Azure Data Lake Store: A Hyperscale Distributed File Service for Big Data Analytics

数据库课堂报告

2001213077 陈俊达

2020年12月8日

Azure Data Lake Storage

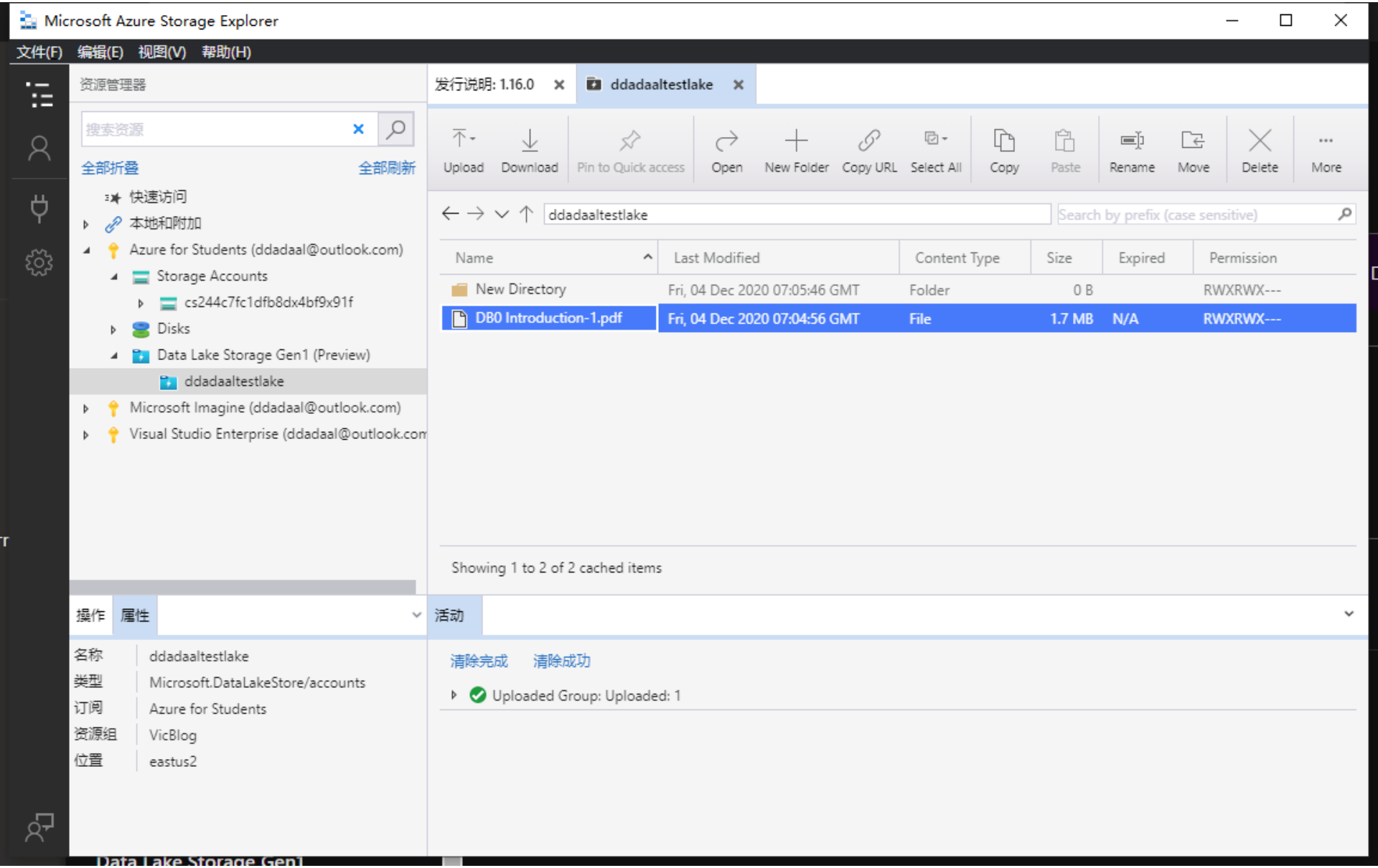
Massively scalable and secure data lake for your high-performance analytics workloads

Start free

Product overview | Features | Security | Pricing | Resources | Customer stories | Updates | FAQs

Build a foundation for your analytics

Eliminate data silos with a single storage platform. Optimize costs with tiered storage and policy management. Authenticate data using Azure Active Directory (Azure AD) and role-based access control (RBAC). And help protect data with security features like encryption at rest and advanced threat protection.

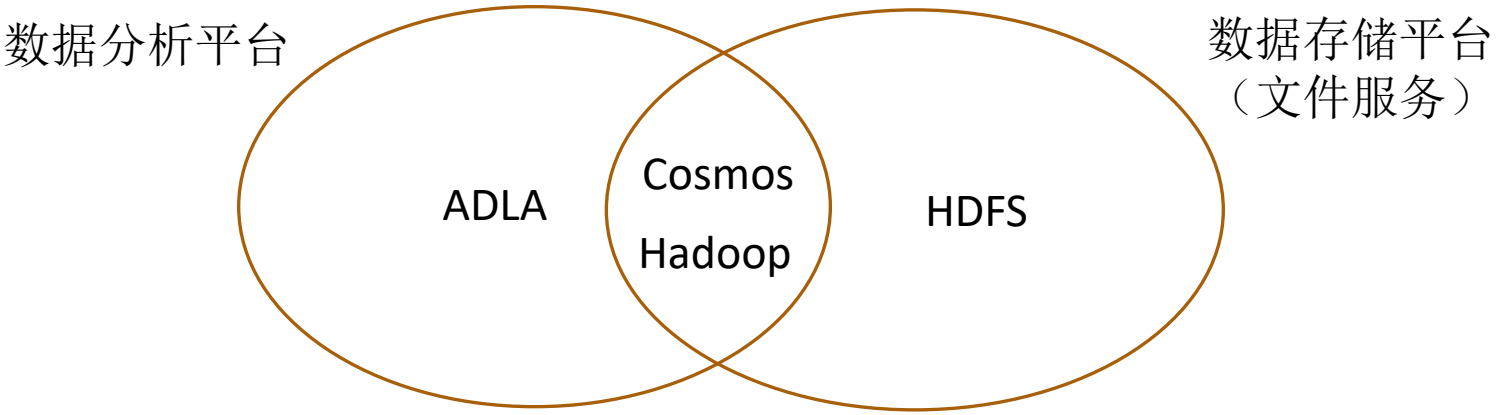

Microsoft Azure Storage Explorer
文件(F) 编辑(E) 视图(V) 帮助(H)
资源管理器
搜索资源
全部折叠
全部刷新
快速访问
本地和附加
Azure for Students (ddadaal@outlook.com)
Storage Accounts
cs244c7fc1dfb8dx4bf9x91f
Disks
Data Lake Storage Gen1 (Preview)
ddadaaltestlake
Microsoft Imagine (ddadaal@outlook.com)
Visual Studio Enterprise (ddadaal@outlook.com
发行说明: 1.16.0
ddadaaltestlake
Upload
Download
Pin to Quick access
Open
New Folder
Copy URL
Select All
Copy
Paste
Rename
Move
Delete
More
ddadaaltestlake
Search by prefix (case sensitive)
Name
Last Modified
Content Type
Size
Expired
Permission
New Directory
Fri, 04 Dec 2020 07:05:46 GMT
Folder
0 B
RWXRWX---
DB0 Introduction-1.pdf
Fri, 04 Dec 2020 07:04:56 GMT
File
1.7 MB
N/A
RWXRWX---
Showing 1 to 2 of 2 cached items
操作
属性
名称 ddadaaltestlake
类型 Microsoft.DataLakeStore/accounts
订阅 Azure for Students
资源组 VicBlog
位置 eastus2
活动
清除完成
清除成功
Uploaded Group: Uploaded: 1

