Try to grab recruitment data from lagou

本人於年前立下目標，在Linux運維培訓結束之前利用已學知識寫一個爬蟲，抓取招聘網站(拉勾)的職位信息，爲自己求職應聘做準備。結業在即，本人花了將近4天時間來實現該需求。最終是通過Bash Shell、PHP、MariaDB實現。爲了達成這個目標，本人重新拾起已經半年多未用過的PHP。開發期間遇到很多技術問題，大部分都成功解決，少部分採用折衷方案。

本想寫一個Web頁面，用於進行數據展示、篩選，暫時擱置吧。畢竟重新拾起HTML、CSS、jQuery、Bootstrap這些東西需要時間，而現在最缺的就是時間。

已經代碼上傳至GitHub

Table Of Contents

Thinking Processes

開始是想寫一個爬蟲，用Golang或PHP，但是Golang自學還沒學會，PHP已經忘的差不多，故想通過Bash實現。使用`curl`抓取頁面，通過`sed`, `awk`, `grep`等工具提取所需數據。

分析完網頁結構，使用curl抓取首頁，發現抓回的是模版頁面，沒有具體數據。後在QQ羣中求教，有人提示使用開發者工具查看網站數據調用方式。經過排查測試，找到了調用數據的API，並通過測試得到了需要的搜索參數。

該API返回的是`json`格式數據，而本人不知如何通過Bash Shell來解析`json`數據。考慮之後，選擇使用PHP來調用API，解析`json`格式數據並寫入數據庫。

爲防止IP被封，選擇通過代理IP進行相關操作。

因網頁佈局的關係，職位描述、需求和公司地址使用Bash Shell通過抓取頁面方式獲取。職位數據直接從數據庫讀取，獲取到相關信息後，更新到對應數據條目中。

代理IP可在免費的IP代理網站獲取。

Final Design Ideas

- 手動在IP代理網站獲取代理IP地址，通過Bash Shell進行數據的提取，並寫入數據庫中；
- 使用PHP通過代理調用API接口，處理返回的`json`格式數據，提取需要的數據寫入數據庫；通過cron定時執行該文件，自動進行數據獲取操作；
- 通過Bash Shell腳本直接抓取職位頁的 *職位描述* 和 *公司地址* 信息，更新入數

據庫，通過cron定時執行該腳本；

Project Architecture

代碼架構

```
[flying@lemp lagoutest]$ tree
.
├── getAddrRequest.sh
├── impotProxyIP.sh
├── index.php
├── lagousql.sql
└── proxyList.txt

0 directories, 5 files
[flying@lemp lagoutest]$
```

- `index.php`: 調用接口獲取數據，寫入數據庫（放置在Web服務器目錄下）
- `lagousql.sql`： 數據庫文件，含有該項目的相關數據表和初始數據
- `proxyList.txt`: 將手動獲取到的代理IP數據寫入該文件
- `impotProxyIP.sh`: 手動執行該腳本，獲取文件`proxyList.txt`中數據，並寫入數據庫proxy表中
- `getAddrRequest.sh`: 抓取職位頁，獲取職位描述和公司地址，更新入數據庫

