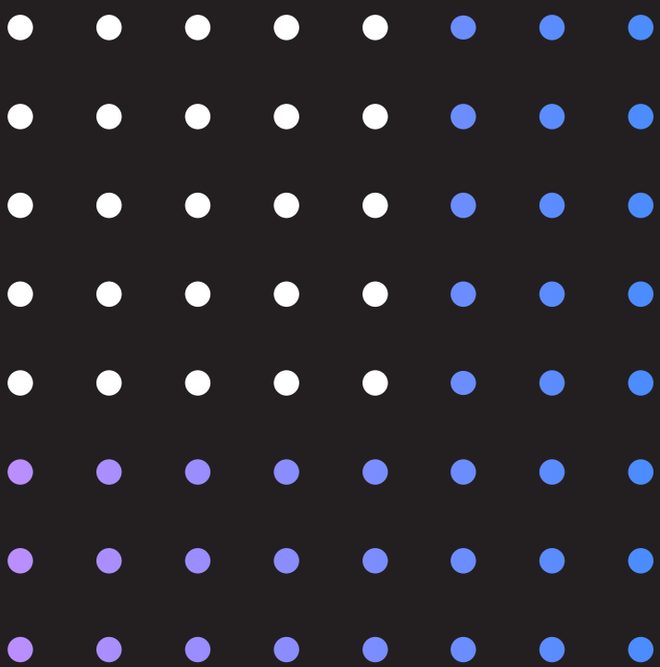
Entregue dados prontos para os negócios com a catalogação de dados inteligente e a governança de data lake.

O IBM Watson Knowledge Catalog fornece uma plataforma de governança de dados habilitada por machine learning para ajudar com os desafios do data lake.



IBM

Índice

03

Resolva os desafios do data lake com uma abordagem de DataOps

03

Desafios para o uso de data lakes corporativos

05

IBM Watson Knowledge Catalog

06

Uma única fonte da verdade e um único ponto de acesso

08

Quatro benefícios da construção de um data lake governado para IA

09

Conclusão

Principais conclusões

- Poucas organizações estão vendo o valor que esperavam dos data lakes que construíram para armazenar e analisar seus dados em busca de insights confiáveis.
- O DataOps resolve os desafios que as organizações enfrentam devido a ineficiências ao acessar, preparar, integrar e disponibilizar dados aos consumidores e, ao mesmo tempo, manter conformidade com as políticas corporativas e regulatórias.
- Os desafios comuns do data lake incluem a dificuldade e o custo de importar novas fontes de dados para o data lake; incapacidade de integrar conjuntos de dados internos e externos; falta de confiança em relação à governança de dados; ausência de acesso a ferramentas de preparação de dados de autoatendimento; e incapacidade de encontrar e entender os dados que estão no data lake.
- Uma plataforma de governança de dados corporativos com catalogação, qualidade e descoberta de dados pode transformar um projeto falho de data lake em uma verdadeira fonte de valor para os negócios.
- O IBM Watson® Knowledge Catalog, habilitado para IBM Cloud Pak™ for Data, fornece um catálogo de machine learning (ML) para descoberta, catalogação, qualidade e governança de dados. Ele ajuda os usuários a descobrir rapidamente, fazer curadoria, categorizar e compartilhar ativos de dados, conjuntos de dados e modelos analíticos.
- Quando as organizações não têm um entendimento profundo de seus dados, fica mais difícil confiar e usar essas informações com todas as formas de inteligência artificial (IA), incluindo ML e aprendizagem profunda.

Resolvendo os desafios do data lake com uma abordagem de DataOps

Há dez anos, começou a jornada para encontrar uma abordagem flexível e versátil para construir um armazenamento de dados central, em que todos os dados corporativos pudessem residir. A solução foi o data lake — um ambiente de armazenamento de dados de uso geral que armazenaria praticamente qualquer tipo de dados. Ele também permitiria aos analistas de negócios e cientistas de dados aplicar os mecanismos e ferramentas de análise mais apropriados para cada conjunto de dados, em seu local original.

