



Apache Paimon



StarRocks

饿了么基于Paimon&StarRocks 的实时湖仓探索

王沛斌/大数据架构师、 Apache Flink Contributor

Streaming Lakehouse Meetup

CONTENT

目录 >>

01 /

饿了么实时数仓演进之路

02 /

实时湖仓方案选型与探索

03 /

实时湖仓规划及展望

01 饿了么实时数仓演进之路

饿了么典型实时应用场景



实时ETL

- 实时入湖入仓
- 实时数据建模
- 实时流量归因
- ...



实时报表

- 营销活动直播
- 商家生意参谋
- 实时流量大盘
- 大促实时大屏
- 实时AB实验



在线应用

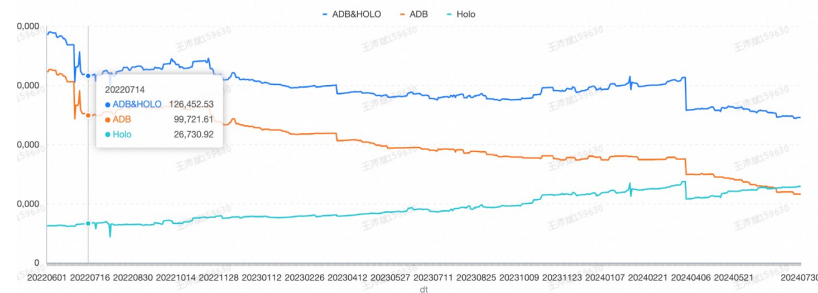
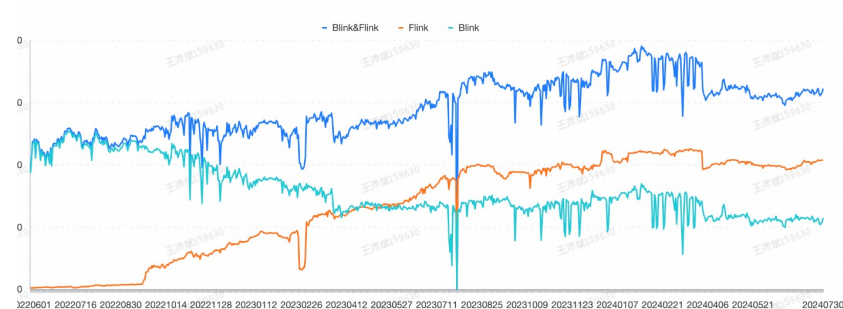
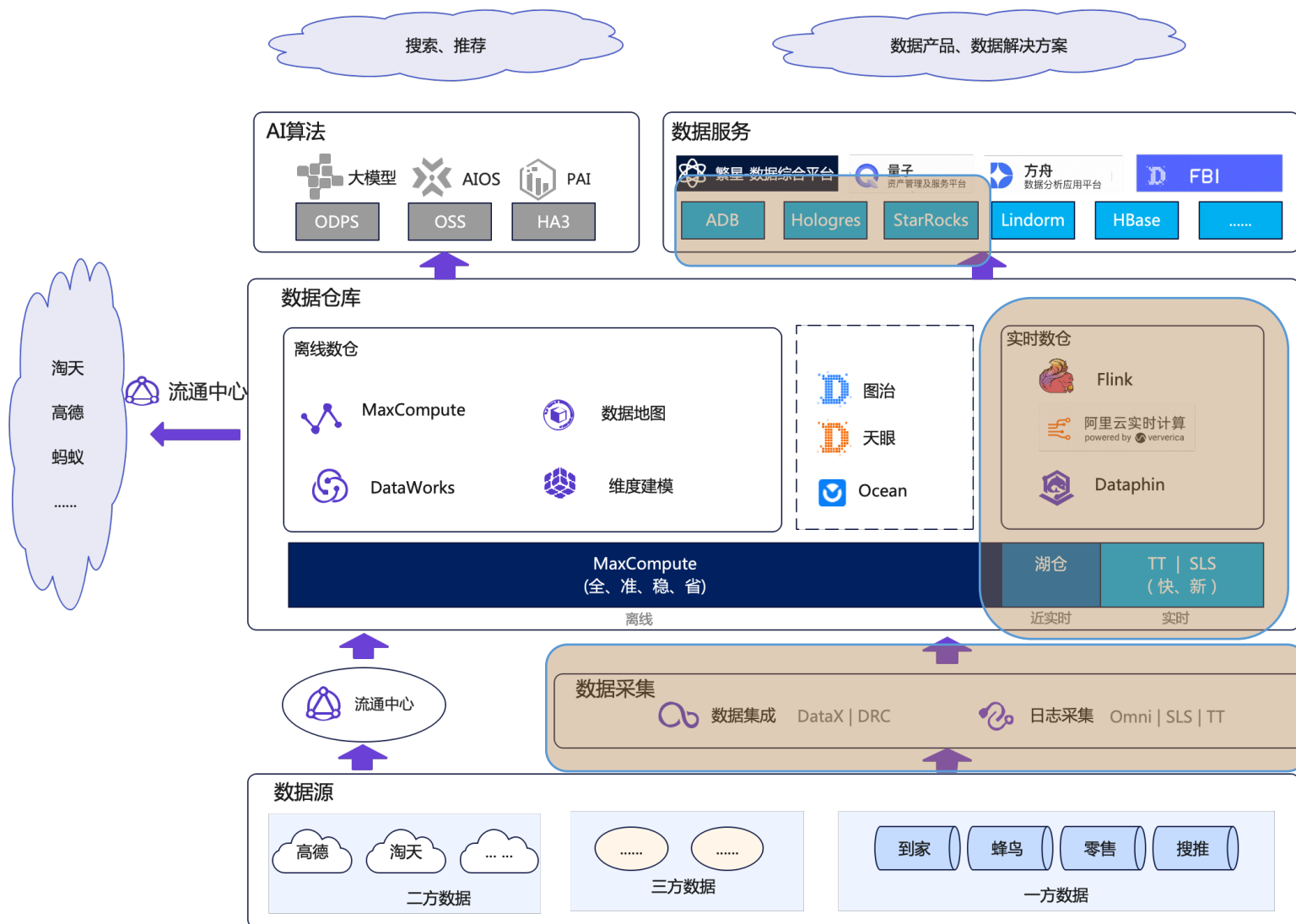
- 商物流实时联动
- 实时人群特征及投放
- 个性化推荐
- IOT信息同步
- 风控实时拦截



