Efficient Data Retrieval for Privacy Preserved Data in Public Cloud

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Mahima Gupta

(801332013)

Under the supervision of:

Dr. Damandeep Kaur

Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Efficient Data Retrieval For Privacy Preserved Data in Public Cloud*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr.Damandeep Kaur* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Signature: Mahima Gupta

(Mahima Gupta)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Damandeep

Dr. Damandeep Kaur

Assistant Professor, CSED

Countersigned by

(Dr. Deepak Garg)

Head

Computer Science and Engineering Department

Thapar University

Patiala

(Dr. S. S. Bhatia)

Dean (Academic Affairs)

Thapar University

Patiala

Acknowledgement

I would like to acknowledge all the people who supported me through this journey on master thesis, whose inputs have proven to be valuable to achieve the desired objectives.

I would like to sincerely thank my mentor **Dr. Damandeep Kaur**, Assistant Professor, Department of Computer Science and Engineering, Thapar University, Patiala, who enlightened me through her motivational thoughts and support throughout the completion of this thesis report. Without her encouragement and guidance this work would not have been possible.Her co-operation and understanding was very grateful that helped me and made this experience honoring and enjoyable.

I am also thankful to **Dr. S.S. Bhatia**, Dean of Academic Affairs, **Dr. Deepak Garg**, Head of Computer Science & Engineering Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, for the motivation and inspiration that trigger me for the thesis work.

Finally, I would like to thank my parents and my friends for supporting me every time in ups & downs and keep me moving in adverse situation.

Mahima Gupta

Mahima Gupta

(801332013)

Abstract

Cloud Computing is a growing technology which enable the organization to focus on their product rather than on physical infrastructure. At very low capital investment, organization can utilize their IT cycle to sharpen their product to establish themselves and to move ahead in the race. Public cloud provide great opportunity for small and establishing organization to focus on their core values and to get their more product and services in the market frequently to win the race. With all these facilities, security and privacy pull away the customer to store their private data on cloud. As they have no clues about where their data is store, who is accessing it? So, before outsourcing their private data in the public cloud they need to ensure that their data is save and does not access by unauthorized users.

In this thesis, a framework has been proposed has been proposed to store the data in secure manner and also securely access the data present in the cloud. The integrity of the data is maintained by encrypting the data. Data is store in an encrypted manner and can only be accessed by authorized users. Data is encrypted before outsourcing it to public cloud and decrypted it after retrieving it from public cloud.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1Cloud Computing Overview

No one has a doubt about the fact that organizations are adopting cloud service at a rapid rate as they have understood the benefits that are delivered by the cloud services. In 2012, a study is conducted by KPMG and found that 81% of the business either assessing cloud services or planning for cloud implementation or had already implemented a cloud solution [1]. Without increasing the IT budget, organizations can enjoy the tractability, scalability and agility provided by the cloud. Now, among the IT managers and even CIO's concern is whether to choose private or public or hybrid cloud.

1.2 Cloud Computing Deployment Model

1.1.1 Private Cloud

Private cloud also known as "internal cloud" or an "enterprise cloud"[1]. Private Cloud can also be depict as a cloud computing platform which is carried out within the corporate firewall, and operate by the IT department [2]. Public Cloud is defined as a "single-tenant" environment where the network, software and the hardware are committed to a particular organization[5].

Private cloud is fully dedicated to a particular organization, in which an organization is responsible for managing all the resources and application. It is not shared with other organizations and have more control on infrastructure. There are two variation in private cloud, on-premise private cloud and externally hosted private cloud. Externally hosted private cloud are handle by third party but can be used by one organization only[3].

Private Cloud is an obvious choice when[4]:

- User want to have consistency across the services
- User want to have data center more effective
- User want to have more control over data along with cloud efficiencies

- Organization server capacity is more than it can use

1.2.2 Public Cloud

Public Cloud is the cloud which is handled and owned by the cloud service provider. Individual does not have any control over the cloud infrastructure. Any number of organizations can share the resources and service provided by the cloud. Public cloud is based on "pay-as-you-go" model. Public Cloud is defined as a "multi-tenant" environment where the user buy the "server-slice" in cloud computing environment which is shared with other tenants [5]. Organizations does not worry about the infrastructure cost, as it is spread across all the users and allow each client to operate on low cost basis[4].

Public Cloud is an obvious choice when[4]:

- User or an organization want to work on project in a collaborate manner
- User or an organization only want to develop and test the application code
- User or an organization want the power to increase compute resources for peak time
- Lots of people are using the standardized workload for application, like e-mails

1.2.3 Community Cloud

Community cloud is a cloud in which infrastructure is shared among various organization having same concerns. Community Cloud are controlled, handled and secured commonly by all the organizations participated or by the third party service provider. All the organizations having similar requirement and their main goal is to work together to accomplish their business target collaborate together to form a community[2]. Cost of the community cloud is spread over the few users which are less than public cloud but more the single tenant. The best example of community cloud is it belong to the government of a single country[5].

Community Cloud is an obvious choice when[4]:

- Resources need to be shared within a state by government organizations

- Specific FCC regulations need to be meet by Telco community cloud for Telco DR
- For a group of clinics and hospital private HIPPA complaint cloud is used

1.2.4 Hybrid Cloud Service

A hybrid cloud is a combination of two or more cloud whether it is public or private or community cloud and designed to use by single organization. The hybrid cloud concept help an organization to attain maximum usability while reducing the faultdegree. Hybrid cloud is used by an organization to increase efficiency by utilizing public cloud for non-sensitive data operations and private cloud for sensitive data operations and ensures that their all platforms must be integrated seamlessly[6].

Hybrid Cloud is an obvious choice when[4]:

- An organization is providing public cloud for the customer and using private cloud for exclusively internal work.
- An organization want to use Software-as-a-Service but they are also concerned about the security

1.3 Benefits of Public Cloud

Here are some of the reason which proves that why public cloud are better than any other cloud.

i) *Utility Model*- This is the most attractive benefit of public cloud computing. Public cloud typically based on pay-as-you model.Cloud users need to pay for computing according to the hours. For example, if cloud user run the test server for 1hour regression testing, the user need to pay less than $0.10[7].

ii) *Elasticity* - Public cloud provide great elasticity as user is not able to consume whole capacity of a public cloud[9].

iii) *Scalability*- Public cloud has vast pool of resources so it provides cloud resources on demand that helps the applications which are running on cloud to respond quickly the variations in the activity very easily[8].

iv) *Reduce Time* - Maintaining an in-house servers take time. For example-if the configurations of hardware and software need to be change or there is need to restart the server or the server crashes, depending upon certain situation this

will need a couple of hours or days. But with public cloud, as everything is virtualized, reconfiguring the environment hardly takes minute. Secondly, if one server fails another server can be activated instantly reducing the down time as servers are hosted within the cloud environment [1].

v) *Reliability* - As public cloud consist of large number of servers and networks, and has redundancy configuration that is even if any physical component fails to work, cloud service will run ineffectively using remaining components. So, there is not a single failure which can make public cloud services vulnerable.

vi) *Mobility and Storage* - Public cloud are best for start-ups , small businesses and even for individual department in large organizations which require cheaper and scalable services for the flexible service needs. The staff can be worked from anywhere independent of the location and store all the data which is backed up in the cloud for free[10]. The storage of the public cloud is long-lasting. They backed-up all the data and store replicated copies in multiple location for disaster recovery.

vii) *No Maintenance* - Internal IT employees are not responsible for the maintenance of the system as public cloud is maintain by third party. Updating the technologies is not the task of IT employees, the design update and introduce the latest technologies to enable the users to use as fast as possible [1].

viii) *No Contracts* - As public cloud is based on pay-as-you-model, so long-term commitments are not there. User is not under any responsibility to continue public cloud once yearly or monthly agreement is over[1].

ix) *Big Data Analytics* - Organization can get the big data benefit by making use of big data and analytics infrastructure solution which areavailable on cloud. Example- Amazon Elastic Map-reduce [11].

1.4 Data Privacy

In our day to day life many times we use this word privacy, then what actually privacy means?? Privacy is secrecy that everyone want in their life. People want that their data, item, information which they want must be private, secrete from others. Data privacy in cloud also means the same, that user want that the data which they store on cloud must be private from other users, their data might contain some

information which if leaked, can cause problems to the users. Privacy concept is different in different cultures, countries or jurisdiction.

In general, Data Privacy which is also known as information privacy, is an Information Technology(IT) aspect that deals with an organization or an individual ability that has to determine what data can be shared with third parties present in the computer system.[14]

According to the Organization for Economic Cooperation and Development (OECD) [13]:

"Privacy is any information relating to an identified or identifiable individual"

According to Generally Accepted Privacy Principle (GAPP) standard by American Institute of Certified Public Accountants (AICPA) and the Canadian Institute of Chartered Accountants [13]:

"Privacy is the right and obligations of individuals and organizations with respect to the collection, use, disclosure of personal information."

1.5 Importance of Data Privacy in Public Cloud

With the widespread of technology and its awareness, people have resorted to the use of latest technology in all spheres of work. And when talk about technology, cloud services always come in mind. People make use of cloud services to store their data on cloud. And the data which people stores on cloud contain much more sensitive personal information. These information must be store in a manner which limits or restrict the disclosure of the sensitive personal information. Now, to frame such system firstly two question comes in mind:

a) What do you mean by sensitive information which is need to be protected?

b) Up to what extent disclosure of sensitive information need to be limited? [12]

Data privacy is very crucial for both business and home computer users also. Mishandling data can have serious rebound for organizations and even for their employee and supporters. One example of mishandling of data include mislaid laptops and USB sticks left on train. Privacy failure can lead to the unlimited financial

punishments, damaged reputation, revenue loss, bad press and supporters trust loss[15].

Data privacy concern for users can also be measured on the scale of embarrassment and financial. For example, a user may not want to share their health records publically because it might be embarrassing for some people to know that how bad blood pressure of user is and insurance company, if they knew might charges more[16].

In cloud, privacy is an important aspect one should work upon. Cloud service provider is responsible to make sure that user's personal information is secured from other users. Provider must also have complete knowledge that is presently accessing the data and who is maintaining the server so that provider provide the privacy to user's personal information. Providers should make sure that only the authorized users can access the data[17].

In cloud, the network that interconnect the system has to be secure. Moreover, several security concern are there in cloud due to the virtualization paradigm like mapping of physical machine to virtual machine has to done securely. Also, algorithms like resource allocation and memory management also need to be secure.Figure 1.1 shows key areas in cloud computing where privacy should be checked as:

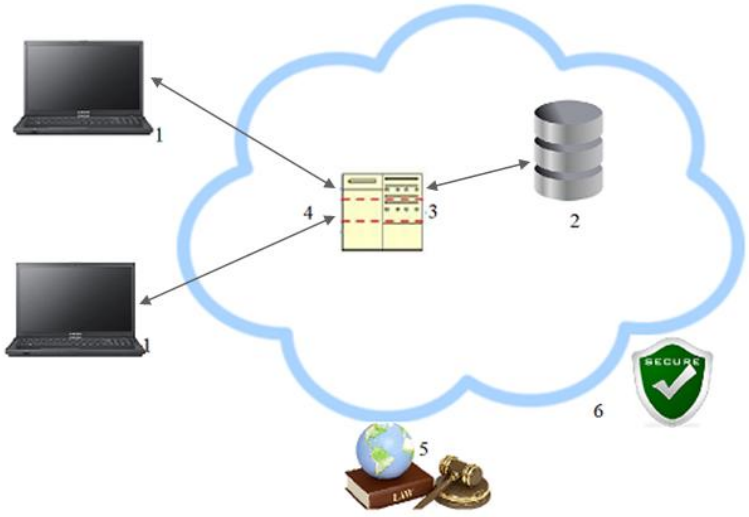


Figure 1.1: Six Areas of Privacy Concern in Cloud Computing: 1)Authentication 2)Privacy of data at rest 3)Privacy of data in transit 4)Separation of data belong to different users 5)Cloud legal and regulatory issue 6)Response to the accidents [17]

Six areas are: 1) user/processes/application authentications 2) privacy of data at rest 3) privacy of data in transit 4) clear separation of data belong to different users 5) Cloud legal and regulatory issues and 6) response to the accident [18].

Consider a case where hospital maintains the record of all patients. Now, hospital wants to expose some information to the pharmaceutical company in such a manner that pharmaceutical company cannot find out that which patient has which diseases. Here, some application or framework must be there which can accomplish this so that privacy of all patients will be maintained.

1.6 Cryptography

Cryptography which can also translates as "secret writing", is an art of converting plain text into an unreadable format called cipher text. Cryptography is a science of hiding the data meaning so that only the specified parties are able to understand the meaning of data transferred.

Cryptography is a science which implements both logics and mathematics to design strong encryption method.A cryptographic algorithm is a set of functions or a process to perform encryption and decryption of the data. Figure 1.2: show the cryptography overview.

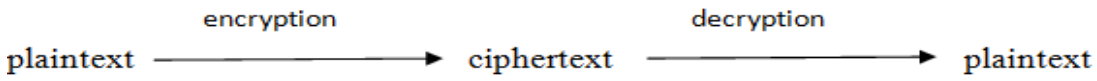


Figure 1.2: Overview of Cryptography

With modern cryptography, for encryption and decryption of the data a variable value which is known as key is used to generate an encrypted result or to decrypt the encrypted result. Cryptography enable the user to get some confidence over their data to store on cloud. Aim of the cryptographic algorithm is to encrypt the original data in such a manner that it will become difficult for unauthorized users to decrypt it without using key.

1.6.1 Encryption

Encryption is a process of converting original data into unrecognizable and useless form to an unauthorized person. For encryption process, a secret key is used. Original

data and the secret key is an input to an encryption algorithm and gives Cipher text as an output. Encryption is a best mean to secure data at rest as well as in transit. Figure 1.3 show the encryption process.

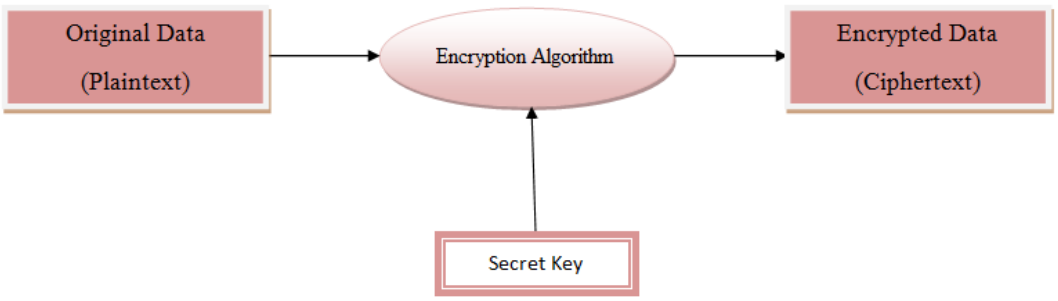


Figure 1.3: Encryption Process

1.6.2 Decryption

Decryption is the reverse of Encryption. Decryption is a process of converting encrypted data into to the original form which has some meaning. For decryption algorithm, encrypted data or the cipher text and the secret key (which may or may not be same as of encrypted key) is the input and the output of the algorithm is the original plain text. It is difficult to decrypt the data without the original secret key. Decryption process is shown in Figure 1.4.

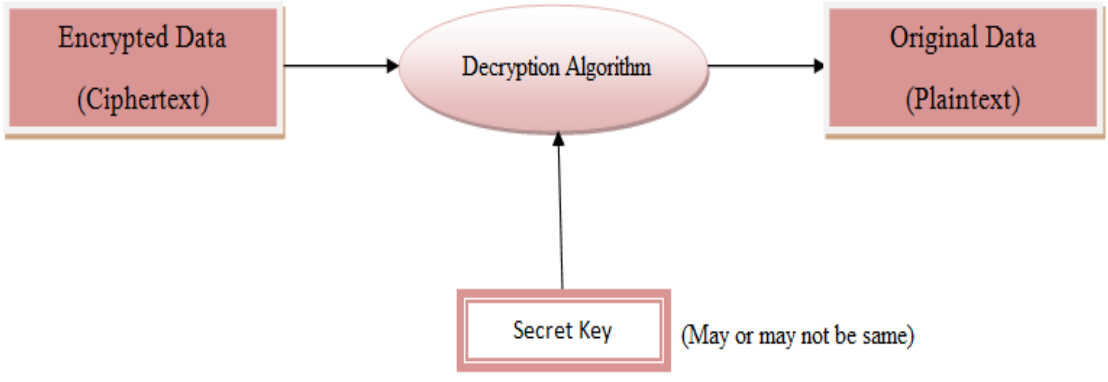


Figure 1.4: Decryption Process

1.7 Goals of Cryptography

The purpose of cryptography is protection of data from an adversary. Modern cryptography has 4 main goals:

i) *Confidentiality*: Confidentiality means making information unavailable to those who are not authorized users or entities. In communication process when two or more parties are communicating, then the purpose of confidentiality is to make sure that only those communicating parties must be able to get information, no other party understand the data exchanged. It also make sure that no one can identify the source and destination of the message. Confidentiality is applied by encryption[19].

ii) *Integrity*:Integrity means the information should not be alter or destroy in unauthorized manner. Integrity required as data is transmitted over the network such as internet were the attack can be happened in the middle. One should be able to detect is there is any data manipulation while transferring the data. Data manipulation include insertion, deletion or replacement of original data[19].

iii) *Non-repudiation*: Repudiation means any of the entities involved in communication can deny. Non- repudiation is a service by which any of the entity involved in communication cannot deny from their commitments. So, in case of disputes due to the denying by any entity certain actions can take place which is necessary to resolve dispute[20].

iv) *Authentication*: Authentication is a service which is associated with identification. Both the parties which are communicating must authenticate themselves based on either password or key or biometric measurements such as voice recognition or retina scan or combination of any of these. Data integrity is implicitly provided by data authentication[20].

