SNOWFLAKE DATA SCIENCE WORKLOADS WITH SNOWPARK PYTHON AND DATAOPS.LIVE

A #TrueDataOps white paper

CONTENTS

About DataOps.live

DataOps.live—the Data Products company, delivers productivity breakthroughs for data teams by enabling agile DevOps automation **(#TrueDataOps)** and a powerful Developer Experience (DX) to modern data platforms.

The DataOps.live SaaS platform brings automation, orchestration, continuous testing and unified observability to deliver the Data Products you want at the speed the business needs. DataOps.live is a global company funded by Snowflake Ventures, Notion Capital, and Anthos Capital, with enterprise clients including Snowflake, Roche Diagnostics and OneWeb.

For more information, visit **www.dataops.live** or connect with the team on **LinkedIn** or **Twitter**.

Snowpark is a powerful tool, and it does bring even more value when it is connected with operational best practices that support the deployment and monitoring of Snowpark code.

With Snowflake acknowledged for its performance, scalability, and concurrency, Snowpark provides an exciting way to program in this environment: promoting code to data in Snowflake Python. Snowpark also supports Java and Scala.

Before Snowpark, users interacted with Snowflake predominately through SQL. Now, data engineers and data scientists can execute more workflows entirely within the Snowflake Data Cloud without managing additional processing systems.

DataOps.live has supported Snowpark since its initial release, and enables you to leverage all of Snowpark's capabilities for maximum productivity.

The DataOps.live platform enables Snowpark source code to be fully managed and lifecycled through development, test and production environments alongside all other objects in Snowflake

AN EXCITING NEW WAY TO PROGRAM

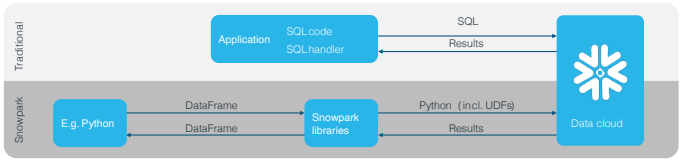
Snowpark takes advantage of custom warehouses and allows for familiar languages such as Python, Java, Scala, and the SQL interface from Snowflake. In practice, it offers an exciting option for data engineers and architects to design solutions without moving data out of Snowflake.

DataOps.live supports Snowpark implementations for Java, Scala, and our preferred environment, Python. It manages your code's development, testing, deployment, and promotion that would otherwise need to be performed manually to get a Snowpark implementation off to the races.

HOW SNOWPARK WORKS: DATAFRAMES

In conjunction with Java UDFs and other innovations, Snowflake became the first fully programmable data cloud in history. In this environment, Snowpark enables users comfortable with popular languages to write code using a widely used and familiar DataFrame model.

Snowpark libraries allow you to use standard DataFrame paradigms in code (e.g. Python, Scala). When these are executed, the Snowpark libraries transparently compile these to a combination of SQL and Python, execute them, then take the results and convert them back to DataFrames. This approach has been game-changing for the Snowflake developer experience.



DataOps.live enables the Python, Scala, or Java code to be stored and fully lifecycle managed depending on the environment context it is run against, e.g. development, test, or production.

MAKING A GAME-CHANGING EXPERIENCE EVEN BETTER

Increasing volumes of data are coming into enrichment platforms from an ever-growing variety of sources (streaming, third parties, IOT, etc.). Leveraging the speed, performance, and cost-savings of keeping the data in a specific location is a significant advantage: the time it takes to move data around the enterprise will only increase as data volumes continue to rise.

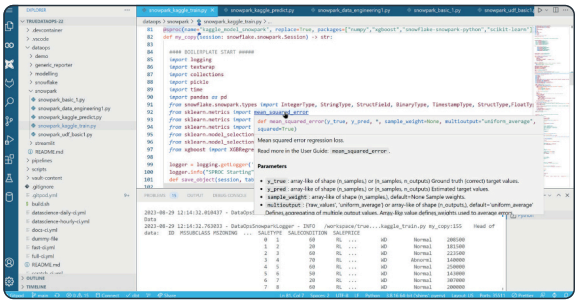
The DataOps.live platform helps you to leverage Snowpark further to deliver even greater benefits.

Data-Ops.live has built-in helper functions for, say, the Python data scientist to save models in easily retrieved locations for use within a Snowpark function. Its repository-based infrastructure ensures that configuration is stored as code, checked in, audited, and approved before releasing it to production—naturally supporting the incorporation of Snowpark code.

A DataOps.live orchestration can perform many functions, including data ingestion, transformation, automated testing, and more. To ensure everything meets high-quality standards, as required by the data product or application, DataOps.live can seamlessly orchestrate code running within Snowpark. This approach opens the door to many use cases when building data applications.

