International Journal of Grid and Distributed Computing
Vol. 10, No. 7 (2017), pp.1-10
http://dx.doi.org/10.14257/ijgdc.2017.10.7.01

Enhancement of Data Storage and Retrieval by Classification using Probability Distribution Function

Debabrata Sarddar[1], *Sougata Chakraborty[2] and Priyajit Sen[3]

[1]*Assistant Professor, Department of Computer Science & Engineering, University of Kalyani, West Bengal, India*

[2]*Senior System Engineer at IBM India Private Limited, Kolkata, West Bengal, India*

[3]*Student of Master of Technology, Department of Computer Science & Engineering, University of Kalyani, West Bengal, India*
[1]*dsarddar1@gmail.com*, [2]*me.sougata.chakraborty@gmail.com*,
[3]*priyajit91@gmail.com*

Abstract

Today's most challenging concern is the data storage and retrieval in the cloud server. In our daily life, we need to upload and download data in the cloud server but sometimes, the speed of uploading and downloading hamper the performance of the cloud server. Hence, the speed of searching any content in the cloud server can be enhanced by adopting modern technologies. In this paper, we have tried to solve the problem of data storage and retrieval with the existing system and enhance the speed also. We have proposed a new technique "Enhancement of data storage and retrieval by classification using probability distribution function" in which data will be stored in the cloud storage server by classifying them as per their types and at the same time whenever we need to search data, we will go to the specific segment of the cloud server by checking its type so that it will increase the speed of searching.

Keywords*: Data storage, Probability distribution function, Bayes' theorem, Amdahl's law, Single server, classified server.*

1. Introduction

Cloud storage is a cloud computing model in which data is stored by the cloud user and the service providers to get efficient and any time access over the internet. Data is stored on the storage server using virtualization techniques. Cloud server can be of several types:

1.1. Public Cloud

In public cloud, data is made available to the common people by the service provider. Sometimes, data is free and pay per use.

1.2. Private Cloud

In private cloud, data is secured for a person or an individual privately. Sometimes, it is dedicated to a single organization also. It provides self-service mechanism.

1.3. Hybrid Cloud

Hybrid cloud provides the mix architecture of public and private cloud. It is an integrated infrastructure. Hybrid cloud offers scalability, cost efficiency, security and flexibility.

ISSN: 2005-4262 IJGDC

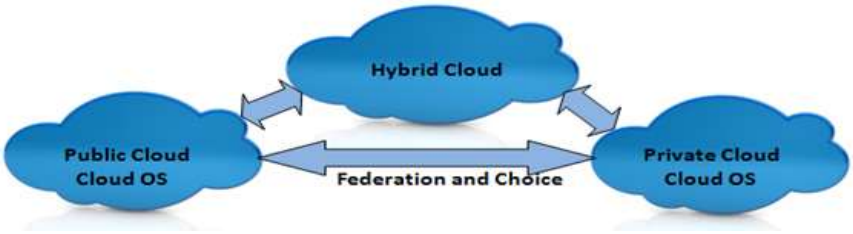

Figure 1. Public, Private and Hybrid Cloud Infrastructure

The cloud server provides reliability and security over the data. Sometimes, the data owners are not sure about the safety of data. Data is kept secure in the cloud server by encryption. Encryption is used to change the form of the data so that no one can steal or read data directly. The eavesdroppers are also unable to decrypt the encrypted data. Two ways are available by which data can be kept secure in the cloud server:

1.4. Authentication

In Authentication, the secure user id and its password are required to access the data. For example, if we need to access data from drop box account or store data into it then we need to login to the system first and thereafter we can perform the needful work.

1.5. Authorization

Sometimes, the data owner gives authorization to other users to access the data. Prior to access the data, authorization process must be checked.

Data store in cloud network is rapid and continuously being done worldwide. Data is stored in the server virtually so that no one can understand the physical location of the data where it is being stored but the efficient access to data is possible from anywhere in the globe. Therefore, the clients of cloud network are the responsible for the management of data practically and they can provide additional services of cloud network [5].

2. Characteristics of Cloud Server

1. Cloud storage server can manage a system having minimum number of resources.
2. There is some protocol of storing and retrieving data in the cloud server. The protocol through which cloud server is exposed is also present in the system.
3. Cloud server shows multi user characteristics. The cloud server can meet the higher demands or loads.
4. Data should be available all the time so that whenever it is required, it can be produced in front of the user.
5. The cloud server can control the overall system and improve performance.

