Hadoop存储与计算分离实践

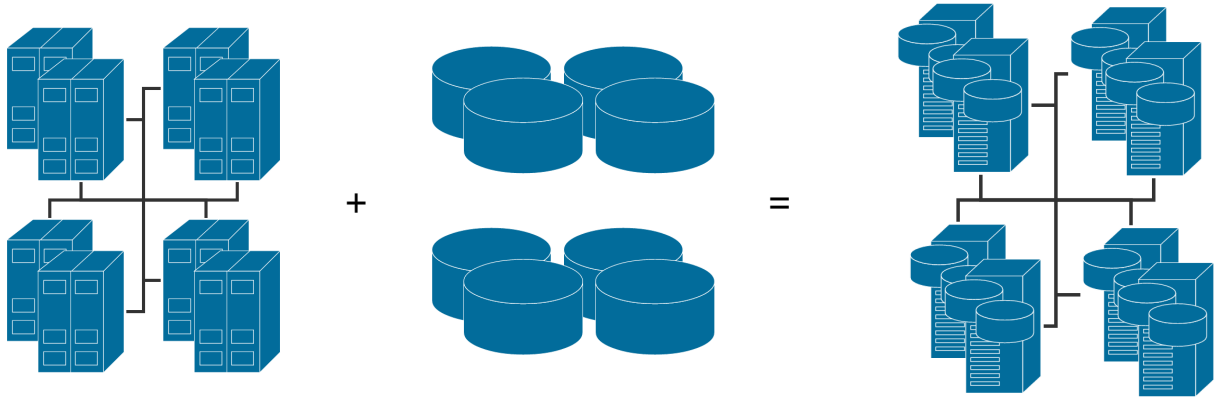
余根茂
阿里云E-MapReduce团队

主办单位：杭州 Alibaba Group 阿里巴巴集团
战略合作伙伴：intel

- 传统集群部署实践
- 云上集群部署实践

传统集群部署实践

■ 存储和计算



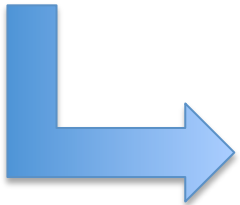
计算能力　　存储能力　　集群能力

■ 数据“中心”

