# Elise采集打点统计概要设计

|  |  |  |  |
| --- | --- | --- | --- |
| 版本号 | 作者 | 日期 | 说明 |
| V1.0 | 聂嘉文 | 2017-10-27 | 初始版本 |
|  |  |  |  |

## 1、背景与目的

采集任务的进度信息不明，处在”黑盒子”中，运维或者开发很难掌握采集任务的采集情况。

因此在采集的过程中记录一些统计信息，可供运维系统查询，前端展示，方便知晓采集任务的进度信息。

## 2、数据结构设计与存储

## 2.1 数据存储

Redis能满足实时统计的需求，且采集任务也存储在Redis中，因此统计信息也存储在Redis。

## 2.2 数据结构

### 2.2.1 方案一

数据采用Redis里的Hash，Key-Value，Value为field-fieldValue的形式。

Key的名称格式：system\_childesystem\_module\_status。

统计信息表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Key | Field | Value | 说明 |
| 1 | elise\_crawler\_crawl\_success | taskId | 采集数量 | 采集成功的数量 |
| 2 | elise\_crawler\_crawl\_failed | taskId | 采集数量 | 采集失败的数量 |
| 3 | elise\_crawler\_download\_success | taskId | 下载数量 | 下载成功的数量 |
| 4 | elise\_crawler\_download\_failed | taskId | 下载数量 | 下载失败的数量 |
| 5 | elise\_crawler\_extract\_success | taskId | 抽取数量 | 抽取成功的数量 |
| 6 | elise\_crawler\_extract\_failed | taskId | 抽取数量 | 抽取失败的数量 |
| 7 | elise\_crawler\_pipeline\_success | taskId | 持久化数量 | 持久化成功的数量 |
| 8 | elise\_crawler\_pipeline\_failed | taskId | 持久化数量 | 持久化失败的数量 |
| 9 | elise\_crawler\_taskschdule | taskId | 总任务下载数量 |  |

### 2.2.2 方案二

数据采用Redis的String，Key-Value形式；

Key的名称格式：system\_childesystem\_module\_status\_taskId。

统计信息表：

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Key | Value | 说明 |
| 1 | elise\_crawler\_crawl\_success\_taskId | 采集数量 | 采集成功的数量 |
| 2 | elise\_crawler\_crawl\_failed\_taskId | 采集数量 | 采集失败的数量 |
| 3 | elise\_crawler\_download\_success\_taskId | 下载数量 | 下载成功的数量 |
| 4 | elise\_crawler\_download\_failed\_taskId | 下载数量 | 下载失败的数量 |
| 5 | elise\_crawler\_extract\_success\_taskId | 抽取数量 | 抽取成功的数量 |
| 6 | elise\_crawler\_extract\_failed\_taskId | 抽取数量 | 抽取失败的数量 |
| 7 | elise\_crawler\_pipeline\_success\_taskId | 持久化数量 | 持久化成功的数量 |
| 8 | elise\_crawler\_pipeline\_failed\_taskId | 持久化数量 | 持久化失败的数量 |
| 9 | elise\_crawler\_taskschdule\_taskId | 总任务下载数量 |  |

### 2.2.3 方案选择

方案一比方案二方便统计所有任务的总量信息；

已知方案二比方案一在Redis事务方面(并发操作)更容易实现一下，待实际测试后进行选择；

## 3、流程图

在采集的各个关键点增加统计信息，流程图和采集流程类似，略；