DataFun.

流批一体的实时多维分析

郑德来　　百度资深研发工程师

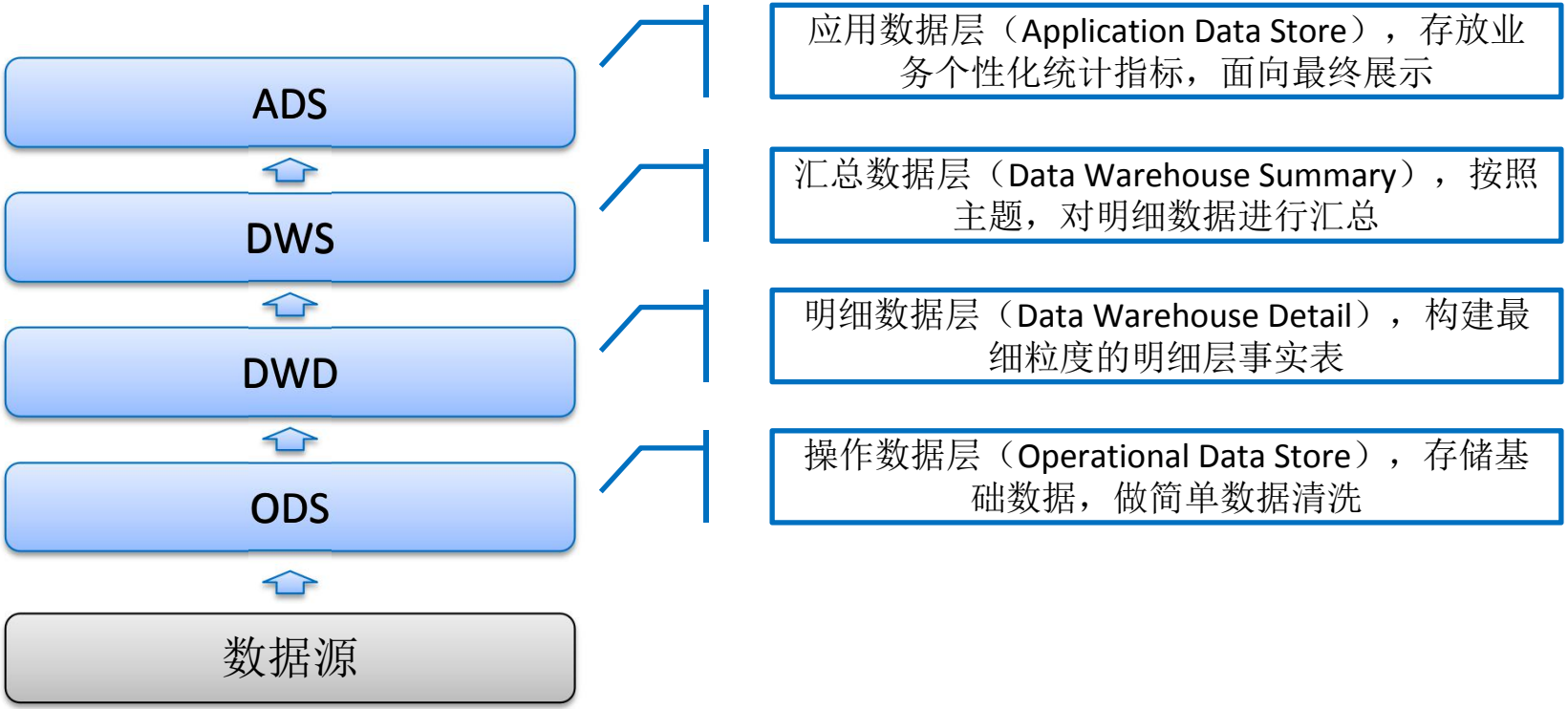
目录 CONTENT

DataFun.

