1. INTRODUCTION

Most of the organizations have stored huge amounts of data over long periods of operation. Data mining is able to extract very valuable knowledge from this data. Data mining is defined as the process of analyzing data from different sources into useful information. A search engine is a system that is designed to search for information on the computer hard disks. The information may be a mix of web pages, images, audio, documents, videos and other types of files. Clustering is defined as it is a process of dividing datasets into smaller groups such that the members of each small group are as similar as one another. Clustering data is a fundamental problem in a variety of areas of storage data and related fields. Clustering makes data easier to process, mine for information and more human readable. The types of clustering algorithms are Hierarchical and Partitioned. Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters.

Types of hierarchical clustering are- Agglomerative is a "bottom up" approach in which each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive is a "top down" approach: in which all observations start in one cluster, and splits are performed recursively as one move down the hierarchy. Partitioned clustering algorithm is a method of cluster analysis which seeks to build a partition of clusters.

Types of partitioned clustering are k-means and self-organizing map.Map-Reduce function is used for processing huge datasets. Map-Reduce can take advantage of locality of data, processing it on or near the storage assets in order to reduce the distance over which it must be transmitted.

Map-Reduce function has two steps-Map step and Reduce step. The Map function takes a series of key/value pairs, processes each, and generates zero or more output key/value pairs. The input and output types of the map can be different from each other. If the application is doing a word count, the map function would break the line into words and output a key/value pair for each word. Each output pair would contain

the word as the key and the number of instances of that word in the line as the value. The Reduce function takes the input values, sums them and generates a single output of the word and the final sum.

1.1 Motivation

The Google is having their web crawler which searches for all documents in all the servers connected to Internet, we are going to implement crawler which will be searching for all the types of files in our personal computers.

1.2 Scope

In regular day today life user can use this software for searching data and retrieving data.

1.3 Overview

Clustering is defined as it is a process of dividing datasets into smaller groups such that the members of each small group are as similar as one another. Clustering data is a fundamental problem in a variety of areas of storage data and related fields. Clustering makes data easier to process, mine for information and more human readable.

Map-Reduce function is used for processing huge datasets. Map-Reduce function has two steps-Map step and Reduce step .The Map function takes a series of key/value pairs, processes each, and generates zero or more output key/value pairs. The input and output types of the map can be different from each other. The Reduce function takes the input values, sums them and generates a single output of the word and the final sum

2. LITERATURE SURVEY

Puppala Priyanka[2014][1]

Author has proposed the issue in the clustering is that - how to determine the similarity between two objects, and have introduced a clustering algorithm called “EMaRC” for similarity measurement.

Kiran jyoti ,Dr. Satyaveer Singh et.al [2011][2]

Author has proposed the different conventional and fuzzy based clustering techniques for fault detection and isolation in process plant monitoring. The author implements k-means algorithm and fuzzy c means algorithm to cluster the relevant data.

S.Ayyasamy , S.N. Sivanandam et.al [2010][3]

Author has purposed, a cluster based replication architecture for load-balancing in peer-to-peer content distribution systems. In the intelligent replica placement technique, peers are grouped into strong and weak clusters based on their weight vector which comprises available capacity, CPU speed, access latency and memory size. In order to achieve complete load balancing across the system, an intra- cluster and inter-cluster load balancing algorithms are proposed.

Wang Kay Ngai ,Ben Kao ,Chun Kit et.al [2006][4]

Author has proposed that one methods to cluster uncertain objects is to apply the UK mean algorithm which is based on the traditional K- means algorithm. In UK-means, and object is assigned to the cluster whose representative has the smallest excepted distance to the object

2.1 Limitation of Existing System

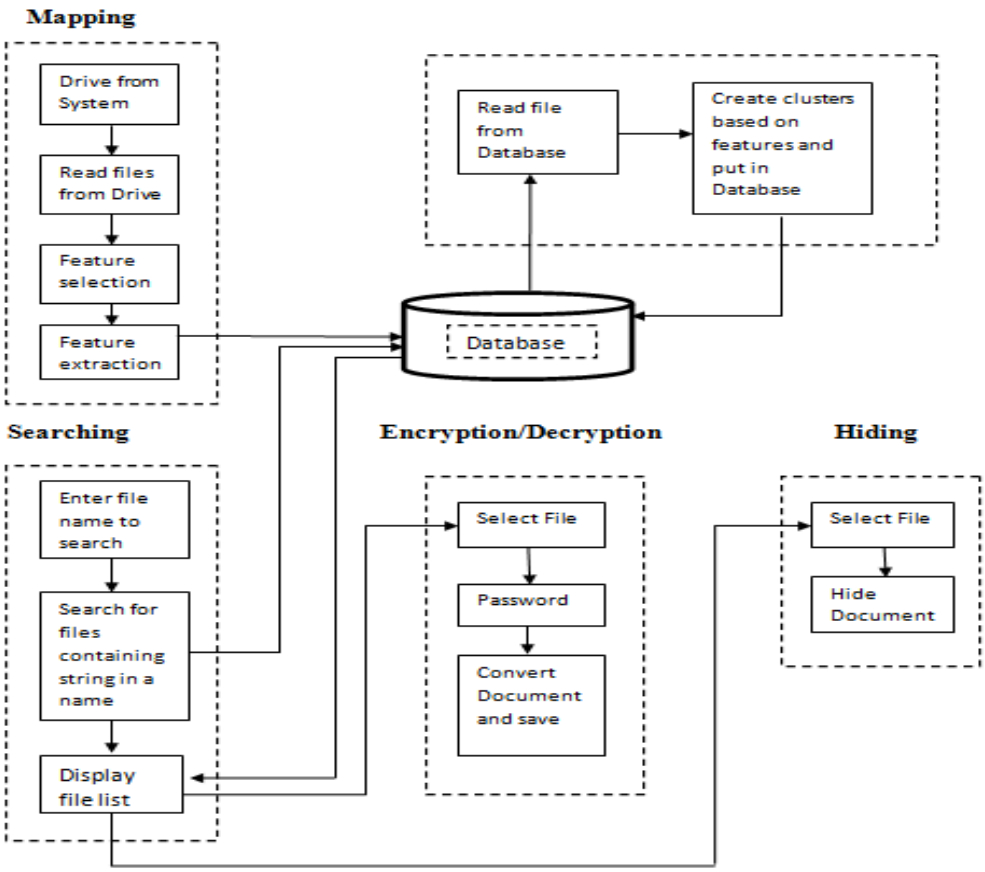
- Searching required more time..
- Word count based searching is not possible.

