DATACAKE
streamline data & AI workflow

AI算法在大数据治理中的应用

李卫民

2022.12.08

个人介绍

李卫民

- 360大数据开发工程师
- 猫眼风控算法工程师
- 贝壳推荐算法工程师
- 茄子科技算法工程师

DATACAKE
streamline data & AI workflow

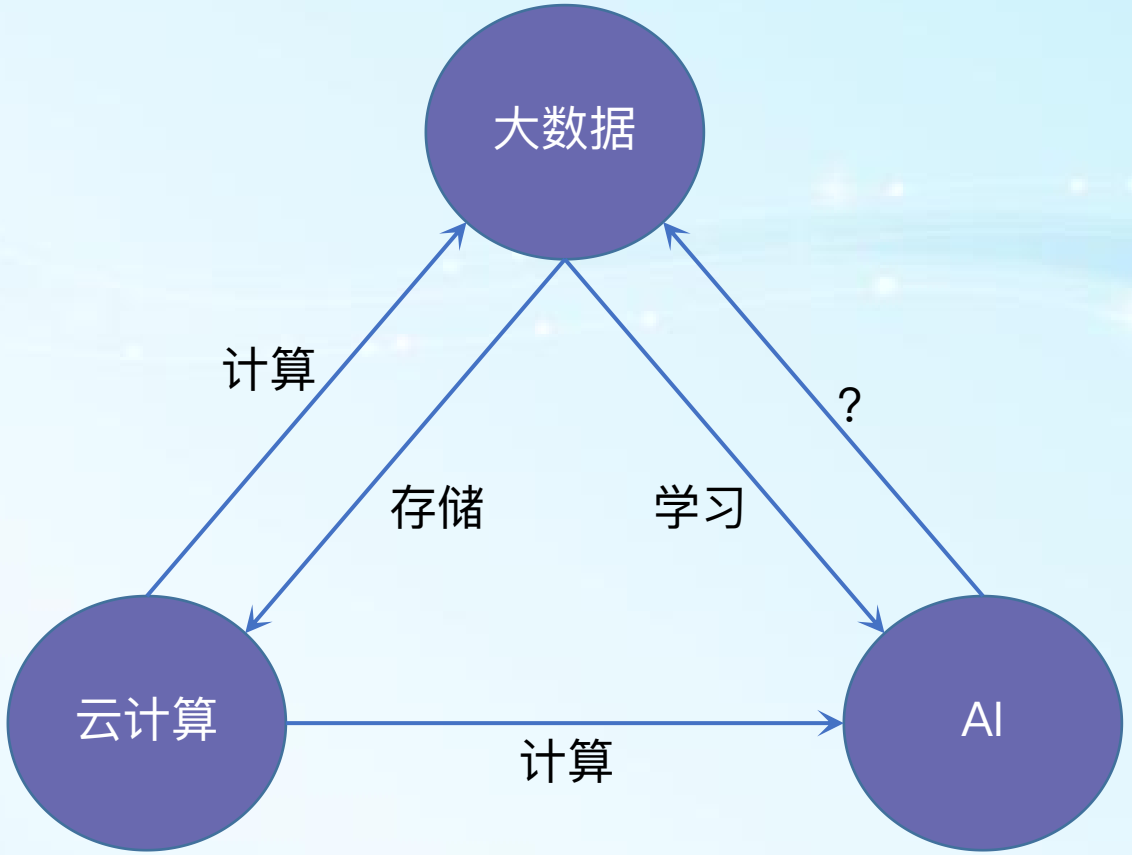
目录

DATACAKE
streamline data & AI workflow

1. 大数据与AI

大数据与AI的关系

- 大数据仅服务于人工智能吗?
- 人工智能也可以反哺大数据。



DATACAKE
streamline data & AI workflow

大数据全生命周期

周期阶段	数据采集	数据传输	数据存储	数据处理	数据交换	数据销毁
面临风险	采集质量 采集频率 加密脱敏	可用性 完整度 安全性	存储结构 占用资源 容错备份	资源消耗 处理效率 易操作性	共享 安全 边界	合规性 必要性 关联性
AI作用	日志评估 异常报警	故障诊断 入侵检测	任务评估 任务优化	资源优化 效率提升	联邦学习 智能打码	时机判断 关联分析

