云采集技术白皮书

深圳比一比网络科技有限公司

20170524

提要

- 本白皮书阐述云采集的应用场景
- 本白皮书阐述云采集的技术特性
- 本白皮书阐述云采集的系统介绍
- 本白皮书阐述云采集的系统部署

目录

1. 概述

随着网络的迅速发展，互联网成为大量信息的载体，如何能够获取全面的数据，是件极其重要且不容易的事情。要真正做好大数据时代的分析，仅仅靠企业内部的数据是远远不够的，还需要借助互联网上的数据资源。从互联网上爬取数据资源，成为了非常关键的一环。

云采集，即是一款网络爬虫系统软件，主要是面向互联网及移动网络进行数据抓取。在本白皮书中，将从云采集系统的应用场景、技术特性、系统介绍、应用部署几个方面，对云采集系统进行全面的阐述。

1.1 什么是爬虫

网络爬虫（又被称为网页蜘蛛，网络机器人，在 FOAF 社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。

1.2 什么是正则表达式

又称规则表达式。（英语：Regular Expression，在代码中常简写为 regex、regexp 或 RE），计算机科学的一个概念。正则表通常被用来检索、替换那些符合某个模式(规则)的文本。

2. 云采集的应用场景

2.1 静态互联网数据抓取

静态抓取主要是对海量网络公开数据，进行抓取，将数据进行统一存储，便于企业对数据进行统一的操作处理，让网络的沉默数据，变成企业可以使用的有价值的数据。

2.2 动态互联网数据抓取

动态抓取顾名思义是对动态的网络数据进行抓取，这些数据会因为时间的不同而发生变化。云采集会根据不同的需求，按照设定的时间周期，对数据进行抓取，从而为企业累积有价值的数据，进行商业分析。

2.3 移动数据抓取

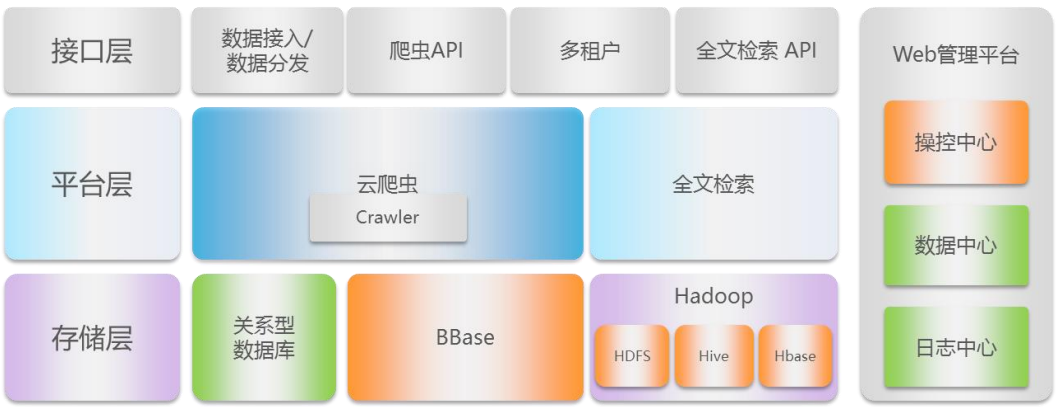
目前中国手机网民超过 7 亿，越来越多的商业活动通过手机应用达成，很多商家的手机活动和 PC 活动有很大的区别，手机端数据也酝酿着越来越大的商业力量蓄势待发。而手机端数据有壁垒高、难抓取、时间短的特性。云采集支持对移动端网页及 APP 数据抓取，

