



(12) 发明专利申请

(10) 申请公布号 CN 112199935 A

(43) 申请公布日 2021.01.08

(21) 申请号 202011015168.1

(22) 申请日 2020.09.24

(71) 申请人 建信金融科技有限责任公司
地址 200120 上海市浦东新区自由贸易试
验区银城路99号12层、15层

(72) 发明人 张同虎 王文娟

(74) 专利代理机构 北京市兰台律师事务所
11354

代理人 张峰

(51) Int. Cl.

G06F 40/194 (2020.01)

G06F 40/205 (2020.01)

G06F 16/182 (2019.01)

G06Q 40/02 (2012.01)

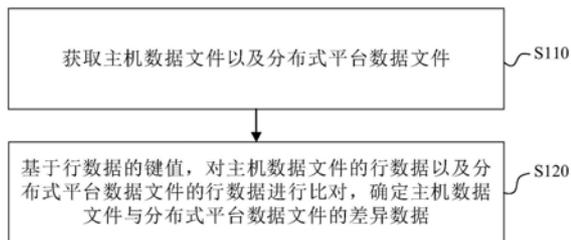
权利要求书2页 说明书10页 附图3页

(54) 发明名称

数据的比对方法、装置、电子设备及计算机
可读存储介质

(57) 摘要

本申请实施例提供了一种数据的比对方法、
装置、电子设备及计算机可读存储介质。该方法
包括：获取主机数据文件以及分布式平台数据文
件；基于行数据的键值，对主机数据文件的行数
据以及分布式平台数据文件的行数据进行比对，
确定主机数据文件与分布式平台数据文件的差
异数据。基于本方案，能够实现对中批量账务数
据量的自动对比，能够满足主机下移的过程中批
量账务数据量的对比需求。



1. 一种数据的比对方法,其特征在于,包括:
获取主机数据文件以及分布式平台数据文件;
基于行数据的键值,对所述主机数据文件的所述行数据以及所述分布式平台数据文件的行数据进行比对,确定所述主机数据文件与所述分布式平台数据文件的差异数据。
2. 根据权利要求1所述的方法,其特征在于,所述获取主机数据文件以及分布式平台数据文件,包括:
从NAS存储节点分别获取主机数据文件以及分布式平台数据文件。
3. 根据权利要求1所述的方法,其特征在于,所述基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,包括:
确定所述主机数据文件的行数据的键值是否唯一,以及所述分布式平台数据文件的行数据的键值是否唯一;
基于所述主机数据文件的行数据的键值是否唯一,以及所述分布式平台数据文件的行数据的键值是否唯一,对所述主机数据文件的行数据以及所述分布式平台数据文件的行数据进行比对。
4. 根据权利要求3所述的方法,其特征在于,所述基于所述主机数据文件的行数据的键值是否唯一,以及所述分布式平台数据文件的行数据的键值是否唯一,对所述主机数据文件的行数据以及所述分布式平台数据文件的行数据进行比对,包括:
若所述主机数据文件以及所述分布式平台数据文件中存在键值唯一的行数据,则将所述主机数据文件与所述分布式平台数据文件中相同键值的行数据进行比对;
若所述主机数据文件以及所述分布式平台数据文件中存在键值不唯一的行数据,则将主机数据文件与所述分布式平台数据文件中键值不唯一的行数据中相同键值的行数据进行比对。
5. 根据权利要求4所述的方法,其特征在于,所述将所述主机数据文件与所述分布式平台数据文件中相同键值的行数据进行比对,包括:
通过Shuffle操作获取所述主机数据文件与所述分布式平台数据文件中相同键值的行数据,并进行比对。
6. 根据权利要求1所述的方法,其特征在于,所述对主机数据文件的行数据以及所述分布式平台数据文件的行数据进行比对,包括:
将所述主机数据文件以及所述分布式平台数据文件中的行数据解析成列,并提取所述列的列值;
基于所述列值进行所述主机数据文件的行数据以及所述分布式平台数据文件的行数据的比对。
7. 根据权利要求6所述的方法,其特征在于,所述将所述主机数据文件以及所述分布式平台数据文件中的行数据解析成列,包括:
确定所述主机数据文件以及所述分布式平台数据文件所属的文件类型;
基于所述文件类型确定对应的解析规则;
基于所述解析规则将所述主机数据文件以及所述分布式平台数据文件中的行数据解析成列。
8. 根据权利要求7所述的方法,其特征在于,若所述文件类型为字符分隔型,所述基于

所述解析规则将所述主机数据文件以及所述分布式平台数据文件中的行数据解析成列,包括:

基于预配置的分隔符对所述主机数据文件以及所述分布式平台数据文件中的行数据进行切割,得到列。

9. 根据权利要求7所述的方法,其特征在于,若所述文件类型为定长型,所述基于所述解析规则将所述主机数据文件以及所述分布式平台数据文件中的行数据解析成列,包括:

基于字段的起始位置以及字段的字符长度,将所述主机数据文件以及所述分布式平台数据文件中的行数据解析成列。

10. 一种数据的比对装置,其特征在于,包括:

文件获取模块,用于获取主机数据文件以及分布式平台数据文件;

数据比对模块,用于基于行数据的键值,对所述主机数据文件的行数据以及所述分布式平台数据文件的行数据进行比对,确定所述主机数据文件与分布式平台数据文件的差异数据。

11. 一种电子设备,其特征在于,包括处理器和存储器;

所述存储器,用于存储操作指令;

所述处理器,用于通过调用所述操作指令,执行权利要求1-9中任一项所述的方法。

12. 一种计算机可读存储介质,其特征在于,所述存储介质上存储有计算机程序,所述计算机程序被处理器执行时实现权利要求1-9中任一项所述的方法。

数据的比对方法、装置、电子设备及计算机可读存储介质

技术领域

[0001] 本申请涉及计算机技术领域,具体而言,本申请涉及一种数据的比对方法、装置、电子设备及计算机可读存储介质。

背景技术

[0002] 近年来,全球IT市场发生了巨大的变化,主机由于有高可用、高吞吐率的特点备受银行业青睐,但随着分布式架构的崛起,主机面临着严峻的挑战。目前,各大型商业银行都在积极探索“主机下移”的解决方案。“主机下移”指的是将主机上的部分系统部署到分布式平台,从集中式部署向分布式部署转变。“主机下移”旨在减少对主机的单方面依赖,实现自主可控,降低安全风险和成本。

[0003] 在主机下移的过程中,为保证下移后的数据和现行生产数据一致,验证开放平台的程序是否可以达到和现行生产上主机程序有同样的效果,需要进行账务一致数据比对,即针对相同的输入数据,经过分布式平台的程序和主机程序跑批,比较两个环境结果数据,若在某测试时间段内持续保持数据一致,那么说明系统重构正确,若有差异数据,说明系统重构有缺陷,如:主档数据、财会流水文件等进行数据比对,确保重构正确性。

[0004] 为了满足主机下移的过程中批量账务数据量的对比需求,亟需提供一种能够对金融领域批量账务一致数据比对方案。

发明内容

[0005] 本申请的目的旨在至少能解决上述的技术缺陷之一。本申请所采用的技术方案如下:

[0006] 第一方面,本申请实施例提供了一种数据的比对方法,该方法包括:

[0007] 获取主机数据文件以及分布式平台数据文件;

[0008] 基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,确定主机数据文件与分布式平台数据文件的差异数据。

[0009] 可选地,获取主机数据文件以及分布式平台数据文件,包括:

[0010] 从NAS存储节点分别获取主机数据文件以及分布式平台数据文件。

[0011] 可选地,基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,包括:

[0012] 确定主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一;

[0013] 基于主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对。

[0014] 可选地,基于主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,包括:

- [0015] 若主机数据文件以及分布式平台数据文件中存在键值唯一的行数据,则将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对;
- [0016] 若主机数据文件以及分布式平台数据文件中存在键值不唯一的行数据,则将主机数据文件与分布式平台数据文件中键值不唯一的行数据中相同键值的行数据进行比对。
- [0017] 可选地,将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对,包括:
- [0018] 通过Shuffle操作获取主机数据文件与分布式平台数据文件中相同键值的行数据,并进行比对。
- [0019] 可选地,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,包括:
- [0020] 将主机数据文件以及分布式平台数据文件中的行数据解析成列,并提取列的列值;
- [0021] 基于列值进行主机数据文件的行数据以及分布式平台数据文件的行数据的比对。
- [0022] 可选地,将主机数据文件以及分布式平台数据文件中的行数据解析成列,包括:
- [0023] 确定主机数据文件以及分布式平台数据文件所属的文件类型;
- [0024] 基于文件类型确定对应的解析规则;
- [0025] 基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列。
- [0026] 可选地,若文件类型为字符分隔型,基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列,包括:
- [0027] 基于预配置的分隔符对主机数据文件以及分布式平台数据文件中的行数据进行切割,得到列。
- [0028] 可选地,若文件类型为定长型,基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列,包括:
- [0029] 基于字段的起始位置以及字段的字符长度,将主机数据文件以及分布式平台数据文件中的行数据解析成列。
- [0030] 第二方面,本申请实施例提供了一种数据的比对装置,该装置包括:
- [0031] 文件获取模块,用于获取主机数据文件以及分布式平台数据文件;
- [0032] 数据比对模块,用于基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,确定主机数据文件与分布式平台数据文件的差异数据。
- [0033] 可选地,文件获取模块具体用于:
- [0034] 从NAS存储节点分别获取主机数据文件以及分布式平台数据文件。
- [0035] 可选地,数据比对模块在基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对时,具体用于:
- [0036] 确定主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一;
- [0037] 基于主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对。
- [0038] 可选地,数据比对模块在基于主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一,对主机数据文件的行数据以及分布式平台数据

