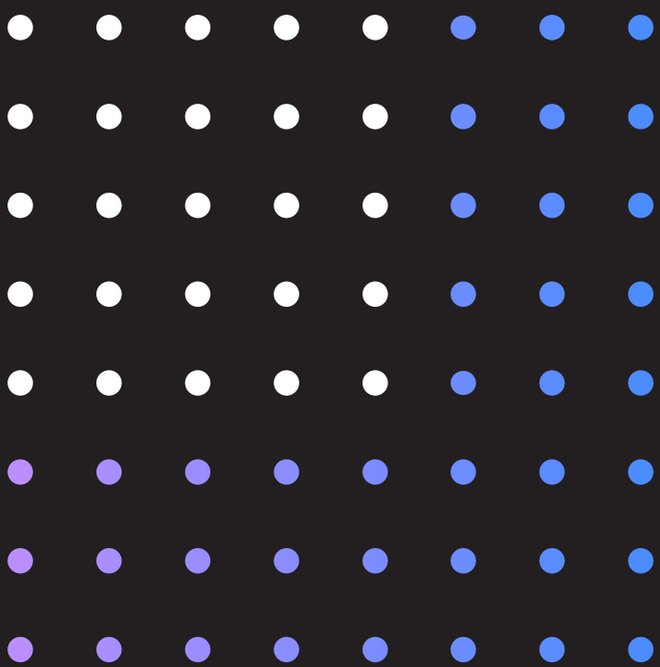
Entregue datos listos para el negocio con catalogado de datos inteligente y gobernanza del data lake

IBM Watson Knowledge Catalog proporciona una plataforma de gobernanza de datos impulsada por el machine learning para ayudarle con los retos del data lake



IBM

Índice

Aspectos clave

- Pocas organizaciones ven el valor que esperaban conseguir de los data lakes que han construido para almacenar y analizar sus datos y con ello conseguir información de confianza.
- DataOps resuelve los retos a los que se enfrentan las organizaciones con ineficiencias en el acceso, la preparación, la integración y la disponibilidad de los datos para los consumidores, a la vez que cumplen con las políticas normativas y corporativas.
- Los retos comunes de los data lakes incluyen la dificultad y el coste de la importación de nuevas fuentes de datos en el propio data lake; la incapacidad de integrar los conjuntos de datos externos e internos; la falta de confianza en lo que respecta a la gobernanza de datos; la falta de acceso a las herramientas de preparación de datos con autoservicio; y la incapacidad de encontrar y entender los datos que se encuentran en el data lake.
- Una plataforma de gobernanza de datos empresariales con catalogado, calidad de los datos y descubrimiento de los datos puede transformar un proyecto fallido de data lake en una auténtica fuente de valor empresarial.
- IBM Watson® Knowledge Catalog, impulsado por IBM Cloud Pak™ for Data, proporciona un catálogo de machine learning (ML) para el descubrimiento, el catalogado, la calidad y la gobernanza de los datos. Ayuda a los usuarios de los datos a descubrir, proteger, categorizar y compartir rápidamente los activos y conjuntos de datos, así como los modelos analíticos.
- Cuando a las organizaciones les falta una comprensión profunda de sus datos, se vuelve más complicado fiarse de esta información y utilizarla con todas las formas de Inteligencia Artificial (IA), incluyendo el machine learning y el deep learning.

Resuelva los retos de los data lakes con un enfoque de DataOps

Hace diez años empezó el viaje para encontrar un enfoque flexible y versátil que permitiera construir un almacén central de datos en el que pudieran alojarse todos los datos empresariales. La solución fue el data lake -un entorno de almacenaje de datos de uso general que pudiera almacenar casi cualquier tipo de dato. También permitiría a los analistas de datos y científicos de datos aplicar los motores y herramientas de análisis más adecuados para cada conjunto de datos, en su ubicación original.

Normalmente estos data lakes se construían utilizando Apache Hadoop y Hadoop Distributed File System (HDFS), combinados con motores tales como Apache Hive y Apache Spark. A medida que estos data lakes empezaron a crecer, surgieron una serie de problemas. Pese a que la estructura era físicamente capaz de adaptarse para capturar, almacenar y analizar vastas y variadas colecciones de datos estructurados y no estructurados, se prestó muy poca atención a la práctica de cómo incluir estas funciones en los flujos de trabajo empresariales.