- 数据生命周期管理的目标通常有以下几点：高效、低成本、安全。
- 依赖规则、策略、专家经验？AI算法！

DATACAKE
streamline data & AI workflow

2. 大数据任务健康度评估

DATACAKE
streamline data & AI workflow

背景

- 数据开发任务owner不了解任务健康度；
- 云资源消耗成本黑盒，治理方向不明确；
- 任务治理门槛及效果无量化标准不同；
- 任务的体量及资源消耗存在天然差别，评估指标规则制定困难。

需求

- 根据任务资源用量等执行情况，给出量化评估结果。

功能模块方案

- 对已上线任务依据历史执行情况给出健康度评分；
- 突出显示任务主要问题，引导owner进行任务治理。

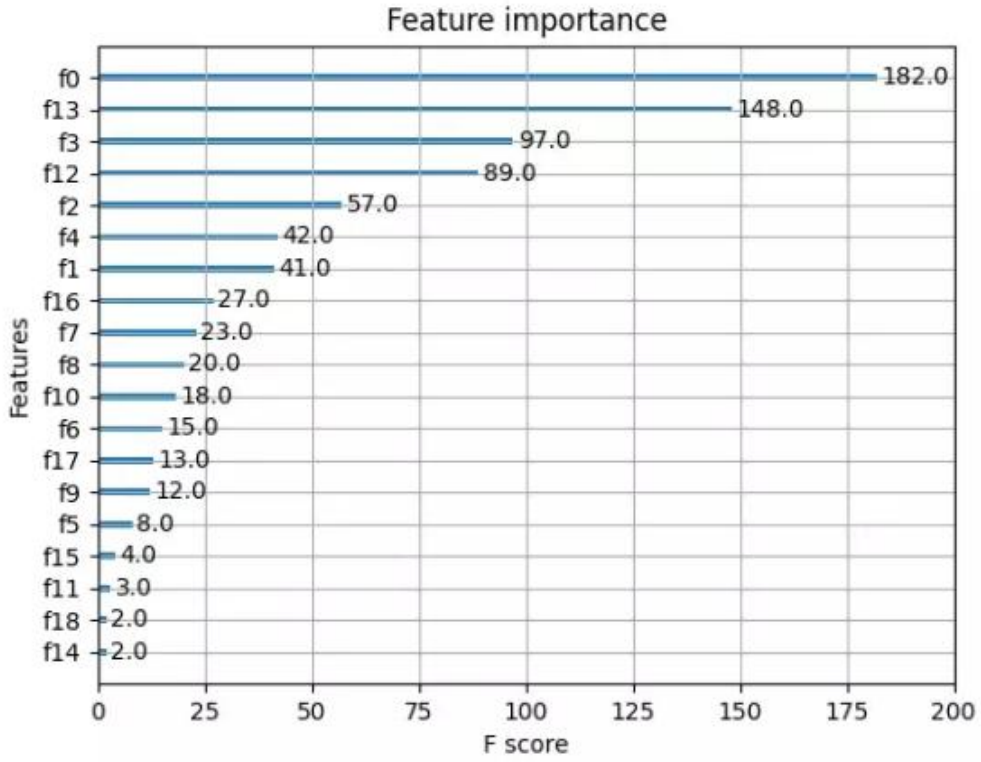
任务名称	评分	排名	负责人	部门	前日总用量	最近实例用量	cpu 平均利用率	cpu 最大利用率	cpu 最小利用率	操作
dws_san_sdk_ad_... 数据倾斜	32.6317 任务评分中等	953	renhantao	广告分析	147	147	68.07% CPU使用率中等	78.28%	4.09%	提醒
DA_chenchen_del... 数据倾斜	72.7963 任务评分高	71	wangxiaoying	广告中心	56.41	56.41	76.35% CPU使用率中等	99.30%	2.27%	提醒
DA_renhantao_d... 数据倾斜	31.1054 任务评分中等	1001	renhantao	广告分析	442.13	442.13	64.39% CPU使用率中等	88.53%	2.52%	提醒
DA_chenchen_CH... 数据倾斜	74.0243 任务评分高	55	chenchen	广告分析	48.93	48.93	71.11% CPU使用率中等	85.72%	3.86%	提醒
DA_caowei_dws_... 数据倾斜	64.5665 任务评分高	173	caowei	广告分析	49.97	49.97	69.61% CPU使用率中等	83.74%	1.88%	提醒
ab_extract_bayle_...	31.3219 任务评分中等	992	yangping	核心服务	202.41	202.41	70.70% CPU使用率中等	97.47%	3.89%	提醒
ab_gen_shareit_w...	57.7354 任务评分中等	248	yangping	核心服务	97.22	97.22	48.37% CPU使用率中等	84.45%	12.10%	提醒
ab_shareit_ch_ret...	69.172 任务评分高	112	yangping	核心服务	13.76	13.76	69.92% CPU使用率中等	97.89%	9.87%	提醒
ab_shareit_ch_ret...	74.9151 任务评分高	48	yangping	核心服务	47.83	47.83	103.56% CPU使用率高	137.43%	20.06%	提醒
ab_shareit_ch_ret...	68.3342 任务评分高	119	yangping	核心服务	24.59	24.59	57.91% CPU使用率中等	97.30%	4.27%	提醒

