# DATA SCIENCE HANDS-ON LAB

## Credit Card Fraud Detection using Snowpark and Java UDF

CARLOS CARRERO, GSI Sales Engineering EMEA I OCT 2021

```
docker pull ccarrero71/sf:Fraud-Detection-Lab-GSI
```

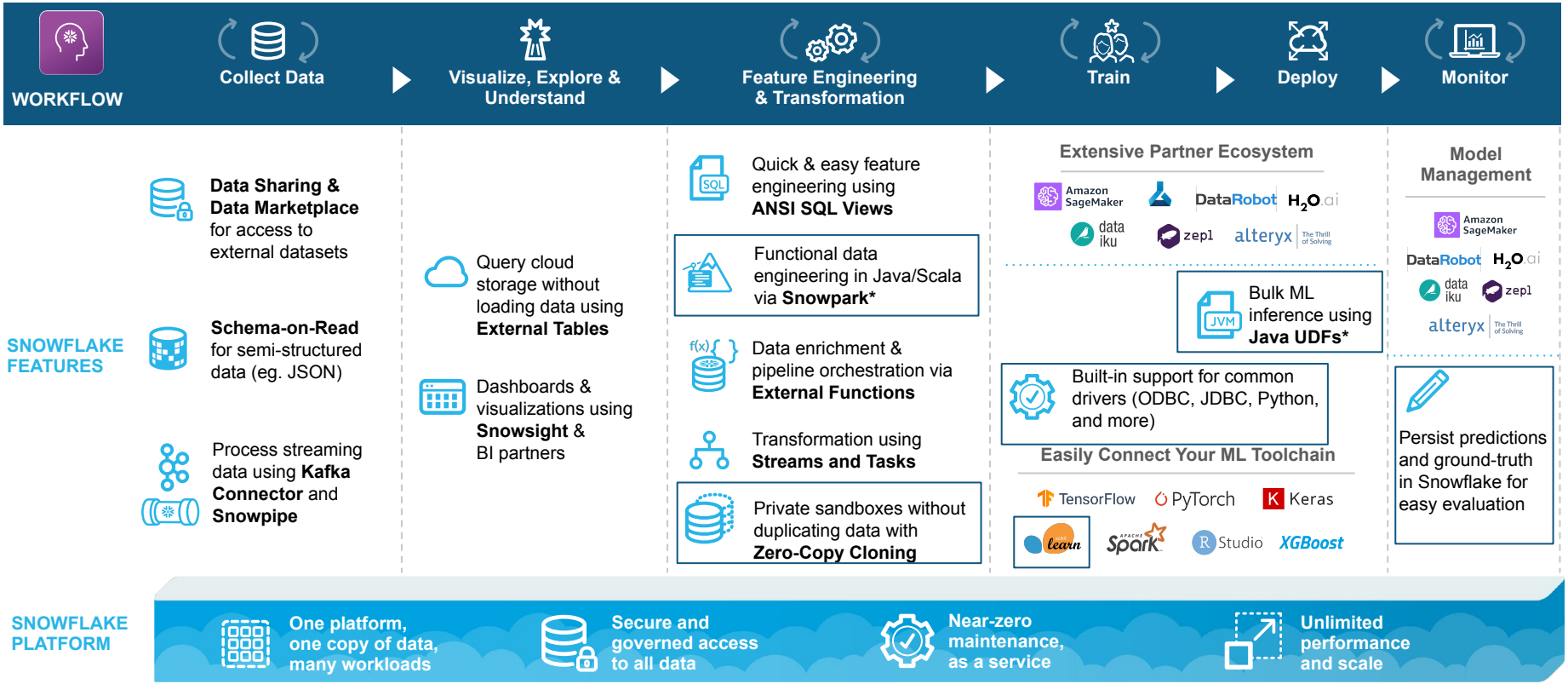
# AGENDA

- ❏ Snowflake for Data Science Intro - 15 minutes
- ❏ Lab Introduction - 15 minutes
- ❏ Credit Card Fraud Detection using Snowpark and Java UDF Lab - 1h

