

货拉拉基于DORIS的 OLAP体系演进及建设方法

杨秋吉 大数据引擎负责人
张斌 大数据工程师



目录 CONTENT

01 背景介绍

02 OLAP体系演进（上）

03 OLAP体系演进（下）

04 总结思考与后续规划

01

背景介绍



货拉拉介绍

352

国内城市

58万

月活司机

760万

月活用户

8+

业务线

3+

IDC

1000+

机器数

20PB+

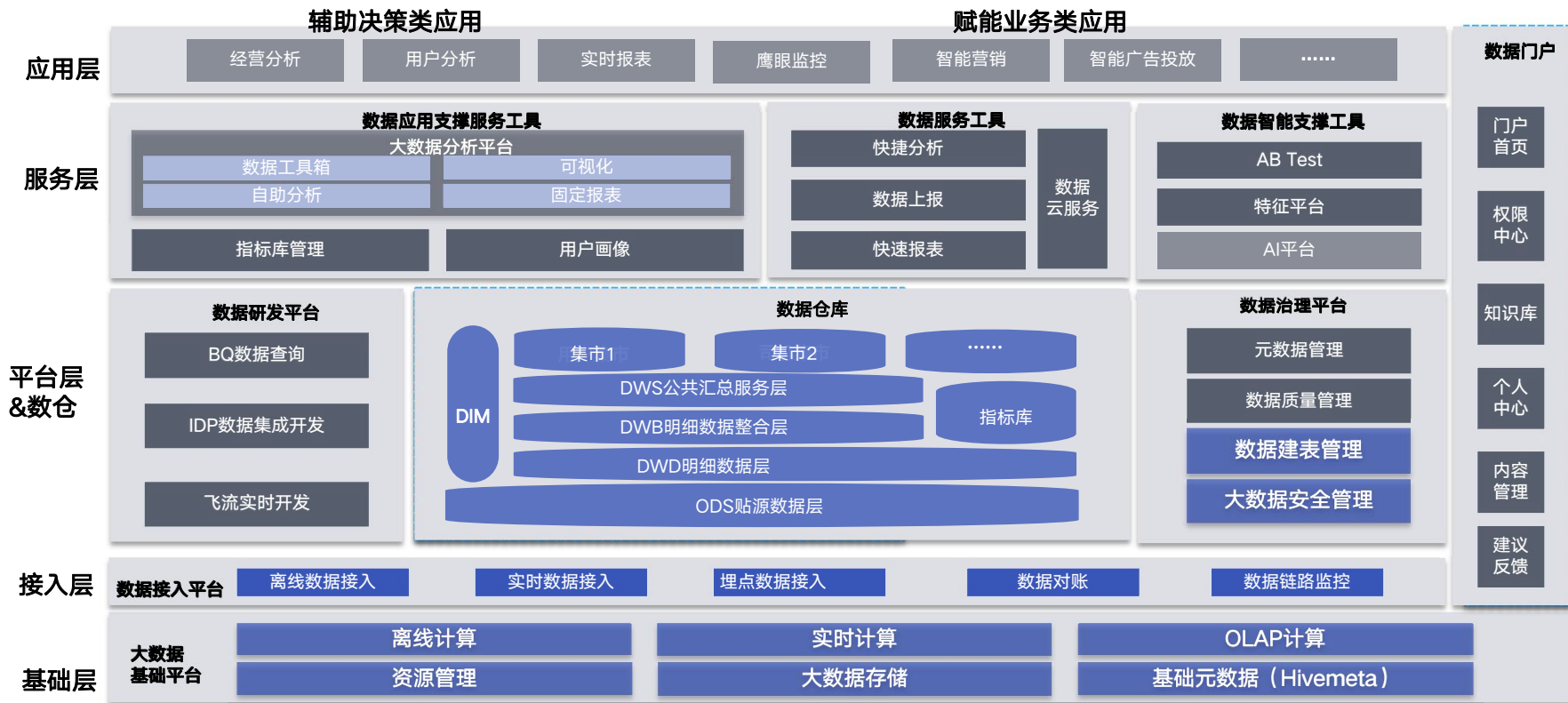
存储量

20K+

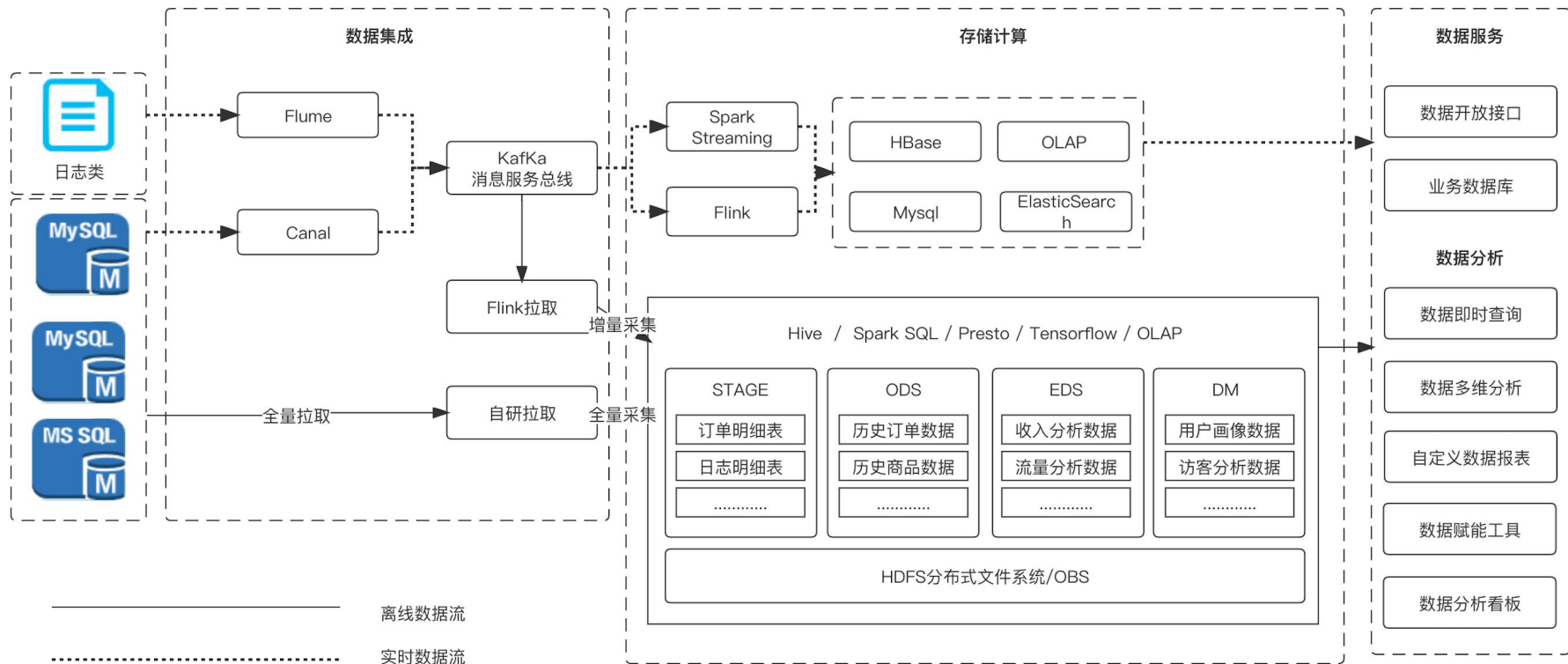
日均任务数

货拉拉-大数据

DataFun



货拉拉-大数据



02

OLAP体系演进（上）



OLAP 演进简介

货拉拉
DataFun

支撑业务：罗盘（实时智能决策系统，支持实时分析、诊断和策略以及复盘）

需求特点：数据实时导入、自由组合维度、实时聚合分析

引入引擎：Druid，提供单表预聚合查询能力

支撑业务：智能定位工具（基于埋点数据提供司机和订单的汇总和明细数据查询功能）

需求特点：单表明细查询和聚合分析、海量埋点数据实时导入

