Apache Flink

Flink Table Store 典型应用场景

李劲松/Flink PMC 2022-9-24

Apache Flink meetup 北京站

About Me – 李劲松 Jingsong Li

- 就职于阿里云 – 开源大数据
  - 高级技术专家
  - Team Leader of Flink Table Store
  - 流计算 -> 批计算 -> 存储

- 开源社区
  - Apache Flink PMC & Committer
  - Apache Iceberg Committer
  - Apache Beam Committer

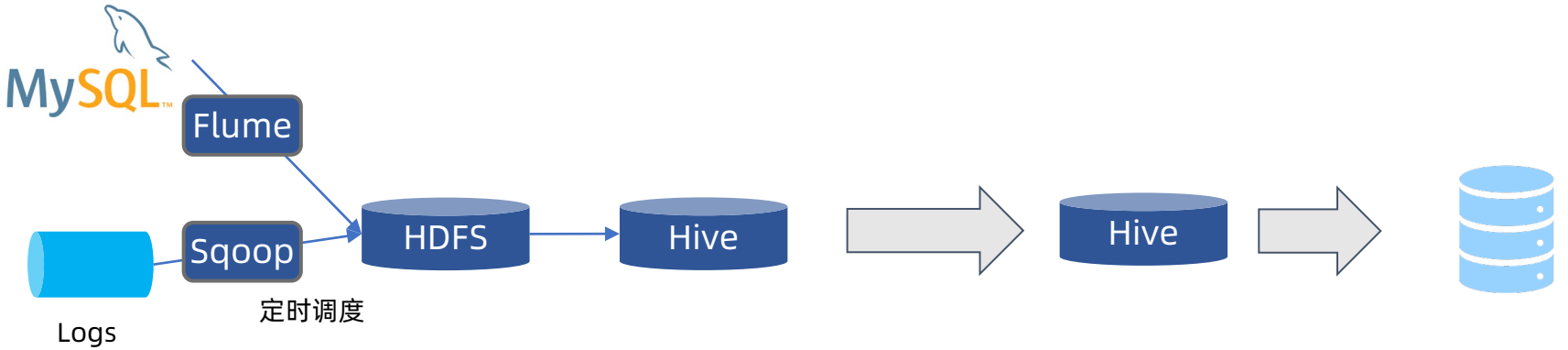
