(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 112650739 A
(43)申请公布日 2021.04.13

(21)申请号 202011530168.5

(22)申请日 2020.12.22

(71)申请人 煤炭科学研究总院
地址 100013 北京市朝阳区和平街13号煤炭大厦12层

(72)发明人 苏上海 张晓霞 施展 李昊 袁慧 王雅琨

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 韩海花

(51)Int.Cl.
G06F 16/215(2019.01)
G06F 16/27(2019.01)
G06F 16/28(2019.01)

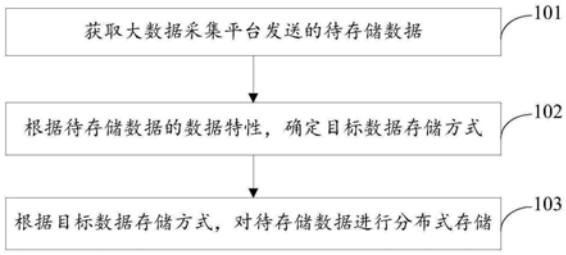
权利要求书2页 说明书7页 附图2页

(54)发明名称

煤矿数据中台的数据存储处理方法和装置

(57)摘要

本申请提出一种煤矿数据中台的数据存储处理方法和装置，其中，煤矿数据中台包括大数据采集平台和大数据基础平台，该方法应用于大数据基础平台，包括：获取大数据采集平台发送的待存储数据；根据待存储数据的数据特性，确定目标分布式存储方式；根据目标分布式存储方式，对待存储数据进行分布式存储。由此，针对数据特性不同的数据，选择不同的数据存储方式进行分布式存储，不仅提高了存储效率，而且提高了查询效率。



CN 112650739 A

1.一种煤矿数据中台的数据存储处理方法,其特征在于,所述煤矿数据中台包括大数据采集平台和大数据基础平台,所述方法应用于所述大数据基础平台,包括:

获取所述大数据采集平台发送的待存储数据;

根据所述待存储数据的数据特性,确定目标数据存储方式;

根据所述目标数据存储方式,对所述待存储数据进行分布式存储。

2.如权利要求1所述的方法,其特征在于,还包括:

获取待处理数据;

对所述待处理数据进行分布式计算。

3.如权利要求2所述的方法,其特征在于,所述对所述待处理数据进行分布式计算,包括:

对所述待处理数据进行预处理;和/或,

对所述待处理数据进行数据清洗;和/或,

对所述待处理数据进行数据建模。

4.如权利要求2所述的方法,其特征在于,其中,所述分布式计算包括离线批处理和在线实时处理,所述对所述待处理数据进行分布式计算,包括:

根据所述待处理数据对应的业务需求,从所述离线批处理和所述在线实时处理中,选出目标处理方式;

根据所述目标处理方式,对所述待处理数据进行分布式处理。

5.如权利要求1-4任一所述的方法,其特征在于,所述数据特性包括工业协议、采集频率、数据格式、数据分析方式中的至少一种。

6.一种煤矿数据中台的数据存储处理装置,其特征在于,所述煤矿数据中台包括大数据采集平台和大数据基础平台,所述装置应用于所述大数据基础平台,包括:

第一获取模块,用于获取所述大数据采集平台发送的待存储数据;

确定模块,用于根据所述待存储数据的数据特性,确定目标数据存储方式;

存储模块,用于根据所述目标数据存储方式,对所述待存储数据进行分布式存储。

7.如权利要求6所述的装置,其特征在于,还包括:

第二获取模块,用于获取待处理数据;

计算模块,用于对所述待处理数据进行分布式计算。

8.如权利要求7所述的装置,其特征在于,所述计算模块,具体用于:

对所述待处理数据进行预处理;和/或,

对所述待处理数据进行数据清洗;和/或,

对所述待处理数据进行数据建模。

9.如权利要求7所述的装置,其特征在于,所述分布式计算包括离线批处理和在线实时处理,所述计算模块,具体用于:

根据所述待处理数据对应的业务需求,从所述离线批处理和所述在线实时处理中,选出目标处理方式;

根据所述目标处理方式,对所述待处理数据进行分布式处理。

10.如权利要求6-9任一所述的装置,其特征在于,所述数据特性包括工业协议、采集频率、数据格式、数据分析方式中的至少一种。

11.一种计算机设备,其特征在于,包括处理器和存储器;

