(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113609374 A
(43)申请公布日 2021.11.05

(21)申请号 202110160293.X

(22)申请日 2021.02.05

(71)申请人 腾讯科技(深圳)有限公司
地址 518044 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 刘刚

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 朱佳

(51)Int.Cl.
G06F 16/9535(2019.01)

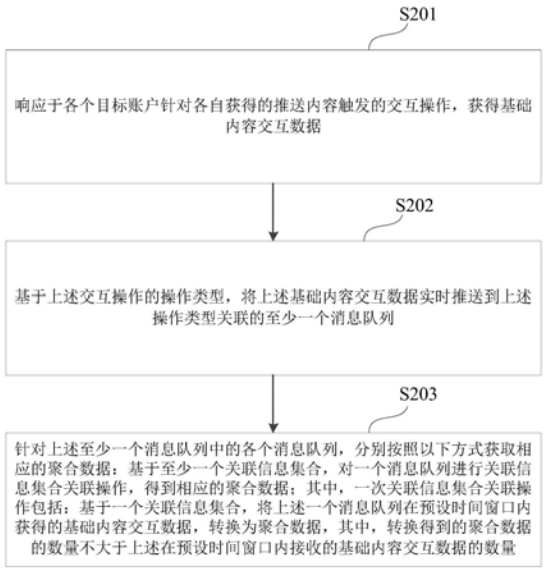
权利要求书3页 说明书27页 附图11页

(54)发明名称

基于内容推送的数据处理方法、装置、设备及存储介质

(57)摘要

本申请提供一种基于内容推送的数据处理方法、装置、设备及存储介质，涉及数据处理技术领域，以提升对内容交互数据进行处理的及时性。该方法包括：基于交互操作的操作类型，将针对推送内容触发的交互操作确定的基础内容交互数据，实时推送到对应的消息队列，针对各个消息队列，分别按照以下方式获取相应的聚合数据：基于至少一个关联信息集合，对一个消息队列进行关联信息集合关联操作，得到相应的聚合数据；一次关联信息集合关联操作包括：基于一个关联信息集合，将一个消息队列在预设时间窗口内获得的基础内容交互数据，转换为聚合数据。该方法中能够实时对基础内容交互数据进行多维度的关联操作处理。



CN 113609374 A

1.一种基于内容推送的数据处理方法,其特征在于,包括:

响应于各个目标账户针对各自获得的推送内容触发的交互操作,获得基础内容交互数据;

基于所述交互操作的操作类型,将所述基础内容交互数据实时推送到所述操作类型关联的至少一个消息队列;

针对所述至少一个消息队列中的各个消息队列,分别按照以下方式执行数据关联操作,获取相应的聚合数据:

基于至少一个关联信息集合,对一个消息队列进行数据关联操作,得到相应的聚合数据;其中,一次数据关联操作包括:基于一个关联信息集合,将所述一个消息队列在预设时间窗口内获得的基础内容交互数据,转换为聚合数据,其中,转换得到的聚合数据的数量不大于所述在预设时间窗口内接收的基础内容交互数据的数量。

2.如权利要求1所述的方法,其特征在于,每个基础内容交互数据包含触发所述交互操作的目标账户关联的账户标识和触发所述交互操作的推送内容关联的内容标识,所述基于一个关联信息集合,将所述一个消息队列在预设时间窗口内获得的基础内容交互数据,转换为聚合数据,包括以下操作中的任意一种或组合:

若所述关联信息集合包括账户画像集合,则将所述一个消息队列在第一预设时间窗口内容获得的基础内容交互数据,确定为第一交互数据集合,针对所述第一交互数据集合中基础内容交互数据包含的各个账户标识,分别执行如下操作:基于所述账户画像集合中记录的各个目标账户的账户画像数据,获取一个账户标识关联的目标账户的账户画像数据,通过获取的账户画像数据,将所述第一交互数据集合中包含所述一个账户标识的基础内容交互数据进行聚合,得到一个第一聚合数据;

若所述关联信息集合包括内容信息集合,则将所述一个消息队列在第二预设时间窗口内获得的基础内容交互数据,确定为第二交互数据集合,针对所述第二交互数据集合中基础内容交互数据包含的各个内容标识,分别执行如下操作:基于所述内容信息集合中记录的各个推送内容的内容信息,获取一个内容标识关联的推送内容的内容信息,通过获取的内容信息,将所述第二交互数据集合中包含所述一个内容标识的基础内容交互数据进行聚合,得到一个第二聚合数据。

3.如权利要求2所述的方法,其特征在于,所述通过获取的账户画像数据,将所述第一交互数据集合中包含所述一个账户标识的基础内容交互数据进行聚合,得到一个第一聚合数据,包括:

确定所述第一交互数据集合中包含所述一个账户标识的基础内容交互数据;

获取确定的基础内容交互数据中包含的内容标识,生成内容标识集合;

将获取的账户画像数据、所述一个账户标识和所述内容标识集合进行关联,获得相应的一个第一聚合数据。

4.如权利要求2所述的方法,其特征在于,所述通过获取的内容信息,将所述第二交互数据集合中包含所述一个内容标识的基础内容交互数据进行聚合,得到一个第二聚合数据,包括:

确定所述第二交互数据集合中包含所述一个内容标识的基础内容交互数据;

获取确定的基础内容交互数据中包含的账户标识,生成账户标识集合;

将获取的内容信息、所述一个内容标识和所述账户标识集合进行关联，获得相应的一个第二聚合数据。

5.如权利要求2所述的方法，其特征在于，所述账户画像集合存储于第一键值对数据库，所述第一键值对数据库是基于第一周期周期性更新的；

所述内容信息集合存储于第二键值对数据库，所述第二键值对数据库是基于第二周期对内容数据库进行周期性备份得到的，所述内容数据库用于实时记录所述推送内容的内容信息。

6.如权利要求1所述的方法，其特征在于，所述针对所述至少一个消息队列中的各个消息队列，分别按照以下方式执行数据关联操作，获取相应的聚合数据之后，还包括：

针对获得的各个聚合数据，分别按照以下方式执行数据存储操作，将所述各个聚合数据分别存储到对应的磁盘分片中：将一个聚合数据存储至所述一个聚合数据关联的推送内容映射的磁盘分片中，其中，一个磁盘分片用于存储一个推送内容关联的聚合数据；

以及

响应于针对待查询推送内容的数据查询请求，从所述待查询推送内容映射的磁盘分片中，获取所述待查询推送内容关联的聚合数据；

按照目标业务需求关联的数据处理规则，对获取的聚合数据进行数据处理，获得查询数据，并返回所述查询数据。

7.如权利要求6所述的方法，其特征在于，所述数据查询请求中携带所述待查询推送内容的内容标识和查询时间段，所述从所述待查询推送内容映射的磁盘分片中，获取所述待查询推送内容关联的聚合数据，包括：

基于所述内容标识，将所述待查询推送内容映射的磁盘分片确定为目标磁盘分片；以及

通过预设时间粒度，将所述查询时间段划分为至少一个子时间段；

根据所述待查询推送内容关联的聚合数据中，各个子时间段映射的聚合数据在所述目标磁盘分片中的存储地址，确定所述数据查询请求对应的数据查询索引信息；

基于所述数据索引信息，从所述各个子时间段映射的聚合数据在所述目标磁盘分片中的存储地址中，获取所述各个子时间段映射的聚合数据。

8.一种基于内容推送的数据处理装置，其特征在于，包括：

数据采集单元，用于响应于各个目标账户针对各自获得的推送内容触发的交互操作，获得基础内容交互数据；

数据拆分单元，用于基于所述交互操作的操作类型，将所述基础内容交互数据实时推送到所述操作类型关联的至少一个消息队列；

数据聚合单元，用于针对所述至少一个消息队列中的各个消息队列，分别按照以下方式执行数据关联操作，获取相应的聚合数据：基于至少一个关联信息集合，对一个消息队列进行关联信息集合关联操作，得到相应的聚合数据；其中，一次关联信息集合关联操作包括：基于一个关联信息集合，将所述一个消息队列在预设时间窗口内获得的基础内容交互数据，转换为聚合数据，其中，转换得到的聚合数据的数量不大于所述在预设时间窗口内接收的基础内容交互数据的数量。

9.一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计

算机程序,其特征在于,所述处理器执行所述程序时实现权利要求1-7中任一权利要求所述的方法。

10.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机指令,当所述计算机指令在计算机上运行时,使得计算机执行如权利要求1-7中任一项所述的方法。

基于内容推送的数据处理方法、装置、设备及存储介质

技术领域

[0001] 本申请涉及数据处理技术领域，尤其涉及一种基于内容推送的数据处理方法、装置、设备及存储介质。

背景技术

[0002] 在自媒体时代，内容推送系统可以对账户针对推送内容进行交互操作产生的内容交互数据进行离线数据处理，将离线数据处理的结果作为中间数据，存储到数据仓库中，进而账户可以基于上述中间数据查询与其创作的内容相关的数据；但上述过程中需要对海量的内容交互数据延后一定时间如一天才进行数据处理，导致处理内容交互数据的实时性低，因此如何提升对内容交互数据进行处理的及时性，以提升账户基于数据处理得到的中间数据查询内容相关的数据的及时性，成为了需要考虑的问题。

发明内容

[0003] 本申请实施例提一种基于内容推送的数据处理方法、装置、设备及存储介质，用于提升对针对推送内容的交互数据进行处理的及时性。

[0004] 本申请第一方面，提供一种基于内容推送的数据处理方法，包括：

[0005] 响应于各个目标账户针对各自获得的推送内容触发的交互操作，获得基础内容交互数据；

[0006] 基于所述交互操作的操作类型，将所述基础内容交互数据实时推送到所述操作类型关联的至少一个消息队列；

[0007] 针对所述至少一个消息队列中的各个消息队列，分别按照以下方式执行数据关联操作，获取相应的聚合数据：

[0008] 基于至少一个关联信息集合，对一个消息队列进行关联信息集合关联操作，得到相应的聚合数据；其中，一次关联信息集合关联操作包括：基于一个关联信息集合，将所述一个消息队列在预设时间窗口内获得的基础内容交互数据，转换为聚合数据，其中，转换得到的聚合数据的数量不大于所述在预设时间窗口内接收的基础内容交互数据的数量。

[0009] 本申请第二方面，提供一种基于内容推送的数据处理装置，包括：

[0010] 数据采集单元，用于响应于各个目标账户针对各自获得的推送内容触发的交互操作，获得基础内容交互数据；

[0011] 数据拆分单元，用于基于所述交互操作的操作类型，将所述基础内容交互数据实时推送到所述操作类型关联的至少一个消息队列；

[0012] 数据聚合单元，用于针对所述至少一个消息队列中的各个消息队列，分别按照以下方式执行数据关联操作，获取相应的聚合数据：基于至少一个关联信息集合，对一个消息队列进行关联信息集合关联操作，得到相应的聚合数据；其中，一次关联信息集合关联操作包括：基于一个关联信息集合，将所述一个消息队列在预设时间窗口内获得的基础内容交互数据，转换为聚合数据，其中，转换得到的聚合数据的数量不大于所述在预设时间窗口内

接收的基础内容交互数据的数量。

[0013] 在一种可能的实现方式中，每个基础内容交互数据包含触发所述交互操作的目标账户关联的账户标识和触发所述交互操作的推送内容关联的内容标识，所述数据聚合单元具体用于执行以下操作中的任意一种或组合：

[0014] 若所述关联信息集合包括账户画像集合，则将所述一个消息队列在第一预设时间窗口内容获得的基础内容交互数据，确定为第一交互数据集合，针对所述第一交互数据集合中基础内容交互数据包含的各个账户标识，分别执行如下操作：基于所述账户画像集合中记录的各个目标账户的账户画像数据，获取一个账户标识关联的目标账户的账户画像数据，通过获取的账户画像数据，将所述第一交互数据集合中包含所述一个账户标识的基础内容交互数据进行聚合，得到一个第一聚合数据；

[0015] 若所述关联信息集合包括内容信息集合，则将所述一个消息队列在第二预设时间窗口内获得的基础内容交互数据，确定为第二交互数据集合，针对所述第二交互数据集合中基础内容交互数据包含的各个内容标识，分别执行如下操作：基于所述内容信息集合中记录的各个推送内容的内容信息，获取一个内容标识关联的推送内容的内容信息，通过获取的内容信息，将所述第二交互数据集合中包含所述一个内容标识的基础内容交互数据进行聚合，得到一个第二聚合数据。

[0016] 在一种可能的实现方式中，所述数据聚合单元具体用于：

[0017] 确定所述第一交互数据集合中包含所述一个账户标识的基础内容交互数据；

[0018] 获取确定的基础内容交互数据中包含的内容标识，生成内容标识集合；

[0019] 将获取的账户画像数据、所述一个账户标识和所述内容标识集合进行关联，获得相应的一个第一聚合数据。

[0020] 在一种可能的实现方式中，所述数据聚合单元具体用于：

[0021] 确定所述第二交互数据集合中包含所述一个内容标识的基础内容交互数据；

[0022] 获取确定的基础内容交互数据中包含的账户标识，生成账户标识集合；

[0023] 将获取的内容信息、所述一个内容标识和所述账户标识集合进行关联，获得相应的一个第二聚合数据。

[0024] 在一种可能的实现方式中，所述账户画像集合存储于第一键值对数据库，所述第一键值对数据库是基于第一周期周期性更新的；

[0025] 所述内容信息集合存储于第二键值对数据库，所述第二键值对数据库是基于第二周期对内容数据库进行周期性备份得到的，所述内容数据库用于实时记录所述推送内容的内容信息。

[0026] 在一种可能的实现方式中，所述数据聚合单元还用于：

[0027] 针对所述至少一个消息队列中的各个消息队列，分别按照以下方式执行数据关联操作，获取相应的聚合数据之后，针对获得的各个聚合数据，分别按照以下方式执行数据存储操作，将所述各个聚合数据分别存储到对应的磁盘分片中：将一个聚合数据存储至所述一个聚合数据关联的推送内容映射的磁盘分片中，其中，一个磁盘分片用于存储一个推送内容关联的聚合数据；

[0028] 以及

[0029] 响应于针对待查询推送内容的数据查询请求，从所述待查询推送内容映射的磁盘

分片中，获取所述待查询推送内容关联的聚合数据；

[0030] 按照目标业务需求关联的数据处理规则，对获取的聚合数据进行数据处理，获得查询数据，并返回所述查询数据。

[0031] 在一种可能的实现方式中，所述数据查询请求中携带所述待查询推送内容的内容标识和查询时间段，所述数据聚合单元具体用于：

[0032] 基于所述内容标识，将所述待查询推送内容映射的磁盘分片确定为目标磁盘分片；以及

[0033] 通过预设时间粒度，将所述查询时间段划分为至少一个子时间段；

[0034] 根据所述待查询推送内容关联的聚合数据中，各个子时间段映射的聚合数据在所述目标磁盘分片中的存储地址，确定所述数据查询请求对应的数据查询索引信息；

[0035] 基于所述数据索引信息，从所述各个子时间段映射的聚合数据在所述目标磁盘分片中的存储地址中，获取所述各个子时间段映射的聚合数据。

[0036] 本申请第三方面，提供一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现第一方面所述的方法。

[0037] 本申请第四方面，提供一种计算机程序产品，该计算机程序产品包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行上述第一方面中提供的方法。

[0038] 本申请第五方面，提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机指令，当所述计算机指令在计算机上运行时，使得计算机执行如第一方面所述的方法。

[0039] 由于本申请实施例采用上述技术方案，至少具有如下技术效果：

[0040] 本申请实施例中，一方面，实时的将获得的基础内容交互数据推送到对应的消息队列中进行处理，获得相应的聚合数据，实现了对基础内容交互数据的实时处理，提升了数据处理的及时性，进而可以提升基于聚合数据查询目标账户对推送内容的反馈信息的及时性以及准确度；另一方面，本申请实施例中基于关联信息集合，将各个消息队列中在各个时间窗口内获得的大量的基础内容交互数据，转换为较为少量的聚合数据，减少了反映目标账户对推送内容的反馈信息的聚合数据的数量，进而账户基于聚合数据查询目标账户对推送内容的反馈信息时，需要查询和处理的聚合数据的数量明显减少，降低了查询目标账户对推送内容的反馈信息的时延，以及提升查询的效率。

