

QUALIDADE E GOVERNANÇA DE DADOS

Maximizando o valor dos dados na era da IA

Os produtos de dados aceleram os resultados
e aprimoram a governança

No atual cenário acelerado dos negócios, os dados são ativos estratégicos — vitais para otimizar as operações, impulsionar a inovação e o crescimento. Além de reunir números e estatísticas, os dados oferecem insights, tendências e padrões que podem apontar o caminho para o sucesso. Agora, mais do que nunca, com o destaque da IA generativa e do aprendizado de máquina, as empresas estão lidando com uma pergunta fundamental:

De que forma elas podem desbloquear todo o valor dos dados para acelerar os resultados de negócios, enquanto aprimoram a confiança e a governança dos dados?

A velocidade é um elemento-chave nesta equação. Segundo uma pesquisa recente da IDC, 75% dos tomadores de decisão acreditam que os dados perdem seu valor em questão de dias.¹

O desafio vai além de gerenciar volumes imensos de dados. Inclui refinar as informações para que virem ativos de alta qualidade, com curadoria e prontamente acessíveis, adequados a domínios de negócios específicos. Porém, como quem trabalha com dados já sabe, isso é mais difícil do que parece. Os líderes de dados e análise estão lidando com diversos desafios que impedem a realização das iniciativas de dados:

- **O gerenciamento de dados orientado a projetos carece de escalabilidade.**

Cada equipe de projeto tem seus próprios parâmetros, incluindo escopo, objetivos e stakeholders. Esta abordagem fragmentada impede a colaboração entre áreas e limita a extração de valor dos dados. E há diversos riscos: desalinhamento dos stakeholders nos projetos, trabalho duplicado, lacunas de governança, utilização ineficiente dos dados, aumento de custos e, em última análise, atrasos ou resultados de negócios não atingidos.

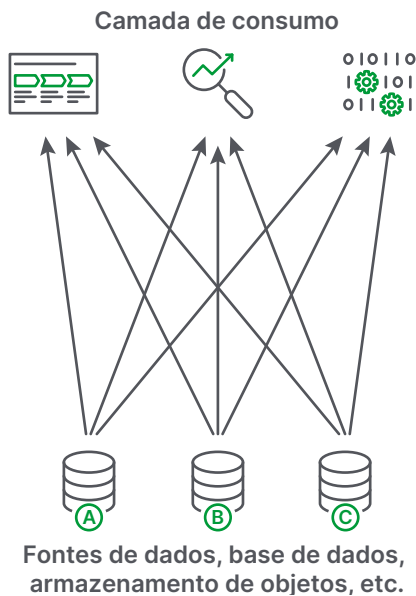
- Em todas as áreas da empresa, **há falta de consistência na qualidade dos dados entre os sistemas**, minando a tomada de decisões e a confiança nos insights gerados pelos dados.
- Em toda a cadeia de suprimento de dados e de valor, **há falta de clareza sobre a propriedade e responsabilidade** de consumidores e produtores de dados, frustrando a tentativa de extrair seu valor na velocidade dos negócios.
- **Depender de soluções fragmentadas e abordagens ad-hoc e do tipo "faça você mesmo"** de gerenciamento de dados atrasa a extração de valor. E sem profissionais treinados, os custos de escalabilidade e manutenção ficam altos.

Esses desafios ampliam a distância entre consumidores e produtores de dados. De um lado, os produtores de dados — com foco em TI — priorizam as plataformas de dados em vez de entender o uso deles e as necessidades dos consumidores. Por outro lado, os consumidores de dados, focados em suas atividades de negócios, têm dificuldades para transmitir suas necessidades e não confiam nos dados fornecidos. **Para preencher essa lacuna cada vez maior, as empresas precisam de uma abordagem mais simples e estratégica de gerenciamento de dados.**

¹ Blog da IDC, "Navigating the Planes of Enterprise Intelligence Architecture" - Junho/2023.

Mudando a abordagem do gerenciamento de dados

Tradicionalmente, as empresas têm empregado duas abordagens distintas para extrair valor dos dados: uma **fragmentada** e uma **centralizada**.



