

数据来源：数据库产品上市商用时间



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



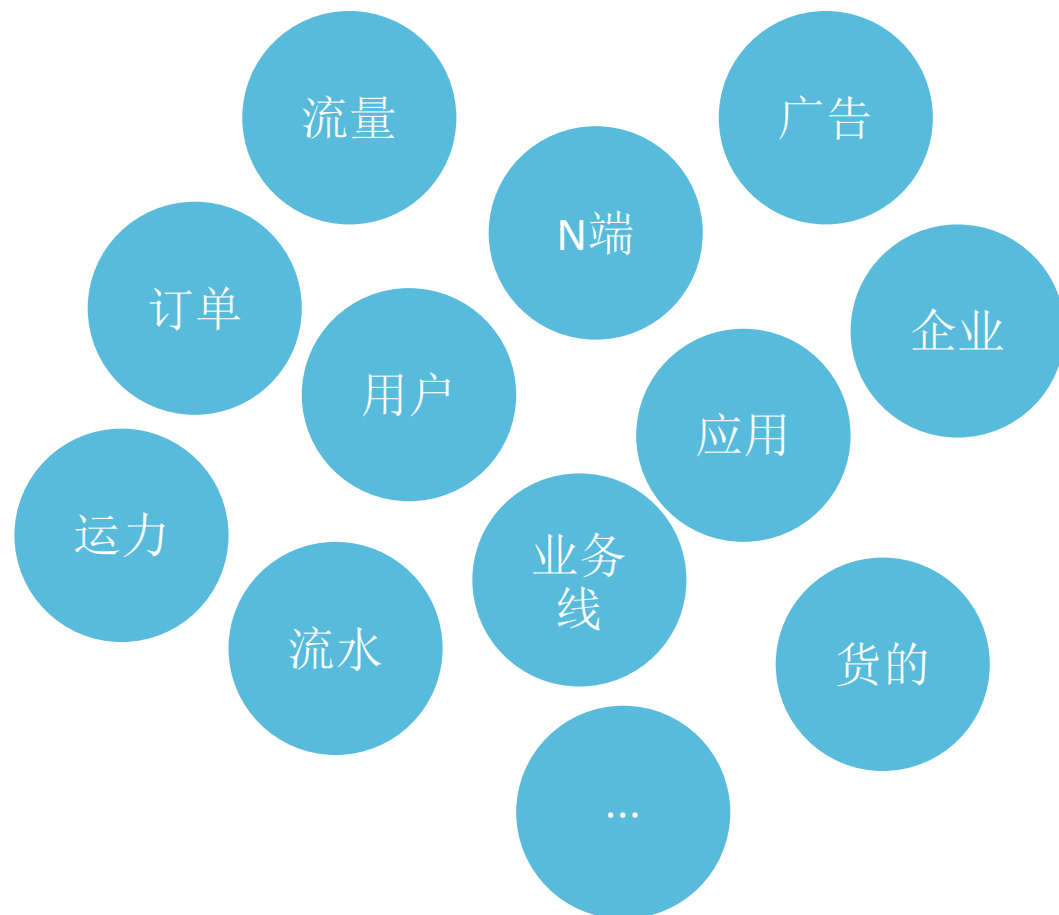
线上直播 | 2022/12/14-16



快狗打车实时数仓演进之路

