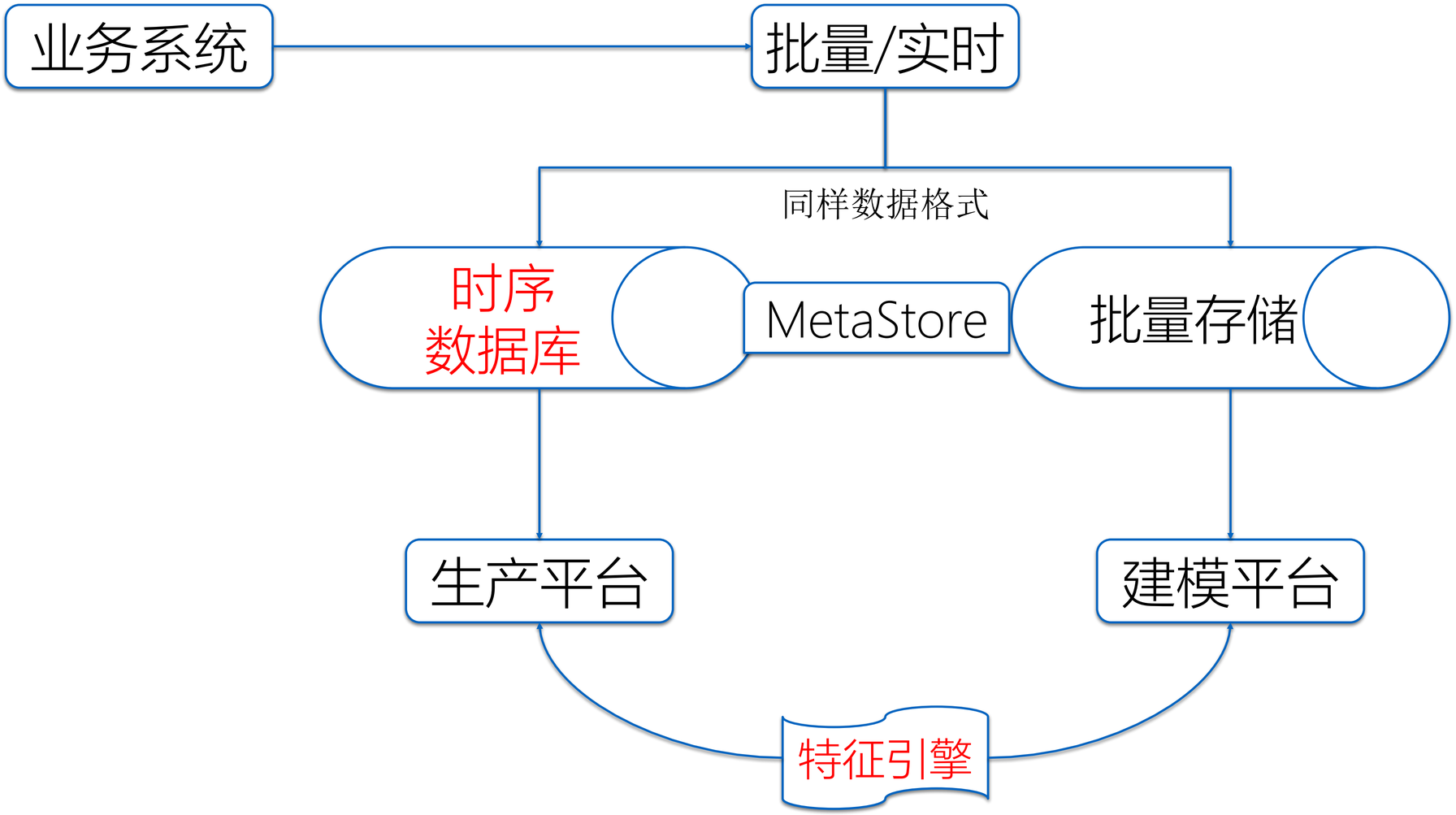
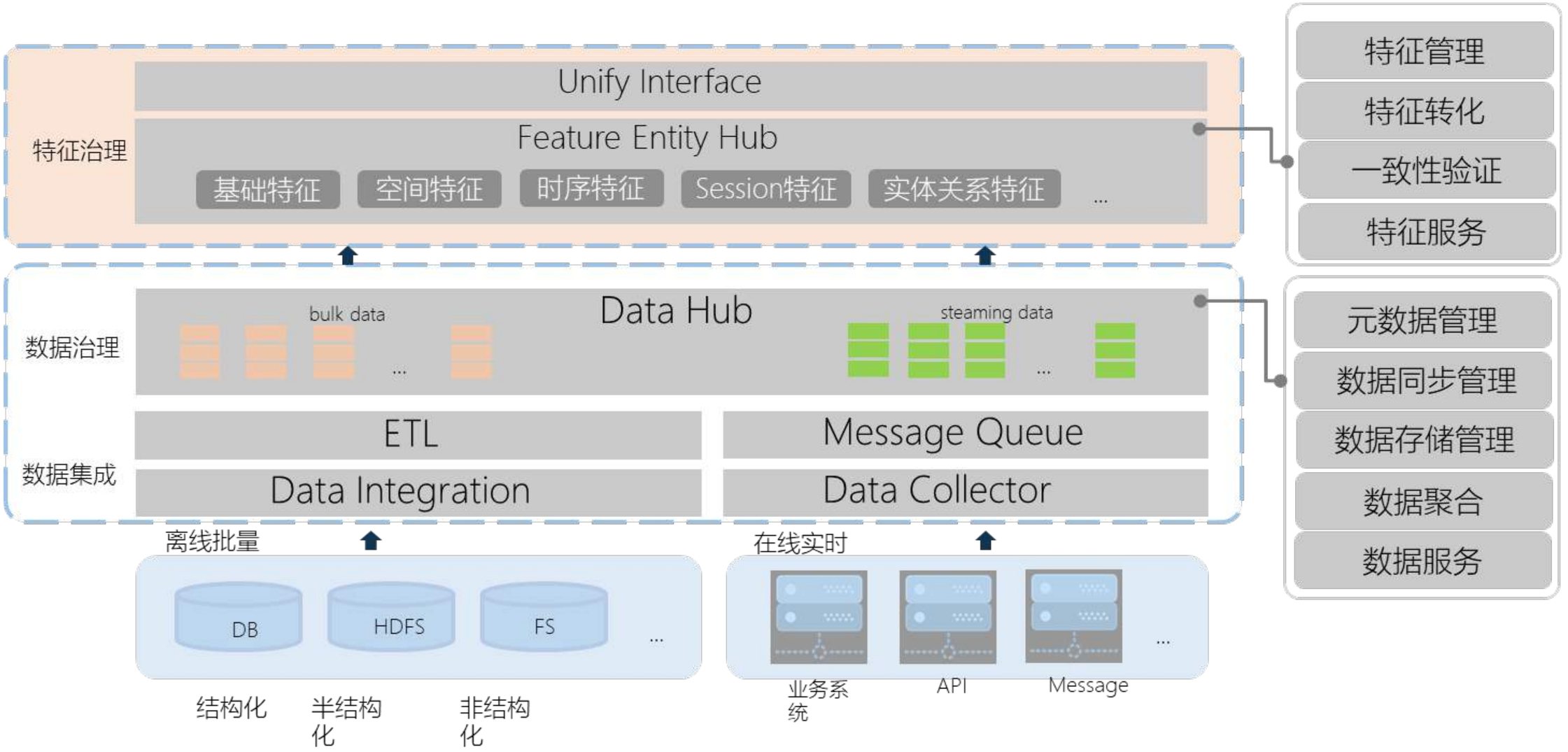
AI数据平台 – 蓄水池架构