index.php

```php
<?php
header("Content-Type:text/html;charset=UTF-8");
// ini_set("display_errors", "On");
// error_reporting(E_ALL | E_STRICT);
// error_reporting(1);


    $dbuser = 'flying';
    $dbpass = '12345';
    try {
        $pdo = new PDO('mysql:host=127.0.0.1;port=3306;dbname=l
agou',$dbuser,$dbpass);
        $pdo->exec('set names utf8');//如果不設置，中文是亂碼

    } catch (Exception $e) {
        echo '數據庫連接失敗，報錯信息： '.$e->getMessage();
    }

    $data_api = 'http://www.lagou.com/jobs/positionAjax.json';

    #定义查詢变量
    $order='default'; //px 排序 'default','new'
    $city='北京';//city 查选城市
    $salary_range='10k-15k';//yx月薪范围
    $job_nature='全职';//gx 工作性质
    $industry_field='移动互联网';//hy 行业领域
    $finance_stage='成长型';//jd 公司阶段
    $work_experience='1-3年';//gj 工作经验
    $educational_background='本科';//xl学历要求
    $keywords='运维工程师';//kd 搜索关键词
    $first='true';//first
    // $now_page_num=1;//pn 当前页面数

    #curl 参数设置
    $user_agent = "Mozilla/4.0";

    $search_paras = "?px=$order&city=$city&gx=$job_nature&kd=$k
eywords&pn=";
    // $curl_url=$data_api.$search_paras.$now_page_num;
    $curl_url=$data_api.$search_paras;
```

```
# 從數據庫中獲取city列表，名稱爲key，id值爲value
    $sql = 'select id,name from city';
    $stmt = $pdo->prepare($sql);
    $stmt->execute();
    $rows = $stmt->fetchAll(PDO::FETCH_ASSOC);
    foreach ($rows as $key => $val) {
        $city_arr[$val['name']] = $val['id'];
    }
    unset($sql);
    unset($stmt);
    unset($rows);

    # 從數據庫獲取proxy列表

    // $sql = 'select ipaddr,port from proxy where id='.rand(1,
100);
    $sql = 'select ipaddr,port from proxy order by rand() limit
1';
    $stmt = $pdo->prepare($sql);
    $stmt->execute();
    $rows = $stmt->fetchAll(PDO::FETCH_ASSOC);
    if (empty($rows)){
        $proxy = "117.135.251.133";
        $proxyPort = "82";
    }else {
        $proxy=$rows[0]['ipaddr'];
        $proxyPort=$rows[0]['port'];
    }
    unset($sql);
    unset($stmt);
    unset($rows);


    $totalPageCount = max_page_count($curl_url,$now_page_num=
1,$user_agent,$proxy,$proxyPort);//最大页面数

    // echo $totalPageCount;exit;

    for ($i=1; $i <= $totalPageCount; $i++) {
```

```
        // $sql = 'select ipaddr,port from proxy where id='.ran
d(1,100);
        $sql = 'select ipaddr,port from proxy order by rand() l
imit 1';
        $stmt = $pdo->prepare($sql);
        $stmt->execute();
        $rows = $stmt->fetchAll(PDO::FETCH_ASSOC);
        if (empty($rows)){
            $proxy = "117.135.251.133";
            $proxyPort = "82";
        }else {
            $proxy=$rows[0]['ipaddr'];
            $proxyPort=$rows[0]['port'];
        }
        unset($sql);
        unset($stmt);
        unset($rows);

        $result_arr = curl_data($curl_url.$i,$user_agent,$prox
y,$proxyPort);//返回content中信息
        $job_list_arr=$result_arr["result"];//職位列表詳細信息

        foreach ($job_list_arr as $key => $val) {
            $now = date('Y-m-d H:i:s');

            #判斷positionId是否存在表jobs中
            $sql = 'select id,publish_time from jobs where posi
tionId='.$val['positionId'];
            $stmt = $pdo->prepare($sql);
            $stmt->execute();
            $rows = $stmt->fetchAll(PDO::FETCH_ASSOC);
            unset($sql);
            unset($stmt);

            //如果不爲空，即表示在數據表jobs中存在，判斷發佈時間是否一
致，若不一致，update_times加1
            if (!empty($rows) && $rows[0]['id']>0){
                $job_id=$rows[0]['id'];
                $job_publish_time=$rows[0]['publish_time'];
                if ($val['createTime'] != $job_publish_time){
                    $sql = "update jobs set update_times=update
```

```
_times+1 where id=$job_id";
                    $res = $pdo->exec($sql);