3. PROPOSED SYSTEM

3.1 The problem domain

Problem with time required to retrieve the data is more and its does not provide security but by using map-reduce algorithm improves the speed of performance and provides security.

3.2 System overview



. .

Fig 1. System Architecture

Fig.1 illustrates, the system consist of three modules such as mapping, reducing and searching. The method we present here is simple, less complex and efficient and can meet the high speed requirements in practical applications. The overall goal of this project is to extract information from data sets and transform it into an understandable

structure for future use. In this project the input may be any kind of data such as our computer hard disk, external hard disk, mobile data, etc. From this input the feature selection process is used to identify the most effective subset of the features to use in clustering. Then feature extraction function is used to produce new most important features. Then a clustering algorithm is used to find out whether two objects are similar to each other. This clustering algorithm will do two functions mapping function and reducing function. The main database is connected to these modules like mapping, clustering, searching, encryption/decryption and hiding of the data.

4. SYSTEM REQUIREMENT SPECIFICATION

4.1 Hardware requirements

- Memory: Min. 1 GB RAM
- Processor: Min. Pentium IV

4.2 Software requirements

- Window7/Window8
- .NET Framework 3.5
- Front End : C#.Net
- Back End: SQL Server

4.3 System requirements

4.3.1 Functional requirements

- Operating System: Windows
- CPU: Pentium 4
- Memory: 1 GB RAM
- Environment: Microsoft Visual Studio 2008.

4.3.2 Non-functional requirements

- **User Friendly:**

  This application can handle by any user easily.

- **Security:**

  This application provides security to data.

- **Reliability:**

  This application can be used in any environment and at any given amount of time.

4.4 Project Scheduling

TABLE.1 Project Scheduling

Sr.No.	Month	Task
1.	July	Overview Design of the Project
2.	August	Initiated the First Module – Mapping.
3.	September	Completion of Mapping Module.
4.	October	Initiated Implementation of Clustering Module(K-Mean) & Searching Module.
5.	November	Completed development of Clustering Module and Searching Module.
6.	December	Initiated implementation of Security Module (Encryption & Decryption of Files, Hiding of Files).
7.	January	Completed development of Security Module.
8.	February	Development of additional module of clustering based on count of word in file.
9.	March	Modification in Searching Module based on Content based clustering algorithm.
10.	April	Formatting of the report and final testing of the project.

4.5 Development Process

For the development process of the system we have implemented the software according to the Waterfall model. The waterfall model is a sequential design process, used in software development processes, in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of Conception, Initiation, Analysis, Design, Construction, Testing, Production/Implementation and Maintenance. In the requirement analysis phase we decided the system requirements i.e. the operating system, processor, memory required by the software, software's required for the implementations of the modules, environment on which our application could be executed.

