Fundamentos da Qualidade de Dados

Guia prático para criar pipelines de dados confiáveis

Barr Moses, Lior Gavish, e Molly Vorwerck

ALTA BOOKS
GRUPO EDITORIAL
Rio de Janeiro, 2024

Sumário

CAPÍTULO 1

Por que a Qualidade de Dados Merece Atenção Imediata

Levante a mão (ou engasgue com o café, suspire profundamente e balance a cabeça) se o seguinte cenário lhe parece familiar.

Os dados são prioridades para sua CEO [Diretora-executiva], como ocorre muitas vezes com empresas essencialmente digitais, e ela domina as melhores e mais recentes ferramentas de business intelligence. Já seu CTO [Diretor de Tecnologia] está animado com a migração para a nuvem e envia constantemente artigos à sua equipe, ressaltando as métricas de desempenho relacionadas a algumas das tecnologias mais atuais. Em contrapartida, seus usuários de dados [data consumers], incluindo analistas de produtos, líderes de marketing e equipes de vendas, dependem de ferramentas orientadas a dados, como plataformas de gerenciamento de relacionamento com o cliente/experiência do cliente (CRMs/CXPs), sistemas de gerenciamento de conteúdo (CMSs) e qualquer outro acrônimo que se possa imaginar para realizar o trabalho de forma rápida e eficaz.

Como analista ou engenheiro de dados responsável por gerenciar esses dados e torná-los utilizáveis, acessíveis e confiáveis, você raramente passa um dia sem que seja necessário atender a alguma solicitação de suas partes interessadas. Mas o que acontece quando os dados estão errados?

Por um acaso, você já esteve prestes a encerrar o expediente após um longo dia executando queries ou criando pipelines de dados, quando seu chefe de marketing avisa que "existem dados ausentes" em um relatório decisivo? Já recebeu um e-mail alucinado do seu CTO sobre "dados duplicados" em um dashboard de business intelligence? Ou um recado de sua CEO, sempre tão otimista em relação aos dados, sobre algum número confuso ou sem acurácia em seus últimos slides de uma apresentação ao conselho da empresa? Se qualquer uma dessas situações lhe caiu como uma luva, você não é o único.

Esse problema, muitas vezes chamado de "data downtime [tempo de inatividade de dados, em tradução livre]", ocorre até mesmo com as companhias mais inovadoras e que priorizam os dados e, em nossa opinião, é um dos maiores desafios enfrentados pelas empresas no século XXI. O tempo de inatividade de dados diz respeito aos períodos de tempo em que os dados estão ausentes, imprecisos ou errados e são apresentados em dashboards obsoletos,

em relatórios inexatos e até mesmo em tomadas inadequadas de decisão. Qual é a causa do tempo de inatividade de dados? Dados não confiáveis, uma infinidade deles.

O tempo de inatividade de dados pode custar às empresas mais de centenas de milhões de dólares por ano (*https://oreil.ly/FF8kC)* sem mencionar a confiança do cliente. Na verdade, em 2019, a ZoomInfo descobriu que uma em cada cinco empresas perdeu um cliente devido a problemas relacionados à qualidade de dados (termo também conhecido como data quality, conforme mencionado no início deste livro).

Como todos devem saber, os resultados financeiros da sua empresa não são os únicos afetados pelo tempo de inatividade de dados. Lidar com problemas de qualidade de dados consome mais de 40% do tempo de uma equipe de dados (*https://oreil.ly/HEpED*), que poderia ser gasto em projetos mais interessantes ou na busca de inovações para a empresa. É bem provável que você não tenha ficado surpreso com essas informações. Nós com certeza não ficamos.

Anteriormente, Barr Moses era vice-presidente de operações em uma empresa customer success de software. Sua equipe era responsável por gerenciar os relatórios dos clientes maiores, desde a geração de dashboards para o CEO usar durante as reuniões gerais até a definição de estratégias a fim de reduzir a rotatividade de clientes com base nas métricas do usuário. Ela era responsável por gerenciar as operações de dados da empresa e garantir que as partes interessadas fossem bem-sucedidas ao trabalhar com dados. Barr nunca esquecerá o dia em que retornou à sua mesa depois de passar horas em uma sessão extenuante de planejamento e se deparou com o seguinte post-it "Os dados estão errados" no monitor de seu computador. Uma descoberta não apenas constrangedora; infelizmente, também não era incomum. Vez após vez, ela e sua equipe se deparavam com esses pequenos problemas silenciosos, mas potencialmente nocivos, em seus dados.

Deveria existir uma solução melhor.

Há décadas, as organizações enfrentam problemas relacionados à baixa qualidade de dados e a dados não confiáveis, quer sejam causados por relatórios precários, informações falsas ou erros técnicos. E à medida que elas utilizam cada vez mais dados e constroem ecossistemas e infraestruturas de dados cada vez mais complexos esse problema tende a aumentar.

