International Journal of Grid and Distributed Computing
Vol. 10, No. 4 (2017), pp.21-32
http://dx.doi.org/10.14257/ijgdc.2017.10.4.03

An Efficient Technique to Design Elasticity and Reliable Content Based Publish/Subscribe System in Cloud

Nakka Thirupathi Rao[1], Pilla Uday Bhaskar[1], Pilla Srinivas[1],
Debnath Bhattacharyya[1] and Hye-jin Kim[2]

[1]Department of Computer Science & Engineering
Vignan's Institute of Information Technology
Visakhapatnam, Andhra Pradesh, India
[2]Sungshin Women's University, Bomun-ro 34da-gil,
Seongbuk-gu, Seoul, Korea
{Nakkathiru, comudayc92, debnathb}@gmail.com, hyejinaa@daum.net
(Corresponding Author)

Abstract

As the data arriving from the day to day applications was being increasing from the internet, several problems arise to disseminate the data. The major problem that was being observed recently was to supply the required data to the user who want the exact data from the vast existing data in the internet. The present proposed model was one of the mostly used mechanisms for achieving this task. As the data is huge, the size of the system increases numerously. The cloud computing is the latest technology which was being used to solve such types of problems and provides great opportunities for the users. In this paper, the ESCC Technique which was used to solve such type of problems was discussed in detail. A two layer pub/sub system was developed and analyzed to achieve this task in the environment of cloud computing.

Keywords*: ESCC technique, Cloud Computing, Two layer Publish/Subscribe System*

1. Introduction

In software architecture, the pub/sub system consists of publishers who are senders of messages, and the people who receive the messages directly through specific users. Likewise, subscribers express interest in one or more classes and only receive messages that are of interest. Most of the systems using for the message transmission and receiving and analyzing the messages are the publish/subscribe systems and the models which work on their Java Message Service. Some messages are typically identified for reaction and treating for reducing the redundancy of data, which is also known as the filtering of the data. The most common techniques used for reducing the redundancy of data in a publish/subscribe system are topic based system and content based system.

Publish/Subscribe (pub/sub) system is one of the mostly used asynchronous communication models which could be used mostly for the applications like clouds and other related applications of the cloud environment. The senders who send the messages are separated from other sources like receivers and the interaction between the users can be made by using the present system. Similarly, the receivers of the application or the present system uses the same phenomenon of separating the users and the receivers and the interaction between the users will be made from the present system. The receivers who were connected to the present system with proper connection with a valid identification of themselves and the identity provided by the service provider will register the topics or the messages which were interested to the users. The receivers will register all their requests and their needs and these needs were maintained and represented in the

ISSN: 2005-4262 IJGDC

form of a subscription. The senders who were connected to the present system are used to send the data and the present data was published by the senders to the present system.

The working of the system was very simple and easy to understand by any sort of users. The present system counterparts the messages that were needed to be sent or to be processed for the receiver to subscriptions and distributes the same messages and the required data to the users who were interested in these kind of messages with the help of a mechanism called notifications. Conversely, the development or the establishment of a pub/sub as a service to the interested users gives us a vast in amount of challenges that we were going to face. Some of the modifications or the changes we must consider whenever we are developing such type of a huge system was that it must provide the service should be tremendously scalable in nature such that to maintain or to enhance the good number of subscriptions made by various users and the rate of messages that were being delivered should be at an high rates. The second and most important thing to be considered while working on these systems, the present or the modern developed systems should have a feature of elastic in nature such that making the system more useful and very attractive by changing the working mode of the system which might happen or take place at very less period of time. The final result or the service must be elastic in nature in terms of the working conditions of the system and to rapidly acclimate to workload changes that might occur in a little volume of time and finally the service must be robust to be used by both types of features like the failures in network and the server.

Topic Based System: The messages which are processed in these types of systems are In a Topic-based system, messages are issued in terms of "topics". The second type of users in the system is the subscribers and they receive the messages which were already issued to the topics that they had already selected. All the users or subscribers present or using this system will receive the similar type of messages.

Content Based System: The working of this system is somewhat different compared to the topic based system. Here, the messages that were selected will be delivered or dropped to the users or subscribers only when the attributes of the data selected or the contents of the data that was being selected by a user.

