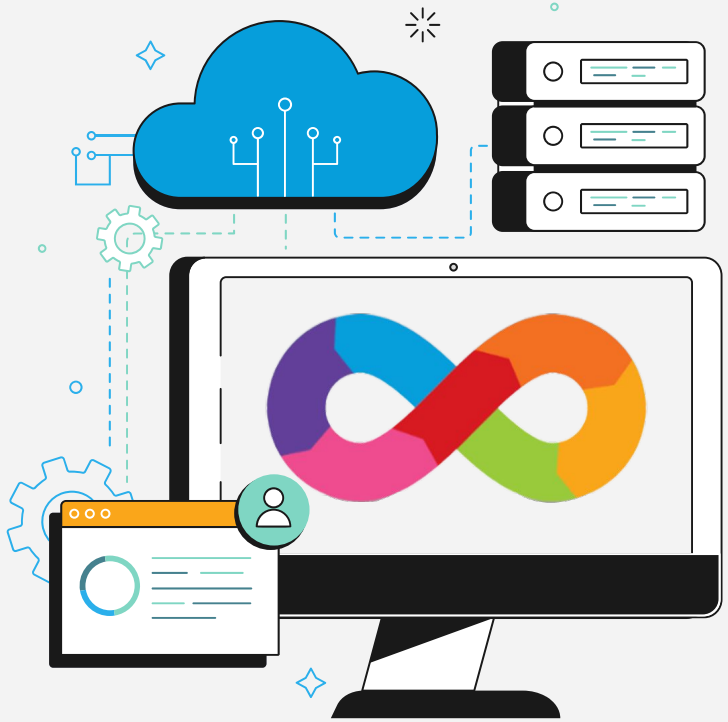
Introduction to DevOps

@ IBA - SMCS

Week 14 - 2

Disaster Recovery & MLOPs



Obaid ur Rehman
Software Architect / Engineering Manager @ Folio3

What is a Disaster Recovery

- Any event that has negative impact on a business continuity or finance is a Disaster.
- Disaster Recovery (DR) is all about recovering from a disaster and resume normal operations.
- Let's discuss two terms: RPO & RTO

RPO & RTO

RTO

The recovery time objective (RTO) is the targeted duration of time between the event of failure and the point where operations resume.

RPO

A recovery point objective (RPO) is the maximum length of time permitted that data can be restored from, which may or may not mean data loss.

RPO & RTO

Business continuity

How much data can you afford to recreate or lose?

How quickly must you recover? What is the cost of downtime?

Source: https://docs.aws.amazon.com/wellarchitected/latest/reliability-pillar/disaster-recovery-dr-objectives.html

Disaster Recovery Strategies

- Backup and Restore
- Pilot Light
- Warm Standby
- Hot site / Multi site approach

RTO & RPO of Disaster Recovery Strategies

Backup & Restore	Pilot Light	Warm Standby	Multi Site
RTO/RPO: Hours	**RTO/RPO: 10s of Minutes**	**RTO/RPO: Minutes**	**RTO/RPO: Real-Time**
• Lower priority use cases • Solutions: Cloud Storage, Backup Solutions • Cost: $ to $$	• Lower RTO/RPO requirements • Solutions: Database Service, Replication Solutions • Cost: $$	• Core Applications and Services • Solutions: Cloud Storage, Database Service, Replication Solutions • Cost: $$$	• Mission Critical Applications and Services • Solutions: Database Service, Replication Solutions • Cost: $$$$

Backup & Restore

- Cheapest and oldest method of DR Strategy.
- High RPO & possibly High PTO.
- Simple case could be: You take backup of database. All infrastructure is IaC. Recovery would be a manual process.

Pilot Light

- In this approach, we basically keep a minimal set of resources running in another environment or region, which will be ready to go live. We can spin it up when it’s needed.

- Whatever takes longest to spin up is kept in stand by mode and rest of the infrastructure we spin when needed. For example, database takes the longest to spin up in another env or to restore its backup, then that needs to be kept up with a small machine but in sync with the prod data, so, when needed, we scale the system up and it’s ready to go live. Rest of the infrastructure can be spin up using IaC.

