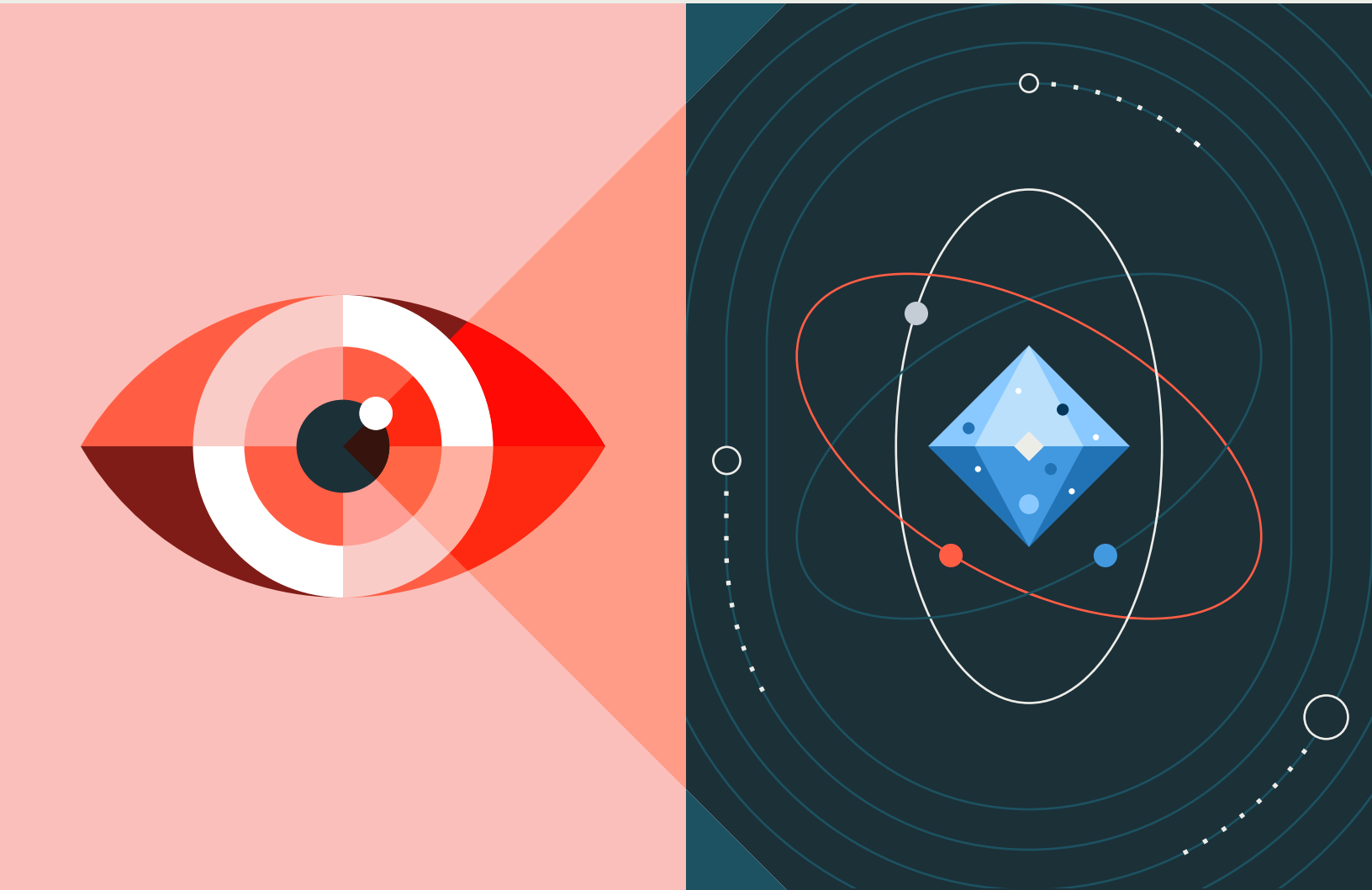
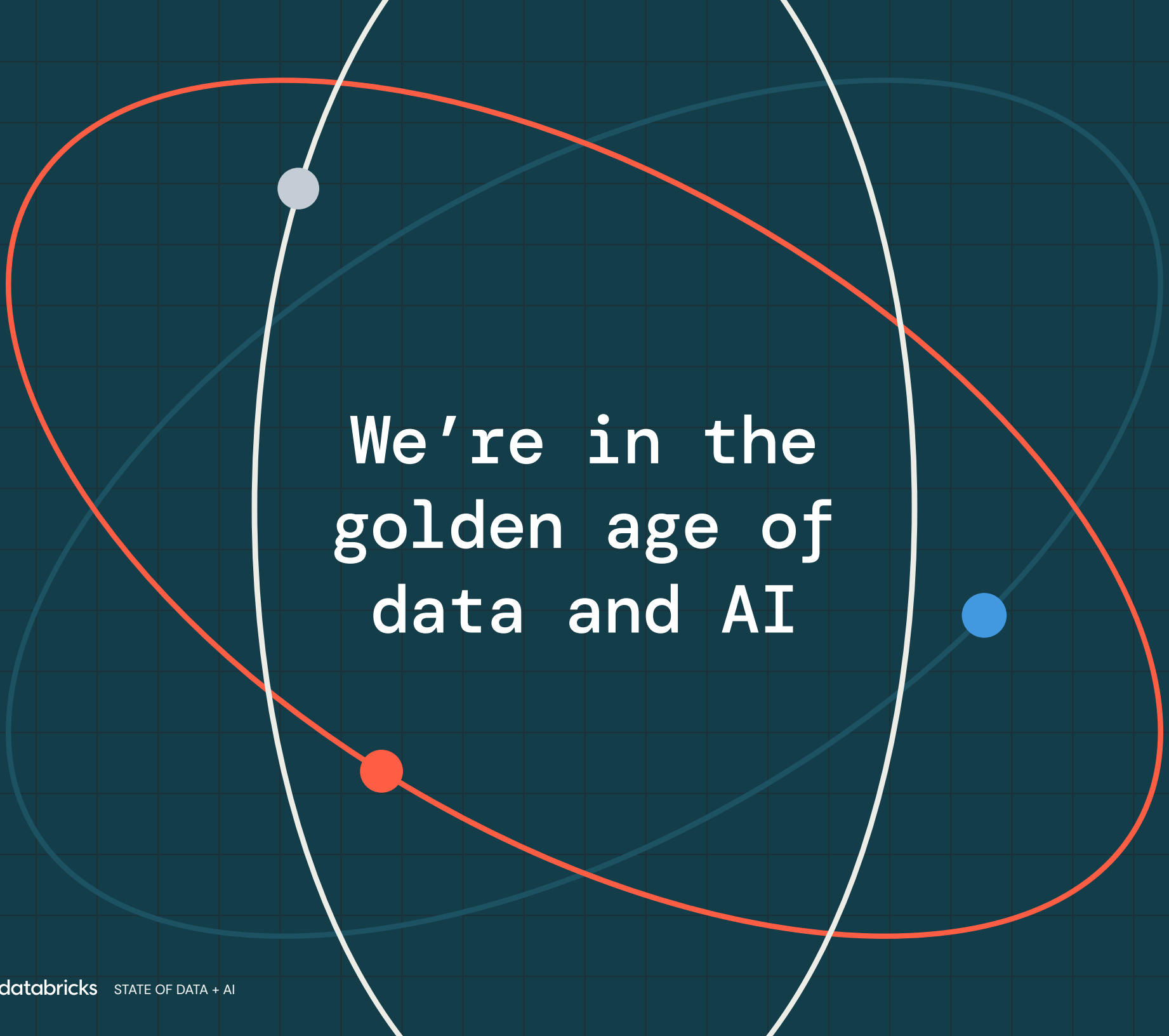
databricks

# State of Data + AI




We're in the golden age of data and AI

# INTRO

In the 6 months since ChatGPT launched, the world has woken up to the vast potential of AI. The unparalleled pace of AI discoveries, model improvements and new products on the market puts data and AI strategy at the top of conversations across every organization around the world. We believe that AI will usher in the next generation of product and software innovation, and we're already seeing this play out in the market. The next generation of winning companies and executives will be those who understand and leverage AI.

In this report, we examine patterns and trends in data and AI adoption across more than 9,000 global Databricks customers. By unifying business intelligence (BI) and AI applications across companies' entire data estates, the Databricks Lakehouse provides a unique vantage point into the state of data and AI, including which products and technologies are the fastest growing, the types of data science and machine learning (DS/ML) applications being developed and more.

