ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Techno Research Publishers, Bhopal, India.**
www.computerscijournal.org

ISSN: 0974-6471
April 2015,
Vol. 8, No. (1):
Pgs. 09-12

Big Data Solution by Divide and Conquer technique in Parallel Distribution System using Cloud Computing

RAVIKUMAR H. ROOGI

Dept of Computer Science, Karnatak University, Dharwad, India.

(Received: March 10, 2015; Accepted: April 10, 2015)

ABSTRACT

Cloud computing is a type of parallel distributed computing system that has become a frequently used computer application. Increasingly larger scale applications are generating an unprecedented amount of data. However, the increasing gap between computation and I/O capacity on High End Computing machines makes a severe bottleneck for data analysis. Big data is an emerging paradigm applied to datasets whose size or complexity is beyond the ability of commonly used computer software and hardware tools. Such datasets are often from various sources (Variety) yet unstructured such as social media, sensors, scientific applications, surveillance, video and image archives, Internet texts and documents, Internet search indexing, medical records, business transactions and web logs; and are of large size (Volume) with fast data in/out (Velocity). More importantly, big data has to be of high value (Value) and establish trust in it for business decision making (Veracity). To handle the dynamic nature of big data successfully, architectures, networks, management, mining and analysis algorithms should be scalable and extendable to accommodate the varying needs of the applications. In this paper we propose a big data solution through cloud computing by using divide and conquer technique in parallel distribution system.

Key word: Big Data, Cloud Computing, Divide and Conquer Technique.

INTRODUCTION

Big Data: Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. Big data has the potential to help companies improve operations and make faster, more intelligent decisions. While the term may seem to reference the volume of data, that isn't always the case. The term big data, especially when used by vendors, may refer to the technology (which includes tools and processes) that an organization requires handling the large amounts of data and storage facilities. The term big data is believed to have originated with Web search companies who needed to query very large distributed aggregations of loosely-structured data. An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

Cloud Computing

The term Cloud refers to a Network or Internet. In other words, we can say that Cloud is something, which is present at remote location. Cloud can provide services over network, i.e., on public networks or on private networks, i.e., WAN, LAN or VPN. Applications such as e-mail, web conferencing, customer relationship management (CRM), all run in cloud. Cloud Computing refers to manipulating, configuring, and accessing the applications online. It offers online data storage, infrastructure and application.

Benefits

Cloud Computing has numerous advantages. Some of them are listed below:

1. One can access applications as utilities, over the Internet.
2. Manipulate and configure the application online at any time.
3. It does not require to install a specific piece of software to access or manipulate cloud application.
4. Cloud Computing offers online development and deployment tools, programming runtime environment through Platform as a Service model.
5. Cloud resources are available over the network in a manner that provides platform independent access to any type of clients.
6. Cloud Computing offers on-demand self-service. The resources can be used without interaction with cloud service provider.
7. Cloud Computing is highly cost effective because it operates at higher efficiencies with greater utilization. It just requires an Internet connection.
8. Cloud Computing offers load balancing that makes it more reliable.

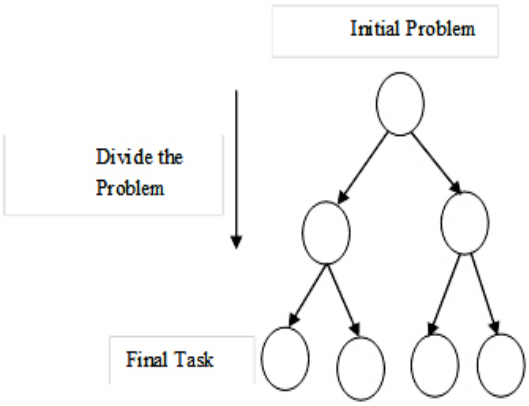

Fig. 1: Tree Construction

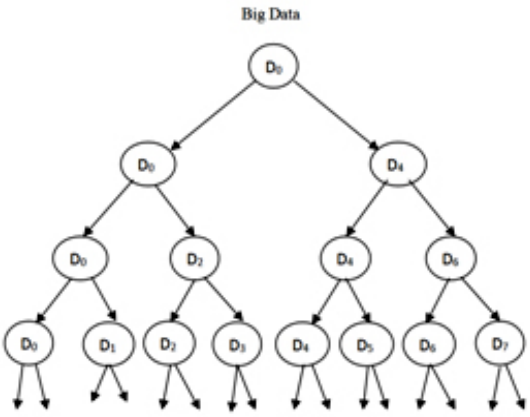

Fig. 2: Dividing Data into Small Data

Why is parallel computing important?