Para el 2022, casi el 80 % de los proyectos de data lakes no podrán proporcionar el valor esperado, ya que encontrar, inventarizar y proteger los datos demostrará ser el mayor inhibidor para el éxito de ciencias de datos y análisis.[1] Como resultado, preguntas como: "¿Qué datos deberíamos poner en los data lakes?", "¿Quién los utilizará?", "¿Cómo hacemos que sea fácil para los usuarios encontrarlos?", "¿De dónde vienen estos datos?" y "¿Cómo evitamos el mal uso de los datos?" a menudo quedaban sin responder. Estas críticas limitaciones en el tratamiento de los problemas tecnológicos, de proceso y de usuarios llevó, efectivamente, a implementaciones fallidas de data lakes.

Hoy en día muchas organizaciones han reconocido sus errores, han cambiado los equipos de liderazgo por la implementación del data lake y están lanzando un segundo, tercer o incluso cuarto intento de implementar con éxito los data lakes -esta vez dirigidos por las operaciones de datos DataOps.

Este estudio realiza una evaluación de los retos comunes a los que se enfrentan los data lakes y proporciona nuevos enfoques como DataOps, que pueden ayudar a convertir los pantanos de datos en una pieza central del canal de datos listos para la empresa.

DataOps es una práctica de gestión de datos colaborativa que se centra en la mejora de la comunicación, la integración y la automatización de los flujos de datos entre gestores y consumidores de datos a través de una organización.

Presentación de DataOps

DataOps le trae las mejores prácticas de DevOps, la gestión de datos y la gobernanza de datos en un marco de trabajo común, con una forma de desarrollo colaborativa y manteniendo los flujos de datos a través de múltiples socios. DataOps está diseñado para resolver los retos a los que se enfrentan las organizaciones y que se asocian con ineficiencias en el acceso, la preparación, la integración y la disponibilidad de los datos para los consumidores, a la vez que cumplen con las políticas normativas y corporativas. Estas ineficiencias pueden encontrarse en una unidad empresarial, un equipo de análisis o incluso un proceso operativo.

Seguir esta metodología requiere el tratamiento de los problemas tecnológicos, de procesos y de usuarios que marcan la diferencia entre implementaciones de data lakes con o sin éxito. Desde el punto de vista tecnológico, DataOps recalca la importancia de usar una plataforma completa totalmente integrada para la ingestión e integración de los datos, así como su calidad, gobernanza y consumo para crear un data lake gobernado. Las normas de validación de la calidad de los datos deberían ejecutarse automáticamente como parte del proceso de ingestión para sostener un canal de datos continuos a lo largo de toda la empresa. El proceso de ingestión debería integrarse plenamente con el catálogo de datos, que se convierte en el centro de su canal. Los consumidores de datos deberían poder acceder a las puntuaciones de calidad de datos y los resultados de creación de perfiles del catálogo de datos, y confiar en que la organización esté trabajando con los mismos datos en contexto.

El crecimiento de los datos está superando la capacidad de las organizaciones de sacar provecho de estos. Cuando se les preguntó a las organizaciones cuáles eran los mayores retos para usar sistemas de información, estas respondieron: 1) El 40 % están fusionando procesos empresariales existentes con fuentes de datos para analizarlos y 2) el 39 % está abasteciéndose, recogiendo, gestionando y rigiendo los datos a medida que crecen.[2] Hoy en día ya no se trata solo de proteger las enormes inversiones de tiempo y recursos que ya se han realizado en tecnologías de data lake—es el hecho de que no hay alternativa posible. Para realizar la implementación de la IA, o incluso para realizar un análisis exhaustivo, es de vital importancia tener una vista plena de tantos datos como sea posible, lo que significa que necesitará una arquitectura que sea capaz de alojar, analizar y gobernar todos esos datos en un único sitio. En la mayoría de los casos, un data lake gobernado es la única opción realista para cumplir tales requisitos.

Los negocios de hoy en día pueden -y deben- encontrar una forma de extraer valor de su data lake al asegurar que dan soporte a un canal de datos listos para la empresa para DataOps.

