snowflake®

TOP 6 DATA SCIENCE AND ANALYTICS TRENDS FOR 2023

How the Data Cloud accelerates machine learning

CHAMPION GUIDES

EBOOK

TABLE OF CONTENTS

INTRODUCTION

Data science has evolved dramatically over the last 10 years. It has grown to encompass the emergent fields of artificial intelligence (AI) and machine learning (ML) and has fueled the rise of big data—large volumes of data arriving at high velocity. As data science expands its reach and scale, cloud computing has kept pace, meeting practitioners' needs by providing nearly unlimited resources to support ambitious and complex projects.

Despite an increased investment in data science and machine learning (ML), very few organizations have experienced the full business impact or competitive advantage from their advanced analytics. The reason? Getting to production has remained a significant hurdle.

Data engineers, scientists, and ML engineers often work in isolation, with separate copies of data, leading to greater complexity and a lack of collaborative solutions. Data silos—wherein data is scattered across systems and data lakes—weigh data scientists down with time-consuming busywork, while processing complexity puts an organization's focus on managing infrastructure instead of driving value.

But change is now afoot. Recent technology advancements are poised to significantly impact the way in which data scientists and data analysts work. In 2023, six trends have the potential to accelerate ML and move organizations from descriptive and diagnostic analytics—explaining what happened and why—toward predictive and prescriptive analytics that forecast what will happen and provide powerful pointers on how to change the future.

In this ebook, you will learn how:

- Unified infrastructure supporting multiple programming languages is allowing data scientists, engineers, and analysts to leverage the maximum potential of each.
- Snowflake's Data Cloud can expand data access, data sharing, and the use of various data types, including unstructured data (in preview), through a secure ecosystem with access to ready-to-use third-party data.
- Feature stores enable data scientists to manage and deploy ML features at scale by delivering reproducibility, discoverability, and scalability.
- Data analysts and scientists are increasingly able to utilize the power of systems once reserved for ML engineers in order to facilitate and participate in more savvy production processes.
- Increasingly accessible web app development in Python is empowering data scientists to make their models more comprehensible and actionable.
- Rapid advancements in open source libraries, tools, and frameworks demonstrate the need for a solution that future-proofs data science and ML investments.

FOCUS ON REPEATABLE PRACTICES FOR PRODUCTION ML

In recent years, huge investments have been made in data science, AI, and ML, guided by the promise of higher financial returns, more efficient processes, and greater overall business resilience. Adoption of AI has more than doubled since 2017, according to McKinsey's Global Survey on AI, with companies that make greater investments in AI pulling ahead of their competitors. The value of data science is a given; in 2023, companies will turn their attention toward streamlined production and optimization.

Businesses will also focus on maximizing the impact of AI, ML, and data science initiatives by refining processes, modernizing legacy systems, and adopting cohesive tool sets available to leverage these fields. Barriers that have long divided data science and business users, developers of different programming languages, and more, are quickly being broken down.

Advancements made in 2022 point to six exciting trends for data science and ML in 2023. Emerging tools and technologies can accelerate the work of data scientists, remove data silos, and expand the possibilities of structured and unstructured data alike. It's an exciting time to work in this field.

The cloud underlies this acceleration. Data scientists, data engineers, and data analysts benefit from cloud technologies that provide elastic, and virtually unlimited amounts of compute resources. In addition, the cloud enables the elimination of data silos by consolidating data lakes, data warehouses, and data marts for fast, secure, and easy data sharing and analysis in a single location to strengthen collaboration across data teams.

In short, data is more actionable than ever and new tools are using that reality to move the needle forward. As a result, organizations are on the brink of mobilizing data to not only predict the future but also to increase the likelihood of certain outcomes through prescriptive analytics.

Here are six trends that will shape data science in 2023 and continue the evolution of analytics towards ML.

TREND #1: UNIFIED TOOLS AND INFRASTRUCTURE FOR SQL AND PYTHON

As the amount of data and number of applications grew, on-premises and other legacy data warehouses impeded scalability. The industry solved for this by copying data into cloud object stores and setting up processing infrastructure for programming languages like Python, SQL, and Java. This turned into complex infrastructure management and limited collaboration across teams.

