Performance Evaluation of Big Data by Applying Ant Colony Optimization Techniques

Prasad Suman Sourav[1], Mohanty Anita[1], Mishra Sambit Kumar[2,*]

[1]Department of MCA, A.B.I.T., Cuttack
[2]Department of Computer Sc.&Engg, Gandhi Institute for Education & Technology, Bhubaneswar
*Corresponding author: sambitmishra@gietbbsr.com

Abstract Big data is a collection of huge amount of data. As the world is changing rapidly, many new technologies, devices such as smart phones, social networking sites have been evolved due to which the amount of data produced day by day is increasing rapidly. It has become a problem for many companies to process such a huge amount of data using traditional computing techniques. The collection of each and every data of a company(homogeneous or heterogeneous data) is called big data. Research is being carried out to find an appropriate algorithm to find an optimal solution when the size of the database increases. Most of the data we are handling today are of unstructured type like the data in social sites, research engines, blogs etc. The challenges we face with big data today is not only to store or link but also to retrieve, update and analyze them too. Now this big data is needed to be processed on some platform. This platform on which big data is operated is known as cloud computing. Anyone may process big data on cloud computing without the need of any specific software. Cloud computing can expand and shrink as per the need of storage. Cloud computing mainly provides resources as and when needed. As big data is also a kind of resource so it is also available through cloud computing. In this paper, ant colony optimization technique may be applied to evaluate the performance while processing queries in big data.

Keywords: big data, ant colony optimization, cloud computing, query optimization

Cite This Article: Prasad Suman Sourav, Mohanty Anita, and Mishra Sambit Kumar, "Performance Evaluation of Big Data by Applying Ant Colony Optimization Techniques." *Journal of Computer Sciences and Applications*, vol. 3, no. 6 (2015): 134-136. doi: 10.12691/jcsa-3-6-5.

1. Introduction

Today big data is a hot topic in large organizations. Many organizations collect, store and analyze huge amounts of data. This data is commonly known as "big data" because of its volume, the velocity with which it arrives and the variety of data it stores. It can be better defined as

- High volume—means amount of data
- High velocity—the rate at which data created
- High variety—the different types of data(both homogeneous and heterogeneous)

As big data has all the above three characteristics so new technologies and techniques are required to capture, store and analyze it.

Big data is captured from many sources. For example, each and every mouse click on a web site can be captured in a file called Web log files and then analyzed to understand the customers buying behaviours in a better way and to influence their shopping by dynamically recommending products. Another example is social media sites such as facebook, twitter etc. generate huge amounts of comments and tweets every second. This data is captured and analyzed to understand what people think about new products introduction.

In today's date another topic is in boom i.e cloud computing. The cloud makes it possible for users to access information from anywhere anytime. It removes the need for users to be in the same location as the hardware that stores data. Once the internet connection is established either with wireless or broadband, user can access services of cloud computing through various hardwares. This hardware could be a desktop, laptop, tablet or phone. Cloud provides a reliable online storage space. It is the way to store your software or data in Internet (server) and you simply use this either free or sometimes paid.
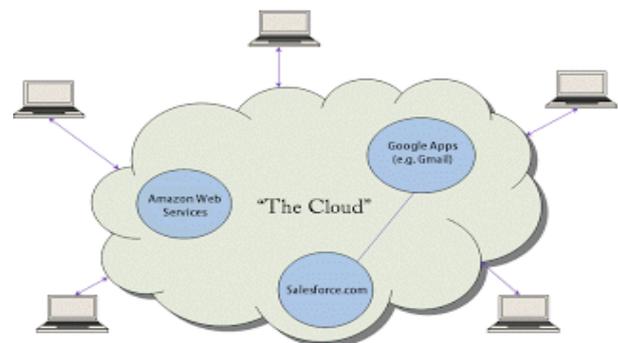
Big data can store all types of data like structured, semi-structured and unstructured. Now a days the data stored in big data can be operated on clusters of computers at the same time. The user always thinks that he is

operating on a single system which is actually working through a number of computers. Big data is always associated with cloud computing. Big data is implemented on cloud computing platform to do all the necessary operations on data. Cloud is present in a remote location about which user is unaware of.

Now a question arises how the processing of big data will be faster and secure. In this paper we may implement ant colony optimization technique on big data to access it in a faster and secure way.