Los retos de utilizar data lakes empresariales

Compartir datos

Cuando un equipo dentro de una empresa adquiere o crea un nuevo conjunto de datos, es muy probable que tengan una fuerte convicción del valor de los datos, y de la sensibilidad que los rodea. Si contiene información comercial confidencial, información que permita la identificación personal (IPIP) o datos de clientes, por ejemplo, el equipo sabrá cómo debería o no usar esa información, y tomará las precauciones necesarias para asegurar que nadie del equipo hace un mal uso de dicha información.

También serán conscientes de que fuera de su equipo otros potenciales usuarios de datos pueden no compartir la misma concepción del valor de los mismos, o de los riesgos asociados a su mal uso. Estos riesgos harán, por supuesto, que el equipo sea extremadamente cauteloso respecto a compartir los datos o almacenarlos en algún sitio que no esté bajo su control.

Esto son malas noticias para los data lakes. Si la empresa ve el data lake como un simple depósito de datos sin control, serán muy reticentes a confiar sus valiosos datos a esta. Como resultado, otras partes de la empresa no se podrán beneficiar de esos datos, y la totalidad del concepto del uso del data lake como repositorio de auto-servicio para compartir los datos empresariales queda anulado.

Integración de los datos
Incluso cuando un equipo se pone de acuerdo a que se integren sus datos en el data lake, la integración puede ser un proceso tortuoso. El concepto original del data lake implica importarlos en su formato propio sin procesar, de modo que no se necesiten los complejos procesos de extracción, transformación y carga (ETL) de un almacén de datos tradicional. Sin embargo, la realidad es que casi todas las fuentes de datos requieren algún grado de procesamiento previo antes de que puedan ser útiles para cualquier tipo de análisis que sea significativo.

Como resultado, integrar una nueva fuente de datos en un data lake puede, a veces, tardar meses. Y debido a que gran parte de estos datos han sido alojados previamente en pequeños silos operacionales y no en sistemas empresariales, puede haber docenas o incluso centenares de fuentes a integrar en total.

Esto significa que en muchos casos la información que necesitan los analistas empresariales o los científicos de datos no se ha añadido todavía al data lake, y probablemente siga así durante meses o incluso años. De nuevo, esto puede ser una importante barrera para la adopción.

Almacenamiento de datos
Pese a que el coste de los recursos computacionales y el almacenamiento de bienes ha disminuido radicalmente durante los últimos años, los clústeres de Hadoop no son gratuitos. Almacenar cantidades masivas de datos en un data lake es mucho más económico que hacerlo en maquinaria de almacén de datos de alto rendimiento, pero el coste todavía puede resultar ser significativo.

Además, a diferencia de los datos que se han almacenado tradicionalmente en almacenes de datos, la relación valor-volumen del Big Data alojado en el data lake es comparativamente bajo. Es posible que necesite almacenar un pajar enorme para localizar el puñado de agujas altamente valiosas en su interior.

Si no sabe qué conjuntos de datos serán realmente útiles y valiosos para sus científicos de datos, probablemente invertirá considerables sumas en la integración y almacenamiento de datos que están destinados a hundirse en el fondo de su data lake y nunca tener uso.

Encontrar datos
Incluso en el supuesto de que haya identificado los conjuntos de datos más valiosos para almacenar, haya persuadido a sus socios que los compartan y haya tenido éxito al integrarlos en su data lake, todavía deberá hacer posible que otros usuarios los puedan encontrar, entender y utilizar adecuadamente.

Los retos de utilizar data lakes empresariales

Figura 1. Las empresas que han adoptado tecnologías de data lakes pueden encontrarse con uno o más de estos problemas comunes.

La calidad de los datos en el data lake es otro desafío añadido. No está seguro de si los datos son de alta o baja calidad, pero son introducidos en el data lake.

Desgraciadamente esto no es tarea fácil en la mayoría de los data lakes. Los datos se almacenan a menudo sin ningún contexto, lo que hace que sea difícil o imposible para un nuevo usuario descodificarlos sin consultar al propietario original. La tecnología es a menudo tan específica con los dominios que una métrica utilizada en un área de la empresa puede conocerse con un nombre totalmente distinto -o definirse de una forma sutilmente diferente- en otra. La probabilidad de confusión y mala interpretación puede ser tan grande que muchos conjuntos de datos son de hecho inservibles, o incluso peligrosos, para un analista que no esté familiarizado con los mismos.

