocio nueva es un beneficio clave de los entornos de data lake, fundamentales para los mercados de datos. En un entorno competitivo en el que las líneas de negocio necesitan velocidad, los procesos que no estén automatizados de forma sistemática retrasarán la producción de información nueva. Sin un proceso de producción de alta velocidad, los activos de datos nunca llegarán a tiempo a las líneas de negocio. Como tales, los procesos rápidos son otro principio fundamental de los mercados de datos.

Es más, los métodos de codificación manual anticuados pueden cohibir el mantenimiento a largo plazo de la lógica de negocio. Las soluciones de codificación manual integradas en lenguajes de bajo nivel suponen un riesgo: si el lenguaje deja de ser compatible o los desarrolladores con los conocimientos necesarios dejan de trabajar en la empresa, tales soluciones deben reescribirse.

Incluso las soluciones generadoras de código, que automatizan la codificación manual, plantean un riesgo enorme en torno a la compatibilidad y el mantenimiento. Una vez que los artefactos de la lógica de negocio se almacenan en paradigmas de desarrollo especializados, el negocio depende de tales paradigmas en todo momento.

Más allá de la falta de mantenimiento, las soluciones de codificación manual suponen un riesgo en cuanto a auditabilidad y gobierno. La mayoría de las organizaciones se rigen por preceptos internos y externos para realizar el seguimiento de las modificaciones y el uso de datos. Sin una vista lógica, las operaciones ejecutadas por las soluciones de codificación manual son difíciles de controlar y supervisar con fines de auditoría.

Por último, hay una cuestión de orden práctico a la hora de basarse en procesos manuales y especializados para gestionar el volumen de datos que se requiere actualmente. Dado que las organizaciones se enfrentan a un gran crecimiento de la cantidad de datos que deben procesar, resulta inviable esperar un crecimiento similar de los recursos para gestionar estos datos como respuesta: las organizaciones necesitan encontrar una solución automatizada que se adapte a esta explosión continua de datos.

Preguntas que puede formularse:

¿Qué cantidad de lógica de negocio se crea y gestiona mediante soluciones de codificación manual que requieren conocimientos especializados de codificación? Si los desarrolladores especializados dejan la compañía, o el lenguaje cae en desuso, ¿cuáles serán el riesgo y el coste de su iniciativa?

Si utiliza soluciones generadoras de código, ¿son sostenibles a largo plazo? ¿Ofrecen transparencia de metadatos, a la vez que permiten la automatización y la agilidad empresarial con las necesidades cambiantes?

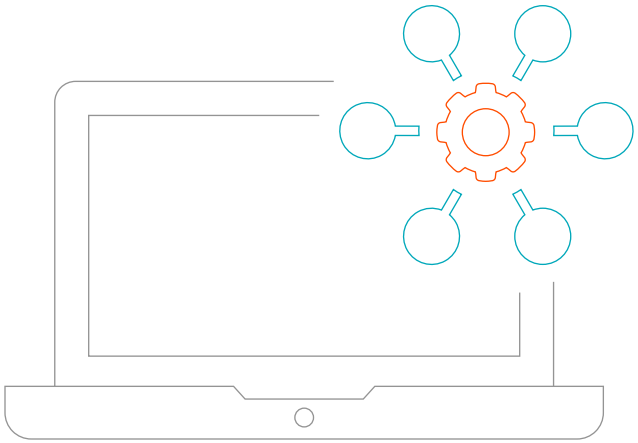
¿Cuenta con medios para asumir un aumento del 30 al 50 % del volumen de datos durante los próximos cinco años?

Automatización de la incorporación y la transformación de los datos

El aspecto más táctico de cualquier entorno de data lake es la automatización de la incorporación y la transformación de datos. La incorporación y la transformación de los datos de forma manual es un complejo proceso de varios pasos que conduce a resultados incoherentes e irrepetibles.

