

(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 104243564 A

(43) 申请公布日 2014. 12. 24

(21) 申请号 201410444864. 2

(22) 申请日 2014. 09. 03

(71) 申请人 河海大学

地址 211100 江苏省南京市江宁区佛城西路  
8号

(72) 发明人 陆佳民 冯钧

(74) 专利代理机构 南京苏高专利商标事务所

(普通合伙) 32204

代理人 李玉平

(51) Int. Cl.

H04L 29/08 (2006. 01)

G06F 17/30 (2006. 01)

权利要求书1页 说明书3页 附图1页

(54) 发明名称

一种面向 AWS 平台的分布式数据管理方法

(57) 摘要

本发明公开一种面向 AWS 平台的分布式数据管理方法，属于信息技术处理领域。本方法公开了面向 AWS (AmazonWebServices 亚马逊网络服务系统) 平台，进行快速且廉价的数据分布方法。本发明明确了基于 AWS 所提供的 EC2 (ElasticComputeCloud 弹性计算云) 和 EBS (ElasticBlockStorage 弹性块存储) 服务，采用集中式数据划分与分配、并行配对挂载、存储与处理服务分离等手段，实现了在大规模集群内进行批量快速的数据导入准备，同时在集群计算资源被回收后，实现低成本的数据存储与复用，从而将集群数据的准备时间与费用成本降到最低。



1. 一种面向 AWS 平台的分布式数据管理方法，其特征在于，利用集中式数据分配与批量式数据挂载的方式来进行大规模集群数据准备，包括如下步骤：

步骤 1，创建单台实例并挂载大量小容量 EBS 卷组，实现大数据的划分与分配；

步骤 2，创建全部集群实例，与得到数据分配后的 EBS 卷组进行并行匹配挂载；

步骤 3，待计算任务结束后，对存储服务与计算服务进行分离，卸载所有 EBS 卷组并终止集群实例，将并行数据结果单独存留在 EBS 卷组中。

2. 根据权利要求 1 所述的一种面向 AWS 平台的分布式数据管理方法，其特征在于，步骤 1 的数据分配，具体步骤为：

步骤 1-1，依据卷组标签扫描所需 EBS 卷组 ID，采用循环模式分配存储设备名称，在实例外部利用 EC2 命令行工具进行挂载后向实例提交所有设备名；

步骤 1-2，实例内部创建所有挂载文件路径并赋予相应权限，与接收到的存储设备进行系统级别挂载；

步骤 1-3，利用用户自定义方法对集群数据进行划分，并对应转移到相应的小容量存储设备中去。

3. 根据权利要求 1 所述的一种面向 AWS 平台的分布式数据管理方法，其特征在于步骤 2 的并行配对挂载，具体步骤为：

步骤 2-1，依据资源标签扫描可用的 EBS 卷组与集群实例 ID，利用循环模式进行配对挂载后，对每一实例上传分配的存储设备名；

步骤 2-2，并行访问所有实例，创建挂载文件路径并赋予读写权限，在系统级别挂载设备。

## 一种面向 AWS 平台的分布式数据管理方法

[0001]

### 技术领域

[0002] 本发明涉及一种面向 AWS 平台的分布式数据管理方法，属于信息技术处理领域。

### 背景技术

[0003] AWS(Amazon Web Services 亚马逊网络服务系统)是由亚马逊公司于2002年开始运营的云计算平台,提供多种类型的远端 Web 服务。其中的 EC2 (Elastic Compute Cloud 弹性计算云) 服务可以让用户以租用虚拟计算机(实例)方式,远程使用不同类型的计算机系统,并且运行任何自己所需要的软件。租用虚拟计算机的价格随提供的硬件资源多少而区分,并且用户可以随时创建、运行和终止服务,因此 AWS EC2 被视为一个灵活的、相对廉价的云计算平台。EBS (Elastic Block Storage 弹性块存储) 服务亦是由 AWS 提供的,用于支持持久性数据块级存储需求。EBS 卷组独立于 EC2 实例的生命周期,通过网络驱动伪装成块设备的方式使其能够成为任意实例的虚拟存储设备,并可在实例终止后继续存在,实现数据复用。AWS 同时提供了 S3 (Simple Storage Service) 服务来进行网络数据存储,但其更适用于静态的数据存储,S3 设备可以转换为 EBS 卷组来进行使用。

[0004] 由于受到 AWS 自身服务的限制,在利用申请大量 EC2 实例组建计算机集群时,实例内的 EBS 卷组资源在实例被暂停或终止后即被收回,相应地存储在 EBS 卷组内的数据也会同时丢失。

