DTCC 2010-2022

第十三届中国数据库技术大会

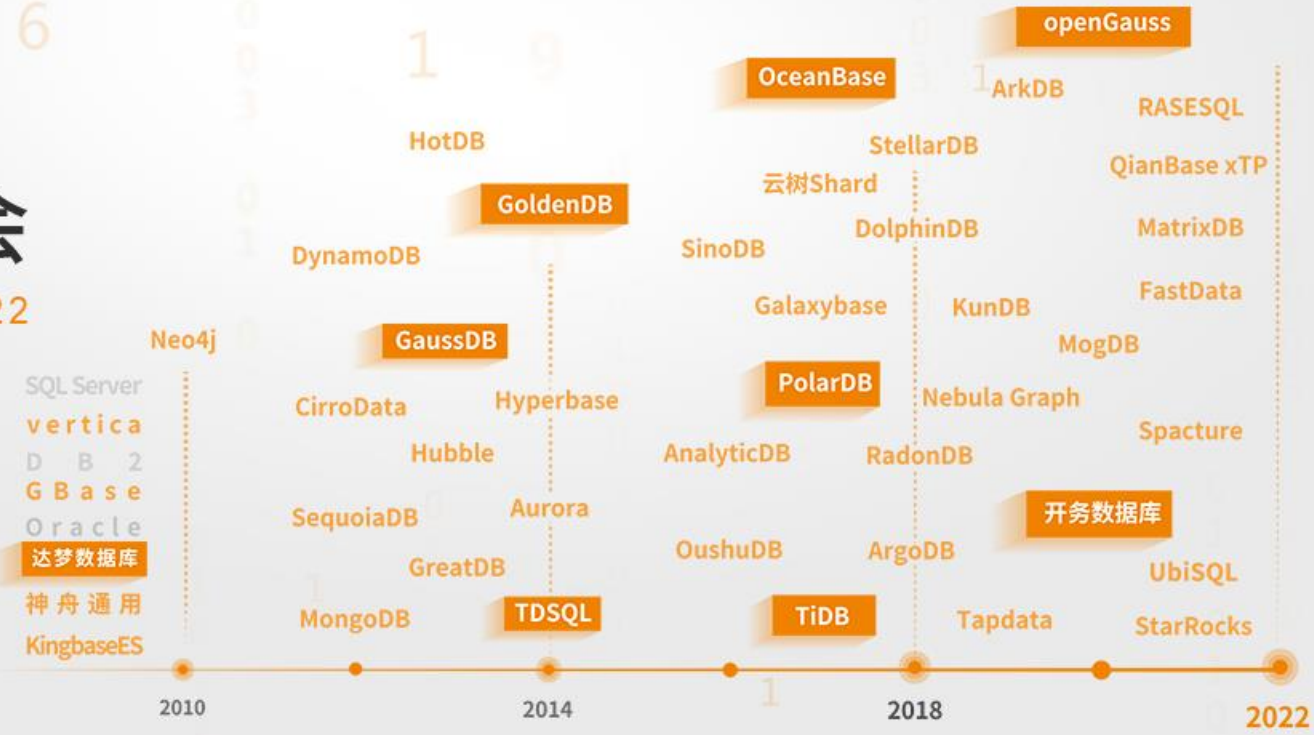
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



