Fraud Detector using Machine Learning

Sarbani Maiti

AVP, AI/ML & Cloud in MagicFinserv

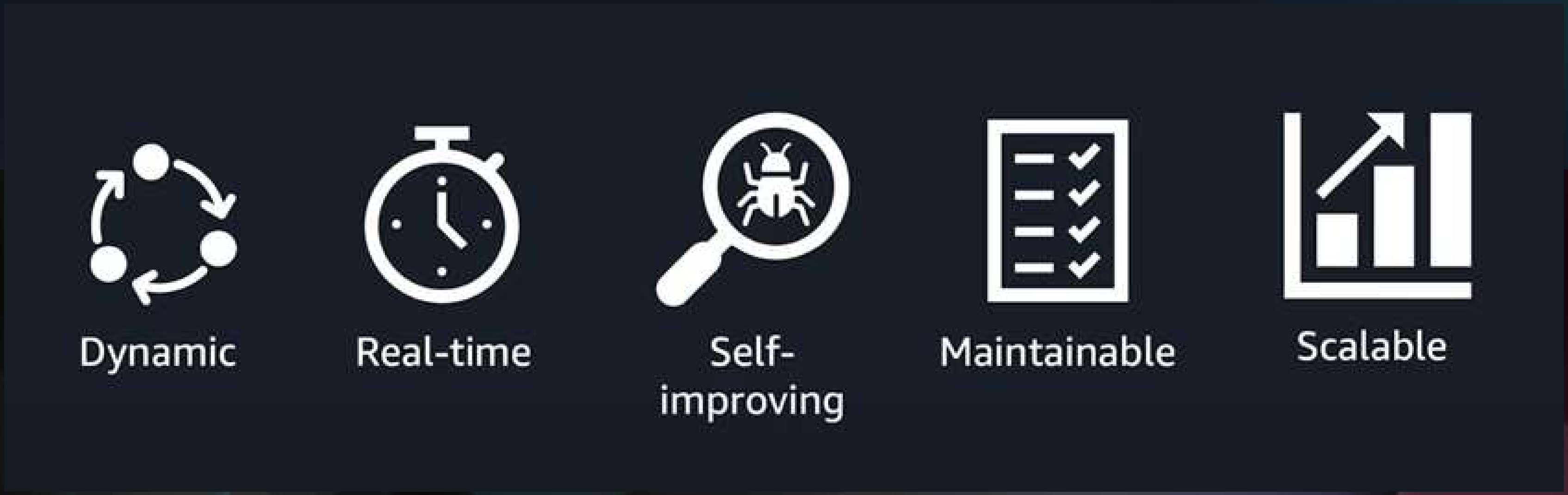
sarbaniiiitb2020@gmail.com

Fraud Detection with ML – Why Machine Learning

Rule based ML Fraud Detector challenges

- Static Handcrafted Rules
- Always Behind
- Bug-prone
- Complicated Code
- Cannot Scale