## Here are the major stories we uncovered:

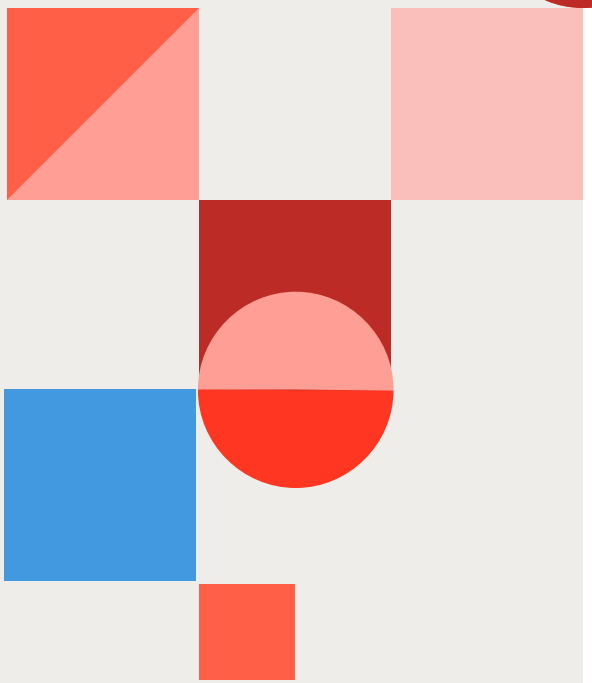
Companies are adopting machine learning and large language models (LLMs) at a rapid pace. Natural language processing (NLP) is dominating use cases, with an accelerated focus on LLMs.

Open source wins in today's data and AI markets. Eight out of 10 of our most widely adopted AI and machine learning products are based on open source.

Organizations are increasingly using the Lakehouse for data warehousing, as evidenced by the high growth of data integration tools dbt and Fivetran, and the accelerated adoption of Databricks SQL.

We hope that by sharing these trends, data leaders will be able to benchmark their organizations and gain insights that help inform their strategies for an era defined by data and AI.

# Summary of Key Findings

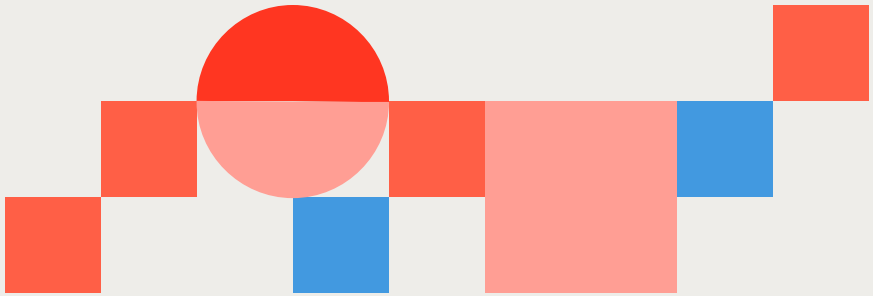


## 1 DATA SCIENCE AND MACHINE LEARNING: NLP AND LLMS ARE IN HIGH DEMAND

- The number of companies using SaaS LLM APIs (used to access services like ChatGPT) grew 1310% between the end of November 2022 and the beginning of May 2023
- NLP accounts for 49% of daily Python data science library usage, making it the most popular application
- Organizations are putting substantially more models into production (411% YoY growth) while also increasing their ML experimentation (54% YoY growth)
- Organizations are getting more efficient with ML; for every three experimental models, roughly one is put into production, compared to five experimental models a year prior

## 2 FASTEST-GROWING DATA AND AI PRODUCTS

- BI is the top data and AI market, but growth trends in other markets show that companies are increasingly looking at more advanced data use cases
- The fastest-growing data and AI product is dbt, which grew 206% YoY by number of customers
- Data integration is the fastest-growing data and AI market on the Databricks Lakehouse with 117% YoY growth



## 3 ADOPTION AND MIGRATION TRENDS

- 61% of customers migrating to the Lakehouse are coming from on-prem and cloud data warehouses
- The volume of data in Delta Lake has grown 304% YoY
- The Lakehouse is increasingly being used for data warehousing, including serverless data warehousing with Databricks SQL, which grew 144% YoY

# Methodology: How did Databricks create this report?

The *State of Data + AI* is built from fully aggregated, anonymized data collected from our customers based on how they are using the Databricks Lakehouse and its broad ecosystem of integrated tools. This report focuses on machine learning adoption, data architecture (integrations and migrations) and use cases. The customers in this report represent every major industry and range in size from startups to many of the world's largest enterprises.

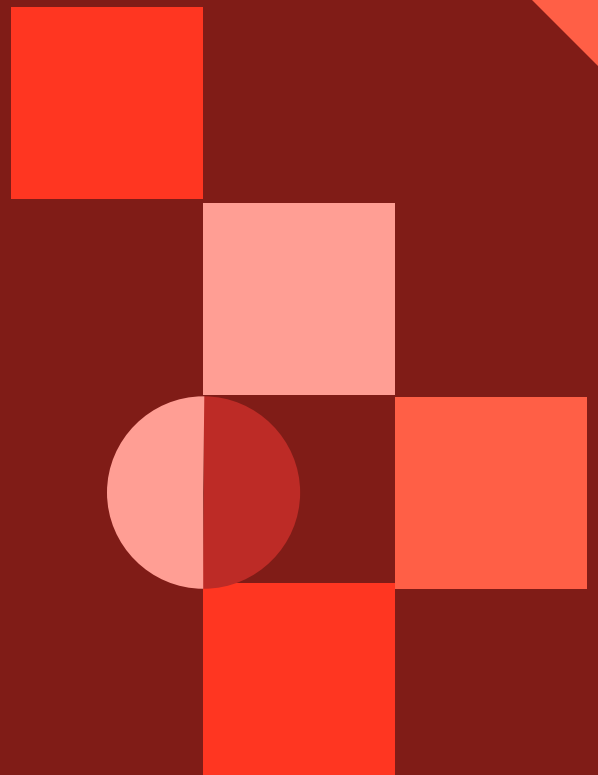
Unless otherwise noted, this report presents and analyzes data from February 1, 2022, to January 31, 2023, and usage is measured by number of customers. When possible, we provide YoY comparisons to showcase growth trends over time.