                    // if ($resl){
                    //  echo '更新成功';
                    // }
                    unset($sql);
                    unset($res);
                }
            }
            unset($rows);

            //如果爲空，即表示在數據表jobs中不存在，須進行入庫操作
            //先判斷companyId是否存在表company中,不存在則先入庫
            $sql = 'select id from company where companyId='.$v
al['companyId'];
            $stmt = $pdo->prepare($sql);
            $stmt->execute();
            $rows = $stmt->fetchAll(PDO::FETCH_ASSOC);
            unset($sql);
            unset($stmt);
            if (!empty($rows)){
                $company_id=$rows[0]['id'];
            }else {
                //如果爲空，即不存在表company中,先入庫
                $city_id = $city_arr[$val['city']];

                $sql = "insert into company set
                    city_id=$city_id,
                    companyId='".$val['companyId']."',
                    companyShortName='".$val['companyShortNam
e']."',
                    companyName='".$val['companyName']."',
                    companyLogo='".$val['companyLogo']."', indu
stryField='".$val['industryField']."',
                    financeStage='".$val['financeStage']."',
                    companySize='".$val['companySize']."',
                    leaderName='".$val['leaderName']."',
                    create_time='$now'";

                unset($city_id);
                if ($pdo->exec($sql)){
```

```
                    $company_id = $pdo->lastInsertId();
                }
                unset($sql);
            }
            unset($rows);

            //將職位寫入數據庫
            list($salary_low,$salary_high) = explode('-',str_re
place('k','',$val['salary']));


            $sql = "insert into jobs set
                company_id=$company_id,
                positionName='".$val['positionName']."',
                positionType='".$val['positionType']."',
                positionId='".$val['positionId']."',
                work_city='".$val['city']."',
                jobNature='".$val['jobNature']."',
                education='".$val['education']."',
                salary_low=$salary_low,
                salary_top=$salary_high,
                postionAdvantage='".$val['positionAdvantag
e']."', publish_time='".$val['createTime']."',
                create_time='$now'";

            if ($pdo->exec($sql)){
                $new_job_id = $pdo->lastInsertId();
            }
            unset($sql);

        } //end foreach

    }//end for


    #抓取数据，返回json格式数据
    function curl_data ($url,$user_agent,$proxy,$proxyPort){
        $ch = curl_init();
        curl_setopt($ch,CURLOPT_PROXY,$proxy); //代理地址
        curl_setopt($ch,CURLOPT_PROXYPORT,$proxyPort); //代理地址
端口
```

```
        curl_setopt($ch, CURLOPT_PROXYAUTH, CURLAUTH_BASIC); //
代理认证模式
        curl_setopt ($ch, CURLOPT_URL, $url); //目標地址
        curl_setopt ($ch, CURLOPT_USERAGENT, $user_agent); //瀏
覽器類型
        curl_setopt ($ch, CURLOPT_HEADER, 0);//是否取得返回头信息
        curl_setopt ($ch, CURLOPT_RETURNTRANSFER, 1);
        curl_setopt ($ch, CURLOPT_FOLLOWLOCATION, 1);
        curl_setopt ($ch, CURLOPT_TIMEOUT, 120);
        $temp = curl_exec ($ch);
        curl_close($ch);
        // return $result;
        $result = json_decode($temp,true);
        return $result['content'];
    }


    //獲取最大頁面數
    function max_page_count ($curl_url,$now_page_num=1,$user_ag
ent,$proxy,$proxyPort){
        $curl_url.=$now_page_num;
        $result_arr = curl_data($curl_url,$user_agent,$proxy,$p
roxyPort);//返回content中信息
        $result=$result_arr["totalPageCount"]; //最大页面数
        return $result;
    }

 ?>
```

