基于 NBA 球员资质综合分析系统的设计和实现

蒋浩杰 张铷钫

四川大学锦城学院 计算机与软件学院 四川 成都 611731

【摘 要】：现在时代快要进入了一个大数据时代，在这个信息量巨大的时代，信息数据也是一种潜在的资源，这个资源的开发就需要自己去挖掘。本文要开发的系统就是基于 hadoop、spark、kafka、flume 等大数据软件的相互连接配合，再利用这个 NBA 球员资质产生的大量数据来分析数据的发展，一定程度上预测未来的球员数据，以及在海量数据中发现一些肉眼难以看见的数据，利用自己设计的计算方法来找到最能代表球员能力的数据，最终以非常通俗易懂的方式来展示数据和利用数据。

【关键词】数据处理；海量数据；可视化；hadoop

1.概述

在中国日新月异的科技发展速度下，中国的很大一部分生活、教育、工作都是信息化的作用点。不管是接收信息的新闻报纸、杂志期刊、热点时事、电视节目，还是散播信息的网课、直播、比赛他们都是信息产生的源头。本文的想法就是在最近十几年非常火热的地方——NBA 这个篮球比赛上去挖掘有用的信息，并且进行分析甚至预测。本文做的方面主要就是将现有的数据进行综合展示，以一种非常直观的方式带给大家，并且在现有的数据上利用本文的预测方法结合大量的数据来预测一些数据增减的趋势。

2.数据分析技术

数据的分析是一个非常巨大的工程，不是成百上千条的数据综合，而是将百万甚至千万上亿天数据进行计算，所以一般的普通软件的数据处理最多算是分析数据，而如果想要做到数据预测，数据量越大就与真实数据越接近。

2.1 Hadoop

Hadoop 是现在最流行的大数据软件之一，它的工作倾向性就是处理巨大量的数据，这也是们想要的。它的工作原理就是将大量的数据分散到不同的区域，并且分布在不同的计算机上计算，利用多个计算机的算力来增加工作的效率，实时检查节点的工作状态，这都很好的保证了数据在计算时的精确性。MapReduce 就是 HDFS 计算上最关键的部分，它也是 Hadoop 中重要的计算单元。[1]

2.2Flume

Flume 也是支持分区（block）筛选的，因为数据量的巨大，分区才能保证数据在高吞吐量的工作下不会出现错误。它的数据处理能力是非常强大的，在巨大的数据量下也可以将需要的数据整理到指定位置储存的，与 HDFS 的交互使用。最后就可以根据条件清理筛选出想要的指定数据，清理掉大量的垃圾数据。

2.3 Kafka

Kafka 是同样是分布式的主要功能是订阅和发布消息。因为它的高数据传输量，和水平扩展方面功能突出所以很多时候都要用到。它有能够监听和处理流动数据，这样保证了大量数据流动时不出错，设定适量的消费者，可以缓存消息，减少因为数据量过大产生的程序奔溃。

2.4 Hive

Hive 在是 Hadoop 上 HDFS 是一种数据结构框架，它有自己的格式，存储的时候数据会进行整理，将数据按照规则的数据结构进行存储，基本上就是整理成一张数据库表，然后就可以像操作数据库表格一样来做增删改查等操作，区别就在于 hive 处理数据量的庞大。它还可以很方便的查看 HDFS 上的表格内容。

2.5 Spark

Spark 主要的功能就是数据的计算，与 hive 一起进行数据计算，他们都是使用 SQL 的计算方法，与 MySQL 的区别主要就在于数据的运算速度和数据量的不同，关键就在于它也可以和 Hadoop 的 HDFS 配合使用。这就联合起来各个软件的优点最终获得最精确的计算数据。

2.6 虚拟机

在实验阶段很多软件需要用到 Linux 系统，这是就需要用到虚拟机，需要在虚拟机和前端之间起到一个连接的作用，首先遇到的难题就是，需要连接虚拟机，并且可以传送数据，运行文件。

