GOTC

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

OPEN SOURCE , OPEN WORLD

「分布式数据库与存储」专场

分布式文件系统在云原生时代的挑战与趋势

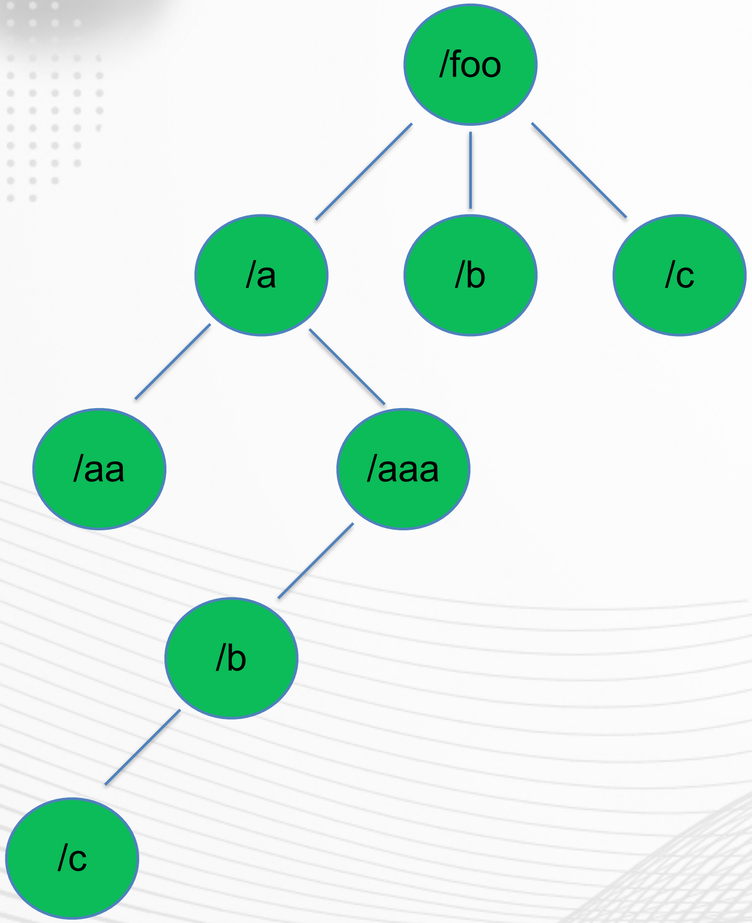
苏锐 2021年8月1日

议程

- 比较几种流行的分布式文件系统
- 云原生环境下遇到的挑战
- 如何为云原生环境设计分布式文件系统
- 在业务场景中的收益
- 未来展望

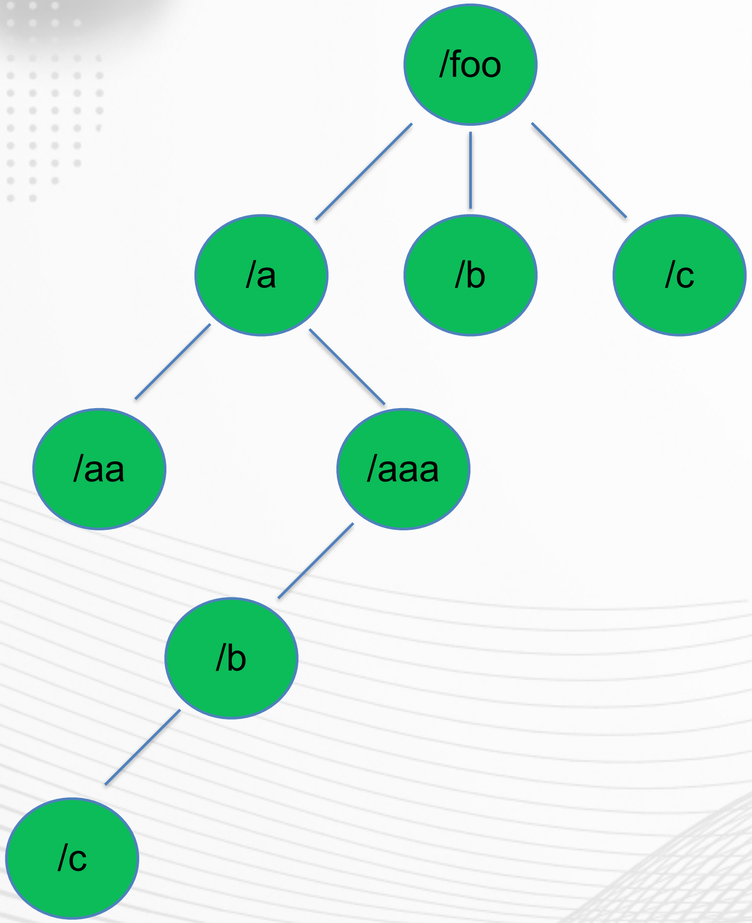
比较几种流行的分布式文件系统

强调下什么是文件系统 👇



比较几种流行的分布式文件系统

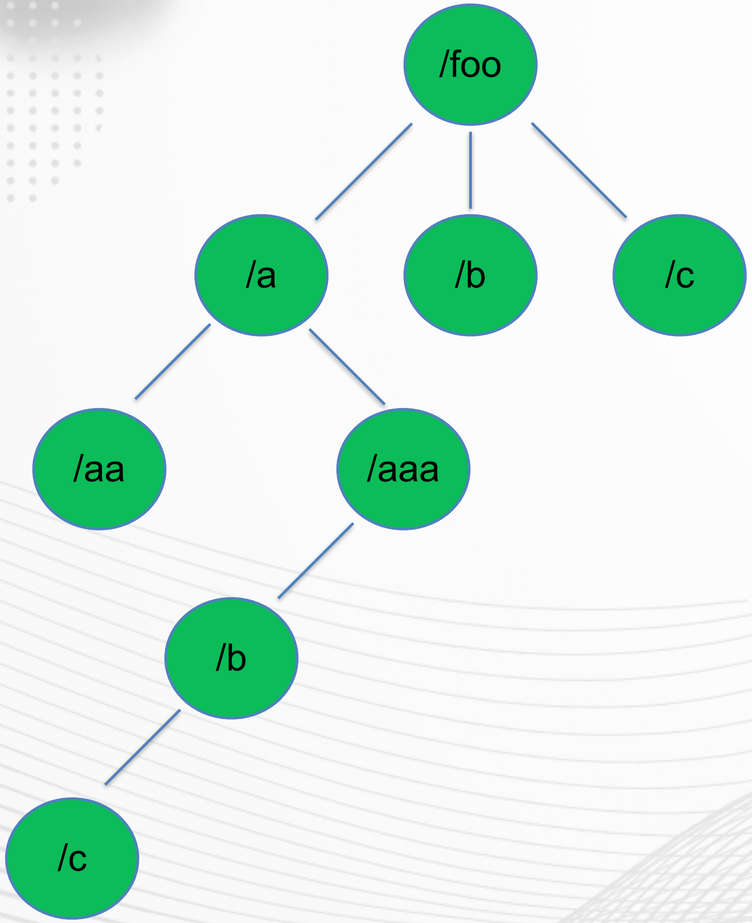
文件系统 👇



对象存储 👇

- /foo/a
- /foo/b
- /foo/c
- /foo/a/aa
- /foo/a/aaa
- /foo/a/aaa/b
- /foo/a/aaa/b/c