Fraud Detection with Machine Learning



Dynamic

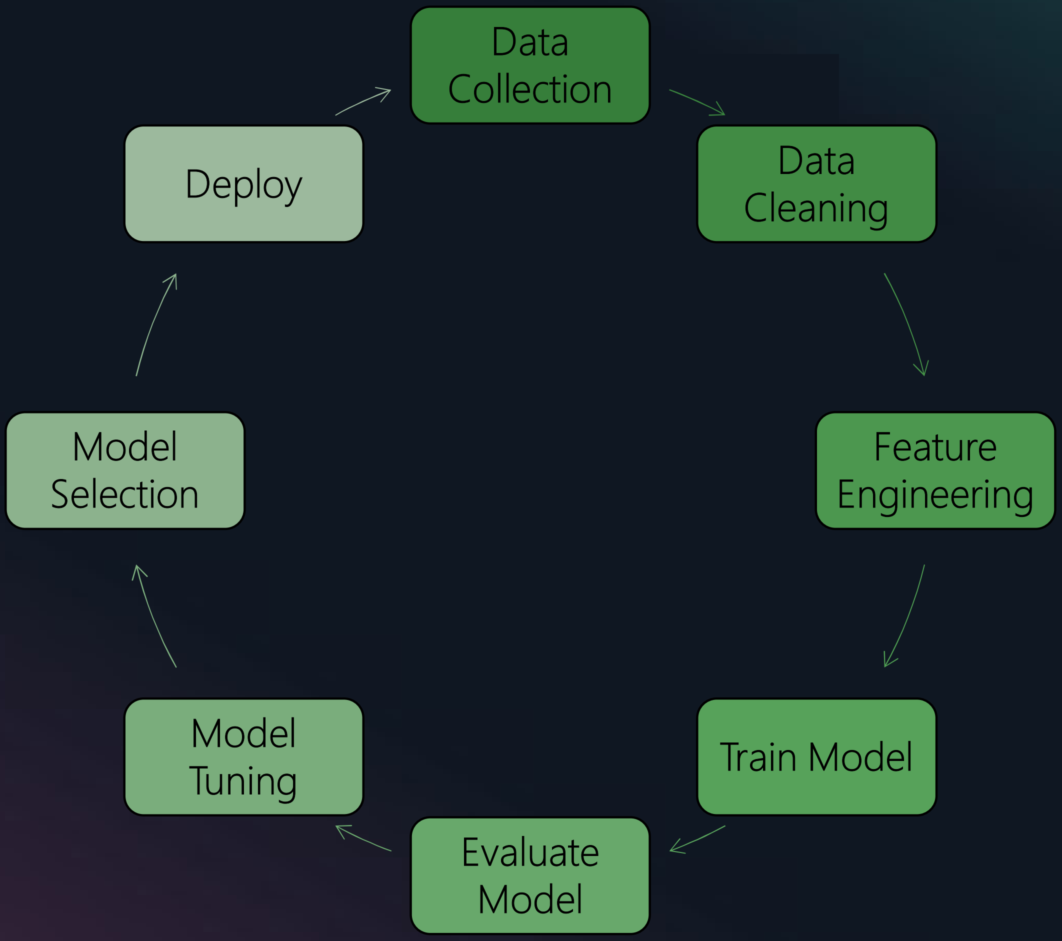
Real-time

Self-improving

Maintainable

Scalable

Machine Learning Lifecycle



Machine Learning Lifecycle in Production

Once the model is deployed, we need a production ready solution to serve the consumer applications.

Availability

- Availability of the ML system is critical for production system
- Onboarding the right kind of tool and technologies which are easy to adopt as per the change management system of the organization.
- Maintenance and reusability of the tools and models.
- Tracking, monitoring, alerting and feedback loop are other important aspects of the model in production

