百度数据仓库体系介绍

刘立萍
liuliping@baidu.com
2012年11月27日

Bai du 百度

提纲

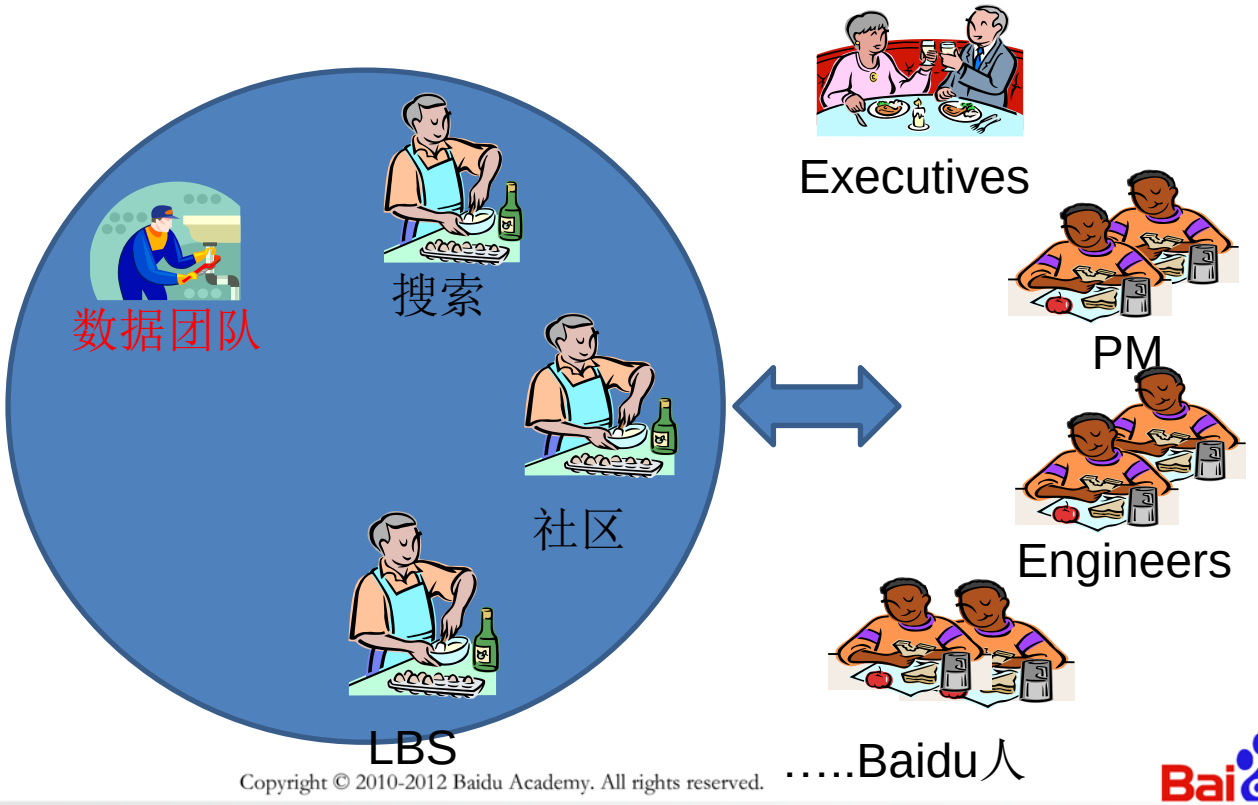
百度大数据应用

- 网页和超链 10PB-50PB
- 日志 100PB
- 数据仓库 80PB
- 广告 1TB

百度基础架构部数据团队

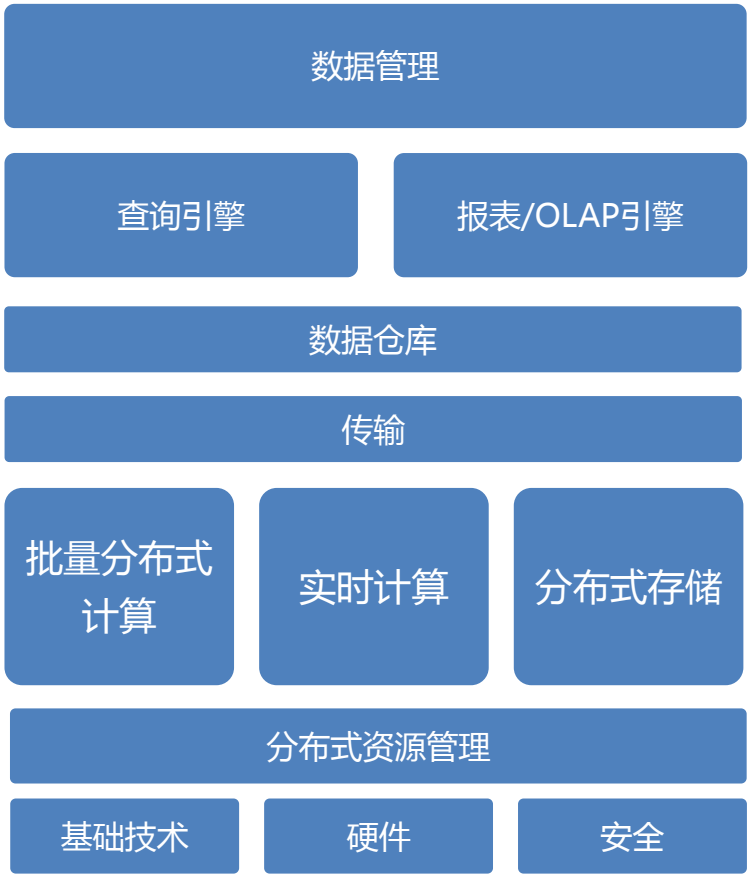
- 我们的职责：
  - 整合Baidu基础数据，构建数据平台，提供技术服务，推动数据处理，挖掘和应用

- 我们的用户：



基础架构部大数据平台

- 分布式存储
  - KV： Mola、
  - Table： CCDB
- 计算
  - 批量计算：Abaci、Peta
  - 小批量计算: Mini-Batch Process
  - 流式计算： Stream Process
- 调度
  - 底层资源管理：Matrix
  - 上层通用调度：Long-Scheduler
- 数据仓库体系
  - 格式化：Logging/PB
  - 传输：BigPipe、LogSaver
  - 数据仓库：DW
  - 报表&多维分析引擎：Doris
  - Ad Hoc 查询引擎：Query Engine
  - BI：Baidu Insight

