Unlocking FMOps/LLMOps using Apache Airflow : A guide to operationalizing and managing Large Language Models

Parnab Basak (he/him)

Senior Solution Architect and Apache Airflow Specialist

Amazon Web Services

Agenda

for the next 20 mins

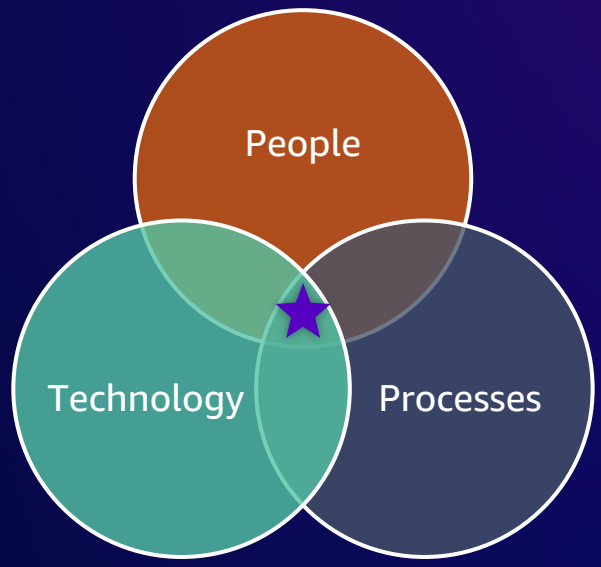
- Machine Learning (ML) Ops overview
- Machine Learning Ops foundation
- Intro to Foundation Model (FM)/Large Language Model (LLM) Ops
- Airflow for FM Ops/LLM Ops
- ML Ops vs FM/LLM Ops Differentiators
- QnA

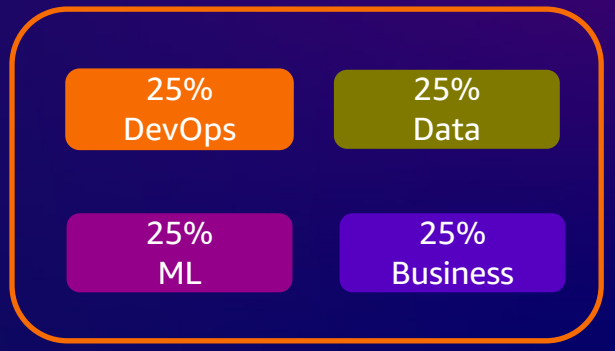
Machine Learning (ML) Ops

What is MLOps?

MLOps
Machine Learning
& Operations

The combination of people, processes, and technology to productionize ML solutions efficiently.






Productivity


Reproducibility

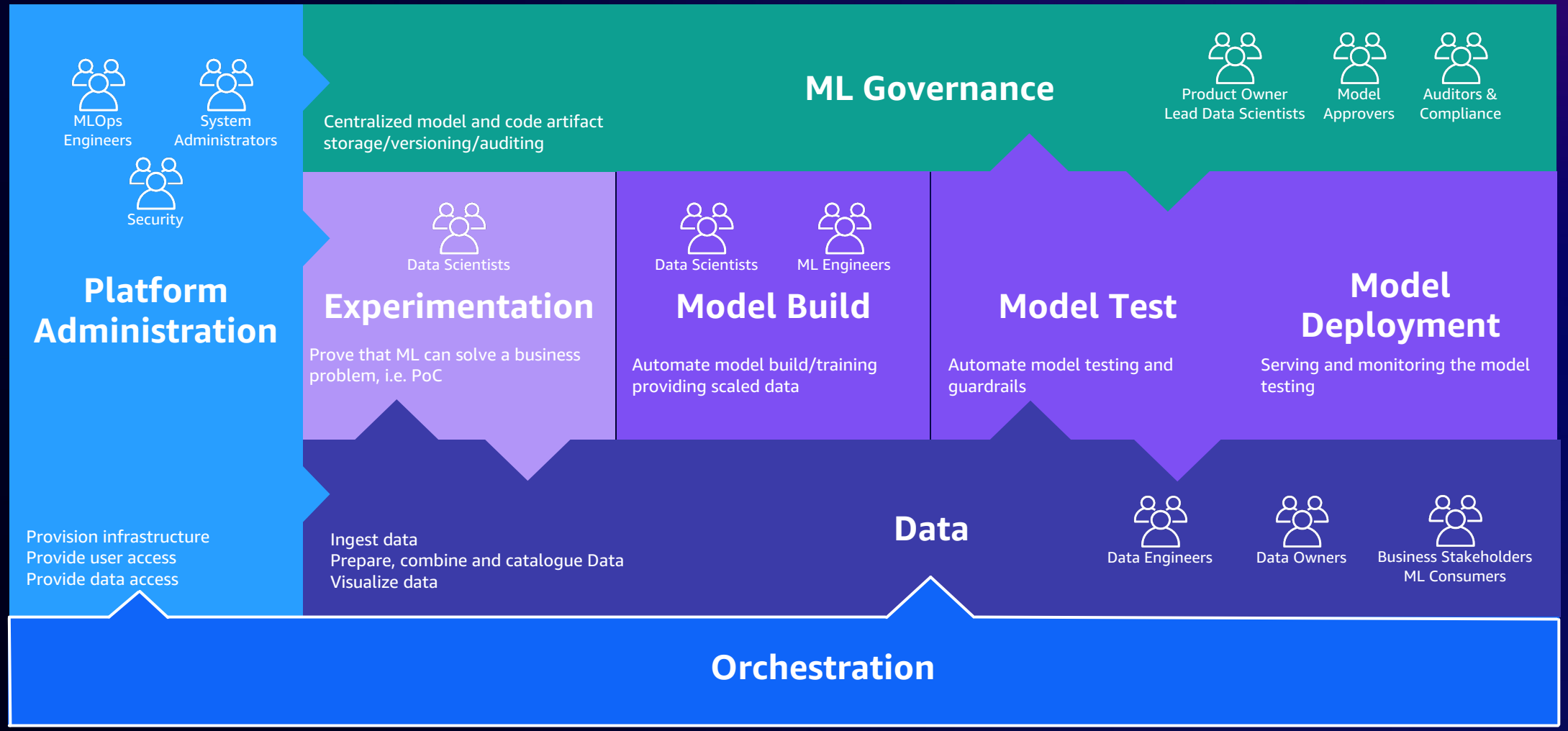

Reliability


Observability


Lower TCO

MLOps Foundation People & Processes

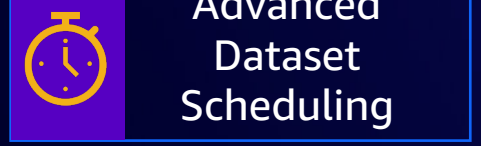
SEPARATION OF CONCERNS IS KEY FOR SUCCESS

