

50%. *K-means* 方法在时间上的优化次之, *CCDP4WF* 最差, 这两种方法的传输时间都高于 2 000s.

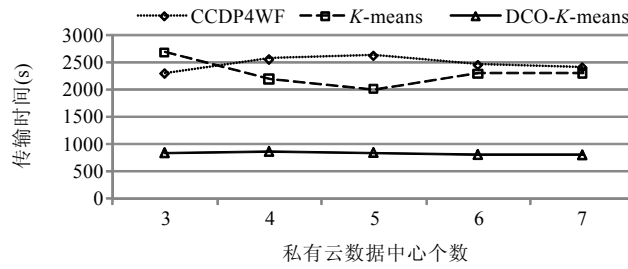


Fig.12 Comparison of transmission time with the number of private datacenters

图 12 不同私有云数据中心个数的传输时间

DCO-K-means 的布局方案仅与隐私数据的存放位置和数据依赖程度相关,在隐私数据集个数确定的情况下,非隐私数据集根据存储在不同数据中心之间的隐私数据集依赖关系,将依赖度高的非隐私数据集布局到对应的私有云数据中心,将依赖度低的非隐私数据集存储到公有云数据中心.因此,私有云数据中心的增加对 *DCO-K-means* 方法的数据布局不会造成大的影响,基本维持在 832s 左右的范围.然而,基于负载均衡的 *CCDP4WF* 和 *K-means* 方法会将所有的非隐私数据集均匀布局到逐渐增加的数据中心,破坏了数据集之间的依赖关系,导致数据传输时间不仅大于 *DCO-K-means* 方法,得到的传输时间稳定性也比 *DCO-K-means* 要差.

5.2.3 隐私数据集百分比变化对传输时间的影响

图 13 表示隐私数据集比例逐渐增加时数据传输时间的变化趋势,其中, workflow 数据集为 40 个,每个数据集 500MB,所有数据集布局到 1 个公有云、3 个私有云数据中心组成的混合云环境中.隐私数据集百分比为 0 时,3 种方法的传输时间为 0s,此时,所有的数据集均布局到公有云数据中心.随着隐私数据集百分比的增加,3 种方法的传输时间逐渐增加,并且增长的幅度逐渐减小,在隐私数据集占 70% 时趋于相同(4 950s).从图 13 可以看出:*CCDP4WF* 和 *K-means* 方法的传输时间大致一致,并在隐私数据集为 10%~40% 时明显高于本文的 *DCO-K-means* 方法,平均多消耗 40% 的时间.

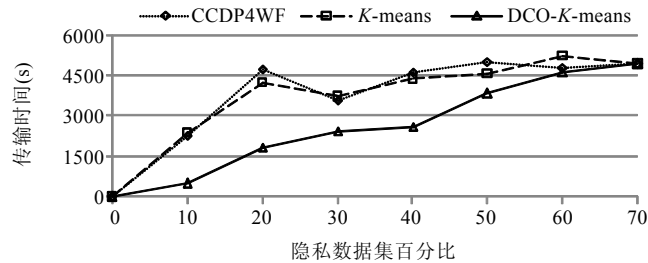


Fig.13 Comparison of transmission time with the rates of privacy datasets

图 13 不同隐私数据集百分比的传输时间

数据总量为 40 个,随着隐私数据集比例的增加,更多的隐私数据集存放在私有云数据中心,更少的非隐私数据集可供灵活布局,并且 workflow 数据和任务的依赖关系没有改变.所以, workflow 在执行时会产生更多的数据跨数据中心传输,导致数据的传输时间显著增加.当隐私数据集比率为 0 时,3 种方法均采用相同的聚类布局算法,将所有的非隐私数据集布局到同一个公有云数据中心上,数据的传输时间也为 0.当隐私数据集比率逐渐增加时, *CCDP4WF* 和 *K-means* 方法以负载均衡为目标,没有充分考虑非隐私和隐私数据集之间的依赖关系,导致数据传输时间显著增加.而 *DCO-K-means* 方法在非隐私和隐私数据集依赖关系的基础上对数据集布局,因此缩短了数据的传输时间,提升了 workflow 的执行效率.当隐私数据比率逐渐上升到 70% 左右时,3 种方法的传输时间逐

渐趋于相同,因为非隐私数据集和更多的隐私数据集存在相同的依赖度,同一个非隐私数据集有多个相同效果的布局方案,并且可供布局的非隐私数据集个数也较少.这两个主要因素导致 DCO-K-means 方法较 CCDP4WF 和 K-means 方法的优势不够明显,数据传输时间逐渐趋于相同.