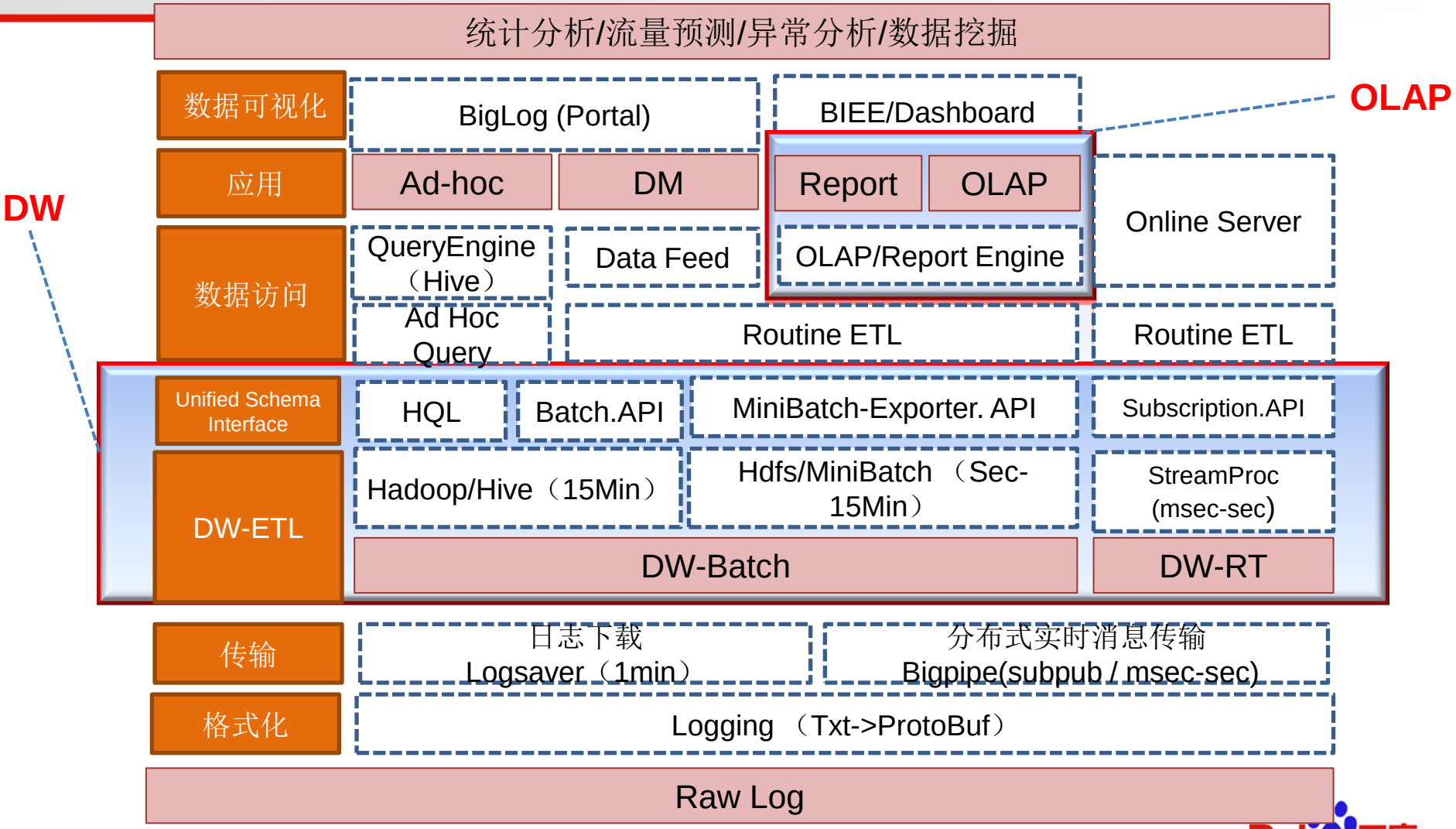


DW

DW设计目标

- 目标：
  - 应用场景： 策略分析，统计分析，挖掘程序，DataFeeds
  - 用户： 可以通过类SQL工具进行即席(Ad-hoc)查询，提高访问效率
- 思路：
  - 数据全，准确，一致
  - 易于理解 ：数据建模
  - 一致的数据结构
  - Fast IN ：数据写入，由ETL效率决定
  - Fast OUT ：数据访问，由吞吐率决定
- 价值：非结构化高价值数据治理，数据分层
  - 基础数据提供者
  - 数据集市（Data-Mart）：增加特定领域（Domain Specific）数据，逻辑，与DW搭配成为数据管理基本结构;特定领域数据主题分片, 提速，固化领域逻辑
  - 可以隐藏99%的对原始日志访问需求