2.1. A High Level Architecture of Content-based Systems

Information filtering systems are the systems which work on the basis of content. These systems basically work on filtering the information based on the content. These systems are used for presenting the profiles of the users, matching the user profiles. The architecture of these systems are represented in Figure 2. The above discussed process can be implemented in three stages and these stages are maintained and monitored by various devices in the area of usage. The following are the several devices that were being used in the working environment of these systems. They are,

- **Content Analyzer**– The major task performed in this device is to identify the items of data related to word documents, newspapers, journals, webpages, news and research articles which was arriving from various sources of information production in a format which was suitable to other applications and accepted working and processing of data in this article. By arranging the data in the present format the data that was processed in this device should be accepted by other system such that to process further. The items in the data are analyzed in various forms such that it should be in an easily formatted and working condition. The data which was obtained here was considered as the input to the next stages of the system or other devices of the system.
- **Profile Learner**– This is the second module or the second stage of the working of the system. The present module gathers data and its related items from various sources and the user preferences and choices such that to regularize this particular

data to design a user friendly profile. The user friendly profile should be easy to understand and easy to work in any system. The generalization policy or the mechanism which was identified and implemented in that module uses the machine learning methods. The present module is a recommender module (4) which works on the basis of feedback method which was used to join the likes and dislikesof the document or the topic (5). This list of topics will be treated as the vector which demonstrates the user profile with his likes, dislikes and his choices.

- **Filtering Component**– The major task of this module is to recommend the similar items of data which matches the data that was represented at the profile. The results that were being obtained from this module are in the form of either binary format or in continuous format of data. This data can be represented in terms of ranking. The similarity index cosine is used to identify the matching ratio between the two files property vector and the item vector.

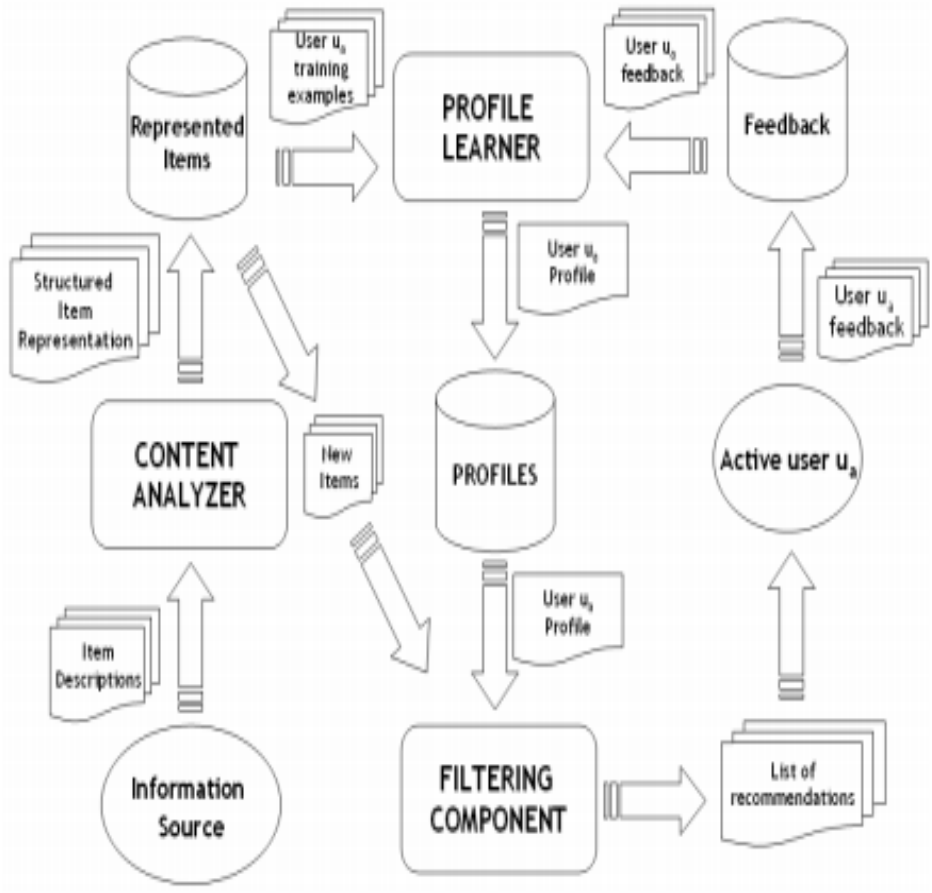

Figure 2. High Level Architecture of a Content-based System

2.2. Advantages and Drawbacks of Content-based Filtering

The content based filtering mechanism was highly used to identify the wanted and unwanted data such that to make the documents more useful and for better performance. Several advantages were present compared to the disadvantages for using this type of filtering mechanism. Some of the advantages of these systems are explained below,

- **Independence to the User in finding the similarity:** Content-based recommenders systems deed merely for the ratings given by the several users who

were in the mode of active such that to develop their own profile. Conversely the collaborative filtering approaches require more ratings from various other users such that to identify the users who were at locations of neighbors at nearest places to the active users. Hence, the items which were interested and likes by most of the users and their neighbors related to the active users were being endorsed or suggested.

