AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

AI驱动软件研发
全面进入数字化时代

中国·深圳 11.24-25

AI+
software
Development
Digital
summit

OpenMLDB：以实时特征驱动实时智能决策

陈迪豪 第四范式

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 深圳站

K+ 全球软件研发行业创新峰会

会议时间：2024.05.24-25

K+峰会 上海站

K+ 全球软件研发行业创新峰会

会议时间：2024.09.20-21

AiDD 峰会

AiDD峰会 深圳站

AI+ 软件研发数字峰会

会议时间：2023.11.24-25

AiDD峰会 北京站

AI+ 软件研发数字峰会

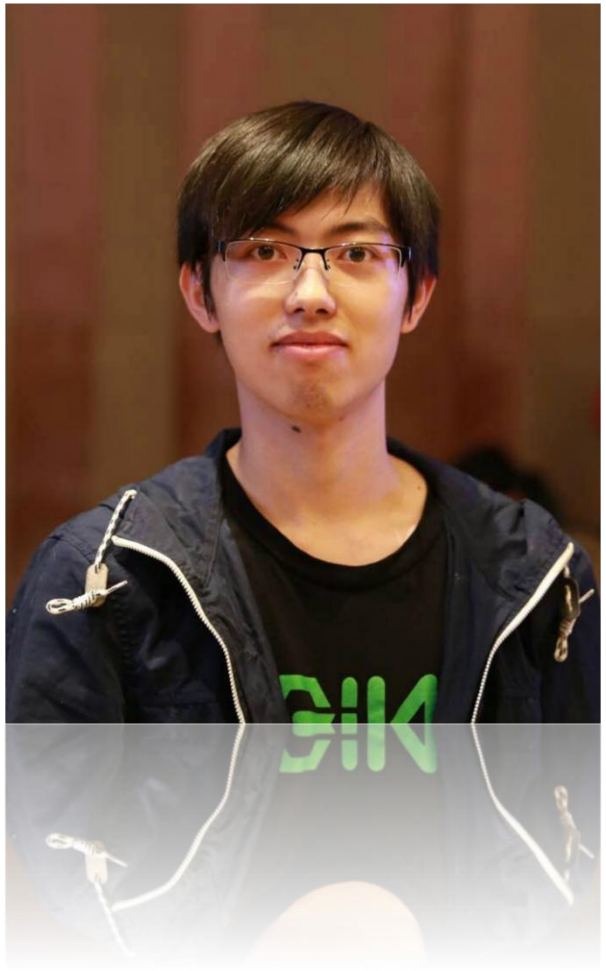
会议时间：2024.07.19-20

AiDD峰会 深圳站

AI+ 软件研发数字峰会

会议时间：2024.11.15-16

演讲嘉宾



陈迪豪

第四范式平台架构师，OpenMLDB PMC Member

目前担任 OpenMLDB PMC 以及第四范式平台架构师，曾担任小米云深度学习平台架构师以及优思德云计算公司存储和容器团队负责人。活跃于分布式系统、机器学习相关的开源社区，也是HBase、OpenStack、TensorFlow、TVM等开源项目贡献者。

目录
CONTENTS

PART 01
实时智能决策的工程化挑战

实时智能决策的工程化挑战

基于机器学习的实时智能决策，需要毫秒级的实时计算能力

两大 AI 应用：感知类、决策类

硬实时计算真正满足实时决策需求 – 实时数据、实时计算

流式计算为 Big Data 和 BI 设计

分钟/小时 级别 | Seconds 秒级 | Milliseconds 毫秒级

批量计算 Batch → 流式计算 Streaming → 硬实时计算 Hard Real Time

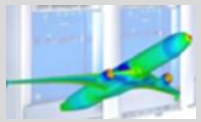
python, Spark Streaming, Flink, sas, APACHE STORM, kafka, TensorFlow

现在市面上所谓的AI实时计算大都是流式计算


AI无人车

AI事中反欺诈


量化交易


航空航天

硬实时场景蕴藏巨大商业价值，鲜有通用商业化产品

银行要求毫秒级业务响应

以某银行反欺诈场景为例

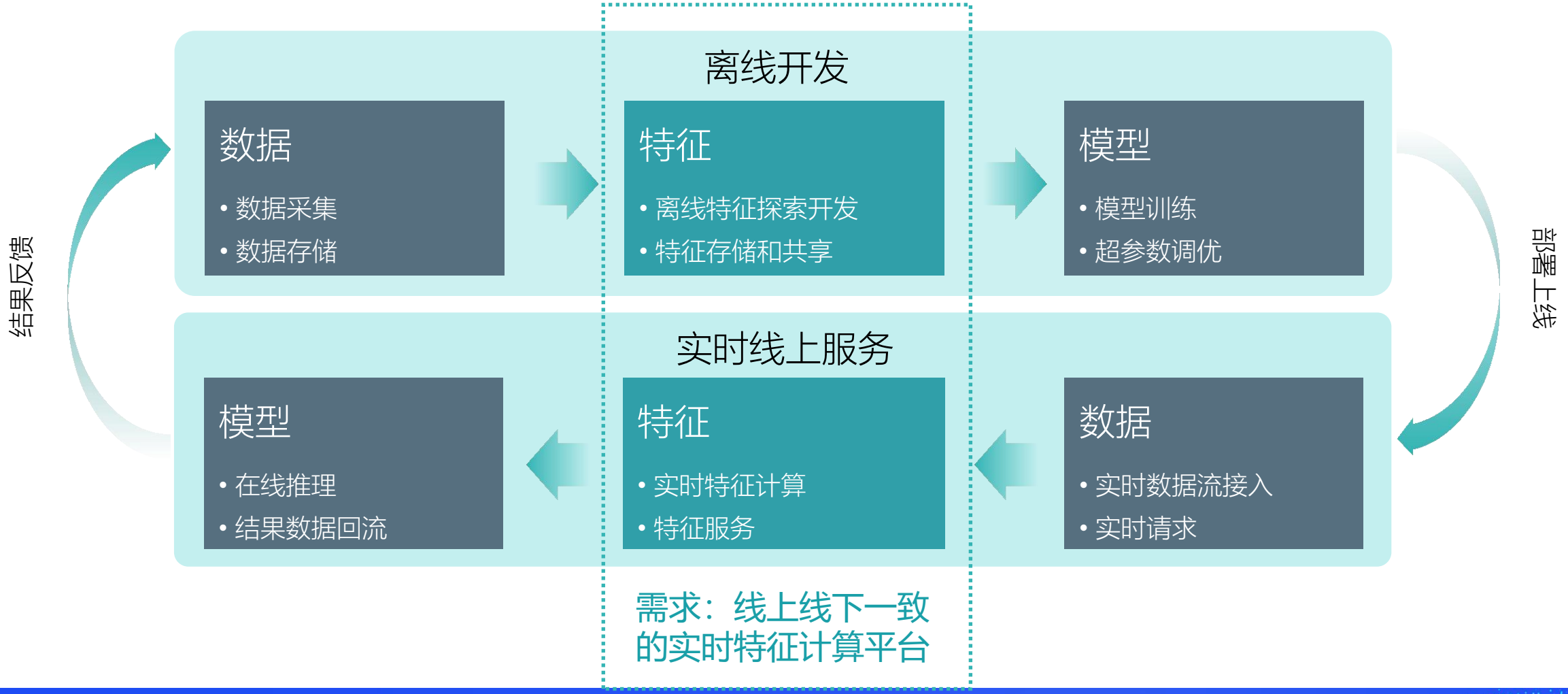
客户需求：

特征计算响应时间 20ms 内，高准召率的事中反欺诈系统

解决方案	响应时间	准召率
传统规则系统	~200ms	较差
客户自研系统	~50ms	中等
第四范式先知	<20ms	优等

实时智能决策的工程化挑战

基于机器学习的智能决策从离线开发到上线全流程



实时智能决策的工程化挑战

事中反欺诈交易的实时特征计算

刷卡记录

卡号	刷卡金额	刷卡时间
012159	1000	2022/01/12 08:00:00

虚拟插入

历史交易表

卡号	刷卡金额	刷卡时间（已排序）
012112	223	2022/01/12 02:00:00
012159	15	2022/01/12 06:00:00
012159	1000	2022/01/12 07:59:55
012159	2000	2022/01/12 07:59:57
012159	1000	2022/01/12 08:00:00