1.8 Cryptographic Technique

Cryptographic techniques can be classified as shown in Figure 1.5

i) Symmetric Key Cryptography

ii) Asymmetric Key Cryptography

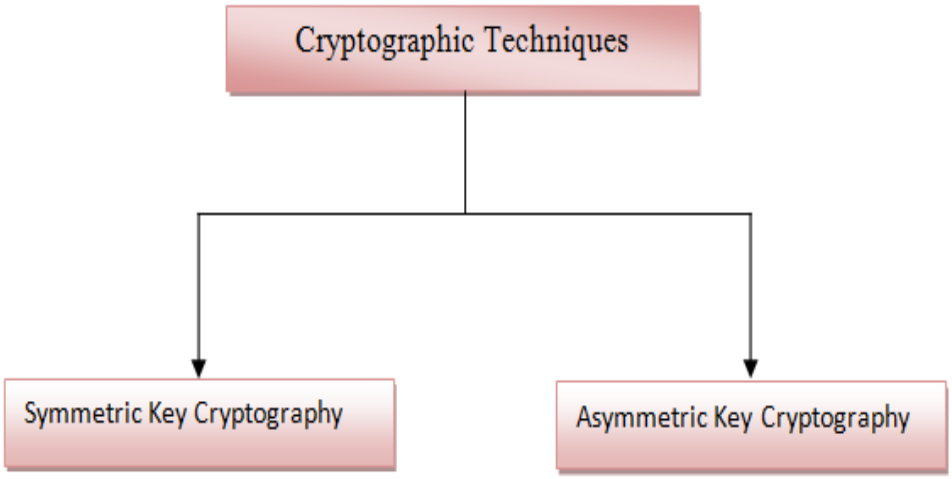


Figure 1.5: Cryptographic Techniques

i) *Symmetric Key Cryptography*: Symmetric encryption also known as conventional encryption or single key encryption. Symmetric encryption is a mechanism in which single common key is used for both encryption and decryption process.

Symmetric Key Encryption mechanism involves 5 components is shown in Figure 1.6:

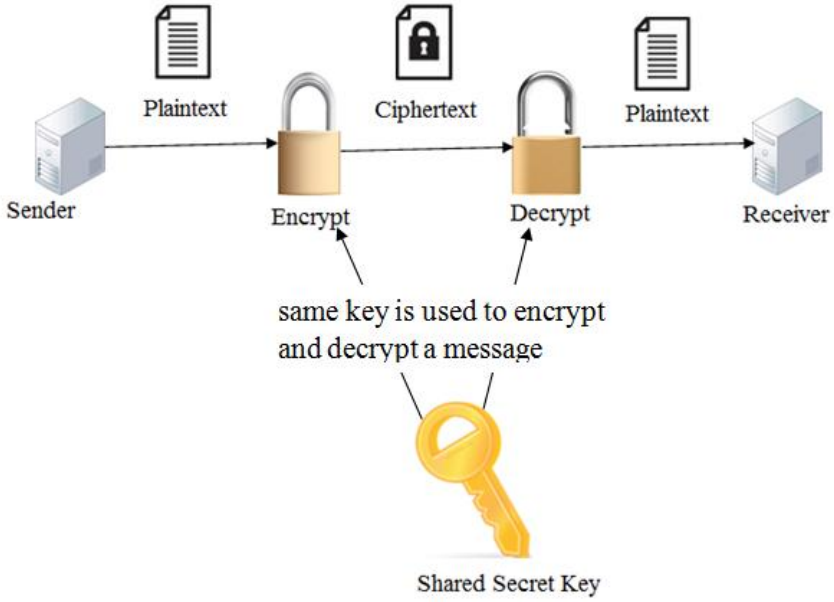


Figure 1.6: Symmetric Key Encryption Mechanism[21]

i. Plaintext- It is the input to the algorithm which is in readable form.

ii. Encryption Algorithm- It is used to encrypt the plain text so that it will get converted into a non-readable form.

iii. Secret Key- The secret key is an input to the encryption algorithmwhich is used to convert the plain text into ciphertext. Depending on secret key same algorithm can produce different encrypted output.

iv. Cipher Text- It is an encoded/encrypted message which is obtained asa result of transformation of plaintext.

v. Decryption Algorithm- Decryption algorithm is used to decrypt the cipher text back to the readable form using the same secret key which is used for encrypting the data.

ii) *Asymmetric Key Cryptography*: Asymmetric encryption also know as public key encryption. In asymmetric key encryption, different but related keys are used separately for encryption and decryption.In this technique, sender's encryption key is made public and decryption key is private. Private Key is send to receivers in secure manner. Only with the sender's private key, the data can be decrypted.

Asymmetric key encryption mechanism involve 5 components as shown in Figure 1.7:

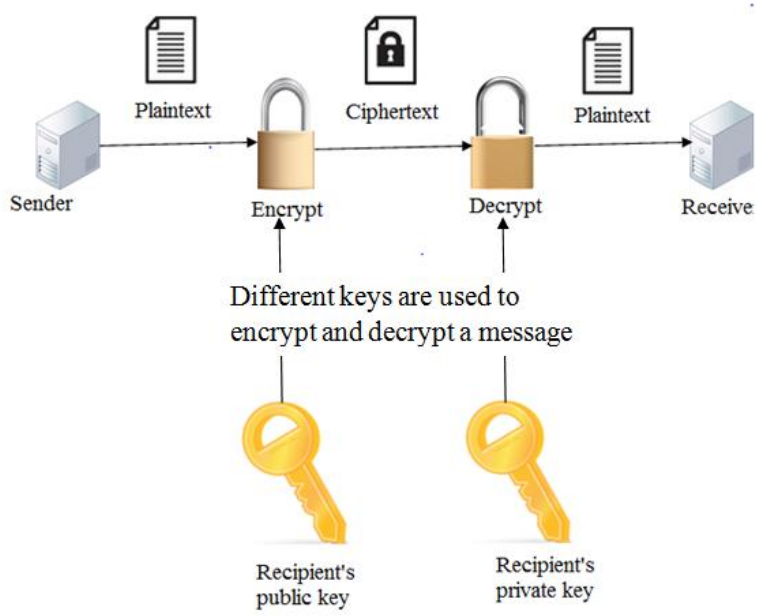


Figure 1.7: Asymmetric Key Encryption Mechanism[21]

i. Plaintext: It is the original data which users want to encrypt. It is a readable form and input to the algorithm
ii. Encryption Algorithm: It is use to encode the plaintext which is input so that no one can read the private information thereby maintaining the privacy of data.
iii. Public and private key: These are the pair of keys which is used for encryption and decryption. Public key is used as an input for the encryption algorithm and private key is used as an input for the decryption algorithm.
iv. Ciphertext: Ciphertext is in non-readable form so that no information can be obtain from it. It is an output of the encryption algorithm.
v. Decryption Algorithm: To get the original data back from the ciphertext, decryption algorithm is used. For decryption, private key is used. Without using private key, decryption cannot be done.

1.9 Structure of the Thesis

The rest of the thesis is organized in the following order:

Chapter 2 provides the literature review and chapter 3 defines the problem statement for the work and chapter 4 describes the proposed framework and chapter 5 provides the implementation and the result of the proposed system. Finally, chapter 6 gives the conclusion and the future scope of the work.

Thesis concludes with the references and publication.

Chapter 2

Literature Survey

Privacy of the data present in cloud is such a big issue that many techniques, frameworks, methodologies *etc* has been proposed so that the privacy of the data can be ensured and user's can outsourced their data on cloud without any hesitation.

In [22], Zhang Wei*et al.* describe a new method, called Bit Split-Bit Combination(BSBC) to protect the privacy of data. In this method, data privacy protection does not depend on any of the original encryption algorithm. According to Wei, if privacy protection is provided to data using encryption algorithm, then user's need a key to decrypt the data, and if key is lost user will not be able to decrypt the data. Also, a large number of keys are required for data encryption and maintaining those key is also to big task. So, to solve all these problem BSBC method is there in which using bit-split algorithm data is recoded then split into many files before storing it to multiple cloud storage platform. When user want to access the original data, then all the files are downloaded from multiple cloud storage platform and then decoded using bit-combination method.

In [23], Aswin*et al.*, proposed a cloud intelligent track system with the help of privacy manager and risk manager. They provide this technique for hybrid cloud in which privacy manager resides in private cloud not in public cloud. Privacy manager will receive the data from the user and segmented it according to size, generate key,encrypt the data and finally store in random location in cloud. All those location information will be store in privacy manager database. Risk manager deals with the risk analysis of the data while receiving or transmitting to public cloud. In their system, client will not directly access public cloud, data to the client will be received via privacy manager.

In [24], information leak risk on analyzing the data on the third party cloud server due to the unauthorized access or criminal activity within in a cloud service provider has been discussed. To surmount this problem, Hitachi proposed a searchable encryption privacy preserving analysis technique. Searchable encryption technique allow text matching to performed on encrypted data itself and also used for statistical analysis

or analysis of correlation rule of encrypted data. As this privacy preserving technique required only encrypted data and encrypted query to perform analysis, the risk of the information leak or unauthorized access is reduced as without disclosing the content of the data to the service provider analysis has been done.

In cloud environment, group signature using Identity-based cryptosystem method [25] has been introduced by S.Kuzhalvaimozhi and G.Raghavendra Rao. They say that many authentication algorithm are there but no such algorithm is there which provide anonymity to the user. If the group of users want to access the data without revealing their identity to the service provider, theycannot convince the provider that the user sending request is an authorized user. So, they propose a secure anonymous method which uses identity based group signature for the cloud service in which the cloud user can access the data store on cloud without disclosing their identity. In this method, the real signer must be traced by Group Manager. Group Manager assures the receiver that the signature was develop by legitimate user and the manager can expose the identity of the signer in case of any dispute. In the group, everyone knows that the sign was produced by one of their group member but no one knows who the actual signer is. In this manner cloud provider does not know the actual identity of the individual user.

2.1 Searchable Encryption

Searchable encryption is a general term for encryption technique, in which not only encryption and decryption can be done but also text matching can be done using encrypted query on the encrypted data. For encryption and decryption process keys are required but for text matching no keys are required so it can also be performed by cloud service provider without the knowledge of keys. Still, some technique uses separate private key for text matching so that can be performed by authorized user's only[24].

A number of searchable encryption techniques are there. Some of them are:

Song *et al.* [26] initiated searchable encryption in which they provide a technique for remote searching on encrypted data without decrypting it and also divide the cryptographic components into client side and server side in which client will encrypt/decrypt the data and generate a query while the keyword based search is done

by server side. Under a specified two-layer encryption each word is encrypted independently.

Later, Secure Index was define by Goh[27], using which develop a security model for the indexes, the model is known as semantic security against adaptive choose keyword attack(IND-CKA). Secure indexes for the keyword search was constructed using Bloom filter in which the server will be able to assure if keyword is present in the document without decrypting the whole document.

Curtmolawt al, [28] built a searching algorithm for the encrypted index based on hash table. Each entry of the hash table consist of trapdoor and encrypted file identifiers. In earlier work on searchable symmetric Encryption (SSE), each query is submittedonly by the owner of the data. This work is a natural extension where the parties other than the owner of the data can also submit the search queries.

Based on Asymmetric Cryptography, searchable encryption algorithm was proposed by Boneh*et al.*[29]. In their scheme, public key will be used for encrypting and uploading the data to the cloud server but for searching the encrypted files, private key will be used by authorized users only.

Golle*et al.*[30], proposed conjunctive keyword search throughout the encrypted data. This scheme conduct search in a secure manner. In conjunctive keyword search, all those files containing all the keyword required by the users will be retrieve. This is the first scheme for which communication cost is linear. This scheme trust Decisional Diffie-Hellman (DDH) assumption for security.

Wang *et al.* [31], proposed an efficient ranked keyword over the encrypted data store on cloud. In ranked keyword search based on ranked relevance value the files will be return to the user. Searching through ranked keyword improves the system usability by modifying the search result instead of sending identical results and also ensures accuracy of the file retrieval. Also the time cost of the file decryption as well as communication cost of returning them can be reduced using ranked search. Drawback of this scheme is that it support only one keyword per query.

Cao *et al.*[32], then work on the drawback of Wang *et al.*[31] by proposing a scheme for multi-keyword ranked search for the encrypted data(MRSE) store on cloud. Their scheme further improves the keyword search result accuracy by retrieving specific

files but for that cloud users required to provide exact keywords. Between several multi-keyword semantics, they choose "coordinating matching" for efficient similarity measure and "inner product similarity" for evaluating quantitatively similarity measure. Basic idea of MRSE is proposed using secure inner product computation.

Li *et al.* [33], says that searchable encryption over the encrypted data in public cloud for privacy preserving keyword search does not support simultaneously fine-grained access control. So, they accept the challenge and solve the issue by introducing hybrid architecture in which private cloud was introduce between public cloud and user as an interface. They allow both keyword search as well as fine-grained access control simultaneously over the encrypted data. Keyword search improved by using fuzzy keyword search. In their scheme, all the overhead is on the private cloud, only encryption and decryption will be done at user's side.

Li*et al.* in [34],provide novel framework so that encrypted data can be outsourced and shared securely on hybrid cloud. Their framework provides two features : i) authorized users without sharing same private key can perform keyword based search on the encrypted data directly and ii) for achieving fine-grained sharing of encrypted data two-layer access control is provided. In their framework, storage service offered by cloud service provider can be used by group of users so that they can outsource their data on the cloud so that it can be shared in a secure manner. Only the authorized users will be able to perform keyword based search over the encrypted files.

Later on, Li *et al.* in [35], enhance the above system in two steps: i) in terms of functionality, the system will support advance fuzzy keyword based search. Overhead computation due to generation of fuzzy keyword set and the decryption at the data user side can be eliminated with the help of private cloud and ii) in terms of efficiency, outsourcing attribute based encryption issues to the private cloud can further be discussed to lighten the user's side computational cost.

Chapter 3

Problem Statement

In today's era, delivering new features frequently and promptly in the market is key to success for an organizations that they required to play these days. Establishing and supervising datacenters is a trade off and spending IT cycles on handling physical infrastructure barely adds any value to the bottom line of the business. Public cloud provide great opportunity for organizations to focus on their core values rather than on physical infrastructure. Specifically for the startup organization with very less capital and great engineering resources, it is beneficial to adopt public cloud and build master product quickly and efficiently at much lower cost and provide services for lower cost. With this, a startup organization can take a good start and can give a tough competition to already established organizations successfully. In spite of economy not a impulsion for many organization, they are still moving to public cloud to outsource their product quickly to market. One of the reasons to organizations for moving to public cloud is unlimited scalability and resources offer on demand. Organization also need not to worry about the latest technologies used, only they need to focus on their services which they are providing to the user.

At the top of the list security and privacy of the data store in a public cloud is in fact a major concern. Having no geographical restriction on accessing data from anywhere is often seen as an advantage of public cloud. This does not means that data in the public cloud does not have any security. , but data servers will be store in different country which is ruled by completely different set of security and privacy regulations. But users hesitate to store their personal or highly confidential sensitive data on public cloud. Organization is not aware about where their data is stored, if or how it is backed up and whether an unauthorized user is restrict to access the data or not. With this, data privacy become the major concern of an organization as they are storing their sensitive data on public cloud leaking of which can lead to many problems to them.

So, before outsourcing their sensitive data, organization need to be assure that their data must be:

- secure
- access at high speed
- should not access by unauthorized users
- data storage is reliable
- avoid data loss, leakage
- avoid unauthorized modification
- guarantee replication of data in a jurisdiction and uniform state
- identify, control that to what extent cloud sub-contractors are involved in processing

Best mechanism to ensure privacy or/and confidentiality of the data is encryption. Organizations must store their data on public cloud in an encrypted format. They should also apply some criteria, agreeing of which only, users are allow to access the data.

Chapter 4
Proposed Technique

User when outsource the data to the public cloud they assumed to trust the cloud service provider. However, server cannot be trusted with the sensitive data. It is necessary to prevent the data from access by unauthorized users. It is also important to maintain the data privacy so that no other users can read the private data of the data owner without their permission. So, data must be outsourced to the public cloud in an encrypted format.

4.1 Encryption Algorithm Used

Various encryption algorithm are there which can be used to encrypt and decrypt the data in the proposed framework, AES(Advance Encryption Technique) is used to show the workflow of the proposed framework.

AES is a symmetric key encryption algorithm with a block length of 128 bits. It is developed by cryptographer Joan Daemen and Vincent Rijmen. It has 3 different key length *i.e.* 128 bits, 192 bits or 256 bits. Variable number of round is used by AES which are fixed. For 128-bit keys 10 rounds of processing are there, for 192-bit keys 12 rounds of processing are there and for 256-bit keys 14 rounds of processing are there. In each case all the rounds are identical, expect for last round. For each rond four steps are there. One single-byte based substitution step, a column-wise mixing step , addition of round key and a row-wise permutation step is included in each round of processing. Execution order of the steps differ for encryption and decryption process.

4.1.1 AES Overall Structure

AES structure is shown in Figure 4.1 for case of encryption key 128-bit long, consist of 10 rounds. Firstly, prior to any round based processing, array input state is XORed with first four words of key schedule. For decryption, same process follow the only difference is ciphertext state array is XORed with last four words of key schedule.