比较几种流行的分布式文件系统




```
mv foo bar
```

/foo/a

/foo/b

/foo/c

/foo/a/aa

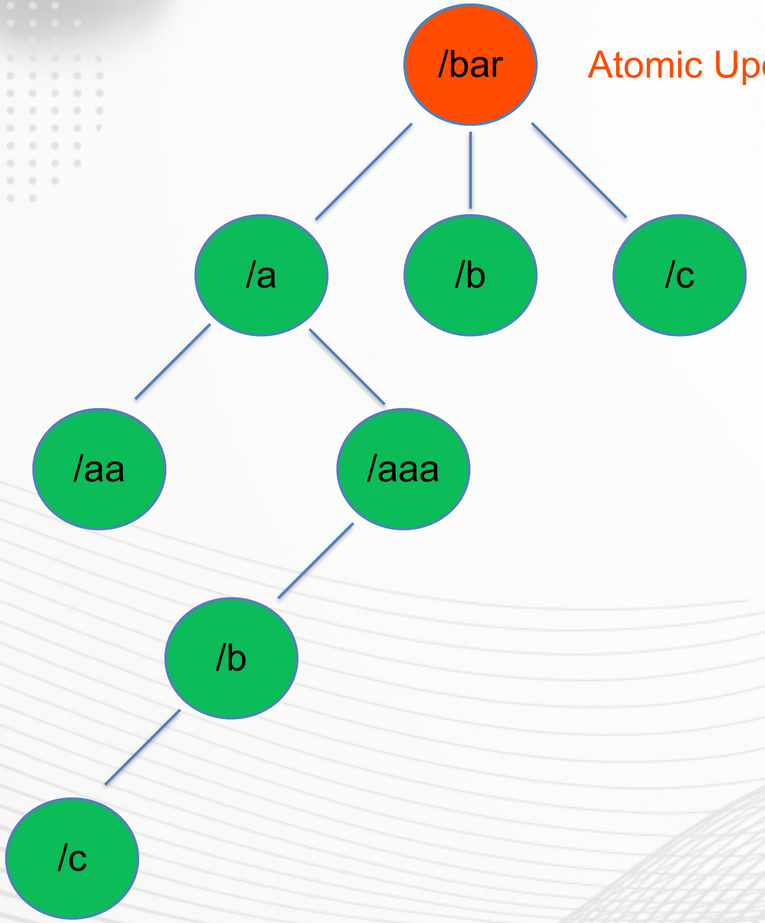
/foo/a/aaa

/foo/a/aaa/b

/foo/a/aaa/b/c

比较几种流行的分布式文件系统

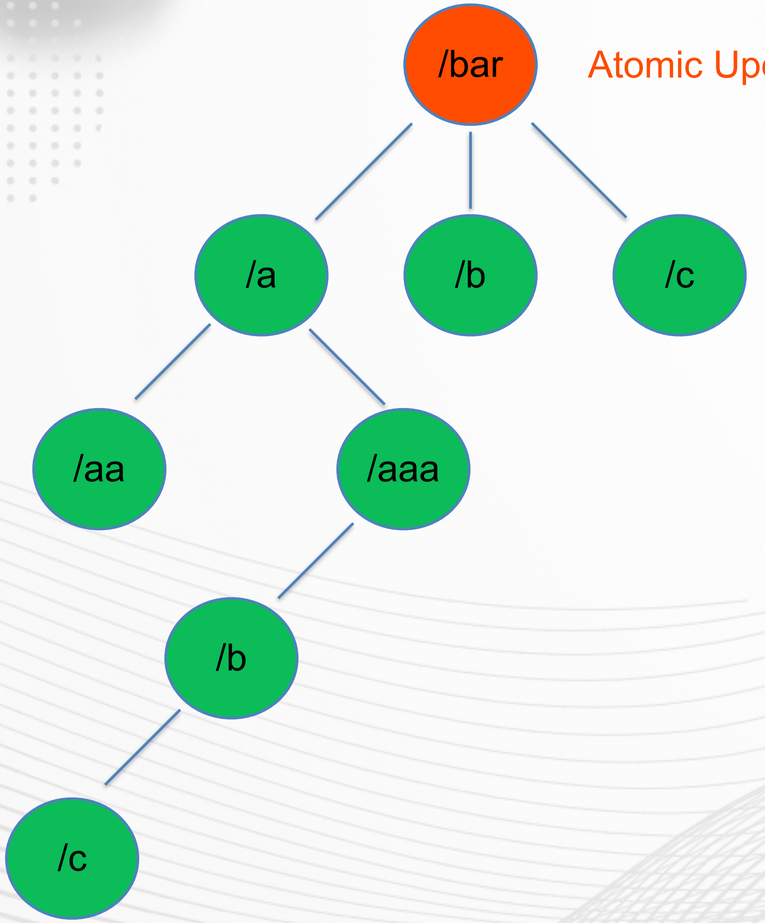

```
mv foo bar
```

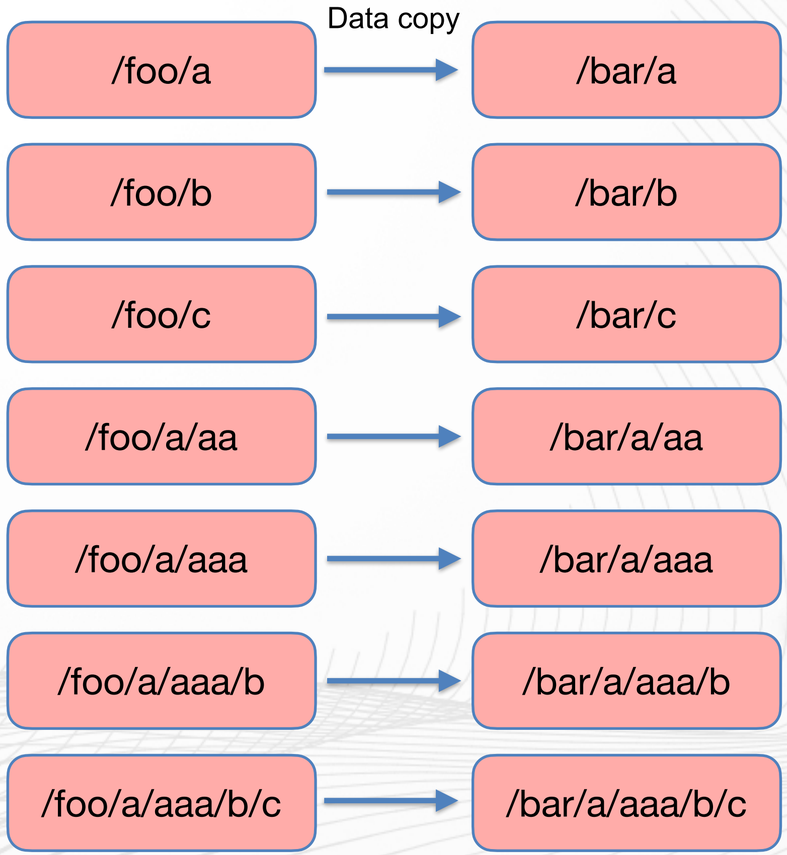


- /foo/a
- /foo/b
- /foo/c
- /foo/a/aa
- /foo/a/aaa
- /foo/a/aaa/b
- /foo/a/aaa/b/c

