SNOWFLAKE FOR DATA SCIENCE

snowflake

WHITE PAPER

INTRODUCTION: IT'S ALL ABOUT THE DATA

Machine learning (ML) technologies have entered the mainstream. According to a 2019 TDWI survey on artificial intelligence (AI) and ML use, 92 percent of respondents reported using machine learning technology, and 85 percent said they are building predictive models using tools for ML.[1]

Data scientists require massive amounts of data to build and train these machine learning models. In the age of AI, fast and accurate access to data has become an important competitive differentiator. Data management (discovering, securing access, cleaning, combining, and preparing the data for analysis) is commonly recognized as the most time consuming aspect of the process.

An efficient data platform is paramount

According to Forbes, scientists spend up to 80 percent of their time finding, retrieving, consolidating, cleaning, and preparing data for analysis and training.[2] The same study from Forbes found that data scientists not only spend most of their time massaging rather than mining or modeling data, but that 76 percent of these highly skilled professionals view data preparation as the least enjoyable part of their work.[3]

This white paper will help you identify the data requirements driving today's data science and ML initiatives, and explain how you can satisfy those requirements with a platform that supports industry-leading tools from Snowflake and its partners.

THE MOST COMPLETE PLATFORM FOR DATA SCIENCE

Snowflake's platform combines the power of data warehousing, the flexibility of big data platforms, the elasticity of the cloud, and live data sharing at a fraction of the cost of traditional data platform solutions. Snowflake delivers the performance, concurrency, and simplicity needed to store and analyze all your data in one location, both for internal use and to create a data exchange. Thousands of customers are standardizing on this platform because it satisfies three essential needs:

- **A single consolidated source for all data:** Snowflake helps data scientists access structured and semi-structured data from one consistent source, making it easy to find, consolidate, clean, and use more of your organization's data assets. Output from data science can seamlessly be incorporated back into Snowflake for access by business users.
- **Efficient, high-speed data preparation:** Snowflake provides efficient, dedicated virtual warehouses that can ingest, transform, and query data using SQL without impacting other users or departments. SQL in Snowflake is in many cases 10x more efficient at data preparation than other tools such as Spark, resulting in reduced latency between ML tasks.
- **An extensive partner ecosystem:** Snowflake has connectors to all the established and emerging data science technologies. This allows customers to choose the best data science tools for their needs, and all tools access a unified and consistent data platform. Snowflake seamlessly exports data to Amazon S3 and other blob stores for universal access by data science tools.

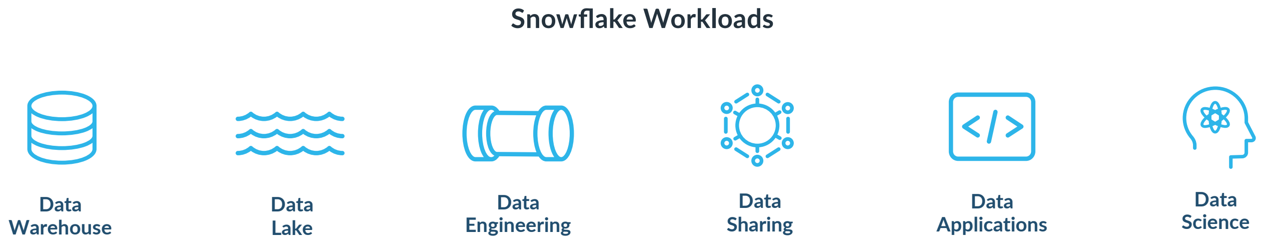

Figure 1: Snowflake enables many use cases and workloads in addition to data science, so your organization can leverage the power of a single platform for data, analytics, and predictive analytics.

KEY DATA SCIENCE CONCEPTS AND PERSONAS

Data scientists use machine learning technology to identify patterns, relationships, correlations, outcomes, and inferences in their data. These data-driven discoveries are incorporated into models that can detect fraud, predict maintenance cycles, mitigate customer churn, forecast sales, and automate many other forward-looking tasks. The key roles and personas in this process include the following:

- **Data scientists** build models and train them with data. They use notebooks such as Jupyter and Zeppelin and languages like R, Python, Java, and Scala.
- **Data analysts/citizen data scientists** use these models to conduct predictive and prescriptive analytics for business decisions, based on their working understanding of machine learning.
- **Data engineers** prepare data and establish automated data pipelines that feed ML models on a continuous basis.