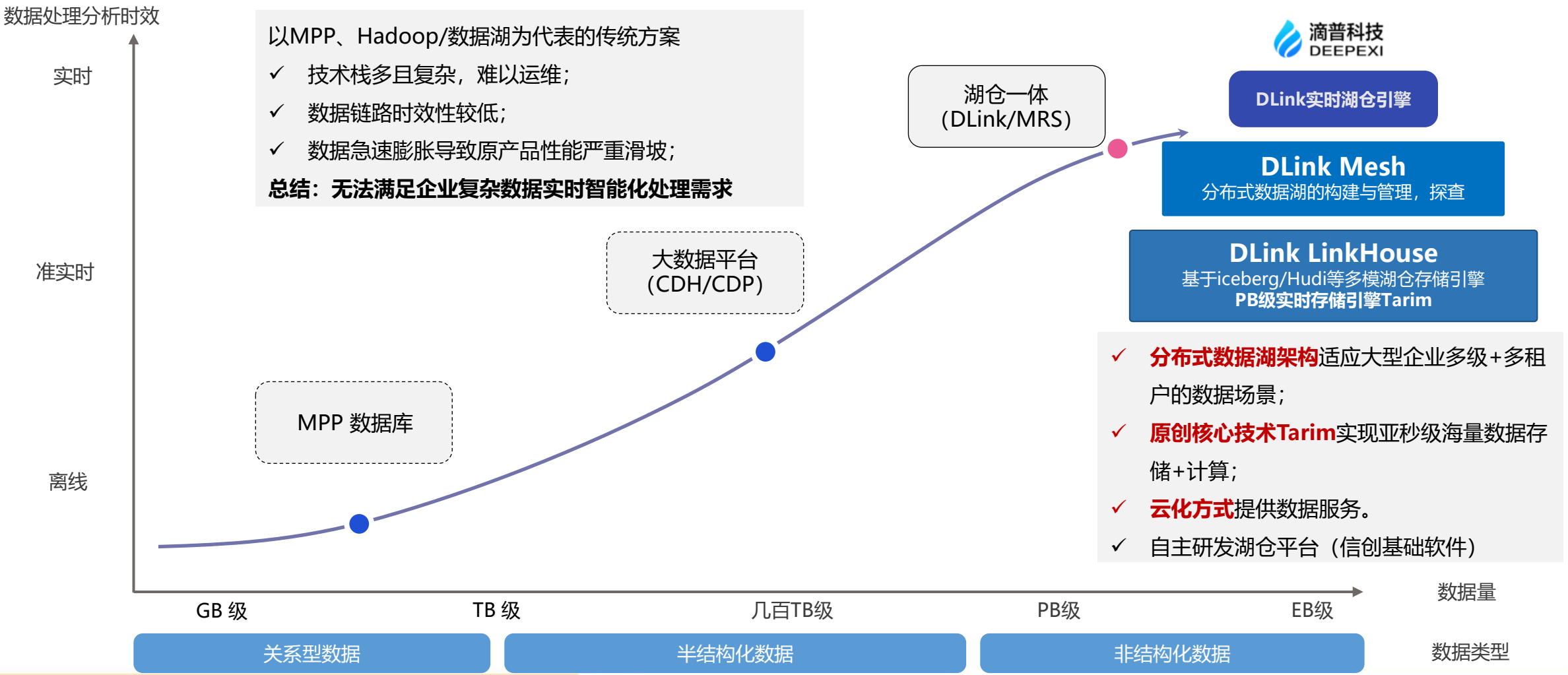
FastData DLink 实时湖仓引擎的架构设计与实践

北京滴普科技有限公司
FastData DLink PDT 总经理　冯森

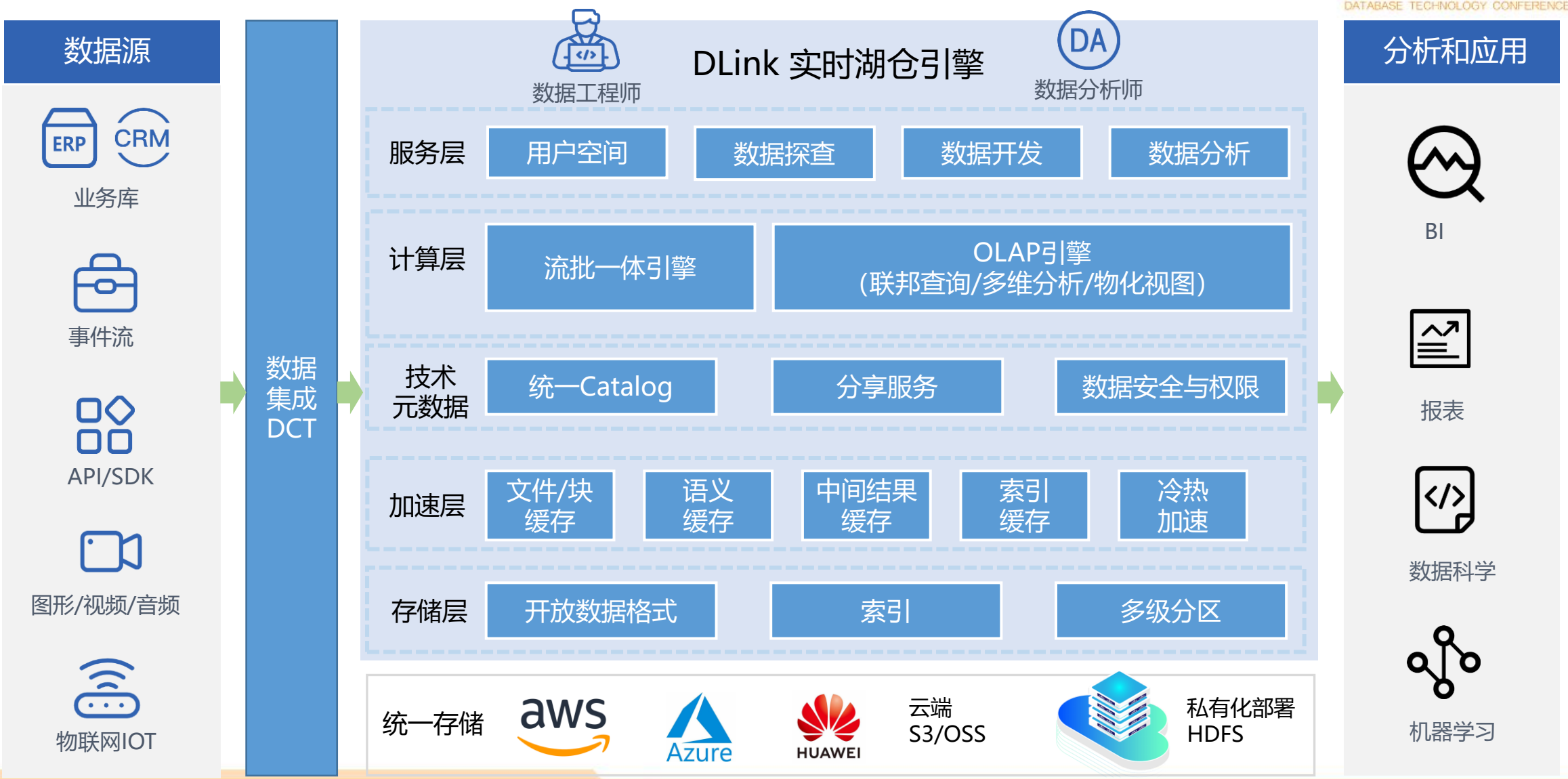
目录

DLink 架构介绍

湖仓一体平台演进趋势

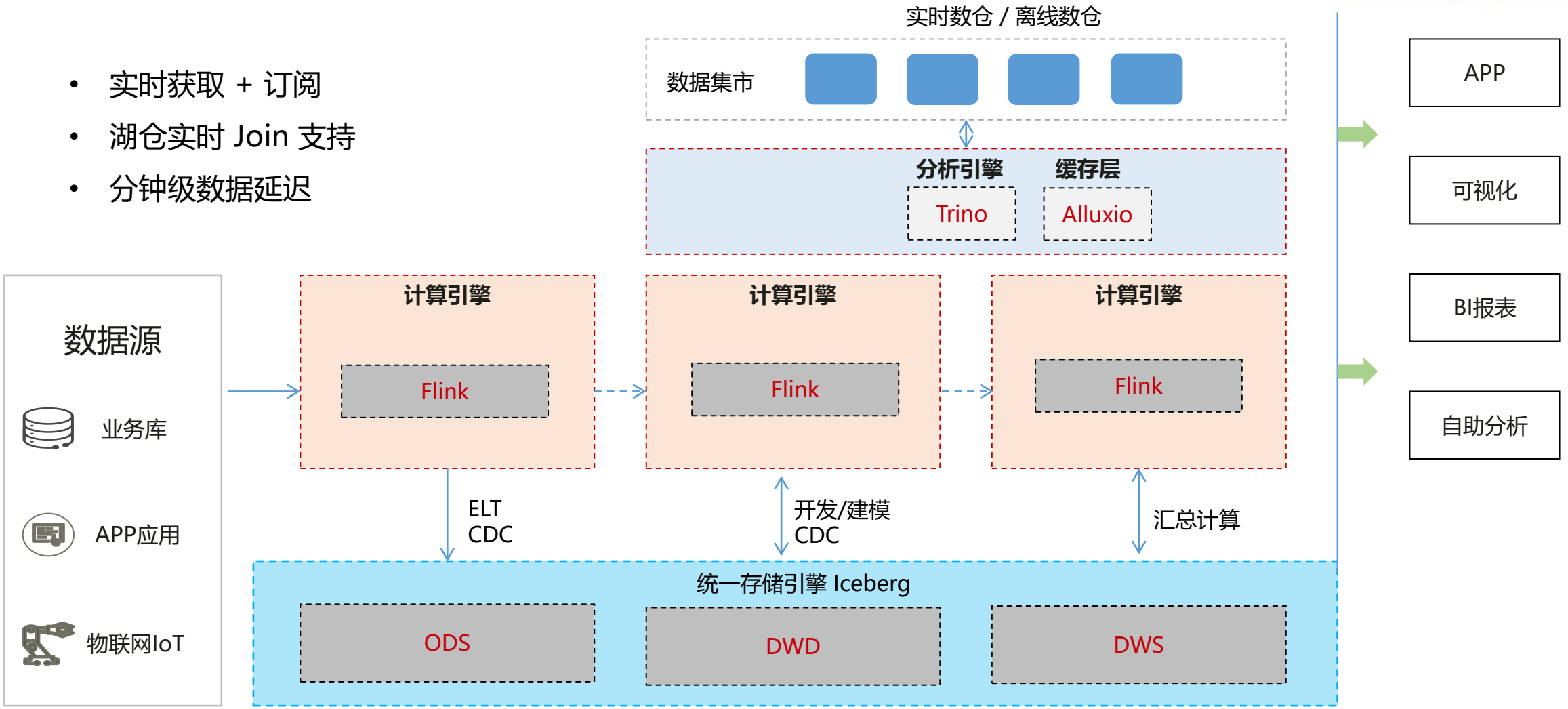

数据处理分析时效
实时
准实时
离线
以MPP、Hadoop/数据湖为代表的传统方案
✓ 技术栈多且复杂，难以运维；
✓ 数据链路时效性较低；
✓ 数据急速膨胀导致原产品性能严重滑坡；
总结：无法满足企业复杂数据实时智能化处理需求
MPP 数据库
大数据平台
(CDH/CDP)
湖仓一体
(DLink/MRS)
滴普科技
DEEPEXI
DLink实时湖仓引擎
DLink Mesh
分布式数据湖的构建与管理，探查
DLink LinkHouse
基于iceberg/Hudi等多模湖仓存储引擎
PB级实时存储引擎Tarim
✓ 分布式数据湖架构适应大型企业多级+多租户的数据场景；
✓ 原创核心技术Tarim实现亚秒级海量数据存储+计算；
✓ 云化方式提供数据服务。
✓ 自主研发湖仓平台（信创基础软件）
GB 级
TB 级
几百TB级
PB级
EB级
数据量
关系型数据
半结构化数据
非结构化数据
数据类型

DLink 产品架构图


数据源
ERP
CRM
业务库
事件流
API/SDK
图形/视频/音频
物联网IOT
数据集成 DCT
数据工程师
DLink 实时湖仓引擎
DA
数据分析师
服务层
用户空间
数据探查
数据开发
数据分析
计算层
流批一体引擎
OLAP引擎
(联邦查询/多维分析/物化视图)
技术元数据
统一Catalog
分享服务
数据安全与权限
加速层
文件/块缓存
语义缓存
中间结果缓存
索引缓存
冷热加速
存储层
开放数据格式
索引
多级分区
统一存储
aws
Azure
HUAWEI
云端 S3/OSS
私有化部署 HDFS
分析和应用
BI
报表
数据科学
机器学习

DLink 实时湖仓架构图

- 实时获取 + 订阅
- 湖仓实时 Join 支持
- 分钟级数据延迟



DLink 产品关键特性

通过DLink实现流批一体、湖仓一体、实时数仓，同时整合了数据湖与数据仓库的优点

云原生	云中立架构，支持计算资源弹性扩缩容，存储高可用，自动监控，可独立部署任意规模的工作负载，满足快速增长的业务需求
流批一体	统一数据基础架构与开发逻辑，同时支持系统数据流作业与批作业，统一Connector，降低学习、使用、维护成本。
多模数据的存储和管理	支持结构化、半结构化、非结构化数据存储，内置包括S3/OSS/HDFS/OBS的多模态存储引擎与分级存储；支持Parquet、ORC行列存储兼顾，支持隐藏分区和分区进化
存算分离	存储计算分离架构，可根据业务特性动态升降配和扩缩容；支持直接读取离线数仓数据，系统负载均衡调度更加灵活，利用率更高，并以更低成本交付部署生产系统。
安全与隐私	实现了数据端到端加密与落盘再加密，统一身份鉴别、访问控制、安全认证，支持租户隔离、安全审计与数据全生命周期管理，支持跨区域实时灾备，全面保障数据安全与隐私。
统一元数据管理	集成大规模元数据管理，表和分区的所有元数据通过统一的元数据访问，并持久化，使用高性能表格管理，可向多计算引擎添加表，单表可达数十PB，保障跨语言和兼容性。
统一数据接口	多种异构数据的实时或离线批量同步传输与计算，解决了架构复杂度高，数据格式不统一等问题，打通多业务系统数据，满足多种数据查询与应用。
即席数据查询	支持海量数据即席查询，支持多catalog的联邦查询，支持在线编写 SQL 语句、语法检查、调试和发布部署，并提供线上运维管理。整个过程无需复杂编程，降低了用户使用实时流计算的门槛。
统一工作空间	机器学习（Python，R，各种机器学习库），强SQL标准支持（Spark SQL、Flink SQL、Hive SQL等），其他工具对接（BI工具、IDE等）
支持事务与Schama	基于Iceberg支持ACID事务能力，支持同时读取和写入数据，同时提供upsert/merge into的能力；支持schema的执行和演变，支持数仓的星型/雪花模型；

基于使用场景

租户管理	可创建多个项目，为每个项目分配单独资源和权限，管理数据源链接，多个项目间资源和权限完全隔离。
空间概览	提供通用运维监控能力，支持通过WEB、外接消息机制对故障进行告警的能力，以及是否能够以日志记录告警信息。
数据探查	提供已连接的外部数据源信息对应数据源存储的数据结构以及数据样例，探测数据连接连接性，提供自动生成ddl功能
湖仓管理	提供数据集模块，统一管理湖内数据；提供过期快照删除、小文件合并、删除孤儿文件等表运维功能，支持参数配置、策略配置及手动触发。
实时计算	提供可视化Dlink SQL作业提交和任务管理能力。支持在流上执行类SQL任务，SQL能力至少包括：过滤、转换、基于窗口的计算能力、提供窗口数据的统计能力、关联能力、流数据的拆分与合并。
数据分析	支持即席查询，交互分析，物化视图，并支持多catalog的联邦查询，并可以JDBC、HTTP等方式支持数据分析结果输出
机器学习支持	支持非结构化数据入湖、管理及任务运维，支持特征存储，向量索引模型和特征实时查询。
作业管理	支持多种作业类型：SQL 作业、JAR 作业等。支持在每个项目空间内提供作业管理运维、作业包上传等功能。并可为作业配置CPU及内存资源。

DLink 核心功能

DLink 优于开源功能

优于开源Iceberg

BloomFilter 索引	在等值查询和范围查询场景下性能大幅提升
Z-Order 数据排序	在多维数据分析场景下性能大幅提升
Hive 存量数据快速迁移	支持生成 Iceberg 元数据的方式对 Hive 历史数据快速迁移，避免数据搬迁
Iceberg CDC 能力	Iceberg 通过流读变更数据构建实时数仓的重要能力
小文件自动合并	通过内置合并策略，可以自动进行后台小文件合并，快照清理等
隐藏分区/计算列	支持 Flink 在iceberg上创建带有计算列/隐藏分区的表
支持 iceberg 维表	支持通过 iceberg 存储维表功能

