

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2021 年 9 月 30 日 (30.09.2021)



(10) 国际公布号
WO 2021/189695 A1

(51) 国际专利分类号:
G06F 16/215 (2019.01)

(21) 国际申请号: PCT/CN2020/099088

(22) 国际申请日: 2020 年 6 月 30 日 (30.06.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010222406.X 2020年3月25日 (25.03.2020) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 潘翔(PAN, Xiang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理有限公司(普通合伙)(SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,

(54) **Title:** DISTRIBUTED DATABASE DYNAMIC EXPANSION METHOD AND APPARATUS, AND DEVICE AND STORAGE MEDIUM

(54) 发明名称: 分布式的数据库动态扩容方法、装置、设备及存储介质

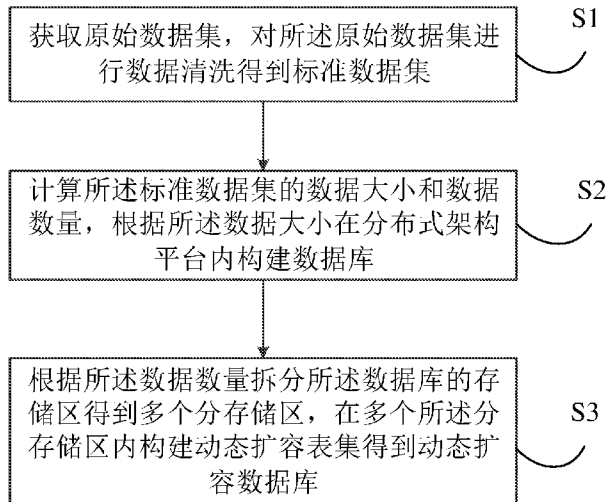


图 1

S1 Acquire an original data set, and perform data cleaning on the original data set to obtain a standard data set

S2 Calculate the data size and data quantity of the standard data set, and construct a database in a distributed architecture platform according to the data size

S3 Split a storage area of the database according to the data quantity to obtain multiple sub-storage areas, and construct dynamic expansion table sets in the multiple sub-storage areas to obtain a dynamically expanded database

(57) **Abstract:** A distributed database dynamic expansion method and apparatus, and an electronic device and a computer-readable storage medium, which relate to the technical field of cloud storage. The method comprises: acquiring an original data set, and performing data cleaning on the original data set to obtain a standard data set (S1); calculating the data size and data quantity of the standard data set, and constructing a database in a distributed architecture platform according to the data size (S2); and splitting a storage area of the database according to the data quantity to obtain multiple sub-storage areas, and constructing dynamic expansion table sets in the multiple sub-storage areas to obtain a dynamically expanded database (S3). According to the method, the problems of low storage efficiency, and the uneven allocation of storage memory during data storage can be solved.

MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种分布式的数据库动态扩容方法、装置、电子设备以及计算机可读存储介质, 涉及云存储技术领域。所述方法包括: 获取原始数据集, 对原始数据集进行数据清洗得到标准数据集(S1); 计算所述标准数据集的数据大小和数据数量, 根据所述数据大小在分布式架构平台内构建数据库(S2); 根据所述数据数量拆分所述数据库的存储区得到多个分存储区, 在多个所述分存储区内构建动态扩容表集得到动态扩容数据库(S3)。所述方法可以解决存储效率低下、存储数据时存储内存分配不均匀的问题。

分布式的数据库动态扩容方法、装置、设备及存储介质

[0001] 本申请要求于2020年3月25日提交中国专利局、申请号为CN202010222406.X、发明名称为“分布式的数据库动态扩容方法、装置、设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

[0002] 本申请涉及云存储技术领域，尤其涉及一种分布式的数据库动态扩容的方法、装置、设备及存储介质。

背景技术

[0003] 随着大数据兴起，数据库存储的数据量也越来越大，发明人意识到当存储的数据量达到千万级别甚至上亿级别时，普通的关系型数据库应对存储时会出现明显的存储效率低下、存储数据时存储内存分配不均匀的现象。

发明概述

技术问题

[0004] 本申请提供一种分布式的数据库动态扩容方法、装置、电子设备及计算机可读存储介质，其主要目的在于解决存储效率低下、存储数据时存储内存分配不均匀的问题。

问题的解决方案

技术解决方案

[0005] 为实现上述目的，本申请提供一种分布式的数据库动态扩容方法，包括：

[0006] 对原始数据集进行数据清洗得到标准数据集；

[0007] 计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库；

[0008] 根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。

[0009] 为了解决上述问题，本申请还提供一种分布式的数据库动态扩容装置，所述装置包括：

- [0010] 数据清洗模块，用于对原始数据集进行数据清洗得到标准数据集；
- [0011] 数据库构建模块，用于计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库；
- [0012] 动态扩容模块，用于根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。
- [0013] 为了解决上述问题，本申请还提供一种电子设备，所述电子设备包括：
- [0014] 存储器，存储至少一个指令；及
- [0015] 处理器，执行所述存储器中存储的指令时实现如下步骤：
- [0016] 对原始数据集进行数据清洗得到标准数据集；
- [0017] 计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库；
- [0018] 根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。
- [0019] 为了解决上述问题，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一个指令，所述至少一个指令被电子设备中的处理器执行时实现如下步骤：
- [0020] 对原始数据集进行数据清洗得到标准数据集；
- [0021] 计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库；
- [0022] 根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。

发明的有益效果

有益效果

- [0023] 本申请先计算数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库，由于数据库是基于分布式架构平台和数据大小搭建的，所以数据库更具有针对性，同时根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库，因为整个数据库扩容都依赖于数据数量和数据大小，所以每个分存储区的存

