

Manual para uma engenharia de dados moderna

Mudança de mentalidade: por que você deve tratar os dados como um produto	4
Práticas de engenharia para acelerar a entrega do seu produto de dados	9
Os dados são um esporte de equipe: desenvolvendo uma equipe de dados eficaz	14
Três princípios de planejamento de entrega para ajustes em direção ao produto de dados certo	19
Arquitetura para sistemas de dados: como equilibrar as compensações nas decisões sobre tecnologia	25
A qualidade é fundamental: encontrando o valor em sua estratégia de teste de dados	29
Mudar para a esquerda em termos de segurança e privacidade: porque é fundamental para a velocidade, qualidade e confiança da cliente	34



Introdução

Os dados foram apelidados de “o novo petróleo” - altamente valiosos e capazes de impulsionar indústrias inteiras. Mas, a menos que sejam tratados com cuidado e de forma estratégica, podem ser inúteis ou até mesmo prejudiciais para sua empresa.

Não importa em que negócio você esteja, os dados podem lhe dar um ponto claro de diferença e vantagem competitiva. E agora que eles também são um subproduto de cada passo digital que damos, suas equipes têm a capacidade de coletar e armazenar mais dados do que nunca. Usando-os para tomar melhores decisões estratégicas e criar experiências de cliente personalizadas e sem atritos.

No entanto, à medida que nossa necessidade de insights aumenta em rápida escala, as arquiteturas de plataformas de dados centralizadas não estão conseguindo acompanhar esse ritmo. Elas não têm a flexibilidade necessária para permitir que equipes dispersas tomem decisões informadas e oportunas. Com os desafios adicionais relacionados à governança, privacidade, segurança e qualidade dos dados, as organizações estão lutando para gerenciar a complexidade da operacionalização de seus ativos de dados.

A abordagem moderna de dados (MDS, modern data stack em inglês) - um conjunto de ferramentas e padrões usados para a integração de dados - surgiu para enfrentar esses desafios. Ela ajuda você a analisar dados, melhorar a eficiência e descobrir novas oportunidades. E há uma variedade cada vez maior de ferramentas prontas para escolher, evitando a necessidade de soluções personalizadas. Por sua vez, isso permite custos mais baixos e maior tempo de produção.

No entanto, para obter os benefícios do MDS, é preciso mais do que ter as ferramentas certas. Seu sucesso é sustentado por práticas modernas de engenharia e princípios de fornecimento de software que permitem acelerar o desenvolvimento, reduzir o risco de grandes projetos e melhorar continuamente os produtos. Você também precisa dos conjuntos de habilidades corretos em suas equipes, para que elas possam tornar os dados mais acessíveis, direcionar os problemas certos e criar os produtos certos. E, é claro, uma inteligência de dados valiosa depende inteiramente da qualidade dos dados.

Esses são os aspectos que abordamos neste manual, ajudando você a economizar tempo, reduzir riscos e melhorar o retorno sobre o investimento de seus projetos de dados.

Saiba como a mudança para uma mentalidade de produto de dados o ajudará a criar a coisa certa e a criar a coisa certa - e como montar a equipe certa para que isso aconteça. Explore práticas e princípios que te ajudarão a acelerar a produção e descubra como economizar tempo ao detectar problemas de qualidade de dados com antecedência. Além disso, descubra como você pode incorporar a segurança e a privacidade desde o início para melhorar a qualidade do seu produto, criar confiança com suas clientes e permitir que você avance mais rapidamente.

Vamos começar.

Mudança de mentalidade: porque você deve tratar os dados como um produto

Por Keith Schulze e Kunal Tiwary

Atualmente, as organizações estão reconhecendo cada vez mais o valor potencial dos dados, mas muitas não conseguem obter um retorno sobre o investimento de seus ativos de dados.

Quando se trata de desenvolver ativos de dados, muitas organizações adotam uma abordagem do tipo “construa e eles virão”. Embora essa filosofia possa ter valido a pena para Kevin Costner em Campo dos Sonhos, as organizações podem não ter a mesma sorte. Porque há uma falha inerente à abordagem: ela não leva em consideração as necessidades das pessoas usuárias de dados.

E se mudássemos a mentalidade e considerássemos algumas valiosas lições centradas na pessoa usuária das nossas equipes de produtos? E se gerenciássemos os dados como um produto, e não apenas como um ativo? Estamos vendo essa mudança de percepção ganhar força, permitindo que as organizações obtenham mais valor dos projetos de dados.

Repensando os dados

A mentalidade de dados como um produto é um dos quatro princípios do data mesh, um estilo de gerenciamento de dados que descentraliza os modelos de arquitetura de projetos. Os dados como um produto tratam os usuários de dados como clientes, desenvolvendo produtos de dados para agregar valor a eles e ajudá-los a atingir suas metas finais. Por exemplo, se a meta final da sua cliente for reduzir a taxa de rotatividade em 10%, você precisará começar com essa meta e trabalhar de trás para frente - e desenvolver um produto de dados de previsão de rotatividade que atenda a essa necessidade. Pensar nos dados como um produto significa colocar essas necessidades da pessoa usuária no centro do seu design. Eles são projetados para serem compartilhados, não controlados.

Zhamak Dehghani, autor do livro Data Mesh, fornecendo valor de dados em escala e fundador do conceito de Data Mesh, descreve-o da seguinte forma: “Os dados como produto são muito diferentes

Os dados como um produto tratam as pessoas usuárias de dados como clientes, desenvolvendo produtos de dados para agregar valor e ajudá-los a atingir suas metas finais.





dos dados como ativo. O que você faz com um ativo? Você a coleta e o acumula. Com um produto, é o contrário. Você os compartilha e torna a experiência com esses dados mais agradável.”

O pensamento voltado ao produto requer um profundo conhecimento e compreensão da cliente. Suas equipes podem, então, construir para problemas do mundo real e desenvolver continuamente produtos que ofereçam mais valor.

Construa da *forma correta* e eles virão

Muitos produtos de dados fracassam porque são uma solução em busca de um problema - por exemplo, incluir um novo conjunto de dados na plataforma porque “alguém” achará útil. Adicionar mais dados não necessariamente resolve os problemas da cliente ou agrega valor a ela.

É por isso que é tão importante começar sabendo quem é a sua cliente e o que é mais valioso para ela. Que problemas elas estão tentando resolver? O que está em jogo para elas se não puderem usar ou acessar os dados com facilidade? Essas clientes podem ser internas ou externas - a chave é pensar além de simplesmente oferecer fontes de dados e esperar que as pessoas usuárias adaptem ou comprometam a maneira como trabalham para usá-los.

Infelizmente, não há solução mágica para isso. Leva tempo para entender sua base de clientes e suas metas, e envolve testes no mundo real e refinamento constante. E uma vez que você tenha resolvido isso para um grupo de clientes, como dimensionar e expandir isso? Você pode tornar esses produtos reutilizáveis, satisfazendo as necessidades de uma gama mais ampla de clientes?

Na Thoughtworks, adaptamos o modelo de processo de design Double-Diamond para **garantir que vamos construir a coisa certa e da forma correta**. Isso começa com a identificação das necessidades da cliente. Usamos um processo estruturado de descoberta e criação para descobrir esses requisitos para qualquer novo produto de dados. Em seguida, aplicamos um conjunto de práticas e ferramentas bem compreendidas que são conhecidas por fornecer software e dados de alta qualidade.

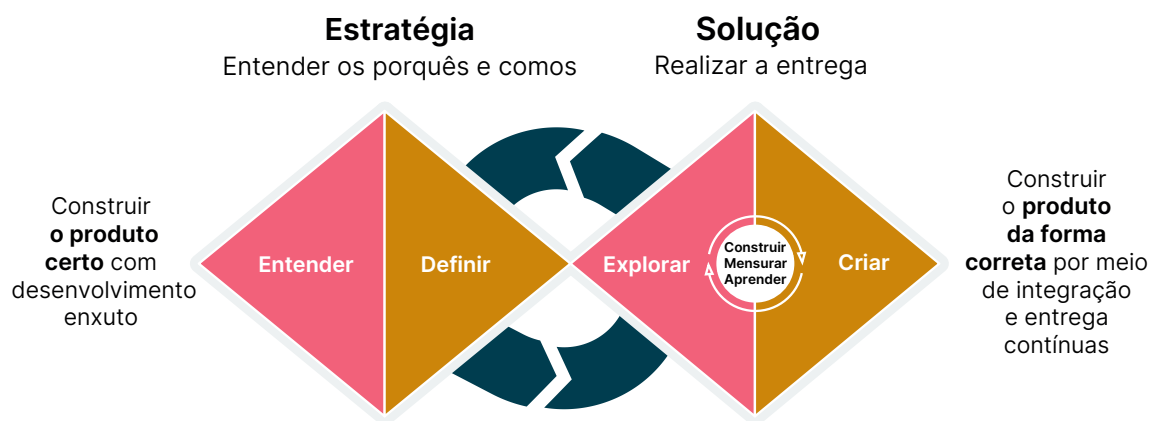


Figura 1: O diamante duplo da Thoughtworks, modelo de processo de design

“Os dados como um produto são muito diferentes dos dados como um ativo. O que você faz com um ativo? Você coleta e acumula. Como um produto é do jeito contrário. Você compartilha e transforma a experiência com esses dados mais agradável.”

Zhamak Deghani
Autora o Data Mesh, Delivering Data Value at Scale,
e fundadora do data mesh concept



Incentivar a propriedade e a responsabilidade

Se, na mentalidade mais tradicional, os projetos terminam quando um conjunto de dados ou relatório é entregue, o pensamento do produto exige que as equipes mantenham a propriedade sobre um produto de dados durante todo o seu ciclo de vida. Isso significa que as pessoas proprietárias do produto de dados são responsáveis por evoluir e adaptar o produto para garantir que ele continue a atender às necessidades das clientes, mesmo que seus requisitos mudem..

Assim como os produtos de software, os produtos de dados também se beneficiam de uma equipe responsável que melhora continuamente o desempenho e lança novos recursos em um ambiente seguro. Isso também reduz os ciclos de feedback necessários para evoluir ou aprimorar esses produtos. Ele incentiva a comunicação direta entre produtores e consumidores dos produtos de dados, eliminando processos de planejamento central longos e complicados.

As pessoas proprietárias de um produto de dados também são responsáveis por manter os níveis de serviço acordados. Isso é importante porque, sem uma responsabilidade clara, podem haver processos complexos e prioridades conflitantes a serem enfrentadas quando os serviços forem interrompidos.

As pessoas proprietárias dos produtos de dados são responsáveis por evoluir e adaptar o produto de dados para garantir que ele continue a atender às necessidades das clientes mesmo que seus requisitos mudem



Por exemplo, as organizações de varejo usam uma série de métricas para facilitar o planejamento de demanda (por exemplo, precisão da previsão, taxa de atendimento de pedidos). Diferentes equipes dependem dessas métricas para prever e fornecer estoque para atender à demanda. Quaisquer atrasos ou erros nos relatórios podem ter impactos graves nos processos comerciais posteriores, levando a clientes insatisfeitas e, a uma perda de receita ou a um excedente de estoque com custo para os negócios.

À medida que uma empresa evolui, podem haver outras métricas de planejamento de demanda que permitiriam previsões mais precisas; qualquer atraso na implementação dessas métricas também significa um sacrifício no lucro potencial. As empresas precisam evoluir continuamente seu processo de planejamento de demanda para usar as métricas mais precisas e garantir que elas sejam confiáveis e de alta qualidade. Qualquer erro deve ser corrigido imediatamente para minimizar o impacto sobre o público consumidor final.

Ter uma equipe com responsabilidades claras para desenvolver, evoluir e manter essas métricas como um produto de alto nível de serviço garante que:

1. Você está sempre fornecendo ao público consumidor downstream as métricas mais precisas para facilitar o planejamento da demanda, e
2. Você minimiza os impactos das interrupções em outras equipes no processo de planejamento da demanda.

O que é necessário para fazer a mudança

Uma mudança de mentalidade como essa geralmente requer mudanças culturais e comportamentais também. Se a sua organização quiser colher os benefícios dos produtos de dados centrados no usuário, precisará adotar uma cultura mais centrada no produto e no cliente e formar equipes multifuncionais para apoiar essa abordagem.

Uma das maneiras de chegar lá é deixar de ter equipes alinhadas a arquétipos ou conjuntos de habilidades e passar a ter pequenas equipes orientadas a produtos com metas bem focadas. Essas equipes podem exigir uma combinação de diferentes capacidades, como pessoas engenheiras de dados, cientistas de dados, QAs e designers, para desenvolver um produto que atenda às necessidades dos clientes.

