# Try to grab recruitment data from lagou v1.1

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。~~

**更新：**重構代碼，優化數據表設計，棄用PHP，純Bash Shell Script實現，採用jq解析json格式數據。–2016.03.17 18:50 Thu

代碼已經上傳至GitHub

# Table Of Content

# 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。~~
採用jq解析json格式數據，在epel源中有安裝包。

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腳本處理獲取的API數據，使用`jq`解析json格式數據；
- 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；

腳本是`jq`解析json格式數據，故須先使用epel源安裝`jq`

```
sudo yum -y install epel-release
sudo yum -y jq
#用於數值計算
sudo yum -y bc
```

# Project Architecture

代碼架構

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

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

- `crawler.sh`: 調用接口獲取數據，寫入數據庫
- `getAddrRequest.sh`: 抓取職位頁，獲取職位描述和公司地址，更新入數據庫
- `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
- `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

## crawler.sh

```
##!/bin/bash
#lemp-馬雪東
#https://lempstacker.com
#2016.03.17 10:40 Thu
#腳本抓取拉勾數據，jq解析json數據

# 判斷jq是否安裝，須以root權限執行腳本
type jq &> /dev/null || sudo yum -q -y install jq


dbname='lagou'
lagouapi='http://www.lagou.com/jobs/positionAjax.json'

#curl參數 retry重試次數 retryDelay時間間隔
declare i retry=5
declare i retryDelay=5
# curl -s --retry $retry --retry-delay $retryDelay -x ipaddr:
port

#獲取代理IP
function getProxyIP () {
    # temp=`mysql -Bse "select ipaddr,port from $dbname.proxy
order by rand() limit 1;"`
    # arr=($temp)
    arr=(`mysql -Bse "select ipaddr,port from $dbname.proxy o
rder by rand() limit 1;"`)
    ipaddr=${arr[0]}
    port=${arr[1]}
    echo $ipaddr:$port
}

#調用自定義函數getProxyIP
curltool="curl -s --retry $retry --retry-delay $retryDelay -x
"`getProxyIP`

#定义查詢变量
order='new'    #px 排序 'default','new'
# city='北京'    #city 查选城市
city_arr=('北京' '上海' '杭州');
rand=`echo "$RANDOM%${#city_arr[*]}" | bc`
city=${city_arr[${rand}]}

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

#拼接查詢參數
search_paras="$lagouapi?px=$order&city=$city&gx=$job_nature&k
d=$keywords&pn="


#獲取城市列表
tempfile=`mktemp -t tempXXXXXX.txt`
mysql -Bse "select name,id from $dbname.city;" > $tempfile

declare -A cityarr
while read line; do
    temp=($line)
    tempname=${temp[0]}
    tempid=${temp[1]}
    cityarr[$tempname]=$tempid
    unset temp
    unset tempname
    unset tempid
done < $tempfile
rm -f $tempfile



#獲取最大頁面數
totalPageCount=`$curltool $search_paras'1' | jq '.content.tot
alPageCount'`

for (( i=1; i<=$totalPageCount; i++ )); do
    tempfile=`mktemp -t tempXXXXXX.txt`
    $curltool $search_paras$i > $tempfile
    pageSize=`cat $tempfile | jq '.content.pageSize'`

    for (( j=0; j<$pageSize; j++ )); do
        createTime=`cat $tempfile | jq -r ".content.result |
.[${j}].createTime"`
```

```
        city=`cat $tempfile | jq -r ".content.result | .
[${j}].city"`
        companyId=`cat $tempfile | jq -r ".content.result | .
[${j}].companyId"`
        companyName=`cat $tempfile | jq -r ".content.result |
.[${j}].companyName"`
        companyShortName=`cat $tempfile | jq -r ".content.res
ult | .[${j}].companyShortName"`
        companyLogo=`cat $tempfile | jq -r ".content.result |
.[${j}].companyLogo"`
        industryField=`cat $tempfile | jq -r ".content.result
| .[${j}].industryField"`
        financeStage=`cat $tempfile | jq -r ".content.result
| .[${j}].financeStage"`
        companySize=`cat $tempfile | jq -r ".content.result |
.[${j}].companySize"`
        leaderName=`cat $tempfile | jq -r ".content.result |
.[${j}].leaderName"`

        positionId=`cat $tempfile | jq -r ".content.result |
.[${j}].positionId"`
        positionName=`cat $tempfile | jq -r ".content.result
| .[${j}].positionName"`
        positionType=`cat $tempfile | jq -r ".content.result
| .[${j}].positionType"`
        positionFirstType=`cat $tempfile | jq -r ".content.re
sult | .[${j}].positionFirstType"`
        jobNature=`cat $tempfile | jq -r ".content.result | .
[${j}].jobNature"`
        education=`cat $tempfile | jq -r ".content.result | .
[${j}].education"`
        positionAdvantage=`cat $tempfile | jq -r ".content.re
sult | .[${j}].positionAdvantage"`
        # workYear=`cat $tempfile | jq -r ".content.result |
.[${j}].workYear"`
        salary=`cat $tempfile | jq -r ".content.result | .
[${j}].salary" | tr -d k`
        salaryhigh=${salary#*-}
        salarylow=${salary%-*}

        #1判斷positionId是否存在表jobs中

        #1.1通過判斷數組長度，判斷是否已經存在
        jobarr=(`mysql -Bse "select id,publish_time from $dbn
```

```
ame.jobs where positionId=$positionId;"`)
        #1.1.1存在
        if [[ ${#jobarr[*]} -gt 0 ]]; then
            jobid=${jobarr[0]}
            jobpubtime=${jobarr[1]}' '${jobarr[2]}
            #比對publish_time，不一致則更新字段update_times，last
_update_time
            if [[ "$jobpubtime" != "$createTime" ]]; then
                mysql -Bse "update $dbname.jobs set update_ti
mes=update_times+1,last_update_time='$createTime' where id=$j
obid;" &> /dev/null

            fi
            unset jobid
            unset jobpubtime
        else
            #1.1.2 不存在，入庫
            #1.1.2.1 companyId是否存在表company中,不存在則先入庫,
获取表company id
            comparr=(`mysql -Bse "select id from $dbname.comp
any where companyId=$companyId;"`)
            if [[ ${#comparr[*]} -eq 0 ]]; then
                cityid=${cityarr[$city]}
                mysql -Bse "insert into $dbname.company set c
ity_id=$cityid, companyId='$companyId', companyShortName='$co
mpanyShortName', companyName='$companyName', companyLogo='$co
mpanyLogo', industryField='$industryField', financeStage='$fi
nanceStage', companySize='$companySize', leaderName='$leaderN
ame';" &> /dev/null && compid=`mysql -Bse "select id from $db
name.company where companyId=$companyId;"`

            else
                compid=${comparr[0]}
            fi
            unset comparr

            if [[ $positionId -ne 0 ]]; then
                mysql -Bse "insert into $dbname.jobs set comp
any_id=$compid, positionName='$positionName', positionType
='$positionType', positionFirstType='$positionFirstType', pos
itionId='$positionId', work_city='$city', jobNature='$jobNatu
re', education='$education', salary_low='$salarylow', salary_
top='$salaryhigh',positionAdvantage='$positionAdvantage',publ
ish_time='$createTime';"
```

```
            fi
        fi
        unset jobarr


    done

    rm -f $tempfile
done
```

