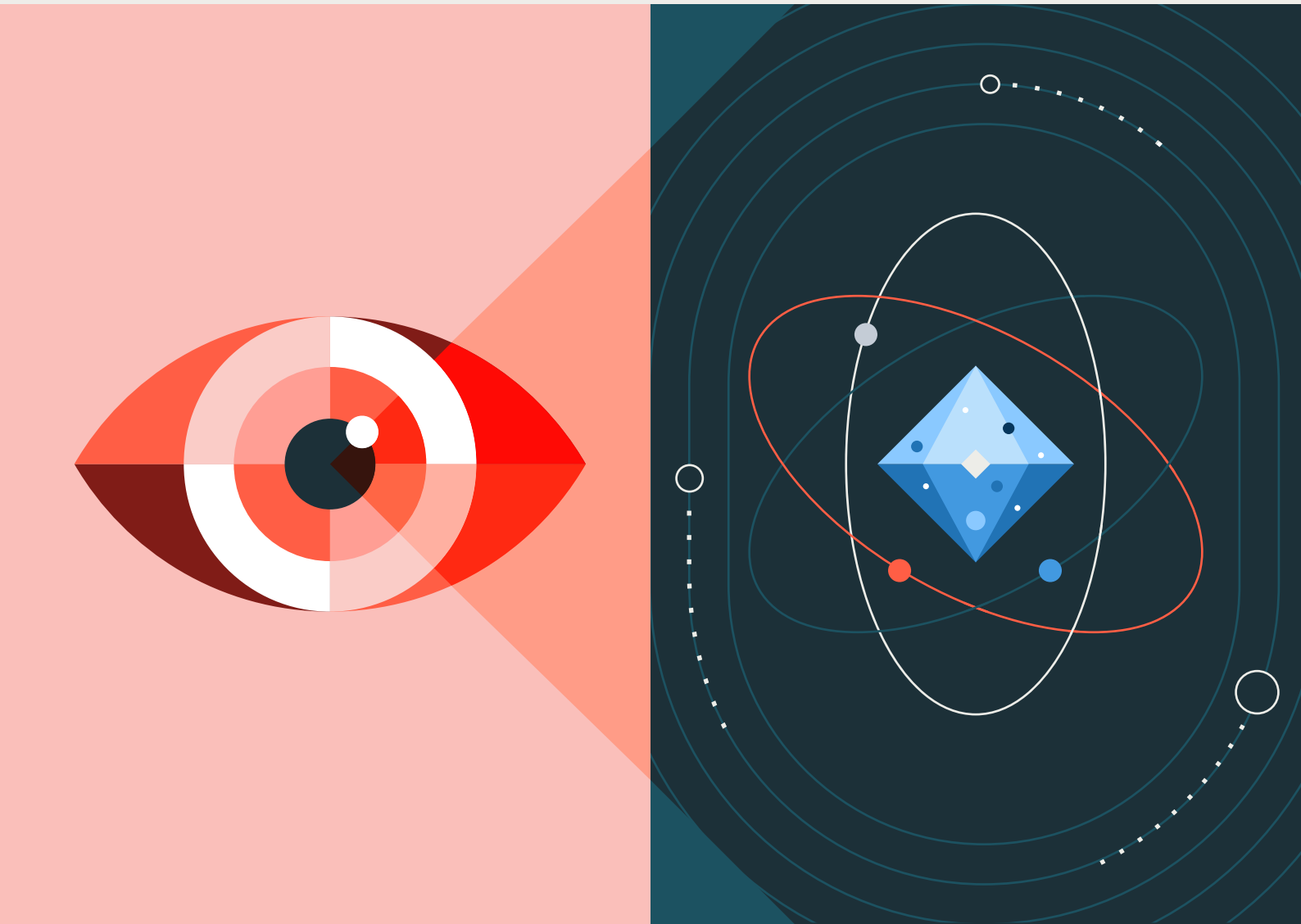
databricks

# データ + AI のトレンド 2023 年

Databricks レイクハウスの活用状況からのインサイト



# データと AI
# 黄金時代の到来

# はじめに

ChatGPT が公開されて半年、世界は AI の大きな可能性に目覚めました。AI による発見、モデルの改善、新製品の市場投入がかつてないペースで進行し、データ・AI 戦略は世界中の組織で話題の中心となっています。Databricks は、AI が次世代の製品やソフトウェアのイノベーションの先駆けとなると考えており、その動きは市場で既に始まっています。次世代をリードしようとする企業と経営幹部にとって、AI を理解し、活用する術を知ることは、欠くことのできない条件となります。

本レポートは、9,000 社を超える世界中の Databricks のお客さまを対象に、データと AI の活用状況を調査した結果をまとめたものです。Databricks レイクハウスは、企業の全データエステートに関わるビジネスインテリジェンス（BI）アプリケーションと AI アプリケーションを統合することで、どの製品やテクノロジーが広く使用され急成長しているか、どのような種類のデータサイエンス・機械学習（DS／ML）アプリケーションが開発されているかなど、データと AI の活用状況について独自の優れた視点を提供します。

# 今回の調査で明らかになった主なポイント

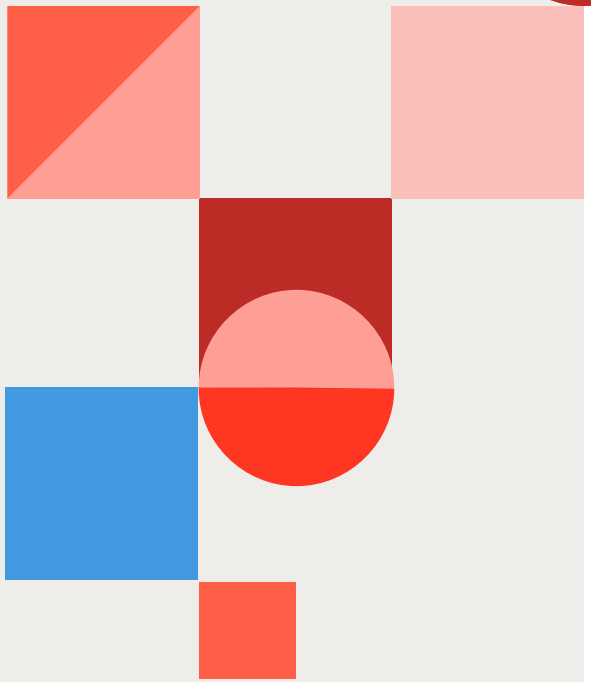
企業による機械学習と大規模言語モデル（LLM）の導入が急速に進んでいる。自然言語処理（NLP）がユースケースの大半を占め、LLMへの注目が加速している。

現在のデータ・AI市場において優位性が高いのはオープンソースである。最も広く利用されているデータ・AIの10製品中8製品がオープンソースをベースにしている。

レイクハウスを利用してデータのウェアハウス化を行う組織が増えている。Microsoft Power BIがデータ・AI製品で最も多く利用されている製品であること、dbtとFivetranが急成長しているデータ統合ツールであること、Databricks SQLの導入が加速していることが、これを証明している。

Databricksでは、これらのトレンドを共有することで、データリーダーが自社組織をベンチマークし、データとAIの時代に必要な戦略策定の一助となることを願っています。