本地磁盘比网络传输快

任务处理中数据获取开销大

数据本地性

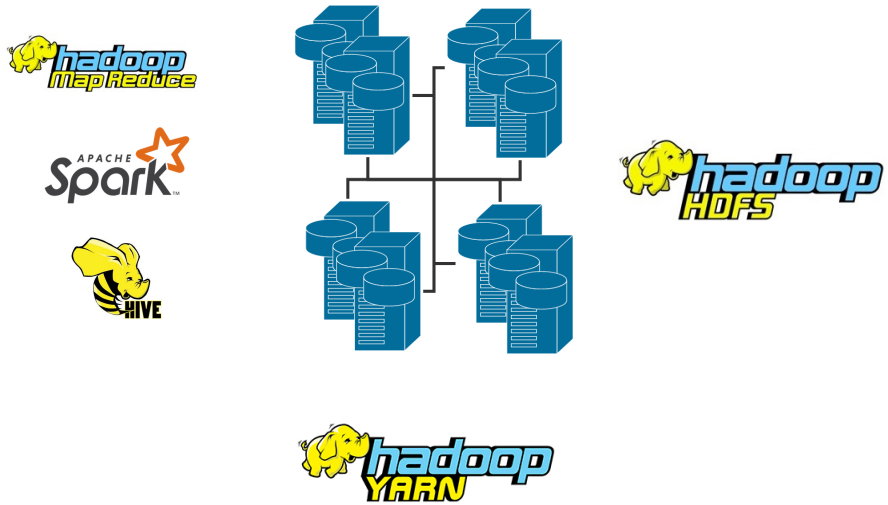


计算找数据

以数据为中心

扫码观看大会视频

■ 集群混部



■ 理想

更少的数据迁移

更高的资源利用率

现实

带宽逐渐不是稀缺资源

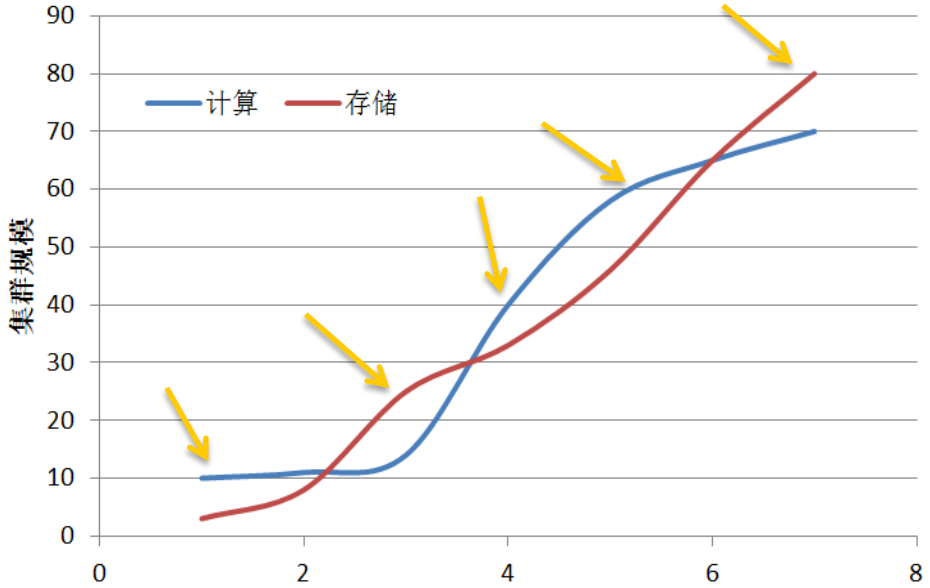
网络	1Gbps	10Gbps~20Gbps	↑ 10~100
内存	12g	96g~192g	↑ 8~16
磁盘	800Mbps	1200Mbps	↑ 1~2
	2009	2016	

