SNOWFLAKE FOR DATA SCIENCE & ML

June 2022

Challenges Preventing ML at Scale

Prepare

66%

Data scientist's time spent collecting and improving data

Anaconda. State of Data Science 2020

Build

Ungoverned copies of data sitting in complex-to-manage systems

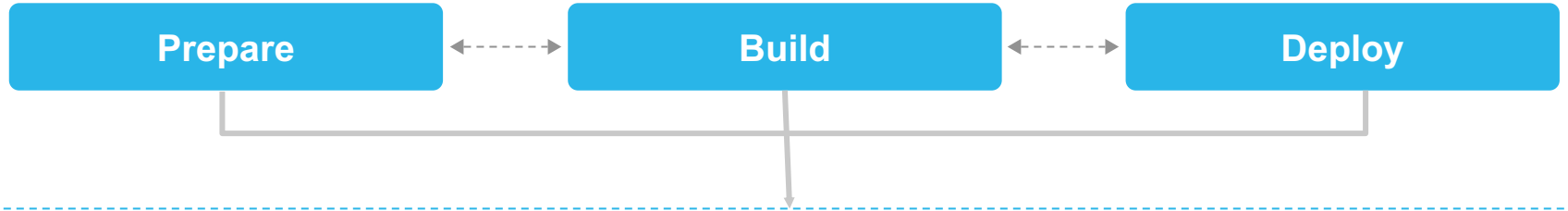
Deploy

62%

Organizations struggle to operationalize ML models

Forrester. Operationalize ML 2020

Snowflake for Data Science & ML



Make relevant data instantly accessible

- Reduce data collection time
- Support for all data structures
- Use shared and third-party data

Accelerate processing with near-zero operations

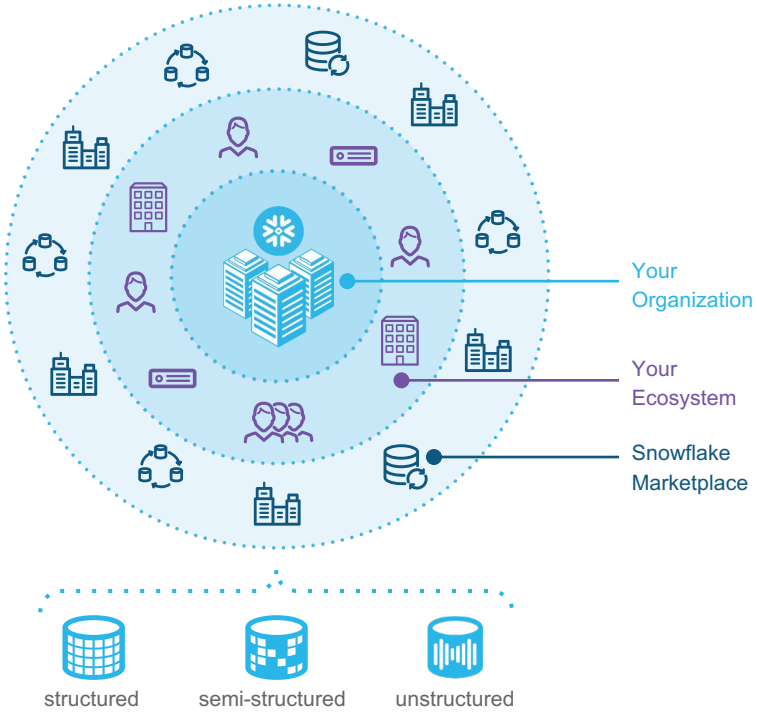
- Use language and tool of choice*
- Process data where it lives
- Focus on building not managing

Unify teams around governed data

- Streamline your architecture
- Simplify path to production
- Advanced data governance & security

*Snowpark Python in preview

One Place to Instantly Access Relevant Data



Reduce data collection time

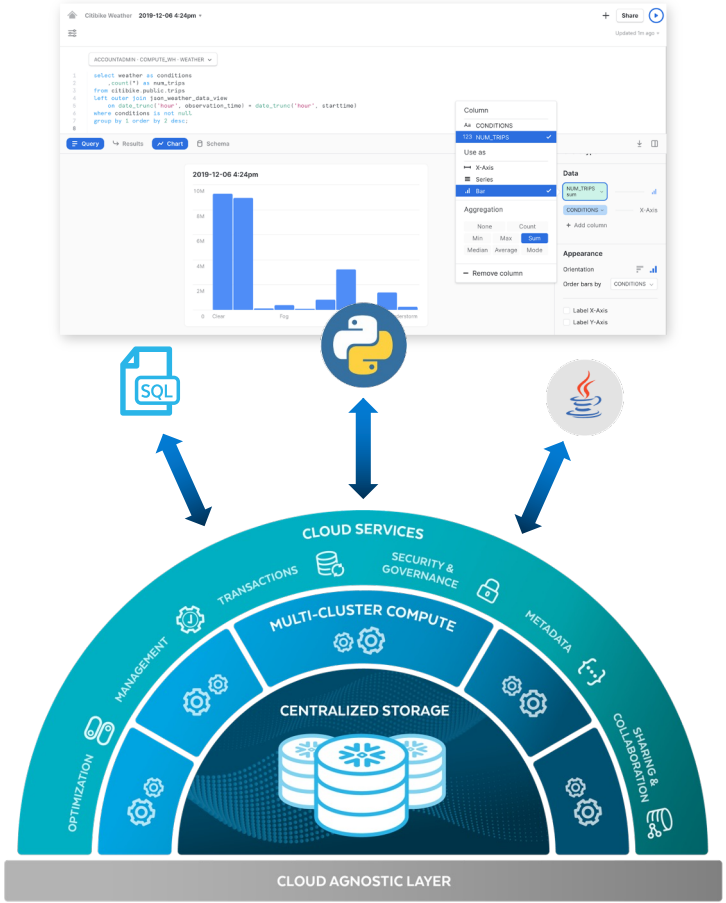
Single point for discovery and access to a global network of high-quality data

Bring all data types into your model with ease

Native support for structured, semi-structured and unstructured data

Build powerful models with shared data/services

Easily incorporate shared data, and third-party data & data services via Snowflake Data Marketplace



Fast Processing Engine With No Operational Overhead

Prepare data with your language of choice

Support for ANSI SQL and Java/Scala & Python with Snowpark* for feature engineering

