数据的重要性

拼命优化算法不如给数据增加一个字段

数据源

公开数据
黑客数据
爬虫

获取数据

爬虫定义

爬虫实战

爬虫进阶

爬虫流程

工具/lib

python **Scrapy Pyspider**

java **Nutch Crawler4j**

node.js **Node-crawler**

软件 **火车头 / 八爪鱼 / Hawk**

语言生态

crawler

Languages

Python	5,872
Java	3,266
JavaScript	2,009
Ruby	1,070
PHP	896
C#	614
HTML	460
Go	363
C++	319
Scala	194

spider

Languages

Python	3,489
Java	1,249
JavaScript	1,019
HTML	357
PHP	341
Ruby	254
C++	208
C#	207
Go	128
C	100

Node.js + JS + MongoDB / MySQL / Postgres

爬虫分类

通用型 vs **定向 / 聚焦 / 垂直**

单机 vs **分布式**

dom解析 vs **ajax** vs **正则表达式**

写个网站

知己知彼 百战不殆

chrome 没有秘密的前端世界

network
寻找请求
过滤请求

element
右击 —> 检查进入element 查看相应结构

sources
查看源代码
Ctrl + U / option + command + U

console
调试代码
js命令对话框

百度 FE助手

最简页面 (不用server)

```
<!DOCTYPE html>
<body>
    <div>a hahaha!</div>
</body>
</html>
```

head

中文字符
css的出现
title

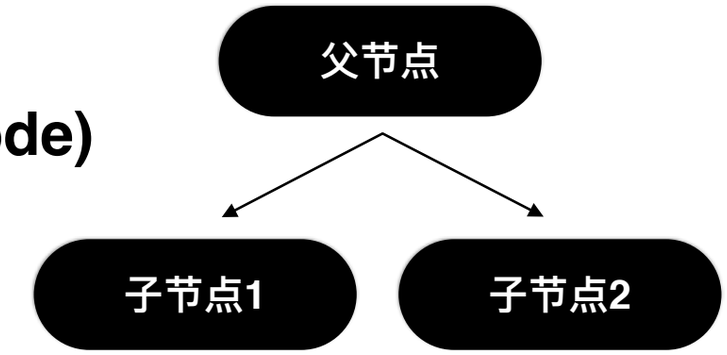
最简server: (解决跨域问题)

```
➜  ~ sudo npm i anywhere -g
Password:
```

dom (Document Object Model)

节点(node)

万物皆块



标签体系 div a span img video iframe

jquery(cheer.io)体系

document.gerElementById(‘id’) vs **$(‘#id’)**

node.attr() node.text()

xpath

异步

请求服务器

大促活动把服务器拖慢

今天小区上网的人超多

许多请求 / 串联和并联

请求1

请求2

请求3

VS

请求1 请求2 请求3

函数可以成为参数

```
function cb(){
  console.log('异步执行..');
}

function fn(cb){
  setTimeout(cb, 100);
}

fn(cb);
```

ajax请求

定义

AJAX = 异步 JavaScript 和 XML。

好处

无需刷新网页

步骤

创建AJAX对象 ——> 发HTTP请求 ——> 接收数据 ——> 更新网页

原生

```
var xmlhttp = new XMLHttpRequest();
xmlhttp.open( 'GET', 'http://example.com' , true );
xmlhttp.onreadystatechange = function () {
  if ( XMLHttpRequest.DONE != xmlhttp.readyState ) {
    return;
  }
  if ( 200 != xmlhttp.status ) {
    return;
  }
  console.log( xmlhttp.responseText );
};
xmlhttp.send();
```

jquery

```
$.ajax({
  url: 'http://www.example.com/',
  success: function(ds){
    cosnole.log(ds);
  },
  type:'GET',
  dataType: 'json'
});
```

```
$.getJSON('http://www.example.com/', function(ds) {
  cosnole.log(ds);
});
```

开始实战

一些sample

链家　　高德　　益动gps
饿了么

莆田系黑医院

https://github.com/langhua9527/BlackheartedHospital

https://github.com/zhouningyi/BlackheartedHospital

百度fe助手

低频、易被封杀

搜索接口

三月爬虫

setTimeout 的坏处

队列的实现(原生与async)