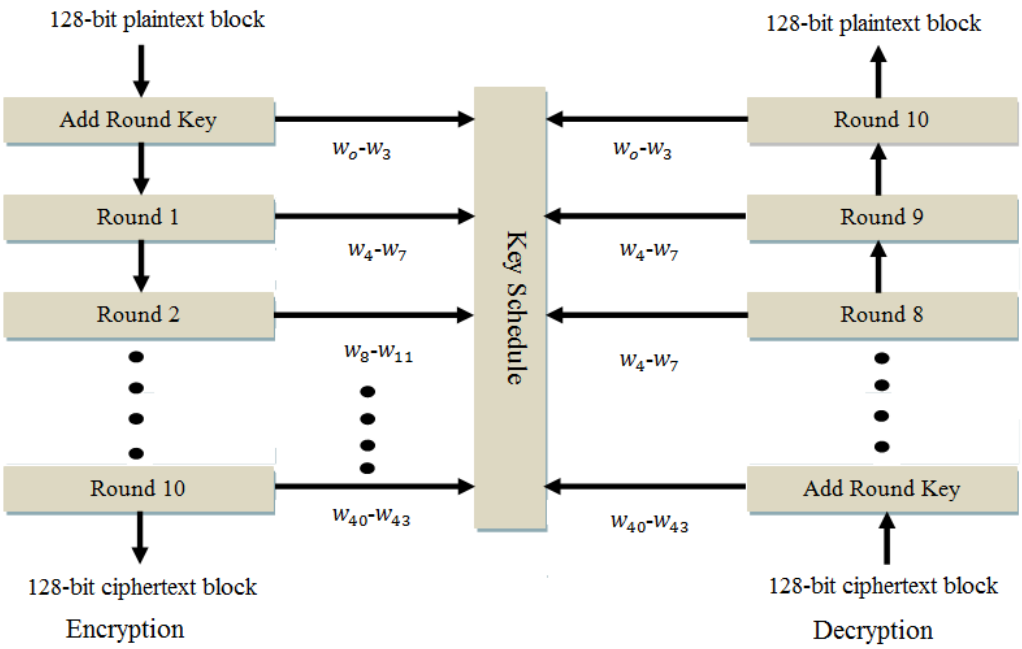


Figure 4.1: AES Structure

4.1.2Steps in each Round of Processing

Figure 4.2 show the four step in each round of encryption and decryption.

Step-1: It is called as Sub-Byte for byte-by-byte substitution during forward process. In this step16 × 16 lookup table is used to find replacement byte for the byte given in input state array.

The corresponding substitution step at time of decryption is called InvSubBytesi.e Inverse Substitution Bytes.

Step-2: It is called ShiftRows for shifting the state array rows during forward process. This transformation goal is to scramble the order of the byte inside each 128-bit block.

The corresponding transformation during decryption process is called InvShiftRows*i.e.* Inverse Shift-Row Transformation.

Step-3: It is called MixColumns where mixing up of bytes is carried out in each column separately at the time of forward process. The aim of this step is to further scramble up the input block 0f 128-bit.

The corresponding transformation at the time of decryption is called InvMixColumns*i.e.* Inverse Mix Column transformation.

Step-4: It is called AddRoundKey in which round key is added to the output of the previous step at the time of forward process.

The corresponding step at the time of decryption is called InvAddRoundKeyi.e Inverse Add Round Key transformation.

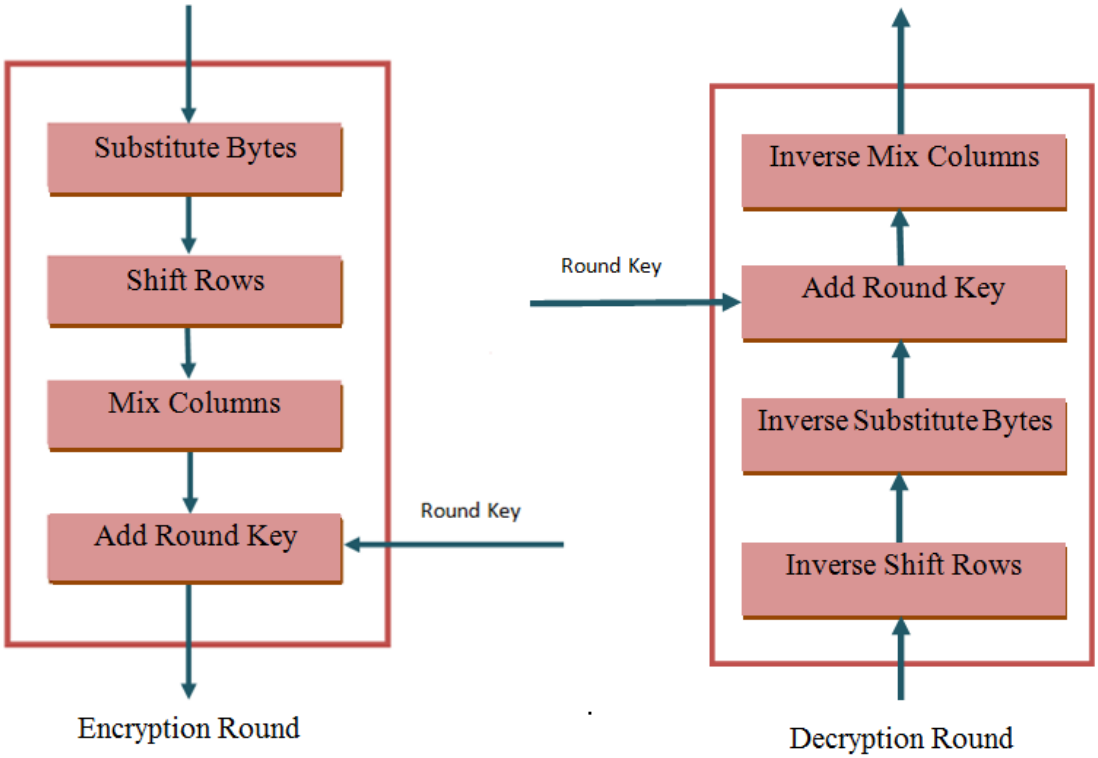


Figure 4.2: One Round of Encryption and Decryption

4.2Proposed Framework

In the proposed framework, the basic focus is to maintain the data privacy while extracting the information from the public cloud. Proposed framework is shown in Figure 4.3.

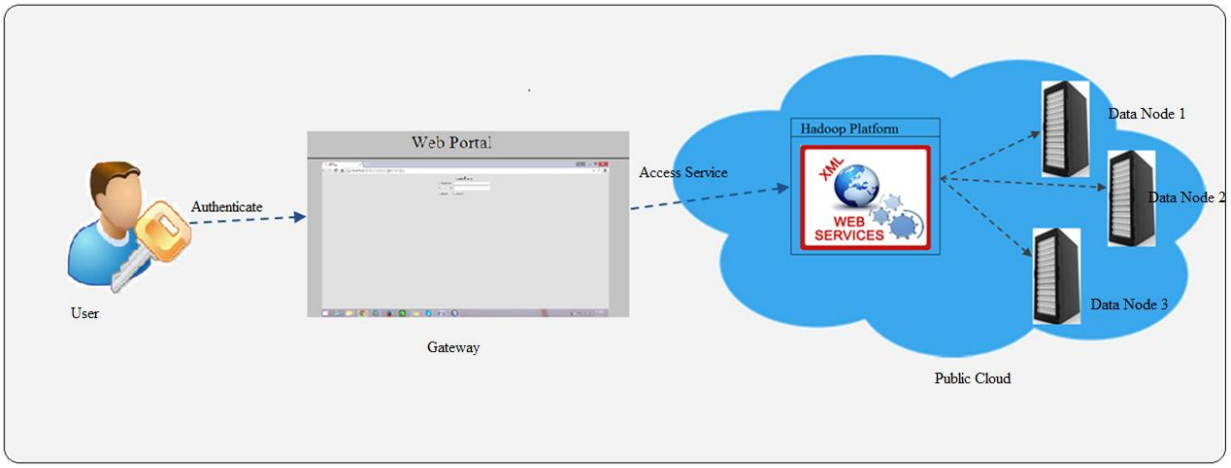


Figure 4.3: Proposed Framework

1. The data provider stores the data on public cloud to save the cost of cloud infrastructure. The data on public cloud is store in encrypted form to secure its accessibility as public cloud can be accessed by anyone.
2. A web portal is introduced as a gateway between the actual data consumer and data provider or public cloud. Indirectly only the authenticated user are allowed to extract the information from the public cloud and can extract the data only through a service.
3. When the data consumers asked for the desired data, the service is invoked by the web portal and the public cloud sends the desired information through the service so that privacy of the data can be ensured.
4. To speed up the search for the desired data parallel search as a service has been implemented on public cloud.

The proposed approach can be divided into two parts:

i) Gateway: The web portal of our proposed system will act as gateway so that secure access to the web service will be there. User are not allowed to access the public cloud directly. They will send the query to the web service, the service will then communicate with the cloud to search for the data required by the users, and it will reduce the overhead.

ii) Parallel Search Algorithm: When the user entry the query, then to search for the data required by them web service will invoked map-reduce, HDFS to search the files in parallel.

4.3 Parallel Search Algorithm

i) Keyword list generation: Firstly, we will extract the words str_i from the query Q_i enter by the user and them to an array A[i].

ii) Trapdoor Generation: Taking as input the master key(MK) and the words str_i from an array A[i], apply Trapdoor Generating Algorithm (TGA) to generate trapdoorT_i.

iii) Search Document: Now, by invoking map-reduce, HDFS through web service T_i will be searched on all the files present in cloud in parallel. If the T_i is found in any of the document, then it will return the document iddoc_{id} and all the document id's will be added in List[j].

iv) Identify: Finally, those doc_{id} will be selected which contain all theT_i . Then, those documents will be retrieved from public cloud and send to the user.

BEGIN

1-/*Keyword list Generation(Q_i , A[i])*/

Q_i ⟵ Obtain query from user

for all str_i in Q_i

    if(str_i == punctuation :: special character :: stop words)

        removestr_i

    end if

end for loop

for all str_i in Q'_i

    if(str_i not in A[i])

        str_i.add(A[i])

    end if

end for loop

2- /* Trapdoor Generating Algorithm */

k_p ⟵ generate random_key /*or MK⟵ generate master_key */

for all w_i in A[i] do

T_i ⟵ generate_trapdoor_TGA(w_i,k_p or MK)

end for loop

3- /* Search Documents*/

for all T_i

/* first find direct distance between nodes */

for nodes k ⟵1 to n do

C({i,k},k) ⟵ $d_{i,k}$

end for loop

for S ⟵ 2 to n do

for all S= {1,2....n} ||S|| = S do

for all k € S do

{C(S,k)⟵$min_{m\neq k, m\in S}$ [C(S-{k},m) + $d_{m,k}$] }

opt ⟵ min [C({1,2...n},k) ,$d_{i,k}$]

end for

end for

end for

return opt

ifT_i is found in k

doc_{id}.Add [List[j]]

end if

end for

4- /*Identify*/

for all doc_{id} in List[j]

if (doc_{id} contains all T_i)

returndoc_{id}

end if

end for

END

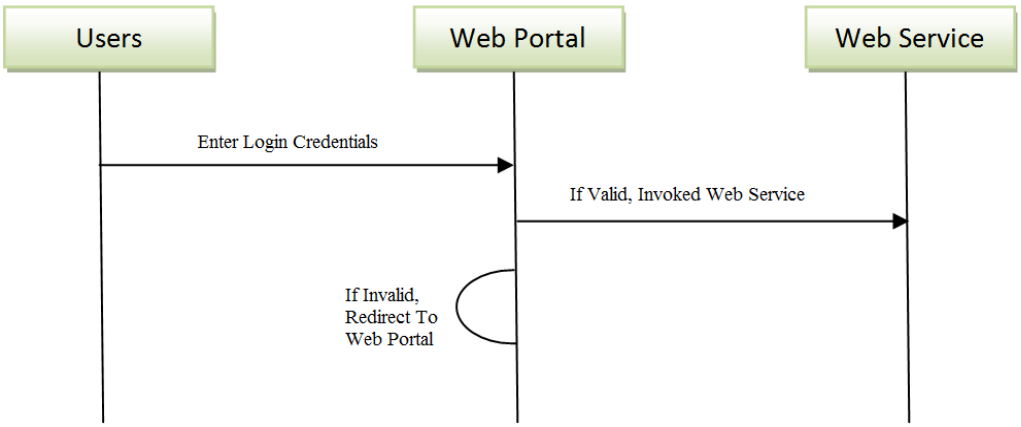
where,

Table 4.1: List of Symbols used in Algorithm

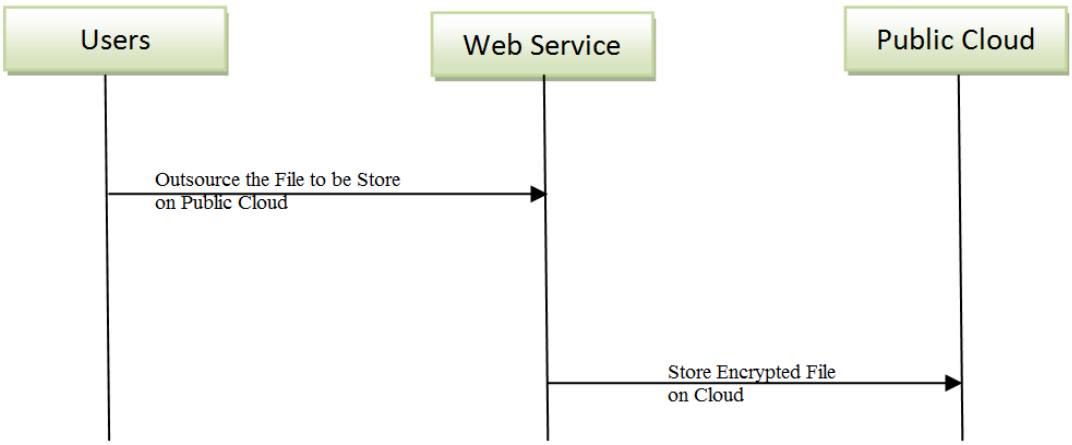
Symbol	Meaning
Q_i	query enter by user
Q'_i	refined query
str_i	words of which trapdoor has been generated
A[i]	array containing str_i
T_i	trapdoor of str_i
C(k,m)	distance between two nodes via other nodes
$d_{i,k}$	distance between consecutive nodes
Opt	optimal distance time between two nodes
doc_{id}	document containing all T_i
List[j]	list of doc_{id}

4.4Workflow of Proposed Framework

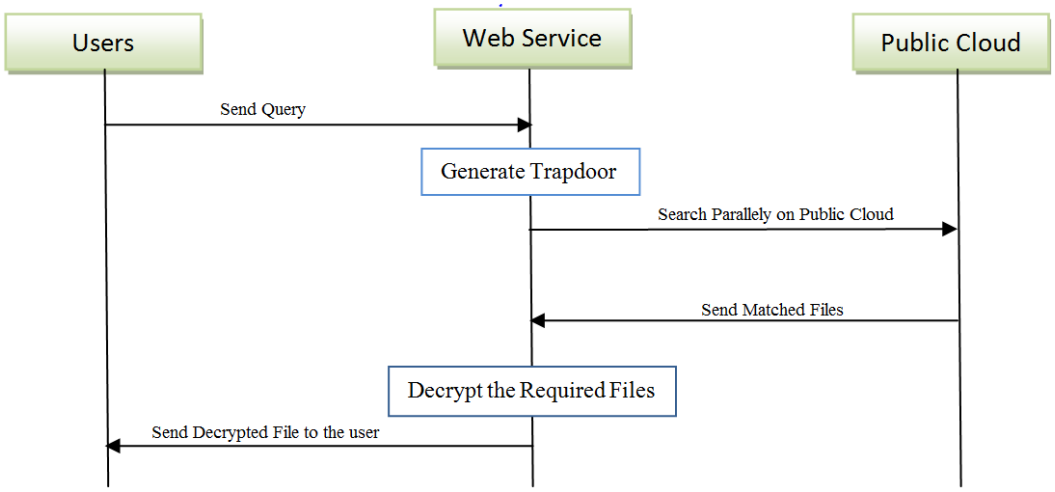
User will login through a secure web portal. Only authorized users will be able to access web service through which they can outsourced their data to the public cloud. Figure show the procedure in the proposed framework in which only the authorized user will e able to use service. User need to outsource their data to the web service. Web service will then encrypt the data and store on cloud. If a user want to retrieve some desired information then user need to enter the query , which then later used by web service to generate trapdoor so that required file can be search on public cloud. After the required file obtained web service then decrypt the file required by the user and then send it to the user. Figure 4.4 shows the work flow of the proposed framework.



a) Initialization



b) Submission



c) Query

Figure 4.4: Workflow of Proposed Framework

Chapter 5

Implementation and Result

To implement the proposed framework firstly multi-node Hadoop cluster need to be installed.

Hadoop is a java-based free programming framework that endorses processing of a large amount of data present in an distributed computing environment *i.e.* support application running on big data[32].Hadoop is inspired from Goggle map-reduce framework in which an application is broken down in small parts for fast and quick processing. It is named by Doug Cutting and Michael Cafarella in 2005. The main advantage of Hadoop is that it is scalable.

5.1 Steps to install Hadoop multi-node cluster

To install the multi-node Hadoop cluster on two Ubuntu machine, the most appropriate way is to install, configure and test Hadoop setup for each of the two Ubuntu machine "locally" and then "merge" these two single node cluster into multi-node cluster, in which one will act as a master and other as slave as shown in figure 5.1.