Normalmente, esses data lakes foram criados usando o Apache Hadoop e o Hadoop Distributed File System (HDFS), combinados com mecanismos como o Apache Hive e o Apache Spark. À medida que esses data lakes começaram a crescer, um conjunto de problemas se tornou aparente. Embora a tecnologia fosse fisicamente capaz de ser dimensionada para capturar, armazenar e analisar coleções vastas e variadas de dados estruturados e não estruturados, pouca atenção foi dedicada aos aspectos práticos de como incorporar essas capacidades aos fluxos de trabalho de negócios.

Até 2022, mais de 80% dos projetos de data lake falharão em agregar valor, pois a descoberta, o inventário e a curadoria de dados serão os maiores inibidores do sucesso da análise e da ciência de dados.[1] Como resultado, surgiram as perguntas: “Quais dados devemos colocar no data lake?”, “Quem os usará?”, “Como facilitar sua localização?”, “De onde vêm esses dados?” e “Como impedimos que os dados sejam mal utilizados?” frequentemente ficavam sem resposta. Essas limitações críticas ao lidar com problemas referentes a pessoas, processos e tecnologia efetivamente levaram a falhas nas implementações de data lakes.

Hoje, muitas organizações reconheceram suas falhas, mudaram as equipes de liderança da implementação do data lake e estão lançando uma segunda, terceira ou quarta tentativa de implementar um data lake com sucesso — desta vez, liderando-as com as operações de dados DataOps.

Este whitepaper avalia os desafios comuns que os data lakes enfrentam e fornece novas abordagens, como o DataOps, que podem ajudar a transformá-los de um pântano de dados na peça central do pipeline de dados prontos para os negócios de uma organização.

DataOps é uma prática colaborativa de gerenciamento de dados, focada em melhorar a comunicação, integração e automação dos fluxos de dados entre gerenciadores e consumidores de dados em uma organização.

Introdução ao DataOps

O DataOps traz as melhores práticas de DevOps, gerenciamento de dados e governança de dados para uma estrutura comum, com uma maneira colaborativa de desenvolver e manter os fluxos de dados entre as diversas partes interessadas. O DataOps foi desenvolvido para resolver os desafios que as organizações enfrentam, associados a ineficiências ao acessar, preparar, integrar e disponibilizar dados aos consumidores e, ao mesmo tempo, manter conformidade com as políticas corporativas e regulatórias. Essas ineficiências podem ser encontradas em uma unidade de negócios, em uma equipe de análise ou mesmo em um processo operacional.

Seguir essa metodologia exige resolver os problemas referentes a pessoas, processos e tecnologia que significam a diferença entre o sucesso e a falha nas implementações de data lake. Do lado da tecnologia, o DataOps enfatiza a importância de usar uma plataforma totalmente integrada e de ponta a ponta para a ingestão e integração, qualidade, governança e consumo de dados para criar um data lake governado. As regras de validação da qualidade dos dados devem ser executadas automaticamente como parte do processo de ingestão para manter um pipeline de dados contínuo em toda a empresa. O processo de ingestão deve ser totalmente integrado ao catálogo de dados, que se torna o cerne de seu pipeline. Os consumidores de dados devem poder acessar as pontuações de qualidade de dados e os resultados da criação de perfil de dados do catálogo de dados e confiar que a organização está trabalhando com os mesmos dados em contexto.

O crescimento dos dados está superando a capacidade das organizações de obter valor com eles. Quando foi perguntado às organizações quais são os maiores desafios para o uso de sistemas de insight, elas responderam: 1) 40% estão mesclando os processos de negócios existentes com os dados da fonte para analisá-los e 2) 39% estão adquirindo, coletando, gerenciando e governando os dados à medida que eles crescem.[2] Hoje, não é apenas um caso de proteger os enormes investimentos em tempo e recursos que já foram feitos em tecnologias de data lake — o fato é que não há alternativa. Desde a implementação da IA ou até a realização de análises abrangentes, é fundamental ter uma visão completa do máximo de dados possível, o que significa que você precisa de uma arquitetura capaz de manter, analisar e governar todos esses dados em um único local. Em muitos casos, um data lake governado é a única opção realista para atender a esses requisitos.

As empresas de hoje podem — e devem — encontrar uma maneira de extrair valor de seu data lake, garantindo que ele ofereça suporte a um pipeline de dados prontos para os negócios de DataOps.