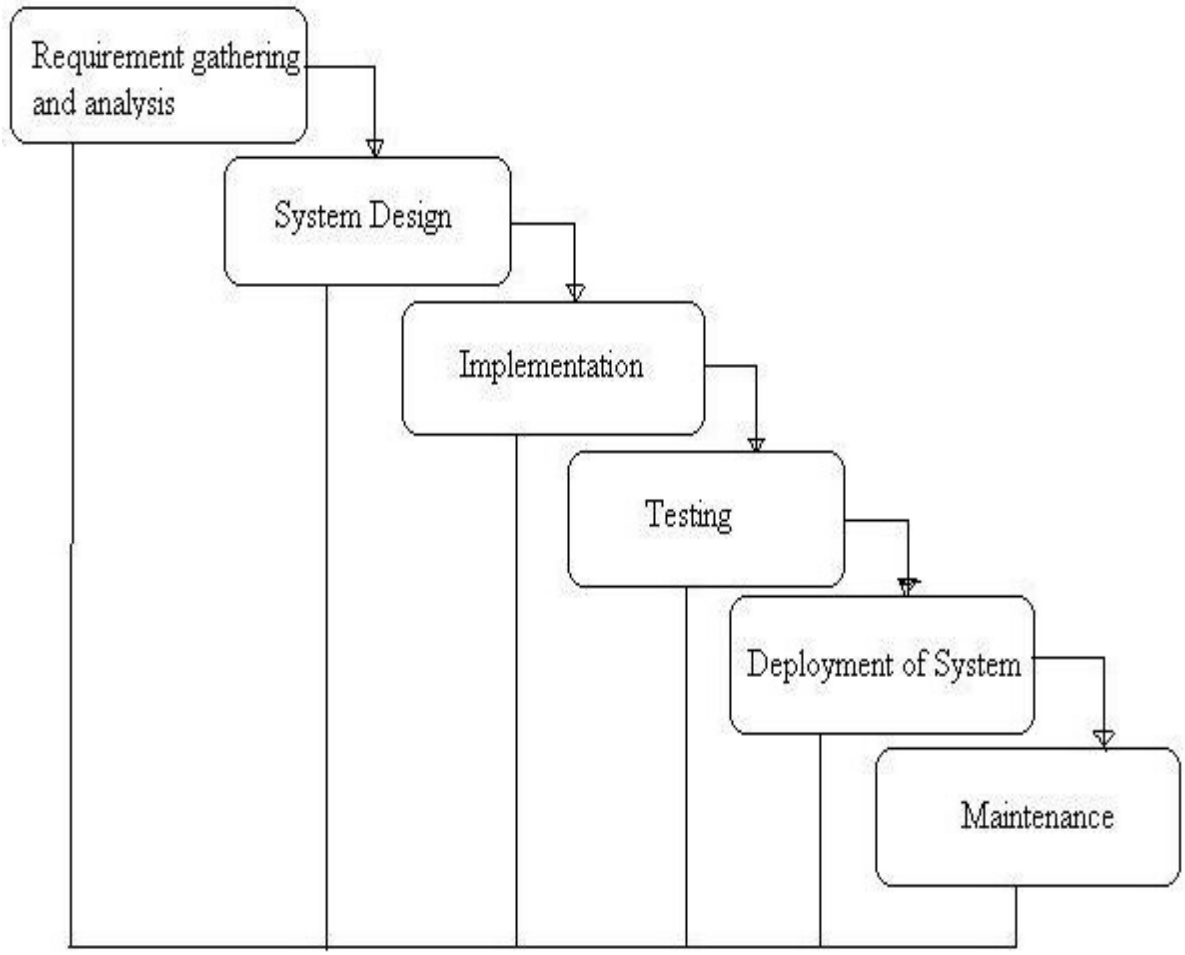


Fig 2: Waterfall Model

Project planning involved a well planned flow chart of work to be done in the restricted time period, where which part of the modules to implemented in which month that was decided so as to complete our project in a swift way.System design phase is in which we designed the overall architecture of the system. In this phase we worked on the system architecture which helped us to decide the various modules of our applications.

Detailed design was the detailed workflow of various modules. It included the various modules of our system like the mapping, reducing, searching. In this phase each module when given an input should display the correct output was required while designing the detailed structure of the system. Coding of the application is done by using the C# programming language which is part .net framework 3.5.

Testing ensured that the output of a phase is consistent with its input i.e. the output of the previous phase and that the output of the phase is consistent with the overall requirements of the system. Deployment of system was done by using the azure cloud of Microsoft which provides a wide range of facilities as well as services which the software needed.

Operations and maintenance include the proper check of the software whether it is working in required manner. It is the future scope of the developer to keep in mind the basic check of the software and its functionality. If any bugs or faults occur in the software, it should be recovered.

5. SYSTEM DESIGN

5.1 System Description

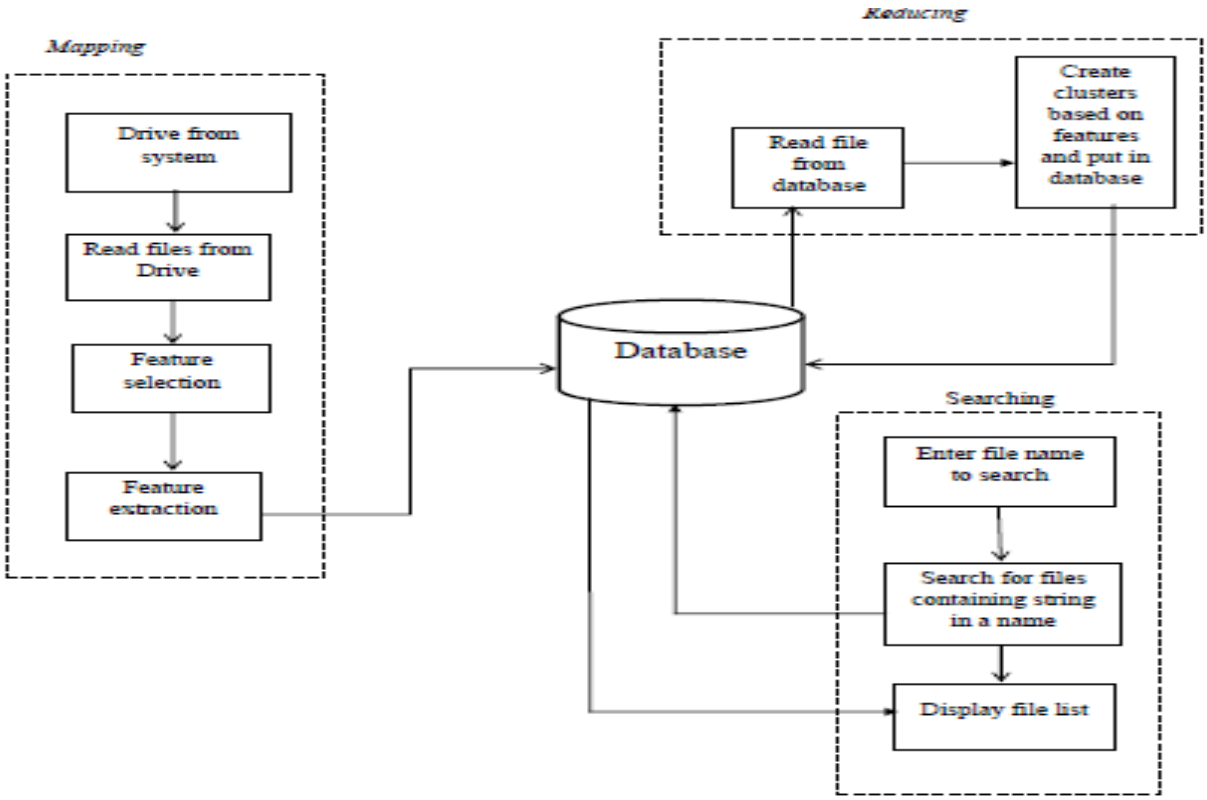


Fig 2. System Architecture

Fig.2 illustrates, the system consist of three modules such as mapping, reducing and searching. The method we present here is simple, less complex and efficient and can meet the high speed requirements in practical applications. The overall goal of this project is to extract information from data sets and transform it into an understandable structure for future use. In this project the input may be any kind of data such as our computer hard disk, external hard disk, mobile data, etc. From this input the feature selection process is used to identify the most effective subset of the features to use in clustering. Then feature extraction function is used to produce new most important features. Then a clustering algorithm is used to find out whether two objects are similar to each other. This clustering algorithm will do two functions mapping function and reducing function.

