杨华磊 - 大数据&AI应用开发工程师

杨华磊 丨 统招本科 丨 山东师范大学|亦庄

17600505796(微信同号)丨yyqq188@foxmail.com

个人简述

10年开发经验,能针对业务找准技术要点并积极解决技术难点。善于找到需求中的通用技术点，并封装为工具。善于研究或预研技术上的问题，并积极分享,根据业务要求设计技术架构，并开发框架，分拆任务模块。

个人网站 :https://yanghl.top

博客: https://blog.csdn.net/yyqq188

github: https://github.com/yyqq188

技能特长

- 熟练使用python,掌握常用的数据分析的包pandas,数据可视化的库matplotlib,plotly,bokeh以及数据展示web工具streamlit,熟悉常用的pythonweb开发框架flask,fastapi,tornado,以及爬虫框架scripy.
  并在csdn上发表过《pandas优化方法的总结》《bokeh与tornado结合的三种方式》等关于python的文章.
- 了解原型设计工具figma,熟悉基于React的CharkraUI, 并能够通过figma+chakraUI kit工具快速实现具备可交互页面的开发,熟悉tremorUI,专注于后台dashboard的UI组件库,并能够基于该组件库快速构建dashboard管理页面,实现后台数据的产品化封装.

- 关注基于大模型的应用场景,包括基于RAG的客服服务助手,基于大模型function call的实时数据处理和text2SQL的报表开发,以及建立在其上的精准营销
- 熟悉基于大模型的文本转sql框架Vanna AI 以及 pandasai, Vanna AI通过RAG实现特定场景下的复杂sql生成,并通过自定系统提示词和RAG的方式提高文本转sql的准确度.pandasai 通过将csv等格式的数据转为pandas,提供数据语言交互分析以及数据的可视化分析.
- 熟悉传统机器学习算法,并能通过scikit-learn库实现数据的建模,围绕存量保单未来的退单风险,通过逻辑回归,随机森林,Adaboost等算法给出不断优化的分析.
- 熟悉tensorflow2深度学习模型开发框架,能利用tf2的keras模块构建和训练模型,并能基于tensorflow2 serving进行模型的部署使用.

工作经历

- 2021.3-至今　　中电金信　　数据分析挖掘
- 2016.11-2021.3　　亚信科技CTC数据应用开发部　　数据技术经理
- 2015.2-2016.11　　北京数字新思科技有限公司　　大数据工程师

工具&项目

基于RoBERTa& BiLSTM的智能客服服务开发

bertServer　kashgari　flask

为了提高人保在智能客服服务上回答的准确性和针对性,通过收集保险领域知识以及生成保险领域相关的对话数据,通过bertServer加载的模型,利用 kashgari包的BLSTMCRFModel (BiLSTM + CRF)命名实体识别算法,对保险的语料数据再训练得到优化后的模型,并通过flask封装成API接口对外提供模型服务.

保险险种选择分析

sklearn pandas matplotlib

围绕保险产品复购场景展开,基于用户历史购买数据、市场动态等多维度信息。深入分析用户购买险种组合销售关联度、用户生命周期价值等关键指标,构建选品预测模型。
项目职责:

1、使用Pandas对数据进行清洗,对不规整数据进行标准化处理
2、使用GBDT对特征重要性进行排序,筛选特征
3、使用Matplotlib可视化销量复购率、销量关系等关键数据,并形成报告
4、选取Decision Tree训练数据,画出Precision-Recal趋势图
5、使用Random Forest、LogicRegression训练数据,并对比训练效果
6、使用GirdSearchCV交叉验证和参数自动调优,输出结果提供决策支持

