

Apache Doris × 阿里云联合 Meetup

🕒 10月26日 (周六) 13:30-17:15



浙江电信 Doris 实战之路

喻志强

浙江电信 大数据中心

2024年10月

浙江电信大数据平台建设历程



TERADATA



建设内容：基于 TD 建设 B 域
数据仓库与数据集市应用
规模：20台
支持存储：400TB

建设内容：实时经分和网络大数据集群，
基于 CDH 构建人力、业务稽核、网管、
端到端应用、跨域分析等应用。
规模：700+ 台
支持存储：20PB

建设内容：构建**数据中台**能力，基
于中台开展作业、模型、报表迁移，
数据治理及培训推广工作，并实现
开发运营模式优化。
规模：20+ 台

建设内容：基于电信自研 PaaS 翼
MR/TDP+Iceberg+Doris 新建**湖
仓一体**架构进行 BMO 域数据统一
汇聚。逐步转型为以自有人员为主。
规模：640+ 台

2004

2010

2016

2017

2021

2022

2023

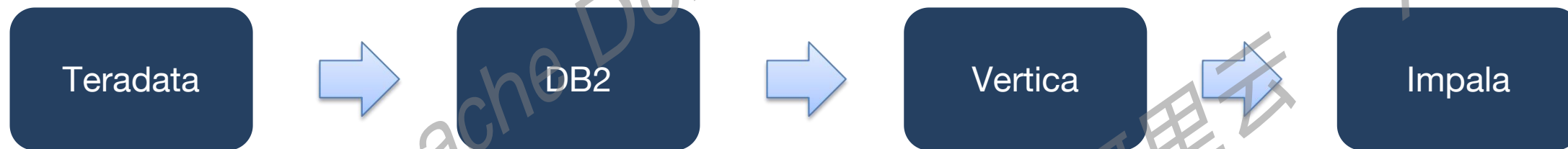


建设内容：基于 DB2 建设数据集市、
数据挖掘集市应用
规模：8 台（IBM小型机）+18 台天玑
存储阵列：456TB

建设内容：由 Vertica 承接 TD 和 DB2
的 B域**数据仓库与数据集市**应用。
规模：108 台
支持存储：2.5PB
日处理数据量：1.3TB

建设内容：基于 SR 构建实时数仓，
提升实时战报、反欺诈等应用时效性，
构建支撑报表等高并发 OLAP 场景
查询库。
规模：69 台
支持存储：1PB

浙江电信作为国内较早构建数据仓库系统的企业，是深度 MPP 数据库的使用者，商业化数据库产品从 Teradata 到 DB2 再到 Vertica，再到开源 Impala 组件。



浙江电信自 2004 年建设数据仓库起，便开始采用 MPP 数据库

2010 年开始考虑建设数据集市，同时因成本等因素，建设 DB2 数据库

2017 年因行式存储的 IO 瓶颈，对高性能 MPP 的需求，引入 Vertica (Teradata 和 DB2 随即改造下线)