实时监控与补偿

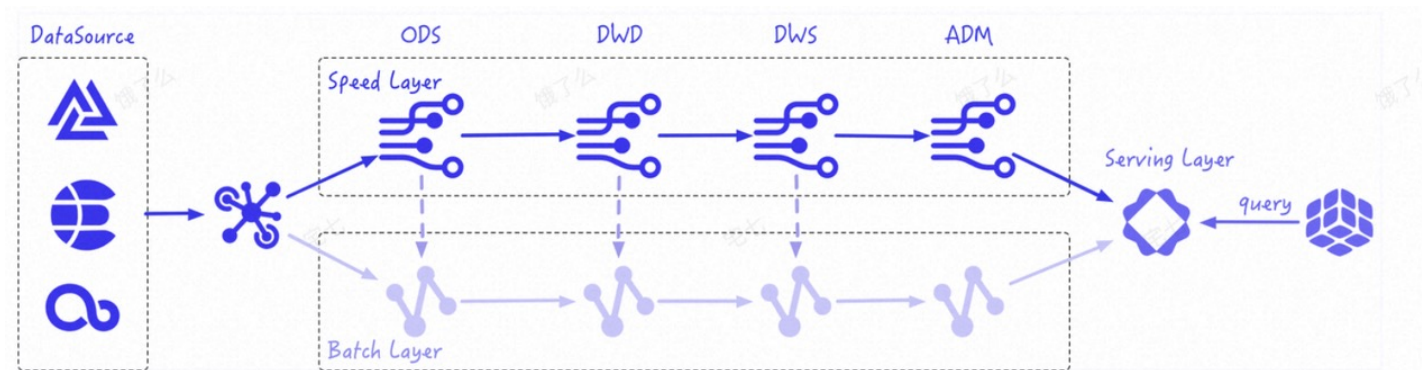
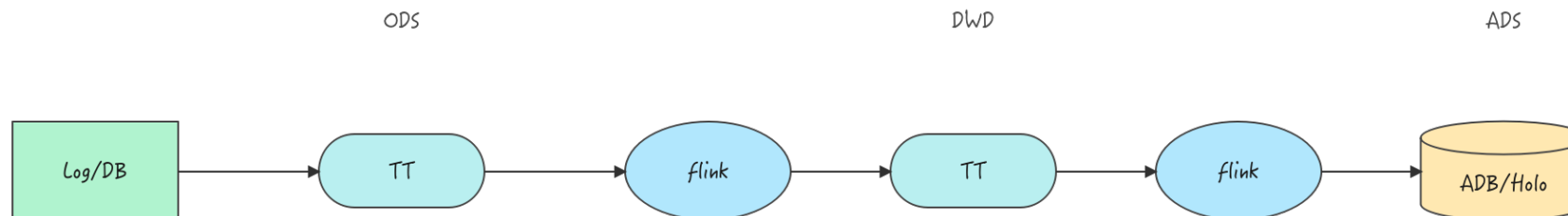
- 实时数据核对&订正
- 业务诊断预警
- 服务器异常监控
- ...

