



# **CINCO PRÁTICAS RECOMENDADAS PARA O DESENVOLVIMENTO DE DATA WAREHOUSE**

EBOOK

# ÍNDICE

- 3** Introdução
- 4** Criar um modelo de dados
- 7** Adotar uma metodologia Agile para data warehouses
- 8** Escolher o método ELT em vez de ETL
- 9** Adotar uma ferramenta de automação
- 10** Treinar sua equipe em novas abordagens
- 11** Resumo
- 12** Sobre o Snowflake



# INTRODUÇÃO

A tecnologia de nuvem revolucionou a maneira como os dados são armazenados, acessados e analisados pelas empresas. As organizações que estão criando uma nova plataforma de dados do zero ou reprojando um sistema legado de data warehouse para aproveitar novos recursos poderão contar com este conjunto de diretrizes e práticas recomendadas para ajudar a garantir o sucesso do projeto. Algumas dessas práticas recomendadas parecem óbvias, mas é comum vermos empresas que não dedicam tempo para estabelecer e documentar esses pontos de decisão, e isso causa problemas e ineficiência no futuro.

Neste ebook, destacamos cinco recomendações para você estruturar sua estratégia de dados e obter alinhamento em toda a empresa. Assim, o data warehouse que você criar atenderá às necessidades atuais e futuras. Estas práticas recomendadas para o desenvolvimento de data warehouse aumentarão as chances de que os usuários da empresa obtenham mais valor do data warehouse criado, bem como vão estabelecer as bases para uma plataforma de dados corporativos mais ampla, capaz de crescer e se adaptar às mudanças das necessidades da empresa.

# 1. CRIAR UM MODELO DE DADOS

A primeira etapa importante em qualquer programa de dados é a criação de um modelo de dados. Trata-se de uma representação abstrata que organiza elementos de dados e descreve o relacionamento entre eles e com as propriedades de suas entidades no mundo real. Um modelo de dados estabelece definição e entendimento comuns sobre quais informações são importantes para a empresa, assim como o cenário geral dos dados da empresa. A existência de um modelo de dados oferece uma forma de documentar os conjuntos de dados que serão integrados ao data warehouse, o relacionamento entre esses conjuntos de dados e os requisitos de negócios que a plataforma precisa atender.

Seria possível criar um data warehouse sem um modelo de dados? Sim, mas ao decidir ignorar essa etapa básica, você perde muitos insights importantes. A criação de um modelo de dados abrangente é um exercício revelador para as empresas, pois força equipes de diferentes funções a concordar com a definição e o delineamento de ativos de dados e requisitos de negócios do data warehouse antes do início do desenvolvimento.

Um modelo de dados bem-definido causa um impacto positivo muito depois de o data warehouse (ou data mart) estar ativo. Por exemplo, um modelo de dados estabelece a linhagem dos dados para todos os objetos de um data warehouse, o que facilita a integração de novos membros da equipe ou traz novos objetos de dados para o data warehouse, à medida que as necessidades de negócios mudam.