## getAddrRequest.sh

```
#!/bin/bash
#lemp-馬雪東
#https://lempstacker.com
#2016.03.17 19:16 Thu
#通過職位頁面獲取工作描述和工作地址

dbname='lagou'
limit=100

#curl參數 retry重試次數 retryDelay時間間隔
declare i retry=5
declare i retryDelay=5
# curl -s --retry $retry --retry-delay $retryDelay -x ipaddr:
port

#獲取代理IP
function getProxyIP () {
    arr=(`mysql -Bse "select ipaddr,port from $dbname.proxy o
rder by rand() limit 1;"`)
    ipaddr=${arr[0]}
    port=${arr[1]}
    echo $ipaddr:$port
}

#調用自定義函數getProxyIP
curltool="curl -s --retry $retry --retry-delay $retryDelay -x
"`getProxyIP`

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

mysql -Bse "select id,positionId from $dbname.jobs where addr
ess 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
    $curltool -o $tempfile $url$positionId'.html'

    #使用sed地址定界獲取指定標籤內容
```

```
    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
```

## 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';
sed -i '/^$/d;s@High Anonymous@HighAnonymous@g' $file
#createTime設置屬性timestamp default current_timestamp，無需手動指定入庫時間

mysql -e "truncate table $dbname.proxy;"

awk '{print $1,$2,$3,$4,$5,$8,$9,$10}' $file | while read line;do
    # now=`date +'%Y-%m-%d %H:%M:%S'`
    # now=`date +'%F %T'`
    #存入數組
    arr=(${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',port='$port',protocol='$protocol',anonymity='$anonymity',country='$country',region='$region',city='$city',uptime='$uptime';"

