普林科技云爬虫 v1.0

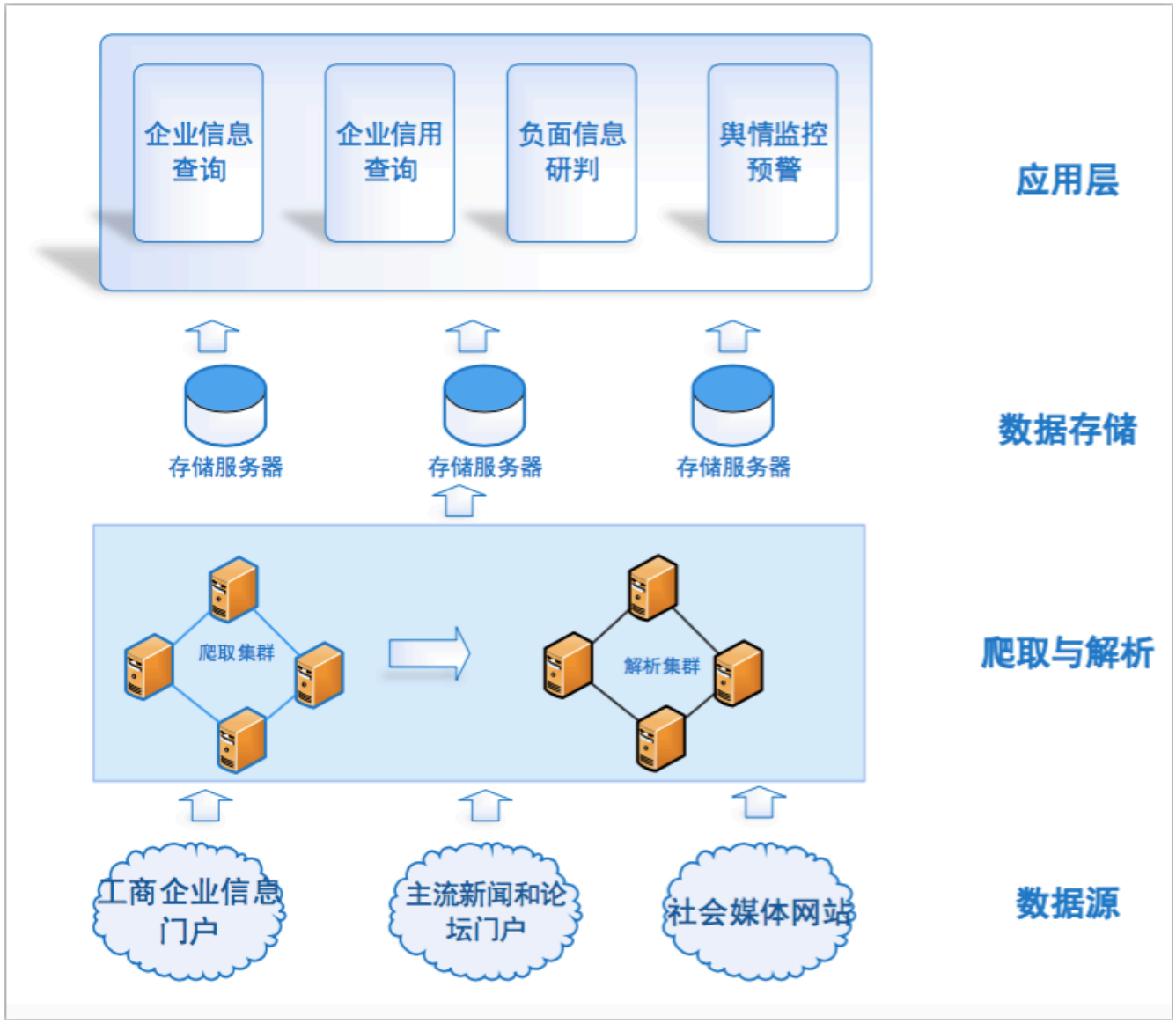
分布式可水平扩展的爬虫集群

北京至信普林科技有限公司

2015.09.1

目录

*云爬虫系统框架一览图


企业信息查询
企业信用查询
负面信息研判
舆情监控预警
应用层
存储服务器
存储服务器
存储服务器
数据存储
爬取集群
解析集群
爬取与解析
工商企业信息门户
主流新闻和论坛门户
社会媒体网站
数据源