2019 因 Vertica 对实时类应用支撑能力有限，加上资源拥挤，开始尝试开源 MPP 组件

原架构特点



+



- 易用性不足
- 稳定性不足
- 异构对象访问不友好
- SQL 语法支持度有限
- 主键和 Tablet 限制

商业 MPP 数据库
- Vertica

经分大数据

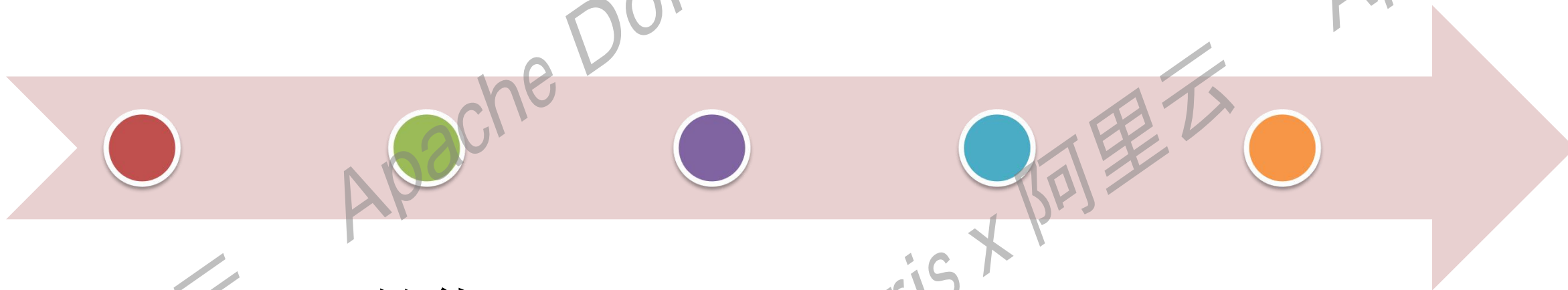
网络大数据

- 节点总量大超过1000+
- 架构**复杂**、应用场景**复杂**
- 数据**孤岛**
- 数据**冗余度过高**
- 数据**一致性问题突出**

易用性

稳定性

创新

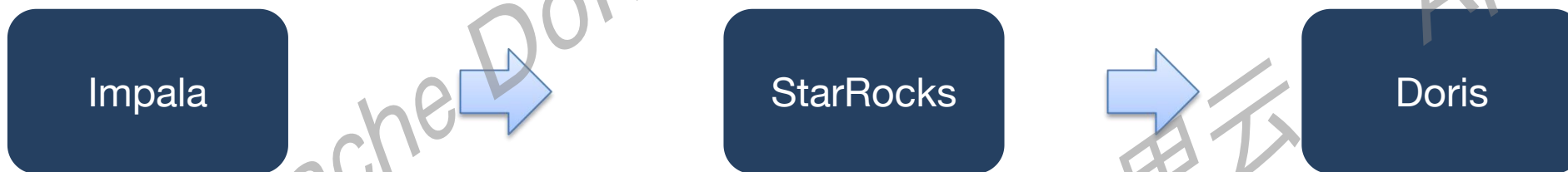


性能
提升

易维护

在使用 Vertica 和 Impala 时，新的瓶颈产生了，思考未来引进的方向：

- BI 报表类应用对底层库有高并发场景需求，原有 Vertica 并发能力太弱
- 实时战报类应用基于 Impala+Kudu 性能、稳定性不足，运营成本较大



Impala+Kudu 的组合存在的问题：

- 底层实时汇聚的表项和数据量越来越大，稳定性问题突出，性能也随之下滑厉害
- 流批数据一致性问题越来越突出
- 运营维护成本过大，本地人员也存在对应组件技能以及经验有限
- 能力自主化，提升运营管理效率

2022 年开始验证探索，选型需要满足 MPP、列式、高性能、高并发、维护便捷、标准 SQL、稳定等指标。

2023 年开始基于文档和可靠性需求开始进行 Doris 集群建设。

痛点和驱动： SR 使用到后期随着支撑应用场景逐步增加，稳定性问题异常突出基本上一周得有 1-2 次问题，也伴随着社区支持资源向其合作伙伴倾斜，大部分问题得不到解决或解决效率很差，同步随着全面自研 PaaS 改造的需求，开始全面改造 Doris 的实施。

监控层

Grafana

Prometheus

- 软件版本: prometheus-2.32.0、grafana-8.2.5
- 架构组成: 单机部署
- 硬件配置: (单机) 2*10C、256G、4*12TSATA

计算层

FE (3节点)

BE (5节点)

- 软件版本: Doris
- 架构组成: FE (3节点) + BE (5节点)
- 硬件配置: (单机) 2*10C、256G、4*12TSATA

初期实时战报应用上线



- **本地验证：**参照网上已有的一些标准测试对比基础上，直接拿实时战报任务实地验证
- 在计算资源压降的情况下，原实时战报任务相比 Impala 有了倍数的提升