基于窗口聚合

10s

3h

特征计算

生成的特征

卡号	刷卡金额	过去10秒内：刷卡次数 | 刷卡最大金额 | 最小金额 | 平均金额	过去三小时内：刷卡次数 | 刷卡最大金额 | 最小金额 | 平均金额
012159	1000	3 | 2000 | 1000 | 1333	4 | 2000 | 14 | 1003

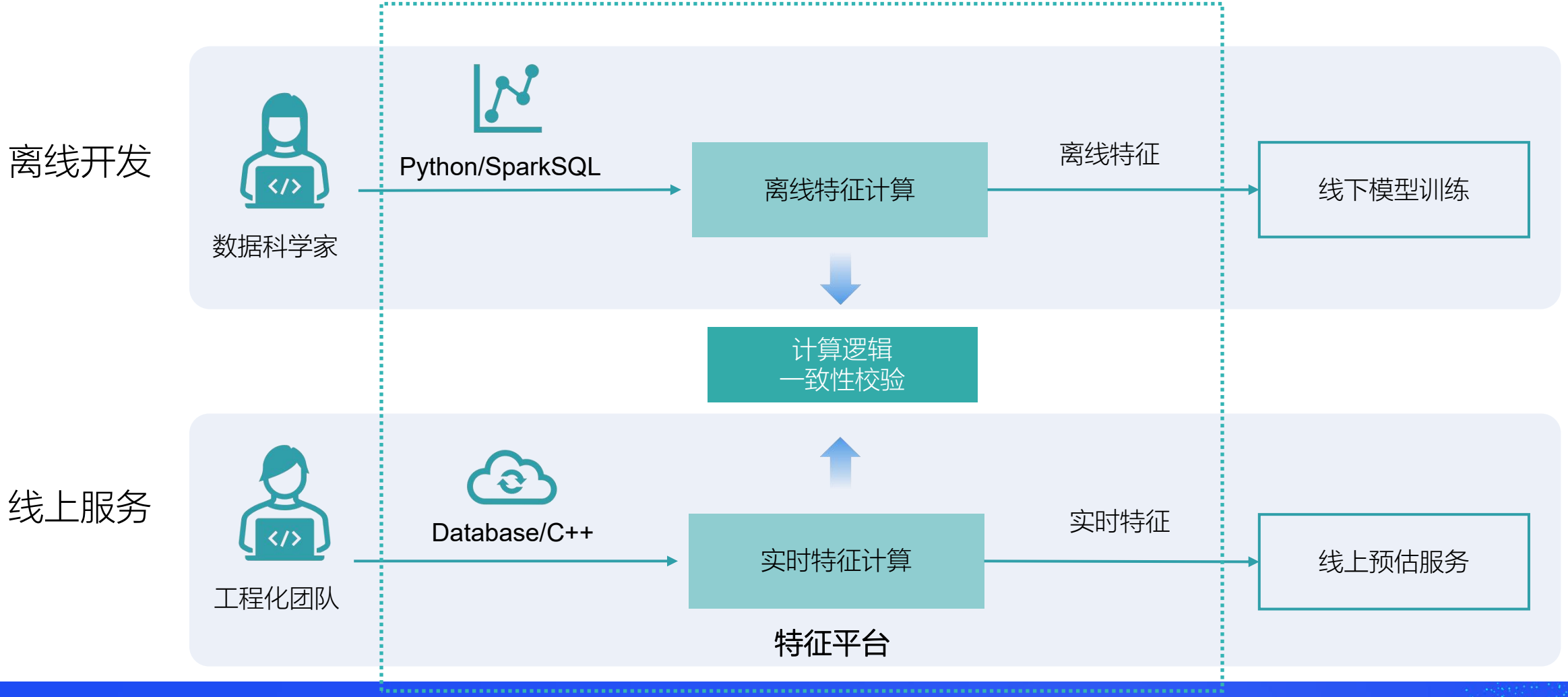
模型推理 → 欺诈交易？

工程化需求

1. 线上线下一致性
2. 低延迟、高并发、高可用

实时智能决策的工程化挑战

传统特征开发：离线开发和线上服务分离，高成本投入



实时智能决策的工程化挑战

线上线下不一致性可能的原因

工具能力的不一致性

离线开发

Python

```python
import pandas as pd
t1 = pd.read_csv("data.csv")
account_feat = t1['account'].std()
```

$$\sqrt{\sum_{i=0}^{N}\frac{(x_i-\mu)^2}{N-1}}$$

标准差

(Bessel's Correction)

线上应用

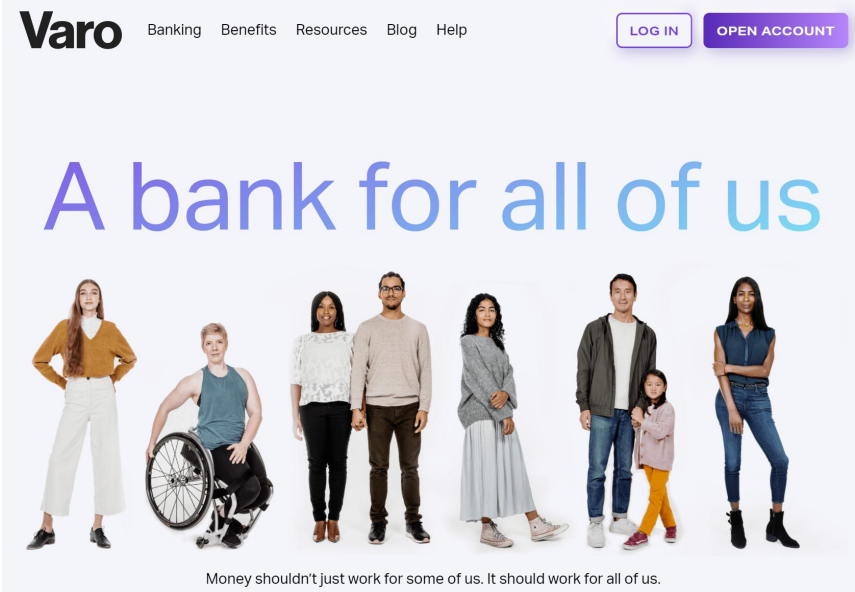
MySQL

```sql
select std(account) as account_feat from t1;
```