如何导出数据（json、csv、mongodb）

高频、不易被封杀

GeoCoding

多线程

啥是进程和线程 真正的线程　　食堂

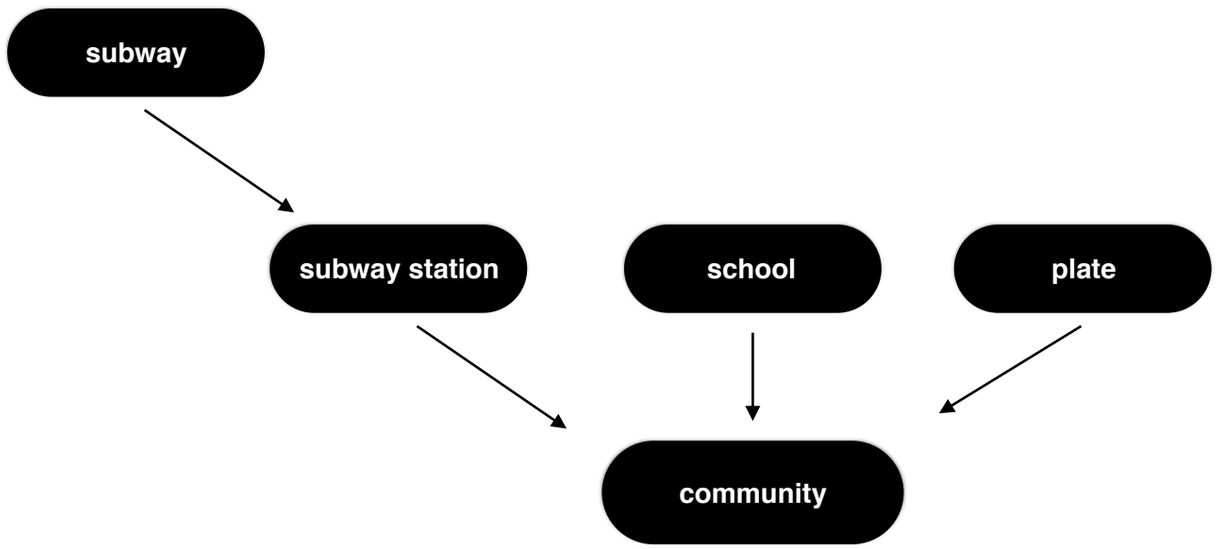
process

原生和async

链家网

小区主页

二手交易主页 / 已完成的交易



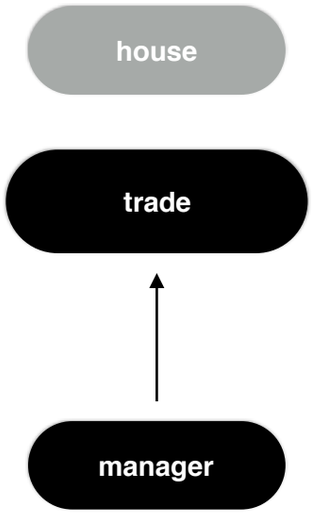
技巧

cheer.io解析dom

如何越过100页限制

常用的正则表达式

多对多的数据库存储



如何模拟登录

phantomjs

模拟user-agent

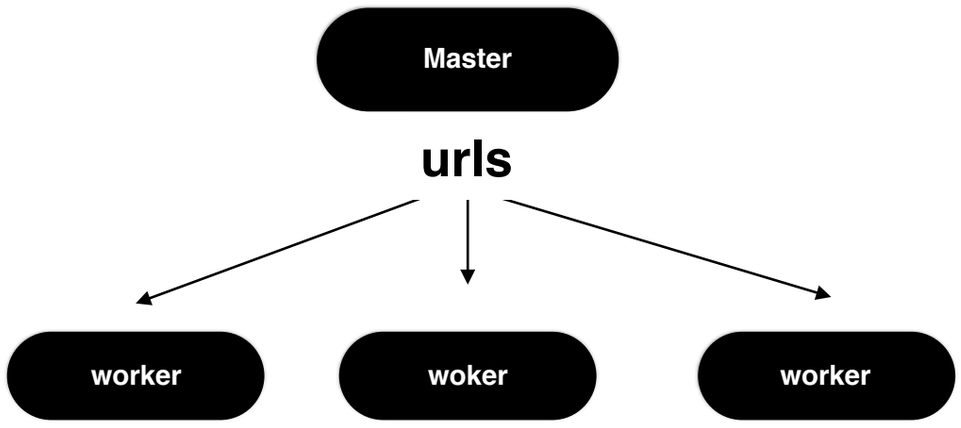
random ua

