# 软件工程实践项目报告书

组别：第五组 组长：吴健 031804128

## 一、项目目标及意义

### 1. 项目概述

随着互联网和电子支付的应用及普及，电子商务行业发展迅猛，同时越来越多信息 被聚集起来，电子商务进入了信息大爆炸的时代。从用户角度来看，用户在面临越来越多的选择的情况下显得无所适从，无法尽快地锁定自己想要购买的物品；从企业角度来看，海量信息的汇入使得企业无法对其进行全量的存储和及时的利用，难以抽取有效信息支持决策，造成用户和资本的流失。而电商数据处理与分析平台应运而生并逐步发展，成为有利于企业分析实现精准营销的有效方式。

本次项目主要实现了面对电商企业的数据仓库的开发，视为软件应用领域的应用软件。应用软件是指解决特定业务需要的独立应用程序。这类应用软件处理商务或技术数据，以协助业务操作或协助做出管理和技术决策。数据仓库（Data Warehouse，可简写为 DW 或 DWH），是面向分析的集成化数据环境，为企业决策制定过程，提供系统数据支持的战略集合。该项目主要目标为企业制定科学决策提供数据依靠，以及为数据的存储和管理提供保障。

项目开发需求主要分为以下几个模块：

1.用户行为数据采集平台

2.业务数据采集平台

3.数仓分层建模、数据分析（针对电商业务，面向用户、商品、销售、活动等相关指标）

4.BI 报表展现、实时查询和分析等。

## 二、可行性分析

### 1. 项目的规模、难度和成本

本次项目规模适中，在开发涉及的框架上，主要包含当前 Apache 原生框架和 CDH 框架的大数据组件，包括 Hadoop、Hive、Flume、Kafka、Mysql 等等。在数据维度和数据量方面，数据库设计包含了电商企业的常见业务指标，包括：日、周、月用户活跃度、用户相关属性统计、成交额、购买率、以及相关交易数据分析等等，数仓总表在 100 张表以内，可对比中小型电商企业。

在项目开发的难度和技术要求上，需要开发成员有以下能力：

1.熟悉 Java 编程语言和 SQL 语句，熟悉 Linux 环境下 Shell 脚本语言。

2.了解并熟悉相关大数据组件，如：Flume、Kafka、Hadoop、Hive、Presto、Kylin 等。

3.熟悉数仓维度建模理论，了解数据仓库/OLAP 相关理论知识。

在项目的开发成本上，该项目实际投入具体开发需要较大的硬件支持，以中小型企业场景为例，初步估计集群规模需要 6~10 台（每台约 8TB）左右硬件服务器。作为模拟开发，本次项目可能会使用阿里云等云计算提供商的云服务器资源，或者直接在本地集群上完成。

数仓整体架构：


## 2. 开发团队

| 姓名 | 学号 | 班级 |
|---|---|---|
| 吴健 | 031804128 | 数据科学与大数据技术 01 班 |
| 陈世益 | 031804105 | 数据科学与大数据技术 01 班 |
| 邓憧 | 031804107 | 数据科学与大数据技术 01 班 |
| 林文伟 | 031702434 | 计算机 04 班 |
| 张海龙 | 031804136 | 数据科学与大数据技术 01 班 |
| 张世博 | 031804138 | 数据科学与大数据技术 01 班 |
| 张荣龙 | 031804137 | 数据科学与大数据技术 01 班 |
| 黄玲玥 | 031804116 | 数据科学与大数据技术 01 班 |

## 3. 时间要求

本项目开发周期时长暂定为 6-8 周左右。

## 4. 风险

经过整体的团队评估，本项目主要面临的风险大概在需求分析、开发环境以及技术要求方面。由于应用领域不同，该软件需求要针对于所面向的电商企业，在用户行为数据和业务数据设计方面有所缺陷，需求分析无法涵盖全部指标。在开发环境上，该数仓所需要的硬件配置比较高，作为模拟开发，所需要购买的服务器以及计算资源存在不确定性。最后，作为不是很成熟的开发团队，该项目在待开发时的复杂性和技术难度也存在一定程度的风险。

# 三、项目计划

## 1. 软件开发工具

本次项目主要运用的开发工具为：
操作系统：Linux
编程环境：IDEA、Shell
技术组件：Flume、Kafka、Mysql、Hadoop、Hive、Presto、Superset 等

## 2. 人员分工及时间安排

| 姓名 | 负责模块 | 时间进度 |
| --- | --- | --- |
| 吴健 | 整体架构设计 | 第 6-14 周 |
| 陈世益 | 数仓建模 | 第 8-12 周 |
| 邓憧 | 数据采集 | 第 6-8 周 |
| 林文伟 | 数据采集 | 第 6-8 周 |
| 张海龙 | 数仓建模 | 第 8-12 周 |
| 张世博 | 数据可视化、即时查询 | 第 12-14 周 |
| 张荣龙 | 数据可视化、即时查询 | 第 12-14 周 |
| 黄玲玥 | 数仓建模 | 第 8-12 周 |