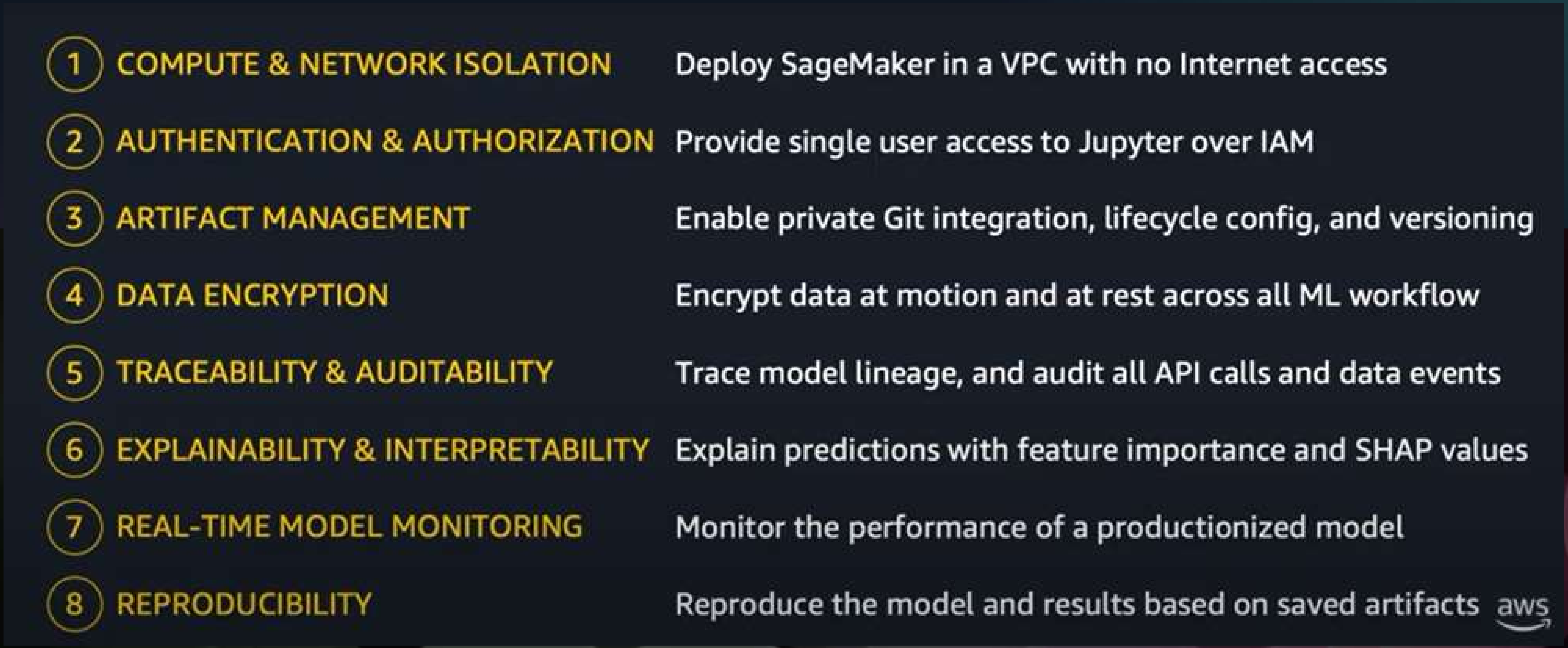
Scalability

- Model must support automatic scaling in production. Autoscaling dynamically adjusts the number of instances provisioned for a model in response to changes in your workload.
- When the workload increases, autoscaling brings more instances online.
- When the workload decreases, autoscaling removes unnecessary instances so that users don't pay for provisioned instances that are not in use.

Security

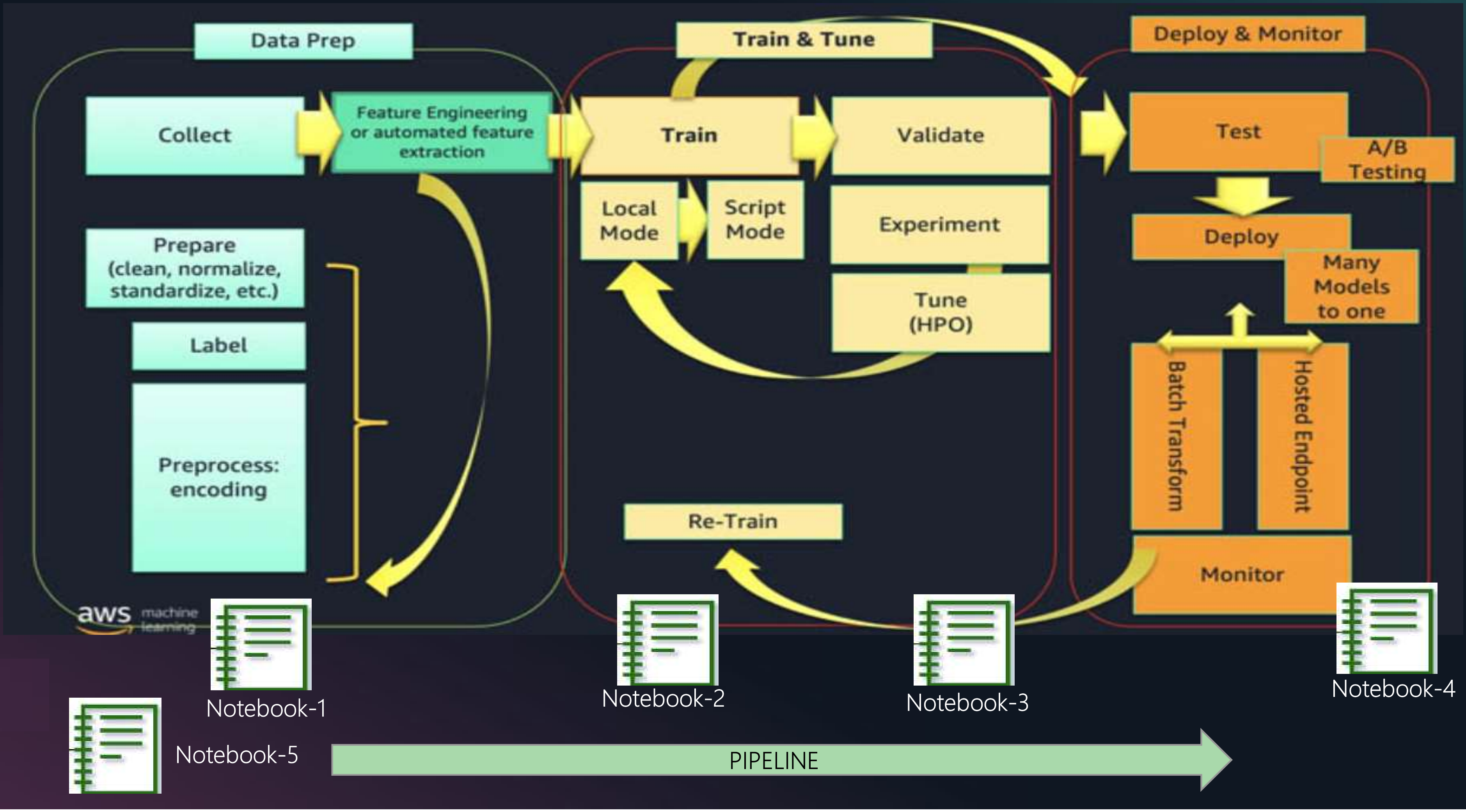
- The model has to be compliant as per the regulatory requirement.
- Create environments with the least privileged access to sensitive data
- Protect & Encrypt sensitive data
- Audit and trace activity in your environment
- Reproduce results in your environment by tracking the lineage of ML artifacts throughout the lifecycle and using source and version control tools