引入引擎：ClickHouse，提供单表明细查询且有数据高压缩率

支撑业务：AB Test和实时数仓

需求特点：多数据源（试验埋点数据、订单数据、用户数据、司机数据）关联分析

引入引擎：Doris，提供多张大表关联分析能力

2021 H1

OLAP 1.0: 孕育期

2021 H2

OLAP 2.0: 完善期

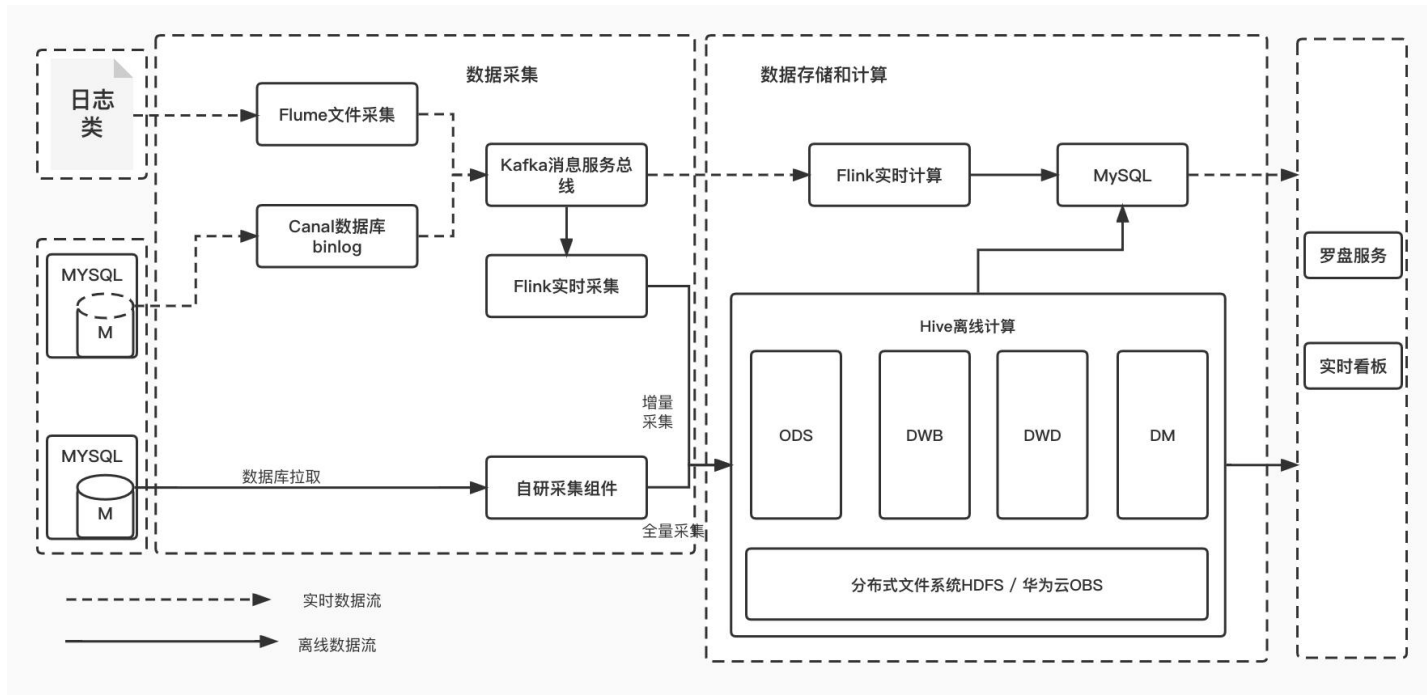
2022

OLAP 3.0: 成熟期

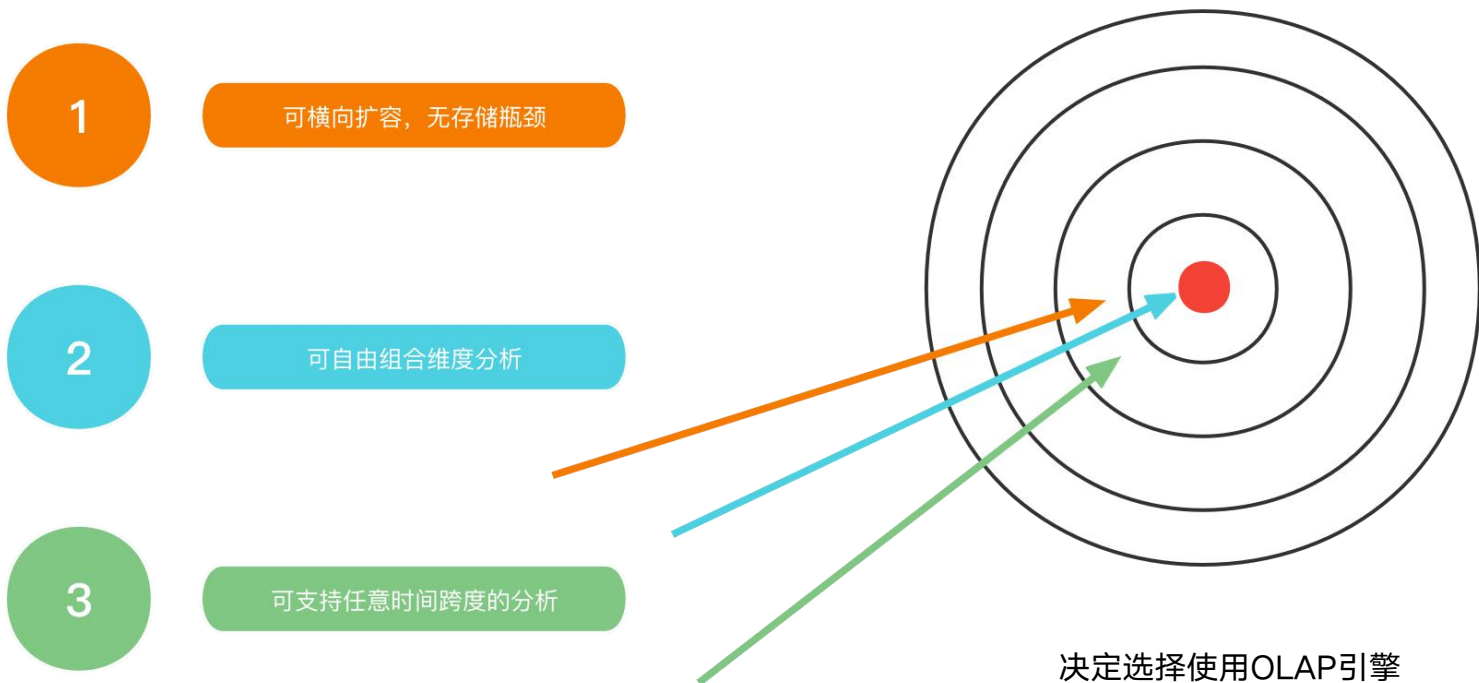
OLAP 1.0 - 业务场景

存在问题

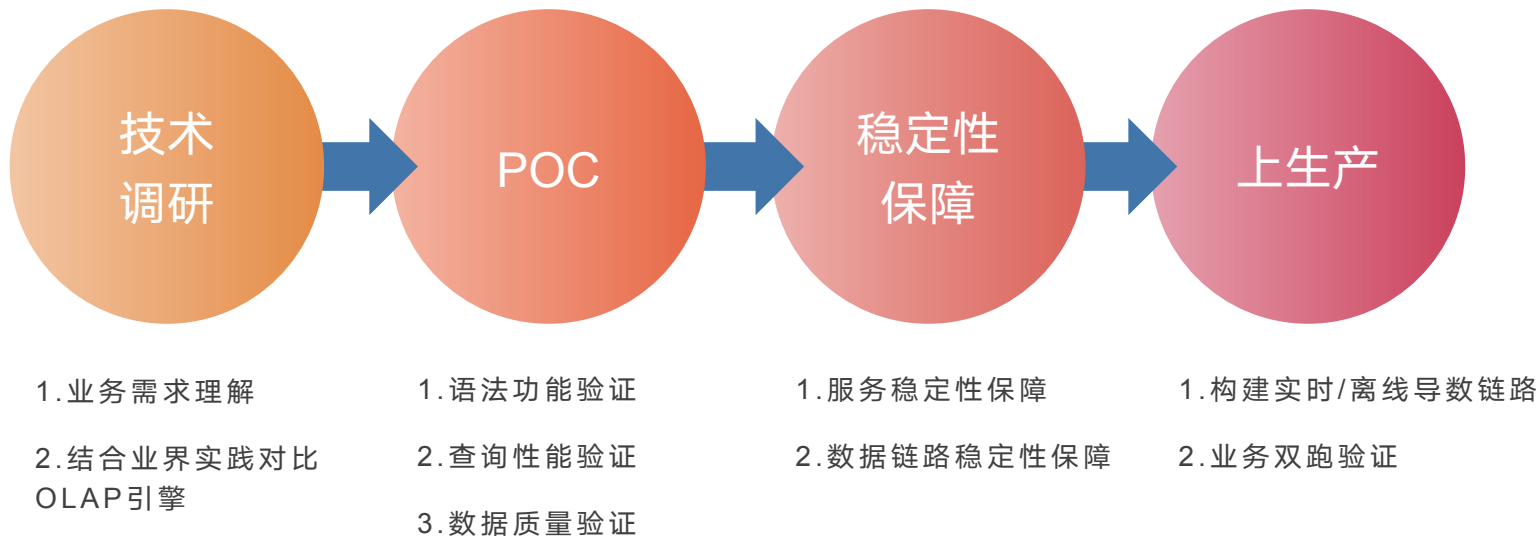
1. Mysql存储瓶颈
2. 开发成本高、效率低
3. 部分聚合需求不支持
(如长时间窗口聚合分析)



OLAP 1.0 - 需求分析



OLAP 1.0 - 解决思路



OLAP 1.0 - 技术调研

业务需求

1. 可横向扩容，无存储瓶颈；2. 可自由组合维度分析；3. 可支持任意时间跨度的分析

