

Hadoop跨机房架构实践

陈昱康



陈昱康

- 携程大数据平台基础架构负责人
- 曾就职于大众点评，阿里云
- 16年加入携程，负责Hadoop, Spark研发

目录

1

跨机房项目背景

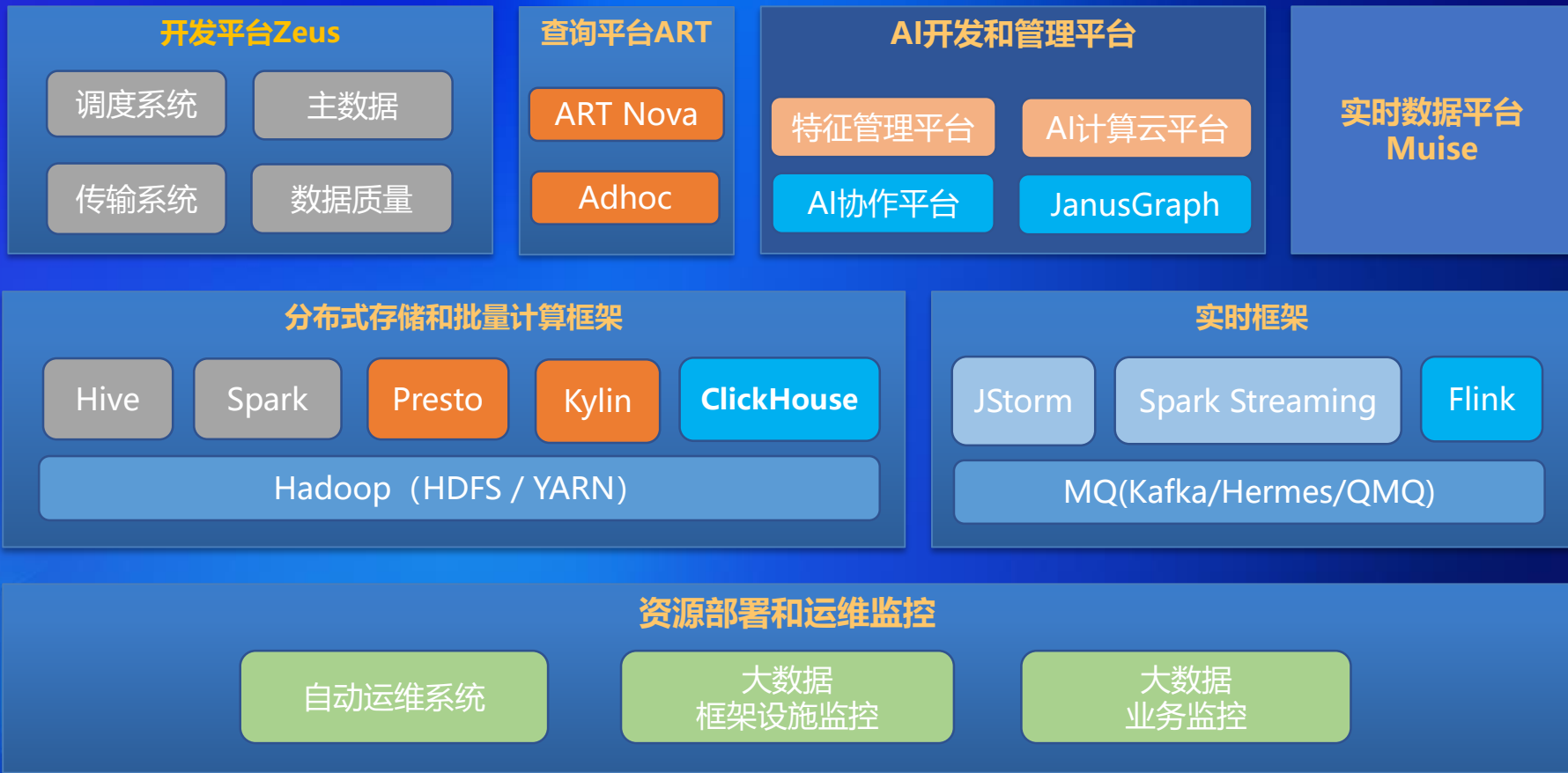
2

跨机房项目落地实践

3

总结和展望

携程大数据平台架构



Hadoop在携程

- 2014年落地
- 数据量每年以两倍速度增长
- 2019年
 - 近200PB, 2000+计算存储节点, 4个ns, nnproxy, 冷热存储分离, 一套EC集群
 - 两套计算集群, 近12w vcore, 每天30w Hadoop job, 90% spark
 - 四个机房, 在线离线混部