Combinación de datos internos y externos
Finalmente, incluso los mayores data lakes as mayores no deberían intentar alojar todos los conjuntos de datos posible que los científicos de la empresa quieran utilizar. Por ejemplo, no tendría sentido importar una réplica completa de Google Maps, Weather.com® o Bloomberg en su data lake solo porque uno de sus científicos de datos quiere realizar un análisis geoespacial, o integrar datos del tiempo o precios de la bolsa en un algoritmo.

Debido a que su data lake no podrá alojar todos los datos que sus analistas empresariales necesitan para los análisis, tendrán que pasar más tiempo buscándolos en múltiples aplicaciones. Puesto que gran cantidad de los análisis útiles

probablemente implicarán la combinación de conjuntos de datos internos y externos, esto eleva la barrera de nuevo en la entrada y desde la perspectiva del usuario reduce el valor que se percibe del data lake.

Preparación de los datos
Existen muchos factores que hacen que la preparación de los datos sea todo un reto -desde entender dónde encontrarlos, hasta formatearlos. La preparación de los datos para su uso en análisis es la tarea más ineficiente y que más tiempo consume a los usuarios de los datos. Estos se pasan la mayoría de su tiempo encontrando, limpiando y formateando información, en lugar de centrarse en el análisis, modelado y derivación de conocimientos para el impacto empresarial.

La accesibilidad limitada a los conjuntos de datos gobernados ha causado también un exceso de confianza en la TI durante la fase de preparación. Este acceso limitado señala la necesidad de mejorar las capacidades de auto-servicio y de uso de los datos a través de toda la empresa para aliviar este bloqueo en el camino.

Calidad de los datos
Tirar los datos en un data lake puede volverlo inservible. Puesto que no se aplican normas de calidad o de validación de los datos antes de que se transfieran al data lake, este no proporciona datos fiables y que puedan utilizarse. Los datos de alta calidad son una característica esencial que determinan la fiabilidad de los datos para tomar decisiones. Los datos son un valioso bien que debe ser gestionado a medida que se mueven por una organización. A medida que las fuentes de información crecen en número y diversidad, y las iniciativas de cumplimiento normativo se vuelven más específicas, la necesidad de integrar y acceder a la información desde estas dispares fuentes de formas consistentes, fiables y reutilizables se ha vuelto de vital importancia.

Un enfoque holístico a la construcción de data lakes gobernados

La mayoría de data lakes aprovechan Apache Hadoop y su amplio ecosistema de proyectos de fuente abierta para sus motores de análisis y capas de almacenamiento de datos. Tal y como es de esperar, la comunidad de fuentes abiertas relacionadas con Hadoop ha reconocido los problemas a los que se enfrentan las implementaciones de data lakes actuales, y de pronto han aparecido muchos proyectos que persiguen tratar los diversos problemas de forma individual. De forma similar, existe un número de herramientas de propietario en el mercado que se destinan a resolver estos mismos problemas.

Puede resultar tentador, pues, remediar los problemas de su data lake sobre la marcha, según aparecen. Cuando el número de conjuntos de datos se eleva demasiado como para poder gestionarse, añada una herramienta de catalogado. Cuando los usuarios se quejen de que no pueden encontrar los datos que necesitan, añada una portada con una función de búsqueda. Y cuando sus agentes de datos ya no puedan saber de dónde vinieron sus datos o quién los está utilizando, implemente herramientas de ascendencia de los datos y un marco de trabajo de gobernanza de los datos.

Suena fácil, pero en la práctica este enfoque desordenado tiende a tener el precio de una complejidad que va en aumento de forma masiva y una reducción de la capacidad de mantenimiento, especialmente a medida que aumentan el alcance y la escala del data lake. Del mismo modo que añadir nuevas fuentes de datos a un data lake aumenta la complejidad de sus requisitos de ETL, la adición de nuevas herramientas tiende a aumentar la complejidad de los requisitos no funcionales del data lake.

En lugar de tener una plataforma integrada de extremo a extremo que pueda integrar datos, llevar a cabo operaciones de calidad en sus datos y catalogar los mismos para su uso eficaz por parte de sus analistas empresariales, lo que normalmente encontrará es que cada herramienta tiene sus propias formas de gestionar los fallos, y su propio enfoque respecto al registro. Como resultado, la detección y resolución de problemas puede convertirse en una tarea que consuma mucho tiempo.

