ANALYSIS OF BIG DATA THROUGH DE-DUPLICATION TECHNIQUE

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

Master of Engineering

in

Software Engineering

Submitted By

Sanjeev Garg

(Roll No. 801431024)

Under the supervision of:

Dr. Anju Bala

Assistant Professor

Thapar University, Patiala

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2016

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "***Analysis of Big Data through De-Duplication Technique***", in partial fulfillment of the requirements for the award of degree of Master of Engineering in ***Software 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. Anju Bala*** 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:
Sanjeev Garg
Sanjeev Garg
801431024
ME(SE)

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

Anju Bala
Dr. Anju Bala
Assistant Professor
CSED, Thapar University

Countersigned by

Dr. Maninder Singh
Head & Associate Professor
Computer Science and Engineering Department
Thapar University
Patiala

Dr. S. S. Bhatia
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. It is a great privilege to express my gratitude and admiration towards my respected supervisor **Dr. Anju Bala** Assistant Professor Computer Science & Engineering Department. She has been an esteemed guide and great support behind achieving this task. This work would not have been possible without the encouragement and able guidance of her. I also thank my supervisor for her time, patience, discussions and valuable comments. Her enthusiasm and optimism made this experience both rewarding and enjoyable. I am truly grateful to her for extending her total co-operation and understanding whenever I needed help and guidance from her. I am also heartily thankful to **Dr. Maninder Singh**, Associate Professor and Head, Computer Science & Engineering Department and **Rupali Bhardwaj**, PG coordinator, for motivation and providing uncanny guidance and support throughout the preparation of the thesis report.

I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academic Affairs, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable. Last but not least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Sanjeev Garg

Sanjeev Garg

(80143102)

Abstract

As the data available on the web is in heterogeneous formats such as text, video, audio etc. Hence, there is need to integrate the data from the different sources and analyze the data which can be utilized for efficient query execution. If data is not analyzed properly then execution time for the user query processing will be more and result also will not be according to user need. So, there is need to analyze the data after combining different formatted data into same format. After integration, data becomes large and there is need to used different type of de-duplication techniques to analyze data. Because the different formatted data may contain same record so there is chance of redundancy of data. There are different data de-duplication techniques for removal of redundant or similar data. There is another a de-duplication technique has been introduced in which format comparison of data is checked after integrating heterogeneous data in same format. Finally, the experimental results validate the efficiency in terms of execution time, storage space and success.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Big Data

With advancement of technology in recent years and embedding up of the technology in the day-to-day lives of people has created lump sum amount of data in magnitudes of terabytes. The introduction up of Big Data is due to the fact that with increase in complexity of data the process of analyzing the data cannot be achieved using traditional data processing applications. Data is called big when it becomes difficult to handle it by any database management tool. But having bigger data consequently requires new tools and technologies to solve old and new problems of big data in a better way.

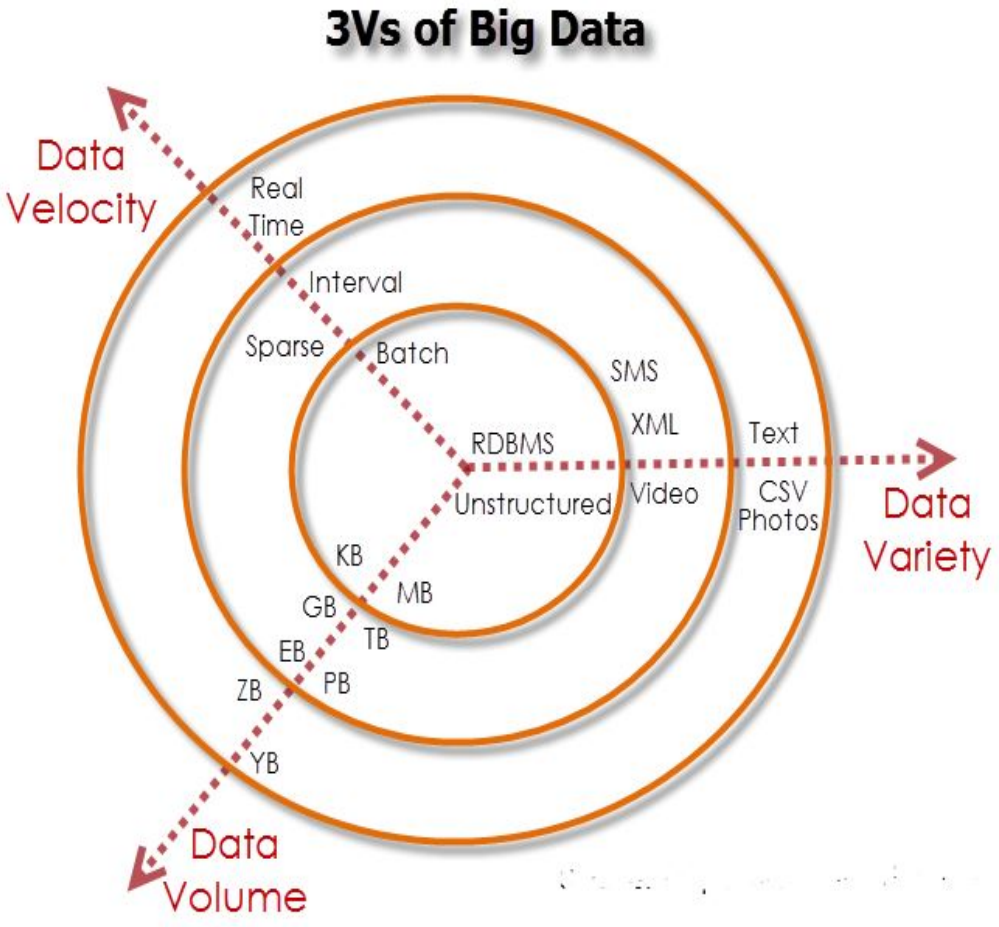


Figure 1.1 3 Vs of Big Data [23]

- More than 1 million customers transactions handled by Walmart in every hour.
- 40 Billion photos are handled by Facebook from its client base.
- 10 years ago translation of the human genome was very hard process. It took 10 years to process but now it can be accomplished just within a week.

1.2 Characteristics of Big Data

1.2.1 Volume

• An ordinary Personal computer can store only 10 GB data in storage.

• Now days, only Facebook produce 500 terabytes(TB) of new data or information consistently.

• 240 terabytes(TB) of flight information amid a solitary flight over the United State by Boeing 737.

• The advanced or propelled mobile phones, the information they make or produce and devour; sensors installed into regular articles will soon bring about billions of new, continually overhauled information encourages containing natural, area, and other information, including video.

1.2.2 Velocity

• On every occasion velocity of promotion and click streams increased to impress the client in every second.

• High-recurrence stock exchanging or stock trading calculations reflect market changes inside microseconds

• Form trading information is there between machine to machine i.e between billions of devices

• Huge log information is produced by base and sensors every day.

• On-line gaming systems or gadgets and structure bolster a great many concurrent and simultaneous clients, each client is creating various inputs every second.

1.2.3 Variety

- Big Information isn't simply numbers, dates, and strings. Huge Information or tremendous data is likewise geospatial information, 3D information, sound and video, and unstructured content, including log records and internet organizing.
- Traditional database frameworks were intended or proposed to address littler volumes of organized information, less redesigns or an anticipated, steady or relentless information structure or data structure.
- Big Information examination fuses diverse sorts of information

1.3 Why Big Data

Key enablers for the growth of "Big Data" are:

- Increase of storage limits or capacities
- Increase of handling power or processing power
- Availability of data (Availability of information)

1.3.1 Data Store

Data storage is growing rapidly from analog to digital day by day. As shown in the Figure 1.2 from 1996 to 2007 data growth rate has increased.

Data storage has developed significantly, moving markedly from analog to digital or simple to computerized after 2000. Global or worldwide installed optimally compressed, storage.

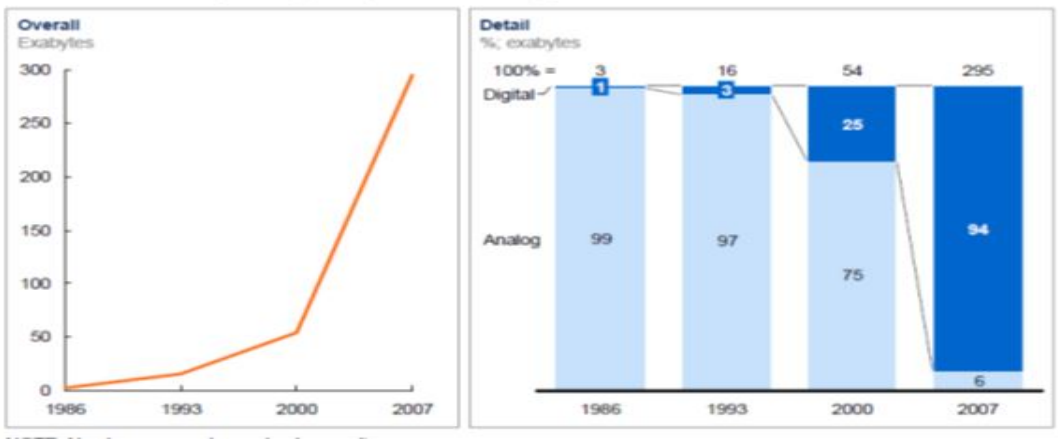


Figure 1.2 Data Storage[24]

1.3.2. Computation Capacity

Computation capacity has also risen since 1986.For computation of information we have used the following .Computation capacity has also risen sharply. As shown in the Figure 1.3 there are two graph .First graph represents the overall computation capacity of information and second graph shows the detail view of the computation capacity of the information.

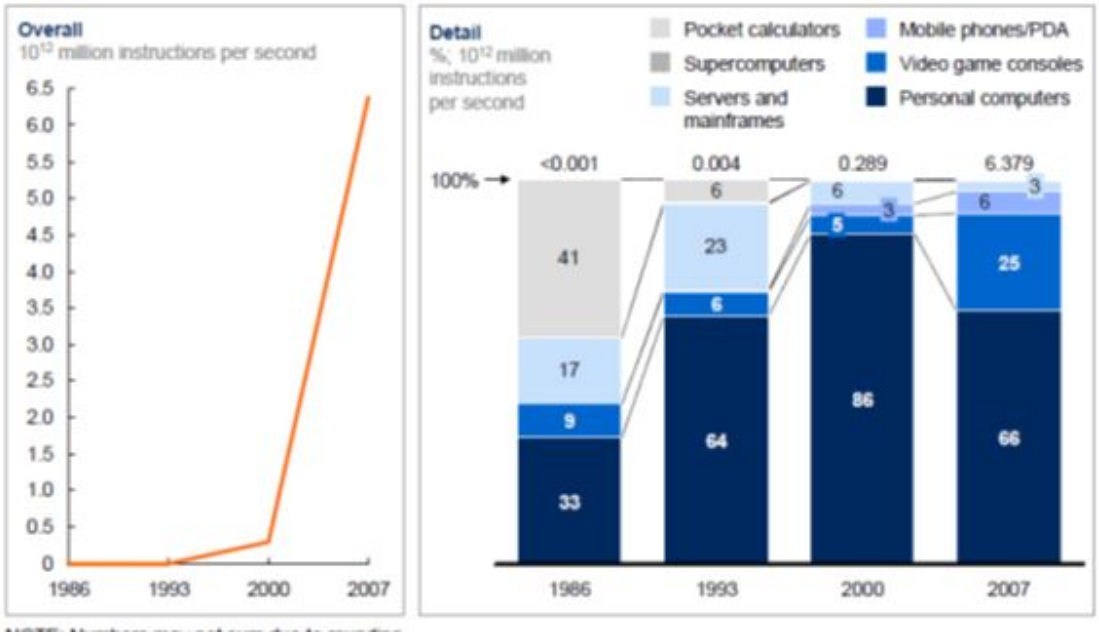


Figure 1.3 Computation Capacity [24]

- Pocket Calculator
- Supercomputer
- Server and Mainframes
- Mobile phones
- PDA
- VGC(Video game console)
- PC(Personal Computers)

1.3.3 Data Availability

Now a days availability of data is increasing day by day. Data is available in all fields as

- Government
- Communication Media
- Process Manufacturing
- Health Care Providers
- Security and investment services
- Retail

Companies in all parts have at least 100 terabytes of capacity information or storage data in inited states; many have more than 1pb (petabyte).

