基于 Flume+Kafka 分析球员数据研究

青山科 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】：随着社会信息量的爆炸，在这海量的数据当中，人们如何准确地获取所需要的信息以及将这些信息进行各种各样的处理，如何精确高效直接地获取对自己有帮助的数据显得尤为重要。为了满足以上的需求，本文着重以 flume 和 kafka 之间的数据流通为重点，以爬取和模拟生成的球员信息为样例，快速获取球员资料

【关键词】：数据处理；hive；kafka

1.概述

随着科技的日益革新，在近十年来，网民们大量使用互联网进行信息的涉猎，各式各样的数据涌入人们的眼球，人们通过微博，博客，微信，QQ 等社交平台，获取不同种类的信息。在每年双十一的时候，人们对于网购的需求日益增加，而在网购的背后，又是如何保证数据实时的同步更新的呢？在这些海量信息数据的背后，有着大量大数据工具的支撑。本文以大数据背后的数据传输为重点，着重渗透 Flume 和 Kafka 之间的逻辑关系，以及 flume 和 kafka 之间的接口连接方式，介绍如何配置组件之间的逻辑以及使用的方式。

2.数据的相关分析技术应用组件

2.1 Hadoop

首先 Hadoop 是一种分布式处理的文件系统，主要由各式各样地常应用于大数据开发的软件构成。首先 HDFS 具有非常高的容错性。大数据分析中所有需要处理的数据都将放在 HDFS 中，而这些数据信息也将被 Mapreduce 解决[1]。在 Hadoop 中，还有一些经常会应用到的组件，例如名节点数据节点,以及受主节点支配的 slaves，它的强大之处在于当集群资源过大时，它可以分配给其它节点来完成庞大的任务[2]。但它的缺点在于运行速度相比于 Spark 来讲过慢，时效性不够强，这也是 Hadoop 逐渐被淘汰的重要原因。

2.2 Flume

Flume 也是同 Hadoop 一样是一个分布式的、运行有保障的，处理效率极高的、即可将日志进行合并以及分发的传输系统。它的优势在于它不仅能够收集简单的日志信息，也能收集庞大信息的日志。并且 flume 和 kafka 的结合以及 flume 与 hive，flume 与 hbase 的结合在实际场景中的应用十分广泛。Flume 的缺陷在于不能够实现副本的保存数据如果丢失将无法再次找回。在本研究中主要使用的就是 kafka 与 flume 相结合。

2.3 Kafka

Kafka 同 Hadoop 机制一样可分布式的、可支持分布主题区域的、可复制每个主题内容的消息处理系统。单个 Kafka 服务器能处理几百 MB 的海量数据，并且一个单一集群可作为一个大数据的中心，集中处理；消息被存储到磁盘上时有着强大的容错机制；生产者产生的数据在生产到主题内后能够立刻被消费者消费。它对网络的占用也比较小，对 CPU 和内存的耗费也比较小，因此应用起来特别的稳定。但它是将数据一批一批地发送，数据不能够达到真正的实时发送，有一定的延迟。其次它不具有排除重复消费数据能力，并且一个主题如果有多个分区，顺序将会被打乱无法保证数据的准确。

2.4 Spark

Spark 的整个生态系统成为数据分析栈[3]，spark 的目标就是将批处理、交互处理、数据流式处理等汇聚再一起，其主要使用的编程语言是 Scala。Spark 最为关键的技术应用是 RDD，其中算子是对传入到 spark 中的数据进行操作的函数。并且 spark 的处理及运行速度效率相较于 Hadoop 特别快。并且对于操作性而言 spark 比 Hadoop 操作更加简单。还有一大优点在于 spark 的可靠性，容错性特别强。Spark 的内核 RDD 是一种只对对象进行操作的弹性分布集。Spark 不同于 Hadoop 的地方在于 spark 执行时采用的是多线程模式，而 Hadoop 是多进程模式。并且 spark 处理数据是流式数据，这导致它处理数据时只能处理小批量的数据。

2.5 Hive

Hive 是一种类似于 MySQL 但是存储在 Hadoop 当中的元件。它的优势在于它与 MYSQL 查询语言相似，其此它比 MySQL 强大的地方在于它有着良好的容错性，它的任意节点出问题仍然可以继续工作。Hive 的缺陷在于它的效率比较低，并且在自动生成 MapReduce 作业时，显得不够智能化。