5.2 Use case Diagram

In the following use case diagram, user and system are the actors. Drives, mapping, clustered group, clustering and display data are the use cases. In this project, user can select a drive for searching a file/folder. The system maps the drives and fuzzy clustering based on the extension of file is done by the system. As a result the file/folder searched by the user is displayed.

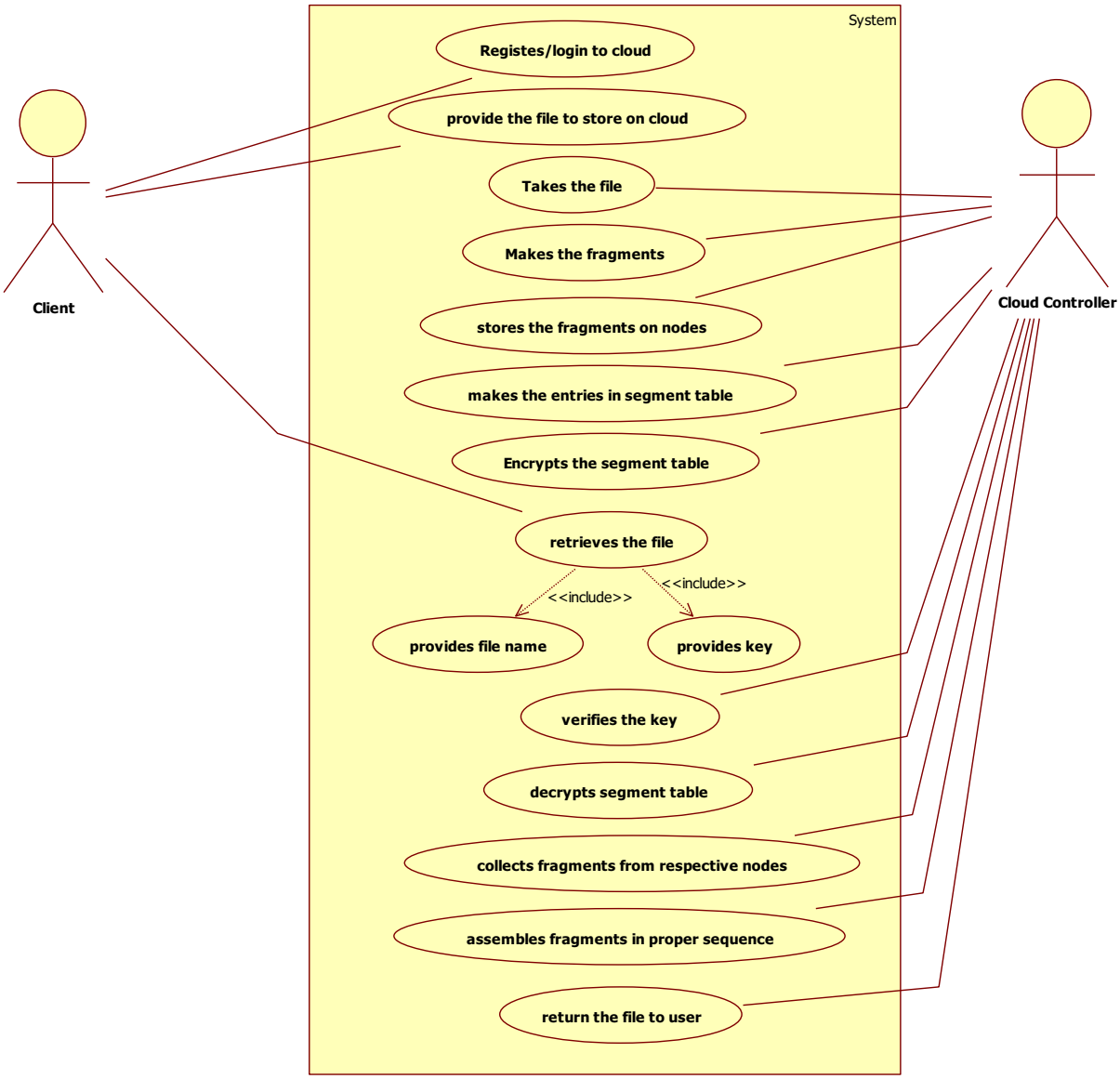


Figure : Use case Diagram

5.3 Sequence Diagram

In the following sequence diagram, when user selects a drive or drives to map system map all selected drives for files/folders. As user enter file name to search, the system read file extension and clusters it as per the corresponding file extension. As a result all possible result i.e files/folders are displayed to the user.