其中,所述处理器通过读取所述存储器中存储的可执行程序代码来运行与所述可执行程序代码对应的程序,以用于实现如权利要求1-5中任一所述的煤矿数据中台的数据存储处理方法。

12.一种非临时性计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现如权利要求1-5中任一所述的煤矿数据中台的数据存储处理方法。

煤矿数据中台的数据存储处理方法和装置

技术领域

[0001] 本申请涉及数据处理技术领域，尤其涉及一种煤矿数据中台的数据存储处理方法和装置。

背景技术

[0002] 随着科学技术的发展，煤矿的现代化程度越来越高，煤矿中系统数量也逐渐增加，数据源的数量也在增加，采集的煤矿数据量也比较大，通过信息的采集，可以较好的掌握井下生产各个环节的运行情况。因此，如何提高煤矿数据的存储效率和处理效率是亟待解决的问题。

发明内容

[0003] 本申请提出一种煤矿数据中台的数据存储处理方法和装置。

[0004] 本申请一方面实施例提出了一种煤矿数据中台的数据存储处理方法，所述煤矿数据中台包括大数据采集平台和大数据基础平台，所述方法应用于所述大数据基础平台，包括：

[0005] 获取所述大数据采集平台发送的待存储数据；

[0006] 根据所述待存储数据的数据特性，确定目标分布式存储方式；

[0007] 根据所述目标分布式存储方式，对所述待存储数据进行分布式存储。

[0008] 本申请实施例的煤矿数据中台的数据存储处理方法，应用于煤矿数据中台的大数据基础平台，通过获取大数据采集平台发送的待存储数据，根据待存储数据的数据特性，确定目标数据存储方式，并根据目标数据存储方式，对待存储数据进行分布式存储。由此，针对数据特性不同的数据，选择不同的数据存储方式进行分布式存储，不仅提高了存储效率，而且提高了查询效率。

[0009] 在本申请一方面实施例一种可能的实现方式中，该方法还包括：获取待处理数据；

[0010] 对所述待处理数据进行分布式计算。

[0011] 在本申请一方面实施例一种可能的实现方式中，所述对所述待处理数据进行分布式计算，包括：

[0012] 对所述待处理数据进行预处理；和/或，

[0013] 对所述待处理数据进行数据清洗；和/或，

[0014] 对所述待处理数据进行数据建模。

[0015] 在本申请一方面实施例一种可能的实现方式中，所述分布式计算包括离线批处理和在线实时处理，所述对所述待处理数据进行分布式计算，包括：

[0016] 根据所述待处理数据对应的业务需求，从所述离线批处理和所述在线实时处理中，选出目标处理方式；

[0017] 根据所述目标处理方式，对所述待处理数据进行分布式处理。

[0018] 在本申请一方面实施例一种可能的实现方式中，所述数据特性包括工业协议、采

集频率、数据格式、数据分析方式中的至少一种。

[0019] 本申请另一方面实施例提出了一种煤矿数据中台的数据存储处理装置，所述煤矿数据中台包括大数据采集平台和大数据基础平台，所述装置应用于所述大数据基础平台，包括：

