(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107103064 B
(45)授权公告日 2021.06.01

(21)申请号 201710248134.9

(22)申请日 2017.04.17

(65)同一申请的已公布的文献号
申请公布号 CN 107103064 A

(43)申请公布日 2017.08.29

(73)专利权人 北京五八信息技术有限公司
地址 100083 北京市海淀区学清路甲18号中关村东升科技园学院园三层301室

(72)发明人 丰宗军

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
代理人 陈文香 刘芳

(51)Int.Cl.
G06F 16/2458(2019.01)
G06F 16/22(2019.01)
G06F 16/25(2019.01)
G06F 16/27(2019.01)

(56)对比文件
CN 101039211 A,2007.09.19
CN 101039211 A,2007.09.19
CN 105868197 A,2016.08.17
US 9256761 B1,2016.02.09
郑耀东.基于Hadoop的百度游戏数据平台的设计与实现.《中国优秀硕士学位论文全文数据库 信息科技辑》.2015,(第2期),第46-49页.

审查员 张彪

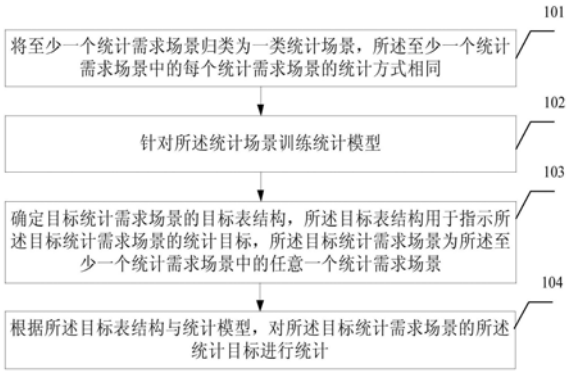
权利要求书2页 说明书10页 附图3页

(54)发明名称

数据统计方法及装置

(57)摘要

本申请实施例提供一种数据统计方法及装置,将至少一个统计需求场景归类为一类统计场景,至少一个统计需求场景中的每个统计需求场景的统计方式相同,针对统计场景训练统计模型,确定目标统计需求场景的目标表结构,目标表结构用于指示目标统计需求场景的统计目标,目标统计需求场景为至少一个统计需求场景中的任意一个统计需求场景,根据目标表结构与统计模型,对目标统计需求场景的统计目标进行统计。该过程中,通过抽象归类,将统计方式相同的至少一个统计需求场景归类为一类统计场景,仅针对该统计场景训练统计模型,并基于该统计模型进行数据统计以满足用户对海量数据的统计需求。



CN 107103064 B

1.一种数据统计方法,其特征在于,包括:

将至少两个统计需求场景归类为一类统计场景,所述至少两个统计需求场景中的每个统计需求场景的统计方式相同,所述至少两个统计需求场景包括第一统计需求场景和第二统计需求场景,所述第一统计需求场景对应的表结构和所述第二统计需求场景对应的表结构不同;

针对所述统计场景训练统计模型;

确定目标统计需求场景的目标表结构,所述目标表结构用于指示所述目标统计需求场景的统计目标,所述目标统计需求场景为所述至少两个统计需求场景中的任意一个统计需求场景;

根据所述目标表结构与统计模型,对所述目标统计需求场景的所述统计目标进行统计;

所述根据所述目标表结构与统计模型,对所述目标统计需求场景的所述统计目标进行统计之前,还包括:

对所述目标统计需求场景配置统计任务。

2.根据权利要求1所述的方法,其特征在于,所述统计任务指示所述目标统计需求场景的即时统计任务,所述根据所述目标表结构与统计模型,对所述目标统计需求场景的所述统计目标进行统计,包括:

根据所述目标表结构与统计模型,对所述目标统计需求场景的所述统计目标进行即时统计。

3.根据权利要求1所述的方法,其特征在于,所述统计任务指示所述目标统计需求场景的定时统计任务,所述根据所述目标表结构与统计模型,对所述目标统计需求场景的所述统计目标进行统计,包括:

根据所述目标表结构与统计模型,对所述目标统计需求场景的所述统计目标进行定时统计。

4.一种数据统计装置,其特征在于,包括:

归类模块,用于将至少两个统计需求场景归类为一类统计场景,所述至少两个统计需求场景中的每个统计需求场景的统计方式相同,所述至少两个统计需求场景包括第一统计需求场景和第二统计需求场景,所述第一统计需求场景对应的表结构和所述第二统计需求场景对应的表结构不同;

训练模块,用于针对所述统计场景训练统计模型;

确定模块,用于确定目标统计需求场景的目标表结构,所述目标表结构用于指示所述目标统计需求场景的统计目标,所述目标统计需求场景为所述至少两个统计需求场景中的任意一个统计需求场景;

统计模块,用于根据所述目标表结构与统计模型,对所述目标统计需求场景的所述统计目标进行统计;

配置模块,用于在所述统计模块根据所述目标表结构与统计模型,对所述目标统计需求场景的所述统计目标进行统计之前,对所述目标统计需求场景配置统计任务。

5.根据权利要求4所述的装置,其特征在于,所述统计任务指示所述目标统计需求场景的即时统计任务,所述统计模块,具体用于根据所述目标表结构与统计模型,对所述目标统

计需求场景的所述统计目标进行即时统计。

6.根据权利要求4所述的装置，其特征在于，所述统计任务指示所述目标统计需求场景的定时统计任务，所述统计模块，具体用于根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行定时统计。