# GOALS

- ❏ Get familiar with new Snowpark capabilities
- ❏ Experience first hand how Snowpark can bring big performance benefits to Feature Engineering
- ❏ Get familiar with Java UDFs for ML Scoring
- ❏ Understand data management capabilities to facilitate features consumption

# DATA SCIENCE WITH SNOWFLAKE LAB FOCUS



# DATA SCIENCE WITH SNOWFLAKE
# BEST PRACTICES

Enrich datasets using **Data Marketplace** for improved model accuracy


Use **Streams & Tasks** to build end-to-end ML pipelines


Create datasets without loading data into Snowflake via **External Tables**


Leverage **External Functions** to trigger training or get predictions

Use **Zero-Copy Clones** for training snapshots

Use regular or Materialized **Views** to create repository of ML features used for training and prediction

Optimize training instance memory usage by using **Snowflake SQL** for aggregation & sampling

# SNOWPARK

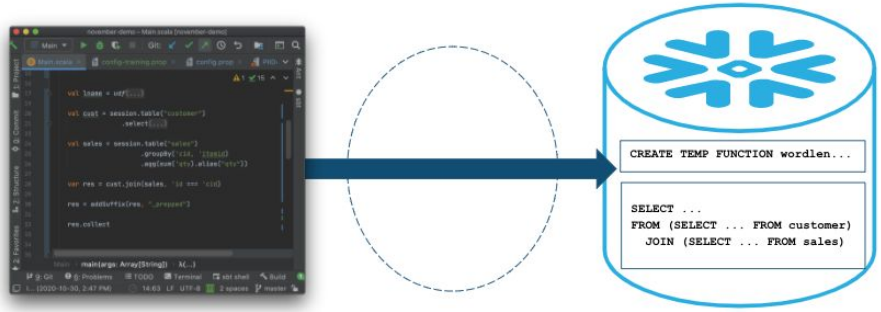
A new developer experience that allows you to write Snowflake code in your preferred way, and execute it directly within Snowflake

## Example Use Cases:

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

## Allows coders to:

- Write in your language with your preferred tool
- Easily complete and debug data pipelines with familiar constructs such as DataFrames, and bring in third-party libraries.
- Eliminate the need to have other processing systems, and run directly on Snowflake.




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

**Support for Scala is in Public Preview. Plan to add other languages in future.*

# JAVA FUNCTIONS

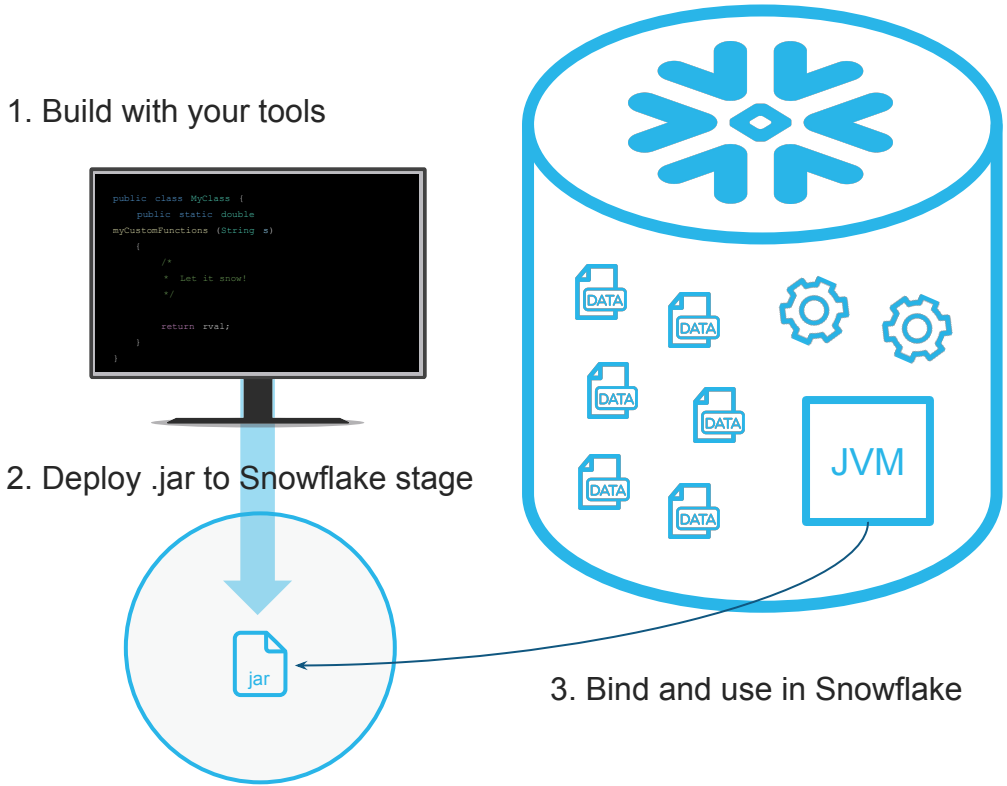
Transform and augment your data using custom logic running right next to your data, with no need to manage a separate service.

