

**2024-2025学年第2学期**

**《软件开发应用基础实践》**

**课程报告**

学 院 人工智能与信息工程学院

专业班级 大数据222

学 号 1221004045

姓 名 朱俞恺

# 基于Spark的1996-2022多赛季NBA球员数据分析

1.摘要

近年来，体育数据分析在职业篮球领域发挥着越来越重要的作用。本研究基于Spark分布式计算框架，对1996-2022年共26个NBA赛季的球员数据进行系统性分析，旨在挖掘球员表现、薪资趋势、球队战术演变等关键规律。

研究采用Spark SQL进行高效的数据清洗与聚合，利用机器学习库构建球员聚类模型和表现预测模型，并分析球员间的合作关系网络。实验结果表明，不同位置的球员在效率值（PER）、真实命中率（TS%）等关键指标上存在显著差异，薪资水平与球员贡献并非完全线性相关。此外，联盟整体进攻节奏（Pace）呈现上升趋势，三分球占比显著提高，印证了现代NBA向"小球时代"的战术转型。

本研究不仅为球队管理层提供数据驱动的决策支持，如球员评估和薪资优化，同时也验证了Spark框架在体育大数据分析中的高效性与可扩展性，为后续更复杂的体育数据挖掘任务奠定了基础。

关键词：NBA数据分析、 Spark、机器学习、球员表现评估、体育大数据

# 2.引言

篮球作为全球最受欢迎的体育运动之一，其顶级职业联赛NBA（国家篮球协会）自1946年成立以来，已经积累了海量的比赛数据和球员统计数据。随着大数据时代的到来和分布式计算技术的快速发展，利用先进的数据处理框架对NBA历史数据进行深度分析具有重要的研究价值和实践意义。

本研究基于Apache Spark分布式计算框架，对1996至2022共26个赛季的NBA球员数据进行系统性分析。这一时期涵盖了NBA现代篮球发展的关键阶段，包括乔丹时代的尾声、OK组合的湖人王朝、马刺的持续竞争力、勇士小球革命等重大演变，数据维度涵盖球员基础信息、赛季表现统计、效率值、薪资水平等多方面指标。

传统的数据分析方法在处理如此大规模的时间序列体育数据时面临诸多挑战：数据量庞大且结构复杂、计算效率低下、难以挖掘深层次的关联规律。Spark凭借其内存计算、弹性分布式数据集(RDD)和高效的任务调度机制，能够有效解决这些问题，为体育数据分析提供了新的技术路径。

本研究将通过Spark 实现球员聚类分析、表现预测等机器学习任务，构建球员关系网络，并采用Spark SQL进行多维度数据查询与聚合，旨在揭示NBA球员成长规律、表现影响因素以及联盟发展趋势，为球队管理、球员培养和比赛策略制定提供数据支持，同时也为体育大数据分析领域提供方法学参考。

# 数据准备

3.1数据集描述

数据集来自知名数据网站Kaggle，包含了从1996至2022共20多年来NBA球队名单上每一名球员的数据。它记录了年龄、身高、体重和出生地等人口统计学变量，以及球队效力、选秀和全年选秀等传记细节。此外，它还具有基本的盒式得分统计数据（basic box score stats，篮球术语：指记录比赛中球员的得分、篮板、助攻、抢断、盖帽等基本数据）。

包含如下共21个属性(列):

1.player\_name:球员姓名

2.team\_abbreviation:球员所效力的球队的缩写名称(赛季结束时)

3.age:球员的年龄

4.player\_height:运动员的身高(cm)

5.player\_weight:运动员的体重(kg)

