(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 108599992 A
(43)申请公布日 2018.09.28

(21)申请号 201810235937.5

(22)申请日 2018.03.21

(71)申请人 四川斐讯信息技术有限公司
地址 610100 四川省成都市龙泉驿区龙泉街道公园路125号

(72)发明人 魏晓林

(74)专利代理机构 上海硕力知识产权代理事务所(普通合伙) 31251
代理人 郭桂峰

(51)Int.Cl.
H04L 12/24(2006.01)
H04L 12/26(2006.01)
H04L 29/08(2006.01)

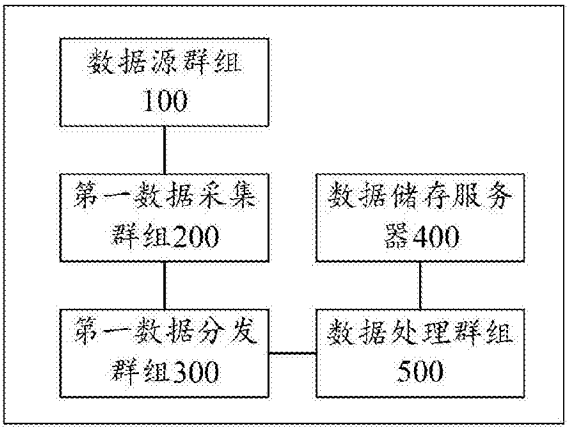
权利要求书2页 说明书9页 附图4页

(54)发明名称

一种数据处理系统及方法

(57)摘要

本发明提供了一种数据处理系统及方法，其系统包括：数据源群组，第一数据采集群组，第一数据分发群组，数据处理群组和数据储存服务器；第一数据采集群组分别与数据源群组和第一数据分发群组连接，数据处理群组分别与第一数据分发群组和数据储存服务器连接；第一数据采集群组，采集数据源群组上传的上传文件，并上传上传文件至第一数据分发群组；第一数据分发群组，分发上传文件至数据处理群组；数据处理群组，实时分析上传文件得到日志信息；日志信息包括特征值；数据储存服务器，根据特征值，分类储存日志信息至对应的储存分区。本发明实现提升数据处理效率和可靠性。



CN 108599992 A

1.一种数据处理系统，其特征在于，包括：数据源群组，第一数据采集群组，第一数据分发群组，数据处理群组和数据储存服务器；所述第一数据采集群组分别与所述数据源群组和所述第一数据分发群组连接，所述数据处理群组分别与所述第一数据分发群组和所述数据储存服务器连接；

所述第一数据采集群组，采集所述数据源群组上传的上传文件，并上传所述上传文件至所述第一数据分发群组；

所述第一数据分发群组，分发所述上传文件至所述数据处理群组；

所述数据处理群组，实时分析所述上传文件得到日志信息；所述日志信息包括特征值；

所述数据储存服务器，根据所述特征值，分类储存所述日志信息至对应的储存分区。

2.根据权利要求1所述的数据处理系统，其特征在于，所述第一数据采集群组包括：若干个第一采集器；所述第一采集器包括：第一获取模块、第一判断模块、第一采集模块和第一控制模块；所述第一获取模块与所述第一判断模块连接，所述第一判断模块分别与所述第一采集模块和所述第一控制模块连接；

所述第一获取模块，获取当前时刻自身的第一状态值；

所述第一判断模块，判断所述第一状态值是否与第一预设状态值匹配；

所述第一采集模块，当判断所述第一状态值与所述第一预设状态值匹配时，采集所述上传文件；

所述第一控制模块，当判断所述第一状态值与所述第一预设状态值不匹配时，标记工作状态为异常，并下发采集开启指令至第一目标采集器。

3.根据权利要求1所述的数据处理系统，其特征在于，还包括：第二数据采集群组，第二数据分发群组；所述第二数据采集群组与所述数据处理群组和所述第二数据分发群组连接，所述第二数据分发群组与所述数据储存服务器连接；

所述第二数据采集群组，从所述数据处理群组获取所述日志信息，并将所述日志信息下发至所述第二数据分发群组；

所述第二数据分发群组，将所述日志信息分发至所述数据储存服务器。

4.根据权利要求3所述的数据处理系统，其特征在于，所述第二数据采集群组包括：若干个第二采集器；所述第二采集器包括：第二获取模块、第二判断模块、第二采集模块和第二控制模块；所述第二获取模块与所述第二判断模块连接，所述第二判断模块分别与所述第二采集模块和所述第二控制模块连接；

所述第二获取模块，获取当前时刻自身的第二状态值；

所述第二判断模块，判断所述第二状态值是否与第二预设状态值匹配；

所述第二采集模块，当判断所述第二状态值与所述第二预设状态值匹配时，采集所述日志信息；

所述第二控制模块，当判断所述第二状态值与所述第二预设状态值不匹配时，下发采集开启指令至下一个第二采集器。

5.根据权利要求4所述的数据处理系统，其特征在于，所述第二数据分发群组包括：一个主控服务器和若干个分发服务器；所述第二数据采集群组分别与所述分发服务器连接，所述分发服务器均与所述主控服务器连接，所述主控服务器与所述数据储存服务器连接；

所述分发服务器，获取所述第二数据采集群组下发的所述日志信息，并将所有日志信

息根据所述特征值分别上传至所述主控服务器；

所述主控服务器，发送所述日志信息至所述数据储存服务器。

6.根据权利要求5所述的数据处理系统，其特征在于，所述分发服务器包括：收集模块、缓存模块、发送模块；所述收集模块与所述收集模块连接，所述缓存模块分别与所述收集模块和所述调用模块连接；