数据统计方法及装置

技术领域

[0001] 本申请实施例涉及数据统计技术，尤其涉及一种数据统计方法及装置。

背景技术

[0002] 随着越来越激烈的市场竞争，数据作为业务精细化运营与管理决策的重要依据，成为驱动互联网行业的根本力量。商家通过数据统计可以准确的把握用户的品牌感知，然后针对性的进行品牌塑造。

[0003] 数据统计过程中，针对特定的统计需求训练统计模型，利用该统计模型对特定的需求进行数据统计。常见的数据统计包括页面浏览量(page view，PV)、访客量(Unique View，UV)、交易额、订单量等。

[0004] 然而，随着互联网的迅速发展，使得数据越来越海量，海量数据越来越不规则且处于动态变化之中，商家对数据统计的精细化要求程度越来越高，除了关注数据统计的静态结果外，更关注数据的变化趋势。显然，针对特定的统计需求训练处的统计模型，是远远无法满足用户对海信数据的统计需求的。

发明内容

[0005] 本申请实施例提供一种数据统计方法及装置，通过抽象归类训练统计模型，基于该统计模型进行数据统计以满足用户对海量数据的统计需求。

[0006] 第一方面，本申请实施例提供一种数据统计方法，包括：

[0007] 将至少一个统计需求场景归类为一类统计场景，所述至少一个统计需求场景中的每个统计需求场景的统计方式相同；

[0008] 针对所述统计场景训练统计模型；

[0009] 确定目标统计需求场景的目标表结构，所述目标表结构用于指示所述目标统计需求场景的统计目标，所述目标统计需求场景为所述至少一个统计需求场景中的任意一个统计需求场景；

[0010] 根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计。

[0011] 在一种可行的设计中，所述根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计之前，还包括：

[0012] 对所述目标统计需求场景配置统计任务。

[0013] 在一种可行的设计中，所述统计任务指示所述目标统计需求场景的即时统计任务，所述根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计，包括：

[0014] 根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行即时统计。

[0015] 在一种可行的设计中，所述统计任务指示所述目标统计需求场景的定时统计任

务，所述根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计，包括：

[0016] 根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行定时统计。

[0017] 在一种可行的设计中，所述至少一个统计需求场景包括第一统计需求场景和第二统计需求场景，所述第一统计需求场景对应的表结构和所述第二统计需求场景对应的表结构不同。

[0018] 第二方面，本申请实施例提供一种数据统计装置，包括：

[0019] 归类模块，用于将至少一个统计需求场景归类为一类统计场景，所述至少一个统计需求场景中的每个统计需求场景的统计方式相同；

[0020] 训练模块，用于针对所述统计场景训练统计模型；

[0021] 确定模块，用于确定目标统计需求场景的目标表结构，所述目标表结构用于指示所述目标统计需求场景的统计目标，所述目标统计需求场景为所述至少一个统计需求场景中的任意一个统计需求场景；

[0022] 统计模块，用于根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计。

[0023] 在一种可行的设计中，上述的装置还包括：

[0024] 配置模块，用于在所述统计模块根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计之前，对所述目标统计需求场景配置统计任务。

[0025] 在一种可行的设计中，所述统计任务指示所述目标统计需求场景的即时统计任务，所述统计模块，具体用于根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行即时统计。

[0026] 在一种可行的设计中，所述统计任务指示所述目标统计需求场景的定时统计任务，所述统计模块，具体用于根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行定时统计。

[0027] 在一种可行的设计中，所述至少一个统计需求场景包括第一统计需求场景和第二统计需求场景，所述第一统计需求场景对应的表结构和所述第二统计需求场景对应的表结构不同。

[0028] 本申请实施例提供的数据统计方法及装置，将至少一个统计需求场景归类为一类统计场景，至少一个统计需求场景中的每个统计需求场景的统计方式相同，针对统计场景训练统计模型，确定目标统计需求场景的目标表结构，目标表结构用于指示目标统计需求场景的统计目标，目标统计需求场景为至少一个统计需求场景中的任意一个统计需求场景，根据目标表结构与统计模型，对目标统计需求场景的统计目标进行统计。该过程中，通过抽象归类，将统计方式相同的至少一个统计需求场景归类为一类统计场景，仅针对该统计场景训练统计模型，并基于该统计模型进行数据统计以满足用户对海量数据的统计需求。同时，通过抽象归类是抽象出不同统计需求场景的共性，训练统计模型时将公用的处理逻辑封装，对外提供统一的调用，能够很大程度上减少重复开发、提高训练模型的可维护性和通用性。

附图说明

[0029] 图1为本申请数据统计方法实施例一的流程图；

[0030] 图2为本申请数据统计方法所适用的处理过程示意图；

[0031] 图3为本申请数据统计方法所适用的数据统计系统的架构示意图；

[0032] 图4为本申请数据统计方法中所适用的表映射示意图；

[0033] 图5为本申请数据统计装置实施例一的结构示意图；

[0034] 图6为本申请数据统计装置实施例二的结构示意图。

具体实施方式

[0035] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例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，本领域技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。以下内容为结合附图及较佳实施例，对依据本申请申请的具体实施方式、结构、特征及其功效的详细说明。

[0036] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等(如果存在)是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例例如能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0037] 图1为本申请数据统计方法实施例一的流程图，包括：