[0020] 第一获取模块，用于获取所述大数据采集平台发送的待存储数据；

[0021] 确定模块，用于根据所述待存储数据的数据特性，确定目标数据存储方式；

[0022] 存储模块，用于根据所述目标数据存储方式，对所述待存储数据进行分布式存储。

[0023] 在本申请另一方面实施例一种可能的实现方式中，该装置还可包括：

[0024] 第二获取模块，用于获取待处理数据；

[0025] 计算模块，用于对所述待处理数据进行分布式计算。

[0026] 在本申请另一方面实施例一种可能的实现方式中，所述计算模块，具体用于：

[0027] 对所述待处理数据进行预处理；和/或，

[0028] 对所述待处理数据进行数据清洗；和/或，

[0029] 对所述待处理数据进行数据建模。

[0030] 在本申请另一方面实施例一种可能的实现方式中，所述分布式计算包括离线批处理和在线实时处理，所述计算模块，具体用于：

[0031] 根据所述待处理数据对应的业务需求，从所述离线批处理和所述在线实时处理中，选出目标处理方式；

[0032] 根据所述目标处理方式，对所述待处理数据进行分布式处理。

[0033] 在本申请另一方面实施例一种可能的实现方式中，所述数据特性包括工业协议、采集频率、数据格式、数据分析方式中的至少一种。

[0034] 本申请实施例的煤矿数据中台的数据存储处理装置，应用于煤矿数据中台的大数据基础平台，通过获取大数据采集平台发送的待存储数据，根据待存储数据的数据特性，确定目标数据存储方式，并根据目标数据存储方式，对待存储数据进行分布式存储。由此，针对数据特性不同的数据，选择不同的数据存储方式进行分布式存储，不仅提高了存储效率，而且提高了查询效率。

[0035] 本申请另一方面实施例提出了一种计算机设备，包括处理器和存储器；

[0036] 其中，所述处理器通过读取所述存储器中存储的可执行程序代码来运行与所述可执行程序代码对应的程序，以用于实现如上述一方面实施例所述的煤矿数据中台的数据存储处理方法。

[0037] 本申请另一方面实施例提出了一种非临时性计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如上述一方面实施例所述的煤矿数据中台的数据存储处理方法。

[0038] 本申请附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本申请的实践了解到。

附图说明

[0039] 本申请上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

[0040] 图1为本申请实施例提供的一种煤矿数据中台的数据存储处理方法的流程示意图；

[0041] 图2为本申请实施例提供的另一种煤矿数据中台的数据存储处理方法的流程示意图；

[0042] 图3为本申请实施例提供的一种分布式计算方式的示意图；

[0043] 图4为本申请实施例提供的一种煤矿数据中台的数据存储处理装置的结构示意图。

具体实施方式

[0044] 下面详细描述本申请的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，旨在用于解释本申请，而不能理解为对本申请的限制。

[0045] 下面参考附图描述本申请实施例的煤矿数据中台的数据存储处理方法和装置。

[0046] 图1为本申请实施例提供的一种煤矿数据中台的数据存储处理方法的流程示意图。

[0047] 本实施例中，煤矿数据中台包括大数据采集平台和大数据基础平台。其中大数据平台可对接煤矿的各系统，用于采集各个数据源的数据，大数据基础平台可对大数据平台中的数据进行存储和处理。

[0048] 煤矿数据中台还可包括大数据管理平台和大数据分析平台，其中，大数据管理平台可对数据进行质量管理、元数据管理、安全管理等，大数据分析平台用于对煤矿数据进行分析，可进行柱状图、折线图、条形图等可视化分析，也可进行多维数据分析。

[0049] 本申请实施例的煤矿数据中台的数据存储处理方法，应用于煤矿数据中台的大数据基础平台，以实现对大数据采集平台中的数据进行分布式存储和处理。

[0050] 如图1所示，该煤矿数据中台的数据存储处理方法包括：

[0051] 步骤101，获取大数据采集平台发送的待存储数据。

[0052] 本实施例中，可对大数据采集平台采集的数据进行存储，具体地，大数据采集平台可将采集的数据直接发送给大数据基础平台，或者在采集过程中创建元数据及收集数据的元数据，发送给大数据基础平台。由此，大数据基础平台可获取大数据采集平台发送的待存储数据。

[0053] 步骤102，根据待存储数据的数据特性，确定目标数据存储方式。

[0054] 为了提高存储和查询效率，本实施例中，可先确定待存储数据的特性，根据待存储数据的数据特性，从多种数据存储方式中选出目标数据存储方式。

[0055] 本实施例中，可针对不同数据的数据特性，选择不同的数据存储方式。其中，数据特性包括但不限于工业协议、采集频率、数据格式、数据分析方式等维度或类型。

[0056] 比如，数据格式可以是关系型的，也可以是键值型等。以煤矿工作面数据为例，煤矿工作面数据主要包括工作面传感器数据和监控视频数据。工作面传感器数据是点状时序数据，每个传感器只监测单个设备的某个或某几个状态信息，且只在数据发生变化时产生数据，因此在存储工作面传感器数据时，需要支持快速查询某个时间段内某个设备的数据。