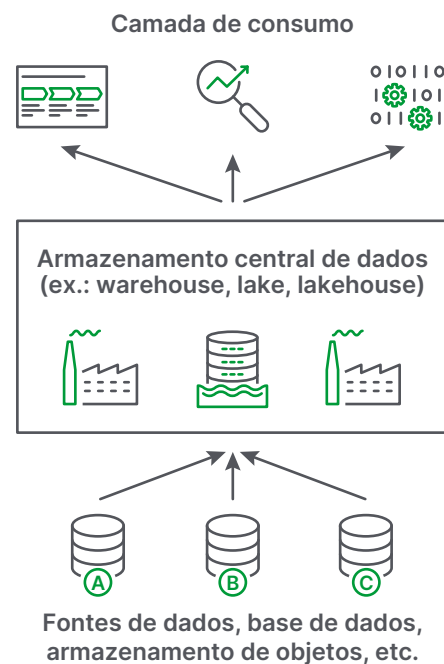
A ABORDAGEM FRAGMENTADA

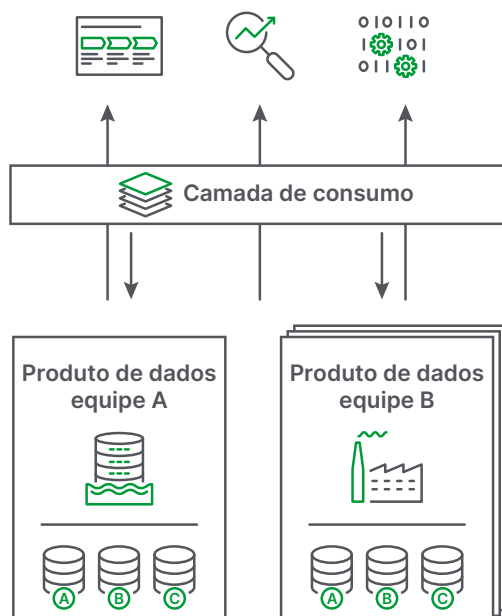
Na abordagem fragmentada, cada equipe de "projeto de dados" reúne dados de diversas fontes para atender casos de uso específicos. Cada equipe consome dados de modo independente, usando as próprias ferramentas e tecnologias para extrair, integrar e analisar os dados para as suas necessidades específicas. Embora esse método gere valor a curto prazo, ele não é escalável para empresas grandes a longo prazo, resultando em mais silos de dados, esforços duplicados, aumento de custos e atrasos nos resultados dos negócios.

A ABORDAGEM CENTRALIZADA

Na abordagem centralizada, uma única equipe de dados é responsável por extrair, limpar e agregar os dados em larga escala. O objetivo dessa equipe central é abordar as exigências analíticas de toda a empresa de forma abrangente. A equipe cria pipelines de dados elaborados para obter dados a partir de vários sistemas, limpá-los e consolidá-los em data lakes baseados em nuvem, warehouses ou lakehouses (como Databricks Delta Lake, Snowflake, BigQuery, S3, etc.).

A centralização pode parecer a solução para a fragmentação - e até certo ponto, é - mas ela tem um problema sério: como exige uma ótima coordenação entre os departamentos para fazer as mudanças ou as atualizações necessárias, ela dificulta a agilidade organizacional. Além disso, ao focar em exigências de dados comuns em toda a empresa, a equipe de dados central não percebe as necessidades específicas e as preferências dos stakeholders e consumidores, resultando em uma solução incompleta e que não atende às exigências do negócio.





A ABORDAGEM FEDERADA

Estas duas abordagens tradicionais — tanto a fragmentada quanto a centralizada — têm limitações e não atendem às necessidades de consumo de dados das empresas. No entanto, para superar esses desafios, uma quebra de paradigma está ocorrendo. Cada vez mais popular, a **abordagem federada** distribui estrategicamente a entrega dos dados, utilizando os produtos de dados e gerenciando de forma centralizada as funções de engenharia da plataforma.

Diversas equipes de produtos de dados podem trabalhar em paralelo, e cada uma foca em seus produtos de dados específicos do domínio. Essa abordagem acelera a obtenção dos resultados ao estimular a colaboração entre as áreas, incluindo stakeholders técnicos e de negócios.

Além disso, a abordagem federada oferece várias vantagens importantes:



Agilidade e flexibilidade

Ao simplificar o gerenciamento de dados de ponta a ponta, a abordagem federada permite que as empresas **desbloqueiem rapidamente o valor** dos dados para os negócios. Essa abordagem promove agilidade ao adotar uma mentalidade voltada para o produto, em que a entrega e a criação de valor são distribuídas entre equipes de produtos de diversas áreas, enquanto as funções de engenharia da plataforma são executadas de forma centralizada. Isso possibilita **respostas mais rápidas para as necessidades em evolução do negócio**, garantindo a integridade dos dados e as responsabilidades ao longo do processo.



Custo-benefício

O modelo federado se baseia na **reusabilidade**, possibilitando que as empresas maximizem seus investimentos e reduzam os custos operacionais. Além disso, o compartilhamento central de componentes fundamentais entre as diversas equipes de dados **minimiza a execução de atividades redundantes e facilita a manutenção a longo prazo**.



Entrega acelerada de caso de uso

A colaboração entre equipes de diversas áreas gera **resultados rápidos e específicos do domínio**, personalizados para as necessidades em evolução do negócio e às novas iniciativas orientadas por dados. A abordagem federada pode ser **escalável e permitir a customização de casos de uso específicos do domínio**, por meio de desenvolvimento de streams paralelos entre as diversas equipes de produtos de dados.



Mais confiança

O modelo federado **promove confiança com governança, transparência e conformidade aprimorados**, fomentando a segurança na tomada de decisões orientada por dados. A propriedade clara do domínio de dados promove o alinhamento e a responsabilidade entre os stakeholders analíticos e de dados, fortalecendo a estrutura de governança de dados da empresa.

Apresentando os produtos de dados

A abordagem federada é o ponto de partida mais adequado para os produtos de dados já que, para muitas empresas, ela significa uma mudança evolutiva — do controle à colaboração, do projeto ao produto, da rigidez à agilidade —, enquanto entrega valor em escala. **Data Mesh** e **Data Fabric**, distintos em suas abordagens em relação ao gerenciamento de dados, podem formar uma sinergia poderosa para desbloquear todo o potencial de produtos de dados. No paradigma de Data Mesh, a propriedade dos dados é descentralizada — ela pertence e é gerenciada pelas equipes que utilizam os dados. O produto de dados é o aspecto mais tangível e impactante do paradigma Data Mesh. Já o Data Fabric é a arquitetura de base de gerenciamento de dados que permite a entrega precisa dos produtos de dados às equipes de domínio.²



No Gartner CDAO Agenda Survey de 2024, 50% das empresas que participaram disseram já ter implementado produtos de dados, e outras 29% estão comprometidas em fazer um piloto ou considerar a implementação no próximo ano.³

