ISSN 2229-3531 (Print)

JUCIT Vol. 8(3), 20-25 (2017). Periodicity-2-Monthly

(Online) ISSN 2455-9997

Estd. 2010

JOURNAL OF ULTRA COMPUTER & INFORMATION TECHNOLOGY

An International Open Free Access Peer Reviewed Research Journal of Computer Science Engineering & Information Technology

website:- www.compitjournal.org

Data Aggregation in Cloud Using Map Reduce Framwork

SUCHI SINGH PARIHAR[1] and M.A. RIZVI[2]

[1]Department of Computer Technology and Application
National Institute of Technical Teacher's Training and Research Bhopal (India)
[2]Head of Computer Engineering and Applications
National Institute of Technical Teacher's Training and Research Bhopal (India)
Under Ministry of Human Resource and Development
Email of Corresponding Author :- singhsuchi71@gmail.com

http://dx.doi.org/10.22147/jucit/080301

Acceptance Date 8th June, 2017, Online Publication Date 20th June, 2017

Abstract

Cloud computing is one of the important component of today's tech savvy society.it creates a new paradigm for information exchange without investing in a new infrastructure or licensing new software. It expands the capabilities of existing traditional way of accessing the application software, system software, storage and others through internet. In the last few years, cloud computing has grown tremendously from a promising business concept to fastest growing segments of the IT industry.

At a remarkable pace, cloud computing has transformed how business and government function by providing various services like IaaS, PaaS and SaaS. Huge amount of data is being generated by various applications on the cloud called big data applications. Big data applications need appropriate framework and techniques to store, aggregate and retrieve the data.

Consequently the objective of this research paper has divided into two sections; firstly to identifying the traditional methods of data aggregation and optimization, then evaluate aggregation of data through minimization of links association between Mapper and Reducer to reduce the data traffic in the network.

Secondly to present a viable solution to overcome these major problems using high level scripting language pig on Apache Hadoop Framework.

This paper proposes data aggregation and optimization for big data applications within a cloud environment.

Key words : Data Aggregation, Big Data , Cloud Computing

Abbreviations –
IaaS- Infrastructure as a Service
Paas- Plateform as a Service
SaaS- Software as a Service

I. Introduction

Traditional application integration technologies are performed in a rigid and slow process that usually takes a long time to build and deploy, requiring professional developers and domain experts. Since the face of the Internet is continually changing, as new services and novel applications appear and become globally noteworthy at an increasing pace. Now a days the locus of computation is changing, with functions migrating to remote datacenters via Internet based communication.

Cloud computing is the idea that data and programs can be stored centrally, in the cloud, and accessed anytime from anywhere through thin clients and lightweight mobile devices. This brings many advantages, including data ubiquity, ûexibility of access, and resilience. The services of cloud computing are broadly divided into three categories: Infrastructure-as-a-Service(IaaS),Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) Cloud computing also is divided into five layers including clients, applications, platform, infrastructure and servers. In this paper I focus on the typical application of cloud computing, comparison of various cloud computing models, cloud computing characteristics, security challenges, review the several cloud deployment and service models. It also explore certain benefits of cloud computing over traditional IT service environment -including scalability, flexibility, reduced capital and higher resource utilization -are considered as adoption reasons for cloud computing environment. I also include security, privacy, internet dependency and availability as avoidance issues. Finally, I concludes the paper.

II. Typical Application of Cloud Computing

Cloud computing has very wide area of applications. The main objective behind any application[1] of this domain is to provide seamless connectivity from multiple locations and on multiple different devices. A 2009 survey from Evans Data shows that 40 percent of developers working on open source plan to deliver their applications as web services offerings using cloud providers. Few of them are listed below.

*A. **Google App Engine :***

It is one of the perfect tool which offers servers, load balancers, or DNS tables to get an app on the cloud. The database is integrated well with open source programming languages like python and others. It has various features which helps developers to build an app without much difficulty.

*B. **Force.com and Google :***

Force.com for Google App engine is a set of tools and services to enable developer success with application development in the cloud. This helps to create entirely new web and business applications like social networks *i.e.* linkedin.com and facebook.com

*C. **Google Gears :***

An open source technology for creating offline web applications. It is a single standard for offline applications. Google offers Google Gears as a free, fully open source technology in order to help every web application not just Google applications.

*D. **Microsoft Azure services :***

It is a tool provided for developers who want to write applications that are going to run partially or entirely in a remote datacenter. It is an Internet scale cloud services platform hosted in Microsoft datacenters, which provides an operating system and a set of developer services.