AI数据平台 – 整体架构视图



高性能特征存储 - RtiDB

越来越多的AI场景需要实时时序特征

- 金融领域用户交易行为历史
- IoT设备的历史行为

....

(key, timestamp) => value

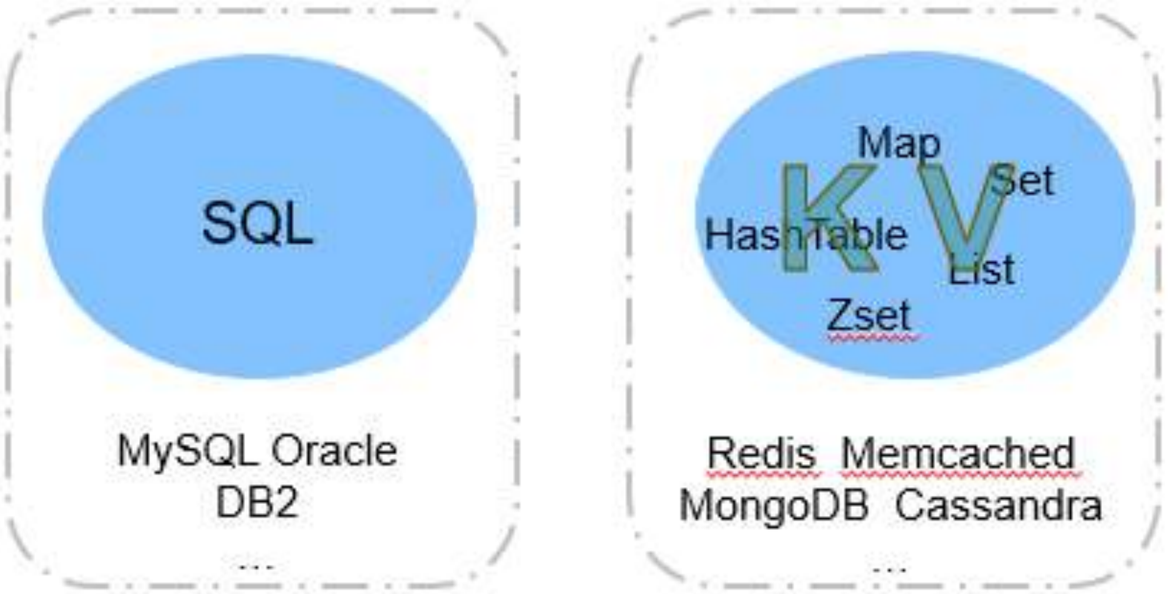
KEY

时序特征的生成需要按key抽取时序数据

- scan(key, start_time, end_time)

RTIDB

传统数据库并非为时序抽取而设计，性能无法满足需求



毫秒级海量高维时序特征抽取

NO!

现有的时序数据库(Time Series DB)并非为硬实时场景设计，也无法满足性能需求

VoltDB

InfluxDB

…

硬实时 = TP99 5ms

NO!

RTIDB：为硬实时而生

高性能

- 纯c++实现，GC-free，核心代码高度优化
- 全内存
- 采用skiplist数据结构和无锁原子命令
- 读写隔离

高可用

- 主从架构
- 持久化（snapshot + binlog）与灾难恢复
- 自动failover机制(V1.3+)

分布式(V1.3+)

- 可通过增加节点实现扩容

灵活、多用途

- 支持表级存储（可创建不同的表实现key的隔离）
- 可以当k-v查询使用（get、put操作）
- 支持TTL（过期自动删除）
  - ✓按时间窗口TTL
  - ✓按记录条数TTL（V1.2+）
- 支持表内schema，支持字段存取（V1.2+）
- 支持多维度（V1.2+）
  - ✓可以按照多个字段进行时序查询
  - ✓优化的存储结构，零数据冗余