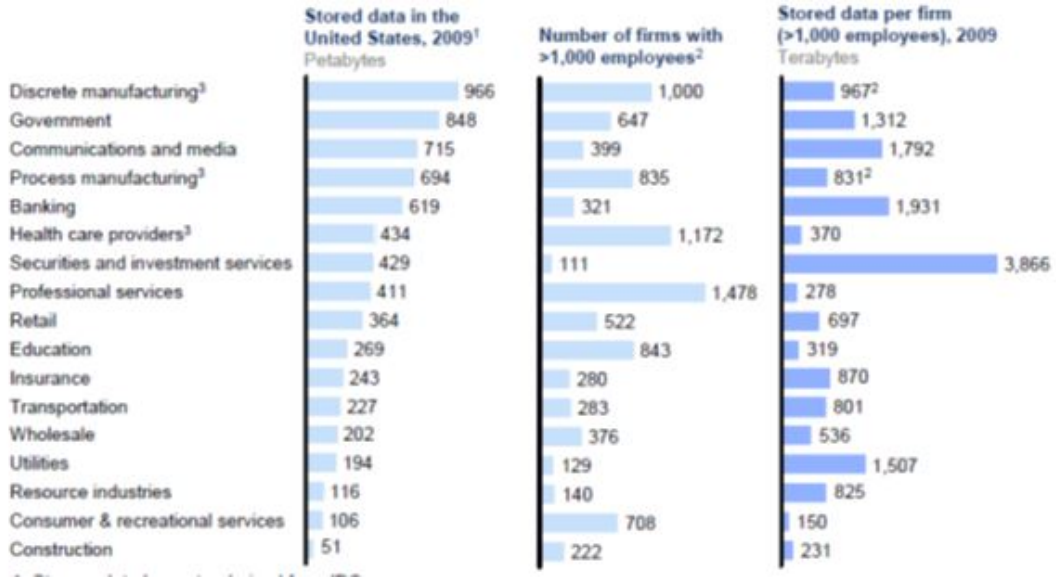


Figure 1.4 Availability of Data [24]

- Storage data or information by area from IDC.
- Firm data split into areas when required using employment.
- Particularly huge number of firms in manufacturing and human care provider or supplier sectors makes the capacity per company much littler.

  These are the following firms

  - Education

- Insurance
- Transportation
- Wholesale
- Resources Industries
- Construction etc

1.4 Type of Data stored

In each sector data can be stored in different format. As it can be in video, image, audio and text format. For each type of data there penetration can be different. It can be High, Low and Medium.

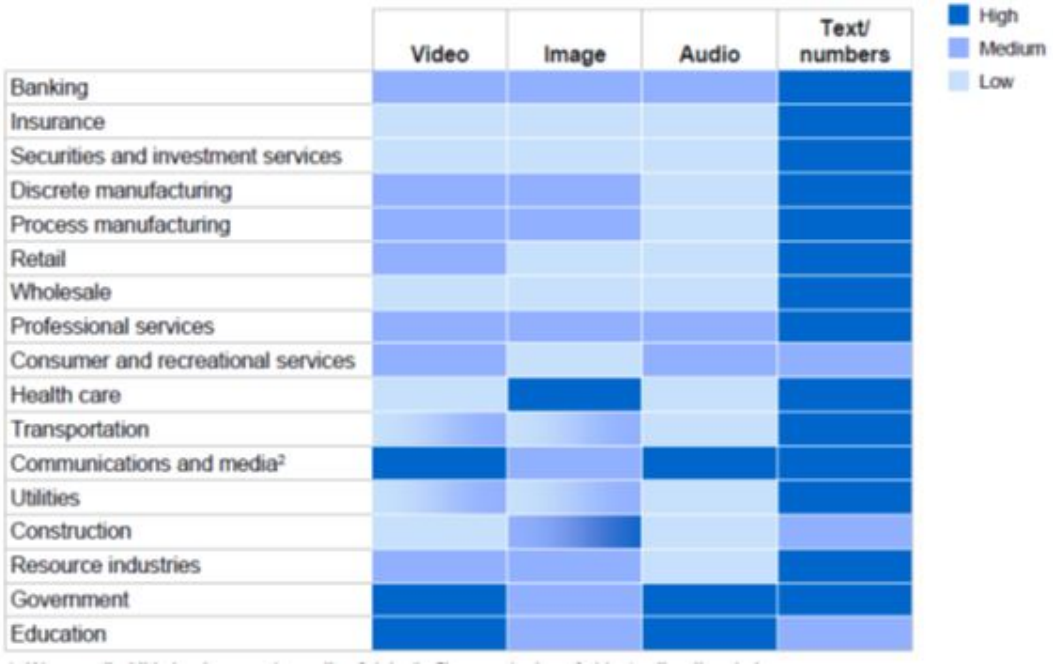


Figure 1.5 Types of Data [24]

- We compiled this warmth map using units of data(in files records or minutes of video) rather than bytes.
- Video and audio sound are high in some subsectors

There are different tools and technologies used for analysis of big data. As the amplitude of data that is increasing on internet started to question that how to handle this data.

1.5 De-duplication

Data de-duplication is the process of removal of similar data or redundant data is called data de-duplication.

Data de-duplication is basically divided into three parts based on location where data de-duplication is to be performed, data unites and disk placement based.

I. Data Unit Based
II. Location Based
III. Disk Placement Based

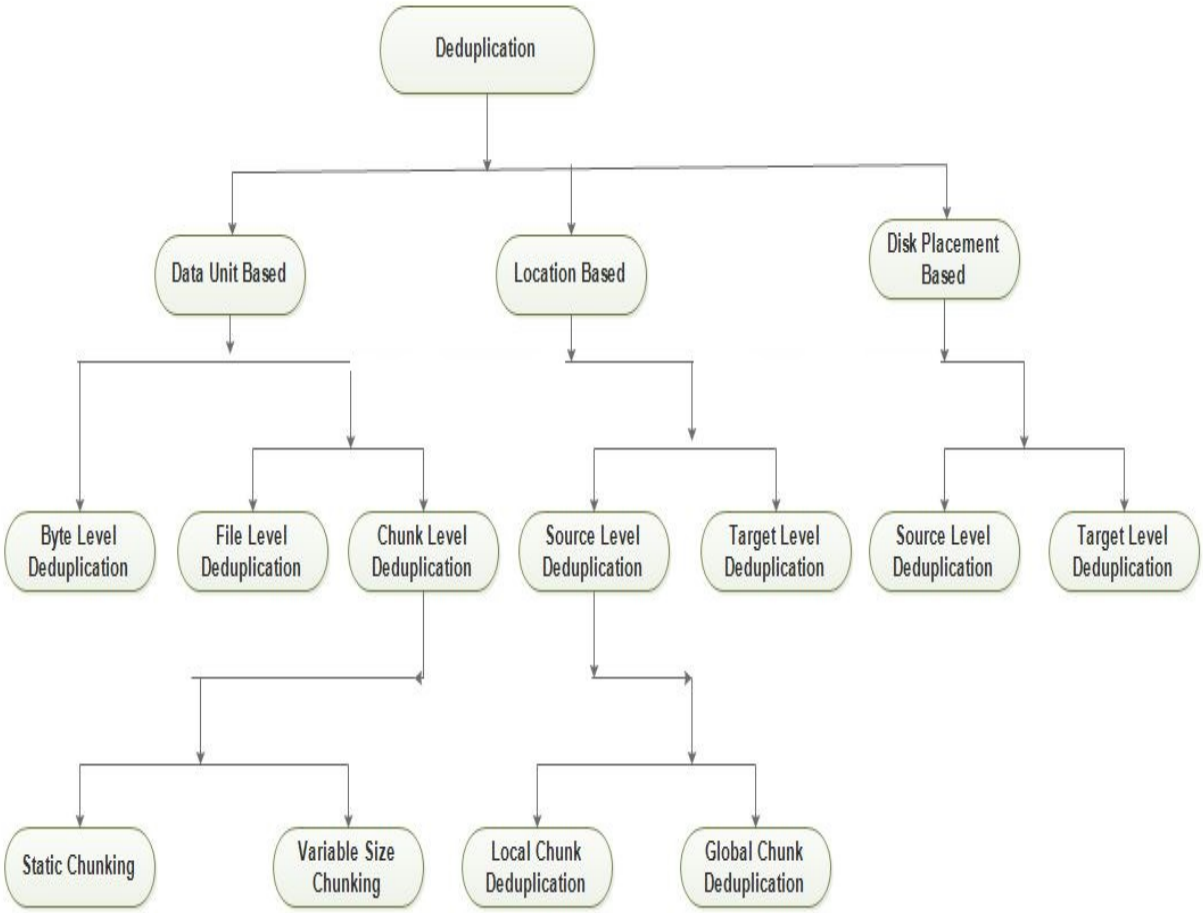


Figure 1.6 Different Data De-duplication Techniques

1.5.1 Data Unit Based

Data unit based data de-duplication technique is divided into three parts based on data units.

1.5.1.1 Byte Level De-duplication

In this type of data de-duplication technique the data is to be send in form of bytes over the network. Every byte of data is to be checked for finding similarity of data. If a byte is similar to its previous data then this byte of data is not to be placed on the network. If a byte of data is not similar to its previous byte of data then this data is to be placed over the network. This type of data de-duplication is said to be byte level data de-duplication technique [9].

1.5.1.2 File Level De-duplication

File level data de-duplication is related to files of data. In this only one copy of file is to be stored if the hash value of the file is to be same i.e if two files have the same hash value then that file is to be said as identical file and if both the files have different hash value then these file is to be said as not identical or unique files [8][2].

1.5.1.3 Chunk Level De-duplication

In the Block Level data de-duplication or Chunk Level Data de-duplication the each file is fragmented into number of blocks. Only one copy of each block is to be stored on storage device. If a block with same id number is present on storage device then it is not stored again. If a block with same id is not present on storage device then this block is to be stored on stored device. It is further divided into two parts [7].

- **Static chunking-:** In static chunking each block is to be divided into same size of block. Every block that is to be divided is of the same size.
- **Variable size Chunking-:** In the variable size of chunking each block is to be divided into different size.

1.5.2 Location Based

Duplication of data can be eliminated on different location. If redundancy of data is to be eliminated on different location then that type of data de-duplication technique is called location based data de-duplication technique.

1.5.2.1 Source Level De-duplication

When elimination of redundant data is to be performed on client site or where data is created rather than where data is stored is called source level data de-duplication. Further source level data de-duplication is to be divided into two parts.

i. **Local Chunk De-duplication**

In the local chunk level data de-duplication redundant data is to be removed before Sending it to the destination where the data is to be stored.

ii. **Global Chunk De-duplication**

In the global chunk level data de-duplication technique redundant data is removed at global for each client.

1.5.2.2 Target Level De-duplication

When elimination of redundant or similar data is to be performed on target site where the data is stored is called target level data de-duplication technique. In this data de-duplication the client does not know any technique of removal of redundant data. This type of technique increase the processing time but increase the bandwidth also.

1.5.3 Disk Placement Based

Based on how data is to be placed on disk data de-duplication technique is to be used either forward reference or backward reference technique.

1.5.3.1 (Forward-Reference)Source Level

In the forward reference recent data chunks are maintained and all the old data chunk are associated with pointers that points forwards to the recent chunks.

1.5.3.2 (Backward-Reference)Target Level

It introduces the more fragmentation for the past data chunks.

1.6 Karma Data Integration Tool

In order to apply Karma to the problem of big data, we plan to begin with the same center capabilities to be able to quickly model or show sources, which permits us to automate or robotize many of the required changes or transformations, and then develop new information restructuring capabilities and then execute this rebuilding on big or huge datasets.

