Data Leakage Detection System for Diabetes Patients DB

Sonali Patil[#1], Hemlata Bhole[#2]

[#]BE Student, Information Technology Department, PCCOE, Pune, India

***Abstract*- In both the industrial and defence area, a forceful need is rising for fast, yet secure, propagation of Information. We centre on field with one information source (sender) and many information sinks (recipients) where: (i) contribution is equally useful for the one who sends and for the one who receives data, (ii) disclosing a pooled information is beneficial to the addressee but adverse to the sender, and (iii) information sharing decisions of the sender are determined using imperfect monitoring of the (un)intended information leakage by the recipients.**

We study the following problem: A data distributor (a physician) has given sensitive data to a set of supposedly trusted agents (research labs). Some of the data are leaked and found in an unauthorized place (e.g., on the web or somebody's laptop). The physician must assess the likelihood that the leaked data came from one or more research labs, as opposed to having been independently gathered by other means. We propose data allocation strategies (across the research labs) that improve the probability of identifying leakages. These methods do not depend on alterations of the released data (e.g., watermarks). In some cases, we can also inject "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party.

***Keywords*-** Fake object, Data leakage, Allocation strategy

I. INTRODUCTION

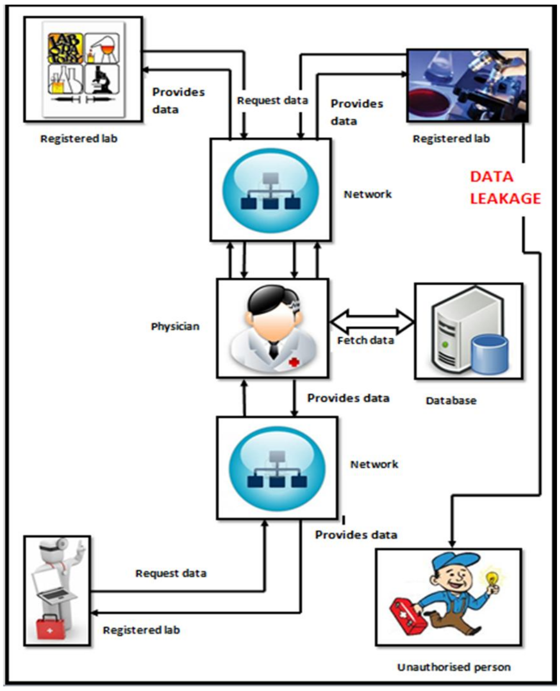
A. Problem Definition

A Physician maintains the database of all patients. This data is requested by different labs for the purpose of analysis or research.

Physician gives the data to labs as per their requests. If some data are leaked and found in an unauthorized place (e.g., on the web or somebody's laptop).And physician comes to know about this leakage he will require finding the guilty. We propose data allocation strategies (across the agents) that improve the probability of identifying leakages. These methods do not depend on alterations of the released data (e.g., watermarks). In some cases, we can also inject "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party.

B. What is data leakage?

Data Leakage is the unauthorized transmission of data (or information) from within an organization to an external destination or recipient. Data Leakage is same as the term Information Leakage. It is shown below:



C. Project Objectives

A physician has given susceptible data to a set of supposedly confidential research labs (third parties). A quantity of the data is leaked and found in an illegal place (e.g., on the web or somebody's laptop). The physician must measure the probability that the leaked data came from one or more labs, as opposed to having been separately gathered by other means. We suggest data allocation strategies (across the labs) that get better probability of discovering leakages. These techniques do not depend on alterations of the distributed data (e.g., watermarks). We can also inject fake patient data records to improve our chances of identifying leakage and find guilty.

Our aim is to detect who leaked the physician's data, and if achievable, to identify the lab that leaked the data.

D: ***Solution***

In this, we consider the option of adding “fake” objects to the scattered set. Such objects do not match to real entities but appear realistic to the agents. The fake objects acts as a watermark for the whole set, without modifying any individuals. If it found that an agent was given one or more fake objects that were leaked, then the distributor can be more self-confident that agent was guilty.

This system gives the better solution over watermarking. The system provides a new approach for detecting the guilty research lab who has leaked the data. The system can be used in different fields where some data is outsourced for processing.

The rest of the paper is organized as follow. Section II contains literature survey. Section III covers proposed data leakage detection System using Fake objects.