OLAP引擎	数据导入延迟	实时数据导入语义	数据查询延迟	支持多维分析	SQL支持程度	支持明细查询	JOIN支持度	支持复杂数据类型	集群成本	可控性	扩展性	可运维性
Druid	支持实时	Exactly-Once	低 (亚秒~秒级)	支持	较完善	可支持 (关闭rollup)	很低	不支持	中 (角色较多, 依赖HDFS)	高 (JAVA开发, 社区活跃, 应用多)	高	高
ClickHouse	支持实时	At-least-once	低 (亚秒~秒级)	支持	较完善	支持	一般 (内存JOIN)	支持 (MAP/JSON/Array等)	低 (数据压缩率高)	中 (C++开发, 社区一般活跃, 应用较多)	高	中
Kylin	分钟级/天级	N/A	非常低 (亚秒级)	支持	非常完善	不擅长	不支持	不支持	高	高 (JAVA开发, 社区活跃)	高	高
Presto	小时级/天级	N/A	一般 (秒~分钟级)	支持	非常完善	不擅长	支持	支持	低 (无存储)	高 (JAVA开发)	高	高
Doris	支持实时	Exactly-Once	低 (亚秒~秒级)	支持	非常完善	支持	支持	不支持 (2022 Roadmap 已有规划)	中	中 (JAVA/C++开发, 社区较活跃, 应用较多)	高	高