- **Working of the system at full Transparency:** The working of the recommender system can be explained in detail with the help of several features or similes of the content which causes an item to be placed in the list of recommendations or suggestions. Hence, these features can be considered as the list of features to be considered such that to verify whether the given recommendation was trusted or untrusted recommendations. Collaborative systems works in full converse with these systems as they follow the mechanism of black box concept. Here the item was explained as a recommendation made by the users who were unknown the system and has given related tastes which like the item that was selected.

Drawbacks or Disadvantages of Content-based Systems:

- **Partial Content Analysis**: Content based systems will have several features which were related either manually or automatically with several objects and the items related to those features. Domain knowledge must be required to propose or decide several features and several decisions. Some of the examples to be considered for the knowledge like recommendations for the movie. The system in the present scenario should know the several details regarding the recommendations like the actors in the movie and their casts, directors of the movie, producer of the movie and sometimes the knowledge of the several ontologies used for the system should be known. These systems cannot be able to provide or suggest the required or the good suggestions if the data required for making the decisions was in a position of inadequate, inappropriate or incomplete. Some of the examples like the Web pages, where the feature extraction techniques from text which may completely flout some aesthetic potentials and supplementary multimedia information.
- **Specialization:** Content-based systems which work on the basis of the recommendations have very negative essential method for finding something unexpected. The system suggests items whose scores are high when matched against the user profile, hence the user is going to be recommended items similar to those already rated. This drawback is also called serendipity problem to highlight the tendency of the content-based systems to produce recommendations with a limited degree of novelty. To give an example, when a user has only rated movies directed by Stanley Kubrick, she will be recommended just that kind of movies. A "perfect" content-based technique would rarely find anything novel, limiting the range of applications for which it would be useful.

4. Proposed System

The main objective is to design and implement elastic and scalable pub/sub system that was familiarized with the workloads of data that were churn in behavior and produces high delivery latency at high arrival rate of real time content such that it amend the balance of the server which was based on the workloads of type churn.

In this part, we had developed a two layer pub/sub system which is distributed based on the cloud computing model such that to deliver accessible and flexible data in the present system.

The currently developed model consists of two layers, one is the matching layer and the other is the delivery layer. At first the matching layer is used for identifying the events

that are going to be matched with respect to the subscriptions and sends the data to the delivery layer. And secondly the delivery layer is used for the events to be ordered based on the semantics from the total orders and sending them to the users who are interested in the same events or same topics of data. Some of the important factors are to be considered for reducing the scalability in terms of subscriptions and the numbers of events that are going to be take place, the dimension of the data that is going to be used in several data based applications. Hence, by following the several workload conditions and situations that were available to us are used for improving the performance of the system. The proposed or the developed model consists of the following layers.