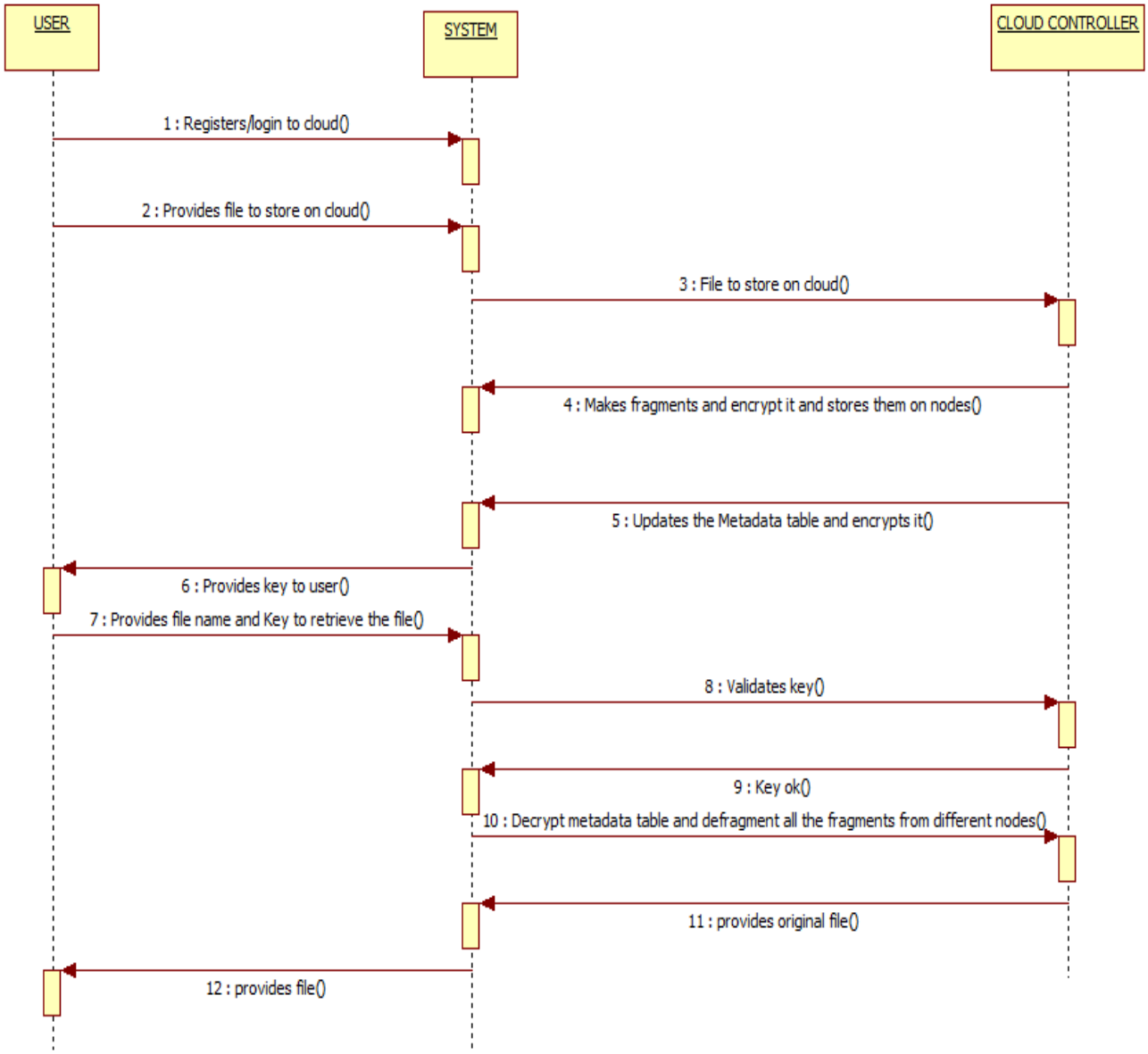


Figure : Sequence Diagram

5.4 Activity Diagram

The following activity diagram shows the step by step flow of the system. The user selects the drive or drives for searching file/folder. User creates the cluster group and puts extension to the groups. User can select a particular group or all the groups for searching file/folder. User inputs keyword to search the file and if extension does not match the selected group, system shows negative result like 'file not found' otherwise the file is displayed.

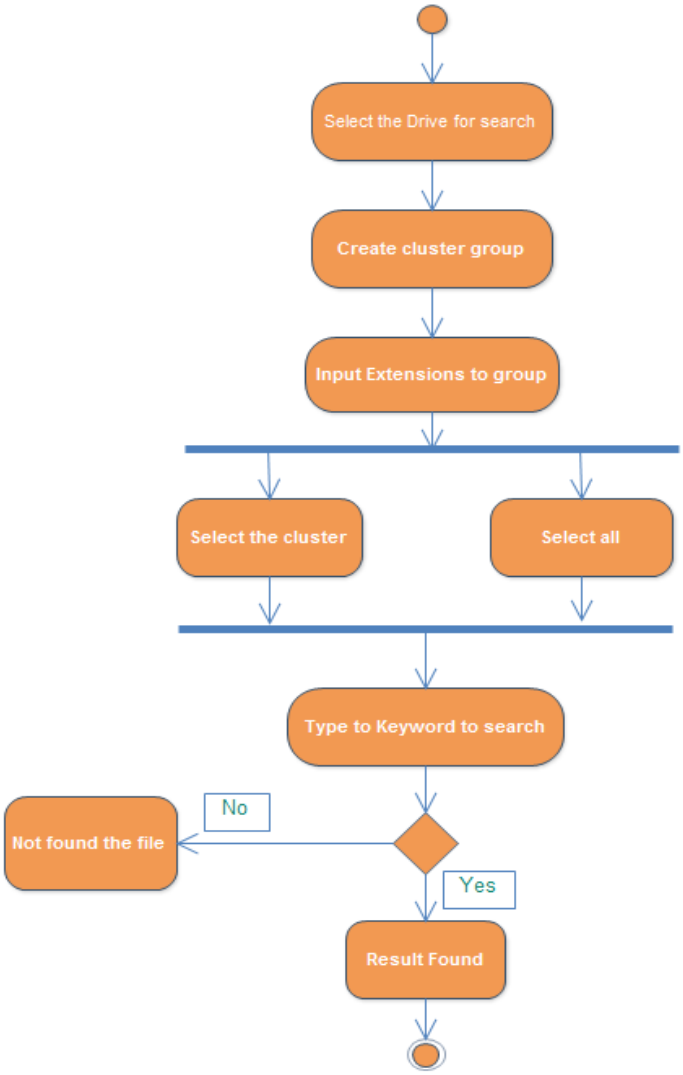


Figure: Activity Diagram

6. IMPLEMENTATION

6.1 Module Specification

Module 1: Mapping

Mapping module takes the input as the data of drive and analyzes the data and finds the similarities between the data. This module performs the feature selection and feature extraction process.

