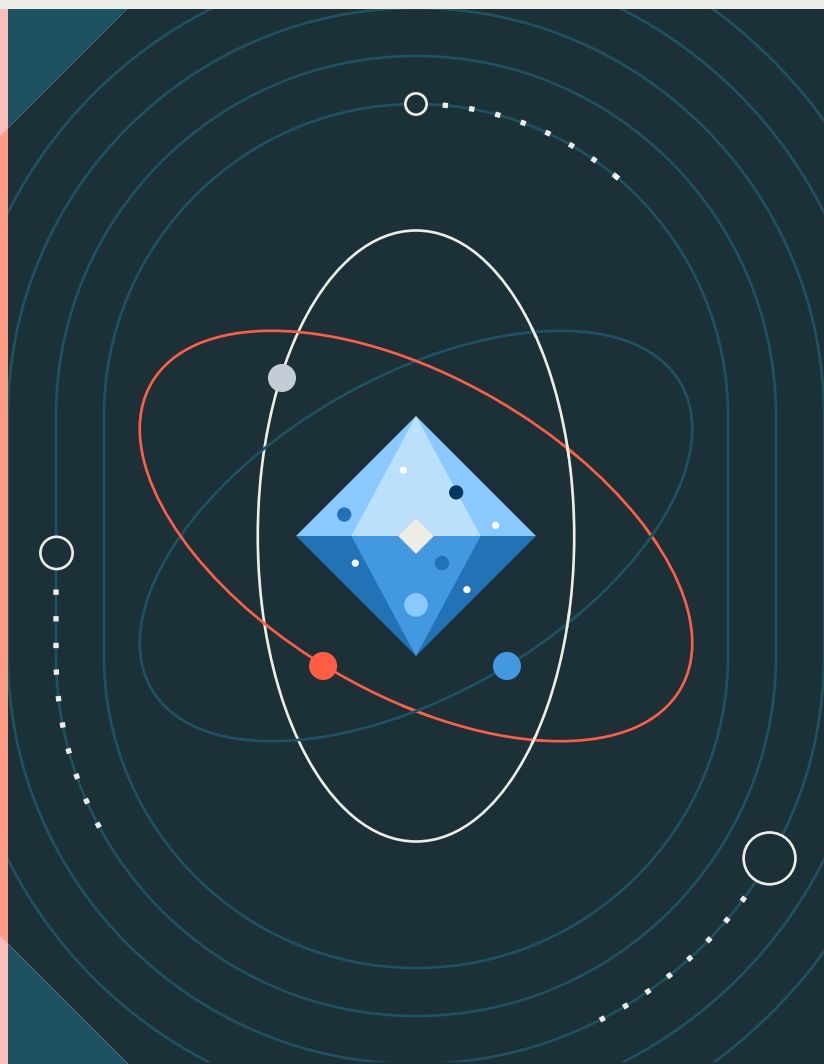
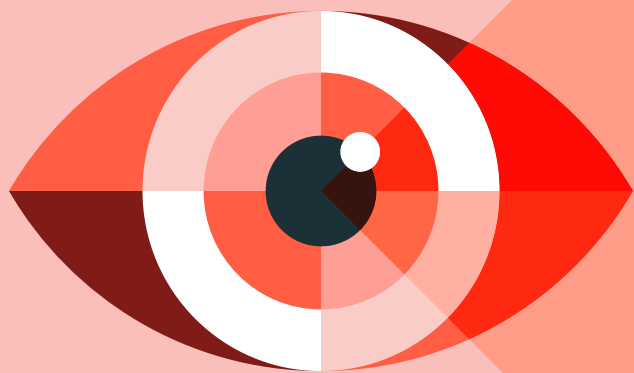
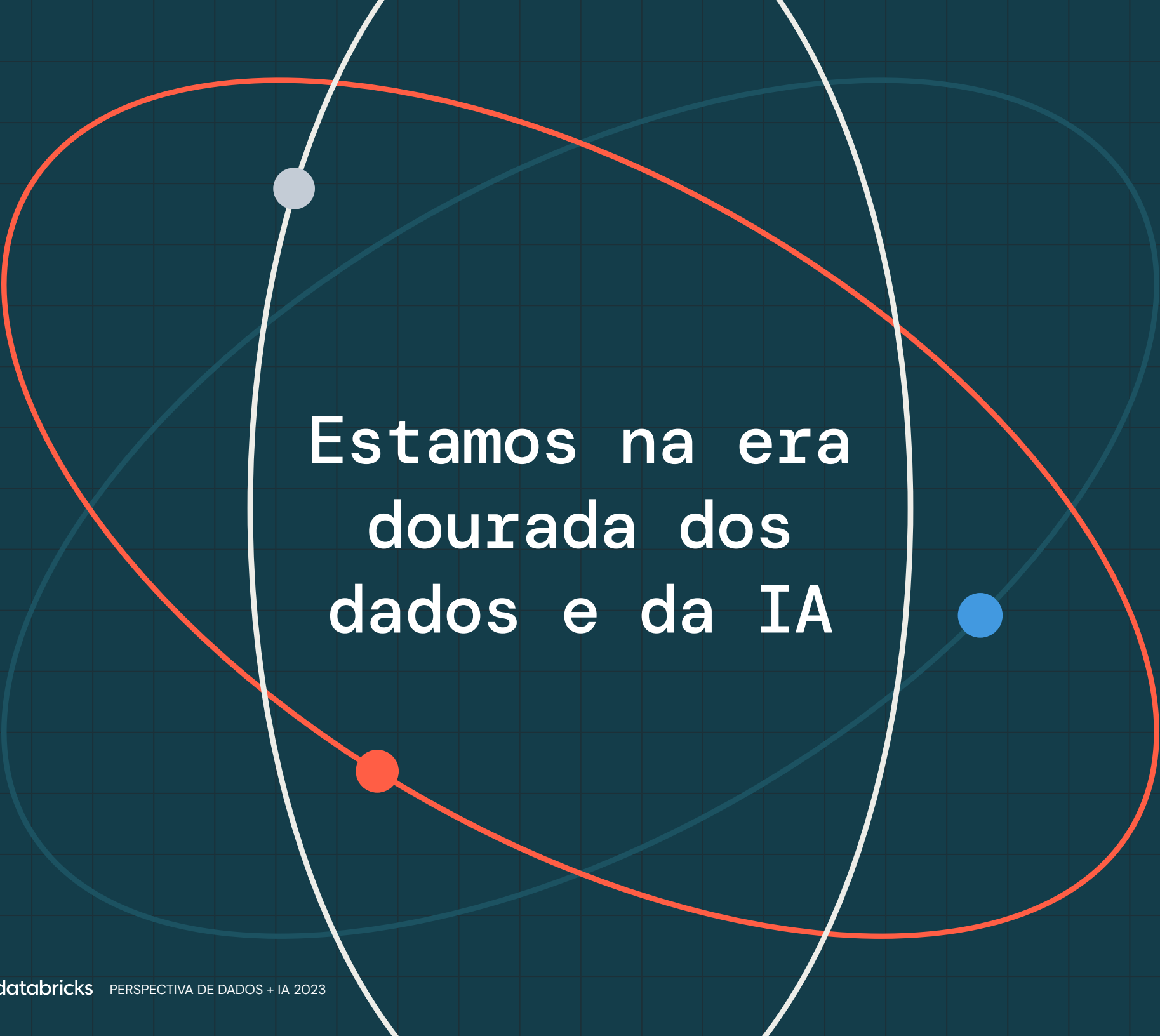


Perspectiva de dados + IA 2023

Fornecido pelo Databricks Lakehouse





Estamos na era dourada dos dados e da IA

INTRODUÇÃO

Nos 6 meses desde o lançamento do ChatGPT, o mundo acordou para o vasto potencial da IA. O ritmo incomparável das descobertas de IA, melhorias de modelo e novos produtos no mercado colocam os dados e a estratégia de IA no topo das conversas em todas as organizações do mundo. Acreditamos que a IA inaugurará a próxima geração de inovação de produtos e softwares, e já estamos vendo isso acontecer no mercado. A próxima geração de empresas e executivos vencedores será aquela que entende e potencializa a IA.