E: Existing Technique-

Conventionally, leakage detection is controlled by watermarking, e.g., a unique code is rooted in each distributed copy. If that copy is later exposed in the hands of an illegal party, the leaker can be recognized. Watermarks can be very valuable in some cases, but again, include some change of the original data. Additionally, watermarks can sometimes be demolished if the data receiver is nasty. E.g. A hospital may provide patient records to research labs who will devise new treatments. Similarly, a company may have businesses with other companies that require input customer data. Another enterprise may subcontract its data handling, so this must be given to various other companies. We call the owner of the data the physician and the supposedly trusted third parties the research labs.

II. PROPOSED SCHEME

A. Block diagram

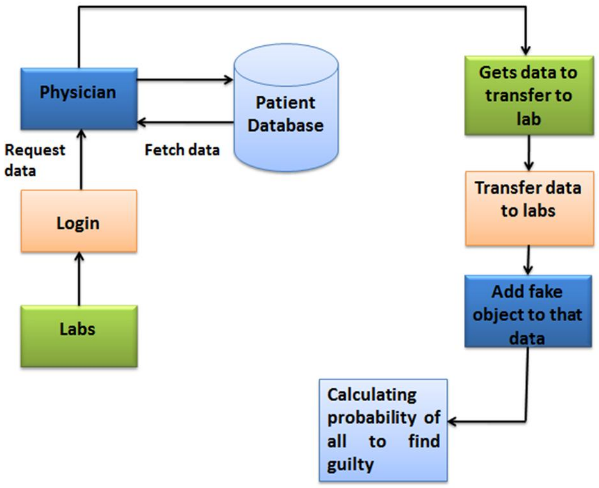

Physician
Patient Database
Request data
Fetch data
Login
Labs
Gets data to transfer to lab
Transfer data to labs
Add fake object to that data
Calculating probability of all to find guilty

Fig.1.Block Diagram

The system is composed of several modules each providing some functionality.

The first module is concerned with allocating the requested data to the labs. Research labs request the data from a physician with any specific condition (Explicit request) or number of records (Sample request). According to the request ID, the physician searches the data requested by the lab in his database using the condition. Physician prepares the fetched data to give it to the lab.

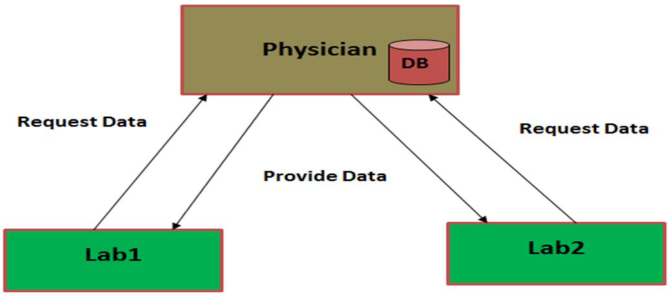


Fig.2 Data Allocation

The second module consists of creation of fake object; According to the condition, the physician adds the record from fake record table to the original data. The physician maintains the fake database along with original database.

The third module is data distribution. Now the physician is ready to send the requested data along with the fake one. But the research lab does not know about the fake one. Physician sends the data to research lab. The status changes to data sent.

B: Fake objects-

The physician may be able to add fake objects to the circulated data in order to improve his utility in detecting guilty research labs. However, fake objects may influence the correctness of what research labs do, so they may not always be acceptable. The idea of perturbing data to detect leakage is old technique. However, in most cases, individual objects are perturbed, e.g., by adding a watermark to text file. In our case, we are perturbing these of physician objects by adding fake elements.

C: Algorithm for Explicit Data Requests with Fake objects

Input: $U_1,\ldots,U_n$, $cond_1,\ldots,cond_n$, $r_1,\ldots,r_n$, R

Output: $U_1,\ldots,U_n$, $F_1,\ldots,F_n$

```
U <- Φ
for i=1, …. ,n do
        if $r_i > 0$ then
              U <- U U {i}
$F_i$ <- Φ
while R > 0 do
      i <-  SELECTAGENT($U_1,U_2,\ldots,U_n$)
      f <- CREATEFAKEOBJECT($U_i$ , $F_i$ , $cond_i$)
$U_i$  <-  $U_i$ U {f}
 $F_i$  <-  $F_i$ U {f}
  $r_i$  <-  $r_i - 1$
If $r_i$ =0 then
 U  <-  U \ {$U_i$}
 R  <- R - 1
```

The fourth module is data leakage detection. The physician comes to know about the leakage(he finds data that he has given to registered labs on any website or on anybody's laptop).Then he calculates the probability of each lab by comparing the leaked data to the data given to the labs .He declares the lab with highest probability as a Guilty one.