SNOWPARK DEVELOPMENT

As well as all the normal **source of truth** for 'everything data' stored in the DataOps' Git repository, we now have full software management requirements related to your Python code:



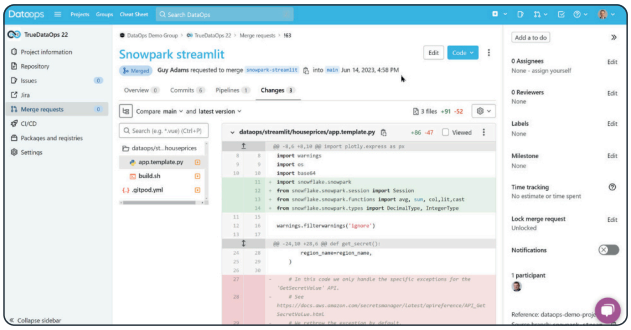
We need to **branch, version, compile, test and deploy** the software and produced artifacts just like any other software project.

As Snowflake continues its journey of becoming the programmable data cloud, the ability to lifecycle and manage software code as well as data objects is critical.
The programmable data cloud means that the future of DataOps is **Data + Software (code)**.

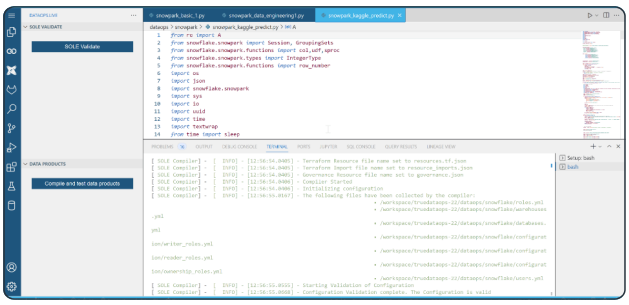
This fits perfectly with the #TrueDataOps philosophy, which advocates "starting with pure DevOps and Agile principles (which have been battle hardened over 20+ years) and determining where they don't meet the demands of Data and adapting accordingly".

The key requirements for developing Snowpark data science applications are:

- **Support for full code lifecycle, including the ability to work with code diffs, merge requests, rollbacks, and reverts, empowering you to work as a scaled data engineering and data science team and not just as an individual:**

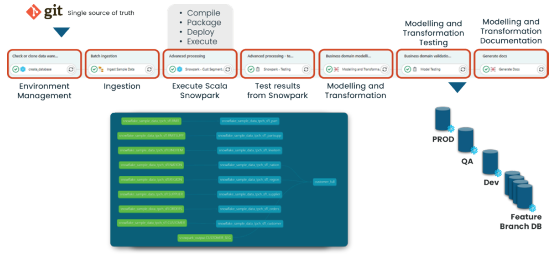


- A fully git-compatible DataOps Development Environment to work on your Snowpark data science workload
- An IDE to develop your surrounding Snowflake workloads, e.g. building analytical data applications powered by Streamlit.
- A built-in terminal environment that allows you to use your preferred set of command line tools to automate you workflows

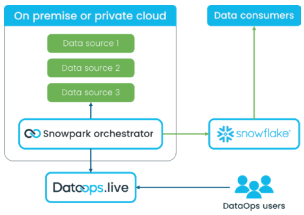


SNOWPARK EXECUTION

Running a Snowpark Application means running the application where the Snowpark libraries are being used. For example, if Python is being used, an environment is needed with all the runtime tools and libraries (plus the Snowpark libraries), e.g.

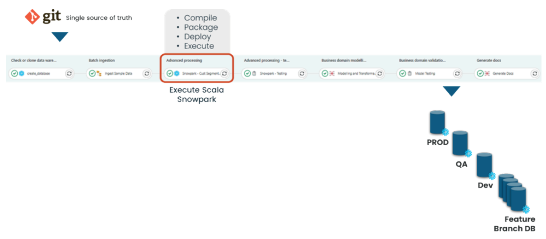


Fortunately, DataOps.live provides an execution environment (termed orchestrators) that can be run anywhere a customer needs, including in a private cloud or on-premises e.g.



In a real-life Snowpark scenario, many Snowpark applications may need executing at different points in a pipeline, interspersed with ingestion, transformation, testing, and data sharing —with all the correct dependencies modeled and tested. This orchestration of Snowpark applications is part of a complete end-to-end data pipeline.

The use of this Snowpark orchestrator abstracts all of the complexity and dependency management. It makes running a Snowpark application as part of a DataOps pipeline as simple as any other orchestration.



In many cases, a Snowpark application will be used to do advanced data manipulation, particularly manipulation beyond what can be achieved within SQL, but with results still stored back into Snowflake. In these cases, the automated data testing within the DataOps.live Modelling and Transformation Engine can be used to validate the results of the Snowpark application:

CONCLUSION:

Making Snowpark a walk in the park

With Snowpark, you can streamline your data processing needs and enhance collaboration across data teams that use multiple programming languages, but it also creates new requirements in terms of management and deployment.

DataOps.live is a transformational way for data engineers and data scientists to create new value for their organizations—and provides the only purpose-built platform to help modern data teams take the next steps forward with Snowpark.

DataOps.live enriches the capabilities of Snowpark because it manages all of the things around it: you can 'do more with less' and make even more of the opportunities presented by Snowpark through automation, orchestration, and so on. These capabilities enable complete lifecycle management and deployment of the software source code used in Snowpark (Python, Java, Scala). The flipside is that without DataOps.live, additional steps and further manual effort are required of the user.