3. 云采集的技术特性

3.1 云采集的主要技术特性：

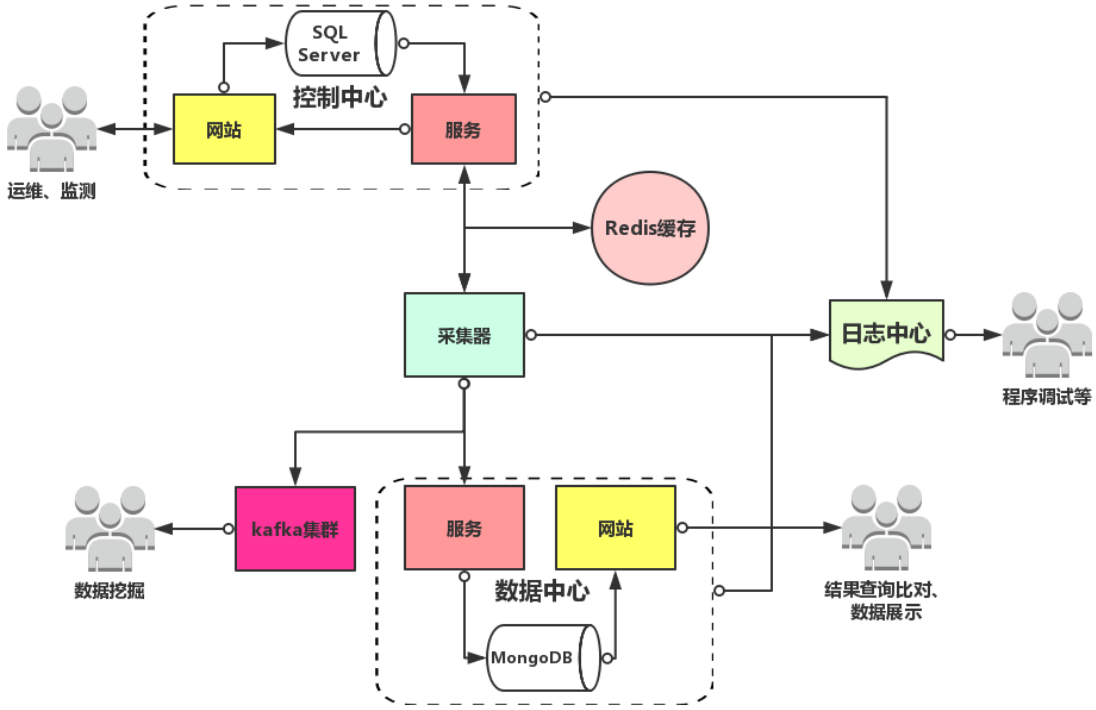
- 适用性：即可以采集 PC 数据，也可以采集 APP 数据
- 可靠性：可靠性达 99.99%，业内平均不间断运行时间为 72 小时
- 大吞吐：每天可采集一亿 URL

3.2 云采集的架构图



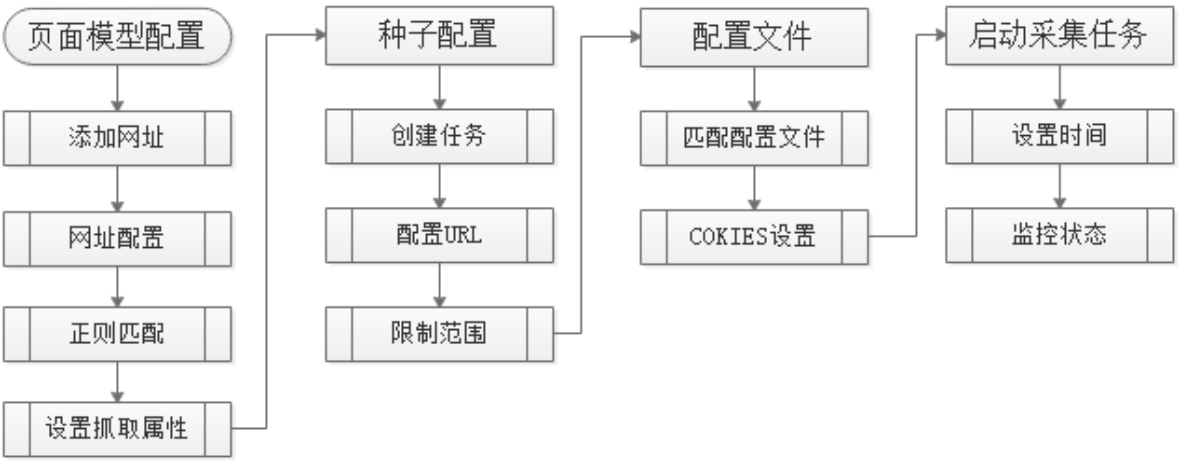
4. 云采集的系统介绍

“一套软件、两种语言、三个中心”，简短的一句话，清楚的表述了云采集系统的基本特征。三个中心指构成系统软件的三个结构模块，分别是：操控中心、数据中心、日志中心。两种语言指云采集系统，包括 JAVA 版和.NET 版两个版本，客户可以根据情况灵活选择。



