



(12) 发明专利

(10) 授权公告号 CN 102314460 B

(45) 授权公告日 2014. 05. 14

(21) 申请号 201010222602. 3

说明书第 29-31、89-102、128-131、257-259 段 .

(22) 申请日 2010. 07. 07

CN 1207186 A, 1999. 02. 03, 全文 .

(73) 专利权人 阿里巴巴集团控股有限公司

审查员 胡平

地址 英属开曼群岛大开曼岛资本大厦一座
四层 847 号邮箱

(72) 发明人 岑文初

(74) 专利代理机构 北京集佳知识产权代理有限
公司 11227

代理人 遂长明 王宝筠

(51) Int. Cl.

G06F 17/30 (2006. 01)

(56) 对比文件

US 2007/0118491 A1, 2007. 05. 24, 摘要、说
明书第 14-15、30-32 段 .

US 2008/0059392 A1, 2008. 03. 06, 摘要、说

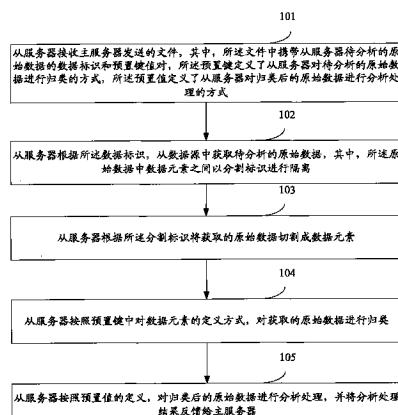
权利要求书3页 说明书14页 附图5页

(54) 发明名称

数据分析方法、系统及服务器

(57) 摘要

本申请实施例公开了一种数据分析方法、系
统及服务器。其中，所述方法包括：从服务器接收
主服务器发送的文件，其中，所述文件携带待分析
的原始数据的数据标识和预置键值对，所述预置
键定义了从服务器对待分析的原始数据进行归类
的方式；从服务器根据所述数据标识，从数据源
中获取待分析的原始数据；从服务器根据所述分
割标识将获取的原始数据切割成数据元素；从服
务器按照预置键中对数据元素的定义方式，对获
取的原始数据进行归类；从服务器按照预置值的
定义，对归类后的原始数据进行分析处理，并将分
析处理结果反馈给主服务器。根据本申请实施例，
可以实现对并行的数据处理架构中的海量数据进
行分析。



1. 一种数据分析方法,其特征在于,包括:

从服务器接收主服务器发送的文件,其中,所述文件携带待分析的原始数据的数据标识和预置键值对,所述预置键值对中的预置键定义了从服务器对待分析的原始数据中的数据元素进行归类的方式,所述预置键值对中的预置值定义了从服务器对归类后的原始数据中的数据元素进行分析处理的方式;

从服务器根据所述数据标识,从数据源中获取待分析的原始数据,其中,所述原始数据中数据元素之间以分割标识进行隔离;

从服务器根据所述分割标识将获取的原始数据切割成数据元素;

从服务器按照预置键中对数据元素的定义方式,对获取的原始数据中的数据元素进行归类;

从服务器按照预置值的定义,对归类后的原始数据中的数据元素进行分析处理,并将分析处理结果反馈给主服务器。

2. 根据权利要求1所述的数据分析方法,其特征在于,所述从服务器按照预置键中对数据元素的定义方式,对获取的原始数据进行归类之后,还包括:

从服务器从归类后的原始数据中筛选出符合第一预置过滤条件的原始数据;

则按照预置值的定义,对归类后的原始数据键进行分析处理为:按照预置值的定义,对筛选出的原始数据进行分析处理。

3. 根据权利要求1所述的数据分析方法,其特征在于,所述从服务器按照预置值的定义,对归类后的原始数据进行分析处理之后,还包括:

从服务器从分析处理得到的分析处理结果中筛选出符合第二预置过滤条件的分析处理结果;

则所述将分析处理结果反馈给主服务器为:将筛选出的分析处理结果反馈给主服务器。

4. 根据权利要求1-3中任意一项所述的数据分析方法,其特征在于,所述方法还包括:

当主服务器对接收到的分析处理结果进行合并处理后,将得到的合并处理结果与同一时间下的历史合并结果进行对比分析,根据对比分析的结果产生预警信号。

5. 一种数据分析方法,其特征在于,包括:

多线程中子线程接收主线程发送的文件,其中,所述文件携带待分析的原始数据的数据标识和预置键值对,所述预置键值对中的预置键定义了子线程对待分析的原始数据中的数据元素进行归类的方式,所述预置键值对中的预置值定义了子线程对归类后的原始数据中的数据元素进行分析处理的方式;

子线程根据所述数据标识,从数据源中获取待分析的原始数据,其中,所述原始数据中数据元素之间以分割标识进行隔离;

子线程根据所述分割标识将获取的原始数据切割成数据元素;

子线程按照预置键中对数据元素的定义方式,对获取的原始数据中的数据元素进行归类;

子线程按照预置值的定义,对归类后的原始数据中的数据元素进行分析处理,并将分析处理结果反馈给主线程。

6. 一种数据分析系统,其特征在于,包括:一主服务器和至少两个从服务器,其中,

所述主服务器，用于向从服务器发送文件，并对接收到的分析处理结果进行合并，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键值对中的预置键定义了从服务器对待分析的原始数据中的数据元素进行归类的方式，所述预置键值对中的预置值定义了从服务器对归类后的原始数据中的数据元素进行分析处理的方式；

所述从服务器，用于接收主服务器发送的文件，根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离，根据所述分割标识将获取的原始数据切割成数据元素，按照预置键中对数据元素的定义方式，对获取的原始数据中的数据元素进行归类，按照预置值的定义，对归类后的原始数据中的数据元素进行分析处理，并将分析处理结果反馈给主服务器。

7. 根据权利要求 6 所述的数据分析系统，其特征在于，当所述主服务器对接收到的分析处理结果进行合并后，所述主服务器还用于将得到的分析处理结果与同一时间下的历史合并结果进行对比分析，根据对比分析的结果产生预警信号。

8. 一种数据分析系统，其特征在于，包括：一主线程模块和至少两个子线程模块，其中，

所述主线程模块，用于向子线程模块发送文件，并对接收到的分析处理结果进行合并，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键值对中的预置键定义了子线程模块对待分析的原始数据中的数据元素进行归类的方式，所述预置键值对中的预置值定义了子线程模块对归类后的原始数据中的数据元素进行分析处理的方式；

所述子线程模块，用于接收主线程模块发送的文件，根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离，根据所述分割标识将获取的原始数据切割成数据元素，按照预置键中对数据元素的定义方式，对获取的原始数据中的数据元素进行归类，按照预置值的定义，对归类后的原始数据中的数据元素进行分析处理，并将分析处理结果反馈给主线程模块。

9. 一种数据分析装置，其特征在于，包括：

第一文件接收模块，用于接收主服务器发送的文件，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键值对中的预置键定义了从服务器对待分析的原始数据中的数据元素进行归类的方式，所述预置键值对中的预置值定义了从服务器对归类后的原始数据中的数据元素进行分析处理的方式；

第一数据获取模块，用于根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离；

第一数据切割模块，用于根据所述分割标识将获取的原始数据切割成数据元素；

第一数据归类模块，用于按照预置键中对数据元素的定义方式，对获取的原始数据中的数据元素进行归类；

第一数据计算模块，用于按照预置值的定义，对归类后的原始数据中的数据元素进行分析处理，并将分析处理结果反馈给主服务器。

10. 根据权利要求 9 所述的装置，其特征在于，还包括：

第一过滤模块，用于从归类后的原始数据中筛选出符合第一预置过滤条件的原始数据；

则所述第一数据计算模块按照预置值的定义,对筛选出的原始数据进行分析处理。

11. 根据权利要求 9 所述的装置,其特征在于,还包括:

第二过滤模块,用于从分析处理得到的分析处理结果中筛选出符合第二预置过滤条件的分析处理结果;则所述第一数据计算模块将筛选出的分析处理结果反馈给主服务器。

12. 一种数据分析装置,其特征在于,包括:

