

(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
2020 年 11 月 26 日 (26.11.2020)



(10) 国际公布号
WO 2020/233252 A1

(51) 国际专利分类号:
G06F 11/36 (2006.01)

(21) 国际申请号: PCT/CN2020/083381

(22) 国际申请日: 2020 年 4 月 3 日 (03.04.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910432603.1 2019年5月23日 (23.05.2019) CN

(71) 申请人: 深圳前海微众银行股份有限公司
(WEBANK CO., LTD.) [CN/CN]; 中国广东省深圳市前海深港合作区前湾一路1号A栋201室, Guangdong 518027 (CN)。

(72) 发明人: 王和平(WANG, Heping); 中国广东省深圳市前海深港合作区前湾一路1号A栋201室, Guangdong 518027 (CN)。 尹强(YIN, Qiang); 中国广东省深圳市前海深港合作区前湾一路1号A栋201室, Guangdong 518027 (CN)。 刘有(LIU, You); 中国广东省深圳市前海深港合作区前湾一路1号A栋201室, Guangdong 518027 (CN)。 黄山(HUANG, Shan); 中国广东省深圳市前海深港合作区前湾一路1号A栋201室, Guangdong 518027 (CN)。 杨峙岳(YANG, Zhiyue); 中国广东省深圳市前海深港合作区前湾一路1号A栋201室, Guangdong 518027 (CN)。 邸帅(DI, Shuai); 中国广东省深圳市前海深港合作区前湾一路1号A栋201室, Guangdong 518027 (CN)。 卢道和(LU, Daohe); 中国广东省深圳市前海深港合作区前湾一路1号A栋201室, Guangdong 518027 (CN)。

(54) Title: METHOD AND APPARATUS FOR DIAGNOSING SPARK APPLICATION

(54) 发明名称: 一种诊断Spark应用的方法及装置

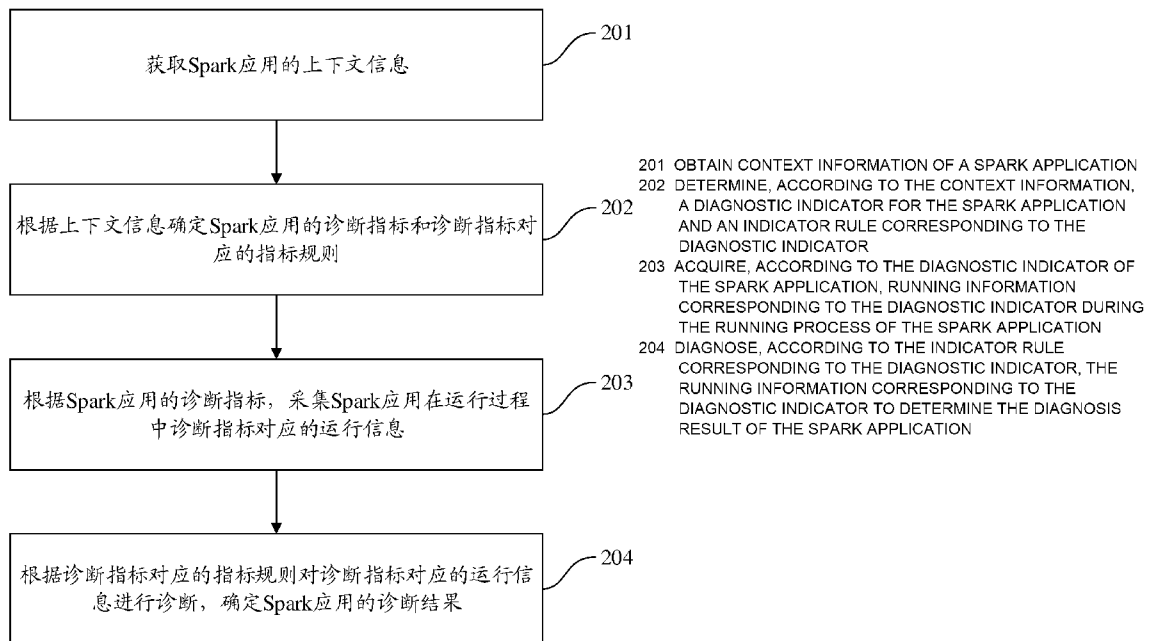


图 2

(57) Abstract: The present invention relates to the technical field of finances. Disclosed are a method and an apparatus for diagnosing a Spark application. The method comprises: obtaining context information of a Spark application; determining, according to the context information, a diagnostic indicator for the Spark application and an indicator rule corresponding to the diagnostic indicator; acquiring, according to the diagnostic indicator of the Spark application, running information corresponding to the diagnostic indicator during the running process of the Spark application; and diagnosing, according to the indicator rule corresponding to the diagnostic indicator,

(74) 代理人: 北京同达信恒知识产权代理有限公司
(TDIP & PARTNERS); 中国北京市西城区裕民路
18号北环中心A座2002, Beijing 100029 (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,
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))。

the running information corresponding to the diagnostic indicator to determine the diagnosis result of the Spark application. In the present technical solution, the running indicator of the Spark application is acquired in real time during the running process of the Spark application, real-time diagnosis is conducted for problems during the running of the Spark application, and an effective diagnostic measure is provided.

(57) 摘要: 本发明涉及金融科技领域, 并公开了一种诊断Spark应用的方法及装置, 其中, 方法包括: 获取Spark应用的上下文信息, 根据上下文信息确定Spark应用的诊断指标和诊断指标对应的指标规则, 根据Spark应用的诊断指标, 采集Spark应用在运行过程中诊断指标对应的运行信息, 并根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断, 确定Spark应用的诊断结果。本技术方案在Spark应用运行过程中, 实时采集Spark应用的运行指标, 对Spark应用的运行中的问题进行实时诊断, 并提供有效的诊断措施。

一种诊断Spark应用的方法及装置

相关申请的交叉引用

本申请要求在2019年05月23日提交中国专利局、申请号为201910432603.1、
申请名称为“一种诊断Spark应用的方法及装置”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

