

爱奇艺数据治理中的 数据湖应用实践

爱奇艺 杜益凡

信息革命极大地提升了社会生产效率，人们节约出了越来越多的时间，这都需要以健康的方式消耗掉。**娱乐是健康地消耗掉剩余时间最主要的方式之一。**

爱奇艺是一家科技公司，也是一家娱乐公司，是科技创新驱动的一家娱乐公司，用技术创新降低娱乐成本，让用户更便捷地获得更多快乐，这是爱奇艺与传统娱乐公司最大的不同。

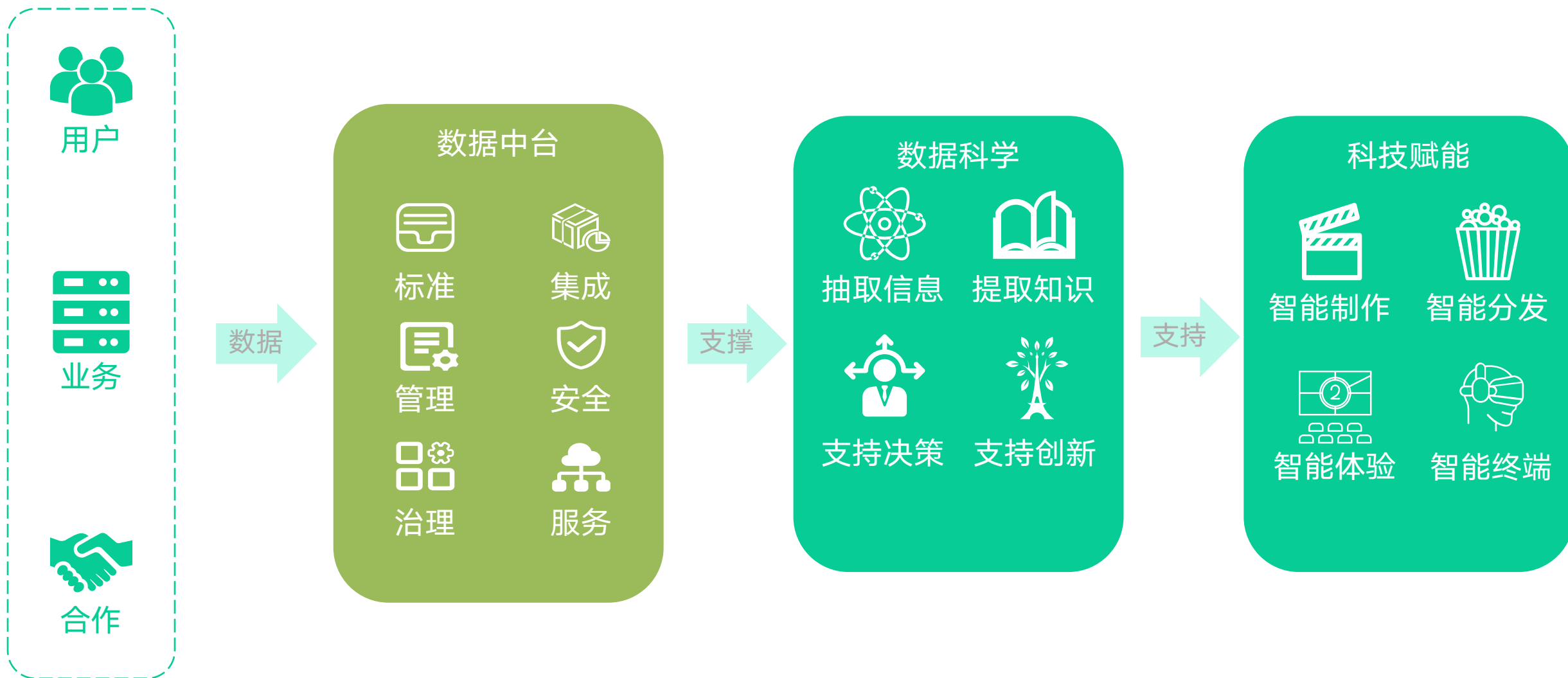
伟大有两方面内涵：**一是商业层面的**，是指获得卓越的商业发展，做到用户多、收入多、规模大、服务和影响范围大；**二是精神层面的**，是指我们要以内容价值观影响千千万万的人，特别是年轻人，帮助他们树立健康的价值观，让他们的成长过程变得更快乐，精神更充实，这是更有意义的事情。



做一家以科技创新
为驱动的伟大娱乐公司

iQIYI 爱奇艺 | 企业愿景







数据工作遇到的问题



接入成本高



使用门槛高



质量低



可靠性低



跨业务难度大



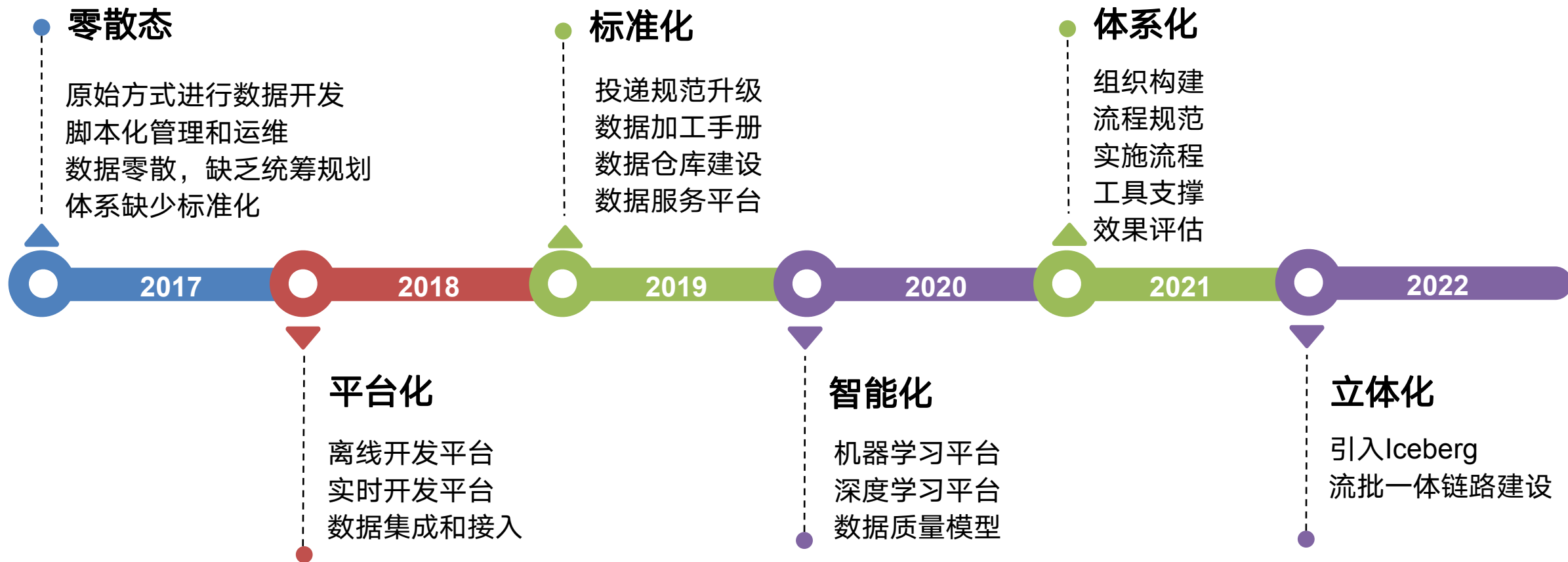
口径不一致



数据资产模糊



资源成本高



数据沼泽

数据池

数据湖

数据湖是一种大数据技术实现

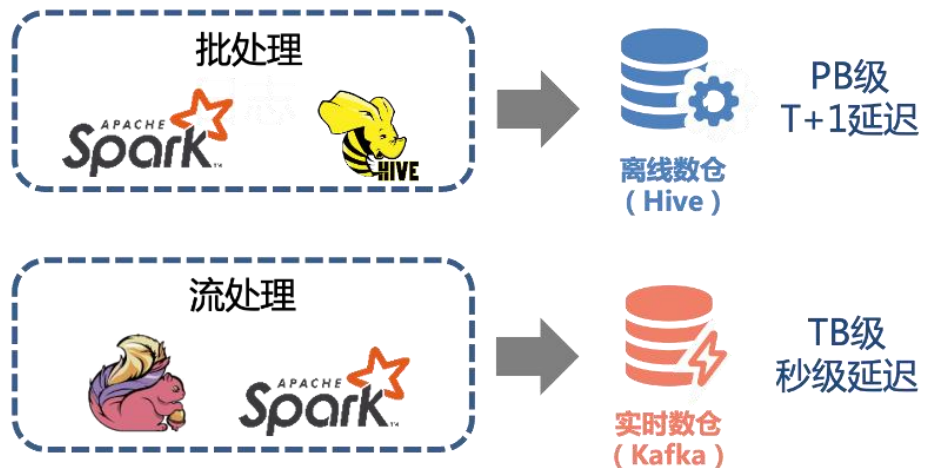
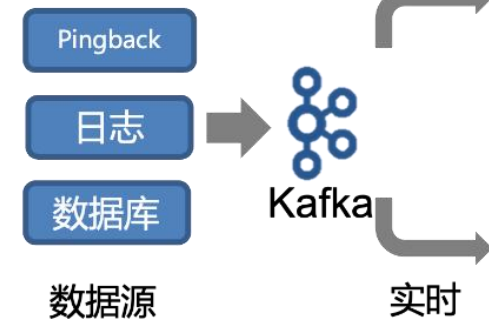
数据存储

