## 大数据软件应用与技术实践

### 一、实验要求

1. 结合Netflix Dataset或ml-25m，按照实验步骤完成实验。

2. Netflix数据集包含了1999.12.31至2005.12.31期间由网站用户提供的超过一亿条电影评价。Netflix Dataset.7z压缩文件包含电影信息、training set（训练集）、probe set（探测集）和qualifying set(评估集)。压缩文件的详细信息如下图1所示：

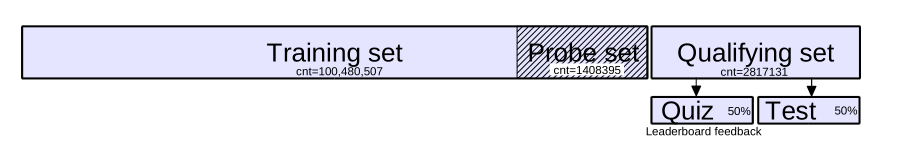


图1 Netflix Dataset说明

3. 实验成绩按照各步骤完成情况给分。

4. 大数据软件需包括教材当中有的，可以是介绍过的也可以是没介绍过的，尽可能地使用本学期学习并实验过的大数据软件和方法。

### 二、实验步骤

1. 独立设计一套切实可行的数据挖掘任务task，并简要介绍该任务的各个步骤step，使用的关键技术与软件应用。（10分）

1. 步骤：
2. 数据收集和预处理：首先，我们需要收集用户的历史评分数据。这些数据可以从提供的training\_set训练集中获取。预处理步骤包括数据清洗（例如，处理缺失值和异常值）, 使用Hadoop的MapReduce模式来并行处理数据，这可以大大提高数据预处理的效率。使用HDFS持久化存储数据.
3. 数据分析：在这个步骤中，我们可以进行一些描述性统计分析，使用Hadoop的MapReduce模式来并行计算描述性统计量, 例如，计算每部电影的平均评分、评分的标准差等。
4. 模型训练：我们可以使用Mahout协同过滤算法来训练我们的推荐模型。协同过滤算法可以根据用户的历史评分来预测他们可能喜欢的电影。Mahout是一个开源的机器学习库，它提供了一系列的算法，包括协同过滤、分类、聚类等。
5. 推荐生成：最后，我们可以使用我们的推荐模型来为用户生成推荐。我们可以为每个用户推荐他们最可能喜欢的前N部电影。
6. 关键技术与软件应用：HDFS、MapReduce和Mahout.

2. 详细介绍数据准备过程，例如：如何对数据进行预处理，如何做数据持久化存储。（10分）

3. 详细介绍数据统计分析的过程，例如：统计、分析电影评价的数量、用户的数量、电影的数量等，统计、分析电影评价的平均值、方差等。（10分）

4. 详细介绍数据挖掘的算法、过程和结果。例如：使用了某个数据挖掘的算法，或编辑了某段代码，接着对多少用户和多少电影进行了协同过滤，得到了什么结果或什么成果，有什么结论。（10分）