**Example Scenarios:**

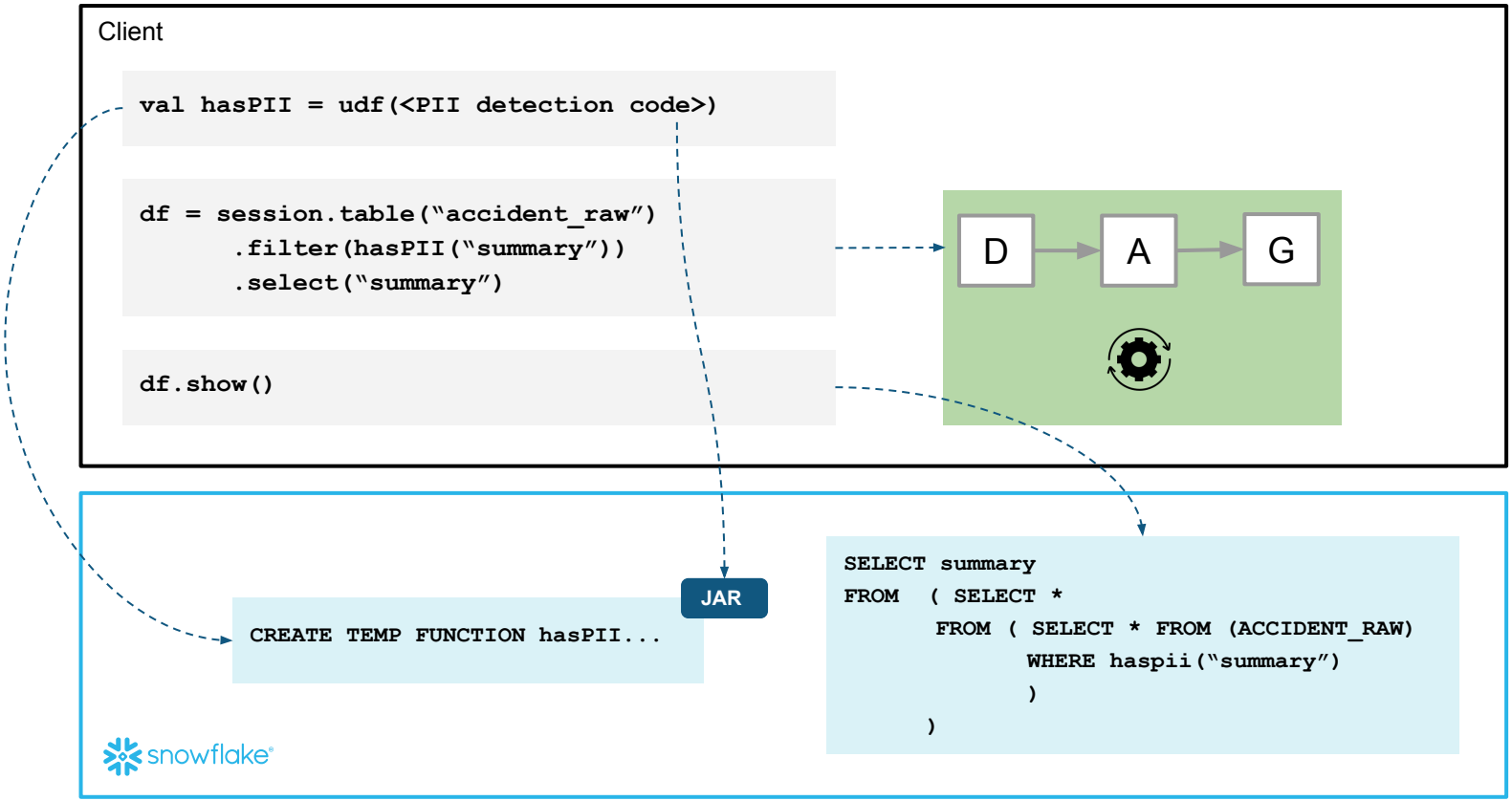
- ML Scoring
- Apply custom code
- Use third-party libraries

