国 家 知 识 产 权 局

200233

上海桂平路 435 号 上海专利商标事务所有限公司
施浩(021-34183200)

发文日：

2025 年 01 月 10 日

申请号：202210094546.2 **发文序号：**2025011001865340

申请人：上海金融期货信息技术有限公司

发明创造名称：一种基于流式处理的舆情实时监测系统

第 一 次 审 查 意 见 通 知 书

1.☒应申请人提出的实质审查请求，根据专利法第 35 条第 1 款的规定，国家知识产权局对上述发明专利申请进行实质审查。

☐根据专利法第 35 条第 2 款的规定，国家知识产权局决定自行对上述发明专利申请进行审查。

2.☐申请人要求以其在：

☐申请人已经提交了经原受理机构证明的第一次提出的在先申请文件的副本。

☐申请人尚未提交经原受理机构证明的第一次提出的在先申请文件的副本，根据专利法第 30 条的规定视为未要求优先权要求。

3.☐经审查，申请人于_____提交的修改文件，不符合专利法实施细则第 57 条第 1 款的规定，不予接受。

4.审查针对的申请文件：

☒原始申请文件。☐分案申请递交日提交的文件。☐下列申请文件：

5.☐本通知书是在未进行检索的情况下作出的。

☒本通知书是在进行了检索的情况下作出的。

☒本通知书引用下列对比文件(其编号在今后的审查过程中继续沿用)：

编号	文 件 号 或 名 称	公开日期（或抵触申请的申请日）
1	基于流式计算的网络舆情分析模型研究，《情报学报》，高欢，第 35 卷第 7 期	2016-07-24
2	CN112765294A	2021-05-07
3	CN111368165A	2020-07-03

6.审查的结论性意见：

关于说明书：

☐申请的内容属于专利法第 5 条规定的不授予专利权的范围。

☐说明书不符合专利法第 26 条第 3 款的规定。

纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路 6 号 国家知识产权局专利局受理处收

电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

☐说明书不符合专利法第 33 条的规定。
☐说明书的撰写不符合专利法实施细则第 20 条的规定。
☐_____

关于权利要求书：

☐权利要求_____不符合专利法第 2 条第 2 款的规定。
☐权利要求_____不符合专利法第 9 条第 1 款的规定。
☐权利要求_____不具备专利法第 22 条第 2 款规定的新颖性。
☒权利要求 1-9 不具备专利法第 22 条第 3 款规定的创造性。
☐权利要求_____不具备专利法第 22 条第 4 款规定的实用性。
☐权利要求_____属于专利法第 25 条规定的不授予专利权的范围。
☐权利要求_____不符合专利法第 26 条第 4 款的规定。
☐权利要求_____不符合专利法第 31 条第 1 款的规定。
☐权利要求_____不符合专利法第 33 条的规定。
☐权利要求_____不符合专利法实施细则第 22 条的规定。
☐权利要求_____不符合专利法实施细则第 23 条的规定。
☐权利要求_____不符合专利法实施细则第 24 条的规定。
☐权利要求_____不符合专利法实施细则第 25 条的规定。
☐_____

☐申请不符合专利法第 26 条第 5 款或者实施细则第 29 条的规定。
☐申请不符合专利法第 19 条第 1 款的规定。
☐申请不符合专利法实施细则第 11 条的规定。
☐分案申请不符合专利法实施细则第 49 条第 1 款的规定。

上述结论性意见的具体分析见本通知书的正文部分。

7.基于上述结论性意见，审查员认为：

☐申请人应当按照通知书正文部分提出的要求，对申请文件进行修改。
☐申请人应当在意见陈述书中论述其专利申请可以被授予专利权的理由，并对通知书正文部分中指出的不符合规定之处进行修改，否则将不能授予专利权。
☒专利申请中没有可以被授予专利权的实质性内容，如果申请人没有陈述理由或者陈述理由不充分，其申请将被驳回。
☐_____

8.申请人应注意下列事项：

(1) 根据专利法第 37 条的规定，申请人应在收到本通知书之日起的 4 个月内陈述意见，如果申请人无正当理由逾期不答复，其申请被视为撤回。

(2) 申请人对其申请的修改应当符合专利法第 33 条的规定，不得超出原说明书和权利要求书记载的范围，同时申请人对专利申请文件进行的修改应当符合专利法实施细则第 57 条第 3 款的规定，按照本通知书的要求进行修改。

(3) 申请人的意见陈述书和/或修改文本应邮寄或递交国家知识产权局专利局受理处，凡未邮寄或递交给受理处的文件不具备法律效力。

(4) 未经预约，申请人和/或代理师不得前来国家知识产权局专利局与审查员举行会晤。

(5) 对进入实质审查阶段的发明专利申请，在第一次审查意见通知书答复期限届满前（已提交答复意见的除外），主动申请撤回的，可以请求退还 50%的专利申请实质审查费。

9.本通知书正文部分共有 9 页，并附有下述附件：

☒引用的对比文件的复印件共 1 份 7 页。
☐_____

审查员：乔晋　　联系电话：028-62967682　　审查部门：专利审查协作四川中心

国家知识产权局 专利审查业务章

纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路 6 号　国家知识产权局专利局受理处收。
电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

第 一 次 审 查 意 见 通 知 书

申请号:2022100945462

本发明涉及一种基于流式处理的舆情实时监测系统，经审查，提出如下审查意见：

1.权利要求 1 请求保护一种基于流式处理的舆情实时监测系统，对比文件 1(基于流式计算的网络舆情分析模型研究，《情报学报》，高欢，第 35 卷第 7 期，公开日 2016 年 7 月 24 日)公开了一种基于流式计算的网络舆情分析模型，并具体公开了以下内容(参见第 725-728 页，图 4)：

基于流式计算的网络舆情分析(相当于一种基于流式处理的舆情实时监测系统)主要由数据收集(相当于系统包括数据获取模块)、舆情分析(相当于数据实时计算模块)与舆情治理三部分构成，各功能模块如图 4 所示。

网络舆情的数据收集是网络舆情分析的第一步，为分析提供所需的数据。收集全面、真实、准确的舆情信息，是消除信息不对称和确保分析结果准确、客观的关键。数据收集过程应尽量多地扩大信息源，包括门户网站，贴吧、论坛、微博、微信等社交媒体，QQ、MSN 等即时通信软件， 新闻网站的报道及评论等。由于对时效性的追求，网络舆情分析需要进行实时数据收集(相当于数据获取模块，用于舆情数据的实时采集)。

网络舆情分析要变被动响应为主动分析，最好的方式是对收集的数据进行实时计算(相当于数据实时计算模块用于舆情数据的分布式实时计算)，而不是先存储再利用。由于数据源实时不间断导致数据量大且无法预算， 收集的数据先进入流计算平台进行处理的方式，优于利用分布式批处理技术进行数据分析。流处理无法实时计算的，又需存储以备利用的数据，存入混合型数据库（如 HDFS）(相当于数据存储模块)，混合型数据库中，系统可分为两层，下层使用分布式数据库管理系统进行任务的分解和调度，上层用关系型数据库管理系统进行数据的查询和处理(相当于数据存储模块，用于数据的持久化存储和查询)。