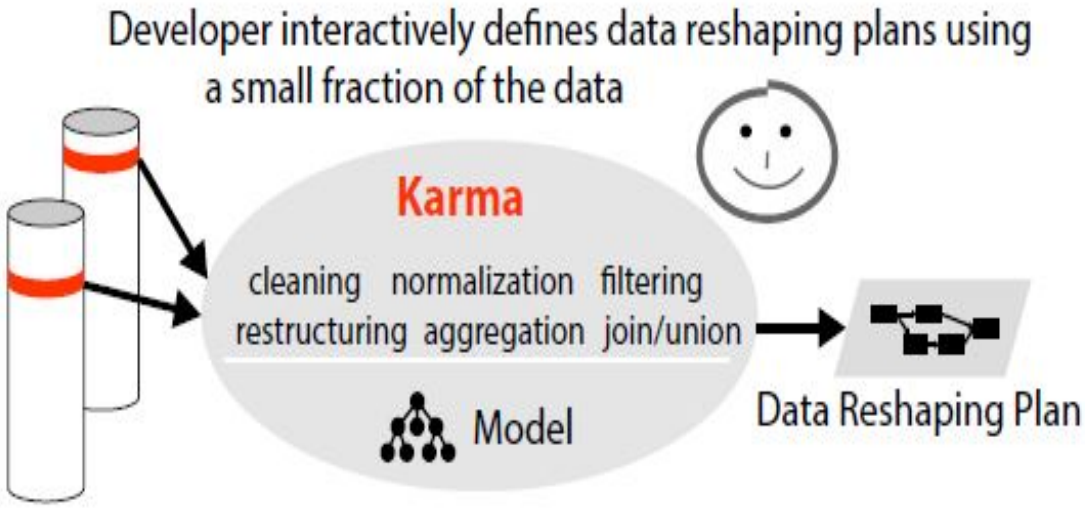


Figure 1.7 Data Restructuring plan [1]

The way that the system or framework can rapidly build a model of a source means that Karma would enable or empower a user(client) to quickly define a rebuilding plan. As shown in Figure 1.7, the user would define the restructuring plan on a small subset of the information or data,

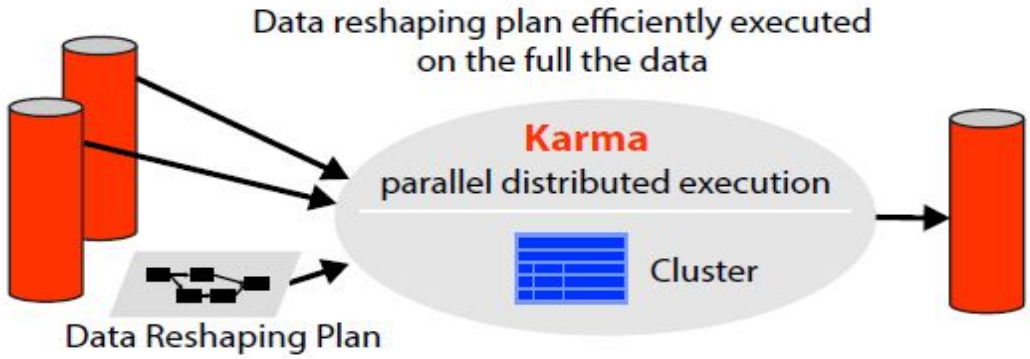


Figure 1.8 Execution plan [1]

and then Karma would build or fabricate the general plan and execute that plan in a circulated or distributed environment over the entire dataset.

Map Reduce is an implementation that is used to handle large amount of data.

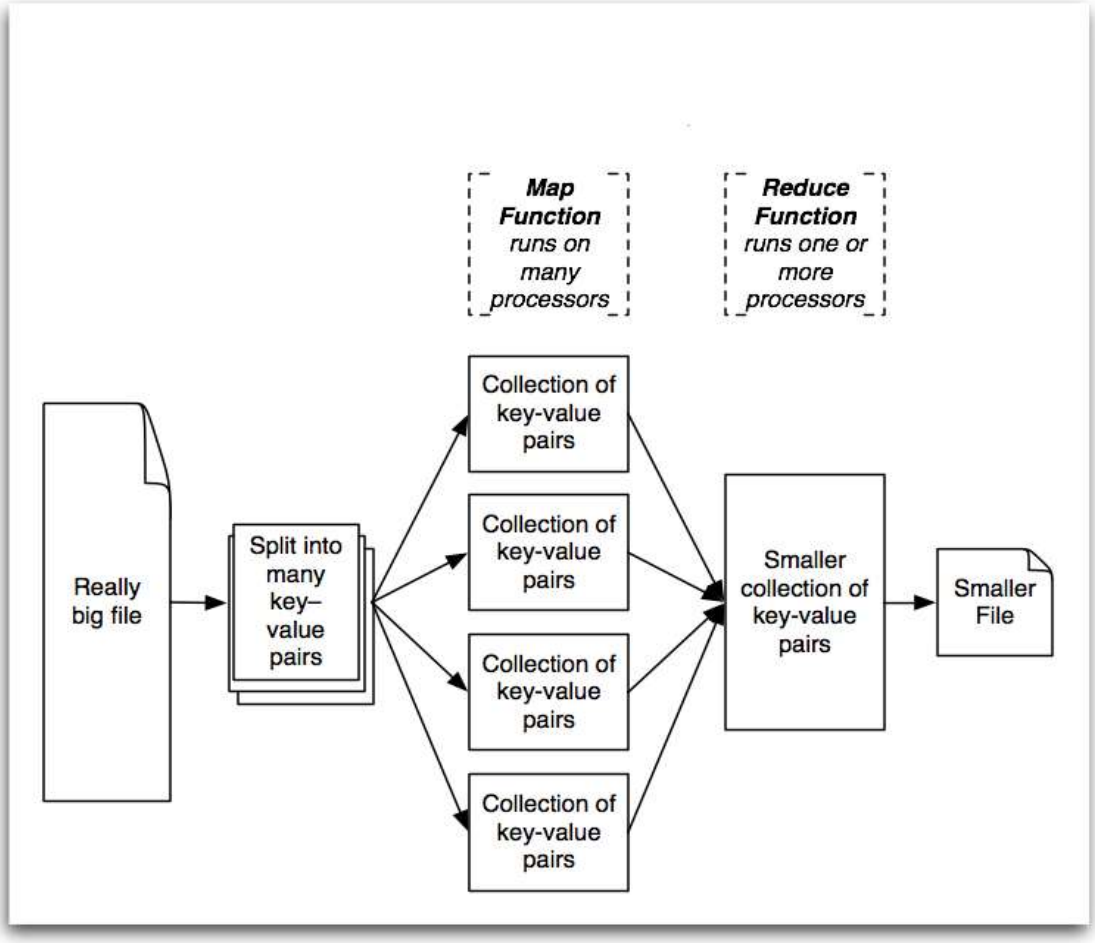


Figure 1.9 Mapreduce Framework [25]

The most popular technology that is able to sort and mine the data is called Hadoop. It is the most implemented solution for handling the big data. Hadoop also use Hadoop file distributed system .Most of the companies as Google , Facebook uses Hadoop for analysis of big data.

There are another techniques used for analysis of big data that are data de-duplication techniques.

1.7 Research Motivation

Even today Facebook, twitter and other social networking site produces large amount of data in every second. So, there is large scope on big data. As there are many tools and technologies to analyze the data. There are another techniques to analyze the big data i.e Data de-duplication techniques.

As when the data is increasing then there are more chances that data is repeated or redundant that needs to requirement of large storage devices. It also increasing the processing time of the query. So there is a large scope to analyze the data by data de-duplication techniques so that we can reduce the requirement of large storage media and we can reduce the processing time to execute the query.

1.8 Thesis Outline

Thesis structure for the research work done is shown in Figure 1.10.

 Chapter 1 Introduces thesis work.

 Chapter2 Literature review is briefed in this chapter, the basics of the research topic and existing approaches are discussed.

 Chapter 3 Research gaps, identified problem and objectives of thesis are mentioned in this chapter.

 Chapter 4 The Proposed approach for fault prediction and reliability improvement is described in this chapter, fault monitoring system configuration and statistical analysis distribution model is also given.

 Chapter 5 Implementation and Results are discussed. Interface design of the proposed methodology is described and graphical results are given in this chapter.

 Chapter 6 Conclusion of the thesis work, its contribution and future scope of the proposed approach is given in this chapter.

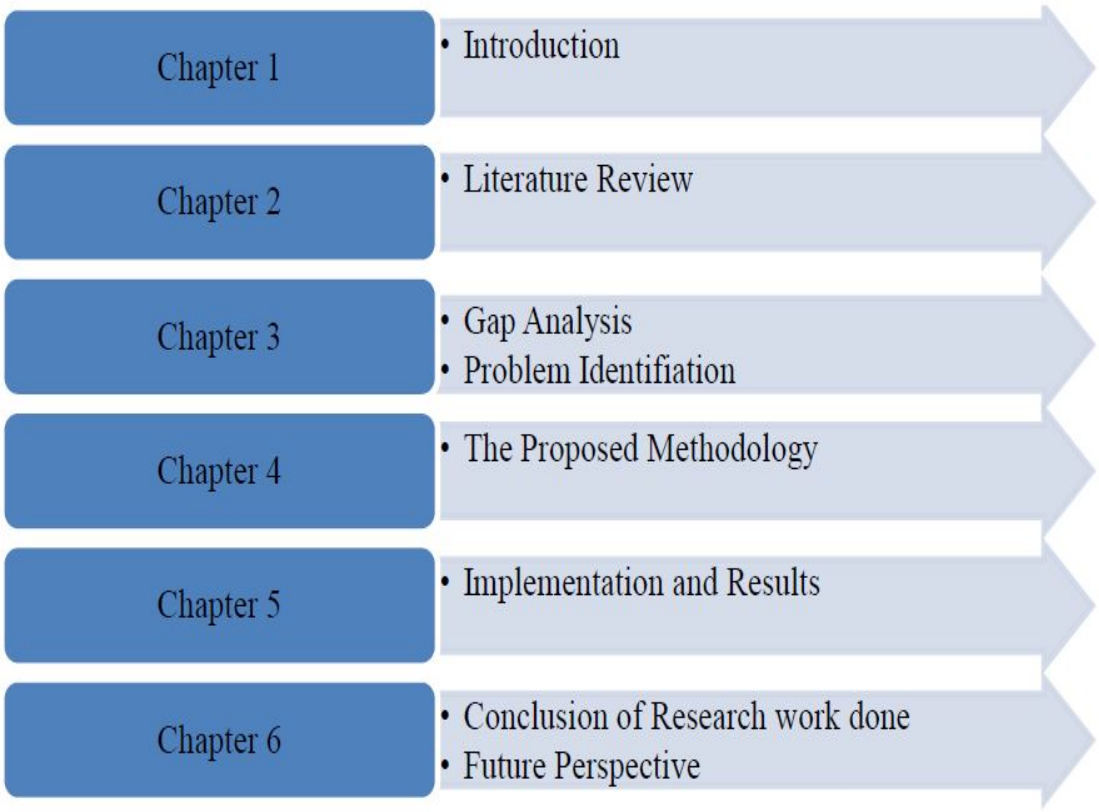


Figure 1.10 Thesis Outlines

In this chapter Big Data and Data de-duplication techniques is discussed. Research motivation and thesis outline are also given. Conclusion of this chapter is that Data de-duplication in Big data is an efficient approach so that processing time of query can be improved.

Chapter 2

Literature Survey

2.1 Introduction

In the accompanying part, investigation is performed on different information de-duplication strategies. The following are edited compositions of those investigations performed by different exploration colleagues.

Organizations are growing day by day and they are following their own format for storing big data. However, the organization needs to merge big data sets in different format and analyze it for better query processing. There are different tools and technologies that are being used for analysis of big data. Various tools are being used to integrate different data format. Data is converted into different format, analyzed and then we perform query processing on this filtered data

2.2 Related Work

Zillow [9] is a commercial website that is helping home buyers and sellers by offering information about real estate like property tax and historical pricing by batching the data from various websites [1]. Main drawback of this website is that it does not cover all aspects of the data integration.