本申请实施例涉及金融科技（Fintech）领域，尤其涉及一种诊断 Spark
应用的方法及装置。

背景技术

随着计算机技术的发展，越来越多的技术应用在金融领域，传统金融业
正在逐步向金融科技转变，Spark 技术也不例外，但由于金融、支付行业的安全
性、实时性要求，也对 Spark 技术提出的更高的要求。

Spark 技术是专为大规模数据处理而设计的快速通用的计算引擎，Spark
使用了内存运算技术，能在数据尚未写入硬盘时在内存分析运算。现有的
Spark 应用诊断是在 Spark 应用运行完成后，收集运行过程中的日志进行分析，
基于预设规则方法确定 Spark 应用在运行中存在的问题，从而进行相应的调整。

现有技术中，在 Spark 应用运行完成后对日志进行诊断分析，不能及时发
现运行过程中的问题，并提供有效诊断措施。

发明内容

本申请实施例提供一种诊断 Spark 应用的方法及装置，用以实现在 Spark
应用运行过程中，实时采集 Spark 应用的运行指标，对 Spark 应用的运行中的
问题进行实时诊断，并提供有效的诊断措施。

第一方面，本申请实施例提供一种诊断 Spark 应用的方法，该方法可以通

过在金融科技领域中提供诊断 Spark 应用的方法所适用的运行诊断器(runtime diagnoser) 执行, 包括: 获取 Spark 应用的上下文信息; 根据上下文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则; 根据 Spark 应用的诊断指标, 采集 Spark 应用在运行过程中诊断指标对应的运行信息; 根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断, 确定 Spark 应用的诊断结果。

上述技术方案中, 通过获取 Spark 应用的上下文信息, 生成诊断指标和指标规则, 实现在 Spark 应用过程中实时获取诊断指标对应的运行信息, 并对运行信息进行诊断, 确定出 Spark 应用的诊断结果, 从而可以对 Spark 应用运行中出现的运行故障进行实时诊断, 并获取到诊断结果, 进一步的, 在 Spark 应用运行中获取运行信息, 可以较全面的采集到 Spark 应用运行中的参数指标等, 相比于在 Spark 应用运行完成后的运行日志, 运行参数更为全面且反应 Spark 应用的当前运行状态。

可选的, 诊断指标为多个; 针对任一个诊断指标, 根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断, 确定诊断指标对应的诊断结果; 从多个诊断指标对应的诊断结果中确定出符合预设规则的诊断结果, 并确定为 Spark 应用的诊断结果。

上述技术方案中, 针对任一个诊断指标, 都设置有与之相对应的诊断结果, 同时预设规则, 用于从多个诊断结果中确定出符合预设规则的诊断结果, 进一步将其确定为 Spark 应用的诊断结果。从多方面诊断 Spark 应用在运行中的运行指标, 多维度评估 Spark 应用, 可以及时发现运行中的故障, 且根据预设规则, 将具有代表性的诊断指标对应的诊断结果作为当前 Spark 应用的诊断结果, 从而方便的用户直观了解到当前 Spark 应用运行问题, 提高用户体验。

可选的, 在确定 Spark 应用的诊断结果之后, 还包括: 根据符合预设规则的诊断结果中的诊断码, 从预设数据库中获取符合预设规则的诊断结果中的诊断码对应的诊断措施并报告给用户; 预设数据库中预先设置有诊断码与诊断措施的对应关系。

上述技术方案中，提供预设数据库，该预设数据库中预先设置有诊断码与诊断措施的对应关系，从而可以在确定出 Spark 应用的诊断结果后，向用户提供有针对性的诊断措施，即解决方案，方便用户根据诊断措施自主解决问题，从而及时解决 Spark 应用在运行中的问题。该技术方案，无需用户查询相关资料用于解决 Spark 应用的运行问题，而是直接设置好相关的解决方案并提供给用户，提高用户解决问题的效率，且提高用户体验。

可选的，将诊断指标对应的运行信息进行统一化处理后，生成诊断指标对应的运行指标；根据诊断指标对应的指标规则对诊断指标对应的运行指标进行诊断。

上述技术方案中，将诊断指标对应的运行信息进行统一化处理，封装成可进行诊断的运行指标，从而对运行指标进行诊断。

可选的，在根据上下文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则之前，还包括：获取用户配置信息；根据用户配置信息、上下文信息，确定 Spark 应用的诊断指标和诊断指标对应的指标规则。

上述技术方案中，支持用户选择诊断指标和诊断指标对应的指标规则，即用户可以自行选择指标采集器和诊断规则器来对 Spark 作业进行实时诊断，满足不同用户的需求。

第二方面，本申请实施例提供了一种诊断 Spark 应用的装置，该装置可以为上述第一方面中的运行诊断器 (runtime diagnoser)，或者包含上述运行诊断器的装置，或者为具有运行诊断器的相应功能的芯片等。该装置包括实现上述方法相应的模块、单元、或手段 (means)，该模块、单元、或 means 可以通过硬件实现，软件实现，或者通过硬件执行相应的软件实现。该硬件或软件包括一个或多个与上述功能相对应的模块或单元。其中，该装置包括：获取单元，用于获取 Spark 应用的上下文信息；处理单元，用于根据上下文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则；根据 Spark 应用的诊断指标，采集 Spark 应用在运行过程中诊断指标对应的运行信息；根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断，确定 Spark 应用的诊

断结果。

可选的，诊断指标为多个；处理单元具体用于：针对任一个诊断指标，根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断，确定诊断指标对应的诊断结果；从多个诊断指标对应的诊断结果中确定出符合预设规则的诊断结果，并确定为 Spark 应用的诊断结果。

可选的，处理单元还用于：在确定 Spark 应用的诊断结果之后，根据符合预设规则的诊断结果中的诊断码，通过获取单元从预设数据库中获取符合预设规则的诊断结果中的诊断码对应的诊断措施并报告给用户；其中，预设数据库中预先设置有诊断码与诊断措施的对应关系。

可选的，处理单元具体用于：将诊断指标对应的运行信息进行统一化处理后，生成诊断指标对应的运行指标；根据诊断指标对应的指标规则对诊断指标对应的运行指标进行诊断。

可选的，处理单元还用于：在根据上下文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则之前，通过获取单元获取用户配置信息；根据用户配置信息、上下文信息，确定 Spark 应用的诊断指标和诊断指标对应的指标规则。

