

QUALIDADE E GOVERNANÇA DE DADOS

Os seis princípios dos dados prontos para IA

Criando uma base confiável de dados para IA

Índice

Sumário executivo	3
Introdução	3
Os seis princípios dos dados prontos para IA	4
O AI Trust Score	12
Base de dados Qlik Talend para IA	13
Conclusão	14

Sumário executivo

- Este documento apresenta seis princípios para garantir que os dados estejam prontos para uso com a inteligência artificial (IA).
- Esses princípios são os seguintes: os dados precisam ser diversos, oportunos, precisos, seguros, identificáveis e facilmente consumíveis pelas máquinas.
- O documento também descreve o AI Trust Score, que ajuda a avaliar o quão aderentes os seus dados são em relação a esses princípios.

Introdução

A inteligência artificial (IA) promete levar grandes avanços a segmentos como saúde, manufatura e atendimento ao cliente, proporcionando experiências de mais qualidade para consumidores e colaboradores. De fato, tecnologias de IA, como o aprendizado de máquina (ML), já auxiliam os profissionais de dados a produzir previsões matemáticas, gerar insights e melhorar a tomada de decisões. Além disso, tecnologias emergentes de IA, como a IA generativa (GenAI), podem criar conteúdo extremamente realista, com potencial de aumentar a produtividade em praticamente todos os aspectos do negócio.

De acordo com o Gartner, até 2025, a GenAI será uma parceira de trabalho das equipes de 90% das empresas em todo o mundo, e até 2026, mais de 80% das companhias terão implementado aplicações habilitadas por IA generativa em produção¹

No entanto, a IA não vai funcionar bem sem dados de qualidade, e este paper descreve os seis princípios para garantir que os seus dados estejam prontos para a IA.

¹ "Analysts to Discuss Generative AI Trends and Technologies", Gartner, Outubro, 2023.

Os seis princípios dos dados prontos para IA

Seria ingênuo pensar que você poderia apenas jogar dados em diversas iniciativas de IA e esperar que ocorresse uma mágica, mas é isso o que muitos profissionais fazem. Embora essa abordagem pareça funcionar nos primeiros projetos de IA, os Cientistas de Dados passam cada vez mais tempo corrigindo e preparando os dados à medida que os projetos amadurecem.

Além disso, os dados usados para IA precisam ser de alta qualidade e preparados especificamente para essas aplicações inteligentes. Isso significa passar muitas horas aprimorando e limpando manualmente os dados para garantir a precisão e a integridade deles, organizando-os de tal forma que as máquinas consigam facilmente entendê-los. Ainda, esses dados geralmente exigem informações adicionais — como definições e rótulos — para enriquecer o significado semântico do aprendizado automatizado e para ajudar a IA a realizar as tarefas com mais eficiência.

Portanto, quanto mais cedo os dados puderem ser preparados para os processos de IA downstream, maiores serão os benefícios. Usar dados preparados e prontos para IA é como dar vegetais já limpos e picados para um chef de cozinha, em vez de uma sacola cheia de mantimentos — isso economiza esforço e tempo, e ajuda a garantir que o prato seja entregue prontamente. O diagrama abaixo define seis princípios críticos para garantir a “prontidão” dos dados e a adequação deles para o uso da IA.

Os capítulos a seguir abordam cada princípio com mais detalhes.