lagousql.sql

數據庫表結構，共4張表

```
MariaDB [lagou]> show tables;
+-----------------+
| Tables_in_lagou |
+-----------------+
| city            |
| company         |
| jobs            |
| proxy           |
+-----------------+
4 rows in set (0.00 sec)

MariaDB [lagou]>
```

```
-- MySQL dump 10.16  Distrib 10.1.12-MariaDB, for Linux (x86_64)
--
-- Host: localhost    Database: lagou
-- ------------------------------------------------------
-- Server version   10.1.12-MariaDB

/*!40101 SET @OLD_CHARACTER_SET_CLIENT=@@CHARACTER_SET_CLIENT */;
/*!40101 SET @OLD_CHARACTER_SET_RESULTS=@@CHARACTER_SET_RESULTS */;
/*!40101 SET @OLD_COLLATION_CONNECTION=@@COLLATION_CONNECTION */;
/*!40101 SET NAMES utf8 */;
/*!40103 SET @OLD_TIME_ZONE=@@TIME_ZONE */;
/*!40103 SET TIME_ZONE='+00:00' */;
/*!40014 SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0 */;
/*!40014 SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0 */;
/*!40101 SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='NO_AUTO_VALUE_ON_ZERO' */;
/*!40111 SET @OLD_SQL_NOTES=@@SQL_NOTES, SQL_NOTES=0 */;

--
-- Table structure for table `city`
--

DROP TABLE IF EXISTS `city`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `city` (
  `id` int(10) unsigned NOT NULL AUTO_INCREMENT COMMENT '城市自增id',
  `name` char(20) NOT NULL COMMENT '城市名稱',
  `create_time` datetime NOT NULL COMMENT '入庫時間',
  PRIMARY KEY (`id`),
  KEY `index_city_id` (`id`,`name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='城市表';
/*!40101 SET character_set_client = @saved_cs_client */;
```

```
insert into city (name,create_time) values('北京',now()),('上海',
now()),('廣州',now()),('深圳',now()),('杭州',now());

--
-- Table structure for table `company`
--

DROP TABLE IF EXISTS `company`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `company` (
  `id` int(10) unsigned NOT NULL AUTO_INCREMENT COMMENT '公司自增
id',
  `city_id` int(10) unsigned NOT NULL COMMENT '所在城市id，對應表c
ity',
  `companyId` int(10) unsigned DEFAULT NULL COMMENT 'lagou公司i
d',
  `companyShortName` varchar(80) DEFAULT NULL COMMENT '公司簡称',
  `companyName` varchar(200) DEFAULT NULL COMMENT '公司全称',
  `companyLogo` varchar(200) DEFAULT NULL COMMENT '公司logo，前綴
http://www.lagou.com/',
  `industryField` varchar(60) DEFAULT NULL COMMENT '行業類型',
  `financeStage` varchar(60) DEFAULT NULL COMMENT '公司階段',
  `companySize` varchar(60) DEFAULT NULL COMMENT '公司人數規模',
  `leaderName` varchar(30) DEFAULT NULL COMMENT '公司老闆',
  `address` varchar(200) DEFAULT NULL COMMENT '公司地址',
  `create_time` datetime NOT NULL COMMENT '入数据库时间',
  PRIMARY KEY (`id`),
  UNIQUE KEY `companyId` (`companyId`),
  KEY `index_company_cityid_companyid` (`id`,`city_id`,`company
Id`),
  KEY `fkey_city_companty` (`city_id`),
  CONSTRAINT `fkey_city_companty` FOREIGN KEY (`city_id`) REFER
ENCES `city` (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='公司列表';
/*!40101 SET character_set_client = @saved_cs_client */;

--
-- Table structure for table `jobs`
--
```

```
DROP TABLE IF EXISTS `jobs`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `jobs` (
  `id` int(10) unsigned NOT NULL AUTO_INCREMENT COMMENT '職位列表
自增id',
  `company_id` int(10) unsigned DEFAULT NULL COMMENT '公司id,對應
表company',
  `positionName` varchar(60) NOT NULL COMMENT '職位名稱',
  `positionType` varchar(60) NOT NULL COMMENT '職位類型',
  `positionId` int(10) unsigned DEFAULT NULL COMMENT 'lagou職位i
d，http://www.lagou.com/jobs/ID.html',
  `work_city` varchar(20) NOT NULL COMMENT '工作城市',
  `jobNature` varchar(20) NOT NULL COMMENT '工作性質',
  `education` varchar(20) NOT NULL COMMENT '學歷要求',
  `salary_low` tinyint(3) unsigned NOT NULL COMMENT '薪資最低值',
  `salary_top` tinyint(3) unsigned NOT NULL COMMENT '薪資最高值',
  `postionAdvantage` varchar(200) DEFAULT NULL COMMENT '職位優
勢',
  `publish_time` datetime NOT NULL COMMENT '公佈時間',
  `create_time` datetime NOT NULL COMMENT '入庫時間',
  `update_times` tinyint(3) unsigned DEFAULT '0' COMMENT '職位刷
新次數',
  `duty_and_request` text COMMENT '職位描述，崗位職責和任職資格',
  `address` varchar(255) DEFAULT NULL COMMENT '公司地址',
  PRIMARY KEY (`id`),
  UNIQUE KEY `positionId` (`positionId`),
  KEY `index_jobs_companyid_postionid` (`id`,`company_id`,`posi
tionId`),
  KEY `fkey_company_jobs` (`company_id`),
  KEY `index_addr` (`address`(18)),
  CONSTRAINT `fkey_company_jobs` FOREIGN KEY (`company_id`) REF
ERENCES `company` (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='職位表';
/*!40101 SET character_set_client = @saved_cs_client */;

--
-- Table structure for table `proxy`
--
```

```
DROP TABLE IF EXISTS `proxy`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `proxy` (
  `id` int(10) unsigned NOT NULL AUTO_INCREMENT COMMENT '代理列表
自增id',
  `ipaddr` char(20) NOT NULL COMMENT '代理IP地址',
  `port` smallint(5) NOT NULL COMMENT '代理IP地址端口',
  `protocol` char(20) DEFAULT NULL COMMENT 'http類型',
  `anonymity` varchar(20) DEFAULT NULL COMMENT '匿名等級',
  `country` varchar(20) DEFAULT NULL COMMENT 'IP所屬國家',
  `region` varchar(20) DEFAULT NULL COMMENT 'IP所屬省份地區',
  `city` varchar(20) DEFAULT NULL COMMENT 'IP所屬城市',
  `uptime` varchar(10) DEFAULT NULL COMMENT '正常運行時間',
  `create_time` datetime NOT NULL COMMENT '入数据库时间',
  PRIMARY KEY (`id`),
  KEY `index_id_ipaddr` (`id`,`ipaddr`,`port`),
  KEY `index_ipaddr_city` (`ipaddr`,`city`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='代理IP列表';
/*!40101 SET character_set_client = @saved_cs_client */;
/*!40103 SET TIME_ZONE=@OLD_TIME_ZONE */;

/*!40101 SET SQL_MODE=@OLD_SQL_MODE */;
/*!40014 SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS */;
/*!40014 SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS */;
/*!40101 SET CHARACTER_SET_CLIENT=@OLD_CHARACTER_SET_CLIENT */;
/*!40101 SET CHARACTER_SET_RESULTS=@OLD_CHARACTER_SET_RESULTS
*/;
/*!40101 SET COLLATION_CONNECTION=@OLD_COLLATION_CONNECTION */;
/*!40111 SET SQL_NOTES=@OLD_SQL_NOTES */;

-- Dump completed on 2016-03-09 11:53:57
```

