elastic 中国开发者大会 2023

elasticsearch整合机器学习强化排序

彭晟，技术专家
上海哈啰普惠科技有限公司 ，2023/04/08

分享嘉宾

嘉宾简介

四轮司乘匹配引擎负责人，
es在搜广推领域大规模应用以及与机器学习
在线预测整合有多年的实践优化经验

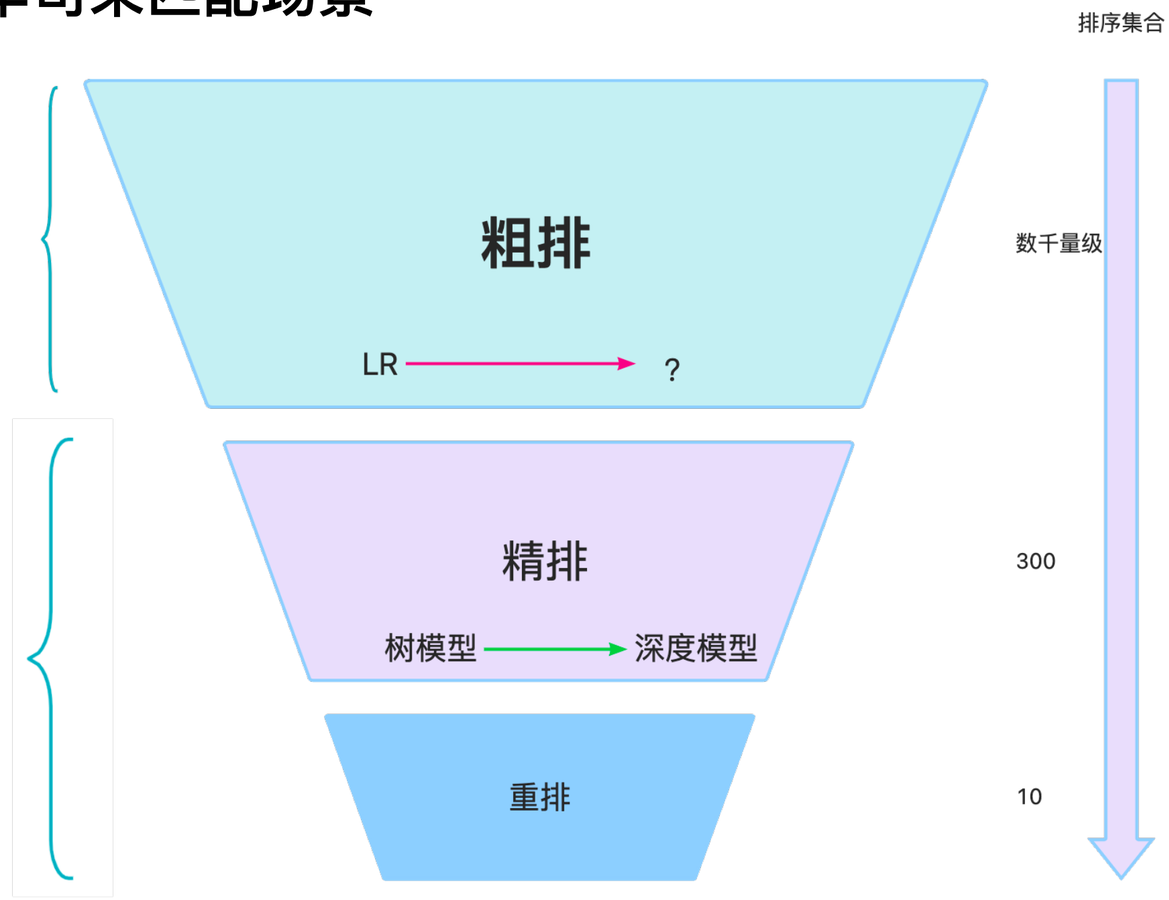
- 背景介绍
- 整体方案
- 关键组件
- 后续

统

背景介绍

顺风车司乘匹配场景

service



