snowflake®

TOP 6 DATA SCIENCE AND ANALYTICS TRENDS FOR 2022

How the Data Cloud accelerates machine learning

CHAMPION GUIDES

EBOOK

TABLE OF CONTENTS

INTRODUCTION

Data science has evolved dramatically over the last 10 years. However, very few organizations have experienced the full business impact or competitive advantage from their advanced analytics, despite significant investments in data science and machine learning (ML). The reason? Many of the tools needed to scale ML are too complicated, and necessary skill sets are in short supply. But change is now afoot. Recent technology advancements are poised to significantly impact the way in which data scientists and data analysts work. In 2022, 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:

- The gap between ML and analytics can be bridged when data scientists have programming language flexibility and data analysts benefit from easy-to-use ML tools and consolidated data platforms
- 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
- New distributed training frameworks offer an alternative superior to Spark while delivering up to 2,000x faster performance
- Rapid advancements in ML libraries, tools, and frameworks demonstrate the need for a solution that future-proofs data science and ML investments

SHIFT TOWARD PREDICTIVE AND PRESCRIPTIVE ANALYTICS CONTINUES

In 2022, the field of data science is poised to finally live up to the high expectations that many organizations have held for years. Over the last 10 years, huge investments have been made in data science and ML, guided by the hope that they would transform the way companies do business. However, many organizations continue to feel challenged to drive real business impact with analytics, as evidenced by the fact that only 10% of organizations are seeing significant financial benefits from their investments in AI, according to a report from *MIT Sloan Management Review* and Boston Consulting Group.[1]

Organizations invest in data science because it promises to bring competitive advantages, but many of the tools and skill sets needed to scale ML have been missing or in short supply. Data scientists continue to be a sought-after and expensive resource, and their valuable efforts tend to be relegated to time-consuming tasks such as data selection and data preparation. Conversely, data analysts are in abundant supply in most companies and already know how to address business problems directly, but they lack the technical background required to make the jump from analytics to data science to build their own ML models.

Remarkably, advancements made in 2021 point to six exciting trends for data science and ML in 2022. New tools and technologies have emerged—and continue to be released every month—that accelerate the work of data scientists and enable data analysts to move beyond descriptive analytics and conduct light data science and ML.

Underlying this acceleration is the cloud. Data scientists and data analysts benefit from cloud technologies that provide 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.

In short, data is transforming into an actionable asset, and new tools are using that reality to move the needle with ML. 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 2022 and continue the evolution of analytics toward ML.

TREND #1: EASY-TO-USE ML TOOLS EMPOWER DATA ANALYSTS AND DATA SCIENTISTS

Most organizations employ an abundance of data analysts and a limited number of data scientists, due in large part to the limited supply and high costs associated with data scientists. Since analysts lack the data science expertise required to build ML models, data scientists have become the de facto bottleneck for broadening the use of ML.

However, new and improved ML tools are opening the floodgates on ML by automating the technical aspects of data science. Data analysts now have access to powerful models without needing to build them manually, and data scientists have the ability to automate multiple steps in the process and increase their own productivity. Specifically, automated machine learning (AutoML) and AI services via APIs are removing the need to prepare data manually and then build and train models.

AUTOML

AutoML tools are aptly named: They automate the tasks associated with developing and deploying ML models, which are traditionally done by data scientists. AutoML is game changing for both data scientists and data analysts because these tools enable data users to automate one or more parts of the ML workflow, including data preparation, model training and selection, and more.

But AutoML isn't just for analysts. Thanks to its power, AutoML is making a huge difference for data scientists by addressing the busy work (loading, selecting, preparing, and cleaning data) that previously took up to 80% of their time but is now estimated to take 45%, according to a survey of data scientists conducted by Anaconda and reported by Datanami.[2] By eliminating these time-consuming data chores, AutoML increases data scientists' productivity and provides more time to conduct analysis. Human errors found in manual modeling processes are also reduced, which improves accuracy.

In addition, the one historical flaw of AutoML was that it was seen as a black box, but that challenge has been solved. AutoML services now provide transparency and explanations for their models, which is key for auditing and detecting bias. For data scientists, AutoML transforms how quickly they can build and test multiple models simultaneously.

In the last couple of years, AutoML tools from providers such as DataRobot, Dataiku, and H2O have experienced significant advancements, and solutions such as Amazon SageMaker Autopilot have been introduced and gained traction. And, in 2021, the increased automation of model deployment and monitoring helped with the productionizing of models.

AI SERVICES VIA APIS

Another approach growing in popularity is AI services, which are ready-made models available through APIs. Rather than use your own data to build and train models for common activities, organizations can access pre-trained models that accomplish specific tasks. Whether an organization needs natural language processing (NLP), automatic speech recognition (ASR), or image recognition, AI services simply plug-and-play into an application through an API, which requires no involvement from a data scientist.

Amazon provides a variety of fully managed AI services, including Amazon Lex, Polly, Rekognition, Forecast, and Translate.[3] To illustrate the value, Rekognition allows an image to be sent from an app through the API to Amazon; the AI service then returns a classification and description of what the image is. These types of utilities not only save time and effort, but they also free up data scientists to focus on building and training models that are highly customized to their business, rather than re-creating commonly used services.