Então, o que é um produto de dados? Um produto de dados entrega datasets de alta qualidade, com curadoria, prontos para o uso e preparados para IA, que as pessoas podem acessar e aplicar facilmente a diversos casos de uso em toda a empresa. Um produto de dados pode ser visto como um microserviço para os dados. Assim como os microserviços expõem os recursos de uma aplicação por meio de unidades menores, mais contextuais e combináveis, o produto de dados fornece acesso a dados relevantes e orientados ao domínio por meio de interfaces de serviço. Um produto de dados segue diversos princípios fundamentais:



Orientação e propriedade de domínio

Cada produto de dados deve se alinhar a um domínio específico e ter um proprietário de dado (pessoa ou equipe) responsável pelo seu sucesso — a entrega de valor e a adoção abrangente, por exemplo.



Product thinking

Os produtos de dados são criados com base no conceito de 'product thinking', que analisa a propriedade, o versionamento e os clientes (neste caso, os consumidores de dados). Os produtos são aprimorados por meio da compreensão dos padrões de uso e da obtenção de feedback dos clientes. Eles têm um ciclo de vida, assim como os produtos físicos.



Padronização e interoperabilidade

Os produtos de dados seguem convenções de metadados, formatos e schemas padronizados. Eles facilitam a interoperabilidade e a integração com outros componentes do ecossistema de dados.



Localizável e identificável

Os produtos de dados são registrados em um catálogo para que possam ser facilmente encontrados e identificados pelos consumidores de dados. Um produto de dados se autodescreve, seja os dados que ele representa, como pode ser acessado, sua semântica de negócio, sua linhagem e sua qualidade.



Reusabilidade

Os produtos de dados são criados uma vez e usados diversas vezes em vários casos de uso. Outros produtos de dados são construídos usando opções já existentes.

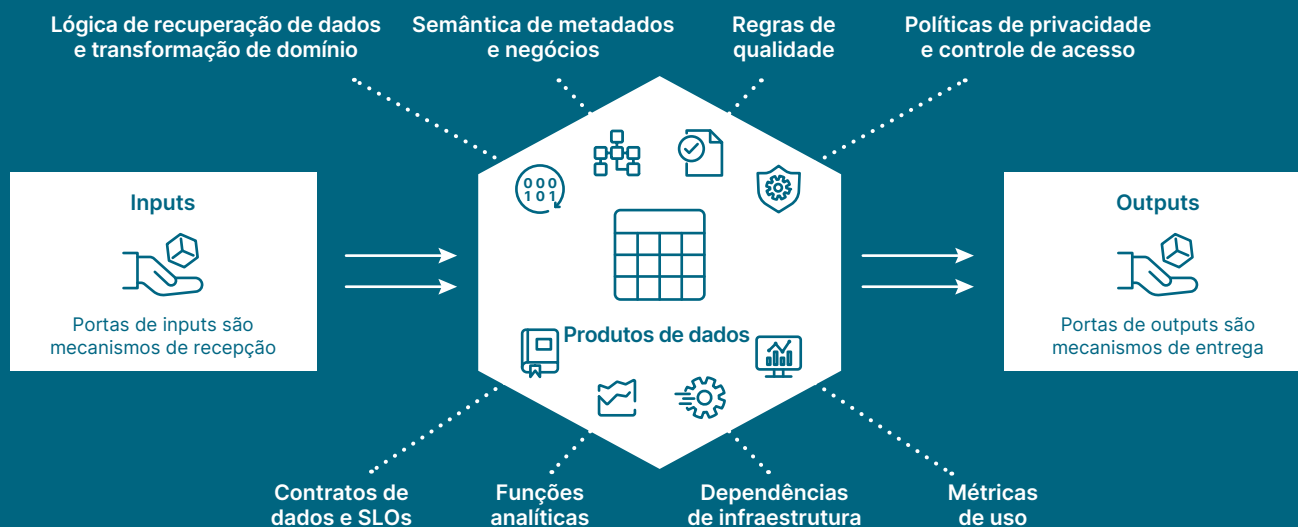


Acessibilidade

Os produtos de dados são acessados por meio de um conjunto padrão de interfaces, como SQL ou REST API, que são definidas e exibidas quando os produtos de dados são criados. As interfaces ocultam a complexidade presente no produto de dados.

² Gartner: '6 Lessons Data Leaders Can Learn from Adopters of Data Mesh by Robert Thanaraj', por Robert Thanaraj, et al., 4 de março de 2024. GARTNER é uma marca comercial e de serviço registrada do Gartner, Inc. e/ou de suas afiliadas nos EUA e internacionalmente, usada aqui com permissão. Todos os direitos reservados.

³ Gartner: 'Gartner Chief Data and Analytics Officer Agenda Survey for 2024', 12 de março de 2024. Os resultados desse estudo não representam o cenário global ou do mercado como um todo, mas refletem o sentimento dos entrevistados e das empresas pesquisadas. GARTNER é uma marca comercial e de serviço registrada do Gartner, Inc. e/ou de suas afiliadas nos EUA e internacionalmente, usada aqui com permissão. Todos os direitos reservados.



Um produto de dados é um conjunto autônomo que reúne dados, transformações, códigos, políticas de governança, SLAs, padrões de acesso e dependências de infraestrutura. O diagrama acima mostra a anatomia do produto de dados.