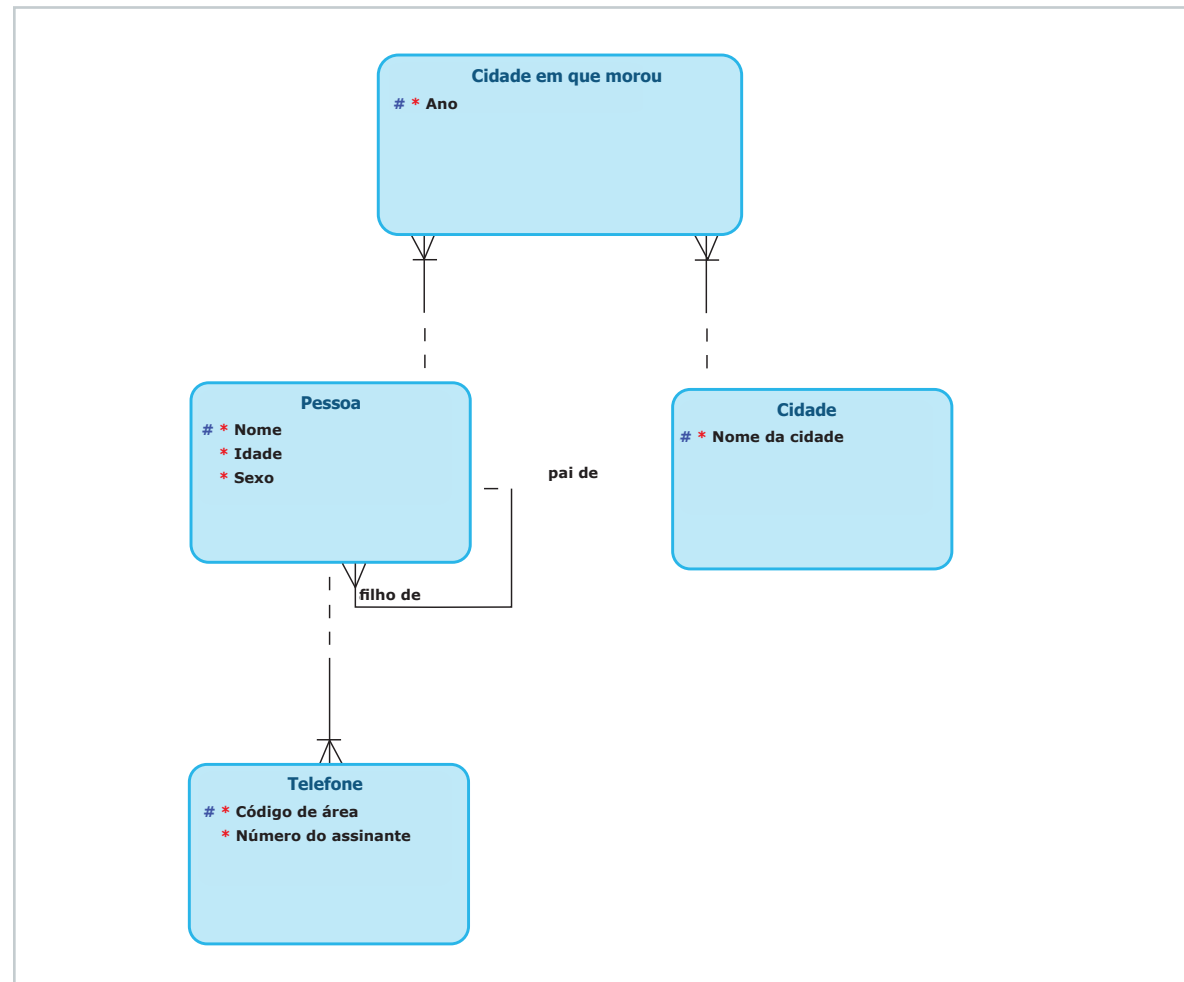


Figura 1: um modelo de dados lógico típico 3NF (terceira forma normal)

O modelo de dados também fornece documentação clara sobre o conteúdo, o contexto e as fontes. Isso facilita a auditoria ou o cumprimento de novos requisitos de dados, como os apresentados pelo regulamento geral de proteção de dados (General Data Protection Regulation, GDPR), ou seja, a estrutura de regulamentação e proteção de dados geral da eu que define as diretrizes para a coleta e o processamento de informações pessoais de indivíduos.

Um modelo de dados forte também ajuda a evitar confusões e reprojatos onerosos no futuro. É sempre bom associar uma camada de integração independente da origem, que permita a análise em vários conjuntos de dados com base nas respectivas semelhanças.

Um data warehouse reúne muitas fontes e tipos de dados diferentes. Ele inclui conjuntos de dados tradicionais, como de gerenciamento do relacionamento com o cliente (customer relationship management, CRM) e de planejamento de recursos empresariais (enterprise resource planning, ERP), assim como conjuntos de dados de blogs, feeds do Twitter, dados de Internet das Coisas (IoT) e até conjuntos de dados que ainda não foram inventados. É por isso que ter uma camada de integração flexível e que não esteja fortemente ligada a qualquer sistema único vai ajudar a manter o data warehouse preparado para o futuro.

