

OSS: Apache Doris 存算分离的数据基座

周翱

阿里云 技术专家

目录

1. OSS 核心特点和关键技术
2. Doris 冷热分层和存算分离
3. OSS 对存算分离的应对办法

1 OSS 核心特点和关键技术

OSS 天然优势

存储成本低

极低的存储成本



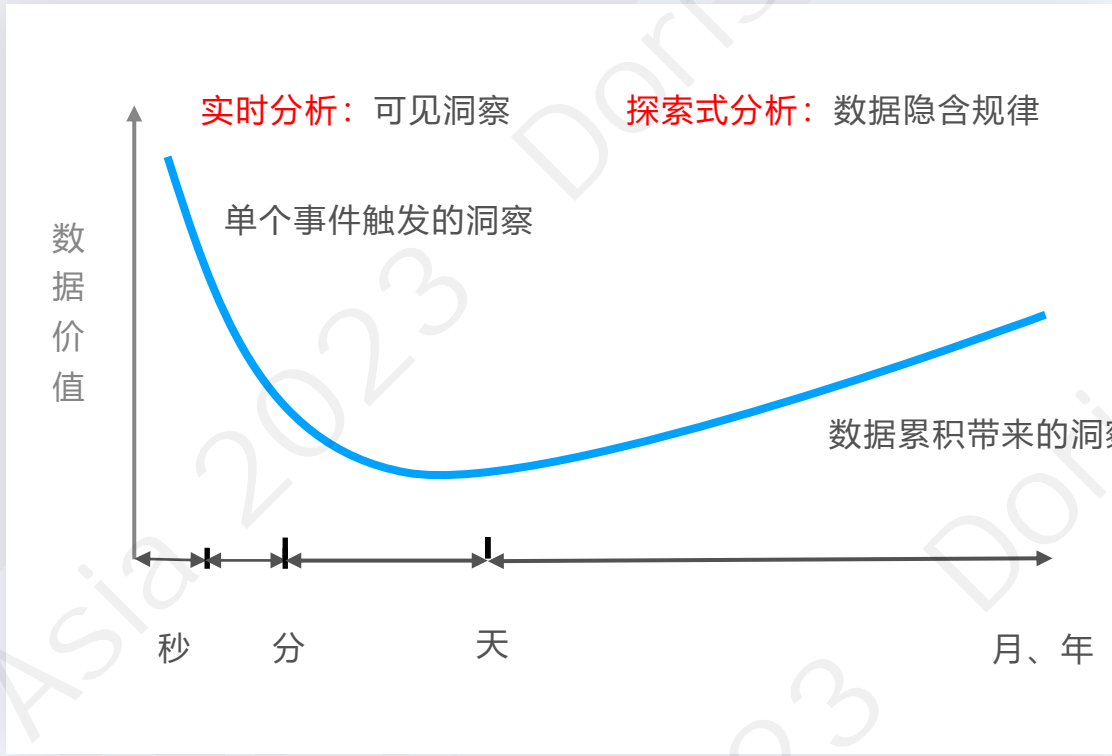
支持半结构和非结构化数据

能够存储，网站，媒体，视频，音频
等各种数据



弹性能力强

性能容量无限扩展



基于公共协议

通过http协议访问，使用简洁方便，
便于理解和调试

OSS 关键特性



多级存储

通过生命周期管理来优化成本



安全可靠

需要企业级高可用，可用性 99.995%
数据不丢不错，持久性12个9



数据质量

确保数据有效性



存算分离

资源灵活扩展



一源多用

支撑多业务对数据同时查询分析



灵活分析

同时支持多种计算引擎

OSS: 助力云上数据湖仓构建

计算层

(弹性计算引擎)

数据湖计算



优化层

(湖构建与加速)

数据湖构建与优化



DLF数据湖构建

DELTA LAKE

Apache Hudi

ICEBERG

开放数据湖格式

分布式缓存加速

存储层

(数据湖统一存储)

数据湖存储

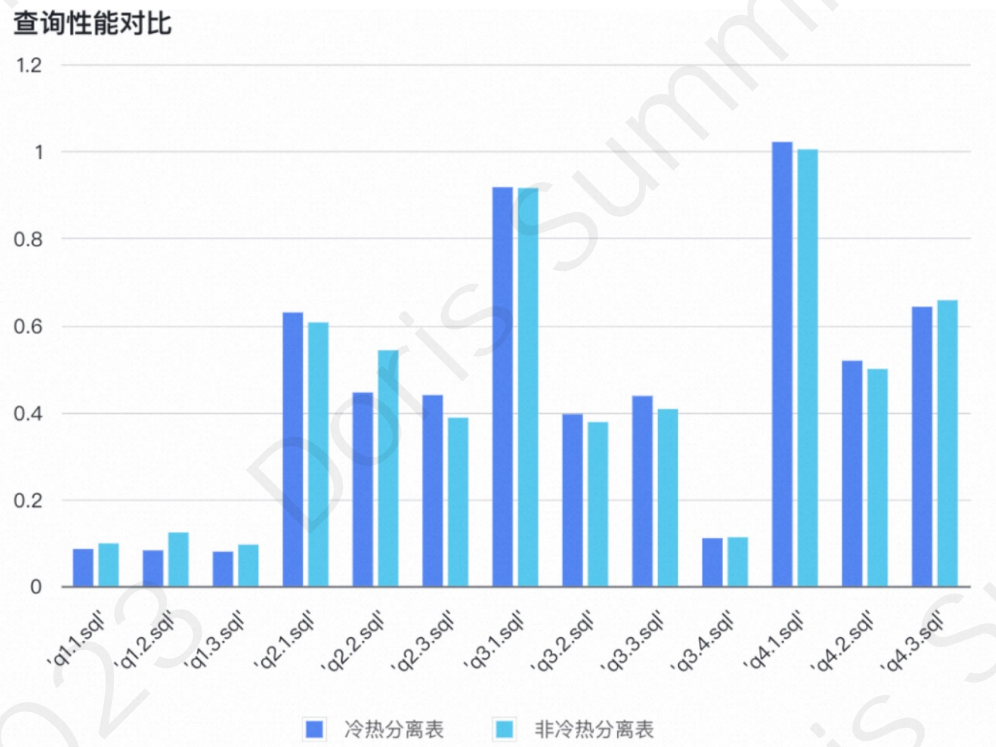
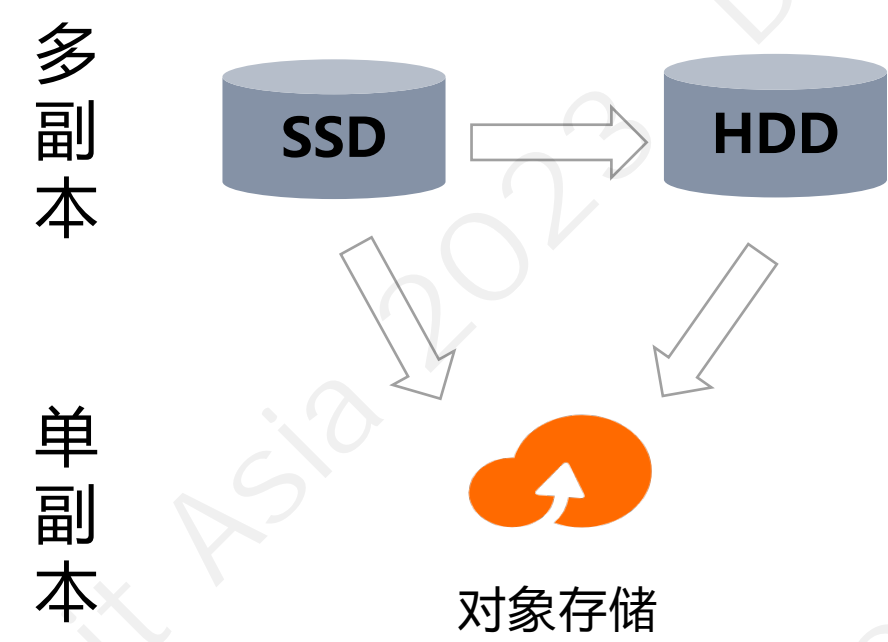