[0038] 101、将至少一个统计需求场景归类为一类统计场景，所述至少一个统计需求场景中的每个统计需求的统计方式相同。

[0039] 通常来说，很多不同的统计需求场景都可以用完全相同的处理逻辑去处理，例如：统计一个电商平台的即时交易额的统计需求场景和统计一个菜单项的点击次数的统计需求场景，都是累加的统计需求场景，所以属于同一类统计场景。再如，对订单表按照商户维度统计出各个商户的订单总金额的统计需求场景，和对订单表按照地区维度统计出各个地区的订单总金额的统计需求场景，这两个统计需求场景本质上也是同一类统计需求，实际实现时，该两个统计需求场景的订单表存储在服务器，只是统计过程中使用不同的字段而已，其他的处理逻辑都是完全相同的。本步骤中，将至少一个统计需求场景抽象归类为一类统计场景，所述至少一个统计需求场景中的每个统计需求场景的统计方式相同。其中，抽象归类是抽象出不同统计需求场景的共性，将公用的处理逻辑封装，对外提供统一的调用，能够很大程度上减少重复开发、提高程序的可维护性和通用性。

[0040] 102、针对所述统计场景训练统计模块。

[0041] 本步骤中，对抽象归类出的统计场景进行模型训练，训练出统计模型。训练过程中，对抽象归类出的公用的处理逻辑封装，得到统计模型，并对外提供统一的调用。

[0042] 103、确定目标统计需求场景的目标表结构，所述目标表结构用于指示所述目标统

计需求场景的统计目标，所述目标统计需求场景为所述至少一个统计需求场景中的任意一个统计需求场景。

[0043] 针对至少一个统计场景中的任意一个突击需求场景，以下称之为目标统计需求场景，确定该目标统计需求场景的目标表结构，该目标表结构用于指示目标统计需求场景的统计目标，即需要统计什么。

[0044] 104、根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计。

[0045] 在确定出目标统计需求场景的目标表结构后，根据目标表结构与统计模型，对目标统计需求场景的统计目标进行统计。

[0046] 本申请实施例提供的数据统计方法，将至少一个统计需求场景归类为一类统计场景，至少一个统计需求场景中的每个统计需求场景的统计方式相同，针对统计场景训练统计模型，确定目标统计需求场景的目标表结构，目标表结构用于指示目标统计需求场景的统计目标，目标统计需求场景为至少一个统计需求场景中的任意一个统计需求场景，根据目标表结构与统计模型，对目标统计需求场景的统计目标进行统计。该过程中，通过抽象归类，将统计方式相同的至少一个统计需求场景归类为一类统计场景，仅针对该统计场景训练统计模型，并基于该统计模型进行数据统计以满足用户对海量数据的统计需求。同时，通过抽象归类是抽象出不同统计需求场景的共性，训练统计模型时将公用的处理逻辑封装，对外提供统一的调用，能够很大程度上减少重复开发、提高训练模型的可维护性和通用性。

[0047] 可选的，在本申请一个示例中，所述根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计之前，还对所述目标统计需求场景配置统计任务。

[0048] 具体的，目标统计需求场景的目标表结构，该目标表结构用于指示目标统计需求场景的统计目标，即需要统计什么。而如何统计，即要怎样统计，可通过配置统计任务实现。例如，通过配置任务，配置目标统计需求场景的即时统计任务，根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行即时统计；再如，通过配置任务，配置目标统计需求场景的定时统计任务，根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行定时统计。

[0049] 可选的，在本申请一个示例中，所述至少一个统计需求场景包括第一统计需求场景和第二统计需求场景，所述第一统计需求场景对应的表结构和所述第二统计需求场景对应的表结构不同。

[0050] 具体的，属于同一类统计场景的至少一个统计需求场景中，各个统计需求场景的表结构不同。也就是说，属于同一类统计场景的至少一个统计需求场景都采用相同的统计模型，指示统计过程中针对不同的统计需求场景，设置的表结构不一样。

[0051] 图2为本申请数据统计方法所适用的处理过程示意图。请参照图2，本申请实施例中，数据统计包括对数据源进行数据采集、数据存储、数据处理与数据展示、生成报表等环节。

[0052] 图3为本申请数据统计方法所适用的数据统计系统的架构示意图。请参照图3，本申请实施例中，数据统计系统是一套按需求自定义配置的服务，是基于hbase(如图中的的持久化)、流式处理(storm)、hive、impala、搜索引擎解决方案(elasticsearch)、kafka等开源框架之上的集成服务。该数据系统的具有(1)、基于自定义配置和统计模型，与业务系统

解耦;(2)、支持实时统计和定时统计需求;(3)、具有高可扩展性,即:当数据量增加时,可以通过增加节点进行横向扩容;(4)、插件式服务,各子服务之前相互解耦,各子服务内部的实现方案可以随时迭代替换。

[0053] 请参照图3,来自Web端采集模块、手机端采集模块和服务端采集模块的数据经过数据采集后,进行数据清洗,然后依次进行storm流式处理、Hbase持久化、以及Hive/Mapreduce等处理,最后可通MySQL/搜索引擎解决方案(Elasticsearch)进行展示,如Web报表展示引擎、数据导出服务、监控告警服务以及系统监控模块进行展出。另外,该数据统计系统还包括权限认证服务,用于对不同的用户设置不同的权限;任务配置模块,用于配置统计任务;系统配置模块,用于对数据统计系统进行配置。

[0054] 下面,结合图3,以统计每天各地区的订单金额为例,对上述的数据统计系统处理数据的流程进行介绍。

[0055] 首先,创建元数据表,即表结构。

[0056] 具体的,在数据统计系统的管理界面创建出表结构,该表结构例如是订单表,通过表映射机制将表结构自动同步到数据统计系统中所有的数据流环节中。该过程中,在管理界面创建表结构,可以理解为定义了一个基于xml规范的表结构信息,在管理界面创建完成后,数据统计系统会自动更新配置到其他子服务模块,并且也会在hbase和hive中同时创建出与值对应的表结构,即订单表。