SQL and Python are the preeminent languages of the modern data landscape, and each boasts unique benefits. SQL's speed of querying and aggregation is a huge asset, while Python's ability to manage complex analytics and transformations using its rich, open-source ecosystem is a necessity for many organizations. There are clear benefits of each one, yet they essentially live in different worlds. Each runs on separate infrastructure and is developed with different tools, which have long prevented data scientists from capitalizing on the best of both.

This lack of interoperability has created a profound siloing effect wherein users of one language aren't able to collaborate on analysis or workflows with users of the other. Even for those confident in both languages, code-switching and the friction between querying in each language can be both frustrating and time-consuming.

Yet, the division between SQL and Python is diminishing thanks to tools like dbt, Hex, and Snowflake's Snowpark, which combine the languages for any data task, drawing on the advantages of both for different operations during analysis. Here's how:

dbt: Data transformation workflow for data teams following software engineering best practices like modularity, portability, CI/CD, and documentation in their cloud warehouse. While a SQL-first transformation workflow, in 2022, dbt introduced Python as a second language to meet the growing demand for seamless solutions to work between languages on the same project.

Hex: Hex is a modern platform for analytics and data science that makes it easy to connect to data, analyze it in collaborative SQL and Python-powered notebooks, and share work as interactive data apps and stories. To provide near-unlimited processing scalability, its approach is not to load all data into a notebook but rather to push compute down to the warehouse.

Snowpark: Snowpark is a new developer framework for Snowflake. It allows data engineers, data scientists, and data developers to write code in their preferred language, and run that code in Snowflake. Supporting interfaces for development in SQL, Python, Java, and more, Snowflake allows easy context switching without moving data or setting up separate clusters.

Tools that unify programming languages are crucial for continued growth through collaboration. In 2023, data engineers, scientists, and analysts no longer have to work in isolation for lack of a shared language; they can work together to move from raw data to insight. This knowledge-sharing ultimately creates more agile projects with better long-term results.

TREND #2: MANAGING AND DEPLOYING ML FEATURES AT SCALE WITH FEATURE STORES

When data scientists build new ML models, they face the arduous tasks of preparing data and creating features. Features are created by sourcing and preparing data columns in a specific format that can be fed into machine learning models. Once features are generated for one model, data scientists encounter the additional challenge to either rewrite the same features or spend time searching for and finding existing features to use for the next model.

Thankfully, 2022 saw a sharp uptick in the adoption of feature stores—a central repository that helps increase searchability, collaboration, and scalability of ML features. Data scientists can quickly find features that are transformed and ready for use, resulting in both faster experimentation and faster time to production. The benefits of a feature store include the ability to increase collaboration across teams through the reuse of work from other data scientists. Additionally, feature stores reduce the time and effort required to deploy a trained model in a production environment because data scientists no longer need to redefine what is often an existing data pipeline.

Today, feature stores are viewed as the best way to improve ML models because teams can more easily access enhanced and refined data that is relevant to their models. However, building a feature store is no small feat. Operationalizing features is challenging because reproducibility, discoverability, and scalability must be built into the feature store.

1. **Model reproducibility** requires features to be centralized in a single location and for data and features to be versioned. Data scientists must be able to go back in time to discover the features and data used to train a model. Features must also be defined once and then available for all future use cases, which means feature stores must update regularly and manage version control.
2. **Discoverability** requires feature creation to be taken out of individual notebook instances and centralized in a unified repository. To enable collaboration and assist in the efficient reusability of the feature store, a catalog must exist that makes features easily discoverable and searchable.
3. **Scalability** means items in the feature store can keep up without heavy operational burden as the ML use cases grow. A feature store should be able to scale from hundreds to millions of features and continue to efficiently serve both training and inference workflows. Rather than running duplicate pipelines for training and inference, centralized processing accelerates access to features and reduces redundant processing.

Snowflake provides two approaches for building feature stores, both of which avoid creating new systems or new silos of data between data scientists and data analysts.

The first approach is to leverage an open-source solution such as Feast as the feature store interface while Snowflake becomes the store and engine for features. Features persist on the single data platform, supported by any existing ingestion, ELT, and cataloging tools. With this solution, organizations own both the management of the data pipelines and the interface on which data scientists discover and access data and features in one centralized, scalable location.

