蚂蚁开源 ANT OPEN SOURCE RAY

AnalyticDB Ray: Data+AI Architecture

李 伟

阿里云 高级技术专家

2025/12/20

蚂蚁开源 ANT OPEN SOURCE RAY

CONTENT
目录

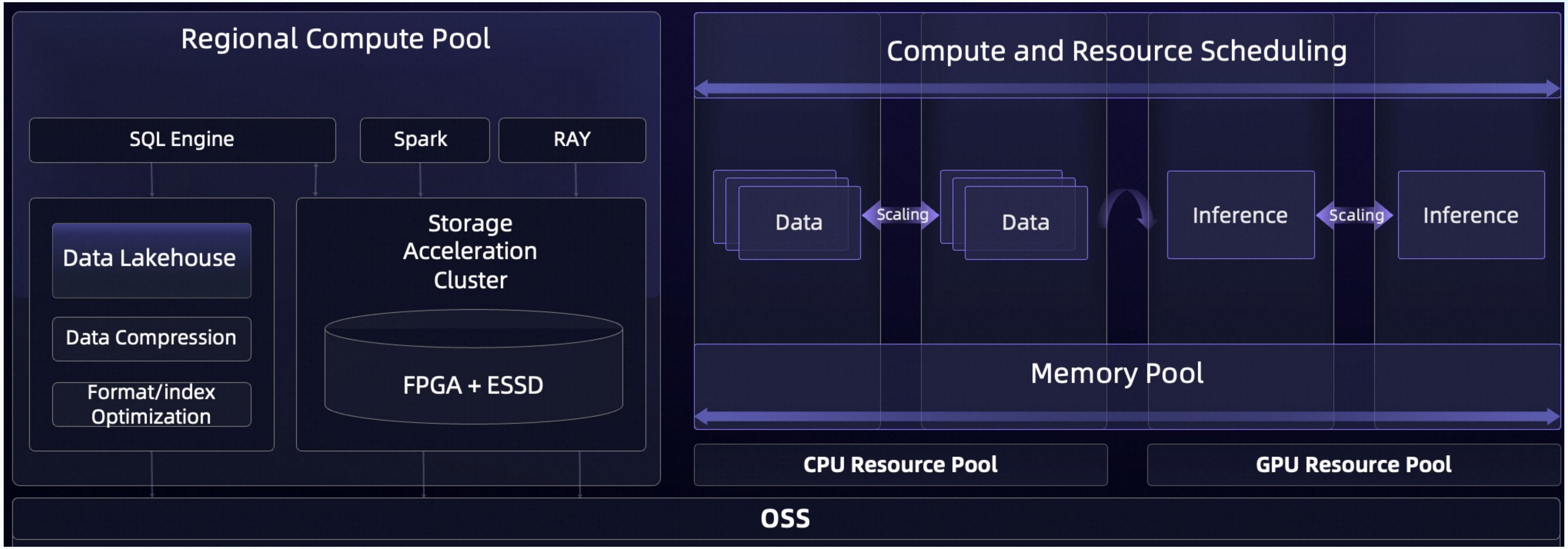
01 **AnalyticDB Data + AI云原生架构**

02 AnalyticDB Ray特性

03 多模态处理调度技术

04 典型案例介绍

AnalyticDB Ray: Data+AI Architecture



Unified Resource Pool

Heterogeneous computing resources

Enterprise-level Service

Seamlessly Integration

蚂蚁开源
ANT OPEN SOURCE

RAY

CONTENT
目录

01 AnalyticDB Data + AI云原生架构

02 **AnalyticDB Ray特性**

03 多模态处理调度技术

04 典型案例介绍

AnalyticDB Ray:特性增强

Usability

- One-Click Deployment
- Observability

Stability

- Seamless Migration
- Self-healing
- High Availability

Cost Performance

- Cache Pool
- Auto Scaling
- Fine-grained Scheduling

Ecosystem

- Data Warehouse
- Data Lake (Lance etc.)
- Tools (Llama-factory etc.)