OLAP 1.0 - POC验证

01 语法功能验证

1. 收集业务SQL，提取SQL Pattern
2. Druid建表和SQL改写：UDF、rollup语义、count distinct语义

02 性能验证

1. 采用业务真实数据和SQL测试
2. 关闭Cache，统计P75、P90、P99的查询时间
3. 结合Arthas火焰图分析性能
4. 性能调优：优化建表数和索引逻辑、参数调整、物化视图

03 数据准确性验证

1. 选择基准值：hive表
2. hive和druid双跑验证
(发现StringLast函数在特定场景下计算值不稳定)

OLAP 1.0 - 稳定性保障



事前【故障预防】

1. 容量规划：压测（导数和查询）、容量评估
2. 容灾演练：扩容、缩容、停服务、HA验证
3. 恢复预案：前期业务双跑，链路随时可切换



事中【故障处理】

1. 发现能力：全链路监控告警（机器、服务、任务）
2. 定位能力：研究引擎原理，关注业务分享，定位大盘
3. 恢复和规避能力：业务双跑



事后【故障整改】

1. 故障复盘
2. 整改落地

OLAP 1.0 - 上生产



2. OLAP 上线观察

1. 业务查询走Druid
2. 业务随时能切回MySQL

3. OLAP 稳定运行

1. MySQL链路下线

1. OLAP 测试阶段

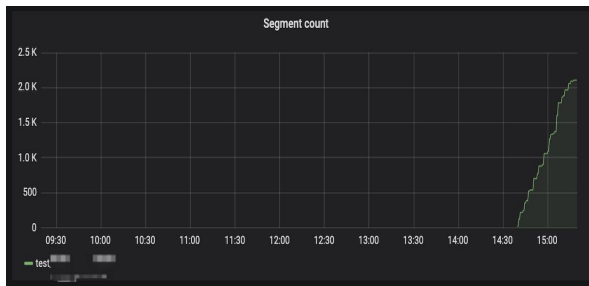
1. 业务数据接入Druid
2. 线上查询走MySQL库
3. 验证Druid数据质量和稳定性

OLAP 1.0 - 问题总结

01

实时数据乱序

- ✓ 影响：
会产生大量的小文件（Segment），影响查询效率，增大元数据压力
- ✓ 解决办法：
在上游Flink里过滤异常数据



02

StringLast函数结果值不稳定

- ✓ 影响：
多次查询的结果值不一致
- ✓ 解决办法：
新增StringLastMax和StringLastMin函数

time	order_id	big_bus_data	br_city	city	eilb	s	k	l	c	y	vehicle_type	n	m	mon	total_subsidy_flow	amounts
2021-05-22T01:00:00.0002			深圳	0	1	1	1	0	其它		0	0	0		128.0	128.0
2021-05-22T01:00:00.0002			深圳	0	1	1	1	0	其它		0	0	0		113.0	113.0
2021-05-22T01:00:00.0002			深圳	0	1	1	1	0	其它		0	0	0		128.0	128.0
2021-05-22T01:00:00.0002			深圳	0	1	1	1	0	其它		0	0	0		113.0	113.0
2021-05-22T01:00:00.0002			深圳	0	1	1	1	0	其它		0	0	0		128.0	128.0
2021-05-22T01:00:00.0002			深圳	0	1	1	1	0	其它		0	0	0		113.0	113.0
2021-05-22T01:00:00.0002			深圳	0	1	1	1	0	其它		0	0	0		128.0	128.0
2021-05-22T01:00:00.0002			深圳	0	1	1	1	0	其它		0	0	0		113.0	113.0

03

无高效的精准去重函数

- ✓ 影响：
离线场景业务需要精准去重能力
- ✓ 解决办法：
1. 引入社区里快手提供的patch，合入0.20版本
2. 另外新增SQL API，并支持导入hive的bitmap二进制字符串类型

OLAP 2.0 - 业务需求分析

能查询司机详情信息

支持明细查询

能支持单天近10亿实

时数据写入

实时写入吞吐高

业务需求分析

能统计分析趋势指标

支持多维聚合分析

能支持复杂数据结构

需支持Map和Json格式数据的高效写入和查询

OLAP 2.0 - 业务需求分析

1、点击汇总情况

汇总情况 司机明细记录 订单明细 问题定位

2、选择需要查询的时间段

时间: 2021-09-13 00:00:00 至 2021-09-13 23:59:59 数据模型: 查询所有

司机ID: 司机ID 汇总类型: pk司机抢单汇总指标 搜索

3、输入司机ID

4、选择需要查询的指标

司机ID	接单日期	收到推送订单总数	多人PK单总数	多人PK单中的
------	------	----------	---------	---------

Dashboard / 接单查询 / 智能定位工具

Dashboard 智能定位工具

汇总情况 司机明细记录 订单明细 问题定位

1、选择查询时间段

2、输入订单ID (可选)

