基于华为云OBS服务的大数据实践

存算分离架构

CSDN

大数据基础架构演进：走向计算存储分离

大数据架构变革

Hadoop 1.0
计算存储一体

大数据处理基本能力
- HDFS三副本可靠性
- MapReduce计算框架
- 计算节点和存储节点融合部署

面向单一MR分析业务

第一代：融合型
2011

Hadoop 2.0
资源解耦

解耦的大数据平台
- 独立的资源管理引擎Yarn
- 各计算引擎蓬勃发展
- 面向不同场景的nosql数据库蓬勃发展

面向复合型分析业务

第二代：扩展型
2013

Hadoop 3.0
存算分离

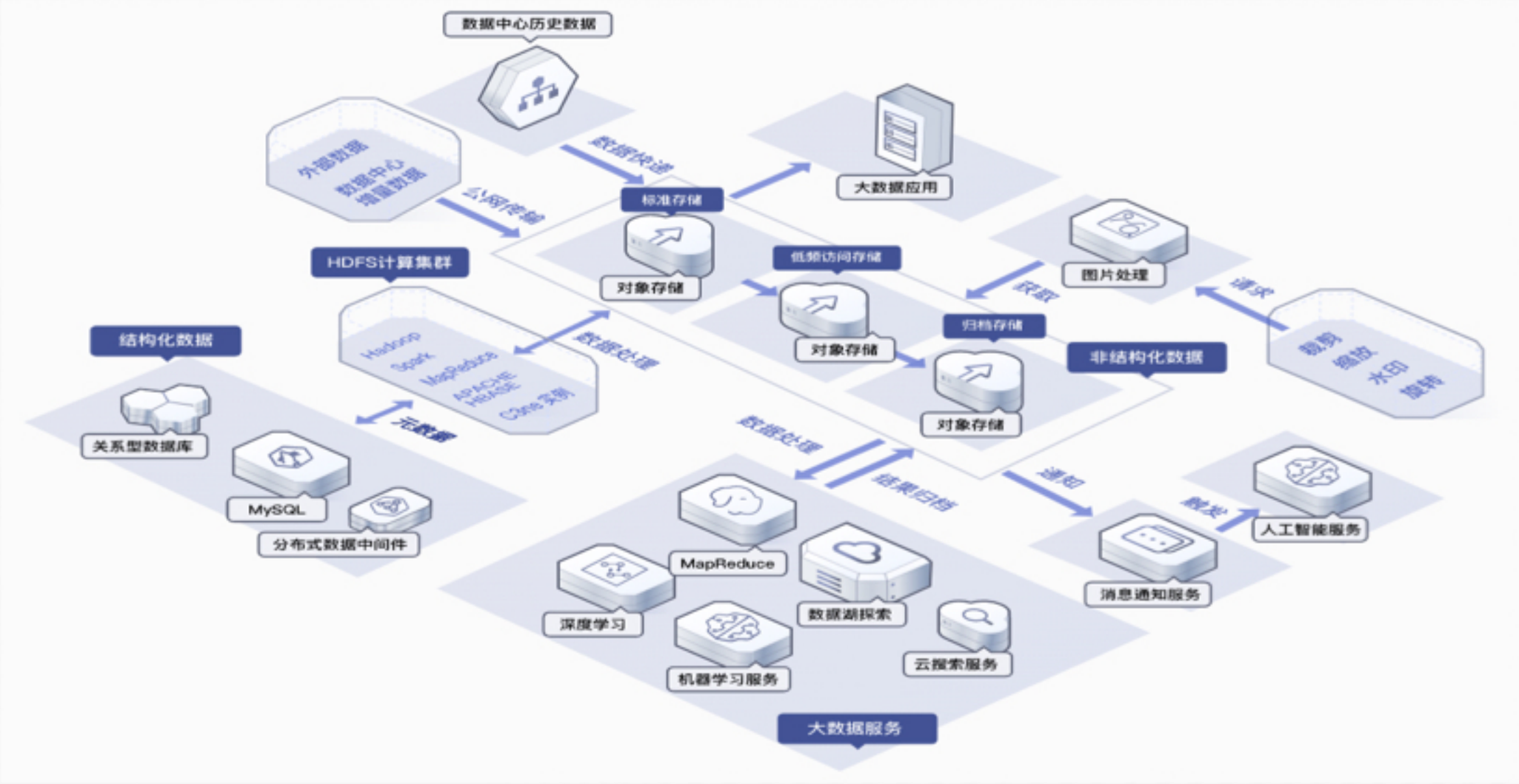
开放的云生态
- 基于对象存储的存算分离架构
- 支持容器化资源管理
- 分析性数据仓库的发展