杨铮 快狗打车 数据负责人

业务背景



业务复杂度高

流量大

应用场景多

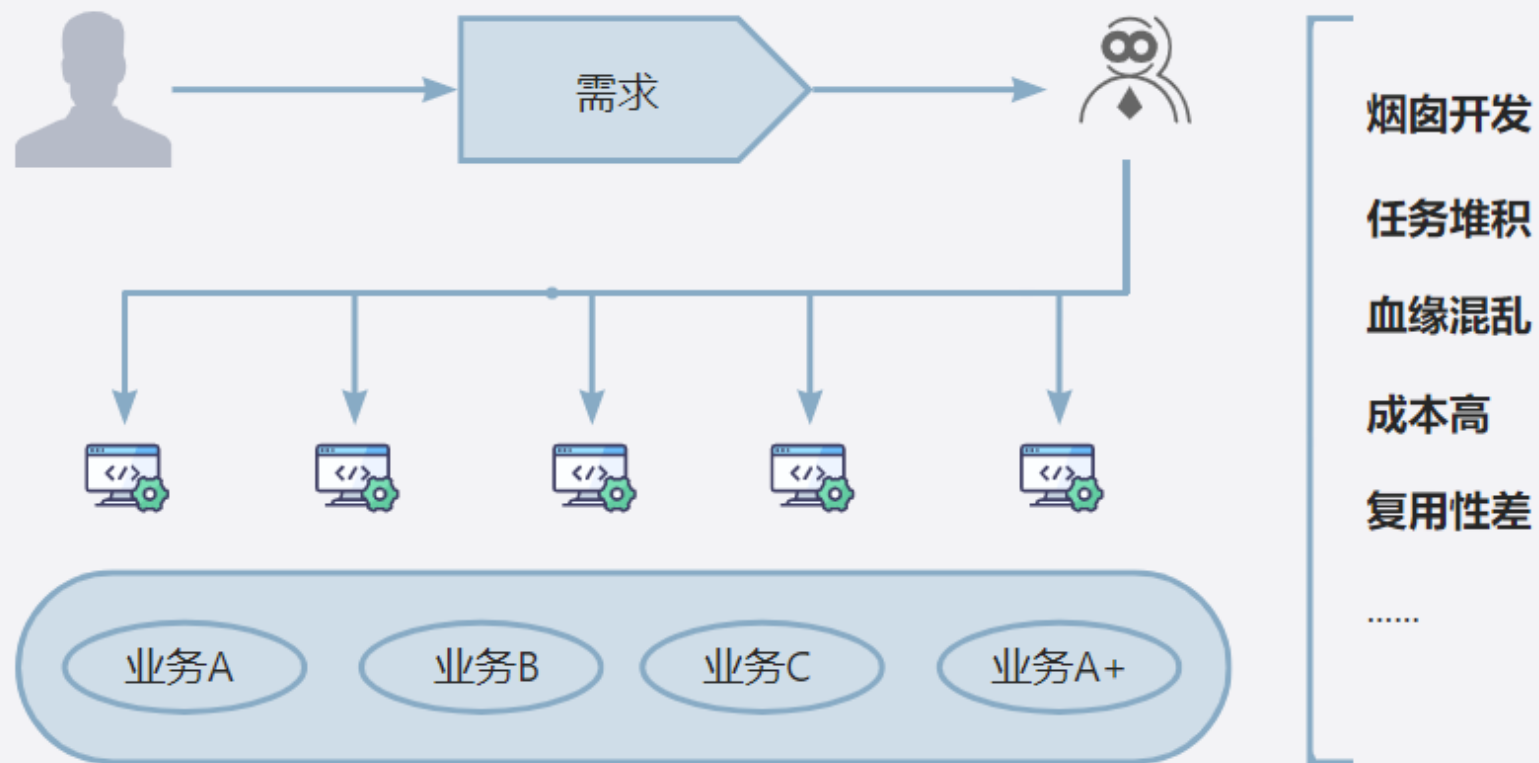
对实时数据需求多

开发时长存在问题

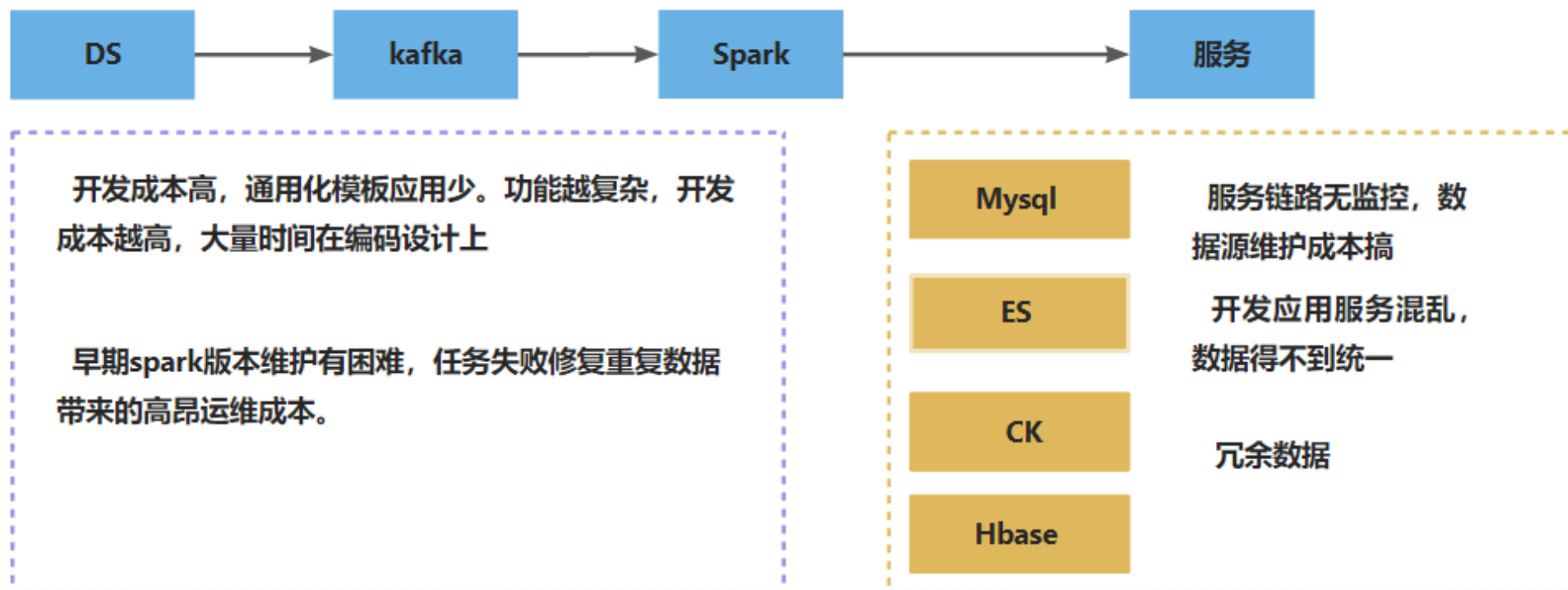
.....

