

# 国产开源湖仓框架 LakeSoul 设计理念和落地应用

陈绪  
北京数元灵科技有限公司  
CTO

DataFunCon # 2023



# 我们是谁



## 湖仓数据中台

自研开源湖仓框架  
统一数据口径



## 数据智能

面向企业智能场景  
释放数据价值



## 一站式服务

从数据源到业务落地  
一站式解决方案

北京数元灵科技有限公司

专注湖仓数据智能新基建

### 产品解决方案

- 实时数据中台解决方案
- 实时湖仓BI分析解决方案
- 低代码个性化生成解决方案
- 智能文案生成引擎解决方案

# Contents

## 目录

- LakeSoul  
设计理念和技术解读

- LakeSoul  
应用场景和案例

- LakeSoul  
核心功能和优势

- LakeSoul  
开源社区进展和未来规划



# 01 • LakeSoul 设计理念和技术原理解读



# LakeSoul 设计理念解析 —— 定位

现代数据栈：



# LakeSoul 设计理念解析 —— 时间线

起源于大型推荐和广告  
业务实时数据流场景

**2021.12**

LakeSoul 国产自  
研流批一体湖仓框  
架开源

**2022.07**

重构元数据，提升  
并发更新和事务能  
力

**2022.10**

发布 Flink CDC  
多表自动入湖，支  
持整库同步，自动  
DDL变更

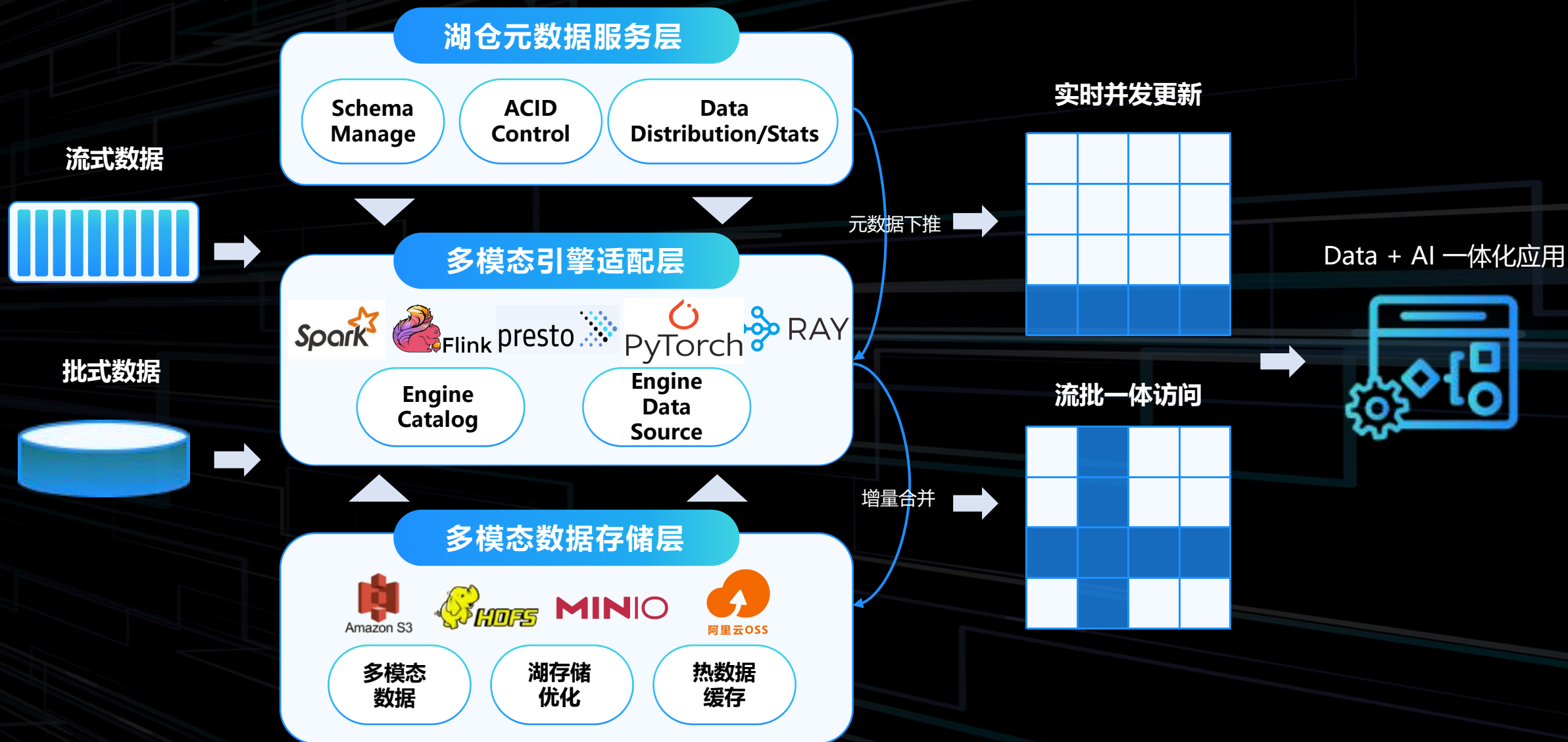
**2023.05**

发布 Native IO，  
扩大性能领先优势。  
LakeSoul 项目捐  
赠给 Linux 基金  
会孵化

**2023.06**

发布全链路流式增  