[0057] 对于监控视频数据，在存储时，除了要保证视频文件本身的存储外，还可进行视频

文件元数据的提取、存储和检索等。

[0058] 井下设备中还有一些高频数据，对于高频数据需要使用支撑高速写入的存储引擎进行存储。

[0059] 步骤103，根据目标数据存储方式，对待存储数据进行分布式存储。

[0060] 在获取目标数据存储方式后，根据目标数据存储方式，利用多台存储服务器，对待存储数据进行分布式存储。比如，将关系型数据可存储到MySql数据库中，键值型数据可存储到Redis、HBase中等。

[0061] 本实施例中，分布式存储采用可扩展的系统结构，使用多台存储服务器存储数据，可以横向扩展，适用于存储海量的煤矿数据。

[0062] 在进行分布式存储时，可根据待存储数据的类型，选择相应的分布式存储引擎进行存储。

[0063] 比如，对于关系型数据，可采用关系型分布式存储引擎，将数据存储到分布式数据环境内；对于键值型数据，可采用键值型分布式存储引擎，将数据存储到分布式数据环境内；对于全文检索型数据，可采用全文检索型分布式引擎，将数据存储到分布式数据环境内。

[0064] 其中，关系型数据分布式存储，可支持数据的分析统计等交互式查询；键值型数据分布式存储，可支持根据Key值快速找到Value值的查询；全文检索型数据分布式存储，可支持中英文分词，能够对数据中的任意字段内容进行快速检索查询。

[0065] 本实施例中，根据待存储数据的类型，选择合适的存储引擎进行存储，不仅提高了存储效率，而且提高了查询效率。

[0066] 本申请实施例中，煤矿数据中台包括大数据采集平台和大数据基础平台，该方法应用于大数据基础平台，通过获取大数据采集平台发送的待存储数据，根据待存储数据的数据特性，确定目标数据存储方式，并根据目标数据存储方式，对待存储数据进行分布式存储。由此，针对数据特性不同的数据，选择不同的数据存储方式进行分布式存储，不仅提高了存储效率，而且提高了查询效率。

[0067] 在本申请的一个实施例中，大数据基础平台还可对数据进行分布式处理。下面结合图2进行说明，图2为本申请实施例提供的另一种煤矿数据中台的数据存储处理方法的流程示意图。

[0068] 如图2所示，该方法还包括：

[0069] 步骤201，获取待处理数据。

[0070] 本实施例中，大数据基本平台除了可以对数据进行分布式存储，还可对数据进行分布式处理。本实施例中，可从分布式存储的数据中获取待处理数据。具体地，在获取待处理数据时，可以从数据库系统、日志系统等获取待处理数据。

[0071] 步骤202，对待处理数据进行分布式计算。

[0072] 本实施例中，分布式计算构建在分布式存储之上，大数据基本平台可对分布式存储的数据进行分布式计算。其中，分布式计算是一种计算方法，和集中式计算是相对的。分布式计算将大量的数据分割成多个小块，由多台计算机分工计算，然后将结果汇总。这样可以节约整体计算时间，大大提高计算效率。

[0073] 在一种可能的实现方式中，在待处理数据进行分布式计算时，可对待处理数据进

行预处理、数据清洗和数据建模中一种或多种处理。也就是说，分布式计算可应用数据预处理、数据清洗和数据建模等。

[0074] 为了提高分布式计算效率，作为另一种可能的实现方式，分布式计算可包括离线批处理和在线实时处理，在对待处理数据进行分布式计算时，可根据待处理数据对应的业务需求，从离线批处理和在线实时处理两种处理方式中，选出目标处理方式，利用目标处理方式，对待处理数据进行分布式处理。

[0075] 其中，分布式离线数据的批处理引擎，需支持良好的扩展性、容错性及高吞吐率；分布式数据的实时处理引擎，需确保数据处理的及时性，并保证对每条数据处理不重复不漏。

[0076] 比如，对于实时性要求不高的数据可采用离线批处理的方式，比如地质数据等，对于实时性要求较高的数据比如采集的煤矿工作面上数据，可采用在线实时处理。

[0077] 本实施例中，针对不同类型的数据，采用相应的分布式处理方式进行处理，不仅满足了实际需求，也提高了处理效率。

[0078] 下面结合图3，对上述离线批处理和在线实时处理两种分布式计算方式进一步说明。图3为本申请实施例提供的一种分布式计算方式的示意图。

[0079] 如图3所示，RDBMS(Relational Database Management System，关系数据库管理系统)和NOSQLs(Not Only Sqls，非关系型数据库)，可通过流式数据的处理平台DataHub，向离线批处理分发流式数据。