面向数据湖，企业级大数据业务

第三代：开放型
2017

存算分离本质：构建计算和存储分离架构，能够扩存储不扩计算，并需要进一步降低存储成本

云上大数据架构



围绕“对象存储”构建面向不同场景和需求的数据分析服务

为什么是对象存储

环顾云上存储

- 云盘：成本和高吞吐低时延的矛盾
- 本地盘：计算和存储资源耦合，较高的成本和运维周期
- 文件存储服务：面向传统存储领域，不适用于大数据

对象存储

- 大容量和高带宽：提供PB级容量和TB级带宽和ms级别的时延
- 元数据管理能力：可以管理百亿级别的对象
- 面向多场景：支持多协议，对象存储rest协议，POSIX协议，HDFS协议
- 数据管控：存储任意非结构化数据，便于统一存储和安全管控
- 多级存储和EC策略：有效降低TCO

存算分离大数据实践

解决方案	实践
计算架构灵活	• 单集群：在单集群算力不足时可以轻易的扩容，无需过多顾虑存储配置和成本 • 多集群：多集群隔离，互不影响
桶命名空间隔离	• 通过桶级别的命名空间进行业务隔离 • 通过桶级别的命名空间进行访问控制 • 通过桶级别的命名空间进行容量带宽控制 • 通过桶级别的命名空间方便“存储扩容”
数据权限和审计	• 细粒度的权限控制：可以精细到访问接口和对象级别 • 审计：通过桶级别的审计日志满足审计要求
数据保护	• 安全的秘钥管理服务 • 支持多种加密算法：满足不同场景的数据安全要求
数据可靠性	• 3AZ特性支撑数据的高可靠性 • AZ之间具有独立的风火水电，物理隔离，故障隔离
成本治理	• 灵活的生命周期策略：定期转低频，归档，删除；有效的降低存储成本 • 灵活的存储策略：标准存储，低频访问存储，归档存储，适应不同的业务和数据生命周期场景

组件优化	含义
Hive权限优化	
hive.metastore.pre.event.listeners	*Because the object store does not have the concept of directory permission, set hive metastore. pre. event. Listeners to reduce the number of directory permission checks in the object store*
Hive读优化	
hive.exec.input.listing.max.threads	*Sets the maximum number of threads that Hive uses to list input files. Increasing this value can improve performance when there are many partitions being read.*
mapreduce.input.fileinputformat.list-status.num-threads	*Sets the number of threads used by the FileInputFormat class when listing and fetching block locations for the specified input paths.*
Hive写优化	
hive.mv.files.threads	*Sets the number of threads used to move files in a move task. Increasing the value of this parameter increases the number of parallel copies that can run on object storage.* *A separate thread pool is used for each Hive query. When running only a few queries in parallel, you can increase this parameter for greater per-query write throughput. However, when you run a large number of queries in parallel, decrease this parameter to avoid thread exhaustion.* *To disable multi-threaded file moves, set this parameter to 0. This can prevent thread contention on HiveServer2.* *This parameter also controls renames on HDFS, so increasing this value increases the number of threads responsible for renaming files on HDFS.*
Hive元数据优化	
hive.metastore.fshandler.threads	*Sets the number of threads that the Hive metastore uses when adding partitions in bulk to the metastore. Each thread performs metadata operations for each partition added, such as collecting statistics for the partition or checking if the partition directory exists.* *This parameter is also used to control the size of the thread pool that is used by MSCK whe4n it scans the file system looking for directories that correspond to table partitions. Each thread performs a list status on*