1. 以往的开发流程和实时计算

历史开发流程



历史实时计算



2. 从上云开始转变

- spark+多源
- 存在诸多问题

2019

- 上云完成
- OneData
- OneService

2020

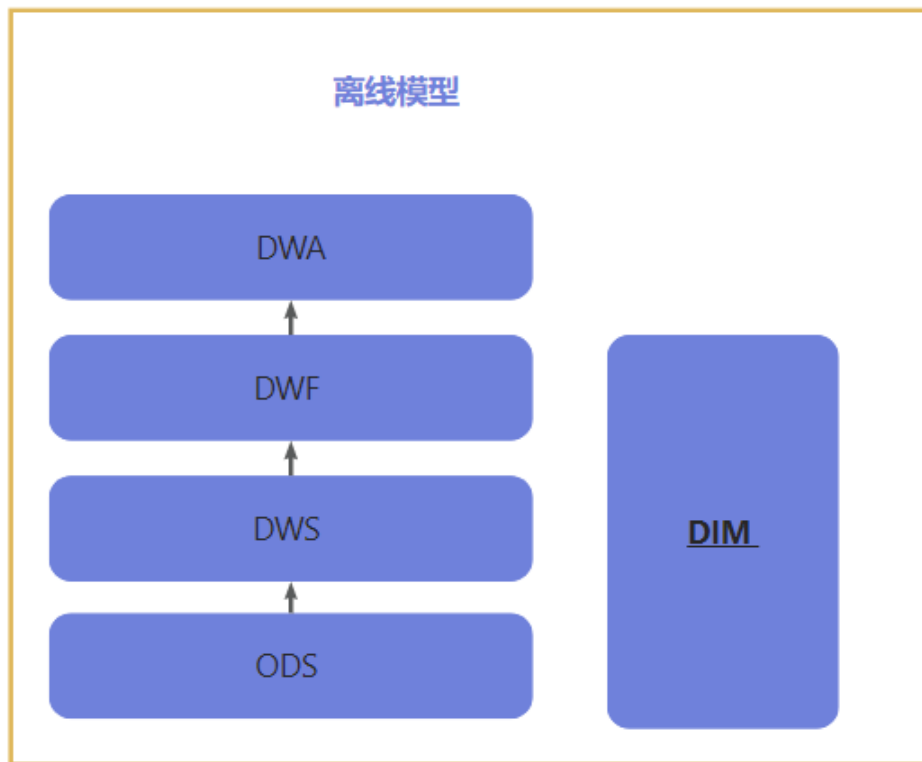
- 智能化系列

2022

3.解决痛点

模型升级-分层模型

摆脱混乱开发，建设分层模型，数据重复利用。采取实时==离线的方案



ODS：基础数据层，在实时仓库，ODS数据层在消息中间件：kafka

DWS：服务数据层，作用为数据脱敏，轻汇。在业务上应用于分析，数仓进行数据查询补充。在实时仓库，DWS数据层在消息中间件：kafka或者OLAP系统。

DWF：事实数据，即宽表所在。该层次数据主要为三种类型宽表建设，用于日常分析洞察，数据查询等。在实时仓库，DWF数据层在消息中间件：kafka或者OLAP系统。

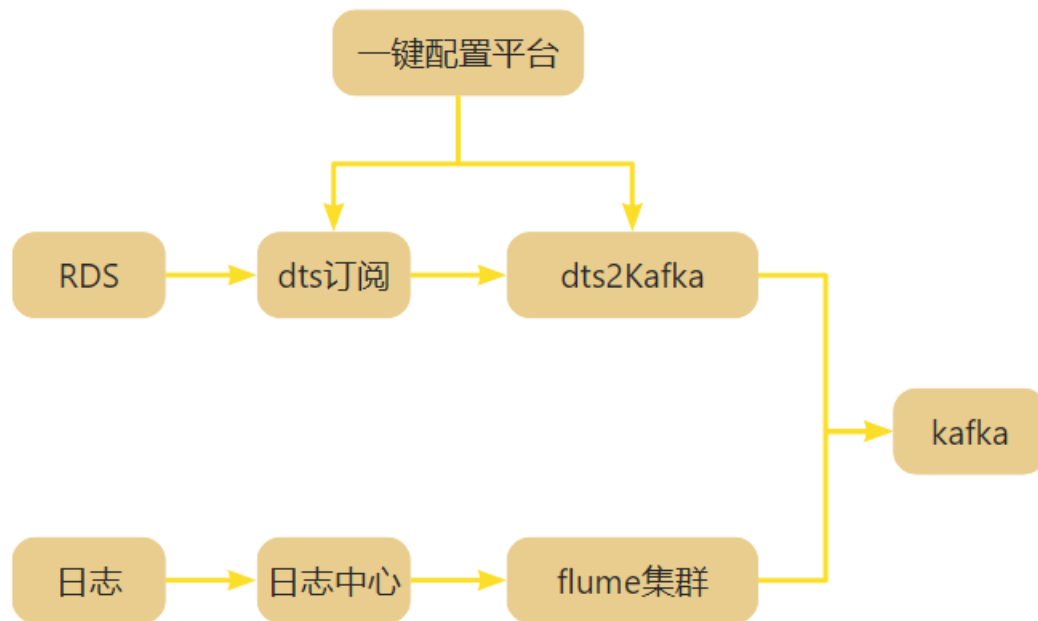
DWA：高度汇总数据，沉淀指标数据。在实时仓库，DWA数据层在OLAP系统。

DIM：维度数据。分为不同类型存于Redis和Mysql

引擎

对比	Spark	Flink
数据处理	微批	批/流处理
SQL	Spark SQL	Table&SQL API
