QCon
全球软件开发大会

腾讯云存储数据湖架构

程力

腾讯云 存储数据湖负责人

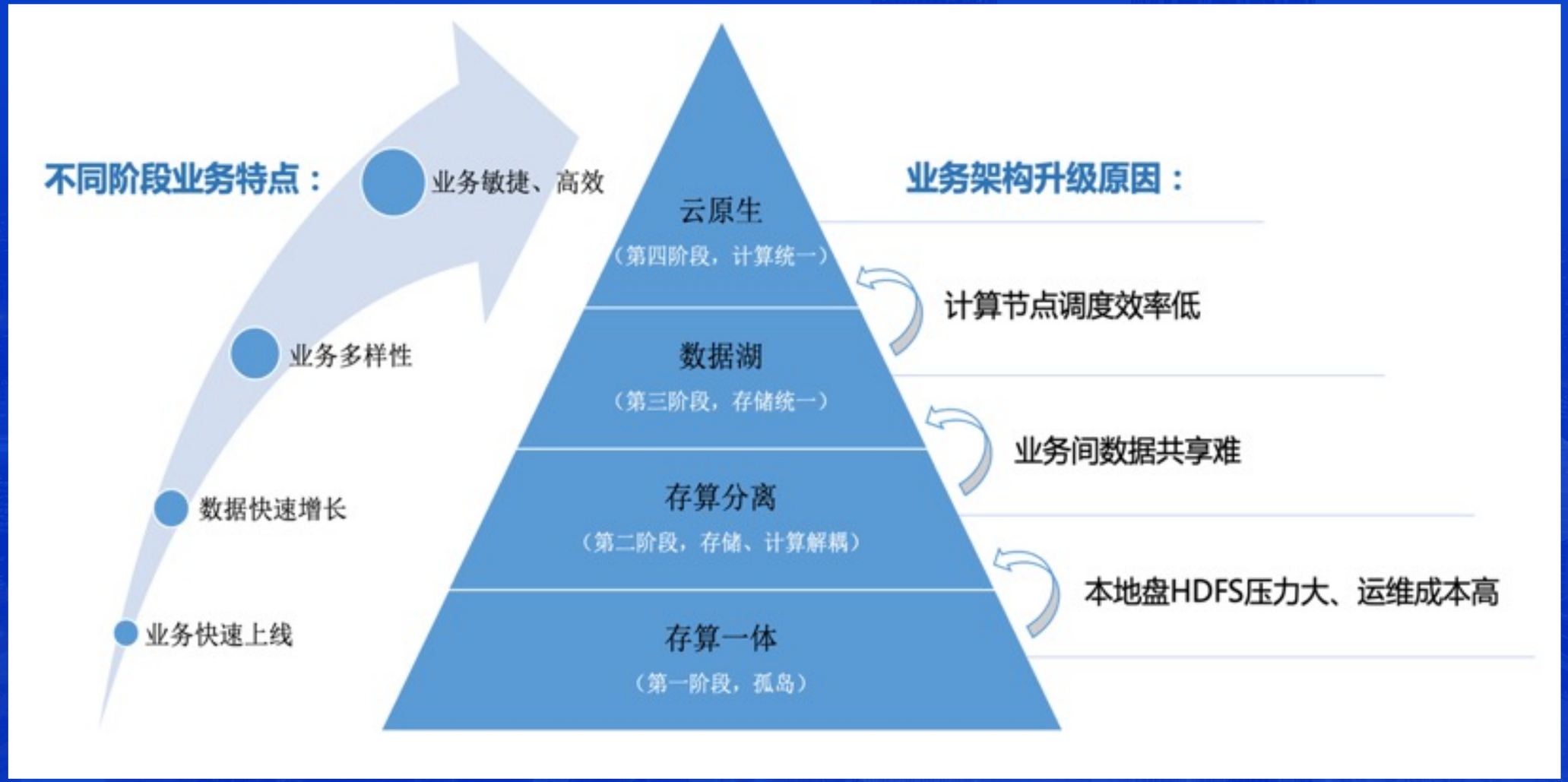
Apache Ozone PMC / Apache Hadoop Committer

QCon
全球软件开发大会

- 云原生生态下的存算分离
- 云原生数据湖三层加速
- 大数据和AI下的数据湖架构

QCon
全球软件开发大会

数据存储发展的趋势



对象存储的优势

存储可靠性
提供多副本或者纠删码冗余存储，最高可达 *12个9* 的数据持久性。

高可用
提供高达 *99.95%* 的服务可用性，为您的业务保驾护航。

开放兼容
提供*全兼容*行业标杆 AWS S3 的接口，提供 Verraform 等多种生态工具支持。

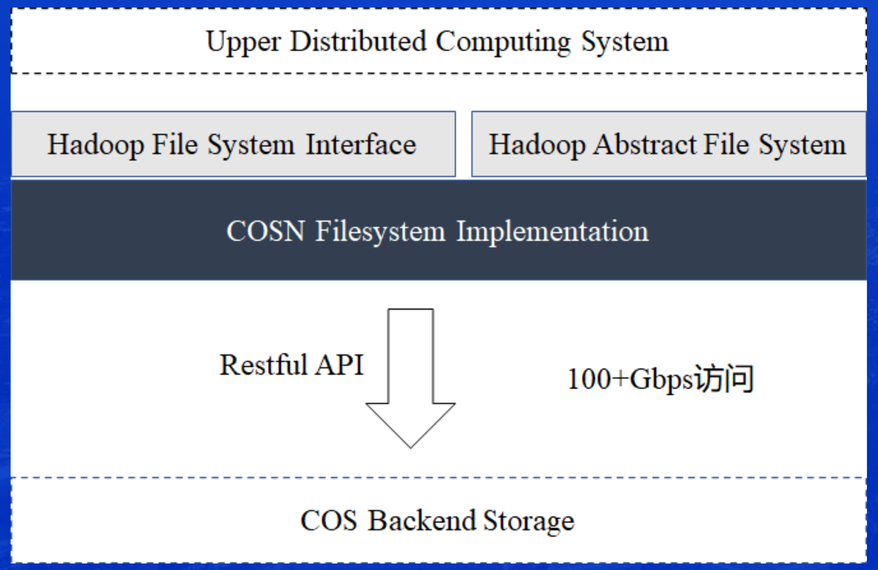
高性能
提供高达 *30,000* QPS的请求性能；保障客户业务在高并发情况下稳定运行。

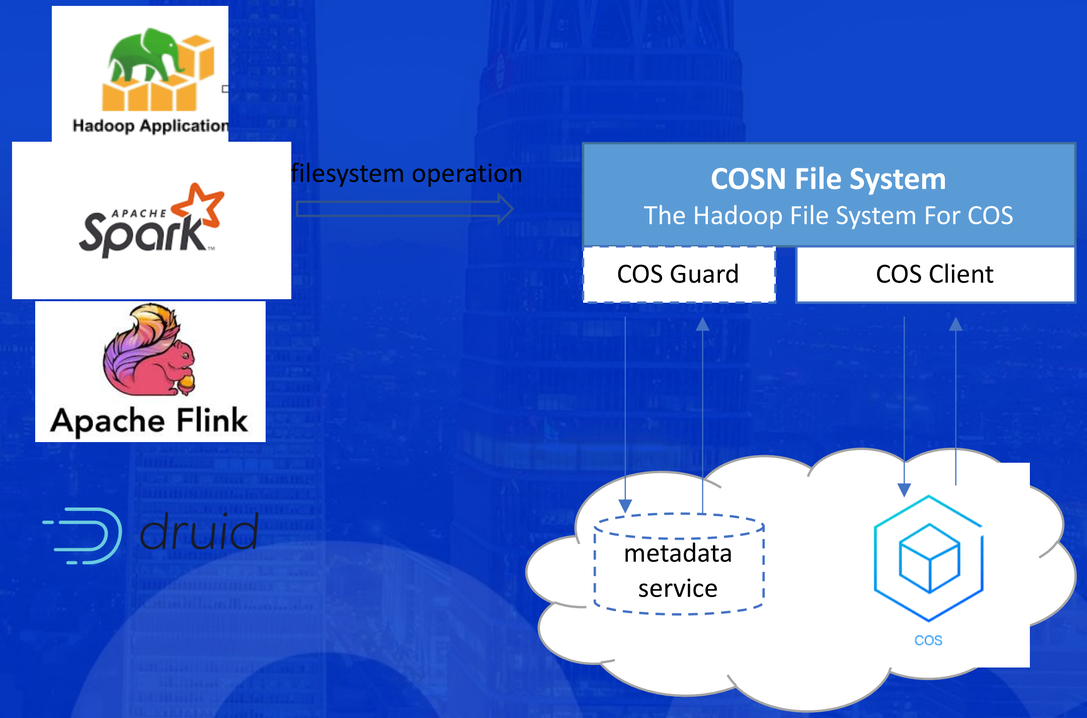
数据安全
提供传输、存储、访问*全链路安全保障*，支持多租户权限隔离，支持加密传输，支持数据加密存储。

低成本
提供低至 *￥0.01/GB* 的深度归档存储，支持数据生命周期管理，进一步降低云存储成本。

QCon
全球软件开发大会

以对象存储为底座的存算分离架构





腾讯云COSN对象文件系统接口

- 实现了HCFS接口，全覆盖HDFS大数据计算应用；
- 实现了文件系统的扩展属性管理接口，允许用户对文件和目录设置xAttr的扩展属性；
- 实现了包含CVM/EMR instance 角色授权以及临时密钥访问的凭证获取机制；