优于开源Trino

Catalog热加载	支持动态加载catalog能力
支持 local cache	对 IO 密集型 query 进行性能优化
支持多租户	基于ranger支持多租户能力
支持物化视图	支持物化视图动态刷新
数据加密	支持通过masking的方式对数据进行加密
数据权限	支持库、表和字段级数据权限
支持CBO优化	支持基于 iceberg 统计信息进行优化执行计划

优于开源Flink

UI界面	DLink在UI界面集成了作业提交、管理运维、数据视图Metrics等能力
Flink 引擎支持多版本	支持 Flink1.12-1.14 版本
流批一体	Flink 支持 SDK 的方式提供被批调度能力
整库入湖	支持整库数据入湖，提升入湖效率
算子调优	支持 flink 算子自动调优、算子拆分，算子并发
数据连接	支持丰富的connector
支持 yarn/K8S	Flink 支持 on yarn 和 on k8s 资源调度

DLink 在开源社区的贡献

总PR 数： 48个

总Contributor ： 14位

Iceberg 社区

- PR总数： 31个
- Contributor总人数： 9位

Hudi 社区

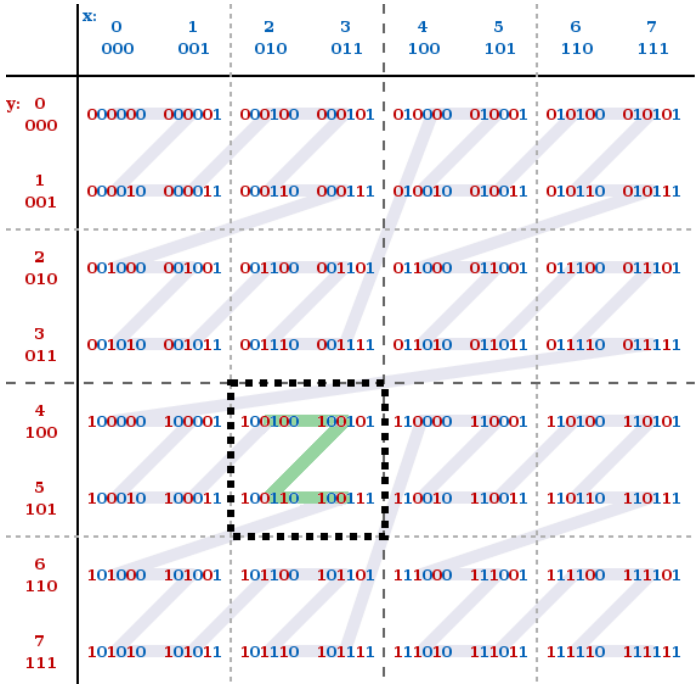
- PR总数： 13个
- Contributor总人数： 3位

Trino 社区

- PR总数： 4个
- Contributor总人数： 2位

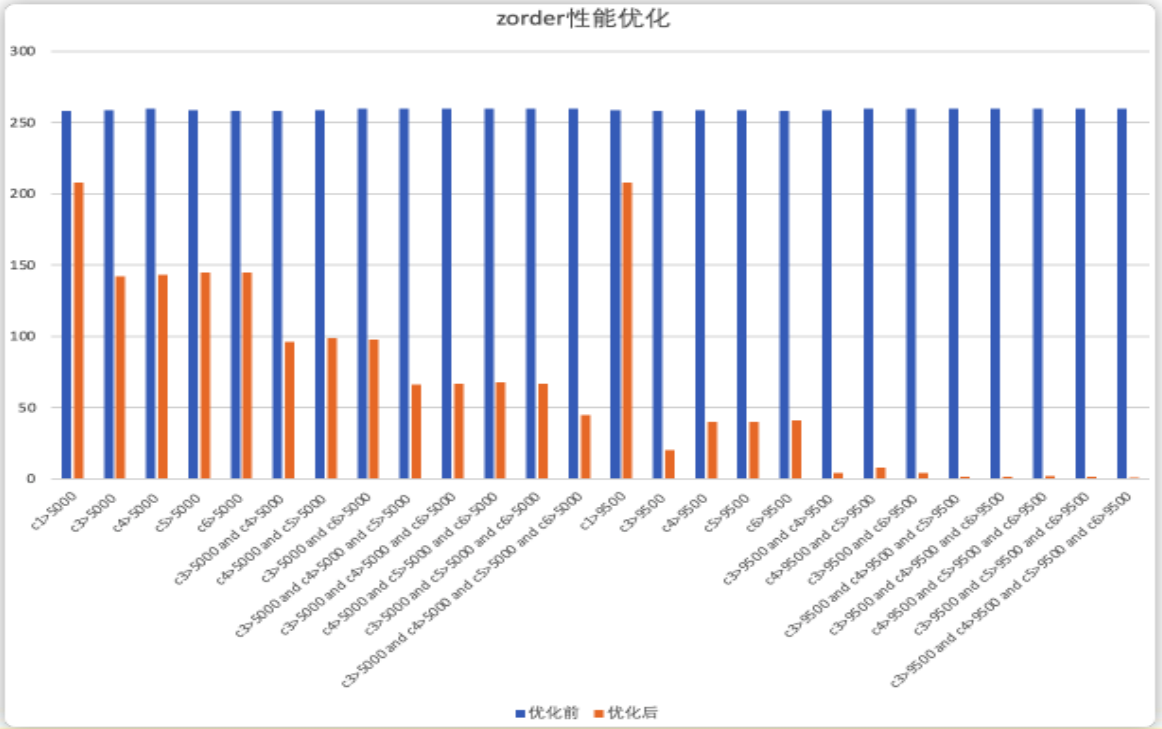
DLink Iceberg Z-Order 排序

Z-Order是一种特殊的将多维数据映射到一维的方法，如右图所示，对于一个二维的查询条件来说，无论对A还是对B进行范围查询，都能至少过滤掉50%的数据量。在多维分析场景性能会有大幅提升。



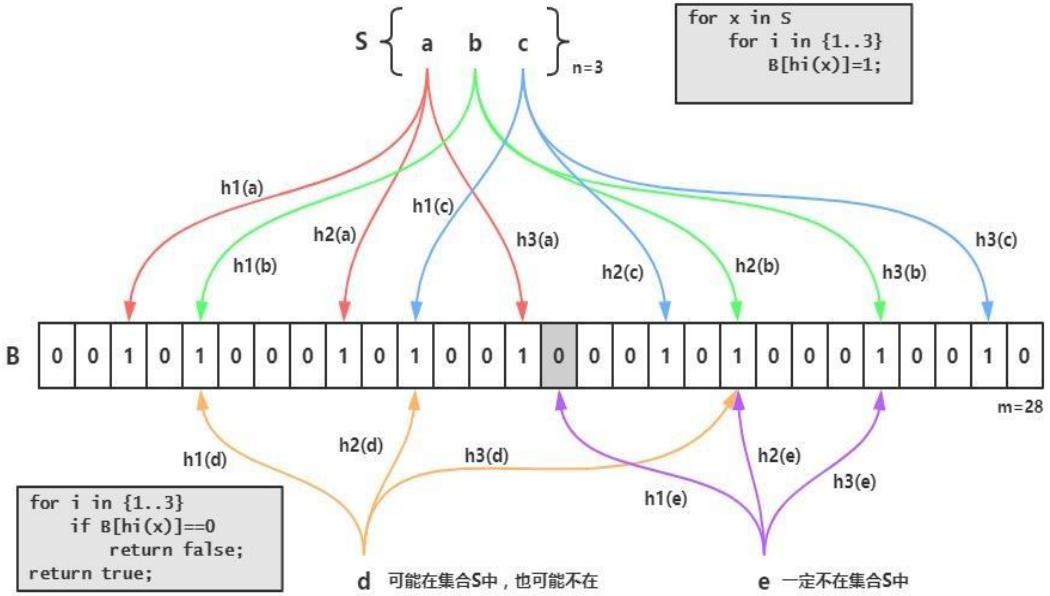
表格 1 两种不同顺序数据堆叠方式

序号	(A,B)自然顺序	Z-Order 顺序
0	a1,b1	a1,b1
1	a1,b2	a1,b2
2	a1,b3	a2,b1
3	a1,b4	a2,b2
4	a2,b1	a1,b3
5	a2,b2	a1,b4
6	a3,b3	a2,b3
7	a4,b4	a2,b4