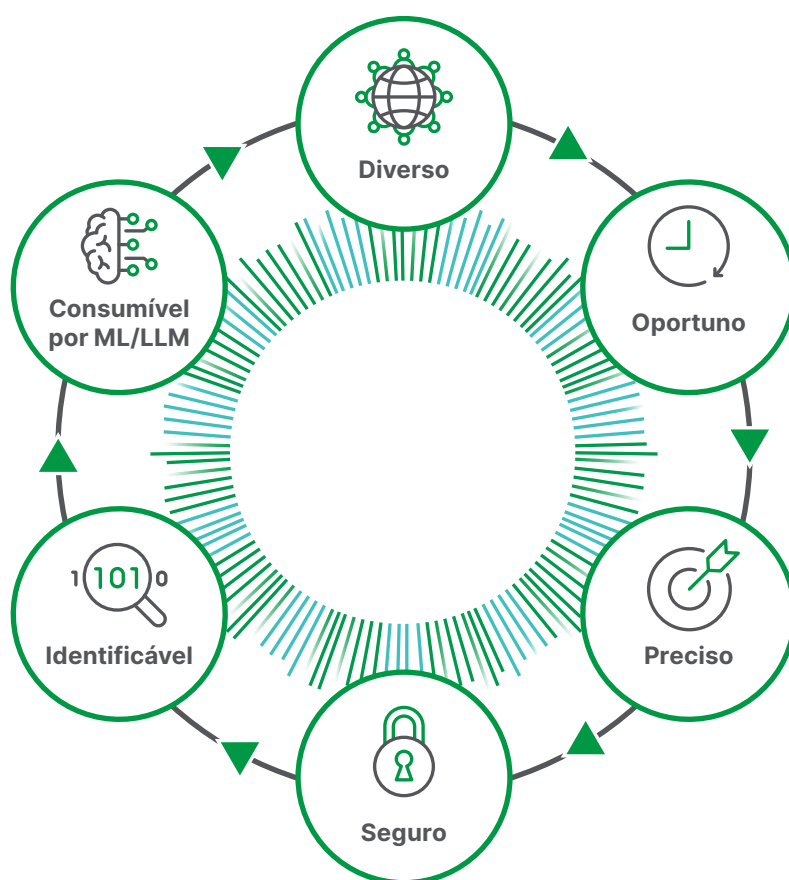


Figura 1. Os seis princípios dos dados prontos para IA

1 Os dados precisam ser diversos.

Os vieses nos sistemas de IA, também conhecidos como 'vieses de algoritmo' ou 'vieses de aprendizado de máquina', ocorrem quando aplicações de IA produzem resultados que refletem os vieses humanos, como a desigualdade social. Isso acontece quando o processo de desenvolvimento do algoritmo conta com suposições prejudiciais ou, o que é mais comum, quando os dados de treinamento têm vieses. Por exemplo: um algoritmo de score de crédito pode negar um empréstimo caso ele utilize um espectro restrito de atributos financeiros.

Consequentemente, o nosso primeiro princípio foca em oferecer uma ampla variedade de dados aos modelos de IA, o que aumenta a diversidade dos dados e reduz os vieses, ajudando a garantir que as aplicações de IA sejam menos propensas a tomarem decisões injustas.

Contar com dados diversos significa que os seus modelos de IA não serão construídos sob datasets restritos e em silos. Em vez disso, você conta com uma gama ampla de fontes de dados, com diversos padrões, perspectivas, variações e cenários relevantes para o domínio do problema. Esses dados podem ser bem estruturados e ficar armazenados na nuvem ou localmente. Eles também podem ser armazenados em um mainframe, banco de dados, sistema SAP ou aplicação 'Software as a Service' (SaaS). Por outro lado, os dados de origem também podem ser não estruturados e ficar armazenados como arquivos ou documentos em um drive corporativo.