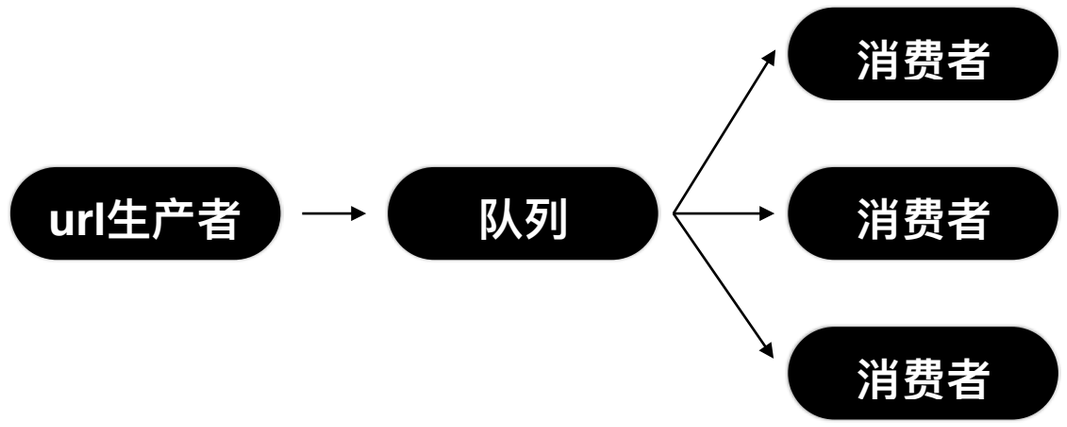
登录

回到搜索接口

代理 ip池 ip的回收策略

分布式





生成URL：

类比：数列的通项公式和递推公式 社交网站广度优先

不用公式

一些奇葩的例子　自增id　geohash

Bloom Filter

移动app charles 加代理

发出请求

请求的返回

URL Error 没网
HTTPError 服务器不给数据
1xx 临时返回
2xx 基本成功
3xx 转跳为主
4xx 基本挂了
5xx 服务器灭了
其他，被服务器发现是爬虫

请求的结构

User-Agent：浏览器型号
Referer：请求从哪来
Cookie：模拟登陆
Content-Type：请求体的类型(POST)

