云计算环境下基于虚拟数据代理的云模型数据布局策略

**摘要** 当前，大量数据密集型应用开始部署到各类云计算平台中，面对多数据中心的云计算环境，这类应用在数据布局方面遇到了新的挑战：数据布局的过程中存在着不确定性和随机性，如何恰当的描述这一过程，并能缩减跨数据中心的数据传输时间开销和数据移动次数成为亟待解决的关键问题。对此，本文提出一种“基于云模型的数据布局策略”：通过建立云模型，设计一种面向数据密集型应用的数据布局策略。该策略解决了数据布局过程中的不确定性和随机性问题，并考虑到数据集间的相互依赖关系与数据中心的存储大小，减少数据跨数据中心传输的时间开销和移动次数。

**关键词**：云计算；数据密集；云模型；数据布局

1. 引言

数据密集型应用[1]是依赖于海量数据的大数据应用，这些应用形成了日益重要的高性能计算工作负载段，应用程序通常通过科学工作流来处理大型外部数据集，并且通常需要超过主内存容量的非常大的工作集[2,3]。例如，生物信息学、天文学、核能等科学型工程应用，以及社交网络服务、移动视频监控、金融数据处理等信息型服务应用等等，它们都通过对海量数据进行复杂计算、分析挖掘与处理从而创造丰富的数据智慧，服务于社会生活生产以及科学研究领域。云计算是数据共享与服务共享计算模式的结合体[4]，为数据密集型应用提供了一种全新的部署和执行方式。当前，许多地理上分布的私有云可以对外提供一部分计算和存储资源，可将此类私有云看作单独数据中心并通过互联网技术形成更大的云计算平台。该云计算平台允许用户执行其应用，并将所需数据集上传到平台中。由于数据集规模可能非常庞大、部分数据集只能存放在指定的数据中心并且云计算平台的地理分布节点之间存在带宽限制，不可能将所有数据集上传到单一数据中心或为每个数据中心上传所有数据集，需要将不同数据集分别上传到不同数据中心使用户的数据密集型应用的多个子任务并行执行。由于任务之间存在较强的数据依赖关系，数据密集型应用在运行时也必然会产生跨多个数据中心的数据移动任务，涉及大量数据传输、数据同步等数据开销，不合理的数据放置将产生高额成本并严重影响工作流执行效率。因此，如何在云计算环境下面向数据密集型应用设计合理的数据布局策略成为亟待解决的问题。

本文将定性概念和定量论域之间的不确定性转换工具云模型应用到数据布局策略中以解决数据布局过程中的不确定性和随机性问题。在数据布局过程中充分考虑数据数据密集型应用所涉及数据集之间的依赖关系以及数据中心的存储量，从而减轻数据密集型应用运行时所涉及的数据集因跨数据中心所产生的时间消耗和传输次数，提高应用执行效率和系统性能。

本文第2节介绍相关工作，在此基础上进一步阐明本文与相关工作的差异与研究意义；第3节介绍数据布局问题，给出相关概念定义；第4节介绍云模型数据布局策略；第5节通过仿真实验部分对本文提出的策略进行测试和对比分析；最后一节，将对本文工作进行总结并对后续研究进行展望。

1. 相关工作

由于云计算日益受到重视，基于其上的数据管理系统层出不穷，例如Hadoop[5]、Google File System[6]等。常见的数据密集型应用系统都有其各自的数据管理策略，在数据布局方面也有研究。文献[7]提出了一种HDFS集群数据块动态配置算法，能够有效提高HDFS工作效率；Krish等人[8]针对Hadoop系统提出了一个完整的分层存储系统，其主要思想是依据物理存储设备的性能和工作模式来匹配数据服务请求，从而实现效率和能效的双目标优化；宋杰等人[9]提出了一种高能效的数据布局算法，能够有效地降低MapReduce集群的能耗。上述研究都能促进高能耗计算和大数据分析的应用，但并没有对跨数据中心产生的数据传输开销这一问题进行考虑和优化。