对象存储

- 标准型
- 低频型
- 归档型
- 冷归档

2 Doris 冷热分层和存算分离

Doris 冷热分层，极致成本



三级存储，更高存储效率

Apache Doris 2.0 版本中支持三级存储，分别是 SSD、HDD 和对象存储。冷数据多副本变为单副本，存储成本进一步降至原先的三分之一

支持对象存储，超低存储成本

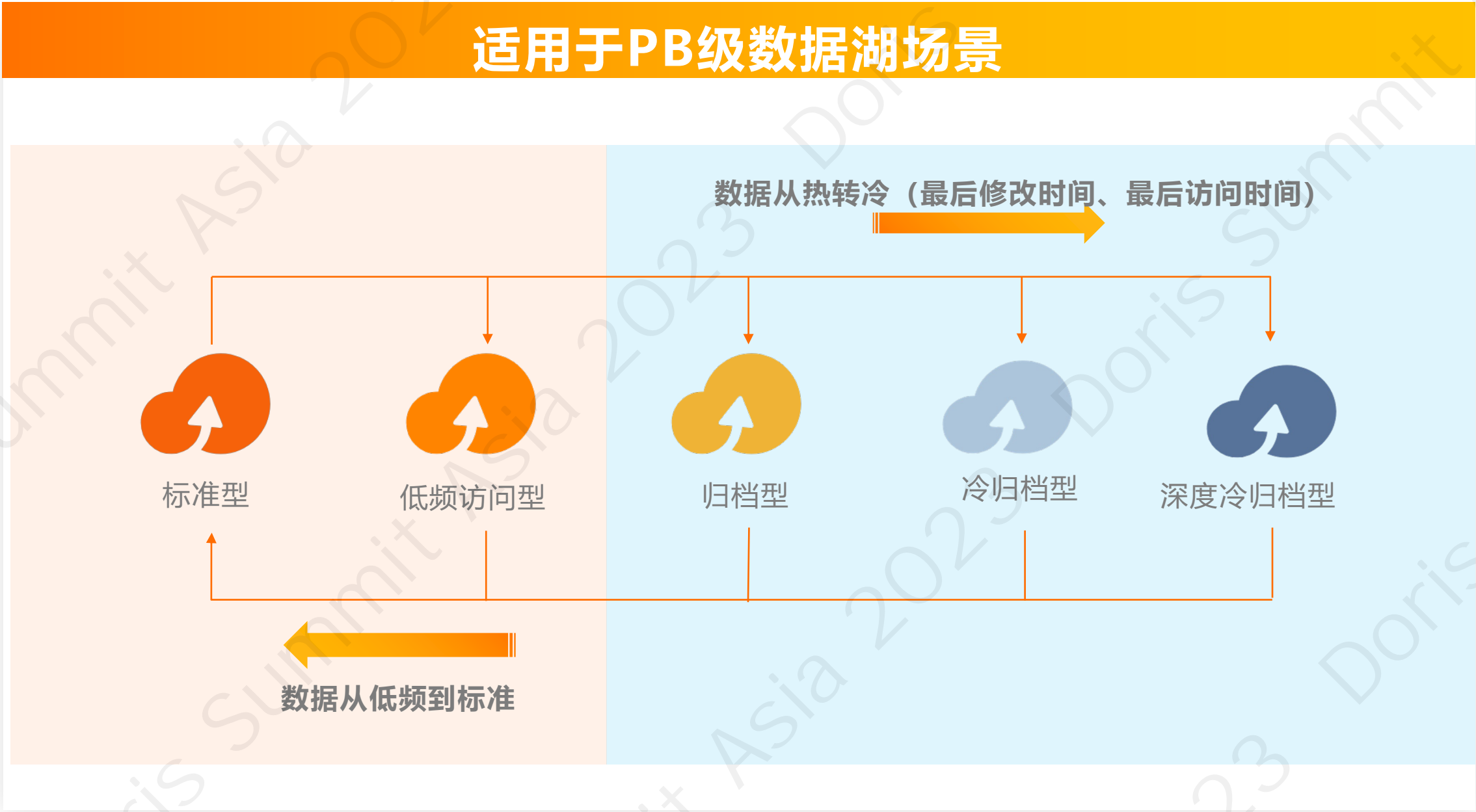
支持多种存储类型，无限容量。如将 80% 的冷数据保存到对象存储中，存储成本至少可降低 70%。