Handle any amount of data or users

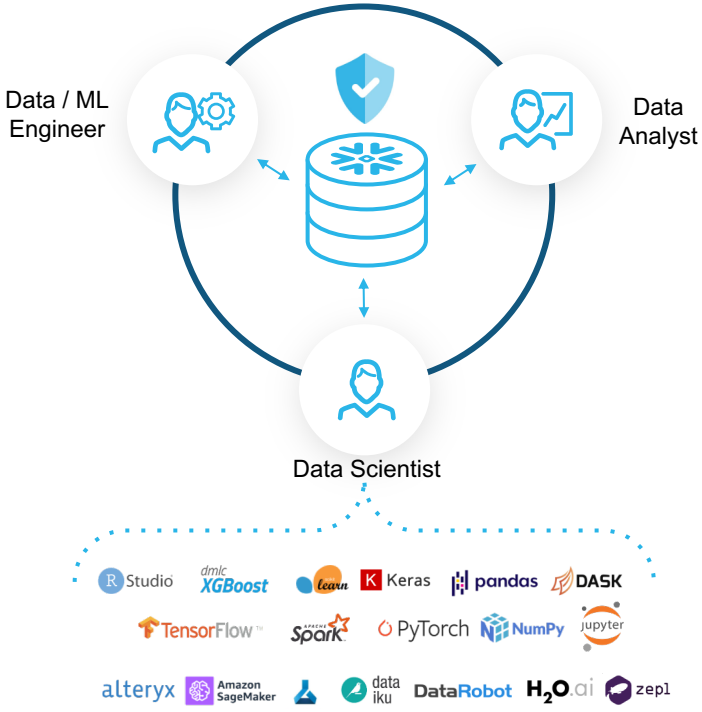
Intelligent multi-cluster compute infrastructure instantly scales to meet your data preparation demands without bottlenecks of user concurrency limitations

Automate and scale feature pipelines

Use Streams & Tasks to automate feature engineering pipelines for model inference

*Python in preview

Single Platform to Unify Teams Around Governed Data



Connect your ML tool of choice to Snowflake data

Native integrations along with Python, R and Spark connectors effortlessly extend Snowflake to cutting edge ML tools

Simplify MLOps with secure model inference

Run inference in Snowflake with models as UDFs* or trigger request to secure model endpoint with External Functions

Increase trust in your models with data governance

Advanced data security & governance features in Snowflake to understand, classify, and protect the data going into your models

*Python UDFs in preview

No.1 restaurant for health & safety compliance during COVID-19[1]

"Curated, reliable data sets from Snowflake Data Marketplace just make things so much easier, and we're super excited to leverage those data sets to enhance the performance of our machine learning models.

Mash Syed
Lead Data Scientist

Preventing fraud for over 9,000 businesses worldwide

"The elasticity and near-zero maintenance of Snowflake enables our data science team to elevate our productivity by spending less time preparing data and spending more time building models.

Matthew Jones
Data Science Manager

Predicting real-time availability of +200M grocery items every hour

"The new scoring architecture [using Snowflake] that we built from scratch scores 15x more items, using one-fifth of the resources in one-quarter of the time.

Abhay Pawar
Senior Machine Learning Engineer

[1]Source: Restaurant Dive. 2020

Chipotle is a fast casual restaurant chain with a radical belief that there is a connection between how food is raised and prepared, and how it tastes. Real is better.

Challenge

- Shift from reactive to preventive operations for restaurant safety practices during pandemic

With Snowflake

- Effortlessly combined internal data with third-party COVID-19 case counts to leverage machine learning to drive decisions.
- Live, ready-to-query COVID-19 data from Snowflake Data Marketplace simplified Chipotle's data pipeline and reduced administrative effort.
- Recognized as top restaurant for health & safety compliance during COVID-19 thanks to up-to-date, ML-driven operations

> Curated, reliable data sets from Snowflake Data Marketplace just make things so much easier, and we're super excited to leverage those data sets to enhance the performance of our machine learning models.
>
> **– Mash Syed, Lead Data Scientist**

Kount®
An Equifax Company

Kount, an Equifax company, exists to protect digital innovation. Kount's Identity Trust Platform analyzes signals from 32 billion interactions per year to prevent fraud and enable personalized customer experiences.

Challenge

- Kount's on-premises data environment could not easily scale to handle the company's data volumes and data science workloads
- This inhibited feature generation

With Snowflake

- Native support for structured and semi-structured data, Snowflake as data lake makes all relevant data for our machine learning models easily accessible
- The elasticity and near-zero maintenance of Snowflake enables the data science team to elevate their productivity by spending less time preparing data, so they can spend more time building models

The elasticity and near-zero maintenance of Snowflake enables our data science team to elevate our productivity by spending less time preparing data, so they can spend more time building models.

– Matthew Jones, Data Science Manager, Kount

Instacart is an online grocery delivery service in the United States and Canada. Instacart employees shop for your items at grocery stores and deliver them to your door.

Challenge

- Rapidly growing demand requiring predictions for real-time availability of +200 million grocery items