第三方面，本申请还提供了一种计算设备，包括：处理器和存储器；处理器，用于与存储器耦合，通过调用并执行存储器中存储的存储器存储有计算机程序或指令，当该处理器执行该计算机程序或指令时，以使该通信装置可以执行上述第一方面的方法。该计算设备可以为上述第一方面中的运行诊断器，或者包含上述运行诊断器的装置，或者为具有运行诊断器的相应功能的芯片等。

第四方面，本申请还提供了一种计算机可读非易失性存储介质，包括计算机可读指令，当计算机读取并执行计算机可读指令时，使得计算机执行上述诊断 Spark 应用的方法。

第五方面，本申请提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机可以执行上述第一方面的方法。

其中，上述第二方面至第五方面中任一种可能的实现方式所带来的技术效果可参见上述第一方面中不同的实现方式所带来的技术效果，此处不再赘述。

5

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简要介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

10

图 1 为本申请实施例提供的一种系统架构的示意图；

图 2 为本申请实施例提供的一种诊断 Spark 应用的方法的流程示意图；

图 3 为本申请实施例提供的一种诊断 Spark 应用的装置的结构示意图；

图 4 为本申请实施例提供的一种计算设备的结构示意图。

15

具体实施方式

为了使本发明的目的、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。

20

Spark 提供了方便的监听器（Listener）和指标系统（Metrics）等 API（Application Programming Interface，应用程序编程接口），用于收集 Spark 应用在运行过程中的运行信息。在此基础上，图 1 示例性的示出了本申请实施例在金融科技领域中提供诊断 Spark 应用的方法所适用运行诊断器（Runtime Diagnoser）100，该运行诊断器 100 可以包括指标收集器（Metric Collector）101、指标规则器（Metric Ruler）102、规则结果合并器（Rule Result Merger）

25

103、诊断通知器 (Diagnostic Notifier) 104、数据库 105; 运行诊断器 100 与监测器 (Monitor) 200 相连接。

运行诊断器 100 用于整个 Spark 应用在诊断过程中的调度, 具体的, 运行诊断器 100 获取 Spark 应用的上下文信息, 根据上下文信息实例化出诊断上下文信息 (Diagnostic Context); 根据诊断上下文信息, 注册指标收集器 101 和指标规则器 102, 并通过定时触发或主动触发的方式, 触发诊断 Spark 应用的任务, 即触发指标收集器 101 根据指标规则收集 Spark 应用在运行过程中的指标信息, 并将收集到的指标信息发送至指标规则器 102, 指标规则器 102 对指标按照相应的指标规则生成规则结果, 随后, 指标规则器 102 将规则结果发送至规则结果合并器 103, 规则结果合并器 103 根据接收到的多个规则结果, 生成该 Spark 应用的诊断结果, 将诊断结果发送至诊断通知器 104, 诊断通知器 104 根据诊断结果从数据库 105 中获取相应的诊断措施, 将诊断结果和诊断措施发送至监测器 200, 以使监测器 200 将 Spark 应用在运行中的诊断结果和诊断措施展示给用户。

基于上述描述, 图 2 示例性的示出了本申请实施例提供的一种诊断 Spark 应用的方法的流程, 该流程可以由诊断 Spark 应用的装置执行, 该装置可以位于上述运行诊断器中, 可以是上述运行诊断器。如图 2 所示, 该流程具体包括:

步骤 201, 获取 Spark 应用的上下文信息。

此处, Spark 应用在运行时, 生成 Spark 应用的上下文信息, 也可以说是 Spark 应用的诊断上下文信息。Spark 应用的上下文信息可以包括 Spark Context, Spark Context 在 Spark 应用程序的执行过程中起着主导作用, 它负责与程序和 Spark 集群进行交互, 包括申请集群资源、创建 RDD (Resilient Distributed Datasets, 弹性分布式数据集)、Accumulators (累加器) 及广播变量等。

步骤 202, 根据上下文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则。

具体的, 根据 Spark 应用的上下文信息获取 Spark 应用的基本信息用于在

运行诊断中进行指标采集。也可以说，诊断上下文信息通过 Spark Context 生成指标收集器、诊断规则器，以及将 Spark Context 里面的接口（Listener 和/或 Metrics）传输入指标收集器，用于指标收集。

本申请实施例中，还可以支持用户选择诊断指标和诊断指标对应的指标规则，即用户可以自行选择指标采集器和诊断规则器来对 Spark 作业进行实时诊断。具体的，在根据上下文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则之前，可以先获取用户配置信息，随后根据用户配置信息、上下文信息，确定 Spark 应用的诊断指标和诊断指标对应的指标规则。

步骤 203，根据 Spark 应用的诊断指标，采集 Spark 应用在运行过程中诊断指标对应的运行信息。

诊断指标可以包括 Task 相关指标、Executor 相关指标、Job 相关指标。示例性的，Spark 应用在运行过程中诊断指标对应的运行信息可以有，Task 相关指标对应的运行信息，如 Task 执行时长、Task 尝试次数、Task 启动时间、Task 输入记录数、Task 输出记录数、Task 状态等；Executor 相关指标对应的运行信息，如 Executor 参数设置个数、现有的 Executor 个数、Executor 退出个数、数据读取量、数据输出量等；Job 相关指标对应的运行信息，如 Job 运行时长、Job 的 Stage 总数和成功的总数、Job 的 Task 总数和成功的总数等。

本申请实施例中，可以根据 Spark 应用的诊断指标，按照一定的采集频率采集 Spark 应用在运行过程中诊断指标对应的运行信息，如在 Spark 应用运行过程中，设定每隔 1min 采集诊断指标对应的运行信息。采集频率可以根据经验进行设定，也可以根据用户需求设定。不同诊断指标的采集频率可以相同或不同。

步骤 204，根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断，确定 Spark 应用的诊断结果。

本申请实施例中，可以将诊断指标对应的运行信息进行统一化处理后，生成诊断指标对应的运行指标，再根据诊断指标对应的指标规则对诊断指标对应的运行指标进行诊断。统一化处理可以包括单位统一化、格式转换等处

