© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 01 | January - March 2018
Paper is available at www.jrps.in | Email : info@jrps.in

Survey of Big Data Management on a Distributed Cloud

Namrata Mahakalkar,Asst Prof,PIET ,Nagpur

Abstact: In todays world more and more users share data and analysis results with each other to save the user cost of big data analytics. Communication and collaboration are increasingly important to modern big data applications. Big data management in a distributed cloud is a major challenge, which is to determine the efficient location of users' big data and other dynamic application data, so that as many users as possible can be served. Fair allocation of cloud resources to different users is crucial, otherwise, unfair allocation may result in unsatisfied users no longer using the service, the service provider may then fall into disrepute, and its revenue will be significantly reduced. There are numerous difficulties in the load balancing techniques, problem with sharing of resources such as security, fault tolerance etc. in cloud computing environments. Many researchers have been proposed several techniques to enhance the Big data management in a distributed cloud. This paper portrays presents a review of the current big data research, exploring applications, opportunities and challenges, as well as the state-of-the-art techniques and underlying models that exploit cloud computing technologies.

Keywords: Big data management, Cloud computing, Distributed cloud

I. Introduction

Cloud is the cluster of distributed computers that provides on-demand computational resources over a network .Cloud is a technology discontinuity that, with in next 10 years, is likely to dramatically change IT organizational missions, structures, roles, skills and operations. The cloud is changing our life by providing users with new types of services. User gets service from a cloud without paying attention to the details.. A Cloud computing is becoming an advanced technology in recent years. It is conceptually distributed system where computing resources distributed through the network (Cloud) and services pooled together to provide the users on pay-as-needed basis.

A. Cloud Computing

Cloud Computing became widely spread in last some years. Due to versatile use of internet cloud computing is becoming the back bone of soft computing.When a server is overloaded the arriving job should be diverted to the server which is in normal(underloaded) state such that there will be maximum utilization of the available resources. Cloud computing has given the IT sector new direction for utilization of resources in a organized manner as a user pay for usage, Cloud computing is a combined technique from the Grid Computing , utility computing and autonomic computing.

Cloud provides stretchy software as service (SaaS), Platform as Service(PaaS), Infrastructure as Service (IaaS) as shown in below fig.1.

Service is a delivery of a computing platform over the web where users can create and install their own application as they need. Configuration of computing platform and server is managed by the vendor or cloud provider. Example of PaaS is Google App Engine. Infrastructure as a Service (IaaS), where servers, software, and network equipment is provided as an on-demand service by the cloud provider [1].

1• Software as a Service (SaaS):

SaaS, sometimes referred to as "software on demand". Cloud application services, or Software as a Service (SaaS), represent the largest cloud market and are still growing quickly. SaaS uses the web to deliver applications that are managed by a third-party vendor and whose interface is accessed on the clients' side. Most SaaS applications can be run directly from a web browser without any downloads or installations required, although some require plugins. SaaS is software that is owned, delivered and managed remotely by one or more providers and is offered in a pay-as-per-use manner [3]. SaaS focuses on providing users with business specific capabilities such as e-mail or customer management [4]. The typical user of SaaS offering usually has neither knowledge nor control about the underlying infrastructure [3]. SaaS applications run on a SaaS provider's servers.

The provider manages access to the application, including security, availability, and performance. Improved access to data from any networked device while making it easier to manage privileges, monitor data use, and ensure everyone sees the same information at the same time.One of the examples of SaaS provider is Google Apps that provides large suite of web based applications for many business applications including accounting, enterprise resource management (ERP), human resource management (HRM), customer relationship management (CRM) and security device manager (SDM).

