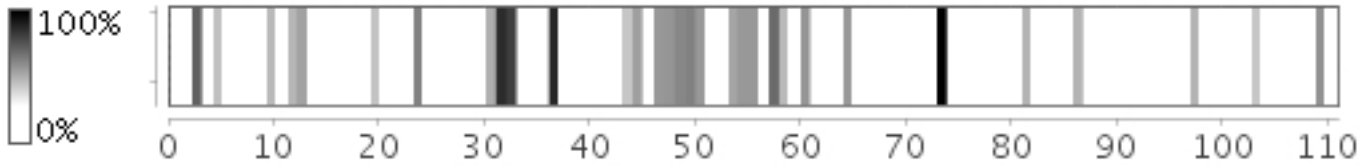
PaperPass专业版检测报告
简明打印版

比对结果（相似度）：
总　体：16 % *（总体相似度是指本地库、互联网的综合比对结果）*
本地库：11 % *（本地库相似度是指论文与学术期刊、学位论文、会议论文数据库的比对结果）*
期刊库：7 % *（期刊库相似度是指论文与学术期刊库的比对结果）*
学位库：5 % *（学位库相似度是指论文与学位论文库的比对结果）*
会议库： 1 % *（会议库相似度是指论文与会议论文库的比对结果）*
互联网：10 % *（互联网相似度是指论文与互联网资源的比对结果）*

编　号：58C0012D9ED31JMR5
版　本：专业版
标　题：第一章
作　者：王江波
长　度：5503 字符(不计空格)
句子数：111句
时　间：2017-3-8 21:03:41
比对库：学术期刊、学位论文（硕博库）、会议论文、互联网资源
查真伪：http://www.paperpass.com/check

句子相似度分布图：



本地库相似资源列表（学术期刊、学位论文、会议论文）：
1. **相似度：4 % 篇名:《大数据流式计算:关键技术及系统实例》**
   来源：学术期刊 《软件学报》 2014年4期 作者: 孙大为 张广艳 郑纬民
2. **相似度：1 % 篇名:《云平台下分布式文件系统评测技术研究》**
   来源：学位论文 哈尔滨工业大学 2014 作者: 胡军杰
3. **相似度：1 % 篇名:《面向海量遥感影像数据的分布式文件系统管理技术研究》**
   来源：学位论文 兰州交通大学 2012 作者: 王旭东
4. **相似度：1 % 篇名:《应用于OpenStack平台的无元数据服务器的海量网络存储系统设计》**
   来源：学位论文 浙江大学 2014 作者: 单旦骏
5. **相似度：1 % 篇名:《基于Nginx高性能Web服务器性能优化与负载均衡的改进与实现》**
   来源：学位论文 电子科技大学 2015 作者: 王永辉

6. 相似度：1％ 篇名:《一种平台无关的并行编程模型的设计与实现》
   来源：学位论文 中国科学技术大学 2014 作者: 李婷
7. 相似度：1％ 篇名:《MapReduce并行编程模型研究综述》
   来源：学术期刊 《电子学报》 2011年11期 作者: 李建江 崔健 王聃 严林 黄义双
8. 相似度：1％ 篇名:《实时数据处理在钢铁生产质量监控中的应用》
   来源：学位论文 武汉科技大学 2015 作者: 李保连
9. 相似度：1％ 篇名:《基于云平台的城市照明设备分布式综合管理系统设计》
   来源：学术期刊 《软件导刊》 2015年9期 作者: 喻宏进 徐源 李朋

互联网相似资源列表：

1. 相似度：8％ 标题：《大数据流式计算：关键技术及系统实例 - cador的专栏...》
   http://blog.csdn.net/u013524655/article/details/41073759
2. 相似度：3％ 标题：《研究 大数据分析在金融领域是如何应用的？-搜狐》
   http://mt.sohu.com/20160904/n467599346.shtml
3. 相似度：2％ 标题：《在 5 分钟内使用 Node-RED 构建实时的聊天应用程序 - 推酷》
   http://www.tuicool.com/articles/Jfaylj
4. 相似度：1％ 标题：《使用 Node-RED 和 IBM Bluemix Push 服务自...》
   http://www.tuicool.com/articles/iiaQniz