THE ROLE OF MACHINE LEARNING IN DATA SCIENCE

Machine learning deals mainly with the data modeling aspect of the much broader data science discipline that encompasses data preparation, data discovery, analytics, and data modeling. Today's ML and data science tools can handle many aspects of parsing data, generating predictive and prescriptive models, placing models into production, and maintaining those models over time. Predictive and prescriptive analytics apps can often make their own decisions without human intervention, such as monitoring web browsing patterns to recommend products and services to visitors.

Data scientists use analytics tools to formulate a hypothesis, and then use programming languages and ML libraries to create their predictions. Types of ML include linear regression, logistic regression, classification, decision trees, deep learning, and many others. Some popular ML libraries include XGBoost, TensorFlow, scikit-learn, and PyTorch.

While data scientists are responsible for creating and training models that can make reliable predictions, data engineers are responsible for data pipelines that feed ML models with data needed for inference. The results of these ML/AI processes are made available to business users for data-driven decision-making.

THE MACHINE LEARNING PROCESS

Successful machine learning initiatives depend on getting the right data at the right time to the correct models. That's not always easy, since most machine learning cycles consist of multiple stages, from discovery and development into production. Data is added and prepared multiple times during each stage of the ML cycle, often with different data requirements. Success in ML is predicated on getting the right data in the right condition into the right analytic platforms to generate business results.

As shown in Figure 2, data scientists begin by finding, collecting, understanding, and preparing the data (steps 1-3). They may use business intelligence tools to better understand the data and formulate a hypothesis. Data scientists experiment with many data sets throughout this iterative process. Whenever they broaden or extend the scope of the data set, they have to wait for data engineers to load and prepare the data. This causes a delay and introduces significant latency between iterations. They also must "shape" the data into a normalized form, and many algorithms require nuanced formats.

Next they run training data through the models prepared in earlier steps (4), evaluating the outcomes to determine each model's effectiveness, then further tune the models through cycles of feature engineering (3) and hyper parameter tuning (4).

The resulting trained models are then deployed to production (5) to empower business users with predictive and prescriptive tools. Once deployed into production, the models receive ongoing evaluation (6) to identify model-drift and determine whether the models are out of date. Models must be periodically retrained with fresh training data. These model updates represent yet another iteration of the ML cycle, and require processing more data which is time consuming and prone to errors. Depending on the use case, the models may need to be retrained as frequently as every few hours, days, or weeks.

ML workflow

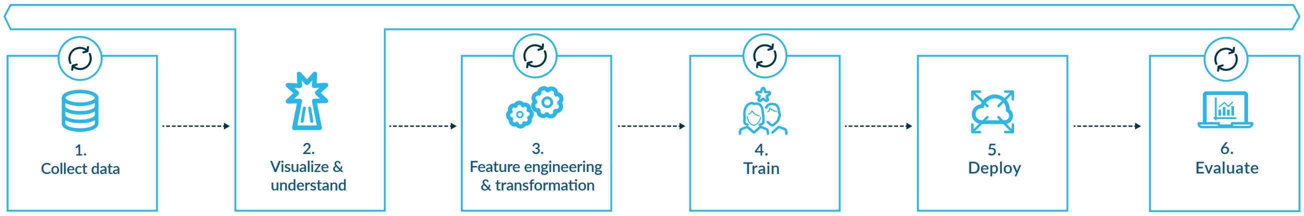

Figure 2: Data drives the ML process, from collection and preparation through training, predictions, and productization.

THE ROLE OF THE DATA CLOUD

Machine learning is a data-intensive activity, and the success of each predictive model depends on large volumes of diverse data that must be collected, persisted, transformed, and presented in many different ways. This involves large volumes of data characterized by many dimensions and details, and arising from many contexts. For example, if you have built a machine learning model to predict customer churn, you will likely have data about customer behaviors relating to sales, service, purchasing, and app interactions, both historic and real-time.

Snowflake's Data Cloud allows you to consolidate your data from data warehouses, data marts, and data lakes into a single source of truth that powers multiple types of analytics and data science applications. It makes it easy for diverse teams to share governed data, internally and externally, by allowing team members to collaborate without having to copy data and move it from place to place. Raw, structured, and semi-structured data is easily discoverable and immediately accessible for data science workflows, with native support for JSON, AVRO, XML, ORC, and Parquet.

Being able to use one set of tools to manage both structured and semi-structured data shortens the data discovery and preparation cycle. Furthermore, the data that is output from the ML algorithms is placed back into the repository for access by business users, alongside the source data. This means all data is always up to date and consistently maintained for business users, analysts, and data scientists.