附图说明

[0041] 图1为本申请实施例提供的一种基于内容推送的数据处理的应用场景的示意图；

[0042] 图2为本申请实施例提供的一种基于内容推送的数据处理方法的流程的示例图；

[0043] 图3为本申请实施例提供的一种更新账户画像集合和内容信息集合的原理的示意图；

[0044] 图4为本申请实施例提供的一种账户维度关联的过程的示意图；

[0045] 图5为本申请实施例提供的一种信息维度关联的过程的示意图；

[0046] 图6为本申请实施例提供的一种将聚合数据写入一个磁盘分片的过程的示意图；

[0047] 图7为本申请实施例提供的一种基于数据查询请求获取聚合数据的过程的示意图；

[0048] 图8为本申请实施例提供的一种从目标磁盘分片中获取聚合数据的过程的示意图；

[0049] 图9为本申请实施例提供的一种内容推送系统的框架的示意图；

[0050] 图10为本申请实施例提供的一种基于推送内容的数据处理方法的流程的示例图；

[0051] 图11为本申请实施例提供的一种对基础内容交互进行聚合的过程的示例图；

[0052] 图12为本申请实施例提供的一种基于内容推送的数据处理装置的结构示例图；

[0053] 图13为本申请实施例提供的一种计算机设备的结构图。

具体实施方式

[0054] 为了更好的理解本申请实施例提供的技术方案，下面将结合说明书附图以及具体的实施方式进行详细的说明。

[0055] 为了便于本领域技术人员更好地理解本申请的技术方案，下面对本申请涉及的部分概念进行说明。

[0056] 1)内容和推送内容

[0057] 在自媒体时代，内容一般可以指代音频、视频、图文等；本申请实施例中推送内容为内容推送系统向账户推送的内容；本申请实施例中的推送内容可以是单个用户的账户创作发布的内容，也可以是多个用户组成的团体对应的账户发布的内容等，推送内容也可以是专业生产内容(Professionally-generated Content，PGC)或者用户生产内容(User-generated Content，UGC)的账户主动发布的；

[0058] 本申请实施例中的推送内容可以但不局限于包括文本、音频、视频、文章、图文、图片等中至少一种信息或任意组合得到的多媒体资源；其中，文章可以但不局限于是自媒体创建的公众号所主动编辑发布的图文，图文可以但不局限于包括竖版的小图文、横版的短图文，或能上下滑动或左右滑动的图文等；视频可以但不局限于是专业生产内容或者用户生产内容的用户提供，最后以信息流(即Feeds)流的形式提供。

[0059] 2)内容发布账户、目标账户和查询账户

[0060] 一般情况下账户是用户的身份表示，本申请实施例中涉及的各种账户又可成为用户；本申请实施例中将向内容推送服务器发送推送内容的账户成为内容发布账户(又可以称为内容生产者或内容生产端)，将接收内容推送服务器分发的推送内容的账户称为目标账户(又可以称为内容消费者或内容消费端)，将触发针对推送内容的数据查询的账户称为查询账户，其中，查询账户可以是上述内容生产者，也可以是内容推送系统中各个内容空间关联的账户，上述内容可以但不局限于是内容推送系统中发布推送内容的内容频道或内容社区等。

[0061] 3)基础内容交互数据和关联信息集合

[0062] 本申请实施例中基础内容交互数据可以是目标账户针对推送内容进行的交互操作产生的数据，基础内容交互数据可以表征目标账户对推送内容的反馈信息；本申请实施例中，基础内容交互数据可以但不局限于包括触发交互操作的目标账户关联的账户标识和触发上述交互操作的推送内容关联的内容标识。

[0063] 本申请实施例中的关联信息集合中包含用于将多个基础内容交互数据进行聚合的信息，如关联信息集合中可以包括触发上述交互操作的目标账户的有关信息，关联信息集合中也可以包括触发上述交互操作的内容的有关信息等；其中本申请实施例中涉及的关联信息集合可以有多种表现形式，比如包含目标账户的有关信息或内容的有关信息的维表，后续的实施例中，将以维表作为关联信息集合的例子进行说明。

[0064] 4)专业生产内容(Professionally-generated Content，PGC)和用户生产内容(User-generated Content，UGC)

[0065] 上述PGC是一种互联网术语，表示专业生产内容的机构或者组织，例如视频网站、专家生产内容，例如微博等。

[0066] UGC指用户原创内容，是伴随着以提倡个性化为主要特点的Web2.0概念而兴起的，它并不是某一种具体的业务，而是一种用户使用互联网的新方式，即由原来的以下载内容为主的互联网使用方式，逐渐转变成下载内容和上传内容并重的互联网使用方式，指的是用户生产内容；在UGC平台中，任何人都可以上传相关的图文、视频、图片等内容，目前的即时通讯应用中用户分享的文章、小视频，以及各类短视频应用中的短视频都属于UGC的方式产生的内容。

[0067] 5)Feeds流(即信息流)

[0068] Feeds流是持续更新并呈现给用户内容的信息流，即消息来源，又可以称为源料、馈送、资讯提供、供稿、摘要、源、新闻订阅、网源(web feed、news feed、syndicated feed)等；Feed是一种给用户持续提供内容的数据形式，是由多个提供内容的消息源组成的资源聚合器，由用户主动订阅提供内容的消息源源并向用户提供内容，即Feed是将用户主动订阅的若干消息源组合在一起形成内容聚合器，帮助用户持续地获取订阅的消息源的最新的内容Feeds流，即Feeds流是一种资料格式，网站透过Feeds流将最新的资讯信息传递给用户，通常以时间轴方式排列，时间线(Timeline)是Feeds流中最原始最直觉也最基本的展示形式。

[0069] 其中用户能够订阅网站的先决条件是，该网站提供了消息来源；将Feeds流汇流于一处称为聚合(Aggregation)，而用于聚合的软体称为聚合器(Aggregator)。对最终用户而言，聚合器是专门用来订阅网站的软件，一般亦称为RSS阅读器、feed阅读器、新闻阅读器等。

[0070] 6)独立访问者带来的页面访问量PV/UV

[0071] 页面访问量(Page View，PV)，指用户每次对网站的访问均被记录，用户对同一页面的多次访问，指网站的页面浏览量或点击量，访问量累计；独立访问用户数(Unique Visitor，UV)访问网站的一台电脑客户端为一个访客，24小时之内，同一地址，多次访问，只算一次；

[0072] PV/UV＝单位UV带来的页面访问量(独立访问者所浏览的页面访问量)，反映了页面访问质量的其中一个因素。

[0073] 为了使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请作进一步地详细描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

[0074] 本申请实施例涉及人工智能(Artificial Intelligence,AI)和机器学习技术,基于人工智能中的计算机视觉技术和机器学习(MachineLearning,ML)而设计,尤其涉及人工智能中的大数据处理技术。

[0075] 人工智能是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。

[0076] 机器学习是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。

[0077] 本申请实施例中涉及的基础内容交互数据是海量的大数据(Big Data),大数据一般指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合,是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产,大数据的主要特点是数量大、种类多、复杂、难处理、价值大;本申请实施例中提供的数据处理方法涉及大数据处理技术,其中大数据处理的流程主要包括数据收集、数据预处理、数据存储、数据处理与分析、数据展示或数据可视化、数据应用等环节,其中数据质量贯穿于整个大数据处理的流程,每一个数据处理环节都会对大数据的质量产生影响作用,大数据处理的技术可以但不局限于包括大规模并行处理(Massively Parallel Processing,MPP)数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统等;其中云计算平台也称为云平台,是指基于硬件资源和软件资源的服务,提供计算、网络和存储能力。

[0078] 其中,大数据处理技术包括离线数据处理和实时数据处理,其中:离线数据处理也叫离线计算、批量计算或批处理计算,是指先将数据抽取、存放到本地存储,数据一经抽取,就是静态不变的,再进行后续的加工、分析。在计算开始前已知所有输入数据,输入数据不会产生变化,且在解决一个问题后就要立即得出结果的前提下进行的计算;

[0079] 实时数据处理又称实时计算,是指将数据的生产看作是一个持续的动态的数据流,预先定义好加工规则,当数据流过时按预先定义的规则进行加工分析方向;一般都是针对海量数据进行的,要求数据处理的效率为秒级;实时计算主要分为数据的实时入库和数据的实时计算两部分;实时计算应用在数据源是实时的不间断的,要求用户的响应时间也是实时的(比如对于大型网站的流式数据:网站的访问PV/UV、用户访问了什么内容、搜索了什么内容等,实时的数据计算和分析可以动态实时地刷新用户访问数据,展示网站实时流量的变化情况,分析每天各小时的流量和用户分布情况),数据量大且无法或没必要预算,但要求对用户的响应时间是实时的。

[0080] 本申请实施例在进行数据关联操作的过程中,可以基于大数据处理技术中的实时数据处理的技术,对实时获取的基础内容交互数据进行数据处理,以提升对基础内容交互数据进行处理的及时性。

[0081] 下面对本申请的设计思想进行说明。

[0082] 在自媒体时代,账户可以将自己创作的内容随时随地发布到内容推送系统,进而内容推送系统可以将上述内容推送给其他账户,其他账户可以对获得的内容进行交互操作

产生内容交互数据，并将内容交互数据上报给内容推送系统；内容推送系统可以通过Spark等数据分析工具，对接收到的内容交互数据进行多层的离线数据处理，最终将离线数据处理得到的中间数据，存储到Mysql或者ES等数据仓库中；进而账户可以基于上述中间数据查询与其创作的内容相关的数据；但上述过程中常对海量的内容交互数据延后进行离线处理，导致内容交互数据处理的及时性低，无法及时的基于上述中间数据得知其他账户对内容的反馈信息，则无法及时的基于反馈信息发现异常的内容(如可以但不局限于包括一些涉及敏感话题等的内容)；另外目前上述中间数据的处理维度低，基于中间数据查询内容相关的数据时，需要对中间数据进行较为复杂的统计分析处理后，才能得到查询数据，导致数据查询耗时长、效率低。

[0083] 鉴于此，发明人设计了一种基于内容推送的数据处理方法、装置、设备及存储介质，用于提升对账户针对内容的内容交互数据进行处理的实时性和处理维度，以便提升账户基于处理后的内容交互数据，获取其他账户对内容的反馈信息的效率和实时性；该方法中考虑到为降低对内容交互数据进行离线数据处理的时延，按照交互操作的操作类型，实时对获取的基础内容交互数据进行拆分，将其推送到对应的消息队列(Message Queue，MQ)中；并基于一个或多个关联信息集合，对各个消息队列中各个时间窗口内获得的大量的基础内容交互数据，转换为少量的聚合数据，以提升数据处理的维度，降低得到的聚合数据的数量，进而可以降低查询相关数据时需要进行统计分析的聚合数据的数量，降低数据查询的时延和提升数据查询的效率。

[0084] 作为一种实施例，为了进一步提升账户查询推送内容相关的内容交互数据的效率，本申请实施例中还可以将同一个推送内容关联的聚合数据存储到同一个磁盘分片中，在查询推送内的相关数据时，可以直接从对应的磁盘分盘中获取聚合数据进行统计分析，降低查询聚合数据的时间，进而可以提升基于聚合数据返回查询数据的时间。

[0085] 为了更清楚地理解本申请的设计思路，以下对本申请实施例中的应用场景进行示例介绍。

[0086] 请参照图1，表示一种基于内容推送的数据处理的应用场景，该场景中包括终端设备110、内容推送服务器120和数据处理服务器130；终端设备110、内容推送服务器120和数据处理服务器130之间可以通过网络进行通信，其中：

[0087] 终端设备110用于接收内容发布账户上传的推送内容，并将推送内容发送给内容推送服务器120；终端设备110也可以接收上述内容推送服务器120分发的推送内容，终端设备110还可以接收目标账户针对推送内容触发的交互操作。

[0088] 作为一种实施例，终端设备110上可以安装内容推送系统的客户端，终端设备110可以通过上述客户端向内容推送服务器120上传推送内容，以及通过上述客户端接收内容推送服务器120分发的推送内容，或接收目标账户针对推送内容触发的交互操作等。

[0089] 内容推送服务器120用于接收内容发布账户通过终端设备110上传的推送内容，以及将推送内容分发给一个或多个目标账户使用的终端设备110。

[0090] 数据处理服务器130用于响应于各个目标账户针对各自获得的推送内容触发的交互操作，获得基础内容交互数据，基于上述交互操作的操作类型，将上述基础内容交互数据实时推送到上述操作类型关联的至少一个消息队列；以及基于至少一个关联信息集合，对各个消息队列接收到的基础内容交互数据进行关联信息集合关联操作，得到相应的聚合数

据。

[0091] 本申请实施例涉及的消息队列可以但不局限于是一个存放数据(或消息)的容器，即存放要传输的基础内容交互数据的队列;消息队列是一种异步的服务间通信方式，是分布式系统中重要的组件，可以用于发布和订阅消息，主要解决应用耦合，异步消息、流量削峰等问题，实现高性能、高可用、可伸缩和最终一致性架构;本申请实施例中涉及的消息队列可以但不局限于包括RocketMQ、RabbitMQ、Kafka等中的至少一种消息队列。

[0092] 本申请实施例中的终端设备100可以是移动终端、固定终端或便携式终端，例如移动手机、站点、单元、设备、多媒体计算机、多媒体平板、互联网节点、通信器、台式计算机、膝上型计算机、笔记本计算机、上网本计算机、平板计算机、个人通信系统(PCS)设备、个人导航设备、个人数字助理(PDA)、音频或视频播放器、数码相机或摄像机、定位设备、电视接收器、无线电广播接收器、电子书设备、游戏设备或者其任意组合，包括这些设备的配件和外设或者其任意组合。

[0093] 本申请实施例中的内容推送服务器120和数据处理服务器130可以是相同的服务器，也可以是不同的服务器;且内容推送服务器120和数据处理服务器130可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是云服务技术中提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN、以及大数据和人工智能平台等基础云计算服务的多个云服务器(如内容推送服务器120可以但不局限于包括图中示意出的服务器120-1、服务器120-2或服务器120-3;如数据处理服务器130可以但不局限于包括图中示意出的服务器130-1、服务器130-2或服务器130-3);上述内容推送服务器120的功能可以由一个或多个云服务器实现，还可以由一个或多个云服务器集群实现等;上述数据处理服务器130的功能可以由一个或多个云服务器实现，还可以由一个或多个云服务器集群实现等。

[0094] 其中，云服务技术(Cloud technology)是基于云计算商业模式应用的网络技术、信息技术、整合技术、管理平台技术、应用技术等的总称，可以组成资源池，按需所用，灵活便利。云服务技术为重要支撑;技术网络系统的后台服务需要大量的计算、存储资源，如视频网站、图片类网站和更多的门户网站。伴随着互联网行业的高度发展和应用，将来每个物品都有可能存在自己的识别标志，都需要传输到后台系统进行逻辑处理，不同程度级别的数据将会分开处理，各类行业数据皆需要强大的系统后盾支撑，只能通过云服务技术来实现。

[0095] 作为一种实施例，上述数据处理服务器130可以但不局限于是数据流处理引擎中的服务器，其中数据流处理引擎对流式数据(即数据流)进行处理的数据处理引擎，可以但不局限于包括Apache、Flink、Storm、Samza等;其中Flink是一个针对无界和有界数据流进行有状态计算处理的框架，是一款比较高效的数据流处理引擎，其认为一切的数据都是以流的形态存在的。

[0096] 作为一种实施例，本申请实施例中数据处理服务器130获取相应的聚合数据后，还可以但不局限于存储获取的聚合数据，以便查询账户可以针对推送内容进行数据查询，以基于推送内容关联的聚合数据，获取与目标业务需求关联的查询数据，其中，查询数据可以但不局限于是按照目标业务需求关联的数据处理规则，对获取的聚合数据进行数据处理，上述目标业务需求可以但不局限于是一些业务指标，业务指标可以但不局限于包括目标账

户浏览推送内容的浏览率、目标账户分享推送内容的概率、推送内容的热度排名、推送内容的PV/UV等至少一个数据指标。

[0097] 作为一种实施例，为了提高针对推送内容的数据查询的效率，本申请实施例中还可以通过数据分析系统实现数据查询，具体地，可以将获得聚合数据推送到数据分析系统中的存储空间中进行存储，由数据分析系统响应数据查询请求，并按照目标业务需求关联的数据处理规则，对相应的聚合数据进行数据处理，获得查询数据，以及向查询账户返回获得的查询数据等；上述数据分析系统可以但不局限于包括Druid、列式存储数据库ClickHouse或Zookeeper中，其中：

[0098] 上述Druid是一个分布式的、支持实时多维联合分析处理(On-Line Analytical Processing，OLAP)分析的数据处理系统；它既支持高速的数据实时摄入处理，也支持实时且灵活的多维数据分析查询，因此本申请实施例中使用Druid对数据量大的聚合数据进行灵活快速的多维OLAP分析，以提升账户查询推送内容关联的数据的效率。

[0099] 上述ClickHouse是一个MPP架构的面向列的数据库管理系统(Database Management System，DBMS)，用于OLAP分析且使用列式存储的数据库，数据按列进行组织，属于同一列的数据会被保存在一起，列与列之间也会由不同的文件分别保存，本申请实施例中使用ClickHouse存储上述聚合数据，具有如下优点：可以在不重启服务的情况下动态地创建、修改或删除聚合数据的数据库、表和视图，可以动态查询、插入、修改或删除上述聚合数据，可以按照用户粒度设置存储聚合数据的数据库或者表的操作权限，保障聚合数据的安全性，可以灵活的对聚合数据进行备份的导入和导出，且ClickHouse中提供集群模式，能够自动管理多个聚合数据的数据库节点；进而提升账户查询推送内容关联的聚合数据的灵活度。