理，统一化处理后，将处理后的运行信息封装为诊断指标对应的运行指标，以使得诊断规则器对诊断指标对应的运行指标进行诊断，从而获取到 Spark 应用的诊断结果。

获取到的诊断指标可以为多个，各诊断指标对应运行指标，在对各诊断指标对应运行指标进行诊断后，可以生成各诊断指标对应的诊断结果，从多个诊断指标对应的诊断结果中确定出符合预设规则的诊断结果，并确定为 Spark 应用的诊断结果。

为了更好的解释上述生成 Spark 应用诊断结果的实现方式，以下提供在具体实现场景下的实施例。该实施例可以有三种场景，数据倾斜场景、队列资源不足场景和内存超出场景，并对三种场景下的诊断指标进行诊断，生成记录有诊断评分的诊断结果。

数据倾斜场景：如当某个阶段由于数据倾斜导致 Task 的执行时长异常时，诊断规则器会通过从 Task 指标中获取的所有 Task 的执行时长，取多个执行时长的中位数和最大数，若最大数大于中位数的十倍（参数可配置），则获取最大执行时长的 Task 的输入记录数、中位数执行时长的 Task 的输入记录数，若最大执行时长的 Task 的输入记录数仍是中位数执行时长的 Task 的输入记录数的十倍（参数可配置），则确定此时存在数据倾斜情况，且通过执行时长的倍数和输入记录数的倍数确定诊断评分。例如，当前所有 Task 的执行时长为 1min、2min、4min、5min、45min，则确定出最大执行时长为 45min，中位数执行时长为 4min，确定 45min 大于 4min 的十倍，则进一步确定最大执行时长的 Task 的输入记录数、中位数执行时长的 Task 的输入记录数，假设分别为 300 条和 40 条，则确定 300 条大于 40 条的十倍，即此时存在数据倾斜情况，通过执行时长确定打分为 1 分，通过输入记录数确定打分为 1 分，则数据倾斜情况的诊断评分为 1+1=2 分。

队列资源不足场景：如当队列没有资源导致执行器个数远小于用户设置的值时，诊断规则器会通过从执行器指标中拿到现有 Executor 个数和设置的 Executor 个数，通过判断现有的 Executor 个数是否小于用户设置的 2/3（参数

可配置), 若是, 则确定此时存在队列资源不足情况, 且通过欠缺的 Executor 数确定诊断评分。例如, 现有的 Executor 个数为 10 个, 用户设置的 Executor 个数为 30 个, 则现有的 Executor 个数小于用户设置的 $2/3$, 则此时存在队列资源不足情况, 并通过欠缺的 Executor 数除以 5 进行打分, 当前欠缺的 Executor 数为 20 个, 则队列资源不足情况的诊断评分为 4 分。

内存超出场景: 如当用户执行大数据量的复杂查询时, 多个 Executor 由于内存超出退出的情况时, 诊断规则器会通过从执行器指标中拿到现有 Executor 个数和失败的 Executor 个数, 并通过判断失败的 Executor 个数是否超出现有个数的 $1/4$ (参数可配置), 若是, 则确定此时存在内存超出情况, 且通过失败的 Executor 数确定诊断评分。例如, 现有的 Executor 个数为 10 个, 失败的 Executor 个数为 3 个, 此时存在内存超出情况, 确定根据失败的 Executor 数除以 3 进行打分, 则内存超出情况的诊断评分为 1 分。

根据上述各诊断指标对应的诊断结果中的诊断评分, 可以确定将诊断评分最高的诊断指标对应的诊断结果确定为 Spark 应用的诊断结果。上述例子中, 数据倾斜情况的诊断评分 2 分、队列资源不足情况的诊断评分为 4 分、内存超出情况的诊断评分为 1 分, 则将队列资源不足情况的诊断结果确定为 Spark 应用的诊断结果。

此外, 诊断指标对应的诊断结果不仅可以包括各诊断指标的诊断评分, 还可以包括各诊断指标的诊断码、诊断信息等, 例如, 数据倾斜情况的诊断码为 d10001, 诊断信息为当前的 Task 的最大执行时长是 45min, 中位数执行时长为 4min, 对应的输入记录数分别为 300 条和 40 条。

当然, 本申请实施例中, 还可以采用其他预设规则, 从多个诊断指标对应的诊断结果中确定符合预设规则的诊断结果, 并确定为 Spark 应用的诊断结果, 其他预设规则如取诊断评分中最小值对应的诊断结果, 或取诊断码中权重最大的诊断码对应的诊断结果等, 当然, 也可以根据多个诊断指标对应的诊断结果, 综合权重或者其他参数确定出 Spark 应用的诊断结果, 在此不做限制。

本申请实施例中，还设置有预设数据库，该预设数据库中预先设置有诊断码与诊断措施（解决方案）的对应关系。在确定出 Spark 应用的诊断结果之后，还可以根据上述诊断结果中的诊断码，从预设数据库中获取与诊断结果中的诊断码相对应的诊断措施，从而报告给用户。举例来说，数据倾斜场景会获取到数据倾斜处理解决方案，队列资源不足场景会获取到队列资源不足处理解决方案。该技术方案，可以在确定出 Spark 应用的诊断结果后，向用户提供有针对性的诊断措施，即解决方案，方便用户根据诊断措施自主解决问题，从而及时解决 Spark 应用在运行中的问题。无需用户查询相关资料用于解决 Spark 应用的运行问题，而是直接设置好相关的解决方案并提供给用户，提高用户解决问题的效率，且提高用户体验。

本申请实施例可应用于金融科技（Fintech）领域，金融科技领域是指将信息技术融入金融领域后，为金融领域带来的一种新的创新科技，通过使用先进的信息技术辅助实现金融作业、交易执行以及金融系统改进，可以提升金融系统的处理效率、业务规模，并可以降低成本和金融风险。示例性的，可以在银行中使用 Spark 做用户的白名单分析和黑名单分析，可以在银行中基于 Spark 执行 ETL（Extract-transform-load，数据抽取、清洗、转换、装载）作业，在执行各种 Spark 应用时，可以执行 Spark 运行过程中的诊断，进行实时诊断以便能够实时的监测 Spark 应用运行的正常性。