量计算，自动合并  
等领先功能  
通过国产信创认证

# LakeSoul 技术解析 —— 整体架构





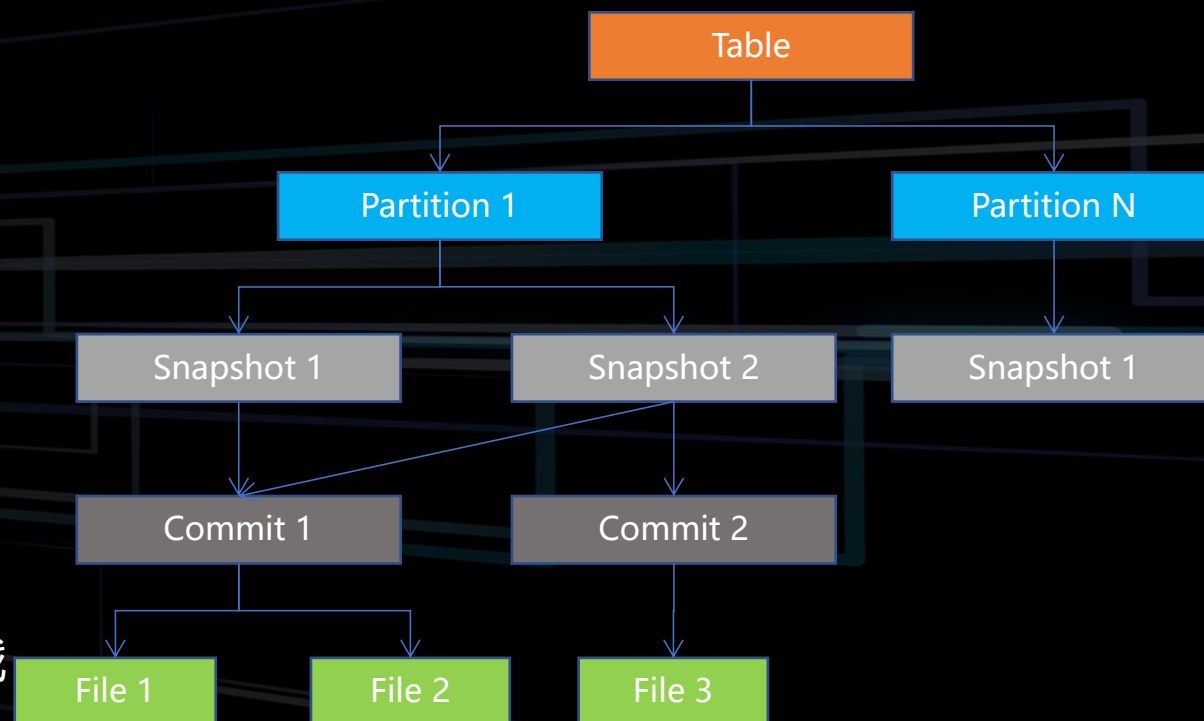
# LakeSoul 技术解析 —— 数据格式

## • 表物理数据组织

- 在文件系统上以Parquet格式存储
- 主键：按主键哈希分片，文件内按主键排序
- 分区：多级Range分区

## • 表元数据组织

- Commit: 文件序列
- Snapshot: Commit 序列
- Version: 递增序号，标识一个 Snapshot 及其时间戳

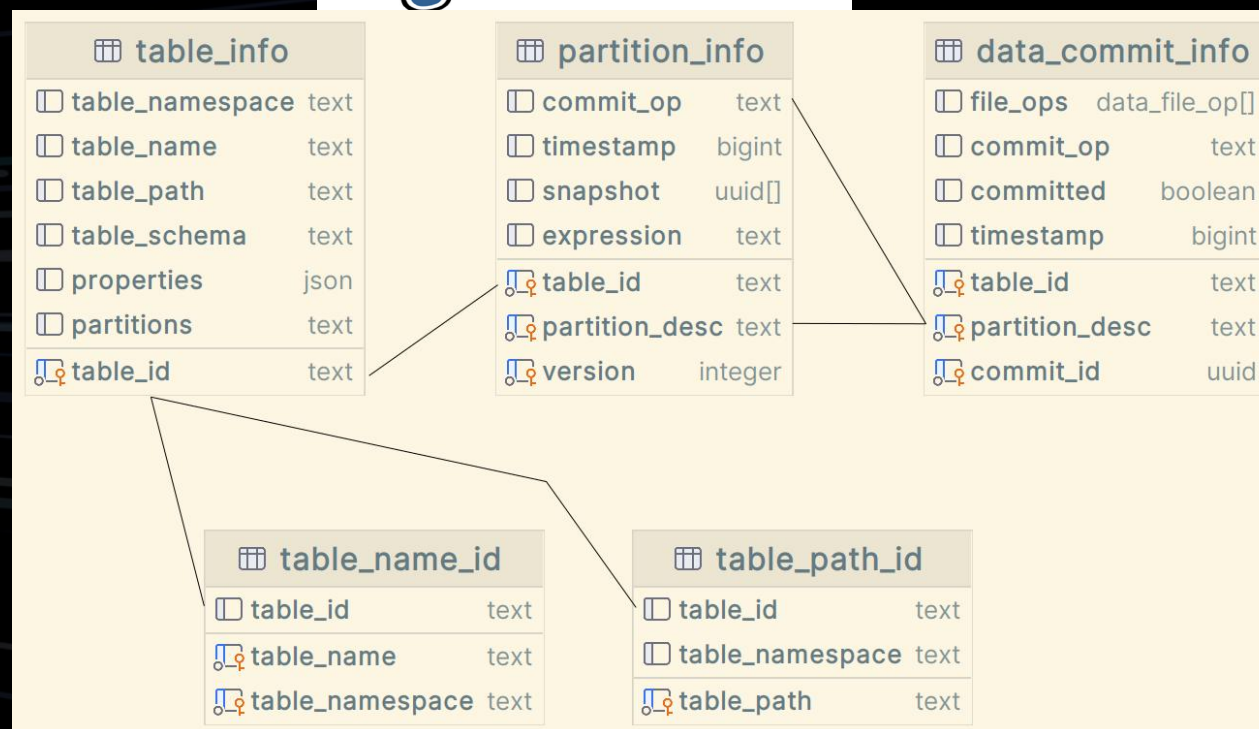




# LakeSoul 技术解析 —— 元数据

## • 中心化的元数据管理

- 使用 PostgreSQL 作为存储
- 使用 PG 事务实现并发控制、ACID
- 两阶段提交协议
- 细粒度写冲突自动解决机制
- 使用 PG Trigger 机制实现事件订阅



# LakeSoul 技术解析 —— 元数据

## • 两阶段提交协议

- 在 Spark/Flink 流/批作业写数据时执行

- Prepare Phase - Insert entries into data\_commit\_info:
  - file\_ops: "s3://bucket/file1,add"
  - partition\_desc: "date=202305054"
  - timestamp: 1682234381
  - committed: false

- Commit Phase
  - BEGIN TRANSACTION
  - Change status iff committed == false:
    - file\_ops: "s3://bucket/file1,add"
    - partition\_desc: "date=202305054"
    - timestamp: 1682234381
    - committed: true
  - Insert new snapshot entry into partition\_info with version incremented by 1 iff version has not been changed
  - END TRANSACTION

Persisted Checkpoint State

Conflict Resolver



# LakeSoul 技术解析 —— 元数据

## • 自动并发冲突解决机制

- 直接重试提交：兼容的写冲突（Append、Merge）
- 重新排列 Commit：Compaction、Update的部分情况
- 不兼容冲突：并发全量 Update 等，提交失败

Operation	Append	Merge	Compaction	Update
Append	Retry	X	Retry	Retry
Merge	X	Retry	Reorder	Retry
Compaction	Reorder	Reorder	Ignore	Ignore
Update	Reorder	Reorder	Overwrite	Fail

# LakeSoul 技术解析 —— 元数据

- **Schema 自动演进**

- 在写时自动处理 Schema 变更
- 允许并发进行变更，不需要停作业再执行 DDL
- 读数据自动兼容
  - 新增加列：旧数据自动补充 null
  - 删除列：旧数据自动过滤该列
  - 提升类型精度：旧数据自动执行 Upcast