储内存分配较为均匀，且当需要进行数据存储时同时存储至多个动态扩容表集内，提高了存储效率。因此本申请提出的分布式的数据库动态扩容方法、装置、电子设备及计算机可读存储介质，可以解决存储效率低下、存储数据时存储内存分配不均匀的问题。

对附图的简要说明

附图说明

- [0024] 图1为本申请一实施例提供的分布式的数据库动态扩容方法的流程示意图；
- [0025] 图2为本申请一实施例提供的分布式的数据库动态扩容方法中S2的流程示意图；
- [0026] 图3为本申请一实施例提供的分布式的数据库动态扩容方法中S3的流程示意图；
- [0027] 图4为本申请一实施例提供的分布式的数据库动态扩容装置的模块示意图；
- [0028] 图5为本申请一实施例提供的分布式的数据库动态扩容方法的电子设备的内部结构示意图；
- [0029] 本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

发明实施例

本发明的实施方式

- [0030] 应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。
- [0031] 本申请提供一种分布式的数据库动态扩容方法。参照图1所示，为本申请一实施例提供的分布式的数据库动态扩容方法的流程示意图。该方法可以由一个电子设备执行，该电子设备可以由软件和/或硬件实现。
- [0032] 在本实施例中，分布式的数据库动态扩容方法包括：
- [0033] S1、获取原始数据集，对所述原始数据集进行数据清洗得到标准数据集。
- [0034] 本申请的目的是根据原始数据集的数据量对预构建的数据库进行动态扩容，当完成动态扩容后的数据库不管是存储其他数据集时，还是存储完成后查询其他数据集时，速度都会更快更高效。
- [0035] 所述原始数据集的数据量一般非常庞大，目的是为了最大化动态扩容数据库，

防止后续动态扩容后的数据库存储其他数据集时，由于扩容的存储区数量不够导致需要再次动态扩容的现象发生，如利用爬虫技术爬取某大型购物网站的用户交易额数据，所述用户交易额数据即为原始数据集，且大型购物网站的用户交易额数据的量级一般都为亿级，满足数据量庞大的要求。

[0036] 进一步地，如上述利用爬虫技术爬取得到用户交易额数据，但由于爬取过程会爬取到不符合要求的数据，因此对原始数据集进行数据清洗是必不可少的步骤。

[0037] 详细地，对所述原始数据集进行数据清洗得到标准数据集，包括：对所述原始数据集进行缺失值填充、异常值去除、属性区间划分得到所述标准数据集。

[0038] 进一步地，所述缺失值填充的方法众多，如填充固定值、填充均值、填充众数等。

[0039] 所述属性区间划分是依据原始数据集内数据的属性不同，根据数据的属性对原始数据集内数据进行划分。所述属性区间划分包括：首先对原始数据集内数据进行排序，同时利用等宽的区间对排序后的原始数据集进行划分得到多个区间的原始数据集，分别计算每个区间的原始数据集的均值，并用均值替代每个区间内的原始数据集，完成属性区间划分。如上述用户交易额数据共有1亿个数据量，对每笔交易额进行排序得到从小到大的用户交易额数据，按照每100元一个区间对从小到大排序的用户交易额数据进行划分，分别计算每个区间内的交易额的平均值，并替代每个区间的每笔交易额，如0-100元区间的交易额为[1.2, 3.7, 4.8, 10, 99]，则计算出平均值为23.74，则将23.74替代[1.2, 3.7, 4.8, 10, 99]得到[23.74, 23.74, 23.74, 23.74, 23.74]。

[0040] S2、计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库。

[0041] 所述分布式架构平台可采用当前已公开的hadoop、Pregel、Spark、Storm等，本申请中分布式架构平台是对数据库进行动态扩容的基础，所有的动态扩容都是在分布式架构平台内完成，因为多数分布式架构平台具有高可靠性的特点，在动态扩容中不容易产生崩溃的现象，其次分布式架构平台内包括多个节点，使得在动态扩容时保证多个节点之间具有联系且可以自由移动交换数据，从而

保证各个节点的动态平衡。

[0042] 详细地，所述S2可参阅图S2的流程示意图所示，包括：

[0043] S21、统计所述分布式架构平台内各存储节点的存储量；

[0044] S22、根据所述标准数据集的数据大小和所述存储量计算得到存储节点数量；

[0045] S23、根据所述存储节点数量从所述分布式架构平台内选择节点构建得到所述数据库。

[0046] 例如上述利用爬虫技术爬取并数据清理得到的标准用户交易额数据，通过计算得到标准用户交易额数据的数据大小为200GB，而分布式架构平台内各存储节点的存储量为1GB，则计算得到需要200个存储节点，则从所述分布式架构平台内选择200个存储节点构建为数据库。

[0047] S3、根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。

[0048] 详细地，所述S3可参阅图S3的流程示意图所示包括：

[0049] S31、计算所述数据数量以指数形式表达的底数和指数；

[0050] S32、根据所述底数划分所述存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表；

[0051] S33、根据预先设定的命名规则，对所述分存储区和所述动态扩容表命名得到动态扩容数据库。

[0052] 如上述200个存储节点构建成的数据库，假设上述标准用户交易额数据的数据数量为56800235584，由于56800235584以指数形式表达就是62的6次方，因此可将上述构建的数据库划分为62个分存储区，然后在每一个分存储区内建立相同指数的62张动态扩容表。

