Scientific workflows perform computations exceeding single workstation’s capabilities. When running such data intensive workflows in the cloud distributed across several physical locations, the execution time and the resource utilization efficiency highly depends on the initial placement and distribution of the input datasets across these multiple virtual machines in the Cloud. The ideal data placement scheme optimizes the execution of the data intensive scientific workflows in cloud by assigning the tasks to the execution site in such a way that the file transfers and the cost associated are reduced. Several data placement strategies in cloud based scientific workflows are reviewed. A data placement scheme which uses big data to improve the performance and also the data movement cost is studied. BDAP (Big Data Placement strategy), improves workflow performance by minimizing data movement across multiple virtual machines.
8. Qiang Li, Kun Wang, Member, Suwei Wei, Xuefeng Han, Lili Xu, “A data placement strategy based on clustering and consistent hashing algorithm in Cloud Computing,” In 9th International Conference on Communications and Networking in China (CHINACOM), pp 278-283, 2014.

Index Terms
Computer Science Information Sciences

Keywords
Cloud computing, Big data, Scientific workflow, Data placement, Virtual machine.