- **快照管理**

- 快照读、快照回滚、快照清理
- 默认读取最新的快照



# LakeSoul 技术解析 —— IO

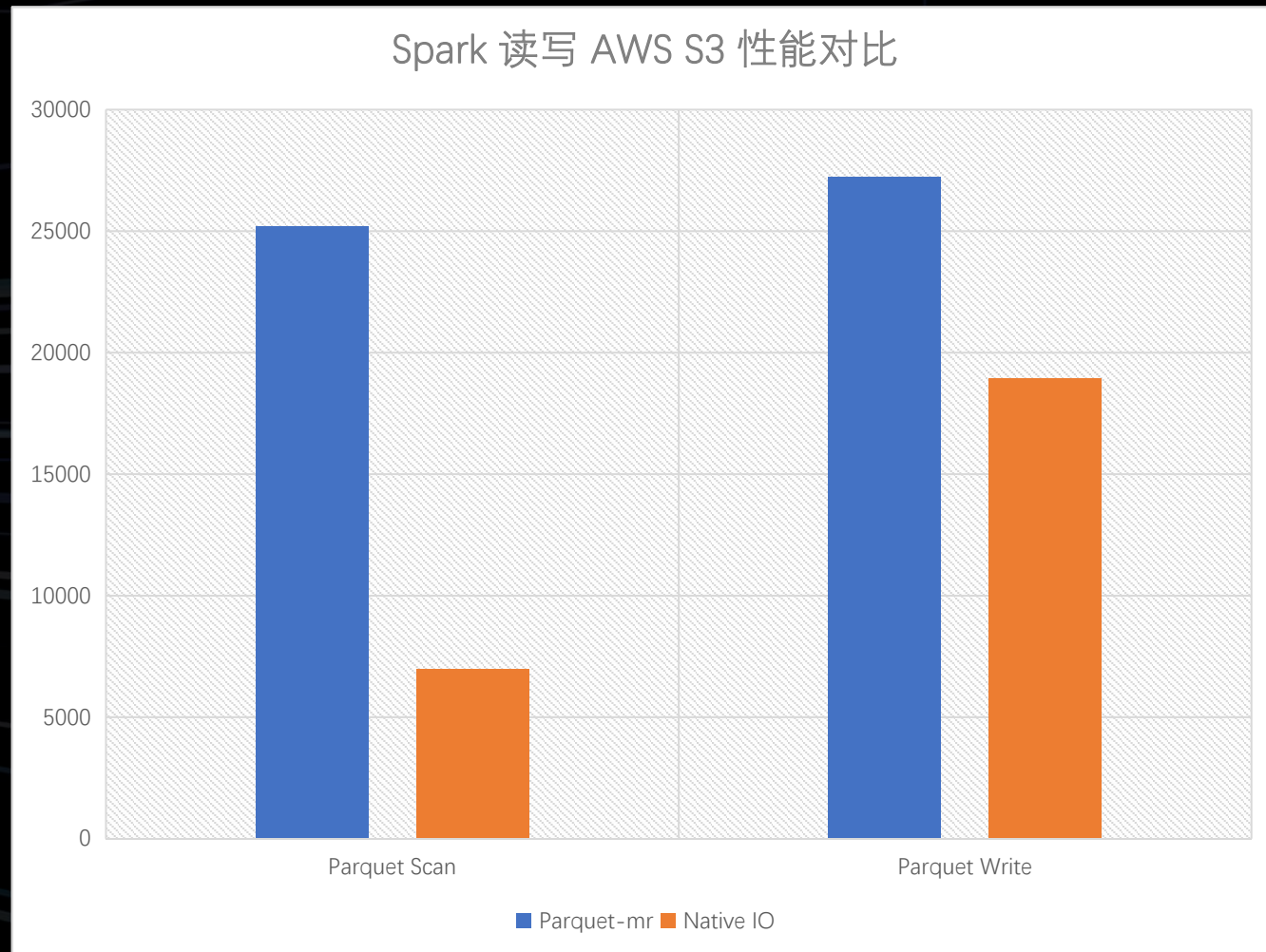
## 核心优势:

### 1. 高性能

1. Rust 原生实现向量化 MOR
2. 返回 Arrow Batch, 跨语言零拷贝
3. 对象存储优化 (Parallel Multipart Upload, Async RowGroup Prefetch, .....

### 2. 多语言、多引擎对接

1. Rust: DataFusion Source&Sink
2. Java: 对接 Spark/Flink/Presto, 已兼容 Gluten
3. C++: Arrow C++ Dataset, Velox(开发中), Doris(开发中)
4. Python: PyArrow/Pandas Dataset
5. PyTorch、Ray, 支持分布式作业



## 02• LakeSoul 核心功能和优势



# LakeSoul 核心功能 —— 入湖

- 多源数据实时入湖

- 数据库多表实时入湖、Kafka 多 topic 实时入湖
- 支持 Flink CDC、Debezium 等多种 CDC 采集工具
- 自动发现新表、自动 DDL 变更
- Exactly Once