CONTENT
目录

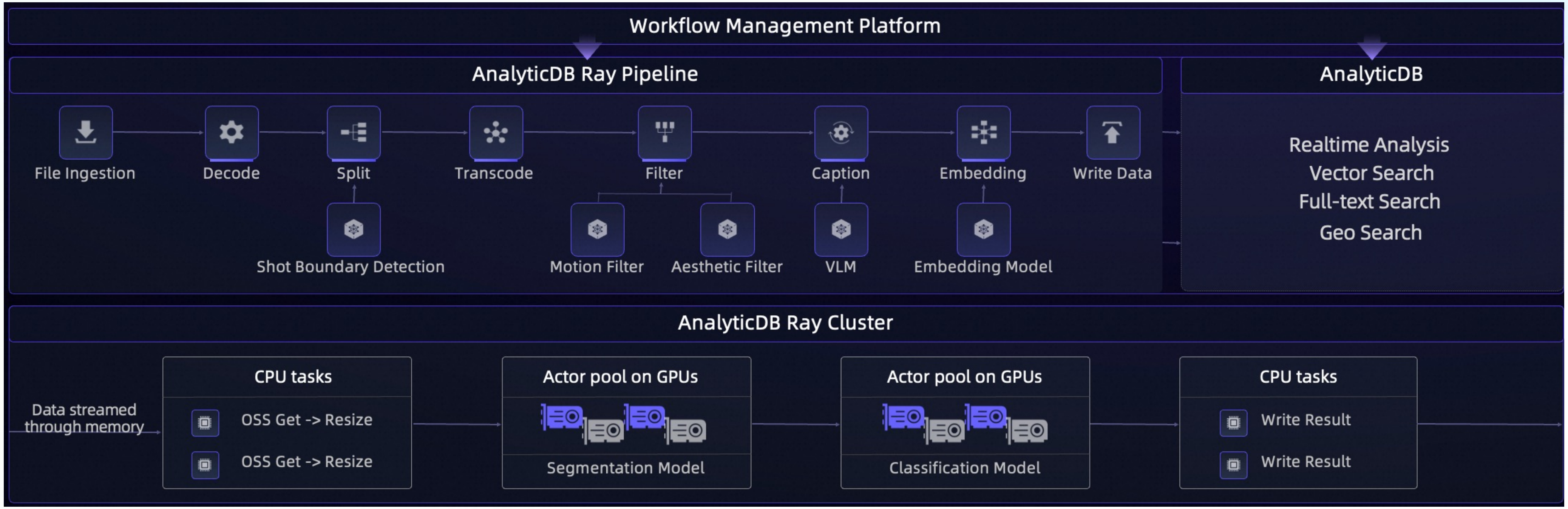
01 AnalyticDB Data + AI云原生架构

02 AnalyticDB Ray特性

03 **多模态处理调度技术**

04 典型案例介绍

AnalyticDB Ray: 多模态处理调度技术



Streaming Pipeline

80% ↑

efficiency improvement

Load-Aware Auto-scaling

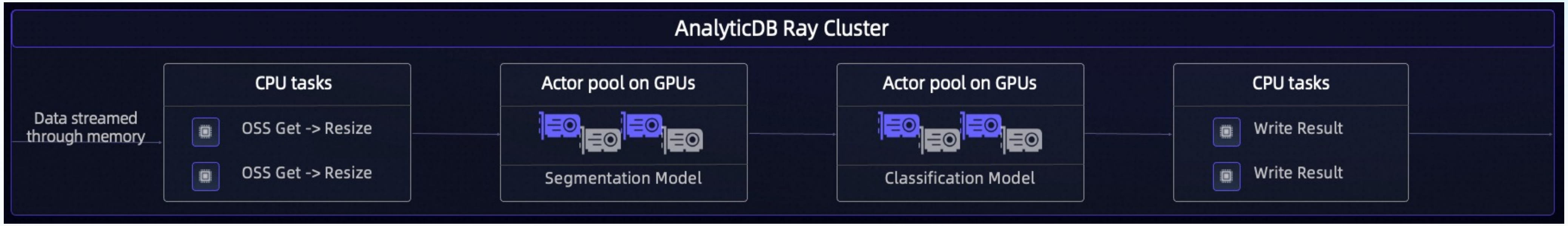
80% ↑

GPU utilization

Fine-grained CPU+GPU co-scheduling