文件的行数据进行比对时,具体用于:

[0039] 若主机数据文件以及分布式平台数据文件中存在键值唯一的行数据,则将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对;

[0040] 若主机数据文件以及分布式平台数据文件中存在键值不唯一的行数据,则将主机数据文件与分布式平台数据文件中键值不唯一的行数据中相同键值的行数据进行比对。

[0041] 可选地,数据比对模块在将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对时,具体用于:

[0042] 通过Shuffle操作获取主机数据文件与分布式平台数据文件中相同键值的行数据,并进行比对。

[0043] 可选地,数据比对模块在对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对时,具体用于:

[0044] 将主机数据文件以及分布式平台数据文件中的行数据解析成列,并提取列的列值;

[0045] 基于列值进行主机数据文件的行数据以及分布式平台数据文件的行数据的比对。

[0046] 可选地,数据比对模块在将主机数据文件以及分布式平台数据文件中的行数据解析成列时,具体用于:

[0047] 确定主机数据文件以及分布式平台数据文件所属的文件类型;

[0048] 基于文件类型确定对应的解析规则;

[0049] 基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列。

[0050] 可选地,若文件类型为字符分隔型,数据比对模块在基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列时,具体用于:

[0051] 基于预配置的分隔符对主机数据文件以及分布式平台数据文件中的行数据进行切割,得到列。

[0052] 可选地,若文件类型为定长型,数据比对模块在基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列时,具体用于:

[0053] 基于字段的起始位置以及字段的字符长度,将主机数据文件以及分布式平台数据文件中的行数据解析成列。

[0054] 第三方面,本申请实施例提供了一种电子设备,该电子设备包括:处理器和存储器;

[0055] 存储器,用于存储操作指令;

[0056] 处理器,用于通过调用操作指令,执行如本申请的第一方面的任一实施方式中所示的数据的比对方法。

[0057] 第四方面,本申请实施例提供了一种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本申请的第一方面的任一实施方式中所示的数据的比对方法。

[0058] 本申请实施例提供的技术方案带来的有益效果是:

[0059] 本申请实施例提供的方案,通过获取主机数据文件以及分布式平台数据文件,并且基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,从而确定主机数据文件与分布式平台数据文件的差异数据。基于本方案,能够实现对中

批量账务数据量的自动对比,能够满足主机下移的过程中批量账务数据量的对比需求。

附图说明

[0060] 为了更清楚地说明本申请实施例中的技术方案,下面将对本申请实施例描述中所需要使用的附图作简单地介绍。

[0061] 图1为本申请实施例提供了一种数据的比对方法的流程示意图;

[0062] 图2为本申请实施例提供了一种数据的比对方法的具体实施方式流程示意图;

[0063] 图3为本申请实施例提供了一种数据的比对装置的结构示意图;

[0064] 图4为本申请实施例提供了一种电子设备的结构示意图。

具体实施方式

[0065] 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,仅用于解释本申请,而不能解释为对本发明的限制。

[0066] 本技术领域技术人员可以理解,除非特意声明,这里使用的单数形式“一”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是,本申请的说明书中使用的措辞“包括”是指存在所述特征、整数、步骤、操作、元件和/或组件,但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、组件和/或它们的组。应该理解,当我们称元件被“连接”或“耦接”到另一元件时,它可以直接连接或耦接到其他元件,或者也可以存在中间元件。此外,这里使用的“连接”或“耦接”可以包括无线连接或无线耦接。这里使用的措辞“和/或”包括一个或多个相关联的列出项的全部或任一单元和全部组合。

[0067] 下面以具体地实施例对本申请的技术方案以及本申请的技术方案如何解决上述技术问题进行详细说明。下面这几个具体的实施例可以相互结合,对于相同或相似的概念或过程可能在某些实施例中不再赘述。下面将结合附图,对本申请的实施例进行描述。