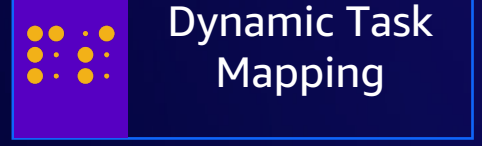


Airflow as the Orchestrator

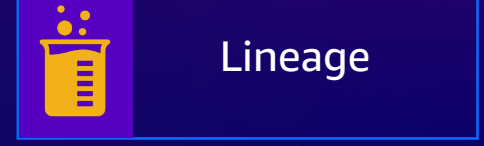
OUT OF THE BOX FEATURES TO AID MLOPS


Macros & Jinja Templates


Advanced Dataset Scheduling


Dynamic Task Mapping


TaskFlow API


Lineage


Provider Packages


Backfills


Automatic Retries


Task Group


Airflow Plugins


Setup and Teardown


Reruns

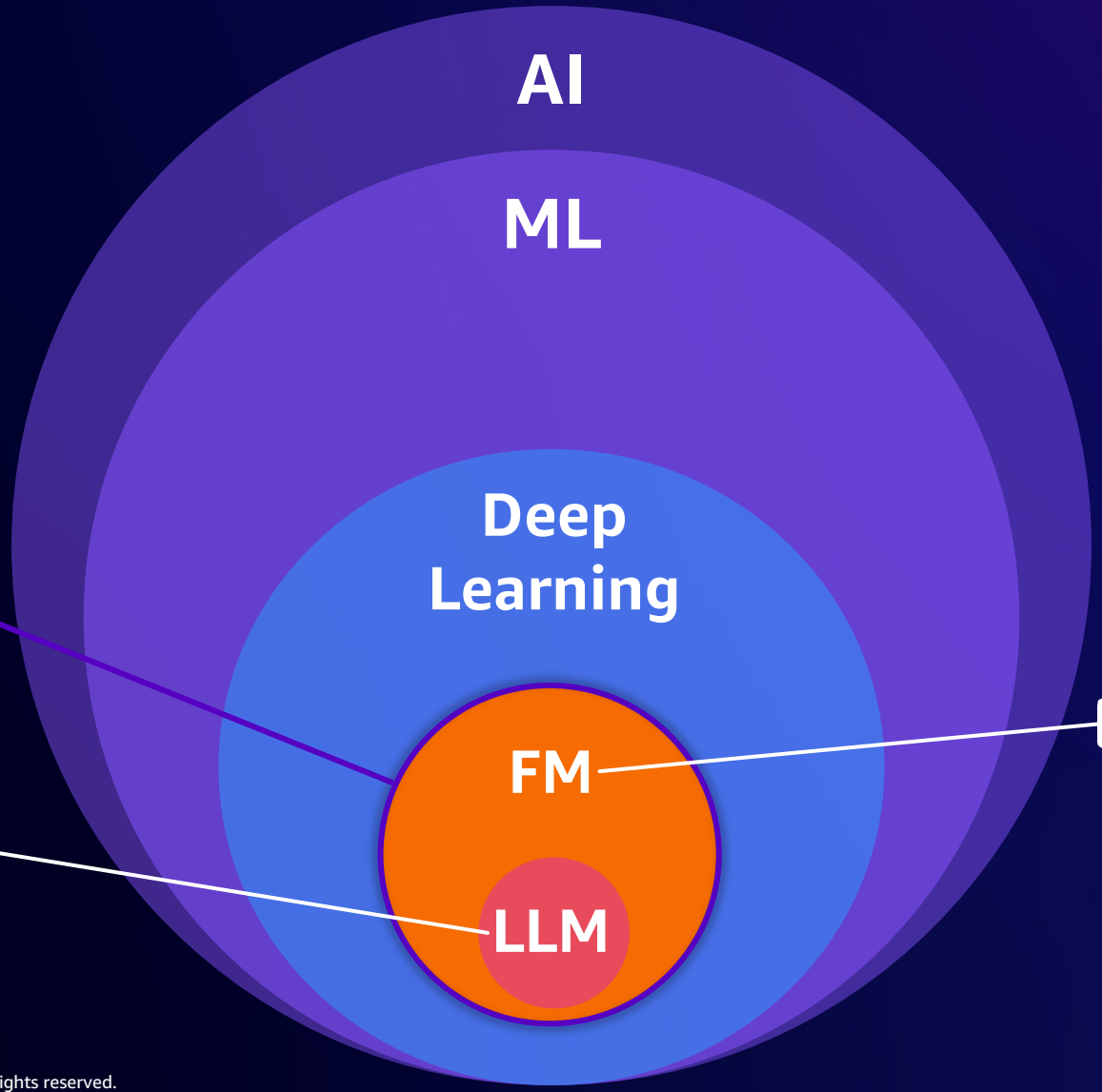

Dynamic Compute


DS Toolkit & IDE Integration


Monitoring & Alerting

Generative AI & Foundation Model Ops/Large Language Model Ops

What is Generative AI?

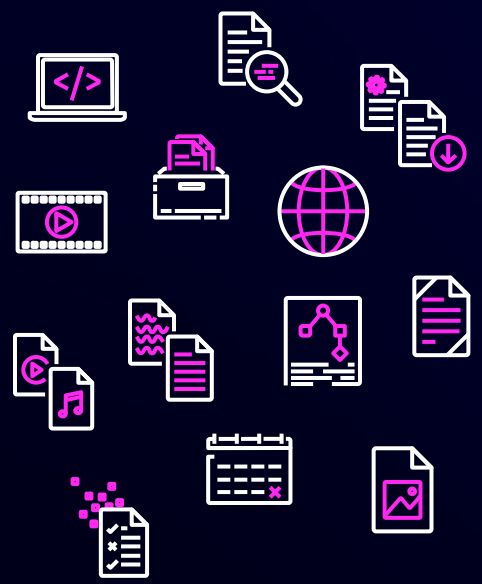


Generative AI

Large Language Models
Pre-trained Generative AI Models to understand and generate/predict human-like text

Foundation Models
Pre-trained Generative AI Models based on trillions of data for broad spectrum of tasks

Generative AI overview



Unlabeled data

Pretrain

FMs

Broad range of general tasks

Adapt

- Text generation
- Summarization
- Information extraction
- Q&A
- Chatbot

LLMs

Broad range of specific tasks

Generative AI User Types & Skills

Generative AI User Types	Providers	Fine-Tuners	Consumers
			
	Entities who pre-train FMs from scratch themselves and provide them as a product to **fine-tuner** and **consumer**.	Customize (e.g. fine-tune) **providers'** pre-trained FMs with their own data and run inference, while provide access to **consumers**.	Interact with Generative AI services from **provider** or **fine-tuner** by text prompting or visual interface to complete desired actions.
Skills	Deep **end-to-end ML**, NLP expertise and data science, labeler "squad".	Strong end-to-end ML expertise and **domain knowledge for tuning** including prompt engineering.	No ML expertise required. Focus on prompt engineering and retrieval augmented generation.

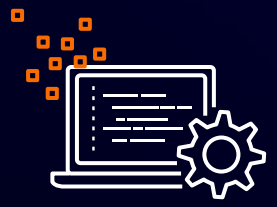
can be also

can become

Productionize large models
leveraging ML & Operations (FMOps)

Productionize applications
leveraging Generative AI & Operations

Generative AI Processes – Consumers



Select, evaluate, & use FM as a black-box & adapt context

Using multiple chained models and prompt engineering techniques to achieve context adaptation (if necessary). Expose the solution to the end users

Inputs/Outputs & Rating

Interaction with the Generative AI Solutions. Aim to improve outcomes by penalizing or rewarding Generative AI solution outputs providing insights for prompt engineering

