CLOUD NATIVE INDORE: TECH TALK

Build AI Cloud using Kubernetes

Anjul Sahu, CEO, CloudRaft

CLOUDRAFT

Anjul Sahu
CEO, CloudRaft

About Me

- Founder & CEO, **CloudRaft** - an AI & Cloud Native Consulting
- Organizer, Cloud Native Indore
- More than 16 years in Industry building large scale systems. Previously worked for Telco, Banks, Product & Startups
- Passionate about new technology

256 GH200 DGX Cluster
1 Exaflop, 144 TB GPU
2023

In this Presentation

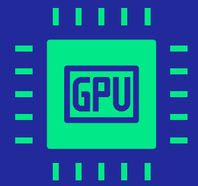
Overview

An AI Cloud simplifies AI implementation for organizations by integrating it into daily operations. AI Clouds cover the AI lifecycle, from creating features and models to operating, monitoring, and sharing them throughout the organization. Platforms supporting the full AI lifecycle are known as AI platforms, and when available in scalable environments, they are termed AI Clouds.

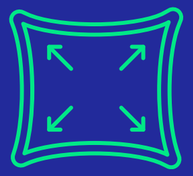
Features of AI Cloud

On-prem , Hybrid or Cloud

Support end-to-end lifecycle of AI



GPUs & High Performance



Scalable

Reliable

Billing or Chargeback

AI/ML Frameworks

Full stack: IaaS, PaaS, SaaS

Self-service

Current Trends in AI Infrastructure

01

Specialized Cloud
Eg: CoreWeave, Salad, RunPod, Nebius, Lambda labs etc

02

2x Data in every 18 months
The demand for data to build better AI/ML models is increasing faster than Moore's Law, doubling every 18 months

03

Cloud Native and Kubernetes is an accelerator for AI

04

GenAI: Bigger Models
model size is increasing that means more powerful infrastructure is required

05

Data Sovereignty Requirements
Enterprise data loss risk, AI Safety and new Govt policies to keep data local

AI Runs on ~~GPUs~~ Accelerators

- AI = matrix multiplications which is massively parallelizable
- GPUs are great at parallel programming
- CPU < 32 cores/threads, GPUs> 4000 cores/threads
- CPU is 10x slower at least
- Impractical to train or even run any reasonable AI model outside ASICs

GEN AI APP INFRASTRUCTURE STACK

SAPPHIRE VENTURES

Data Management & Processing

DATA PIPELINE
ASTRONOMER, chalk, databricks, dataloop, DATAVOLO, INSTILL AI, LlamaIndex, matillion*, NEUM, NOMIC, Prophecy, Superlinked, Unstructured, Vector Flow, Visual Layer

DATA LABELING, CLASSIFICATION & CURATION
Alation*, Cleanlab, datasaur.ai, datologyai, dataloop, ENCORD, Fortify AI, HIVE, HumanSignal, Labelbox, Lilac, databricks, refuel, scale, Snorkel, SuperAnnotate, surge*

DATA STORAGE & RETRIEVAL
activeloop, Baseplate, chroma, cohere, DATASTAX, drant, GROUNDX, EYELEVEL, LanceDB, marqo, mindsdb, neo4j, nuclia, Pinecone, PostgresML, seekr, turbopuffer, vectara, Weaviate, VOYAGE AI, zilliz

SYNTHETIC DATA GENERATION
aindo, datacebo, datagen, gretel, MOSTLY AI, parallel domain, Synthesis AI, SYNTHESIZED, YData

Model Training & Deployment

MODEL DISCOVERY
Hugging Face, DVCX, Replicate

MODEL TESTING & EVALUATION
Argilla, Adaptive ML, deepchecks, distributional, Galileo, gentrace, kolena, Patronus AI, Prolific, RagaAI, scale, surge*, W&B*

MODEL EXPERIMENTATION
comet, DataRobot*, DVCX, Galileo, mlflow, Outerbounds, neptune.ai, W&B*

MODEL SERVING & INFERENCE
CentML, deci, Exafunction, FlexAI, HippoML, Lightning AI, Modular, NEURAL MAGIC, OctoAI, RECOGNI, run:ai, NVIDIA, Stochastic, Titan