Fraud Detection Model –
Secure & Complaint ML Workflow



1	COMPUTE & NETWORK ISOLATION	Deploy SageMaker in a VPC with no Internet access
2	AUTHENTICATION & AUTHORIZATION	Provide single user access to Jupyter over IAM
3	ARTIFACT MANAGEMENT	Enable private Git integration, lifecycle config, and versioning
4	DATA ENCRYPTION	Encrypt data at motion and at rest across all ML workflow
5	TRACEABILITY & AUDITABILITY	Trace model lineage, and audit all API calls and data events
6	EXPLAINABILITY & INTERPRETABILITY	Explain predictions with feature importance and SHAP values
7	REAL-TIME MODEL MONITORING	Monitor the performance of a productionized model
8	REPRODUCIBILITY	Reproduce the model and results based on saved artifacts

Fraud Detection with ML – AWS Sagemaker Solution



Data Prep

Collect

Feature Engineering or automated feature extraction

Prepare (clean, normalize, standardize, etc.)

Label

Preprocess: encoding

Train & Tune

Train

Validate

Local Mode

Script Mode

Experiment

Tune (HPO)

Re-Train

Deploy & Monitor

Test

A/B Testing

Deploy

Many Models to one

Batch Transform

Hosted Endpoint

Monitor

Notebook-1

Notebook-2

Notebook-3

Notebook-4

Notebook-5

PIPELINE

Fraud Detection with ML – AWS Sagemaker Solution

- Notebook 1: Data Prep, Process, Store Featur
  - Data Wrangler
  - Datasets Processing
  - Sagemaker Feature Store
  - Create train and test datasets