01

大数据架构演进

经典离线数仓架构介绍



经典离线数仓架构优缺点分析

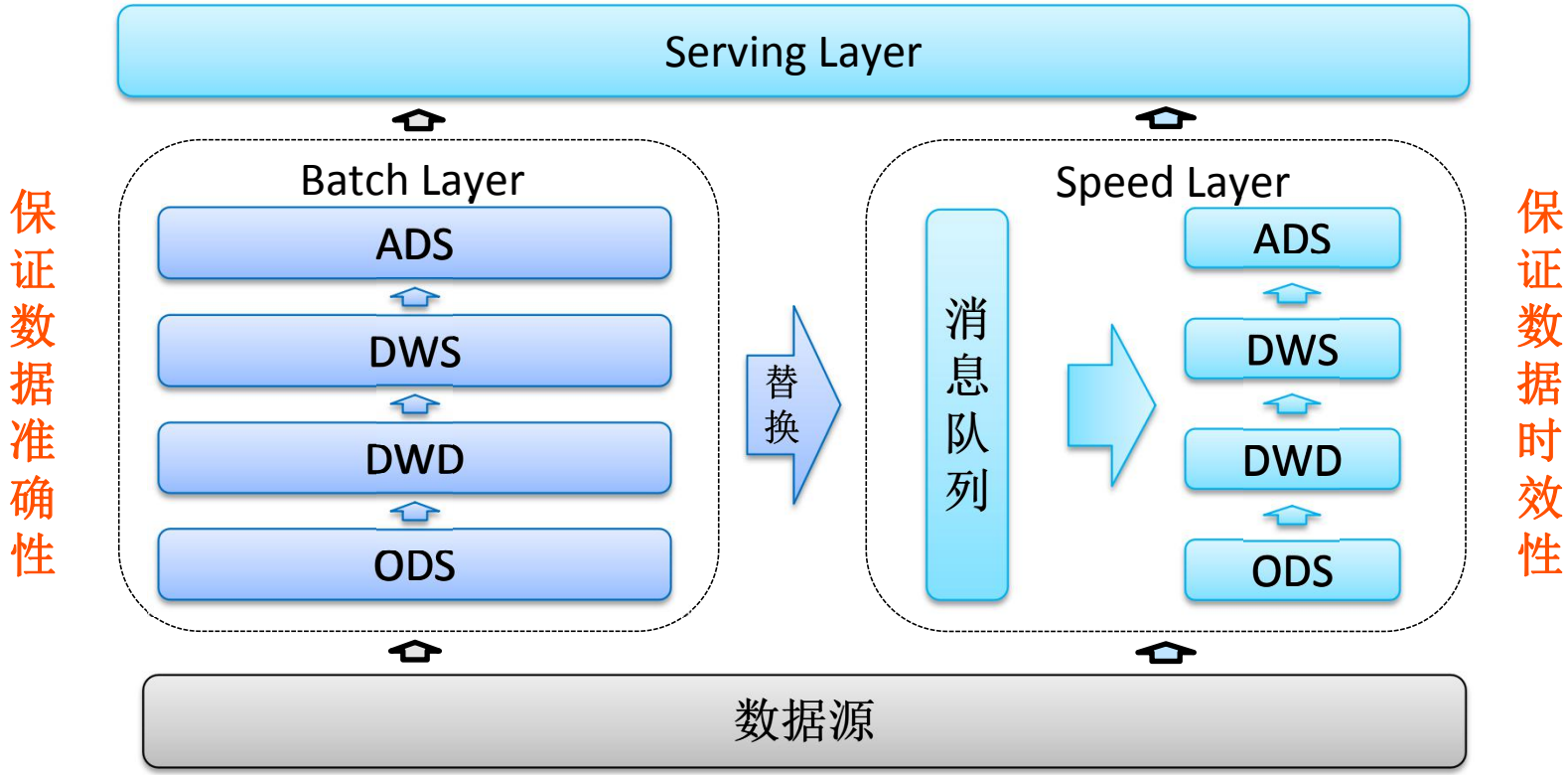
优点

- 架构简单，开发成本低
- 资源成本低
- 数据易管理，diff少

缺点

- 数据时效性差
- 缺少实时数据
- 表数量太多

Lambda架构介绍



Lambda架构优缺点分析

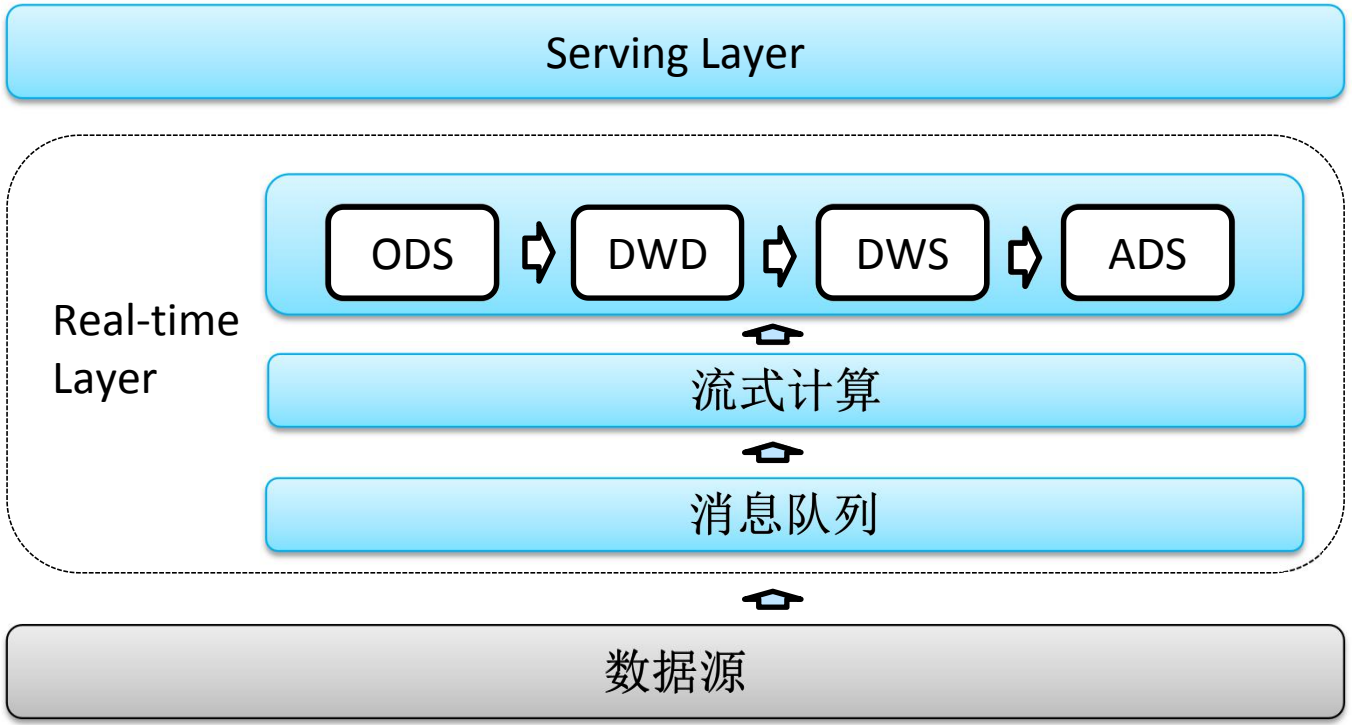
优点

- 引入实时数据