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 01 | January - March 2018
Paper is available at www.jrps.in | Email : info@jrps.in

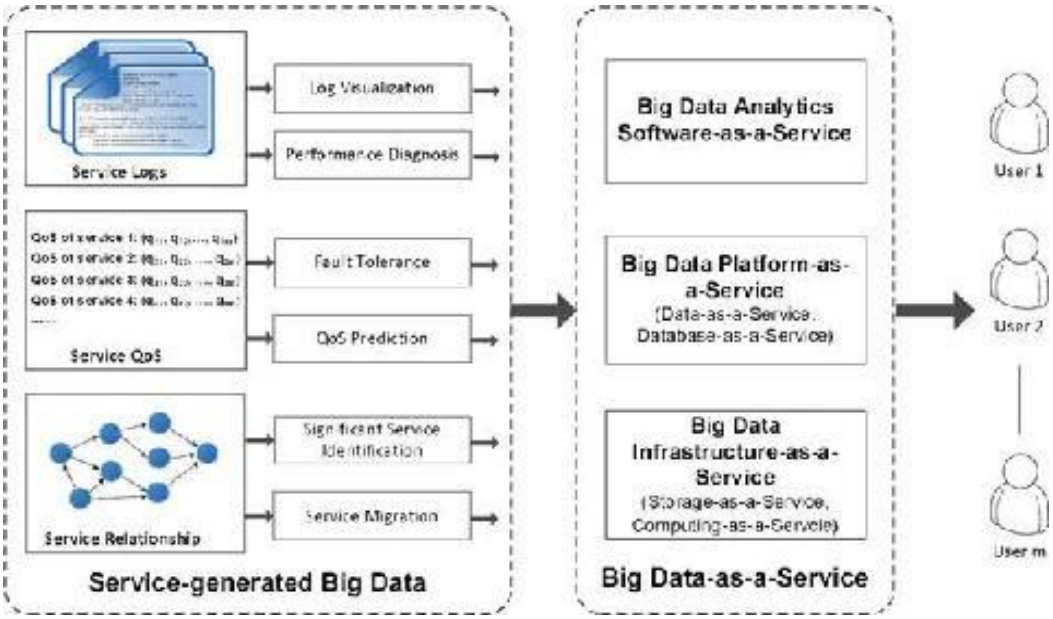


Fig. 1 Service-generated big data and big data-as-a-service

In this context, Zheng et al. [27] critically review the service-generated big data and big data-as-a-service Fig. 1) towards the proposal of an infrastructure to provide functionality for managing and analyzing different types of service-generated big data. A big data-as-a-service framework has been also employed to provide big data services and data analytics results to users, enhance efficiency and reduce cost.

- Platform as a Service (PaaS):

PaaS is a service model cloud computing. In this model, client creates the software using tools and libraries from the provider[4]. The client controls the applications that run in the environment, but does not control the operating system, hardware and network infrastructure on which they are running [3]. The provider provides the network, servers and storage. Risks associated with PaaS are provider downtime or a provider changing its development roadmap. If a provider stops supporting a certain programming language, users may be forced to change their programming language, or the provider itself. Both are difficult and disruptive steps. One of the examples of PaaS is Google App

Engine that provides clients to run their applications on Google's infrastructure [4].PaaS services include application design, development, testing, deployment and hosting. Other services include team collaboration, web service integration, database integration, security, scalability, storage, state management and versioning. PaaS also supports web development interfaces such as Simple Object Access Protocol (SOAP) and Representational State Transfer (REST), which allows the construction of multiple web services, sometimes called smashups. A downfall to PaaS is a lack of interoperability and portability among providers [2].

1• Infrastructure as a Service (IaaS):

IaaS, also known as cloud infrastructure services, delivers computer infrastructure-typically a platform virtualization outsourced service. IaaS model provides a virtual data center within the cloud. IaaS provides servers (physical and virtualized), cloud-based data storage, etc. IaaS platforms offer highly scalable resources that can be adjusted on-demand. This makes IaaS well-suited for workloads that are temporary, experimental or change unexpectedly. The client need not purchase the required servers, data center or the network resources. The key advantage is that customers need to pay only for the time duration they use the service [3]. One of the examples of IaaS providers is Amazon Elastic Compute Cloud (EC2). It provides users with a special virtual machine that can be deployed and run on EC2 infrastructure [4].