全文简明报告：

第一章绪 论

1.1 研究背景与意义

{71％：近年来，由于云计算[1]、物联网、移动互联、社交媒体等信息技术和应用模式的快速发展，不断地推动人类社会迈向大数据时代。} 早在2010年，全球的数据量就已经具有ZB级的规模，有预测显示，到2020年全球的数据量将达到35 ZB， {42％：大量数据无时无刻地影响着人们的生活、工作，甚至是社会的发展和国家经济，大数据时代已经到来。} 而近年来，有关大数据方面的研究和应用也越来越广泛，新形式下的大数据技术为我们分析问题和解决问题提供了新的思路和方法，其研究已经成为业界的热点。

大数据的分析计算模式主要分为批量计算（batch computing）、流式计算（stream computing）、交互式计算（interactive computing）、图形计算（graph computing）等等。 其中批量计算和流式计算[2，3，4]这两种计算模式不管是在学术界还是在工业界都是主要的研究模式，同时各自都有广泛的大数据应用场景。 其中批量计算是一种适用于大估摸并行批量处理作业的分布式计算模式，也就是我们大家都十分熟悉的MapReduce计算模式。 {45％：MapReduce本身是一种编程模型，这种编程思想有着广泛的应用，尤其在大规模数据集的并行计算中，} 由于其简单易用性的特点使得它成为目前最为流行的大数据并行处理模型[5，6]。 {45％：后来，在开源社区的努力下，Hadoop系统[5]应运而生，在Hadoop系统中包括HDFS（hadoop分布式文件系统）和MapReduce两个核心组件，} {52％：HDFS用于存储海量的数据，而MapReduce是用于海量数据的并行处理。} Hadoop平台的应用也十分广泛，国内外许多企业都在用Hadoop平台来进行大数据处理。 此外，Spark系统[7]也具备批处理计算的能力。 而对于流式数据计算，它是一种对实时性要求极高的计算模式，由于数据的到来是不确定的、无序的、不间断的， 为了避免在数据处理过程中造成数据的大量堆积或者数据丢失，这就要求流式计算必须在指定时间限度内对系统所产生的新数据完成实时处理。 在许多行业的大数据应用系统中，比如金融银行业的业务监控系统、政府政

务管理系统、道路监控系统、互联网行业的访问日志处理等， 在这些应用系统中不仅大量累计的历史数据，同时还具有高流量的实时流式数据，因而在提供批处理计算模式的同时， {41%：系统还需要能具备高实时性的流式计算能力。} 因此，研究和设计一套高效，稳定的流式数据处理模型具有广泛的应用价值，目前也有比较流行的流式计算系统， 比如像 Twitter公司的 Storm、 Yahoo公司的 S4以及 Apache Spark Steaming[7]。

在传统的流式计算模型中，绝大多数都是利用数据库来实现的，而在大数据时代下的流式计算有了新的需求，表现在低时延、高带宽等。 {61%：所以，如何构建一个低时延、高带宽、持续可靠、长期运行的大数据流式计算系统成为了当前亟待解决的问题。} Redis这种基于内存计算的、可进行数据持久化的Key-Value存储系统[8]的诞生，为大数据流式计算提供了一个很好的解决方案。 Redis数据库最初是为了解决像 SNS类网站在数据存取过程中的实时性等刚性需求的，而传统的关系型数据库越来越难以胜任了， 这也使得 Redis这种数据库也越来越受到人们的关注。 如今Redis数据库已经得到了广泛的应用，不论是在高速缓存系统中，还是在海量文件的实时检索中，甚至是在如何如荼的各种推荐系统中，Redis都起着举足轻重的作用。 Redis基于内存的数据计算和高效的数据存储策略也能够很好的满足实时流计算问题中的低时延的刚性需求。 因此，研究redis的内存计算以及存储策略并将其运用到实时流式计算模型中具有重要的意义和实用价值。