- 兼顾准确性和时效性
- 兼容经典离线数仓体系

缺点

- 一个需求两套代码
- 资源占用多
- 实时数据和离线数据diff

Kappa架构介绍



Kappa架构优缺点分析

优点

- 一套数据流，开发成本低
- 省掉离线数据流计算资源
- 实时离线数据逻辑统一

缺点

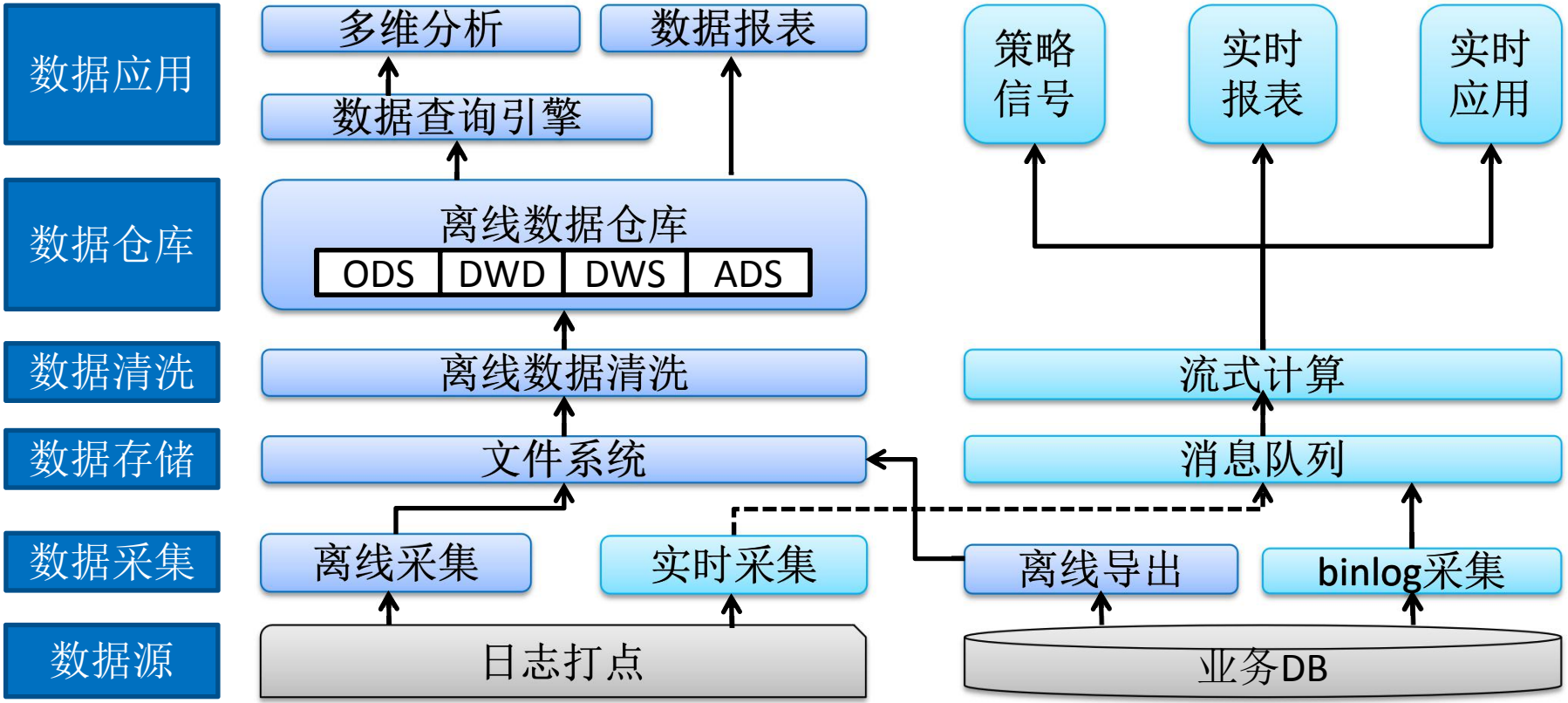
- 数据回溯成本高
- 复杂关联场景开发维护成本高
- 历史包袱的迁移成本高

DataFun.