[0068] 图1示出了本申请实施例提供了一种数据的比对方法的流程示意图,如图1所示,该方法主要可以包括:

[0069] 步骤S110:获取主机数据文件以及分布式平台数据文件;

[0070] 步骤S120:基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,确定主机数据文件与分布式平台数据文件的差异数据。

[0071] 本申请实施例中,可以在进行数据比对时,获取待比对的主机数据文件以及分布式平台数据文件,并且分别读取主机数据文件的行数据以及分布式平台数据文件的行数据。主机数据文件可以为主机中生成的数据文件,分布式平台数据文件可以为分布式平台中生成的数据文件。

[0072] 本申请实施例中,可以读取各行数据的键值,通过键值区分各行数据,从而对主机数据文件以及分布式平台数据中对应的行数据进行比较。可以通过预设置配置文件,配置行数据的键值,从而读取键值。

[0073] 本申请实施例提供的方法,通过获取主机数据文件以及分布式平台数据文件,并且基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,从而确定主机数据文件与分布式平台数据文件的差异数据。基于本方案,能够实现对中

批量账务数据量的自动对比,能够满足主机下移的过程中批量账务数据量的对比需求。

[0074] 本申请实施例的一种可选方式中,获取主机数据文件以及分布式平台数据文件,包括:

[0075] 从NAS存储节点分别获取主机数据文件以及分布式平台数据文件。

[0076] 本申请实施例中,主机数据文件以及分布式平台数据文件可以被存储于NAS存储节点,以便于在进行数据比对时,直接从NAS存储节点获取主机数据文件以及分布式平台数据文件,节约了数据比对程序中准备数据的步骤,提升比对效率。

[0077] 本申请实施例的一种可选方式中,基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,包括:

[0078] 确定主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一;

[0079] 基于主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对。

[0080] 本申请实施例中,主机数据文件或者分布式平台数据文件中可能存在键值不唯一的行数据,因此可以针对键值是否唯一进行主机数据文件的行数据以及分布式平台数据文件的行数据进行比对。

[0081] 本申请实施例的一种可选方式中,基于主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,包括:

[0082] 若主机数据文件以及分布式平台数据文件中存在键值唯一的行数据,则将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对;

[0083] 若主机数据文件以及分布式平台数据文件中存在键值不唯一的行数据,则将主机数据文件与分布式平台数据文件中键值不唯一的行数据中相同键值的行数据进行比对。

[0084] 本申请实施例中,针对键值唯一的行数据,可以将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对。

[0085] 本申请实施例中,针对键值不唯一的行数据,即存在多条相同键值的行数据,可以将主机数据文件中键值不唯一的行数据以及分布式平台数据文件中键值不唯一的行数据分别进行汇总,而后逐一进行比对,具体而言,在逐一对比的过程中,若确定出主机数据文件中键值不唯一的行数据与分布式平台数据文件中键值不唯一的行数据中相同的行数据,则可以排出该条记录。将主机数据文件中键值不唯一的行数据与分布式平台数据文件中键值不唯一的行数据中存在差异的行数据中作为差异数据输出。

[0086] 本申请实施例的一种可选方式中,将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对,包括:

[0087] 通过Shuffle操作获取主机数据文件与分布式平台数据文件中相同键值的行数据,并进行比对。

[0088] 在主机下移中,由于批量账务数据量对比时,动辄进行上百GB的文件的比对,需比对的数据量巨大,如通过多线程文件比对程序进行文件的比对,则不能合理的使用集群资源,可能会导致无法在规定时间内完成文件比对。

[0089] 本申请实施例中,可以对待比对的批量数据进行分组,并分配至多台不同的节点

进行处理。针对一台节点,在进行数据比对时,可以通过Shuffle操作获取主机数据文件与分布式平台数据文件中相同键值的行数据,从而进行行数据的比较,合理的使用了集群资源,提升了比对效率。

[0090] 本申请实施例的一种可选方式中,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,包括:

[0091] 将主机数据文件以及分布式平台数据文件中的行数据解析成列,并提取列的列值;

[0092] 基于列值进行主机数据文件的行数据以及分布式平台数据文件的行数据的比对。

[0093] 本申请实施例中,在将主机数据文件以及分布式平台数据文件中的行数据进行比对时,可以将行数据解析成列,并且提取列值,从而根据列值完成行数据的比对。

[0094] 本申请实施例的一种可选方式中,将主机数据文件以及分布式平台数据文件中的行数据解析成列,包括:

[0095] 确定主机数据文件以及分布式平台数据文件所属的文件类型;

[0096] 基于文件类型确定对应的解析规则;

[0097] 基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列。

[0098] 本申请实施例中,可以在配置文件中配置对文件类型对应的解析规则,从而在进行数据比对时,确定述主机数据文件以及分布式平台数据文件所属的文件类型,并基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列。

[0099] 本申请实施例的一种可选方式中,若文件类型为字符分隔型,基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列,包括:

[0100] 基于预配置的分隔符对主机数据文件以及分布式平台数据文件中的行数据进行切割,得到列。

[0101] 本申请实施例中,可以在配置文件中配置分隔符,例如“|”或“|@|”,通过分隔符实现对行数据的切割。

[0102] 本申请实施例的一种可选方式中,若文件类型为定长型,基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列,包括:

[0103] 基于字段的起始位置以及字段的字符长度,将主机数据文件以及分布式平台数据文件中的行数据解析成列。

[0104] 本申请实施例中,也可以在配置文件中配置字段的起始位置以及字段的字符长度,从而实现对行数据的切割。

[0105] 本申请实施例中,切割后的数据,可以按照配置文件中是否为key区分,将主机数据文件和分布式平台数据文件的每行按照<Key,Value>的格式输出。

[0106] 图2中示出了本申请实施例提供的一种数据的比对方法的具体实施方式的流程示意图。如图2中所示,在生产环境构建Spark Standalone计算集群,通过master配置对集群的资源进行管理和分配。

[0107] 准备主机数据文件和分布式平台数据文件,主机文件通过跑批生成,分布式平台数据存储于数据库中,需要通过数据库卸数生成,最终数据保存在生产环境各节点的共享NAS上,共享NAS每个节点都可以访问到,可以模拟输入数据在本地情况,节约了将数据再上传到HDFS的时间;

[0108] 准备配置文件,配置文件通过人工配置,文件的字段定义有两种类型:定长型和字符分隔型。定长型需要配置列的开始位置和长度,字符分隔型需要配置索引位置,配置文件中type属性配置了选择哪种方式。配置文件中包含的内容可以具体如表1所示:

[0109] 表1

标签名	标签意义
Name	列名
Index	字段索引,适用于分隔型配置文件,初始值为1
Pos	字段起始位置,适用于定长型配置文件,初始值为1
Length	字段所占字符大小,适用于定长型配置文件
Datatype	目前只支持String,BigDecimal类型
Ifcompare	true/false,若为false说明该列不比对
Iskey	true/false,是否为key

[0111] 执行Spark分布式程序。

[0112] 在程序中,按行读入主机文件、数据库文件和配置文件。根据配置文件确定主机文件和数据库文件类型为字符分隔型还是定长型,若文件为字符分隔型,那么根据配置文件的分隔符(例如:|或|@|)对每行数据进行切割,把每列数据的值取出。若文件为定长型,那么根据配置文件中的字段起始位置和字段所占字符大小对每行数据进行切割,把每列数据的值取出。其中比对分为key值唯一和key值不唯一两种情况,若key值唯一确定一条记录,那么该条记录可以每个字段进行比对。若key值不能唯一确定一条记录,那么对来自两个文件的多条记录逐一比对,若有相同的,则排除该条记录,若有不同的依然作为差异输出。

[0113] 解析切割后的数据,按照配置文件中是否为key区分,将主机文件和数据库文件的每行按照<Key,Value>的格式输出。

[0114] Spark分布式系统进行Shuffle操作,将相同key的数据聚合到同一个结点上,即可对比两个文件中相同key的数据差异,若无差异则略过,若有差异则输出。

[0115] Spark分布式系统的结果会保存在HDFS上,即分布式文件系统,为了方便留存数据,落数到本地NAS。差异具体内容文件的第一行列出了比对该条记录所用的主键字段(在配置文件中通过iskey指出),在这一行下面,使用>>>缩进的部分,是该条记录差异明细,详细列出来该条记录的差异字段在主机文件中的内容和在数据库中的内容。差异分析人员可以定位数据并分析差异原因。

[0116] 本文所阐述的发明要解决的技术问题,能够提供一种金融领域批量账务一致数据比对方法,基于Spark分布式集群实现该方法,可以对贷记卡主机核心处理结果和贷记卡分布式核心处理结果的处理结果进行比对,可有效提高数据比对效率,完成主机下移的时间需求。

[0117] 基于与图1中所示的方法相同的原理,图3示出了本申请实施例提供的一种数据的比对装置的结构示意图,如图3所示,该数据的比对装置20可以包括:

[0118] 文件获取模块210,用于获取主机数据文件以及分布式平台数据文件;

[0119] 数据比对模块220,用于基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,确定主机数据文件与分布式平台数据文件的差异数据。

[0120] 本申请实施例提供的装置,通过获取主机数据文件以及分布式平台数据文件,并

且基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,从而确定主机数据文件与分布式平台数据文件的差异数据。基于本方案,能够实现对中批量账务数据量的自动对比,能够满足主机下移的过程中批量账务数据量的对比需求。

[0121] 可选地,文件获取模块具体用于:

[0122] 从NAS存储节点分别获取主机数据文件以及分布式平台数据文件。

[0123] 可选地,数据比对模块在基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对时,具体用于:

[0124] 确定主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一;

[0125] 基于主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对。

[0126] 可选地,数据比对模块在基于主机数据文件的行数据的键值是否唯一,以及分布式平台数据文件的行数据的键值是否唯一,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对时,具体用于:

[0127] 若主机数据文件以及分布式平台数据文件中存在键值唯一的行数据,则将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对;

[0128] 若主机数据文件以及分布式平台数据文件中存在键值不唯一的行数据,则将主机数据文件与分布式平台数据文件中键值不唯一的行数据中相同键值的行数据进行比对。

[0129] 可选地,数据比对模块在将主机数据文件与分布式平台数据文件中相同键值的行数据进行比对时,具体用于:

[0130] 通过Shuffle操作获取主机数据文件与分布式平台数据文件中相同键值的行数据,并进行比对。

[0131] 可选地,数据比对模块在对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对时,具体用于:

[0132] 将主机数据文件以及分布式平台数据文件中的行数据解析成列,并提取列的列值;

[0133] 基于列值进行主机数据文件的行数据以及分布式平台数据文件的行数据的比对。

[0134] 可选地,数据比对模块在将主机数据文件以及分布式平台数据文件中的行数据解析成列时,具体用于:

[0135] 确定主机数据文件以及分布式平台数据文件所属的文件类型;

[0136] 基于文件类型确定对应的解析规则;

[0137] 基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列。

[0138] 可选地,若文件类型为字符分隔型,数据比对模块在基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列时,具体用于:

[0139] 基于预配置的分隔符对主机数据文件以及分布式平台数据文件中的行数据进行切割,得到列。

[0140] 可选地,若文件类型为定长型,数据比对模块在基于解析规则将主机数据文件以及分布式平台数据文件中的行数据解析成列时,具体用于:

[0141] 基于字段的起始位置以及字段的字符长度,将主机数据文件以及分布式平台数据

文件中的行数据解析成列。

[0142] 可以理解的是,本实施例中的数据的比对装置的上述各模块具有实现图1中所示的实施例中的数据的比对方法相应步骤的功能。该功能可以通过硬件实现,也可以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。上述模块可以是软件和/或硬件,上述各模块可以单独实现,也可以多个模块集成实现。对于上述数据的比对装置的各模块的功能描述具体可以参见图1中所示实施例中的数据的比对方法的对应描述,在此不再赘述。

[0143] 本申请实施例提供了一种电子设备,包括处理器和存储器;

[0144] 存储器,用于存储操作指令;

[0145] 处理器,用于通过调用操作指令,执行本申请任一实施方式中所提供的数据的比对方法。

[0146] 作为一个示例,图4示出了本申请实施例所适用的一种电子设备的结构示意图,如图4所示,该电子设备2000包括:处理器2001和存储器2003。其中,处理器2001和存储器2003相连,如通过总线2002相连。可选的,电子设备2000还可以包括收发器2004。需要说明的是,实际应用中收发器2004不限于一个,该电子设备2000的结构并不构成对本申请实施例的限定。

[0147] 其中,处理器2001应用于本申请实施例中,用于实现上述方法实施例所示的方法。收发器2004可以包括接收机和发射机,收发器2004应用于本申请实施例中,用于执行时实现本申请实施例的电子设备与其他设备通信的功能。

[0148] 处理器2001可以是CPU(Central Processing Unit,中央处理器),通用处理器,DSP(Digital Signal Processor,数据信号处理器),ASIC(Application Specific Integrated Circuit,专用集成电路),FPGA(Field Programmable Gate Array,现场可编程门阵列)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2001也可以是实现计算功能的组合,例如包含一个或多个微处理器组合,DSP和微处理器的组合等。