项目背景

- 四个机房，机房搬迁
- 预计到2024年底集群规模将达到万台
- 单机房机架数物理瓶颈
- 跨机房带宽仅200Gbps
- 同地域网络延迟约1ms(带宽打满后，延迟10ms，丢包率10%)




原生Hadoop架构

- Shuffle读写
- hdfs读
- hdfs pipeline写






可选方案

多机房多集群

- 不需要改代码 
- 对用户不透明
- 运维成本高
- 数据一致性难以保证

多机房单集群

- 需要改Hadoop core
- 对用户透明 
- 运维简单 
- 保证数据一致性 

在线离线混部跨机房

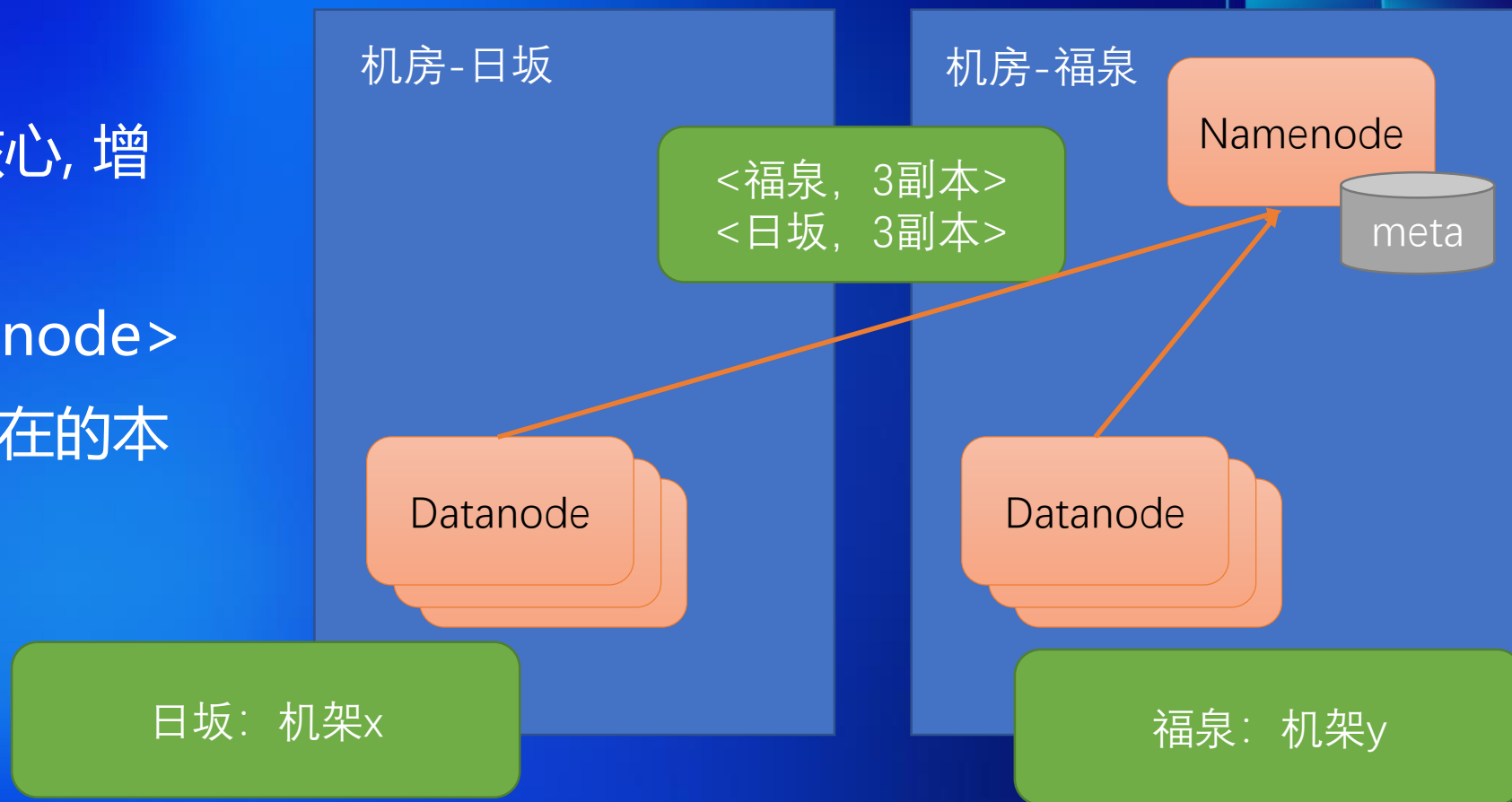
- 在线离线跨机房
- 多机房多集群，在线服务部署独立Yarn集群
- 通过作业的资源画像，分配低shuffle量和hdfs读写量的作业到在线机房Yarn集群
- 动态反压
- 对用户透明
- Yarn支持基于label调度，一个机房一个label，禁止shuffle
- 缓解集群计算压力百分之八