饿了么数据架构大图



实时数仓1.0

实时数仓1.0：



01

现象

研发效率低
数据不一致
运维保障难
计存成本高

02

原因

- 实时指标需要细粒度同环比
- TT不支持OLAP查询
- 实时中间层缺失（烟囱开发）
- 快照数据导致计算量大
- 研发门槛高 VS 研发水平参差不齐
- 任务峰值与业务峰值重叠
- 保障要求更高（7*24H）
- AP调优不稳定、需人肉调优
- 元数据缺失，上下游依赖不清晰

03

目标

1. 提供更快、准、稳、一致的实时数据，满足实时业务数据的诉求，提升核心指标的实时覆盖率
2. 降低实时数据使用门槛，提升实时数据研发效率

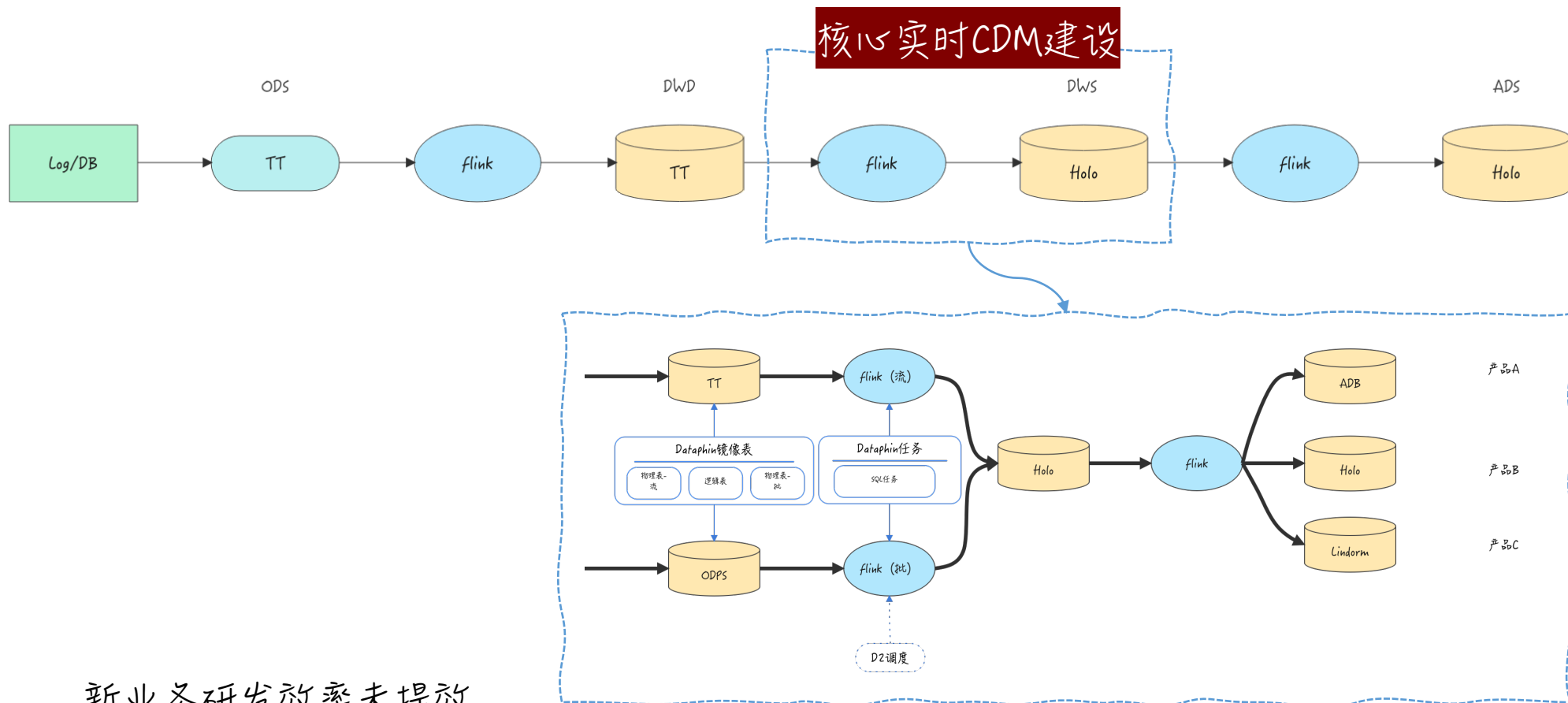
04

解决方案

1. 数据产品能力升级，收敛实时需求
2. 夯实实时CDM资产，收口指标加工逻辑
3. 实时数仓架构方案升级，获取技术红利、降低研发复杂度
4. 研发规范化及工具沉淀（流程卡点&实时基线等）

实时数仓2.0

实时数仓2.0 :



新业务研发效率未提效

“流批一体”：DWD两份存储

TT：不支持检索&更新、无法列裁剪

Holo内表高昂的存储成本



Streaming Lakehouse

02 实时湖仓方案选型与探索

选型与测评方案

软件信息

集群版本

软件信息

EMR-5.15.1

Celeborn (0.3.1)

HADOOP-COMMON (3.2.1)

Hive (3.1.3)

Hudi (0.13.1)

JindoCache (6.1.1)

Knox (1.5.0)

OpenLDAP (2.4.46)

OSS-HDFS (1.0.0)

Paimon (0.6)

Presto (0.283)

Spark3 (3.3.1)

Tez (0.10.2)

EmrRuntime (1.1.0)

