Current State and Challenges of the SAP GenAI Hub for LLMs

Lize Cai, SAP Labs Singapore

THE BEST RUN SAP

Intro

- SAP Labs Singapore
- SAP AI Core Team
- Open Source Champions Network
- Open Source Contributor
- KServe Committer

Lize Cai
Senior Software Engineer in SAP

Agenda

1. SAP GenAI Hub Overview
2. SAP AI Core Overview
3. Challenges of Serving LLMs
4. Open Source Solutions

AI is most valuable when it is **operationalized at scale**.

Manasi Vartak (2022)
in Harvard business Review

Source: Vartak, M. (2022). How to Scale AI in Your Organization. In Harvard Business Review.

The reality of today's

Organizational AI Challenges

Pace of Innovation

Requires organizations to continually adapt to new tools, methodologies, and frameworks.

Resources

High resource requirements for onboarding AI models and managing orchestration tasks such as authorizations, metering, and monitoring.

Time

Big portion of AI engineers' time is dedicated to preparatory tasks like selecting and fine-tuning LLMs or engineering prompts.

Productivity

Lacking empowerment to fully utilize AI, often results in low productivity and value realisation from AI initiatives

Operationalization

Gap between AI development and operational deployment due to lack of frameworks, tools, and integration.

Our Business AI is embedded across the portfolio

Relevant | Reliable | Responsible

Joule

A copilot that truly understands your business

Embedded AI capabilities

- SAP Cloud ERP
- SAP Supply Chain Management
- Human Capital Management
- Spend Management and SAP Business Network
- SAP Customer Relationship Management
- SAP Business Technology Platform