O conceito de "dados ruins" e de dados de baixa qualidade existe desde a existência dos humanos, embora de formas diferentes. A baixa qualidade de dados (ou melhor, a tomada de decisões desinformadas pelos dados) levou o capitão Robert Falcon Scott e outros primeiros exploradores a preverem inadequadamente onde e quanto tempo demorariam para chegarem ao Polo Sul, seu destino-alvo.

Na história recente, diversos casos também merecem destaque. Um deles foi o famigerado acidente da sonda Mars Climate Orbiter, em 1999. A sonda foi destruída por causa de um erro de entrada de dados: as saídas foram geradas em unidades que não faziam parte do Sistema Internacional de Unidades (SI), fazendo com que a sonda se aproximasse demais do planeta. Esse acidente custou à NASA a quantia colossal de US$125 milhões. Assim como as espaçonaves, os pipelines analíticos podem ser extremamente vulneráveis às mudanças

mais inofensivas em qualquer etapa do processo. Outro exemplo seria a crise financeira de 2008, estimulada em parte por dados imprecisos que sobreavaliavam a quantidade de títulos com garantia hipotecária e outros derivativos. E isso é apenas a ponta do iceberg.

O infeliz incidente com o post-it de Barr a fez pensar: "Não devo ser a única a passar por isso!" Com Lior Gavish, Barr começou a investigar a causa-raiz do problema de "tempo de inatividade de dados". Juntos, entrevistaram centenas de equipes de dados sobre os maiores problemas e a qualidade de dados aparecia repetidamente como problema principal. Do e-commerce aos serviços de assistência médica, empresas de diversos setores enfrentavam problemas semelhantes: mudanças de esquema estavam fazendo com que pipelines de dados quebrassem, linhas ou colunas duplicadas apareciam em relatórios críticos de negócios e dados desapareciam dos dashboards, ocasionando perda substancial de tempo, dinheiro e recursos para corrigi-los. Eles também se deram conta de que era necessário uma forma melhor de comunicar e abordar os problemas relacionados à qualidade de dados, como parte de um ciclo iterativo para melhorar a confiabilidade de dados — e criar uma cultura orientada à data trust. Essas entrevistas nos inspiraram a escrever este livro. Assim, podemos difundir algumas das melhores práticas que aprendemos e desenvolvemos, relacionadas à gestão da qualidade de dados em cada etapa do pipeline de dados, desde a ingestão até a análise, e podemos compartilhar como equipes de dados em situações semelhantes podem evitar o tempo de inatividade de dados.

Nesta obra, "dados em produção" se refere a dados oriundos de sistemas de origem (como CRMs, CSMs, bancos de dados ou qualquer uma das outras analogias mencionadas anteriormente), ingeridos por data warehouses, data lakes ou outras soluções de armazenamento e processamento de dados, assim como fluxos por meio de pipeline de dados (ETL: Extrair, Transformar e Carregar), apresentados pela camada de análise para usuários de negócios. Os pipelines de dados podem manipular dados em lote e dados em streaming, e, em alto nível, os métodos para mensurar a qualidade de dados em qualquer tipo de ativo são praticamente os mesmos.

O tempo de inatividade de dados traça paralelos com a engenharia de software e DevOps, um mundo em que o tempo de disponibilidade [uptime] ou de indisponibilidade [downtime] das aplicações (ou seja, com que frequência o software ou o serviço fica "disponível" ou "ativo" ou "indisponível" ou "inativo") é calculado minuciosamente para garantir que o software seja acessível e performático. Muitos engenheiros de confiabilidade de site usam o "tempo de disponibilidade" como medida, pois se correlaciona diretamente com o impacto de baixo desempenho do software para o cliente nos negócios. Em um mundo no qual "cinco dígitos nove" (ou seja, 99,999% de disponibilidade) de confiabilidade está se tornando o padrão da indústria, como podemos aplicar isso aos dados?

Neste livro, abordaremos como as equipes modernas de dados podem desenvolver tecnologias, equipes e processos mais resilientes para garantir a alta qualidade e confiabilidade de dados em toda a organização.

E neste capítulo, começaremos definindo o que significa qualidade de dados no contexto deste livro. Em seguida, delimitaremos o momento atual a fim de entender melhor por que

a qualidade de dados é mais importante do que nunca para os líderes de dados. Finalmente, analisaremos como as melhores equipes da categoria podem alcançar elevada qualidade de dados em cada estágio do pipeline de dados e o que é necessário para manter o data trust em escala. Nosso foco principal é a qualidade de dados que alimentam os pipelines e sistemas de produção, ao contrário das plataformas de ciência de dados ou outros trabalhos mais voltados à área de pesquisa acadêmica.