- 个人爱好：爬山、海滩躺平

CONTENT

目录 >>

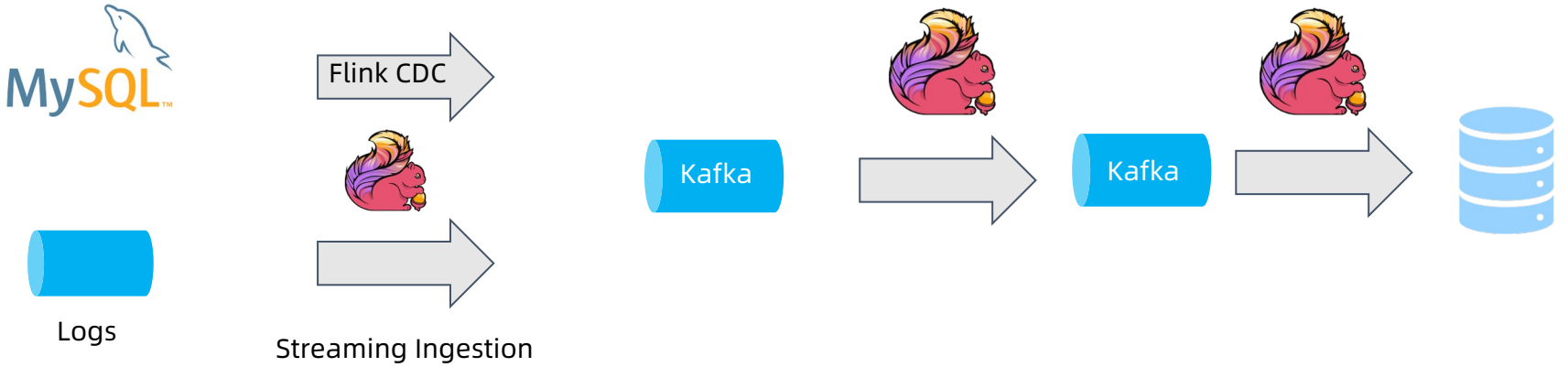
01 介绍 Flink Table Store

两种数仓形态



离线数仓

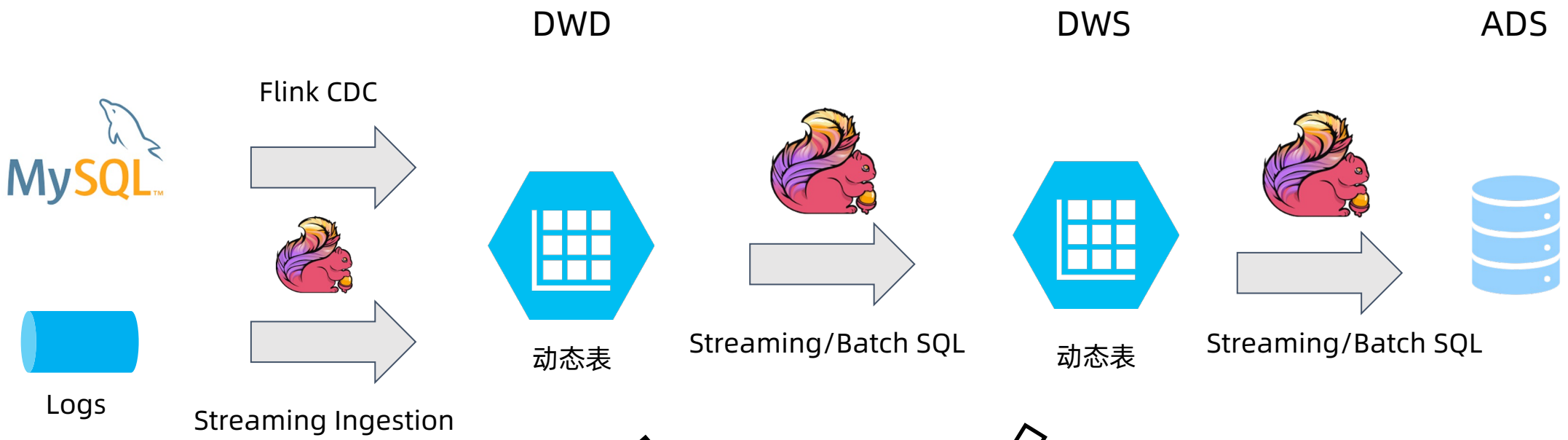
- 延迟高
- 更新：全量合并



实时数仓

- 中间数据不可查
- 没有历史数据

Streaming Warehouse



- 全链路实时流动
- 沉淀数仓中间数据
- 实时离线分析一体化

动态表需要的能力

Table Format

更新写：面向数据库 CDC 和流处理产生的**大量更新**数据

批读：提供**高效**的批查询

Streaming Queue