第二文件接收模块,用于接收主线程发送的文件,其中,所述文件携带待分析的原始数据的数据标识和预置键值对,所述预置键值对中的预置键定义了子线程对待分析的原始数据中的数据元素进行归类的方式,所述预置键值对中的预置值定义了子线程对归类后的原始数据中的数据元素进行分析处理的方式;

第二数据获取模块,用于根据所述数据标识,从数据源中获取待分析的原始数据,其中,所述原始数据中数据元素之间以分割标识进行隔离;

第二数据切割模块,用于根据所述分割标识将获取的原始数据切割成数据元素;

第二数据归类模块,用于按照预置键中对数据元素的定义方式,对获取的原始数据中的数据元素进行归类;

第二数据计算模块,用于按照预置值的定义,对归类后的原始数据中的数据元素进行分析处理,并将分析处理结果反馈给主线程。

数据分析方法、系统及服务器

技术领域

[0001] 本申请涉及通信和计算机技术领域,特别涉及一种数据分析方法、系统及服务器。

背景技术

[0002] 随着 web2.0 技术的发展,互联网应用或者互联网平台中的业务数据,如用户行为数据和平台系统数据,都呈现出海量增长的趋势。为了便于海量业务数据的处理,挖掘其内在价值,通常采用一种并行的数据处理架构来支撑海量数据的处理工作,即利用多个分布式的计算机相互协作工作,共同完成对海量数据的处理。

[0003] 当前,在大型的互联网网站平台中,应用最为广泛的一种并行的数据处理架构为 Hadoop 系统框架。在 Hadoop 的系统架构中包括有一个主服务器和多个从服务器组成的集群,主服务器将海量数据分割成多个数据块,再将分割后的数据块分配给多个并行的从服务器,由每个从服务器处理各自的数据块,并将处理的结果发送至主服务器,主服务器将处理的结果合并后输出。此外,当前阶段主服务器输出的合并结果又可以作为下一阶段主服务器进行数据处理的一个输入,得到下一阶段的合并结果。这种并行和串行相结合的处理方式可以使并行的数据处理系统高效地处理海量数据。

[0004] 目前,对于数据的分析方法主要为基于关系型数据库的数据分析方法,然而,这种方法很难基于并行的数据处理架构对关系型数据库的数据进行分析,特别是在需要进行归类、报表生成等复杂的数据分析处理工作时,难以满足实际需要。因此,基于关系型数据库的数据分析方法并不适用于对并行的数据处理架构中的海量数据进行分析。

发明内容

[0005] 为了解决上述技术问题,本申请实施例提供了一种数据分析方法、系统及服务器,以实现对并行的数据处理架构中的海量数据进行分析。

[0006] 本申请实施例公开公开了如下技术方案 :一种数据分析方法,包括 :

[0007] 从服务器接收主服务器发送的文件,其中,所述文件携带待分析的原始数据的数据标识和预置键值对,所述预置键定义了从服务器对待分析的原始数据进行归类的方式,所述预置值定义了从服务器对归类后的原始数据进行分析处理的方式;从服务器根据所述数据标识,从数据源中获取待分析的原始数据,其中,所述原始数据中数据元素之间以分割标识进行隔离;从服务器根据所述分割标识将获取的原始数据切割成数据元素;从服务器按照预置键中对数据元素的定义方式,对获取的原始数据进行归类;从服务器按照预置值的定义,对归类后的原始数据进行分析处理,并将分析处理结果反馈给主服务器。

[0008] 本申请还提供另一种数据分析方法,包括 :多线程中子线程接收主线程发送的文件,其中,所述文件携带待分析的原始数据的数据标识和预置键值对,所述预置键定义了子线程对待分析的原始数据进行归类的方式,所述预置值定义了子线程对归类后的原始数据进行分析处理的方式;子线程根据所述数据标识,从数据源中获取待分析的原始数据,其中,所述原始数据中数据元素之间以分割标识进行隔离;子线程根据所述分割标识将获取

的原始数据切割成数据元素；子线程按照预置键中对数据元素的定义方式，对获取的原始数据进行归类；子线程按照预置值的定义，对归类后的原始数据进行分析处理，并将分析处理结果反馈给主线程。

[0009] 本申请还提供一种数据分析系统，包括：一主服务器和至少两个从服务器，其中，所述主服务器，用于向从服务器发送文件，并对接收到的分析处理结果进行合并，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键定义了从服务器对待分析的原始数据进行归类的方式，所述预置值定义了从服务器对归类后的原始数据进行分析处理的方式；所述从服务器，用于接收主服务器发送的文件，根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离，根据所述分割标识将获取的原始数据切割成数据元素，按照预置键中对数据元素的定义方式，对获取的原始数据进行归类，按照预置值的定义，对归类后的原始数据进行分析处理，并将分析处理结果反馈给主服务器。

[0010] 本申请还提供另一数据分析系统，包括：一主线程模块和至少两个子线程模块，其中，所述主线程模块，用于向子线程模块发送文件，并对接收到的分析处理结果进行合并，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键定义了子线程模块对待分析的原始数据进行归类的方式，所述预置值定义了子线程模块对归类后的原始数据进行分析处理的方式；所述子线程模块，用于接收主线程模块发送的文件，根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离，根据所述分割标识将获取的原始数据切割成数据元素，按照预置键中对数据元素的定义方式，对获取的原始数据进行归类，按照预置值的定义，对归类后的原始数据进行分析处理，并将分析处理结果反馈给主线程模块。

[0011] 本申请还提供一种从服务器，包括：第一文件接收模块，用于接收主服务器发送的文件，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键定义了从服务器对待分析的原始数据进行归类的方式，所述预置值定义了从服务器对归类后的原始数据进行分析处理的方式；第一数据获取模块，用于根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离；第一数据切割模块，用于根据所述分割标识将获取的原始数据切割成数据元素；第一数据归类模块，用于按照预置键中对数据元素的定义方式，对获取的原始数据进行归类；第一数据计算模块，用于按照预置值的定义，对归类后的原始数据进行分析处理，并将分析处理结果反馈给主服务器。

[0012] 本申请还提供另一种服务器，包括：第二文件接收模块，用于接收主线程发送的文件，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键定义了子线程对待分析的原始数据进行归类的方式，所述预置值定义了子线程对归类后的原始数据进行分析处理的方式；第二数据获取模块，用于根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离；第二数据切割模块，用于根据所述分割标识将获取的原始数据切割成数据元素；第二数据归类模块，用于按照预置键中对数据元素的定义方式，对获取的原始数据进行归类；第二数据计算模块，用于按照预置值的定义，对归类后的原始数据进行分析处理，并将分析处理结果反馈给主线程。

[0013] 由上述实施例可以看出，首先从数据源中获取待分析的原始数据，然后按照分割

标识将原始数据切割成数据元素，并将切割得到的数据元素作为键值对中的键，再从切割得到的数据元素中，提取出符合预置键值对中的键定义的数据元素，最后按照预置键值对中的值定义，对提取出的数据元素进行分析处理，并将分析处理结果反馈给主服务器，以便主服务器对接收到的分析处理结果进行合并。因此，为并行的数据处理架构中的海量数据进行分析提供了具体的实现方案。

附图说明

[0014] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

- [0015] 图 1 为本申请一种数据分析方法的一个实施例的流程图；
- [0016] 图 2 为本申请一种数据分析方法的另一个实施例的流程图；
- [0017] 图 3 为本申请一种从服务器的一个实施例的结构图；
- [0018] 图 4 为本申请一种从服务器的另一个实施例的结构图；
- [0019] 图 5 为本申请一种从服务器的另一个实施例的结构图；
- [0020] 图 6 为本申请一种服务器的一个实施例的结构图；
- [0021] 图 7 为本申请一种数据分析系统的一个实施例的结构图；
- [0022] 图 8 为本申请一种数据分析系统的另一个实施例的结构图。

