

Integration of H2O Driverless AI with Snowflake



With the integration between H2O.ai and Snowflake, data engineers and dataOps users can score and re-train predictive models with SQL commands from within Snowflake, reducing costs and management resources to run the end-to-end automation of ML pipelines.

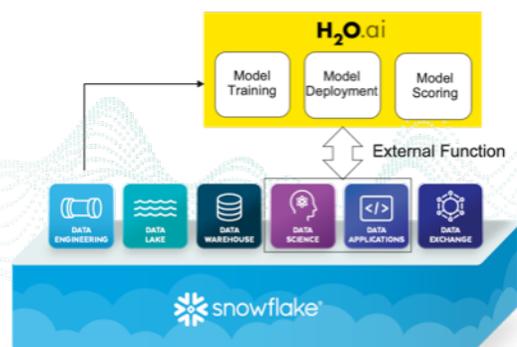
Challenges for Production AI

Companies have invested in data management and AI platforms to create value from unique data and domain-specific insights. However, separate systems for data management, data preparation, and data science can make production operations difficult for machine learning projects. Actions such as data wrangling, training models, scoring records using a model, refitting the model on new data and retraining a model happen outside the data platform leaving data engineers using multiple tools, writing brittle integration code, and without consistent logging across systems for governance.

H2O.ai Integration

An operations manager or data engineer in charge of machine learning models in production using Snowflake can now manage the end-to-end ml pipeline using SQL in Snowflake. Snowflake users can conduct scoring jobs, refit models on newer data, and rebuild models. Snowflake data engineers can prepare and clean the data for machine learning using SQL. Once the data is prepared then a data scientist can also build new models using H2O Driverless AI based on Snowflake data. These capabilities are enabled using external function calls from Snowflake to H2O

Driverless AI and Driverless AI scoring pipeline(MOJO). This seamless integration keeps operations engineers in Snowflake and lets them use SQL to manage machine learning using the interface they already know.



Google Cloud aws Azure

Figure 1: H2O.ai integrates with Snowflake

Batch Scoring Scenarios

Existing batch scoring methods are slow, error-prone, and resource-intensive for operations teams to manage. Using a traditional database approach, users would connect to the database using JDBC or some language-dependent component, issue SQL to query the table, and then score

the data and write back the results. Each step is typically a separate process. When breakdowns occur in a process, the entire flow must be restarted for the job to complete successfully.

With the H2O.ai integration to Snowflake, the predictive model acts like it is embedded in the SQL logic, and executes from within the Snowflake process. This integration does not require the batch extract and loading steps. In tests, Snowflake and Driverless AI scoring was 3X faster than traditional database scoring methods. With the external function, Snowflake schedules the table to update at a specific time using Snowflake Tasks. Snowflake users can also issue a SQL Update command on a table to get predictions. Either option makes it a simple database operation rather than requiring DevOps setup of the batch scheduler, which saves on setup time and management resources.

The integration also helps in situations where scoring jobs run periodically based on new records creating an active database that is triggered to score new data as it arrives. With a traditional approach, a batch job runs every few minutes or hours to create a new view in the database and then starts the scoring process. This process runs even if there are no further records to be scored. Snowflake can look at the existing table and determine if changes happened and then invoke the H2O.ai model via the external function. Snowflake reduces the operational complexity of creating a view for each of the tables and reduces the need to have a job run to look for changes, which means less monitoring is needed to ensure the batch job is running correctly. Overall, this process is more streamlined to set up and requires fewer resources to manage, saving time and money.

Application Scoring Scenarios

Commercial and customer written applications that use Snowflake as the system of record can now automatically leverage predictions within applications with external function integration. For example, a customer service application that uses Snowflake to retrieve customer's details can also retrieve an H2O.ai model score on the likelihood of churn at the same time, just by changing the SQL call when retrieving the customer details from Snowflake.

Model Training and Refitting

One significant difference between machine learning models and traditional software applications is that models can decay over time as data in production environments deviates from the training data. This decay can reduce prediction accuracy and create issues for downstream applications. Typically, IT operators running ML pipelines will submit a request to the data science team to retrain or refit the newer data model. This process can take days or weeks, during which the model's results continue to decay.

With the external function integration between Snowflake and H2O Driverless AI, IT operators can now rebuild models using simple commands from within Snowflake. Driverless AI automatic machine learning (AutoML) will retrain the model or refit it based on the specified configuration. The new model object is then tested and moved into production as part of the typical operations process. Removing the need to go to other systems or even teams to retrain models, saves the operations team valuable time, and keeps high performing models running in production.

About H2O.ai

H2O.ai is an open source leader in AI with a mission is to democratize AI for everyone. H2O.ai is transforming the use of AI with software with its category-creating visionary open source machine learning platform, H2O. More than 18,000 companies use open-source H2O in mission-critical use cases for Finance, Insurance, Healthcare, Retail, Telco, Sales and Marketing. H2O Driverless AI uses AI to do AI in order to provide an easier, faster and cheaper means of implementing data science. H2O.ai partners with leading technology companies such as NVIDIA, IBM, Intel, AWS, Microsoft Azure and Google Cloud Platform and is proud of its growing customer base which includes Capital One, Progressive Insurance, Comcast, Walgreens and MarketAxess. For more information about how H2O.ai is driving an AI transformation, visit www.h2o.ai