Neste relatório, examinamos padrões e tendências na adoção de dados e IA em mais de 9.000 clientes globais da Databricks. Ao unificar os aplicativos de inteligência de negócios (BI) e IA em todo o patrimônio de dados das empresas, a Databricks Lakehouse oferece um ponto de vista exclusivo sobre a perspectiva de dados e da IA, incluindo quais produtos e tecnologias são os mais populares e de crescimento mais rápido, os tipos de aplicativos de ciência de dados e machine learning (DS/ML) que estão sendo desenvolvidos e muito mais.

Aqui estão as principais histórias que descobrimos:



As empresas estão adotando machine learning e modelos de linguagem grandes (LLMs) com rapidez. O processamento de linguagem natural (NLP) está dominando os casos de uso, com um foco acelerado em LLMs.



O código aberto vence nos mercados atuais de dados e IA. 8 de 10 de nossos produtos de IA e dados mais amplamente aceitos são baseados em código aberto.



As organizações estão cada vez mais usando o Lakehouse para armazenamento de dados, conforme evidenciado pelo principal produto de IA e dados Microsoft Power BI, o alto crescimento das ferramentas de integração de dados dbt e Fivetran, e a rápida adoção do Databricks SQL.

Esperamos que, ao compartilhar essas tendências, os líderes de dados possam comparar suas organizações e obter insights que ajudem a informar suas estratégias para uma era definida por dados e IA.

Resumo das principais conclusões

1

DATA SCIENCE E MACHINE LEARNING: NLP E LLMS ESTÃO EM ALTA

- O número de empresas que usam APIs de LLM SaaS (usadas para acessar serviços como ChatGPT) aumentou 1310% entre o fim de novembro de 2022 e o início de maio de 2023
- O NLP responde por 49% do uso diário da biblioteca de ciência de dados Python, tornando-o o aplicativo mais popular
- As organizações estão colocando mais modelos em produção (crescimento anual de 411%) enquanto também aumentam sua experimentação de ML (crescimento anual de 54%)
- As empresas estão ficando mais eficientes com ML; para cada três modelos experimentais, aproximadamente um é colocado em produção, em comparação com cinco modelos experimentais um ano antes

2

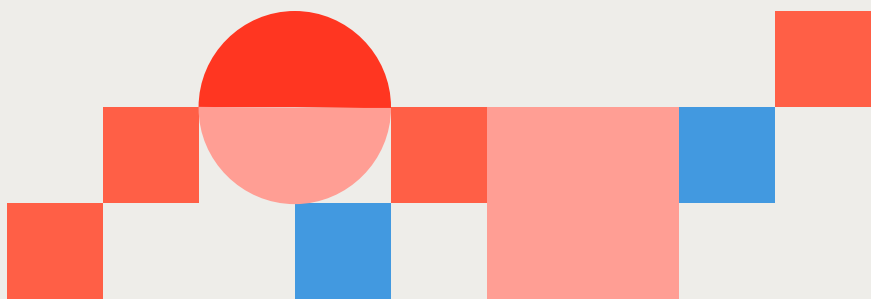
PRINCIPAIS PRODUTOS DE DADOS E IA E MERCADOS

- O Microsoft Power BI é o produto de IA e dados mais popular de 2023, além do Databricks Lakehouse
- O produto de dados e IA que mais cresce é o dbt, que expandiu 206% em relação ao ano anterior em número de clientes
- 8 dos 10 principais produtos de IA e dados são baseados em código aberto
- A integração de dados é o mercado de IA e dados que mais se desenvolve no Databricks Lakehouse, com crescimento de 117% em relação ao ano anterior

3

TENDÊNCIAS DE ADOÇÃO E MIGRAÇÃO

- 61% dos clientes que migram para o Lakehouse vêm de data warehouses locais e na nuvem
- O volume de dados no Delta Lake cresceu 304% em relação ao ano anterior
- O Lakehouse está sendo cada vez mais usado para data warehousing, incluindo serverless data warehousing com Databricks SQL, que cresceu 144% em relação ao ano anterior



Metodologia: Como a Databricks criou esse relatório?

O *Perspectiva de Dados + IA 2023* foi criado a partir de dados anônimos e totalmente agregados coletados de nossos clientes com base em como eles estão usando o Databricks Lakehouse e seu amplo ecossistema de ferramentas integradas. Esse relatório se concentra na adoção de machine learning, arquitetura de dados (integrações e migrações) e casos de uso. Os clientes nesse relatório representam todos os principais setores e variam de startups a muitas das maiores empresas do mundo.

Salvo indicação em contrário, este relatório apresenta e analisa dados de 1º de fevereiro de 2022 a 31 de janeiro de 2023, e o uso é medido pelo número de clientes. Quando possível, fornecemos comparações anuais para mostrar as tendências de crescimento ao longo do tempo.

