HIVE在腾讯分布式数据仓库实践分享

赵伟

Tencent 腾讯 TDW 腾讯分布式数据仓库 TENCENT DISTRIBUTED DATA WAREHOUSE

自我介绍

- 赵伟
- 2009年加入腾讯
- 任职于数据平台部
- 一直从事海量数据处理平台研发工作
- 熟悉hive、hadoop、postgreSQL等技术

提纲

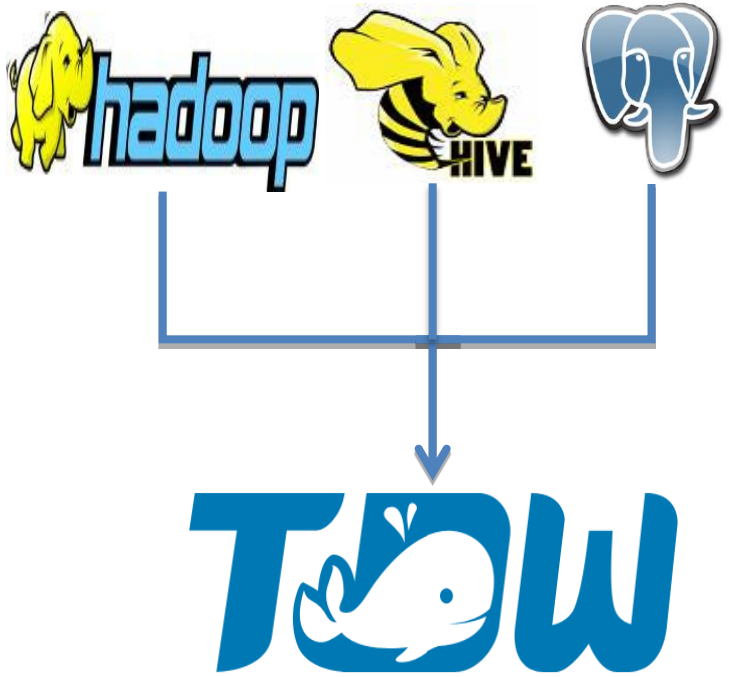
腾讯分布式数据仓库介绍