磁盘不再是承载计算的主战场

■ 现实

集群的木桶效应

集群资源浪费



■ 现实

Data Locality vs. Remote Data

混部的劣势

更多的集群资源浪费

更差的集群扩展性

不再万能的Data Locality

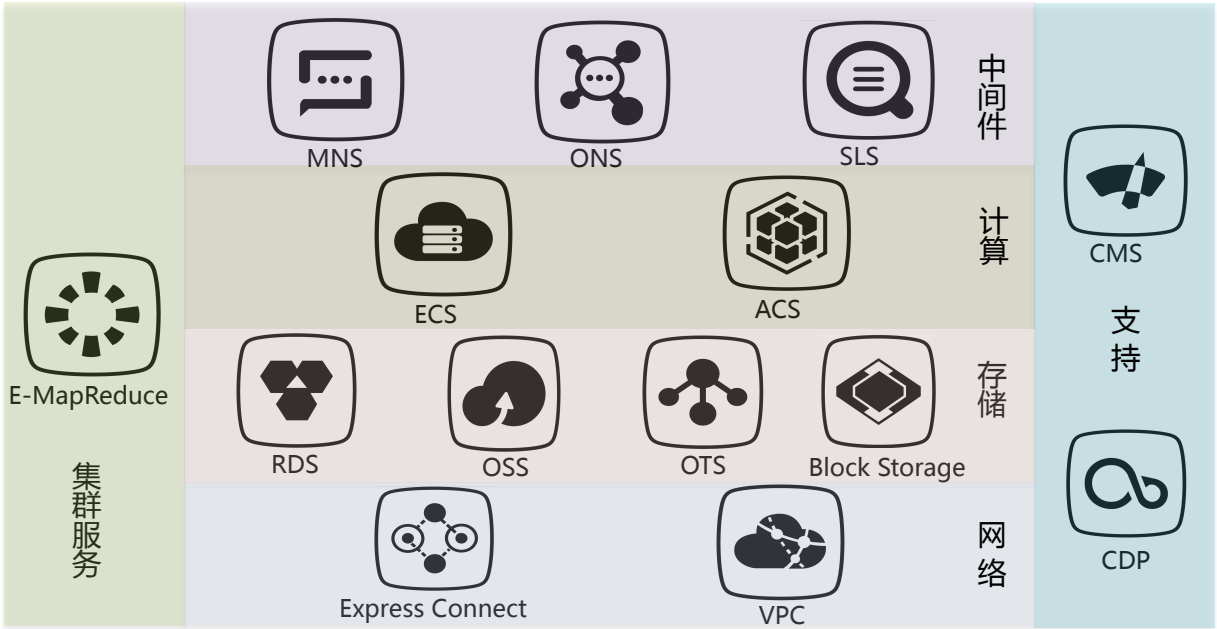
混合部署的合理性？

扫码观看大会视频

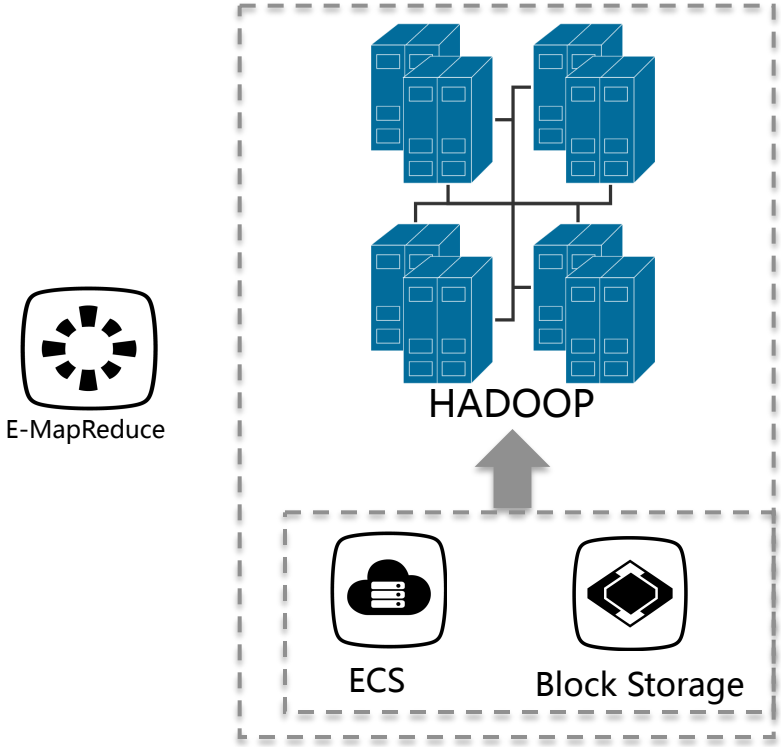
云上集群部署实践

扫码观看大会视频

云计算基础设施

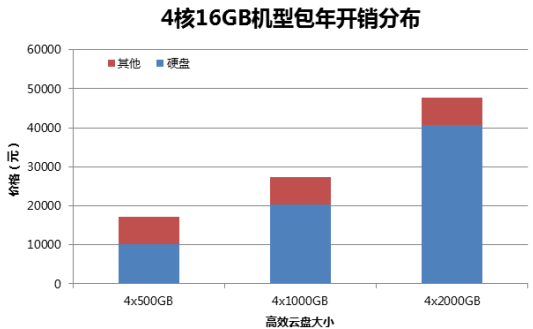


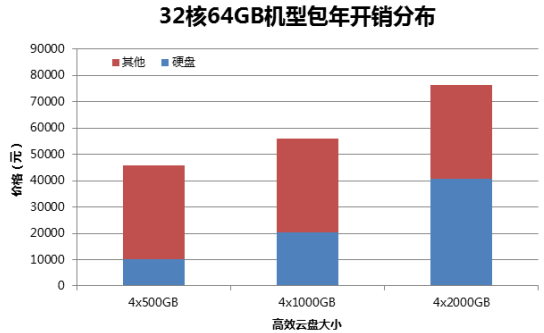
■ 云上集群部署

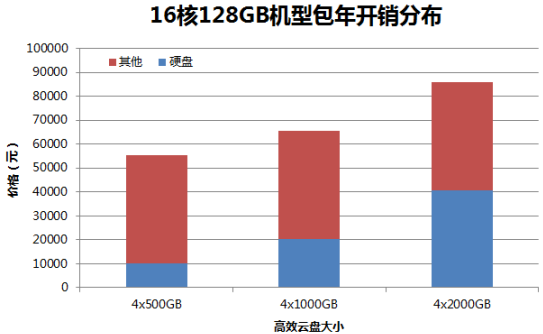


一键部署，即开即用

新的挑战



