Abstract—In this paper, the approach of data retrieval is proposed from Intelligent and Distributed Data Warehouse (IDDW), which is an N-Levels hierarchal Distributed Data Warehouse as proposed by us earlier. On applying the proposed approach, data from the IDDW for the corresponding Unique Identification Number (UIN) entered by the user is retrieved. The approach of data retrieval starts with the user entering the UIN, corresponding to which the data warehouse is located in IDDW. Once, the data warehouse is located the desired data is retrieved.

Keywords—Data retrieval; Intelligent; data warehouse; Unique Identification Number

I. INTRODUCTION

In the world, where things are changing with leaps and bound, so is the technology. The change can also be seen in the geographical expansion of the organizations. To keep pace with the change, organizations are also changing their way of storing the data as well mining of the data from Data Warehouse. The basic definition of Data Warehouse was given by Inmon W. H. in [19] that Data Warehouse is a subject-oriented, integrated, time-varying, non-volatile collection of data in support of the management’s decision-making process. Organization’s expanding to various locations in a geographical area needs the distribution of the centralized Data Warehouse. Many methods were proposed by many authors for both distributions of data warehouses. In [2] we have proposed IDDW which is a hierarchal distributed data warehouse. In [3] we have presented a novel approach of storing the data in the Common table of a most suitable data warehouse in IDDW. The approach has been shown to perform well, which is verified experimentally.

The fast and correct retrieval of information for proper decision making has become an important issue. Frawley W. J. et.al. in [13] refers to the data retrieval the overall process of discovering new patterns or building models from a given dataset. Data retrieval techniques also come in the category of analytical systems that help to give insight into hidden information.

Data retrieval techniques evolved as a requirement when enormous data started accumulating in digital format. The existing data retrieval algorithms can work in three different computing environments: Centralized, Parallel and Distributed.

In this paper, the approach of data retrieval from IDDW is designed and developed. On applying the developed approach, data from the IDDW for the corresponding UIN entered by the user is retrieved from the IDDW.

The next section presents the related works on data retrieval in context of Centralized, Parallel and Distributed approaches of data retrieval.

II. RELATED WORKS

A. Centralized Approach of Data Retrieval

Organizations may have multiple repositories of the transactional data depending on the location of their office. In the centralized approach, data is extracted and accumulated on a centralized store after cleaning and pre-processing. From, this central store, task relevant data is selected and retrieval techniques are applied. Initially, data retrieval techniques were restricted to centralized processing only, as discussed by Agrawal R. et.al. in [1], Bradley P. S. et.al. in [5], Chiang D. et.al. in [8], Duan L. et.al. in [10], Duan L. et.al. in [11].

B. Distributed Approach of Data Retrieval

According to Fu, Y. in [14] and Park B. et.al. in [22] data retrieval from distributed data warehouse is concerned with the application of the classical data retrieval procedure in a distributed computing environment trying to make the best of the available resources (communication network, computing units and databases). Data retrieval takes place both locally at each distributed site and at a global level.

Most methods for distributed retrieval of data assume that data is horizontally partitioned among sites. Each site mines its local data and generates locally valid concepts. These concepts are exchanged among all the sites to obtain the globally valid concepts. Some algorithms and systems used for distributed retrieval of data are as follows: The Partition Algorithm by Savasere A. et.al. in [26]; JAVA based multi-agent system JAM by Stolfo S. et.al. in [27], Prodromidis A. et.al. in [25], Guo Y. et.al. in [17]; Parthasarathy S. et.al. in D-DOALL [23] uses distributed do-all primitive for easy scheduling of independent retrieval task on a network of workstations; Grossman R. l. et.al. in [16] proposed The Papyrus, a JAVA based system targets wide area distributed data over cluster and meta-clusters; and JAVA based system for distributed enterprise by Chattratichat J. et. al. in [7].

C. Parallel Approach of Data Retrieval

Parallel approach for data retrieval deals with tightly-coupled systems including shared-memory systems,
distributed-memory machines, or clusters of shared-memory systems workstations with a fast interconnect. The parallel formulation must address the issue of efficiency and scalability both in terms of memory requirements and parallel runtime. Data retrieval in a highly parallel environment over multiple processors was explained by Wang L. et.al. in [28].