**Benefits:**

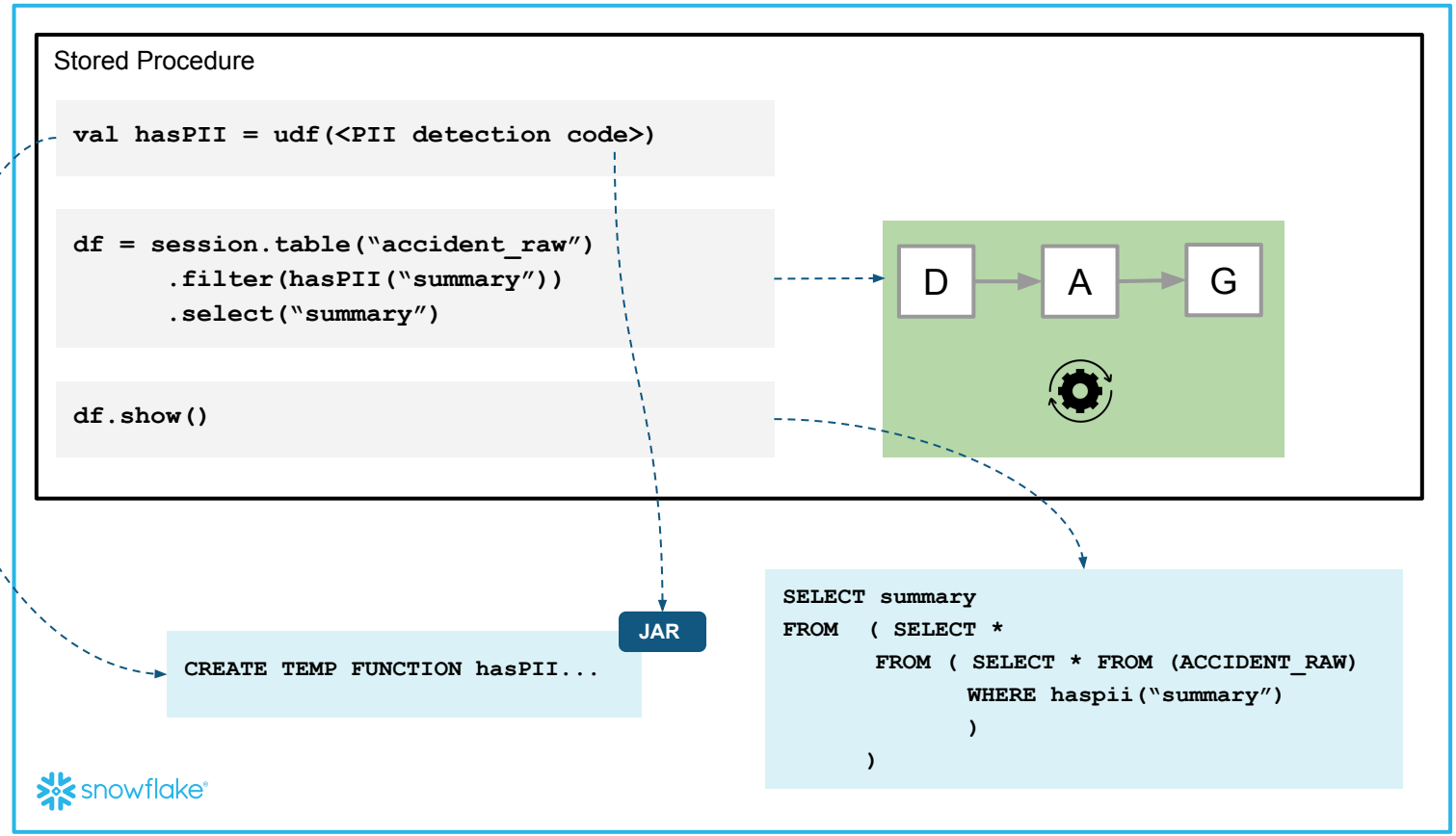
- Developers can build functionality into Snowflake using the popular Java language and libraries.
- Users can access this functionality as if it were built into Snowflake.
- Administrators can rest easy: data never leaves Snowflake.