# 重要な発見

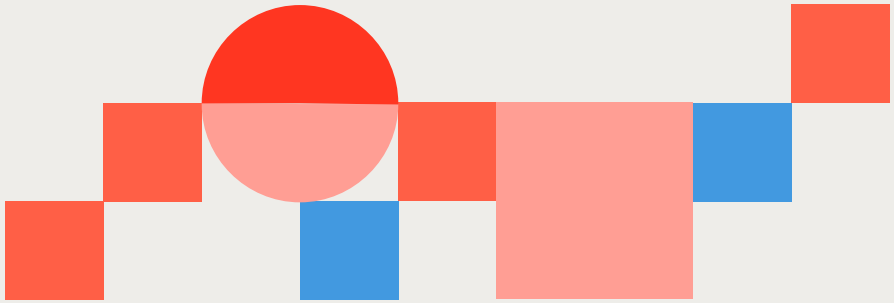


## 1 データサイエンス・機械学習：NLP と LLM の需要が高まっている

- SaaS LLM API（ChatGPT のようなサービスへのアクセスに使用される API）を利用する企業数は、2022 年 11 月末から 2023 年 5 月初めまでの間に 1310% 増加している。
- NLP は、Python データサイエンスライブラリの 1 日あたりの使用量の 49% を占めており、最も利用されているアプリケーションである。
- より多くのモデルが本番運用されている（前年比 411% 増）。同時に、機械学習の実験も増えている（前年比 54% 増）。
- 機械学習の利用が効率化している。実験と本運用のモデルの比率は、1 年前は 5 対 1 であったが、現在は 3 対 1 になっている。

## 2 データ・AI のトップ製品と市場

- 2023 年現在、Databricks レイクハウスと連携するデータ・AI 製品のうち、最も多く利用されているのは Microsoft Power BI である。
- 導入数が著しく増えているデータ・AI 製品は dbt であり、利用ユーザー数は前年比で 206% 増加している。
- 導入数が多い 10 製品中 8 製品は、オープンソースベースである。
- データ統合製品の市場が、Databricks レイクハウスとの連携において最も急成長しており、前年比 117% の成長率を示している。



## 3 導入と移行の傾向

- レイクハウスに移行するユーザーの 61% は、オンプレミスまたはクラウドのデータウェアハウスからの移行である。
- Delta Lake のデータ量は、前年比で 304% 増加している。
- レイクハウスを利用してデータのウェアハウス化を行うケースが増えており、Databricks SQL を活用したサーバーレスデータウェアハウスの導入は、前年比で 144% 増加している。

# 本レポートにおける調査手法

「データ＋ AI のトレンド 2023 年」は、Databricks のお客さまから収集した完全に匿名化されたデータを分析し、Databricks レイクハウスと、統合ツールの広範なエコシステムをどのように利用しているかについてまとめた調査結果レポートです。本レポートは、機械学習の導入、データアーキテクチャ（統合と移行）、ユースケースに重点を置いています。本レポートが調査対象としたお客さまは、各主要業界のスタートアップ企業から大規模エンタープライズ企業まで、多岐にわたります。

注意書きのあるものを除き、本レポートで使用・分析したデータは 2022 年 2 月 1 日～2023 年 1 月 31 日までの期間を対象としており、利用数を顧客数で測定しています。また、時間の経過に伴う成長傾向を示すために、可能な限り前年比を記載しています。

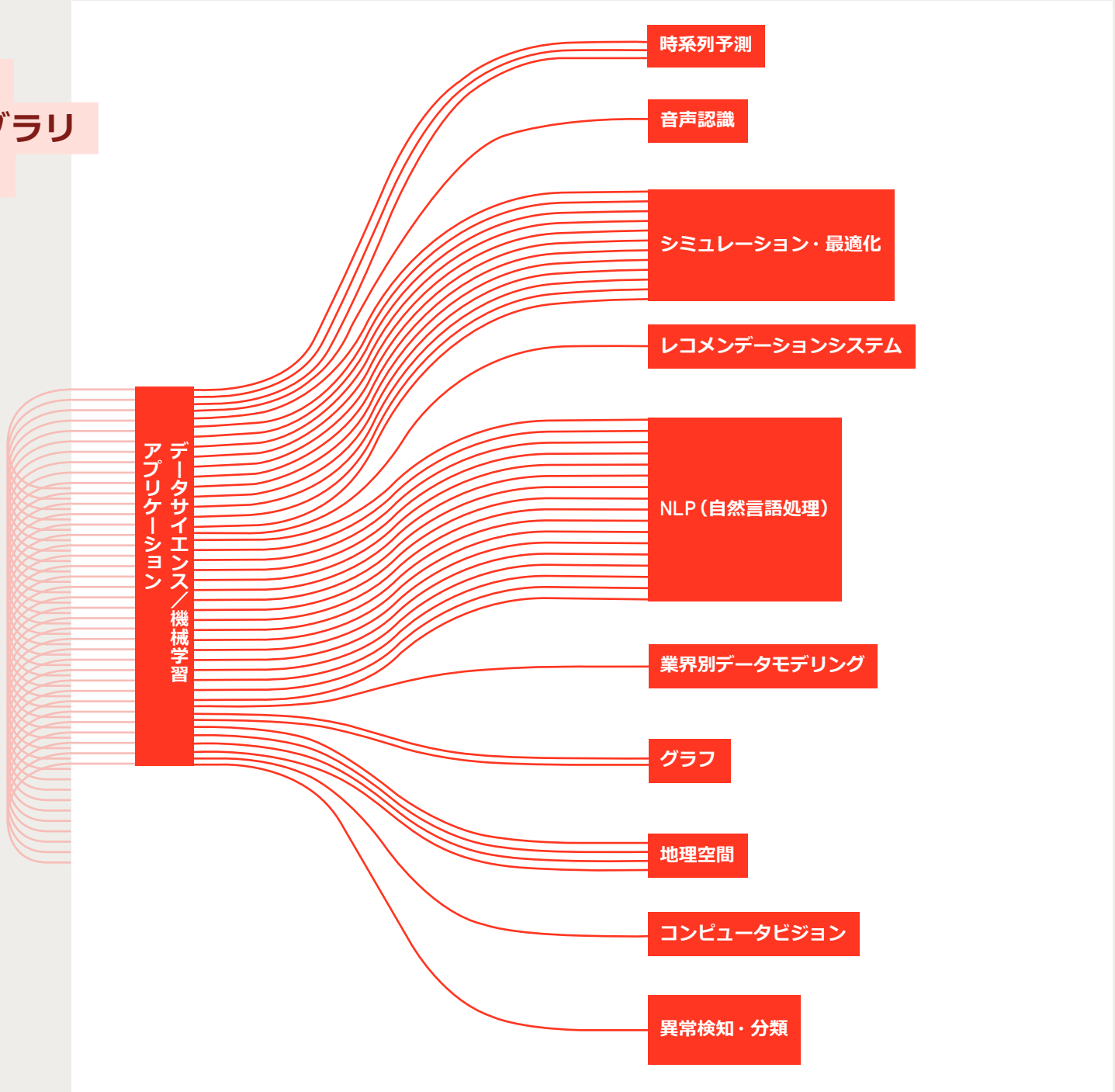
# データサイエンスと機械学習

## NLP（自然言語処理）と LLM（大規模言語モデル）の需要が高まっている

業界を問わず、多くの企業が、データサイエンスと機械学習（DS／ML）の活用を通じて、成長の加速、予測可能性の改善、顧客体験の向上を図っています。大規模言語モデル（LLM）の最近の進歩は、企業が AI を再評価し、自社のデータ戦略として活用する契機を生み出しています。急速に進化する DS／ML の環境を踏まえ、Databricks では、市場における次のような側面を解明したいと考えました。

- 企業はどのような DS／ML アプリケーションに投資しているか？特に、最近話題の LLM に関連するデータはどのようなものか？
- 企業は機械学習モデル（MLOps）の運用化を進めているか？