O Que É Qualidade de Dados?

O conceito de qualidade de dados não é inédito — "qualidade de dados" existe desde que os humanos começaram a coletar dados!

Nas últimas décadas, no entanto, a definição de qualidade de dados começou a se concretizar como função de mensurar a confiabilidade, a completude [data completeness] e a acurácia de dados em relação ao estado do que está sendo informado. Como dizem, não podemos gerenciar o que não podemos mensurar, e alta qualidade de dados é o primeiro estágio de qualquer programa robusto de análise. A qualidade de dados também é uma forma extremamente poderosa de entender se nossos dados atendem às necessidades do negócio.

Neste livro, definimos a qualidade de dados como o estado dos dados em qualquer estágio de seu ciclo de vida. A qualidade de dados pode sofrer impacto em qualquer estágio do pipeline de dados, antes da ingestão, na produção ou mesmo durante a análise. Em nossa opinião, a qualidade de dados muitas vezes tem má reputação. As equipes de dados sabem que precisam priorizá-la, mas o conceito não parece tão natural quanto "aprendizado de máquina", "ciência de dados" ou até mesmo "analytics [inteligência analítica]", e muitas equipes não têm a capacidade ou os recursos para contratar alguém em tempo integral para gerenciá-la. As empresas com recursos limitados dependem dos analistas e dos engenheiros de dados para gerenciá-la, desviando-os de projetos, considerados mais interessantes ou inovadores.

Mas se não podemos confiar nos dados e nos produtos de dados fornecidos, como os usuários de dados podem confiar em sua equipe para agregar valor? A expressão "nenhum dado é melhor do que dados ruins" é muito usada por profissionais do setor, e embora tenha mérito, geralmente não é realidade. Problemas de qualidade de dados (ou tempo de inatividade de dados) são basicamente inevitáveis, considerando o ritmo de crescimento e consumo de dados da maioria das empresas. Contudo, ao entendermos como definimos a qualidade de dados, fica mais fácil mensurar e impedir problemas nas etapas seguintes.

Delimitando o Momento Atual

Equipes técnicas têm rastreado e buscado melhorar a qualidade de dados há tanto tempo quanto têm acompanhado o analytical data, mas somente em 2020 a qualidade de dados se tornou prioridade de faturamento para muitas empresas. À medida que os dados se tornam não apenas saída, como também commodity financeira para muitas organizações, é importante que essas informações sejam confiáveis.

Como resultado, as empresas estão cada vez mais tratando dados como código, aplicando estruturas e paradigmas há muito consagrados entre as equipes de engenharia de software às organizações e às arquiteturas de dados. As Operações de Desenvolvimento (DevOps), área técnica dedicada a reduzir o ciclo de vida de desenvolvimento de sistemas, deu origem às melhores e principais práticas do setor, como a engenharia de confiabilidade de sites (SRE), CI/CD (integração contínua/entrega contínua) e arquiteturas baseadas em microsserviços. Em resumo, o objetivo do DevOps é lançar software mais confiável e com melhor desempenho por meio da automação.

Nos últimos anos, mais empresas têm aplicado esses conceitos aos dados na forma de "DataOps". DataOps se refere ao processo de melhorar a confiabilidade e desempenho dos dados por meio de automação, reduzindo os silos de dados e fomentando análises mais rápidas e tolerantes à falhas.

Desde 2019, empresas como Intuit (*https://oreil.ly/NhMtB*), Airbnb (*https://oreil.ly/fbHlY*), Uber (*https://oreil.ly/0GhQC*) e Netflix (*https://oreil.ly/Ai2zC*) têm escrito prolificamente sobre o compromisso de garantir dados confiáveis e altamente disponíveis para as partes interessadas em toda a empresa, aplicando as melhores práticas de DataOps. Além de alimentar a tomada de decisões baseada em analytics (ou seja, estratégia de produto, modelos financeiros, estratégia growth marketing etc.), os dados gerados por essas empresas alimentam suas aplicações e serviços digitais. Dados imprecisos, ausentes ou errados podem custar tempo, dinheiro e a confiança dos clientes.

À medida que os gigantes tecnológicos destacam a importância e os desafios de alcançar alta qualidade de dados, outras empresas de todos os tamanhos e setores estão começando a prestar atenção e a replicar essas iniciativas, desde a implementação de testes mais robustos até o investimento em melhores práticas de DataOps, como monitoramento e observabilidade de dados. Mas o que levou a essa necessidade de maior qualidade de dados? O que mudou no cenário de dados para facilitar o surgimento do DataOps e, assim, a ascensão da qualidade de dados? A seguir, investigaremos a fundo essas questões.