Fagin et al.(2009) proposed similar work in schema mapping[10]. This research covers seminal work done on Clio which provides a mechanism that supports additional schema constraint (Marette al. 2011). Similarly Alexe et al. (2011) created schema mapping by using examples of prime source data elements and the corresponding elements over the required target schema.

Knoblock and Carman [1] in 2007 also used various sources of data to create a GLAV mapping for a target source. The main limitation of this work is that this approach could only learn descriptions having conjunctive combinations from known source descriptions.

By searching paths in the field of ontology along with patterns of known sources, it is possible to hypothesize required target mappings which are more basic than previous source information and their combinations. Many other systems exsits supporting data cleaning and data transformation. Artermis [15] And Agent wizard [16] utilize the inquiry noting procedure to construct an arrangement that incorporates different information sources.

Potter's Wheel and OpenRefine5(Raman and Hellerstein) in 2001 make possible for users to specify and edit operations. OpenRefine is a analysis tool for clearing redundant data. Potter's Wheel provides sets of transformation operation that lets user to gradually make transformations by submitting or undoing listed transformations through interactive GUI.

There are different data de-duplication techniques that are used for analysis of data. Compression is different from data de-duplication technology [13][14].

Table 2.1 Different Data de-duplication techniques

De-Duplication	***Des.***	***I/O***	***Eff.***	***Thro.***	***FC***
Chunk Level	Block compared	H	H	L	N
File Level	File Compared	M	M	L	N
Byte Level	Every Stream of Byte is compared	L	M	M	N

Here

I/O-> input operation.

Des-> description

Eff-> efficiency

Thro->throughput

FC->format comparison.

Here H represents high, M represents medium, L represents low and N represents not applicable.

De-duplication is the technique used for removal of redundant data from the cloud. There are three types of data de-duplication technique.

The de-duplication is a very important technique for the shared storage [6] .Chunk level (Block-level) data de-duplication technique is utilized to stream data that is divided into blocks [7]. Each block is provided with unique identification number depending on which different blocks are to be stored on the storage media. If a block is already available then that block is not stored again on the storage media otherwise that block is stored on the storage media.

File-level de-duplication [8][2] is usually referred as Single Instance Storage (SIS) [11]. It verifies index backup of the file by comparing the indexes. If the file has same index as the file that is to be placed then that file will not be placed again otherwise that file is to be placed in the storage media and index of the file is to be updated.

Data de-duplication technique at Byte level [9] required analysis of data at the level of byte stream. It is a data de-duplication technique in which each byte stream is checked for data de-duplication i/o operation. Input Output operation is the number of operation required to transfer the data from secondary memory to primary memory.

Efficiency is the total efficient work that is to be done. The disadvantage in chunk level data de-duplication, file level data de-duplication and byte level data de-duplication techniques is that we cannot do the format comparison in all these methods. It means that if in a column a name is as William J. Smith and in the same column another record is given as Smith, W. J. being common in both the records are same but the format is

different so this type of data duplication is not identified by the chunk level data de-duplication, file level data de-duplication and byte level data de-duplication technique.
It is observed that throughput of de-duplication is better than compression [27]. Keeping compression ahead, measure of hash estimation is diminished. Then again, de-duplication before compression diminishes the volume of information to pack. Table 2.2 compares the existing de-duplication techniques.

Table 2.2 Comparison of Existing De-duplication Techniques

Techniques	**Meta Data Processing**	**Chunking Method**	**Chunk Granularity**	**De-duplication Scalability**
De-duplication using byte index chunking method[29]	Index-matrix table	Fixed-Size	Chunk	Small scale storage
Probabilistic De-duplication for Cluster-based storage[29]	Bitmap Vector	Content-based	Super-Chunk	Cluster based storage
Scalable De-duplication using data routing technique [29]	Similarity Index	Variable-length	Super-Chunk	Cluster based storage
De-duplication using multi-layer metadata[29]	Tree map Global and local metadata	Variable-length	Chunk	Small scale cloud based storage

Application aware De-duplication for cloud backup[29]	Local and Global metadata	All types of chunking methods	Chunk	Large scale cloud based storage

Concerning De-duplication, throughput alludes to measure of information that can be de-copied in a given measure of time. It has been watched that accomplishing throughput may influence de-duplication proportion (i.e. proportion of information before applying de-duplication to information after the de-duplication process). [17][18][22][19][20] achieves throughput to some extent. Jingwei Mama. et.al [21] proposes a pipelined and directed model for the de-duplication errands (piecing, hash count, compression, duplication recognition) in de-duplication. In the above table we have defined different method for the data de-duplication techniques. It shows the comparison of different data de-duplication technique on the basis of following parameter as

- **Processing time**
  it is time required to process the data from a cloud storage for a better query execution.
- **Scalability**
  Scalability defines that how much storage space we can increase for large amount of data by removing redundant or similar data.

Chapter 3

Research Problem

This chapter tells about the gaps encountered during the research by reviewing the already existing literature in the area of semantic analysis of big data problem in cloud storage.

3.1 Problem Statement

As the data available on the web is in heterogeneous formats such as text, video, audio etc. Hence, there is need to integrate the data from the different sources and analyze the data which can be utilized for efficient query execution. After integration the data becomes large and there is need to use different techniques to analyze data. Thus, a de-duplication technique has been introduced by integrating heterogeneous data in same format. Further, to analyze large amount of data, record level data duplication is being applied on integrated data that is also used for removing similar type of data. At long last, the exploratory results accept the effectiveness as far as execution time, storage room and achievement.

3.2 Research Gaps

This section tells about the gaps encountered during the research by reviewing the already existing literature in the semantic analysis of big data.

- For the semantic analysis of big data de-duplication is not checked at record level to find the relationship between two records [1].
- Optimized storage capacity for data is taken for one format of file i.e text.De-duplication is to be checked for text file [26] .

3.3 Objectives

The objectives used for this work are discussed below:

- To analyze the big data so that query processing can be fast for the useful information.

- To apply the data de-duplication techniques at record level for semantic analysis of big data.
- To apply the data de-duplication techniques on other format of data rather than only text format of data.
- To reduce the storage space for data by removing redundant or similar record or line.

Chapter 4

The Proposed Methodology

In this chapter the proposed methodology for solving the problem identified during gap analysis is described.

4.1 Tree-Similarity Search using Edit Distance

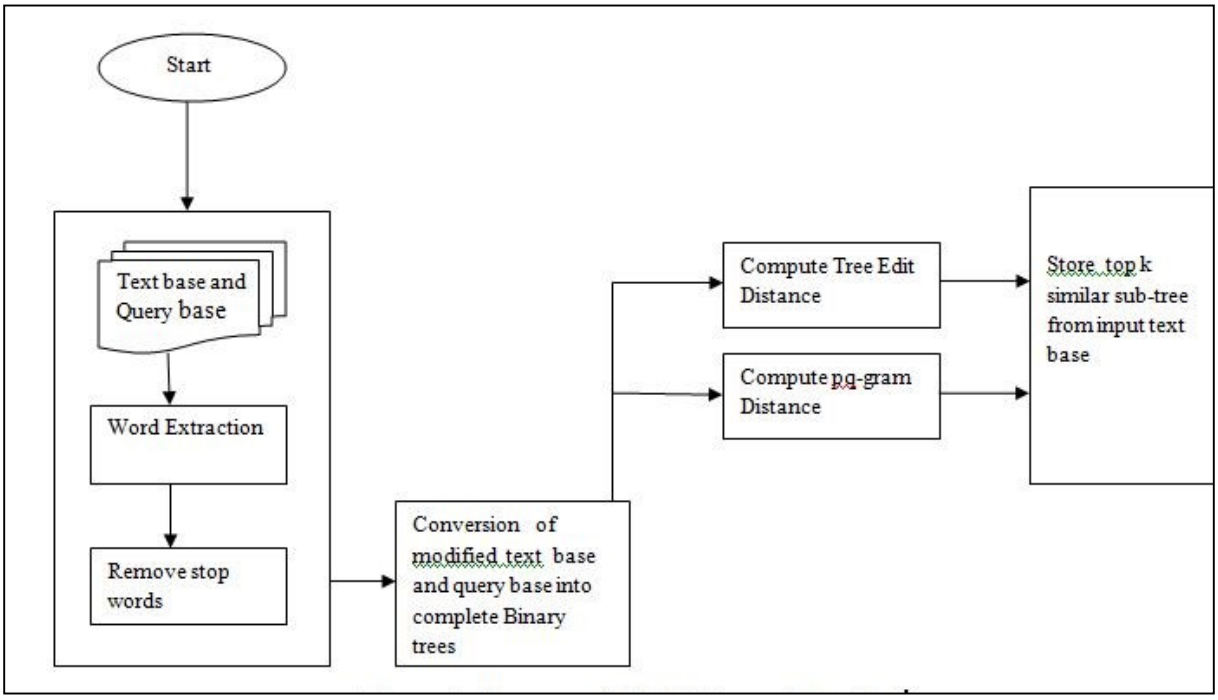


Figure 4.1 Tree-Similarity Search using Edit Distance

PSEUDO-CODE of Tree-Similarity Search using Edit Distance

Input: input documents/input text paragraphs and query text paragraph

Output: Most Similar tree in comparison with query tree

Step1: Ontology identification after scanning each of the input documents

Collect Input documents/text (Di) where i=1, 2, 3….n;

For each input Di;

Extraction of Word (EWi) = Di; // for all document apply extraction word process i=1, 2, 3…n in and extract words// For each EWi;

Stop Word (SWi) =EWi; // To eliminate the word like am, is, to, as etc apply stop word /

Step2: Construct Complete binary tree for each of input documents for each (SWi); where i =1, 2, 3….n;

Construct Complete Binary Tree (Ti)= (SWi);

Construct Complete Query Binary Tree(Q) from Query text paragraph after Removing stop words.

Step 3: Evaluate Edit distance for each (Ti, Q); where i =1, 2, 3….n and Q is Query text tree .

Edit distance can be find out using two approaches [28]:

- Unit cost function where all operation insert , delete and update have same cost.
- Multi cost function where all operation insert , delete and update have different cost.

Step 4: Find out and return Ti such that having minimum Edit Distance with Q , where i = 1, 2, 3…n

S(ind,value).value = min(S(ind,value).value, Edit Distance(Ti,Q)) .

Where S(ind,value) is pair of index and edit distance of most similar tree in.

comparison to query tree.

PSEUDO-CODE of Edit Distance

Input : Complete Binary Tree of query tree and text tree

Output : Edit Distance between two input tree [28]

Step1 : Traverse both input tree in Level Order Traversal .

S1 : Level Order Traversal of Query Tree

S2 : Level Order Traversal of Text Paragraph

Where S1 and S2 are strings of node , where each node represents a word from tree

Step 2 : Calculate 2-Dimensional array of Edit Distance

EDistance[i][j] = minimum number of operations to convert string S1[1..i] to S2[1…j]. where operations are insertion(I), deletion(D) and updation(U) . If I = D, D =U and U = I unit cost function ,otherwise multi cost function .

PSEUDO-CODE of Calculation of Hop-Count of Query tree in Text Paragraph

Input: input documents/input text paragraph and query text paragraph

Output: Most Similar tree with query tree

Step1: Ontology identification after scanning each of the input documents

Collect Input documents/text (Di) where i=1, 2, 3….n;

For each input Di;

Extraction of Word (EWi) = Di; // for all document apply extraction word process i=1, 2, 3…n in and extract words// For each EWi;

Stop Word (SWi) =EWi; // To eliminate the word like am, is, to, as etc apply stop word /

Step2: Construct Complete binary tree for each of input documents for each (SWi); where i =1, 2, 3….n;

Construct Complete Binary Tree (Ti)= (SWi);

Construct Complete Query Binary Tree(Q) from Query text paragraph after removing stop word .

Step 3: Traverse Query Tree and Text Paragraph in Level Order Traversal .