[0057] 其次,数据采集。

[0058] 具体的,调用方通过工具包,如客户端UDK(client UDK)工具包,发送订单数据到服务端接口,storm流式处理模块解析表结构,并存储到hbae持久化模块对应的表结构中。其中client UDK工具包提供了表结构的增加、修改、删除定相应的操作方法,只需要指定表结构的名称和参数,就能直接修改hbase中的元数据。

[0059] 再次,配置统计任务。

[0060] 具体的,统计任务中包含了表结构、结果输出表、统计描述信息、统计周期、统计维度、日期以及统计策略等信息。

[0061] 接着,定时统计。

[0062] 统计任务根据任务类型,从统计模型中选择相应的统计模型,hive以及impala支持类sql的查询语言,可将HBase中的数据映射为关系数据库中的数据,大大简化了对Hadoop分布式文件系统(Hadoop Distributed File System,HDFS)上数据处理的难度。数据统计系统自动加载任务信息,并由quartz自动调用统计Job的执行。

[0063] 统计Job将任务信息转化成与Hive、impala相对应的类sql语言,hive/impala执行结束后再将统计结果存储到hbase中。

[0064] 接着,即时统计。

[0065] 数据统计系统中的即时统计是根据用户定义的计算规则对元数据更新的操作,操作事件除了对元数据表的增加、修改和删除操作之外,还包括一类单独的运算事件,可以由调用方提供运算规则,由Storm解析运算规则并修改Hbase中的数据。

[0066] 最后,数据展现。

[0067] 本申请实施例中,数据统计系统中提供了元数据和统计结果的查询服务,它支持mysql和elasticsearch两种方案,可根据需要选择。默认使用elasticsearch,HBase中的元

数据和统计结果数据会自动同步到搜索引擎解决方案(elasticsearch,ES)当中,由ES对外提供查询服务。

[0068] 接下来,对数据统计系统的具体实现进行详细讲解。

[0069] 首先、数据表映射机制。

[0070] 具体的,数据统计系统是一套与调用方业务逻辑完全解耦的系统,它在为调用方提供数据统计功能之前,首先需要的一套统计规则。数据统计系统参考数据库的设计思想,设置表结构,表结构是数据存储的单元。数据统计系统中所有类型的元数据和统计数据都要存储到对应的表中。数据统计系统中的表包含两种:实表和虚表。实表是必须手动去创建的表,而虚表则是程序自动创建的表。实表中存储的是对调用方有实际意义的数据,如订单表、商户表、用户表等,针对于实表可以有不同的统计计划。而虚表则一般都是为了应对一些简单的统计需求,为简化用户操作,而程序自动创建出的表。

[0071] 数据统计系统中的数据表映射是指管理员在web配置界面创建表结构,自动将表结构同步到数据统计系统中的各个环节。表映射的目的是为了实现相同的配置信息在各个数据流转环节以不同的形态的存在,从而让各个环节之间的数据交互成为可能。例如:软件开关工具包(Software Development Kit,SDK)是为调用方提供的工具包,工具包中需要根据用户创建的表结构来检查录入数据的合法性。Storm也会需要相同的数据,而Hbase、Hive中本身就有表结构的概念,数据统计系统会自动根据用户的表结构配置,将表结构数据同步到SDK工具包和storm服务中,并且同时在Hbase和Hive中创建出相对应的表。具体的,可参见图4,图4为本申请数据统计方法中所适用的表映射示意图。请参照图4,表映射模块根据表配置信息,创建Hbase表和Hive表。

[0072] 数据统计系统根据用户定义创建出HBase和Hive表结构,由于HBase中只能存储简单的字符类型,数据统计系统在对数据查询和修改之前会做统一的封装,将HBase实际存储的字符类型转化成用户定义的参数类型。对于上层应用来说对HBase和Hive中表的操作都是完全按照用户定义来操作表的。

[0073] 其次,统一配置任务。

[0074] 具体的,数据统计系统的整个配置管理服务包含三个部分:zookeeper集群、数据统计系统web端管理服务和其他分布式应用服务。通过在web端新增或修改配置文件,同时将配置信息上传到zookeeper的永久节点,其他分布式应用服务监听响应配置节点,当接收到配置变化通知时再去zookeeper上获取最新的配置信息。另外,web管理服务有配置文件版本回溯机制,便于操作的回退。因此,统一配置的优点在于,第一、配置发布统一化,降低运维成本;第二、配置下发后及时更新,配置热加载。

[0075] 再次,数据采集。

[0076] 数据统计系统将每一条元数据的发送定义为一个事件,这个事件包含了事件的类型、相关表、事件名称、事件内容、发生时间等信息。

[0077] 数据统计系统中的事件目前包含了四种类型:

[0078] ADD_EVENT:表数据增加事件,是最常用的一种即元数据入库事件。

[0079] UPD_EVENT:表数据修改事件,元数据修改事件。

[0080] DEL_EVENT:表数据删除事件,元数据删除事件。

[0081] CALC_EVENT:表数据计算修改事件,CALC_EVENT是在UPD_EVENT事件之上做了额外

的自定义的扩展。数据统计中有很多的计算方式是基于原始数据的，即在原始数据上进行修改从而得到新的数据，如计数操作是在原来的基础上+1。CALC_EVENT是可以由调用方自定义计算公式，而数据统计系统则按照调用方的计算公式来处理。

[0082] 对于元数据的采集，业内比较通用的做法是基于flume、fluentd、scribe等开源框架，而这一类的开源框架往往基于日志文件的采集方式，这种数据采集方式优点是与业务完全解耦，数据采集流程也不会影响业务自身的正常运行。但其实这种方案也有明显的弊端，因为大多系统日志中的数据相对杂乱，其输出内容完全由业务方决定，对这种元数据其数据清洗和处理难度较高。同时这种实现方式需要在每个调用方服务器上部署相应的采集模块，其管理维护成本也非常高。