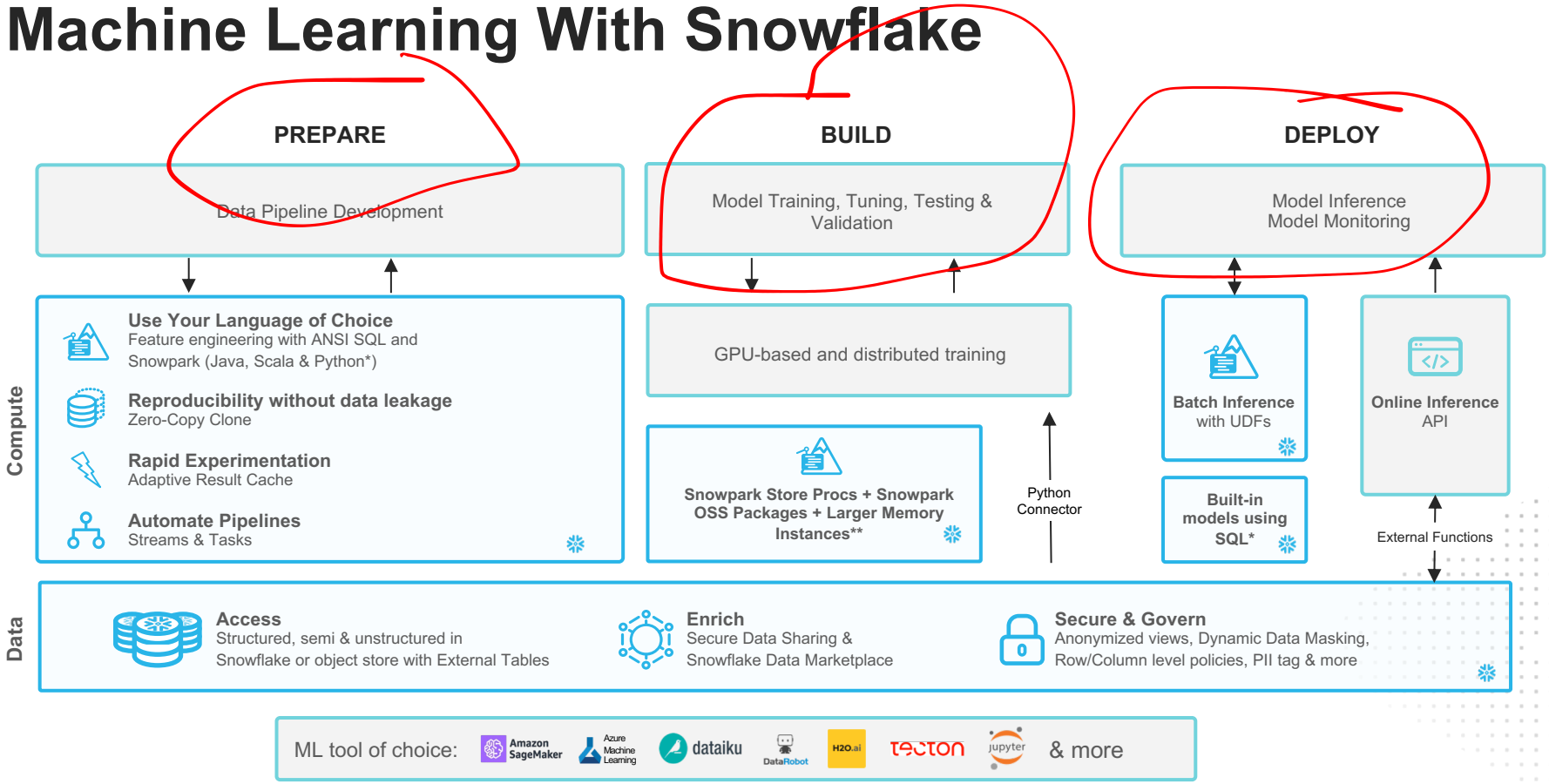
With Snowflake

- Scalable scoring pipeline with 130 features that are created for each item amounting to 10s of TB of data is processed every 60 minutes
- Enhanced customer experience with accurate expectations for out-of-stock items and recommend appropriate replacements for items likely to be out-of-stock.

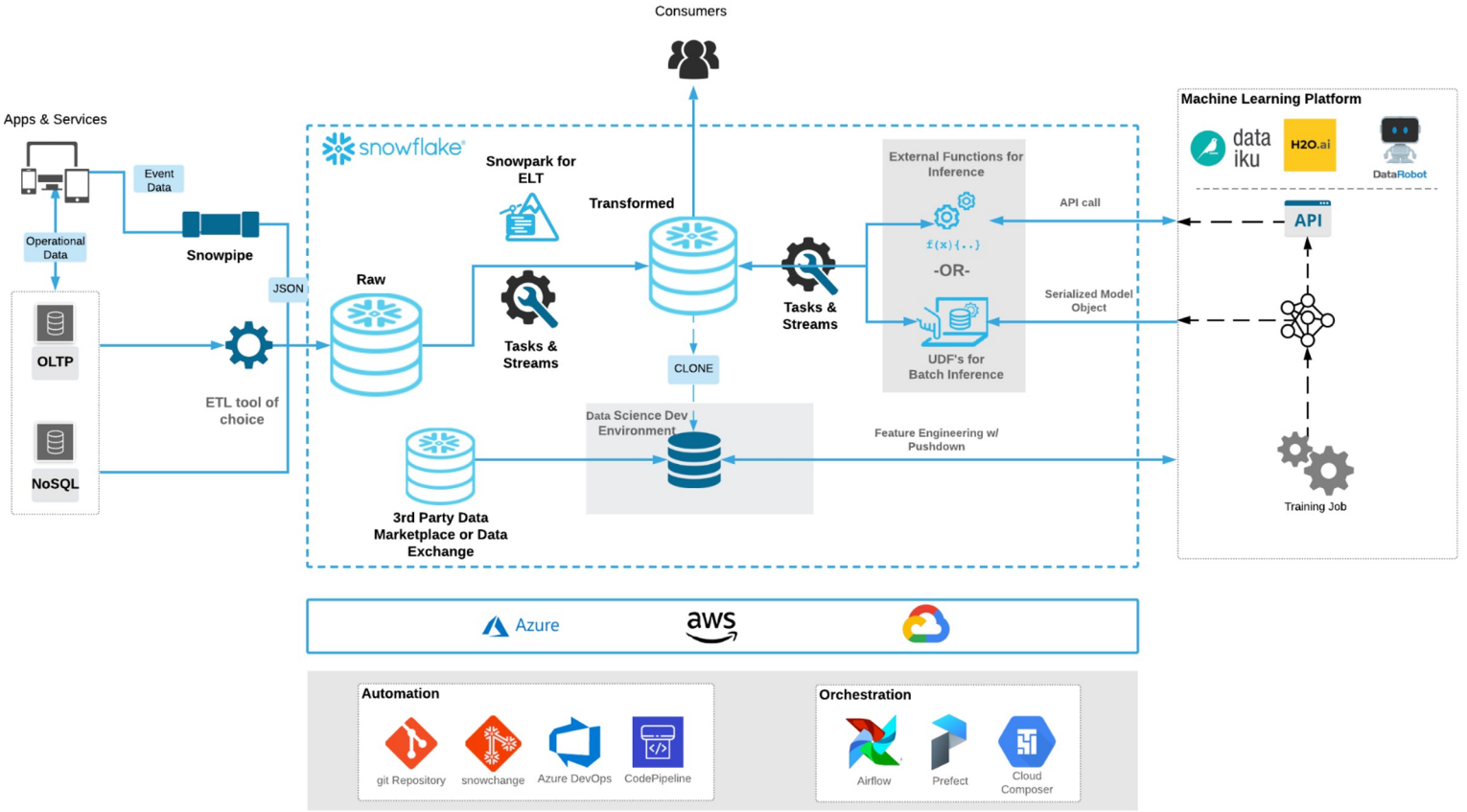
> An item's availability changes in near real-time as they get sold and restocked, and as such we want to predict availability as often as is possible.
>
> **– Abhay Pawar, Senior Machine Learning Engineer**