Entendendo a "Ascensão do Tempo de Inatividade de Dados"

Com foco maior na monetização de dados, aliado ao desejo constante de aumentar a acurácia de dados, precisamos compreender melhor alguns dos fatores que podem resultar no tempo de inatividade de dados. A seguir, analisaremos as variáveis que podem impactar nossos dados.

Migração para a nuvem

Há 20 anos, um data warehouse (lugar para transformar e armazenar dados estruturados) provavelmente ficaria em um porão de escritório, não na AWS ou no Azure. Atualmente, devido à ascensão do analytics baseado em dados, equipes de dados multifuncionais e, o mais importante, a nuvem, as soluções de data warehousing em nuvem, como Amazon Redshift, Snowflake e Google BigQuery, tornaram-se opções cada vez mais populares para empresas com alta demanda por dados. Em muitos aspectos, a nuvem facilita o gerenciamento e o acesso a uma variedade maior de usuários e processamento mais rápido.

Pouco tempo após os data warehouses migrarem para a nuvem, os data lakes (lugar para transformar e armazenar dados não estruturados) também migraram, fornecendo às equipes de dados ainda mais flexibilidade quando se trata de gerenciar seus ativos de dados. À medida que as empresas e seus dados migraram para a nuvem, a tomada de decisão baseada em analytics (e a necessidade de dados de alta qualidade) tornou-se prioridade ainda maior para os negócios.

Mais fontes de dados

Hoje em dia, as empresas usam de dezenas a centenas de fontes de dados internas e externas a fim de gerar análises e modelos de aprendizado de máquina. Qualquer uma dessas fontes pode mudar inesperadamente e sem aviso prévio, comprometendo os dados usados para tomar decisões. Por exemplo, uma equipe de engenharia pode fazer uma mudança no site da empresa, modificando, assim, a saída de um conjunto de dados fundamental para as análises de marketing. Como resultado, métricas importantes de marketing podem estar erradas, levando a firma a tomar decisões insatisfatórias sobre campanhas publicitárias, metas de vendas e outros projetos indispensáveis que geram receita.

Pipelines de dados cada vez mais complexos

Os pipelines de dados têm se tornado cada vez mais complexos, com múltiplos estágios sofisticados de processamento e dependências entre diversos ativos de dados, consequência de ferramentas mais avançadas (e díspares), mais fontes de dados e maior diligência proporcionada aos dados pela liderança executiva. No entanto, sem visibilidade dessas dependências, qualquer mudança efetuada em um conjunto de dados pode ter consequências imprevistas, impactando a veracidade dos ativos dependentes de dados.

Em suma, muita coisa acontece em um pipeline de dados. As fontes de dados passam por extração, ingestão e transformação. São carregadas, armazenadas, processadas e entregues, dentre outras possíveis etapas, com muitas APIs e integrações entre diferentes estágios do pipeline. Em cada circunstância existe oportunidade para o tempo de inatividade de dados, assim como existe oportunidade para tempo de inatividade da aplicação sempre que fazemos o merge do código. Além do mais, as coisas podem dar errado mesmo quando não se trata de circunstâncias críticas. Por exemplo, quando os dados são migrados entre warehouses ou são inseridos manualmente em um sistema de origem.

Equipes de dados mais especializadas

Como dependem cada vez mais de dados para orientar a tomada inteligente de decisões, as empresas estão contratando mais analistas de dados, cientistas de dados e engenheiros de dados para criar e fazer a manutenção de pipelines de dados, análises e modelos de aprendizado de máquina, que alimentam seus serviços, produtos e operações de negócio.

Apesar de os analistas de dados serem os principais responsáveis por coletar, limpar e consultar conjuntos de dados a fim de ajudar as partes interessadas a gerar insights acionáveis sobre o negócio, os engenheiros de dados são responsáveis por garantir que as tecnologias e os sistemas fundamentais que alimentam essas análises sejam rápidos, performáticos e con-

fiáveis. No setor, os cientistas de dados normalmente coletam, processam o data wrangling e o data augmentation, além de darem sentido aos dados não estruturados para melhorar os negócios. A distinção entre analistas de dados e cientistas de dados pode ser um pouco vaga, já que o nome do cargo e as responsabilidades variam com frequência, dependendo das necessidades da empresa. Por exemplo, no final da década de 2010, a Uber mudou o nome do cargo de todos os analistas de dados para cientistas de dados após reestruturação organizacional.