具体实施方式

[0023] 为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图对本申请实施例进行详细描述。

[0024] 本申请实施例中的数据分析方法可以对任何并行的数据处理架构中的海量数据进行分析，例如，Hadoop 系统框架中的海量数据。本申请实施例对并行的数据处理架构并不进行限定。

- [0025] 实施例一

[0026] 请参阅图 1，其为本申请一种数据分析方法的一个实施例的流程图，其应用于包括一个主服务器和多个从服务器组成的集群系统中，该方法包括以下步骤：

[0027] 步骤 101：从服务器接收主服务器发送的文件，其中，所述文件中携带从服务器待分析的原始数据的数据标识和预置键值对，所述预置键定义了从服务器对待分析的原始数据进行归类的方式，所述预置值定义了从服务器对归类后的原始数据进行分析处理的方式；

[0028] 例如，在一个并行的数据处理架构中，主服务器向各个从服务器发送一个文件，在文件中携带有数据标识和预置键值对，其中的预置键值对可以有多个。其中，所述数据标识指示了对应的从服务器需要获取的待分析的原始数据，例如，数据的地址信息等可以作为数据标识，指示对应的从服务器待分析的原始数据。所述预置键值对包括预置键和预置值，预置键定义了从服务器对待分析的原始数据进行归类的方式；预置值定义了从服务器对归类后的原始数据进行分析处理的方式。例如，假设一预置键值对中，预置键为 :key=“1,2,

3”，预置值为 :value=max (\$a\$+\$b\$+\$c\$)。则该预置键值对具体定义了从服务器需要对待分析的原始数据按照第 1 至 3 列数据元素进行归类，并按照预置值的定义，对归类后的原始数据中第 a 列、第 b 列和第 c 列的数据元素的分析处理方法为求和后再取最大值。

[0029] 需要说明的是，预置值定义的分析处理方法可以包括但不限于：统计最小值 (min)、统计最大值 (max)、计算平均值 (average)、计数 (count)、求和 (sum) 及直接显示 (plain) 等，直接显示 (plain) 一般用于主键列的显示。当然，分析处理过程也可以包括其他的计算方法，本申请实施例对分析处理的方式并不进行限定。

[0030] 步骤 102：从服务器根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离；

[0031] 例如，各个从服务器根据自身接收的数据标识，从数据源处获取数据标识所指示的待分析的原始数据，如，数据源可以是 FTP 服务器、数据库 (DB) 或文件系统，具体数据的格式可以是数据表、记录、日志等。并且，在本申请的原始数据中，各个数据元素之间以分割标识进行隔离。该分割标识可以是逗号、分号、空格、冒号等。本申请实施例中以逗号作为数据元素之间的分割标识进行举例说明。

[0032] 例如，以下为一段从数据源处获取原始数据，该原始数据为日志片段。在这个日志片段中，每个数据元素之间以逗号隔离。并且，在这个日志片段中，共有两段记录，每个记录以回车换行符作为记录的结束。

[0033] 0,203.171.227.117,null,xml,12005554,taobao.user.get,0,0,0,

172.24.14.65,小郭 cc,1.0,null,3,null,0,1274803197776,0,0,0,1,0,0,0,0,8,0,0,0,9

[0034] 0,97.74.215.111,null,xml,12028711,taobao.taobaoke.items.detail.get,0,

0,null,172.24.14.65,null,2.0,md5,4,null,221000,1274803197765,0,2,-1,1,0,0,0,

0,23,0,0,0,26

[0035] 步骤 103：从服务器根据所述分割标识将获取的原始数据切割成数据元素；

[0036] 例如，如果各个数据元素之间以逗号隔离，则可以按照逗号将第一条记录切割成以下共 30 个数据元素：第一个数据元素为 0，第二个数据元素为 203.171.227.117，第三个数据元素为 null，第四个数据元素为 xml，第五个数据元素为 12005554，第六个数据元素为 taobao.user.get，……，第 30 个数据元素为 9。

[0037] 同样，可以按照上述方式将第二条记录切割成 30 个数据元素。

[0038] 步骤 104：从服务器按照预置键中对数据元素的定义方式，对获取的原始数据进行归类；

[0039] 仍以上述第一条记录为例，如果在预置键值对中定义的预置键为 :key=“1,2,3”，则从切割得到的数据元素中提取出的符合预置键值对中预置对应的数据元素，即原始数据中的第 1 至 3 列数据元素 0、203.171.227.117 和 null。

[0040] 例如，仍旧以预置键为 :key=“1,2,3”，预置值为 :value=max (\$a\$+\$b\$+\$c\$) 为例来说明上述预置键值对的作用。如，对于一个从服务器上的 10 个待分析的原始数据而言，当通过预置键的归类后，发现在 10 个待分析的原始数据中，有 7 个原始数据中的第 1 至 3 列数据元素相同，另外 3 个原始数据中的第 1 至 3 列的数据元素相同，则分别对 7 个原始数据中第 a 列、第 b 列和第 c 列的数据元素求和，再取最大值，并且分别对另外 3 个原始数据中第 a 列、第 b 列和第 c 列的数据元素求和，再取最大值。

[0041] 但是,当记录中的数据元素较多时,或者数据元素在原始数据中的顺序发生变化时,容易发生数据元素操作错误的现象。例如,当数据元素在原始数据中的顺序发生变化,相应地,该数据元素对应的预置键的定义也会发生变化。如,假设数据元素 0 对应的预置键的定义为 :key= “1”,当其在原始数据中的顺序向右移动一位后,其对应的预置键的定义变为 key= “2”。此时,如果要提取数据元素 0,必须修改其在预置键值对中的预置键的定义,即由 key=1 修改为 key=2,否则就会提取错误的数据元素。为了保证当数据元素移位时,不必重新修改其在预置键值对中的预置键的定义,优选地,为每个数据元素设定一个别名,如下所示,每个数据元素都对应一个别名。

[0042]

```
<aliases>

<alias name="appStatus" key="1"/>
<alias name="remoteIp" key="2"/>
<alias name="partnerId" key="3"/>
<alias name="format" key="4"/>
<alias name="appKey" key="5"/>
<alias name="apiName" key="6"/>
<alias name="readBytes" key="7"/>
<alias name="errorCode" key="8"/>
<alias name="subErrorCode" key="9"/>
<alias name="localIp" key="10"/>
<alias name="nick" key="11"/>
<alias name="version" key="12"/>
```

[0043]

```
<alias name="signMethod" key="13"/>
<alias name="tag" key="14"/>
<alias name="id" key="15"/>
<alias name="responseMappingTime" key="16"/>
<alias name="timestamp0" key="17"/>
<alias name="timestamp1" key="18"/>
<alias name="timestamp2" key="19"/>
<alias name="timestamp3" key="20"/>
<alias name="timestamp4" key="21"/>
<alias name="timestamp5" key="22"/>
<alias name="timestamp6" key="23"/>
<alias name="timestamp7" key="24"/>
<alias name="timestamp8" key="25"/>
<alias name="timestamp9" key="26"/>
<alias name="timestamp10" key="27"/>
<alias name="timestamp11" key="28"/>
<alias name="timestamp12" key="29"/>
<alias name="timestamp13" key="30"/>
</aliases>
```

[0044] 由上述内容可知,在一个记录中,第一个数据元素的别名为“appStatus”,第二个数据元素的别名为“remoteIp”,……,依此类推。此时,上述预置键值对中定义的预置键相应地被别名替换为 :key=“appStatus, remoteIp, partnerId”。可见,即使第一个数据元素 0 在记录中向右移动一位后,其在记录中的顺序发生变化,但是其别名仍为“appStatus”,因此,不必修改预置键值对中的预置键的定义。

[0045] 步骤 105 :从服务器按照预置值的定义,对归类后的原始数据进行分析处理,并将分析处理结果反馈给主服务器,主服务器对接收到的分析处理结果进行汇总。并且还可以进一步执行对应的分析处理工作,例如采用与从服务器相同的处理方式,对收到的分析结果进行分析、合并等工作。