AutoML tools and AI services lower the barrier to entry for ML, so almost anyone can now access and use data science without requiring an academic background. However, the true power of these tools is unleashed when they are integrated seamlessly with your existing technologies. With **Snowflake Partner Connect,** organizations can receive faster insights from their data through pre-built integrations between Snowflake and technology partners' products. Snowflake Partner Connect makes it quick and easy to try new ML tools and services and then adopt those that best meet your business needs.

TREND #2: A CONSOLIDATED PLATFORM CLOSES THE GAP BETWEEN ANALYTICS AND ML

Everyone knows data silos exist within and across organizations. However, few realize that these silos also take the form of "analytics silos," particularly between data scientists and data analysts. These analytics silos have formed as a result of the different ways the two roles work and their respective skill sets. Data silos are just one part of the difference: Data scientists and data analysts use different data (raw versus processed), data sources (data lakes versus warehouses and marts), languages (Python and Java versus SQL), and tools (ML versus BI).

Much like organizational silos, analytics silos thwart collaboration and integration opportunities between data scientists and data analysts. This situation results in organizations missing out on the combined power of these two teams, which is exponentially stronger than simply the sum of the two parts.

For example, data analysts leverage data to provide key business metrics and answer questions around why something happened. According to Sisu, data analysts' superpower is speed, which is used to analyze data sets quickly and work with business stakeholders to uncover potential insights.[4] While their goals are to help companies monetize market opportunities and improve competitive advantage, most of the work data analysts do is backwardlooking because they lack the data science skills necessary to build predictive ML models. Instead, data analysts rely on BI tools whose dashboards have built-in limitations. While they can use data to understand what has already happened, it's challenging for data analysts to be proactive and explore data deeply to figure out what will happen and how to influence it.

Conversely, data scientists have the ability to build ML models that not only predict but also influence business outcomes. However, they are not as well versed in the dynamic and fluctuating business environment as data analysts are. Sisu describes data scientists as "narrow-and-deep workers," and their focus frequently results in organizations trying to focus data science efforts at known problems (often uncovered by data analysts) to maximize their value and contributions rather than potentially wasting time and effort on the unknown.[5]

Snowflake's Data Cloud provides the tools to help deliver stronger outcomes and scale. Through Snowflake, analytics silos are eliminated. The same consistent, governed metrics and data are available for both analytics and ML through a shared feature store and reuse of data engineering pipelines. When data science insights are shared in Snowflake's platform, data analysts can access and incorporate them into dashboards and analysis, thus broadening the scope of impact of the models the data science team builds.

In addition, **Snowpark** (in public preview) offers a developer framework that brings data programmability and flexibility to the Data Cloud. Snowpark bridges the language divide between data scientists and data analysts by extending language support beyond SQL. Data scientists are empowered to code in their programming language of choice (Python [in preview], Java, Scala) and can access, visualize, and process data as part of their ML workflows without moving data. As a result, data scientists and data analysts collaborate on the same data within the Data Cloud.

TREND #3: SNOWFLAKE'S DATA CLOUD EXPANDS ACCESS TO NEW DATA

According to IDC, 64.2 zettabytes of data was created or replicated in 2020, which exceeded forecasts and demonstrated the impact of the global pandemic on digital usage.[6] 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.[7]

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. Unfortunately, unstructured data includes digital files that contain complex data such as images, video, audio, .pdf files, and industry-specific file formats. This complexity makes it extremely challenging to join with other data types, and, 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 results.

With Snowflake, data scientists and data analysts have access to a global, unified system for managing all data types, including unstructured data (in preview). As more and more unstructured data continues to be produced, the Data Cloud will continue to provide a single, consolidated source for data that enables data scientists and data analysts to accelerate data modeling and extract value from all data.

Hand in hand with this ability to use various data types, Snowflake also enables secure, governed, compliant, and seamless access to third-party data in three ways.

1. **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 data services shared by 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.
2. Empowering the Data Cloud is **Snowflake Secure Data Sharing,** which removes traditional data transfer barriers. With Snowflake, data is generally never copied and transmitted. Instead, users can share live data from its original location. Those granted access simply reference data in a controlled and secure manner without latency or contention from concurrent users. Because changes to data are made to a single version, data remains up to date for all consumers, which ensures data models are always using the latest version.
3. Snowflake Secure Data Sharing is the technology foundation for **Snowflake Data Marketplace,** which serves as a single location to access live, ready-to-query data. Secure, governed data can be shared with, and received from, an ecosystem of business partners, suppliers, and customers, or from third-party data providers and data service providers. Snowflake Data Marketplace removes the arduous processes involved in locating the right data sets, signing contracts with vendors, and managing the data to make it compatible with internal data. Instead, data scientists and data analysts can source new data with ease. In addition to Snowflake Data Marketplace, organizations can use private data exchanges to share data with trusted partners, suppliers, vendors, and customers through Snowflake Secure Data Sharing.