意识到数据布局的重要性，一些学者考虑到数据集之间存在的依赖关系，使用聚类方法进行数据布局的研究。Yuan等人[10]数据集之间的依赖关系，提出了一种结合K-means算法的数据布局方案，有效的减少了数据集移动次数。Den等人[11]利用数据集和任务之间的依赖关系，在Yuan提出的策略的基础上提出了一种高效的数据集和任务的协同调度策略，但此方法主要考虑负载均衡，导致同一类任务相关的数据存储在不同的数据中心上。应用在执行过程中，这些相关联的数据会被传输到任务执行时的数据中心，产生高额的跨数据中心时间开销。

同时，也有考虑到数据依赖关系，用智能方法对数据布局问题进行研究。郑湃等人[12]面向流程的数据密集型应用设计了一种基于遗传算法的三阶段数据布局策略，减少了数据集跨数据中心的传输次数和时间开销。张甜甜等人[13]提出一种基于释放和重构的数据布局策略，使得在超大规模的解空间中尽快找到更加接近全局最优的数据布局方案。文献[14]又提出了一种基于粒子群算法的数据布局策略，该方法在数据中心的负载均衡、算法收敛方面有很好的效果。但是上述方法时间复杂度较高，实用性不强。

综上分析可知，关于云计算环境下数据密集型应用的数据布局问题已经成为研究热点，但针对该问题的研究也存在一定的局限性：或者为了拥有很好的负载均衡，而难以对跨数据中心数据传输所导致的时间开销进行有针对性的优化；或者其数据布局策略本身时间开销较大，导致数据布局策略在灵活性方面有待改进；以及对于数据布局的过程中存在着的不确定性和随机性缺乏考虑。本文考虑到云环境中数据中心的存储大小，以及数据集之间的数据依赖关系，提出一种基于云模型的数据布局策略，实现减少跨数据中心传输所产生的时间开销和移动次数。

1. 问题概念与定义

3.1 数据布局相关概念

为了便于研究数据密集型应用在云计算环境下的数据布局问题，本文把云计算环境表示为多个分布的数据中心组成的集合。多个数据中心集合构成的云计算环境表示为,。其中表示编号为的数据中心，数据中心的可用存储空间记为。

将数据密集型应用表示为三元组。其中是中所有任务的集合；是各任务间的控制流的集合；是中所有数据集的集合，,。数据密集型应用中的某些数据只能存在固定的数据中心上，称这类数据集为固定位置数据集；对于位置没有特殊要求的数据集，称为可变位置数据集。本文不考虑固定位置数据集，为可变位置数据集。单个数据集定义为，。其中，表示第个数据集的随机位置，表示第个数据集的大小。

数据密集型应用在跨数据中心的数据传输过程中存在请求，响应，连接等将造成时间开销，任务调度过程也有时间消耗。考虑到数据密集型应用的数据规模巨大，这些时间开销远小于因数据布局跨数据中心产生的时间开销，故本文假设这部分时间开销忽略不计。设源数据中心、目标数据中心和目标数据集分别为，对，用表示数据中心和之间的网络带宽值。那么该数据集单次跨数据中心传输的时间开销可表示为

（1）

云计算环境下的数据布局，即指在多数据中心的云存储环境下为应用数据文件选择存放位置的过程。数据密集型应用的一个数据布局策略，是从集合到的映射，使对，存在唯一与之对应，定义如下：

**定义1. 数据布局策略**。在由已知的多数据中心组成的云环境下，数据密集型应用的一个数据布局策略为，并且满足：，有，如果，那么一定不存在。

3.2基于虚拟数据代理的数据布局模型

数据布局为每一个数据集在大量数据中心中选择合适的位置，往往涉及到数据集在各数据中心的多次移动，是一个复杂的多次择优过程，过程中产生时间及数据移动等各类开销。本文为了减少数据密集型应用因跨数据中心传输而产生的时间消耗和移动次数，在云计算系统中设计了引入虚拟数据代理后数据布局框架如图2所示。数据集不直接选择存储在哪个数据中心，而是先选择最适于自己的虚拟数据代理，而虚拟数据代理选择到最适宜的数据中心来存储。