容错	WAL算法	Chandy-Lamport算法
成本	开发成本高	开发成本低
优化	手动	自动
语言	Scala, Java, Python	Java, Scala, Python, SQL
窗口	定时	时间，计数
延迟	秒	亚秒

ODS - 格式预处理



RDS: 主要为Binlog订阅，添加中间处理，统一数据格式

日志: 各个端上的日志传输，统一规范，日志中心格式处理

数据一键集成

配置

DTS标识

输入DTS标识

* 库名

输入库名

* 表名

输入表名

* Topic

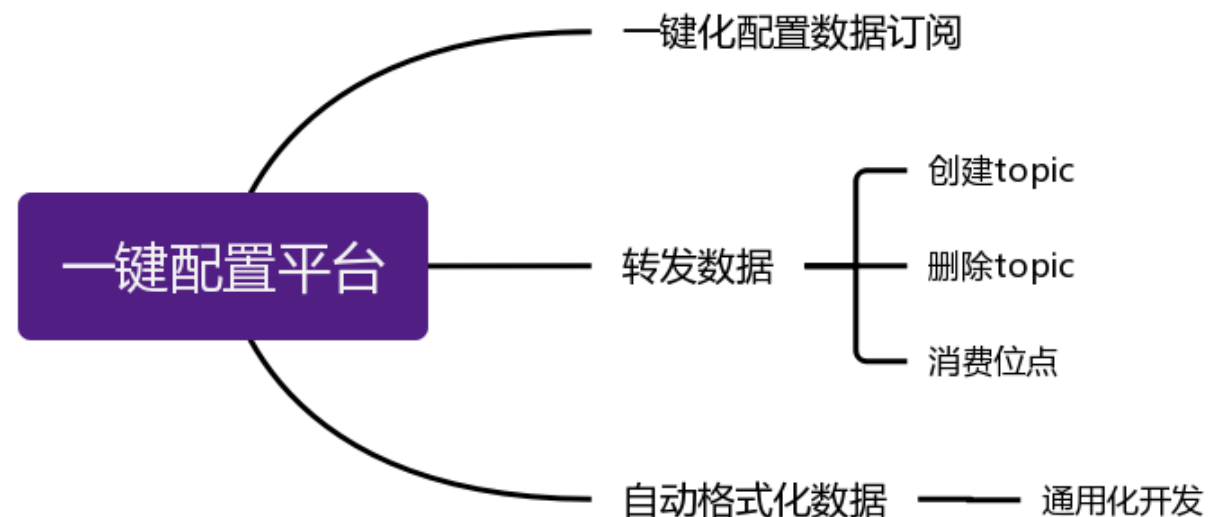
请选择Topic

备注

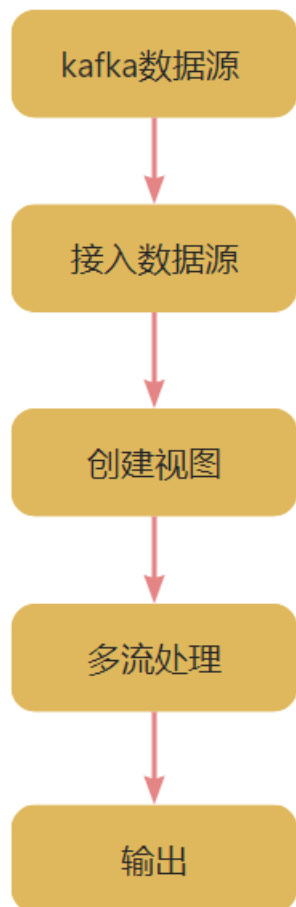
输入备注

取消

确定



开发模板



- ① Flink SQL读取kafka数据源格式固定，可变的是topic参数和读取位点，group等；
- ② 创建视图，利用核心UDF统一离线和实时Schema信息，任务启动阶段进行校验两方的shcema信息（类型，名称等），严格一致
- ③ 多流处理阶段，一般无法定制模板，需要注意state，资源等
- ④ 输出阶段，分为输出至OLAP， Mysql， Kafka。输出至Kafka利用核心UDF固定格式。