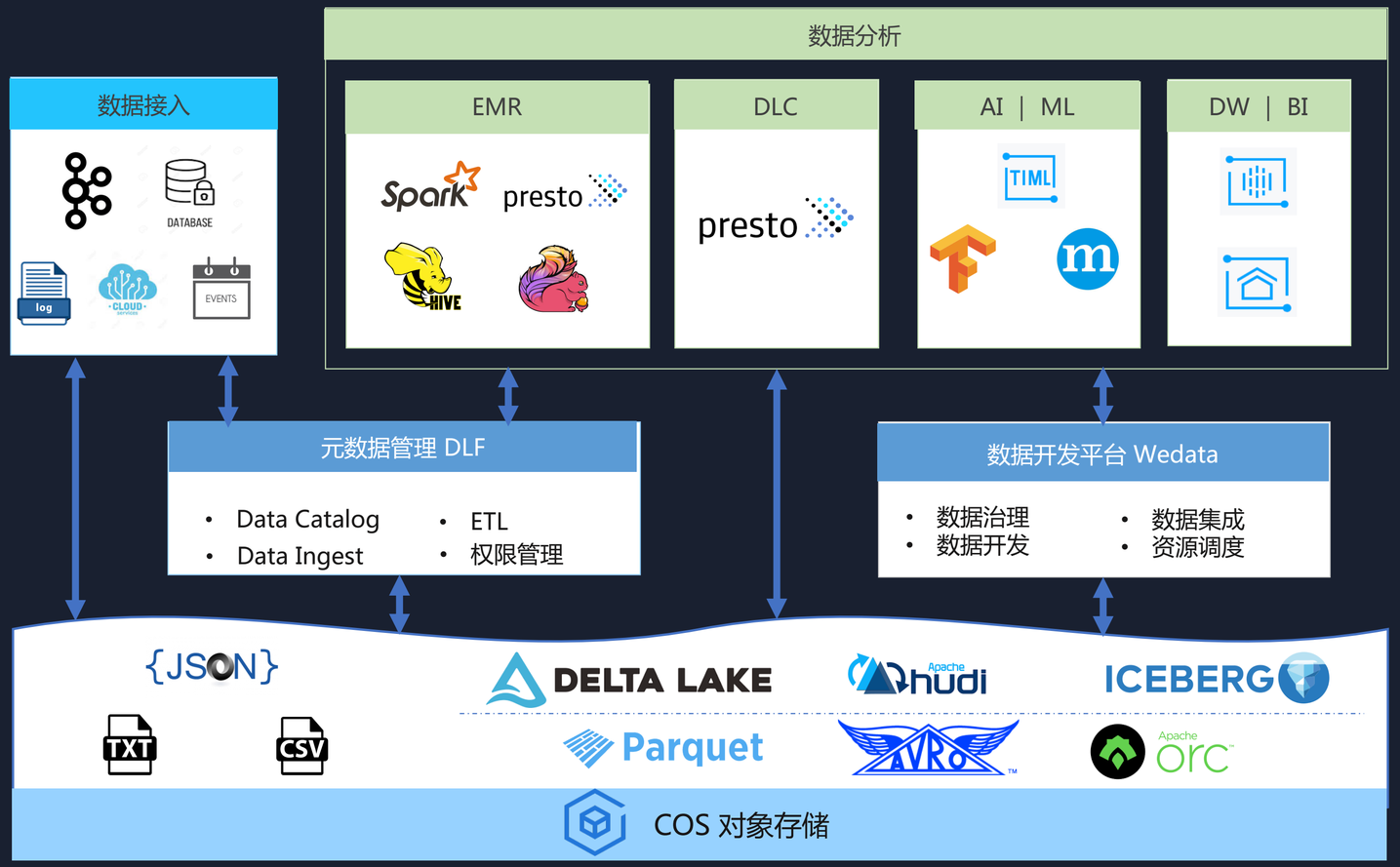
腾讯云上的数据湖生态

数据湖存储终端COS：

- 云原生：serverless架构，免运维；
- 数据共享：通过统一的COS对象存储作为弹性底座，结合三层加速器接入多种生态。
- 结构化数据管理：感知数据Table格式，支持按照Hive Table预热，支持Iceberg Table管理等。
- 高性价比：弹性、按需扩容
- 生态支持：支持Hadoop生态，K8S生态等多种生态的部署、运维、鉴权等。

面向业务场景：

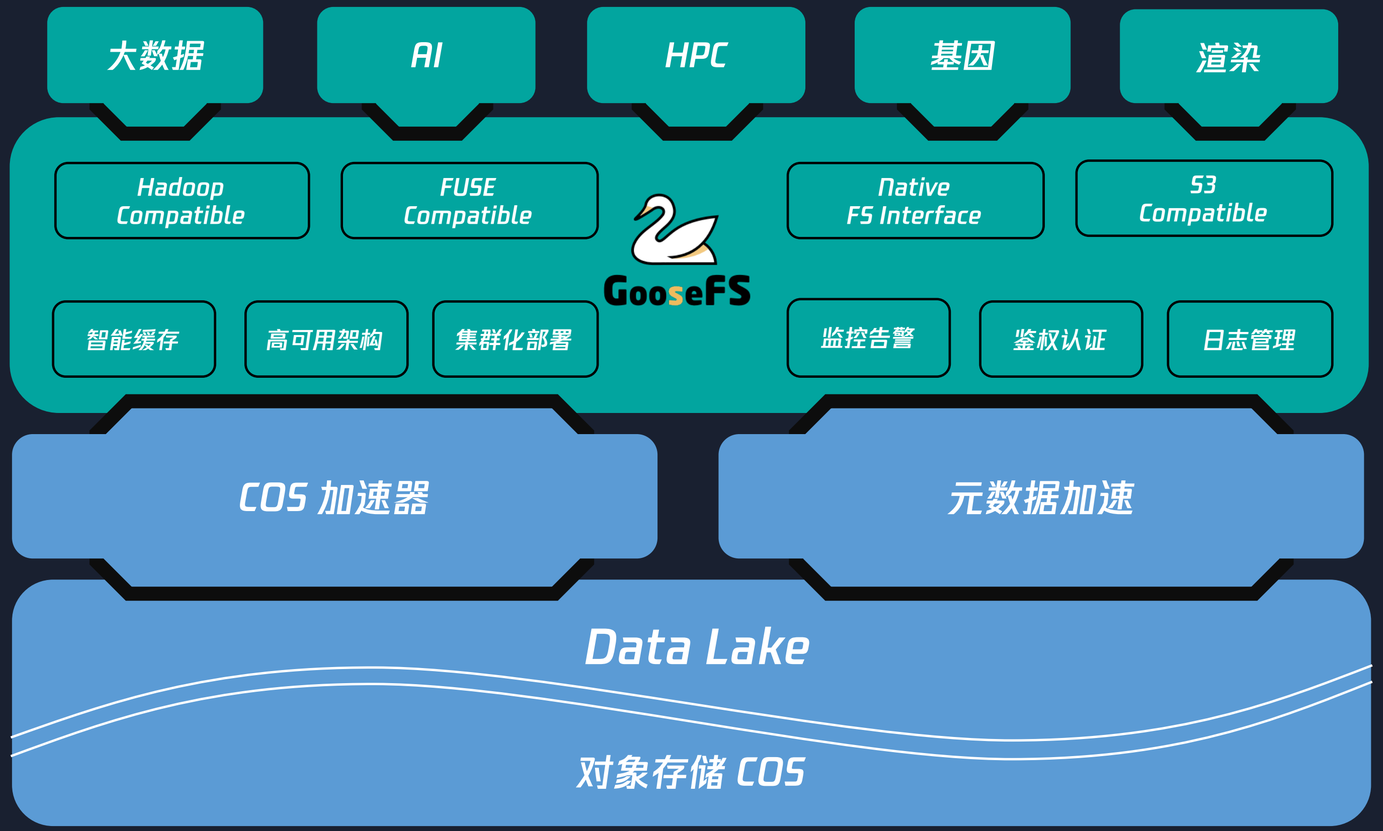
- **数据本地性加强**
- **数据湖结构化**
- **容器化调度**



腾讯云数据湖加速

数据湖三层加速：

- GooseFS ： 计算端 – 湖仓缓存加速
- 元数据加速： 数据端 – 元数据加速
- COS加速器： 存储端 – 数据加速



数据湖三层加速

GooseFS：Cache Accelerator:

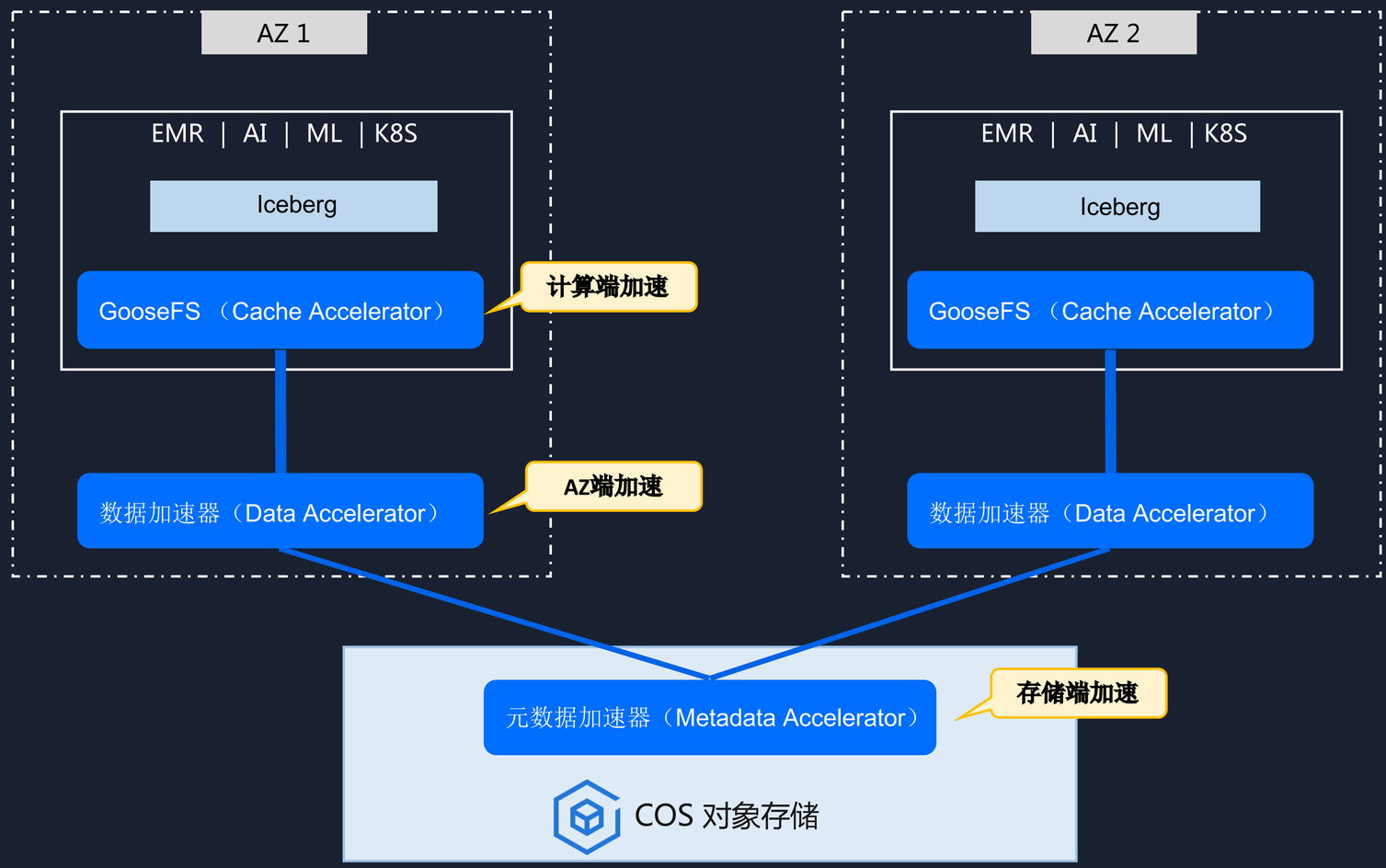
- 运行在EMR/AI/ML/K8S集群内，基于集群MEM/SSD资源，提供Data Cache能力；
- 热数据缓存在Cache中，对象存储保存全量数据；
- 针对各种计算引擎，提供Data Locality能力；
- 提供磁盘模式和内存模式，支持淘汰

数据加速器：Data Accelerator：

- AZ级部署，全SSD存储介质，热数据读加速；
- 提供Tbps带宽，满足高吞吐需求；
- 提供ms级别时延；

元数据加速器：Metadata Accelerator:

- 提供文件系统级别元数据操作能力；
- Rename操作，无需Copy/Delete数据；
- List操作，无频控；
- 每个Bucket，提供10万 QPS；



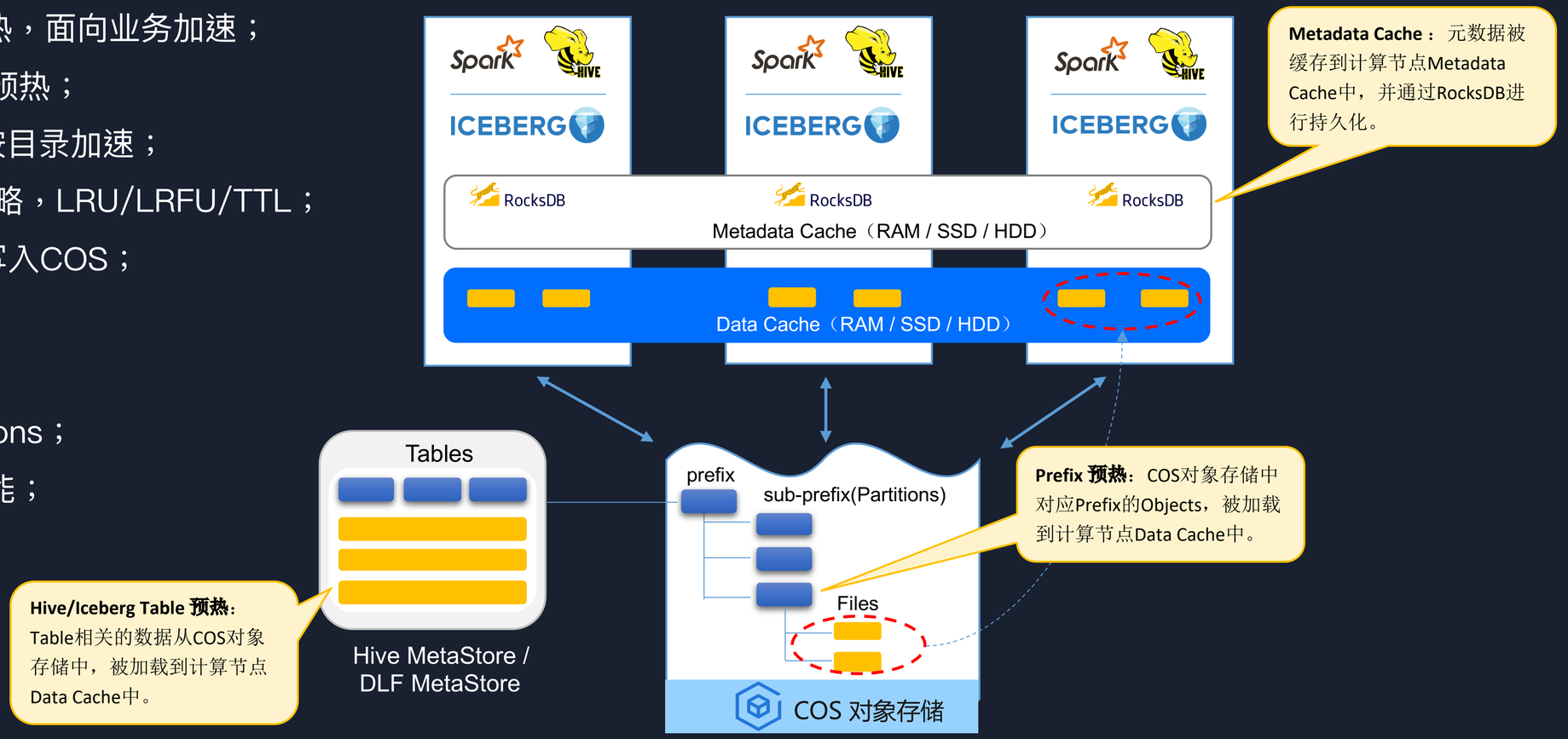
GooseFS Cache加速

Data Cache:

- 支持Hive Table Level预热，面向业务加速；
- 支持Iceberg Table Leve预热；
- 支持Prefix Level预热，按目录加速；
- 支持多种数据缓存淘汰策略，LRU/LRFU/TTL；
- 支持缓存数据同步/异步写入COS；