# データサイエンス・機械学習に利用されている Python ライブラリ（2022 年 2 月～2023 年 1 月）




注：このチャートは、各カテゴリの 1 日あたりの ML ライブラリを使用するノートブックの数を示しており、特定の問題解決ユースケースに使用されたライブラリも含まれています。データの準備やモデリングのためのツールに使用されるライブラリは含まれていません。

# 自然言語処理が機械学習ユースケースの主流となっている

どのような AI/ML がレイクハウスで実行されているかを知るため、NLTK、Transformer、FuzzyWuzzy などの Python ライブラリの使用状況を、多く利用されているデータサイエンスのユースケースで分類しました。[1] ライブラリのデータを調査対象とした理由は、Python が、ML、高度な分析、AI の開発の最先端にあり、近年最も使用されているプログラミング言語の 1 つとして常に上位にランク付けされているためです。

最も多いユースケースは自然言語処理（NLP）です。構造化されていないテキストデータからの価値創出を可能にする NLP は、急速に成長している領域です。NLP により、コンテンツの要約やカスタマーレビューからの感情の抽出など、これまでコーディングでは困難だった抽象的なタスクをユーザーが実行できるようになります。調査対象のデータセットでは、使用されているライブラリの 49% が NLP に関連しています。これには LLM も含まれています。ここ数か月の間に発表されたイノベーションからも、NLP は、チャットボットや研究支援、不正検知、コンテンツ生成といったユースケースに利用され、今後さらに普及することが予測できます。

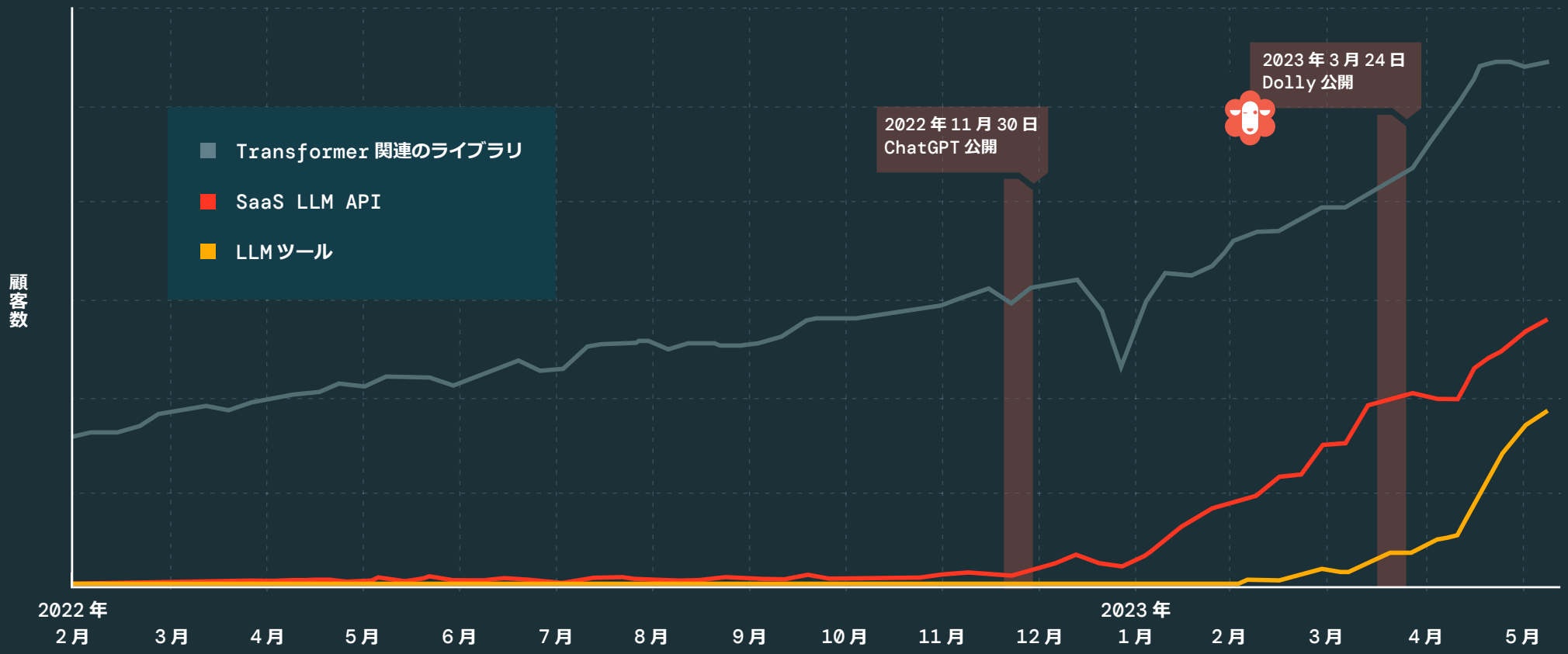
NLP の次に多く利用されている DS／ML アプリケーションはシミュレーションと最適化であり、全体の 30% を占めています。これは、組織がデータを利用してプロトタイプをモデル化し、コスト効率のよい問題解決をめざしていることを示しています。

**調査対象のデータセットでは、使用されている Python ライブラリの 49% が NLP に関連している**

DS／ML のユースケースの多くは、主に特定の業界で活用されています。全体に占める割合は小さくても、多くの組織にとってミッションクリティカルなユースケースです。例えば、時系列予測の予測機能は、各商品・各店舗の需要予測が必要なリテール・消費財などの業界でよく使用されています。

1. このデータには、scikit-learn や TensorFlow などの汎用 ML ライブラリは含まれていません。

# 大規模言語モデル（LLM）利用状況の推移




注：

- LLM には、よく使用される Python ライブラリが数種類あります。これらのライブラリは、事前学習済みのモデルのほか、LLM の構築やトレーニング、デプロイメントのためのツールを提供します。上の図は、これらのライブラリを機能のタイプに基づいてグループ化したものです。
- 12 月最終週の使用量がどれも少ないのは、季節的要因によるものです。

# 大規模言語モデル（LLM）がトレンドを牽引

LLM は現在、NLP の分野で最も活気があり注目されている領域のひとつです。LLM は、従来の手法では困難だったマシンによる人間の言語の理解・解釈・生成に大きく貢献してきました。今日では、機械翻訳からコンテンツ作成、バーチャルアシスタントやチャットボットに至るまで、さまざまな機能を支えています。

Transformer 関連のライブラリは、ChatGPT によって LLM が世間に認識される前から広く採用されてきました。過去 6 か月間のデータは、2 つの顕著な傾向を示しています。第一に、組織において独自の LLM の構築が進んでおり、Dolly のような、アクセスが容易で安価なモデルが採用されていること。第二に、ChatGPT のようなプロプライエタリなモデルが使用されていること。LLM のトレーニングに使用される Hugging Face などの Transformer 関連のライブラリは、レイクハウス内で最も多く採用されています。

2 番目に多く採用されているタイプは SaaS 型 LLM で、OpenAI のようなモデルにアクセスするために使用されます。このカテゴリは、ChatGPT の公開と並行して指数関数的に成長しています。SaaS 型 LLM を利用しているレイクハウスの顧客数は、2022 年 11 月末から 2023 年 5 月初めにかけて、1310% という驚異的な伸びを示しています。（一方で、同期間における Transformer 関連のライブラリの伸びは 82% でした。）

企業が LLM を活用するには、SaaS LLM API を介して OpenAI の ChatGPT のようなサービスを利用するか、自社開発の LLM を社内で運用する方法があります。