4.1 操控中心

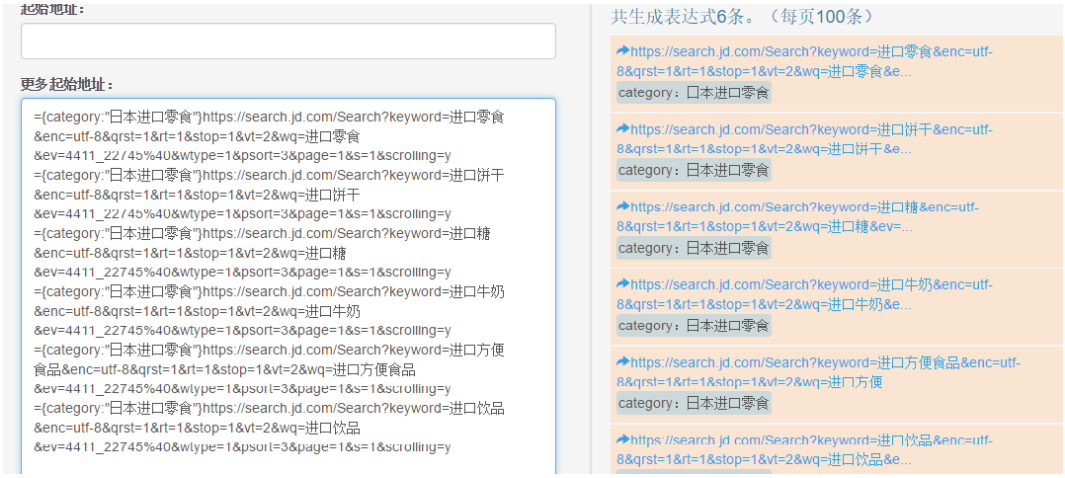
操控中心任务配置整体流程图：



操控中心功能特性：

1) 支持多起始地址

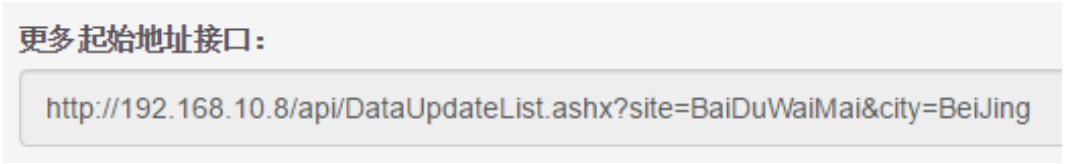
支持一次性配置多个采集的链接地址



2) 支持 API 接口作为起始地址

API 接口中，可以包括多种格式的 URL 地址，可以提高灵活度，减少配置复杂性。

应用场景：百度外卖-店铺详情-北京



3) 支持增量不为 1 的翻页

支持网站的分页接口不是按照页码参数做+1 进行递增，而是根据数据偏移量以页面大小做递增。

应用场景：饿了么商铺列表数据

https://mainsite-restapi.ele.me/shopping/restaurants?extras[]=activities&latitude=22.5483&longitude=113.94444&limit=24&offset=0

该接口参数 limit 表示页面大小， offset 表示数据偏移量，使用将 offset 每次增加 24 的方式实现翻页。

4) 支持探索式翻页

应用场景：百度外卖店铺列表页：

http://waimai.baidu.com/waimai/shoplist/2364115e829cbc29?taste=68&display=json&page=1&count=40

该接口获取的是某地点第一页的店铺列表，分页大小是 40。云采集会根据配置进入下一页（page 参数的值做+1），直到获取到的店铺数量小于 40

5) 支持对要采集的 URL 做预处理

该功能将收集到的 URL 处理成标准格式，可以预防对内容页面进行重复采集，以及预防采集路径进入死循环，最终只能采集到少量的重复数据。

应用场景：当当网产品详情页

http://product.dangdang.com/1265277530.html#ddclick?act=click&pos=1265277530_0_1_m&cat=4010476&key=&qinfo=&pinfo=&minfo=122_1_58&ninfo=&custid=&permid=2017 list

