**Database Technology Review**

**Data Warehousing and OLAP: Design Methodologies and Techniques**

**Written Assignment #1**

### By:

**Keerthi Vardhan Sovenahalli Karanam Student ID : 885935502**

**CPSC 531-03 (22474)**

**Spring, 2022 Professor: Dr. Chun-l Phillip Chen Department of Computer Science**

**California State University, Fullerton March 27, 2022**

**Table of Contents**

Abstract …………………………………………………………………………………………………………………………. 4

1.0 Introduction …………………………………………………………………………………………………………….. 5

2.0 Literature Review ……………………………………………………………………………………………………. 6

2.1 Architecture and End-to-End Process …………………………………………………………… 7

2.2 Working of Conceptual Model ……………………………………………………………………… 9

2.3 Utilities and Tools for Back-end ……………………………………………….…………………. 11

2.4 Database Design Methodology …………………………………………………..……………… 12

3.0 Discussions ……………………………………………………………………………………………………………. 14

3.1 OLTP Use Cases ……….…………………………………………………………………………………. 14

3.2 OLAP Use Cases …………………………………………………………………………………………. 15

3.3 Advantages and Disadvantages of OLTP …………………………………………………….. 15

3.4 Advantages and Disadvantages of OLAP …………..……………………………………….. 16

4.0 Conclusion and Implications …………………………………………………………………………………. 17

5.0 References ……………………………………………………………………………………………………………. 17

**List of Figures**

* + 1. Figure 1: Data Warehousing and its Components 6
    2. Figure 2: Feature of Cloud Computing… 6
    3. Figure 3: Application of DBMS… 9
    4. Figure 4: Service Model of Cloud Computing… 11
    5. Figure 5: Deployment of Cloud Computing… 12
    6. Figure 6: OLTP and OLAP comparison 15
    7. Figure 7: OLAP Architecture 17

### ABSTRACT

Data warehousing systems are the skeletons of decision making process in any large business based organizations. These systems provide a base platform for decision making in the current global world (Asrani, Jain, 2015).When ever there arises a question of making a feasible decision, from the data existing, we bring database, data warehouse and on-line analytical processing (OLAP) into picture. In todays scenario, we find every small scale industry, let’s say a commercial service, is using databases in their functioning. In such an industry, business men will have to analyze the customer behavior and observe the trends or patterns in their purchase history. I in this research work would like to make some discussions on OLAP operations and its requirements. OLAP and OLTP applications typically eliminate manual / clerical data processing tasks such as entry and banking transactions that are day-to-day operations of/(in) an organization. We require some tools to perform different operations as part of knowledge mining. In the mean while, we will also have to mind the efficiency in processing the query. This paper is based on few other presenters view points, taken as reference.

### KEYWORDS

Central database, ETL (extract, transform, load) tools, Metadata, Access tools, Data warehousing, Decision making, OLAP, Data mining

### INTRODUCTION

Data warehousing (DW) is the collection of data from various sources that is integrated into one repository (Troy, 2022). Operating on a data warehouse describes the process of designing, maintaining, modeling and presenting the knowledgeable data information after going through proper cleaning and cleansing operations being perform on the data. Thus obtained piece of information should maintain the ability to solve a complex business problem of the service provider or the organization. As [data sets](https://en.wikipedia.org/wiki/Data_set) have grown in size and complexity, direct data analysis has increasingly been augmented with indirect, and is made to automate data processing, being aided by discoveries by other means in computer science, specially in the field of machine learning, such as [neural networks](https://en.wikipedia.org/wiki/Neural_networks), [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis), [genetic algorithms](https://en.wikipedia.org/wiki/Genetic_algorithms), [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) and [decision rules](https://en.wikipedia.org/wiki/Decision_rules) (1960s), and [support vector machines](https://en.wikipedia.org/wiki/Support_vector_machines) (1990s) (Mehmed, 2003).Systems present at the warehouse are capable to process a wide variety of operations, starting from the simple ones to the data mining requirements where in OLAP operations are also one of them. System store the business information since ages and comes up with a large multi-dimensioned tables forming a data warehouse. Having such huge hunks of data makes the query processing slow. Thus there is a high requirement of pre-analyzed data, which can be stored in a repository to save the run time. We have many techniques available in which materialized view maintenance is mainly used as it contains some pre-defined aggregate functions. This saves a lot of server

load, and processing is efficient. Considering operational systems and data warehousing systems, both of them have a large variation. Operational systems are maintained for the purpose of system operations and data warehousing systems support in providing decision based analysis from the data taken from the company. Even the design of these two systems is entirely different. Systems for operations are designed to receive single atomic type data of a particular business transaction. Whereas data warehousing system is manufactured to handle a non atomic data, and also a minor changes made as per the company requirement. Here, we have the sequence in mining knowledge from a given instance of data.

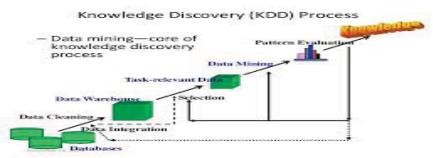


Figure 1: Data Warehousing and its Components (Asrani, Jani, 2015)

|  |  |
| --- | --- |
| **Major Category** | **Sub Category** |
| Modeling & Design | Conceptual / Logical / Physical Design |
| Extract Transfer Load Process Design |
| Complete Architecture Design |
| Standards for Modeling and Design |

|  |  |
| --- | --- |
| Technologies to access Data Warehouse | Query Processing |
| OLAP Tools and Techniques |
| Reporting Tools |
| Data Mining Operations |
| Data Warehouse Maintenance | Materialized View Maintenance |
| Meta Data Management |
| Data Security |
| Indexing |

Table-1 Key categories and research areas in the field of data warehousing (Asrani, Jain,2015)

The table above shows the key areas of research in the fields of data warehouse security, quality and other ETL and its visualization actions.