S1 : Level Order Traversal of Query Tree

S2 : Level Order Traversal of Text Paragraph

Where S1 and S2 are strings of node , where each node represents a word from tree

Step 4: Linearly search each element of S1 in S2 and store hope count.

for each index S1i

Find match of S1i in S2j and store hop-count of two consecutive node.

where i=1,2,3…size of S1 and j = 1,2,3…size of S2.

The overall architecture of the technique is shown in Figure 4.2 Overall architecture is divided into four layers,

Interface Layer - Interface layer provides the user interface to select the file for de-duplication .It also provides interface to select the type of de-duplication and also to specify the split length.

Record Level Layer- In the record level layer first content is stored after that format comparison of each record is checked to find the redundant data that is related to format of record. We can do compression of data also here to save the storage space

De-duplication Layer - It involves the detection of duplicate chunks by comparing hash values generated in the chunk layer. The chunks are compressed after eliminating the duplicate values.

Storage layer - After eliminating the duplicate values the compressed file is stored in the Amazon S3 bucket by using Amazon APIs. The compressed file is uploaded to Amazon storage.

This method is mainly used for format comparison or finding same type of duplicate element. This method is used for analysis of big data by data de-duplication technique.

Lets take some sample of input data and apply data de-duplication techniques on it and proposed method and shows the result. In given Table II it shows details of a person but it is in different format. As shown date of birth format is different in each tuple of same person. City of residence format is also different. It is considered the three different data sets for the same record. So a new approach is being introduced to remove this type of duplication

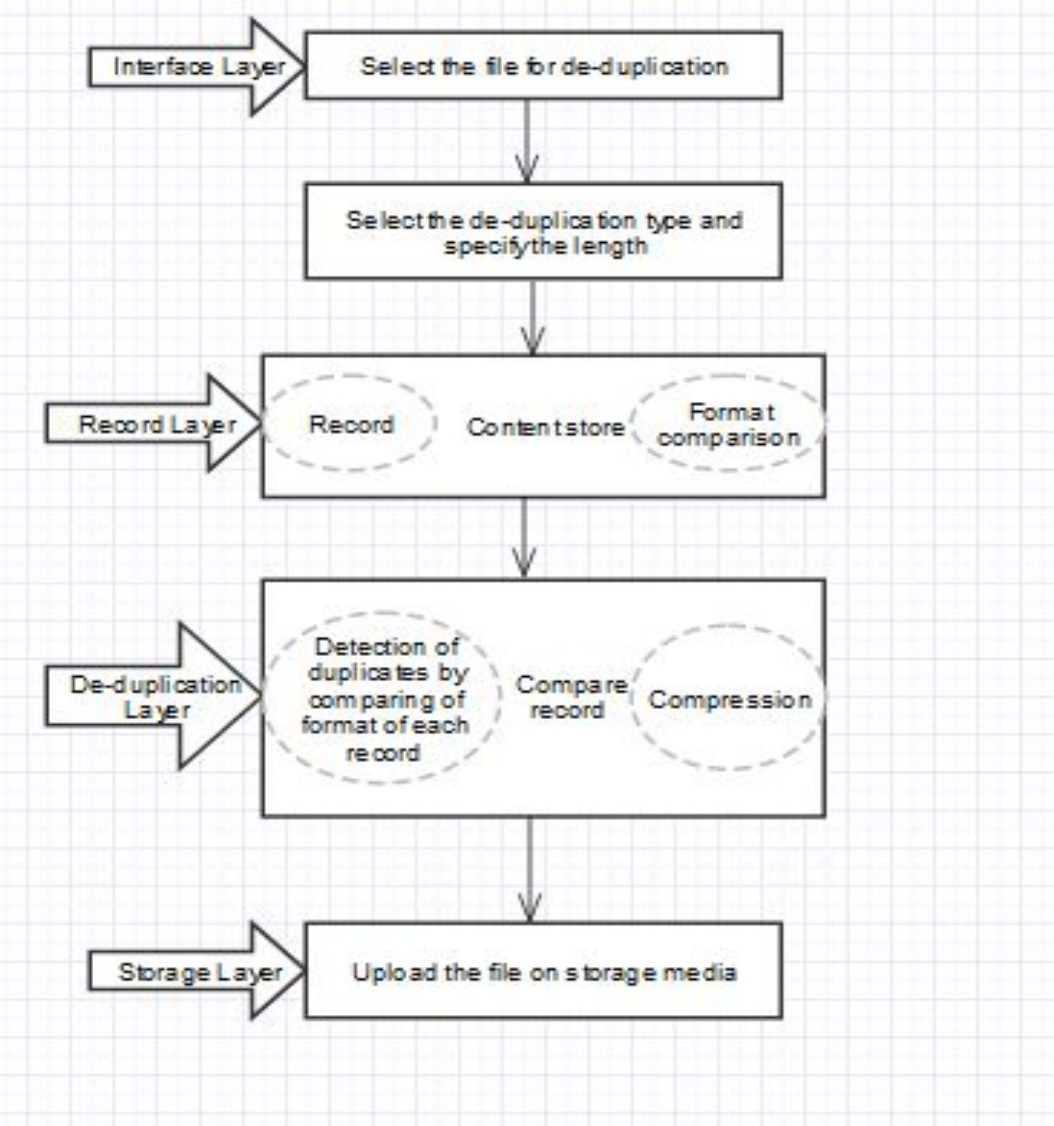


Figure 4.2 Flow of Proposed Technique

4.2 Proposed Technique

Step1:- Initialize the variables T[0…n-1] of n character, P[0…m-1] of m character representing a pattern.

Step2:-Get string from text box and save it into arr.

Step3:-Convert string to lowercase and store in another variable Z.

Step4:- Break string into token and store in array p.

Step5:- Run loop for array p from first element to last element.

Step6:- Match array elements with its own elements in groups of two sequentially. If match is found then increment counter to represent number of matches.

Step7:- Store number of match in ar2 and string match with count in ar1.

Step8:- Run loop for each element in ar1 for index x1. if any element of ar2 > 1 then there is duplicate for that x2 print the string from ar1 that has duplicate

Chapter 5

Implementation and Results

This chapter contains the implementation details of research performed.

5.1 Installation

5.1.1 Karma Data Integration tool :

a. First we downloaded Apache Tomcat 7 and installed it in our system.

b. JDK 1.7 or above is to be installed.

5.2 Implementation and Snapshot

As shown in the below Table 5.1 there are three data set. All the data sets contain same name but in different data format which is an example of data duplication. This type of data duplication technique is not solved or considered by chunk level data de-duplication technique, file level data de-duplication technique and byte level data de-duplication technique, so a new technique is being introduced in which we analyze data on record level. By applying a new technique this type of redundancy is considered.

Table 5.1 Input of sample data

Data set	Name	Date of birth	City of residence
Data set 1	William J. Smith	1/2/73	Berkeley, California
Data set 2	Smith, W. J.	1973.1.2	Berkeley, CA
Data set 3	Bill Smith	Jan 2, 1973	Berkeley, Calif.

As shown in the Table 5.2 we will divide the data set in two sets. We will pick a single identifier that is assumed to be uniquely identifiable, say SSN, and declare that sharing of records with same value identify that it is same person while sharing of records with not same value identify that it is different people.

In the above example, a Deterministic linkage based on SSN would create entities based on P1 and P2; P3 and L1; and P4

Table 5.2 Output of sample data

Data Set	#	*SSN*	*Name*	*DOB*	*Sex*	*Zip*
	1	00956723	Smithh.William	1975/01/02	Male	9471
Set P	2	00956723	Smithh.William	1975/01/02	Male	9473
	3	00005555	Jone.Robert	1943/08/14	Male	9471
	4	23001234	Suee.Marvy	1972/11/19	Female	9419
Set L	1	00005555	Jonees.Bob	1943/08/14		
	2		Smithh.Bill	1975/01/02	Male	9471

While P1, P2, and L2 appear to represent the same entity, because L2 is missing a value for SSN number it would not be included into this type of match.

In Figure 5.2 data is taken in the excel format of crime. These are the data of Chicago crime branch from 2000 to present.

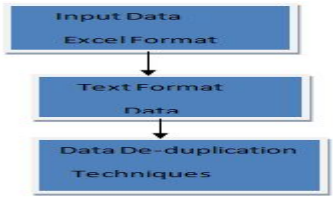
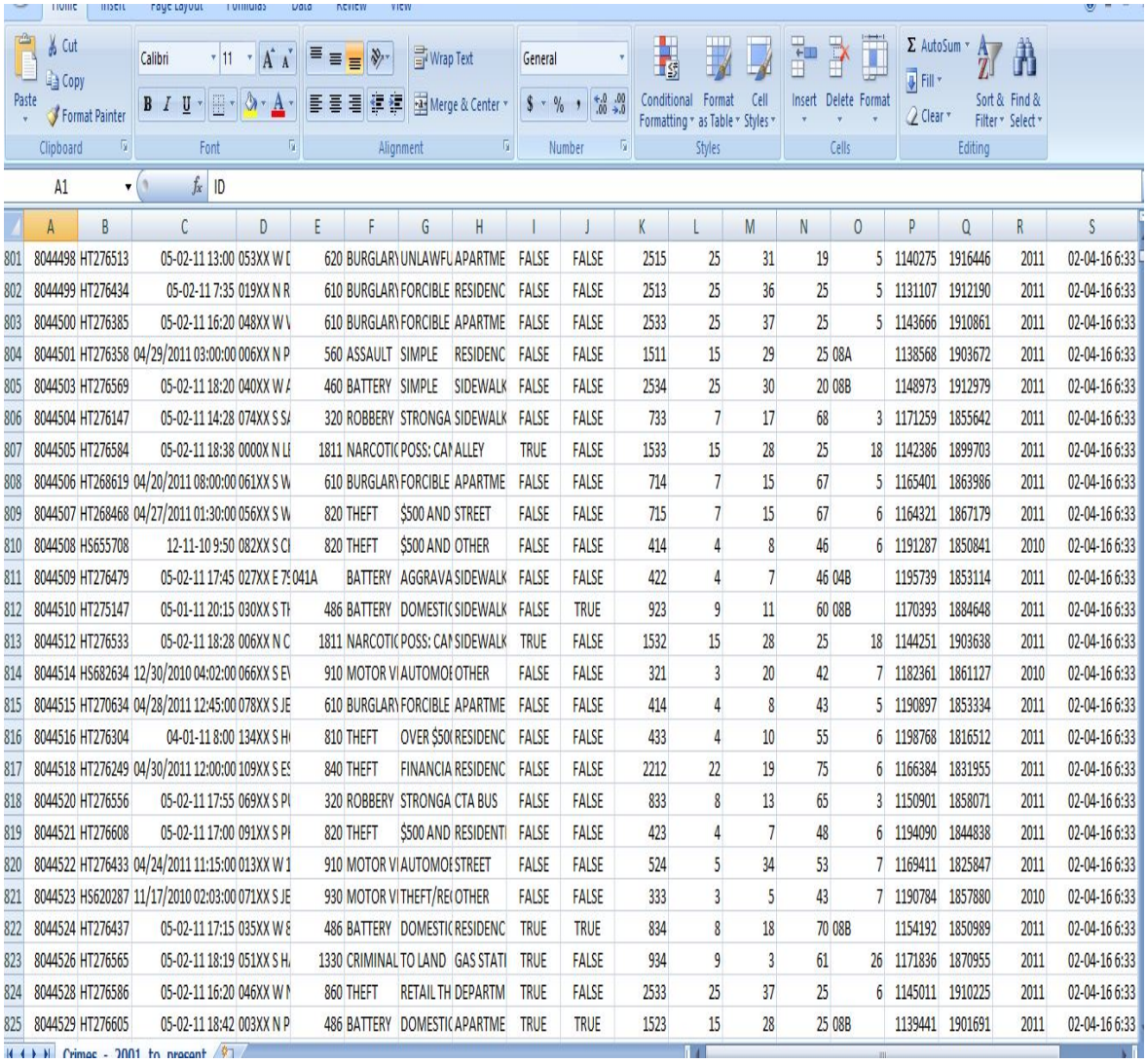