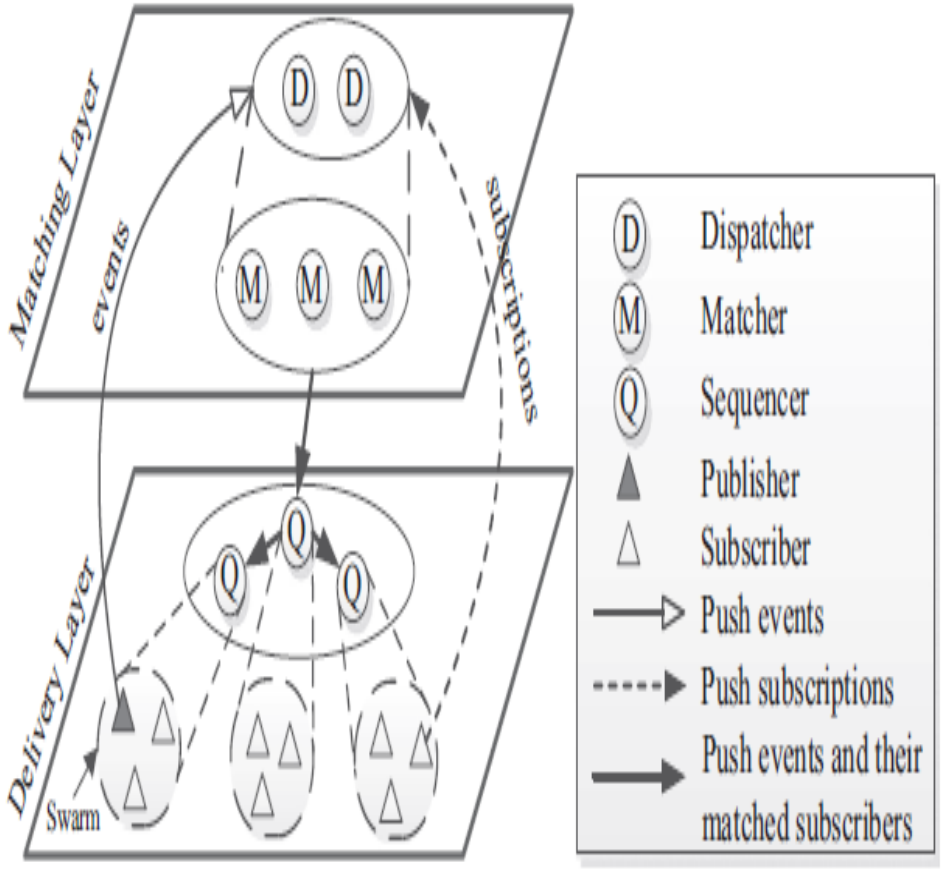

Figure 4. Elastic and Scalable Content based Cloud Pub/Sub System

Matching Layer

The present layer is used mostly and it is responsible for identifying the events that are in same by comparing them and submitting those files to the other layer such as delivery layer. SREM technique is used in this layer to identify the throughput of the documents which were in matched conditions. In the present technique by using a partition technique called hierarchical multi-attribute space partition technique the space is made into several cubes and these were managed by a single server. The number of events that were available in the application and the subscriptions which were available in the current application are kept in the same cube. By following this technique the reduction in the matching latency was observed in a high rate. Also a new technique called SEMAS was developed and considered in this phase of the system.

Delivery Layer

The current layer is used for the purposes of delivery related issues and this layer is considered and focused mainly on keeping the documents in the order either in ascending or in descending order and delivers them to the users who were in need of those events and applications. The important consideration in the present layer was PG Builder which was used to design and develop a graph and the technique which was used to control the provision of the documents was P provision.

PGBuilder algorithm:

1. Initialize cluster C with Q
2. if C==0 then initialize C as new Cluster
3. Process the list L as CPL and if size (CPL)! = 0 the
4. addDPL.get (Q).

```
int I = 0;
while (i<tmpList.size ()) do
e” = tmpList.get (i);
for (int k = 0; k < e”.DPL.size (); k++) do
        tempE = e”.DPl.get (k);
        if (! tmpList.contains (tempE)) then
         — — tmpList.add (tempE);
```

5. i++;

At first, the events that are going to arrive are sent to the several other clusters which were separated and we use a queue called global queue which is used to control all these clusters by adding the e to the trial clusters. The clusters which were present in the GQ are clashed and the cluster which was present here are considered as sliding window cluster. The sequencer in the system model will have a compatibility of processing the cluster which is to be taken as the head one. Once the events are dispatched to their respective users and subscribers, the cluster which considering a head one is removed from the queue which is taken as the global one. Hence, the events that are arriving at a cluster will need to be detected conflicts with the events already present in the system or in the previous documents but not all the clusters only the required and matched clusters. By implementing this model, the reduction in the detection of conflicts in the data matching or events matching was observed with respect to the arrival rate of the events.

Delivery Strategy:

In the present ESCC technique, we can identify or we can observe two types of things one is the sequencer and the second one is the subscriber. Hence, it is observed that as the linking or parting of the roles might seriously obstruct the piece of orders, it is required for us to study and discuss the process of keeping constant and well-organized entire order of systems by following the concept of dynamism in networks.

Subscriber

The major duty or the major task performed by this subscriber is that it transmits the subscriptions made by the users to the dispatchers in the frame work that was being used in our system shown in Figure 4. Whenever a fresh subscriber adds to the existing system, it is postedto the other sequencer. The subscriber is assigned to the sequencer whenever the hash value of the subscriber is approximate or in near to its value by using some hash

model or some hashing technique. The guaranteed delivery can be obtained when the sequences receives the total data that has to be received from the other users.

Sequencer

The working of this sequencer is just like a tree or it follows the working of a tree structure. Whenever a new sequencer comes to join in the existing system, the new sequencer will be treated as a child sequencer to the existing system by a root sequencer. Hence, every sequencer in the system has to verify on its own whether the existing sequencer should join the new sequencer or it should lie with the existing or the current sequencer. This process is done with the existing hashing technique which is a consistent one.