1.系统架构

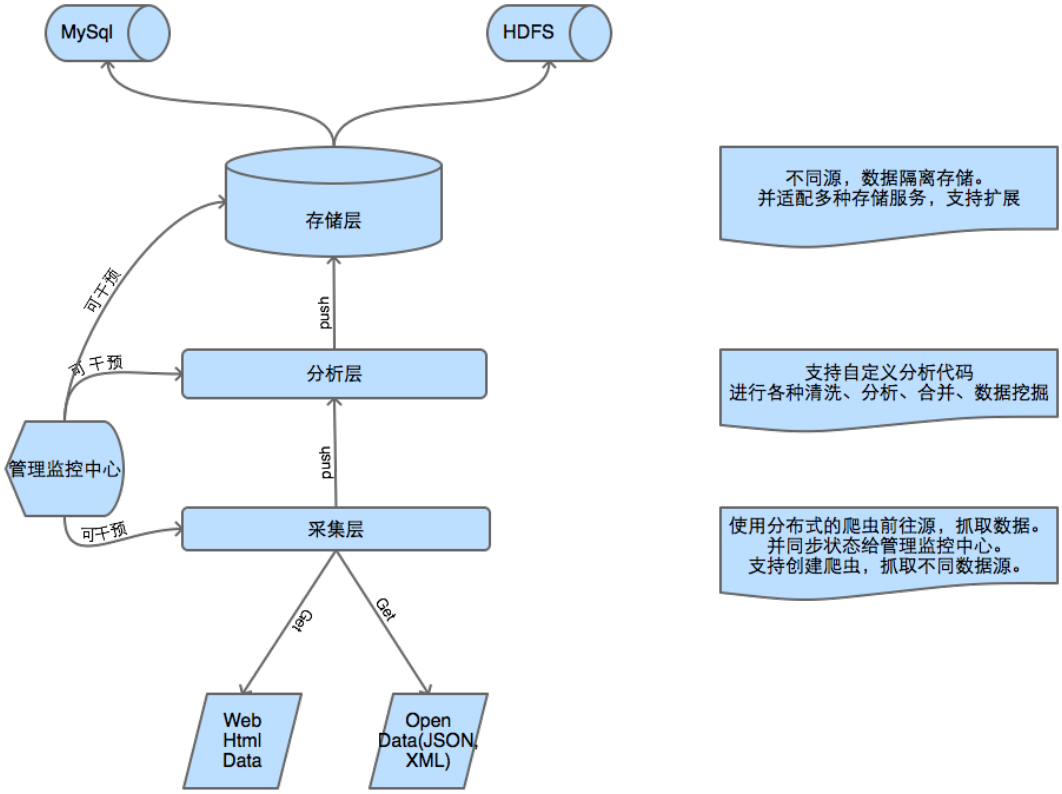
云爬虫使用业内最为成熟的、分布式、可水平扩展方案设计完成。能够支持上***P***级数据抓取、分钟级的更新粒度，并且允许开发者增加特殊插件。

采用了分层架构设计，能够降低系统的复杂度，并且提升了系统的稳定性：

- 采集层。主要支持***HTTP***、***HTTPS***协议，原始数据自动隔离。
- 分析层。允许自定义分析代码，主要支持***Python***开发
- 存储层。支持***HDFS***、***MySQL***等主流存储服务器