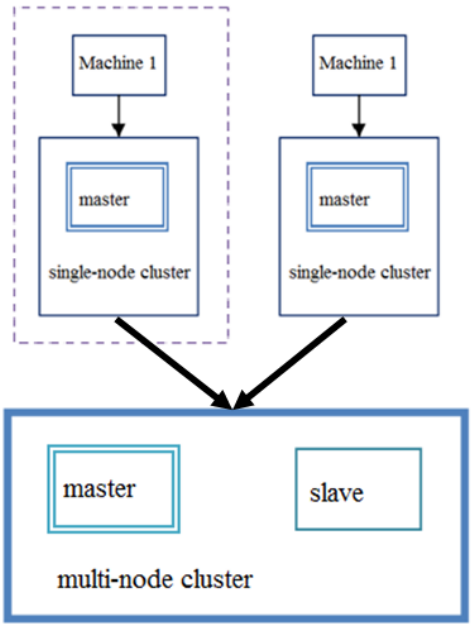


Figure 5.1: Multi-node Cluster Approach and Structure

i) Prerequisites

Configure single-node cluster on both machine - For this, the prerequisites are:

- JAVA
- Adding a dedicated Hadoop system user
- Configure SSH
- Disable IPV6
- Hadoop

For all the above steps see Appendix:A

- Networking- Networking plays an important in multi-node setup. As before merging the nodes it is important that nodes must ping each other and for that they need to be connected on the same network. After completing this, select one node as a master node and other as a slave node. For example, here 192.168.0.1 is selected as a master node and 192.168.0.2 is selected as a slave node. Add this on both machine master as well as slave in "/*etc*/hosts" file.

Figure 5.2: Editing /*etc*/hosts

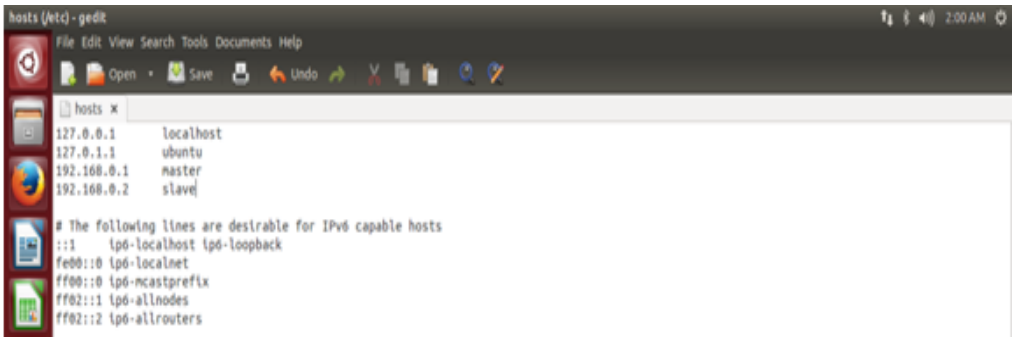


Figure 5.3: Adding nodes to /*etc*/hosts

- Enabling SSH - hduser on the master node need to be connect to
  a) its own master account user and
  b) the hduser account on the slave node via SSH login

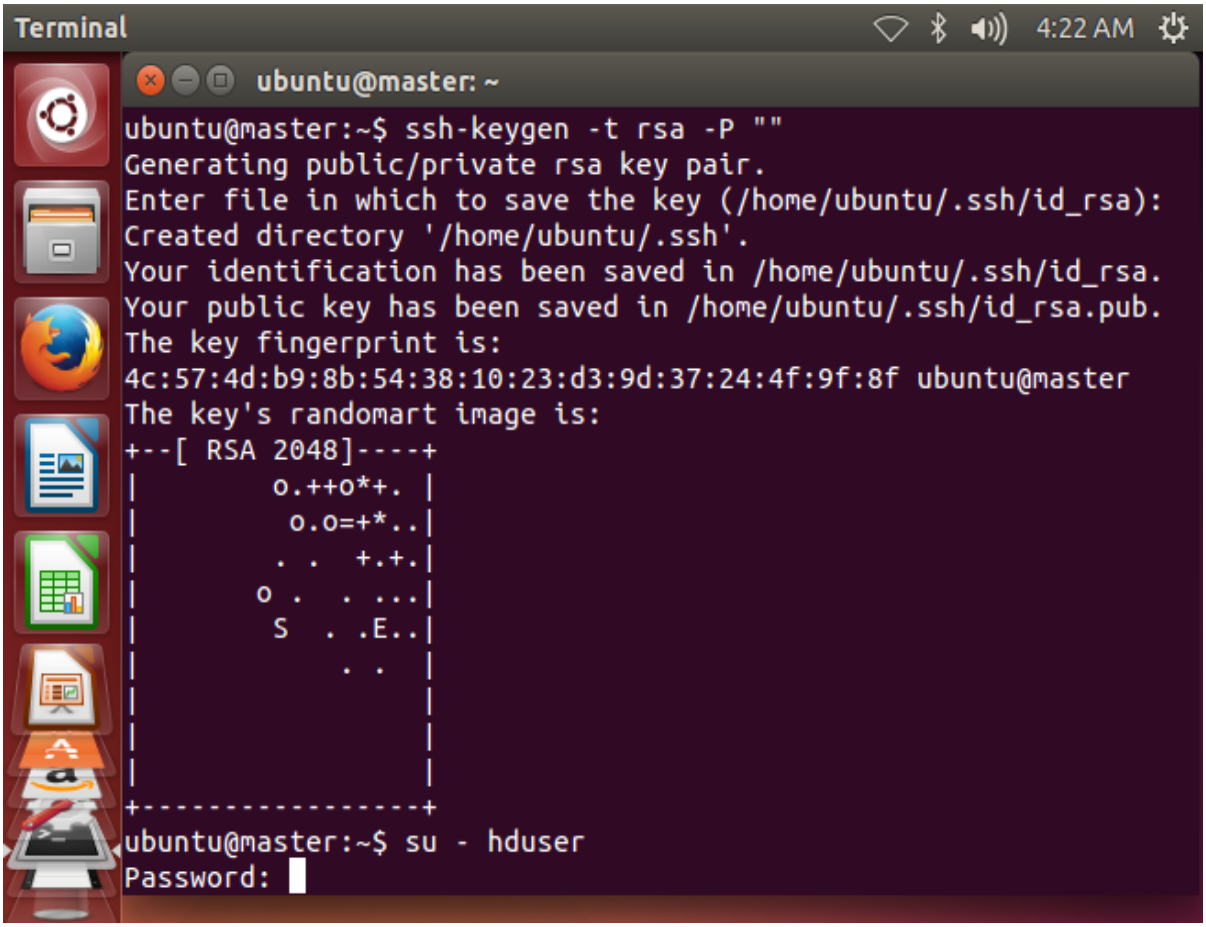


Figure 5.4: Generating SSH Key for User

a) connecting from master to master

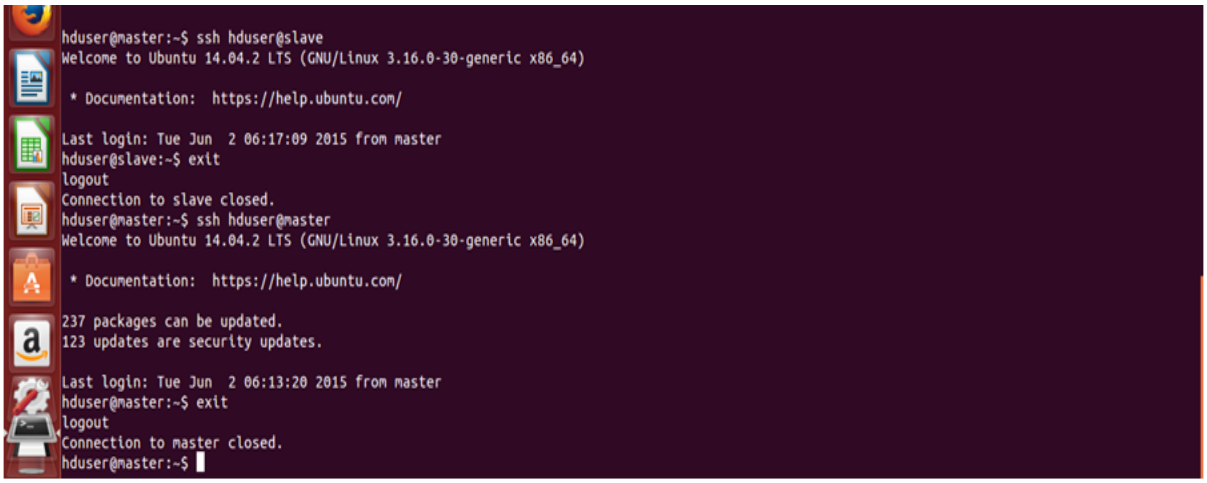


Figure 5.5: Connecting from Master to Master

b) connecting from master to slave

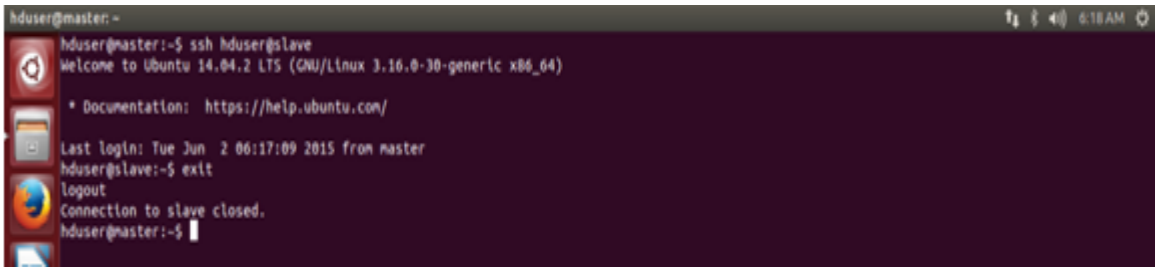


Figure 5.6: Connecting from Master to Slave

ii) Configuration

Now, configure the one Ubuntumachines as Master Node and other as Slave Node. Master Node will also act as slave node, Master node will run the "master" daemons for each layer *i.e.* Namenode and Job Tracker for HDFS storage and MapReduce processing layers respectively. "Salve" daemons will be run by both machine *i.e.* DataNode and TaskTracker for HDFS and Map-Reducing layer respectively. Responsibility of coordination and management is done by master node while slave node is responsible for data storage and data processing work. Figure 5.7 show muli-node cluster.

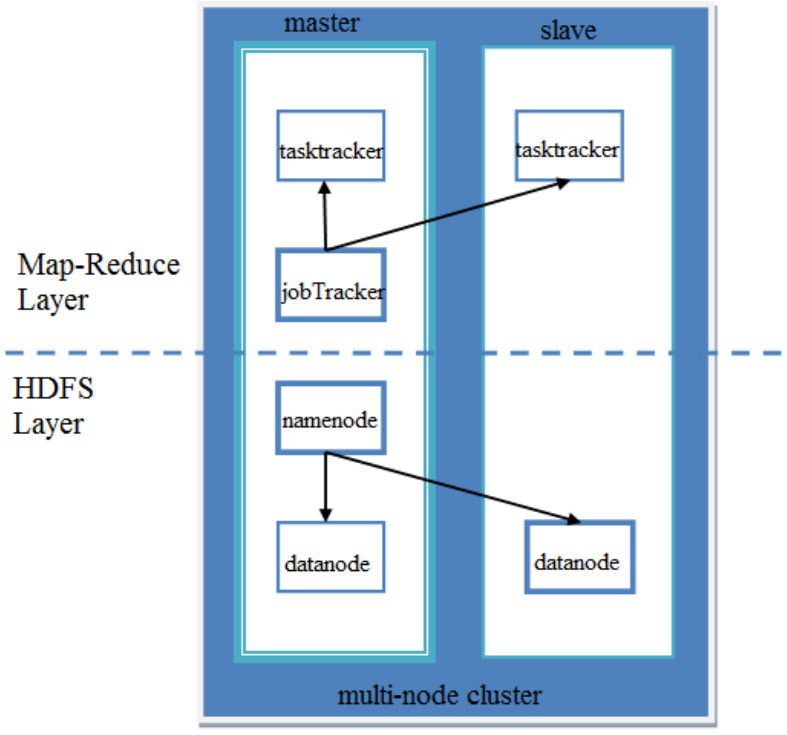


Figure 5.7 Multi-node Cluster

The required files which need to be configure are:

a) masters: In master machine, configure master file and add master Namenode .

Figure 5.8: Editing Master File

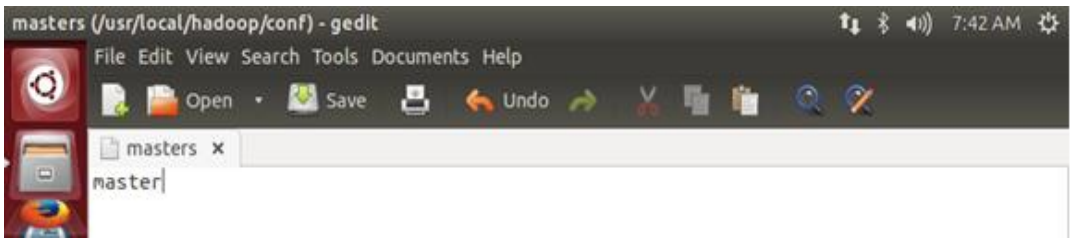


Figure 5.9: Adding master to Master File

b) slaves: Lists the hosts, one per line, where Hadoop slaves daemons will be run in master machine.

Figure 5.10: Editing Slave File

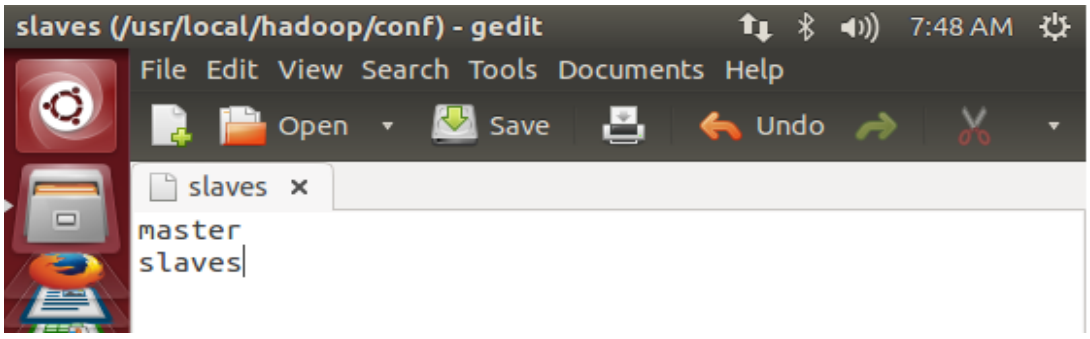


Figure 5.11: Adding master and slave to Slave File

c) core-site.xml: In core-site.xml file, change the hostname from localhost to master which determines the Namenode host and port on all machines.

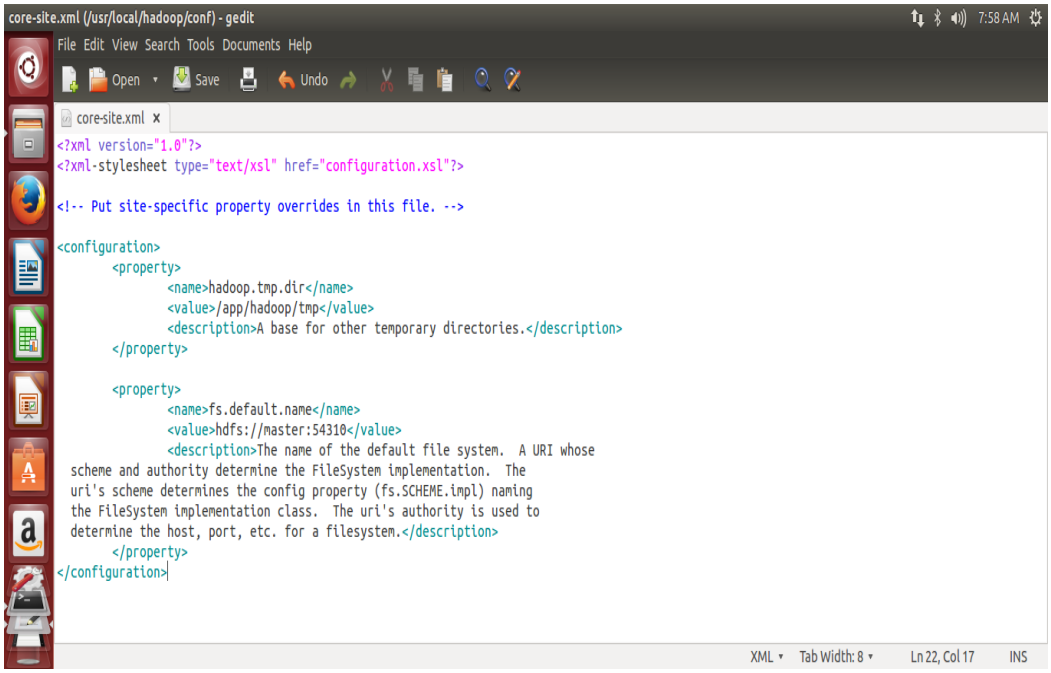


Figure 5.12: core-site.xml

d) mapred-site.xml: In mapred-site.xml, change the host name from "localhost" to "master" which specifies the Job Tracker host and port.