时间: 2021-09-13 00:00:00 至 2021-09-13 23:59:59 订单ID: 订单ID

司机ID: 司机ID 业务类型: 后置过滤分析 搜索

4、选择业务类型

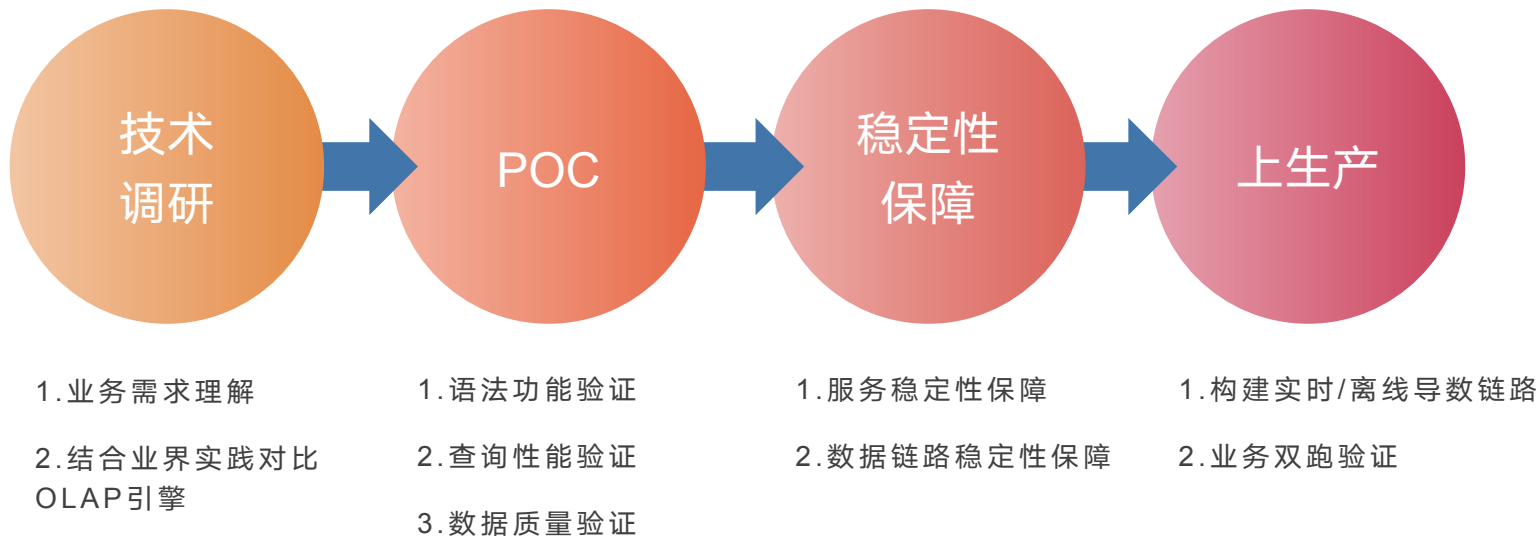
3、输入司机ID

订单ID	司机ID	AB单	导航距离	直	前置过滤分析	累计推送轮数	过滤原因	过滤条件
------	------	-----	------	---	--------	--------	------	------

暂无数据

5、点击“搜索”查询

OLAP 2.0 - 解决思路 (复用1.0的解决思路)



OLAP 2.0 - 技术调研

业务需求

1. 能同时支持明细查询和聚合分析；2. 实时数据写入吞吐高；3. 支持Map和Json格式数据的高效写入和查询

OLAP引擎	数据导入延迟	实时数据导入语义	数据查询延迟	支持多维分析	SQL支持程度	支持明细查询	JOIN支持度	支持复杂数据类型	集群成本	可控性	扩展性	可运维性
Druid	支持实时	Exactly-Once	低 (亚秒~秒级)	支持	较完善	可支持 (关闭rollup)	很低	不支持	中 (角色较多, 依赖HDFS)	高 (JAVA开发, 社区活跃, 应用多)	高	高
ClickHouse	支持实时	At-least-once	低 (亚秒~秒级)	支持	较完善	支持	一般 (内存JOIN)	支持 (MAP/JSON/Array等)	低 (数据压缩率高)	中 (C++开发, 社区一般活跃, 应用较多)	高	中
Kylin	分钟级/天级	N/A	非常低 (亚秒级)	支持	非常完善	不擅长	不支持	不支持	高	高 (JAVA开发, 社区活跃)	高	高
Presto	小时级/天级	N/A	一般 (秒~分钟级)	支持	非常完善	不擅长	支持	支持	低 (无存储)	高 (JAVA开发)	高	高
Doris	支持实时	Exactly-Once	低 (亚秒~秒级)	支持	非常完善	支持	支持	不支持 (2022 Roadmap 已有规划)	中	中 (JAVA/C++开发, 社区较活跃, 应用较多)	高	高

03

OLAP体系演进（下）



OLAP 3.0 - 需求分析

多表关联场景需求强烈

随着公司业务的发展，多个产品线对于多数据源关联场景下在线多维分析需求越来越迫切。

AB-Test:

大数据量的多表关联场景（各类埋点数据）进一步做业务分析

实时数仓:

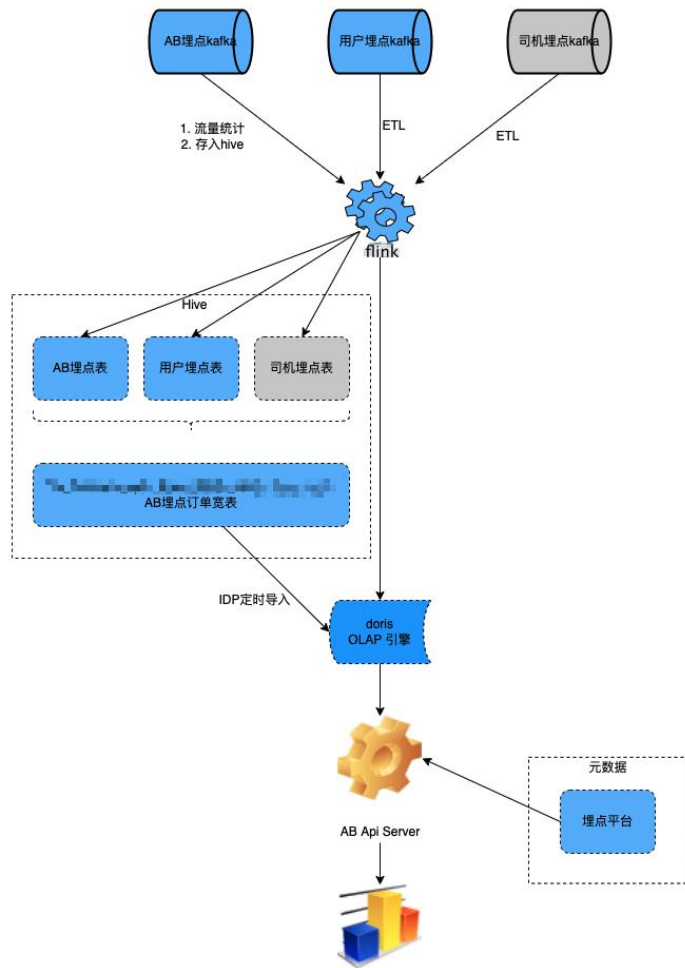
多表关联场景支持，任意时间跨度聚合分析

OLAP 3.0 - 需求分析

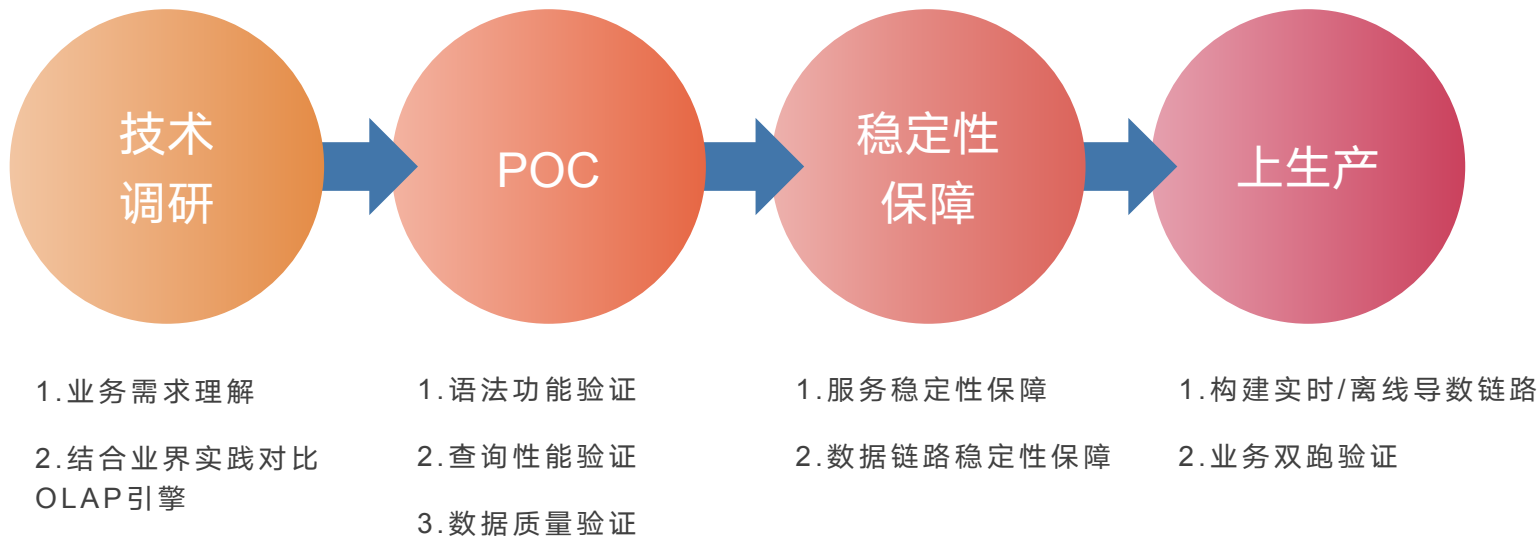
AB-Test:

- 业务持续增长更多靠产研驱动、离不开科学的AB实验；
- AB平台提供科学分流、智能统计能力，助力业务决策、实现业务增长；

通过AB数据与用户埋点数据**关联**，可以更直接有力的证明AB策略的优劣。



OLAP 3.0 - 解决思路 (复用1.0的解决思路)



OLAP 3.0 - 技术调研

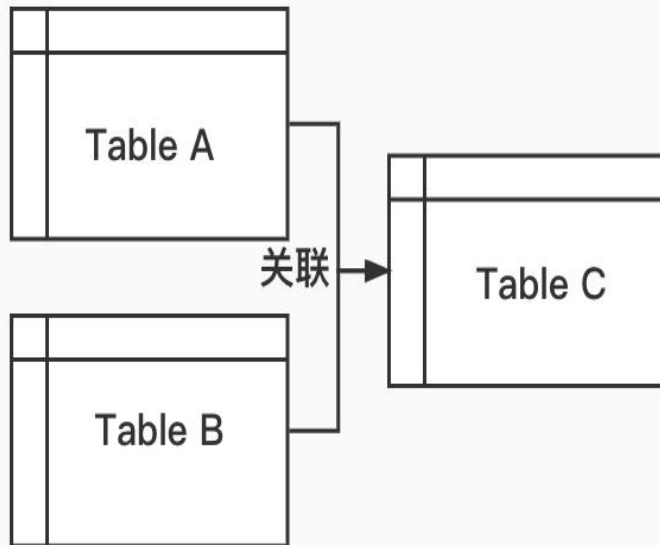
现有OLAP引擎支持大表JOIN能力较弱，druid/clickhouse均不支持千万级甚至亿级数据量下的大表JOIN。

Druid

采用内存字典（KV格式）方式只能支持简单维表JOIN；

Clickhouse

基于内存做MapJoin支持少量数据下的JOIN；



OLAP 3.0 - 技术调研

业务需求

1. 数据导入准确性; 2. 支持大表join;

OLAP引擎	数据导入延迟	实时数据导入语义	数据查询延迟	支持多维分析	SQL支持程度	支持明细查询	JOIN支持度	支持复杂数据类型	集群成本	可控性	扩展性	可运维性
Druid	支持实时	Exactly-Once	低 (亚秒~秒级)	支持	较完善	可支持 (关闭rollup)	很低	不支持	中 (角色较多, 依赖HDFS)	高 (JAVA开发, 社区活跃, 应用多)	高	高
ClickHouse	支持实时	At-least-once	低 (亚秒~秒级)	支持	较完善	支持	一般 (内存JOIN)	支持 (MAP/JSON/Array等)	低 (数据压缩率高)	中 (C++开发, 社区一般活跃, 应用较多)	高	中
Kylin	分钟级/天级	N/A	非常低 (亚秒级)	支持	非常完善	不擅长	不支持	不支持	高	高 (JAVA开发, 社区活跃)	高	高
Presto	小时级/天级	N/A	一般 (秒~分钟级)	支持	非常完善	不擅长	支持	支持	低 (无存储)	高 (JAVA开发)	高	高
Doris	支持实时	Exactly-Once	低 (亚秒~秒级)	支持	非常完善	支持	支持	不支持 (2022 Roadmap 已有规划)	中	中 (JAVA/C++开发, 社区较活跃, 应用较多)	高	高