Benefits of the Data Cloud

TDWI recommends acquiring a modern data platform, built for the cloud, that can satisfy the entire data life cycle of machine learning, artificial intelligence, and predictive application development. What should you look for in such a platform? For data preparation you want to be able to work with large data sets with interactive response times. For training, you want to plow through those data sets iteratively. For production, you want a reliable, repeatable, and scalable data pipeline.

Here are a few of the primary reasons to use Snowflake for your data science endeavors:

- **Simplicity:** No need to manage multiple compute platforms and constantly maintain integrations.
- **Security:** One copy of data is stored securely in the Snowflake environment, with user credentials carefully managed and all transmissions encrypted.
- **Performance:** Query results are cached and can be used repeatedly during the ML process, as well as for analytics.
- **Workload isolation:** Each user and workload can receive dedicated compute resources.
- **Elasticity:** It only takes seconds to scale-up capacity to accommodate large data processing tasks and then it's just as easy to release it once completed, minimizing costs with pay per second pricing.
- **Support for structured and semistructured data:** Easily load, integrate, and analyze all types of data inside a unified repository.
- **Concurrency:** Run massively concurrent workloads at scale across shared data.

CONSOLIDATING DATA FOR MACHINE LEARNING AND ANALYTICS

There are many ways to provision data for machine learning applications, and flexibility is essential. For example, some organizations use a data lake in conjunction with a data warehouse. This allows them to store vast amounts of raw data in its native form, which can then be repurposed for a wide range of analytics when it's needed. Most data science tools rely on data lakes as their source of data, but today's analytic strategies increasingly use multiplatform data architectures that include a mix of big data platforms, clouds, data lakes, and data warehouses. Many leading organizations choose to skip the data lake altogether and instead consolidate their data entirely into a cloud data platform. This approach eliminates the complexity of managing a separate data lake, and it also removes the need for a data transformation pipeline between the data lake and the data warehouse. Having a unified repository, based on a versatile cloud data platform, allows them to select the appropriate storage, processing, and economics for each data set and workload, optimizing the options for ML and analytics.

Once you have collected and prepared your data, you need to be able to discover patterns and insights via analytics and predictive analytics tools. Snowflake allows you to combine general analytics with predictive analytics, so your business intelligence tools and data science tools have one consistent view of the same governed data. All data science tools reference the same data definitions, so you can consistently reproduce the content of queries, forecasts, dashboards, and reports. Both the raw data and ML results reside in the data platform for easy access. This unified approach allows data scientists to output the results of machine learning activities back into the data platform for general-purpose analytics, as well as to embed these results in the decision-making process.

Having common semantics, data definitions, and data models keeps everybody on the same page. For example, a sales manager might look at a BI report that shows the historic performance of the sales team. An ML model could also forecast expected sales results for upcoming quarters based on target account propensity, and highlight both booked and forecasted revenue through the same report.

AUTOMATED DATA ENGINEERING, DATA INTEGRATION, AND DATA SHAPING

Achieving success with ML means creating efficient and reliable data pipelines that feed accurate and timely data to business users, as well as populate the apps and services they use.

Ingesting data

Snowflake includes a serverless ingestion service called Snowpipe that asynchronously loads data and makes it available immediately. Manual data flattening tasks are fully automated: The platform transforms data into the type and shape required for each target table.

Standard connectors and adapters allow you to easily ingest event streams from Kafka and other messaging systems, while Snowflake streams and tasks make it easy to schedule data loads for SQL jobs.

By "productizing" the ML model with an automated data ingestion service, the pipeline simplifies complex data integration tasks. Data scientists can find and prepare data on demand—without waiting days or hours between tests. Once an automated data pipeline service is put into production, raw data is immediately available without requiring ETL from a data lake. As data comes in, it is automatically run through the model to make predictions. And because it's all based in the cloud, data scientists can use dedicated virtual warehouse compute resources without impacting other users.

Universal SQL capabilities

Snowflake customers can leverage a central source of truth with universal SQL capabilities that power robust and efficient ETL and ELT workloads. Data transformations using SQL are faster, easier, and less expensive than the same operations using Spark. Because data can be transformed as part of a SQL query, the transformation becomes part of the analysis. Due to Snowflake's architecture and compression, you can rapidly ingest large amounts of streaming data and store it indefinitely at nominal cost. Data engineers can utilize many types of integration tools including Alteryx, Alooma, Matillion, Fivetran, Alation, Informatica, and many others.

Dedicated compute resources

With Snowflake, ML data ingestion, data management, and data preparation workloads receive dedicated resources that don't contend with non-ML data engineering and analytics workloads. By removing contention for resources, live data can be ingested and transformed in stream and immediately made available for analytics. You can customize the size of your data warehouse for each workload, scale up as needed, and turn off the cloud services upon completion. Thanks to linear scaling, you can request the exact amount of resources you need to execute queries in a predictable timeframe. With instant elasticity and per-second billing, each user and workgroup pays only for the precise compute resources they use. Ultimately, this architecture allows you to maximize the performance and efficiency of each team, while providing consistent data.