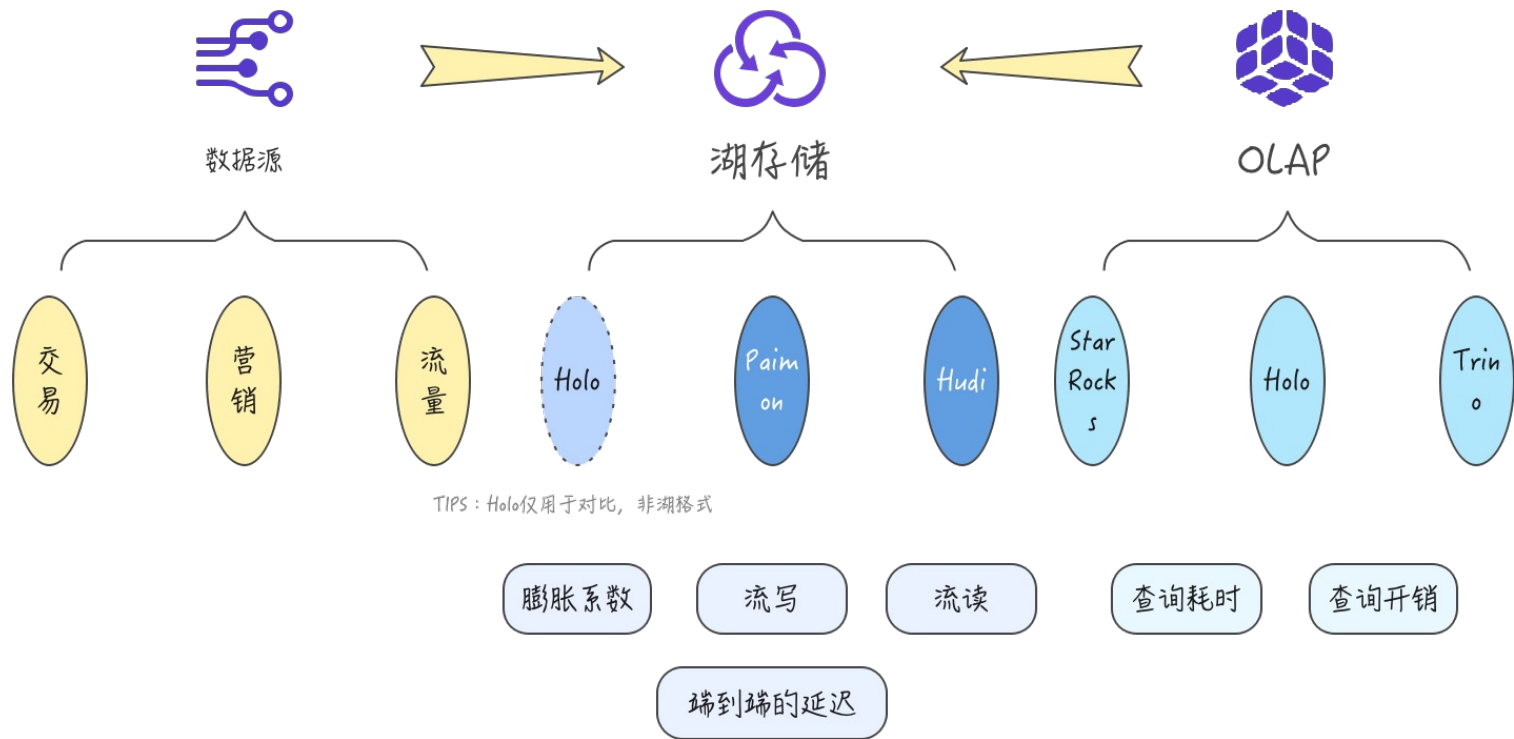
EmrRuntime-Extension (1.1.0)

Trino (422)

YARN (3.2.1)

ZooKeeper (3.6.3)