{48%：在流式数据处理中，因为无法确定数据是什么时候到来，按什么顺序到来，因此，不需要事先对流式数据进行存储，} {88%：而是当流动的数据到来后在内存中直接进行数据的实时计算和分析。} {81%：就像我们熟悉的Twitter的Storm、Yahoo的S4就是典型的流式数据处理框架，数据在任务拓扑中被计算，最后输出有价值的信息。} 目前这些流行的流式处理框架都有一个共同的缺点就是，没有一个方便的能够快速根据业务构建数据任务的拓扑计算流程，也就是我们所说的计算流（flow），同时也缺乏数据的流化功能。 Node- red本身是 node. js开发的，支持 node. js的事件驱动和非阻塞 IO机制， 是一种可视化流程编辑框架[9]，它允许开发人员仅仅使用一个基于浏览器的可视化界面流程编辑器来完成设备、服务器以及 API应用的连接。 {88%：Node-red本身是IBM Emerging Technology团队创建的一个新型开源工具，它允许用户通过组合各种部件来编写应用程序，这些部件可以是硬件设备、Web API或者是在线服务。} Node-red被广泛用于物联网领域，实现数据的流式传输。 在 Node-red中从数据的接入，到数据的解析分析，最后到结果的输出都是通过各种各样的节点来完成的， IBM Emerging Technology团队在开发这个工具的时候只引入了少量的具有特殊功能的节点，比如常用的 http节点、 tcp节点、 udp节点、 debug等数据输入输出节点， 还有一些用于数据分析的节点比如 sentiment节点，还有一些用于访问存储设备的节点，如 mongodb节点； Node-red除了原始已经提供的这些节点外，还允许用户自己按照开发原则开发自己需要的节点。 为了能够充分利用 Node- red的可视化流程编辑的直观性，结合 Rredis数据库的内存计算的特点，{41%：探索开发适应于流式数据分析的数据输入节点、数据输出节点、数据处理节点以及 Redis数据库访问节点，} {53%：这对流式数据分析有着重要的实际意义。}

1.2 国内外研究现状

{55%：1.2.1 实时流数据处理模型的研究应用现状}

{56%：大数据时代下的数据处理主要的两种方式就是实时流数据处理和批量数据处理。} {60%：实时流数据处理主要适合于那些无需事先进行数据存储，可以直接进行数据分析处理，实时性要求比较严格，但数据的准确度要求比较宽松的应用场景。} {62%：而对于传统的批量数据处理，首先要进行数据的存储，然后再对存储的静态数据进行集中或者分布式计算。} {55%：目前，对于传统的批量数据处理模型的技术和研究成果已经相对成熟了，最初有 Google公司的 MapReduce并行编程模型[5]的提出，} 再有后来在开源社区的努力下开发的 Hadoop系统为代表的批处理系统，都已经是稳定而高效的批处理系统。 而对于流式数据处理模型的研究仅仅处于一个初级阶段，在早期关于流式数据的研究也主要集中在以数据库为中心而开展的， {51%：主要是研究了数据计算的流式化，数据规模也比较小，数据对象也比较单一，很难适应在大数据时代下流式数据处理所呈现出来的

新特性。｝｛54％：因为，在新时期的流式数据主要呈现出实时性、突发性、无序性等特点，对新的流式计算系统就有了更高更严格的要求。｝

｛54％：在国外，Yahoo推出了S4流式数据处理系统，随后在2011年，Twitter也推出了自己的流式数据处理系统Storm，｝还有就是近年来开源社区新兴的MOA（Massive Online Analysis）、Spark Stream都是流式处理系统，｛69％：这在一定程度上推动了流式数据处理的发展和应用。｝｛45％：但是像S4、Strom这样的流式数据处理系统在可伸缩性、容错性、数据吞吐量等方面存在着明显的不足，｝而对于MOA，Spark Stream这样的系统，虽然功能和API十分丰富，但是在稳定性和易用性上不尽如人意。｛56％：所以，如果构建一个低延迟、高吞吐、易用且能持续可靠地运行的流式数据处理系统，是一个亟待解决的问题。｝

在国内，目前关于流式数据处理模型的研究还比较少，但目前国内主要有百度公司自主研发的Dstream和TM实时计算平台，在学术界主要是有一些关于流式数据挖掘算法的研究。 但是，流式数据的可视化分析已经在很多场景得到了应用，比如各大银行都陆续建立的大屏监控系统， 就是实时地监控银行的业务状况、系统运行状况、用户行为分析等，又比如政府网站群的监控， ｛54％：也是通过实时监控网站的访问数据，分析用户的行为。｝在这些应用的背后，如何建立一个高效、稳定、易于维护的实时处理模型显得尤为重要。

1.2.2 Node-red的研究应用现状

Node-red作为一种在物联网时代的新型产物，是一种用来快速搭建物联网应用程序的流式处理框架，在信息无处不在的时代，Node-red也越来越受到业界的关注和研究。

它是由IBM Emerging Technologies团队发起的一个开源项目，其中Nick Leary和Dave Conway-Jones工程师为Node-red的设计和开发做出了巨大的贡献。 2013年，Node-red以开源项目的形式被发布，经过短短几年的发展，Node-red已经拥有了一大批活跃的用户和开发人员。 Node-red依然是一个新型科技，时至今日，但凡用过Node-red的制造商、实验人员和一大批大大小小的公司，都已经见证了Node-red极具价值的应用之处。