模型训练过程

- **整体方案：**使用xgBoost模型计算任务为“好任务”的概率，进一步转化为量化评分；
- **样本：**约10万历史任务执行数据，按照二八比例分成数据集与测试集；标签由资源用量、人工评估、部分规则等综合生成；
- **特征：**任务历史执行的原始数据及部分复合、统计特征，如任务处理的数据量、执行失败次数、资源利用率等。

模型效果评估

- 模型筛选出19个决定一个任务是否为“好任务”的关键特征。

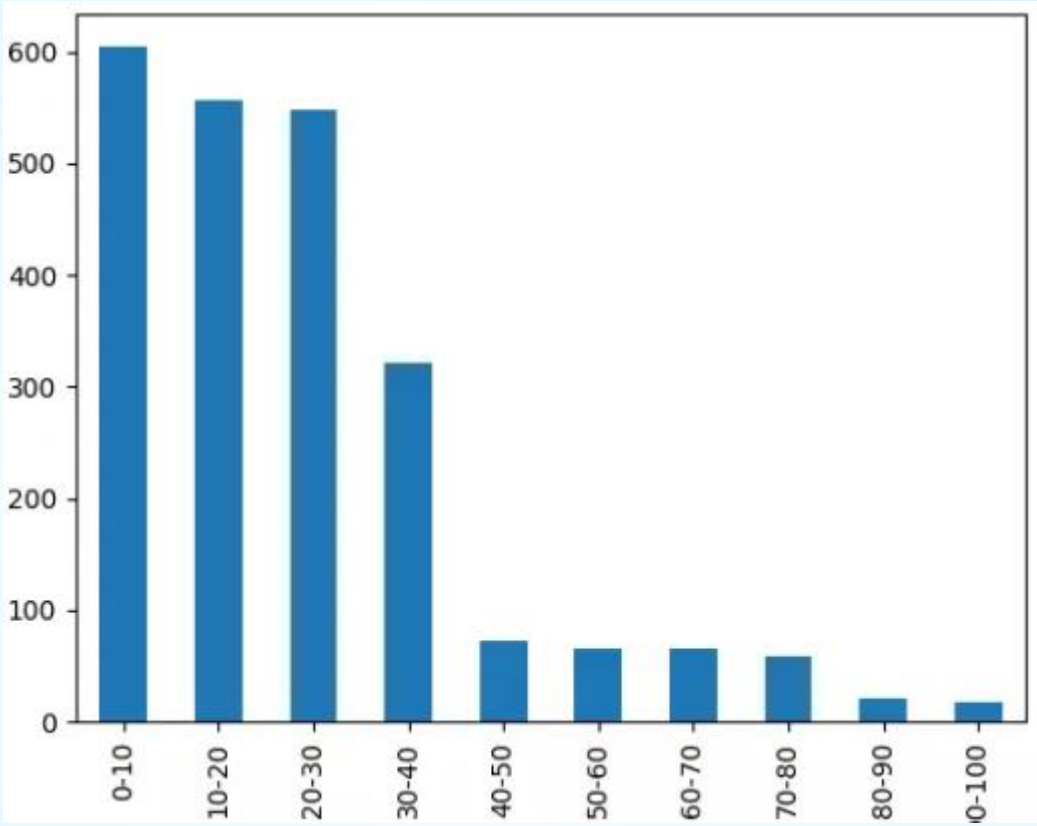


f0	sumNumFailedTasks	f10	spark.sql.files.maxPartitionBytes
f1	avgShuffleWriteBytes	f11	spark.executor.cores
f2	sumNumCompleteTasks	f12	avgMemoryUsed
f3	avgNumCompleteTasks	f13	avgCpuUsed
f4	avgDuration	f14	runCount
f5	failedStageCounts	f15	totalExecutorZeroTaskcount
f6	allStageCounts	f16	totalExecutorDurationTime