比较几种流行的分布式文件系统


```
mv foo bar
```

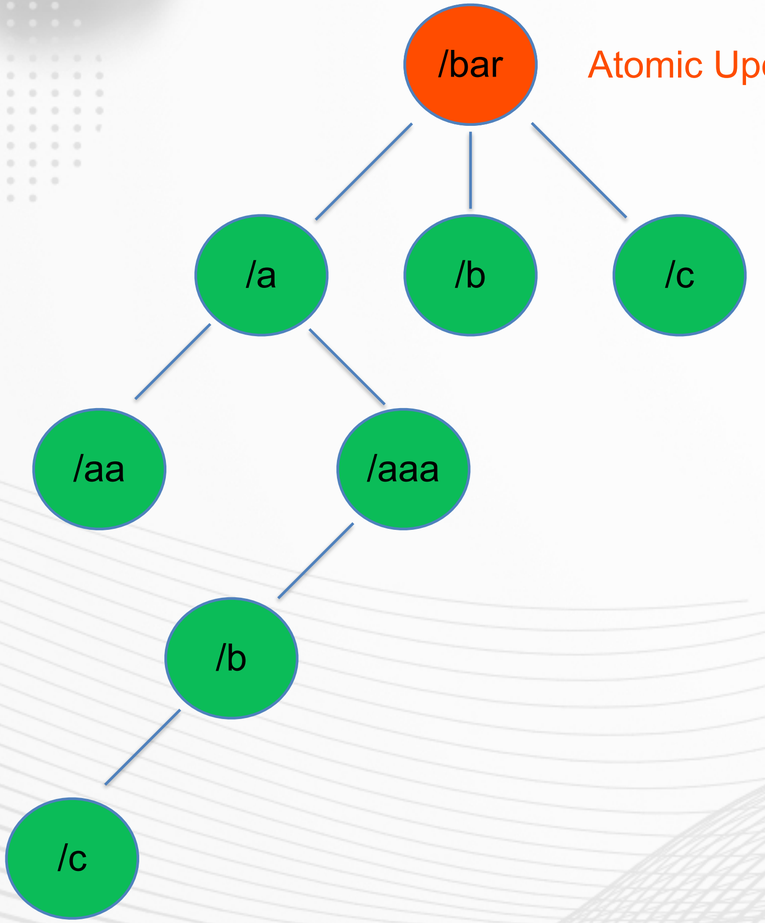


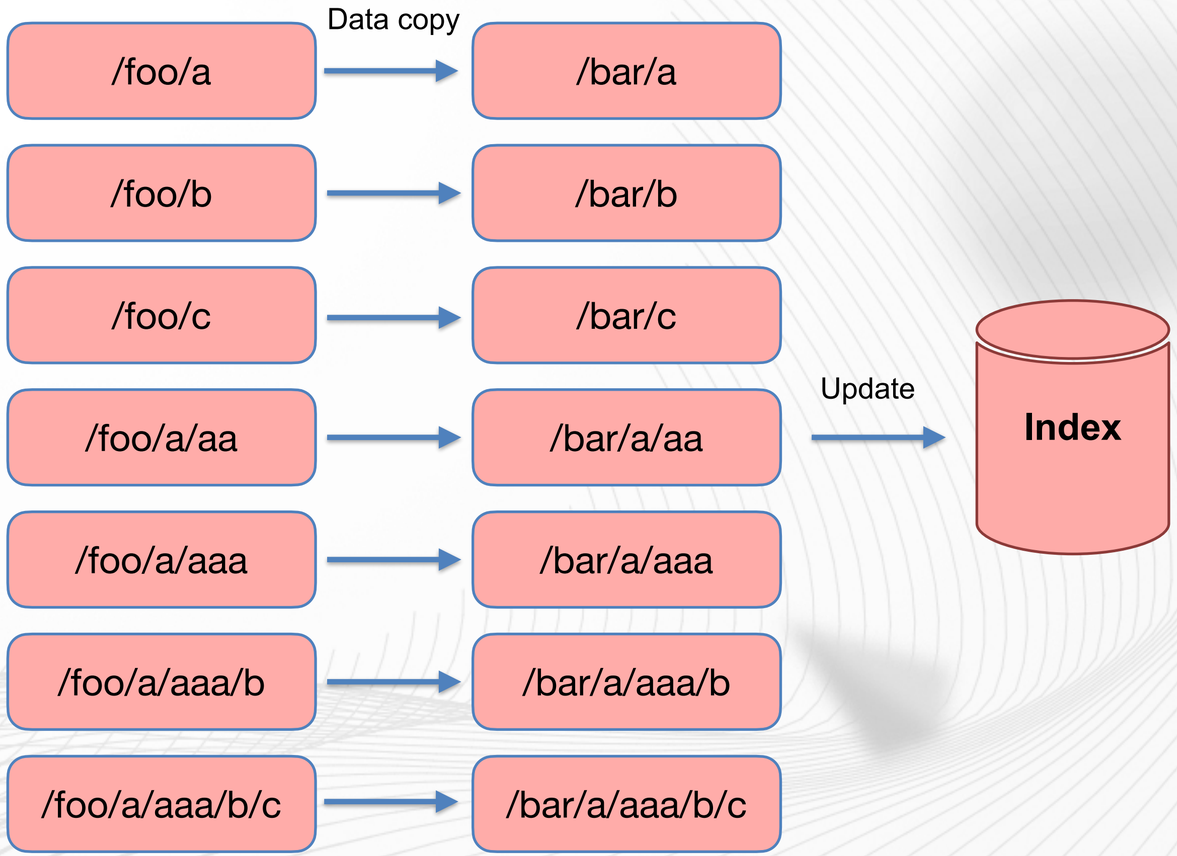


比较几种流行的分布式文件系统

github.com/juicedata/juicefs

```
mv foo bar
```

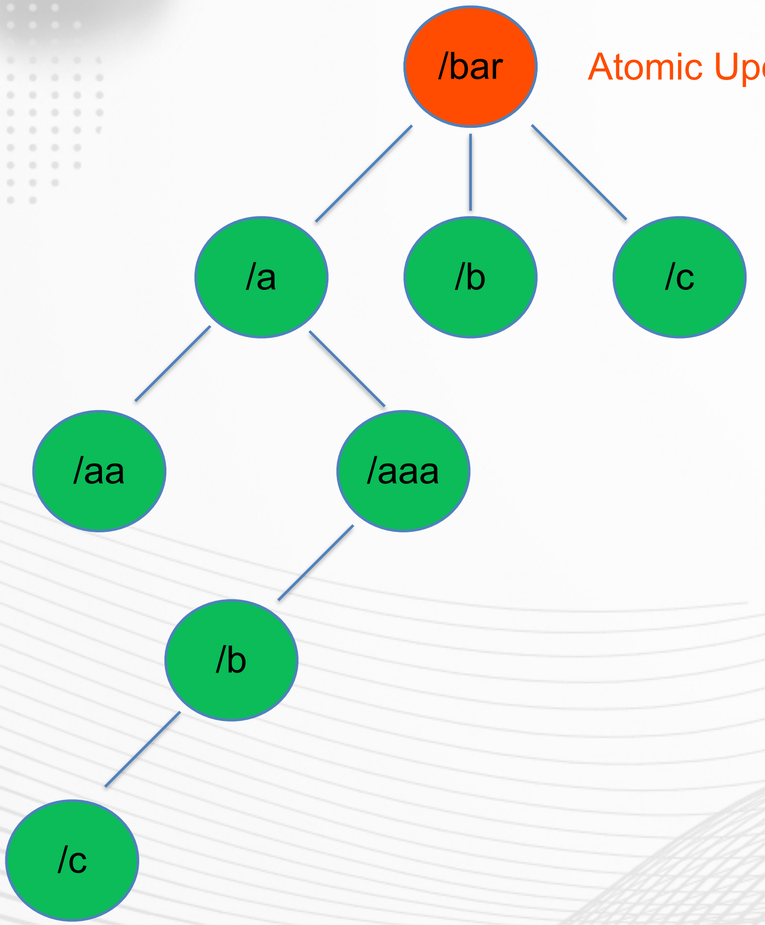




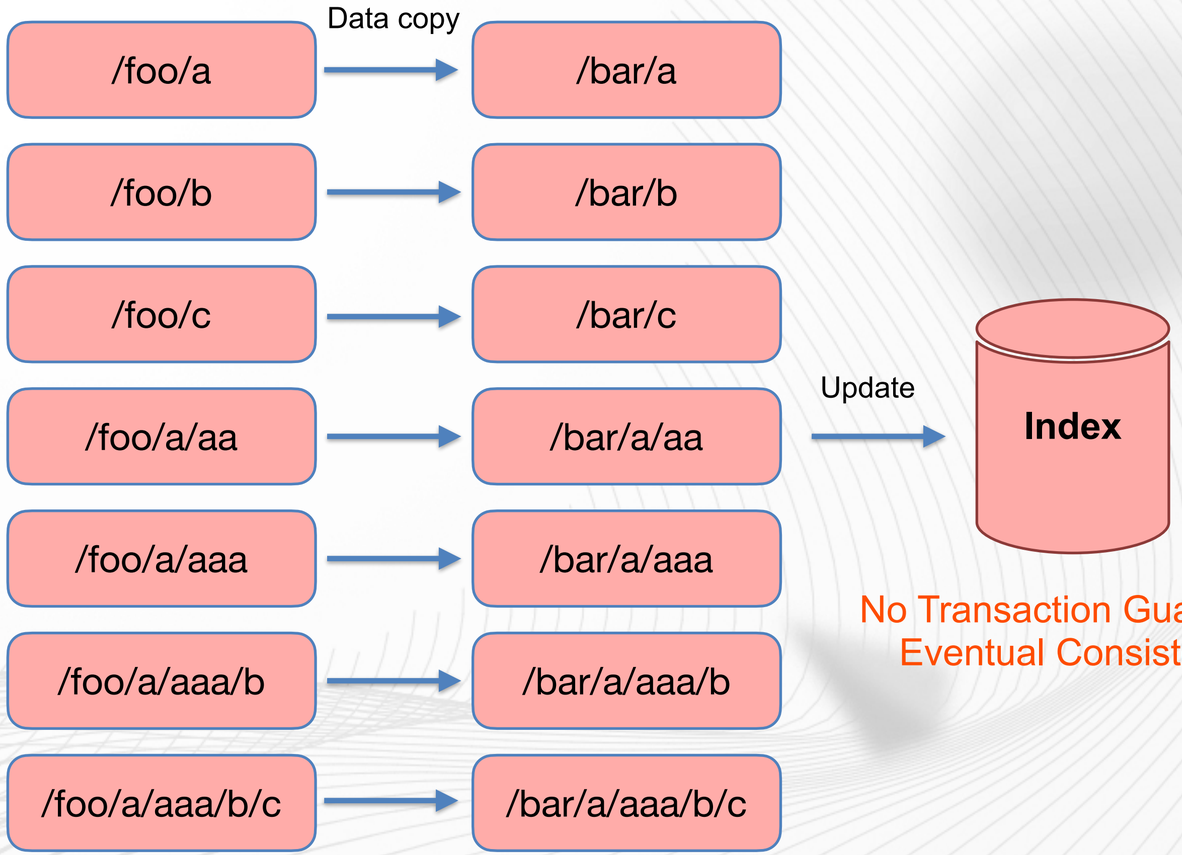
比较几种流行的分布式文件系统

github.com/juicedata/juicefs

```
mv foo bar
```



Atomic Update



No Transaction Guarantee
Eventual Consistency

比较几种流行的分布式文件系统

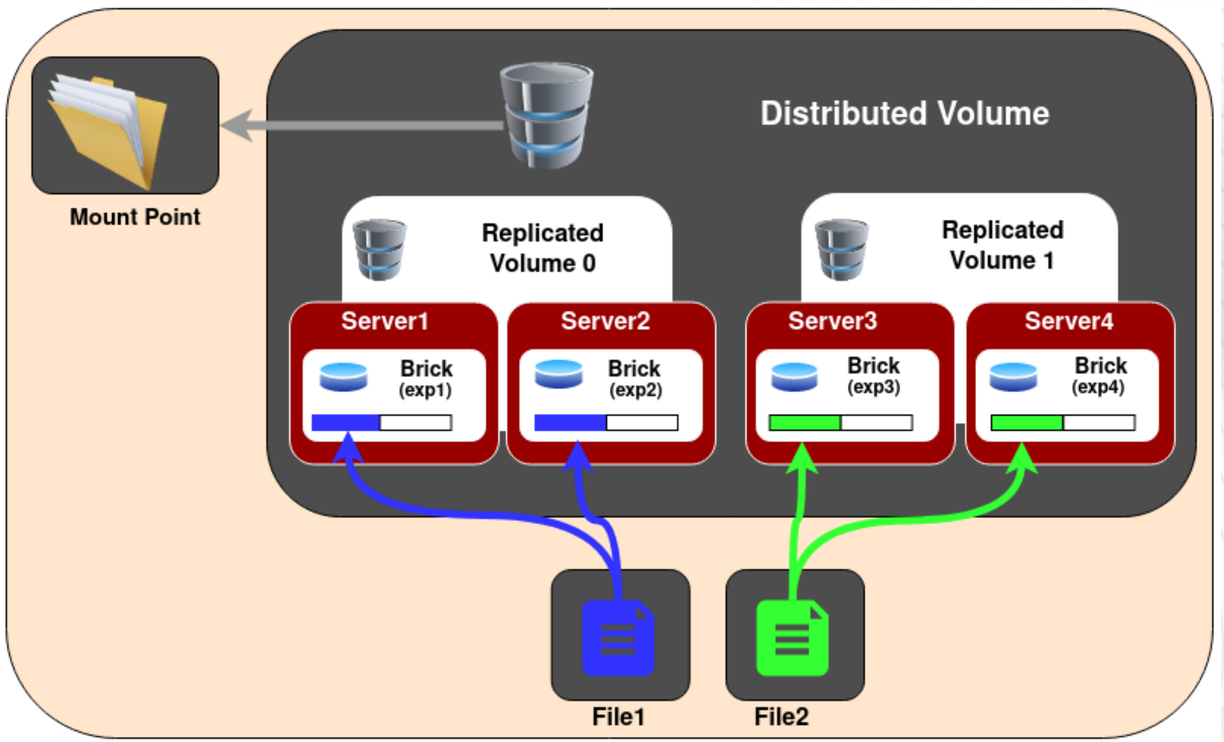
GlusterFS

优势

- 数据文件最终以相同目录保存在指定节点的本地文件系统上，集群不可用时仍能在该节点上直接访问到数据。

劣势

- 集群结构相对静态，不易调整；
- 因为没有独立元数据，元数据性能随集群规模扩大而下降。



https://docs.gluster.org/en/latest/Quick-Start-Guide/Architecture/

比较几种流行的分布式文件系统

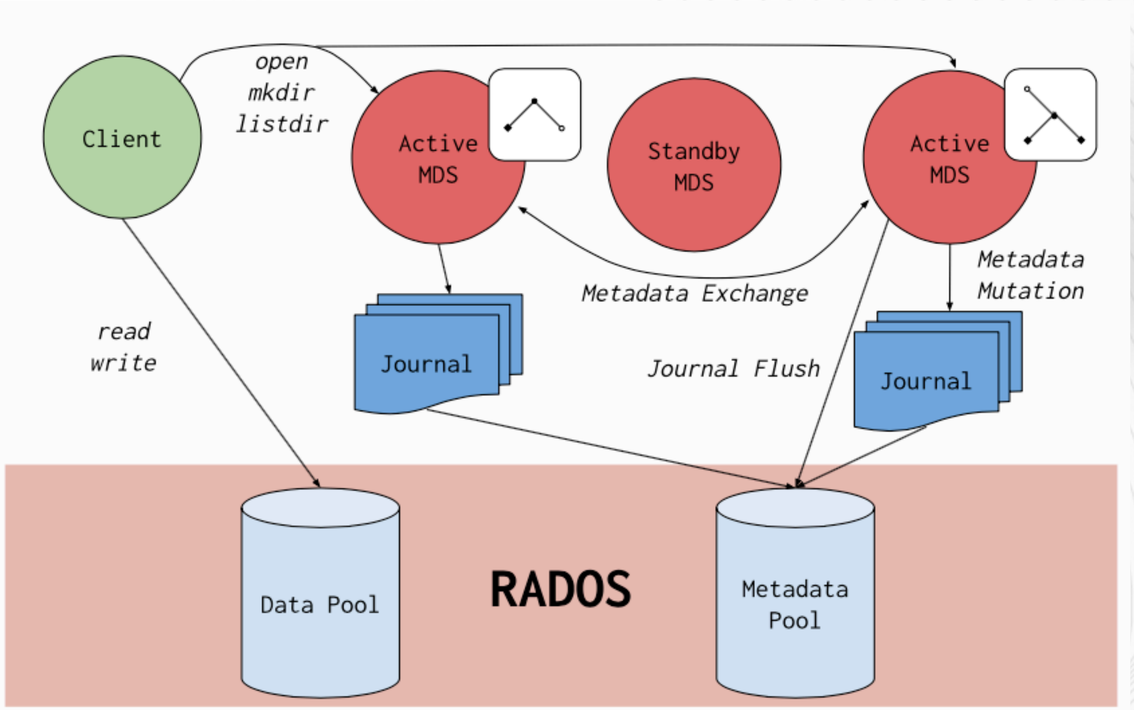
CephFS

优势

- 一套系统能同时满足 Block、Object、File 三种形态。

劣势

- 运维复杂；
- 元数据持久化在 RADOS 中，如果要高性能同样需要很大的内存缓存；
- MDS 内使用多线程提高吞吐，在文件操作的并发处理上复杂度大幅提升。