モダンな LLM アプリケーションを独自に構築するには、モデルのトレーニング専用の Transformer 関連の Python ライブラリを使用したり、LangChain のような LLM ツールを使用してプロンプトのインターフェースの開発や他のシステムとの統合を行う必要があります。

### LLM の定義

- **Transformer 関連のライブラリ：** LLM のトレーニングに使用される Python ライブラリ（例：Hugging Face）
- **SaaS LLM API：** サービスとしての LLM へのアクセスに使用されるライブラリ（例：OpenAI）
- **LLM ツール：** プロプライエタリな LLM の使用や開発を支援するツールチェーン（例：LangChain）

# 機械学習の実験と本番運用が各業界で本格化している

ML ソリューションに対する需要の高まりとテクノロジーの普及により、実験と本番運用の件数が大幅に増加しています。これらは、ML モデルのライフサイクルにおいて全く異なる位置を占めています。ML のトレンドと組織内での採用形態を理解するため、Databricks が開発したオープンソースプラットフォームである MLflow の各モデルについて、ログ記録済みと登録済みのモデルの状況を検討します。

MLflow モデルレジストリは 2021 年 5 月に公開されました。ログ記録されたモデル数は 2022 年 2 月から 54% 増加しています。一方、登録済みモデル数は同期間に 411% の増加を見せています。この増加は、ML に投資し、より多くの人員を割り当てることの価値を組織が理解していることを示唆しています。

### ログに記録されたモデルと ML の実験

ML の実験段階において、データサイエンティストは与えられたタスクを解決するためのモデルを開発します。モデルはトレーニング後にテストされ、正答率、適合率、再現率（実際のポジティブなインスタンスのうち、正しく予測されたインスタンスの割合）などの評価が行われます。評価結果はログとして記録され、さまざまなモデルのパフォーマンスの分析と、どのアプローチが与えられたタスクに最も適しているかの判定に使用されます。

ML 実験を測定するプロキシにログに記録されたモデルを選択したのは、MLflow 追跡サーバーが実験の追跡と再現性を容易にするように設計されているためです。

### 登録済みモデルと ML の本番運用

本番運用モデルは、実験段階を経て、実際のアプリケーションに統合されます。本番運用モデルは通常、新たなデータに基づいて予測や意思決定を行うために使用されます。モデルの登録とは、学習済みモデルに関するメタデータを一元管理された場所に記録・保存するプロセスです。これにより、ユーザーは既存のモデルに容易にアクセスして再利用できます。本番運用前にモデルを登録することで、モデルのデプロイメントとスケーリングにおける一貫性と信頼性を確保できます。

ML の本番運用を表すモデルに登録済みモデルを選択したのは、MLflow モデルレジストリが、実験段階を終了したモデルをライフサイクルの残りの期間にわたって管理するのに適しているためです。

1つの ML モデルが本番運用されるまでには、さまざまなアプローチや可変的要素を考慮したテストが行われます。私たちは、「1つのモデルを本番運用するまでに、データサイエンティストはいくつのモデルを実験しているか」を知りたいと考えました。

2023 年 1 月現在、ログ記録済みモデルと登録済みモデルの比率は 2.9 対 1 でした。すなわち、3 つの実験モデルに対して、1 つのモデルが本番運用の候補として登録されたことになります。この比率は、1 年前は 5 対 1 であったのと比べると、大幅に改善されています。

MLflow や Hugging Face のような改良されたオープンソースライブラリなど、ML の最近の進歩により、モデルの構築と本番環境への投入が大幅に簡素化されています。その結果、現在ログに記録されたモデルの 34% が本番運用の候補となっています。わずか 1 年前の 20% 超に比べて改善が見られます。