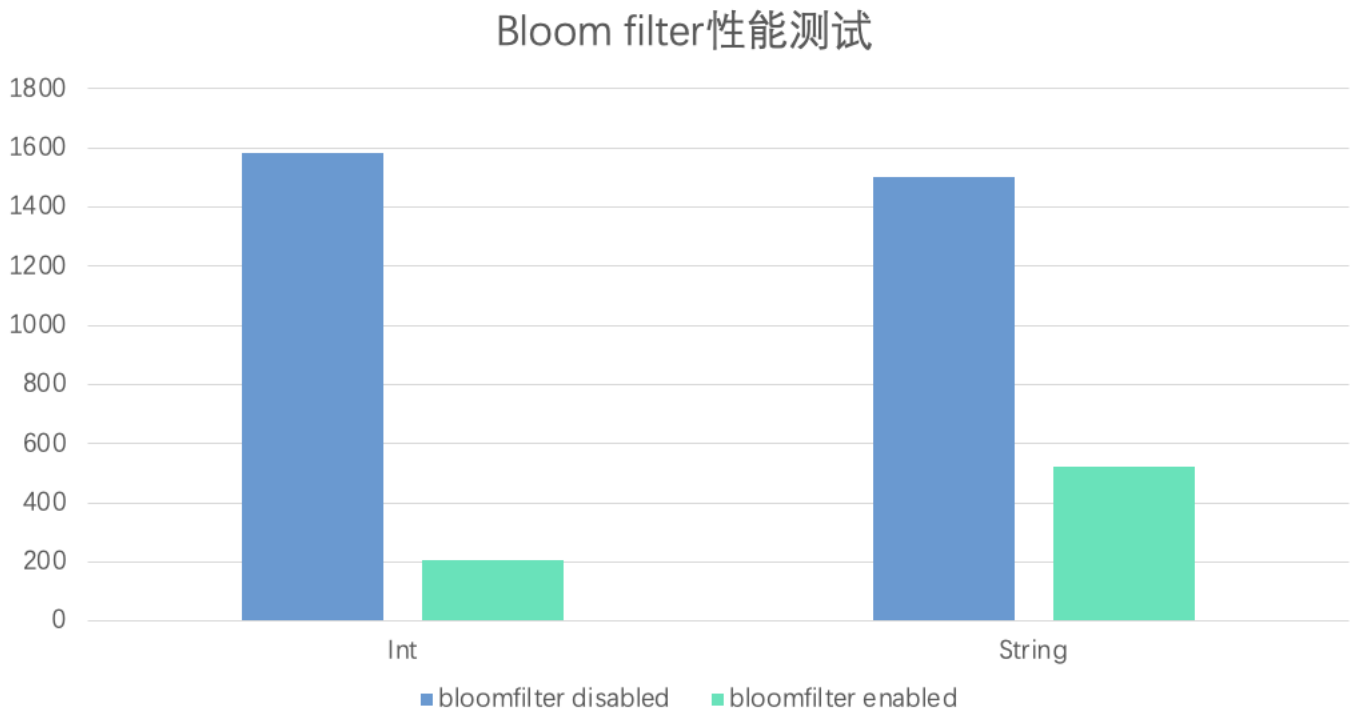


DLink BloomFilter 索引

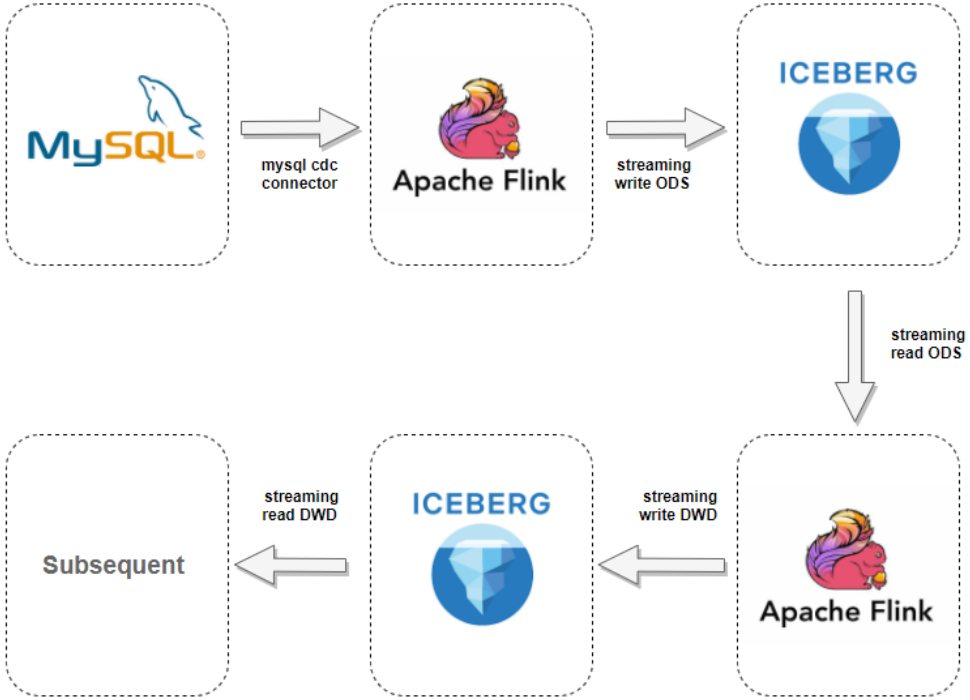
布隆过滤器可以用于检索一个元素是否在一个集合中。它的优点是空间效率和查询时间都远远超过一般的算法，查询性能很优

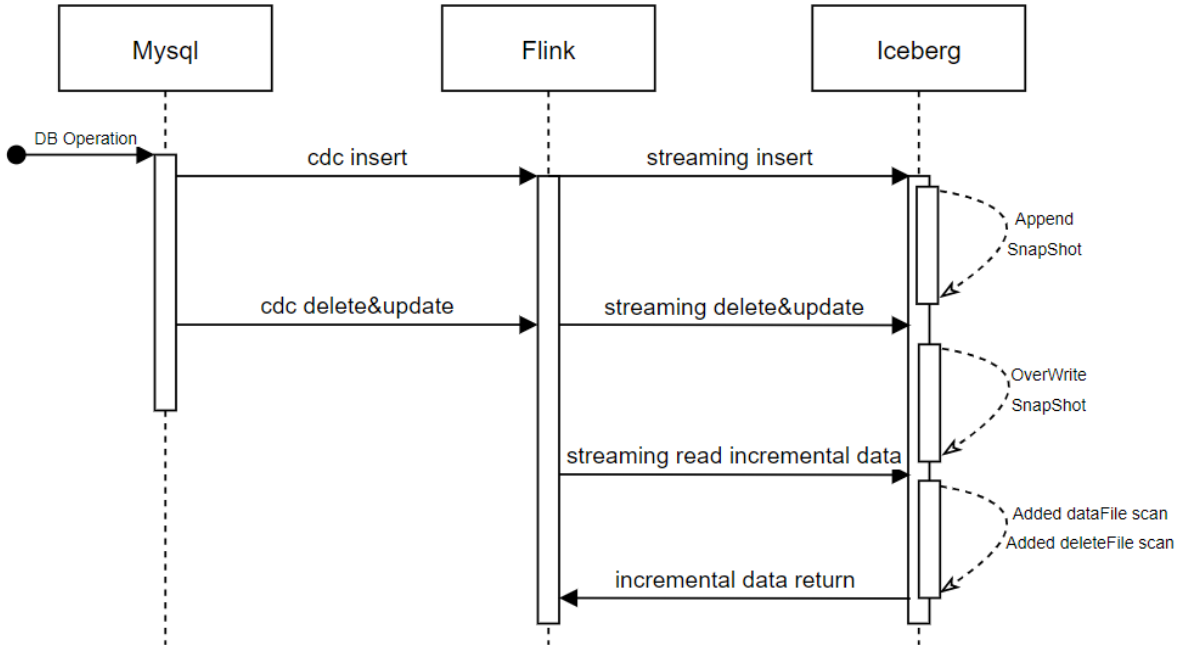


10亿数据规模点查测试：



DLink 支持 Iceberg CDC 流读 Insert/Update/Delete





streaming read sequence chart

SQL示例：select * from hive_catalog.test1.test35 /*+ OPTIONS('streaming'='true', 'monitor-interval'='1s')*/ ;

DLink 自动化合并Iceberg 小文件

DLink 在 iceberg 上构建后台自动合并小文件任务，让用户忘记繁重的表运维任务，让Iceberg开箱即用。

支持独立部署和应用

支持独立部署运维服务

无缝接入DLink

自动生成运维任务

定义任务生成的规则，根据监控结果生成运维任务

资源隔离

DLink运维任务的资源与业务资源隔离

运维任务之间资源隔离

多种任务调度策略

支持FIFO和任务权重优先策略

DLink Hive历史数据快速入Iceberg

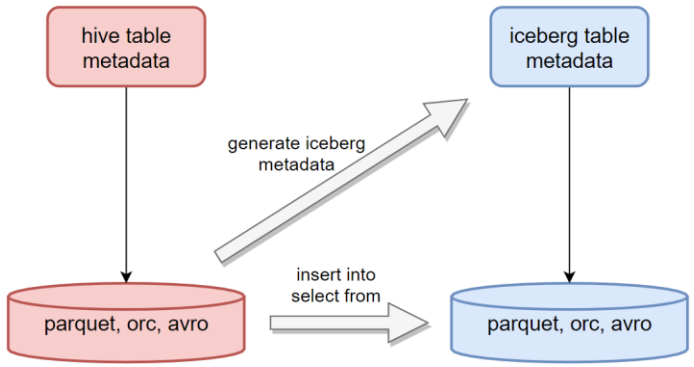
核心价值：

Hive数据可以在不迁移数据文件的情况下，直接构建Iceberg元数据，转换为Iceberg表，并在此基础上做了性能优化，降低了迁移成本。

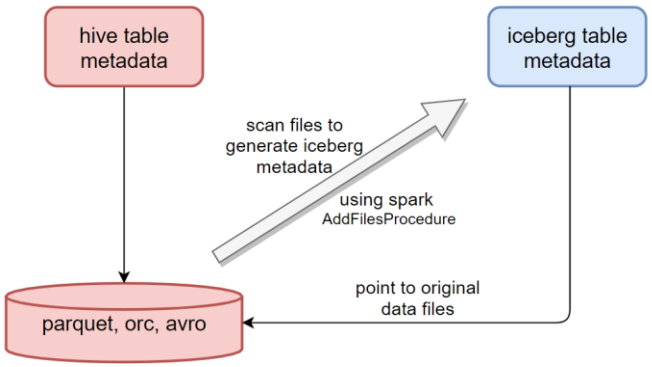
测试结果：

1、经过多次对比测试，说明数据轻量级迁移任务的执行还是很快速、稳定的，理论上与文件数量、大小、是否压缩均有关，但其中文件数量的影响最大，当分区文件数大到万级别时，迁移时间有变长，但也是在分钟级别。