Desafios para o uso de data lakes corporativos

Compartilhamento dos dados
Quando uma equipe de uma empresa adquire ou cria um novo conjunto de dados, é provável que eles tenham um forte senso do valor dos dados e das sensibilidades circunjacentes. Se os dados contiverem informações comercialmente confidenciais, informações de identificação pessoal (PII) ou dados do cliente, por exemplo, a equipe saberá como essas informações devem e não devem ser usadas e tomará precauções para garantir que ninguém da equipe as use indevidamente.

Eles também terão consciência de que, fora da equipe, outros usuários dos dados em potencial podem não ter o mesmo entendimento do valor dos dados ou dos riscos associados ao seu uso indevido. Esses riscos naturalmente tornarão a equipe extremamente cautelosa ao compartilhar os dados ou armazená-los em qualquer lugar que não esteja sob seu controle.

Isso é uma má notícia para os data lakes. Se a empresa enxerga o data lake simplesmente como um depósito não controlado de dados, ficará muito relutante em confiar seus dados valiosos a ele. Como resultado, outras partes da empresa não poderão se beneficiar desses dados, e todo o conceito de usar o data lake como um repositório de autoatendimento para compartilhar dados corporativos se desfaz.

Integração dos dados
Mesmo quando uma equipe concorda com a integração de seus dados no data lake, o processo pode ser tortuoso. O conceito original do data lake é importar dados em seu formato bruto, sem exigir os processos complexos de extração, transformação e carregamento (ETL) de um data warehouse tradicional. No entanto, a realidade é que quase todas as fontes de dados exigem um certo grau de pré-processamento antes que possam ser úteis para qualquer tipo de análise significativa.

Como resultado, a integração de uma nova fonte de dados a um data lake pode frequentemente levar meses. E uma vez que muitos desses dados eram previamente mantidos em pequenos silos operacionais, e não em sistemas corporativos, pode haver dezenas ou mesmo centenas de fontes para integrar no total.

Isso significa que, em muitos casos, as informações de que os analistas de negócios ou os cientistas de dados precisam ainda não foram adicionadas ao data lake e não poderão ser adicionadas por meses ou até anos. Novamente, isso pode ser uma barreira significativa à adoção.

Armazenamento dos dados
Embora o custo do armazenamento de mercadorias e dos recursos de computação tenha diminuído drasticamente nos últimos anos, os clusters do Hadoop não são gratuitos. Armazenar enormes quantidades de dados em um data lake é muito mais econômico do que armazená-los em um dispositivo de armazenamento de dados de alta performance, mas o custo ainda pode ser significativo.

Além disso, diferentemente dos dados que são tradicionalmente armazenados em data warehouses, a razão valor/volume de big data mantidos em um data lake é comparativamente baixa. Você pode precisar armazenar um palheiro gigantesco para localizar um punhado de agulhas de alto valor dentro dele.

Se você não souber quais conjuntos de dados são realmente úteis e valiosos para seus cientistas de dados, poderá investir quantias consideráveis em integrar e armazenar dados que estão destinados a ficar depositados no fundo de seu data lake, sem nunca serem usados.

Localização dos dados
Supondo que você já tenha identificado os conjuntos de dados mais valiosos para armazenar, persuadido seus participantes

Desafios ao uso de data lakes corporativos

Figura 1. As empresas que adotaram tecnologias de data lake podem encontrar um ou mais desses problemas comuns.

a compartilhá-los e conseguido integrá-los ao seu data lake, você ainda precisa possibilitar que outros usuários os encontrem, entendam e usem adequadamente. A qualidade dos dados no data lake é outro desafio. Você não sabe com certeza se os dados são de alta ou baixa qualidade, mas eles estão sendo alimentados no lake.

Infelizmente, na maioria dos data lakes, isso não é fácil de alcançar. Os dados frequentemente são armazenados sem qualquer contexto, dificultando ou impossibilitando que um novo usuário os decodifique sem consultar o proprietário original. A terminologia é frequentemente tão específica do domínio que uma métrica usada em uma área da empresa pode ser conhecida por um nome completamente diferente — ou definida de uma maneira sutilmente diferente — por outra. O potencial de confusão e má interpretação pode ser tão grande que muitos conjuntos de dados são efetivamente inúteis ou mesmo perigosos para um analista que ainda não está familiarizado com eles.

