EMAIL DATA DE-DUPLICATION SYSTEM

A Final Project

Presented to

The Faculty of the Department of General Engineering

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Engineering

By

Sharanjeet Hundal

Tanveer Singh

Basavasai Konuru

May 2012

SAN JOSÉ STATE UNIVERSITY

The Undersigned Committee Approves the Final Project Titled

EMAIL DATA DE-DUPLICATION SYSTEM

By

Sharanjeet Hundal

Tanveer Singh

Basavasai Konuru

APPROVED FOR THE DEPARTMENT OF GENERAL ENGINEERING

_____

Dr. Leonard Wesley, Department of General Engineering Date

_____

Prof. Morris Jones, Department of Electrical Engineering Date

_____

Mr. Surendra Gutlapalli, Brocade, Inc Date

APPROVED FOR THE UNIVERSITY

_____

Associate Dean Office of Graduate Studies and Research Date

ABSTRACT

EMAIL DATA DE DUPLICATION

By

Tanveer Singh

Sharanjeet Hundal

Basavasai Konuru

Data De Duplication is one of the hottest topics in the backup storage systems and Data De-duplication is a way of detecting and eliminating duplicate data and also optimizes network bandwidth. Data De-duplication technique is applied mainly in the permanent storage devices such as backup recovery and data centers. There are many areas from startups to large enterprises that maintain data centers such as financial, Educational, pharmaceutical, Information Technology. Data De-duplication provides lot of benefits to these companies who maintain Data centers in the form of cost. There are two different places where Data De-duplication is done and there are lots of ways to implement Data De-duplication.

Our Company Ultimate Solutions provides the best place and best technique to save more disk space by eliminating large amount of redundant data compared to the other companies that provide De duplication services. Our Services will provide benefits greatly to small to large companies that need Data De-duplication benefits. Our service includes maintaining the backup data for the small companies and providing Data De-duplication software for midsize and large enterprises.

ACKNOWLEDGEMENT

We would like to express our appreciation and gratitude to our beloved committee members for their continuous support throughout the project.

We would like to thank and express our respect to our Prof. Morris Jones, Dept. of Electrical Engineering, San Jose State University for guiding, and helping us to achieve our goal.

We would like to thank Surendra gutlapalli, Senior Engineer at Brocade for his guidance, support, and direction in project.

Our sincere thanks to Prof. Leonard Wesley, Associate Professor, San Jose State University for providing us an opportunity to do our project and we would like to extend out thanks to our family members for giving us this wonderful opportunity, support, help, and encouragement.

Tanveer Singh

Sharanjeet Hundal

Basavasai Konuru

Table of Contents

Table of Figures

1. Introduction

Data De-duplication is the hottest and important advanced technology, which allows customers to scale their storage capacity outside traditional storage such that present-day storage system provides different methods of data reduction and compression techniques that saves large amount of disk space by removing redundant data on the disk. There are several benefits of using Data De-duplication to the medium and large organizations, which saves more cost and more over no one is providing Data De-duplication services to the start-up companies because they have less data to take care and the software license is very expensive to buy.

We at Ultimate Solutions try to fill this vacuum in providing software for mid-size and large companies and services to small companies, which has extremely large potential. Based on our research from different sources by independent groups has verified to the fact that this area has a huge potential, which is growing steadily. Based on the product and services that we are working on right now, we will be targeting all the small, midsized and big companies and industries including the global money banks and leading financial services firms, healthcare and life sciences organizations, manufacturers, airlines, Educational institutions and transportation companies.

Ultimate Solutions provide different services for Startup and large enterprises as follows.

1) Data Back-up and Recovery

2) Software Application

3) Consulting Services

Ultimate Solutions will provide data storage services to the customers considering all security related issues to data storage. We will be storing data at our storage location farm for smaller companies because maintaining data will be expensive for startups. As large companies does not want others to maintain their data taking this into consideration we develop software and license software and provide support and training for the customers.

We believe in establishing the best customer relations and satisfying their demands within the promised time. Our Company is doing something, which is not present in the Data De-dupe market right now and we hope that once this development phase is over, we can definitely attract all the big and small fishes to our Ultimate Solutions. We are going to provide the ARCHIVING service, which is to gain a cost- effective, online archive for our client's organization's non-changing data assets so that we can minimize the risk, maintenance and control costs and eventually increase the content reuse.

According to our research based on research done on the De duplication, we will break even in the last quarter of the third year. For each and every penny, the investor spends, will be returned in a period of 4 years. The relatively unexplored market segment and the continuous growth of companies adopting our new technique model promise well for our business. In addition, this will help the small companies to concentrate on their core business.

Our project provides great knowledge about what it takes to start a consulting company in the Archiving and data storing industry. Complete idea about the organization, human resource, and other features of starting a business has been provided. Moreover, our estimated break even, return of investment, Cash flow and other financial details can be found, which will provide information for somebody planning to start a related business.

1.1 Scope

Our project provides with a good idea of details by setting up a consulting as well as software providing company. This project report gives an idea, which the people can be able to get an inspiration of how to proceed with providing De duplication services, Moreover we have also dealt with what will be a good market to concentrate based on the market estimation. Other business plans are also provided. We have also arrived at a possible break-even point, ROI and other economic factors, which go into evaluating a business idea.

2. Literature Review

After doing some research and studying different definitions of Data De-duplication from different journals and research papers, we found that most of them have backup, recovery keywords in common. The De-dupe method that we are proposing for our business will also include all this keywords.

To understand our business model and implementation, of data de-duplication technique minimum understanding of these keywords is important. In the literature review, we will cover this and other important terms related to De duplication.

2.1 Introduction to Data De-duplication:

With the development of information technology and computer networking, there is a rapid increase in the size of the datacenters .As Information Technology managers and executives are looking at increasing business throughput and steadiness, data availability became high priority. Traditional tape backup had replaced with disk-to-disk backup to improve the backup and recovery process and also improve operating efficiency. But

companies moving to disk-to-disk backup have encountered with unpredictable data growth, increasing costs of storage for data backup and data recovery. Every time whenever backup is performed, it has many duplicate files and also data is backed up, with multiple copies of similar data, which takes more disk space. Moreover energy consumption ratio of IT companies spending is increasing. In the great green environment many companies are eyeing the green store, hoping thereby to reduce the energy storage system.

Data de-duplication is a cutting-edge technology that can reduce the amount of data backup stored by removing duplicate data. Data de-duplication increases storage use allowing information Technology (IT) to maintain more backup data for a longer time. This enormously increases the efficiency of backup disks, replacing the method of how data is protected and also data de-duplication technology optimizes the storage system can greatly reduce the amount of data, thereby reducing energy consumption and reduce heat emission. Data compression can reduce the number of disks used in the operation to reduce disk energy consumption costs.

Data de-duplication compares new data with existing data from last backup process, and eliminates the redundant data. Data de-duplication is a technique that works by comparing blocks of data or files in to detect redundancy. At Ultimate Solutions we provide, data de-duplication to companies to reduce storage costs and minimizes network bandwidth.

2.1.1 Simple Definition of Data De-duplication with example:

De-duplication is a method of eliminating duplicate data in backup storage systems. De-duplication stores a single unique copy and eliminates the redundant data and creates a reference link in the place of duplicate data. De-duplication can be explained with single example, consider the normal situation of a email server where the similar copy which is of size 10 MB is present in 35 peoples Inbox. Instead of storing 35 copies of each 10 MB file, after applying de-duplication to the file the disk only store a single copy of the 10 MB file attachment, and have references to the saved files for any successive backup of the same presentation.

2.1.2 Data De-duplication Methods

There are three methods where De-duplication can be applied whole file level De-duplication and sub-file level De-duplication or block level and byte-level de-duplication. A hashing algorithm generates a unique hash number called hash identifier for each chunk of data analyzed. It is then stored in an index and used for figuring out duplicate hashes. The duplicate portions have the same hash values

- **File level De-duplication:**

File level de-duplication is the simplest of all the three methods of de-duplication and can be implemented easily and also it does not take much processing power because generating hash values is much easier than other methods. The drawback of file level de-duplication is that if there is single byte difference in the file while comparing the hash value also will differ. As the hash values are not same two different copies will be saved on the disk. So the duplicate data can be removed to some extent by using file level de-

duplication. In order to have better disk and cost savings compared to file level de-duplication, block level de-duplication can be used. The following picture shows file level de-duplication before and after de-duplication which duplicate files are eliminated.

Figure-1: File Level Data De-duplication before and after De-duplication.

- **Block level De-duplication**:

Block level De-duplication overcomes the drawback of file level duplication. When block level De-duplication is method applied to the data, the files inside the data will be divided into fixed length blocks. If we make small changes to large file, the backup system only stores the changed bits. On an average if file level De-duplication saves at $1/5^{th}$ rate of the total disk space then block level De-duplication saves disk space rate of about $1/20^{th.}$

But when compared to file level De-duplication block level De-duplication requires more processing power because of increase in the unique hash values. In tern the index values also get increases to seek the individual iteration. The picture below shows the De-duplication method applied at block level. In the Figure below the boxes with same color

is the duplicate data, which is DE, duplicated and redundant blocks are removed and reference link is created to the duplicate data.

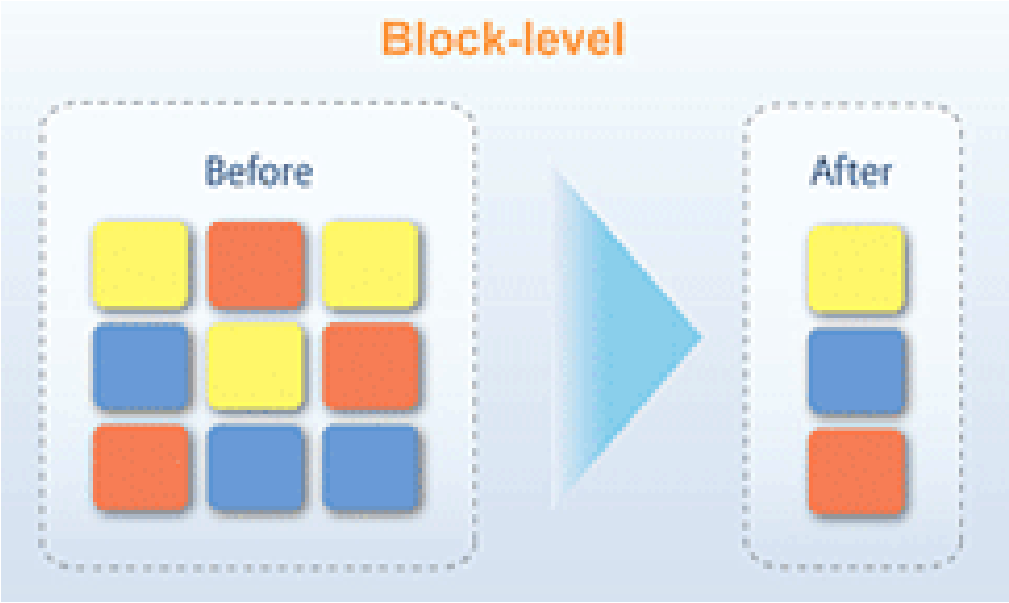


Figure-2: Block Level De-duplication

There are two different types of block level De-duplication. These blocks are also known as chunks. They are fixed level chunking and variable length chunking.