目录

系统概述

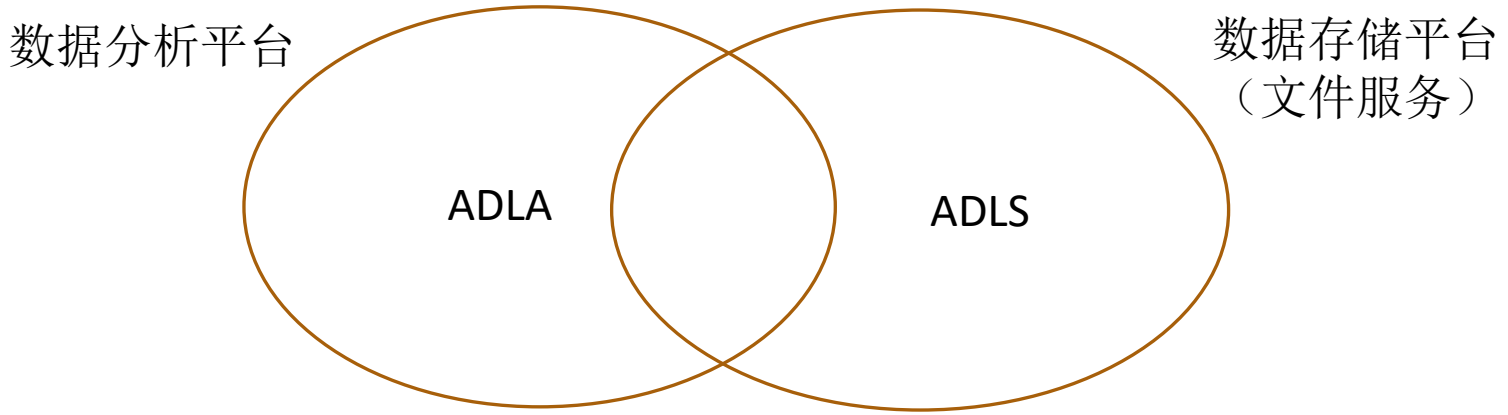
相关技术

- **Cosmos**：微软自己的大数据设施
  - 文件服务 + 基于SQL方言U-SQL的数据分析
- **ADLA**：Azure Data Lake Analytics
  - 数据分析平台，执行计算任务
- **Hadoop**：开源的大数据存储和分析平台标杆
  - HDFS：Hadoop使用的文件系统



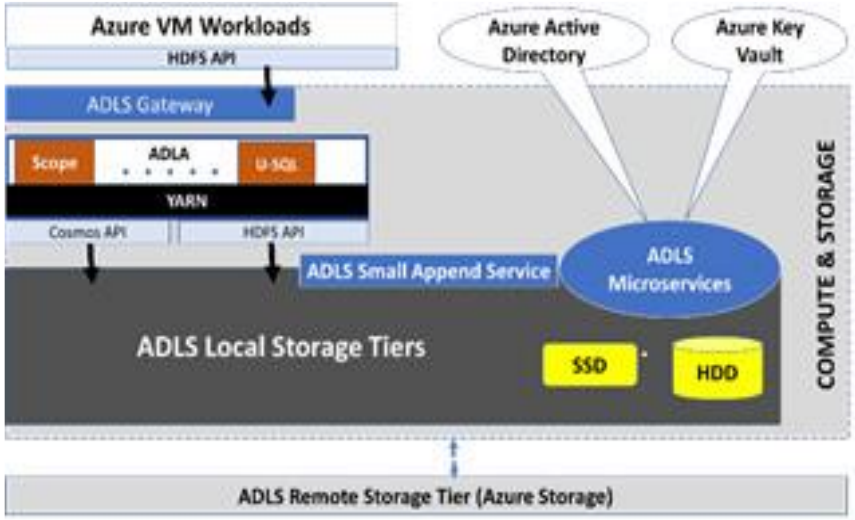
ADLS

- **ADLS：** Azure Data Lake Store
- 分布式文件服务/数据存储平台
- ADLS自Cosmos演化而来，兼容Hadoop
  - 如HDFS协议



ADLS的特点

- **多层次存储**（multiple storage tiers）
  - 数据越来越多
  - 要求数据能够被低代价地存储
  - **本地存储**和**远端存储**
  - 任务**需要使用**的数据放在**本地存储**上
    - 离计算任务较近的地方
    - 任务所在的机器、机房上
  - 当前**不需要使用的数据**放在**远端存储**
  - 开始计算任务前生成计划
    - 找到计算所需数据的存放的位置
    - 把计算任务放到离数据最近的节点上
    - 如果数据只存在于远端节点上，则按需获取



ADLS的特点

- **规模巨大**（exabyte scale， 1 EB = 1百万 TB）
  - Cosmos的要求
  - 最大的hadoop集群：5000节点；Cosmos每个集群：**超过50000节点**
  - 单个文件**PB**级（1000 TB）
  - 每天处理**数百PB**的数据
  - 单个任务可以在**10000个节点**上执行
  - 集群容量不够时需要迁移数据，保证迁移过程中任务仍然可以正常执行

ADLS的特点

- **完整的安全和数据共享措施**
  - Comprehensive security and data sharing
  - 数据可以按**用户**、**目录**和**文件**设定权限
  - 使用**Azure Active Directory**进行鉴权操作，与其他Azure服务和传统企业生态集成
  - 加密存储数据，密钥可以由系统或者用户拥有
  - 模块化的设计使得可以应用多种**安全措施**

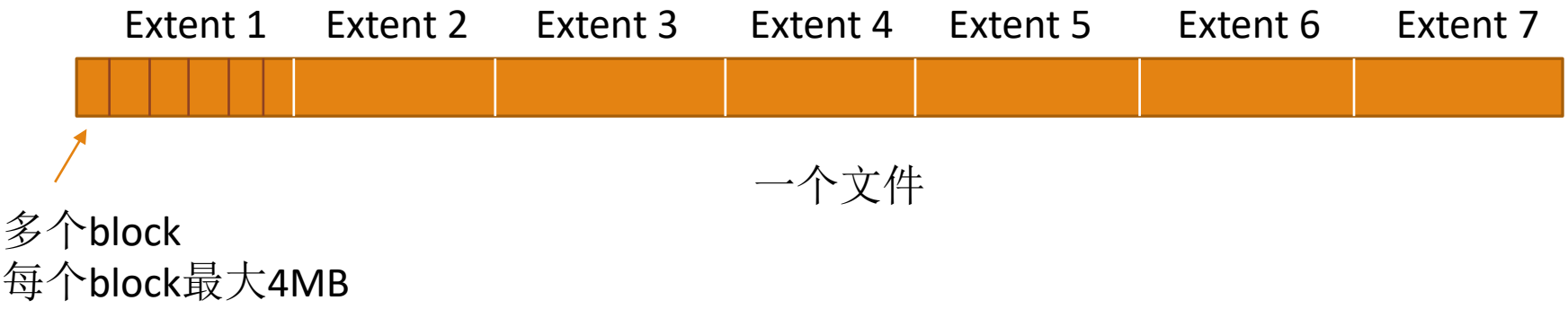
Naming Service

File Name	File ID	Parent	Children	ACL
myfolder	100	(none)	120, 123	744
ABC	120	100		744
XYZ	123	100		744