智能反欺诈分析与检测

sklearn matplotlib

1、对接其他技术部门进行数据同步、完成采集校验
2、解决模型欠采样问题，补充缺失数据
3、选取LogicRegression训练数据，且输出包含概率的结果，提供决策依据
4、预测结果存库，使用神经网络模型拟合新数据
5、通过 Matplotlib 可视化

awesome ai 项目说明

langchain gemini deepseek streanlit

效果展示地址:http://8.219.247.119:8501/

这是一个MVP项目,利用免费模型以及高效的可视化界面,快速实现大模型agent应用开发的效果展示,方案可行,过程代码可以快速进行正式的后端开发.模型使用的google的 `gemini` 模型,以及groq部署的 `deepseek` 模型,可视化上使用 `streamlit` ,agent开发框架使用的是 `langchain`

基于新华实时数仓,以sparksql的方式写流式作业的脚手架

`sparkstreaming` `hbase` `java`

github地址：https://github.com/yyqq188/spark_sql_helper

工具描述:

此工具解决了之前实时业务指标开发中,**开发效率低,维护难度大的痛点**,原来项目基于java 或 scala来开发实时业务指标,开发周期长,代码冗余且维护难度越来越大,该工具以sql的形式开发实时指标,并以json文件的形式来部署实时任务的实时作业脚手架工具. 如果要开发新的算子,利用脚手架工具,只需要实现脚手架的transform接口并实现其中的方法,就可以在json 文件中指定并调用.日常的实时任务开发就转为sql开发,进而可以实现流批一体开发.该脚手架工具利用阿里开源的datax库,并参考了apache的seatunnal项目的实现思路.

通用FlinkSQL数据开发平台

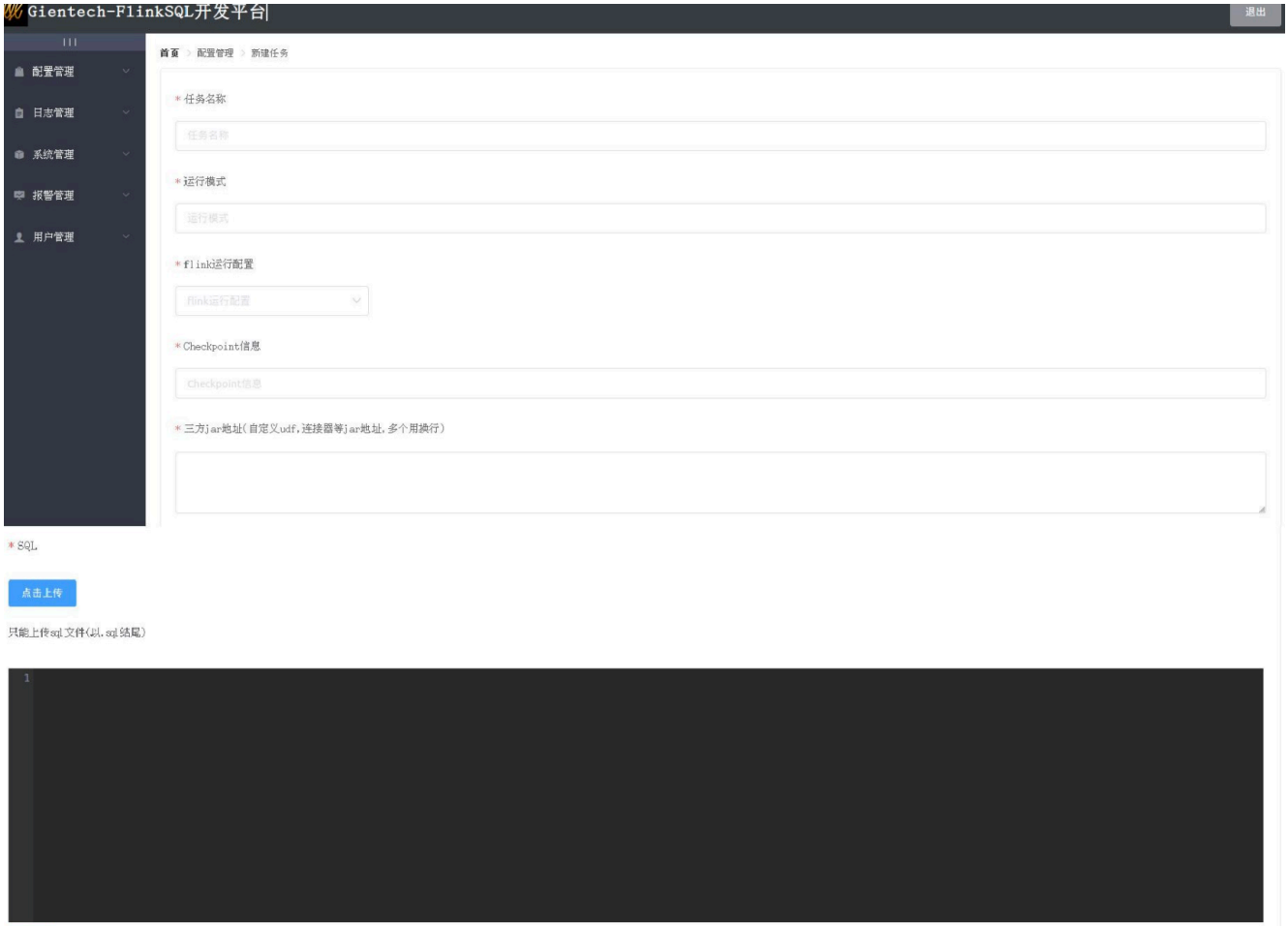
`flink` `springboot` `hibernate` `vue` `codemirrorjs`