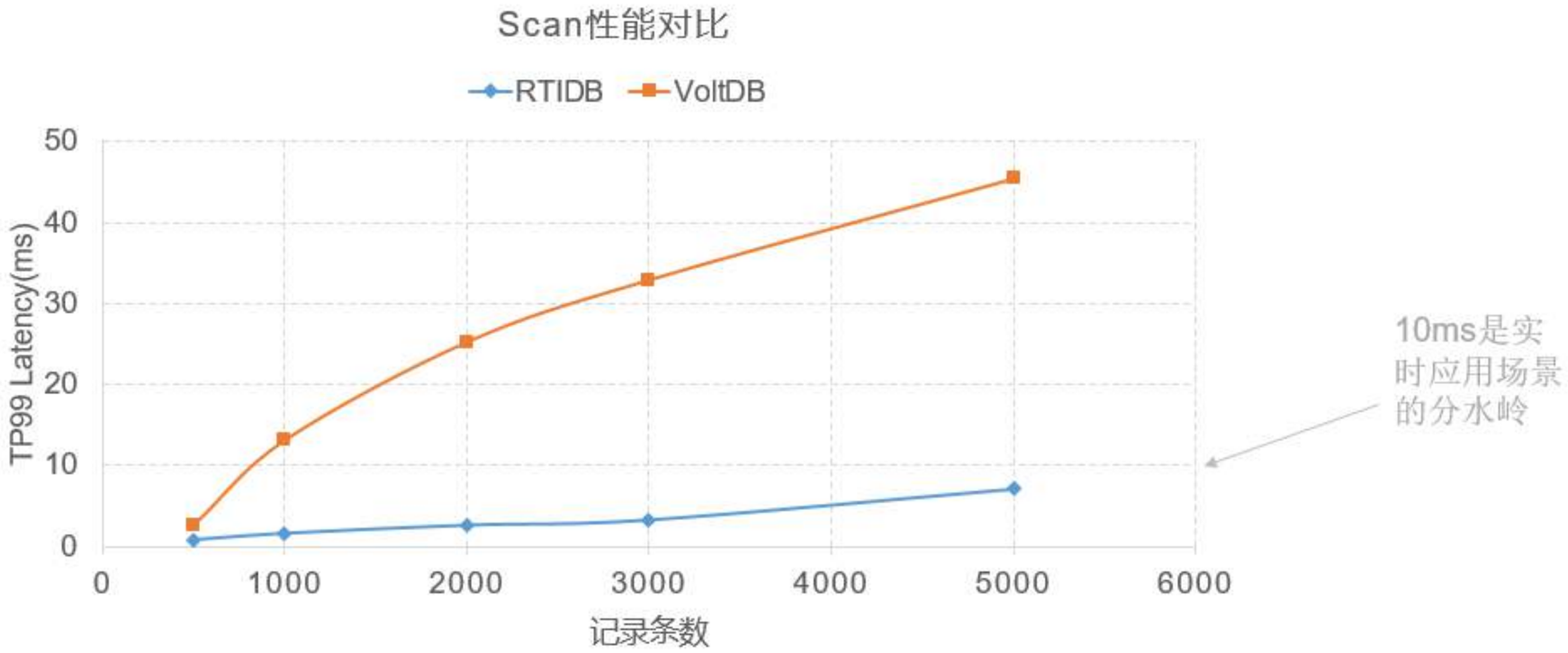
性能对比 RTIDB VS VoltDB

VoltDB是一款流行的内存数据库，具有高性能延时低等优点，有时用于时序类数据的场景中。

内存对比	占用内存	数据量(条)	存储结构
VoltDB	251GB	30,661,433	每条数据16个字段，所有字段值累加之后为256B
RTIDB	22GB	31,169,350	每条数据的value为256B