Fixed Block level chunking:

In fixed level chunking method as I said before all the files are divided into blocks with same size of each block for example say 10Kbytes. After dividing calculate the hash value for all the blocks. If the hash value of the data is same then the redundant block is identified and the chunk value is copied on the disk.

This method has three important steps as follows

1. Dividing the file into fixed blocks.

2. Generate the hash values for each block

3. Identify the redundant data from the hash values

Based on the facts and research from different sources, fixed-sized chunking can effectively reduce 1/20th of the disk space. Since chunking boundaries are decided by the offset, this method is very sensitive to the insertion and deletion operation

The figure-3 shows the fixed size blocks from different data files. The top four blocks are the blocks that are unique blocks already on the disk. The figure below shows the blocks when single change to block-A is made, which is called an insertion. Despite the fact that the same color sequence of information is identical in the upper and lower blocks in the figure-4, each block have changed content and no duplication is detected. If we stored both sequence blocks, we would have 8 unique blocks.

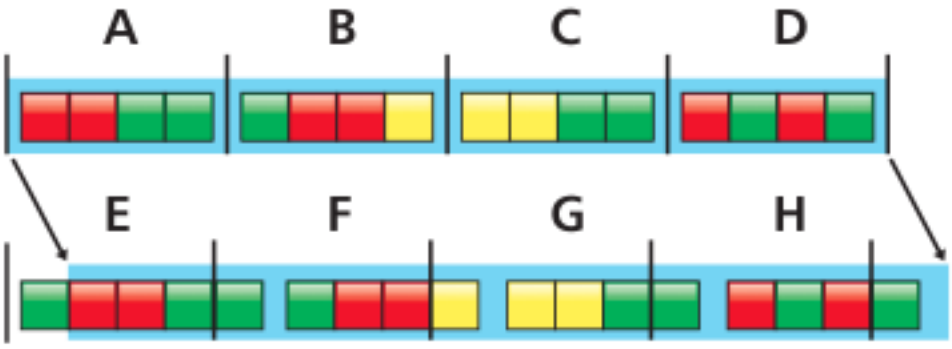


Figure-3: Fixed level Data De-duplication blocks

Variable size chunking:

Variable block chunking is different from fixed block chunking. Chunking boundaries are Determined based on contents of the file, so it is more resistant to the insertion and deletion. Now it is believed as the best algorithm for backup system. In the figure-4 below A to D represents different blocks. Whenever new data is added Block A will change. When the new data is added then it is E, but none of the other blocks are changed. Blocks B, C, and D are all known as duplicate blocks in the first line. If we stored both sequences, we would have only 5 unique blocks.

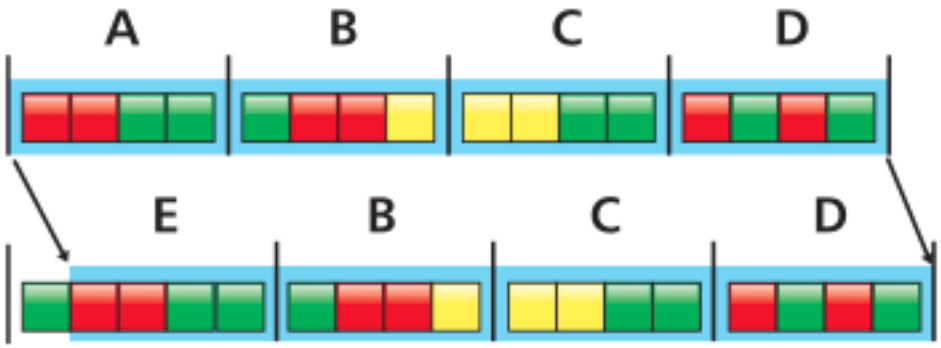


Figure-4: Variable length data De-duplication

Similar to fixed size chunking method variable size chunking has three important steps as follows

1. Dividing the file into variable blocks based on the chunk boundaries.

2. Generate the hash values for each block

3. Identify the redundant data from the hash values

The algorithm we use for finding chunk boundaries is Rabin finger printing algorithm, and each chunk is converted into hash values using common hashing technique MD5 or SHA1. We use SHA1 for generating hash values and identify the redundant data.

Below is the Flow chart that explains the working of Data De-duplication.

2.2 Rabin finger printing algorithm:

The efficiency of data DE duplication occurs because of our algorithm used to eliminate de duplication. The algorithm is Rabin finger printing algorithm. This Rabin fingerprinting algorithm is as follows and also is an effective way of computing fingerprints for every string of length n.

Every sequence of n characters in the string is considered as a number in some base ‘b’. Typically base ‘b’ equals to 256 for 8-bit characters but other values are also possible.

The number encodes as an integer value that can be computed by as a product of integers stored in individual bytes by appropriate powers of base ‘b’ and by adding them up.

The fingerprint is the remainder of this integer when divided by a constant number K. The value of the fingerprint will be an integer in the range 0 to K -1 and can be encoded in $\log_2 K$ bits.

The values constant ‘K’ and base ‘b’ can be selected to decrease the probability of fingerprints that are agreed when the substrings does not. The designers of the algorithm choose base ‘b’ to be 256 and change constant ‘K’ when a wrong match occurs, as the text is being starts scanning. This is achievable because of scanning a big text file for the existence of a single character string. We do not have the option of changing constant ‘K’, so the special care should be taken while choosing constant ‘K’ and base ‘b’.

When calculating fingerprints for all the substrings of certain length ‘l’ in a text, an algorithm is much more effective than the other distinct evaluation method. Based on the

observation, the integer computed from the substring at some offset ‘o’ is easily computed from the integer computed at offset ‘o -1’. Because of the arithmetic properties of remainder, this computation can be done on remainders to get new fingerprint.

Exactly by considering the sequence of characters starting at position ‘o’ as follows

$$C_o\, C_{0+}1\text{…….}C_p\text{……..}C_{o+r-1}$$

Based on the above character sequence integer value is generated as below

$$C_o\ K^{r-1} + C_{0+1}\ K^{r-2}\text{+…….+}C_p\ K^{(0+r-1)-j}\text{+…….+}C_{o+r-1}$$

When this is reduced mod K we get the fingerprint ‘o’.

$F_o = (C_o\ K^{r-1} + C_{0+1}\ K^{r-2}\text{+…….+}C_p\ K^{(0+r-1)-j}\text{+…….+}C_{o+r-1})$ mod K

Then F_o can be computed incrementally based on F_{o-1} using the following equation

$F_o = ((F_{o-1}\ b) + C_{o+r-1}\ \ b^r \text{ mod } K) \text{ mod } K.$

Here (b^r mod K) is a constant that can be precompiled

2.3 SHA1 hashing technique:

After getting the data boundaries from the Rabin finger printing algorithm then hash value is generated to find the hash value. SHA-1 is a cryptographic hash function used mostly and also used in several applications we are using the same technique to find the hash value to identify the duplicate. When the hashing method processes data, a hash is created that represents the data. A hash is a bit string, which is either 128 bit for MD5 or 160 bit for SHA-1, which represents the data processed. When you try to find the hash value of the same data it will generate same hash value.

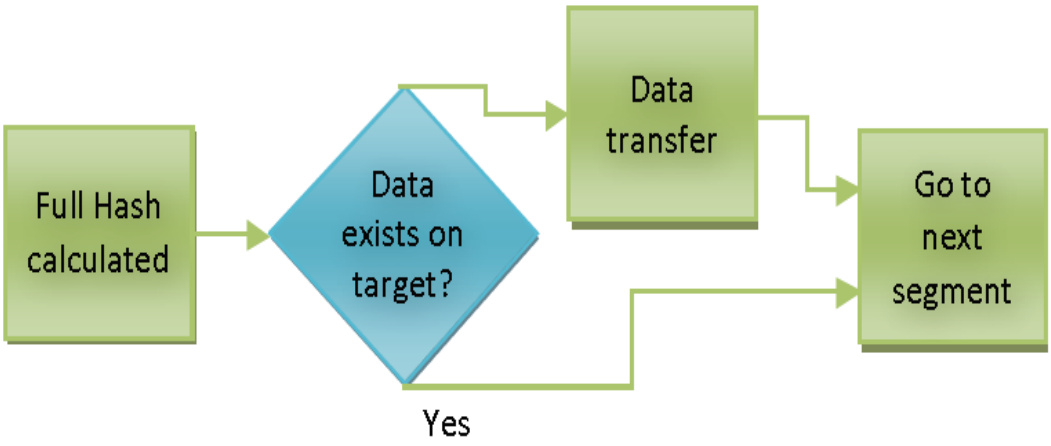


Figure-6: E-mail data De-duplication working process

Ultimate Solutions uses variable block size chunking that can be specified based on the environment requirements 64 or 512 KB or any other. A larger block size requires fewer resources, but in some cases provides very small compression. A smaller block provides better compression, but requires more resources.

- **Byte-level De-duplication:**

Byte-level De-duplication is a method in which the De-duplication is done at very granular level in this case data blocks are compared byte by byte. It checks for redundant bytes, which is more precise. Byte-level De-duplication takes much more time. It is very complex to implement. The picture below shows the byte level De-duplication. The picture has all binary data and De-duplication is done at bit level.

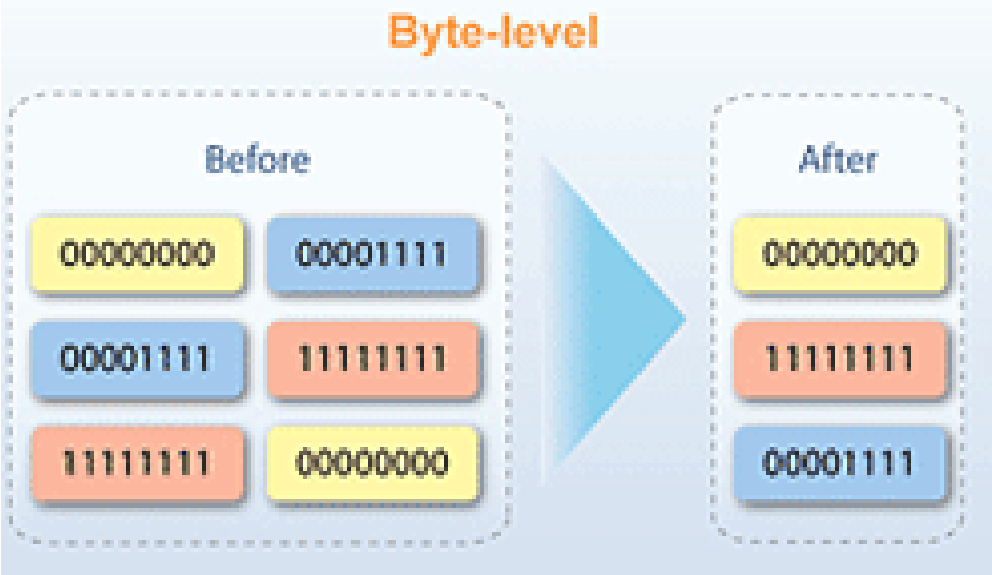


Figure-5: Byte level De-duplication.

2.4 Data De-duplication Approaches:

There are mainly two different types of approaches where Data De-duplication can be done. The data can be DE duplicated before the data is placed in the network at the client side and de-duplication can also be done after the data is stored at the server side after the data has been sent to the server.