3. Related Work

In the paper, "Research on Cloud Data Storage Technology and Its Architecture Implementation", Kun Liu *et al*. proposed the idea of cloud storage as well as its architecture firstly.

In the paper, "Creating optimal cloud storage systems", Josef Spillner *et al*. proposed cloud storage management system to achieve optimality in cloud storage services.

In the paper, "An Efficient Cloud Storage Model for Heterogeneous Cloud Infrastructures", Dejun Wang proposed an efficient cloud storage model for the heterogeneous cloud infrastructures.

In the paper, "Building a Cloud Storage Service System", Chengzhang Peng *et al*. proposed a solution to build a Cloud Storage Service System based on the open-source distributed database.

In the paper, "Cloud Storage Standards Overview and Research Ideas Brainstorm", Mark Carlson proposed the overview of cloud storage, architecture, cloud data management, and cloud peering.

In the paper, "TCLOUD: A Trusted Storage Architecture for Cloud Computing", Sultan Ullah *et al*. have proposed a trusted architecture of cloud data storage and data retrieval in a unique way from the cloud data center.

In the paper, "Cloud Storage Reference Model for Cloud Computing", Pravin O. Balbudhe have proposed a cloud storage architecture.

In the paper, "A Survey on Data Storage and Security in Cloud Computing", V. Spoorthy has proposed different aspects of cloud data storage.

In the paper, "Hybrid Approach for Cloud Storage with Attribute based Encryption", Shivam Patole *et al*. have proposed homomorphic encryption using improved KP-ABE system to achieve data confidentiality.

In the paper, "Data Exchange Service using Google Drive API", Risky CahyaDinatha *et al*. have proposed Google Drive cloud storage as an intermediate data storage media.

In the paper, "Multimedia Data in Cloud Storage System using DE Duplication Verification", Kotte Ajay *et al*. have proposed a hybrid cloud architecture.

In the paper, "A Review on Cloud Computing Architectures", Sreelakshmi P. S. *et al*. have discussed about centralized cloud, Federated Cloud and P2P cloud, their advantages and limitations.

In the paper, "An Efficient and Trusted Data Storage Process for Cloud Computing", Kalyan Singh Meena has proposed a data storage mechanism to achieve data confidentiality in cloud computing.

In the paper, "A survey on cost effective multi-cloud storage in cloud computing", Nitesh Shrivastava *et al*. have surveyed over the multi-cloud storage in cloud computing.

In the paper, "A Survey of Public Auditing for Secure Data Storage in Cloud Computing", Wei-Fu Hsien, *et al*. have proposed a survey about the public auditing for Secure Data Storage.

In the paper, "Cloud Computing Security - Data Storage and Transmission", Mrs. C. Theebendra *et al*. have focused on cloud data storage and transmission security.

4. Proposed Work

We are going to propose a method in which data in cloud network is stored in the storage server by classifying the data as per the type of the data. Users of cloud network may upload and download valuable information anytime. The cloud network should be flexible that every time users can access the information, they may add some information and delete also. They can fetch required information. The overall speed of searching the valuable information can be increased by changing the style of searching the information. In this case, the only thing we must do is that, we must classify the data as per the type while uploading the data in the server. Data can be of many types like as image, text, audio and video. The process is described in the flowchart given below.