[0100] 下面对本申请实施例的技术方案进行进一步的介绍。需要说明的是，以下介绍的技术方案只是示例性的。

[0101] 在介绍详细的技术方案之前，首先对本申请实施例涉及的基础内容交互数据进行说明：

[0102] 本申请实施例中涉及的基础内容交互数据为包含目标账户对推送内容的反馈信息的数据，即基础内容交互数据可以但不局限于为目标账户对获得的推送内容触发的交互操作产生的数据，该基础内容交互数据中可以但不局限于包括目标账户关联的账户标识、推送内容关联的内容标识和目标账户对推送内容进行的交互操作的操作信息等中的至少一个信息。

[0103] 本申请实施例中可以但不局限于将基础内容交互数据设置为n元组的形式，n为正整数，其中n元素中一个元素表征该基础内容交互数据的一个信息；如本申请实施例中可以将基础内容交互数据表示为三元组$\{X_1,X_2,X_3\}$、多元组$\{X_1,X_2,X_3,X_4\cdots\}$等形式，其中上述$X_1$可以但不局限于为触发交互操作的目标账户关联的账户标识，X_2可以但不局限于为上述交互操作针对的推送内容关联的内容标识，X_3可以但不局限于为交互操作的操作信息(如可以但不局限于包括交互操作的操作类型、操作名称等)，X_4可以但不局限于是上述交互操作的时间、地点或方式等信息；也可以将基础内容交互数据设置为“账户A1对推送内容A2进行了交互操作A4”的形式等；本申请实施例中对基础内容交互数据的具体形式不做限定，本领域的技术人员可根据实际需求灵活设置。

[0104] 其次,对本申请实施例中针对推送内容的交互操作进行说明。

[0105] 本申请实施例中的交互操作指目标账户针对接收的推送内容进行的操作,本申请实施例中可以但不局限于按照目标账户对推送内容的喜好程度,对交互操作进行分类,这种情况下交互操作的操作类型可以包括正反馈操作、负反馈操作等,交互操作的操作类型还可以包括不展示目标账户对推送内容的喜好程度的类型,如针对推送内容的曝光操作和互动操作等中的至少一个操作类型,其中:

[0106] 上述曝光操作可以但不局限于是对推送内容进行曝光的操作。

[0107] 上述正反馈操作指目标账户针对接收的推送内容进行正向反馈的操作,本申请实施例中正反馈操作可以但不局限于包括点击推送内容的点击操作、查看推送内容全文的全文查看操作、关注发布推送内容的账户的关注操作、针对推送内容的点赞操作、针对推送内容转移电子货币的电子货币转移操作、针对推送内容转移多媒体资源的多媒体资源转移操作、目标账户将推送内容保存至该目标账户的收藏文件夹的收藏操作、下载推送内容的下载操作、转发推送内容的网络链接的分享操作等中的至少一个操作;上述电子货币是指以电子形式存储在账户所持有的电子钱包(如支付类应用中的钱包或银行类应用中的钱包等)中的货币,电子货币可以但不局限于包括电子票据、数字货币(一种不受管制的、数字化的货币)、游戏资源(如游戏币、游戏装备)等等;上述多媒体资源可以但不局限于包括动态表情符、游戏资源、电子形式的爱心表情符等。

[0108] 上述负反馈操作指目标账户针对接收的推送内容进行负向反馈的操作,本申请实施例中负反馈操作可以但不局限于包括屏蔽推送内容的屏蔽操作、取消对发布推送内容的账户的关注的取消关注操作、查看推送内容的时长小于预设时长的负向阅读操作等中的至少一个操作。

[0109] 上述互动操作指目标账户与推送内容进行互动的操作,或目标账户与发布推送内容的账户进行互动的操作,本申请实施例中互动操作可以但不局限于对推送内容发表评论的评论操作等;若推送内容为直播媒体流,则互动操作还可以是目标账户在发布直播媒体流的直播间中与其他账户进行聊天的聊天操作等。

[0110] 作为一种实施例,本申请实施例中还可以直接基于交互操作的具体操作内容,对交互操作进行分类,如将上述曝光操作、点击操作、电子货币转移操作、关注操作、多媒体资源转移操作、收藏操作、下载操作、分享操作、屏蔽操作、负向阅读操作、评论操作等分别视为一个操作类型等。

[0111] 基于图1的应用场景,下面对本申请实施例中涉及的一种用于即时通讯的应用的启动方法进行示例说明;请参照图2,表示本申请实施例涉及的一种基于内容推送的数据处理方法的示意图,具体包括如下步骤:

[0112] 步骤S201,响应于各个目标账户针对各自获得的推送内容触发的交互操作,获得基础内容交互数据。

[0113] 作为一种实施例,本步骤中可以基于一个原始消息队列,接收各个目标账户触发的交互操作指示的基础内容交互数据,上述原始消息队列可以但不局限于包括RocketMQ、RabbitMQ、Kafka等中的至少一种消息队列。

[0114] 步骤S202,基于上述交互操作的操作类型,将上述基础内容交互数据实时推送到上述操作类型关联的至少一个消息队列。

[0115] 具体地，本步骤中一个消息队列可以关联一个操作类型，也可以关联多个操作类型；为了提升对基础内容交互数据处理的准确度，本申请实施例中可以针对每个操作类型，创建一个消息队列作为其关联的消息队列，且一个消息队列接收一个操作类型的交互操作产生的基础内容交互数据。

[0116] 步骤S203，针对上述至少一个消息队列中的各个消息队列，分别按照以下方式执行数据关联操作，获取相应的聚合数据：基于至少一个关联信息集合，对一个消息队列进行关联信息集合关联操作，得到相应的聚合数据；其中，一次关联信息集合关联操作包括：基于一个关联信息集合，将上述一个消息队列在预设时间窗口内获得的基础内容交互数据，转换为聚合数据，其中，转换得到的聚合数据的数量不大于上述在预设时间窗口内接收的基础内容交互数据的数量。

[0117] 本申请实施例中涉及的各个预设时间窗口，可以但不局限于是划分时间段的参考时间段长度，即步骤203中可以以预设时间窗口为参考时间段长度对时间进行划分，得到各个时间段；进而针对一个消息队列在各个时间段获得的基础内容交互数据，分别转换为对应的聚合数据；其中，对上述预设时间窗口不做限定，本领域的技术人员可根据实际需求设置，如可以但不局限于将预设时间窗口设置为1分钟、5分钟或10分钟等。

[0118] 作为一种实施例，为了提升对基础内容交互数据进行聚合处理的准确度和维度，上述关联信息集合中可以包括账户画像集合和内容信息集合中的至少一种信息，本申请实施例中可以但不局限于基于账户画像集合，对各个消息队列中的基础内容交互数据进行账户维度关联，得到表征各个目标账户分别针对那些推送内容进行交互操作的信息的第一聚合数据；也可以基于内容信息集合，对各个消息队列中的基础内容交互数据进行内容维度关联，得到表征与各个推送内容进行交互操作的具体账户的信息的第二聚合数据；其中本步骤中可以对各个消息队列中的基础内容交互数据进行单独的账户维度关联或内容维度关联，也可以对各个消息队列中的基础内容交互数据同时进行账户维度关联和内容维度关联。

[0119] 本申请的以下内容中，对上述账户维度关联和内容维度关联作进一步说明。

[0120] 首先，对本申请实施例涉及的账户画像集合和内容信息集合作说明。

[0121] 作为一种实施例，为了提升账户维度关联后的数据的信息丰富度，本申请实施例中账户画像集合中包括在内容推送系统中已注册的各个目标账户的账户画像数据，本申请实施例中为了提升获取各个目标账户的账户画像数据的效率，可以将账户画像数据与目标账户的账户标识进行关联，进而可以基于目标账户的账户标识，快速获取目标账户的账户画像数据。

[0122] 本申请实施例中涉及的账户画像数据又称为账户画像或用户画像(User Profile)，其指将账户关联的用户的信息进行标签化，目前，主要通过目标账户与推送内容的交互操作(比如曝光操作、点击操作、点赞操作、评论操作等)抽取目标账户的账户画像。账户画像沉淀在推送内容的标签上，包括静态信息和动态信息。其中，静态信息可以为目标账户首次注册时提供的，比如性别、年龄、常住地、籍贯、身高、学历、婚恋状态、受教育程度、资产情况、收入情况、职业等人口属性信息和社会属性；动态信息可以从目标账户的行为数据中挖掘，包括通过内容日志或第三方数据获取目标账户关联的摄影、运动、美食、美容、服饰、旅游、教育等兴趣，以及包括账户心理(即使用目标账户的用户的心理)、动机、价值观、

生活态度、个性等意识认知。

[0123] 作为一种实施例，为了提升内容维度关联后的数据的信息丰富度，本申请实施例中内容信息集合中包括在内容推送系统中已发布的各个推送内容的内容信息(又称为内容元信息)；本申请实施例中推送内容的内容元信息可以但不局限于是描述推送内容一些特征及属性的数据，内容元信息可以但不局限于包括推送内容的文件大小、封面图链接、内容标题、内容格式、发布时间、发布推送内容的账户的账户信息、推送内容中的图片信息(如可以但不局限于包括图片大小、图片格式、图片创造者)、表征推送内容是否是原创的原创标记、表征推送内容是否是首次发布的首发标记、推送内容的分类信息等中的至少一个信息。

[0124] 作为一种实施例，上述推送内容为视频时，上述内容元信息还可以包括视频的封面图的链接、视频的文件格式、视频的播放时长、视频的码率等中的至少一个信息。

[0125] 其中，上述分类信息可以是人工审核推送内容时，为推送内容划分的分类以及推送内容的标签信息，也可以是通过机器对推送内容进行自动审核时，为推送内容划分的分类；本申请实施例中可以基于推送内容的格式、发布来源、内容领域等划分推送内容的分类，如推送内容为文章时，可以但不局限于基于推送内容涉及的内容领域进行分类，且可以对推送内容进行多级分类，如一篇讲解某品牌某型号的手机的文章，其一级分类可以是科技，二级分类可以是智能手机，三级分类可以是国产手机等，标签信息为“某品牌，某型号手机”。

[0126] 作为一种实施例，为了提升访问上述账户画像集合和内容信息集合的效率，本申请实施例中可以将账户画像集合存储于一个键值对(key-value)数据库中，将内容信息集合存储于另一个键值对数据库中，以下内容中将存储账户画像信息集合的键值对数据库称为第一键值对数据库，将存储内容信息集合的键值对数据称为第二键值对数据库。

[0127] 本申请实施例中涉及的键值对数据库是一种非关系型数据库，使用简单的键值方法存储相关数据，本申请实施例中涉及的键值对数据库可以但不局限于包括Redis缓存，Redis是一个分布式缓存中间件；本申请实施例中使用键值对数据库存储上述账户画像集合和内容信息集合，一方面，可以支持账户画像数据和内容元信息的持久化缓存，可以将内存中的账户画像数据和内容元信息保存在磁盘中，重启的时候可以再次加载账户画像数据和内容元信息进行使用；另一方面，键值对数据库不仅仅支持简单的key-value的数据结构存储上述账户画像数据和内容元信息，还支持以List、Set、Zset、Hash等数据结构存储上述账户画像数据和内容元信息数据；另外，本申请实施例中键值对数据库还支持账户画像数据和内容元信息的备份，且支持快速的大量数据的访问，键值对数据库Redis是毫秒级的，访问Redis的速度基本是访问其他数据库如Hbase的速度的近1000倍，能明显加快访问账户画像集合和内容信息集合的效率。

[0128] 作为一种实施例，请参见图3，账户画像集合可以是维表的形式，为了提升从账户画像集合中获取的账户画像数据的准确度，本申请实施例中可以但不局限于基于第一周期，周期性的更新上述账户画像集合中各个目标账户的账户画像数据；其中对上述第一周期不做过多限定，本领域的技术人员可根据实际需求设置，如可以但不局限于将其设置为1天、7天、10天等。

[0129] 作为一种实施例，请继续参见图3，内容信息集合可以是维表的形式，内容推送系统接收到推送内容的内容信息后，可以将获得的内容信息实时记录在内容数据库(如可以

但不局限于包括Hbase数据库)中,本申请实施例中,为了提升从内容信息集合中获取内容信息的准确度,可以基于第二周期,对内容数据库进行周期性备份得到上述第二键值对数据库;本申请实施例中可以在访问记录内容信息的HBase数据库中的内容之前设置一层Redis缓存作为上述第二键值对数据库,当HBase数据库中的1000条内容信息推送到上述Redis缓存中时,由于1000条数据访问HBase是秒级的,而访问Redis是毫秒级的,访问Redis的速度基本是访问HBase的1000倍;同时为了防止过期的内容信息的数据浪费缓存,可以但不局限于将缓的第二周期设置成24小时或48小时等,同时可以通过监听写HBase Proxy来保证缓存的一致性。这样将访问时间从十几分钟变成了秒级;对上述第二周期不做过多限定,本领域的技术人员可根据实际需求获得。

[0130] 作为一种实施例,为了降低发生缓存穿透的风险,本申请实施例中,在实实时的将内容标识关联的推送内容的内容信息记录到内容数据库的过程中,可以检测异常内容标识,进而不在内容数据库中记录上述异常内容标识关联推送内容的内容信息,以便在对基础内容交互数据进行内容维度关联的过程中,直接过滤掉上述异常内容标识关联推送内容的内容信息;其中,异常内容标识可以但不局限于是因为内容安全或者推送内容的版权监管原因,已经从内容推送系统中删除的推送内的内容标识,如一些因为安全或者策略或者一致性为出现不存在的推送内容的内容标识;上述缓存穿透指所有缓存都无法命中,访问压力全部传到下一层存储的现象。

[0131] 作为一种实施例,为了降低周期性缓存上述第一键值对数据库和第二键值对数据库的过程中出现缓存雪崩的现象,可以在周期性缓存第一键值对数据库和第二键值对数据库的过程中,进行削峰填谷的操作,即对不同的键值对数据库,设置不同的周期性更新的时间周期,错开缓存的时间,以降低周期性更新上述第一键值对数据库和第二键值对数据库的过程中出现雪崩的可能性;上述缓存雪崩指所有缓存内容同时失效了,缓存某个时刻都都不起作用的现象。

[0132] 本申请实施例的以下内容,对上述账户维度关联和内容维度关联的具体过程进行说明。

[0133] (一)账户维度关联

[0134] 该关联方式中,可以基于账户画像集合,将各个消息队列在各个预设时间窗口内获得基础内容交互数据进行聚合,得到第一聚合数据;具体可以参见图4,可以但不局限于针对至少一个消息队列中的各个消息队列,执行如下操作:

[0135] 步骤S401,将一个消息队列在第一预设时间窗口内容获得的基础内容交互数据,确定为第一交互数据集合。

[0136] 上述第一预设时间窗口为账户维度关联的过程中,划分时间段的参考时间段长度,对第一预设时间窗口的具体数据不做限定,本领域的技术人员可根据实际需求设置,如可以但不局限于将第一预设时间窗口设置为1分钟、5分钟或8分钟等。

[0137] 本步骤中,针对一个消息队列而言,将在各个第一预设时间窗口内获得的基础内容交互数据,分别确定为一个第一交互数据集合,如若第一预设时间窗口为1分钟,则将该消息队列在每分钟内获得的基础内容交互数据,分别确定为一个第一交互数据集合。

[0138] 作为一种实施例,考虑到不同消息队列与交互操作的不同操作类型关联,在某些业务需求上,对不同操作类型关联的基础内容交互数据的聚合维度的要求可能是不同的,

因此为了提升对不同消息队列获得的基础内容交互数据进行账户维度关联的灵活度，本申请实施例中可以针对不同的消息队列设置不同的第一预设时间窗口。

[0139] 步骤S402，针对上述第一交互数据集合中基础内容交互数据包含的各个账户标识，获得与各个账户标识关联的第一聚合数据。

[0140] 具体地，第一交互数据集合中可能包含多个基础内容交互数据，不同的基础内容交互数据中包含的账户标识可能是不同的，因此一个第一交互数据集合中基础内容交互数据可能包含一个或多个账户标识，进而针对一个第一交互数据集合而言，分别将包含同一个账户标识的各个基础内容交互数据进行聚合处理，得到一个与上述同一个账户标识关联的第一聚合数据。