XXX客户TPCDS ORC 500GB基准测试query4优化：

```
2020-07-09 11:27:35,962 | INFO  | 3e40aa12-4077-485a-a8b6-05e3b50ceb3a HiveServer2-Handler-Pool: Thread-364 | ObsClient [getObjectMetadata] cost 8 ms | sun.refle
2020-07-09 11:27:35,962 | INFO  | 3e40aa12-4077-485a-a8b6-05e3b50ceb3a HiveServer2-Handler-Pool: Thread-364 | 2020-07-09 11:27:35 961|com.obs.services.internal.R
2020-07-09 11:27:35 961|com.obs.services.internal.RestStorageService|performRequest|818|Storage|1|HTTP+XML|performRequest||||2020-07-09 11:27:35|2020-07-09 11:27
2020-07-09 11:27:35 962|com.obs.services.ObsClient|doActionWithResult|2824|Storage|1|HTTP+XML|getObjectMetadata||||2020-07-09 11:27:35|2020-07-09 11:27:35|||0|
2020-07-09 11:27:35 962|com.obs.services.ObsClient|doActionWithResult|2827|ObsClient [getObjectMetadata] cost 8 ms

2020-07-09 11:27:50,767 | INFO  | 3e40aa12-4077-485a-a8b6-05e3b50ceb3a HiveServer2-Handler-Pool: Thread-364 | Cache Content Summary for obs://
2020-07-09 11:27:50,767 | INFO  | 3e40aa12-4077-485a-a8b6-05e3b50ceb3a HiveServer2-Handler-Pool: Thread-364 | Cache Content Summary for obs://
```

hive.exec.input.listing.max.threads：调整为20，耗时将从504s降低为441s

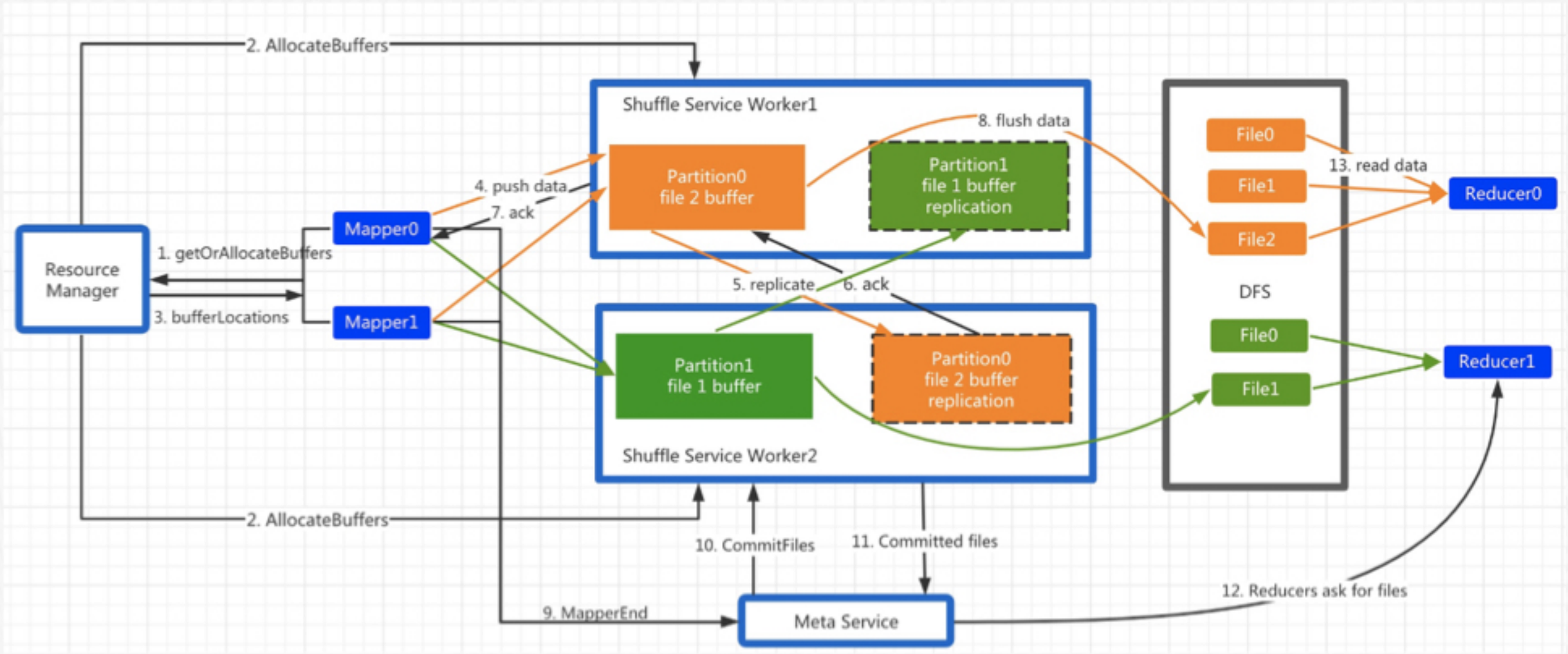
```
Line 1: 2020-07-10 11:14:50,703 | INFO  | Thread-100 | 2020-07-10 11:14:50 702|com.obs.services.internal.RestStorageService|performRequest|671|OkHttp cost 6
Line 6: 2020-07-10 11:14:50,707 | INFO  | Thread-100 | OkHttp cost 4 ms to apply http request | sun.reflect.GeneratedMethodAccessor33.invoke(Unknown Source)
Line 7: 2020-07-10 11:14:50,707 | INFO  | Thread-100 | Storage|1|HTTP+XML|performRequest||||2020-07-10 11:14:50|2020-07-10 11:14:50|||[responseCode: 200][requ

Line 62180: 2020-07-10 11:15:44,520 | INFO  | Thread-100 | Storage|1|HTTP+XML|listObjects||||2020-07-10 11:15:44|2020-07-10 11:15:44|||0| | sun.reflect.Gener
Line 62181: 2020-07-10 11:15:44,520 | INFO  | Thread-100 | ObsClient [listObjects] cost 11 ms | sun.reflect.GeneratedMethodAccessor33.invoke(Unknown Source)
Line 62182: 2020-07-10 11:15:44,520 | INFO  | Thread-100 | 2020-07-10 11:15:44 517|com.obs.services.internal.RestStorageService|performRequest|671|OkHttp cos
Line 62187: 2020-07-10 11:15:44,520 | INFO  | Thread-100 | Total input files to process : 5482 | org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listSt
```