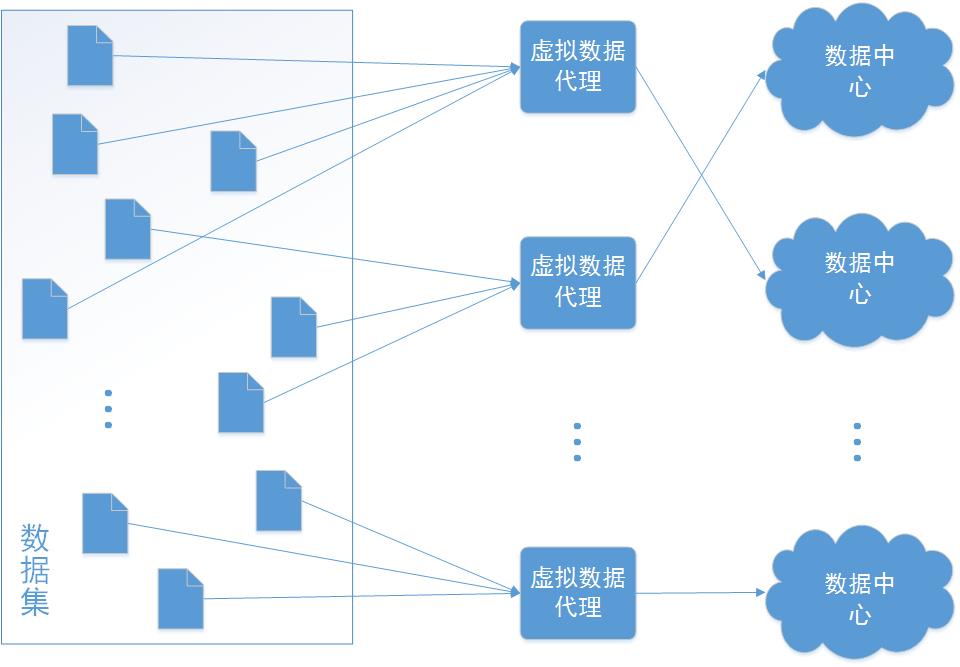


图2. 引入虚拟数据代理的数据布局框架

虚拟数据代理定义如下：

**定义2.虚拟数据代理。**虚拟数据代理是一个数据集集合，同一虚拟数据代理中的数据集特征相似，不同虚拟数据代理中的数据集特征相异。虚拟数据代理集合表示为：，其中，表示编号为的虚拟数据代理。

集合和集合之间满足满射,对在中必有与之对应。本文中，每一个虚拟数据代理存储在特定的数据中心，虚拟数据代理数量等同于数据中心的数量，即。集合与集合之间满足双射,使得对中的每个元素，在中有唯一确定的元素与之对应，有。

图2中的首先是数据集映射到数据中心，需通过虚拟数据代理作为中间件来辅助完成，根据数据集彼此间的数据特性，虚拟数据代理将关系密切的数据集拉拢过来，而虚拟数据代理则存储在数据中心中。定义2中的虚拟数据代理是一个存储在数据中心上的数据集集合，规定它先由一部分数据集开始，而后根据隶属度的计算不断加入新的数据集，直至全部数据集参与运算。隶属度是未布局的数据集隶属于某虚拟数据代理的判定值，原则为隶属度计算值越大，数据集隶属虚拟数据代理的可能性越高。为找出确立虚拟数据代理的数据集以及描述高依赖度的数据布局策略，需要对依赖关系等进行定义。

**定义3. 依赖关系。**数据请求可能需要多数据，同一数据也可能被多个请求处理，即数据集间存在依赖关系。如果数据集和被同一数据中心请求，则称数据集和相互依赖，记作，且，。通过以上定义易知：依赖关系具有传递性，即如果，，则。

**定义4．依赖度。**对，且，二者间的数据依赖度是一个非负整数，为其同时请求的数据中心的个数。记为，表示与的依赖度为。

**定义5．高依赖关系。**设T表示依赖度阈值，对于依赖关系，有，若时，称高度依赖于。

**定义6. 基于云模型的数据布局策略。**在由已知的多数据中心组成的云环境下，数据密集型应用的一个数据布局策略，是从集合到集合的映射,到的映射，使对，存在唯一与之对应，从而减少应用运行时数据跨数据中心传输时产生的时间消耗和移动次数。记数据密集型应用的一个数据布局方案为，则有，其中，。