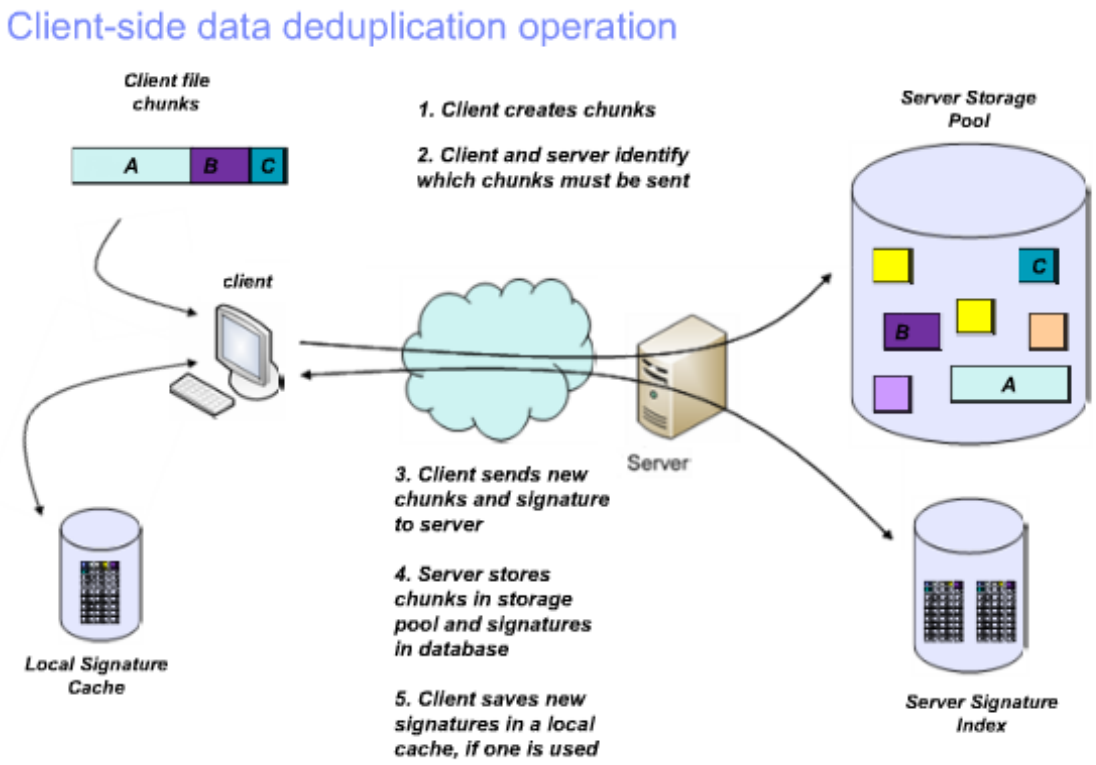
- **De-duplication at the client side**:

When De-duplication is done at the source or the client side the network bandwidth is optimized. Formatting and usage of the data can provide more efficient data reduction and De-duplication at source can facilitate scale out.

The main disadvantage is that De-duplication CPU cycles are done at client side and also requires special software's.

The picture below shows the operation of client side data De-duplication. As the picture depicts the client software breaks up the file into small chunks and create a unique

signature or value for each chunk. Before storing the chunk value is sent to the server and checks whether the chunk is unique or is a duplicate value, which is already in the storage system. The local signature cache is used in the client side to reduce the frequency of server queries. If the chuck value does not exist the client sends both the chunk and the signature that is chunk value to the server. The new chunks are stored in the chunk pool and new signatures are stored in the index databases of that server. The signatures are also stored in the local signature cache memory for client for use in the later backup.



Fgure-6: Performing Client side data De-duplication

- **De-duplication at server side:**

Data de-duplication done after the data is placed at the server side, all the data has been sent to the storage location and then de-duplication will take place after sent the data to our servers.. The biggest advantage of this kind of approach is that whenever de-duplication occurs the system will have a fixed view of the complete file system and knows everything about the data it has access to and can increase de-duplication. De-duplication performance can be slow because it may compare file data to all data stored on disk. Also, data written to the storage system must be batched until the next scheduled de-duplication time. This creates a delay between when the data is written and when eliminating duplicate data reclaims space. It does not require additional client software's. CPU cycles take place at the server side.

The picture below shows the De-duplication process at the server side. In server side De-duplication all of the process done by the server. When the server identifies the chunk after the client processing has copied the data to a disk storage pool that has been setup for De-duplication. The server creates an index value in the server database and places a pointer to chunks in the storage pool that the pointers refer to the actual data file. Before data De-duplication the data server should be backed up. The duplicate chunks are removed from the primary storage pool.

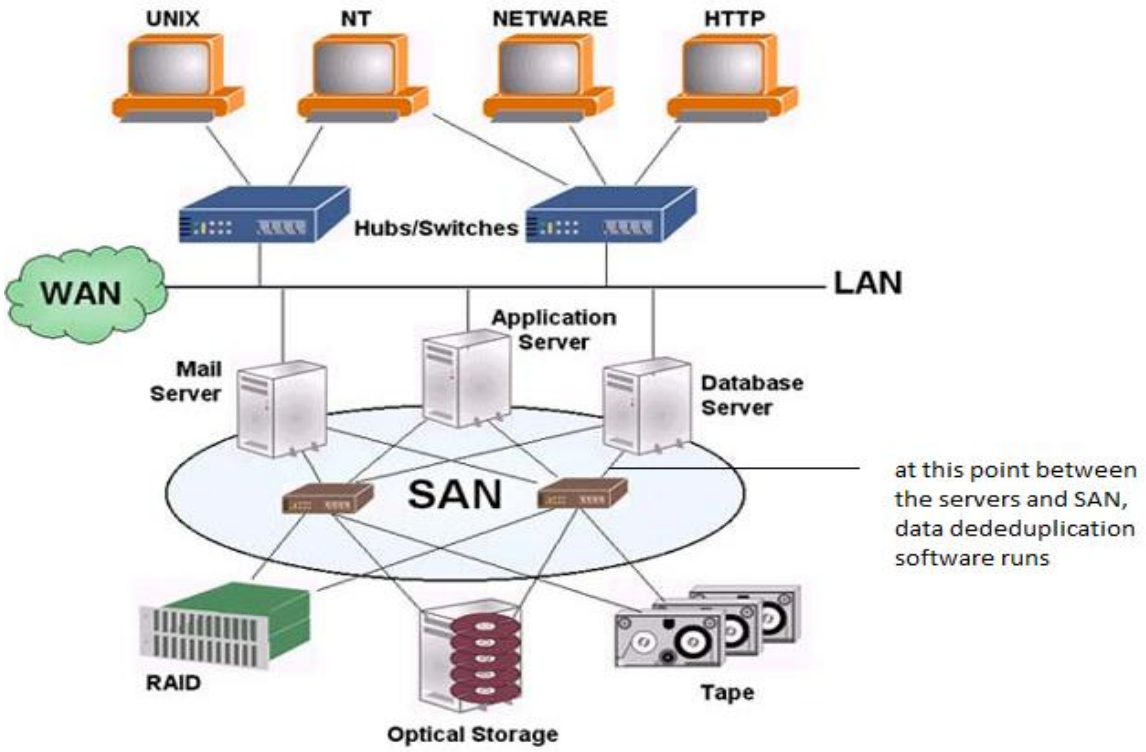


Figure-7: De-duplication process at the server side

At Ultimate solution we prove De-duplication services at both client side and server side in parallel. Whenever the data fails to De-duplicate at client side de-duplication is performed at server side.

3. High level Architectural overview of E-mail Data De-duplication:

Data De-duplication can be applied to different type of data servers like cloud data centers, web data centers and mail data centers. As a start-up company we are going to provide services and software for e-mail servers which uses fixed length data De-duplication also we improve our services further in the future to all kind of data servers and provide complete variable length chunking method to provide much effective data De-duplication.

In this section we describe the architecture of our email systems.

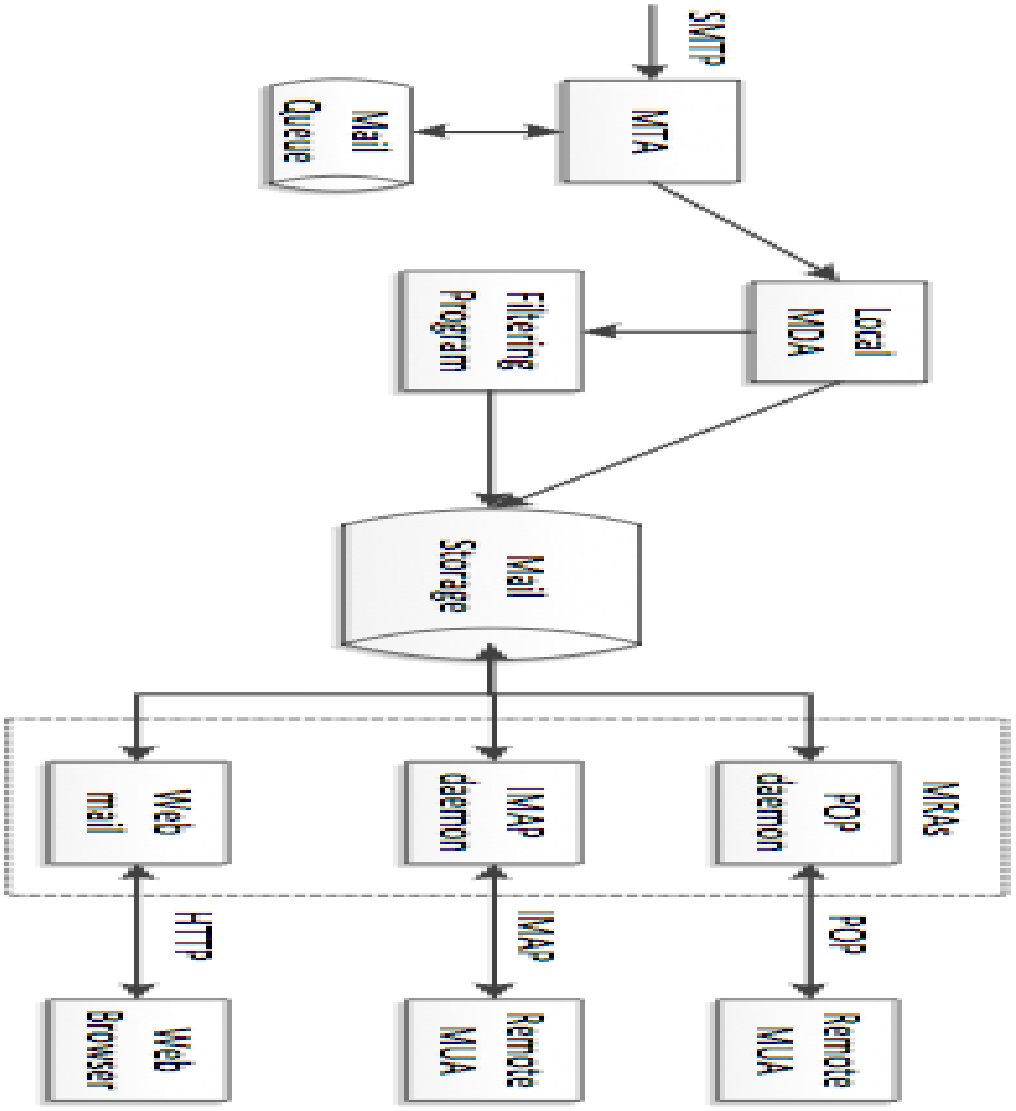


Figure-8: Architecture of Mail Server without De-duplication

3.1 Introduction

Email servers traditionally are used to provide simple text services from one computer to another for small range of computers. As Internet is developing enormously many popular mail servers came into existence such as Excite, Lycos, and Hotmail etc., that brought web technology through internet. After that email became most popular and important service. Email service is a special kind of service among different kinds of Internet services like blogging, social networking, micro blogging etc. In the beginning of email services there are different email application available for the users like Sendmail, Qmail that are called email clients. These clients work with email servers to view mails for the users. Slowly these email clients became very popular. After the development of Webmail facilitated email service for public. Popular email service providers support thousands of clients with high availability, reliability and good usability. After that several company specific email servers have been provided for small and midsize companies for cheap prices. Based on the reports from various sources in the year 2011 there are 275 yahoo mail users and 360 Windows Live Hotmail users. Gmail who started their email service in 2004 has almost 200 million users in a period of month. For group chat or team discussions normally every company uses groupware that has a built in email server. For example if there is a company team working for a project. One of the team members after working with some part sends his part of work to all the other members of the team as an attachment. As the entire group has the same attached document multiple copies of the same content will be stored in the email servers. Before the project finishes lot of copies of files will be created before the project completes.