[0053] 进一步地，所述预先设定的命名规则可多种多样，本申请较佳实施例中，所述分存储区和所述动态扩容表的命名规则均使用52个大小写字母和0-9的十位数字组成的字符串，如分存储区命名为db_0、db_1、…、db_zz，动态扩容表名命名为table_0、table_1、…、table_zz，所有的分存储区和所有的动态扩容表作为一个整体称为动态扩容数据库。

[0054] 当完成所述动态扩容数据库的建立后，若需要在所述动态扩容数据库内存储数

据集，则存储步骤包括：

- [0055] 获取待存储数据集，根据所述分存储区的数量将所述待存储数据集分成相同数量的分待存储数据集，将所述分待存储数据集依次存储至所述动态扩容表集内。
- [0056] 当用户需要访问待存储数据集时，通过命名规则依次访问待存储数据集。同时待存储数据集可为所述原始数据集，也可以为新的数据集，如利用上述用户交易额数据进行动态扩容，当动态扩容完成后将用户交易额数据存储至动态扩容后的数据库内，也可以重新存储新的数据集，如用户在某贷款平台的贷款金额数据集。
- [0057] 如图4所示，是本申请分布式的数据库动态扩容装置的模块示意图。
- [0058] 本申请所述分布式的数据库动态扩容装置100可以安装于电子设备中。根据实现的功能，所述分布式的数据库动态扩容装置100可以包括数据清洗模块101、数据库构建模块102、动态扩容模块103。本发所述模块也可以称之为单元，是指一种能够被电子设备处理器所执行，并且能够完成固定功能的一系列计算机程序段，其存储在电子设备的存储器中。
- [0059] 在本实施例中，关于各模块/单元的功能如下：
- [0060] 所述数据清洗模块101，用于对原始数据集进行数据清洗得到标准数据集。
- [0061] 所述数据库构建模块102，用于计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库。
- [0062] 所述动态扩容模块103，用于根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。
- [0063] 如图5所示，是本申请实现分布式的数据库动态扩容方法的电子设备的结构示意图。
- [0064] 所述电子设备1可以包括处理器10、存储器11和总线，还可以包括存储在所述存储器11中并可在所述处理器10上运行的计算机程序，如分布式的数据库动态扩容程序12。
- [0065] 其中，所述存储器11至少包括一种类型的可读存储介质，所述可读存储介质包

括闪存、移动硬盘、多媒体卡、卡型存储器（例如：SD或DX存储器等）、磁性存储器、磁盘、光盘等。所述存储器11在一些实施例中可以是电子设备1的内部存储单元，例如该电子设备1的移动硬盘。所述存储器11在另一些实施例中也可以是电子设备1的外部存储设备，例如电子设备1上配备的插接式移动硬盘、智能存储卡（Smart Media Card， SMC）、安全数字（Secure Digital， SD）卡、闪存卡（Flash Card）等。进一步地，所述存储器11还可以既包括电子设备1的内部存储单元也包括外部存储设备。所述存储器11不仅可以用于存储安装于电子设备1的应用软件及各类数据，例如分布式的数据库动态扩容的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

[0066] 所述处理器10在一些实施例中可以由集成电路组成，例如可以由单个封装的集成电路所组成，也可以是由多个相同功能或不同功能封装的集成电路所组成，包括一个或者多个中央处理器（Central Processing unit， CPU）、微处理器、数字处理芯片、图形处理器及各种控制芯片的组合等。所述处理器10是所述电子设备的控制核心（Control Unit），利用各种接口和线路连接整个电子设备的各个部件，通过运行或执行存储在所述存储器11内的程序或者模块（例如执行分布式的数据库动态扩容等），以及调用存储在所述存储器11内的数据，以执行电子设备1的各种功能和处理数据。

[0067] 所述总线可以是外设部件互连标准（peripheral component interconnect， 简称PCI）总线或扩展工业标准结构（extended industry standard architecture， 简称EISA）总线等。该总线可以分为地址总线、数据总线、控制总线等。所述总线被设置为实现所述存储器11以及至少一个处理器10等之间的连接通信。

[0068] 图5仅示出了具有部件的电子设备的，本领域技术人员可以理解的是，图5示出的结构并不构成对所述电子设备1的限定，可以包括比图示更少或者更多的部件，或者组合某些部件，或者不同的部件布置。

[0069] 例如，尽管未示出，所述电子设备1还可以包括给各个部件供电的电源（比如电池），优选地，电源可以通过电源管理装置与所述至少一个处理器10逻辑相连，从而通过电源管理装置实现充电管理、放电管理、以及功耗管理等功能。

电源还可以包括一个或一个以上的直流或交流电源、再充电装置、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。所述电子设备1还可以包括多种传感器、蓝牙模块、Wi-Fi模块等，在此不再赘述。

[0070] 进一步地，所述电子设备1还可以包括网络接口，可选地，所述网络接口可以包括有线接口和/或无线接口（如WI-FI接口、蓝牙接口等），通常用于在该电子设备1与其他电子设备之间建立通信连接。

[0071] 可选地，该电子设备1还可以包括用户接口，用户接口可以是显示器（Display）、输入单元（比如键盘（Keyboard）），可选地，用户接口还可以是标准的有线接口、无线接口。可选地，在一些实施例中，显示器可以是LED显示器、液晶显示器、触控式液晶显示器以及OLED（Organic Light-Emitting Diode，有机发光二极管）触摸器等。其中，显示器也可以适当的称为显示屏或显示单元，用于显示在电子设备1中处理的信息以及用于显示可视化的用户界面。