Criar uma cultura em que o aprendizado com o fracasso seja aceito e comemorado também é fundamental para o sucesso do desenvolvimento de produtos de dados eficazes. Descobrir o que não funciona ou onde estão os pontos de atrito permite que as equipes ajustem seu pensamento e abordagem para projetos futuros e melhorem continuamente os produtos e a experiência da cliente ao longo do caminho.



Criar uma cultura em que o aprendizado com o fracasso seja aceito e celebrado também é fundamental para o sucesso do desenvolvimento de um produtos de dados

Colocando em prática os dados como um produto

Um produto de dados eficaz deve ser:

- Descobrível:** Se você deseja que uma cliente use um produto de dados, ele precisa ser capaz de encontrá-lo. A capacidade de descoberta pode assumir várias formas, desde uma lista primitiva de conjuntos de dados em um sistema wiki interno até um catálogo de dados completo. Independentemente da implementação, os catálogos devem conter meta informações importantes sobre os produtos de dados, como suas proprietárias, fonte de origem, linhagem e conjuntos de dados de amostra.
- Endereçável:** Os produtos de dados também devem ter um único endereço exclusivo onde possam ser encontrados. Isso facilita a localização para a cliente e reduz o tempo que as equipes de dados gastam para ajudar as pessoas a localizá-los



Auto Descritivos e interoperáveis: Os produtos de dados devem fornecer metadados às pessoas usuárias e aos sistemas automatizados de forma a permitir o autoatendimento das usuárias. Por exemplo, um produto de dados deve expor metadados que sejam capazes de descrever as fontes de dados usadas para criar o produto de dados, o esquema e as informações sobre os resultados do produto de dados.



Confiável e seguro: Para que um cliente use um produto de dados com confiança, o produto deve se comprometer com um nível acordado de confiabilidade (ou qualidade). Isso significa definir antecipadamente um conjunto de Objetivos de Nível de Serviço (SLO) e Indicadores de Nível de Serviço (SLI) mensuráveis, além de implementar mecanismos automatizados para testar e relatar regularmente as métricas de SLI.



Seguro e governado por um controle de acesso global: Com o rápido aumento das violações de dados, nunca foi tão importante proteger seus produtos de dados e criar segurança. Embora os conjuntos de dados registrados devam ser automaticamente descobertos por todos os clientes, eles não devem ser acessíveis por padrão. As pessoas usuárias precisarão solicitar acesso a cada conjunto de dados de que precisam, com os proprietários de dados concedendo ou negando acesso individualmente, usando controles de acesso federados.



Obtendo benefícios em toda a organização

A adoção de uma mentalidade de dados como produto é um exercício que envolve toda a organização e exige uma mudança não apenas nas perspectivas, mas também na cultura e nas práticas. O esforço certamente vale a pena. Os princípios do pensamento de produto permitem que você desenvolva vários produtos de dados que podem ser usados dentro da organização e, por fim, ajudam a formar uma rede eficaz e simplificada de produtos de dados. E, quando se torna incorporado à sua empresa, ajuda a elevar o nível das equipes de tecnologia, ajudando-as sempre a pensar em criar valor e trabalhar para obter resultados para cada pessoa usuária.

Práticas de engenharia para acelerar a entrega de seus produtos de dados

Por David Tan e Mitchell Lisle

Abordar o desenvolvimento de produtos de dados da mesma forma que você abordaria a criação de software é um bom ponto de partida. No entanto, os produtos de dados costumam ser mais complicados do que os aplicativos de software, pois fazem uso intensivo de software e dados. As equipes não só precisam navegar pelos diversos componentes e ferramentas diferentes que fazem parte do desenvolvimento de software, como também precisam lidar com a complexidade dos dados. Devido a essa dimensão adicional, pode ser muito fácil para as equipes ficarem atoladas em processos de desenvolvimento e implementações de produção complicados, o que gera ansiedade e atrasos no lançamento.

Na Thoughtworks, descobrimos que a aplicação intencional de práticas de engenharia “padrão sensatas” nos permite entregar produtos de dados de forma sustentável e rápida. Neste artigo, vamos nos aprofundar em como isso pode ser feito.

Aplicação de padrões sensatos na engenharia de dados

Muitas práticas de padrão sensatas têm suas raízes na **entrega contínua (CD)** – um conjunto de práticas de desenvolvimento de software que permite que as equipes liberem as alterações para a produção de forma segura, rápida e sustentável. Esse conjunto de práticas reduz o risco de erros nas versões, reduz o tempo de lançamento no mercado e os custos e, por fim, melhora a qualidade do produto. **As práticas de entrega contínua delivery** (como pipelines automatizados de construção e implantação, infraestrutura como código, CI/CD e desenvolvimento baseado em tronco) também **se correlacionam positivamente com a entrega de software e o desempenho comercial de uma organização**.

Desde o desenvolvimento e a implementação até a operação, as práticas de padrão sensatas nos ajudam a criar o que é certo. Essas práticas incluem:

- Criação rápida e automatizada
- Desenvolvimento baseado em troncos
- Programação em pares
- Construir segurança em
- Construção rápida e automatizada
- Pipeline de implantação automatizada
- Gerenciamento eficaz da qualidade e do débito
- Criação para produção



Como explicaremos mais adiante neste capítulo, essas práticas são essenciais para gerenciar a complexidade das pilhas de dados modernas e acelerar a entrega de valor, pois fornecem às equipes as seguintes características que as ajudam a entregar qualidade com rapidez:

- ▶▶ **Feedback rápido:** Descubra se uma mudança foi bem-sucedida em instantes, não em dias. Seja para saber se os testes unitários foram aprovados, se você não interrompeu a produção ou se a cliente está satisfeito com o que você criou.
- ⦿ **Simplicidade:** Crie para o que você precisa agora, não para o que você acha que pode vir a acontecer. Isso permite que você limite a complexidade e, ao mesmo tempo, faça escolhas que permitam que o software mude rapidamente e atenda aos requisitos futuros.
- 🔄 **Repetibilidade:** Tenha a confiança e a previsibilidade resultantes da remoção de tarefas manuais que podem introduzir inconsistências e dedique seu tempo ao que importa, não à solução de problemas.

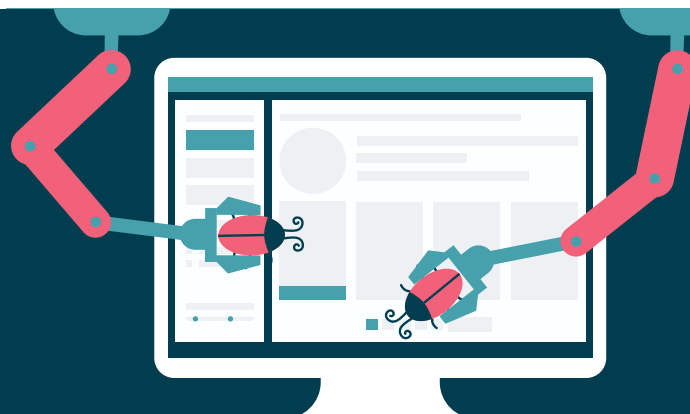
Práticas de engenharia para a engenharia de dados moderna

Embora exista uma **grande quantidade de trabalhos** detalhando como aplicar a entrega contínua no desenvolvimento de soluções de software, pouco está documentado sobre como usar essas práticas na engenharia de dados moderna. Aqui estão três maneiras pelas quais adaptamos essas práticas para criar e fornecer rapidamente produtos de dados eficazes.

1. Automação de testes e o gerenciamento de dados de testes

A automação de testes é a chave para o feedback rápido, pois permite que as equipes desenvolvam sua solução sem os gargalos resultantes de testes manuais e defeitos de produção. Além das práticas bem conhecidas de desenvolvimento orientado por testes (orientar o desenvolvimento de software por meio da criação de testes), também é importante considerar os testes de dados.

A automação de teste é a chave para o feedback rápido, pois permite que as equipes desenvolvam suas soluções sem os gargalos resultantes de teste manuais e defeitos de produção.



Semelhante à pirâmide de teste prático para entrega de software, a **grade de dados de teste prático** (Figura 1) ajuda a orientar como e onde você investe seu esforço para obter uma imagem clara e oportuna da qualidade dos dados ou da qualidade do código, ou de ambos. A grade considera as seguintes camadas de teste de dados:

- **Os testes de dados** pontuais capturam um único cenário que pode ser analisado logicamente, por exemplo, uma função que conta o número de palavras em uma publicação de blog. A implementação desses testes deve ser barata e deve haver muitos deles, para definir as expectativas em uma série de circunstâncias específicas.

- **Sample data tests** fornecem feedback valioso sobre os dados como um todo sem processar grandes volumes. Eles permitem que você compreenda as expectativas mais imprecisas e a variação nos dados, especialmente ao longo do tempo. Embora tragam complexidade adicional e exijam algum ajuste de limite, eles revelam problemas que os testes pontuais não capturam. Considere o uso de amostras **sintéticas** para esses testes.
- **Os testes de dados globais** revelam cenários imprevistos por meio de testes com todos os dados disponíveis. Eles também são os menos direcionados, mais sujeitos a mudanças externas e mais caros do ponto de vista computacional.

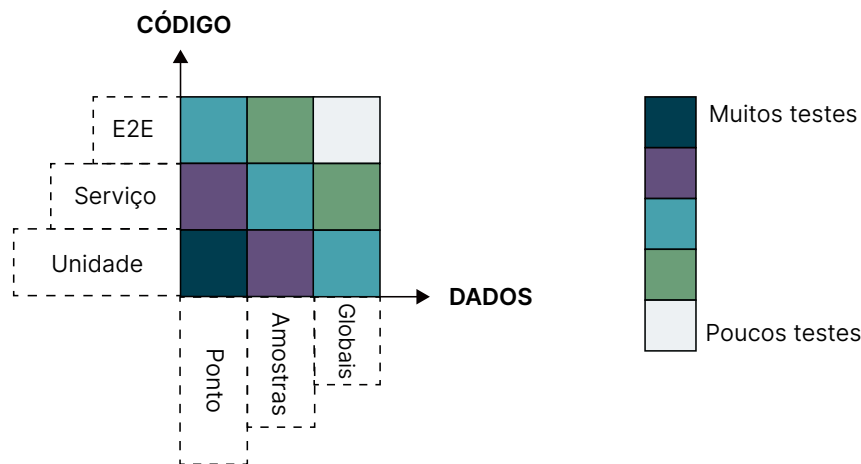


Figura 1: Grade de teste de dados práticos

Você pode aplicar esses testes somente aos dados ou combiná-los com testes de código para verificar vários estágios da transformação de dados - nesse caso, você consideraria as duas dimensões como um teste prático de dados. Novamente, isso não é uma receita, você não precisa preencher todas as células e os limites nem sempre são precisos, mas essa grade ajuda a direcionar nosso esforço de teste e monitoramento para obter feedback rápido e econômico sobre a qualidade em sistemas com uso intensivo de dados.

Um ponto importante sobre o gerenciamento de dados de teste

Você precisará de dados semelhantes aos de produção para usar a grade de teste de dados práticos. Comece pensando em **three planes of flow** (Figura 2):

- **No plano de código**, o código flui ao longo do eixo Y, de baixo para cima, entre ambientes (por exemplo, desenvolvimento, teste, produção). Em termos tradicionais de engenharia de software, trata-se de um pipeline de CI/CD - o fluxo com o qual os engenheiros de software normalmente estão familiarizados.
- **No plano de dados**, os dados fluem ao longo do eixo X, da esquerda para a direita, em cada ambiente em que os dados são transformados de uma forma para outra. Isso é um pipeline de dados e é algo que as pessoas especialistas em dados entendem muito bem.
- **No plano de dados reverso**, os dados fluem ao longo do eixo Y na direção oposta ao plano de código. Você pode criar amostras de dados de produção em ambientes de teste usando diferentes técnicas de preservação de privacidade ou ofuscação, como mascaramento e privacidade diferencial, ou pode criar dados puramente sintéticos, usando amostras de dados.