5.2.4 数据集大小变化对传输时间的影响

图 14 为每个数据集大小逐渐改变时传输时间的变化趋势,其中, workflow 数据集为 40 个,隐私数据集占有数据集的 10%,所有数据集布局到 1 个公有云、3 个私有云数据中心组成的混合云环境中.本文假设该实验中所有数据集大小相同,若仅改变数据集大小而不改变 workflow 中任务和数据集之间的依赖关系,则不能充分反映数据集大小改变对 workflow 数据布局的影响状况.因此,本实验在数据集大小改变时,按照参数要求重新生成所有 workflow,并在相同环境下做 100 次相同实验取平均值.可以看到:随着数据集大小的逐渐增大,3 种布局方法的数据传输时间呈近似线性增长趋势.在数据集大小为 100MB~700MB 时,CCDP4WF 方法的时间优化最差,K-means 次之,DCO-K-means 最优,分别较 CCDP4WF 和 K-means 方法平均节约 71%和 67%左右的时间.

CCDP4WF 和 K-means 方法主要从负载均衡的角度布局,破坏了原有数据中隐私和非隐私数据集之间的依赖性,导致传输时间迅速增加.而 DCO-K-means 方法充分考虑了隐私和非隐私数据集的依赖关系,布局时尽量减少布局对数据依赖关系的破坏,数据传输时间随着数据集的增大而小幅增加,在传输时间上,明显优于前两种布局方法.每组实验均重新生成满足相同参数约束的 workflow,其中,数据集之间的依赖关系、隐私数据集的选取存在不确定性,可能导致数据集大小在 800MB 时,CCDP4WF 方法得到的布局方案突然优于 K-means 方法.

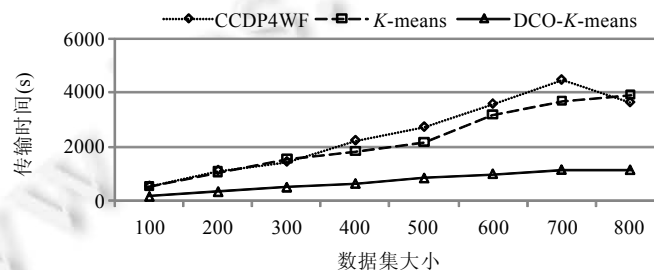


Fig. 14 Comparison of transmission time with the size of datasets

图 14 不同数据集大小的传输时间比较

综合上述 4 个方面的实验结果,DCO-K-means 方法的布局方案均优于 K-means 和 CCDP4WF 方法,有效地缩短了数据传输时间.对于数据规模大、隐私数据集比率为 10%~50%的科学 workflow 应用,DCO-K-means 优化效果会更明显.

6 结论与未来工作

本文首先介绍混合云环境的特点,数据密集型应用程序部署到混合云环境带来了新的机遇与挑战.对于科学 workflow 的数据布局,传统的数据布局方法通常使用聚类方法将数据布局到混合云中.该方法使用基于负载均衡的矩阵划分模型将数据集均匀地布局到所有的数据中心,但是得到的数据布局方案传输时间均非最优.针对传统方法的不足,我们提出一种新的面向数据中心的数据布局优化方法 DCO-K-means,该方法使用基于数据依赖破坏度的矩阵划分模型,将依赖度高的数据集布局到同一个数据中心,从而减少数据传输时间.该方法包括构建阶段静态布局算法和运行阶段动态布局算法,它们分别在工作流的构建阶段和执行阶段对数据进行布局,获得传输时间最优的布局方案.最后,与传统的 CCDP4WF 和 K-means 方法进行对比实验分析,表明 DCO-K-means 方法得到的布局方案优化效果明显.

本文在实验时假定所有数据集大小相同,所有数据中心之间的带宽相同,而实际情况下,数据集的大小差异很大,并且数据中心之间的带宽是动态的,这些因素会对本文的数据布局算法造成一定的影响.对于这些局限,

我们将继续研究能够反映数据集大小的加权数据依赖矩阵、能够表示数据中心之间带宽变化的数学模型,并最终将该方法应用于具体的科学 workflow 领域中。

References:

- [1] Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee GH, Patterson D, Rabkin A, Stoica I, Zaharia M. Above the clouds: A Berkeley view of cloud computing. Technical Report, No.UCB/EECS-2009-28, Berkeley: Department of Electrical Engineering and Computer Sciences, University of California, 2009.
- [2] Buyya R, Pandey S, Vecchiola C. Cloudbus Toolkit for Market-Oriented Cloud Computing. Berlin, Heidelberg: Springer-Verlag, 2009. 24–44. [doi: 10.1007/978-3-642-10665-1_4]
- [3] Sumit G. Public vs. private vs hybrid vs community-cloud computing: A critical review. *Int'l Journal of Computer Network and Information Security (IJCNIS)*, 2014,6(3):21–29. [doi: 10.5815/ijcnis.2014.03.03]
- [4] Grewal RK, Pateriya PK. A rule-based approach for effective resource provisioning in hybrid cloud environment. In: *Proc. of the New Paradigms in Internet Computing*. Berlin, Heidelberg: Springer-Verlag, 2013. 41–57. [doi: 10.1007/978-3-642-35461-8_5]
- [5] Fu J, Wang JC, Lu J, Chen ZH, He MQ. Research on meteorology indices forecasting framework based on hybrid cloud computing platforms. In: *Proc. of the Ubiquitous Information Technologies and Applications*. Netherlands: Springer-Verlag, 2013. 727–735. [doi: 10.1007/978-94-007-5857-5_78]
- [6] Szabo C, Sheng QZ, Kroeger T, Zhang YH, Jian Y. Science in the cloud: Allocation and execution of data-intensive scientific workflows. *Journal of Grid Computing*, 2014,12(2):245–264. [doi: 10.1007/s10723-013-9282-3]
- [7] Deelman E, Singh G, Sua M, Blythe J, Gil Y, Kesselman C, Mehta G, Vahi K, Berriman GB, Good J, Laity A, Jacob JC, Katz DS. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming*, 2005,13(3): 219–237. [doi: 10.1155/2005/128026]
- [8] Yuan D, Yang Y, Liu X. A data placement strategy in scientific cloud workflows. *Future Generation Computer Systems*, 2010, 26(8):1200–1214. [doi: 10.1016/j.future.2010.02.004]
- [9] Yuan D, Yang Y, Liu X. On-Demand minimum cost benchmarking for intermediate dataset storage in scientific cloud workflow systems. *Journal of Parallel & Distributed Computing*, 2011,71(2):316–332. [doi: 10.1016/j.jpdc.2010.09.003]
- [10] Keahey K, Figueiredo R, Freeman T, Tsugawa M. Science clouds: Early experiences in cloud computing for scientific applications. In: *Proc. of the Cloud Computing and its Applications*. 2008. 825–830. https://www.researchgate.net/publication/240918762_Science_Clouds_Early_Experiences_in_Cloud_Computing_for_Scientific_Applications
- [11] Wang LZ, Tao J, Kunze M, Castellanos AC, Kramer D, Karl W. Scientific cloud computing: early definition and experience. In: *Proc. of the 10th IEEE Int'l Conf. on High Performance Computing and Communications (HPCC 2008)*. 2008. 825–830. [doi: 10.1109/HPCC.2008.38]
- [12] Deelman E, Singh G, Livny M, Berriman B, Good J. The cost of doing science on the cloud: The montage example. In: *Proc. of the 2008 ACM/IEEE Conf. on Supercomputing*. 2008. 1–12. [doi: 10.1109/SC.2008.5217932]
- [13] Hoffa C, Mehta G, Freeman T, Deelman E, Keahey K, Berriman B, Good J. On the use of cloud computing for scientific workflows. In: *Proc. of the IEEE 4th Int'l Conf. (eScience 2008)*. 2008. 640–645. [doi: 10.1109/eScience.2008.167]
- [14] Zhang P, Wang GL, Xu XH. A data placement approach for workflow in cloud. *Journal of Computer Research and Development*, 2013,50(3):636–647 (in Chinese with English abstract).
- [15] Zheng P, Cui LZ, Wang HY, Xu M. A data placement strategy for data-intensive applications in cloud. *Chinese Journal of Computers*, 2010,33(8):1472–1480 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2010.01472]
- [16] Pandey S, Wu LL, Guru SM, Buyya R. A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments. In: *Proc. of the 24th IEEE Int'l Conf. on Advanced Information Networking and Applications (AINA)*. 2010. 400–407. [doi: 10.1109/AINA.2010.31]
- [17] Özsu MT, Valduriez P. *Principles of Distributed Database Systems*. 3rd ed., Berlin: Springer Science & Business Media, 2011. 31–41. [doi: 10.1007/978-1-4419-8834-8]

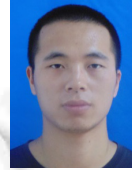
- [18] Deng KF, Ren KJ, Song JQ, Yuan D, Xiang Y, Chen JJ. A clustering based coscheduling strategy for efficient scientific workflow execution in cloud computing. *Concurrency and Computation-Practice & Experience*, 2013,25(18):2523–2539. [doi: 10.1002/cpe.3084]
- [19] McCormick WT, White T. Problem decomposition and data reorganization by a clustering technique. *Operations Research*, 1972, 20(5):993–1009. [doi: 10.1287/opre.20.5.993]

附中文参考文献:

- [14] 张鹏,王桂玲,徐学辉.云计算环境下适于 workflows 的数据布局方法.计算机研究与发展,2013,50(3):636–647.
- [15] 郑湃,崔立真,王海洋,徐猛.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.计算机学报,2010,33(8):1472–1480. [doi: 10.3724/SP.J.1016.2010.01472]



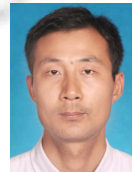
李学俊(1976—),男,安徽金寨人,博士,副教授,CCF 会员,主要研究领域为云计算,智能软件.



程慧敏(1990—),男,硕士生,主要研究领域为 workflow 系统.



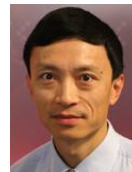
吴洋(1991—),男,硕士生,主要研究领域为云计算.



朱二周(1981—),男,博士,讲师,主要研究领域为虚拟化,云计算.



刘晓(1982—),男,博士,副教授,CCF 会员,主要研究领域为软件工程,云计算.



杨耘(1963—),男,博士,教授,博士生导师,主要研究领域为软件技术,云计算.