f7	job-duration	f17	totalExecutorCount
f8	spark.executor.memory	f18	spark.sql.autoBroadcastJoinThreshold
f9	spark.driver.memory		

模型效果评估

- 对历史任务评估分数分布如下，符合直观评估标准。

	score
count	2725.000000
mean	26.558874
std	20.230616
min	1.018567
25%	11.708370
50%	22.611744
75%	32.401488
max	99.861675



收益

- 任务owner对名下任务的健康情况做到心中有数；
- 为后续开展任务治理提供依据；
- 任务治理效果量化展示。

DATACAKE
streamline data & AI workflow

3. Spark任务智能调参

DATACAKE
streamline data & AI workflow

背景

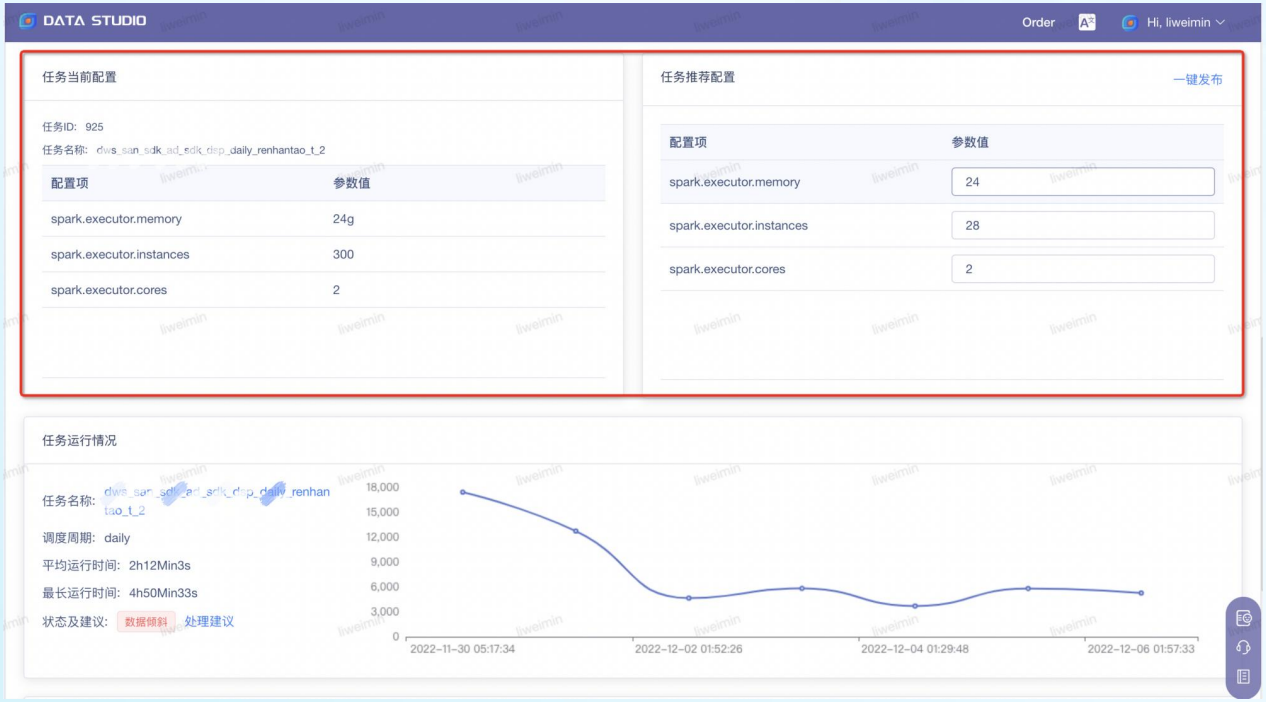
- Gartner的调研显示，云客户消耗的70%的云资源都存在各种不必要的浪费；
- 任务运行参数的配置可以直接影响资源的消耗，但参数配置项多而复杂；
- 依靠专家经验制定的参数配置规则覆盖范围有限，且无法动态调整。