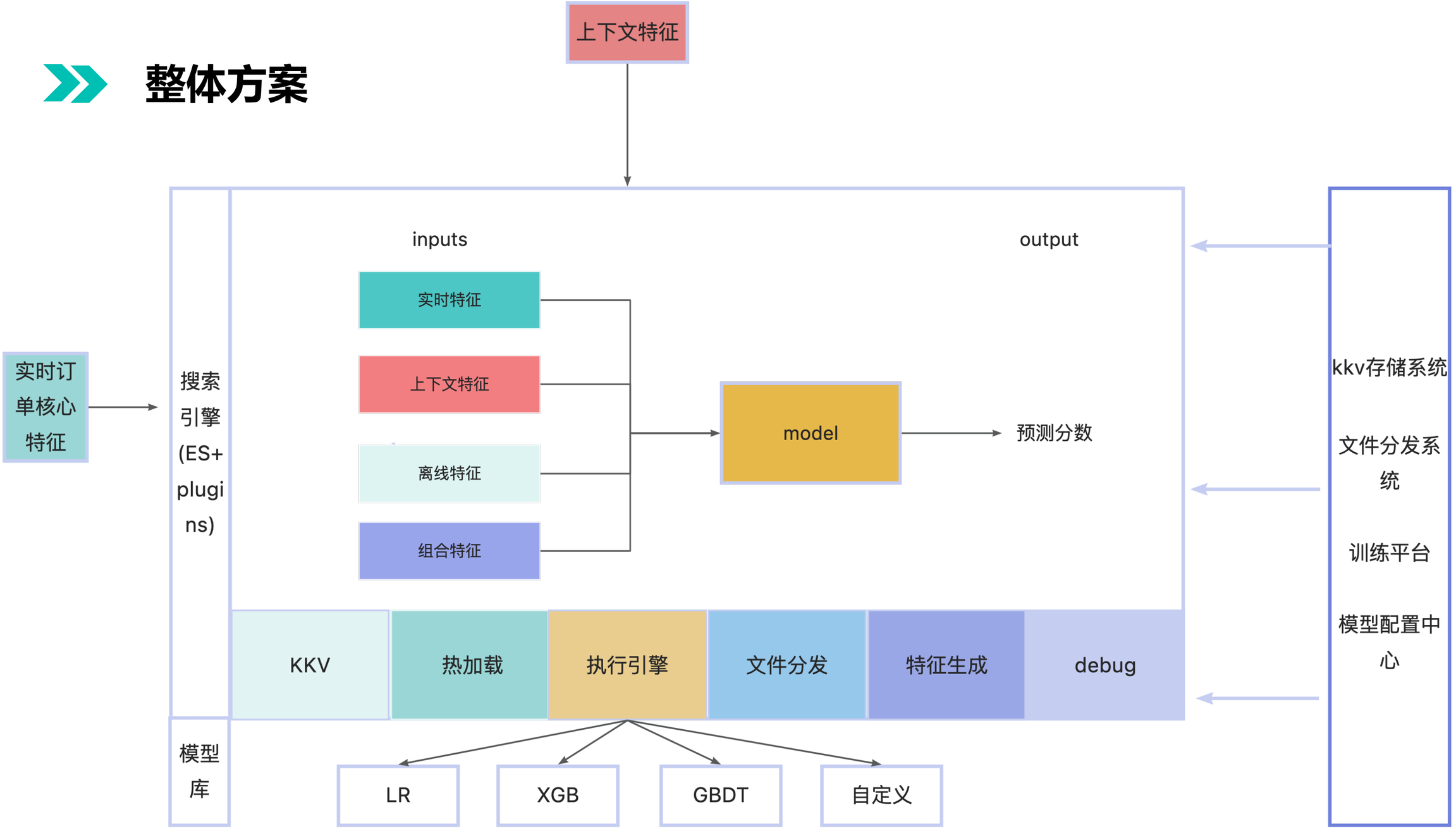
目标:

支持模型排序

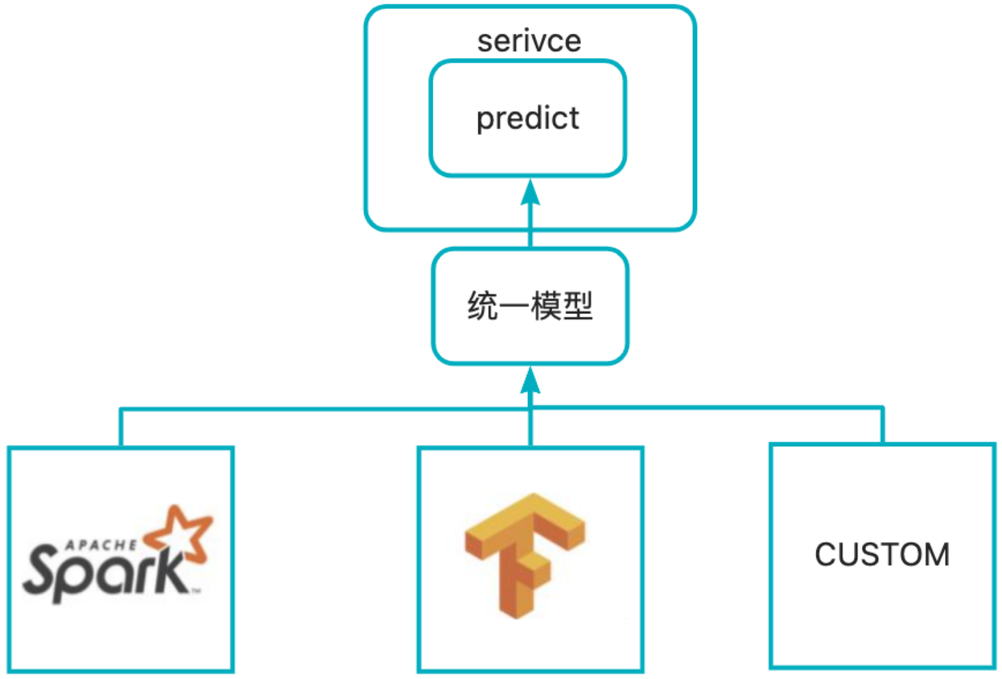
支持树模型全排序

支持配置化迭代

整体方案


上下文特征
inputs
output
实时特征
上下文特征
离线特征
组合特征
model
预测分数
实时订单核心特征
搜索引擎(ES+plugins)
KKV
热加载
执行引擎
文件分发
特征生成
debug
模型库
LR
XGB
GBDT
自定义
kkv存储系统
文件分发系统
训练平台
模型配置中心