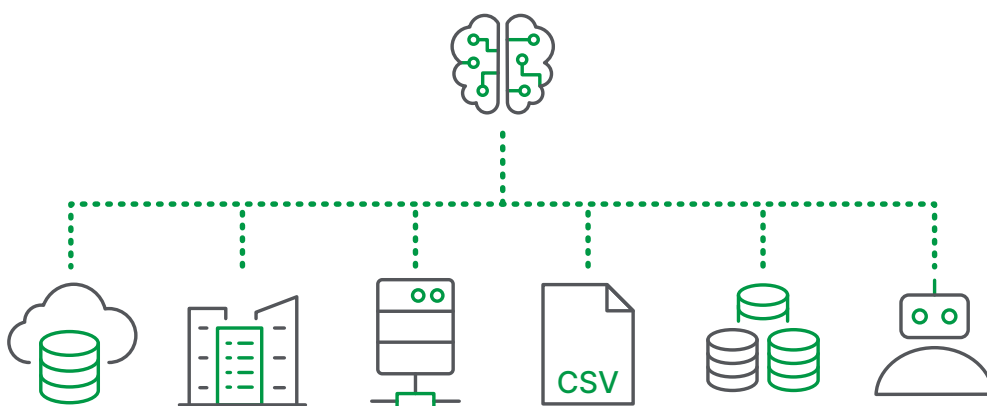


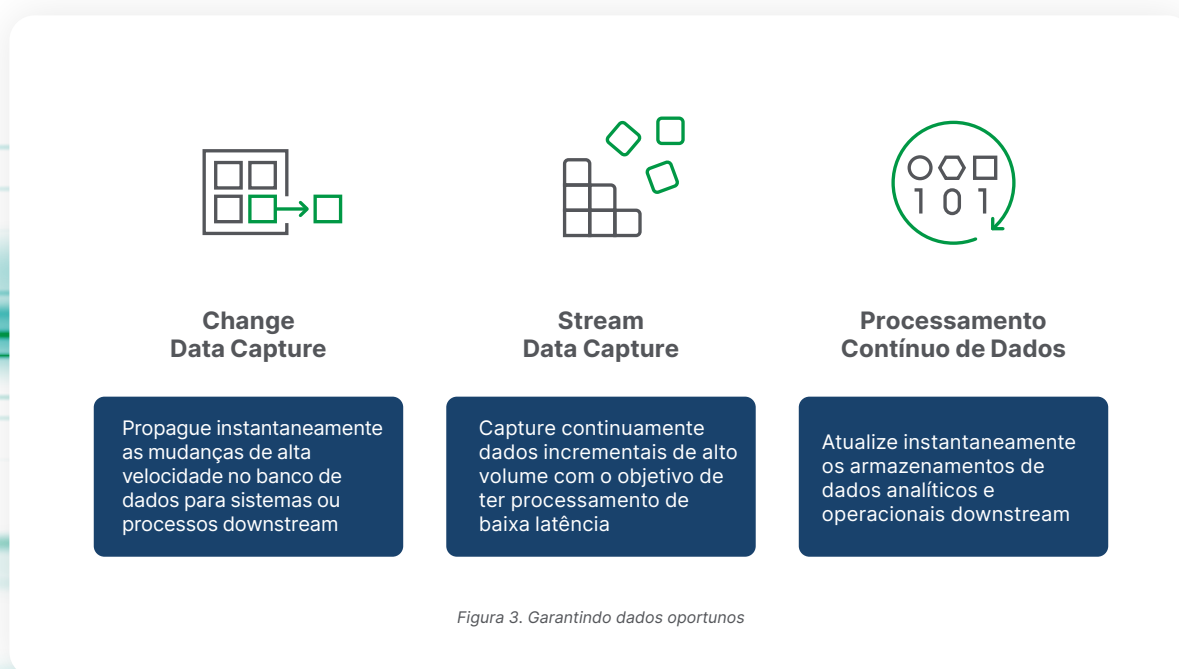
Figura 2. Diversidade de dados

É fundamental coletar dados diversos em inúmeros formatos para integrá-los às suas aplicações de aprendizado de máquina e IA generativa.

2 | Os dados precisam ser oportunos.

Embora seja verdade que as aplicações de aprendizado de máquina (ML) e IA generativa prosperem com dados diversos, o frescor dos dados também é crucial. Assim como a previsão do tempo baseada nas condições de ontem não é válida para você planejar a viagem que fará hoje, os modelos de IA treinados com informações desatualizadas podem produzir resultados imprecisos ou irrelevantes. Além disso, dados atualizados permitem que os modelos de IA acompanhem as tendências, se adaptem às circunstâncias em constante mudança e entreguem os melhores resultados possíveis. Portanto, o segundo princípio dos dados prontos para IA é ser oportuno.

É fundamental que você crie e implemente pipelines de dados em tempo real e de baixa latência para as suas iniciativas de IA para garantir dados oportunos. **Change data capture (CDC)** é geralmente usado para entregar dados oportunos a partir de sistemas de bancos de dados relacionais, enquanto o **stream capture** é usado para dados que se originam de dispositivos IoT, que exigem processamento de baixa latência. Assim que os dados são capturados, os repositórios-alvo são atualizados e as **mudanças são continuamente aplicadas** quase tempo real para obter os dados mais recentes possíveis.



Lembre-se: dados oportunos permitem previsões mais precisas e embasadas.