数据流入和流出阶段，进行严格的格式控制，利用通用模板提高效率，同时保持离线实时一致。

日志处理

参数化配置

仅需传入离线日志表，任务自动获取离线任务所有信息，自动配置到实时任务

01

自动创建

自动创建topic，初步清洗好的日志数据自动传入topic，并且优化格式

02

资源优化

内部核心为任务清洗程序，配置后台根据任务资源，日志数据切分任务

03



主存储系统的演进

Hbase + ES

01

Hbase存储数据 + ES构建加速查询索引

ADB

02

云原生数据仓库，同时也是即席分析平台
支持存算分离，动态扩展，高并发等

Hologres

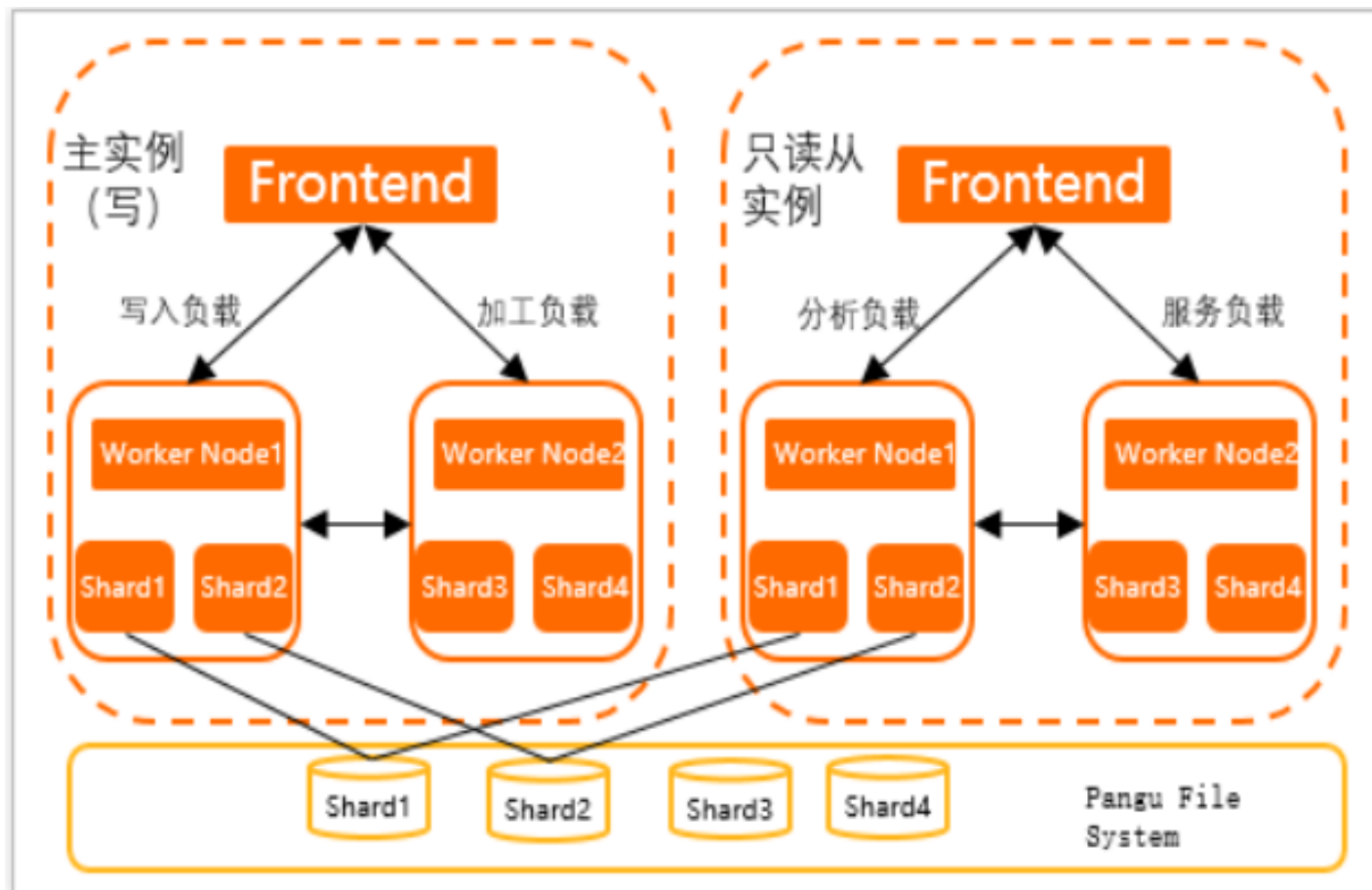
03

当前主要使用的HOLAP系统。

PB级别，支持高并发

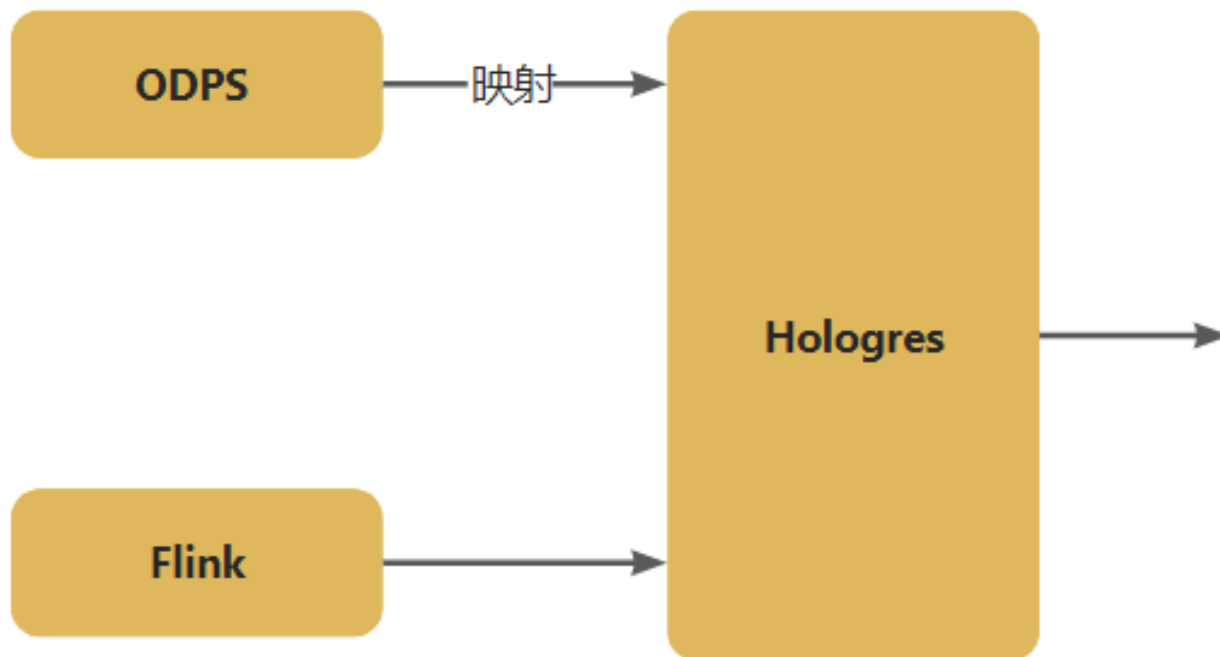
Ad Hoc

Hologres



- 1 共享存储
- 2 故障隔离
- 3 读写分离
- 4 权限可控

Hologres - 实时 + 离线



1

联邦查询，实时和离线数据混合使用

2