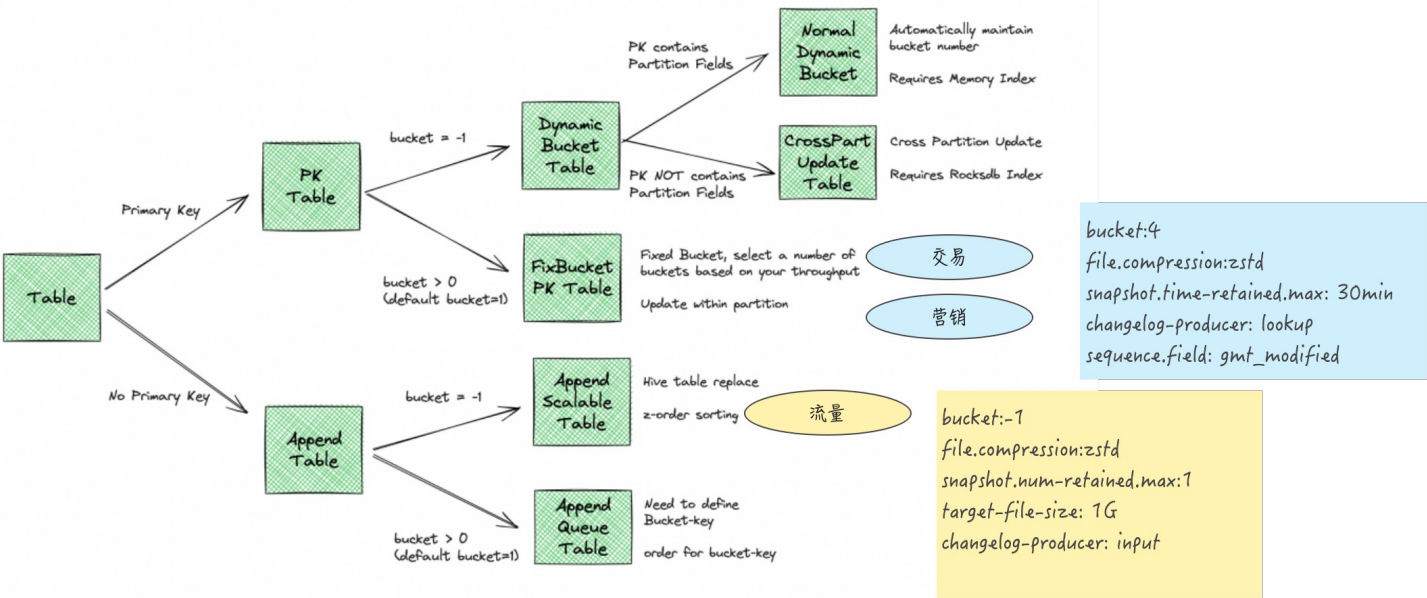
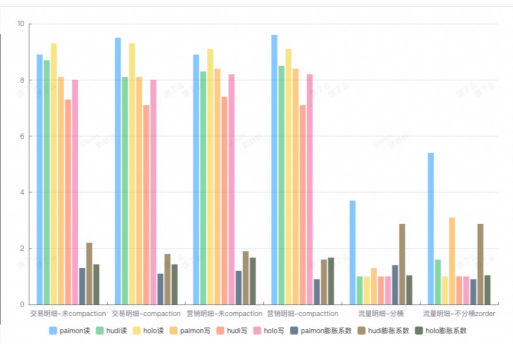
- flink: vvr-8.0.4-flink-1.17
- paimon: 0.6
- Hudi: 0.13.1
- StarRocks:
 - 版本: 3.1.5-1.12-1.4.4
 - 集群规格: 16 CU *12 = 192CU
- Trino:
 - 版本: trino-422
 - 集群规格: 200CU 3master +25core + 25个task : ecs.g6.xlarge
- holo
 - 版本: r2.0.33
 - 集群规格: 200CU



TIPS: 此测评结果基于饿了么自身业务属性结合部分专家经验得出, 仅供参考

Paimon VS Hudi

	读性能			写性能			存储膨胀系数		
	paimon	hudi	holo(内表)	paimon	hudi	holo(内表)	paimon	hudi	holo(内表)
交易明细-未compaction	8.9	8.7	9.3	8.1	7.3	8	1.3	2.2	1.43
交易明细-compaction	9.5	8.1	9.3	8.1	7.1	8	1.1	1.8	1.43
营销明细-未compaction	8.9	8.3	9.1	8.4	7.4	8.2	1.2	1.9	1.67
营销明细-compaction	9.6	8.5	9.1	8.4	7.1	8.2	0.9	1.6	1.67
流量明细-分桶	3.7	1	1.4	1.3	1	1.3	1.4	2.87	1.04
流量明细-不分桶zorder	5.4	1.6	1.4	3.1	1	1.3	0.9	2.87	1.04



时效性：
Paimon 1-5分钟 VS Hudi 10分钟



TIPS：此测评结果基于饿了么自身业务属性结合部分专家经验得出，仅供参考

StarRocks VS Holo VS Trino

- 各取3-5个线上常见SQL然后加权评分

	Trino	StarRocks	Holo
交易-点查	2.87	0.66	2.88
交易-简单聚合	7.07	0.98	8.87
营销-点查	1.45	0.06	1.57
营销-简单聚合	6.45	2.25	11.12
流量-点查	3.68	1.45	4.9
流量-简单聚合	17.85	6.23	50.2