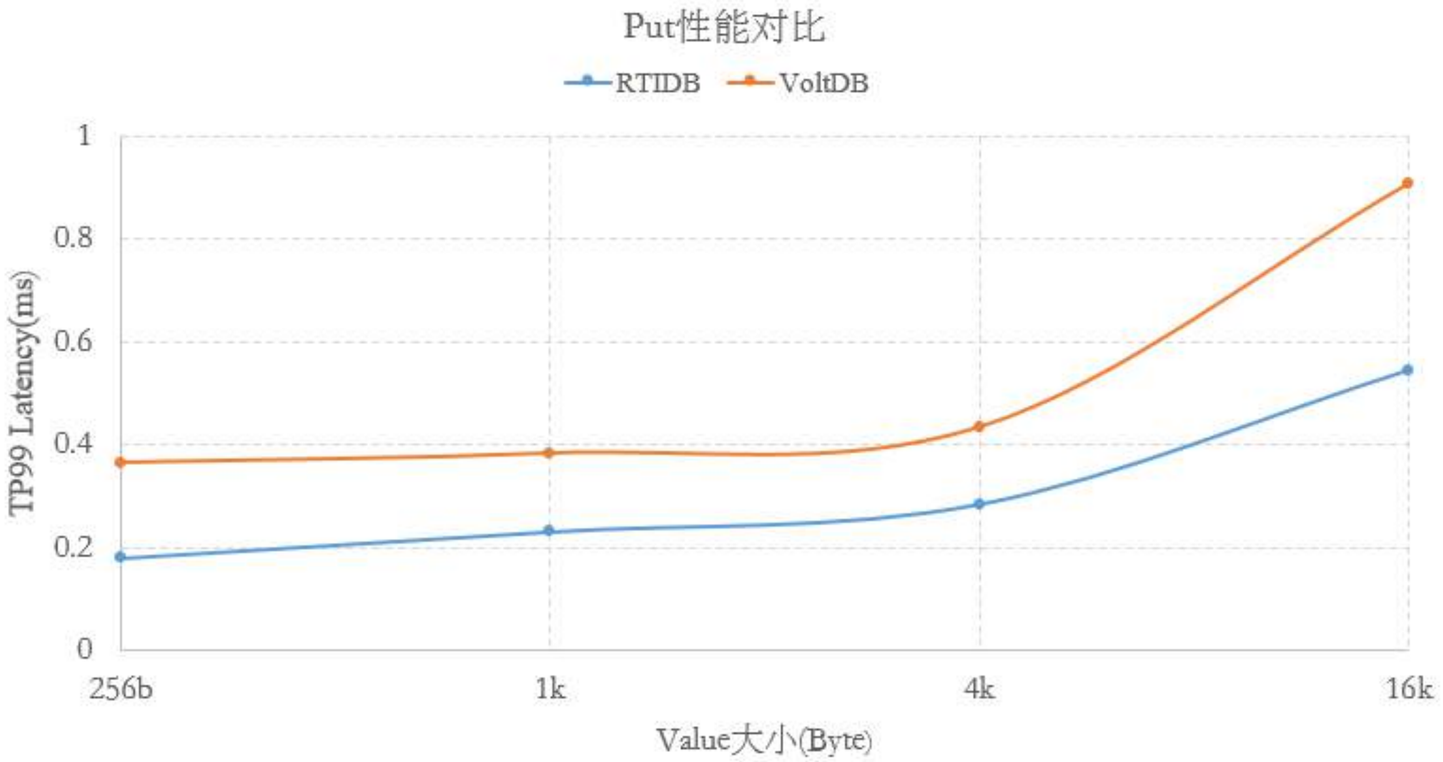
存储等量的数据，RTIDB对内存的消耗约为VoltDB的 10%

性能对比 RTIDB VS VoltDB



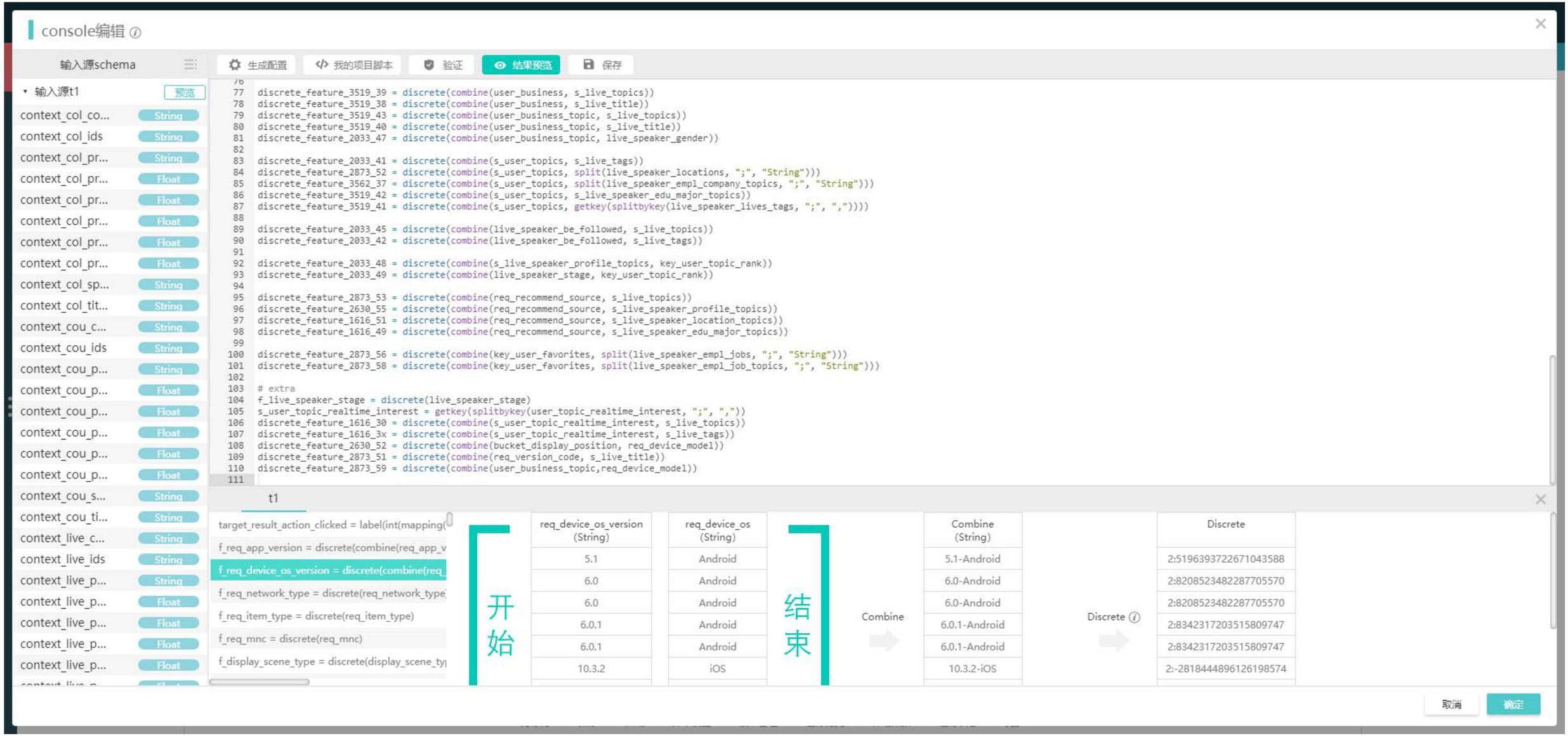
RTIDB的时序抽取耗时性能约约为VoltDB的 5~10X

性能对比 RTIDB VS VoltDB



RTIDB的灌入(put)性能约为VoltDB的 2X

高维特征抽取引擎 – Feature Extractor



高维特征抽取引擎 – Feature Extractor

```
10  f_user_id = discrete(user_id)
```

离散编码

t1

target_result_action_clicked = label(int(mapping(	None	Discrete ⓘ	5:3128268333250715758
f_req_app_version = discrete(combine(req_app_v	None		5:3128268333250715758
f_req_device_os_version = discrete(combine(req_	None		5:3128268333250715758
f_req_network_type = discrete(req_network_type	07		5:-7963361378781268083
f_req_item_type = discrete(req_item_type)	02		5:-7067486826460585577
f_req_mnc = discrete(req_mnc)	None		5:3128268333250715758
f_display_scene_type = discrete(display_scene_ty	07		5:-7963361378781268083
	None		5:3128268333250715758