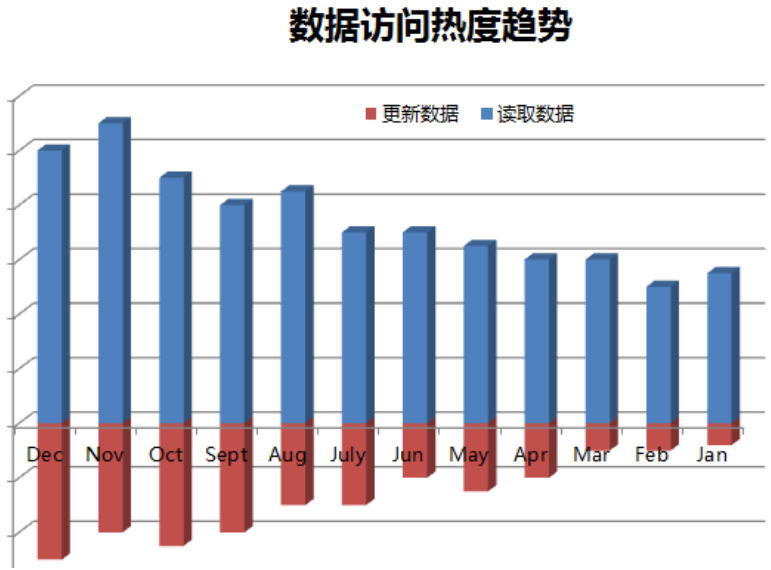




存储成本比例较高

* 采用2016.10.9当天价格

新的挑战



热数据逐渐变冷

冷数据逐渐堆积

存储质量下降

数据Balance代价上升

新的挑战

冷数据积累

磁盘存储费用高

数据存储成本高昂

集群维护成本

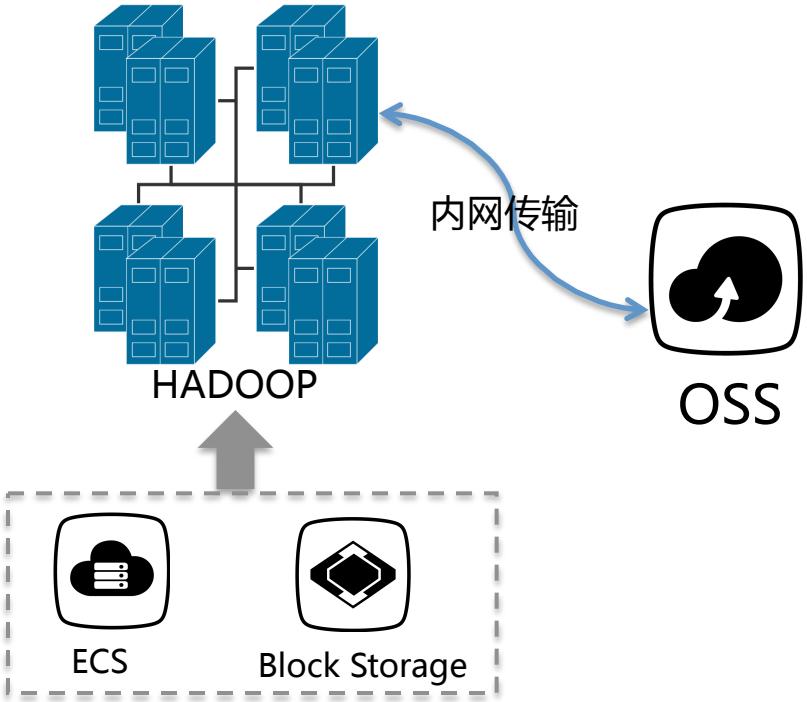
数据服务可用性要求高

…

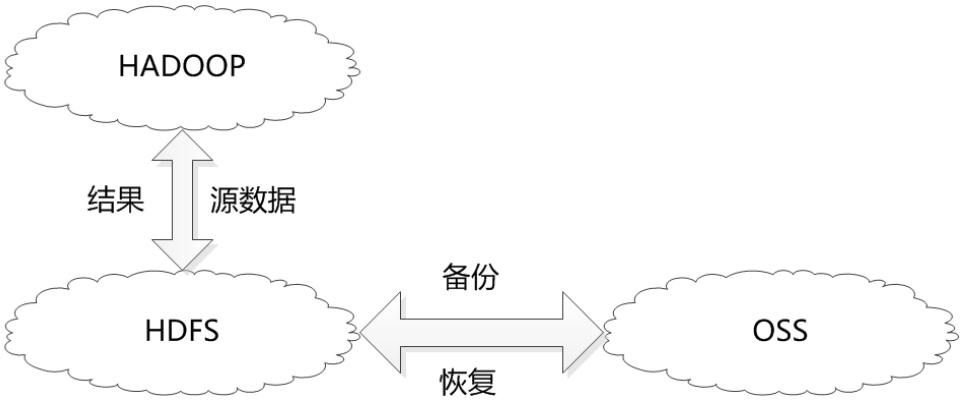
替代方案？

Hadoop + OSS

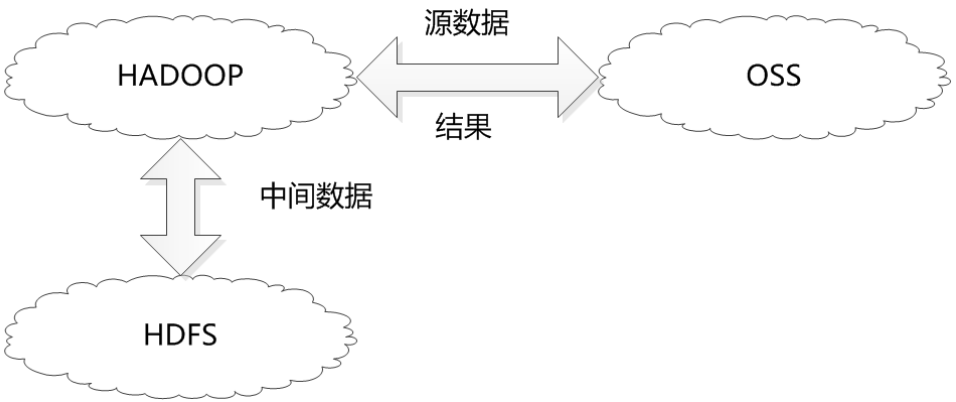
基于OSS的分离部署



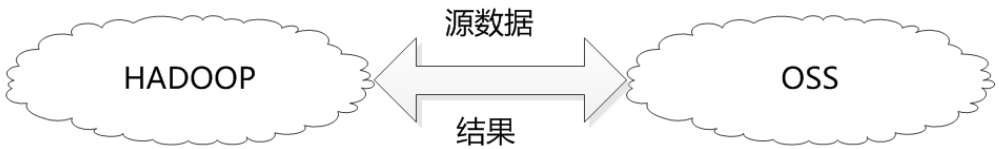
OSS数据使用方式



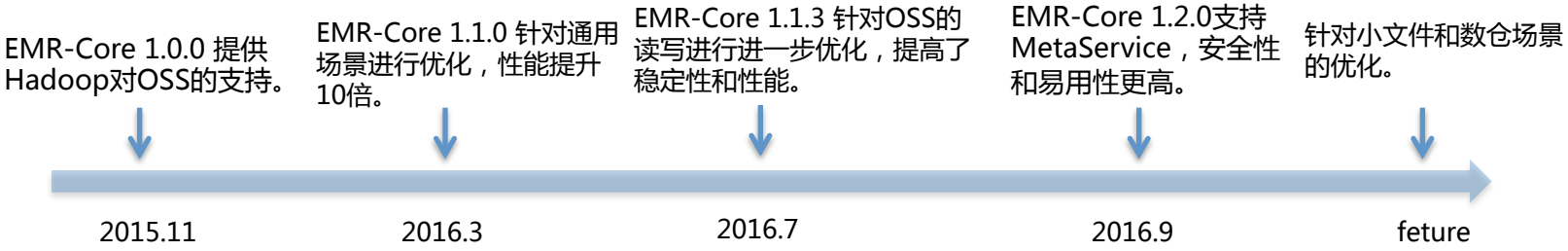