[0046] 例如,如果在预置键值对中定义的预置键为 :key=“version, apiName, format”,定义的预置值为 :value=“average(\$responseMappingTime\$)”,当从服务器按照预置键的定义从获取的原始数据中提取出数据元素 version、apiName 和 format 相同的原始数据(记录)后,即,对原始数据进行归类后,按照预置值的定义,从服务器对提取出的原始数据中的

数据元素 responseMappingTime 进行求平均计算。

[0047] 以下为一个文件中的预置键值对中对预置键和预置值的定义。

[0048]

```
<entryList>
    <entry name="服 务 名 称 " key="version,apiName,format"
value="plain($apiName$)">
        <entry name=" 版 本 号 " key="version,apiName,format"
value="plain($version$)">
            <entry name="返 回 格 式 " key="version,apiName,format"
value="plain($format$)">
                <entry name="Mapping 时 间 " key="version,apiName,format"
value="average($responseMappingTime$) />
                <entry name="Mapping 时 间 最 大 " key="version,apiName,format"
value="max($responseMappingTime$) />
                <entry name="业 务 平 均 消 耗 时 间 (ms)"
key="version,apiName,format" value="average($timestamp9$)" />
                <entry name="处 理 总 数 " key="version,apiName,format"
value="count()"/>
</entryList>
```

[0049] 其中,在第一条预置键值对中,预置键定义了从服务器对数据元素 version、apiName 和 format 相同的原始数据进行归类,预置值定义了从服务器对归类后的原始数据中的数据元素 apiName 进行显示;

[0050] 在第二条预置键值对中,预置键定义了从服务器对数据元素 version、apiName 和 format 相同的原始数据进行归类,预置值定义了从服务器对归类后的原始数据中的数据元素 version 进行显示;

[0051] 在第三条键值对中,预置键定义了从服务器对数据元素 version、apiName 和 format 相同的原始数据进行归类,预置值定义了从服务器对归类后的原始数据中的数据元素 format 进行显示;

[0052] 在第四条键值对中,预置键定义了从服务器对数据元素 version、apiName 和 format 相同的原始数据进行归类,预置值定义了从服务器对归类后的原始数据中的数据元素 responseMappingTime 进行求平均计算;

[0053] 在第五条键值对中,预置键定义了从服务器对数据元素 version、apiName 和 format 相同的原始数据进行归类,预置值定义了对归类后的原始数据中的数据元素 responseMappingTime 求最大值;

[0054] 在第六条键值对中,预置键定义了从服务器对数据元素 version、apiName 和

format 相同的原始数据进行归类, 预置值定义了从服务器对归类后的原始数据中的数据元素 timestamp9 进行求平均计算;

[0055] 在第七条键值对中, 预置键定义了从服务器对数据元素 version、apiName 和 format 相同的原始数据进行归类, 预置值定义了从服务器统计(count) 归类后预置键相同的原始数据(记录) 的数量。

[0056] 另外, 上述七条键值对中还指定了预置值结果的显示名称, 如, “服务名称”、“版本号”、“返回格式”、“Mapping 时间”、“Mapping 时间最大”、“业务平均消耗时间(ms)”和“处理总数”等。

[0057] 经过上述数据分析处理后, 下面为数据分析处理结果的一个数据片段。

[0058]

服务名称	版本号	返回格式	Mapping 时间	Mapping 时间最大	业务平均消耗时间(ms)	处理总数
taobao.areas.get	1	xml	0	0	88.73333	15
taobao.delivery.send	1	json	0	0	417.2395	3561
taobao.delivery.send	1	xml	0	0	423.9512	1210
taobao.fenxiao.alipay.user.get	1	json	0	0	128.5	10

[0059]

taobao.fenxiao.delivery.send	1	json	0	0	306.25	16
taobao.fenxiao.distributor.add	1	json	0	0	158.2	5
taobao.fenxiao.supplier.punish	1	json	0	0	13.5	4
taobao.fenxiao.supplier.update	1	json	0	0	7	1

[0060] 上述数据片段中, 第一行数据表示, 按照预置键对应的数据元素“version”、“apiName”和“format”进行归类, 即按照“服务名称”、“版本号”、“返回格式”分别为“taobao.areas.get”、“1”和“xml”对数据记录进行归类, 相同的数据记录的处理总数为 15 条, 其 Mapping 时间和 Mapping 时间最大均为 0, 15 条记录统计的业务平均消耗时间为 88.73333ms。

[0061] 从上述实例中可以看出, 本申请实施例通过文件中的预置键值对中对预置键和预置值的定义, 可实现数据的归类、统计分析、报表生成(例如可根据设定的预置键值对的顺序生成报表) 等复杂功能, 例如适合于各种不同类型数据的海量分析、处理, 为并行的数据处理架构中的海量数据的分析、处理以及报表的生成提供了一种方便、灵活、直观、具体的实现方案。

[0062] 当各个从服务器对原始数据进行分析处理后, 将各自的分析处理结果反馈给主服务器, 由主服务器对接收到的分析处理结果进行合并。其中, 主服务器也可按照预置键值对中预置键的定义对从多个从服务器处得到分析处理结果进行归类, 并按照预置键值对中预

置值的定义对归类的分析处理结果进行合并处理。例如，在主服务器中，预置键定义了主服务器对分析处理结果进行归类的方式，预置值定义主服务器对归类后的分析处理结果进行合并处理的方法。例如，假设主服务器接收到了来自 5 个从服务器上报的分析处理结果共 10 个，按照预置键的定义，其中的 7 个分析处理结果可以进行归类，另外的 3 个分析处理结果可以进行归类，则主服务器分别可以对归类后的 7 个分析处理结果按照预置值的定义进行合并处理，以及，对归类后的另外 3 个分析处理结果按照预置值的定义进行合并处理。由于前面已经详细说明了预置键的归类方法和预置值的处理方法，故此处不再赘述。

[0063] 需要说明的是，上述实施例一除了应用于由一个主服务器和多个从服务器所组成的集群系统中外，还可以应用于由一个主线程和多个子线程所组成的一个数据分析服务器中。此时，主线程用于实现主服务器的功能，子线程用于实现从服务器的功能。

[0064] 由上述实施例可以看出，首先从数据源中获取待分析的原始数据，然后按照分割标识将原始数据切割成数据元素，并将切割得到的数据元素作为键值对中的键，再从切割得到的数据元素中，提取出符合预置键值对中的键定义的数据元素，最后按照预置键值对中的值定义，对提取出的数据元素进行分析处理，并将分析处理结果反馈给主服务器，以便主服务器对接收到的分析处理结果进行合并。因此，为并行的数据处理架构中的海量数据进行分析提供了具体的实现方案。

[0065] 实施例二

[0066] 当从服务器按照预置键中对数据元素的定义方式对获取的原始数据进行归类之后，根据用户的实际使用需求，还可对归类后的原始数据进行进一步的过滤，以过滤掉用户不需要的一部分原始数据，保留用户需要的原始数据。因此，本实施例与实施例一的区别在于：为了筛除掉归类后的原始数据中不需要处理的原始数据，在执行完步骤 104 后，还包括：从归类后的原始数据中筛选出符合第一预置过滤条件的原始数据。请参阅图 2，其为本申请一种数据分析方法的另一个实施例的流程图，该方法包括以下步骤：

[0067] 步骤 201：从服务器接收主服务器发送的文件，其中，所述文件中包括待分析的原始数据的数据标识和预置键值对，所述预置键定义了从服务器对待分析的原始数据进行归类的方式，所述预置值定义了从服务器对归类后的原始数据进行分析处理的方式；

[0068] 步骤 202：从服务器根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离；

[0069] 步骤 203：从服务器根据所述分割标识将获取的原始数据切割成数据元素；

[0070] 步骤 204：从服务器按照预置键中对数据元素的定义方式对获取的原始数据进行归类；

[0071] 需要说明的是，上述步骤 201 至步骤 204 的执行过程可以参见实施例一中步骤 101 至步骤 104，此处不再赘述。

[0072] 步骤 205：从服务器从归类后的原始数据中筛选出符合第一预置过滤条件的原始数据；

[0073] 例如，当从服务器初始时从数据源处获取的 10 个原始数据，并经过预置值的归类后，将其中的 7 个原始数据进行了归类，将另外的 3 个原始数据进行了归类。而根据用户的实际使用需求，7 个归类后的原始数据经过第一预置过滤条件的过滤后，从中筛选出 5 个符合第一预置过滤条件的原始数据，从服务器将按照预置值的定义对这 5 个筛选出的原始数