所述收集模块，获取所述第二数据采集群组采集的所述日志信息；

所述缓存模块，储存所述日志信息；

所述调用模块，根据所述特征值，调用并发送所述日志信息至所述数据储存服务器。

7.根据权利要求6所述的数据处理系统，其特征在于，所述分发服务器还包括：监测模块和调整模块；所述监测模块分别与所述收集模块和所述调用模块连接，所述调整模块与所述监测模块连接；

监测模块，每间隔预设时长获取自身获取所述日志信息并上传所述日志信息的负荷值；

调整模块，根据所述负荷值，动态调整工作状态。

8.根据权利要求1所述的数据处理系统，其特征在于，还包括：存储服务器，所述存储服务器分别与所述第一数据分发群组和所述数据处理群组连接；

所述存储服务器，获取所述第一数据分发群组分发的所述上传文件；

所述数据处理群组，从所述存储服务器实时获取并分析所述上传文件得到日志信息。

9.根据权利要求1-7任一项所述的数据处理系统，其特征在于，所述数据储存服务器包括：

分析模块，分析所述日志信息的特征值；

储存模块，根据所述特征值，分类储存所述日志信息至对应的储存分区。

10.一种数据处理方法，其特征在于，应用于权利要求1-9任一项所述的数据处理系统，所述数据处理方法包括步骤：

S100第一数据采集群组采集所述数据源群组上传的上传文件；

S200第一数据分发群组将所述上传文件分发至所述数据处理群组；

S300数据处理群组分析所述上传文件得到日志信息；所述日志信息包括特征值；

S400数据储存服务器根据所述特征值，分类储存所述日志信息至对应的储存分区。

11.根据权利要求10所述的数据处理方法，其特征在于，所述步骤S300之后，S400之前包括步骤：

S310第二数据采集群组从所述数据处理群组获取所述日志信息，并将所述日志信息下发至第二数据分发群组；

S320所述第二数据分发群组，将所述日志信息分发至所述数据储存服务器。

一种数据处理系统及方法

技术领域

[0001] 本发明涉及数据处理领域，尤指一种数据处理系统及方法。

背景技术

[0002] 随着信息化的发展，关于上传文件呈海量增长，而且这些上传文件往往需要很长的保存期，而随着时间的增长和数据的增加，对数据存储空间的需求会越来越大，传统的关系数据库恐怕难以满足存储需求，Hadoop分布式技术的发展正好可以解决以上问题。

[0003] Hadoop(hdfs)是Apache开源组织的一个分布式计算框架，可以在大量廉价的硬件设备组成的集群上运行应用程序，构建一个高可靠性和良好扩展性的并行分布式系统。HDFS、MapReduce编程模型和Hbase分布式数据库是其三大核心技术。其中，HBase-HadoopDatabase，是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，利用HBase技术可在廉价PCServer上搭建起大规模结构化存储集群。

[0004] Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供完整的SQL查询功能，可以将SQL语句转换为MapReduce任务进行运行。

[0005] Flume是Cloudera提供的一个高可用的，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。

[0006] 由于hadoop分布式技术的强大优势，越来越多的使用hadoop分布式进行实时储存海量数据，但是在大数据业务中，数据的时效性和精确性是两个重要指标，其中时效性是大数据架构中，上传文件处理策略中一直追求的，尽管现在已经存在许多实时数据处理架构，但是数据处理和数据入库储存的时效性仍然不够。

发明内容

[0007] 本发明的目的是提供一种数据处理系统及方法，实现提升数据处理效率和可靠性。

[0008] 本发明提供的技术方案如下：

[0009] 本发明提供一种数据处理系统，包括：数据源群组，第一数据采集群组，第一数据分发群组，数据处理群组和数据储存服务器；所述第一数据采集群组分别与所述数据源群组和所述第一数据分发群组连接，所述数据处理群组分别与所述第一数据分发群组和所述数据储存服务器连接；所述第一数据采集群组，采集所述数据源群组上传的上传文件，并上传所述上传文件至所述第一数据分发群组；所述第一数据分发群组，分发所述上传文件至所述数据处理群组；所述数据处理群组，实时分析所述上传文件得到日志信息；所述日志信息包括特征值；所述数据储存服务器，根据所述特征值，分类储存所述日志信息至对应的储存分区。

[0010] 进一步的，所述第一数据采集群组包括：若干个第一采集器；所述第一采集器包括：第一获取模块、第一判断模块、第一采集模块和第一控制模块；所述第一获取模块与所

述第一判断模块连接，所述第一判断模块分别与所述第一采集模块和所述第一控制模块连接；所述第一获取模块，获取当前时刻自身的第一状态值；所述第一判断模块，判断所述第一状态值是否与第一预设状态值匹配；所述第一采集模块，当判断所述第一状态值与所述第一预设状态值匹配时，采集所述上传文件；所述第一控制模块，当判断所述第一状态值与所述第一预设状态值不匹配时，标记工作状态为异常，并下发采集开启指令至第一目标采集器。

[0011] 进一步的，还包括：第二数据采集群组，第二数据分发群组；所述第二数据采集群组与所述数据处理群组和所述第二数据分发群组连接，所述第二数据分发群组与所述数据储存服务器连接；所述第二数据采集群组，从所述数据处理群组获取所述日志信息，并将所述日志信息下发至所述第二数据分发群组；所述第二数据分发群组，将所述日志信息分发至所述数据储存服务器。

[0012] 进一步的，所述第二数据采集群组包括：若干个第二采集器；所述第二采集器包括：第二获取模块、第二判断模块、第二采集模块和第二控制模块；所述第二获取模块与所述第二判断模块连接，所述第二判断模块分别与所述第二采集模块和所述第二控制模块连接；所述第二获取模块，获取当前时刻自身的第二状态值；所述第二判断模块，判断所述第二状态值是否与第二预设状态值匹配；所述第二采集模块，当判断所述第二状态值与所述第二预设状态值匹配时，采集所述日志信息；所述第二控制模块，当判断所述第二状态值与所述第二预设状态值不匹配时，下发采集开启指令至下一个第二采集器。