There are two Parallel Programming Models in common use: Thread (POSIX Threads by Butenhof D. R. [6]) and Message Passing (OpenMP by Dagum L. et.al. in [9]). Modern programming languages are also structured so as to efficiently utilize novel architectures. There exist dedicated parallel programming paradigms for parallelizing the algorithms over multiprocessor and networked systems, OPENMP and MPI are used to achieve shared and distributed memory parallelization. CUDA is a programming language that is designed for parallel programming used by Nickolls J. et.al. in [21]. In CUDA, thread access different memories of GPU. CUDA offers a data parallel programming model. Parallel programming is incomplete without discussing the most recent approach called MAP Reduce. It can process large sized data in a highly parallel manner as shown by Bhaduri K. et.al. in [4].

Various data retrieval algorithms has been modified for parallel processing architectures as discussed by Parthasarathy S. et.al. in [24], Zaiane O. R. et.al. in [29], Zaki M. J. et.al. in [30], Huang Z. in [18], Kuok T. et.al. in [20], Foti D. et.al. in [12] and Goil S. et.al. in [15].

In this section, the approach of data retrieval is being presented. The flow chart of the approach, the role of various components used in the approach and the illustration of interconnections of various components with overall working is presented.

### A. Flowchart of the Approach

The flowchart of the approach of retrieving data from IDDW is shown in Fig. 1. It is the modified flowchart of the approach to store user data in the Common table of a most suitable data warehouse in IDDW presented in [3].

The user enters the UIN for which data is to be retrieved. It enters the UIN via a Form (GUI) designed for it. The entered UIN is passed to Identifier Finder in Data Warehouse Locator Bridge (DLB) to calculate the Identifier.

Data Warehouse Locator uses this Identifier to locate the data warehouse in IDDW that store the data corresponding to the UIN entered by the user. The information of availability of the data warehouse for calculated Identifier is obtained from Central Look-Up Data Warehouse. Identifier Finder recursively calculates Identifier until the desired information is obtained from Central Look-Up Data Warehouse. The UIN and desired information obtained from Central Look-Up Data Warehouse is then validated and verified for the error if any occurred in it.

After the UIN is validated and verified with no error the data warehouse in IDDW is identified from the information obtained from Central Look-Up Data Warehouse. The identified data warehouse is same, which contains data in its Common table corresponding to the UIN. The data corresponding to this UIN is then returned to the desired user. A temporary connection is also established as shown in Fig. 1, between the user and identified data warehouse. This temporary connection is used for further retrieving of data from the Local tables of identified data warehouse.

If on validating and verifying UIN an error is found then the entire process gets repeated.
B. Design and Role of different Components used in the Approach of Data Retrieval from IDDW

The components used in the approach of data retrieval from IDDW are same as used in the approach, to store user data in the Common table of a most suitable data warehouse in IDDW [3] except Mapping Table and ETL. These two components are not required here in the approach of data retrieval from IDDW.

The design and role of all other components except User Interface Module are same in the two approaches. The User Interface Module is designed differently here and its design and role is presented in next sub-section.

1) User Interface Module

To retrieved the data from IDDW, there is a need of an interface through which user can enter the UIN. The UIN to be enter is one for which the data is to be retrieved from IDDW. Thus, the User Interface Module designed to retrieve data from IDDW comprises of a Form, like shown in Fig. 2.
C. Components Interconnections and Data Flow in the Data Retrieval Approach

This section presents the interconnection of various components used and the flow of data and information in the approach.

Components inter-connections and data flow in the approach of locating data warehouse in IDDW is modified accordingly, so that using the same approach data can be retrieved from IDDW. Fig 3 shows the Components inter-connections and data flow in the approach of data retrieval. In it, the content presented in bold are new/ modified. The content that is un-highlighted is not needed in the approach of retrieving data from IDDW. Rest of the portion of Fig. 3 is same as shown [3].