ICEBERG

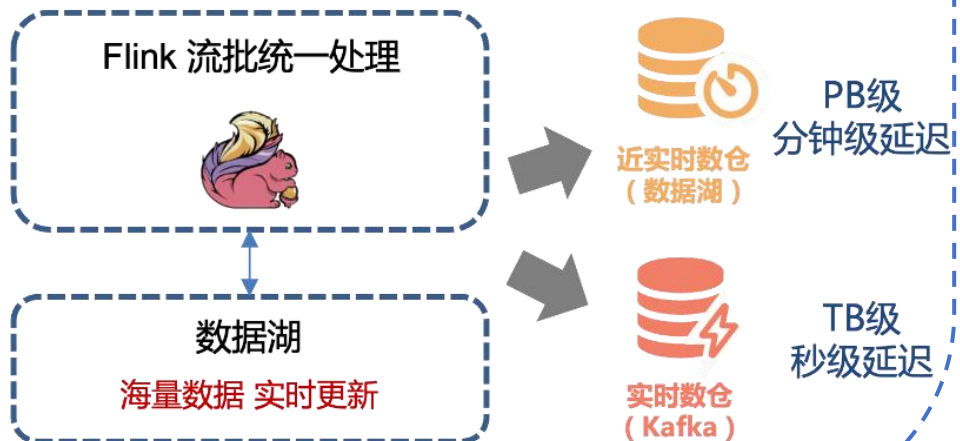
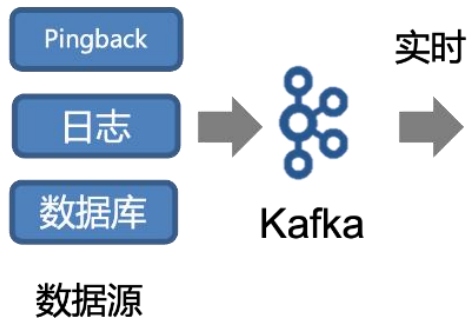


数据处理

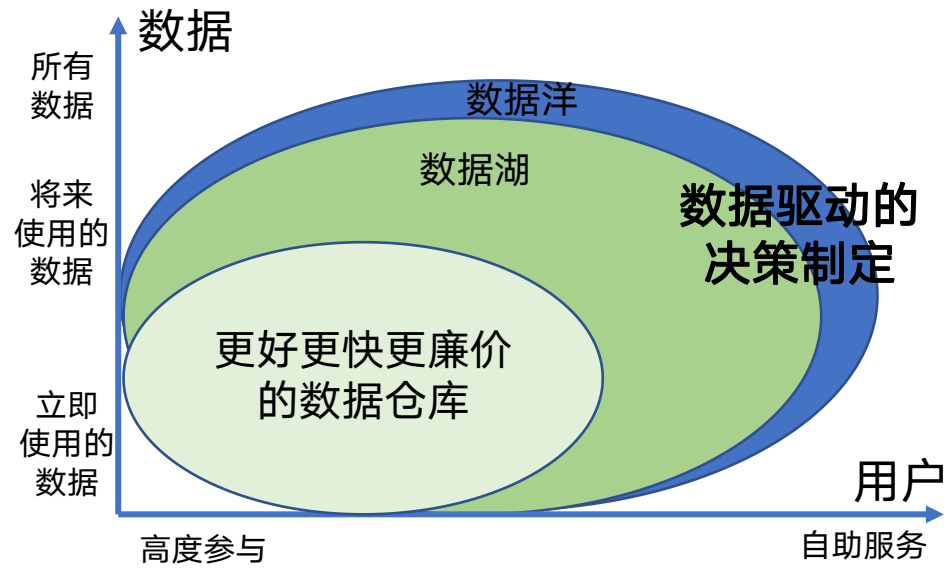
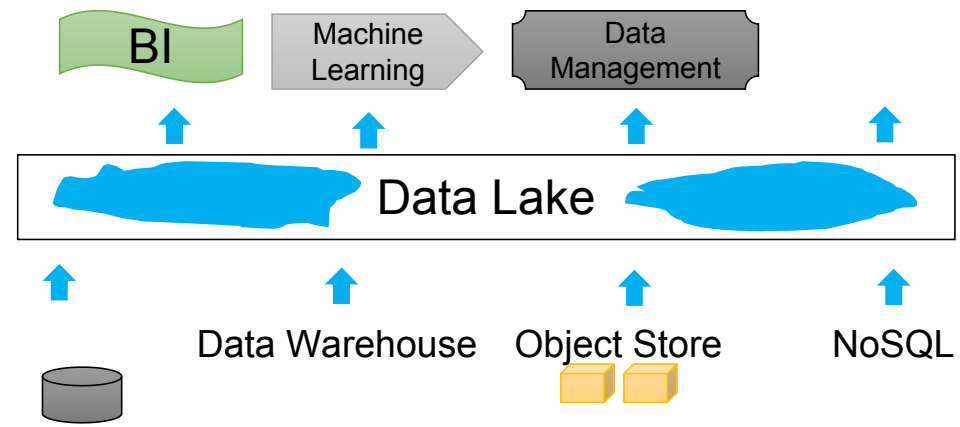
Lambda 架构