Internet web email is designed without the uses dedicated workstation. User occupies some amount of space to store their mails in the server. The user can access their mails from any kind of devices such a personal computer or a mobile device such as smart phones or any hand held device. Large companies like Google, Microsoft, yahoo etc. provide email service with certain amount of storage space that depends on the service provider. In order to utilize more space the user has to spend some extra bucks. The mail can have different file types such as documents, images, audio, video files etc. and these types must support by the service provider. There will be more volumes of duplicate data. Suppose in the case of Gmail if the user needs 1GB of space on an average it takes almost 200petabytes. As the time passes the number users will be more and the disk space needed by the users also increases. If the duplicate data is removed the disk space savings will be more. A single mail in the user mailbox can be a file or part of a file.

Applying De-duplication on email messages is used to eliminate duplicate data parts of large collection of files. As there are many methods to detect duplicate data that is explained in the previous sections that files are divided into small blocks and apply the hashing method to find the redundant data and eliminate this redundancy. In our project we have proposed a method that provides De-duplication to the mail servers.

3.2 Email server architecture without De-duplication:

Figure-8 depicts the architecture of the email system how mails are stored in the server. There are different components in the mail server such as

- Mail Transfer Agent

Mail Transfer agent (MTA) acts as both client and server that send messages through the Internet using a messaging protocol known as SMTP (Simple mail transfer protocol). The MTA collects the messages and store them in mail queues and does some further processes and MTA pass them to the Mail delivery agent.

- Mail Delivery Agent

Mail Delivery agent (MDA) collects the mails from the mail transfer agent and takes the responsibility to store the mail in the users inbox. The users can then login from the browsers to access their inbox. Both MTA and MDA is a structure that lets email servers to run in connected applications.

- Message Retrieval Agent

Message Retrieval Agent is a component of mail server designed for MUA (Message User Agent) to access a email remotely. POP is a protocol used to fetch the mails from the server and in the same way IMAP is a protocol that makes access possible when all mails are stored at the server side. Webmail is also similar to IMAP but it uses HTTP protocol.

As there are lot of duplicate data and also there are lot of emails that has same email but changes it with small changes like changing the name of the file or changing the subject of the mail etc. To find duplicate data more accurately we use variable length chunking taking some considerations like the start and end of the subject as chunk boundaries and similarly the message body from start to end as another chunk boundary. In our project we apply De-duplication on user emails which we follow unique method. So to implement the architecture will be changed slightly. The architecture of our email De-duplication, which we implement, is specified in next section. In our email De-duplication system, all emails that belong to the users are organized in different format to include the requirements to DE duplicate redundancy in the mails and mail attachments. To interact with Message User Agents and Message Transfer Agents of other systems, Message Transfer Agent and Message Retrieval Agent are also modified in the same way.

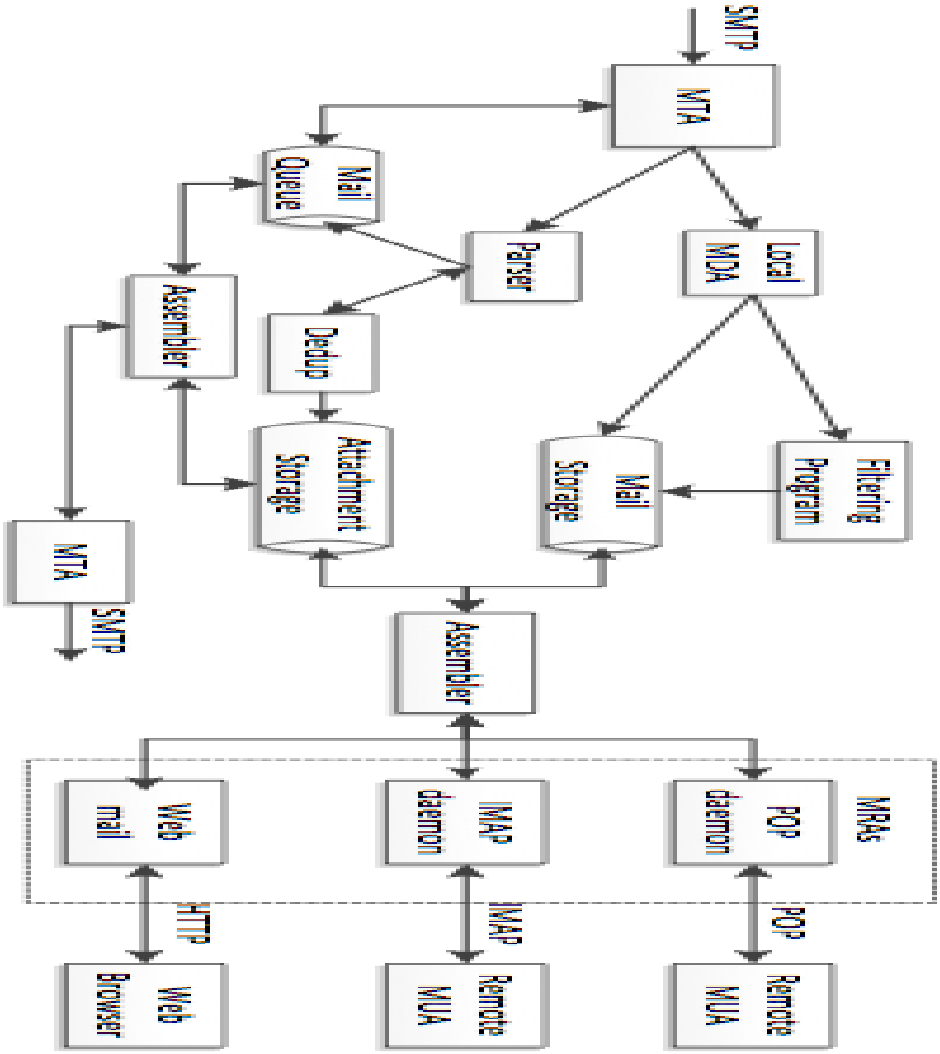


Figure: 9: E-mail Data De-duplication implementation at server.

3.3 Email server architecture with De-duplication:

As shown in the above figure-9 there are different components added in the new architecture. The parser component takes the attachments from the email and changes the message in another format. De-dup component first check the chunk boundaries such as subject of the message, body of the message and attachment of the message and finds the fingerprint of each and every chunk and stores the copy in the server index database. All the attachments are stored in the different storage pool. Assembler is the other new component that combines the whole message again when the user likes to view in his browser and finally SHA1 is used to detect duplication of mails. Digest is the unique identifier of a file and is used by parser to represent the attachment of a message, body of the message and subject of the message. Data structure is used to manage all the identifiers created by the SHA1 algorithm. It also supports insertion, deletion and lookup. Lookup is nothing but a method that makes sure whether an attachment is already there in the server or not. Parser component examines messages that come to the server. Parser first tries to match a specific field named Content message and body and then an attachment is taken and given to De-dup component to calculate hash value. The attachment is replaced by its hash value with an extra field as an anchor. Now the email will be as different parts that is modified and other parts such as attachment, body and subject.

3.3.1 Email storage Management:

As the storage method of emails in the server will be modified when De-duplication is applied. We manage the storage and index's using data structures as follows. Emails are stored in the disk in two different formats that is in a separate directory, which is called mail directory, or it can also be as a part of other file such as mailbox. In these two types of storage emails of different users can be stored on different hosts.

For mails that are already dispensed to local receivers, they are stored in some dedicated storage space assigned to the corresponding users, which makes implementation simple.

A directory is a regular file that has all immediate sub files in the form of tree structure. It takes much time to access a directory with lot of sub files in it. All the attachments are placed in the directory based on the identifier.

The identifier is divided into numerous portions with equivalent lengths. The name for the directory will be first character. The depth of the tree will be equal to length of the identifier. The files in these directory structures are stored at leaf level.

With in very less amount of time the mail server has thousands of mails created and trashed so in our project we maintain dynamic data structure to all the index attachments. There are different types of data structures to answer a request to make sure the index attachment already exist.

The SHA1 hash value attachment is taken as identifier, which requires about 32 bytes. To save the space we used binary version that occupies 16bytes.

If there is a small startup company that maintains mail system and for instance there are 100 million attached files stored at max and if the average size of the attachment is 100KB. It needs about 10 disks of each disk 1TB. The space needed to store the indexes is 1.6GB. When considering modern systems the complete index can be copied into the main memory.

For big email service providers such as Google, Hotmail, and Yahoo etc., there will be billions of attached files, which need 16GB of space for index, which does not fit in the main memory.

Different data structures can be adopted according to the number of attachments hosted by email server. We here only implement internal memory data structure for small-scale system at present.

The data structure for storing and managing the index's we used is red black trees. All operations such as insert, delete, and lookup time complexity is about O (log n) time comparisons Moreover, we place all the data structure in the external drives which helps when the system shutdowns and loads into the main memory when-ever the system is initializes back. Data structure should be copied when it is updated.

For big database systems, it is not possible to place the index values in the main memory. External memory is required for the data structure. We used B-Tree to store indexes for large systems. The index is nothing but hash value. We use External hashing for large systems, as it is a good choice.

4. Hardware and software components involved in the E-mail De-duplication

The requirements for our project include both software as well as the hardware requirements. In the software requirements we will specify the software modules and hardware modules.