2. Review of Literature

Marcos D. Assuncao et.al. [1] have discussed approaches and environments for carrying out analytics on clouds for Big data applications.They have identified possible gaps in technology and provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions.

Khairul Munadi et.al. [2] have proposed a conceptual image trading framework that enables secure storage and retrieval over internet services. The aim is to facilitate secure storage and retrieval of original images for commercial transactions, while preventing untrusted server providers and unauthorized users from gaining access to true contents.

Rupali S. Khachane et.al. [3] have focused on Privacy Homomorphism technique which emphasize to resolve the security of query processing from client side, cloud with R-tree index query and distance re-coding algorithm.

Badrish Chandramouli et.al. [4] have proposed a new progressive analytics system based on a progress model called Prism that allows users to communicate progressive samples to the system and efficient and deterministic query processing over samples.

Satoshi Tsuchiya et.al. [5] have discussed about two fundamental technologies : distributed data store and complex event processing, and workflow description for distributed data processing.

Divyakant Agrawal et.al. [6] have focused on an organized picture of the challenges faced by application developers and DBMS designers in developing and deploying internet scale applications.

Ms.Preeti Tiwari et.al. [7] have discussed that the performance of distributed query optimization is improved when ACO is integrated with other optimization algorithms.

Haibo Hu et.al. [8] have proposed a holistic and efficient solution that comprises a secure traversal framework and an encryption scheme based on privacy homomorphism. The framework is scalable to large datasets by leveraging an index-based approach. Based on this framework, we devise secure protocols for processing typical queries such as k-nearest-neighbor queries (kNN) on R-tree index.

Ku Rahane et. al. [9] have proposed about a framework for big data clustering which utilizes grid technology and ant-based algorithm.

Sudipto das et.al. [10] have discussed to clarify some of the critical concepts in the design space of big data and cloud computing such as: the appropriate systems for a specific set of application requirements, the research challenges in data management for the cloud, and what is novel in the cloud for database researchers.they have provided comprehensive background study of state-of-the-art systems for scalable data management and analysis

Divyakant agrawal et.al. [11] have analyzed the design choices that allowed modern scalable data management systems to achieve orders of magnitude higher levels of scalability compared to traditional databases. With this understanding, we highlight some design principles for systems providing scalable and consistent data management as a service in the cloud.

3. Big Data on Cloud

Big data mainly termed as a large and complex dataset which is difficult to handle by traditional method. To maintain such a huge amount of data so many challenges have to be faced. The whole world find many difficulties with large data sets as they mainly interact with internet. There are many benefits of using big data like history review of patients, analysis and decision making for future use etc. There are two types of big data available i.e. Operational big data and Analytical big data. Big data is mainly important for its volume and analysis.

Big data and cloud computing are like two sides of a coin. In cloud third party is used as a reservoir of data. Any kind of enterprise store their data in cloud computing. Cloud computing providers provide their facility through different models e.g. infrastructure (IaaS), Platform (PaaS) and software (SaaS). To access big data Iaas provides virtual machines, servers, storage, netwrok etc, Paas provides database, web server, development tools etc and Saas provides CRM, email, games etc.

3.1. Ant Colony Optimization

Ant colony optimization is a population based search technique implemented for the solution of combinatorial optimization problems which is inspired by the ants behavior which finds the shortest path between their nest and a food source using pheromone trails. Real ants find the shortest routes between food and nest. They never use their vision. Ants lay a chemical known as pheromone on the ground which acts as a signal to other ants. If an ant decides to follow the pheromone trail with some probability, it itself also lays more pheromone, thus reinforcing the trail. The more ants follow the trail, the pheromone deposition also increases, the more ants are to follow it. Pheromone strength decays over time and pheromone builds up on shorter path faster as it doesn't have so much time to decay, so ants start to follow it.
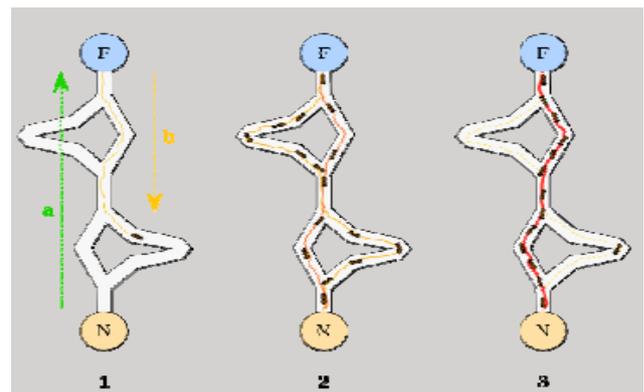
3.2. General Ant Colony Pseudo Code