架构

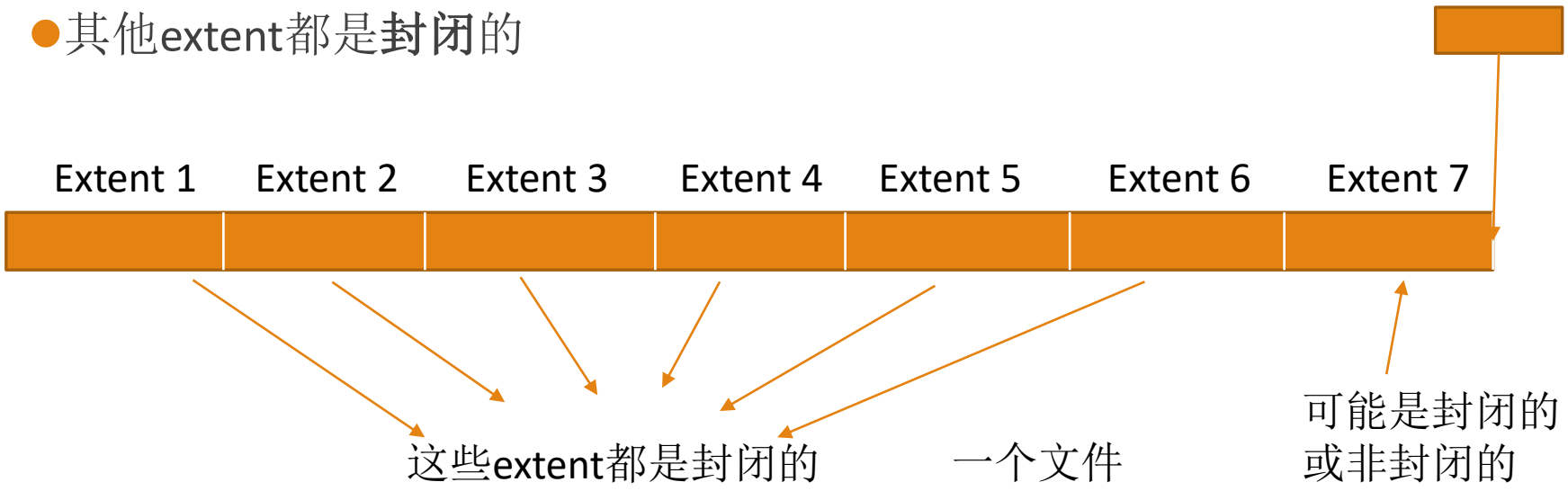
ADLS文件结构

- 一个文件由**extent**序列组成
- 每个extent包含多个**block**
- 每个block最大4MB



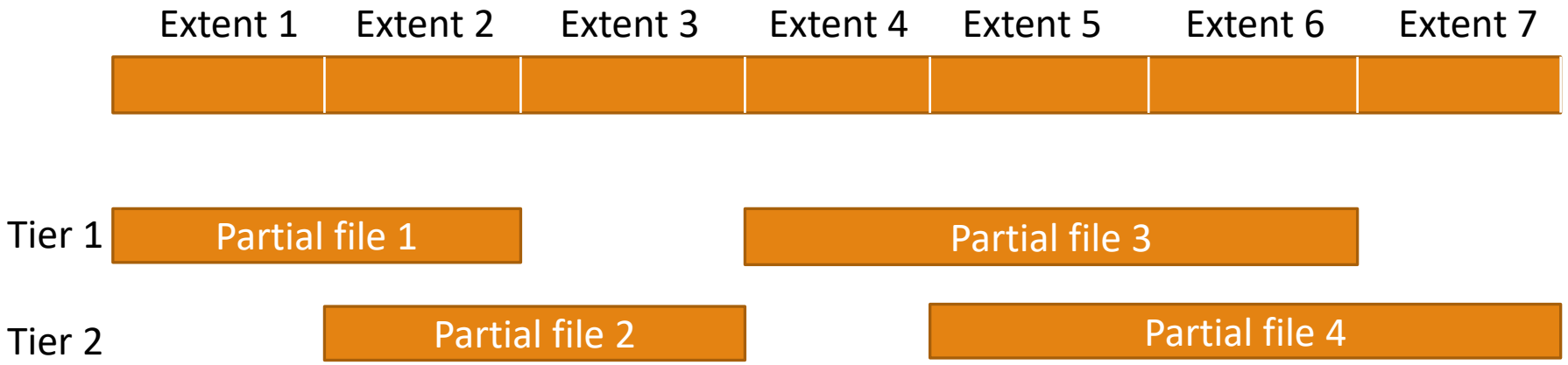
ADLS文件结构

- 每个时刻，一个文件的最后一个extent（**尾extent**）可能是
  - **封闭**的（seal）
  - **非封闭**的（unsealed）
- 只有尾extent是未封闭的文件才可以被追加（append）
- 其他extent都是**封闭**的



ADLS文件结构

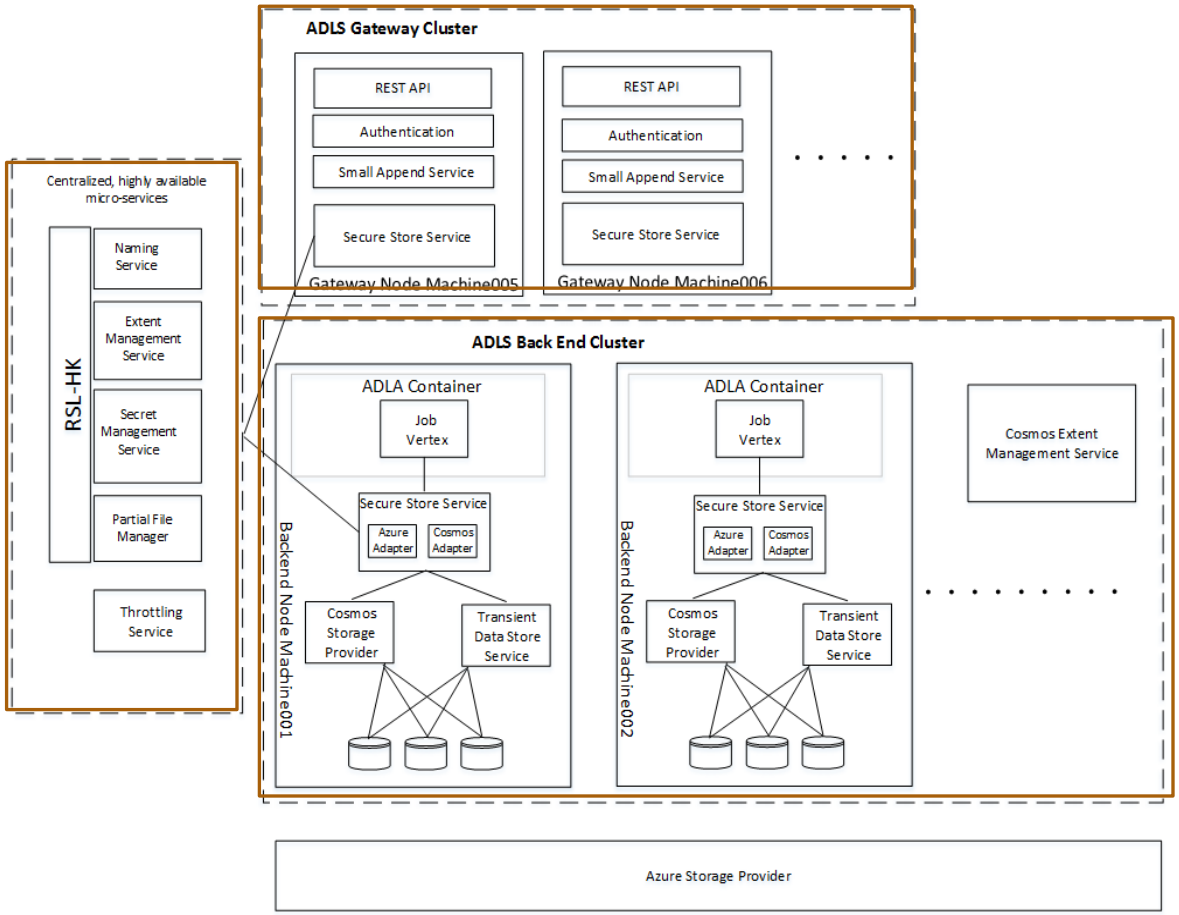
- 文件被分为多个**部分文件**（partial file）
- 一个部分文件是一个连续的extent序列
- 不同部分文件可放在不同存储层中，一个部分文件存在在一个存储层



只有一个尾（tail）部分文件

ADLS系统架构

- 微服务架构
- 一个集群中有3种节点
- **后端节点**
  - 最多
  - 数据的本地**存储**
  - **执行计算**任务
- **前端节点**
  - 网关
  - 路由请求和系统控制
- **微服务节点**
  - 托管微服务



ADLS系统组件

组件	作用简介
Secure Store Service	API的入口，编排各个微服务
RSL-HK	内部状态的数据库
Naming Service	映射文件名字到ID
Partial File Management Service	管理部分文件
Small Append Service	处理追加小文件的使用场景
Transient Data Store Service	临时（中间）数据存储
Throttling Service	管理账号使用限额
Storage Provider	具体的底层存储
Extent Management Service	管理Extent
Secret Management Service	管理系统秘密，处理数据加密解密

访问/myfolder/ABC文件

1. **路径 -> 文件：** NS中查找/myfolder
   1. Myfolder有两个children：120，123
   2. ABC文件ID是120

Naming Service

File Name	File ID	Parent	Children	ACL
myfolder	100	(none)	120, 123	744
ABC	120	100		744
XYZ	123	100		744

访问/myfolder/ABC文件

2. **文件 -> 部分文件：**
   2. PFM中查找文件ID为120的部分文件
   - 部分文件ID是4
   - 提供者是AS（Azure Storage，一个远端存储）

Partial File Management Service

Partial File ID	File ID	Provider ID	Partial File Properties	Partition ID
1	123	AS	-	P1
2	123	AS	-	P2
3	123	Cosmos	-	P2
4	120	AS	-	P1

访问/myfolder/ABC文件

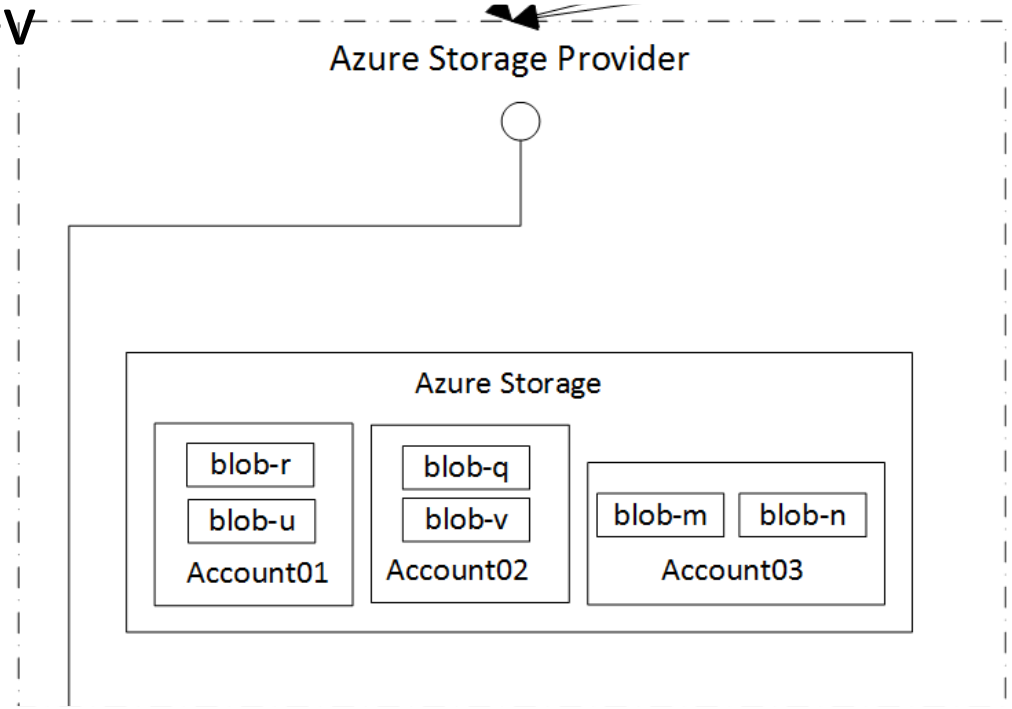
3. **部分文件 -> Extent：**
   3. 在AS的EMS中查找部分文件ID为4的extent
   - Account01, blob-u
   - Account02, blob-v