为什么StarRocks更快？

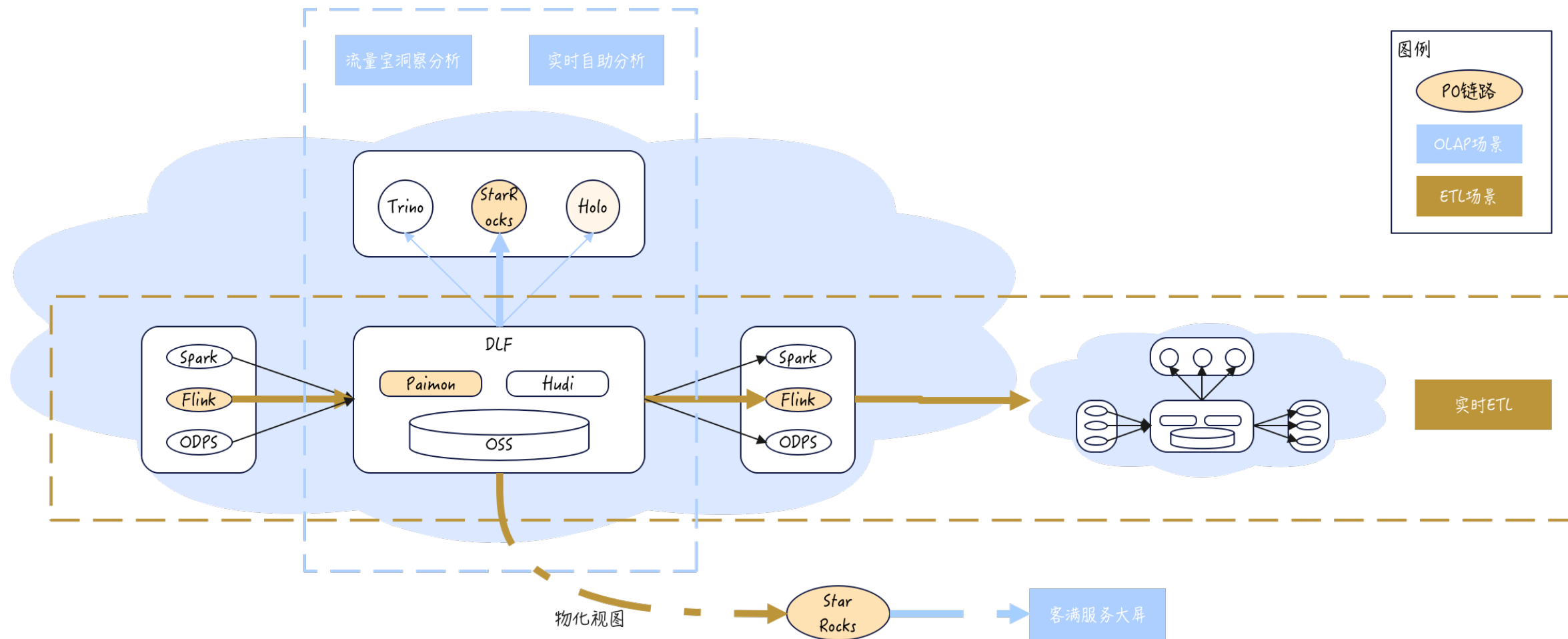
- JNI Connector
- 支持过滤下推
- 向量化执行引擎
- Paimon RO表支持
- 物化视图 + SQL透明改写
- Data Cache

QUERY PLAN

```
Limit (cost=0.00..220.50 rows=1 width=24)
-> Gather (cost=0.00..220.48 rows=1 width=24)
-> Limit (cost=0.00..220.43 rows=1 width=24)
-> Filter (cost=0.00..220.43 rows=1 width=24)
    Filter: ((ds >= '20231211'::text) AND (ds <= '20231217'::text) AND (city_id = '7'::bigint) AND (is_valid_order = '1'::bigint))
-> Forward (cost=0.00..164.96 rows=1 width=48)
    -> Seq Scan on Partitioned Table dl_dwd_ele_mkt_ord_subsidy_ri (cost=0.00..23.33 rows=71936681 width=48)
        Partitions selected: 4 out of 22
        Foreign Table Type: DLF
Optimizer: HQO version 2.0.0
```

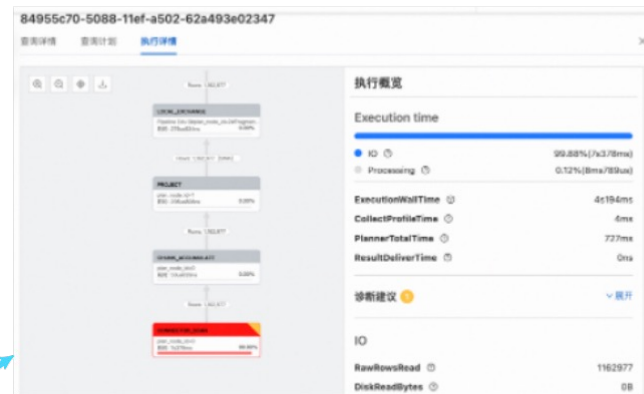
TIPS：此测评结果基于饿了么自身业务属性结合部分专家经验得出，仅供参考

实时湖仓落地探索



落地探索-DWD自助分析

35CU 支持峰值40K+的RPS, 较双流Join方案减少 50% +资源



VS Hologres实时数仓:

- 1-5分钟写入时效性
- 增加约15%查询RT
- + 减少约90%存储成本
- + 减少约50%flink开销

VS 离线ODPS数仓

- + 写入时效性小时级->分钟级
- + 提供增量更新能力
- + 查询RT N分钟级->秒级

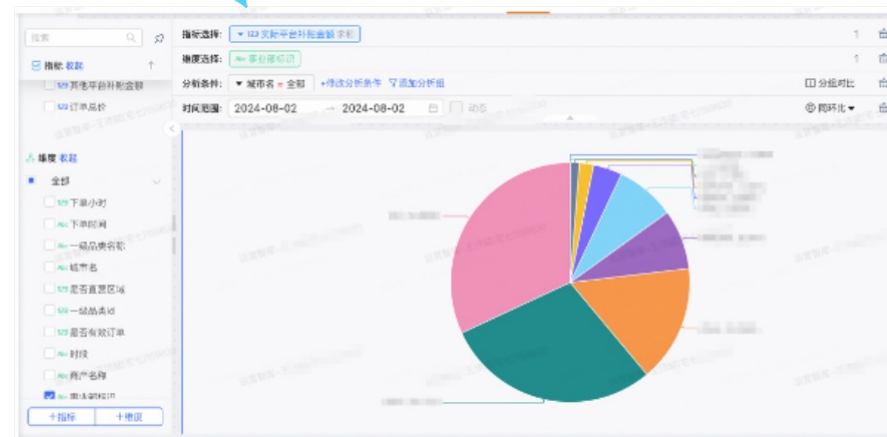
订单
补贴

PAIMON

StarRocks

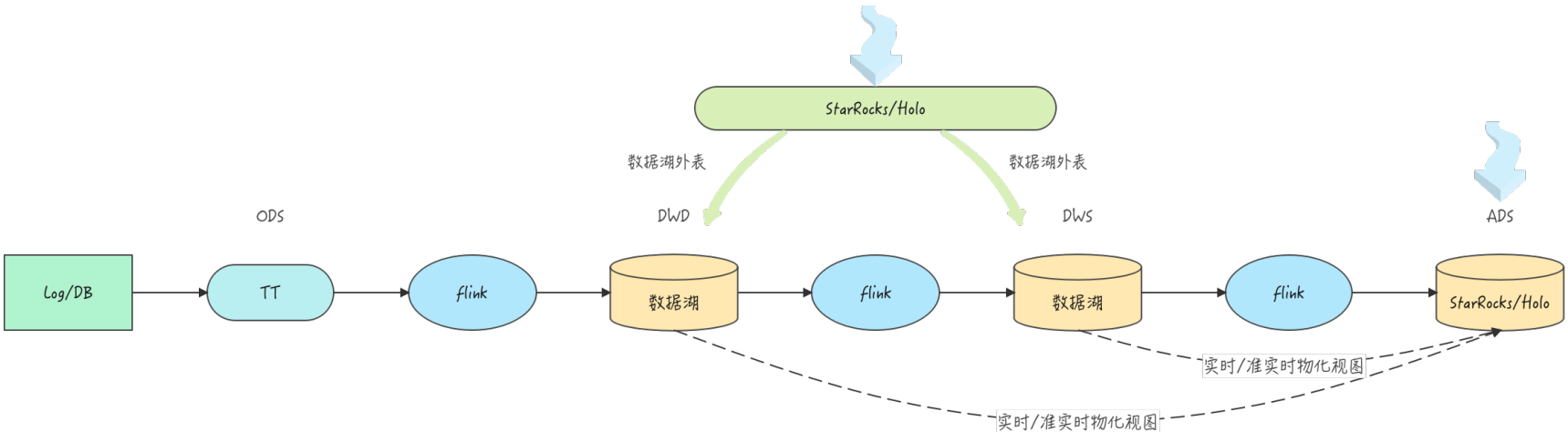
方舟自助分析

```
merge-engine:partial-update
fields.mkt_gmt_modified_cp.sequence-group:subsidy_amt,shop_subsidy_amt,...
fields.order_gmt_modified_cp.sequence-group:order_net_amt,order_source,shop_name...
snapshot.num-retained.min:1
snapshot.time-retained: 120 min
changelog-producer:lookup
```



03 实时湖仓规划及展望

实时数仓3.0 展望

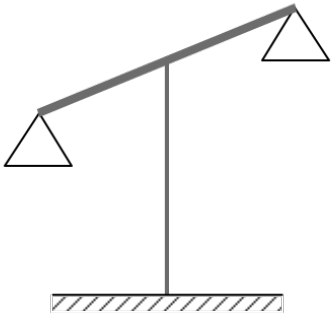


🔒 服务类型	📄 查询复杂度	📄 QPS	📄 交付方式 (排前面的优先)
数据产品	简单 复杂	高	flink任务+繁星服务 物化视图
数据产品	简单	中	数据湖底表 物化视图 flink任务+繁星服务
数据产品	复杂	中	物化视图 flink任务+繁星服务
数据产品	简单 复杂	低	数据湖底表 物化视图
临时分析	简单 复杂	低	数据湖底表
大促战报	简单 复杂	中	物化视图 flink任务+繁星服务