Generative AI Processes for LLM – Consumers

Generative AI Developers & Prompt Engineers/Testers

DevOps/AppDevs

LLM-based Generative AI Solution

Backend

1. Select FM

2. Prompt Engineering

3. Test & Test Prompt lineage (input & outputs)

4. Chain Prompts & Applications

5. Input/ Output Filtering & Guardrails

6. Rating Mechanisms (thumbs up/down, rating, text)

Front-end

New Test Set

Develop & Deploy Web Application

Test Functionality

Input/ Output & Rating Interaction

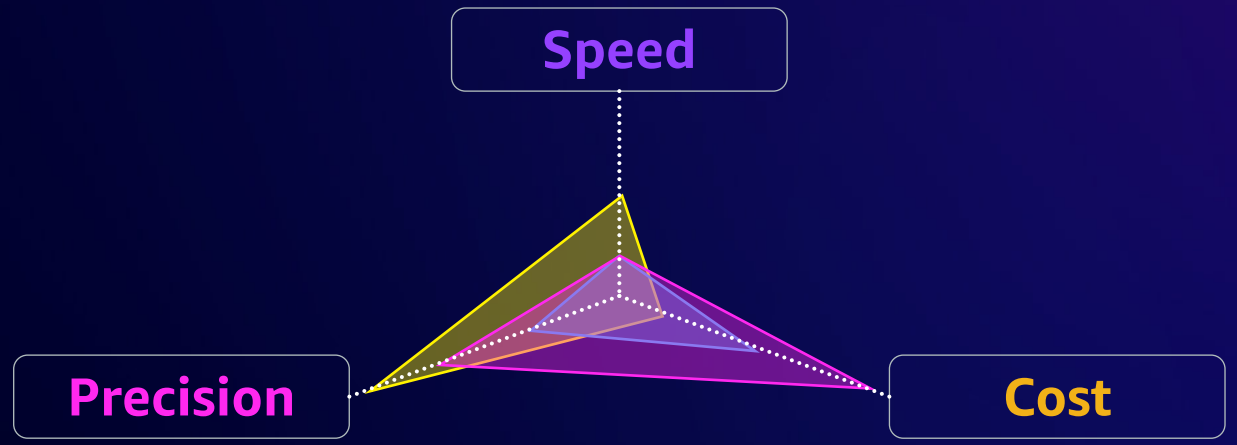
WebUI

Generative AI End-users

Use web application rate the quality of output

Model/Prompt Selection

Model Selection



Automatic evaluation

Accuracy

Robustness

Toxicity