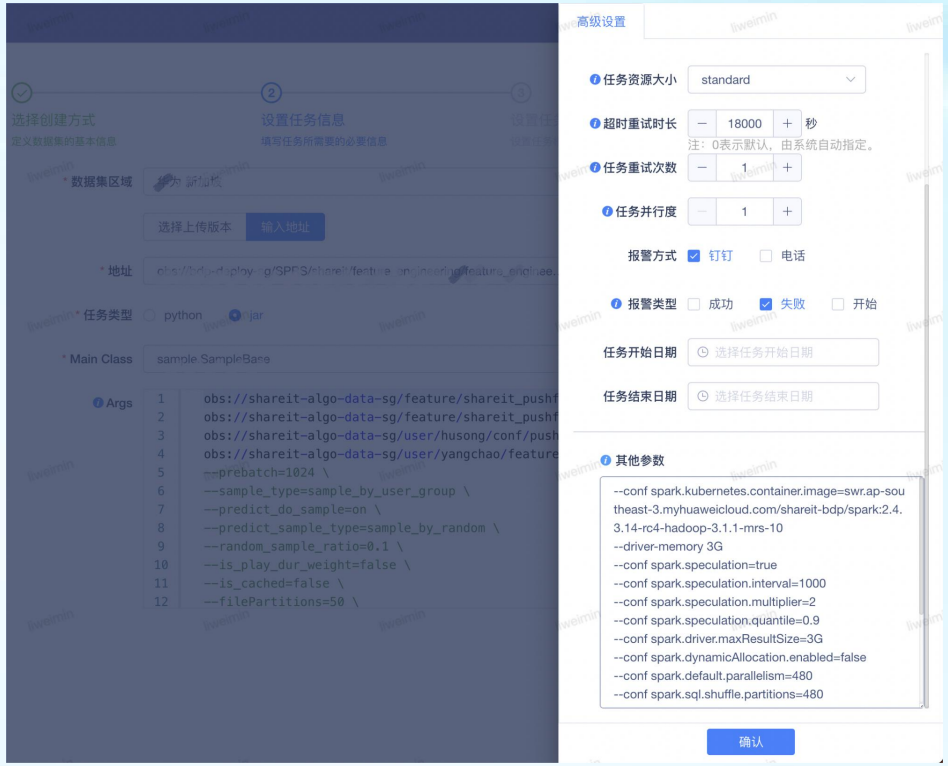
需求

- 模型智能推荐任务运行的参数配置，使得在保持任务原有运行时间不变长的前提下，提高云资源CPU和内存的使用率。

功能模块方案

- 对已上线任务依据历史执行情况进行参数配置推荐；
- 对新创建任务依据语法解析进行参数配置推荐。





模型训练过程

- **模型输出/参数空间：**Spark任务可配置参数多达三百余项，选择了三项对性能影响最大的参数作为调优对象。

参数	含义	类型	默认值	调优范围
spark.executor.cores	每个executor可使用的CPU核数	资源参数	1	1–用户设置的上限（默认上限6）
spark.executor.memory	每个executor可使用的内存总量	资源参数	1G	1g–用户设置的上限（默认上限60g）
spark.executor.instances	executor的实例个数	资源参数	1	1–用户设置的上限（默认上限60）