Customized AI

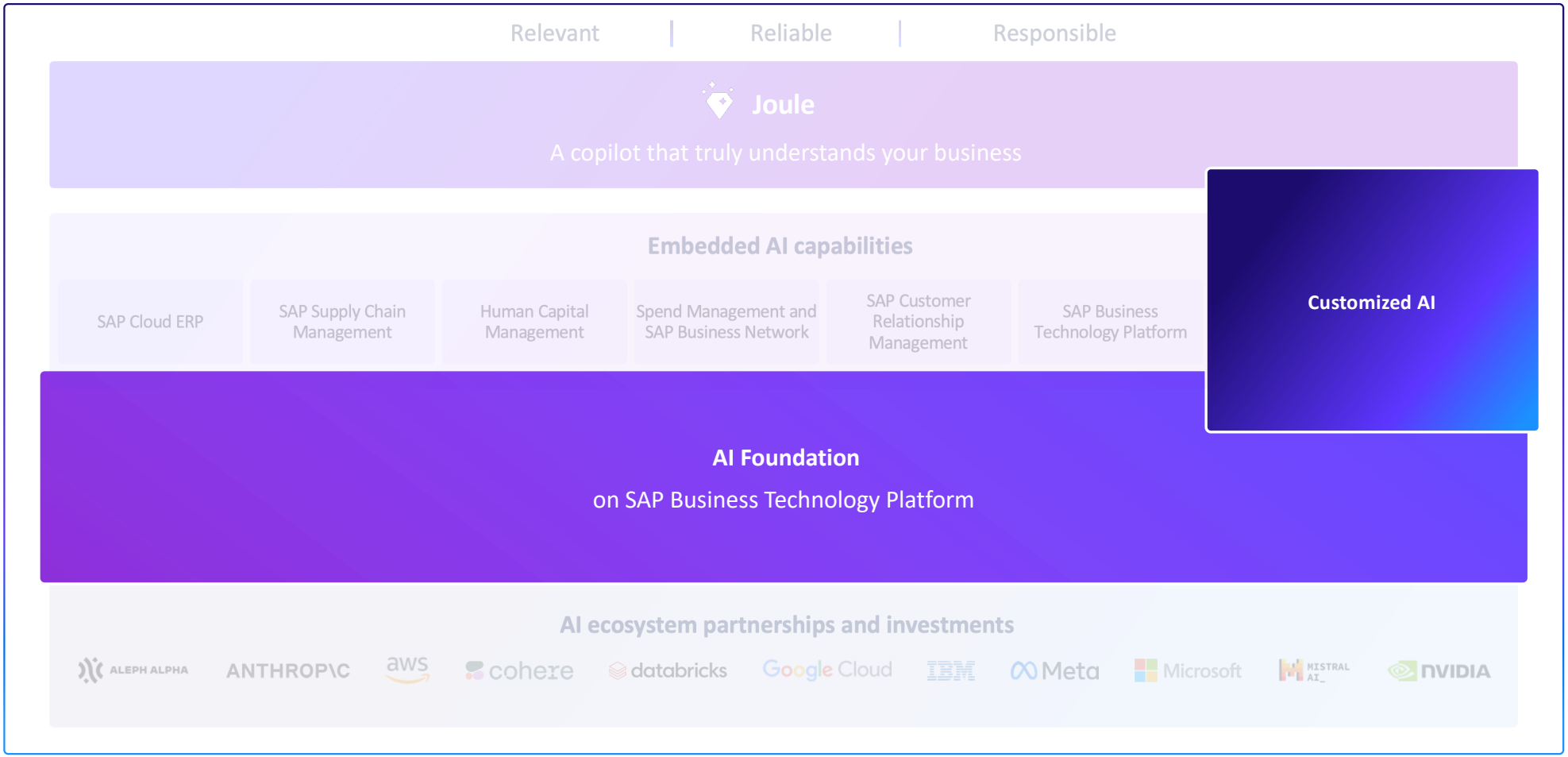
AI Foundation

on SAP Business Technology Platform

AI ecosystem partnerships and investments

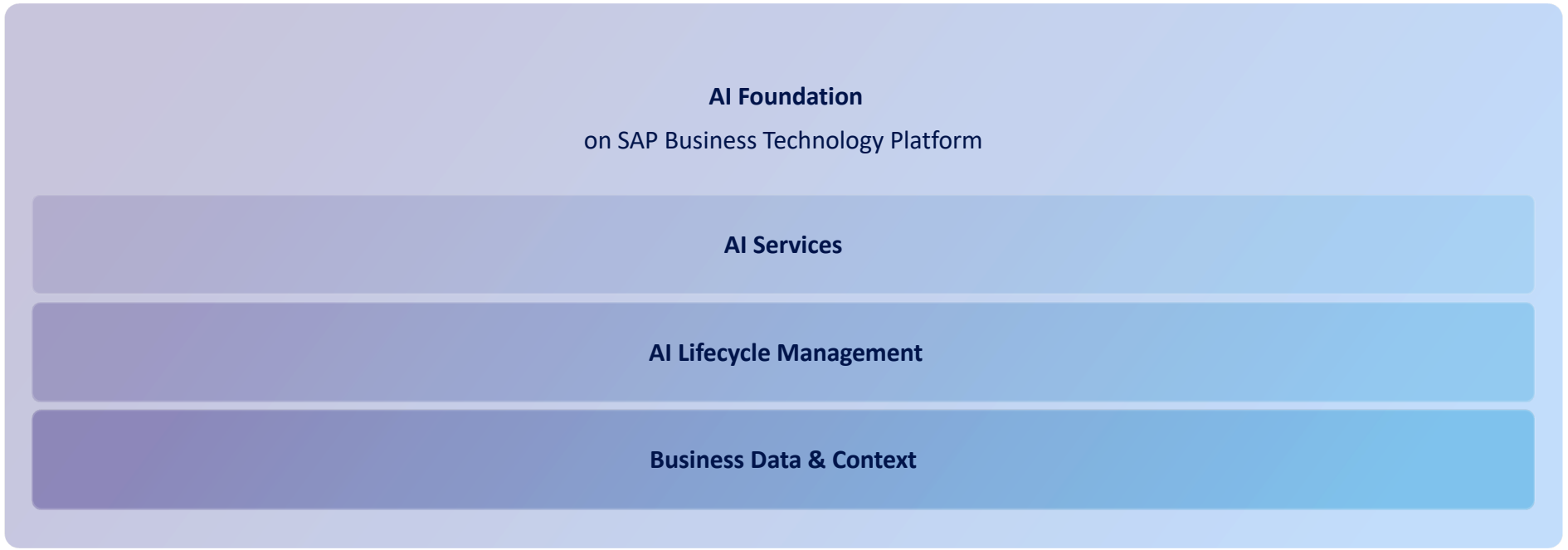
ALEPH ALPHA · ANTHROP\C · aws · cohere · databricks · Google Cloud · IBM · Meta · Microsoft · MISTRAL AI_ · NVIDIA

AI Foundation is a complete set of services for AI developers on SAP BTP

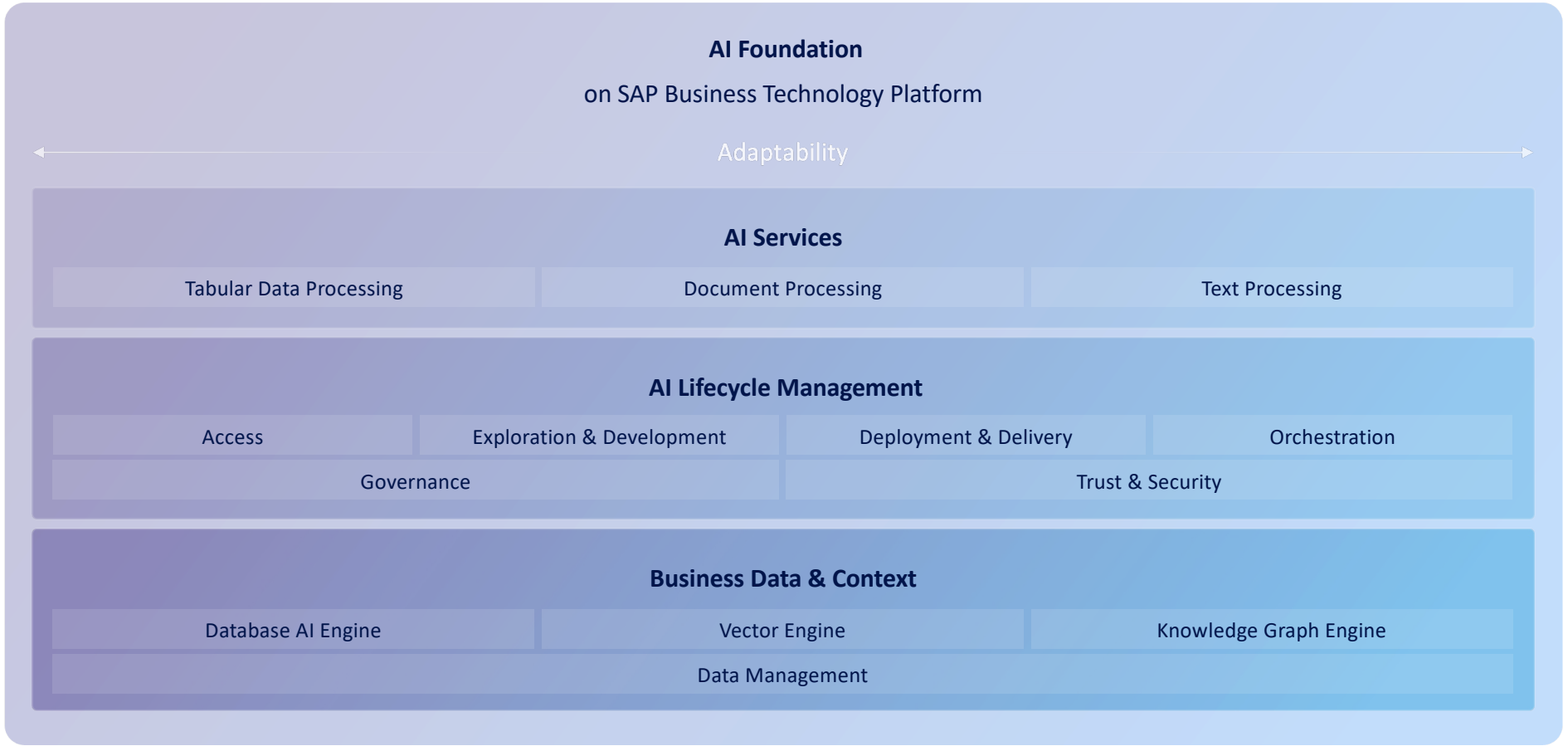


AI Foundation is a complete set of services for AI developers on SAP BTP

to build powerful AI-driven extensions and applications, leveraging the same robust and responsible technology that powers SAP applications.