4.1 Software Requirements:

a) Software modules we will use are Advance compression module where actual DE duplication occurs.

b) Name server, which resolves the names.

c) Operating system we support is all windows and Unix variants

4.2 Hardware Requirements:

a) A chip for DE duplication.

b) Another chip for the flow of the data and control central CPU.

c) Data processing units required based on design we used two for De duplication.

d) Disk subsystem, which is used for storing the data.

e) Data de-duplication powered by Hyper Factor technology

f) Powerful multi core virtualization and de-duplication engine

g) Fiber Channel Ports for host and server connectivity

h) Flexible storage choices and options

4.3 Our De duplication software components:

- Ultimate Solution backup Manager V1.0
- Server Configuration:

- Storage pool

Use Device class of FILE

Specify De-duplication

Windows 2008 server

- Options for the server

Client De-duplication: The user should specify the max size of disk to be DE

Duplicated at when performing client side De-duplication the default is 40GB

Server De-duplication: The size for the disk to be DE duplicated on the Server side should be specified and the default is 250GB.

- Command set: De-duplication Verification level

Percentage of De-duplication done will be displayed. The default will be zero.

Setup:

E-mail Data DE duplication is done on the top of physical hardware is placed.

To confirm the documented features of our DE duplication environment, we setup our own datacenter environment from third party. Following are the steps we used for our prototype our DE duplication setup:

1) We used available 2 Dell R900 servers with 4 Xeon 7350 processors and 64 GB of Random Access Memory. After connecting these servers with standard layer 3 switches we installed both servers with ESX 4.0. ESX 4.0 installation was straight forward as we used all the default options for configuration. We assigned both servers with IP addresses from same VLAN.
2) Next step is to prepare hosts email DE duplication setup needs 3 stand alone systems. Active Directory and DNS server, Microsoft SQL Server, and host for installation of our algorithms and email agent components. All this hosts are physical hosts. We decided to use physical hosts for our setup.
3) We used available copy of windows 2008 enterprise server as a base operating system for installation of all the email DE duplication architecture.
4) Purpose of active directory server is to provide single secure sign on to all the resources and domain. vCenter can validate each user against Active Directory using LDAP (Light Weight Directory Access Protocol). To promote windows server to domain controller we run “dcpromo” command on the command prompt. We also added a DNS (Domain Naming Service) role on the active

directory server. Following “Add Role” wizard from server management in windows 2008 server can do that.

5) Next we installed Microsoft SQL server 2008 on one of the 2008 infrastructure servers. The SQL server will be responsible for storing the emails of the users
6) Next we installed Microsoft SQL native client on the third infrastructure
7) A parser software component takes the attachments from the email and changes the message in another format and DE duplication algorithm that is Rabin fingerprint. Chunk a boundary that are subject of the message, body of the message and attachment of the message and find the fingerprint of each and every chunk and stores the copy in the server index database. All the attachments are stored in the Dell R900 servers.
8) Parser component examines messages that come to the server. Parser first tries to match a specific field named Content message and body and then an attachment is taken and given to De-dup component to calculate hash value. The attachment is replaced by its hash value with an extra field as an anchor. Now the email will be as different parts that is modified and other parts such as attachment, body and subject.
9) Finally SHA1 is used to detect duplication of mails. Digest is the unique identifier of a file and is used by parser to represent the attachment of a message, body of the message and subject of the message. Data structure is used to manage all the identifiers created by the SHA1 algorithm. It also supports insertion, deletion and lookup. Lookup is nothing but a method that makes sure whether an attachment is already there in the server or not.

10) Then we installed 2 Ubuntu Linux in this resource pool, which we later used for security validation.

4.4 Testing our project with a stand-alone computer:

The test has been done with the following requirements

- Personal stand-alone computer is used as a test system.
- Intel Quad core Processer with 2.33GHz
- Main memory with 8 GB
- Hard disk of size 200GB

We use Perl as the programming language for developing our DE duplication method.

4.4.1 Testing DE duplication Performance

As shown in the figure 9 above the parser component parses all the emails that come to the users mailbox. As predicted, the space for user mailboxes falls to a percentage that is nearer to duplication percentage of attachments. Because of security problems it is not possible for us to collect practical email messages. At present, we generate several data sets with 30,000 emails and 100GB of disk space in each dataset. The duplication percentages savings for the attachments are chosen as 1%, 5%, 10%, 20%, 30%, and 50%.

5. Company Summary

5.1 Introduction

We are planning to start an IT company of our own rather than selling the software product to another firm. So, this section is going to deal with all the details about the company that we are planning to start. This section contains the details about the ownership of the company, company location, the start-up plans and all the requirements for the company to function efficiently and make profits.

5.2 Company Ownership

Ultimate Solutions Inc. is the Invention of Sharanjeet Hundal, Tanveer Singh and Basavasai Konuru. We have approached Venture Capitalists for the initial funding to get us on the track to start the company. The funding Venture Capitalists will also be a part of the Board members along with the three mentioned above. All team members will be employed at Ultimate Solutions Inc.

5.3 Company Locations and Facilities

Our Company is registered by the name "Ultimate Solutions, Inc.". It was established on September 1st 2011 and is located in Palo Alto.

Our company believes in sincere work and building best customer relations. We specialize in providing complete De-duplication solutions at the customer satisfaction. Ultimate Solutions is mainly focusing on the development of the de-dupe software which is more reliable than the current available software's. For the mid-sized market, we have

developed a cheaper substitute by providing the data de-duplication service. Our customers will include global money banks and leading financial services firms, healthcare and life sciences organizations, manufacturers, airlines and transportation companies, Internet service and telecommunications providers, public-sector agencies and educational institutions.

The present market trying to reduce their storage disk size is adapting fixed size chunking methodology in which each file is broken into fixed size chunks and if these chunk differ by only a byte, the data de-duplication doesn't takes place for that files. So to overcome this drawback of the current technology, we are mainly focusing on the use of variable sized chunking technique along with Rabin Fingerprinting Algorithm. In this technique we are dividing the files into variable size chunks through which the chunks that are known to be stored on the disk are de-duplicated even when there is a small byte difference between the chunks.

5.4 Organization Hierarchy

The organization will have the following employees:

The Chief Technology Officer (CTO), Data De-Duplication Consultants and the Chief Marketing Officer (CMO) would report directly to the Chief Executive Officer. The Marketing and Sales professionals along with the Finance Manager would in turn report directly to the Chief Marketing Officer. The Customer Support Professionals, Developers and the IT Support Engineers would report directly to the Chief Technology

Officer. The organization lawyer and the Junior Data De-Duplication Consultant would report to the Senior Data De-Duplication Consultant.

The below diagram depicts the hierarchy of the Ultimate Solutions:

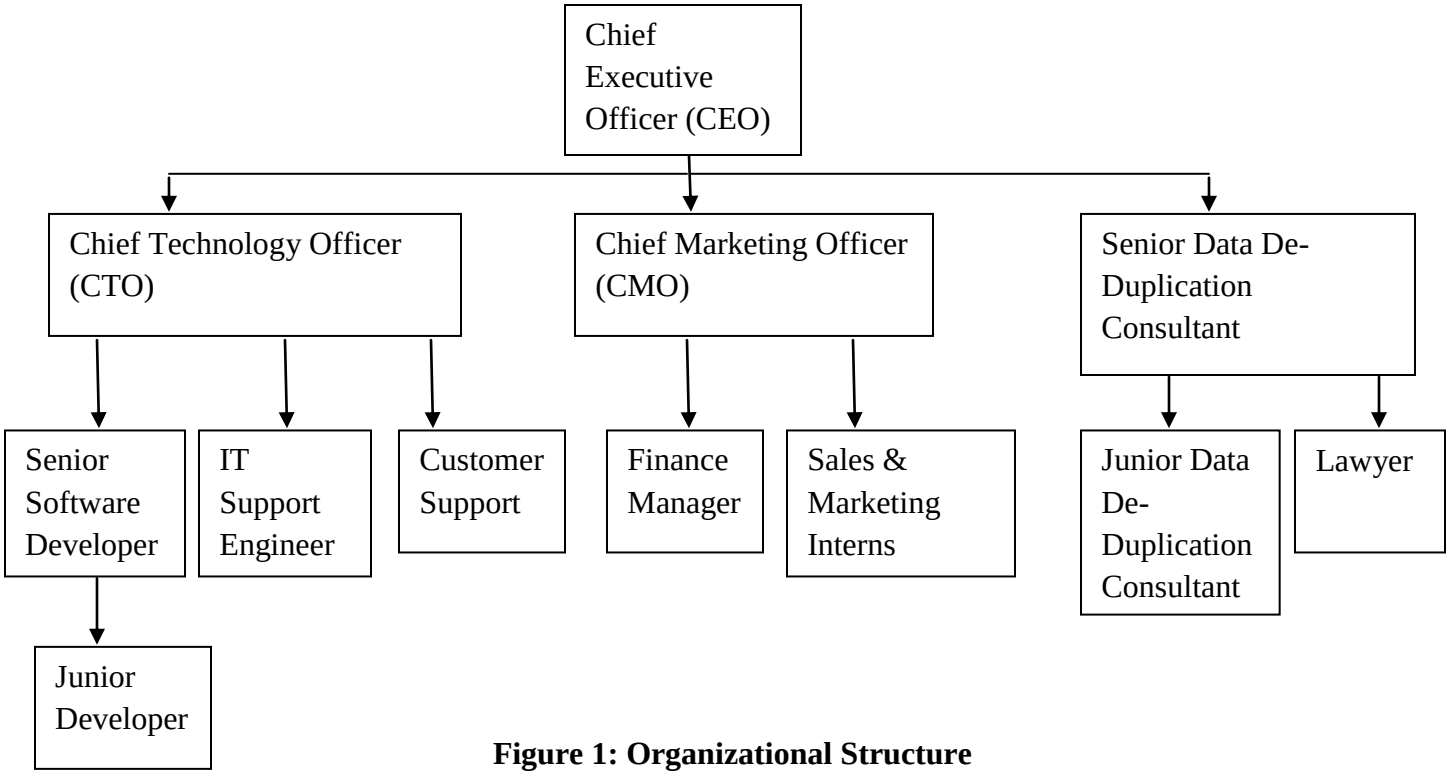


Figure 1: Organizational Structure

5.5 Business Model

A lot of IT companies today are seriously looking at the "Data De-Duplication" options to check if it could provide a solution to their challenges. There are a lot of data de-duplication benefits that can be availed by running the Data De-Duplication software between the servers and the data centers like magnetic tapes, disks etc . A lot of companies are not very clear on the impact of their data to be de-duplicated when the data is entering the storage servers. Also, most of the companies have serious concerns about

the security and safety of their precious data and other issues which delay or deny their moving to the use of data de-duplication software.

In addition to all of these challenges discussed above, there are no customized packages for the small- scale businesses. This sector has been largely remained unused by the major IT companies. The companies like Data Domain, Net App etc. are busy going after the bigger firms, which can reap them, more money. However, the forecast shows that these big and small companies are very likely to move into the use of data de-duplication because of their problems in managing their IT data as the data centers are growing rapidly and the data de-duplication is needed to remove the redundant data.