HIVE在TDW中的实践

TDW HIVE接下来的工作

TDW简介

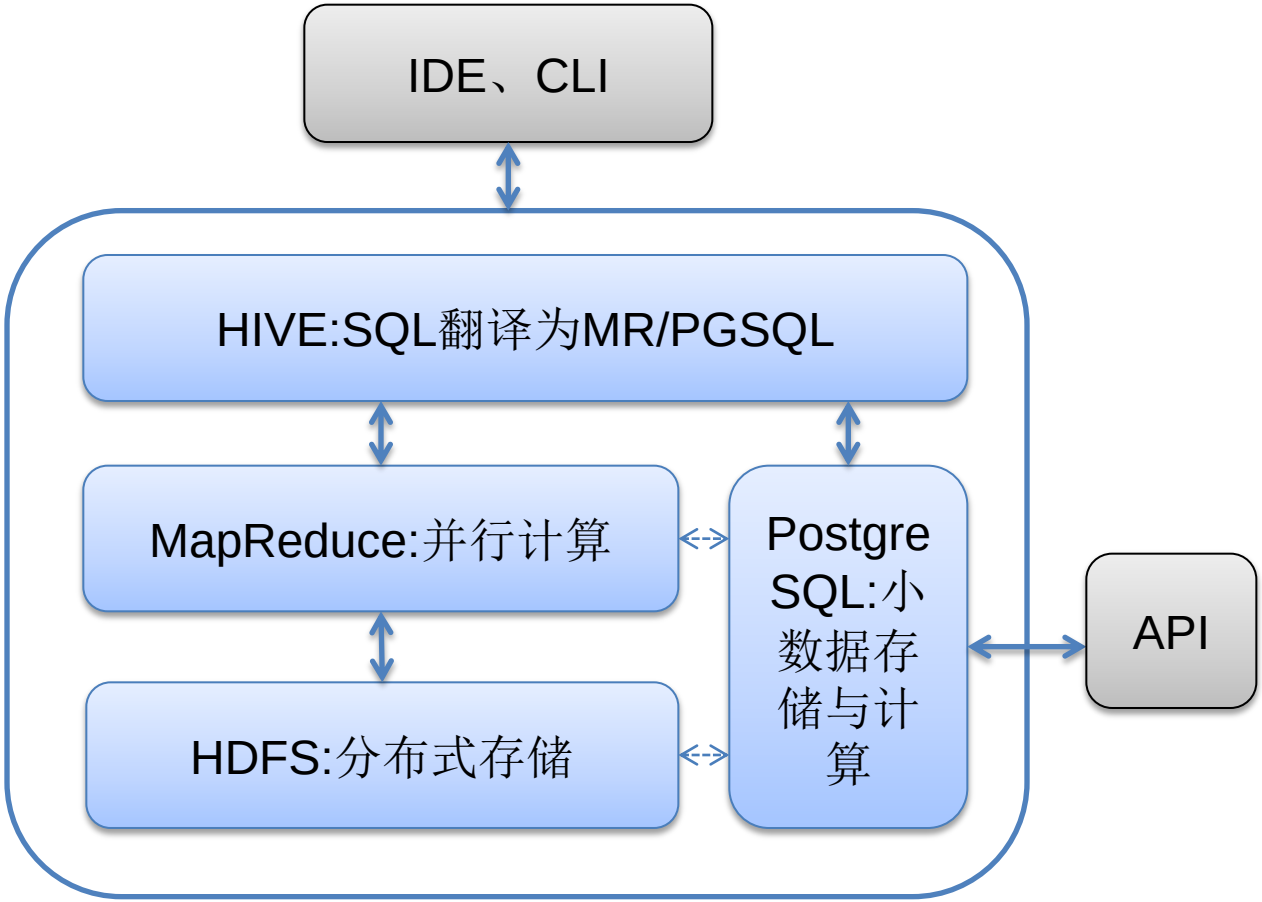
- 腾讯分布式数据仓库，简称TDW
- 基于Hadoop、Hive和PostgreSQL之上，进行了大量定制和优化
- 腾讯内部最大的分布式系统
- 公司级数据仓库，集中了各业务有价值的数据
- 对腾讯内部提供离线海量数据分析服务
  - ✓ 数据挖掘
  - ✓ 产品报表
  - ✓ 经营分析