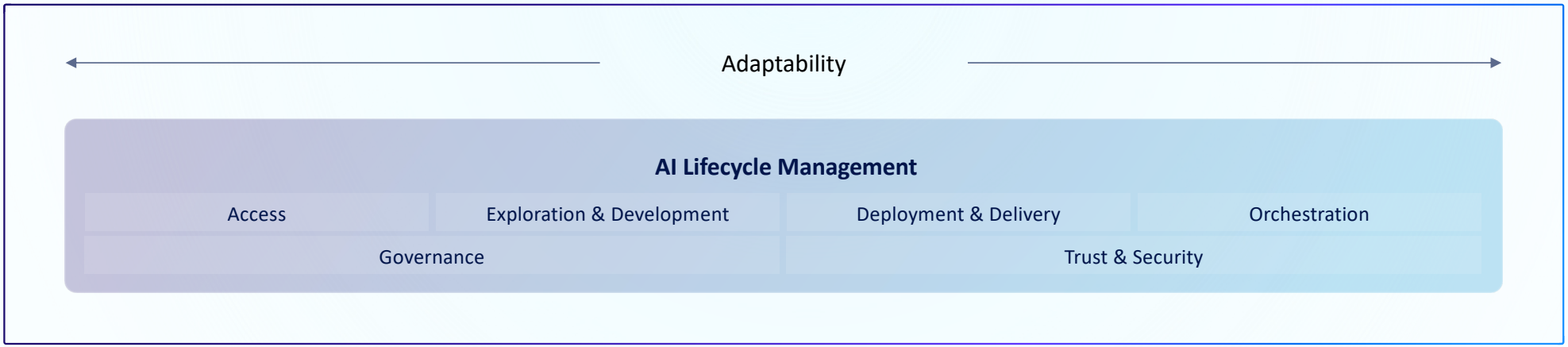


AI Foundation is a complete set of services for AI developers on SAP BTP



Generative AI Hub

Develop, deploy, and manage custom-built AI solutions
and AI-powered extensions of SAP applications.



Benefit today

Get flexible access to broadest range of models & compute capacity

Speed up AI development with access to the broadest set of frontier AI models, infrastructure and tooling.

Extend SAP applications with AI & build custom AI solutions

Combine AI models with your unique data to build powerful custom AI solutions or extend SAP applications.

Safeguard your data & develop AI responsibly

Safeguard your data and maintain full control of your AI lifecycle with SAP's trusted privacy and security policies, and a centralized orchestration approach.

Get flexible access to broadest range of models & compute capacity

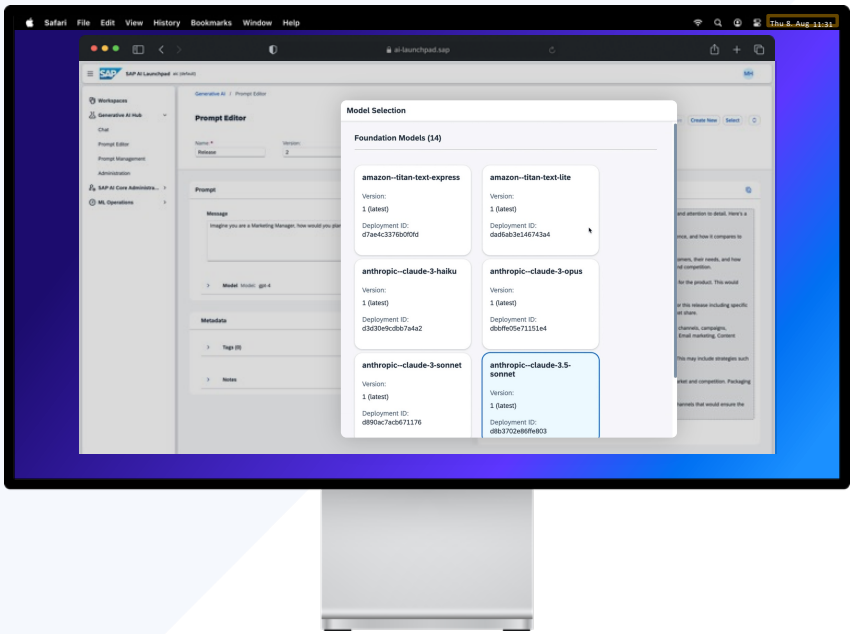
Tailor AI to your Needs

Foundation Model Access

Connect easily and seamlessly to any supported foundation model or bring your own model.

Switch easily between models to find and upgrade to the best suited technology for your needs.
No need for individual contracts and no lock-in to ultimately boost the ROI of your AI projects.

Learn more.



25 models available

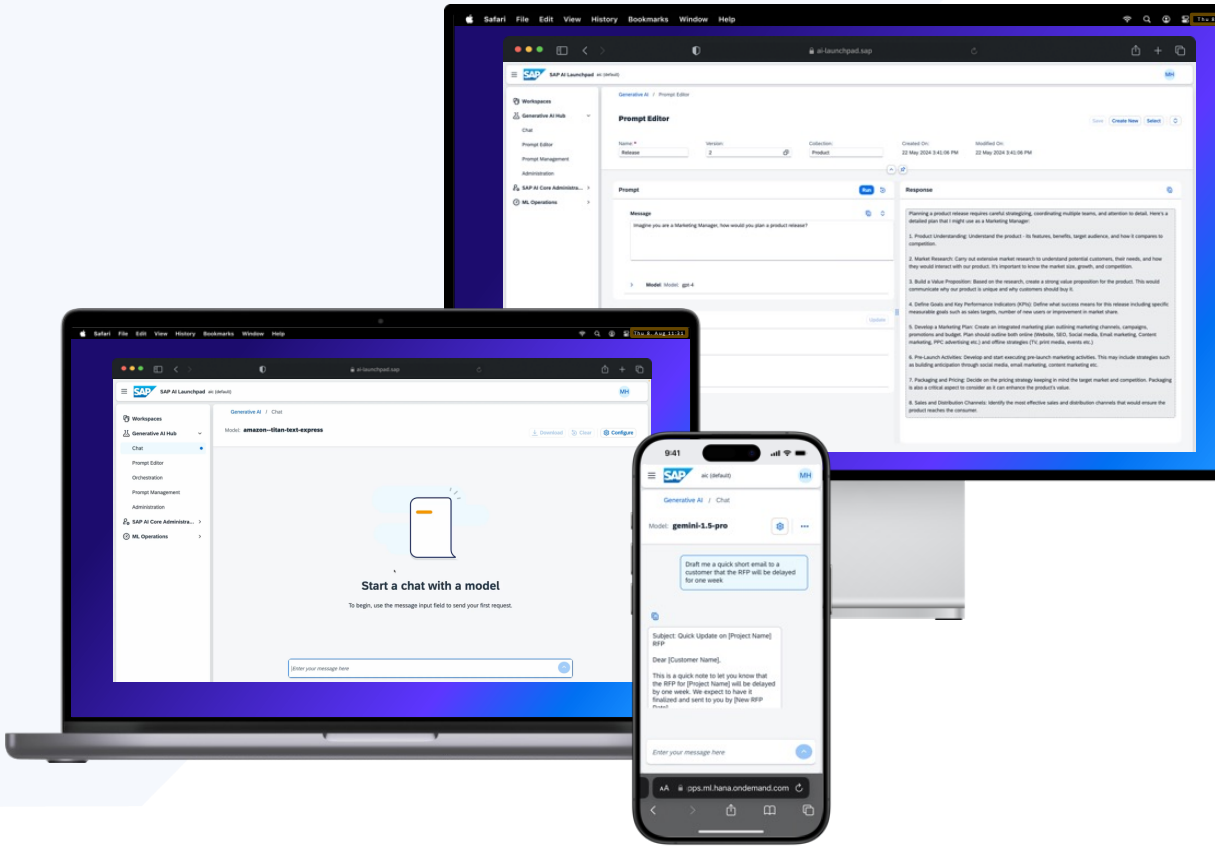
Managed by SAP | Built by Partners | Built by SAP

Playground and Prompt Engineering

Experiment with a prompt engineering playground and explore different models, meta data and parameter changes or generative AI capabilities.