Data Science e machine learning

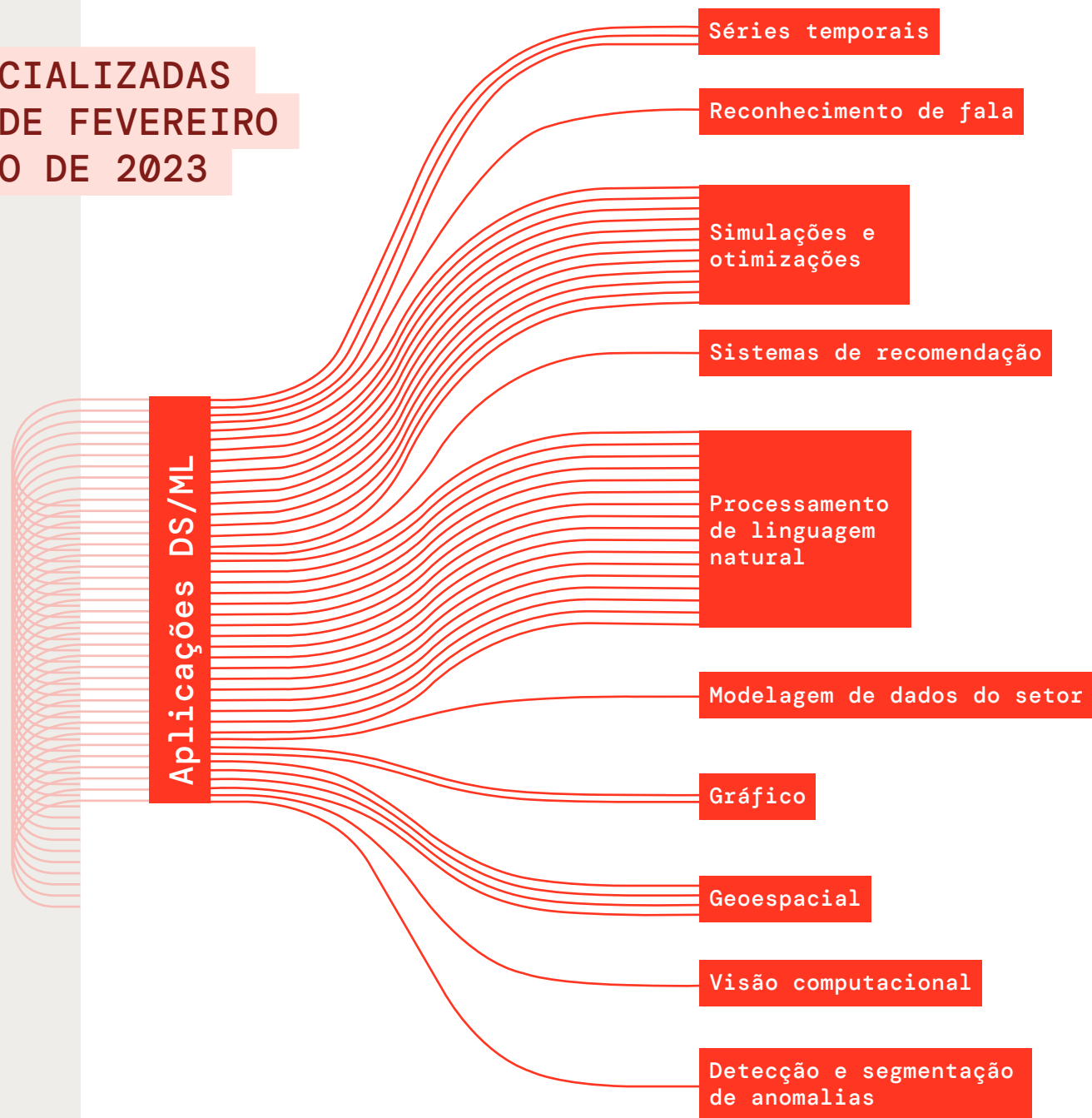
PROCESSAMENTO DE LINGUAGEM NATURAL E GRANDES MODELOS DE LINGUAGEM ESTÃO EM ALTA

Em todos os setores, as empresas aproveitam data science e machine learning (DS/ML) para acelerar o crescimento, melhorar a previsibilidade e aprimorar as experiências do cliente. Avanços recentes em grandes modelos de linguagem (LLMs) estão impulsionando as empresas a repensar a IA em suas próprias estratégias de dados.

Dado o cenário de DS/ML em rápida evolução, queríamos entender vários aspectos do mercado:

- Em quais tipos de aplicativos DS/ML as empresas estão investindo? Em especial, como são os dados dos LLMs?
- As empresas estão avançando na operacionalização de seus modelos de machine learning (MLOps)?

BIBLIOTECAS ESPECIALIZADAS EM PYTHON DS/ML DE FEVEREIRO DE 2022 A JANEIRO DE 2023



Observação: este gráfico apresenta o número exclusivo de notebooks usando bibliotecas de ML por dia em cada uma das categorias. Inclui bibliotecas usadas para os casos de uso específicos de resolução de problemas mencionados. Ele não inclui bibliotecas usadas em ferramentas para preparações de dados e modelagem.

O processamento de linguagem natural domina os casos de uso de machine learning

Para entender como as organizações estão aplicando IA e ML dentro do Lakehouse, agregamos o uso de bibliotecas Python especializadas, que incluem NLTK, Transformers e FuzzyWuzzy, em casos de uso populares de data science.¹ Analisamos os dados dessas bibliotecas porque o Python está na vanguarda dos novos desenvolvimentos em ML, análise avançada e IA, e tem sido consistentemente classificado como uma das [linguagens de programação mais populares](#) nos últimos anos.

Nosso caso de uso mais popular é o processamento de linguagem natural (NLP), um campo de rápido crescimento que permite que as empresas obtenham valor de dados textuais não estruturados. Isso possibilita que os usuários realizem tarefas que antes eram muito abstratas para código, como resumir o conteúdo ou extrair o sentimento das avaliações dos clientes. Nossos dados mostram que 49% das bibliotecas usadas estão associadas ao NLP. Os LLMs também se enquadram aqui. Dadas as inovações lançadas nos últimos meses, acreditamos que o NLP crescerá ainda mais nos próximos anos, pois é aplicado a casos de uso como chatbots, assistência de pesquisa, detecção de fraudes, geração de conteúdo e muito mais.

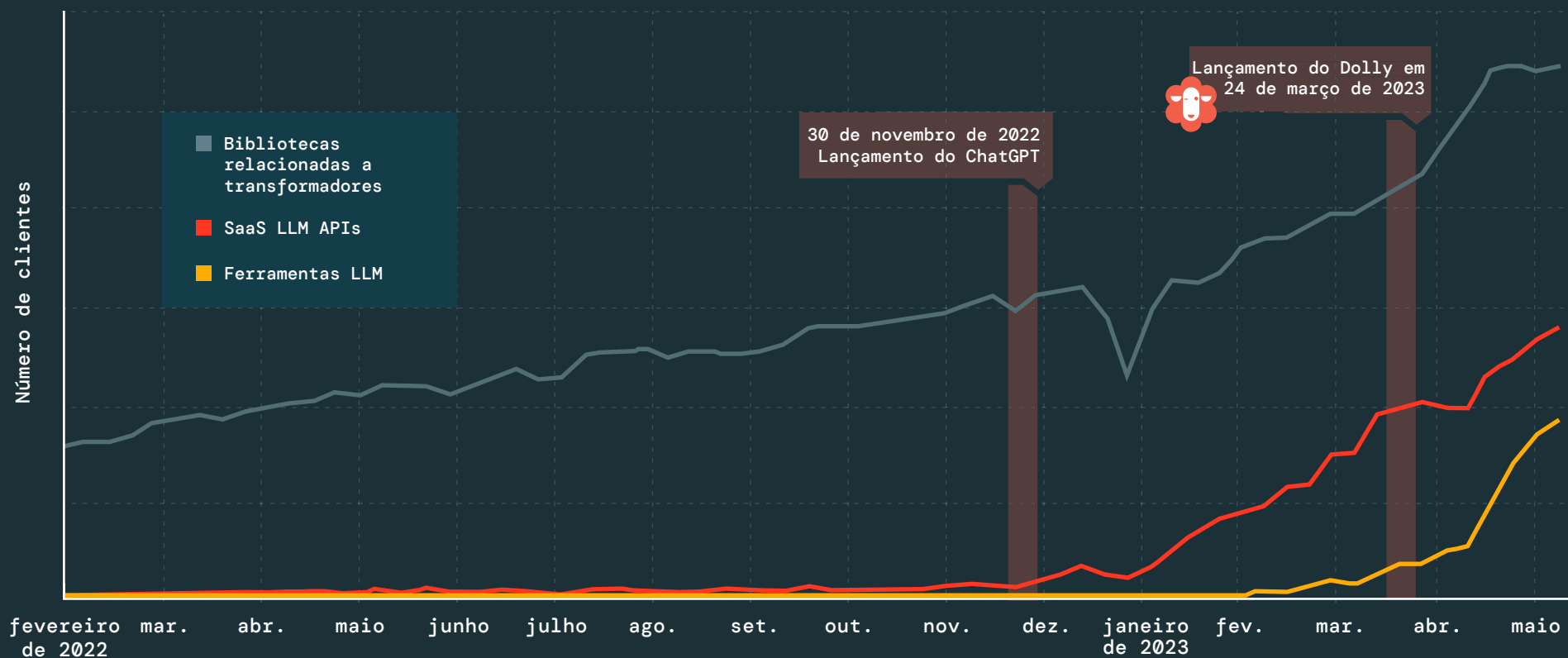
Nosso segundo aplicativo de DS/ML mais popular é o de simulações e otimização, que representa 30% de todos os casos de uso. Isso mostra que as organizações estão usando dados para modelar protótipos e resolver problemas de forma econômica.

Em nosso conjunto de dados, 49% das bibliotecas especializadas Python usadas estão associadas ao NLP

Muitos dos casos de uso de DS/ML são predominantemente utilizados por setores específicos. Embora representem uma parcela menor do total, são essenciais para muitas organizações. Por exemplo, a série temporal inclui previsões, um caso de uso especialmente popular em setores como varejo e bens de consumo embalados (CPG), que dependem muito da capacidade de prever a necessidade de cada item em cada loja.

¹ Esses dados não incluem bibliotecas de ML de uso geral, incluindo scikit-learn ou TensorFlow.

USO DE GRANDES MODELOS DE LINGUAGEM (LLMs)



Nota: existem vários tipos populares de bibliotecas Python que são comumente usadas para LLMs. Essas bibliotecas fornecem modelos e ferramentas pré-treinados para criar, treinar e implantar LLMs. Agrupamos essas bibliotecas com base no tipo de funcionalidade que elas oferecem.

