BDTC 2017 中国大数据技术大会
Big Data Technology Conference 2017

阿里云大规模结构化云存储HBase架构解析

封神　阿里云高级技术专家

曹龙(封神)

阿里云技术专家、架构师

专注在大数据领域，有6年分布式引擎研发经验

先后研发上万台Hadoop、ODPS集群

先后负责阿里YARN、spark及自主研发内存计算引擎

目前为广大公共云用户提供专业的云Hadoop服务及云HBase服务

提纲

- 阿里大数据三大组件
- 云　　最佳实践
- 云　　部署模式
- 云　　真实案例
- 云　　内核特性
- 云　　未来

阿里大数据三大件


70年代
凤凰牌、永久牌自行车；梅花牌、上海牌手表；
蝴蝶牌缝纫机，70年代结婚有了这几样
简直就可以称得上世纪婚礼了
自行车
手表
缝纫机

阿里巴巴集团大数据三大件

组件	内部规模	公有云产品	主要功能
ODPS	7w	MaxCompute	离线计算&机器学习
HBase	1.2+w	云HBase	实时数仓&在线存储&实时更新查询
Flink(Blink)	数千	StreamCompute	实时计算

阿里巴巴HBase数百个集群 从4台 到 2000台

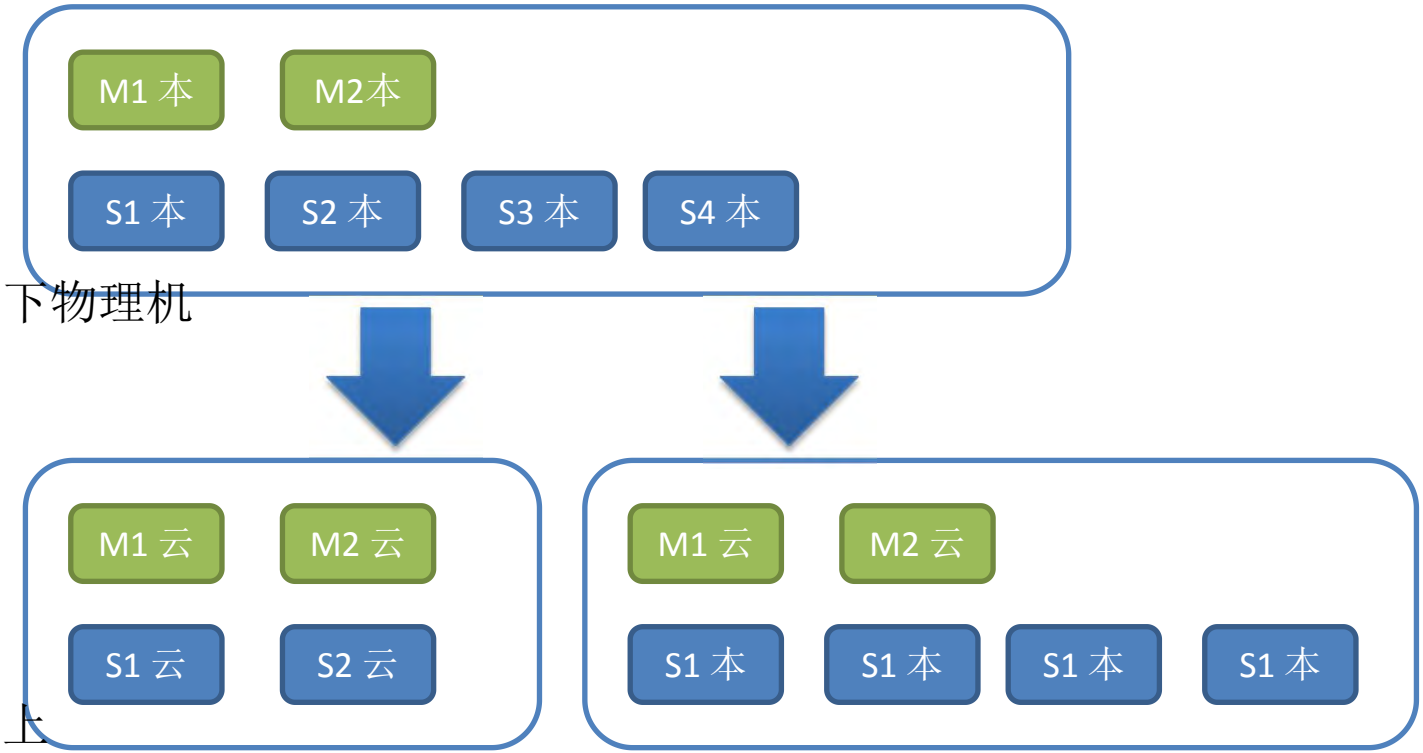
单集群数据从 几百G 到 10P

阿里巴巴及业界HBase使用场景

场景	搭配	核心需求
机器学习	Spark	存储特征等，高纬度，强调吞吐量
大数据风控	ODPS	存储用户画像等热数据信息 对QPS要求比较高，从离线导入数据速度
消息订单物联网数据 等存储	ECS	对SLA、延迟(99.9)有要求 数据量巨大，不断增多，成本比较看重
物联网 时序数据	HiTSDB OpenTSDB	时序场景，写多读少，写速度快，成本敏感
图数据库	JanusGraph	图的需求，并发高
多维分析	Kylin、Hadoop MR	存cube，需要离线搭配
操作型分析	Phoenix	倒排索引，HBase端本地分析
检索	Elasticsearch\solr	数据实时写入，HBase存放原始数据，ES/slor存索引数据

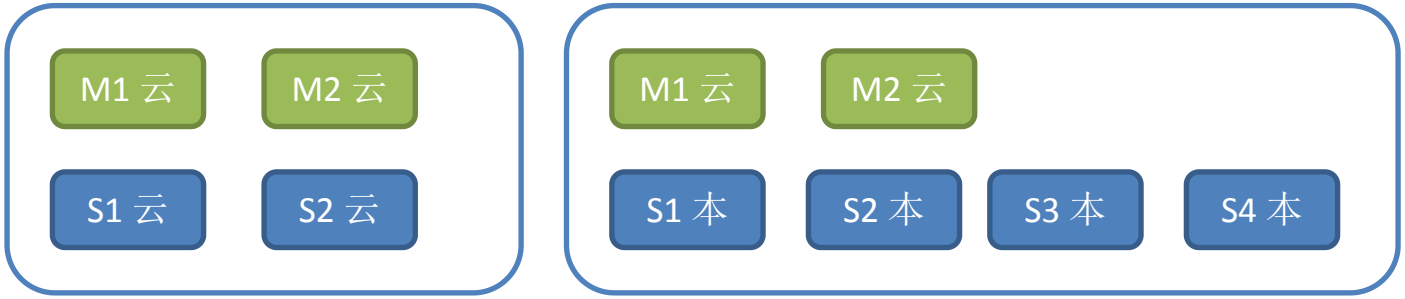
部署模式

云HBase基础架构 – 基于VM



	Master	系统盘	模式 灵活性
线下	规模大，一般空闲较多	物理磁盘，有损坏风险，且占用一个磁盘	起步较高，往往会浪费空间
云	规格可以很小	使用云磁盘，没有损坏风险，不占数据盘	有云盘模式及本盘模式，大小通吃，可以不同业务级别业务不同集群。还可以弹性加减少节点。