模型训练过程

方案一：LR模型学习规则调参

- **整体方案：**学习已有的经验调参规则；
- **样本：**规则引擎计算出的7w余条任务配置；
- **特征：**Spark引擎记录的历史任务执行情况，包括任务数据量、资源使用量、任务耗时等直接特征，以及过去七日平均耗量、最大耗量等部分统计特征；
- **模型训练：**采用多因变量的多元回归模型，标签为三个参数的配置值，训练最终输出model模型文件。

上述数学模型写成矩阵形式如下：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \beta_{01} & \beta_{11} & \cdots & \beta_{m1} \\ \beta_{02} & \beta_{12} & \cdots & \beta_{m2} \\ \vdots & \vdots & & \vdots \\ \beta_{0p} & \beta_{1p} & \cdots & \beta_{mp} \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

统计问题就是从已知的m个自变量，p个因变量的n组实测数据出发，求未知常数β_{ij}的估计值$\hat{\beta}_{ij}$，并对误差ε_j作出估计和推断。

和一元统计分析一样，将略去误差项而得到的关系式：

$$\hat{y}_j = \hat{\beta}_{0j} + \hat{\beta}_{1j}x_1 + \cdots \hat{\beta}_{mj}x_m \qquad j = 1, \cdots, p$$

称为回归方程，称$\hat{\beta}_{ij}(i = 1, \cdots, m;\ j = 1, \cdots, p)$为回归系数，称$\hat{\beta}_{0j}(j = 1, \cdots, p)$为常数项。

和一个因变量的多元回归分析一样，这里用最小二乘法求β的估计，选择$\hat{\beta}_{ij}$的值使误差阵ε各元素平方和相加达到最小，即使

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{p}\varepsilon_{ij}^2 \text{ 最小}$$

模型训练过程

方案二：贝叶斯优化

- **整体方案：**使用贝叶斯框架以尽量少的配置验证次数寻找较优位置；
- **训练过程：**在参数空间中采样一个配置，对采样的配置进行验证，即新建一个该配置对应的任务然后执行，获取执行指标等结果，然后使用获得的结果更新算法。以上步骤重复执行，直到调优完成；
- **模型应用：**完成参数调优后，对于相同的任务或者相似的spark任务，直接配置已找到的最好的配置进行执行。

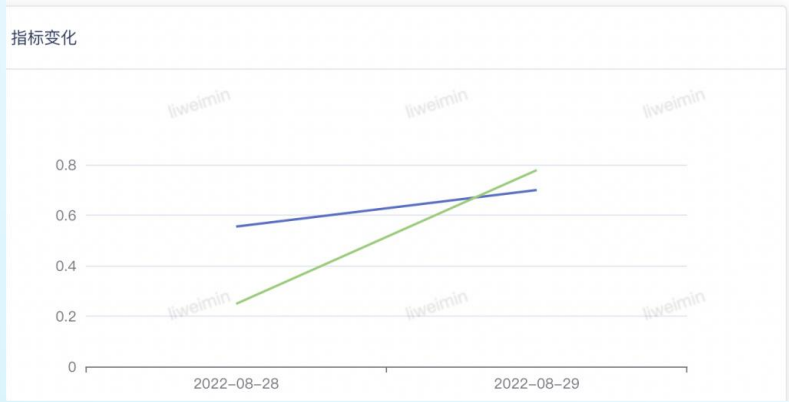
模型效果评估/收益

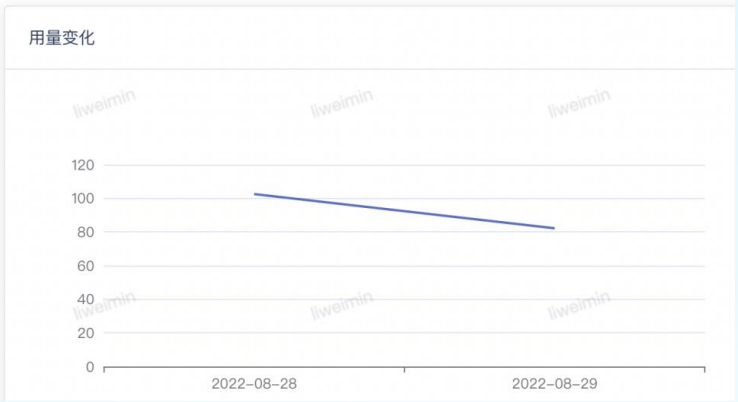
- 对于已有任务，按照模型推荐参数修改任务配置后，80%以上的任务实现约15%的资源利用率提升，部分任务资源使用率翻番。