À medida que os dados se tornam cada vez mais os alicerces das empresas, as equipes de dados tendem a crescer. Na verdade, empresas grandes podem custear funções adicionais, como data stewards [gestores de dados de negócios, em tradução livre], líderes de governança de dados, analistas de operações e até mesmo analytics engineers [engenheiros analíticos, em tradução livre] — função híbrida de engenheiro de dados e analista, popular em startups e empresas de médio porte que não têm recursos para sustentar uma equipe grande de dados. Como muitos usuários diferentes acessam os dados, a falta de comunicação ou coordenação insuficiente é inevitável e fará com que esses sistemas complexos quebrem à medida que as mudanças são efetuadas. Por exemplo, um campo novo adicionado a uma tabela de dados por uma equipe pode fazer com que o pipeline de outra equipe falhe, resultando em dados ausentes ou parciais. Quando se trata de downstream, dados ruins podem levar a milhões de dólares em perda de receita, perda da confiança do cliente e até mesmo pôr em risco os processos de compliance.

Equipes descentralizadas de dados

À medida que os dados se tornam essenciais às operações de negócios, mais equipes funcionais em toda a empresa se envolvem na gestão e análise de dados para otimizar e acelerar o processo de obtenção de insights. Como resultado, cada vez mais equipes de dados estão adotando o modelo distribuído e descentralizado, que imita a migração generalizada da arquitetura monolítica para a arquitetura de microsserviços, que dominou o mundo da engenharia de software em meados dos anos 2010.

O que é uma arquitetura descentralizada de dados? Não devemos confundi-la com data mesh (*https://oreil.ly/Vga7I*), paradigma organizacional que faz uso de um design distribuído e orientado a domínio. A arquitetura descentralizada de dados pressupõe que as equipes de dados trabalhem a partir de uma infraestrutura de dados centralizada (ETL) gerenciada por uma equipe de plataforma de dados, com equipes analíticas e de ciência de dados distribuídas pela empresa. Progressivamente, descobrimos que cada vez mais equipes que adotam o modelo de analista de dados incorporado dependem desse tipo de arquitetura.

Por exemplo, talvez a empresa em que trabalhe tenha duzentas pessoas e uma equipe com três engenheiros de dados e dez analistas de dados distribuídos entre as equipes funcionais para melhor atender às necessidades do negócio. Esses analistas podem se reportar às equipes operacionais ou às equipes de dados centralizadas, mas têm os próprios conjuntos de dados e se reportam a pessoas específicas. Como múltiplos domínios gerarão e utilizarão os dados, isso leva à inevitabilidade de que os conjuntos de dados usados por múltiplas equipes sejam duplicados, fiquem ausentes ou obsoletos com o tempo. Como leitor deste

livro, provavelmente você já passou pela experiência de usar um conjunto de dados que não é mais relevante sem saber!

Outras Tendências que Contribuem para o Momento Atual

Além dos fatores mencionados anteriormente que costumam resultar em tempo de inatividade de dados, diversas mudanças no setor também ocorrem em decorrência de inovações tecnológicas que estão impulsionando a transformação do cenário de dados. Essas mudanças contribuem para maior atenção à qualidade de dados.

Data mesh

Da mesma forma que as equipes de engenharia de software passaram de aplicações monolíticas para arquiteturas de microsserviços, o data mesh é, em muitos aspectos, a versão da plataforma de dados de microsserviços. É importante ressaltar que o conceito de data mesh é incipiente e a comunidade de dados discute bastante sobre como implementar um data mesh (ou se faz sentido), considerando os níveis culturais e técnicos.

Conforme definido inicialmente por Zhamak Dehghani, consultora da ThoughtWorks e criadora do termo, o data mesh, ilustrado na Figura 1-1, é um paradigma sociotécnico que reconhece as interações entre pessoas e a arquitetura técnica e soluções em organizações complexas. O data mesh adota a ubiquidade de dados na empresa, aproveitando um projeto self-serve orientado a domínio. Dehghani recorre à teoria de design orientado a domínios de Eric Evans, paradigma flexível e escalável de desenvolvimento de software que combina a estrutura e a linguagem do seu código com seu domínio de negócios correspondente. Ao contrário das infraestruturas de dados monolíticas tradicionais que lidam com o consumo, armazenamento, transformação e saída de dados em um data lake central, um data mesh suporta usuários de dados distribuídos e específicos de domínio e visualiza os "dados como um produto", em que cada domínio gerencia os próprios pipelines de dados. O tecido que conecta esses domínios e seus ativos de dados associados é uma camada de interoperabilidade universal que usa a mesma sintaxe e padrões de dados. Os data meshes federam a propriedade de dados entre os proprietários de dados do domínio, responsáveis por fornecer dados como produtos, ao mesmo tempo que facilitam a comunicação entre dados distribuídos em diferentes locais.

Embora a infraestrutura de dados seja responsável por fornecer a cada domínio as soluções para processá-los, os domínios são encarregados de gerenciar a ingestão, limpeza e agregação de dados, gerando ativos que podem ser usados por aplicações de business intelligence. Mesmo que cada domínio seja responsável pela propriedade de seus pipelines, usa-se um conjunto de capacidades em todos os domínios visando armazenar, catalogar e manter os controles de acesso para os dados brutos. Após os dados terem sido atendidos e transformados por um determinado domínio, os proprietários do domínio podem então usá-los conforme suas necessidades analíticas ou operacionais.