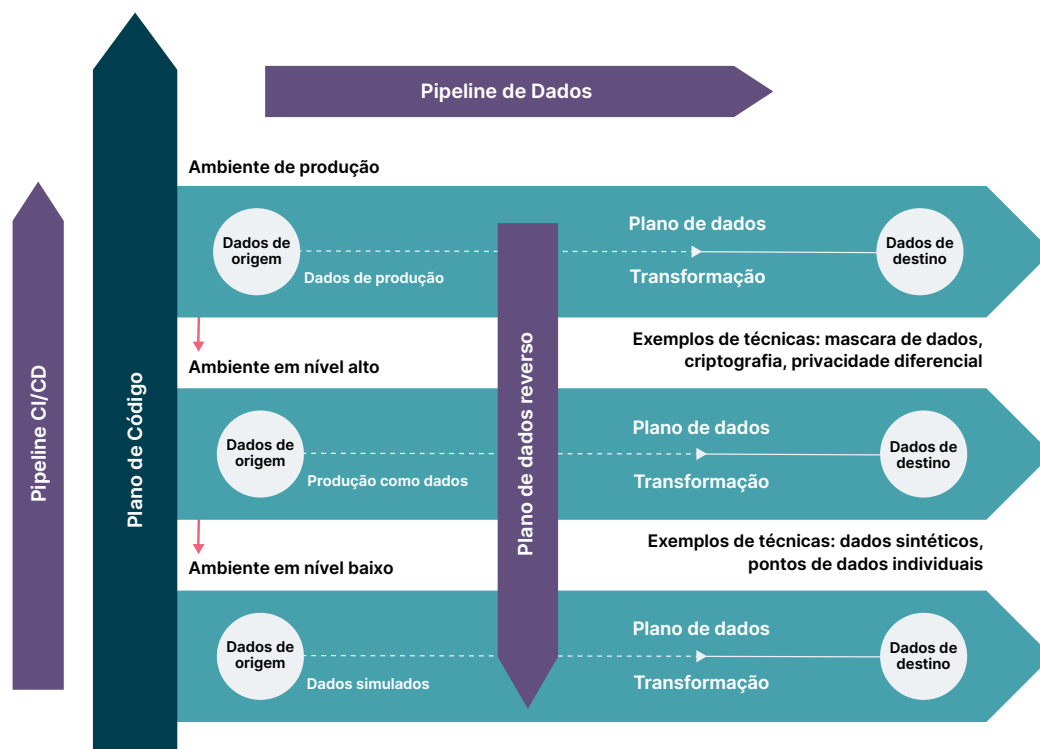


Figure 2: Visualization of a reverse data plane in a data pipeline

2. Estudo de caso: Aceleração da entrega com as práticas de engenharia padrão sensatas

A Empresa Y é uma empresa de software de grande porte cujos clientes usam seu software para fins de relatórios. A empresa Y queria criar um novo recurso que permitisse que seus usuários vissem dados históricos e de previsão que os ajudassem a tomar decisões mais bem informadas.

No entanto, os dados históricos necessários para treinar modelos de aprendizado de máquina para criar previsões estavam bloqueados em bancos de dados operacionais, e o armazenamento de dados existente não podia ser dimensionado para atender a esse novo caso de uso.

Aplicamos práticas sensatas de engenharia padrão para criar uma arquitetura de dados de streaming escalonável para a Empresa Y. Os pipelines ingeriram dados do armazenamento de dados operacionais em um armazenamento de dados analíticos para atender ao novo recurso do produto.

Padrões sensatos em ação

- Adotamos a **programação em pares**, em que as pessoas desenvolvedoras escrevem códigos juntos e trabalham no novo recurso, fornecendo feedback em tempo real uns aos outros por meio da **automação de testes**. Isso envolveu testes unitários, de integração e de ponta a ponta executados localmente.
- Os pares acessaram amostras de **dados de teste** que representam os dados que esperamos na produção e usaram ganchos do git que fazem lint, verificam se há segredos e executam os conjuntos de testes. Se algo fosse falhar na integração contínua, os ganchos mudavam o feedback para a esquerda e avisavam as pessoas desenvolvedoras antes que eles enviassem o código e isso causasse uma compilação vermelha.
- Garantimos que o código esteja sempre em um estado implementável adotando a **integração contínua** (CI), uma prática que exige que as pessoas desenvolvedoras integrem o código em

um repositório compartilhado várias vezes ao dia. Isso acelerou a implementação, enquanto um **pipeline de construção automatizado** e rápido que executa os conjuntos de testes automatizando no servidor de CI forneceu feedback rápido sobre a qualidade

- Nossa equipe criou os artefatos uma vez e os implantou em cada ambiente por meio de **implantação automatizada**. Aplicamos a **infraestrutura como código** (ou seja, a infraestrutura do aplicativo, a configuração da implantação e a observabilidade são especificadas no código) e provisionamos tudo automaticamente. Se a compilação for interrompida no ambiente de pré-produção, nós a corrigimos rapidamente (em 10 minutos ou menos) ou revertermos a alteração.
- Os **testes pós-implantação** garantiram que tudo estava funcionando conforme o esperado, dando-nos a confiança necessária para permitir que o pipeline de CI/CD implantasse automaticamente as alterações na produção (ou seja, **implantação contínua**).
- A **observabilidade** e as **notificações proativas** permitiram que todos da equipe soubessem a qualquer momento a integridade do nosso sistema. E o **registro com** identificadores de correlação ajudou a rastrear um evento por meio de sistemas distribuídos e a observar os pipelines de dados.
- Quando necessário, **refatoramos** (reescrivendo pequenos trechos de código em recursos existentes) e **gerenciamos a dívida técnica** como parte das histórias e iterações, e não como uma reflexão tardia enterrada no backlog.

Desempenho em um nível de elite

Em apenas 12 semanas, fornecemos um conjunto de streaming de ponta a ponta que atualizava continuamente o repositório analítico e alimentava o recurso de previsão que fornecia aos usuários informações em tempo real. Essas práticas ajudaram a equipe da cliente a acelerar a entrega e a se tornar uma **empresa de "elite"**, conforme medido pelas quatro principais métricas:

- Frequência de implementação: sob demanda (várias implementações de produção por dia)
- Tempo de espera para alterações (tempo entre o commit do código e a execução bem-sucedida do código na produção): 20 minutos
- Tempo para restabelecer o serviço: menos de 1 hora
- Taxa de falha de alteração: < 15%

Para alterações mais complexas, nossa equipe usa **alternância de recursos**, alternância de dados ou **implementação azul-verde**.

Uma observação sobre o desenvolvimento baseado em tronco

Durante toda a entrega, aplicamos o **desenvolvimento baseado em tronco** (TBD), que é uma prática de controle de versão onde todos os membros da equipe mesclam suas alterações em um branch dedicado ou o principal. Quando todo o time trabalha na mesma filial, o TBD aumenta a visibilidade e a colaboração, reduz o esforço duplicado e nos dá um feedback muito mais rápido do que a prática alternativa de revisões de solicitações pull.

Se nos opusemos ao envio código para o branch dedicado, geralmente este é um sinal de que estamos perdendo algumas das práticas essenciais mencionadas acima. Por exemplo, se você tem medo que suas alterações possam causar problemas de produção, torne seus testes mais abrangentes. E se você precisar de uma segunda revisão para as solicitações pull, considere a **programação em pares** para acelerar o feedback. É importante observar que o TBD é uma prática que só é possível se tivermos a rede de segurança e as barreiras de qualidade fornecidas pelas práticas anteriores.

Os dados são um esporte de equipe: criando uma equipe de dados eficaz

Por Keith Schulze e Kunal Tiwary

Considere sua equipe esportiva favorita: cada pessoa traz um conjunto exclusivo de habilidades e experiências para o time e desempenha uma função específica. Poderosos e precisos, jogadores quarterbacks dão as ordens no campo, enquanto jogadores linebackers defendem com força e velocidade. Os versáteis meio-campistas conectam os dois e mantêm a bola em movimento sem problemas. Para que a equipe seja bem-sucedida, essas funções e habilidades devem funcionar em conjunto e sem problemas.

As equipes de dados trabalham de forma semelhante. Como parte integrante de processos mais amplos, uma equipe bem-sucedida reunirá habilidades, experiências e conhecimentos complementares para lidar com questões críticas, como:

- Em quais novos mercados, serviços ou otimizações de custos a empresa está embarcando?
- Como os dados poderiam ajudar nossas pessoas a tomar essas decisões?
- Esses dados estão disponíveis para as principais tomadoras de decisão em um formato compreensível?

À medida que a moderna cadeia de valor de dados evolui, as organizações estão fazendo uma mudança fundamental na forma como pensam sobre os dados e como criam equipes em torno deles. Ao se concentrar na democratização dos dados em toda a organização, pense em como alinhar suas estruturas internas com as pessoas que compõem suas equipes de dados. Veja o que achamos útil na formação de equipes de dados para produtos específicos.

Formação de equipes eficazes

A divisão de domínios ou áreas de interesse é uma ótima maneira de começar a identificar as equipes que você precisará formar. Procure áreas de alta coesão - em que os elementos estão intimamente relacionados entre si e têm um objetivo comum - e baixo acoplamento - módulos que funcionam independentemente uns dos outros - entre os domínios. Por exemplo, a Netflix mapeia alguns de seus domínios e produtos de dados da seguinte forma:

- **Assinaturas:** prever a rotatividade e ajudar a prever quais são as clientes que provavelmente cancelarão suas assinaturas
- **Conteúdo:** recomendações e classificação de conteúdo
- **Jogadores:** estatísticas relacionadas ao cliente
- **Pagamento:** detecção de fraudes

Assim como as equipes de software, as equipes de produtos de dados são proprietárias de seus produtos coletivamente. Cada produto deve ter uma pessoa proprietária de produto nomeada que atue como embaixadora da equipe e principal comunicadora para as partes interessadas e outras equipes de produtos de dados. Elas conduzem o roteiro e o ciclo de vida do produto, comunicando as expectativas e facilitando a colaboração.

Elementos de uma equipe de dados vencedora

O desenvolvimento de produtos de dados eficazes exige uma equipe especializada com um conjunto de habilidades, experiências e conhecimentos multidisciplinares, incluindo pessoas engenheiras de dados, cientistas de dados, gerentes de produtos de dados, pesquisadoras de UX de dados e pessoas engenheiras de análise.

Uma das maneiras mais eficazes de formar sua equipe de dados é desenvolver funções que se concentrem em aspectos específicos da infraestrutura, do desenvolvimento e do ciclo de vida do produto. Cada função traz um conjunto diferente de habilidades, pontos fortes e abordagens necessárias para criar valor a partir dos dados. A natureza do produto de dados que você deseja criar determinará as funções necessárias.



O desenvolvimento de produtos de dados eficazes requer uma equipe especializada com um conjunto de habilidades, experiência e conhecimento.

É importante observar que uma função é diferente de um indivíduo que desempenha um papel na sua equipe. De fato, algumas pessoas podem se enquadrar em várias funções. Você deve considerar as funções necessárias para criar o produto e aquelas que o operarão e manterão. Diferentes equipes terão que:

- Criar produtos de dados e certifique-se de que eles sejam entregues por meio de pipelines de dados confiáveis;
- Usar produtos de dados e combine-os com análises avançadas para criar um novo valor comercial;
- Garantir que os produtos de dados sejam confiáveis e funcionem sem problemas;

Em algumas organizações, também é comum ver equipes de plataforma com conhecimento especializado em determinadas competências técnicas, como infraestrutura ou ciência de dados. Isso ajuda a reduzir a carga cognitiva em uma equipe de produtos de dados, pois os membros da equipe não precisam ser especialistas fora de seus conjuntos de habilidades..

Funções na prática

Considere um cenário em que você precisa criar uma nova oferta de serviços financeiros para um varejista. Seu objetivo é fornecer informações agregadas do histórico de crédito da cliente a uma entidade externa, de forma segura e protegida, para ajudá-la a tomar decisões de empréstimo de crédito.

Os **requisitos de alto nível** incluem:

- Adicionar dados de histórico de crédito de várias fontes de dados internas
- Transformar e agregar dados do histórico de crédito em um formato que dê suporte à tomada de decisões de empréstimo de crédito
- Fornecer uma API segura para atender a solicitações em tempo real de informações agregadas do histórico de crédito de uma cliente com base em um identificador exclusivo da cliente.

Os **requisitos multifuncionais** incluem:

- Os dados do histórico de crédito não podem ter mais de um dia
- Os dados não devem sair de um ambiente de rede seguro (ou seja, estar na internet pública)
- A soberania e a governança dos dados do histórico de crédito do cliente devem permanecer com o varejista. Todos os dados transferidos para o provedor de empréstimo externo devem ser auditados e regidos por políticas para sua reutilização e manuseio.

Vamos considerar algumas das funções que você precisará para criar e desenvolver o produto de dados do histórico de crédito do cliente no futuro.

- **Proprietária do produto:** a proprietária do produto é a pessoa responsável por maximizar o valor do seu produto de dados. Ela também desempenha uma função importante no apoio à equipe nos principais pontos de decisão e priorização. Ele ajudará a priorizar qual insight deve ser criado primeiro com base no retorno do investimento do recurso e no esforço envolvido.
- **Analista de negócios:** desempenha um papel fundamental na compreensão e no alinhamento do valor do seu produto de dados com as necessidades das clientes e da empresa em geral.
- **Pessoa engenheira de dados:** elas constroem os pipelines que obtêm dados de vários sistemas internos, transformam, e agregam os dados em um formato que dá suporte à tomada de decisões de crédito.
- **Pessoa engenheiro de infraestrutura:** ela cria uma infraestrutura reutilizável e dimensionável em torno do produto de dados para facilitar a reprodutibilidade, a integração/implantação contínua e automatizar o máximo possível.
- **Pessoa engenheira de back-end:** elas criam lógica comercial na forma de APIs de dados para facilitar as integrações de dados com a interface da pessoa usuária, outros produtos e ferramentas de visualização.
- **Garantia de qualidade:** responsável por manter a **qualidade dos dados e dos produtos**. Essa função é essencial para criar confiança com a cliente.