执行引擎



XGB、Light GBM、GBDT+LR

KKV系统

1.海量离线特征存储查询的挑战

检索rt要求高
1000*100=10w

↓

特征本地化

特征量大
百亿特征

↓

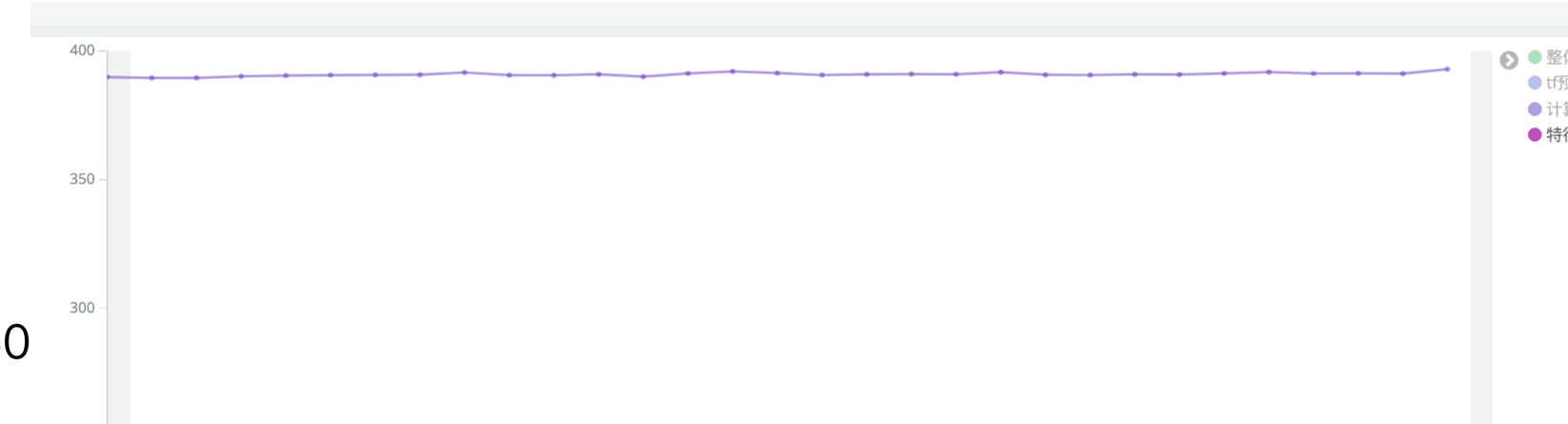
MMAP(读取二进制,实时反序列化)

2.解决方案:

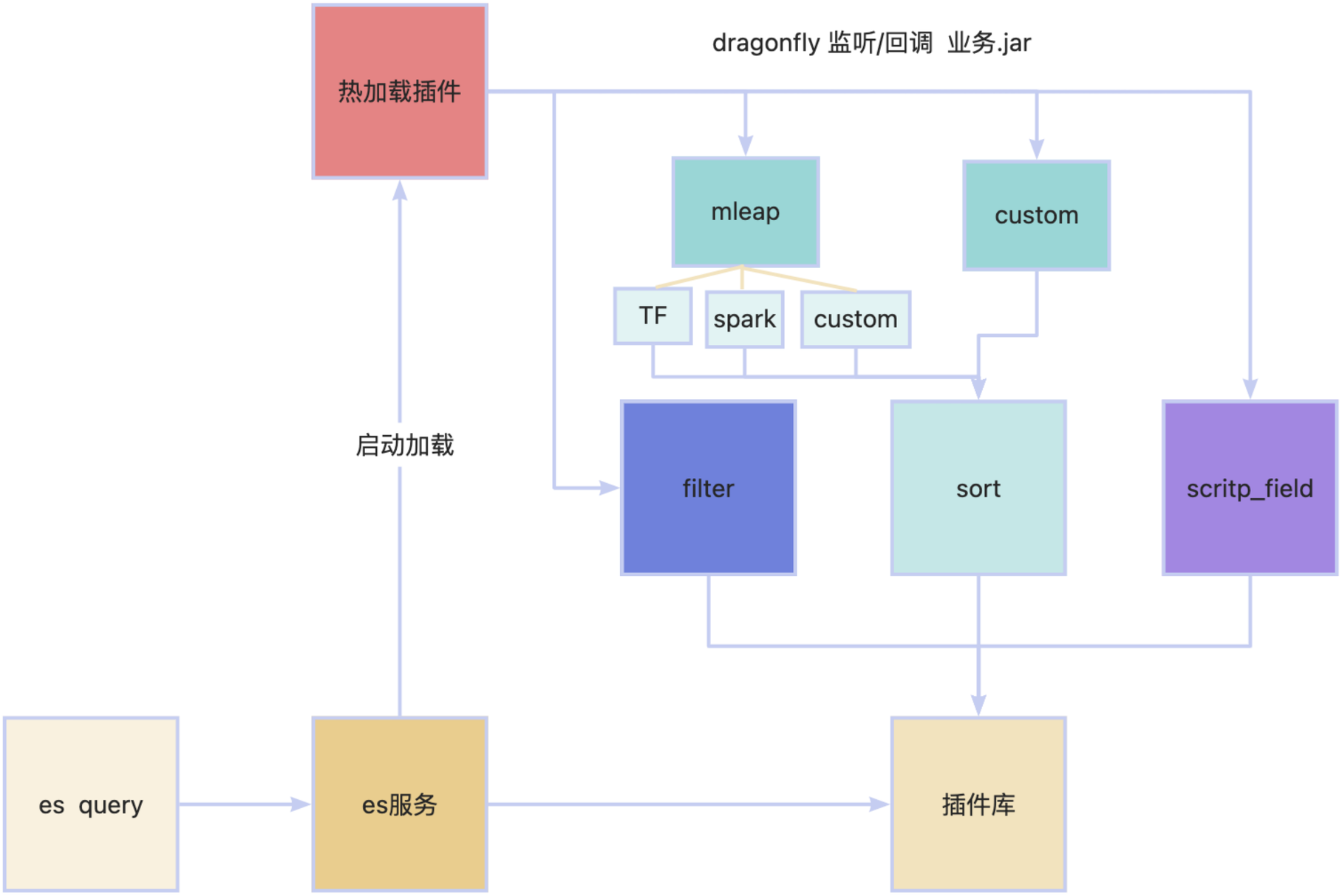
linkedin/PalDB

3.线上的一些数据
50G –> 20G –> 10G

5.6ms 390*(100离线+100组合+50



热加载



tips:

是否存在外部资源
需要手动关闭

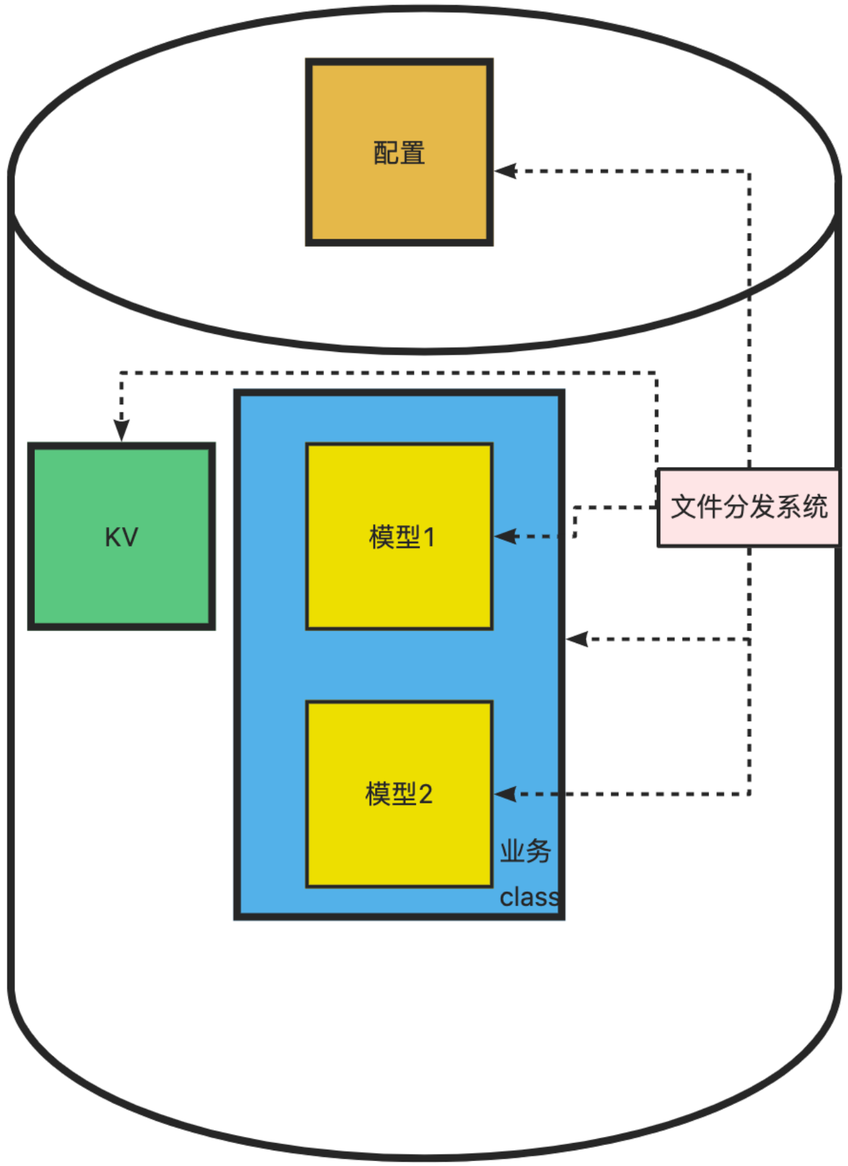
是否存在第三方jar,存在内存泄漏

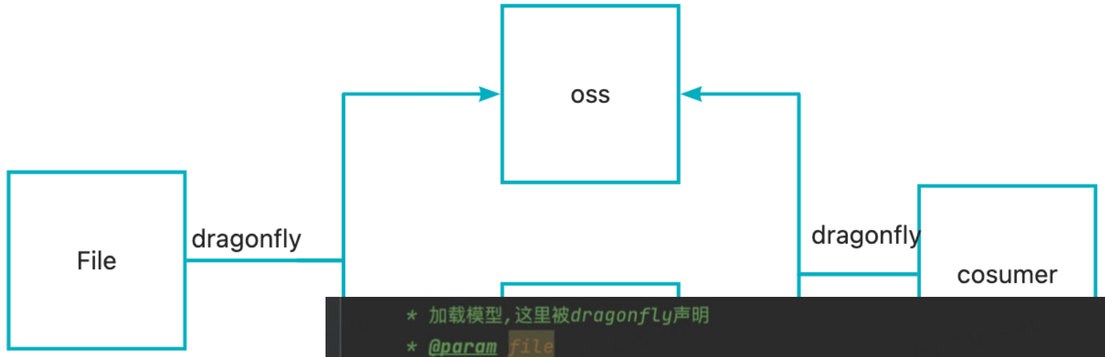
提前加载预热,防止突刺

分层热加载
biz热加载中, kv 独立

错误日志限制输出

文件分发系统-dragonfly






```
* 加载模型,这里被dragonfly声明
* @param file
* @param path
* @throws IOException
* Dragonfly 自动使用当前的class为ranker
*
*/
@DragonflyInitialization(value = "deep",filterBean = ApolloFilter.class )
public void getDeepFMModel(File file, String path) throws IOException {
    super.getDeepFMModel(file, path);
}
```

功能:

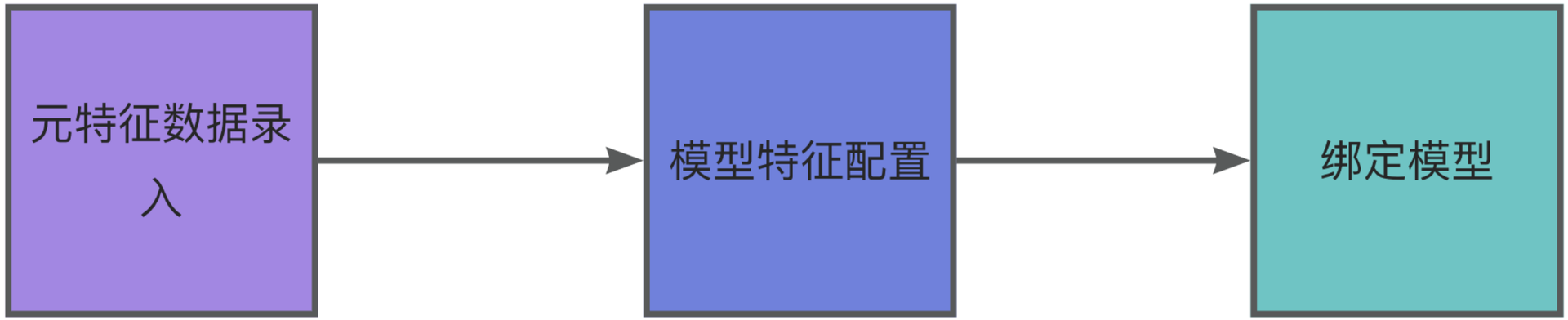
1. 文件变更自动下载最新文件,触发业务回调
2. 极速MD5校验　0.5h–> 1m
3. 易用性　支持注解驱动
4. 支持灰度加载　apollo,自定义配置
5. 更新回调状态
6. 支持多环境
7. 支持多回调