上述技术方案中，通过获取 Spark 应用的上下文信息，生成诊断指标和指标规则，实现在 Spark 应用过程中实时获取诊断指标对应的运行信息，并对运行信息进行诊断，确定出 Spark 应用的诊断结果，从而可以对 Spark 应用运行中出现的运行故障进行实时诊断，并获取到诊断结果，进一步的，在 Spark 应用运行中获取运行信息，可以较全面的采集到 Spark 应用运行中的参数指标等，相比于在 Spark 应用运行完成后的运行日志，运行参数更为全面且反应 Spark 应用的当前运行状态。

基于同一构思，图 3 示例性的示出了本申请实施例提供的一种诊断 Spark 应用的装置的结构，该装置可以执行诊断 Spark 应用的方法的流程。该装置可

以以软件或硬件的形式存在。装置可以包括：处理单元 302 和获取单元 301。作为一种实现方式，该获取单元 301 可以包括接收单元，装置还可以包括发送单元。处理单元 302 用于对装置的动作进行控制管理。获取单元 301 和发送单元用于支持装置与其他网络实体的通信。

5 其中，处理单元 302 可以是处理器或控制设备，例如可以是通用中央处理器（central processing unit, CPU），通用处理器，数字信号处理（digital signal processing, DSP），专用集成电路（application specific integrated circuits, ASIC），现场可编程门阵列（field programmable gate array, FPGA）或者其他可编程逻辑器件、晶体管逻辑器件、硬件部件或者其任意组合。其可以实现或执行结
10 合本申请公开内容所描述的各种示例性的逻辑方框，模块和电路。处理器也可以是实现计算功能的组合，例如包括一个或多个微处理器组合，DSP 和微处理器的组合等等。获取单元 301 是一种该装置的接口电路，用于从其它装置接收信号。例如，当该装置以芯片的方式实现时，该获取单元 301 是该芯片用于从其它芯片或装置接收信号的接口电路，发送单元是该芯片用于向其
15 它芯片或装置发送信号的接口电路。

 该装置可以为上述实施例中的运行诊断器 100，还可以为用于运行诊断器 100 的芯片。例如，当装置为运行诊断器 100 时，该处理单元 302 例如可以是处理器，该获取单元 301 例如可以是收发器。可选的，该收发器可以包括射频电路，该存储单元例如可以是存储器。例如，当装置为用于运行诊断器 100
20 的芯片时，该处理单元 302 例如可以是处理器，该获取单元 301 例如可以是输入/输出接口、管脚或电路等。该处理单元 302 可执行存储单元存储的计算机执行指令，可选地，该存储单元为该芯片内的存储单元，如寄存器、缓存等，该存储单元还可以是第一转发服务器内的位于该芯片外部的存储单元，如只读存储器（read-only memory, ROM）或可存储静态信息和指令的其他类型
25 的静态存储设备，随机存取存储器（random access memory, RAM）等。

 在一种实施例中，该装置为上述实施例中的运行诊断器 100。其中，获取单元 301，用于获取 Spark 应用的上下文信息；处理单元 302，用于根据上下

文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则；根据 Spark 应用的诊断指标，通过获取单元 301 采集 Spark 应用在运行过程中诊断指标对应的运行信息；根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断，确定 Spark 应用的诊断结果。

5 可选的，诊断指标为多个；处理单元 302 具体用于：针对任一个诊断指标，根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断，确定诊断指标对应的诊断结果；从多个诊断指标对应的诊断结果中确定出符合预设规则的诊断结果，并确定为 Spark 应用的诊断结果。

10 可选的，处理单元 302 还用于：在确定 Spark 应用的诊断结果之后，根据符合预设规则的诊断结果中的诊断码，从预设数据库中获取符合预设规则的诊断结果中的诊断码对应的诊断措施并报告给用户；其中，预设数据库中预先设置有诊断码与诊断措施的对应关系。

15 可选的，处理单元 302 具体用于：将诊断指标对应的运行信息进行统一化处理后，生成诊断指标对应的运行指标；根据诊断指标对应的指标规则对诊断指标对应的运行指标进行诊断。

 可选的，处理单元 302 还用于：在根据上下文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则之前，通过获取单元 301 获取用户配置信息；根据用户配置信息、上下文信息，确定 Spark 应用的诊断指标和诊断指标对应的指标规则。

20 基于同一构思，如图 4 所示，本申请实施例还提供了一种计算设备 400，该计算设备 400 可以是上述实施例中的运行诊断器。计算设备 400 包括：处理器 402 和通信接口 403，可选的，计算设备 400 还可以包括存储器 401。可选的，计算设备 400 还可以包括通信线路 404。其中，通信接口 403、处理器 402 以及存储器 401 可以通过通信线路 404 相互连接；通信线路 404 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。通信线路 404 可以分为地址总线、数据总线、控制总线等。为便于表示，图 4

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。

处理器 402 可以是一个 CPU，微处理器，ASIC，或一个或多个用于控制本申请方案程序执行的集成电路。

一种可能的实施例中，处理器 402，可以用于：根据上下文信息确定 Spark 应用的诊断指标和诊断指标对应的指标规则；根据 Spark 应用的诊断指标，通过通信接口 403 采集 Spark 应用在运行过程中诊断指标对应的运行信息；根据诊断指标对应的指标规则对诊断指标对应的运行信息进行诊断，确定 Spark 应用的诊断结果。

通信接口 403，使用任何收发器一类的装置，用于与其他设备或通信网络通信，如以太网，无线接入网（radio access network, RAN），无线局域网（wireless local area networks, WLAN），有线接入网等。

存储器 401 可以是 ROM 或可存储静态信息和指令的其他类型的静态存储设备，RAM 或者可存储信息和指令的其他类型的动态存储设备，也可以是电可擦可编程只读存储器（electrically erasable programmable read-only memory, EEPROM）、只读光盘（compact disc read-only memory, CD-ROM）或其他光盘存储、光碟存储（包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等）、磁盘存储介质或者其他磁存储设备、或者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。存储器可以是独立存在，通过通信线路 404 与处理器相连接。

存储器也可以和处理器集成在一起。

其中，存储器 401 用于存储执行本申请方案的计算机执行指令，并由处理器 402 来控制执行。处理器 402 用于执行存储器 401 中存储的计算机执行指令，从而实现本申请上述实施例提供的方法。