High Availability / Fault Tolerance

AnalyticDB Ray: Adaptive Streaming Pipeline



AnalyticDB Ray Cluster

Data streamed through memory

CPU tasks

OSS Get -> Resize

OSS Get -> Resize

Actor pool on GPUs

Segmentation Model

Actor pool on GPUs

Classification Model

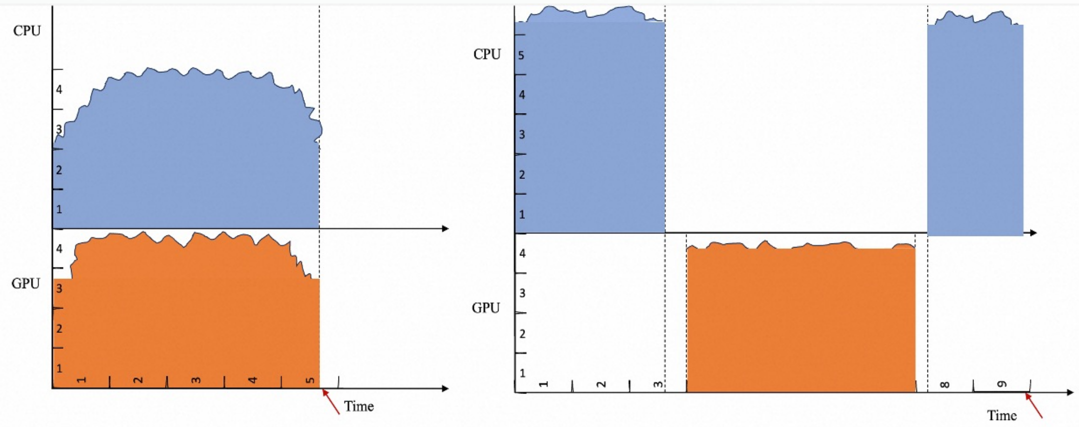
CPU tasks

Write Result

Write Result

minimal resource idleness