OLAP 3.0 - POC

功能验证

TPC-DS数据集验证，业务数据真实场景验证；

性能验证

多表关联场景，单天数据查询，TP75耗时9s；

数据质量

TPC-DS数据集、业务侧真实数据分别在hive和doris侧双跑比对；



OLAP 3.0 - 稳定性保障



事前【故障预防】

1. 容量规划确认容量指标
2. 压测确认容量最大水位



事中【故障处理】

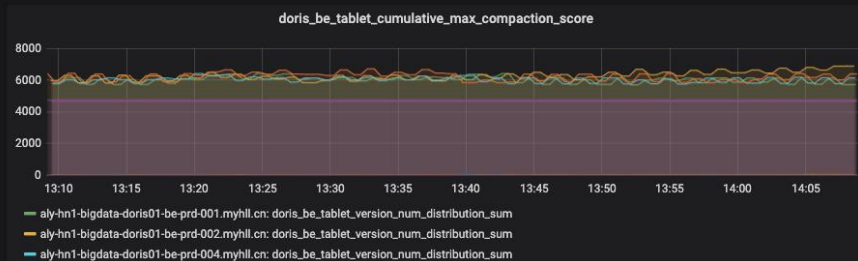
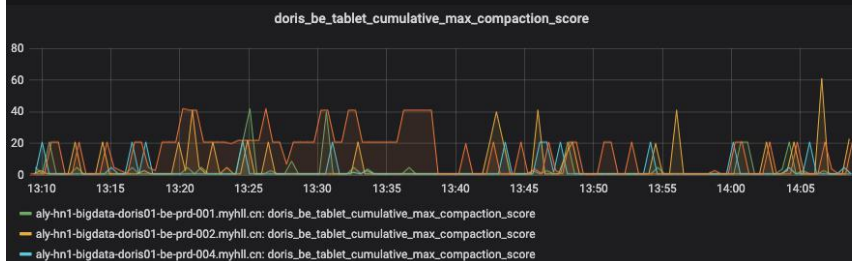
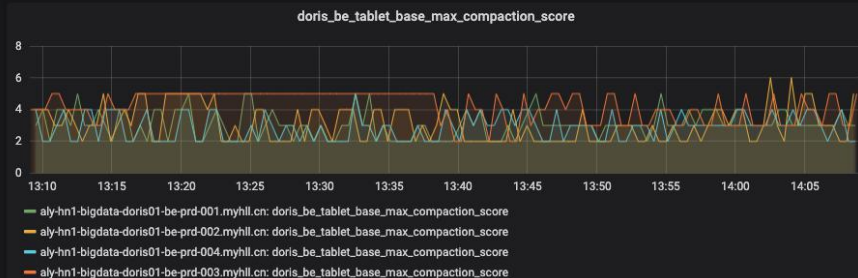
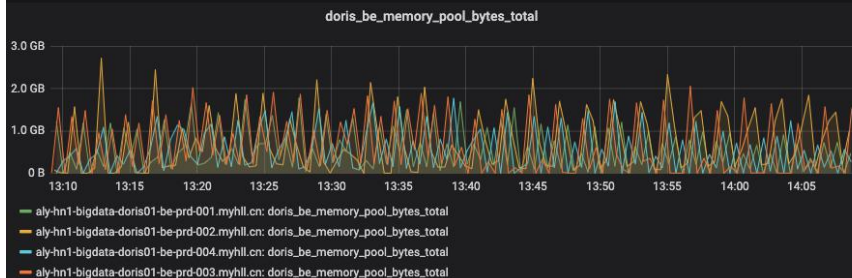
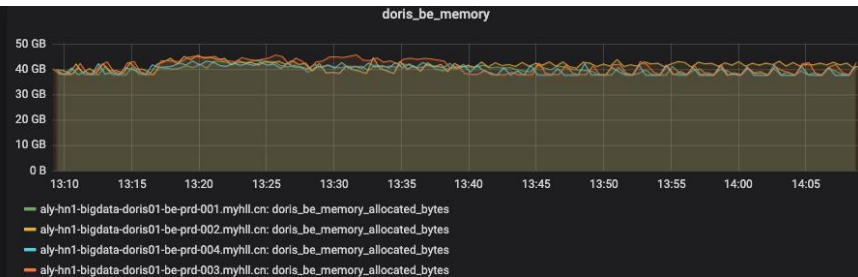
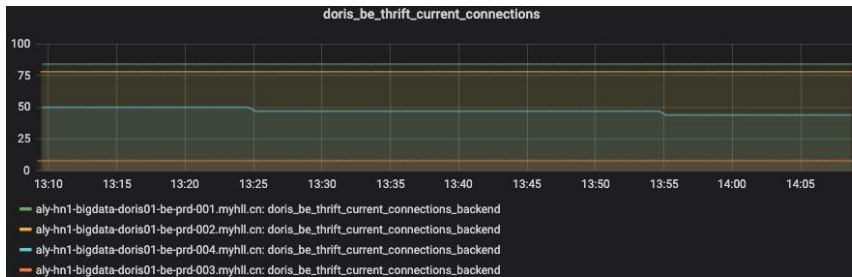
1. 发现能力：全链路监控告警
2. 定位能力：研究引擎原理，关注业务分享，定位大盘



事后【故障整改】

1. 故障复盘
2. 整改落地

OLAP 3.0 - 稳定性保障



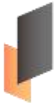
OLAP 3.0 - 问题总结

问题1-查询性能优化:

需求: 查询7天数据, $RT \leq 5s$

优化前: 查询7天数据耗时30s

	优化方法	原理	效果
1	小表 join 大表 -> 大表 join 小表	doris默认使用右表数据构建hashtable	30s --> 16s
2	t1 join (t2 union all t3) -> (t1 join t2) union all (t1 join t3)	利用RuntimeFilter特性 运行时采用BloomFilter, 将HashKey条件下推到 大表Scan时过滤	16s --> 5s



OLAP 3.0 - 问题总结

问题2-UnhealthyTablet不下降，查询报错-230

场景：不停flink写任务，be机器交替重启，重启完后出现unhealthyTablet

This page lists the system info, like /proc in Linux.

Current path: //statistic

DbId	DbName	TableNum	PartitionNum	IndexNum	TabletNum	ReplicaNum	UnhealthyTabletNum	Incon
11001		1	1	1	1	3	0	0
18642		2	2	2	2	6	0	0
13032		6	13	13	290	872	18	0

原因：1. coordinator be在两阶段提交执行Commit后publish前被重启了
2. max_running_txn_num_per_db参数配置过大，compaction压力大