[0013] 进一步的，所述第二数据分发群组包括：一个主控服务器和若干个分发服务器；所述第二数据采集群组分别与所述分发服务器连接，所述分发服务器均与所述主控服务器连接，所述主控服务器与所述数据储存服务器连接；所述分发服务器，获取所述第二数据采集群组下发的所述日志信息，并将所有日志信息根据所述特征值分别上传至所述主控服务器；所述主控服务器，发送所述日志信息至所述数据储存服务器。

[0014] 进一步的，所述分发服务器包括：收集模块、缓存模块、发送模块；所述收集模块与所述收集模块连接，所述缓存模块分别与所述收集模块和所述调用模块连接；所述收集模块，获取所述第二数据采集群组采集的所述日志信息；所述缓存模块，储存所述日志信息；所述调用模块，根据所述特征值，调用并发送所述日志信息至所述数据储存服务器。

[0015] 进一步的，所述分发服务器还包括：监测模块和调整模块；所述监测模块分别与所述收集模块和所述调用模块连接，所述调整模块与所述监测模块连接；监测模块，每间隔预设时长获取自身获取所述日志信息并上传所述日志信息的负荷值；调整模块，根据所述负荷值，动态调整工作状态。

[0016] 进一步的，还包括：存储服务器，所述存储服务器分别与所述第一数据分发群组和所述数据处理群组连接；所述存储服务器，获取所述第一数据分发群组分发的所述上传文件；所述数据处理群组，从所述存储服务器实时获取并分析所述上传文件得到日志信息。

[0017] 进一步的，所述数据储存服务器包括：分析模块，分析所述日志信息的特征值；储存模块，根据所述特征值，分类储存所述日志信息至对应的储存分区。

[0018] 本发明还提供一种数据处理方法，包括步骤：S100第一数据采集群组采集所述数据源群组上传的上传文件；S200第一数据分发群组将所述上传文件分发至所述数据处理群组；S300数据处理群组分析所述上传文件得到日志信息；所述日志信息包括特征值；S400数

据储存服务器根据所述特征值，分类储存所述日志信息至对应的储存分区。

[0019] 进一步的，所述步骤S300之后，S400之前包括步骤：

[0020] S310第二数据采集群组从所述数据处理群组获取所述日志信息，并将所述日志信息下发至第二数据分发群组；

[0021] S320所述第二数据分发群组，将所述日志信息分发至所述数据储存服务器。

[0022] 通过本发明提供的一种数据处理系统及方法，能够带来以下至少一种有益效果：

[0023] 1)本发明通过第一数据采集群组和第一数据分发群组边采集获取到上传文件，边由数据处理服务器进行实时处理，即实时采集和实时处理，能够提升数据处理的效率。

[0024] 2)本发明能够避免数据源群组直接将上传文件上传至数据处理群组进行分析处理，减少数据处理并储存入库的故障概率，提升数据处理高效性。

[0025] 3)本发明通过检测第一采集器或者第二采集器的工作状态，根据工作状态判断自身是否工作异常，从而及时的切换替换故障第一采集器或第二采集器，避免故障而出现数据处理的停滞，提升数据处理的可靠性。

附图说明

[0026] 下面将以明确易懂的方式，结合附图说明优选实施方式，对一种数据处理系统及方法的上述特性、技术特征、优点及其实现方式予以进一步说明。

[0027] 图1是本发明一种数据处理系统的一个实施例的结构示意图；

[0028] 图2是本发明一种数据处理系统的另一个实施例的结构示意图；

[0029] 图3是本发明一种数据处理系统的另一个实施例的结构示意图；

[0030] 图4是本发明一种数据处理系统的另一个实施例的结构示意图；

[0031] 图5是本发明一种数据处理系统的另一个实施例的流程图；

[0032] 图6是本发明一种数据处理方法的一个实施例的流程图。

具体实施方式

[0033] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对照附图说明本发明的具体实施方式。显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，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，并获得其他的实施方式。

[0034] 为使图面简洁，各图中只示意性地表示出了与本发明相关的部分，它们并不代表其作为产品的实际结构。另外，以使图面简洁便于理解，在有些图中具有相同结构或功能的部件，仅示意性地绘示了其中的一个，或仅标出了其中的一个。在本文中，“一个”不仅表示“仅此一个”，也可以表示“多于一个”的情形。

[0035] 本发明一种数据处理系统的一个实施例，如图1所示，包括：数据源群组 100，第一数据采集群组200，第一数据分发群组300，数据处理群组500和数据储存服务器400；所述第一数据采集群组200分别与所述数据源群组100和所述第一数据分发群组300连接，所述数据处理群组500分别与所述第一数据分发群组300和所述数据储存服务器400连接；