# LakeSoul 核心功能 —— 增量计算

- LakeSoul 表 CDC 读写和增量计算
  - 读写均兼容 Flink Changelog Stream 格式
  - 全链路增量事件驱动
  - 降低延迟, 节省资源, 无需离线 DAG 调度组件

```
INSERT INTO lakesoul_table SELECT * FROM mysql_cdc_stream;
```

```
SELECT sum(revenue) FROM lakesoul_table  
/*+ OPTIONS('readstarttime'='2023-04-21 10:00:00','readtype'='incremental')*/  
GROUP BY city;
```

Row Kind	city (pk)	revenue
+I	BJ	100
+I	SH	200
U	BJ	150
-D	SH	



city	sum(rev)
BJ	100
SH	200



city	sum(rev)
BJ	150
SH	200



city	sum(rev)
BJ	150



# LakeSoul 核心功能 —— 多流拼接

- 原生支持多流并发写入，读取时合并 (partial update)

- 多个流有相同主键列，其余列可以不同
- 消除大表 Join/双流 Join 状态开销
- 降低延迟，减少资源消耗

Stream A

PK	Field 1	Field 2
key1	1	"abc"

Stream B

PK	Field 1	Field 3	Field 4
key2	2	9.99	"xyz"

Stream C

PK	Field 2	Field 1	Field 3
key1	"def"	3	0.99

具有相同主键的多个流可以并发 Upsert 到同一张目标表

PK	Field 1	Field 2	Field 3	Field 4
key1	3	"def"	0.99	null
key2	2	null	9.99	"xyz"

Target Table

Stream A

PK_A	Field 1	Field 2	FK_B
key1	1	"abc"	key2

Stream B

PK_B	Field 3	Field 4
key2	2	9.99

不同主键的 Join 可以转为

- 一个流 Upsert
- 第二个流 broadcast/lookup join 后 Upsert

PK_A	PK_B	Field 1	Field 2	Field 3	Field 4
key1	key2	1	"abc"	2	9.99

Target Table

# LakeSoul 核心功能 —— 权限和血缘

- **内置 RBAC 权限管理**

- 原生支持多租户空间
- 元数据隔离：PG RBAC、PG Row Level Security Policy
- 对 SQL/Python/Jar 作业均可隔离