流批一体架构



数据湖也是一个大数据治理理念



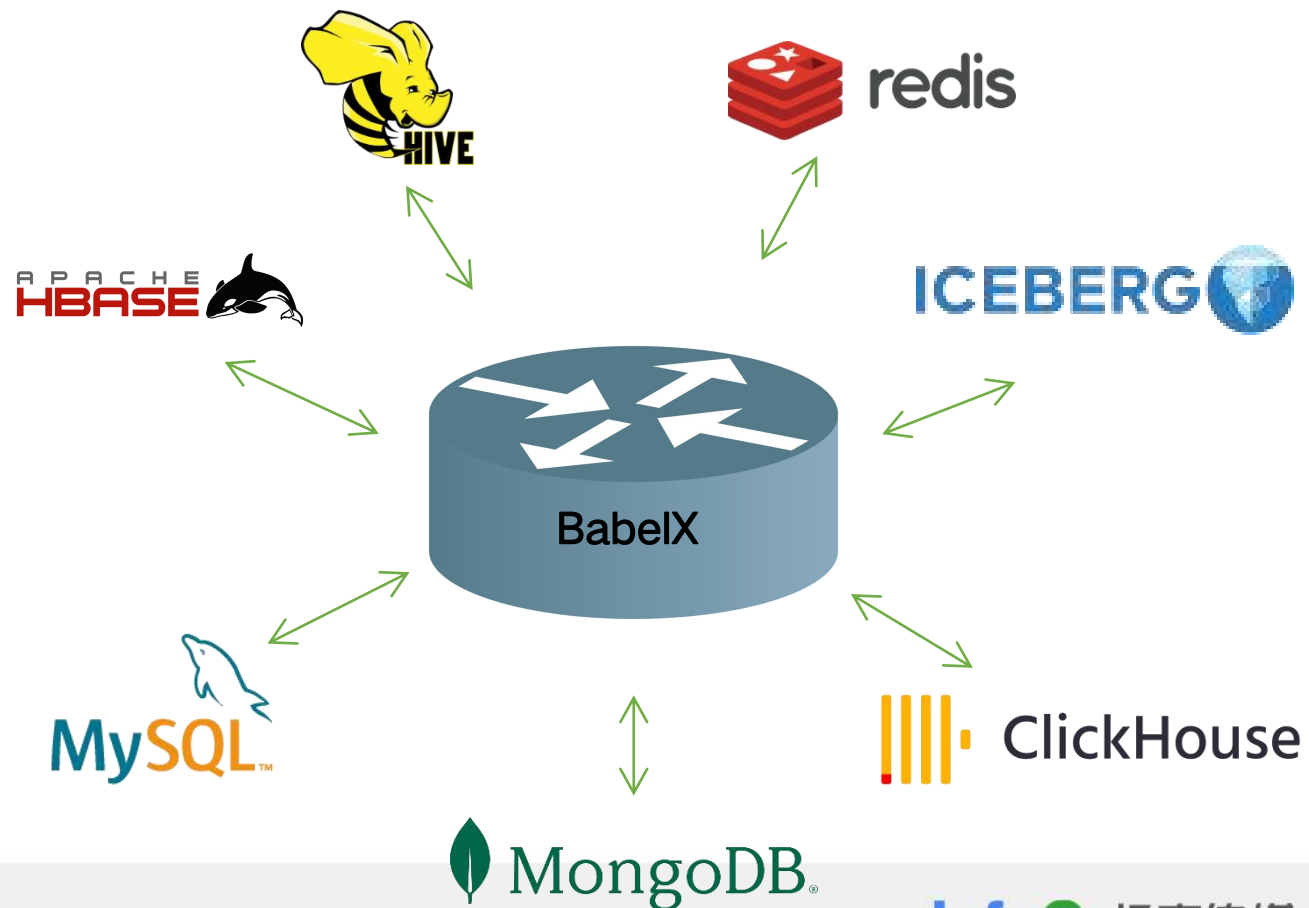
数据湖的价值

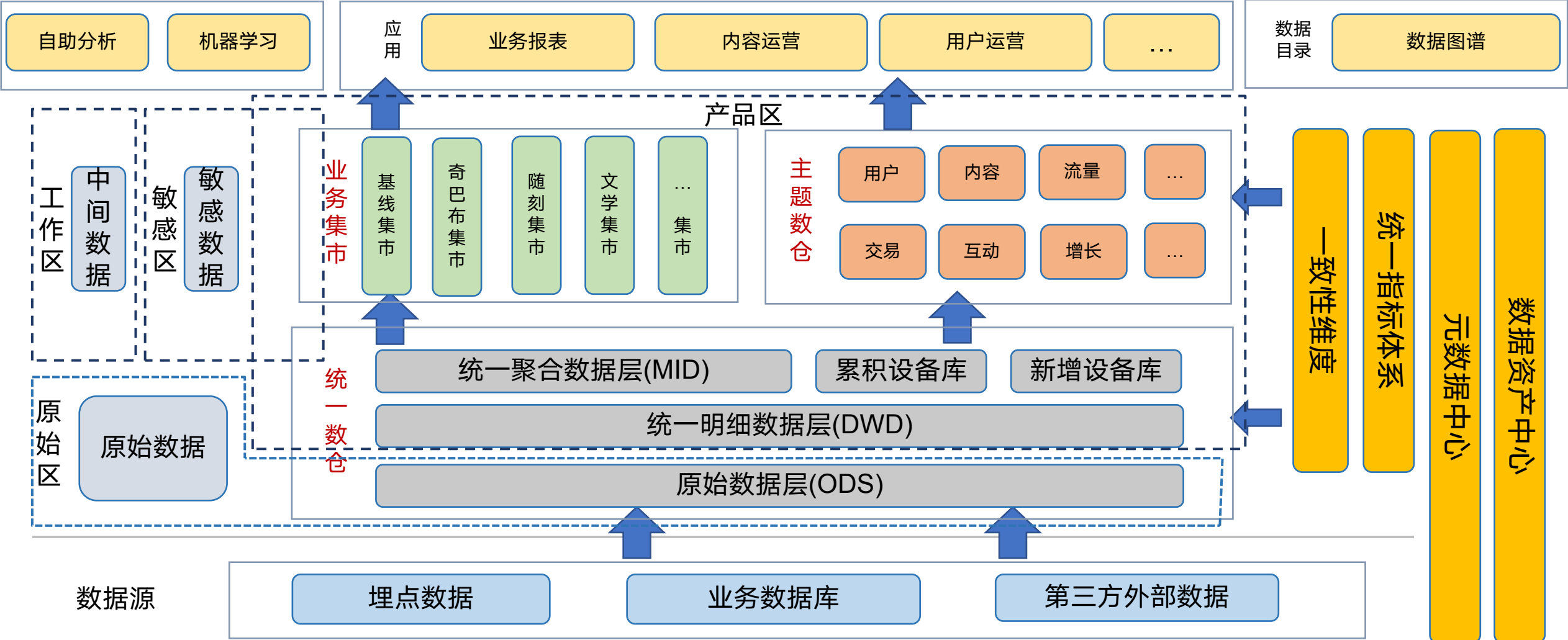
集成数据

- Hive、MySQL、MongoDB 等 15 种数据源之间的数据交换
- 多集群、多云间的数据同步

应用场景

- 数据集成：集成业务数据，打通数据孤岛
- 数据迁移：同步Hive、ClickHouse、MySQL 等数据库数据到不同DC
- 数据备份：全托管式的定时数据同步和备份







Pingback埋点数据


业务数据




合作方数据



ICEBERG 

 **HIVE**

APACHE
HBASE 


alluxio  **HDFS**

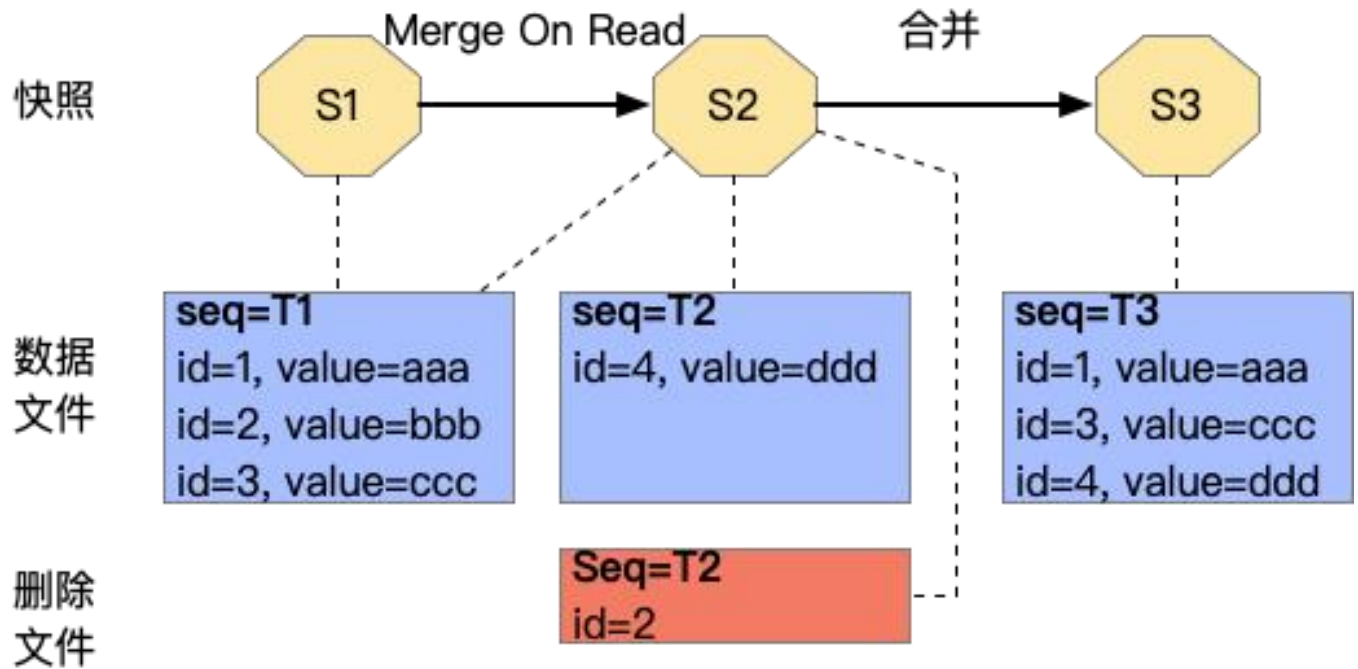
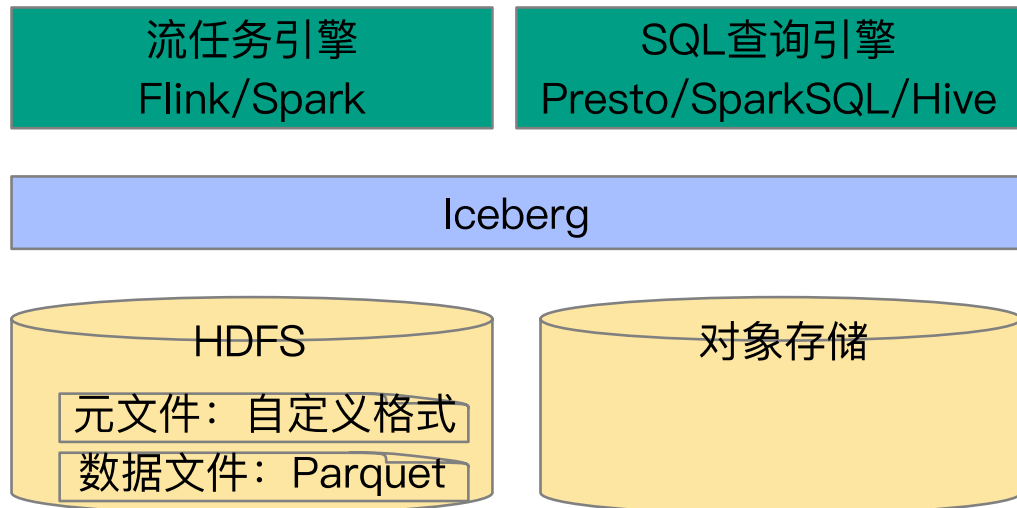
Delta Lake vs Hudi vs Iceberg

维度	Delta Lake	Hudi	Iceberg *
规模/成本	深度绑定 Spark、高级特性仅在商业版支持，不考虑	PB级、低成本	PB级、低成本
行级更新		支持	支持
增量拉取		支持	支持
时效性		近实时（5分钟）	近实时（5分钟）
查询引擎		PrestoDB、SparkSQL	PrestoSQL、SparkSQL
计算引擎		Spark友好、Flink较差	Spark友好、Flink支持好
Flink写性能		遇到瓶颈	写性能好
发展潜力		已有功能更完善	抽象较好，潜力更强