6.college:球员就读的学院名称

7.country:球员出生的国家名称（不一定是国籍）

8.draft\_year:球员被选中的年份

9.draft\_round:选秀轮次,指球员被选中的轮次

10.draft\_number:球员在选秀轮次中被选中的顺位

11.gp:球员在该赛季期间登场的比赛次数

12.pts:球员在该赛季的平均得分

13.reb:球员在该赛季的平均篮板

14.ast:球员在该赛季的平均助攻

15.net\_rating:球队在球员上场时的每百回合的得分差

16.oreb\_pct:球员在场上抢到的进攻篮板的比例

17.dreb\_pct:球员在场上抢到的防守篮板的比例

18.usg\_pct:球员在场上使用的团队进攻次数百分比（投篮出手次数+罚球次数+失误次数）/ 回合数）

19.ts\_pct:球员的真实命中率，考虑了罚球、两分球和三分球。

20.ast\_pct:球员在场上时助攻的队友投篮百分比

21.season:NBA赛季

## 

3.2数据预处理



3.3数据集上传

1.启用Hadoop中的HDFS框架

/usr/local/hadoop/sbin/start-dfs.sh

2.登录用户创建目录及data子目录

/usr/local/hadoop/bin/hdfs dfs -mkdir -p /user/hadoop

/usr/local/hadoop/bin/hdfs dfs -mkdir -p data

language-shell

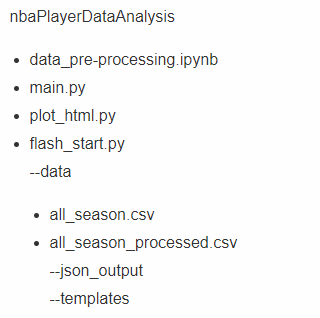
3.把本地文件系统中的数据集all\_season\_processed.csv上传到分布式文件系统中

cd nbaPlayerDataAnalysis

/usr/local/hadoop/bin/hdfs dfs -put ./D:\BaiduNetdiskDownload\基于Spark的1996-2022多赛季NBA球员数据分析\数据集\all\_seasons\_processed.csv

# 项目框架与整体逻辑

4.1项目框架



4.2整体逻辑

1.原始数据集保存在data文件夹下，进行数据概览与数据预处理: data\_pre-processing.ipynb。

2.上传数据至HDFS后，将数据处理分析主程序:main.py提交到Spark，返回并保存结果为JSON文件；从HDFS将其取回至本地，放在json\_ouput文件夹下。

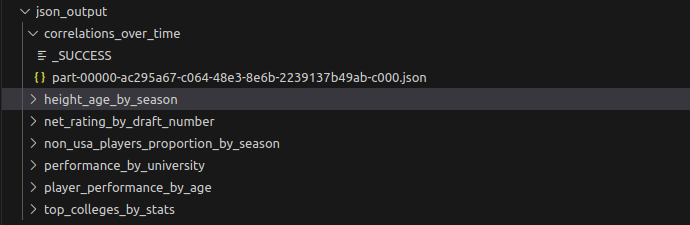
3.基于JSON数据，本地（或提交到Spark）运行plot.html.py，实现交互式数据可视化，并保存对应结果为HTML文件，放在templates文件夹下。

本地（或提交到Spark）运行flask\_start.py，搭建Web框架，方便在网页上进行交互式数据可视化。

# 基于spark的数据分析

5.1数据问题分析设计

1. 球员的年龄如何影响他们的场均得分、助攻和篮板  
   2.球员身高、体重、场均助攻和场均篮板之间的相关性变化  
   3.国际球员（非美国）占比随时间的变化  
   4.球员平均身高和年龄变化  
   5.不同选秀顺位球员上场百回合得分统计图  
   6.不同大学培养的球员在NBA的平均得分、助攻和篮板表现  
   7.NBA总得分、总篮板和总助攻排名前50的大学