1. We can justify the importance of parallel computing for two reasons.
   - Very large application domains, and
   - Physical limitations of VLSI circuits
2. Though computers are getting faster and faster, user demands for solving very large problems is growing at a still faster rate.
3. Some examples include weather forecasting, simulation of protein folding, computational physics etc

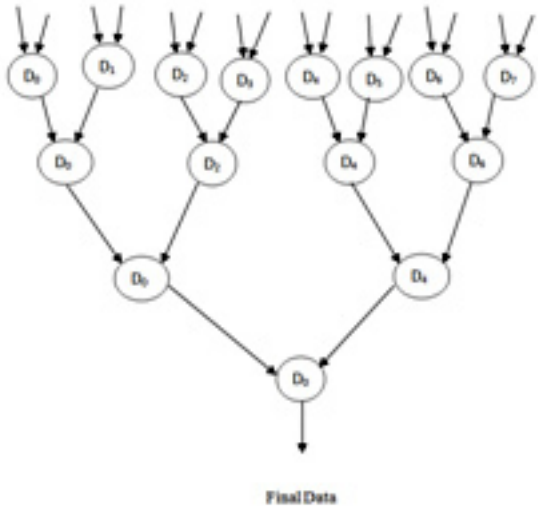

Fig. 3: Partial Summation

Fig. 4: Big Data solution through Cloud in Parallel Distribution System

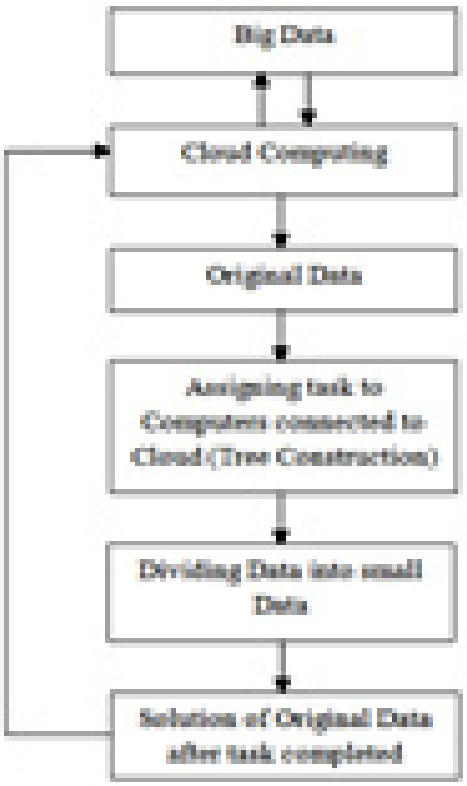

Fig. 5: Flow Chart of Divide and Conquer Technique in Parallel Computer

representing the solution for big data through Cloud Computing by using divide and conquer technique because cloud computing is a type of parallel distributed computing system that has become a frequently used computer application. It will process different computers parallely then returns the solution to the Cloud. Now a day's data is growing rapidly in each and every field, solution for this Big Data is very difficult. We don't know where the server is placed in geographical area which collects all the data for solution. Collecting this data for task completion is difficult so we are using cloud technique because through cloud we can collect all the data stored in server and distribute it to systems in different geographical areas for solution through networks connected to cloud for solution and collect it back after task completion of original data. Data is secured in Cloud and here easily we can encrypt and decrypt the data.

Algorithm

1. Instance of problem Data D_0.
2. Collecting Data D_0 to Cloud.
3. Assigning instance of problem Original Data to different systems in different geographical area through Cloud.
4. D is divided into $D_1, D_2, D_3, \text{——}, D_n$ Assign instance D_1 to computer 1

Assign instance D_2 to computer 2
Assign instance D_3 to computer 3
——
——
——
Assign instance D_n to computer n

5. Return the solution back to Cloud.
6. Solution for the problem D_0.

Analysis

In designing a parallel algorithm, it is more important to make it efficient than to make it asymptotically fast. The efficiency of an algorithm is determined by the total number of operations, or *work* that it performs. On a sequential machine, an algorithm's work is the same as its time. On a parallel machine, the work is simply the processor-time product. Hence, an algorithm that takes time t on a P-processor machine performs work $W = Pt$. In either case, the work roughly captures the actual cost to perform the computation, assuming that the cost of a parallel machine is proportional to the number of processors in the machine. We call an algorithm *work-efficient* (or just efficient) if it performs the same amount of work, to within a constant factor, as the fastest known sequential algorithm. For example, a parallel algorithm that sorts n keys in $O(\sqrt{n}\log n)$ time using $\sqrt{n}$ processors is efficient since the work, $O(n\log n)$, is as good as any (comparison-based) sequential algorithm. However, a sorting algorithm that runs in $O(\log n)$ time using n^2 processors is not efficient. The first algorithm is better than the second - even though it is slower - because it's work, or cost, is smaller. Of course, given two parallel algorithms that perform the same amount of work, the faster one is generally better.

REFERENCES

1. S P Sajjan, Jitendra Rohilla, Vijakumar Badiger, Girish Badiger, Shivaraj Angadi. "Implementing Divide and Conquer Technique in Parallel Computer Using Network." *international journal of electronics & communication technology* 5.1: 1 (2014).
2. Baidari, Ishwar, SP Sajjan, and G. Vijaykumar. "Implementing Divide and Conquer Technique for a Big-Data Traffic." (2014).
3. http://www.cs.cmu.edu/~scandal/html-papers/short/short.html.
4. Cloud Computingtutorial *by tutorialspoint. com*
5. Ellis Horowitz, Sartaj Sahni, Sanguthevar Rajasekaran: Fundamental of computer algorithms, 2nd Edition, University press, (2007).
6. A.M. Padma reddy Desgin and analysis of algorithms 6th Edition, (2012).