[0005] 针对这一问题,常见的解决方法是待集群建立完毕后,再利用主实例进行数据的导入与分布。在这一过程中,其它实例因为没有参与到实际的数据处理过程当中,处于被动等待状态,因此造成了大量计算资源的浪费。

[0006] 最后,当集群被终止后,产生的并行数据结果也需要再收集到额外的 EBS 或 S3 设备中去,同样造成了 AWS 资源的浪费。

[0007] 因此,这种 EC2 实例内部 EBS 的不稳定性,造成了在大规模实例集群是进行数据分布与保存的技术难点。

[0008] 目前,也没有公开发表的技术或手段来对此问题加以解决。

### 发明内容

[0009] 发明目的 :针对现有技术中存在的问题与不足,本发明提供一种面向 AWS 平台的分布式数据管理方法。首先采用集中式的数据划分与分配,利用一台 EC2 实例来完成所有的数据分布任务,降低了整体 AWS 资源的使用。其次通过并行配对挂载的方法,将大量实例资源与已分配数据的 EBS 卷组进行配对挂载,快速构建起可用的计算机集群。最后待处理任务结束后,统一卸载外部挂载的 EBS 卷组,达到了计算资源与存储设备的分离,在不消耗额外计算与存储资源的前提下实现了数据复用的目的。

[0010] 技术方案 :一种面向 AWS 平台的分布式数据管理方法,包括如下步骤 :

步骤 1, 创建单台实例并挂载大量小容量 EBS 卷组, 实现大数据的划分与分配, 具体方法为 :

步骤 1-1, 依据卷组标签扫描所需 EBS 卷组 ID, 采用循环模式分配存储设备名称, 在实例外部利用 EC2 命令行工具进行挂载后向实例提交所有设备名;

步骤 1-2, 实例内部创建所有挂载文件路径并赋予相应权限, 与接收到的存储设备进行系统级别挂载;

步骤 1-3, 利用用户自定义方法对集群数据进行划分, 并对应转移到相应的小容量存储设备中去;

步骤 2, 创建全部集群实例, 与得到数据分配后的 EBS 卷组进行并行匹配挂载;

步骤 2-1, 依据资源标签扫描可用的 EBS 卷组与集群实例 ID, 利用循环模式进行配对挂载后, 对每一实例上传分配的存储设备名;

步骤 2-2, 并行访问所有实例, 创建挂载文件路径并赋予读写权限, 在系统级别挂载设备;

步骤 3, 待计算任务结束后, 对存储服务与计算服务进行分离, 卸载所有 EBS 卷组并终止集群实例, 将并行数据结果单独存留在 EBS 卷组中。

[0011] 以上发明内容中所述的 EC2 命令行工具是指由 AWS EC2 服务提供的, 通过文本命令方式来运行服务的脚本; 所述 EBS 卷组是指由 AWS 提供的, 具有独立生命周期的网络存储设备, 可以与 EC2 实例通过网络驱动的方式的连接, 来提供持久性的数据块级存储。每个 EBS 卷组与 EC2 实例, 都具有惟一的 ID 编号, 并可通过添加标签的方式来进行语义上的区分。

[0012] 本发明采用上述技术方案, 具有以下有益效果:

1、方法在集群建立之前就可利用一台 EC2 实例来完成所有的数据分布任务, 显著降低了整体 AWS 的资源消耗;

2、待集群创建后, 方法采用并行配对挂载的方法, 可将所需数据快速分布到整个集群系统内, 提高了数据分布准备的效率;

3、存储有并行处理结果数据的 EBS 卷组, 可以在处理任务结束后, 继续持久保存, 以极低的成本达到了数据复用的目的。

## 附图说明

[0013] 图 1 为本发明实施例的面向 AWS 平台的分布式数据的划分与配置流程图。

## 具体实施方式

[0014] 下面结合具体实施例, 进一步阐明本发明, 应理解这些实施例仅用于说明本发明而不用于限制本发明的范围, 在阅读了本发明之后, 本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

[0015] 利用面向 AWS 平台的分布式数据管理方法进行并行处理系统的数据分配与并行挂载的最小对象为一台设置有 EC2 命令行工具的可联网的计算机设备和一个可以有效申请 AWS 各类资源的 Amazon 账号。

[0016] 在对数据进行分布管理之前, 需要进行预处理, 具体操作如下: 检测当前计算机系

统内是否准备有 EC2 命令行工具, 工具均为可执行的系统脚本, 以“ec2-”为命名前缀, 也可通过命令如“ec2-describe-regions”来进行实地验证, 此命令用于打印所有可用的 EC2 数据中心名称; 检测当前计算机内是否存在 AWS EC2 账户私有密钥, 如不存在, 则可通过访问 EC2 图形化网页界面来获取。