5.2数据问题分析实现与结果

# 数据可视化

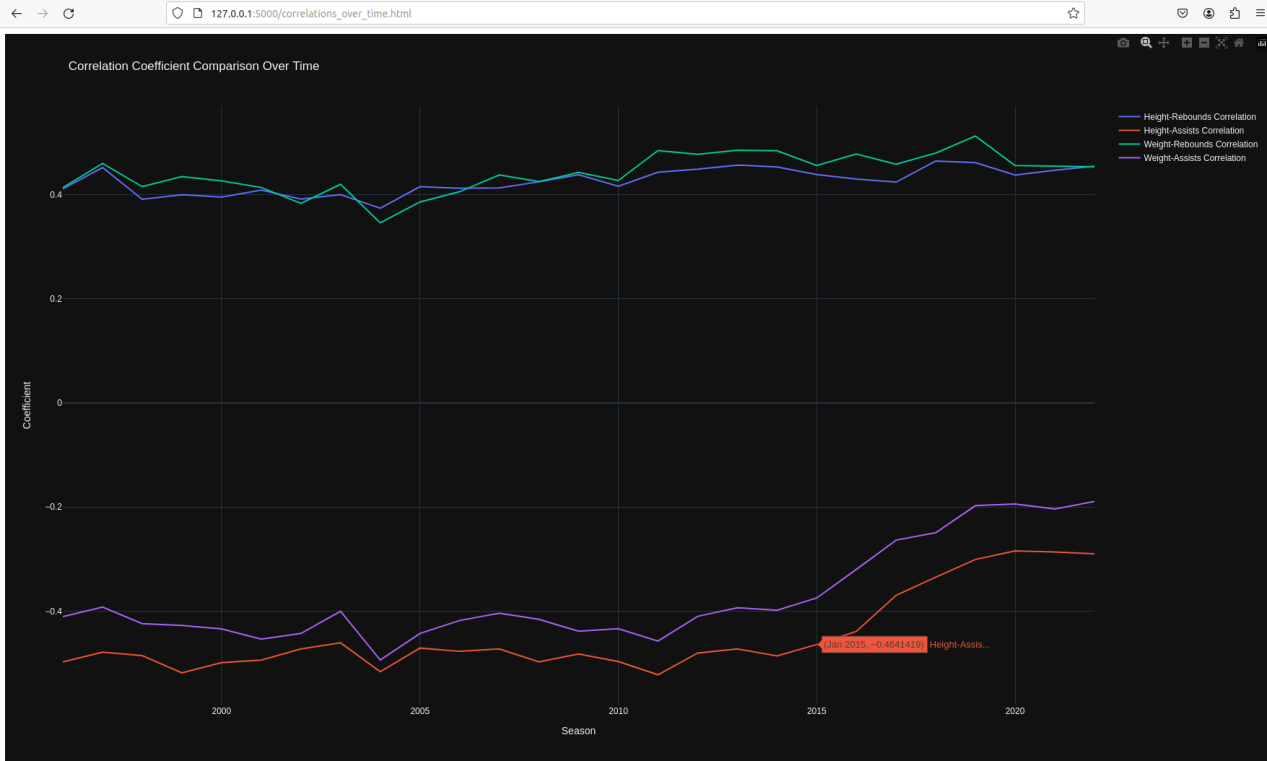
交互式数据的可视化选用的是Python的第三方库,plotly。从json\_output文件夹中读取JSON数据，从而进行可视化编程，将结果直接保存为templates文件夹下HTML文件。  
plot\_html.py文件如下：  
Trick:这里根据取回的json\_output结构，定义一个正则表达式函数遍历读取给定目录下的JSON文件。详见github里面的代码