App Development & Orchestration

PROMPT ENGINEERING
Humanloop, orq, PromptLayer, FLOWGPT, promptfoo, PROMPTBASE, Promptstacks

PROMPT CACHING
Redis, zilliz

ORCHESTRATION & ROUTING
Kong*, LlamaIndex, LangChain, lastmile AI, Martian, n8n, ONE AI, vellum

APP & AGENT PLATFORM
acorn, Auto-gpt, beam, ChainML, Continual, crewAI, FIXIE, gradio, graft, GRADIENT, gradient, griptape, Hugging Face, imbue, KLU, Langflow, DATASTAX, ONE AI, portkey, pulze.ai, Relevance AI, re:tune, stack, Steamship, wand

PRODUCT ANALYTICS
Adaptive ML, context, eppo, langfuse, Nebuly AI, pendo*, PostHog, STATSIG, Tidepool

Validation & Security

MODEL SECURITY
Deepkeep, Guardrails AI, HIDDENLAYER, PROTECT AI, ROBUST INTELLIGENCE, TROJ.AI

GOVERNANCE & COMPLIANCE
credo ai, CRANIUM, Citrus, Holistic AI, PureML, VALIDMIND, MONITAUR

LLM APP SECURITY & MODERATION
CALYPSOAI, Dynamo AI, Enkrypt AI, Guardrails AI, GenNeuron, Lasso, LAKERA, PROTECTO, privacera*, PRIVATEAI

END-USER SECURITY
AIM, cadea, CREDAL, credo ai, cyberhaven, harmonic, Lasso, Nightfall, netskope*, privacera*, Prompt:, skyflow

Monitoring & Observability
aporia, arize, Arthur, baserun.ai, BRAINTRUST, context, DataRobot*, fiddler, GANTRY, Giskard, helicone, HoneyHive, LangChain, literal, mona, nomos, pulze.ai, rubbrband, traceloop, truera, snowflake, Wallaroo.AI, WHYLABS, W&B*

End-to-End Training & Fine-Tuning Platform
Akkio, Chima, CLEARML, deepset, DataRobot*, LAMINI, Qwak, Wallaroo.AI

Foundation Models

OPEN SOURCE MODELS & HOSTING
ALEPH ALPHA, anyscale, arcee.ai, FalconLLM, Fireworks AI, gradient, groq, Hugging Face, LMSYS, Liquid, Meta, mosaicML, databricks, MISTRAL AI_, OctoAI, OpenPipe, Predibase, stability.ai, together.ai, xl

CLOSED SOURCE MODELS & HOSTING
ADEPT, AI21labs, ANTHROPIC, aws, cohere, contextual.ai, poolside, Google, Inflection, OpenAI, Reka, SambaNova, Magic