等价于

http://product.dangdang.com/1265277530.html

6) 支持在起始地址中添加自定义属性

在采集四大外卖平台数据时，通过在起始地址中，对起始地址做若干标记，云采集在整个采集过程中保持该标记，以便于收到数据时，获取业务上需要的信息。

应用场景：四大外卖平台城市信息

更多起始地址：

```
={city:"北京市",point:"苹果园地铁站"}http://i.waimai.meituan.com/mti/home?lat=39.926502&lng=116.177611&category_type=21&category_text=鲜果购
={city:"北京市",point:"古城地铁站"}http://i.waimai.meituan.com/mti/home?lat=39.907419&lng=116.190147&category_type=21&category_text=鲜果购
```

通过添加自定义属性{city:”北京市”,point:”苹果园地铁站”}，使采集到的数据与自定义属性相匹配，此例中可实现数据与城市的直接关联。

7) 支持事务型采集 Step

在采集过程中，很多情况是 FS 模式的， 即只有采集了前面的页面，才能够采集后面的页面，实现按照事务顺序进行采集，能够对采集路径进行精准控制。

应用场景：拉勾网公司列表

更多起始地址：

```
(=https://www.lagou.com/gongsi/),(=https://www.lagou.com/jobs/list_江西博莱农业高科技股份有限公司?cl=false&fromSearch=true&labelWords=&suginput=)
(=https://www.lagou.com/gongsi/),(=https://www.lagou.com/jobs/list_苏州黑盾环境股份有限公司?cl=false&fromSearch=true&labelWords=&suginput=)
```

每个括号中，表示一项采集任务，每个任务中，通过“,”分隔为多个事务，通过确定事务的顺序，完成对目标数据的采集。

8) 支持 JS 注入

云采集支持在采集到的页面中，使用 JS 执行加载数据，修改页面内容等行为

应用场景：美团外卖 • 店铺列表页

页面下载完成后执行JS：

```
var AllCompleted = false;
//var maxPageCount = 100;
var classify_type = 21;//
var page_offset = 1;
//var categoryString = window.location.href.match("category_type=\\d+&category_text=[^&]+")[0];

var shopsOl = document.createElement("ol");
shopsOl.id = "zaneList";
shopsOl.className="root";
document.getElementsByClassName("page-header")[0].insertBefore(shopsOl,null);
```

9) 支持变身

可以设置能换 IP 的采集器对采集中词进行采集变身。

应用场景：天眼查任务采集

在采集过程中，目标网站会采取一些拦截措施，阻止爬虫进行采集，通过变身术，变换 IP 可以有效应对目标网站防采。

10) 支持采集配置

即可以进行采集配置设置，一方面控制采集的最大时长、最大采集次数、最大深度，最大失败次数。另一方面如果出现防采情况，可以通过采集配置，调整采集间隔时间和采集线程，应对防采。