3 Os dados devem ser precisos.

O sucesso de qualquer iniciativa de aprendizado de máquina ou IA generativa depende de um único ingrediente: dados corretos. Isso acontece porque os modelos de IA agem como esponjas sofisticadas, absorvendo as informações para aprender e realizar tarefas. Se a informação não for precisa, é como se a esponja estivesse absorvendo água suja, levando a resultados com vieses, criações sem sentido e, em última análise, a um sistema de IA que não funciona. Dessa forma, a precisão dos dados é o terceiro princípio, fundamental para construir aplicações de IA confiáveis.

A precisão dos dados tem três aspectos. O primeiro é o **profiling dos dados de origem**, para entender suas características, integridade, distribuição, redundância e forma. O profiling também é conhecido como 'Análise Exploratória de Dados', ou 'EDA'.

O segundo aspecto é a **operacionalização de estratégias de correção**, que ocorre por meio da construção, implementação e monitoramento contínuo da eficiência das regras de qualidade dos dados. Os seus administradores de dados talvez precisem se envolver aqui para ajudar com a deduplicação e fusão dos dados. Como alternativa, a IA também pode ajudar a automatizar e acelerar o processo por meio de sugestões de qualidade de dados recomendadas por máquina.

O aspecto final é permitir a linhagem de dados e a análise de impacto — com ferramentas para Cientistas e Engenheiros de Dados que vão destacar o impacto das potenciais mudanças de dados e traçar a origem deles, visando impedir a modificação acidental dos dados usados pelos modelos de IA.



Dados precisos e de alta qualidade garantem que os modelos consigam identificar relações e padrões relevantes, levando a decisões, geração e previsões mais precisas.

4 | Os dados precisam ser seguros.

Geralmente, os sistemas de IA usam dados sensíveis — incluindo Informações de Identificação Pessoal (PII), registros financeiros ou dados proprietários de negócio —, e a utilização desses dados exige responsabilidade. Deixar dados sem segurança em aplicações de IA é como deixar a porta do cofre aberta. Pessoas mal-intencionadas podem roubar informações sensíveis, manipular os dados de treinamento para gerar resultados com vieses ou até mesmo causar danos a sistemas inteiros de IA generativa. A segurança dos dados é essencial para proteger a privacidade, manter a integridade do modelo e garantir o desenvolvimento responsável de aplicações poderosas de IA. E assim, a segurança é o quarto princípio dos dados prontos para IA.

Novamente, três táticas podem ajudar você a automatizar a segurança dos dados em escala, já que é praticamente impossível fazer isso manualmente. A **classificação dos dados** detecta, categoriza e rotula os dados que alimentam o estágio seguinte. A **proteção dos dados** define políticas como o mascaramento, a tokenização e a encriptação para ofuscar os dados. Finalmente, a **segurança dos dados** define políticas que descrevem o controle de acessos (ex.: quem pode acessar os dados). Esses três conceitos atuam juntos, da seguinte forma: primeiro, as camadas de privacidade devem ser definidas e os dados devem ser rotulados com as designações de segurança 'sensíveis', 'confidenciais' ou 'restritos'. Depois, uma política de proteção deve ser aplicada para mascarar os dados restritos. E finalmente, uma política de controle de acesso deve ser usada para limitar os direitos de acesso.



Estas três táticas protegem os seus dados e são cruciais para melhorar a confiança geral em seu sistema de IA e proteger o seu valor reputacional.

5 | Os dados precisam ser identificáveis.

Os princípios que abordamos até agora se concentraram principalmente em entregar prontamente os dados certos, no formato correto, às pessoas, sistemas e aplicações de IA certos. Mas reunir dados não é o suficiente. Os dados prontos para IA precisam ser identificáveis e estar prontamente acessíveis no sistema. Imagine uma biblioteca com todos os livros trancados — o conhecimento está lá, mas ele não é acessível. Os dados identificáveis revelam o verdadeiro potencial do aprendizado de máquina e da IA generativa, possibilitando a essas ferramentas aprender, se adaptar e produzir resultados inovadores. Assim, a capacidade de ser identificável é o quinto princípio dos dados prontos para IA.