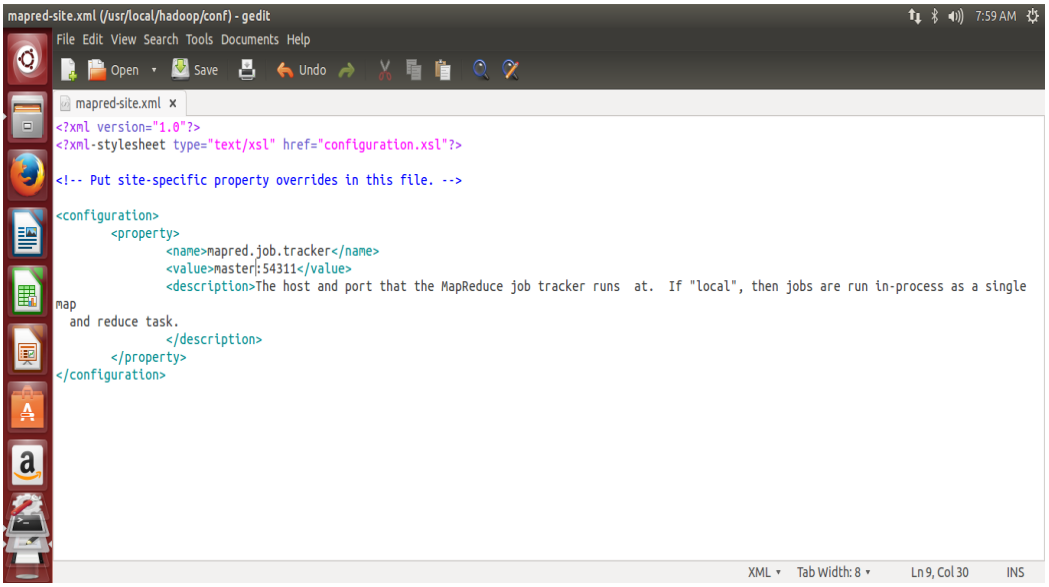


Figure 5.13: mapred-site.xml

e) HDFS-site.xml: In HDFS-site.xml, change replication factor from 1 to 2 as we have two nodes here.

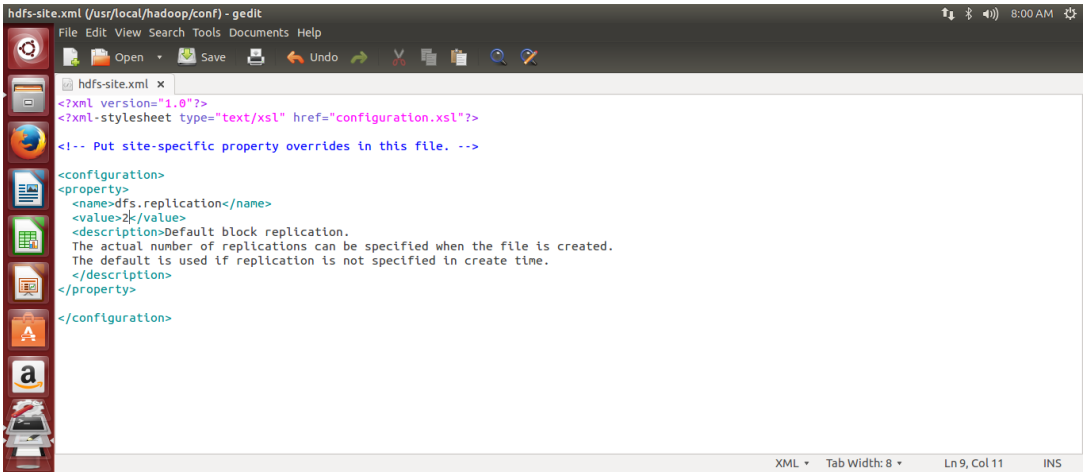


Figure 5.14: HDFS-site.xml

iii) Formatting HDFSfile system via Namenode

Run the following command, to format the file system in master node.

- /usr/local/Hadoop/bin/Hadoopnamenode -format

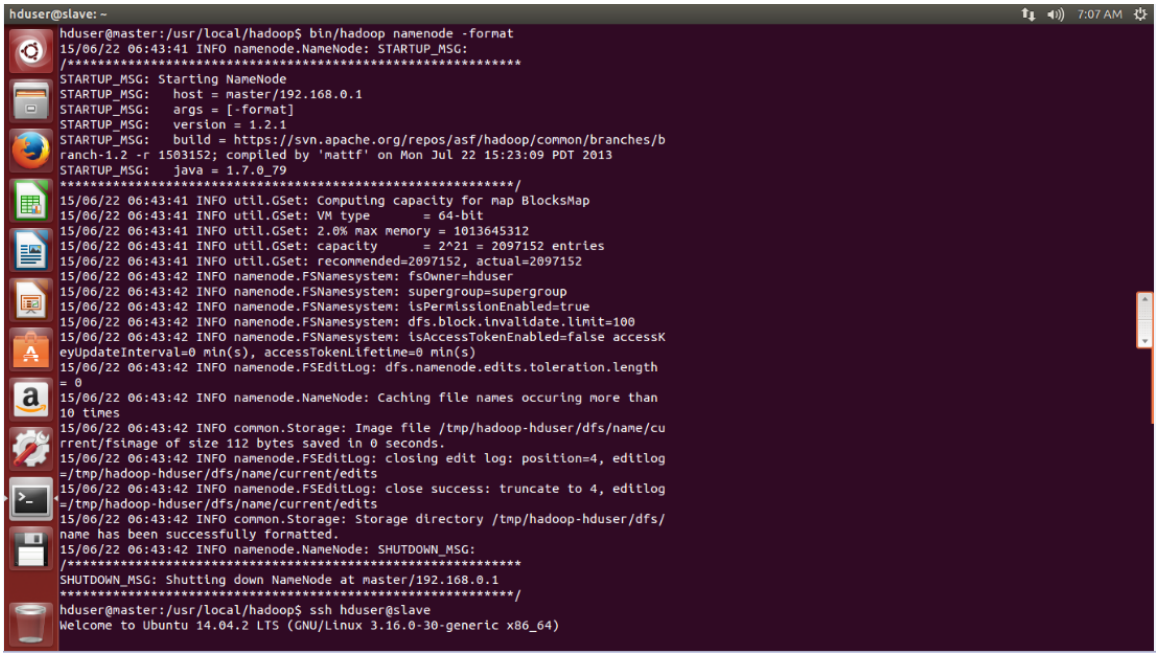


Figure 5.15: Formatting Master Node

Also format the file system in slave node using same command. For that first enter the slave node using following command:

- sshhduser@slave

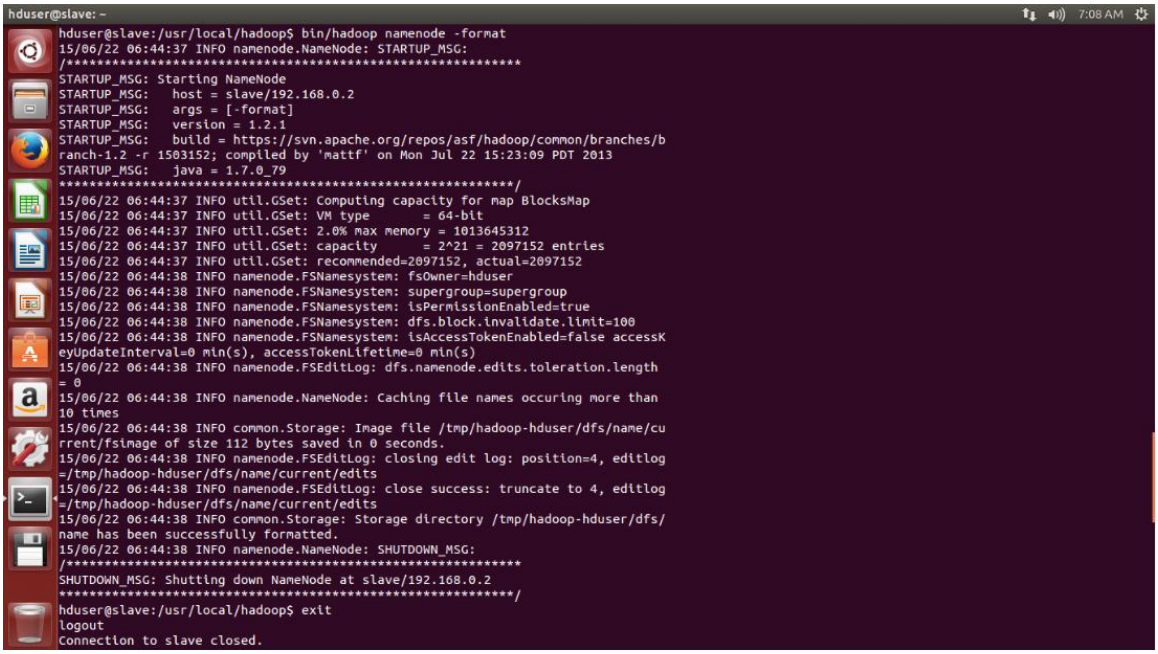


Figure 5.16: Formatting Slave Node

iv) Starting the multi-node cluster

Start multi-node cluster by running command

- start-dfs.sh
- start-mapred.sh

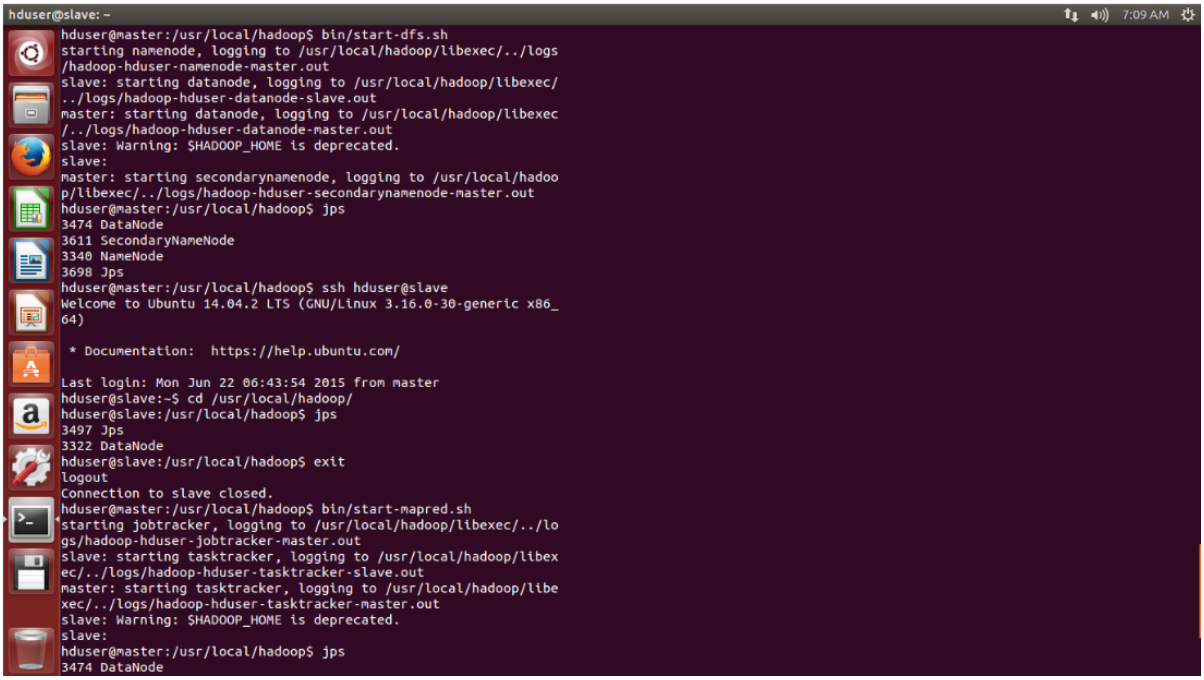


Figure 5.17: Starting Multi-node Cluster

To check daemons has started, browse web interface, by default they are available at:

- for Namenode- http://localhost:50070
- for JobTracker- http://localhost:50030
- for TaskTracker- http://localhost:50060