02

流批一体方案

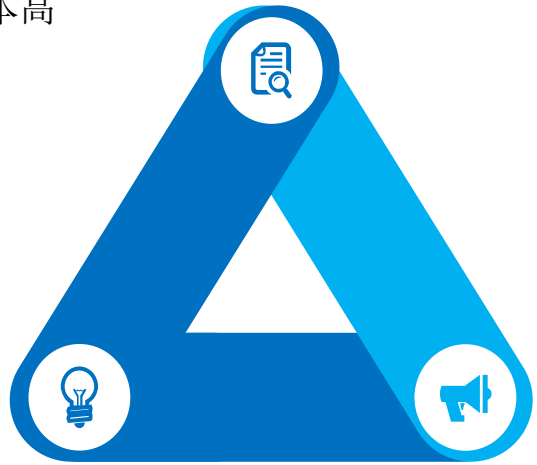
流批一体背景-旧架构



流批一体背景-旧架构问题

表太多

数仓分层建设，表数量太多，使用成本高



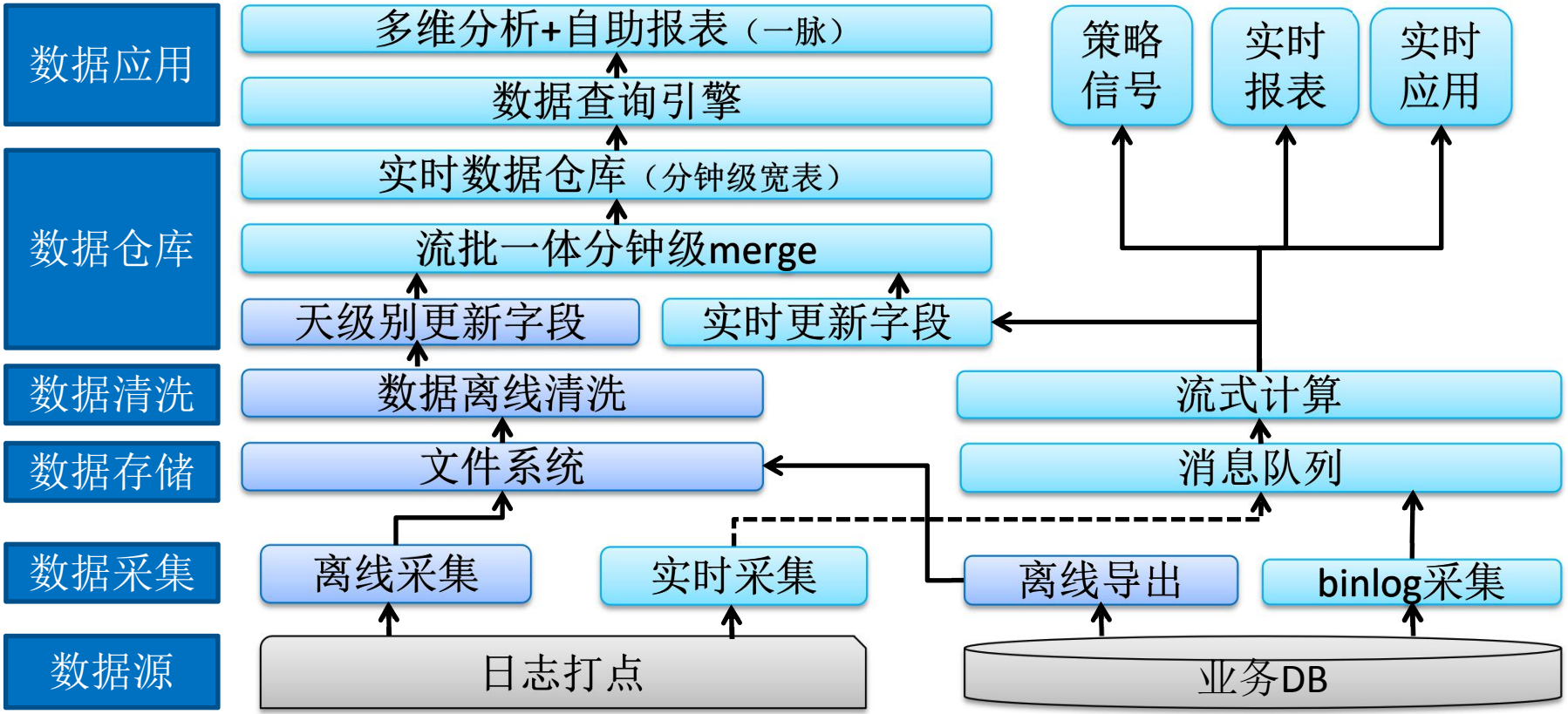
实时分析弱

实时报表太定制化，缺少多维分析能力

查询慢

表关联场景众多，查询时效慢

流批一体整体方案

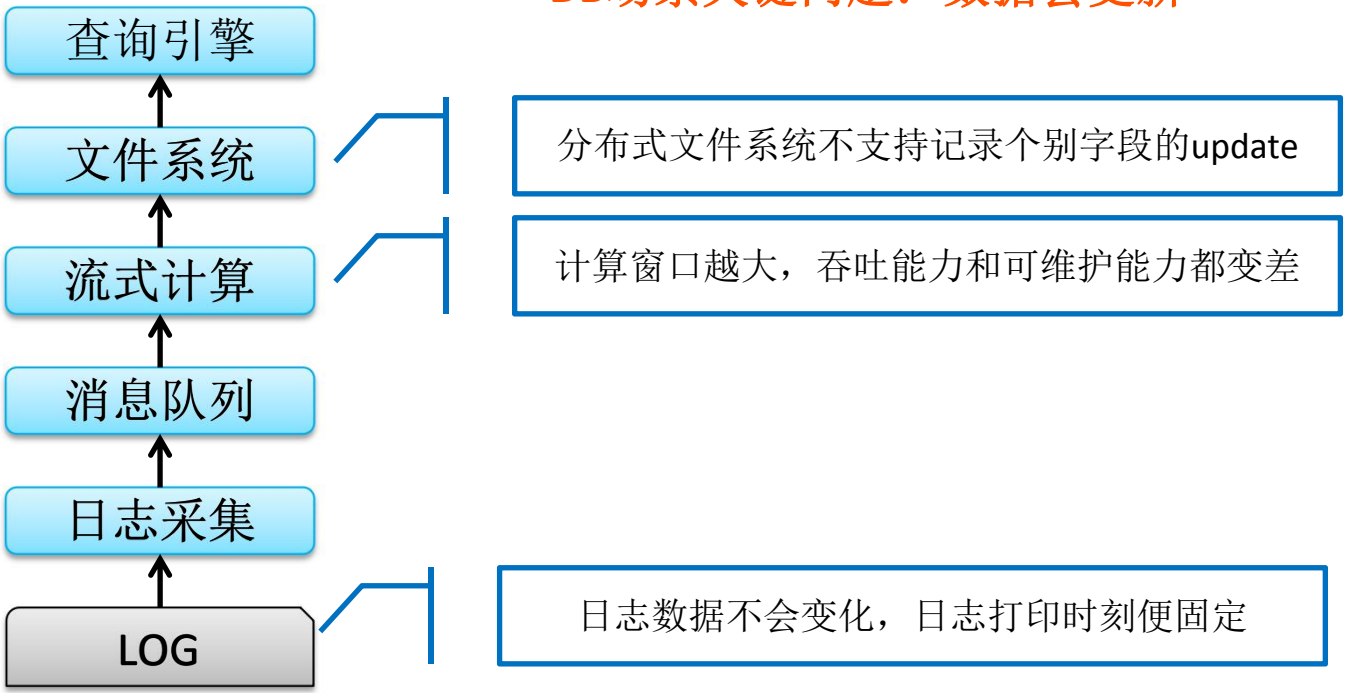


DataFun.

