DevOps for Data: Spark on Containers

CDF / LF Workshops

Speaker: Muhammad Danyal "Sage" Khan

Date: August 28

Why DevOps/CD for Spark

- Business■critical workloads need CI/CD guardrails
- Immutable artifacts, promotions, rollbacks, SLOs
- Containers + CD + CDEvents

CD Blueprint

commit → CI build → tests (code+data+deps) → SBOM & sign → CD apply → observe & rollback

Packaging & Quality Gates

- One job = one image
- Tests: pytest + pandera/GE
- Security: Trivy, SBOM (Syft), Cosign
- Policy-as-code gates

Orchestration Targets

- Kubernetes (Job, CronJob, SparkApplication)

- Others: OpenShift, Nomad, managed Spark

Live Demo

1) kind-up 2) build & test 3) sbom & sign 4) deploy & logs 5) cron-deploy

Observability & SLOs

- Prometheus/Grafana metrics
- Centralized logs
- SLOs: latency, success %, cost/run

Platform Guardrails

Namespaces, quotas, secrets, cost controls, immutable images, rollbacks

Risks & Fixes

Pods Pending, OOMKilled, cold starts, config drift, data regressions

Production Mapping

Local: kind + Job

Prod: EKS/GKE/AKS + Operator/Argo + GitOps + Vault + Autoscaling