Iceberg是一种数据表格式

Iceberg：一种新设计的开源**表格式**用于大规模数据集分析

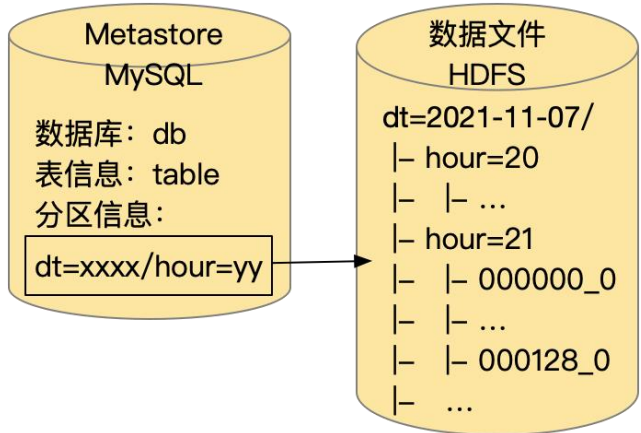
- 不是存储引擎：支持HDFS、对象存储
- 不是文件格式：使用Parquet存储数据
- 不是查询引擎：支持Spark/Flink/Presto/Hive



Iceberg表支持行级更新和ACID

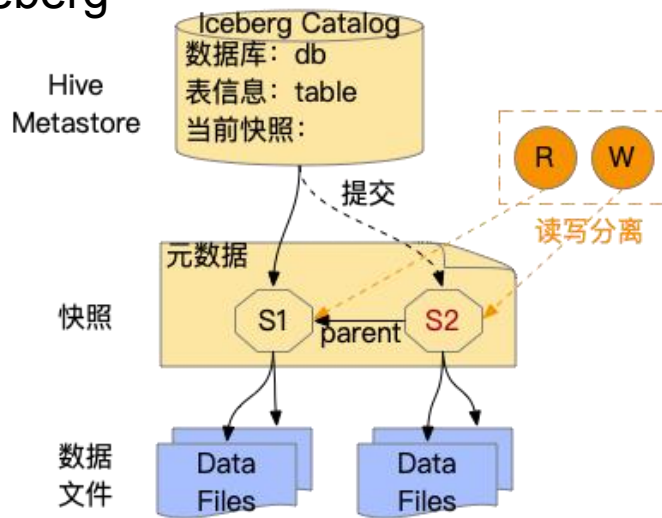
Hive vs Iceberg

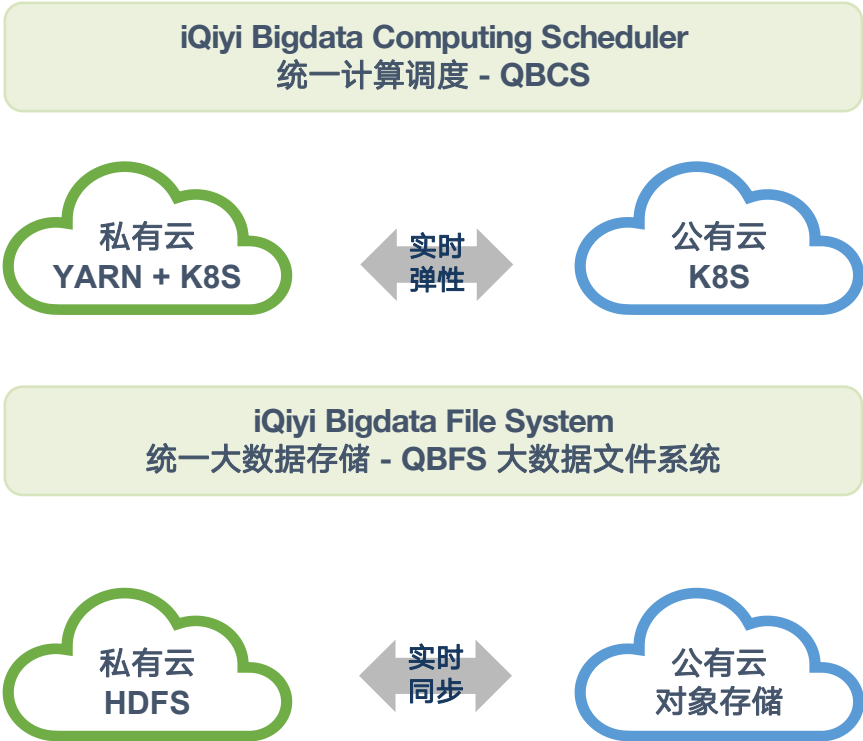
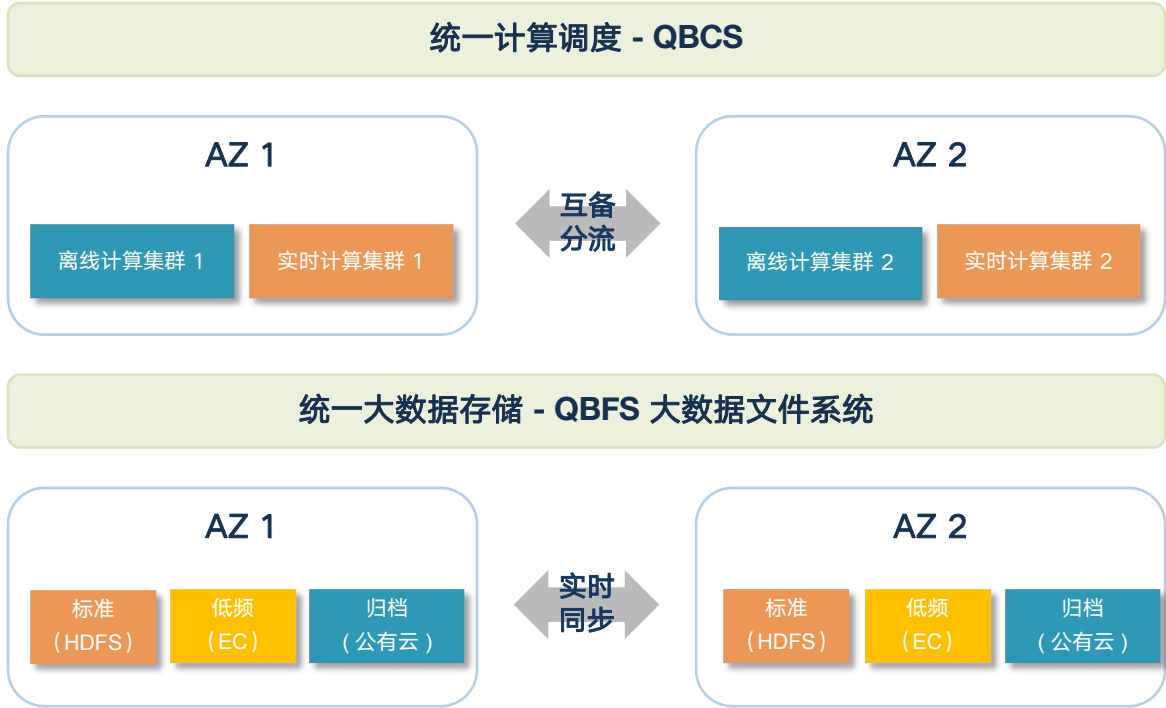
特性	Hive	Iceberg
规模/成本	PB级、廉价	PB级、廉价
修改粒度	分区级覆盖	文件级 更新，支持跨分区修改
提交机制	重命名数据文件，对象存储需复制	无需重命名
下游重新计算	修改涉及所有分区	增量拉取，仅快照间 增量 部分
时效性	离线，天级/小时级	近实时 ，5分钟级
制定执行计划 列举目录调用数	O(N)，N=查询命中分区数量，对象存储前缀列举	常数级 ，元文件+Snapshot文件
分区过滤	分区值	分区值、 隐式分区
文件过滤	无	列值基于统计、字典等 过滤
版本控制	无	可回滚到特定快照版本
行级更新	无	V2格式支持
ACID	高版本可以分区级ACID	支持行级

