基于 Kafka 对 Python 模拟产生的动态金融数据的分析

◆晋　宇　邵　煜

（宁波财经学院　浙江宁波　315175）

摘要：随着时代的发展，数据已经渐渐地渗透于我们生活的每一个地方，而我们对于数据的使用和分析也越来越频繁。对于数据的抓取与分析显得格外的重要。如今，传统的数据的抓取 与分析已经无法满足日益增长的科技发展了。我们需要一个快速，简洁，方便，高吞吐量，可实时消费的高性能分布式消息系统。本文从 Python 对数据的抓取，Kafka 对数据的整合，以及 NS3 对数据的分析来简单描述 Kafka 的消息系统。

关键词：Kafka 分布式发布订阅消息系统；Python；NS3

1.研究意义

时代的发展，电子数据已经渐渐地渗透于我们生活的每一个地方，传统的数据收集和分析模式已经无法满足企业的发展。因此，用 Python 对数据进行抓取，用 Kafka 对数据进行分析，可以了解我国国内金融行业的现状，小而言之，也可以了解金融行业中的一部分，比如对股票进行分析，以判断可以购买哪支股票。

2.Kafka 的概念及优势

Kafka 最初由 LinkedIn 公司开发，之后成为 Apache 软件基金开发的一个开源流处理平台。它使用 Scala 编写，因其高吞吐率而被广泛使用。Kafka 凭借着自身的优势，受到互联网企业的青睐。在国内，唯品会也采用 Kafka 作为其内部核心消息引擎之一。

Kafka 是一个新颖的分布式的消息订阅和发布的系统，能够实时和离线对数据进行处理。同时也具有自己独特的设计优势：

1）高吞吐量（主要优势）。Kafka 被创立出来的初衷就是为了能够有效、快速的提高大量数据抓取和分析。并且，Kafka 即使在普通的硬件上，也能够支持每秒数百万的消息。

2）数据的持久化存储。对数据可持久化到磁盘，用于批量消费，防止数据丢失。

3）利用 zookeeper 确保服务的可行性。通过 zookeeper 管理协调数据的请求，将数据进行转发并进行备份。

3.Kafka 应用于 Python 模拟产生动态的金融数据的分析

3.1 Python 对模拟产生动态的金融数据的采集

通过 Python 中使用 urllib2 来支持 HTTP 通信协议的实现。用 URL 参数指示一个要下载的资源路径。当数据参数为空时，表示将发出一个 GET 类型的请求，该请求不包含任何实体；当数据参数为非空时，预示着将发出一个 POST 类型的请求，数据的内容为请求的实体内容。可以自动地进行抓取网络的金融数据，并对数据进行采集[2]。或者通过网络爬虫按照一定的规则对数据进行自动的抓取。按照行业领域划分，分为股票、证券、债券、期货等理财数据，P2P 数据，电子商务数据等类型[3]。但以这样的方式取得的金融网页的页面数据，很有可能出现数据粗糙，错误的字符编码和无序的结构等现象。所以，首先要确定文档的字符编码，可以通过<head>中的 content-type 元得到。然后将其解码成 uncode 类型[4]，以保证数据存储的方便。

3.2 Kafka 对 Python 抓取数据的处理

3.2.1 Kafka 和 Python 产生的问题与解决方案

Kafka 和 Python 均可以对数据进行抓取，均需要一定的 java 编程基础，甚至于 Kafka 和 Python 均对数据可以进行深入的分析。但在数据采集上，Python 更加的方便。因此，如若对于编程不是特别熟悉的人，可以选择用 Python 进行数据采集，它使用的语言清晰简练，而且易于理解，即使不是专业的编程人员也能够理解程序的含义。但是同样的，Python 语言存在性能不足的缺点。在面对大量的数据时，Python 的数据分析效率不是很高，甚至于可能会崩溃。而 Kafka 正好可以弥补这一缺点，为数据的分析提供强大的支持。并且 Python 经过一代代的开发研究，生成了 kafka-python 库，可以通过一定的方式与 Kafka 进行连接，实现了与 Kafka 之间的数据交互。

当然，在数据的传递时也会产生一定的问题，比如，生产的消息因多次创建 Kafka-Producer 产生的问题。这会使得抓取的数据因为这个问题而丢失。

3.2.2 多次创建 Kafka 的 Producer 产生的问题与解决方案

由于 Kafka-Python 将数据传输给 Kafka，它将产生一条消息，发布者需要多次创建该消息才能单独发送给消费者，但在多次创建发布者时会产生一定的错误，无法继续创建新 Kafka 生成器。产生错误的原因是因为每次创建一个新的 Kafka 生成器都会占用一个文件符号，这是因为 controllen 结束时，没有释放导致的。因此，我们可以创建一个用于控制的全局 Kafka 生成器。

3.3 Kafka 对模拟产生动态的金融数据的处理

通过上述数据采集的方法，采集而来的数据都是粗糙的，因此，我们可以通过 Kafka 对数据进行一遍整理。用 NS3 节点类[5]的方式对数据进行简单的处理。根据 Kafka 的分布式发布订阅消息系统基本构架，可以分别设置生产者、代理者、消费者这 3 个节点。

针对于大数据的交互会有一个管理者来对这样的大型分布式的系统进行协调服务[6]，用它来协调控制分布式网络中各个节点的通信，维护系统的负载均衡[7]，保证最大程度减轻代理系统的通信压力，提高系统的性能。

最后我们可以设计一个特定的场景，比如添加 2 个或者以上的生产者，3 个或以上的代理者，2 个或以上的消费者，设置消息大小为 100 字节，让生产者分别发布 80、100、300 条消息，并让消费者以随机的方式进行分配，最后，通过选取其中一个代理点和一个消费者进行数据的分析，并实时抓取的不同时间点的数据分析图或表。

结束语

每一款软件具有它的优点，我们应该发挥的优点，与其他可以相关联的软件一起用，使得数据得到有效的分析。像 Python 用于捕捉数据速度算快，也方便，不过对于数据的整合上却显得很无力，因而我们可以选用 Kafka 来对数据进行整合，并进行分析。在动态数据上，Kafka 对于动态数据的整合也能够使它达到我们预期的效果。如若对于 Kafka 使用并不熟练者也可以通过数据整合之后，将数据导出放置于 Spass 中进行简单的数据分析。

参考文献：

[1]赫特兰. Python 基础教程[M].2 版.北京：人民邮电出版社，2010.

[2]齐 鹏，李隐峰，宋玉伟.基于 Python 的 Web 数据采集技术[J].2012，25(11):118-120.

[3]王蕾，安英博，刘佳杰.基于 Python 的互联网金融数据采集[J],2017，(9):47-49.

[4]鲁特兹 . Python 学习手册[M]. 北京：机械工业出版社，2009.

[5]马浩然. 基于 NS3 的分布式消息系统 Kafka 的仿真实现[J].2015，(1):94-99.

[6]莫磊，胥布工. 基于分布式估计及任务分配的 WSANs 协同机制[J].新型工业化,2013，(12):15-27.

[7]蒋占军，李成，李磊等. 分布式无线通信系统中并行 Round Robin 调度算法研究[J].新型工业化,，2011，(10):103-111.

[8]杨国龙.企业间大数据推荐引流系统研究与设计[D].湖南大学，2016.

[9]周铁峰.基于大数据的用户电信息采集系统的设计与实现[D].华北电力大学，2018.

基金项目：宁波财经学院校内科研项目（1042218022）。