B. Distributed Cloud

As the next-generation cloud platforms, due to their rich cloud resources, Distributed clouds, consisting of multiple datacenters located at different geographical locations and interconnected by high-speed communication routes or links, are emerging, resilience to disasters, and low access delay [14, 19, 20, 21]. Big data management i.e. processing their petabyte-scale data and applications that rely on the rich resources provided by distributed clouds to store is increased with the escalation of data-intensive applications. The European radio telescope, LOFAR (Low-Frequency Array) based

on a vast array of omni-directional antennas located in Netherlands, Germany,the Great Britain, France and Sweden, produces up to five petabyte of raw data every four hours [17].

Sharing of collected data and obtaining valuable insights and scientific findings from such huge volume of data generated from different locations, data users need to upload and process their big data in a distributed cloud for cost savings. Thus, collaborative researchers at different geo-graphic locations can share the data by accessing and analyzing it. An application, Large Hadron Collider in physics research, generates over 60 TB data per day [16]. A naive placement for such large volume of big data may incur huge costs on data transmission among not only the datacenters but also the collaborated researchers. The data generated at different locations must be fairly placed to the distributed cloud. The potential revenue of the service provider would reduce if biased data is taken into consideration and its placements may severely degrade the reputation of the service provider. There are many reasons that encourage the use of distributed systems including:

1. **Resource sharing:** Resource sharing in distributed computer systems provides mechanisms for sharing and using remote hardware and software resources where a number of different computers are connected to each other via a network, then a user at any of these computers is able to use the resources available at the other computers.

2. Performance Improvement: Availability of distributed computing systems allow us to distribute the computation among various computing nodes to run it concurrently, if the problem to be solved is partitioned into a number of independent sub- problems that can run concurrently**.**

3. **Application Nature:** The use of high performance distributed systems for parallel and distributed applications is beneficial as compared to a single Central Processing Unit (CPU) machine The nature of parallel and distributed applications suggests the use of a communication network that connects several computers. Such networks are necessary for producing data that are required for the execution of tasks on remote resources. Also, most of the parallel and distributed applications have multiple processes that run concurrently on many computers (nodes) communicating over a high-speed interconnect.

TheFair allocation of cloud resources to different users is crucial, otherwise, unfair allocation may result in unsatisfied users no longer using the service, the service provider may then fall into disrepute, and its revenue will be significantly reduced. Clouds can be classified into three main types [5,6]: **a. Public Cloud**

This is the dominant form of current Cloud computing deployment model where resources are dynamically provisioned on a fine-grained, self-service basis over the internet, via web applications/web services, from an offsite third-parity provider who share resources. A public cloud sells services to anyone on the Internet. (Currently, Amazon Web Services is the largest public cloud provider.)

b. Private Cloud

A private cloud is a proprietary network or a data center that supplies hosted services to a limited number of people. Data and processes are manipulated within the organization without any restrictions such as network bandwidth, security exposures, and legal requirements that using public cloud services might entail.

c. Hybrid Cloud

This cloud infrastructure is a combination of two or more clouds (private, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability. To be private, public, or hybrid, the goal of cloud computing is to provide easy, scalable access to computing resources and IT services.Organizations use the hybrid cloud model in order to optimize their resources to increase their core competencies by margining out peripheral business functions onto the cloud while controlling core activities on-premise through private cloud. Fig2. Illustrates the cloud computing types.