    # mysql -e "insert into $dbname.proxy set ipaddr='$ipaddr',port='$port',protocol='$protocol',anonymity='$anonymity',country='$country',province='$region',city='$city',uptime='$uptime',createTime='$now';"

done

mysql -e "select count(*) from $dbname.proxy;"
```

# lagousql.sql

數據表時間字段由`datetime`該爲`timestamp`,`timestamp`使用須知參見本人MariaDB-TIMESTAMP初窺。

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_RESUL
TS */;
/*!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_CHECK
S=0 */;
/*!40014 SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FO
REIGN_KEY_CHECKS=0 */;
/*!40101 SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='NO_AUTO_VALU
E_ON_ZERO' */;
/*!40111 SET @OLD_SQL_NOTES=@@SQL_NOTES, SQL_NOTES=0 */;

--
-- Current Database: `lagou`
--

CREATE DATABASE /*!32312 IF NOT EXISTS*/ `lagou` /*!40100 DEF
AULT CHARACTER SET utf8 */;

USE `lagou`;

--
-- 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` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP
COMMENT '數據入庫時間 YYYY-MM-DD HH:MM:SS',
  PRIMARY KEY (`id`),
  KEY `indcity_name` (`name`)
) ENGINE=InnoDB AUTO_INCREMENT=6 DEFAULT CHARSET=utf8 COMMENT
='城市表';
/*!40101 SET character_set_client = @saved_cs_client */;

LOCK TABLES `city` WRITE;
/*!40000 ALTER TABLE `city` DISABLE KEYS */;
INSERT INTO `city`(name) VALUES ('北京'),('上海'),('广州'),('深
圳'),('杭州');
/*!40000 ALTER TABLE `city` ENABLE KEYS */;
UNLOCK TABLES;


--
-- 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city',
  `companyId` int(10) unsigned DEFAULT NULL COMMENT 'lagou公司
id',
  `companyShortName` varchar(80) DEFAULT NULL COMMENT '公司简
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` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP
COMMENT '數據入庫時間 YYYY-MM-DD HH:MM:SS',
  PRIMARY KEY (`id`),
  UNIQUE KEY `companyId` (`companyId`),
  KEY `indcomp_compid` (`companyId`),
  KEY `fkey_city_companty` (`city_id`),
  CONSTRAINT `fkey_city_companty` FOREIGN KEY (`city_id`) REF
ERENCES `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',
  `positionFirstType` varchar(60) DEFAULT NULL COMMENT '職位類
型FirstType',
  `positionId` int(10) unsigned DEFAULT NULL COMMENT 'lagou職
位id，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',
  `positionAdvantage` varchar(200) DEFAULT NULL COMMENT '職位
優勢',
  `publish_time` timestamp NULL DEFAULT NULL COMMENT '公佈時
間',
  `create_time` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP
COMMENT '數據入庫時間 YYYY-MM-DD HH:MM:SS',
  `update_times` tinyint(3) unsigned DEFAULT '0' COMMENT '職位
```

```
刷新次數',
  `last_update_time` timestamp NULL DEFAULT NULL COMMENT '最近
更新時間',
  `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 `indjobs_posid` (`positionId`),
  KEY `fkey_company_jobs` (`company_id`),
  KEY `index_addr` (`address`(18)),
  CONSTRAINT `fkey_company_jobs` FOREIGN KEY (`company_id`) R
EFERENCES `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` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP
COMMENT '數據入庫時間 YYYY-MM-DD HH:MM:SS',
  PRIMARY KEY (`id`),
  KEY `indpro_ip_port` (`ipaddr`,`port`)
) 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 2016-03-17 19:01:22
```

## 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   B
eijing Beijing 100.0%
117.135.251.135 82   HTTP    Anonymous   China Proxy China   B
eijing Beijing 100.0%
111.12.83.150   103 HTTP    Anonymous   China Proxy China   B
eijing Beijing 99.8%
117.135.251.130 80   HTTP    Anonymous   China Proxy China   B
eijing Beijing 99.8%
117.135.251.131 80   HTTP    Anonymous   China Proxy China   B
eijing Beijing 99.8%
...
...
...
117.177.250.151 8085    HTTP    Anonymous   China Proxy China
Beijing Beijing 99.3%
```

# 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

先安裝epel源，再執行該命令，安裝`bc`、`jq`

```
yum install -y bc jq
```

# 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
proxyList.txt
[flying@lemp lagoutest]$ ssh root@23.105.199.121 -p 27454 'te
st -d /root/lagou || mkdir -p /root/lagou'
[flying@lemp lagoutest]$ scp -P 27454 ./* root@23.105.199.12
1:/root/lagou
crawler.sh                                    100% 5953
5.8KB/s   00:00
getAddrRequest.sh                             100% 2208
2.2KB/s   00:00
impotProxyIP.sh                               100% 1162
1.1KB/s   00:00
lagousql.sql                                  100% 6690
6.5KB/s   00:00
proxyList.txt                                 100%   12KB  1
2.0KB/s   00:00
[flying@lemp lagoutest]$
```