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 #4: 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, 2021 saw a sharp uptick in the adoption of feature stores, which act as a repository for ML features that helps increase searchability, collaboration, and scalability of features. Data scientists can quickly find features that are transformed and ready for use, which results in 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 many times will be an existing data pipeline.

Increasingly, 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 features in the feature store can continue to grow as the ML use cases grow. A feature store should be able to scale from hundreds to thousands of features and continue to efficiently serve both training and inference workflows.

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 build a feature store directly on Snowflake. Features persist on the single data platform, supported by any existing ingestion, ELT, and cataloging tools. Data scientists discover and access data and features in one centralized, scalable location, which improves the speed and ease of machine model training and machine model inference or scoring.

Alternatively, the second approach is to leverage Snowflake partners for building feature stores. Customer data remains in its raw or modelled form in Snowflake's Data Cloud, and Snowflake partner applications build a semantic layer on top of that data to manage, monitor, and present features. Snowflake partners in this space include Tecton, Rasgo, AtScale, Iguazio, and Hopsworks.

TREND #5: NEW GENERATION OF DISTRIBUTED TRAINING FRAMEWORKS OFFERS COMPELLING ALTERNATIVE TO SPARK

Data scientists are always looking for strategic ways to inject efficiency into training and deploying models. Recently, a new generation of distributed training engines has surfaced that delivers on that goal by providing tremendous speed and performance gains over Apache Spark.

One approach that's gaining attention is Dask, a distributed training framework built in Python.[8] Dask is designed to enable data scientists to improve model accuracy faster. Data scientists can do everything in Python end to end, which means they no longer need to convert their code to execute in Spark. The result is reduced complexity and increased efficiency.

Another open-source Python framework is RAPIDS, which is built on top of Dask.[9] RAPIDS optimizes compute time and speed by providing data pipelines and executing data science code entirely on graphics processing units (GPUs) rather than CPUs. Saturn Cloud recently compared RAPIDS to Spark and discovered that model training with RAPIDS took one second on a 20-node GPU cluster, while Spark took 37 minutes on a similarly priced 20-node CPU cluster. Saturn Cloud concluded that RAPIDS enables 2,000x faster processing using GPUs while costing a fraction of the price.[10]

The impact of these distributed training frameworks is already being seen in the real world. Walmart uses RAPIDS with Dask and XGBoost (an ML algorithm) for its data analytics and ML, and NVIDIA reports that Walmart has found that "one GPU server requires only 4% of the time needed to run the same forecasting models versus a 20-node CPU server."[11] That translates to Walmart running models in four hours that previously took several weeks using CPUs.

While organizations are thinking strategically about training frameworks, some have run into barriers in the past. Today, new technologies are unlocking what's possible and demonstrating how much faster things can be when everything is done directly with Python. By eliminating the need to convert models into Spark, organizations are reducing complexity and increasing efficiency. And it's easy to try different distributed training frameworks on Snowflake's platform to find what works best.

TREND #6: CONTINUOUS RELEASES PROVIDE NEW OPTIONS FOR ML LIBRARIES, TOOLS, AND FRAMEWORKS

The field of data science is evolving rapidly. Not only are new ML and AI developments released every month, but new startups, tools, and solutions emerge regularly. With the rapid pace of innovation 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 platform that is vendor-, framework-, and algorithm-agnostic. 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 nearly infinite scalability, instant elasticity, and extremely high levels of concurrency to power the Data Cloud.

In addition to the underlying architecture, Snowflake supports data science in a variety of ways.

- Snowflake's **External Functions** allow any third-party, hosted, or custom ML service to be accessed easily using SQL.
- Recognizing that various teams may prefer languages other than SQL, **Snowpark** (in public preview) extends language support for Java, Scala, and Python (in preview). Snowpark allows data scientists to write code in their language of choice using familiar programming concepts, such as DataFrames, and then execute data preparation and workloads directly on Snowflake.
- **Java user-defined functions (UDF)** (in public preview) are supported to enable trained models to run within Snowflake. That means models built and trained in an ML partner's technology can be brought into and run directly on Snowflake's elastic performance engine.
- With the growth of unstructured data comes the parallel development of methods to manage and process unstructured data. For example, new labelling 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 have tags added 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 2022

It's remarkable how quickly data science has become mainstream. In the last 10 years, companies have shifted their focus from reporting and historical analysis to conducting data science with advanced mathematical models and ML. The cloud changed everything. With the ability to inexpensively collect and store more and more data came the need to build data models powered by ML.

Today, a modern data platform is a necessity if you want to analyze and share data quickly and scalably with security and governance built in. Snowflake provides an architecture that enables data consolidation, efficient data preparation, and an extensive partner ecosystem. Your data is mobilized, which allows you to benefit immediately from new trends in data science and ML.

With Snowflake, limitations on data science are removed. Are you ready to accelerate your machine learning?

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. on.bcg.com/2Kh3grW
2. bit.ly/3lg5EZp
3. amzn.to/3sv5qFx
4. bit.ly/2XJOa1u
5. bit.ly/2XJOa1u
6. bit.ly/3D7u3wh
7. bit.ly/3lkt1qx
8. dask.org
9. rapids.ai
10. bit.ly/3srNbRv
11. bit.ly/3FZozW8