Um modelo de dados altamente eficaz deve empregar definições e estruturas semânticas definidas de acordo com o domínio da empresa, e não baseadas nas definições específicas de um único sistema de origem. Por exemplo, um sistema CRM pode se referir a clientes como "cust", e outros como "cust\_ID". A chave para o sucesso do data warehouse é a definição, para toda a empresa, de uma regra semântica sobre como os usuários devem nomear, acessar e analisar esses dados em um conjunto de dados.

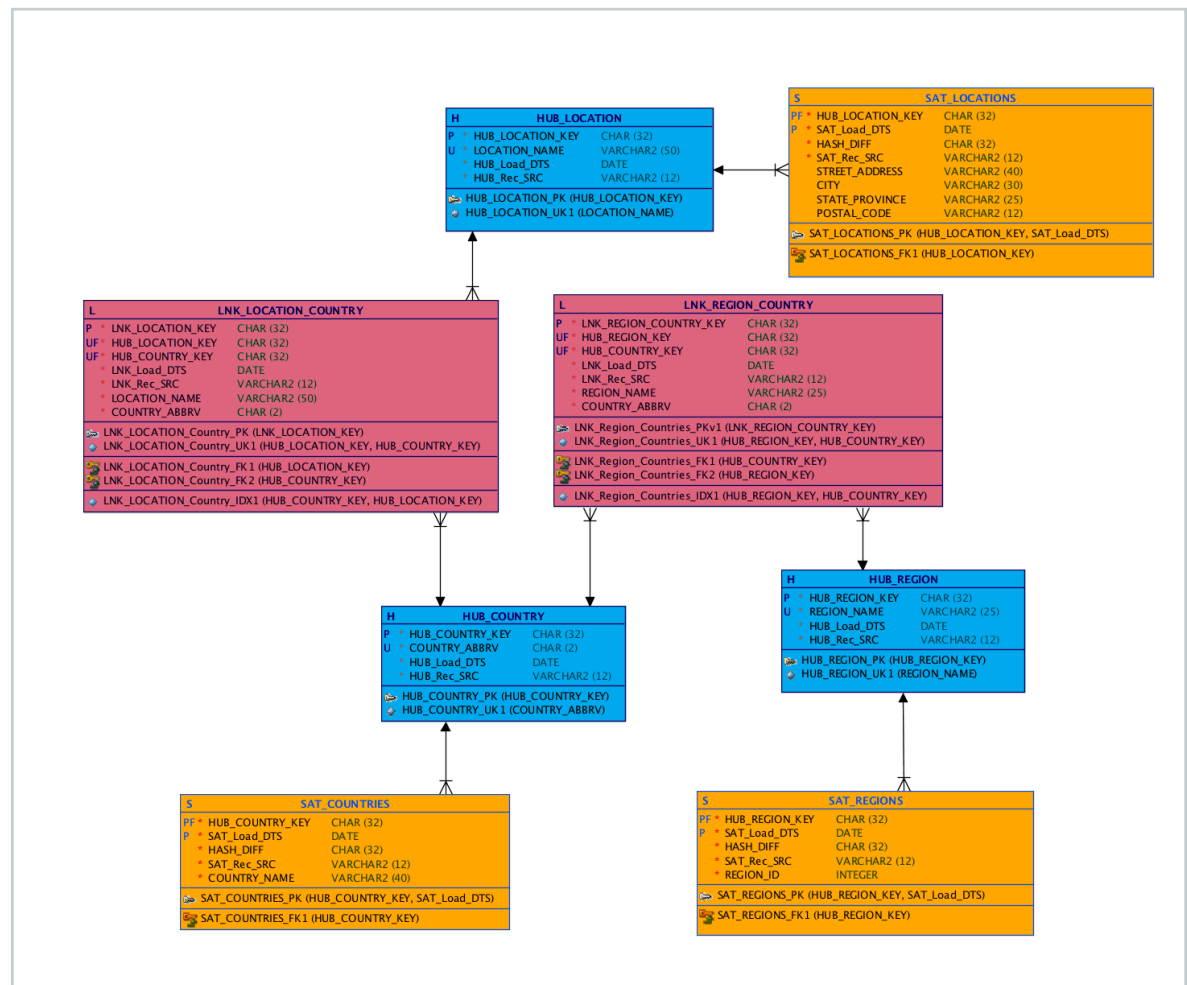


Figura 2: um exemplo de modelo de dados que usa o método de modelagem Data Vault

As mudanças, fusões e aquisições que a empresa vier a enfrentar poderão exigir a substituição do sistema de CRM atual. Se o modelo de dados atual estiver fortemente acoplado a um sistema de origem específico, você precisará realizar uma completa reengenharia para integrar o segundo sistema de origem que vai substituir o sistema legado. Uma camada de integração independente da origem facilita o mapeamento de dados e permite a troca do sistema de origem antigo pelo novo sem afetar a geração de relatórios posteriores ou exigir a mudança de comportamento do usuário.

No modelo de dados, é essencial que seja feita a seleção de uma abordagem padrão. Os principais tipos de padrões de modelagem de dados usados em projetos de data warehouse incluem:

- **3NF:** que significa "terceira forma normal" ("third normal form", 3NF), é um padrão de arquitetura projetada para reduzir a duplicação de dados e garantir a integridade referencial do banco de dados.<sup>1</sup>
- **Esquema em estrela:** a arquitetura mais simples e amplamente usada para desenvolver data warehouses e data marts dimensionais. O esquema em estrela consiste em uma ou mais tabelas de fatos que se referem a qualquer número de tabelas de dimensões.<sup>2</sup>