Robust data security

Many legacy data science projects depend on Apache Hadoop, an open-source framework for storing and processing data in a distributed framework. However, the Hadoop architecture employs only rudimentary access controls, and it was not designed to comply with important industry standards governing the security and privacy of data, including HIPAA, PCI DSS, and GDPR. Other data science projects leverage general-purpose object stores such as Amazon S3, which also lack robust data security.

By contrast, Snowflake's Data Cloud is built on a multilayered security foundation that includes encryption, access control, network monitoring, and physical security measures, in conjunction with comprehensive monitoring, alerts, and cybersecurity practices. In addition to industry-standard technology certifications such as ISO/IEC 27001 and SOC 1/ SOC 2 Type 2, Snowflake complies with important government and industry regulations such as PCI DSS, HIPAA/Health Information Trust Alliance (HITRUST), and FedRAMP certifications. Snowflake customers can securely access data for all types of data science activities.

EFFICIENT DATA SHARING

The Snowflake platform simplifies the exchange of data between partners, suppliers, vendors and customers through Snowflake Data Marketplace and Data Exchange. This offers access to unique data sets that can increase the effectiveness of models and provide additional feature engineering possibilities. Secure data sharing in Snowflake doesn't require data transfer via FTP or the configuration of APIs to link applications. It simplifies ETL integration and automatically synchronizes "live" data among data providers and data consumers. Because the source data is shared rather than copied, consumers don't require any additional cloud storage. Snowflake Data Marketplace and Data Exchange enable data scientists to easily collaborate on models by sharing raw and processed data.

Snowflake Data Marketplace and Data Exchange

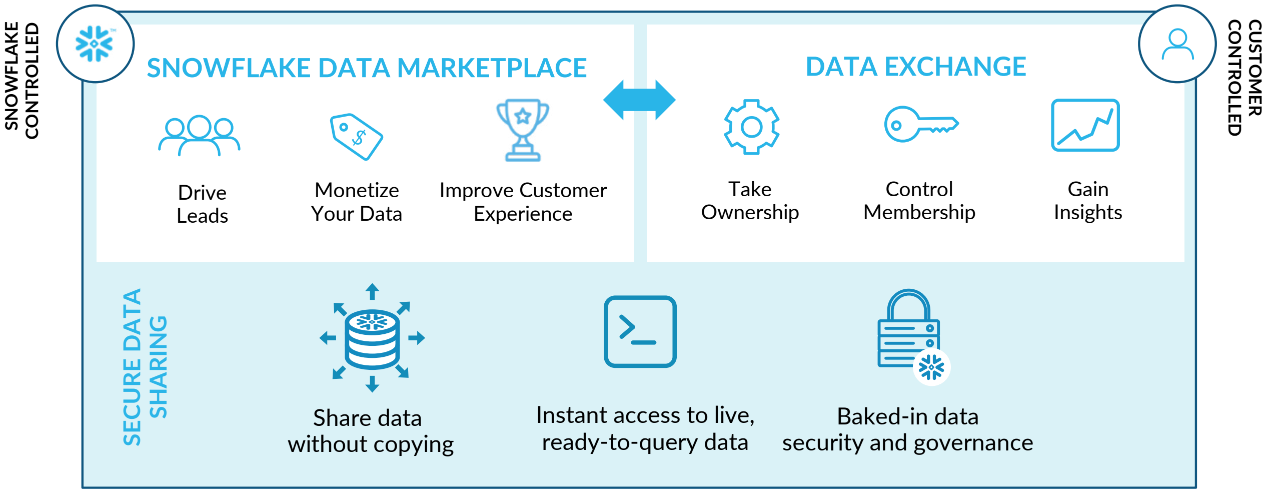
SNOWFLAKE CONTROLLED
SNOWFLAKE DATA MARKETPLACE
Drive Leads
Monetize Your Data
Improve Customer Experience
DATA EXCHANGE
Take Ownership
Control Membership
Gain Insights
CUSTOMER CONTROLLED
SECURE DATA SHARING
Share data without copying
Instant access to live, ready-to-query data
Baked-in data security and governance

Figure 3: Snowflake Secure Data Sharing enables you to share data externally via Snowflake Data Markeplace, and create your own Data Exchange with customers, suppliers, and other business partners.