高维特征抽取引擎 – 特征组合

```
70
71  discrete_feature_1616_44 = discrete(combine(user_gender, s_live_topics))
72  discrete_feature_1616_46 = discrete(combine(user_gender, s_live_title))
```

t1

```
t_context_cou_prices_aver = continuous(context_
f_context_live_scores_aver = continuous(context_
discrete_context_tags_com = discrete(combine(s_
discrete_context_titles_com = discrete(combine(s
discrete_context_topics_com = discrete(combine
discrete_feature_1616_44 = discrete(combine(use
discrete_feature_1616_46 = discrete(combine(use
discrete_feature_1616_43 = discrete(combine(use
```

开始

user_gender (String)	s_live_topics (String)						
None	5482	478	922	7085	100613		
None	880	2537	232	68882	61750	800	10110
1	880	2537	232	68882	61750	800	10110
1	1115	26256	1121	47177	47782	26257	21989
None	99	30960	131787	1354	20726	14455	707
None	4287	25992	30924	33814			

高维特征抽取引擎 – 特征组合

Combine (String)								
None-5482	None-478	None-922	None-7085	None-100613				
None-880	None-2537	None-232	None-68882	None-61750	None-800	None-10110	None-5425	Nc
1-880	1-2537	1-232	1-68882	1-61750	1-800	1-10110	1-5425	
1-1115	1-26256	1-1121	1-47177	1-47782	1-26257	1-21989	1-28041	
None-99	None-30960	None-131787	None-1354	None-20726	None-14455	None-707		
None-4287	None-25992	None-30924	None-33814					

Discrete						
36:-723505695647947440	36:-6246060836289859215	36:-6120979238635011810	36:-5379817706696544399	36:-6328783138438860660		
36:5403487815155201766	36:4336146748246454225	36:4154763130572795127	36:3175899881697131549	36:-1279973750079827986	36:2406957574578818269	36:26
36:-524536395708680644	36:-7187923850190159953	36:5629875406353654	36:1562490522654724790	36:30599300676171133330	36:1183744966692707500	36:-63
36:-4203365330514059597	36:378699440253556061	36:4592193145375254851	36:7723842825800924317	36:-7049670865585042441	36:8644568718213528210	36:-43
36:539448191356196093	36:-4419407291765645488	36:1853413784846976363	36:-5259487958040925882	36:-3304133625097634750	36:-4031746924900441513	36:-52
36:7715882958414768928	36:6504154887891798317	36:1605331534877383257	36:7935331807494662615			

基于AI的数据平台架构优势

- 确保线下调研所用数据线上可获取
- 特征脚本自动编译，模型上线无需另行开发，线上线下天然一致
- 基于计算图的底层执行效率优化，减少中间步骤和空间占用
- 高性能存储与特征计算组件带来响应时间和实时性提升
- 支持版本管理和团队协作，沉淀知识资产

企业AI核心系统——算法平台特性

AI算法能力有以下几种不同类型，算法平台需要对以下三种算法提供统一的管理和支撑

1

统一算法

- 面向常识性公共领域的算法，如人脸识别、语音识别、地址验真
- 对社会通用数据要求较高，可以引入外部能力如领域知识图谱
- 企业可以在通用模型的基础上，通过自身数据与限定范围优化提升效果

2

Model On Demand

- 面向具体业务，根据不同的反馈数据构建模型
- 必须通过企业自身的过程与反馈数据加以建设
- 高维、闭环、实时、低门槛

3

支撑算法

- 不直接提供业务决策
- 为业务决策的AI应用提供特征辅助
- NLP算法、图特征挖掘算法

Model On Demand的定义标准

L3-AssistML

假定给系统输入足够的 *目标-反馈数据* 和 *信息-过程数据*，建模专家可以无需编程，就创造业务上拥有效果的机器学习模型。

L4-AutoML

假定给系统输入足够的 *目标-反馈数据* 和 *信息-过程数据*，无需建模专家，对数据有常识性理解的业务人员就可以创造业务上拥有效果的机器学习模型。

L3-AssistML

建模全流程Web化

数据导入 -> 特征工程 -> 模型训练 -> 模型评估

建模全过程辅助

模型可视化 | 模型DEBUG | 特征DEBUG | 特征重要性分析

模型训练过程辅助 – 自动调参

L4-AutoML

分析传统建模中的关键步骤以及如何避免

- 特征选择 – 数据量足够，可以通过构建高维模型和算法，让机器自动选择有效特征

- 特征变换 – 传统树/NN模型需要根据业务经验将稀疏离散变量转化为连续变量
- 对策1：采用支持高维离散变量的模型
- 对策2：采用自动变量连续化算法降低特征变换成本

- 高级特征工程 – 事实证明，寻找有效的组合/时序/分桶特征是提升模型效果的关键
- 对策1：FeatureGo – 自动搜索有效特征组合和特征分桶
- 对策2：AutoTFE（滚动训练、LSTM） - 自动搜索有效的时序特征

- 数据拼接 – 有了“宽表”就成功了一半，但是宽表应当如何拼接成功
- 探索中 …

逻辑回归模型可视化

特征数: 27　　总特征维度: 137

复制访问链接

搜索

序号	特征名	箱线图
1	f_col_1	
2	f_col_2	
3	f_col_3	
4	f_col_4	
5	f_col_5	
6	f_col_6	
7	f_col_7	
8	f_col_8	
9	f_col_9	
10	f_col_10	
11	f_col_11	
12	f_col_12	
13	f_col_13	