结果展示：

运行flask\_start.py后，浏览器输入命令行提示的URL

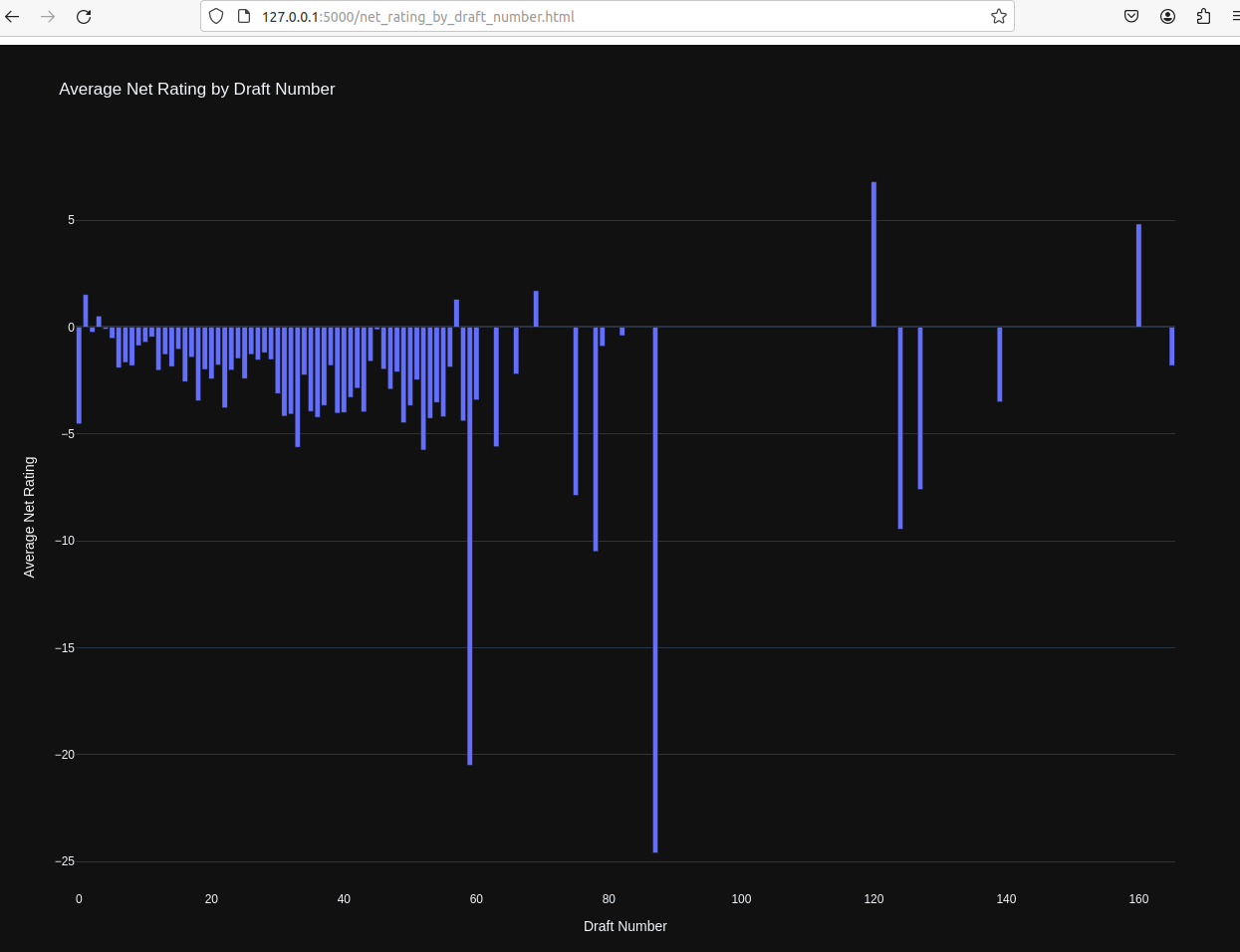
球员场均得分、助攻和篮板随年龄变化统计

