MapReduce 课程设计 1 —— 比赛日志分析

1. 课程设计目标

本课程设计通过使用 MapReduce 实现比赛日志分析。

通过本课程设计的学习，利用 MapReduce 工具实现大数据下的数据分析方法。

2. 学习技能

本次课程设计可以掌握以下 MapReduce 编程技能：

- 海量日志数据的统计分析
- 基于 MapReduce 的预测模型设计

3. 题目描述

各项体育赛事中，根据运动员在场上的具体表现情况，会产生大量的数据。在职业体育赛事中，对赛事过程中产生的日志进行分析，可以有效分析对手的技战术特点，从而可以帮助教练团队制定相应策略予以应对。

本课程设计数据中记录了一系列篮球赛事的比赛日志，要求学生按照要求进行比赛日志的统计、分析，并根据已有的比赛日志预测后续赛事的比赛结果。

数据集

本实验的数据集为单个文件 dataset，存储了单个赛季 1230 场比赛所有的比赛记录。每条记录由多个字段组成，记录一个事件。

表格数据结构如图所示：

Date	Quarter	SecLeft	AwayTeam	HomeTeam	PlayBy	Shooter	ShotType	ShotOutcome	Assister	Blocker	FoulType	Fouler	Fouled	Rebounder	ReboundType	ViolationPlayer	ViolationType	FreeThrowShooter	FreeThrowOutcome	EnterGame	LeaveGame	TurnoverPlayer	TurnoverType	TurnoverCauser
11-1	1	720	team004	team021	team004																			
11-1	1	701	team004	team021	team004	Jeff Bowen	2-pt layup	miss		Thomas Smith														
11-1	1	699	team004	team021	team021									Chris Williams	defensive									
11-1	1	697	team004	team021	team021																	Chris Williams	bad pass	
11-1	1	681	team004	team021	team004	Michael Perez	2-pt jump shot	make	Jeff Bowen															
11-1	1	660	team004	team021	team021	William Wolfe	2-pt jump shot	make	Samuel Sanders															
11-1	1	644	team004	team021	team004	Jared Hawkins	2-pt jump shot	make	Ryan Carpenter															
11-1	1	627	team004	team021	team021	Thomas Smith	2-pt jump shot	miss																
11-1	1	625	team004	team021	team004									Randy Robertson	defensive									
11-1	1	613	team004	team021	team004	Jeff Bowen	2-pt hook shot	miss																
11-1	1	611	team004	team021	team021									Chris Williams	defensive									
11-1	1	601	team004	team021	team021	William Wolfe	2-pt jump shot	make	Thomas Smith															
11-1	1	582	team004	team021	team004						shooting	Chris Williams	Michael Perez											
11-1	1	582	team004	team021	team004													Michael Perez	make					
11-1	1	582	team004	team021	team004													Michael Perez	miss					
11-1	1	581	team004	team021	team021									William Wolfe	defensive									
11-1	1	569	team004	team021	team021	Robert Martin	2-pt jump shot	miss																
11-1	1	568	team004	team021	team004									Jeff Bowen	defensive									

数据列信息如下：

字段名	字段含义	出现场景	备注
EventID	事件 ID	所有	
Date	比赛日期	所有	比赛日期使用“月-日”格式，如 11-1
Quarter	比赛节数	所有	1-4 分别表示第 1 节到第 4 节 加时赛从 5 开始一直编号下去
SecLeft	剩余时间	所有	表示当前节剩余的比赛时间
AwayTeam	客队	所有	由 team001-team030 不等
HomeTeam	主队	所有	由 team001-team030 不等
PlayBy	日志生成方	所有	该条日志由哪支球队生成
Shooter	投篮者	投篮事件	
ShotType	投篮类型	投篮事件	包括分值（2 分或 3 分）和投篮方式 如 3-pt jump shot，2-pt layup 等
Shot Outcome	投篮结果	投篮事件	命中（make）或不中（miss）
Assister	助攻者	投篮事件 助攻事件	投篮命中时可能出现
Blocker	盖帽者	投篮事件 盖帽事件	投篮不中时可能出现 注意：盖帽者所在球队是日志生成方的对手
FoulType	犯规类型	投篮事件 犯规事件	犯规事件可能随投篮事件一起发生，也可能独立出现 注意：犯规类型决定了哪方对哪方犯规；offensive 和 offensive 属于攻方犯规，其它属于守方犯规
Fouler	犯规球员	投篮事件 犯规事件	
Fouled	被犯规球员	投篮事件 犯规事件	
Rebounder	篮板球员	篮板事件	
Rebound Type	篮板类型	篮板事件	可能为进攻篮板（offensive）或防守篮板（defensive）
Violation Player	违例球员	违例事件	
Violation Type	违例类型	违例事件	
FreeThrow Shooter	罚球球员	罚球事件	
FreeThrow Outcome	罚球结果	罚球事件	命中（make）或不中（miss）
Turnover Player	失误球员	失误事件	
Turnover Type	失误类型	失误事件	
Turnover Causer	抢断者	失误事件 抢断事件	抢断者通过抢断的方式制造对手失误 注意：抢断者所在球队是日志生成方的对手
EnterGame	换上球员	换人事件	
LeaveGame	换下球员	换人事件	