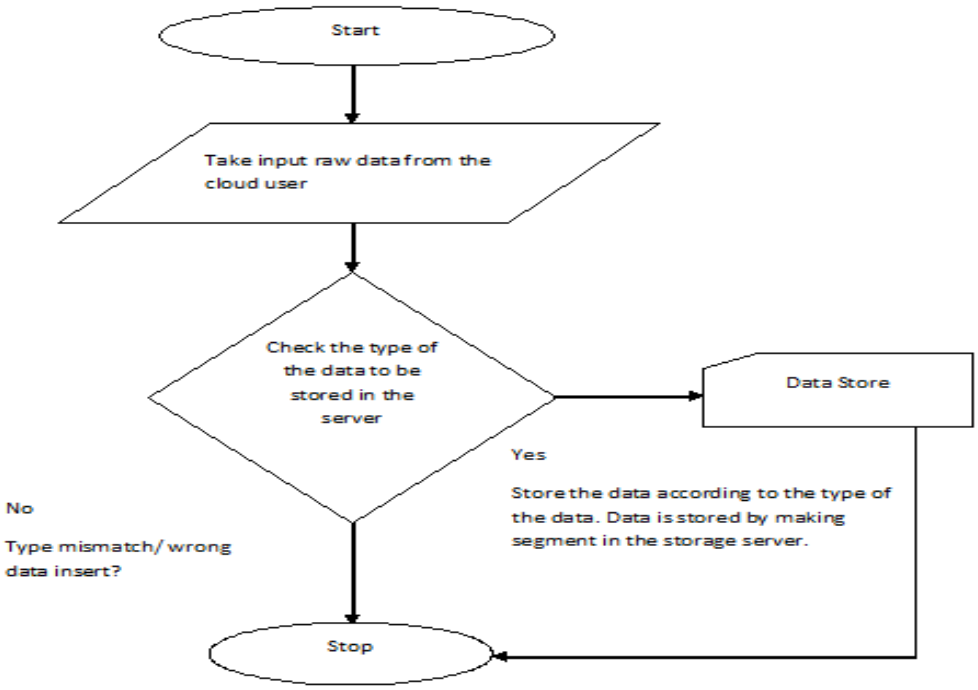

Figure 2. Flowchart of Data Store in the Cloud Server

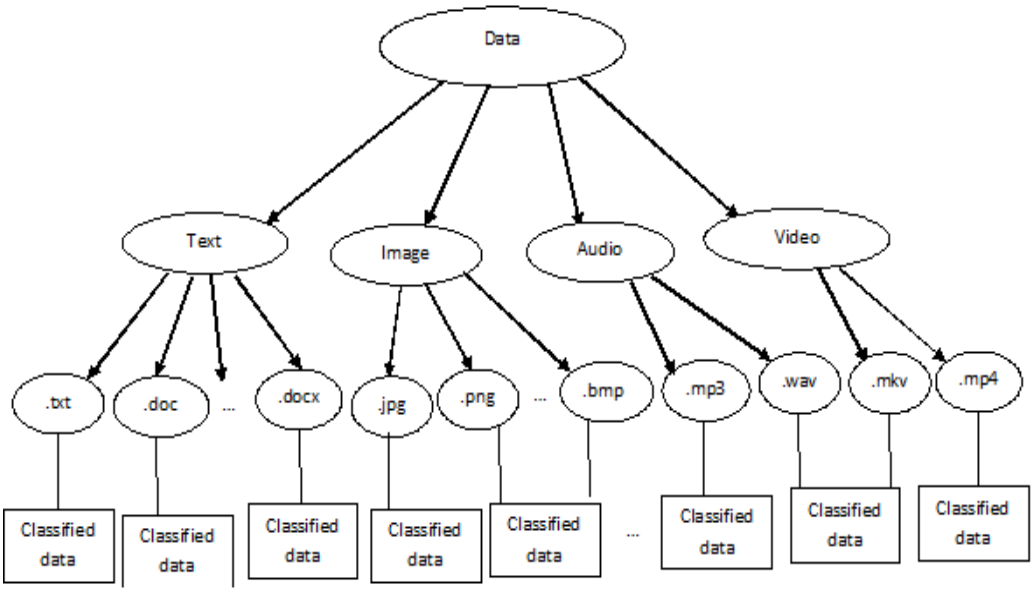

Figure 3. Hierarchy of Storing Data in Cloud Server

We know that, the cloud storage server has a maximum capacity to store data. A random variable X is continuous if, possible values of X comprise either a single interval on the number line (for some $A<B$, any number x between A and B is a possible value) or a union of disjoint intervals. Let $f(x)$ be the distribution function for any two numbers a and b with $a \le b$. a is the lower limit of storage and b is the upper limit of storage.

$$P(a \le X \le b) = \int_a^b f(x)dx$$

$$= \int_a^p f(x)dx + \int_p^q f(x)dx + \int_q^r f(x)dx + \int_r^b f(x)dx$$

The interval is from *a* to *p* is suppose for text data, similarly *p* to *q* is for image data, *q* to *r* is for audio data, *r* to *b* is for video data that is classified by classifying the store function *f(x).*

Here we will go for the conditional probability, which will specify the occurrence of new data to be stored in the cloud server with some data is previously stored in the server. Bayes' theorem suitably explains this matter,

Let *A* and *B* two events and *S* be the sample space of cloud server where all data is stored. S = *{data1, data2, data3, data4, data5}*. Suppose, *data1* is audio, *data2* and *data3* are text, *data4* and *data5* are audio. Also, consider that the extension of *data1* is uniquely different from *data4* and *data5*. Let *A* be the event in which *data1* appears during new data store and *B* is event in which an audio data occurs.