多机房单集群方案

- 多机房单HDFS集群
- 多机房多Yarn集群
- 自动化迁移工具
- 跨机房带宽监控&限流

多机房单HDFS架构

- Datanode拓扑
 - 改造namenode核心, 增加机房感知
 - <机房, 机架, Datanode>
 - 优先读写客户端所在的本地机房副本



多机房单HDFS架构

- 跨机房多副本管理
 - 增加设置目录多机房副本命令
 - <文件路径, 机房, 副本数>
 - 按hive账号设置默认机房
- namenode一份元数据管理多机房副本
 - 数据一致性
 - 增加Editlog Op, 写到fsimage, 无外部依赖, nn切换后不变

```
-bash-4.1$ hadoop fs -setremoterep -s -w -R crossdcconfig idc1=3,idc3=3 op MERGE_R hdfs://ns3/tmp/hadooptest
Try to call setRemoteReplication.Path:hdfs://ns3/tmp/hadooptest,crossdcconfig:[CrossIdcConfigEntry{idcName='idc1',
replication=3}, CrossIdcConfigEntry{idcName='idc3', replication=3}],op:MERGE_R
setRemoteReplication Success.Result:true
```

Balancer & Mover & EC

- Balancer
 - 支持多实例
 - 增加IP范围列表，每个机房起一个，只balance本机房的数据
- Mover
 - 支持多实例
 - 按照跨机房的副本策略move数据到archive节点
 - Proxy和target节点尽量保证在同一机房
- EC
 - Hadoop 3.0，不跨机房，只部署在新机房