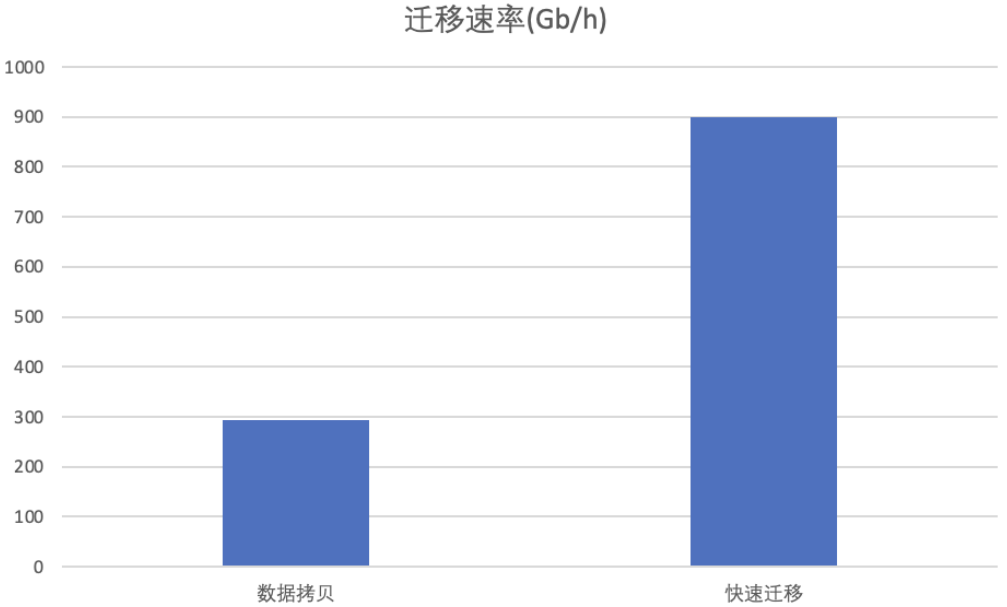
2、Hive快速迁移的效率是数据直接拷贝的三倍。



普通数据迁移

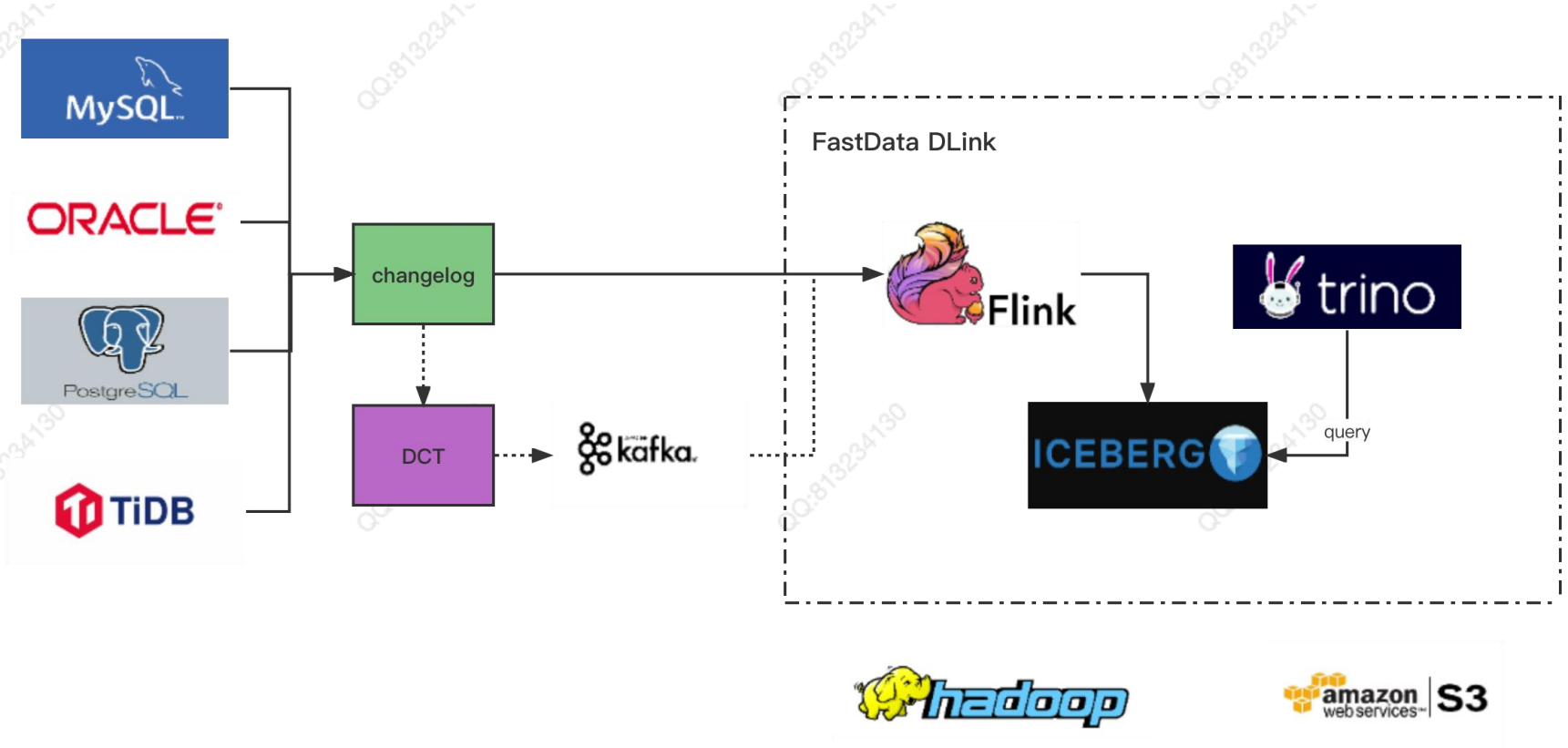


Hive表原地迁移到Iceberg表



DLink 整库数据入湖

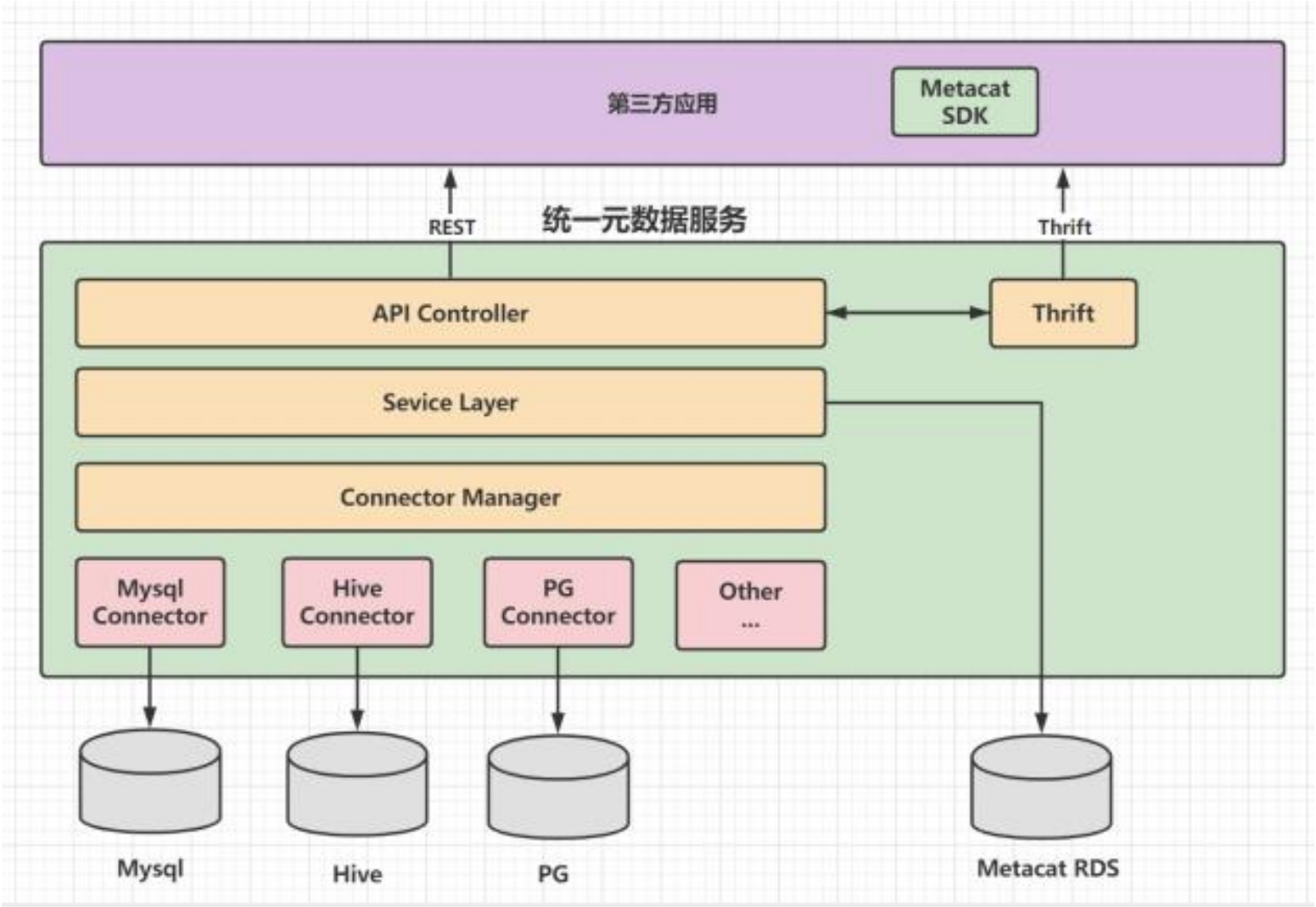
- 支持整库多表入湖
- 支持存量和增量数据的一体化入湖
- 支持运行时的DDL变更（新增列，新增表）
- 支持并行化入湖
- 支持断点续传
- 支持按照指定时间戳回溯