Advantages

- Supports to any number of subscriptions
- High message rate
- Best performance
- Adaptable to any environment
- Receives various data sets without disturbances

5. Comparison between SRRM and ESCC Techniques

Table 5. Comparison between SREM and ESCC Techniques

Techniques	SREM	ESCC
1.	It provides only scalability	It provides scalability and also Elasticity
2.	It does not adaptable to any other environment	It adapts to any environment
3.	Network failure occurs	No Network failure occurs
4.	It does not implement elastic strategies for adjusting churn workloads	It implements elastic strategies for adjusting churn workloads
5.	No guarantee that data receives correctly in various data sizes	Data receives correctly in various data sizes
6.	It performs low price ratio	It provides good performance and price ratio

In this section, a comparison has been provide between the two techniques among which the one technique is existing one and the other is the new technique which was implemented. Several differences were found and observed during the working of these techniques. The first difference or the property which was considered was the scalability. In terms of scalability, the present technique SREM provides the best solutions and the results compared with the existing technique ESCC. The second difference that was considered was the adaptability of the technique with the surrounding environment. The present technique was not up to the expectation of the users. The existing technique was more adaptable to the surrounding environment by the ESCC technique. The most important technique or the feature to be considered was the failure that was going to be observed in the networks. From the results that it was observed as the failures in the network was more in present technique of SREM with the comparison of the existing technique ESCC. Very few network failures were observed during the implementation of the existing technique with the several features and several data sets and various applications.

The other important difference or the factor that should be considered to identify the performance of the developed system with the proposed technique was the size of the data it is going to transmit and the same time the size of the data being received at the receiver end. The observations were made on several data sizes or the various files of various combinations. These observations reveal that the proposed or the new technique was not working clearly on the receiving of the files on various sizes. Whereas the old model or the existing technique was able to receive the files of various sizes and works on the files without considering the size of the files.

6. Results and Analysis

The analysis illustrates that the Multiple clouds like Microsoft Azure [5] and Amazon AWS (S3) [4] has more security when compared to the single cloud. The multi-cloud data is replicated so available even if one cloud fails. Integrity of data is also maintained in our proposed method. The timings are represented for key management technique of key generation, encryption and decryption activities for both single and multiple clouds. This scenario is illustrated in below figures. During the decryption by the user takes some time, this is not constant. It depends on the user in which how much of time that he required to satisfy the specified attributes and policies by the data owner.

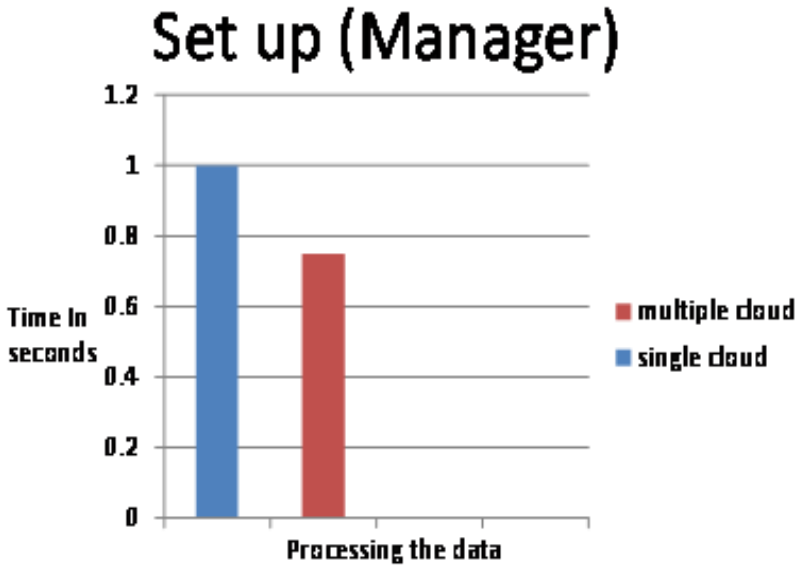

Figure 1. Comparison Graph (Setup manager)

The above Figure 1 is representing the time in seconds to processing the data in the setup manager with cloud. Here the single cloud takes more time to setup the manager then the multiple clouds (AWS). The red color representation s for the multiple cloud and the blue color representation was made for the single cloud. The difference for the working on the various types of the cloud was observed under this section and it was represented in the above diagram.

