Allocating Resources using IaaS Cloud Service for Big Data Applications

Mrs Madhura J
Asst. Professor, Dept. of ISE, DSCE

Asha D K
Department of Information Science
Dayananda Sagar College of Engineering
Bangalore, India

Deepa S Doddagoudar
Department of Information Science
Dayananda Sagar College of Engineering
Bangalore, India

Mamata
Department of Information Science
Dayananda Sagar College of Engineering
Bangalore, India

Raunak Kumar Sahu
Department of Information Science
Dayananda Sagar College of Engineering
Bangalore, India

ABSTRACT
The wide increase in the popular big data applications will lead to invaluable information and also with the challenges to IT industry. The solution to the problem is cloud computing with unlimited resources. Arrangement of fine allocation for cloud infrastructure resources is an important task to achieve this. In this paper, we have presented an algorithm which satisfies the constraints like execution, accessibility and expense of Big Data Application running on cloud. After analysing the relations among these objectives, we have implemented our idea on experimental environment. Our approach will run about 20% faster than custom approaches and will achieve about 15% higher execution than other analytical algorithms while saving 4%-20% expense.

Key-words: Cloud computing, Big data application, Virtual machines, Hadoop.

INTRODUCTION
The term “Big Data” means innovative techniques and technologies to store, capture, distribute, manage and analyse petabyte and larger-sized datasets with varying velocity and different structures. Big data can be organized, unorganized or semi-organized, which leads to impotence of traditional data management methods. Data arrives in the system at various rates which are reproduced from various different sources. [1] In order to process these bulky datasets of data in an economical and effective way, parallelism is used. Big Data is a data whose scale, variety, and multiplicity need new architecture, techniques, and algorithms to organize it and fetch value and hidden knowledge from it. To organize Big Data, Hadoop is used which is the core platform to solve the problem of making Big Data useful for analytics purposes. [2] Hadoop which is an open source software project allows the distributed processing of bulkier data sets across group of commodity servers. It is built to gear up from a single server to thousands of machines, which results in a high degree of fault tolerance. [3] Big data is a concept that points to data sets or combinations of data sets whose volume, variability and velocity make them difficult to be captured, organized, executed or analysed by custom technologies and tools, such as relational databases. [4] Big data is definitely one of the popular phrases in IT today. Integrated with virtualization and cloud computing, big data is a technological calibre that will enforce data centres to transpose and develop within the next five years. Associated with virtualization, big data infrastructure is unique and can create an architectural upheaval in the way systems, storage, and software infrastructure are associated and maintained. Unlike earlier business analytics results, the real-time capacity of novel big data solutions can offer essential business intelligence
that can transpose the structure and speed of enterprise decision building forever.

The cloud proposes new levels of scalability, adaptability and accessibility, and grants easy access to data from different locations and from different devices. In addition, the cloud represents a data model of objects that stores data integrated among its user-defined and system-defined metadata as a single part. [5] Thus, the cloud is an appealing model for a recent form of scalable pre-determined content applications that involve rich metadata. Together trends, ubiquitous sensing and Cloud computing, balance each other in a normal way. [6] Sensor networks fetch information related to physical environment, but in general sensor networks require the resources to store and process the collected data over long periods of time. Cloud computing is flexible to present the missing storage and computing resources. Particularly, it grants storing, processing, and accessing the collected sensor data effectively via cloud-based services. Big data and cloud computing are together the rapidly-moving technologies. Cloud computing is connected with new concept to provide computing infrastructure and big data processing technique for all kinds of resources. [7] One important quality of cloud computing is an aggregation of resources and aggregation of data into data centres on the Internet. The present cloud services like IaaS, PaaS and SaaS take in better performance effectiveness by aggregating application execution environments at different levels involving server, OS and middleware levels for circulating them. Meanwhile, a different approach of aggregating data into clouds has also been introduced, and it is to analyse such data with the efficient computational capability of clouds. As cloud computing services become well-known, more and more academic, business, and personal computing applications are deployed in the shared computing environment. [8] Clients of cloud services want to minimize the execution time of their jobs without crossing the expected expense under the given requirements, while cloud providers attempt to maximize the use of resources and gain more profits. Infrastructure as a Service (IaaS) clouds have greatly minimized the investment risk of owning an infrastructure. It offers computers as physical or more often as virtual machines (VMs). To run parallel or distributed applications such as Map-Reduce, a cluster of VMs, virtual cluster, is often requested as a platform for users.