治理记录

任务ID: 11062 任务名称：DA_renhantao_dm_ads_funnel_revenue_sdk_ads_funnel 负责人：renhantao

治理策略	治理后配置	CPU使用率	MEM使用率	最近实例用量	治理时间	执行时间	发布状态	操作
推荐配置	--conf spark.executor.cores=1 --conf spark.executor.memory…	0.6993006434	0.7782203415	82.17	2022-08-29 14:14:14	2022-08-29	已完成	版本恢复
原始基准	--conf spark.kubernetes.executor.node.selector.lifecycle=On…	0.5546318667	0.2484933488	102.5	2022-08-29 14:14:12	2022-08-28	已完成	版本恢复





模型优化方向

当前模型还存在如下问题

- 学习规则的回归模型局限性较强、全局寻优的贝叶斯优化模型成本较高；
- Spark应用程序拥有多种代码结构和语义，代码特征显著影响Spark性能和配置选择 ——> 语义分析；
- Spark支持各种分析应用，调优系统需要适应不同的应用 ——> 分类调优；
- 实验样本少，测试成本较高，效果评估困难 ——> 结合工程实现线上调优。

DATACAKE
streamline data & AI workflow

4. SQL查询任务引擎智能选择

DATACAKE
streamline data & AI workflow

背景

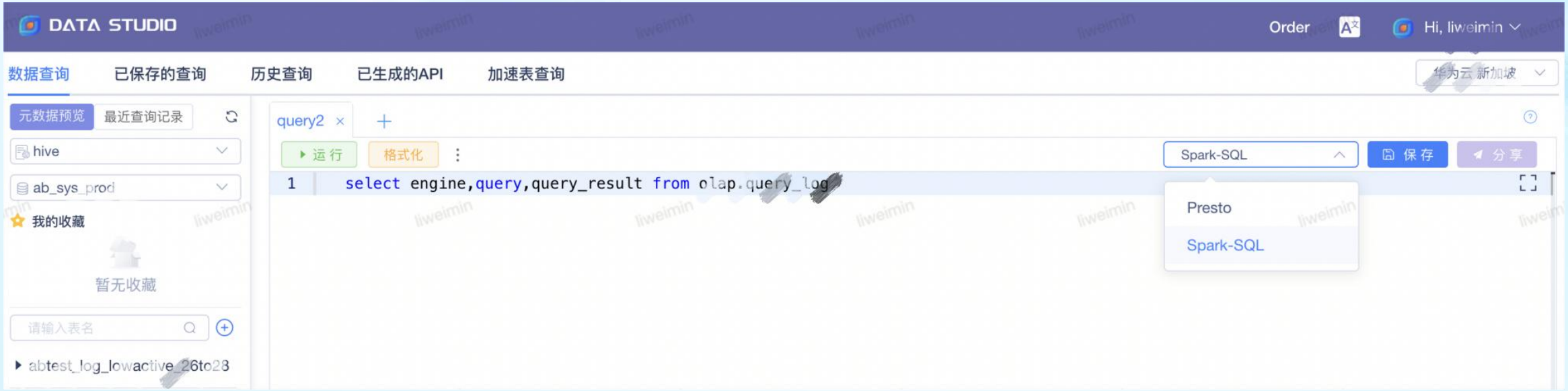
- SQL查询平台是BI、RD、PM使用最多、体验最明显的大数据产品；
- 不同的引擎适合执行不同SQL查询任务：Presto纯内存计算，在简单查询场景下执行速度较快；Spark不易出现OOM，适合大数据量的复杂计算场景；
- SQL查询的效果要综合考虑任务执行时间与资源消耗；
- 传统引擎选择方式有RBO（Rule-Based Optimizer，基于规则的优化器）、CBO、HBO等方式，各自存在不同的缺点。

