**数据仓库实验报告**

秦博 1352961

实验一：数据库表转移

实验过程：

* 利用工具在数据库中建立一张新表格，并根据待用表的模式，建立列；
* 利用表输入得到数据；
* 利用表输出向新建立的数据库中填充数据；

实验小结：

本次实验比较简单，能让我们充分理解kettle的工作模式，所以实验比较顺利。

实验二：从已有文本文件中读取信息，并从网页中得到数据

实验过程：

* 利用工具从文本文件中得到电影ID，具体过程如下：

1. 观察文本文件数据格式，发现可以总结成一个8列的表；
2. 根据列的内容，可以删去以review开头的行，只剩下带有ID的行进行字符串截取；
3. 将得到的ID去重；
4. 文本输出；

* 将电影ID写入数据库（利用JavaScript）；
* 从数据库中拿到ID并且从网页中得到数据（利用JavaScript），并把得到的电影名存入相应的行（此步骤应该重复）；
* 将得到的电影名称去重（利用sql语句在数据库中进行）；

遇到的问题：

* 利用Kettle软件基本没有什么问题，速度很快；
* 在从网页中得到数据的过程中会遇到无法访问的情况，这时候需要利用数据库记录此ID是否查找过或者是否找到名字（state），这样一旦程序崩溃，不至于从第一个记录重新开始，而是查找状态为“未找到”的记录开始；
* 查找过后得到的许多电影名字为“robot check”，这种情况应该是遇到了输入验证码或者其他情况导致无法访问网页，所以应该重复进行此过程直到取到数据,最终得到185732；