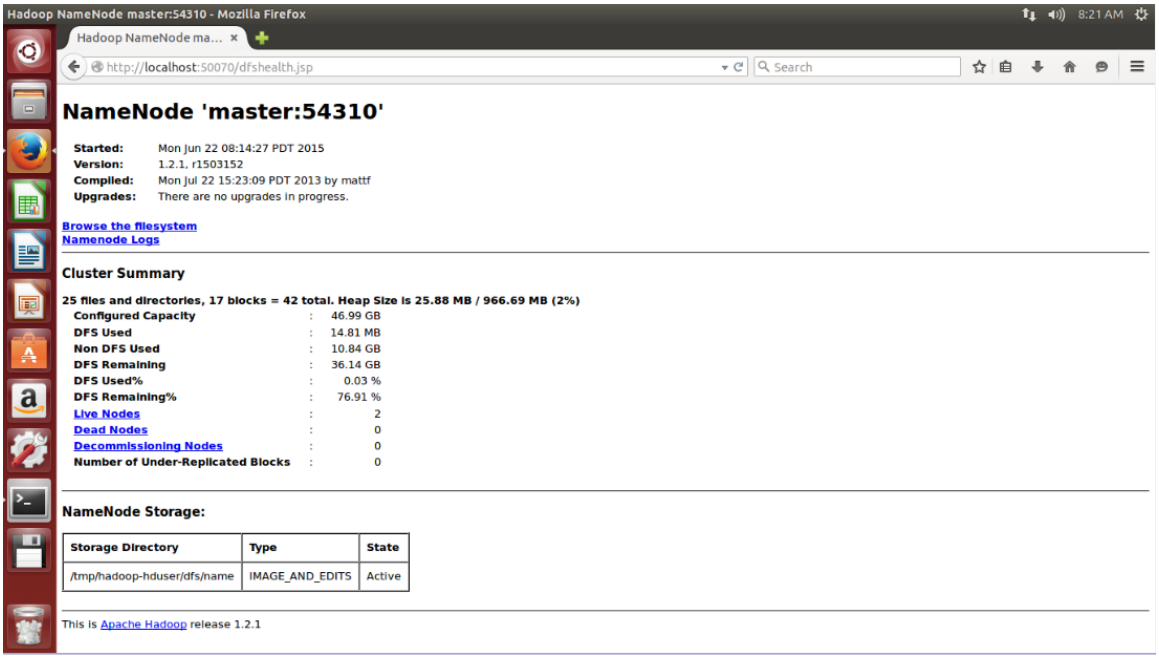


Figure 5.18: Namenode

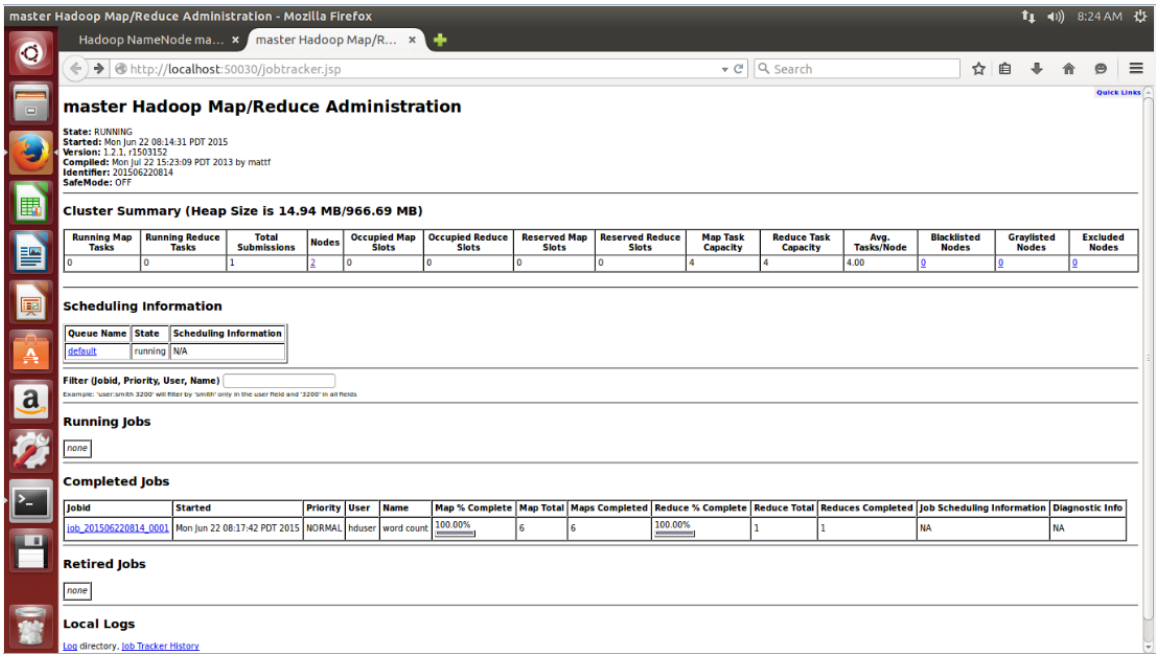


Figure 5.19: Job Tracker

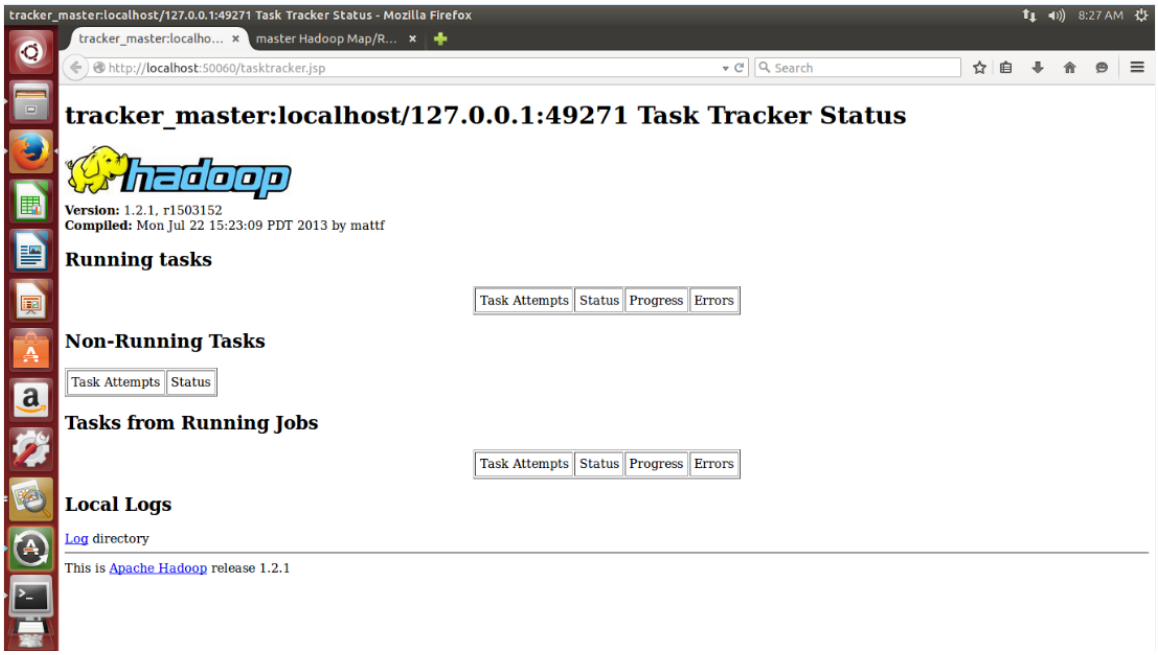


Figure 5.20: TaskTracker

v) Stopping the multi-node cluster

- stop-dfs.sh
- stop-mapred.sh

or

- stop-all.sh

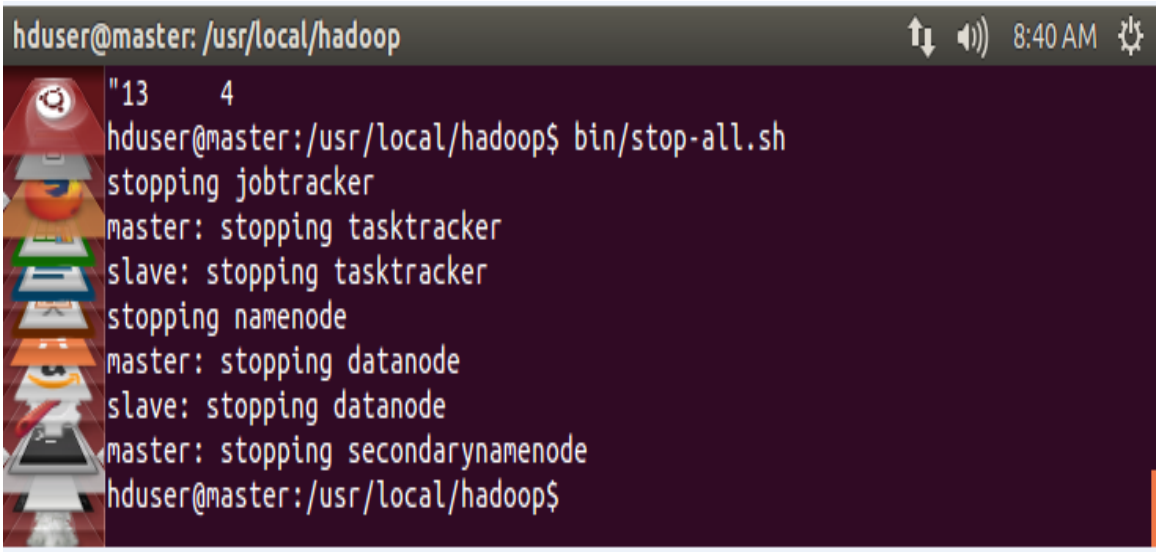


Figure 5.21: Stopping Multi-node Cluster

5.2 Results

5.2.1 Authentication Phase

a) Login Form

Registered users if want to use the service will enter the login credentials into the web portal. If the details enter by the user are verified, then only the user can proceed further and web service will be invoked. If the details are not verified, then the user will not be able proceed further and is notified accordingly.

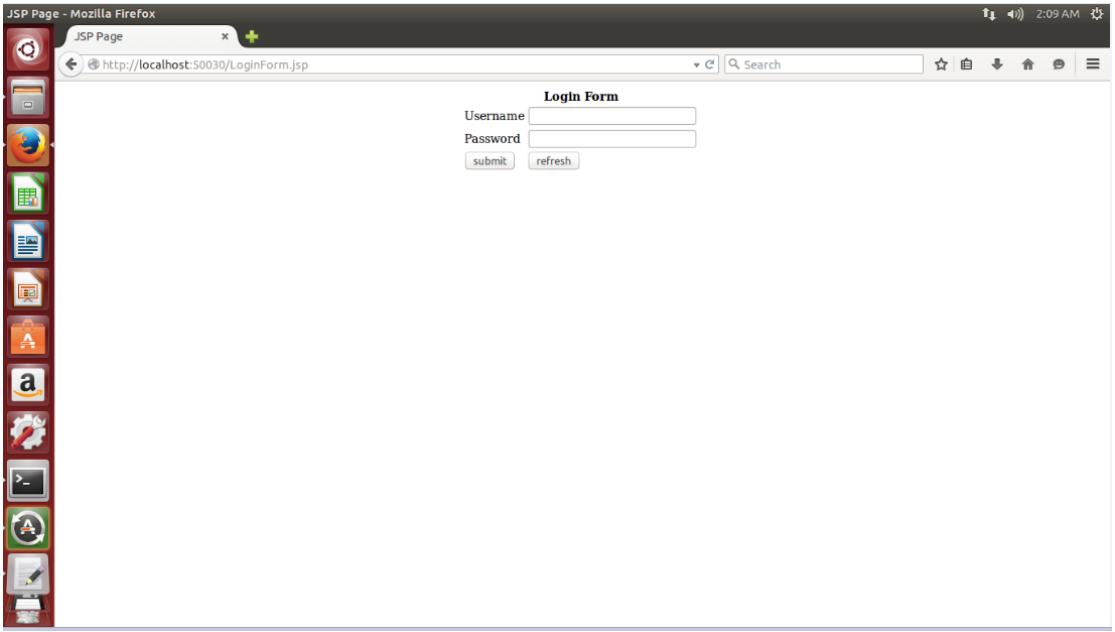


Figure 5.22: Login Form

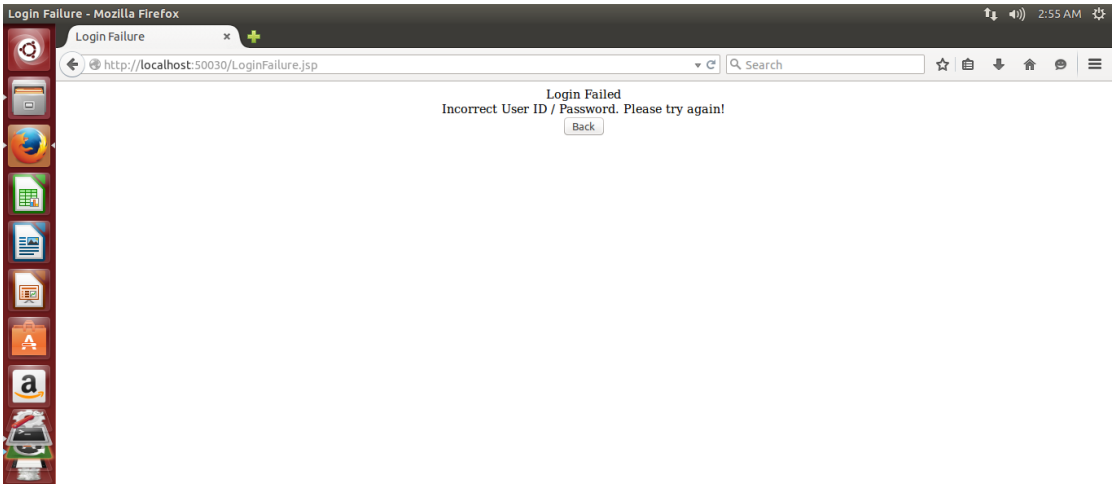


Figure 5.23: Login Failed

5.2.2 Operation Selection Phase

After the successful login, the user will be directed to the operation selection phase, where the user can either upload the data or download the required data. User will select the required operation accordingly.

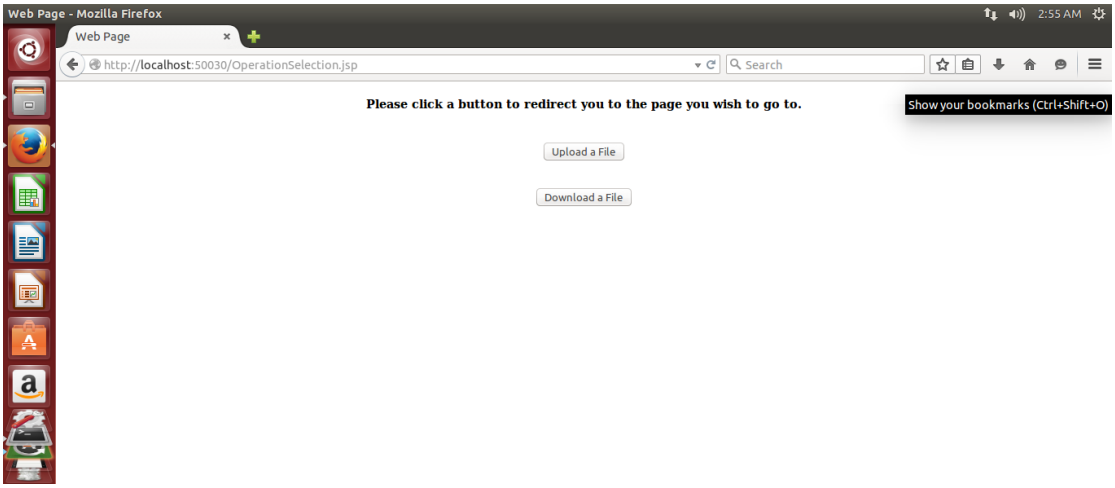


Figure 5.24: Operation Selection Form

5.2.3 Uploading phase

If a user want to upload a file, then user can select the file by browsing and then click on upload button to upload the file. When user click on upload button then the file will be uploaded in an encrypted form.

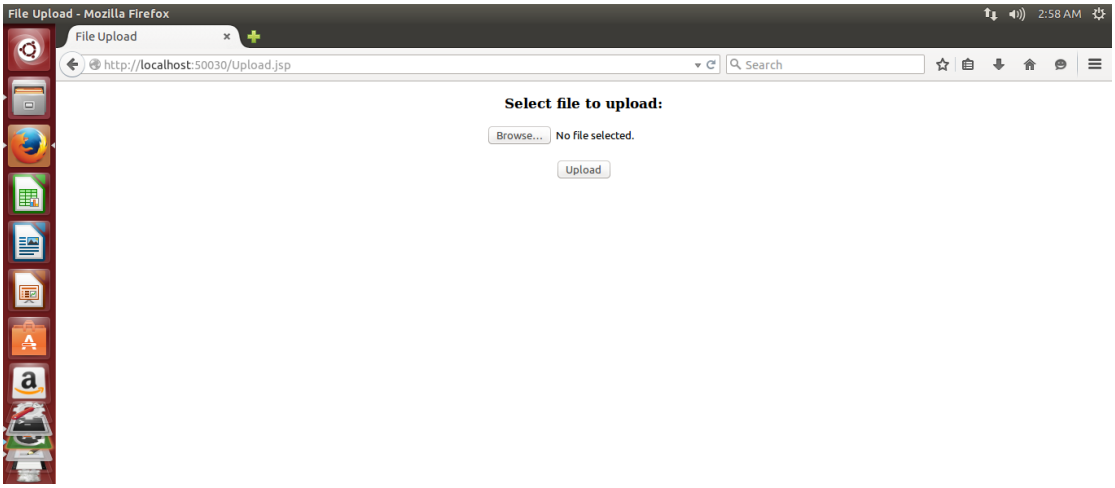


Figure 5.25: Upload Form

5.2.4 Downloading phase

In this phase, user will first enter the keyword want to search. Then, the files containing the keywords will be retrieve and then user select the file which want to

search and then download the file which required. Suppose, user enter the words 'user'. Then, firstly the word will be encrypted and parallel search algorithm will be executed to search the file containing keyword.

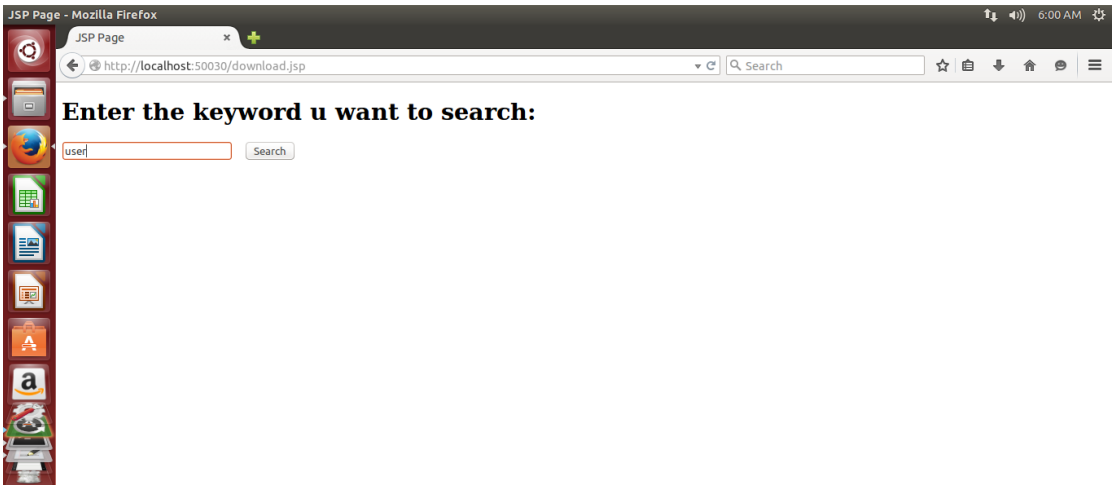


Figure 5.26: Enter the Keyword to Search(ex- user)

Node 1 contain a file named work.txt containing keyword user.

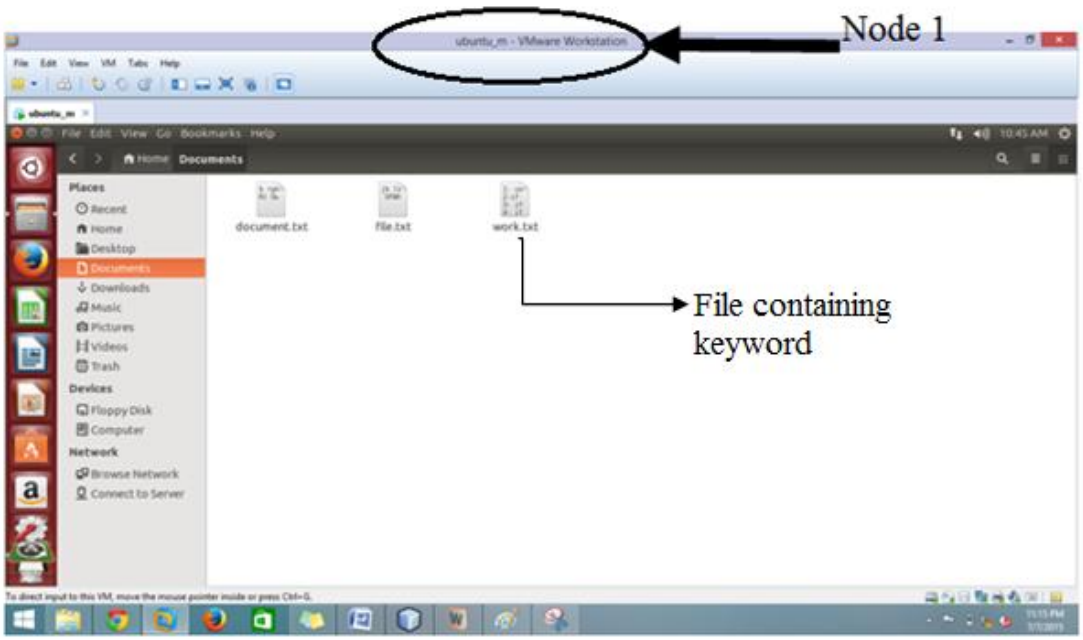


Figure 5.27: Node 1 containing Keyword

Node 2 contain a file named doc.txt which contain a keyword user.