DW集成环境



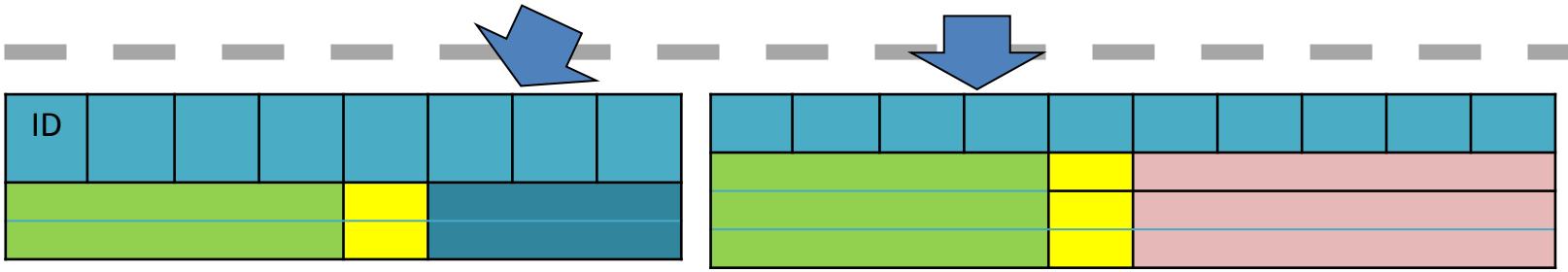
DW 新特性

- 逻辑模型
  - 概念层、逻辑层、物理层
- ETL时效性
  - 引入实时流式计算模型：DW-RT （DAG）
  - 引入增量计算模型 Mini-Batch Computing：[Plan]
- 存储、访问优化[Doing]
  - Index
  - 列式存储
  - P.API性能优化