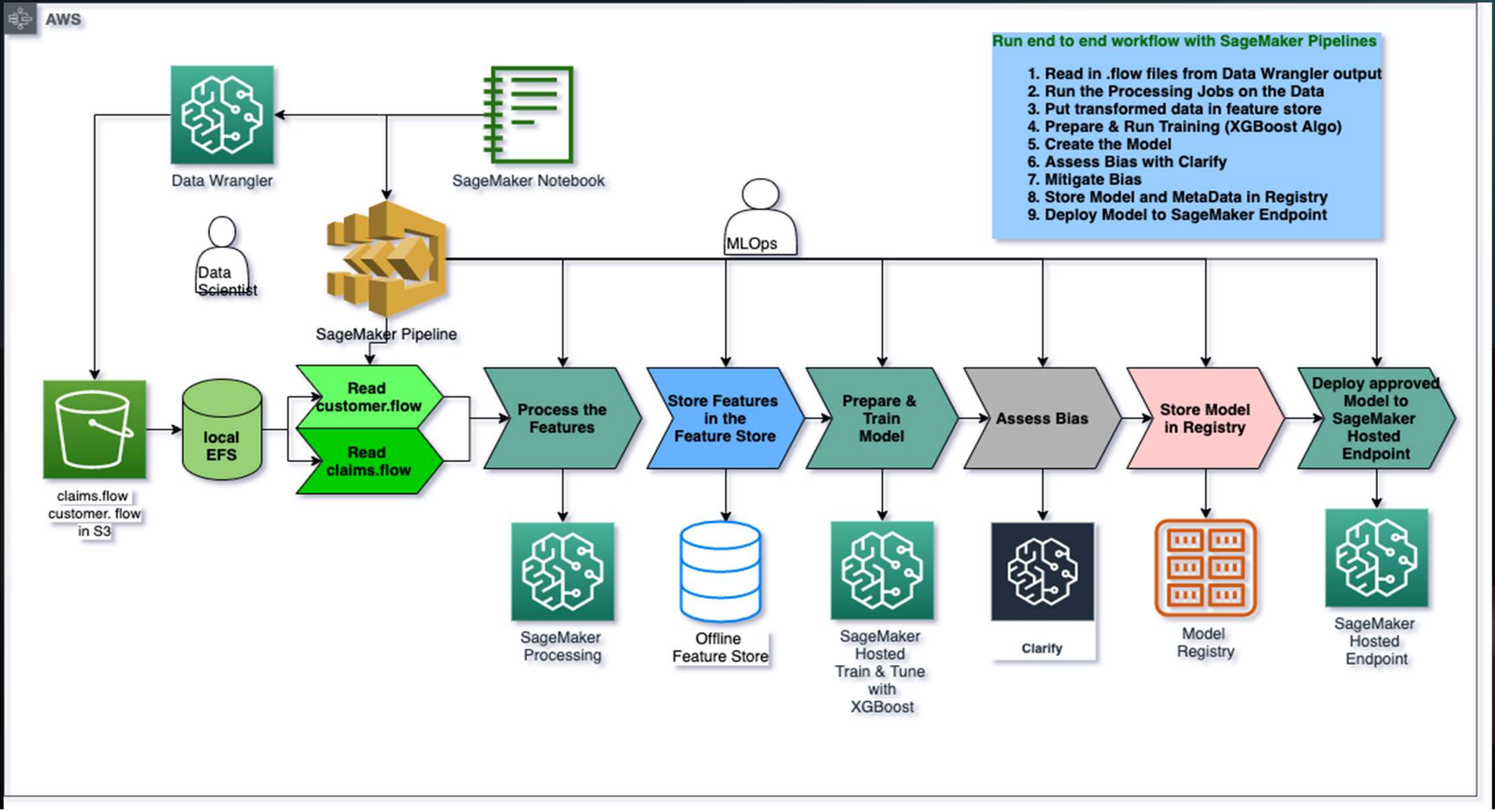
- Notebook 2: Train, Check Bias, Tune, Record Lineage, and Register a Model
  - Train a model using XGBoost, ECR
  - Model lineage with artifacts and associations
  - Evaluate the model for bias with Clarify
  - Deposit Model and Lineage in Sagemaker Model Registry

- Notebook 3: Mitigate Bias, Train New Model, Store in Registry
  - Develop a second model
  - Analyze the Second Model for Bias
  - View Results of Clarify Bias Detection Job
  - Configure and Run Clarify Explainability Job
  - Create Model Package for second trained model

- Notebook 4: Deploy Model, Run Predictions
  - Deploy an approved model and Run Inference via Feature Store
  - Create a Predictor
  - Run Predictions from Online FeatureStore

- Notebook 5: End to end pipeline

Fraud Detection with ML – AWS Architecture



Run end to end workflow with SageMaker Pipelines

1. Read in .flow files from Data Wrangler output
2. Run the Processing Jobs on the Data
3. Put transformed data in feature store
4. Prepare & Run Training (XGBoost Algo)
5. Create the Model
6. Assess Bias with Clarify
7. Mitigate Bias
8. Store Model and MetaData in Registry
9. Deploy Model to SageMaker Endpoint

Fraud Detection with ML – Demo

I will show these critical AWS components & services in demo setup.

- Dataset – Claim & Customer data sets from Car Insurance claim an detect the fraud claim
- Sagemaker Studio – secured set ML environment
- Instance types, data volume
- Data Wrangler to process the data
- Clarify to check bias in data
- Feature store, Model registry
- Sagemaker Pipeline service
- Monitoring Workflow
- Fraud Detector endpoint