Os dados caem consistentemente na última semana de dezembro devido à sazonalidade.

Grandes modelos de linguagem são a ferramenta do momento

Os LLMs são atualmente uma das áreas mais populares e investigadas no campo do PLN. Os LLMs têm sido fundamentais para permitir que as máquinas entendam, interpretem e gerem linguagem humana de uma maneira que antes era impossível, alimentando tudo, desde tradução automática até criação de conteúdo para assistentes virtuais e chatbots.

As bibliotecas relacionadas a transformadores têm se tornado mais populares mesmo antes de o ChatGPT colocar os LLMs em evidência. Nos últimos 6 meses, nossos dados mostram duas tendências em ascensão: as organizações estão criando seus próprios LLMs, que modelos como o GPT-4 mostram que podem ser bastante acessíveis e baratos. Além disso, estão usando modelos proprietários como o ChatGPT. As bibliotecas relacionadas a transformadores, como a Hugging Face, que são usadas para treinar LLMs, têm a maior adoção dentro do Lakehouse.

O segundo tipo mais popular são os LLMs SaaS, usados para acessar modelos como o OpenAI. Esta categoria cresceu exponencialmente em paralelo com o [lançamento do ChatGPT](#): o número de clientes Lakehouse que usam SaaS LLMs cresceu impressionantes 1310% entre o final de novembro de 2022 e o início de maio de 2023. (Em contraste, as bibliotecas relacionadas a transformadores cresceram 82 % neste mesmo período.)

As organizações podem aproveitar os LLMs usando APIs de LLM SaaS para usar serviços como ChatGPT da OpenAI ou operar seus próprios LLMs internamente.

Pensando em construir seu próprio aplicativo de LLM moderno? Essa abordagem poderia envolver o uso de bibliotecas Python especializadas relacionadas a transformadores para treinar o modelo, bem como ferramentas de LLM como LangChain para desenvolver interfaces imediatas ou integrações com outros sistemas.

DEFINIÇÕES DE LLM

- ♦ **Bibliotecas relacionadas a transformadores:** bibliotecas Python usadas para treinar LLMs (exemplo: Hugging Face)
- ♦ **SaaS LLM APIs:** bibliotecas usadas para acessar LLMs como um serviço (exemplo: OpenAI)
- ♦ **Ferramentas de LLM:** Cadeias de ferramentas para trabalhar e construir LLMs proprietários (exemplo: LangChain)

Experimentação e produção de machine learning crescem em todos os setores

A crescente demanda por soluções de ML e a maior disponibilidade de tecnologias levaram a um aumento significativo na experimentação e produção, duas partes distintas do ciclo de vida do modelo de ML. Analisamos o logging e *registro* de modelos no MLflow, uma plataforma de código aberto desenvolvida pela Databricks, para entender como o ML está em evidência e sendo adotado dentro das organizações.

O MLflow Model Registry foi lançado em maio de 2021. No geral, o número de modelos com log cresceu 54% desde fevereiro de 2022, enquanto o número de modelos registrados cresceu 411% no mesmo período. Esse crescimento no volume sugere que as organizações estão entendendo o valor de investir e alocar mais pessoas em ML.



MODELOS REGISTRADOS E EXPERIMENTAÇÃO DE ML

Durante a fase de experimentação de ML, cientistas de dados desenvolvem modelos projetados para resolver tarefas específicas. Depois de treinar os modelos, eles os testam para avaliar sua exatidão, precisão, recuperação (a porcentagem de instâncias positivas previstas corretamente de todas as instâncias positivas reais) e muito mais. Essas métricas são registradas (gravadas) para analisar o desempenho dos vários modelos e identificar qual abordagem funciona melhor para a tarefa.

Escolhemos modelos com log como proxy para medir a experimentação de ML, pois o servidor do MLflow Tracking foi projetado para facilitar o rastreamento e a reprodutibilidade do experimento.

MODELOS REGISTRADOS E PRODUÇÃO DE ML

Os modelos de produção passaram pela fase de experimentação e são implementados em aplicativos do mundo real. Normalmente, eles são usados para fazer previsões ou tomar decisões com base em novos dados. Registrar um modelo é o processo de gravação e armazenamento de metadados sobre um modelo treinado em um local centralizado que permita aos usuários acessar e reutilizar facilmente os modelos existentes. Registrar modelos antes da produção permite que as organizações assegurem consistência e confiabilidade na implementação e escalabilidade do modelo.

Escolhemos modelos registrados para representar a produção de ML porque o MLflow Model Registry foi projetado para gerenciar modelos que deixaram a fase de experimentação durante o resto de seu ciclo de vida.

As organizações testam inúmeras abordagens e variáveis antes de utilizar um modelo de ML. Queríamos entender: "Quantos modelos os cientistas de dados experimentam antes de migrar para a produção?"

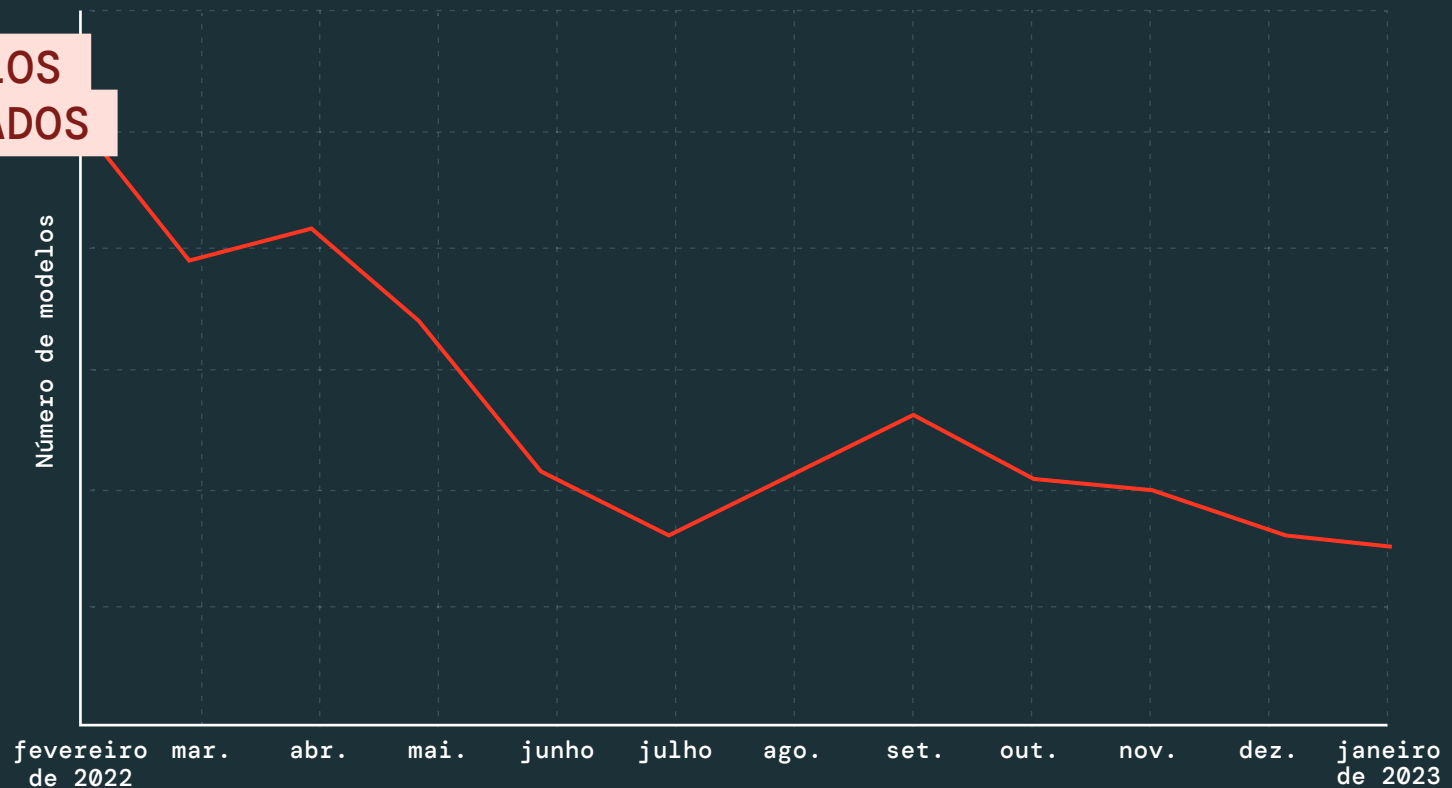
Nossos dados mostram que a proporção de modelos com log para registrados é de 2,9: 1 a partir de janeiro de 2023. Isso significa que, de aproximadamente três modelos experimentais, um será registrado como candidato à