*E. **Bungee Connect :***

It provides development, testing, deployment and hosting in a single on-demand platform. Delivered entirely via browser with no download or plug-in for developers without compromising accessibility and security.

All the applications of cloud are categorized as cloud computing services i.e. Infrastructure as a Service, Platform as a Service and Software as a Service.

Table 1. Cloud Computing Services

S.	Cloud Computing Services and Examples		
No.	*Services*	*Few Examples*	
1	IaaS	Amazon Elastic Compute Cloud (EC2)	GoGrid
2	PaaS	Right Scale	Salesforce.com
3	SaaS	Dropbox	Microsoft Office 365

III. Comparison of Cloud Computing Model :

Four deployment models[2] have been identified for cloud architecture solutions, described below:

- **Public cloud**. This cloud is basically used to offer services for general public users or large group of users which may belong to one industry and is owned by an organization like Google which is going to provide services based on the requirements, demands of general public.
- **Private cloud**. It is completely isolated cloud which provides its unique services to the employees of particular organization that may exist on premise or off premise. This cloud is accessible only from private users from within the organization. It is managed by the organization or a third party.
- **Community cloud.** When the cloud infrastructure is shared by several organizations those are having common mission, interest, security requirements and policy. It is managed by organizations or third party.

- **Hybrid cloud**. It is a combination of two or more public, private or community cloud and offer their services to them as and when needed.

Cloud Computing Services

S. No.	Cloud Computing Models and Features	
	Models	*Features*
1	Public	Available to the general public
2	Private	Available only within an organization
3	Community	Available to common interest organizations
4	Hybrid	Non critical activities for public and private cloud.

IV. Cloud Computing Characteristics :

Characteristics of cloud computing depends upon the type of cloud and the deployment models used by the organization. Five essential characteristics[3] offered to businesses today are as follows:

A. *On-demand capabilities :*

On-demand (OD) computing is an increasingly popular enterprise model in which computing resources are made available to the user as needed. The resources may be maintained within the user's enterprise, or made available by a service provider. The on-demand model was developed to overcome the common challenge to an enterprise of being able to meet fluctuating demands efficiently.

B. *Broad Network access :*

Broad network access refers to resources hosted in a private cloud network (operated within a company's firewall) that are available for access from a wide range of devices, such as tablets, PCs, Macs and smartphones. These resources are also accessible from a wide range of locations that offer online access.

C. *Resource Pooling :*

Multi-tenants environments where multiple customers share adjacent resources in the cloud with their peers and creates a pool of multiple resources to provide different kind of services like computing ,networks and storage services and offers software solutions also reduces the operational cost of these resources is called resource pooling.

D. *Rapid Elasticity :*

Elastic computing is one of the critical characteristics which fulfill the immediate requirements of any business with minimum delay. An organization can easily add or remove users, software features and other resources. Amazon named their cloud platform "Elastic Compute Cloud" ("EC2").

E. *Measured Service*

All the services which is offered by polled resources are measured on run time basis .Depends upon the usage it is measured and bill is generated. It indicates visibility and transparency to consumption rates and costs. It helps cross-departmental reporting and budgeting.

V. Big Data Challenges :

Cloud computing can provide infinite computing resources on demand due to its high scalability in nature, which eliminates the needs for Cloud service[4] providers to plan far ahead on hardware provisioning. Big companies such as Google, Microsoft and Amazon rapidly involved in developing cloud computing applications and add more functionality into it to cater large amount of users. Cloud Computing has already started to revolutionize the way we store and access data. There are various types of issues raised when we discuss about data aggregation in cloud *i.e.*

Capturing data, curation, storage, searching, sharing, transfer, analysis and presentation. **To solve the above challenges we have two approaches.**

A. *Traditional Approach :*

In this approach, an enterprise will have a computer to store and process big data. Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated software's can be written to interact with the database, process the required data and present it to the users for analysis purpose.