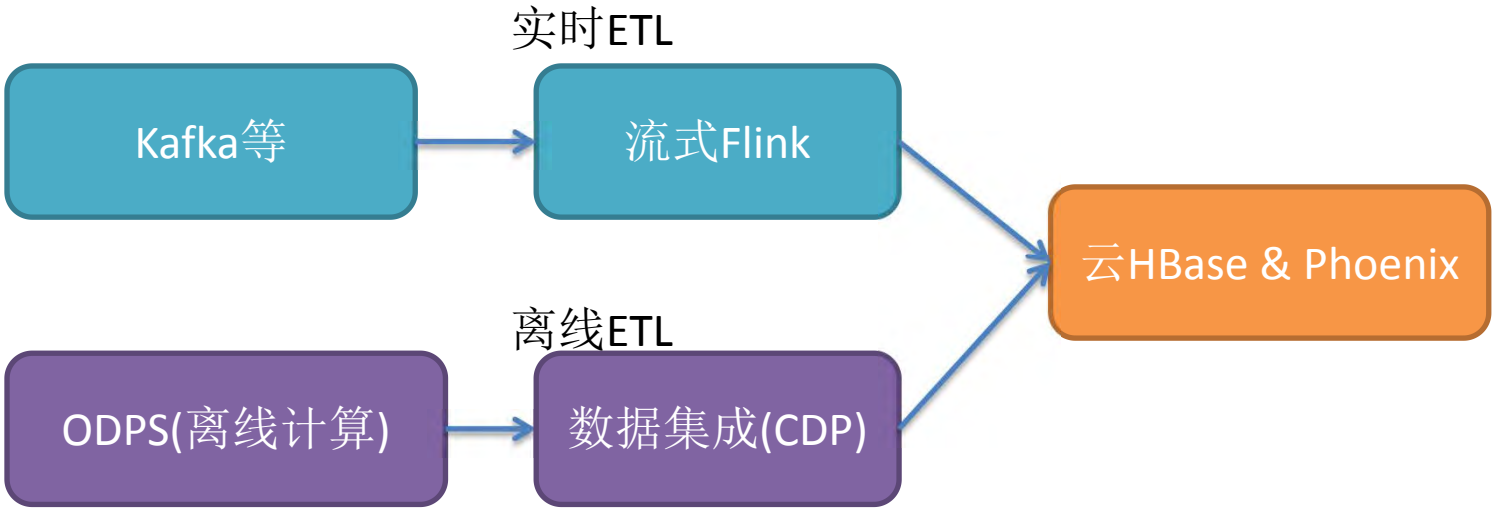
云HBase基础架构 – 基于VM



	云盘存储（中小客户）	本地盘存储（大客户）
起步存储	400G	50T（大约20T 最好用本地盘存储）
成本(110T) 一年，（以实际为准）	120W	35W（成本下降到 30%）
灵活性	计算 与 存储可以分别扩容	计算与存储绑定
稳定性	云盘通过隔离保障	磁盘通过物理磁盘隔离 更加稳定

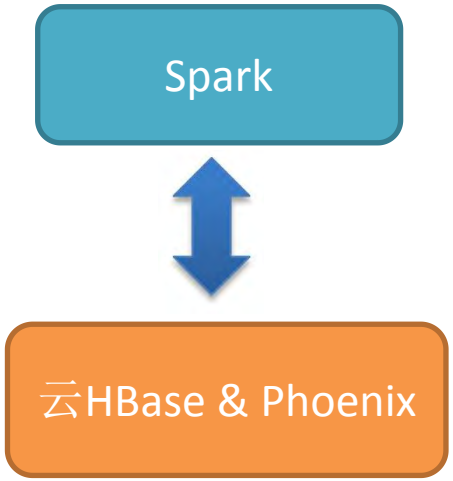
最佳组合

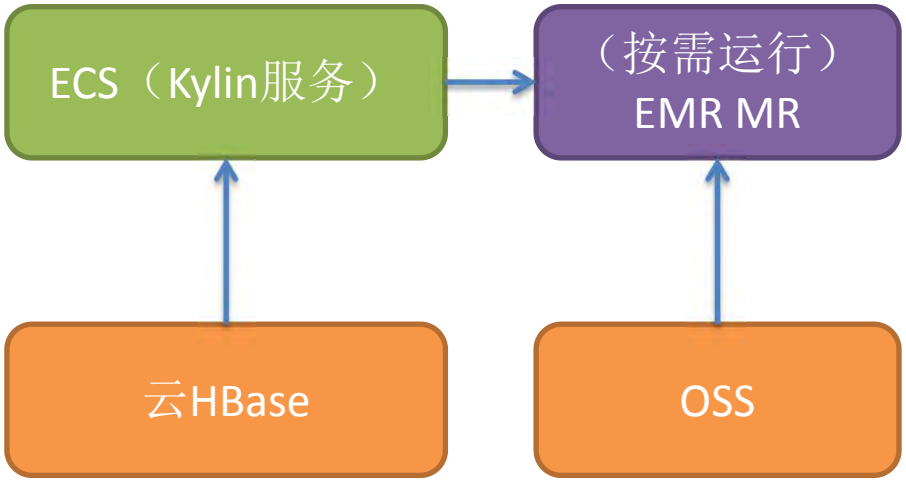
云HBase最佳组合-在线离线方式



HTAP实践（HBase Phoenix Spark）

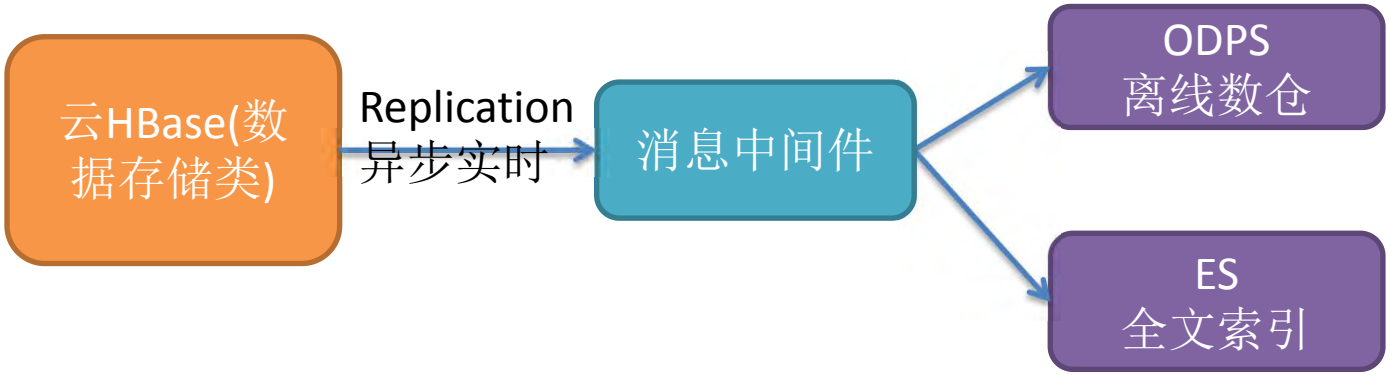
- 满足实时写入、更新
- Spark可以连接Phoenix
- 满足高性能Spark分析
  - 谓词下推
  - 分区裁剪
  - 直接生成RDD\DF\DS等