- **Data Vault (DV):** desenvolvido especificamente para lidar com problemas de agilidade, flexibilidade e escalabilidade que são encontrados em outras abordagens, a modelagem DV foi criada como um repositório de dados corporativos granular, não volátil, auditável e facilmente ampliável. É altamente normalizado e combina elementos dos modelos 3NF e em estrela.<sup>3</sup>

Cada arquitetura tem suas vantagens, mas a escolha de qual adotar está atrelada às necessidades de negócios da organização.

Mais importante do que a arquitetura a ser escolhida é o fator de que sua organização também precisa selecionar, documentar e oferecer suporte contínuo a essa arquitetura como parte do desenvolvimento de um modelo de dados para o warehouse. No futuro, isso resultará em eficiência e permitirá a adoção de uma metodologia única de suporte e solução de problemas que facilitará o crescimento mais rápido dos novos membros da equipe.



## 2. ADOPTAR UMA METODOLOGIA AGILE PARA DATA WAREHOUSE

No passado, a criação de data warehouse (ou até mesmo de data marts) representava um esforço grande, monolítico, de vários trimestres ou anos e sujeito ao tradicional processo de "cascata". Essa não é mais a norma atual, pois muitas organizações estão optando por adotar um método de projeto mais flexível e iterativo ou a metodologia Agile.

As necessidades dos negócios estão mudando mais rápido do que nunca e, com a mesma rapidez, novas fontes de dados estão chegando online, o que exige das empresas capacidade de adaptação e de aproveitamento desses dados de forma concisa e rápida. Isso significa aprender a criar soluções de dados e soluções analíticas de maneira incremental e com a metodologia Agile. Com um planejamento adequado que se alinha a uma camada única de integração independente da origem, grandes projetos de dados agora podem ser divididos em partes menores a serem entregues com mais frequência, fornecendo valor incremental aos negócios de maneira mais rápida.