冷存 Cache，高效访问

通过冷数据 Cache 技术，可以将冷数据缓存在本地磁盘中，提高数据读取速度，从而提高查询效率

Doris冷热分层技术结合OSS分层存储

目前 Doris 冷热分层支持标准型存储



五种不同形态存储

满足用户不同热度存储的性价比要求，其中深度冷归档价格仅为0.0075元/GB/月

智能分层能力

根据访问情况自动使用不同形态存储分层存放数据，成本优化最高可达90%

归档直读

在对象存储OSS中，直接访问归档存储类型的文件，而无需先对其解冻

Doris存算分离，极致弹性

拥抱云计算，极致弹性

拥抱云计算

Apache Doris 针对云计算这种新型基础设施提供更加深度的适配。

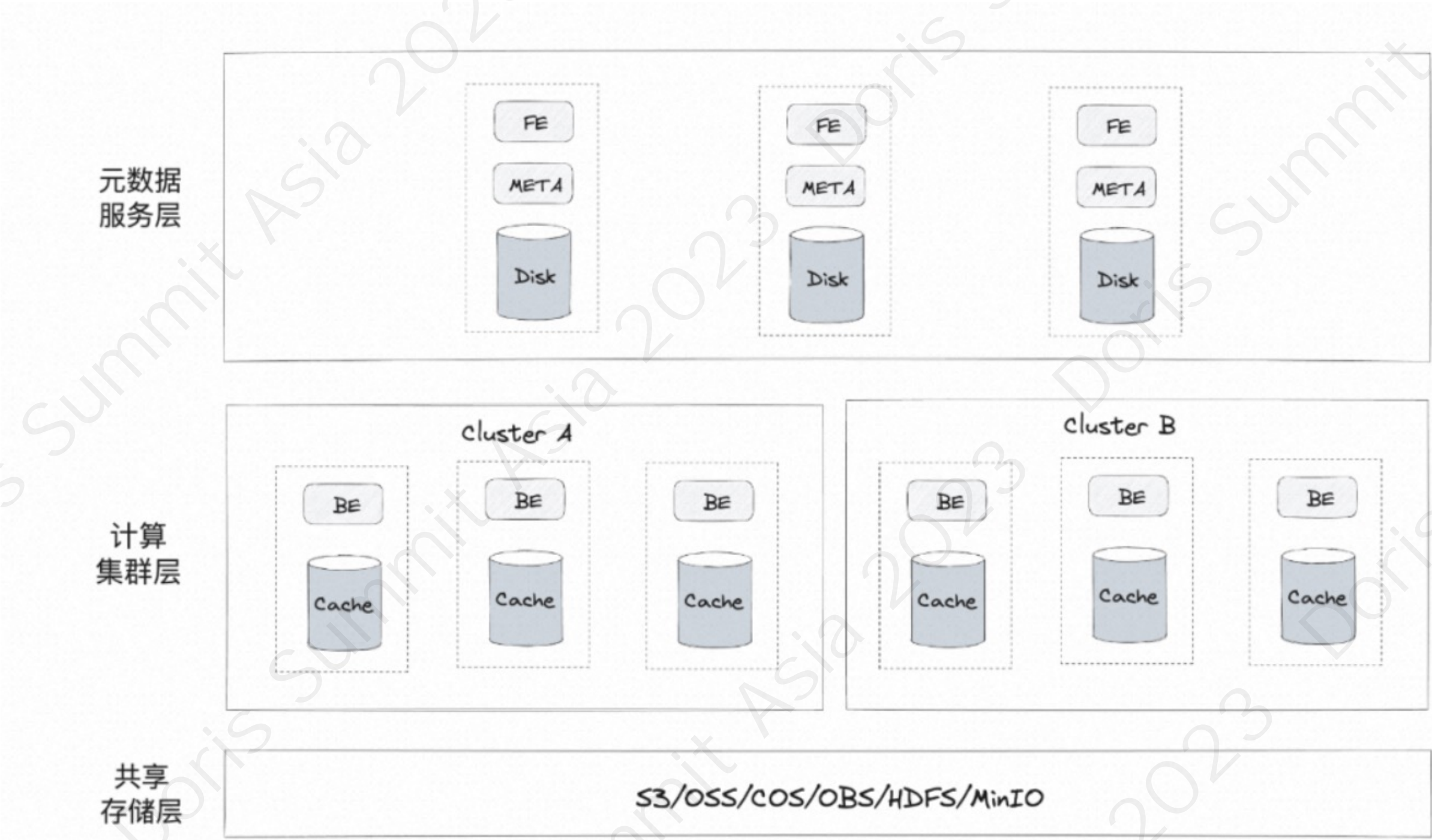
极致弹性,带来极大的成本经济优势

- 1. 计算节点：无状态，可以非常方便的横向进行扩展和释放。
- 2. 共享存储：极低的存储成本和极高的数据可靠性，并且大大简化上层计算节点的实现复杂度。

基于本地的高速缓存

存算分离依赖从网络上读取存储系统的数据来进行计算，在一定程度上会造成计算性能的下降，这也是相较于存算一体架构的主要劣势。

为了解决这一问题，可以在本地利用 SSD 提供高速缓存。



常见的云上存算分离几个痛点



数据下载变慢

存算分离后，一旦没有命中本地缓存中的热数据，下载速度就会慢很多

带宽不够

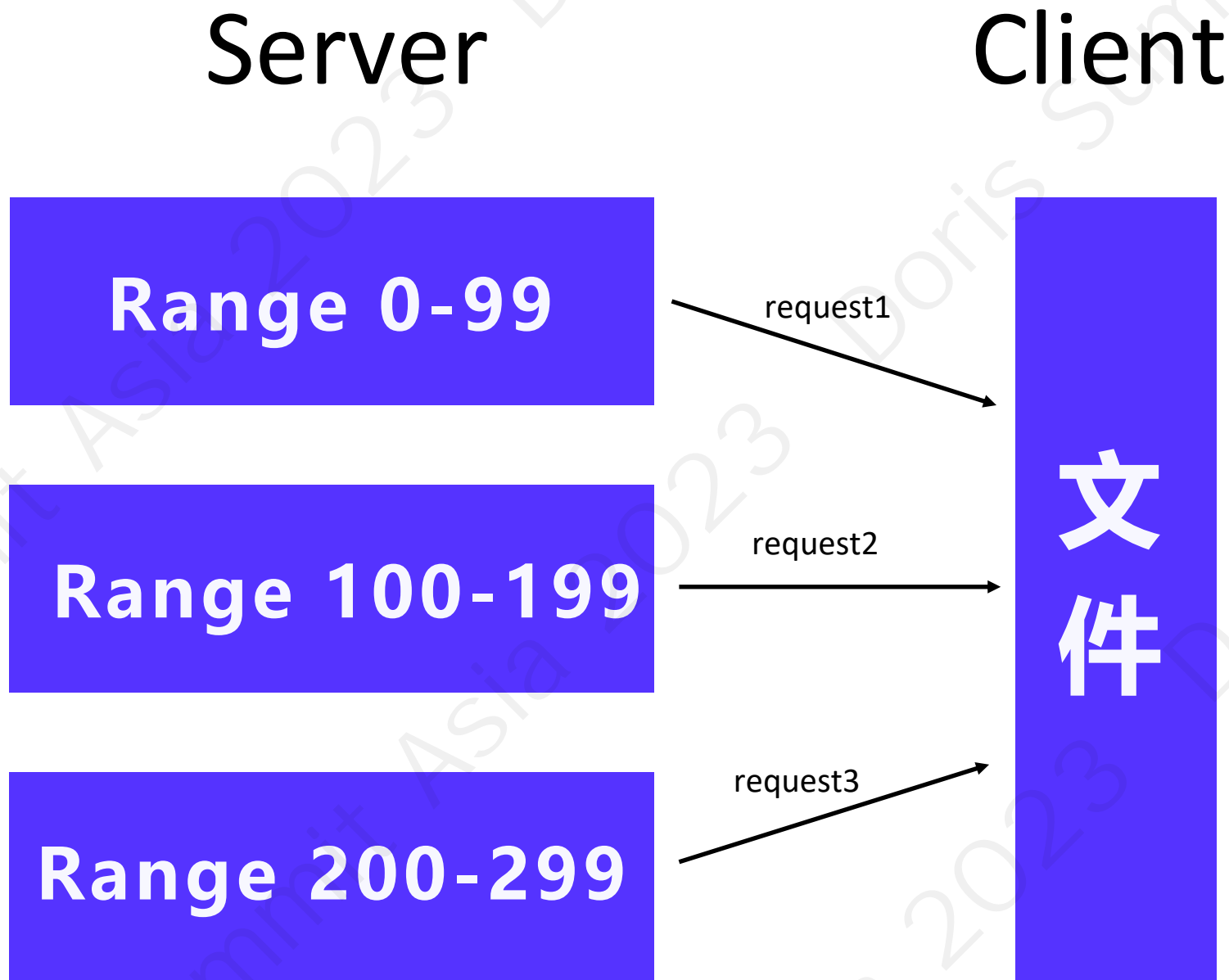
需要大量扫描数据，读写带宽成为瓶颈

成本更低的归档类型无法使用

在降级为归档后，虽然成本更低，但是需要解冻才能使用

3 OSS 对存算分离的应对办法

并发下载能力-Range下载



```
GET /oss.jpg HTTP/1.1
Host:oss-example.oss-cn-hangzhou.aliyuncs.com
Date: Fri, 28 Feb 2012 05:38:42 GMT
Range: bytes=100-900
Authorization: *****
```

```
HTTP/1.1 206 Partial
Content x-oss-request-id: 28f6-15ea-8224-234e-c0ce407*****
x-oss-object-type: Normal
Date: Fri, 28 Feb 2012 05:38:42 GMT
Last-Modified: Fri, 24 Feb 2012 06:07:48 GMT
ETag: "5B3C1A2E05E1B002CC607C*****"
Accept-Ranges: bytes
Content-Range: bytes 100-900/344606
Content-Type: image/jpg
Content-Length: 801
Server: AliyunOSS [801 bytes of object data]
```