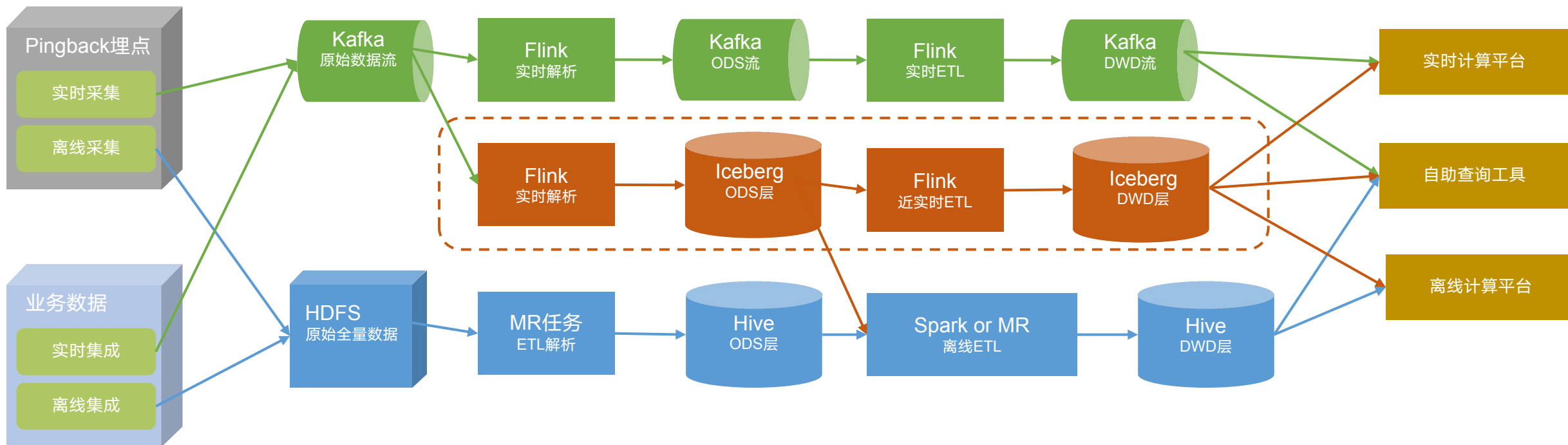


Hive

Iceberg





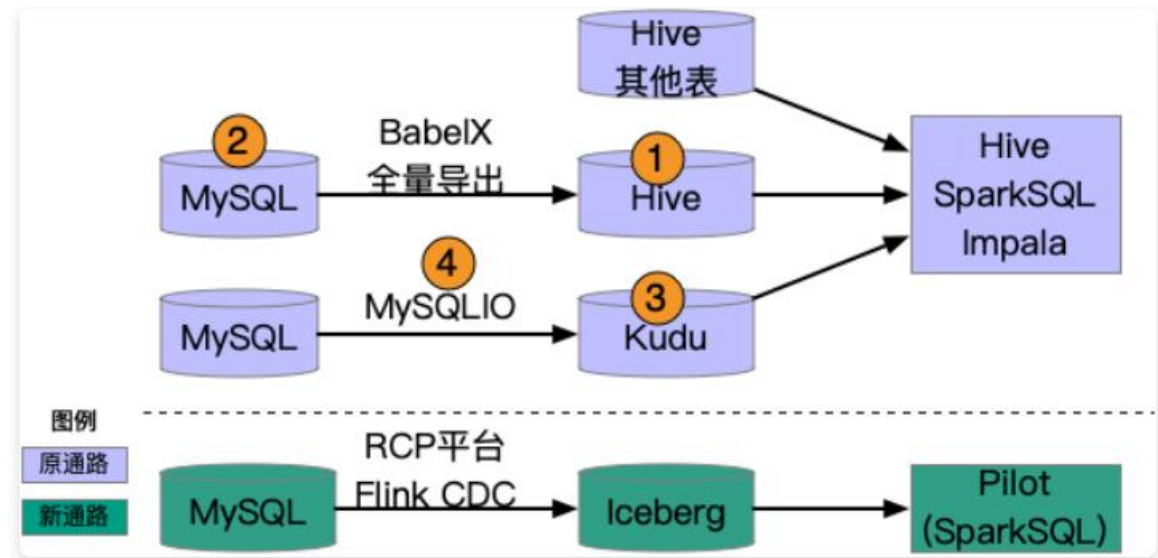
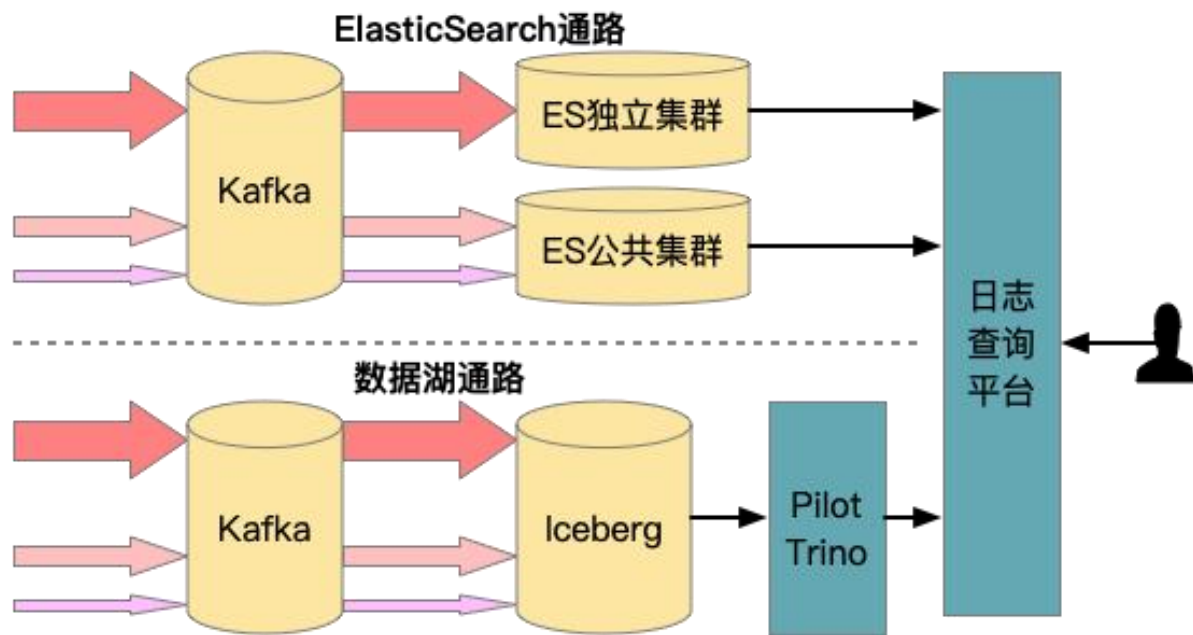