[0036] 所述第一数据采集群组200，采集所述数据源群组100上传的上传文件，并上传所述上传文件至所述第一数据分发群组300；

[0037] 所述第一数据分发群组300，分发所述上传文件至所述数据处理群组500；

[0038] 所述数据处理群组500，实时分析所述上传文件得到日志信息；所述日志信息包括特征值；

[0039] 所述数据储存服务器400，根据所述特征值，分类储存所述日志信息至对应的储存分区。

[0040] 具体的，本实施例中，数据源群组100包括若干个手机、电脑等移动终端，第一数据分发群组300包括若干个第一Flume服务器，数据处理群组500包括若干个数据处理服务器500；数据源群组100根据用户的访问请求访问目标网址对应的服务器，生成对应的上传文件，然后数据源群组100将生成的上传文件发送至第一数据采集群组200，第一数据采集群组200收集数据源群组100上传的上传文件，并且根据第一Flume服务器的负载性能，将收集到的上传文件发送至与数据采集群组对应连接的数据第一Flume服务器中，由每个数据第一 Flume服务器将各自接收的上传文件发送至各自连接的数据处理服务器500，每个数据处理服务器500实时将各自接收的上传文件进行分析处理(如解压、解密和清除过滤)得到对应的日志信息，即一旦获取到上传文件后就立刻进行分析处理得到对应的日志信息，将各自处理的日志信息发送至数据储存服务器400，即数据储存服务器400根据日志信息的特征值进行分类储存于对应的储存分区。不同于现有技术中当天处理前一天的上传文件，本发明是由第一数据采集群组200和第一数据分发群组300边采集获取到上传文件后，就边由数据处理服务器500进行处理，即数据处理服务器500实时处理采集到的上传文件，而且，由于将数据源群组100的上传文件上传至第一数据采集群组 200，再由第一数据分发群组300分发第一数据采集群组200转发的上传文件，能够避免向现有技术那样，由数据源群组100直接将上传文件上传至数据处理群组500进行分析处理，由于数据处理群组500的处理能力不够而导致数据处理群组500出现处理缓慢甚至“死机”的现象，本发明的数据处理效果不断接近实时性处理，数据处理的效率提升，能够减少数据处理并储存入库的故障，提升数据处理可靠性和高效性。

[0041] 本发明一种数据处理系统的另一个实施例，是上述实施例的优化实施例，如图1和图3所示，本实施例与上述实施例相比，主要改进在于，所述第一数据采集群组200包括：若干个第一采集器210；所述第一采集器210包括：第一获取模块211、第一判断模块212、第一采集模块213和第一控制模块214；所述第一获取模块211与所述第一判断模块212连接，所述第一判断模块212分别与所述第一采集模块213和所述第一控制模块214连接；

[0042] 所述第一获取模块211，获取当前时刻自身的第一状态值；

[0043] 所述第一判断模块212，判断所述第一状态值是否与第一预设状态值匹配；

[0044] 所述第一采集模块213，当判断所述第一状态值与所述第一预设状态值匹配时，采集所述上传文件；

[0045] 所述第一控制模块214，当判断所述第一状态值与所述第一预设状态值不匹配时，标记工作状态为异常，并下发采集开启指令至第一目标采集器。

[0046] 具体的，本实施例中，第一目标采集器是第一状态值与第一预设状态值匹配的，并且优先级最高的第一采集器210，第一采集器210可以是第一Nginx 服务器，每个第一采集器210在进行采集数据源群组100上传的上传文件前，需要获取当前时刻自身的第一状态值，判断第一状态值是否与第一预设状态值匹配，当判断第一状态值与第一预设状态值匹

配时，采集数据源群组100上传的上传文件；当判断第一状态值与第一预设状态值不匹配时，下发采集开启指令至第一目标采集器。示例性的，如有三个第一采集器210，分别为第一采集器210A1，第一采集器210B1和第一采集器210C1，第一采集器210A1的第一状态值为active和fault，表明第一采集器210A1正在被使用，并且出现故障；第一采集器210B1的第一状态值为back，并且第一采集器210B1的当前负载为 S1，第一采集器210C1的第一状态值为back，并且第一采集器210C1的当前负载为S2，如果S1大于S2，而负载越大数据处理能力越弱，即第一采集器210B1 的优先级低于第一采集器210C1，所以选择第一采集器210C1作为第一目标采集器，第一采集器210A1标记自身的工作状态为异常，并且发送开启指令至第一采集器210C1，第一采集器210C1接收到开启指令后就转换自身的状态为 active，并且开始获取采集数据源群组100上传的上传文件。本发明能够在第一采集器210出现工作异常时，及时的切换替换故障第一采集器210，避免因为第一采集器210的故障而出现数据处理的停滞，减少了上传文件收集的丢失，提升数据处理的可靠性和高效性。

[0047] 本发明一种数据处理系统的另一个实施例，是上述实施例的优化实施例，如图2所示，本实施例与上述实施例相比，主要改进在于，还包括：第二数据采集群组600，第二数据分发群组700；所述第二数据采集群组600与所述数据处理群组500和所述第二数据分发群组700连接，所述第二数据分发群组700与所述数据储存服务器400连接；

[0048] 所述第二数据采集群组600，从所述数据处理群组500获取所述日志信息，并将所述日志信息下发至所述第二数据分发群组700；

[0049] 所述第二数据分发群组700，将所述日志信息分发至所述数据储存服务器 400。

[0050] 具体的，本实施例中，第二数据分发群组700包括若干个第二Flume服务器，第二数据采集群组600将数据处理群组500分析处理后的日志信息进行采集归总，再由第二数据分发群组700中的各个第二Flume服务器将各自获取到的日志信息发送至数据储存服务器400，然后数据储存服务器400根据特征值，分类储存日志信息至对应的储存分区。

[0051] 本发明一种数据处理系统的另一个实施例，是上述实施例的优化实施例，如图2和图4所示，本实施例与上述实施例相比，主要改进在于，所述第二数据采集群组600包括：若干个第二采集器610；所述第二采集器610包括：第二获取模块611、第二判断模块612、第二采集模块613和第二控制模块614；所述第二获取模块611与所述第二判断模块612连接，所述第二判断模块612分别与所述第二采集模块613和所述第二控制模块614连接；

[0052] 所述第二获取模块611，获取当前时刻自身的第二状态值；

[0053] 所述第二判断模块612，判断所述第二状态值是否与第二预设状态值匹配；

[0054] 所述第二采集模块613，当判断所述第二状态值与所述第二预设状态值匹配时，采集所述日志信息；

[0055] 所述第二控制模块614，当判断所述第二状态值与所述第二预设状态值不匹配时，下发采集开启指令至下一个第二采集器610。

[0056] 具体的，本实施例中，第二目标采集器是第二状态值与第二预设状态值匹配的，并且优先级最高的第二采集器610，第二采集器610可以是第二Nginx 服务器，每个第二采集器610在进行采集数据处理群组500下发的日志信息前，需要获取当前时刻自身的第二状态值，判断第二状态值是否与第二预设状态值匹配，当判断第二状态值与第二预设状态值匹配时，采集数据处理群组500 下发的日志信息；当判断第二状态值与第二预设状态值不匹