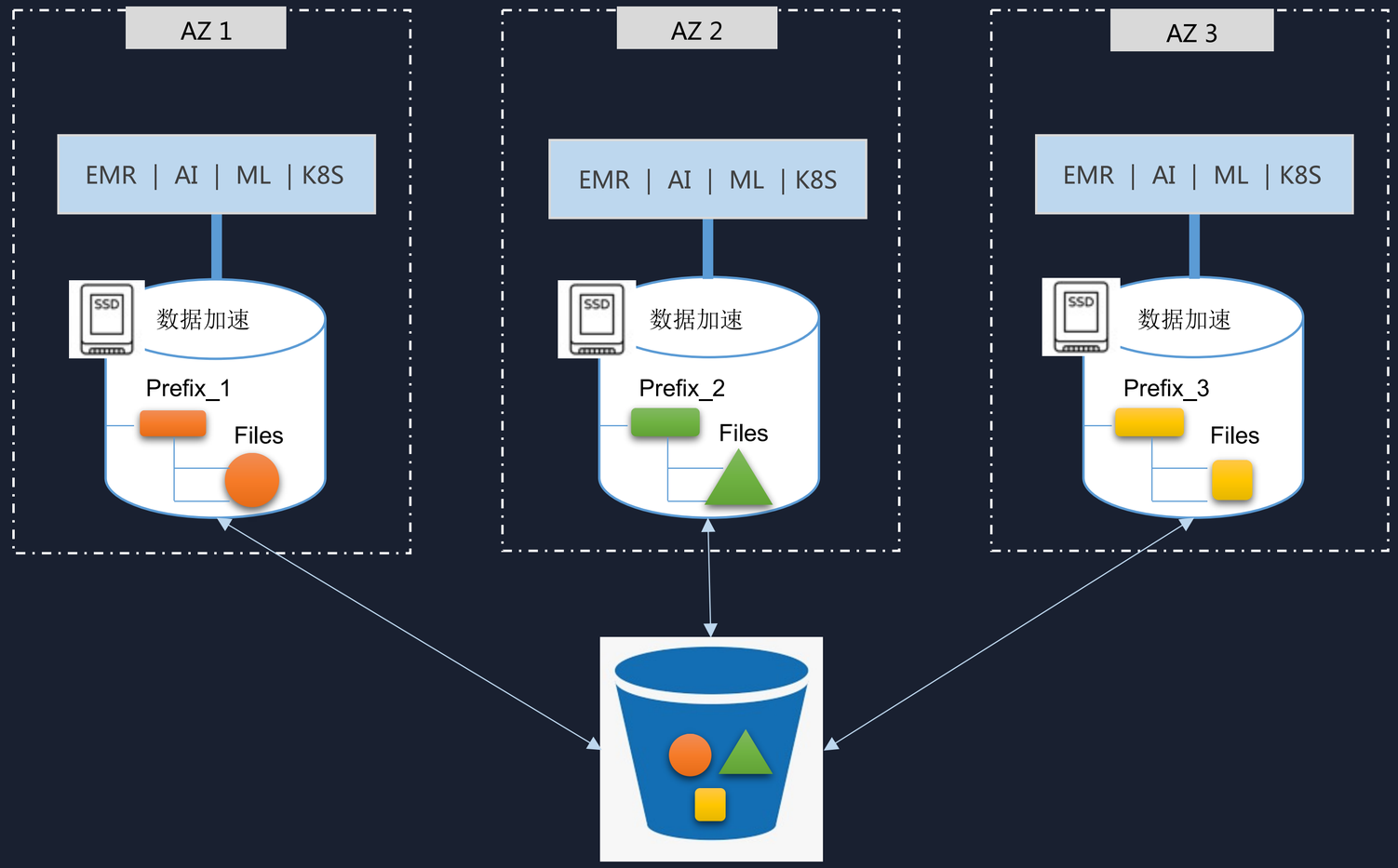
Metadata Cache：

- 避免了大量的list operations；
- 提高了Metadata 访问性能；



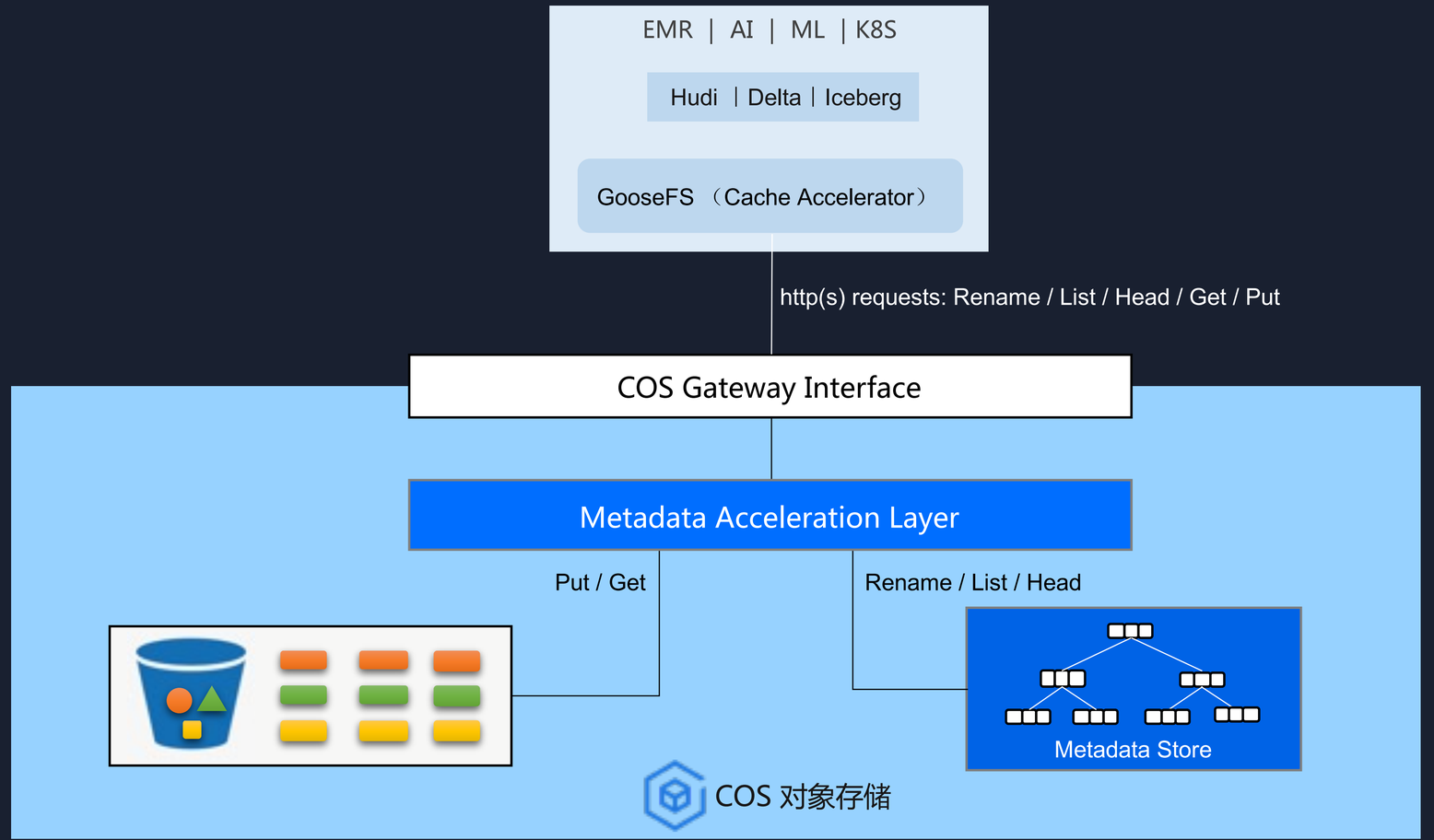
数据加速器（Data Accelerator）— AZ级别数据加速

- AZ Locality
- 专有加速域名访问资源；
- 缓存数据强一致；
- 可以加速Bucket，或者prefix；
- 同一个Bucket，支持多个加速器
- 支持存量Bucket，随时Enable/Disable；
- 如果miss cache，从COS回源；



元数据加速器（Metadata Accelerator）— 文件桶

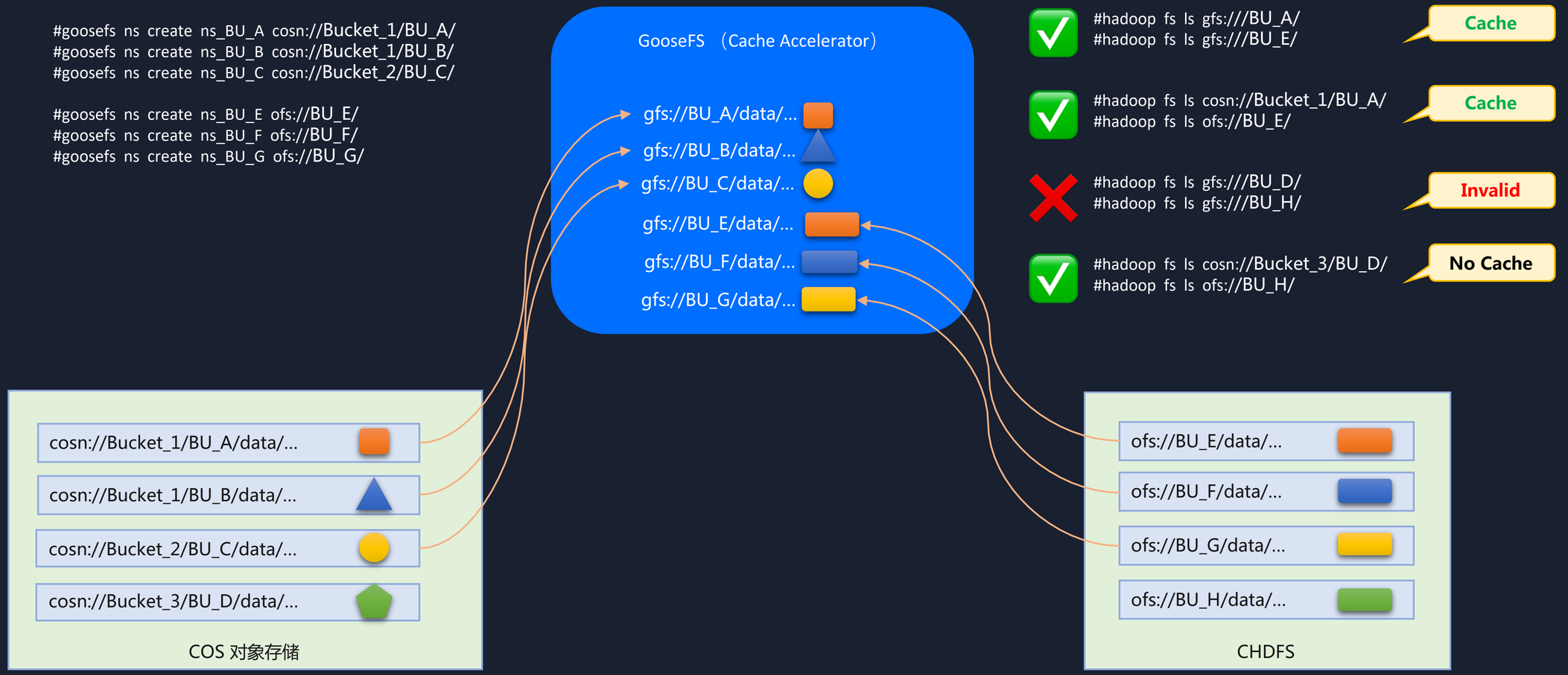
- 文件系统级别元数据操作；
- 提供Rename API，Rename 无需Copy / Delete数据，直接在Metastore完成；
- List、Head操作，直接查询 Metadate Store，避免对象存储QPS问题；
- 性能：10 万 QPS；