Otro defecto más importante del enfoque desordenado se vuelve aparente cuando toma una visión menos técnica y más conceptual de los problemas que los data lakes experimentan normalmente. La idea clave es que las capacidades de escalado, hallazgo, integración, calidad de los datos y gobernanza no son problemas separados: están interrelacionados de forma inextricable. Resolverlos requerirá un enfoque mucho más holístico.

Las capacidades de escalado, hallazgo, integración, calidad de los datos y gobernanza no son problemas separados: están interrelacionados de forma inextricable. Resolverlos requerirá un enfoque de la gestión de la información mucho más holístico.

IBM Watson Knowledge Catalog Descubrimiento, catalogado y calidad de los datos

IBM Watson Knowledge Catalog, impulsado por IBM Cloud Pak for Data, ayuda a los usuarios de los datos a descubrir, proteger, categorizar y compartir rápidamente activos de datos, conjuntos de datos, modelos analíticos y sus relaciones con otros miembros de la organización. Ayuda a los equipos de gobernanza de datos a definir el glosario, las políticas y las normas empresariales, además de proporcionar flujos de trabajo avanzados para la gobernanza. El catálogo sirve de fuente única de la verdad para los ingenieros, agentes y científicos de datos, así como los analistas empresariales, para conseguir el acceso de auto-servicio a los datos que pueden usar con total confianza.

Las soluciones como IBM Watson Knowledge Catalog impulsadas por IBM Cloud Pak for Data pueden proporcionar todas las funciones necesarias para resolver los problemas más importantes de los data lakes en una única y exhaustiva plataforma. El catálogo ayuda a tratar de raíz estos problemas interrelacionados: el fallo generalizado de los data lakes en la provisión de herramientas eficaces para capturar, almacenar y gestionar metadatos, y realizar el seguimiento de la ascendencia de los datos.

En muchos sentidos, el valor del data lake depende de los metadatos que este contiene, tanto como de los datos mismos. Sin los metadatos que expliquen de dónde viene un conjunto de datos, quién lo ha creado, qué contiene, quién tiene permiso para utilizarlo y cómo se utiliza, los datos en sí son prácticamente inútiles. Los usuarios no podrán encontrarlos, e incluso en el caso de que lo hagan, no podrán entender qué significan o confiar con seguridad en los mismos, o saber cómo pueden utilizarlos.

Watson Knowledge Catalog

Proporcionando datos significativos y de confianza

Organizar sus datos

Saber

Los datos deben ser completos, aplicables y accesibles desde cualquier sitio. Descubra, clasifique y entienda todos los tipos de datos.

Gobernar sus datos

Confiar

Los datos deben ser seguros, limpios y fáciles de encontrar para incentivar el acceso al auto-servicio de confianza. Entender de dónde vienen los datos y su calidad

Democratizar sus datos

Consumir

Capacidad de impulsar el descubrimiento auto-servicio y automatizar la toma de decisiones para hacer evolucionar el negocio. Proporcionar una vista de toda la información a aquellos que la necesitan y permitirles el acceso.

Figura 2. IBM Watson Knowledge Catalog proporciona un amplio abanico de funciones para el descubrimiento, el catalogado y la gobernanza de los datos.

Una única fuente de la verdad y un único punto de acceso

IBM Watson Knowledge Catalog impulsado por IBM Cloud Pak for Data trata estos problemas al convertir los metadatos en una prioridad clave. En su núcleo hay un potente motor de catalogado que indexa todos los conjuntos de datos y activos analíticos a los que su empresa tiene acceso, independientemente de si sus datos se alojan “tal cual” en su data lake, almacén de datos o sistema transaccional, o incluso en un grupo de hojas de datos. Sin importar si son estructurados o no, o si se almacenan in situ o se alojan en la nube. Además, el catálogo puede incluir también conjuntos de datos externos y fuentes, tales como los servicios de datos propietarios a los que se adscribe su empresa, o las APIs de datos abiertos.

El catálogo de datos no solo proporciona una única fuente de la verdad acerca de todos sus conjuntos de datos, sino que también le otorga un único punto de acceso. Las funciones de búsqueda y sugerencia de búsqueda impulsadas por la IA ayudan a los analistas de empresa, científicos de datos, ingenieros de calidad de datos y equipos de gobernanza de datos a encontrar activos más fácilmente, y a presentar los metadatos disponibles para ayudar a los usuarios a entender si lo que han encontrado y evaluado es útil para ellos o no.