DLink 统一元数据服务

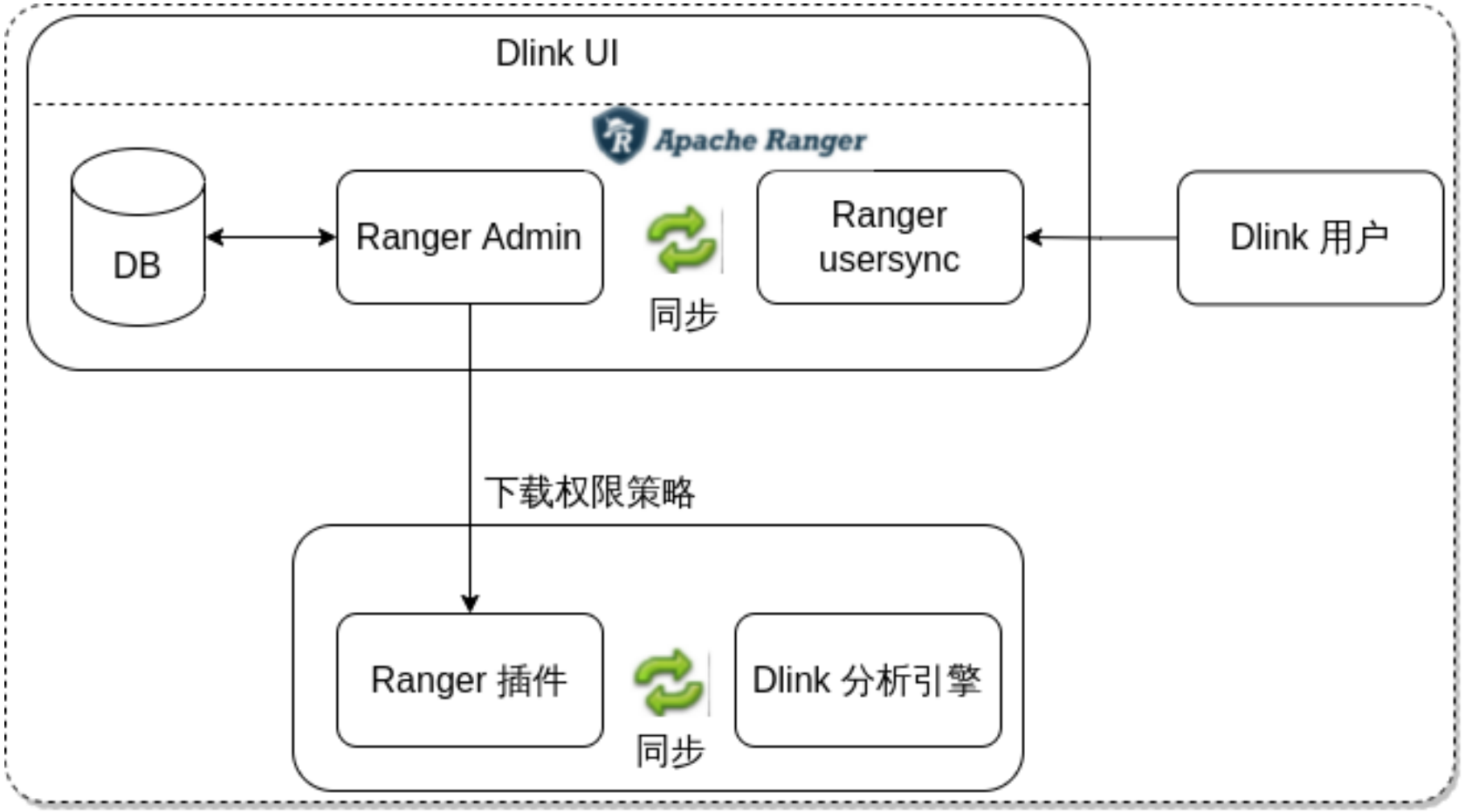
价值点

- 协议兼容：通过HMS多实例，支持多版本HMS协议（2.x/3.x）；
- 存储兼容：通过HMS多实例，支持多种存储介质（S3和HDFS）；
- 统一数据目录管理：提供针对不同数据源如Hive、Iceberg、MySQL 等的统一数据目录管理。
- 权限控制：所有DDL 操作鉴权在统一元数据层完成，不依赖引擎。
- 多租户：支持通过租户（和/或项目空间）对catalog进行隔离；
- 多引擎兼容：不同引擎能够使用同一份元数据，比如Spark、Flink、Hive、Trino不需要独立维护自己的元数据。



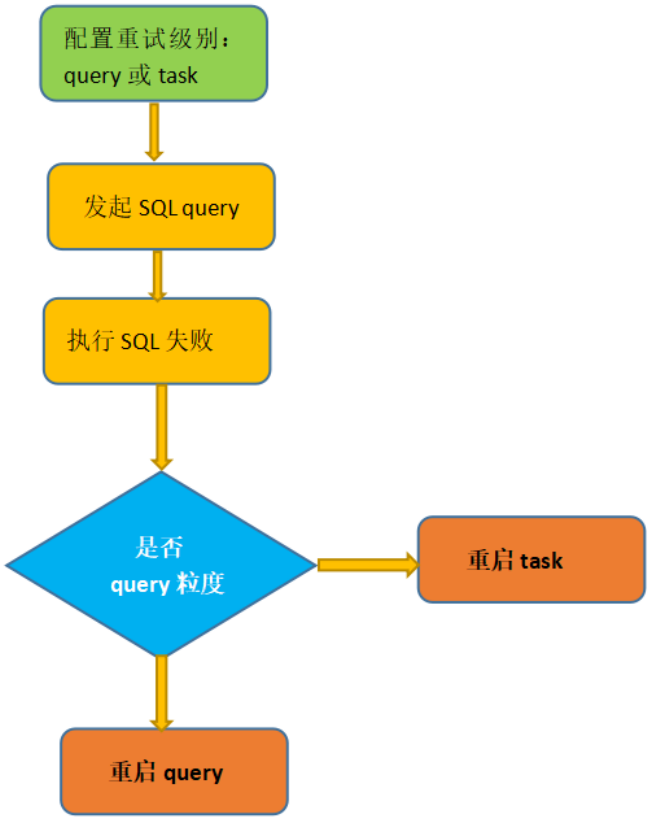
DLink 数据权限控制

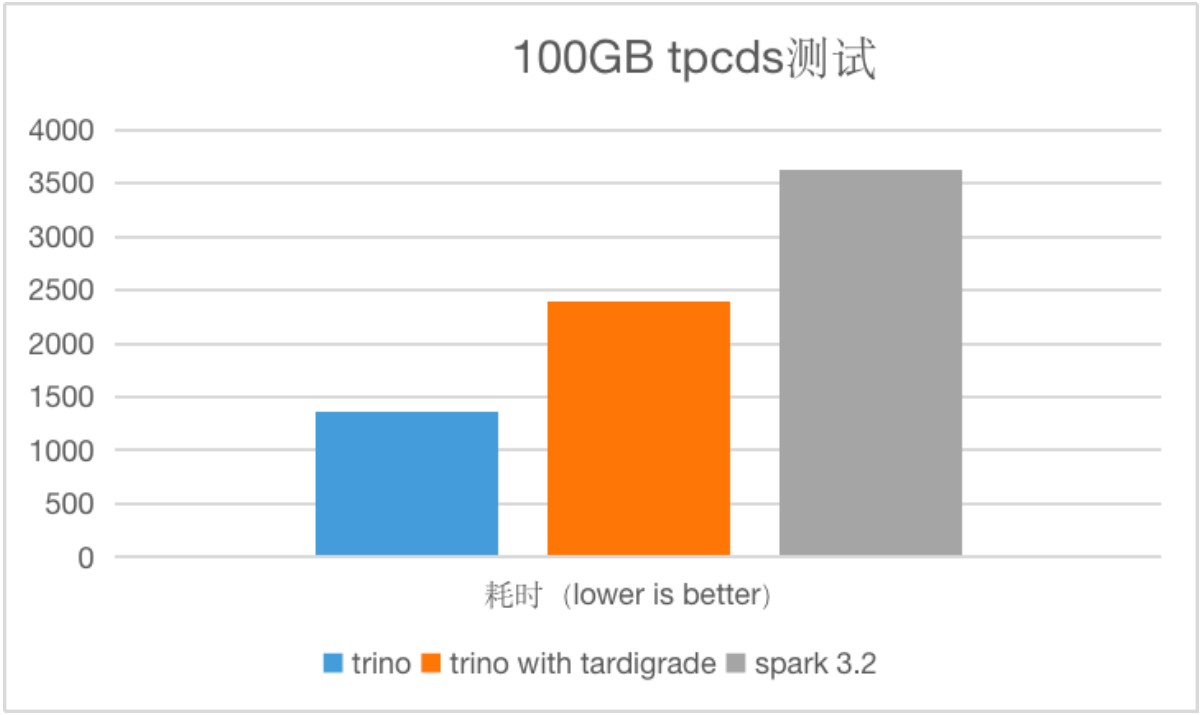
- 实现基于Ranger 对Iceberg对象如用户角色等权限访问控制，具体实现逻辑如左图
- Dlink 分析引擎执行 SQL 时，通过权限策略判断当前用户是否拥有执行权限
- 数据权限体系能和现有大数据权限系统打通或兼容