Now,

$$P(A) = \frac{1}{5}$$

And

$$P(B) = \frac{3}{5}$$

The probability of *data1* appears among audio data,

$$P\left(\frac{A}{B}\right) = \frac{1}{3}$$

Therefore, by Bayes' theorem, the probability of occurrence of event *B* over *A* is

$$P(B/A) = \frac{P(A/B).P(A)}{P(B)}$$

We need to go for a new search process where the user will specify the extension of the data that he/she wants to fetch from the cloud server. A composite query will be there to find the data of that extension.

Algorithm:

1. Input data with extension (.txt, .doc, .docx etc.) from the user.
2. Break the input line to get the extension, where extensions are some keywords that will be stored previously in the cloud server.
3. Run a query to find the data of that extension.
4. Another query will be executed as per the type of the data as the extension specifies the type of the data.

5. Finally, data is obtained correctly of the extension and type.
6. End

Search Query execution is given below:

$$\Pi_{data}\left(\sigma_{type='typ'}\left(\sigma_{extension='ext'}\right)^{extension\ INNERJOIN^{type}}\right)$$

The above query explains that how data is searched from a data set by running a composite query over the data set. Let us consider, we are going to classify our main storage server into four distinct types such as audio, text, image and video. These classification details are stored in type table and each type data has several extensions and for that reason we need to store their details within another table as per the extension found in the data. Now, we sample run a composite query in which INNERJOIN is performed over 'type' and 'extension' to get the overall data. The inner query first finds the data as per the extension specified by the user and the outer query will select the type of the data. Thus, information retrieval will be faster.

The SEO which stands for Search Engine Optimization can optimize while we are searching for a website. When the selection of the type and extension will be done then SEO can be used for searching the information from the selected server. Search engine performs a few distinct functions when it searches content. Those are,

Crawling: This is a process to find the related websites of a website with the help of web crawler or spider that crawl over the web pages to perform the same.

Indexing: With this process, web pages are fetched as per the index by locating the words, expressions and keywords in the giant database.

Processing: The searching engine compares data in the search result versus data in the search request with the help of index page table.

Finding Relevancy: Sometimes, many pages contain the same data, so search engine performs relevancy checking with each pages of the index table.

Retrieving Result: In this step, the search engine retrieves the best matched result and shows them in front of the users.

We are going to describe the concept of parallel and distributed computing as our concept suggests the concept of parallel storing and fetching of data in the cloud server. Parallel computing allows us to solve large problems by splitting them into smaller ones and solving them concurrently. Parallel systems allow us to solve problems by engaging more resources or classifying the single system and reduce the time required to store and fetch data. The Speed up measures the effectiveness of parallel computing. Let us consider, *T(1)* is the execution time for single server and *T(N)* is execution time for classified server. Therefore, Speedup is calculated as,

$$S(N)=\frac{T(1)}{T(N)}$$

Amdahl's Law finds the potential speedup of a parallel computation. By this law, the portion of the computation which cannot be parallelized determines the overall speed-up.

If α is the fraction of running time a sequential program spends on non-parallelizable segments of the computation then,