[0083] 数据统计系统中元数据要存储到对应的表结构中去，元数据的格式与表结构是对应的。数据统计系统采取的是一种基于API的采集机制。Server端完全由thrift接口来提供元数据的采集服务。可以说相对于flume/fluentd那种主动获取的采集方式，数据统计系统使用的是被动接收的数据获取方式，即“调用方想统计什么数据，就传什么数据过来”，另外，数据统计系统对于元数据的采取强验证处理，对于不符合表结构规则的数据则全部过滤掉。

[0084] Thrift是Apache下的一个子项目，最早是脸书(Facebook)的项目，具有以下特征：

[0085] 第一、拥有自己的跨机器的通信框架，并提供一套库。

[0086] 第二、是一个代码生成器，按照它的规则，可以生成多种编程语言的通讯过程代码。

[0087] 第三、提供多语言的编译功能，并提供多种服务器工作模式；用户通过Thrift的IDL(接口定义语言)来描述接口函数及数据类型，然后通过Thrift的编译环境生成各种语言类型的接口文件，可以根据自己的需要采用不同的语言开发客户端代码和服务器端代码。

[0088] Thrift简化了不同语言间基于socket的通讯流程，提供阻塞、非阻塞、单线程和多线程的模式运行在服务器上，大大提高了程序的通用性。本申请实例的数据统计系统基于thrift的接口方案充分考虑不同调用方的语言环境，对调用方提供相应语言的SDK工具包，简化调用复杂度。

[0089] 元数据存储这种应用场景数据量极大、要求低延迟、此外对消息的处理顺序和处理可靠性有较高的要求。Storm是一个开源分布式实时计算系统，可以简单、可靠的处理大量数据流。Storm支持水平扩展，具有高容错性，保证每个消息都会得到处理。

[0090] Storm是一套流式处理框架，它包含了nimbus和supervisor两种服务节点，nimbus负责任务的分配和集群管理，而supervisor则负责了具体的任务执行，supervisor中包含了若干个worker进程去处理具体的任务。

[0091] storm一般用于处理在线及时流数据，它把每一条消息定义为一个tuple原语，连续的多个tuble原语即为stream(流)。Storm有三种抽象类型：Spout、螺栓(Bolt)以及拓扑(Topology)，Spout是计算流的来源。通常可以从各种消息队列中获取数据。Spout即为数据源，可以从各种消息队列中获取数据，而bolt则负责具体的逻辑，spout可以根据时间类型分发消息到不同的bolt处理。Topology可以理解为某一项特定的处理任务，是由很多spout和bolt组成的网络，它规定了spout如何读取实时流数据，并交由哪一个或几个bolt去处

理。

[0092] Storm可以灵活的设置每一个spout和bolt的并行度,以多进程的方式去处理批量任务,大大提高了任务的处理效率。数据统计系统基于storm的流式处理方案之上,将事件类型与bolt进行关联,不同的事件类型交由不同的bolt处理,bolt将结果一一处理存储到HBase中去。数据统计系统中除了支持固定元数据采集之外,还支持算式运算,对应的处理逻辑有CALCBolt实现。调用方可自己定义计算表达式,CALCBolt根据用户定义表达式计算出结果再保存到HBase中去。

[0093] 自定义流式计算可以实现大部分的及时数据统计需求,比如统计每天实时订单总金额,首先在cuber web管理界面配置及时任务信息,程序在每天0点创建出对应HBase指定表中的数据项,每笔订单触发CALC_EVENT事件,而运算表达式是hbase表中的当前值加上该笔订单金额。CALCBolt通过Aviator表达式引擎解析出用户自定义算法,计算完成后再将结果保存到HBase中,从而实现实时统计。除了这种简单的累加场景之外,Aviator支持更加复杂多样的计算方法满足用户的各种需求。

[0094] 接着,数据存储。

[0095] 本申请实例中,元数据写入的场景的特点是:第一、海量;第二、即时性要求高;第三、便于统计;第四、对数据丢失容忍度低。

[0096] 数据统计系统基于HBase的实现数据重传。HBase是一个构建在HDFS上的分布式列存储系统,基于Google BigTable模型开发,是Apache Hadoop生态系统中的重要一员,主要用于海量结构化数据存储。其中,Hbase具有如下特点:第一、大,一个表可以有数十亿行,上百万列;第二、无模式,每行都有一个可排序的主键和任意多的列,列可以根据需要动态的增加,同一张表中不同的行可以有截然不同的列;第三、面向列,面向列(族)的存储和列(族)独立检索;第四、稀疏,空(null)列并不占用存储空间,表可以设计的非常稀疏;第五、数据多版本:每个单元中的数据可以有多个版本,默认情况下版本号自动分配,是单元格插入时的时间戳。

[0097] 在hbase中包含HMaster和HReginServer两种节点,使用zookeeper作为分布式协调服务,HBase是运行在Hadoop上的NoSQL数据库,数据由HDFS做了数据冗余,具有高可靠性。HRegionServer内部管理一系列HRegion对象,每个HRegion对应表格(table)中的一个范围(region),HRegion由多个HStore组成。每个HStore对应了Table中的一个列族(column family)的存储。因此,每个columnfamily其实就是一个集中的存储单元,具备共同IO特性的column放在一个column family中。