3.数据处理

3.1 数据采集和数据模拟

数据的采集就是在网站上进行爬取数据（声明：合法的

利用和爬取数据，仅用于实践研究，不投入商业用途）爬取的数据就是一些源代码，利用 Python 爬虫软件爬取。

爬虫程序的流程：部分代码

url='http://www.nba.com/player/'+str(page)+'.html'

这段代码就是对一个网站的全体网页的一个网址选择，利用变量决定网页的子页，先确定好网页的爬取页面。

然后将爬取的源代码保存到的本地问价中去。这个时候的文件还是很多“杂质”的数据，这个时候需要用到 flume 进行数据的清洗。

3.2 数据的清洗

数据清洗分为两个步骤，第一个步骤就是将数据从网页源代码中分离出来，就像一个值:

<divname=”” value=”” />

这里需要取到 name 和 value 值就需要相应的正则表达式：

.sinks.k[].serializer.regex=[正则表达式]

将数据清洗下来就是想要的类似于：

[姓名][值]

数据清理出来基本上就是类似于表格的数据了。

3.3 数据结构化

存数据之前需要做的就是将数据的格式统一，一般情况下就是统一为 UTF-8 形式的，这样的话在不同环境下进行操作的话就不会出现乱码的情况。

这个时候的数据就是类似于表格但是还不是表格，这一步工作及时将这段数据以表格的形式存入 HDFS，上面说过，hive 有自己的结构，并且就是类似于 SQL 的表格，就可以用到 hive 表进行数据的存储，存为一张 hive 表（HDFS 中）：

LoaddataINPATH'[路径]'OVERWRITEINTOTABLE[表名];

live 表中的数据就可以进行简单的运算了。

3.4 数据运算

在这里有多种方法进行简单的运算：

首先在 spark 中就可以利用 SQL 的部分语句进行增删改查，进行简单的排序或者复杂的数据计算方法，比如说要数据的综合排名时，不是数据的大小进行完全的排序，而是通过参数进行调整计算出一个相对值来排序。

还有一种就是 hive 对数据的查询操作，其实原理和 spark 是一样的，但是操作时的代码不一样，语法不同但是意思是一样的。

3.5 数据分析与预测

通过计算大量的数据可以从中找出一些存在有线性规律的数据，比如说在研究球员的年龄和数据变化规律，这时候咱们能够找到一些规律，这些规律是由大量的数据运算而成的。

根据上述方法，可以大胆的尝试将球队的各项数据进行大量计算得到一定的结果，进行预测。

3.6 数据结果展示

展示是基于 springboot 的前后端连接。下面就是在项目中的一部分代码：

PublicStringgetDate(

HttpServletRequestrequest,

HttpSessionsession,

Modelmodel)throwsIOException{}……}

上面的代码就是在获取网页上用户提交的数据获取，用户可以根据提交的数据可以自定义自己看到的数据排名。

3.5.1 数据图像化

数据展示最直观的展示就是将数据放入表格进行展示，这里用到了网页常用的表格的 Echarts 软件，的代码写在网页的源码里面用 Java 代码注入变量，然后将变量放入 Echarts 的代码中，这样就可以动态展示出计算出数据的结果。