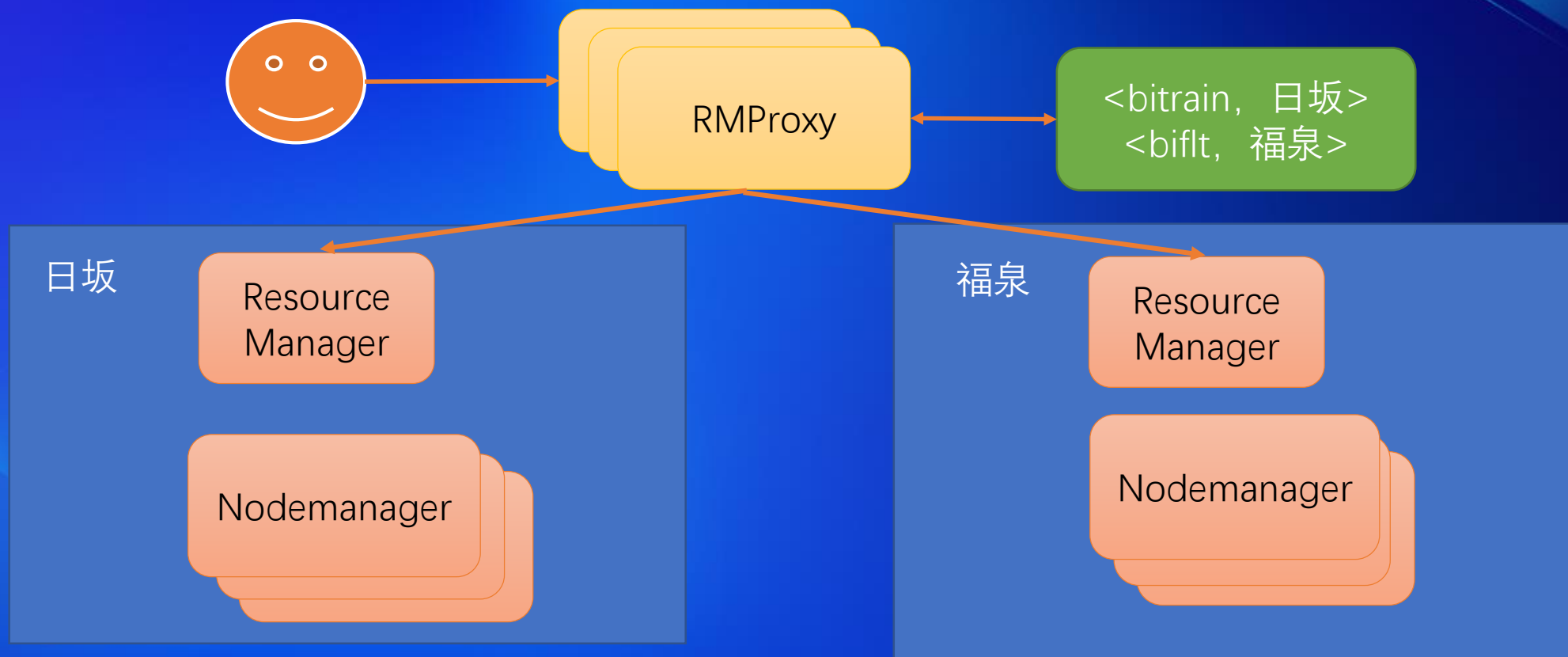
Cross Fsck

- 跨机房HDFS版本按namespace灰度上线
- 感知跨机房配置策略，修正不正确放置的副本
- 大量getBlockLocations rpc，从standby nn读
- 错误副本调用reportBadBlocks rpc，block manager删除错误副本，重新选择新的副本
- 限流

多机房多Yarn集群

- 基于RM Proxy和Yarn federation的跨机房调度
 - 作业同机房内调度, 禁止跨机房shuffle
 - 用户机房mapping管理&持久化, eg. <bitrain, 日坂>
 - 一键账号对应机房切换
 - 收拢Hadoop客户端, 作业提交统一走rm proxy
 - 降级策略, 本地cache

多机房多Yarn集群



Spark/Hive service/Presto

- 每个机房各部署一套服务
- 客户端改造接入rmproxy, 建立jdbc连接前, 先从rmproxy中拿到该用户对应机房的服务链接URL, 再连接
- 对用户透明

自动化迁移工具

- 做到自动化迁移，迁移流程常态化
- 按照BU->账号粒度进行迁移



自动化迁移工具

Hive账号

库权限

表权限

权限查询

Hive账号申请

Hive Owner变更

申请记录

审批

OPS-DI审批

账号回收

NSMapping

添加

紧急停止迁移任务

添加白名单

查看白名单

HiveAccount Ns Mappings

CSV

Hive账号跨数据中心迁移

DB 迁移状态

Hive账号	Path	src	dest	名称	容量(TB)	类型	状态	更新时间
bitrain	hdfs://ns/	fq	rb		3.09	0:3迁移	迁移成功	2019-05-20 12:58
bitrain	hdfs://ns/	fq	rb		439.54	0:3迁移	迁移成功	2019-05-20 18:06
bitrain	hdfs://ns3/	fq	rb	用户Path	121.18	0:3迁移	迁移成功	2019-05-21 15:16
bitrain	hdfs://ns/	fq	rb		0	0:3迁移	迁移成功	2019-05-20 12:58
bitrain	hdfs://ns/	fq	rb		1.53	0:3迁移	迁移成功	2019-05-20 12:58
bitrain	hdfs://ns/	fq	rb		13.85	0:3迁移	迁移成功	2019-05-20 12:58
bitrain	hdfs://ns/	fq	rb		21.81	0:3迁移	迁移成功	2019-05-20 12:58

关闭

	ns3	rb	fq	rb	0:3 成功	修改删除0:3迁移迁移进度	陈昱康	2019-05-22 11:
	ns	fq	fq	rb	3:3 成功	修改删除0:3迁移迁移进度	系统	2019-02-27 13:
	ns	fq	fq	rb	3:3 成功	修改删除0:3迁移迁移进度	系统	2019-02-27 13:
	ns	fq	fq	rb	3:3 成功	修改删除0:3迁移迁移进度	系统	2019-02-27 13:
	ns	fq	fq	rb	3:3 RUNNING	修改删除迁移进度	系统	2019-02-27 13:
	ns	fq	fq	rb	等待3:3调度	修改删除迁移进度	系统	2019-02-27 13:
	ns	fq	fq	rb	等待3:3调度	修改删除迁移进度	系统	2019-02-27 13:

1 2 3 4 5 20

自动化迁移工具

- 实践中的注意点
 - 集群低峰时间执行
 - 控制迁移速率，实时监控nn的UnderReplicatedBlocks和跨机房流量 metrics
 - 实时监控被迁移机房的hdfs可用容量，包括不同的storage type
 - 公共目录可单独设置多机房副本

跨机房带宽监控&限流

- 读写block/文件流量监控
 - dfsclient, datanode埋点实时汇报
 - dfsclient remote block read、data streamer写
 - datanode, balance, block recovery/transfer
 - 路径, blockId, 读写大小, 类型, pipeline, 优先级, zeusid等

跨机房带宽监控&限流

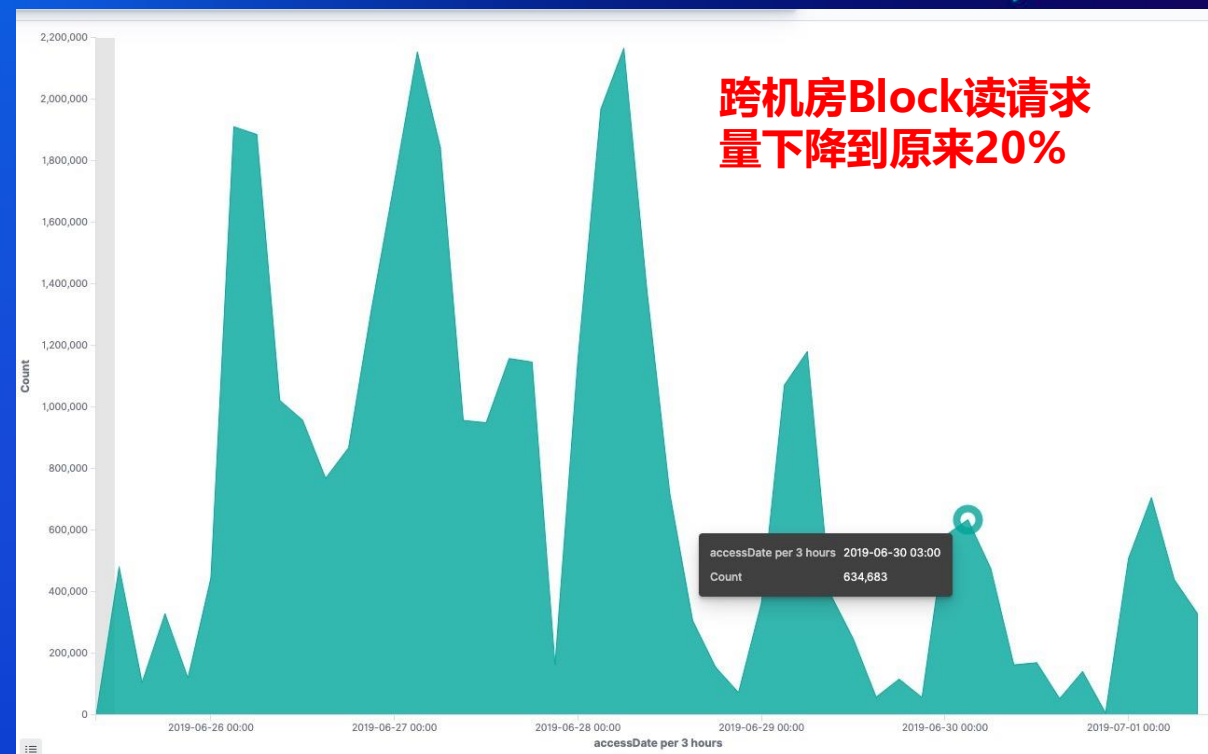


- 流控服务

- 基于Permit(上下行), 客户端获得后才可继续执行, 否则sleep
- 优先级排序
- 实时监控Hickwall网络流量作为开启条件

- 流量分析和副本调整

- 流量数据实时进ES和HDFS, 实时分析流量
- 按需调整公共目录的跨机房副本数



效果



总结和未来规划

- 总结

- 实现单hdfs集群机房感知功能，跨机房副本设置
- 实现基于rm proxy和yarn federation的计算调度
- 实时自动化存储和计算迁移工具
- 实现跨机房流量监控和限流服务

- 未来规划

- 智能决定迁移哪些账号
- 智能公共路径跨机房副本设置和回收
- Hadoop 3支持跨机房

Thanks For Watching



本PPT来自2019携程技术峰会
更多信息请关注“携程技术中心”微信公众号~