1. 基于云模型的数据布局策略

本文基于虚拟数据代理的云模型数据布局包含两个关键内容：第一部分是确定数据集与虚拟数据代理间映射关系。首先，利用逆向云发生器实现从定量数值到定性概念的转换，确定虚拟数据代理云模型三元组；其次，利用正向云发生器，计算各数据集对各虚拟数据代理的隶属度，实现数据集到虚拟数据代理的映射。第二部分是确定虚拟数据代理所映射的数据中心。考虑数据集与数据中心的依赖关系实现虚拟数据代理与数据中心的映射。

4.1基于虚拟数据代理的数据布局策略分析

4.1.1虚拟数据代理云模型计算

云模型[15,16]是李德毅院士在概率理论和模糊集合理论基础上，重点考虑随机性和模糊性的关联性发展起来的定性知识描述和定性概念与其定量数值表示之间的不确定性转换模型，已经在智能控制、模糊评测、进化计算等多个领域得到应用[17,18,19]。它由期望,熵和超熵三个数值特征来表征一个定性概念。云的数字特征如图2所示：是所表征概念的中心值，它最能有效描述这个定性概念。综合反映概念的模糊性和概率，表达云的离散程度，在图中反映为云的跨度。是熵的熵，表达云的偏离程度，反映为云的厚度。其中，横轴表示某一概念的不确定性度量的范围，纵轴表示隶属度。

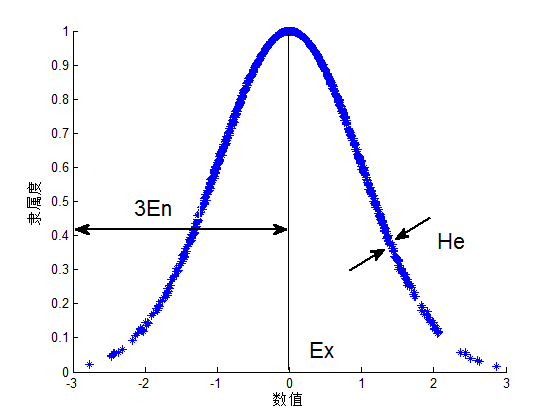


图2 云的数字特征

云模型有两个云产生算法[20]，即云发生器：正向云发生器(forward cloud generator)是用语言值描述的某个基本概念与其数值表示之间的不确定性转换模型，是从定性到定量的映射。逆向云发生器(backward cloud generator)是某个基本概念的数值表示和其语言值描述之间的不确定性转换模型，是从定量到定性的映射。

建立虚拟数据代理云模型，本文首先根据逆向云发生器生成三个特征值，以三元组表达虚拟数据代理。并依此来描述这些精确数据所反映的云滴的整体，云滴形成每一个云相当于一个虚拟数据代理。定性概念表达其所在区域的整个特征，为了尽可能准确表达一个概念，需要一定数量的数值特征，即提取一定数量的数据集来确定一个虚拟数据代理。想要获得一个高质量的虚拟数据代理，并充分利用数据密集型应用中数据集间的依赖性，本文考虑数据特性找出作为“云滴组”的数据集来共同表达定性概念。

利用逆向云发生器来确立虚拟数据代理云模型三元组，原则为具有高依赖关系的数据集优先作为“云滴组”。本文随机选出个数据集,,作为初始数据集，并利用定义5的依赖度阈值为每个初始数据集找到一定数量的相关数据集，构造组“云滴组”。在这里，为各初始数据集找到的相关数据集的个数可能不同，在下文特征值计算公式中将每一组的数据集数量统称为。利用逆向云发生器实现从依照数据特性选定的数据集到定性概念的转换，确立出个虚拟数据代理。对于每一个虚拟数据代理，本文按下式计算期望值、熵和超熵：

 （2）

是云模型表达概念的中心值，它最能有效描述这个定性概念，在本文数据布局策略中，在数据集映射到虚拟数据代理这一过程中起着至关重要的作用，为数据集的随机当前位置。

4.1.2数据集到虚拟数据代理的映射

本文使用正向云发生器，计算数据集对虚拟数据代理的隶属度，实现数据集到虚拟数据代理的映射，完成数据布局。

正态云模型将云期望曲线方程为正向云发生器的判断准则来生成云滴，本文考虑到数据密集型应用中数据集之间的依赖关系，期望高依赖度的数据集能存储在同一虚拟数据代理中，将依赖度引入云期望曲线方程，作为隶属度计算准则。根据虚拟数据代理云模型三元组中期望值和熵，定义虚拟数据代理云模型的云期望曲线方程如下：