Figure 5.1 Steps for Excel file Data in de-duplication

To apply the data de-duplication technique first we have to convert it in text format. After converting it in text format we have to apply data de-duplication technique.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
801	8044498	HT276513	05-02-11 13:00	053XX W D	620	BURGLAR\	UNLAWFU	APARTME	FALSE	FALSE	2515	25	31	19	5	1140275	1916446	2011	02-04-16 6:33
802	8044499	HT276434	05-02-11 7:35	019XX N R	610	BURGLAR\	FORCIBLE	RESIDENC	FALSE	FALSE	2513	25	36	25	5	1131107	1912190	2011	02-04-16 6:33
803	8044500	HT276385	05-02-11 16:20	048XX W V	610	BURGLAR\	FORCIBLE	APARTME	FALSE	FALSE	2533	25	37	25	5	1143666	1910861	2011	02-04-16 6:33
804	8044501	HT276358	04/29/2011 03:00:00	006XX N P	560	ASSAULT	SIMPLE	RESIDENC	FALSE	FALSE	1511	15	29	25	08A	1138568	1903672	2011	02-04-16 6:33
805	8044503	HT276569	05-02-11 18:20	040XX W A	460	BATTERY	SIMPLE	SIDEWALK	FALSE	FALSE	2534	25	30	20	08B	1148973	1912979	2011	02-04-16 6:33
806	8044504	HT276147	05-02-11 14:28	074XX S SA	320	ROBBERY	STRONGA	SIDEWALK	FALSE	FALSE	733	7	17	68	3	1171259	1855642	2011	02-04-16 6:33
807	8044505	HT276584	05-02-11 18:38	0000X N LI	1811	NARCOTIC	POSS: CAN	ALLEY	TRUE	FALSE	1533	15	28	25	18	1142386	1899703	2011	02-04-16 6:33
808	8044506	HT268619	04/20/2011 08:00:00	061XX S W	610	BURGLAR\	FORCIBLE	APARTME	FALSE	FALSE	714	7	15	67	5	1165401	1863986	2011	02-04-16 6:33
809	8044507	HT268468	04/27/2011 01:30:00	056XX S W	820	THEFT	$500 AND	STREET	FALSE	FALSE	715	7	15	67	6	1164321	1867179	2011	02-04-16 6:33
810	8044508	HS655708	12-11-10 9:50	082XX S CI	820	THEFT	$500 AND	OTHER	FALSE	FALSE	414	4	8	46	6	1191287	1850841	2010	02-04-16 6:33
811	8044509	HT276479	05-02-11 17:45	027XX E 7	041A	BATTERY	AGGRAVA	SIDEWALK	FALSE	FALSE	422	4	7	46	04B	1195739	1853114	2011	02-04-16 6:33
812	8044510	HT275147	05-01-11 20:15	030XX S TI	486	BATTERY	DOMESTIC	SIDEWALK	FALSE	TRUE	923	9	11	60	08B	1170393	1884648	2011	02-04-16 6:33
813	8044512	HT276533	05-02-11 18:28	006XX N C	1811	NARCOTIC	POSS: CAN	SIDEWALK	TRUE	FALSE	1532	15	28	25	18	1144251	1903638	2011	02-04-16 6:33
814	8044514	HS682634	12/30/2010 04:02:00	066XX S EV	910	MOTOR VI	AUTOMOE	OTHER	FALSE	FALSE	321	3	20	42	7	1182361	1861127	2010	02-04-16 6:33
815	8044515	HT270634	04/28/2011 12:45:00	078XX S JE	610	BURGLAR\	FORCIBLE	APARTME	FALSE	FALSE	414	4	8	43	5	1190897	1853334	2011	02-04-16 6:33
816	8044516	HT276304	04-01-11 8:00	134XX S HI	810	THEFT	OVER $50	RESIDENC	FALSE	FALSE	433	4	10	55	6	1198768	1816512	2011	02-04-16 6:33
817	8044518	HT276249	04/30/2011 12:00:00	109XX S ES	840	THEFT	FINANCIA	RESIDENC	FALSE	FALSE	2212	22	19	75	6	1166384	1831955	2011	02-04-16 6:33
818	8044520	HT276556	05-02-11 17:55	069XX S PI	320	ROBBERY	STRONGA	CTA BUS	FALSE	FALSE	833	8	13	65	3	1150901	1858071	2011	02-04-16 6:33
819	8044521	HT276608	05-02-11 17:00	091XX S PI	820	THEFT	$500 AND	RESIDENTI	FALSE	FALSE	423	4	7	48	6	1194090	1844838	2011	02-04-16 6:33
820	8044522	HT276433	04/24/2011 11:15:00	013XX W 1	910	MOTOR VI	AUTOMOE	STREET	FALSE	FALSE	524	5	34	53	7	1169411	1825847	2011	02-04-16 6:33
821	8044523	HS620287	11/17/2010 02:03:00	071XX S JE	930	MOTOR VI	THEFT/RE	OTHER	FALSE	FALSE	333	3	5	43	7	1190784	1857880	2010	02-04-16 6:33
822	8044524	HT276437	05-02-11 17:15	035XX W E	486	BATTERY	DOMESTIC	RESIDENC	TRUE	TRUE	834	8	18	70	08B	1154192	1850989	2011	02-04-16 6:33
823	8044526	HT276565	05-02-11 18:19	051XX S HA	1330	CRIMINAL	TO LAND	GAS STATI	TRUE	FALSE	934	9	3	61	26	1171836	1870955	2011	02-04-16 6:33
824	8044528	HT276586	05-02-11 16:20	046XX W M	860	THEFT	RETAIL TH	DEPARTM	TRUE	FALSE	2533	25	37	25	6	1145011	1910225	2011	02-04-16 6:33
825	8044529	HT276605	05-02-11 18:42	003XX N P	486	BATTERY	DOMESTIC	APARTME	TRUE	TRUE	1523	15	28	25	08B	1139441	1901691	2011	02-04-16 6:33

Figure 5.2 Excel Format Input Data

5.2.1 Comparative Analysis

There are a number of technologies under data de-duplication and data reduction .There are three major data de-duplication techniques and fourth one which is a proposed

technique at record level. Table 5.3 shows the space required for storage of data in cloud. Estimated number of resources required for these different data de-duplication techniques.

File Edit Format View Help

ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	Ward	Communi
8043118	HT271778	04/29/2011 11:45:00 AM	052XX S INDIANA AVE	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT	FALSE	TRUE	232	2	3	
8043119	HT275244	04/16/2011 07:00:00 AM	008XX N LARRABEE ST	820	THEFT	$500 AND UNDER	RESIDENCE PORCH/HALLWAY	TRUE	FALSE	1823	18	27	
8043120	HT275243	04/24/2011 09:45:00 PM	008XX N LARRABEE ST	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE PORCH/HALLWAY	TRUE	FALSE	1823	18		
8043121	HT275230	05-01-11 21:28	032XX N LEAVITT ST	1330	CRIMINAL TRESPASS	TO LAND	RESIDENCE	FALSE	FALSE	1913	19	32	5
8043122	HT275281	05-01-11 15:00	082XX S ARTESIAN AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET	FALSE	FALSE	835	8	18	70
8043123	HT275291	05-01-11 22:20	074XX S SOUTH SHORE DR	2825	OTHER OFFENSE	HARASSMENT BY TELEPHONE	RESIDENCE	FALSE	TRUE	334	3	7	
8043124	HT275268	04/20/2011 10:00:00 PM	024XX S MILLARD AVE	890	THEFT	FROM BUILDING	APARTMENT	FALSE	FALSE	1013	10	22	30
8043125	HT275242	05-01-11 21:50	017XX W JULIAN ST	337	ROBBERY	ATTEMPT: ARMED-OTHER DANG WEAP	SIDEWALK	FALSE	FALSE	1433	14	1	
8043126	HT275280	04/25/2011 05:00:00 PM	050XX W PARKER AVE	890	THEFT	FROM BUILDING	RESIDENCE	FALSE	FALSE	2521	25	31	19
8043127	HT275311	05-01-11 9:30	012XX S MICHIGAN AVE	2825	OTHER OFFENSE	HARASSMENT BY TELEPHONE	APARTMENT	FALSE	FALSE	132	1	2	
8043128	HT275217	05-01-11 21:29	045XX S DREXEL BLVD	2820	OTHER OFFENSE	TELEPHONE THREAT	RESIDENCE	FALSE	FALSE	2123	2	4	
8043129	HT275323	04/15/2011 10:00:00 AM	018XX W 46TH ST	2820	OTHER OFFENSE	TELEPHONE THREAT	RESIDENCE	FALSE	FALSE	914	9	20	
8043131	HT275285	05-01-11 20:00	008XX N MICHIGAN AVE	810	THEFT	OVER $500	PARKING LOT/GARAGE(NON.RESID.)	FALSE	FALSE	1833	18	42	
8043132	HT275233	05-01-11 19:00	021XX E 93RD ST	610	BURGLARY	FORCIBLE ENTRY	OTHER	FALSE	FALSE	413	4	7	48
8043133	HT275140	05-01-11 19:30	052XX N DAMEN AVE	1152	DECEPTIVE PRACTICE	ILLEGAL USE CASH CARD	STREET	FALSE	FALSE	2012	20	40	
8043134	HT273730	04/29/2011 07:45:00 AM	027XX W EVERGREEN AVE	560	ASSAULT	SIMPLE	STREET	FALSE	TRUE	1423	14	26	24
8043135	HT275279	05-01-11 22:18	116XX S MICHIGAN AVE	430	BATTERY	AGGRAVATED: OTHER DANG WEAPON	SIDEWALK	FALSE	FALSE	532	5	9	
8043136	HT273976	05-01-11 0:20	014XX N FAIRFIELD AVE	460	BATTERY	SIMPLE	STREET	FALSE	FALSE	1423	14	26	24
8043137	HT275282	05-01-11 0:03	028XX W NORTH AVE	620	BURGLARY	UNLAWFUL ENTRY	RESIDENCE PORCH/HALLWAY	FALSE	FALSE	1421	14	1	
8043138	HT275064	05-01-11 19:00	104XX S UNION AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	FALSE	TRUE	2233	22	34	49
8043139	HT275247	05-01-11 21:53	057XX S HOYNE AVE	143A	WEAPONS VIOLATION	UNLAWFUL POSS OF HANDGUN	STREET	TRUE	FALSE	715	7		
8043140	HT275333	05-01-11 22:50	022XX N NATCHEZ AVE	820	THEFT	$500 AND UNDER	SIDEWALK	FALSE	FALSE	2512	25	36	19
8043141	HT275328	05-01-11 23:07	103XX S AVENUE N	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE	FALSE	TRUE	432	4	10	52
8043142	HT275258	05-01-11 22:03	021XX N HAMLIN AVE	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT	FALSE	FALSE	2525	25	26	22
8043143	HT272332	04/29/2011 08:30:00 PM	105XX S WENTWORTH AVE	560	ASSAULT	SIMPLE	SIDEWALK	FALSE	FALSE	512	5	34	49
8043144	HT275260	05-01-11 21:55	030XX N MONTICELLO AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	FALSE	TRUE	2523	25	35	21
8043145	HT275253	05-01-11 21:35	025XX N MANGO AVE	031A	ROBBERY	ARMED: HANDGUN	SIDEWALK	FALSE	FALSE	2515	25	30	19
8043146	HT275143	04/30/2011 10:00:00 PM	039XX N CLARK ST	265	CRIM SEXUAL ASSAULT	AGGRAVATED: OTHER	CHA APARTMENT	FALSE	FALSE	1923			
8043147	HT274956	05-01-11 17:28	057XX S LOOMIS BLVD	496	BATTERY	AGGRAVATED DOMESTIC BATTERY: KNIFE/CUTTING INST	RESIDENCE	TRUE	TRUE	713			
8043148	HT275297	05-01-11 22:35	001XX W ILLINOIS ST	460	BATTERY	SIMPLE	RESTAURANT	FALSE	FALSE	1831	18	42	8
8043149	HT273952	04/30/2011 02:00:00 PM	072XX S SOUTH SHORE DR	890	THEFT	FROM BUILDING	APARTMENT	FALSE	FALSE	334	3	7	43
8043150	HT272439	04/29/2011 10:00:00 PM	010XX W 63RD ST	820	THEFT	$500 AND UNDER	SIDEWALK	FALSE	FALSE	712	7	16	68
8043151	HT275186	05-01-11 20:14	101XX S PRAIRIE AVE	143B	WEAPONS VIOLATION	UNLAWFUL POSS OTHER FIREARM	RESIDENCE	TRUE	FALSE	511			
8043152	HT275319	05-01-11 22:30	019XX E 74TH ST	031A	ROBBERY	ARMED: HANDGUN	SIDEWALK	FALSE	FALSE	333	3	5	43
8043153	HT275264	05-01-11 22:06	119XX S NORMAL AVE	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT	FALSE	FALSE	522	5	34	53

Figure 5.3 Text Format Of Input Data

The data from the Figure 5.2 is to be converted into Text format as shown in Figure 5.3.