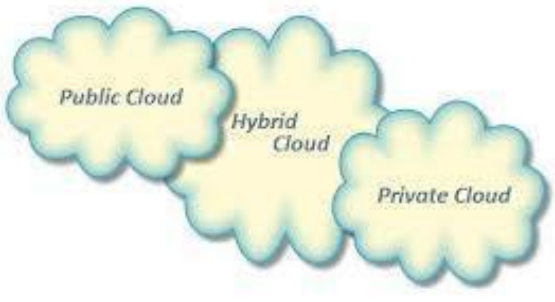


Fig2.Cloud computing types

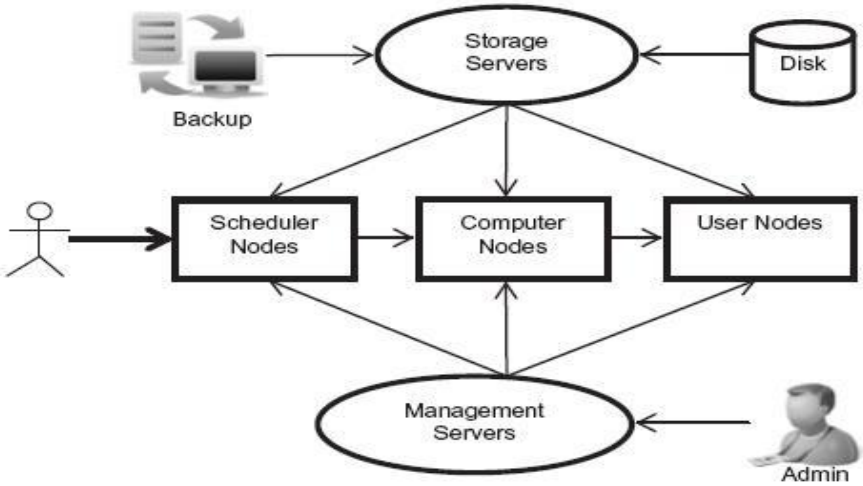
© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 01 | January - March 2018
Paper is available at www.jrps.in | Email : info@jrps.in

II. Literature Survey

High performance distributed Computing System has been divided into Cluster computing systems, Grid computing systems, and Cloud computing systems.

- **Cluster Computing Systems:**

The cluster systems based on load balancing integrate their nodes so that all requests from clients are distributed evenly across the nodes. The systems do not work together in a single process but redirecting requests independently as they arrive based on a scheduler and an algorithm.Load balancing between servers that have the same ability to respond to a client, have problems because one or more servers can respond to requests made and communication is impaired. The element that will make balancing between servers and users, and configured to do so, however multiple servers on one side, for the customers, appear to be only one address. A classic example would be the Linux Virtual Server, or simply prepare a DNS load balancer. The element of balance will have an
address, where customers try to make contact, called Virtual Server (VS), which redirects traffic to a server in the server pool. This element should be a software dedicated to doing all this management, or may be a network device that combines hardware performance and software to make the passage of the packages and load balancing in a single device.



Examples of CCS

1. **ALICE**: ALICE stands for the Ames Lab-ISU Computing Environment. All PCs are connected through a central Fast Ethernet switch, providing a flat network topology. In addition, there is a master node, a file server, and 4 development nodes [8]. It consists of 64 dual- processor Pentium Pros running at 200 MHz..

2. **G4 Cluster**: It consists of 16 single processor G4 computers each with 512 MB RAM and 16 dual processor G4s with 1 GB RAM. Each G4 is running Black Lab Linux and is connected via Fast Ethernet for management and Myrinet for communications [8].

3. **Sysplex**: The word sysplex comes from the first part of the word system and the last part of the word complex. It is a cluster-based approach for a mainframe system designed by IBM. It consists of multiple computers (the systems) that make up the complex. A sysplex is designed to be a solution for business needs involving any or all of the following: parallel processing; online transaction processing (OLTP); very high transaction volumes; very numerous small work units - online transactions, for example (or large work units that canbe broken up into multiple small work units); or applications running simultaneously on separate systems that must be able to update to a single database without compromising data integrity [7, 9].

- **Grid Computing**

Grid computing is made up of applications used for computational computer problems that are connected in a parallel networking environment. The hardware developments from recent years have made multi-core architectures a common place in the industry[26]. It connects each PC and combines information to form one application that is computation-intensive. Grid computing is standardized by the Global Grid Forum and applied by the Globus Alliance using the Globus Toolkit, the de facto standard for grid middleware that includes various application components. GCS supports the sharing and coordinated use of resources independently of their type and locationin dynamic virtual organizations (VOs) consisting of individuals, institutions, and resources solving computationally intensive applications. It uses common interface to link computing clusters or LANs together. These clusters are shared between many users or VOs and a local policy is applied to each cluster that defines their rules for resource sharing. Such rules

can be: what is shared on the basis of what condition and to whom, etc. Moreover, Grid guarantees the secure access by user identification [24]. What makes GCS different from CCS is that GCS tend to be more loosely coupled, heterogeneous, and geographically dispersed [25].

