基于爬虫数据的粮油市场数据仓库构建

一、项目背景

传统的市场调研主要以实地考察、面对面访谈、问卷调查等形式展开，互联网与电子商务的出现为市场调研提供了新的途径。通过实时采集行业及电商平台数据，如电商平台商品数据、交易数据、评论数据、工商注册数据，能够分析特定行业市场竞争状态。

然而，这些实时采集的数据（又称“爬虫数据”），在数据的组织上呈现“批量文本”的特征，既缺乏精确的“字段”管理，又存在数据项不明确、重复、缺失等问题，这对市场竞争分析带来了不少困难。因此，将通过各种途径获取的爬虫数据进行整理，将其关联起来，并用关系数据库进行存储与管理，是有效利用爬虫数据进行市场分析的一种解决方案。

二、数据文件

本项目获取的爬虫数据为粮油行业市场数据，包括大米、面粉、食用油和杂粮四种，数据来源主要分两大类：

**1.电商平台数据**

（1）商品交易数据（每周采集）

采集平台：天猫、京东、苏宁

数据项："\_id",SKU ID,批次,Product ID,SKU链接,与描述相符,价格,优惠券信息,到手价,单位价格,发货地,品类,商品上下架的记录,商品介绍信息,商品是否为广告,商品规格,商品评价数,好评度,平台名称,店铺名称,店铺链接,月销量,标题,货物产地,货物种类,采集时间

注意：不同平台的数据项略有不同

（2）商品评论文本（累计采集）

采集平台：天猫、京东

数据项："\_id",SKU\_ID,评论ID,评价者,评论内容,评论回复数,评论图片,评论时间,评论视频,评论评分

注意：不同平台的数据项略有不同

（3）店铺数据

采集平台：京东

数据项："\_id",店铺ID,企业名称,企业注册资金,店铺名称,店铺链接,法定代表人姓名,联系地址,营业执照所在地,营业执照有效期,营业执照注册号,营业执照经营范围

**2.企业工商信息数据**

采集平台：企查查

（1）文件1：粮油公司基础信息

数据项：\_id,url,enterprise\_name,Current\_Time,annualDate,authority,district, entType,enterprise\_name.1,industry,insuranceNum,legalPerson,licenseNumber,openStatus,openTime,orgNo,pid,prevEntName,qualification,realCapital,regAddr,regCapital,regNo,scope,startDate,unifiedCode

（2）文件2：粮油公司ICP备案

数据项：\_id,序号,首页地址,网站名称,域名,备案号

（3）文件3：粮油公司变更记录

数据项：\_id,序号,变更日期,变更项目,变更前,变更后

（4）文件4：粮油公司股东信息

数据项：\_id,序号,股东名称,股份类型,持股数量(股),持股比例,认缴出资额,认缴出资日期

（5）文件5：粮油公司监管记录

数据项：\_id,序号,日期,类型,结果,检查机构

（6）文件6：粮油公司软件著作

数据项：\_id,序号,软件名称,软件简称,版本号,软件著作分类,行业分类,登记日期,软件著作人,首次发表日期,国籍

（7）文件7：粮油公司注册商标

数据项：\_id,序号,商标,商标名称,注册号,国际分类,申请时间

（8）文件8：粮油公司专利

数据项：\_id,序号,专利名称,公布/公告号,专利类型,公布/公告日期

（9）文件9：粮油公司作品

数据项：\_id,序号,登记号,作品类别,作品名称,创作完成日期,首次发表日期,登记日期

三、项目任务

1.设计关系数据库的表结构，管理系统为MySQL，将上述爬虫数据文件存储在数据库中，便于管理与应用。

2.对爬虫数据进行清洗，包括去重、缺失值数据、数据项拆分、样本标记。

3.将清洗后的数据存入数据库。