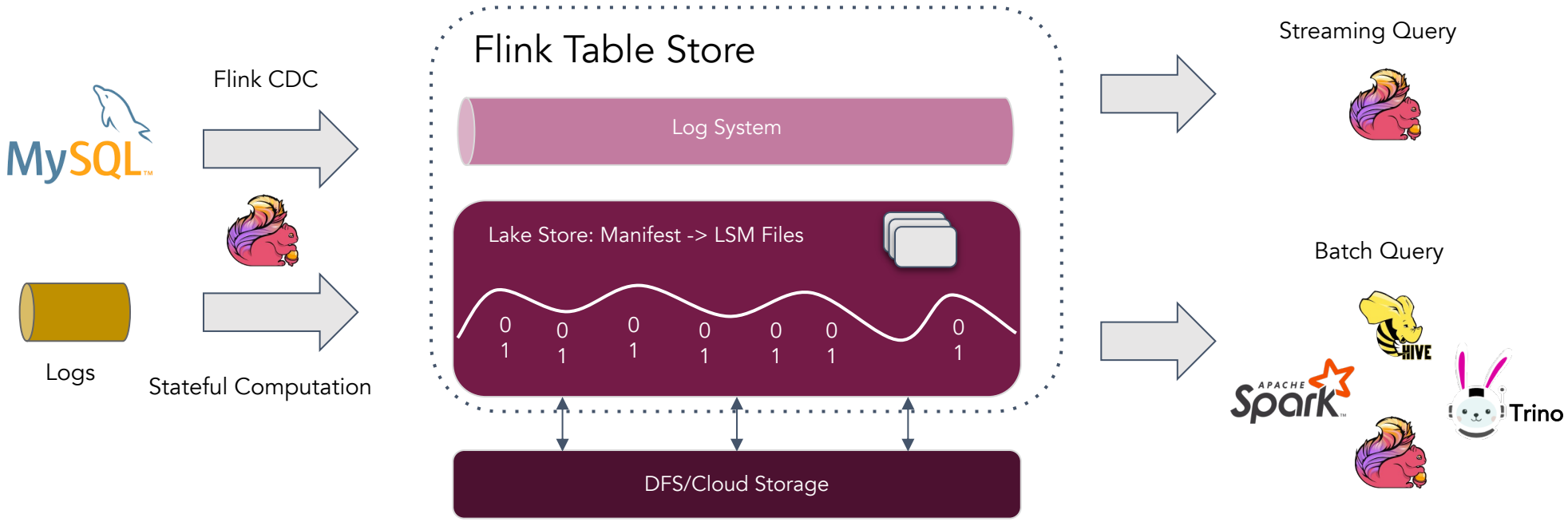
流写、流读

建立**增量处理**的 Pipeline

Lookup Join

面向 Flink 的**维表连接**，提供 Lookup Join

Flink Table Store



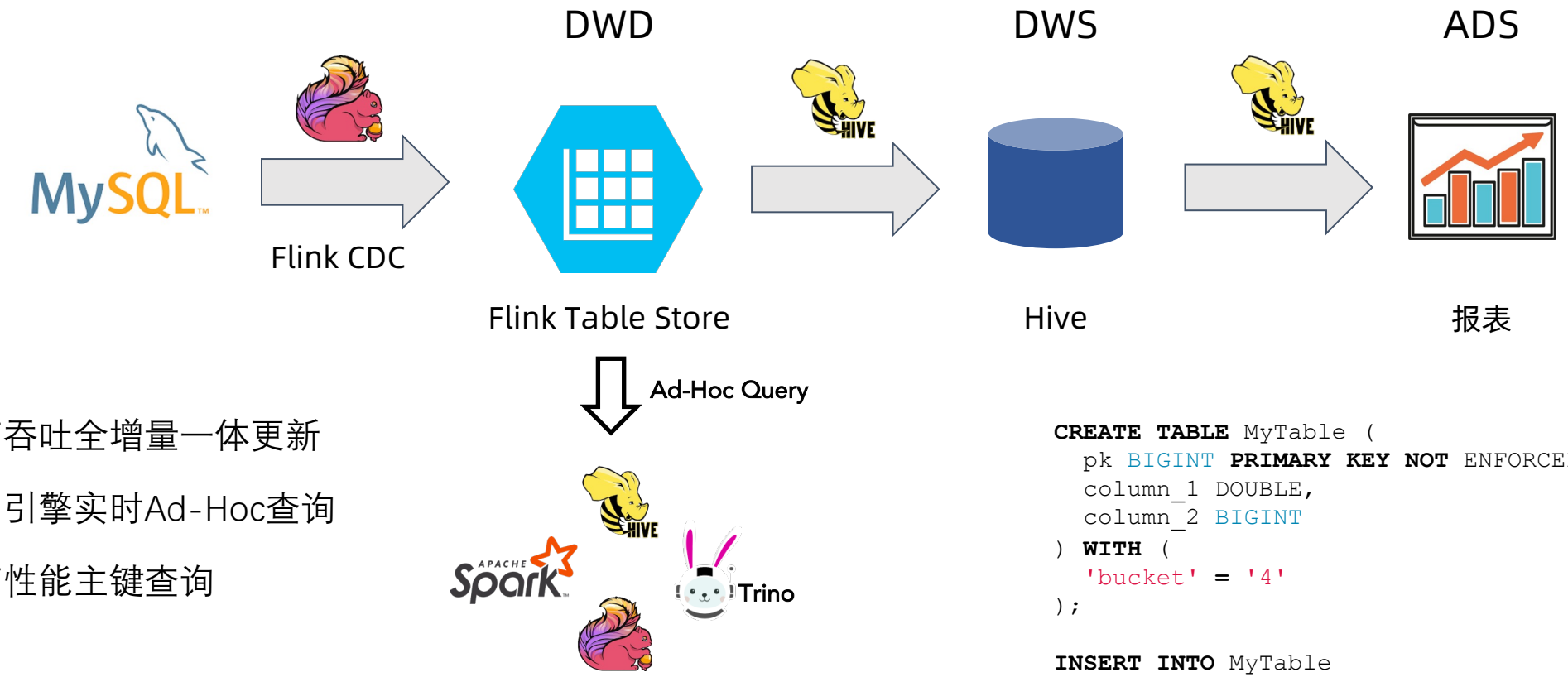
Flink Table Store

更新系统	存储成本	更新时延	更新方案	数据定位	高性能点/范围查
Iceberg	低	小时级	COW	批Join	需手动排序
Delta	低	小时级	COW	批Join	需手动排序
Hudi	低	分钟级	COW/MOR	KV Index	需手动排序
FTS	低	分钟级	MOR	LSM	无需排序