-0.6　-0.4　-0.2　0　0.2　0.4

f_col_4	最小值	后25%	中位数	前25%	最大值	平均值	方差
权重	-0.23	0.07	0.10	0.19	0.27	0.11	0.03
特征值	unknown	high.school	university.degree	professiona .course	basic.4y		

f_col_4

特征维度：6

1.0000　有效特征占比

0.0438　特征维度/总特征维度

权重前500

下载全部

特征值	权重
basic.4y	0.2[illegible]367
basic.9y	0.[illegible]510
professional.course	0.187205
university.degree	0.098464
high.school	0.073893
unknown	-0.232505

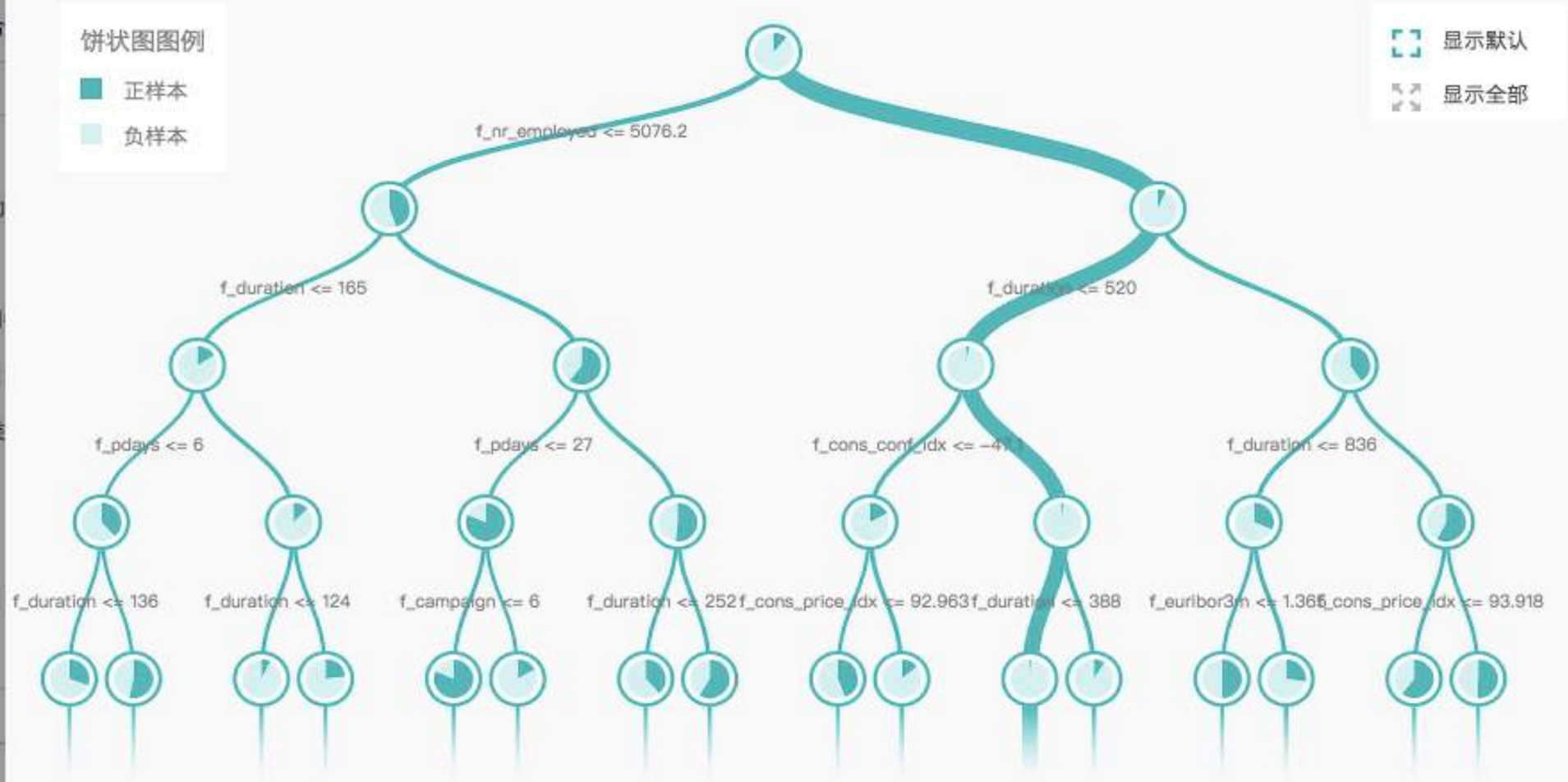
页码 1 / 1

GBDT模型可视化

复制访问链接

选择树图: < 1 >

线的含义: 总样本数 正样本数 负样本数



决策名称: f_nr_employed

缺省值路径: right

缺省值比例: 0.000000%

增益: 2493.74105063353

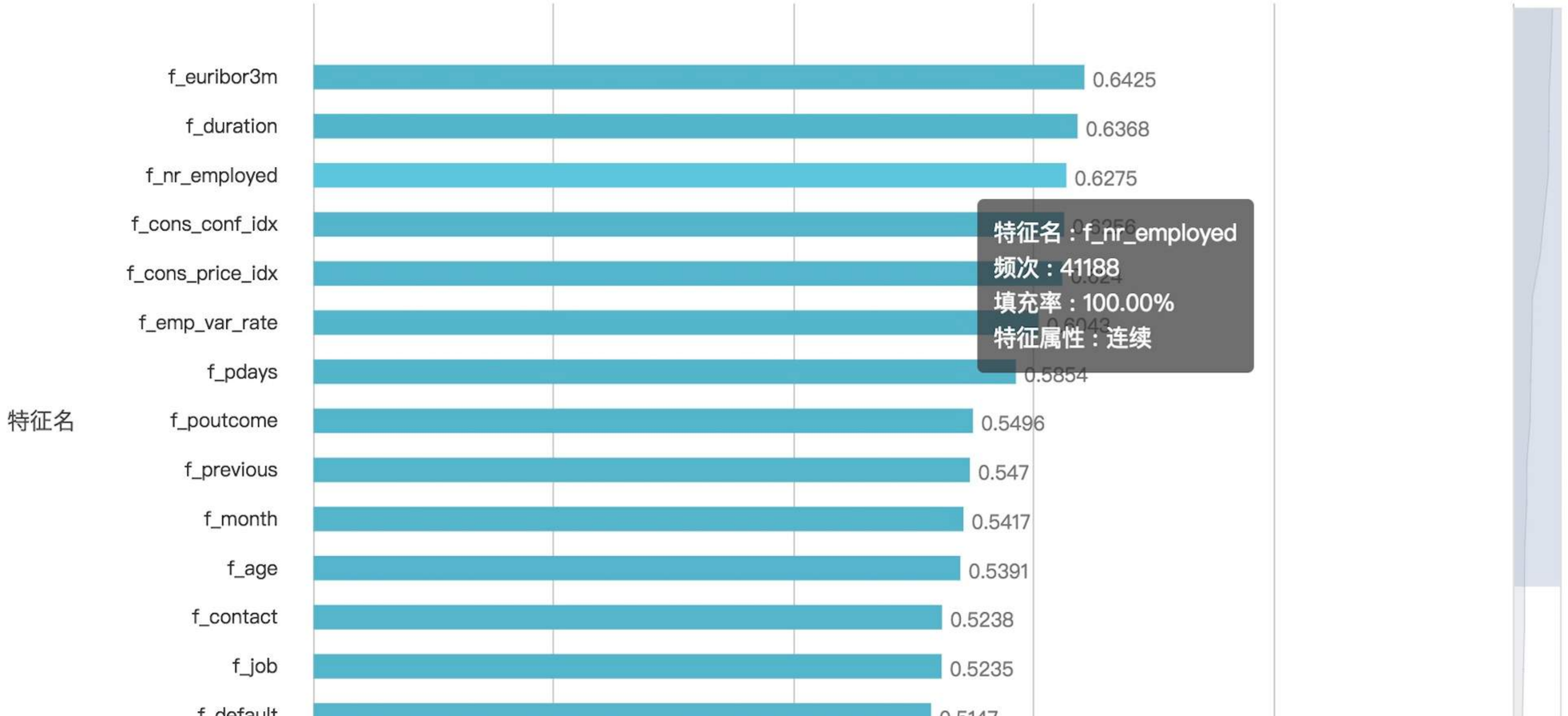
权重: -1.54937237477845

决策路径: 复制

特征重要性分析报告

特征名 支持模糊筛选特征字段 特征重要性系数范围 0.5015 — 0.6425 筛选 复制图标特征

特征数 823760 样本数 41188



特征组合分析报告

特征数 58

显示统计评分后的输入特征配置

特征池序	特征抽取配置	slotwise_auc
0	discrete_feature_440_0=discrete(duration) # duration	0.782146
0	discrete_feature_440_1=discrete(euribor3m) # euribor3m	0.778596
0	discrete_feature_440_2=discrete(cons_conf_idx) # cons_conf_idx	0.770638
0	discrete_feature_440_3=discrete(cons_price_idx) # cons_price_idx	0.770638
0	discrete_feature_440_4=discrete(nr_employed) # nr_employed	0.767778

特征池序	特征抽取配置	slotwise_auc
1	discrete_feature_440_23=discrete(combine(duration,pdays)) # duration pdays	0.859792
1	discrete_feature_440_24=discrete(combine(duration,date1)) # duration date1	0.859479
1	discrete_feature_440_25=discrete(combine(duration,date2)) # duration date2	
1	discrete_feature_440_26=discrete(combine(duration,date3)) # duration date3	0.859474