Uma observação importante sobre segurança

Como você estará compartilhando dados potencialmente confidenciais entre diferentes organizações, a segurança é uma preocupação fundamental para esse produto de dados. Embora seja fundamental que haja alguém na equipe que lidere a segurança, ela também deve ser vista como responsabilidade de toda a equipe, e você deve incorporar a segurança ao seu produto com o apoio de uma função de segurança central na empresa.

Lembre-se de que você não precisa de pessoas dedicadas para cada função. Você pode ter pessoas experientes na equipe que se encaixam em várias funções, ou algumas de suas plataformas principais podem desempenhar uma função de apoio. Por exemplo, você pode pedir a alguém para ser uma pessoa defensora da segurança e da privacidade e torná-lo responsável por garantir que a equipe siga as boas práticas de segurança. Essa pessoa pode trabalhar em estreita colaboração com uma equipe de plataforma de segurança principal para executar atividades importantes, como definir objetivos de segurança ou executar sessões de modelagem de ameaças.

As habilidades e os conhecimentos necessários em cada estágio do ciclo de vida do produto de dados são diferentes, portanto, as funções que você precisará mudarão ao longo do caminho.

Embora seja vital que haja alguém na equipe liderando a segurança, isso também deve ser visto como responsabilidade de todas.



Sobre a importância das habilidades interpessoais

Embora haja uma sobreposição de habilidades em algumas funções da equipe de dados, a importância das habilidades sociais abaixo para completar as funções é de extrema importância.



Liderança: É fundamental para estabelecer uma cultura de tratamento de dados como um produto. Especialmente se a organização estiver fazendo a transição de uma equipe de dados centralizada para um modelo descentralizado em que equipes autônomas são formadas em torno de um produto de dados. Durante essa transição, haverá períodos de incerteza e dúvidas sobre se o retorno do investimento justifica o esforço. Ter líderes experientes com uma visão clara de como um produto de dados agregará valor para as clientes, e a empresa pode ajudar uma equipe a navegar por esses tempos incertos e transmitir os benefícios de longo prazo de um produto de dados para a organização como um todo.



Coragem para se manifestar: A fala pode assumir muitas formas. Por exemplo, questionar por que estamos fazendo coisas que não estão alinhadas com nossos valores e metas. Ou levantar desafios e fornecer sugestões construtivas em uma retrospectiva da equipe (pressupondo um ambiente psicologicamente seguro) como o primeiro passo para a melhoria contínua.



Communication and storytelling: A comunicação e a narrativa eficazes ajudam a tornar o trabalho com dados, que tende a ser de natureza técnica, acessível a partes interessadas não técnicas, incentivam a colaboração e melhoram os resultados gerais.

Alinhamento das equipes de dados às estruturas organizacionais

Equipes de dados descentralizadas formadas em torno de produtos de dados devem ter seu próprio executivo C-Suite e ser tratadas como parte da engenharia convencional. A liderança precisa definir uma direção e fornecer governança - o suficiente para capacitar as equipes e permitir autonomia para que elas possam experimentar e descobrir livremente. Isso permite que eles forneçam valor real aos clientes. Os KPIs da pessoa executiva também devem estar alinhados com as metas da equipe de dados.

Definir o que é sucesso para a sua equipe de dados é fundamental, e as metas devem se concentrar em possibilitar decisões (resultados) em vez do número de painéis criados (atividades). Para os domínios de assinatura no exemplo da Netflix acima, uma métrica de sucesso pode ser reduzir a rotatividade em 10% no próximo trimestre, em vez de medir o número de pontos de dados adquiridos para produzir essa métrica.

A realização de sessões com a equipe de dados é uma ótima maneira de comunicar essas metas e como elas se encaixam no restante da organização, além de criar um entendimento interno das funções e dos papéis das equipes.



Figura 1: Cultura da mentalidade de times descentralizados

Como a gerência e técnicos de equipes esportivas de elite sabem, uma equipe desequilibrada pode ficar aquém do esperado quando chega a hora de atuar. Temos observado muitas organizações que não percebem a importância de equilibrar vários conjuntos de habilidades ao formar suas equipes de dados. Isso geralmente resulta em produtos ou experiências da pessoa usuária mal projetadas ou na transformação de produtos em uma fábrica de recursos sem considerar o caso de uso e o valor para a cliente.

Quando você pode recorrer a uma ampla gama de conhecimentos especializados e tem as estruturas certas para obter o máximo de valor de cada conjunto de habilidades, suas equipes podem atacar os problemas certos e prosperar. Como diz o famoso ditado, *o trabalho em equipe divide a tarefa e multiplica o sucesso*.

Três princípios de planejamento de entrega

Por David Tan, Keith Schulze e Mitchell Lisle

Em um capítulo anterior, analisamos a fundo as práticas de engenharia que poderão ajudar a entregar seus produtos de dados de forma rápida, segura e sustentável. Agora, exploraremos os princípios de planejamento de entrega para orientar como moldamos e sequenciamos nosso trabalho. Isso permite que as equipes agreguem valor incremental, reduzam o risco de grandes projetos e encontrem oportunidades de melhoria contínua ao fornecer produtos com uso intensivo de dados.

Nota: não abordaremos as atividades típicas de planejamento de iteração e planejamento de lançamento, como estimativas de histórias e acompanhamento do tempo de ciclo, pois elas estão se tornando cada vez mais comuns no setor. Essas práticas ainda são importantes, e queremos complementá-las compartilhando princípios e práticas adicionais que proporcionam um foco intencional na criação de valor.

Princípio 1: Fatiamento fino vertical

Uma armadilha comum na engenharia de dados é a entrega sequencial de baixo para cima de camadas funcionais de uma solução técnica (pense em data lake, data warehouse, pipelines de aprendizado de máquina e aplicativos voltados para o usuário) - ou **fatiamento horizontal**.

A desvantagem dessa abordagem é que os usuários só podem fornecer feedback valioso após investimentos significativos de tempo e esforço. Isso também leva a problemas de integração tardia quando as partes horizontais acabam se unindo, aumentando o risco de atrasos no lançamento.

Então, como podemos planejar e sequenciar nosso trabalho para liberar valor antecipadamente - e com frequência? Como podemos evitar a areia movediça da engenharia de dados pela engenharia de dados e criar uma cadência de valor comercial demonstrável em cada iteração? A resposta é mudar nosso modo de pensar e dividir o trabalho verticalmente.

As fatias verticais finas ajudam a garantir isso:

1. **No nível das histórias**, você articula e demonstra o valor comercial em todas as histórias e garante que a maioria delas sejam unidades de valor que possam ser enviadas de forma independente.
2. **No nível das iterações**, você demonstra regularmente o valor para os usuários, fornecendo uma coleção de histórias divididas verticalmente dentro de um prazo razoável.
3. **No nível das versões**, você planeja, sequencia e prioriza uma coleção de histórias orientadas para a criação de valor comercial demonstrável.

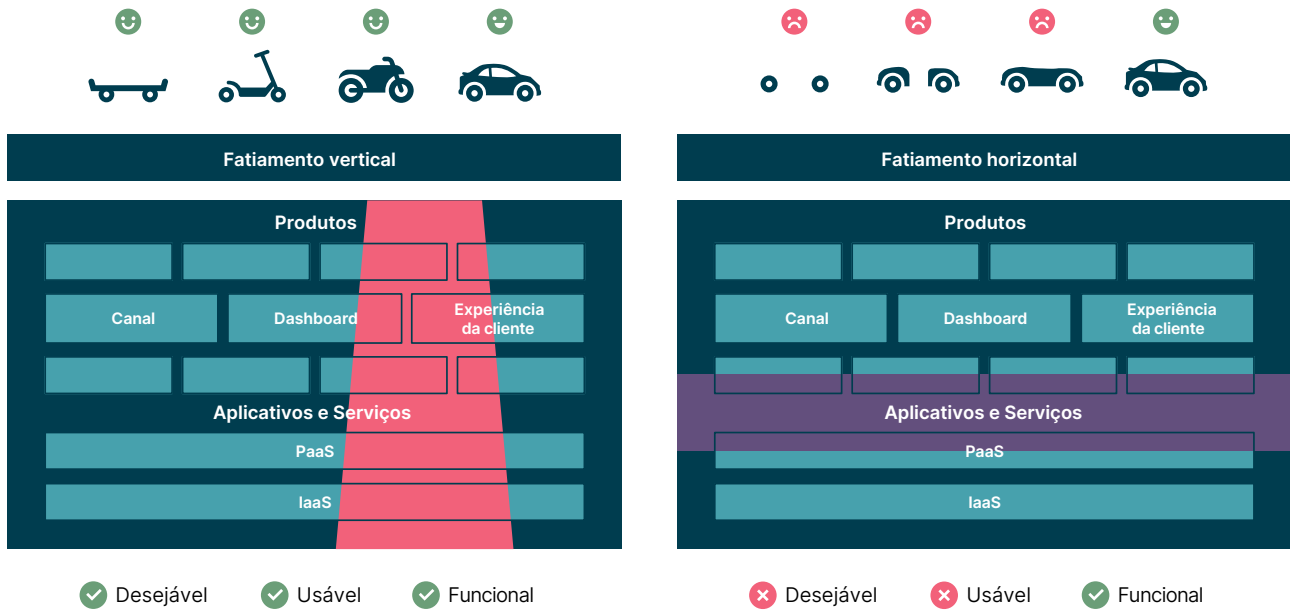


Figura 1: Entrega antecipada com fatias verticais finas

O rastreamento de uma fina fatia de valor por meio dos componentes de um ecossistema de dados e o fornecimento de valor em histórias de pessoas usuárias divididas verticalmente permitem que você teste e aprenda de forma mais econômica, resolva os problemas dos clientes com mais eficiência e forneça valor mais rapidamente. Descreveremos como isso se parece em ação no estudo de caso abaixo.

Princípio 2: Desenvolvimento de hipóteses com base em dados

Ao embarcar em produtos de dados, muitas vezes nos deparamos com a solução de problemas com poucos dados (é por isso que estamos investindo em engenharia de dados!) e altos níveis de riscos (as “incógnitas conhecidas” e as “incógnitas desconhecidas”, consulte a Figura 2). Nesse cenário, devemos nos concentrar em encontrar o caminho mais curto para as soluções certas e eliminar as soluções “incorretas” o mais rápido possível.



Figura 2: As quatro categorias de problemas

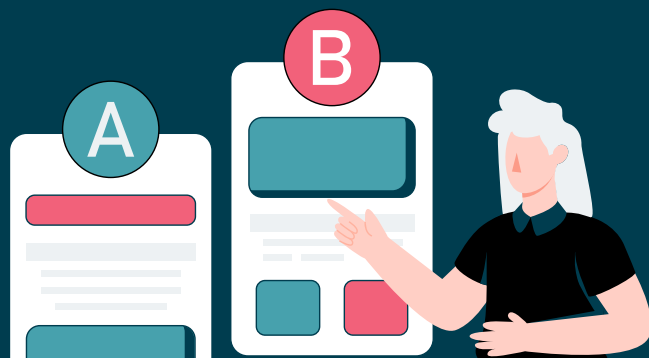
O desenvolvimento de hipóteses orientado por dados (DDHD) é uma maneira eficaz de abordar esses problemas. O teste de hipóteses é uma ferramenta poderosa que pode ajudar a reduzir o risco de um grande trabalho e deve ser usado não apenas antes, mas também durante a entrega.

Em essência, o DDHD consiste em formular hipóteses, realizar pequenos experimentos com resultados e critérios claros e usar os dados que coletamos para contar histórias e compartilhar as lições com a equipe, a empresa e as partes interessadas. O estudo de caso abaixo ilustra isso em ação, mas, por enquanto, esta é a aparência de uma hipótese:

- Acreditamos que **<esse recurso>**
- Resultará em **<este resultado>**
- Saberemos que fomos bem-sucedidos quando **<identificarmos um sinal mensurável>**.

O DDHD cria um espaço e uma estrutura para que as equipes realizem experimentos curtos, aprendam à medida que agregam valor de forma incremental e apliquem continuamente as lições aprendidas para obter maior impacto e reduzir os riscos de investimentos caros e não validados.

O teste de hipóteses é uma ferramenta poderosa que pode ajudar a reduzir o risco de um grande trabalho, e deve ser usado não apenas antes, mas também durante a entrega.