https://docs.ceph.com/en/latest/cephfs/index.html

比较几种流行的分布式文件系统

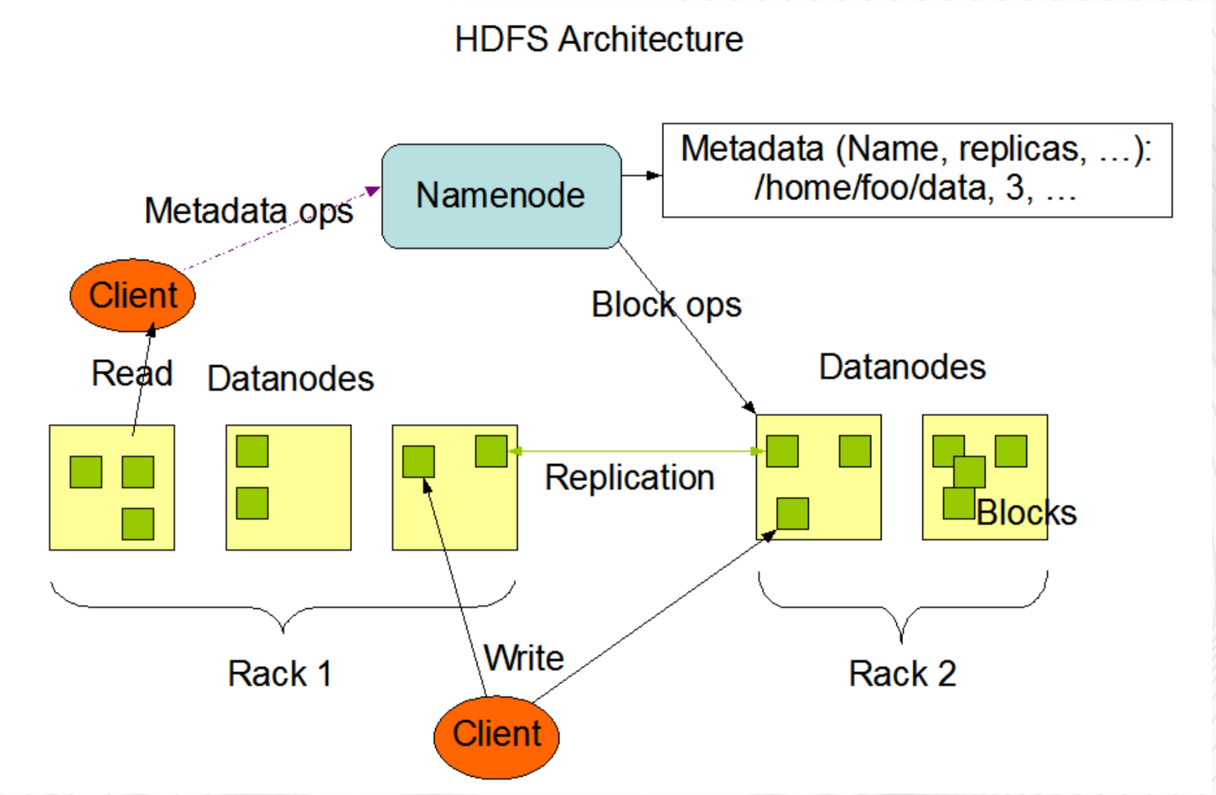
HDFS

优势

- Hadoop 生态标准组件；
- 有大量实践经验积累。

劣势

- NameNode 缺少横向扩展方案；
- 大集群的运维挑战，如：Full GC；
- 无 POSIX 访问协议，非 Java 客户端成熟度也不够。



https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

云原生环境下遇到的挑战

云原生环境下分布式文件系统的需要

- 支持 10亿 ~ 100亿文件管理；
- 支持 100PB ~ 1000PB 容量规模；
- 全托管服务，开箱即用；
- 弹性伸缩；
- 多租户；
- Kubernetes CSI 支持；
- 访问协议？POSIX，HDFS，S3，NFS，CIFS，Samba

云原生环境下遇到的挑战

云原生环境下分布式文件系统的需要

- 支持 10亿 ~ 100亿文件管理；
- 支持 100PB ~ 1000PB 容量规模；
- 全托管服务，开箱即用；
- 弹性伸缩；
- 多租户；
- Kubernetes CSI 支持；
- 访问协议？POSIX，HDFS，S3，NFS，CIFS，Samba