Extent Management Service (Shared)

Partial File ID	Extent ID	Extent Location
1	1	Account03,blob-m
1	2	Account03,blob-n
1	3	Account02,blob-q
2	1	Account03,blob-m
2	2	Account03,blob-n
2	101	Account01,blob-r
4	151	Account01,blob-u
4	152	Account02,blob-v

访问/myfolder/ABC文件

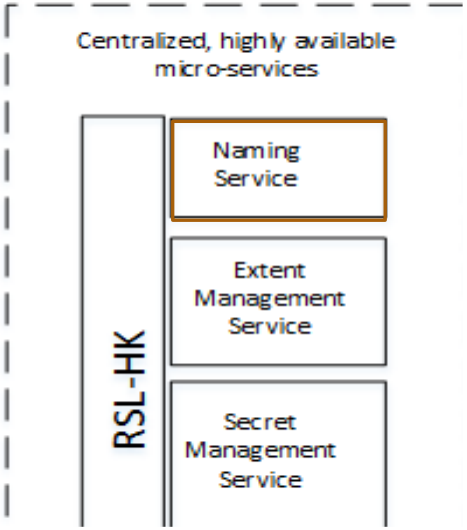
4. 读取内容
   - Account01, blob-u
   - Account02, blob-v



系统组件介绍

Naming Service（NS）

- 分布式名字服务
- **映射**文件名和文件ID
  - 内部服务通过ID使用文件
  - 文件系统的inode号
- **不移动内容，重命名**和**移动**文件
- **分层**的
  - /myfolder/myfile
- 支持POSIX风格的**权限控制**
  - owner, group, other, read, write, execute

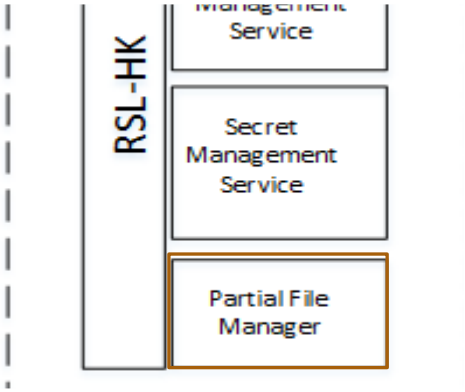


Naming Service

File Name	File ID	Parent	Children	ACL
myfolder	100	(none)	120, 123	744
ABC	120	100		744
XYZ	123	100		744

Partial File Management Service（PFM）

- 把文件ID和以下信息对应
  - 每个部分文件ID
  - 每个部分文件ID的存储提供者
- PFM存储
  - 每个部分文件的**开始**和**结束偏移量**
  - 部分文件的**创建**和**修改时间**
  - **文件大小**（所有部分文件封闭时）
  - 文件和部分文件是否**封闭**

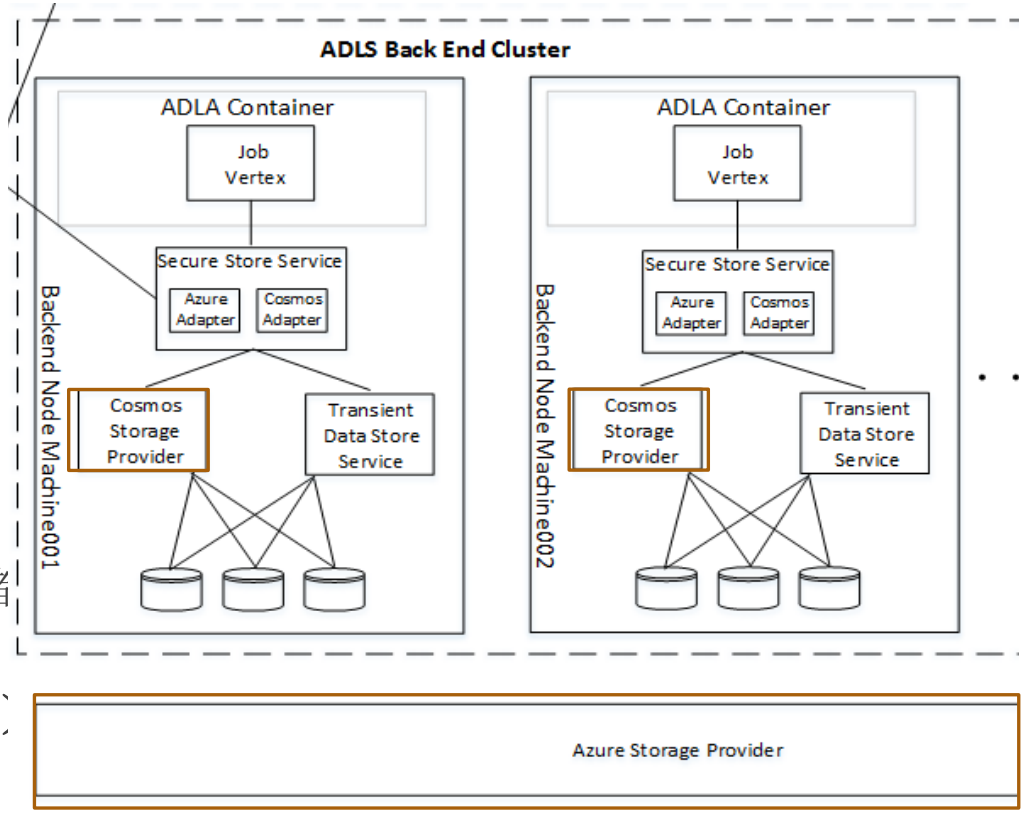


Partial File Management Service

Partial File ID	File ID	Provider ID	Partial File Properties	Partition ID
1	123	AS	-	P1
2	123	AS	-	P2
3	123	Cosmos	-	P2
4	120	AS	-	P1

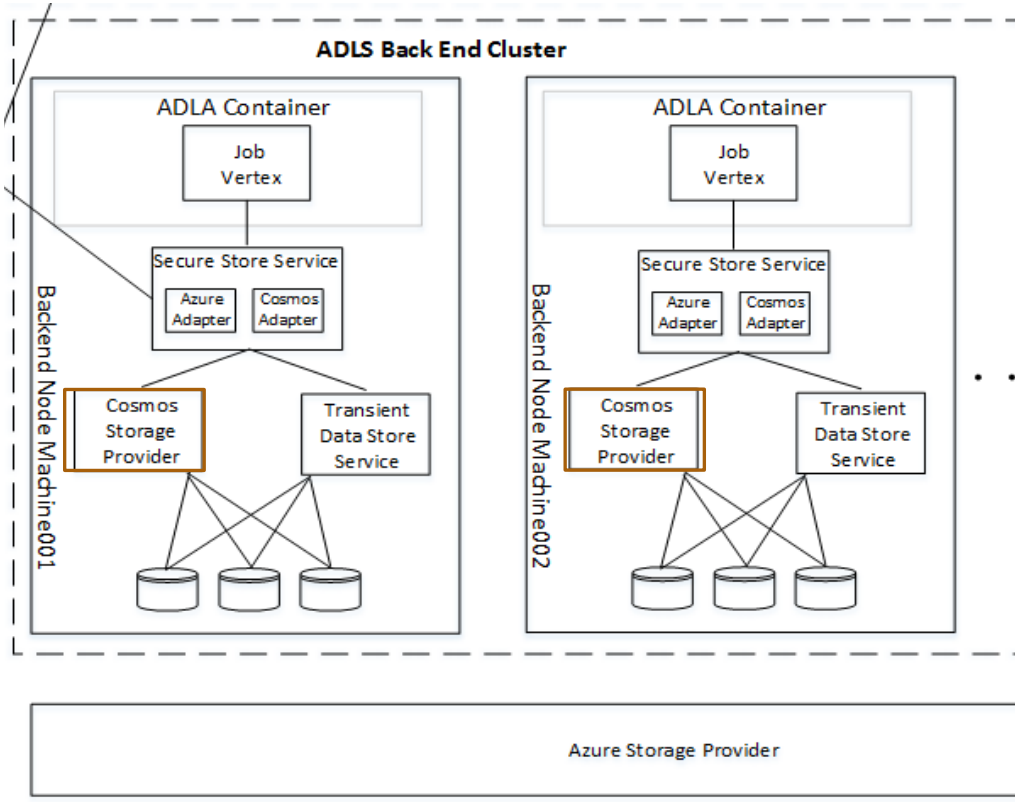
Storage Provider

- 存储提供者
- **单一命名空间**
- **只能追加**
- **对象存储**
- 存储**部分文件**
- 需要实现**存储提供者接口**
  - Storage provider interface
  - 运行在SSS上
  - 把文件操作请求转换成底层存储的请求
- **本地（local）和远端（remote）**