1	discrete_feature_440_27=discrete(combine(duration,previous)) # duration previous	0.856762
2	discrete_feature_440_29=discrete(combine(duration,pdays,date2)) # duration date2 pdays	0.901664
2	discrete_feature_440_28=discrete(combine(duration,pdays,date3)) # duration date3 pdays	0.901663
2	discrete_feature_440_30=discrete(combine(duration,pdays,date1)) # duration date1 pdays	0.901656
2	discrete_feature_440_31=discrete(combine(duration,date2,date3)) # duration date3 date2	0.901445
2	discrete_feature_440_32=discrete(combine(duration,date1,date3)) # duration date3 date1	0.901394
3	discrete_feature_440_33=discrete(combine(duration,pdays,date1,date2,date3)) # duration date3 date2 date1 pdays	0.909985
3	discrete_feature_440_36=discrete(combine(duration,pdays,date1,date2)) # duration date2 date1 pdays	0.909985
3	discrete_feature_440_34=discrete(combine(duration,pdays,date2,date3)) # duration date3 date2 pdays	0.909984

discrete_feature_440_24=discrete(combine(duration,date1)) # duration date1

高维算法+AutoML支撑下的业务表现

银行交易实时反欺诈

- 历史两年1亿条数据，8000万维度模型，世界上最好的反欺诈模型

个性化内容推荐

- 数十亿特征，基本完成L4-AutoML原型验证，基本达成甚至超出人工建模水平

“范式大学”计划

- 经过一个月培训即可基本完成常见营销/反欺诈/信用评分模型建设过程

算法平台的其他能力

基于知识图谱的特征增强

- 教育、位置、金融知识图谱，为地址、教育背景、公司名等字段扩展高维特征

NLP相关应用组合

- 提供切词、词性标注、实体识别、新词发现、词向量、情感分析等基础能力
- 将基础能力与高维机器学习能力结合，形成端到端文本分类、地址验真、情感分类等决策性应用

图像相关算法

- 算法平台出支持Model On Demand的相关算法外，流程还支持嵌入Tensorflow等开源机器学习框架算子，可以将图像特征提取/高维算法组合成为端到端决策性应用
- 基于算法平台支撑行业特定OCR识别、古迹识别等应用解决方案