Para atingir esse objetivo, os arquitetos de armazenamento de dados estão adotando a metodologia Agile, que foi elaborada inicialmente no mundo do desenvolvimento de software. Na metodologia Agile, os requisitos e as soluções evoluem por meio do esforço colaborativo de equipes e clientes auto-organizados e multifuncionais. Quando aplicada à concepção de criação de data warehouses, ela permite que as empresas atuem novos conjuntos de dados e resolvam desafios de negócios rapidamente.<sup>4</sup>

Na cenário mundial de Agile, vários métodos surgiram para ajudar a entregar valor mais rapidamente, dentre eles:

- **Scrum:** é como se chama a formação do rugby em que os jogadores entrelaçam os braços e avançam. Aqui, é usado como nome da estrutura de processo mais amplamente usada para o desenvolvimento com a metodologia Agile. O Scrum é uma estrutura leve, que enfatiza a comunicação diária e a reavaliação flexível de planos que são executados em fases de trabalho curtas e iterativas.<sup>5</sup> Ralph Hughes codificou a aplicação de Scrum para armazenamento de dados em uma série de trabalhos produtivos que são úteis para as empresas que adotam essa metodologia.
- **Kanban:** é um método para gerenciar a criação de produtos com ênfase na entrega contínua, sem sobrecarregar a equipe de desenvolvimento. Da mesma forma que o Scrum, o Kanban é um processo projetado para ajudar as equipes a trabalhar juntas de maneira mais eficiente. Nomeado como os cartões "Kanban" que rastreiam a produção em uma fábrica, o Kanban foi criado por Taiichi Ohno, um engenheiro industrial da Toyota, para melhorar a eficiência da produção.
- **BEAM:** o método Business Event Analysis and Modelling foi apresentado por Lawrence Corr e Jim Stagnitto em seu trabalho inovador, Agile Data Warehouse Design. O BEAM se concentra em eventos de negócios, em vez de requisitos de relatórios conhecidos, para modelar toda a área de processos empresariais. Ele usa sete tipos dimensionais, os sete Ws: who, what, when, where, how, how many e why (quem, o que, quando, onde, como, quantos e por que), para identificar e elaborar eventos de negócios.<sup>6</sup>

Uma plataforma de dados Agile é de grande utilidade para aproveitar melhor as vantagens do desenvolvimento em Agile. As plataformas de dados baseadas na nuvem oferecem essa flexibilidade e elasticidade estruturais ao permitir um escalonamento rápido, compatível com a evolução das necessidades dos negócios. Além de exigirem menos esforços de manutenção e administração para serem úteis, as plataformas de dados baseadas na nuvem podem crescer e se adaptar às mudanças que ocorrem nos requisitos de negócios. Ao utilizar um serviço de nuvem moderno, as equipes podem passar menos tempo ajustando consultas e provisionando armazenamento e se dedicar à abordagem de desafios imediatos de negócios e à agregação de valor aos negócios.

O emprego de metodologias e estruturas Agile não é uma tarefa simples. O processo requer um compromisso cultural da organização e geralmente implica em mudança significativa da mentalidade e do fluxo de trabalho, isto é, dos fluxos de trabalho de data warehousing tradicionais. A reestruturação de uma equipe de TI para trabalhar confortavelmente em um ambiente Agile pode levar de 6 a 12 meses. Isso talvez pareça contraditório, tendo em vista o objetivo da metodologia Agile de agregar valor mais rapidamente. Essa transição pode ser acelerada com a contratação de um treinador experiente em Agile. Depois de realizada a mudança, as equipes podem começar a implementar novas alterações no data warehouse de forma incremental em questão de semanas, em vez de meses.

### 3. ESCOLHER O MÉTODO ELT EM VEZ DE ETL

No passado, o método adotado para o desenvolvimento de data warehousing era o de extrair-transformar-carregar (extract-transform-load, ETL), em que os dados a serem importados para o data warehouse eram extraídos dos sistemas de origem e carregados no data warehouse de destino após limpeza ou aplicação de regras de negócios em um servidor externo. O aumento da capacidade e da eficiência dos recursos de computação da plataforma de dados gerou o método de extrair-carregar-transformar (extract-load-transform, ELT), que se tornou a opção favorita.