为了分析方便，数据集保证以下几个假设成立：

- 所有球员没有重名现象；
- 球员姓名中没有特殊字符；
- 一支球队一天只有一场比赛。

为了更方便的进行本地调试，本实验提供了一个数据量较小的样本文件 dataset_sample。

任务 1：计算每条日志对比赛比分的影响

已知：

- 当 FreeThrowMade 为 make 时，PlayBy 球队得 1 分；
- 当 ShotOutcome 为 make 时，若 ShotType 为 2-pt***，PlayBy 球队得 2 分；若 ShotType 为 3-pt***，PlayBy 球队得 3 分。

针对每条日志，输出“XXX 得 X 分”或“不得分”。

要求输出一个文本文件，每行内容包括的字段包括：EventID、得分结论，举例：

```
1,不得分
2,不得分
3,不得分
4,不得分
5,不得分
6,team004 得 2 分
...
```

任务 2：统计每场比赛的比赛结果

提示：各支球队每天只会有一场比赛。

输出格式：日期，主队，主队比分，客队，客队比分。举例：

```
11-1,team021,94,team004,106
...
```

任务 3：计算数据集中各项技术统计的前五名球员

提示：根据每条日志产生行为的制造队员进行统计。

要求得到得分、篮板、助攻、抢断、盖帽最多的 5 名球员（如有数据相同的，名次并列）

输出格式：数据类型、排名、球员名称、数值。举例：

```
得分,1,Zhang San,2400
得分,2,Li Si,2300
...
篮板,1,Wang Wu,1111
篮板,1,Zhao Liu,1111
```

任务 4：预测给定比赛中，主队和客队的胜率

根据已有的数据作为训练数据，设计预测算法，预测给定几组对阵中主队和客队的胜率。

要求预测的几组对阵存储在 predict 文件中，包括两支球队的名称，前者为客队、后者为主队。predict 文件内容如下：

```
AwayTeam,HomeTeam
team004,team001
team008,team025
team009,team030
team011,team020
team013,team007
team015,team012
team017,team021
team027,team014
```

助教将比较各个预测值与数据生成模型本身预测胜率的差值，判断模型的准确度。

注意：本实验对预测算法设计的评估是主要的，对于预测结果是否准确的评估是次要的。

输出格式：客队，主队，客队胜率，主队胜率。胜率保存为小数，保留四位小数。举例：

```
team004,team001,0.1453,0.8547
...
```

任务 5：设计合理的评价标准，评选出赛季最有价值球员（选做）

可以考虑的因素：

- 球员场均数据
- 球员所在球队的战绩
- 合理设计现成的或新的高阶指标，以评价球员表现

根据设计的评价标准，选出最好的 5 名球员。其中，第 1 名为最有价值球员，第 2 至第 5 名为最有价值球员候选人。

输出格式：名次，球员名称，最有价值球员评分（如果有）。举例：

```
1,Zhang San
2,Li Si
```

如果是有评分的：

```
1,Zhang San,31.5
2,Li Si,28.2
```

该任务为选做任务，不强制要求完成。

4. 提交作业

要求各位同学提交以下内容：

- “源代码”文件夹：提交源代码；
- “JAR 包”文件夹：包含可执行 JAR 包及 JAR 包的执行方式。
- “程序输出”文件夹：包含上述所有任务的最终输出结果，每个任务单独存储一个文件，分别命名为“任务 X.txt”。
- “程序设计报告”文件：文件包括方法设计、执行流程、算法优化、实验结果、代码运行过程截图等内容。