produção. Essa proporção melhorou significativamente em relação ao ano anterior, quando vimos que, para cada cinco modelos experimentais, aproximadamente um foi registrado. Avanços recentes em ML, como bibliotecas de código aberto melhoradas como MLflow e Hugging Face, simplificaram radicalmente a construção e a colocação de modelos em produção. O resultado é que, hoje, 34% dos modelos com log são candidatos à produção, uma melhoria de mais de 20% há apenas um ano.

PROPORÇÃO DE MODELOS COM LOG X REGISTRADOS

2.9 : 1

Proporção de modelos com log
para registrados em janeiro
de 2023

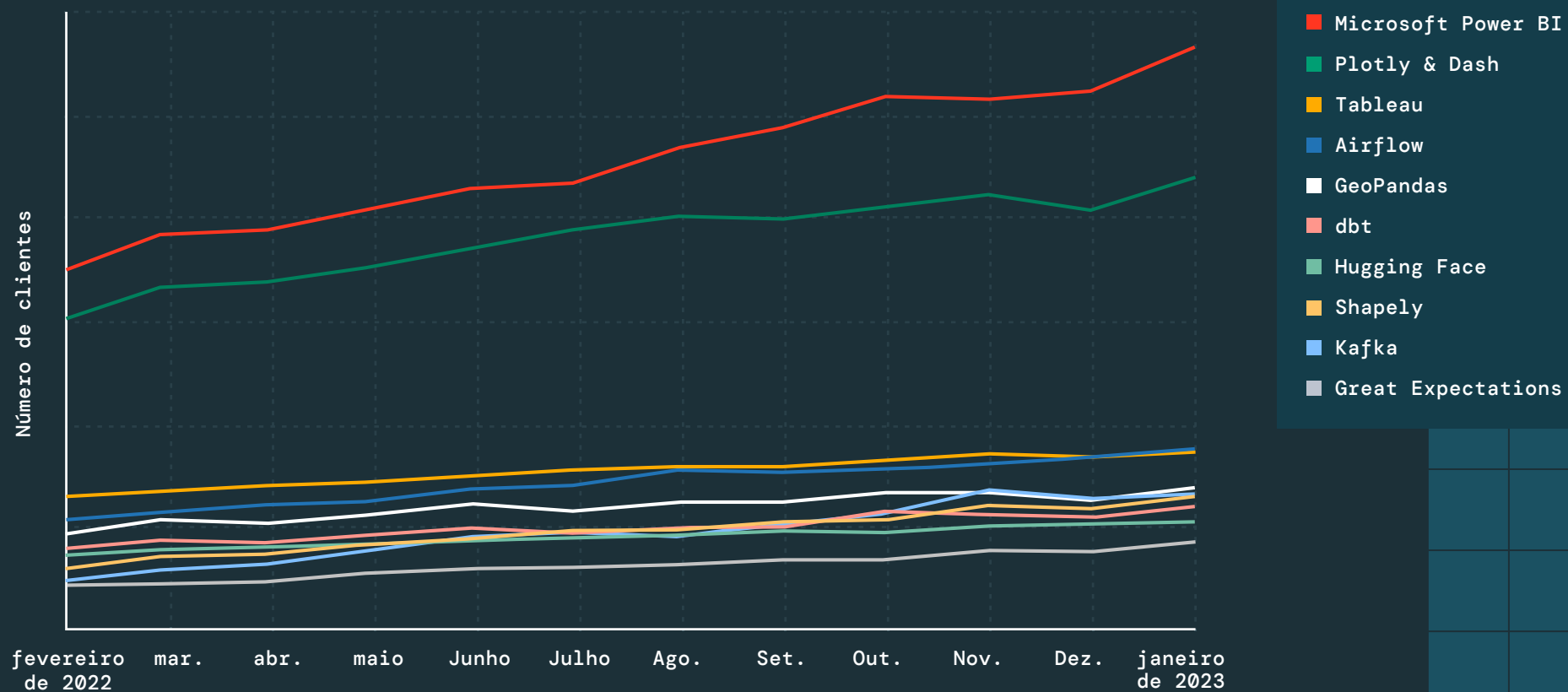


Dados e produtos de IA

O CONJUNTO MODERNO DE DADOS E IA

Nos últimos anos, surgiu a tendência de construir arquiteturas de dados unificadas e abertas em nossos próprios dados. Observamos que os líderes de dados estão optando por preservar a escolha, aproveitar os melhores produtos e oferecer inovação em suas organizações, democratizando o acesso aos dados para mais pessoas.

OS 10 PRINCIPAIS PRODUTOS DE DADOS E IA



Principais produtos de IA e dados

Uma pergunta comum dos clientes é: "Quais produtos de IA e de dados as outras empresas estão usando?" Como o Databricks Lakehouse é usado amplamente em todo o ecossistema, temos insights exclusivos sobre a adoção de centenas de produtos e serviços de dados pelo cliente.

Uma das nossas descobertas mais interessantes é que os produtos de código aberto estão dominando as principais classificações dos produtos de IA e dados; 8 dos nossos 10 produtos de IA e dados mais amplamente aceitos no Lakehouse são baseados em código aberto. As organizações estão escolhendo mais flexibilidade e compartilhamento de dados mais fácil, evitando estruturas e restrições proprietárias. Isso indica um sentimento crescente em todos os setores: plataformas abertas e produtos são essenciais para as estratégias de dados atuais.