[0141] 作为一种实施例，可以但不局限于针对上述各个账户标识，分别执行如下步骤S4021和S4022的操作：

[0142] 步骤S4021，基于上述账户画像集合中记录的各个目标账户的账户画像数据，获取一个账户标识关联的目标账户的账户画像数据。

[0143] 具体地，账户画像集合中的账户画像数据可以是与账户标识关联后的，因此本步骤中可以直接基于一个账户标识，获得该账户标识关联的目标账户的账户画像数据。

[0144] 步骤S4022，通过获取的账户画像数据，将上述第一交互数据集合中包含上述一个账户标识的基础内容交互数据进行聚合，得到一个第一聚合数据。

[0145] 作为一种实施例，为了提升获得的第一聚合数据中的信息的维度，本申请实施例中，可以将获得的账户画像数据和上述第一交互数据集合中包含上述一个账户标识的基础内容交互数据进行聚合，具体地，可以但不局限于通过如下方式得到第一聚合数据：

[0146] 确定上述第一交互数据集合中包含上述一个账户标识的基础内容交互数据；获取确定的基础内容交互数据中包含的内容标识，生成内容标识集合；将获取的账户画像数据、上述一个账户标识和上述内容标识集合进行关联，获得相应的一个第一聚合数据。

[0147] 本申请实施例中对第一聚合数据的具体表示形式不做限定，本领域的技术人员可根据实际需求设置，如可以将第一聚合数据表示为{账户标识，账户画像数据，内容标识集合}三元组，也可以通过数据表的形式标识获得的各个第一聚合数据；

[0148] 此处为了便于理解进行账户维度关联前后相关数据的数据量的差异，请参见表1，给出一个第一交互数据集合包含的基础内容交互数据的示例，表1中每行数据为一个基础内容交互数据，请参见表2，给出将表1的基础内容交互数据聚合成第一聚合数据后的示意，表2中每行数据为一个第一聚合数据。

[0149] 表1：一个第一交互数据集合包含的基础内容交互数据

[0150]

账户标识	内容标识
账户标识1	内容标识1
账户标识1	内容标识3
账户标识1	内容标识5
账户标识2	内容标识2
账户标识2	内容标识3
账户标识2	内容标识4
账户标识2	内容标识6

账户标识3	内容标识1
账户标识3	内容标识2
…	…

[0151] 表2:将表1中的基础内容交互数据进行账户维度关联,确定的第一聚合数据

[0152]

账户标识	账户画像数据	内容标识集合
账户标识1	账户画像数据1	内容标识1、内容标识3、内容标识5
账户标识2	账户画像数据2	内容标识2、内容标识3、内容标识4、内容标识6
账户标识3	账户画像数据3	内容标识1、内容标识2
…	…	…

[0153] 表1中进行账户维度关联前,上述第一交互数据集合中共有9个基础内容交互数据,进行账户维度关联后的第一聚合数据只有3条,表2中第一聚合数据的数量明显比表1中基础内容交互数据的数量少,且考虑到实际应用过程中,推送到各个消息队列的基础内容交互数据是海量的,因此采用本申请实施例提供的方法进行上述账户维度关联后,包含目标账户对推送内容的反馈信息的数据的数据量会明显减少,可以明显减轻数据的存储资源占用,且能够减少后期数据查询时,访问相关的第一聚合数据的数量,以提升数据查询的效率和减少数据查询的时延。

[0154] (二)内容维度关联

[0155] 该关联方式中,可以基于内容信息集合,将各个消息队列在各个预设时间窗口内获得基础内容交互数据进行聚合,得到第二聚合数据;具体可以参见图5,可以针对至少一个消息队列中的各个消息队列,执行如下操作:

[0156] 步骤S501,将一个消息队列在第二预设时间窗口内容获得的基础内容交互数据,确定为第二交互数据集合。

[0157] 上述第二预设时间窗口为账户维度关联的过程中,划分时间段的参考时间段长度,对第二预设时间窗口的具体数据不做限定,本领域的技术人员可根据实际需求设置,如可以但不局限于将第二预设时间窗口设置为1分钟、5分钟或8分钟等。

[0158] 本步骤中,针对一个消息队列而言,将在各个第二预设时间窗口内获得的基础内容交互数据,分别确定为一个第二交互数据集合,如若第二预设时间窗口为5分钟,则将该消息队列在每5分钟内获得的基础内容交互数据,分别确定为一个第二交互数据集合。

[0159] 作为一种实施例,考虑到不同消息队列与交互操作的不同操作类型关联,在某些业务需求上,对不同操作类型关联的基础内容交互数据的聚合维度的要求可能是不同的,因此为了提升对不同消息队列获得的基础内容交互数据进行内容维度关联的灵活度,本申请实施例中可以针对不同的消息队列设置不同的第二预设时间窗口。

[0160] 步骤S502,针对上述第二交互数据集合中基础内容交互数据包含的各个内容标识,获得与各个内容标识关联的第二聚合数据。

[0161] 具体地,第一交互数据集合中可能包含多个基础内容交互数据,不同的基础内容交互数据中包含的内容标识可能是不同的,因此一个第二交互数据集合中基础内容交互数据可能包含一个或多个内容标识,进而针对一个第二交互数据集合而言,分别将包含同一个内容标识的各个基础内容交互数据进行聚合处理,得到一个与上述同一个内容标识关联的第二聚合数据。

[0162] 作为一种实施例，可以但不局限于针对上述各个账户标识，分别执行如下步骤S5021和S5022的操作：

[0163] 步骤S5021，基于内容信息集合中记录的各个推送内容的内容信息，获取一个内容标识关联的推送内容的内容信息。

[0164] 具体地，内容信息集合中的内容信息可以是与内容标识关联后的，因此本步骤中可以直接基于一个内容标识，获得该内容标识关联的推送内容的内容信息。

[0165] 步骤S5022，通过获取的内容信息，将上述第二交互数据集合中包含上述一个内容标识的基础内容交互数据进行聚合，得到一个第二聚合数据。

[0166] 作为一种实施例，为了提升获得的第二聚合数据中的信息的维度，本申请实施例中，可以将获得的内容信息和上述第一交互数据集合中包含上述一个内容标识的基础内容交互数据进行聚合，具体地，可以但不局限于通过如下方式得到第二聚合数据：

[0167] 确定上述第二交互数据集合中包含上述一个内容标识的基础内容交互数据；获取确定的基础内容交互数据中包含的账户标识，生成账户标识集合；将获取的内容信息、上述一个内容标识和上述账户标识集合进行关联，获得相应的一个第二聚合数据。

[0168] 本申请实施例中对第二聚合数据的具体表示形式不做限定，本领域的技术人员可根据实际需求设置，如可以将第二聚合数据表示为{内容标识，内容信息，账户标识集合}三元组，也可以通过数据表的形式标识获得的各个第二聚合数据；

[0169] 此处为了便于理解进行内容维度关联前后相关数据的数据量的差异，请参见表3，给出一个第二交互数据集合包含的基础内容交互数据的示例，表3中每行数据为一个基础内容交互数据，请参见表4，给出将表1的基础内容交互数据聚合成第一聚合数据后的示意，表4中每行数据为一个第一聚合数据。

[0170] 表3：一个第二交互数据集合包含的基础内容交互数据

[0171]

账户标识	内容标识
账户标识1	内容标识3
账户标识2	内容标识2
账户标识2	内容标识3
账户标识3	内容标识1
账户标识3	内容标识2
账户标识4	内容标识1
账户标识5	内容标识2
账户标识5	内容标识3
账户标识6	内容标识3
…	…

[0172] 表4：将表1中的基础内容交互数据进行内容维度关联，确定的第二聚合数据

[0173]

内容标识	内容信息	账户标识集合
内容标识1	内容信息1	账户标识3、账户标识4
内容标识2	内容信息2	账户标识2、账户标识3、账户标识5
内容标识3	内容信息3	账户标识1、账户标识2、账户标识5、账户标识6
…	…	…

[0174] 表3中进行内容维度关联前，上述第二交互数据集合中共有9个基础内容交互数据，进行内容维度关联后的第二聚合数据只有3条，表4中第二聚合数据的数量明显比表3中基础内容交互数据的数量少，且考虑到实际应用过程中，推送到各个消息队列的基础内容交互数据是海量的，因此采用本申请实施例提供的方法进行上述内容维度关联后，包含目标账户对推送内容的反馈信息的数据的数据量会明显减少，可以明显减轻数据的存储资源占用，且能够减少后期数据查询时，访问相关的第二聚合数据的数量，以提升数据查询的效率和减少数据查询的时延。

[0175] 作为一种实施例，在获得聚合数据之后，对聚合数据进行存储的过程中，随着推送内容的数据量的增长，信息流海量内容分发的推送内容的基础内容交互数据的数据量很大，基于基础内容交互数据获得的聚合数据的数据量每条可能达到百亿级的数据，将海量的聚合数据直接写入数据分析系统如Clickhouse的话，则会导致Zookeeper集群的QPS过高，因此本申请实施例中，可以采用的是Batch批量方式，将聚合数据写入存储空间；同时为了缓解Zookeeper集群的压力，本申请实施例中可以选用大小为几十万(如可以但不局限于包括三十万)的Batch，将聚合数据写入存储空间。

[0176] 一般情况下，在存储数据的时候，常将数据写一张分布式表，这就会造成单台机器出现磁盘的瓶颈，如将数据写入Clickhouse时，由于Clickhouse底层运用的是Mergetree，原理类似于HBase的底层LSM-Tree，在合并的过程中会存在写放大的问题，加重磁盘压力；峰值每分钟几千万条数据，写完耗时几十秒，如果正在做Merge，就会阻塞写入请求，查询也会非常慢；鉴于此，本申请实施例中在存储聚合数据之前，可以对上述数据处理引擎或数据分析系统中的存储空间进行磁盘Raid，得到多个磁盘分片，提升磁盘的IO；同时可以在写入聚合数据之前，可以但不局限于通过Hash的方式，对聚合数据进行分表，将大量的聚合数据写入到不同的磁盘分片中；另外，考虑到通过上述方式，将聚合数据写入到分布式系统的存储空间中，可能出现就是局部的最高值Top并非全局Top的问题，尤其是统计监控全局结果，比如同一个推送内容关联的聚合数据分别存储到了不同的存储分片上，在计算阅读量的全局前100(Top100)的推送内容时，有一个推送内容在磁盘分片1上是Top100，但是在其它磁盘分片上不是Top100，导致数据汇总的时候，会丢失一部分数据，影响最终的统计结果，鉴于此，为了提升针对推送内容的数据查询的准确度，本申请实施例中，可以将一个推送内容关联的聚合数据存储到同一个磁盘分片中。

[0177] 鉴于上述内容，本申请实施例中可以但不局限于针对获得的各个聚合数据，分别按照以下方式执行数据存储操作，将上述各个聚合数据分别存储到对应的磁盘分片中：将一个聚合数据存储至上述一个聚合数据关联的推送内容映射的磁盘分片中，其中，一个磁盘分片用于存储一个推送内容关联的聚合数据，上述磁盘分片可以但不局限于是将上述数据处理引擎或数据分析系统中的存储空间进行磁盘Raid得到的，聚合数据关联的推送内容映射的磁盘分片可以但不局限于通过对聚合数据关联的推送内容的内容标识进行Hash路由获得。

[0178] 进而，查询账户通过数据查询请求查询与推送内容相关的数据时，查询账户可以触发针对待查询推送内容的数据查询请求，进而上述数据处理服务器130述数据处理引擎或数据分析系统上述可以响应于针对待查询推送内容的数据查询请求，从上述待查询推送内容映射的磁盘分片中，获取上述待查询推送内容关联的聚合数据；并按照目标业务需求

关联的数据处理规则，对获取的聚合数据进行数据处理，获得查询数据，并返回上述查询数据。

[0179] 作为一种实施例，请参见图6，在获得聚合数据之后，对聚合数据进行存储的过程中，为了保证记录的聚合数据的一致性，本申请实施例中还可以但不局限于通过Zookeeper集群实现高可用的方案，将聚合数据写入一个磁盘分片，仅写入一个副本，然后再写如Zookeeper集群，通过Zookeeper集群告诉同一个磁盘分片的其他副本，其他副本再过来拉取聚合数据，保证聚合数据的一致性；由于Zookeeper集群是轻量级的，而且写入聚合数据的时候，任意写一个副本，其它副本都能够通过Zookeeper集群获得一致的数据，另外就算其它节点第一次来获取数据失败了，后面只要发现它跟Zookeeper集群上记录的聚合数据不一致，就会再次尝试获取聚合数据，保证一致性。

[0180] 作为一种实施例，为了提升数据查询的效率和降低数据查询的时间，本申请实施例中数据查询请求中携带上述待查询推送内容的内容标识和查询时间段，进而还可以基于数据查询请求携带的信息，确定数据查询请求对应的数据查询索引信息，进而直接从数据索引信息中指示的各个存储地址中，获取对应的聚合数据进行处理，具体可参见图7，具体可以但不局限于包括如下步骤：

[0181] 步骤S701，基于数据查询请求携带的待查询推送内容的内容标识，将上述待查询推送内容映射的磁盘分片确定为目标磁盘。

[0182] 本申请实施例中，存储聚合数据的时候，已经对内容标识做过路由，一个内容标识关联的聚合数据只存在于一个磁盘分片上，因此该步骤按照同样的规则先对数据查询请求中携带的内容标识进行Hash路由，确定待查询推送内容映射的磁盘分片为目标磁盘分片。

[0183] 步骤S702，通过预设时间粒度，将上述数据查询请求携带的查询时间段划分为至少一个子时间段。

[0184] 上述预设时间粒度可以但不局限于是一个预设的时间长度，本申请实施例中对预设时间粒度的具体数值不做限定，本领域的技术人员可根据实际的业务需求设置，如可以但不局限于将上述预设时间粒度设置为1分钟、5分钟或10分钟等。

[0185] 应当说明的是，上述步骤S701和步骤S702并无固定的执行顺序。

[0186] 步骤S703，根据上述待查询推送内容关联的聚合数据中，各个子时间段映射的聚合数据在上述目标磁盘分片中的存储地址，确定上述数据查询请求对应的数据查询索引信息。

[0187] 具体地，可以但不局限于将上述各个子时间段映射的聚合数据在上述目标磁盘分片中的存储地址确定为数据查询索引信息，或将各个子时间段的时间段信息和各个子时间段映射的聚合数据在目标磁盘分片中的存储地址，确定为上述数据查询索引信息，上述时间段信息中可以包括子时间段的时间范围，或者包括子时间段的时间范围和待查询推送内容的内容标识等信息；为便于理解，此处给出一个数据查询请求对应的数据查询索引信息的示例，请参见表5，该数据查询请求是针对待查询推送内容M触发的。

[0188] 表5：数据查询索引信息的示例

[0189]

Hash(内容标识+hash(M的日期M的子时间段1))	M的子时间段1的聚合数据的保存地址1
Hash(内容标识+hash(M的日期M的子时间段2))	M的子时间段2的聚合数据的保存地址2
…	…

[0190] 其中，表5中的内容标识可以但不局限于是待查询推送内容的内容ID；表中的聚合数据可以包括待查询推送内容关联的第一聚合数据和第二聚合数据中的至少一种聚合数据。

[0191] 步骤S704，基于上述数据索引信息，从上述各个子时间段映射的聚合数据在上述目标磁盘分片中的存储地址中，获取上述各个子时间段映射的聚合数据。

[0192] 具体地，可以参见图8，此处给出从目标磁盘分片中获取聚合数据的示例图，从图中可以看出，通常情况下查询单个内容ID(内容标识)相关的聚合数据，分布式表会将查询下发到所有的磁盘分片上，然后再返回查询结果进行汇总，而本申请实施例中，由于对内容ID做过路由，一个内容ID关联的聚合数据只存在于一个磁盘分片上，剩下的磁盘分片都在空跑而不会接收上述数据查询请求；针对这类数据查询请求，首先按照同样的规则对数据查询请求中携带的内容ID进行路由(一个分发寻址策略)，直接查询目标磁盘分片(如图中示意出的磁盘分片-2)，减少了N-1/N的负载，大量缩短数据查询时间；同时由于是提供的OLAP查询，数据满足最终一致性即可，通过主从副本读写分离，可以进一步提升性能，同时通过缓存针对同样结果缓存，可以明显提升对外服务的性能。

[0193] 作为一种实施例，在步骤S703中确定数据查询请求对应的数据查询索引信息之后，可以对确定的数据查询索引进行缓存预设时长，以便在确定数据查询索引之后的预设时长内，接收到相同的数据查询请求后，可以直接基于缓存的数据查询索引，获取对应的聚合数据；对上述预设时长不做限定，本领域的技术人员可根据实际需求设置，如将其设置为1分钟、5分钟、10分钟或1小时等。

[0194] 根据的本申请实施例涉及的基于内容推送的数据处理方法的场景，大部分数据查询请求都是和时间以及内容ID(即上述内容标识)相关的，如针对某个推送内容，查询发布后的N(上述N为整数)分钟内，接收到该推送内容的目标账户对该推送内容的反馈信息是怎么样的，本申请实施例中，可以按照日期，预设时间粒度和内容ID建立了数据查询索引，针对某个推送内容的查询，建立数据查询索引之后，可以减少近99％的文件扫描；另外，如果一些业务场景下，查询的数据量太大，维度太多，以视频作为推送内容而言，一天推送的视频有上百亿条，有些维度有几百个类别，如果一次性把所有维度进行预聚合，数据量会指数膨胀，查询反而变慢，并且会占用大量内存空间；解决方案针对不同的维度，建立对应的预聚合物化视图，用空间换时间，这样可以缩短查询的时间。