为了监控控制系统行为，增加了

- 管理监控中心。可以干预采集层、分析层、存储层的行为

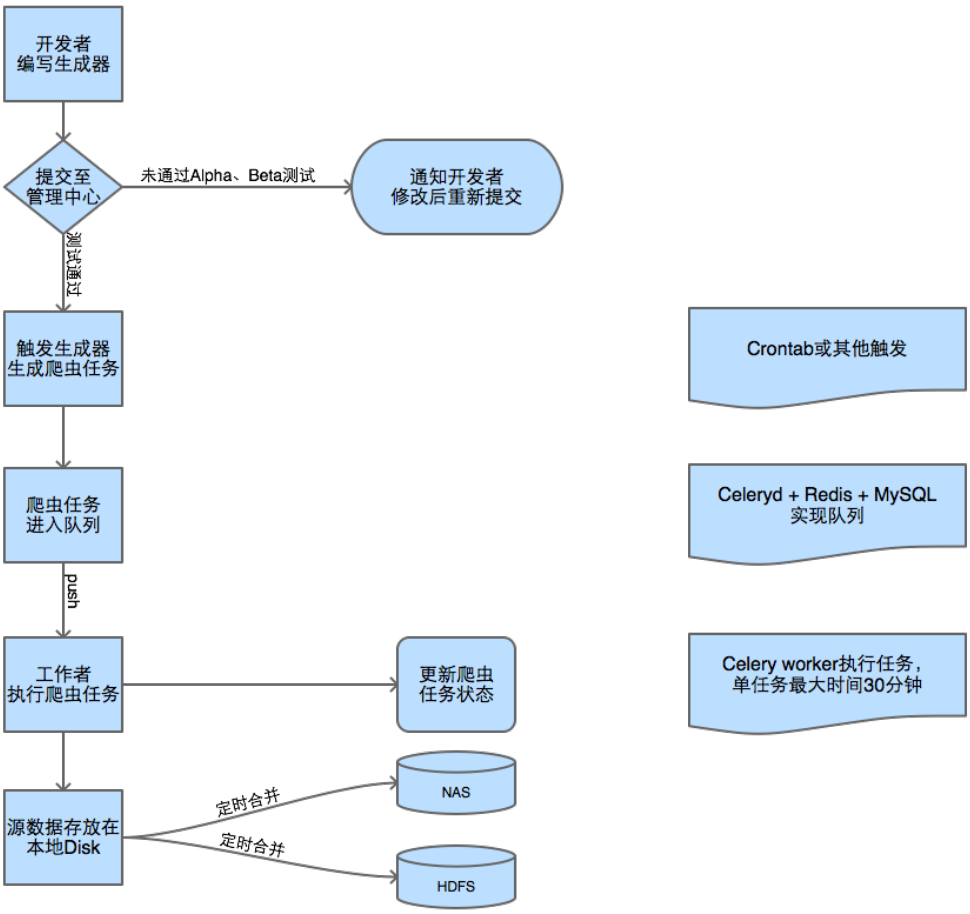

MySql
HDFS
存储层
push
可干预
可 干预
分析层
push
管理监控中心
可干预
采集层
Get
Get
Web
Html
Data
Open
Data(JSON,
XML)
不同源，数据隔离存储。
并适配多种存储服务，支持扩展
支持自定义分析代码
进行各种清洗、分析、合并、数据挖掘
使用分布式的爬虫前往源，抓取数据。
并同步状态给管理监控中心。
支持创建爬虫，抓取不同数据源。

1.1 工作流

主要分如下几个大流程：

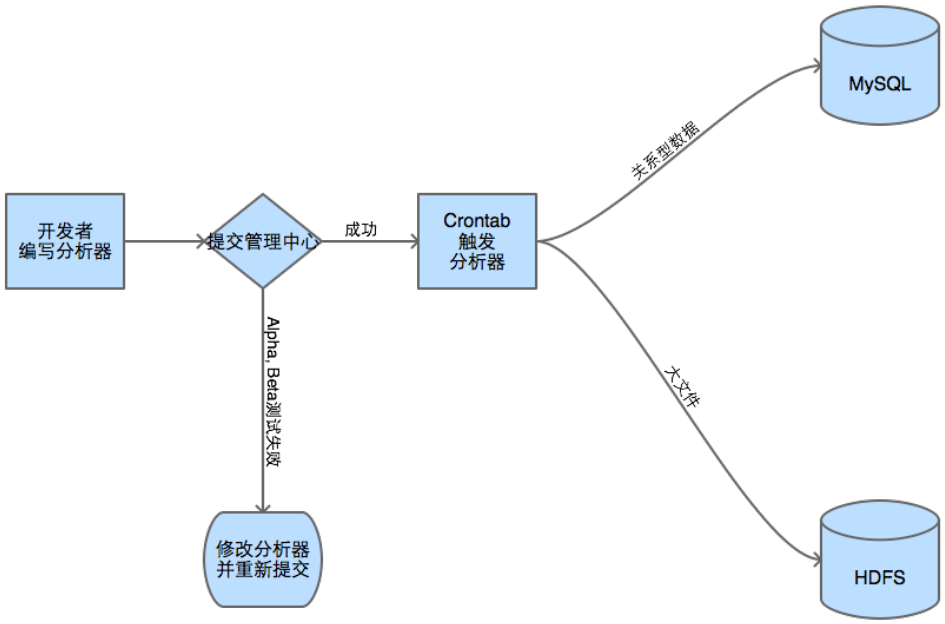
- 采集工作流。管理员创建爬虫，并提交任务生成器。***Crontab***驱动爬虫任务生成器，使用队列服务器统一调度，队列***worker***执行任务。