小时级->近实时

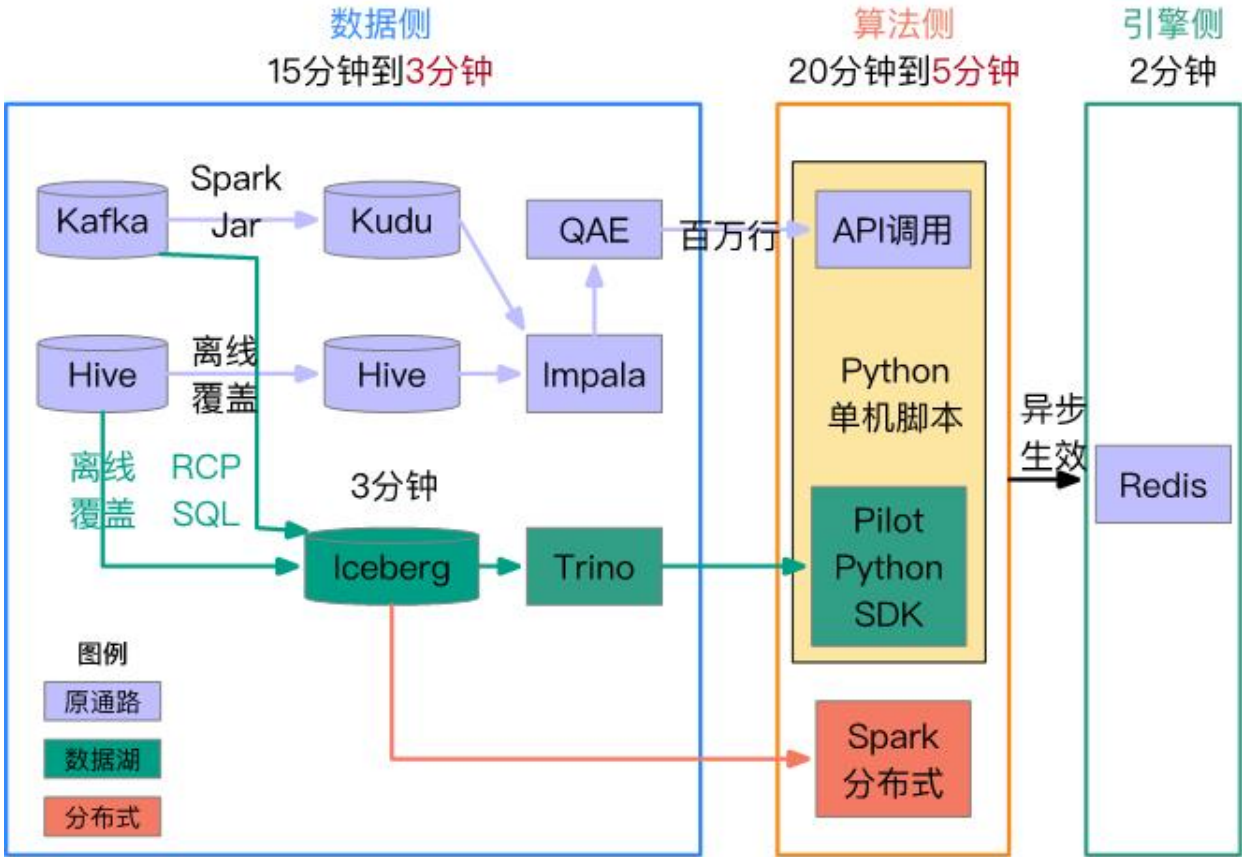
链路维护成本降低

节省60%实时链路资源

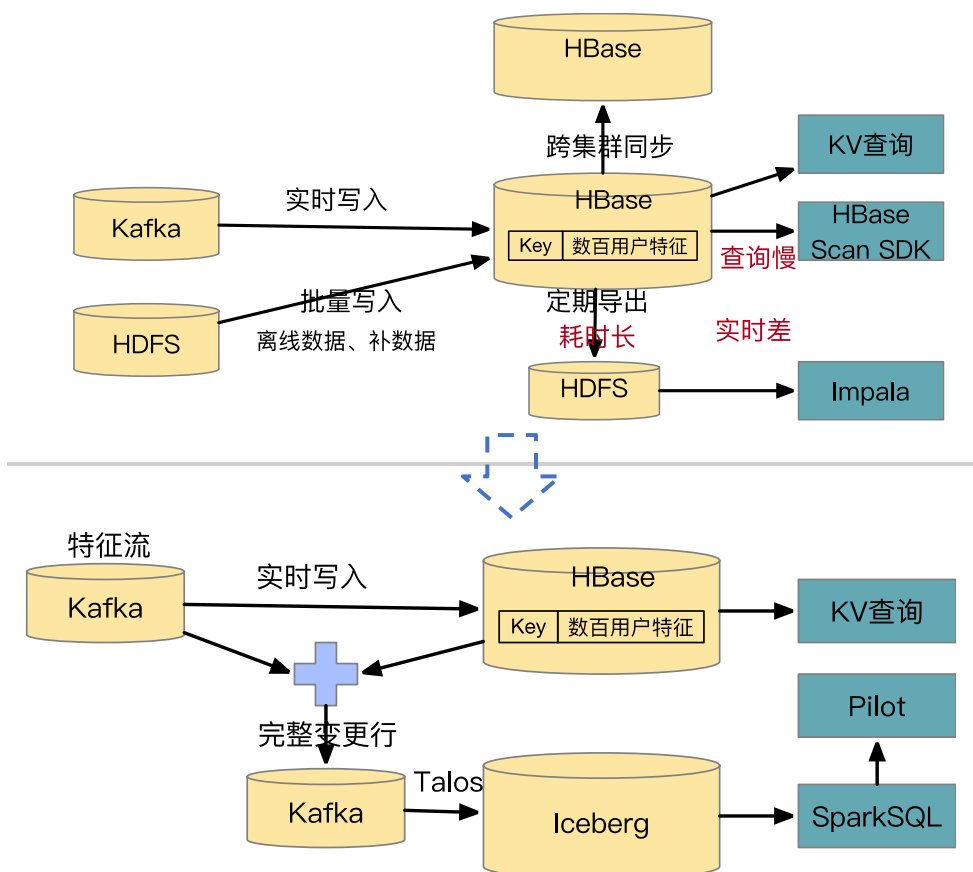
流批一体链路广泛应用



流批一体链路广泛应用

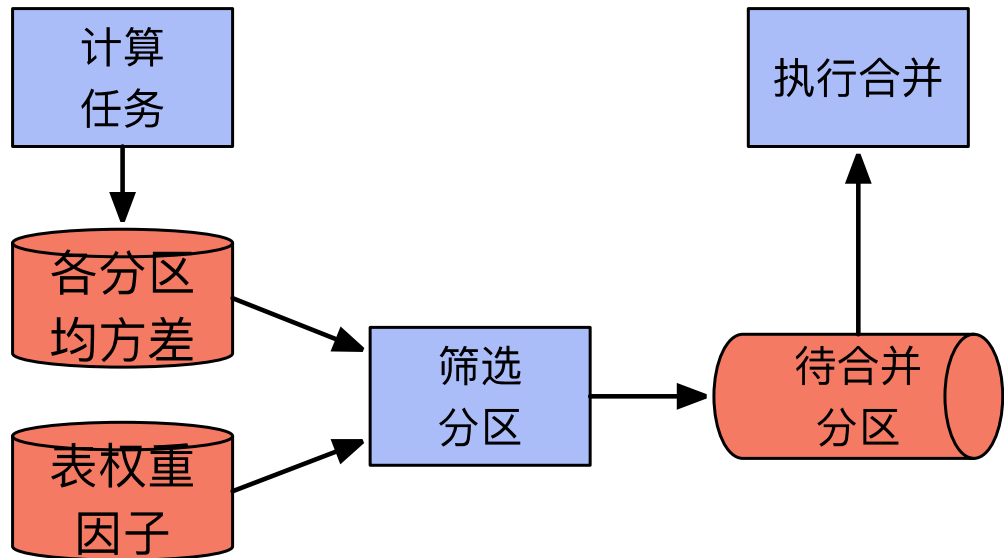


广告数据

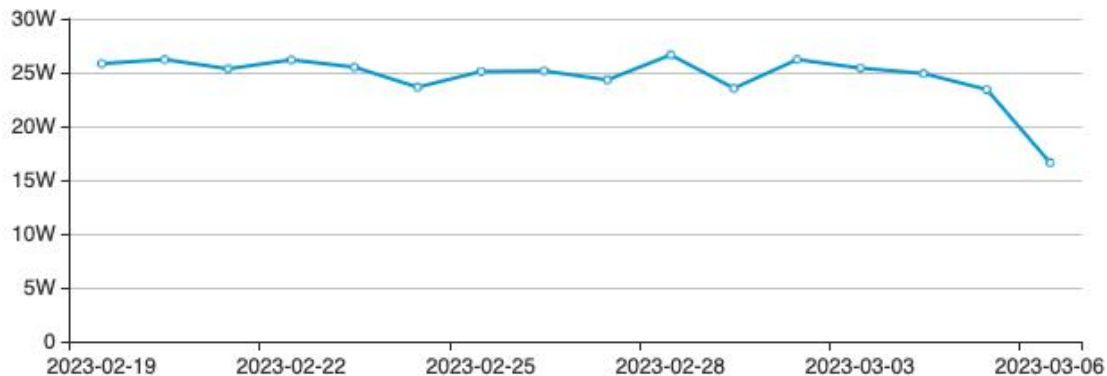


标签数据

性能优化——小文件智能合并



smallFileReduceCountInfo (单位: 个)



定时合并

- 合并任务参数复杂，配置困难
- 合并时机、合并范围：譬如3小时后合并小时分区，一天后合并天分区
- 如合并范围过小：则小文件过多，查询性能下降
- 如合并范围过大：则有重复合并，写放大

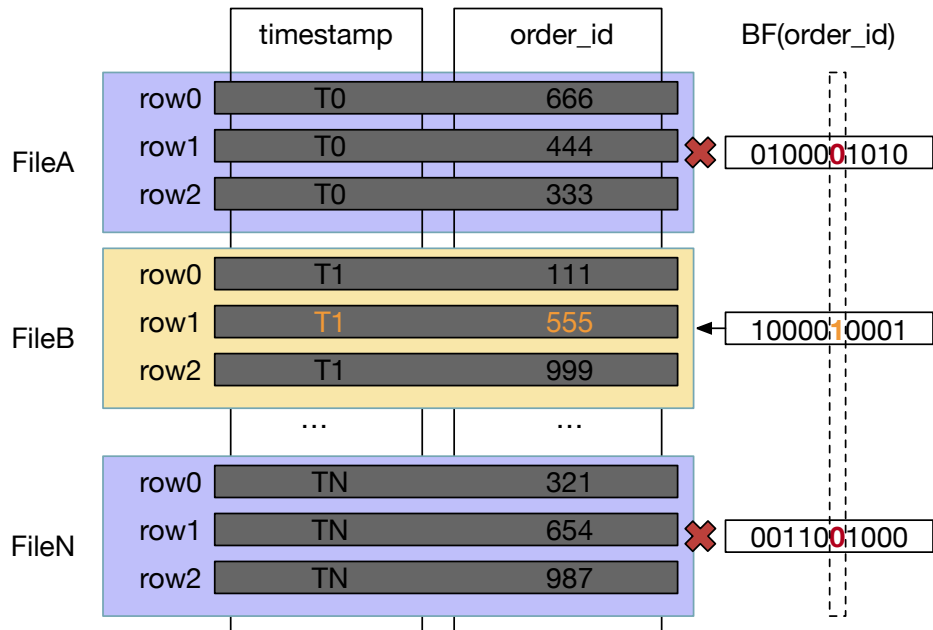
智能合并

- 基于分区下文件大小均方差自动选择待合并分区
- $MSE = \sum_{i=0}^n (Target_i - Actual_i)^2 \div N$
- 微调：业务设置权重、执行失败权重降级
- 业务无需任何配置

参考：[Netflix- Optimizing data warehouse storage](#)

性能优化 - BloomFilter

`SELECT * FROM order_table WHERE order_id = '555';`



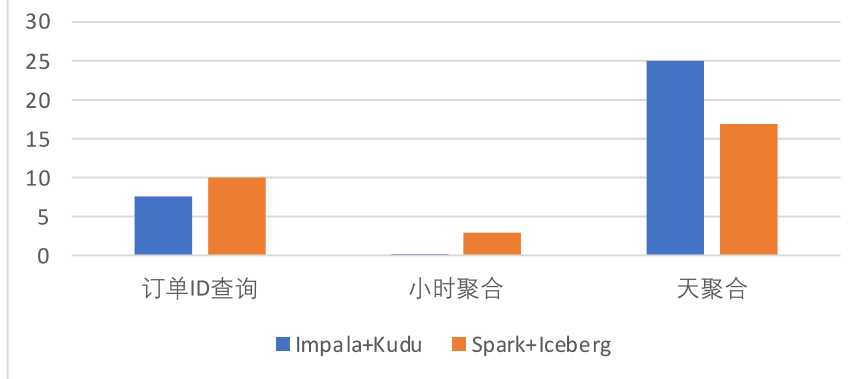
场景说明

- 查询指定ID（如订单ID，用户ID）
- 将ID映射到bitmap一位，出现置1
- 若文件该ID对应bit为0，则必定不包含该值

代码实现

- Parquet 1.12支持Bloom Filter
- 修改Iceberg源码开启BF，修改Spark/Trino应用过滤 (已贡献给社区ISSUE-4831)