[0195] 请参见图9，本申请实施例还提供的一种内容推送系统的框架，该内容推送系统中包括：内容生产端和内容消费端、上下行内容接口服务器、内容数据库、调度中心服务、人工审核系统、排重系统、内容分发出口服务、数据运营系统、实时分发统计接入层、实时存储引擎层、实时计算层、实时数据聚合模型、实时数据监控及展示服务。

[0196] 如图9所示基于实时多维度聚合计算的内容分发监控方法和系统的流程图。在信息流海量内容分发的信息流场景下，当数据量巨大的情况下，一天上报的数据量达到万亿级的规模，要实现极低延迟的实时计算和秒级的多维实时查询和监控是本发明要处理的核心问题。高效的实时数据监控分析和聚合数据的高效处理，实时分发统计数据包括：曝光、PV/VV、评论、负面评论、举报、负反馈等；需要复核的内容通过数据运营系统推送到人工审核系统，并且获取人工审核系统对推送内容的审核结果，来确定后续将推送内容继续留在推荐池分发还是禁用或下架上述推送内容；其中：

[0197] 实时分发统计接入层可以用于对目标账户上传的基础内容交互数据和内容推送系统分发推送内容的统计进行监控分析，比如推送内容在内容消费端(又可以称为C侧)表现异常的统计数据，如针对推送内容的评论数据快速增长、PV/VV增速过快、转发推送内容的次数增长过快、对推送内容进行点赞的操作快速增长等；进而可以通过实时数据监控及展示服务统计监控到某一推送内容的统计信息满足异常条件后，调用送审接口，将推送内容推送到人工审核系统进行送人工复核，对于复核确认在内容消费端表现异常的推送内容，直接将该推送内筒下架；

[0198] 本申请实施例中对推送内容进行监控的统计数据可以但不局限于包括表6中的数据等，表6中的统计数据仅为示例性说明，本领域的技术人员可根据实际需求设置对推送内容进行监控的统计数据。

[0199] 表6:对推送内容进行监控的统计数据的示例

[0200]

PV/VV快速增长	评论快速增长	内容举报或负反馈
负面评论	高跳出	低时长
低阅读完成率	大盘top PV/VV	敏感类目(比如社会)top PV/VV
大盘top转化率	敏感类目top转化率	大盘top评论量
大盘top点赞量	大盘top Biu量	

[0201] 作为一种实施例，本申请实施例提供的基于内容推送的数据处理方法可以通过图9中的实时分发统计接入层、实时存储引擎、实时计算层和实时数据聚合模型实现，其中：

[0202] 实时分发统计接入层主要是将海量的基础内容交互数据，推送到对应的与操作类型关联的消息队列中，对于某一多媒体频道的推送内容(如视频等)来说，拆分过后，数据就只有百万级/s了；实时计算层和实时数据聚合模型主要负责，对各个消息队列获得的基础内容交互数据，实时基于账户画像集合进行账户维度关联，以及基于内容信息集合进行内容维度关联，将多行的基础内容交互数据转换为一行多列的聚合数据等；实时存储引擎主要是设计出符合目标业务需求的，下游好用的实时消息队列，本申请实施例中可以提供两个消息队列，作为实时存储引擎的两层。一层数据仓库中间(Data WareHouse Middle，DWM)层的消息队列，DWM层的消息队列中存储有对账户画像集合中的账户画像数据以及内容信息集合中的内容信息做轻度的聚合操作，生成一系列的中间表，DWM层用以提升公共指标的复用性，减少重复加工；另一层是数据仓库服务(Data WareHouse Servce，DWS)层的消息队列，DWS层存储账户维度关联得到的第一聚合数据以及内容维度关联得到的第二聚合数据的，其中的数据可以但不局限于包含内容标识、B侧数据和C侧数据；可以看到DWS层的流量进一步减小到十万级/s，内容标识聚合后的数据更是万级/s，并且格式更加清晰，维度信息更加丰富。最后可以通过上述实时数据监控和展示服务提供数据查询功能；其中B侧数据指与内容生产者创作的推送内容本身关联的数据或信息，其可以但不局限于包括推送内容的内容信息(又可以称为内容元信息)；C侧数据可以包括本申请实施例中涉及的基础内容交互数据。

[0203] 接下来对实时分发统计接入层、实时存储引擎、实时计算层、实时数据聚合模型实现和实时数据监控和展示服务做进一步说明：

[0204] 其中，实时分发统计接入层主要负责基础内容交互数据接入，实现每秒海量的基础内容交互数据的实时接入，并且进行极低延迟关联信息集合关联；同时实时分发统计接

入层与实时存储引擎交互支持高并发写入、高可用分布式和高性能的数据查询索引，这里是将数据将小时级别的延时减少到分钟级延迟的关键，实时分发统计接入层与实时计算层以及实时存储引擎的逻辑关系可参见图9；

[0205] 如图10所示，实时分发统计接入层的关键是将原始消息队列中的基础内容交互数据，分别推送到与产生基础内容交互数据的交互操作的操作类型关联的消息队列中，如图所示，将原始消息队列中的基础内容交互数据进行微队列拆分和微进程部署，进而可以加快对各个消息队列中的基础内容交互数据进行账户维度关联和内容维度关联的效率；实时分发统计接入层对外就是几个与上述交互操作的操作类型关联的消息队列，不同的消息队列里面存放的就是不同聚合粒度的基础内容交互数据，包括内容标识(如内容ID)、账户标识(如账户ID)、C侧数据、B侧数据和账户画像数据等；实时存储引擎存储的是上图实时计算层输出的聚合数据(可以但不局限于包括上述第一聚合数据和第二聚合数据中的至少一种数据)，将实时计算层输出的聚合数据保存到DWS层的消息队列中，以提供给下游多账户复用；如此使用本申请实施例提供的方法之前。需要对千万级/秒的原始消息队列获得的基础内容交互数据，进行复杂的数据清洗，然后再进行账户级别的关联和信息级别的关联，才能拿到符合要求格式的聚合数据，处理效率很低，但采用本申请实施例提供的方法后，可以基于内容标识直接申请DWS层的消息队列中存储的聚合数据，提升响应针对待查询推送内容的数据查询请求的速度。

[0206] 请继续参见图10，实时存储引擎需要有维度索引、支持高并发、预聚合、高性能实时多维OLAP查询；本申请实施例中，实时存储引擎可以实现分布式的高可用、水平扩展的功能要求，也可以将海量的聚合数据写入对应的磁盘分写，其中将聚合数据写入磁盘分片的过程可参见上述内容，此处不再重复叙述；实时存储引擎还可以高性能的查询，如构建上述数据查询索引，以及基于数据查询索引获得查询数据物化视图构建等，其中数据查询索引的有关内容可参见上述描述，此处不再重复叙述。

[0207] 作为一种实施例，本申请实施例中，实时存储引擎可以但不局限于分为实时写入层、OLAP存储层和后台接口层；其中，实时写入层主要是负责Hash路由将聚合数据写入对应的磁盘分片；OLAP存储层利用MPP存储引擎，设计符合业务的索引和视图，高效存储海量的聚合数据；后台接口层，用于直接查询和检索，作为数据服务接口访问数据使用，提供高效的多维实时查询接口。

[0208] 作为一种实施例，实时计算层和实时数据聚合模型用于对各个消息队列中的基础内容交互数据，进行账户维度关联和内容维度关联，具体地，可参见图11，给出了实时计算层和实时数据聚合模型的关系，以及进行账户维度关联和内容维度的主要流程的示意；如图所示，在实时计算层的过程中，可以基于实时的维表(上述账户画像集合或内容信息集合)，先按照预设时间窗口(即上述第一预设时间窗口或第二预设时间窗口)将各个消息队列获得的基础内容交互数据进行了窗口聚合，得到交互数据集合(即上述第一交互数据集合或第二交互数据集合)，将同一交互数据集合中的基础内容交互数据进行聚合，得到聚合数据(即上述第一聚合数据或第二聚合数据)，得到聚合数据的具体过程可参见上述描述，此处不再重复叙述。

[0209] 请继续参见图11，作为一种实施例，实时计算层和实时数据聚合模型用的处理中，使用了Redis缓存作为存储账户画像集合第一键值对数据库，以及使用Redis缓存作为存储

内容信息的第二键值对数据库，能够加快访问账户画像数据和内容信息的速度；同时通过监听写HBase Proxy来保证缓存的一致性，有关账户画像集合、内容信息集合、第一键值对数据库和第二键值对数据库的详细内容可参见上述描述，此处不再重复叙述。

[0210] 上述实时数据监控和展示服务用于响应上述数据查询请求，创建数据查询请求对应的数据查询索引信息，基于上述数据索引信息，获得对应的聚合数据，按照目标业务需求关联的数据处理规则，对获取的聚合数据进行数据处理，获得查询数据，并返回上述查询数据，具体过程可参见上述内容，此处不再重复叙述。

[0211] 下面对上述内容推送系统中的各个模块的功能进行介绍：

[0212] 1)内容生产端和内容消费端

[0213] PGC、UGC或者PUGC为多频道网络(Multi-Channel Network，MCN)内容生产者，通过移动终端或者后端接口API系统，提供本地编辑的或者web发布系统提供的图文内容或者视频内容等作为推送内容，视频内容包括短视频和小视频，是推荐分发内容的主要内容来源；上述MCN是一种多频道网络的产品形态，将PGC内容联合起来，在资本的有力支持下，保障内容的持续输出，从而最终实现商业的稳定变现。

[0214] 其中，内容生产端通过和上下行内容接口服务器的通讯，获取上下行内容接口服务器的接口地址，通过接口地址上传图文内容或者视频内容，图文内容来源通常是轻量级发布端和编辑内容入口，视频内容发布通常是图像采集设备，拍摄过程当中本地视频内容可以选择搭配的音乐，滤镜模板和视频的美化功能等。

[0215] 内容消费端和上下行内容接口服务器通讯，获取推送内容的索引信息，索引信息以Feeds流的方式展示。当内容消费端发送具体的图文内容或者视频内容请求消息时，内容消费端与内容分发出口服务通讯，获取索引信息中对应的图文内容或者视频内容。

[0216] 此外，内容消费端还可以将目标账户针对推送内容触发的交互操作确定的基础内容交互数据(如可以但不局限于包括目标账户对推送内容的评论、点赞、转发、收藏、浏览、跳出、播放、曝光等信息等信息)，实时上报给统计上报接口服务器用于统计分析，例如，卡顿，加载时间，播放点击等。

[0217] 2)上下行内容接口服务器和内容分发出口服务

[0218] 上下行内容接口服务器与内容生产端直接通讯，将内容生产端提交的推送内容的内容元信息存入内容数据库，以及将内容生产端提交的推送内容同步给调度中心服务器，调度中心服务进行的推送内容的处理和流转；其中，有关内容元信息的描述可参见上述内容，此处不再重复描述。

[0219] 内容分发出口服务将获取的推送内容以Feeds形式发送给内容消费端，内容分发出口服务通常是一组地域上就近部署下用户附件的接入服务。

[0220] 3)内容数据库

[0221] 内容数据库是推送内容的核心数据库，所有内容生产端发布的推送内容的内容源信息都保存在内容数据库中；即本申请实施例中的内容数据库可以保存内容生产端产生的推送内容的内容元信息，其中，有关内容元信息的描述可参见上述内容，此处不再重复描述。

[0222] 内容处理主要包括机器处理和人工审核处理，内容特征建模服务都需要从内容数据库当中获取推送内容的内容源信息依据不同的内容标记，内容数据库分为不同的内容

池。推荐排序服务、排重服务等都需要从内容数据库当中获取推送内容的内容信息，比如排重服务会依据业务需求加载过去一段时间(如一周)已经入库启用的推送内容，对于重复重新入库的推送内容将加上过滤标记不再提供给内容分发出口服务展示给用户；其中，针对图文类的推送内容，图文排重服务和优质内容优质识别服务都是机器处理过程，处理的结果保存在内容数据库当中。

[0223] 作为一种实施例，人工审核过程当中会可以从内容数据库中读取推送内容的内容元信息，同时人工审核的结果和状态也会回传进入内容数据库当中保存，人工审核结果也是后续衡量算法过滤模型效率的一个重要依据。

[0224] 4)调度中心服务

[0225] 调度中心服务负责推送内容流转的整个调度过程，控制上下行内容接口服务器接收上传的推送内容，以及从内容数据库中获取推送内容的内容元信息；调度排重服务对重复入库的推送内容进行标记和过滤；

[0226] 调度中心服务还可以对于机器无法处理的内容，比如政治敏感，安全问题内容需要人工审核的，调用人工审核系统进行人工审核的处理；最后通过人工审核系统的推送内容启用通过内容出口分发服务通常是推荐引擎或者搜索引擎或者运营直接的展示页面提供给终端的内容消费端；

[0227] 作为一种实施例，调度中心服务还可以和实时数据监控及展示服务通讯，获取端分发的实时统计和监控数据，用于调度策略的调整，比如消费数据后验好的类目数据优先进入审核调度队列头部等。

[0228] 5)人工审核系统

[0229] 人工审核系统是人工服务能力的载体，主要用于审核过滤政治敏感、色情、法律不允许等机器无法确定判断的内容，同时还对进行内容的标签标注和二次确认，审核的内容来自于自媒体应用发布的和从公共网络上获取的；人工审核的结果通过调度中心服务写入内容数据库中；由于图文类的推送内容本身完全通过机器学习比如深度学习还不完全成熟，所以通过在机器处理的机器上进行二次的人工审核处理，通过人机协作，提升图文本身标注的准确性和效率。

[0230] 作为一种实施例，人工审核系统还可以在接收调度中心同步的审核任务同时，也接收数据运营系统同步统计监控到数据变化异常的可疑的推送内容，对这部分推送内容复核通过后，直接下架或者继续分发等。

[0231] 6)排重服务

[0232] 排重服务提供图文类、视频和图集类的推送内容的排重服务，主要是对图文和图集及视频进行向量化，然后建立向量的索引，然后通过比较向量之间的距离来确定相似程度；具体地，排重服务和调度中心服务通讯，用于标题去重、封面图的图片去重、内容正文去重及视频指纹和音频指纹去重等。可采用simhash或BERT对正文向量、图片向量去重，对于视频内容可抽取视频指纹和音频指纹构建向量，然后计算向量之间的距离(比如欧式距离)来确定是否重复。具体的排重方法本申请实施例中不作赘述。

[0233] 7)实时分发统计接入层

[0234] 实时分发统计接入层和内容消费端通讯比如对内容的评论，点赞，转发，收藏，浏览，跳出，播放，曝光等信息通过实时统计接口服务上报；按照实时分发统计接入层的功能

和数据接入的策略，实现数据实时接入和预处理，其中时分发统计接入层的功能可参见上文描述。

[0235] 8)实时计算层和数据聚合模型

[0236] 实时计算层和数据聚合模型层的核心是处理实时关联信息集合与实时数仓的关系，按照上面描述详细策略和方案，处理接入层和数据存储引擎之间的关系，提升数据处理和计算的效率，以及降低数据处理和计算的资源消耗等。

[0237] 9)实时存储引擎层

[0238] 实时存储引擎层可以实现分布式-高可用的水平扩展，且能将实时计算层和数据聚合模型得到的聚合数据存储到对应的磁盘分片中，具体内容可参见上述描述，此处不再重复叙述。

[0239] 10)实时数据监控及展示服务

[0240] 实施数据监控及展示服务可以将获得的聚合数据进行存储，以及将聚合数据的计算结果服务化，提供实时数据展示及对外服务；以及响应于数据查询请求，创建对应的数据查询索引并返回查询结果，具体过程可参见上述内容，此处不再重复叙述。

[0241] 实时数据监控及展示服务也可以针对内容消费端针对推送内容的基础内容交互数据进行监控，可以但不局限于监控在C侧表现异常的推送内容的统计数据，统计数据的描述可参见上述内容，此处不再重复叙述。

[0242] 需要说明的是，上述应用场景仅是示例的，并不构成对本申请保护范围的限定。

[0243] 本申请实施例中，一方面，能够对实时获得的基础内容交互数据进行处理，提升了对基础内容交互数据进行处理的及时性；另一方面，本申请实施例中在查询与推送内容相关的数据的过程中，可以针对数据查询请求建立对应的数据查询索引，能够直接基于数据查询索引获取相关的聚合数据进行处理，降低了数据查询的时延，且本申请实施例中的聚合数据可以是进行用户维度关联和内容维度关联后的数据，在对获得的聚合数据处理的过程中，能够减少计算能力的消耗，进而提升基于聚合数据获得查询数据的效率；能够在第一时间发现在内容消费端表现异常的推送内容，对于实时数据分析场景来说，数据查询响应的速度有明显提升，返回查询数据的时延明显降低。