Las funciones integradas de preparación de datos de auto-servicio aceleran el tiempo necesario para transformar los datos para su uso productivo en el análisis y las aplicaciones de IA. Los analistas y los científicos de datos no pierden el tiempo preparando y analizando los datos. La integración con una solución de preparación de datos a lo largo de toda la empresa, como IBM® InfoSphere® Advanced Data Preparation, le ayuda a asegurar que los conjuntos de datos gobernados que se crean a través de la superficie del catálogo para aquellos con el mayor contexto para impulsar los conocimientos empresariales y las acciones para los usuarios de la empresa. Esta integración aumenta la colaboración en todo el canal de datos.

Las capacidades de escalado, hallazgo, integración, calidad de los datos y gobernanza no son problemas separados: están interrelacionados de forma inextricable. Resolverlos requerirá un enfoque de la gestión de la información mucho más holístico.

El catálogo también ayuda a los agentes de los datos en la oficina del chief data officer (CDO) al etiquetar y clasificar los conjuntos de datos y hacer el seguimiento de su ascendencia y uso, y al equilibrar el glosario empresarial integrado para estandarizar la terminología empresarial a través de los datos. Como resultado, es más fácil para los agentes entender qué contiene cada conjunto de datos, dónde se encuentran los datos sensibles o información que permita la identificación personal, y quién debería tener acceso a estos.

Un único catálogo para múltiples fuentes de datos dentro y fuera de la organización