$$S = \frac{1}{\alpha}$$

5. Simulation Result

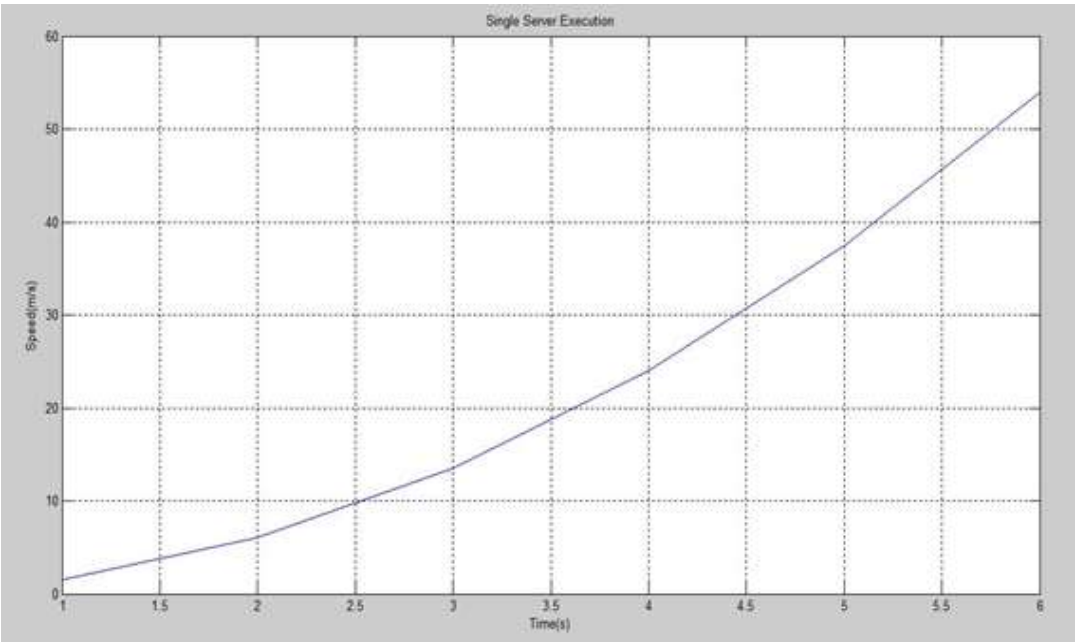

Figure 4. Single Server Execution Graph: Speedup Vs Time

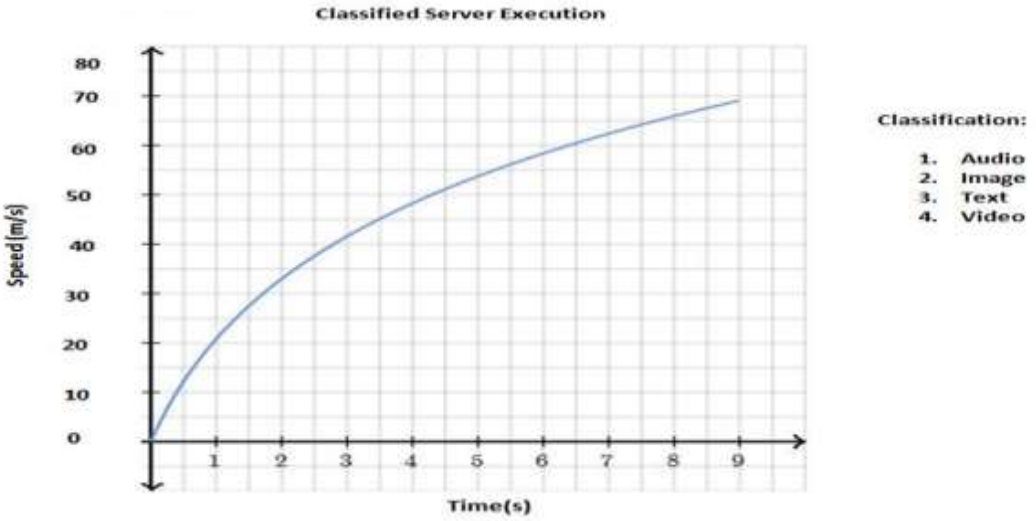

Figure 5. Classified Server Execution Table: Speedup Vs Time

Table 1. Speedup for Single Server Execution

Speedup for single server execution	
Time	**Speedup**
0 s	1 m/s
1.5 s	4 m/s
2.5 s	10 m/s
3.5 s	19 m/s
4.5 s	31 m/s
5.5 s	46 m/s
	Average Speedup=111/6=18.5

Table 2. Speedup for Classified Server Execution

Speedup for classified server execution	
Time	**Speedup**
0 s	0 m/s
1.5 s	25 m/s
2.5 s	35 m/s
3.5 s	45 m/s
4.5 s	51 m/s
5.5 s	58 m/s
	Average Speedup=214/6=35.67

6. Conclusion

The proposed concept shows the relatively newer approach of data storage and retrieval in the cloud server. The implementation result is showing the difference between average speed up of the single server and classified server execution. Here, we can see the average speed of single server execution is 18.5 whereas it is 35.67 in classified server execution. Therefore, the proposed concept of storing the data separately within classified server and retrieving it in a faster way is more feasible than existing approach. In this case, we need to keep it in our mind that, data must be stored separately to adopt the proposed technique of data retrieval. The future scope of our research is to improve the proposed scheme by engaging the newer dimension in the data storage technique more elaborately so that the new data storage technique may work faster and be potentially efficient. The distributed and parallel execution concept will be discussed more specifically then.