C: Data Leakage-

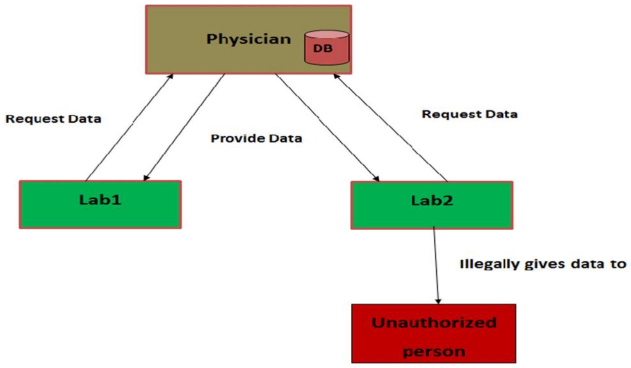


Fig.3.Data Leakage

D: Finding guilty by calculating probability:

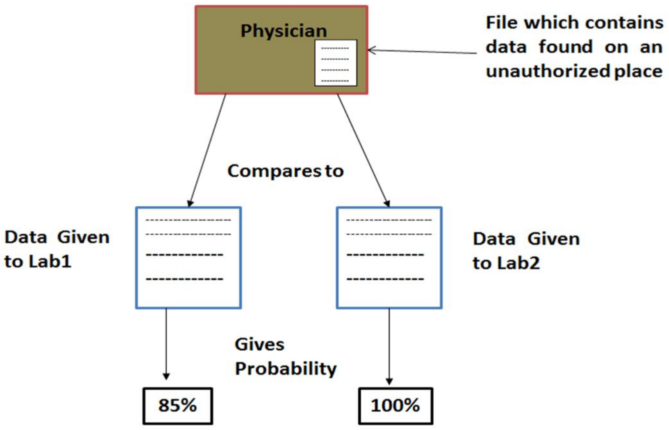
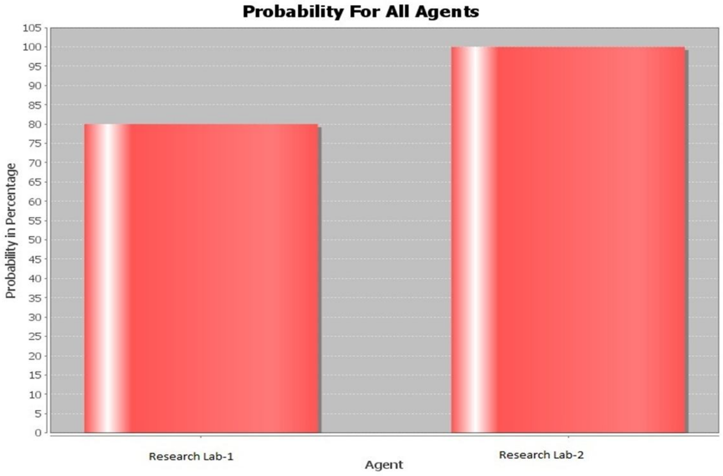
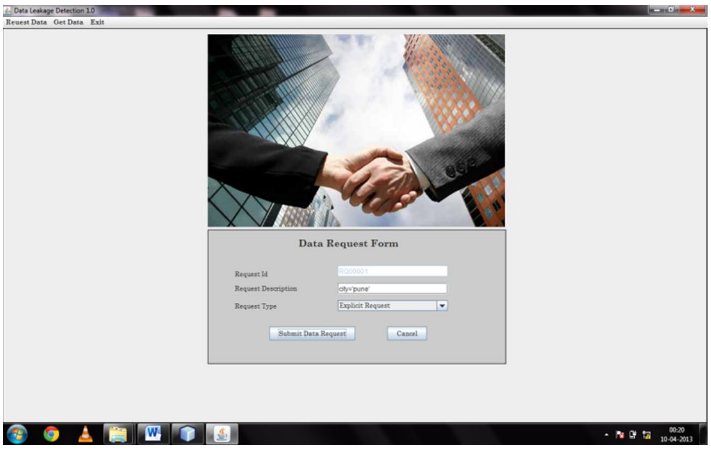