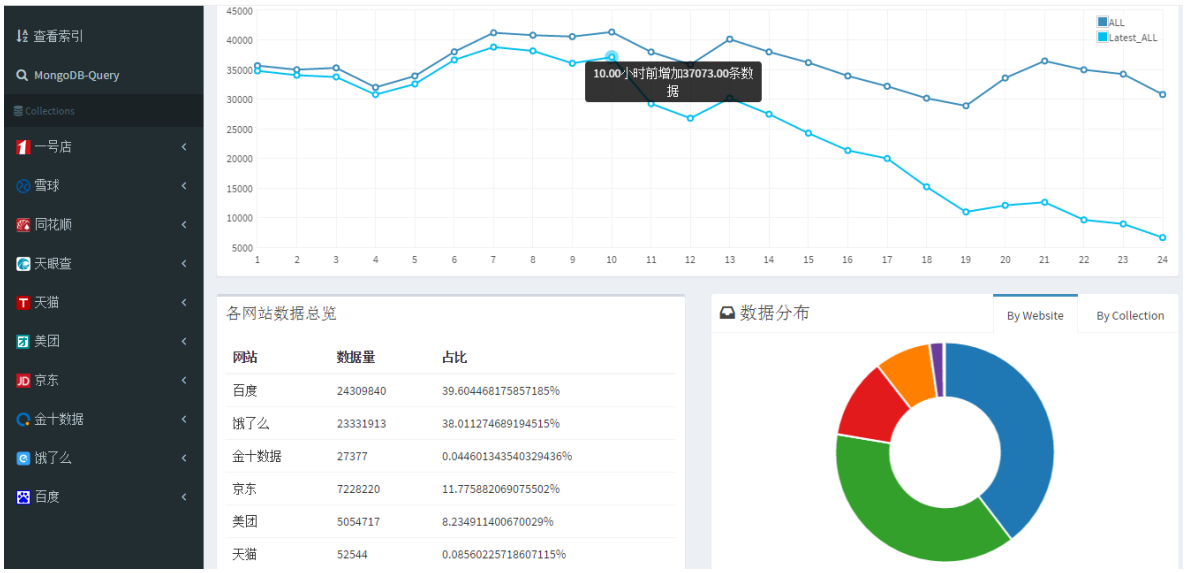
应用场景：美团外卖-店铺详情页

11) 支持单任务 HTTP 设置

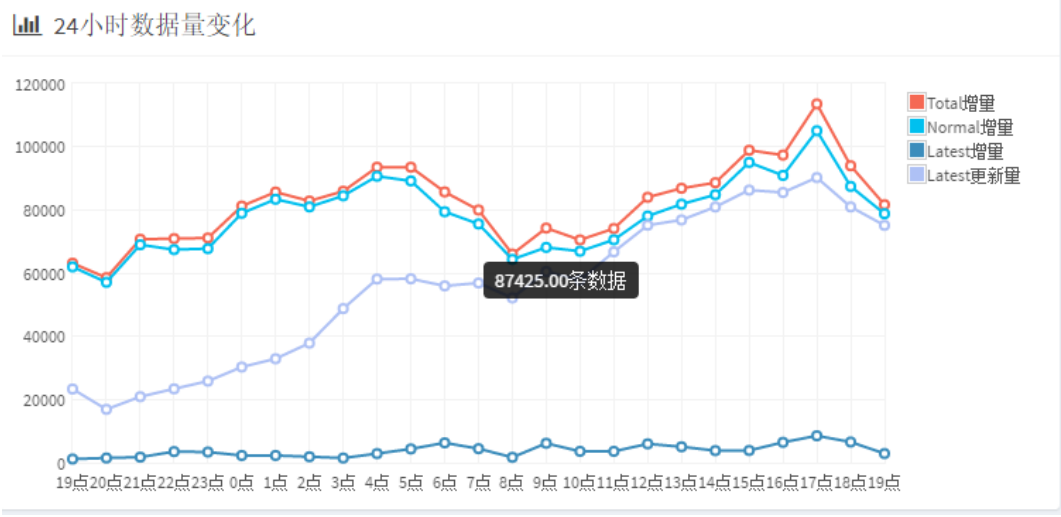
HTTP 设置包括设置 HTPP 请求的超时时长、设置 HTTP 失败时，重试前的等待时间、设置 HTTP 失败时，允许的重试次数、设置可接受的最大 HTTP Size、以及设置是否使用公共 Cookie。

应用场景：金十数据-快讯

4.2 数据中心



1) **支持 24 小时数据变化监控**



- 通过此图能看到过去 24 小时数据量的变化
- 每个时间节点统计的数据都是该节点到下一个节点之间的数据量
- Total 增量：所有数据集 24 小时数据量变化
- Normal 增量：所有普通数据集 24 小时数据量变化
- Latest 增量：所有 Latest 数据集 24 小时数据量变化
- Latest 更新量：所有 Latest 数据集 24 小时数据量的更新情况