客户端工具数据加速

OSSFS

简单易用，以文件形式提供服务，不需要业务上修改代码逻辑，基于Fuse框架。属于Fuse-like类的文件系统。

FastImage

使用Ublk，对外体系为块设备，并且是一种只读块，元数据全部缓存在本地，元数据操作性能高。

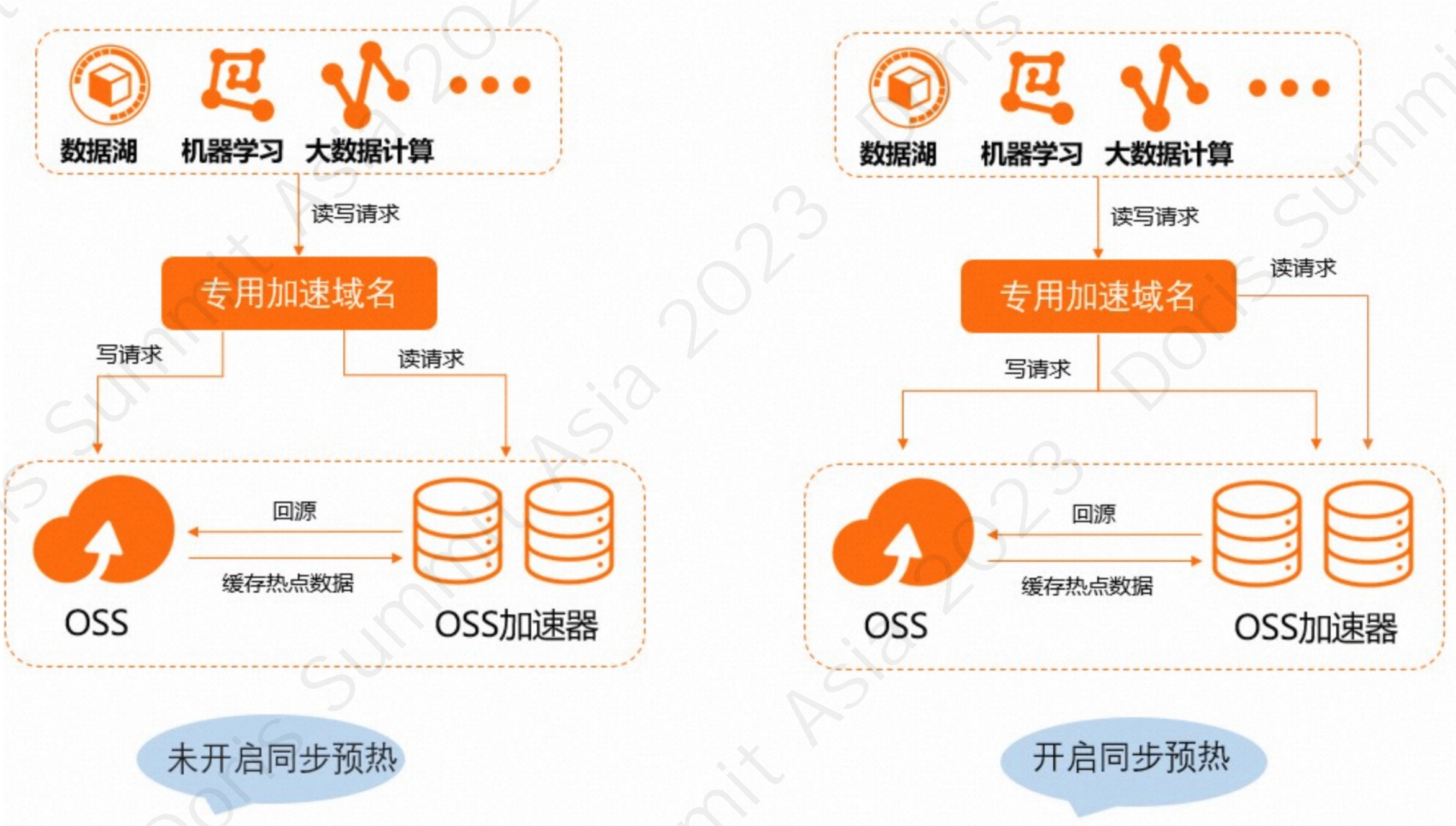
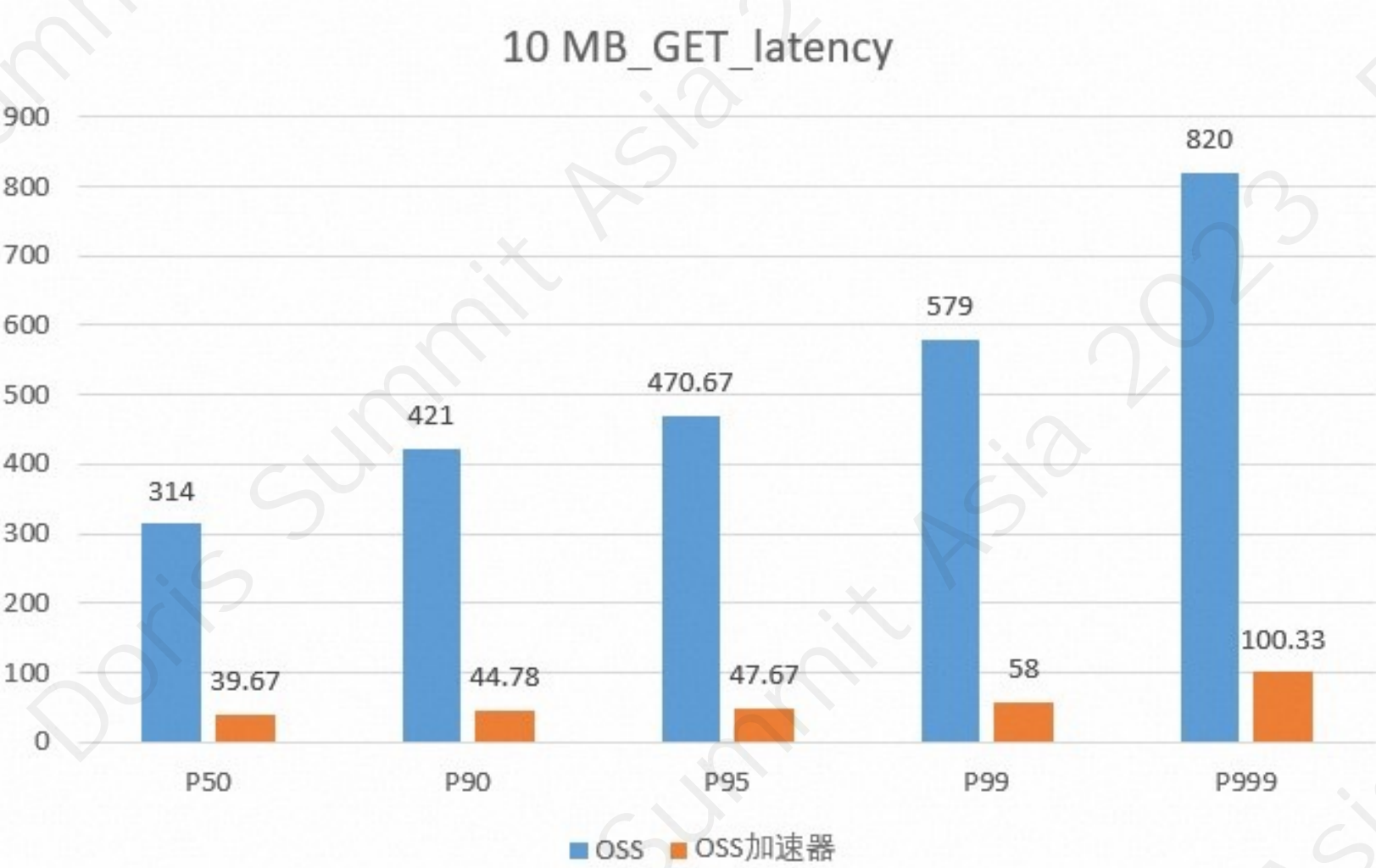
OSS-SDK