1 Dataset: existe no centro do produto. Ele pode ser um conjunto de tabelas, arquivos, gráficos ou representações vetoriais. Pode haver diversas camadas de dados dentro de um produto de dados — um conjunto (set) bruto, que representa os dados como eles foram ingeridos a partir de um sistema-fonte; um conjunto com curadoria, que representa dados limpos e integrados; e um dataset pronto para consumo, modelado para facilitar o uso.

2 Inputs: representa as portas de input ou a fonte de dados. Podem ser tabelas em uma aplicação SAP que precisam ser ingeridas; tabelas no Snowflake que precisam ser usadas como fonte; outro produto de dados do qual um novo está sendo derivado ou uma fila de mensagens para receber feeds de dados em tempo real.

3 Outputs: representa os mecanismos de entrega de dados para consumo. Um produto de dados pode exibir uma grande variedade de interfaces padronizadas, como SQL, REST API, busca vetorial, mensagens e arquivos. As portas de output também podem variar com base no tipo de serviço que o produto de dados oferece. Um output pode oferecer acesso a dados históricos, enquanto o outro pode oferecer acesso ao último snapshot de dados.

4 Lógica de recuperação de dados e transformação de domínio: representa a lógica e o código de pipeline para obter os dados (ex.: quando os dados são movidos de um banco de dados de origem para uma plataforma na nuvem ou quando é feita a leitura de um conjunto de tabelas). Posteriormente, a lógica de transformação é aplicada para moldar os dados do formato-fonte para o formato-alvo de dataset, que pode ser na forma de tabelas, visualizações ou embeddings vetoriais.

5 Semântica de metadados e negócios: contém os metadados técnicos (nomes de tabela, nomes de arquivo e definições de coluna) e metadados de negócios (domínio, nome do produto, tipo de produto, propósito, propriedade, tag, etc.), para permitir um melhor entendimento do produto de dados.

6 Regras de qualidade de dados: definem as diversas verificações de qualidade de dados que precisam ser executadas no produto de dados. Cada regra tem uma informação de configuração sobre quando executar, o seu escopo, se deve ser executada em todos os dados ou de modo incremental, etc. As regras de qualidade de dados serão executadas no cronograma prescrito para garantir a precisão dos dados e a confiabilidade.

7 Políticas de privacidade e controle de acesso: estabelecem as políticas que controlam o acesso ao dataset dentro do produto de dados. Por exemplo, se o dataset contém informações de identificação pessoal (PII), uma política de privacidade pode mascarar isso para todos os usuários, ou apenas determinados grupos de usuários podem ter acesso.

8 Contratos de dados e SLOs: definem os valores-alvo para o objetivo do serviço que o produto de dados deve executar. Por exemplo: um SLO pode exigir que uma métrica de precisão de qualidade de dados seja maior do que 95%, ou que o os dados sejam atualizados a cada cinco minutos.

9 Funções analíticas: reúnem a lógica de negócios personalizada adicional para tornar a função reutilizável. Uma função em um produto de dados de pedido de venda pode se referir a pedidos que não foram realizados ou entregas da forma correta. Ela irá reunir a lógica complexa de domínio para determinar quais são esses pedidos. As funções analíticas podem ser exibidas via uma porta de output, como uma API.

10 Dependências de infraestrutura: as informações sobre os requerimentos de infraestrutura para os datasets associados ao produto de dados. Por exemplo: os datasets são armazenados em uma plataforma de dados (como Snowflake ou Databricks), com o código associado (ex.: lógica de transformação), exigindo um ambiente computacional (ex.: SQL) para executar.

Ciclo de vida do produto de dados

Os produtos de dados não são criações únicas. Assim como os produtos físicos, eles exigem gerenciamento e monitoramento contínuos. Cada produto de dados é de propriedade e gerido por uma equipe de dados alinhada ao domínio, responsável por entregar valor, ampliar a adoção e manter do ciclo de vida dele.



Veja os estágios-chave do ciclo de vida do produto de dados:

- 1 Conceitualizar:** o processo começa ao identificar as necessidades de negócio específicas do domínio, como entender os usuários, seus casos de uso e como eles costumam consumir os dados.
- 2 Projetar:** a próxima etapa é moldar a solução, o que inclui identificar os datasets necessários e suas fontes, além de definir as características do produto de dados por funcionalidade, interatividade, abordagem de apresentação, etc.
- 3 Construir:** após projetar, é hora de construir a infraestrutura técnica e as funcionalidades do produto de dados. Isso inclui reunir e integrar datasets brutos a partir das suas fontes, garantir a qualidade dos dados e transformá-los.
- 4 Disponibilizar:** esse é o 'estágio da acessibilidade'. Ele envolve concluir a documentação do produto, atribuir a propriedade, fazer a integração com os sistemas de negócio existentes ou fluxos de consumo para acesso contínuo, e ativar o produto de dados ou torná-lo pronto para consumo.
- 5 Promover:** após a ativação, é importante disseminar e promover o uso. Essa etapa envolve treinar os consumidores e definir canais de comunicação para receber feedback.
- 6 Usar:** após um lançamento bem-sucedido, os consumidores vão começar a interagir com o produto de dados para buscar, encontrar, entender e ter acesso. O foco desse estágio é monitorar o uso e reunir feedback. Geralmente, os dados coletados se referem a como os usuários interagem com o produto de dados, o desempenho e se as necessidades centradas no domínio da empresa estão sendo atendidas.
- 7 Manter:** os produtos de dados não são entidades estáticas; eles precisam ser aprimorados para manterem seu valor. Com base no feedback coletado dos usuários e nas necessidades mutáveis do negócio, novas dimensões de dados, recursos e funcionalidades são propostas para a próxima versão do produto de dados.