2) 支持各网站数据总览

各网站数据总览

数据的数量（单位：条）

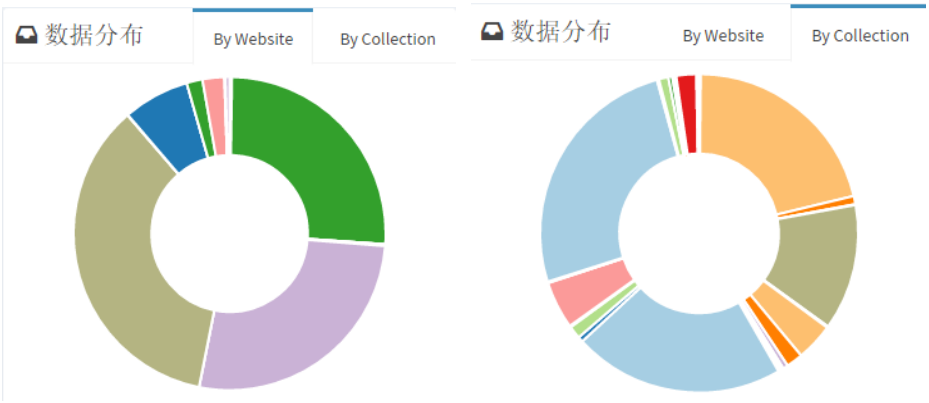
网站 ← 网站

占比 ← 当前网站的数据量在总数据量中所占的百分比

网站	数据量	占比
kaoshi100	4	0.00007602744438687477%
百度	1191830	22.65294726090224%
饿了么	2052414	39.0099478109608%
金十数据	100	0.0019006861096718694%
京东	793560	15.083084691912086%
看准网	68	0.0012924665545768711%
拉勾网	10528	0.2001042336262544%
美团	865833	16.456767563955236%
天猫	67424	1.2815186025851613%
天眼查	215494	4.095864525176298%
同花顺	9136	0.173646682979622%
雪球	8727	0.16587287679106405%
一号店	46026	0.8748097888375745%

- 显示各网站当前的数据量
- 显示各网站总数据量占比

3) 支持数据分布预览



- 支持按照网站进行分类的数据分布查询

- 支持按照集合进行分类的数据分布查询

4) 支持集合详情页查询

- 支持支持查看所有采集到的数据和采集过程的相关记录。如上图，一号店集合详情包括：一号店团购商品详情页、一号店商品详情页、一号店类别商品列表页、一号店评论页。

4.3 日志中心

- 支持网站模型异常监控，监控内容包括：采集器、采集状态、异常信息说明。
- 支持配置种子异常监控，监控内容包括：下载总数、抽取总数、命中缓存、抽取比例、下载成功比、抽取成功比、开始时间。
- 支持采集网址异常监控，监控内容包括：下载失败、不满足页面模型、JSON 解析失败、JSONP 解析失败、JS 执行异常。

5. 云采集的系统部署

5.1 数据中心：

1) 安装、配置并启动 MongoDB。
2) 将 Beyebe.DataHub.RemotingLauncher（数据中心-服务）生成的文件复制到本地，将 MongoDBAddress 设置为第一步中设置的 MongoDB 连接字符串；将 RemotingPort 设置为一个固定端口值，此为数据中心向外提供服务的端口；若有最新数据中心的功能，则将 MongoDBAddress_Temp 设置为相应 MongoDB 数据库的连接字符串，否则将该行注释掉。
3) 将 Beyebe.DataHub.MainSite（数据中心-网站）生成的文件复制到本地，将 MongoDBAddress 设置为 MongoDB 数据库的连接字符串；将 ControllerRemotingAddress 设置为控制中心服务的地址。将 UserName 和 Password 设置好。通过 IIS 服务部署该网站。

5.2 控制中心：

1) 安装并配置 SQL Server。
2) 将 Beyebe.DataHub.Controller.RemotingLauncher（控制中心-服务）生成的文件复制到本地，将其配置文件中的 DataBaseConnectionString 连接字符串改为第一步设置好的 SQL Server 连接字符串；将 RemotingPort 设置为一个固定端口值，此为控制中心向外提供服务的端口；将 DatahubRemotingAddress 字段设置为数据中心服务的地址。
3) 将 Beyebe.DataHub.Controller（控制中心-网站）生成的文件复制到本地，将

其配置文件中的 DataBaseConnectionString 连接字符串改为第一步设置好的 SQL Server 连接字符串；将 DatahubRemotingAddress 字段设置为数据中心服务的地址；将 ControllerRemotingAddress 设置为第二步中设置的控制中心服务的地址。通过 IIS 服务部署该网站。

5.3 Redis 缓存：

1) 安装、配置并启动 Redis。
2) 将 Beyebe.DataHub.CacheHub.RemotingLauncher 生成的文件复制到本地，将 RedisAddress 设置为 Redis 所在 IP 地址；将 RemotingPort 设置为一个固定端口值，此为缓存向外提供服务的端口。

5.4 采集器：

1) 将上述服务和网站都启动。
2) 将 Beyebe.BCrawler.Bee 生成的文件复制到本地，将 ClientName 设置为当前采集器的名字；将 ControllerRemotingAddress 设置为控制中心服务的地址；将 DatahubRemotingAddress 设置为数据中心服务的地址；将 CachehubRemotingAddress 设置为缓存服务的地址。
3) 启动。