统一数据出口，无论是即席分析还是实时接口查询等，数据出口均在 Hologres

4.应用

应用接口



Http接口

灵活性高，可拓展性强



表映射形式

解耦，无感知变更



接口监测

响应时长，ip，查询频率等进行
资源监控



平台一站式开发

内部研发的接口管理平台，上线
接口从测试到上线达到分钟级别



慢查询监控

慢查询及时监控预警

接口开发平台

添加配置 +

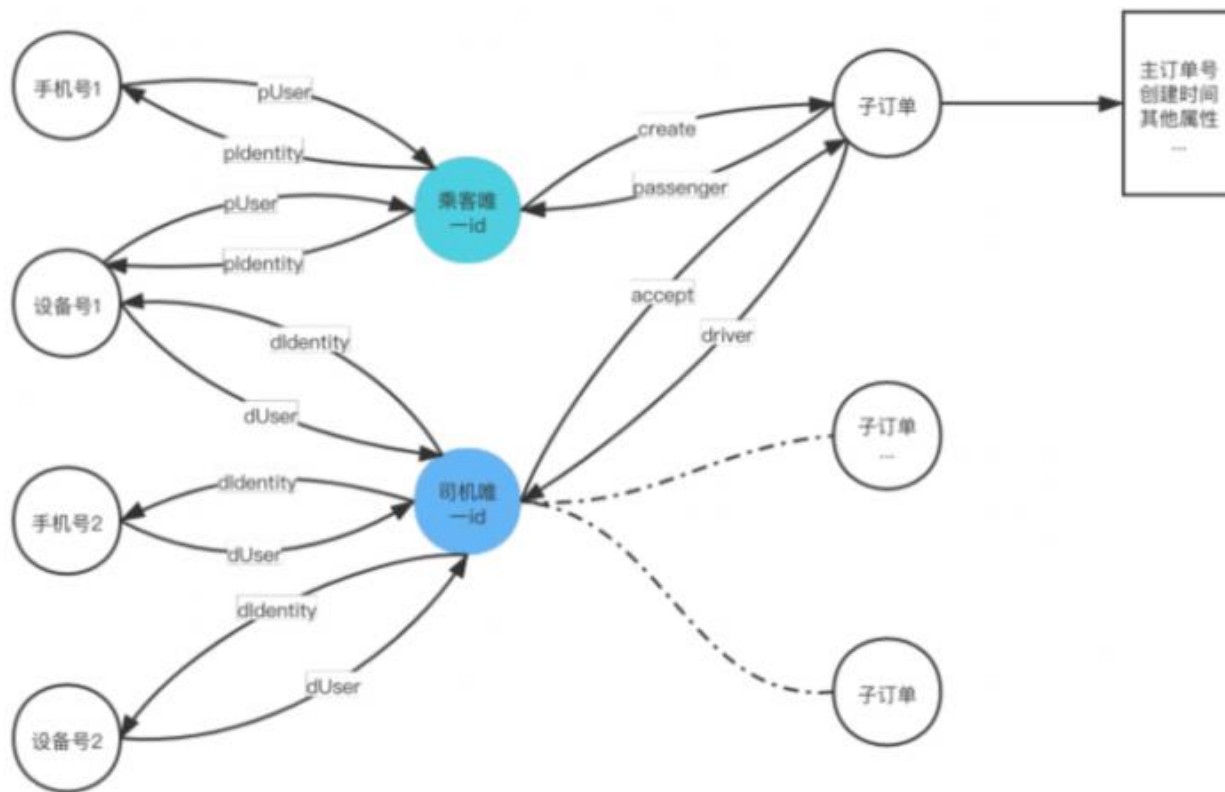
queryId	SQL	SQL参数	业务场景	创建人	更新时间	操作
b1b3751431708c54858	SELECT ...ci...id as city ,...		实时...司机...数据	fanc...	2022-11-22 18:52:41	 
d9f150715b19ddC	select ci...id , ..._type_ic...a...		实时...订单...数据	fanc...	2022-11-22 14:07:23	 
fc4148335b1c7656	SELEC ...lgo...y_id,ca...ype...		实时...司机	yan...	2022-09-29 18:13:26	 
711b27159175aff1a4...	SELEC ...city... ,car_t..._id ,...		实时...订单	ya...	2022-09-29 18:12:29	 

接口配置为SQL开发，测试之后自动生成接口id，分钟级别上线。

目前接口规模300+，平均查询时长在毫秒级别。

风控应用





构建出1个实体表和8个关系表

指标预警

添加预警任务

×

* 名称

请输入预警任务名称

启停状态



* 描述

请输入预警任务描述

* 规则类型

☐ SQL

☐ 指标

* 颗粒度

☐ 小时级

☐ 天级

* 数据逻辑

请输入SQL

* 预警方式

自定义预警

智能预警

预警规则

运算符



阈值

通知方式

☐ 钉钉

☐ 邮件

* 接收人邮箱

请输入接收人邮箱，多个邮箱以英文逗号分隔

确定

取消

自定义规则定制实时预警和离线预警方式

目前支持SQL和指标形式进行监控

小时级别和天级别的预警

智能预警形式结合算法，自动测算，实时预警

OneData-指标管理

指标名称: 英文名称: 事实表: 指标级别:

业务线: 周期: 主题: 分城市:

是否实时:

添加指标

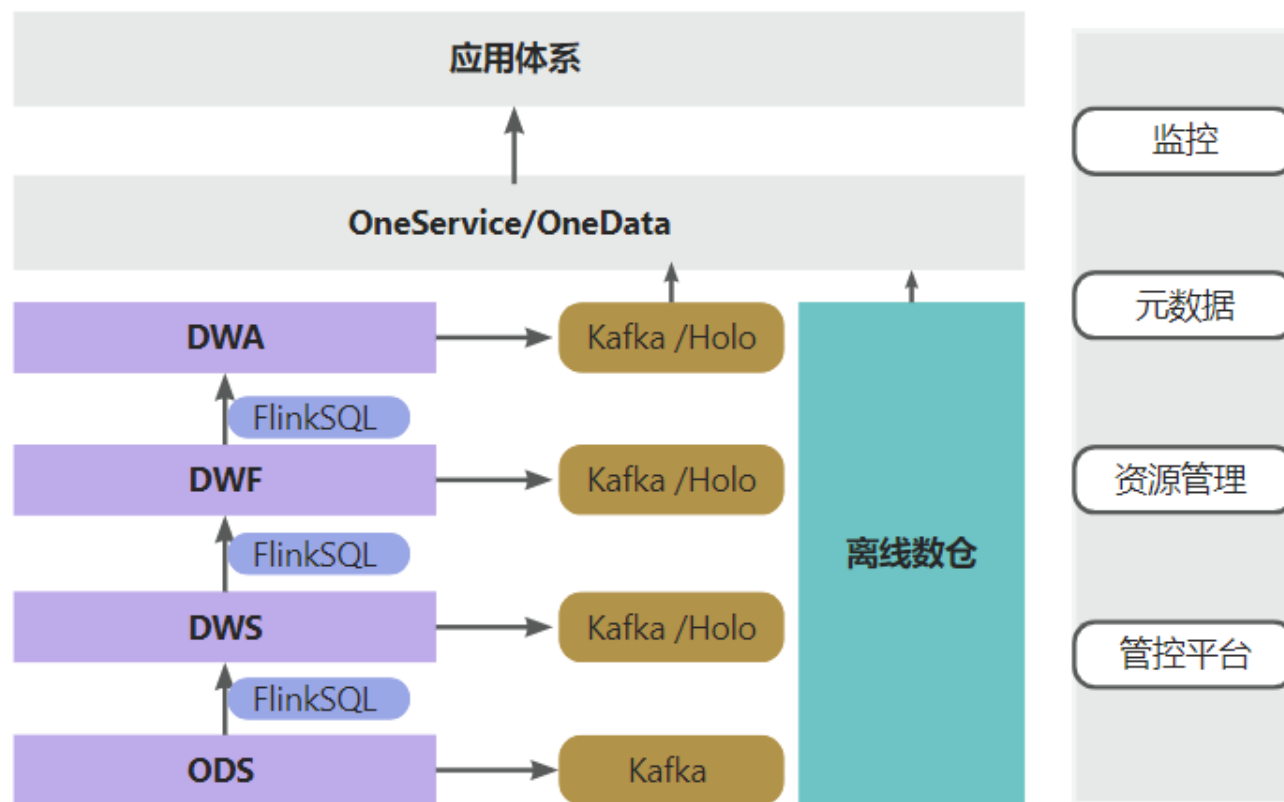
	ID	指标名称	英文名称	是否实时	指标级别	业务线	定义	指标口径	操作
1	207	下单量_实时	order_cnt_r	实时	核心	货的		count(distint order_id)	维度 血缘 版本 编辑 删除
2	208	需求单量_实时	need_order_r	实时	核心	货的		sum(need_order)	维度 血缘 版本 编辑 删除

实时指标和离线指标统一管理平台；

指标血缘，版本等各种管理功能。

无论实时或者离线指标，首先是指标管理平台收录，再进行三方嵌入使用

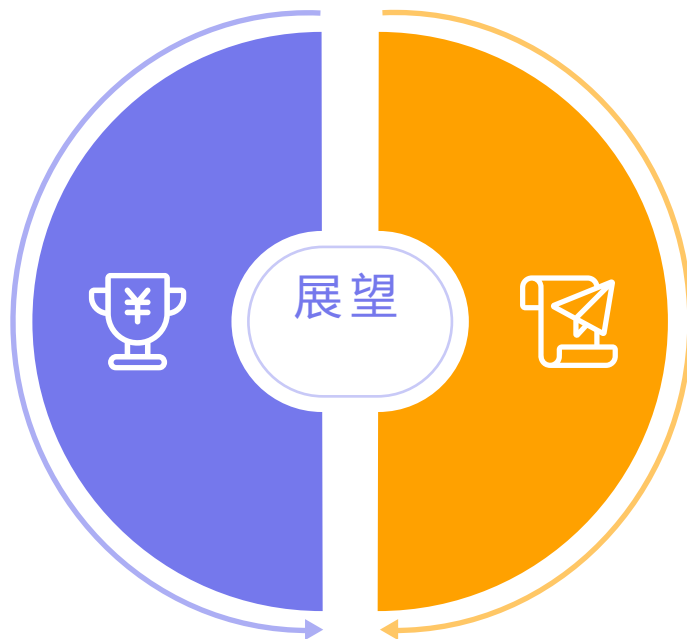
整体架构





未来

流批一体
一套系统
一个逻辑



动态规则
智能营销

THANKS

SQL Server
vertica
D B 2
G B a s e
O r a c l e
达梦数据库
神舟通用
KingbaseES

2010

2014

2018

openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
MatrixDB
DynamoDB
SinoDB
DolphinDB
Galaxybase
KunDB
FastData
GaussDB
PolarDB
GDB
Spacture
Sequoiadb
OushuDB
ArgoDB
开务数据库
GreatDB
MongoDB
TDSQL
TiDB
Tapdata
UbiSQL
StarRocks