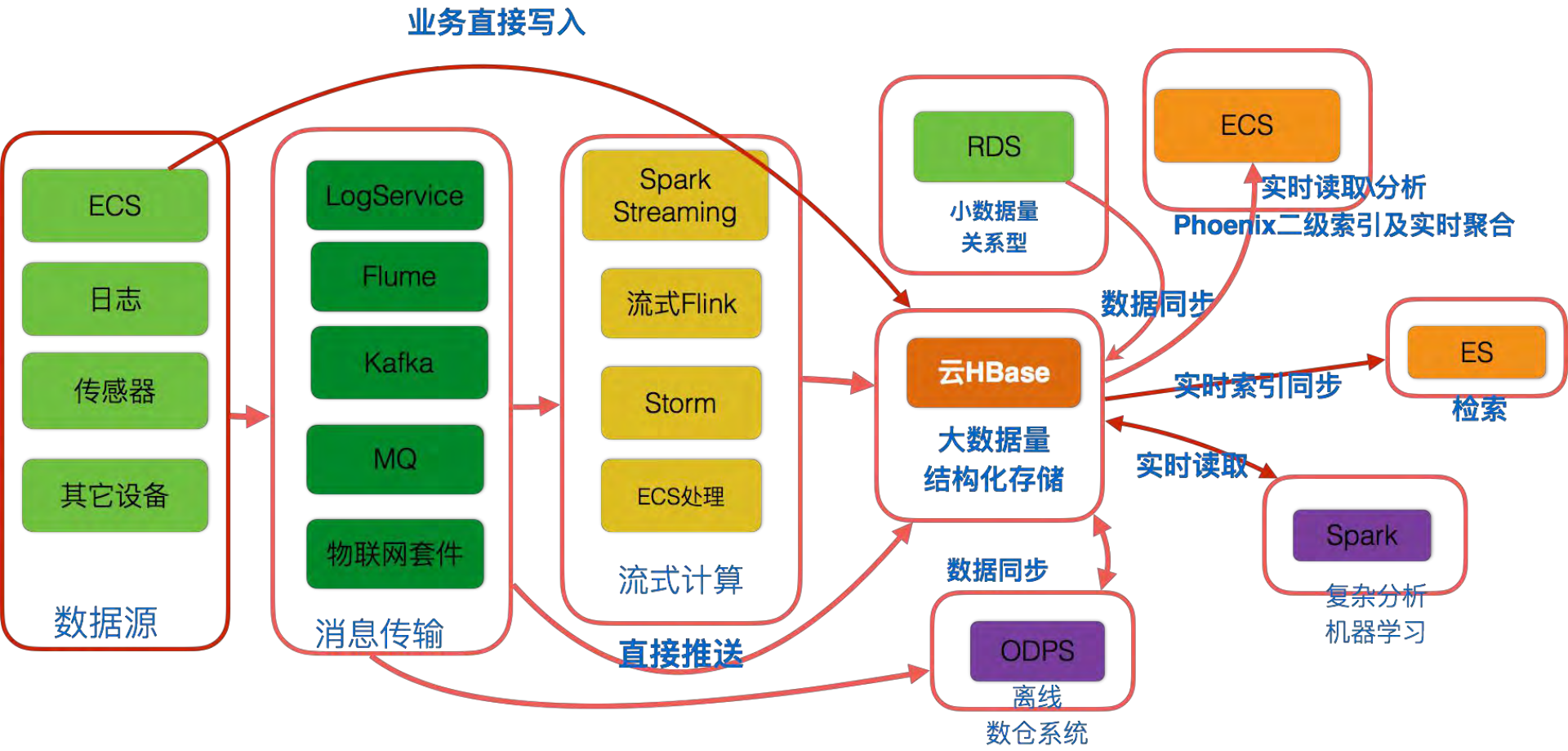


服务	
Kylin	多维度实时查询
云HBase	满足高效的写入查询
EMR MR	按需、离线build
OSS	历史数据列式存储

- 数据到ODPS
  - 离线分析，出报表的需求
- 数据到ES
  - 检索的需求
  - 原始数据存放HBase，检索字段存放ES

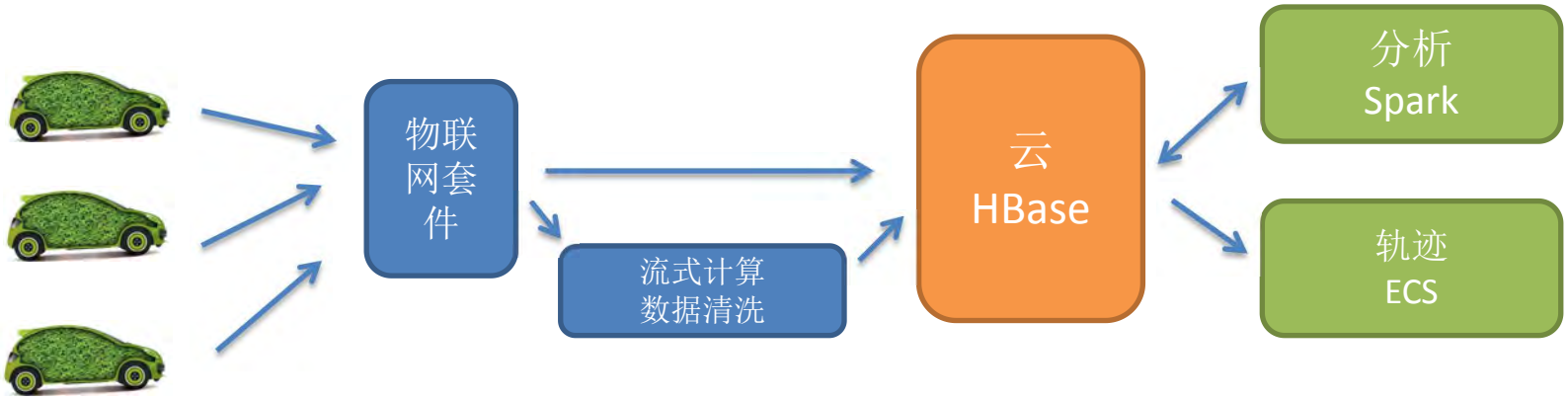