[0098] HStore存储是HBase存储的核心,由两部分组成,一部分是MemStore,一部分是StoreFile。MemStore是Sorted Memory Buffer,用户写入的数据首先会放入MemStore,当MemStore满了以后会Flush成一个StoreFile(底层实现是HFile)。

[0099] HBase基于预写日志系统(Write-Ahead Logging,WAL)机制进行数据写入,WAL是一种高效的日志算法,基本原理是在数据写入之前首先顺序写入日志,然后再写入缓存,等到缓存写满之后统一落盘。WAL机制之所以能够提升写性能,是因为WAL将一次随机写转化为了一次顺序写加一次内存写。提升写性能的同时,WAL可以保证数据的可靠性,即在任何情况下数据不丢失。假如一次写入完成之后发生了宕机,即使所有缓存中的数据丢失,也可以通过恢复日志还原出丢失的数据。HBase的WCL机制保证了其在高并发读写场景下的性

能。

[0100] 传统的行式数据库，是按照行存储的，维护大量的索引和物化视图无论是在时间(处理)还是空间(存储)方面成本都很高。而列式数据库恰恰相反，列式数据库的数据是按照列存储，每一列单独存放，数据即是索引。只访问查询涉及的列，大大降低了系统I/O，每一列由一个线来处理，而且由于数据类型一致，数据特征相似，极大方便压缩。行式数据库擅长随机读操作，列式数据库则更适合大批量数据量写入和查询，非常适合大数据统计系统的应用场景。通过上述存储方式，使得所有的rowkey、列、列族都有match属性，match属性是行名、列名、列族名的简写，也是存储到HBase中的数据，可以极大的降低hbase中的资源占用。同时，列族与列的划分由调用方定义，便于调用方将具有相同IO特性的列放在同一列族，从而提高读写效率。

[0101] 接着，定时统计。

[0102] 定时任务是由系统加载定时任务信息，然后由quartz调用统计Job的执行，统计Job会将任务配置信息转化成类sql语言，交由hive和Impala来处理。

[0103] 在描述数据统计系统的定时统计之前，首先介绍下MapReduce和Hive。MapReduce是Hadoop的核心组件之一，hadoop的另外一个核心组件是hdfs，hdfs是分布式存储引擎，而MapReduce就是建立在hdfs上的批量数据处理引擎。

[0104] MapReduce采用"分而治之"的思想，把对大规模数据集的操作，分发给一个主节点管理下的各个分节点共同完成，然后通过整合各个节点的中间结果，得到最终结果。简单地说，MapReduce就是"任务的分解与结果的汇总"。在Hadoop中，用于执行MapReduce任务的机器角色有两个：一个是JobTracker；另一个是TaskTracker，JobTracker是用于调度工作的，TaskTracker是用于执行工作的。一个Hadoop集群中只有一台JobTracker。

[0105] 在Hadoop中，每个MapReduce任务都被初始化为一个Job，每个Job又可以分为两种阶段：map阶段和reduce阶段。这两个阶段分别用两个函数表示，即map函数和reduce函数。map函数接收一个<key,value>形式的输入，然后同样产生一个<key,value>形式的中间输出，Hadoop函数接收一个如<key,(list of values)>形式的输入，然后对这个value集合进行处理，每个reduce产生0或1个输出，reduce的输出也是<key,value>形式的。

[0106] 在分布式计算中，MapReduce框架负责处理了并行编程中分布式存储、工作调度、负载均衡、容错均衡、容错处理以及网络通信等复杂问题，把处理过程高度抽象为两个函数map和reduce，map负责把任务分解成多个任务，reduce负责把分解后多任务处理的结果汇总起来。

[0107] Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供完整的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。其优点是学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。

[0108] Hive是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制。Hive定义了简单的类SQL查询语言，称为HQL，允许熟悉SQL的用户查询数据。同时，该语言也允许熟悉MapReduce开发者的开发自定义的mapper和reducer来处理内建的mapper和reducer无法完成的复杂的分析工作。统计过程中，首先，统计系统加载定时任务信息，由Quartz在指定时刻调度；然后，任务解析模块判断是否可将任务信息解析成可被Impala或Hive可以处理

的类SQL语句。接着，若可以转化成类SQL语句的任务，由Impala和Hive对HBase元数据进行统计，并将统计结果存储到HBase结果表中；若不可转化，则需要有自定义开发的MR程序统计执行。

[0109] 图5为本申请数据统计装置实施例一的结构示意图，包括：

[0110] 归类模块11，用于将至少一个统计需求场景归类为一类统计场景，所述至少一个统计需求场景中的每个统计需求场景的统计方式相同；

[0111] 训练模块12，用于针对所述统计场景训练统计模型；

[0112] 确定模块13，用于确定目标统计需求场景的目标表结构，所述目标表结构用于指示所述目标统计需求场景的统计目标，所述目标统计需求场景为所述至少一个统计需求场景中的任意一个统计需求场景；

[0113] 统计模块14，用于根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计。

[0114] 本申请实施例提供的数据统计装置，将至少一个统计需求场景归类为一类统计场景，至少一个统计需求场景中的每个统计需求场景的统计方式相同，针对统计场景训练统计模型，确定目标统计需求场景的目标表结构，目标表结构用于指示目标统计需求场景的统计目标，目标统计需求场景为至少一个统计需求场景中的任意一个统计需求场景，根据目标表结构与统计模型，对目标统计需求场景的统计目标进行统计。该过程中，通过抽象归类，将统计方式相同的至少一个统计需求场景归类为一类统计场景，仅针对该统计场景训练统计模型，并基于该统计模型进行数据统计以满足用户对海量数据的统计需求。同时，通过抽象归类是抽象出不同统计需求场景的共性，训练统计模型时将公用的处理逻辑封装，对外提供统一的调用，能够很大程度上减少重复开发、提高训练模型的可维护性和通用性