网络舆情治理的功能在于提高网络信息的真实性,维护网络秩序的正常化和网络行为的有序性，本模型中，通过流式计算对网络舆情进行实时分析，政府部门可以及时监测部分舆情事件

走势，预测舆情走向，有效化解潜在矛盾，占据舆情高地，引导舆论走向。由图 4 可知，模型中包含了舆情分析和舆情展示(相当于数据实时推送模块)的流程，其中，舆情展示可以将舆情信息推送给有关部门进行展示(相当于数据实时推送模块用于数据的实时接收及推送)。

权利要求 1 要求保护的技术方案与对比文件 1 相比，其区别技术特征为：系统还包括数据总线模块，其中：数据总线模块，用于为系统内的各模块之间的交互提供传输通道，实现数据的传输、数据的持久化以及数据的分发。

基于该区别技术特征，权利要求 1 实际解决的技术问题为：如何实现各功能模块的信息交互。

对于上述区别技术特征，对比文件 2(CN112765294A，公开日 2021 年 5 月 7 日)公开了一种气象大数据处理调度系统，并具体公开了以下特征(参见权利要求 1-3)：

一种气象大数据处理调度系统，其特征是：包括数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块；所述数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块之间通过消息总线通信；所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。

可见，对比文件 2 公开了消息总线用于各模块之间的数据转发。而且该技术特征在对比文件 2 中所起的作用与其在本申请中为解决技术问题所起的作用相同，都是用于实现模块间的信息交互，即对比文件 2 给出了将该技术特征用于对比文件 1 以解决技术问题的启示。在上述对比文件的基础上，本领域技术人员容易想到系统还包括数据总线模块，其中：数据总线模块，用于为系统内的各模块之间的交互提供传输通道，实现数据的传输、数据的持久化以及数据的分发，这属于本领域的惯用技术手段。

由此可见，在对比文件 1 的基础上结合对比文件 2 以及本领域的惯用技术手段得到权利要求 1 所要求保护的技术方案，对本领域技术人员来说是显而易见的，因此，权利要求 1 所要求保护的技术方案不具备突出的实质性特点和显著的进步，因而不具备专利法第二十二条第三款规定的创造性。

2.权利要求 2 对权利要求 1 做了进一步限定，对比文件 1 公开了以下内容(参见第 725-728 页)：

网络舆情的数据收集是网络舆情分析的第一步，为分析提供所需的数据。收集全面、真实、准确的舆情信息，是消除信息不对称和确保分析结果准确、客观的关键。数据收集过程应尽量多地扩大信息源，包括门户网站，贴吧、论坛、微博、微信等社交媒体，QQ、MSN 等即时通信软件， 新闻网站的报道及评论等。由于对时效性的追求，网络舆情分析需要进行实时数据收集。

可见，对比文件 1 已经公开了可以对多信息源进行实时信息的采集。在此基础上，本领域技术人员容易想到数据获取模块进一步配置为：利用分布式爬虫技术自动采集互联网文本信息，再利用 XML 路径语言方法对抓取到的互联网文本信息进行解析，从中获取所关注的信息，形成结构化的舆情数据，这属于本领域的惯用技术手段。

因此，当其引用的权利要求不具备创造性时，权利要求 2 也不具备专利法第二十二条第三款规定的创造性。

3.权利要求 3 对权利要求 2 做了进一步限定，对比文件 1 公开了以下内容(参见第 725-728 页)：

网络舆情的数据收集是网络舆情分析的第一步，为分析提供所需的数据。收集全面、真实、准确的舆情信息，是消除信息不对称和确保分析结果准确、客观的关键。数据收集过程应尽量多地扩大信息源，包括门户网站，贴吧、论坛、微博、微信等社交媒体，QQ、MSN 等即时通信软件， 新闻网站的报道及评论等。由于对时效性的追求，网络舆情分析需要进行实时数据收集。

可见，对比文件 1 已经公开了可以对多信息源进行实时信息的采集。在此基础上，本领域技术人员容易想到数据获取模块采用任务调度框架定时触发爬虫程序进行互联网文本信息的抓取，这属于本领域的惯用技术手段。

因此，当其引用的权利要求不具备创造性时，权利要求 3 也不具备专利法第二十二条第三款规定的创造性。

4.权利要求 4 对权利要求 1 做了进一步限定，对比文件 2 公开了以下特征(参见权利要求 1-3，图 3)：

所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述 AMQP 协议从流式通信管线中根据同一主题数据进行路由选择，路由选择方式包括数据优先级别，数据类型，过滤数据或者插值订正处理。由图 3 可知，其包含了流失处理器 Kafka（相当于分布式消息系统）。

可见，对比文件 2 公开了消息总线用于各模块之间的数据转发，其中，包含了流式处理器 Kafka。而且该技术特征在对比文件 2 中所起的作用与其在本申请中为解决技术问题所起的作用相同，都是用于实现模块间的信息交互，即对比文件 2 给出了将该技术特征用于对比文件 1 以解决技术问题的启示。在上述对比文件的基础上，本领域技术人员容易想到数据总线模块使用分布式消息系统进行数据的传输、分发和持久化，这属于本领域的惯用技术手段。

因此，当其引用的权利要求不具备创造性时，权利要求 4 也不具备专利法第二十二条第三款规定的创造性。

5.权利要求 5 对权利要求 4 做了进一步限定，对比文件 2 公开了以下特征(参见权利要求 1-3，图 1)：

一种气象大数据处理调度系统，其特征是：包括数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块；所述数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块之间通过消息总线通信；所述数据生产者模块用于将数据生产者产生的数据主题发布到数据集市模块；所述数据消费者模块用于将数据消费者从订阅的数据主题中提取数据；所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式

通信管线构成。所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述 AMQP 协议从流式通信管线中根据同一主题数据进行路由选择，路由选择方式包括数据优先级别，数据类型，过滤数据或者插值订正处理。由图 3 可知，其包含了流失处理器 Kafka。

可见，对比文件 2 公开了消息总线用于各模块之间的数据转发，所述数据生产者模块用于将数据生产者产生的数据主题发布到数据集市模块；所述数据消费者模块用于将数据消费者从订阅的数据主题中提取数据，由图 1 可知，该系统包含了数据仓库。而且该技术特征在对比文件 2 中所起的作用与其在本申请中为解决技术问题所起的作用相同，都是用于实现模块间的信息交互，即对比文件 2 给出了将该技术特征用于对比文件 1 以解决技术问题的启示。在上述对比文件的基础上，本领域技术人员容易想到数据总线模块进一步配置为：接收数据并将接收到的数据进行持久化的存储，利用分布式高可用技术实现数据的转发和存储，其中数据总线模块支持消息的发布和订阅功能，数据的生产者向数据总线模块发布数据，数据的消费者从数据总线模块订阅消费数据，这属于本领域的惯用技术手段。

因此，当其引用的权利要求不具备创造性时，权利要求 5 也不具备专利法第二十二条第三款规定的创造性。

6.权利要求 6 对权利要求 1 做了进一步限定，对比文件 1 公开了以下内容(参见第 725−728 页，图 4)：

网络舆情分析要变被动响应为主动分析，最好的方式是对收集的数据进行实时计算，而不是先存储再利用。由于数据源实时不间断导致数据量大且无法预算， 收集的数据先进入流计算平台进行处理的方式，优于利用分布式批处理技术进行数据分析。流处理无法实时计算的，又需存储以备利用的数据，存入混合型数据库（如 HDFS），混合型数据库中，系统可分为两层，下层使用分布式数据库管理系统进行任务的分解和调度，上层用关系型数据库管理系统进行数据的查询和处理。