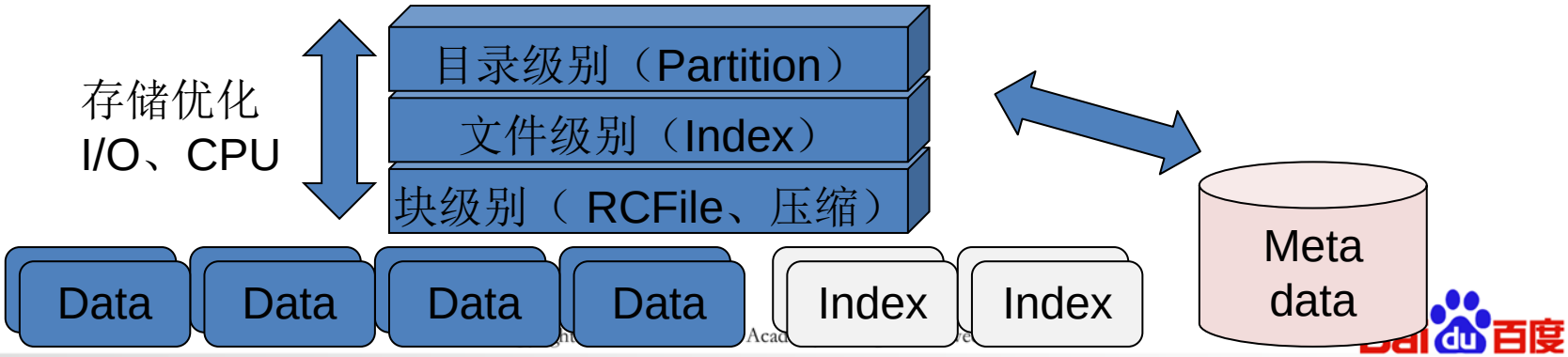
逻辑层次

主题:概念表

=

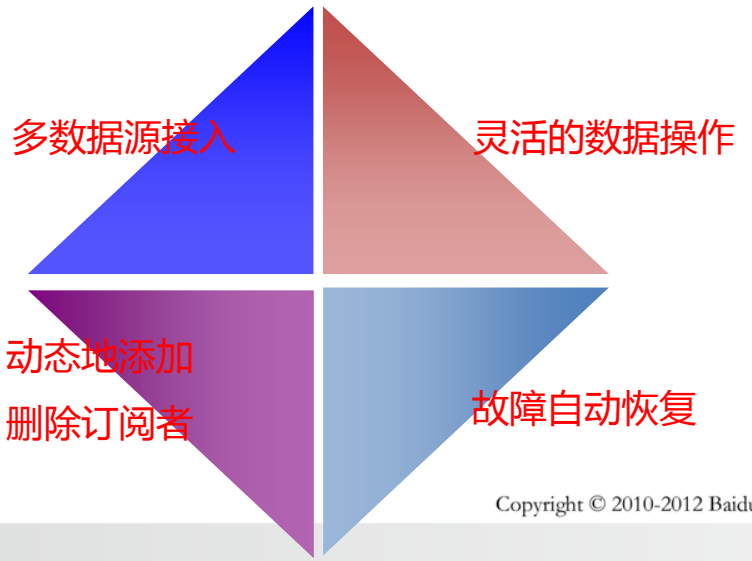


物理层次



ETL时效性：DW-RT

- 内核基于百度流式计算系统
  - 高性能DAG流式处理系统
  - Process Node ←→ Task tracker
  - Process Element (PE) ←→ Mapper/Reducer
  - Processor → N * PE （多并发）
  - 故障的PE会被自动重启或迁移，在恢复过程中丢失数据



- 灵活数据操作：
  - 物理源 → 逻辑源 → 新的逻辑源
  - SQL-LIKE、三种数据操作
    - **filter** …
    - **project** …
    - **union**
- 动态修改订阅者： 调研任务支持
  - 逻辑/物理源都可被订阅
  - 通过客户端订阅数据源，即时获得数据
  - 订阅者隔离
  - 动态扩容

DW-RT工作流图

HY

WISE

.....

nsclk

BigPipe

BigPipe

BigPipe

IMPORTER-PE For HY

IMPORTER-PE For WISE

.....

IMPORTER-PE For nsclick

RT-PE

RT_SELECT

2012-6-27

.....

EXPORTER-PE for UFS

EXPORTER-PE for Experi

.....

EXPORTER-PE for RP

KV Storage

Bai du 百度

DW TODOs

- 提高基础数据覆盖率：
- 数据模型完善：
- 提供更好/更高效的ETL 框架: DW ETL，DM ETL，统计ETL，Data Feed ETL，DataMart ETL
- 新的存储技术的考虑，如Column IO etc.
- 增量小批量处理技术在ETL中的应用
- 跨机房/集群的数据存储和访问，数据同步
- 异构数据的ETL和访问

OLAP

OLAP设计目标

- 问题：
  - 提供用户数据报表后台引擎：商业产品/用户产品分析报表
  - 定制报表 + OLAP 报表

- 要求：
  - 95%查询响应时间 < 10 sec
  - 可以对最细粒度数据进行统计和访问
  - 7*24h，服务可用性>99.999%
  - 容量100T+

OLAP问题域

- 星型多维数据模型：
  - Fact Table：维度id与事实（指标）
  - Dimension Table：维度id与分层维度描述
- 计算模型
  - Rollup/Drilldown
  - 计算相对简单
    - 事实表：select / group by / aggregate / filter / sort
    - 维度表：select / filter
    - 事实表结果与维度表结果大小表Join
- 思路
  - 并行查询 + 行列式存储

OLAP问题域(cont.)

- 优化方法
  - 单机扫描能力做到极致
  - 多机并行汇聚查询
  - 减少扫描量
    - 物化视图：对某些维度的组合聚合结果的预计算
    - 索引
    - 压缩：根据数据特点的高压缩比算法
    - 按列存储
  - 提高写入速度
    - 上游接小批量实时计算系统
    - 增量数据更新