Princípio 3: Medição das métricas de entrega

Pesquisas descobriram que as organizações de tecnologia de alto desempenho têm um bom desempenho em **quatro métricas principais**:

1. prazo de entrega
2. frequência de implantação
3. tempo médio para restabelecer o serviço
4. alteração da porcentagem de falhas

As quatro métricas principais fornecem informações sobre o fluxo e o atrito da entrega de valor e são um ótimo ponto de partida para identificar o que está funcionando bem e o que precisa ser melhorado. Para as organizações que não têm as ferramentas da plataforma que permitem que as equipes meçam as quatro métricas principais, é possível começar simplesmente pesquisando as equipes regularmente com a **ferramenta de verificação rápida da DORA**. Embora a precisão não seja tão alta, esse método lhe dá uma primeira indicação de como está a sua organização e quais são as tendências.

Além disso, você também deve medir **métricas orientadas para resultados** que sejam relevantes para a sua organização, como melhoria na eficiência e satisfação da cliente. As métricas orientadas para os resultados ajudam a alinhar as equipes às atividades que contribuem para as metas organizacionais, em vez de serem um trabalho de rotina. O diagrama abaixo dá um exemplo de como as métricas orientadas para resultados podem se parecer no domínio de seguros.

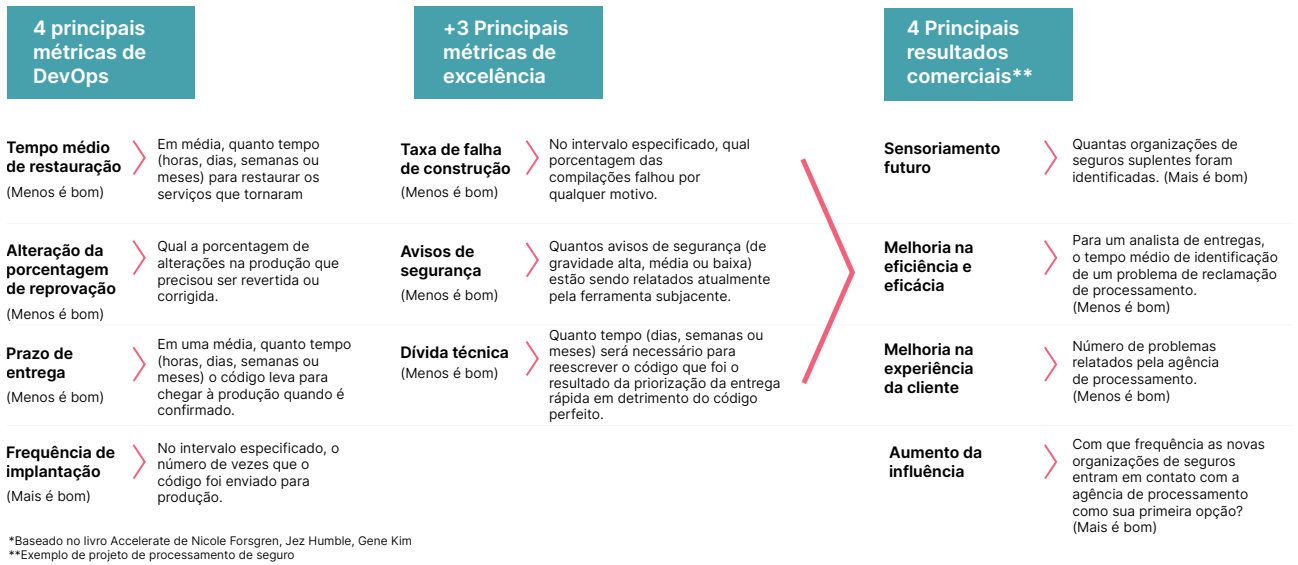


Figura 3: Medição de métricas de entrega e métricas orientadas para resultados (por exemplo: melhoria na experiência de cliente, melhoria na eficiência) ajudam as equipes a se concentrarem no trabalho impactante, em vez de no trabalho ocupado

No entanto, essas métricas devem ser sustentadas por uma cultura organizacional saudável e uma mentalidade orientada por valores. Caso contrário, podemos cair na armadilha das métricas disfuncionais. Como diz a lei de Godhart: “Quando uma medida se torna um alvo, ela deixa de ser uma boa medida”. E tenha em mente o **efeito Hawthorne** - Quando a sua equipe sabe que está sendo avaliada, pode haver um reflexo para burlar as regras e encontrar brechas para atingir as metas.

Estudo de caso: Aplicação dos princípios de planejamento de entrega para reduzir os ciclos de feedback e criar o produto certo

A Uma empresa B2B, vamos chamá-la de Empresa X, queria ajudar suas clientes a administrar e expandir seus negócios com uma nova oferta de serviços financeiros. A Empresa X sabia que, usando os dados históricos de transações de sua base de clientes, poderia oferecer uma experiência de empréstimo melhor do que a de outros provedores de serviços financeiros. Assim, trabalhamos com a Empresa X para criar um produto de dados de histórico de crédito da cliente que permitisse recomendar produtos financeiros adequados a sua base de clientes. Os clientes que consentiram que a Empresa X usasse seus dados para esse produto teriam uma experiência mais tranquila na obtenção de um financiamento adequado para suas compras comerciais de varejo.



Corte vertical em ação

A Empresa X não tinha uma plataforma de dados para extrair, processar e criar novos produtos de dados – os dados de que precisávamos para construir isso estavam isolados em repositórios de dados transacionais na produção.

Em vez de esperar que uma plataforma de dados fosse concluída (ou seja, fatiamento horizontal), aplicamos o **desenvolvimento de produtos enxutos e os princípios de malha de dados** para criar um produto de dados que fornecesse valor real para o cliente no curto prazo, ao mesmo tempo em que oferecia suporte à extensibilidade para uma plataforma compartilhada no médio e longo prazo. Isso nos permitiu fornecer rapidamente uma solução de baixo risco (em pouco mais de quatro meses) e, ao mesmo tempo, fornecer informações valiosas sobre um esforço contínuo de desenvolvimento de plataforma de dados..

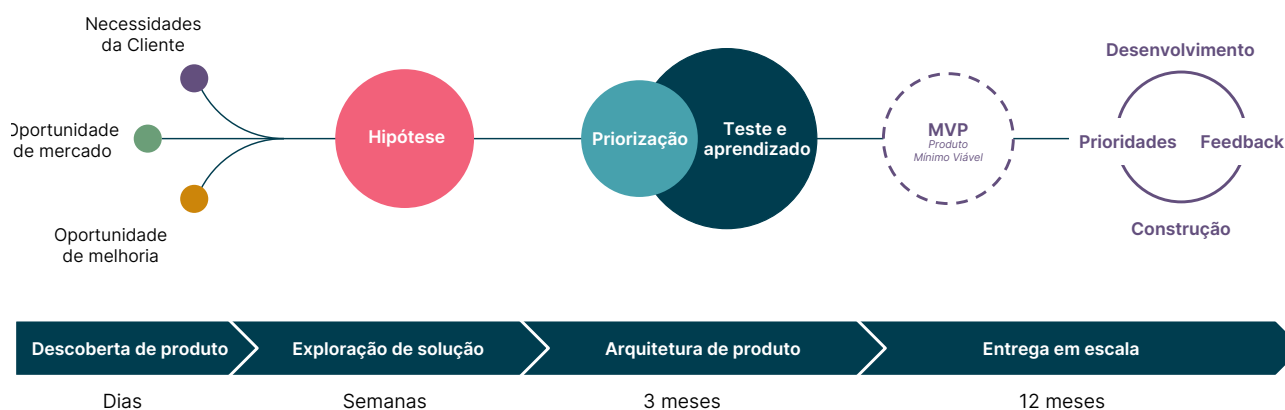


Figura 4: Construir, medir, aprender no contexto. Uma abordagem Lean para a entrega de produtos.

Fase 1: Descoberta

Com a ajuda de especialistas no domínio, refinamos os requisitos comerciais da Empresa X para delinear as fontes de dados necessárias para o produto de histórico de crédito. Descobrimos que só precisávamos de sete (de 150) tabelas do banco de dados de origem para atender aos requisitos mínimos. Isso reduziu os esforços de ingestão de dados, pois não precisávamos processar ou limpar dados desnecessários. Durante seis semanas, também refinamos os recursos e os requisitos multifuncionais do produto de dados de histórico de crédito do cliente e nos alinhamos ao valor comercial pretendido.

Articulamos hipóteses para nos ajudar a encontrar o caminho mais curto para as soluções “corretas”. Essas hipóteses nos ajudaram a permanecer no caminho certo em direção à nossa meta de criar o produto certo. Por exemplo, poderíamos validar nossa abordagem executando um experimento e coletando dados sobre uma de nossas hipóteses:

- **Acreditamos que** o estabelecimento de uma pré-triagem automatizada baseada em regras e em várias dimensões do histórico de transações de uma cliente
- **O resultado será em** uma forma escalável de identificar clientes com capacidade de crédito
- **Saberemos que fomos bem-sucedidos quando** um aplicativo de pré-triagem automatizado for capaz de rejeitar clientes que não são dignos de crédito com uma margem de erro de X% em relação às avaliações de crédito feitas por especialistas no assunto treinados profissionalmente.

Fase 2: Entrega

Quando todas as pessoas estavam alinhadas quanto à forma do produto e ao valor que ele deveria oferecer, começamos a desenvolver um produto mínimo viável (MVP). O escopo de um MVP pode ser difícil. Nosso objetivo era obter a **fatia “vertical” mais fina** que fornecesse feedback sobre a viabilidade do produto de dados e garantir que ele estivesse próximo o suficiente do produto final do ponto de vista da cliente para testar continuamente nossas hipóteses. O MVP também revelou casos de vantagem em potencial, oportunidades de produto ocultas ou perdas e possíveis obstáculos. Esse feedback inicial ajudou a identificar riscos e onde podemos concentrar nossos esforços para mitigá-los ao desenvolver mais o produto.

Isso também ajudou a definir as fontes de dados e as transformações que poderíamos aproveitar ao iterar em futuras versões do produto. Nosso foco na fase de entrega da produção foi implementar transformações bem controladas em uma infraestrutura de dados suportável e extensível e fornecer os resultados ao público consumidor do produto de dados. Aplicamos nossas práticas de engenharia padrão sensatas, como desenvolvimento orientado por testes (TDD), infraestrutura como código, CI/CD, observabilidade nos planos de código e de dados, entre outras.

Fornecimento de um produto de dados independente e abrangente

Em 10 iterações (em um período de mais de 4 meses), entregamos um produto de dados consumível, totalmente automatizado e governado de forma abrangente, sem dependência de uma plataforma de dados centralizada. A equipe também mediu as **quatro principais métricas** (por exemplo, tempo de entrega, taxa de falha de alteração) e outras **métricas de entrega** (por exemplo, velocidade, taxa de burnup etc.) para fornecer informações sobre como estamos progredindo em direção à meta. As métricas nos ajudaram a recalibrar os parâmetros de entrega quando necessário.

Amplie o impacto, reduza o tempo de entrega

Essas práticas têm ajudado a Thoughtworks a agregar valor para clientes repetidas vezes e são padrões sensatos que trazemos para cada engajamento de dados para acelerar a entrega e gerar um impacto extraordinário. Onde quer que você esteja em sua jornada de entrega agora, você pode traçar um caminho para sucesso na entrega por:

- **Conscientização:** Há alguma lacuna ou oportunidade em suas práticas atuais de planejamento de entrega?
- **Abertura para o que precisa ser mudado:** Como você aplicaria os princípios e práticas descritos neste capítulo para ajudá-lo a melhorar seu planejamento de entrega?
- **Executando a mudança:** Conecte as práticas recomendadas e testadas pelo setor com a experiência prática no fornecimento bem-sucedido de produtos de dados.

No próximo capítulo, compartilharemos como você pode economizar horas gerenciando melhor a qualidade dos seus dados.

Arquitetura para sistemas de dados: como equilibrar as compensações nas decisões sobre tecnologia

Por Simon Aubury e Kunal Tiwary

Os dados são parte integrante da entrega de cada produto e do envolvimento da cliente. Ao projetar sistemas de dados, você precisa entender tanto os dados quanto a tecnologia, ao mesmo tempo em que aprecia o valor final que um produto trará as clientes. Não importa o tamanho do seu produto de dados, estabelecer padrões sensatos ajuda a equilibrar as compensações de decisões tecnológicas específicas.

Avaliar as práticas recomendadas de gerenciamento de dados é um bom ponto de partida que pode orientar suas escolhas de design de dados. Como uma pessoa engenheira de dados, você deve selecionar ferramentas tecnológicas que funcionem bem em conjunto para o projeto e são adequadas para as necessidades mais amplas de dados em toda a organização. Em sua essência, toda decisão tecnológica precisa ser orientada pelo valor que proporcionará aos negócios.