等等

```
@Dragonfly(storagePath = "model/freq_route_deep_model_v3.tar.gz")
public void getDeepFMModel(File file, String path) throws IOException {
    super.getDeepFMModel(file,path);
}

@Dragonfly(storagePath = "model/freq_route_deep_model_v3.tar.gz")
public void test(File file, String path) throws IOException {
    System.out.println("单体测试3");
}
```

配置化的迭代



建时间	feature_name	dictName	expression	stage
021-08-25 16:47:33	apply_method	driver_portrait_temporary_travel	--	user
021-08-25 16:47:33	vehicle_model_name	driver_portrait_temporary_travel	--	user
021-08-25 16:47:33	vehicle_company	driver_portrait_temporary_travel	--	user
021-08-25 16:47:33	driver_usual_route_lines	driver_portrait_temporary_travel	--	user

导入 导出

绑定策略Id *

driver_fadan_v1

绑定策略的id,可以使用xx算法 + 下划线fg ,例如: deepFM_fg

特征表1

序号	* feature_name	expression	* dataType	defaultValue	stage
1	time_diff	computeTimeDiff:item.sta	int	0	item
		录入表达式后,后续就不用填写了		特征默认值,如果不填写会使用0 代替	
2	start_time_diff	timeDiff:user.start_time	int	0	item
		录入表达式后,后续就不用填写了		特征默认值,如果不填写会使用0 代替	
3	end_time_diff	timeDiff:user.start_time	int	0	item
		录入表达式后,后续就不用填写了		特征默认值,如果不填写会使用0 代替	

⊕ 新增一项 批量导入

在ES中模型预测进行debug

原始user特征

原始item特征

模型入参特征

```
"size": 1,
"timeout": "1500ms",
"explain": true,
"query": {
    "bool": {
        "must": [
            {
                "script_score": {
                    "query": {
                        "match_all": {
                            "boost": 1.0
                        }
                    },
                    "script": {
                        "source": "score",
                        "lang": "hello_hitch",
                        "params": {
                            "type_name": "MleapScore",
                            "start_time_from": 1672718700000,
                            "sortPolicy": "psg_fadan_v1.rf",
                            "end_point": [
                                31.194311,
                                121.436829
                            ],
                            "start_time_to": 1672719000000,
                            "estimate_price": 3851,
                            "passenger_count": 1,
                            "userId": "1200364780",
                            "start_time": 1672718850000,
                            "order_create_time": 1672717412521,
                            "start_point": [
                                31.121676,
                                121.355867
                            ],
                            "now": 1672717413522,
                            "start_city_code": "021",
                            "end_city_code": "021"
                        }
                    },
                    "boost": 1.0
                }
            }
        ],
```

user

item

```
"gender" : 1,
"e  p  nt" : "31.195515,121.4366",
"n  e_plate" : "粤D9589N",
"o d r_ui " : "JP20230309033627000012003368 1",
"s  t_ti " : "2023-03-09 18:15:52.828",
"o d r_re te_time" : "2023-03-09 18:13:25.464",
"s  _ un " : 4,
"o d r_ s ctive" : 1,
"s  t_p i t" : "31.123389,121.36506",
"s  t_ dd ess" : "旭辉·莘庄中心",
"s  t_ it _code" : "021",
"u e_ w_ d" : 1200336811,
"o d r_ pd e_time" : "2023-03-09 18:13:25.809",
"e  ' ci y_ ode" : "021",
"en  ddres " : "11号线;1号线;9号线",
                : 1
},
"_explanation" : {
  "value" : 0.71413577,
  "description" : "sum of:",
  "details" : [
    {
      "value" : 0.7141357660293579,
      "description" : "mleap(policyName=psg_fadan_v1.rf,params={type_name=MleapScore,
        start_time_from=1672718700000, sortPolicy=psg_fadan_v1.rf, end_point=31.194311,121
        .436829, start_time_to=1672719000000, estimate_price=3851, passenger_count=1, userId
        =1200364780, start_time=1672718850000, order_create_time=1672717412521, start_point=31
        .121676,121.355867, now=1672717413522, start_city_code=021, end_city_code=021},detail
        ={distance_diff=-953.0, start_end_rate_psg=0.09228564405377388, hitch_rate=0
        .8280468259626927, same_start_city=1.0, end_distance=155.0, driver_order_whether_same=1
        .0, end_time_diff=563795 .0, start_end_rate_driver=0.097918055951 5953, start_rate_driver
        =0.08508413285617014, plog_start_time_di f=5639539.0, start_driver_angel=145
        .23456427892827, driver_distance_rate=0.9 24782991409302, driver_hour=18.0,
        psg_order_whether_same=1.0, end_rate_psg=0.012095690704862145, price=3851.0,
        driver_distance=14827.0, end_rate_driver=0.012833919376134472, start_rate_psg=0
        .08018995076417923, end_driver_angel=55.851865118506474, time_diff_length=300.0,
        same_end_city=1.0, start_time_diff=5638253.0, start_distance_mhd=895.0,
        psg_rate_psg_start_end=0.915511429309845, plog_start_time_diff=5638103.0,
        driver_rate_driver_start_end=0.9108147621154735, psg_driver_angel=3.4246394060168033,
        time_diff=5637952.0, passenger_count=1.0, share_rate=0.8789617334150281,
        pstart_publish_time_diff=1437.47900390625, psg_distance=15780.0, plog_publish_time_diff
        =5639540.0, psg_distance_rate=1.0610324144363403, start_distance=1065.0, seat_count=4.0,
        driver_weekday=3.0, result=0.7141357943259036})",
      "details" : [
```