LITERATURE SURVEY


The effectiveness and scalability of MapReduce-based implementations of complex data-intensive tasks depend on an even redistribution of data between maps and reduce tasks. [9] In the presence of skewed data, sophisticated redistribution approaches become important to carry out load balancing among all reduce tasks to be processed in parallel. For the complex problem of entity resolution, the two approaches for such skew handling and load balancing are proposed and evaluated. [10] The approaches support blocking techniques to decrease the search space of entity resolution, use a preprocessing MapReduce job to consider the data distribution, and distribute the entities of large blocks among multiple reduce tasks. The assessment on a real cloud infrastructure presents the value and effectiveness of the proposed load balancing approaches.

Advantages:
The Block-Split approach is simpler than PairRange but carry out already excellent results.

Disadvantages:
It does not divide large blocks but still processes all data sharing the same key at the same reduce task which results in unbalanced reduce workloads.


Huge companies like Google and Microsoft and a number of small and medium enterprises everyday process large amounts of data in batch jobs and in real time applications. This leads to high network traffic, which is hard to support using conventional, oversubscribed, network infrastructures. To deal with this issue, most network topologies have been proposed, attempting to enhance the bandwidth available in enterprise clusters. It is observed that in many of the commonly used work-loads, data is combined during the process and the output size is a division of the input size. This inspired us to get into a different point in the design space. Instead of enhancing the bandwidth, the focus is to decrease the traffic by pushing aggregation from the edge into the network.
Advantages:
Camdoop on CamCube would still outperform them in many situations, due to the capability of significantly reducing network traffic by aggregating packets on-path.

Disadvantages:
Switches are utilized only to connect CamCube servers to the external networks but are not utilized to route internal traffic.


As a distributed data - parallelization (DDP) pattern, MapReduce has been chosen by many new big data analysis tools to achieve good scalability and execution in Cluster or Cloud environments. How two binary DDP patterns, i.e., CoGroup and Match, can also be used in these tools is explored. An existing bioinformatics tool, called CloudBurst, with three different DDP pattern combinations is implemented. The two factors, namely, input data balancing and value sparseness is identified which could greatly affect the executions using different DDP patterns. The experiments show: (i) a simple DDP pattern switch can gear up execution by two times; (ii) the identified factors can show the dissimilarities well. CloudBurst is a parallel seed-and-extend sequence mapping tool. Bychoosing Map and Reduce patterns, CloudBurst has Seed Extraction, Data Shuffle and Seed Extension phases, as expected. The Seed Extraction phase utilizes the Map pattern to execute both the query and reference sequences and discharges seeds in parallel. Its result is a list of <key, value> pairs where key is the seed identifier and value is detailed information of the seed, containing its sequence string and offset in original sequence. The Seed Extension phase lengthens the shared seeds by allowing mismatches. This phase utilizes the Reduce pattern, so different seeds can be executed in parallel. In this phase, CloudBurst first divides the value list into two lists by examining whether it is query or reference. Next, it fetches all pairs of elements from the Cartesian result of the two lists, and extends seeds for these pairs to produce mapping results. Based on the results in [4], CloudBurst is selected as a good case study to try different DDP patterns and evaluate their dissimilarities. The query and reference datasets of CloudBurst have to be categorized throughout the phases. Since Map and Reduce only encourage one input dataset, the query and reference datasets have to be made a distinction in an indirect way when utilizing Map and Reduce patterns.

PROPOSED SYSTEM
The infrastructure data allocation problem in clouds contains three major goals, which are expense, execution and accessibility. The solution for data allocation deployment is to accomplish the greater execution and accessibility with the lowest expense. In current cloud deployment solutions, different data mediums can be utilized for different purposes. Different types of data medium have different expense, execution and accessibility respectively. In general, the higher the execution and accessibility are, the higher the expense will be. VMs are used as the basic working node in cloud infrastructure, to quantify data allocation problem. Imagining that all the VMs are with the same configuration, the more the number is, the greater execution and accessibility are, the greater the cost will be. Meanwhile, the topological structure of data allocation has close relation with execution and accessibility. Because of the communication cost among working Virtual Machines, the longer the distance between them is, the lower execution will be. However, the longer the distance is, the greater accessibility and security will be. In conclusion, the expense, execution and accessibility interplay with each other in form of the topological structure. Heterogeneous data is an important feature of cloud-based applications. Consequently, the VMs used in cloud-based big data applications are with the different configuration. They are with different CPU, memory, OS, networking bandwidth and geographic location. As a result, heterogeneity has to be considered, which data allocation problem extremely complicated to solve. The system proposed a distributed algorithm for big data applications by breaking down the actual large-scale problem into many subproblems that can be executed in parallel. The system design an online algorithm whose idea is to delay the migration operation until the cumulative traffic price crosses a threshold. The network topology is based on three-tier architectures. An access tier, an aggregation tier and a core tier. The system checks network traffic
reduction within MapReduce jobs by jointly misusing traffic-aware intermediate data partition and data aggregation between multiple map tasks. It presents computers as physical or more just as virtual machines. A cluster of virtual machines, virtual cluster, is usually requested as a platform for users to run parallel or distributed applications such as MapReduce. To achieve greater throughput, fast response, load balance and low price, most concepts on VM configuration, VM placement, VM consolidation, and VM migration are discovered. The network topology of a virtual cluster has a definite impact on the performance of applications running on it because the physical nodes where VMs are situated can be tagged in different ways. For example, some nodes are situated in the same rack while others in different racks along a slow link. Another special architecture is the hierarchical network where two physical nodes may be situated in different local area networks. Furthermore, the features of different applications have different needs for the network topology. Some applications develop tasks running on various VMs which are required to interchange large amount of data frequently, while others develop tasks which process individually or interchange a little data. MapReduce and MapReduce-like models are widely used to execute “Big Data”. Applications based on such models allow heavily data-dependency or communication on VMs, so network traffic achieves the bottleneck of jobs.