Neste capítulo, abordamos os princípios básicos para você começar e algumas considerações para equilibrar as vantagens e desvantagens da tecnologia

Princípios padrão

Considerando o complexo cenário tecnológico e de dados de hoje, a noção de que pode haver uma única e melhor maneira de fazer qualquer coisa parece ambiciosa. Mas as práticas padrão sensatas são um ótimo ponto de partida, pois são uma maneira eficaz de criar uma arquitetura com base em valores compartilhados. Elas também permitem que você seja agnóstico em relação à tecnologia, ao mesmo tempo em que se concentra nos elementos de um bom design. E para projetos de dados, elas podem fornecer um conjunto inicial de princípios básicos: práticas e técnicas eficazes para você começar.

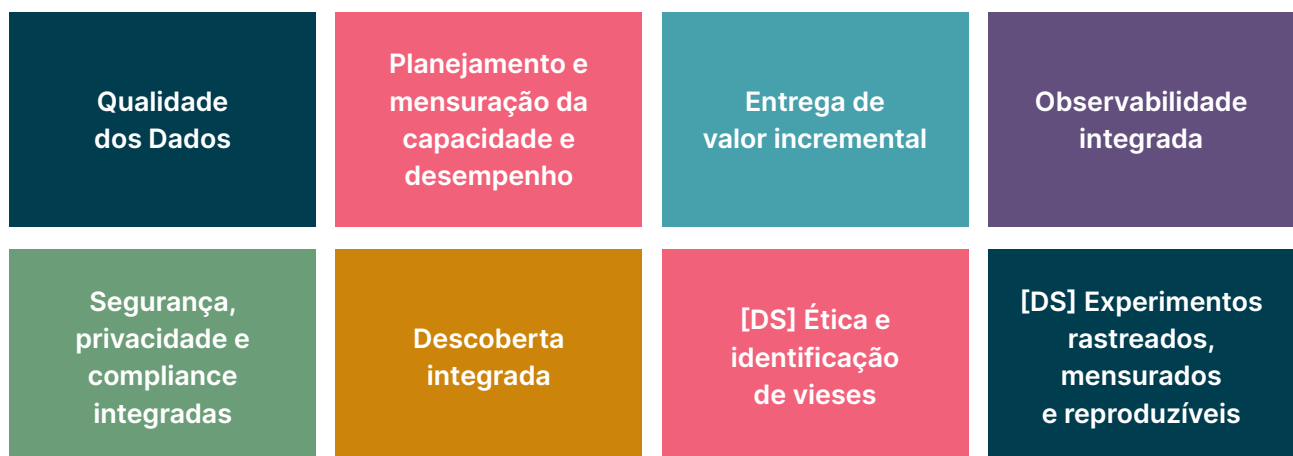


Figura 1: Práticas padrão de dados (além dos padrões sensíveis da engenharia principal)



Assim como nos projetos de software, acreditamos que é essencial estabelecer práticas recomendadas de gerenciamento de dados para plataformas de dados modernas, incluindo:

- **Data quality**
- Capacidade de descoberta de planejamento e medição de capacidade e desempenho
- Entrega de valor incremental medidos e reproduzíveis
- Observabilidade
- Segurança e conformidade
- Descoberta
- Ética e vieses
- Experimentos rastreados
- Medição da adequação arquitetônica

No entanto, podem haver circunstâncias que tornem a escolha padrão inválida ou, pelo menos, abaixo do ideal. Por exemplo, se a diferenciação do seu produto exigir leituras de latência ultrabaixa em detrimento da consistência, você precisará usar um armazenamento de dados de nicho especializado para o seu caso de uso.

Medição da adequação arquitetônica

A arquitetura de dados precisa crescer e evoluir de acordo com as necessidades da organização. **Arquitetura evolutiva** é uma abordagem que permite que a arquitetura mude de forma incremental, permitindo que sua empresa responda rapidamente a novas demandas. Para garantir que a mudança não comprometa a qualidade nem cause problemas de arquitetura, avalie como a arquitetura atende ao caso de uso original ao longo do tempo. **As funções de adequação** fornecem uma medida objetiva, informando o processo de desenvolvimento à medida que ele acontece, em vez de depois do fato.

Algumas funções de adequação para a arquitetura de dados a serem consideradas incluem:



Custo



Data latency



Volume de dados

Colocar os casos de uso antes da tecnologia

Muitas vezes, há uma fixação em rotular um problema de dados como sendo uma carga de trabalho transacional ou analítica, ou um caso de uso como sendo um sistema em tempo real ou em lote. Isso geralmente leva à caracterização de um problema comercial como “adequado” para uma tecnologia. No entanto, a tecnologia existe para dar suporte aos negócios, e não o contrário. Tomemos como exemplo o processamento de dados.

Aguardar até o fim do dia para processar os dados em blocos é como comprar o jornal para entender o que aconteceu ontem.



Os processos de negócios são semelhantes a um fluxo de eventos, e praticamente todos os dados com os quais você lida são transmitidos nesse fluxo. Os dados são quase sempre produzidos

e atualizados continuamente em sua fonte, estão sempre chegando. Esperar afinal do dia para processar os dados em lote é como comprar um jornal para ler o que está acontecendo descobrir o que aconteceu no mundo ontem. Embora para alguns casos de uso, como sistemas de faturamento e folha de pagamento, isso seja aceitável, outros exigem um processamento de dados de streaming mais imediato.

Ao projetar a arquitetura adequada, concentre-se no desenvolvimento de soluções que processem dados para atender ao resultado comercial que você está buscando. Você precisa dar um passo atrás e avaliar o que a arquitetura deve suportar. Você está criando um sistema para facilitar as “transações” (pense em vendas em um site de comércio eletrônico ou em um sistema de processamento de pagamentos)? Ou está tentando “analisar” o histórico para identificar tendências e usar dados agregados para obter insights? Comece com o problema que está tentando resolver. Em seguida, observe as características dos dados comerciais relevantes e considere como a tecnologia e a arquitetura do projeto podem oferecer melhor suporte a essas cargas de trabalho.

Embora seja fundamental usar o armazenamento de dados certo para o caso de uso certo, você quer resolver o problema, e não construir de acordo com as restrições da tecnologia. As escolhas erradas de tecnologia podem direcionar mal o esforço de engenharia e prejudicar a probabilidade de sucesso comercial futuro.

Equilibrando as vantagens e desvantagens da tecnologia

A tecnologia muda rapidamente; isso é especialmente verdadeiro para os sistemas de dados. A tecnologia selecionada precisa atender às crescentes demandas e expectativas das plataformas de dados, atender a uma ampla gama de necessidades - desde transacionais e operacionais até analíticas - e permitir a exploração interativa de dados em tempo real.

Mas encontrar a tecnologia certa pode ser demorado. Para acelerar o processo, o fundador da Amazon, [Jeff Bezos sugere](#) você não delibere sobre decisões facilmente reversíveis, do tipo “porta de mão dupla”. Simplesmente passe pela porta e veja se você gosta dela - se não gostar, volte atrás. Você pode tomar essas decisões rapidamente e até mesmo automatizá-las. Poucas ou nenhuma decisão sobre dados é difícil de reverter. Mas as que são precisam ser tomadas com cuidado.

Um [technology radar de dados](#) é uma ótima maneira de equilibrar o risco em seu portfólio de tecnologia e polinizar a inovação entre as equipes, fazendo as devidas experiências. Ele permite que você descubra que tipo de organização tecnológica você deseja ser - e avalie objetivamente as ferramentas de dados que estão funcionando e as que não estão.



Também é importante investir e usar soluções personalizadas somente quando elas forem um diferencial para a empresa ou proporcionarem uma vantagem competitiva. Considere o que é importante para sua empresa:

- Segurança, esquemas, design e linhagem – os itens não negociáveis para o design do sistema
- Um teste inicial de latência, correção e durabilidade – decidir a classificação ainda é importante
- Lote, harmonize e consolide a entrega – porque os custos vão e voltam

Agir rapidamente pode levar à duplicação, enquanto a consolidação de esforços em uma plataforma de dados centralizada eliminará a duplicação, mas levará mais tempo. É preciso entender as vantagens e desvantagens e tomar decisões sobre a arquitetura de dados em nível empresarial rapidamente. Decisões lentas podem fazer com que sua infraestrutura de dados se prolifere desnecessariamente, o que pode ser difícil de reverter se você quiser consolidar sua infraestrutura.

A otimização para padrões de dados sensatos pode facilitar a seleção da tecnologia certa e fazer boas escolhas de arquitetura. Também pode ajudá-lo a validar se o seu investimento está proporcionando à empresa a vantagem competitiva esperada, ajudando-o a tomar decisões mais informadas.

No próximo capítulo, compartilharemos como você pode implementar uma estratégia de dados eficaz que forneça as bases para o gerenciamento da qualidade dos dados.

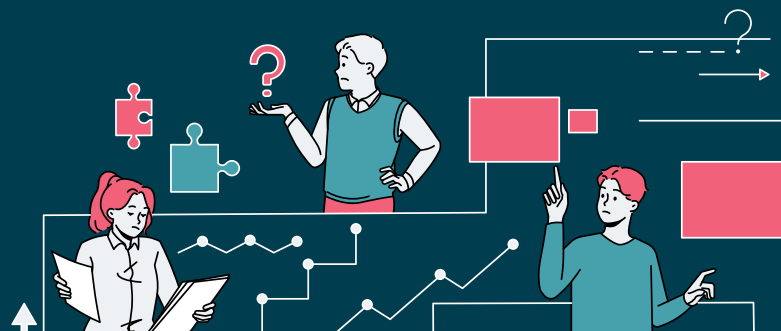
A qualidade é fundamental: encontrando o valor em sua estratégia de teste de dados

Por Simon Aubury e Kunal Tiwary

Imagine o que suas equipes de dados poderiam conseguir com dois dias a mais por semana. A ideia é empolgante, não é? Mas de onde viria esse tempo adicional? A resposta pode estar no melhor gerenciamento da qualidade de seus dados. A melhor maneira de fazer isso é detectar os problemas antecipadamente com a ajuda de uma estratégia rigorosa de teste de dados.

O impacto da mudança do teste de dados para o upstream não deve ser subestimado. Pesquisas sugerem que as equipes de dados gastam **de 30 a 40% do seu tempo concentrando-se em problemas de qualidade de dados**. Esse é um tempo significativo que elas poderiam estar gastando em atividades que geram receita, como a criação de produtos e recursos melhores ou a melhoria do acesso a insights mais rápidos e precisos em toda a organização. E, além da produtividade e da eficácia organizacional, o tempo de inatividade dos dados – causado por dados ausentes, imprecisos ou comprometidos – pode **custar milhões de dólares às empresas** e corroer a confiança da organização na sua equipe de dados como geradora de receita para a organização.

Pesquisas sugerem que as equipes de dados gastem de 30% a 40% do seu tempo concentrando-se em problemas de qualidade de dados.



Valores ausentes em conjuntos de dados podem levar a falhas nos sistemas de produção, dados incorretos podem levar à tomada de decisões comerciais erradas e alterações na distribuição de dados podem degradar o desempenho dos modelos de aprendizado de máquina. Recomendações irrelevantes de produtos podem afetar a experiência do cliente e levar a uma perda de receita. Em setores como o de saúde, as consequências podem ser muito mais significativas. Dados incorretos podem levar à prescrição de medicamentos errados, desencadeando reações adversas ou até mesmo a morte.

Neste capítulo, vamos entender como implementar uma estratégia de dados eficaz. Mas, primeiro, veremos as principais considerações que estabelecem as bases para a qualidade dos dados.

As considerações e as vantagens e desvantagens da qualidade dos dados

Inspirar a confiança da organização na qualidade de seus dados e da equipe de dados é vital. No entanto, isso só pode ser feito descrevendo o que realmente queremos dizer com qualidade de dados: quais recursos e dimensões são fundamentais para ela. Aqui estão cinco áreas a serem consideradas:



Atualidade: A importância da recenticidade dos dados é específica ao contexto. Um aplicativo de monitoramento de segurança ou de detecção de fraude requer dados muito recentes para garantir que as irregularidades sejam detectadas e possam ser tratadas rapidamente, enquanto o treinamento de um modelo de aprendizado de máquina pode tolerar dados mais latentes.



Precisão: Decisões que afetam a vida, como a eficácia de medicamentos e decisões financeiras, têm pouco espaço para dados imprecisos. No entanto, podemos sacrificar a precisão pela velocidade ao oferecer sugestões sobre uma loja on-line de varejo ou serviço de streaming.



Consistência: As definições de termos comuns precisam ser consistentes. Por exemplo, o que significa “atual” quando falamos de clientes - comprados na semana passada ou há dois meses? Ou o que constitui uma “cliente” - já registrada, autenticada, um ser humano real?



Compreensão da fonte de dados: A fonte de dados ou a forma como os dados são capturados pode afetar sua precisão. Se um representante do atendimento ao cliente de um banco selecionar um campo suspenso com pressa ou sem validação, um erro manual poderá levar a relatórios incorretos de encerramento de conta.