Não é de se surpreender que as boas práticas de metadados estejam no centro da capacidade de identificação. Além dos metadados técnicos associados aos datasets de IA, os metadados de negócios e semantic typing também precisam ser definidos. **Semantic typing** oferece um significado extra para os sistemas automatizados, enquanto os termos de negócios adicionais entregam contexto extra para auxiliar a compreensão humana. Uma boa prática é criar um **glossário de negócios**, capaz de conectar os termos de negócios aos itens técnicos nos datasets, garantindo um entendimento comum de conceitos. A ampliação assistida por IA também pode ser usada para gerar documentação automaticamente e adicionar semântica de negócios do glossário. Finalmente, todos os metadados são indexados e tornados pesquisáveis por meio de um **catálogo de metadados**.



Esta abordagem garante que os dados sejam diretamente identificáveis, aplicáveis, práticos e significativos para a tarefa de IA que está sendo executada.

6 Os dados precisam ser facilmente consumíveis por MLs ou LLMs.

Já mencionamos que as aplicações de ML e IA generativa são ferramentas poderosas, mas o potencial delas reside na capacidade de consumir os dados prontamente. Diferentemente dos humanos, que conseguem ler anotações escritas à mão e planilhas caóticas, estas tecnologias exigem que as informações sejam representadas em formatos específicos. É como cozinhar para uma pessoa exigente — se ela não comer o que você preparou, vai ficar com fome. De modo similar, as iniciativas de IA não serão bem-sucedidas se os dados não estiverem nos formatos certos para os experimentos de ML ou para as aplicações de LLM. Tornar os dados facilmente consumíveis ajuda a desbloquear o potencial destes sistemas de IA, permitindo que eles ingiram as informações tranquilamente e as traduzam em ações inteligentes para gerar resultados criativos. Consequentemente, tornar os dados facilmente consumíveis é o último princípio dos dados prontos para IA.

Tornando os dados consumíveis para o aprendizado de máquina

A transformação dos dados é uma tarefa subestimada quando falamos de dados consumíveis para ML. Embora os algoritmos, como a regressão linear, sejam o centro das atenções, a qualidade e a forma dos dados nos quais eles são treinados são igualmente importantes. Além disso, os esforços investidos para limpar, organizar e tornar os dados consumíveis pelos modelos de ML geram recompensas significativas. Os dados preparados permitem que os modelos aprendam de modo eficiente, levando a previsões precisas, resultados confiáveis e, em última análise, ao sucesso de todo o projeto de ML.

No entanto, os formatos dos dados de treinamento dependem bastante da infraestrutura de ML. Os sistemas de ML tradicionais são baseados em discos. E grande parte do fluxo de trabalho dos Cientistas de Dados foca na definição de melhores práticas e procedimentos de codificação manual para lidar com grandes volumes de arquivos. Mais recentemente, os sistemas de ML baseados em lakehouse têm usado recursos de armazenagem similares a bancos de dados, enquanto o fluxo de trabalho dos Cientistas de Dados fez a transição para SQL como linguagem de primeira classe. Como resultado, estruturas de dados tabulares, bem-formadas e de alta qualidade são os formatos de dados mais consumíveis e convenientes para os sistemas de ML.