Combinação entre dados internos e externos
Por fim, até o maior dos data lakes não deve tentar armazenar todos os conjuntos de dados possíveis que os cientistas de dados de uma empresa desejam usar. Por exemplo, não faria sentido importar uma réplica completa do Google Maps, Weather.com® ou Bloomberg para o seu data lake, apenas porque um de seus cientistas de dados deseja executar análises geoespaciais, integrar dados climáticos ou preços de ações em um algoritmo.

Como o data lake não armazenará todos os dados de que os analistas de negócios precisam para fazer as análises, eles terão que gastar um tempo procurando-os em diversas aplicações. Uma vez que uma proporção muito grande de análises úteis provavelmente envolverá a combinação entre conjuntos de dados internos e externos, isso mais uma vez aumenta a barreira à entrada e, da perspectiva do usuário, reduz o valor percebido do data lake.

Preparação dos dados
Existem muitos fatores que tornam a preparação dos dados desafiadora — desde entender onde encontrar os dados até, em seguida, formatá-los. Preparar os dados para uso em análises é a tarefa mais ineficiente e demorada para os usuários de dados. Os usuários de dados passam a maior parte do tempo encontrando, limpando e formatando informações, em vez de se concentrarem na análise de dados, na modelagem e na obtenção de insights sobre o impacto nos negócios.

A acessibilidade limitada aos conjuntos de dados governados também causou uma dependência excessiva da TI durante a fase de preparação. Esse acesso limitado sinaliza a necessidade de melhorar as capacidades de autoatendimento e as habilidades de alfabetização de dados em toda a empresa para amenizar esse obstáculo.

Qualidade dos dados
Despejar os dados em um data lake pode torná-los inutilizáveis. Uma vez que as regras de qualidade ou validação não estão sendo aplicadas aos dados antes de serem alimentados em um data lake, ele não fornece dados que possam ser confiáveis e usados. Os dados de alta qualidade são uma característica essencial que determina a confiabilidade dos dados para a tomada de decisões. Os dados são um ativo valioso, que deve ser gerenciado à medida que se move pela organização. À medida que as fontes de informações estão se tornando cada vez mais numerosas e diversas, e as iniciativas de conformidade regulatória são mais focadas, é fundamental a necessidade de integrar e acessar informações dessas fontes díspares de maneira consistente, confiável e reutilizável.

Uma abordagem holística para a construção de data lakes governados

A maioria dos data lakes utiliza o Apache Hadoop e seu amplo ecossistema de projetos de código aberto para suas camadas de armazenamento de dados e mecanismos de análise. Não é surpresa que a comunidade de código aberto do Hadoop reconheceu os problemas enfrentados pelas atuais implementações de data lake e recentemente surgiram muitos projetos com o objetivo de resolver os diversos problemas individualmente. Da mesma forma, existem diversas ferramentas de propriedade exclusiva no mercado que pretendem resolver os mesmos problemas.

Sendo assim, pode ser tentador corrigir os problemas do seu data lake de maneira fragmentada, à medida que eles surgirem. Quando o número de conjuntos de dados aumentar demais para ser gerenciável, adicione uma ferramenta de catalogação. Quando os usuários reclamarem que não conseguem encontrar os dados de que precisam, instale um front-end com uma função de pesquisa. E quando seus administradores de dados não puderem mais rastrear de onde os dados vieram ou quem os utiliza, implante ferramentas de linhagem de dados e uma estrutura de governança de dados.

Isso parece simples, mas, na prática, essa abordagem fragmentada tende a ocorrer à custa do aumento massivo da complexidade e da redução da capacidade de manutenção, especialmente à medida que a escala e o escopo do data lake aumentam. Da mesma forma que adicionar novas fontes de dados a um data lake aumenta a complexidade dos requisitos de ETL, a adição de novas ferramentas tende a aumentar a complexidade dos requisitos não funcionais do data lake.

Em vez de ter uma plataforma de ponta a ponta integrada que pode integrar dados, executar operações de qualidade em seus dados e catalogá-los para uso efetivo pelos analistas de negócios, você normalmente observará que cada ferramenta tem suas próprias maneiras de gerenciar as falhas e sua própria abordagem à criação de logs. Como resultado, a solução de problemas pode consumir muito tempo.

