**数据任务分发平台初步方案**

# 一、方案背景

# 二、平台开发

## 2.1 平台架构



数据流程任务分发平台.pdf

### 2.1.1工具

Azkaban —开源任务调度程序

### 2.1.2一级监控模块

从服务平台如DOService、POC、Operation等获取任务名单，以机构为单位形成任务名单表（读取逻辑发生改变，机构获取数据的优先程度可控）；

根据账户级别管理表获取机构级别（针对不同机构，数据处理流程差异可控），读取数据处理流程模型表，根据机构级别初始化对应的数据处理流程；

任务级别爬取优先级控制通过控制机构名称考虑。

### 2.1.3 数据获取模块

自营爬虫程序循环读取数据处理过程表，寻找是否有自己流程下已开始的任务分配，如果有已开始任务则获取该任务对应机构号，则去一级监控生成的任务名单表根据机构号获取该机构下所有企业，后台生成自己的爬虫名单（自主进行企业去重）。循环往复。数据爬取完毕生成供第三方去重的去重对照表，更新数据处理过程表状态，以便后续流程进行，同时进行数据入CB表的过程。

法海数据情况：循环读取数据处理过程表，寻找是否有自己流程下已开始的任务分配，在一级监控初始化名单时要进行注意同时开启爬虫爬取任务和法海诉讼任务。如果爬虫方面已经生成去重对照表，则进行去重清洗合并进CB表，更新流程状态表。

爬虫数据该机构下所有企业全部爬取完毕之后入CB表之后，再次更新流程表状态。

数据获取模块结束，进入数据加工模块。

### 2.1.4 数据加工模块

每个数据加工模块都是轮询数据处理过程表，查看是否有自己流程下的任务，有则获取机构号，开始任务。所有去重都放在自模块进行，企业名称作为各模块去重标准。

## 2.2 主要表结构设计

### 2.2.1 数据处理流程任务表

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程任务表 | | DataProcessTask | |
| 序号 | 中文名称 | 英文名称 | 备注 |
|  | 序列号 | SerialNo | 主键 |
|  | 机构号 | BankId |  |
|  | 模型ID | ModelId |  |
|  | 任务阶段 | TaskStage |  |
|  | 执行状态 | Status | init,running,success |
|  | 开始时间 | BeginTime |  |
|  | 结束时间 | EndTime |  |
|  | 发起人 | Operator | 记录平台操作发起人，若无法确定，则填写system |
|  | 通知对象 | Notifies | 该值来源于通知对象信息表，csrbjt库中已存在该表operation\_notification\_conf |
|  | 创建时间 | CreateTime |  |
|  | 备注 | Remark |  |
|  | 预留一 | Attribution1 |  |
|  | 预留二 | Attribution2 |  |
|  | 预留三 | Attribution3 |  |

### 2.2.2 数据处理流程模型表

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程模型表 | | DataProcessModel | |
| 序号 | 中文名称 | 英文名称 | 备注 |
|  | 序列号 | SerialNo | 主键 |
|  | 机构类型 | BankType |  |
|  | 机构号 | BankId | 配置会麻烦 |
|  | 模型ID | ModelId |  |
|  | 任务阶段 | TaskStage |  |
|  | 创建时间 | CreateTime |  |
|  | 备注 | Remark |  |
|  | 预留一 | Attribution1 |  |
|  | 预留二 | Attribution2 |  |
|  | 预留三 | Attribution3 |  |

## 2.3主要阶段定义

### 2.3.1 诉讼结构化产品主要阶段定义

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程主要阶段定义 | | | |
| 序号 | 阶段名称 | 具体说明 | 前置 |
| 1 | 任务初始化 | 一级监控根据机构初始化数据处理流程表。 | 无 |
| 2 | 诉讼数据采集 | 实时爬虫程序依据机构下的任务名单表中企业名单和指定爬取网站采集数据。 | 1 |
| 3 | 诉讼监控数据同步至生产 | 将采集诉讼监控数据同步至苏州生产库，并录入solr。 | 2 |
| 4 | 法海阶段一 | 根据名单获取列表概要信息。 | 1 |
| 5 | 法海阶段二(接口调用) | 根据solr对概要去重，调用详情入诉讼结构化表，并建立关联关系。 | 3,4 |
| 6 | 自有诉讼数据生成关联关系 | 根据名单检索solr和法海阶段二中的关联关系，生成自有诉讼数据关联关系明细表。 | 5 |
| 7 | 内容去重 | 后续会加入。根据去重明细表做内容去重。 |  |
| 8 | 生成运营任务 | 依据自有诉讼数据关联关系明细表生成运营任务。 | 6 |
| 9 | 数据运营结束 |  | 8 |