DLink Trino 支持批处理和容错机制

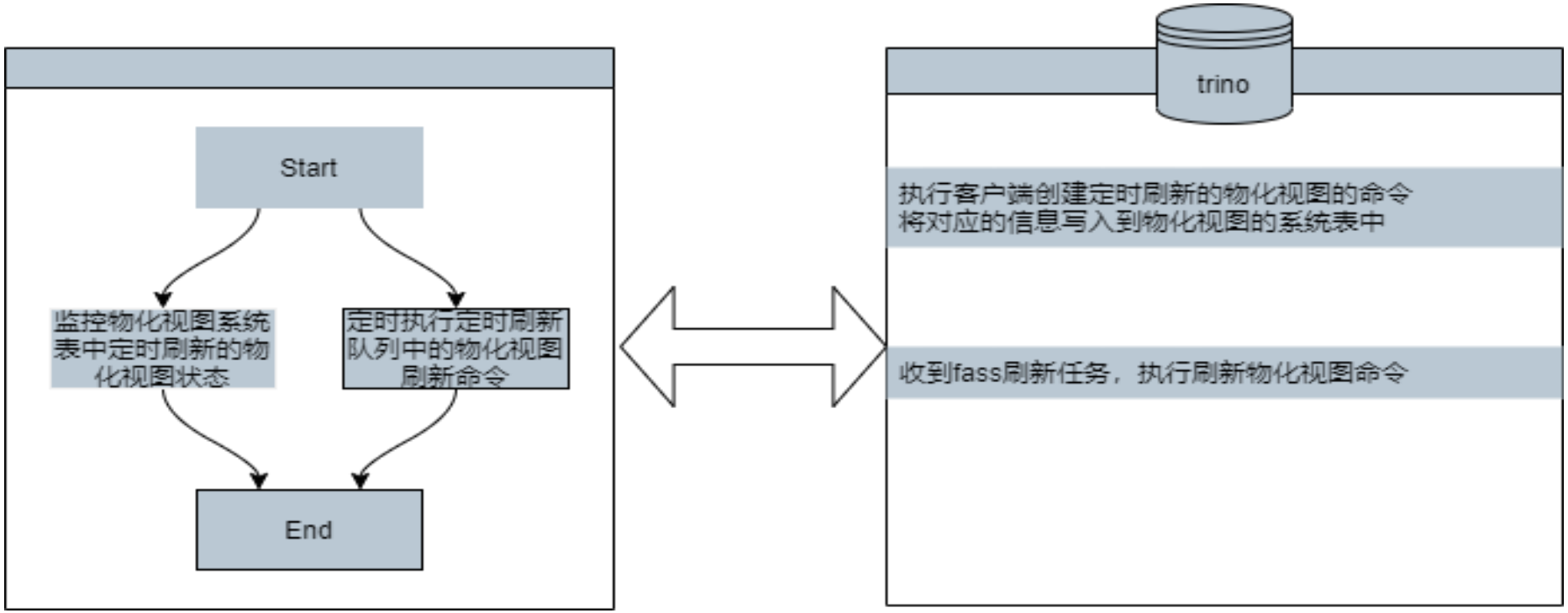
容错执行是 Trino 中的一种机制，它使集群能够通过在发生故障时重试查询或其组件任务来减轻查询故障。支持Query和Task级别重试，同时基于Tardigrade 实现了强大的批处理能力。





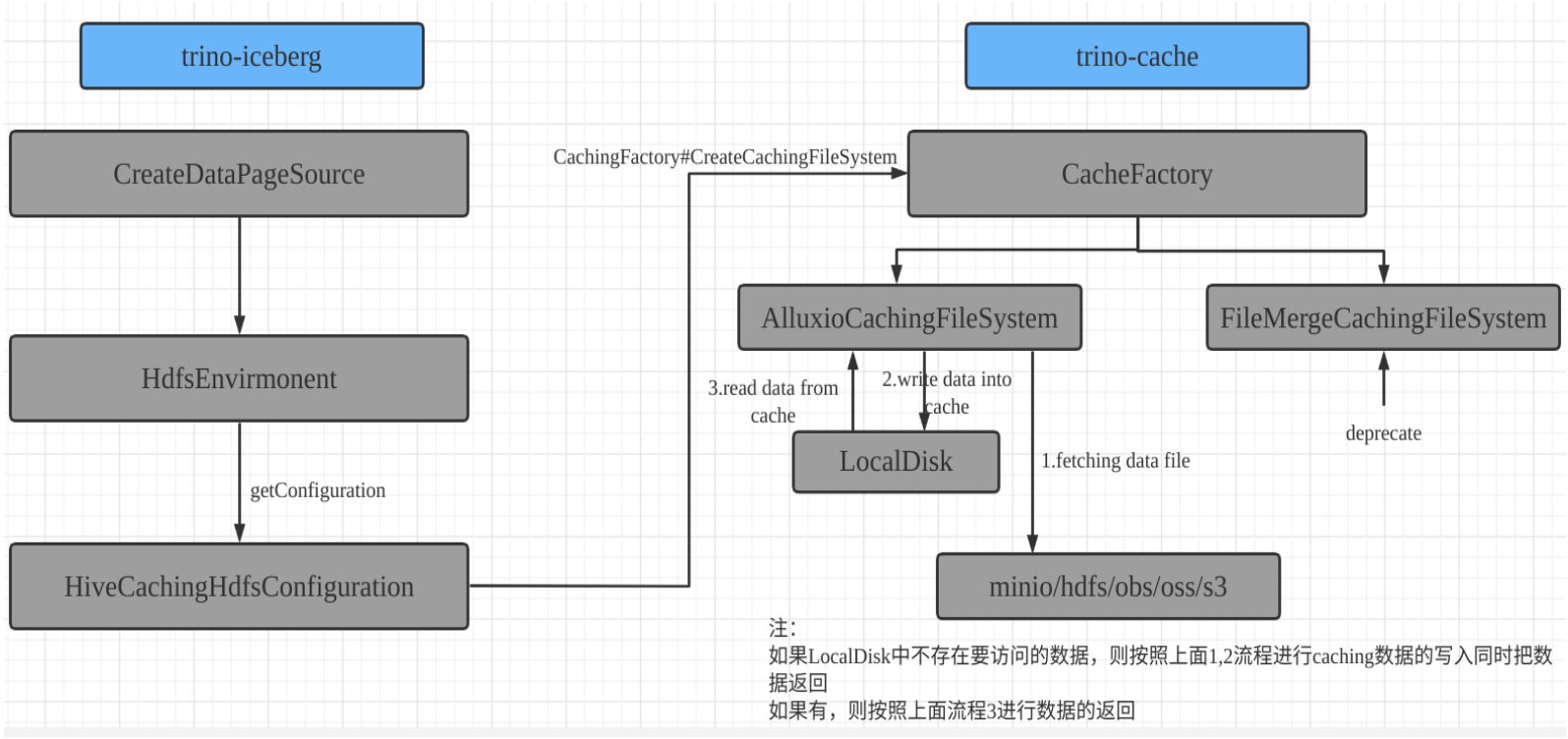
DLink 支持物化视图

物化视图的全量刷新很慢，其次当我们对物化视图关联的表进行dml操作的时候，数据会进行变化，但是物化视图无感知，导致物化视图查询的结果可能不是最新数据。定时刷新的好处在于我们刷新之后，把表的数据同步到了物化视图，使查询的数据不是旧的数据。



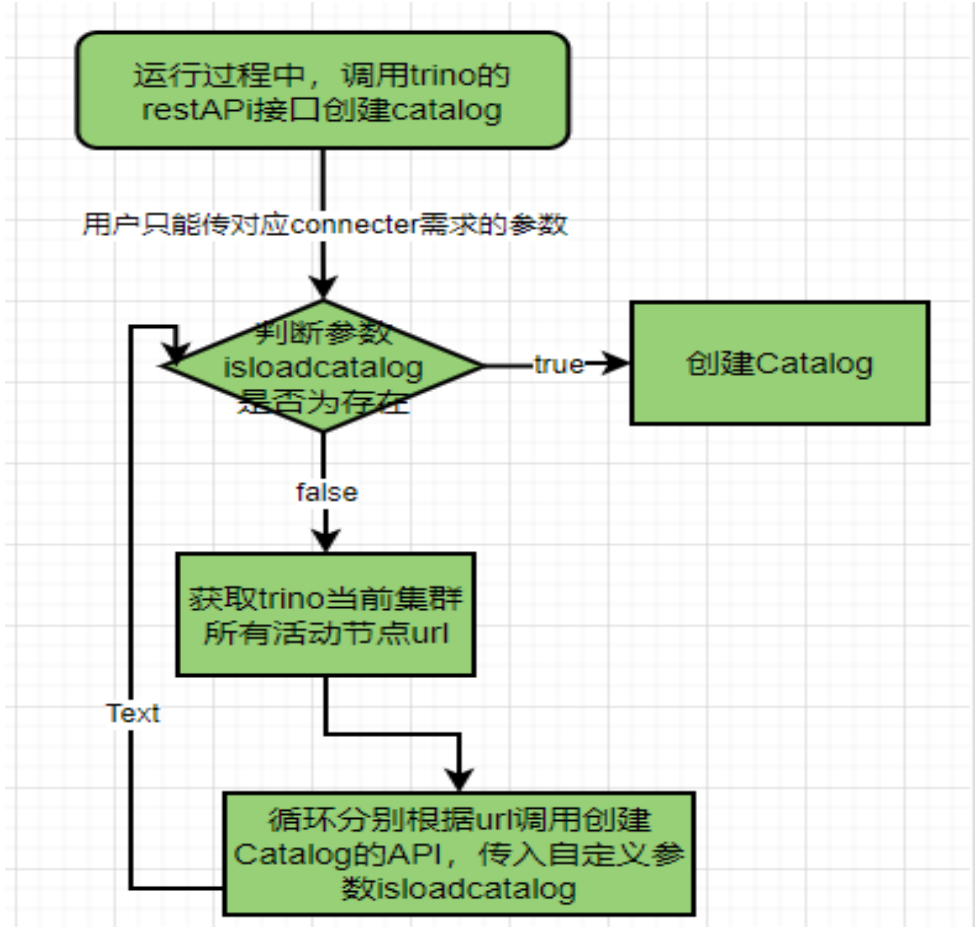
DLink 支持数据缓存

Alluxio Local data cache，轻量级仅本地节点访问的缓存，将数据缓存在计算Node的本地SSD中，不考虑集群节点间数据共享，依赖于soft affinity schedule，增加缓存命中率，尽量本地node处理本地的数据。