## 登録済みモデル vs. 記録済みモデルの比率

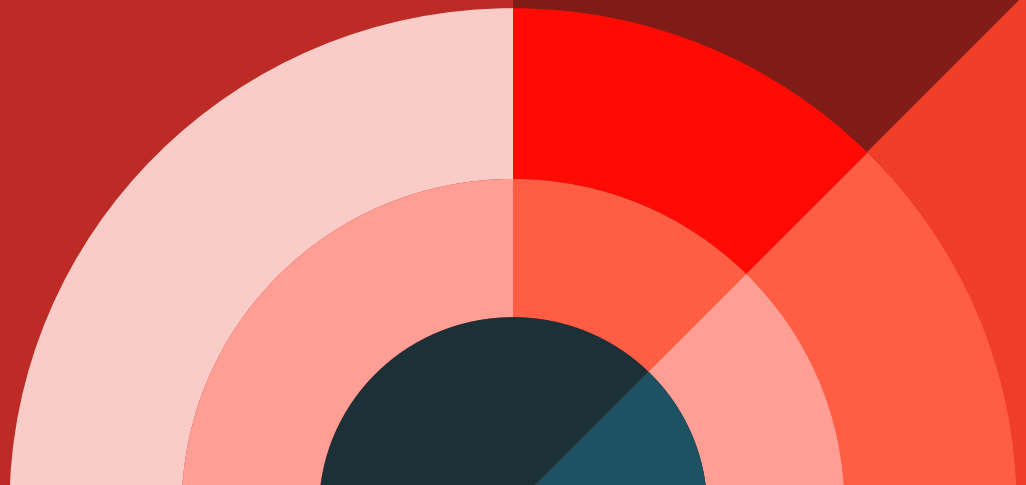
**2.9 : 1**

**登録済みモデル vs. 記録済みモデルの比率（2023 年 1 月現在）**

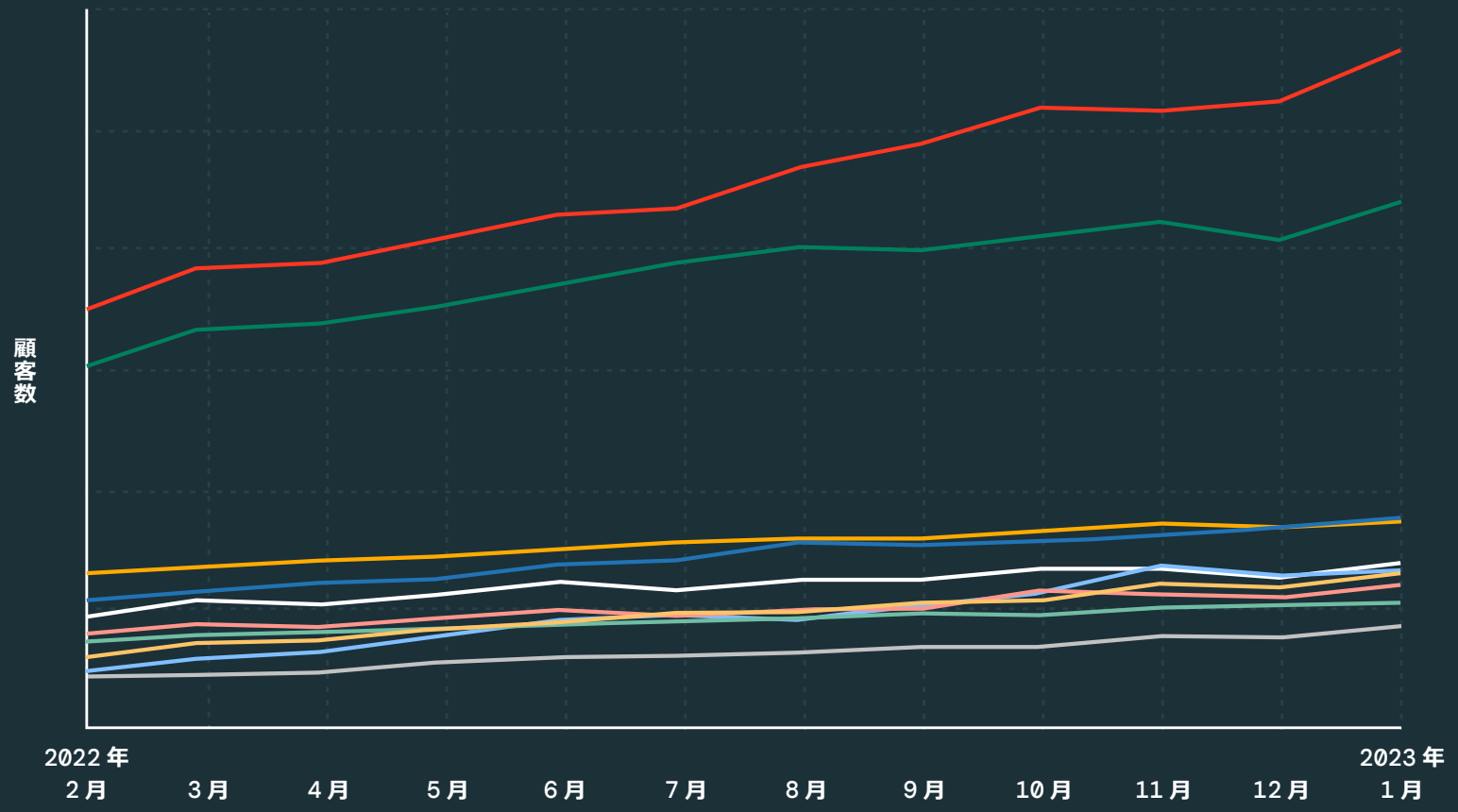
# データ・AI 製品

## データ+ AI スタックのモダナイズ

ここ数年、オープンで統合されたデータアーキテクチャを構築しようという傾向が、調査対象のデータで明らかになっています。データリーダーは、採用する製品の選択肢があることを望み、優れた製品を選択して活用し、データアクセスを民主化することで組織全体におけるイノベーションの促進を図っています。



# データ・AI 製品トップ 10



# 主要なデータ・AI 製品

Databricks では、お客さまから、「他社ではどのようなデータ・AI 製品を使用しているのかを知りたい」というお問い合わせを数多くいただきます。Databricks のレイクハウスは、エコシステム全体で幅広く利用されているため、私たちは、何百ものデータ製品やサービスの導入事例から得られる独自のインサイトを持っています。

特筆すべき発見の1つは、広く利用されているデータ・AI 製品の上位を独占しているのがオープンソース製品であることです。レイクハウスで最も採用されているデータ・AI 製品では、10 製品のうち 8 製品がオープンソースを基盤としています。組織の多くが、柔軟性の高さとデータ共有の容易さを優先し、プロプライエタリなテクノロジーによる障壁や制限を回避する傾向にあります。このことは、現在のデータ戦略において、どの業界においてもオープンなプラットフォームと製品が不可欠であることを示しています。

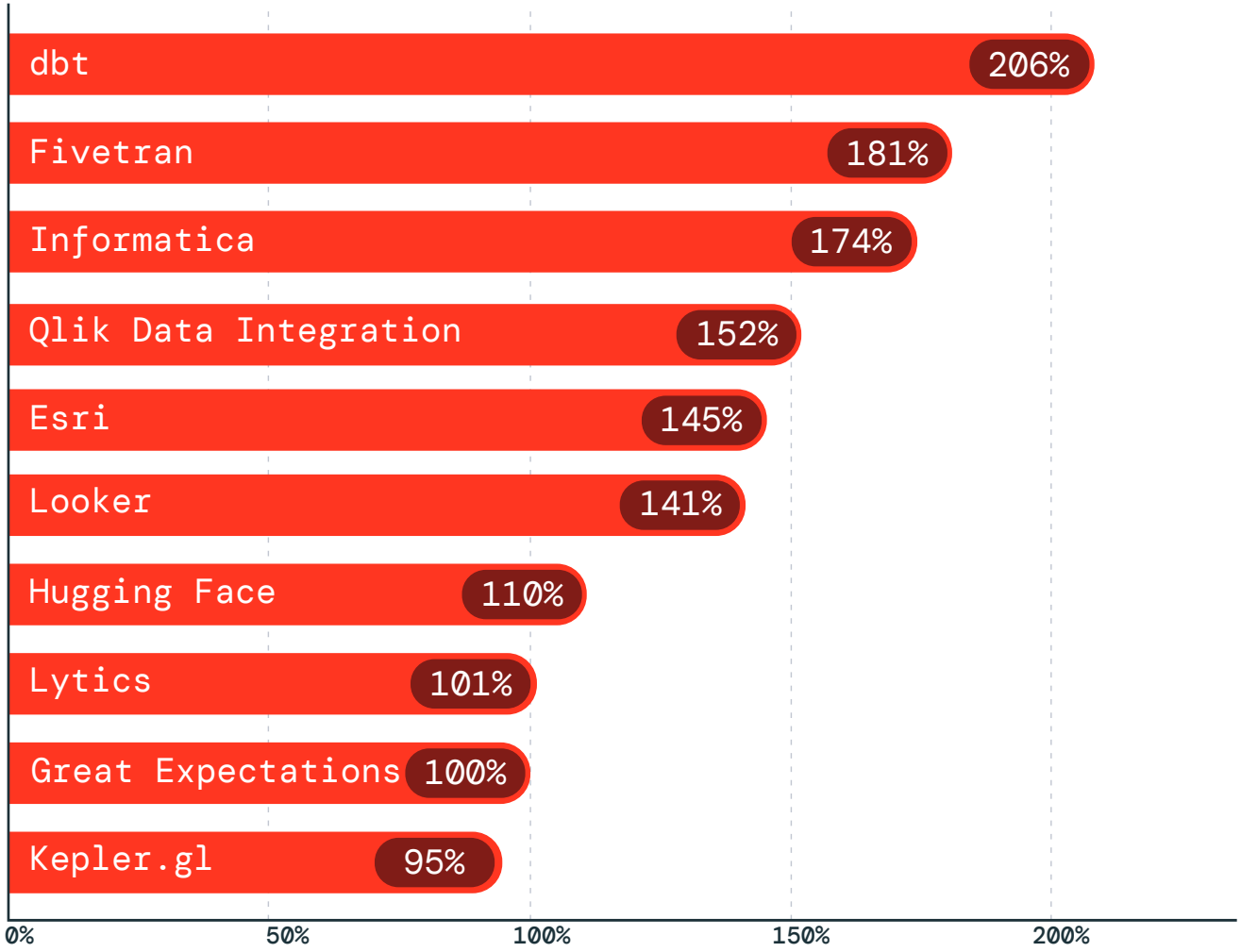
データ・AI 製品のトップ 10 に入る製品には、今日のデータ・AI スタックの多面的な性質が反映されています。企業は、経済不安にもかかわらず軒並みスタックへの投資を続けています。Databricks のユーザーが使用しているデータ・AI 製品の上位リストには、BI ツールと DS／ML が各 2 製品ずつ、データガバナンス・セキュリティに関する製品が 3 つ、データ統合の製品が 1 つが入っています。また、地理空間データという小規模な市場を扱う製品も 2 つが上位に入っています。