Module 2: Reducing

Reducing module takes the input as the output of the mapping module. And creates clusters based on the feature selection and update the database which results into clustered data.

Module 3: Searching

Searching module takes the query from the user and search for that particular data in the updated database which contains clustered data and displays the result to the user.

6.2 Methodology

6.2.1 Divisive Algorithm

The following agglomerative algorithm is used for searching the files.

Dir=get directory to traverse**;**

N=no. of files in **Dir**;

For i=1->**N**

If (!exist(i)) (check if entry exist in database)

Then put entry of i to database;

End If

End For

M=no. of directories in **Dir**;

For i=1->**M**

If (!exist(i)) (check if entry exist in database)

Then put entry of i to database;

End If

repeat for all i;

End for

7. TESTING

7.1 Testing Resources

- Operating System: Windows
- CPU: Pentium 4
- Memory: 1 GB RAM
- Environment: Microsoft Visual Studio 2008.

7.2 Objective of testing

- To prevent defects.
- To make sure that end result meets the user requirements.
- To provide a good quality product.

7.3 Testing strategy

7.3.1 Unit Testing

For this testing, overall system is being decomposed into number of modules and then these modules are tested separately.

7.3.2 Integration testing

In integration testing many units tested modules are combined into subsystems which are then tested.

7.3.3 System testing

In system level testing, entire software system is tested.

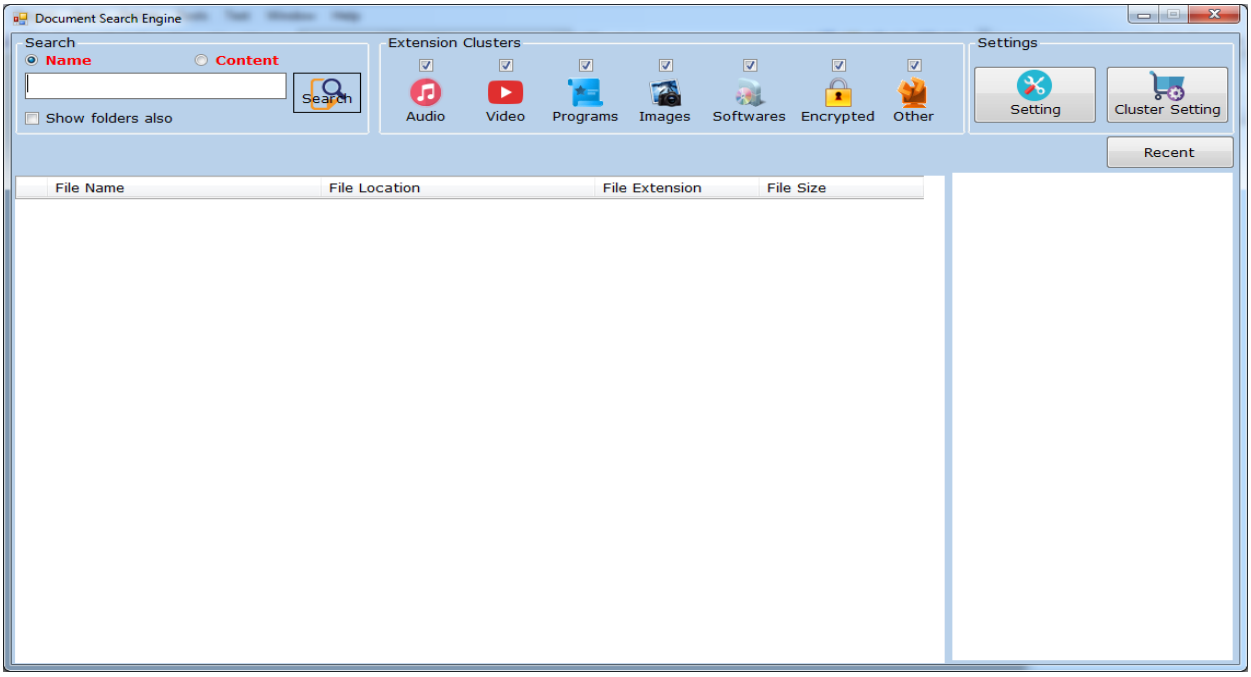
7.3.4 Acceptance testing

Accepting testing essentially test if the system satisfactory solves the problem for which it is designed.

7.4 Test Results

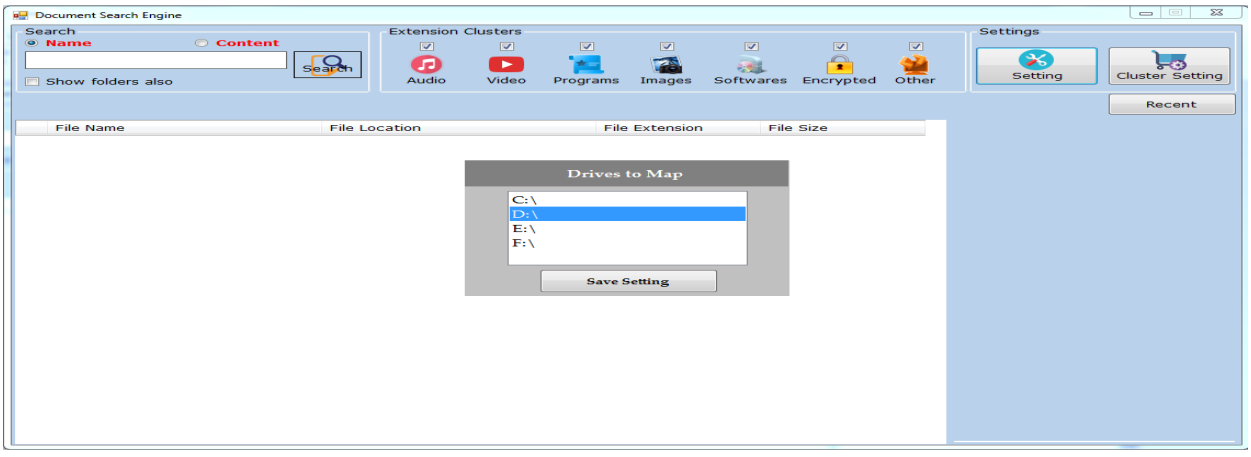
Case number	Case name	Actual Result	Expected Result	Status
1	Searching File in Drives (all)	Display all files	Display all Files in selected Drives	Pass
2	Select particular Clustered group and find files included in that group	Display files included in that group	Display expected files	Pass
3	Select particular Clustered group and find file, not included in that group	File is not found	This file is not found	Pass
4	When the user connects USB	This drive is displayed on setting	Displayed	Pass
5	Enter file *.any extension	Display all files for that extension	Displayed	Pass
6	Enter file *any word *. extension	Display files included that word in file	Displayed	Pass