[0072] 应该了解，所述实施例仅为说明之用，在专利申请范围上并不受此结构的限制。

[0073] 所述电子设备1中的所述存储器11存储的分布式的数据库动态扩容12是多个指令的组合，在所述处理器10中运行时，可以实现：

[0074] 对原始数据集进行数据清洗得到标准数据集；

[0075] 计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库；

[0076] 根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。

[0077] 具体地，所述处理器10对上述指令的具体实现方法如下所述：

[0078] 步骤一、获取原始数据集，对所述原始数据集进行数据清洗得到标准数据集。

[0079] 本申请的目的是根据原始数据集的数据量对预构建的数据库进行动态扩容，当完成动态扩容后的数据库不管是存储其他数据集时，还是存储完成后查询其他数据集时，速度都会更快更高效。

[0080] 所述原始数据集的数据量一般非常庞大，目的是为了最大化动态扩容数据库，防止后续动态扩容后的数据库存储其他数据集时，由于扩容的存储区数量不够

导致需要再次动态扩容的现象发生，如利用爬虫技术爬取某大型购物网站的用户交易额数据，所述用户交易额数据即为原始数据集，且大型购物网站的用户交易额数据的量级一般都为亿级，满足数据量庞大的要求。

[0081] 进一步地，如上述利用爬虫技术爬取得到用户交易额数据，但由于爬取过程会爬取到不符合要求的数据，因此对原始数据集进行数据清洗是必不可少的步骤。

[0082] 详细地，对所述原始数据集进行数据清洗得到标准数据集，包括：对所述原始数据集进行缺失值填充、异常值去除、属性区间划分得到所述标准数据集。

[0083] 进一步地，所述缺失值填充的方法众多，如填充固定值、填充均值、填充众数等。

[0084] 所述属性区间划分是依据原始数据集内数据的属性不同，根据数据的属性对原始数据集内数据进行划分。所述属性区间划分包括：首先对原始数据集内数据进行排序，同时利用等宽的区间对排序后的原始数据集进行划分得到多个区间的原始数据集，分别计算每个区间的原始数据集的均值，并用均值替代每个区间内的原始数据集，完成属性区间划分。如上述用户交易额数据共有1亿个数据量，对每笔交易额进行排序得到从小到大的用户交易额数据，按照每100元一个区间对从小到大排序的用户交易额数据进行划分，分别计算每个区间内的交易额的平均值，并替代每个区间的每笔交易额，如0-100元区间的交易额为[1.2, 3.7, 4.8, 10, 99]，则计算出平均值为23.74，则将23.74替代[1.2, 3.7, 4.8, 10, 99]得到[23.74, 23.74, 23.74, 23.74, 23.74]。

[0085] 步骤二、计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库。

[0086] 所述分布式架构平台可采用当前已公开的hadoop、Pregel、Spark、Storm等，本申请中分布式架构平台是对数据库进行动态扩容的基础，所有的动态扩容都是在分布式架构平台内完成，因为多数分布式架构平台具有高可靠性的特点，在动态扩容中不容易产生崩溃的现象，其次分布式架构平台内包括多个节点，使得在动态扩容时保证多个节点之间具有联系且可以自由移动交换数据，从而保证各个节点的动态平衡。

- [0087] 详细地，所述步骤二包括：
- [0088] 统计所述分布式架构平台内各存储节点的存储量；
- [0089] 根据所述标准数据集的数据大小和所述存储量计算得到存储节点数量；
- [0090] 根据所述存储节点数量从所述分布式架构平台内选择节点构建得到所述数据库。
- [0091] 例如上述利用爬虫技术爬取并数据清理得到的标准用户交易额数据，通过计算得到标准用户交易额数据的数据大小为200GB，而分布式架构平台内各存储节点的存储量为1GB，则计算得到需要200个存储节点，则从所述分布式架构平台内选择200个存储节点构建为数据库。
- [0092] 步骤三、根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。
- [0093] 详细地，所述步骤三包括：
- [0094] 计算所述数据数量以指数形式表达的底数和指数；
- [0095] 根据所述底数划分所述存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表；
- [0096] 根据预先设定的命名规则，对所述分存储区和所述动态扩容表命名得到动态扩容数据库。
- [0097] 如上述200个存储节点构建成的数据库，假设上述标准用户交易额数据的数据数量为56800235584，由于56800235584以指数形式表达就是62的6次方，因此可将上述构建的数据库划分为62个分存储区，然后在每一个分存储区内建立相同指数的62张动态扩容表。
- [0098] 进一步地，所述预先设定的命名规则可多种多样，本申请较佳实施例中，所述分存储区和所述动态扩容表的命名规则均使用52个大小写字母和0-9的十位数字组成的字符串，如分存储区命名为db_0、db_1、…、db_zz，动态扩容表命名为table_0、table_1、…、table_zz，所有的分存储区和所有的动态扩容表作为一个整体称为动态扩容数据库。
- [0099] 当完成所述动态扩容数据库的建立后，若需要在所述动态扩容数据库内存储数据集，则存储步骤包括：

- [0100] 获取待存储数据集，根据所述分存储区的数量将所述待存储数据集分成相同数量的分待存储数据集，将所述分待存储数据集依次存储至所述动态扩容表集内。
- [0101] 当用户需要访问待存储数据集时，通过命名规则依次访问待存储数据集。同时待存储数据集可为所述原始数据集，也可以为新的数据集，如利用上述用户交易额数据进行动态扩容，当动态扩容完成后将用户交易额数据存储至动态扩容后的数据库内，也可以重新存储新的数据集，如用户在某贷款平台的贷款金额数据集。
- [0102] 进一步地，所述电子设备1集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个非易失性或易失性计算机可读存储介质中。所述计算机可读存储介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器（ROM，Read-Only Memory）。
- [0103] 本申请计算机可读存储介质具体实施方式与上述分布式的数据库动态扩容方法各实施例基本相同，在此不作累述。
- [0104] 在本申请所提供的几个实施例中，应该理解到，所揭露的设备，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。
- [0105] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。
- [0106] 另外，在本申请各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。
- [0107] 对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且