Several studies on data placement in clouds have been conducted in the past [11,23,13, 14]. However, most of these studies did not consider the placement of dynamically generated big data [13,14], and focused only on the communication costs [13]. Furthermore, they took neither fairness of resource allocations [11] nor the intermediate results of processed data into consideration [11,14]. For example, Golab et al. [13] studied the problem of data placement to minimize data communication cost for data-intensive tasks. Their goal was to determine where to store data and where to evaluate tasks in order to minimize data communication costs.

Jiao et al. [14] investigated multi-objective optimization for placing users' data for socially aware services over multiple clouds; they aimed to minimize the cost of updating and reading data by exploring trade-offs among the multiple optimization objectives. They solved the problem by decomposing it into two sub-problems- placement of master replicas and placement of slave replicas - and solving these two sub-problems separately, thereby deriving a sub-optimal solution to the problem. An algorithm for data placement in scientific cloud workloads, which grouped a set of datasets in multiple datacenters first, and then dynamically clustered newly generated datasets to the most appropriate datacenters based on data dependencies. Liu et al. [18] proposed a data placement strategy for scientific workflows by exploring data correlation, aimed at minimizing the communication overhead incurred by data movement. Agarwal et al. [11] proposed an automated data placement mechanism named Volley for geo-distributed cloud services, where the objective was to minimize the user-perceived latency.

III. Proposed Approach

A. Big Data Management in Distributed Clouds

The development of efficient solutions to the mentioned big data management problem is challenging, which lies in several aspects:

(i) The big data applications /collaboration-aware users dynamically and continuously generate data from different geographical locations.While managing such data due to their geo-distributions and large volumes the high cost will be incurred.

(2) Users typically have Quality of Service (QoS) requirements. Fair usage of cloud services is crucial, otherwise, biased allocation of cloud resources may result in that unsatisfied users no longer use the service, the service provider may fall into disrepute, and its revenue will be significantly reduced.

(3) The computing resource in each datacenter typically is limited [10]. Processing and analyzing the placed big data require massive computing resource. However, If the data placed in a datacenter cannot be processed as required, the overhead on migrating the placed data to other datacenters for processing will be high.

(4) Provisioning adequate computing and network resources for big data applications usually incurs a high operational cost, including the energy cost of powering servers in datacenters, the hardware cost on switches and routers between datacenters, and the communication cost for transmitting data along Internet links.

The collaboration- and fairness aware big data management problem in a distributed cloud to fairly place continuously generated data to the datacenters of the distributed cloud, the placed data are then processed, and the generated intermediate results finally are utilized by other collaborative users. Our objective is to maximize the system throughput, while keeping the operational cost of the service provider minimized, subject to the resource capacity and the fairness among users constraints, where the system throughput is the ratio of the amount of generated data that are successfully placed and processed to the total amount of data generated by each user. In other words, the system throughput is identical for each user, i.e., the proposed algorithm can fairly place the same percentage of data for each user into the system.

Despite that several studies in literature focused on big data management [22,18, 12, 13, 11, 15], we are not aware of any studies on the collaboration- and fairness aware big data management problem yet. For example, the studies in [22, 12, 13, 15] focused on data management based on given static data, while the work in [11, 15] did not consider collaborations and the fairness issue among users. They neither take the intermediate results of processed data nor the computing capacity of data centers into account.

B. Allocation of Big Data in Distributed Cloud:

Big data management in a distributed cloud is a major challenge, which is to determine the efficient location of users' big data and other dynamic application data, so that as many users as possible can be served. Checking a user's location, and migrate the user's data to his/her closest datacenter is the only issue. However, this simple heuristic ignores the cost of cloud service providers, and neglects the features of big data applications of users.