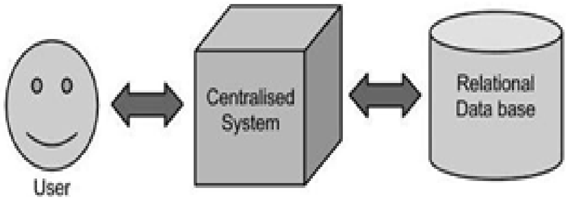


Figure 1 Traditional Approach

B. *Modern Approach :*

Google solved this problem using an algorithm called Map Reduce. This algorithm divides the task into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result dataset

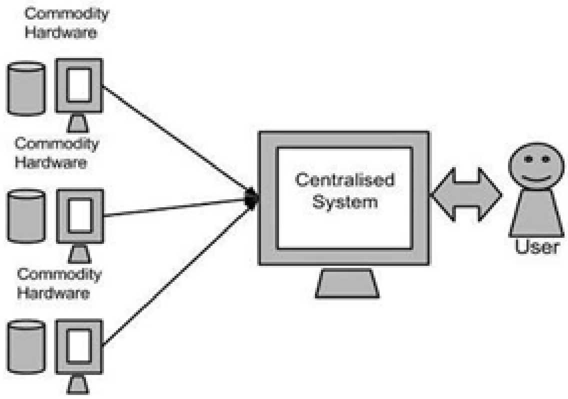


Figure 2 Modern Approach

Cloud computing has ‘‘unique attributes that require risk assessment in areas such as availability and reliability issues, data integrity, recovery, and privacy and auditing’’[5].

VI. Data aggregation :

Data aggregation is a type of data and information mining process where data is searched, gathered and presented in a report-based, summarized format to achieve specific business objectives or processes and/or conduct human analysis. Data aggregation may be performed manually or through specialized software. Data aggregation is a component of business intelligence (BI) solutions. Data aggregation personnel or software search databases find relevant search query data and present data findings in a summarized format that is meaningful and useful for the end user or application.

Data aggregation generally works on big data or data marts that do not provide much information value as a whole. Data aggregation’s key applications are the gathering, utilization and presentation of data that is available and present on the global Internet.[6]

Data aggregation is the process of collecting and aggregating the useful data. Data aggregation is considered as one of the fundamental processing procedures for saving the energy.

Two kinds of data aggregation schemes, intra-machine and inter-machine, which are elaborated in the following.-

A. Intra-Machine Data Aggregation :

The most straightforward way to reduce data traffic is to aggregate the same key-value pairs generated by map tasks within the same machine before they are sent over the network where an aggregator is created to merge the intermediate results generated by each map task.

B. Inter-Machine Data Aggregation :

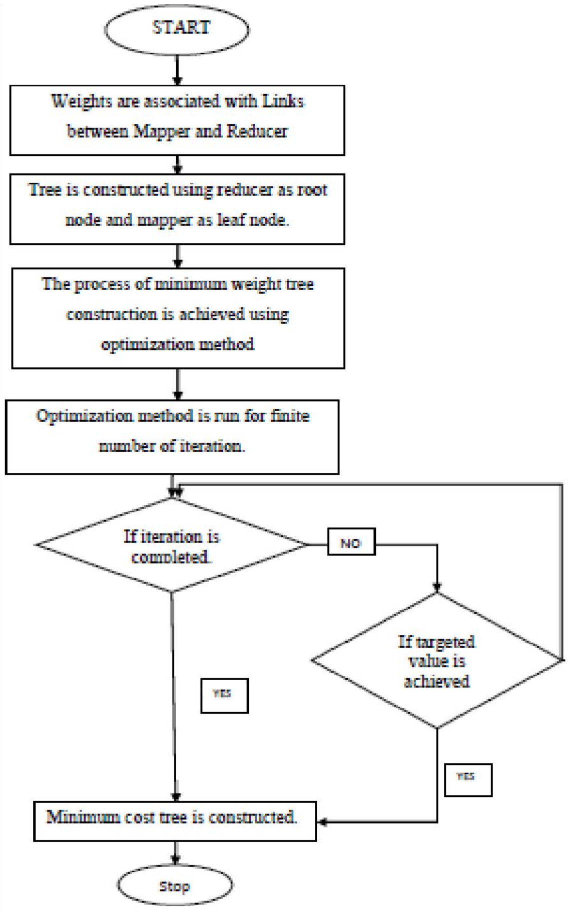
In addition to intra-machine data aggregation, we can further reduce the data traffic by aggregating the intermediate results from different machines, referred to as inter-machine data aggregation.

That the selection of nodes conducting inter machine data aggregation[7] can affect the performance, and thus becomes an additional challenge we need to handle.

VII. Proposed Work :

The objectives of work idea are as follows:

1. Modify the existing data aggregation approach in Map-Reduce frame-work.
2. Analyse the proposed approach for required environment.
3. Compare the performance of proposed approach with existing approach on various parameters.
4. Verify the robustness of proposed approach is better than existing approach.