Ultimate Solutions Inc would like to address some of these challenges and thus make best data de-duplication software to meet the expectations of the IT companies. Our target customers are the small and big sized companies as we are providing data de-duplication services to the small-scale companies and selling the software to the big companies like Google, Microsoft etc. We will be coming up with the special packages and business plans around them

Ultimate Solutions Inc. would provide three kinds of services. They are:

- Data De-duplication Hosting
- Data De-duplication free trial software
- Data De-duplication Consulting

For the "Data De-duplication ignorant", i.e. those who are not very clear as to whether the data de-duplication might be a good solution to their IT needs, we have the special consultancy service. Our Expert Data De-duplication Consultants will be able to guide

them on whether the Data De-duplication is a suitable choice for their nature of IT tasks. For the "Data De-duplication Skeptic", i.e. those who are not very sure about the performance of the Data De-duplication software in spite of knowing that it can act as a very good solution for saving their data centers memory by removing the redundant data, we have the free trial software for the 30 days. Finally, for those who have become "Data De-duplication Confident", we have the installation of the software solution. Our installation team will install the software between the servers and the data centers of our clients and we will make sure that it will be in a secure and easy to use environment.

This business model can manage over a long time as we will have recurring revenue from those who have opted to use our Data De-duplication product. The costs we are charging will be based on month basis depending on the amount of data to be de-duplicated. In addition, we also have revenue from the data de-duplication consultancy services.

5.6 Service Consultancy

As explained previously, we expect to have a part of our revenue from the consultancy service provided to the small businesses, which who wants to use our services to remove the redundant data from their data centers. We will advertize our consultancy services and promote other services by the following way:

1) **Providing secure Data De-duplication:** The IT customers can temporarily host our data de-duplication software by using the free trial version for a month. This period will help them to analyze the benefits the business benefits of the data de-duplication. The main reason for the small companies to use our product is due to

the use of the latest variable sized chunking technology behind our de-duplication product.

2) **End to End deployment:** If the De-duplication Service Seeker wants to buy our software, we will provide an End to End deployment of the software including all the services ranging from making sure that all the IT requirements of the small and large business are satisfied and the data is de-duplicated in a well manner.

In a nutshell, we are planning to prepare organizations to become ready for the data de-duplication by educating them on various changes on their data centers and also help those who are ready, to transition to the data de-duplication services. Our highest priority is the customer satisfaction and data de-duplication accuracy due to the nature of our business.

5.7 Website as a Sales tool

We are developing a website that would act as a major tool for sales. The website would be like any other company's website that provides the information regarding the various services and us we offer. It would provide all the details about our company's product and the services to our potential customers so that they can make a wise decision. In addition to that, there would also be a cost table for the data to be de-duplicated per terabyte for the customers. Based on the details of the customer requirements, we would be able to provide an estimate service cost for the customer.

5.8 Company Objectives

- Help small and large businesses prepare for the data de-duplication

- Provide free trial version of data de-duplication to verify performance.
- Undertake a company's IT maintenance and support.
- Help companies to easily transition to the Data De-duplication services.
- Organize a culture of regular improvement in customer satisfaction.

6. Economic Justification

6.1 Executive Summary

Ultimate Solutions Inc. (USI) is a startup founded by Tanveer Singh, Sharanjeet Hundal, Basavasai Konuru in September 2011 in Palo Alto, CA. We help business eliminate duplicate data, thus resulting in savings due to reduced storage and data integrity and maintenance costs. We have a proprietary algorithm and procedure for detecting and preventing the storage of redundant data that is 70% more effective than what competitors can achieve.

The total of $7 million funding is required by USI that will take the company to break even in 2 years. Funds will be used for salary, equipment, servers, magnetic tape, lease and other expenditures. To reach the breakeven point our company will need to sell 20 software units. We anticipate securing $4,000,000 in the first year by selling around 11 of our de-duplication software units.

Revenue will be generated on a license fee per Terabyte/year. The anticipated ROI is expected to be 10-fold by the fifth year of operation. During the first quarter of funding, the CEO and key VP positions will be filled with qualified and experienced personnel. The founders will transition to technical leadership positions.

6.1.1 Solution and value proposition

Many companies like Data Domain and Sepatan offer a fixed size chunking (FSC) technique as a solution to eliminate data duplication. This approach is not as efficient and flexible as what can be provided by USI. FSC partitions files into fixed sizes. If the data within two partitions are identical, one of the partitions is not saved. However, if two partitions differ by just one "bit," then both partitions are saved, and a significant amount of data is duplicated depending on the size of the fixed partitions. USI offers a proprietary and more efficient variable block-chunking (VBC) scheme based on the Rabin Fingerprinting algorithm. An analysis of the file's content is made and is partitioned into "chunks" of different sizes to better identify areas of difference and similarity between files. VBC facilitates reducing the amount of data that is duplicated on storage up to 70% more than can be achieved by the FSC scheme.

We will provide the companies with the best Software solution, targeting the big size segment market and for the small business; USI will be focusing on the data de-duplication service which will be done using our servers and setup. The complete solution will be offered at quite a lower price than what other companies are offering right now.

6.1.2 Market Analysis

The need of Data De-duplication is tempting and totally driven by the customer needs

and the industry trends. The key factors that are driving our company to attain a successful position in the market are:

- We believe in establishing the best customer relations and satisfying their demands within the promised time.
- Our company is doing something, which is not present in the Data De-dupe market right now and we hope that once this development phase is over, we can definitely attract all the big and small fishes to our Ultimate Solutions.
- We are going to provide the ARCHIVING service, which is to gain a cost-effective, online archive for our client's organization's non-changing data assets so that we can minimize the risk, maintenance and control costs and eventually increase the content reuse.

The current size of our company is very small which consists of only 23 employees including the owners of our company, which is just a team of three people. With the growing demand of our software and the service, we hope to achieve our targets on time and then we can definitely increase the size of our company according to the needs. The projected size in the coming years is that the USI will have around 200 employees with the customer count around 2000.

1. **Market Size:** The potential customers who are currently using data de-duplication services are paying in the range of $9000 - $12000 per Terabyte according to the data range. Even though they are using the old technology we are planning to sell our product in the same price range and we are expecting to capture at least 70%

of the market. As we are using better and efficient technique that will find 60% more redundant data than the de-duplication techniques that are currently being used by our competitors like Data Domain, EMC, and Net app. So we are planning to sell our service at a price range of $11000 - $14000 per Terabyte depending on the data that's need to be compressed and duplicated.

In 2009, the data de-duplication market exceeded $1 billion and projected to grow over 10% annually. Data domain, Net App, IBM, EMC, Exagrid, FalconStor Software, ExaGrid Systems, Sepaton, NEC and Quantum all have products competing in some way in this space. Potential customers currently are paying $9K - $12K per Terabyte for data de-duplication solutions. Even though most are using the older FSC technology, we are planning to penetrate the market by 2% within 2 years by offering a more effective solution at a more competitive price.

Over the period of 2011–2015, the Global Data De-duplication market is showing the signs of growing at a compound annual growth rate of 31.71 percent. One of the main factors, which have been contributing to this market growth, is the gradual increase in the implementation of unstructured data applications on the storage platforms. The de-duplication market has been experiencing the increase in the use of cloud based computing systems for compression of the redundant data on the cloud to save the disk space and other security features.

6.1.3 Competitors

Data De-duplication is a big market with a lot of giants, which are already well

established and maintain a good position in the world of data de-duplication. Being a startup with a new advancement in this technology we are already prepared to face these giants by advertising and marketing our software at quite a large scale.

6.1.4 Profiles of Target Markets

Based on the product and services that we are working on right now, we will be targeting all the small, midsized and big companies and industries including the global money banks and leading financial services firms, healthcare and life sciences organizations, manufacturers, airlines and transportation companies.

6.1.5 Marketing Strategies, Sales Plans & Projections

Our software and service will be available online on our website. We will approach the database system administrators and give demo to them to meet their needs and provide them with the complete solution. We will hire the marketing agency to advertise our product and services and reach to our customers with the best possible solution for their organizational needs. Once the product is sold, our employee's will follow up regularly to see if there are any further issues that need to be dealt with.

In order to compete with the existing Data De-dupe software's, we will provide promotion like free service for 6 months and money back guarantee if the client is not satisfied with the software.

Talking about the sales, we are projecting to target at least 70% of the market. USI will effectively access each market segment through distributors, a captive sales force, using

E-commerce and direct mail. Customer's price sensitivity will play a main role in establishing our market. The cost of acquiring and retaining customers depends all on the current software's and their price. We at USI, believe in 100% customer satisfaction and for that we will price our software at quite a lower price than the available products and services.

6.1.6 The Company and Its Services and Strategy

Our Company is registered by the name "Ultimate Solutions, Inc.". It was established on September 1st 2011 and is located in Palo Alto.

Our company believes in sincere work and building best customer relations. We specialize in providing complete De-duplication solutions at the customer satisfaction. Ultimate Solutions is mainly focusing on the development of the de-dupe software which is more reliable than the current available software's. For the mid sized market, we have developed a cheaper substitute by providing the data de-duplication service. Our customers will include global money banks and leading financial services firms, healthcare and life sciences organizations, manufacturers, airlines and transportation companies, Internet service and telecommunications providers, public-sector agencies and educational institutions.

The present market trying to reduce their storage disk size is adapting fixed size chunking methodology in which each file is broken into fixed size chunks and if these chunk differ by only a byte, the data de-duplication doesn't takes place for that files. So to overcome this drawback of the current technology, we are mainly focusing on the use of variable

sized chunking technique along with Rabin Fingerprinting Algorithm. In this technique we are dividing the files into variable size chunks through which the chunks that are known to be stored on the disk are de-duplicated even when there is a small byte difference between the chunks.

6.1.7 Customers

The potential customers for data de-duplication are all those small and large enterprises which have a large database to maintain as well as their database is increasing gradually day by day. With the help of our software, they can easily reduce the space on their disks to maintain this data by reducing it using the best de-dupe solution available in the market. Moreover, the recent global recession has triggered every company to think about their finances and reducing their expenses, specially the IT solutions which includes the deployment of this software and erasing the redundant data and save a lot on their disk space and this in fact will stop them from buying new disks or spending money on servers and tapes etc. In addition to this, most of the enterprises whether it's a mid-sized or large segment, are concentrating on their core activity rather than spending all their time and money in buying and investing in IT hardware and maintaining it because that also takes quite a lot of effort and manpower.

Another important aspect of our software and service is that it comes with a lot of security features, which will help them in not spending on the security.

6.2 SWOT Analysis

SWOT analysis is an important part for any organization that is planning or has already started operations. It will give us an idea about what our strategic edge is, over our competitors and will also help us in planning for remedial measures considering our weaknesses and therefore let us manage our resources economically. Strength of our project is that the technique we are using can detect 70% of the redundant data whereas other data de-duplication techniques in the market can detect only up to 25% of the redundant data. The only weakness associated with our project is that we are just the start-up company. The companies that want to erase the duplicate data and to save the large amount of disk space on their data centers is the big opportunity for our company, whereas new techniques that would be employed by our competitors can be a big threat for our company.

6.3 Goals

- To broaden the software and service line.
- To expand the market penetration through advertisements and opening up more offices around the globe.
- To expand the business to the point where it becomes a dominant player in the Data De-dupe industry.

6.4 Operational Plan

The operational plan considers the many details of converting inputs to outputs that Customers value. Our plan is to focus on the continuous flow operation of the service provided and look after the software development and production. We will be looking

after the Customers response to our product and services and enforce the changes as soon as possible to make it the best in the market. Inventory of the software will be maintained at our site and shipping and handling will be taken care off at a fast pace within the promised time. For our software we are going to maintain the inventory of at least 100 software copies for the first 6 months and then the production will be paced up according to the market response. The day-to-day production of the software will be handled internally and providing the data de-dupe as a service will be outsourced.