GlusterFS vs. CephFS vs. HDFS

- 都是面向物理机设计，无法弹性伸缩；
- 没有全托管服务；
- 元数据能力是文件系统的挑战；
- 单集群规模挑战以及多租支持改造难度大。

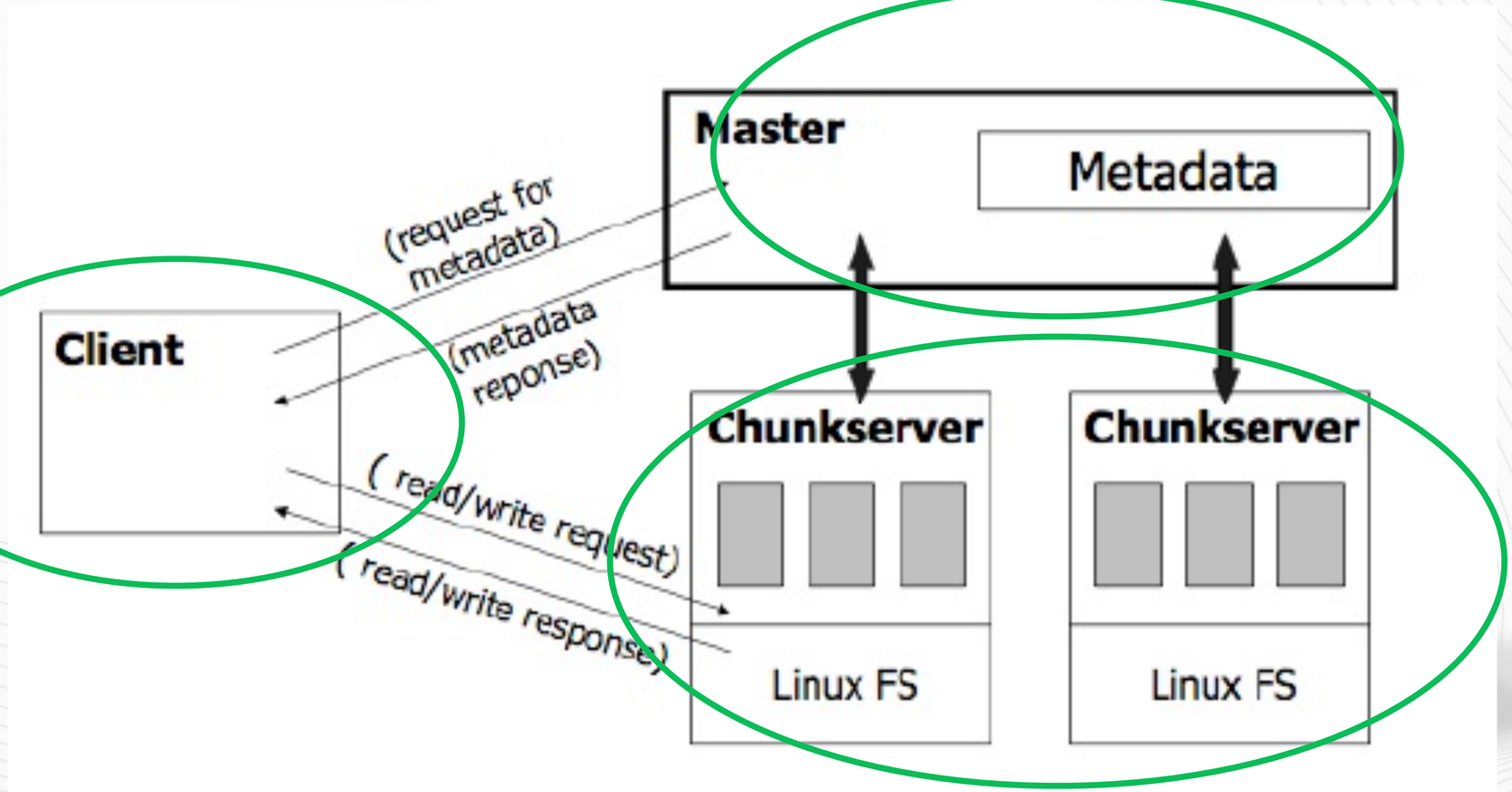
云原生环境下，如何设计分布式文件系统

文件存储三要素：Meta - Data - Client

Meta（NameNode）

Client
（Java SDK）

Data（DataNode）



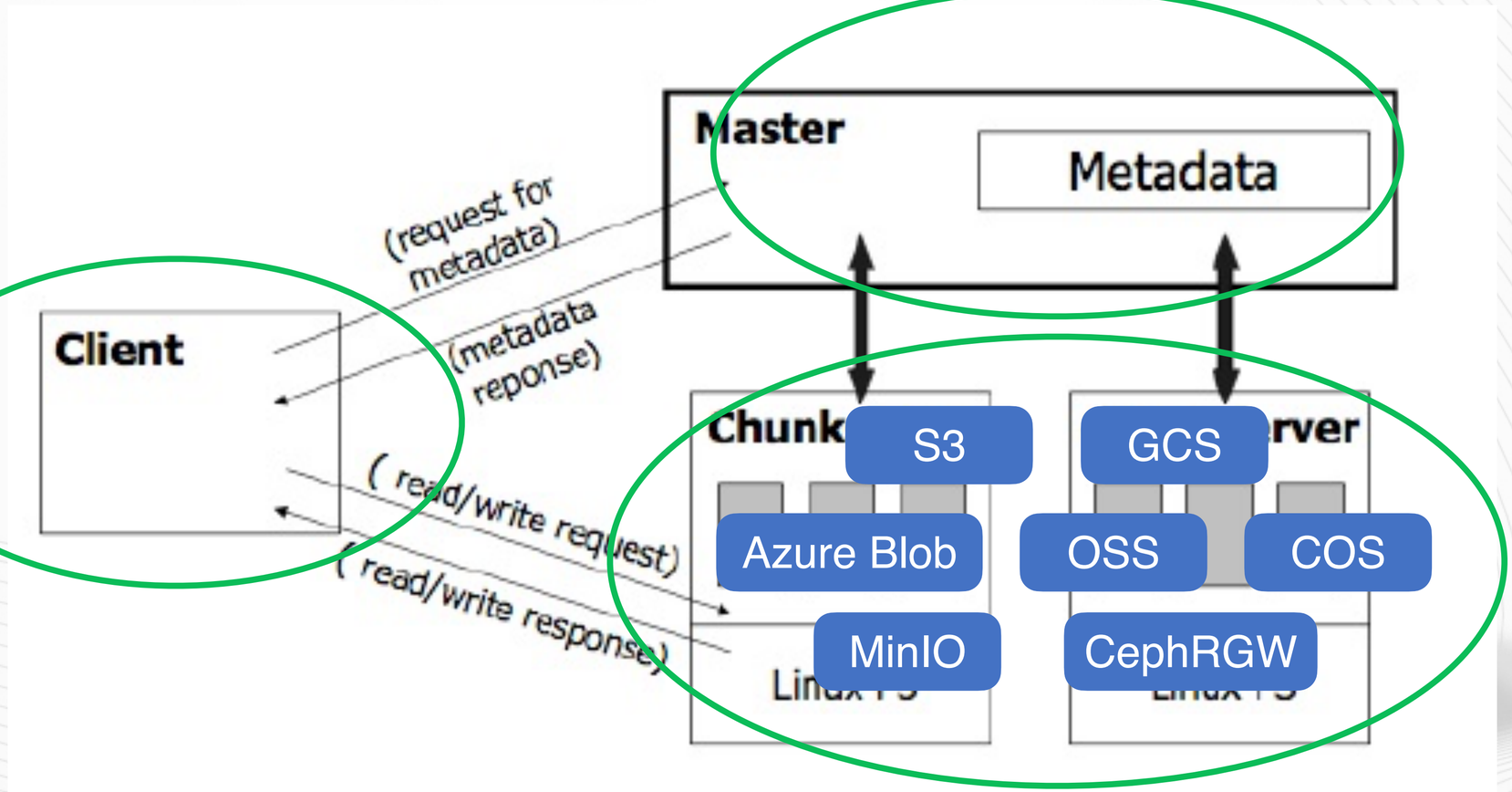
Google File System Architecture

云原生环境下，如何设计分布式文件系统

文件存储三要素：Meta - Data - Client

Meta（NameNode）

Client
（Java SDK）

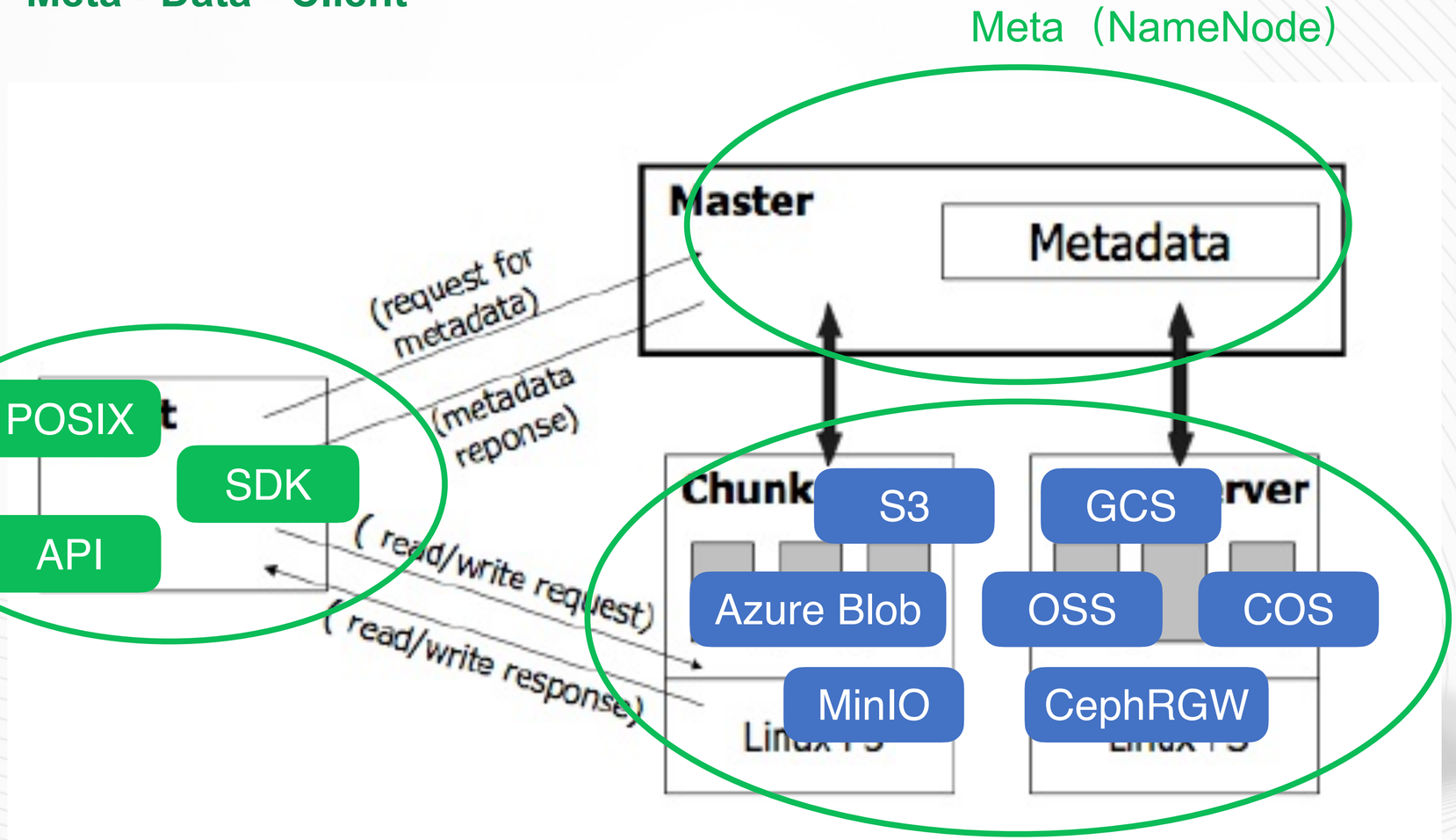


Data（DataNode）

Google File System Architecture

云原生环境下，如何设计分布式文件系统

文件存储三要素：Meta - Data - Client

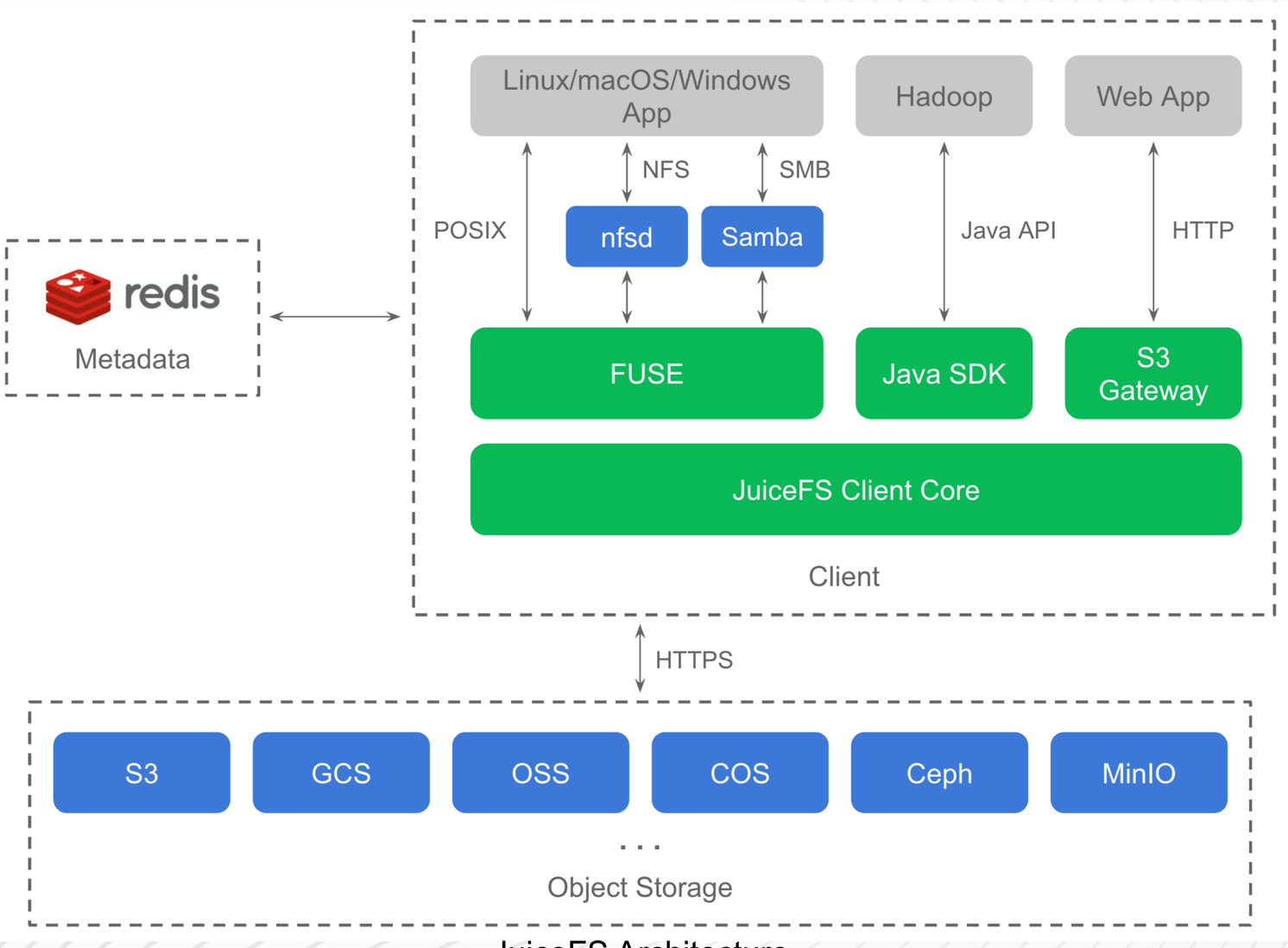


Google File System Architecture

云原生环境下，如何设计分布式文件系统

github.com/juicedata/juicefs