DLink Trino Catalog 热加载

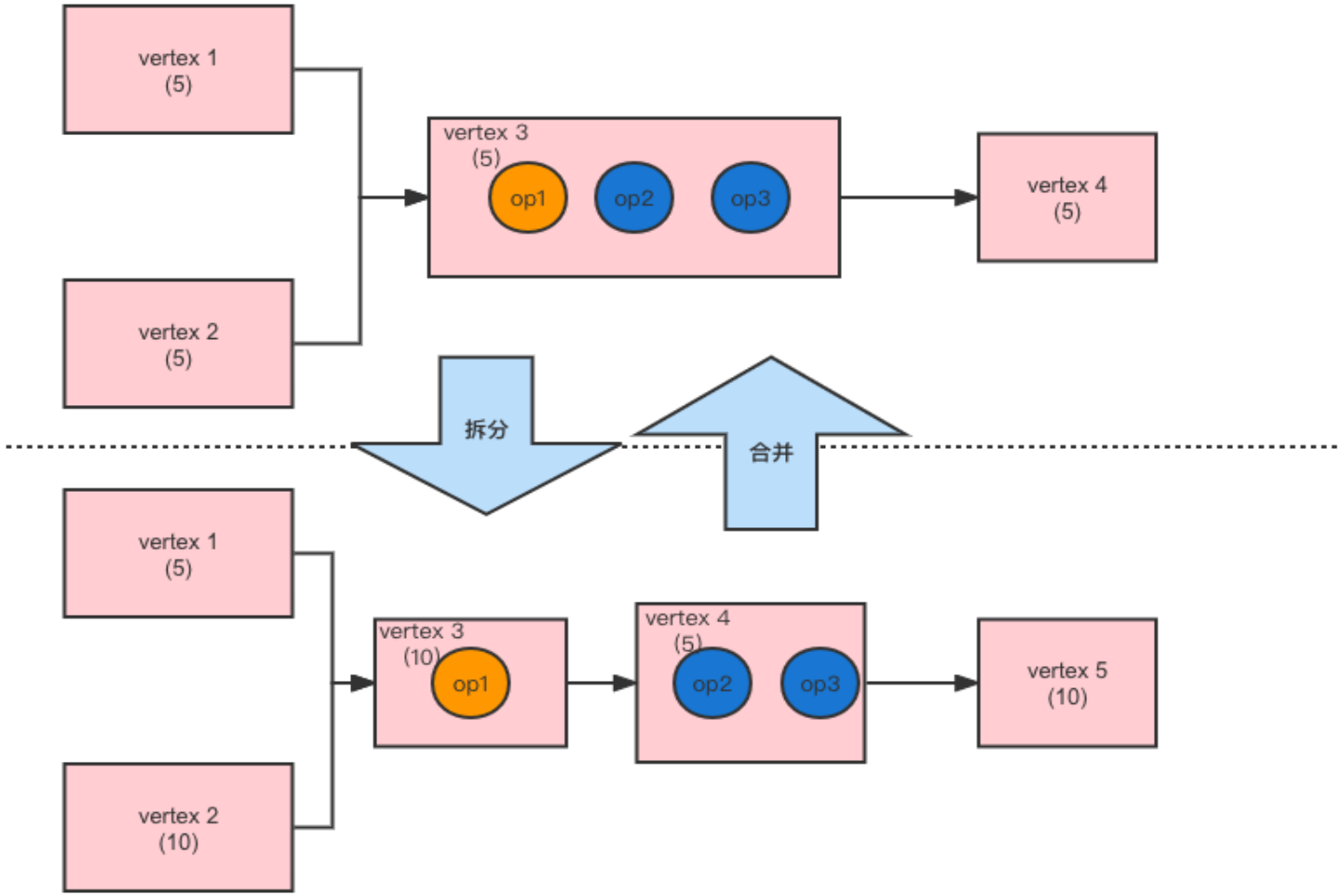
Catalog热加载主要解决了在Trino运行过程中，可以通过注册的方式，给Trino动态增加Catalog，无需重启。



左图：Catalog热加载的流程.
大体思路是:
1.扩展Trino的restApi接口，新增注册Catalog接口；

2.Catalog注册接口，在接收到用户请求后，先查找到当前Trino集群所有活动的节点。获取节点url，然后循环调用每个节点的注册CatalogAPI，分别热加载Catalog；最好再通过Announcement进行所有节点状态同步。

DLink 支持Flink Job 算子并发拆分和调优



- 方便排查瓶颈算子
- 节约资源
- 提升作业性能

DLink 支持在湖内构建维表和缓存加速

- 维表join的批量化加快join速度

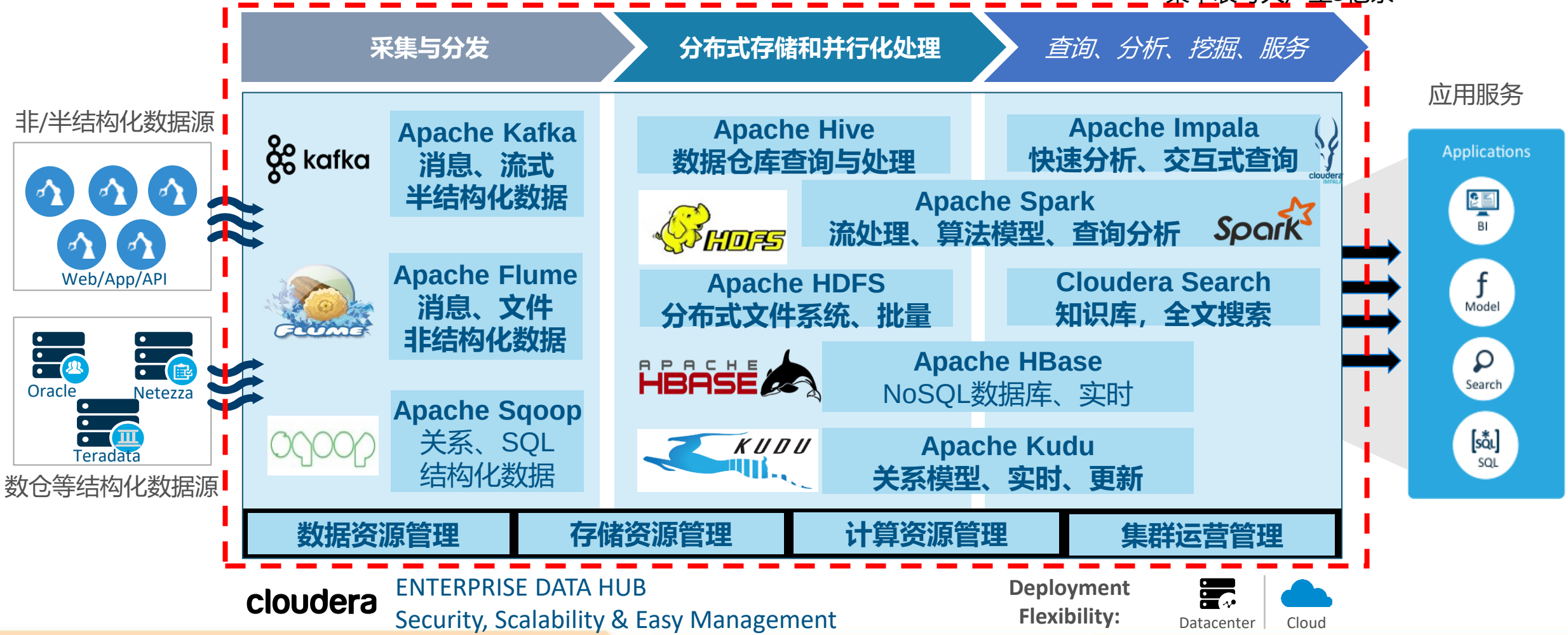
```
32  create table dim_table (
33    id int,
34    name string,
35    age int,
36    gender int,
37    phone string,
38    address1 string,
39    address2 string,
40    company string
41  ) with (
42    'connector' = 'redis',
43    'url' = 'ip:port',
44    'user' = 'xxxxx',
45    'password' = 'yyyyyy',
46    'batchLookupMaxCacheSize' = '10000', --最多缓存10000条
47    'batchLookupMaxWaitTime' = '5' --单位 s
48  );
```

DLink 落地实践

某存储客户数据能力建设现状

- xx存储 现有大数据平台以CDH（6.3.2）技术栈为主，基于本地IDC物理架构构建
- 沿用业界经典的Lambda架构，以离线采集+离线数仓为主核心技术架构

数据量：
20万亿条数据，xPB级
某单表每天产生5亿条



cloudera ENTERPRISE DATA HUB
Security, Scalability & Easy Management

Deployment Flexibility: Datacenter | Cloud

某存储客户关注点

客户需求

稳定性、可用性

- 希望能做到大数据组件到数据本身的全链路监控

灵活性、兼容性

- 各技术组件支持灵活的组合，需要与现有的CDH兼容
- 随着业务的发展，数据的增加，支持横向扩展

数据准确可靠

- Hive 历史数据快速迁移入湖
- 历史数据和新增数据去重
- 流式任务出现异常后，需要有相应的补数方案

解决方案

DataOps理念

- DataOps理念，提供可靠的数据采集到上线运维的流程监控，实现数据的持续稳定交付

湖仓一体架构

- 支持 Dlink on YARN / K8S 的部署方式，兼容现有CDH，各组件之间解耦。
- 云原生架构，支持计算和存储资源弹性扩容

面向数据质量提升的技术架构

- 提供Hive历史数据生成 Dlink iceberg元数据的方式快速迁移数据入湖
- 提供历史数据快速批式去重，新增数据流式去重
- 提供流式任务异常情况补数据的方案

某零售行业客户案例

行业趋势


海量数据


大数据精准营销


智能机器人

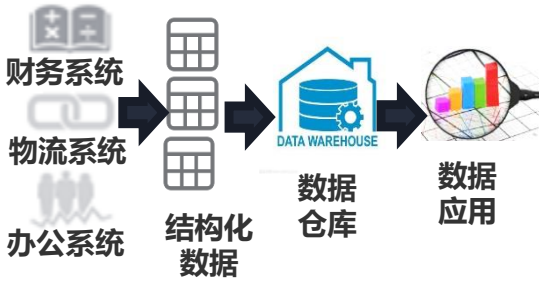
业务


“品牌+零售”战略


品牌大脑

数据智能挖掘

数据现状



会议、直播、行为数据等半/非结构化数据缺少统一存储与挖掘

湖仓架构优点

统一存储

结构化数据和非结构化数据可以统一存储，减少ETL，减少数据不一致，提升分析效率。

流批一体

统一数据开发接口，存储和计算层实现流批一体，提升开发效率和时效性；

存算分离

云原生架构，支持存储和计算资源弹性扩展，提升资源利用率

Data + AI

支持非结构化数据分析，对接算法和机器学习应用，挖掘数据价值；

DLink 未来规划

DLink 未来规划

- 支持亚秒级实时数仓
- 支持企业级多租户多级数据湖
- 支持IMT二级索引
- 支持自适应实时物化视图
- 支持数据加密和查询脱敏
- 支持 Hudi 表格式
- 支持机器学习

加入技术交流群

THANKS

关注社区

加入技术交流群