03

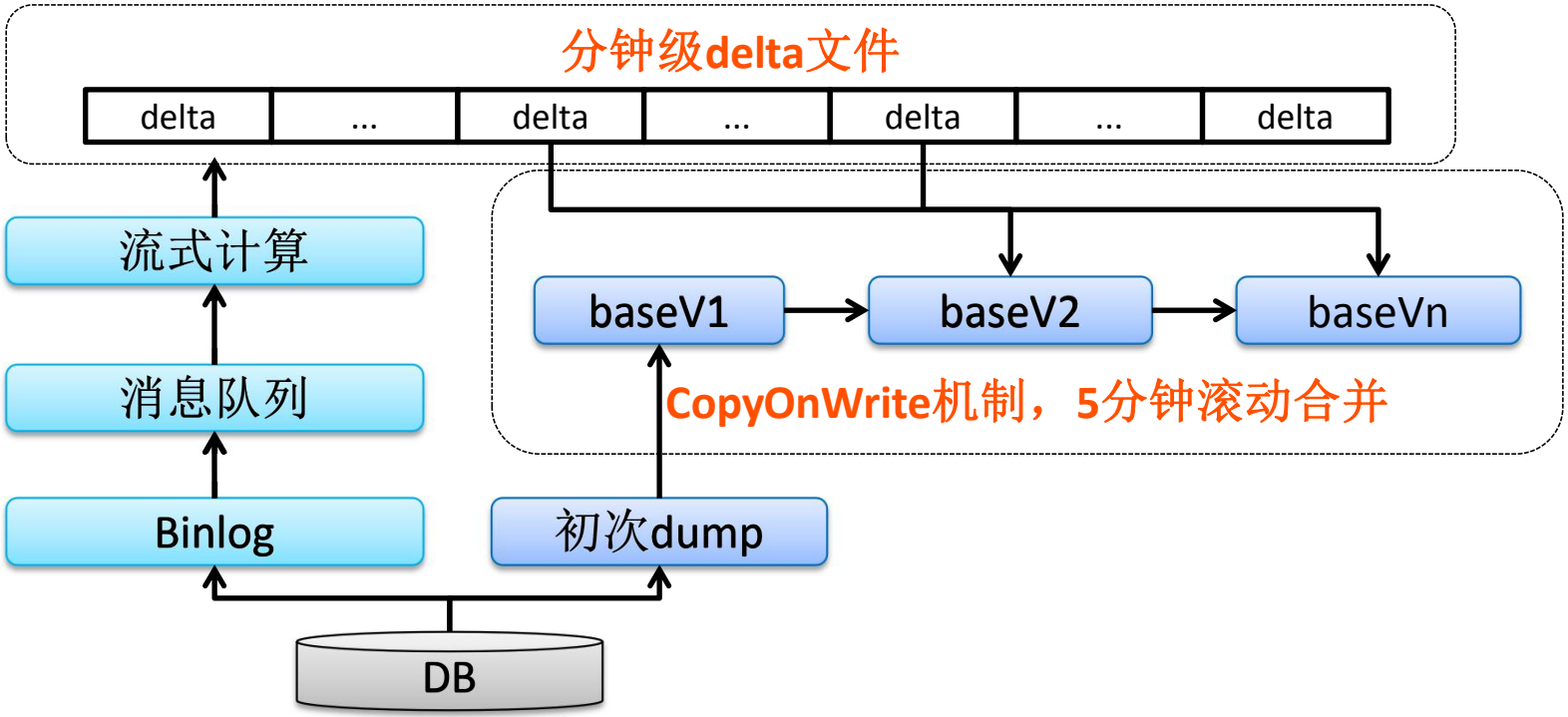
关键问题突破

1、DB数据更新问题-背景



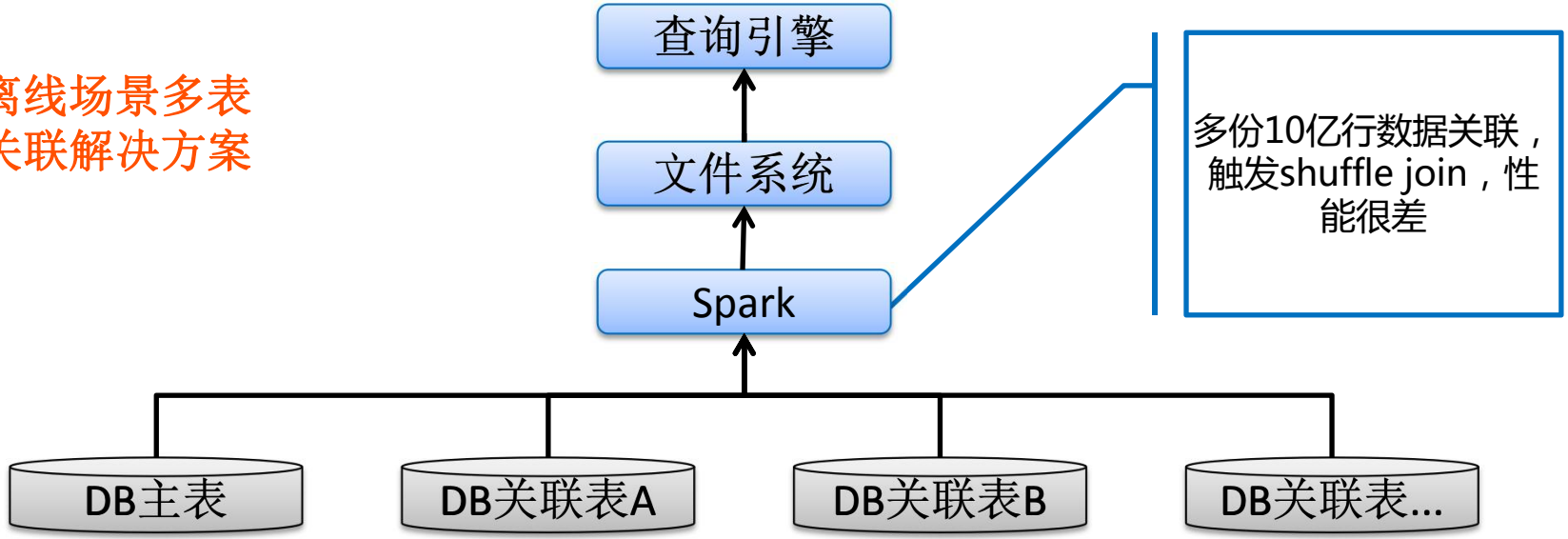
日志场景实时数仓的典型方案

1、DB数据更新问题-解决方案