Infrastructure

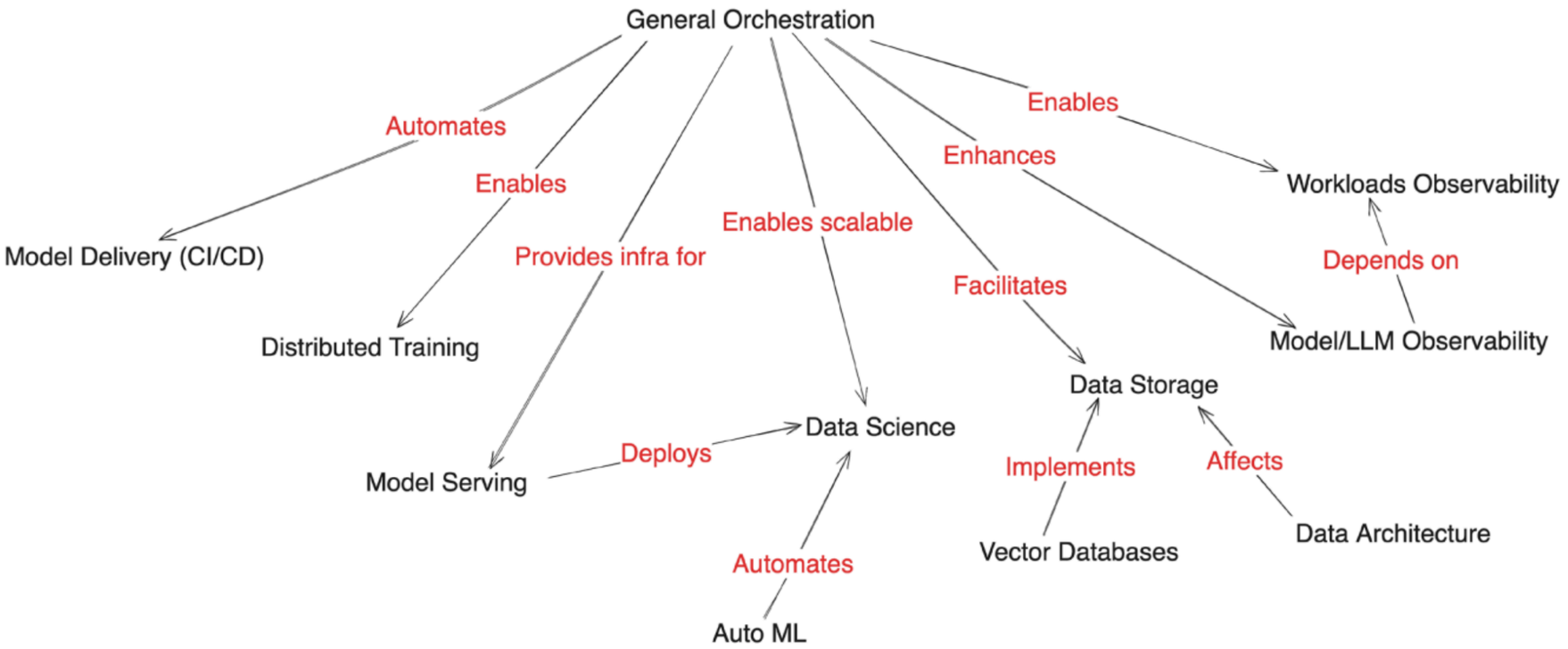
GPU CLOUD HOSTING
aws, Azure, baseten, CoreWeave, Crusoe, databricks, DeepInfra, DigitalOcean, Paperspace, Google Cloud, Genesis Cloud, io.net, kinetica, Lambda, NVIDIA, ORACLE, salad, snowflake, vast.ai, VULTR

PAAS & SERVERLESS GPUS
beam, BentoML, cerebrium, fal, Fly.io, foundry Accelerate, FOUNDRY, Inference.ai, Lepton AI, Modal, mystic, OctoAI, PoplarML, Replicate, RunPod, Shadeform