在国外，IBM公司率先将Node-red应用起来，Node-red被集成到IBM公司的最新的云产品Bluemix上。 通过Bluemix提供的云服务，用Node-red来建立和管理一个实例（也就是一个应用流程），就可以实现消息的推送服务。｛100％：Node-RED的使用，与Bluemix中简单的Push服务相结合，使整个流程变得非常简单，需要调整的部分也少得多。｝

在国内，目前也有很多智能设备制造公司在使用Node- red，可以很方便地通过Node- red节点来控制硬件设备的状态， 比如拿Node- RED搭配Arduino，是一个快速原型化的好用工具，例如控制RPI的某根管脚位去点亮LED， 只要简单的拉四个节点，串一串再写一点程序代码即可做到。 因为Node-red还在进一步完善当中，原始开发的节点可能很难满足实际的需求，所以，我们在运用Node-red来管理数据流程的时候，还需要自己开发需要的功能节点。 在这一点上，目前在不少银行的业务监控系统中引入了Redis的访问节点。

1.2.3 Redis的研究应用现状

Redis作为存储系统[11]之中的后起之秀，由于其数据结构丰富、基于内存计算、支持网络又可进行数据持久化等特点，迅速为许多企业和开发者所爱戴。 ｛47％：不论是在学术界还是在工业界，对Redis的研究都从未停止过。｝

Redis是由Salvatore Sanfilippo为实时统计系统LLOOGG量身定制的一个数据库，在2009年的时候将Redis开源发

布， 并开始于另外一位 Redis代码贡献者 Pieter Noordhuis一起继续 Redis的开发，直到现在， Redis的代码托管在 GitHub上，并且开发也十分活跃。 随着Redis内存数据库的发布，经过短短几年的发展，Redis已经拥有了一大批活跃的用户和开发人员。 {46%：在国外，像GitHub、Viacom、Pinterest等都是Redis的用户，Github利用Redis集群，来统计用户项目的跟进状况。} 而在国内，新浪在研究了 Redis数据库的源码后，搭建了有号称史上最大的 Redis集群，实现了 传统的 SQL数据库难以实现的计数分析（counting）、反向缓存（reverse cache）、 top10 list等功能。 近年来，也有不少银行，在自己的实时数据监控平台引入了 Redis数据库，实现了数据的实时处理和分析， 还有就是随着国家电子政务系统的逐渐推行，不少的地方政府也在自己的数据中心监控系统中引入了 Redis数据库， 来实现数据实时计算和处理。

1.3 论文主要工作和研究内容

本文对大数据背景下流式数据处理过程中所遇到的挑战和难题进行了研究分析，详细研究了 Node- red流式处理框架的编程模型和消息推送机制， Redis数据库的实现原理及其基于内存计算的原理。 设计了一种新的基于 Node-red的流式管理和Redis的内存计算的流式数据处理模型，并通过实现网站访问实时监控系统来验证了该模型的可行性。 主要工作内容如下：

{46%：（1）本文首先对当前实时流数据处理模型的研究应用现状以及 Node- red与 Redis的研究应用现状进行分析，} 同时结合 node. js的事件驱动与非阻塞机制详细阐述 Node- red的消息推送原理。

（2）对Redis数据库做了深入研究。 因为在流式数据处理中，经常会遇到关于最大值，最小值，累计求和等指标的计算，而去重统计是计算这些指标的基础。 因此，本文通过分析Redis有序集合的源码，结合Skip List的基本原理，提出了基于Redis有序集合的去重统计方法。

（3）在研究分析了流式数据的特点和流式数据处理的基本原理后，结合 Node- red的编程模型和消息推送机制， {40%：设计了一种新的基于 Node- red的流式管理和 Redis的内存计算的流式数据处理模型。} 由于原始的 Node- red缺乏对 Redis数据库的访问节点以及 Redis的 pub/ sub节点， 重新设计了新的数据输入、输出节点以及数据处理函数节点（function_ node）， 并安装部署到 Node- red框架当中，实现数据的流式处理和数据流的管理。

（4）本文最后还将设计好的流式数据处理模型，应用到实际生产环境中加以验证。 使用该模型对某政府网站的访问流量数据进行实时监控分析，设计了一套实时数据监控系统，该系统包括了数据的实时采集、实时分析和处理， {56%：以及最后的数据可视化展示，并对结果进行了有效性分析。} 实现了从模型设计到模型应用的全过程。