- **内置数据血缘功能**

- 采用 OpenLineage 协议上报血缘关系
- 支持 Spark、Flink 流、批作业



# LakeSoul 核心功能 —— 自动维护

- **自动维护 (Auto Maintenance)**

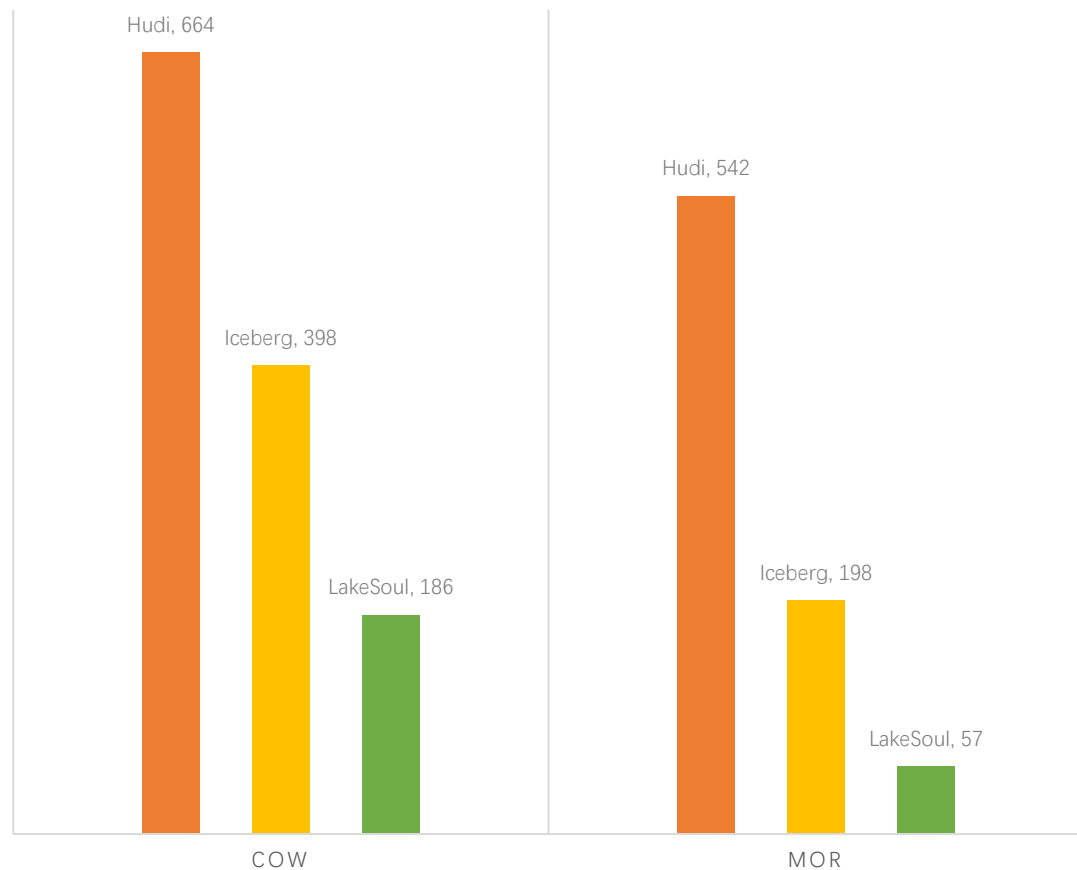
- 自动全局 Compaction 服务
- 自动清理过期数据服务

- **使用 PostgreSQL 的 Trigger 功能**

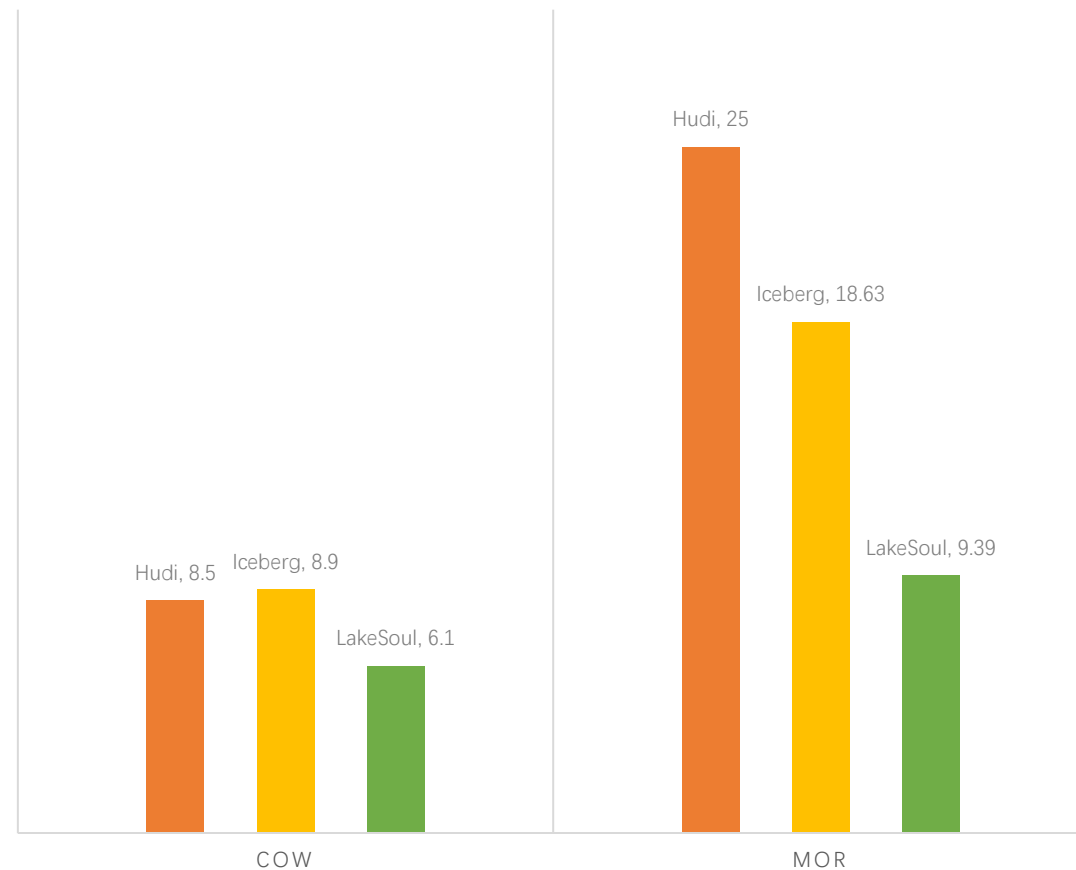
- 达到 Compaction/清理 条件时触发事件
- 使用 Spark 作业监听事件并执行 Compaction/清理操作
- Spark 作业可以弹性伸缩

# LakeSoul 性能评测

## WRITE TIME(SECONDS)



## READ TIME(SECONDS)



测试代码和数据:

<https://github.com/meta-soul/ccf-bdci2022-datalake-contest-examples/tree/mor>

<https://github.com/meta-soul/ccf-bdci2022-datalake-contest-examples/tree/cow>

测试方式:

- 第一批插入 1000 万行数据
- 分10次插入 100 万行数据
- MOR 读取时没有执行过 Compaction

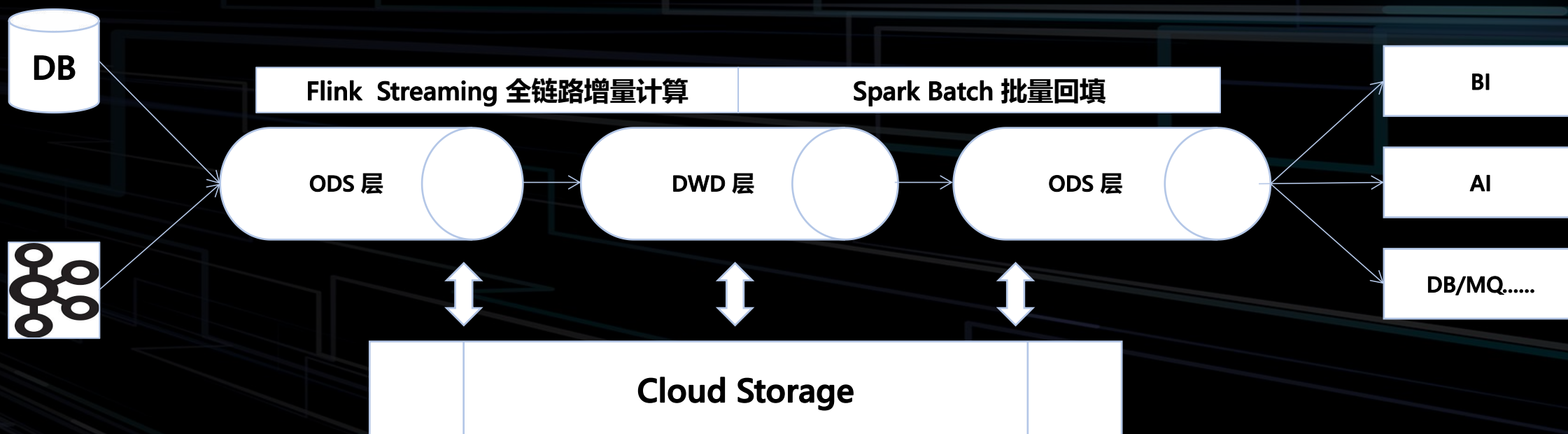


## 03• LakeSoul 应用场景

# LakeSoul 应用场景 —— 构建实时湖仓

## • 构建完整的实时湖仓一体链路

- 多源数据实时入湖
- 实时增量计算
- 全量、增量一体化分析
- BI、AI多种上层应用





# LakeSoul 应用场景 —— 实时机器学习

## • 构建实时机器学习样本

- 使用 LakeSoul 多流拼接功能
- 将多个特征流、标签流实时拼接
- 将样本流式传入机器学习训练，实现在线学习

## • 在 PyTorch、Ray 等框架中读取数据

- 消除额外格式转换，跨引擎共享数据
- 支持 MOR 表读取
- 支持分布式作业读取



# LakeSoul 应用场景 —— 实时机器学习

```
dataset_table = "imdb"

def read_text_table(datasource, split):
    dataset = datasets.IterableDataset.from_lakesoul(datasource, partitions={"split": split})
    for i, sample in enumerate(dataset):
        yield {"text": sample["text"], "label": sample["label"]}
```

```
# Tokenize the IMDb dataset
train_tokenized_imdb = IterableDataset\
    .from_generator(read_text_table, gen_kwargs={"datasource": dataset_table, "split": "train"})\
    .map(preprocess_function, batched=True)\
    .shuffle(seed=1337, buffer_size=25000)
test_tokenized_imdb = IterableDataset\
    .from_generator(read_text_table, gen_kwargs={"datasource": dataset_table, "split": "test"})\
    .map(preprocess_function, batched=True)
```