$$\sqrt{\sum_{i=0}^{N}\frac{(x_i-\mu)^2}{N}}$$

标准差

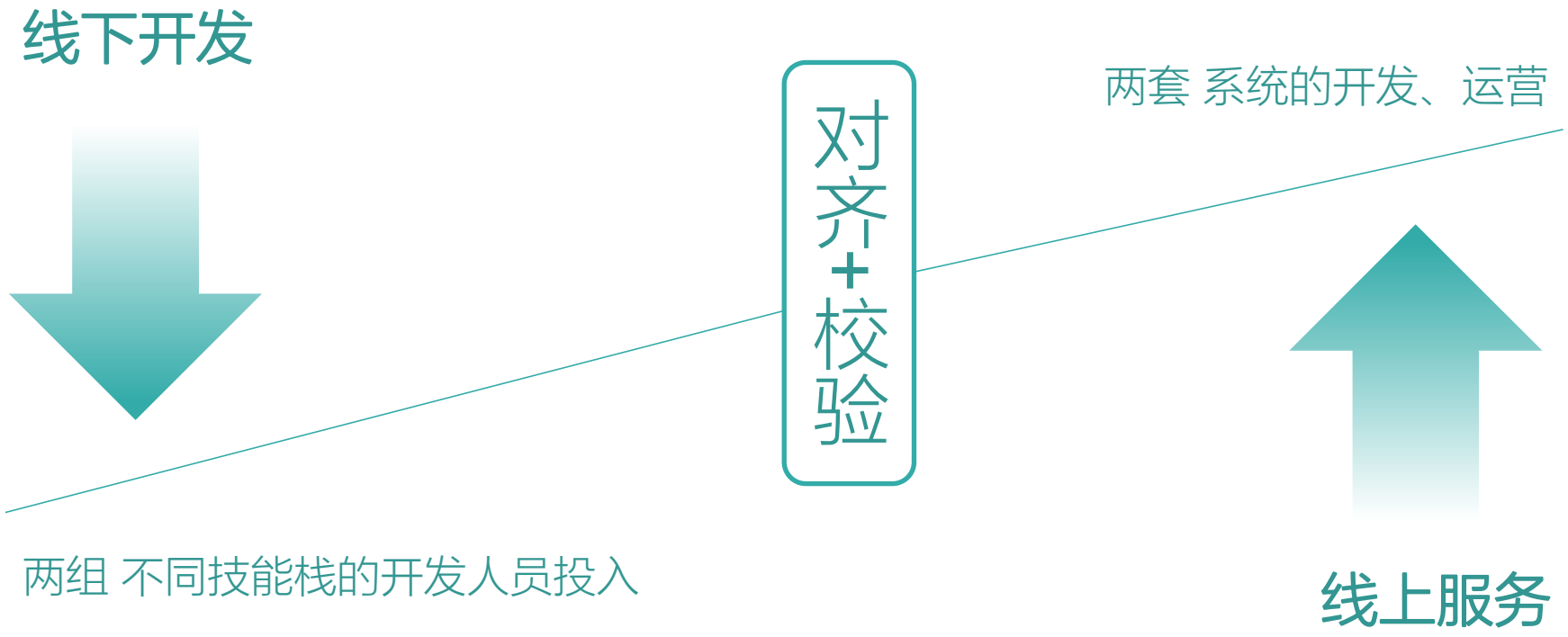
需求沟通的认知差



	Account Balance
线上应用	current “account balance”
离线开发	“account balance” as of yesterday

实时智能决策的工程化挑战

线上线下一致性校验带来的高昂工程化落地成本



PART 02

OpenMLDB 提供线上线下一致的实时特征计算

OpenMLDB 提供线上线下一致的实时特征计算

OpenMLDB 发展历程：从闭源走向开源

过往5年 RTIDB/FEDB（第四范式闭源） **2021.6** → OpenMLDB（开源/商业化）

开源前，跟随第四范式 先知 平台，在 100+场景 落地，覆盖超过 300个节点。

开源后，以开放姿态积极拥抱社区开发者、整合开源生态，提供商业化定制和支持。

akulaku 37手游 JDT 京东科技 HUAWEI 唯品会 品牌特卖 UBiX

主要使用场景

信用卡现金分期精准营销	贷前风险评分	营销获客	个性化推荐	反洗钱可疑交易智能识别
信用卡账户风险预警	合规额度决策	风险管理	投顾客户挖掘	信用卡申请反欺诈
欺诈养卡防控	理财个性化推荐	零售贷款反欺诈	历史客户激活	客户流失预警
网点流量预测	交易欺诈评分	现金分期个性化推荐	信用卡交易反欺诈	金融产品推荐

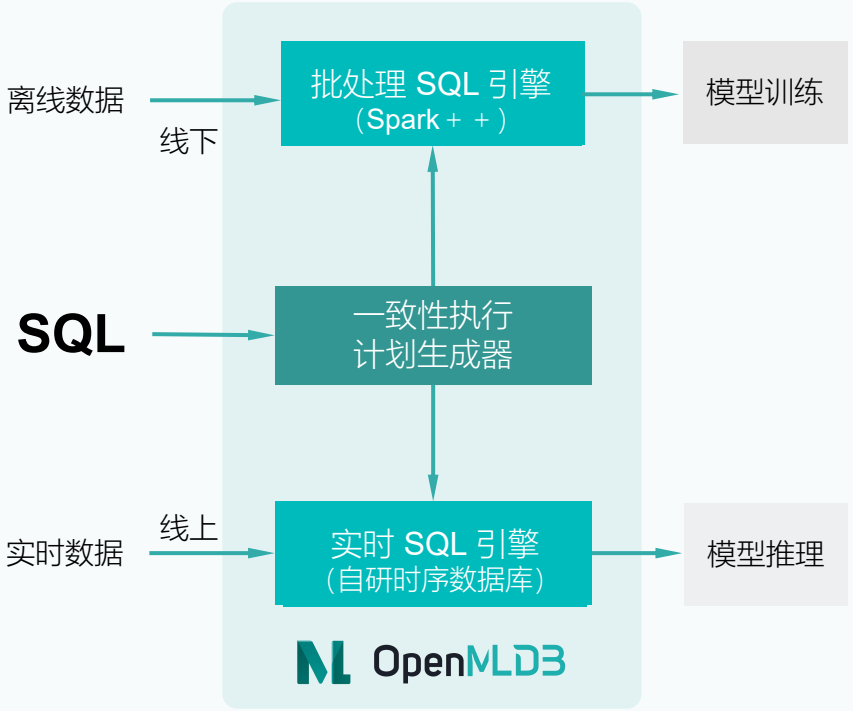
OpenMLDB 提供线上线下一致的实时特征计算

OpenMLDB：开源机器学习数据库，线上线下一致的特征平台

原有流程

- Step 1：特征脚本开发 — 科学家
- Step 2：重构，满足线上低延时、高吞吐、高可用 — 研发工程师
- Step 3：线上线下一致性校验 — 科学家+研发工程师
- Step 4：生产上线 — 运维工程师