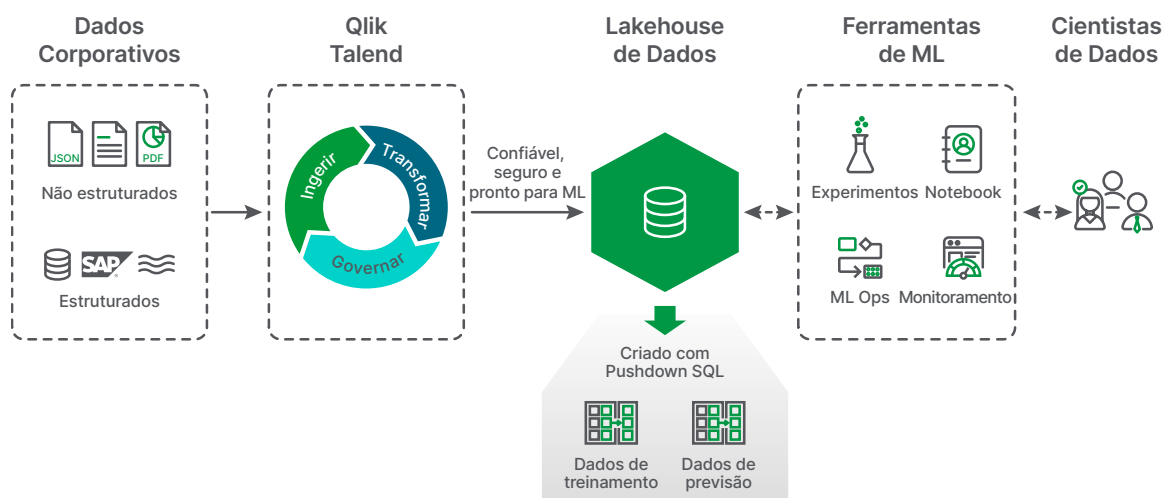


Figura 7. Tornando os dados consumíveis para ML

Tornando os dados consumíveis para a IA generativa

Large Language Models (LLMs), como GPT-4 (OpenAI), Anthropic (Claude) e LaMDA e Gemini (Google), foram pré-treinados com uma grande quantidade de dados de texto e compõem a base da IA generativa. Estima-se que modelo GPT-3, da OpenAI, tenha sido treinado com cerca de 45 TB de dados, superando 300 bilhões de tokens. Apesar desse grande volume de inputs, os LLMs não podem responder a questões específicas sobre o seu negócio, já que eles não têm acesso aos dados da sua empresa. A solução é ampliar esses modelos com as suas próprias informações, resultando em aplicações de IA mais corretas, relevantes e confiáveis.

O método para integrar os seus dados corporativos a uma aplicação baseada em LLM é chamado de 'Geração Aumentada de Recuperação', ou 'RAG'. Geralmente, a técnica usa informações de texto derivadas de fontes não estruturadas e baseadas em arquivos, como apresentações, mensagens de e-mail, documentos de texto, PDFs, transcrições, etc. Depois, o texto é dividido em partes gerenciáveis e convertido em uma representação numérica, usada pelo LLM em um processo chamado de 'embedding'. Esses 'embeddings' são posteriormente armazenados em um banco de dados vetorial, como Chroma, Pinecone e Weviate. É interessante notar que diversos fornecedores de bancos de dados tradicionais — como PostgreSQL, Redis e SingleStoreDB — também suportam vetores. Além disso, plataformas de nuvem como Databricks, Snowflake e Google BigQuery também incorporaram suporte para vetores.