proxyList.txt

IP代理網站地址 http://www.freeproxylists.net/

數據格式如下，通常存100條

```
117.135.251.133 82  HTTP    Anonymous   China Proxy China   Bei
jing Beijing 100.0%
117.135.251.135 82  HTTP    Anonymous   China Proxy China   Bei
jing Beijing 100.0%
111.12.83.150   103 HTTP    Anonymous   China Proxy China   Bei
jing Beijing 99.8%
117.135.251.130 80  HTTP    Anonymous   China Proxy China   Bei
jing Beijing 99.8%
117.135.251.131 80  HTTP    Anonymous   China Proxy China   Bei
jing Beijing 99.8%
...
...
...
117.177.250.151 8085    HTTP    Anonymous   China Proxy China
Beijing Beijing 99.3%
```

impotProxyIP.sh

```
#!/bin/bash
#lemp-馬雪東
#https://lempstacker.com
#2016.03.08 22:50 Tue
#獲取代理IP列表並寫入數據庫

# http://www.freeproxylists.net/
dbname='lagou'
file='./proxyList.txt';

awk '{print $1,$2,$3,$4,$5,$8,$9,$10}' $file | while read line;
do
    now=`date +'%Y-%m-%d %H:%M:%S'`
    arr=(${line})
    ipaddr=${arr[0]}
    port=${arr[1]}
    protocol=${arr[2]}
    anonymity=${arr[3]}
    country=${arr[4]}
    region=${arr[5]}
    city=${arr[6]}
    uptime=${arr[7]}
    mysql -e "insert into $dbname.proxy set ipaddr='$ipaddr',po
rt='$port',protocol='$protocol',anonymity='$anonymity',country
='$country',region='$region',city='$city',uptime='$uptime',crea
te_time='$now';"

done

#輸出條目總數
mysql -e "select count(*) from $dbname.proxy;"
```

getAddrRequest.sh

```
#!/bin/bash
#lemp-馬雪東
#https://lempstacker.com
#2016.03.08 22:50 Tue
#通過職位頁面獲取工作描述和工作地址


# 從數組表中找取duty_and_request和address爲空的數據，拿id，positionId，通過positionId拼接URL爲http://www.lagou.com/jobs/positionId.html
# 通過代理使用curl抓取頁面存儲到/tmp目錄下，使用  `mktemp -t tmpXXXXXX.txt`
#通過管道命令 提取职位描述和工作地址
#update數據庫，where條件是id

dbname='lagou'
limit=20

# 獲取代理 管道會fork一個shell子進程，變量不會保存，故通過創建臨時文件實現
tempfile=`mktemp -t tempXXXXX.txt`
mysql -Bse "select ipaddr,port from $dbname.proxy order by rand() limit 1" > $tempfile
while read i; do
    arr=($i)
    ipaddr=${arr[0]}
    port=${arr[1]}
done < $tempfile
rm -f $tempfile

# mysql -Bse "select ipaddr,port from $dbname.proxy order by rand() limit 1" | while read i; do
#     arr=($i)
#     export ipaddr=${arr[0]}
#     export port=${arr[1]}
# done

url='http://www.lagou.com/jobs/'

mysql -Bse "select id,positionId from $dbname.jobs where address is null limit $limit;" | while read line; do
```

```
    arr=($line)
    id=${arr[0]}
    positionId=${arr[1]}

    #使用代理curl抓取頁面
    tempfile=`mktemp -t tempXXXXX.txt`
    # -s quiet靜默模式 --retry 重試次數 --retry-delay 間隔時間 -x 代理 -o保存路徑
    curl -s --retry 5 --retry-delay 5 -x $ipaddr:$port -o $tempfile $url$positionId'.html'

    #使用gzip，gunzip仍無法解決某些數據入庫亂碼問題
    # curl -H "Accept-Encoding: gzip" -s --retry 5 --retry-delay 5 -x $ipaddr:$port $url$positionId'.html' | gunzip > $tempfile


    duty_and_request=`sed -n '/<dd class="job_bt">/,/<\/dd>/ p' $tempfile | grep -Evi "job_bt|</dd>|职位描述" | grep -v '^$' | sed  -r 's@</?(p|strong|br|span|class|ul|li)[[:space:]]{0,}/?>@@g;s@(<br class="">|<span class="">|<p class="">|<ul class="">| )@@g;' | sed -r "s@'@@g"`

    address=`grep -E -A 1 -i '工作地址' $tempfile | tail -1 | tr -d '</div>[[:space:]]'`

    #將數據更新入數據庫
    mysql -e "update $dbname.jobs set duty_and_request='$duty_and_request',address='$address' where id=$id;"
    rm -f $tempfile
done
```