云HBase大数据组件流通性



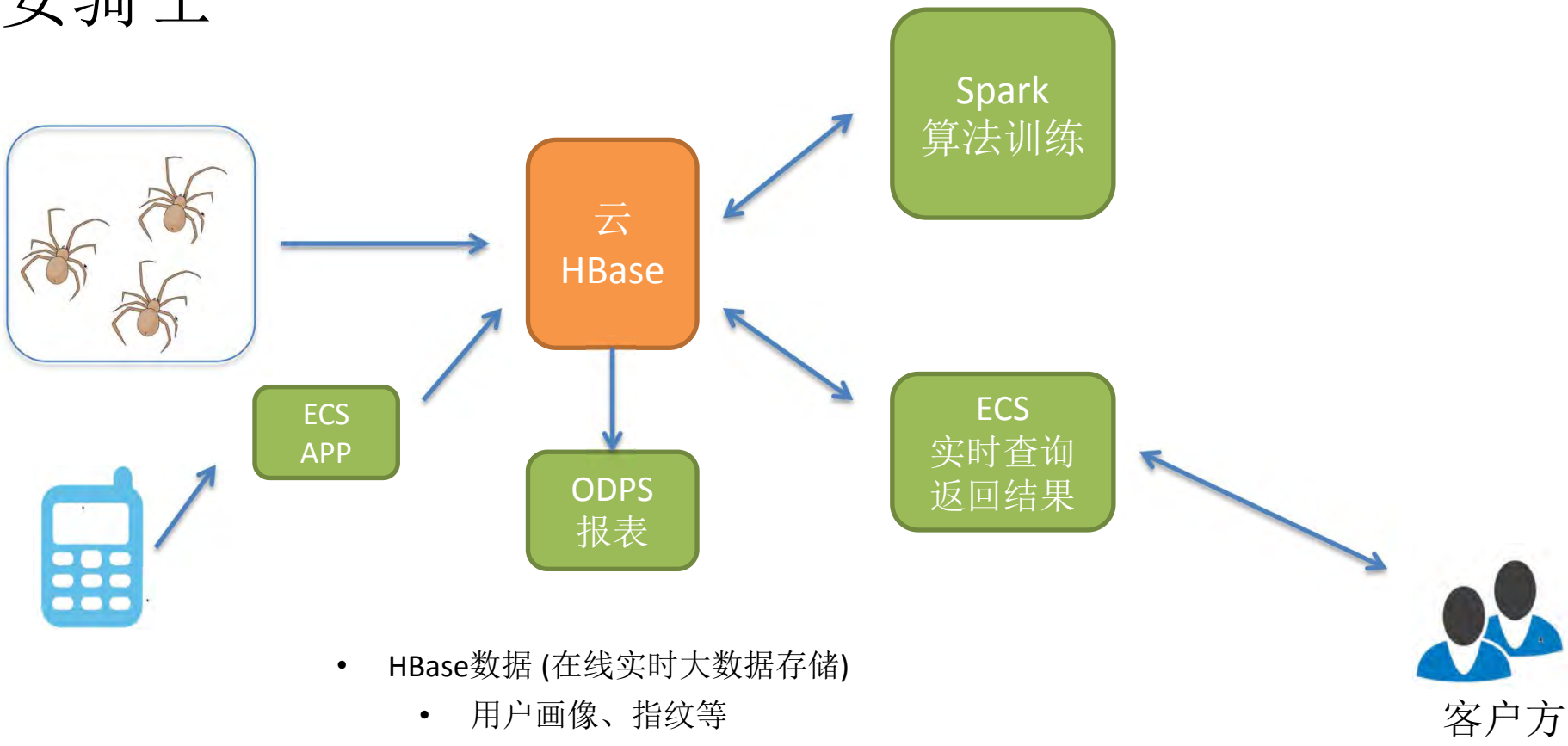
真实案例

某车联网公司



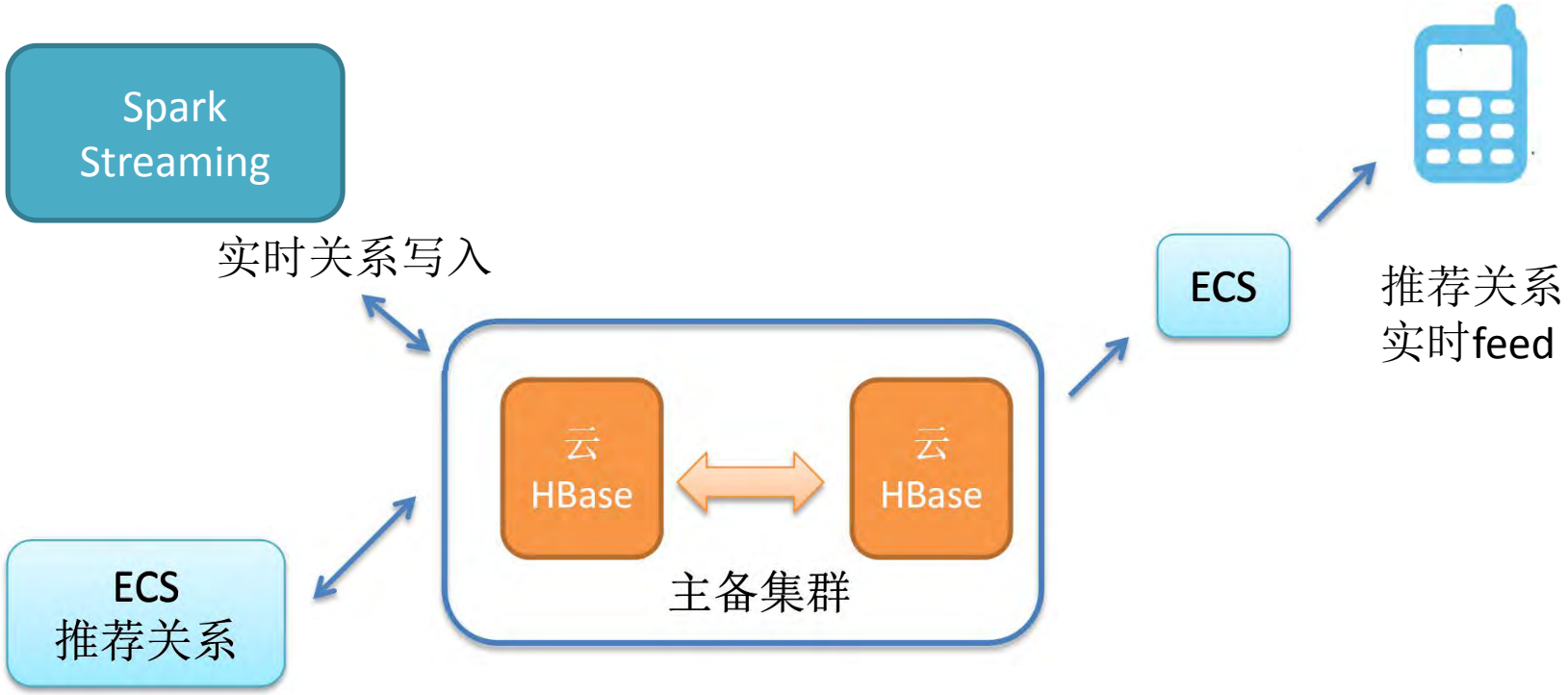
- Rowkey设计：Sub(Hash(车辆ID),5) + 车辆ID + 时间
- 每辆车 10s上传一次，每次1KB
- GeoHash存放轨迹信息
- 100万台车
- 1年数据存储3P
- 读写请求：100w+