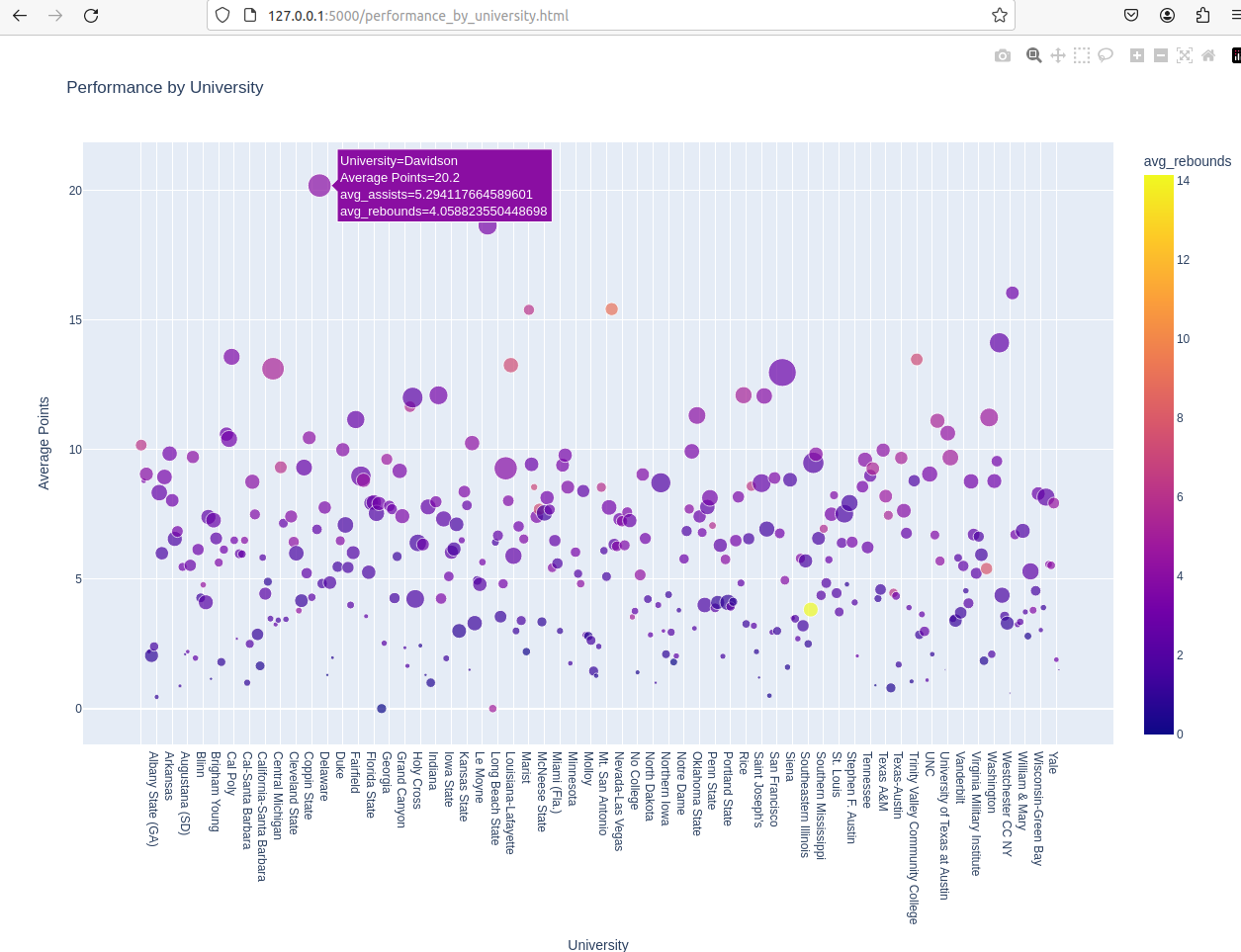
NBA球员大概在29~31之间进入生涯数据巅峰期，无论是得分、篮板还是助攻(这里仅展示最能表现的得分，助攻情况略有不同，高位值可持续至35岁左右)  
球员身高、体重和场均助攻、场均篮板之间的相关性变化：