OpenMLDB 抽象架构



基于 OpenMLDB 的流程

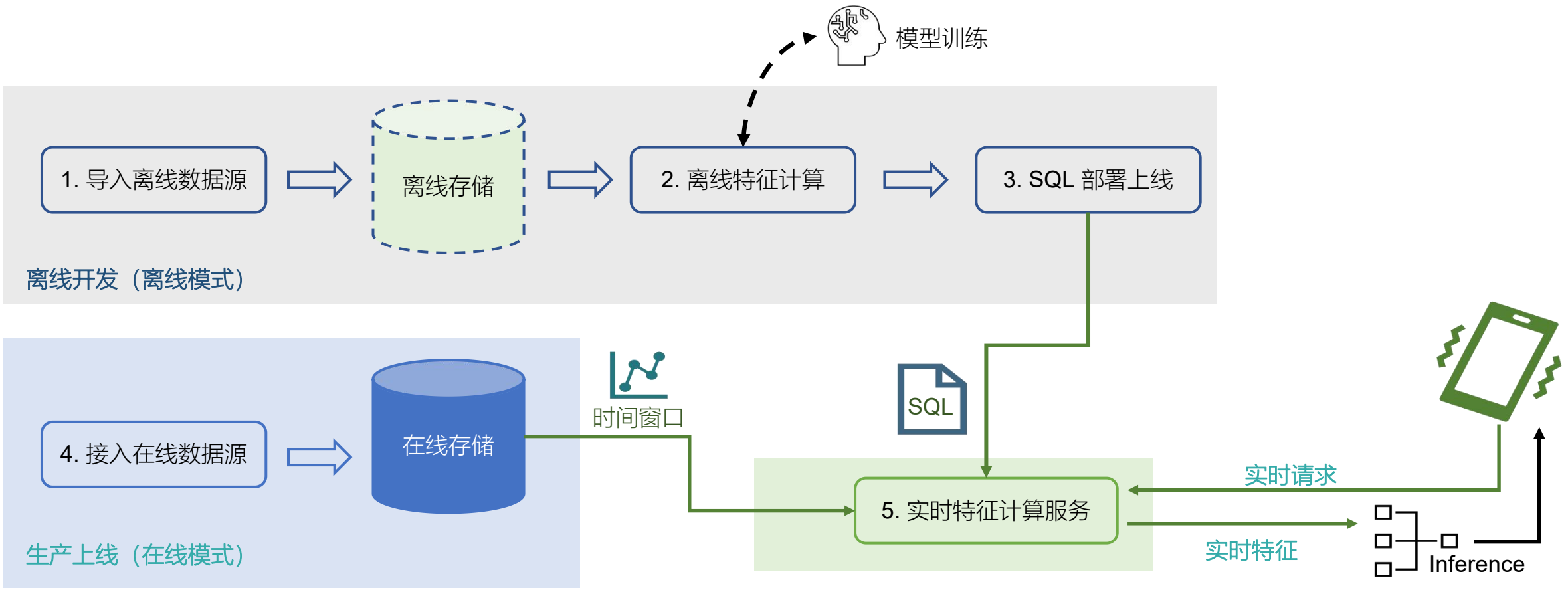
- Step 1：特征脚本开发 — 科学家

使用OpenMLDB，
实现开发即上线，
节省数月人天成本

- Step 2：一键生产上线 — 运维工程师

OpenMLDB 提供线上线下一致的实时特征计算

从离线开发到线上服务完整流程



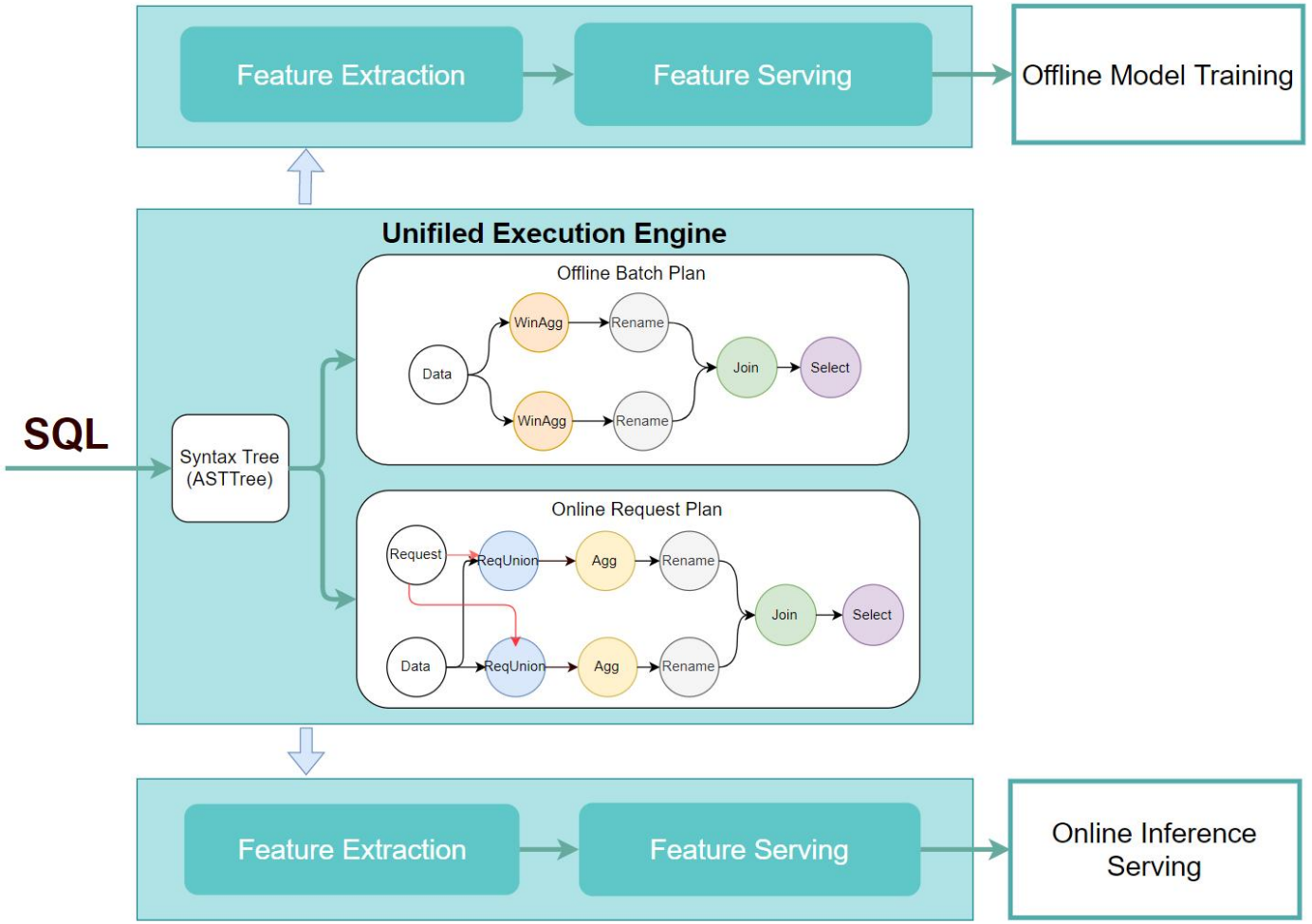
OpenMLDB 提供线上线下一致的实时特征计算

OpenMLDB 提供了一个 线上线下一致 的
毫秒级 实时特征计算平台

- 基于实时数据按需计算 (on-demand)
- 基于 SQL 定义特征
- 生产级平台，分布式、可扩展、高可用

OpenMLDB 提供线上线下一致的实时特征计算

核心组件一：线上线下一致性执行引擎



- 统一的底层计算函数
- 逻辑计划到物理计划的线上线下执行模式自适应调整

线上线下一致性得到 天然保障

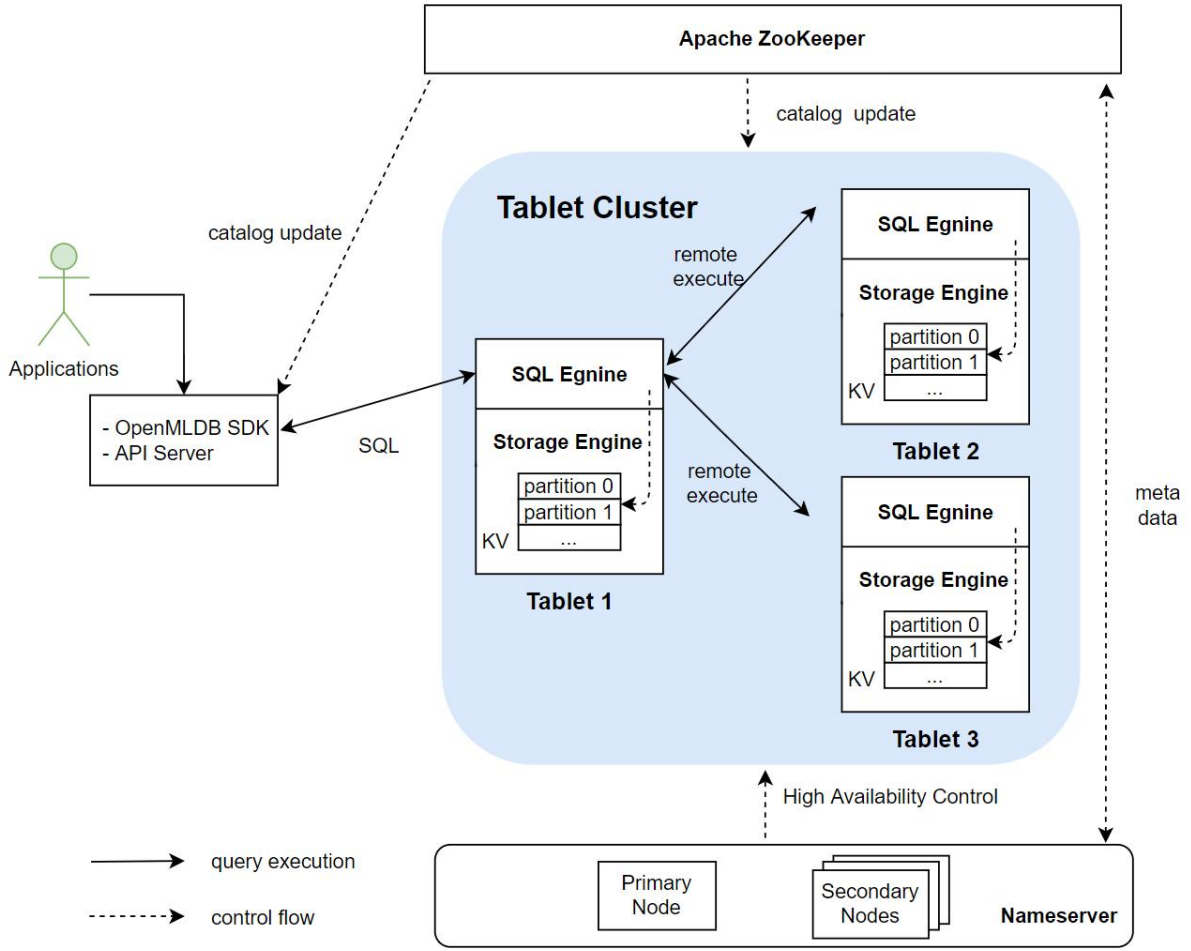
OpenMLDB 提供线上线下一致的实时特征计算

核心组件二：高性能实时 SQL 引擎

分布式实时 SQL 引擎主要模块

- **ZooKeeper** – 元数据存储和管理
- **Nameserver** – tablet 管理和故障转移
- **Tablets**
  - 分布式 SQL 执行引擎
  - 分布式存储引擎：
    内存、磁盘双存储引擎
- 高性能、可扩展、高可用