JuiceFS 基于 Redis 做元数据管理



JuiceFS Architecture

云原生环境下，如何设计分布式文件系统

github.com/juicedata/juicefs

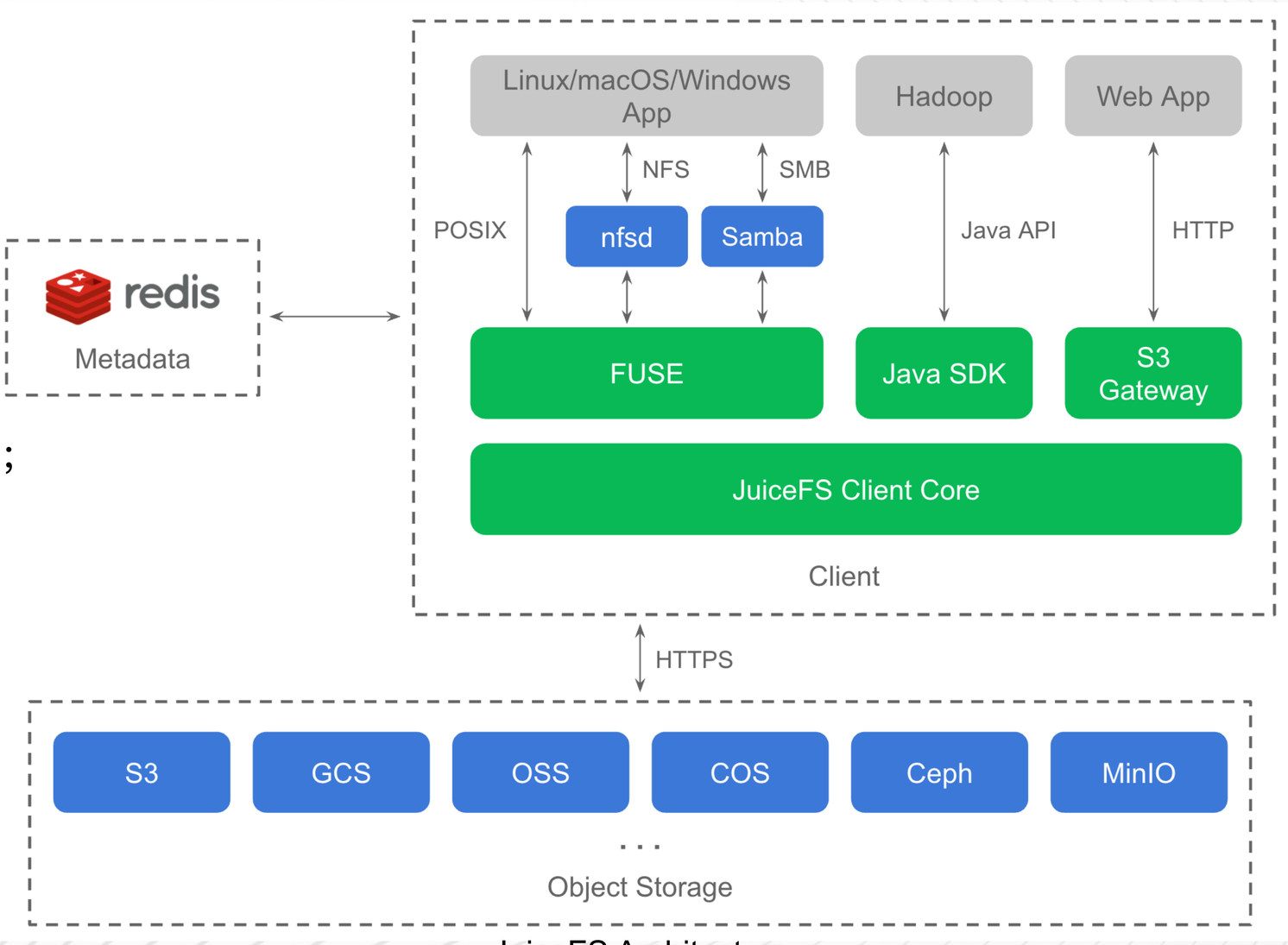
JuiceFS 基于 Redis 做元数据管理

优势

- Redis 很流行，大家对它很熟悉；
- 内存引擎，快；
- 支持事务。

劣势

- HA 方案不够完美，关注 RedisRaft 进展；
- 不能用 Redis Cluster，无法水平扩展；
- AOF 持久化仍有风险。

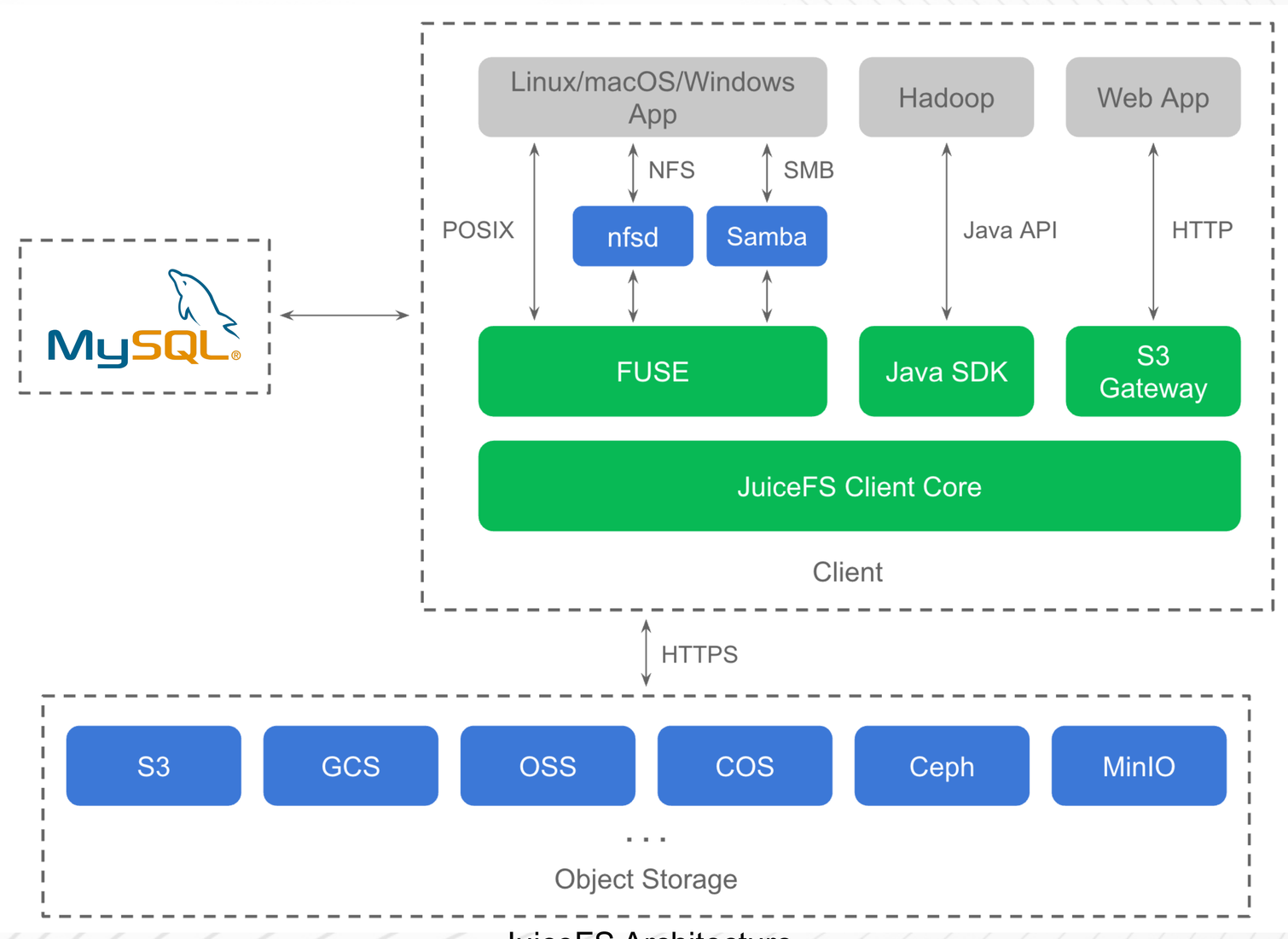


JuiceFS Architecture

云原生环境下，如何设计分布式文件系统

github.com/juicedata/juicefs

JuiceFS 基于 MySQL 做元数据管理



JuiceFS Architecture

云原生环境下，如何设计分布式文件系统

github.com/juicedata/juicefs

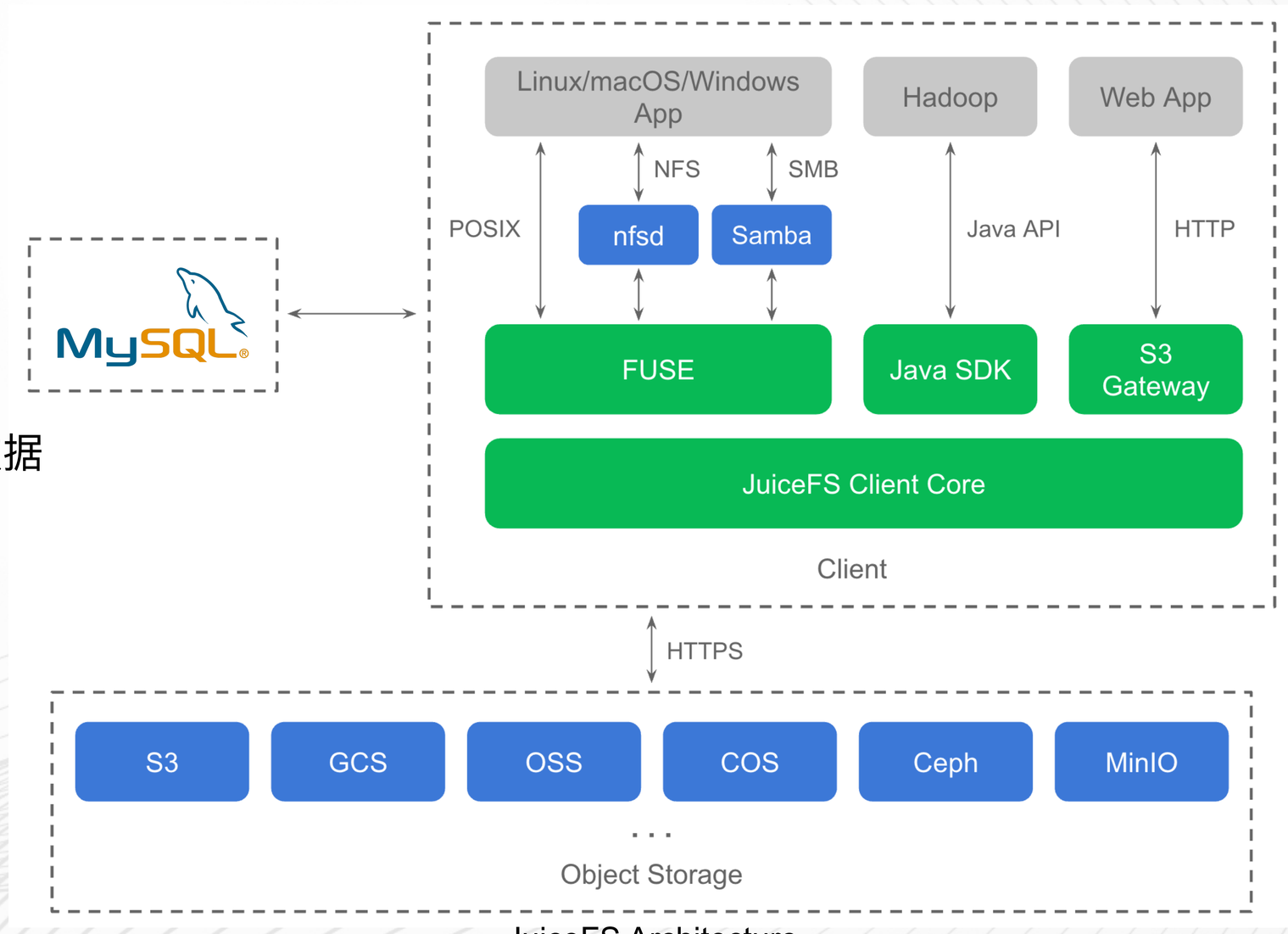
JuiceFS 基于 MySQL 做元数据管理

优势

- MySQL 同样流行，大家对它很熟悉；
- 持久化问题解决了；
- 规模瓶颈解决了。

劣势

- 相比内存引擎，时延高，不适合对元数据性能要求高的场景。