OSS数据使用方式



OSS数据使用方式



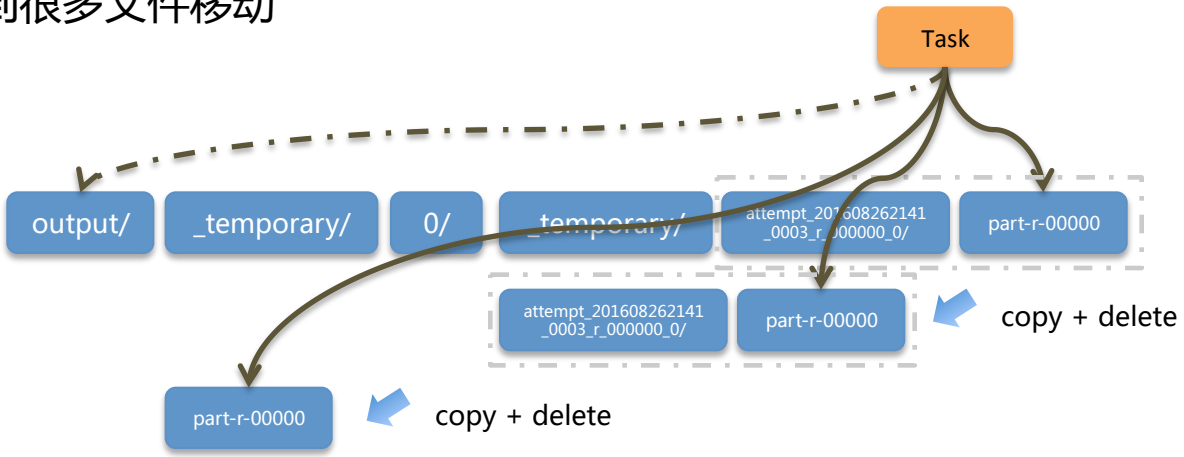
Hadoop对OSS支持演进



针对性优化

文件移动对OSS来说是比较重的操作

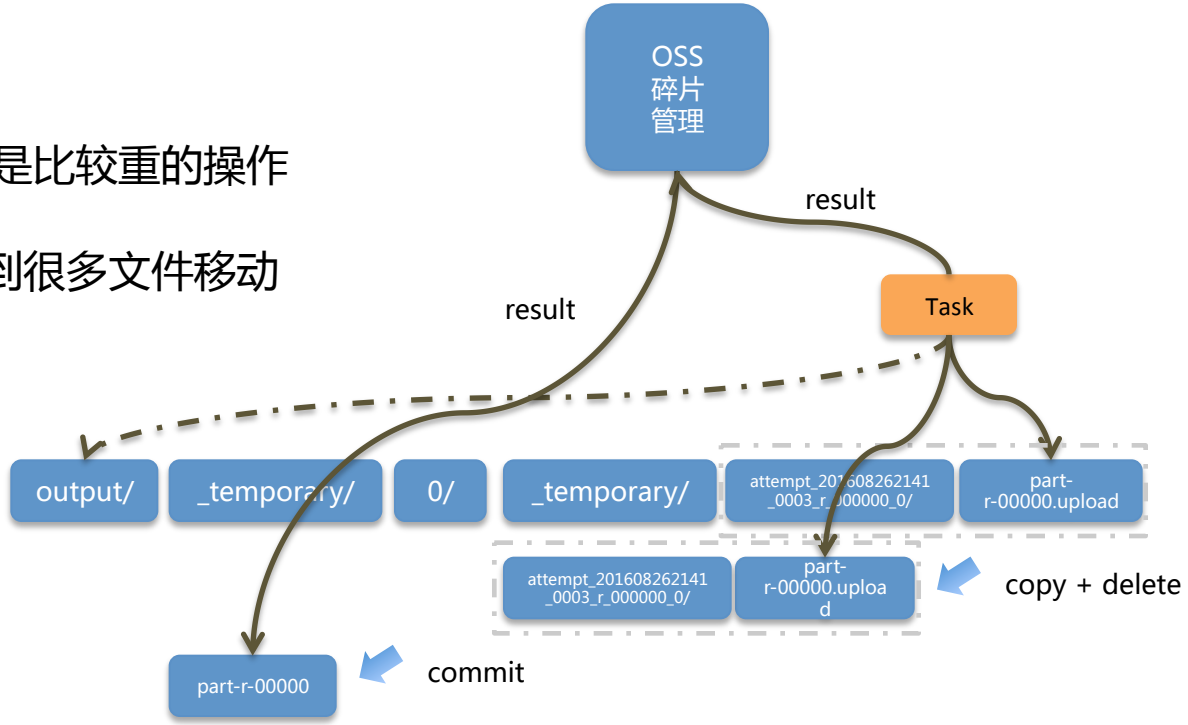
Hadoop执行中涉及到很多文件移动



针对性优化

文件移动对OSS来说是比较重的操作

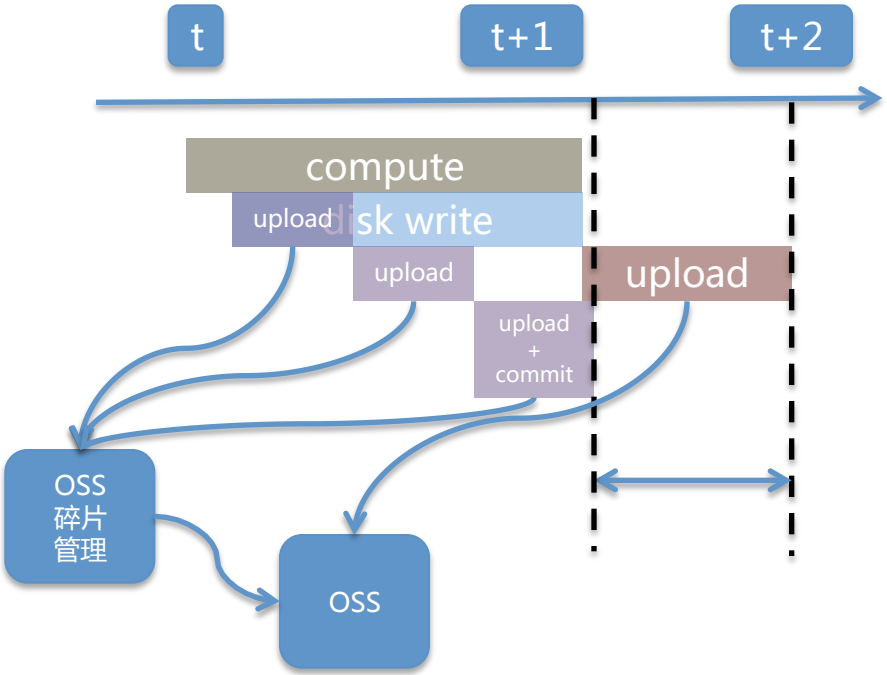
Hadoop执行中涉及到很多文件移动



■ 针对性优化

结果数据先写本地再上传OSS

结果数据边计算边上传OSS



针对性优化-未来

- 小文件预取和缓存
- 元数据视图系统
- ...

On the way

■ 成本和性能

成本能节省多少？

性能能达到要求？

性能测试

	IO	TeraGen/TeraSort