据进行分析处理。

[0074] 其中,第一预置过滤条件包括大于、小于、不等于、大于或者等于和小于或者等于等条件表达式。当然,第一预置过滤条件为用户根据实际使用需求而设置的条件,本申请实施例对其不进行限定。

[0075] 步骤 206 :从服务器按照预置键值对中预置值的定义,对筛选出的原始数据进行分析处理,并将分析处理结果反馈给主服务器。

[0076] 另外,当按照预置键值对中的预置值定义,对筛选出的原始数据进行分析处理之后,根据用户的实际使用需求,有些分析处理结果是不符合使用条件的分析处理结果。优选地,为了筛除掉不符合使用条件的分析处理结果,在按照预置键值对中的预置值的定义,对筛选出的数据元素进行分析处理之后,且将分析处理结果反馈给主服务器之前,还包括:从分析处理得到的分析处理结果中筛选出符合第二预置过滤条件的分析处理结果。

[0077] 其中,第二预置过滤条件为用于根据实际使用需求而设置的条件,支持大于、小于、不等于、大于或者等于、小于或者等于和是否是数字等表达式。

[0078] 另外,对于一个报表数据的分析来说,除了对数据本身作分析以外,还可能需要能够对数据与其他数据作对比分析,产生一些预警,避免出现的问题或者关注的内容被埋没在海量的数据之中。优选地,本申请实施例中,当主服务器对接收到的分析处理结果进行合并处理后,将得到的合并处理结果与同一时间下的历史合并结果进行对比分析,根据对比分析的结果产生预警信号。例如,用户可以根据各自的使用需求设定各种预警条件,当主服务器对接收到的分析处理结果进行合并处理后,将合并处理结果与同一时间下的历史合并结果进行对比分析,判断对比分析的结果是否满足预警条件,如果是,生产预警信号。其中,

[0079] 具体地,可以包括四种对比分析方式:

[0080] 将今天(day)合并处理后的数据与昨天合并处理后的数据进行对比。例如,将今天合并处理后的数据与昨天合并处理后的数据进行比对,预警条件是前者小于后者时,产生预警信号。

[0081] 预警条件为将将今天的数据和上周(week)同一时间的数据进行对比。

[0082] 预警条件为将将今天的数据与上月(month)的同一时间数据进行对比分析。

[0083] 将今天合并处理后的数据与定义的时间同期合并处理后的数据进行对比。

[0084] 当然,根据具体的应用需求,还可以包括其他的对比分析方式,本申请对对比分析方式及预警条件的设立并不进行限定。

[0085] 需要说明的是,上述实施例二除了应用于由一个主服务器和多个从服务器所组成的集群系统中外,同样可以应用于由一个主线程和多个子线程所组成的一个数据分析服务器中。此时,主线程用于实现主服务器的功能,子线程用于实现从服务器的功能。其中,优选地,子线程按照预置键中对数据元素的定义方式,对获取的原始数据进行归类之后,还包括:子线程从归类后的原始数据中筛选出符合第一预置过滤条件的原始数据;则按照预置值的定义,对归类后的原始数据键进行分析处理为:按照预置值的定义,对筛选出的原始数据进行分析处理。

[0086] 优选的,子线程按照预置值的定义,对归类后的原始数据进行分析处理之后,还包括:子线程从分析处理得到的分析处理结果中筛选出符合第二预置过滤条件的分析处理结果;则所述将分析处理结果反馈给主线程为:将筛选出的分析处理结果反馈给主线程。

[0087] 优选的，当主线程对接收到的分析处理结果进行合并处理后，将得到的合并处理结果与同一时间下的历史合并结果进行对比分析，根据对比分析的结果产生预警信号。

[0088] 由上述实施例可以看出，首先从数据源中获取待分析的原始数据，然后按照分割标识将原始数据切割成数据元素，并将切割得到的数据元素作为键值对中的键，再从切割得到的数据元素中，提取出符合预置键值对中的键定义的数据元素，最后按照预置键值对中的值定义，对提取出的数据元素进行分析处理，并将分析处理结果反馈给主服务器，以便主服务器对接收到的分析处理结果进行合并。因此，为并行的数据处理架构中的海量数据进行分析提供了具体的实现方案。并且通过过滤条件可以过滤掉原始数据中不符合条件的数据，使分析处理后的数据更加准确有效。此外还通过设定的预警条件，避免出现的问题或者关注的内容被埋没在海量的数据之中。

[0089] 实施例三

[0090] 与上述一种数据分析方法相对应，本申请实施例还提供了一种数据分析装置。请参阅图3，其为本申请一种从服务器的一个实施例的结构图，该从服务器包括第一文件接收模块301、第一数据获取模块302、第一数据切割模块303、第一数据归类模块304和第一数据计算模块305。下面结合该装置的工作原理进一步介绍其内部结构以及连接关系。

[0091] 第一文件接收模块301，用于接收主服务器发送的文件，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键定义了从服务器对待分析的原始数据进行归类的方式，所述预置值定义了从服务器对归类后的原始数据进行分析处理的方式；

[0092] 第一数据获取模块302，用于根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离；

[0093] 第一数据切割模块303，用于根据所述分割标识将获取的原始数据切割成数据元素；

[0094] 第一数据归类模块304，用于按照预置键中对数据元素的定义方式，对获取的原始数据进行归类；

[0095] 第一数据计算模块305，用于按照预置值的定义，对归类后的原始数据进行分析处理，并将分析处理结果反馈给主服务器。

[0096] 优选的，请参阅图4，其为本申请一种从服务器的另一个实施例的结构图，所述从服务器还包括第一过滤模块306，用于从归类后的原始数据中筛选出符合第一预置过滤条件的原始数据；则第一数据计算模块305按照预置值的定义，对筛选出的原始数据进行分析处理。

[0097] 优选的，请参阅图5，其为本申请一种从服务器的另一个实施例的结构图，所述从服务器还包括：第二过滤模块307，用于从分析处理得到的分析处理结果中筛选出符合第二预置过滤条件的分析处理结果；则第一数据计算模块305将筛选出的分析处理结果反馈给主服务器。

[0098] 由上述实施例可以看出，首先从数据源中获取待分析的原始数据，然后按照分割标识将原始数据切割成数据元素，并将切割得到的数据元素作为键值对中的键，再从切割得到的数据元素中，提取出符合预置键值对中的键定义的数据元素，最后按照预置键值对中的值定义，对提取出的数据元素进行计算，并将计算结果反馈给主服务器，以便主服务器对接收到的计算结果进行合并。因此，为并行的数据处理架构中的海量数据进行分析提供

了具体的实现方案。

[0099] 实施例四

[0100] 与上述一种数据分析方法相对应,本申请实施例还提供了一种数据分析装置。请参阅图6,其为本申请一种服务器的一个实施例的结构示意图。所述服务器包括:第二文件接收模块601、第一数据获取模块602、第一数据切割模块603、第一数据归类模块604和第一数据计算模块605。下面结合该装置的工作原理进一步介绍其内部结构以及连接关系。

[0101] 第二文件接收模块601,用于接收主线程发送的文件,其中,所述文件携带待分析的原始数据的数据标识和预置键值对,所述预置键定义了子线程对待分析的原始数据进行归类的方式,所述预置值定义了子线程对归类后的原始数据进行分析处理的方式;

[0102] 第二数据获取模块602,用于根据所述数据标识,从数据源中获取待分析的原始数据,其中,所述原始数据中数据元素之间以分割标识进行隔离;

[0103] 第二数据切割模块603,用于根据所述分割标识将获取的原始数据切割成数据元素;

[0104] 第二数据归类模块604,用于按照预置键中对数据元素的定义方式,对获取的原始数据进行归类;