All in a secure and safe environment to interact with cutting-edge technology.

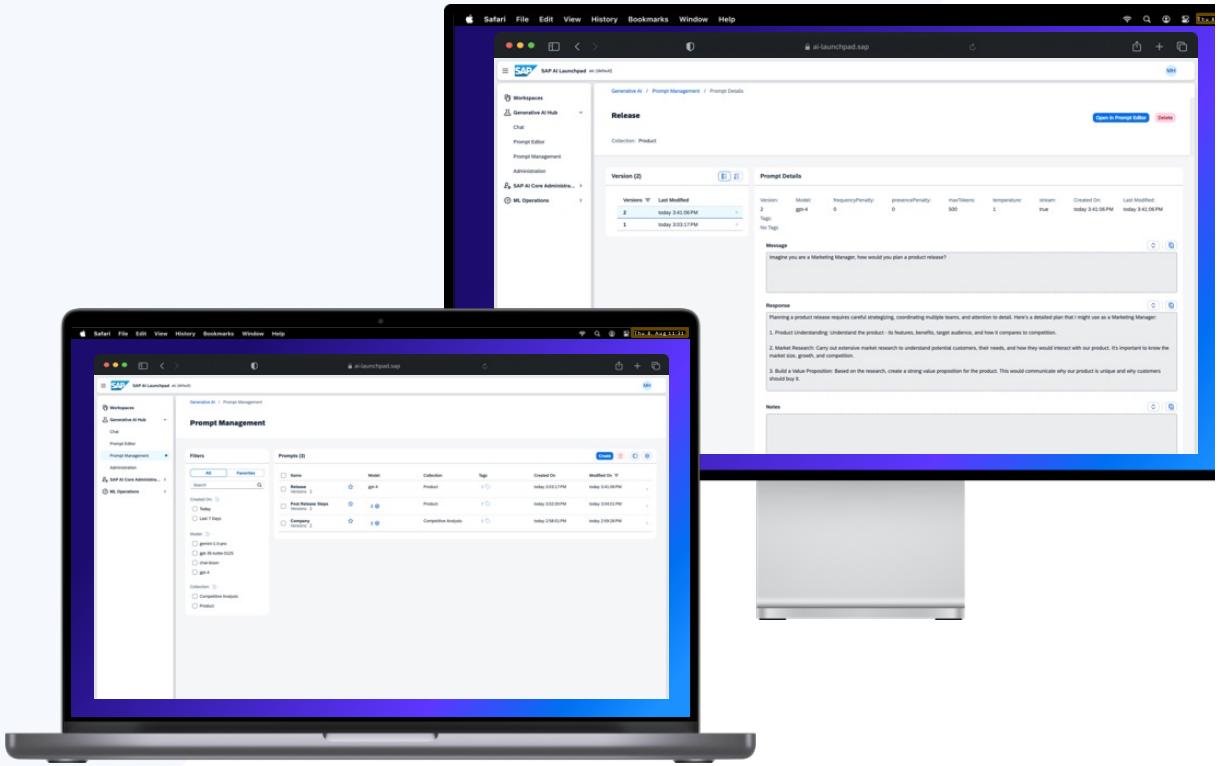
Find out more and see it in action.



Prompt Management and Registry

Effectively manage prompt lifecycles, save prompts and use prompt templates to kick-start the productization of LLM-centric applications.

Find out more and see it in action.

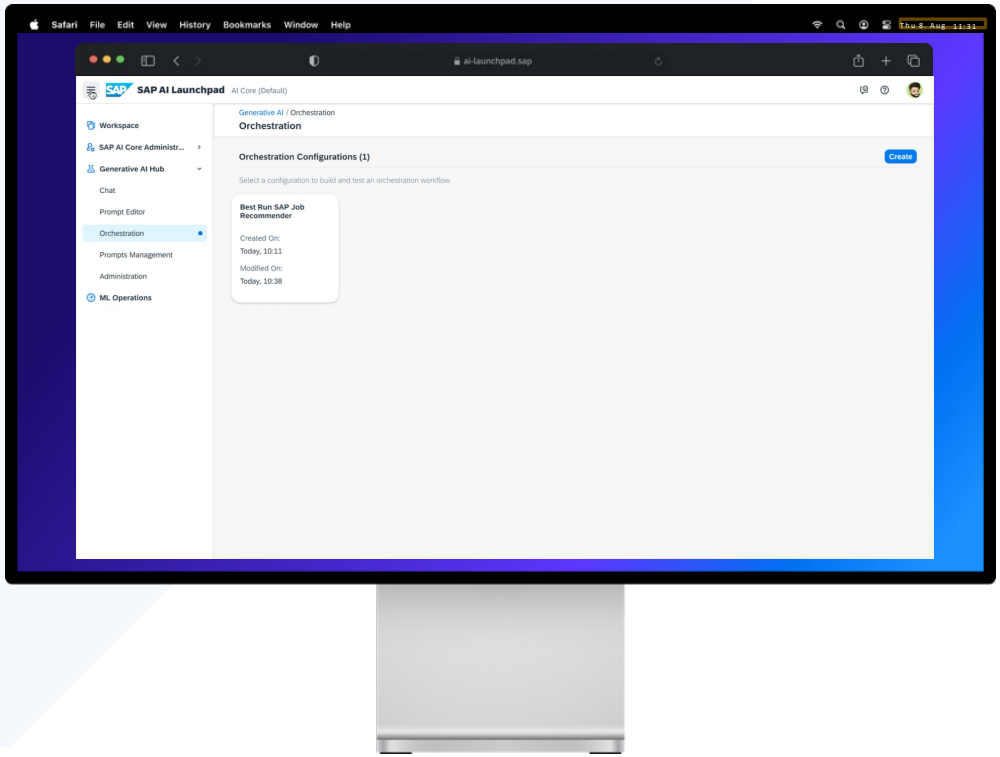


Orchestration Workflows

Design powerful AI workflows, connecting diverse components like data pipelines, AI models, and pre-built modules (grounding, content filtering, and more) and gain peace of mind with less maintenance.