# SNOWPARK + UDFs



# SNOWPARK + UDFs + SPs

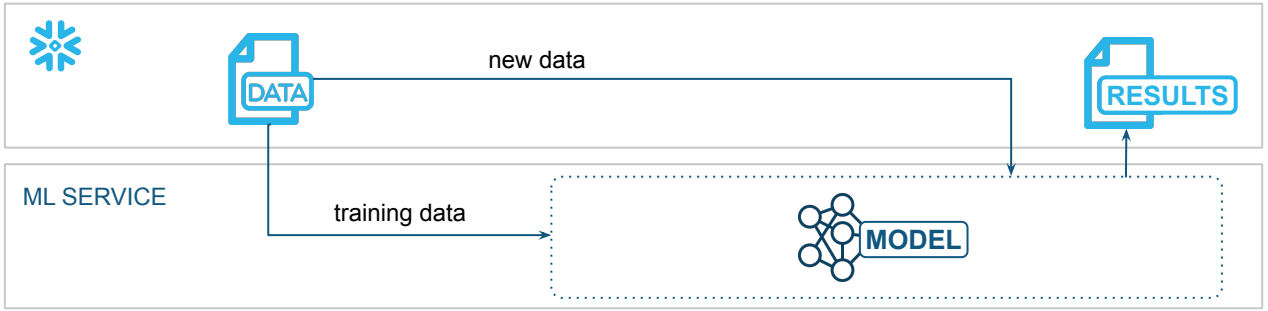


# JAVA UDFs FOR MODEL INFERENCE

Prepare | Build | Deploy

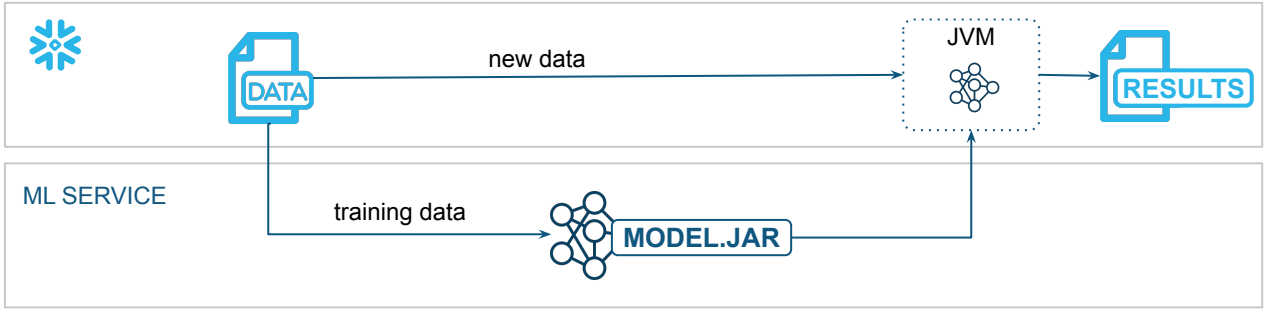
## EXTERNAL SERVING

Data continuously travels to externally hosted model



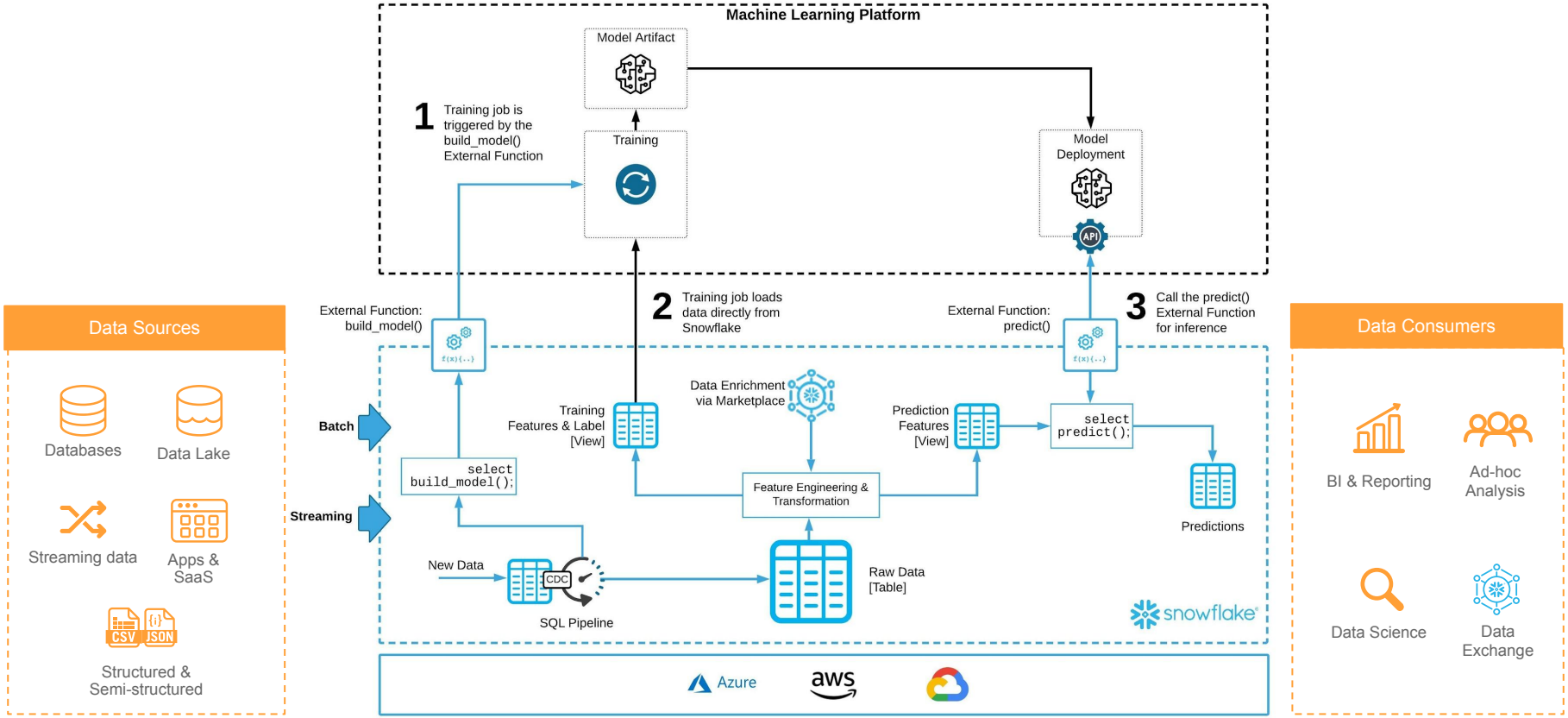

## WITH JAVA UDF

Model packaged as java file (.jar) runs where data lives



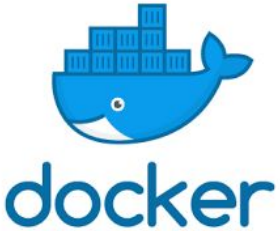

ML partners with .JAR models: DataRobot, Dataiku, H2O

# DATA SCIENCE REFERENCE ARCHITECTURE



# Lab Introduction

# How to Run the Lab




Get a Snowflake Trial Account (3 days in advance)

Available as a Docker Image. Install Docker Desktop and run:

```
docker pull ccarrero71/sf:Fraud-Detection-Lab-GSI
```

```
docker run --rm -p 8888:8888 -e JUPYTER_ENABLE_LAB=yes ccarrero71/sf:Fraud-Detection-Lab-GSI
```