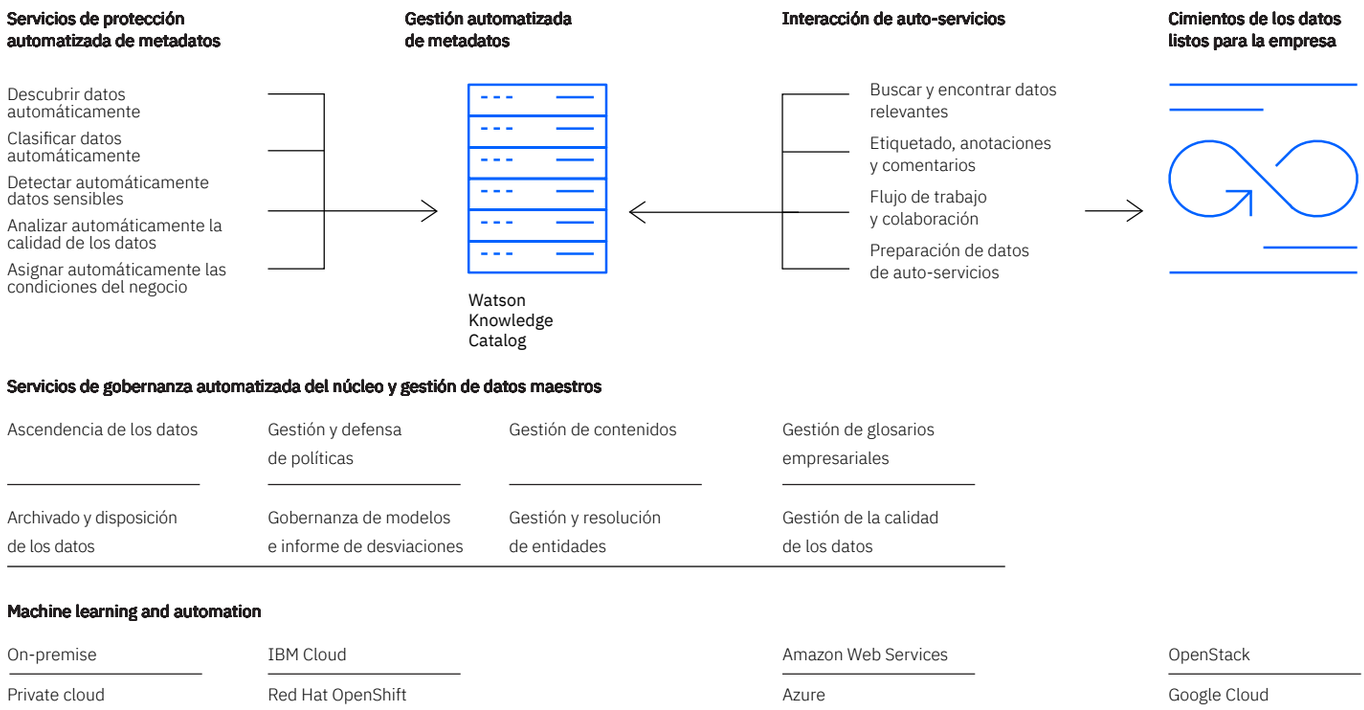


Figura 3. Con el índice inteligente de metadatos de IBM Watson Knowledge Catalog, los datos - tanto estructurados como no estructurados- pueden alojarse en sistemas originales, pero los usuarios pueden descubrirlos rápidamente para unos análisis más inteligentes.

IBM Watson Knowledge Catalog hace de los metadatos una prioridad clave, proporcionando una única fuente de la verdad y un único punto de acceso a todos los conjuntos de datos a los que su empresa tiene acceso.

Descubrimiento inteligente de datos integrado

Para mejorar la capacidad de búsqueda aún más, el catálogo permite a los usuarios etiquetar y comentar los conjuntos de datos y los activos analíticos, enriqueciendo los metadatos y añadiendo contenidos extra para ayudar a los cotrabajadores a encontrar lo que necesitan. La solución también incluye algoritmos integrados de descubrimiento de datos que utilizan ML para clasificar de forma automática los contenidos de cada conjunto de datos. Al identificar los tipos de campos comunes como nombres, direcciones, códigos postales y números de la seguridad social, la solución reduce la necesidad de anotar los datos manualmente para los autores. Proporciona automatización y ML para automatizar la protección de los datos y la gestión de los metadatos. Con funciones de calidad integradas, la solución permite la creación de perfiles de datos profundos, la calidad de los datos y las normas de validación.

Las operaciones de datos automatizadas proporcionan un canal de datos protegido, con calidad de los datos y gobernanza, y ayuda a asegurar que existe un flujo continuo de datos gobernados de alta calidad hasta el data lake.

De modo similar, la adición de un modelo de metadatos inteligente de sus activos proporciona una forma única de defender automáticamente, como el Reglamento General de Protección de Datos (RGPD) y la Ley de Privacidad del Consumidor de California (CCPA).

IBM Watson Knowledge Catalog, impulsado por IBM Cloud Pak for Data, le ayuda a proporcionar datos fiables, de alta calidad y listos para la empresa a, esencialmente, todos los usuarios de los datos.

Todos los componentes de la solución han sido diseñados como microservicios, con un único conjunto de principios de diseño y un enfoque común a los requisitos no funcionales, como la escabilidad, la gestión de errores, la seguridad y el registro.

IBM Watson Knowledge Catalog proporciona una plataforma de gobernanza empresarial de ML -para que esté listo para la IA a escala.

En lugar de los confusos errores y los cuellos de botella en el rendimiento que probablemente surgirán a causa de la adopción de un enfoque del tipo “hágalo usted mismo”, sobre la marcha, IBM Watson Knowledge Catalog proporciona una plataforma de gobernanza empresarial de ML, así que está listo para la IA a escala.

IBM Watson Knowledge Catalog está disponible en tres variaciones:

- Como solución de software como servicio (SaaS) en la IBM Cloud™
- En IBM Cloud Pak for Data
- Integrado en IBM Watson Studio

Las soluciones como IBM Watson Knowledge Catalog pueden desbloquear el valor que las iniciativas de data lakes prometían originalmente. Watson Knowledge Catalog con funciones de gobernanza y catalogado inteligentes ayuda a construir un data lake gobernado y fiable para la IA.

Los cuatro beneficios de construir un data lake gobernada para la IA

1. Genera confianza y seguridad en los datos a través de la calidad y la gobernanza.

- Las funciones de calidad de los datos le ayudan a mejorar la calidad de sus datos y a hacer que los datos de alta calidad estén disponibles en su data lake.
- Las políticas de gobernanza se configuran y defienden de forma automática -así que cuando encuentre un conjunto de datos, ya sabe si puede o no, y cómo, usarlo.
- Puede proteger sus datos mientras los usuarios añaden valoraciones, comentarios y otra información que ayudará a otros a determinar si un conjunto de datos les será de utilidad o no.