# Data Science and Machine Learning

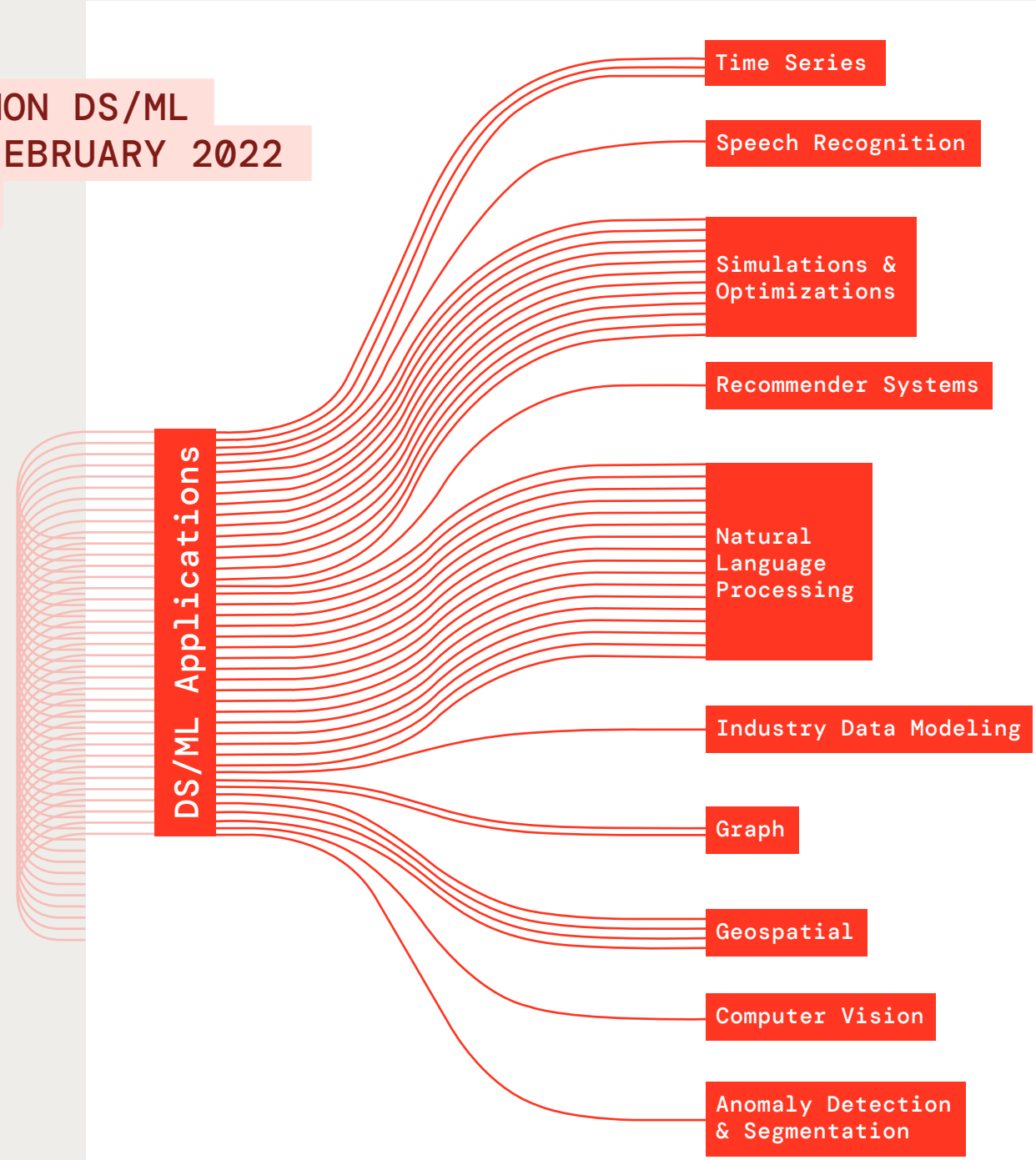
## NATURAL LANGUAGE PROCESSING AND LARGE LANGUAGE MODELS ARE IN HIGH DEMAND

Across all industries, companies leverage data science and machine learning (DS/ML) to accelerate growth, improve predictability and enhance customer experiences. Recent advancements in large language models (LLMs) are propelling companies to rethink AI within their own data strategies. Given the rapidly evolving DS/ML landscape, we wanted to understand several aspects of the market:

- Which types of DS/ML applications are companies investing in? In particular, given the recent buzz, what does the data around LLMs look like?
- Are companies making headway on operationalizing their machine learning models (MLOps)?



# SPECIALIZED PYTHON DS/ML LIBRARIES FROM FEBRUARY 2022 TO JANUARY 2023




Note: This chart reflects the unique number of notebooks using ML libraries per day in each of the categories. It includes libraries used for the particular problem-solving use cases mentioned. It does not include libraries used in tooling for data preparations and modeling.

# Natural language processing dominates machine learning use cases

To understand how organizations are applying AI and ML within the Lakehouse, we aggregated the usage of specialized Python libraries, which include NLTK, Transformers and FuzzyWuzzy, into popular data science use cases.[1] We look at data from these libraries because Python is on the cutting edge of new developments in ML, advanced analytics and AI, and has consistently ranked as one of the most popular programming languages in recent years.

Our most popular use case is natural language processing (NLP), a rapidly growing field that enables businesses to gain value from unstructured textual data. This opens the door for users to accomplish tasks that were previously too abstract for code, such as summarizing content or extracting sentiment from customer reviews. In our data set, 49% of libraries used are associated with NLP. LLMs also fall within this bucket. Given the innovations launched in recent months, we expect to see NLP take off even more in coming years as it is applied to use cases like chatbots, research assistance, fraud detection, content generation and more.

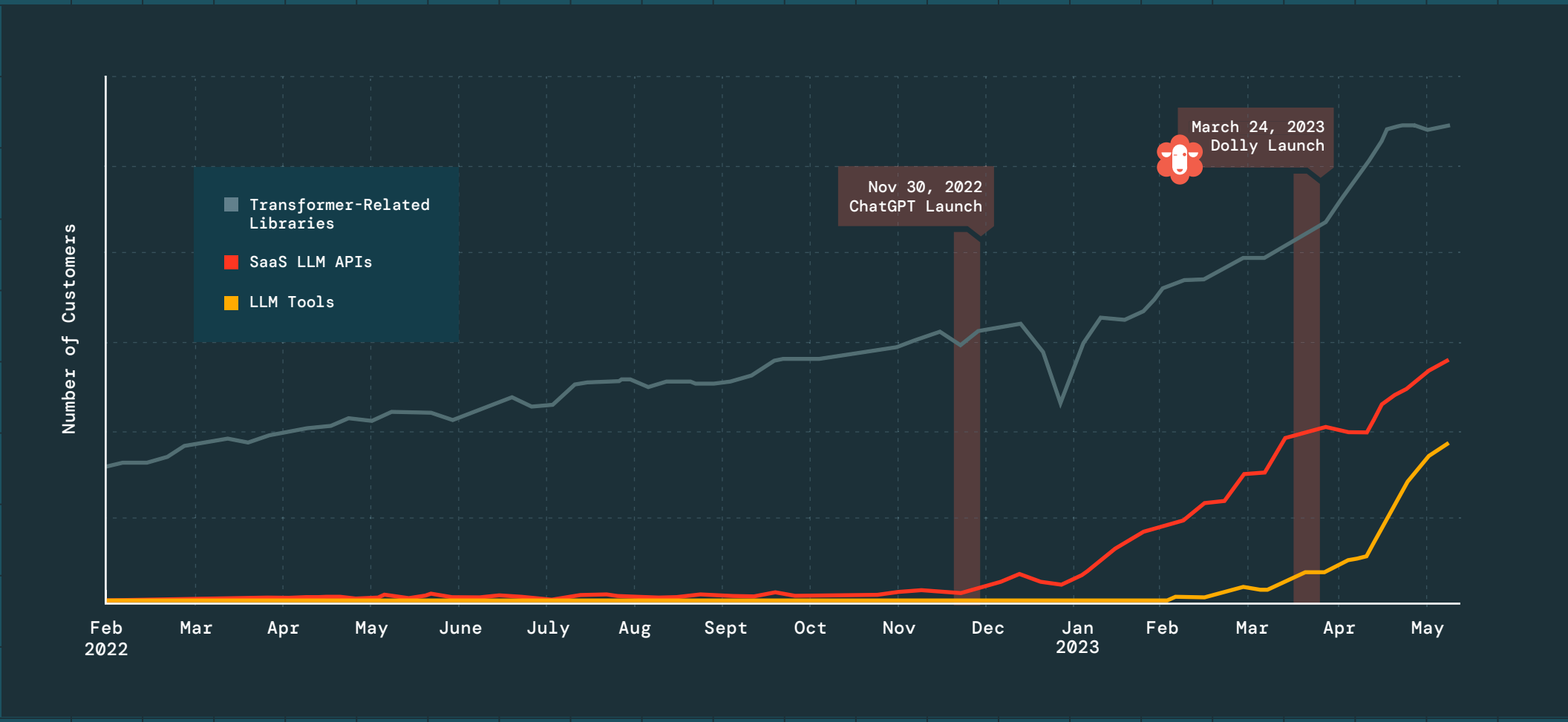
Our second most popular DS/ML application is simulations and optimization, which accounts for 30% of all use cases. This signals organizations are using data to model prototypes and solve problems cost-effectively.

**In our data set, 49% of specialized Python libraries used are associated with NLP**

Many of the DS/ML use cases are predominantly leveraged by specific industries. While they take up a smaller share of the total, they are mission-critical for many organizations. For example, time series includes forecasting, a use case that is especially popular in industries such as Retail and CPG, which rely heavily on the ability to forecast the need for every item in every store.

1. This data does not include general-purpose ML libraries, including scikit-learn or TensorFlow.

# USE OF LARGE LANGUAGE MODELS (LLMS)




Note: There are several popular types of Python libraries that are commonly used for LLMs. These libraries provide pretrained models and tools for building, training and deploying LLMs. We have rolled these libraries up into groupings based on the type of functionality they provide.

Data consistently dips in the last week of December due to seasonality.

# Large language models are the "it" tool

LLMs are currently one of the hottest and most-watched areas in the field of NLP. LLMs have been instrumental in enabling machines to understand, interpret and generate human language in a way that was previously impossible, powering everything from machine translation to content creation to virtual assistants and chatbots.