详细线上引擎架构描述参见：
https://openmldb.feishu.cn/wiki/wikcnavULzxKH5Aka3ox0871R2f

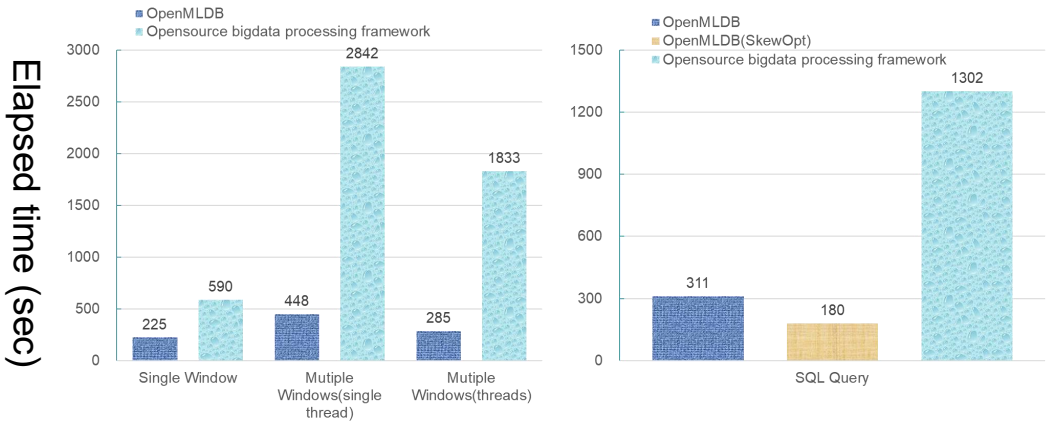


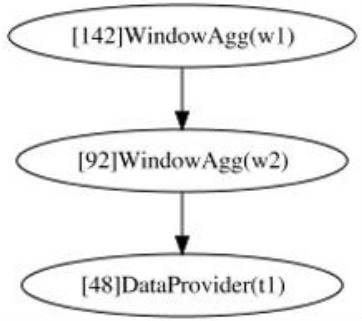
OpenMLDB 提供线上线下一致的实时特征计算

核心组件三：面向特征计算的优化的离线计算引擎

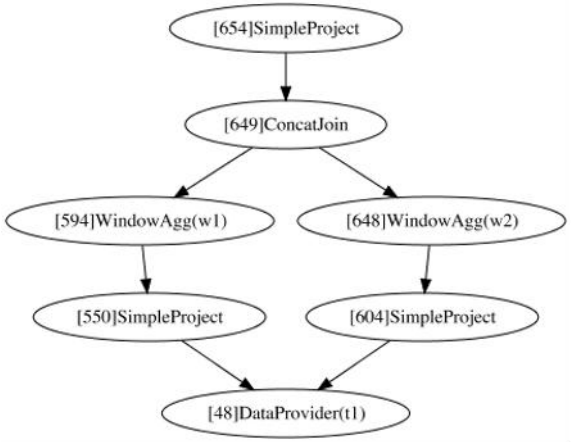
- 多窗口并行计算优化
- 数据倾斜计算优化
- SQL 语法扩展
- 针对特征计算优化的 OpenMLDB Spark 发行版

```
SELECT
  min(age) OVER w1 as w1_min_age,
  min(age) OVER w2 as w2_min_age
FROM t1
WINDOW
  w1 as (PARTITION BY name ORDER by age ROWS BETWEEN 10 PRECEDING AND CURRENT ROW),
  w2 as (PARTITION BY age ORDER by age ROWS BETWEEN 10 PRECEDING AND CURRENT ROW)"
```





Spark 3.0.0



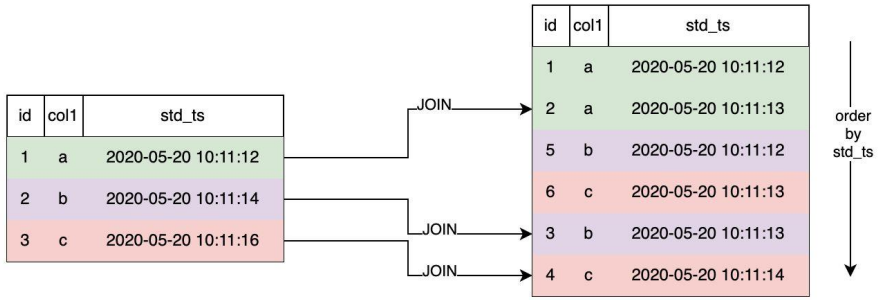
OpenMLDB

OpenMLDB 提供线上线下一致的实时特征计算

核心组件四：针对特征工程的 SQL 扩展

LAST JOIN

多行匹配时，仅匹配最新记录



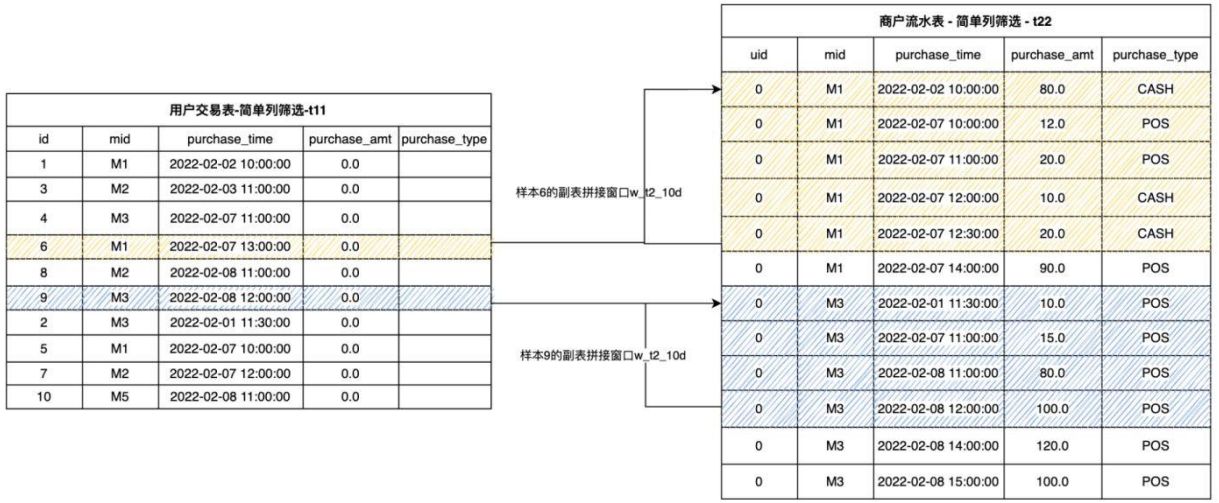
id	col1	std_ts
1	a	2020-05-20 10:11:12
2	b	2020-05-20 10:11:14
3	c	2020-05-20 10:11:16

id	col1	std_ts
1	a	2020-05-20 10:11:12
2	a	2020-05-20 10:11:13
5	b	2020-05-20 10:11:12
6	c	2020-05-20 10:11:13
3	b	2020-05-20 10:11:13
4	c	2020-05-20 10:11:14

id	col1	std_ts	id	col1	std_ts
1	a	2020-05-20 10:11:12	2	a	2020-05-20 10:11:13
2	b	2020-05-20 10:11:14	3	b	2020-05-20 10:11:13
3	c	2020-05-20 10:11:16	4	c	2020-05-20 10:11:14

WINDOW UNION

跨表的 join 和窗口聚合操作 (point-in-time)，避免特征穿越



用户交易表-简单列筛选-t11

id	mid	purchase_time	purchase_amt	purchase_type
1	M1	2022-02-02 10:00:00	0.0	
3	M2	2022-02-03 11:00:00	0.0	
4	M3	2022-02-07 11:00:00	0.0	
6	M1	2022-02-07 13:00:00	0.0	
8	M2	2022-02-08 11:00:00	0.0	
9	M3	2022-02-08 12:00:00	0.0	
2	M3	2022-02-01 11:30:00	0.0	
5	M1	2022-02-07 10:00:00	0.0	
7	M2	2022-02-07 12:00:00	0.0	
10	M5	2022-02-08 11:00:00	0.0	