Pilot Light

- It’s cheap to run this minimal infrastructure (+operational costs) and we are not paying for the rest of the infrastructure, until we need them.

Warm Standby

- Kind of similar to Pilot light but additional services are duplicated.
- Full System is duplicated but at a minimum size.
- Upon Disaster we scale it up.

Multi-site (active - active)

- This is the costliest approach of all the solutions, but have the shortest RTO/RPO.
- We basically have identical copies of production infrastructure which is running 24 * 7 side by side both in active status fronting by a load balancer. Best suitable for business generating critical applications.

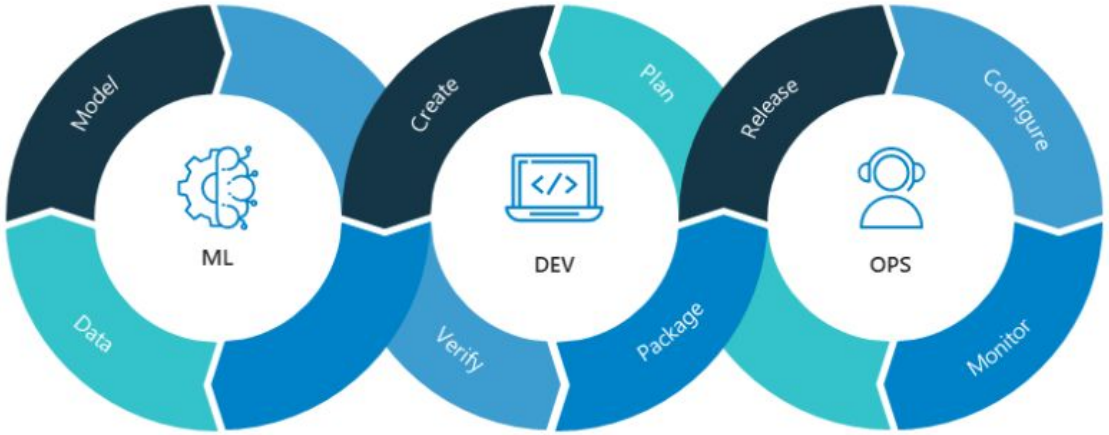
Conclusion

- Whatever the solution we choose, that needs to be tested thoroughly to ensure, that, during actual DR situation, it works.

- Monitoring tools tied up with alerts/events, using which, remediate the application back to its desired state automatically, is a very significant/elegant process, which requires a lot of testing. Automating these strategies saves a lot of time, which directly means meeting RTO/RPO SLAs.

What is MLOps

Machine learning operations (MLOps) are a set of practices that automate and simplify machine learning (ML) workflows and deployments.