企业AI核心系统——生产平台特性

- 将“开发时”的解决方案，变为“运行时”的应用服务
- 服务需要满足高可用、可监控、可扩展等企业级特性
- 针对AI特点，该应用应当拥有自我迭代的能力

机器智能构建的关键 – 自学习

自学习的变

- 输入数据的内容随业务发展和时间推移而变化
- 输入数据的规模随业务发展和时间推移而变化

自学习的不变

- 使用数据的方式通常在同一个模型方案的生命周期中不变
- 面向问题的目标通常保持不变

自学习过程的抽象

- 使用同样的算法、特征工程方案，通过持续输入数据和反馈，使模型能够自我迭代

自学习的关键技术

数据管理

- 在数据平台中，对样本数据进行Schema定义和管理，定义“数据组”的概念
- 数据需要有时间戳、版本

增量机器学习

- 机器学习训练算法服务化，支持流式数据接入和模型Dump

灾备

- 定期进行全量训练，生成基准
- 版本管理机制，维护基准和增量的关系，支持随时回滚

增量服务发布

- 支持对正在运行的服务进行不停机参数更新
- 服务支持随时快速回滚到历史状态

生产平台的其他特性

元素仓库与自学习DAG发布

- 元素仓库统一对所有资源进行版本、元数据管理，支援灵活二次开发
- DAG管理：在建模平台内设计的符合规范的DAG图，可以免开发直接发布为自学习流程

多租户与资源隔离：

- 支援企业内不同部门工作空间隔离与资源分配的需求
- 基于LDAP的身份认证API，以及针对Yarn/Kubernetes的资源隔离支持

高可用：

- Workflow：封装Yarn/K8S的调用过程，支援对自学习流程的调度、灾备功能
- 常驻内存服务（例如在线预估、RtiDB）基于K8s实现高可用
- Gateway：OpenResty

监控与报警：

- Grafana / ElasticSearch

企业级AI核心系统——第四范式·先知

数据核心

数据接入和处理

- 结构化数据接入
- 非结构化数据接入
- 批量数据处理
- 实时数据处理
- 图关系数据库
- RTiDB

数据管理与访问

- 元数据管理
- 统一数据访问
- 特征知识库

算法核心

模型调研

- 深度学习
- 高维特征生成
- 高维机器学习
- 时序特征生成
- NLP
- 特征增强
- 模型诊断评估
- 模型解释

智能建模

- AutoML
- 行业通用模板建模

生产核心

应用服务

- 高维模型计算服务
- 在线特征计算服务
- NLP模型服务
- 深度学习模型服务
- 知识图谱模型服务
- 服务灰度发布
- A/B Test
- 服务编排和管理
- 模型自学习
- 实时决策引擎

模型仓库

- 模型生命周期管理
- 外部模型兼容支持
- 模型仓库

系统资源管理

- 任务调度
- 服务调度
- 工作流引擎
- 资源隔离
- 权限管理
- 监控和报警
- 服务高可用
- 动态资源调度

企业级AI核心系统——第四范式·先知

数据管理产品

- 实时数据集成平台
- 元数据管理平台
- 非结构化数据平台
- 实时数据访问平台

数据核心

模型调研产品

- 专业版建模平台
- 智能版建模平台
- 资源调度与隔离
- 用户间共享与协作

算法核心

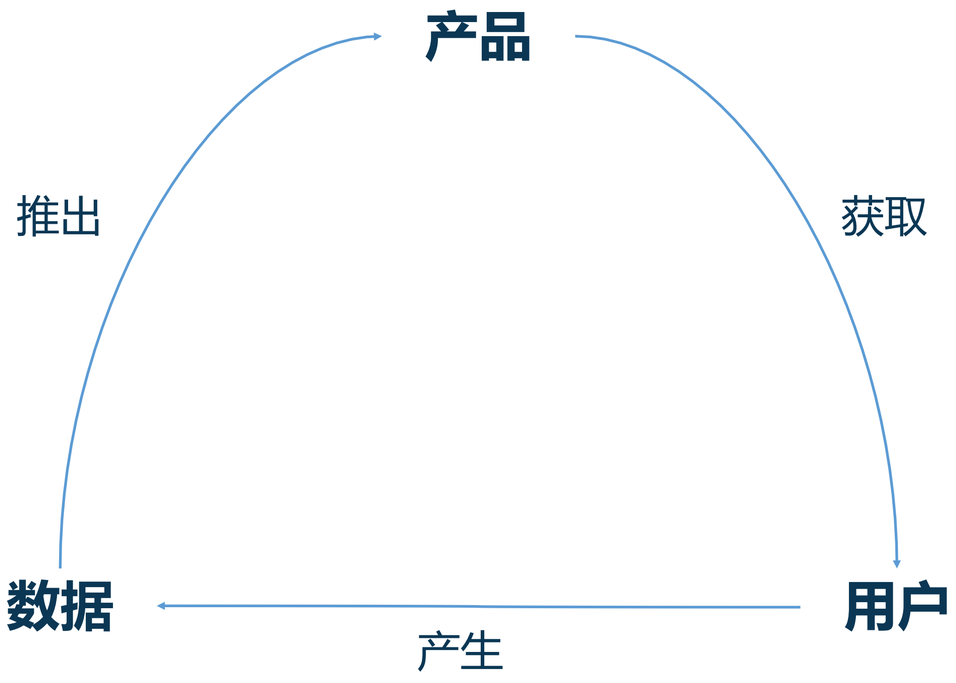
模型应用服务产品

- 服务管理平台
- 自学习平台
- 决策引擎
- 运维管理平台

生产核心

TABLE OF CONTENTES

总结：如何成为一个AI企业



业务能力

- 在组织内部建立数据驱动意识的能力
- 业务目标分解的能力
- 策略性地持续获取数据的能力

技术能力

- 数据能力

实时数据采集与获取、数据描述、非结构化数据管理等

- 算法能力

高维机器学习算法、特征工程与特征增强、智能建模等

- 应用能力

模型全生命周期管理、自学习、A/B Test、高维模型计算等

Thanks!