VPS

```
[root@localhost ~]# ls
lagou
[root@localhost ~]# tree
.
└── lagou
    ├── crawler.sh
    ├── getAddrRequest.sh
    ├── impotProxyIP.sh
    ├── 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)]>
```

# Import SQL File

~~創建名爲 `lagou` 的數據庫，導入 `lagousql.sql` 文件（文件中含有創建數據庫命令）~~

- ~~`mysql < /root/lagou/lagousql.sql`~~
- ~~`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 < /root/lagou/lagousql.sql
[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 -e "show tables from lagou;"
+-----------------+
| Tables_in_lagou |
+-----------------+
| city            |
| company         |
| jobs            |
| proxy           |
+-----------------+
[root@localhost ~]# mysql -Bse "select id,name from lagou.cit
y;"
2   上海
1   北京
3   广州
5   杭州
4   深圳
[root@localhost ~]# mysql -Bse "select count(*) from lagou.ci
ty;"
5
[root@localhost ~]# mysql -Bse "select count(*) from lagou.pr
oxy;"
0
[root@localhost ~]# mysql -Bse "select count(*) from lagou.jo
bs;"
0
[root@localhost ~]# mysql -Bse "select count(*) from lagou.co
```

```
mpany;"
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(*) |
+-----------+
|       166 |
+-----------+
[root@localhost lagou]# mysql -Bse "select count(*) from lago
u.proxy;"
166
[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 or
der by rand() limit 2\G"
*************************** 1. row ***************************
*
                id: 226
        company_id: 184
      positionName: 运维工程师
      positionType: 运维
positionFirstType: 技术
        positionId: 1339518
         work_city: 上海
         jobNature: 全职
         education: 学历不限
        salary_low: 6
        salary_top: 12
positionAdvantage: 环境优 晋升制度完善 员工福利好
      publish_time: 2016-03-14 10:12:28
       create_time: 2016-03-17 19:50:02
      update_times: 0
 last_update_time: NULL
 duty_and_request: NULL
           address: NULL
*************************** 2. row ***************************
*
                id: 347
        company_id: 269
      positionName: 运维工程师
      positionType: 运维
positionFirstType: 技术
        positionId: 1568767
         work_city: 上海
         jobNature: 全职
         education: 大专
        salary_low: 10
        salary_top: 20
positionAdvantage: 完善的薪酬福利体系
      publish_time: 2016-03-09 19:51:21
       create_time: 2016-03-17 19:50:28
      update_times: 0
 last_update_time: NULL
 duty_and_request: NULL
           address: NULL
[root@localhost lagou]# mysql -Bse "select count(*) from lago
```

```
u.jobs;"
610
[root@localhost lagou]# mysql -Bse "select count(*) from lago
u.company;"
459
[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

```
*/15 * * * * /bin/bash /root/lagou/crawler.sh > /dev/null 2>&
1
*/5 * * * * /bin/bash /root/lagou/getAddrRequest.sh > /dev/nu
ll 2>&1
```

- 執行`crontab -l`查看

```
[root@localhost lagou]# crontab -l
*/15 * * * * /bin/bash /root/lagou/crawler.sh > /dev/null 2>&
1
*/5 * * * * /bin/bash /root/lagou/getAddrRequest.sh > /dev/nu
ll 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@23.10
5.199.121:/root/lagou
lagousql.sql
100% 4918     4.8KB/s   00:00
[flying@lemp lagoutest]$ scp -P 27454 /var/www/html/lagou/ind
ex.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.
jobs;"
+----------+
| count(*) |
+----------+
|      610 |
+----------+
[root@localhost lagou]# mysql -Bse "select count(*) from lago
u.jobs;"
610
[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 $dbname.proxy order by ra
nd() limit 1" | while read i; do
     arr=($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 ra
nd() 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
- jq: Cannot index array with string
- Using jq to parse and display multiple fields in a json serially

# Change Logs

- 2016.03.09 16:24 Wed
    - 初稿完成
- 2016.03.17 19:56 Thu
    - 代碼重構，棄用PHP，使用jq解析json格式數據

---

- GitHub: code
- PDF Version 1.0: Google Drive
- Note Time: 2016.03.09 16:24 Wed
- Update Time: 2016.03.17 19:56 Thu
- Note Location: Beijing
- Writer: lemp-馬雪東
- Blog: https://lempstacker.com