Outra desvantagem mais importante da abordagem fragmentada se torna aparente quando você adota uma visão menos técnica e mais conceitual dos problemas que os data lakes comumente enfrentam. O principal insight é que escalabilidade, capacidade de localização, integração, qualidade dos dados e governança não são problemas separados: eles estão inextricavelmente inter-relacionados. Resolvê-los exigirá uma abordagem muito mais holística.

Escalabilidade, capacidade de localização, integração, qualidade dos dados e governança não são problemas separados: eles estão inextricavelmente inter-relacionados. Resolvê-los exigirá uma abordagem muito mais holística ao gerenciamento de informações.

IBM Watson Knowledge Catalog, descoberta de dados, catalogação de dados e qualidade de dados

O IBM Watson Knowledge Catalog habilitado para IBM Cloud Pak for Data ajuda os usuários a descobrir rapidamente, fazer curadoria, categorizar e compartilhar ativos de dados, conjuntos de dados, modelos analíticos e suas relações com outros membros da organização. Ele ajuda as equipes de governança de dados a definir glossário, políticas e regras de negócios e fornece fluxos de trabalho avançados para governança. O catálogo atua como uma única fonte da verdade para que engenheiros de dados, administradores de dados, cientistas de dados e analistas de negócios obtenham acesso de autoatendimento aos dados em que podem confiar e usar com confiança.

Soluções como o IBM Watson Knowledge Catalog habilitado para IBM Cloud Pak for Data podem fornecer todas as capacidades exigidas para resolver os principais problemas dos atuais data lakes em uma plataforma única e abrangente. O catálogo ajuda a resolver a causa raiz desses problemas inter-relacionados: a falha generalizada dos data lakes para fornecer ferramentas eficazes para capturar, armazenar e gerenciar metadados e rastrear a linhagem de dados.

De diversas maneiras, o valor de um data lake depende dos metadados que ele contém, na mesma medida que depende dos dados em si. Sem os metadados para explicar de onde veio um conjunto de dados, quem o criou, o que ele contém, quem tem permissão para usá-lo e como está sendo usado, os dados em si são praticamente inúteis. Os usuários não conseguirão encontrá-los e, mesmo que o façam, não entenderão o que eles significam nem confiarão neles ou saberão como usá-los.

Watson Knowledge Catalog
Fornecendo dados confiáveis e significativos

Organize seus dados

Conhecer

Os dados devem ser completos, aplicáveis e acessíveis em qualquer lugar. Descubra, classifique e entenda todos os tipos de dados.

Governe seus dados

Confiar

Os dados devem ser seguros, limpos e fáceis de encontrar para incentivar um acesso de autoatendimento confiável. Entenda de onde os dados vieram e sua qualidade.

Democratize seus dados

Consumir

Capacidade de impulsionar a descoberta do autoatendimento e automatizar a tomada de decisões para avançar os negócios. Forneça uma visão de todas as informações para aqueles que precisam delas e permita que eles as acessem.

Figura 2. O IBM Watson Knowledge Catalog fornece uma ampla variedade de capacidades para descoberta, catalogação e governança de dados.

Uma única fonte da verdade e um único ponto de acesso

O IBM Watson Knowledge Catalog habilitado para IBM Cloud Pak for Data resolve esses problemas, tornando os metadados uma prioridade essencial. Em seu cerne, está um potente mecanismo de catalogação que indexa todos os conjuntos de dados e ativos analíticos aos quais sua empresa tem acesso, independentemente de onde residam, seja no data lake, no armazém de dados ou no sistema transacional ou até em um conjunto de planilhas. Independentemente de estarem estruturados, não estruturados ou armazenados no local ou hospedados na nuvem. Além disso, o catálogo também pode incluir conjuntos e fontes de dados externos, como serviços de dados de propriedade exclusiva que sua empresa assina ou APIs de dados abertas.

Além de fornecer uma única fonte da verdade sobre todos os seus conjuntos de dados, o catálogo de dados também fornece um único ponto de acesso. As capacidades de pesquisa e sugestão habilitadas para IA ajudam os analistas de negócios, cientistas de dados, engenheiros de qualidade de dados e equipes de governança de dados a encontrar os ativos mais facilmente e a apresentar os metadados disponíveis para ajudar os usuários a entender o que encontraram e avaliar se é útil para eles.