TDW特性

特性	说明
存储和计算容灾	集群中个别节点down机不影响存储和计算
存储和计算线性扩展	通过添加节点线性扩展存储和计算能力
SQL语言	select、insert、join、where、groupby、having、limit、orderby、分区、视图等
SQL函数	简单函数、聚合函数、窗口函数、数据挖掘函数
过程语言	以python语言为母体的PL/python
多维分析	rollup、cube、grouping
MapReduce	允许提交MR任务
多种存储结构	文本/结构化/列存储/ProtoBuf/DB存储
SQL/MED	可访问和管理PostgreSQL、Oracle数据
开发工具	集成开发环境TDW IDE、命令行工具PLClient
任务调度系统	图形化的任务依赖配置、数据流转配置
系统DB	元数据与普通表一样可以通过TDW SQL进行访问
其他	Show processlist、kill query、select expr、insert values、show create table、comment on操作等

TDW核心架构



TDW运营现状

- 机器总量5000+，最大集群约2000个节点
- 覆盖腾讯90%+的产品
- TDW集成开发环境活跃用户数：200+
- 每日运行的分析SQL数：50000+
- 每日SQL翻译成的MR job数：100000+
- 最近半年SLA：99.99%

TDW

提纲

腾讯分布式数据仓库介绍

HIVE在TDW中的实践

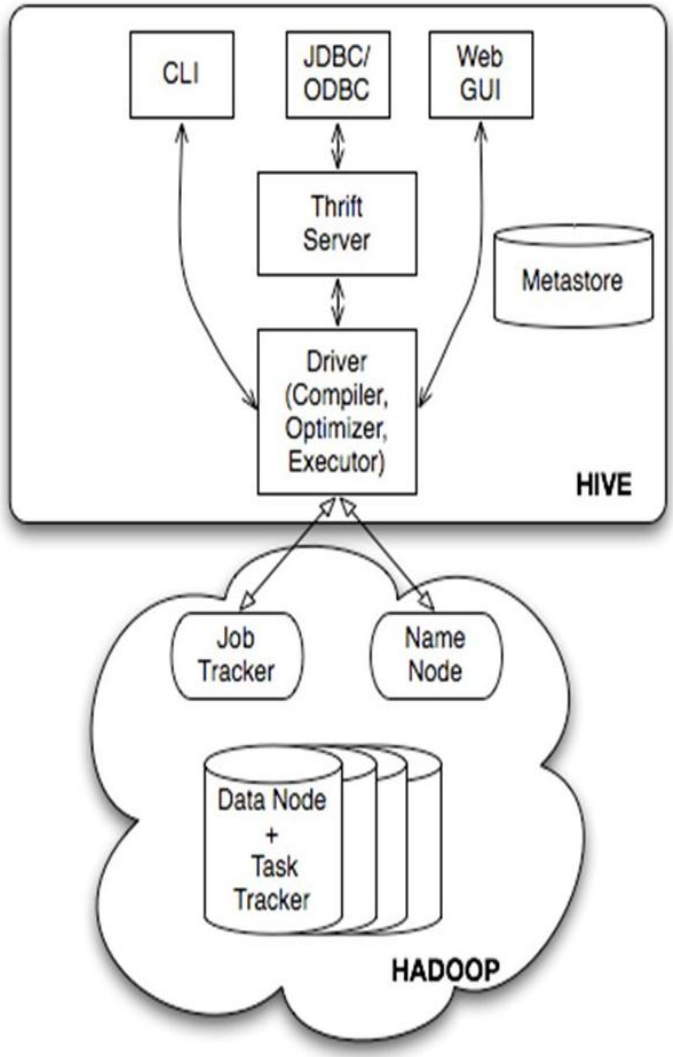
TDW HIVE接下来的工作

HIVE简介

- HIVE是什么

  HIVE是一个在Hadoop上构建数据仓库的软件，它支持通过类SQL的HQL语言操作结构化的数据

- HIVE的优势
  - ✓ 实现了基本的SQL功能
  - ✓ 可扩充UDF/UDAF
  - ✓ 可自定义SerDe
  - ✓ Thrift协议，支持多语言客户端



HIVE的不足

- 数据仓库功能不够完善
  - ✓ 缺乏权限管理、过程语言、窗口函数、多维分析等功能
- 使用门槛高
  - ✓ 用户界面简陋、运行调试麻烦、问题定位困难、查询计划难看
- 性能有提升空间
  - ✓ SQL翻译成的MR任务效率低或者不合理
- 不够稳定
  - ✓ 在生产环境中经常会出现卡死、元数据损坏、进程异常退出等

对HIVE定制和优化

- 功能扩充
- 易用性提升
- 性能优化
- 稳定性优化

功能扩充

- 基于角色的权限管理
  - ✓ 参考Oracle与MySQL的功能进行设计
  - ✓ 增加元数据相关的表结构、增加权限管理SQL语法
- 兼容Oracle的分区功能
  - ✓ 增加分区相关的元数据
  - ✓ 实现了Oracle建分区表的语法
  - ✓ 修改查询优化器，使它支持显式和隐式分区优化
- 窗口函数
  - ✓ 借鉴UDAF框架，实现了UDWF窗口函数框架
  - ✓ 在UDWF基础上，实现了lag、lead、rank、row_number等常用窗口函数

功能扩充-续

- 多维分析功能
  - ✓ 通过变换抽象语法树实现cube、rollup、grouping等
- 公用表表达式（CTE）
  - ✓ 将with固化为临时表，作为后面语句的输入
- DML（update/delete）
  - ✓ update和delete都使将结果数据保存为临时表，然后替换原表。
- 入库数据校验
  - ✓ 入库检查数据合法性
  - ✓ 通过hadoop counter返回入库成功条数与reject条数