Impala

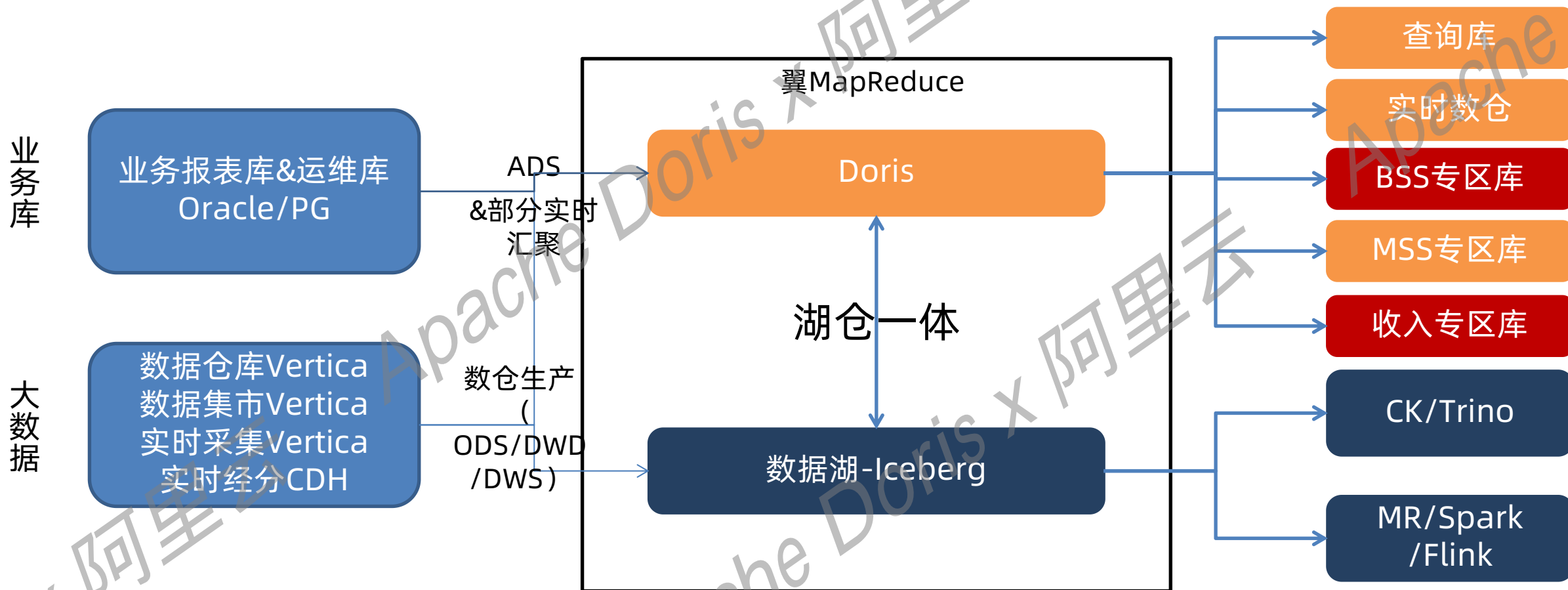
- CPU: 448C
- 内存: 3.75T
- 平均耗时: 高于5分钟

Doris

- CPU: 180C
- 内存: 2.25T
- 平均耗时: 低于2分钟

层次	任务名称	Impala+kudu 耗时	Doris 耗时	效率提升
基础层	RT_OFFER_INST_ORD_INFO_Z	0.02.59	0.00.43	3.11
基础层	RT_ORD_OFFER_PROD_INST_REL	0.00.51	0.00.08	5.55
基础层	RT_PROD_INST_ORD_INFO_Z	0.12.27	0.02.22	4.25
应用层	PRT_REAL_ISE_APP_DAY_Z	0.07.17	0.06.05	0.20
应用层	PRT_REAL_ISE_REPORT_DAY_Z	0.05.57	0.03.17	0.81
应用层	PRT_REAL_LIVE_ACT_DAY_Z	0.03.00	0.00.34	4.22
应用层	PRT_REAL_LST_JO_DAY_Z	0.25.43	0.05.03	4.10

基于Doris的湖仓一体架构转型



因不确定各类场景的支持程度，以及压力上涨带来的稳定性问题，采取方式是逐块验证，逐块使用：

- 一期：验证实时数仓相关应用，上线了实时数仓类相关应用；
- 二期：待稳定后开启验证 BI 类高并发类分析型应用场景，上线 BI 报表类平台应用；
- 三期：开启标签类宽表应用验证，上线画像标签宽表应用，地市相关数据集市应用也陆续上线；
- 四期：MSS 上云改造，实现通过 Flink+Doris 对接 Kafka。
- 版本探索：从2.0.1 持续到 2.1.5，性能提升超 30%

实时库

(日实时汇聚：超百亿条
任务总量：超3000个)

实时数仓：Flink 数据实时写入、实时战报任务执行。后续陆续上线省市其他实时应用，如反欺诈、电渠实时营销等。

查询库

(任务总量：超3000个
并发均值：超万级)

数据集市：省市 BI 类高并发分析应用、画像标签等大宽表复杂 SQL 应用，以及其他省市 ADS 层应用。

MSS专区库

(实时汇聚：1300张表
任务总量：超4000个)