商户流水表 - 简单列筛选 - t22

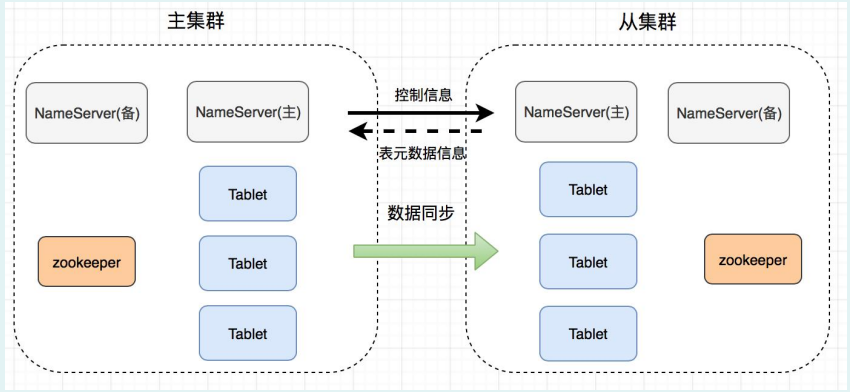
uid	mid	purchase_time	purchase_amt	purchase_type
0	M1	2022-02-02 10:00:00	80.0	CASH
0	M1	2022-02-07 10:00:00	12.0	POS
0	M1	2022-02-07 11:00:00	20.0	POS
0	M1	2022-02-07 12:00:00	10.0	CASH
0	M1	2022-02-07 12:30:00	20.0	CASH
0	M1	2022-02-07 14:00:00	90.0	POS
0	M3	2022-02-01 11:30:00	10.0	POS
0	M3	2022-02-07 11:00:00	15.0	POS
0	M3	2022-02-08 11:00:00	80.0	POS
0	M3	2022-02-08 12:00:00	100.0	POS
0	M3	2022-02-08 14:00:00	120.0	POS
0	M3	2022-02-08 15:00:00	100.0	POS