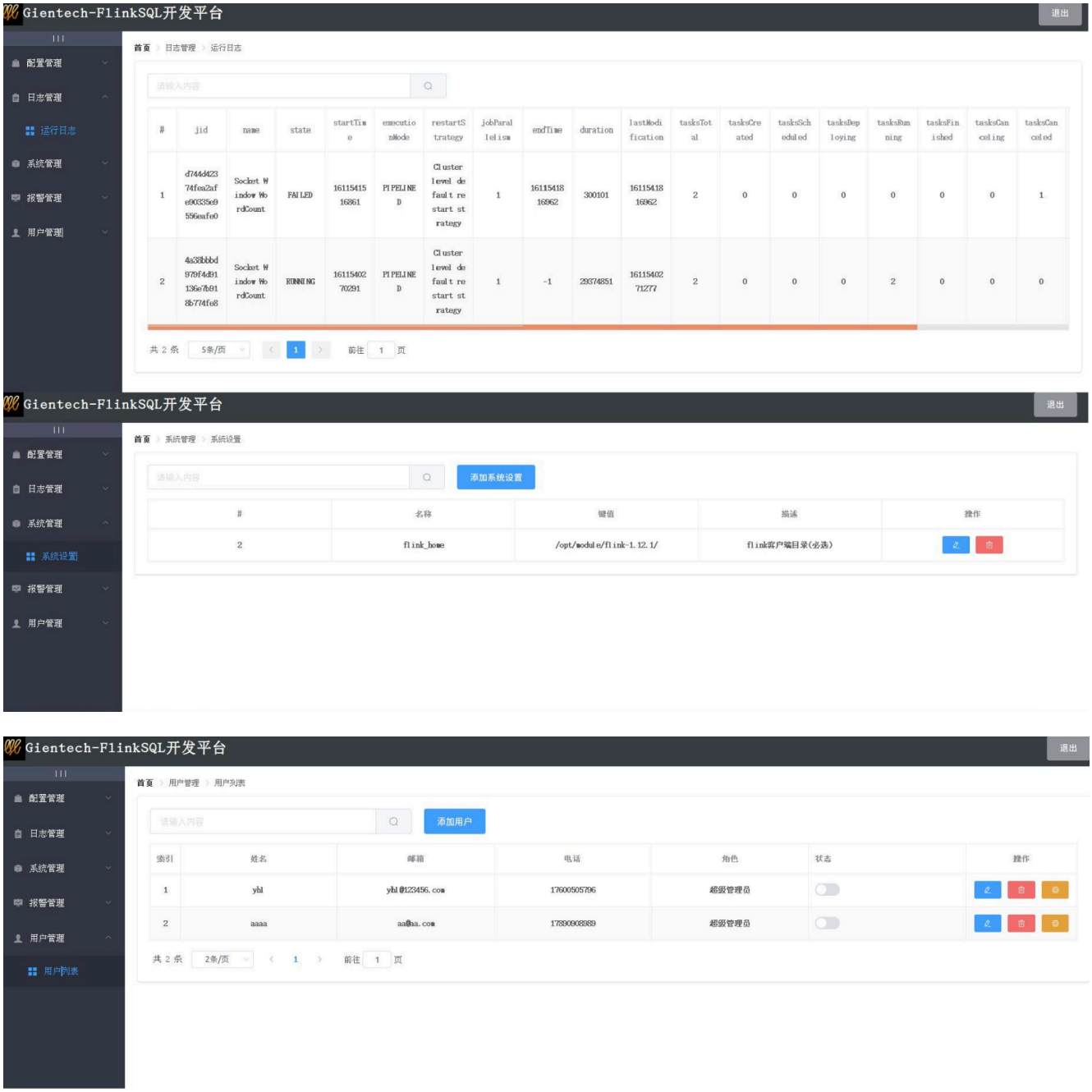
工具描述:

结合实际的工作场景，开发此基于flink的 sql开发平台， 目标用户为业务人员。可以通过写flinksql一站式开发实时程序并能监控job的执行情况。产品以后台管理平台的形式开发，包括配置管理，新建任务，任务列表，日志管理，系统配置管理，告警配置，用户配置五个模块的功能。前后端分离架构，前端基于vue框架开发,后端基于springboot,结合flink rest api接口开发任务和集群监控的后台逻辑，通过codemirrorjs收集前端sql语句并分离出DDL DML语句，配合前端的任务设置，生成

job执行命令。该项目是采用前后端分离的方式开发的，由于技术实现上有前端vue后端springboot数据上flink的技术，技术综合性高，难点在数据层如何将sql以flink的job的形式提交到yarn上来执行。

页面展示：





基于flink的同步Gaussdb和DWS数据工具开发

flink kafka dws

工作描述:

需求沟通,需求分析,拆解技术点并具体开发实施

项目描述:

该工具实现消费DSG和OGG格式的kafka数据,并可分别同步到Gaussdb和DWS数据库中.微批同步,可以调节微批的时间跨度以及微批的数据量大小,提高同步效率.采用配置文件的方式,增加同步的新表,配置新表 的字段信息和主键信息,无主键也可以同步,无需开发同步逻辑.