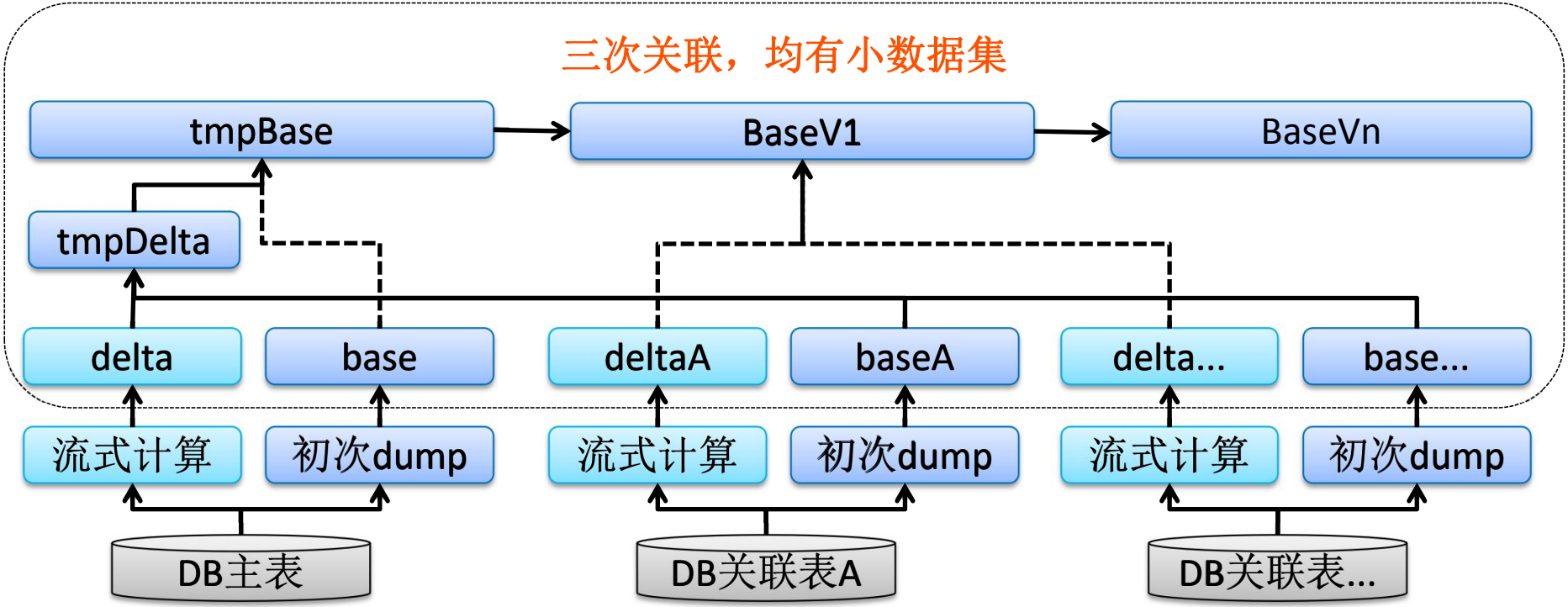


2、多表关联问题-背景

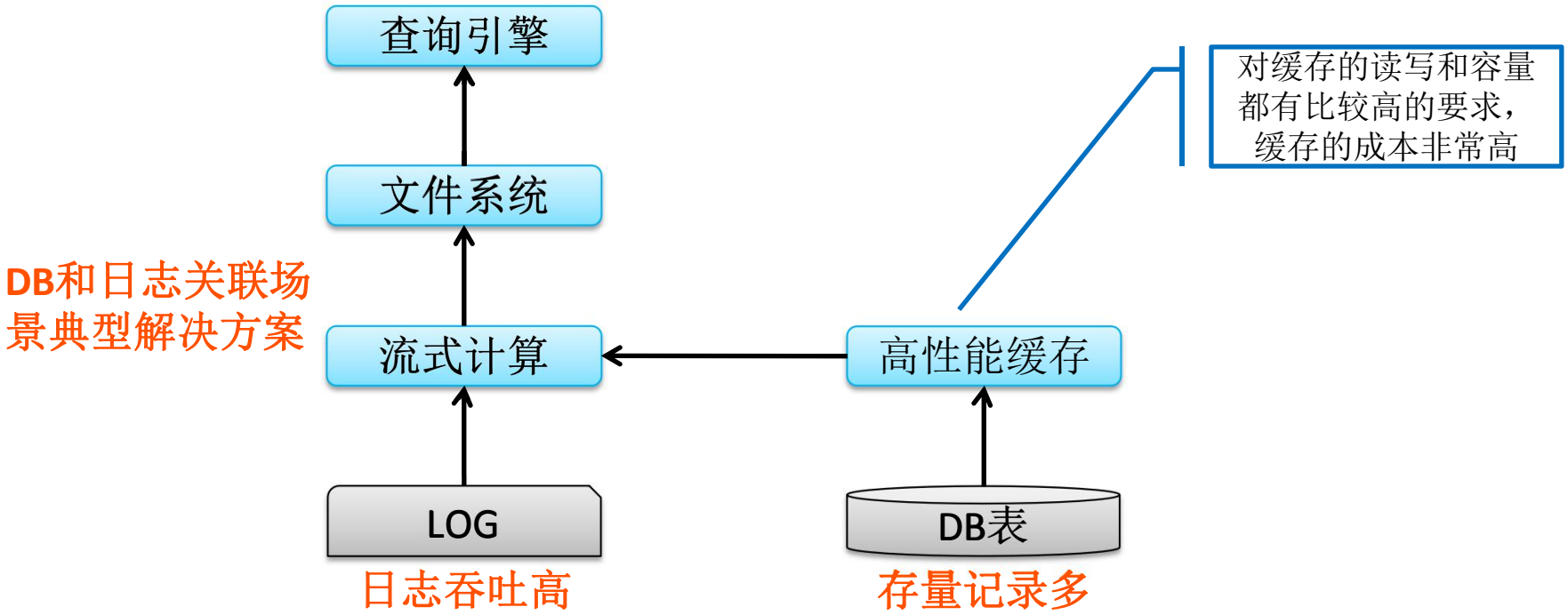
离线场景多表关联解决方案



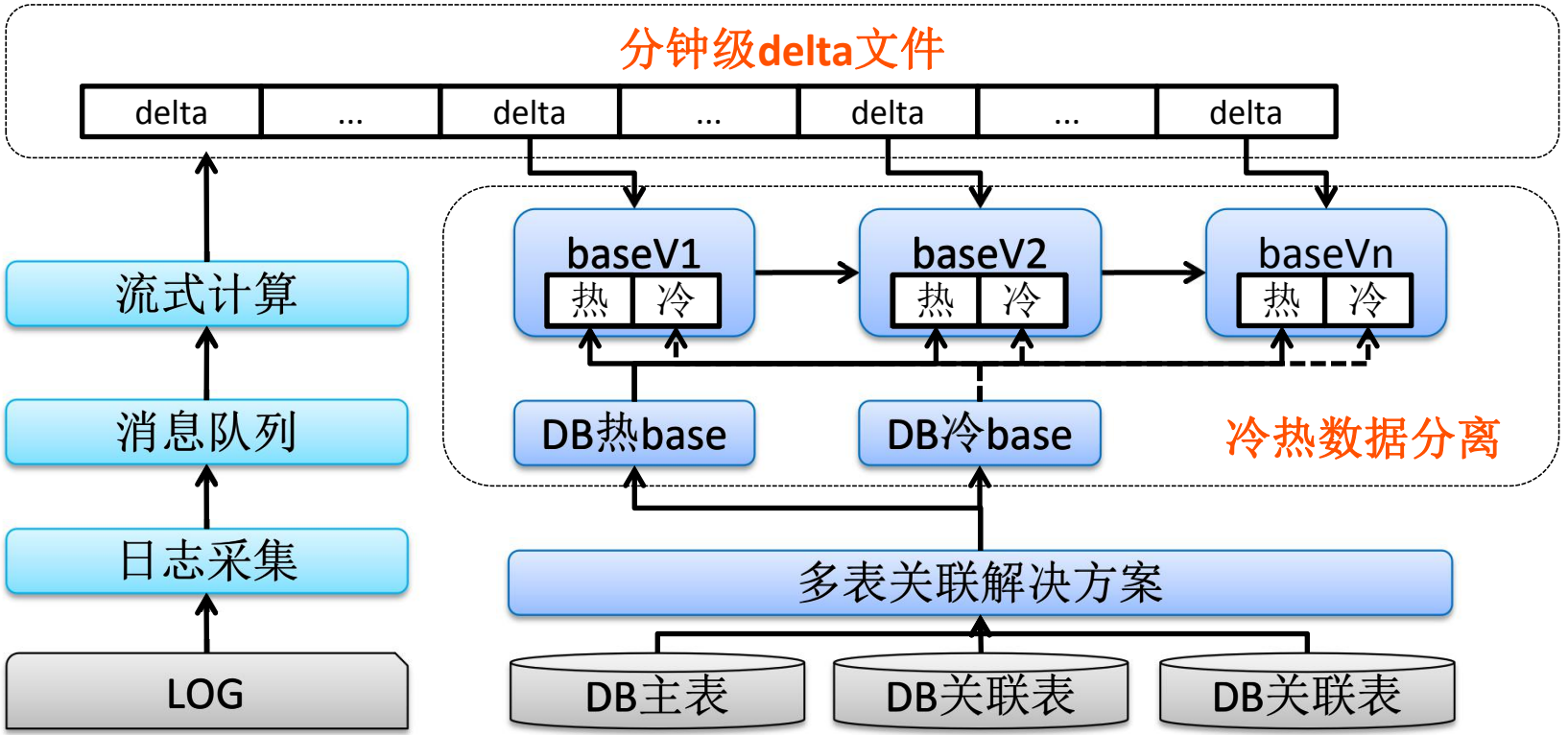