（3）

为避免云模型生成过程虚拟数据代理大小的差异过大，保证所有虚拟数据代理云模型能够均匀增长，并考虑到数据中心的剩余存储问题，本文采取了一个策略：当一个数据集相对几个不同的虚拟数据代理具有相同的隶属度，该数据集分配到数据中心剩余存储量最大的虚拟数据代理。

我们以20个数据集和3个数据中心为例，将上述算法运用到给定条件中。随机抽取三个数据集分别为和，数据之间的依赖关系如表1所示。在这里，选取依赖度为3的数据集构成云滴组建立虚拟数据代理云模型。

表1. 数据集依赖度

|  |  |
| --- | --- |
|  |  |
|  | 1 2 1 2 1 2 3 2 1 0 1 -  2 3 0 2 3 2 0 1 3 1 -    3 1 1 1 0 -  3 1 2 3 -  2 3 **3** -  0 1 -  2 -  - |

通过上述方法，生成三个虚拟数据代理云模型，并将二十个数据集布局到三个虚拟数据代理中，结果如表2所示。

表2 数据集到虚拟数据代理的映射

|  |  |  |
| --- | --- | --- |
| 虚拟数据代理组成  （数据集编号） | 虚拟数据代理  （Ex,En,He） | 隶属该虚拟数据代理的数据集  （编号） |
| 2,5,9,19 | =（1.825, 0.219, 0.12） | 2,5,7,9,19 |
| 3,4,11,14,15 | =（2.52, 0.521, 0.292） | 3,4,11,13,14,15 |
| 1,6,10 | =（1, 0.251, 0.081） | 1,6,8,10,12,16,17,18,20 |

4.1.3 虚拟数据代理到数据中心的映射

由定义1可知，每个虚拟数据代理最终将与一个数据中心一一映射。每个虚拟数据代理是一组特征相似数据集的集合。云环境中，数据中心会发出数据请求，数据请求可能需要多个数据，同一数据集也可能被多个请求处理，即数据中心和数据集之间存在依赖关系。本文依据数据集和数据中心的依赖关系来实现这一映射。在虚拟数据代理云模型的计算过程中，最初已经随机抽取了个数据集作为个虚拟数据代理的第一个数据集，随后为这个初始数据集找到依赖度最强的数据中心即能找出虚拟数据代理最高依赖度的数据中心从而实现虚拟数据代理到数据中心的映射。假设有数据中心和，数据集和。发出请求需要2次数据集和1次，发出请求需要1次数据集，记每次被请求值加1，则会有如下矩阵：

那么和将分别作为两个虚拟数据代理，的开始，且虚拟数据代理，映射到的数据中心分别是和，其中和是来自两个数据中心的随机数据集。更一般地将数据中心与随机抽取的数据集依赖关系采用m\*m的矩阵表述如下：

|  |  |
| --- | --- |
|  |  |
|  |  |

其中行向量表示数据中心、列向量表示随机抽取的个数据集，为每个初始数据集行向量找到最大值，即为每个初始数据集所代表的虚拟数据代理找到对应的数据中心。同样采用4.1.1中的实例，随机选择的三个初始数据集与数据中心的依赖关系如下列矩阵所示：

运用本节方法得到结果如表3所示。

表3 虚拟数据代理到数据中心的映射

|  |  |
| --- | --- |
| 数据中心  （编号） | 虚拟数据代理  （Ex,En,He） |
| 3 | =（1, 0.251, 0.081） |
| 1 | =（1.825, 0.219, 0.12） |
| 2 | =（2.52, 0.521, 0.292） |