Transformer-related libraries had been growing in popularity even before ChatGPT thrust LLMs into the public consciousness. Within the last 6 months, our data shows two accelerating trends: organizations are building their own LLMs, which models like Dolly show can be quite accessible and inexpensive. And, they are using proprietary models like ChatGPT. Transformer-related libraries, such as Hugging Face, which are used to train LLMs, have the highest adoption within the Lakehouse.

The second most popular type is SaaS LLMs, which are used to access models like OpenAI. This category has grown exponentially in parallel with the launch of ChatGPT: the number of Lakehouse customers using SaaS LLMs grew an impressive 1310% between the end of November 2022 and the beginning of May 2023. (In contrast, transformer-related libraries grew 82% in this same period.)

Organizations can leverage LLMs either by using SaaS LLM APIs to call services like ChatGPT from OpenAI or by operating their own LLMs in-house.

Thinking of building your own modern LLM application? This approach could entail the use of specialized transformer-related Python libraries to train the model, as well as LLM tools like LangChain to develop prompt interfaces or integrations to other systems.

## LLM DEFINITIONS

- **Transformer-related libraries:** Python libraries used to train LLMs (example: Hugging Face)
- **SaaS LLM APIs:** Libraries used to access LLMs as a service (example: OpenAI)
- **LLM tools:** Toolchains for working with and building proprietary LLMs (example: LangChain)

# Machine learning experimentation and production take off across industries

The increasing demand for ML solutions and the growing availability of technologies have led to a significant increase in experimentation and production, two distinct parts of the ML model lifecycle. We look at the *logging* and *registering* of models in MLflow, an open source platform developed by Databricks, to understand how ML is trending and being adopted within organizations.

MLflow Model Registry launched in May 2021. Overall, the number of logged models has grown 54% since February 2022, while the number of registered models has grown 411% over the same period. This growth in volume suggests organizations are understanding the value of investing in and allocating more people power to ML.

## LOGGED MODELS AND ML EXPERIMENTATION

During the experimentation phase of ML, data scientists develop models designed to solve given tasks. After training the models, they test them to evaluate their accuracy, precision, recall (the percentage of correctly predicted positive instances out of all actual positive instances), and more. These metrics are logged (recorded) in order to analyze the various models' performance and identify which approach works best for the given task.

We have chosen logged models as a proxy to measure ML experimentation because the MLflow Tracking Server is designed to facilitate experiment tracking and reproducibility.

## REGISTERED MODELS AND ML PRODUCTION

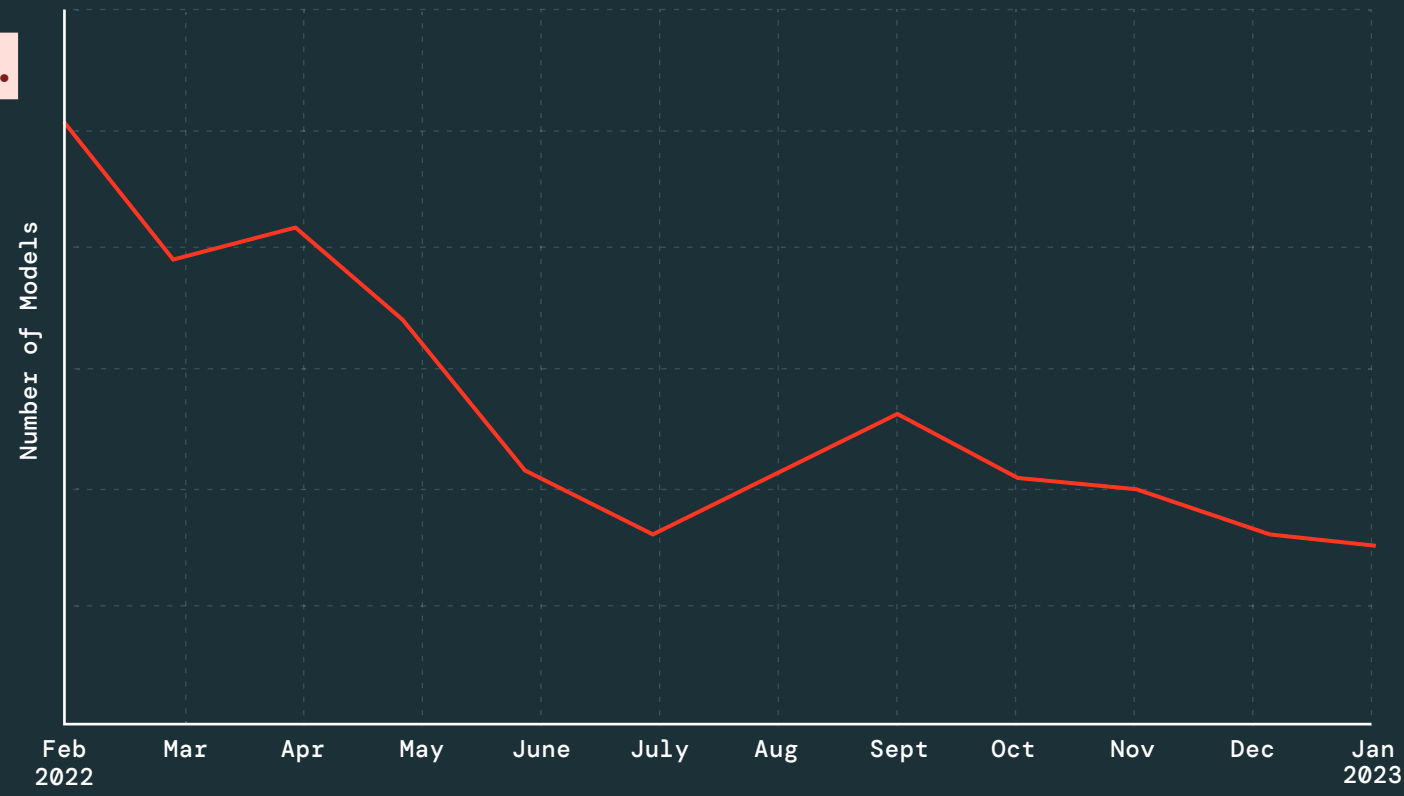
Production models have undergone the experimentation phase and are then deployed in real-world applications. They are typically used to make predictions or decisions based on new data. Registering a model is the process of recording and storing metadata about a trained model in a centralized location that allows users to easily access and reuse existing models. Registering models prior to production enables organizations to ensure consistency and reliability in model deployment and scale.

We have chosen registered models to represent ML production because the MLflow Model Registry is designed to manage models that have left the experimentation phase through the rest of their lifecycle.

Organizations test numerous approaches and variables before committing an ML model to production. We wanted to understand, "How many models do data scientists experiment with before moving to production?"

Our data shows the ratio of logged to registered models is 2.9 : 1 as of January 2023. This means that for roughly every three experimental models, one model will get registered as a candidate for production. This ratio has improved significantly from just a year prior, when we saw that for roughly every five experimental models, one was registered. Recent advances in ML, such as improved open source libraries like MLflow and Hugging Face, have radically simplified building and putting models into production. The result is that 34% of logged models are now candidates for production today, an improvement from over 20% just a year ago.