[0017] 通过上述预处理后, 即可按照图 1 所示的流程图, 实现面向 AWS 平台的分布式数据的划分与配置流程, 包括以下步骤:

步骤 1, 利用 EC2 命令行工具或 AWS 提供的网页控制终端, 启动一台 EC2 实例与若干小容量 EBS 卷组, 卷组的数量由用户自行定义并添加描述标签(如“data”), 用以实现大数据的划分与分配, 具体方法为:

步骤 1-1, 利用 EC2 命令行工具中的“ec2-describe-volumes”, 依据卷组标签扫描所需 EBS 卷组 ID, 采用循环模式分配存储设备名称, 在实例外部利用 EC2 命令行工具进行挂载后向实例提交所有设备名, 设备名称一般由“sd”开头, 加上 f-p 之间单个字母, 以及由 1-15 内的数字;

步骤 1-2, 实例内部创建所有挂载文件路径并赋予相应权限, 利用 mount 指令与接收到的存储设备进行系统级别挂载;

步骤 1-3, 导入大数据后, 利用用户自定义的划分方法对数据进行分割, 并将相应需要的文件一一拷贝入挂载后的小容量 EBS 卷组内;

步骤 1-4, 卸载所有已分配数据的 EBS 卷组;

步骤 2, 利用 EC2 命令行工具或 AWS 提供的网页控制终端来创建大量 EC2 实例, 来构建计算机集群, 并与步骤 2 中产生的 EBS 卷组进行并行匹配挂载, 具体方法为:

步骤 2-1, 依据资源标签扫描可用的 EBS 卷组与集群实例 ID, 利用 EC2 命令行工具中的“ec2-describe-instances”以及“ec2-describe-volumes”,

步骤 2-2, 利用 EC2 命令行工具“ec2-attach-volume”, 基于循环模式对实例与卷组进行配对挂载, 对每一实例上传分配的存储设备名, 每个终端在保证设备名称惟一性的前提下可以获得一个或多个卷组, 匹配后的设备名被上传至实例中的特定文件“/tmp/.matchedVolumes”;

步骤 2-3, 并行访问所有实例, 上传集群内加密传输的密钥, 得以从控制端在无需密码下的情况下远程访问所有相关实例;

步骤 2-4, 在每个实例中, 创建挂载文件路径并赋予读写权限, 读取上传的卷组匹配文件并利用 mount 指令在系统级别对相关设备进行挂载。

[0018] 步骤 3, 待计算任务结束后, 对存储服务与计算服务进行分离, 卸载所有 EBS 卷组并终止集群实例, 将并行数据结果单独存留在 EBS 卷组中。

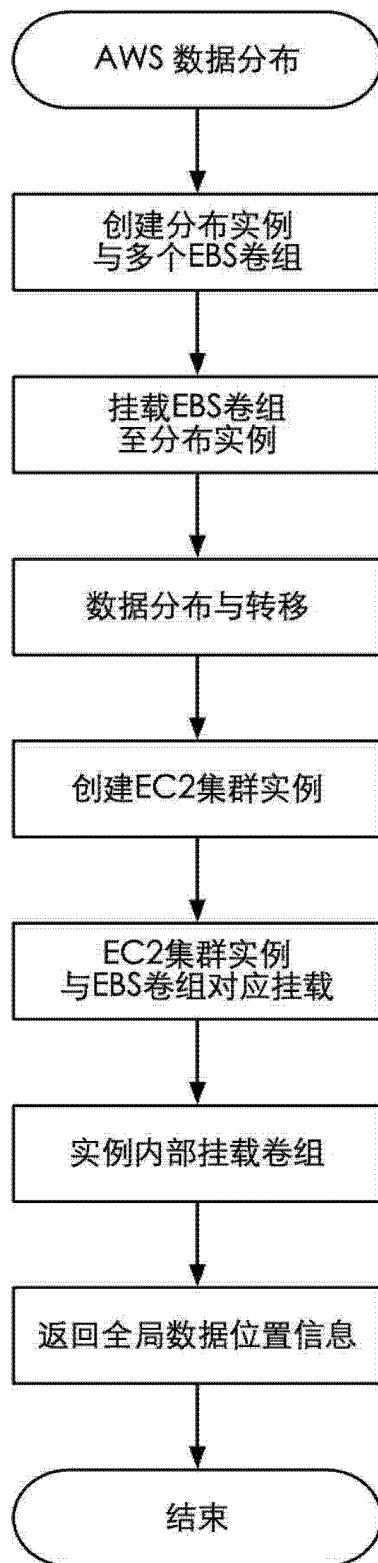


图 1