安骑士



- HBase数据 (在线实时大数据存储)
  - 用户画像、指纹等
  - 爬虫、手机等原始信息
- 数据量
  - 200T+

Soul社交



- SLA要求高99.95，双集群保障
- 单集群读写高峰QPS 800w+
- 数据量：30T

某金融公司

- 单表 10000亿 +
- 二级索引 多个索引字段
- 数据量：100T

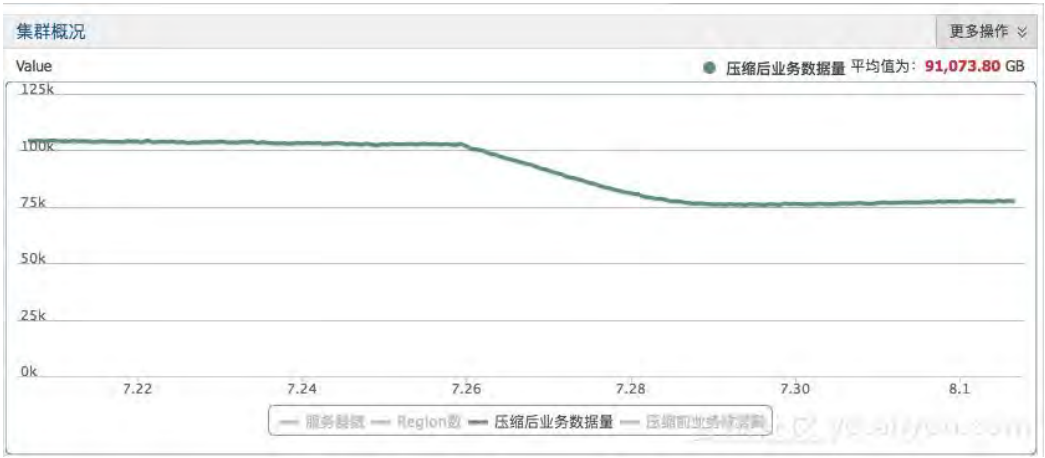
内核优化


工匠精神
大国工匠 有态度守匠心
Craftsman of great powers The spirit of originality

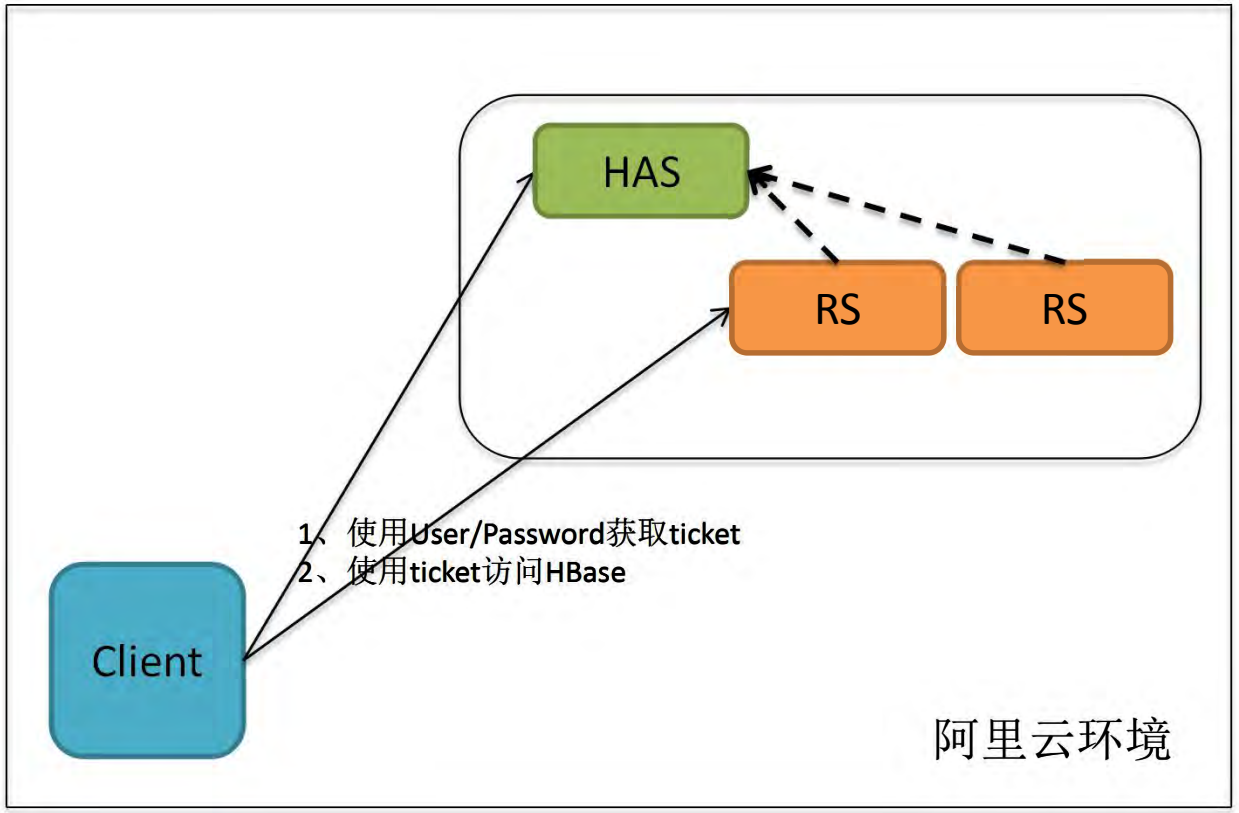
- 更快更省的CCSMap
  - 自己来管理写缓存的生命周期
- 永不晋升的Cache – BucketCache
  - 向JVM深圳一块用不归还的内存作为BlockCache，内存进行规定大小分段
- AliGC
  - 中等生命周期对象中最”大头"的部分，将这些对象在生成时直接分配到中等生命周期租户的old区，避免RSet标记
- HDFS串行pipeline改为并发Quorum机制
  - 降低写抖动

