数据仓库大项目：

1. 数据来源：

<http://snap.stanford.edu/data/web-Movies.html>

1. 项目要求：
   1. 针对电影及其周边信息，建立基于关系型数据仓库和分布式文件型数据仓库系统，并进行系统性能比对
      1. 能够从数据来源处获取数据
      2. 建立关系型数据仓库存储模型，存储数据
      3. 建立分布式文件系统存储模型，存储数据
      4. 在数据展现的界面上能够执行数据应用中的查询，并将在两种不同存储模型上的执行时间以数值的方式和图表的方式显示在界面上
   2. 数据来源：数据来源自Snap的文本文件和Amazon网站，数据包括但不限于以下信息：

电影ID，评论用户ID，评论用户ProfileName，评论用户评价Helpfulness，评论用户Score，评论时间Time，评论结论Summary，评论结论Text，电影演员，电影上映时间，电影风格，电影导演，电影主演，电影演员，电影版本等信息。

* 1. 数据应用：常见查询及统计（占总查询数目）=80%）：
     1. 按照时间进行查询及统计（例如XX年有多少电影，XX年XX月有多少电影，XX年XX季度有多少电影，周二新增多少电影等）
     2. 按照电影名称进行查询及统计（例如 XX电影共有多少版本等）
     3. 按照导演进行查询及统计（例如 XX导演共有多少电影等）
     4. 按照演员进行查询及统计（例如 XX演员主演多少电影，XX演员参演多少电影等）
     5. 按照电影类别进行查询及统计（例如 Action电影共有多少，Adventure电影共有多少等）
     6. 按照上述条件的组合查询和统计

1. 提交内容
   1. ETL脚本
   2. E-R图
   3. 物理存储模型（Schema）
   4. 数据仓库导出文件（请按照每张表分别导出csv格式文件并压缩。文件名同表名）
   5. 分布式文件系统导出文件，schema定义文件
   6. 查询和统计程序
   7. 每组准备一个答辩ppt，说明整体思考过程和解决方案。