[0080] 其中，RDBMS是包括相互联系的逻辑组织和存取这些数据的数据库管理系统软件。RDBMS用于管理关系数据库，并将数据逻辑组织的系统。其中，数据库是表格式的，因此存储在表的行和列中，它们之间很容易关联协作存储，提取数据很方便。而NOSQLs数据库则与RDBMS相反，它是大块的组合在一起，通常存储在数据集中，像文档、键值对或者图结构等。

[0081] 在离线批处理时，可利用HDFS(Hadoop Distributed File System，Hadoop分布式文件系统)、Hive、Spark SQL、MR对从RDBMS、NOSQLs等获取的数据进行分布式处理。

[0082] 其中，Hive是基于Hadoop的一个数据仓库工具，用来进行数据提取、转化、加载，这是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制；Spark SQL是Spark的一个模块，主要用于进行结构化数据的处理；MR是分布式数据处理工具。

[0083] Spark是一种与Hadoop相似的开源集群计算环境，是UC Berkeley AMP lab(加州大学伯克利分校的AMP实验室)开源的类Hadoop MapReduce的通用并行框架，Spark拥有Hadoop MapReduce所具有的优点，且有更好的性能表现。

[0084] 图3中，离线批处理结果可用于机器学习、联机分析处理引擎等。对于机器学习，可采用Spark MLilb进行机器学习。其中，Spark MLilb是一个包含通用机器学习功能的包，其包含分类、聚类、回归等，还包含模型评估和数据导入。

[0085] 对于联机分析处理引擎，可将离线批处理结果存入HBase，采用开源的分布式分析引擎Apache Kylin进行数据查询。其中，HBase是一个分布式的、面向列的开源数据库，也是一个结构化数据的分布式存储系统。

[0086] 日志系统Logs可通过为日志搜集系统Flume，向在线实时处理传输日志。其中，Flume是一个高可用的、高可靠的、分布式的海量日志采集、聚合和传输的系统，Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，

并写到各种数据接受方(可定制)的能力。

[0087] 在进行在线实时处理时,可通过Kafka、Spark Streaming、Flink等进行在线实时处理。

[0088] 其中,Kafka是由Apache软件基金会开发的一个开源流处理平台,由Scala和Java编写,Kafka可以处理动作流数据。

[0089] Spark Streaming是Spark核心应用编程接口的一个扩展,可以实现高吞吐量的、具备容错机制的实时流数据处理。Spark Streaming可接收Kafka、Flume等各种来源的实时输入数据,并进行处理。

[0090] 在线实时处理的结果,可保存在Business RDBMS(业务关系数据库管理系统)和Redis(Remote Dictionary Server,远程字典服务)。其中,Redis是一个key-value存储系统。

[0091] Flink是由Apache软件基金会开发的开源流处理框架,其核心是用Java和Scala编写的分布式流数据流引擎。

[0092] 另外,通过Kafka处理后的数据,也可以进行批处理。

[0093] 为了实现上述实施例,本申请还提出一种煤矿数据中台的数据存储处理装置。图4为本申请实施例提供的一种煤矿数据中台的数据存储处理装置的结构示意图。

[0094] 本申请实施例中,煤矿数据中台可包括大数据采集平台和大数据基础平台。其中,大数据采集平台可对接煤矿的各系统,比如监控系统、生产系统、业务系统等,以采集各数据源中的数据。本申请实施例的煤矿数据中台的数据存储处理装置应用于大数据基础平台,以实现对大数据采集平台采集的数据进行存储和处理。

[0095] 如图4所示,该煤矿数据中台的数据存储处理装置300包括:第一获取模块310、确定模块320、存储模块330。

[0096] 第一获取模块310,用于获取大数据采集平台发送的待存储数据;

[0097] 确定模块320,用于根据待存储数据的数据特性,确定目标数据存储方式;

[0098] 存储模块330,用于根据目标数据存储方式,对待存储数据进行分布式存储。

[0099] 在本申请实施例一种可能的实现方式中,该装置还可包括:

[0100] 第二获取模块,用于获取待处理数据;

[0101] 计算模块,用于对待处理数据进行分布式计算。

[0102] 在本申请实施例一种可能的实现方式中,计算模块,具体用于:

[0103] 对待处理数据进行预处理;和/或,

[0104] 对待处理数据进行数据清洗;和/或,

[0105] 对待处理数据进行数据建模。

[0106] 在本申请实施例一种可能的实现方式中,述分布式计算包括离线批处理和在线实时处理,计算模块,具体用于:

[0107] 根据待处理数据对应的业务需求,从离线批处理和在线实时处理中,选出目标处理方式;

[0108] 根据目标处理方式,对待处理数据进行分布式处理。

[0109] 在本申请实施例一种可能的实现方式中,数据特性包括工业协议、采集频率、数据格式、数据分析方式中的至少一种。

[0110] 需要说明的是，上述对煤矿数据中台的数据存储处理方法实施例的解释说明，也适用于该实施例的煤矿数据中台的数据存储处理装置，故在此不再赘述。

[0111] 本申请实施例的煤矿数据中台的数据存储处理装置，煤矿数据中台包括大数据采集平台和大数据基础平台，该装置应用于大数据基础平台，通过获取大数据采集平台发送的待存储数据，根据待存储数据的数据特性，确定目标数据存储方式，并根据目标数据存储方式，对待存储数据进行分布式存储。由此，针对数据特性不同的数据，选择不同的数据存储方式进行分布式存储，不仅提高了存储效率，而且提高了查询效率。

[0112] 为了实现上述实施例，本申请实施例还提出一种计算机设备，包括处理器和存储器；

[0113] 其中，处理器通过读取存储器中存储的可执行程序代码来运行与所述可执行程序代码对应的程序，以用于实现如上述实施例所述的煤矿数据中台的数据存储处理方法。

[0114] 为了实现上述实施例，本申请实施例还提出一种非临时性计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如上述实施例所述的煤矿数据中台的数据存储处理方法。

[0115] 在本说明书的描述中，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。在本申请的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

[0116] 尽管上面已经示出和描述了本申请的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本申请的限制，本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