可选的，本申请实施例中的计算机执行指令也可以称之为应用程序代码，本申请实施例对此不作具体限定。

基于同一发明构思，本申请实施例还提供了一种计算机可读非易失性存储介质，包括计算机可读指令，当计算机读取并执行计算机可读指令时，使

得计算机执行上述诊断 Spark 应用的方法。

本申请是参照根据本申请实施例的方法、设备（系统）、和计算机程序产品的流程图和 / 或方框图来描述的。应理解可由计算机程序指令实现流程图和 / 或方框图中的每一流程和 / 或方框、以及流程图和 / 或方框图中的流程和 / 或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

尽管已描述了本申请的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

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

权 利 要 求

1、一种诊断 Spark 应用的方法，其特征在于，包括：

获取 Spark 应用的上下文信息；

根据所述上下文信息确定所述 Spark 应用的诊断指标和所述诊断指标对
5 应的指标规则；

根据所述 Spark 应用的诊断指标，采集所述 Spark 应用在运行过程中所述
诊断指标对应的运行信息；

根据所述诊断指标对应的指标规则对所述诊断指标对应的运行信息进行
诊断，确定所述 Spark 应用的诊断结果。

2、如权利要求 1 所述的方法，其特征在于，所述诊断指标为多个；

所述根据所述诊断指标对应的指标规则对所述诊断指标对应的运行信息
10 进行诊断，确定所述 Spark 应用的诊断结果，包括：

针对任一个诊断指标，根据所述诊断指标对应的指标规则对所述诊断指
标对应的运行信息进行诊断，确定所述诊断指标对应的诊断结果；

15 从多个所述诊断指标对应的诊断结果中确定出符合预设规则的诊断结果，
并确定为所述 Spark 应用的诊断结果。

3、如权利要求 1 或 2 所述的方法，其特征在于，在所述确定所述 Spark
应用的诊断结果之后，还包括：

根据所述符合预设规则的诊断结果中的诊断码，从预设数据库中获取所
20 述符合预设规则的诊断结果中的诊断码对应的诊断措施并报告给用户；所述
预设数据库中预先设置有诊断码与诊断措施的对应关系。

4、如权利要求 1-3 任一项所述的方法，其特征在于，所述根据所述诊断
指标对应的指标规则对所述诊断指标对应的运行信息进行诊断，包括：

将所述诊断指标对应的运行信息进行统一化处理后，生成所述诊断指标
25 对应的运行指标；

根据所述诊断指标对应的指标规则对所述诊断指标对应的运行指标进行

诊断。

5、如权利要求 1 至 4 任一项所述的方法，其特征在于，在所述根据所述上下文信息确定所述 Spark 应用的诊断指标和所述诊断指标对应的指标规则之前，还包括：

5 获取用户配置信息；

所述根据所述上下文信息确定所述 Spark 应用的诊断指标和所述诊断指标对应的指标规则，包括：

根据所述用户配置信息、所述上下文信息，确定所述 Spark 应用的诊断指标和所述诊断指标对应的指标规则。

10 6、一种诊断 Spark 应用的装置，其特征在于，包括：

获取单元，用于获取 Spark 应用的上下文信息；

处理单元，用于根据所述上下文信息确定所述 Spark 应用的诊断指标和所述诊断指标对应的指标规则；根据所述 Spark 应用的诊断指标，通过所述获取单元采集所述 Spark 应用在运行过程中所述诊断指标对应的运行信息；根据所述诊断指标对应的指标规则对所述诊断指标对应的运行信息进行诊断，确定所述 Spark 应用的诊断结果。

7、如权利要求 6 所述的装置，其特征在于，所述诊断指标为多个；所述处理单元具体用于：针对任一个诊断指标，根据所述诊断指标对应的指标规则对所述诊断指标对应的运行信息进行诊断，确定所述诊断指标对应的诊断结果；从多个所述诊断指标对应的诊断结果中确定出符合预设规则的诊断结果，并确定为所述 Spark 应用的诊断结果。

8、如权利要求 6-7 任一项所述的装置，其特征在于，所述处理单元还用于：在所述确定所述 Spark 应用的诊断结果之后，根据所述符合预设规则的诊断结果中的诊断码，通过所述获取单元从预设数据库中获取所述符合预设规则的诊断结果中的诊断码对应的诊断措施并报告给用户；其中，所述预设数据库中预先设置有诊断码与诊断措施的对应关系。

9、如权利要求 6-8 任一项所述的装置，其特征在于，所述处理单元具体

用于：将所述诊断指标对应的运行信息进行统一化处理后，生成所述诊断指标对应的运行指标；根据所述诊断指标对应的指标规则对所述诊断指标对应的运行指标进行诊断。

5 10、如权利要求 6 至 9 任一项所述的装置，其特征在于，所述处理单元还用于：在所述根据所述上下文信息确定所述 Spark 应用的诊断指标和所述诊断指标对应的指标规则之前，通过所述获取单元获取用户配置信息；根据所述用户配置信息、所述上下文信息，确定所述 Spark 应用的诊断指标和所述诊断指标对应的指标规则。

11、一种计算设备，其特征在于，包括：

10 存储器，用于存储程序指令；

处理器，用于调用所述存储器中存储的程序指令，按照获得的程序执行权利要求 1 至 5 任一项所述的方法。

12、一种计算机可读非易失性存储介质，其特征在于，包括计算机可读指令，当计算机读取并执行所述计算机可读指令时，使得计算机执行如权利
15 要求 1 至 5 任一项所述的方法。