虚拟数据代理存储在数据中心上，同样的，虚拟数据代理则分别选择数据中心来存储隶属于自身的数据集。

4.2 基于虚拟数据代理云模型的数据布局

基于虚拟数据代理云模型的数据布局算法描述如下：

**云模型数据布局策略**

输入: 个随机数据集

开始：

**步骤1：for** (=0;<;++) **do**

找出满足与输入数据集相互之间有高依赖关系的数据集；

**步骤2：for** (=0;<;++) **do**

计算 公式（2）**and**

生成三个数字特征（，，），生成云滴的个数

**步骤3：for** (=0;<;++) **do**

找出与输入数据集依赖关系最高的数据中心 **and**

虚拟数据代理映射到数据中心；

**步骤4：for** (=0;<;++) **do**

计算 公式（3） **and**

个云滴及其确定度；

**步骤5： for** (=0;<; ++) **do**

的虚拟数据代理；

输出：数据布局结果（数据集映射到数据中心的情况） 。

依照上述算法步骤和规则执行，所有数据集都完成隶属度计算，实现数据集到虚拟数据代理的映射。使数据集最终放置到其虚拟数据代理映射到的数据中心，完成数据布局。不难看出，依照本文提出的基于云模型的数据布局策略给出的算法，此策略本身的时间复杂度会趋于一个常数，即，在算法收敛方面有很好的效果，实用性很高。

1. 实验与分析

5.1 实验环境与设置

为了验证本文提出的基于云模型的数据布局策略效果，在MATLAB R2014b上对数据密集型应用的执行过程进行了仿真实验。采用标准CyberSahke科学工作流，工作流中包含50个任务，并运用HEFT[21]任务调度算法来进行测试。通过分别改变科学工作流的数据集（文件数据）个数以及数据中心个数两种参数，与随机Random算法和最具代表性的K-means聚类算法[10]进行对比试验，三种数据布局策略从应用执行过程中的数据传输次数以及传输时间情况两方面进行对比和分析，本文策略在本节中简称为Cloud。为保证实验的公平性，5.2节中的结果均为运行500次所得结果的平均值。

5.2 实验结果和分析

5．2．1数据传输次数对比

图1横坐标表示数据集的数量，数据集数量由20渐增到100；纵坐标表示跨数据中心的数据传输次数。随着数据量的增多，三种策略的数据传输次数明显成上升趋势，但本文策略对应的传输次数处于最低。图2横坐标表示数据中心的数量，纵坐标表示跨数据中心的数据传输次数。当数据中心数量由3渐增到10，三种数据布局策略所对应的跨数据中心数据传输次数都呈明显上升趋势，虽然数据中心数量较少时本文策略与聚类策略相差甚微，但综合来看，随着数据中心数量的增多，从跨数据中心的数据移动次数角度来比较，本文策略始终处于最低位置。

图1 数据集个数对移动次数的影响 图2 数据中心个数对移动次数的影响

5．2．2数据传输时间对比

如图3和图4所示，分别从数据集数量和数据中心数量两个方面来对比三种布局策略对跨数据中心产生时间消耗的影响。从结果可以看出，随着数据集以及数据中心数量的增多，三种数据布局策略所对应的跨数据中心数据传输时间都越来越高，但本文策略所对应的数据传输时间开销明显低于其他两种策略。

图3 数据集变化对传输时间的影响 图4 数据中心变化对传输时间的影响

上述实验结果表明：在云计算环境中实现的数据密集型应用，随着数据量的增多和数据中心的增多，应用在执行过程中涉及到的跨数据中心数据移动次数和时间开销都在不断增长，而本文策略相比基于随机算法以及聚类算法的数据布局策略减少了一部分开销，能够得到较优的结果。

（初始m数据集分析）

1. 结束语

本文首先对数据密集型应用在云计算环境下部署并执行进行了简要说明，分析并指出云计算环境下这类应用存在多数据中心之间的数据传输开销这一挑战。针对相关研究存在的一些不足，以及数据布局问题自身存在的随机性与不确定性，本文提出一种基于云模型的数据布局策略，通过与其他数据布局策略进行实验对比，说明本文的数据布局策略具有较好的性能，在降低跨数据中心数据传输次数和时间开销方面，取得较优的结果。

本文数据布局过程中的核心——虚拟数据代理的确立，其结果高度敏感于第一个数据集，而本文采取的是随机抽取方法，对数据布局策略带来一定的影响。下一步将继续研究能够反映出真实情况的云模型数据布局算法，并最后将该方法应用于具体的数据密集型应用领域中。