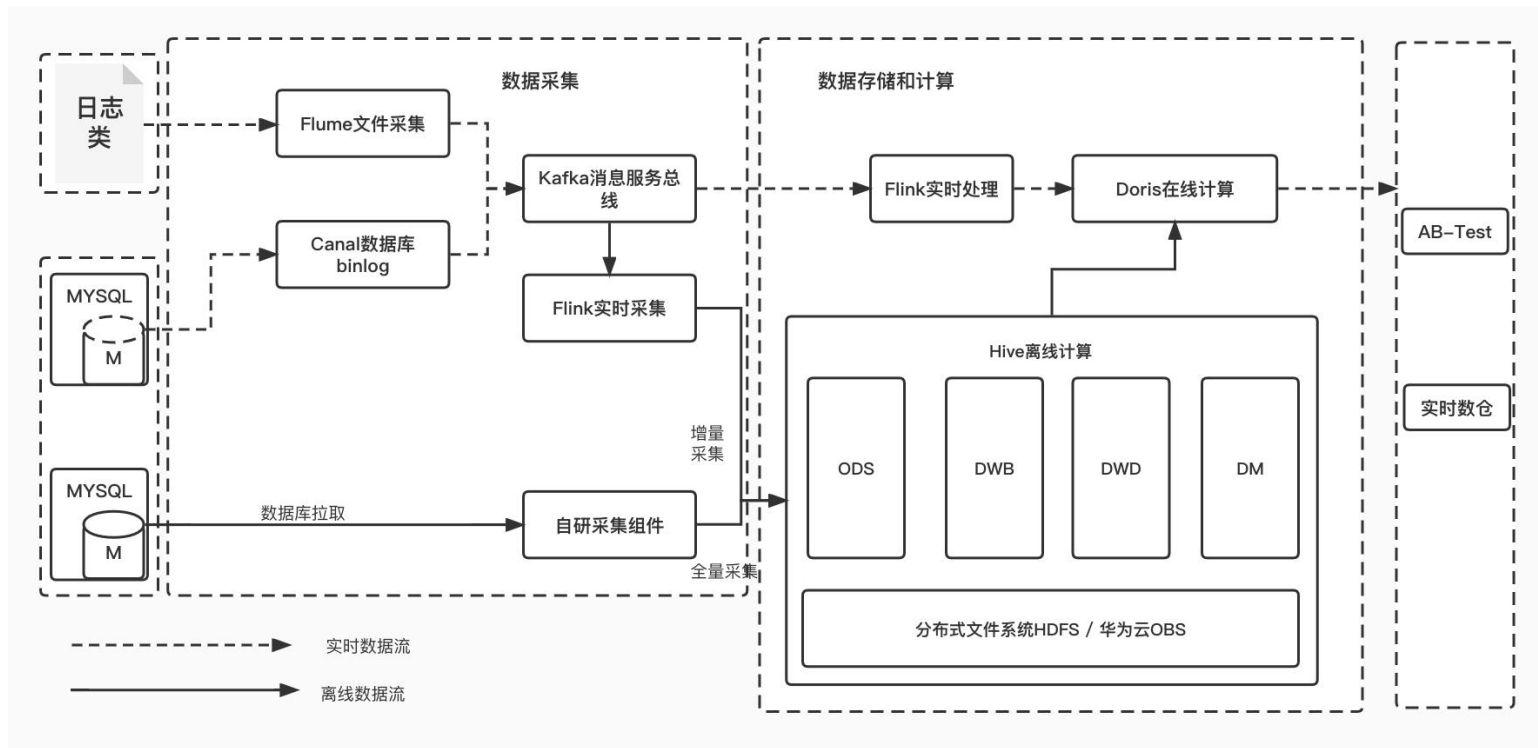
解决办法：1. 引入社区1.10 patch (issue-9267)
2. 数据恢复

OLAP 3.0 - 参数优化

使用过程中的一些参数优化：

配置项	含义	默认值	修改值
enable_profile	进行查询分析	FALSE	TRUE
exec_mem_limit	单个查询的内存限制	2G	8G
parallel_fragment_exec_instance_num	BE上执行实例的个数	1	8
compaction_task_num_per_disk	并发compaction数量	2	4
streaming_load_json_max_mb	控制单次streamLoad数据量	100	150
max_segment_num_per_rowset	限制rowset中segment的数量	200	500
enable_sql_cache	SQL级缓存	FALSE	TRUE
enable_partition_cache	分区级缓存	FALSE	TRUE

OLAP 3.0 - 数据流



04

总结思考与后续规划



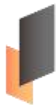
总结与思考

总结

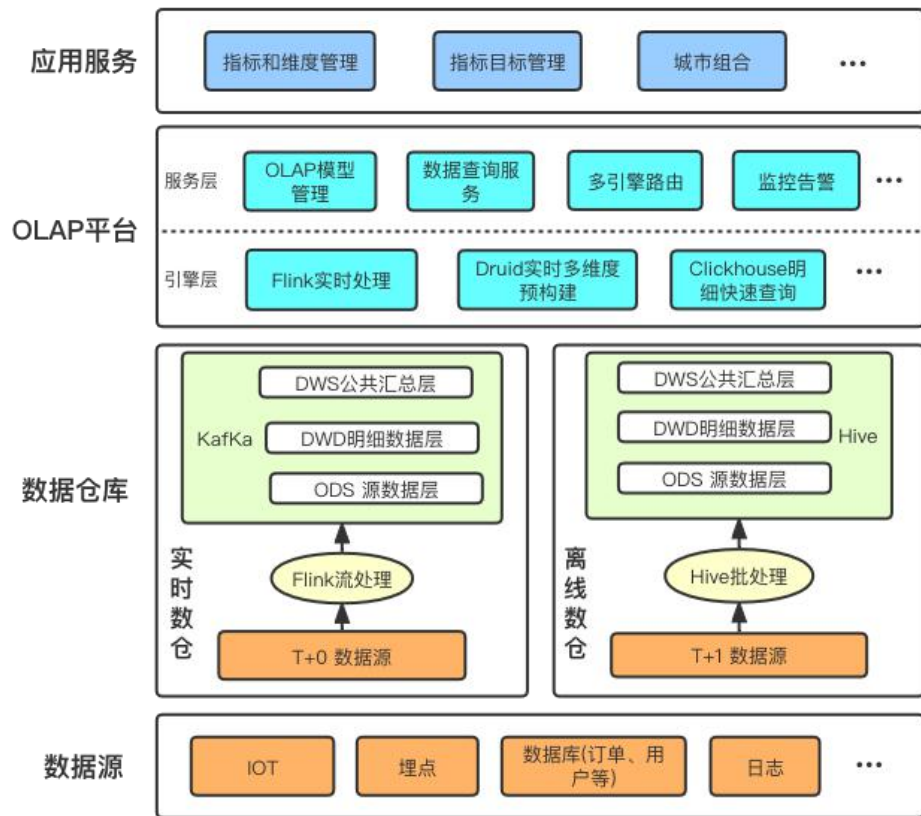
1. 从业务需求出发匹配合适引擎，为业务精细化运营提供技术支持；
2. 摸索出一套较完善的上线流程及稳定性保障体系方案，为业务平稳运行提供能力保障；

思考

1. 没有单种引擎能高效支持各种场景，需要针对需求特点选取合适的引擎；



后续规划



OLAP平台化:
自助化建模; 多引擎路由、支持各类聚合、明细、关联等场景。

后续规划

01

高效

支持更多业务场景，提升开发、决策效率，降本增效

02

稳定

深入内核原理，提供二次开发支持；完善监控告警体系

03

内核演进

Doris逐步替换Druid，以Doris为主引擎、Clickhouse为辅

非常感谢您的观看

货拉拉 | DataFun.