Focus on innovation, not integration, and bring your AI vision to life faster.

Learn more.



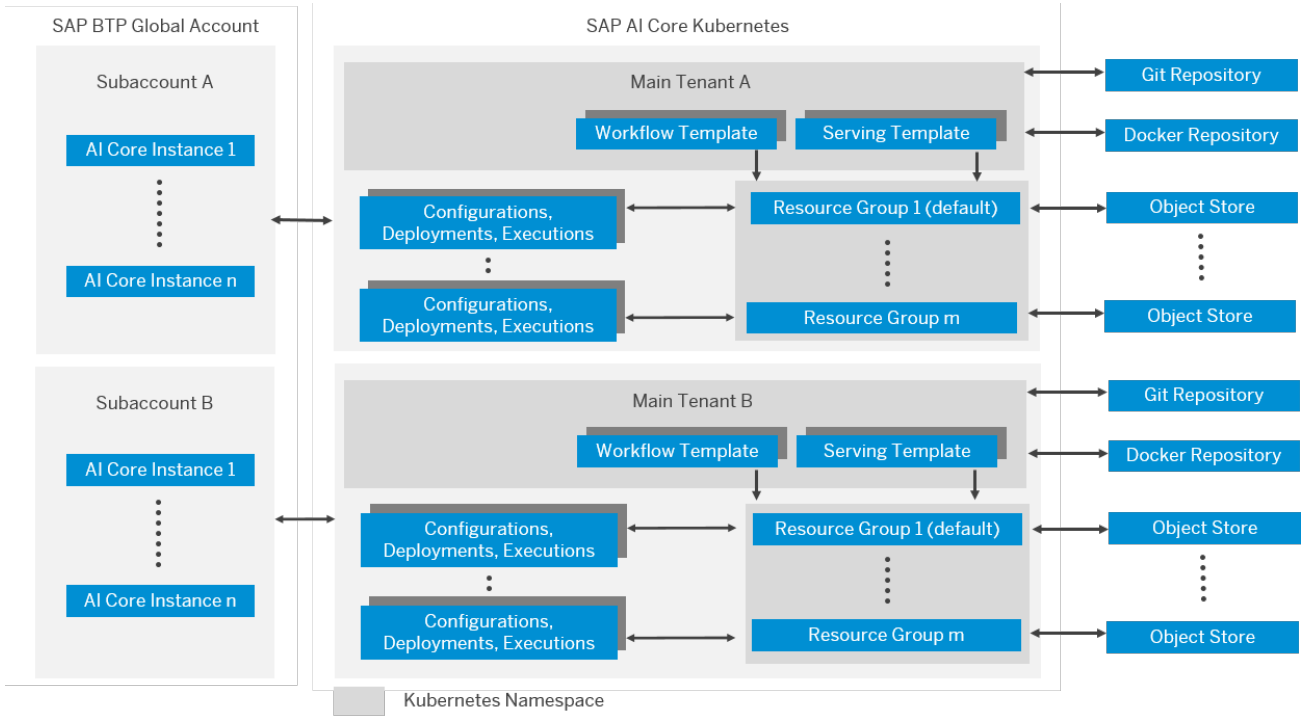
25 models available

Generative AI Hub

What is SAP AI Core?

SAP AI Core is a service in the SAP BTP which is designed to handle the execution and operations of your AI assets in a standardized, scalable, and hyperscaler-agnostic way.

- KServe
- Argo Workflows
- Istio
- ...



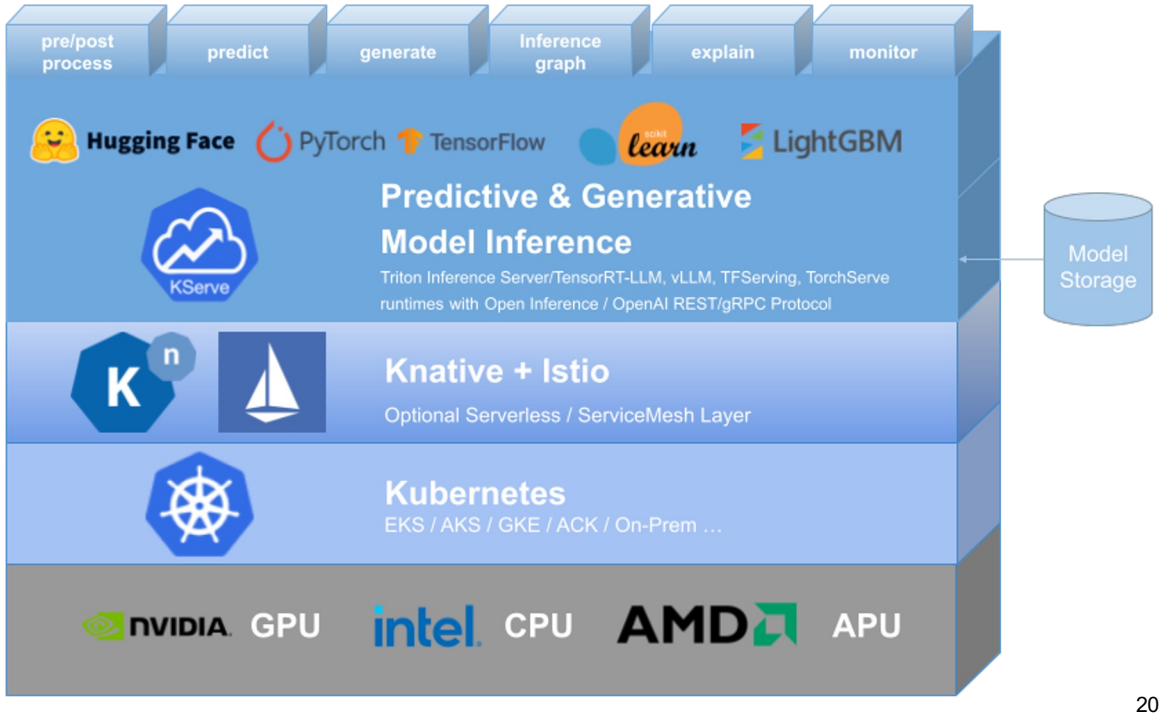
LLM Challenges in Production

- **New requirements on serving LLM**
  - New inference APIs like text generation, embeddings.
  - Streaming response is required for real-time user experience.
- **Variety of models and runtimes**
  - TGI, vLLM, TRT-LLM etc.
  - Llama, Mistral, Phi, Qwen etc.
- **LLM services from cloud providers**
  - Different providers have their own spec (api and token calculation)
- **High computing cost**
  - Expensive hardware, high energy consumption
- **Data privacy**
  - Model and request data can be sensitive and private for inference.