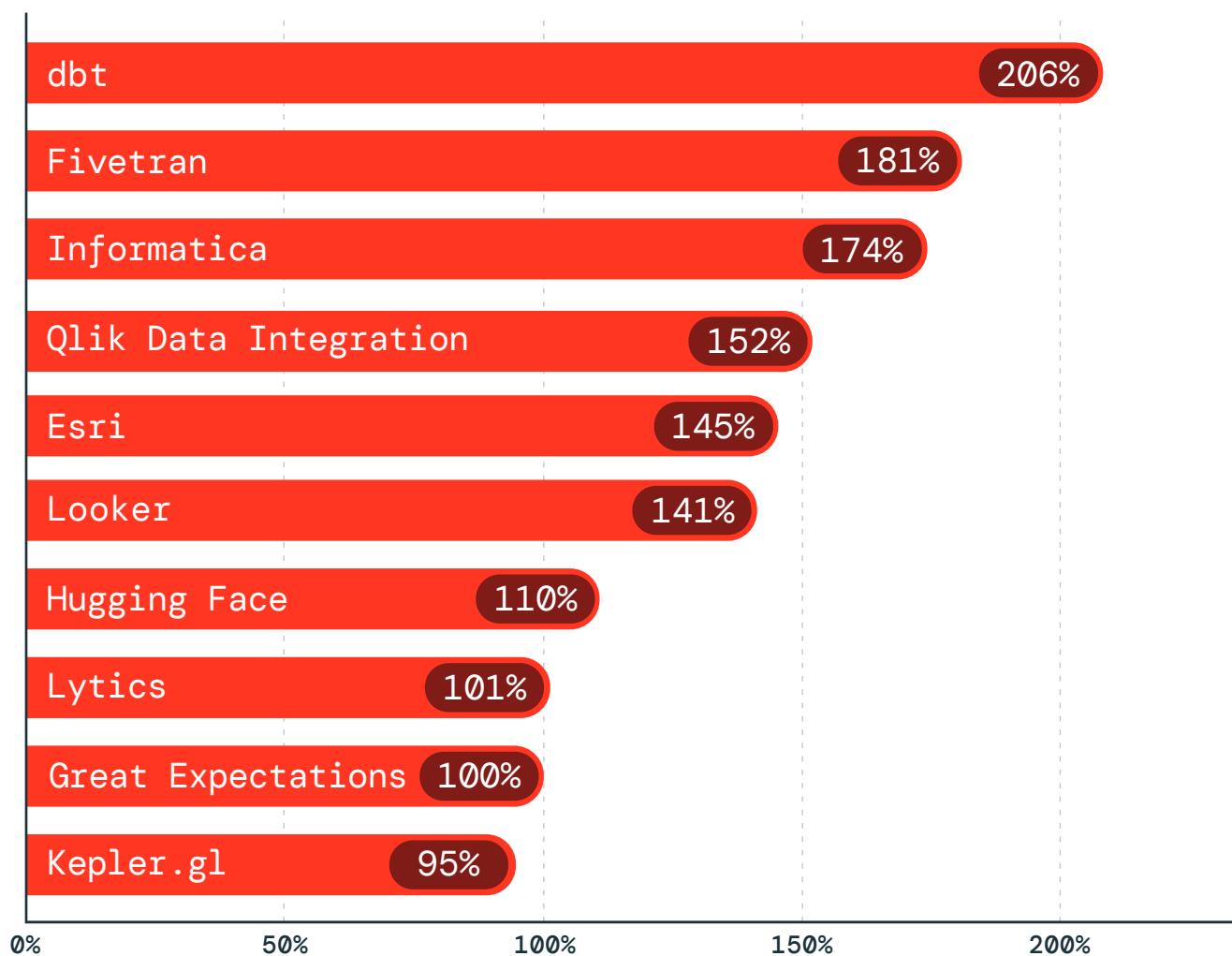
Os 10 principais produtos de IA e dados representam a natureza multifacetada da pilha moderna de dados e IA. Apesar de uma economia instável, as organizações continuam investindo em sua pilha. Nossos principais produtos incluem duas ferramentas de Business Intelligence (BI), dois produtos de DS/ML, três produtos de governança e segurança de dados e um produto de integração de dados. Dois também vêm de um mercado muito menor: o geoespacial.

Produtos e plataformas abertas são essenciais para as estratégias de dados atuais

O MICROSOFT POWER BI É O PRINCIPAL PRODUTO; OS PRODUTOS DE DATA SCIENCE E IA ESTÃO CRESCENDO RAPIDAMENTE

O Microsoft Power BI é o produto de IA e dados mais popular de 2023. Assim como os principais mercados de dados e IA, as ferramentas de BI são procedimentos operacionais padrão na maioria das empresas. Mas também vemos que as empresas estão priorizando data science e os produtos de IA. O segundo produto mais popular, o Plotly, é uma plataforma de baixo código, baseada em Python, que permite que as equipes de data science criem e dimensionem facilmente gráficos e mapas interativos. Classificada em 7º lugar, o código aberto Hugging Face democratiza a IA com sua biblioteca de transformadores que fornece acesso fácil a modelos abertos pré-treinados para uma variedade de tarefas. Essas duas ferramentas são criadas para tornar o ML mais rápido de adotar e fácil de usar em toda a organização.

PRODUTOS DE DADOS E IA QUE MAIS CRESCEM



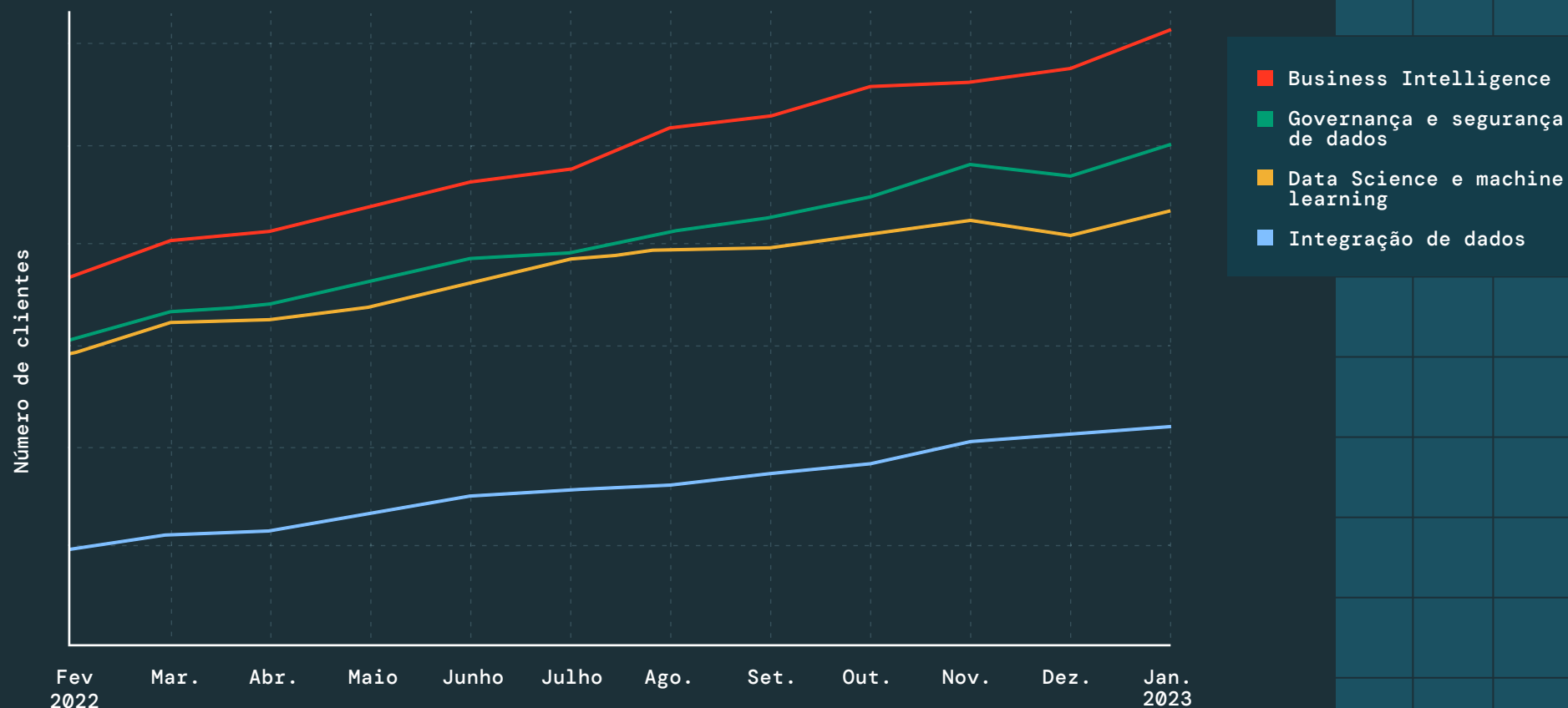
Crescimento anual por número de clientes



O DBT É O PRODUTO DE DADOS E IA DE CRESCIMENTO MAIS RÁPIDO EM 2023

À medida que as empresas se movem rapidamente para desenvolver casos de uso mais avançados com seus dados, elas investem em produtos mais novos que produzem conjuntos de dados confiáveis para relatórios, modelagem de ML e fluxos de trabalho operacionais. Assim, vemos o rápido desenvolvimento dos produtos de integração de dados. O dgt, uma ferramenta de transformação de dados, e o Fivetran, que automatiza pipelines de dados, são nossos dois produtos de IA e dados que crescem mais rápido. Isso sugere uma nova era do mercado de integração de dados com ferramentas desafiadoras avançando à medida que as empresas mudam para priorizar as iniciativas de DS/ML. Com o Great Expectations da Superconductive em nono lugar, um total de 50% de nossos produtos de crescimento mais acelerado representam a categoria de integração de dados.