Las organizaciones de éxito aprovechan los conectores preintegrados y las plataformas de incorporación de datos de alta velocidad para cargar y transformar los conjuntos de datos en el lake, lo que permite a los data lakes escalar a un volumen cada vez mayor de datos entrantes.

La automatización también permite la iteración y la flexibilidad rápidas, así como la agilidad requeridas para satisfacer las cambiantes necesidades del negocio, ya que los procesos automatizados se pueden modificar rápidamente, sin correr el riesgo de interrumpirlos y que ello afecte a los usuarios existentes.



Aprovechar la validación y la puntuación de datos basadas en reglas para detectar problemas en la calidad de los datos en fases tempranas

Como los ejecutivos saben, los problemas que no se detectan de manera temprana causan mayores problemas posteriormente. Con los data lakes, los errores relacionados con la calidad de los datos no identificados de forma temprana pueden afectar considerablemente a la información de negocio debido a las imprecisiones o incoherencias entre los diferentes activos de datos. Con el volumen de datos que las empresas deben gestionar y analizar, es casi imposible detectar los problemas de calidad de datos de forma manual.

Las técnicas de inteligencia artificial que recomiendan e infieren las reglas de negocio son la respuesta: un método de automatización de procesos de calidad de datos. Los data lakes con validación de datos basada en reglas pueden detectar automáticamente los signos de datos incompletos o incoherentes. La detección temprana de estas anomalías puede afectar considerablemente a la fiabilidad de la información de negocio.

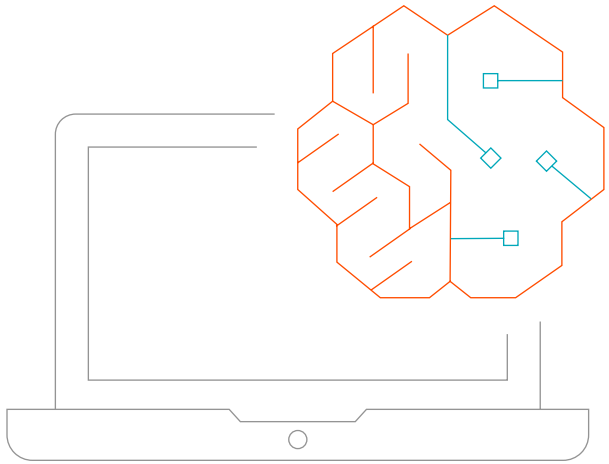
Se debe utilizar un sistema de reglas para perfilar y filtrar los datos a medida que se incorporan y se transforman en el data lake. Cuando las reglas automatizadas identifican datos que se encuentran fuera de los límites, estas instancias se pueden catalogar y escalar para su seguimiento por parte de los analistas y administradores de datos. Este tipo de validación y puntuación de datos basadas en reglas centra el tiempo limitado de los miembros del equipo al poner de relieve las áreas donde los datos pueden tener más problemas. Por lo tanto, los scorecards y los cuadros de mando de calidad de datos impulsan la visibilidad y la comprensión donde debe centrarse el esfuerzo manual.

Uso del aprendizaje automatizado para la detección y administración de datos

Con unos volúmenes de datos que crecen rápidamente, uno de los mayores retos para las organizaciones es obtener visibilidad de los activos de datos disponibles. Si bien el acto de crear un data lake ayuda a centralizar los activos de datos clave en un solo entorno, aún se plantea la cuestión de decidir qué activos deben incorporarse en el data lake en primer lugar.

Al igual que los motores de búsqueda web rastrean e indexan la web, se deben utilizar escáneres de datos automatizados para buscar e indexar de forma proactiva nuevos activos de datos por toda la empresa. Se deben utilizar técnicas de aprendizaje automatizado para identificar las correlaciones y similitudes entre los distintos activos de datos y crear una visión global de los activos de datos para la administración de estos.

Por otra parte, esta visión global de los activos de datos se debe utilizar para formar un catálogo inteligente de todos los activos de datos y las relaciones entre ellos deducidas. Los consumidores de datos, como analistas de negocio, pueden utilizar el catálogo para identificar nuevos activos que pueden ser de interés para ellos.