The approach of data retrieval from IDDW starts by the user entering the UIN for which the data is to be retrieved. The entered UIN by the user via Form designed in User Interface Module is moved to the GRAM, which is stored there. Replica Manager produces two copies of this UIN. One copy of it is moved to Identifier Finder in DLB and other copy is moved to summation block in DLB. The rest of the procedure to obtain Machine Address, for the entered UIN, from Central Look-Up Data Warehouse is same as discussed in the approach of locating data warehouse in IDDW [3]. Once the Machine Address is obtained for the entered UIN than at the summation block in DLB the UIN and Machine Address is clubbed together and moved to LRAM. The procedure of Validation and Verification of UIN is same as discussed in the approach of locating data warehouse in IDDW. Here no ETL tool is needed, so finally, a data warehouse in IDDW is located where the data is stored for the entered UIN. The user gets connected to this data warehouse that may now obtain data from both Common table and Local table of this data warehouse. Similar procedure repeats for all the users who want to retrieve data.
After presenting in the previous sections, the flow chart of proposed approach, designing and role of components used in proposed approach and component interconnections and data flow in proposed approach, the Case Study is presented in the next section.

IV. CASE STUDY

The 8-Level hierarchal structure presented in the case study taken in [2] is utilized here to evaluate the approach presented in this paper. The approach of data retrieval is applied on a hierarchy $H_1$ defined in [3] as $H_1: 2013(13) | \text{Uttar Pradesh}(30) | UPTU(20) | \text{KNIT}(104) | \text{CSE}(10) | \text{Faculty}(5) | \text{Professor}(2) | \text{Unique ID}$. A UIN $= 13302010410520017$ is generated for the data (Location Names) entered by the user belonging to the hierarchy $H_1$. The other possible UIN’s that may be generated for the different users belonging to the hierarchy $H_1$ are (UIN = 1330201041052002C_k to 1330201041052999C_k). In total for each possible 999 different users belonging to the hierarchy $H_1$ an UIN’s may be generated. These number of UIN’s generated are considered to verify the approach of data retrieval from IDDW.

Next section presents the Experimental Work that contains two subsections i.e., Experimental set-up and an Experiment. Conclusion

V. EXPERIMENTAL WORK

The hardware, software and set-up required to implement the approach of data retrieval is presented in the subsection i.e. experimental set-up. In other subsection named as experiment the detail of experiment performed is presented.
A. Experimental Set-Up

The implementation of the presented approach of data retrieval from IDDW is done by writing the programs in JAVA with NetBeans: 6.9 as an IDE, using the Web server: Apache Tomcat 6.0.26. All the tables required are built in the database MySql 5.0.45. The hardware requirements include one 16-port Switch, one database server and ten machines each with the following specifications: Core 2 Duo processor, 3.0 GHz, and 1GB RAM. The implementation details are same as presented in [3].

![Experimental Set-Up for the proposed approach of data retrieval from IDDW.](image)

The set-up is prepared to implement the hierarchy H1. It is same as presented in [1]. Other components like Central Look-Up data warehouse, DLB Validation and Verification module are implemented in same manner as presented in [3]. One of the machines is used to enter UIN by the user for which data is to be retrieved. The same machine also performs all the roles of User Interface Module. Here User Interface Module is designed differently as compared to the User Interface Module designed in [3]. The components GRAM and Replica Manager are used as standard and placed in shared memory and are accessible to all the users. ETL tool is not needed here as compared to in [3]. Thus it is un-highlighted in the Fig. 4.

B. Experiment

Initially, the UIN “13302010410520017” is entered via form developed. The first Identifier, calculated by Identifier Finder for this UIN is “1330201041052001”. Data Warehouse Locator looks for the Machine Address corresponding to this Identifier in the tables of Central Look-Up Data Warehouse.

As in experiment set-up initially a data warehouse is placed at the Level Year, so, corresponding to this Identifier a message, “No DW is found” is returned to Data Warehouse Locator in DLB. The same process of generating Identifier by Identifier Finder and looking for machine address by Data Warehouse Locator repeats until Identifier (13) is generated. For the Identifier (13), machine address 00-1A-4D-1C-AB-49 is returned to Data Warehouse Locator. The obtained machine address is moved to LRAM along with the UIN “13302010410520017” via summation block. The Validation and Verification block re-calculates check digits and compare it with original. Both the digits come out to be same in this case which means no error has occurred in UIN. Once UIN is checked for an error, the record against the entered UIN is retrieved from the Common table of the data warehouse located in IDDW for the corresponding machine address 00-1A-4D-1C-AB-49.