**参考文献**

1. 宫学庆,金澈清,王晓玲,等. 数据密集型科学与工程: 需求和挑战[J]. 计算机学报, 2012, 35(8): 1563-1578.
2. Deelman E, Chervenak A. Data management challenges of data—intensive scientific workflows[J]. CCGRID, Lyon, France 2008：687—692
3. [Brian Van Essen](http://dblp.uni-trier.de/pers/hd/e/Essen:Brian_Van), [Henry Hsieh](http://dblp.uni-trier.de/pers/hd/h/Hsieh:Henry), [Sasha Ames](http://dblp.uni-trier.de/pers/hd/a/Ames:Sasha), et al. DI-MMAP—a scalable memory-map runtime for out-of-core data-intensive applications[J]. Cluster Comput, 2015, 18: 15-28.
4. Armbrust M, Fox A, Griffith R, et al. A view of cloud computing[J]. Communications of the ACM, 2010, 53(4): 50-58.
5. Apache Hadoop．What Is Apache Hadoop[EB/OL]．[2017-02-08]. http://hadoop.apache.org
6. Ghemawat S, Gobioff H, Leung S T. The google file system[J], SOSP,2003,37(5):29—43.
7. N. Maheshwari, R. Nanduri, V. Varma. Dynamic energy efficient data placement and cluster reconfiguration algorithm for MapReduce framework[J]. Future Generation Computer Systems, 2012, 28(1):119-127.
8. K. Krish, A. Anwar, A. Butt. hatS: A heterogeneity-aware tiered storage for Hadoop [C]. CCGRID, 2014:502-511.
9. 宋杰, 王智, 李甜甜, 等. 一种优化MapReduce系统能耗的数据布局算法[J]. 软件学报, 2015, 26(8): 2091-2110.
10. Yuan D, Yang Y, Liu X. A data placement strategy in scientific cloud workflows[J]. Future Generation Computer Systems, 2010,26(8):1200-1214.
11. Deng KF, Ren KJ, Song JQ, Yuan D, Xiang Y, Chen JJ. A clustering based coscheduling strategy for efficient scientific workflow execution in cloud computing[J]. Concurrency and Computation-Practice & Experience, 2013,25(18):2523−2539.
12. 郑湃, 崔立真, 王海洋等. 云计算环境下面向数据密集型应用的数据布局策略与方法[J]. 计算机学报, 2010, 33(8): 1472-1480.
13. 张甜甜,崔立真等. 基于释放和重构的科学工作流数据布局策略[J].计算机研究与发展,2013,50(Suppl.):71—76.
14. Pandey S, Wu LL, Guru SM, Buyya R. A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments[C]. In: Proc. of the 24th IEEE Int’l Conf. on Advanced Information Networking and Applications(AINA). 2010. 400−407.
15. Wang SG, Sun QB, Zhang GW, Yang FC. Uncertain QoS-aware Skyline service selection based on cloud model[J]. Journal of Software, 2012,23(6):1397−1412.
16. C. Xu, G. Wang, Q. Zhang, A new multi-step backward cloud transformation algorithm based on normal cloud model[J]. Fund. Inform. 2014, 133 (1): 55–85.
17. L. Yuchao, Z. Haisu, M. Yutao, L. Deyi, C. Guisheng, Collective intelligence and uncertain knowledge representation in cloud computing[J].China Commun. 2011, 8(6): 58–66.
18. G. Wang, C. Xu, D. Li, Generic normal cloud model[J]. Information Science. 2011, 280: 1–15.
19. J. Chen, Y. Liu, D. Li, Enhancing recommender diversity using gaussian cloud transformation[J]. [International Journal of Uncertainty](http://dblp.uni-trier.de/db/journals/ijufks/ijufks23.html#ChenLL15), Fuzziness Knowl.-Based Syst. 2015,23 (04): 521–544.
20. D.Y. Li, H.J. Meng, X.M. Shi, Membership clouds and membership cloud generators[J]. J. Comput. Res. Dev. 32 (6) (1995) 15–20.
21. Topcuoglu H, Hariri S, Wu M. Performance-effective and low-complexity task scheduling for heterogeneous computing[J]. IEEE transactions on parallel and distributed systems, 2002, 13(3): 260-274.