请求的速度

速度精细化控制

请求的IP

百变IP 修改http头　　x-forward-for

百变IP 使用代理的各种姿势

反爬虫与反反爬虫

网站:
明杀404 / 验证码 && 暗杀

有一定流速控制:
3小时和1分钟检测

要素伪造与发现
假百度爬虫 user-agent与ip穿帮
IP 和 user-agent的绑定关系

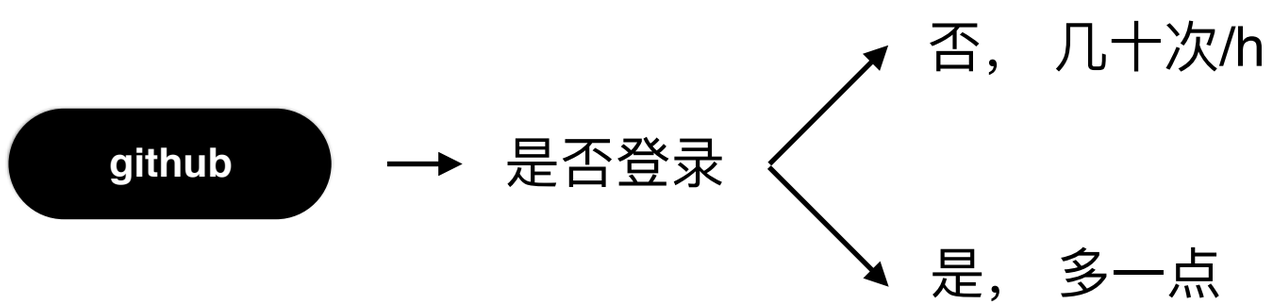
Referer

X-forward-for

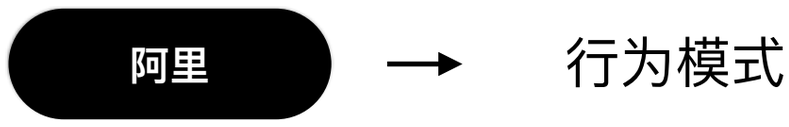
是否访问了js、html

大招
绑定微信 + 限制访问次数

陷阱法
display:none法


github
是否登录
否， 几十次/h
是， 多一点


google
执行js前后端校验
ip黑名单


阿里
行为模式

优化：

解析

dom解析

正则表达式解析