配时，下发采集开启指令至第二目标采集器。示例性的，如有三个第二采集器610，分别为第二采集器610A1，第二采集器610B1和第二采集器610C1，第二采集器610A1 的第二状态值为active和fault，表明第二采集器610A1正在被使用，并且出现故障；第二采集器610B1的第二状态值为back，并且第二采集器610B1的当前负载为S1，第二采集器610C1的第二状态值为back，并且第二采集器610C1的当前负载为S2，如果S1大于S2，而负载越大数据处理能力越弱，即第二采集器610B1的优先级低于第二采集器610C1，所以选择第二采集器610C1作为第二目标采集器，第二采集器610A1标记自身的工作状态为异常，并且发送开启指令至第二采集器610C1，第二采集器610C1接收到开启指令后就转换自身的状态为active，并且开始获取采集数据处理群组500下发的日志信息。本发明能够在第二采集器610出现工作异常时，及时的切换替换故障第二采集器610，避免因为第二采集器610的故障而出现数据处理的停滞，减少了数据处理群组 500下发的日志信息收集的丢失，提升数据处理的可靠性和高效性。

[0057] 本发明一种数据处理系统的另一个实施例，是上述实施例的优化实施例，本实施例与上述实施例相比，主要改进在于，还包括：所述第二数据分发群组700包括：一个主控服务器和若干个分发服务器；所述第二数据采集群组600 分别与所述分发服务器连接，所述分发服务器均与所述主控服务器连接，所述主控服务器与所述数据储存服务器400连接；

[0058] 所述分发服务器，获取所述第二数据采集群组600下发的所述日志信息，并将所有日志信息根据所述特征值分别上传至所述主控服务器；

[0059] 所述主控服务器，发送所述日志信息至所述数据储存服务器400。

[0060] 具体的，本实施例中，主控服务器和分发服务器可以均为Flume服务器，所有的分发服务器均与主控服务器连接，而主控服务器与数据储存服务器400 连接，分发服务器接收第二数据采集群组600发送的日志信息，然后每个分发服务器根据特征值将各自获取的日志信息发送至主控服务器，由主控服务器根据特征值将日志信息分别发送至数据储存服务器400。Flume服务器是一个高可用的，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的能力。本发明由若干个分发服务器进行获取，能够分摊每个分发服务器的负载压力，如特征值为上传文件中的时间戳时，那么每个分发服务器就根据时间戳进行分门别类，从而根据时间戳的先后顺序进行发送对应的日志信息至主控服务器，由主控服务器进行管理发送日志信息至数据服务器，能够减少数据压缩上传的时间，从而提升数据处理效率。

[0061] 本发明一种数据处理系统的另一个实施例，是上述实施例的优化实施例，本实施例与上述实施例相比，主要改进在于，所述分发服务器包括：收集模块、缓存模块、发送模块；所述收集模块与所述收集模块连接，所述缓存模块分别与所述收集模块和所述调用模块连接；

[0062] 所述收集模块，获取所述第二数据采集群组600采集的所述日志信息；

[0063] 所述缓存模块，储存所述日志信息；

[0064] 所述调用模块，根据所述特征值，调用并发送所述日志信息至所述数据储存服务器400。

[0065] 具体的，本实施例中，Flume服务器的核心是收集数据，然后将收集到的数据由调

用模块发送到指定的数据储存服务器400。为了保证输送的过程一定成功，在送到数据储存服务器400之前，会先缓存数据，待数据真正到达数据储存服务器400后，Flume服务器再删除自己缓存的数据。在整个数据的传输的过程中，流动的是日志信息，如果日志信息是文本文件，通常是一行记录。收集模块是专门用来收集数据的，可以处理各种类型、各种格式的日志数据，收集模块把数据收集来以后，临时存放在缓存模块中，即缓存模块是专门用来存放临时数据的——对采集到的数据进行简单的缓存，调用模块把数据发送到数据储存服务器400，数据储存服务器400包括hdfs、logger、avro、 thrift、ipc、file、null、hbase、solr等等。

[0066] 优选的，所述分发服务器还包括：监测模块和调整模块；所述监测模块分别与所述收集模块和所述调用模块连接，所述调整模块与所述监测模块连接；

[0067] 监测模块，每间隔预设时长获取自身获取所述日志信息并上传所述日志信息的负荷值；

[0068] 调整模块，根据所述负荷值，动态调整工作状态。

[0069] 具体的，本实施例中，由于Flume服务器的储存空间一般较小，因此很容易出现日志信息占满Flume服务器的储存空间，导致Flume服务器不能正常工作的现象，因此通过监测模块进行监测收集模块获取日志信息的获取速率和调用模块调用发送日志信息的发送速率，每个Flume服务器根据获取速率和发送速率进行计算得到对应的当前时刻各自对应的负荷值，从而根据负荷值进行动态调整各自对应的工作状态(包括获取速率和/或发送速率)，这样就能够避免Flume服务器由于储存空间占满而影响Flume服务器自身的数据分发处理功能，提升数据的处理效率。

[0070] 本发明一种数据处理系统的另一个实施例，是上述实施例的优化实施例，本实施例与上述实施例相比，主要改进在于，还包括：存储服务器，所述存储服务器分别与所述第一数据分发群组300和所述数据处理群组500连接；