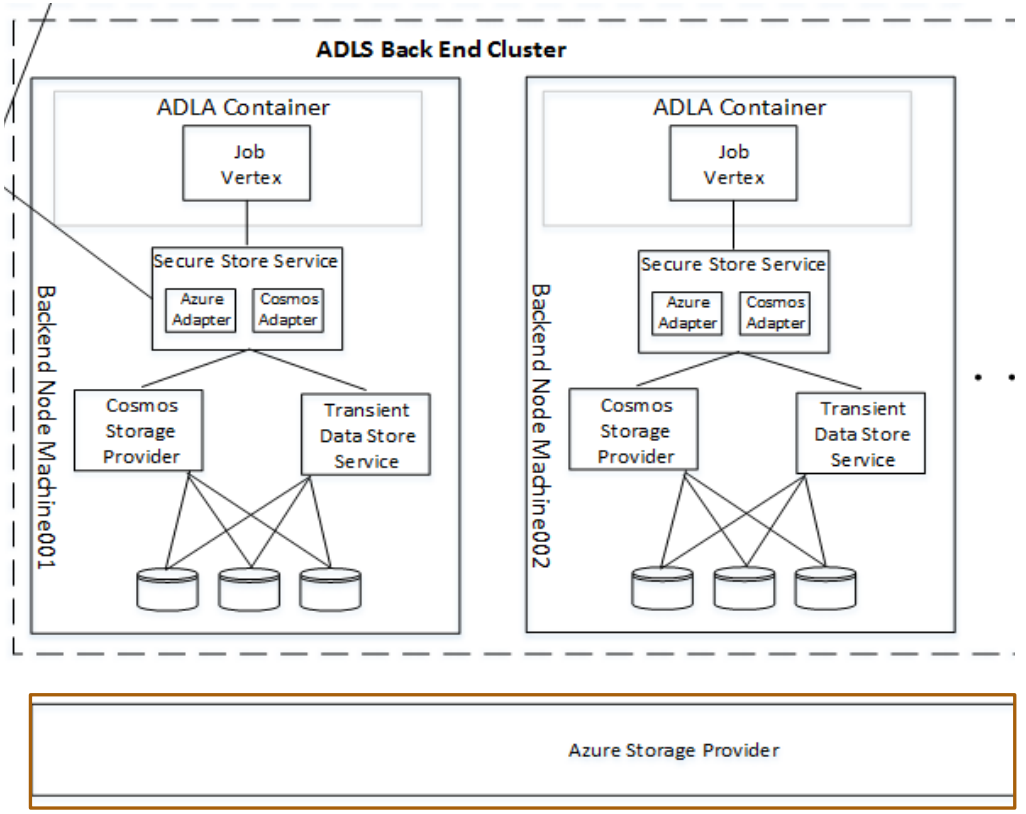
Local Storage Provider

- 本地存储提供者
- 数据存储在ADLS后端节点
- 这些节点上执行的计算可以快速访问这些数据
- 速度快，但是空间有限
- 示例：Cosmos



Remote Storage Provider

- 远端存储提供者
- 存储在集群之外
- 按需获取
- 速度慢，但是灵活，容量大，且便宜
- 示例：Azure Storage

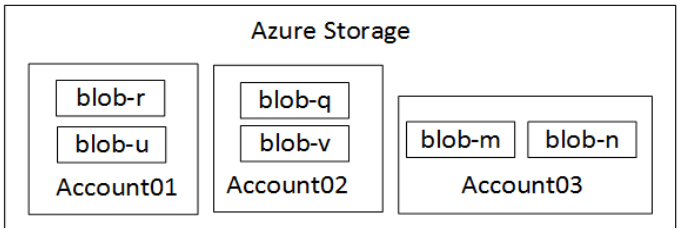


Extent Management Service（EMS）

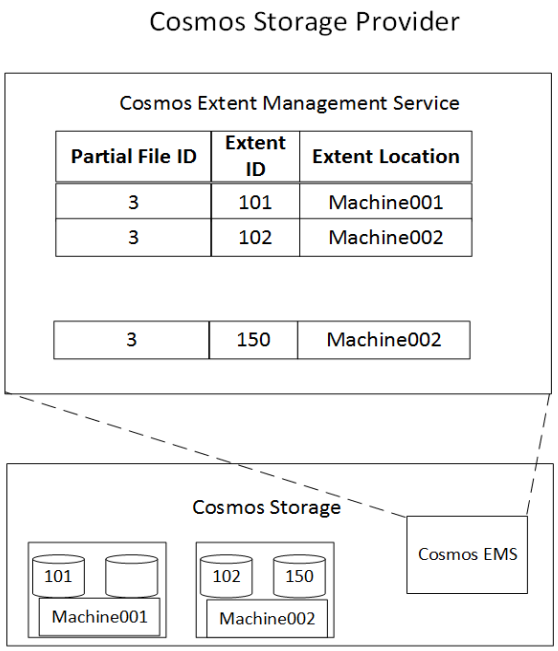
- 每个存储提供者有自己EMS
- 映射**部分文件和对应的extent**

Extent Management Service (Shared)

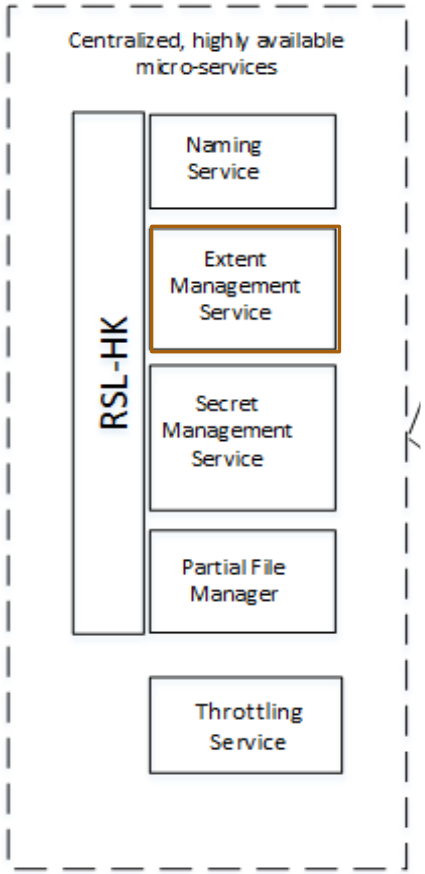
Partial File ID	Extent ID	Extent Location
1	1	Account03,blob-m
1	2	Account03,blob-n
1	3	Account02,blob-q
2	1	Account03,blob-m
2	2	Account03,blob-n
2	101	Account01,blob-r
4	151	Account01,blob-u
4	152	Account02,blob-v