As capacidades incorporadas de preparação de dados de autoatendimento aceleram o tempo necessário para transformar os dados para uso produtivo em aplicações de IA e análise. Os analistas de negócios e cientistas de dados não precisam perder tempo preparando e analisando os dados. A integração com uma solução de preparação de dados para toda a empresa, como o IBM® InfoSphere® Advanced Data Preparation ajuda a garantir que os conjuntos de dados governados, criados por meio do catálogo, sejam exibidos para aqueles com mais contexto para impulsionar insights e ações de negócios para os usuários de negócios. Essa integração promove a colaboração em todo o pipeline de dados.

Escalabilidade, capacidade de localização, integração, qualidade dos dados e governança não são problemas separados: eles estão inextricavelmente inter-relacionados. Resolvê-los exigirá uma abordagem muito mais holística ao gerenciamento de informações.

O catálogo também ajuda os administradores de dados no escritório do diretor executivo de dados (CDO), marcando e classificando os conjuntos de dados e rastreando sua linhagem e uso automaticamente, e também aproveitando o glossário de negócios incorporado para padronizar a terminologia entre os dados. Como resultado, é mais fácil para os administradores de dados entender o que cada conjunto de dados contém, onde estão as informações confidenciais ou PII e quem deve ter permissão para acessá-las.