VI. RESULTS

The time elapsed in the execution of entire process, starting, after entering the UIN till the retrieval of data from the corresponding data warehouse at Level Year of hierarchy H1 comes out to be 6.1 milliseconds. The other 998 UIN’s for same hierarchy H1 with unique ID varying from 002 to 999 were entered. Average time taken by each comes out to be 6.4 milliseconds. Average time 6.4 millisecond is greater than 6.1 because of the following reasons: communication delays; resending of information from GRAM due to identification in error that might have been generated in UIN.

Further, experiment has been performed by placing a data warehouse at each level of hierarchy H1 one-by-one from the top level and same numbers of UIN’s were entered. The following results are obtained, which are shown in table 1.

![The Hierarchy H1 of 8-level Hierarchal Structure](image)

<table>
<thead>
<tr>
<th>Level in hierarchy</th>
<th>Location of Data Warehouse in hierarchy (H1)</th>
<th>Identifier</th>
<th>Average Time in milliseconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Year</td>
<td>13</td>
<td></td>
<td>6.4</td>
</tr>
<tr>
<td>2 Uttar Pradesh</td>
<td>1330</td>
<td></td>
<td>6.1</td>
</tr>
<tr>
<td>3 UPTU</td>
<td>133020</td>
<td>2</td>
<td>5.7</td>
</tr>
<tr>
<td>4 KNIT</td>
<td>13302010401</td>
<td></td>
<td>5.4</td>
</tr>
<tr>
<td>5 CSE</td>
<td>133020104010</td>
<td></td>
<td>4.9</td>
</tr>
<tr>
<td>6 Faculty</td>
<td>1330201040105</td>
<td></td>
<td>4.3</td>
</tr>
<tr>
<td>7 Professor</td>
<td>13302010401052</td>
<td></td>
<td>3.8</td>
</tr>
</tbody>
</table>

The table 1 shows the average time needed to retrieve the data from various specified locations at different levels in hierarchy H1. The identifier calculated by using algorithm in DLB for each location is also shown in the table. As the data
warehouse is placed at lower levels of hierarchy H1, the time to retrieve data gets reduced.

**TABLE II.** The Percentage of Data Retrieved Correctly from the Common Table of Data Warehouse Placed at Different Levels of Hierarchy H1.

<table>
<thead>
<tr>
<th>Level in Hierarchy</th>
<th>Location of Data Warehouse in hierarchy (H1)</th>
<th>Location Identifier</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Year</td>
<td>13</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Uttar Pradesh</td>
<td>1330</td>
<td>91</td>
</tr>
<tr>
<td>3</td>
<td>UPTU</td>
<td>133020</td>
<td>93</td>
</tr>
<tr>
<td>4</td>
<td>KNIT</td>
<td>1330201040</td>
<td>95</td>
</tr>
<tr>
<td>5</td>
<td>CSE</td>
<td>133020104010</td>
<td>96</td>
</tr>
<tr>
<td>6</td>
<td>Faculty</td>
<td>133020104105</td>
<td>98</td>
</tr>
<tr>
<td>7</td>
<td>Professor</td>
<td>1330201041052</td>
<td>99</td>
</tr>
</tbody>
</table>

The table 2 shows the data retrieved correctly (in percentage) from the data warehouse placed in various levels of hierarchy H1. The correct data is the data that must be retrieved for the entered UIN. It may be seen from the values in table 2 that as the data warehouse is placed at lower levels of hierarchy H1, the percentage of correct data retrieved increases. It is because the number of times the Identifier is calculated are less, thus, chances of error are less.

**VII. CONCLUSION**

In this work we have experimentally seen that the data from the data warehouses, available at various levels of IDDW, can be retrieved by using UIN of the user. It is investigated from the experiment that average time to retrieve data is reduced from 6.4 milliseconds to 3.8 milliseconds, when data warehouse is placed at lower most level of hierarchy H1 as compared to data warehouse placed at top-most level of hierarchy H1. As hierarchy H1 is formed from the 8-level hierarchical structure that is analogous to IDDW so it is concluded from the result that as data warehouse is build at lower level of IDDW the time elapses to retrieve data decreases. It is also observed from the results that are obtained by performing experiment that the correct data retrieved also increases from 90% to 99% when data warehouse is placed at a location in lower most level of IDDW as compared to data warehouse placed at the top-most level of IDDW.

**REFERENCES**


[18] Huang Z.,” A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining.”, In Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Department of Computer Science, The University of British Columbia, Canada, pp. 1–8, 1997.