可见，对比文件 1 已经公开了网络舆情分析要变被动响应为主动分析，最好的方式是对收集的数据进行实时计算。此外，对比文件 2 公开了以下特征(参见权利要求 1-3，图 1)：

一种气象大数据处理调度系统，其特征是: 包括数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块；所述数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块之间通过消息总线通信；所述数据生产者模块用于将数据生产者产生的数据主题发布到数据集市模块；所述数据消费者模块用于将数据消费者从订阅的数据主题中提取数据；所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述 AMQP 协议从流式通信管线中根据同一主题数据进行路由选择，路由选择方式包括数据优先级别，数据类型，过滤数据或者插值订正处理。由图 3 可知，其包含了流失处理器 Kafka。

可见，对比文件 2 公开了消息总线用于各模块之间的数据转发，所述数据生产者模块用于将数据生产者产生的数据主题发布到数据集市模块；所述数据消费者模块用于将数据消费者从订阅的数据主题中提取数据，由图 1 可知，该系统包含了数据仓库。而且该技术特征在对比文件 2 中所起的作用与其在本申请中为解决技术问题所起的作用相同，都是用于实现模块间的信息交互，即对比文件 2 给出了将该技术特征用于对比文件 1 以解决技术问题的启示。在上述对比文件的基础上，本领域技术人员容易想到数据实时计算模块采用分布式流式处理框架，数据实时计算模块进一步配置为：数据实时计算模块订阅数据总线模块的数据，数据实时计算模块采用数据驱动的方式将计算结果加以保存，在达到设定的逻辑条件后将计算结果发布到数据总线模块，这属于本领域的惯用技术手段。

因此，当其引用的权利要求不具备创造性时，权利要求 6 也不具备专利法第二十二条第三款规定的创造性。

7.权利要求 7 对权利要求 6 做了进一步限定，对比文件 3(CN111368165A，公开日 2020 年 7

月 3 日)公开了一种时空流数据集成平台，并具体公开了以下特征(参见权利要求 1-8)：

一种时空流数据集成平台，其特征在于，包括：数据传输模块，以分布式消息中间件为基础，建立控制和管理功能，供用户建立“数据源-连接器-消息中间件-连接器-数据目的地”之间的传输通道；用户管理模块，用于注册和审核用户、管理历史操作日志，以及为用户分配权限；任务管理模块，提供任务的新建、控制和通知功能；任务是指通过数据传输模块建立的传输通道执行和完成一次数据传输的过程；运维监控模块，通过可视化运维看板管控所有数据传输过程和进展情况，提供多样化的图表；实时计算模块，结合流式处理的分布式计算中间件，以 API 接口方式提供针对时空流数据的过滤、聚合、汇总实时计算；数据清洗模块，提供自动化的纠正和转换方法，处理存在的无效数据、重复数据、异常数据。

任务管理模块包括：任务与任务组单元，提供任务和任务组的新建和编辑功能；供用户新建和删除任务，编写任务描述信息，说明任务或者任务组的意图，授予用户组管理任务的权限，或将任务分配至任务组中；任务中心单元，呈现任务概览信息和运行状态，监控任务的执行情况，可视化管理数据传输过程，查看与任务相关的操作通知信息；任务控制单元，用于进行，任务取消：终止正在运行的任务并立即使任务失败；任务暂停：阻止新任务或者连接器的运行，目前正在运行的任务、连接器照常进行；任务恢复：恢复暂停执行；任务重试：当任务仍处于活动状态时，重试将重新启动所有失败作业；准备执行：设置任务运行的时间；事件通知单元，在任务出现变更或执行过程中出现重大情况时，通知与任务相关的用户；每个任务均可设置如下通知项：任务变更、任务开始、任务失败、任务完成。

运维监控模块，包括：集群概况监控单元：进行包括主题数量、broker 节点数量和地址、监控分区数量、leader 数量监控项目；节点负载监控单元：进行包括数据流入/流出速度、CPU 占用率、网络流入/流出速度、磁盘读写速度、IO 等待情况、磁盘占有率、内存使用率监控项目；数据传输监控单元：进行包括数据流入/流出总数、数据流入/流出速度、平均速度、1 分钟或多分钟均速监控项目；错误事件管理单元：实时收集所有的数据相关错误，以错误概览和详情方

式呈现。

可见，对比文件 3 公开了平台中包含了任务管理模块，提供任务的新建、控制和通知功能；运维监控模块，通过可视化运维看板管控所有数据传输过程和进展情况，提供多样化的图表，包含了集群概况监控单元；实时计算模块，结合流式处理的分布式计算中间件，以 API 接口方式提供针对时空流数据的过滤、聚合、汇总实时计算。而且该技术特征在对比文件 3 中所起的作用与其在本申请中为解决技术问题所起的作用相同，都是用于实现数据流式实时计算过程中的管理，即对比文件 2 给出了将该技术特征用于对比文件 1 以解决技术问题的启示。在上述对比文件的基础上，本领域技术人员容易想到数据实时计算模块进一步包括集群管理单元、分布式存储单元和数据计算通路，其中集群管理单元用于统一调度和管理，集群管理单元具有包括任务提交、任务调度、错误恢复、资源分配以及权限控制在内的功能，分布式存储单元用于对数据实时计算模块中产生的中间结果进行持久化存储，在计算节点发生故障时通过分布式存储单元对计算结果进行恢复，数据计算通路具有标准化的处理架构，包括：数据源单元、数据预处理单元、多个数据计算单元、数据汇集单元，其中数据源单元用于从包括数据总线模块在内的数据源获取数据，数据预处理单元用于数据的预加工、预处理、数据拼接，数据计算单元用于数据的计算、聚合、排序等，数据汇集单元用于从各个数据计算单元汇聚计算结果并进行全局加工，这属于本领域的惯用技术手段。

由此可见，在对比文件 1 的基础上结合对比文件 2-3 以及本领域的惯用技术手段得到权利要求 7 所要求保护的技术方案，对本领域技术人员来说是显而易见的，因此，当其引用的权利要求不具备创造性时，权利要求 7 也不具备专利法第二十二条第三款规定的创造性。

8.权利要求 8 对权利要求 1 做了进一步限定，对比文件 1 公开了以下内容(参见第 725-728 页，图 4)：

网络舆情分析要变被动响应为主动分析，最好的方式是对收集的数据进行实时计算，而不是先存储再利用。由于数据源实时不间断导致数据量大且无法预算， 收集的数据先进入流计算平

台进行处理的方式，优于利用分布式批处理技术进行数据分析。流处理无法实时计算的，又需存储以备利用的数据，存入混合型数据库（如 HDFS），混合型数据库中，系统可分为两层，下层使用分布式数据库管理系统进行任务的分解和调度，上层用关系型数据库管理系统进行数据的查询和处理。

可见，对比文件 1 已经公开了混合型数据库中系统可分为两层，下层使用分布式数据库管理系统进行任务的分解和调度，上层用关系型数据库管理系统进行数据的查询和处理。此外，对比文件 2 公开了以下特征(参见权利要求 1-3，图 1)：