## オープンなプラットフォームと製品は今日のデータ戦略に不可欠である

### 最も広く利用されている製品は Microsoft Power BI、一方で、データサイエンス・AI 製品は急成長している

2023 年現在、最も広く使用されているデータ・AI 製品は Microsoft Power BI です。データ・AI 市場の上位リストの製品と同様に、BI ツールは多くの企業に業務ツールとして普及しています。しかし、企業がデータサイエンス・AI 製品の採用を優先していることもわかります。上位リスト 2 位の Plotly は、Python ベースのローコードプラットフォームで、データサイエンスチームがインタラクティブなチャートやマップを容易に作成・拡張できるようにする製品です。また、7 位のオープンソースの Hugging Face は、さまざまなタスク向けに事前にトレーニングされたオープンモデルへの容易なアクセスを可能にする Transformers ライブラリによって、AI の民主化を支援する製品です。これら 2 つのツールは、ML の導入を迅速化し、組織全体で使いやすくすることを目的としています。

# 急成長しているデータ・AI 製品

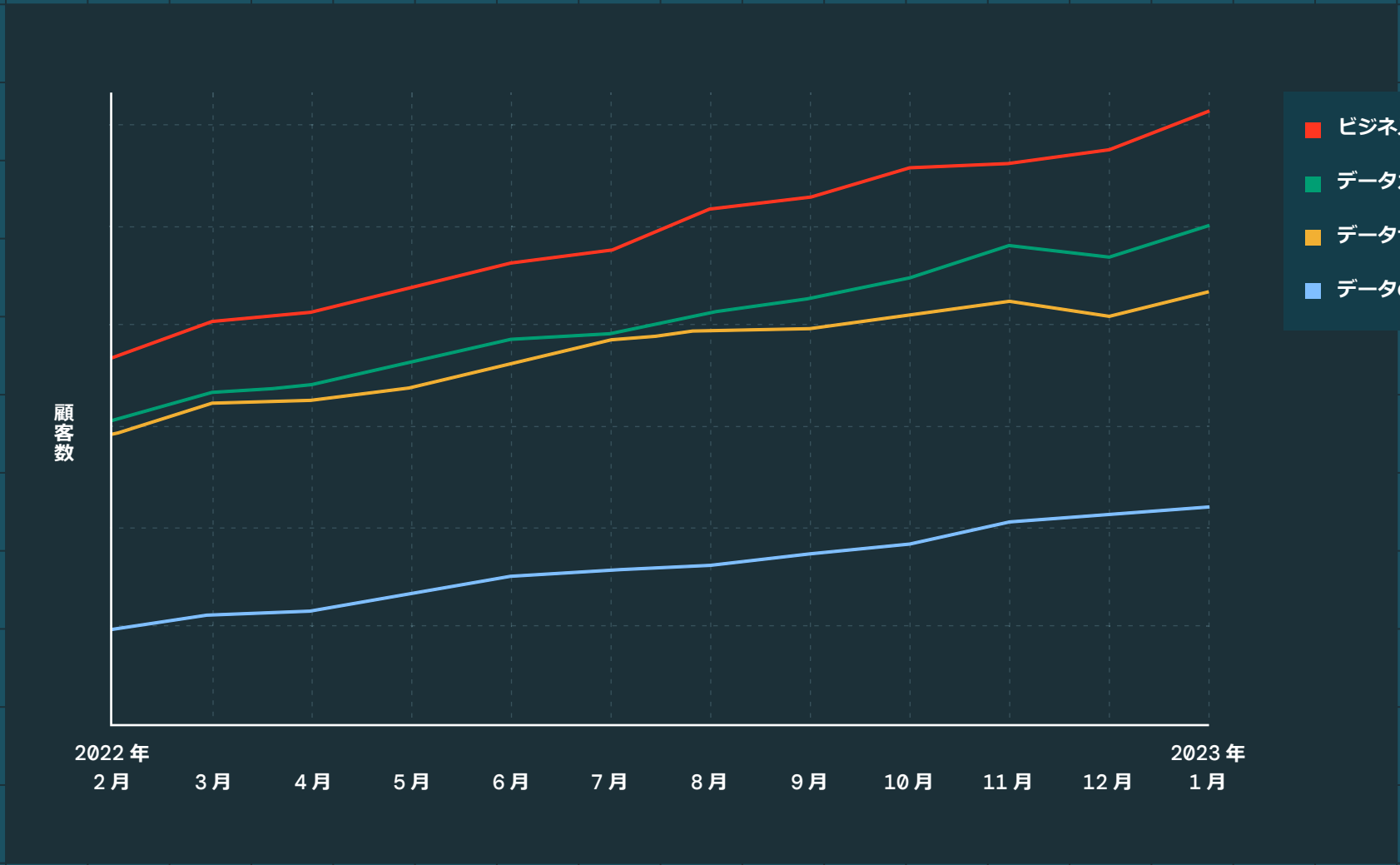



顧客数別の前年比成長率


## dbt は、2023 年に最も急成長しているデータ・AI 製品である

企業は、データを活用したより高度なユースケースの開発を急いでおり、レポーティング、ML モデリング、業務ワークフローに必要な信頼性の高いデータセットを生成する新たな製品への投資が進んでいます。これを受けて、データ統合製品の採用が急増しています。データ・AI 製品のなかでは、データ変換ツールの dbt、データパイプラインを自動化する Fivetran が最も急成長しています。このことは、企業における DS／ML への取り組みの優先度が高まり、競合ツールが躍進するデータ統合市場の新たな時代が進行していることを示唆しています。上位リストの 9 位に位置する Superconductive 社の Great Expectations も含めると、急成長している製品の 50% がデータ統合カテゴリの製品であることになります。

# データ・AI 製品市場の成長状況




注：このグラフは、4 つのカテゴリのそれぞれにおいて、1 つ以上のデータ・AI 製品を導入している顧客の数を示しています。
これら 4 つのカテゴリは、データ・AI 関連の全製品を網羅していたものではありません。Unity Catalog などの Databricks 製品は、
このデータには含まれていません。

# データ・AI 市場：BI が普及し、企業は機械学習の基盤に投資している

企業がデータの取り組みにどのような優先順位をつけているかを知るため、私たちは Databricks のレイクハウスで利用されている全てのデータ・AI 製品を、BI、データガバナンスとセキュリティ、DS／ML、データ統合という 4 つのコア市場に分類しました。調査対象のデータからは、初期段階のカテゴリと比べて、BI ツールがより広く採用され、前年比 66% 増で成長を続けていることがわかります。これは、次のセクション「レイクハウス視点からの展望」で詳述する、データウェア化にレイクハウスを利用する組織が増えているというトレンドに一致するものです。

BI をデータ戦略の始点とする企業は少なくありませんが、現在は、データや AI を活用したより高度なユースケースに着目する企業が増えつつあります。

**データの統合に関する製品の導入が急成長しており、前年比で 117% の成長率である**

### データ統合に関する製品の需要が急増している

現在最も急速に成長しているのはデータ統合市場です。データ統合ツールは、膨大な量の上流・下流データを統合し、単一のビューに集約して視覚化するものです。データ統合製品を導入することで、BI と DS／ML のあらゆる取り組みを堅牢な基盤の上に構築できます。

市場規模が小さいほど急速な成長が見られる傾向にある一方で、データ統合製品の市場は、前年比 117% 増と、BI 関連の製品の市場よりもはるかに急速に成長しています。この傾向は、本レポートのセクション「データサイエンスと機械学習」で述べたレイクハウスで見られる ML 導入の急成長と一致しています。