Clickhouse	高（服务化）	秒级	MOR	LSM	无需排序

FTS ≈ 湖存储版本的 Clickhouse MergeTree Engine

02 应用场景

数据库 CDC 入仓

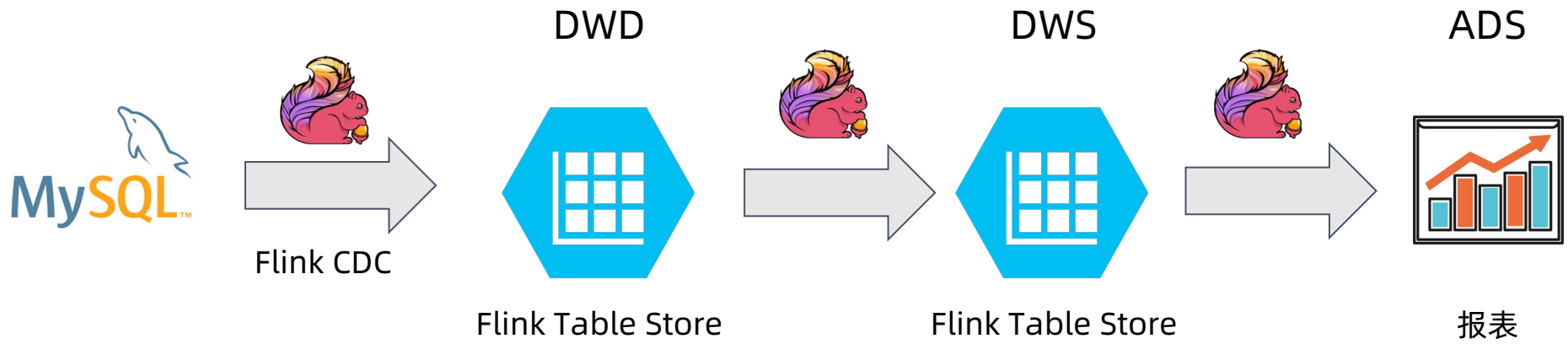


- 高吞吐全增量一体更新
- 多引擎实时Ad-Hoc查询
- 高性能主键查询

```
CREATE TABLE MyTable (
  pk BIGINT PRIMARY KEY NOT ENFORCED,
  column_1 DOUBLE,
  column_2 BIGINT
) WITH (
  'bucket' = '4'
);

INSERT INTO MyTable
SELECT * FROM cdc_table
```