不依赖额外的工具。但是需要高性能的情况下要业务层进行封装和调试。

OSS加速器：加速数据获取

缓存OSS中的热点文件（Object），提供高性能、高吞吐量的数据访问服务。

- 1. 弹性带宽：根据申请容量自动调整带宽。
- 2. 更快的获取：更快的下载速度。
- 3. 使用简便：无需部署，开箱即用。
- 4. 强一致：数据保持和oss强一致关系，无需特殊的同步策略。



OSS加速器：构建温数据层



两种预热模式

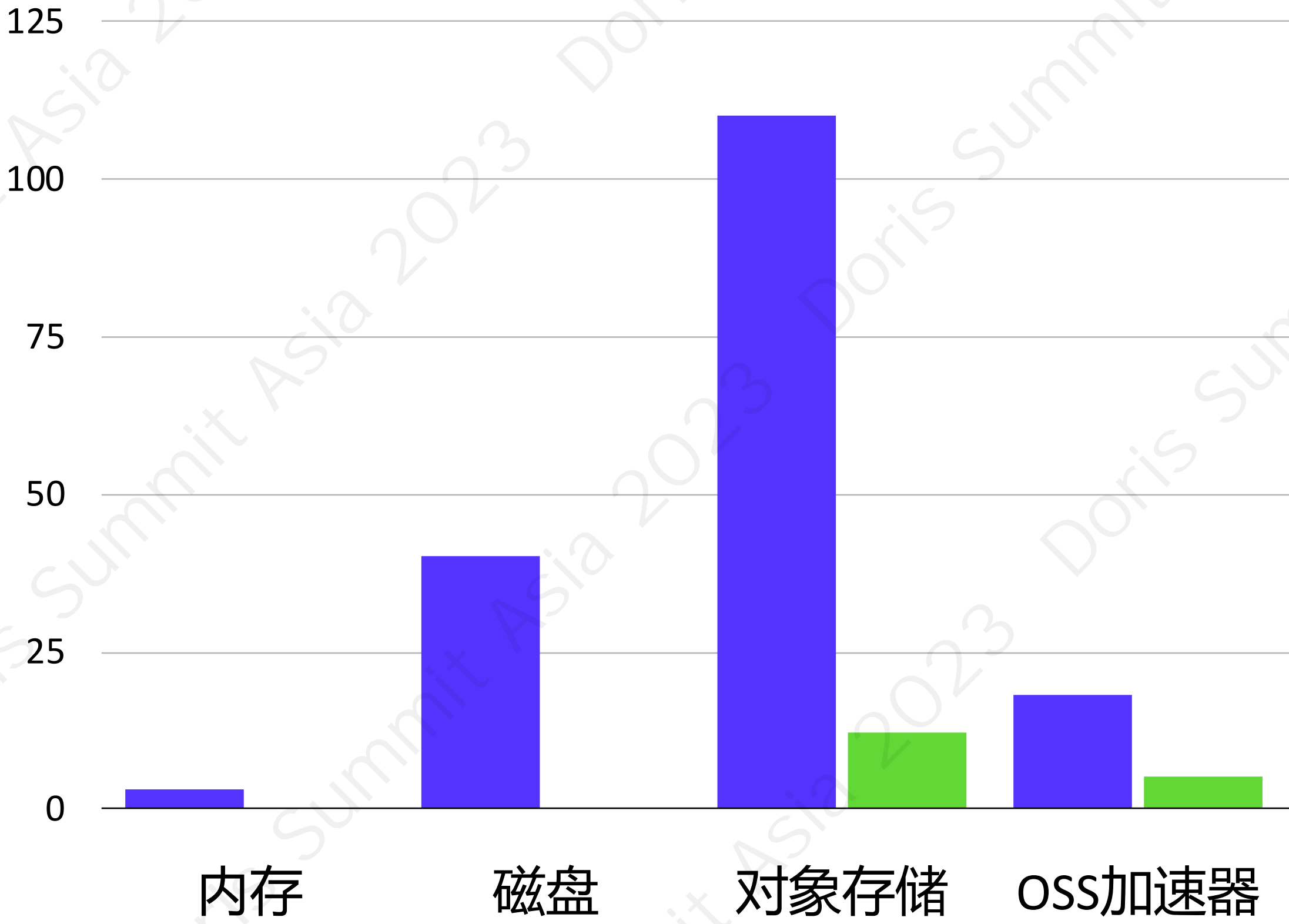
- 写时预热：写入数据同时写入加速器
- 读时预热：读取数据同时写入加速器

构建暖数据层

某数仓冷数据查询场景：可以大幅度降低冷数据性能回退。全量扫描、10%随机查询性能是直读OSS的2 ~ 2.5倍。整体约为本地 Cache 的30%-85%。

下载性能

项目/性能 (5.6GB 单文件 load)	耗时约x秒	说明
内存	3	单机
磁盘	40	单机单盘
对象存储	12-110	并发-单文件
OSS加速器	5-18	并发-单文件



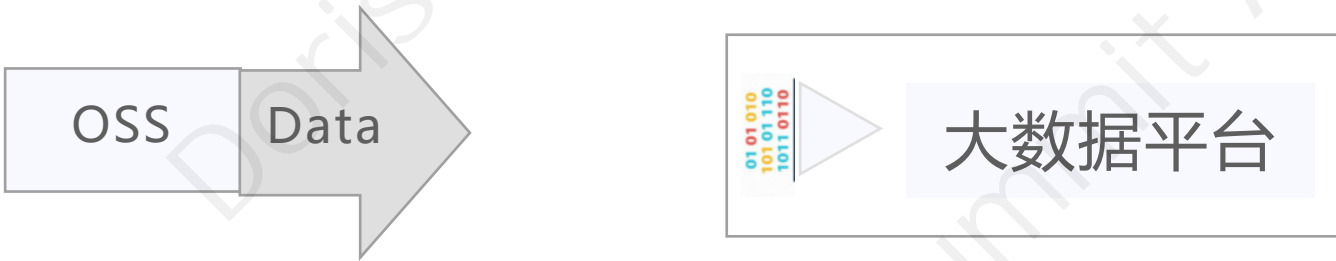
OSS-Select: 轻量级的计算引擎

存算分离模式下网络传输带宽和耗时往往会制约系统性能的发挥，因此即便是 Hadoop、Spark 这种一开始便采用存算分离模式的分布式框架，也会尽量将计算逻辑推送到数据所在的节点以此来提升计算任务的执行性能。

