# 0库存模式方案

将指定的网页上的产品数据抓取后转换成各个平台(ebay,amazon等)的数据结构，并将产品刊登到各个平台。

## 1 业务流程

整个流程大致可分为：抓取、分析、转换、上架4个步骤：

### 抓取

通过网络爬虫技术将指定站点的产品页面抓取并保存下来。

### 分析

将抓取后的页面结构进行分析，提取出产品信息，如：标题、描述、尺码颜色、图片等信息。

### 转换

将分析后的产品数据转换为各个平台接口所需要的数据结构。

### 上架

将转换好的产品数据刊到各个平台。

## 2 技术选型

整个业务作为新的项目，主要选用Spring boot2.2、mybatis、Mysql、mongodb、elasticsearch、redis、selenium、webmagic、httpclient、jsoup、htmlunit等

### 2.1 数据抓取

数据抓取可分为两种：静态页面、动态页面

#### 2.1.1 静态页面

这种数据抓取相对简单，可以使用请求页面完成后，页面HTML代码中有所需要的信息，直接解析HTML页面即可提取出数据。使用 jsoup、Xpath等库实现解析。

#### 2.1.2 动态页面

动态页面指请求完成页面后，所需要的数据是通过Ajax异步请求后动态渲染的页面的，这种页面的数据抓取比较难。可以使用selenium技术模拟浏览器实现。

### 2.2 数据存储

选择使用Mysql + mongodb / elasticsearch / redis 关系型数据库和非关系型数据库结合。mysql存储日常的业务数据，如下载的页面、爬取日志等等；mongodb/elasticsearch/redis用于存储 页面分析后的数据，因为分析后的数据是结构化(JSON)的、且容易发生结构变化的，如果存储在Mysql关系型数据库中，结构发生变化时需要修改数据库结构才能适配，不易维护。

## 3 存在问题

### 3.1 语言问题

产品都是中文的（如：标题、颜色、描述等），是否需要转换为英文？怎样转换为英文？

### 3.2 图片问题

产品相关的图片是否需要下载本地存储重新生成自己的URL地址，还是直接使用电商平台的图片URL