START
Weights are associated with Links between Mapper and Reducer
Tree is constructed using reducer as root node and mapper as leaf node.
The process of minimum weight tree construction is achieved using optimization method
Optimization method is run for finite number of iteration.
If iteration is completed.
NO
If targeted value is achieved
YES
YES
Minimum cost tree is constructed.
Stop

In this paper, we investigate the importance of current cloud computing big data applications. Based on the investigation data aggregation concerns needed by companies nowadays are very important and consequently result in a big obstacle for users to adapt into the cloud computing systems. Hence, more concerns on various aspects of retrieval of accurate data such as confidentiality, integrity, authentication, non-repudiation, control and audit and so on should be taken into account.

Cloud computing is a cluster[8] of diverse key technologies that have evolved and matured over the years. It has immense potential for cost savings to the enterprises but at the same time security risk are more.

This paper discussed the applications and popular models of cloud computing. It also addressed challenges and issues of cloud computing in detail. Identified challenges of big data applications, aggregation problem that need to be overcome by adopting proposed method.

We propose to adopt a novel approach in tackling the data aggregation problem of big data application.

The proposed system is the implementation of an algorithm for big data application for aggregation of data through minimization of links association between Mapper and Reducer to reduce the data traffic in the network. Simulation is performed using Hadoop framework a simulator which provides the environment for big data application. Configuration and programming is done through high level scripting language pig on Apache Hadoop Framework.

VIII. Result Analysis :

The result of my work based on the objective prescribed is achieved by applying data aggregator approach in Map Reduce Framework along with stated optimization technique and minimum cost tree is constructed.

Parameters	Before Aggregation	After Aggregation & Optimization
Data Size	29.6 MB	4 KB
No of Rows	90000	151
Processing Speed	Slow	Fast
Memory Usage	Huge	Less
Complexity	High	Low

Inspite of the several limitations and the need for better methodologies processes, cloud computing is becoming a hugely attractive paradigm [9], especially for large enterprises. Cloud Computing initiatives could affect the enterprises within two to three years as it has the potential to significantly change IT.

IX. Conclusion

It has been observed that after applying MAP functions FOREACH,FILTER and FLATTEN and Reduce functions GROUP,COROUP,JOIN,DISTINCT, data is aggregated and meaningful data is achieved for further analysis. As you have seen the data size is reduced from 29.6 MB to 4 KB.The other parameters like, memory usage will become less and processing speed will become fast. Complexity is also reduced.The approach used in this paper is one of the viable solution for data aggregation purpose.

Acknowledgement

I would like to express my gratitude Dr. M. A. Rizvi **,** Professor and Head Department of Department of Computer Engineering & Applications, National Institute of Technical Teacher’s Training and Research, Bhopal who helped me during the period of entire course including dissertation work and papes.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Santosh Kumar and R. H. Goudar, “Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications: A Survey”, International Journal of Future Computer and Communication, Vol. 1, No. 4, December (2012).
2. Pradeep Kumar Tiwari, Dr. Bharat Mishra, “Cloud Computing Security Issues, Challenges and Solution”, International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 8, August (2012).
3. Rabi Prasad Padhy, Manas Ranjan Patra, Suresh Chandra Satapathy, “Cloud Computing: Security Issues and Research Challenges, International Journal of Computer Science and Information Technology & Security, Vol. 1, No. 2, December 2011.
4. Tanvi Agrawal, S. K .Singh, “Analysis of Security Algorithms in Cloud Computing”, International Conference on Computing for Sustainable Global Development (INDIACom), IEEE (2016).
5. Komal Gandhi, Dr. Parul Gandhi, “Cloud Computing

Security Issues: An Analysis", International Conference on Computing for Sustainable Global Development (INDIACom), IEEE (2016).

6. Vahid, Ashktorab and Seyed Reza Taghizadeh. "Security threats and countermeasures in cloud computing." International Journal of Application or Innovation in Engineering & Management (IJAIEM) 1.2, 2012 pp 234-245.
7. Tsai Chang-Lung and Uei-Chin Lin. "Information security of cloud computing for enterprises." Advances in Information Sciences and Service Sciences 3.1, pp 132 (2011).
8. Rajani Sharma, Rajender Kumar Trivedi, "Literature review: Cloud Computing Security Issues, Solution and Technologies, International Journal of Engineering Research, Vo lu me No.3, Issue No.4, April, pp : 221 (2014).
9. Suruchee V. Nandgaonkar, Prof. A. B. Raut, "A Comprehensive Study on Cloud Computing", International Journal of Computer Science and Mobile Computing, Vol. 3, Issue. 4, April, pg.733 – 738 (2014).