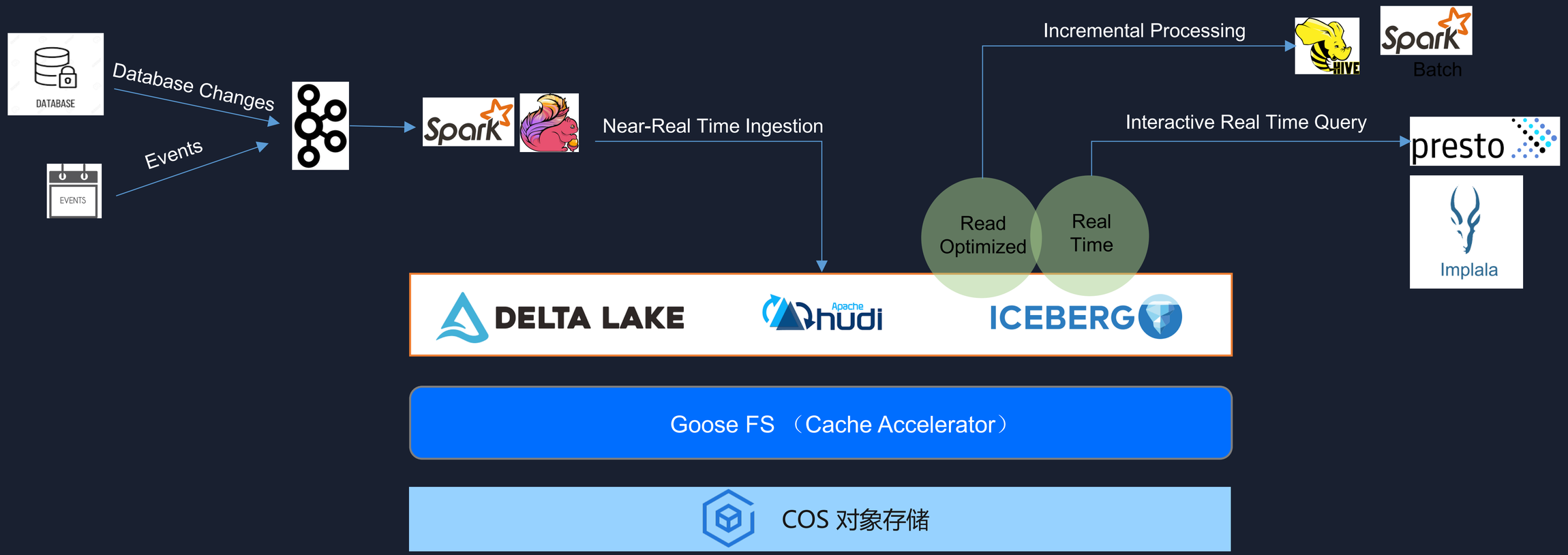
GooseFS Namespace

```
#goosefs ns create ns_BU_A cosn://Bucket_1/BU_A/
#goosefs ns create ns_BU_B cosn://Bucket_1/BU_B/
#goosefs ns create ns_BU_C cosn://Bucket_2/BU_C/

#goosefs ns create ns_BU_E ofs://BU_E/
#goosefs ns create ns_BU_F ofs://BU_F/
#goosefs ns create ns_BU_G ofs://BU_G/
```

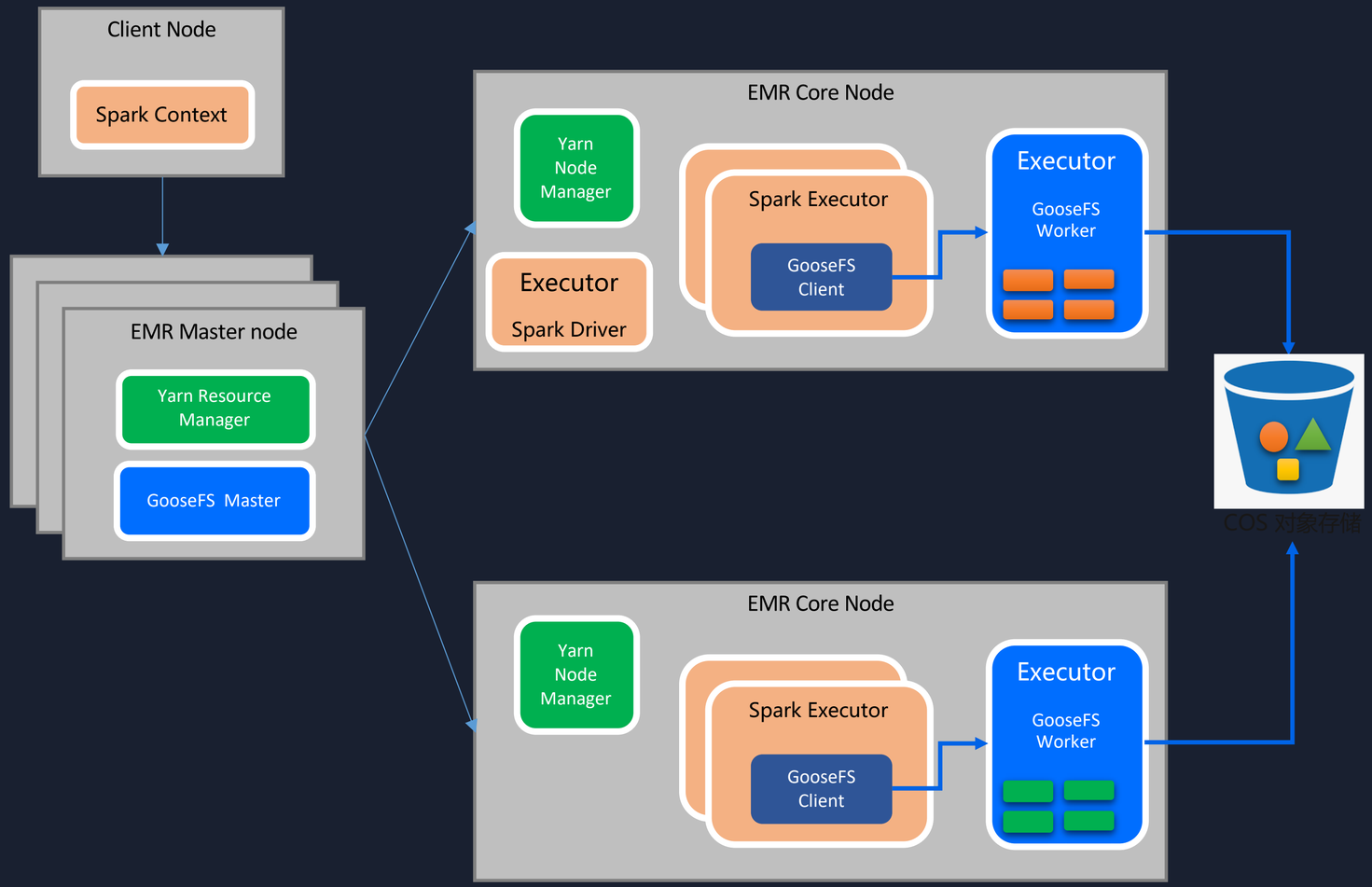


GooseFS 支持数据湖结构化



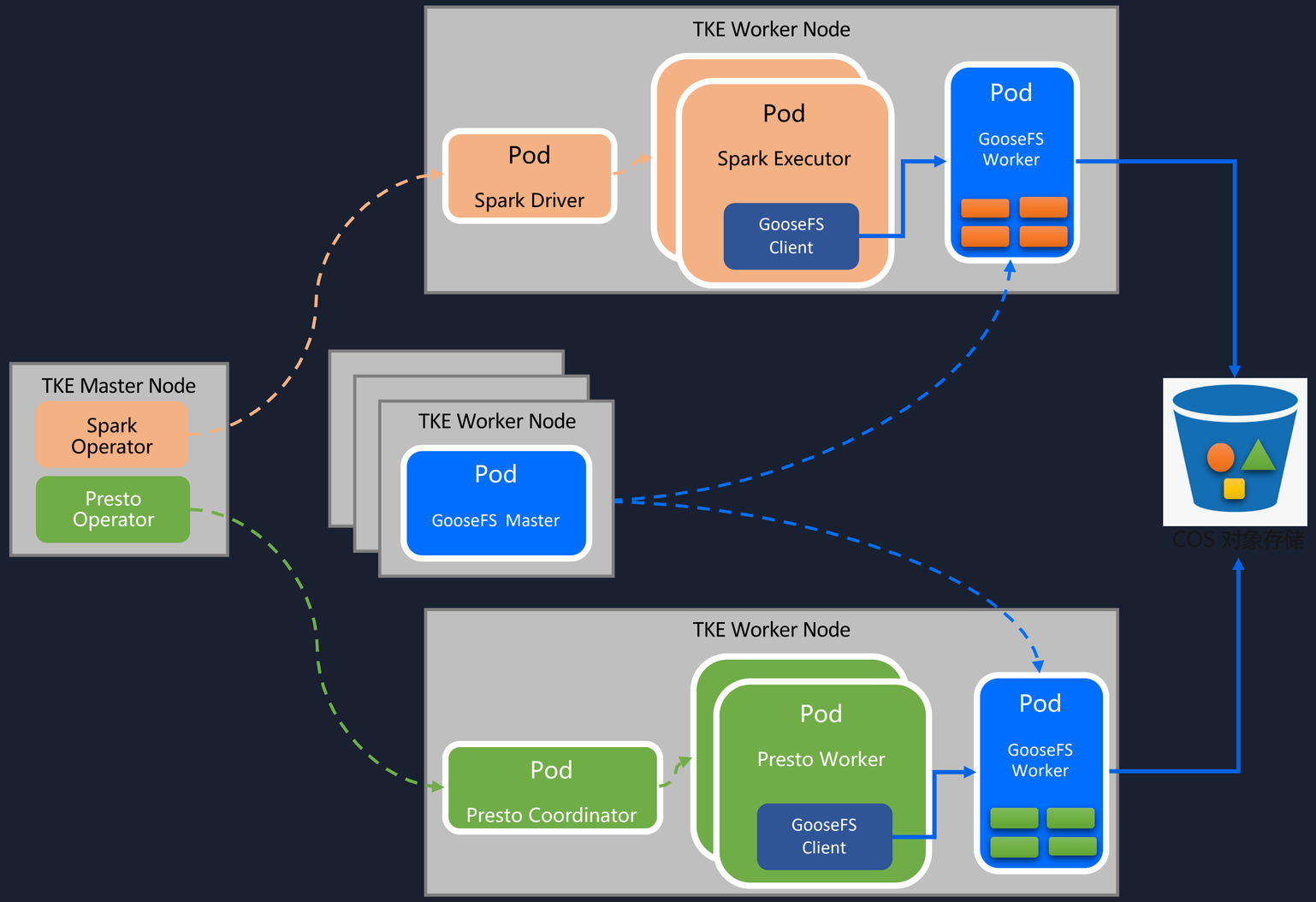
GooseFS on EMR (Spark/Hive/Presto/Impala)