Nesse método, os dados brutos são extraídos quase inalterados da origem e carregados na área de preparação do data warehouse. Metadados, data de carregamento ou informações de origem podem ser adicionados aos dados e levados diretamente para o data warehouse. Uma vez no data warehouse, as empresas podem usar a capacidade do banco de dados para realizar transformações, como a alteração da estrutura dos dados, isto é, aplicar um modelo de dados, aplicar regras de negócios ou implementar medidas de qualidade de dados para limpar os dados (por exemplo, corrigindo endereços incompletos e eliminando duplicidades).

O método ELT tem duas vantagens distintas: economia de custos e maior rastreabilidade. O ELT ajuda a reduzir os custos ao permitir que as empresas aproveitem o poder da plataforma para transformar os dados, em vez de usar um servidor externo. Geralmente, a computação baseada em nuvem é muito menos onerosa do que a realização e a manipulação de dados em um servidor externo. Dessa forma, mover dados diretamente para a nuvem é mais rápido e mais barato. O método ELT também facilita o rastreamento e a auditoria dos dados no futuro, pois fornece uma imagem dos dados originais diretamente na plataforma de dados. Dessa forma, o próprio data warehouse pode desempenhar o papel do que veio a ser conhecido como "data lake", onde os dados brutos são armazenados de maneira permanente.



## 4. ADOPTAR UMA FERRAMENTA DE AUTOMAÇÃO

O objetivo do data warehouse é agilizar a ativação e o fornecimento de dados para auxiliar o processo decisório com base em informações e gerar de mais valor. Uma forma de aumentar a rapidez da entrega é adotar a metodologia Agile. A outra é adotar ferramentas de automação que ajudem na implantação mais rápida dos códigos. Muitas metodologias de data warehouse são baseadas em padrões, portanto, a codificação necessária para carregar e estruturar dados geralmente pode ser repetida, isto é, pode ser automatizada. Há várias ferramentas no mercado que automatizam algumas ou até mesmo todas as tarefas de projeto e criação, e a lista cresce diariamente.

A automação permite que as empresas utilizem seus recursos de maneira quase completa, iterem mais rapidamente e imponham padrões de codificação com facilidade. Ela permite a criação de código padronizado, o que é extremamente útil em organizações em que o código ETL e os modelos de dados são tradicionalmente desenvolvidos de forma manual. A automação oferece um padrão documentado para esses diferentes artefatos, além de um mecanismo de controle de qualidade (quality assurance, QA) para monitorar se todos os desenvolvedores e designers estão seguindo esse padrão.

As ferramentas de automação que usam modelos para gerar códigos são muito úteis, porque reforçam os padrões e os tornam as propriedades preferenciais nos próprios modelos. Isso acelera a integração, pois novos desenvolvedores e designers usarão essas ferramentas padrão, garantindo, assim, uma implementação consistente e diminuindo a curva de aprendizado. Uma implementação consistente oferece o benefício adicional de ser mais fácil de testar e depurar, pois o código é desenvolvido nos mesmos padrões.

A iteração também se torna mais rápida com o uso dessas ferramentas, pois os geradores de códigos automatizados tendem a cometer menos erros de sintaxe. Normalmente, a atualização do código implica na adição de novos objetos à ferramenta ou na alteração das propriedades dos modelos no nível global. Esse processo gera um novo código que fica disponível imediatamente para implementação no ambiente de teste e validação.

## 5. TREINAR SUA EQUIPE EM NOVAS ABORDAGENS

A mudança para a metodologia Agile ou para o desenvolvimento de código automatizado não é apenas uma alteração dos conjuntos de habilidades, mas sim uma mudança de mentalidade. São necessários treinamento e formação para garantir que a equipe faça uso eficaz dessas novas abordagens e tecnologias. Isso significa que pode ser necessário trazer especialistas externos para treinar as equipes com as práticas recomendadas de Scrum ou instruir as equipes sobre os benefícios, as regras e as práticas recomendadas de qualquer arquitetura padrão que a empresa tenha adotado para sua plataforma de dados.