订单Iceberg表结合BloomFilter点查询性能大幅提升，整理和Impala+Kudu接近



落地效果

查询速度提升

- SparkSQL: 订单ID查询由948秒降低到10秒，整体性能接近于Impala查询Kudu
- Trino: 开启BF之后，文件过滤 **98.5%**，总执行时间为 40%，峰值内存为 25%，CPU 时间为 5%

存储空间增加

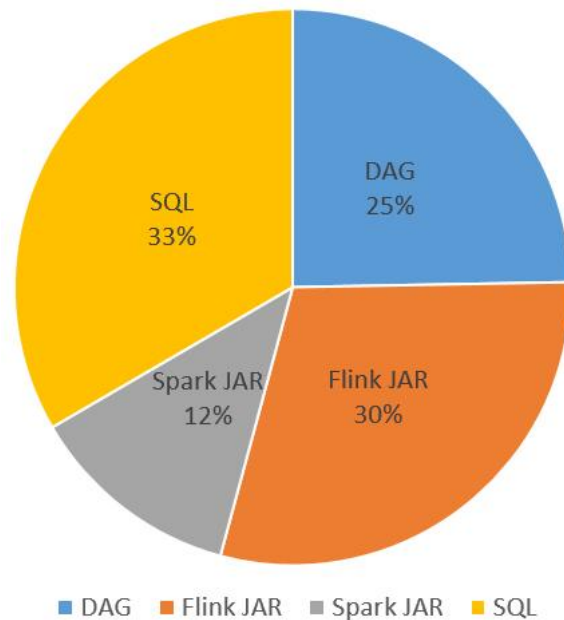
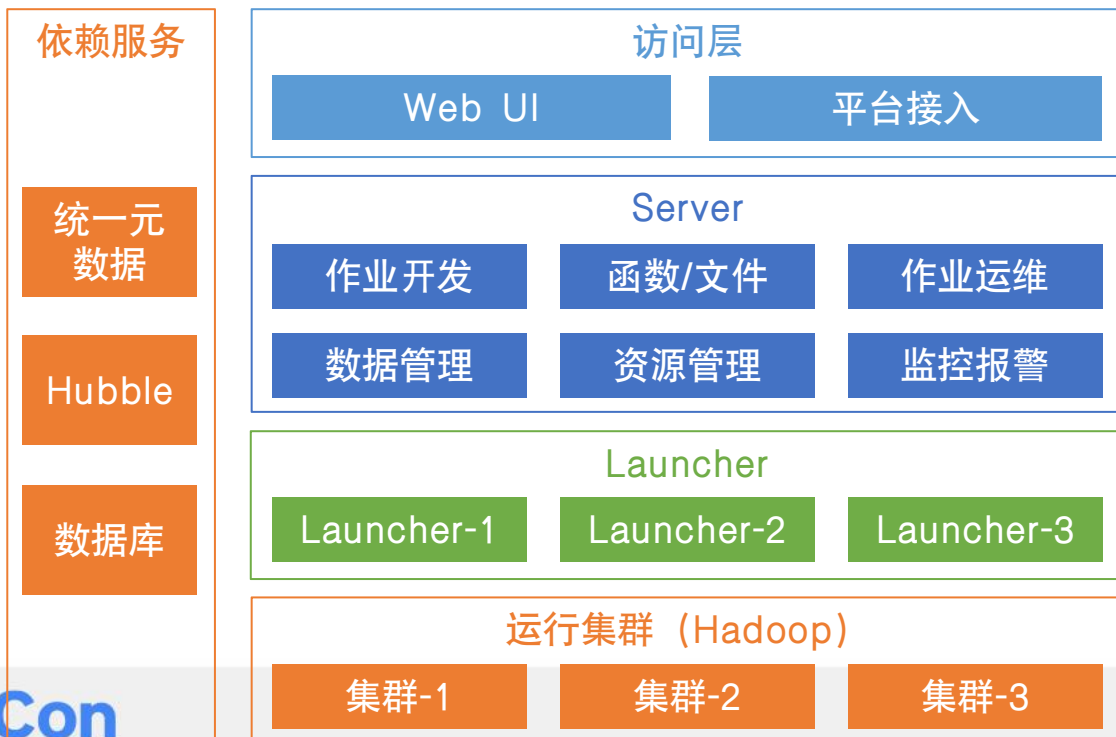
- 原先884G，开启BF后913G，**3%**额外空间

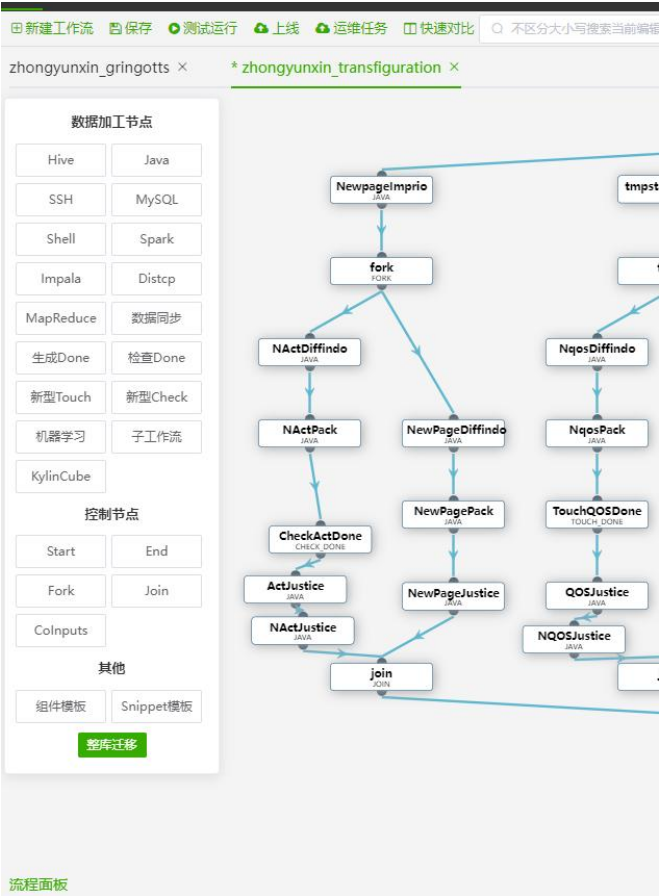
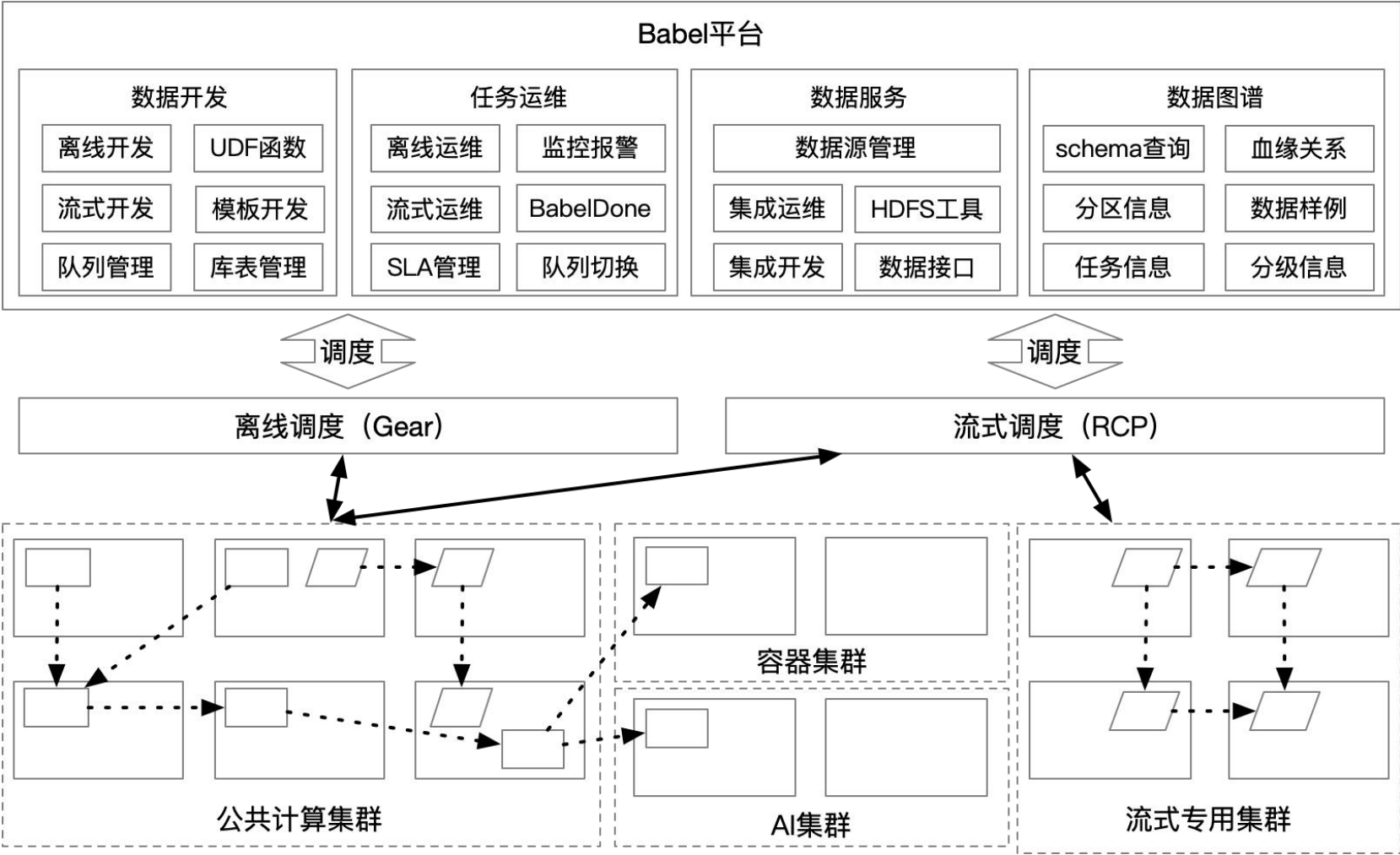
流式计算平台