Manages LLM lifecycle in a K8S way

What is KServe?

Highly scalable and **standards**-based **cloud-native model inference platform** on **Kubernetes** for trusted AI that encapsulates the complexity of deploying AI models to production.



What is KServe?

Core Inference

- Transformer/Predictor
- Serving Runtimes
- Custom Runtime SDK
- Open Inference Protocol
- Serverless Autoscaling
- Cloud/PVC Storage

Advanced Inference

- ModelMesh for Multi-Model Serving
- Inference Graph
- Payload Logging
- Request Batching
- Canary Rollout

Model Explanability & Monitoring

- Text, Image, Tabular Explainer
- Bias Detector
- Adversarial Detector
- Outlier Detector
- Drift Detector

What is KServe?

Serving runtime support matrix

Serving Runtime/ Model Format	scikit-learn	xgboost	lightgbm	TensorFlow	PyTorch	TorchScript	ONNX	MLFlow	Custom	HuggingFace
MLServer (open)	✓	✓	✓					✓	✓	
Triton (open)				✓		✓	✓			
TorchServe (v1, open)					✓	✓				✓
KServe Runtime (v1, open)	✓	✓	✓						✓	
TFServing (v1)				✓						
HuggingFace Server (v1,v2,openAI)										✓
HuggingFace vLLM Server (v2, openAI)										✓

KServe on LLM

Inference Service and Serving Runtime for LLM

KServe User

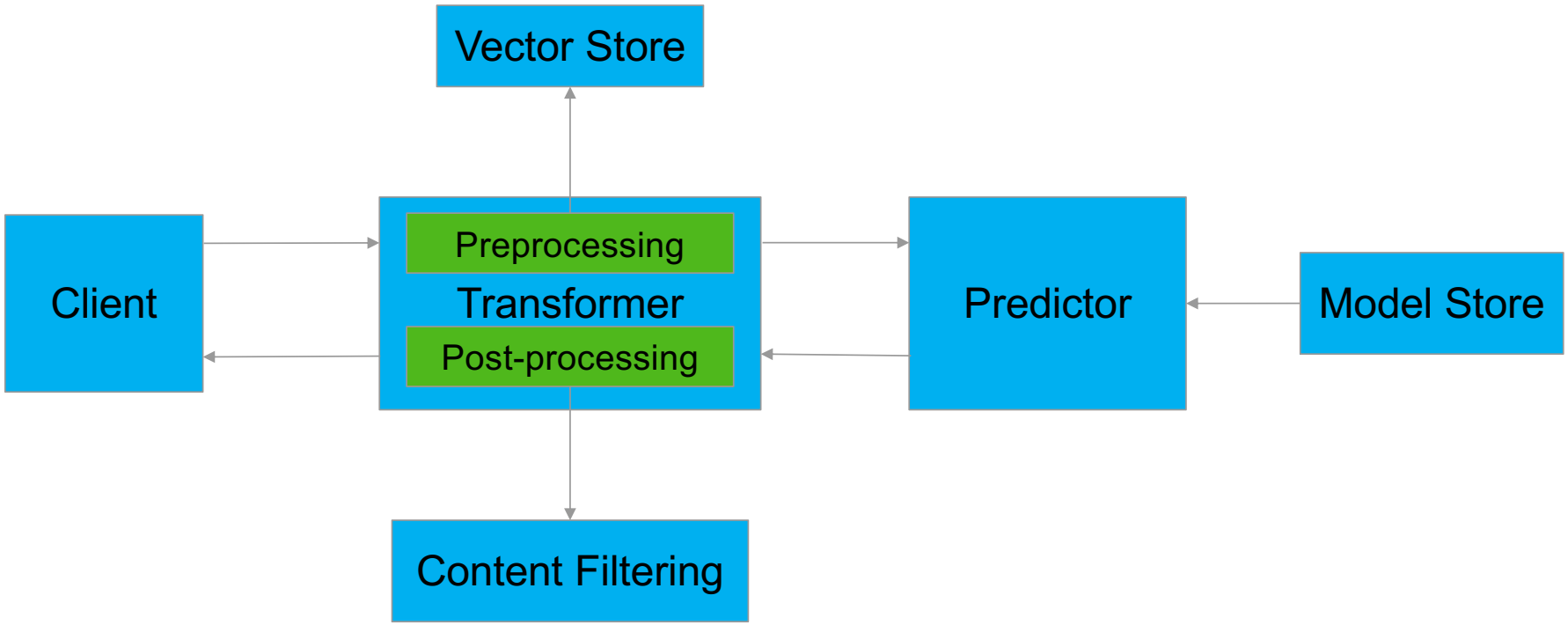
```yaml
apiVersion: serving.kserve.io/v1beta1
kind: InferenceService
metadata:
  name: huggingface-llama3
spec:
  predictor:
    model:
      modelFormat:
        name: huggingface
      args:
        - --model_id=meta-llama/meta-llama-3-8b-instruct
      resources:
        limits:
          cpu: "6"
          memory: 24Gi
          nvidia.com/gpu: "1"
        requests:
          cpu: "6"
          memory: 24Gi
          nvidia.com/gpu: "1"
```

KServe Admin

```yaml
apiVersion: serving.kserve.io/v1alpha1
kind: ClusterServingRuntime
metadata:
  name: kserve-huggingfaceserver
spec:
  annotations:
    prometheus.kserve.io/port: '8080'
    prometheus.kserve.io/path: "/metrics"
  supportedModelFormats:
    - name: huggingface
      version: "1"
      autoSelect: true
      priority: 1
  protocolVersions:
    - v1
    - v2
  containers:
    - name: kserve-container
      image: "kserve/huggingfaceserver:latest"
      args:
        - --model_name={{ .Name }}
      resources:
        requests:
          cpu: "1"
          memory: 2Gi
        limits:
          cpu: "1"
          memory: 2Gi
```

What is KServe?

