

Intégration de H2O Driverless AI avec Snowflake



Grâce à l'intégration entre H2O.ai et Snowflake, les ingénieurs de données et les utilisateurs de dataOps peuvent évaluer et réentraîner des modèles prédictifs avec des commandes SQL depuis Snowflake, réduisant ainsi les coûts et les ressources de gestion pour exécuter l'automatisation de bout en bout des pipelines ML.

Les Défis de l'IA en Production

Les entreprises ont investi dans la gestion des données et les plates-formes d'IA pour créer de la valeur à partir de données uniques et d'informations spécifiques au domaine. Cependant, des systèmes séparés pour la gestion des données, la préparation des données et la science des données peuvent rendre les opérations de production difficiles pour les projets d'apprentissage automatique. Des actions telles que la gestion des données, la formation de modèles, la notation des enregistrements à l'aide d'un modèle, le réaménagement du modèle sur de nouvelles données et le recyclage d'un modèle se produisent en dehors de la plate-forme de données, laissant les ingénieurs de données utiliser plusieurs outils, écrivant un code d'intégration fragile et sans journalisation cohérente entre les systèmes pour la gouvernance.

Intégration H2O.ai

Un responsable des opérations ou un ingénieur de données en charge des modèles d'apprentissage automatique en production avec Snowflake peut désormais gérer le pipeline ML de bout en bout à l'aide de SQL dans Snowflake. Les utilisateurs de Snowflake peuvent effectuer des tâches de scoring, refaire des modèles sur des données plus récentes et reconstruire des modèles.

Les ingénieurs de données Snowflake peuvent préparer et nettoyer les données pour l'apprentissage automatique à l'aide de SQL. Une fois les données préparées, un data scientist peut également créer de nouveaux modèles à l'aide de H2O Driverless AI en se basant sur les données Snowflake. Ces capacités sont activées à l'aide d'appels de fonctions externes de Snowflake à H2O Driverless AI et Driverless AI scoring pipeline (MOJO). Cette intégration transparente permet aux ingénieurs d'exploitation de rester dans Snowflake et leur permet d'utiliser SQL pour gérer l'apprentissage automatique à l'aide de l'interface qu'ils connaissent déjà.

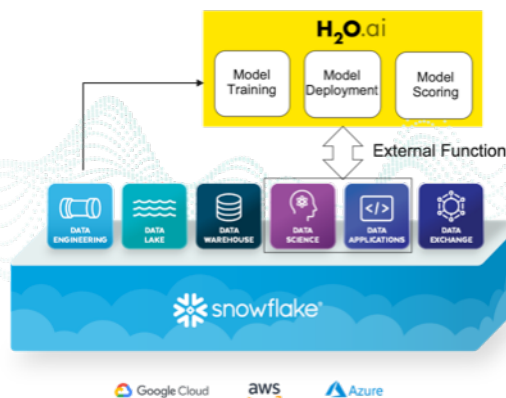


Figure 1: H2O.ai s'intègre avec Snowflake

Scénarios de notation par lots

Les méthodes de notation par lots existantes sont lentes, sujettes aux erreurs et gourmandes en ressources pour les équipes opérationnelles. En utilisant une approche de base de données traditionnelle, les utilisateurs se connecteraient à la base de données à l'aide de JDBC ou d'un composant dépendant de la langue, émettraient du SQL pour interroger la table, puis noteraient les données et réécriraient les résultats. Chaque étape est généralement un processus distinct. Lorsque des pannes se produisent dans un processus, l'ensemble du flux doit être redémarré pour que le travail se termine avec succès. Avec l'intégration H2O.ai à Snowflake, le modèle prédictif agit comme s'il était intégré dans la logique SQL et s'exécute à partir du processus Snowflake. Cette intégration ne nécessite pas les étapes d'extraction par lots et de chargement. Lors des tests, la notation de Snowflake et de Driverless AI était 3 fois plus rapide que les méthodes traditionnelles de notation de base de données. Avec la fonction externe, Snowflake planifie la mise à jour de la table à un moment précis à l'aide des tâches Snowflake. Les utilisateurs de Snowflake peuvent également émettre une commande SQL Update sur une table pour obtenir des prédictions. Les deux options offrent une opération de base de données simple plutôt que de nécessiter la configuration DevOps du planificateur de lots, ce qui économise du temps de configuration et des ressources de gestion. L'intégration est également utile dans les situations où les tâches de scoring s'exécutent périodiquement en fonction de nouveaux enregistrements créant une base de données active qui est déclenchée pour évaluer les nouvelles données à leur arrivée. Avec une approche traditionnelle, un travail par lots s'exécute toutes les quelques minutes ou heures pour créer une nouvelle vue dans la base de données, puis démarre le processus de notation. Ce processus s'exécute même s'il n'y a plus d'enregistrement à noter. Snowflake peut consulter la table existante et déterminer si des changements se sont produits, puis appeler le modèle H2O.ai via la fonction externe. Snowflake réduit la complexité opérationnelle de la création d'une vue pour chacune des tables et réduit la nécessité d'exécuter une tâche pour rechercher les modifications, ce qui signifie que moins de surveillance est nécessaire pour garantir que la tâche par lots s'exécute correctement.

Dans l'ensemble, ce processus est plus simple à mettre en place et nécessite moins de ressources à gérer, ce qui permet d'économiser du temps et de l'argent.

Scénarios de notation des applications

Les applications commerciales et écrites par les clients qui utilisent Snowflake comme système d'enregistrement peuvent désormais exploiter automatiquement les prédictions dans les applications avec intégration de fonctions externes. Par exemple, une application de service client qui utilise Snowflake pour récupérer les détails du client peut également récupérer un score de modèle H2O.ai sur la probabilité de désabonnement en même temps, simplement en modifiant l'appel SQL lors de la récupération des détails du client à partir de Snowflake.

Entraînement et réaménagement des modèles

Une différence significative entre les modèles d'apprentissage automatique et les applications logicielles traditionnelles est que les modèles peuvent se détériorer au fil du temps lorsque les données dans les environnements de production s'écartent des données d'entraînement. Cette dégradation peut réduire la précision des prévisions et créer des problèmes pour les applications en aval. En règle générale, les opérateurs informatiques exécutant des pipelines ML soumettent une demande à l'équipe de science des données pour recycler ou réaménager le nouveau modèle de données. Ce processus peut prendre des jours ou des semaines, pendant lesquels les résultats du modèle continuent de se détériorer. Grâce à l'intégration de fonctions externes entre Snowflake et H2O Driverless AI, les opérateurs informatiques peuvent désormais reconstruire des modèles à l'aide de commandes simples depuis Snowflake. L'apprentissage automatique de Driverless AI (AutoML) recyclera le modèle ou le remodelera en fonction de la configuration spécifiée. Le nouvel objet modèle est ensuite testé et mis en production dans le cadre du processus opérationnel typique. Le fait de ne plus avoir à recourir à d'autres systèmes ou même à des équipes pour recycler les modèles, fait gagner un temps précieux à l'équipe d'exploitation et permet aux modèles hautement performants de fonctionner en 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