Fig.4.Finding Guilty

E: Graph of probabilities:



III. Results



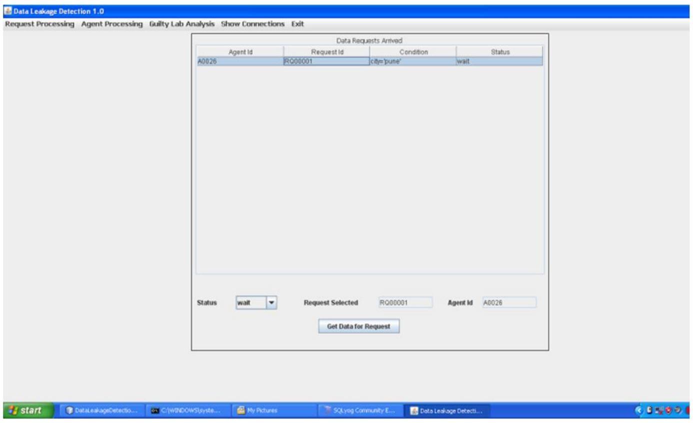
Data Request Form

Request Id RQ00001

Request Description city='pune'

Request Type Explicit Request

Fig. 1 Research lab request data

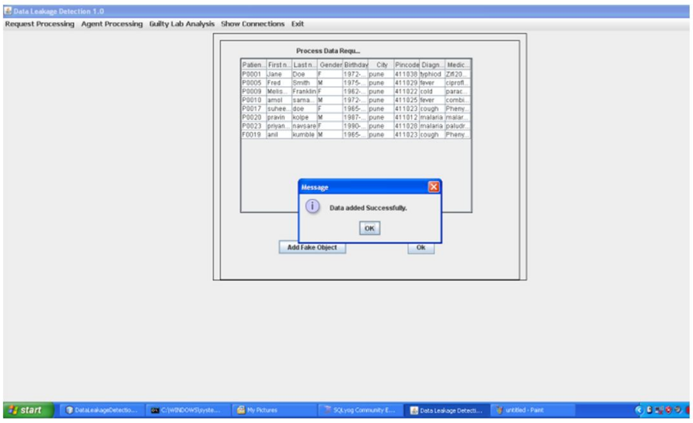


Data Requests Arrived

Agent Id	Request Id	Condition	Status
A0026	RQ00001	city='pune'	wait

Status: wait
Request Selected: RQ00001
Agent Id: A0026

Fig. 2 Physician checks data request

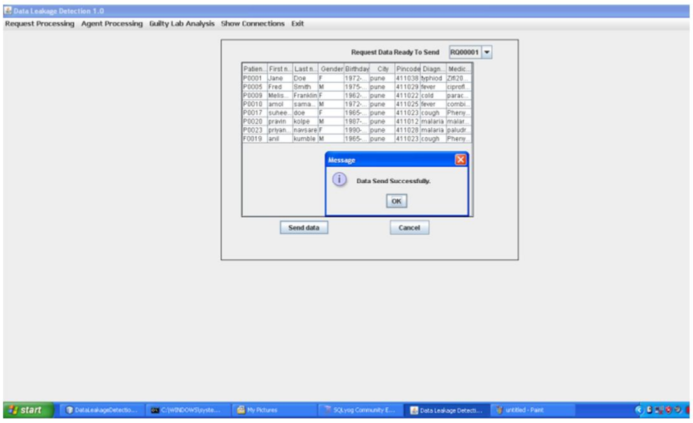


Process Data Requ...

Patien...	First n...	Last n...	Gender	Birthday	City	Pincode	Diagn...	Medic...
P0001	Jane	Doe	F	1972-...	pune	411038	typhiod	Zifi20...
P0005	Fred	Smith	M	1975-...	pune	411029	fever	ciprofl...
P0009	Melis...	Franklin	F	1962-...	pune	411022	cold	parac...
P0010	amol	sama...	M	1972-...	pune	411025	fever	combi...
P0017	suhee...	doe	F	1965-...	pune	411023	cough	Pheny...
P0020	pravin	kolpe	M	1987-...	pune	411012	malaria	malar...
P0023	priyan...	navsare	F	1990-...	pune	411028	malaria	paludr...
F0019	anil	kumble	M	1965-...	pune	411023	cough	Pheny...