Deploying Processes

將代碼部署到VPS上

- IP: `23.105.199.121`
- PORT: `27454`

- System Version: `CentOS Linux release 7.0.1406 (Core)`
- Apache/2.4.6, MariaDB Server 10.1.12, PHP 7.0.4

Push Code To VPS

使用`SCP`將代碼傳到VPS，已經在VPS上安裝有SSH key

- `ssh root@23.105.199.121 -p 27454 'test -d /root/lagou || mkdir -p /root/lagou'`
- `scp -P 27454 ./* root@23.105.199.121:/root/lagou`

Localhost

```
[flying@lemp lagoutest]$ ls
getAddrRequest.sh  impotProxyIP.sh  index.php  lagousql.sql  pr
oxyList.txt
[flying@lemp lagoutest]$ ssh root@23.105.199.121 -p 27454 'test
-d /root/lagou || mkdir -p /root/lagou'
[flying@lemp lagoutest]$ scp -P 27454 ./* root@23.105.199.121:/
root/lagouget
AddrRequest.sh                                100% 2829     2.8KB/
s   00:00
impotProxyIP.sh                                 100%  786     0.8
KB/s   00:00
index.php                                       100% 7739     7.6
KB/s   00:00
lagousql.sql                                    100% 6202     6.1
KB/s   00:00
proxyList.txt                                   100%   11KB  10.8
KB/s   00:00
[flying@lemp lagoutest]$
```

VPS

```
[root@localhost ~]# ls
lagou
[root@localhost ~]# tree
.
└── lagou
    ├── getAddrRequest.sh
    ├── impotProxyIP.sh
    ├── index.php
    ├── lagousql.sql
    └── proxyList.txt

1 directory, 5 files
[root@localhost ~]#
```

Create .my.cnf

數據庫用戶名和Linux當前用戶須名稱相同

在數據庫中創建`root@localhost`賬戶後，在root用戶根目錄下創建文件`.my.cnf`，填入用戶名和密碼，這樣連接數據庫時可直接登錄，不用輸入用戶名、密碼

```
[root@localhost ~]# cat .my.cnf
[client]
user=root
password=rootpassword
[root@localhost ~]# ls -lh .my.cnf
-rw------- 1 root root 40 Mar  4 21:40 .my.cnf
[root@localhost ~]#
```

注意文件權限，設置爲`600`，只有所有者有讀、寫權限。

Change Database Default Character

如果數據庫默認字符集不是`UTF-8`，寫入中文會報錯

```
MariaDB [lagou]> show variables like '%character%';
+--------------------------+-----------------------------+
| Variable_name            | Value                       |
+--------------------------+-----------------------------+
| character_set_client     | utf8                        |
| character_set_connection | utf8                        |
| character_set_database   | latin1                      |
| character_set_filesystem | binary                      |
| character_set_results    | utf8                        |
| character_set_server     | latin1                      |
| character_set_system     | utf8                        |
| character_sets_dir       | /usr/share/mysql/charsets/ |
+--------------------------+-----------------------------+
8 rows in set (0.01 sec)

MariaDB [lagou]>
```

在MariaDB數據庫配置文件`/etc/my.cnf`的`[mysqld]`中添加

```
character_set_server=utf8
collation-server=utf8_general_ci
```

重啓數據庫服務

再次查看

```
MariaDB [(none)]> show variables like '%character%';
+--------------------------+----------------------------+
| Variable_name            | Value                      |
+--------------------------+----------------------------+
| character_set_client     | utf8                       |
| character_set_connection | utf8                       |
| character_set_database   | utf8                       |
| character_set_filesystem | binary                     |
| character_set_results    | utf8                       |
| character_set_server     | utf8                       |
| character_set_system     | utf8                       |
| character_sets_dir       | /usr/share/mysql/charsets/ |
+--------------------------+----------------------------+
8 rows in set (0.01 sec)

MariaDB [(none)]>
```