Ao seguir esses estágios de ciclo de vida, as empresas podem criar produtos de dados capazes de promover insights e tomadas de decisão baseadas em dados e atingir os resultados de negócios esperados.

O que um produto de dados NÃO É

O mundo dos dados é cheio de termos, e o uso deles nem sempre é consistente. Um bom exemplo: o termo "produto de dados" é frequentemente confundido com termos como "dataset" ou "catálogo de dados", que não são sinônimos. Esta seção destaca algumas diferenças fundamentais entre esses conceitos.



Embora o dataset seja um componente fundamental de diversos produtos de dados, é crucial entender que ele é apenas um componente. Um produto de dados vai muito além do dataset bruto.

DATASETS: O MATERIAL BRUTO

Geralmente, um dataset é um grupo de pontos de dados organizados, sejam eles estruturados ou não estruturados. Aqui estão incluídas informações brutas como dados dos clientes, registros financeiros e leituras de sensores. Pense nos datasets como materiais brutos — pilhas de madeira, por exemplo —, que podem se tornar valiosos, mas precisam ser refinados antes. Uma das coisas mais importantes que diferenciam um dataset de um produto de dados é a falta de contexto de negócio específico do domínio.

PRODUTOS DE DADOS: OS BENS ACABADOS

Os produtos de dados utilizam datasets meticulosamente selecionados e processados, frequentemente combinados com artefatos adicionais (abordados na seção anterior) para entregar valor. Indo além na metáfora da pilha de madeira: pense nos produtos de dados como móveis sofisticados, feitos com essa madeira (ex.: os datasets).

Veja o que diferencia os produtos de dados dos datasets:

- **Contexto e centralidade do domínio:** os produtos de dados são customizados para abordar necessidades específicas do negócio ou do usuário. Eles transcendem os dados brutos e incluem informações contextuais como definições, linhagem e verificações de qualidade.
- **Usabilidade e acessibilidade:** os produtos de dados são feitos para aumentar o valor e facilitar o consumo deles. Eles são preparados e embalados para estarem prontamente consumíveis pelos consumidores de dados de domínios específicos, com expertise técnica limitada.
- **Insights acionáveis:** os produtos de dados não apresentam apenas dados. Eles oferecem insights operacionais, analíticos, preditivos e prescritivos. Esses insights podem ser entregues por meio de dashboards, modelos analíticos pré-construídos ou ganchos de integração com outras ferramentas via SQL ou APIs.
- **Propriedade e responsabilidade:** os produtos de dados têm linhas claras de propriedade, garantindo a responsabilidade por diversos aspectos do desenvolvimento, consumo e feedback dentro da organização. A propriedade dos produtos de dados pode depender do contexto do negócio, com o proprietário sendo o produtor que gera o produto de dados, o consumidor dentro da equipe do domínio responsável pelos casos de uso relevantes, ou um cargo recém-criado: o Gerente de Produto de Dados.



Os produtos de dados transformam datasets em ativos valiosos e consumíveis, que promovem decisões informadas e o crescimento do negócio. A Harvard Business Review estima que “as empresas que tratam os dados como um produto reduzem o tempo de implementação deles em novos casos de uso em até 90%, diminuem os custos totais de propriedade (tecnologia, desenvolvimento e manutenção) em até 30% e reduzem os riscos e o ônus da governança de dados.”⁴

⁴ “A Better Way to Put Your Data to Work”, Harvard Business Review, Julho-Agosto/2022.

CATÁLOGO DE PRODUTO DE DADOS: O HUB DOS BENS ACABADOS

Um catálogo de produto de dados serve como um repositório abrangente que abriga produtos de dados e metadados associados — descrições, tags, versão, rótulos de propriedade, etc. — para auxiliar na compreensão e na descoberta de produtos de dados. Ele atende tanto os consumidores quanto os produtores de dados, oferecendo soluções customizadas para as necessidades deles. Para os Gerentes de Produto de Dados, o catálogo oferece ferramentas para configurar, ativar e gerenciar todo o ciclo de vida de um produto de dados, da criação à ativação e ao fim do ciclo de vida. Para os consumidores, o catálogo serve como um marketplace, que permite uma 'experiência de compra' promovendo a utilização.