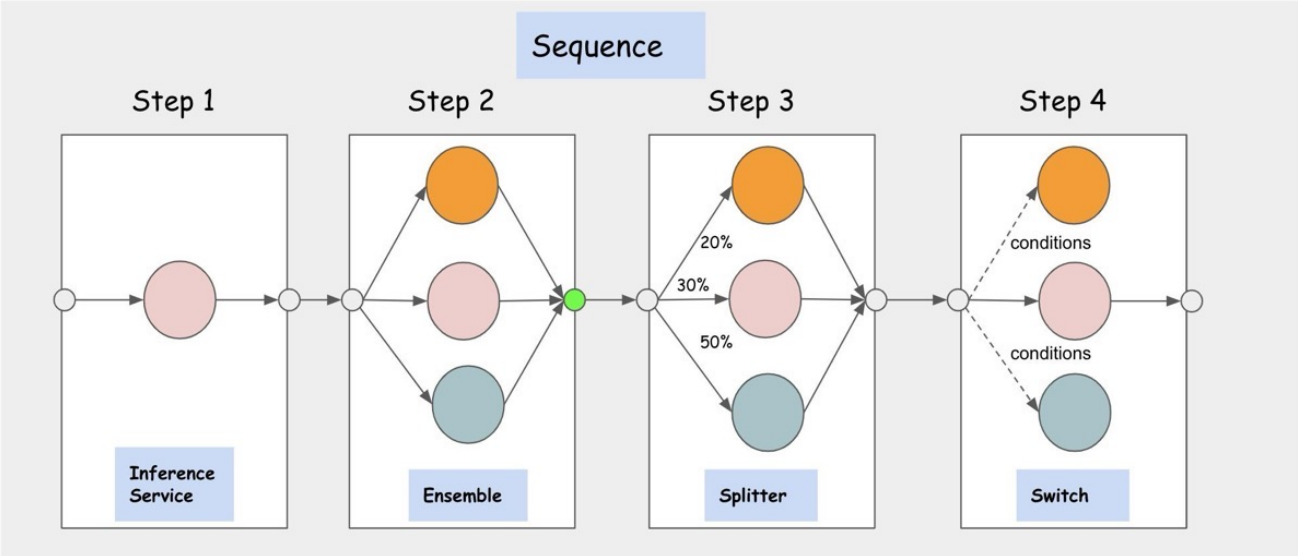
Support of common LLM Use cases



What is KServe?

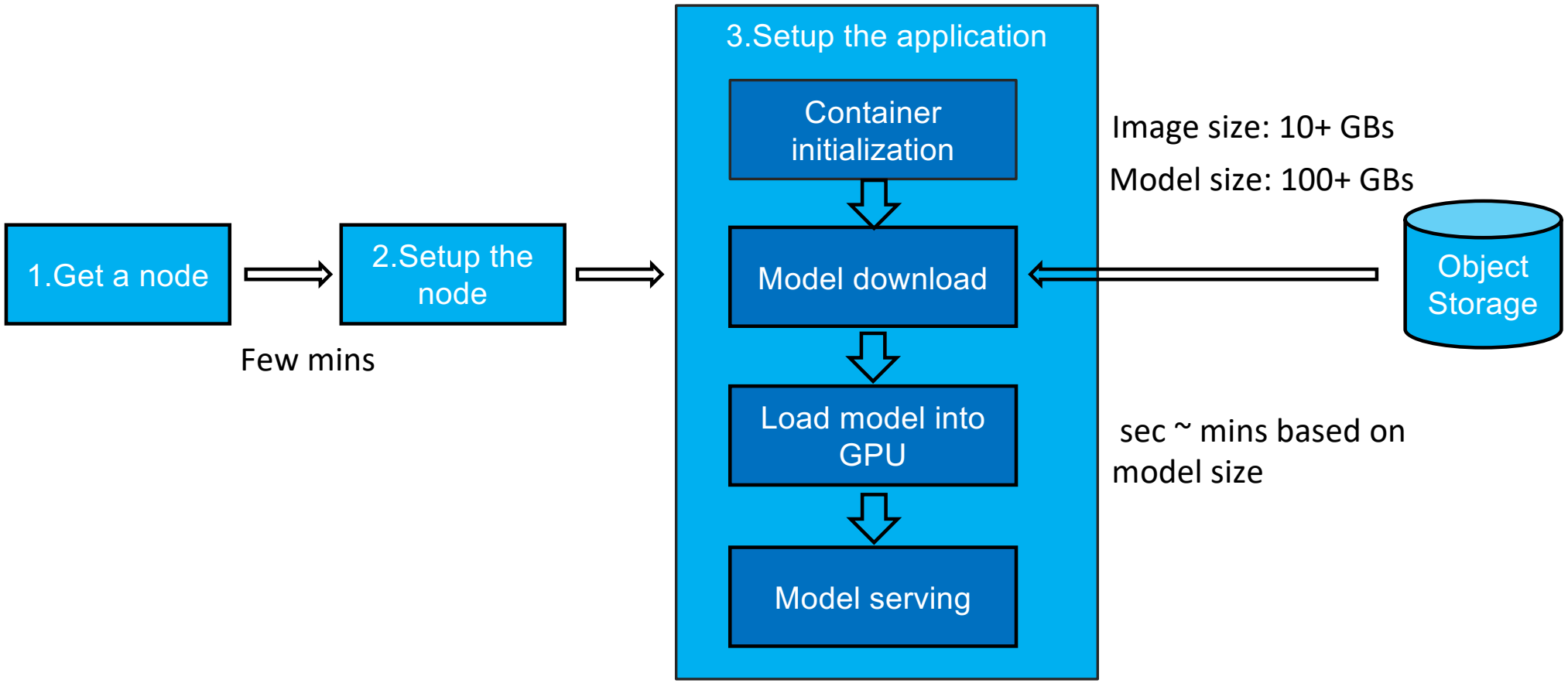
Inference Graph:

- Inference Graph is built for more complex **multi-stage inference pipelines**.
- Inference Graph is deployed in a **declarative way and highly scalable**.
- Inference Graph supports **Sequence, Switch, Ensemble and Splitter** nodes.
- Inference Graph is **highly composable**. It is made up with a list of routing nodes and each node consists of a set of routing steps which can be either route to an InferecenService or another node.