Create Database & Import SQL File

創建名爲`lagou`的數據庫，並導入`lagousql.sql`文件

- `mysql -e "create database if not exists lagou default character set utf8 collate utf8_general_ci;"`
- `mysql -e "show databases like '%lagou%';"`
- `mysql -D lagou < /root/lagou/lagousql.sql`
- `mysql -e "show tables from lagou;"`

```
[root@localhost ~]# mysql -e "create database if not exists lag
ou default character set utf8 collate utf8_general_ci;"
[root@localhost ~]# mysql -e "show databases;"
+--------------------+
| Database           |
+--------------------+
| information_schema |
| lagou              |
| lempstacker        |
| mysql              |
| performance_schema |
+--------------------+
[root@localhost ~]# mysql -e "show databases like '%lagou%';"
+--------------------+
| Database (%lagou%) |
+--------------------+
| lagou              |
+--------------------+
[root@localhost ~]# mysql -D lagou < /root/lagou/lagousql.sql
[root@localhost ~]# mysql -e "show tables from lagou;"
+-----------------+
| Tables_in_lagou |
+-----------------+
| city            |
| company         |
| jobs            |
| proxy           |
+-----------------+
[root@localhost ~]# mysql -Bse "select id,name from lagou.cit
y;"
1   北京
2   上海
3   廣州
4   深圳
5   杭州
[root@localhost ~]# mysql -Bse "select count(*) from lagou.cit
y;"
5
[root@localhost ~]# mysql -Bse "select count(*) from lagou.prox
y;"
```

```
0
[root@localhost ~]# mysql -Bse "select count(*) from lagou.jobs;"
0
[root@localhost ~]# mysql -Bse "select count(*) from lagou.company;"
0
[root@localhost ~]#
```

Import Proxy List Into MariaDB

`proxyList.txt`中默認有150條數據，執行腳本`impotProxyIP.sh`導入數據庫

注意：二者必須在同一目錄下，否則會報錯`awk: fatal: cannot open file './proxyList.txt' for reading (No such file or directory)`

不用賦予腳本執行權限，直接使用`bash impotProxyIP.sh`即可

```
[root@localhost lagou]# bash impotProxyIP.sh
+----------+
| count(*) |
+----------+
|      150 |
+----------+
[root@localhost lagou]# mysql -Bse "select count(*) from lagou.proxy;"
150
[root@localhost lagou]#
```

Copy index.php Under Web Server Root Directory

此處使用Apache Web服務器，默認目錄是`/var/www/html`，創建目錄`lagou`，並將文件`index.php`放置到`/var/www/html/lagou`下

```
[root@localhost lagou]# mkdir -pv /var/www/html/lagou
mkdir: created directory ‘/var/www/html/lagou’
[root@localhost lagou]# cp -v index.php /var/www/html/lagou
‘index.php’ -> ‘/var/www/html/lagou/index.php’
[root@localhost lagou]#
```

在瀏覽器地址欄中輸入`http://23.105.199.121/lagou/`即可手動抓取數據

注意，不要忘記修改index.php的數據庫用戶名和密碼，默認是

```
$dbuser = 'flying';
$dbpass = '12345';
```

會提示`數據庫連接失敗，報錯信息： SQLSTATE[HY000] [1045] Access denied for user 'flying'@'127.0.0.1' (using password: YES)`

改成自己對應的數據庫用戶名和密碼，數據庫用戶名和Linux當前用戶須名稱相同

執行SQL語句查詢入庫情況

```
[root@localhost lagou]# mysql -e "select * from lagou.jobs orde
r by rand() limit 2\G"
*************************** 1. row ***************************
              id: 482
      company_id: 415
    positionName: 运维工程师
    positionType: 运维
      positionId: 1556282
       work_city: 北京
       jobNature: 全职
       education: 大专
      salary_low: 6
      salary_top: 9
postionAdvantage: 弹性工作制、期权、员工旅游、五险一金
    publish_time: 2016-03-07 15:22:22
     create_time: 2016-03-09 06:24:06
    update_times: 0
duty_and_request: NULL
         address: NULL
*************************** 2. row ***************************
              id: 62
      company_id: 58
    positionName: 运维工程师
    positionType: 运维
      positionId: 1439702
       work_city: 北京
       jobNature: 全职
       education: 大专
      salary_low: 4
      salary_top: 5
postionAdvantage: 14薪，100%补充医疗报销，免费食堂班车
    publish_time: 2016-03-01 14:21:14
     create_time: 2016-03-09 06:23:31
    update_times: 0
duty_and_request: NULL
         address: NULL
[root@localhost lagou]# mysql -Bse "select count(*) from lagou.
jobs;"
696
[root@localhost lagou]# mysql -Bse "select count(*) from lagou.
```

```
company;"
585
[root@localhost lagou]#
```