mapreduce.input.fileinputformat.list-status.num-threads：调整为20，耗时将从441s降低为293s

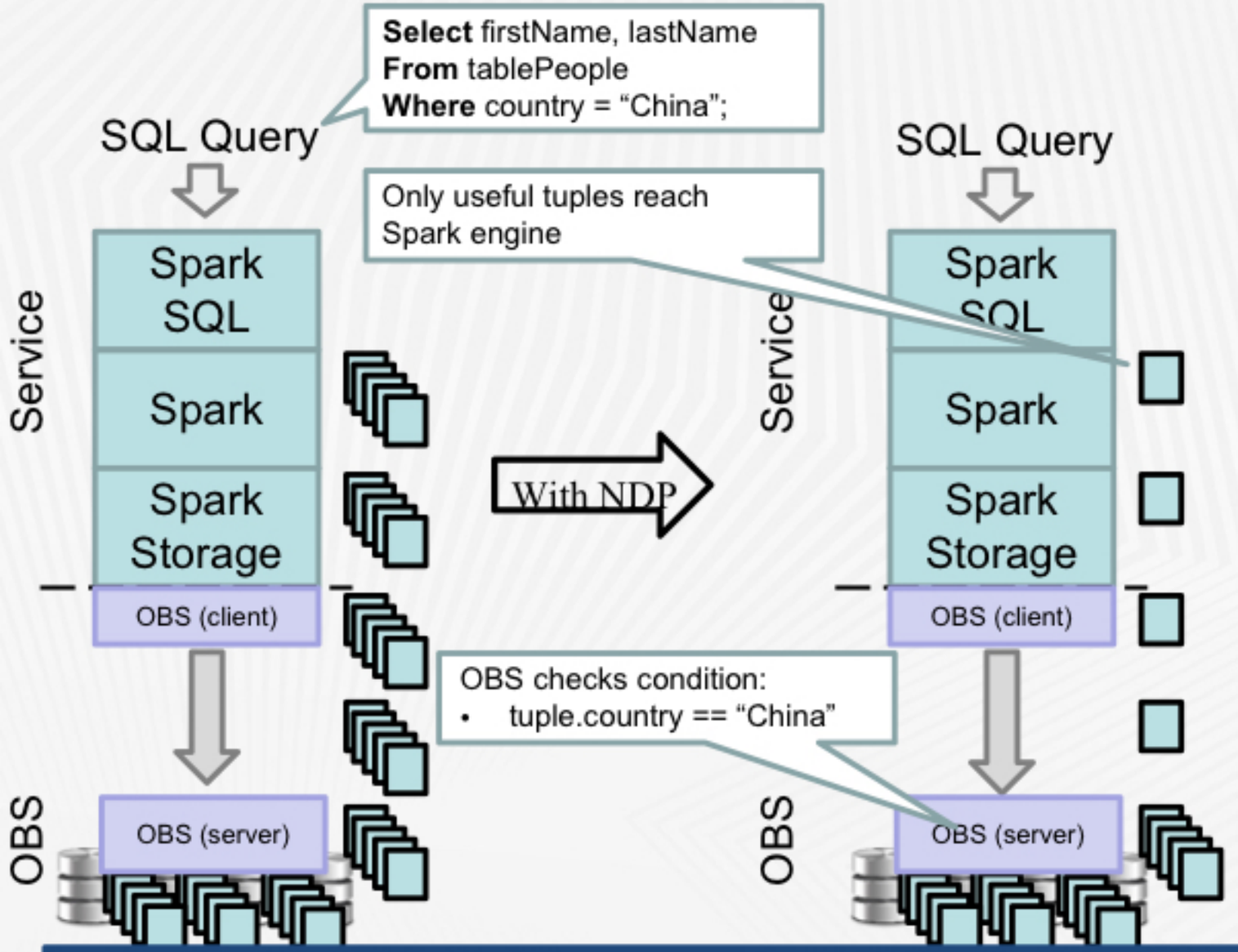
存算分离发展-缓存

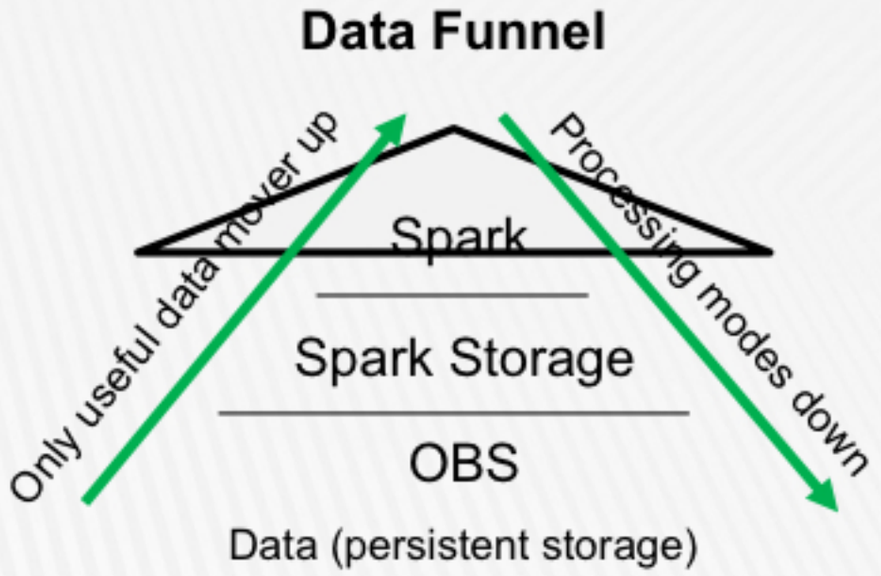

Presto
Spark
MapReduce
MapReduce
HIVE
TensorFlow
PyTorch
HDFS Interface
Java File API
POSIX Interface
S3 Interface
REST API
ALLUXIO
DATA ORCHESTRATION FOR ANALYTICS AND AI
HDFS Driver
S3 Driver
GCS Driver
Azure Driver
NFS Driver
AWS S3
GC Storage
Azure
HDFS
ceph
MinIO
IBM
cleversafe
Net App
DELL EMC
ECS
HITACHI
Cloudian

存算分离发展-remoteShuffle



存算分离发展-近数据计算





Specialization ✗ + Abstract API ✗ =

Implication: Melts traditional rigid boundaries of storage and processing

NDP technology pushes computation down to move only useful data up

成就一亿技术人

成为技术人交流和成长的家园

用户为本 | 求真求是 | 协作共赢 | 极客精神 | 结果导向

CSDN