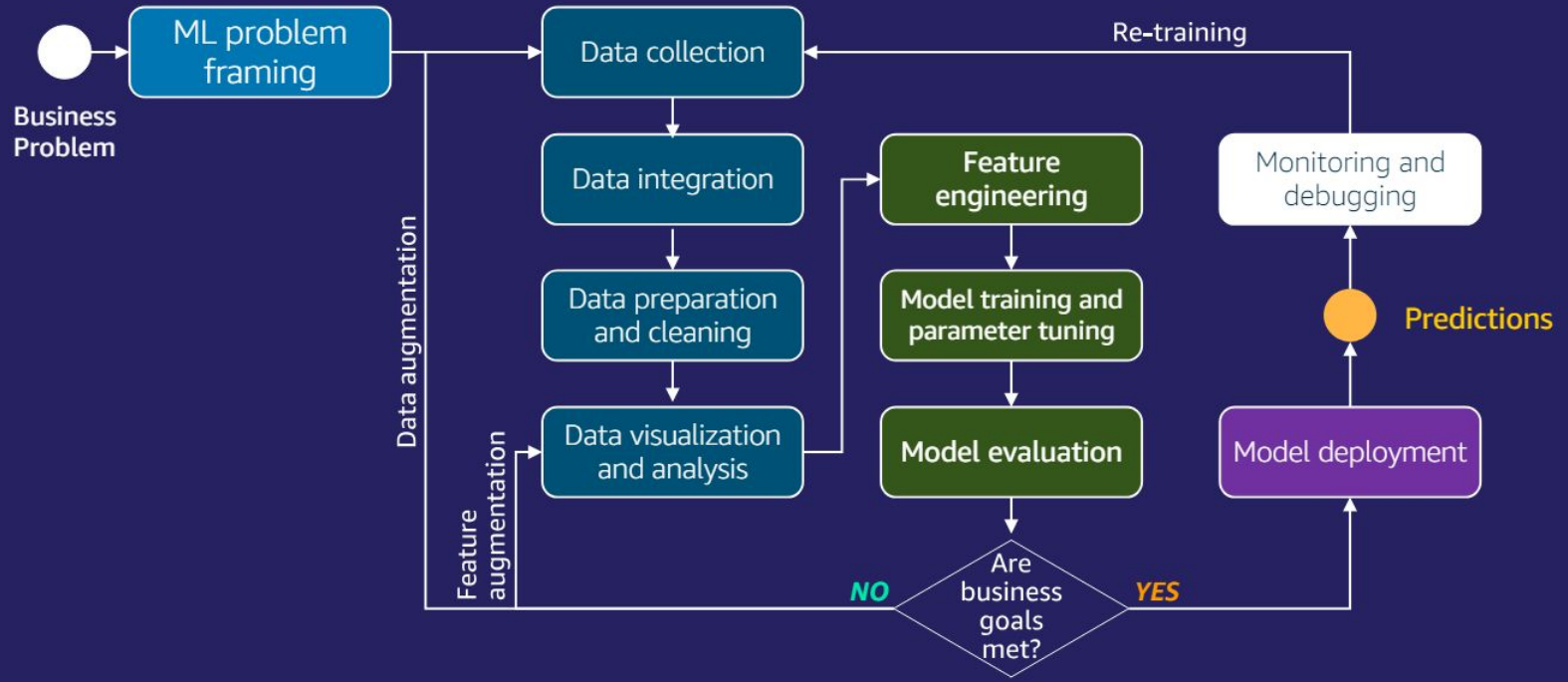


What is MLOps

ML + Dev + Ops = MLOps

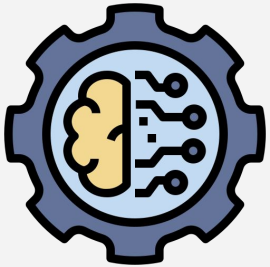
Collaborative and experimental in nature | Automate as much as possible |
Continuous improvement of ML Models | Standardize and Scale

The ML Process



The Friction

- Build
- Train
- Deploy
- Monitor
- Manage
- Re-train

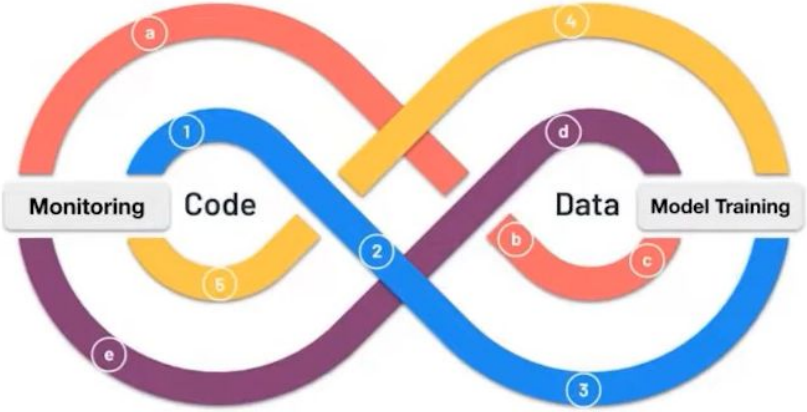


Manual Process can become bottleneck, impact productivity and become costly.