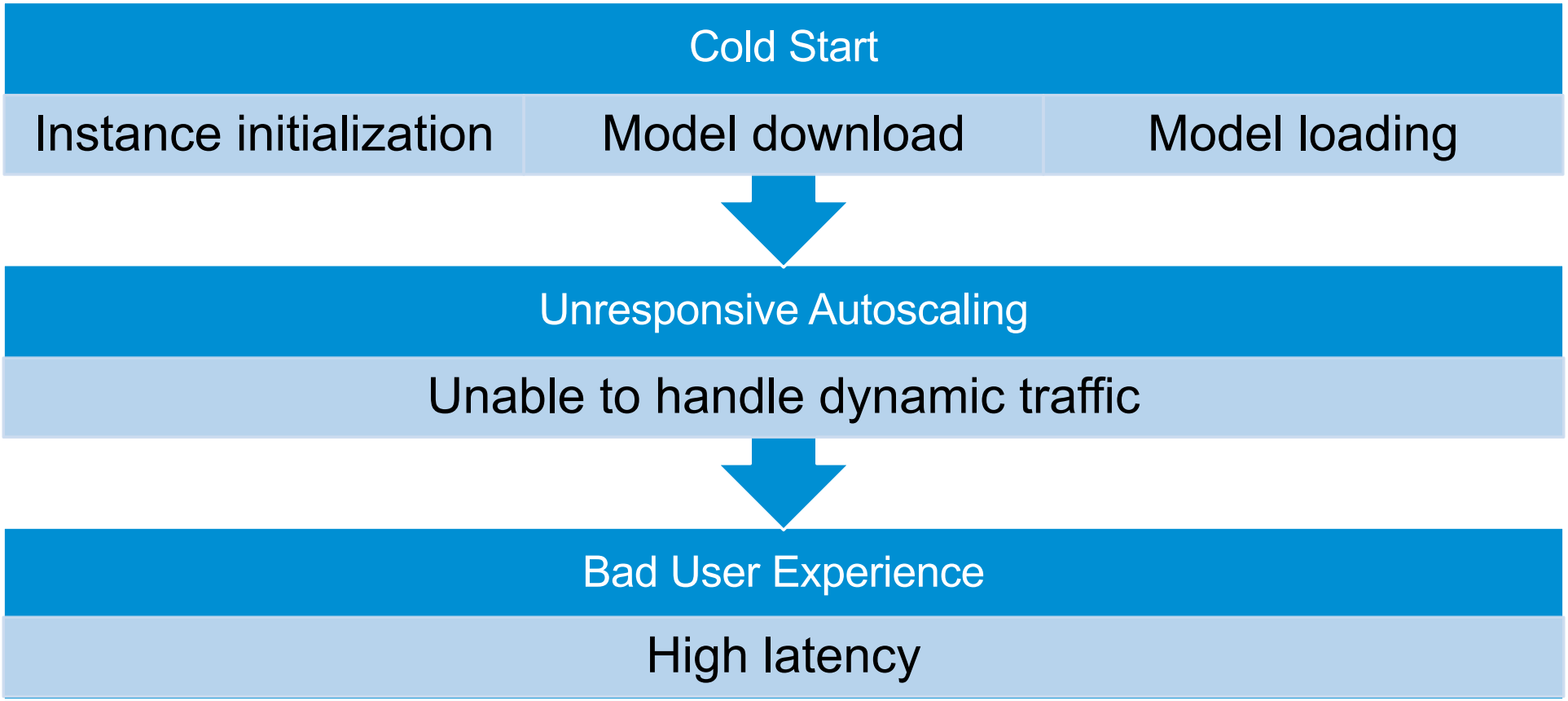



```yaml
apiVersion: "serving.kserve.io/v1alpha1"
kind: "InferenceGraph"
metadata:
  name: "dog-breed-pipeline"
spec:
  nodes:
    root:
      routerType: Sequence
      steps:
      - serviceName: cat-dog-classifier
        name: cat_dog_classifier # step name
      - serviceName: dog-breed-classifier
        name: dog_breed_classifier
        data: $request
        condition: "[@this].#(predictions.0==\"dog\")"
```

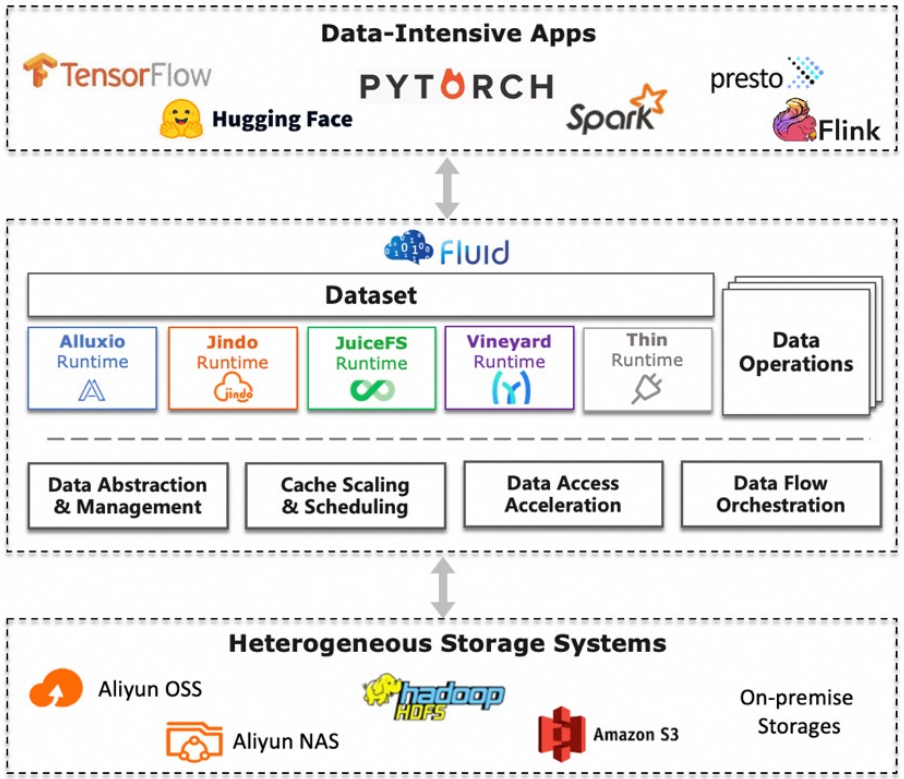
Process of Deploying LLM



Challenges of Autoscaling in LLM



What is Fluid



- **Standardized:** K8s Native APIs for data access and distributed cache management.
- **Extensible**: Runtime plugins for different distributed cache and storage backends.
- **Elasticity**: Scale out and in the distributed cache on demand.
- **Performance:** Accelerate data access via elastic distributed cache
- **Automation**: Operation for Data like. *prefetching processing*, *migration* and *cache scaling*
- **Orchestration**: Data and task co-aware scheduling

Joint launched by Nanjing University, Alibaba Cloud and Alluxio

https://github.com/fluid-cloudnative/fluid

Fluid Optimization for LLMs

Characteristics and Current Issues in LLMs:

- Distributed caching are *complex* and *vary greatly* across environments
- How to balancing Performance and Cost
- Cross-region/zone data access affects performance
- Data Operations are complex and time-consuming.

Capabilities Provided by Fluid:

- Out-of-the-box acceleration capabilities.
- Elastic Compute-Side Distributed Cache
- Affinity Scheduling for Data and Workloads
- Data Flow for Automated Data Management and Consumption Processes

Demo

Future Works

- LLM Serving Runtimes: TGI, TRT-LLM etc
- LLM RAG Pipeline Orchestration
- GenAI Task APIs
- LLM Gateway

Thank you & QA