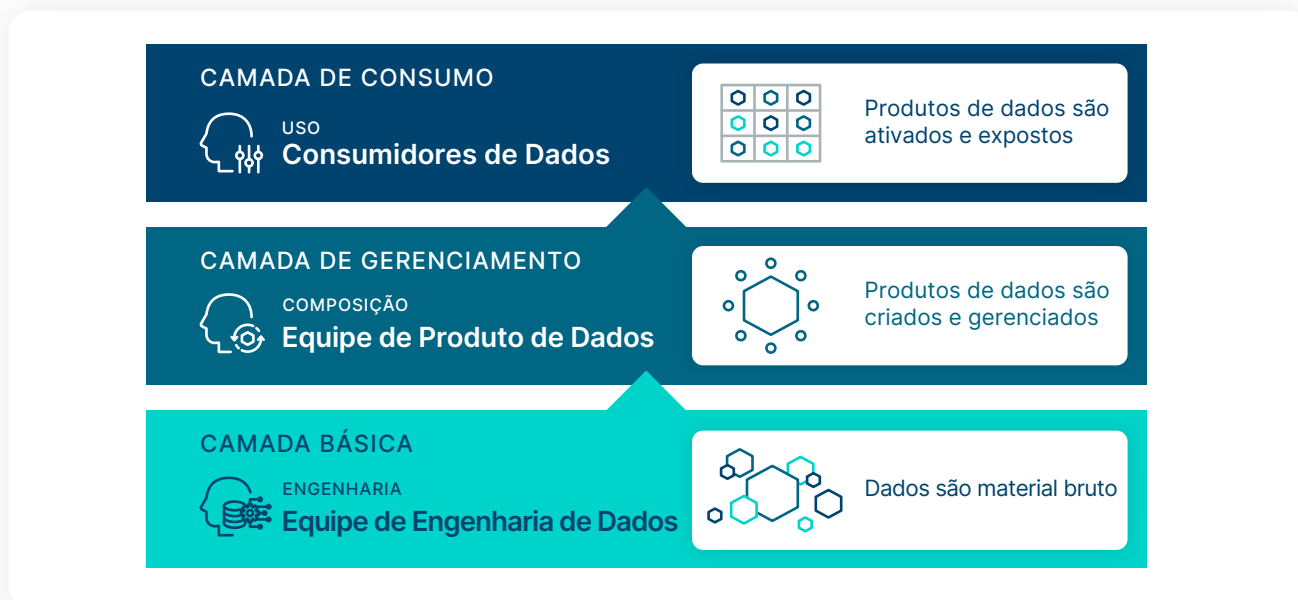
Veja o que diferencia um catálogo de produto de dados de um catálogo de dados:

RECURSO	CATÁLOGO DE DADOS	CATÁLOGO DE PRODUTO DE DADOS
Missão	Fortalecer a governança de dados	Aumentar o valor dos dados
Escopo	Foco na empresa	Focado no domínio
Modelo de gerenciamento	Centralizado	Federado
Propósito	Ser identificável, conformidade, transparência	Usabilidade, confiabilidade, acionabilidade
Conteúdo	Metadados de diversos tipos, incluindo técnicos, operacionais e de negócios, relacionados a todos os ativos de dados	Produtos de dados confiáveis, metadados de ciclo de vida de produtos de dados, documentação de produtos de dados e propriedade
Entidades	Ordem de centenas de milhares	Ordem de centenas
Público	Principalmente governança de dados	Consumidores e produtores de dados

Na área de gerenciamento de dados, o desafio é não apenas criar plataformas de dados, mas garantir a adoção eficaz delas. Frequentemente, as empresas investem recursos substanciais — milhões de dólares — para construir infraestruturas de dados, mas encontram um problema sério: **barreiras de compartilhamento de dados entre produtores e consumidores**. Por quê? Pois os resultados gerados pelas diversas plataformas de tecnologia carecem de confiabilidade e não se alinham aos objetivos de negócios específicos dos consumidores. **Em resumo: na sua empresa, você precisa tanto de um catálogo de dados quanto de um catálogo de produto de dados.** Pense neles como dois corredores distintos em um supermercado — no primeiro, há materiais brutos, como farinha, manteiga e açúcar; no segundo, há itens acabados, como bolos e doces.

Catálogo de produto de dados: camadas e players

O catálogo de produto de dados é o principal hub para a organização de produtos de dados dentro de uma empresa. Ele é estruturado em três camadas, cada uma customizada para um conjunto diferente de stakeholders. Sejam Engenheiros de Dados, Supervisores, Analistas ou novos cargos (como o de Gerente de Produto de Dados), cada grupo interage de forma única com essas camadas, de acordo com seus objetivos e exigências. Essa abordagem garante a separação de interesses com base nas necessidades do usuário, com o nível adequado de abstração para facilitar o consumo e a adoção.



A CAMADA DE BASE

A base do catálogo de produto de dados é uma camada fundamental, que estabelece a infraestrutura técnica e entrega os recursos que vão dar suporte às camadas subsequentes.

Os principais recursos dessa camada são:

- **Engenharia de dados:** mover e processar dados de várias fontes, incluindo arquivos, warehouses de dados na nuvem, data lakes e streams de dados em tempo real.
- **Gerenciamento de metadados:** coletar e organizar informações valiosas sobre os dados, com linhagem, dicionários, schemas, etc.
- **Segurança e controle de acesso:** promover controles de segurança para proteger os dados e garantir que apenas usuários autorizados tenham acesso a eles.
- **Qualidade dos dados e rastreamento da linhagem:** monitorar a qualidade dos dados para garantir a precisão e o rastreamento da linhagem dos dados em nome da confiabilidade.

A camada de base serve principalmente a equipe da plataforma de dados, incluindo cargos técnicos como Arquitetos, Supervisores e Engenheiros de Dados. A meta é fazer com que os recursos nesta camada sejam 'self-service', para que usuários não técnicos consigam utilizá-los com facilidade. Os recursos exigidos nesta camada de base podem ser encontrados em diversas ferramentas de gerenciamento de metadados e de ingestão, transformação, qualidade e segurança de dados.

A CAMADA DE GERENCIAMENTO

A camada de gerenciamento de produto de dados — um novo conceito — conta com recursos que atendem às necessidades do papel emergente de Gerente de Produto de Dados (ou da equipe). Essa camada oferece recursos para desenvolver, gerenciar e manter produtos de dados confiáveis e orientados ao domínio. Ela se ampara nos serviços da camada de base para inventariar, mover, transformar e proteger os ativos de dados. Ela serve como uma ponte entre os metadados brutos que residem na camada de base e os insights acionáveis que os consumidores buscam nos seus dados.