测试工具	Hadoop DFSIO	Spark版本
对比指标	$Thoughput\,(N) = \frac{\sum_{i=0}^{N} filesize_i}{\sum_{i-0}^{N} time_i}$ $Average\ IO\ rate(N) = \frac{\sum_{i=0}^{N} rate_i}{N} = \frac{\sum_{i=0}^{N} \frac{filesize_i}{time_i}}{N}$	时间
测试组	800 x 1KB（1MB，100MB，10GB）	10GB，100GB，500GB

性能测试

测试平台	E-MapReduce
Region	华东1 可用区B
网络环境	VPC网络
集群规模	1 主节点，8 从节点
机器配置	4核16G 机型II，4*200GB高效云盘
镜像版本	EMR-2.1.0

■ 成本对比

$$Cost_{storage} = Time \times Price$$

场景模拟：

- 一次500GB TeraSort排序的存储成本
- 一次年度采购中的存储成本

成本对比

一次排序测试的成本

	Time	Price	Cost(storage)
Hadoop+OSS	7364s	0.000148/GB/h	0.151元 ↓ 81.7%
Hadoop+HDFS	5106s	0.000582/GB/h	0.825元

* 以上为500GB TeraSort测试数据
* 云盘按照包年价格折算
* OSS按照包年价格折算
* HDFS采用2备份配置

■ 成本对比

一次年度采购中的数据存储成本

数据量	5T	10T	20T
HDFS	52224	104448	208896
OSS	6480 ↓87.6%	12924 ↓87.6%	25848 ↓87.6%
HDFS+OSS（热/冷=4/6）	24777.6 ↓52.6%	49555.2 ↓52.6%	99074.4 ↓52.6%