Alternatively, the second approach is to leverage a managed feature store solution on top of Snowflake. Customer data remains in its raw and modeled form in Snowflake's Data Cloud, and while the transformation of the data is executed in Snowflake using SQL or Snowpark for Python, the pipeline orchestration and management are abstracted by the feature store provider. Snowflake partners in this space include Tecton and Iguazio.

TREND #3: ANALYSTS AND DATA SCIENTISTS GET THE POWER OF ML ENGINEERS

Python is increasing in popularity across a wide variety of industries and is being used not just by ML developers—who traditionally rely on it for its ability to handle a massive amount of data requests—but also by data scientists in order to create trustworthy models and systems built on intuitive, flexible code. Python's popularity continues to grow as it becomes the *lingua franca* of data science. In fact, 70% of ML developers and data scientists now report using Python.

Historically, it has been difficult to run Python at an enterprise-level due to its complex infrastructure and security risks from its open-source ecosystem. As such, despite its scalability and performance, the language has been reserved for ML developers who are comfortable managing complex infrastructure and dealing with security patching for mission-critical applications. Yet, new developer frameworks that eliminate the burden of managing infrastructure for Python (alongside other languages) are making it easier than ever for data scientists and analysts to leverage Python at scale, too.

As Python spreads across data functions, data analyst teams will be able to leverage its speed and open-source libraries to participate in cleaning and structuring data. With greater participation in this area, errors are mitigated, productivity is enhanced, and better insights are developed to enable high-quality decision-making.

Cross-collaboration does not end with data analytics teams, however, data scientists well versed in Python will likely expand their roles as well. Instead of leaning on Python for exploratory work, these will take an increasingly active role in production-related work. Here we see a crucial bridge between the world of development and production thanks to access to robust infrastructures that can do both—with a strong data foundation that provides ad hoc access to data living anywhere and compute that is easy to burst on demand. With no lag when creating clusters, data scientists will be able to work at a greater scale with insights into how their models will be taken to production.

AutoML embedded into scalable platforms will also support the empowerment of data scientists and analysts with limited machine learning expertise to quickly train and deploy powerful ML-powered insights. AutoML are tools that automate tasks associated with developing and deploying ML models, traditionally executed only by expert data scientists. They enable more data users to automate one or more parts of the ML workflow, including data preparation, model training and selection, and more.

These tools are making a huge difference for data scientists and analysts alike by addressing busy work (loading, selecting, preparing, and cleaning data) that previously took up 80% of their time, but are now estimated to take just 45%, according to a survey of data scientists conducted by Anaconda and reported by Datanami. As such, AutoML increases productivity and provides more time to conduct analysis.

As AutoML tools become more scalable and transparent, their adoption will likely increase, enabling more players on the data team to leverage the power of ML in production.

TREND #4: MORE USE CASES FOR UNSTRUCTURED DATA

Statista's rejected final amount of data created in 2022 was 97 zettabytes, a number that is expected to increase rapidly in coming years. By 2025, IDC projections, reported by Analytics Insight, also predict that 80% of the world's data will be unstructured, which should sound alarm bells for organizations since only 0.5% of these resources are analyzed today.[1]

These anticipated volumes of unstructured data point to the escalating need for data scientists to be able to analyze unstructured data alongside structured and semi-structured data (i.e., captured data with some structural elements but isn't formatted in conventional ways, such as user logs and web activity delivered in formats like JSON).

Unfortunately, unstructured data includes digital files that contain complex data such as text, images, video, audio, .pdf files, and industry-specific file formats. As mountains of unstructured data are generated through things like written documents, the failure to process these effectively is a missed opportunity to gain a competitive edge. It is the complexity of these unstructured data sources that makes it extremely challenging to analyze them along with other data types. Without a single source that supports all data types, unstructured data gets stuck in silos. As a result, data scientists cannot easily search, analyze, or query unstructured data, and instead must gather it from multiple systems.