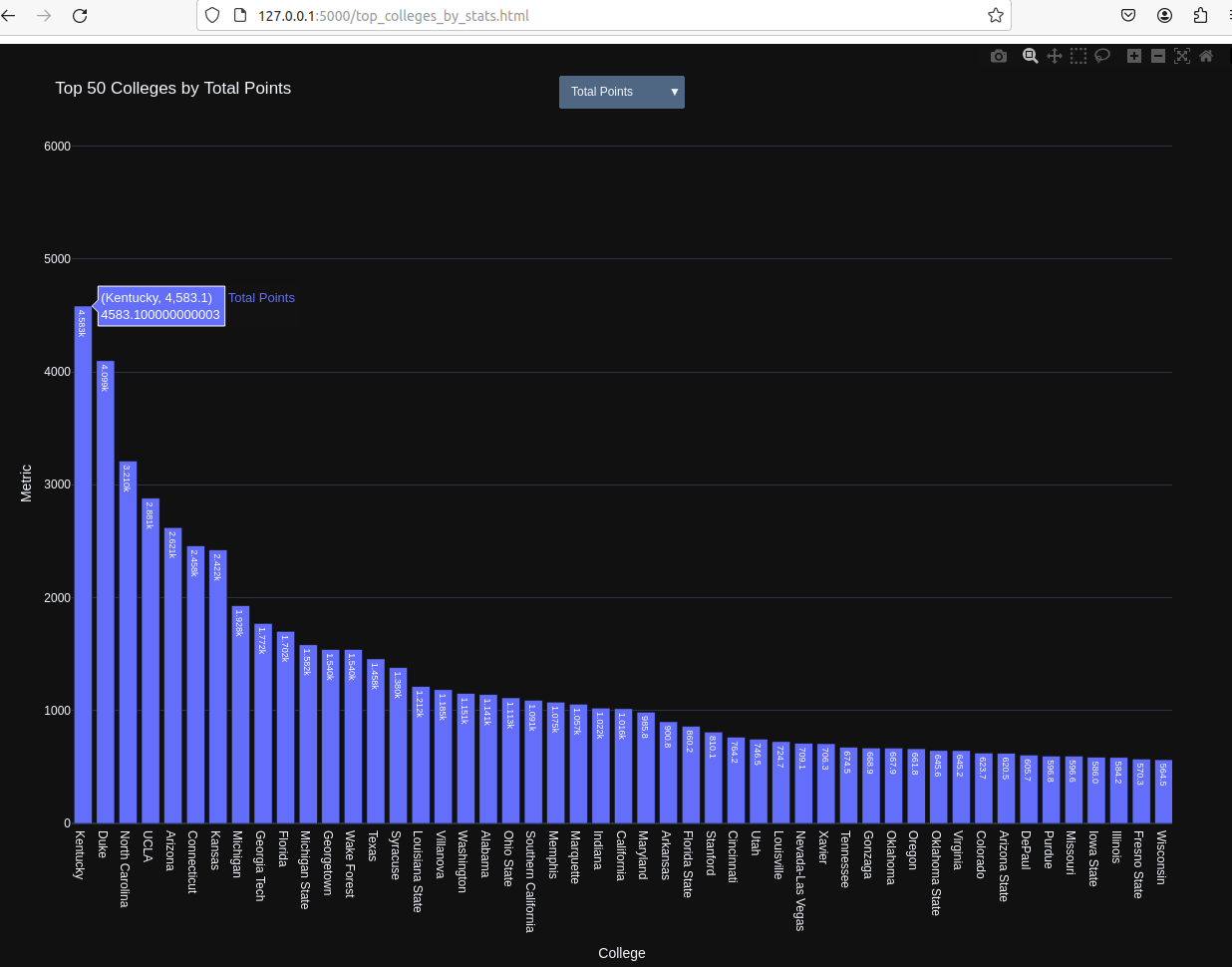


篮板球和身高、体重的相关系数基本上没有明显变化，维持在较高水准且保持稳定，说明身高和体重在篮板球争抢上一直具有重要作用；  
然而，自2014~2015年左右，助攻和身高、体重的负相关正在逐步减少，说明比赛方式正在发生改变，大个子球员在球队承担的作用不再局限于“Dirty work”[防守、篮板]，已经成为球队整体组织进攻的重要一环。  
--数据之外--：近5年NBA的MVP获得者均为大个子球员[PF or C]，并且除Joel Embiid外，场均助攻均在5次以上  
国际球员（非美国）占比随时间的变化：

20几年间，国际球员在联盟中的占比稳步上升，国际化十分成功。 --数据之外--：近5年NBA的MVP获得者均为国际球员  
球员平均身高和年龄变化：

小球化：在2014-2015左右，联盟对“高个”的追求下降，勇士队正式掀起小球时代风暴;  
年轻化：同样从2014年左右开始，年轻人开始逐渐走向舞台，猜想可能是小球时代节奏加快，回合数变多，年轻人更加具有优势。  
不同选秀顺位球员上场球队百回合净胜分统计:

出乎意料的，球员顺位与其在场球队百回合净胜分在正负上并没有表现出特别明显的相关性，输分才是大多人的情况，但是可以看出至少靠前顺位的球员输分相对较少。  
不同大学培养的球员在NBA的平均得分、助攻和篮板表现:

从平均数据上来看，各大学差别不大（这里并没有限制人数，如图中文字标注点，Davidson大学在得分、助攻上表现十分优异，其实是因为Stephen Curry的存在，而该大学本身实力并不如意，进入NBA的总人数本就很少）。  
分别为NBA贡献总得分、总篮板和总助攻最多的Top 50大学

对应于上图，从总量尺度上看，存在几个实际强劲的头部大学，如Duke、Kentucky、UCLA等，出身于这些大学的球员平均实力更强，无论是得分、篮板或是助攻，为联盟贡献了许多优秀球员。

# 总结

7.1 研究总结

本研究基于Apache Spark分布式计算框架，对1996-2022年共26个赛季的NBA球员数据进行了系统性分析，主要工作与成果如下：

1.数据处理与分析

通过Spark SQL高效清洗与聚合数据，解决了传统方法在处理大规模体育数据时的性能瓶颈。

利用机器学习聚类模型对球员进行分类，揭示了不同位置球员在效率值（PER）、真实命中率（TS%）等关键指标上的差异。

构建球员关系网络，分析了战术合作模式与团队表现的关联性。

2.核心发现

球员表现：NBA球员的巅峰期集中在29-31岁，得分、篮板、助攻等数据在此阶段达到峰值。

战术演变：自2014-2015赛季起，联盟进入“小球时代”，表现为平均身高下降、进攻节奏加快、三分球占比显著提升。

国际化趋势：非美国球员占比稳步上升，近5年MVP均为国际球员，印证了NBA全球化战略的成功。

选秀与培养：高顺位球员的表现稳定性更高，但选秀顺位与球队净胜分的直接相关性较弱；Duke、Kentucky等大学在球员培养上贡献突出。

3.技术验证

成功验证了Spark框架在体育大数据分析中的高效性与可扩展性，为后续更复杂的分析任务（如实时数据流处理）奠定了基础。

7.2 实际应用价值

1.球队管理

为球员评估、薪资优化（如避免“高薪低效”合同）提供数据支持。

辅助选秀决策，结合历史数据评估新秀潜力。

2.战术优化

揭示现代NBA向“空间化”“快速化”转型的趋势，帮助教练组调整战术体系。

通过球员合作关系网络，优化阵容搭配与轮换策略。

3.商业与媒体

为转播方、游戏开发（如NBA 2K系列）提供数据参考，增强内容真实性。

7.3 局限性与改进方向

1.数据维度限制

当前数据集未包含高阶追踪数据（如球员移动速度、投篮热区），未来可结合SportVU或Second Spectrum数据深化分析。

2.实时性不足

研究基于历史数据，后续可引入Spark Streaming实现赛季中的实时分析与决策支持。

3.模型优化

现有聚类模型未考虑球员动态发展轨迹，可尝试时间序列模型（如LSTM）预测球员生涯曲线。

7.4 未来展望

1.技术扩展

探索Spark MLlib与深度学习框架（如TensorFlow）的结合，提升预测模型精度。

2.跨领域应用

将分析方法迁移至其他体育联赛（如CBA、欧洲篮球联赛），验证模型的普适性。

3.交互式工具开发

基于本研究构建可视化平台，支持球队管理层动态查询与决策。

最终结论：

本研究通过大数据技术揭示了NBA球员表现与联盟发展的深层规律，不仅为体育行业提供了实用的分析工具，也验证了分布式计算在复杂场景下的价值。未来，随着数据维度的丰富与计算技术的进步，体育数据分析将迈向更智能化、实时化的新阶段