Azure Storage的EMS

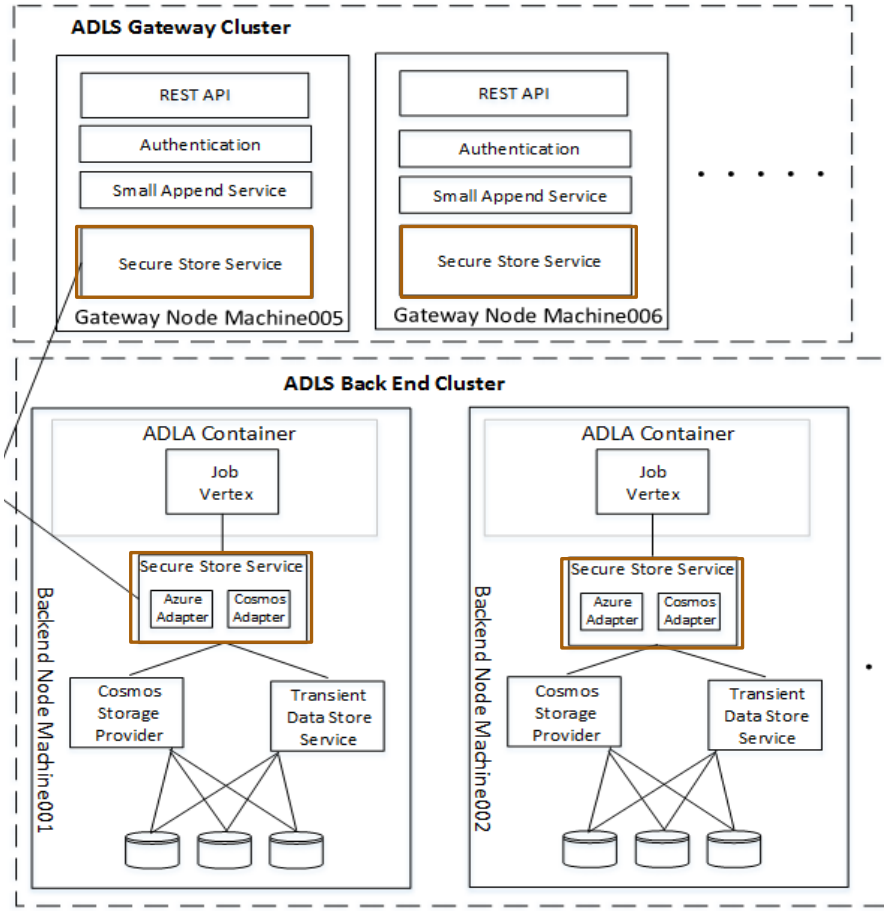


Cosmos Storage的EMS

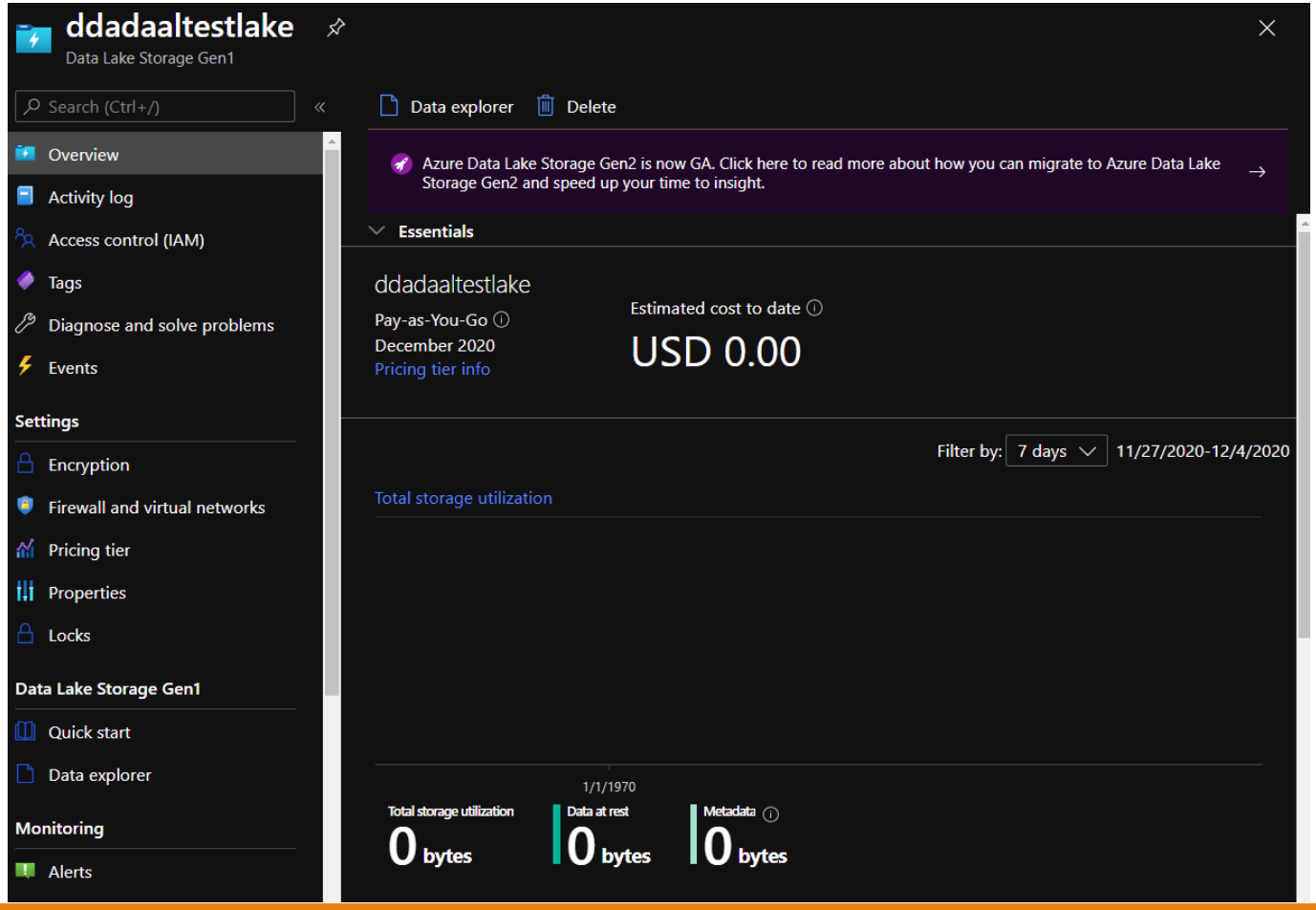


Secure Store Service（SSS）

- ADLS系统的入口
- **编排和协调**微服务
- 处理**请求失败**和**超时**
- 在ADLS系统文件格式和底层存储提供者的**文件格式**之间的**转换**
- 记录统计信息
- 执行带宽和吞吐量限制

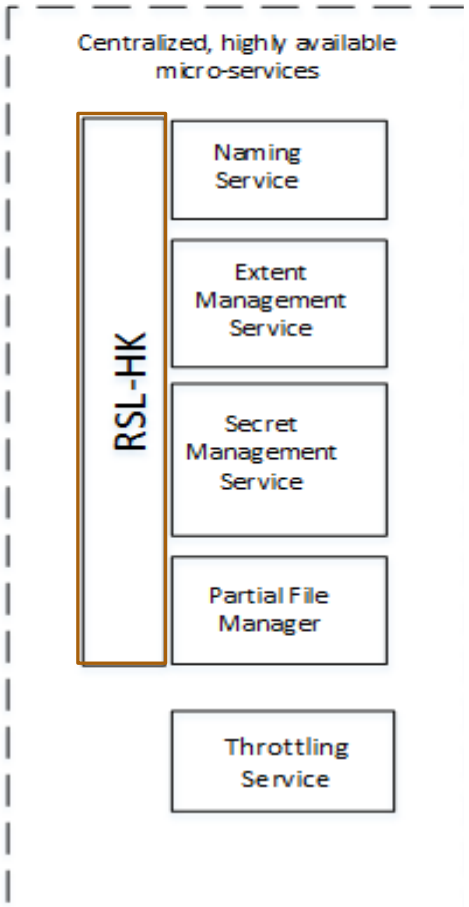


数据统计


ddadaaltestlake
Data Lake Storage Gen1
Search (Ctrl+/)
Data explorer
Delete
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Events
Settings
Encryption
Firewall and virtual networks
Pricing tier
Properties
Locks
Data Lake Storage Gen1
Quick start
Data explorer
Monitoring
Alerts
Azure Data Lake Storage Gen2 is now GA. Click here to read more about how you can migrate to Azure Data Lake Storage Gen2 and speed up your time to insight.
Essentials
ddadaaltestlake
Pay-as-You-Go
December 2020
Pricing tier info
Estimated cost to date
USD 0.00
Filter by: 7 days
11/27/2020-12/4/2020
Total storage utilization
1/1/1970
Total storage utilization
0 bytes
Data at rest
0 bytes
Metadata
0 bytes

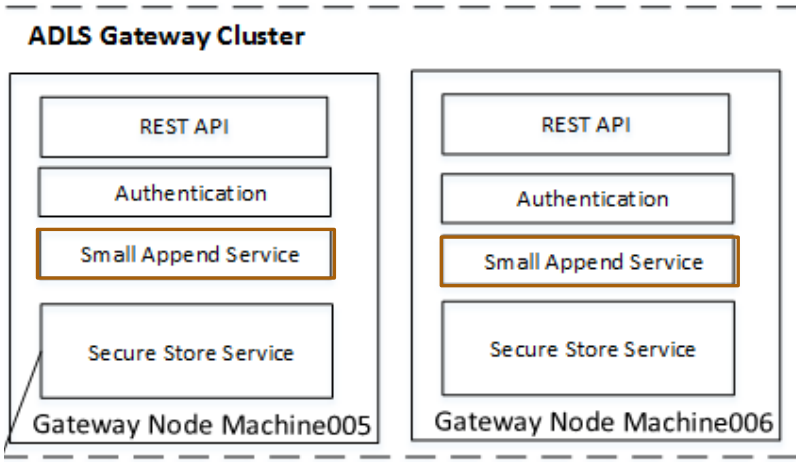
RSL-HK

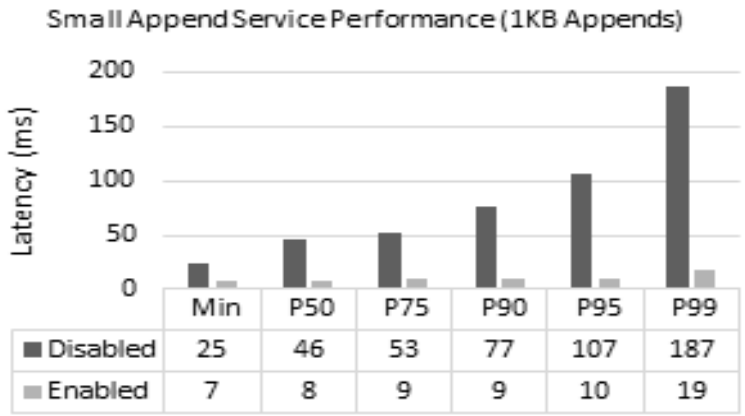
- 简单理解：内部状态的**数据库**
  - 对应Hadoop的Hbase？
- 管理各个服务的**持久状态**
- **满足ACID的事务性地操作数据**
- 检查点、记录、恢复
- 很多服务基于RSL-HK实现
- 基于Paxos协议
- 使用SQL Server的内存数据库引擎Hekaton