[0244] 请参照图12，基于同一发明构思，本申请实施例提供一种基于内容推送的数据处理装置1200，包括：

[0245] 数据采集单元1201，用于响应于各个目标账户针对各自获得的推送内容触发的交互操作，获得基础内容交互数据；

[0246] 数据拆分单元1202，用于基于上述交互操作的操作类型，将上述基础内容交互数据实时推送到上述操作类型关联的至少一个消息队列；

[0247] 数据聚合单元1203，用于针对上述至少一个消息队列中的各个消息队列，分别按照以下方式执行数据关联操作，获取相应的聚合数据：基于至少一个关联信息集合，对一个消息队列进行关联信息集合关联操作，得到相应的聚合数据；其中，一次关联信息集合关联操作包括：基于一个关联信息集合，将上述一个消息队列在预设时间窗口内获得的基础内容交互数据，转换为聚合数据，其中，转换得到的聚合数据的数量不大于上述在预设时间窗口内接收的基础内容交互数据的数量。

[0248] 作为一种实施例，每个基础内容交互数据包含触发上述交互操作的目标账户关联

的账户标识和触发上述交互操作的推送内容关联的内容标识，上述数据聚合单元1203具体用于执行以下操作中的任意一种或组合：

[0249] 若上述关联信息集合包括账户画像集合，则将上述一个消息队列在第一预设时间窗口内容获得的基础内容交互数据，确定为第一交互数据集合，针对上述第一交互数据集合中基础内容交互数据包含的各个账户标识，分别执行如下操作：基于上述账户画像集合中记录的各个目标账户的账户画像数据，获取一个账户标识关联的目标账户的账户画像数据，通过获取的账户画像数据，将上述第一交互数据集合中包含上述一个账户标识的基础内容交互数据进行聚合，得到一个第一聚合数据；

[0250] 若上述关联信息集合包括内容信息集合，则将上述一个消息队列在第二预设时间窗口内获得的基础内容交互数据，确定为第二交互数据集合，针对上述第二交互数据集合中基础内容交互数据包含的各个内容标识，分别执行如下操作：基于上述内容信息集合中记录的各个推送内容的内容信息，获取一个内容标识关联的推送内容的内容信息，通过获取的内容信息，将上述第二交互数据集合中包含上述一个内容标识的基础内容交互数据进行聚合，得到一个第二聚合数据。

[0251] 作为一种实施例，数据聚合单元1203具体用于：

[0252] 确定上述第一交互数据集合中包含上述一个账户标识的基础内容交互数据；

[0253] 获取确定的基础内容交互数据中包含的内容标识，生成内容标识集合；

[0254] 将获取的账户画像数据、上述一个账户标识和上述内容标识集合进行关联，获得相应的一个第一聚合数据。

[0255] 作为一种实施例，数据聚合单元1203具体用于：

[0256] 确定上述第二交互数据集合中包含上述一个内容标识的基础内容交互数据；

[0257] 获取确定的基础内容交互数据中包含的账户标识，生成账户标识集合；

[0258] 将获取的内容信息、上述一个内容标识和上述账户标识集合进行关联，获得相应的一个第二聚合数据。

[0259] 作为一种实施例，上述账户画像集合存储于第一键值对数据库，上述第一键值对数据库是基于第一周期周期性更新的；

[0260] 上述内容信息集合存储于第二键值对数据库，上述第二键值对数据库是基于第二周期对内容数据库进行周期性备份得到的，上述内容数据库用于实时记录上述推送内容的内容信息。

[0261] 作为一种实施例，数据聚合单元1203还用于：

[0262] 针对上述至少一个消息队列中的各个消息队列，分别按照以下方式执行数据关联操作，获取相应的聚合数据之后，针对获得的各个聚合数据，分别按照以下方式执行数据存储操作，将上述各个聚合数据分别存储到对应的磁盘分片中：将一个聚合数据存储至上述一个聚合数据关联的推送内容映射的磁盘分片中，其中，一个磁盘分片用于存储一个推送内容关联的聚合数据；以及

[0263] 响应于针对待查询推送内容的数据查询请求，从上述待查询推送内容映射的磁盘分片中，获取上述待查询推送内容关联的聚合数据；

[0264] 按照目标业务需求关联的数据处理规则，对获取的聚合数据进行数据处理，获得查询数据，并返回上述查询数据。

[0265] 作为一种实施例，上述数据查询请求中携带上述待查询推送内容的内容标识和查询时间段，上述数据聚合单元1203具体用于：

[0266] 基于上述内容标识，将上述待查询推送内容映射的磁盘分片确定为目标磁盘分片；以及

[0267] 通过预设时间粒度，将上述查询时间段划分为至少一个子时间段；

[0268] 根据上述待查询推送内容关联的聚合数据中，各个子时间段映射的聚合数据在上述目标磁盘分片中的存储地址，确定上述数据查询请求对应的数据查询索引信息；

[0269] 基于上述数据索引信息，从上述各个子时间段映射的聚合数据在上述目标磁盘分片中的存储地址中，获取上述各个子时间段映射的聚合数据。作为一种实施例，图12中的装置可以用于实现前文论述的任意一种基于内容推送的数据处理方法。

[0270] 与上述方法实施例基于同一发明构思，本申请实施例中还提供了一种计算机设备。该计算机设备可以用于基于推送内容的数据处理。在一种实施例中，该计算机设备可以是服务器，如图1所示的数据处理服务器130。在该实施例中，计算机设备的结构可以如图13所示，包括存储器1301，通讯模块1303以及一个或多个处理器1302。

[0271] 存储器1301，用于存储处理器1302执行的计算机程序。存储器1301可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统，以及运行即时通讯功能所需的程序等；存储数据区可存储各种即时通讯信息和操作指令集等。

[0272] 存储器1301可以是易失性存储器(volatile memory)，例如随机存取存储器(random-access memory，RAM)；存储器1301也可以是非易失性存储器(non-volatile memory)，例如只读存储器，快闪存储器(flash memory)，硬盘(hard disk drive，HDD)或固态硬盘(solid-state drive，SSD)；或者存储器1301是能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质，但不限于此。存储器1301可以是上述存储器的组合。

[0273] 处理器1302，可以包括一个或多个中央处理单元(central processing unit，CPU)或者为数字处理单元等等。处理器1302，用于调用存储器1301中存储的计算机程序时实现上述基于内容推送的数据处理方法。

[0274] 通讯模块1303用于与终端设备和其他服务器进行通信。

[0275] 本申请实施例中不限定上述存储器1301、通讯模块1303和处理器1302之间的具体连接介质。本公开实施例在图13中以存储器1301和处理器1302之间通过总线1304连接，总线1304在图13中以粗线表示，其它部件之间的连接方式，仅是进行示意性说明，并不引以为限。总线1304可以分为地址总线、数据总线、控制总线等。为便于表示，图13中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0276] 存储器1301中存储有计算机存储介质，计算机存储介质中存储有计算机可执行指令，计算机可执行指令用于实现本申请实施例的内容推荐方法。处理器1302用于执行上述的基于内容推送的数据处理方法。

[0277] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于一计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：移动存储设备、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或

者光盘等各种可以存储程序代码的介质。

[0278] 或者,发明上述集成的单元如果以软件功能模块的形式实现并作为独立的产品销售或使用时,也可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机、服务器、或者网络设备等)执行本发明各个实施例上述方法的全部或部分。而前述的存储介质包括:移动存储设备、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0279] 基于同一技术构思,本申请实施例还一种计算机可读存储介质,该计算机可读存储介质存储有计算机指令,当上述计算机指令在计算机上运行时,使得计算机执行如前文论述的基于内容推送的数据处理方法。

[0280] 本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0281] 显然,本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样,倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内,则本申请也意图包含这些改动和变型在内。

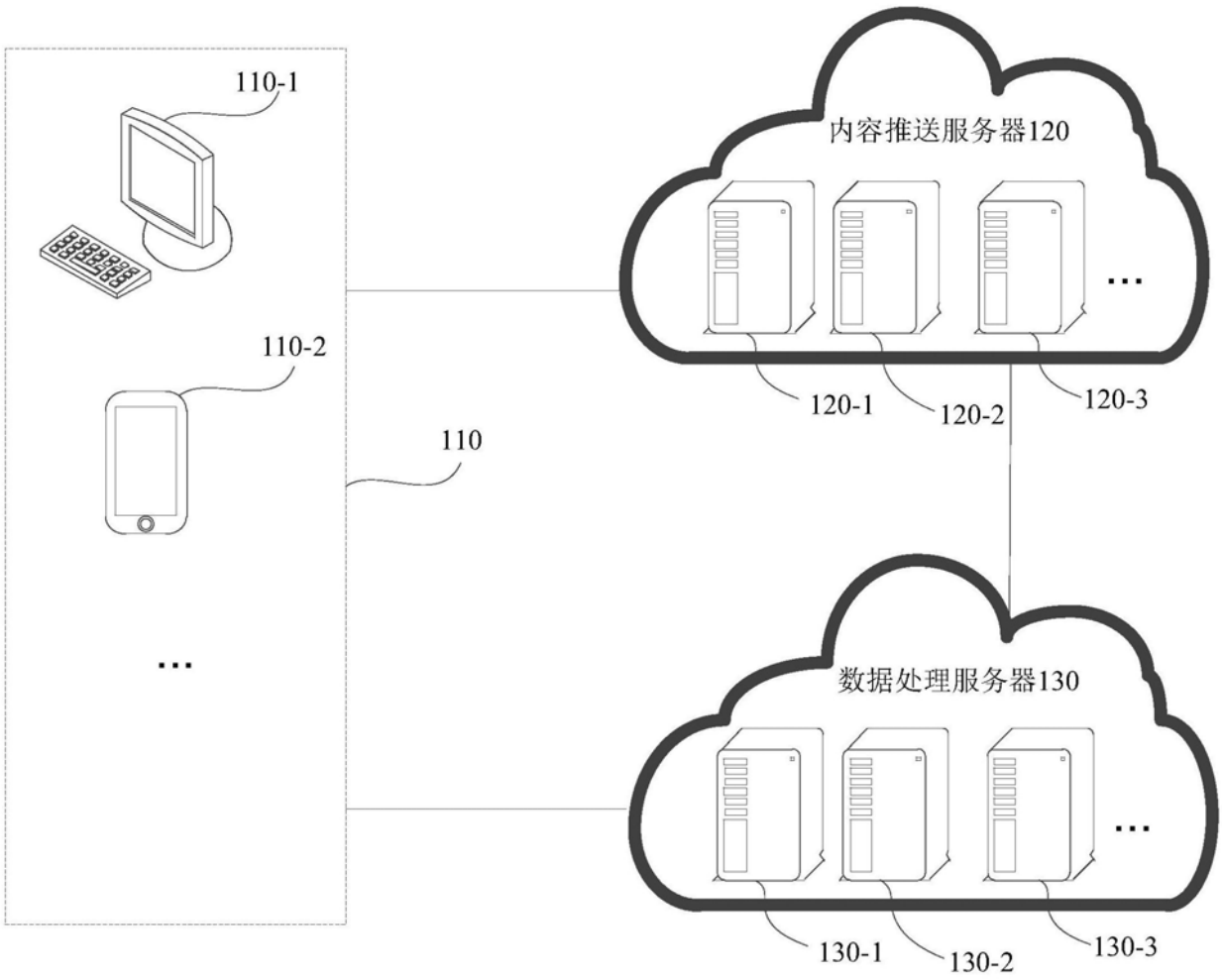

110-1
110-2
110
...
内容推送服务器120
...
120-1
120-2
120-3
数据处理服务器130
...
130-1
130-2
130-3

图1

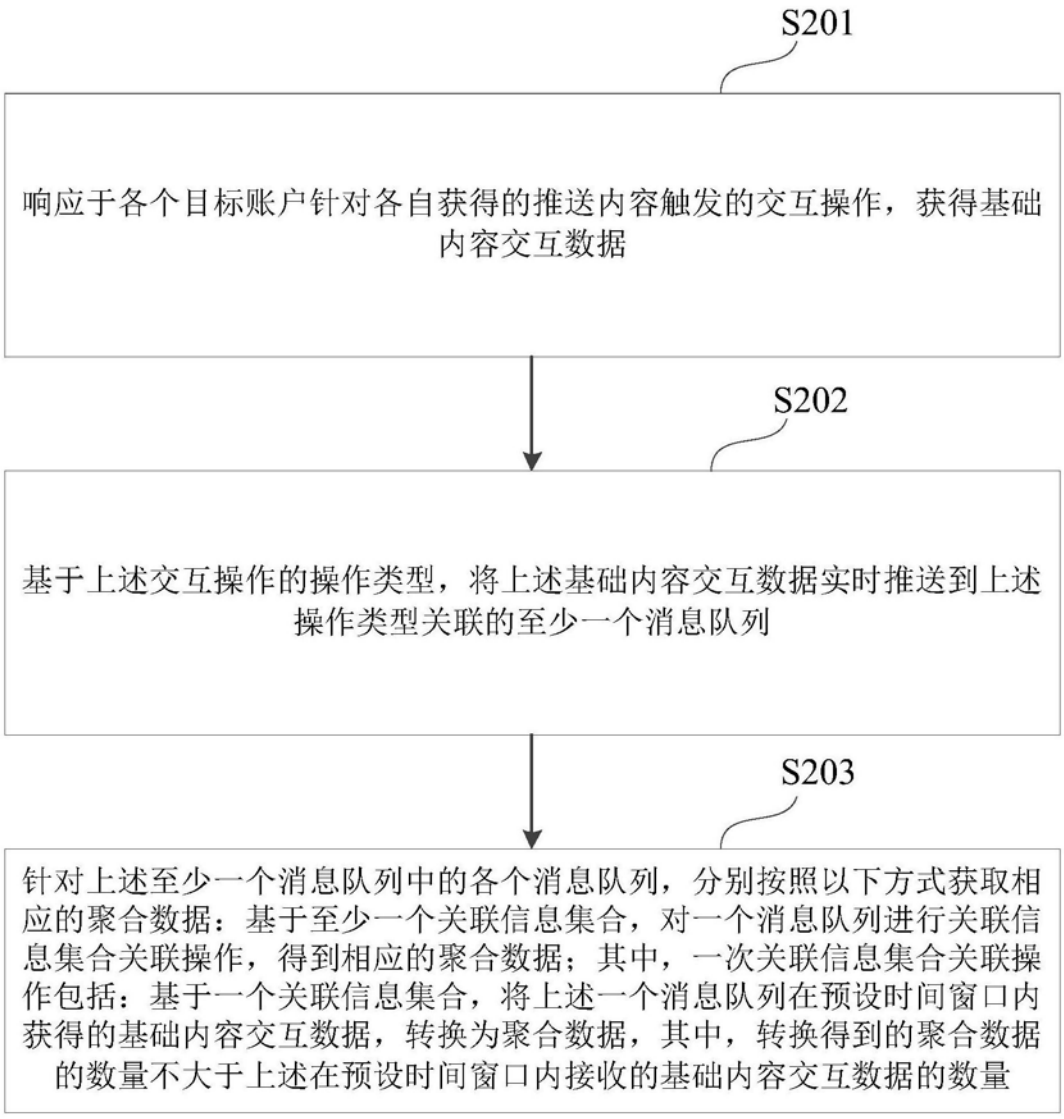

S201
响应于各个目标账户针对各自获得的推送内容触发的交互操作，获得基础内容交互数据
S202
基于上述交互操作的操作类型，将上述基础内容交互数据实时推送到上述操作类型关联的至少一个消息队列
S203
针对上述至少一个消息队列中的各个消息队列，分别按照以下方式获取相应的聚合数据：基于至少一个关联信息集合，对一个消息队列进行关联信息集合关联操作，得到相应的聚合数据；其中，一次关联信息集合关联操作包括：基于一个关联信息集合，将上述一个消息队列在预设时间窗口内获得的基础内容交互数据，转换为聚合数据，其中，转换得到的聚合数据的数量不大于上述在预设时间窗口内接收的基础内容交互数据的数量