6.5 Human resource cost

Along with the infrastructure of our business, human resource is another major and one of important component of our expense. In this following section, we are going to deal with the finances for our human resource segment, which includes all the people working at USI. The following table explains all the labor cost for next five years. Using some really valuable inputs form the professionals in the our enterprise and based on the research that we did of initial salaries, we came out with the decision our startup company USI would be successful with the best quality management and the operational processes. Moreover we made a fair decision of giving equal stakes in our company to all the founding partners of USI. To make our business more successful in the long run, we think that there is a need to spend more amount of money on marketing, high quality service and sales and engineering professionals working in the company.

	Y1		**Y2**		**Y3**		**Y4**		**Y5**	
Employee	**No.**	**Salary**	**No.**	**Total Salary**	**No.**	**Salary**	**No.**	**Salary**	**No.**	**Salary**

(Team)	4	180,000	4	250,000	4	450,000	4	650,000	4	680,000
Database consultant	2	154000	2	165,000	4	393000	6	600950	10	997500
Customer Support	2	15930	2	35230	5	98577.5	7	78740	10	98560
IT Support Engineers	2	49000	3	105000	9	527,000	12	687,000	16	874,000
Software Developer	3	30000	3	32000	7	82600	10	116480	13	151973
Marketing and Sales	8	328000	8	397000	12	823000	15	998750	12	1192500
Total	19	756,930	20	984,230	41	2,374,177	54	3,131,920	65	3,994,533

Table 1 Estimated HR Salary for Five years

As a start up, we will start with a small workforce, which will basically consist of database managers, consultants, customer support professionals including the online and tech support, IT support and software developers and finally the marketing and sales professionals to market and advertise our software. Out of all these professionals, our online technical support will be based in China and their salaries in the above table have been calculated based on the current market rate in China.

Depending on the overall performance and their work, the employees will be receiving a 8-15% hike in their salaries every year. A specific team will be allotted to work on the IT development and operations. In the first couple of years, we are planning to keep our workforce to a small number but gradually we will hire more professionals as we expand our customer base and the operations. We will start seeing a hike in the third year and our customer base will be on a good number list, after we reach the breakeven point for USI.

6.6 Overhead Expenses

As far as the expenses go for the infrastructure of our company, there comes along the overhead expenses, which basically consists of travel expenses, stationary, rent, utilities, hardware, magnetic disks etc. These overhead expenses will be considered starting form the first year and will have a small cut in our budget because some of these things will be only one time investment and we don't need to spend money every month on them.

6.7 Total Cost

The total cost is addition of all the above mentioned costs including the HR, overhead expenses, and licensing fees etc.

Total Cost of Ownership

6.7.1 Start-up Funding

Capital	
Planned Investment	
Owner	$500,000
Investor	$4,000,000
Additional Investment Requirement	$1,000,000
Total Planned Investment	**$6,000,000**
Loss at Start-up (Start-up Expenses)	$500,000
Total Capital	$1,000,000
Total Investment	**$7,000,000**

6.8 ROI

ROI is a measure of the financial return on an investment over a specified period of time, represented as a percentage. After the product development phase, we will be looking at the demand and we will step up and start selling our product and service. According to our analysis and first year's investment of around $7 million, we are supposing that we will be getting at least $2,000,000 in the first year by selling around 30 of our deduplication software's to some of the biggest IT firms. That means that the return on investment (ROI) on our product will be around 30% in the first year and thereafter it will be more and we will reach the break even point after around two and half years since the start of the company.

6.9 Profit and Loss

The following table lists the profit and loss for three years. It also shows us details about the total number of customers we are targeting to reach, the expenditure, revenue and profit/loss value.

6.9.1 Projected Profit and Loss

	Year 1	Year 2	Year 3
Sales	4,000,000	8,500,000	10,000,000
Direct Cost of Sales	5,00,000	$500,000	$500,000
Other Costs of Sales	$100,000	$300,000	$500,000
Total Cost of Sales	$600,000	800,000	1,000,000
Gross Margin	3,400,000	7,700,000	9,000,000
Gross Margin %	566.00%	862.00%	900.00%
Expenses			
Depreciation	$55,000	$41,000	$32,000
Rent	$180,000	$180,000	$180,000
Utilities	$107,000	$1000	$1000
Insurance	$55,000	$55,000	$55,000
Payroll Taxes	$1,125,000	$1,140,000	$1,137,000
Equipment Leasing Cost	$90,000	$90,000	$90,000
Total Operating Expenses	**$1.6 M**	**$1.5M**	**$1.49M**

Budget Table

	Year 1	Year 2	Year 3
Expenses			
Salary	$75,000	$75,000	$75,000
Employee Related Expenses	$1,500	$1,500	$1,500
Marketing & Promotion	$120,000	$120,000	$120,000
Rent	$180,000	$180,000	$180,000
Utilities	$20,729	$22,653	$21,498
Office Supplies and Furniture	$61,176	$10,876	$11,569
Insurance	$60,000	$60,000	$60,000
Total Expenses	**$518,405**	**$470,029**	**$469,567**

Table 2: Budget of USI

Profit and Loss Statement

	Year 1	Year 2	Year 3
Income	**$3,400,000**	**$5,000,000**	**$7,400,000**
Direct Cost	**$170,000**	**$250,000**	**$370,000**
Gross Margin	$3,230,000	$4,750,000	$7,030,000
Gross Margin %	**95%**	**95%**	**95%**
Expenses			
Salary	$75,000	$75,000	$75,000
Employee Related Expenses	$1,500	$1,500	$1,500
Marketing & Promotion	$120,000	$120,000	$120,000
Rent	$180,000	$180,000	$180,000
Utilities	$20,729	$22,653	$21,498
Office Supplies and Furniture	$61,176	$10,876	$11,569
Insurance	$60,000	$60,000	$60,000
Total Expenses	**$518,405**	**$470,029**	**$469,567**
Operating Income	**$2,711,595**	**$4,279,971**	**$6,560,433**
Income Taxes	$223,707	$353,098	$541,236
Net Profit	**$2,487,888**	**$3,926,873**	**$6,019,197**
Net Profit/Sales	**73%**	**79%**	**81%**

Table 3: Profit and Loss statements

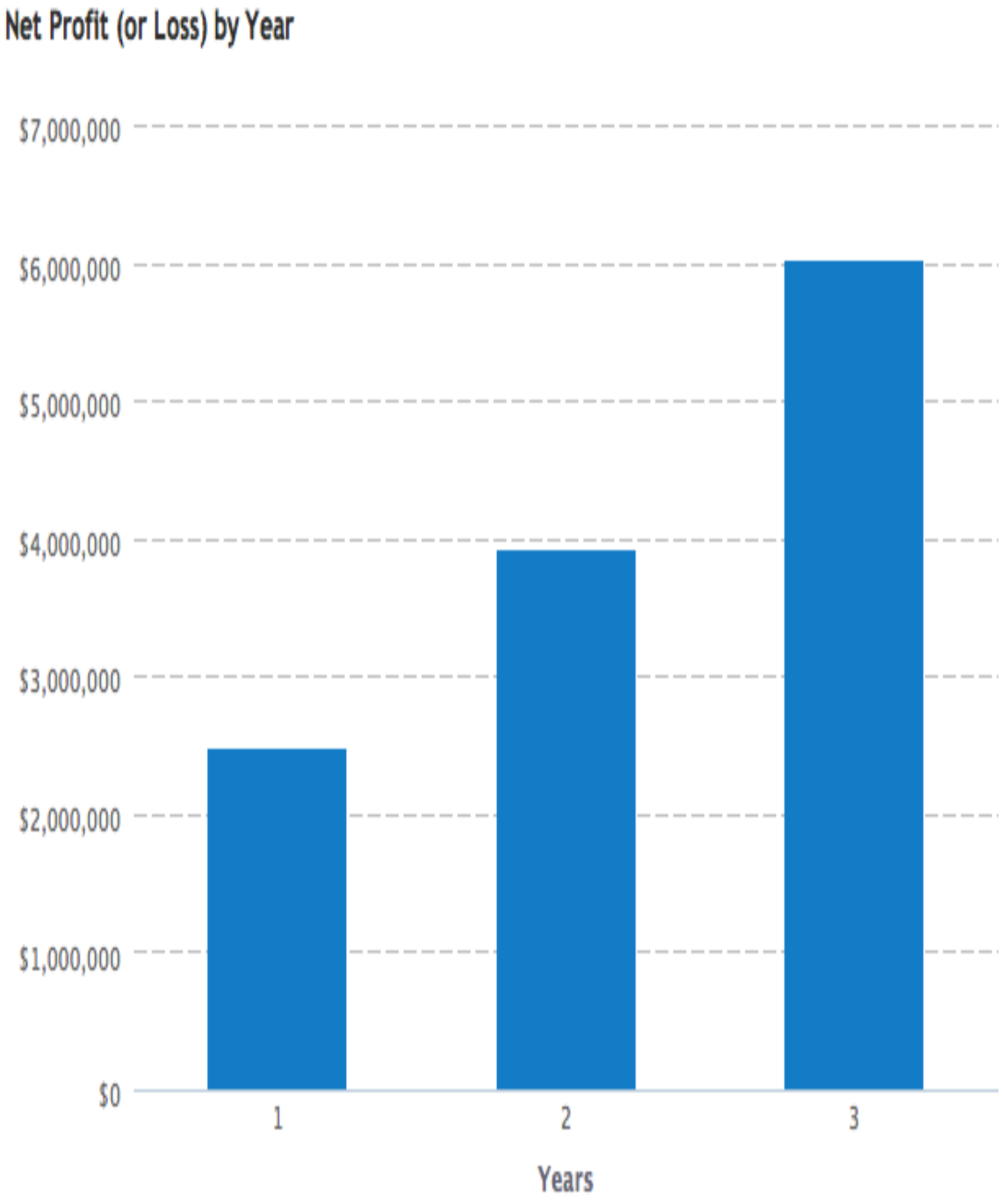


Table 4: Net profit

Sales Forecast Table

	Year 1	Year 2	Year 3
Unit Sales	34	50	74
Price Per Unit	$100,000	$100,000	$100,000
Total Sales	**$3,400,000**	**$5,000,000**	**$7,400,000**
Direct Cost Per Unit	$5,000	$5,000	$5,000
Total Direct Cost	**$170,000**	**$250,000**	**$370,000**
Gross Margin	$3,230,000	$4,750,000	$7,030,000
Gross Margin %	**95%**	**95%**	**95%**