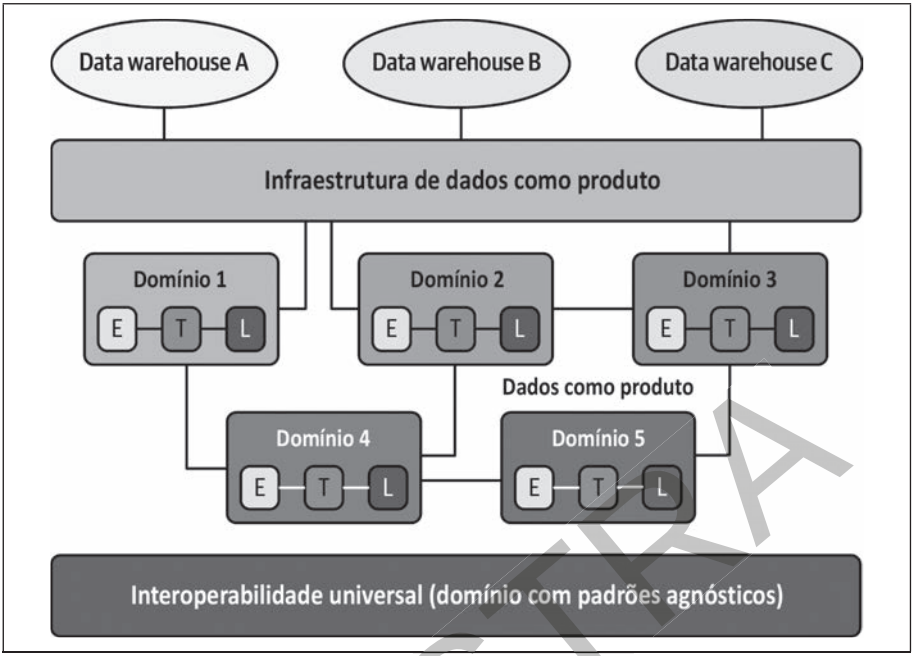

Figura 1-1. O data mesh, introduzido por Zhamak Dehghani, impulsiona a arquitetura de dados descentralizada e orientada a domínios que depende de dados confiáveis de alta qualidade e governança universal.

O paradigma data mesh é bem-sucedido apenas se os dados forem confiáveis e fidedignos e se a chamada "camada de interoperabilidade universal" for aplicada em todos os domínios. Qual é a única maneira de os dados serem confiáveis e fidedignos? Atenção cuidadosa à qualidade de dados por meio de testes, monitoramento e observabilidade.

Muitas empresas estão adotando o paradigma de data mesh, sobretudo organizações maiores com a necessidade de múltiplos domínios de dados. Por exemplo, em um artigo de blog de janeiro de 2021 (*https://oreil.ly/oxTyk*) escrito por Mammad Zadeh, ex-vice-presidente de Engenharia de Dados da Intuit, e Raji Arasu, vice-presidente sênior de Core Services & Experiences da Intuit, a organização se posiciona como "empresa de plataforma especializada e orientada por IA", cuja plataforma "coleta, processa e transforma um fluxo contínuo de dados em um data mesh conectado com alta qualidade de dados". Outro exemplo é o JPMorgan Chase (*https://oreil.ly/Tga4W*), que desenvolveu uma arquitetura de data mesh com a intenção de ajudar a delinear a propriedade de dados entre funções analíticas discretas e melhorar a visibilidade do compartilhamento de dados em toda a empresa. Independentemente de sua opinião, o data mesh sem dúvidas causou grande impacto na comunidade de dados e gerou excelentes discussões e artigos de blog (*https://oreil.ly/rcFTp*) sobre o futuro de nossas arquiteturas de dados distribuídos e estruturas de equipe.

Streaming de dados

Streaming de dados se refere ao processo de transmissão de um fluxo contínuo de dados em um pipeline para gerar rapidamente insights em tempo real. Tradicionalmente, a qualidade de dados foi imposta por meio de testes de dados em lote antes que esses dados fossem inseridos nos pipelines de produção, porém, as empresas cada vez mais buscam análises em tempo real. Embora tenha o potencial de gerar insights mais rápidos, isso também levanta questões e desafios mais importantes relacionados à qualidade de dados, já que streaming de dados são dados "em movimento". Cada dia mais, as organizações estão adotando o processamento em lote e o processamento em fluxo, fazendo com que as equipes de dados repensem sua abordagem de teste e de observabilidade de dados.