Human evaluation

Creativity

Style

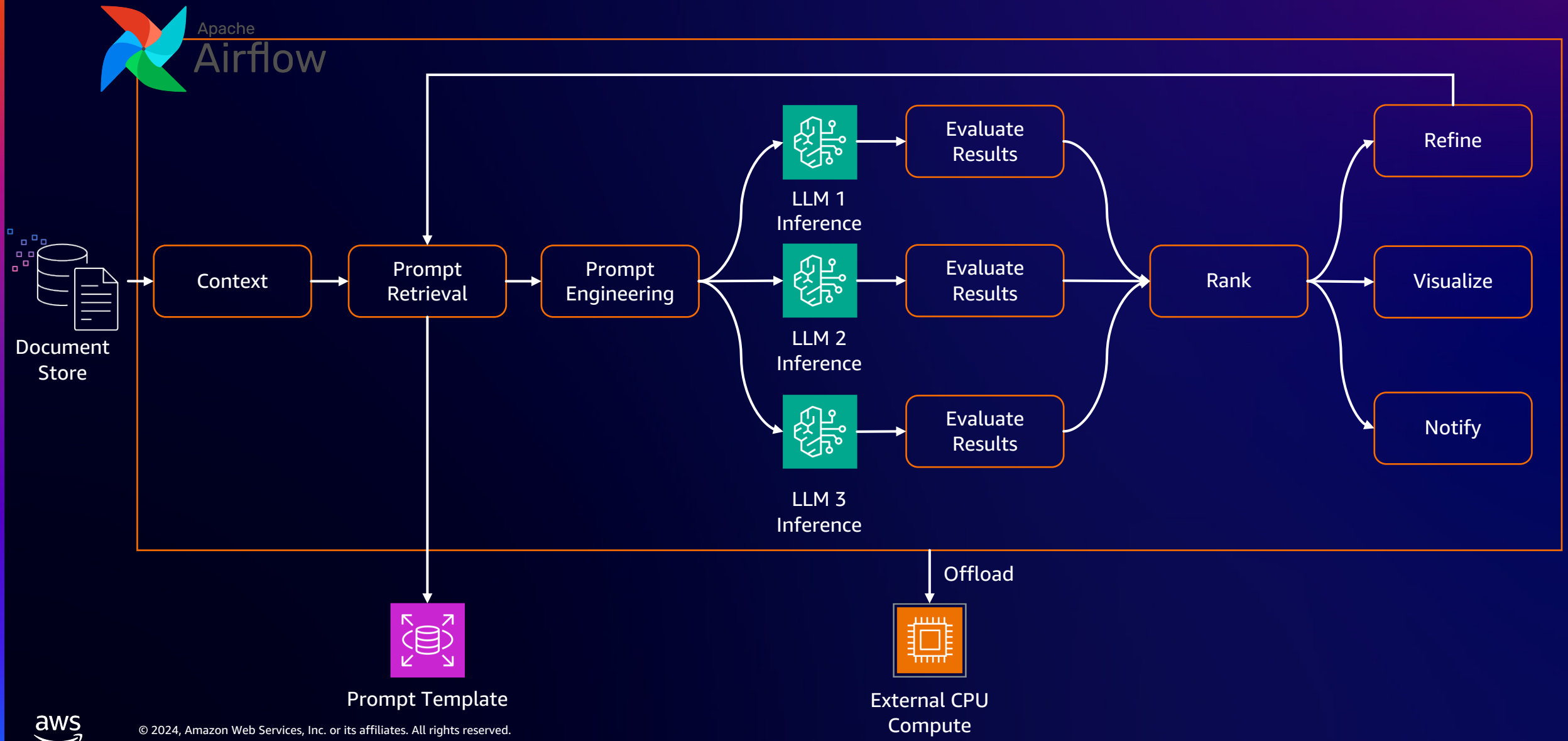
Tone

Relevance

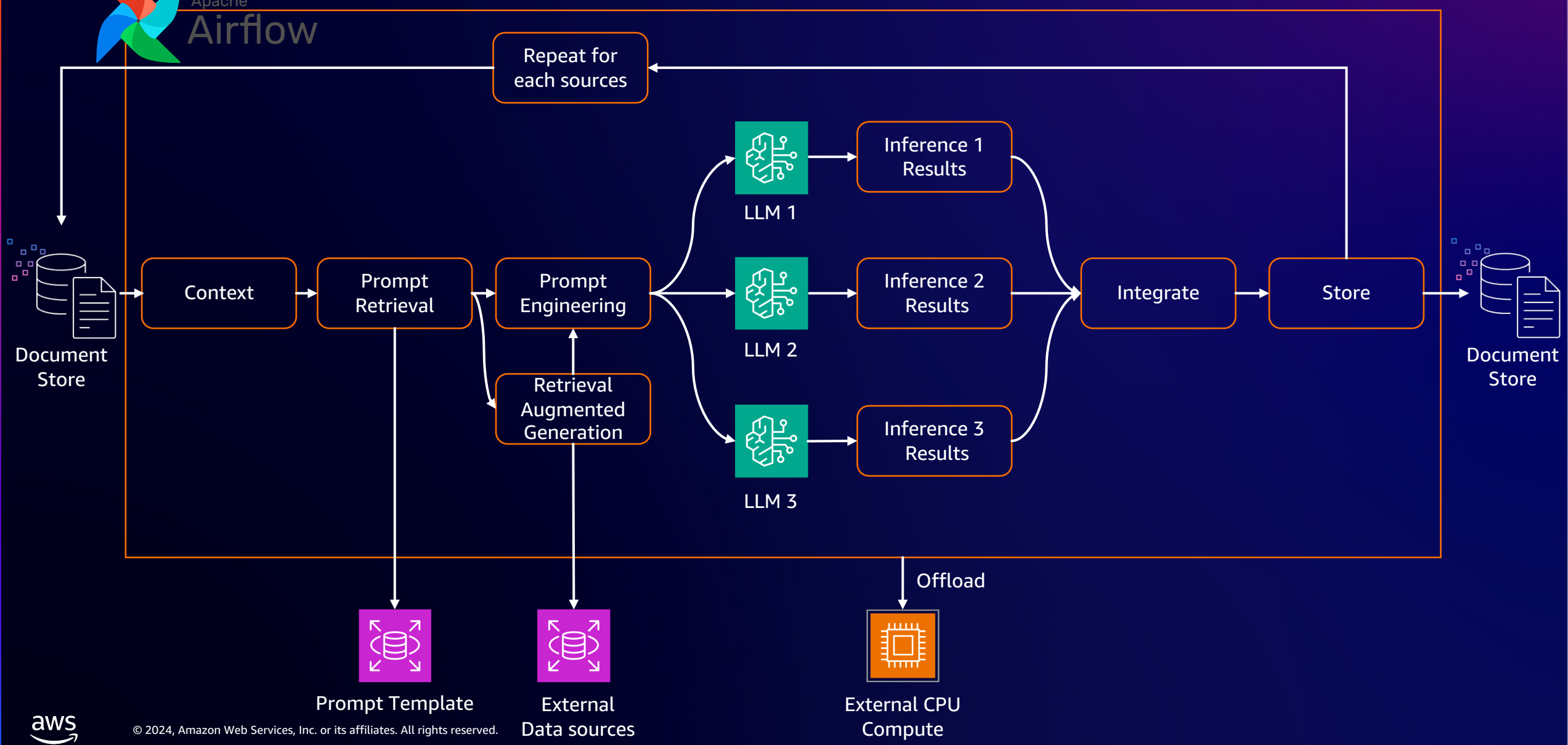
Coherence

Brand voice

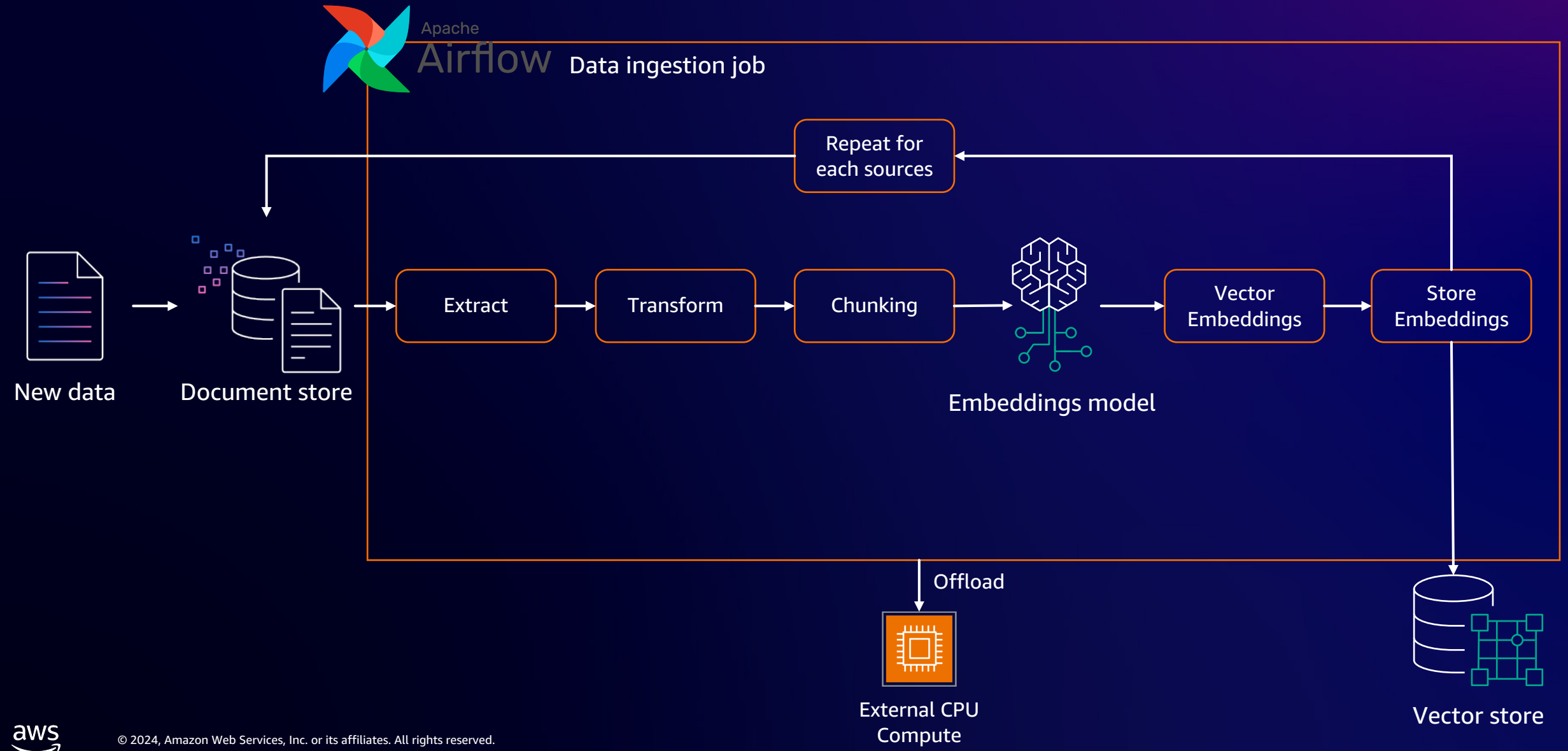
Model/Prompt Evaluation with Airflow



Offline Batch Inferencing with Airflow



RAG Data Ingestion with Airflow



Generative AI Processes - Fine-Tuners

Data Labeling

Human in the loop to label thousands, or hundreds of data

Fine-tune

Customization for specific domains

Deployment & Prompt Engineering

Trade-off between cost, precision, & latency

Chain of models

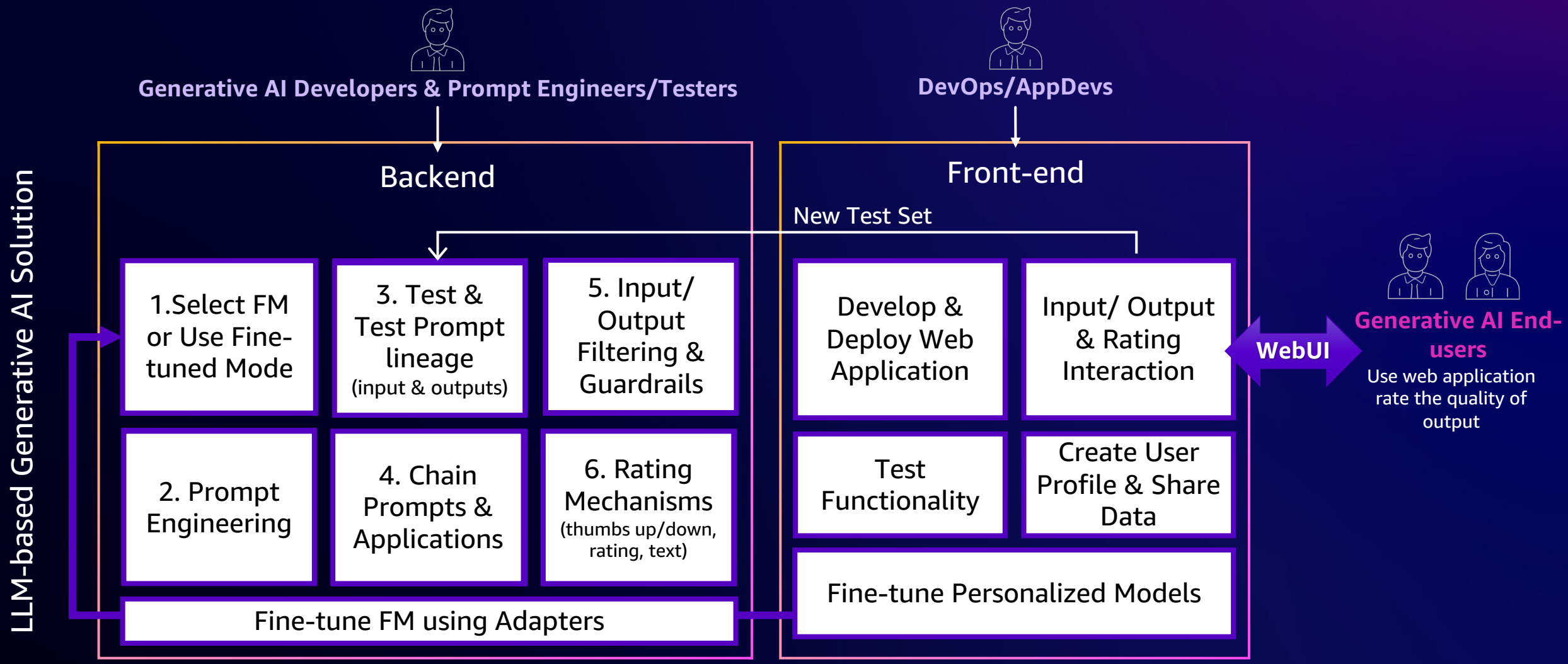
Prompt designing and engineering

Filtering input prompts and results using embeddings