稳定性

1.完善的压测方案

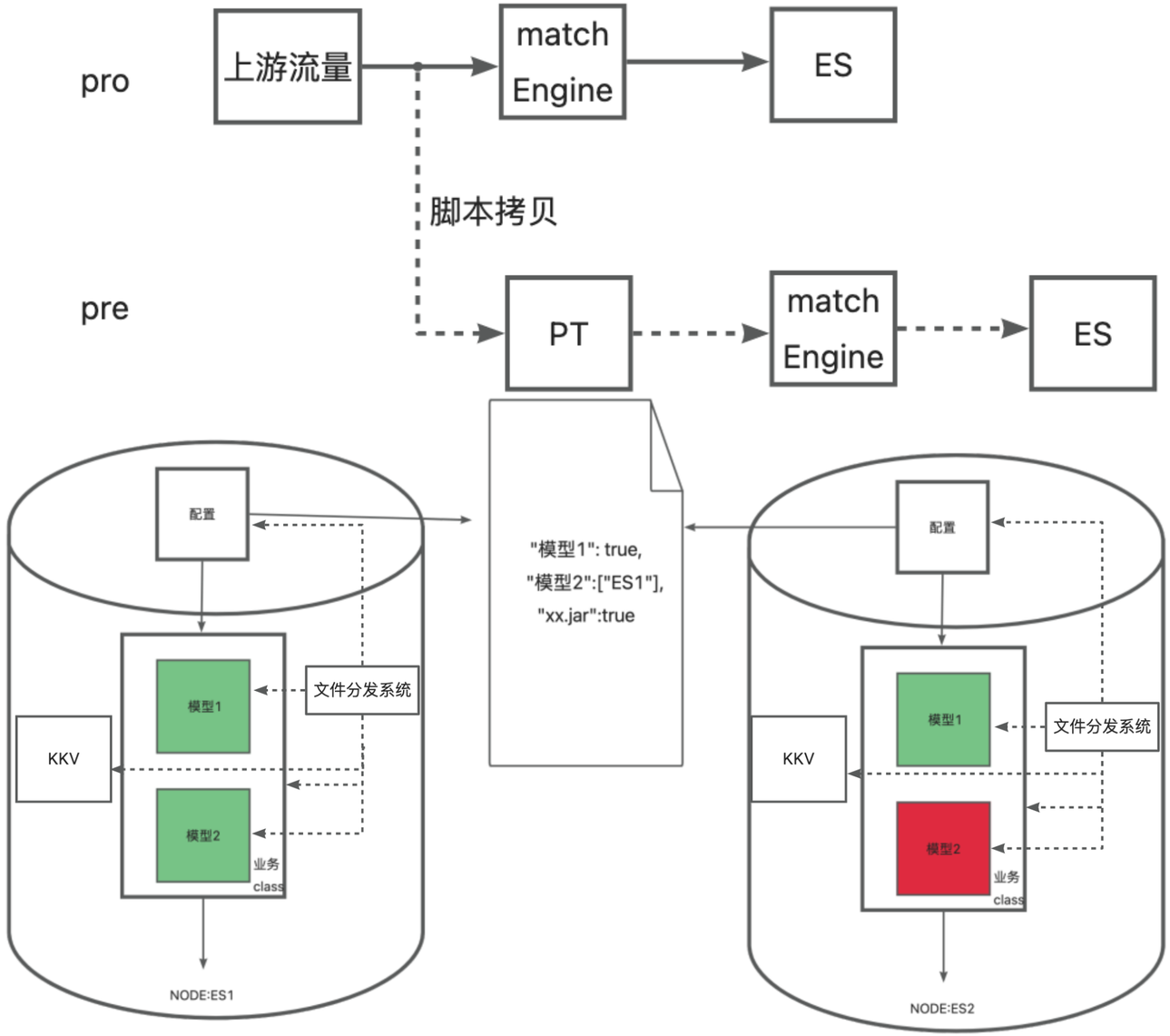
上线前 压测回归&新功能验证

2.风险点,变更点极限压测

biz.jar model ,kv 变更频率 1d–>1m

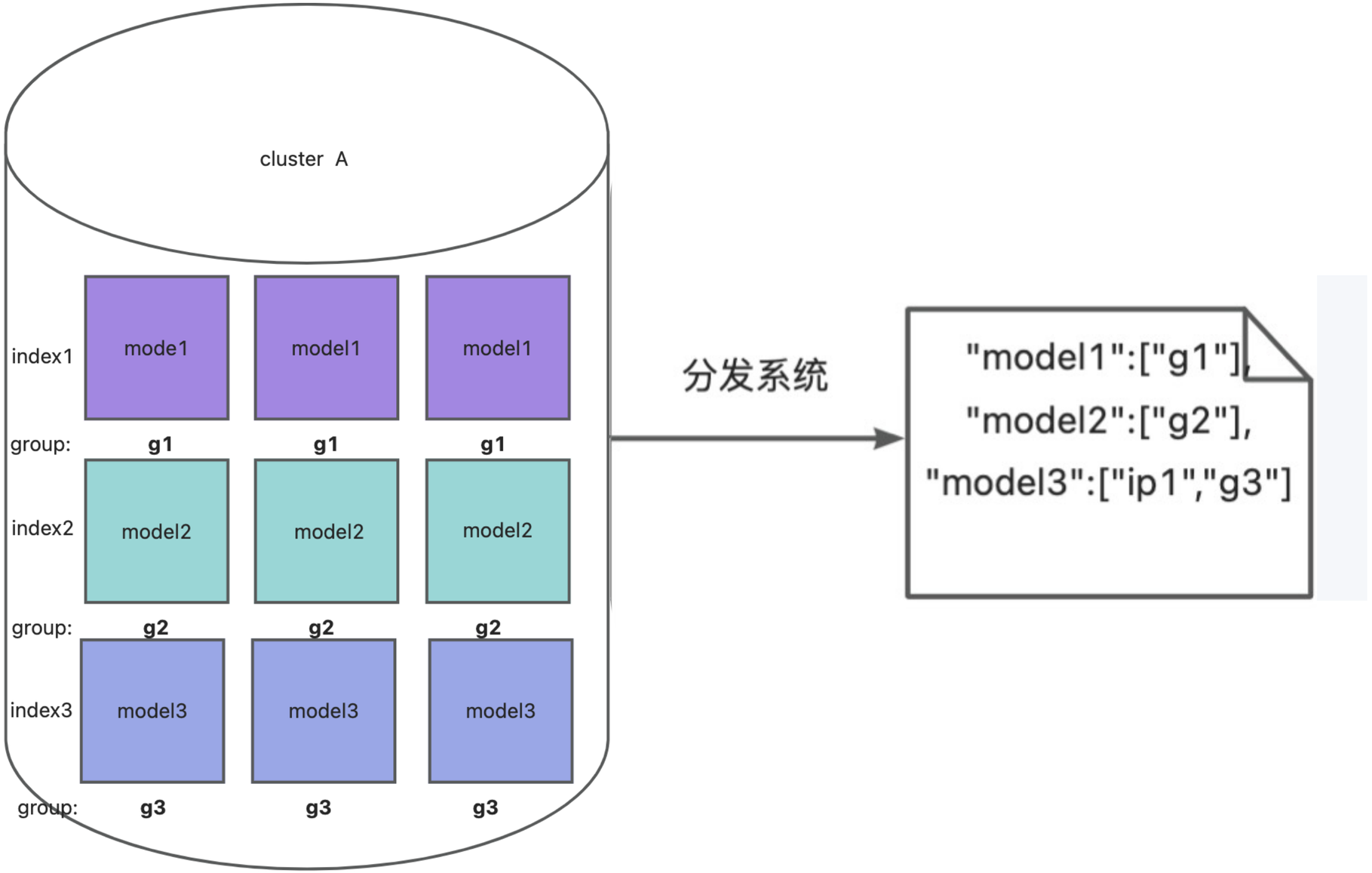
3.变更的灰度&顺序加载

文件分发系统增加了模型,业务插件上线的顺序加载,灰度功能,保证稳定.