Initialize the base attractiveness, T, and visibility, n, for each edge;
for i < IterationMax do:
 for each ant do:
  choose probabilistically (based on previous equation) the next state to move into;
  add that move to the tabu list for each ant ;
  repeat until each ant completed a solution ;
 end ;
 for each ant that completed a solution do:
  update attractiveness T for each edge that the ant traversed;
 end ;
 if (local best solution better than global solution)
  save local best solution as global solution ;
 end ;
end;

3.3. Experimental Analysis

Because this classic design enables ants to find their way around changes, the application of ACO to query optimization may be an alternative to existing soft computing approaches. Therefore, the ant colony optimization algorithm may be proposed to utilize an ordinal number encoding scheme to enable ants to explore the solution space by iteratively constructing solutions, partly guided by global pheromone traces signalling good solutions and partly guided by their own local cost estimations of every next join. The performance of this algorithm along with optimization of query has been assessed on a 32-bit 2.66 GHz Intel Core 2 Duo desktop computer with 2 GB physical memory.

Table 3.1. Database with records

No. of databases	Avg. Records per database
16	369
27	439
34	469
37	477

Each solution may be associated with execution costs, that may be realized by the costs of data transmission from the source to the processor and the costs of processing the data. As this work focuses on a multiple sources, the data transmission costs and data processing costs cannot be ignored. The sum of costs may be associated with all joins within a solution.

While considering the queries with heterogeneity in nature having about 15 joins, it is seen that the ant colony optimization algorithm yields better solutions. Compared to other algorithms, this approach may need approximately 80% less time to converge. As the query size increases, the differences in mean execution times tend to decrease. For queries consisting of more than 20 joins, the ant colony optimization algorithm may need about 60% less time to converge.

4. Conclusion

Today big data is in boom and handling such a large volume and variety of data is a big challenge for us. In this paper we have discussed about big data, cloud computing and ant colony optimization technique. Cloud is a platform which resides in remote location. Big data are implemented on this platform and uses its tools, softwares and hardwares for the manipulation of its data. Further in our research we are going to implement ant colony optimization technique on big data for faster manipulations of data.

References

[1] Marcos D. Assuncao, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya: "Big Data computing and clouds: Trends and future directions". (2015).

[2] Khairul Munadi, Fitri Arnia, Mohd Syaryadhi, Masaaki Fujiyoshi and Hioshi Kiya: "A Secure online image trading system for untrusted cloud environments". (2015).

[3] Ms. Rupali S. Khachane and Dr. Pradeep K. Deshmukh: "Attribute Based Secure Query Processing in Cloud with Privacy Homomorphism". (July 2015).

[4] Badrish Chandramouli, Jonathan Goldstein, Abdul Quamar: "Scalable Progressive Analytics on Big Data in the Cloud". (2013).

[5] Santoshi Tsuchiya, Yoshinori Sakamoto, Yuichi Tsuchimoto and Vivian Lee: "Big Dataa Processing in Cloud Environmenets". (2012).

[6] Divyakant Agarwal, Sudipto Das and Amr El Abbadi: "Big data and Cloud Computing : Current State and Future Opportunities". (2011).

[7] Ms. Preeti Tiwari, D. S. V Chande: "Optimization of Distributed database queries using hybrids of ant colony optimization algorithm". (2013).

[8] Haibo Hu, Jianliang Xu, Chushi Ren, Byron Choi: "Processing private queries over untrusted data clouds through privacy homomorphism". (2011).

[9] Ku Ruhana Ku-Mahamud: "BIG DATA CLUSTERING USING GRID COMPUTING AND ANTBASED ALGORITHM". (2013).

[10] D. Agarwal, S. Das and A. E. Abbadi: "Big data and Cloud Computing: New wine or just new bottles ?". (2010).

[11] D. Agarwal, A. E. Abbadi, S. Das: "Data Management Challeneges in Cloud Computing Infrastructures". (2010).