3.数据传输过程及原理

3.1 kafka 与 flume 整合原理

flume 和 kafka 的架构的目的是为了达到实时的数据处理[4]，而 kafka 组件本身由于分批量的发送数据造成在实际的应用当中有一定的延迟，而如果直接对 flume 直接进行计算，如果当 flume 的数据接收速度大于 flume 的数据处理速度则会发生数据的拥塞和堆积，造成内存空间的浪费。因此如果将 flume 和 kafka 相结合，先将数据传入到 flume 内，再通过 flume 流式的计算传入到 Kafka 内部，数据一方面可以与 HADOOP 中的 HDFS 相结合做到离线计算，另一方面，也可与 flume 的功能想照应做到实时计算数据可以多并行多计算。

3.1.1 flume 的组件

Flume 中第一个重要的组件是 source，它是 flume 的接受传入信息的端口，它既可以监听文件又可以监听目录。另一个组件是 channel 为 flume 提供了缓冲的作用，保证了 flume 运行时的稳定以及可靠的作用，与 channel 相连的是 sink 部分，sink 连接的是可以是下一个 source 也可以直接到达目的地。Flume 的组件部分就像是一个水桶作缓存的作用，而另外两根管子就是一根管子进一根管子出。

3.1.1 kafka 的组件

Kafka 的三个主要部分则是由主题以及生产者和消费者组成，类似于 flume 的水桶和管道原理，flume 的信息传入到 kafka 内时，kafka 的生产者将消息接收传入到主题中，消费者再向主题获得消息，不同于 flume 之处在于 kafka 中的主题起到了资源隔离的作用，生产者只能向指定的主题中获取消息，而消费者也只能想指定的主题中获取消息。

3.2 数据爬取及模拟

本实验是先通过在 Pycharm 上使用爬虫爬取 fake 函数生成伪数据，再将伪书据传入到虚拟机上，在虚拟机上使用 flume 和 kafka 组件，将数据存入到 flume 中再与 kafka 的接口相连接。

数据爬虫主要代码：

```
defsetHtml(firstNum):

url="https://nba.hupu.com/stats/players/pts/"+str(firstNum)

response=requests.get(url=url)
```

对于爬虫代码，需要准备 python 的编译环境，并安装 request 的依赖库包，python 的大部分代码都是通过加载库里面的函数来实现的，之后在 setnum 函数中通过输入球员的姓名便可以直接爬取到球员在虎扑内的所有数据信息。

数据模拟主要代码

```
fake=Factory().create('zh_CN')

withopen("shuju.txt","w")asf:

team=[“所有球队名”]
```

数据模拟中主要通过 faker 函数来完成，faker 函数的作用在于能随机生成球员的姓名，并在 team 数组中加入所有球队名，在 data 函数中包含了球员的各项数据信息

3.3 Flume 的配置主要代码

数据模拟完成后，将生成数据传入虚拟机，再进行 Flume 的配置。

描述和配置 source 组件

```
a1.sources.r1.type=spooldir

a1.sources.r1.spoolDir=/home/test
```

描述和配置 sink 组件

```
a1.sinks.k1.type=KafkaSink

a1.sinks.k1.kafka.topic=flumeTest
```

Flume 配置拦截器过滤系统

拦截器整数超过预期

```
a1.sources.interceptors.i1.regex=[7-9][0-9](?!\.|%)
```

Flume 配置的关键之处在于要指定好监听的文件或者目录，该研究中数据是从网页上爬虫加模拟数据一起传入到虚拟机内，数据传入到虚拟机内时，Flume 就直接开始监听传入的内容，并且通过 Flume 配置的 sink 组件与 kafka 直接相连，传入到 Kafka 的 flumeTest 的主题内，Kafka 的生产者可以直接开始接受 Flume 传出进来的信息，由 Kafka 的消费者传到 hive 数据表内。

3.4Kafka 配置

在启动 flume 后，才能紧接着启动 kafka，kafka 启动时需要先启动 broker，再启动生产者消费者，这个顺序是不能改变的。在 kafka 中还支持多副本多分区的执行方式，这样做的优势在于当一个 broker 崩掉的时候，其他的 broker 还能接着继续工作。