一种气象大数据处理调度系统，其特征是：包括数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块；所述数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块之间通过消息总线通信；所述数据生产者模块用于将数据生产者产生的数据主题发布到数据集市模块；所述数据消费者模块用于将数据消费者从订阅的数据主题中提取数据；所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述 AMQP 协议从流式通信管线中根据同一主题数据进行路由选择，路由选择方式包括数据优先级别，数据类型，过滤数据或者插值订正处理。由图 3 可知，其包含了流失处理器 Kafka。

可见，对比文件 2 公开了消息总线用于各模块之间的数据转发，所述数据生产者模块用于将数据生产者产生的数据主题发布到数据集市模块；所述数据消费者模块用于将数据消费者从订阅的数据主题中提取数据，由图 1 可知，该系统包含了数据仓库。而且该技术特征在对比文件 2 中所起的作用与其在本申请中为解决技术问题所起的作用相同，都是用于实现模块间的信息交互，即对比文件 2 给出了将该技术特征用于对比文件 1 以解决技术问题的启示。在上述对比文件的基础上，本领域技术人员容易想到数据存储模块采用分布式多用户能力的全文搜索引擎引擎，数据存储模块进一步配置为：数据存储模块订阅消费数据总线模块中的原始数据以及经数

据实时计算模块计算后的数据，将订阅到的这些数据按照预先定义的模式存储在磁盘上，在磁盘上的这种预先定义的存储模式具有结构性，用于对存入的舆情数据进行倒排索引，以便后续对舆情数据的分词检索和单点查询，这属于本领域的惯用技术手段。

因此，当其引用的权利要求不具备创造性时，权利要求 8 也不具备专利法第二十二条第三款规定的创造性。

9.权利要求 9 对权利要求 1 做了进一步限定，对比文件 1 公开了以下内容(参见第 725-728 页)：

网络舆情治理的功能在于提高网络信息的真实性，维护网络秩序的正常化和网络行为的有序性，本模型中，通过流式计算对网络舆情进行实时分析，政府部门可以及时监测部分舆情事件走势，预测舆情走向，有效化解潜在矛盾，占据舆情高地，引导舆论走向。由图 4 可知，模型中包含了舆情分析和舆情展示(相当于数据实时推送模块)的流程，其中，舆情展示可以将舆情信息推送给有关部门进行展示。

可见，对比文件 1 已经公开了可以对多信息源进行实时信息的采集，舆情展示可以将舆情信息推送给有关部门进行展示。

此外，对比文件 2 公开了以下特征(参见权利要求 1-3，图 1)：

一种气象大数据处理调度系统，其特征是：包括数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块；所述数据生产者模块，数据消费者模块，数据集市模块，数据调度模块以及 GIS 检索模块之间通过消息总线通信；所述数据生产者模块用于将数据生产者产生的数据主题发布到数据集市模块；所述数据消费者模块用于将数据消费者从订阅的数据主题中提取数据；所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述消息总线用于各模块之间的数据转发，消息总线内部有多个流式通信管线构成。所述 AMQP 协议从流式通信管线中根据同一主题数据进行路由选择，路由选择方式包括数据优先级别，数据类型，过滤数据或者插值订正处理。由图 3 可知，其包含了流失处理器

Kafka。

可见，对比文件 2 公开了消息总线用于各模块之间的数据转发，所述数据生产者模块用于将数据生产者产生的数据主题发布到数据集市模块；所述数据消费者模块用于将数据消费者从订阅的数据主题中提取数据，由图 1 可知，该系统包含了数据仓库。而且该技术特征在对比文件 2 中所起的作用与其在本申请中为解决技术问题所起的作用相同，都是用于实现模块间的信息交互，即对比文件 2 给出了将该技术特征用于对比文件 1 以解决技术问题的启示。在此基础上，本领域技术人员容易想到数据实时推送模块采用分层的一站式轻量级开源框架，数据实时推送模块进一步配置为：数据实时推送模块订阅数据总线模块的原始数据以及经过数据实时计算模块计算后的数据，经过业务逻辑的处理后，利用 websocket 协议将将数据推送给前端进行业务展示，这属于本领域的惯用技术手段。

因此，当其引用的权利要求不具备创造性时，权利要求 9 也不具备专利法第二十二条第三款规定的创造性。

基于上述理由，本申请的独立权利要求以及从属权利要求都不具备创造性，如果申请人不能在本通知书规定的答复期限内提出表明本申请具有创造性的充分理由，本申请将被驳回。

审查员姓名:乔晋
审查员代码:30140602

情报学报 ISSN 1000-0135
第35卷第7期723-729，2016年7月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 7, 723-729 July 2016

doi:10.3772/j.issn.1000-0135.2016.007.006

基于流式计算的网络舆情分析模型研究[1)]

高 欢

（中国人民大学信息资源管理学院，北京 100872）

摘要 互联网时代，网络舆情的庞大数据规模和舆情分析的计算复杂性，使对网络舆情的分析和实时掌控变得愈发困难。面向快速、不断产生的网络舆情采用流式计算进行实时处理的分析模型，在时效性、突发性和无限性三个方面都更加符合网络舆情的自身特性。基于流式计算的网络舆情分析模型分为数据收集、舆情分析和舆情治理三个部分，通过对语义保障和负载控制等关键技术的把控，可以实现个案把握向整体掌控、被动响应向主动分析的转变。基于流式计算的网络舆情分析模型具有可扩展性，能够联合众多服务器及资源，具有平台优势，能够解决地方舆情分析中面临的技术门槛，保障网络舆情分析的准确性与及时性。

关键词 智能信息分析 流式计算 云计算 网络舆情

Research on Model of Network Public Opinion Analysis based on Stream Computing

Gao Huan

(*Information Resource Management College ofRenmin University, Beijing* 100872)

Abstract During the Internet age, the network public opinion analysis and real-time control are becoming more difficult since the large data scale of network public opinion and the computational complexity of public opinion analysis. For rapid, stream data continuously generated in real-time processing, the analysis model has three aspects of advantages including timeliness, sudden and unlimited which is more in line with its own characteristics. The model of network public opinion analysis based on stream computing can be divided into three parts: data collection, analysis of public opinion and public opinion management, through the key technologies such as semantic security and load control, this model has also realized transformation from the case to overall control and the passive response to proactive analysis. The model of network public opinion analysis based on stream computing is scalable, can be combined with many servers and resources, with the advantages of the platform, it is possible to solve the technical barriers faced by local public opinion analysis to ensure network public opinion analysis accurate and timeliness.