[0149] 总线2002可包括一通路,在上述组件之间传送信息。总线2002可以是PCI(Peripheral Component Interconnect,外设部件互连标准)总线或EISA(Extended Industry Standard Architecture,扩展工业标准结构)总线等。总线2002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。

[0150] 存储器2003可以是ROM(Read Only Memory,只读存储器)或可存储静态信息和指令的其他类型的静态存储设备,RAM(Random Access Memory,随机存取存储器)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EEPROM(Electrically Erasable Programmable Read Only Memory,电可擦可编程只读存储器)、CD-ROM(Compact Disc Read Only Memory,只读光盘)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。

[0151] 可选的,存储器2003用于存储执行本申请方案的应用程序代码,并由处理器2001来控制执行。处理器2001用于执行存储器2003中存储的应用程序代码,以实现本申请任一实施方式中所提供的数据的比对方法。

[0152] 本申请实施例提供的电子设备,适用于上述方法任一实施例,在此不再赘述。

[0153] 本申请实施例提供了一种电子设备,与现有技术相比,通过获取主机数据文件以及分布式平台数据文件,并且基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,从而确定主机数据文件与分布式平台数据文件的差异数据。基于本方案,能够实现对中批量账务数据量的自动对比,能够满足主机下移的过程中批量账务数据量的对比需求。

[0154] 本申请实施例提供了一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序,该程序被处理器执行时实现上述方法实施例所示的数据的比对方法。

[0155] 本申请实施例提供的计算机可读存储介质,适用于上述方法任一实施例,在此不再赘述。

[0156] 本申请实施例提供了一种计算机可读存储介质,与现有技术相比通过获取主机数据文件以及分布式平台数据文件,并且基于行数据的键值,对主机数据文件的行数据以及分布式平台数据文件的行数据进行比对,从而确定主机数据文件与分布式平台数据文件的差异数据。基于本方案,能够实现对中批量账务数据量的自动对比,能够满足主机下移的过程中批量账务数据量的对比需求。

[0157] 应该理解的是,虽然附图的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,其可以以其他的顺序执行。而且,附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段,这些子步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,其执行顺序也不必然是依次进行,而是可以与其他步骤或者其他步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0158] 以上所述仅是本发明的部分实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