Only 53% of POCs make it into production
(Source: Gartner)

75% of organizations will shift from piloting to operationalizing AI
(Source: Gartner)

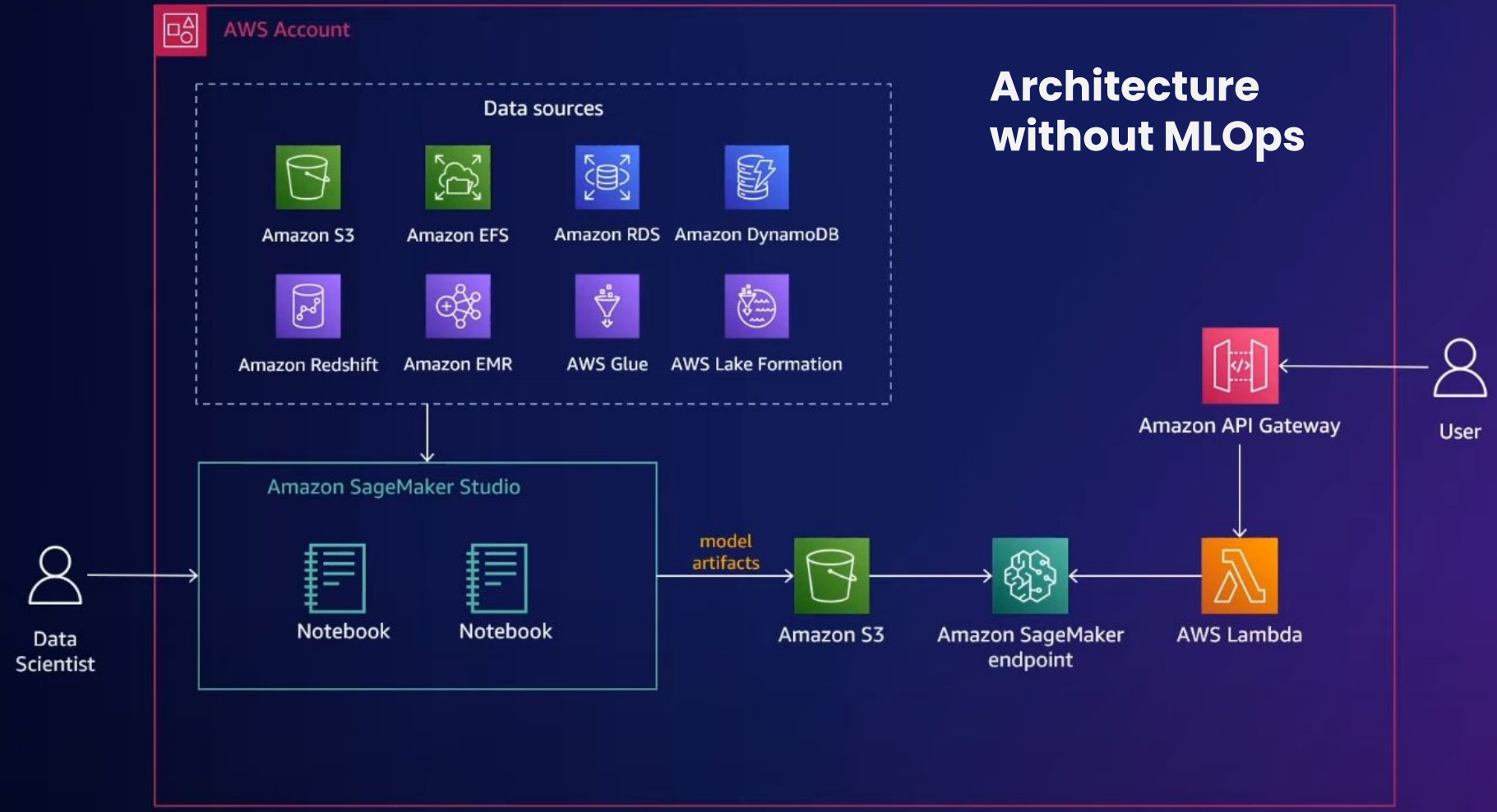
Code and Data Loops

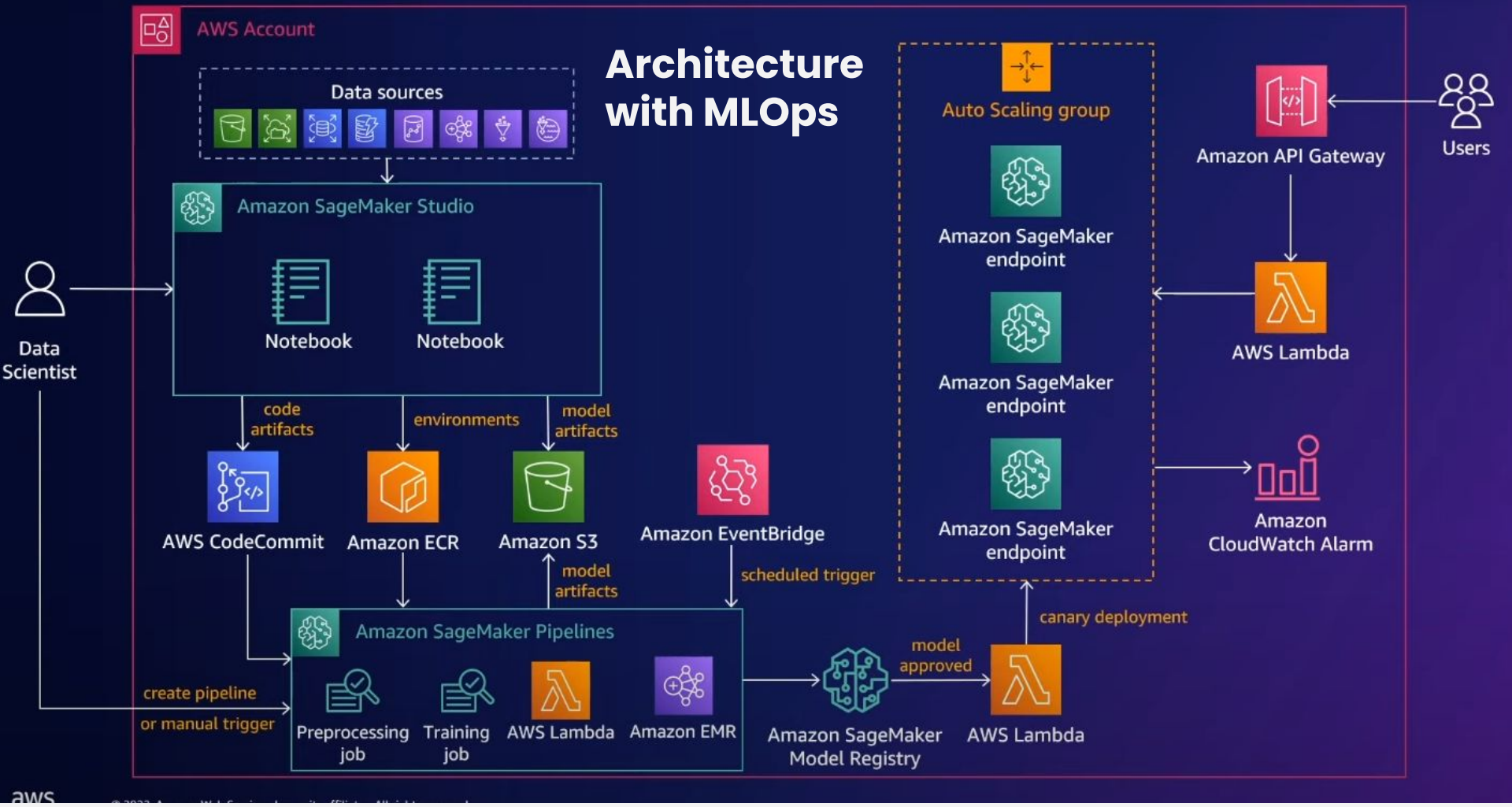


Code Loop:
Iterations on machine learning code

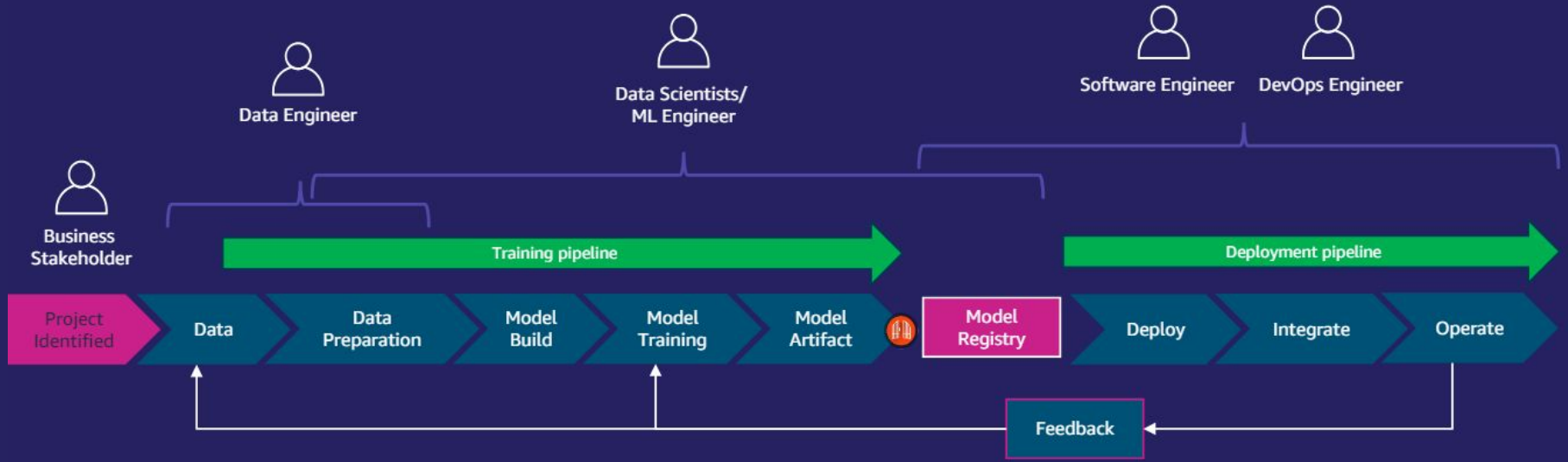
Data Loop:
Developments on data

Dev	1 Define opportunity and desired outcomes	a Identify and collect data from sources
	2 Design and prototype ML techniques	b Clean and label data
	3 Model training, tuning, and selection	c Explore, prepare, and split data
Prod	4 Testing, integration, and productionization	d Evaluate test data performance
	5 Deployment, system and prediction monitoring	e Training vs. production data monitoring


AWS Account
Architecture without MLOps