流式 Pipeline



- 包含历史数据的流读
- 中间数据可查询
- V0.3: 完整 Changelog 的产生


```
-- 建议1分钟 checkpoint interval
-- 准实时 Streaming Pipeline

CREATE TABLE MyTable (
  pk BIGINT PRIMARY KEY NOT ENFORCED,
  column_1 DOUBLE,
  column_2 BIGINT
) WITH (
  'bucket' = '4',
  'changelog-producer' = 'input'
);

INSERT INTO MyTable
SELECT * FROM cdc_table;

SELECT * FROM MyTable;
```

维表连接

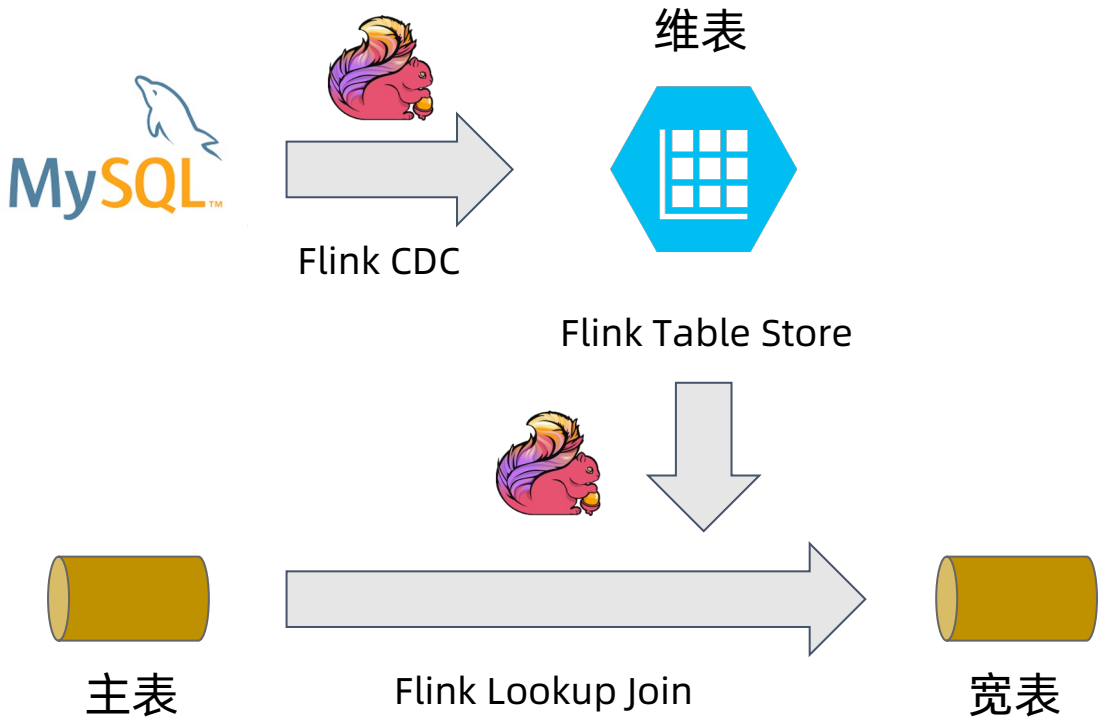
```
SELECT * FROM Main LEFT JOIN dim
FOR SYSTEM_TIME AS OF Main.proctime AS D
ON Main.dimId = D.id;
```