[0105] 第二数据计算模块605,用于按照预置值的定义,对归类后的原始数据进行分析处理,并将分析处理结果反馈给主线程。

[0106] 优选的,子线程按照预置键中对数据元素的定义方式,对获取的原始数据进行归类之后,所述服务器还包括:第三过滤模块,用于从归类后的原始数据中筛选出符合第一预置过滤条件的原始数据;则第二数据计算模块605按照预置值的定义,对归类后的原始数据键进行分析处理为:按照预置值的定义,对筛选出的原始数据进行分析处理。

[0107] 优选的,子线程按照预置值的定义,对归类后的原始数据进行分析处理之后,所述服务器还包括:第四过滤模块,用于从分析处理得到的分析处理结果中筛选出符合第二预置过滤条件的分析处理结果;则第二数据计算模块605将分析处理结果反馈给主线程为:将筛选出的分析处理结果反馈给主线程。

[0108] 由上述实施例可以看出,首先从数据源中获取待分析的原始数据,然后按照分割标识将原始数据切割成数据元素,并将切割得到的数据元素作为键值对中的键,再从切割得到的数据元素中,提取出符合预置键值对中的键定义的数据元素,最后按照预置键值对中的值定义,对提取出的数据元素进行计算,并将计算结果反馈给主服务器,以便主服务器对接收到的计算结果进行合并。因此,为并行的数据处理架构中的海量数据进行分析提供了具体的实现方案。

[0109] 实施例五

[0110] 本申请还提供了一种数据分析系统,请参阅图7,其为本申请一种数据分析系统的一个实施例的结构图。所述系统包括:一主服务器701和至少两个从服务器702,其中,

[0111] 主服务器701,用于向从服务器702发送文件,并对接收到的分析处理结果进行合并,其中,所述文件携带待分析的原始数据的数据标识和预置键值对,所述预置键定义了从服务器702对待分析的原始数据进行归类的方式,所述预置值定义了从服务器702对归类后的原始数据进行分析处理的方式;

[0112] 从服务器702,用于接收主服务器701发送的文件,根据所述数据标识,从数据源

中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离，根据所述分割标识将获取的原始数据切割成数据元素，按照预置键中对数据元素的定义方式，对获取的原始数据进行归类，按照预置值的定义，对归类后的原始数据进行分析处理，并将分析处理结果反馈给主服务器 701。

[0113] 优选的，当主服务器 701 对接收到的分析处理结果进行合并后，主服务器 701 还用于将得到的分析处理结果与同一时间下的历史合并结果进行对比分析，并根据对比分析的结果产生预警信号。

[0114] 由上述实施例可以看出，首先从数据源中获取待分析的原始数据，然后按照分割标识将原始数据切割成数据元素，并将切割得到的数据元素作为键值对中的键，再从切割得到的数据元素中，提取出符合预置键值对中的键定义的数据元素，最后按照预置键值对中的值定义，对提取出的数据元素进行计算，并将计算结果反馈给主服务器，以便主服务器对接收到的计算结果进行合并。因此，为并行的数据处理架构中的海量数据进行分析提供了具体的实现方案。

[0115] 实施例六

[0116] 本申请还提供了一种数据分析系统，请参阅图 8，其为本申请一种数据分析系统的另一个实施例的结构图。所述数据分析系统包括：一主线程模块 801 和至少两个子线程模块 802，其中，

[0117] 主线程模块 801，用于向子线程模块 802 发送文件，并对接收到的分析处理结果进行合并，其中，所述文件携带待分析的原始数据的数据标识和预置键值对，所述预置键值对定义了子线程模块 802 对待分析的原始数据进行归类的方式，所述预置值定义了子线程模块 802 对归类后的原始数据进行分析处理的方式；

[0118] 子线程模块 802，用于接收主线程模块 801 发送的文件，根据所述数据标识，从数据源中获取待分析的原始数据，其中，所述原始数据中数据元素之间以分割标识进行隔离，根据所述分割标识将获取的原始数据切割成数据元素，按照预置键中对数据元素的定义方式，对获取的原始数据进行归类，按照预置值的定义，对归类后的原始数据进行分析处理，并将分析处理结果反馈给主线程模块 801。

[0119] 由上述实施例可以看出，首先从数据源中获取待分析的原始数据，然后按照分割标识将原始数据切割成数据元素，并将切割得到的数据元素作为键值对中的键，再从切割得到的数据元素中，提取出符合预置键值对中的键定义的数据元素，最后按照预置键值对中的值定义，对提取出的数据元素进行计算，并将计算结果反馈给主服务器，以便主服务器对接收到的计算结果进行合并。因此，为并行的数据处理架构中的海量数据进行分析提供了具体的实现方案。

[0120] 需要说明的是，本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory, ROM) 或随机存储记忆体(Random Access Memory, RAM) 等。

[0121] 以上对本申请所提供的一种数据分析方法、系统及服务器进行了详细介绍，本文中应用了具体实施例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于

帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

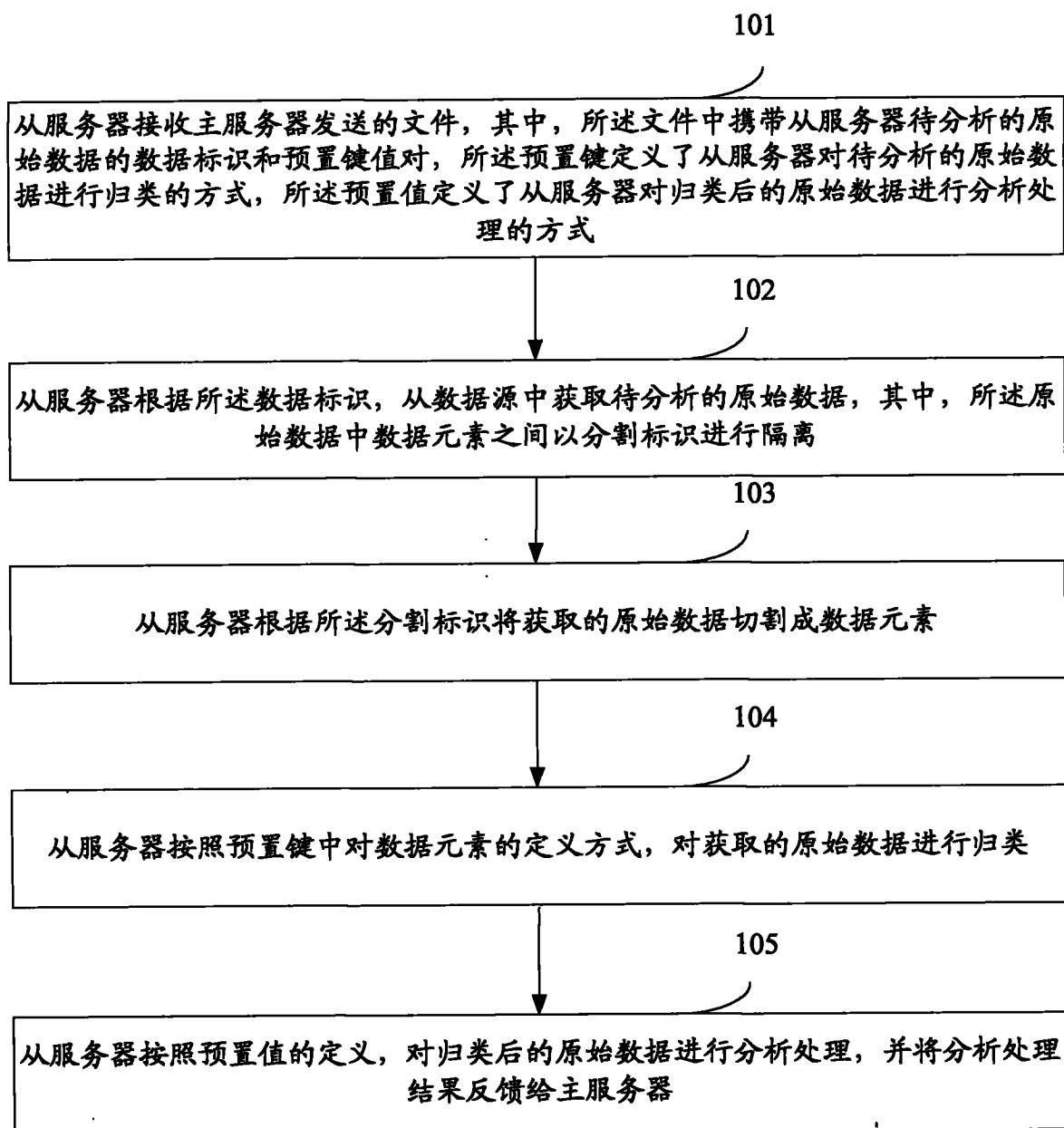


图 1

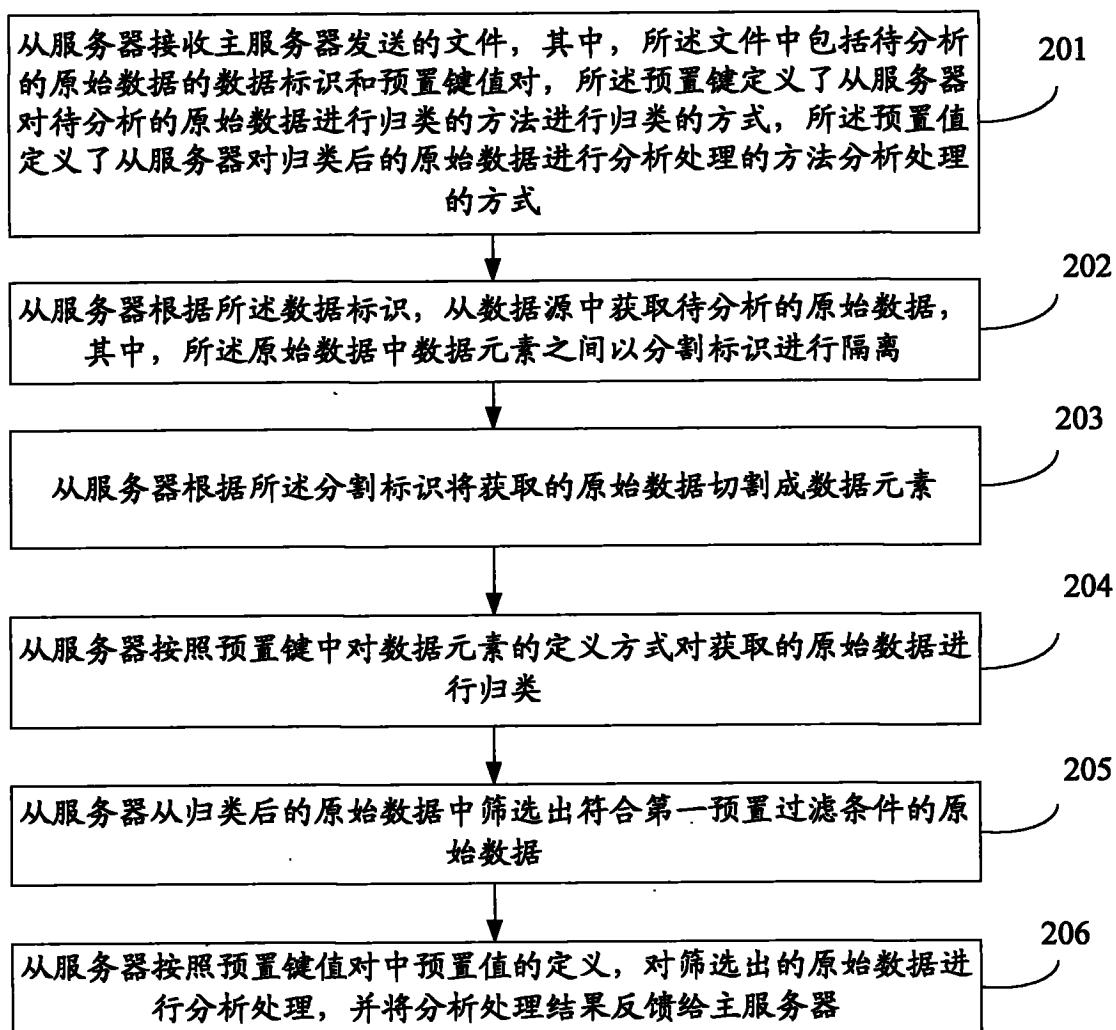
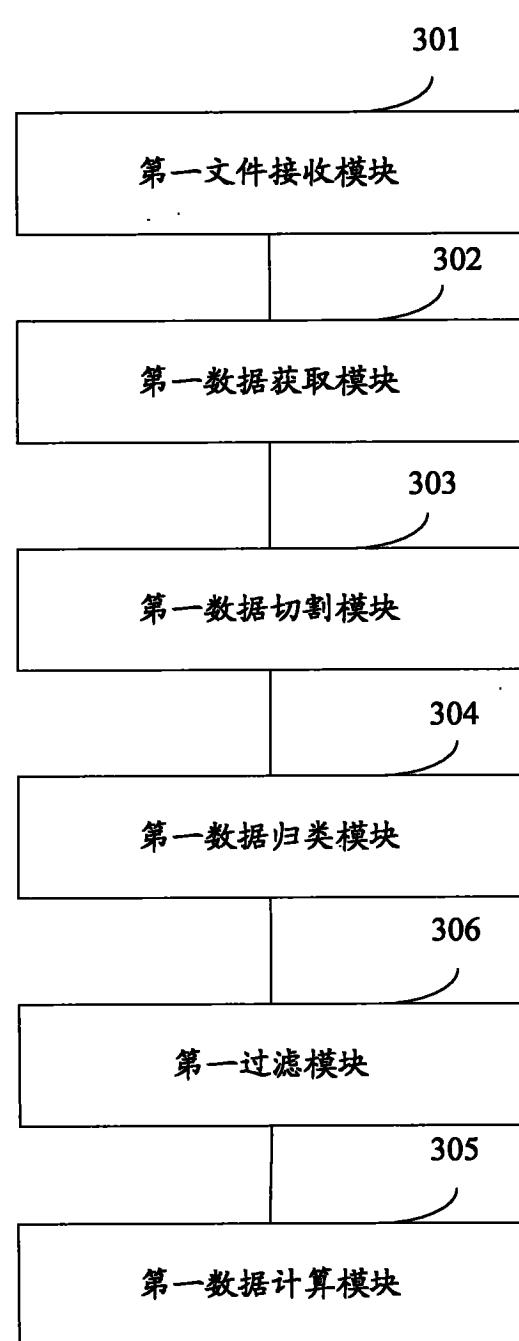
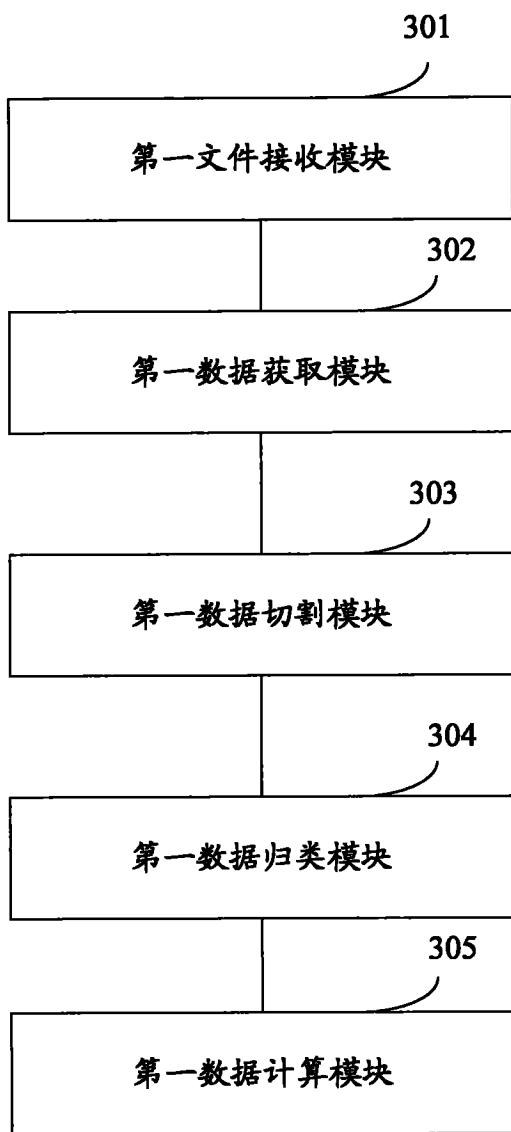