# レイクハウス視点からの展望

## データの移行と形式のトレンド

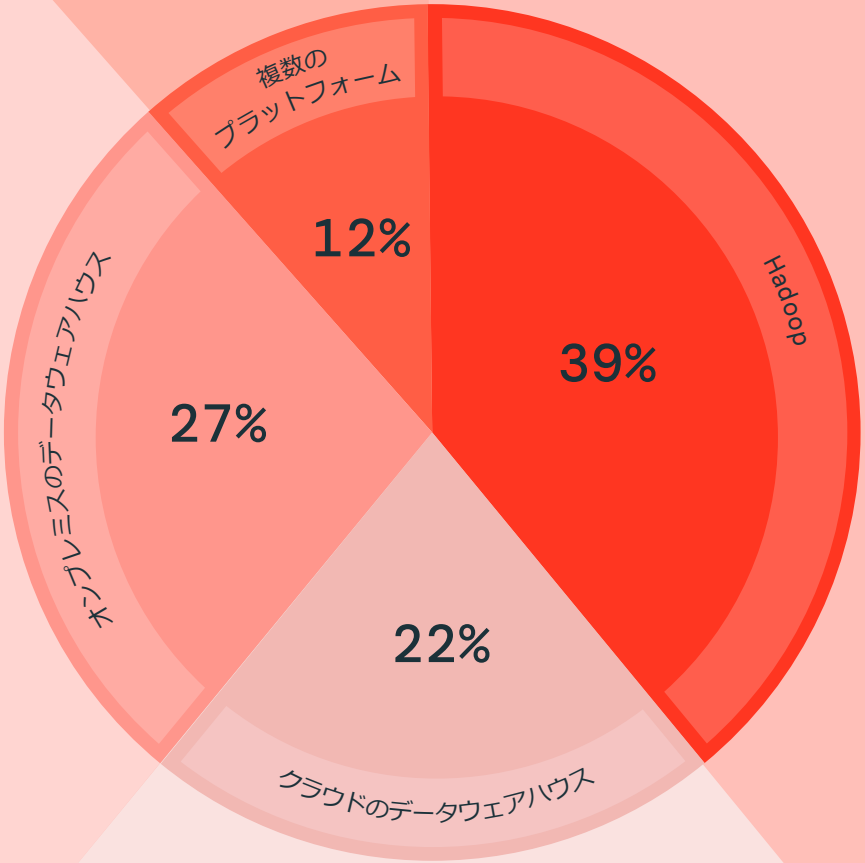
データの移行は大きな課題です。高リスク・高コストの移行プロセスが、企業における計画遅延の要因となり得ます。安易に開始することはできません。しかし、企業がレガシーデータプラットフォームの限界やスケーラビリティの課題、コスト面の高負荷に直面する現在、新たなタイプのアーキテクチャに移行する傾向が高まっています。

# 移行のトレンド：データウェアハウスの最適解は「レイクハウス」

レイクハウスプラットフォームは、高度なユースケースや DS／ML をサポートしており、従来のデータウェアハウスに代わってデータ戦略を包括的に強化するための有力な選択肢となっています。最も利用されているデータと AI 製品が BI とデータ統合ツールであることからもわかるように、データレイクハウスを利用してデータのウェアハウス化を行う組織が増えています。移行元のレガシープラットフォームを知るため、私たちは、Databricks の新規のお客さまの移行状況に着目しました。

特筆すべきは、レイクハウスに移行した企業の約半数がデータウェアハウスからの移行であったことです。これには、クラウドのデータウェアハウスからの移行 22% が含まれています。また、データウェアハウスのワークロードをレイクハウス上で実行し、データプラットフォームを統合することによって、コストの削減を図るという傾向が高まっていることもわかります。

## Databricks に移行した新規顧客の移行元プラットフォームの割合

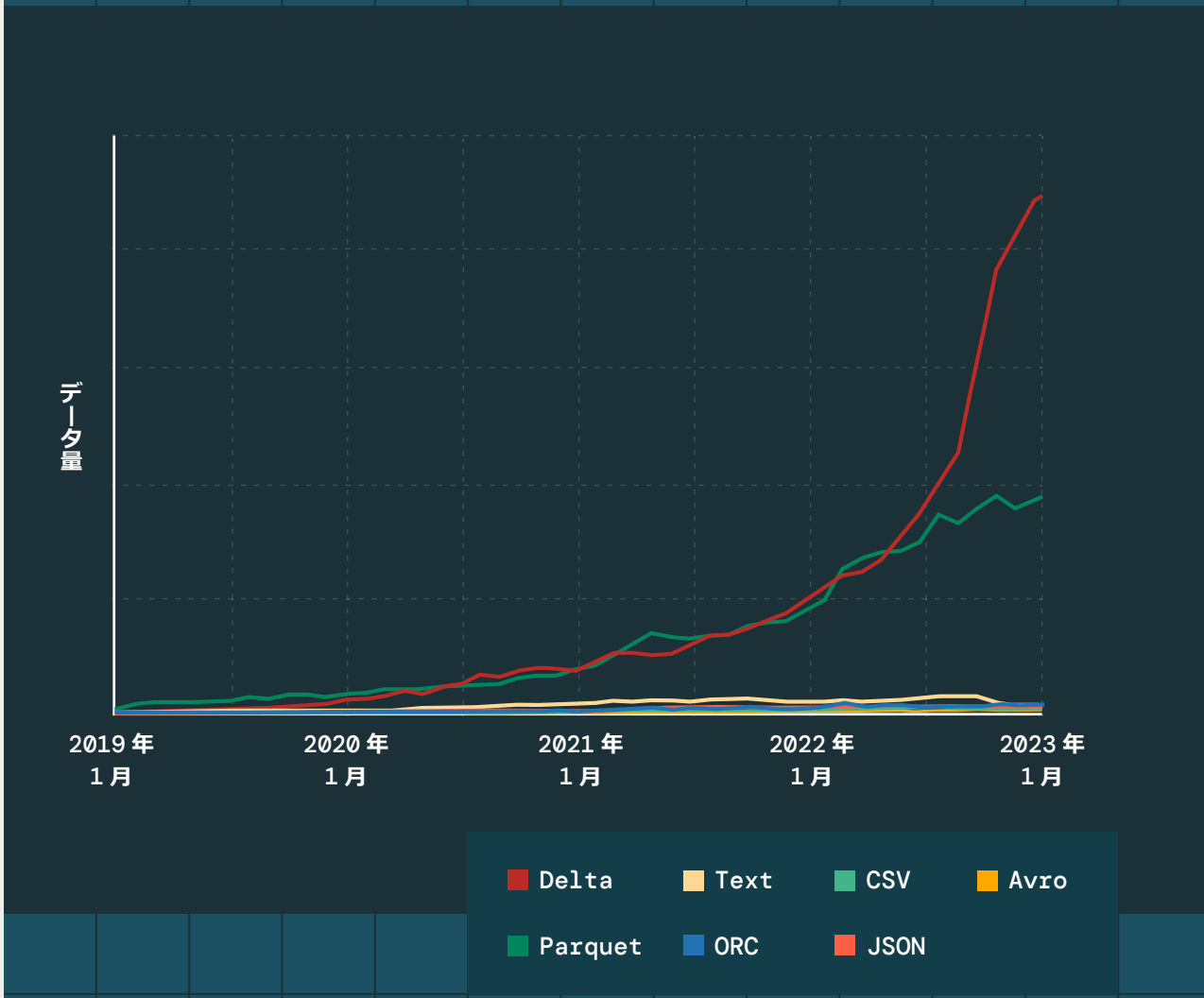


# データ形式のトレンド：Delta Lake のデータ量が前年比で 304% 増大している

データ量の爆発的な増大に伴い、半構造化・非構造化データの占める割合が高まっています。これまでは、構造化・非構造化・半構造化データをそれぞれ別のプラットフォームで管理しなければならず、複雑な作業と高コストという障壁が発生していました。レイクハウスは、あらゆるデータの種類とフォーマットに対応する統合プラットフォームであるため、この問題は、レイクハウスの導入によって解決できます。