2. Empodera a sus usuarios de datos

- Sus equipos línea de negocio (LOB) comparten sus datos voluntariamente porque están seguros de que se gobernarán y protegerán debidamente contra el mal uso.
- Puede impulsar la colaboración y transformar los datos en bienes empresariales de confianza a través de políticas y defensa de datos dinámicos.
- Sus datos son, a medida que pasa el tiempo, más utilizables y fáciles de encontrar, puesto que los usuarios añaden etiquetas relevantes y metadatos para ayudar a otros a extraer valor de los mismos.
- Una única interfaz le da acceso a todos los conjuntos de datos propiedad de su organización, sin importar dónde se alojen.

3. Recupere su tiempo

- El descubrimiento automático de datos reduce el tiempo y esfuerzo que necesita dedicar a la adición de metadatos para nuevos conjuntos de datos.
- La protección automática de datos y la gestión de metadatos reducen el tiempo que se necesita para descubrir los metadatos y asignar términos, y también reduce el tiempo de creación del glosario empresarial.
- Con simples e intuitivas herramientas de preparación de datos auto-servicio, sus usuarios de datos pueden pasar menos tiempo preparandolos y más tiempo descubriendo información.
- Usted dirige a sus científicos de datos y analistas empresariales para proporcionar mejores análisis en menos tiempo.
- La búsqueda inteligente impulsada por la IA le ayuda a encontrar los datos que necesita en segundos, en lugar de esperar durante semanas a que otro equipo se los proporcione.

4. Gestione los crecientes datos y costes

- Puede optimizar los costes de almacenamiento al evitar los gastos de la ingestión de conjuntos de datos de bajo valor en el data lake.
- También puede ver todos los conjuntos de datos externos a los que se adscribe su empresa, lo que reduce el riesgo de pagar por más suscripciones de las que necesita.
- Puede dar prioridad a la ingestión de nuevas fuentes de datos en el data lake basándose en la demanda de los usuarios de los datos, lo que le ayuda a integrar sus fuentes más valiosas primero.

Desbloquear el valor de sus datos

Trabaje como analista o científico de datos, en el departamento de TI o en la oficina del CDO, usted y sus compañeros comparten un objetivo común. Si pueden construir un data lake que realmente proporcione servicio en sus instalaciones, no solo conseguirán que sus trabajos sean mucho más fáciles y productivos; Además, podrían jugar un papel clave en darle a su negocio la ventaja competitiva a la que pocas organizaciones pueden plantar cara actualmente.

Si pueden limpiar las aguas de sus data lakes mientras sus competidores están todavía arrastrándose entre los lodos de su pantano, podrá destapar posibilidades con las que ellos solo pueden soñar. Una gran ventaja les aguarda a aquellos que son los primeros en desbloquear el valor de los datos previamente no utilizados.

Conclusión

Sepa dónde se alojan todos sus datos, quién los está utilizando y su valor analítico para su empresa.

Los catálogos de datos son vitales para las iniciativas de DataOps porque pueden ayudar a proporcionar una gestión automatizada de metadatos abiertos al integrar la gobernanza de datos, su calidad y la gestión de políticas activas.

IBM Watson Knowledge Catalog con funciones de gobernanza y catalogado inteligentes ayuda a construir un data lake gobernado y fiable para la IA. El catálogo incorpora la integración de los datos, su calidad y su gobernanza en el entorno de su data lake para ayudarle a proporcionar datos listos para la empresa a los DataOps -y una única fuente de la verdad.

Información adicional

Si desea más información, visite:.
ibm.com/cloud/watson-knowledge-catalog

IBM España
Santa Hortensia, 26-28
28002 Madrid
España

Producido en los Estados Unidos de América en octubre de 2019, IBM, el logotipo de IBM, **ibm.com**, IBM Cloud, IBM Cloud Pak, IBM Watson e InfoSphere son marcas comerciales de International Business Machines Corp., registradas en diversas jurisdicciones de todo el mundo.

1. Catálogo aumentado de datos: Now an Enterprise Must-Have for Data and Analytics Leaders–Gartner, Sept 2019
2. The Forrester Wave: Machine Learning Data Catalogs, Q2 2018

ASW12449-ESES-03