Keywords intelligent information analysis, stream computing, cloud computing, network public opinion

1 引 言

舆情作为中国社会政治思想的重要组成部分，是民众对社会管理者、企业、个人及其他各类组织，围绕社会事件的发生、发展和变化，表达的信念、态度、意见和情绪等表现的总和。随着互联网的普及和发展，人们的行为习惯发生了改变，虚拟和现实世

收稿日期：2015年10月24日

作者简介：高欢，男，1990年生，中国人民大学信息资源管理学院信息分析专业博士研究生，主要研究方向：信息分析，云计算，E-mail：gaohuanyxh@live.cn。

1） 本文是国家社科基金重大项目《云计算环境下的信息资源集成与服务研究》（项目编号：12&ZD220）的研究成果之一。

界进行交互，人类社会活动以数据形式被记录、存储和传播[1]。各种形式的社会化媒体，如微博、微信等的出现，使互联网成为公众获取信息、传递信息、交流思想、表达意见的重要平台，也成为了政府有关部门获取舆情、了解公众思想动态的重要渠道。

云计算、物联网、移动互连等信息技术的快速发展，促使网络舆情在量上急剧增加，产生、传播的速度也较以往更快，总量已经远远超过历史上的任何时期。在此环境下，舆情分析工作面临比以往更严峻的挑战，如何在复杂而多元化的信息中辨明真伪、把握方向和有效分析网络舆情，已成为政府维护公信力，提高执政能力所面临的重要问题之一。

社会舆情在数据体量、复杂性、产生和传播速度等方面发生了巨大变化，科学监测、分析并正确引导舆情，需要越来越高的数据计算和使用能力。在当今社会关系重构的社交媒体时代，要建构科学有效的社会舆情管理体系，必须正视舆论生态新变化，善用大数据技术预测和引导社会舆论。

网络舆情分析对数据处理过程的整体延迟要求非常苛刻，如果能够在秒级或更短的时间内得到结果，将有利于作出进一步反应。现有技术并不能很好地满足对海量高速数据进行实时处理和分析的需求，即使是HDFS/MapReduce这种近年来被业界广泛采用的海量数据处理架构，也并不适用于如此高速和复杂的实时数据处理和分析场景。HDFS/MapReduce主要是面向静态数据的批处理，使用外存作为中间结果的存储介质，巨大的I/O代价成为影响处理过程实时性的瓶颈。海量、高速数据的实时处理引发了越来越多的关注，一些新的技术已经萌芽。其中，一类面向快速、不断产生的数据进行处理并立即产生结果的流处理模式得到了迅猛发展。流式计算中，数据往往来自最近一个时间窗口，因此延迟较短，能够满足舆情信息分析中的实时性要求。

2 流式计算

流式计算与其他大数据解决方案的处理方式不同，它是一种内存计算。在数据的有效时间内获取其价值，是流式计算的首要目标。因此，当数据到来后将立即对其进行计算，而不是缓存等待后续全部数据到来再进行计算。以往的数据处理方式（图1），先将收集到的数据储存到数据库中，然后在收到请求后搜索这些数据。例如，分布式系统（其实质是一个批处理系统）中，数据被引入文件系统（如HDFS）并分发到各个节点进行处理；当处理完成时，结果数据返回到文件系统以供使用。这是高效的处理方式，能够反复、多方式地使用数据，分析其中信息，但这也容易造成时间的浪费。流式计算（图2）中，运算法则在接收流数据时就开始对其进行分析。流式计算支持创建拓扑结构来转换没有终点的数据流，不同于分布式处理系统，它们会持续处理到达的数据[2]。

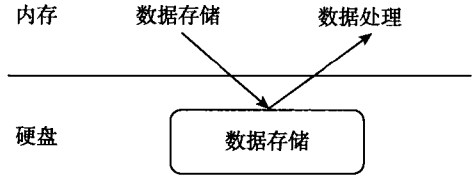

图1 批量计算

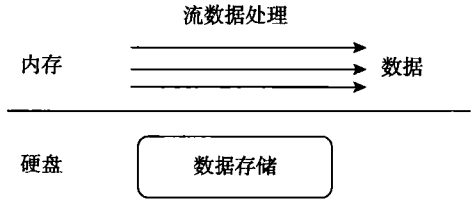

图2 流式计算

流式计算关注的是数据多次处理一次写入，而批量处理关注的是数据一次写入并多次处理使用。流处理系统运行起来后是持续不断的，而批处理系统往往只是在业务需要时调用数据。两者关注及应用的方向不一样，如表1所示。流式计算，是对数据的实时处理，是针对在线业务而存在的计算平台，适用于实时性要求高的情况。

表1 流式计算、批量计算部分性能对比

性能指标	计算方式	常驻空间	处理延时	有序性	数据速率	精确度	重现
流式计算	实时	内存	短	无	突发	较低	难
批量计算	批量	硬盘	长	有	稳定	较高	易

随着数据规模和复杂程度的增加，数据处理在吞吐量和响应时间上的要求越来越高。批处理系统（MapReduce）、大规模并行数据库、流处理系统和内存数据库，在数据吞吐量和响应时间上有所不同，如图3。内存数据库是一种基于磁盘静态数据的细粒度处理模式，在内存中重新设计了体系结构，实现了数据缓存、快速算法、并行操作等，数据处理速度快，适合吞吐量要求不高同时需要快速响应的应用。大

规模并行数据库通过并行使用多个 CPU 和磁盘，将诸如装载数据、建立索引、执行查询等操作并行化。同内存数据库相比，大规模并行数据库拥有更高的吞吐量，但数据处理延迟也会随之增加。随着数据量的不断增加，批处理系统将吞吐量在大规模并行数据库的基础上再次提升了一个数量级，处理速度虽有提升，但在响应时间上仍只适合于实时性要求不高的处理任务。流处理系统面向不断产生的动态数据并进行实时分析。相较于前三类处理模式，流处理系统在拥有高吞吐量的同时，具备了相对最低的处理延时[3]。

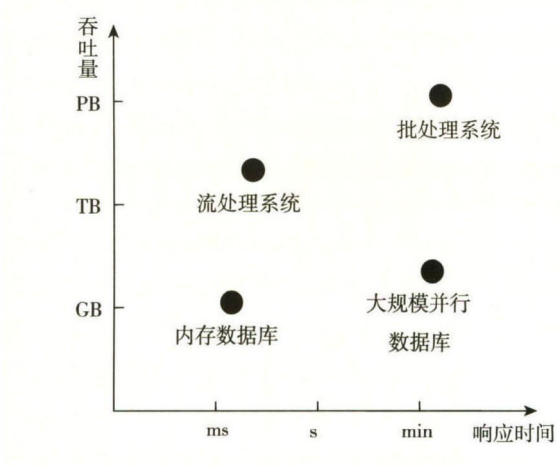

图 3　流处理系统和批处理系统对比

3　基于流式计算的网络舆情分析模型构建

3.1　必要性分析

网络舆情作为舆情的一种形式，具有传统媒介中舆情的特点，也有其自身的特性。互联网的发展，改变了社会的连接方式，降低了信息的交流成本，提升了传播速率，增加了数据总量。网络信息无限性和网民关注能力有限性之间的矛盾，加剧了网络舆论的选择性传播。社会化媒体为信息的传播、分享提供了平台，促进了信息开放，促使沟通更加便捷，凸显传播的个性化，使各类观点更容易找到扎根的土壤，从而相互支持、强化、放大，加剧舆论情绪[4]。网络舆情越来越难以掌控，大量相关性、偶发性因素使舆情复杂多变，传统的舆情监测研判手段和方法难以奏效，新的技术手段和方法要求更高。网络舆情分析，需要起到及时引导舆论的作用。

为此，网络舆情分析需要及时准确地发现舆情。准确取决于数据是否全面、算法模型是否合理，而及时则更多的取决于信息技术平台的处理速度。传统舆情分析技术更多关注于处理的准确性，对处理时间的要求并不高[5]。传统舆情分析的主要步骤是，先建立一套指标体系，再基于网络爬虫等手段获取数据，进行数据预处理，最后进行数据分析[6]。这一方法耗时较长，并不能在引导舆论时及时地发挥作用。