8. SNAPSHOT



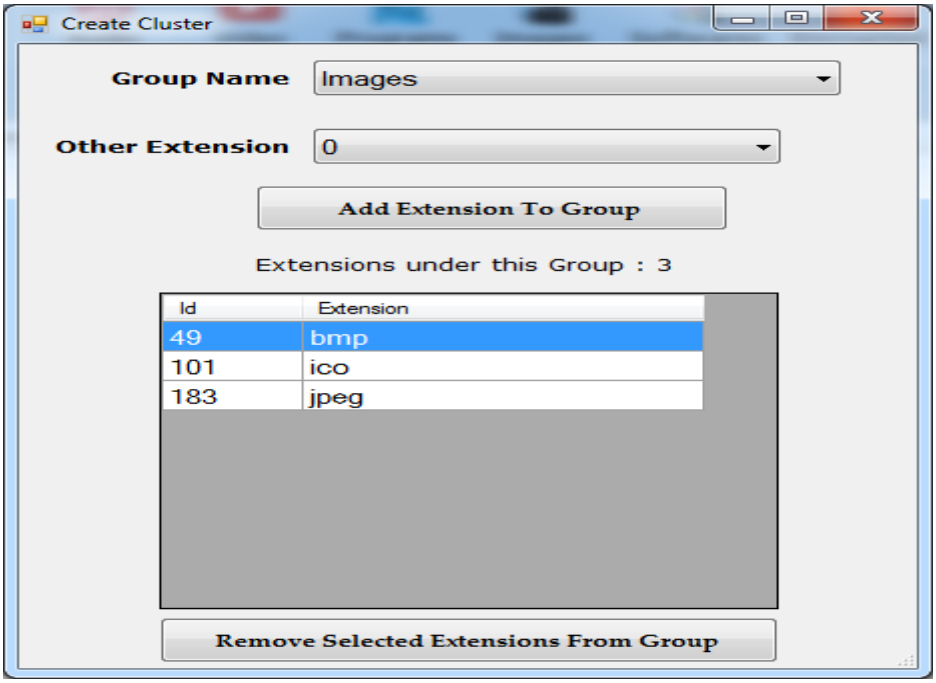
Snapshot 1: "MAIN MENU"

Snapshot 1:- This is the home page of project.



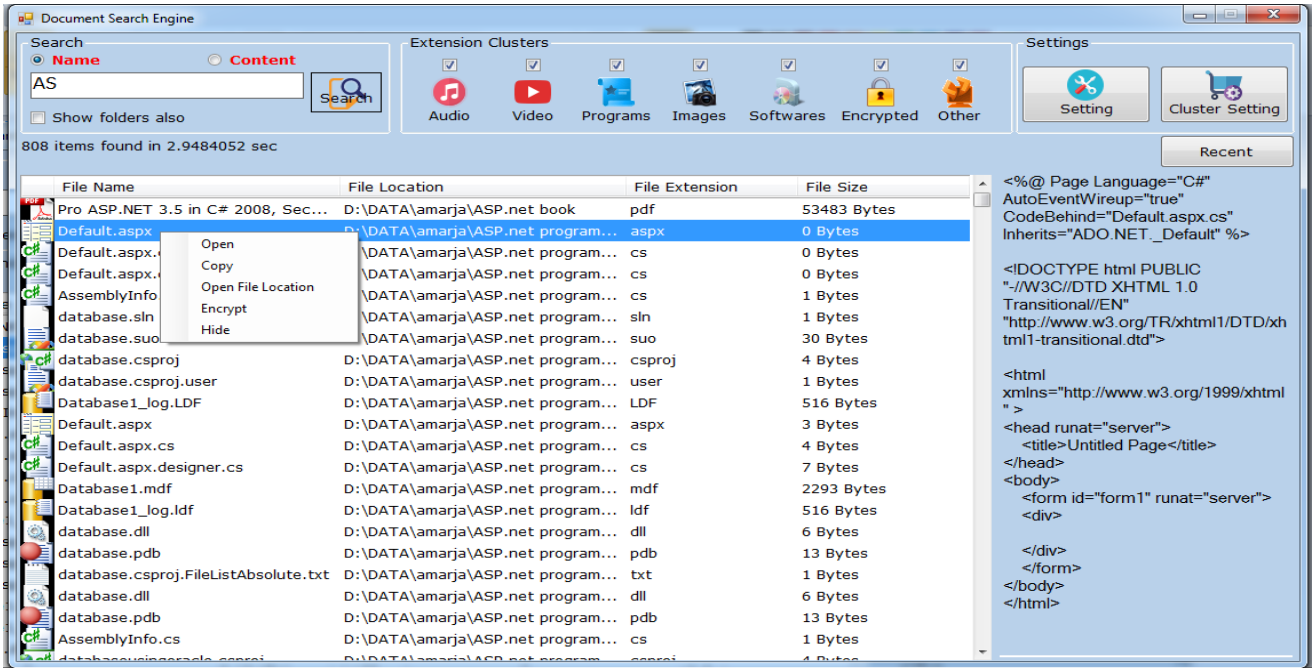
Snapshot 2: "SETTING"

Snapshot 2:- In this window settings are done.