Um único catálogo para diversas fontes de dados dentro e fora da organização

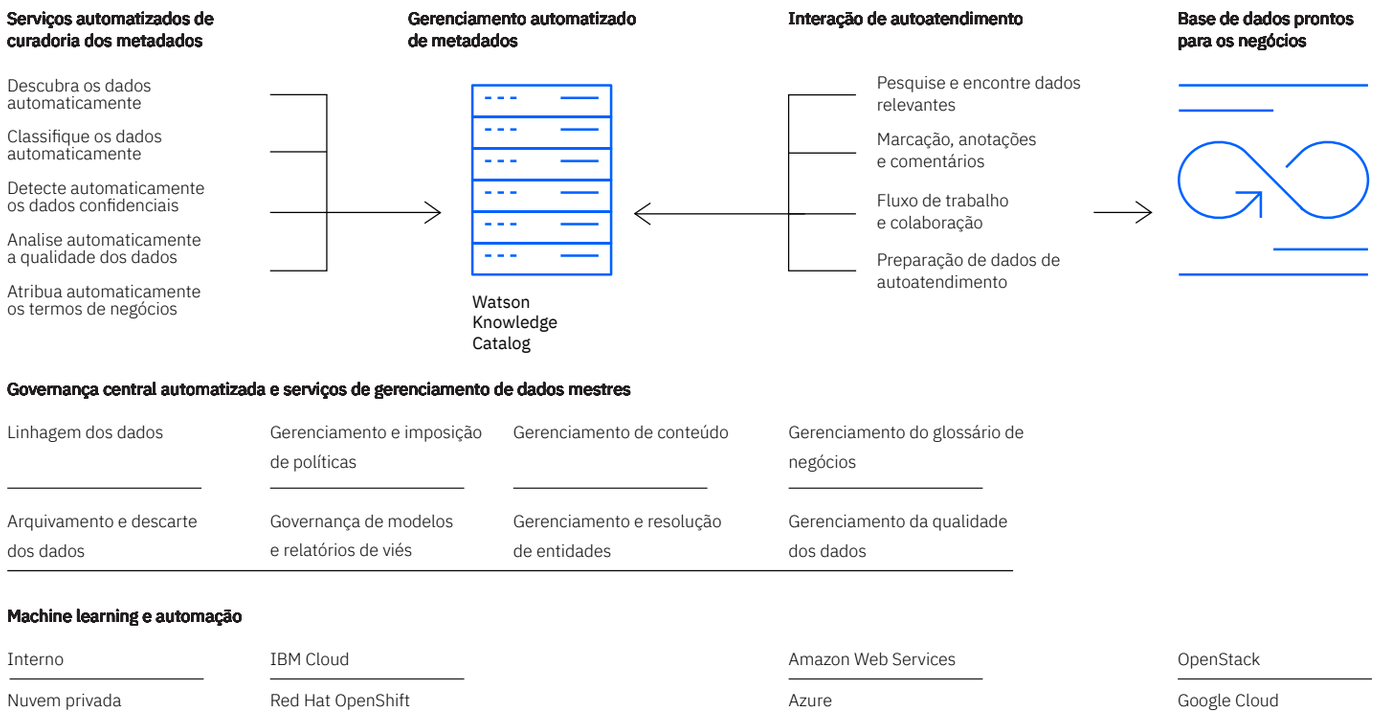


Figura 3. Com o índice de metadados inteligentes do IBM Watson Knowledge Catalog, os dados — tanto estruturados como não estruturados — podem residir em sistemas originais, mas os usuários podem descobri-los rapidamente para análises mais inteligentes.

O IBM Watson Knowledge Catalog torna os metadados uma prioridade essencial, fornecendo uma única fonte da verdade e um único ponto de acesso a todos os conjuntos de dados aos quais sua empresa tem acesso.

Descoberta de dados inteligente incorporada

Para melhorar ainda mais a capacidade de localização, o catálogo permite que os usuários marquem e comentem os conjuntos de dados e ativos analíticos, enriquecendo os metadados e adicionando contexto extra para ajudar os colegas a encontrar o que precisam. A solução também inclui algoritmos de descoberta de dados incorporados que usam o ML para classificar automaticamente o conteúdo de cada conjunto de dados. Ao identificar tipos de campos comuns como nomes, endereços, códigos postais e números de identidade, a solução reduz a necessidade de os autores anotarem os dados manualmente. Ela infunde a automação e o ML para automatizar a curadoria de dados e o gerenciamento de metadados. Com as funções de qualidade de dados incorporadas, a solução permite regras profundas de criação de perfil, qualidade dos dados e validação.

As operações automatizadas de dados fornecem um pipeline de dados submetidos à curadoria, com qualidade de dados e governança, e ajudam a garantir que haja um fluxo contínuo de dados governados de alta qualidade no data lake.

Da mesma forma, a adição de um modelo inteligente de metadados de seus ativos fornece uma maneira única de impor regras automaticamente, como o General Data Protection Regulation (GDPR, Regulamento Geral de Proteção de Dados) e a California Consumer Privacy Act (CCPA, Lei de Privacidade do Consumidor da Califórnia).

O IBM Watson Knowledge Catalog habilitado para IBM Cloud Pak for Data ajuda a fornecer dados confiáveis, de alta qualidade e prontos para os negócios, essencialmente para todos os usuários de dados.

Todos os componentes da solução foram projetados como microsserviços, com um único conjunto de princípios de design e uma abordagem comum aos requisitos não funcionais como escalabilidade, gerenciamento de erros, segurança e criação de logs.

O IBM Watson Knowledge Catalog fornece uma plataforma de governança corporativa de ML — para total prontidão para a IA em escala.

Em vez dos erros confusos e gargalos de performance que provavelmente resultam de uma abordagem fragmentada, do tipo “faça você mesmo”, o IBM Watson Knowledge Catalog fornece uma plataforma de governança corporativa de ML, para total prontidão para a IA em escala.

O IBM Watson Knowledge Catalog está disponível em três variantes:

- Como solução de software como serviço (SaaS) no IBM Cloud™
- No IBM Cloud Pak for Data
- Integrado ao IBM Watson Studio

Soluções como o IBM Watson Knowledge Catalog podem desbloquear o valor que as iniciativas de data lake prometeram originalmente. O Watson Knowledge Catalog, com capacidades inteligentes de catalogação e governança, ajuda a criar um data lake confiável e governado para a IA.

Quatro benefícios da construção de um data lake governado para IA

1. Criar confiança nos dados por meio de qualidade e governança

- As capacidades de qualidade dos dados ajudam a melhorar a qualidade de seus dados e disponibilizar dados de alta qualidade em seu data lake.
- As políticas de governança são definidas e aplicadas automaticamente — assim, quando você encontra um conjunto de dados, sabe se e como tem permissão para usá-los.
- Você pode fazer a curadoria de seus dados à medida que os usuários adicionam classificações, comentários e outras informações que ajudarão outras pessoas a determinar se um conjunto de dados será ou não útil para elas.