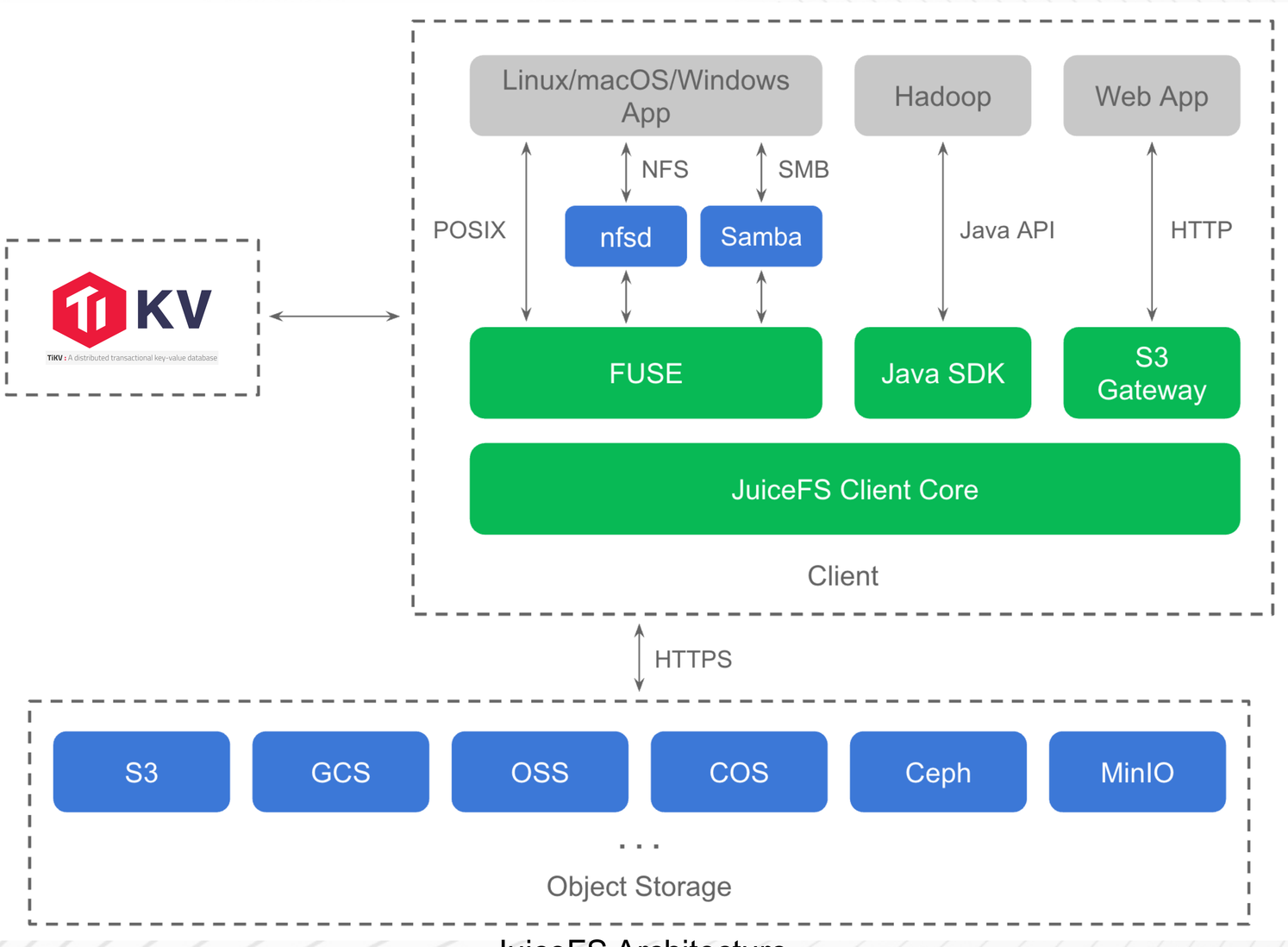


JuiceFS Architecture

云原生环境下，如何设计分布式文件系统

github.com/juicedata/juicefs

JuiceFS 基于 TiKV 做元数据管理



JuiceFS Architecture

云原生环境下，如何设计分布式文件系统

github.com/juicedata/juicefs

JuiceFS 基于 TiKV 做元数据管理

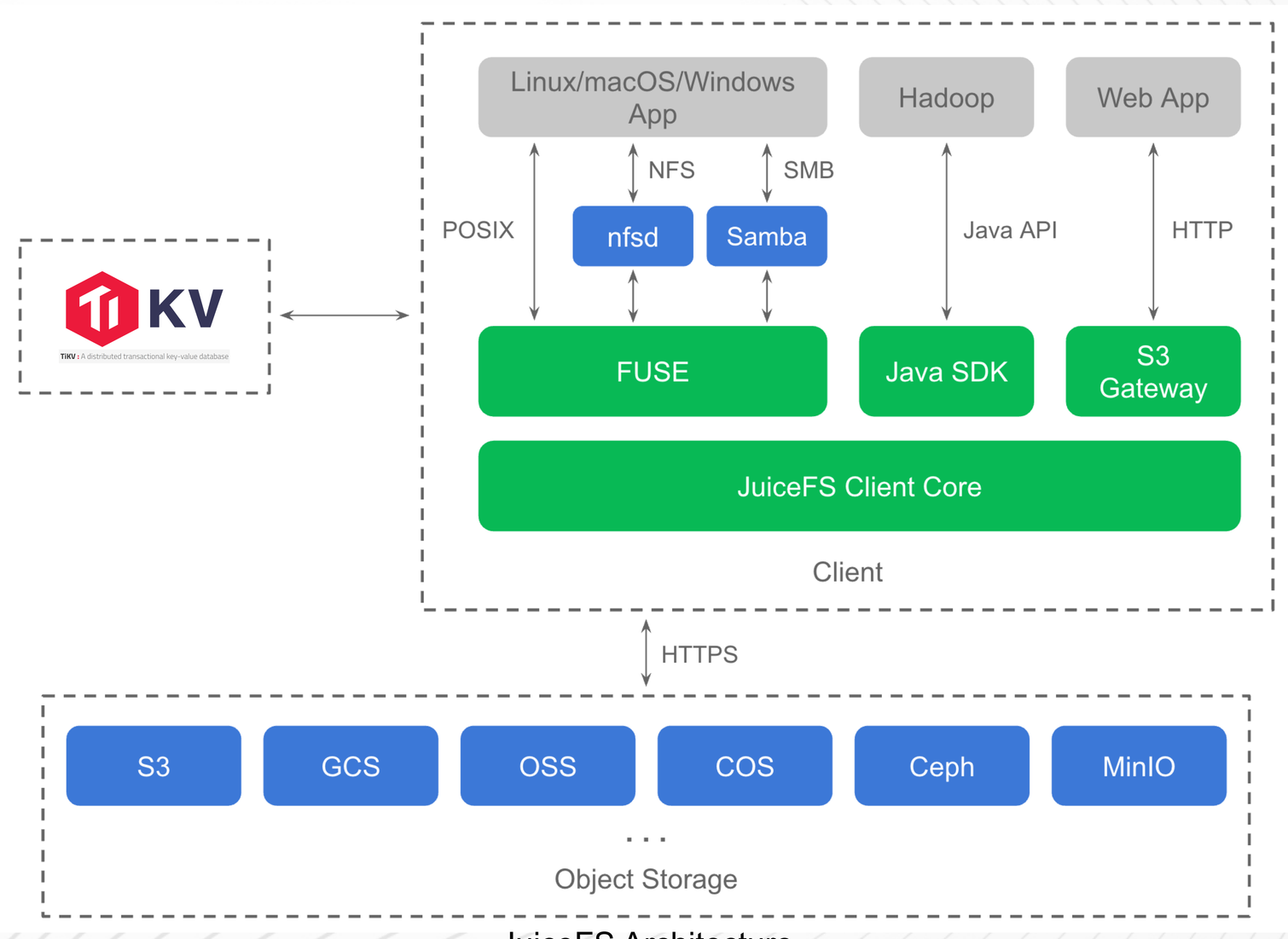
优势

- 持久化问题解决了；
- 规模瓶颈解决了；
- HA 问题解决了。

劣势

- 延迟略高于 Redis。

目前是早期版本，有优化空间，
欢迎大家参与测试、评估、改进
❤️



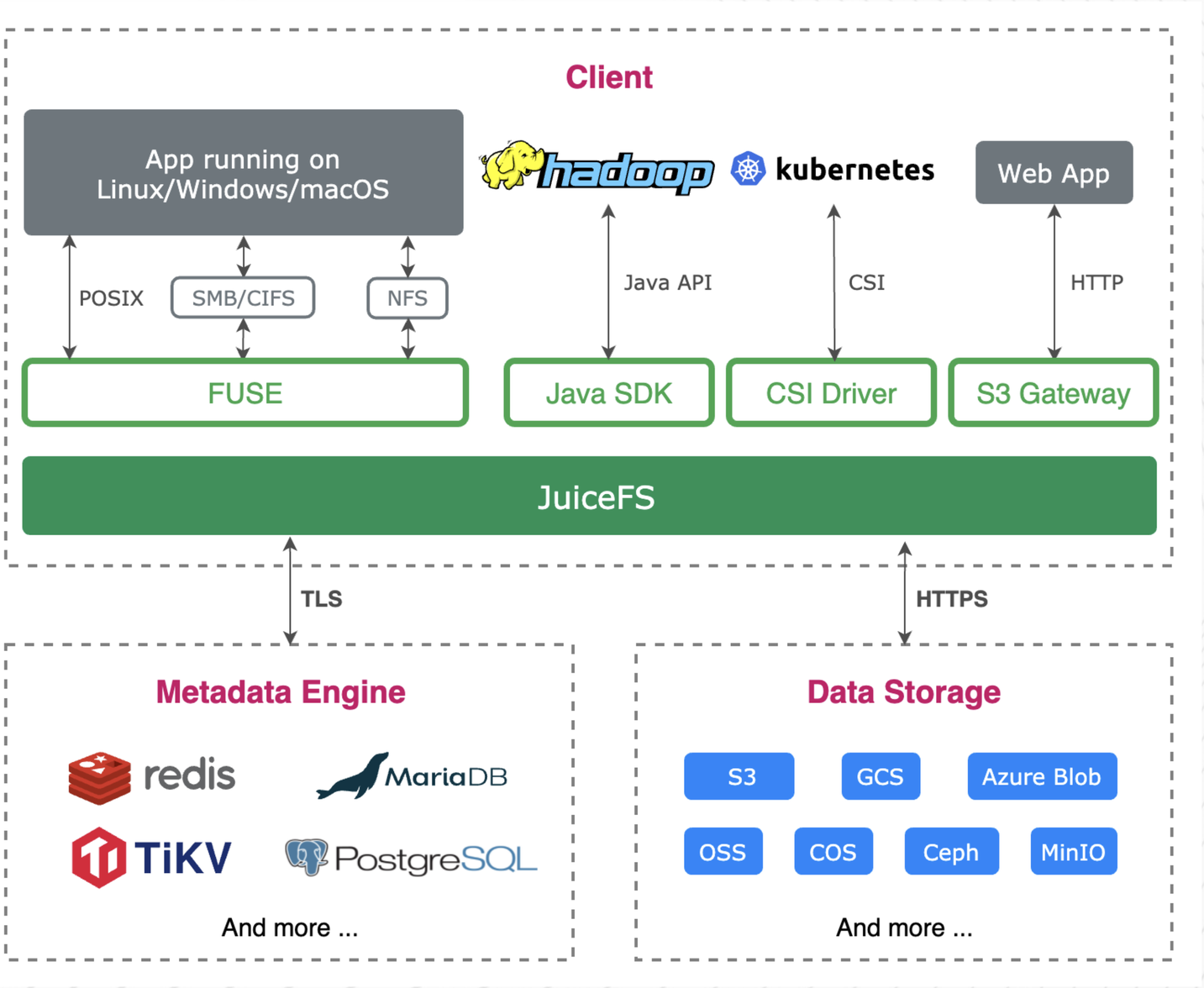
JuiceFS Architecture

云原生环境下，如何设计分布式文件系统

github.com/juicedata/juicefs

JuiceFS

- 多元数据引擎支持，可以在性能、规模、成本等方面取舍选择；
  - Redis
  - MySQL
  - MariaDB
  - PostgreSQL
  - SQLite
  - TiKV
  - …
- 多后端存储选择，轻松适配各种云环境；
  - 已支持 33 中，查看完整列表；
- POSIX、HDFS、S3 协议互通，方便对接各种上层应用。