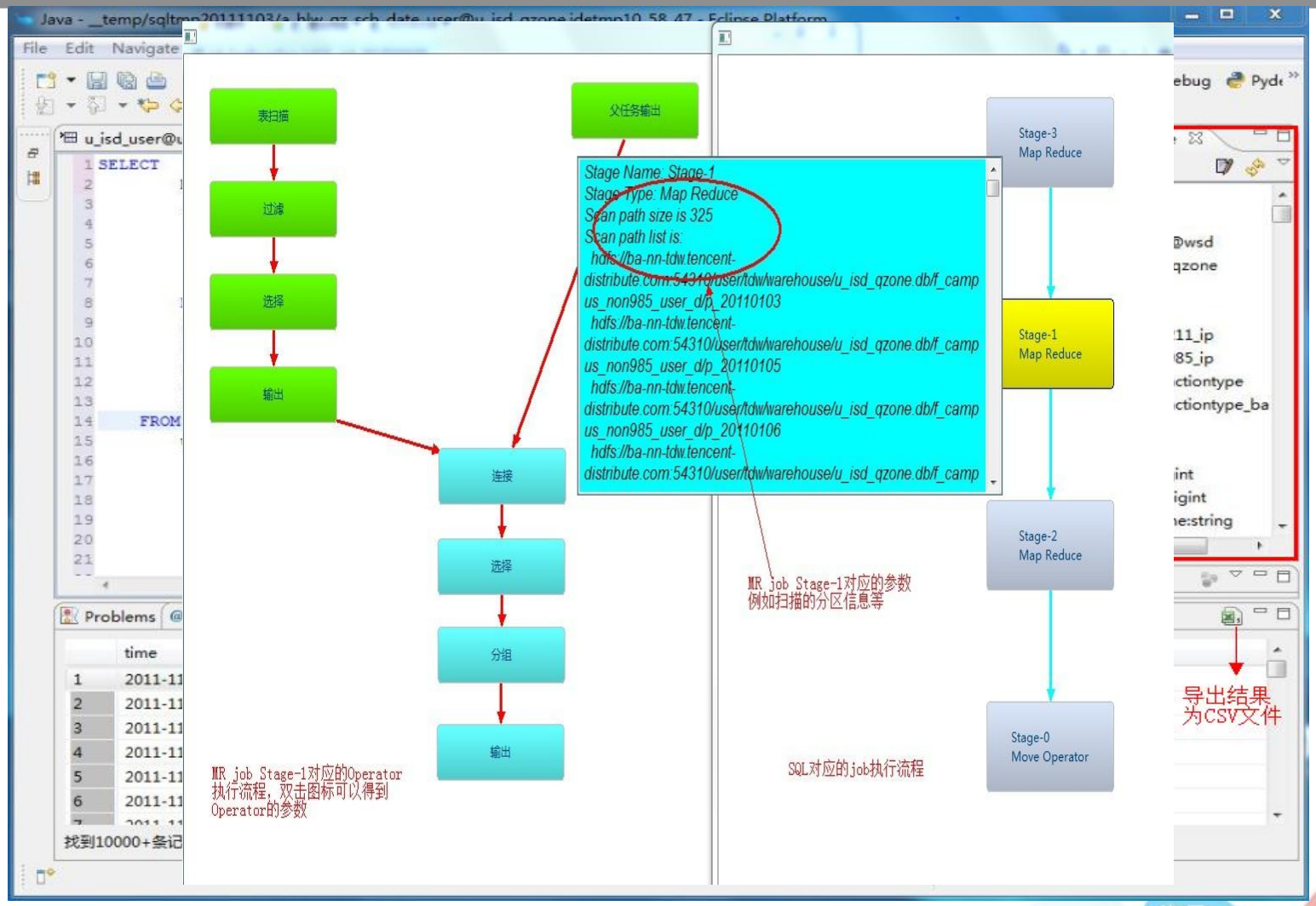
易用性提升

- 命令行工具
  - ✓ 使用Python实现的HiveServer命令行工具
  - ✓ 命令的使用格式借鉴了SQLPlus
- DB存储引擎
  - ✓ 将PG中的表映射到TDW中
  - ✓ 在TDW通过JDBC与PG进行数据交互
  - ✓ 在PG中通过tdwlink功能或者tdw_fdw访问TDW数据
- SQL语法细节
  - ✓ exists、in、not like、insert values、select expression，show create table、show processlist、kill query、comment on操作、系统DB

易用性提升-TDW集成开发环境

- Eclipse提供基本的IDE功能
- PyDev提供过程语言编辑、运行和调试环境
- Jython提供Python与Java的粘合功能
- 借鉴了开源eclipse SQL功能插QuantumDB