Preguntas que puede formularse:

¿Se ha automatizado la incorporación de datos en el data lake?

¿Se han implementado reglas de negocio y un proceso de administración para identificar y mitigar los problemas de calidad de datos?

¿Se está aprovechando todo el potencial del aprendizaje automatizado para la detección y administración de datos?

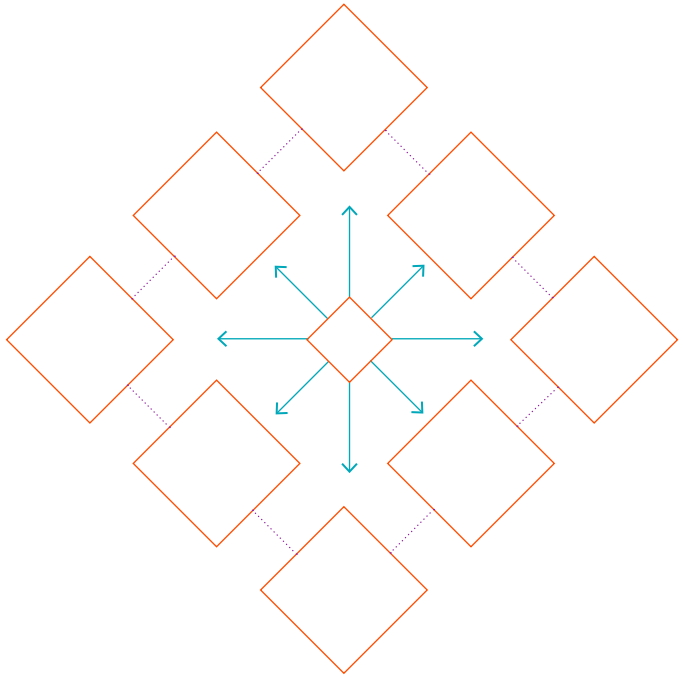
Tercera parte

Organización para un éxito rápido y colaborativo

Problemas de los equipos aislados y descentralizados

A menudo, las organizaciones se enfrentan al reto de trabajar con el equipo de TI y los usuarios de líneas de negocio más allá de los límites geográficos y de la organización. Estos silos en las organizaciones pueden afectar a las ventajas de los entornos de data lake: uno de los objetivos tácticos de un data lake consiste en crear una visión de la realidad única y coherente en torno a los activos de datos para varios consumidores. Con un almacenamiento eficiente, ya no hay necesidad de utilizar data marts departamentales. Un inventario único de los activos de datos es otro principio fundamental de un mercado de datos.

Sin embargo, el legado de estos silos departamentales, combinado con una tendencia general hacia el acaparamiento de datos funcionales, puede limitar las ventajas de los data lakes. Las soluciones de gestión de data lakes pueden ayudar a facilitar la colaboración y a convertir la sabiduría popular en un activo en lugar de ser una responsabilidad.