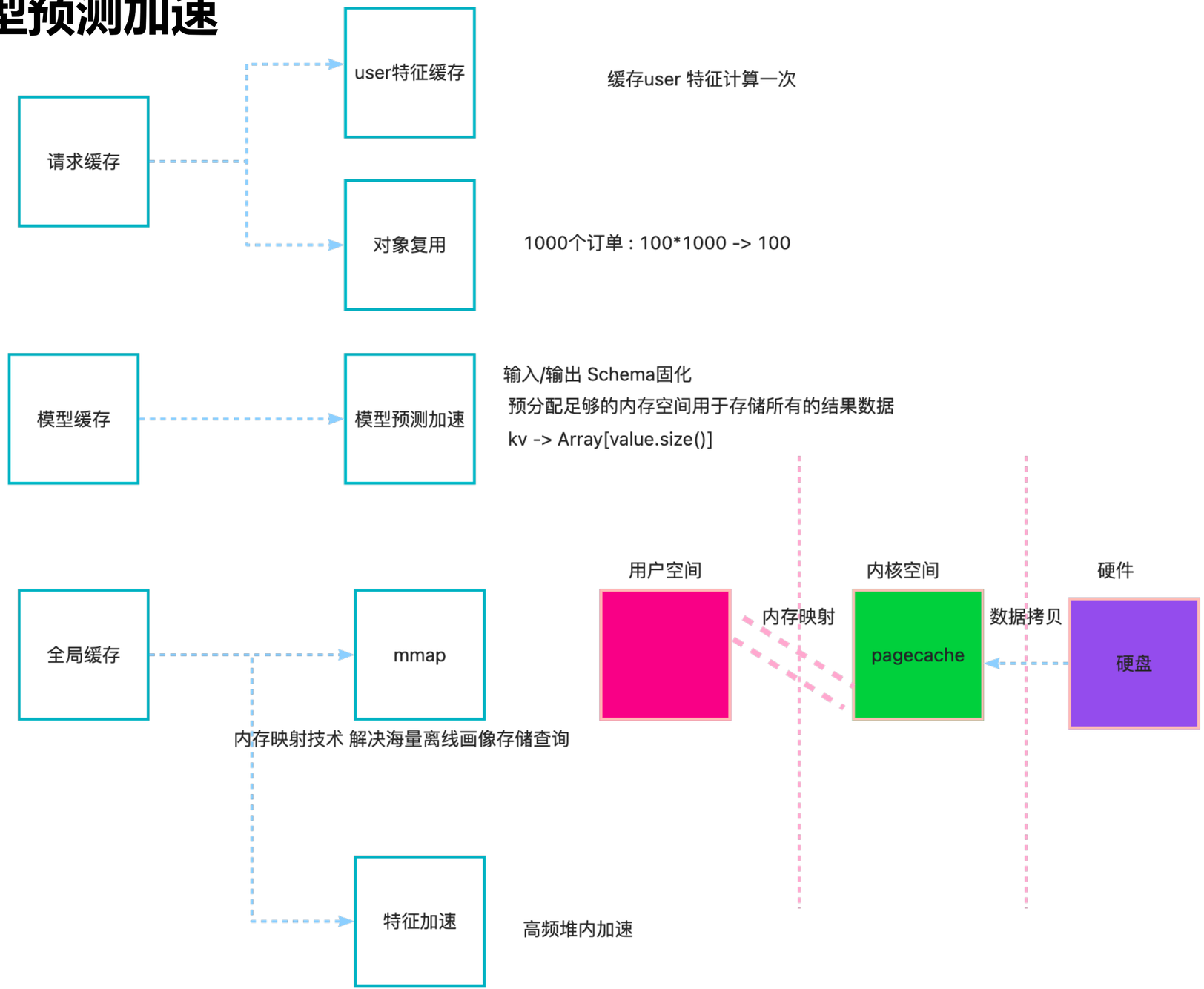


稳定性

4. 机器学习分组加载



模型预测加速

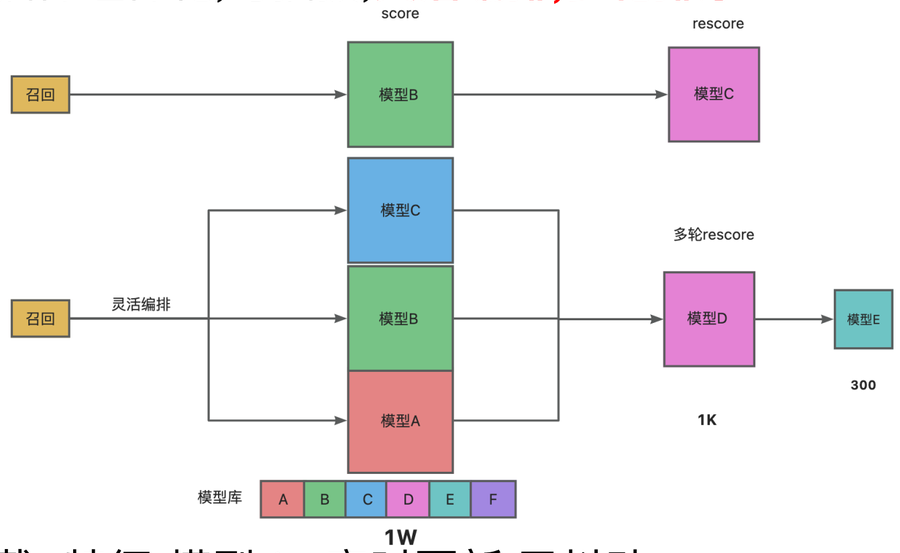

user特征缓存
缓存user 特征计算一次
请求缓存
对象复用
1000个订单：100*1000 -> 100
模型缓存
模型预测加速
输入/输出 Schema固化
预分配足够的内存空间用于存储所有的结果数据
kv -> Array[value.size()]
用户空间
内核空间
硬件
内存映射
数据拷贝
pagecache
硬盘
全局缓存
mmap
内存映射技术 解决海量离线画像存储查询
特征加速
高频堆内加速

上线后业务上的表现

1.支持spark 全部的模型

2.模型迭代,**免开发**,通过特征配置化可以快速稳定上线

3.算法插件组件化,可插拔,**灵活编排,多轮排序**



4.热加载, 特征 模型 jar实时更新,无抖动

5.火焰图,单核心场景,排序只占到7%的cpu消耗

6. 在单机单分片场景 1500深度下，树模型相比LR 多了10ms

7. 全场景 LR –> 树模型　　核心ab 增加 1.2%

```
Try  Settings  Help
98    "from": 0,
99    "size": 10,
00    "query": {
01      "function_score": {
02        "query": {
03          "bool": {⟺}
28        },
29        "functions": [
30          {
31            "script_score": {