Ascensão do data lakehouse

Data warehouse ou data lake? Eis a questão, pelo menos se perguntarmos a um engenheiro de dados. Os data warehouses, repositórios de dados estruturados, e os data lakes, conjuntos de dados brutos não estruturados, dependem de dados de alta qualidade para processamento e transformação. As equipes de dados estão cada vez mais optando por usar data warehouses e data lakes para atender às crescentes necessidades de dados de seus negócios. Apresentamos a todos o data lakehouse.

Os data lakehouses surgiram quando provedores de warehouse em nuvem começaram a adicionar recursos que ofereciam benefícios no estilo data lakes, como o Redshift Spectrum ou Delta Lake. Da mesma forma, os data lakes têm adicionado tecnologias que oferecem recursos no estilo de data warehouse, como funcionalidades e esquema SQL. Hoje, as diferenças históricas entre data warehouses e data lakes estão se estreitando para que possamos acessar o melhor dos dois mundos em um único pacote. Migrar para o modelo lakehouse sugere que os pipelines estão se tornando mais complexos e, embora alguns possam escolher um fornecedor dedicado para lidar com ambos, outros estão migrando os dados para múltiplas camadas de armazenamento e de processamento, resultando em mais oportunidades para os dados do pipeline alcançarem o ponto de equilíbrio com testes amplos.

Recapitulando

A ascensão da nuvem, arquiteturas e equipes de dados distribuídas e a mudança rumo à produtização de dados atribuíram aos líderes de dados a responsabilidade de ajudar suas empresas a buscar dados mais fidedignos (levando a análises mais fidedignas). Obter dados confiáveis é uma maratona, não um sprint, e envolve muitos estágios relacionados ao pipeline de dados. Além disso, comprometer-se com a melhoria da qualidade de dados vai além do desafio técnico; também envolve níveis culturais e organizacionais. No próximo capítulo, analisaremos algumas tecnologias que sua equipe pode usar para evitar pipelines quebrados e criar processos e estruturas repetitivos e iterativos com os quais se comunicar melhor, abordar e até mesmo evitar o tempo de inatividade de dados.

CAPÍTULO 2

Reunindo os Componentes Essenciais de Um Sistema Confiável de Dados

Com Ryan Kearns

Apesar de a resolução de problemas associados à qualidade de dados ser conjunto de habilidades imprescindíveis a qualquer profissional da área, o tempo de inatividade de dados pode ser evitado quase que totalmente com os sistemas e processos adequados. Como os softwares, os dados podem depender de inúmeras influências operacionais, programáticas ou até mesmo relacionadas aos próprios dados em diversas etapas do pipeline. Basta uma mudança de esquema ou um push de código para gerar um relatório de downstream confuso.

Como analisaremos no Capítulo 8, a solução para a qualidade de dados e para a criação de pipelines mais confiáveis se divide em três componentes principais: processo, tecnologias e pessoas. Neste capítulo, abordaremos o componente tecnológico dessa equação, mapeando as partes discrepantes do pipeline de dados e o que é necessário para mensurar, corrigir e prevenir o tempo de inatividade de dados em cada etapa.

Os sistemas de dados são absurdamente complexos, já que diversas etapas do pipeline de dados contribuem para o caos generalizado. E ao mesmo tempo que as empresas investem progressivamente em dados e em analytics, aumenta também a pressão para se desenvolver em escala, fazendo com que os engenheiros de dados considerem a qualidade antes mesmo de os dados entrarem no pipeline.

Neste capítulo, destacaremos os inúmeros componentes essenciais, impulsionados por metadados — desde catálogos de dados até data warehouses e data lakes — para assegurar que sua infraestrutura de dados esteja preparada para o sucesso quando se trata de garantir alta qualidade de dados em cada etapa do pipeline.

Entendendo a Diferença Entre Dados Operacionais e Dados Analíticos

Se perguntarmos a um engenheiro de dados qual é a maior diferença possível entre os dados em sua organização, podemos ouvir os termos "dados operacionais" e "dados analíticos". A diferença entre operacional e analítico tem a ver com as inúmeras maneiras de decompor os dados em um ecossistema. Mas se trata de uma diferença importante, que você precisará entender, caso esteja interessado em adotar a cultura de qualidade de dados. Ainda que nos aprofundemos um pouco mais nos dados operacionais neste capítulo, vale ressaltar que, neste livro, temos focado e continuaremos a focar a qualidade de dados referente aos dados analíticos. O gerenciamento da qualidade e da confiabilidade de dados operacionais normalmente é responsabilidade do DevOps, da engenharia de confiabilidade de site e de outras disciplinas de software mais preocupadas em criar produtos de software baseados em dados analíticos.

Dados operacionais
: Dados gerados operacionalmente, ou seja, gerados pelas operações cotidianas de sua organização (*https://oreil.ly/kZmui*). Snapshots de estoque em momentos específicos, opiniões de clientes e registros de transações são exemplos de dados operacionais.