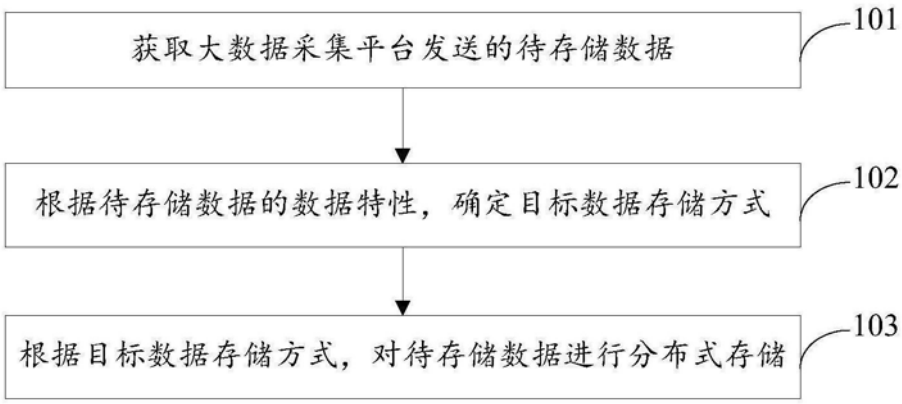

获取大数据采集平台发送的待存储数据
101
根据待存储数据的数据特性，确定目标数据存储方式
102
根据目标数据存储方式，对待存储数据进行分布式存储
103

图1

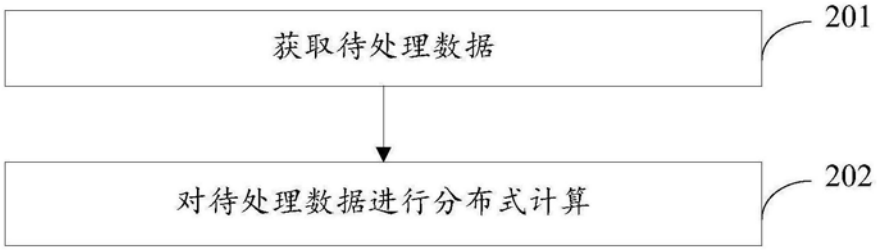

获取待处理数据
201
对待处理数据进行分布式计算
202

图2

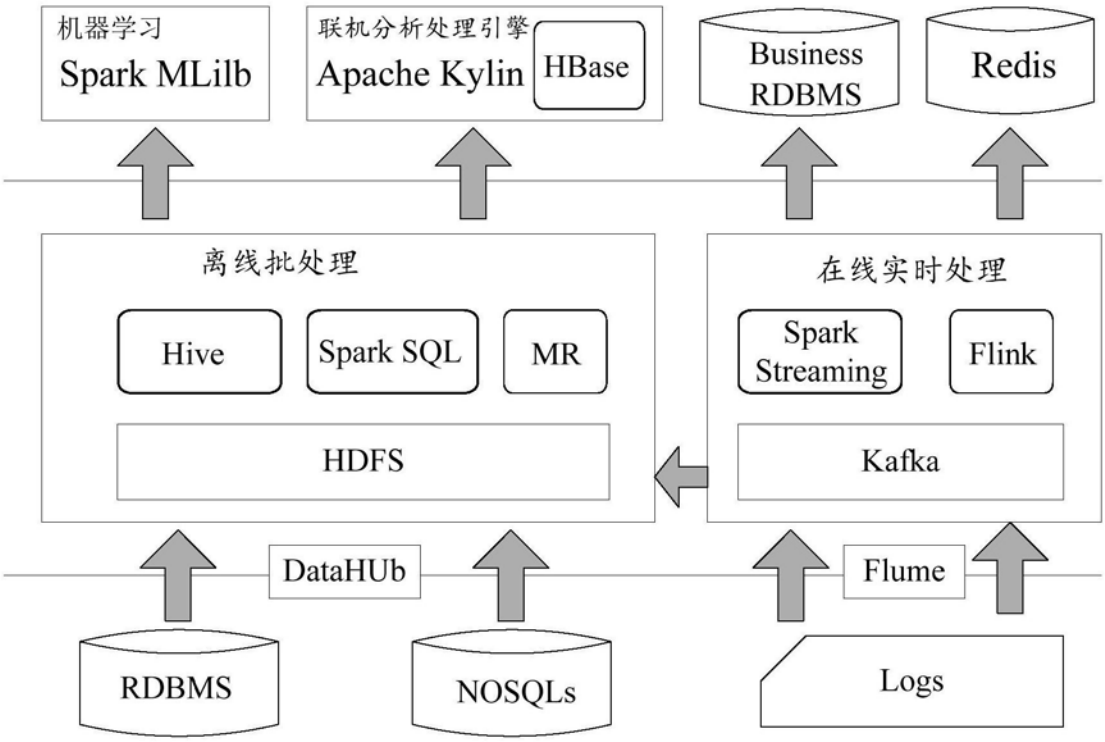

机器学习
Spark MLilb
联机分析处理引擎
Apache Kylin
HBase
Business RDBMS
Redis
离线批处理
Hive
Spark SQL
MR
HDFS
在线实时处理
Spark Streaming
Flink
Kafka
DataHUb
Flume
RDBMS
NOSQLs
Logs

图3

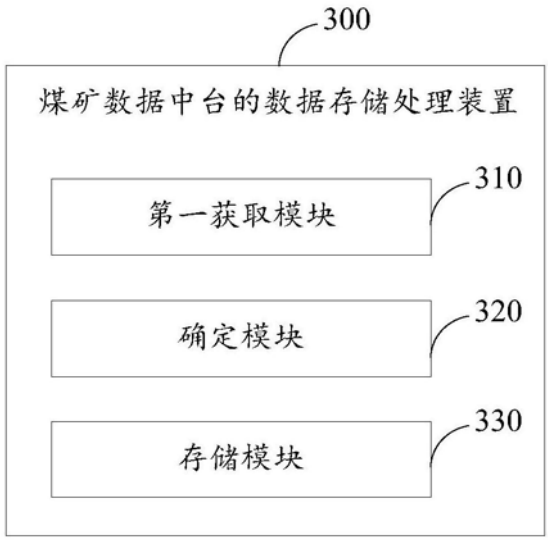


图4