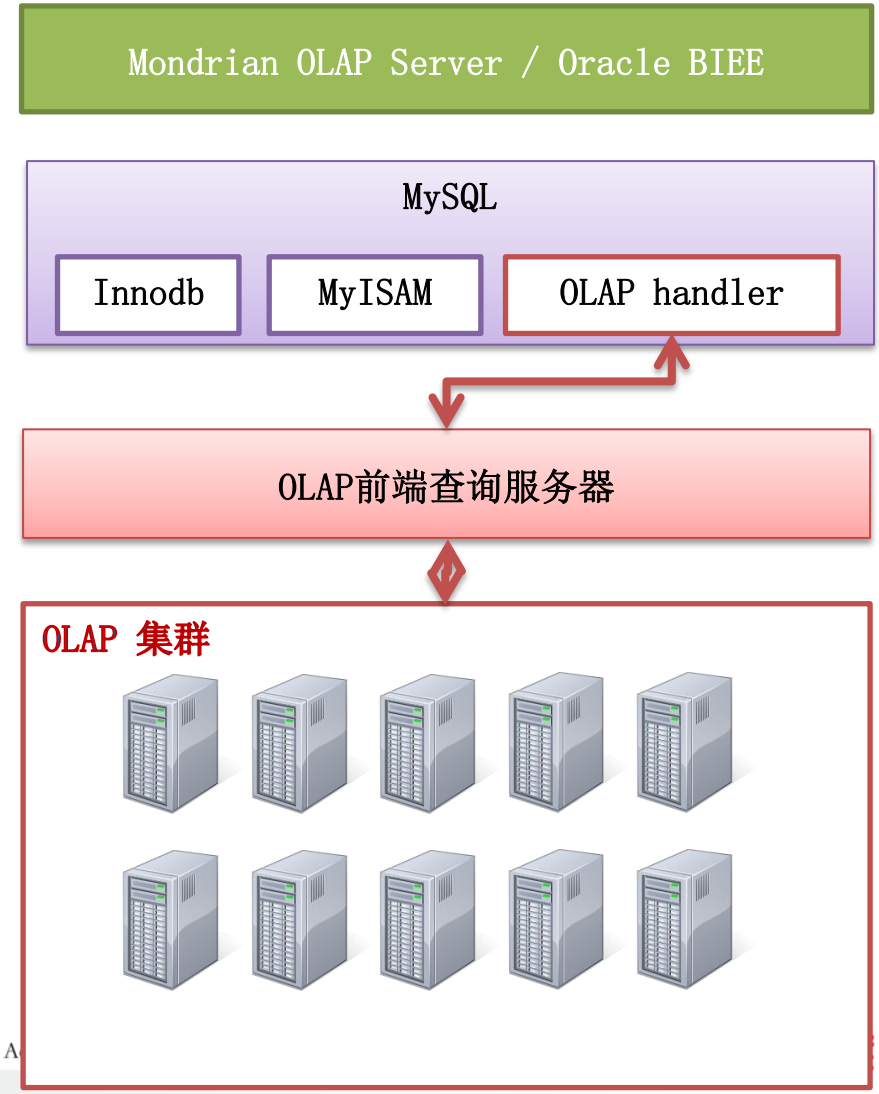
OLAP Engine Intro

- 核心技术
  - 可扩展性：
    - 数据分布式存储
    - 行列组织：
      - Row-group：Sorted Keys Range +数据量切分（1GB）
      - Colum: 压缩比>15:1
  - 并行查询
    - 自动选择最优物化视图
  - 支持SQL92、和Mysql结合
- 设计原则
  - 简单有效
  - 业务需求驱动
- 功能
  - 多视图、多索引
  - Schema change
  - 多版本视图间Transaction更新

OLAP Engine Intro(cont.)

- 系统性能指标（100节点）
  - 规模：总量500T；更新1T/Day;
  - 日查询量：2亿；
  - 最大导入速度：300MB/s
  - 单节点最大扫描速度：1.5GB/s
  - 平均响应时间：msec；大查询<10sec
  - Max QPS：5000;
  - 查询成功率:99.999%

- TODOs
  - 计算层：分布式排序、Join
  - 存储层：按列存储
  - 运维：数据恢复流程优化



谢谢

Bai du 百度