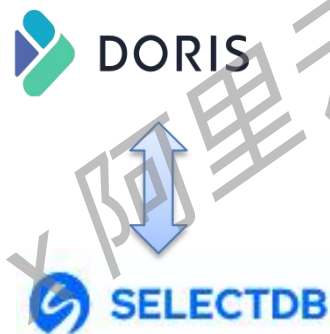
MSS专区库：MSS 上云改造实现 Flink+Doris 对接 Kafka，数据实时写入，数据时效性从离线跨越到实时。

协同与交互



随着 Doris 2.0.1 - 2.1.5 版本深入应用，逐步暴露和解决一些问题和功能需求，主要集中在跨源读写 Iceberg、FE-JVM 异常增长等，在天翼云大数据团队小伙伴的联动下，我们开启了中国电信与 Doris 社区的紧密合作，并得到云公司和社区的大力协助。

问题描述	解决方法	问题描述	解决时间	是否解决	优先级 (1级)
doris iceberg catalog访问空插异常问题	设置 metastore.filter.hook="" org.apache.hadoop.hive.metastore.DefaultMetaStoreFilterHookImpl"			解决	
doris be节点无法启动, 报错Please disable swap memory before	"调整be节点start_be.sh脚本中, 将以下内容注释掉. # [["\$swapon -s wc -l" -gt 1]] then # echo "Please dis			解决	
doris查询时遇见<=>符号查询慢	版本迭代到2.0.1修复后版本		2023/11/19	解决	
doris查询Iceberg时报hive metastore连接错误	版本迭代到2.0.3		2023/12/11	解决	
doris查询Iceberg时会偶尔报rpc超时错	remote_fragment_exec_timeout_ms=30000, 升级版本到2.0.3		2023/12/11	解决	
doris并发起来之后节点间rpc超时	remote_fragment_exec_timeout_ms=30000, 升级版本到2.0.3		2023/12/11	解决	
doris大小写不敏感设置后, 做insert into table_name select ' from	升级版本到2.0.3并进行小版本迭代, 使用时将库名替换为小写		2023/12/11	解决	
doris insert into语句过慢问题	enable_nereids_dmi=true, 使用新优化器			解决	
doris insert into with table_name as表名为大写会报错	将with table_name as中间表名替换为小写			解决	
doris BI报表访问, 特定语句BE节点频繁堵塞, 语句单线程跑BE节点又	set GLOBAL enable_nereids_timeout = false (关闭超时回退到老优化器)		2024/1/19	解决	
自动采集问题fe jvm溢出, 导致fe节点堵塞	set GLOBAL enable_auto_analyze=false.版本升级到2.0.4后不需要设置		2024/1/29	解决	
Truncate table 在集群运行30分钟后变慢					
catalog访问Iceberg data类型查询为空					
sql内存溢出问题					
doris反写Iceberg的版本					
CCR同步-整库同步待解决					
多表物化视图					
doris反写Iceberg分桶表报错问题					
doris访问H delete表报错问题					
表count()数据量会变化					
JVM异常导致FE节点堵塞					
doris访问Iceberg带时区字段, 时间差8个小时					
访问Iceberg用老优化器执行					
访问Iceberg导致be节点堵塞问题					
select array_max(array(split_by_string("a.b.c",""))).执行sql导致be节点全部挂掉					
元数据同步异常, 非master节点查询找不到tablet					
查询库 be节点异常堵塞					
升级2.1.5后Iceberg catalog里面表名和库名只能小写, 回退不会恢复					
任务效率下降, 主要为一个insert into的sql效率逐渐变慢					
升级30分钟后, truncate 效率变慢, 但不是以前debugstring问题引起, 打印火焰图的时候被自动劫后, truncate效率又变正常					
中台离线传输任务, 主要是doris (使用mysql数据源) 下发vatica的任务 (主要表现在集团下发任务) 不兼容, 在读取数据一段时间内					



中国电信浙江公司



后续开展



持续深入使用，构建 BSS 专区库将老旧报表库（Oracle）、运维库改造 Doris，构建收入报表分析专项库，同步其他模块相关应用需求也在对接中

BSS专区库

BSS报表库、运维库

收入报表专项

省市BI类高并发分析应用从Vertica改造查询库

.....

存算分离、国产化适配验证、混合部署等探索

问题与期望



随着使用的深入，数据量和任务量级数增长，新的需求与难点也逐步显现：

稳定性

跨源访问

资源隔离

便捷性

硬件异构化



谢谢

Apache Doris x 阿里云

Apache Doris x 阿里云

Apache Doris x 阿里云

Ap

Apache D