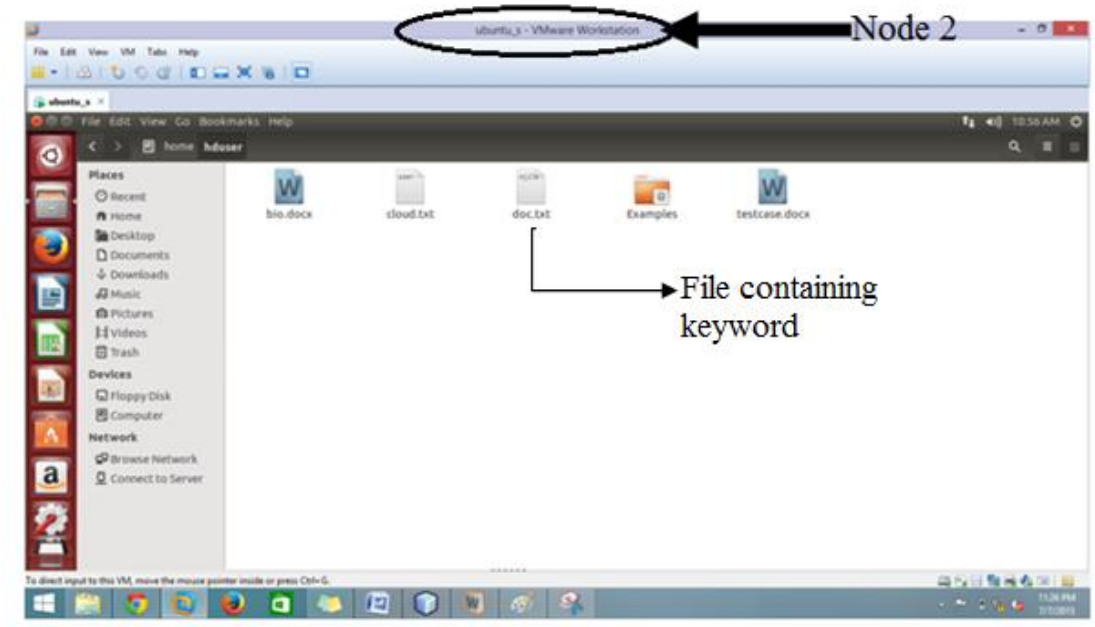


Figure 5.28: Node 2 containing Keyword

After that, all the files containing the keyword will be displayed and user can choose which file user want to download and click on file to download it. The file first decrypted and then downloaded.

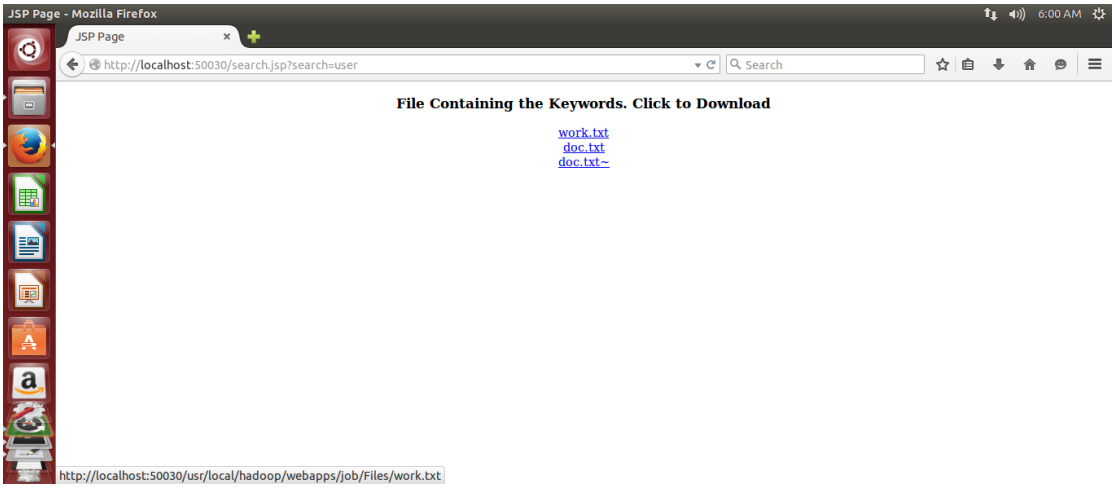


Figure 5.29: File Containing Keyword

5.2.5 Logout Phase

Once the user completed the process, then user finally disconnect the connection and can redirect to login page.

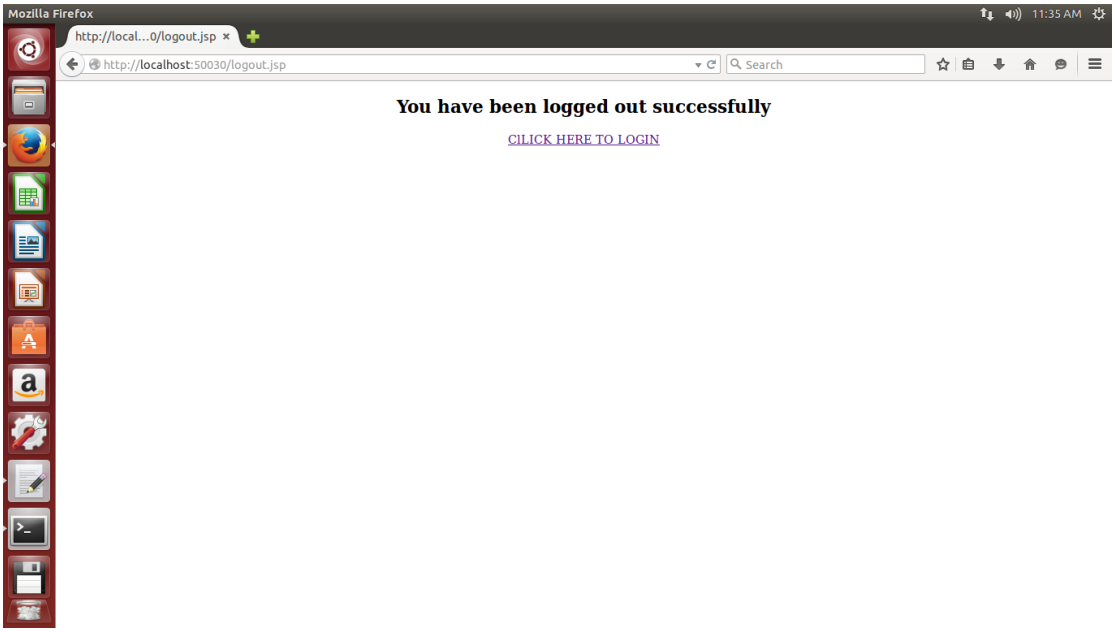


Figure 5.30: Logout Form

Chapter 6

Conclusion and Future Scope

This chapter discusses about the conclusion of the work presented in this thesis. The chapter ends with the discussion of the future scope.

6.1 Conclusion

In today's market the competition is driving the need for the technology, so that new features can often and quickly deliver in the market. In such position specifically, for small and growing organization setting up and maintaining data centers, dealing with physical infrastructure consume more IT cycle rather than in delivering new features in the market. Public cloud enable them to focus more on organization core value rather than on data centers or physical infrastructure. Public cloud even enable them to build new features product at very low capital investment. Organization does not have to worry about upgrading the data centers and structures to the latest technology, public cloud provider will take care of that. The established organization are also moving to public cloud because of its features like scalability , resources on demand so that they can consume the service provided by public cloud and deliver new features product with less investment in the market. All customers are not comfortable about their data being store in public cloud. As they do not know that where their data is store and who are able to access them. All the data are store on large data centers which are not fully trustworthy. This problem can be solved by storing the data in an encrypted form.The proposed framework will ensure that no unauthorized user can access the data. Only the authorize user will be able to access the data for that user will first login into the system then able to upload or download the data. In this framework, user will not be able to access the public cloud directly. All access will be done via webservice. All the data will be store on cloud in an encrypted format using an encryption algorithm. Here, Advance Encryption Standard encryption algorithm is used to encrypt and decrypt the data. User will enter the keyword which he/she want to search. Then, searching will be done using parallel searching algorithm so that searching will take less time. If found the file, then it can be downloaded in a format readable to user.

6.2 Future Scope

Future opportunities could explore following:

i) The proposed framework has been implemented on word and text files. This can be further enhanced to images and pdf files as well audio and video.

ii) Encryption algorithm can be improved further thereby ensuring more privacy of the data.

References

[1] "Public or Private Cloud: The Choice is Yours", AerohiveNetworks,Inc.(2013), May 1, 2013,<http://www.aerohive.com/pdfs/Aerohive-Whitepaper-Public-or-Private-Cloud.pdf>.

[2] VangieBeal, "What is Private Cloud? A Webopedia Definition", http://www.webopedia.com/TERM/P/private_cloud.html, May 2, 2015.

[3] Naveen Gabrani,"Private, public, hybrid clouds", http://thecloudtutorial.com/cloudtypes.html, May 2, 2015.

[4] "Types of Cloud Computing: Private, Public and Hybrid Clouds - Appcore ", http://www.appcore.com/types-cloud-computing-private-public-hybrid-clouds/, May2, 2015.

[5] Goran Candrlic "Types of Cloud Computing Explained | GlobalDots", http://www.globaldots.com/cloud-computing-types-of-cloud/, May 3,2015.

[6] "What is a Hybrid Cloud? | Interoute", http://www.interoute.com/cloud-article/what-hybrid-cloud.

[7] Mike Klein ,"Three Benefits of Public Cloud Computing", http://resource.onlinetech.com/three-benefits-of-public-cloud-computing/, May 3,2015.

[8] "What is a Public Cloud? | Interoute",http://www.interoute.com/cloud-article/what-public-cloud, May 3, 2015.

[9] Joe McKendrick, "12 reasons why public clouds are better than private clouds | ZDNet",http://www.zdnet.com/article/12-reasons-why-public-clouds-are-better-than-private-clouds/, May 5,2015.

[10] Antony Savvas,"The benefits of public cloud computing | ITProPortal.com",http://www.itproportal.com/2014/05/07/benefits-public-cloud-computing/, May 5,2015.

[11] RajkumarSelvaraj, "Top 10 advantages of Public Cloud Service For Start-Ups – Zombies Lounge", http://www.zombieslounge.com/2014/02/11/top-10-advantages-public-cloud-service-for-startups/, May 7, 2015.

[12] "Cornell Database Group - Privacy",

http://www.cs.cornell.edu/bigreddata/privacy/, May 5, 2015.

[13] D. Chen, and H. Zhao, "Data security and privacy protection issues in cloud computing", in *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on*. Vol. 1, pp.647-651. IEEE,2012.

[14] Margaret Rouse, "What is data privacy (information privacy)? - Definition from WhatIs.com",http://searchcio.techtarget.com/definition/data-privacy-information-privacy, May 6,2015.

[15] "ico-think-privacy-toolkit-charities.pdf",https://ico.org.uk/media/for-organisations/think-privacy/2586/ico-think-privacy-toolkit-charities.pdf, June 10, 2015.

[16] Jeff Leek, " Measuring the importance of data privacy: embarrassment and cost | Simply Statistics",http://simplystatistics.org/2013/07/01/measuring-the-importance-of-data-privacy-embarrassment-and-cost/, May 8,2015.

[17] Ashalatha R and VaidehiM.,"The Significance of Data Security in Cloud: A survey on Challenges and Solutions on Data Security", Banglore: International Journal of Internet Computing, 2012, pp. 15-18 [Online]. Available: http://interscience.in/IJIC_Vol1Iss3/15-18.pdf. Accesssed: 10-March-2015].

[18] Jaydip Sen," Security and Privacy Issues in Cloud Computing", Kolkata, Innovation Labs, Tata Consultancy Services Ltd. [Online]. Available: http://arxiv.org/ftp/arxiv/papers/1303/1303.4814.pdf. [Accessed: 8- March-2015].

[19] Larry Koved, Anthony Nadalin, NatarajNagaratnam and Macro Pistoia, "The Theory of Cryptography | The Purpose of Cryptography | InformIT" , http://www.informit.com/articles/article.aspx?p=170808, May 10, 2015.

[20] "Handbook%20of%20Applied%20Cryptography" https://notendur.hi.is/pgg/Handbook%20of%20Applied%20Cryptography.pdf , March 20, 2015.

[21] Prof Alan Woodward, "Alan Woodward: An Emerging Threat to Public-Key Encryption ",http://www.profwoodward.org/2012/01/emerging-threat-to-public-key.html, May10, 2015.

[22] Zhang Wei, Sun Xinwei and Xu Tao, "Data privacy protection using multiple cloud storages", in *Mechatronic Sciences, Electric Engineering and*

Computer(MEC), Proceedings 2013 International Conference ,pp. 1768-1772, IEEE, 2015.

[23] M. R. Aswin and M. Kavitha, "Cloud intelligent risk-Risk analysis and privacy data management in the cloud computing", In *Recents Trends In Information Technology (ICRTIT),International Conference,* pp. 222-227, IEEE, 2012.

[24] Ken Naganuma, Masayuki Yoshino, Hisayoshi Sato and Yoshinori Sato, "Privacy-preserving Analysis Technique for Secure, Cloud-based Big Data Analytics", *Hitachi Review* 63 no. 9 , pp. 50-56, 2014.

[25] S. Kuzhalvaimozhi, and G. Raghavendra Rao, "Privacy protection in cloud using identity based group signature", in *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the*, pp. 75-80. IEEE, 2014.

[26] D.X. Song, D. Wagner and A.Perrig, "Practical techniques for searches on encrypted data", in *Security and Privacy, 2000*. S&P 2000. Proceedings. 2000 IEEE Symposium on, pp. 44-55, IEEE, 2000.

[27] E.J. Goh, Eu-Jin, "Secure Indexes", in *IACR Cryptology ePrint Archive* 2003 (2003): 216.

[28] R. Curtmola,J.Garay, S. Kamara and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions", in *Proceedings of the 13th ACM conference on Computer and communications security*, pp. 79-88, ACM, 2006.

[29] D. Boneh, G. Di Crescenzo,R.Ostrovsky and G. Persiano,"Public key encryption with keyword search", in *Advances in Cryptology-Eurocrypt 2004*,pp. 506-522, Springer Berlin Heidelberg, 2004.

[30] P. Golle, J. Staddon and B. Waters, "Secure conjunctive keyword search over encrypted data" in *Applied Cryptography and Network Security*, pp. 31-45, Springer Berlin Heidelberg, 2004.

[31] C. Wang, N. Cao, J. Li and W. Lou, " Enable Secure and Efficient Ranked Keyword Search over Encrypted Cloud Data" in*Parallel and Distributed Systems,IEEE Transactions* Volume 23, Issue 8, pp.1467-1479, Aug, 2012.

[32] N. Cao, C. Wang, M. Li, K. Ren and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," in *Parallel and Distributed Systems, IEEE Transactions*, pp. 222-233, 2014.

[33] J. Li, J. Li, X. Chen, C. Jia and Z.Liu, "Efficient keyword search over encrypted data with fine-grained access control in hybrid cloud", in *Network and System Security*, pp. 490-502 , Springer Berlin Heidelberg, 2012.

[34] J. Li, C. Jia, J. Li and Z.liu, "A novel framework for outsourcing and sharing searchable encrypted data on hybrid cloud", in Intelligent Networking and Collaborative System (INcoS), 2012 4th International Conference on , pp.1-7, IEEE, 2012.

[35] JingweiLi, Jin Li, XiaofengChen, Zheli Liu, and ChunfuJia, "Privacy-preserving data utilization in hybrid clouds," in *Future Generation Computer Systems* 30, pp. 98-106, Elsevier, 2014.

[36] Margaret Rouse, "What is Hadoop? - Definition from WhatIs.com", http://searchcloudcomputing.techtarget.com/definition/Hadoop, Jan 20,2015.

[37] Michael G. Noll, "Running Hadoop on Ubuntu Linux (Multi-Node Cluster) - Michael G. Noll",http://www.michael-noll.com/tutorials/running-Hadoop-on-Ubuntu-linux-multi-node-cluster/, Feb 10, 2015.

A:Hadoop installation

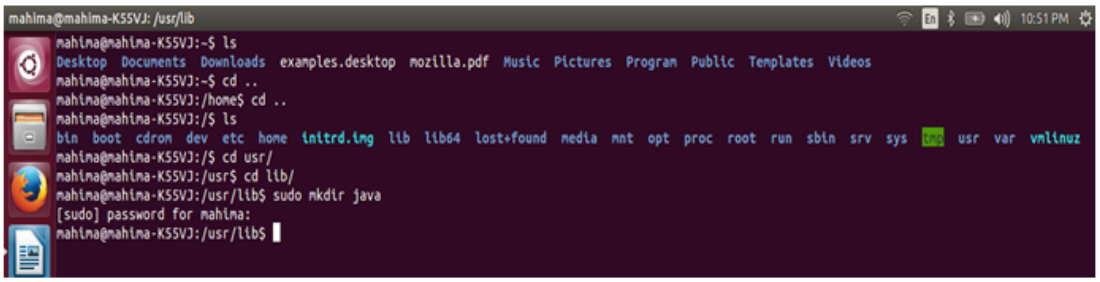
Step-1: Download jdk and extract files

Step-2: All libraries of any software resides in "/usr/lib" so move all jdk libraries in this folder

so that linux operating system can detect all java libraries automatically.