TECHNICAL DEEP DIVE

Machine Learning With Snowflake

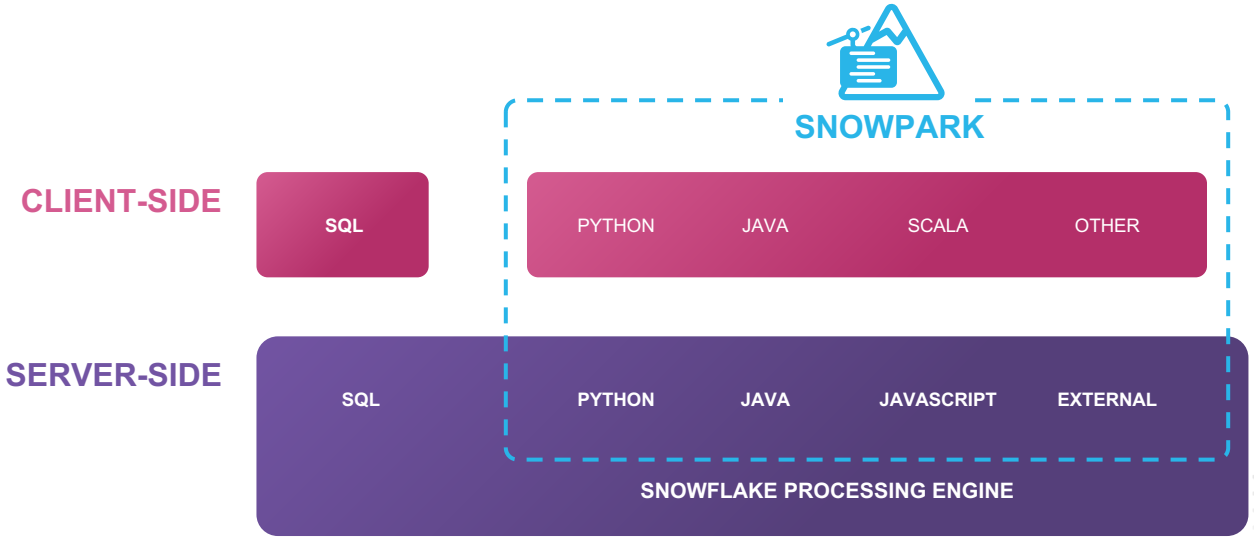


DATA SCIENCE REFERENCE ARCHITECTURE



SNOWPARK DEEP-DIVE

Code the Same Way, Execute Faster With Snowpark

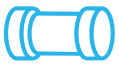


* Snowpark for Python in Preview

Why Snowpark

Streamline Architecture

Collaborate on the same data in a single platform by natively supporting different user's programming language of choice



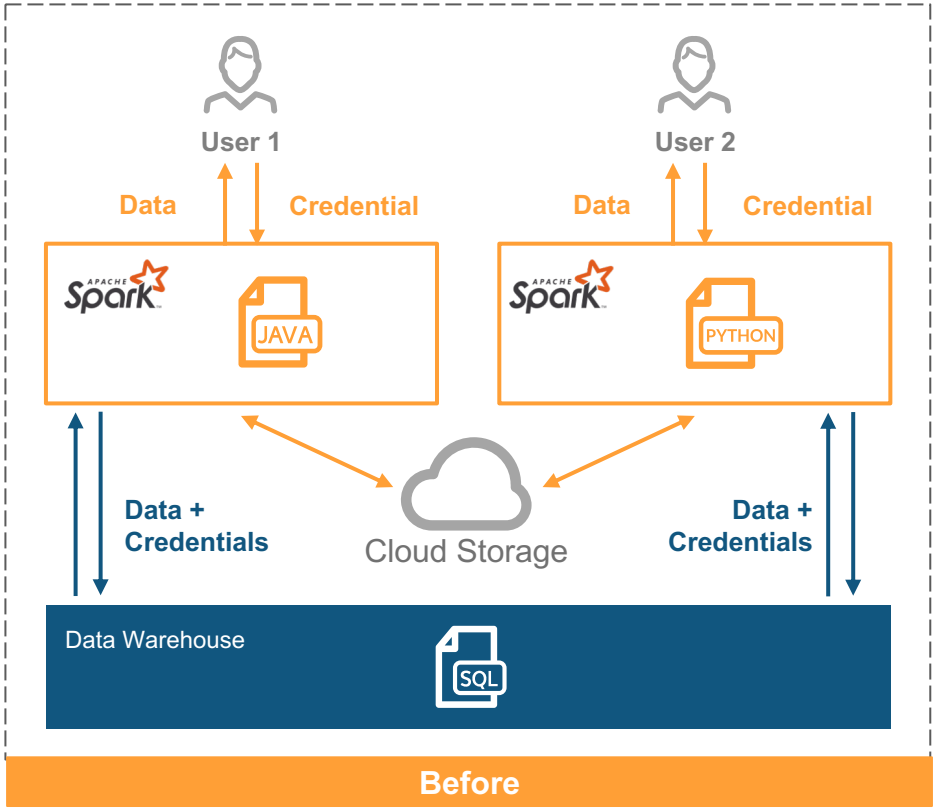
Build Scalable & Optimized Pipelines

Benefit from the Snowflake Data Cloud with superior price/performance and near-zero maintenance

Act With Confidence

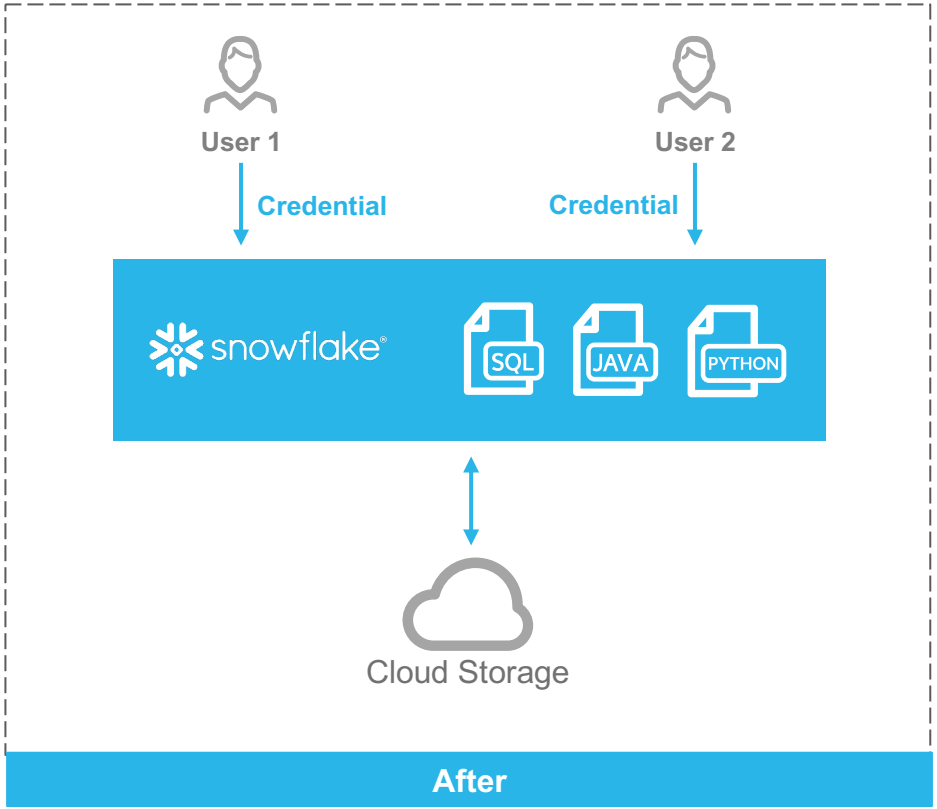
Enforce consistent, enterprise-grade governance controls & security across all your workflows

Complexity With Traditional Approach



- Customers often run separate processing clusters for different languages
- Complex capacity management & resource sizing
- Lots of data movement and data silos
- Loose governance control and security loopholes