2. Capacitar seus usuários de dados

- As equipes de linha de negócios (LOB) compartilham seus dados de bom grado, porque estão confiantes de que eles serão governados e protegidos adequadamente contra o uso indevido.
- Você pode impulsionar a colaboração e transformar os dados em ativos corporativos confiáveis por meio de políticas de dados dinâmicos e imposições.
- Seus dados tornam-se mais fáceis de encontrar e reutilizáveis com o tempo, à medida que os usuários adicionam marcas e metadados relevantes para ajudar outras pessoas a extrair valor deles.
- Uma única interface fornece acesso a cada conjunto de dados que sua organização possui, independentemente de onde estão armazenados.

3. Recuperar seu tempo

- A descoberta automática de dados reduz o tempo e o esforço que você precisa dedicar para adicionar metadados a novos conjuntos de dados.
- A curadoria automática de dados e o gerenciamento de metadados reduzem o tempo necessário para descobri-los e atribuir termos, além de reduzir o tempo da criação do glossário de negócios.
- Com ferramentas simples e intuitivas de preparação de dados de autoatendimento, seus usuários de dados gastam menos tempo preparando os dados e mais tempo descobrindo insights.
- Você libera seus cientistas de dados e analistas de negócios para fornecer análises melhores em um espaço de tempo mais curto.
- A pesquisa inteligente, habilitada para IA, o ajuda a encontrar os dados de que você precisa em segundos, em vez de esperar semanas para que outra equipe os forneça.

4. Gerenciar o aumento de dados e custos

- Você pode otimizar os custos de armazenamento evitando as despesas da ingestão de conjuntos de dados de baixo valor no data lake.
- Você também pode ver todos os conjuntos de dados externos que sua organização assina, reduzindo o risco de pagar por mais assinaturas do que o necessário.
- Você pode priorizar a ingestão de novas fontes de dados no data lake com base na demanda dos usuários pelos dados, o que o ajuda a integrar primeiro as fontes mais valiosas.

Desbloqueie o valor de seus dados

Independentemente de você trabalhar no escritório do CDO, no departamento de TI ou como cientista ou analista de dados LOB, você e seus colegas compartilham um objetivo comum. Se você puder construir um data lake que realmente cumpra suas promessas, você poderia não apenas tornar os trabalhos dessas pessoas muito mais fáceis e produtivos. Você também poderia cumprir uma função fundamental em proporcionar à sua empresa uma vantagem competitiva, que poucas organizações podem rivalizar atualmente.

Se você puder limpar as águas do seu data lake enquanto seus concorrentes ainda estão se debatendo no pântano, você abrirá possibilidades com as quais eles só podem sonhar. Uma vantagem genuína do pioneiro na adoção aguarda aqueles que são os primeiros a desbloquear o valor dos dados anteriormente inexplorados.

Conclusão

Saiba onde todos os seus dados residem, quem os está usando e seu valor para os negócios para análise.

Os catálogos de dados são iniciativas críticas para DataOps, porque podem ajudar a fornecer gerenciamento automatizado de metadados abertos integrando governança de dados, qualidade e gerenciamento ativo de políticas.

O IBM Watson Knowledge Catalog, com capacidades inteligentes de catalogação e governança, ajuda a criar um data lake confiável e governado para a IA. O catálogo incorpora integração, qualidade e governança de dados em seu ambiente de data lake para ajudar a fornecer dados prontos para os negócios para DataOps — e uma única fonte da verdade.

Para mais informações

Para saber mais, acesse:
ibm.com/cloud/watson-knowledge-catalog

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produzido nos Estados Unidos da América, outubro de 2019 IBM, o logotipo IBM, **ibm.com**, IBM Cloud, IBM Cloud Pak, IBM Watson e InfoSphere são marcas comerciais da International Business Machines Corp., registradas em diversas jurisdições no mundo todo.

Red Hat e OpenShift são marcas comerciais ou marcas registradas da Red Hat, Inc. ou de suas subsidiárias nos Estados Unidos e em outros países. Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. Uma lista atual das marcas comerciais da IBM está disponível na Internet em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml (em inglês).

1. Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders – Gartner, setembro de 2019
2. The Forrester Wave: Machine Learning Data Catalogs, T2 2018

ASW12449-BRPT-03