CRESCIMENTO DOS MERCADOS DE DADOS E IA



Observação: neste gráfico, contamos o número de clientes que implementam um ou mais produtos de IA e dados em cada categoria. Essas quatro categorias não englobam todos os produtos. Os produtos Databricks, como o Unity Catalog, não estão incluídos nestes dados.

Mercados de IA e dados: business intelligence é padrão, as organizações investem em sua base de machine learning

Para entender como as organizações estão priorizando suas iniciativas de dados, agregamos todos os produtos de IA e dados no Databricks Lakehouse e os categorizamos em quatro mercados centrais: BI, governança e segurança de dados, DS/ML e integração de dados. Nosso conjunto de dados confirma que as ferramentas de BI são adotadas de forma mais ampla em organizações em relação a categorias mais emergentes — e continuam crescendo, com um aumento anual de 66% na adoção. Isso se alinha à tendência de mais organizações armazenando dados em um Lakehouse, coberto na próxima seção, Visualizações do Lakehouse.

**A integração de dados
é o mercado que se
expande mais rápido, com
crescimento anual de 117%**

Embora o BI seja geralmente o ponto de partida das organizações em sua jornada de dados, as empresas lidam cada vez mais com casos de uso de IA e dados mais avançados.

A DEMANDA POR PRODUTOS DE INTEGRAÇÃO DE DADOS ESTÁ CRESCENDO RAPIDAMENTE

Vemos o crescimento mais acelerado no mercado de integração de dados. Essas ferramentas permitem que uma empresa integre grandes quantidades de dados upstream e downstream em uma visão consolidada. Os produtos de integração de dados garantem que todas as iniciativas de BI e DS/ML sejam desenvolvidas com base sólida.

Embora seja mais fácil para mercados menores experimentarem um crescimento mais rápido, com 117% de aumento na adoção ao ano, o mercado de integração de dados está crescendo substancialmente mais rápido do que o BI. Essa tendência se encaixa com o rápido crescimento da adoção de ML que vemos em toda a Lakehouse, abordada na seção [DS/ML](#) do relatório.

Visualizações de Lakehouse

TENDÊNCIAS DE FORMATO DE DADOS E MIGRAÇÃO

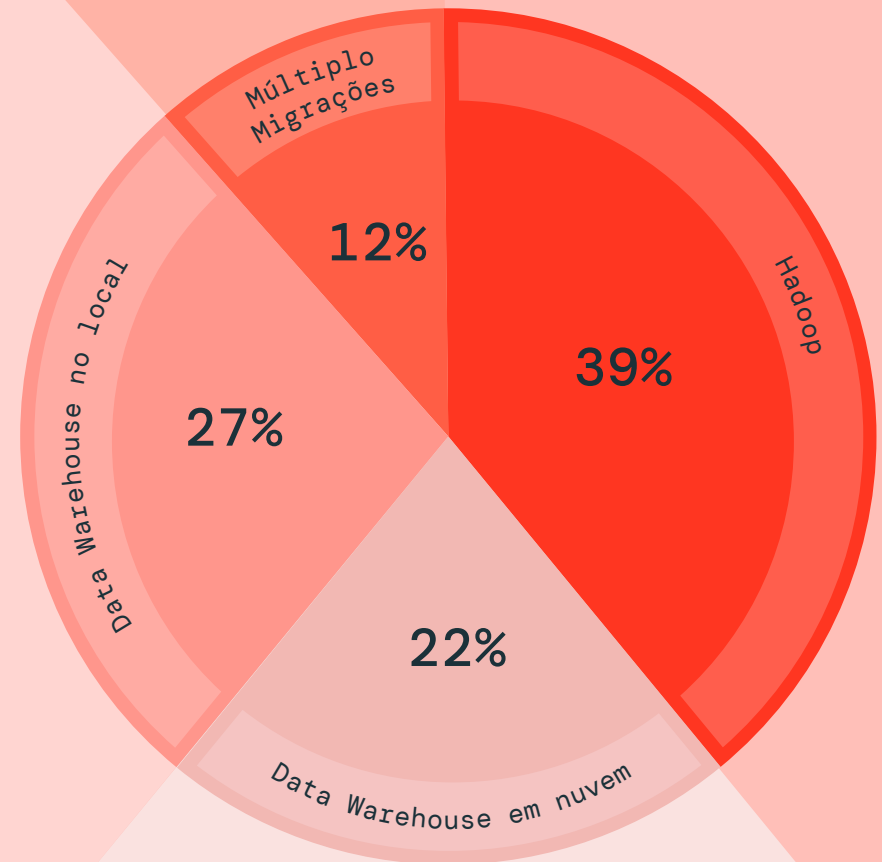
A migração de dados é um grande empreendimento: pode ser arriscada, cara e atrasar os cronogramas das empresas. Não é uma tarefa fácil. À medida que as organizações enfrentam as limitações, os desafios de escalabilidade e a carga de custos das plataformas de dados legadas, é cada vez mais provável que migrem para um novo tipo de arquitetura.

Tendências de migração: o melhor data warehouse é o Lakehouse

A Plataforma Lakehouse é uma alternativa atraente aos data warehouses tradicionais, pois é compatível com casos de uso avançados e DS/ML, permitindo que as organizações impulsionem sua estratégia geral de dados. As organizações estão cada vez mais usando o data lakehouse para armazenamento de dados, com as ferramentas de BI e integração de dados no topo, conforme mostrado pelos produtos de IA e dados mais populares. Para entender melhor de quais plataformas legadas as organizações estão se afastando, analisamos as migrações de novos clientes para Databricks.

Uma conclusão interessante é que cerca de metade das empresas que se mudam para o Lakehouse vêm de data warehouses. Isso inclui os 22% que estão migrando de data warehouses em nuvem. Também demonstra um foco crescente na execução de cargas de trabalho de data warehousing em um Lakehouse e a unificação de plataformas de dados para reduzir custos.

ORIGEM DAS MIGRAÇÕES DE NOVOS CLIENTES PARA DATABRICKS



Expansão: o volume de dados no Delta Lake cresceu 304% em relação ao ano anterior