图 2



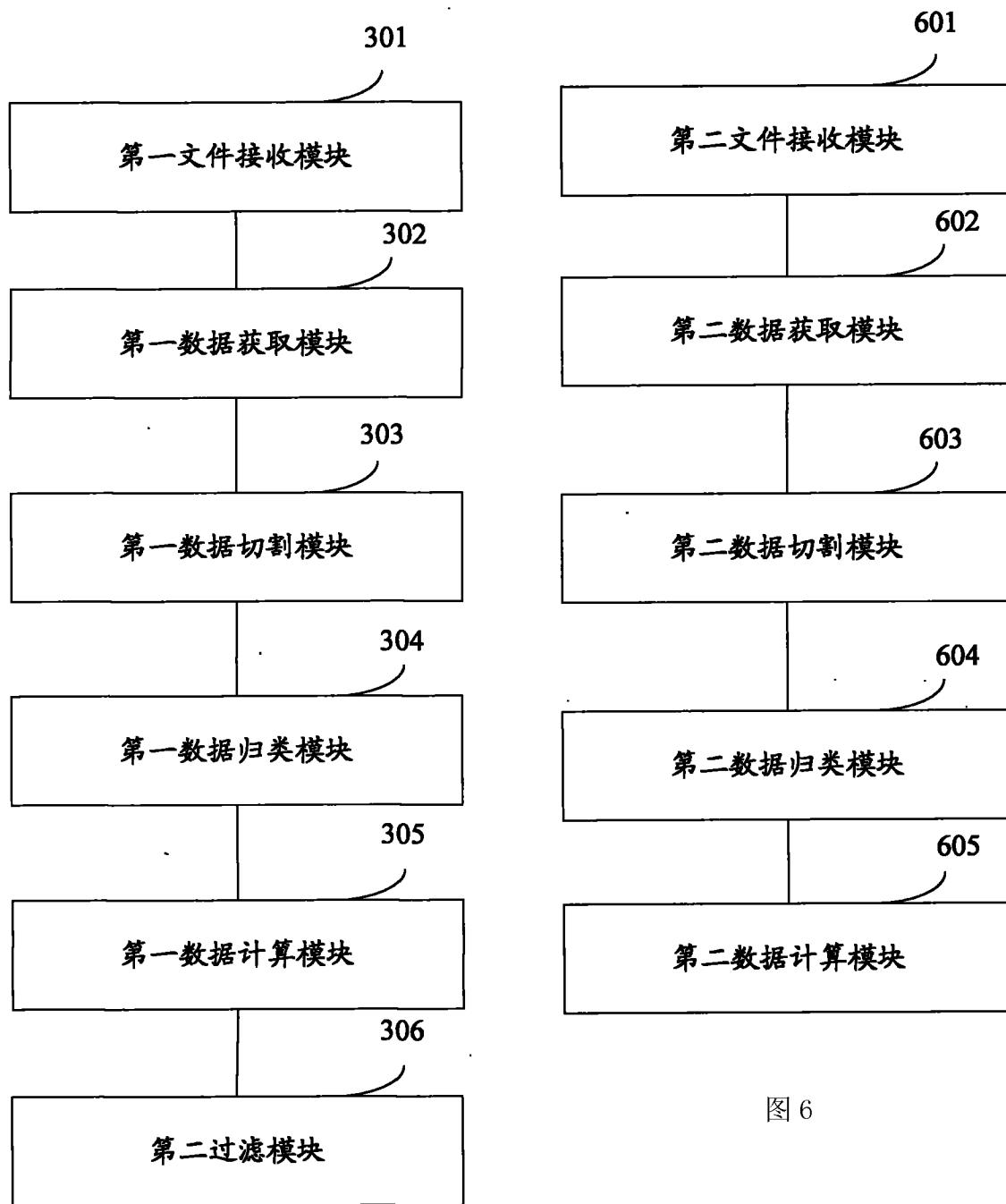


图 5

图 6

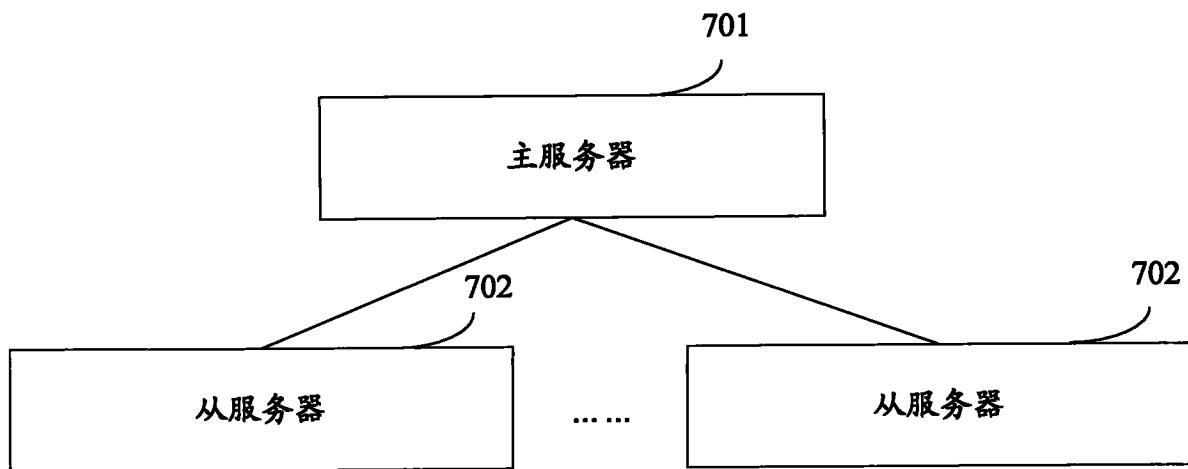


图 7

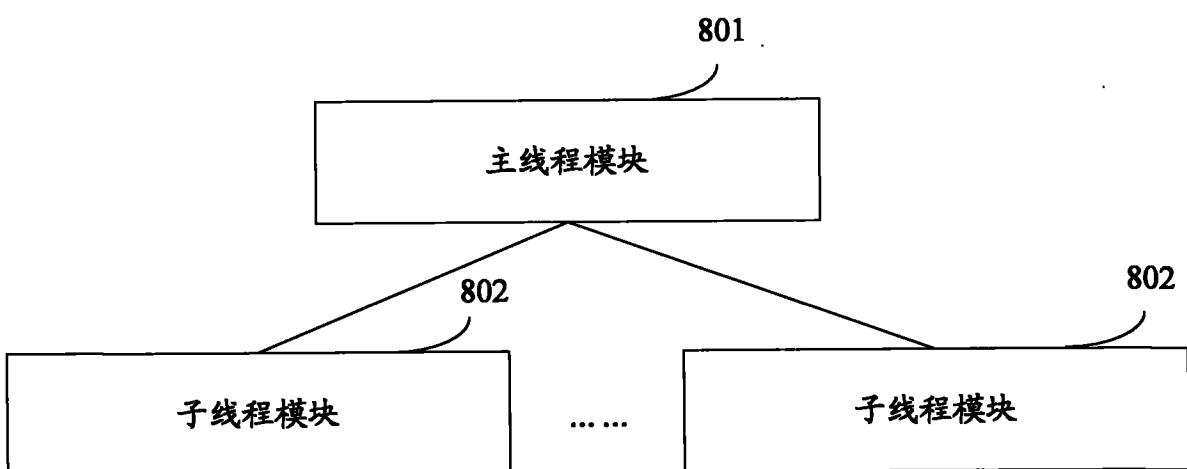


图 8