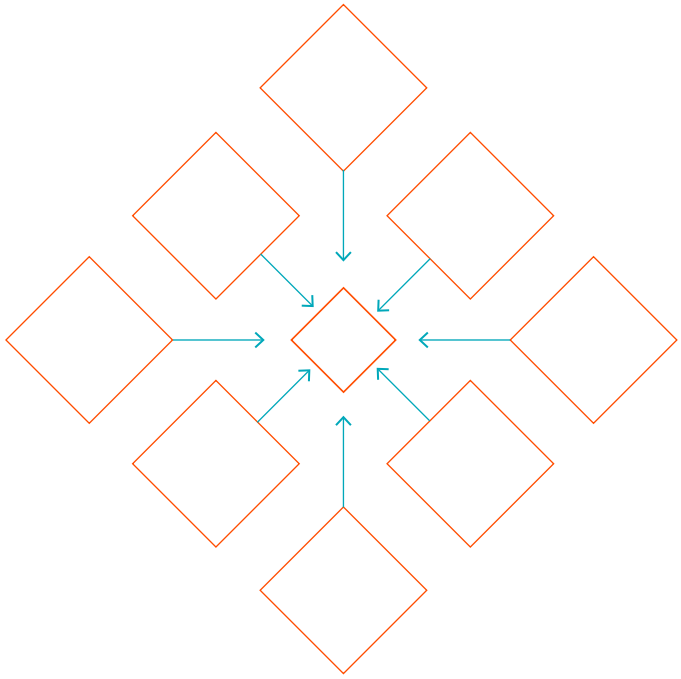


Diseño para la centralización y colaboración

El legado de los modelos anticuados de gestión de datos sigue afectando a muchas organizaciones. Estos modelos anticuados, lentos y manuales, forzaron a las organizaciones a recopilar datos en las líneas de negocio. Posteriormente, los equipos departamentales actuales han empezado a crear data lakes aislados incoherentes y que duplican otros entornos de la organización.

El principio de la coubicación es fundamental para sacar el máximo provecho a los beneficios de un data lake. Debe analizar un número limitado de entornos de data lake organizados de manera exhaustiva alrededor de dominios empresariales críticos. De este modo, se garantiza que los data lakes reflejen visiones únicas de la realidad en toda la organización y minimicen la duplicación innecesaria, ya que solo aumenta la complejidad y el riesgo con respecto al gobierno.

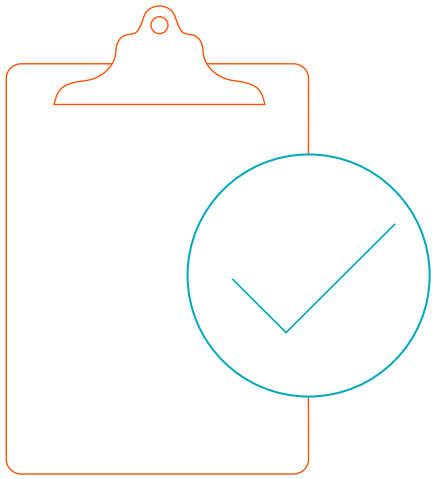
Además, los enfoques de gestión de data lakes que explotan tanto el uso compartido y el etiquetado de los datos como los espacios de trabajo para proyectos facilitan la colaboración necesaria para estos entornos. Los consumidores de datos deberían verse unos a otros como cohortes en transiciones analíticas en las que el trabajo de un analista en el data lake puede publicarse y compartirse con otros analistas para su reutilización.



Normalización del proceso de gestión de datos y fomento de la coherencia en la arquitectura

Las organizaciones a menudo tienen la lacra de sufrir los mismos problemas de gestión de datos una y otra vez. La falta de normalización puede perjudicar de forma permanente los esfuerzos del data lake a medida que las demanda sigue aumentando, ya que los entornos simplemente no están preparados para la escalabilidad: la normalización y la coherencia son fundamentales.

Un proceso normalizado y una arquitectura coherente garantizan también que los recursos de la organización se centren en la innovación y el análisis, y no en la gestión de datos: cuanto más se centran los usuarios de TI y de las líneas de negocio en la gestión de datos, menos se concentran en impulsar las innovaciones que proporcionan la información más valiosa para el negocio.

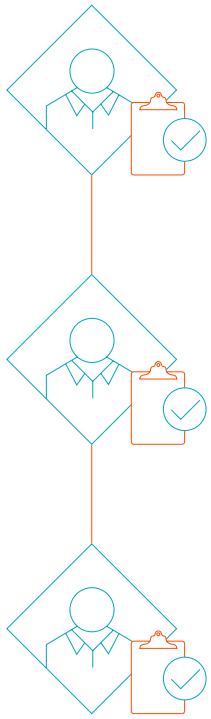


Establecimiento de taxonomías y clasificaciones para que todos los equipos vayan en la misma dirección

Uno de los mayores obstáculos para la velocidad, la agilidad y la colaboración es la ausencia de un lenguaje común. Si el personal de la organización no reconoce los activos de datos de forma coherente, se puede crear un malentendido de datos aislados que no maximiza su uso en toda la empresa. Además, los consumidores y los especialistas en datos informan a menudo de que pasan demasiado tiempo limpiando incoherencias en los datos, en lugar de centrarse en los esfuerzos de valor añadido del análisis.