在不背离本申请的精神或基本特征的情况下，能够以其他的具体形式实现本申请。

[0108] 因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本申请的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本申请内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

[0109] 此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第二等词语用来表示名称，而并不表示任何特定的顺序。

[0110] 最后应说明的是，以上实施例仅用以说明本申请的技术方案而非限制，尽管参照较佳实施例对本申请进行了详细说明，本领域的普通技术人员应当理解，可以对本申请的技术方案进行修改或等同替换，而不脱离本申请技术方案的精神和范围。

权利要求书

- [权利要求 1] 一种分布式的数据库动态扩容方法，其中，所述方法包括：
- 对原始数据集进行数据清洗得到标准数据集；
- 计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库；
- 根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。
- [权利要求 2] 如权利要求1所述的分布式的数据库动态扩容方法，其中，所述计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库，包括：
- 统计所述分布式架构平台内各存储节点的存储量；
- 根据所述标准数据集的数据大小和所述存储量计算得到存储节点数量；
- 根据所述存储节点数量从所述分布式架构平台内选择节点构建得到所述数据库。
- [权利要求 3] 如权利要求1所述的分布式的数据库动态扩容方法，其中，所述根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库，包括：
- 计算所述数据数量以指数形式表达的底数和指数；
- 根据所述底数划分所述存储区得到多个所述分存储区；
- 在多个所述分存储区内构建动态扩容表集，根据预先设定的命名规则，命名所述分存储区和所述动态扩容表集得到动态扩容数据库。
- [权利要求 4] 如权利要求1所述的分布式的数据库动态扩容方法，其中，所述方法还包括：
- 获取待存储数据集，根据所述分存储区的数量，将所述待存储数据集拆分成与所述分存储区相同数量的分待存储数据集；
- 将所述分待存储数据集依次存储至所述动态扩容表集内；
- 接收访问所述待存储数据集的访问指令，通过命名规则依次访问所述

待存储数据集。

- [权利要求 5] 如权利要求1所述的分布式的数据库动态扩容方法，其中，所述对原始数据集进行数据清洗得到标准数据集，包括：
对所述原始数据集进行缺失值填充、异常值去除、属性区间划分得到所述标准数据集。
- [权利要求 6] 如权利要求5所述的分布式的数据库动态扩容方法，其中，所述属性区间划分包括：
对完成所述缺失值填充及所述异常值去除操作的原始数据集内的数据进行排序；
利用等宽的区间对排序后的原始数据集进行划分得到多个区间的原始数据集；
分别计算每个区间的原始数据集的数据均值，并用所述数据均值替代每个区间内的原始数据集内的数据。
- [权利要求 7] 一种分布式的数据库动态扩容装置，其中，所述装置包括：
数据清洗模块，用于对原始数据集进行数据清洗得到标准数据集；
数据库构建模块，用于计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库；
动态扩容模块，用于根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。
- [权利要求 8] 如权利要求7所述的分布式的数据库动态扩容装置，其中，所述根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库，包括：
计算所述数据数量以指数形式表达的底数和指数；
根据所述底数划分所述存储区得到多个所述分存储区；
在多个所述分存储区内构建动态扩容表集，根据预先设定的命名规则，命名所述分存储区和所述动态扩容表集得到动态扩容数据库。
- [权利要求 9] 一种电子设备，其中，所述电子设备包括：

至少一个处理器；以及，
与所述至少一个处理器通信连接的存储器；其中，
所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被
所述至少一个处理器执行时实现如下步骤：
对原始数据集进行数据清洗得到标准数据集；
计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分
布式架构平台内构建数据库；
根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多
个所述分存储区内构建动态扩容表集得到动态扩容数据库。

[权利要求 10] 如权利要求9所述的电子设备，其中，所述计算所述标准数据集的数
据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据
库，包括：
统计所述分布式架构平台内各存储节点的存储量；
根据所述标准数据集的数据大小和所述存储量计算得到存储节点数量
；
根据所述存储节点数量从所述分布式架构平台内选择节点构建得到所
述数据库。

[权利要求 11] 如权利要求9所述的电子设备，其中，所述根据所述数据数量拆分所
述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动
态扩容表集得到动态扩容数据库，包括：
计算所述数据数量以指数形式表达的底数和指数；
根据所述底数划分所述存储区得到多个所述分存储区；
在多个所述分存储区内构建动态扩容表集，根据预先设定的命名规则
，命名所述分存储区和所述动态扩容表集得到动态扩容数据库。

[权利要求 12] 如权利要求9所述的电子设备，其中，所述指令被所述至少一个处理
器执行时还实现如下步骤：
获取待存储数据集，根据所述分存储区的数量，将所述待存储数据集
拆分成与所述分存储区相同数量的分待存储数据集；

将所述分待存储数据集依次存储至所述动态扩容表集内；

接收访问所述待存储数据集的访问指令，通过命名规则依次访问所述待存储数据集。

[权利要求 13] 如权利要求9所述的电子设备，其中，所述对原始数据集进行数据清洗得到标准数据集，包括：
对所述原始数据集进行缺失值填充、异常值去除、属性区间划分得到所述标准数据集。