[0115] 图6为本申请数据统计装置实施例二的结构示意图，请参照图6，本申请实施例提供的数据统计装置，在上述图5的基础上，进一步的，还包括：

[0116] 配置模块15，用于在所述统计模块14根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计之前，对所述目标统计需求场景配置统计任务。

[0117] 可选的，在本申请一实施例中，所述统计任务指示所述目标统计需求场景的即时统计任务，所述统计模块14，具体用于根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行即时统计。

[0118] 可选的，在本申请一实施例中，所述统计任务指示所述目标统计需求场景的定时统计任务，所述统计模块14，具体用于根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行定时统计。

[0119] 可选的，在本申请一实施例中，所述至少一个统计需求场景包括第一统计需求场景和第二统计需求场景，所述第一统计需求场景对应的表结构和所述第二统计需求场景对应的表结构不同。

[0120] 本领域普通技术人员可以理解：实现上述各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，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

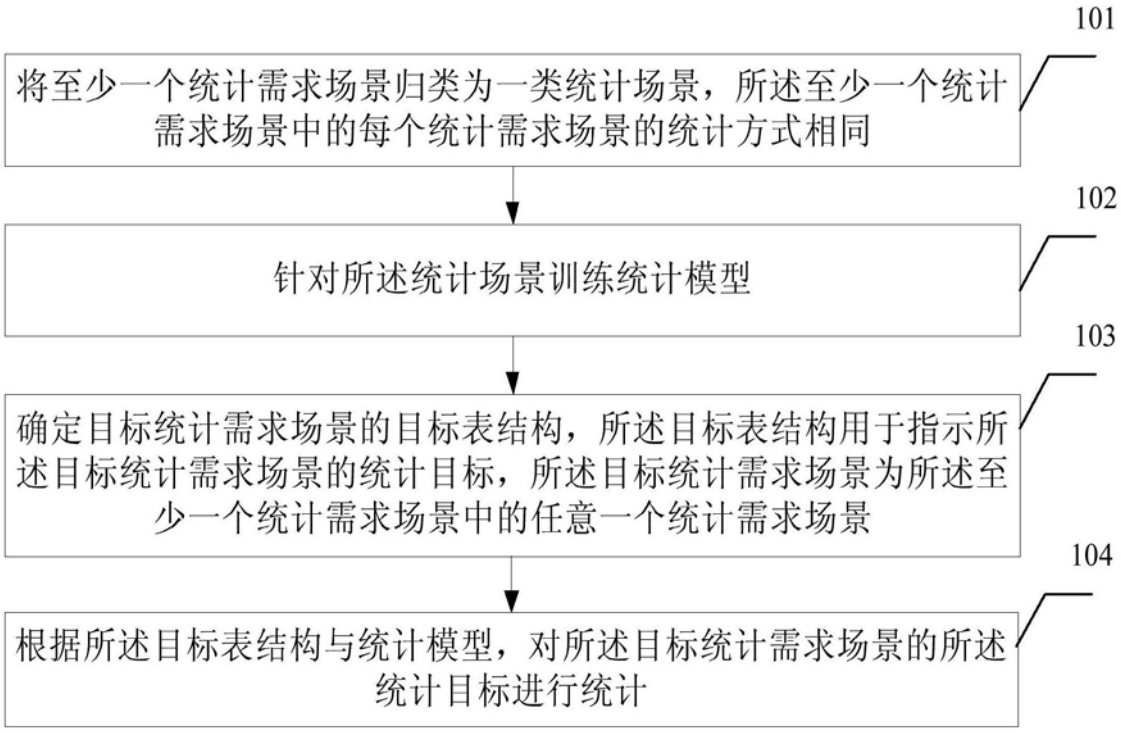

101
将至少一个统计需求场景归类为一类统计场景，所述至少一个统计需求场景中的每个统计需求场景的统计方式相同
102
针对所述统计场景训练统计模型
103
确定目标统计需求场景的目标表结构，所述目标表结构用于指示所述目标统计需求场景的统计目标，所述目标统计需求场景为所述至少一个统计需求场景中的任意一个统计需求场景
104
根据所述目标表结构与统计模型，对所述目标统计需求场景的所述统计目标进行统计