## RATIO OF LOGGED VS. REGISTERED MODELS

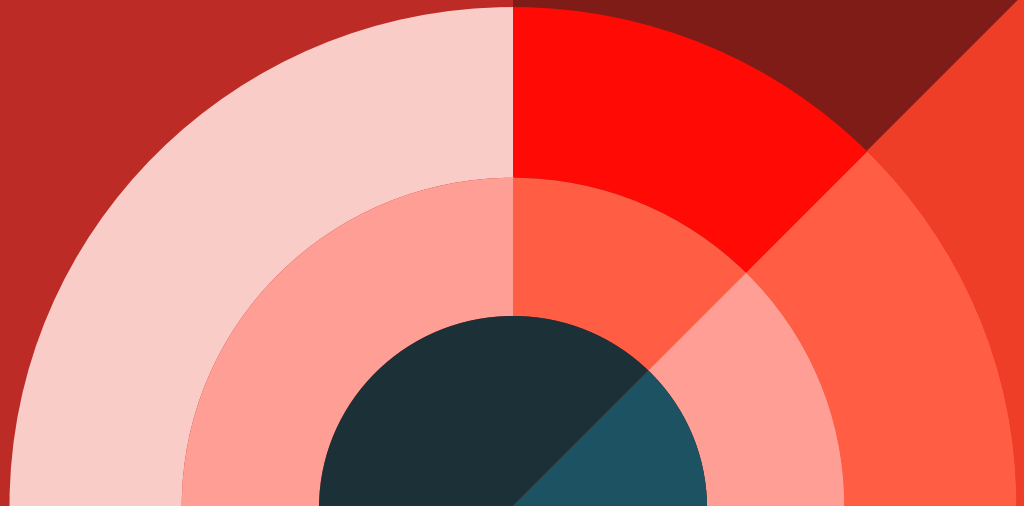



**2.9 : 1**

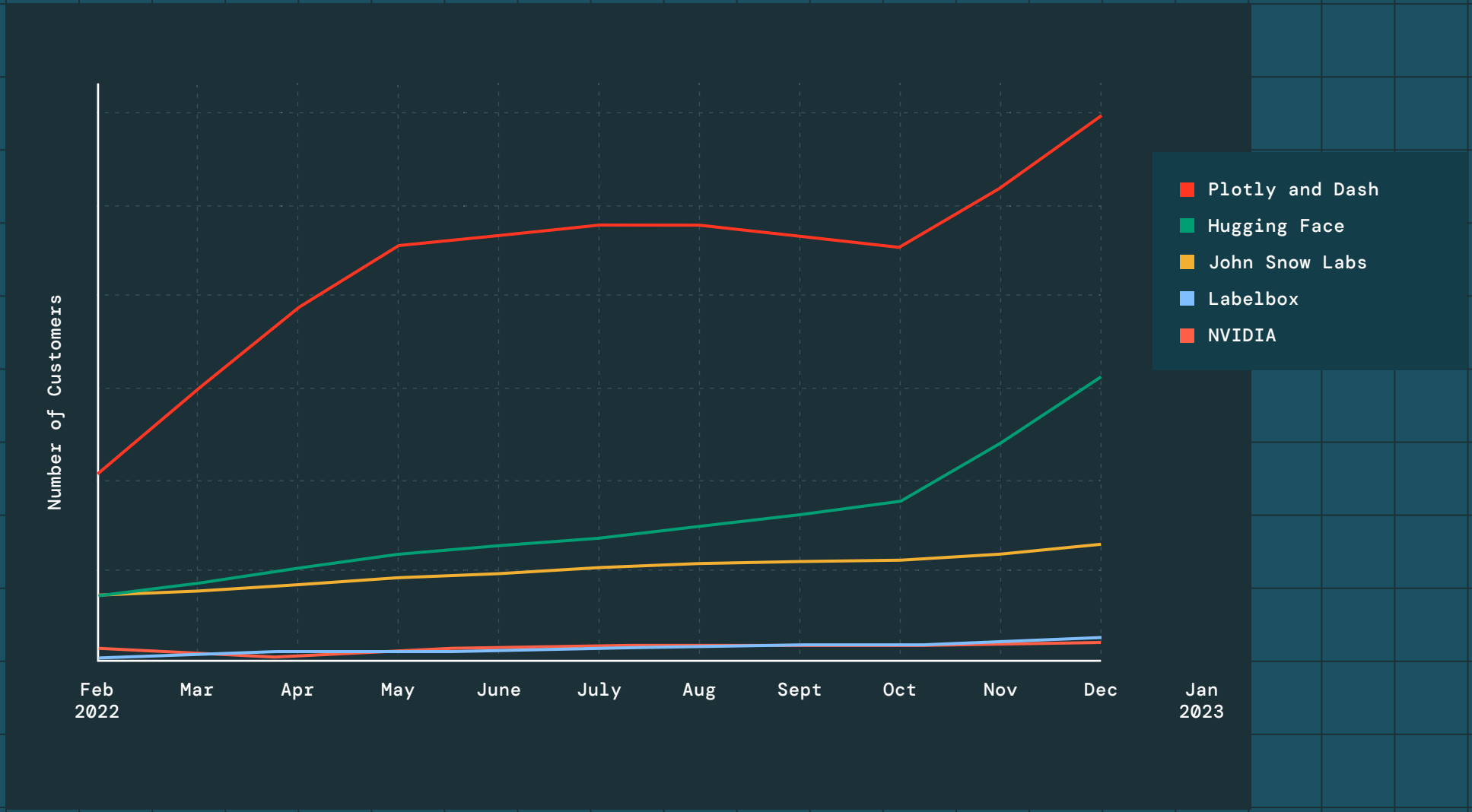
Ratio of Logged to Registered Models in Jan 2023

# The Modern Data and AI Stack

Over the last several years, the trend toward building open, unified data architectures has played out in our own data. We see that data leaders are opting to preserve choice, leverage the best products and deliver innovation across their organizations by democratizing access to data for more people.



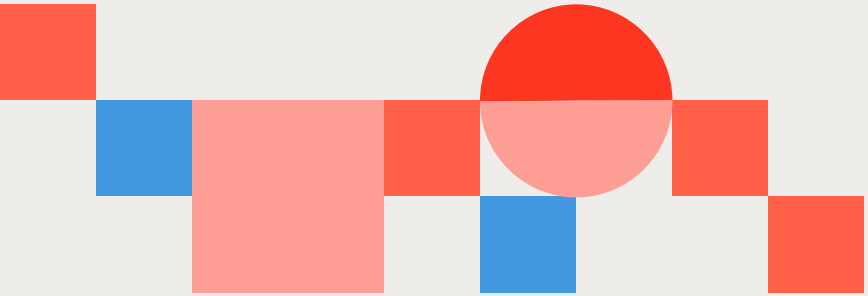
# TOP 5 AI AND MACHINE LEARNING PRODUCTS



# Top AI and ML Products

As companies integrate data science and ML into their business strategies, many leaders are looking for guidance on the right tools to add to their arsenals. The Databricks Lakehouse integrates with a growing number of AI and ML solutions to support these use cases.

One of our most interesting findings is that open source is dominating the top ranks; 3 out of our 5 most widely adopted AI and ML products on the Lakehouse are based on open source. This indicates a growing sentiment across industries: open platforms and products are critical to today's AI and ML strategies. We anticipate this trend to continue with the rise of generative AI. Many organizations want to leverage LLMs but share concerns over issues such as data privacy for their sensitive data. Open source and open models empower organizations to build LLMs without relying on third-party proprietary tools.