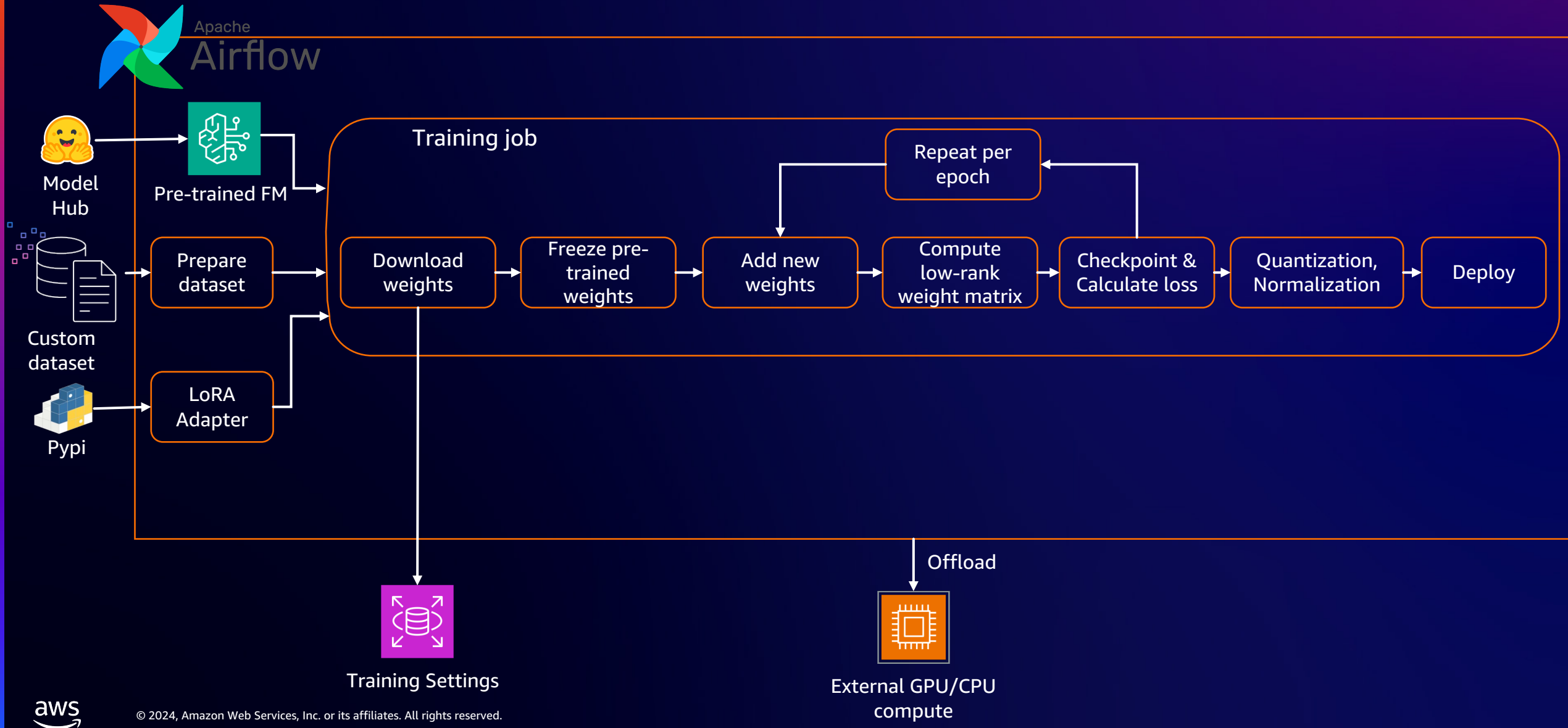
Monitor

Human in the loop feedback/rating, result similarities, toxicity rate, new methods under research

Generative AI Processes for LLM – Fine-Tuners



Fine-tuning PEFT (LoRA or QLoRA) with Airflow



Generative AI Processes - Providers

PRE-TRAINING AN FM IS NOT A TRIVIAL TASK

Massive Data Labeling

Human in the loop to label trillions of data

Weeks to Months of Training

Distribute the train of the Foundation Models

Deployment

Deploy proprietary FMs or Open Source the model artifact and code

Leaderboards

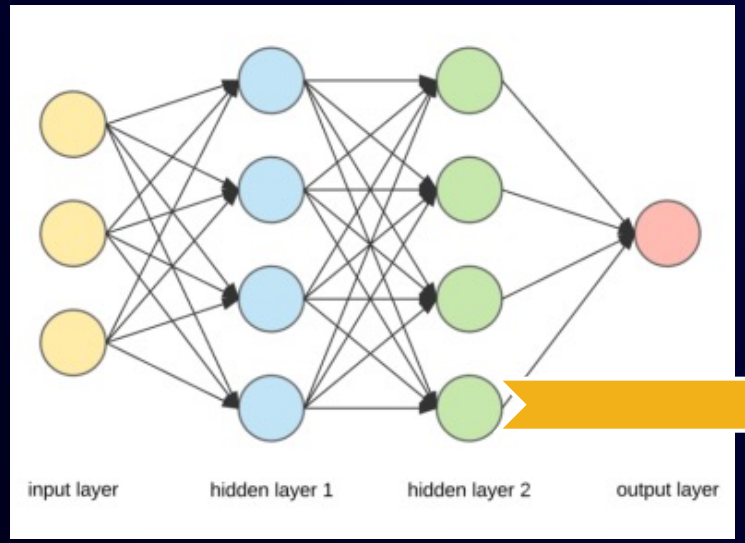
FM are being evaluated by third party organizations

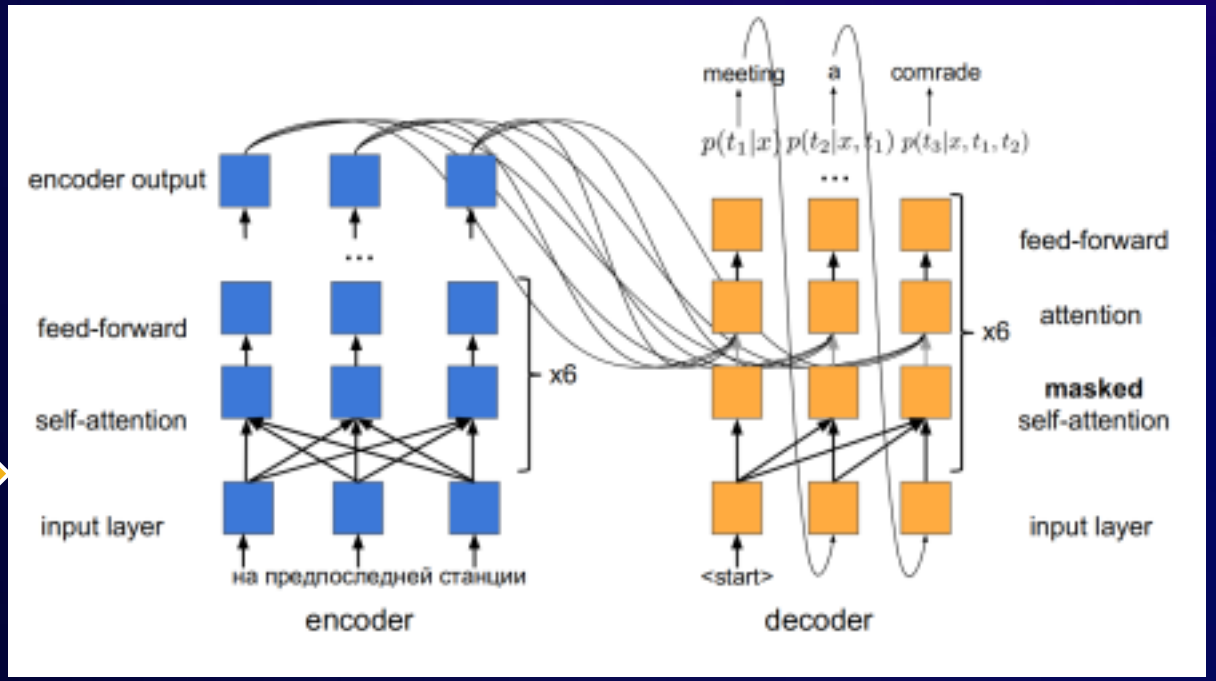
(optional)