图1

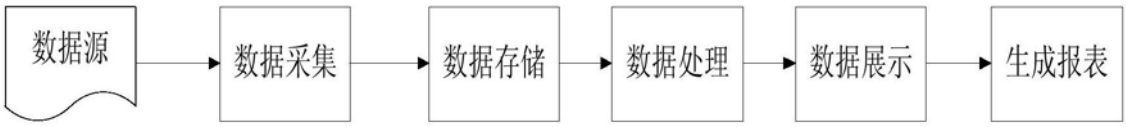

数据源
数据采集
数据存储
数据处理
数据展示
生成报表

图2

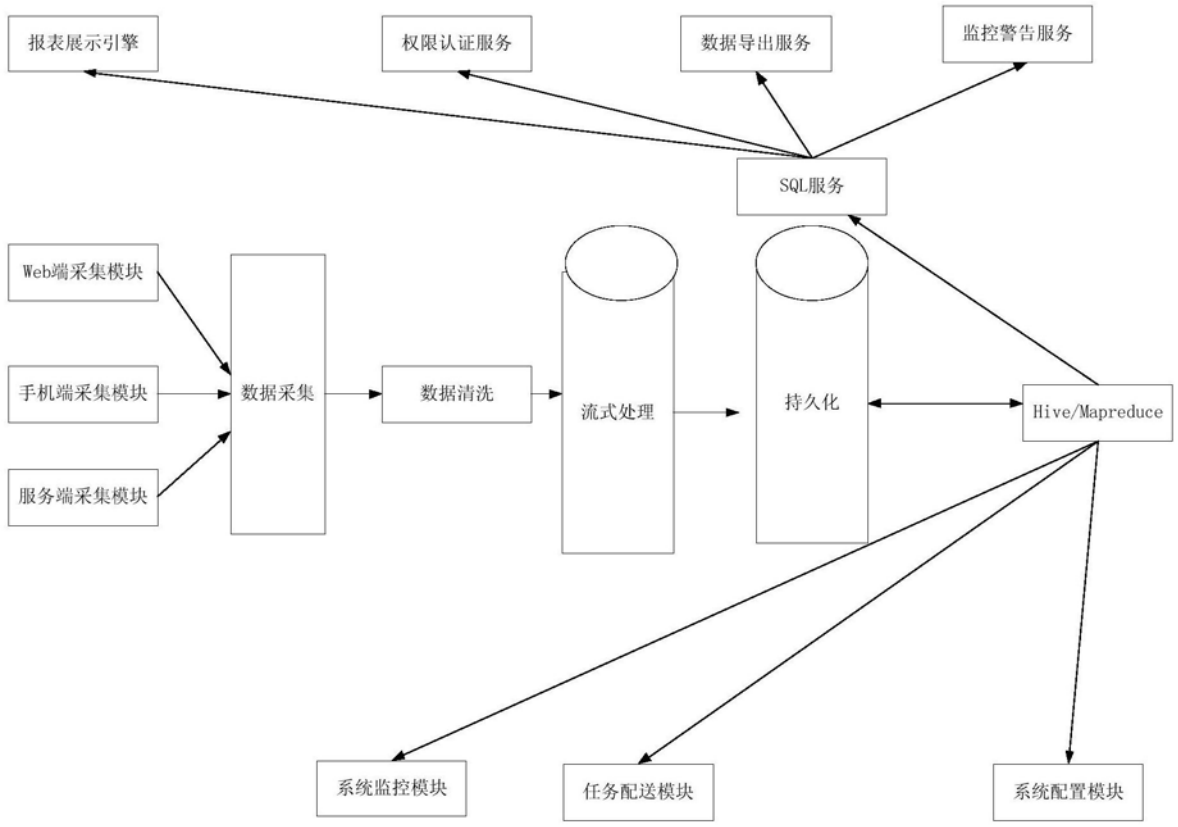

报表展示引擎
权限认证服务
数据导出服务
监控警告服务
SQL服务
Web端采集模块
手机端采集模块
服务端采集模块
数据采集
数据清洗
流式处理
持久化
Hive/Mapreduce
系统监控模块
任务配送模块
系统配置模块

图3

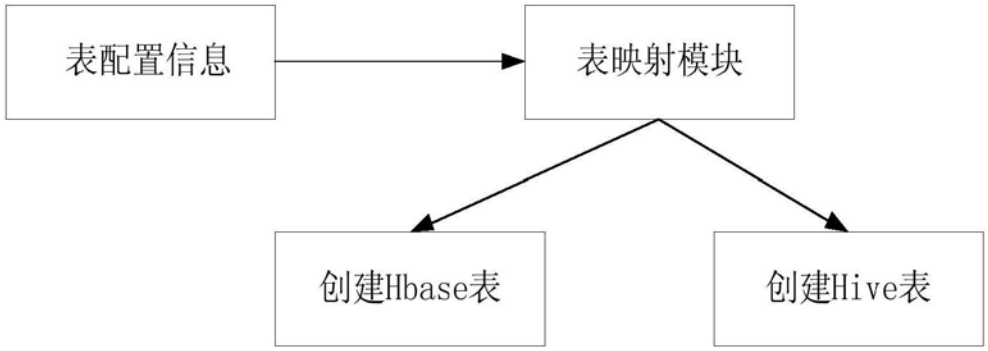

表配置信息
表映射模块
创建Hbase表
创建Hive表

图4

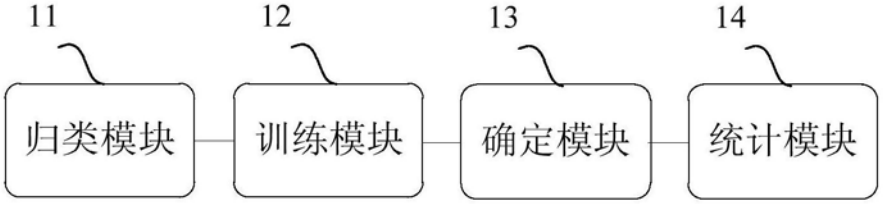

11
12
13
14
归类模块
训练模块
确定模块
统计模块

图5

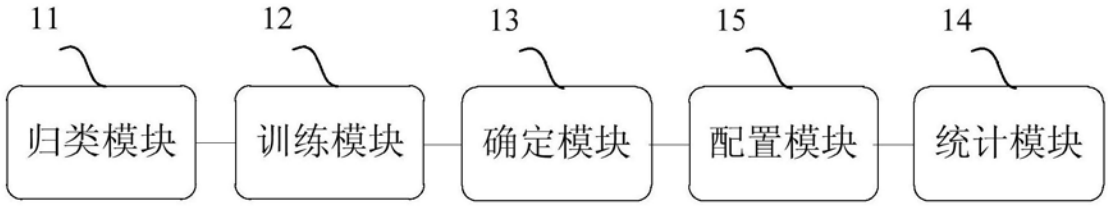


图6