YGCT从120ms降低到5ms
基本消灭GC的影响

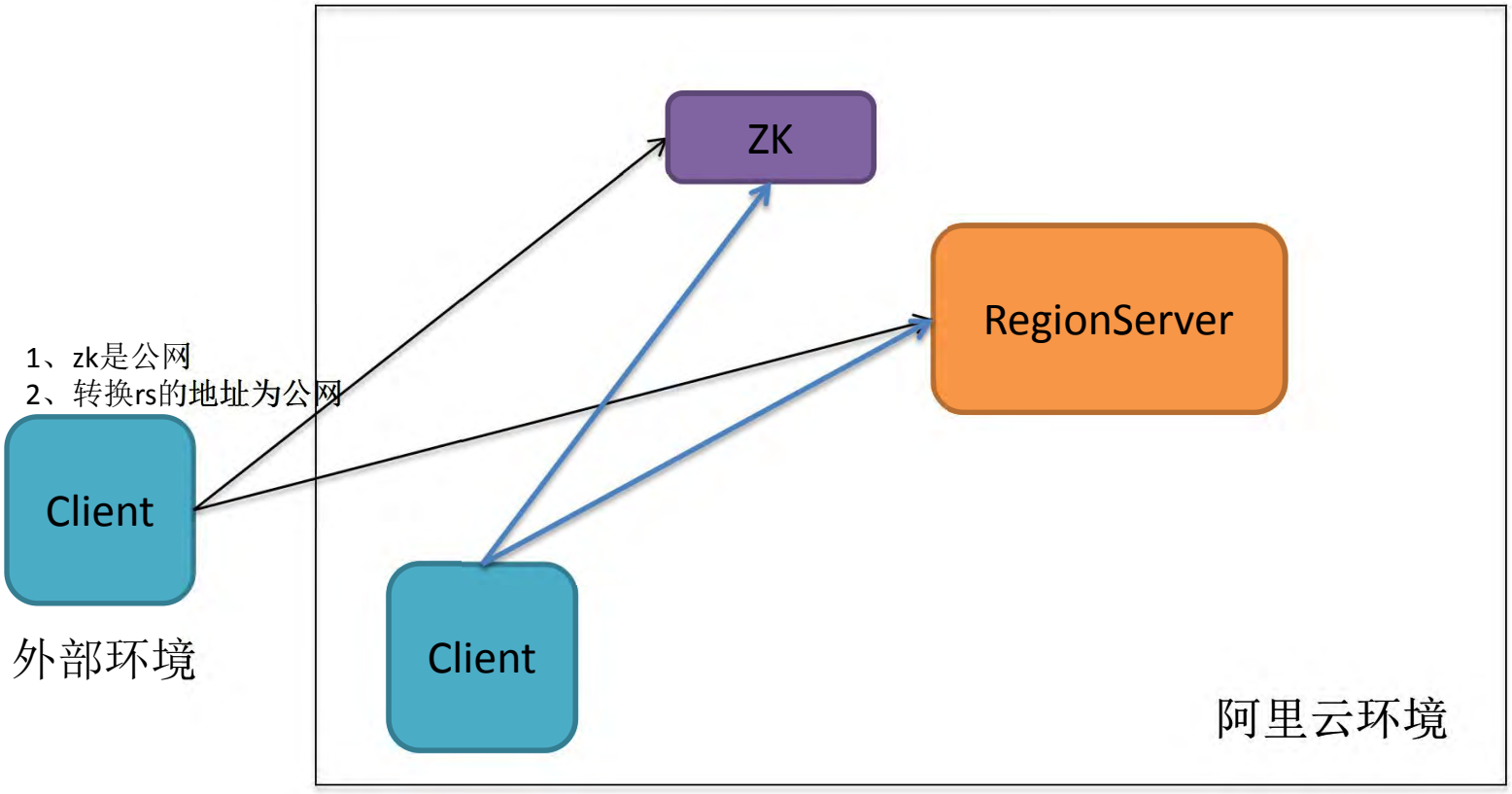
- ZSTD新型无损压缩算法
  - ZSTD的数据压缩率相对于LZO基本可以提高25% - 30%
- 新Indexable Delta Encoding上线
  - 相对于Diff encoding rt下降50%，但存储开销仅仅提高3-5%



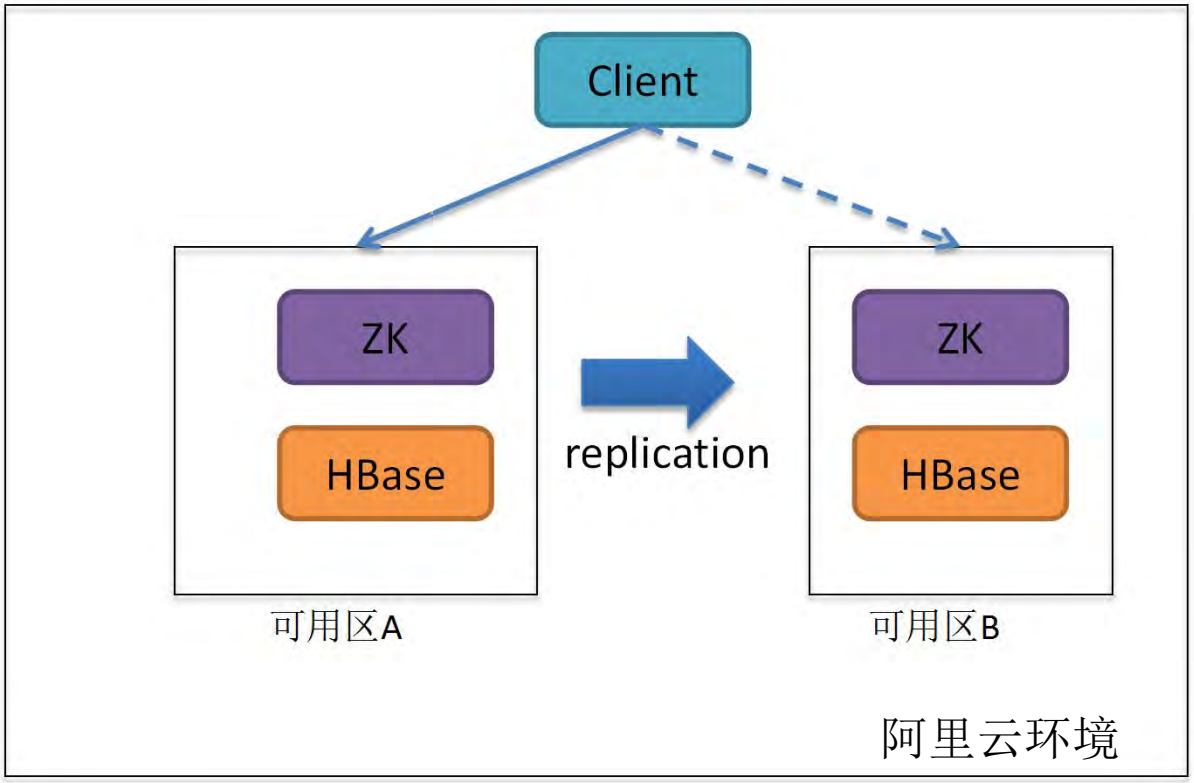
- 安全白名单：机器网络防火墙
- VPC：网络隔离
- 认证：
  - User/PassWord、授权ACL
  - 跟MySQL体验一致
- 授权
  - 可以授权到表及列族
- 加密
  - 数据加密