Encoder/Decoder Transformation Model

HIGH LEVEL IMPLEMENTATION

Apache Airflow





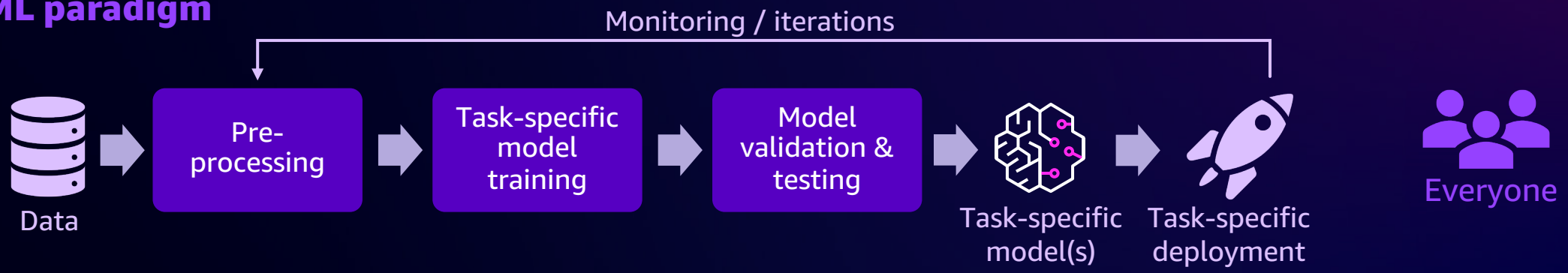
Source: https://arxiv.org/abs/1706.03762