[权利要求 14] 如权利要求13所述的电子设备，其中，所述属性区间划分包括：
对完成所述缺失值填充及所述异常值去除操作的原始数据集内的数据进行排序；
利用等宽的区间对排序后的原始数据集进行划分得到多个区间的原始数据集；
分别计算每个区间的原始数据集的数据均值，并用所述数据均值替代每个区间内的原始数据集内的数据。

[权利要求 15] 一种计算机可读存储介质，存储有计算机程序，其中，所述计算机程序被处理器执行时实现如下步骤：
对原始数据集进行数据清洗得到标准数据集；
计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库；
根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库。

[权利要求 16] 如权利要求15所述的计算机可读存储介质，其中，所述计算所述标准数据集的数据大小和数据数量，根据所述数据大小在分布式架构平台内构建数据库，包括：
统计所述分布式架构平台内各存储节点的存储量；
根据所述标准数据集的数据大小和所述存储量计算得到存储节点数量；
根据所述存储节点数量从所述分布式架构平台内选择节点构建得到所

述数据库。

- [权利要求 17] 如权利要求15所述的计算机可读存储介质，其中，所述根据所述数据数量拆分所述数据库的存储区得到多个分存储区，在多个所述分存储区内构建动态扩容表集得到动态扩容数据库，包括：
- 计算所述数据数量以指数形式表达的底数和指数；
- 根据所述底数划分所述存储区得到多个所述分存储区；
- 在多个所述分存储区内构建动态扩容表集，根据预先设定的命名规则，命名所述分存储区和所述动态扩容表集得到动态扩容数据库。
- [权利要求 18] 如权利要求15所述的计算机可读存储介质，其中，所述计算机程序被处理器执行时还实现如下步骤：
- 获取待存储数据集，根据所述分存储区的数量，将所述待存储数据集拆分成与所述分存储区相同数量的分待存储数据集；
- 将所述分待存储数据集依次存储至所述动态扩容表集内；
- 接收访问所述待存储数据集的访问指令，通过命名规则依次访问所述待存储数据集。
- [权利要求 19] 如权利要求15所述的计算机可读存储介质，其中，所述对原始数据集进行数据清洗得到标准数据集，包括：
- 对所述原始数据集进行缺失值填充、异常值去除、属性区间划分得到所述标准数据集。
- [权利要求 20] 如权利要求19所述的计算机可读存储介质，其中，所述属性区间划分包括：
- 对完成所述缺失值填充及所述异常值去除操作的原始数据集内的数据进行排序；
- 利用等宽的区间对排序后的原始数据集进行划分得到多个区间的原始数据集；
- 分别计算每个区间的原始数据集的数据均值，并用所述数据均值替代每个区间内的原始数据集内的数据。