### 2.3.2被执行人产品主要阶段定义

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程主要阶段定义 | | | |
| 序号 | 阶段名称 | 具体说明 | 前置 |
| 1 | 任务初始化 | 一级监控根据机构初始化数据处理流程表。 | 无 |
| 2 | 被执行人数据采集 | 实时爬虫程序依据机构下的任务名单表中企业名单和指定爬取网站采集数据。 | 1 |
| 3 | 被执行人监控数据同步至生产 | 采集被执行人监控数据同步至苏州生产库。 | 2 |
| 4 | 法海阶段一 | 根据名单获取列表概要信息。 | 1 |
| 5 | 法海阶段二 | 根据已有被执行数据对概要去重，调用详情入被执行人表。 | 3,4 |

### 2.3.3失信被执行人产品主要阶段定义

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程主要阶段定义 | | | |
| 序号 | 阶段名称 | 具体说明 | 前置 |
| 1 | 任务初始化 | 一级监控根据机构初始化数据处理流程表。 | 无 |
| 2 | 失信被执行人数据采集 | 实时爬虫程序依据机构下的任务名单表中企业名单和指定爬取网站采集数据。 | 1 |
| 3 | 失信被执行监控数据同步至生产 | 采集失信被执行人监控数据同步至苏州生产库。 | 2 |
| 4 | 法海阶段一 | 根据名单获取列表概要信息。 | 1 |
| 5 | 法海阶段二 | 根据已有失信被执行人对概要去重，调用详情入失信被执行人表。 | 3,4 |

### 2.3.4舆情预警产品主要阶段定义

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程主要阶段定义 | | | |
| 序号 | 阶段名称 | 具体说明 | 前置 |
| 1 | 任务初始化 | 一级监控根据机构初始化数据处理流程表。 | 无 |
| 2 | 舆情数据采集 | 实时爬虫程序依据机构下的任务名单表企业名单和指定爬取网站采集数据。 | 1 |
| 3 | 舆情监控数据同步至生产 | 采集数据同步至苏州生产库，录入solr。 |  |

### 2.3.5信用报告V4-金融办风险挖掘报告产品主要阶段定义

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程主要阶段定义 | | | |
| 序号 | 阶段名称 | 具体说明 | 前置 |
| 1 | 任务初始化 | 一级监控根据机构初始化数据处理流程表。 | 无 |
| 2 | 监控任务名单表名单生成 | 依据机构下的任务名单表中企业名单调用工商关联关系数据（含主体企业、一层关联企业、二层关联企业，不含三层关联企业）接口，生成含主体企业、一层关联企业、二层关联企业，不含三层关联企业的任务名单表数据，并做诉讼结构化、被执行、失信被执行、舆情预警等数据处理流程的任务初始化。 | 1 |

## 2.4人员职责涉及

### 2.4.1 平台开发组

王军：一级监控模块开发，更改企业名单监控表表结构，以机构号 +企业唯一确定一条记录。对于某个机构下的所有待查企业，根据模型表在流程任务表中初始化生成任务，初始化各阶段任务状态；

**以苏州银行为例：**

|  |  |  |  |
| --- | --- | --- | --- |
| 监控任务名单表 | |  | |
| …… | BankId | EntName | …… |
|  | SZBank | 企业A |  |
|  | SZBank | 企业B |  |
|  | SZBank | 企业C |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程模型表 | | DataProcessModel | |
| …… | BankType/BankId | TaskStage | …… |
|  | 一级/SZBank | 法海：获取概要信息 |  |
|  | 一级/SZBank | Solr：生成去重对照表 |  |
|  | 一级/SZBank | 法海：去重并获取详情 |  |
|  | 一级/SZBank | Solr：数据入C表 |  |
|  | 一级/SZBank | 平台：生成关联关系 |  |
|  | 一级/SZBank | 平台：生成运营任务 |  |
|  | …… | …… |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 数据处理流程任务表 | | DataProcessTask | |
| …… | BankId | TaskStage | …… |
|  | SZBank | 法海：获取概要信息 |  |
|  | SZBank | Solr：生成去重对照表 |  |
|  | SZBank | 法海：去重并获取详情 |  |
|  | SZBank | Solr：数据入C表 |  |
|  | SZBank | 法海：详情入CB表 |  |
|  | SZBank | 平台：生成关联关系 |  |
|  | SZBank | 平台：生成运营任务 |  |
|  | …… | …… |  |

陈晓平：

### 2.4.2 应用交付组

段宪锋：负责法海数据模块的接入，概要信息获取阶段、概要信息去重阶段，在Azkaban上任务分发平台的搭建；定时读取任务表信息，查看是否有已经开始的自己阶段上的任务，如果存在，则启动任务，结束之后，对应任务状态置成success，并将下阶段任务设为已开始（具体可通过Azkaban任务调度）。

### 2.4.3 数据爬虫组

严晨：负责爬虫模块更改；定时读取任务表信息，查看是否有已经开始的自己阶段上的任务，如果存在，则启动任务，结束之后，对应任务状态置为success，并将下阶段任务设为已开始（具体可通过Azkaban任务调度）。