Copy/paste the link provided in a browser and open work/START_HERE notebook for an overview

# Credit Card Fraud Detection Lab

**Data Collection**

Data Ingestion via Snowpark

- 00 - Snowpark - Load Data

**Feature Engineering**

Create new Features for ML Models

- 01 - Snowpark - Feature Engineering

**Model Training**

Train Model and Create PMML

- 02 - Python- Train Model

**Model Deployment**

Score using Java UDF

- 03 - Snowpark - Deploy Model & Scoring

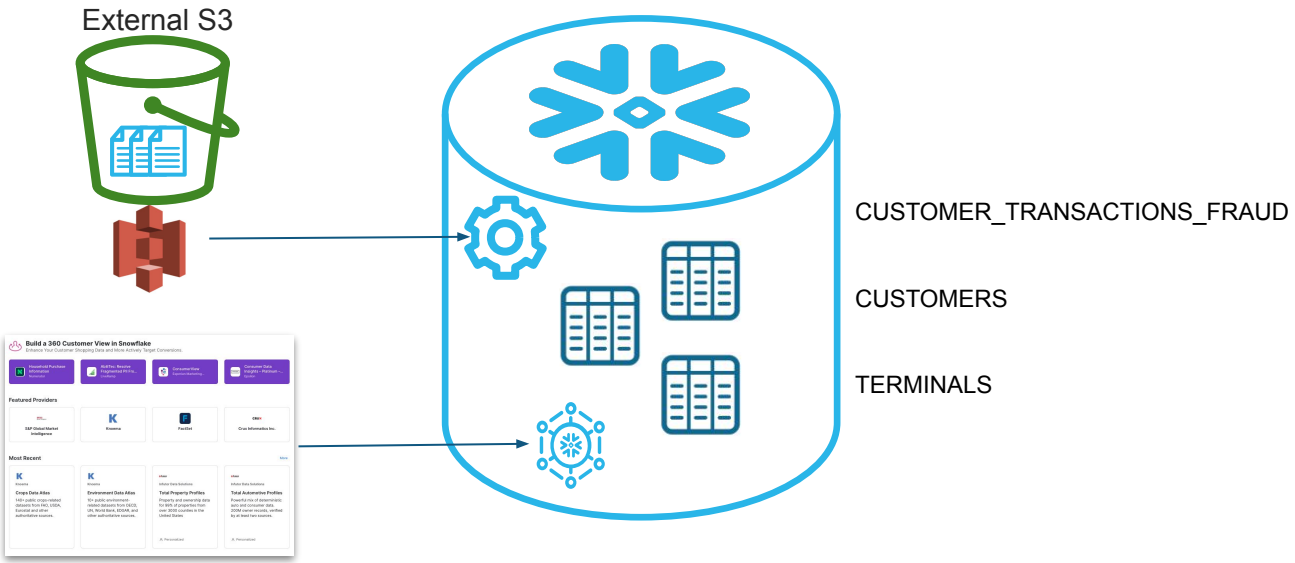
PMML: Predictive Model Markup Language

# Demo for Credit Card Fraud Detection

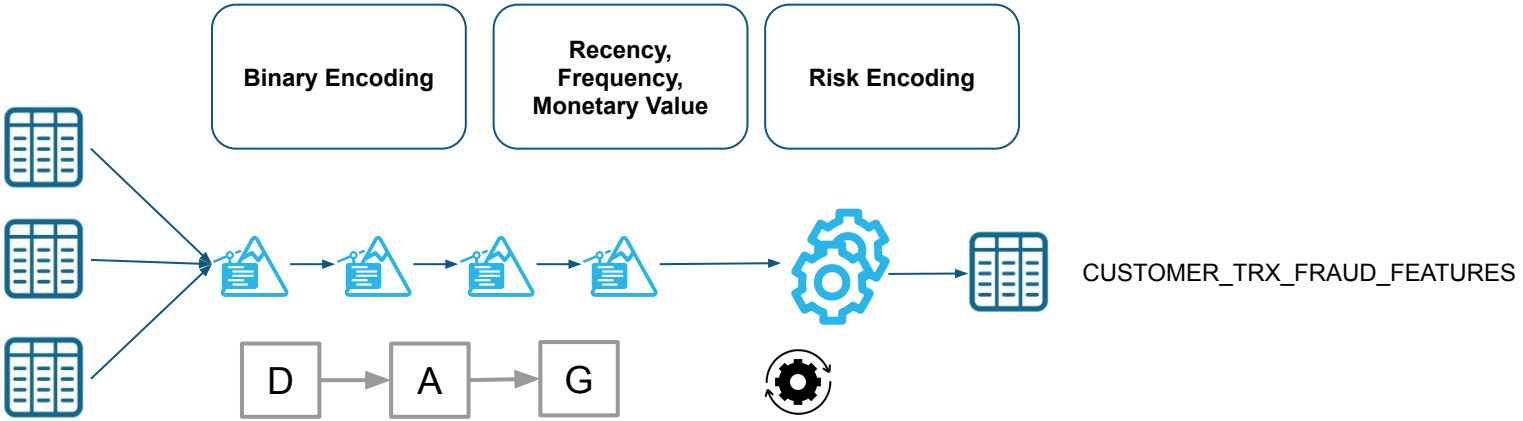
## Data Collection

Data Ingestion via Snowpark

00 - Snowpark - Load Data



# Feature Engineering

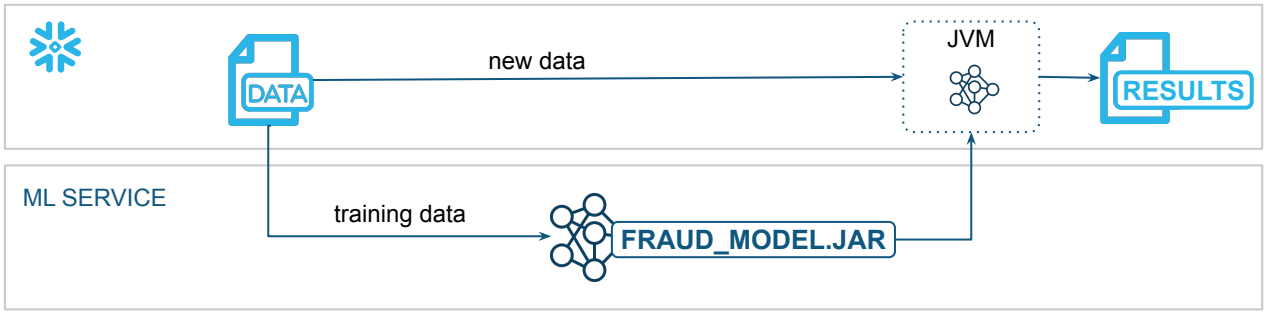


# Training Model

# Deploy Model

WITH JAVA UDF

Model packaged as java file (.jar) runs where data lives



# SUMMARY

- Support of native Data Frames for Snowpark uses lazy evaluation
- Queries only executed at Snowflake when running .collect(), .show(), etc..
- Very efficient method for running transformations
- Models being stored using PMML format
- Java UDF allows code execution within Snowflake

# Additional Links

Feature Engineering with Snowflake, Using Snowpark and Scala

Mats Stellwall Following

Oct 7 · 13 min read

From Zero to Snowpark in 5 minutes

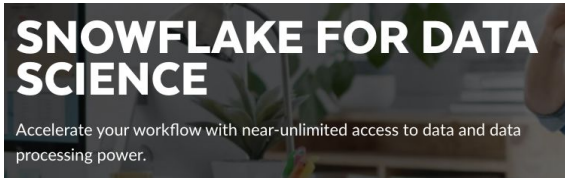
Zohar Nissare-Houssen Following

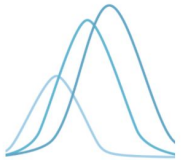
Aug 11 · 3 min read

Build a Recommendation Engine with AWS SageMaker

98 min Updated Sep 1, 2021

START



Machine Learning for Credit Card Fraud detection - Practical handbook