Setting Crontab

Crontab設置2條

- index.php：定時抓取數據
- getAddrRequest.sh： 定時抓取尚未寫入工作地址、職位描述的職位頁面，更新數據庫
- 執行`crontab -e`寫入如下信息

```
*/30 * * * * /usr/bin/php -q /var/www/html/lagou/index.php > /dev/null 2>&1
*/4 * * * * /bin/bash /root/lagou/getAddrRequest.sh > /dev/null 2>&1
```

- 執行`crontab -l`查看

```
[root@localhost lagou]# crontab -l
*/20 * * * * /usr/bin/php -q /var/www/html/lagou/index.php > /dev/null 2>&1
*/4 * * * * /bin/bash /root/lagou/getAddrRequest.sh > /dev/null 2>&1
[root@localhost lagou]#
```

如果需要註釋，在行首添加`#`即可

Problems Meeting

Using SCP with Port

因爲VPS自定義了端口，需要指定，但是將`-P 27454`放在行末會報錯，正確使用方式是放置在`SCP`後,如

```
[flying@lemp lagoutest]$ scp -P 27454 lagousql.sql root@23.105.
199.121:/root/lagou
lagousql.sql
100% 4918     4.8KB/s   00:00
[flying@lemp lagoutest]$ scp -P 27454 /var/www/html/lagou/inde
x.php root@23.105.199.121:/var/www/html/lagou
index.php
100% 7313     7.1KB/s   00:00
[flying@lemp lagoutest]$
```

CLI Database With Formatation

Shell中使用`mysql -e`返回的結果帶有表頭和分隔線，但只想獲取數據值，使用`mysql -Bse`解決。

```
[root@localhost lagou]# mysql -e "select count(*) from lagou.jo
bs;"
+----------+
| count(*) |
+----------+
|     1120 |
+----------+
[root@localhost lagou]# mysql -Bse "select count(*) from lagou.
jobs;"
1120
[root@localhost lagou]#
```

Bash Variable Scope

變量作用域

Bash中從數據庫中獲取數據使用管道符`|`時，定義的變量無法使用，爲空，使用`export`標誌爲全局變量無效。原因時管道會fork一個shell進程，即新打開一個子Shell進程，變量不會保存

此種形式無效

```
mysql -Bse "select ipaddr,port from $dbname.proxy order by rand
() limit 1" | while read i; do
     arr=($i)
     export ipaddr=${arr[0]}
     export port=${arr[1]}
done
```

使用如下形式，直接在當前Shell中進行

```
tempfile=`mktemp -t tempXXXXX.txt`
mysql -Bse "select ipaddr,port from $dbname.proxy order by rand
() limit 1" > $tempfile
while read i; do
    arr=($i)
    export ipaddr=${arr[0]}
    export port=${arr[1]}
done < $tempfile
```

Get Contents From Html

從抓取的html頁面中提取指定標籤中的內容，使用`sed`的地址定界實現。

如：想提取`<dd class="job_bt"></dd>`標籤中的內容，該標籤中還嵌套有其它標籤

```
sed -n '/<dd class="job_bt">/,/<\/dd>/ p' $tempfile
```

Insert Into Databas

Wrapping with Single Quote

在數據入庫時，只要不是整型，都需要用單引號`'`包裹起來，`insert`和`update`都需要，否則會報錯，無法正常提交。

Dealing WIth Special Character

數據入庫時，如果存在特殊字符，如單引號`'`，無法正常入庫，會報錯。本人在Bash Shell下，沒有找到比較好的辦法（不借用其它語言的前提下）。只能採用折衷方案，將特殊字符移除。

使用`sed -r "s@'@@g"`即可。

其它遇到的問題，暫時想不起來了。

References

- shell脚本中可以设置全局变量么？
- shell用curl抓取页面乱码

Change Logs

- 2016.03.09 16:24 Wed
    - 初稿完成

- GitHub: code
- Note Time: 2016.03.09 16:24 Wed
- Note Location: Beijing
- Writer: lemp-馬雪東
- Blog: https://lempstacker.com