- GooseFS Master：和Yarn RM同节点；支持元数据持久化；支持Raft Based HA；支持Ranger
- GooseFS Worker：Worker和计算Executor/Worker同节点，保证类似HDFS的数据本地性；支持内存模式和磁盘模式混合
- GooseFS client：Shaded client打入计算fat jar，同时支持原生的COSN schema和GooseFS schema开启三层加速特性



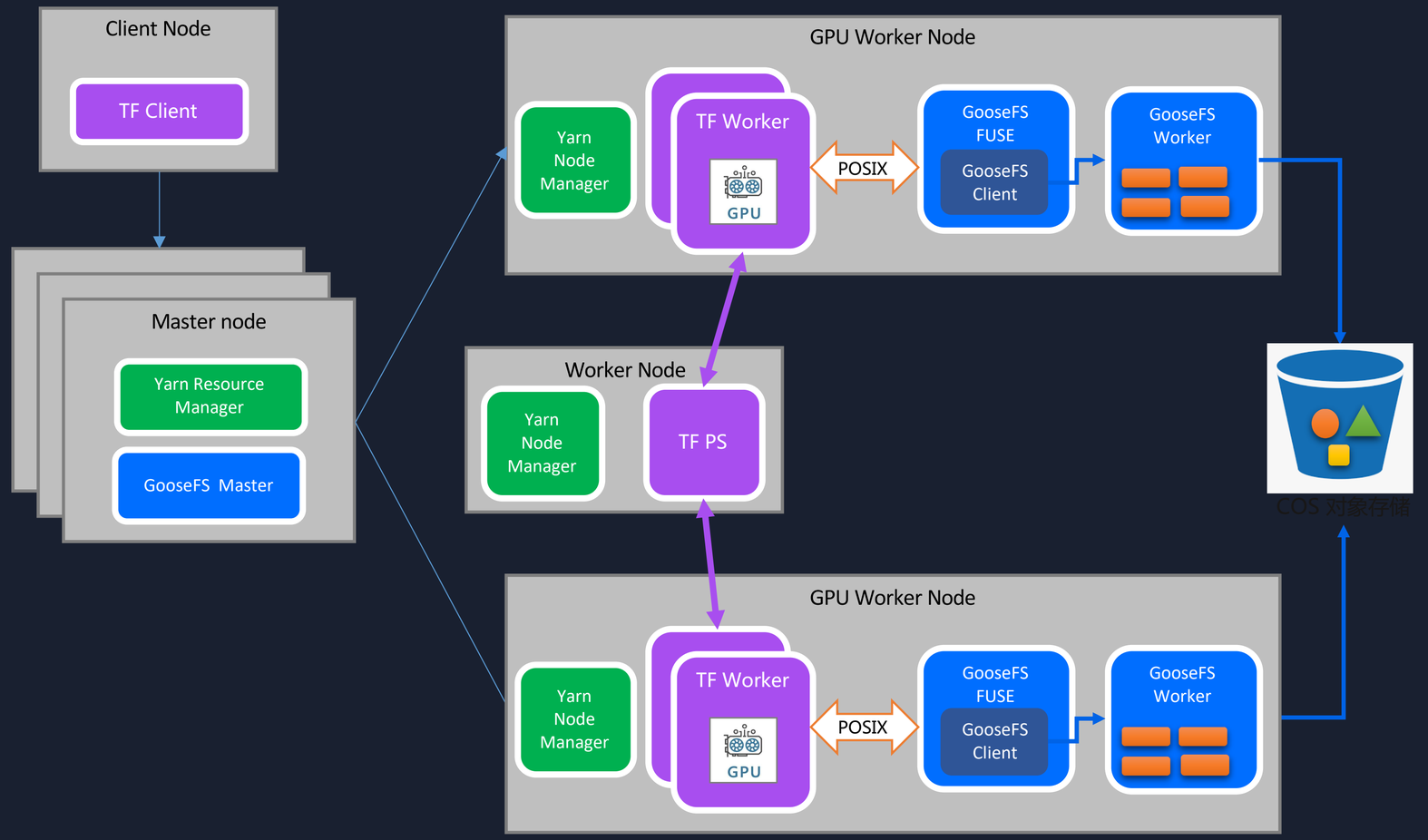
GooseFS on TKE (K8s)

- GooseFS Master：独立部署；支持元数据持久化；支持Raft Based HA；支持Ranger
- GooseFS Worker：通过DaemonSet保证每个宿主机部署一台GooseFS Worker Pod提供数据Locality
- GooseFS Fuse：Master和Worker Pod都可以起Fuse
- 独立部署框架控制GooseFS runtime

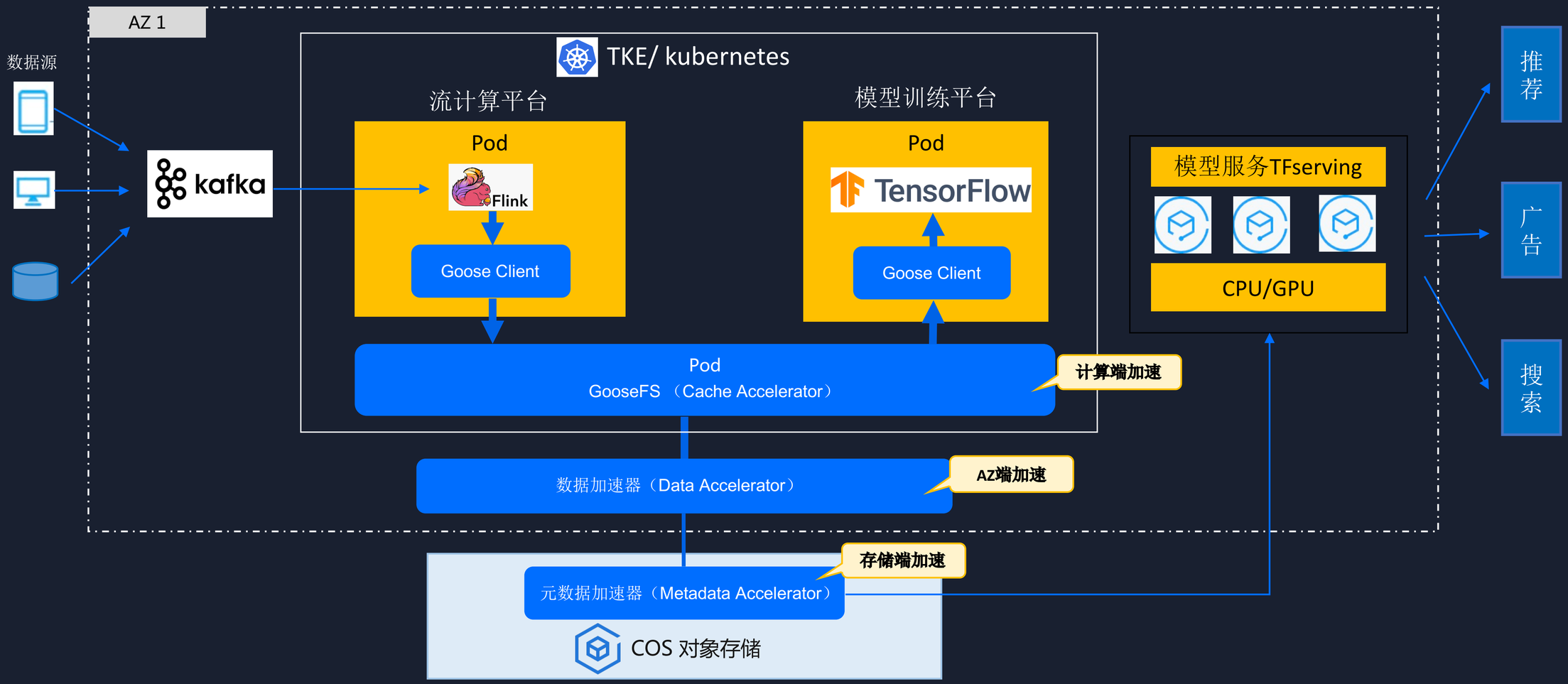


GooseFS on YARN with TensorFlow

- GooseFS Master：和Yarn RM同节点；支持元数据持久化；支持Raft Based HA；支持Ranger
- GooseFS Worker：每个GPU Worker Node部署一个GooseFS Worker 同TF Worker部署在一个Node；
- GooseFS Fuse：Master和Worker Pod都可以起Fuse