- 支持内网公网同时访问


ZK
RegionServer
1、zk是公网
2、转换rs的地址为公网
Client
外部环境
Client
阿里云环境

- Proxy代理切换
  - 人工切换
  - 自动切换
- 断网演练
- 双可用区
- Replication优化
  - 同城100ms
  - 异地1s



99.99%可用性

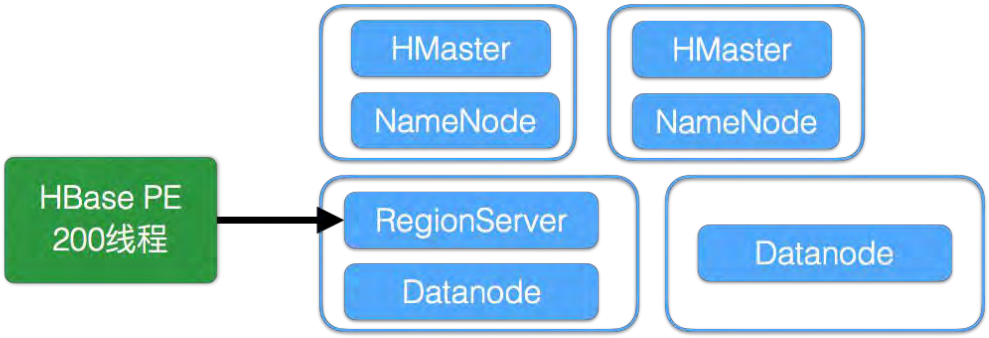
云HBase内核优化 – HBase性能

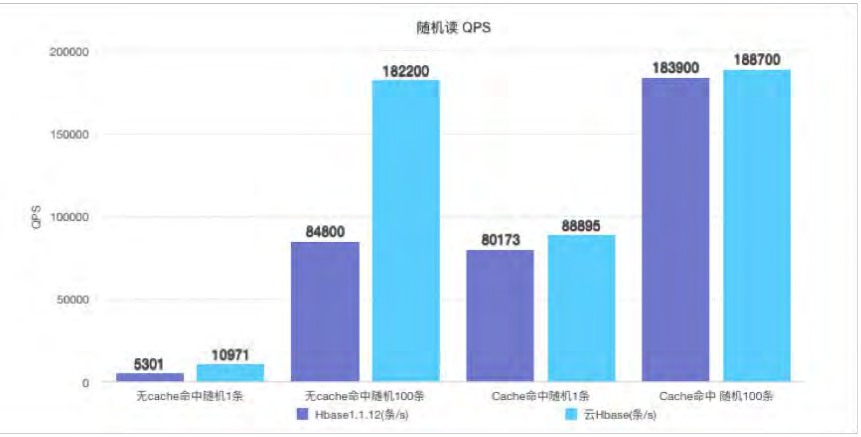
社区版本：1.1.12 VS 阿里云云HBase版本

2 slave 8cpu32g

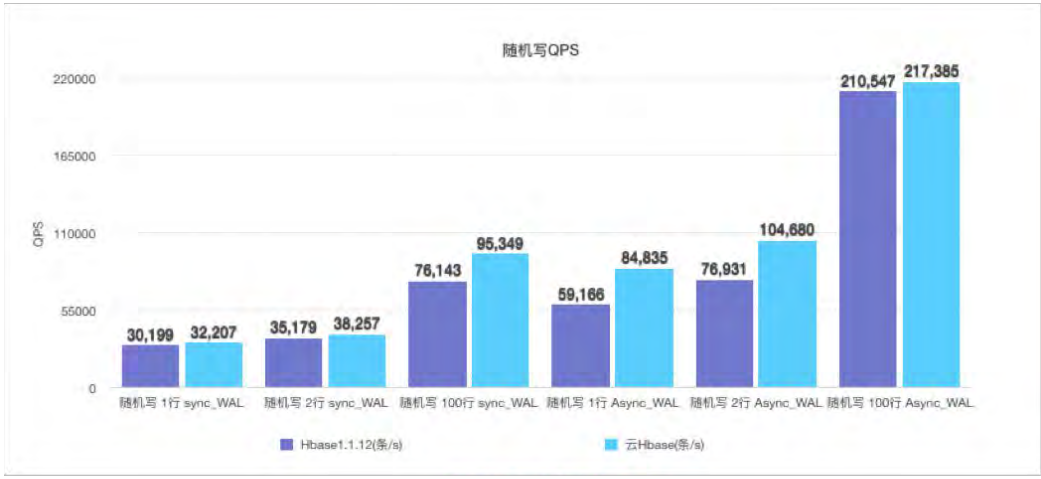
启动单个RegionServer

单条写 1KB





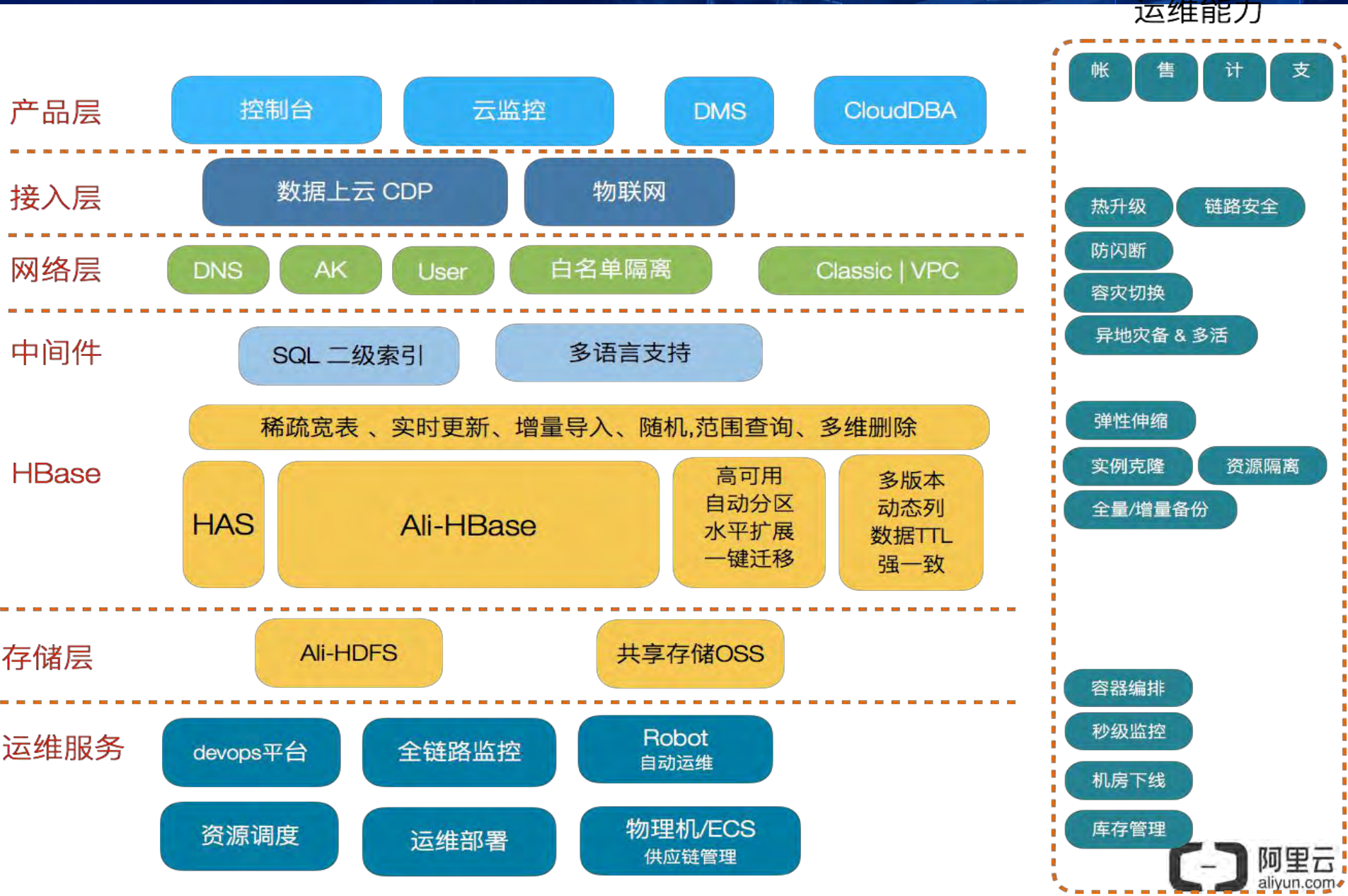
随机读最高提升 200%以上



随机写提升50%

平台能力

云HBase平台能力


运维能力
产品层
控制台
云监控
DMS
CloudDBA
接入层
数据上云 CDP
物联网
网络层
DNS
AK
User
白名单隔离
Classic | VPC
中间件
SQL 二级索引
多语言支持
HBase
稀疏宽表 、实时更新、增量导入、随机,范围查询、多维删除
HAS
Ali-HBase
高可用
自动分区
水平扩展
一键迁移
多版本
动态列
