

Cloud & Hạ tầng AI

Cloud & AI Infrastructure

Kho nhiên liệu và xưởng động cơ của phi đội AI: nền tảng dữ liệu (data lake, feature store) hợp nhất với hạ tầng tính toán (GPU/accelerator, MLOps) — sẵn sàng cho toàn vòng đời AI/ML, từ dữ liệu tới phục vụ mô hình và LLM ở quy mô sản xuất.

Bảng mô tả dịch vụ
 Service description

Mã dịch vụ Service code	NS-AIC
Nhóm dịch vụ Service group	Nền tảng dữ liệu & hạ tầng AI
Phạm vi Scope	Data platform, feature store, quản trị dữ liệu; cluster GPU, MLOps, model serving, tối ưu suy luận
Đối tượng Target customers	Tổ chức đưa AI/ML — kể cả LLM — vào sản xuất
Mô hình hợp tác Engagement models	Dự án nền tảng · vận hành cluster theo tháng · đồng hành đội data science
Cam kết Commitments	SLA theo hợp đồng · hiệu năng pipeline dữ liệu và mức sẵn sàng cluster

SẢN PHẨM BÀN GIAO · DELIVERABLES

- ✓

Kiến trúc dữ liệu và hạ tầng tính toán cho AI
- ✓

Data lake/lakehouse, feature store và pipeline vận hành
- ✓

Cluster GPU và MLOps pipeline từ huấn luyện đến phát hành mô hình
- ✓

Lớp phục vụ mô hình có giám sát và quy tắc quản trị dữ liệu

KẾT QUẢ KỲ VỌNG · OUTCOMES

- Dữ liệu sạch và hạ tầng đủ lực cho mô hình
- Rút ngắn thời gian từ ý tưởng đến mô hình chạy thật
- Chi phí dữ liệu và suy luận kiểm soát được

■ **Hạng mục công việc theo giai đoạn** Consult → Deploy → Operate → Optimize

01

Tư vấn / Consult

- Khảo sát nguồn dữ liệu & workload AI
- Đánh giá mức độ sẵn sàng
- Thiết kế kiến trúc dữ liệu + cluster GPU
- Dự toán TCO

02

Triển khai / Deploy

- Dựng data lake/lakehouse & feature store
- Dựng cluster GPU & scheduler
- MLOps pipeline (huấn luyện → serving)
- Môi trường huấn luyện & model serving

03

Vận hành / Operate

- Vận hành pipeline dữ liệu & cluster 24/7
- Quản trị chất lượng & hạn mức GPU
- Kiểm soát truy cập & dòng dữ liệu
- Theo dõi hiệu suất & chi phí

04

Tối ưu / Optimize

- Tối ưu hiệu năng truy vấn & suy luận
- Quantization & autoscaling
- Tối ưu chi phí lưu trữ theo lớp và GPU
- Chuẩn hoá tái sử dụng dữ liệu & mô hình

CÔNG NGHỆ SỬ DỤNG · TECHNOLOGY STACK

Databricks

Apache Spark

Kafka

Airflow

Feast

MinIO / S3

NVIDIA GPU

Kubernetes

Kubeflow

MLflow

Ray

vLLM

Triton Inference Server

**Bắt đầu với Cloud & Hạ tầng AI
bằng một buổi khảo sát hiện
trạng.**

Start with a current-state assessment.

Điện thoại

+84 973 391 388

Email contact@appcarrier.cloud hoặc dungpv30@fpt.com