实时训练平台

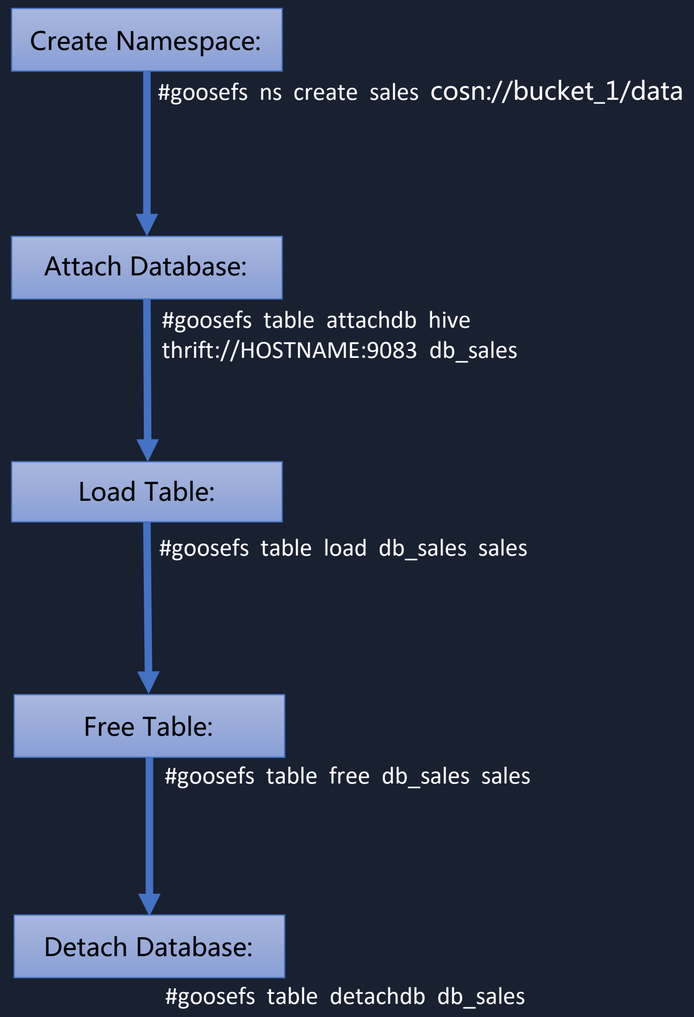

AZ 1
数据源
kafka
TKE/ kubernetes
流计算平台
模型训练平台
Pod
Flink
Goose Client
Pod
TensorFlow
Goose Client
模型服务TFserving
CPU/GPU
推荐
广告
搜索
Pod
GooseFS （Cache Accelerator）
计算端加速
数据加速器（Data Accelerator）
AZ端加速
元数据加速器（Metadata Accelerator）
存储端加速
COS 对象存储

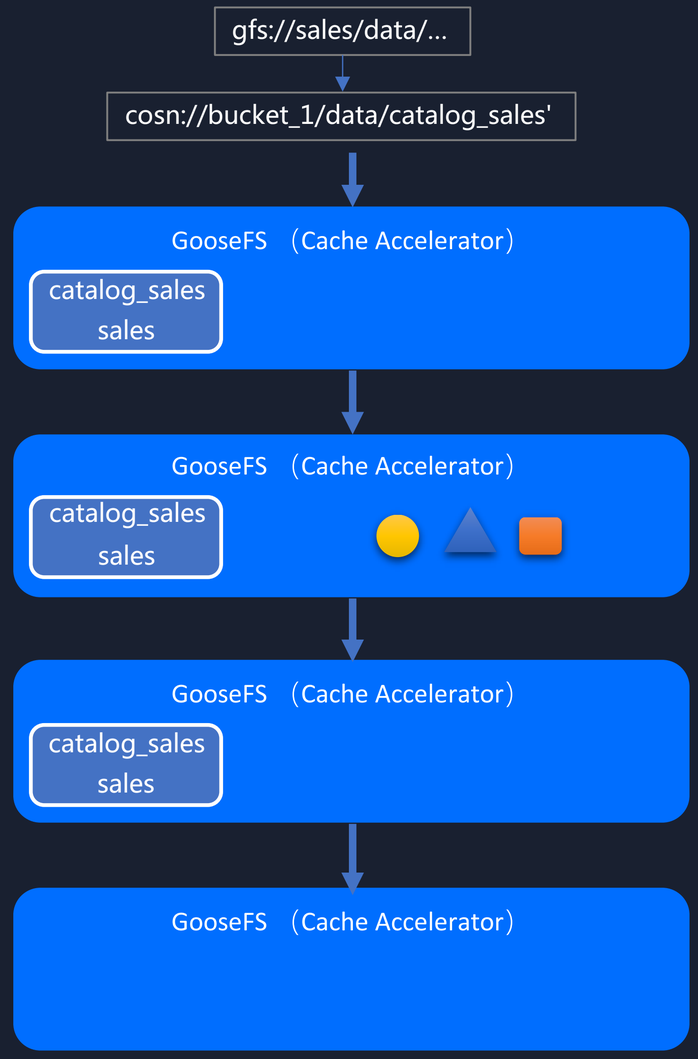
GooseFS Table（Hive/Iceberg）

```
#goosefs table attachdb hive thrift://HOSTNAME:9083 hive_db_name
#goosefs table ls db_name table_name
#goosefs table load db_name table_name
#goosefs table free db_name table_name
#goosefs table stat db_name table_name
```

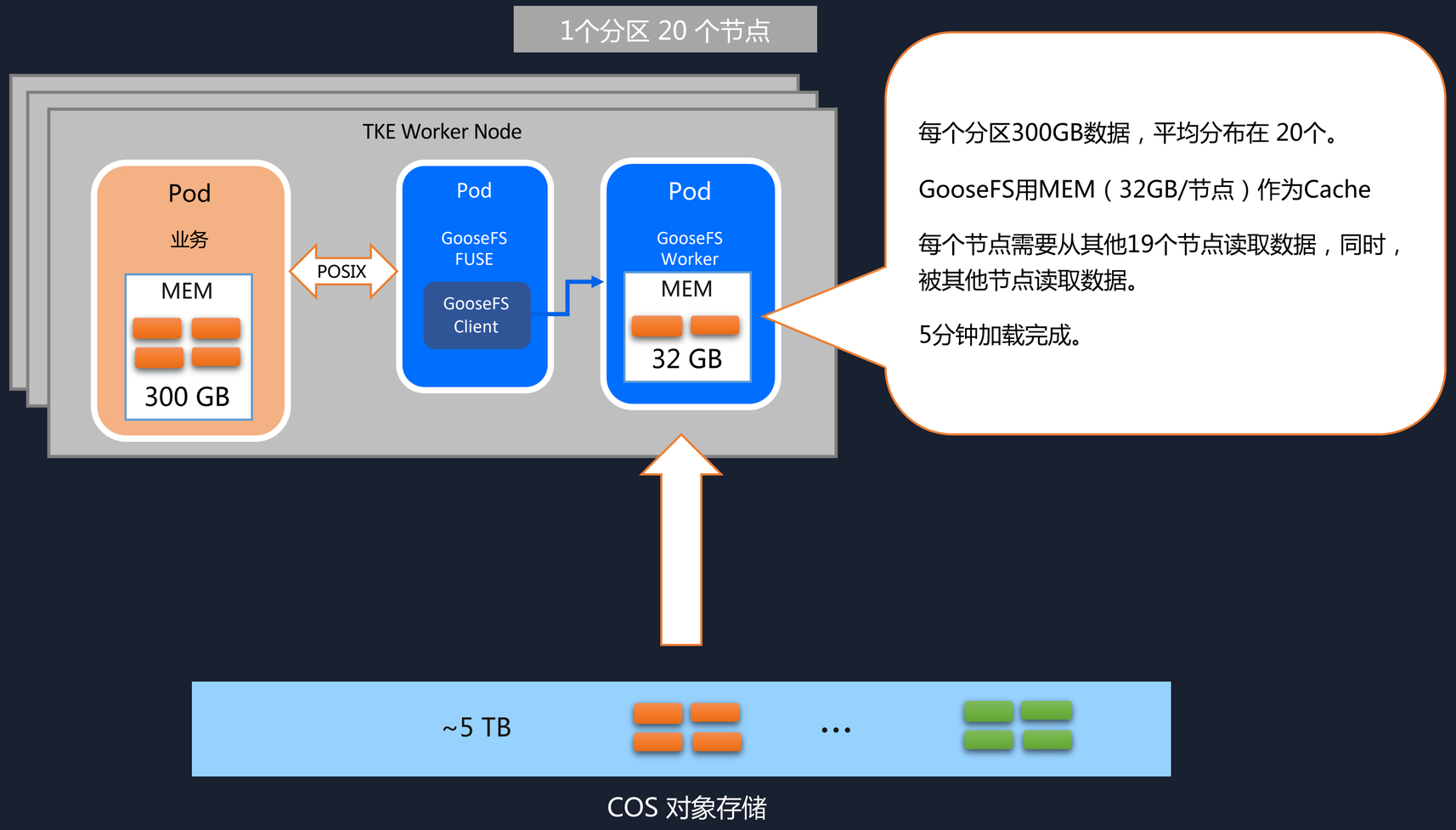
```
CREATE EXTERNAL TABLE `sales`(
  `cs_sold_time_sk` int,
  `cs_ship_date_sk` int,
  `cs_bill_customer_sk` int,
  `cs_bill_cdemo_sk` int,
  `cs_bill_hdemo_sk` int,
    .........
PARTITIONED BY (
  `cs_sold_date_sk` string)
LOCATION
  'cosn://bucket_1/data/catalog_sales'
```