References

[1] K. Liu and L. J. Dong, "Research on Cloud Data Storage Technology and Its Architecture Implementation", International Workshop on Information and Electronics Engineering (IWIEE), SciVerse Science Direct, Elsevier, Procedia Engineering, vol. 29, **(2012)**, pp. 133-137, 2012.
[2] J. Spillner, J. Müller and A. Schill, "Creating optimal cloud storage systems", Future Generation Computer Systems, vol. 29, **(2013)**, pp. 1062–1072.
[3] D. Wang, "An Efficient Cloud Storage Model for Heterogeneous Cloud Infrastructures", Procedia Engineering, vol. 23, **(2011)**, pp. 510-515.
[4] C. Peng and Z. Jiang, "Building a Cloud Storage Service System", 3rd International Conference on Environmental Science and Information Application Technology (ESIAT 2011), Elsevier, Procedia Environmental Sciences vol. 10, **(2011)**, pp. 691-696.
[5] M. Carlson, "Cloud Storage Standards Overview and Research Ideas Brainstorm", SNIA TC and Sun Chair, SNIA Cloud Storage TWG CMU SDI Lecture, **(2009)**.
[6] S. Ullah and Z. Xuefeng, "TCLOUD: A Trusted Storage Architecture for Cloud Computing", International Journal of Advanced Science and Technology, vol.63, **(2014)**, pp.65-72.
[7] P. O. Balbudhe and P. O. Balbudhe, "Cloud Storage Reference Model for Cloud Computing", International Journal of IT, Engineering and Applied Sciences Research (IJIEASR), vol. 2, no. 3, **(2013)**.

[8] V. Spoorthy, M. Mamatha and B. Santhosh Kumar, "A Survey on Data Storage and Security in Cloud Computing", International Journal of Computer Science and Mobile Computing, vol.3, no. 6, **(2014)**, pp. 306-313.
[9] S. Patole and A. Sarkeja, "Hybrid Approach for Cloud Storage with Attribute based Encryption", International Journal of Computer Applications, vol. 154, no.1, **(2016)**, pp. 975 – 8887.
[10] Risky CahyaDinatha, I. Made Sukarsa and A. A. K. Agung Cahyawan, "Data Exchange Service using Google Drive API", International Journal of Computer Applications, vol. 154, no.7, **(2016)**.
[11] K.Ajay and G. N. Beena Bethel, "Multimedia Data in Cloud Storage System using DE Duplication Verification", International Journal of Computer Applications, vol. 149, no.12, **(2016)**.
[12] Sreelakshmi P. S. and Sabitha S., "A Review on Cloud Computing Architectures", International Journal of Computer Applications, vol. 152, no. 7, **(2016)**.
[13] K. S. Meena, "An Efficient and Trusted Data Storage Processfor Cloud Computing", International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 3, **(2014)**.
[14] N. Shrivastava and G. Kumar, "A survey on cost effective multi-cloud storage in cloud computing", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) vol. 2, no. 4, **(2013)**.
[15] W.-F. Hsien, C.-C. Yang, and M.-S. Hwang, "A Survey of Public Auditing for Secure Data Storage in Cloud Computing", International Journal of Network Security, vol.18, no.1, **(2016)**, pp. 133-142.
[16] C. Theebendra and N. Santhini, "Cloud Computing Security- Data Storage and Transmission", International journal of research in computer applications and robotics, vol. 2, no. 12, **(2014)**, pp. 27-35.

Authors

Debabrata Sarddar, Assistant Professor in the Department of Computer Science and Engineering, University of Kalyani, Kalyani, Nadia, West Bengal, INDIA. He has done Ph.D. at Jadavpur University. He completed his M. Tech in Computer Science & Engineering from DAVV, Indore in 2006, and his B.E in Computer Science & Engineering from NIT, Durgapur in 2001. He has published around 200 research papers in different journals and conferences. His research interest includes wireless and mobile system, Cloud Computing and WSN.

Sougata Chakraborty is a Senior System Engineer at IBM India Private Limited in Kolkata. He completed M. Tech in Computer Science & Engineering from Jadavpur University in 2011. He had also completed his B. Tech in Information Technology from Murshidabad College of Engineering & Technology under West Bengal University of Technology in 2008. His research interests include Cloud Computing and Mobile Computing.

Priyajit Sen is presently pursuing M. Tech in Computer Science and Engineering at the Department of Computer Science and Engineering, University of Kalyani, Kalyani, Nadia, West Bengal, India. He has completed his MCA from Department of Computer Science and Engineering, University of Kalyani, Kalyani, Nadia, West Bengal, India in 2015. His research interest includes Mobile Computing, Wireless Sensor Network and Cloud Computing.