**SYSTEM ARCHITECTURE**

According to Architectural Diagram, file is uploaded into the cloud by the data owner. Segmentation of dataset takes place in the cloud by the segmentor. The SLA process module is the input, and it transmits the requirements about execution time, accessibility and expense to next module. [3] The provisioning module receives the requirements from the SLA process module and prepares all parameters for the BRA algorithm linked with server configuration module. Some machine learning techniques are implemented in the provisioning module to fetch some information related to Virtual Machines. Usually it takes some time to execute the workload and produce the solution. Consequently, it is necessary to start the evaluation before the workload really happens. [4] The training data occurs from the server configuration module, which is the connection between practical server information and our approach. [4] The monitor and feedback module is used to guide the real-time status of running VMs, and it will enhance re-configuration request to server configuration module once some exceptional circumstances occur.

**FLOWCHART**

According to flow chart, initially file is uploaded by the data owner to the cloud. The segmentor segments the data file into independent chunks of data. MapReduce technique is applied. Key-value pairs are generated for each individual data. File is divided into datasets and each dataset is assigned with virtual machines. These virtual machines are inter-connected using Ethernet switches. [6] As per the SLA agreement that is Service Level Agreement, there must be the profit for both the parties like for the client and the service provider. Provisioning takes place. [5] The provisioning...
means making the list of all the requirements from the SLA process and preparing all parameters needed by the BRA algorithm. Server configuration also occurs at provisioning module. BRA and GA algorithm is executed in order to meet the requirements of SLA process and also to reduce the network traffic. Finally, evaluation process takes place. [6] The algorithm is executed under the online cases like OHRA and OHNA. The output dataset is obtained which is well-structured, easy to access, manipulate and retrieve the data.

EVALUATION
The optimal cloud infrastructure deployment solution should be economical with greater execution and accessibility. However, the expense is counteractive towards the execution and accessibility in form of Virtual Machines. At the same time, the performance and the availability are mutually restrictive to each other in form of the distance between VMs.

The Execution Time is calculated by the formula:

$$\text{Execution Time} = (\text{Current Time} – \text{Initial Time})/1000 \quad \text{Eqn 1}$$

The Expense is calculated by the formula:

$$\text{Expense} = 100 \times \text{Execution Time} \quad \text{Eqn 2}$$

$$\text{Total Energy Cost} = \text{Computation cost} + (\text{Rate} \times \text{Weight}) \quad \text{Eqn 3}$$

$$\text{Arrival rate} = \text{Total Energy Cost} \times N \text{ datacentres}$$

Where N is number of datacentres $$\text{Eqn 4}$$

Hence the Total Energy Cost and Arrival rate is calculated as shown above.

CONCLUSION AND FUTURE WORK
The relations among the expense, execution and accessibility of one cloud-based big data application are analysed and built three models. Based on these three models, BRA algorithm and GA algorithm are proposed to obtain the optimal solution meeting all requirements. Then a complete approach has been designed and approached to allocate resources of big data application running on cloud. Finally, three sets of SLAs are performed to verify the feasibility of our approach, and compared it with seven other approaches to show the effectiveness. In future, research has to be done in two aspects. The first one is to add more constraints, including the security and data processing preference. The second one is to test our approach on advanced networking environments such as Software-Defined Networking (SDN).

REFERENCES