In addition to data management issues, it's virtually impossible for any organization to produce or collect all the data needed to uncover business and competitive trends. Increasingly, the ability to share and join data sets, both within and across organizations, is viewed as the best way to derive more value from data. That's why data scientists and data analysts are continually on the hunt for more data to supplement their ML models and analysis with external data to improve the accuracy of model predictions.

With Snowflake, data scientists and data analysts have access to a global, unified system for managing all data types, including unstructured data. As more and more unstructured data continues to be produced, theSnowflake Data Cloud will continue to provide a single, consolidated source for data that enables data scientists and data analysts to accelerate the speed at which data is accessed and processed to extract value from all data.

Hand in hand with this ability to use various data types, Snowflake also enables secure, governed, and seamless access to external data, allowing those who use Snowflake to fulfill their compliance obligations under various data protection regimes.

Snowflake's Data Cloud is an ecosystem where Snowflake customers, partners, data providers, and data service providers connect to their own data and seamlessly share and consume data and applications from other users. Underpinned by Snowflake's platform, the Data Cloud eliminates barriers presented by siloed data and enables organizations to unify and connect to a single copy of data. In addition, the Data Cloud is a seamless way to derive value from rapidly growing commercialized data sets with fast, easy, and governed access.

Empowering the Data Cloud is **Snowflake's Secure Data Sharing technology**, which allows companies to easily collaborate with their internal and external business partners by removing traditional data transfer barriers. With Snowflake, data is never copied or moved. Instead, the providing party grants users access to live data from its original location. Thanks to Snowflake's separation of storage and compute, latency or contention from concurrent users is never an issue. Because changes to data are made to a single version, data is always up-to-date for all consumers and ensures data models are consistently using the latest version.

On Snowflake Marketplace, users can access and purchase live, ready-to-query data and applications from third-party providers. Leveraging the same secure data sharing technology, there's no need for ETL processes, allowing data scientists and analysts to start using data from third-party providers more quickly. Snowflake Marketplace simplifies and accelerates the discovery and evaluation of third-party data with actionable data samples that can be seamlessly joined with first-party data to validate prototypes. Once a data set has been evaluated, the purchasing process can be handled directly in the product and the additional charges become part of your Snowflake invoice.

External data is available and accessible to all Data Cloud users with just a few clicks. Once it's in the Data Cloud, data is ready to be shared and consumed. There's no need to send CSV files or deal with manual version control. Data scientists can enrich models with seamless access to almost-unlimited data on any topic, including real-time and evolving circumstances.

TREND #5: DATA SCIENTIST ALSO BECOME APP DEVELOPERS WITH PYTHON

Data scientists at any organization are responsible for sharing data and models in ways that are comprehensible and compelling for their collaborators. Yet, data scientists are often hamstrung to put results in dashboards that don't support new ways of communicating this information, and rarely do data science teams have full stack app developers at their disposal to build internal products to share their work.

In the coming year, we will see a significant change to the status quo as data scientists are also empowered as app developers in Python thanks to open-source development frameworks like Streamlit that make it easy to go from idea to app using only Python, effectively closing the ML-to-action gap.

Data scientists can quickly build interactive applications with rich components like charts, input fields, and more to equip their team to engage with data and models—without the traditional complexity involved in building a web app like defining routes, handling HTTP requests, and writing HTML, CSS, or JavaScript. Facilitated by this platform and others like it, data scientists can develop web apps that bring ML models to life in ways that data science teams have never accessed before.

With the ability to rapidly build applications only with Python and share and iterate on these interactive platforms, teams can more readily mobilize data and put ML-powered insights into the hands of business users to take action—which is the end goal, after all.

Historically, a gulf has separated the work of data scientists and business users. Many data scientists spend a significant amount of their time and energy developing models that may or may not be totally intelligible or actionable by business users. Without accessible, trustworthy data and analysis, action stalls out and the promise of data science can never be fully realized.

Streamlit is changing all that. Already adopted by many Fortune 500 companies, Streamlit effectively democratizes web app development. And since Streamlit was acquired by Snowflake, concerns like infrastructure management, elasticity, and security data governance on these applications are a thing of the past. As part of an ongoing integration, users can seamlessly deploy and securely share their apps leveraging Snowflake's infrastructure security and reliability.

