数据仓库与大数据工程课程实验二

实验报告提交时间要求： 截止日期：4 月 15 日

2.1 实验背景

数据采集是建立数据仓库与大数据平台重要基础性工作。在本实验中，要求将基于该实验所搭建的实验平台的相关环境，针对利用真实的用户行为日志数据来实现大数据平台的增量数据采集过程。所使用的用户行为日志有两种（如下所示：数据集一和数据集二），可任选其一进行实验。

1) 数据集一：某中文 IT 技术社区用户行为日志数据

本课程所使用的数据集一由某中文 IT 技术社区提供，共包含 157,427 位用户在 2015 年期间产生的多种类型的行为数据。数据集基本统计信息如下：

数据类别	数据内容（行为类型）	数据量
用户行为数据	用户发表博客记录	1,000,000 条
	用户浏览博客记录	3,536,444 条
	用户评论博客记录	182,273 条
	用户对博客点赞记录	95,668 条
	用户对博客点踩记录	9,326 条
	用户收藏博客记录	10,4723 条

数据集中用户编号为 U0000001 至 U0157247，文档编号为 D0000001 至 D1000000。原始行为日志数据以文本文件的形式存储，其记录格式为：

每一行代表一条日志记录，包含四个字段，依次为用户行为类型、用户编号、博客编号和行为产生时间，用\001 分开。其中用户行为类型包括：用户发表博客、用户浏览博客、用户评论博客、用户对博客点赞、用户对博客点踩、用户收藏博客等六种用户行为。

2) **数据集二：**某电商平台的用户行为日志数据

http://www.dataju.cn/Dataju/web/datasetInstanceDetail/290

2.2 实验目的

基于前期对用户行为日志数据的理解，熟悉并练习真实企业中数据的采集流程。主要实验目标有：

1) 业务背景问题领域 ER 模型设计；
2) 选择合适的技术，设计并实现原始日志数据的实时增量收集（包括单数据源和多数据源）、多粒度增量收集；
3) 设计并实现原始日志数据的实时解析与结构化存储；

1.3 实验环境

1) Kafka 集群；
2) MySQL 数据库；
3) 编程语言：Java（推荐使用）、Scala、C++等；

2.4 实验与设计内容

1) 设计用于存储结构化的用户行为日志数据的 MySQL ER 模型；
2) 设计并实现三种不同类型的模拟数据抓取的 Kafka 生产者，将文本文件形式的日志数据发送到 Kafka 集群中的相应话题中。三种不同类型的 Kafka 生产者为：
   - 模拟单数据源秒级（每秒一次）抓取的 Kafka 生产者；
   - 模拟多数据源秒级抓取的 Kafka 生产者；
   - 模拟小时级（每小时一次）数据抓取的 Kakfa 生产者；

针对本实验数据集，对于小时级数据抓取的 Kafka 生产者进行如下明确：为 Kafka 生产者设置单个数据包的数据时间跨度为一小时（即最近一小时的数据）。为缩短程序运行时长，实验测试时，Kafka 生产者两次传输数据的时间间隔可人为设定（如 5 秒）。需要说明的是，数据集二中提供的时间戳为 Unix 时间戳，可编写程序将该时间戳转换为标准日期时间戳。

3) 设计并实现一个模拟数据解析的 Kafka 消费者，将 Kafka 集群中相应话

题中的文本形式的日志数据记录解析为结构化数据，并存储到 MySQL 中自己设计的相应关系表中。

2.5 实验步骤

1) 设计存储结构化日志数据的 MySQL 关系表；
2) 根据原始数据情况，创建 Kafka 话题并尝试设置不同的分区数目进行后续实验；
3) 根据原始数据和创建的 Kafka 话题，设计并实现三种不同类型的模拟日志抓取的 Kafka 生产者；
4) 根据 Kafka 中存储日志数据的话题中的数据量情况，设计并实现日志解析与结构化数据存储的 Kafka 消费者；