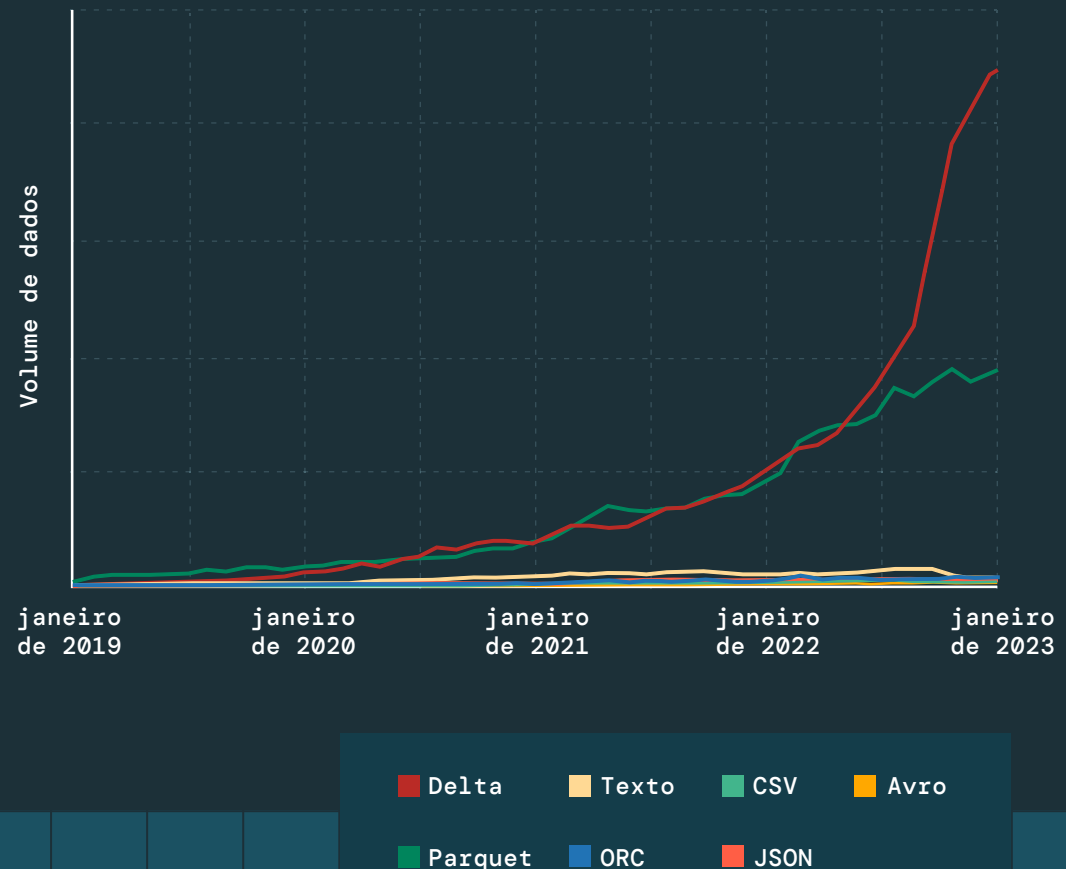
Com a explosão do volume de dados, uma proporção cada vez maior está na forma de dados semiestruturados e não estruturados. Anteriormente, as organizações tinham que gerenciar várias plataformas diferentes para seus dados estruturados, não estruturados e semiestruturados, o que causava complexidade desnecessária e altos custos.

O Lakehouse resolve esse problema fornecendo uma plataforma unificada para todos os tipos e formatos de dados.

Delta Lake é a base do Databricks Lakehouse.

O formato Delta Lake engloba dados estruturados, não estruturados e semiestruturados. Seu uso aumentou nos últimos 2 anos. Quando comparados ao crescimento constante, estável ou decrescente de outros formatos de armazenamento (por exemplo, texto, JSON e CSV), nossos dados mostram que um número cada vez maior de organizações está recorrendo ao Delta Lake para gerenciar seus dados. Em junho de 2022, o Delta Lake ultrapassou o Parquet como a fonte de data lake mais popular, atingindo um crescimento de 304% em relação ao ano anterior.

VOLUME DE DADOS GERENCIADOS, POR FORMATO DE ARMAZENAMENTO



Data warehousing cresce, com ênfase em serverless

Nos últimos 2 anos, as empresas aumentaram muito o uso de data warehousing na Plataforma Lakehouse. Isso é demonstrado principalmente pelo uso do Databricks SQL — o data warehouse serverless no Lakehouse — que mostra um crescimento de 144% em relação ao ano anterior. Isso sugere que as organizações estão abandonando cada vez mais os data warehouses tradicionais e são capazes de realizar todo o seu BI e análises em um Lakehouse.

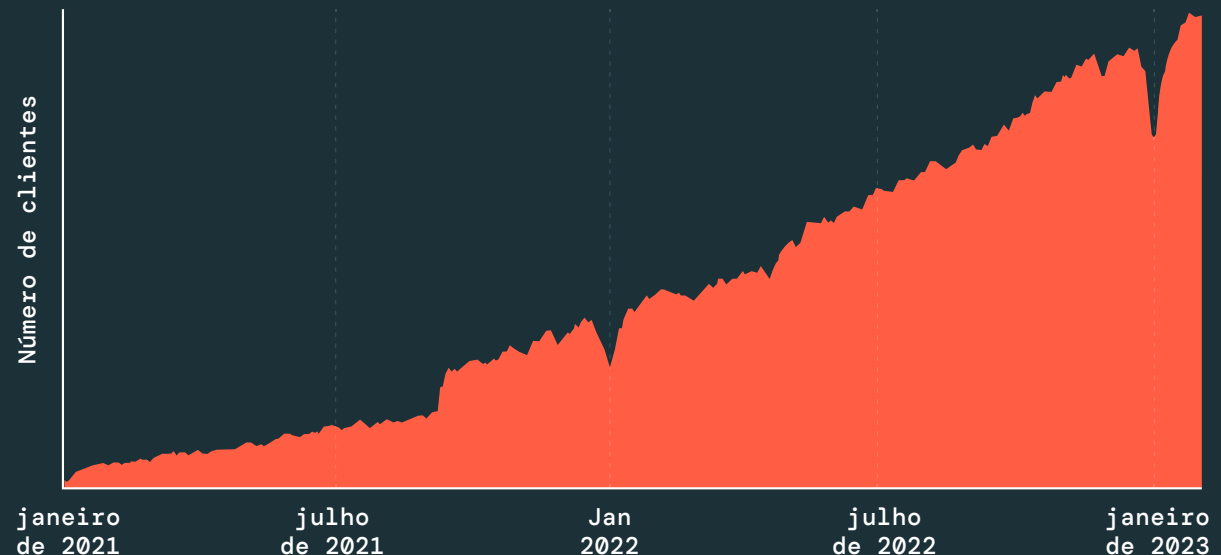
Data
Warehouse

Plataforma
Lakehouse

DATA WAREHOUSING NO LAKEHOUSE COM DATABRICKS SQL

Observação: Há um pico em outubro de 2021 como resultado do lançamento preliminar do Databricks SQL, seguido pela disponibilidade geral em dezembro de 2021.

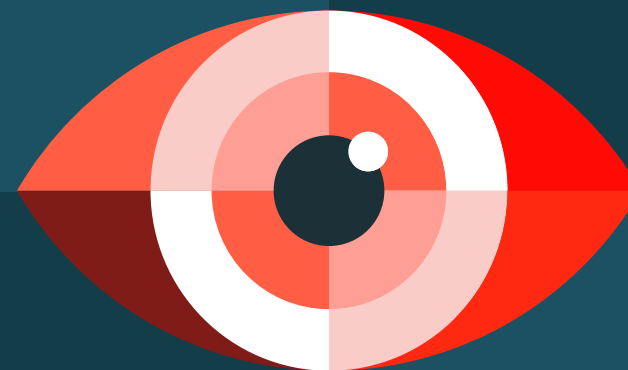
Os dados caem consistentemente na última semana de dezembro devido à sazonalidade.



CONCLUSÃO

Geração de IA

Embora o produto de IA e dados mais popular seja o Microsoft Power BI, estamos entusiasmados com o fato de que as empresas estão progredindo para casos de uso de ML e IA mais avançados, e a pilha moderna de dados e IA está evoluindo para acompanhar. Junto com o rápido crescimento das ferramentas de integração de dados (incluindo nosso DBT de crescimento mais rápido), estamos vendo o rápido aumento do uso de NLP e LLM em nosso próprio conjunto de dados, e não há dúvida de que os próximos anos verão uma explosão nessas tecnologias. Nunca foi tão claro: as empresas que aproveitam o poder do DS/ML liderarão a próxima geração de dados.



Sobre a Databricks

Databricks é a empresa de dados e IA. Mais de 9.000 organizações em todo o mundo — incluindo Comcast, Condé Nast e mais de 50% das empresas da Fortune 500 — contam com a Databricks Lakehouse Platform para unificar seus dados, análises e IA. A Databricks está sediada em São Francisco, com escritórios em todo o mundo. Fundada pelos criadores originais do Apache Spark™, Delta Lake e MLflow, a Databricks tem a missão de ajudar as equipes de dados a resolver os problemas mais difíceis do mundo. Para saber mais, siga Databricks no [Twitter](#), [LinkedIn](#) e [Instagram](#).

DESCUBRA O LAKEHOUSE



© Databricks 2023. Todos os direitos reservados. Apache, Apache Spark, Spark e o logotipo Spark são marcas comerciais da Apache Software Foundation | [Termos de uso](#)