Dados analíticos
: Dados usados de forma analítica. Ou seja, é o tipo de dados por trás das decisões de negócios orientadas a dados. Churn de marketing, taxas de cliques e opiniões por região global são exemplos de categorias de dados analíticos.

Em resumo, os dados operacionais registram dados provenientes dos processos reais de negócio visando atualizações rápidas em sistemas e processos, ao passo que os dados analíticos são usados para análises mais robustas e eficientes. Podemos pensar da seguinte forma: os dados operacionais *operacionalizam* os negócios, enquanto os dados analíticos *gerenciam* os negócios. Considerando que os dados analíticos impulsionam o business intelligence de uma forma que os operacionais não impulsionam, alguém pode cair na tentação de suspeitar que aqueles sejam mais importantes ou mais "essenciais" para o sucesso de uma organização. No entanto, na maioria das vezes, os dados analíticos se baseiam em um backbone de transformações e agregações de dados operacionais.

A distinção entre operacional e analítico é a mesma feita pela comparação entre sistemas de processamento de transações e sistemas de dados analíticos (OLTP [Processamento de Transações Online] versus OLAP [Processamento Analítico Online]), por exemplo, no livro *Designing Data-Intensive Applications* [Projetando Aplicações Com Data Intensive, em tradução livre].

O Que Diferencia Esses Dados?

Como se pode imaginar, os dados analíticos e operacionais se diferenciam de algumas maneiras críticas que informam como gerenciamos sua confiabilidade, conforme ilustrado na Figura 2-1.

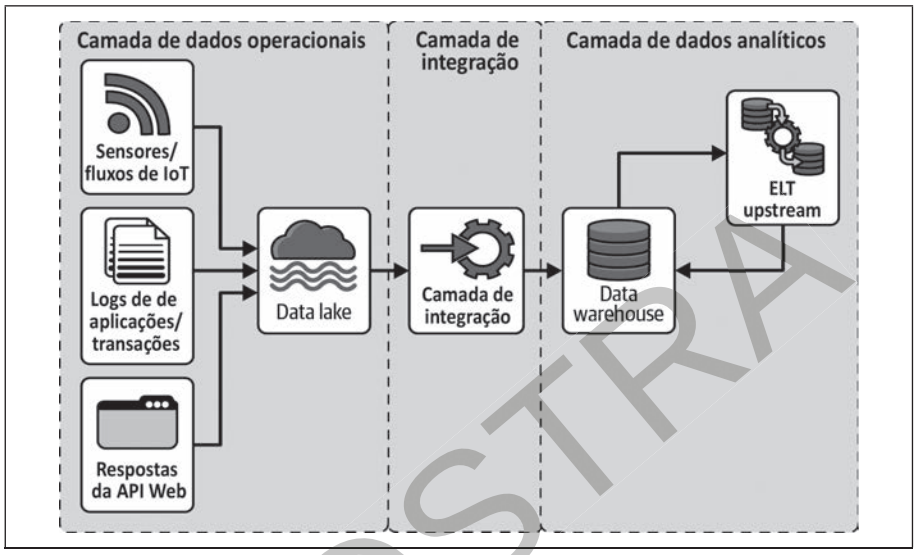

Figura 2-1. Um exemplo de plataforma de dados, ilustrando somente uma maneira de distinguir dados operacionais e dados analíticos.

Quase sempre, os dados operacionais aparecem *upstream* dos dados analíticos nos pipelines de dados. Isso ocorre porque os dados analíticos podem conter, e geralmente contêm, agregações ou expansões (via data augmentation) de data stores de dados operacionais. A taxa de cliques de um usuário em um navegador às 5h da manhã é um dado operacional, e a de uma campanha de marketing de dezembro é um dado analítico correspondente. Uma das razões fundamentais pela qual a diferença entre dados operacionais e analíticos é importante é o chamado *trade-off entre taxa de requisição e latência*. A restrição de taxa de requisição-latência impacta qualquer sistema com poder computacional limitado. Tradicionalmente, a *taxa de requisição* se refere à quantidade de dados processados em alguma unidade de tempo e a *latência* se refere ao delay antes que os dados sejam processados.

Imagine um cibercafé popular com uma fila do lado de fora. Quanto tempo leva para alguém no final da fila receber seu café? Esse processo envolve ficar na fila, fazer o pedido, pagar e esperar que o barista prepare a bebida. A soma total desse tempo representa a *latência* do café. Em contrapartida, o número de clientes bebendo seus cafés dentro do estabelecimento em, digamos, uma hora, representa a *taxa de requisição* do café.

Infelizmente, essas duas medidas de desempenho de processamento de dados estão fadadas a competir. Não podemos ter alta taxa de requisição e baixa latência em nosso cibercafé.