[0071] 所述存储服务器，获取所述第一数据分发群组300分发的所述上传文件；

[0072] 所述数据处理群组500，从所述存储服务器实时获取并分析所述上传文件得到日志信息。

[0073] 具体的，本实施例中，第一数据分发群组300将各自获取到的上传文件发送至存储服务器，由存储服务器统一进行储存，数据处理群组500从存储服务器的储存区域进行读取相应的上传文件，从而将读取到的上传文件进行分析处理得到对应的日志信息。本发明能够避免第一数据分发群组300中的第一Flume服务器将各自接收的上行文件无差别、无针对性的发送至每个数据处理服务器500，能够使得数据处理服务器500直接从储存服务器获取需要调用处理的上传文件，而不是由每个数据处理服务器500获取所有的上传文件进行筛选后，才能进行处理的情况，因此，能够针对性的根据特征值获取相应的上传文件进行分析处理，从而提升数据处理效率。

[0074] 本发明一种数据处理系统的另一个实施例，是上述实施例的优化实施例，本实施例与上述实施例相比，主要改进在于，所述数据储存服务器400包括：

[0075] 分析模块，分析所述日志信息的特征值；

[0076] 储存模块，根据所述特征值，分类储存所述日志信息至对应的储存分区。

[0077] 具体的，本实施例中，特征值包括数据源标识和获取上传文件的时间戳等等，根据

特征值将日志信息分类储存至对应的储存分区，能够便于用户后续调用查询时，直接根据特征值进行查询，提升查询效率。

[0078] 基于上述所有实施例，举例说明一个实例，如图5所示：配置N台第二 Flume服务器Ni8(1<＝i<＝N)和另外一台第二Flume服务器P9，通过第一 Nginx服务器3即第一数据采集群组200接收数据源群组100发送过来的上行文件(.data文件)并发送至第一Flume服务器4，然后第一Flume服务器4 将上行文件发送至数据处理群组500，数据处理群组500对上行文件进行处理得到日志信息(.txt文件)，然后第二Nginx服务器7实时分发日志信息到N 台第二Flume服务器Ni8，N台第二Flume服务器Ni8接收到日志信息并发送到第二Flume服务器P9，通过第二Flume服务器P9将日志信息导入到hdfs 并压缩追加存储到Hive库对应的区表中，即储存至对应的储存分区内。即

[0079] 步骤一、两台第一Nginx服务器3，状态处于active的第一Nginx服务器 3接收移动终端1对应的终端服务器2并行发送过来的上行文件，并通过第一 Flume服务器4分发这些上行文件，存储在特定的存储服务器5；(两台第二 Nginx服务器7是互相备份机制，即一台状态是active，另一台是back，一旦 active状态的第二Nginx服务器7出现故障，则自动启动back状态的第二Nginx 服务器7，back状态的第二Nginx服务器7状态转换为active)

[0080] 步骤二、设定时钟定时器，多台数据处理服务器500每隔一小时从存储服务器5获取并处理上一小时的上行文件；

[0081] 步骤三、另配置两台第二Nginx服务器7和N+1台Flume服务器对处理后的终端用户行为信息(.txt格式)进行上传hdfs和入库hive，具体步骤如下：

[0082] 1、数据处理服务器500在处理完每个上行文件之后，则将日志信息发送给对应的第二Nginx服务器7。

[0083] 2、第二Nginx服务器7接收到日志信息之后，分发日志信息到N台Flume 服务器(两台第二Nginx服务器7是互相备份机制，即一台状态是active，另一台是back，一旦active状态的第二Nginx服务器7出现故障，则自动启动 back状态的第二Nginx服务器7，back状态的第二Nginx服务器7状态转换为active)。

[0084] 3、每台第二Flume服务器Ni8的source组件接收到第二Nginx服务器7 分发的日志信息，将日志信息暂存在channel组件，sink组件不断消耗掉滞留在channel组件处的日志信息，将日志信息以行为单位进行记录，进行发送到 Flume服务器P。

[0085] 4、第二Flume服务器P9的sink组件连接Hdfs服务器10，且第二Flume 服务器P9的source组件接收到各个第二Flume服务器Ni8(1<＝i<＝N)sink 端发送过来的日志信息，在channel组件进行暂存，由sink组件进行整理消耗处理，传送到Hdfs服务器10中，Hdfs服务器10将第二Flume服务器P9上传的日志信息，根据日期等等特征值，压缩存储到hive的对应分区表内。

[0086] 本发明不仅能够实时的将数据上传到hdfs中，同时根据数据的时间戳，将数据压缩存储到hive对应的区表中，从而实现海量小文件实时快速上传hdfs 和入库hive的功能，提升数据处理效率，而且节省了大量的人力资源。

[0087] 本发明一种数据处理方法的一个实施例，如图6所示，包括：

[0088] S100第一数据采集群组采集所述数据源群组上传的上传文件；

[0089] S200第一数据分发群组将所述上传文件分发至所述数据处理群组；

[0090] S300数据处理群组分析所述上传文件得到日志信息;所述日志信息包括特征值;

[0091] S400数据储存服务器根据所述特征值,分类储存所述日志信息至对应的储存分区。

[0092] 具体的,本实施例是上述系统实施例对应的方法实施例,具体效果参见上述系统实施例,在此不再一一赘述。

[0093] 本发明一种数据处理方法的另一个实施例,包括:

[0094] S100第一数据采集群组采集所述数据源群组上传的上传文件;

[0095] S200第一数据分发群组将所述上传文件分发至数据处理群组;

[0096] S300数据处理群组分析所述上传文件得到日志信息;所述日志信息包括特征值;

[0097] S310第二数据采集群组从所述数据处理群组获取所述日志信息,并将所述日志信息下发至第二数据分发群组;

[0098] S320所述第二数据分发群组,将所述日志信息分发至所述数据储存服务器;

[0099] S400数据储存服务器根据所述特征值,分类储存所述日志信息至对应的储存分区。

[0100] 具体的,本实施例是上述系统实施例对应的方法实施例,具体效果参见上述系统实施例,在此不再一一赘述。

[0101] 应当说明的是,上述实施例均可根据需要自由组合。以上所述仅是本发明的优选实施方式,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。