OpenMLDB 提供线上线下一致的实时特征计算

高级生产级特性，保证系统稳定性和可扩展性

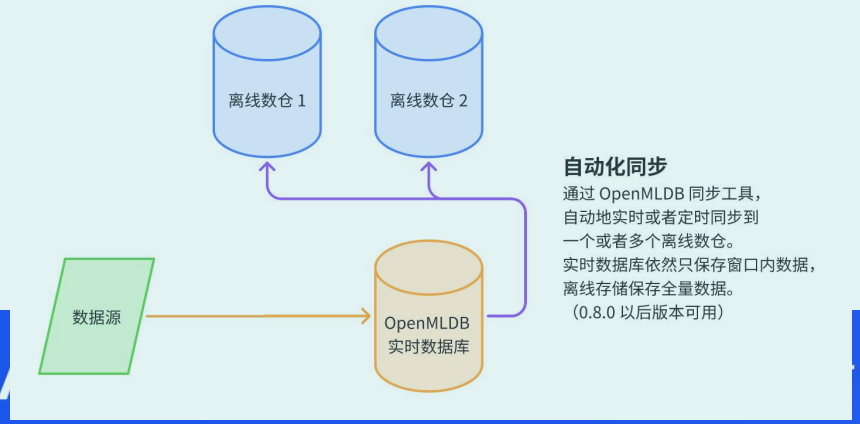
跨机房容灾

构建主从集群，进一步提升可靠性



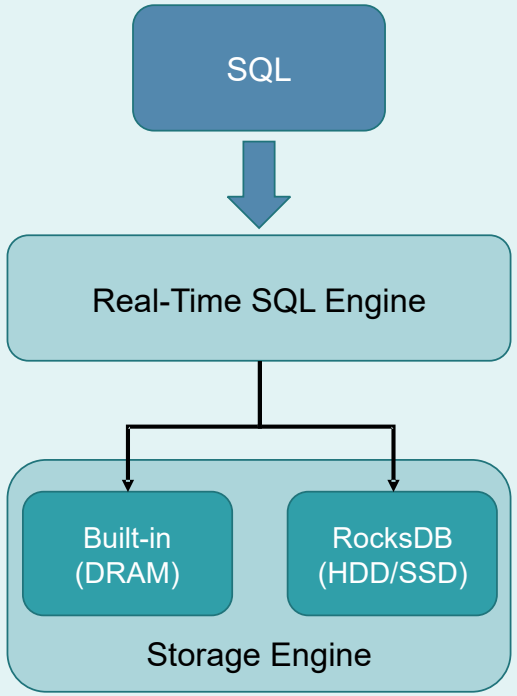
自动化在离线数据同步

简化运维操作，保证数据一致性



线上内存/磁盘双引擎存储架构

平衡性能和成本



智能化运维和诊断

智能诊断　分片自动平衡　一键数据恢复

支持不同部署模式

APACHE hadoop

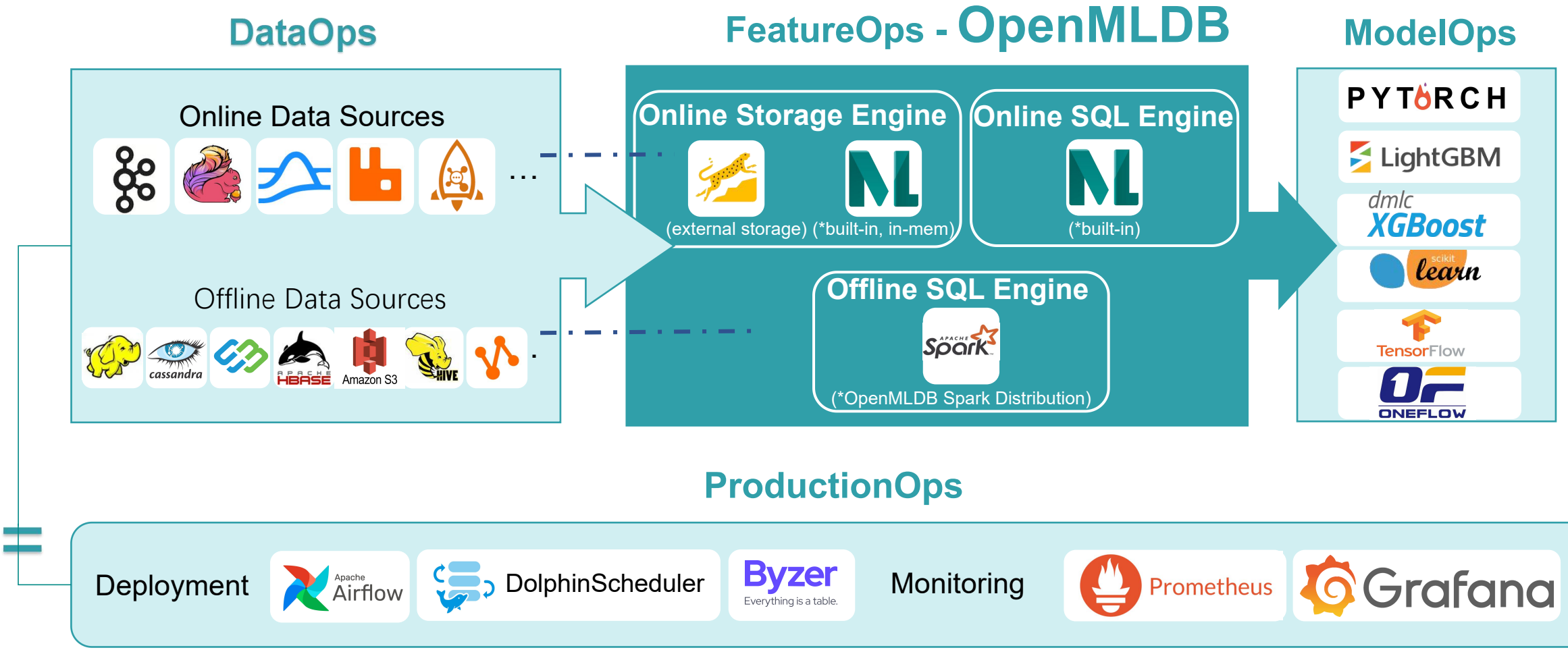
kubernetes

OpenMLDB 原生形态部署

PART 03
社区生态和案例分享

社区生态和案例分享

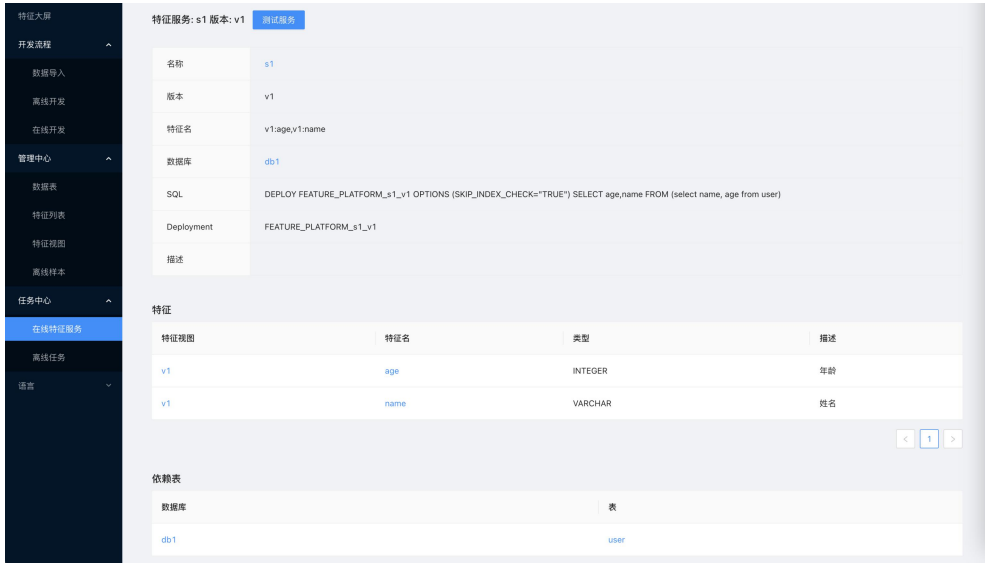
OpenMLDB 上下游开源生态

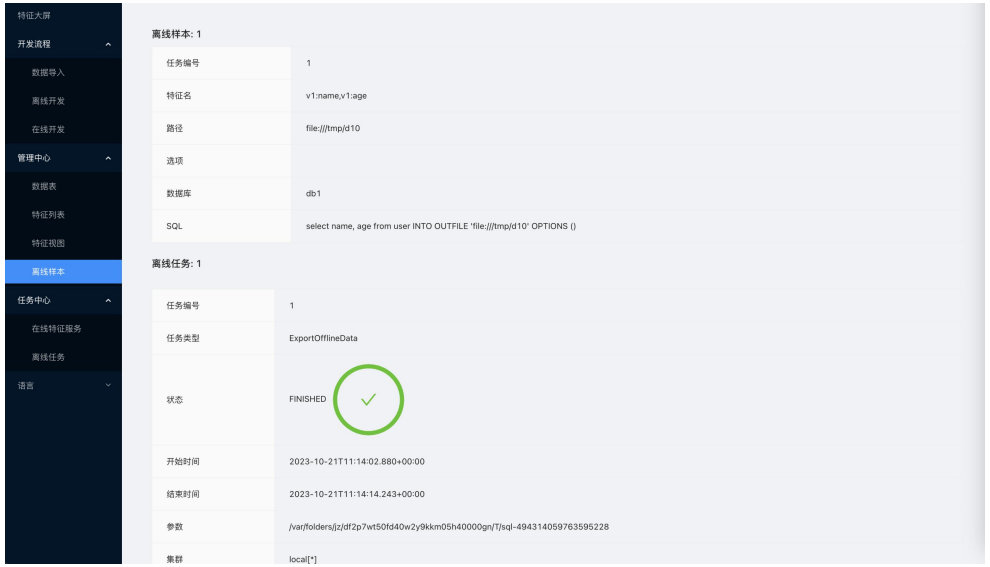


社区生态和案例分享

基于 OpenMLDB 的特征平台

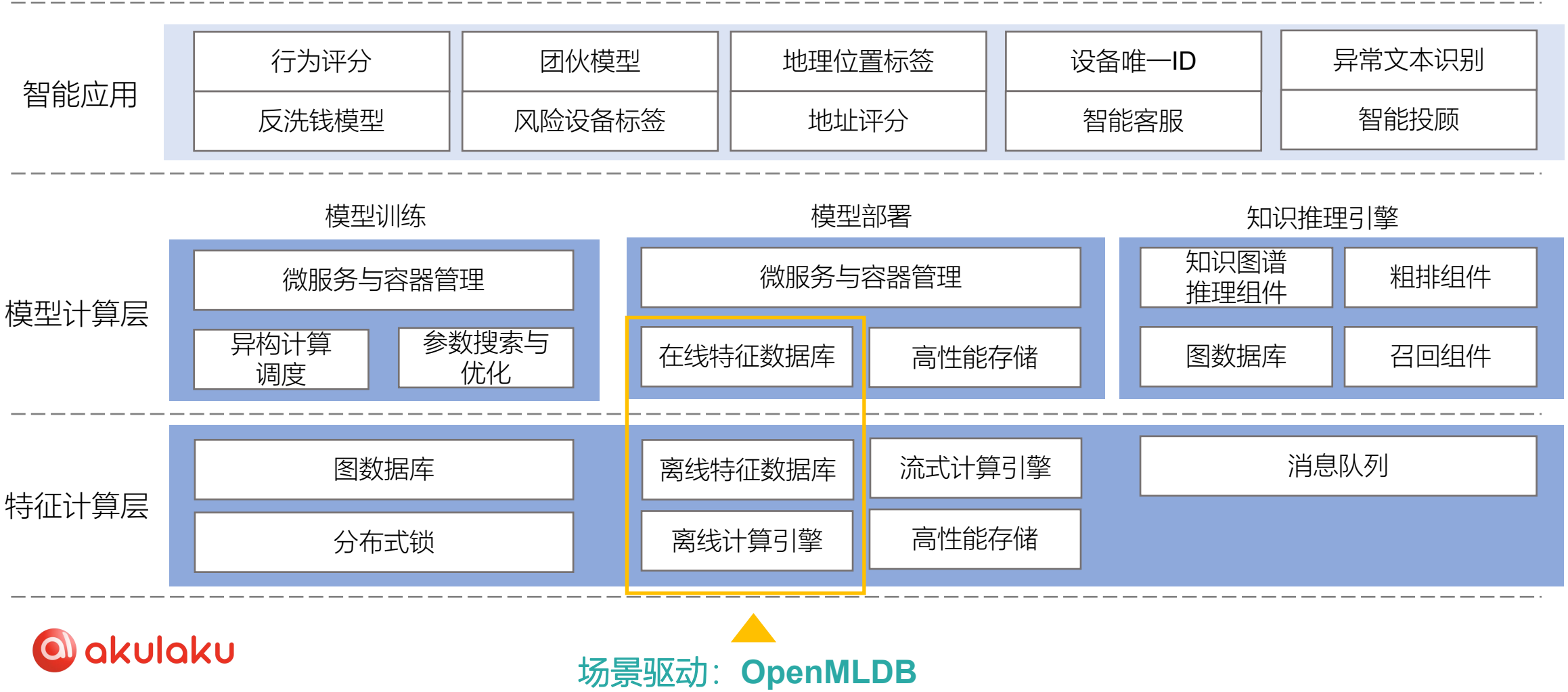
- 可视化特征开发和管理界面
- 基于 DAG 的大型复杂特征开发辅助
- 特征灵活复用
- 特征血缘管理和版本管理
- 同时支持毫秒级实时特征、离线特征





社区生态和案例分享

OpenMLDB 案例 – Akulaku 智能计算架构中的特征平台



社区生态和案例分享

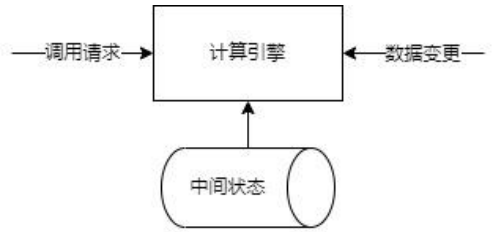
Akulaku 智能风控场景，对 10 亿条订单进行窗口特征计算，达到 4 毫秒延迟性能

