

(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



(10) 国际公布号
WO 2025/153068 A1

(43) 国际公布日
2025 年 7 月 24 日 (24.07.2025)

(51) 国际专利分类号:
H04L 67/10 (2022.01)

(21) 国际申请号: PCT/CN2025/073036

(22) 国际申请日: 2025 年 1 月 17 日 (17.01.2025)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202410077190.0 2024 年 1 月 18 日 (18.01.2024) CN
202410543817.7 2024 年 4 月 30 日 (30.04.2024) CN

(71) 申请人: 华为云计算技术有限公司 (HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.)
[CN/CN]; 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, 550025 (CN)。

(72) 发明人: 钟延辉 (ZHONG, Yanhui); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心 550025 (CN)。白平昌 (BAI, Pingchang); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心 550025 (CN)。伍华涛 (WU, Huatao); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心 550025 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区东晓街道东晓社区太白路 3008 号悦彩城 (北地块) 写字楼 1304 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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI,

(54) Title: VECTOR PROCESSING METHOD BASED ON CLOUD MANAGEMENT PLATFORM, AND CLOUD MANAGEMENT PLATFORM

(54) 发明名称: 一种基于云管理平台的向量处理方法以及云管理平台

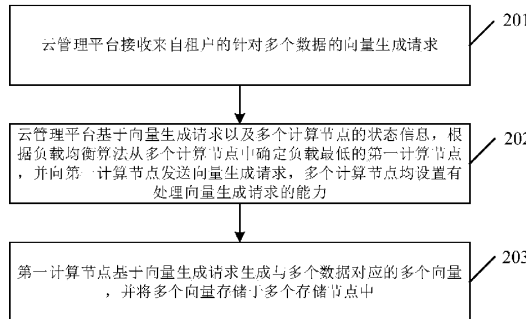


图 2

- 201 A cloud management platform receives a vector generation request from a tenant for a plurality of pieces of data
202 On the basis of the vector generation request and state information of a plurality of computing nodes, the cloud management platform determines a first computing node with the lowest load from among the plurality of computing nodes on the basis of a load balancing algorithm, and sends the vector generation request to the first computing node, wherein the plurality of computing nodes are all provided with capabilities for processing the vector generation request
203 On the basis of the vector generation request, the first computing node generates a plurality of vectors corresponding to the plurality of pieces of data, and stores the plurality of vectors in a plurality of storage nodes

(57) Abstract: Disclosed in the present application are a vector processing method based on a cloud management platform, and a cloud management platform, which can realize global sharing of resources of a plurality of computing nodes, thereby more effectively utilizing the resources of the computing nodes, and thus reducing the cost to a certain extent. The method in the present application comprises: after a cloud management platform receives a vector processing request triggered by a tenant, the cloud management platform acquiring, on the basis of the vector processing request, state information of a plurality of computing nodes serving the tenant; after obtaining the state information of the plurality of computing nodes, the cloud management platform analyzing the states of the plurality of computing nodes by using the states of the plurality of computing nodes, so as to select, from among the plurality of computing nodes, a computing

[见续页]



WO 2025/153068 A1

GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, 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, UY, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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))。

node with an optimal state as a first computing node, and sending the vector processing request to the first computing node; and after obtaining the vector processing request, the first computing node executing the vector processing request, so as to complete vector processing.

(57) 摘要: 本申请公开了一种基于云管理平台的向量处理方法以及云管理平台, 其可令多个计算节点的资源在全局上实现共享, 更加有效地利用了计算节点的资源, 从而在一定程度上降低成本。本申请的方法包括: 当云管理平台接收到租户所触发的向量处理请求后, 云管理平台可基于向量处理请求获取服务于租户的多个计算节点的状态信息。得到多个计算节点的状态信息后, 云管理平台可利用多个计算节点的状态对多个计算节点的状态进行分析, 以从多个计算节点中选择状态最优的计算节点作为第一计算节点, 并向第一计算节点发送向量处理请求。得到向量处理请求后, 第一计算节点可执行向量处理请求, 从而完成向量处理。