3.5 数据展示

3.5.1 程序框架软件

本次研究主要涉及关于后端 flume 以及 kafka 及 hive 的层面。前端使用墨刀构建网页，再将墨刀网页转换成 jsp 与 springboot 相结合，形成一套前端系统，再与虚拟机上后端相连接。这次研究所用到的数据架构模型如图 1 所示。

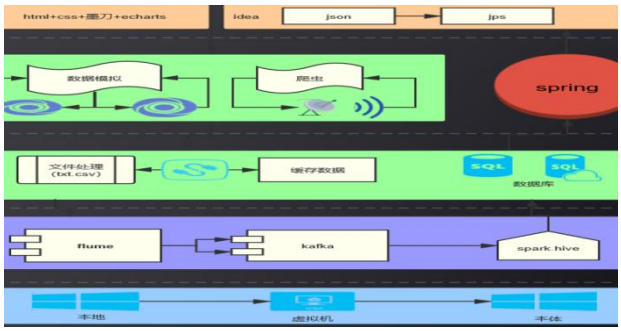

图 1

3.5.2 程序设计和展示

前端展示主要通过 echarts,其优势在于便于进行开发和阅读文档，并且里面有着丰富的图和表格，易于将数据直观，优雅地呈现在面前。在筛选数据时，只需要选择需要查询的信息，数据便可以直接得出。在图 2 中可以清晰地看到 echarts 表格得到的数据。

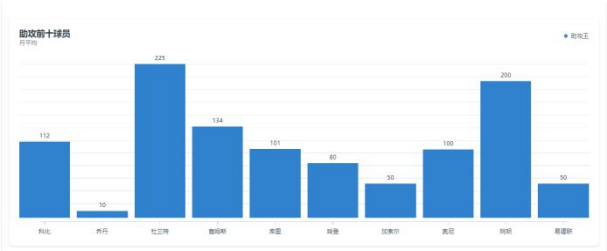

图 2

总结

通过本次关于 Flume 和 Kafka 以及 hive 的相关组件配置及打通，而这一系列的组件当中，有 Flume 的日志监听以及数据的过滤，也有 Kafka 的数据中生产者的保存及消费者的消费，也有类似于 mysql 的 hive 数据库，也要存储更为强大的 hbase，而这一切也都是源于 Hadoop 的普及。本次研究也存在许多的缺陷，由于本次项目的几大运算模块，Flume，Kafka，hive 在之前都是相对孤立的运行模块。因此怎么将每个部分连接起来就成了必须要面对的问题。对于技术层面，还需要做到能够一键式，从前后端一键获取到所需要的信息，而不是先将数据从前端传入到后端，再进行数据分析得出数据。对于这一套流程，如果在以后的开发当中能够使用 shell 脚本来实现，获取信息的速度又将会大大提高。在 echarts 的应用方面，还有更多美观，直观的可视化选择加以应用，它上面的丰富的可视化类型，多渲染方案以及跨平台使用对于做前端开发能提供许多的帮助。

对于未来，大数据对于打造一座城市的科技智慧是起到至关重要的作用[5]，大数据能够对城市服务，工商业的发展提供便利，而大数据发展在未来仍需要攻克的主要问题依旧是如何进行大量数据的清洗以及时效性，只有真正掌握到了数据，才能够了解我们生活的社会是怎么样的。

参考文献：

[1] 董楠楠,单晓欢,牟有静.基于 Hadoop 和 MapReduce 的大数据处理系统设计与实现[J].信息通信,2020(06):29-31.

[2] 张趁香.基于 Hadoop 平台的海量数据分析和处理[J].电脑编程技巧与维护,2019(01):95-97.

[3] 许丹亚,王晶,王利,张伟功.基于 Spark 的大数据访存行为跨层分析工具[J].计算机研究与发展,2020,57(06):1179-1190.

[4] 陈军.基于 Flume 的分布式日志聚合系统的研究[J].科技视界,2017(11):77+114.

[5] 王伟.大数据技术与“互联网+健康”产业发展[J].现代商业,2020(23):39-40.