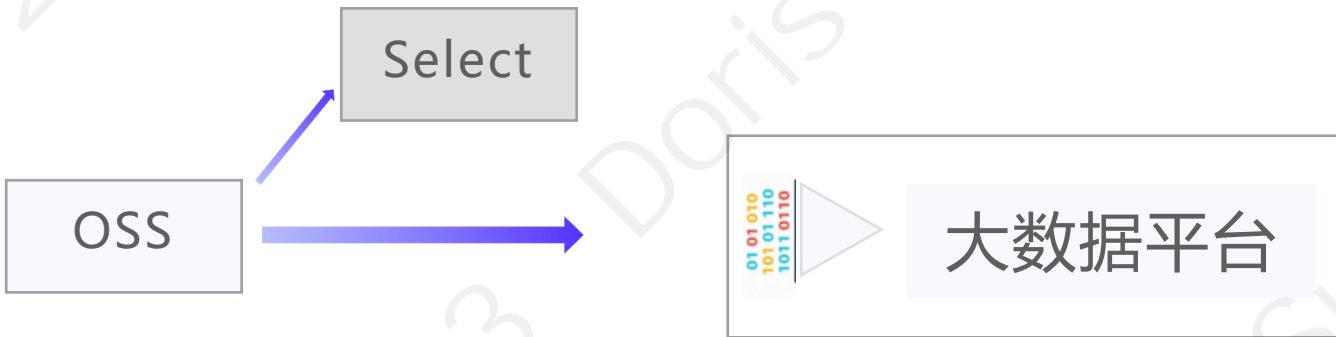
OSS Select是运行在OSS内部的轻量级的计算引擎，它不是用来代替已有的大数据平台，相反它旨在让已有的大数据平台更快更经济地访问OSS数据，大大减少ECS端的带宽消耗，从而降低用户的总体成本。