问题

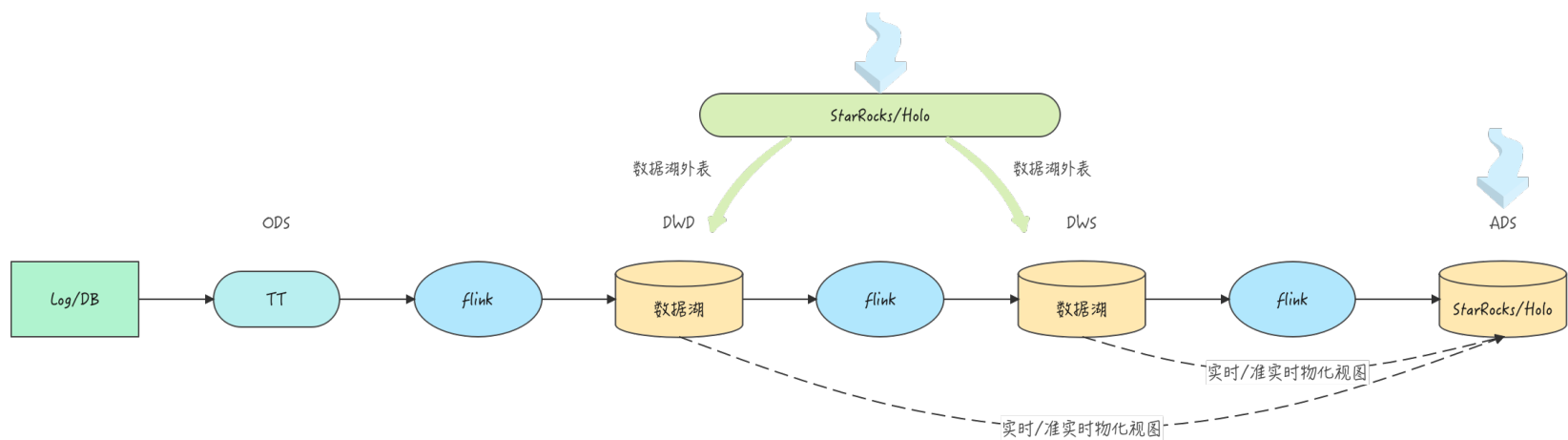
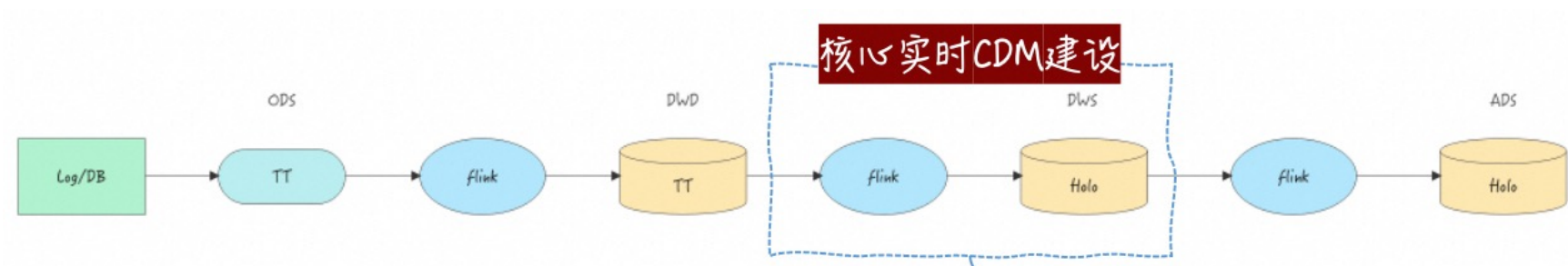
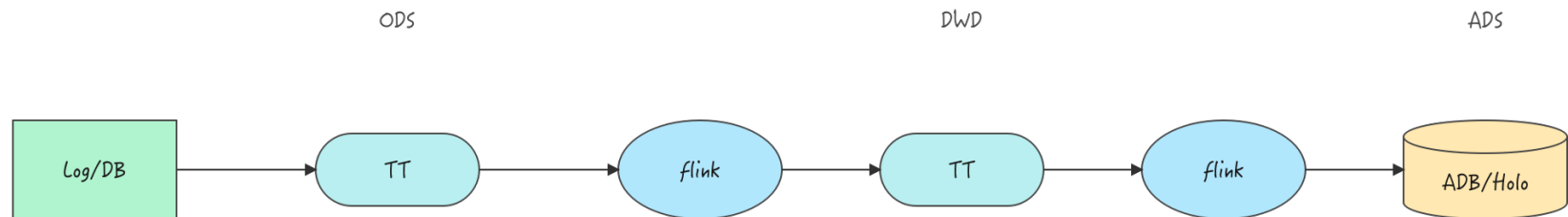
元数据管理
OSS小文件
集群权限控制
OSS带宽瓶颈
Paimon时效性

重点推进



StarRocks物化视图
Deletion Vector
MaxCompute 调度集成
入湖工具
异步Compaction
Paimon 实时中间层
Spark支持
OSS冷热分层

回顾





Apache Paimon



StarRocks

Thanks

Streaming Lakehouse Meetup