甲方名称：

上海某大数据公司（因保密协议限制无法公开具体信息）

甲方需求：

根据需求定制一采集平台，并进行长期维护。

1. 采集目标：淘宝（含天猫）、京东、一号店；
2. 从数据库读取客户的关键词列表（非固定），分别在各电商平台进行搜索；
3. 提取搜索结果商品的"标题","价格","30天销量","总销量","规格参数","评论数","收藏人数","卖家信息", "商品评分","服务评分", "物流评分"等信息存储在数据库中；
4. 采集商品新增的评论信息到数据库中（MongoDB）；
5. 采集频率：每日采集一次，保留每次采集结果；
6. 采集量：淘宝约100万商品、京东约90万商品、一号店约10万商品；

商品信息属性如下JSON文档所示：


```
{
    "people_wrote": [
        {
            "status": 0,
            "count": 331,
            "name": "外观漂亮",
            "created": "2015-06-15 01:31:42",
            "modified": "2017-03-20 12:20:09",
            "rid": 464,
            "type": 0,
            "id": 876913,
            "productId": 1593516
        },
        {
            "status": 0,
            "count": 286,
            "name": "性能不错",
            "created": "2015-06-10 15:11:49",
```

```
        "modified": "2017-03-15 10:25:25",
        "rid": 466,
        "type": 0,
        "id": 871381,
        "productId": 1593516
    },
    {
        "status": 0,
        "count": 260,
        "name": "速度快",
        "created": "2015-06-12 15:36:31",
        "modified": "2017-03-20 12:20:09",
        "rid": 469,
        "type": 0,
        "id": 874091,
        "productId": 1593516
    },
    {
        "status": 0,
        "count": 258,
        "name": "配置不错",
        "created": "2015-06-15 11:40:15",
        "modified": "2017-03-20 12:20:09",
        "rid": 462,
        "type": 0,
        "id": 877296,
```

```
    "productId": 1593516
},
{
    "status": 0,
    "count": 231,
    "name": "屏幕大",
    "created": "2015-06-11 18:00:49",
    "modified": "2017-03-17 11:03:36",
    "rid": 465,
    "type": 0,
    "id": 872937,
    "productId": 1593516
},
{
    "status": 0,
    "count": 214,
    "name": "开机速度",
    "created": "2015-06-15 11:40:15",
    "modified": "2017-03-14 14:11:33",
    "rid": 471,
    "type": 0,
    "id": 877295,
    "productId": 1593516
},
{
```

```
    "status": 0,
    "count": 210,
    "name": "东西不错",
    "created": "2015-06-12 15:36:31",
    "modified": "2017-03-16 21:22:12",
    "rid": 461,
    "type": 0,
    "id": 874092,
    "productId": 1593516
},
{
    "status": 0,
    "count": 198,
    "name": "硬件不错",
    "created": "2015-06-15 01:31:42",
    "modified": "2017-03-20 12:20:09",
    "rid": 472,
    "type": 0,
    "id": 876912,
    "productId": 1593516
},
{
    "status": 0,
    "count": 167,
    "name": "性价比高",
```

```
            "created": "2015-06-17 10:18:54",
            "modified": "2017-03-18 11:23:29",
            "rid": 470,
            "type": 0,
            "id": 880157,
            "productId": 1593516
        },
        {
            "status": 0,
            "count": 104,
            "name": "键盘不错",
            "created": "2015-06-15 01:31:42",
            "modified": "2017-03-17 11:03:36",
            "rid": 467,
            "type": 0,
            "id": 876911,
            "productId": 1593516
        }
    ],
    "description": {
        "待机时长": "9小时以上",
        "特性": "背光键盘",
        "屏幕尺寸": "其他",
        "商品名称": "AppleMacBook Pro",
        "硬盘容量": "256G固态",
```

```
        "厚度": "15.1mm—20.0mm",
        "分类": "轻薄本",
        "系统": "其他",
        "商品产地": "中国大陆",
        "商品毛重": "3.96kg",
        "显卡类别": "集成显卡",
        "显卡型号": "其他",
        "处理器": "Intel i7标准电压版",
        "显存容量": "其他",
        "裸机重量": "2-2.5kg",
        "分辨率": "超高清屏（2K/3k/4K)",
        "商品编号": "1593516",
        "内存容量": "16G"
    },
    "describe_score": "",
    "title": "Apple MacBook Pro 15.4英寸笔记本电脑 银色(Core i7 处理器/16GB内存/256GB SSD闪存/Retina屏 MJLQ2CH/A)",
    "url": "http://item.jd.com/1593516.html",
    "service_score": "",
    "jd_price": "13688.00",
    "shop_name": "京东Apple产品专营店",
    "brand_id": "14026",
    "shop_type": "C",
    "shop_id": "1000000127",
    "props": {
        "指纹识别": "无",
```

"屏幕尺寸": "15英寸",

"净重": "其它",

"触摸板": "有",

"续航时间": "其它",

"颜色": "银色",

"内置蓝牙": "蓝牙4.0",

"三级缓存": "6M",

"CPU类型": "第四代智能英特尔酷睿i7处理器",

"物理分辨率": "其他",

"类型": "英特尔核芯显卡",

"屏幕规格": "15.4英寸",

"电池": "其它",

"音频端口": "其它",

"内置麦克风": "有",

"无线局域网": "其它",

"网络摄像头": "有",

"局域网": "其它",

"键盘": "背光键盘",

"系列": "MacBook",

"平台": "Intel",

"电源适配器": "其它",

"显示比例": "宽屏16：10",

"最大支持容量": "16GB",

"屏幕类型": "LED背光",

"扬声器": "内置扬声器",

"内存容量": "其它"

```
    },
    "category_id": "672",
    "good_rate": "0.972",
    "express_score": "",
    "rate_total": "31741"
}
```


技术难点：

- 淘宝（含天猫）和京东封锁IP非常厉害；
- HTTP请求量非常大，每天超过600万；

实现方案：

- 采用鲲鹏ADSL动态IP代理有效绕过淘宝和京东的封IP限制；
- 采用分布式架构，可以动态扩展采集集群规模（增加采集机器）；

项目状态：

分布式电商数据采集平台于2016年2月1日正式开始上线运行，目前仍在稳定进行中。

其它类似需求客户：

美国某知名传媒有限公司 （因保密协议限制无法公开具体信息）