OSS Select目前仍在不断演进中，未来将支持BZIP等CSV文件，以及ORC等列存文件，同时也将支持更丰富的内置函数。

用 OSS Select 前



用 OSS Select 后



Select功能	支持
支持CSV	✓
支持JSON	✓
支持ORC	X
支持parquet	X
SelectTo	✓
Object Lambda	即将

OSS-Select: 构建混合模式

文件 100GB_TPCCH\lineitem.csv.1, 文件大小 2.27G。

用 OssUtil 下载该文件时间: 66.7s, 其中服务器端处理时间 11.2s

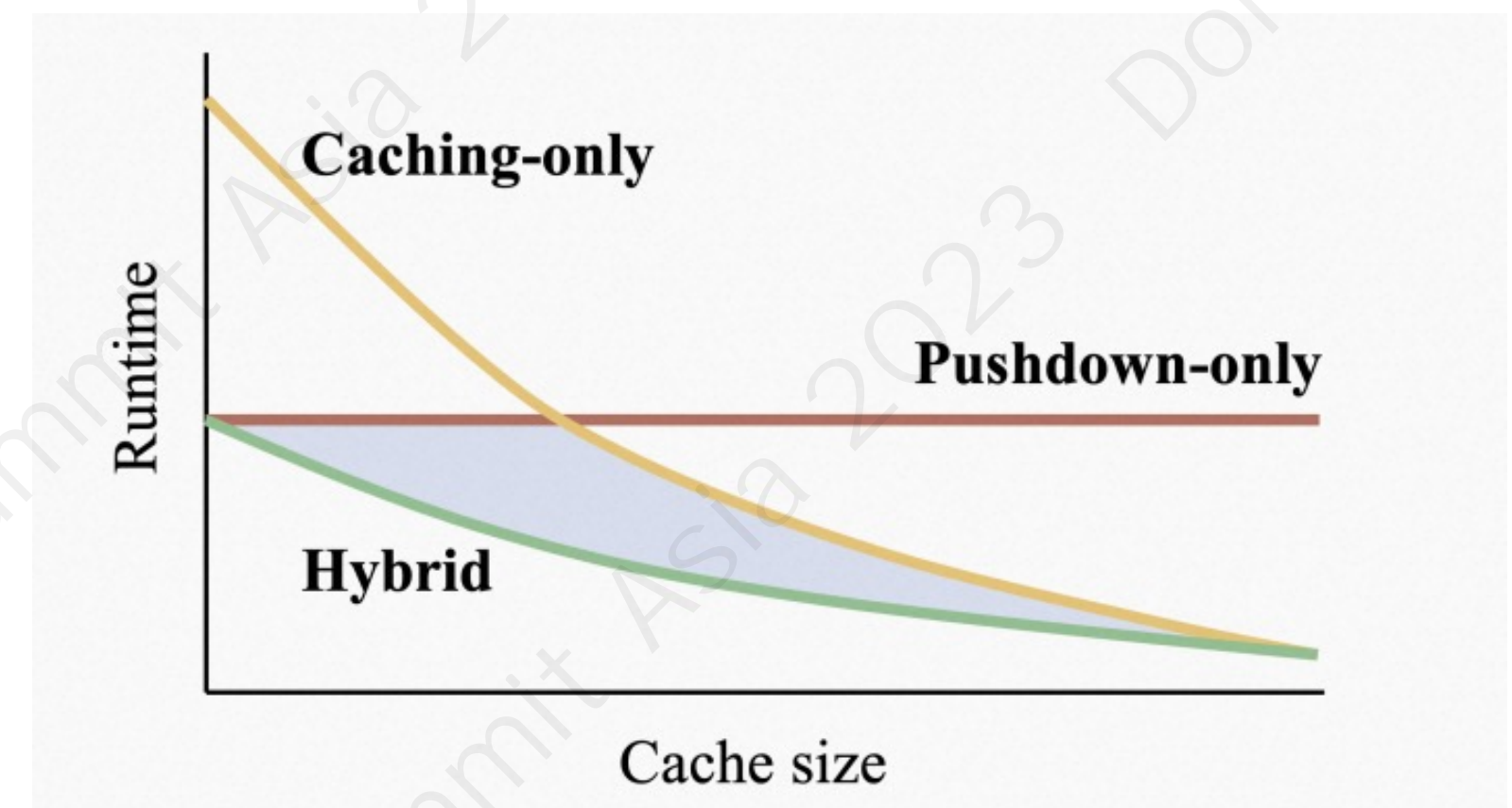
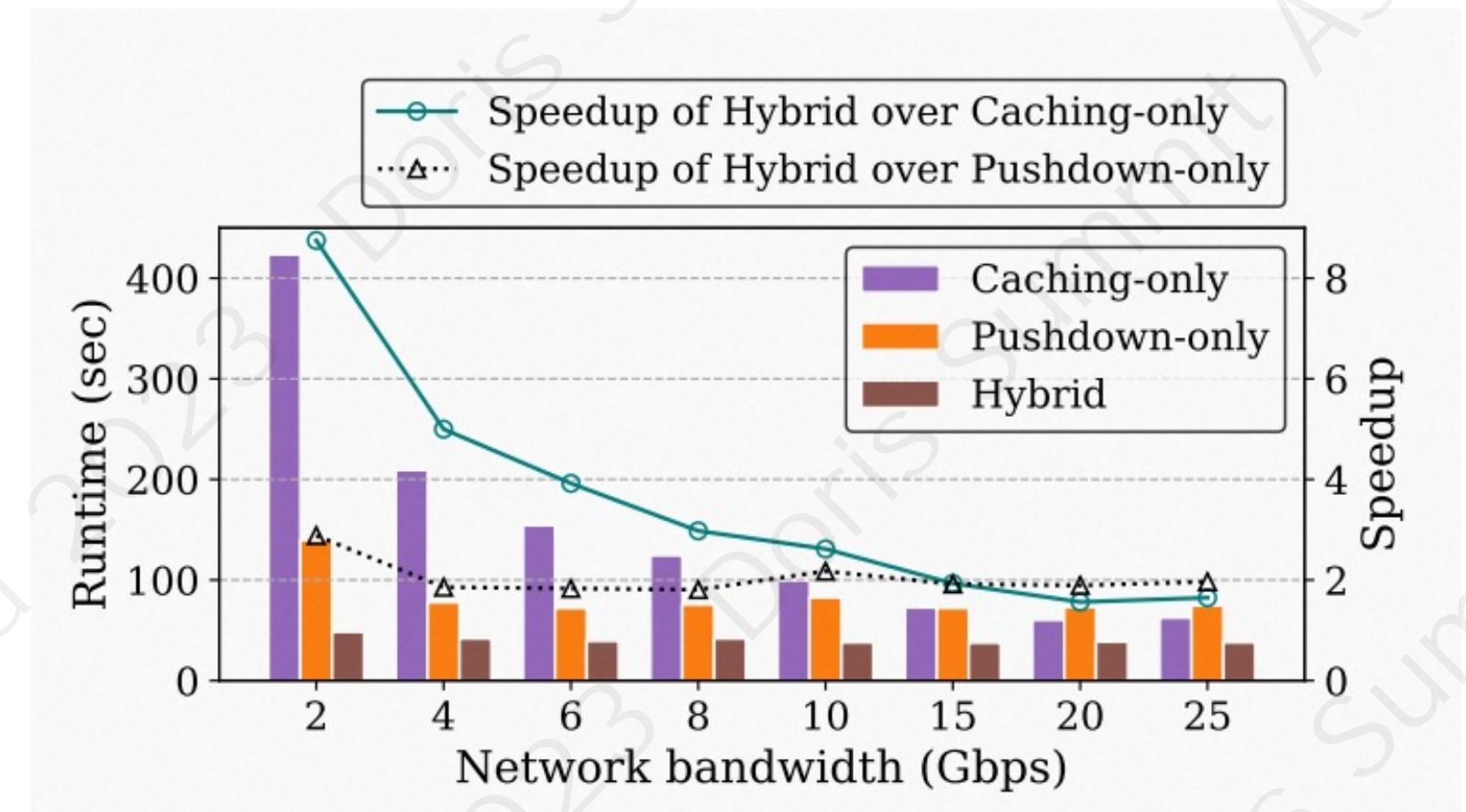
```
select count(*) from ossobject where cast(_12 as timestamp) >  
cast('1996-01-01' as timestamp)
```

总时间: 7.9s. 返回数据大小: 0.002KB

混合模式:

有研究表明, 在云上环境中, 存算分离数仓使用下推后能极大的降低带宽使用和相对小缓存空间的情况下获得相对更快的运行速度。

可以结合 local cache 以及 pushdown, 以便达到最优的效果。



归档直读：更加方便的归档

解决的问题：

归档虽然存储成本更低，但是其需要解冻才能使用，这就导致很多情况下无法直接使用。