图2

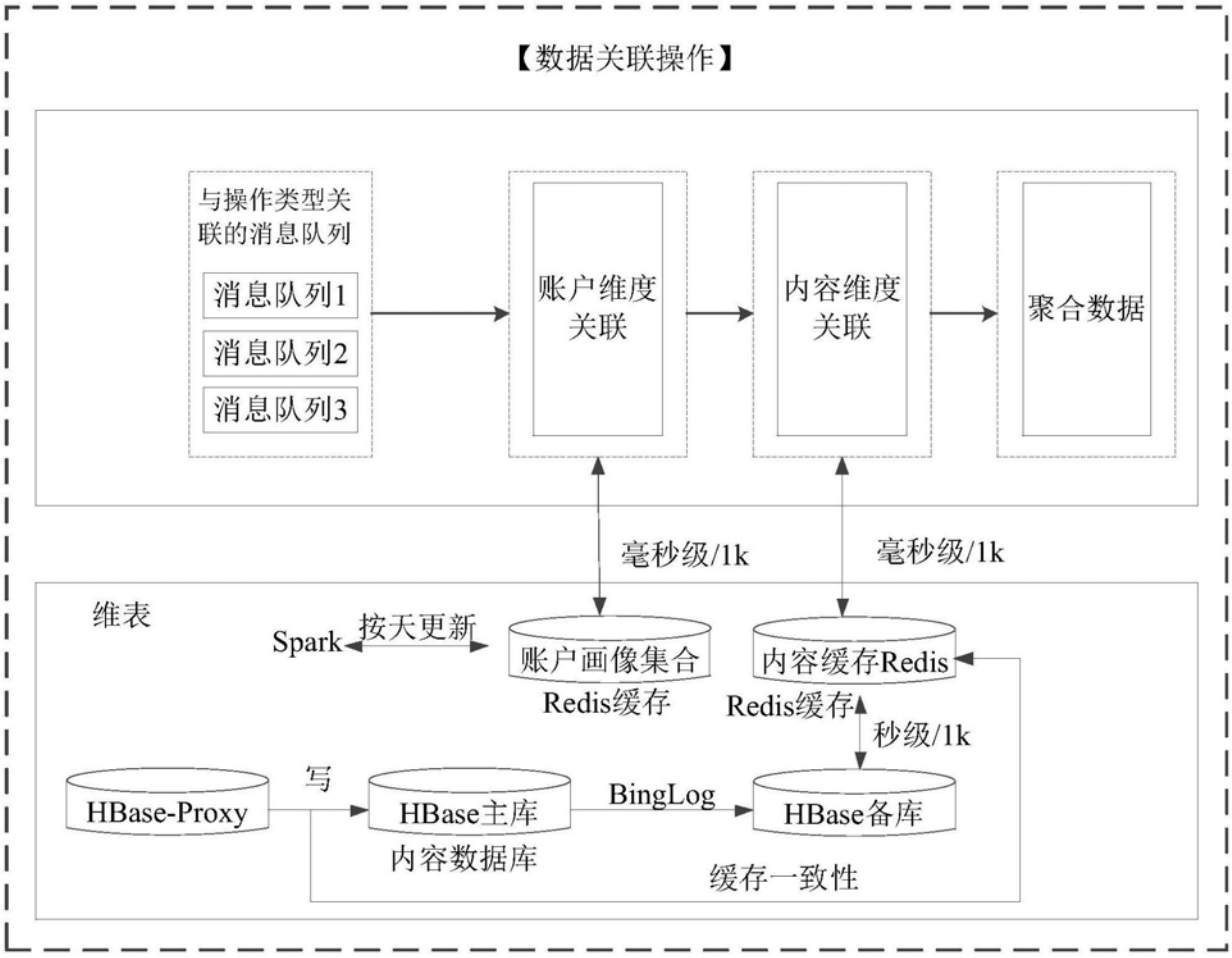

【数据关联操作】
与操作类型关联的消息队列
消息队列1
消息队列2
消息队列3
账户维度关联
内容维度关联
聚合数据
毫秒级/1k
毫秒级/1k
维表
Spark
按天更新
账户画像集合
Redis缓存
内容缓存Redis
Redis缓存
秒级/1k
HBase-Proxy
写
HBase主库
内容数据库
BingLog
HBase备库
缓存一致性

图3

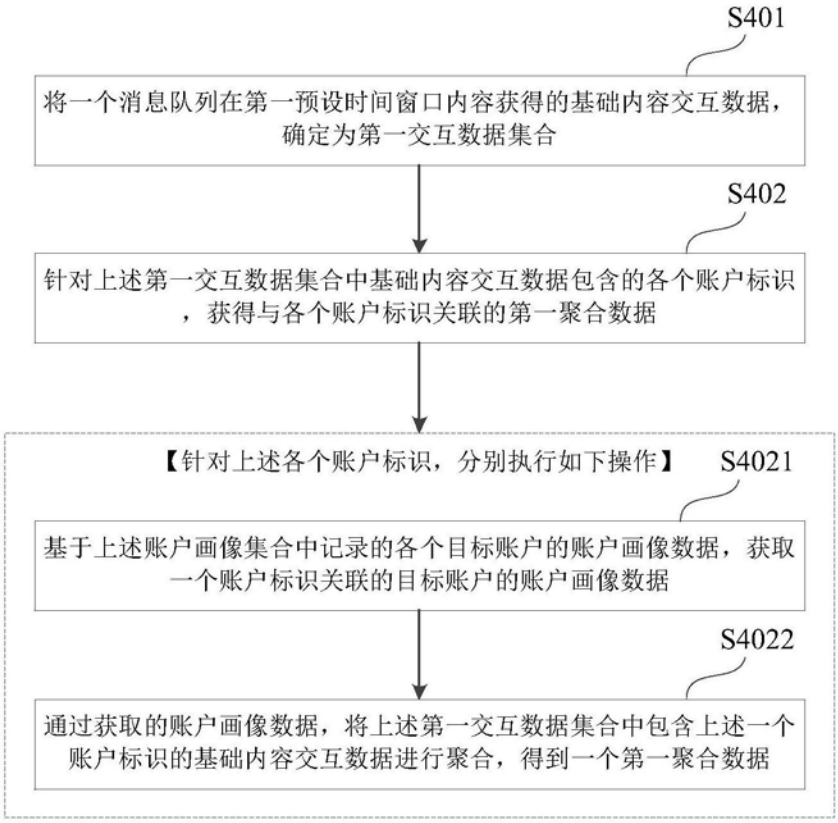

S401
将一个消息队列在第一预设时间窗口内容获得的基础内容交互数据，确定为第一交互数据集合
S402
针对上述第一交互数据集合中基础内容交互数据包含的各个账户标识，获得与各个账户标识关联的第一聚合数据
【针对上述各个账户标识，分别执行如下操作】
S4021
基于上述账户画像集合中记录的各个目标账户的账户画像数据，获取一个账户标识关联的目标账户的账户画像数据
S4022
通过获取的账户画像数据，将上述第一交互数据集合中包含上述一个账户标识的基础内容交互数据进行聚合，得到一个第一聚合数据

图4

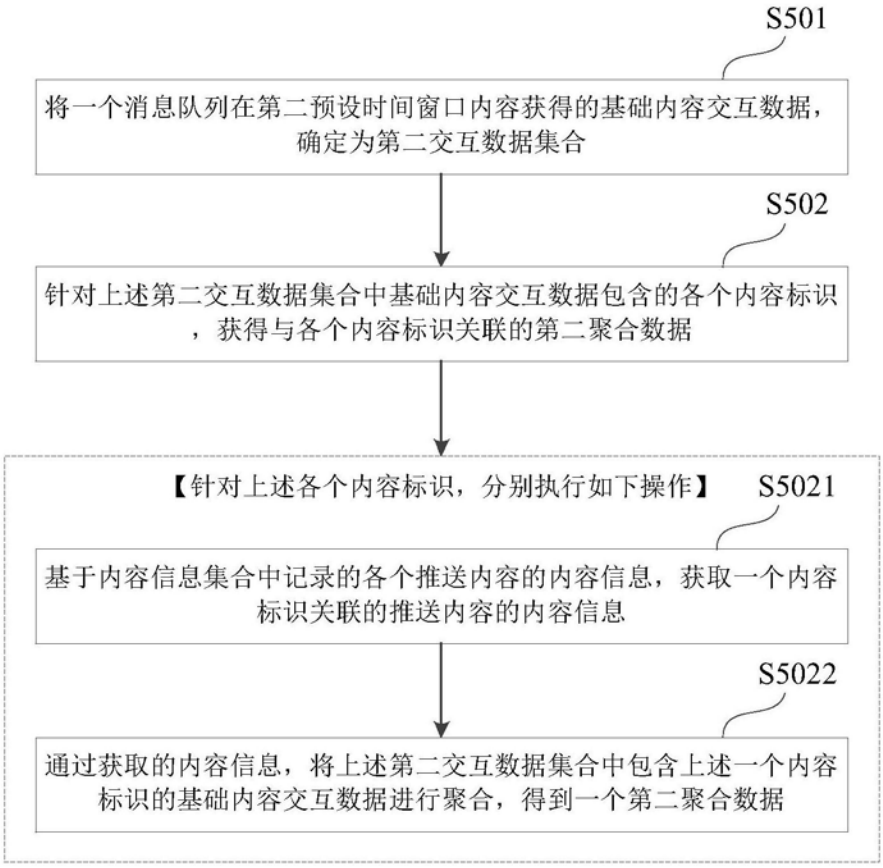

S501
将一个消息队列在第二预设时间窗口内容获得的基础内容交互数据，确定为第二交互数据集合
S502
针对上述第二交互数据集合中基础内容交互数据包含的各个内容标识，获得与各个内容标识关联的第二聚合数据
【针对上述各个内容标识，分别执行如下操作】
S5021
基于内容信息集合中记录的各个推送内容的内容信息，获取一个内容标识关联的推送内容的内容信息
S5022
通过获取的内容信息，将上述第二交互数据集合中包含上述一个内容标识的基础内容交互数据进行聚合，得到一个第二聚合数据

图5

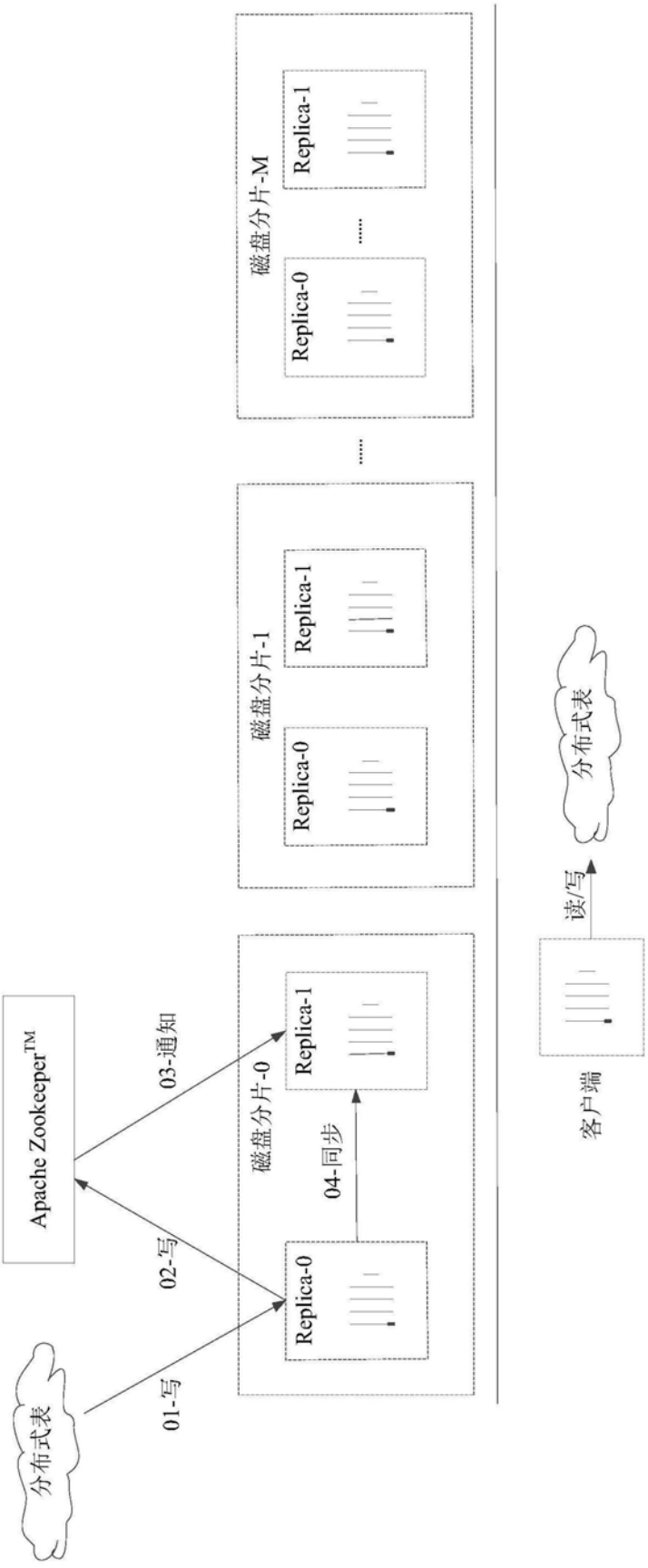

磁盘分片-M
Replica-1
Replica-0
磁盘分片-1
Replica-1
Replica-0
磁盘分片-0
Replica-1
Replica-0
04-同步
Apache Zookeeper™
03-通知
02-写
01-写
分布式表
分布式表
读/写
客户端

图6

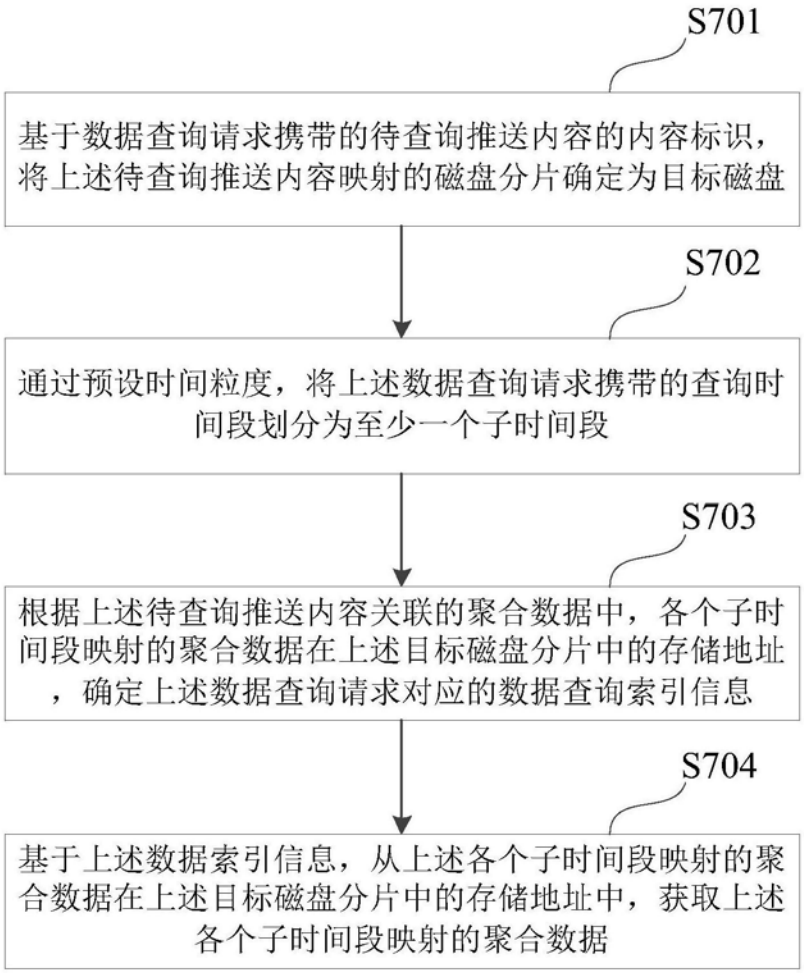

S701
基于数据查询请求携带的待查询推送内容的内容标识，将上述待查询推送内容映射的磁盘分片确定为目标磁盘
S702
通过预设时间粒度，将上述数据查询请求携带的查询时间段划分为至少一个子时间段
S703
根据上述待查询推送内容关联的聚合数据中，各个子时间段映射的聚合数据在上述目标磁盘分片中的存储地址，确定上述数据查询请求对应的数据查询索引信息
S704
基于上述数据索引信息，从上述各个子时间段映射的聚合数据在上述目标磁盘分片中的存储地址中，获取上述各个子时间段映射的聚合数据