Os principais recursos dessa camada são:

- **Composição de produto de dados:** recursos para construir um produto de dados, definir os inputs e outputs e configurar todos os seus diversos artefatos.
- **Implementação de produto de dados:** gerenciamento de versões e fluxo de trabalho para ativar um produto de dados quando ele estiver pronto para o consumo dos usuários finais.
- **Manutenção do produto de dados:** rastreamento do uso, coleta de feedback, mudanças, atualizações e captura de todos os estágios do ciclo de vida do produto de dados conforme descrito nas seções anteriores.

A camada de gerenciamento de produto de dados serve principalmente a equipe de gerenciamento de produto, especialmente os Gerentes de Produto de Dados, que cuidam do ciclo de vida dos produtos de dados para alcançar os objetivos do negócio.

A CAMADA DE CONSUMO (OU O 'MARKETPLACE')

A camada de marketplace fica no topo do catálogo de produto de dados, representando a parte do sistema voltada para o consumidor. Ela permite a exploração e o consumo de dados self-service, estimulando uma cultura orientada por dados dentro da empresa.

Os principais recursos nessa camada incluem:

- **Busca e identificação:** interfaces intuitivas permitem aos usuários explorar e identificar os produtos de dados que melhor atendem às suas necessidades.
- **Aprendizado e confiança:** os consumidores de dados podem entender o produto de dados visualizando tanto os metadados quanto as amostras de dados, aprender como usá-los visualizando uma amostra de código e a documentação dos dados, e passar a confiar neles por meio da revisão da linhagem e da qualidade dos dados.
- **Acessos e permissões:** esta camada controla quem pode acessar produtos de dados específicos, garantindo a segurança e a conformidade com as políticas de governança de dados.
- **Interfaces de consumo de dados:** APIs, ferramentas de visualização ou mesmo pontos de acesso direto ao banco de dados fornecem aos usuários os meios para consumir os produtos de dados. Dependendo do produto, diferentes interfaces podem estar disponíveis para atender às inúmeras necessidades e habilidades dos usuários.

A camada de consumo de dados serve principalmente analistas de negócio, cientistas de dados, usuários dos sistemas e outros tomadores de decisão de negócios importantes.

BENEFÍCIOS DOS PRODUTOS DE DADOS PARA OS DIFERENTES PLAYERS NO ECOSISTEMA DE DADOS

Os dados são os ativos mais importantes de uma empresa. Mas o verdadeiro potencial deles só pode ser explorado se estiverem prontamente acessíveis e forem bem compreendidos e usados de modo eficiente. Os produtos de dados preenchem essas lacunas. Eles transformam dados brutos em ativos consumíveis e com curadoria, disponíveis aos stakeholders de todas as áreas da empresa.

Vamos analisar como os produtos de dados impactam os players-chave no ecossistema de dados:



Para as **equipes de Engenharia de Dados**, os produtos de dados são uma linguagem comum — um acordo bem definido —, que trazem estrutura e facilitam a comunicação entre consumidores e produtores de dados. Isso permite aos produtores desenvolver produtos de dados reutilizáveis de modo iterativo, reduzindo o custo total de propriedade (TCO). Além disso, a equipe de Engenharia de Dados não precisa construir pipelines de dados grandes, complexos e centralizados. Elas podem desenvolver pipelines menores e independentes, dedicados à implementação de cada produto.



Para as **equipes de Governança de Dados**, os produtos de dados são um ponto central de auditoria — e como os produtos são confiáveis e governados, é mais simples minimizar o risco. A equipe de Governança define centralmente as políticas e padrões, enquanto as equipes de Produtos de Dados implementam as políticas de maneira federada.



Para os **Gerentes de Produto de Dados**, os produtos de dados alinham consumidores e produtores, promovendo a colaboração e a responsabilidade em todo o ecossistema de dados. Isso garante propriedade e responsabilidade aos dados.



Para os **consumidores de dados**, os produtos de dados são ativos com curadoria e centrados no domínio, que entregam exatamente o que é preciso para maximizar o valor do negócio. Isso reduz o retrabalho e acelera os resultados.

Juntos, esses stakeholders podem utilizar um catálogo de produto de dados para explorar todo o valor dos seus ativos de dados de modo eficiente, promovendo decisões orientadas por dados e atingindo os objetivos do negócio.

“No segmento de hipotecas, os produtos de dados vão muito além de números e algoritmos — nós os enxergamos como agentes de transformação. Ao definir produtos de dados self-service, queremos empoderar nossos agentes de créditos a oferecer a melhor solução para o cliente certo, permitindo que as equipes de Finanças possam avaliar melhor os custos e os riscos. Explorar os produtos de dados em nosso negócio está realmente redefinindo o futuro das hipotecas para nós.”