Data added Successfully.

Fig. 3 Physician adds fake object

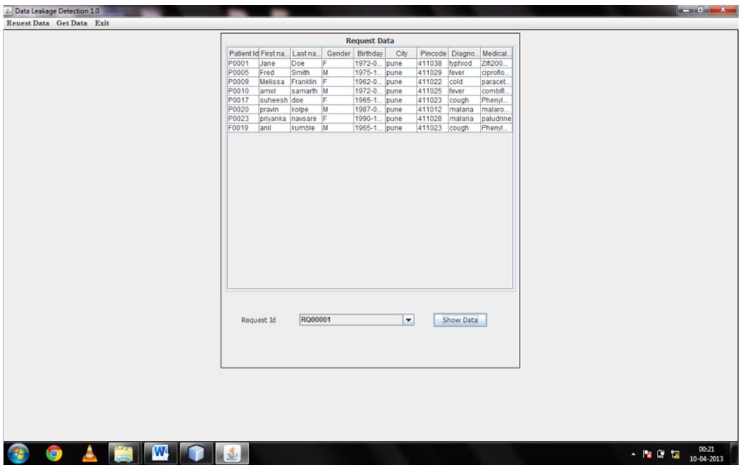


Request Data Ready To Send: RQ00001

Patien...	First n...	Last n...	Gender	Birthday	City	Pincode	Diagn...	Medic...
P0001	Jane	Doe	F	1972-...	pune	411038	typhiod	Zifi20...
P0005	Fred	Smith	M	1975-...	pune	411029	fever	ciprofl...
P0009	Melis...	Franklin	F	1962-...	pune	411022	cold	parac...
P0010	amol	sama...	M	1972-...	pune	411025	fever	combi...
P0017	suhee...	doe	F	1965-...	pune	411023	cough	Pheny...
P0020	pravin	kolpe	M	1987-...	pune	411012	malaria	malar...
P0023	priyan...	navsare	F	1990-...	pune	411028	malaria	paludr...
F0019	anil	kumble	M	1965-...	pune	411023	cough	Pheny...

Data Send Successfully.

Fig. 3 Physician sends data

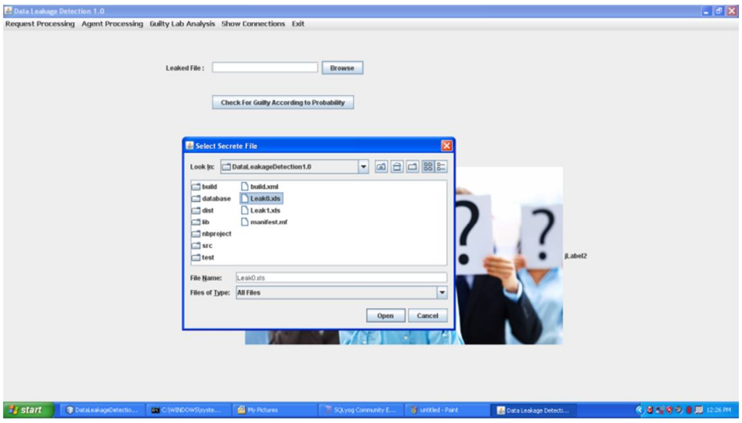


Request Data

Patient Id	First na...	Last na...	Gender	Birthday	City	Pincode	Diagno...	Medical...
P0001	Jane	Doe	F	1972-0...	pune	411038	typhiod	Zifi200...
P0005	Fred	Smith	M	1975-1...	pune	411029	fever	ciproflo...
P0009	Melissa	Franklin	F	1962-0...	pune	411022	cold	paracet...
P0010	amol	samarth	M	1972-0...	pune	411025	fever	combifl...
P0017	suheesh	doe	F	1965-1...	pune	411023	cough	Phenyl...
P0020	pravin	kolpe	M	1987-0...	pune	411012	malaria	malaro...
P0023	priyanka	navsare	F	1990-1...	pune	411028	malaria	paludrine
F0019	anil	kumble	M	1965-1...	pune	411023	cough	Phenyl...