需求

- 模型智能推荐SQL执行引擎，在保持任务原有运行时间不变长的前提下，降低资源消耗。

功能模块方案

- 用户可以自主选择SQL任务执行引擎，亦可使用模型推荐的执行引擎。



模型训练过程

- **整体方案：**基于SQL语句本身推荐执行引擎；
- **样本标签：**数据量大的、执行时间长的定义为适合在Spark上执行的任务；
- **特征提取：**使用自然语言处理中的n-gram TF-IDF方法，将SQL文本转化为数值特征；
- **特征选择：**使用LR模型计算特征重要性，筛选特征；
- **模型训练：**样本标签为使用Spark或Presto引擎，标签由任务实际执行情况及人为倾向引导给出。使用XGBoost梯度提升树算法训练模型。

模型效果评估

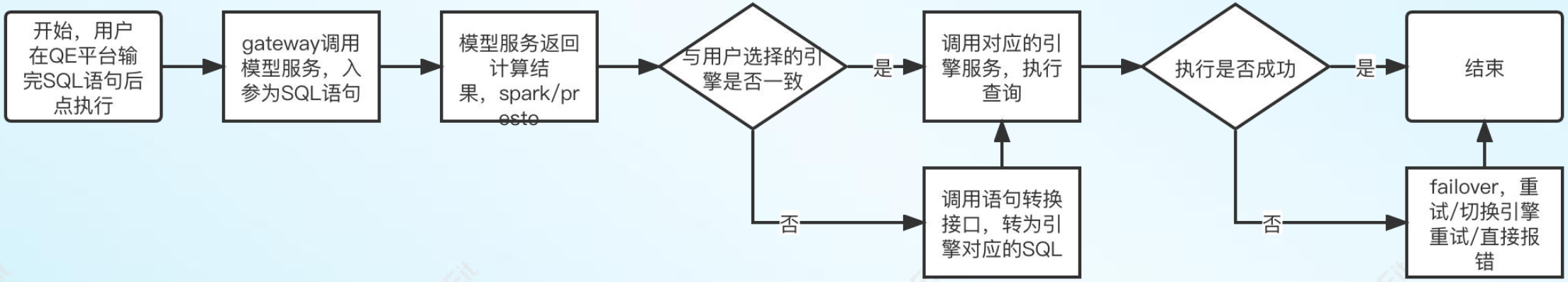
- 模型在测试集上的效果评估指标如下图所示，能够正确识别92%在presto上能够正确执行的SQL，能够正确识别90%在presto上失败failover到spark执行的SQL。

xgboost准确率为: 0.911

	precision	recall	f1-score	support
0	0.92	0.90	0.91	996
1	0.90	0.92	0.91	1004
accuracy			0.91	2000
macro avg	0.91	0.91	0.91	2000
weighted avg	0.91	0.91	0.91	2000

模型应用

- 待推荐SQL语句先进行向量化，后筛选出模型所需特征，送入判断；
- 对于用户初始引擎与推荐引擎不一致的情况，在切换引擎时涉及到SQL语句的转换，调用已有功能模块；
- 切换引擎后存在少量执行失败的情况，failover到原始引擎执行。



收益

- 自动选择最合适的执行引擎，完成语句转换，无需用户额外学习成本；
- 原有执行效率不变，且降低失败概率，用户体验上升；
- 资源消耗成本下降。

5. AI算法在大数据治理中的应用展望

应用场景

- **采集：**如日志评估，日志生产方式是否合理、建议日志生产频率；
- **传输：**如入侵检测，数据是否被篡改、是否发生异常；
- **存储：**如存储建议，表数据结构是否合理、是否有冗余；
- **处理：**进一步降本增效；
- **交换：**如数据共享，保障数据安全；
- **销毁：**如时机判断，何时可以销毁数据、是否会影响关联数据等。

DATACAKE
streamline data & AI workflow

Thanks

Q&A

公众号

微信群