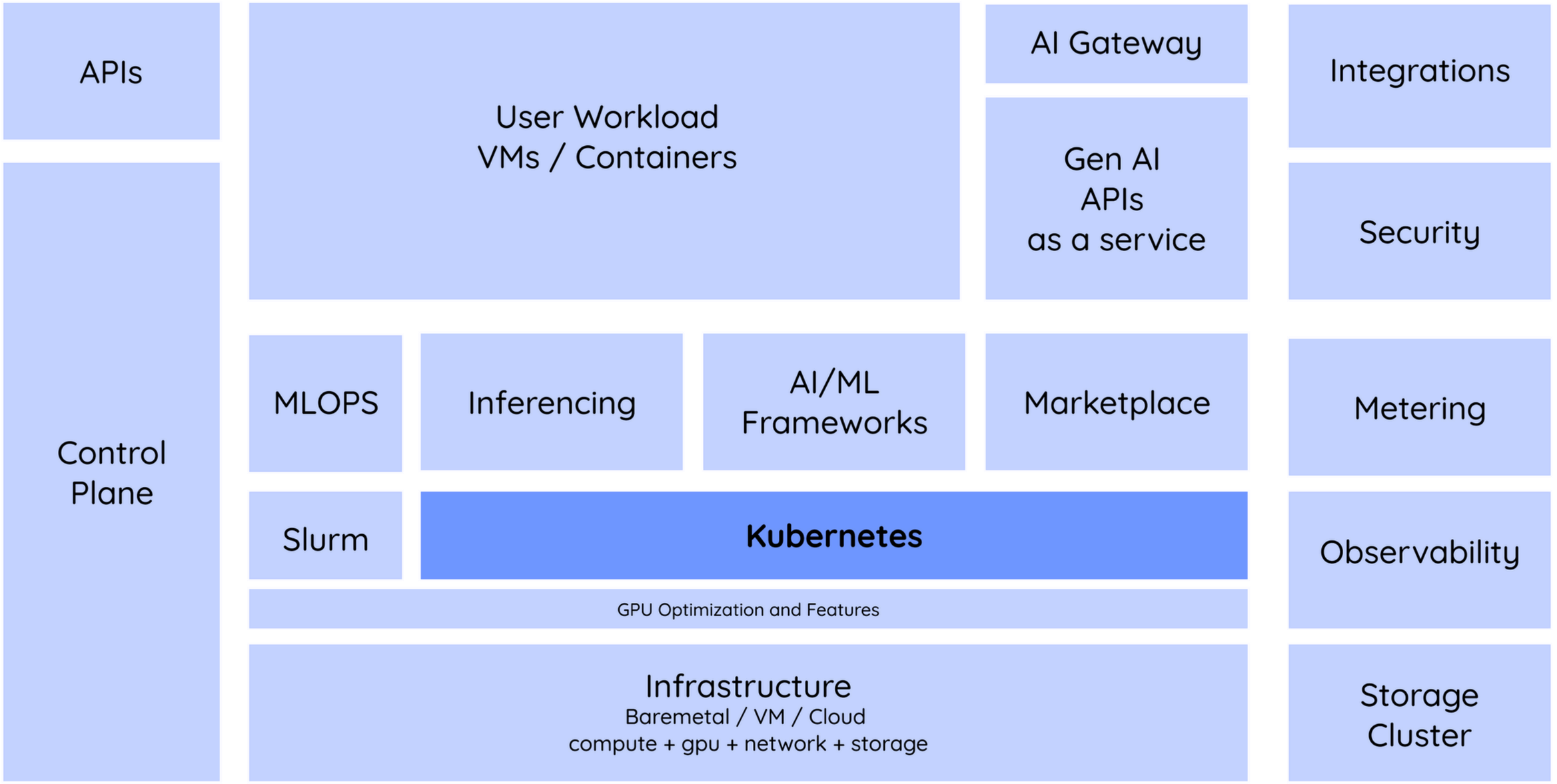
* Denotes current or exited Sapphire portfolio company

Such companies are a representative sample of portfolio companies in which Sapphire has invested and which the author believes fit the criteria of the market, which do not reflect all investments made by Sapphire. Sapphire has internally assigned categories to all companies described above in the market map. Although Sapphire believes that such categorizations are reasonable, they are subjective in nature and may be categorized differently by other market participants.

How Cloud Native helps in running AI Workload



"Research teams can now take advantage of the frameworks we've built on top of Kubernetes, which make it easy to launch experiments, scale them by 10x or 50x, and take little effort to manage."
— CHRISTOPHER BERNER, HEAD OF INFRASTRUCTURE FOR OPENAI



AI Cloud Reference Architecture

Cloud Native Projects for AI

[Distributed Training](#)

[General Orchestration](#)

[Model / LLM Observability](#)

[ML Serving](#)

[CI/CD Delivery](#)

[Vector Databases](#)

[Workload Observability](#)