特征计算环节难点

- **线上部署**：低延迟，高时效性，尽可能反映数据变更
- **线下分析**：高吞吐量
- **逻辑一致**：线下分析和线上部署的逻辑需要完全一致

OpenMLDB 解决方案

- 场景驱动：业务调用环节驱动，实时计算结果，现用现算
- **具体方案**：1）使用SQL作为离线和在线的桥梁；2）在线基于时序数据库做时间滑窗




```
(product_id:1,  price:2    ,update_time:100L )
(product_id:1,  price:3    ,update_time:200L )

(product_id:2,  price:3    ,update_time:1000L )
(product_id:2,  price:4    ,update_time:1201L )
(product_id:2,  price:3.5  ,update_time:1100L )
```

```
SELECT COUNT(*)
FROM w100ms

WINDOW w100ms AS
 (PARTITION BY product_id ORDER BY update_time
 ROWS_RANGE BETWEEN 100ms PRECEDING
AND
CURRENT ROW)
```

基于 OpenMLDB 的业务实现

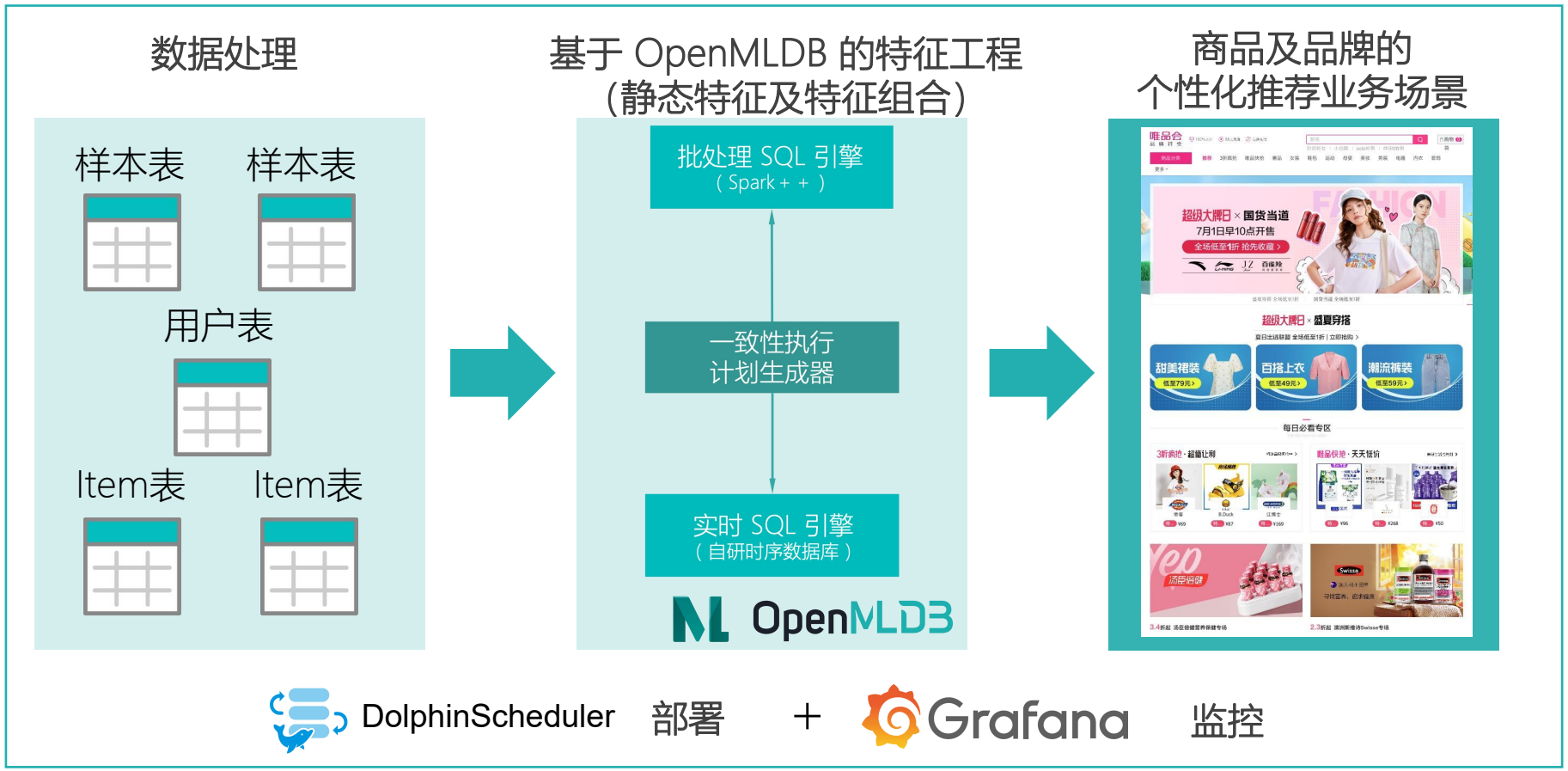
- **场景**：近1天订单个数实时计算
- **数据量**：10亿条订单数据/天
- **需求**：实时更新，时间窗口实时滑动，存在复杂关联需求
- **测试结果**：4毫秒 延迟

akulaku

社区生态和案例分享

唯品会将 OpenMLDB 应用于商品及品牌个性化推荐场景，带来特征开发迭代速度60%的提升

唯品会 VIPSHOP.COM



特征开发迭代速度

5人天 → 2人天

注：
样本表：不同场景下的用户行为表，包括曝光点击收藏
用户表：用户侧所有用户画像信息
Item表：不同物料的全量信息表

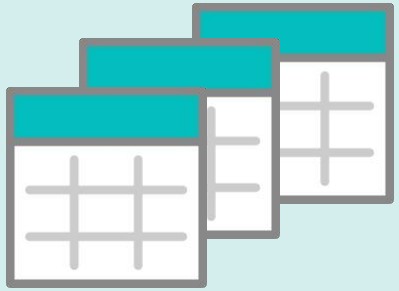
社区生态和案例分享

某头部ICT公司将 OpenMLDB 用于实时商品个性化推荐场景

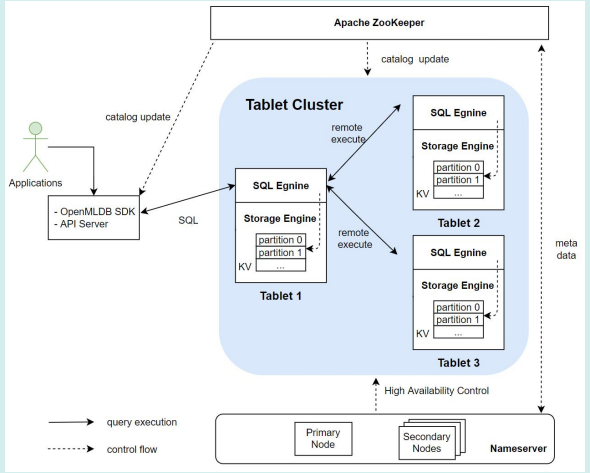
线上实时数据

- 数据分钟级更新
- 7.2亿条订单数据/天

客户数据表

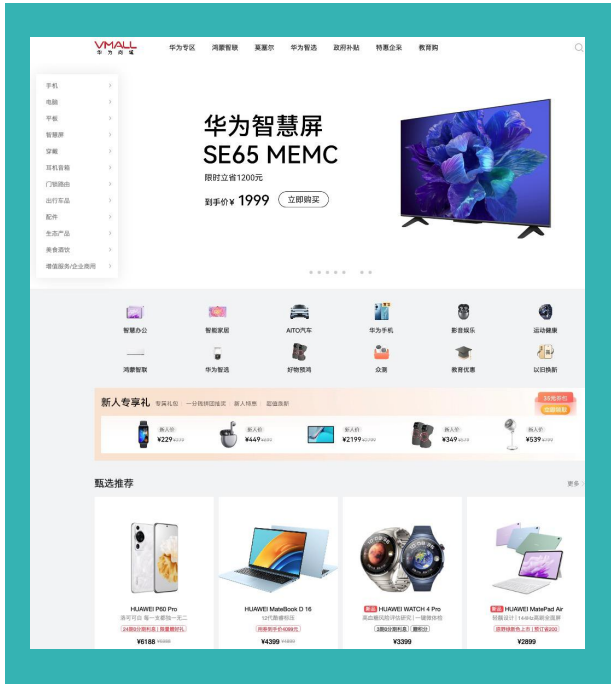


基于OpenMLDB的实时特征抽取



OpenMLDB

小时级特征上线



AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

THANKS