* 以上按照对应产品的包年价格
* HDFS采用2备份配置
* 未计入OSS的接口调用费用

如何选择

$$Cost_{storage} = Time \uparrow \times Price \downarrow$$

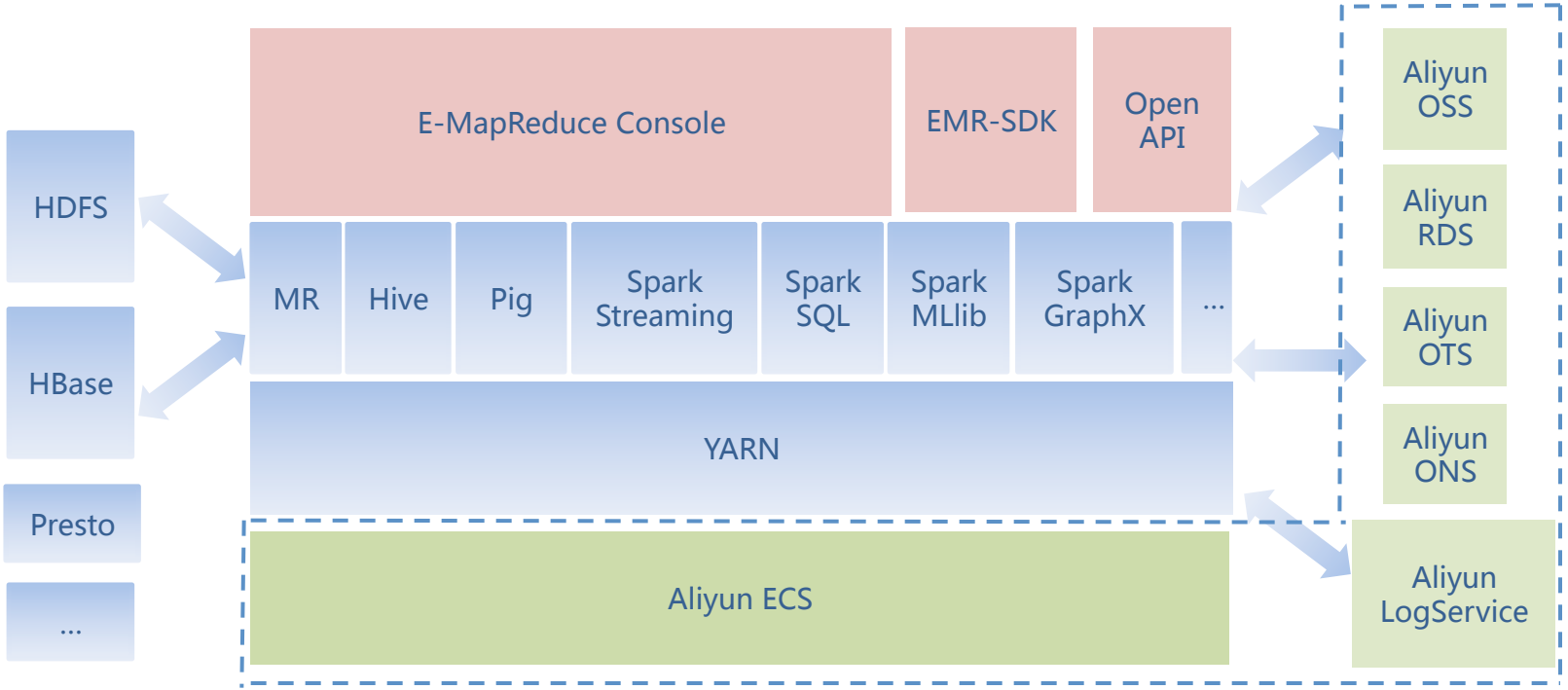
	成本	性能
HDFS	高	高
OSS	低	较高

合理组合才能带来性能和成本的双赢

E-MapReduce

E-MapReduce（Elastic MapReduce）是构建于阿里云ECS虚拟机之上，结合开源生态系统，为用户提供集群，作业和执行计划等管理的一站式大数据处理分析服务。

E-MapReduce


E-MapReduce Console
EMR-SDK
Open API
HDFS
HBase
Presto
...
MR
Hive
Pig
Spark Streaming
Spark SQL
Spark MLlib
Spark GraphX
...
YARN
Aliyun ECS
Aliyun OSS
Aliyun RDS
Aliyun OTS
Aliyun ONS
Aliyun LogService

欢迎大家使用E-MapReduce

2016 The Computing Conference

THANKS

2016 杭州·云栖大会
THE COMPUTING CONFERENCE