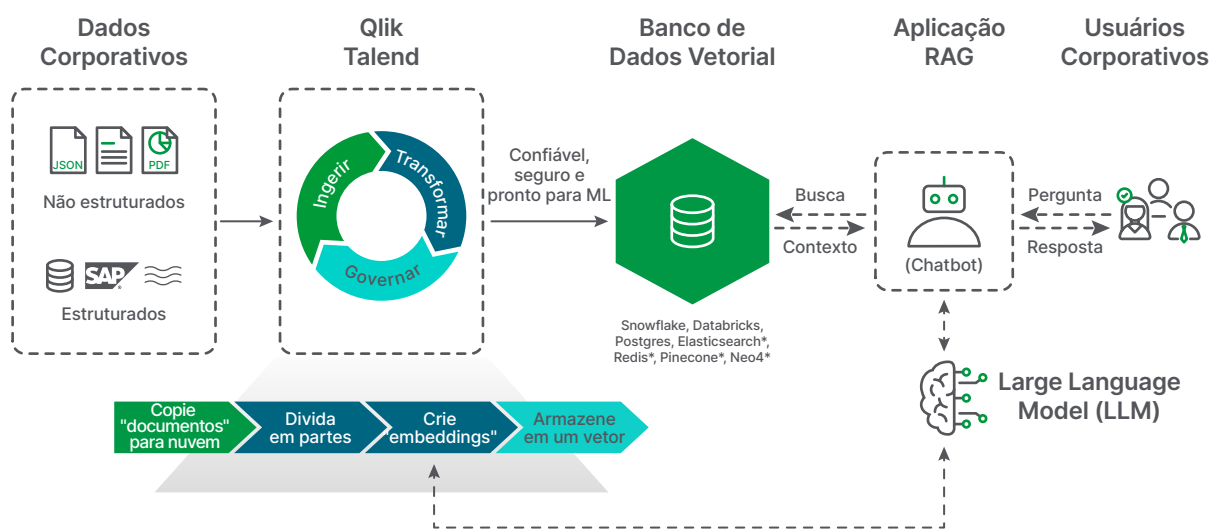


Figura 8. Tornando os dados consumíveis para a IA generativa

Não importa se os seus dados de origem são estruturados ou não estruturados: a abordagem da Qlik garante que dados de qualidade estejam prontamente consumíveis para as suas aplicações de IA generativa, RAG ou LLM.

O AI Trust Score

Após definir os seis princípios-chave de prontidão e adequação de dados, as questões que ficam são: os princípios podem ser codificados e facilmente traduzidos para uso diário? E como a prontidão de dados para IA pode ser rapidamente avaliada? Uma possibilidade é utilizar o AI Trust Score, da Qlik, como um indicador global e compreensível de prontidão.

O AI Trust Score atribui uma dimensão separada para cada princípio e, em seguida, agrega cada valor para criar um score composto, criando um atalho rápido e confiável para você avaliar a prontidão dos seus dados para IA. Além disso, como os dados da empresa mudam continuamente, o score é regularmente verificado e recalibrado para rastrear as tendências de prontidão dos dados.

Composição AI Trust Score: 4.8

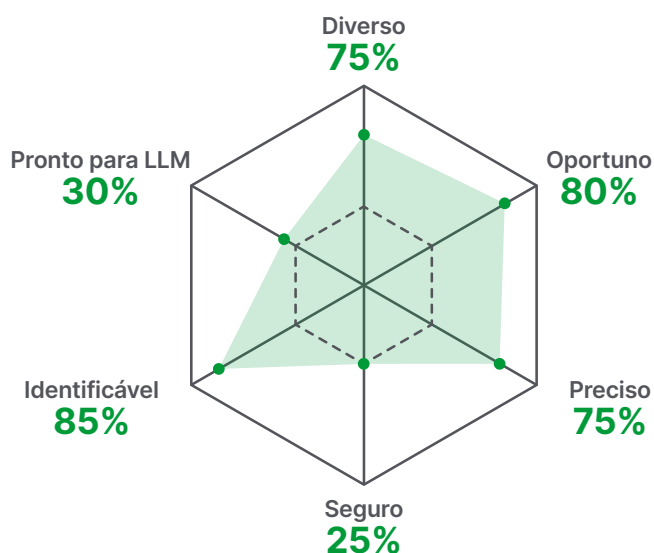


Figura 9. O AI Trust Score para prontidão de dados

O seu AI Trust Score agrega diversas métricas em um score de prontidão único e fácil de entender.

Base de dados Qlik Talend para IA

A necessidade de dados de alta qualidade em tempo real, capazes de promover decisões mais embasadas, eficiência operacional e inovação nos negócios, nunca foi tão grande. É por isso que as empresas bem-sucedidas procuram as soluções de qualidade e integração de dados líderes de mercado da Qlik Talend, com o objetivo de entregar dados confiáveis a warehouses, lakes e outras plataformas de dados corporativas. Abrangentes e de ponta, as nossas soluções usam pipelines automatizados, transformações inteligentes e datasets de qualidade confiáveis para oferecer a agilidade que os profissionais de dados querem e a governança e o compliance que as empresas esperam.

Portanto, se você está criando warehouses ou lakes para analytics perspicaz, modernizando as infraestruturas operacionais de dados para a eficiência dos negócios ou usando dados em múltiplas nuvens para iniciativas de inteligência artificial, a Qlik Talend pode mostrar o caminho a seguir.