Metadados (incluindo linhagem): Os metadados são a base de um resultado de qualidade. Eles ajudam a caracterizar os dados e ajudam sua organização a entendê-los e consumi-los facilmente. Os metadados devem explicar quem, o quê, quando, como e por quê dos dados - eles podem até mesmo fornecer informações sobre coisas como a propriedade do código do produto de dados.

Preparação de uma estratégia de teste: A qualidade começa com uma conversa

Nossa experiência sugere que, quando as equipes produtoras de dados assumem a responsabilidade pelo processo de teste de dados, a qualidade dos dados é mantida de forma mais fácil e consistente. Mas uma estratégia de teste robusta exige a colaboração entre o público consumidor e produtores de dados. O público consumidor de dados precisam abordar as áreas abaixo para desenvolver uma estratégia de teste de dados:

- Quais recursos de qualidade são importantes: completude, distinção, conformidade ou algo mais?
- Que requisitos de negócios estamos criando?
- Qual é o valor-alvo que os produtores devem buscar e otimizar?
- Identificação de métricas de qualidade orientadas por domínio - por exemplo, as necessidades do varejo seriam bem diferentes das necessidades do setor imobiliário

As equipes produtoras de dados também devem ter como objetivo capturar métricas mais refinadas, como:

- Tolerância de erro
- Propriedade - se um controle de qualidade estiver quebrado, quem o consertará e quando?
- Quanto vale a qualidade dos dados? A adição de mais dados poderá melhorar sua análise? Quais são os limites e as tolerâncias apropriados e acordados para a qualidade dos dados da empresa? A precisão de 90%, 99%, 99,9% é esperada ou aceitável para o usuário final dos dados?
- Acordos de nível de serviço (SLAs) - quanto tempo de inatividade a empresa permite por ano?

Como proprietários da qualidade dos dados, as pessoas produtoras de dados são, em última instância, responsáveis por conhecer esses limites e atender às expectativas acordadas.

Implementação da estratégia

Com uma estratégia de testes sólida em vigor, a próxima etapa a se considerar é a implementação de testes de qualidade de dados como um padrão de gravação-auditoria-publicação (WAP). Usando esse padrão, você grava dados e audita os resultados antes de publicá-los. Isso permitirá que você faça correções antes da publicação.

A habilitação da ingestão de novos dados nos pipelines de integração contínua e de integração contínua/entrega contínua (CI/CD) também garante que os dados importados passem por testes de qualidade e não interrompam as verificações existentes. Podem haver casos em que as verificações devam interromper o pipeline e enviar um alerta de alta urgência. Se uma verificação sinalizar um número de casa negativo em um pipeline executado com dados imobiliários, por exemplo, você precisará resolver o problema imediatamente e interromper a execução. Por outro lado, se um número de casa estiver simplesmente fora do intervalo, você poderá continuar a execução do pipeline com alertas simples.

É extremamente importante disponibilizar os resultados dessas verificações de qualidade para a empresa como um todo. Por exemplo, uma lista de endereços que esteja 15% incompleta pode atrasar o lançamento da campanha da equipe de marketing. Já uma variação de 1% em uma medição de engenharia pode colocar em risco um processo de fabricação caro. Tornar os níveis de qualidade visíveis como parte de um catálogo de metadados também pode ser extremamente valioso, pois permite que o público consumidor desses dados tome decisões informadas ao considerar os casos de uso dos dados.



Tornar os níveis de qualidade visíveis como parte de um catálogo de metadados também pode ser extremamente valioso, pois permite que os consumidores de dados tomem decisões informadas ao considerar os casos de uso dos dados.

Atualmente, muitas estruturas de qualidade de dados oferecem relatórios de perfil que incluem a distribuição de erros/falhas. Você pode encontrar algumas boas estruturas de código aberto - tivemos experiências positivas com [Great Expectations](#), [Deequ](#) e [Soda](#) – que podem ajudá-la a implementar testes de qualidade de dados por meio de uma série de recursos. Dependendo do nível de integração necessário, alguns dos principais recursos a serem considerados em sua estrutura incluem:

- Usar uma solução de código aberto para evitar a dependência do fornecedor
- Como os resultados podem ser visualizados e compartilhados para garantir a transparência em toda a organização
- Implementar a validação de dados em cargas incrementais para garantir que as verificações sejam realizadas de forma contínua, independentemente da frequência de ingestão desejada
- Implementar a detecção de anomalias para identificar e gerar alertas automaticamente para desvios inesperados acima ou abaixo de uma determinada tolerância
- Integração com ferramentas de alerta e monitoramento para garantir a visibilidade do sistema sem criar integrações de observabilidade por conta própria (e acelerar o tempo de implementação das verificações)
- Integração com o catálogo de dados para evitar a criação de integrações de descoberta por conta própria e criar visibilidade desde o início
- Suporte à linguagem de programação - escolha de uma estrutura com base na pilha técnica de seu ecossistema de dados atual

Testes suficientes no momento certo

Você deve considerar a implementação de verificações de dados em vários estágios, garantindo que a qualidade seja mantida em todos os pipelines de extração, transformação e carga (ETL) e testes independentes para transformações complexas de dados. Pense nisso como uma mudança de foco da qualidade dos dados para a esquerda ou como a realização de verificações o mais cedo possível. Implemente verificações de esquema e integridade básica durante a ingestão bruta em si, bem como durante o estágio de transformação. A qualidade dos dados organizacionais melhora com a criação de diferentes camadas de testes à medida que você passa pelo pipeline de dados.

Trate seu ambiente de desenvolvimento e os testes de pipeline da mesma forma que os trataria em um ambiente de produção. Comunique-se com as equipes de aplicativos de origem sobre problemas de qualidade de dados e corrija-os no aplicativo de origem. Por exemplo, as verificações de Know Your Customer (KYC) em seu pipeline precisam ter atributos de cliente não nulos. Se o aplicativo de origem não aplicar uma validação no sistema, valores nulos/vazios serão ingeridos, tornando as transformações inúteis ou inválidas para muitas linhas de dados. O monitoramento de métricas, como contagens de linhas, totais e médias, e a definição de alertas de falha em tempo hábil também reduzirão o tempo de resolução.



Crie confiança e economize tempo por meio da qualidade dos dados

Um conjunto robusto de testes de qualidade de dados com foco em confiança, relevância e repetibilidade contribuirá muito para inspirar confiança em seus consumidores de dados e reduzir o tempo gasto com problemas de qualidade.

As organizações precisam criar uma infraestrutura que permita às equipes produtoras de dados corrigir, resolver problemas e implementar mudanças rapidamente para manter e evoluir continuamente o paradigma da qualidade dos dados. Quando fizerem isso, os times poderão começar a se concentrar mais em atividades que agregam valor, em vez de simplesmente corrigir problemas.

A medição e a comunicação da qualidade dos dados ajudará a sua organização a obter alinhamento sobre o estado e a relevância comercial dos dados em toda a organização. Isso também permitirá que as equipes de dados façam melhorias contínuas ao longo do tempo.

Mudar para a esquerda em termos de segurança e privacidade: porque é fundamental para a velocidade, qualidade e confiança da cliente

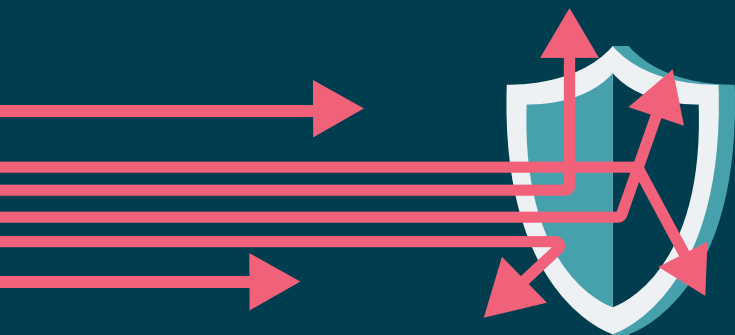
Por Mitchell Lisle e Harmeet Kaur Sokhi

No primeiro semestre de 2022, houve **817 comprometimentos de dados que afetaram mais de 53 milhões de pessoas** nos Estados Unidos. Cada uma dessas violações de segurança custou **em média US\$ 4,35 milhões** (EUA) - um aumento de 12,7% desde 2020.

Governos de todo o mundo estão implementando leis mais rígidas em relação à privacidade de dados, já que mais organizações e indivíduos são afetados por violações todos os dias. Como as equipes de engenharia desempenham um papel cada vez mais importante nesse espaço, não seria correto encerrar nosso mergulho profundo na engenharia de dados sem discutir segurança e privacidade.

Segurança e privacidade são frequentemente usadas de forma intercambiável, mas não são a mesma coisa. A segurança permite a privacidade, mas não a garante. A privacidade normalmente se refere à capacidade de uma pessoa usuária de controlar, acessar e regular suas informações pessoais, enquanto a segurança se refere ao sistema que protege esses dados para que não caiam em mãos erradas. É possível ter segurança de dados sem privacidade de dados, mas não o contrário.

Eles são igualmente importantes e qualquer bom sistema de gerenciamento de informações garantirá que os dados pessoais sejam tratados adequadamente.



A privacidade normalmente se refere à capacidade da usuária de controlar, acessar e regular suas informações pessoais, enquanto a segurança se refere ao sistema que protege esses dados de caírem em mãos erradas.

Por que mudar para a esquerda?

Com muita frequência, a segurança e a privacidade são comprometidas no início dos projetos de desenvolvimento simplesmente por serem negligenciadas. Embora isso possa significar que, no início, você possa se mover rapidamente, com o tempo você precisará investir tempo e energia significativos refatorando o software para segurança e privacidade.

Para tornar as coisas ainda mais complicadas, os desafios de fazer isso em um produto ou solução que já está em produção podem levar a riscos adicionais, pois aumentam a área de superfície para violações de segurança ou de dados. Para qualquer produto que processe dados que possam ser considerados Informações de Identificação Pessoal (PII), a segurança e a privacidade são especialmente importantes a serem consideradas desde o início de um projeto.

E é isso que queremos dizer com “deslocamento para a esquerda”. Na engenharia de software, o deslocamento para a esquerda é um esforço consciente para incorporar determinadas práticas mais cedo no ciclo de vida do desenvolvimento - sendo a esquerda o início do ciclo de vida de um produto e a direita o final.

Benefício em mover-se para a esquerda

Enquanto mantêm os valores na direita

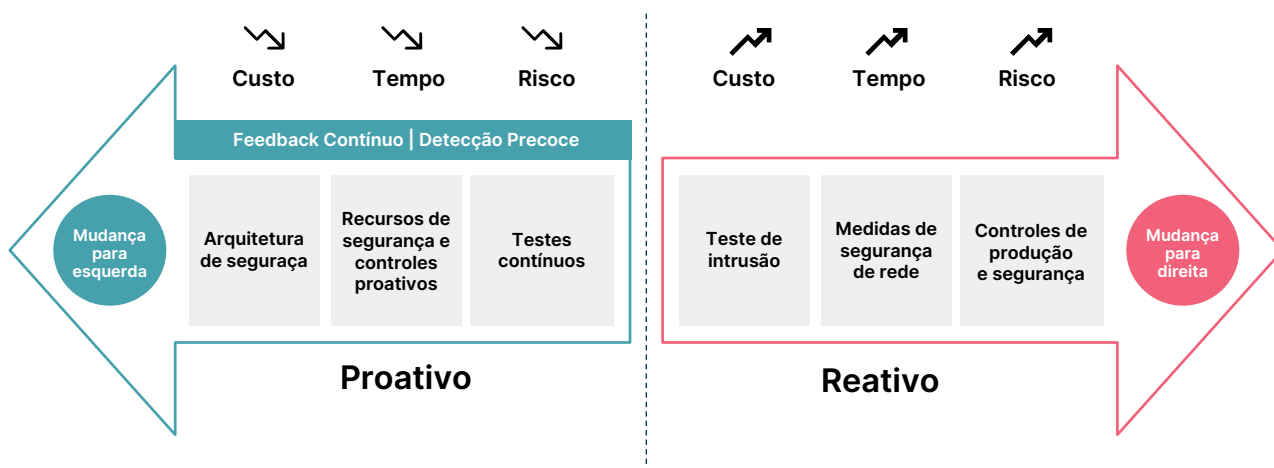


Figura 1: Valor no deslocamento para a esquerda

Muitas organizações têm engenheiros de segurança ou até mesmo um CISO (Chief Information Security Officer, diretor de segurança da informação), mas muitas delas não têm conhecimento técnico quando se trata de privacidade. Isso levou ao surgimento do **engenheiro de privacidade** – uma função especializada em engenharia de software que garante que as considerações de privacidade sejam incorporadas ao desenvolvimento do produto, em vez de serem deixadas para depois. A importância dessa função para as organizações se intensificou, em parte porque hoje há cada vez mais requisitos legislativos com os quais as práticas, os processos e os produtos devem estar em conformidade. A maior conscientização sobre a dimensão ética da tecnologia também torna a função particularmente valiosa. No passado, a coleta de dados era algo feito com pouca consideração pela privacidade pessoal dos usuários. Entretanto, hoje a privacidade pode ser um diferencial: há **muitos exemplos** de privacidade sendo colocada no centro do desenvolvimento de um produto.