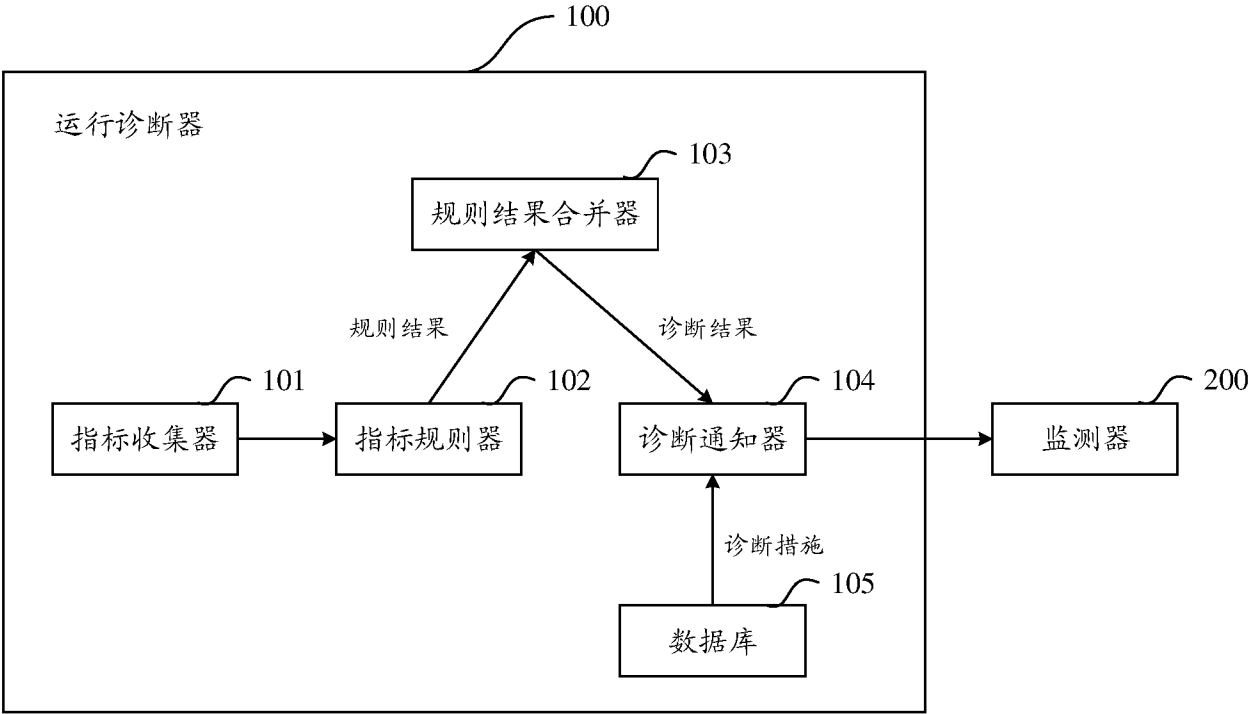


图 1

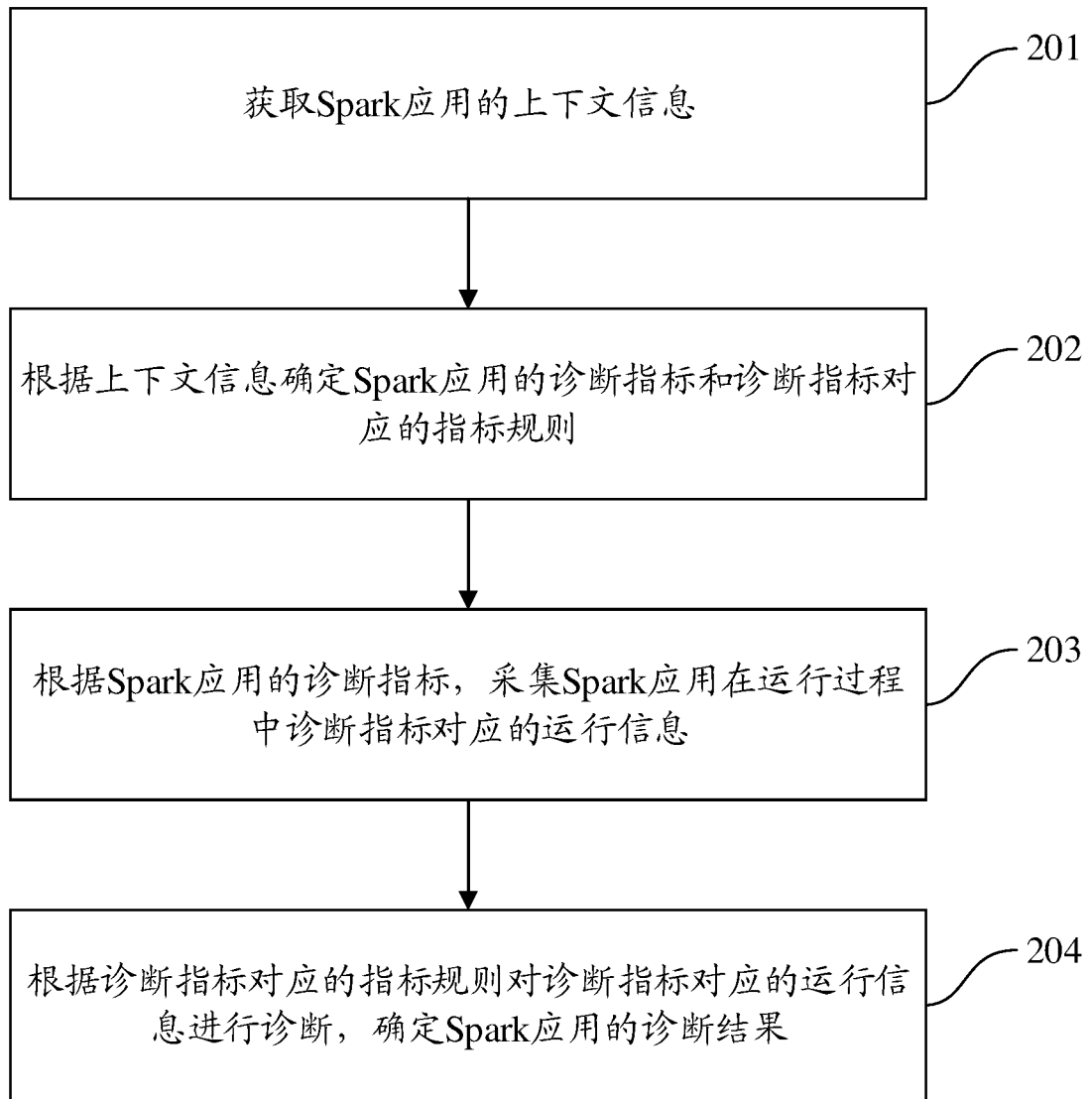


图 2

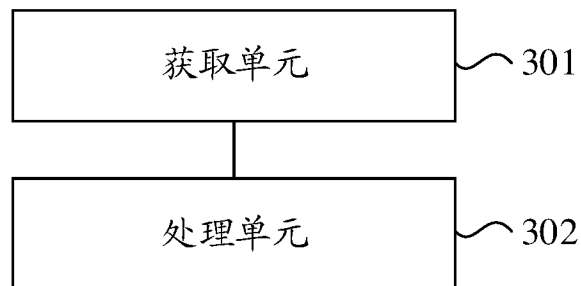


图 3

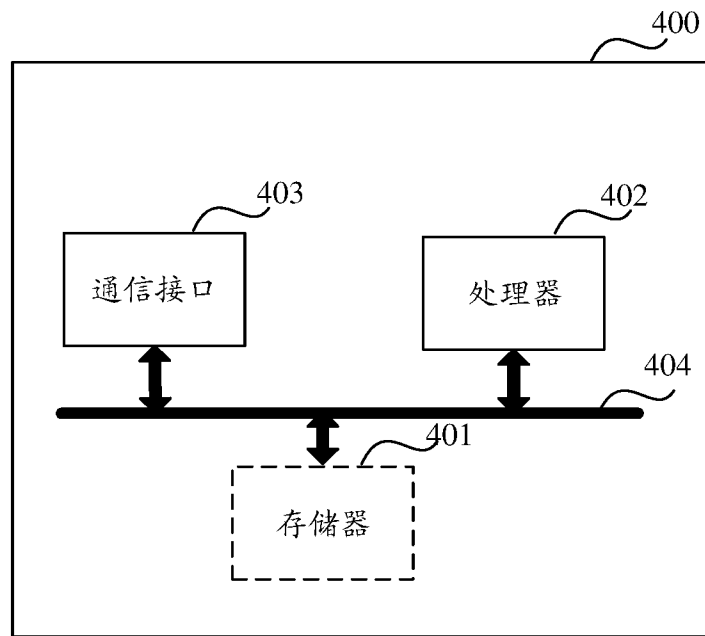


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/083381

A. CLASSIFICATION OF SUBJECT MATTER

G06F 11/36(2006.01)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)

G06F11

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)

CNABS; CNTXT; VEN; USTXT; WOTXT; EPTXT; CNKI: 深圳前海微众, 应用, 软件, 程序, 诊断, 评估, 评测, 王和平, 尹强, 刘有, 黄山, 杨峙岳, 邱帅, 卢道和, 实时, 上下文, 指标, 运行, 日志, 处理, 修复, 方法, 措施, 格式, 分析, 判断, 及时, 问题, 错误, 故障, spark, application, software, program, app, diagnose, assess, evaluate, real, actual, current, time, context, index, run, log, dispose, manage, deal, renovate, method, measure, format, analyze, judge, in time, timely, problem, error, fault, malfunction

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110175124 A (SHENZHEN QIANHAI WEBANK CO., LTD.) 27 August 2019 (2019-08-27) description, paragraphs 6-48	1-12
Y	CN 107992406 A (KYLAND TECHNOLOGY CO., LTD. et al.) 04 May 2018 (2018-05-04) paragraphs 69-142	1-12
Y	马志亨 (MA, Zhiheng). "基于Spark的日志分析工具的设计与实现 (Non-official translation: Design and Implementation of Spark-based Log Analysis Tool)" 中国优秀硕士学位论文全文数据库信息科技辑 (Information & Technology, China Master's Theses Full-Text Database), No. 4, 15 April 2018 (2018-04-15), ISSN: 1674-0246, pages 1138-2067	1-12
A	CN 103412805 A (BANK OF COMMUNICATIONS CO., LTD.) 27 November 2013 (2013-11-27) entire document	1-12
A	US 2009132860 A1 (INVENTEC CORP.) 21 May 2009 (2009-05-21) entire document	1-12



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

08 May 2020

Date of mailing of the international search report