业务场景中的收益 – 大数据

与 HDFS 相比

- 实现存储计算分离架构；
- 弹性伸缩；
- 元数据横向扩展能力；
- 支持 POSIX，方便和更多应用生态对接；
- 支持随机读写，能支持更丰富的需求。

业务场景中的收益 – 大数据

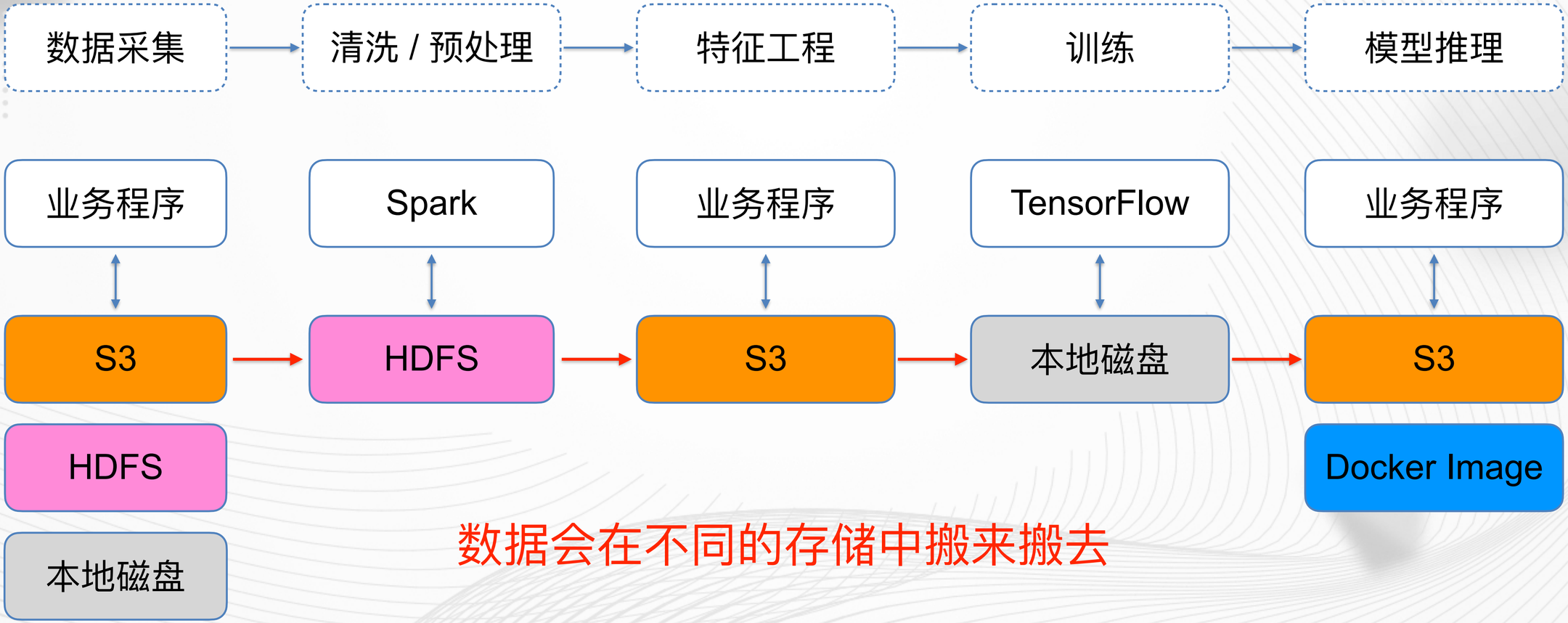
与 HDFS 相比

- 实现存储计算分离架构；
- 弹性伸缩；
- 元数据横向扩展能力；
- 支持 POSIX，方便和更多应用生态对接；
- 支持随机读写，能支持更丰富的需求。

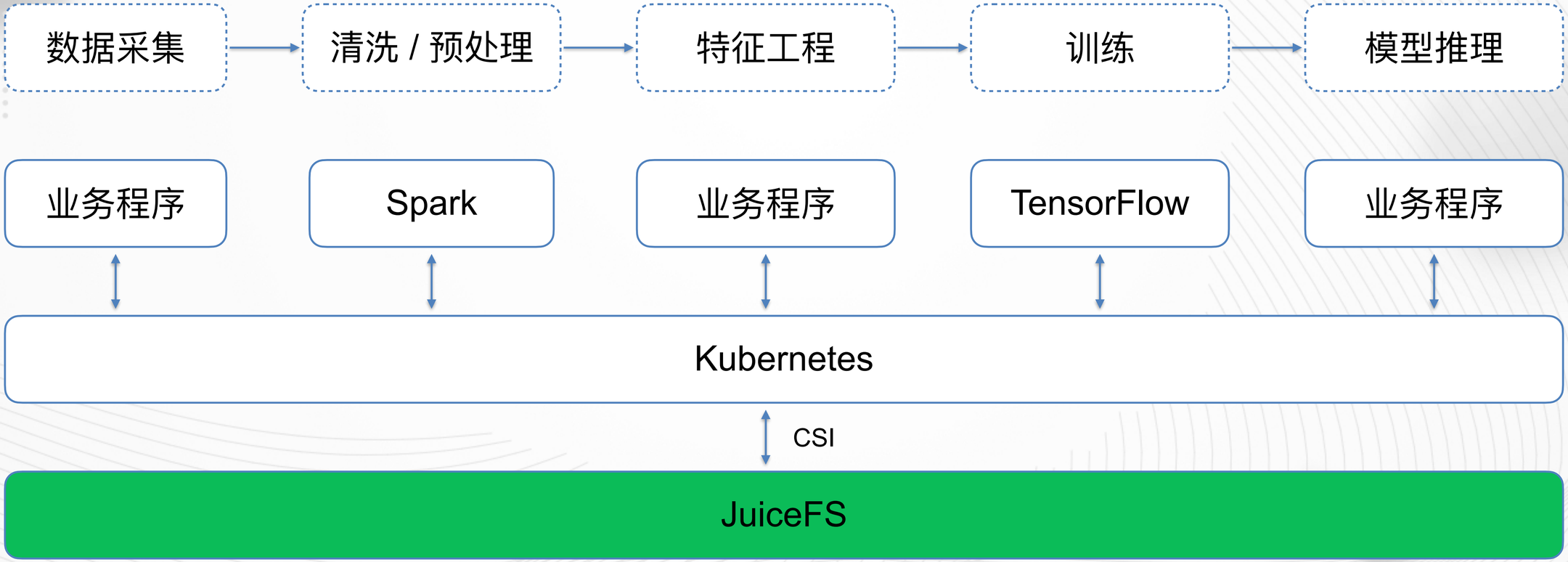
与 对象存储 相比

- 100% HDFS 兼容；
- 强一致性；
- 高性能 Listing；
- 原子 Rename；
- 在不同云环境中无差异；
- 对人来说，文件系统相比 Bucket 更容易做数据管理

业务场景中的收益 – AI

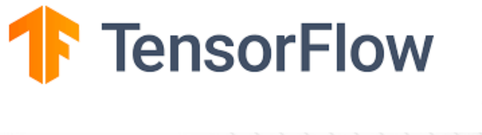

数据采集
清洗 / 预处理
特征工程
训练
模型推理
业务程序
Spark
业务程序
TensorFlow
业务程序
S3
HDFS
S3
本地磁盘
S3
HDFS
Docker Image
本地磁盘
数据会在不同的存储中搬来搬去

业务场景中的收益 – AI


数据采集
清洗 / 预处理
特征工程
训练
模型推理
业务程序
Spark
业务程序
TensorFlow
业务程序
Kubernetes
CSI
JuiceFS

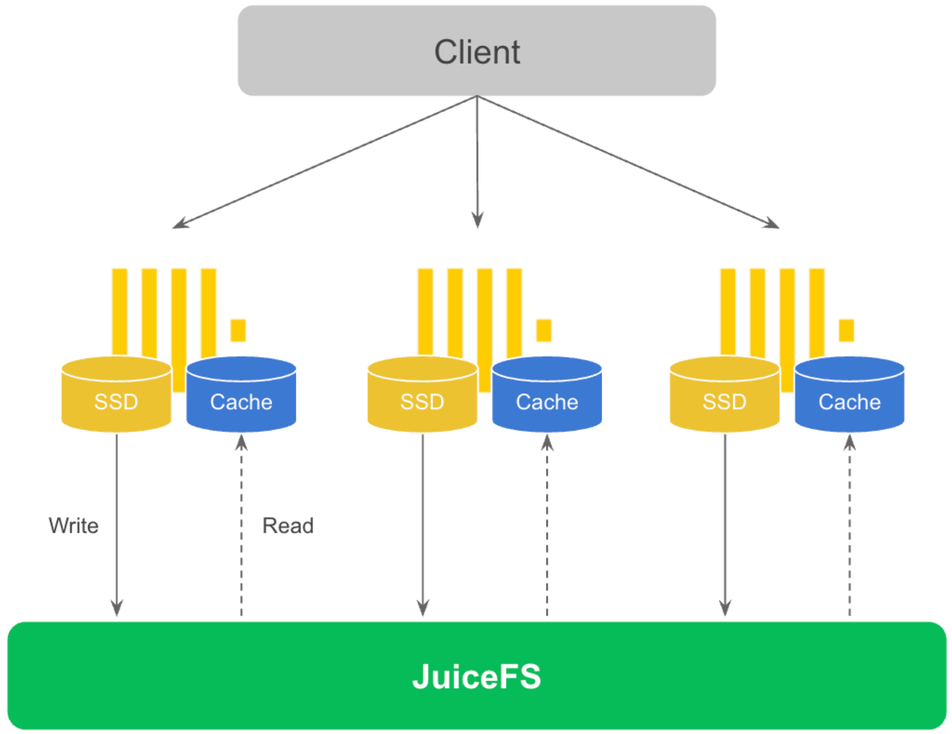
未来展望

github.com/juicedata/juicefs

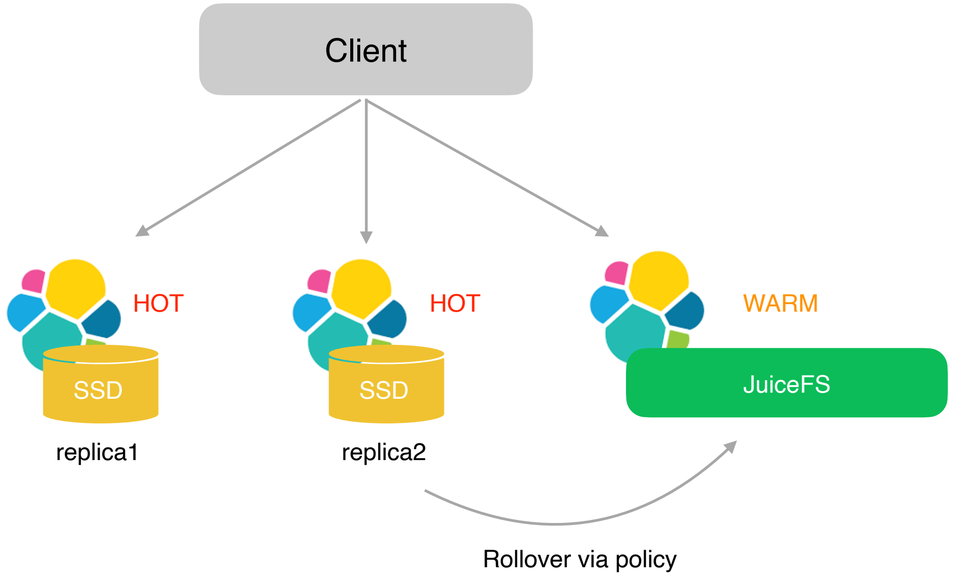


未来展望 – 存算分离带来的可能性

ClickHouse



elastic



THANKS

github.com/juicedata/juicefs

扫一扫上面的二维码图案，加我微信