La conservación de datos sin procesar en datos analizados y preparados sistemáticamente reduce considerablemente los gastos derivados de la preparación de los datos por parte de sus consumidores, como los científicos de datos. Los glosarios y las taxonomías normalizados, como parte de un completo programa de gestión de metadatos, también garantizan que todos los miembros del equipo de proyecto hablen el mismo idioma. Realizar una serie de ejercicios sencillos mediante los catálogos de datos para establecer cuáles son los activos de datos clave y cómo se va a hacer referencia a ellos puede eliminar muchas rotaciones y frustraciones más adelante.

Las taxonomías normalizadas también pueden simplificar radicalmente la auditoría y el seguimiento del linaje cuando los proyectos de data lake utilizan datos sensibles.



Preguntas que puede formularse:

¿Los usuarios del data lake pasan más tiempo con la gestión de datos (acceso, limpieza y transformación de datos para garantizar que los datos son aptos para su uso), o extrayendo valor de la información (ayudando a lograr resultados de negocios)?

¿Se ha definido la estrategia de metadatos para el data lake, asegurándose de configurar una taxonomía normalizada y un glosario de términos para los activos de datos, y garantizando la auditabilidad y transparencia necesarias?

¿Afectan los límites geográficos y de la organización a la estructura y el funcionamiento del data lake?

Conclusión

Los data lakes ofrecen la oportunidad única de proporcionar información de negocio completamente nueva de manera rápida y eficaz. La aplicación de las mejores prácticas en las soluciones y los procesos acelerará la entrega de información nueva. Además, al fomentar la colaboración en el proceso, los data lakes se liberan de una base de espacios de almacenamiento pasivos para utilizar mercados de datos activos para consumidores de datos. Las mejores prácticas descritas en este documento le ayudarán a evitar muchos escollos comunes y le permitirán empezar a trabajar con su data lake de la forma correcta.

Otras lecturas

Lea el informe ejecutivo sobre la gestión de data lakes inteligentes

Descubra cómo Informatica puede ayudarle a abordar los desafíos relacionados con los data lakes y le permite obtener información más precisa y coherente.

LEER MÁS

Acerca de Informatica

La transformación digital modifica las expectativas: mejor servicio, entrega más rápida, menores costes. Los negocios deben transformarse para seguir siendo relevantes y los datos tienen la respuesta.

Como líder mundial en gestión de datos de cloud empresariales, le brindamos ayuda para que encabece la marcha de forma inteligente, en cualquier sector, categoría o nicho. Informatica le aporta perspectiva para que aumente su agilidad, concrete nuevas oportunidades de crecimiento o incluso invente cosas nuevas. Al estar completamente centrados en todo lo relacionado con los datos, ofrecemos la versatilidad necesaria para alcanzar el éxito.

Le invitamos a explorar todo lo que puede ofrecerle Informatica y a desatar el poder de los datos para impulsar su próxima disrupción inteligente.

Sede central mundial
José Echegaray 8, edif. 3, PB 3, 28232 Las Rozas, Madrid
Teléfono: +34 91 787 61 40
Fax: 917 542 950

informatica.com/es
linkedin.com/company/informatica
twitter.com/Informatica
facebook.com/informaticaLLC

PÓNGASE EN CONTACTO CON NOSOTROS

IN18-0718-3492

© Copyright Informatica LLC 2018. Informatica y el logotipo de Informatica son marcas comerciales o marcas comerciales registradas de Informatica LLC en los Estados Unidos de Norteamérica y en otros países.

Informatica™