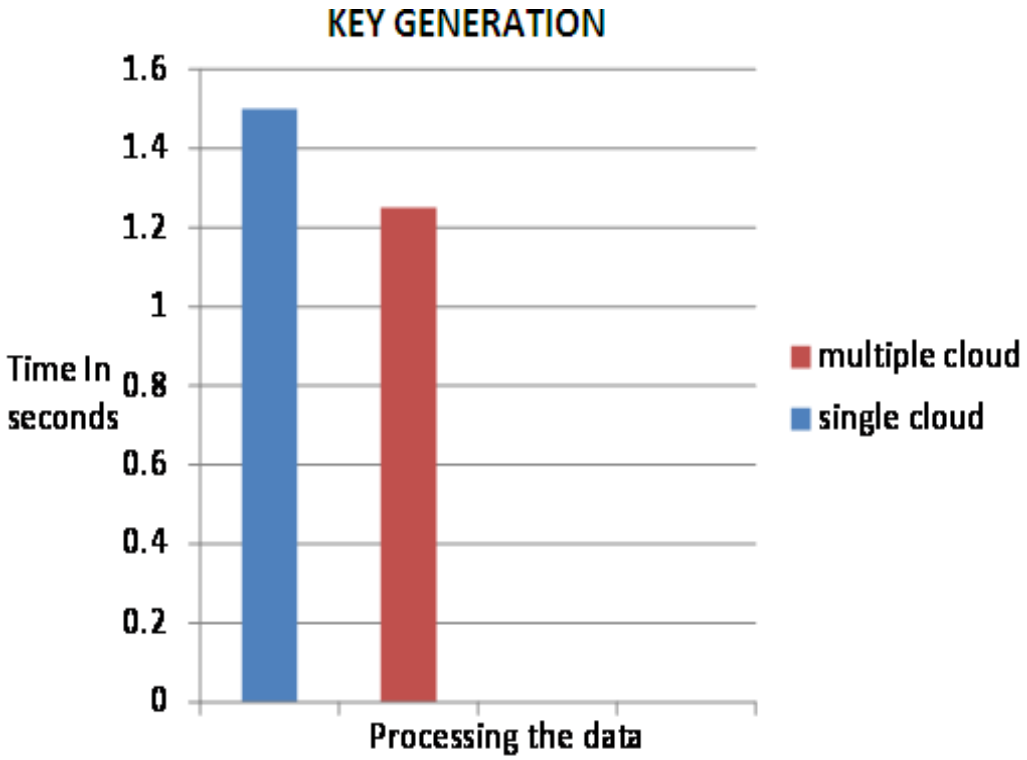

Figure 2. Comparison Graph (Key Generation)

The time processing for key generation to the cloud, here the single cloud takes more time to generate key then the multiple clouds (AWS). The time taking to generate a key was represented in the above diagram. The key generation time was calculated and represented for both the types of clouds. The single cloud and the multiple cloud data and the application was being calculated and represented. The time taking for the single cloud was more than the time taking for generating the key was more for the multiple cloud. The results were calculated and represented for both the types of clouds *i.e.*, the single and the multiple cloud models.

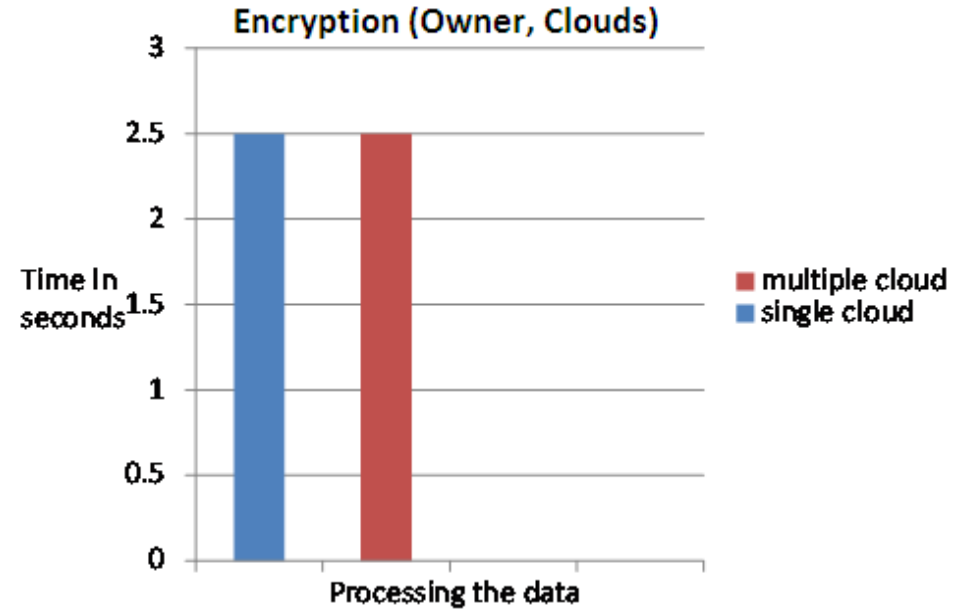

Figure 3. Comparison Graph (Encryption)