Small Append Service（SAS）

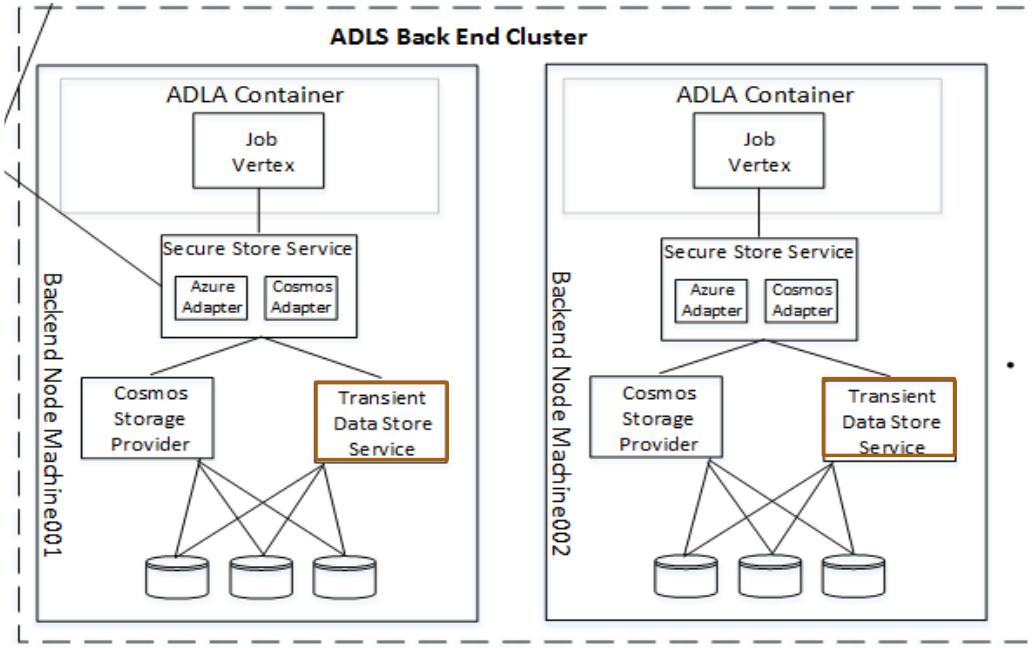
- 小文件追加服务
- 降低追加**大量小文件**的延迟
  - Hadoop具有这个问题
- 统一的文件追加API
- 动态检测追加文件的大小，若文件小则采用SAS
- 方法
  - 使用快速的**存储介质**（如SSD）
  - 把多个小文件**合并**为一个chunk
- 性能提升显著





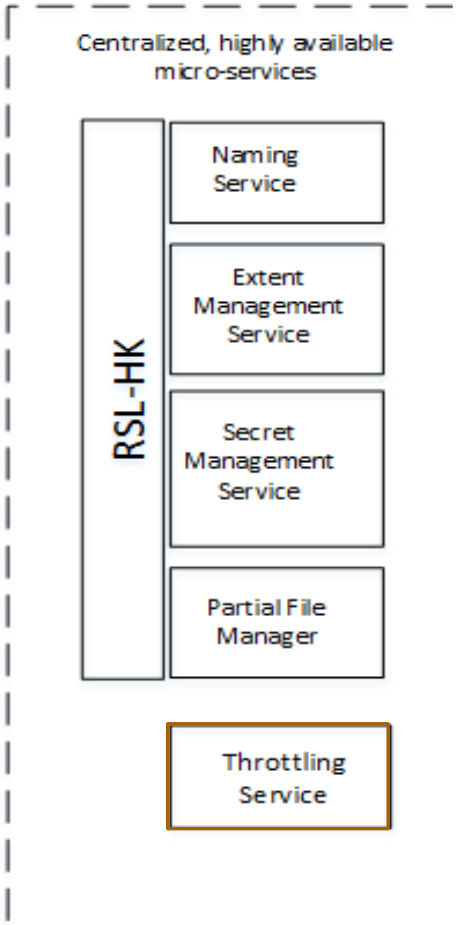
Transient Data Store Service（TDSS）

- 临时文件存储服务
- 临时存储ADLA计算任务的中间结果
- 每个后端节点均有此服务
- 为了支持debug，任务完成后仍保存一段时间的中间数据
  - 成功任务保存几个小时
  - 失败任务保存数天



Throttling Service（TS）

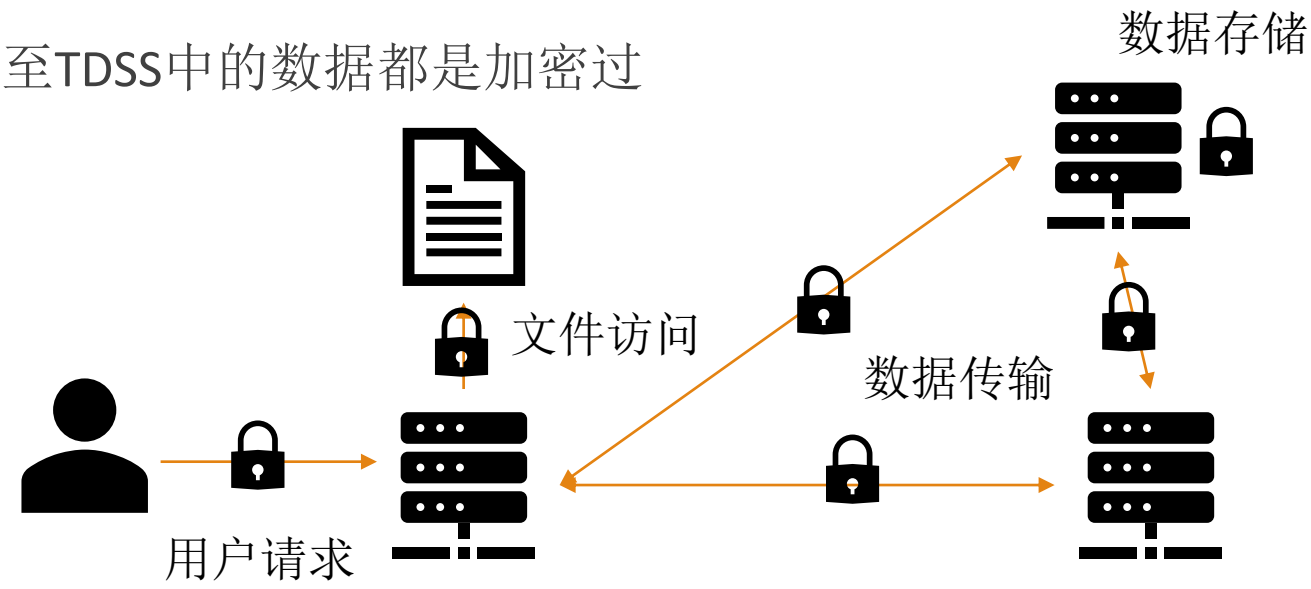
- 限流服务
- 收集各个账号对**带宽**和**读写次数**的数据
- 每50ms，SSS把信息发送给TS
- SSS在接下来50ms**停用**超过限额的账号的请求