存储

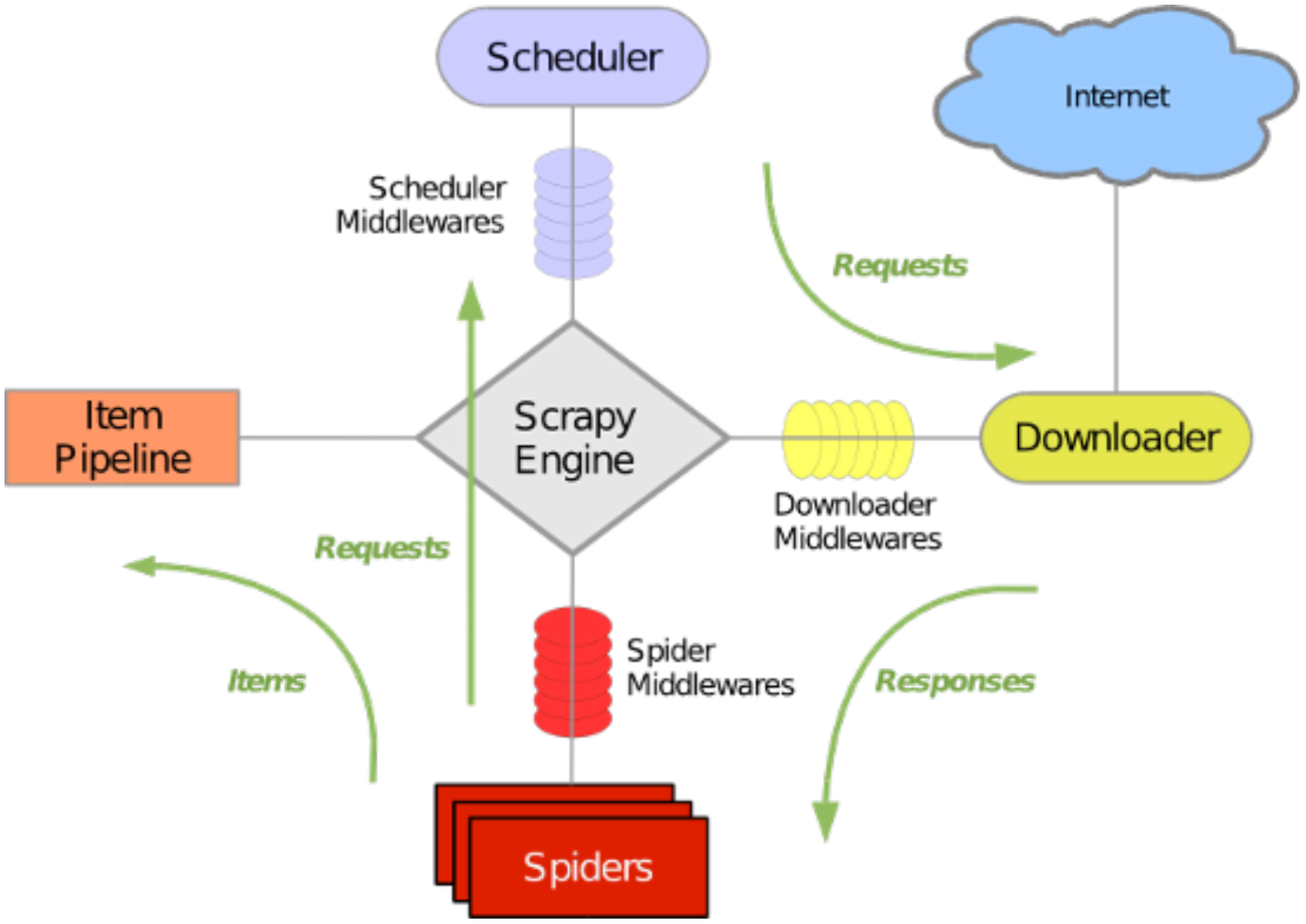
Moogoose / MongoDB的介绍

设计一个schema表结构

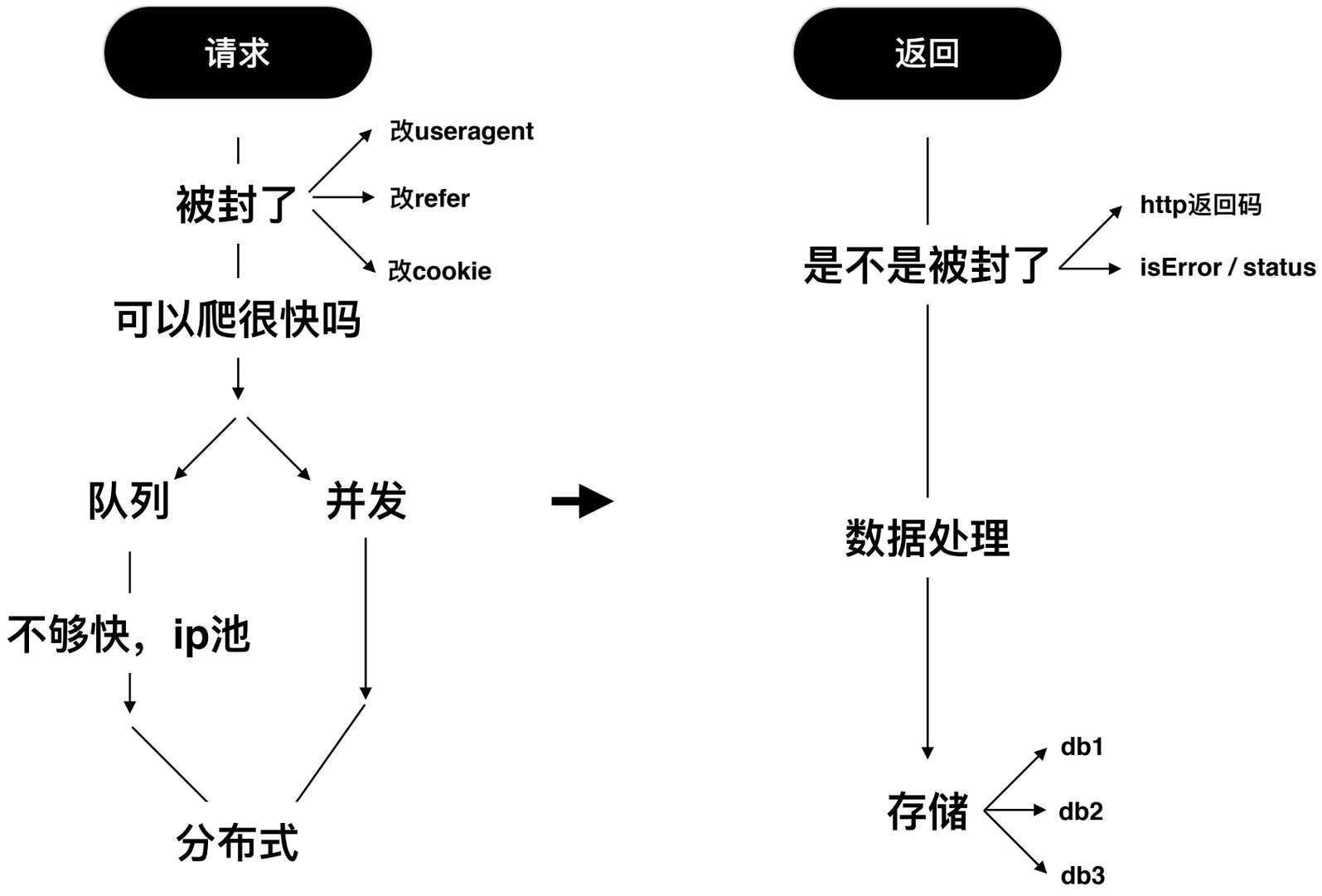
与爬虫对接

持久化，存储url

持久化 存储html



引擎从调度器中取出一个链接(URL)用于接下来的抓取
引擎把URL封装成一个请求(Request)传给下载器
下载器把资源下载下来，并封装成应答包(Response)
爬虫解析Response
解析出实体（Item）,则交给实体管道进行进一步的处理
解析出的是链接（URL）,则把URL交给调度器等待抓取


请求
被封了
改useragent
改refer
改cookie
可以爬很快吗
队列
并发
不够快，ip池
分布式
返回
是不是被封了
http返回码
isError / status
数据处理
存储
db1
db2
db3