32              "script": {⟺}
53            },
54            "weight": 0.5
55          },
56          {
57            "script_score": {
58              "script": {⟺}
79            },
80            "weight": 0.5
81          },
82          {
83        "script_score": {
84          "script": "return 1"
85        },
86          "weight": 1
87        }
88      ],
89      "score_mode": "sum",
90      "boost_mode":"replace"
91    }
92  },
93  "_source": {⟺}
17 }
18
```

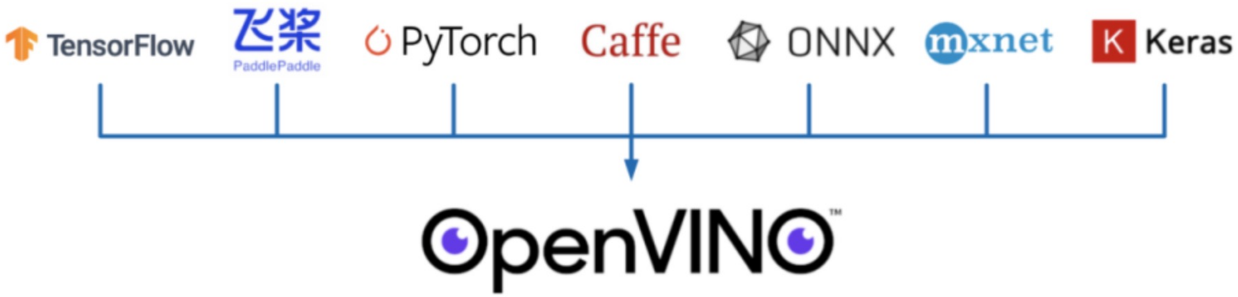
后续动作

elasticsearch 补齐排序短板--深度学习

整合 tensorflow

row by row – batch calculate 解决性能问题

整合openvino



实现最终目标 elasticsearch 一排到底 极致性能 高度灵活

elastic 中国开发者大会 2023

感谢观看

elastic
中文社区

专业、垂直、纯粹的 Elastic 开源技术交流社区

https://elasticsearch.cn/