Data sources
Amazon S3
Amazon EFS
Amazon RDS
Amazon DynamoDB
Amazon Redshift
Amazon EMR
AWS Glue
AWS Lake Formation
Amazon API Gateway
User
Amazon SageMaker Studio
Notebook
Notebook
model artifacts
Amazon S3
Amazon SageMaker endpoint
AWS Lambda
Data Scientist


AWS Account
Architecture with MLOps
Data sources
Amazon SageMaker Studio
Notebook
Notebook
Data Scientist
code artifacts
environments
model artifacts
AWS CodeCommit
Amazon ECR
Amazon S3
Amazon EventBridge
model artifacts
scheduled trigger
Amazon SageMaker Pipelines
create pipeline or manual trigger
Preprocessing job
Training job
AWS Lambda
Amazon EMR
Amazon SageMaker Model Registry
model approved
AWS Lambda
canary deployment
Auto Scaling group
Amazon SageMaker endpoint
Amazon SageMaker endpoint
Amazon SageMaker endpoint
Amazon API Gateway
Users
AWS Lambda
Amazon CloudWatch Alarm

ML Ops Practices



DevOps V/s MLOps

	DevOps	MLOps
Main Purpose	Automation of software processes like quality assurance and feedback loops	Standardize Machine Learning Lifecycle by processes and automated quality checks
Deployment cycles	Frequent stepwise iterations	Continuous, long training cycles
Team composition	Software Development + DevOps + QA Team	Machine Learning Engineers + Data Engineers/Scientist
Deliverable	Code and integration	Model + Training Data + Training parameters
Objectives	Business goals	Exploration of data and Model Experiments

Whats next: LLMOPs

LLMOPS: Large Language Model Ops encompasses the practices, techniques and tools used for the operational management of large language models in production environments.

End

Q&A