- 维表实时拉取 FTS 最新版本
- 支持维表非主键的关联
- 维表 Join 会维护本地 Cache，规模：

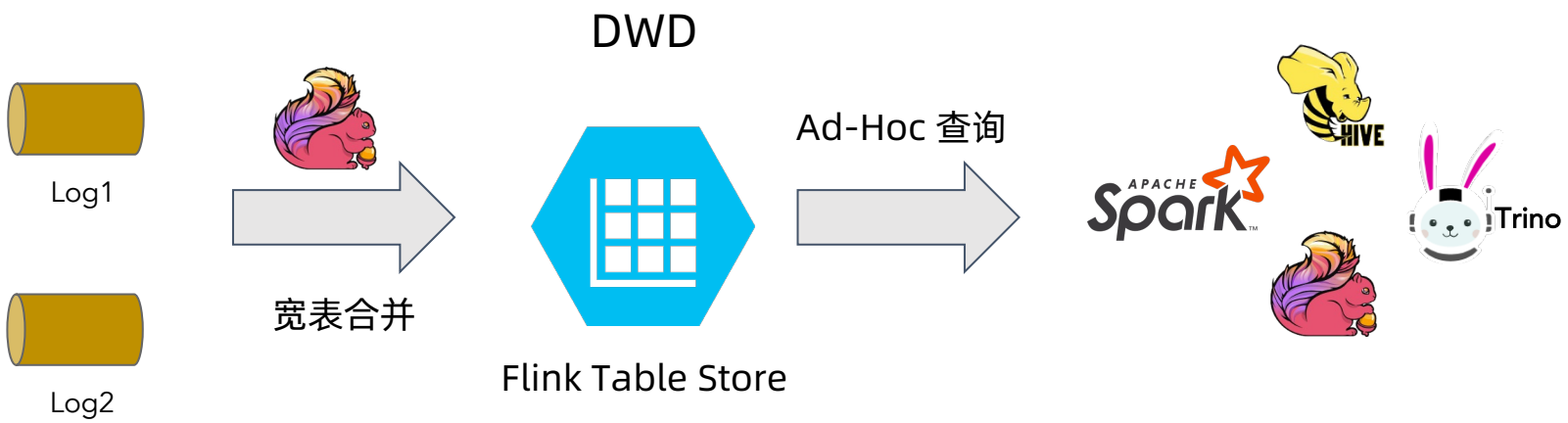
  字段较少，比如 2-3 个：可以支持数千万

  字段较多，最好在千万级以下的规模

  后续逐渐加强中



宽表合并



- 打宽表，实时查询
- V0.3：多作业写入
- V0.3：Changelog 流读

```
CREATE TABLE MyTable (
  pk BIGINT PRIMARY KEY NOT ENFORCED,
  column_1 DOUBLE,
  column_2 BIGINT
) WITH (
  'merge-engine' = 'partial-update'
);

INSERT INTO MyTable
SELECT pk, column_1, NULL FROM Src1
UNION ALL
SELECT pk, NULL, column_2 FROM Src2
```