实时数据即席查询平台 2020

flink clickhouse python

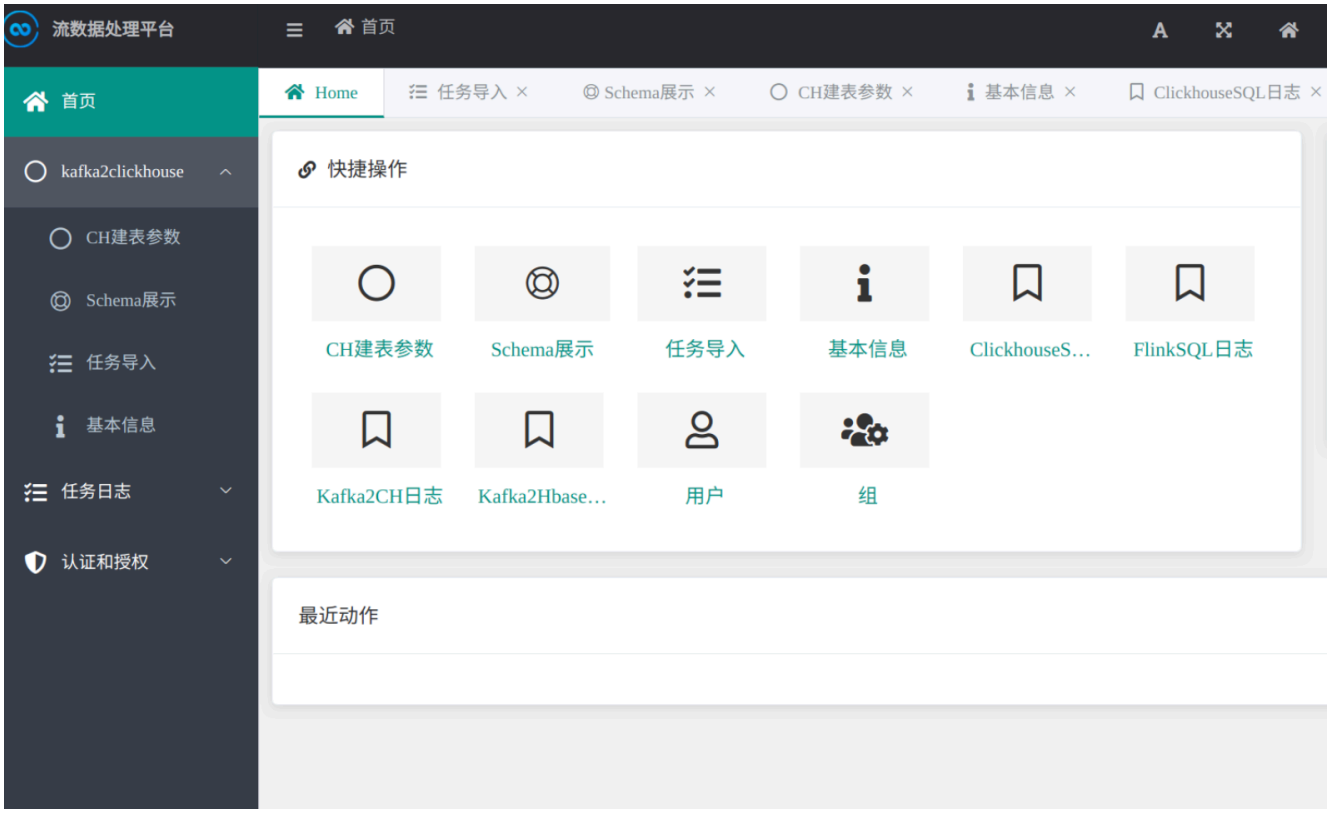
github地址:https://github.com/yyqq188/IStreamBigDataProcessingPlatform

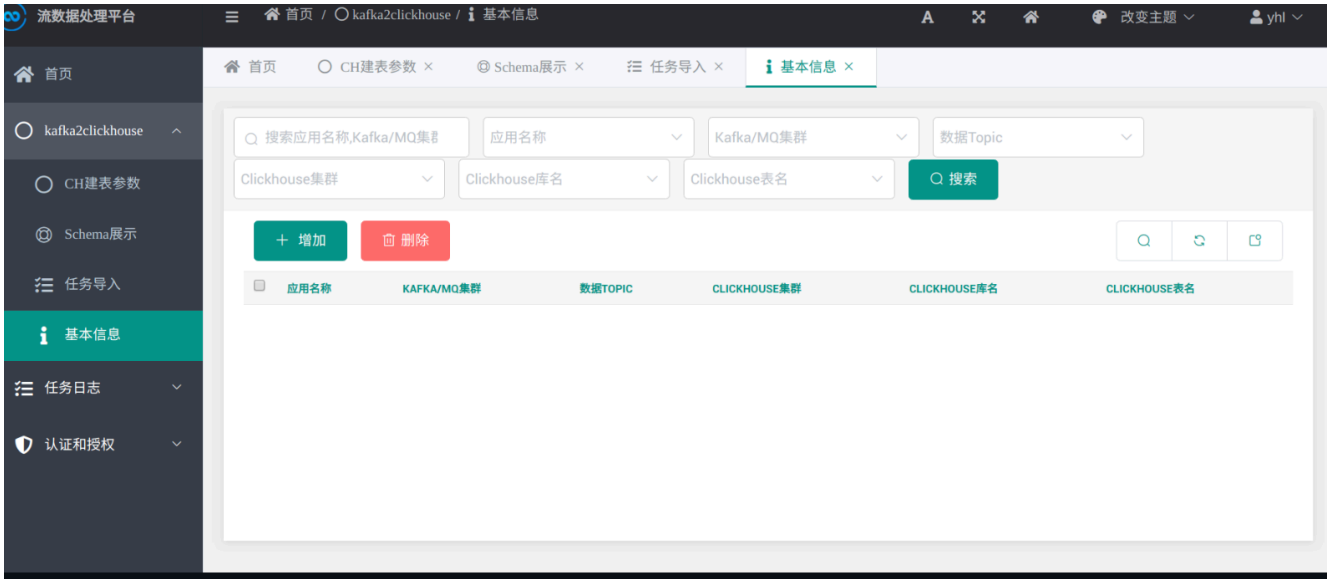
开发背景:

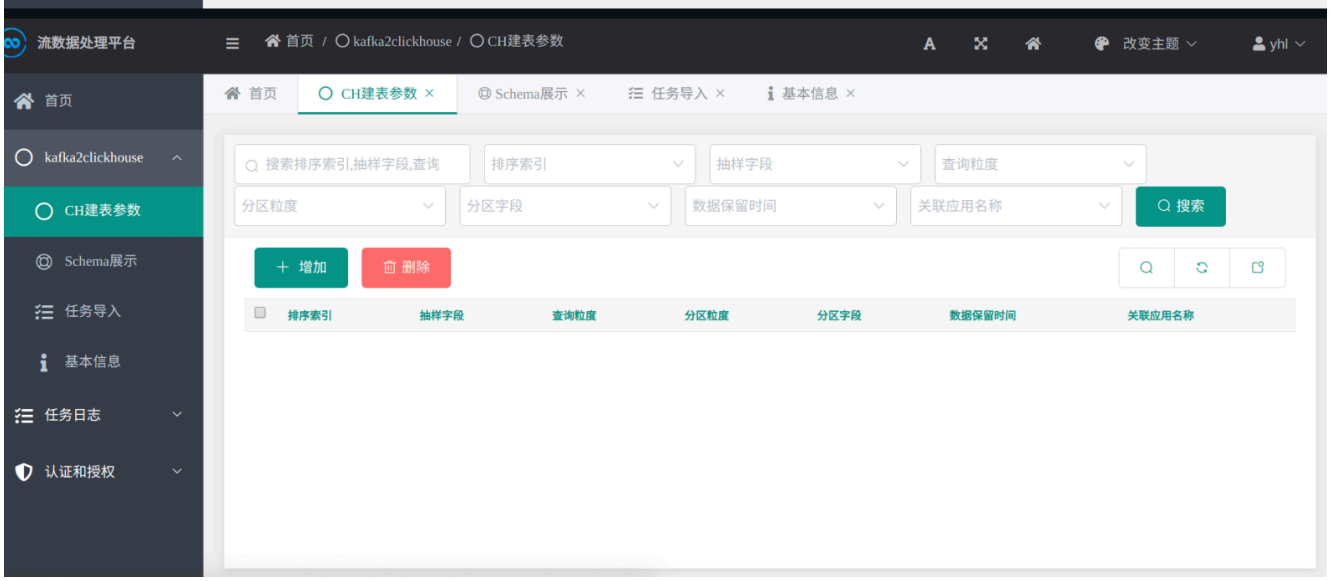
为了满足电信总部对实时数据处理的需求，在完成中国电信全国5G开户数实时统计的基础上，开发实时数据处理平台，

工具描述:

平台主要分三大模块，实时数据导入（hbase/clickhouse），实时计算(flinkSQL)，即席查询(clickhouseSQL)。业务人员只需要在页面上按需配置导入规则，生成的参数信息会存储在mysql 并被实时检测出生成导入程序并检测启动。实时计算和即席查询利用SQL语句描述业务逻辑。该产品可减少开发人员日常工作量，并能更快的响应业务需求。