图7

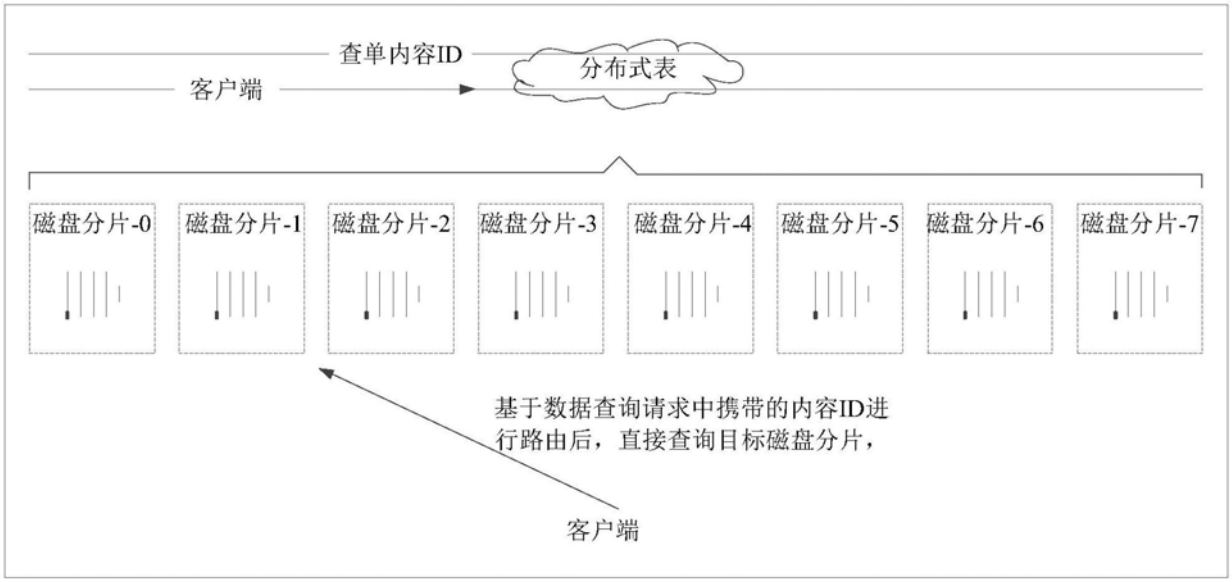

查单内容ID
客户端
分布式表
磁盘分片-0
磁盘分片-1
磁盘分片-2
磁盘分片-3
磁盘分片-4
磁盘分片-5
磁盘分片-6
磁盘分片-7
基于数据查询请求中携带的内容ID进行路由后，直接查询目标磁盘分片，
客户端

图8

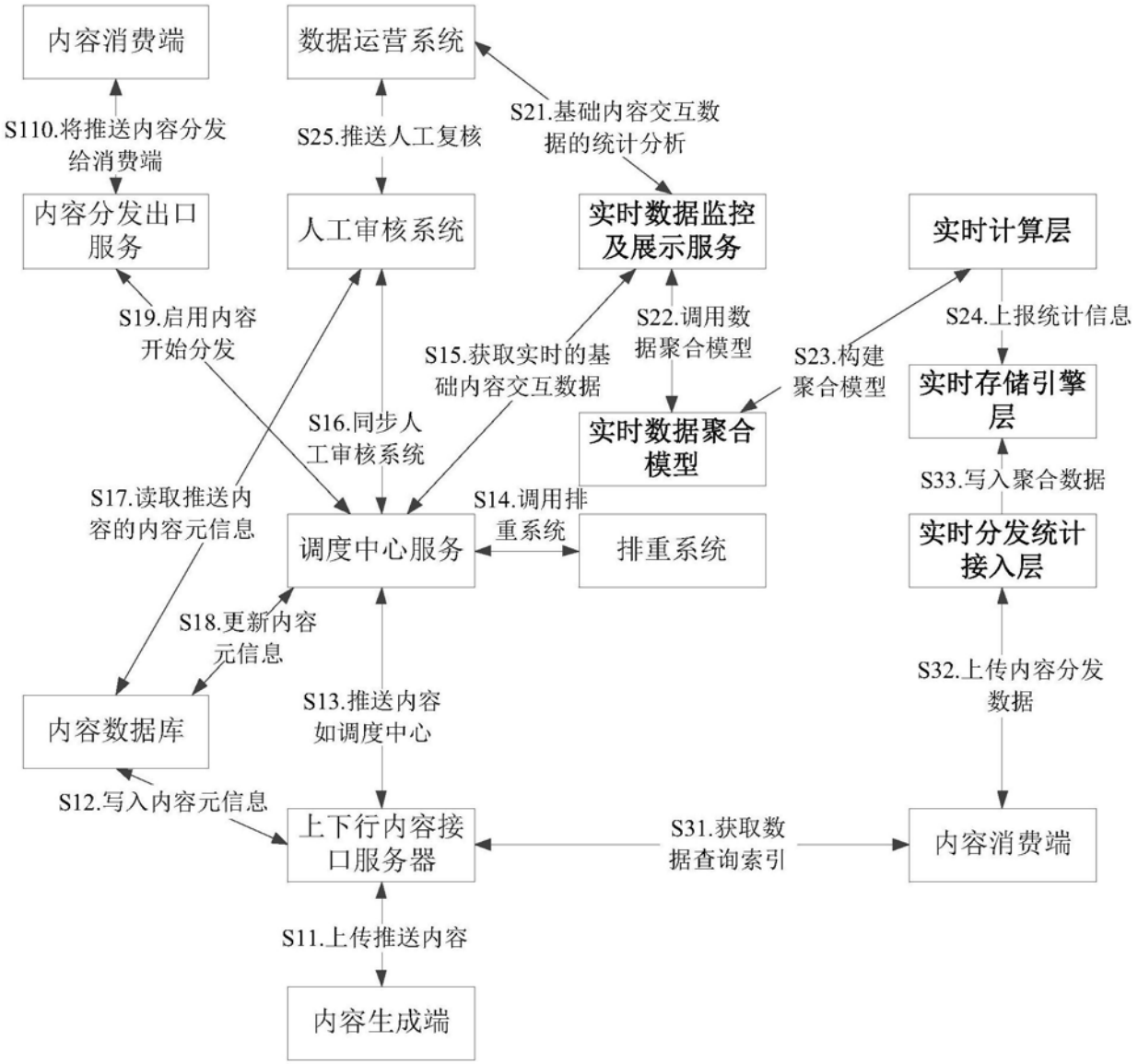

内容消费端
数据运营系统
S110.将推送内容分发给消费端
S25.推送人工复核
S21.基础内容交互数据的统计分析
内容分发出口服务
人工审核系统
实时数据监控及展示服务
实时计算层
S19.启用内容开始分发
S22.调用数据聚合模型
S24.上报统计信息
S15.获取实时的基础内容交互数据
S23.构建聚合模型
实时存储引擎层
S16.同步人工审核系统
实时数据聚合模型
S33.写入聚合数据
S17.读取推送内容的内容元信息
S14.调用排重系统
调度中心服务
排重系统
实时分发统计接入层
S18.更新内容元信息
S32.上传内容分发数据
内容数据库
S13.推送内容如调度中心
S12.写入内容元信息
上下行内容接口服务器
S31.获取数据查询索引
内容消费端
S11.上传推送内容
内容生成端

图9

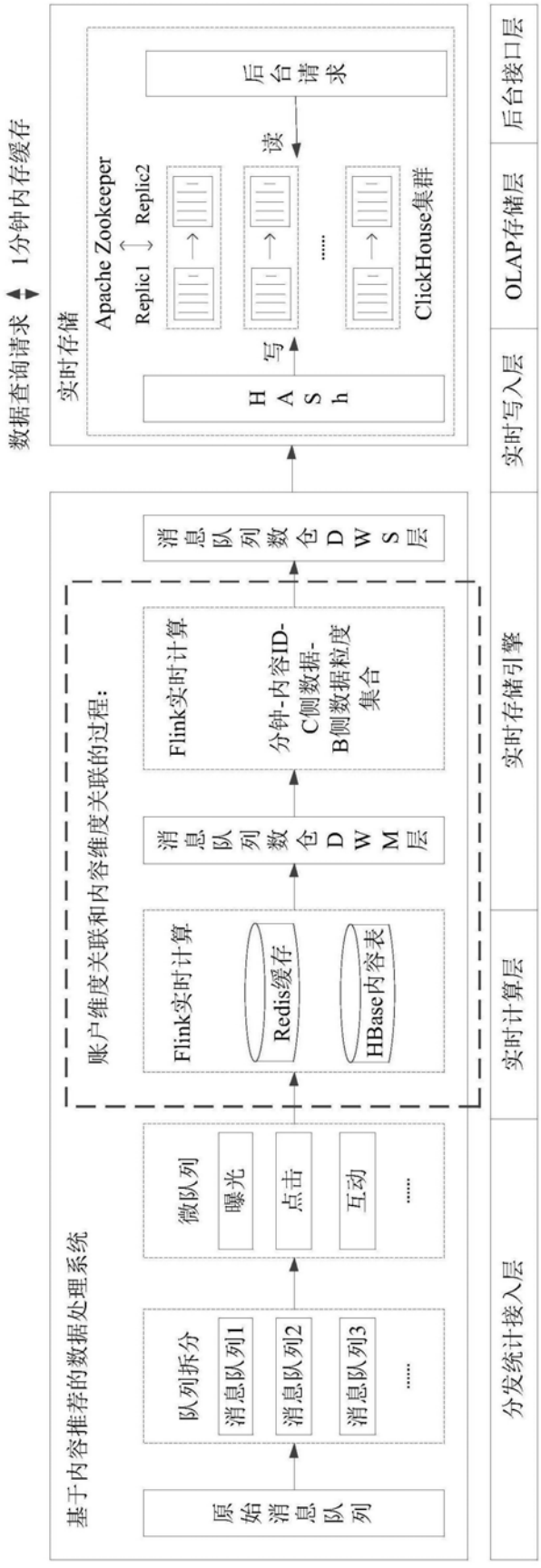

数据查询请求
1分钟内存缓存
实时存储
Apache Zookeeper
Replic1
Replic2
ClickHouse集群
后台请求
读
写
HASh
基于内容推荐的数据处理系统
原始消息队列
队列拆分
消息队列1
消息队列2
消息队列3
微队列
曝光
点击
互动
账户维度关联和内容维度关联的过程：
Flink实时计算
Redis缓存
HBase内容表
消息队列数仓DWM层
Flink实时计算
分钟-内容ID-C侧数据-B侧数据粒度集合
消息队列数仓DWS层
分发统计接入层
实时计算层
实时存储引擎
实时写入层
OLAP存储层
后台接口层

图10

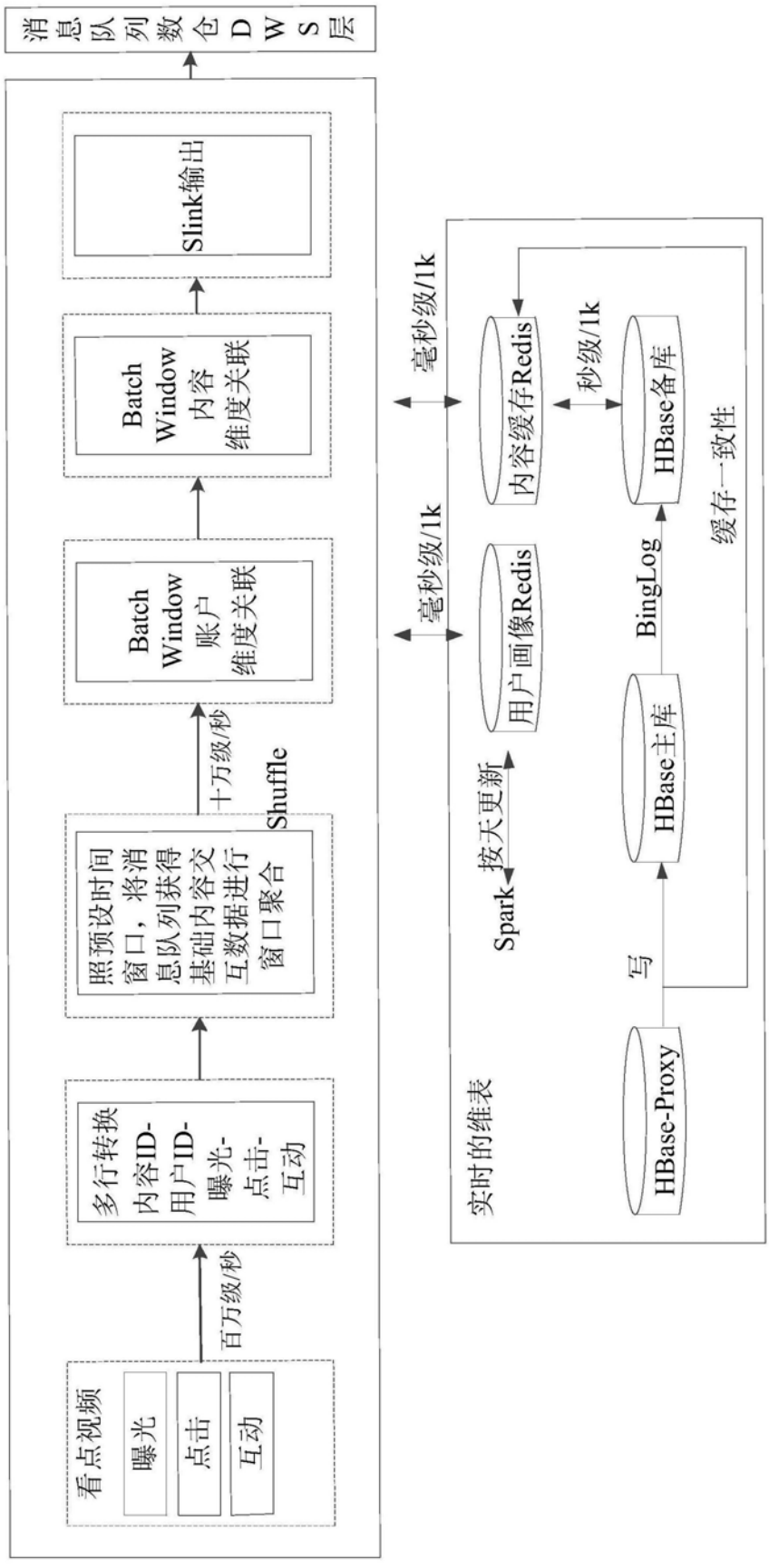

消息队列数仓DWS层
Slink输出
Batch Window 内容维度关联
Batch Window 账户维度关联
照预设时间窗口，将消息队列获得基础内容交互数据进行窗口聚合
多行转换 内容ID-用户ID-曝光-点击-互动
看点视频
曝光
点击
互动
百万级/秒
十万级/秒
Shuffle
毫秒级/1k
毫秒级/1k
秒级/1k
内容缓存Redis
用户画像Redis
HBase备库
HBase主库
HBase-Proxy
BingLog
写
Spark
按天更新
实时的维表
缓存一致性

图11

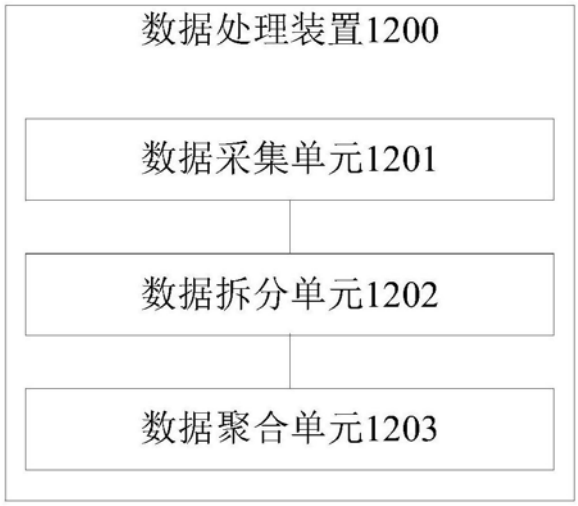

数据处理装置1200
数据采集单元1201
数据拆分单元1202
数据聚合单元1203

图12

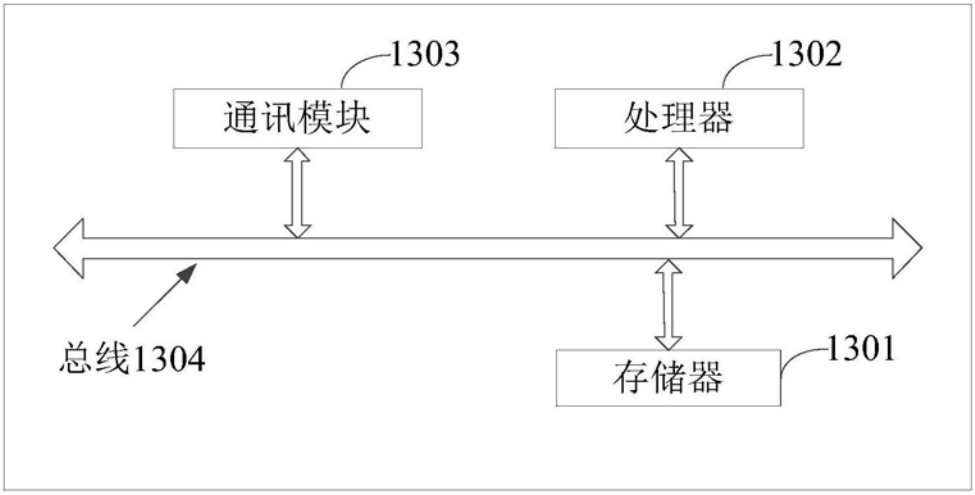

1303
通讯模块
1302
处理器
总线1304
1301
存储器

图13