TDW集成开发环境-续

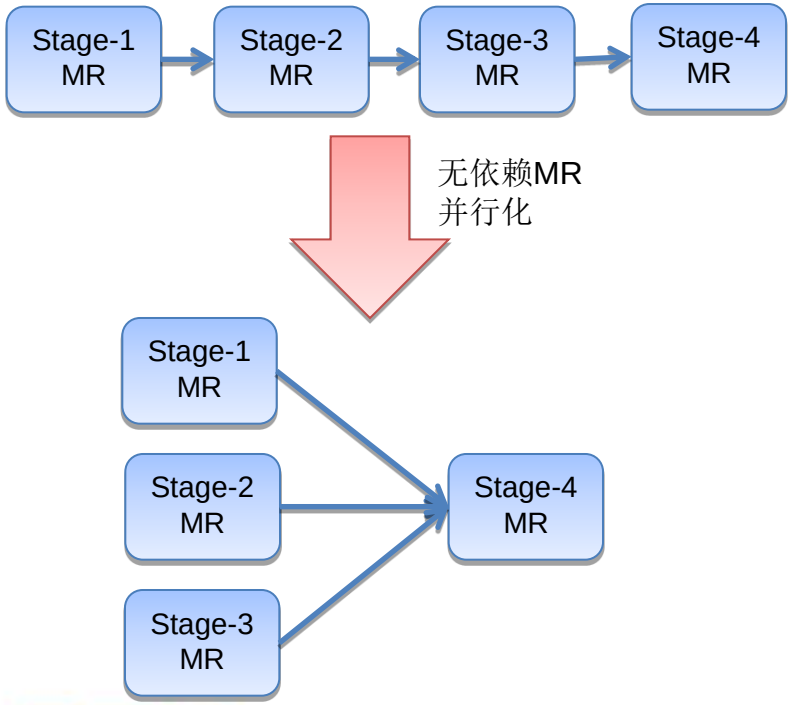


性能优化

- 自定义的存储格式
  - ✓ 二进制存储，读写更高效
  - ✓ 支持Lzo压缩，均衡了压缩比与压缩/解压效率
  - ✓ 优化了随机读取
- Hash Join
  - ✓ 在Map端使用Hash分区进行join
  - ✓ 对共用id的业务数据关联优化效果较好
- 按行split
  - ✓ 使每个map处理的行数相同，避免task长尾
  - ✓ TDW自定义存储格式使得可以做到快速split
- Order by limit优化
  - ✓ 在Map阶段使用堆排序选出top N，减少reduce的输入数据量

性能优化-MR并行优化

- 社区性能优化的补丁移入TDW
- 设置HIVE参数set hive.exec.parallel = true打开
- 原理是HIVE翻译成的MR任务尽量并行化执行
- 已经在TDW大规模应用，优化效果明显



优化前后	对比维度	对n个字段做cube计算	对m个字段做rollup计算
优化前	执行过程	2^n+1个MR逐个串行	2*m+1个MR逐个串行
	执行时间	t1	t2
优化后	执行过程	Stage-1：2^n 个MR并行 Stage-2：1个MR并行	Stage-1：(m+1)个MR并行 Stage-2：1个MR并行
	执行时间	$t1/2^n$	t2/(m+1)

稳定性的优化

- HiveServer容灾与负载均衡
  - ✓ DNS轮训
- 大结果集获取接口优化
  - ✓ 使用FetchN实现FetchAll
- 元数据接口优化
  - ✓ 优化元数据接口，减少元数据DB访问量
  - ✓ Datanucleus-core-2.0.3.jar+补丁NUCCORE-559、NUCCORE-553

稳定性的优化-续

- 内存泄漏解决
  - ✓ 使用jmap、jhat进行剖析和统计
  - ✓ 不再使用的变量赋值为null
- 服务过载保护
  - ✓ HiveServer最大连接数限制
  - ✓ SQL长度限制
- hdfs实例获取接口优化
  - ✓ HADOOP-6231

HIVE在TDW中的实践总结

- 功能：对TDW功能需求数量降低80%
- 易用性：数据分析应用开发效率提升3倍
- 性能：部分SQL性能是社区HIVE的2倍
- 稳定性：HIVE异常告警减少90%
- 仍然需要解决的问题
  - ✓ SQL优化器不够智能
  - ✓ 元数据模块效率低下
  - ✓ 基于eclipse的IDE过于笨重

提纲

腾讯分布式数据仓库介绍

HIVE在TDW中的实践

TDW HIVE接下来的工作

接下来的工作

- SQL优化器
  - ✓ 引入基于cost模型的查询优化
- 元数据
  - ✓ 元数据结构重构
  - ✓ 元数据接口重构，去除低效的ORM层
- 易用性
  - ✓ web版的IDE

Q&A

谢 谢！

Tencent 腾讯 TDW 腾讯分布式数据仓库 TENCENT DISTRIBUTED DATA WAREHOUSE