Request Id: RQ00001

Fig. 4 Research lab views data



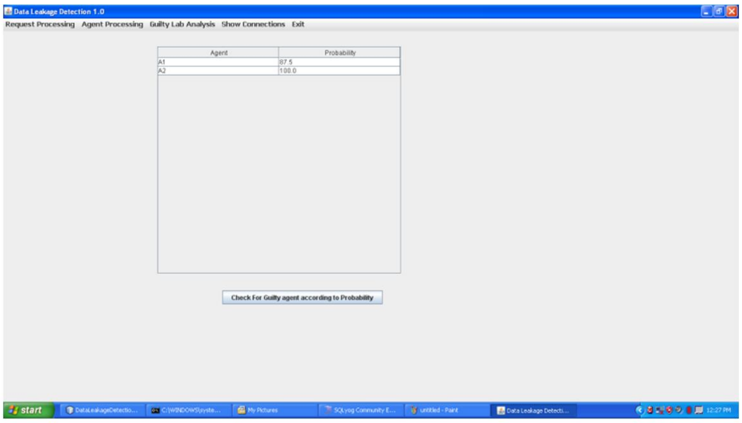
Leaked File:

Look In: DataLeakageDetection1.0

File Name: Leak0.xls

Files of Type: All Files

Fig. 5 Physician browses the leaked file



Agent	Probability
A1	87.5
A2	100.0

Fig. 6 Physician gives probability of being guilty of all research labs

IV. Conclusion and Future Scope

A) Conclusion:

The probability that a lab is guilty for a leak is measured, based on overlie of lab's data with the leaked data and the data of other labs. The algorithms we have given implement a diversity of data distribution strategies that can perk up the physician's chances of finding a leaker. We have mentioned that releasing objects with care can make a important difference in identifying guilty labs, especially in cases where there is big overlap in the data that labs must get.

B) Future Scope:-

Our future work includes the study of agent guilt models that confine leakage situations that are not calculated in this paper. For example, what is the suitable model for situations where agents can plot and identify fake tuples? Another problem is the addition of our allocation strategies so that they can process agent requests in an online way (in our scenario we assume that there is a permanent set of agents with requests known before).

REFERENCES

[1] Panagiotis Papadimitriou (Student Member, IEEE),and Hector Garcia (Molina, Member, IEEE), **'Data Leakage Detection'**, Department of Computer Science, Stanford University, Gates Hall 4A, Stanford, CA 94305-9040, *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011*

[2] Polisetty Sevani Kumari, Kavidi Venkata Mutyalu, **"Development of Data leakage Detection Using Data Allocation Strategies"**, M.Tech, CSE, Sri Vasavi Engineering College, Pedatadepalli, Tadepalligudem W.G.Dt, A.P. India, *INTERNATIONAL JOURNAL OF COMPUTER TRENDS AND TECHNOLOGY-VOLUME3 ISSUE4- 2012*

[3] Rohit Pol, Vishwajeet Thakur, Ruturaj Bhise, Prof. Akash Kate, **"Data leakage Detection ",** *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH AND APPLICATIONS (IJERA)*

[4] Mr. Vaibhav Narawade (Associate Professor), Unnati Kavali (Student of P.V.P.P.C.O.E.), Tejal Abhang (Student of P.V.P.P.C.O.E.), **'DATA ALLOCATION STRATEGIES IN DATA LEAKAGE DETECTION',** *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH AND APPLICATIONS (IJERA)* ISSN: 2248-9622, VOL. 2, ISSUE 2, MAR-APR 2012

[5] http://archive.ics.uci.edu/ml/machine-learning-databases/diabetes/

[6] Sridhar Gade, Kiran Kumar Munde, Department of CSE, DRK Institute of Science and Technology ,Krishnaiah.R.V. Principal, Department of CSE, DRK Institute of Science and Technology, Andhra Pradesh**,"DATA ALLOCATION STRATEGIES FOR LEAKAGE DETECTION"**,IOSR JOURNAL OF COMPUTER ENGINEERING (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 VOLUME 5, ISSUE 2 (SEP-OCT. 2012), PP 30-35

[7] **The Complete Reference Java,** Seventh Edition, Herbert Schildt, Tata McGraw-Hill Education(2006)

[8] **Software Engineering A Practioner's Approach**, Fifth Edition, Roger S. Pressman

[9] The **Unified Modelling Language** User Guide, Second Edition, Grady Booch, James Rumbaugh, Ivar Jacobson