With the advance of information and communication technology, various types of data have grown at exponential rates. Efficiently and effectively managing and analyzing big data become crucial in creating competitive advantages, in answering scientific questions, and making effective decisions. Evaluating queries for big data analytics requires considerable storage, computing, and network resources across multiple data centers.

However, traditional data processing techniques cannot be adopted for big data analytics due to the large volume and complexity of the datasets that are dynamically generated and collected. In contrast, the distributed cloud provides an excellent platform for this by satisfying the resource demands of queries. As the computing capability of a single datacenter in distributed clouds is limited, and query evaluation can be made only when the resource requirement of the query is met, and further, the original data required by the query must be co-located with the query, the original data on which the query will be evaluated have to be replicated to the datacenter that can satisfy the resource requirement of the query.(1) throughput maximization of a cloudlet in mobile cloud computing, (2) collaboration and ensuring fair resource allocation for users for big data management in a distributed cloud, (3) cost minimization of query evaluation for big data analytics in a distributed cloud, and (4) operational cost minimization of a distributed cloud for data placement of online social networks.

In the proposed approach of big data management in the distributed cloud environments, where the problem consists of placing data, processing data, and transmitting the intermediate results of data processing to collaborative users located at different geographical locations.

(1) devise approximation algorithms to enable collaboration-aware big data management in a distributed cloud, given that data users such as enterprises, organizations, and institutes have considerable collaborations with their peers to build more wealth and improve the daily lives of people through jointly analyzing their data in different datacenters.

On one hand, existing studies ignored such collaborations among users, and simply applying their solutions will either incur useless analytic results with partial values, or high costs due to repeated data analysis on identical source data. On the other hand, simple heuristics without performance guarantees may lead to sub-optimal solutions;

(2) development of novel mechanisms to incorporate resource-allocation fairness into collaboration-aware big data management.

Specifically, in a distributed cloud, users are typically located in different geographical locations, and generate data at those locations. Such data must be fairly placed to the distributed cloud, otherwise, biased data placement may severely degrade access time, leading to degradation of the reputation of service providers, thereby potentially reducing the revenue of the service providers; and

(3) Naive resource allocation methods that neglect resource capacities will incur high overheads on migrating the placed data to other datacenters, if its current datacenter not have adequate available resources. Various types of resources with different capacities of a distributed cloud, by designing efficient resource allocation and provisioning mechanisms for collaboration-and fairness-aware big data management.

IV. Conclusion

Cloud computing will grow and with the age of BigData, the survey report proposed some of the key challenges existing in the field of cloud computing. With the existing tool and techniques it is not sufficient to adhere all the challenges relating to big volume of data. We have studied collaboration and fairness aware of Big data management in distributed clouds.Approach for the usages of various cloud resources at different datacenters of the distributed cloud. The paper analyzes how maximum number of users can also be served in the cloud and to minimize the cost of managing Big Data.

V. Future Scope

Future work would consider application of the above stated approach and Consideration of Big data on dynamically changing source data and resource sharing with each data centre.

VI. References

[1] A. Beloglazov, and R. Buyya, Energy efficient resource management in virtualized cloud data centers, Proc. 10th IEEE/ACM international conference on cluster, cloud and grid computing, 2010, 826-831.