Rising Star

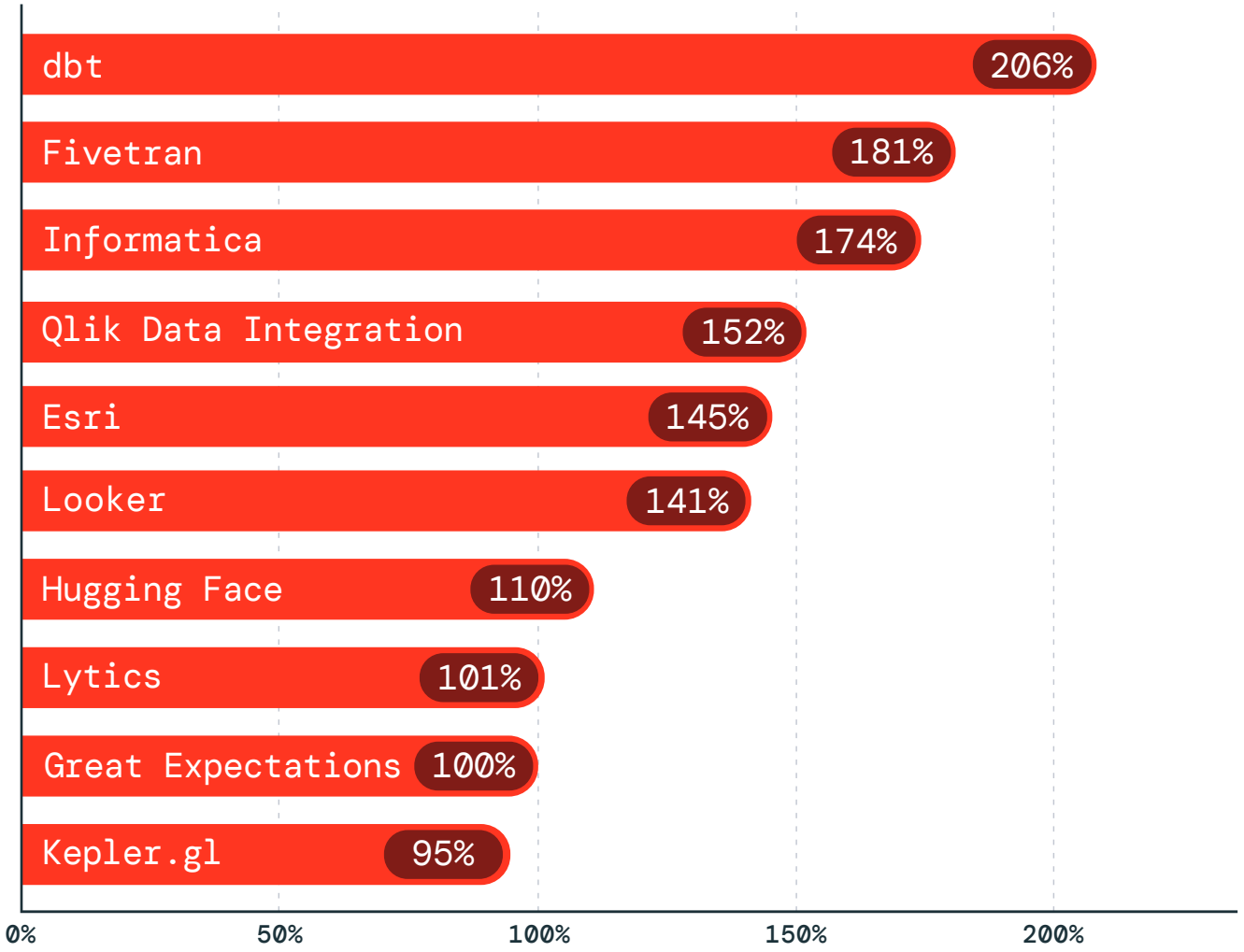
LangChain

Launched in October 2022, LangChain is an open source framework for developing LLM applications. As a new integration, LangChain does not qualify for this year's Top 5 AI and ML Products list. But its accelerated growth with the Databricks Lakehouse is worth highlighting, as it speaks volumes about the current state of the industry.

The LangChain and Databrick integration launched at the end of April 2023. In just three months, LangChain became the **third most popular AI and ML product** with our customers, following Plotly and Dash and Hugging Face.

This data point supports a very clear message: enterprises want to use generative AI within their businesses *now*.

# FASTEST-GROWING DATA AND AI PRODUCTS



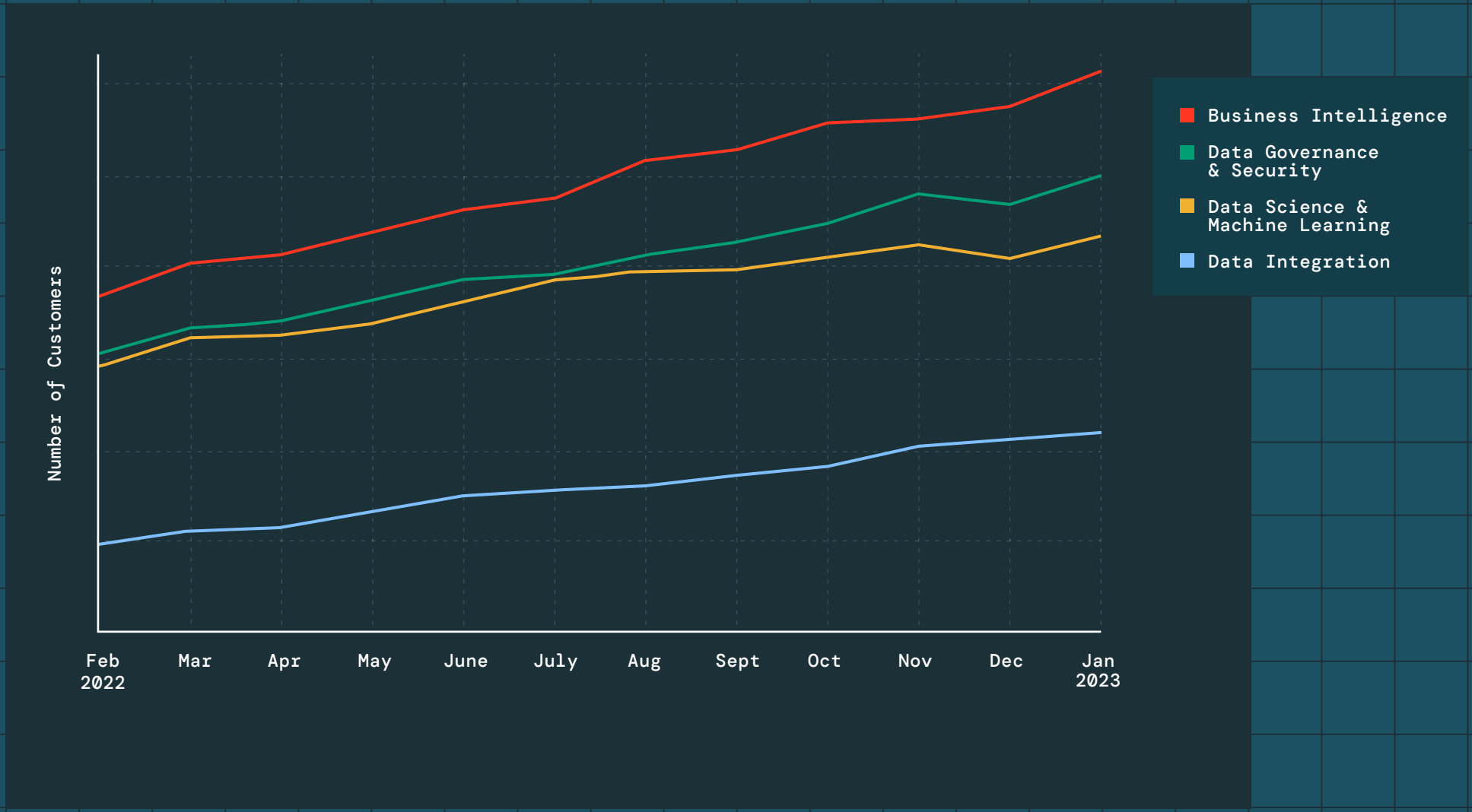

Year-Over-Year Growth by Number of Customers


## DBT IS THE FASTEST-GROWING DATA AND AI PRODUCT OF 2023

The data ecosystem is undergoing a major transition, and selecting the right products is critical for companies looking to take advantage of the newest innovations. Because the Databricks Lakehouse is used broadly across this ecosystem, it provides unique insights into how customers adopt hundreds of data products and services.

We discovered that as companies move quickly to develop more advanced use cases, they are investing in newer products that produce trusted data sets for reporting, ML modeling and operational workflows. Hence, we see the rapid rise of data integration products. dbt, a data transformation tool, and Fivetran, which automates data pipelines, are our two fastest-growing data and AI products. This suggests a new era of the data integration market, with challenger tools making headway as companies shift to prioritize DS/ML initiatives. With Great Expectations from Superconductive in the ninth spot, a full 50% of our fastest-growing products represent the data integration category.

# GROWTH OF DATA AND AI MARKETS




Note: In this chart, we count the number of customers deploying one or more data and AI products in each category. These four categories do not encompass all products. Databricks products, such as Unity Catalog, are not included in this data.

# Data and AI markets: business intelligence is standard, organizations invest in their machine learning foundation

To understand how organizations are prioritizing their data initiatives, we aggregated all data and AI products on the Databricks Lakehouse and categorized them into four core markets: BI, data governance and security, DS/ML, and data integration. Our data set confirms that BI tools are more widely adopted across organizations relative to more nascent categories — and they continue to grow, with a 66% YoY increase in adoption. This aligns with the broader trend of more organizations performing data warehousing on a Lakehouse, covered in the next section, Views from the Lakehouse.

**Data integration is the fastest-growing market, with 117% YoY growth**

While BI is often where organizations start their data journey, companies are increasingly looking at more advanced data and AI use cases.

