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，参考数据使用场景，建立至少2种存储模型，设计其存储架构并进行系统性能比对。
      1. 能够从数据来源处获取数据
      2. 参考数据使用场景，建立适当的存储模型存储数据
      3. 在数据展现的界面上能够进行典型数据使用场景中的查询，并将在多种不同存储模型上的执行性能以数值的方式和图表的方式显示在界面上
      4. 撰写文档说明数据存储模型选择的理由、数据存储模型、系统架构。
   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. 数据使用场景：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. 按照导演-演员的合作关系，或者演员-演员的合作关系进行查询（例如XX演员经常和哪些演员及导演合作，合作了哪些作品；或哪些演员经常一起合作等）
     7. 按照电影上映时间进行查询及统计（例如Action类型电影常见的黄金上映期是什么？恐怖片和动作片的排片日期有没有什么关系等）
     8. 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）/图数据存储模型/分布式文件系统存储模型等
   4. 查询和统计程序
   5. 数据存储模型选择的理由、数据存储模型、系统架构说明文档
   6. 每组准备一个答辩ppt，说明整体思考过程和解决方案