Table 5.3 Space saving and Resources used

Technology	***Typical space saving***	***Resource footprint***
File –Level	10%	Low
Block Level	20%	High

Byte Level	28%	High
Proposed method	35%	Medium

As shown in Table 5.3 File level data de-duplication technique saves only 10% i.e. removal of redundant data features is less in File level data de-duplication technique. But for this type of data de-duplication technique less resources are required. Block level data de-duplication has 20% space saving features but it requires more number of resources as compared to File level data de-duplication technique. Byte level data de-duplication has 35% features to remove redundant data but the method proposed in this paper has 35% features for removal of redundant data and it also requires medium level of resources.

5.2.1.1 De-Duplication Efficiency

The cloud test has been used for measuring the data de-duplication technique. The saving of space ratio is called as de-duplication throughput. The saving of time ratio is called as de-duplication efficiency.

Figure 5.4 De-duplication analysis by Edit Distance algorithm (Existing method)

The Figure shows the output of the edit distance algorithm. Here it ask the following distance

i. **Insertion Cost**

It is the cost required to insert a word in the file or match a word in tree is called insertion cost.

ii. **Deletion cost**

It is the cost required to delete a word from the file is that word is not there in that file to check similarity between two file.

iii. **Updation Cost-**: It is the cost required to update a word in the given file to convert it similar word is called updation cost.

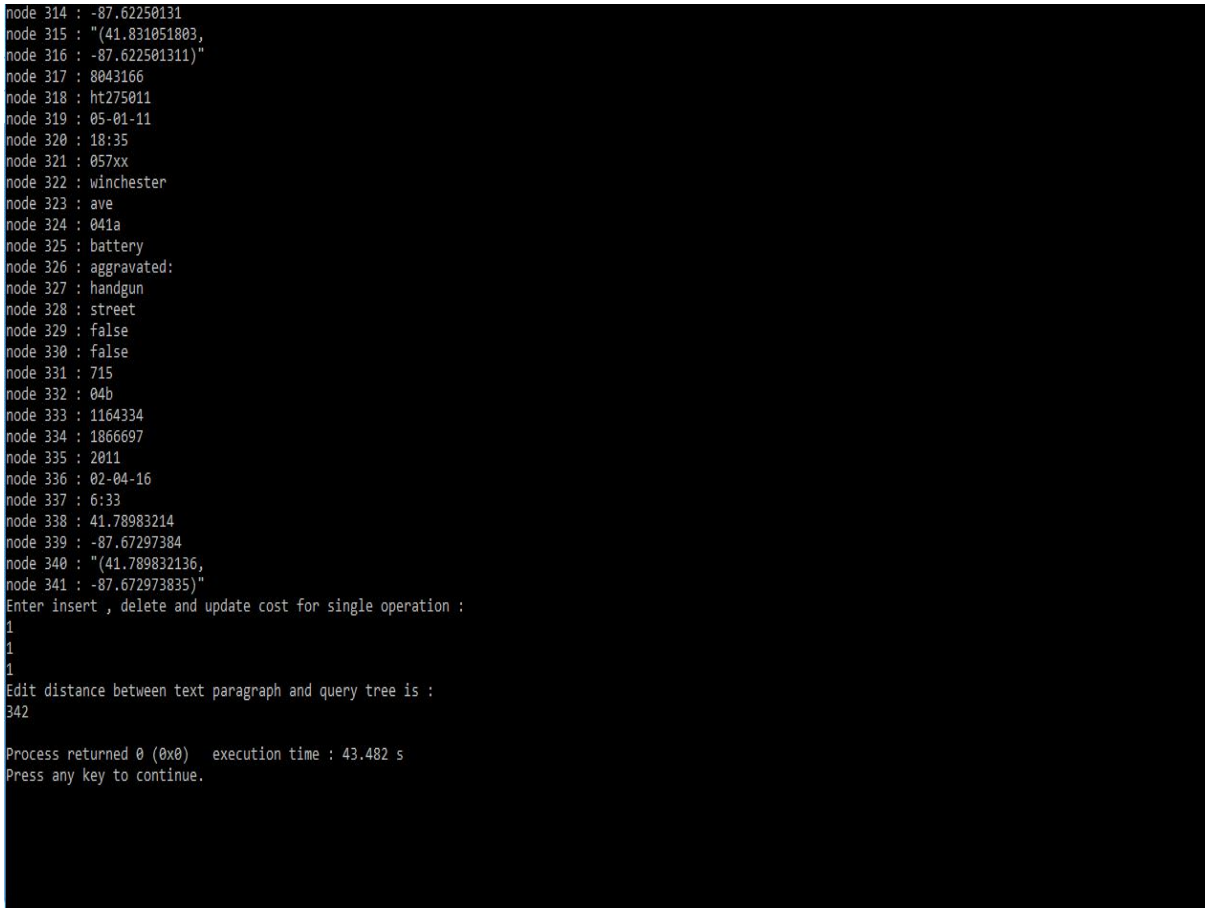


Figure 5.5 De-duplication analysis with cost by Edit Distance algorithm

As shown in Figure 5.5 we give the different cost for checking the similarity in two different file as

Insert cost =1

Delete cost=1

Update cost=1

So edit distance between text paragraph and tree is 342 according to file size.

More the edit distance cost more the percentage of similarity between two file i.e one file is more redundant to other file.

If edit distance cost is 0 means a file is not similar to other file i.e redundant data is removed.

```
description location:2
location description:2
coordinate y:2
y coordinate:2
ave 1310:3
1310 criminal:5
criminal damage:6
damage to:6
to property:5
property apartment:3
apartment false:8
false true:10
2011 02-04-16:35
02-04-16 6:33:35
008xx n:3
n larrabee:2
larrabee st:2
st 820:2
820 theft:3
theft $500:3
$500 and:3
and under:3
residence porch/hallway:3
porch/hallway true:2
true false:10
false 1823:2
```

Figure 5.6 De-duplication analysis at record level (proposed method)

In the Figure 5.6 numeric value on the right side shows that this word is repeated that

number of time in that file .that particular word if checked at each level

.

The Figure.5.7 shows the variation of data de-duplication efficiency with the file size. As in this Figure.5.7 shown that as the file size increase the time taken by the proposed method to store the data is less as compared to Chunk level ,File level and Byte level data de-duplication technique

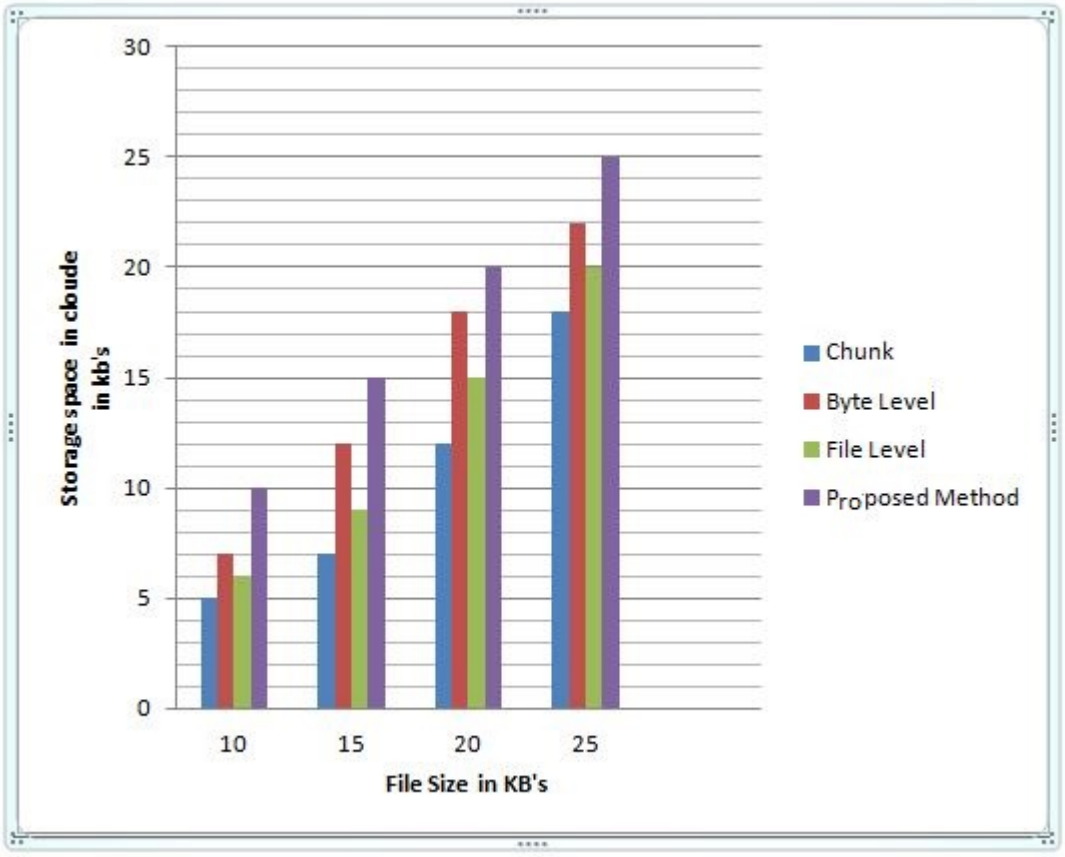


Figure 5.7 Storage space v/s file size.

5.2.1.2 De-Duplication Throughput

The Figure.5.8 shows the variation of data de-duplication throughput with the file size. As shown in the Figure.5.8 that as the file size increase throughput increases for same method. Data stream of numerous users residing to the same group takes less time with file level de-duplication compared to the chunk level supplement.

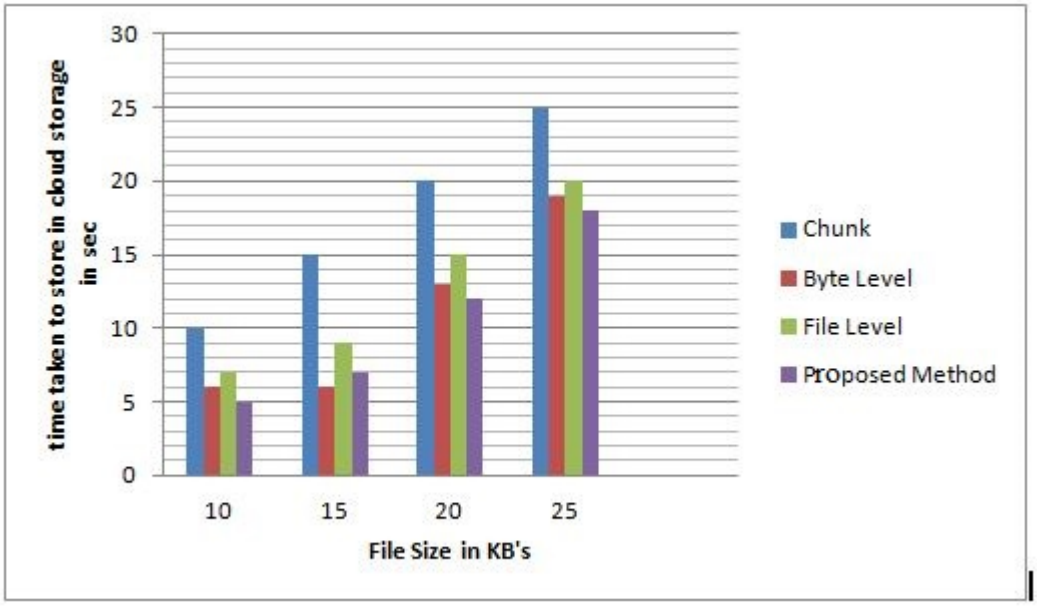


Figure 5.8 Time for storage v/s File size

Space Reduction ratio=Bytes in/Bytes out

These space reduction may be view as data capacity of the system divided by usable storage capacity. For example, if 100 GB of data consumes 10GB of storage capacity then ratio is 10:1.