因此，研究基于流式计算的网络舆情分析构建十分必要，它能够从传感器、网络日志、网络点击流等设备实时采集下来的数据，连续注入到流计算平台，流计算平台部署相应业务规则，从而实现实时的业务分析与判断。

3.2　基于流式计算的网络舆情分析特点

(1)时效性

网络舆情实时产生，如若能够实时计算，就能在结果反馈中保证网络舆情分析的时效性。流式计算中，数据在到来后直接于内存中进行计算，其后将部分数据存储到硬盘中进行长久保存。流处理系统具有足够的低延迟计算能力，可以快速地进行数据计算，同时，对时效性强、潜在价值大的数据优先计算，保证在数据价值有效的时间内，体现数据的有用性，挖掘网络舆情的时效价值。

(2)突发性

网络舆情分析中，舆情数据的产生由数据源确定。由于不同数据源的状态不统一且在不同时空范围内发生动态变化，数据流的产生速率具有突发性。网络舆情的实时分析中，前一时刻数据速率和后一时刻数据速率可能会有巨大的差异，需要系统具有很好的可伸缩性。一方面，在突发高数据流速的情况下，保证不丢弃数据或识别并选择性地丢弃部分不重要的数据；另一方面，在低数据速率的情况下，保证不会太久或过多地占用系统资源。流处理系统能够动态适应突发性数据流，具有很强的系统计算能力和数据流匹配能力。

(3)无限性

舆情数据是实时产生、动态增加的，即潜在的数据量是无限的。在数据计算过程中，既没有足够大的空间来存储这些无限增长的数据，也没有合适的软件来有效地管理所有数据，因此，不会保存全部数据。流处理系统具有很好的稳定性，能够保证系统长期而稳定地分析并选取有价值的数据，不需要保存全部数据。

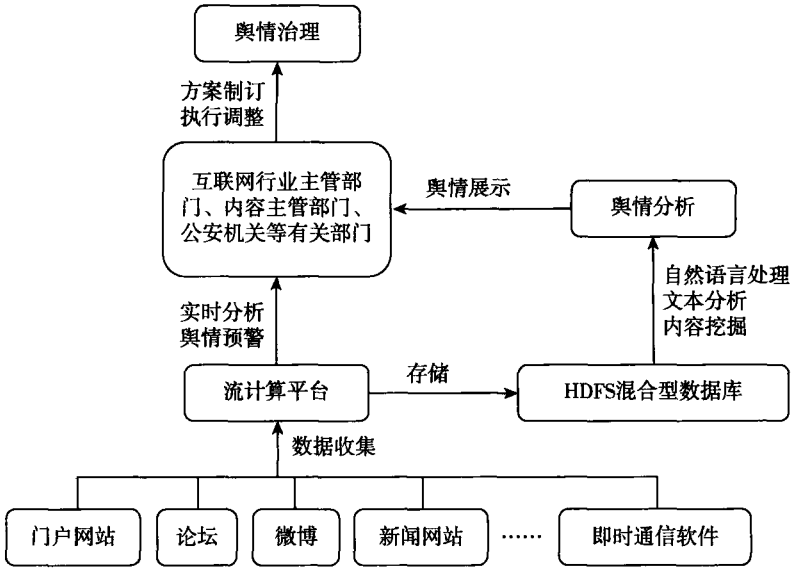

图4 基于流式计算的网络舆情分析模型

3.3 模型框架

基于流式计算的网络舆情分析主要由数据收集、舆情分析与舆情治理三部分构成,各功能模块如图4所示。

网络舆情的数据收集是网络舆情分析的第一步,为分析提供所需的数据。收集全面、真实、准确的舆情信息,是消除信息不对称和确保分析结果准确、客观的关键。数据收集过程应尽量多地扩大信息源,包括门户网站,贴吧、论坛、微博、微信等社交媒体,QQ、MSN等即时通信软件,新闻网站的报道及评论等。网络舆情分析的数据收集若要进一步实现全面、真实和准确,需要多元化的收集渠道,不局限于网民发布的信息,应加强对物联网中信息的收集,必要时可与历史数据对照,做到全面分析、多角度验证,保障数据的全面性和可信性。除数据的全面、真实、准确之外,由于对时效性的追求,网络舆情分析需要进行实时数据收集。数据收集中,数据源实时不间断地形成流式数据,如网站的访问PV/UV、用户访问内容、搜索内容等。

网络舆情分析要变被动响应为主动分析,最好的方式是对收集的数据进行实时计算,而不是先存储再利用。由于数据源实时不间断导致数据量大且无法预算,收集的数据先进入流计算平台进行处理的方式,优于利用分布式批处理技术进行数据分析。网络舆情数据可以分为三种类型:实时增量数据、全量历史数据和统计结果数据。实时增量数据源自于各种操作,如网民评论数据、视频数据、网络日志数据等,具有数据量大、价值密度低、数据价值随时间流逝而降低的特性;历史全量数据则为数据字典等,具有数据量小,修改不多的特征;统计结果数据为上述数据的复杂计算结果,如舆情热点小时统计结果,舆情增长异常分析结果等,该类数据由于大多产生于聚合计算,所以数据量一般不大,且周期产生。上述三类数据中,实时增量数据很少需要精细查询,但统计结果数据需要极快的响应速度,同时需要能够被并发查询。流处理平台的实时计算,较之以往的数据分析,能够更好地处理统计结果数据,针对海量数据进行,一般响应时间为秒级,能够更好地提供舆情预警。过往的数据离线分析是计算前一天来自每个省份不同性别、不同年龄、不同职业、不同名族的访问量分布,分析时距数据的产生已有一定时间。实时统计可以动态实时地刷新用户访问数据,展示网络流量的实时变化情况,分析每天各小时的流量和用户分布情况。

由于无法存储产生的所有数据,流处理系统不再进行流式数据的存储,当数据到来后在内存中直接进行实时计算。流处理无法实时计算的,又需要存储以备利用的数据,存入混合型数据库(如HDFS)。混合型数据库中,系统可分为两层,下层使用分布式数据库管理系统进行任务的分解和调度,上层用关系型数据库管理系统进行数据的查询和处理。利用关系型数据库进行数据存储和查询处理,能够更好地解决查询分析的性能问题。分布式数据库的任务调度机制能够提高系统的扩展性和容错性,以解决大规模网络舆情数据分析的横向扩展问

题,有利于信息共享。同时,分布式数据库进行数据处理具有一个显著优势,它是将计算推向数据,而不是将数据推向计算。由于数据块存放在磁盘上,将数据块传输到内存的时间可能会大于数据块到达内存后所需的处理时间,不能充分利用内存的处理速度。因此,利用分布式数据库存储大规模数据,能够在数据处理阶段,将计算推向数据,提升系统效率[7]。