03 Demo

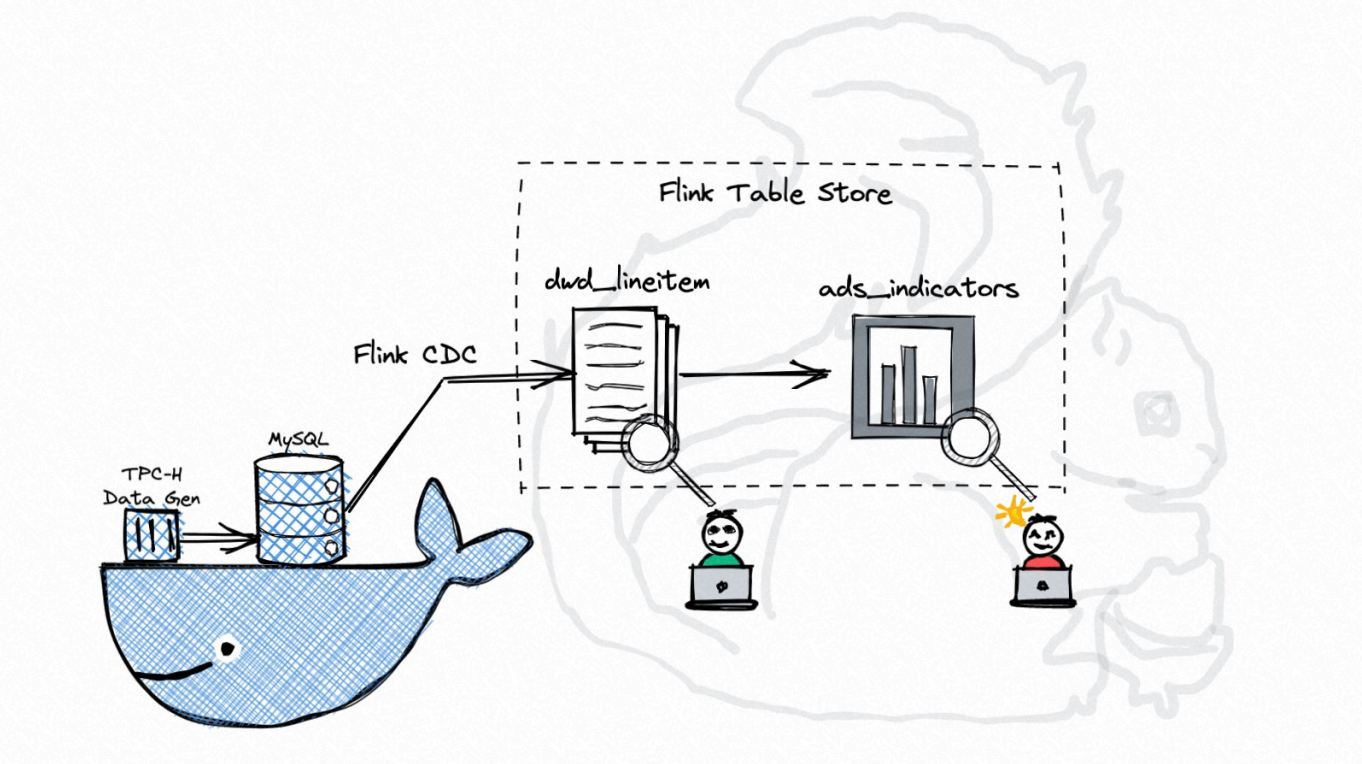
About Speaker - 陈婧敏 Jane Chan

- 目前就职于阿里云 – 开源大数据
  - Apache Flink Contributor
  - 目前专注于 Flink Table Store 研发

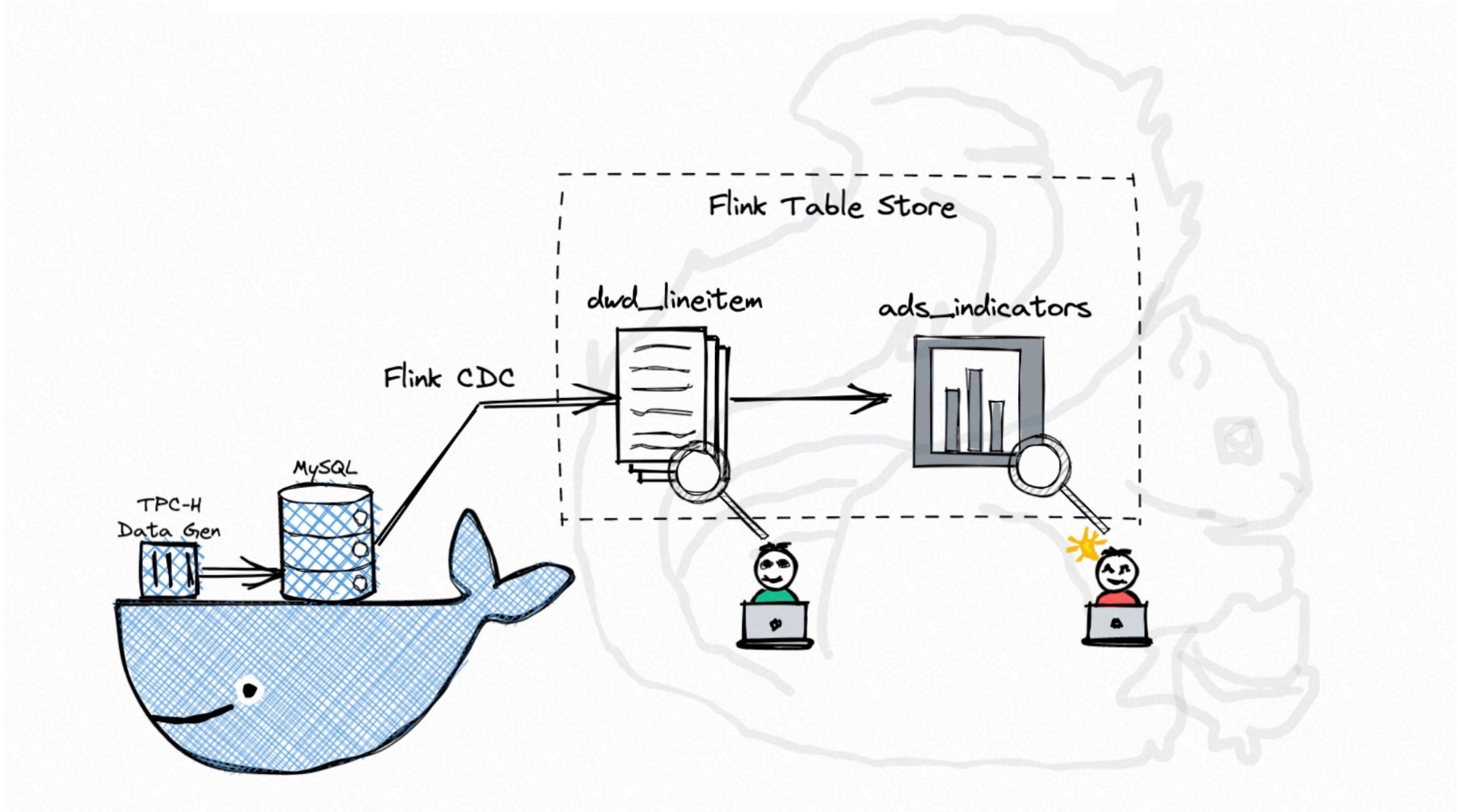
- Previous Role
  - 2015 年复旦大学毕业后加入阿里巴巴数据中台，从事实时数据和平台开发

Demo

- 全增量一体 CDC 实时入湖，单机轻松搞定
  近百个分区 + 6000万CDC数据
- Streaming Data Warehouse 构建
- https://github.com/LadyForest/flink-table-store-101/blob/master/real-time-update/README.zh.md



Flink Table Store 全增量一体 CDC 实时入湖



04 后续挑战

Streaming Data Warehouse 挑战

存储的管控

存储的运维、表管控、元数据、DDLs

流计算准确性

缺少完整的Changelog，Normalize节点代价太高，什么时候应该去重，怎么算错了

头大的 Join

维表Join需要额外系统且有时语义不满足，双流Join保存全量明细代价太高

物化视图一致性

流计算输出的表呈现不一致的视图

Flink Table Store 后续规划

好用的流存储

- 多作业写入支持 & Compaction 分离
- 完整的 Streaming Data Warehouse
  - DDLs & Procedures
  - Update & Delete 语法等
  - Time Travel API 支持

准确的流存储

- 完整 Changelog 生成
  - Input
  - Full Compaction
  - Lookup
- 下游不引入 Normalize Node 或去重的情况下，带来准确的流计算 Pipeline

可连接的流存储

- Lookup Join 加强，解锁超大维表
- 二级索引，增强连接灵活性
- 维表对齐能力，解决维表的不确定

项目信息

- Subproject of Apache Flink
- Github: https://github.com/apache/flink-table-store
- User Docs: https://nightlies.apache.org/flink/flink-table-store-docs-master/
- Mail list:
  - dev@flink.apache.org
  - user@flink.apache.org
  - user-zh@flink.apache.org
- Ding group

钉钉群：10880001919

Thanks

Apache Flink meetup 北京站