## DEMAND FOR DATA INTEGRATION PRODUCTS IS GROWING FAST

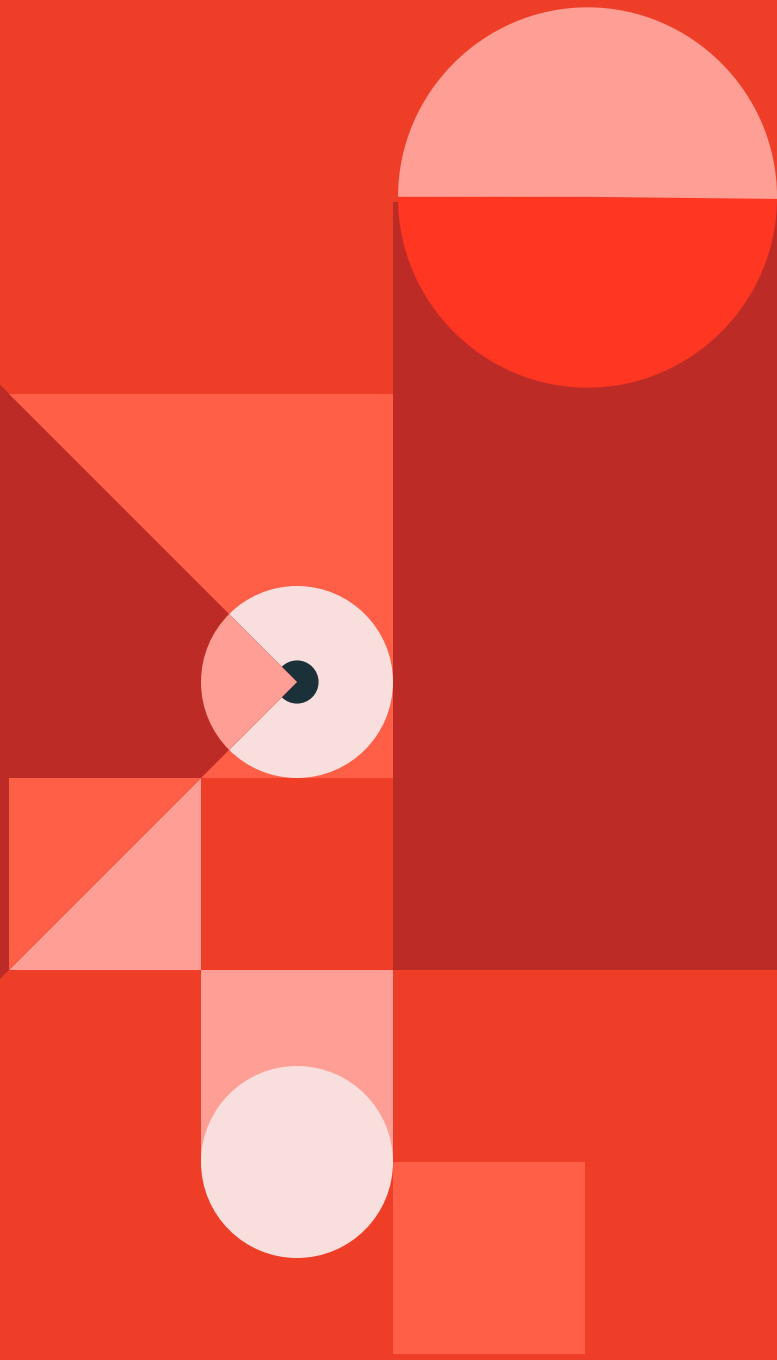
We see the fastest growth in the data integration market. These tools enable a company to integrate vast amounts of upstream and downstream data in one consolidated view. Data integration products ensure that all BI and DS/ML initiatives are built on a solid foundation.

While it's easier for smaller markets to experience faster growth, at 117% YoY increased adoption, the data integration market is growing substantially faster than BI. This trend dovetails with the rapid growth of ML adoption we see across the Lakehouse, covered in the DS/ML section of the report.

# Views from the Lakehouse

## MIGRATION AND DATA FORMAT TRENDS

Data migration is a major undertaking: it can be risky, expensive and delay companies' timelines. It's not a task to jump into lightly. As organizations run into the limitations, scalability challenges and the cost burden of legacy data platforms, they are increasingly likely to migrate to a new type of architecture.

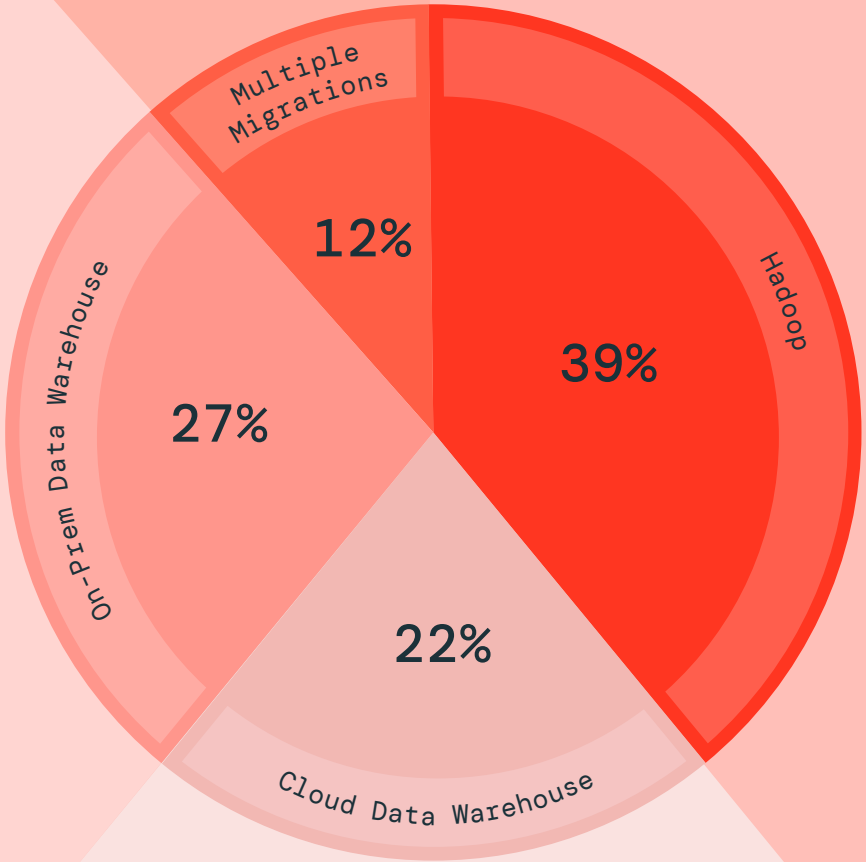


# Migration trends: the best data warehouse is a Lakehouse

The Lakehouse Platform is an attractive alternative to traditional data warehouses because it supports advanced use cases and DS/ML, allowing organizations to boost their overall data strategy. As evidenced by the most popular data and AI products, with BI and data integration tools at the top, organizations are increasingly using the data lakehouse for data warehousing. To better understand which legacy platforms organizations are moving away from, we look at the migrations of new customers to Databricks.

An interesting takeaway is that roughly half of the companies moving to the Lakehouse are coming from data warehouses. This includes the 22% that are moving from cloud data warehouses. It also demonstrates a growing focus on running data warehousing workloads on a Lakehouse and unifying data platforms to reduce cost.