网络舆情治理的功能在于提高网络信息的真实性,维护网络秩序的正常化和网络行为的有序性,主要围绕与政府管理有关的网络热点事件展开。这些事件通常具有传播快、规模大、影响远的特征,容易出现网络群体性事件。本模型中,通过流式计算对网络舆情进行实时分析,政府部门可以及时监测部分舆情事件走势,预测舆情走向,有效化解潜在矛盾,占据舆情高地,引导舆论走向,减少其给人民和社会带来的负面影响。以往的网络舆情治理模式中,各地方多部门分别存储信息,各自挖掘舆情,既不经济也不环保[8]。舆情分析平台,具有不受时空限制的特点,为政府网络舆情协同治理提供了可能,通过平台即服务的形式有效地将各地各级政府的网络舆情治理服务整合到一起,实现实时动态的互联共享。平台可以根据功能设计有效地规划不同层级、不同地域、不同职能部门在应对舆情事件时所承担的职责,从而明确分工,统一调度人、财、物,有效地应对舆情事件,实现网络舆情的高效协同治理。

3.4 关键技术

3.4.1 语义保障

为应对网络舆情的突发性,保障分析的时效性,流处理系统需要对输入数据的处理过程从语义方面进行一定保障。按照每条记录被完全处理(一条记录及其产生的新记录被途经的全部计算单元所处理)的次数,可将语义分为四类:无保障、至多处理一次、至少处理一次以及精确处理一次。其中,无保障即无约束,对记录的语义处理过程不做次数要求,在此无需赘述。现对至多处理一次、至少处理一次以及精确处理一次介绍如下:

(1)至多处理一次,是指每条记录被完全处理一次,或不被完全处理(中途丢失),但系统不会出现重复处理,以保障运行效率。这一语义适合处理数据时,完整性要求不高的应用。例如舆情分析中对热点或趋势的分析,少量记录未被处理不会对分析结果产生太大影响。

(2)至少处理一次,是指系每条记录一定都会被完全处理,但不保障避免重复处理。这一语义适合于重复相同操作不会对系统产生影响的情况下使用,例如处理数据时以相同的 Key 和 Value 多次写入。该语义下,若无可重复获取数据的可靠数据源,系统通常需要对流入的数据进行持久化存储。可重复获取数据的可靠数据源,是指系统若检测到处理失败或超时,能够进行数据重发。相较于至多处理一次语义,至少处理一次其实是通过牺牲效率,以克服流式运算中“一经处理就很难再次获取”的问题[9]。

(3)精确处理一次较上述两种语义更为严格,是指每个记录都被完全处理且仅处理一次。该语义下,系统需要区分处理与未处理的数据,并进行标记。以流处理系统 Storm 的 Trident 为例,它所采取的方法是在保证至少处理一次的前提下,对数据流划分批次,并分别赋予唯一的 ID。处理完成后,Trident 将对应批次产生的状态更新和 ID 以一次原子操作写入存储介质中。若遇到超时或故障,为避免数据未被处理,可以进行数据重发。此时,Trident 将生成与之前完全相同的批次数据和 ID。为避免重复提交,可以根据检测 ID 决定是否接受重发的数据。典型应用是 Word Count(统计数据流中每个单词的出现次数),在应用场景中每条记录都需要被精确处理,重复或遗漏都将导致最终的结果错误。

3.4.2 负载控制

面对网络舆情数据的无限性和分析的时效性要求,负载控制主要是通过一些调度分配策略保障系统高效稳定运行。负载控制策略可以分为静态策略和动态策略。

(1)静态策略,是一种提前确定的策略,在系统运行过程中无法改变。以早期数据库的静态查询策略为例,策略可分为两步:第一步,根据历史数据和现有模型,生成一套静态查询计划;第二步,为第一步查询计划的执行阶段,这一阶段将查询计划按照某个特定的并发度数实例化为多个算子,每个算子实例执行全部数据的一部分[10]。在运行过程中实例数目和算子所对应的数据子集无法改变。流计算系统中采用的静态策略与前述针对数据库的静态策略类似,如 S4 平台在数据划分和路由选择方面,早期版本中通过 XML 文件进行固定配置,在新版本中通过指定并行度后将数据按照 Key 值自动划分,无论哪一版本都是策略确定后就无法在运行过程中

改变。

(2)动态策略,与静态策略相对,是指运行过程中可以改变的负载控制策略。不同时间,数据本身的偏向性和运行环境的变化都会对系统负载产生影响,静态负载控制策略难以保证总是高效,尤其在网络舆情分析中,数据的无限性质会导致处理任务被长期连续执行,因此需要一些动态策略。集中式系统中,可以通过运行时调度和动态查询优化的方式来动态调整负载。分布式系统方面,Shah 等对流处理过程中的动态策略进行了介绍,并将能够给予同类数据划分及路由位置保障的策略称为上下文敏感策略,反之则是随机策略[11]。由于涉及聚合等操作,分布式流处理系统都应支持上下文敏感策略。自适应方面,动态策略可分为自适应和非自适应的。前者允许主节点根据工作节点的 CPU、内存等资源使用情况自动对计算任务进行转移、拆分以及合并,如 MillWheel 平台就采用了这种策略。后者则允许系统在外界干预(如人为)下完成在线的负载均衡等工作,如 Storm 平台。

4 典型流处理系统(Storm)介绍

自 2010 年雅虎公司公开其通用分布式流处理平台 S4 起,许多用途相近又各具特色的平台相继被提出,包括无中心节点的对称式系统架构(如 S4、Puma 等系统)以及有中心节点的主从式架构(如 Storm 系统)。本文将选取由 BackType 公司研发,后由 Twitter 开源的 Storm 流处理平台进行介绍。Twitter 每天约 3.4 亿条的推文均用 Storm 进行分析处理。

Storm[12,13]是一个开源的分布式实时计算系统,实现了一种流式处理模型,支持水平扩展,具有高容错性,能够高效地处理大量数据流。其主要应用场景为实时分析、在线机器学习、持续计算、ETL、分布式 RPC 等。

Storm 主要有两类节点:主节点(Master)和工作节点(Worker)。主节点上通常运行着 Nimbus 后台程序。它是 Storm 框架的管理节点,负责分配工作任务给每一个工作节点并监控其运行状态。工作节点上会运行 Supervisor 程序作为具体的工作节点,负责监听 Nimbus 分配的任务,启动或停止执行任务的工作进程[14]。集群系统中,一个节点一般运行一个或多个工作进程,每一个工作进程执行一个任务的子集。一个任务往往由分布在不同工作节点上的多个工作进程共同执行。

平台采用弱中心化的结构,Nimbus 和 Supervisor 都是无状态的,且两个模块之间没有直接的数据交互,两者的所有状态都保存在 Zookeeper 中,主节点只负责通过 Zookeeper 向工作节点分配任务,不参与实际计算过程。同去中心化结构中每个节点都要掌握全局信息相比,这种模式大大降低了通信和同步代价,有利于在运行时进行任务调度。Nimbus 通过写入 Zookeeper 来发布指令,而 Supervisor 则通过读取 Zookeeper 节点信息来执行这些指令。同时,Supervisor 会定时发送信息到 Zookeeper,使得 Nimbus 可以监控整个 Storm 集群的状态。当有节点挂掉时,Nimbus 能够快速使之重启。这种工作方式使得整个 Storm 集群十分健壮,任何一台工作机器突然失效都不会影响到整个系统的正常运行,只需重启失效节点后再从 Zookeeper 上面重新获取状态信息即可。