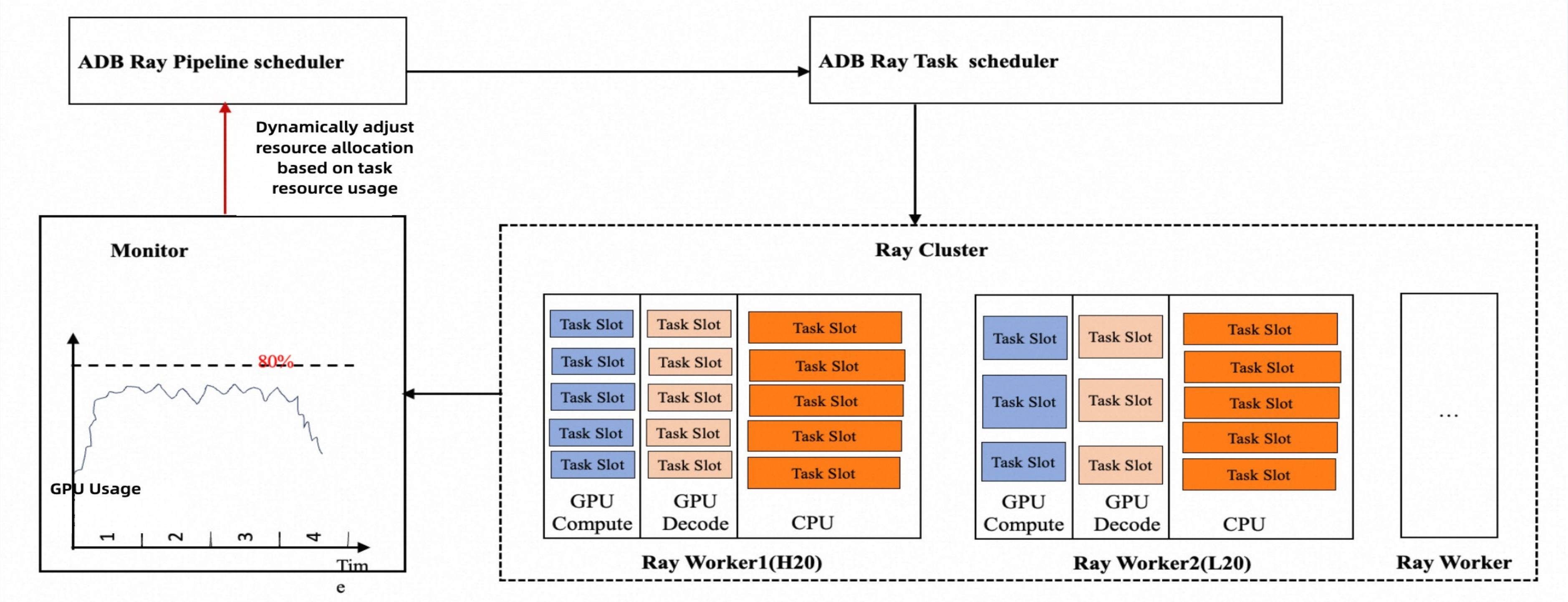
maximize processing throughput



某客户:

Total throughput: 4 Million/Day, Peak 80K/sec

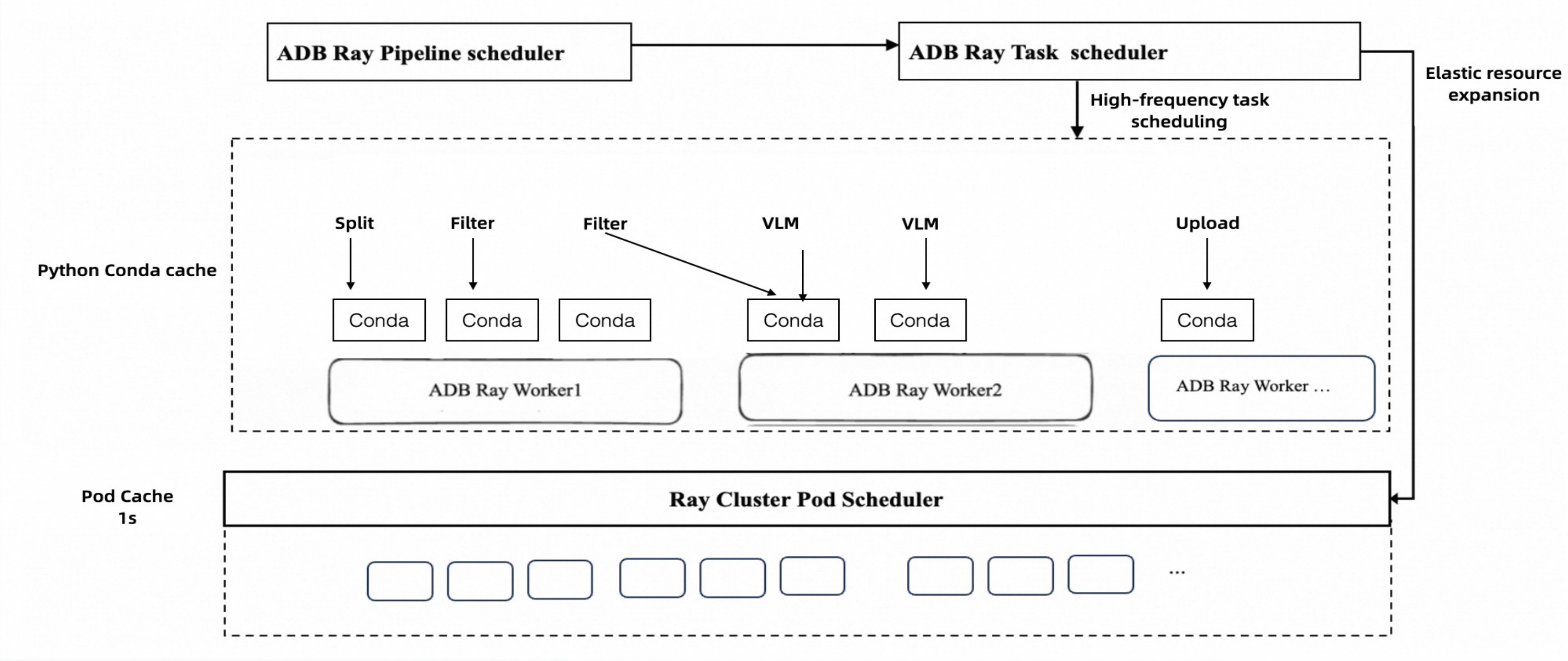
AnalyticDB Ray: 精细化CPU+GPU Scheduling



Profiling-based hybrid scheduling => Maximizes task deployment density

CPU: 1 million cores, with an elastic capacity of 500k cores/day.

