snowflake®

MACHINE LEARNING WITHIN REACH

Activate your data lake for ML and AI with Snowflake and Amazon SageMaker

aws

CRACKING THE MACHINE LEARNING (ML) CODE

You might be eager to take advantage of machine learning (ML) technologies to bring predictive power to your business and analyze data on a deeper level. With new cloud-based ML technologies, you can:

- Innovate and gain a competitive advantage through automated data analysis.
- Provide your IT teams with the latest technologies they need to build new solutions.
- Give your developers the tools to optimize their code and focus on the strategic work of building ML models.

The issue is where to start. You might think that it takes too long and it costs too much. Until recently that has been the case.

With Snowflake—the data warehouse built for the cloud—and Amazon Web Services (AWS), you can quickly start with the data you already have and take advantage of cost-effective, pay-as-you-go pricing. It's easy to get started, doesn't require a team of data scientists and can quickly show value.

This eBook describes how 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. Also read how ConsumerTrack is using Snowflake and AWS to quickly implement a cloud-based data lake and ML solution.

This eBook describes how 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.

BIG DATA AND ML CHALLENGES

With big data and ML, you need a "single source of truth." Synchronized data translates to speed and the ability to process data on the go. Consider computer security—it's not enough to use yesterday's data to predict tomorrow. You need to process while you are deriving insights and dynamically change models.

Performing analytics across structured and semi-structured data is key as the mobile devices, websites, and IoT devices that generate data become a larger part of both personal and corporate life.

Implementing on-premises ML properly, at scale, is a big and costly undertaking. Beyond finding data science resources, and building machines and data storage, you must configure for unique ML system requirements and dependencies. Other challenges include:

- Dispersed data gets replicated and out of synchronization.
- Data preparation takes too long. It's estimated that data science teams spend 80% of their time preparing data for analysis.[1]
- In ML, "data writes the code." It's critically important to analyze the right data and to prepare it well.
- Scalability must be managed to distribute and commercialize models throughout an organization.
- Adding data to existing data sets and organizational issues must be managed to avoid incorrect insights.
- It's hard to move massive amounts of varied data types, and different results can occur during synchronization.
- Developers must interact with multiple systems and programming languages, keeping them from focusing on model development and production.

It's critically important to analyze the right data and to prepare it well.

[1] Gartner, *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says* (March 2016)

THE POWER TO INNOVATE AND SCALE

Snowflake, the data warehouse built for the cloud, complements ML technologies like Amazon SageMaker. SageMaker supports the entire ML workflow. It's a fully managed service that labels and prepares data, chooses an algorithm, trains a model, tunes and optimizes it for deployment, makes predictions, and then takes action. Your developers get models to production faster with much less effort, less process, and lower cost.

Snowflake stores data in one central, enterprise-grade, scalable repository, providing a single source of truth and synchronized data. The same tools manage structured and semi-structured data, shortening the data preparation cycle. Data is modified transactionally, which creates data consistency and reduces risk. Transactional data modification transforms data sets using parallelization. All users get the same query results and jobs run efficiently.

Because Snowflake utilizes AWS, you can build infrastructure when and where you need it. Snowflake's elasticity feature allows you to scale as you go—horizontally as well as vertically—all with the push of a button. Data sharing and exchange capabilities let you bring in complementary third-party data to drive better insights and to enhance the value of your own data.

Because Snowflake utilizes AWS, you can build infrastructure when and where you need it.

CASE STUDY: CONSUMERTRACK

ConsumerTrack is a digital advertiser and publisher, which aggregates and syndicates website performance data from hundreds of providers to portals such as CNN and MSN.

Customer challenge:

ConsumerTrack was operating a ML environment using MySQL, DBA, ETL, and orchestration tools that were laborious, caused data chokepoints and latency, and didn't easily scale.

Solution:

ConsumerTrack implemented a data lake on AWS with Snowflake and Amazon SageMaker. Data flows into the data lake, using AWS Lambda and AWS Glue for ETL (extract, transform, and load). The data streams are configured with custom alert settings and methods for anomalies. Data is curated and then loaded into Snowflake. Amazon SageMaker then connects to Snowflake for developing, testing, and building the ML models.

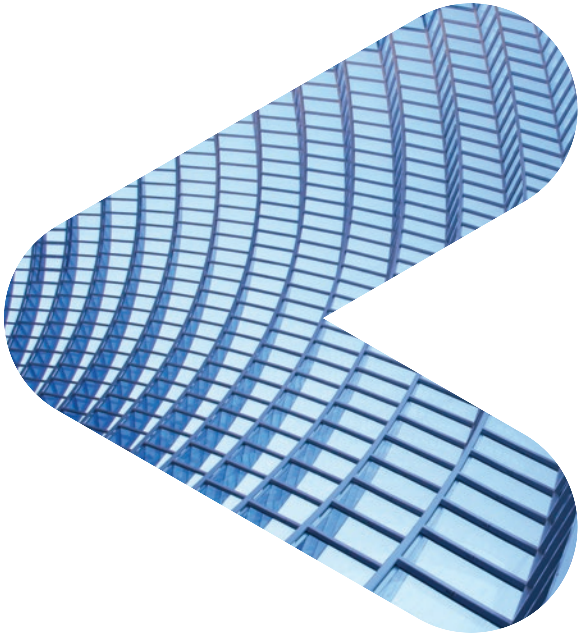
CASE STUDY: CONSUMERTRACK *(CONTINUED)*

Outcomes delivered:

Snowflake and Amazon SageMaker enabled ConsumerTrack to scale data operations with its business. An Amazon SageMaker model was implemented out of the box that lets ConsumerTrack drill down into new levels of dimensions that can detect anomalies based on data such the type of device and zip code.

Data science is delivered as a service, chokepoints are eliminated, and raw data is exposed as it's being acquired. Small changes are visible in the aggregate data set, providing a proactive way to ensure that data doesn't fall through the cracks. With Snowflake and Amazon SageMaker, waiting for access to insight is reduced from hours to minutes, and performance matches the growth in cluster sizes.

MODERN ML WITH SNOWFLAKE AND AMAZON SAGEMAKER



With Snowflake, you can quickly create cost-effective data lakes that leverage Amazon SageMaker and Amazon S3. It gives you direct vision into a single data lake, which is easier to manage, and track spend.

You get the peace of mind from being able to easily see who is using your data lake and how they are using it. You can see the predictability of models that drive customer outcomes, helping your teams more easily determine value from investments and directing focus to your most pressing business initiatives and challenges.

A data lake on AWS and Snowflake enables you to build infrastructure when and where you need it and at any scale. You provide quick and efficient access to unified data that is always analyzed in the same repository—no data subsets required—and you empower all levels of users with data analytics.

Your small teams can explore data as they wish, and you only pay for what you use. Snowflake also connects to Amazon EMR, a hosted big data platform that allows you to process vast amounts of data quickly and at scale, using open source tools including Apache Spark and Apache Hive.

Snowflake leverages native AWS AI components that enhance forecasting, personalization, and interactive dashboards.

A data lake on AWS and Snowflake enables you to build infrastructure when and where you need it and at any scale.

LEARN MORE ABOUT SNOWFLAKE

TRY SNOWFLAKE

October, 2019