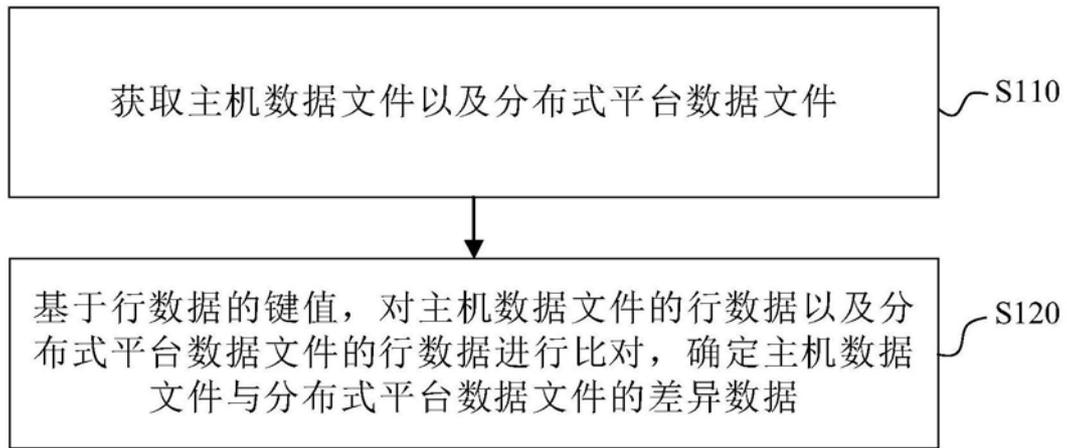


图1

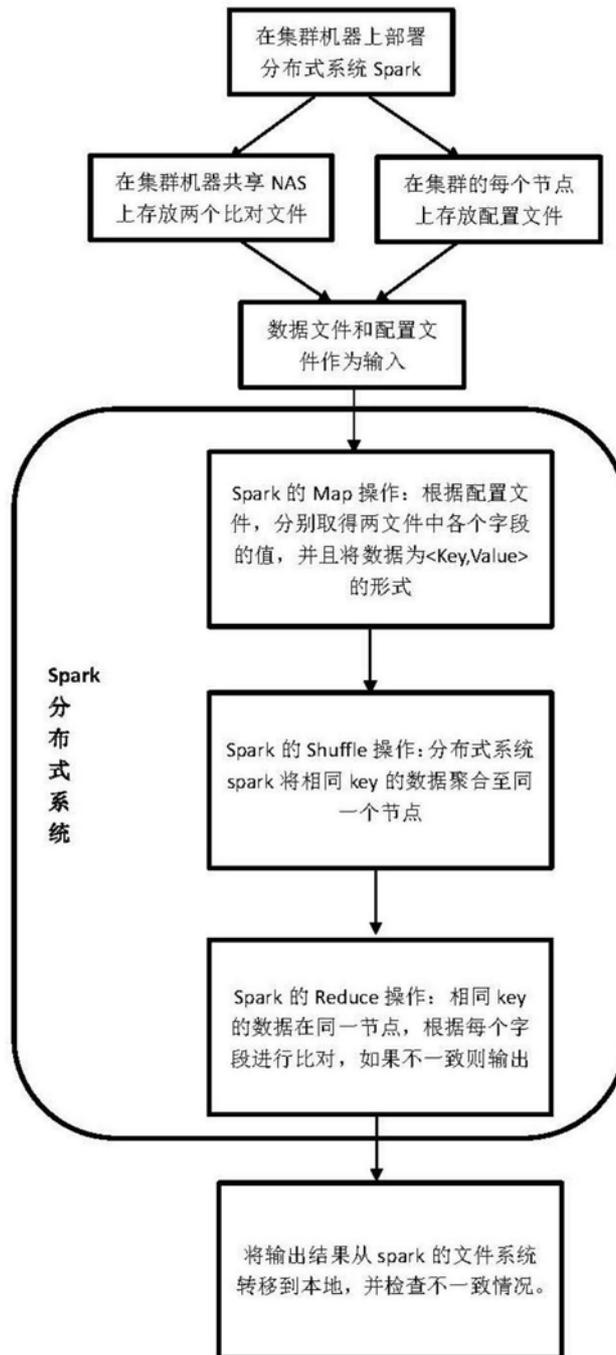


图2

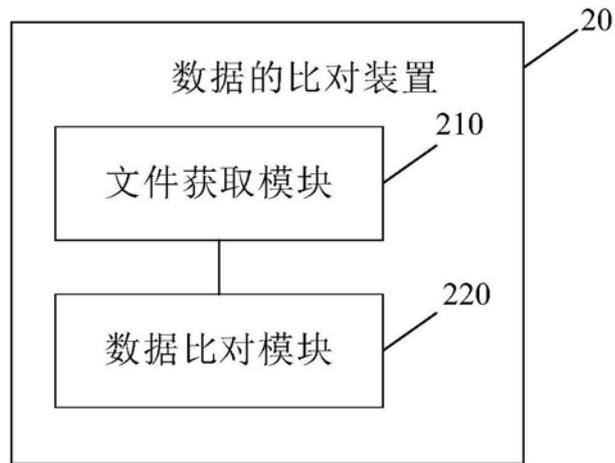


图3

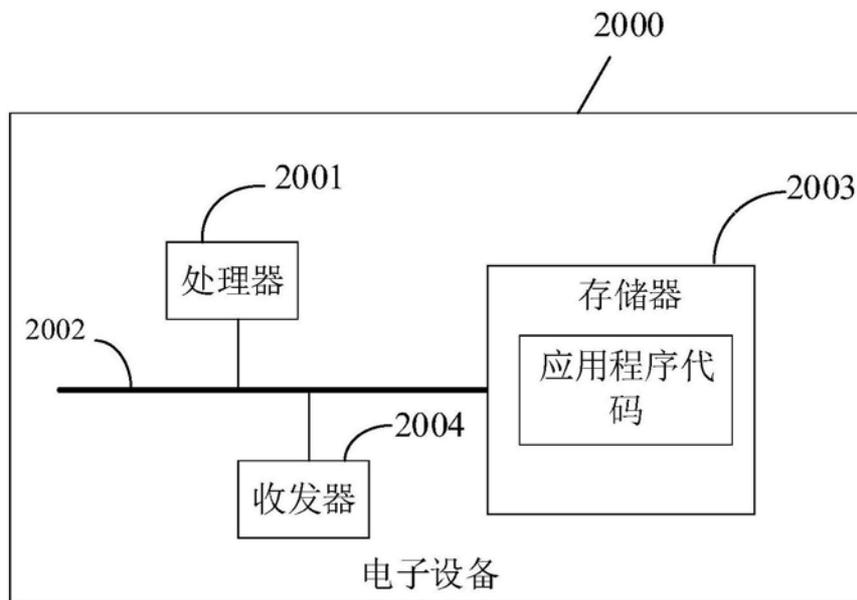


图4