Space reduction %=(1-(1/space reduction ratio))

Table 5.4 shows success rate of different data de-duplication techniques. Success rate is defined as how much redundant data is to be removed by data de-duplication techniques.

Table 5.4. Methods with Success rate.

S.No	*Method*	*Success rate*
1	Chunk Level	80 %
2	File Level	90 %
3	Byte Level	95 %

4	Proposed Method	97%

As shown in above Table 5.4 chunk level data de-duplication technique has success rate of 80 % i.e 80% of the redundant data is removed by this data de-duplication technique, File level data de-duplication technique has success rate of 90 %, Byte level data de-duplication technique has 95 % success rate and proposed method has 97 % success rate because in this format comparison is also consider to remove redundant data.

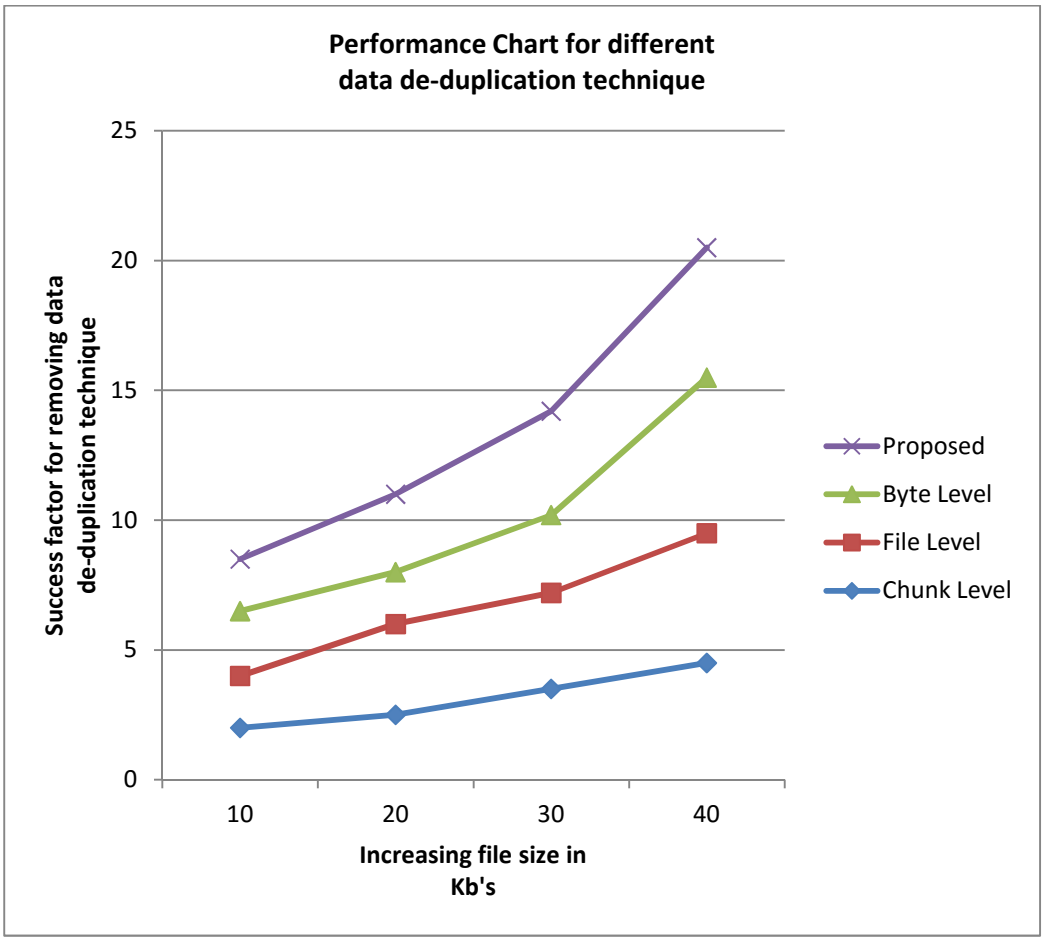


Figure 5.9 Success Rate of data de-duplication techniques.

As shown in above Figure.5.9 a graph is shown between success rate of different data de-duplication techniques and file size i.e increasing in kilo bytes. As shown in Figure.5.9 as

the size of file is increasing the data removal rate is also increasing for the different data de-duplication techniques. Here, proposed technique shows the maximum redundant data removal rate as compared to other data de-duplication techniques.

Chapter 6
Conclusion and Future Work

6.1 Conclusion

In this research, have been discussed the different data de-duplication techniques for analysis of big data . When different format of data combined in one format then there is chance of duplication of data. Where chunk level, file level and byte level data de-duplication technique are not able to find out the redundant data at record level. So a technique is proposed in this thesis for removal of similar data at record level because of format.

6.2 Thesis Contribution

Contribution made by the proposed thesis work is summarized

- A technique for analysis of big data is proposed.
- Throughput of the process is increased for analysis of data.
- Storage space for data is reduced that make fast execution of query.
- Data de-duplication is applied at record level that coordinates in linkage of data at record level.

6.3 Future Work

In future, we would like to explore the de-duplication on images, videos and deploy the same on cloud based storage with proposed storage policies. For analysis of big data we can further apply visualization and filtering technique for more analyzing the data.

References

[1] C. Knoblock and P. Szekely, "semantics for big data integration and analysis", the Proceedings of the AAAI Fall Symposium on Semantics for Big Data, 2013

[2] L. AO, J. SHU and M. LI, "Data Deduplication Techniques", Journal of Software, vol. 21, no. 5, pp. 916-929, 2010.

[3] H. Jung, S. Park, J. Lee and Y. Ko, "efficient data duplication system considering file modification pattern", International Journal of Security and Its Applications, vol. 6, no. 2, pp. 421--426, 2012.

[4] S. Deepu, "Performance Comparison of Deduplication techniques for storage in Cloud computing Environment", Asian Journal of Computer Science \& Information Technology, vol. 4, no. 5, 2014.

[5] Z. Al-sagar, A. Sameen and M. Saleh, "Optimizing the Cloud Storage by Data Deduplication: A Study", 2015

[6] P. Gapat, M. Pise, A. Khiste and S. Khillare, "A Survey Paper on Removal of Data Duplication in a Hybrid Cloud", 2016

[7] A. Banu and C. Chandrasekar, "A survey on deduplication methods", *International Journal of Computer Trends and Technology*, vol. 3, no. 3, pp. 364-368, 2012.

[8] Q. He, X. Zhang and Z. Li, "Data deduplication techniques", *Future Information Technology and Management Engineering (FITME), 2010 International Conference on,IEEE*, vol. 1, pp. 430-433, 2010.

[9] R. Tuchinda, C. Knoblock and P. Szekely, "Building data integration queries by demonstration", *Proceedings of the 12th international conference on Intelligent user interfaces,ACM*, pp. 170-179, 2007

[10] Y. Lee, A. Rosenthal, A. Doan and M. Sayyadian, "eTuner: tuning schema matching software using synthetic scenarios", *The VLDB Journal—The International Journal on Very Large Data Bases,Springer-Verlag New York, Inc*, vol. 16, no. 1, pp. 97-122, 2007

[11] W. Bolosky, J. Douceur, D. Goebel and S. Corbin, "Single instance storage in Windows 2000", *Proceedings of the 4th USENIX Windows Systems Symposium,Seattle, WA*, pp. 13-24, 2000

[12] Z. Ives, D. Weld, A. Levy, M. Friedman and D. Florescu, "An adaptive query execution system for data integration", *ACM SIGMOD Record,ACM*, vol. 28, no. 2, pp. 299-310, 1999.

[13] "Different Data de-duplication techniquies"http://bbs.chinabyte.comlthread-393434-1-I.htrnl

[14] "Different Data de-duplication techniquies"http://storage.chinaunix.netlstor/c/

[15] R. Tuchinda, Y. Gil, S. Thakkar and E. Deelman, "Artemis: Integrating scientific data on the grid", *AAAI*, pp. 892-899, 2004.

[16] R. Tuchinda and C. Knoblock, "Agent wizard: building information agents by answering questions", *Proceedings of the 9th international conference on Intelligent user interfaces,ACM*, pp. 340-342, 2004

[17] D. Frey, K. Kloudas and A. Kermarrec, "Probabilistic deduplication for cluster-based storage systems", *Proceedings of the Third ACM Symposium on Cloud Computing,ACM*, p. 17, 2012.

[18] Y. Fu, N. Xiao and H. Jiang, "A scalable inline cluster deduplication framework for big data protection", *ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing,Springer*, pp. 354-373, 2012.

[19] F. Yinjin, X. Lei, L. Fang, T. Lei, X. Nong and J. Hong, "Application-Aware Source Deduplication for Cloud Backup Services of Personal Storage,", *," IEEE Transactions on Parallel and Distributed Systems,IEEE*, 2013.

[20] G. Wang, L. Liu, X. Xie and Y. Zhao, "Research on a clustering data de-duplication mechanism based on Bloom Filter", *Multimedia Technology (ICMT), 2010 International Conference on ,IEEE*, pp. 1-5, 2010

[21] J. Ma, X. Liu, G. Wang and B. Zhao, "Adaptive pipeline for deduplication", *012 IEEE 28th Symposium on Mass Storage Systems and*

Technologies (MSST),IEEE, pp. 1-6, 2012

[22] San Kong, Y. Ko, W. Lee and M. Kim, "Two-Level Metadata Management for Data Deduplication System", *IST*, vol. 23, pp. 299-303, 2013.

[23] "Characteristics of big data", https://www.google.co.in/search?q=3+volume+of +big+data&biw=1366&bih=659&source=lnms&tbm=isch&sa=X&sqi=2&ved=0ahUK0ahUKEwj7n4GV0ZLNAhVGMY8KHQ8iDGYQ_AUIBigB#imgrc=xHSjpBwX_tiEHM%3A.

[24] "Introduction of growth of big data" www.planet data.eu/sites/default/files presentations/Big_Data_ Tutorial_part4

[25] "MapReduce",Framework"http://tomato.biol.trinity.edu/blog/wpcontent/uploads/2011/01/map_reduce_schematic1.png

[26] Y. Lokeshwari, C. Babu and B. Prabavathy, "Optimized cloud storage with high throughput deduplication approach", *Proceedings of the International Conference on Emerging Technology Trends (ICETT),Citeseer*, 2016.

[27] C. Constantinescu, D. Chambliss and J. Glider, "Mixing deduplication and compression on active data sets", *2011 Data Compression Conference,IEEE*, pp. 393-402, 2011

[28] K. Zhang and D. Shasha, "Simple Fast Algorithms for the Editing Distance between Trees and Related Problems", *SIAM J. Comput.*, vol. 18, no. 6, pp. 1245-1262, 1989.

[29] Malhotra and J. Bakal, "A survey and comparative study of data deduplication techniques", *Pervasive Computing (ICPC), 2015 International Conference on*, pp. 1-5, 2015.

List of publication

S. Garg and A. Bala, "Semantic analysis of big data by data de-duplication techniques," IEEE International Conference on Inventive Computation Technologies(ICICT 2016),Date-26-27 August-16 . [Communicated]

Video Link

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