Julia Fryk

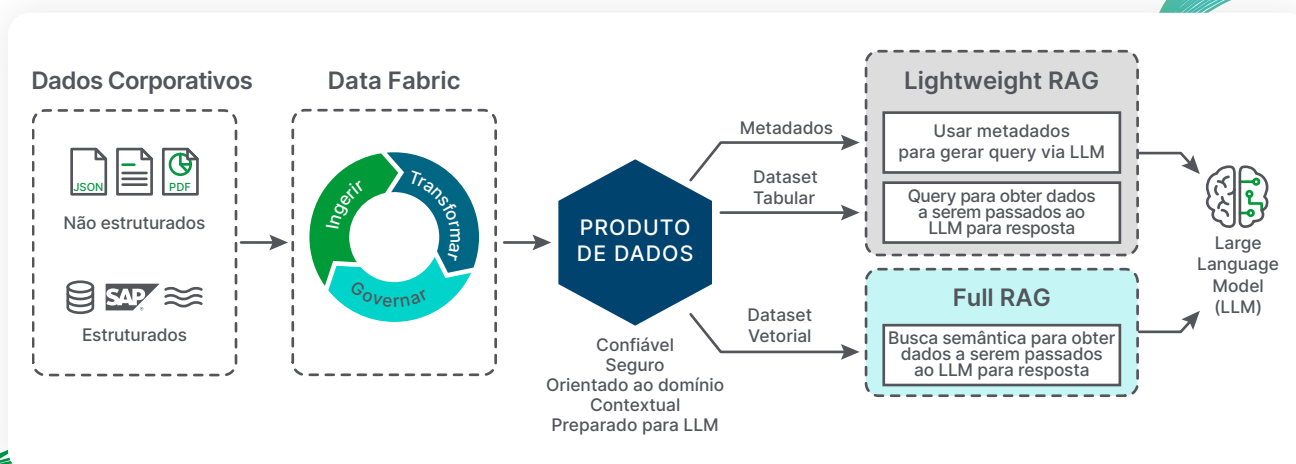
Arquiteta-Chefe de Dados da Waterstone Mortgage

Entregando dados contextuais para LLMs com produtos de dados

A IA generativa está levando grandes oportunidades de aumento de produtividade a todos os aspectos do negócio. No cerne da IA generativa estão os modelos de base — LLMs como GPT-3, Anthropic, Llama e BERT —, que foram pré-treinados com uma quantidade imensa de dados de texto (páginas da internet, documentos, etc.). No entanto, os LLMs de base não entendem os dados internos específicos de uma empresa. É por isso que eles precisam primeiro receber dados corporativos.

Uma das técnicas mais eficientes de fornecer dados corporativos aos LLMs é por meio de um mecanismo chamado Geração Aumentada de Recuperação (RAG - Retrieval-Augmented Generation). Conforme o nome sugere, os dados contextuais relacionados a um prompt do usuário (uma questão) são **recuperados**. Depois, o prompt do usuário é **aumentado** com esses dados e passado ao LLM para a geração de uma resposta. Há dois métodos de RAG: 'lightweight RAG' e 'full RAG'.

- No lightweight RAG (para dados estruturados), a aplicação RAG passa os metadados ao LLM para a geração de uma query SQL. Depois, a query é executada em um dataset tabular e os resultados são passados ao LLM para a geração de uma resposta. Esse método é usado quando é preciso que os dados forneçam uma resposta específica.
- No modo 'full RAG' (para dados estruturados e não estruturados), os dados são vetorizados primeiro. Depois, a aplicação RAG executa uma busca semântica nesses dados vetorizados para recuperar os dados. Esses dados são, então, passados ao LLM para a geração da resposta.



Nos dois métodos, é importante ter dados limpos, contextuais e de alta qualidade para a aplicação RAG.

Os produtos de dados oferecem dados contextuais e orientados ao domínio – de diversas formas – para consumo. Conforme mencionado anteriormente, os produtos de dados contêm metadados ricos (de negócios, técnicos e orientados ao domínio) e datasets com grande curadoria (arquivos, tabelas ou vetorizados) para consumo. Dessa forma, os produtos de dados se tornam um veículo perfeito para as aplicações RAG.

A criação de uma aplicação RAG começa com a busca no catálogo de produto de dados/marketplace pelo produto de dados certo para uso, com base na questão que está sendo proposta. Por exemplo: uma questão relacionada a vendas pode ser melhor atendida por um produto de dados de “Oportunidade”, proveniente do CRM, baseado em metadados detalhados associados ao produto de dados. Ou uma questão relacionada a um pedido pode ser melhor atendida por um produto de dados de “Pedido de venda”, proveniente do SAP, baseado em metadados detalhados associados com aquele produto de dados. Assim que o produto de dados é identificado, a aplicação RAG pode fazer uma query diretamente no dataset de produto de dados — tabular ou vetorizado —, baseada no caso de uso.

Conclusão

À medida que os dados continuam a impulsionar os negócios modernos, os produtos de dados se tornam cada vez mais fundamentais para **desbloquear todo o valor dos dados** e acelerar os resultados da empresa, **aprimorando a governança e a confiança nos dados**. Na Qlik, o nosso compromisso é ajudar você nesta jornada de construção de sua base de dados com recursos 'self-service'. Isso envolve empoderar as suas equipes de dados para construir e gerenciar produtos de dados, garantindo que os consumidores de dados possam rapidamente buscar, identificar e utilizar produtos de dados para alcançar resultados de negócios bem-sucedidos.

Uma nota final: embora os produtos de dados sejam ferramentas poderosas, eles não são uma solução do tipo 'bala de prata'. O verdadeiro poder dos produtos de dados reside na identificação das necessidades do negócio e em ter uma estratégia de dados federada bem definida, garantindo uma governança eficaz e promovendo a colaboração entre as equipes.

“ Cada vez mais, as empresas querem criar produtos de dados centrados no domínio, capazes de transcender a qualidade e a integração de dados. O foco dos produtos de governança e qualidade de dados da Qlik é oferecer uma solução abrangente e madura, alinhada às necessidades de hoje dos clientes. ”

Stewart Bond

VP de Inteligência de Dados e Integração de Software da IDC



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

**Quer desbloquear
todo o valor dos seus
dados?**

Saiba mais