Snapshot 3: "CLUSTERING"

Snapshot 3:-In this window we can create a group and add extensions in it.



Snapshot 4: "SEARCHING"

Snapshot 4:- In this window search results are shown

9. RESULT ANALYSIS (RESULT GENERATED BY YOUR SYSTEM)

9.1 Results & Charts showing results:

- **Searching Graphs**

Figure: Searching Graph for Our Project

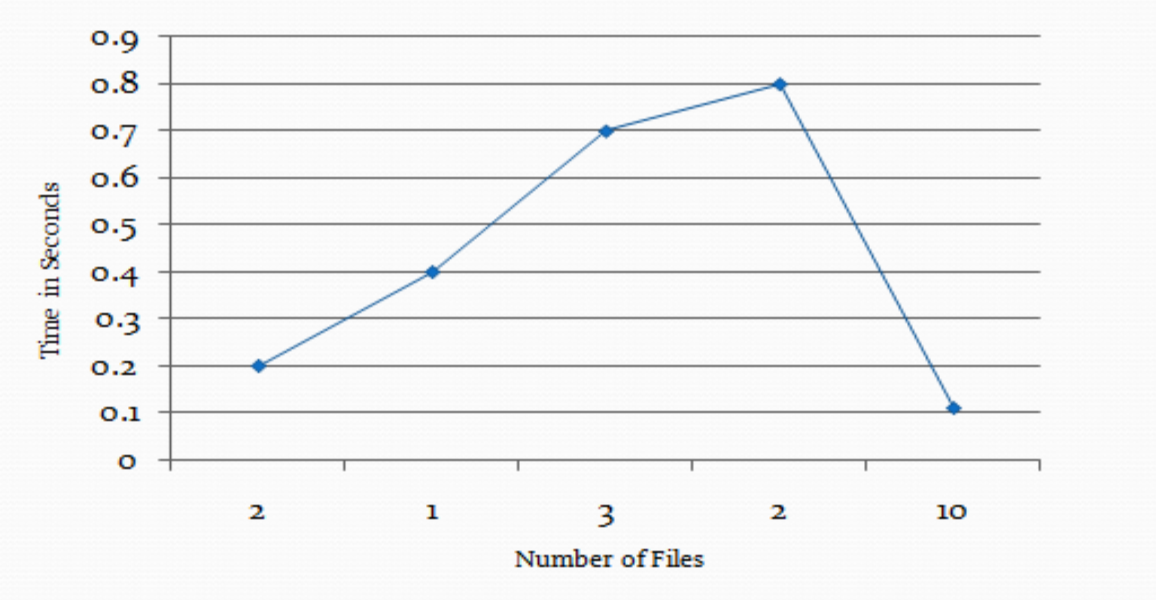


Figure: Searching Graph for Windows 7

9.2 Pros & Cons

Pros

- Fast Searching.
- Provide high security.
- Fast retrieval of data.
- Content based searching

Cons:

- Can not applicable for Distributed System.

10. Conclusion

We are working to develop a project which will be capable of searching a file/folder in a faster way. This will give a better and simple interface to user in every approach the system being compatible. It will be a faster way to search a file/folder in a system. Till now, we have partially developed the system in which the files/folders can be mapped, clustered as per corresponding file extension.

10.1 Future work

Proceeding further, we are going to develop the whole system which will be able to

- Search the files/folders in a faster way as compared to current system.
- Hide/unhide a particular file/folder.
- Provide security to particular file/folder.

11. REFERENCES

[1]Puppala Priyanka, Department of CSE, AVN Inst. of Engg. & Tech., Hyderabad*," An Efficient Algorithm for Clustering Data Using Map-Reduce Approach*", In International Journal of Computer Science and Mobile Computing, Vol.3 Issue.5, May- 2014, pp. 1013-1021.

[2]Kiran Jyoti,Dr. Satyaveer singh, Department of CSE and IT ,GNDEC Ludhiyana,Punjab,India,

" *Data Clustering pproach TO Industrial Process Monitoring,Fault Detection And Isolation ", International Journal OF Computer Application,March 2011.*

[3]R. Subhashin et.al, Research Scholar, Sathyabama University," *The Anatomy of Web Search Result Clustering and Search Engines "*,In Indian Journal of Computer Science and Engineering Vol. 1 No. 4 392-401.

[4]Wang Kay Ngai ,Ben0 Kao,Chun Kit Chui, Reynold Cheng ,Michael Chau,Kevin Y. Yip,University of Hongkong,,"*Efficient Clustering of Uncertain Data*", Proceeding of Sixth International Conferencing on Data Mining(ICDM'06).

[5]S.Ayyasamyet.al, Department of Information Technology, Tamil Nadu College of Engineering*," A Cluster Based Replication Architecture for Load Balancing in Peer-to-Peer Content Distribution "*, International Journal of Computer Networks & Communications (IJCNC) Vol.2, No.5, September 2010.

[6] Ralf L¨ammel, Google's MapReduce Programming Model,Data Programmability Team Microsoft Corp.,Redmond, WA, USA.

[7] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simpli_ed Data Processing on Large Clusters, January 2008/Vol. 51, No. 1.

[8] pros-and-cons-of-hadoop *:*-http://www.guruzon.com/6/introduction/hadoop/pros-and-cons-of-hadoop

12. APPENDIX

Clustering: Dividing datasets into smaller groups such that the members of each small group are as similar as one another.

Data mining: Data mining is defined as the process of analyzing data from different panorama and epitomizing it into useful information

Map: Map function that processes a key/value pair to generate a set of intermediate key/value pairs.

Reduce: Reduce function that merges all intermediate values associated with the same intermediate key.