Table 5: Sales Forecast table

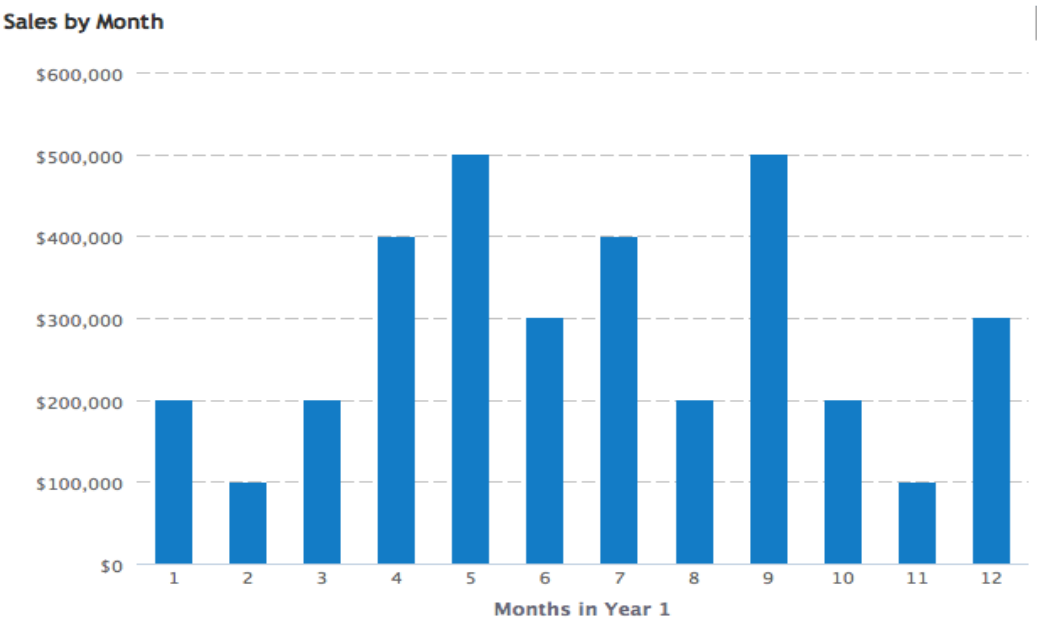


Table 6: Sales by month

7. Break Even Analysis

We calculated the break-even analysis of USI's Software by considering total revenue and total cost for each quarter over a time period of three years.

We calculated it using

Revenue = Total Number of customers*total number of software's purchased

Total Cost = Fixed Cost + Variable Cost

Break Even point = (TC = TR)

The following graph shows that, we obtained the breakeven point at the start of Q7.

The below graph shows the break-even analysis of USI

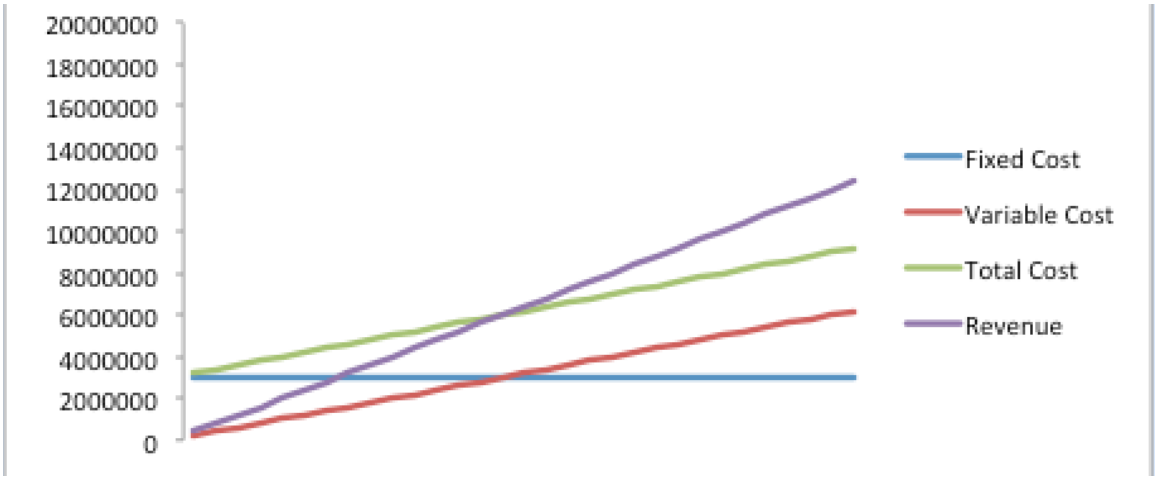


Figure 12: Breakeven Analysis

8. Exit Strategy

"It's not enough to build a business worth a fortune", every business has to have an Exit Strategy or we can say a way to get the money back out. We, at Ultimate solutions Inc. are planning to follow the **liquidation strategy** if in the near future there comes a day, when we are no longer able to run the business. If we liquidate our business, any proceeds from the assets must be used to repay our creditors. The leftover money gets divided among the company shareholders.

9. Project Schedule

Our team divided the tasks and came up with the targets to achieve in certain deadlines for the each task. These tasks and achievable were distributed among our team based on the skill set of the team members of USI. For example, the person who has experience in the software development was assigned the task to develop the software and the IT experts were given tasks related to setting up servers, and intercommunication. The following graph below gives a preview of the various tasks and targets achieved in the timely manner in the form of GANTT chart.

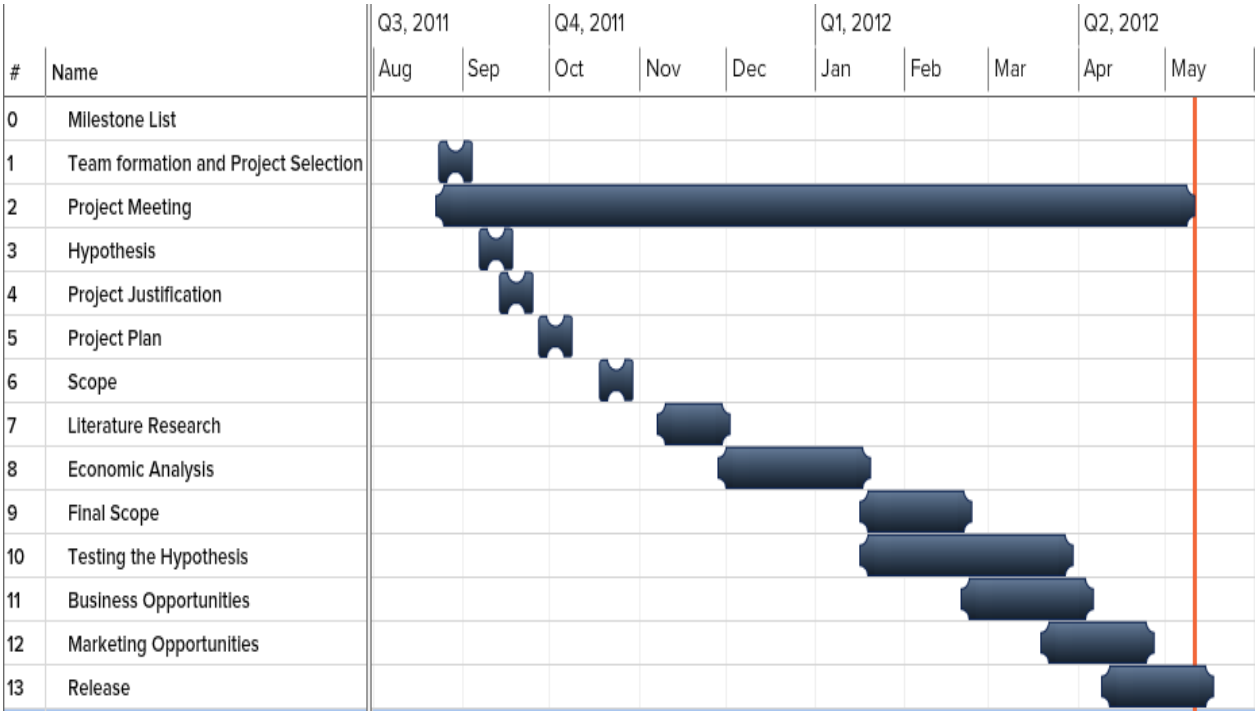


Figure 13: Gantt Chart

10. Intellectual property:

As we are using the variable sized chunking technique along with the Rabin Fingerprinting algorithm to implement and develop our Data De-duplication Software. So this is exclusively our idea and we own the set of exclusive rights for it under the patent law acts. USI has the right to decide who may - or may not - use our Data De-duplication invention for the period in which it is protected. USI may give permission to other firms to use the invention on mutually agreed terms and conditions.

11. Future scope of the project

In the scope of this project we have decided to implement and develop the Data De-duplication software by using the variable sized chunking method. We mainly developed the software for de-duplicating data that is stored in the mail servers. So basically our software mainly focuses on the emails. In the future, we would be designing de-duplication software that will take an important role in disaster recovery that will save disk storage space by copying the same data of one deduplication platform to another located at other off site. This would decrease the need to move magnetic tapes back and forth, which can be particularly meaningful when copying hundreds of terabytes of the data.

12. Conclusion

We are strongly convinced that there is an excellent business potential with Ultimate solutions Inc. Thanks to our market research we have identified the challenges faced by the potential customers that are using the data de-duplication techniques currently available in the market. We have developed a business by overcoming the various

challenges faced by the potential customers. We have confirmed through our research that the small and medium business have been not seriously pursued so far by the big players in the market. Also, estimates predict many small businesses to soon adopt the data de-duplication model for a variety of reasons. Thus we are convinced that the market segment we are focusing is an appropriate one.

Our business model provides a relief from the challenge faced by the companies that are using previous Data De-duplication techniques. In addition, we setup our own Data De-duplication environment and performed various tests to make sure that security precautions are in place. In addition, customers have the opportunity to run the Data De-duplication software for a free trial. Thus, Ultimate Solutions Inc is a one stop place for all those who are either "Data De-duplication Ignorant" or "Data De-duplication Skeptic" or "Data De-duplication Confident".

As a result of our market survey and economic justification, we are confident that there is abundant potential in this business. The profit loss statement, cash flow analysis, break-even analysis, Return on Investment, revenue and expenditure charts assure us that this venture will be profitable to both us and the investors. We are breaking even at the end of the second year and the investors.

13. References

1. Maddodi.S, Attigiri.G, A.K. (2002),"Data Deduplication techniques and analysis"; IEEE Explore, 664-668.
2. Mark W. Storer Kevin Greenan Darrell D. E. Long Ethan L. Miller. 2008. Secure Data Deduplication. StorageSS'08, October 31, 2008, Fairfax, Virginia, USA. 2008, 1-10.
3. Austin Clements, Irfan Ahmad, Murali Vilayannur, and Jinyuan Li. Decentralized deduplication in san cluster file systems. In Proc. of the USENIX Annual Technical Conference, June 2009.
4. Geer,D. (2004),"Reducing the storage burden via data deduplication"; IEEE Explore, 15-17.
5. Lu.G, Jin.Y, H.C. Du (2010)," Frequency Based Chunking for Data De-Duplication"; IEEE Explore, 287-296.
6. Lawrence L. You, Kristal T. Pollack, Darrell D. E. Long. Deep Store: An Archival Storage System Architecture in Proceedings of the 21st International Conference on Data Engineering. April 2005, pp. 804 -- 815.
7. Muthitacharoen, B. Chen, and D. Mazi`eres, A low-bandwidth network file system. In Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP '01). October 2001, pp. 174-187.
8. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicaterecord detection: A survey. Knowledge and Data Engineering, IEEE Transactions on, 19:1-16, 2007.

9. Bolosky WJ, Corbin S,Goebel D,Douceur JR.Single instance storage in Windows 2000.In :Proc.of the 4th Usenix Windows System. Symp.Berkeley: USENIX Association,2000. 13-24.