ALS 算法在真实数据集上的测试方案

1. 测试环境

本实验的实验环境是真实的物理集群，集群规模是 10 台真实物理机器，其中 9 台节点做 slave 节点，一台节点作为 master 节点。节点配置信息如下表 1 所示：

节点基本信息	
处理器	8 Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz
内存	16GB RAM
操作系统	Ubuntu 11.04
网络带宽	1000MBps
磁盘	2* 1TB SATA
Spark version	Spark 2.0

表 1 节点信息

2. 测试数据

测试数据来自于真实的问题和数据集，在算法上选择 ALS 算法作为实验的算法，同时选择了 MovieLens [17,18]和 Netflix[19]作为 ALS 算法测试数据集。算法和数据集的详细描述如下表所示：测试数据集如表 2 所示：

数据集	数据集说明
MovieLens	MovieLens 数据集是由美国 Minnesota 大学的 GroupLens 研究小组创建并维护，其研究项目涉及相关的许多研究项目到信息过滤，协同过滤等领域推荐系统。本数

	据集是 GroupLens 发布的数据集，它包括 260,000 个用户在 4 万部电影中应用了 24,000,000 个评分和 670,000 个标签应用。所有的评分值也都是 1 到 5 中的整数值，其中分数越高表示客户对相应电影的评价越高（越喜欢）
Netflix	Netflix 数据集是由 Netflix-prize 比赛提供的公开数据集，根据历史用户对电影的评价文件对用户进行电影推荐，历史电影评价文件包含超过 1 亿的评分从 48 万随机选择，匿名 Netflix 客户超过 17 万部电影的数据收集于 1998 年 10 月至 2005 年 12 月，并反映了在此期间收到的所有评级的分布。 评分是在规模从 1 到 5 星。 为了保护客户的隐私，每个客户 id 已替换为随机分配的 ID。 每个评级的日期和还提供每个电影 id 的发行的标题和年份。所有的评分值都是 1 到 5 中的整数值,其中分数越高表示客户对相应电影的评价越高（越喜欢）

表 2 算法测试数据集

3．测试方法

ALS 算法的

4. 数据收集方法

Spark 提供了完整的 api 接口,用来帮助用户将数据获取到本地。我们通过相应的 api，可以得到相应的以 json 数据格式呈现的运行后的数据，通过程序解析，将所有运行数据导入本地 excel 表格中,从而可以进行比较分析。具体的,通过 http://master:18080/api/v1 可以访问到相应 history server 里面的数据。

/applications：得到已经运行过的应用信息

/applications/[app-id]/jobs： 得到对应应用的 job 信息

/applications/[app-id]/stages：得到对应应用的 stage 信息

/applications/[app-id]/stages/[stage-id]/[stage-attempt-id]/taskList ：得到对应 stage 的 task 信息

更详细的参数获取可以参考 spark 官方文档的监控子文档：

http://spark.apache.org/docs/latest/monitoring.html

5．数据分析和整理