流数据处理平台
首页
首页
kafka2clickhouse
CH建表参数
Schema展示
任务导入
基本信息
任务日志
认证和授权
Home
任务导入 ×
Schema展示 ×
CH建表参数 ×
基本信息 ×
ClickhouseSQL日志 ×
快捷操作
CH建表参数
Schema展示
任务导入
基本信息
ClickhouseS…
FlinkSQL日志
Kafka2CH日志
Kafka2Hbase…
用户
组
最近动作


流数据处理平台
首页 / kafka2clickhouse / 基本信息
改变主题
yhl
首页
kafka2clickhouse
CH建表参数
Schema展示
任务导入
基本信息
任务日志
认证和授权
首页
CH建表参数 ×
Schema展示 ×
任务导入 ×
基本信息 ×
搜索应用名称,Kafka/MQ集群
应用名称
Kafka/MQ集群
数据Topic
Clickhouse集群
Clickhouse库名
Clickhouse表名
搜索
增加
删除
应用名称
KAFKA/MQ集群
数据TOPIC
CLICKHOUSE集群
CLICKHOUSE库名
CLICKHOUSE表名


流数据处理平台
首页 / kafka2clickhouse / CH建表参数
改变主题
yhl
首页
kafka2clickhouse
CH建表参数
Schema展示
任务导入
基本信息
任务日志
认证和授权
首页
CH建表参数 ×
Schema展示 ×
任务导入 ×
基本信息 ×
搜索排序索引,抽样字段,查询
排序索引
抽样字段
查询粒度
分区粒度
分区字段
数据保留时间
关联应用名称
搜索
增加
删除
排序索引
抽样字段
查询粒度
分区粒度
分区字段
数据保留时间
关联应用名称

政企爬虫工具 2018.10-2019.5

docker python scrapy

github地址:https://github.com/yyqq188/spider_tool_common

dockerhub地址:yyqq188/spider_tool_common

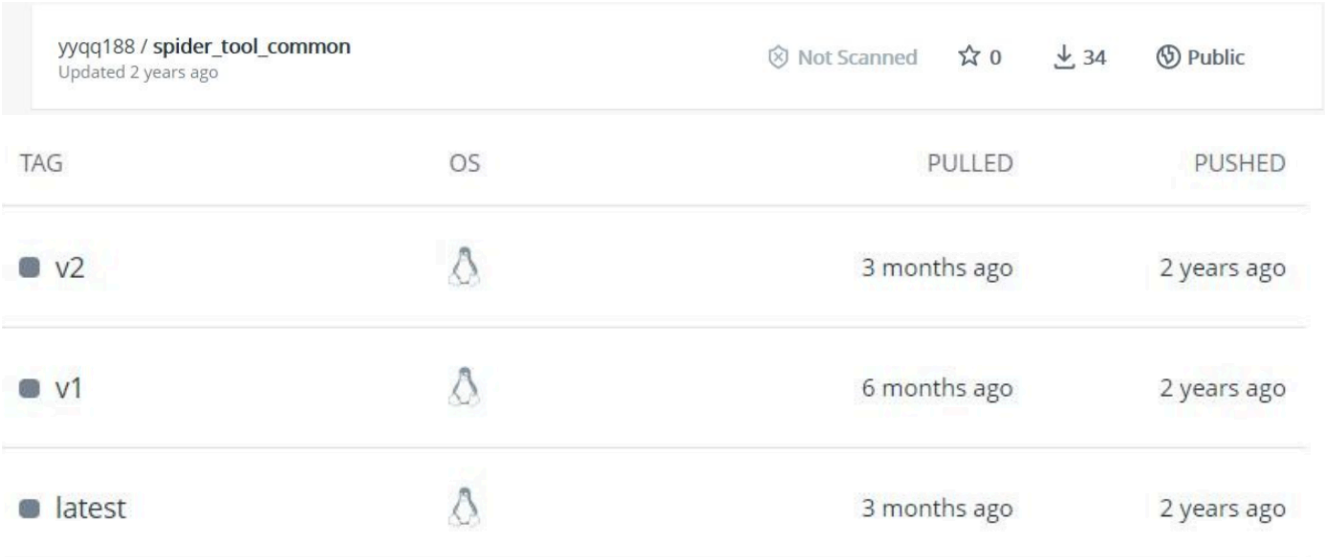
开发背景:

浙江电信政企需要爬取国家，省，市，县各级的公共采购网或招投标网上的招投标信息，由于需要采集的网站数量很多，为了简化爬虫程序的开发工作量，参考开源爬虫框架scrapy开发该工具

工具描述:

把招投标相关的通用逻辑加入到代码生成中， 分别开发通用版和招投标版的爬虫工具，并封装成pip包，并开发人员只需要下载对应工具的docker镜像， 并写好规范的json配置文件，就可以生成初始化好的爬虫程序，可以即可启动爬取或在此基础上进行修改后再爬取。**降低爬虫代码开发工作量,并将开发和部署统一起来,简化维护步骤**

dockerhub页面

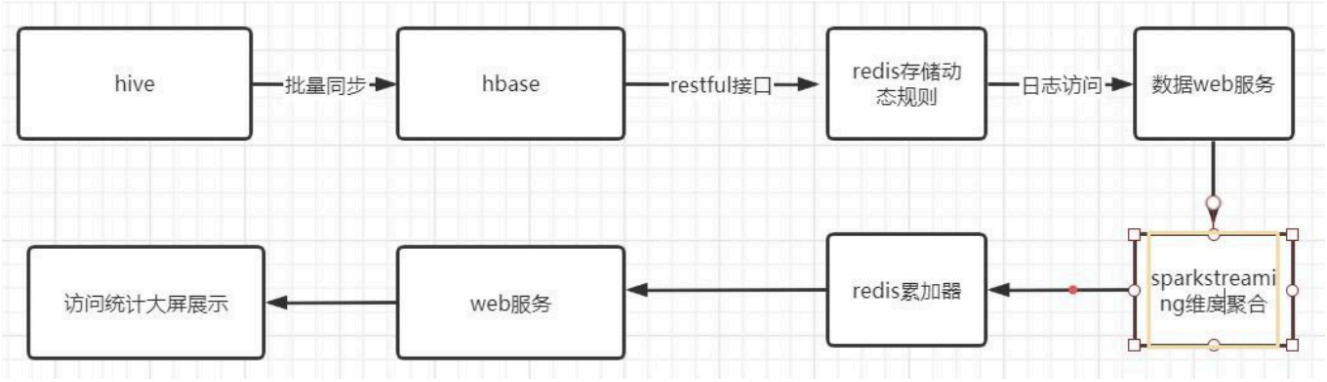


中国电信营销服务数据调用实时展示平台 2017

sparkstreaming kafka flume hbase hive redis

项目描述:

后台利用hive从中间表跑出结果宽表并批量打入hbase，利用redis做动态规则和用户权限设置，开发restful接口服务，设计restful接口规范。配置flume拦截器，将接口服务日志利用flume分别导入到kafka和hdfs。开发sparkstreaming程序根据业务指标分不同维度实时统计指标并封装为json存入redis，供前端实时调取展示.流程如下



电商受众监控平台

java kafka hbase spark hive

项目描述

该监控平台对目前主流的电商平台（京东，一号店，国美，苏宁等）在客户公司现有的销量，市场份额占比，地域分布，商品库等各种指标进行实时以及离线监控，使用hdfs，hive作为底层存储，使用分布式的kafka消息中间件对实时数据进行动态存储，使用sparkstreaming，spark进行各个维度指标实时统计和批处理统计。

主要工作是对当时利用路由采集的log日志，进行维度分析统计，供BI人员进行分析。

职责描述

数据处理代码编写

与产品经理配合，梳理统计口径

工作奖惩

2018-2019 亚信科技股份有限公司**优秀员工**

外派杭州期间,在浙江电信政企项目中及时解决未预估到的技术问题,从而保证项目进度正常完成.