As more organizations adopt platforms like Streamlit to empower data scientists as app developers, model development will be accelerated, collaboration will become more meaningful, and insights will be shared more efficiently. This enables agility and action on the part of organizations, supporting long-term growth.

TREND #6: CONTINUOUS GROWTH IN OPEN SOURCE ML LIBRARIES, TOOLS, AND FRAMEWORKS

The field of data science is evolving rapidly. Not only are new ML and AI developments released every month—many of them open source—but new startups, tools, and solutions emerge regularly. Take, for example, the rapid rise and staggeringly quick adoption of ChatGPT. The chatbot developed by OpenAI with a fine-tuned language model quickly went viral, bringing open-source AI to the public domain for the first time. With the rapid pace of innovation, adoption, and change occurring in this space, it's imperative not to get locked into using a single tool.

That's also why it's important to select a framework-and-algorithm-agnostic platform that can also cohesively interact with other tools. By choosing a future-proof platform, you ensure that upcoming ML tools will continue to work seamlessly with the platform you have. After all, the last thing you want to do is re-platform just to use the next generation of tools.

What makes Snowflake's platform unique is its modern architecture. Designed with separate but logically integrated compute and storage, Snowflake eliminates manual cluster-building efforts that other systems must perform to make separate layers work together. As a result, Snowflake offers a **multi-cluster, shared data architecture** that provides near-infinite scalability, instant elasticity, and extremely high levels of concurrency to power all the necessary processing for data science and ML from data preparation to model inference and deployment in an application.

In addition to the underlying architecture that supports all data types and brings together multiple languages for processing in a single engine, Snowflake supports integrations with the data science ecosystem in a variety of ways.

- Snowflake's **External Functions** allow users to interact with any third-party, hosted, or custom ML service outside of Snowflake using SQL. This could be something like a voice-to-text AI service and other NLP services, or even an ML model that is deployed in an external environment for real-time prediction requests.
- To develop and deploy their Python and Java code with **Snowpark**, developers have always had the flexibility to work from their favorite integrated development environment (IDE) or notebook. Using the Snowpark client library, developers can push down processing from almost any development tool into Snowflake so there is no need to stop using the tools you already love.
- Since the power of Python lies in its rich ecosystem of open-source packages, as part of the Snowpark for Python offering, we are excited to bring seamless, enterprise-grade open source innovation to the Data Cloud via our Anaconda integration. With Anaconda's comprehensive set of open-source packages and seamless dependency management, you can speed up your Python-based workflows.

  In addition, Snowflake also has integrations to popular open-source tools across the many layers of the machine learning stack. This includes integrations into the popular Apache Iceberg tables (in preview) to more easily access all your data to integrations to feature store solutions such as Feast to help manage the end-to-end lifecycle of your ML features.
- With the growth of unstructured data comes the parallel development of methods to manage and process it. For example, new labeling services provide manual tagging of images and other unstructured data. With **Snowflake Secure Data Sharing**, unstructured data (in public preview) can be shared with a provider to add tags to data without moving your data. In addition, unstructured data analysis tools can be used on top of Snowflake to enhance unstructured data using NLP services offered by companies such as Hugging Face and AI cloud services such as AWS Rekognition.

ACCELERATE YOUR MACHINE LEARNING IN 2023

It's remarkable how quickly data science has become mainstream. In the last 10 years, companies have expanded their focus from reporting and historical analysis to conducting data science with advanced mathematical models and ML. But most organizations using machine learning have yet to see a return on their investment as they struggle to take their projects into production.

A modern data platform is a necessary foundation to provide data access and processing in a way that easily scales and meets the most stringent security needs of even the most regulated industries such as financial services and healthcare. Snowflake provides an architecture supportive of many programming languages, including SQL and Python, that enables data consolidation, efficient data preparation, easy access to open-source libraries, and extensive integrations into the data science ecosystem. Your data is mobilized, which allows you to benefit immediately from new trends in data science and ML.

With Snowflake, the complexity of taking data science and machine learning into production is removed. Are you ready to accelerate your machine learning and capture its full value?

ABOUT SNOWFLAKE

Snowflake delivers the Data Cloud—a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake's platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data Cloud. **snowflake.com**

CITATIONS

[1] bit.ly/3lkt1qx