EXTENSIVE PARTNER ECOSYSTEM

The machine learning space is rapidly evolving, with new tools being added each year. Through Snowflake's extensive partner ecosystem, customers can take advantage of direct connections to all existing and emerging data science tools, platforms, and languages such as Python, R, Java, and Scala; open-source libraries such as PyTorch, XGBoost, TensorFlow, scikit-learn; notebooks like Jupyter and Zeppelin; and platforms such as Data Robot, Dataiku, H20.ai, Amazon Sagemaker, and others. By offering a single consistent repository for data, Snowflake removes the need to retool the underlying data every time you switch tools, languages, or libraries. Furthermore, the output from these activities is easily fed into Snowflake and made accessible by nontechnical users to generate business value.

Notebook-based ML tools

Traditional ML notebooks such as Jupyter and Zeppelin power today's leading data science tools, including Amazon Sagemaker, Dataiku, Zepl, and many others. This approach allows data scientists to have ultimate control over the frameworks and algorithms they choose, conduct in-depth feature engineering, tune hyperparameters, and iteratively create, assess, and productize ML models. They can turn intuitions into accurate predictions by iteratively experimenting with algorithms, scoring their performance, and choosing and refining new models. Users of Amazon SageMaker can use the Snowflake Python Connector to directly populate Pandas DataFrames. This high speed connection results in accelerated training speed as well as an optimized data preparation and feature engineering cycle that leverages the full power of ANSI SQL.

AutoML tools

Alternatively, AutoML tools such as RapidMiner, BigSquid, H2o.ai, and DataRobot can automatically select algorithms, conduct model training, and choose the best model. These tools are a great way to democratize access to advanced analytics, enabling data analysts to perform ML functions without requiring advanced programming skills or deep mathematics/statistics knowledge. A few tools bridge these two approaches, allowing data scientists to customize AutoML processes. DataRobot, a leading player in the AutoML space, has a built-in Snowflake integration where its users can quickly connect their DataRobot account to Snowflake and use it as a data store.

Analytics and cloud partners

Regardless of which ML approach you choose, Snowflake allows you to consume the results via dashboards, reports, and business analytics tools by leveraging connections to other ecosystem partners such as Tableau, Looker, ThoughtSpot, and Sigma. Furthermore, Snowflake allows you to store and replicate your data across any region, on any cloud, including popular offerings from Amazon, Microsoft, and Google. Snowflake can seamlessly export data to external tables maintained in Amazon S3, Azure Blob, and Google Cloud Storage for universal access by any tool. For example, you can use Snowflake to complement your data lake on AWS, then connect with Amazon SageMaker to develop, test, and deploy ML models at scale. The platform automates everything, from data storage and processing to transaction management, security, governance, and metadata management.

Get started in minutes

If you're looking for the fastest way to get started with Snowflake and ML, consider the Snowflake Partner Connect program, which simplifies deployment through pre-configured integrations with select technology partners. You can automatically provision and configure partner applications in minutes and load data into Snowflake for immediate use.

CASE IN POINT

ConsumerTrack is a digital advertiser and publisher that aggregates and syndicates website performance data from hundreds of providers to portals such as CNN and MSN. Previously its data science team struggled with an ML environment that used MySQL and various orchestration tools, which led to data chokepoints and latency issues.

ConsumerTrack augmented its existing data lake with Snowflake and chose Amazon SageMaker as it's fully managed service for automating the ML workflow. It labels and prepares data, chooses an algorithm, trains a model, tunes and optimizes the model for deployment, makes predictions, and then takes action.

Now data flows into the data lake via an automated pipeline that uses AWS Lambda and AWS Glue. Data is curated and then loaded into Snowflake, and the data streams are configured with custom alerts. Amazon SageMaker connects to Snowflake to simplify the development, testing, and building of ML models.

ConsumerTrack has eliminated chokepoints and reduced time-to-insight from hours to minutes. Snowflake substantially reduces the amount of time spent on data discovery and preparation. Snowflake's broad ecosystem allows ConsumerTrack to connect with many types of data science platforms and tools, including a native connector for Python. When they need to, the data science team can export data to any blob store for universal access.

WHAT'S NEXT?

To learn more about for machine learning, visit the Snowflake data science page and the Snowflake platform page.

Featured Data Science Partners

DataRobot

Amazon SageMaker

dataiku

zepl

H2O.ai

bigsquid

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] "Best Practices Report: Driving Digital Transformation Using AI and Machine Learning" (tdwi.org/bpreports).

[2] Forbes "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says" (bit.ly/38EbXmN).

[3] Forbes "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says" (bit.ly/38EbXmN).