Generative AI & MLOps

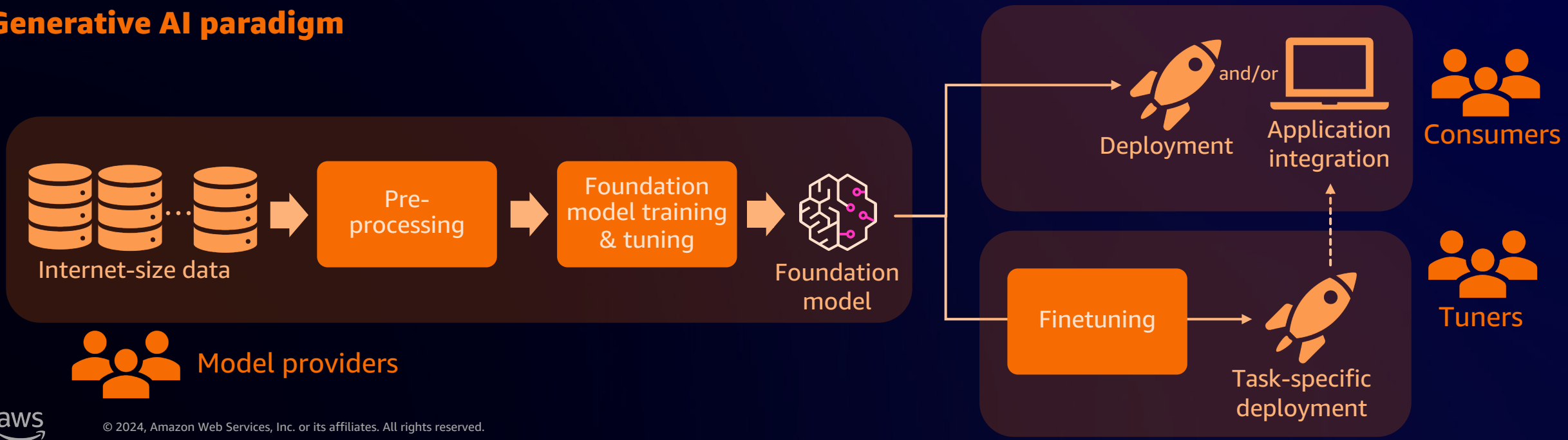
MLOps & FMOps/LLMOps Differentiators

Foundation model lifecycle

Conventional ML paradigm

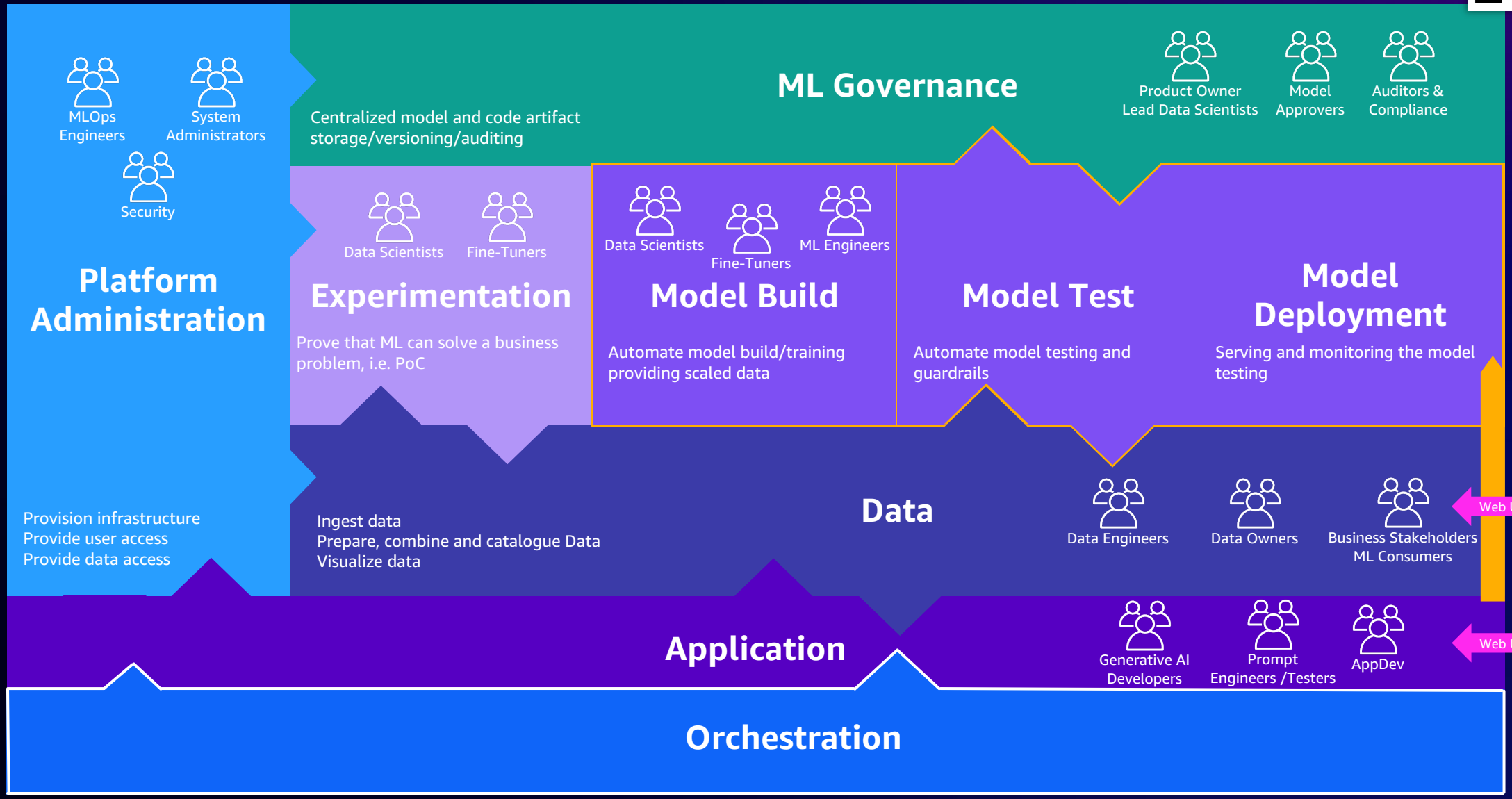


Generative AI paradigm



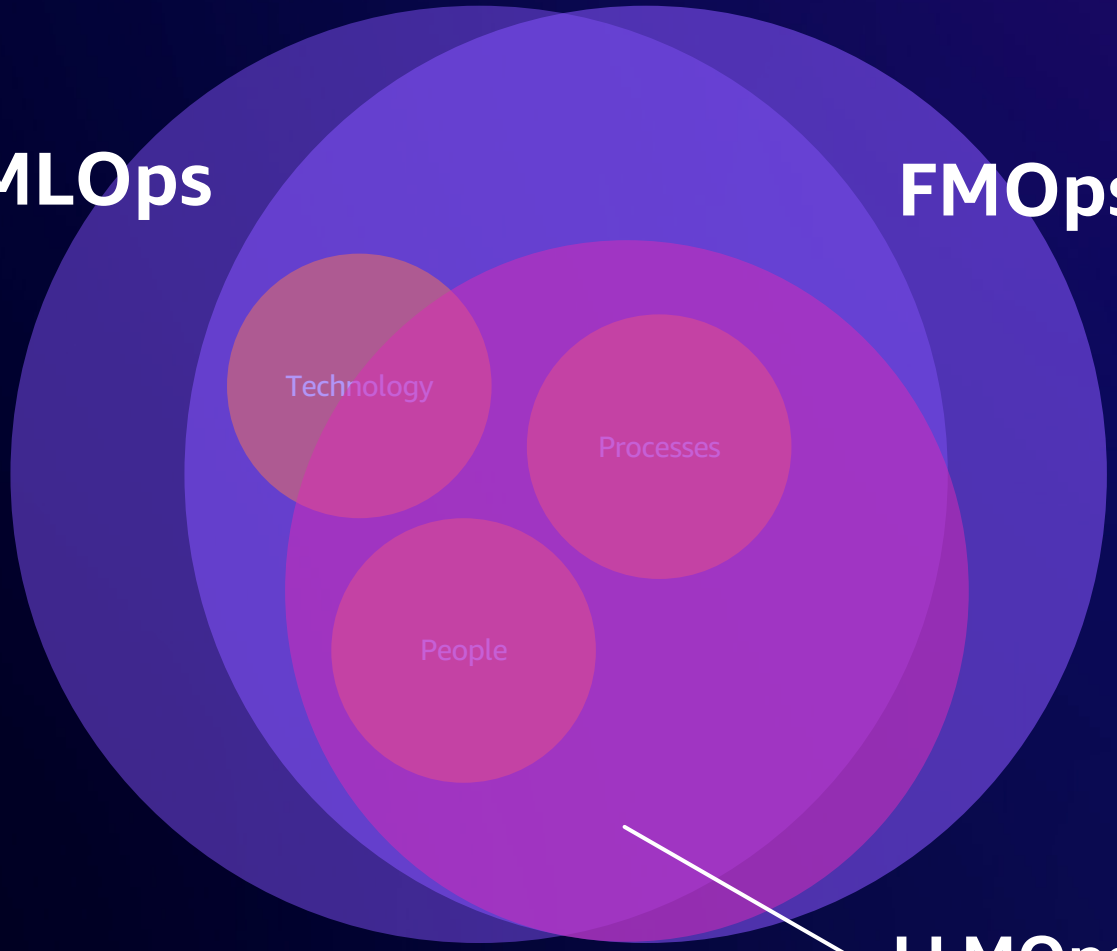
FMOps Foundation People & Processes

SEPARATION OF CONCERNS IS KEY FOR SUCCESS



MLOps/FMOps/LLMOps

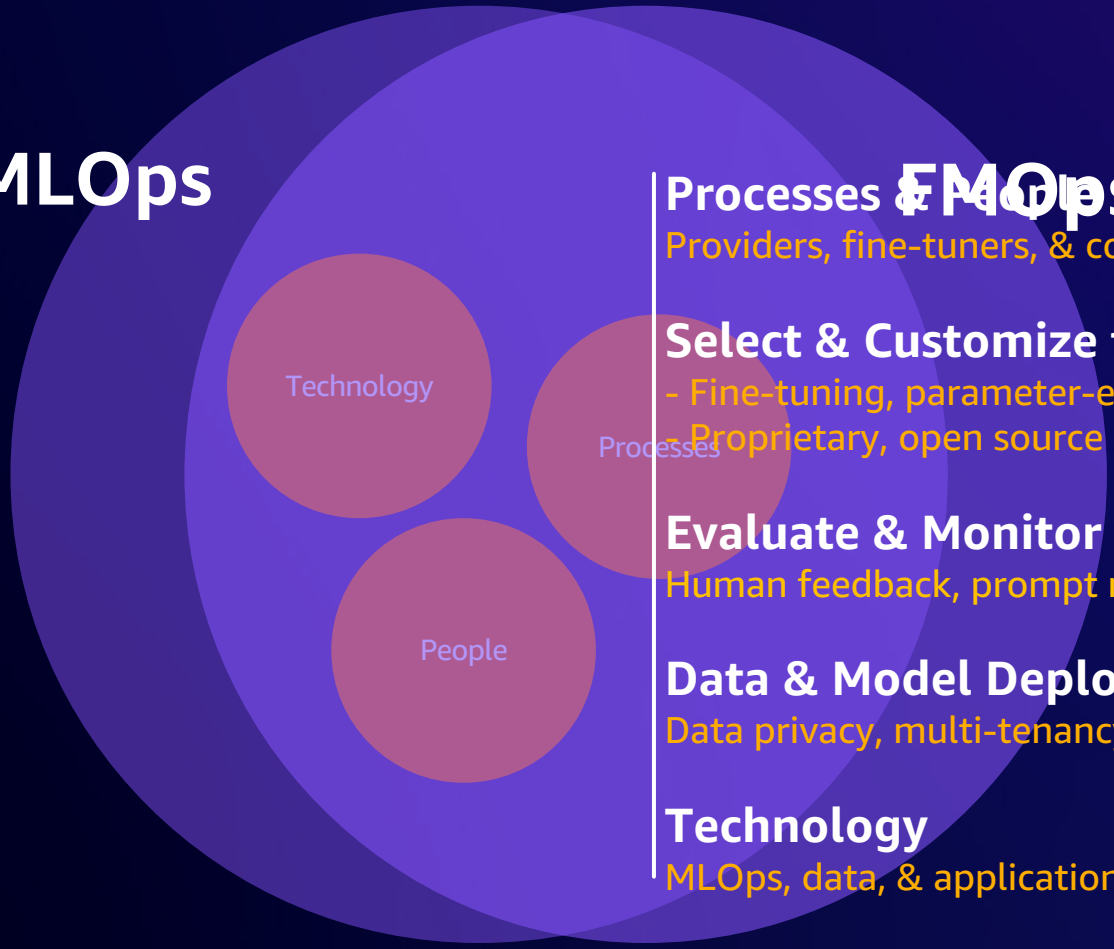
Machine learning operations
Productionize ML solutions efficiently



Foundation model operations
Productionize generative AI solutions (text-text/image/video/audio/…)

Large language model operations
Productionize large language model-based solutions