AnalyticDB Ray: Startup Acceleration / Runtime Isolation



Two-level cache

built-in conda environment

Pod cache

Task startup in 5-10ms
Node startup in 1 sec

蚂蚁开源 ANT OPEN SOURCE

RAY

CONTENT
目录

01 AnalyticDB Data + AI云原生架构

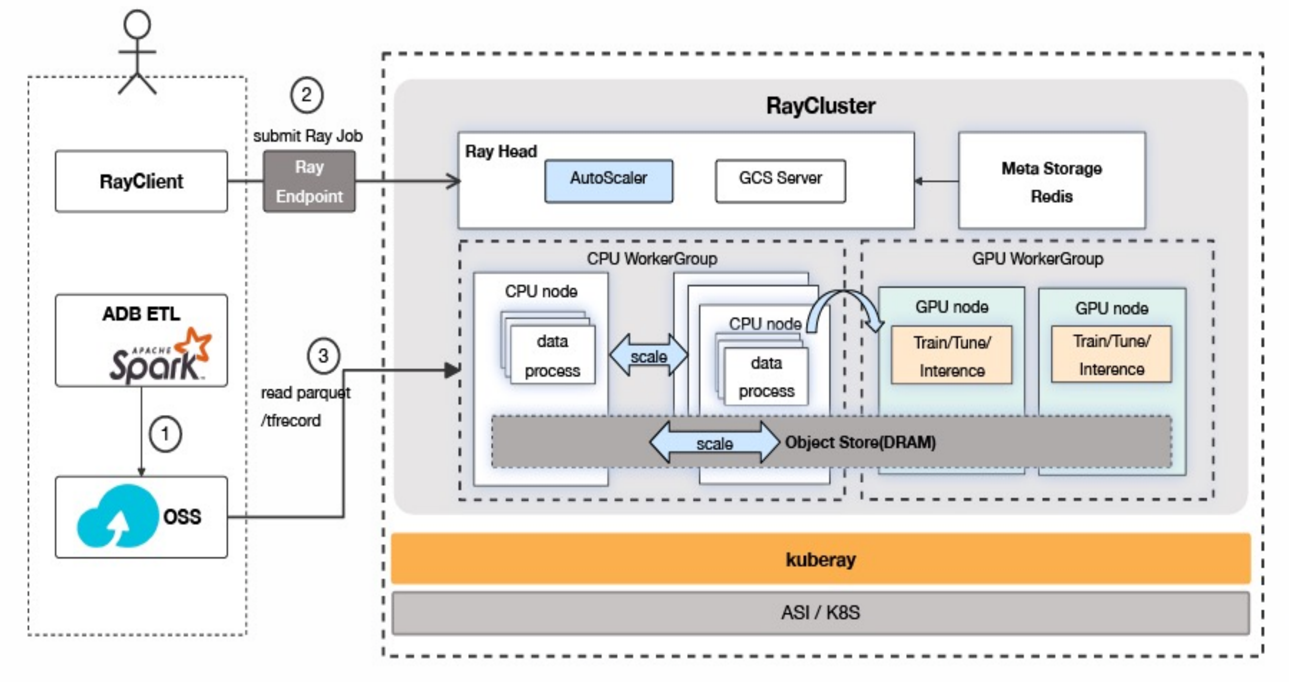
02 AnalyticDB Ray特性

03 多模态处理调度技术

04 **典型案例介绍**

Customer Case: Customer CTR Prediction

Offline Batch Inference



Computing autoscaling

Independent auto scaling of CPU/GPU

GPU utilization 5% -> 80%

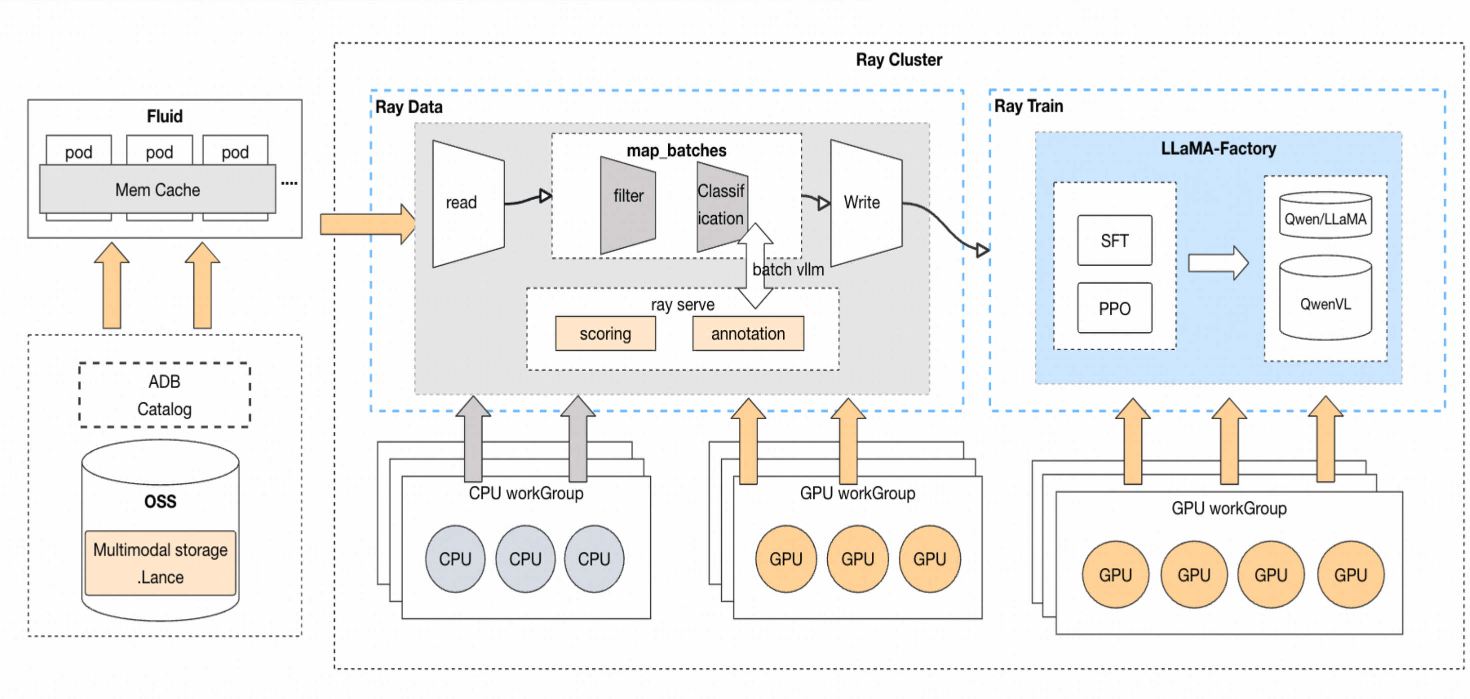
Object Store autoscaling

The object store DRAM auto-scale

Performance Improvement 2 ~ 3X

Customer Case: Game Assistance

RayData： process multimodal image and text data stored in Lance
LLaMA-Factory on Ray： distributed fine-tuning on Qwen-VL.



Multimodal Data Storage

Lance

unified storage of raw data,

metadata, vectors

Computing autoscaling

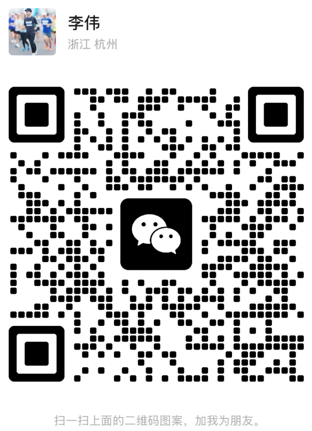
Fine-tuning on Ray

3 ~ 5X

Performance Improvement

谢谢

AnalyticDB产品文档：
https://help.aliyun.com/zh/analyticdb/analyticdb-for-mysql/product-overview/what-is-analyticdb-for-mysql



李伟

浙江 杭州

扫一扫上面的二维码图案，加我为好友。