[2]Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi and Jameela Al-Jaroodi" A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms" 2012 IEEE Second Symposium on Network Cloud Computing and Applications.
[3]. J. Srinivas, K.V.S.Reddy and A.M. Qyser, "Cloud \Computing Basics", International Journal of Advanced Research in Computer and Communication Engineering, 1(5), July 2012.
[4]. S. Ray and A.D. Sarkar, "Execution Analysis of Load Balancing Algorithms in Cloud Computing Environment", International Journal on Cloud Computing Services and Architecture, 2(5), October 2012.
[5] http://searchcloudcomputing.techtarget.com/definition/cloud-computing
[6] Tharam Dillon, Chen Wu and Elizabeth Chang," Cloud Computing: Issues and Challenges", 24th IEEE International Conference on Advanced Information Networking and Applications, 2010.
[7] Parallel sysplex, http://www-03.ibm.com/systems/z/ advantages/pso/.
[8] http://www.scl.ameslab.gov/Projects/parallel_computing/cluster_examples.html
[9] http://searchdatacenter.techtarget.com/definition/sysplex-and-Parallel-Sysplex
[10] M. Alicherry and T.V. Lakshman. Network aware resource allocation in distributed clouds. Proc. of INFOCOM, IEEE, 2012.
[11]. S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan. Volley: automated data placement for geo-distributed cloud services. Proc. of NSDI, USENIX, 2010.
[12]. I. Baev, R. Rajaraman, and C. Swamy. Approximation algorithms for data placement problems. SIAM J. on Computing, Vol. 38, No. 4, pp.1411-1429, SIAM, 2008.
[13]. L. Golab, M. Hadjieleftheriou, H. Karloff, and B. Saha. Distributed data placement to minimize communication costs via graph partitioning. Proc. of SSDBM, ACM, 2014.
[14]. L. Gu, D. Zeng, P. Li, and S. Guo. Cost minimization for big data processing in geo-distributed data centers. Trans. on Emerging Topics in Computing, Vol. 2, No. 3,pp.314-323, IEEE, 2014.
[15]. L. Jiao, J. Li, W. Du, and X. Fu. Multi-objective data placement for multi-cloud socially aware services. Proc. of INFOCOM, IEEE, 2014.
[16]. 'The Large Hadron Collider. http://home.web.cern.ch/topics/large-hadron-collider, 2016.
[17]. LOFAR. http://www.lofar.org/, 2016.
[18]. X. Liu and A. Datta. Towards intelligent data placement for scientific workflows\in collaborative cloud environment. Proc. of IPDPS, IEEE, 2011.
[19]. Z. Xu and W. Liang. Minimizing the operational cost of data centers via geographical electricity price diversity. Proc. of CLOUD, IEEE, 2013.
[20]. Z. Xu and W. Liang. Operational cost minimization of distributed data centers through the provision of fair request rate allocations while meeting different user SLAs. Computer Networks, Vol. 83, pp. 59-75, Elsevier, 2015.
[21]. L. Zhang, Z. Li, C. Wu, and M. Chen. Online algorithms for uploading deferrable big data to the cloud. Proc. of INFOCOM, IEEE, 2014.
[22]. D. Yuan, Y. Yang, X. Liu, and J. Chen. A data placement strategy in scientific cloud workflows. J. of Future Generation Computer Systems, Vol. 26, No, 8, pp.1200-1214, IEEE, 2010.
[23] I. Baev, R. Rajaraman, and C. Swamy. Approximation algorithms for data placement problems. SIAM J. on Computing, Vol. 38, No. 4, pp.1411-1429, SIAM, 2008.
[24]S. F. El-Zoghdy, "An AntNet-based Load Balancing Algorithm for Grid Computing Environments", International Journal of Computer Applications (0975 – 8887, Vol. 83, No 7, December 2013.
[25]Grid computing, http://www.adarshpatil.com/newsite/images/grid-computing.gif
[26]Namrata Mahakalkar,A.R.Mahajan ," Multi-GPU Island-Based Genetic Algorithm", IJACTE,Volume-1, Issue-2, 2012
[27]. Zheng, Z., Zhu, J., Lyu, M.R.: Service-generated big data and big data-as-a-service: anoverview. Proceedings of the 2013 IEEE International Congress on Big Data (BigData Congress), pp. 403–410. Santa Clara, California (2013)
[28] Nitesh Thakur,Namrata Mahakalkar," Artificially Intelligent Chatbot",Universal Research Reports,Volume 1 ,Issue 6,September 2017.
[29] Nitesh Thakur,Namrata Mahakalkar,"Artificially Intelligent Chatbot",Innovative Research Thoughts,Vol 4 Issue 5,April 2018.