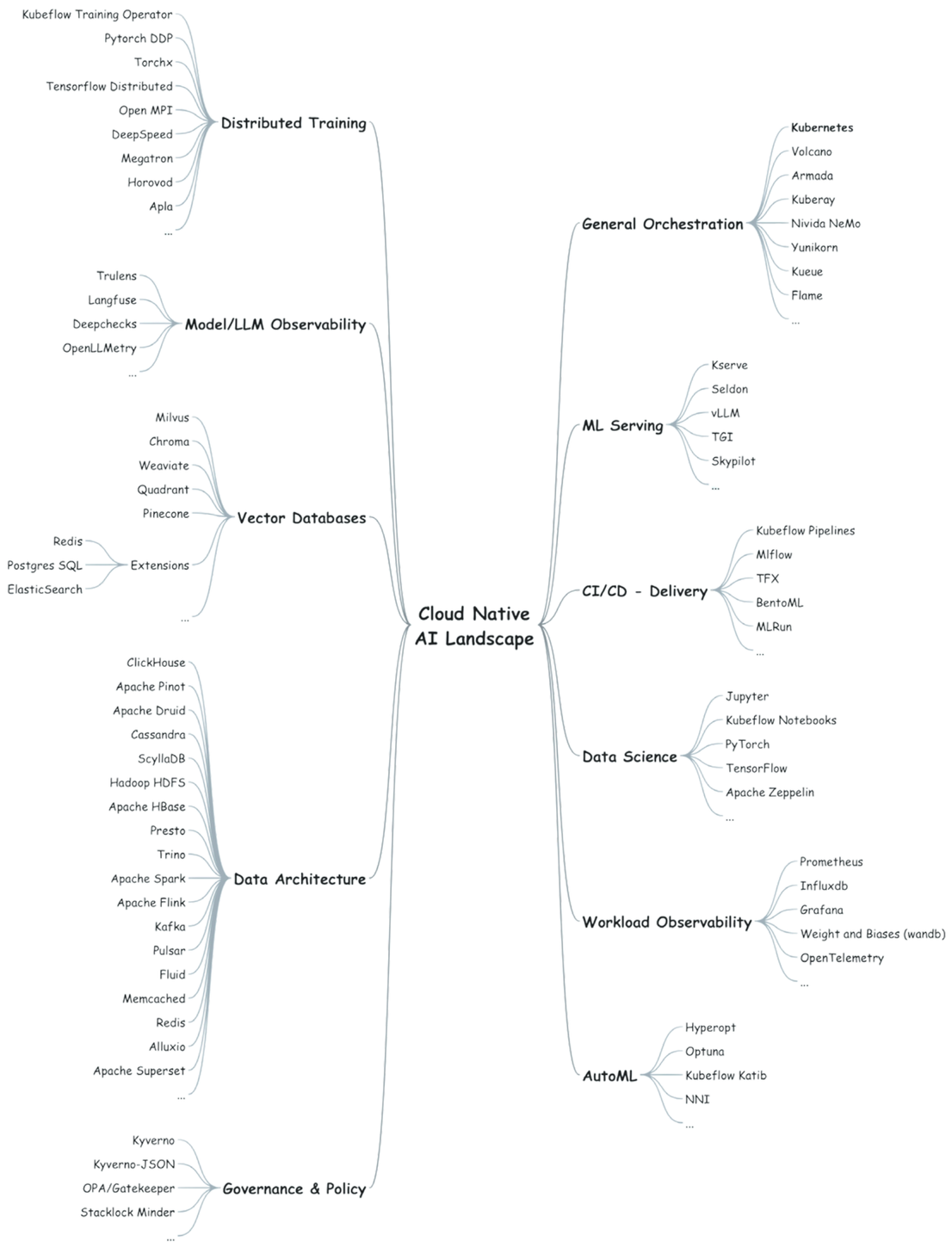
[Data Architectures](#)

[AutoML](#)

[Governance and Policy](#)

[Security](#)

Ecosystem is evolving fast...



Challenges in Building AI Cloud

- Building an AI Cloud is a large investment
- GPU supply chain issues
- Skill issues
- High reliability required for long running distributed training jobs
- Unknown security threats and AI Risk in the fast evolving ecosystem
- Sustainability - Each H100 energy consumption is more than avg household
- Some of the hardware limitations becomes bottlenecks such as storage or the network

Why we need AI Cloud?

- Data Privacy
- AI is making humans more productive
- AGI is possible
- Cost is still less as compared to hyperscalers
- It is a game changer for many enterprises

This talk is based on **our recent work.**
And it was not possible without the ground breaking innovations done by
Kubernetes, NVIDIA and CNCF foundation

See our insights on AI
cloudraft.io/blog

Q & A

"Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks."
-Stephen Hawking, Theoretical Physicist

CLOUDRAFT