Vários recursos do setor estão disponíveis para ajudar a gerenciar a transição para a metodologia Agile. A **Agile Alliance** é uma organização membro global sem fins lucrativos, dedicada a promover os conceitos do desenvolvimento de software conforme descrito no Agile Manifesto. Ela oferece muitas opções de treinamento para apresentação dos conceitos Agile. A **Scrum Alliance** oferece certificações e treinamento em Scrum básico e avançado. Da mesma forma, o bootcamp e a certificação Data Vault são oferecidos por parceiros selecionados por meio da **Data Vault Alliance**.

Como ocorre em qualquer novo processo e mudança cultural, as organizações devem gerenciar a curva de adoção para garantir a mudança consistente e eficaz para a nova abordagem nas operações diárias. A identificação de projetos-piloto ou de prova de conceito para iniciar as equipes nas novas abordagens garantirá que os profissionais desenvolvam e dominem as habilidades em cenários protegidos, mas reais, que vão acelerar a competência e as capacidades nessas novas habilidades.



## RESUMO

Todas as práticas recomendadas descritas neste ebook exigem um investimento inicial para alcançar o valor comercial de longo prazo que podem oferecer. Mas o retorno desse investimento é duplo: ele vai estabelecer as bases para um programa de análise de dados bem-sucedido desde o início e acelerar a entrega de valor comercial incremental para seu ambiente de dados muito depois da primeira versão estar em produção.

À medida que os requisitos de negócios mudam e o desejo de obter mais valor de mais dados e tipos de dados continua acelerando, ter essas práticas recomendadas em vigor permitirá que você pense e cresça muito além dos casos de uso de data warehousing tradicionais. Com uma base sólida e uma plataforma ágil, você poderá expandir para novos domínios de dados e atender às novas demandas, aplicando o programa para auxiliar a ciência de dados, o aprendizado de máquina, a IA e talvez até a monetização de dados. Com os atuais recursos de nuvem flexíveis e escalonáveis, realmente não há limites para o que você pode alcançar com seus dados.





# SOBRE O SNOWFLAKE

O Snowflake permite que todas as empresas impulsionem seus dados, graças ao Snowflake Data Cloud. Os clientes usam o Data Cloud para eliminar silos de dados, descobrir e compartilhar dados com segurança, capacitar aplicativos de dados e executar inúmeras cargas de trabalho analíticas e de IA/ML. Onde quer que os dados ou os usuários estejam, o Snowflake proporciona uma única experiência de dados em inúmeras nuvens e regiões. Milhares de clientes em diversos setores, incluindo 639 das empresas que aparecem na Forbes Global 2000 (G2K) de 2023 (dados de 31 de julho de 2023), usam o Snowflake Data Cloud para impulsionar seus negócios.

Saiba mais em [snowflake.com](https://www.snowflake.com).



© 2023 Snowflake Inc. Todos os direitos reservados. Snowflake, o logotipo da Snowflake e todos os demais nomes de produtos, recursos e serviços da Snowflake mencionados neste documento são marcas registradas ou marcas comerciais da Snowflake Inc. nos Estados Unidos e em outros países. Todos os outros nomes de marcas ou logotipos mencionados ou usados neste documento são apenas para fins de identificação e podem ser marcas comerciais de seus respectivos detentores. A Snowflake não pode ser associada a tais detentores, nem patrocinada ou apoiada por eles.

---

## CITAÇÕES

<sup>1</sup> [en.wikipedia.org/wiki/Third\\_normal\\_form](https://en.wikipedia.org/wiki/Third_normal_form)

<sup>2</sup> [en.wikipedia.org/wiki/Star\\_schema](https://en.wikipedia.org/wiki/Star_schema)

<sup>3</sup> [snowflake.com/blog/data-vault-modeling-and-snowflake](https://snowflake.com/blog/data-vault-modeling-and-snowflake)

<sup>4</sup> [Agiledata.org/essays/dataWarehousingBestPractices.html](https://agiledata.org/essays/dataWarehousingBestPractices.html)

<sup>5</sup> [scrum.org/resources/what-is-scrum](https://scrum.org/resources/what-is-scrum)

<sup>6</sup> [bystembuilders.com/beam](https://bystembuilders.com/beam)