Streamlined Architecture With Snowflake



- One single platform with native support for different languages
- Simpler capacity management & resource sizing
- Streamline architecture and collaborate on the same data
- Consistent governance and security policies

* Snowpark Python in Private Preview

Snowpark for Python

Familiar Programming Constructs

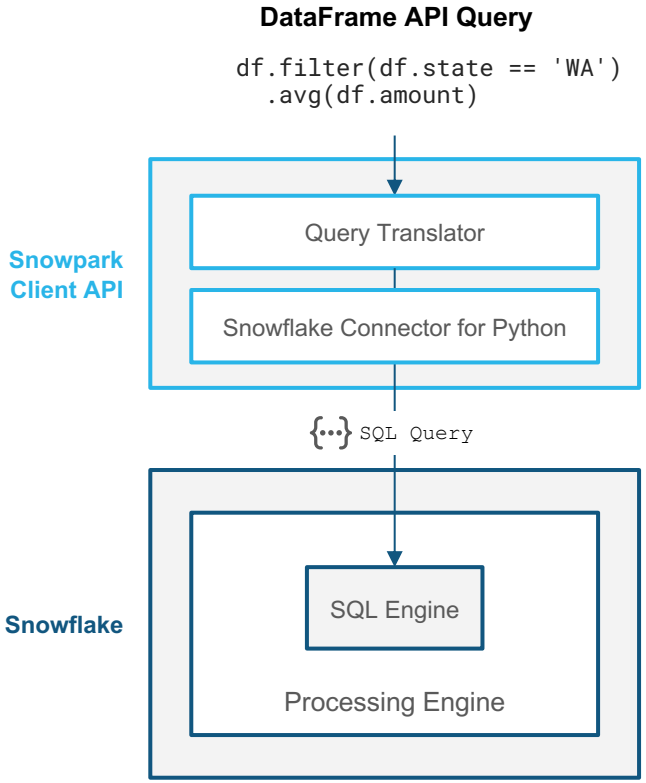
Use familiar syntax with DataFrame abstraction

Rich Ecosystem

Easy access to hundreds of packages with automated dependency management

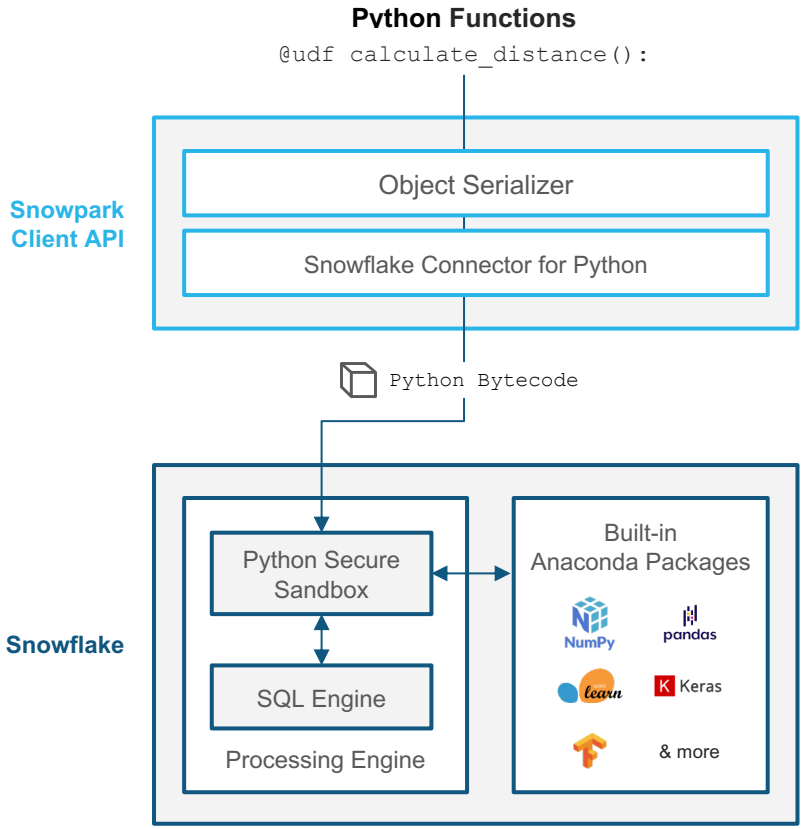
Secure Processing

Build with confidence in a highly secure, sandboxed environment



DataFrame API

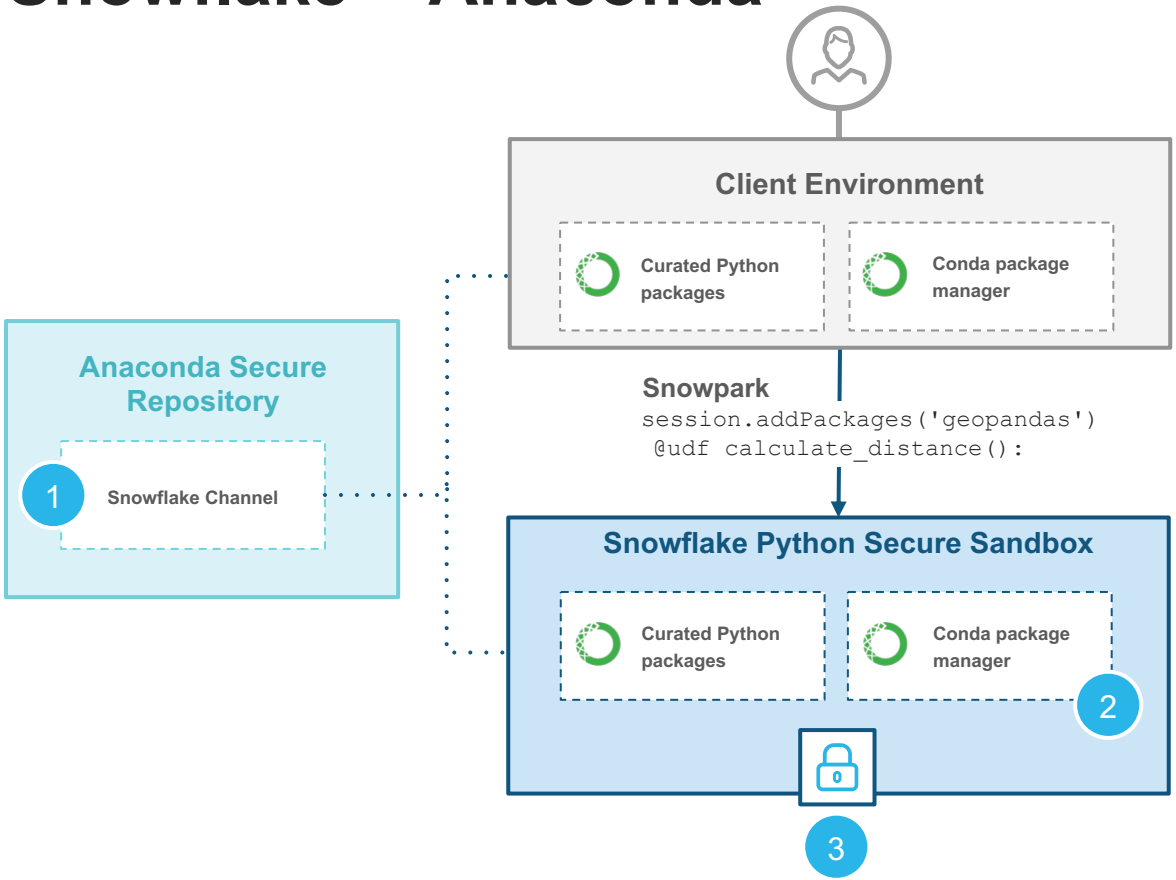
- Query Snowflake data with Python
- Familiar DataFrame API
- 100% push-down to Snowflake
- Native Snowflake performance and scale