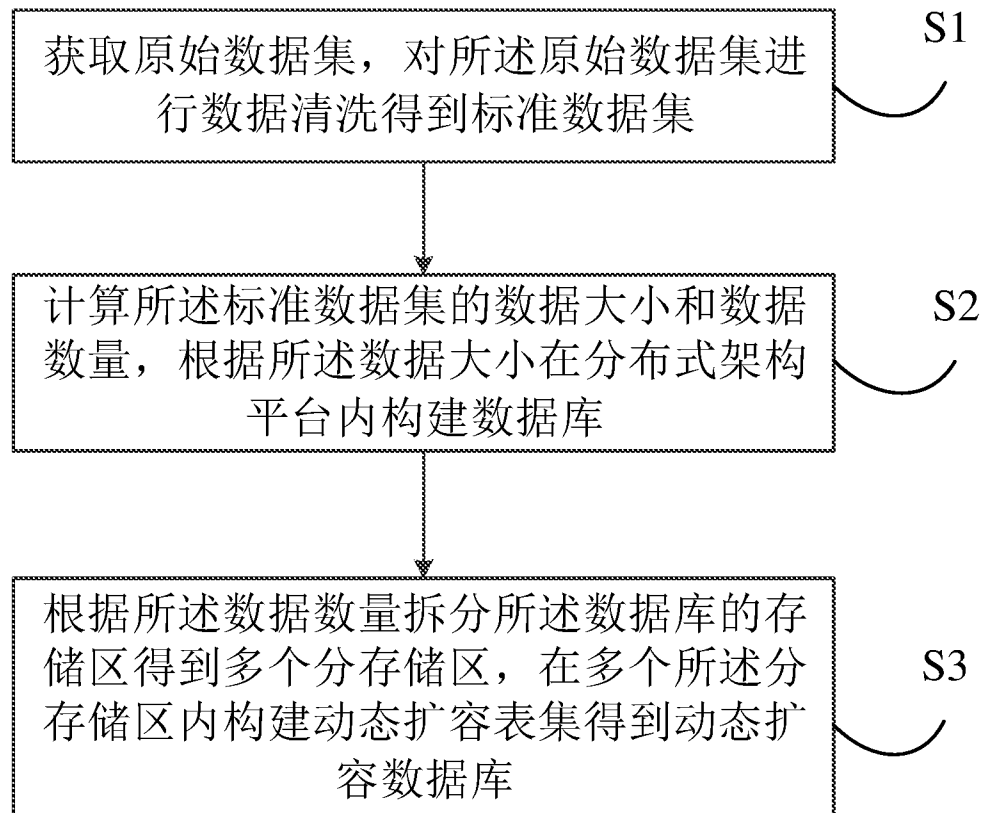


图 1

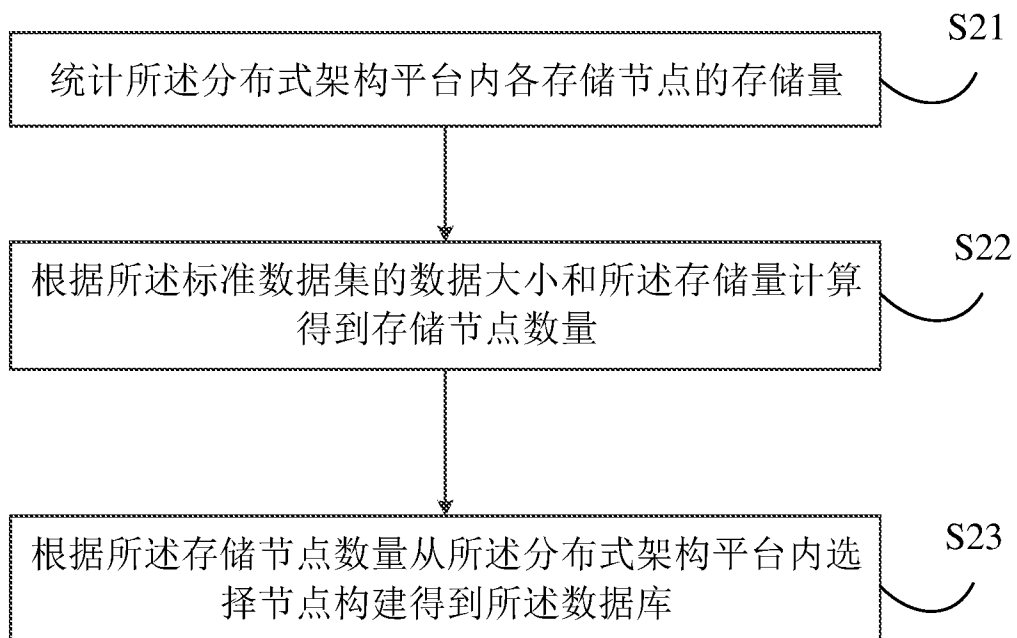


图 2

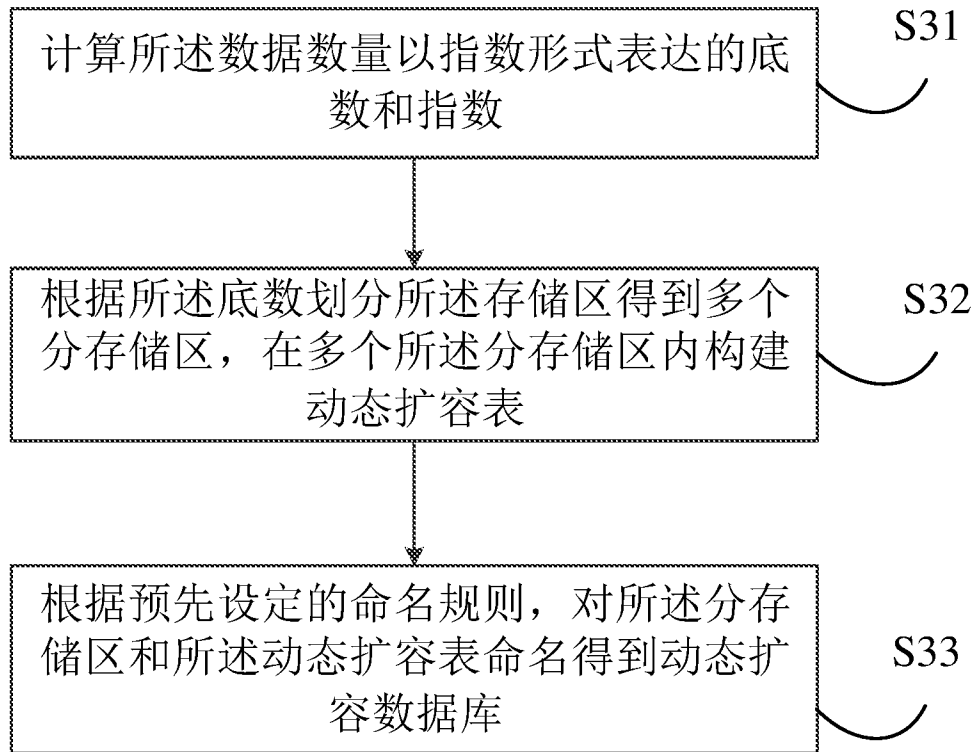


图 3

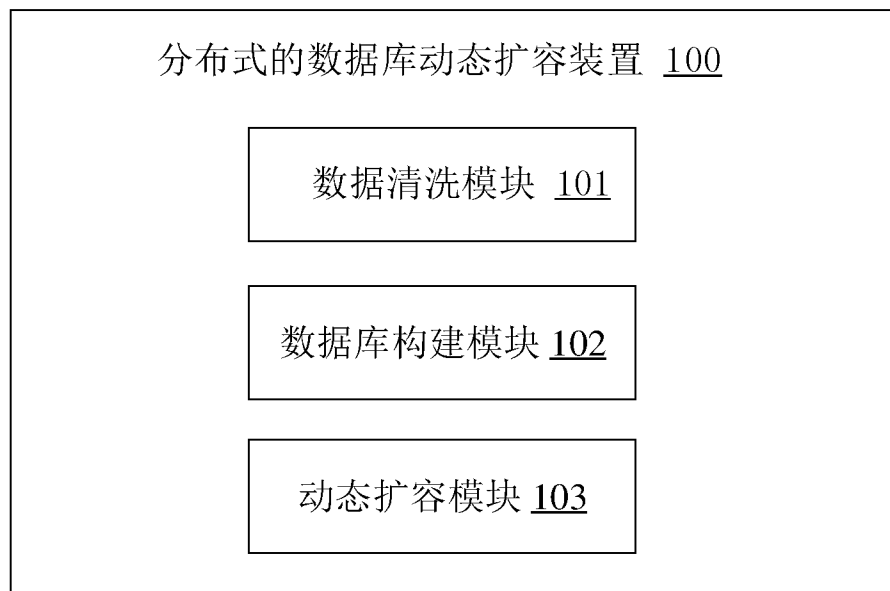


图 4

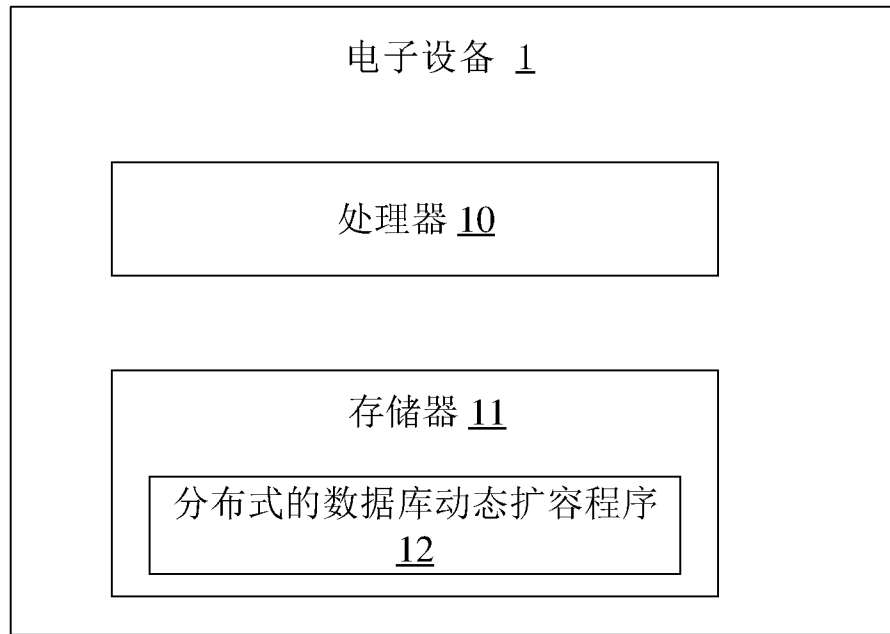


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/099088

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/215(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE: 分布, 数据库, 动态, 扩容, 清洗, 标准, 数据集, 大小, 容量, 数量, 存储, 拆分, 平均, distribution, database, dynamic, expand, clean, standard, set, size, number, capacity, store, split, average

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106959948 A (CHRUST INFORMATION TECHNOLOGY CO., LTD.) 18 July 2017 (2017-07-18) description, paragraphs [0062]-[0106]	1-5, 7-13, 15-19
A	CN 105323111 A (NARI GROUP CORPORATION) 10 February 2016 (2016-02-10) entire document	1-20
A	CN 110674101 A (BEIJING KINGSOFT CLOUD NETWORK TECHNIQUE CO., LTD. et al.) 10 January 2020 (2020-01-10) entire document	1-20
A	US 2018225346 A1 (ZTE CORPORATION) 09 August 2018 (2018-08-09) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

14 December 2020

Date of mailing of the international search report

21 December 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/099088

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106959948	A	18 July 2017	None			
CN	105323111	A	10 February 2016	None			
CN	110674101	A	10 January 2020	None			