采集层工作流



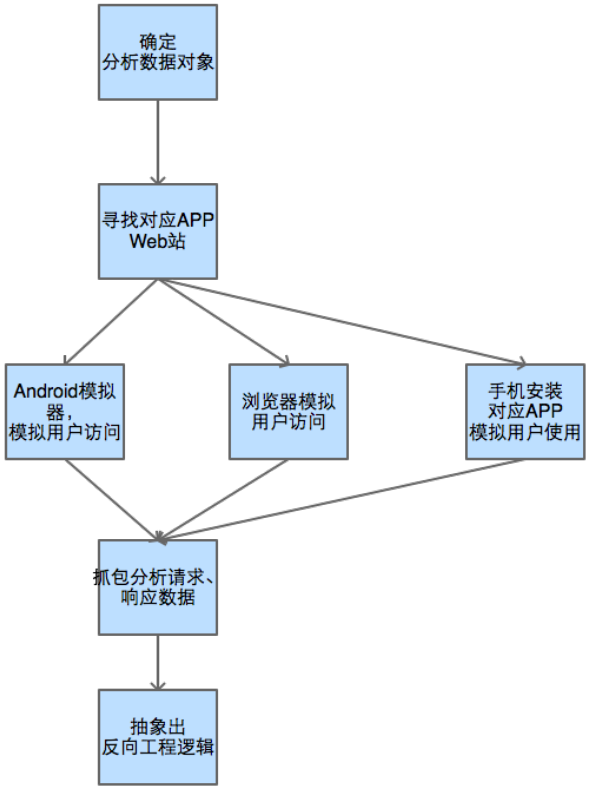
- 分析层工作流。

分析层工作流



- ***ETL***预研。需要提前分析目标对象，制定反向工程逻辑。

ETL预研


确定
分析数据对象
寻找对应APP
Web站
Android模拟
器，
模拟用户访问
浏览器模拟
用户访问
手机安装
对应APP
模拟用户使用
抓包分析请求、
响应数据
抽象出
反向工程逻辑

1.2 系统拓扑硬件要求&成本

系统拓扑硬件主要分为：

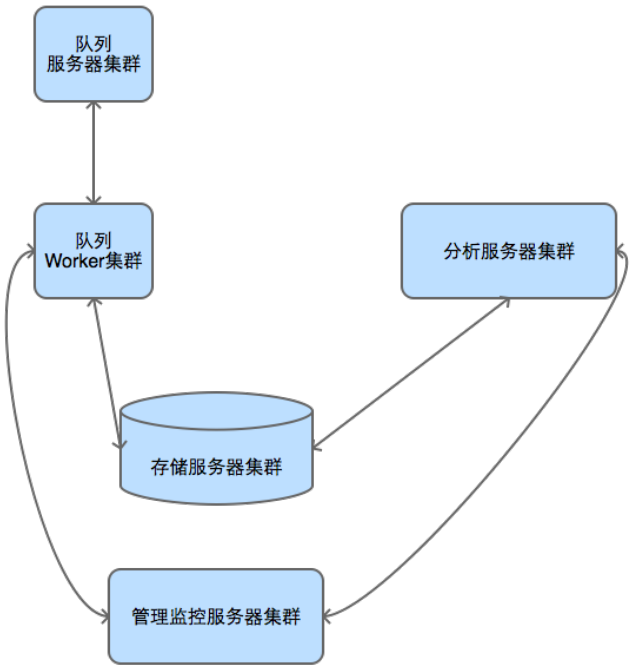
	分钟粒度	小时粒度	天粒度
队列服务器	2	1	1
队列 **worker**	6	4	2
存储服务器	6	4	2
分析服务器	6	4	2
管理监控服务器	2	1	1
总计数量	22	14	8

硬件成本核算

	分钟粒度	小时粒度	天粒度
队列服务器	17	8.5	8.5
队列 **worker** 服务器	51	34	17
存储服务器	60	40	20
分析服务器	51	34	17
管理监控服务器	17	9	8.5
安装费用	58.8	37.5	21.3
维护费用	—	—	—
总计（万元）	254.8	162.5	92.3