11 June 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/CN2020/083381

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	110175124	A	27 August 2019	None	
CN	107992406	A	04 May 2018	None	
CN	103412805	A	27 November 2013	None	
US	2009132860	A1	21 May 2009	None	

国际检索报告

国际申请号

PCT/CN2020/083381

A. 主题的分类

G06F 11/36 (2006.01) 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)

G06F11

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))

CNABS; CNTXT; VEN; USTXT; WOTXT; EPTXT; CNKI: 深圳前海微众, 应用, 软件, 程序, 诊断, 评估, 评测, 王和平, 尹强, 刘有, 黄山, 杨峙岳, 邸帅, 卢道和, 实时, 上下文, 指标, 运行, 日志, 处理, 修复, 方法, 措施, 格式, 分析, 判断, 及时, 问题, 错误, 故障, spark, application, software, program, app, diagnose, assess, evaluate, real, actual, current, time, context, index, run, log, dispose, manage, deal, renovate, method, measure, format, analyze, judge, in time, timely, problem, error, fault, malfunction

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
PX	CN 110175124 A (深圳前海微众银行股份有限公司) 2019年 8月 27日 (2019 - 08 - 27) 说明书第6-48段	1-12
Y	CN 107992406 A (北京东土科技股份有限公司 等) 2018年 5月 4日 (2018 - 05 - 04) 第69-142段	1-12
Y	马志亨. “基于Spark的日志分析工具的设计与实现” 中国优秀硕士学位论文全文数据库信息科技辑, 第4期, 2018年 4月 15日 (2018 - 04 - 15), ISSN: 1674-0246, 第1138-2067页	1-12
A	CN 103412805 A (交通银行股份有限公司) 2013年 11月 27日 (2013 - 11 - 27) 全文	1-12
A	US 2009132860 A1 (INVENTEC CORP) 2009年 5月 21日 (2009 - 05 - 21) 全文	1-12

☐ 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 5月 8日

国际检索报告邮寄日期

2020年 6月 11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

张帅领

电话号码 86-(20)-28950721

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/083381

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	110175124	A	2019年 8月 27日	无	
CN	107992406	A	2018年 5月 4日	无	
CN	103412805	A	2013年 11月 27日	无	
US	2009132860	A1	2009年 5月 21日	无	