US	2018225346	A1	09 August 2018	CN	106708815	A	24 May 2017
				EP	3324304	A1	23 May 2018
				WO	2017008604	A1	19 January 2017
				JP	2018524733	A	30 August 2018
				KR	20180017198	A	20 February 2018

国际检索报告

国际申请号

PCT/CN2020/099088

A. 主题的分类

G06F 16/215 (2019.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, WPI, EPDOC, CNKI, IEEE: 分布, 数据库, 动态, 扩容, 清洗, 标准, 数据集, 大小, 容量, 数量, 存储, 拆分, 平均, distribution, database, dynamic, expand, clean, standard, set, size, number, capacity, store, split, average

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 106959948 A (普华诚信信息技术有限公司) 2017年 7月 18日 (2017 - 07 - 18) 说明书第[0062]-[0106]段	1-5, 7-13, 15-19
A	CN 105323111 A (南京南瑞集团公司) 2016年 2月 10日 (2016 - 02 - 10) 全文	1-20
A	CN 110674101 A (北京金山云网络技术有限公司等) 2020年 1月 10日 (2020 - 01 - 10) 全文	1-20
A	US 2018225346 A1 (ZTE CORPORATION) 2018年 8月 9日 (2018 - 08 - 09) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 12月 14日

国际检索报告邮寄日期

2020年 12月 21日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

胡百乐

电话号码 86-(10)-53961519

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/099088

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	106959948	A	2017年 7月 18日	无	
CN	105323111	A	2016年 2月 10日	无	
CN	110674101	A	2020年 1月 10日	无	
US	2018225346	A1	2018年 8月 9日	CN 106708815 A	2017年 5月 24日
				EP 3324304 A1	2018年 5月 23日
				WO 2017008604 A1	2017年 1月 19日
				JP 2018524733 A	2018年 8月 30日
				KR 20180017198 A	2018年 2月 20日