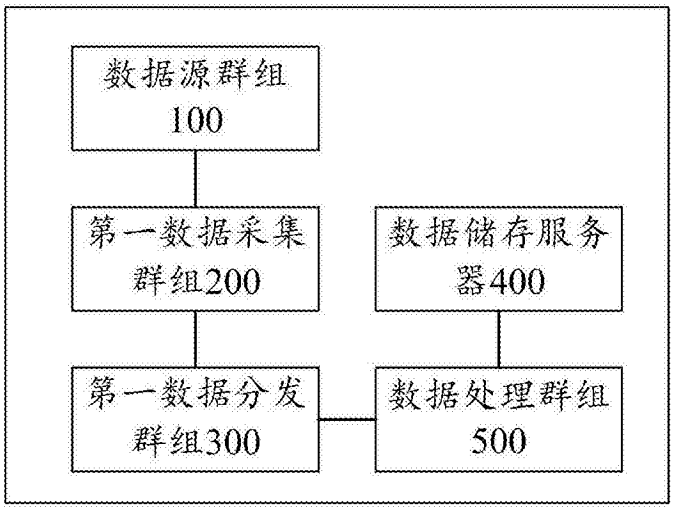

数据源群组
100
第一数据采集
群组200
数据储存服务
器400
第一数据分发
群组300
数据处理群组
500

图1

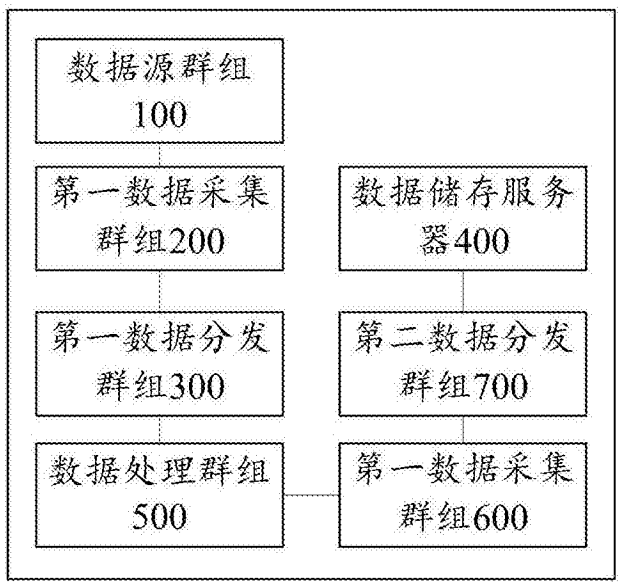

数据源群组
100
第一数据采集
群组200
数据储存服务
器400
第一数据分发
群组300
第二数据分发
群组700
数据处理群组
500
第一数据采集
群组600