For this, create directory at "/usr/lib" using following command

-> sudo mkdir java



Step-3: Move extracted jdk folder to this java folder using command

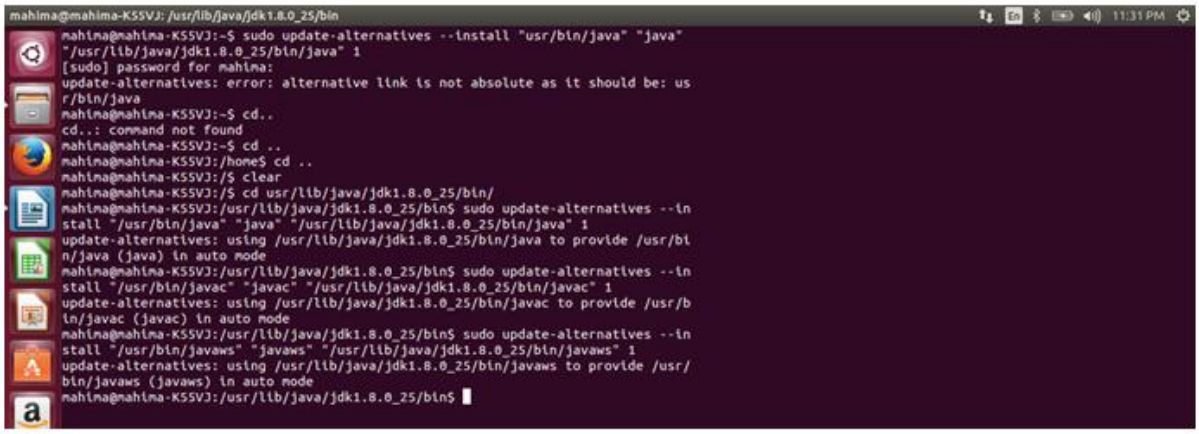
-> sudo mv jdk1.8.0_25/ /usr/lib/java

Step-4: Now, install java by following commands:

->sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/java/jdk1.8.0_25/bin/java" 1

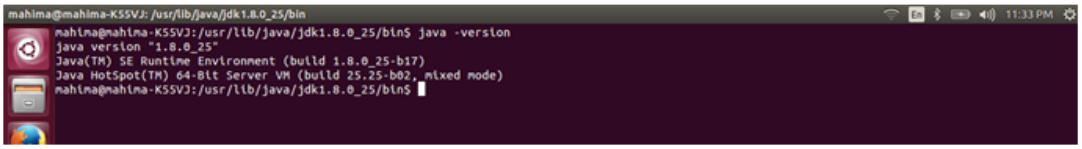
->sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/lib/java/jdk1.8.0_25/bin/javac" 1

->sudo update-alternatives --install "/usr/bin/javaws" "javaws" "/usr/lib/java/jdk1.8.0_25/bin/javaws" 1



To check which java version is installed use following command

->java -version



Step-5: Hadoop environment searches for java environment when run, so set environment variable in "/.bashrc" file.

Open "./bashrc" file and write following:

->gedit ~/.bashrc

#JAVA_HOME directory setup

export JAVA_HOME="/usr/lib/java/jdk1.8.0_25"

set PATH="$PATH:$JAVA_HOME/bin"

export PATH

Step-6: Now, install Hadoop.

First download hadoop and extract files.

Step-7: Now, set hadoop environment similarly java environment in "/.bashrc" file.

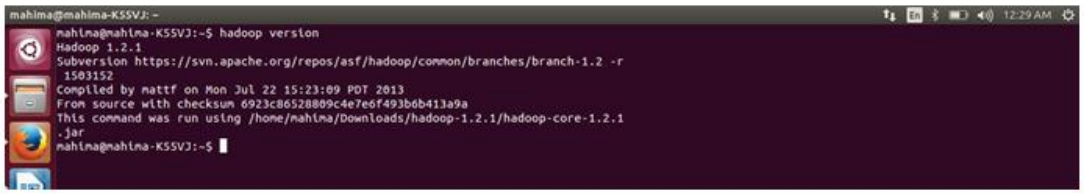
export HADOOP_HOME="Downloads/hadoop-1.2.1"

PATH=$PATH:$HADOOP_HOME/bin

In "hadoop-env.sh" file which is present in conf directory of hadoop enable JAVA_HOME and set java path

Check for hadoop version using command

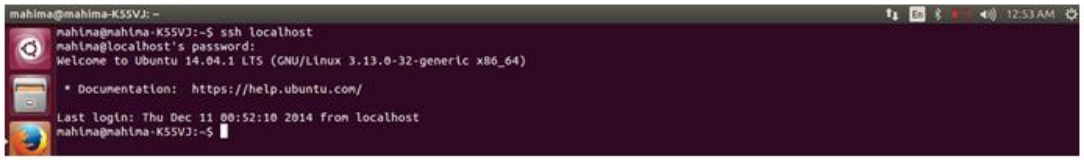
-> hadoop version



Step-8: Hadoop requires ssh access to manage its node, therefore need to configure ssh access to local host. So, install secure shell using command

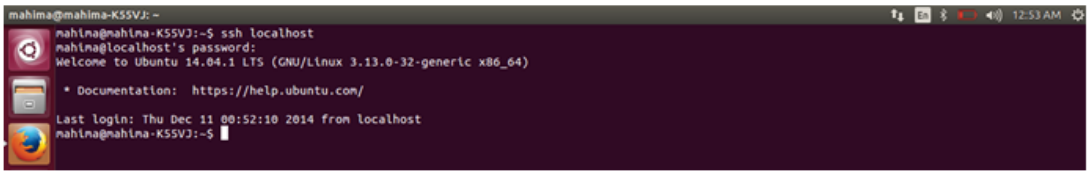
-> sudo apt-get install ssh

-> ssh localhost

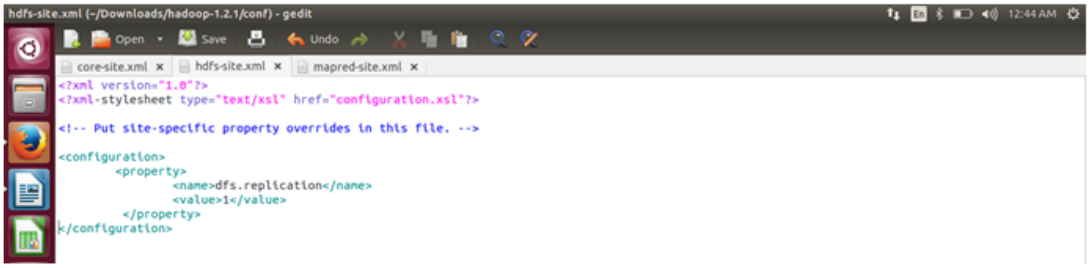


Step-9: Configured hadoop configuration files-core-ste.xml, hdfs-site.xml and mapred-site.xml

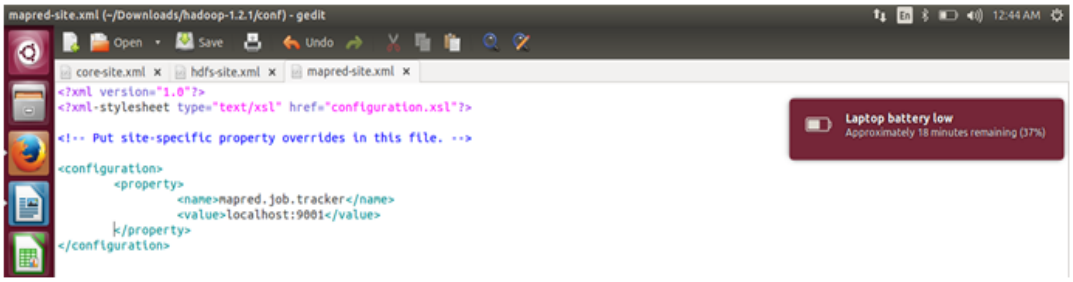
core-site.xml



hdfs-site.xml

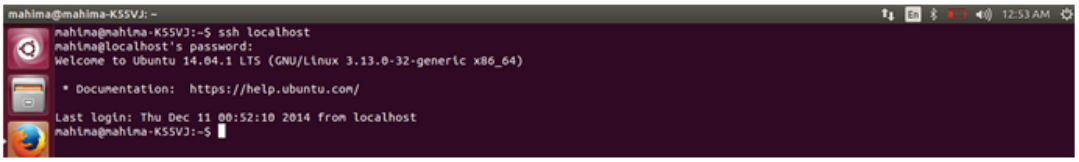


mapred-site.xml



Step-10: Connect to localhost using

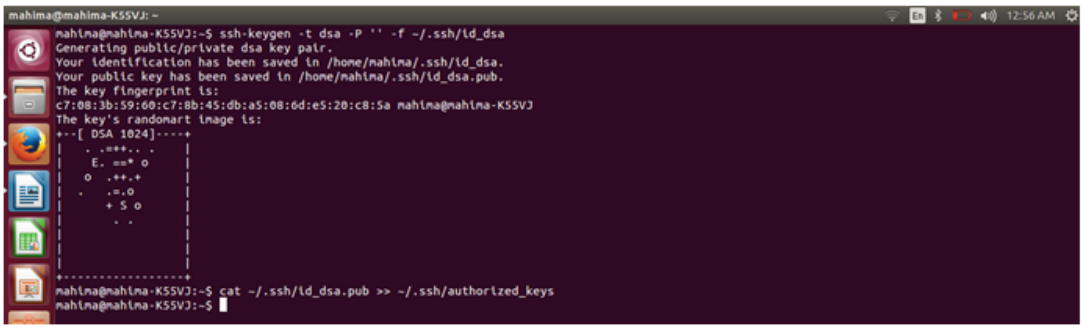
-> ssh localhost



While connecting to localhost we always need to write password, to make is password less execute following command

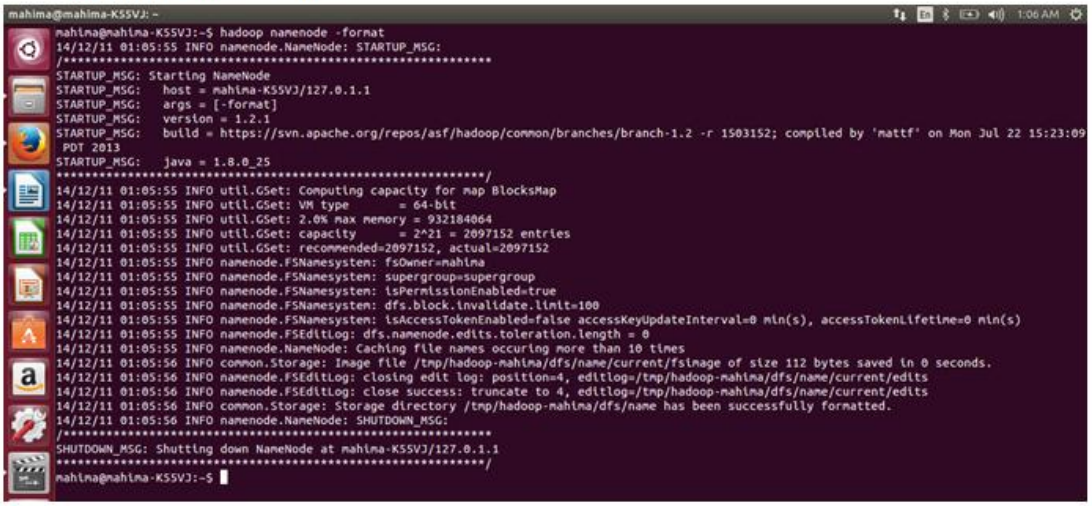
-> ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa

-> cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys



Step-11: Before starting the hadoop daemons for the first time only, need to format the namenode, as it will tell the operating system to provide some namespace for internal purpose, using following command:

-> hadoop namenode -format



Step-12: Start hadoop daemons individually using command:

-> start-dfs.sh

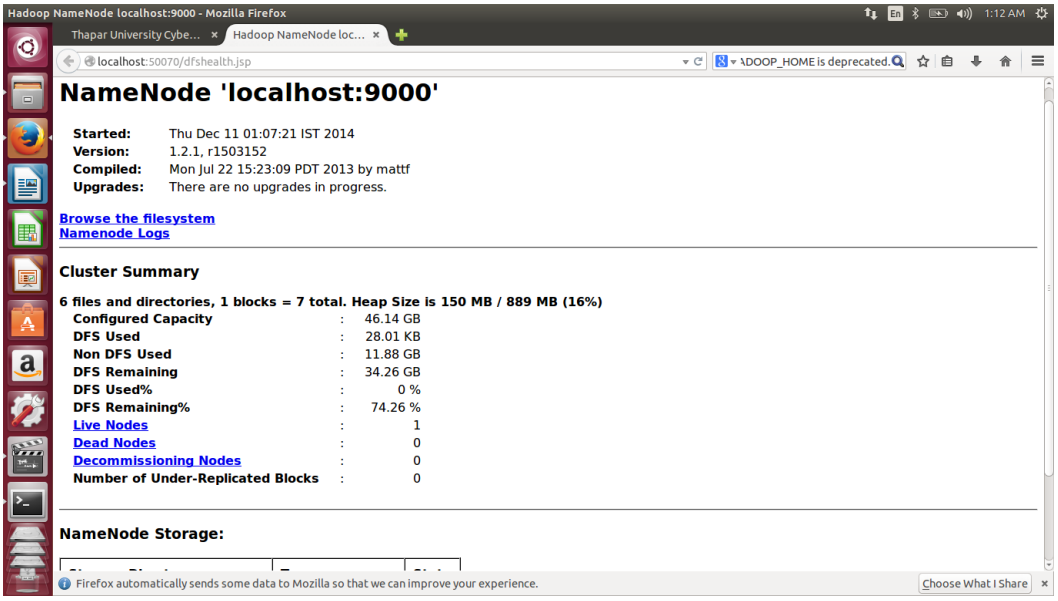
-> start-mapred.sh

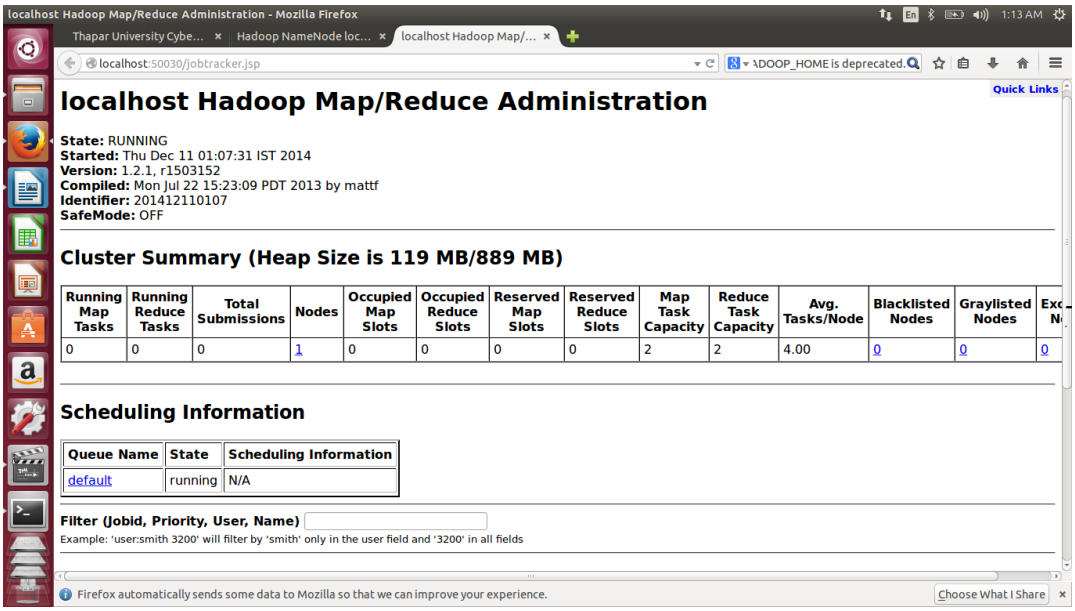
To check daemons has started, browse web interface, by default they are available at;

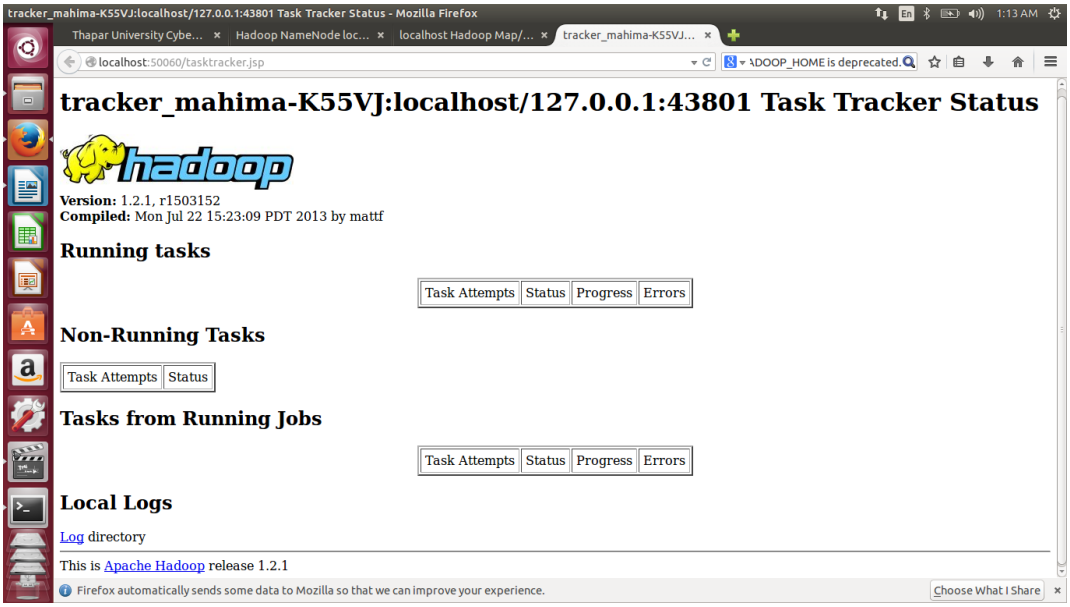
- for NameNode- http://localhost:50070

- for JobTracker- http://localhost:50030

- for TaskTracker- http://localhost:50060







Step-13: To stop all hadoop daemons, execute following command:

-> stop-all.sh

Communicated

Mahima Gupta and Dr. Damandeep Kaur, "Efficient Data Retrieval While Maintaining Privacy of Data", in *IEEE International Conference on Communication Control & Intelligent System(CCIS),* at GLA University, Mathura,2015.

YouTube Link

https://www.youtube.com/watch?v=WEtsREoAVC4&feature=youtu.be