归档直读既可以以归档的价格存储数据，又可以以标准的方式使用，为客户存储数据提供了更大的灵活性。

使用场景：

归档直读适用于需要实时读取极低频访问数据的场景，例如数据湖、云相册、媒体资产归档、医疗影像等。

在这些场景下，归档直读能够满足极低频访问数据的实时读取的业务需求，同时兼顾低存储成本与实时访问能力。

费用说明：

为Bucket开启归档直读后，直接读取Bucket中未解冻的归档存储类型文件，会产生归档直读数据取回容量（RetrievalDataArchiveDirect）费用。对于已解冻的归档存储类型文件，直接读取不会产生归档直读数据取回容量费用。

项目	未开启归档直读	开启归档直读
取回方式	先解冻，再读取	直接读取
取回费用	低	高
取回时间	分钟级	毫秒级

数据取回	0.06元/GB
归档直读数据取回容量	0.2元/GB

标准型单价	低频访问型单价	归档型单价
0.12元/GB/月	0.08元/GB/月	0.033元/GB/月



获取更多社区动态与最佳实践

Apache Doris 官方平台:

- Apache Doris 官网: doris.apache.org
- Apache Doris GitHub: github.com/apache/doris/

获取更多峰会资料:

- Doris Summit 峰会官网: doris-summit.org.cn
- Doris Summit 峰会回放: <https://space.bilibili.com/1196172099/channel/collectiondetail?sid=1824324>