图2

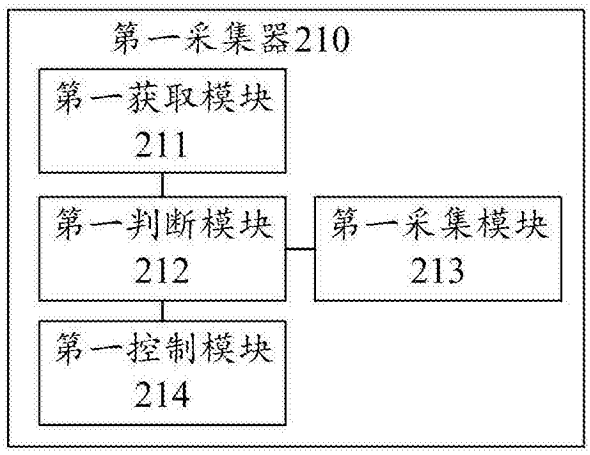

第一采集器210
第一获取模块
211
第一判断模块
212
第一采集模块
213
第一控制模块
214

图3

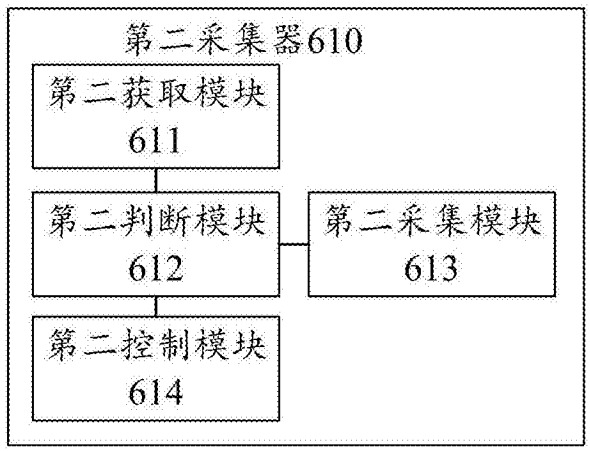

第二采集器610
第二获取模块
611
第二判断模块
612
第二采集模块
613
第二控制模块
614

图4

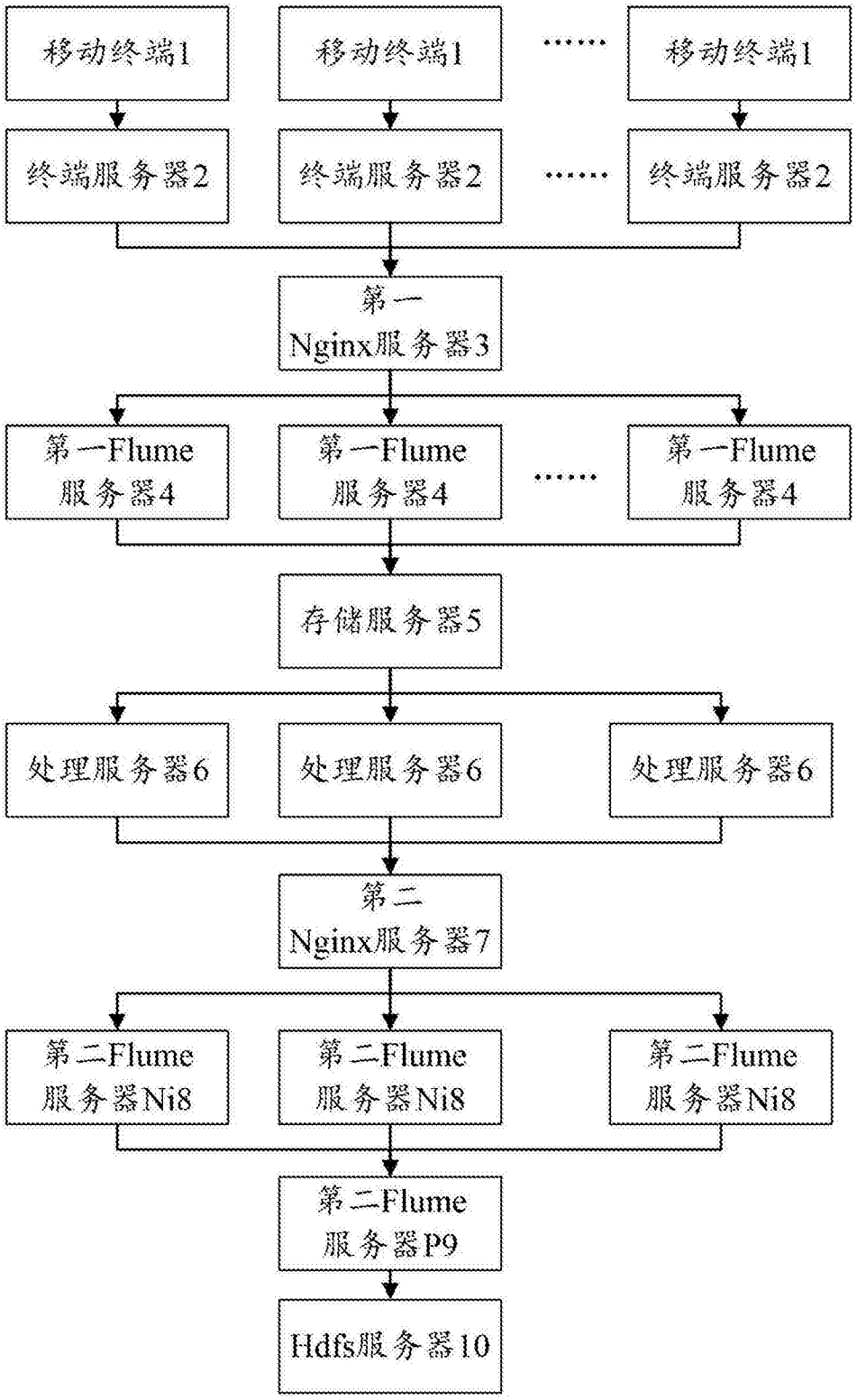

移动终端1
移动终端1
……
移动终端1
终端服务器2
终端服务器2
……
终端服务器2
第一
Nginx服务器3
第一Flume
服务器4
第一Flume
服务器4
……
第一Flume
服务器4
存储服务器5
处理服务器6
处理服务器6
处理服务器6
第二
Nginx服务器7
第二Flume
服务器Ni8
第二Flume
服务器Ni8
第二Flume
服务器Ni8
第二Flume
服务器P9
Hdfs服务器10

图5

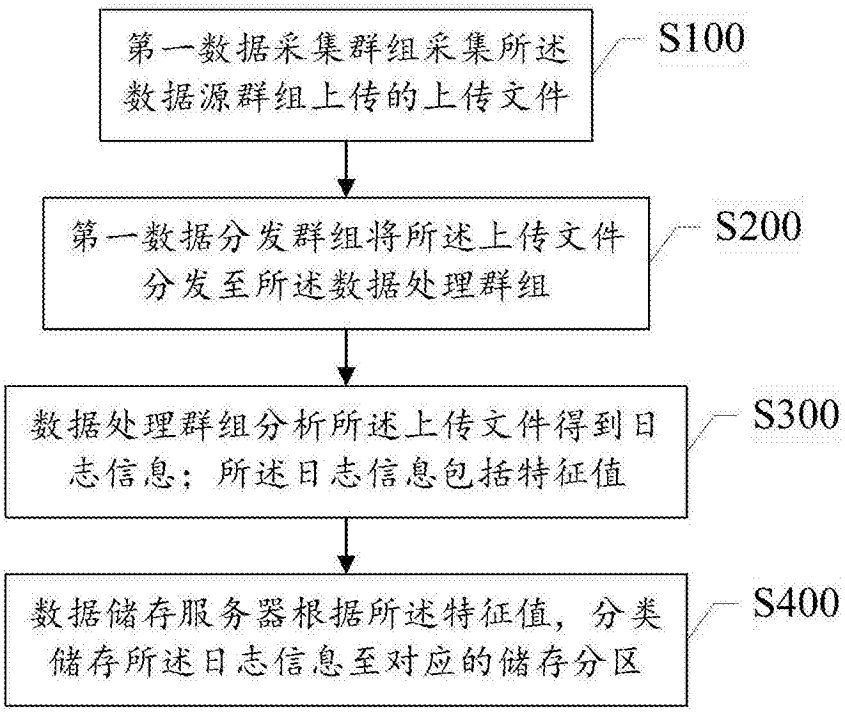

第一数据采集群组采集所述数据源群组上传的上传文件
S100
第一数据分发群组将所述上传文件分发至所述数据处理群组
S200
数据处理群组分析所述上传文件得到日志信息；所述日志信息包括特征值
S300
数据储存服务器根据所述特征值，分类储存所述日志信息至对应的储存分区
S400

图6