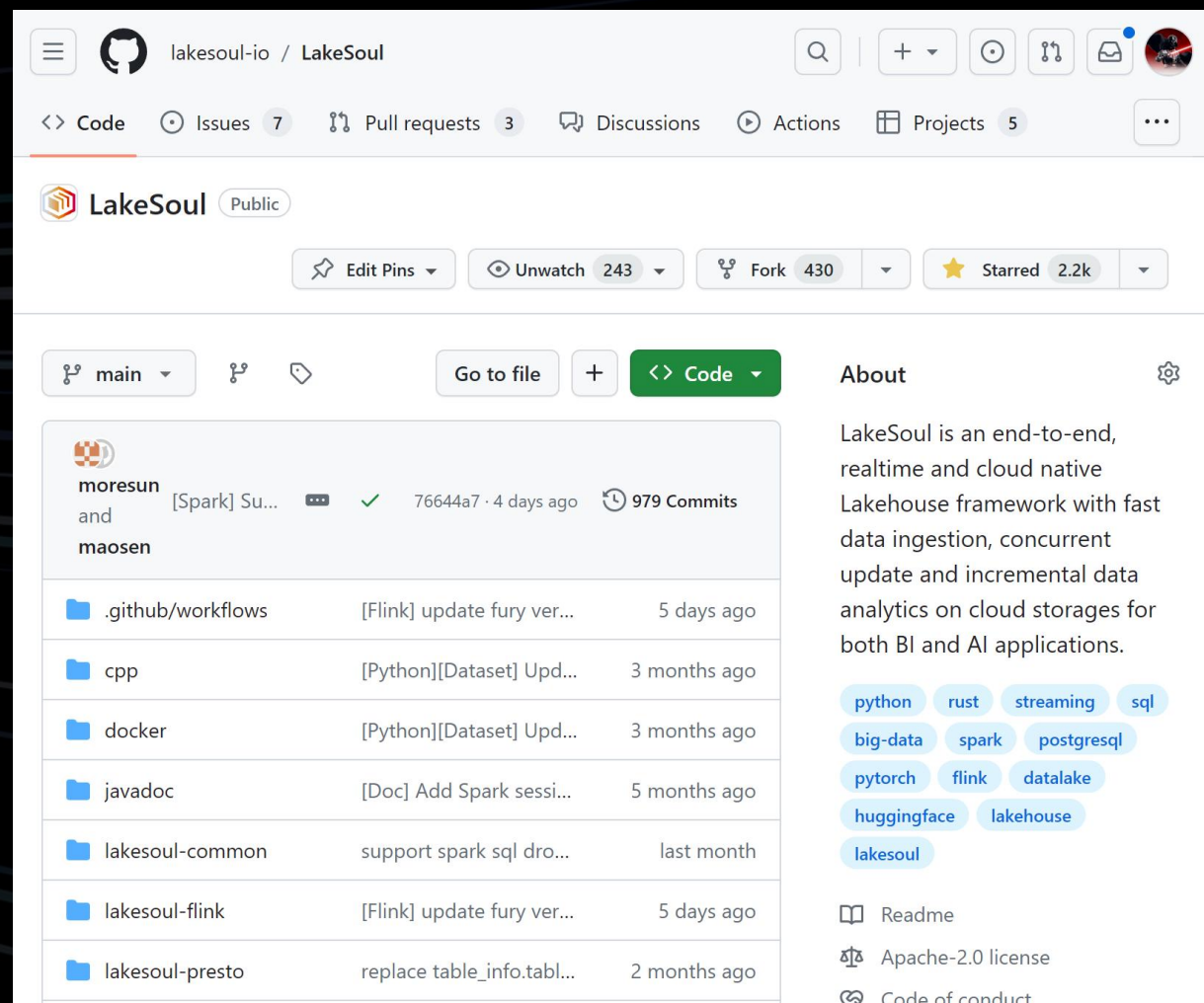


# 04• LakeSoul 开源社区和未来规划



# LakeSoul 开源社区

- 2021年底开源，采用 Apache License 2.0 协议，成为国内首个开源湖仓框架
- 2023年5月将项目捐赠给 Linux 基金会，成为 Linux Foundation AI & Data 旗下 Sandbox 项目
- 欢迎感兴趣的朋友关注和参与到社区中来
- <https://github.com/lakesoul-io/LakeSoul>





# LakeSoul 未来演进方向

## • 功能

- 可插拔 WAL 支持  
亚秒级实时可见性
- 实时数据质量校验
- Hadoop/K8s 自动化部署

## • 生态

- 支持更多数据库入湖
- Kafka Connect Sink
- LogStash Sink

## • 性能

- Minor compaction
- 支持 Presto Velox Worker
- 支持 Apache Doris



# 谢谢



数元灵  
DMetaSoul

## 北京数元灵科技有限公司

### 专注湖仓数据智能新基建

#### 联系我们

官网: [www.dmetasoul.com](http://www.dmetasoul.com)

地址: 北京市朝阳区广顺南大街嘉美中心

商务咨询: [public-contact@dmetasoul.com](mailto:public-contact@dmetasoul.com)



#### 湖仓数据中台

自研开源湖仓框架  
统一数据口径



#### 数据智能

面向企业智能场景  
释放数据价值



#### 一站式服务

从数据源到业务落地  
一站式解决方案



欢迎扫码进群交流

#### 产品解决方案

- 实时数据中台解决方案
- 实时湖仓BI分析解决方案
- 低代码个性化生成解决方案
- 智能文案生成引擎解决方案