Delta Lake は、Databricks のレイクハウスの基盤です。Delta Lake フォーマットは、構造化・非構造化・半構造化データに対応しています。利用はこの 2 年間で急増しています。他のストレージフォーマット（テキスト、JSON、CSV など）が緩やかな増加、横ばい、または減少傾向を見せているのに対して、Delta Lake フォーマットでデータを管理する組織が増加していることが、今回対象としたデータによって示されています。2022 年 6 月の時点で、Delta Lake フォーマットは Parquet を抑えて最も利用されているデータレイクのソースとなり、前年比 304% の成長を示しています。

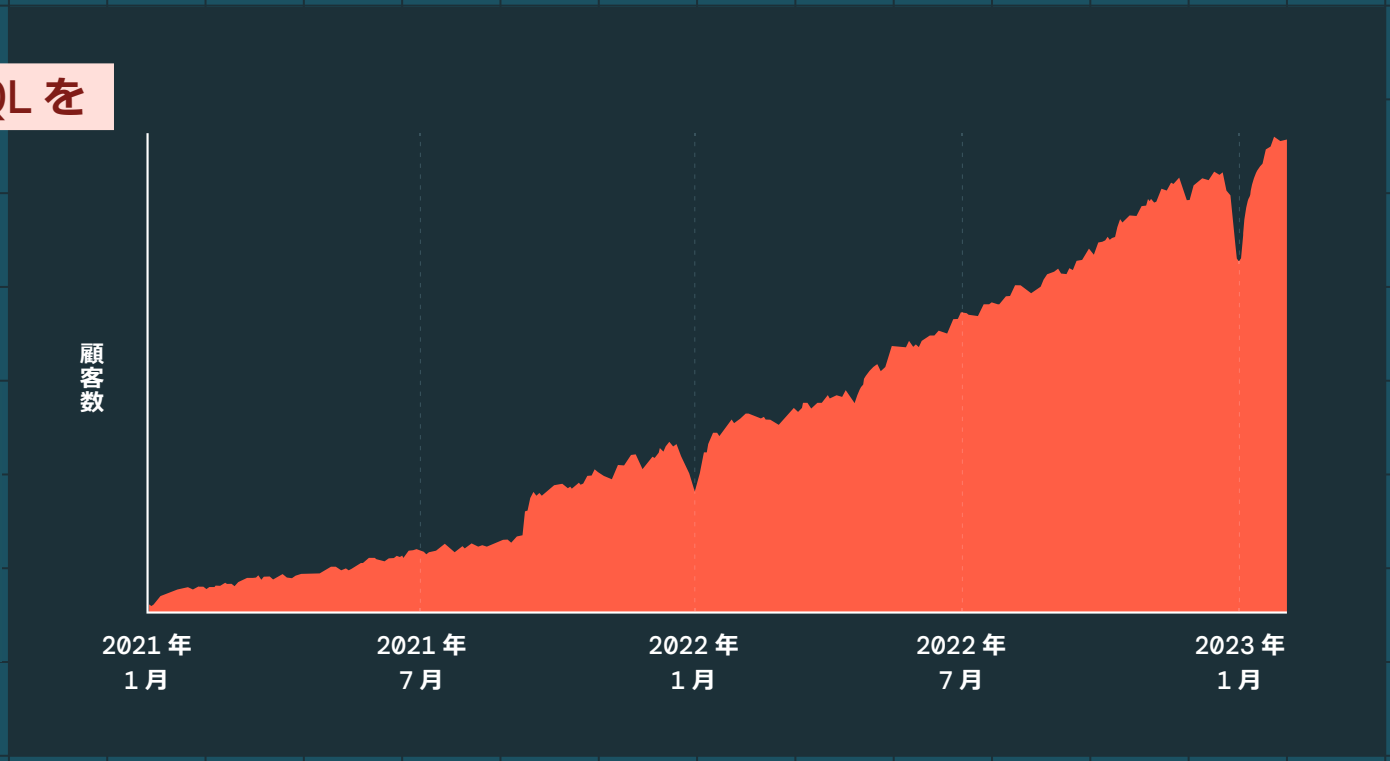
## 管理データ量の推移（フォーマット別）



# データのウェアハウス化が拡大、サーバーレス型がカギ

過去 2 年間で、レイクハウスプラットフォームでデータのウェアハウス化を行うケースが大幅に増加しています。この傾向を裏付ける事実として、レイクハウスのサーバーレスデータウェアハウスである Databricks SQL の利用が前年比で 144% 伸びているという状況があります。この状況はさらに、多くの企業が従来のデータウェアハウスの利用を廃止し、あらゆる BI と分析に対応するレイクハウスに移行していることを示しています。


## データウェアハウスにレイクハウスと Databricks SQL を利用している顧客数の推移



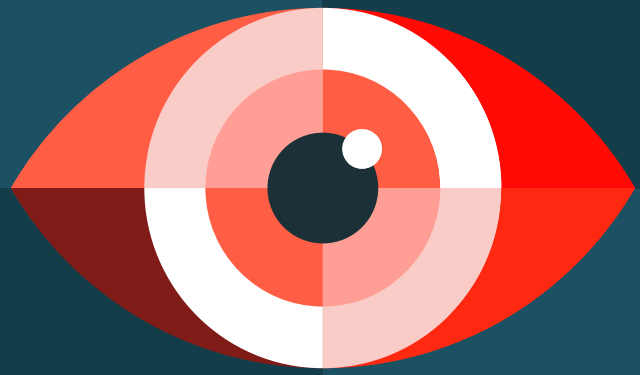

注：

- 2021 年 10 月の顧客数急増は、Databricks SQL の登録不要プレビューの提供開始に起因するものです。同製品は、同年 12 月に一般提供を開始しました。
- 12 月最終週の顧客数が減少しているのは、季節的要因によるものです。

結論

# 生成 AI

最も広く利用されているデータ・AI 製品は Microsoft Power BI である一方で、企業の関心は、より高度な ML と AI のユースケースに向かっており、それに伴ってデータ・AI スタックのモダナイズも進行しているというのは、注目すべきポイントです。データ統合ツール（Databricks での利用が急増している dbt を含む）を導入する企業が急増しているとともに、NLP（自然言語処理）と LLM（大規模言語モデル）の利用が拡大していることも今回のデータが示しています。これらのテクノロジーが今後数年間で爆発的に普及することは間違いないでしょう。DS／ML を最大限に活用する企業がデータの次世代をリードするであろうことが、かつてないほど明確になっています。



# データブリックスについて

データブリックスはデータと AI の企業です。コムキャスト、コンデナストをはじめ、フォーチュン500企業の過半数を含む世界中の9,000を超える企業が、Databricks のレイクハウスプラットフォームを利用して、データ、分析、AI の統合を実現しています。データブリックスは、米国カリフォルニア州サンフランシスコに本社を置き、世界中に事業所を配しています。Apache Spark™、Delta Lake、MLflow のクリエーターによって創立され、企業のデータチームが抱える、世界の最も困難な課題を解決するための支援を提供しています。Twitter、LinkedIn、Facebook での情報発信も行っております。ぜひご覧ください。

**Databricks のレイクハウスプラットフォーム**

databricks