## SOURCE OF NEW CUSTOMER MIGRATIONS TO DATABRICKS

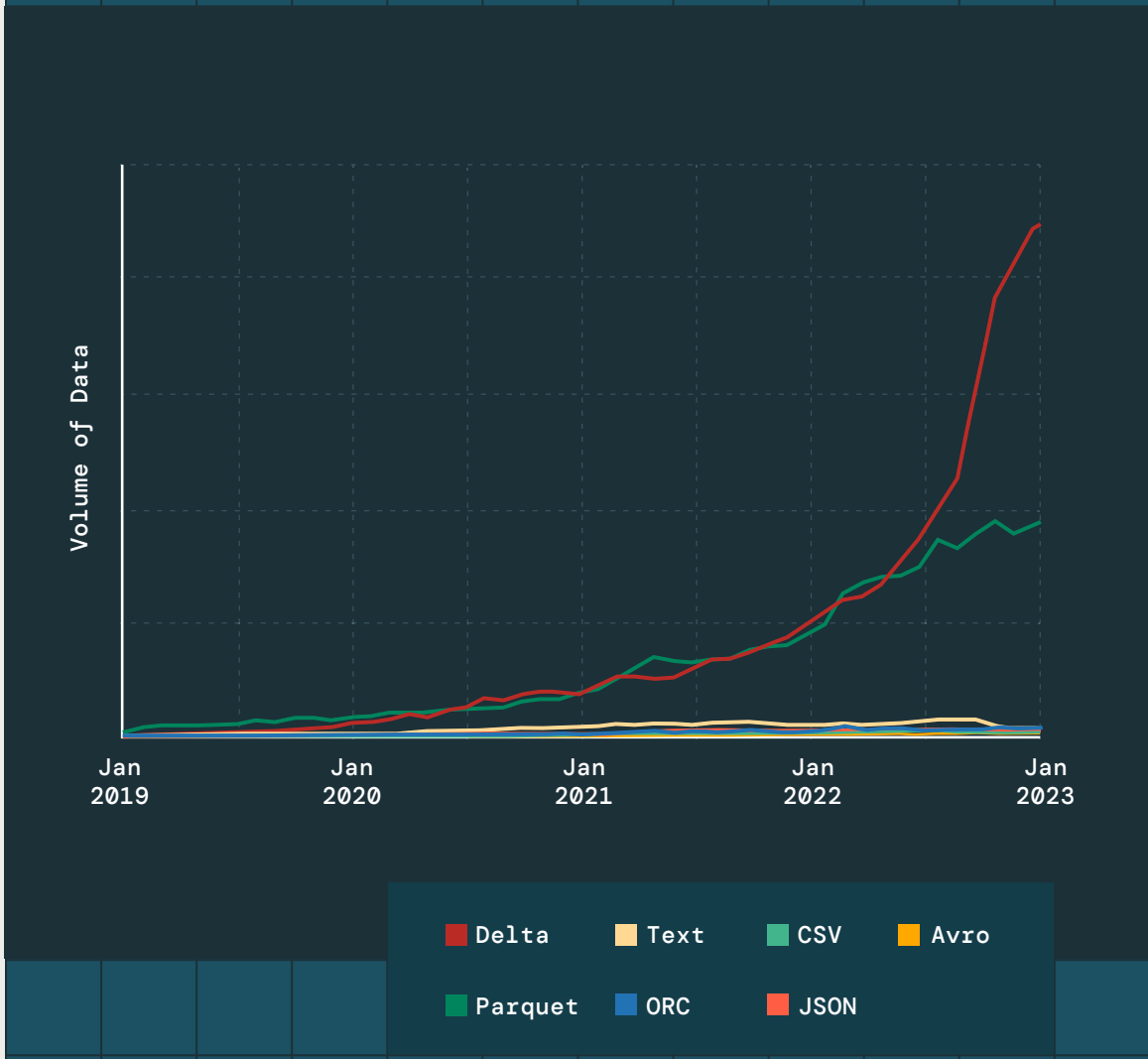


# Rising tides: the volume of data in Delta Lake has grown 304% YoY

As the volume of data explodes, an increasingly large proportion is in the form of semi-structured and unstructured data. Previously, organizations had to manage multiple different platforms for their structured, unstructured and semi-structured data, which caused unnecessary complexity and high costs. The Lakehouse solves this problem by providing a unified platform for all data types and formats.

Delta Lake is the foundation of the Databricks Lakehouse. The Delta Lake format encompasses structured, unstructured and semi-structured data. Use has surged over the past 2 years. When compared to the steady, flat or declining growth in other storage formats (e.g., text, JSON and CSV), our data shows that a growing number of organizations are turning to Delta Lake to manage their data. In June 2022, Delta Lake surpassed Parquet as the most popular data lake source, reaching 304% YoY growth.

## VOLUME OF DATA MANAGED, BY STORAGE FORMAT



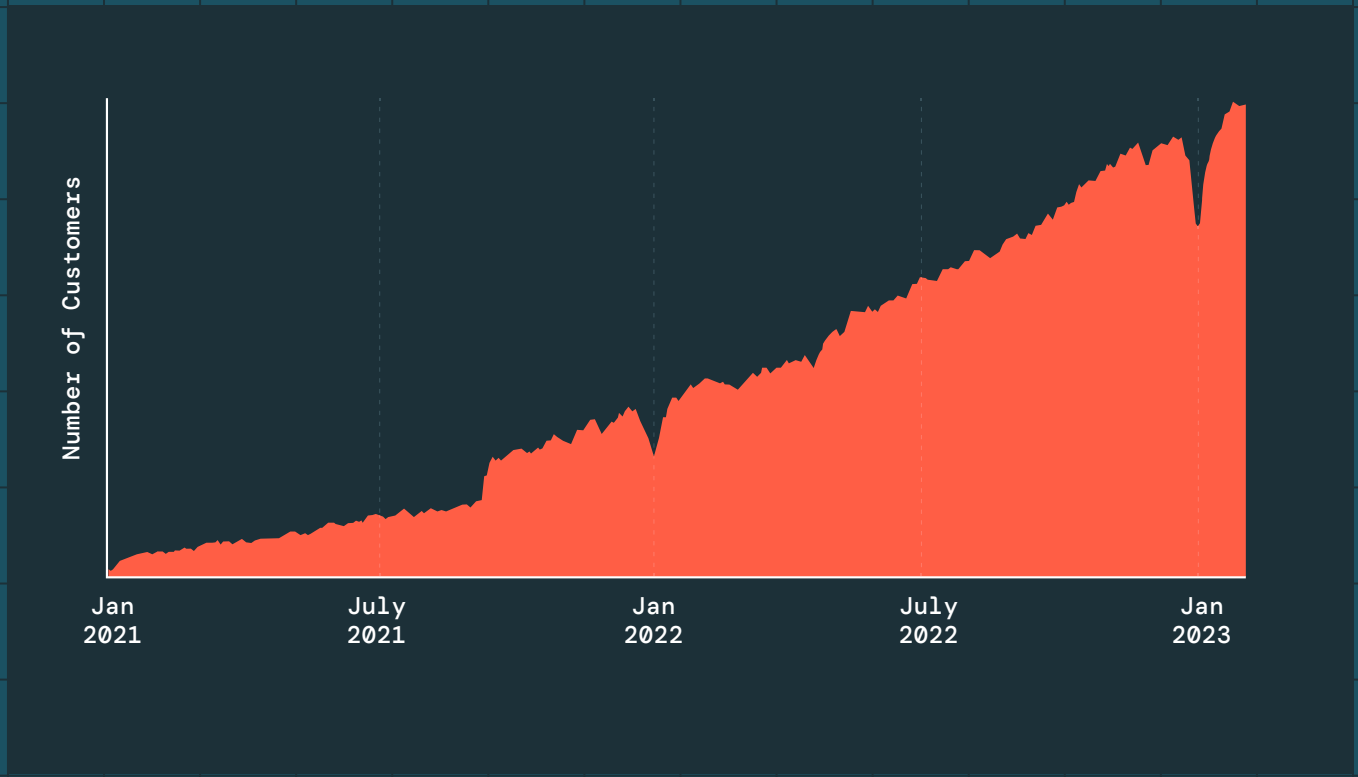
# Data warehousing grows, with emphasis on serverless

Over the past 2 years, companies have vastly increased their usage of data warehousing on the Lakehouse Platform. This is especially demonstrated by use of Databricks SQL — the serverless data warehouse on the Lakehouse — which shows 144% YoY growth. This suggests that organizations are increasingly ditching traditional data warehouses and are able to perform all their BI and analytics on a Lakehouse.


## DATA WAREHOUSING ON LAKEHOUSE WITH DATABRICKS SQL

Note: There is a spike in October 2021 as a result of the ungated preview launch of Databricks SQL, followed by General Availability in December 2021.

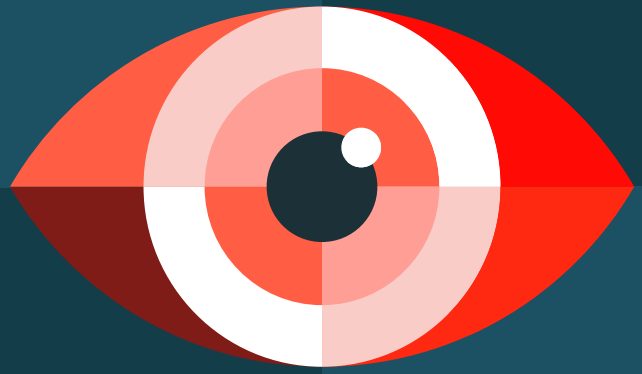
Data consistently dips in the last week of December due to seasonality.



CONCLUSION

# Generation AI

We're excited that companies are progressing into more advanced ML and AI use cases, and the modern data and AI stack is evolving to keep up. Along with the rapid growth of data integration tools (including our fastest growing, dbt), we're seeing the rapid rise of NLP and LLM usage in our own data set, and there's no doubt that the next few years will see an explosion in these technologies. It's never been more clear: the companies that harness the power of DS/ML will lead the next generation of data.



# About Databricks

Databricks is the data and AI company. More than 9,000 organizations worldwide — including Comcast, Condé Nast and over 50% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Instagram.

DISCOVER LAKEHOUSE

databricks