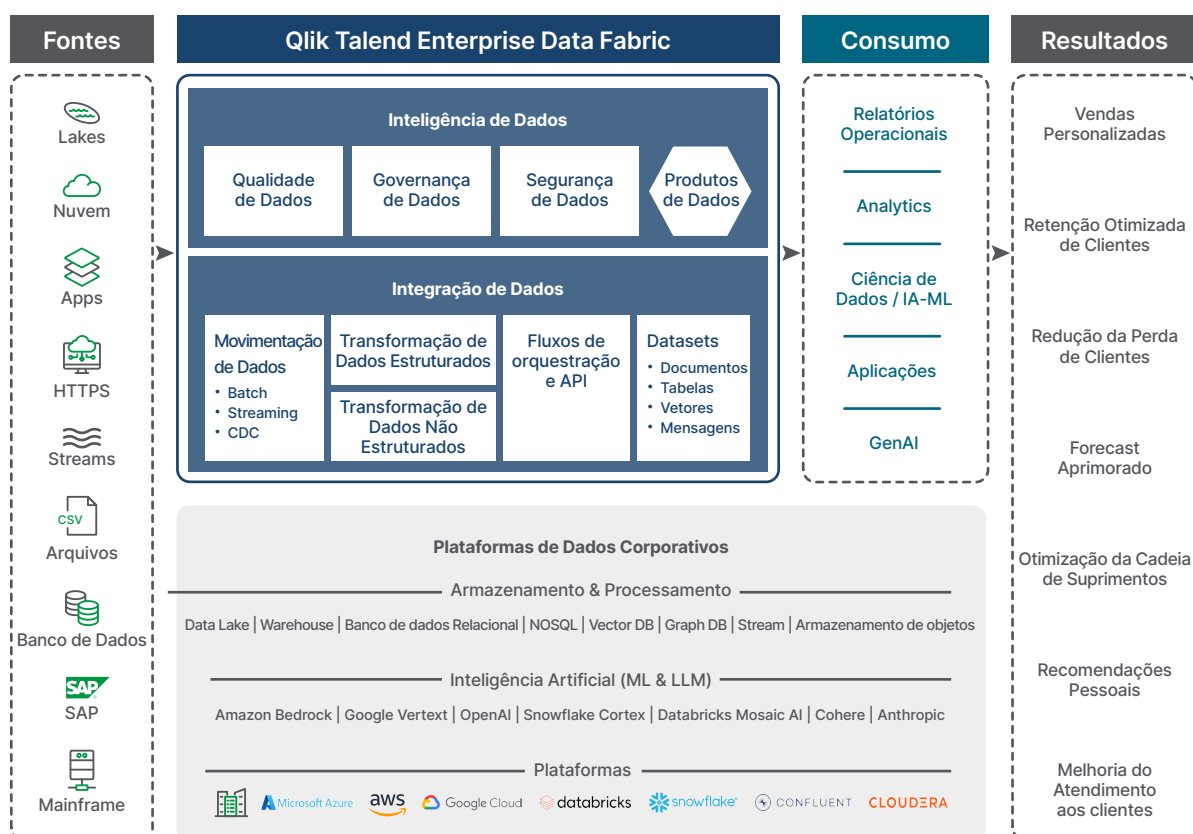


Figura 10. Qlik Talend Enterprise Data Fabric para IA e Analytics

Conclusão

Apesar do poder transformador do aprendizado de máquina e do imenso potencial de crescimento da IA generativa, a prontidão dos dados ainda é a base de uma implementação de IA bem-sucedida. Este paper apresentou os seis princípios-chave para estabelecer uma base de dados robusta e confiável, capaz de ajudar a sua empresa a desbloquear o verdadeiro potencial da IA.

Quer desbloquear todo o potencial de IA da sua organização?

[Comece aqui](#)



Sobre a Qlik

A Qlik transforma cenários de dados complexos em insights acionáveis, impulsionando resultados estratégicos de negócios. Atendendo a mais de 40.000 clientes globais, o portfólio da Qlik utiliza IA/ML avançada de nível empresarial e qualidade de dados abrangente. A Qlik é referência em integração e governança de dados, oferecendo soluções completas que funcionam com diversas fontes de dados. As análises intuitivas e em tempo real da Qlik revelam padrões ocultos, capacitando as equipes a enfrentar desafios complexos e aproveitar novas oportunidades. As ferramentas de IA/ML da Qlik, práticas e escaláveis, levam a decisões melhores e mais rápidas. A Qlik é uma parceira estratégica e sua experiência e plataforma com tecnologia agnóstica tornam seus clientes mais competitivos.

qlik.com