Python Functions

- Bring custom Python code to Snowflake as User Defined Functions (UDFs)
- Code is serialized and pushed down to run in a secure sandboxed environment
- Seamlessly access third-party packages with Anaconda integration

Snowflake + Anaconda



1 Easy Access

Curated packages pre-installed in Snowflake also available for local development

2 No Dependency Hell

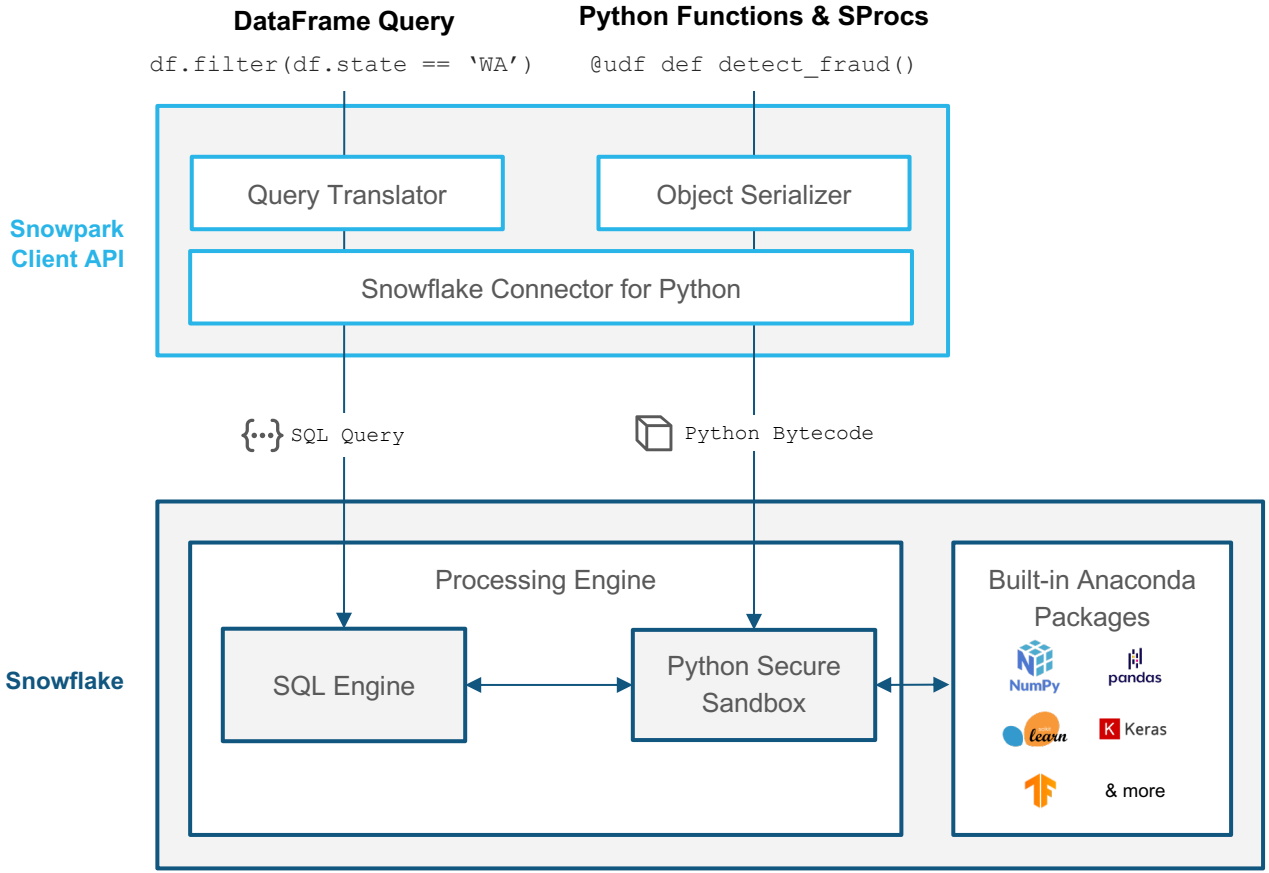
Conda package manager integrated in Snowflake secure sandbox

3 Scalable and Secure

Process with secure sandbox integrated into Snowflake processing engine

All of this with no additional charges beyond warehouse usage

Snowpark for Python



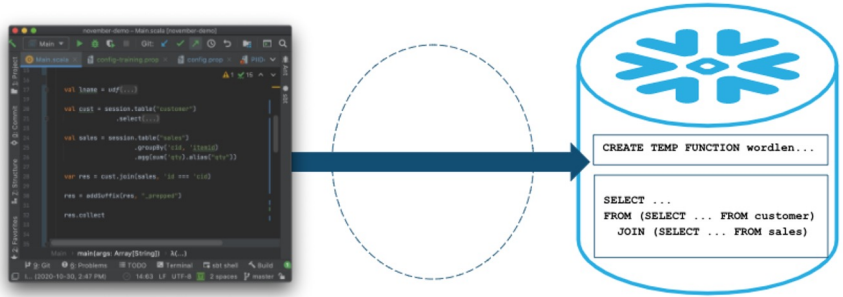
Snowpark

Example Use Cases:

- Data transformation, ELT systems
- Data preparation and feature engineering
- ML Scoring / Inference to operationalize ML models in data pipelines
- Data apps

Allows coders to:

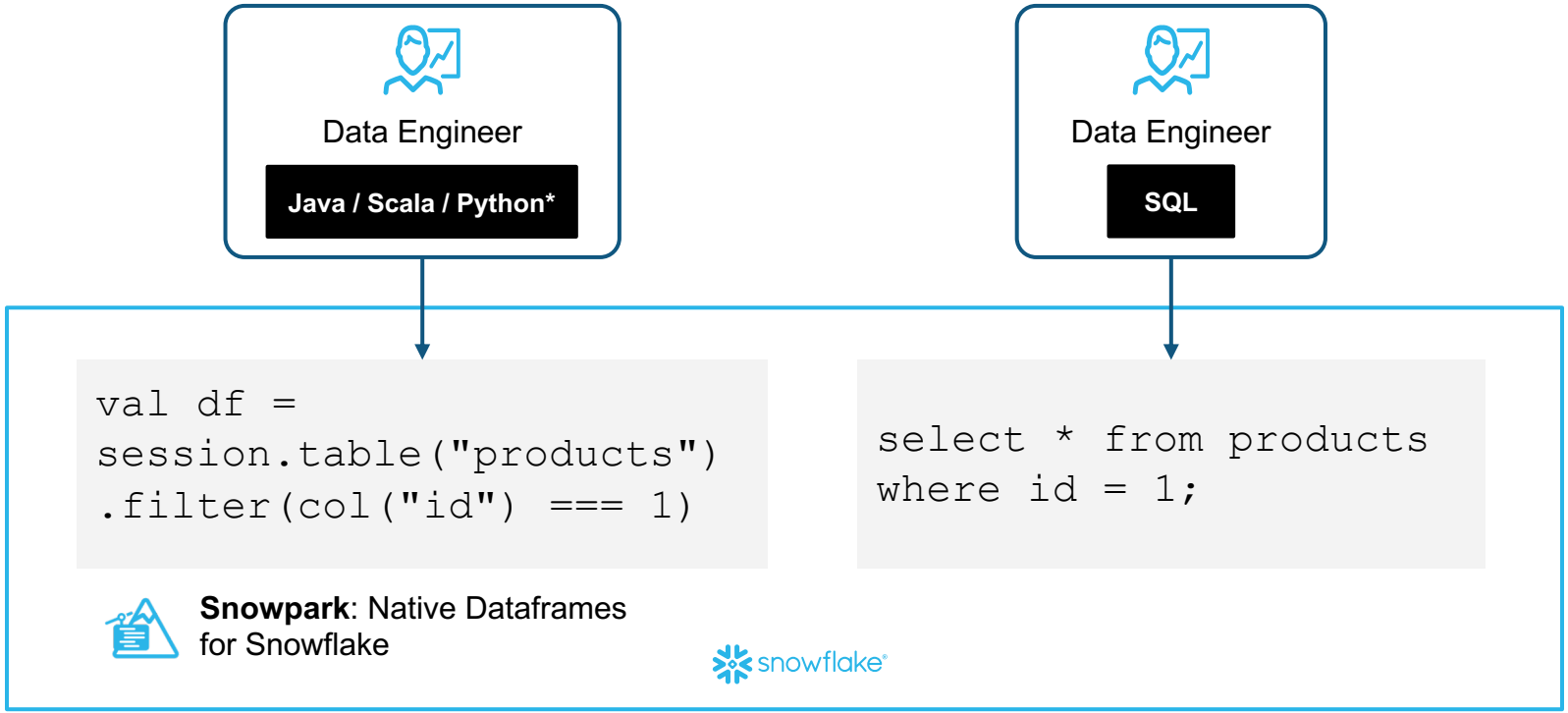
- **Code the same way, execute faster:** Use familiar language, constructs, libraries and IDE to code the same way, but get better price/performance and efficiency by optimizing execution with Snowflake's elastic performance for automated scalability.
- **Focus on what matters with streamlined architecture:** Fully managed platform with automation, no more maintenance and tuning. Fewer systems to interact with, and no need to build and manage unnecessary data pipelines to move data in and out.
- **Eliminate redundant data processing:** Benefit from the Data Cloud to process data once and make it available for all of your use cases.



Snowpark pushes all of its operations directly to Snowflake without Spark or any other intermediary.

* Snowpark Java API & Python in Preview

How it Works



Java Functions

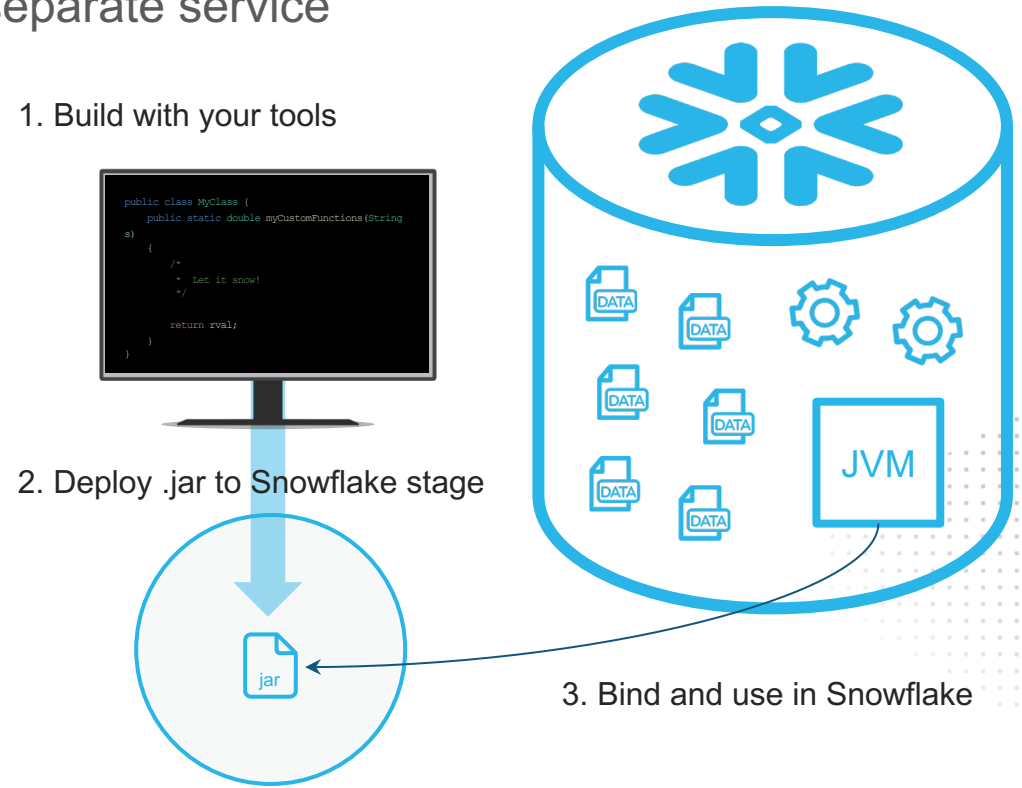
Transform and augment your data using custom logic running right next to your data, with no need to manage a separate service

Example Scenarios:

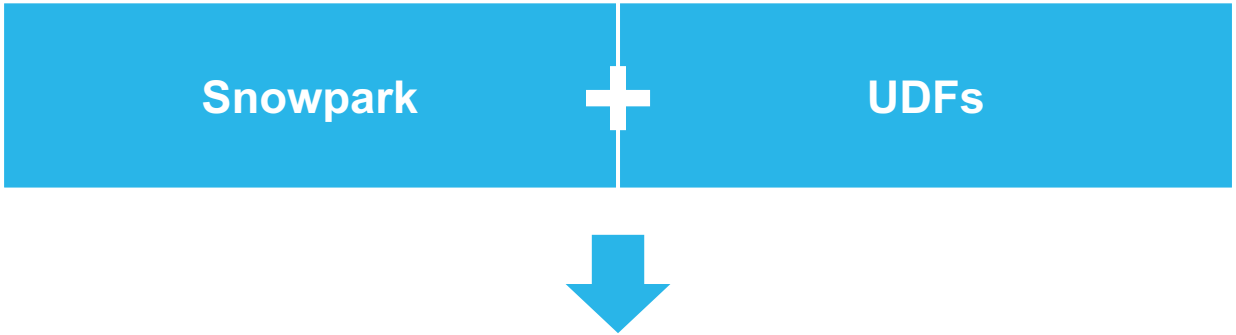
- ML Scoring
- Apply custom code
- Use third-party libraries

Benefits:

- Developers can build functionality into Snowflake using the popular Java language and libraries
- Users can access this functionality as if it were built into Snowflake
- Administrators can rest easy: data never leaves Snowflake

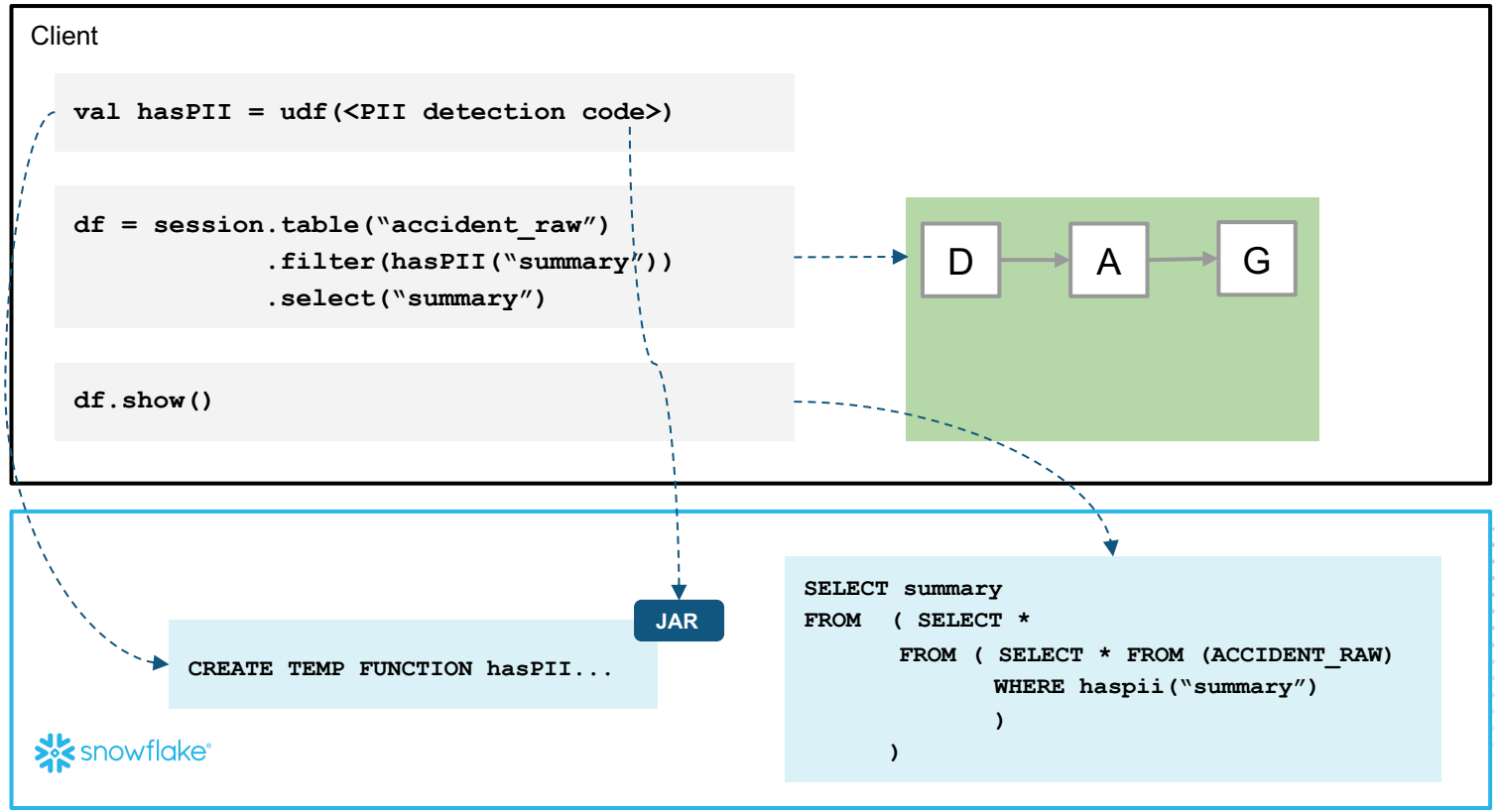


Snowpark + UDFs

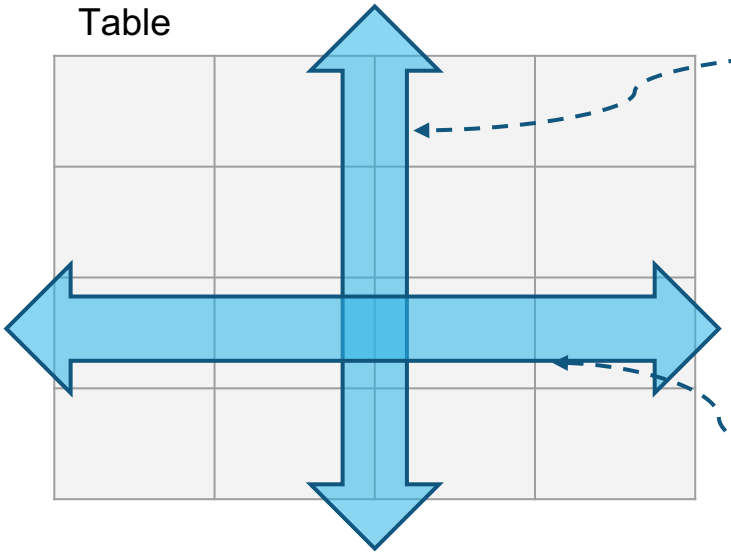


Functional Data Engineering
in Snowflake

Snowpark UDFs



Functions vs. Dataframes



Snowpark

```
df = session.table("accident_raw")
            .filter(callUDF("hasPII",
$"summary")
            .select("summary")
```

UDF

```
public boolean hasPII(String s){
   for(Pattern p : patterns){
            if(p.matcher(s).find()) return
true;
   }
}
```

snowflake

THANK YOU