Hive Meta Store

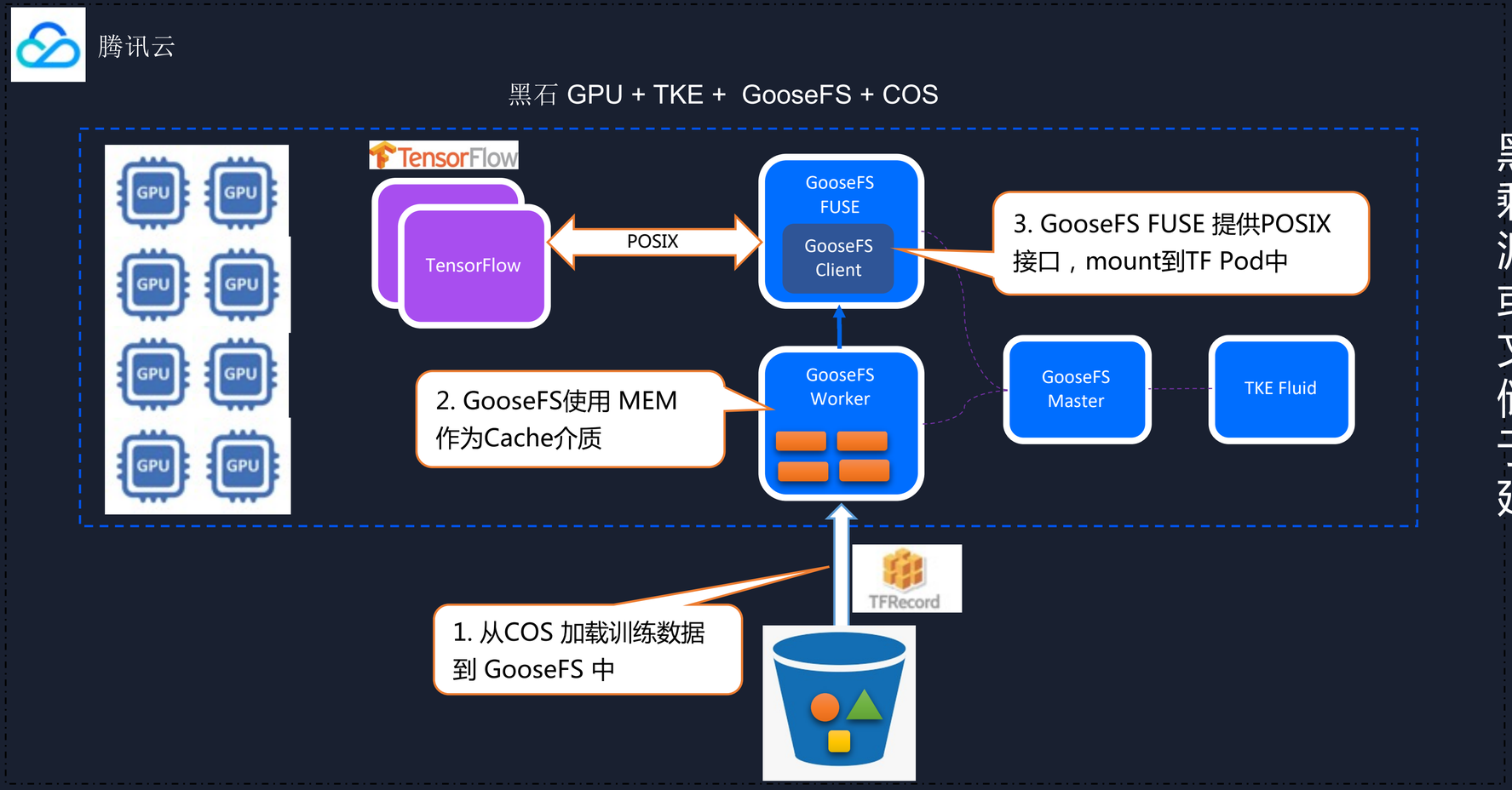




TKE + GooseFS + COS 支持OCR搜索框架实例



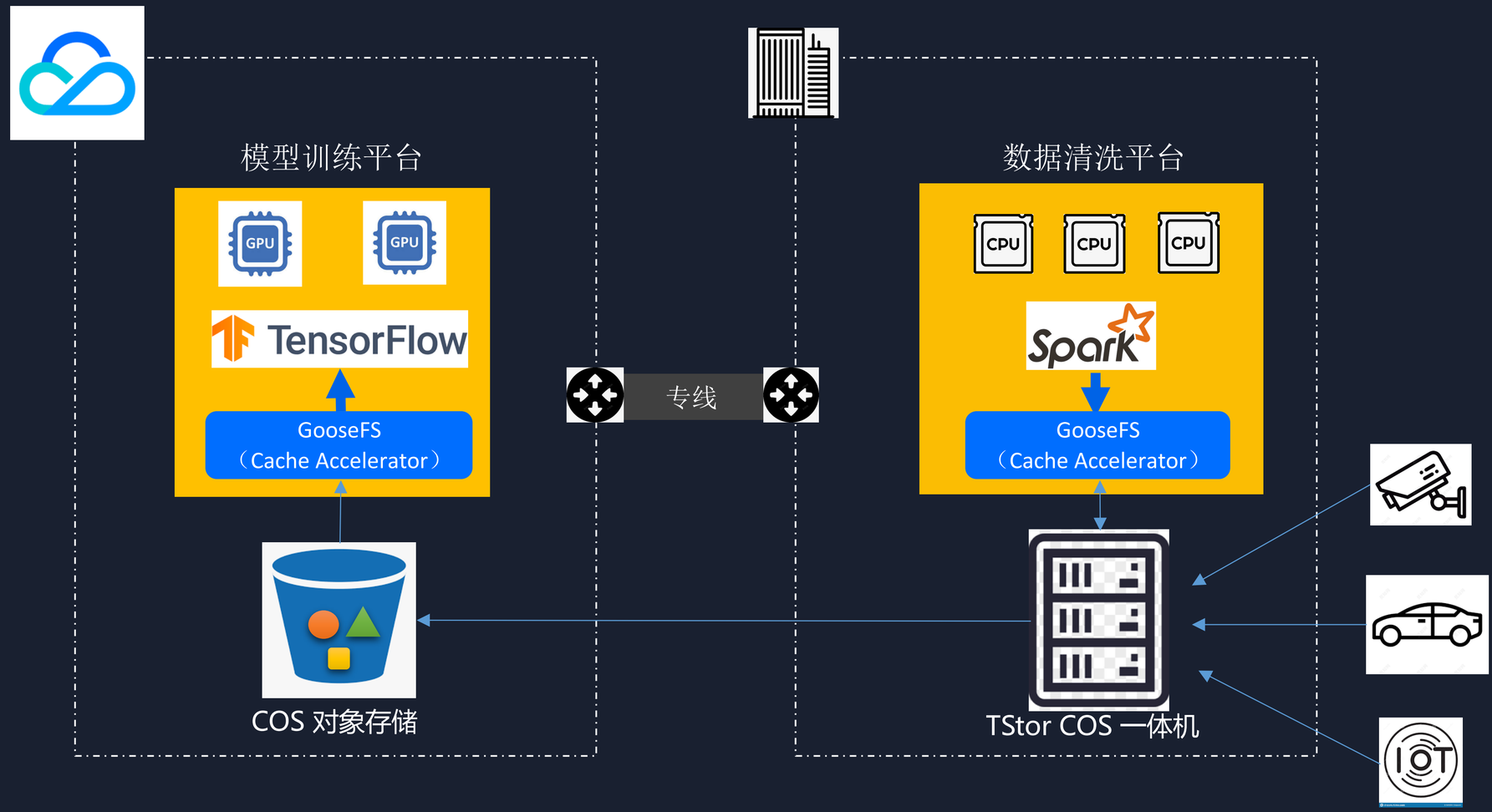
黑石GPU + GooseFS + COS 支持TensorFlow模型训练调优



黑石GPU机器上剩余大量内存资源（单机400GB或800GB），小文件/图片居多存储在COS中，对于小文件读写时延要求高

GooseFS + COS + Tstor支持云上云下打通应用实例

- IOT数据（车载数据、摄像头数据）上传到本地数据中心TStor对象存储；
- 本地大数据集群通过GooseFS加速数据访问，完成数据清洗和标注，生成训练数据集；
- TStor自动同步训练数据集到云上COS对象存储；
- 在云上按需拉起GPU训练集群，通过GooseFS加速，完成AI模型训练；



GooseFS powered by Tencent

THANKS

—licheng@apache.org

更多产品信息
欢迎参考腾讯云存储公众号！
同时欢迎各界英才加入！