数据TTL
强一致
存储层
Ali-HDFS
共享存储OSS
运维服务
devops平台
全链路监控
Robot
自动运维
资源调度
运维部署
物理机/ECS
供应链管理
帐
售
计
支
热升级
链路安全
防闪断
容灾切换
异地灾备 & 多活
弹性伸缩
实例克隆
资源隔离
全量/增量备份
容器编排
秒级监控
机房下线
库存管理

云HBase平台能力 – 运维能力

运维自动化、白屏化

自动守护服务

在线扩容节点\磁盘

内核在线升级

可用性检测及报警

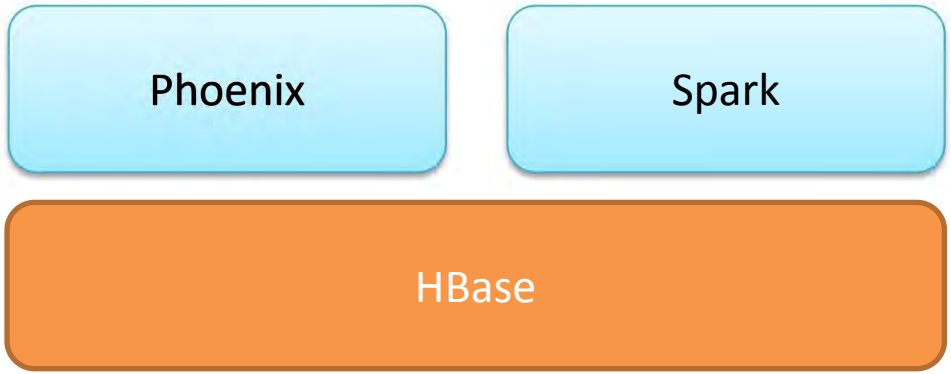
15分钟快速交付

指标可视化

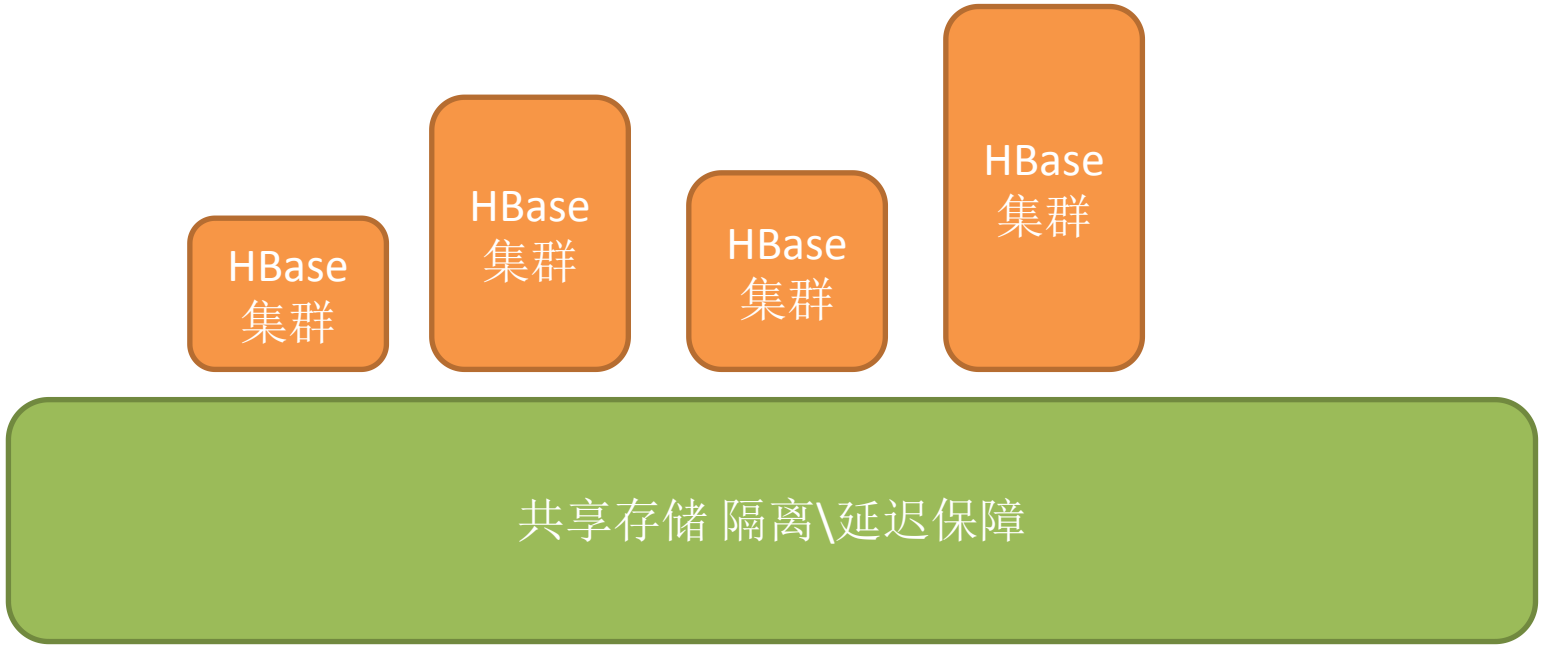
专家坐诊

The Next

- 分析能力 提升
  - 融合列存技术 索引
- HBase + Spark
  - 改善分析能力



从计算存储本地化 到完全计算存储分离！（研究中）



相关资料

- HBase全网最佳资料(实际案例\知识点)：https://yq.aliyun.com/articles/169085
- 云HBase产品首页：https://www.aliyun.com/product/hbase

我们招聘：

HBase\Hadoop\Spark专家 欢迎联系我！ 微信：

钉钉阿里云HBase技术交流群

谢谢

BDTC 2017 中国大数据技术大会
Big Data Technology Conference 2017