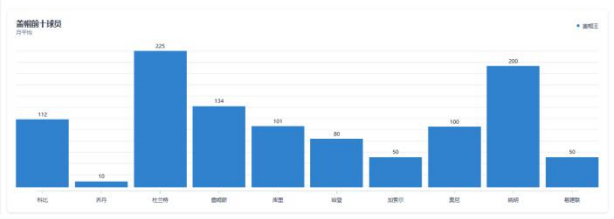
Java 变量注入：

<td><%=data.get(0).getName()%></td>#此处为变量

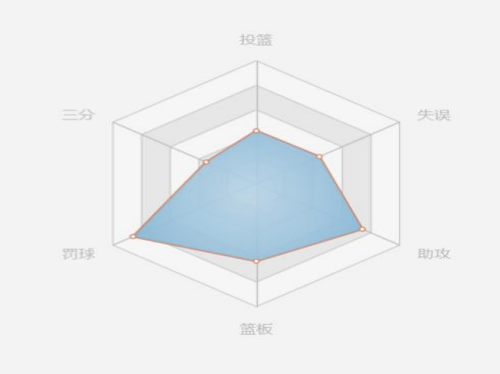
（部分代码，需要完整代码才能实现）

3.5.2Echarts 展示

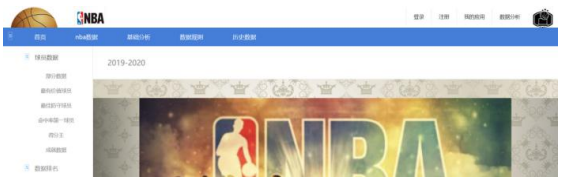
经过后端的变量注入的前端战术如图二，图三



图二



图三



4.成果（维度分析）

这个系统可以做到以下内容：

4.1 可直接展示数据类

可以直接利用 springMVC 的前后端连接的 mybatis 工具进行直接的数据查询，直接连接的数据库输出球员的各项数据，包括排名、平均、众值等都可以直接输出，不需要大量的计算得出规律。

4.2 不可直接展示数据类

4.2.1 分析

首先需要大量的计算之前的所有获得该项球员的数据以及可以量化的表现，得到一些成为综合球员能力值的必要因素，并且计算这些因素占重比，就可以利用大数据的得到各种综合分析的数值系数，球员的各项数据按照不同的系数相计算，得出一个最公正的数值，最终才能表示球员是否有这个能力担任该项奖项。

得分* α +助攻* β +胜场* γ +上场时间* δ +……=综合数据

这里的 α β γ δ 都是变量参数，这些参数就是利用大数据分析得出的，需要用到大量的数据算出一个比较精确的参数，数据量越大得到的数据就越精确。得到这些参数的时候就可以对比所有球员在不同时期的表现最佳等等数据。这也是人们经常头疼的地方，因为没有一个统一的说法，所以利用大数据来计算出数据的参数标准，就可以解决这一问题。

4.2.2 预测

根据数据的各种趋势和一些相关性，分析大量数据还可以得到球员的最佳状态等一系列的问题。

如果预测到一个球员在黄金时期的发展状态就可以计算这样一些参数：1.球员的年龄相对应的球员各项数据和可量化的数据。2.球员的身体素质（体重，身高，臂展，肌肉率等）3.球员的性格。4.球员在球队的地位影响。最后对应的产生一种类似于标准模板的东西，想要预测一个球员的后面几年的价值就可以利用这个模板进行分析，如果数据量大并且精准的话就可以让球队的管理层使用这套系统，可以代替一些人工的分析。[2]

5.总结

这次系统制作目的是将所学的大数据框架运用到 NBA 球员的数据分析与挖掘中。

利用 Hadoop 最重要的一点就是提高计算机的计算能力，大量的数据就需要大量的计算，分布式计算就是在不提升计算机本身的计算能力的情况下，利用软件提高计算效率，这就使大数据的研究不再成为成本的工作。

大数据的发展壮大就是这个时代必然的结果，计算机的普遍应用使得信息量爆炸，每一个行业基本上都能挖掘到有用的信息，如果说将信息利用起来这就是一种宝贵财富，最重要的就是可以直接的创造经济利益。大数据的特点就是数据量巨大，从而找到从混乱之中找到一定的规律,从而达到预测的效果。

参考文献：

[1] 赵健.浅析 Hadoop 的核心技术[A].天津市电子学会、天津市仪器仪表学会.第三十四届中国（天津）2020’IT、网络、信息技术、电子、仪器仪表创新学术会议论文集[C].天津市电子学会、天津市仪器仪表学会:天津市电子学会,2020:4.

[2] 宋薇.基于 Spark 框架与 K-means 的篮球运动数据分析研究[A].中国体育科学学会.第十一届全国体育科学大会论文摘要汇编[C].中国体育科学学会:中国体育科学学会,2019:3.

[3]符添玮.大数据分析关键技术研究[J].大众标准化,2020(02):125-126.