Storm 框架中,主要有两种类计算过程,源头处理过程 Spout 和中间处理过程 Bolt,前者负责读数据,后者负责计算任务。作为 Storm 中的消息源,Spout 组件将不断地从数据源(如 Log File、No-SQL、Message Queue 等)读取数据,并进行异常检查、数据去重等操作,最后把完整的、正常的数据,经直接分组、随机分组、广播分组等分组方式分组后,发送到 Bolt 中。Bolt 接收数据后,将执行过滤、聚合、数据库查询等操作,同时可以根据情况选择储存数据或是把数据传给下一级 Bolt。这一处理过程中,后台的守护线程将负责跟踪任务到达 Bolt 后是否执行成功。通过上述处理流程,Storm 可以保障每一个数据流在任务拓扑中能够被完全执行。

5 总 结

第一,流处理系统的高吞吐量和低处理延时,提升了网络舆情分析的数据处理能力。网络舆情工作需要对大量冗繁的网络信息进行监控、处理,并非几台服务器或几名工作人员就能完成的。通过模型的分布性与可扩展性,众多服务器及资源实现联合,形成超强的运作能力,保障网络舆情信息的及时获取、加工及处理。

第二,基于流式运算的网络舆情分析,具有平台优势,能够解决地方舆情分析中面临的技术门槛,提高经济效益,实现资源节约型、环境友好型的网络舆情分析。各地开展网络舆情工作时,仍存在无法通

过有效的技术手段及时获取、充分利用与安全存储信息的情况;各部门或单位独立运作也增加了更高层机构区分、筛选和剔除信息的工作任务。平台服务中,各地无需反复建设,通过多地合作开发或者国家统一布局,在网络终端即可获取应用服务,提升资源利用率的同时降低了技术门槛。

第三,模型的高共享性提升了网络舆情应对效率,保证多方参与。舆情分析系统的互联互通,带来舆情信息和分析技术的合作共享,简化了多方协作、参与的方式,极大地提高了信息资源的利用率及响应速度,利于多部门或多层级机构出台和制定更具整合性、协调性、一体化和配套性的舆情治理方案,及时引导舆论,保障互联网中信息的真实性,为社会共治提供基础。

参 考 文 献

[1] 李纲,陈璟浩. 突发公共事件网络舆情研究综述[J]. 图书情报知识,2014,(2):111-119.

[2] 孙大为,张广艳,郑纬民. 大数据流式计算:关键技术及系统实例[J]. 软件学报,2014,(4):839-862.

[3] 崔星灿,禹晓辉,刘洋,等. 分布式流处理技术综述[J]. 计算机研究与发展,2015,(2):318-332.

[4] 潘芳,仲伟俊,胡彬,等. 突发事件网络舆情的管控机制及效率测评[J]. 情报杂志,2012,(5):40-45.

[5] 李金海,何有世,熊强. 基于大数据技术的网络舆情文本挖掘研究[J]. 情报杂志,2014,(10):1-6+13.

[6] 蔡立辉,杨欣翥. 大数据在社会舆情监测与决策制定中的应用研究[J]. 行政论坛,2015,(2):1-10.

[7] Das S,Sismanis Y,Beyer K S, et al. Ricardo: Integrating R and Hadoop [C]// Elmagarmid AK, Agrawal D, eds. Proc. Of the SIGMOD. Indiana: ACM Press, 2010: 87-998.

[8] 中共中央宣传部舆情信息局. 网络舆情信息工作理论与实务[M]. 北京:学习出版社,2009.

[9] Babcock B, Babu S, Datar M, et al. Models and issues in data stream systems [C]//Proc of the 21st ACM SIGACT-SIGMOD-SIGART Symp on Principles of Database Systems. New York:ACM,2002:1-16.

[10] Wei Hong,Stonebraker M. Optimization of parallel query execution plans in XPRS[C]//Proc of the 1stIntConf on Parallel and Distributed Information Systems. Berlin: Springer,1991:218-225.

[11] Shah M A,Hellerstein J M,Franklin C S, et al. Flus:An adaptive partitioning operator for continuous query systems [C]//Proc of the 19th IntConf on Data Engineering. Piscataway,NJ:IEEE,2003:25-36.

[12] The Apache Foundation. Storm official website[OL]. [2014-04-08]. http://storm-project. net

[13] Github Inc. Storm Wiki[OL]. [2013-12-07]. https://github. com/nathanmarz/storm/wiki

[14] Petko V. Integrating parallel application development with performance analysis in periscope[J]. IPDPS Workshops, 2010:1-8.

(责任编辑 魏瑞斌)

国 家 知 识 产 权 局

检 索 报 告

申请号：2022100945462	申请日：2022 年 01 月 26 日	首次检索
申请人：上海金融期货信息技术有限公	最早的优先权日：	
权利要求项数：9	说明书段数：45+4	
审查员确定的 IPC 分类号：G06F 16/953,G06F 16/958		
检索记录信息：CN112765294A: 87 CNABS, (分布式 and (流式 or 流处理) and 实时) and 总线 CN111368165A: CNTXT 语义检索 基于流式计算的网络舆情分析模型研究:cnki，流式，流处理，舆情		

相 关 专 利 文 献

类型	国别以及代码[11] 给出的文献号	代码[43]或[45] 给出的日期	IPC 分类号	相关的段落 和 / 或图号	涉及的权 利要求
Y	CN112765294A	2021−05−07	G06F16/29	权利要求 1−3	1−9
Y	CN111368165A	2020−07−03	G06F16/951	权利要求 1−8	7

相 关 非 专 利 文 献

类型	书名（包括版本号和卷号）	出版日期	作者姓名和出版者名称	相关页数	涉及的权利要求
类型	期刊或文摘名称 （包括卷号和期号）	发行日期	作者姓名和文章标题	相关页数	涉及的权利要求
Y	《情报学报》,第 35 卷,第 7 期	2016−07−24	高欢,基于流式计算的网络舆情分析模型研究	第 725−728 页，图 4	1−9

国 家 知 识 产 权 局

类型	网址	网络发布日或公开日	作者姓名和网页标题	相关部分	涉及的权利要求

表格填写说明事项：

1. 审查员实际检索领域的 IPC 分类号应当填写到大组和 / 或小组所在的分类位置。
2. 期刊或其它定期出版物的名称可以使用符合一般公认的国际惯例的缩写名称。
3. 相关文件的类型说明：

   X：单独影响权利要求的新颖性或创造性的文件；

   Y：与本检索报告中其他 Y 类文件组合后影响权利要求的创造性的文件；

   A：背景技术文件，即反映权利要求的部分技术特征或者有关的现有技术的文件；

   R：任何单位或个人在申请日向专利局提交的、属于同样的发明创造的专利或专利申请文件。

   P：中间文件，其公开日在申请的申请日与所要求的优先权日之间的文件，或者会导致需要核实该申请优先权的文件；

   E：单独影响权利要求新颖性的抵触申请文件；

   T：申请日或优先权日当天或之后公布的，可以对所要求保护发明的理论或原理提供清楚解释的文件，或者可显示出所要求保护发明的推理或事实不成立的文件；

   L：除 X、Y、A、R、P、E 和 T 类文件之外的文件。

审 查 员：乔晋　　审查部门：专利审查协作四川中心

2025 年 01 月 08 日