Da mesma forma que nós, como pessoas desenvolvedoras, pensamos sobre a dívida técnica, também precisamos começar a prestar atenção à nossa **“dívida de privacidade”**. Com o aumento das violações de dados, as empresas têm que tomar uma decisão sobre quando lidar com essa dívida. Aquelas que conseguirem mudar para a esquerda em termos de segurança e privacidade reduzirão significativamente a probabilidade de acabar em uma **lista crescente de empresas que não conseguiram proteger seus dados**.

Protegendo os dados

Informações de identificação pessoal (PII) são dados que identificam diretamente, isoladamente ou em combinação com outros dados, um indivíduo. Sem a segurança adequada, os hackers podem acessar esses dados e criar perfis, usando essas informações para se passar por eles ou vendê-las a outros criminosos.

É preciso ponderação ao armazenar qualquer forma de dados pessoais. Pare e reflita : a coleta de PII é necessária para proporcionar a experiência da cliente? Você pode manter a confiança? Por exemplo, um sistema de recomendação para a indústria de varejo pode oferecer uma experiência personalizada para uma ampla faixa etária sem capturar a data de nascimento da cliente. Para as organizações que usam PII, a capacidade de encontrar e identificar dados confidenciais é fundamental para proteger sua base de clientes e sua reputação. Tecnologias como catálogos de dados e estruturas de governança apropriadas são particularmente úteis aqui para garantir que os dados possam ser organizados e protegidos de forma eficaz.

É importante observar que o simples fato de ofuscar ou mascarar os campos de PII em um conjunto de dados não necessariamente retira a identificação dos dados de um indivíduo. Pode ser possível identificar novamente os dados **usando outras informações contextuais**. Por exemplo, os hackers usam a exclusividade como um caminho para explorar vulnerabilidades, portanto, é importante conhecer todas as maneiras pelas quais seus dados podem ser exclusivos de um indivíduo. Um motorista de um carro de cor diferente em uma cidade grande pode ser bastante exclusivo, mas esse mesmo motorista em uma cidade pequena do interior seria facilmente identificado. Isso também se aplica às abordagens de aprendizado de máquina que são treinadas em dados com outliers. Excedentes podem revelar informações confidenciais sobre seus dados e podem vazá-los inadvertidamente por meio de uma API de previsão.

Uma das complexidades que as organizações enfrentam é a necessidade de acesso a PII para permitir que as equipes de desenvolvimento façam experimentos e testes enquanto trabalham. Os ambientes de teste geralmente não estão sujeitos à mesma segurança e privacidade, pois não contêm dados de produção. Mas um fluxo de trabalho de ciência de dados precisa ter dados que representem a produção, para que você possa treinar modelos e fazer análises para entender o que esse modelo fará na produção.

Há muitas maneiras de fazer isso sem dar acesso direto aos dados de produção:

1. Gerar dados falsos que correspondam ao esquema de seus dados de produção. Para alguns casos de uso, como verificações de validação de dados, isso pode ser suficiente para ajudar a garantir que os pipelines funcionem adequadamente e sem erros. Mas se o seu objetivo é treinar e lançar um modelo na produção, você deve evitar dados falsos. Não treinar seu modelo com dados que sejam o mais próximo possível dos reais gera preocupações éticas e de precisão.
2. Avalie se **dados sintéticos**, ou os dados representativos do cenário real, mas gerados por um modelo, podem funcionar para seu caso de uso. Talvez você ainda precise aplicar técnicas adicionais de preservação da privacidade sobre esses dados.
3. Gerar um subconjunto de dados seguros e anônimos. A anonimização é uma tarefa desafiadora e, às vezes, impossível quando se trata de PII. A simples remoção de campos óbvios, como nomes, endereços e outros identificadores, não significa que você não possa identificar novamente um indivíduo nesse conjunto de dados. É importante ter um bom entendimento das práticas de engenharia de privacidade, como mascaramento, privacidade diferencial e computação criptografada, para fazer isso de forma eficaz.

4. Criar um ambiente seguro e isolado especificamente para a criação de modelos e treinamento com acesso a uma cópia dos dados de produção. Essa abordagem é mais cara e introduz riscos, pois você está copiando dados em outro local. Você também precisará de um ambiente separado com todos os mesmos controles de segurança e privacidade da produção.

A maior mudança que você precisa fazer para melhorar a segurança e a privacidade é pensar pequeno. **A minimização de dados** é sua amiga - ajudando-o a criar o que deseja, usando o menor subconjunto de dados de que realmente precisa.

Adotar práticas de segurança sensatas

Há várias práticas que podem ser implementadas para ajudá-lo a tomar decisões quando se trata de desenvolver produtos de dados seguros.

1. A **nomeação de defensores da segurança** com experiência em atividades e processos de segurança ajudará a orientar as equipes de desenvolvimento sobre as decisões corretas a serem tomadas.
2. Implemente um **processo de classificação de dados** que lhe permita marcar dados confidenciais e aplicar políticas de governança em toda a organização com base na sensibilidade dos seus dados. Aqui está um modelo mental simples para seus dados quando eles entram em seus sistemas.



Figura 2: Mentalidade enxuta

3. Realize **workshops de segurança** frequentes, como **modelagem de ameaças**, para que você possa causar impacto de forma incremental, priorizando o trabalho à medida que avança. Concentre-se em mudanças pequenas e práticas que possam ser feitas para garantir que essas sessões continuem a agregar valor à sua infraestrutura de segurança.

4. Use a **tríade CIA** (Confidencialidade, Integridade, Disponibilidade) (*Confidentiality, Integrity, Availability*) para pensar sobre segurança:

- **Confidencialidade:** O ativo não pode ser acessado por pessoas ou sistemas que não deveriam acessá-lo.
- **Integridade:** O ativo não pode ser alterado por pessoas ou sistemas que não deveriam alterá-lo.
- **Disponibilidade:** Todas as pessoas e sistemas que deveriam ser capazes de acessar um ativo podem fazê-lo.

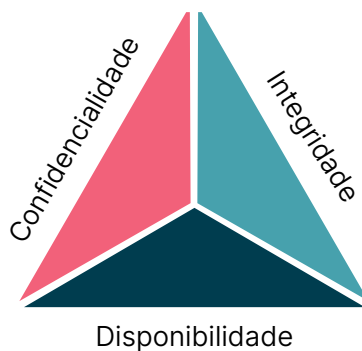


Figura 3: Tríade CIA

Aqui estão algumas considerações de segurança em cada fase do ciclo de vida de um projeto:

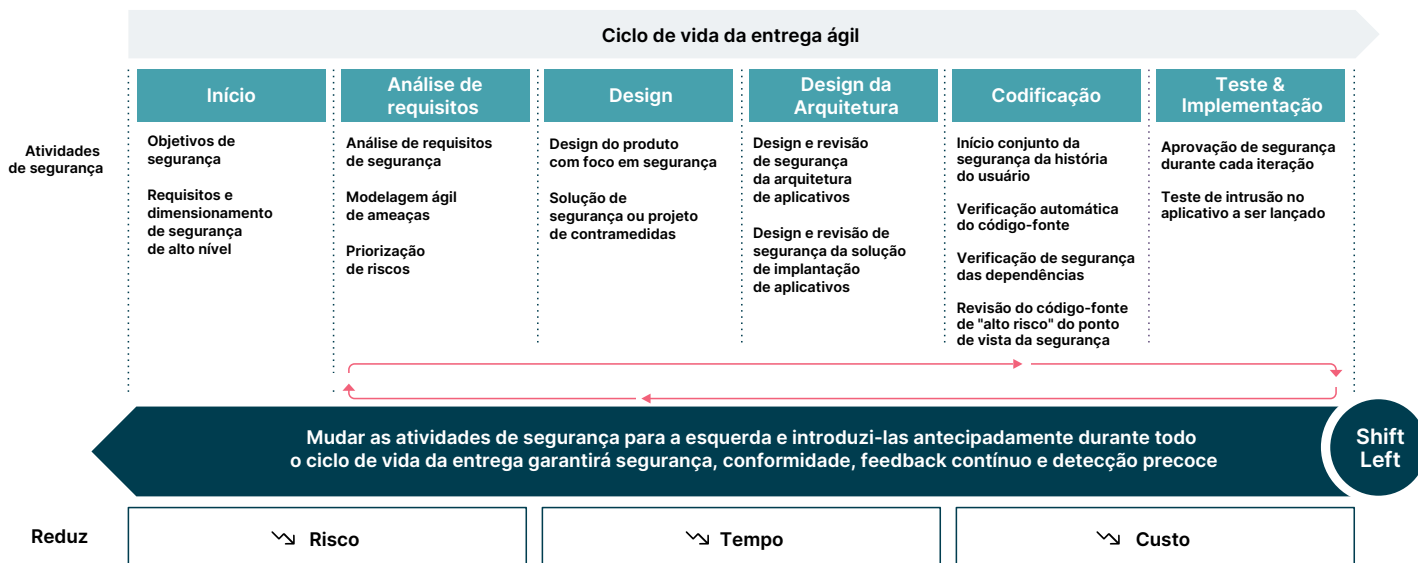


Figura 4: Fases do ciclo de vida de um projeto

Leia nosso manual de tecnologia responsável para obter mais ferramentas, técnicas e princípios úteis. Use-o para informar sua próxima sessão de planejamento e para abordar melhor as considerações críticas de segurança e privacidade; muitas das ferramentas e estruturas encontradas nele o ajudarão a avaliar seus próprios riscos.

Investir em segurança desde o início

A inclusão da segurança e da privacidade no seu processo melhorará a qualidade do seu software e permitirá que você avance mais rapidamente sem ter que depender de grandes refatorações para garantir que seus sistemas estejam de acordo com o padrão.

Quanto mais tarde você deixar a segurança e a privacidade no processo de desenvolvimento, maior será o risco de uma violação de dados significativa ou de um incidente de segurança. Investir em tudo o que é necessário para incorporar a segurança e a privacidade em seu sistema vale a pena no longo prazo: para sua organização e sua base de clientes.



Tudo pronto para desbloquear todo o potencial dos dados?

O potencial dos dados é indiscutível. No entanto, em um cenário acelerado e em constante evolução, você precisa das práticas corretas de engenharia de dados para aproveitá-los e dar à sua organização uma vantagem competitiva.

A engenharia de dados moderna ajuda você a obter o máximo de valor de seus dados, a tomar decisões melhores e a criar experiências mais personalizadas para clientes - com rapidez.

Mas, para realmente aproveitar as práticas modernas, sua empresa precisa adotar uma nova maneira de pensar sobre os dados. A mudança para uma mentalidade de produto de dados exige uma mudança cultural em toda a organização. Talvez também seja necessário realinhar as estruturas e plataformas internas para capacitar as equipes de dados.

Porque uma equipe capacitada com o conjunto certo de habilidades, experiência e conhecimento será capaz de desenvolver as soluções certas para os problemas certos mais rapidamente - e prosperar.

E, ao aplicar padrões sensatos às suas práticas e princípios, suas equipes de dados podem agregar mais valor, reduzir os riscos do projeto e encontrar oportunidades de melhoria, além de fornecer produtos de dados com rapidez.

A mudança para práticas modernas de engenharia de dados pode ser complexa e levar tempo. Na Thoughtworks, ajudamos muitas organizações a fazer essa mudança - e a desbloquear o verdadeiro potencial de seus dados em escala. Se precisar de ajuda com qualquer uma das áreas discutidas neste manual ou quiser compartilhar sua experiência com a engenharia de dados moderna, entre em contato agora.

solutions@thoughtworks.com

[thoughtworks.com/pt-br/what-we-do/data-and-ai](https://www.thoughtworks.com/pt-br/what-we-do/data-and-ai)

Sobre a Thoughtworks

A Thoughtworks é uma consultoria global de tecnologia que integra estratégia, design e engenharia de software para alavancar a inovação digital. Somos mais de 12,5 mil pessoas distribuídas entre 50 escritórios e em 18 países. Há mais de 30 anos, trabalhamos junto a nossas clientes para criar impacto extraordinário, usando a tecnologia como diferenciador para ajudá-las a resolver problemas de negócio complexos.

[thoughtworks.com/pt-br/contact](https://www.thoughtworks.com/pt-br/contact)

Entre em contato!