The time processing for Encrypt a file in the cloud, here the single cloud and multiple clouds takes same processing time to encrypt a file. The encryption was being used here in this application or this system was to provide the security tot the users with the exact type of three data that the user is requesting the user wants to find the data from the servers. The data might be stored at several stores or at the various locations of the servers at various places. Here the diagram representation was made for the findings of the data for the time with respect to the types of clouds like the single cloud and the multiple clouds. The encryption was being used here to find the exact data requested by the user without making any changes to the existing data. The encryption was made here to provide security to the data to be transferred or the data to be presented for the user before going to be finalized for the verification for the exact data or the data with mismatch.

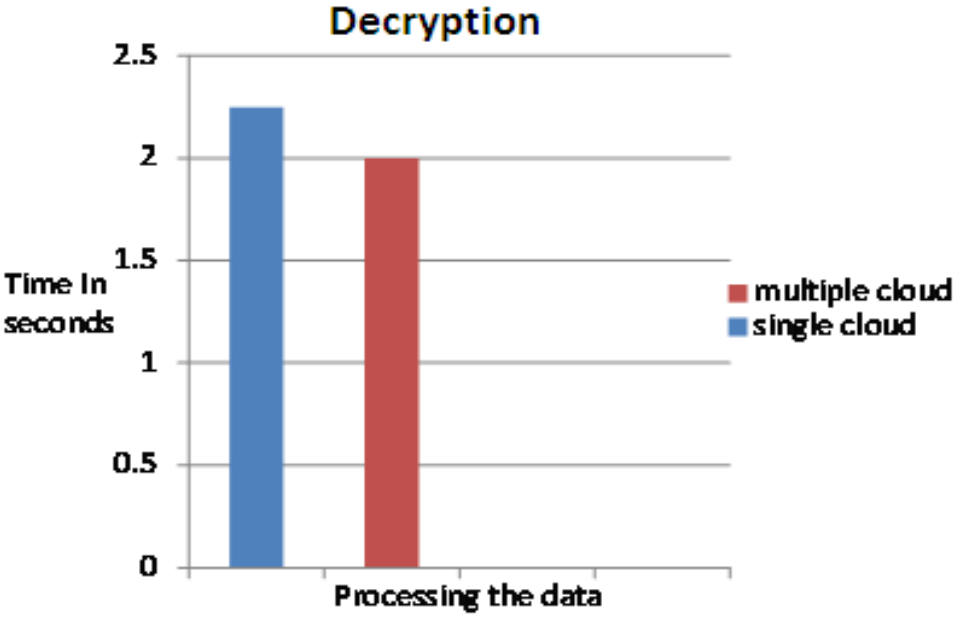

Figure 4. Comparison Graph (Decryption)

The time processing for Decrypt a file in cloud, here the single cloud takes more time then multiple clouds (AWS). The decryption was the model or the technique used here in this model of application such that to decrypt the data that was being implemented or the data to be submitted to the users who requested the exact data or the data items related to the existing data or the data that was being found to be the matched data for the user requirements. The processing time for the generation of the encrypted data the time taking for the decrypting the same data was calculated for both types of clouds. The clouds like the single cloud and the multiple clouds were being used for the applications and the time for these two types of clouds were being made calculated. The time taking for decrypting the messages or the data for the processing of the data for other applications were studied. The time for the single cloud is more compared to the time taking for the decryption was very less.

As we got from the above observations multiple clouds like (AWS) have the more time complexity then the single cloud. From the ESCC Technique the file uploading and download in cloud is fast then the single cloud.

The migration of cloud computing from single toward multi-clouds to ensure the security of user's data is extremely important. Recent research has focused on the multi-clouds environment which control several clouds and avoids dependency on any one individual cloud. Distributing a user's data among multiple clouds is a helpful solution. Approximate timings for encryption and decryption of data in single cloud and multiple clouds of plain text size of 1.2 KB.

Survey on Different Text File Sizes (4kb,40kb,400kb)

The multi clouds have high data transfer to deliver the data to subscriber. The text size are 4kb,40kb, and 400kb are have same time process to upload the file in cloud but as come to downloading the file there are different time processing.