Differences between MLOps and FMOps



FMOps

Processes & People
Providers, fine-tuners, & consumers

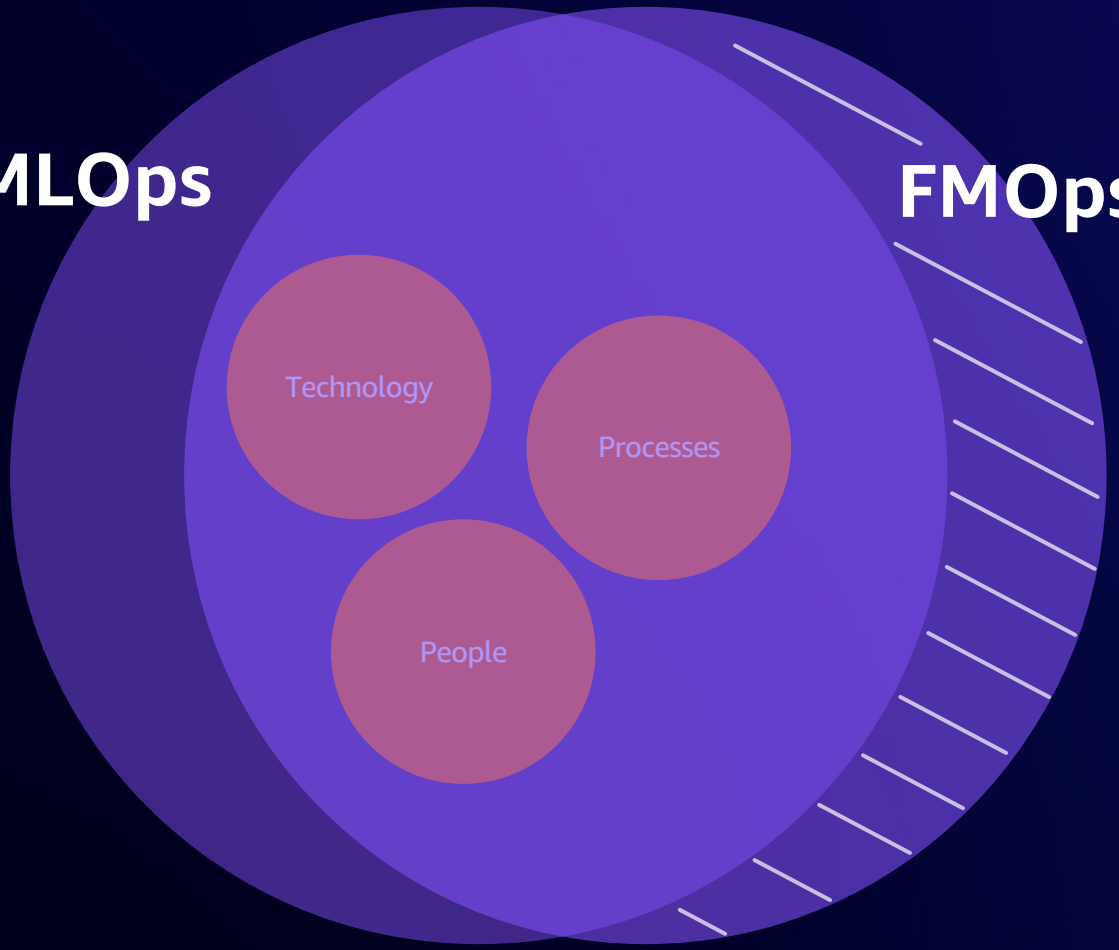
Select & Customize the FM on a Specific Context
- Fine-tuning, parameter-efficient fine-tuning, prompt engineering
- Proprietary, open source based on the application

Evaluate & Monitor Fine-tuned Models
Human feedback, prompt management, toxicity/bias…

Data & Model Deployment
Data privacy, multi-tenancy, & cost, latency, and precision

Technology
MLOps, data, & application layers

MLOps & FMOps Differentiators



Processes & People
Providers, fine-tuners, & consumers

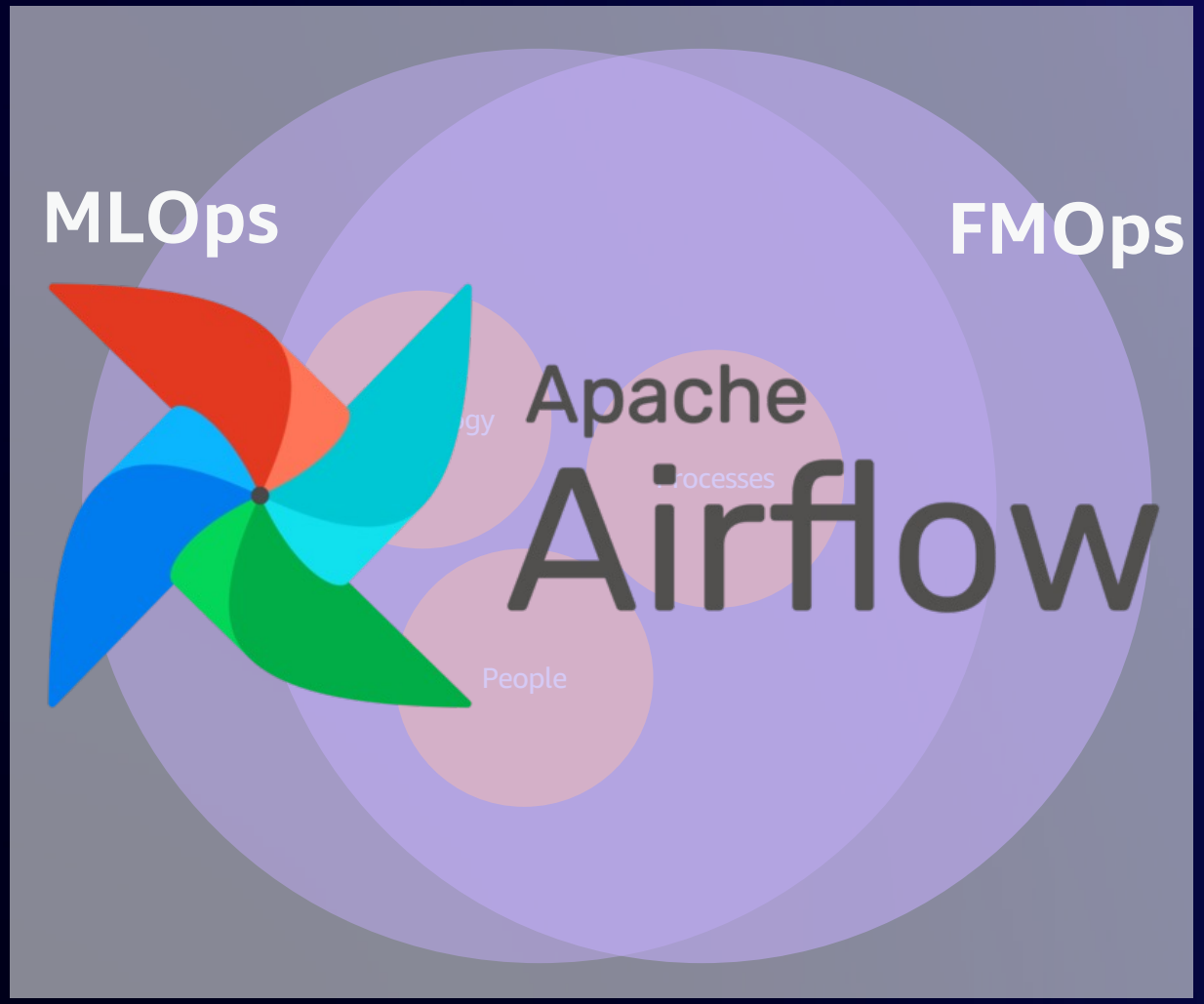
Select & Customize the FM on a Specific Context
- Fine-tuning, parameter-efficient fine-tuning, prompt engineering
- Proprietary, open source based on the application

Evaluate & Monitor Fine-tuned Models
Human feedback, prompt management, toxicity/bias…

Data & Model Deployment
Data privacy, multi-tenancy, & cost, latency, and precision

Technology
MLOps, data, & application layers

Orchestration with Apache Airflow



Workflows as Code

- Dynamic, Extensible & Flexible
- Data and Compute Agnostic
- Scalability & Reliability
- Rich Ecosystem
- Community Driven
- Continuous Innovation

Generative AI & Operations Resources

Scaling AI Workloads with Apache Airflow
Today 4:35 PM @Elizabethan A+B

FMOps/LLMOps: Operationalize generative AI and differences with MLOps
https://aws.amazon.com/blogs/machine-learning/fmops-llmops-operationalize-generative-ai-and-differences-with-mlops

Operationalize LLM Evaluation at Scale using Amazon SageMaker Clarify and MLOps services
https://aws.amazon.com/blogs/machine-learning/operationalize-llm-evaluation-at-scale-using-amazon-sagemaker-clarify-and-mlops-services

Build an internal SaaS service with cost and usage tracking for foundation models on Amazon Bedrock
https://aws.amazon.com/blogs/machine-learning/build-an-internal-saas-service-with-cost-and-usage-tracking-for-foundation-models-on-amazon-bedrock/

Session Survey

https://pulse.aws/survey/UUNNEDF8

Questions ?

Thank you!

Please complete the session survey in the mobile app