机器托管&带宽成本核算

	分钟粒度	小时粒度	天粒度
BGP 机房	Y	Y	Y
机柜数	2	2	1
带宽大小（**Mb**）	50	20	10
带宽费用（元）	—	—	—
托管费（万元）	16	16	8
城市	北京	北京	北京
时长（月）	12	12	12
总计（万元）	**16**	**16**	**8**


队列
服务器集群
队列
Worker集群
分析服务器集群
存储服务器集群
管理监控服务器集群

2.爬虫产品介绍

爬虫架构支持分布式部署，支持单机多线程、多进程运行，大大提高爬虫的可扩展性。独特的用户定制下载功能能够根据用户需求针对性地下载数据。数据采集种类全面：涵盖企业信用信息、国内主流新闻门户、主流论坛网站和主流的微博系统。

2.1 数据覆盖范围

2.1.1 企业信用信息

包括企业基本信息、资本信息、股东信息、变更信息等。

注册号	住所
企业名称	营业期限起始日期
企业类型	营业期限结束日期
注册资本	实缴出资金额
法定代表人	经营范围
成立日期	登记机关
核准日期	登记状态

2）股东信息

股东类型	证件类型
股东姓名	证件号码

3）变更信息

变更事项	变更后内容
变更前内容	变更日期

2.1.2 主流新闻网站

国内的主流新闻网站是系统新闻和评论的主要数据来源。这些网站包括新浪新闻、搜狐新闻、凤凰网新闻、网易新闻等。针对每一个网站制定该网站专用的下载模板。以一定频率从上述新闻网站抓取新闻和评论，将抓取到的新闻和评论与已下载的新闻和评论进行对比消重，将没有重复的内容存入数据库的新闻表和评论表。

新闻网站主要采集两类信息：

1）新闻文章信息

新闻 ID	下载时间

新闻标题	点击数
新闻发布时间	评论数
新闻内容	发布者
新闻来源	

2）新闻的评论信息

评论 ID	评论内容
新闻 ID	评论时间
评论者	

2.1.3 主流论坛网站

雪球论坛、新浪 BBS、搜狐 BBS 等主要论坛网站，更多的论坛网站可以按需添加。论坛网站主要采集两类信息：

1）论坛主贴

主贴 ID	下载时间
主贴标题	点击数
发布时间	回帖数
主贴内容	发布者
主贴来源	

2）论坛回帖

回帖 ID	回复内容
主贴 ID	回复时间
回复者	

2.1.4 微博数据

微博有着十分丰富的用户基础数据，包括分享信息，好友信息，标签信息等都可以让我们更多的了解用户的情绪、态度，对传播领域中舆论方向的把控以及了解用户更多的特征。新浪微博每日活跃用户数超过 5000 万，是规模巨大的数据产生源，所以这里我们主要针对新浪微博数据的抓取，在微博系统中能够抓取到的信息主要有三种：

1）用户基本信息

用户 ID	微博数
微博昵称	关注数
省份	收藏数
城市	创建时间
地址	是否加 V
个人描述	认证类型

用户性别	所在公司
粉丝数	职业信息
兴趣标签	

2）微博信息

发布者 ID	转发数
发布者昵称	评论数
微博 ID	转发微博 ID
微博发布时间	微博来源
微博内容	

3）用户关注关系信息

用户 ID	关注好友列表

2.2 产品介绍

2.2.1 企业信用查询

企业资产查询，企业股东信息查询，企业股东变更信息查询，企业负债查询等。

2.2.2 企业互联网舆情监控与预警

1）企业互联网报道查询

功能描述：统计企业在不同时间段在各大数据来源的报道热度。提供多种查询条件，检索字段包括时间，关键词、来源、用户名等。

2）负面信息自动研判

功能描述：根据负面情感词库，使用自然语言处理技术构建负面新闻研判模块，自动识别负面新闻报道、负面论坛讨论、负面微博信息。

报道标题	负面
中信证券骨干“陨落”	负面
中信证券遭遇空前“难堪”	负面

3）企业负面信息监控与预警

功能描述：根据提供的企业名称（“关键词”），实时下载各大主流新闻、论坛和微博平台的相关信息，支撑负面信息和重点用户的在线监控和预警。