安全性措施

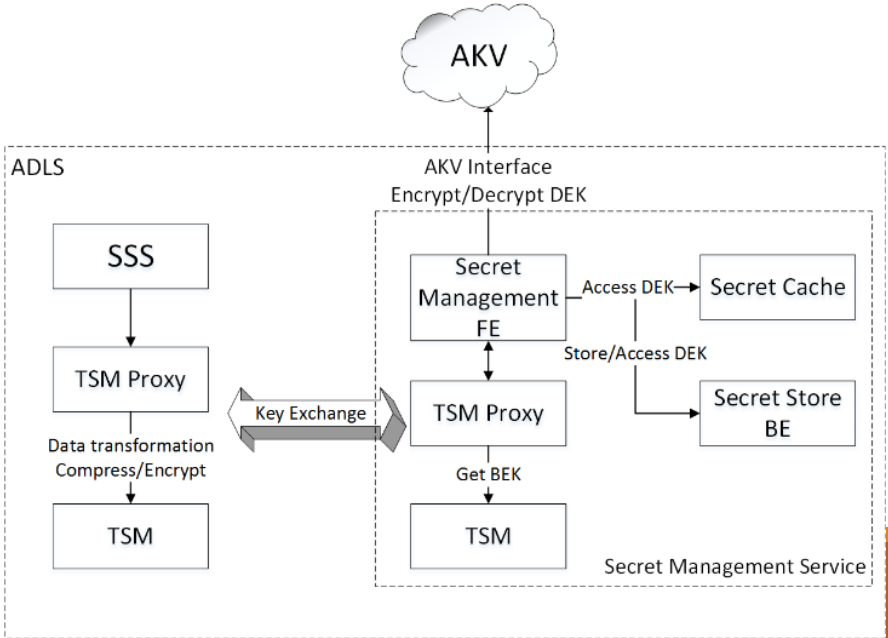
安全措施

- 使用Azure Active Directory实现对用户请求的**认证**（Authentication）和**授权**（Authorization）
- 在NS中实现POSIX风格的**访问控制**
- 进入ADLS后，ADLS之间的**数据传输**都是**加密**的
- 甚至TDSS中的数据都是加密过



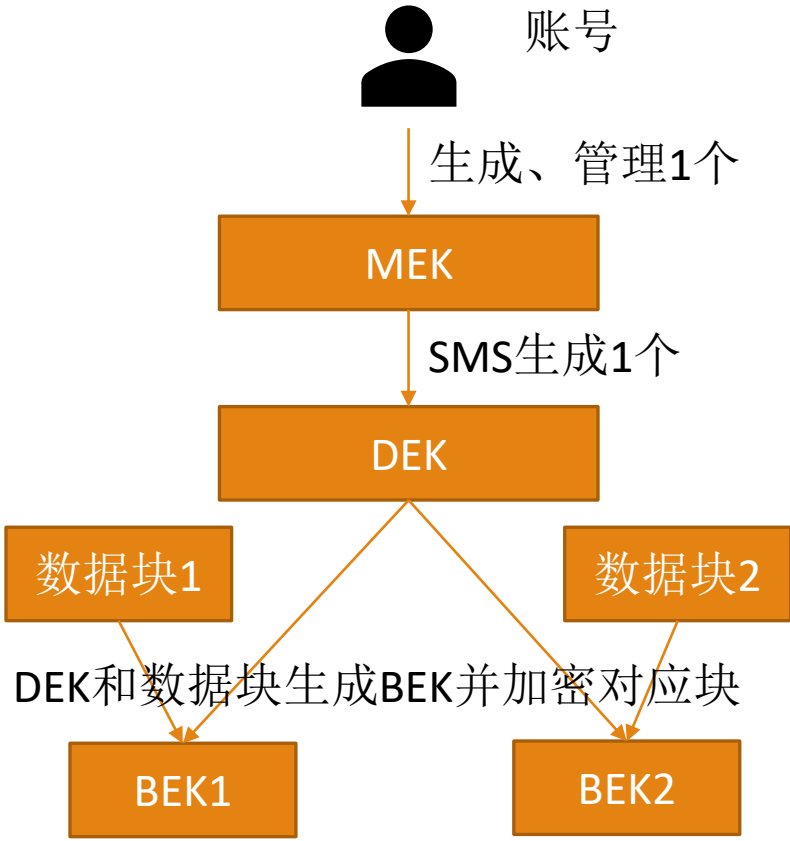
Secret Management Service

- **秘密仓库**
  - 按key-value存储秘密（密码等），使得对秘密的访问更安全
  - 与其他秘密仓库交互（如Azure Key Vault）
- 负责**数据的变换**（加密、解密、压缩、解压缩）
  - 在独立的**可信软件模块（TSM）**中执行
  - 设定安全边界
  - 保证对数据的操作不会泄漏
- 基于RSL-HK

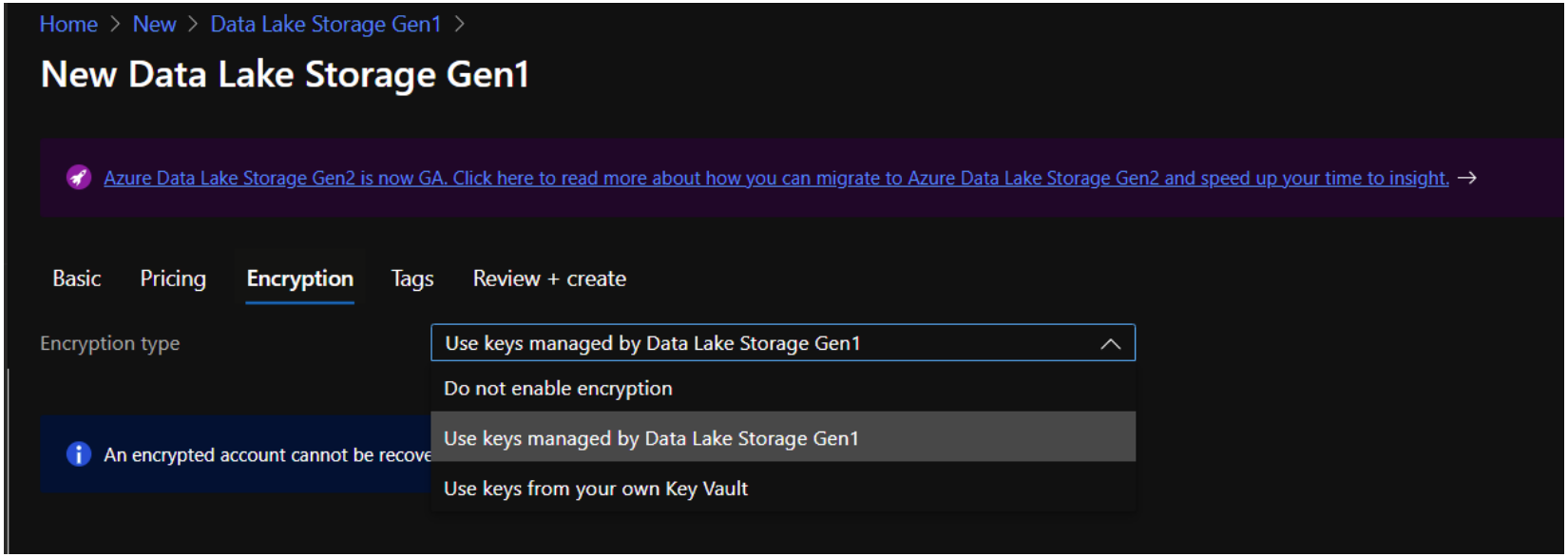


加密

- Master Encryption Key（MEK）
  - 每个ADLS账号一个MEK，存储在AKV中
  - 用户自己管理MEK的生命周期
    - 更换、删除、上传等
- Data Encryption Key（DEK）
  - 一个账号一个DEK，由SMS生成
  - DEK使用MEK加密，存储在ADLS集群中
- Block Encryption Key（BEK）
  - 对于**每个数据块**，使用账号的DEK和块的ID生成BEK并**加密此块**
  - 在TSM中执行
- 在一个**读写比例为6:4**的U-SQL请求中，时长超过3000小时，加解密只增加了**0.22%**的总执行时间。



加密



可以选择使用密钥的管理

总结和思考

总结

- 基于Cosmos、支持HDFS协议的大数据存储平台
- 是第一个公开的**在超大规模上支持完整文件系统**功能的PaaS平台
- 支持**多层次存储**
- 通过AAD、加密、TSM等方式保证了数据的安全和权限控制

ADLS系统组件

组件	作用简介
Secure Store Service	API的入口，编排各个微服务
RSL-HK	内部状态的数据库
Naming Service	映射文件名字到ID
Partial File Management Service	管理部分文件
Small Append Service	处理追加小文件的使用场景
Transient Data Store Service	临时（中间）数据存储
Throttling Service	管理账号使用限额
Storage Provider	具体的底层存储
Extent Management Service	管理Extent
Secret Management Service	管理系统秘密，处理数据加密解密

思考

- 微服务架构带来的可扩展性和伸缩性
  - SAS专门用来解决小文件合并的问题
  - 微服务可部署到各个系统上
  - 自动微服务调度
- 与Azure生态系统的紧密集成：挑战和机遇
  - Azure Active Directory; Azure Key Vault; 基于Azure Blob Storage的Gen2
  - 使用其他LDAP Provider进行鉴权
  - 使用其他Secret Store
  - 公开存储接口，接入第三方对象存储？
- 兼容性
  - 兼容现有生态：Hadoop
  - 使用Hadoop作为底层存储，节省重复造轮子

谢谢！