2、多表关联问题-解决方案



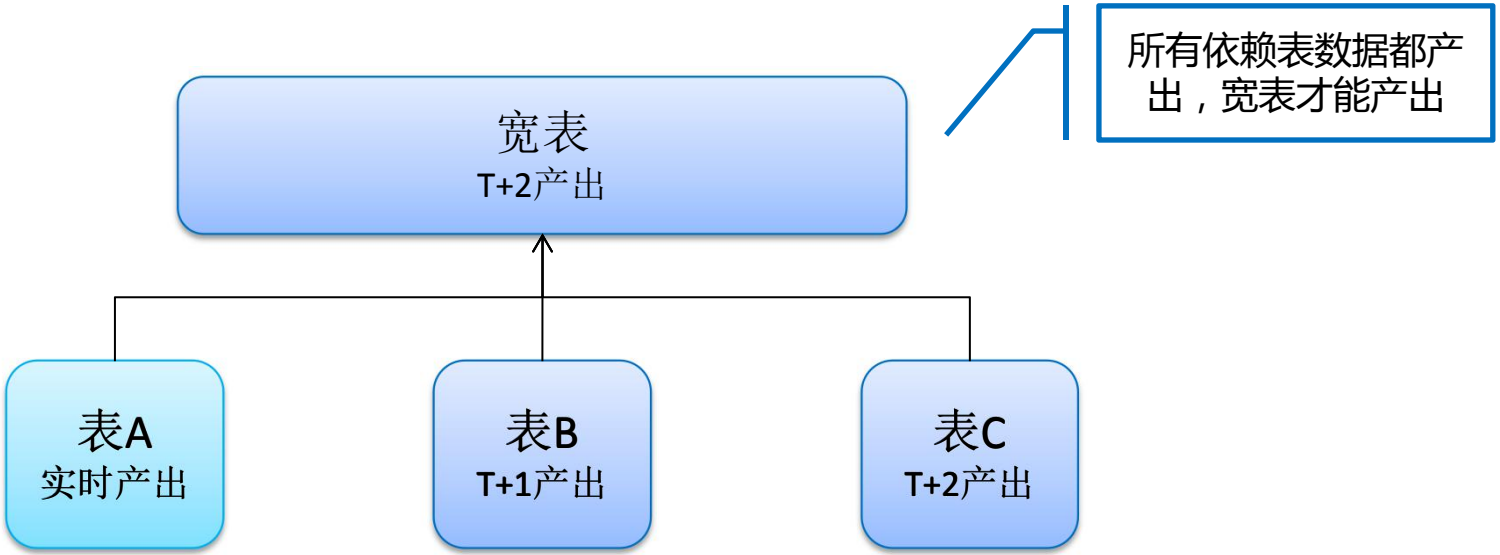
3、DB和日志关联问题-背景



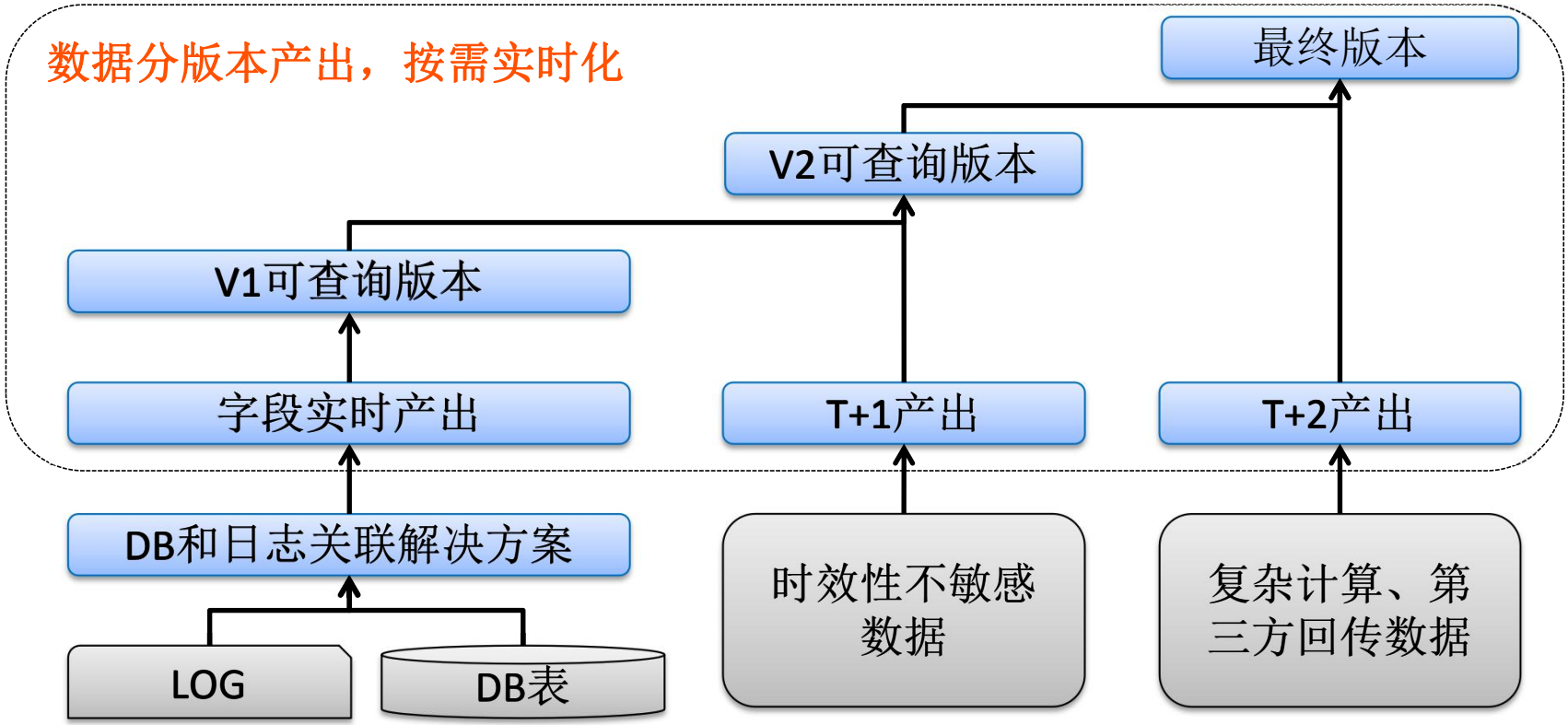
3、DB和日志关联问题-解决方案



4、数据到位时间问题-背景



4、数据到位时间问题-解决方案



DataFun.

04

总结和规划

总结和规划

总结：

（1）架构选型要符合业务现状，解决业务实际问题
（2）架构选型要综合考量资源、复杂度、维护成本

规划：

（1）引擎查询性能持续提升
（2）上层查询工具体验优化

非常感谢您的观看

DataFun.