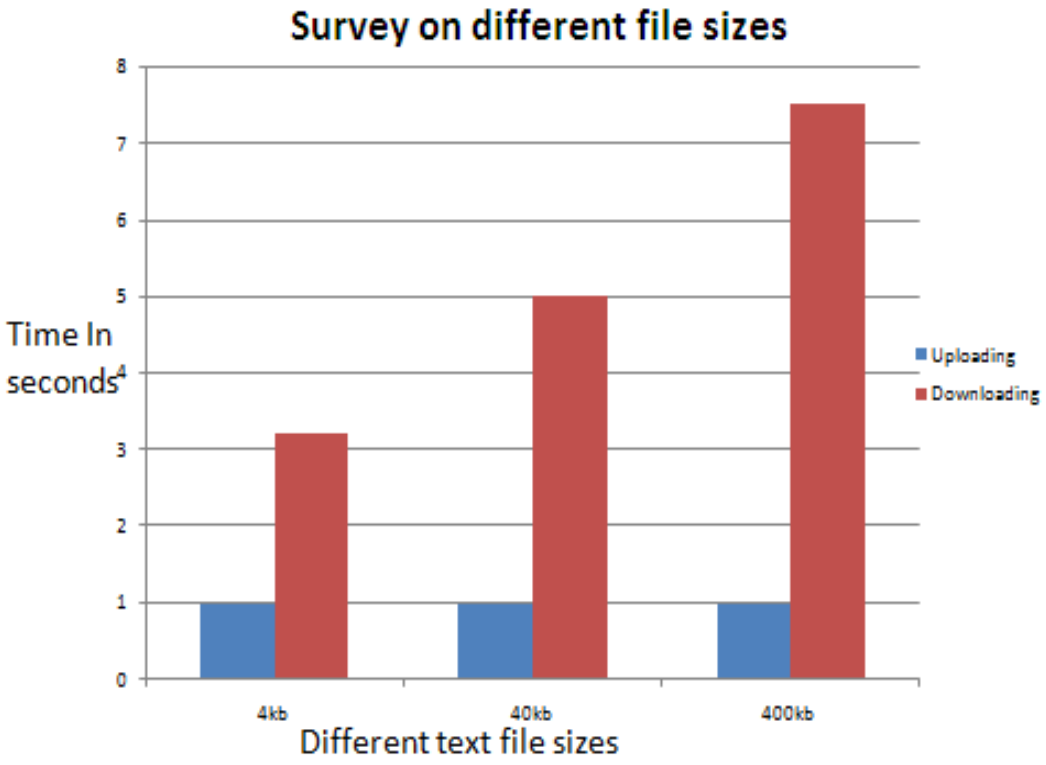

Figure 5. Comparison Graph (Different File Size)

Advantages of Proposed Solution:

1. Reliability in the data can be observed in the data

2. The various services from this method can be available to several users

3. Here the user executes the ritual applications with the help of resources from the service providers

6. Conclusion

From various studies on cloud computing and its allied areas for data analysis and finding the required data from the existing databases that the practice of cloud computing has quickly augmented, its security is quiet reflected as the main concern in the various applications and its allied areas *etc*. The technique that we had discussed in this paper was intended to develop, design and analyze the various elastic strategies from various servers depending on the workloads. In the present work, the multi clouds have high data transfer to deliver the data to subscriber. The text size are 4kb,40kb, and 400kb are have same time process to upload the file in cloud but as come to downloading the file there are different time processing. We had analyzed all these cases through graphical representations in detail. The advantages of this method for selection and implementing this method or the proposed model were explained in detail.

References

[1] J. Raymond. Mooney and R. Loriene, "Content- Based Book Recommendation Using Learning for Text Categorization", Fifth ACM conference on digital libraries, San Antonio, TX, **(June 2000)**, pp.195- 204.
[2] P. Resnick, H. Varian, "Recommender Systems", Communications of the ACM, **(1997),** pp.56-58.
[3] Mitchell T. "Machine Learning", McGraw-Hill, New York, **(1997).**
[4] Rocchio J., "Relevance Feedback Information Retrieval", Prentice-Hall, Englewood Cliffs, **(1971)**, pp. 313–323.
[5] Herlocker L., "Evaluating Collaborative Filtering Recommender Systems", ACM Transactions on Information Systems, Vol. 22, No.1, **(2004),** pp.5–53.
[6] Xingkong, Wang, and Xiaoqiang Pei, "A Scalable and Reliable Matching Service for Content –Based Publish/Subscribe Systems", Vol.3, No. 1, **(2015)**.