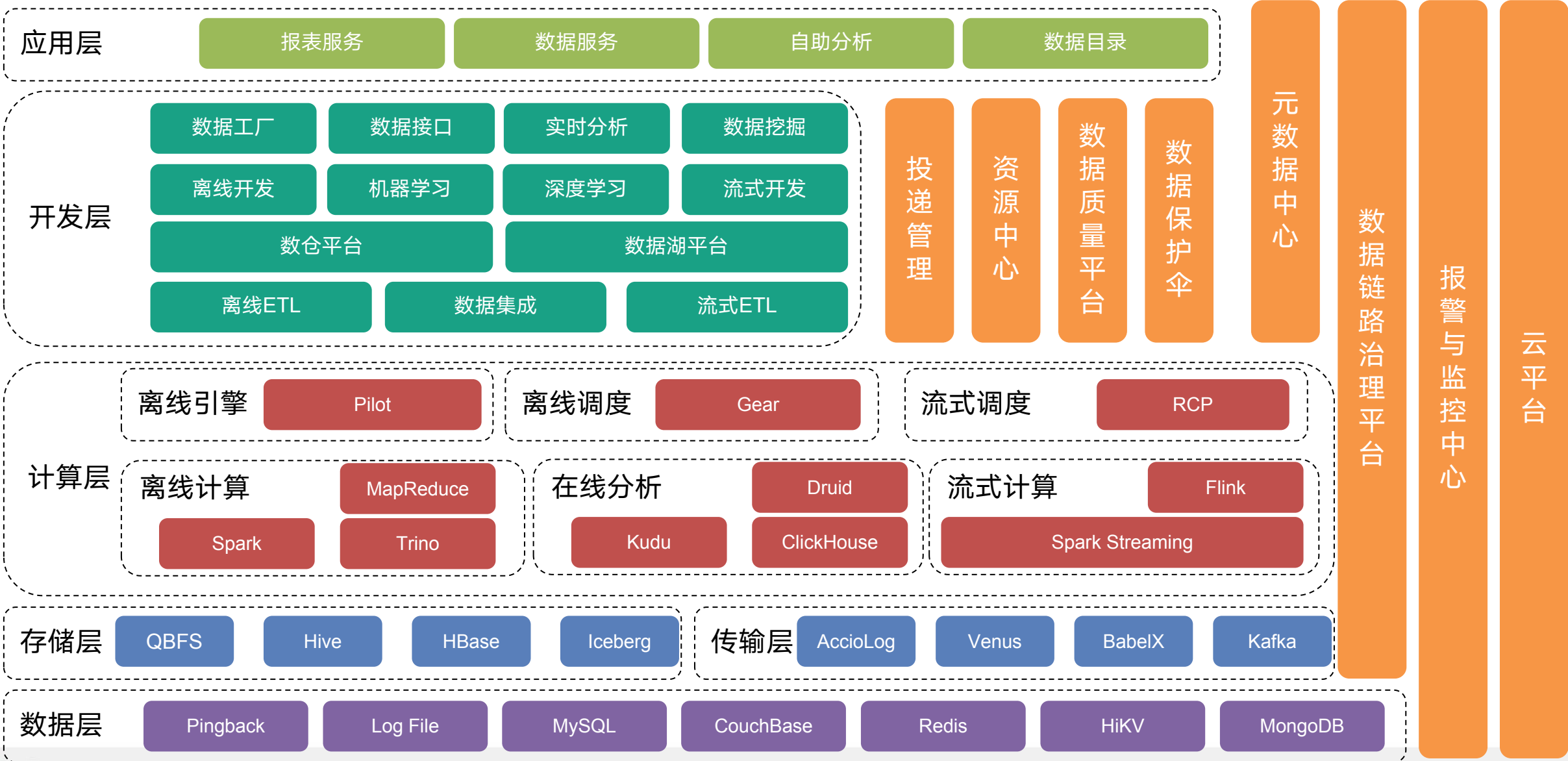
- RCP (Real-time Computing Platform) :

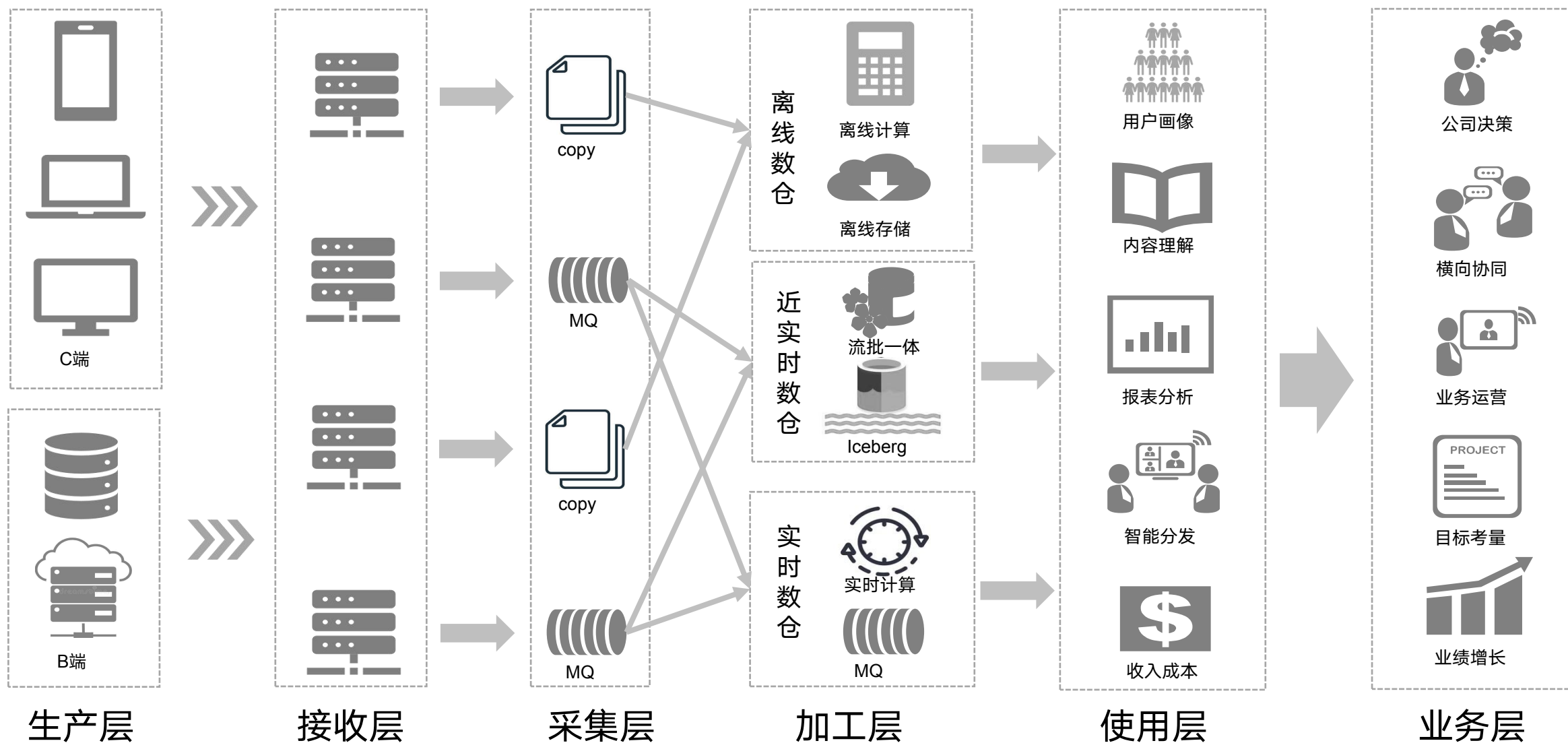
- 流程完整: 数据摄取 → 计算 → 分发
- 多种开发模式: JAR/SQL/DAG
- 对接统一元数据中心

- 架构











THANKS

软件正在重新定义世界

Software Is Redefining The World