### LITERATURE REVIEW

Modeling and design of a data warehouse refers to analysis and diagrammatic representation of data storage structure (Asrani, Jain, 2015). To design and develop a new database, we will have to analyze the representation pattern and the storage structure. Flow of transfer of data from an operational system to a data driven systems should also be implemented and architecture needs to be analyzed. Relational data base is the traditional format to store and retrieve the data but it can also be made used to develop a model for data warehouses. All these conversions and schemas for these data model is completely dependent on the design standards of the databases let say by the NIST Protocols. It can also be known as the Dimensional Model.

A warehouse is a place which gives the users a decision making support with all the technology it has, which is mainly with the intension of improvising the work of the knowledge workers like executives and managers. These days, industries have been moving towards exponential growth rates in terms of product count and service delivery by adopting these newer technologies that

are available in the market. According to a survey by META group, market share for data warehouse is going to face a rise from $1.5 billion to $9 billion in a span of 3-4 years down the lane. Today, these techniques have been successfully initiated in almost every industry to support and retain their service and customer base. Industries are trying to maintain a user profile and management system for inventory and financial services being offered and are also playing an active role in the detection of fraud. Some other techniques in Transport and Telecommunications, Health department can be considered as an addition to them. Our paper, at present likes to give out a roadmap for various data base technologies with a main focus on OLAP operations and other functionalities of Data Mining.

### Architecture and End-to-End Process

Architecture is something which consists of tools and mining techniques to be acting on a piece of data or any other external resources to clean and transform as a process of bringing out frequent patterns. We go through many steps internally, to achieve a perfect usable information from a huge junk of the warehouse. The actions could go like, to clean, to modify(as required), integrate the data, and finally load and keep refreshing periodically for upgrading the database from the source. This is because, a mart may have multiple stores at various locations, to obtain and understand the working of sales in each store and in each department, data operations are well used. While trying to obtain data from such a scenario, we come up with multi-dimensional view points at front end and reports multiple analysis tools.

The concept or the diagrammatic coverage of these tools and functions is as shown in the below figure.

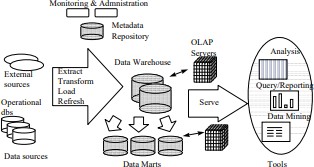


Figure 2. Data Warehousing Architecture (Salley, 2012)

At last, when we come up with a data repository which facilitates us to store and manage data along with its metadata and also the tools of administration of the warehouse, this warehouse may be as a collection of multiple locations and will be accessed for load balancing, study of its scalability and availability. Having a distributed architecture of such type, metadata, the explanatory data of some other data, is replicated with other warehouse and central system for administration. It is also an inclusive of huge expense to the database owner or the organization to design and develop an entire single unit warehouse with appropriate logics to it. It also requires to come up with a own data repository and an administration unit as a whole. We have few pre - defined functions that needs to be performed at any data warehouse.

★ First, a necessity of define and design an architecture, following a plan for the same

★ Identify the servers and their capacities to store and perform OLAP operations on the data defined.

★ Try to combine and integrate the servers.

★ Develop an end - to - end schema and other required structures for the data warehouse.

★ Obtain views for the databases.

★ Define the gateways, connectors and wrap drivers at each server machines.

★ Identify the methods and various functionalities to be performed through out the warehouse at the time of mining.

★ Check the structure to adjust with the organization and placing of the data and its access at retrieval times.

★ Give a proper mechanism and connection to the end users for querying the data as per his/her own needs.

★ Understand how to implement the query and retrieval process.

★ Develop a complete script for data mining operations which include extraction and cleansing followed by transform and load.

★ Provide a conceptual space for metadata where ever appropriate.

★ Try to roll the warehouse for applications access.

### Working of Conceptual Model

A model which can be generalized as the representation of the entire database or the data warehouse which tries to comes up with a diagrammatic representation of the data model is said to be a conceptual data model. Formally known as Entity Relation Model is designed here to identify the characters and its behavior at various times in the databases. A model of such kind which influences the front end design and tries to perform OLAP operations is the view in multi- dimensional methodology in a warehouse. In this multi dimensional view model, user is supplied with a numeric content which are helpful in analyzing the objects of the databases. We can identify sales and have a track of the budget and revenue, profit on investment etc. Dimensions are equipped at each measurable units, provides context of measure. One of the generalized example could be to analyze the population of a country, we can identify and get the data of population in the country as a whole or the user is given access to to dig a particular piece of data for information, say the population of one particular state or city or county and street to the minimum could be access. That is how a multi dimensioned view point tries to provide the necessary details to the end user. Another example could gland retrieve the data with the sales in a particular city and the total purchases, product names and date of purchase. Measurement of each dimension is unique. Every dimension is supplied with an attribute set. For example, a dimension named PRODUCT holds the attributes like category, year of manufacture, average profit, etc. Attributes are always put the order of their hierarchy of relationship.

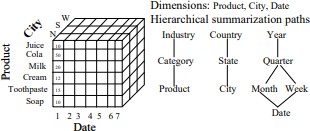


Figure 3. Multidimensional data (Chegg.com)

Another key operations on the dimensional conceptual model of OLAP is stress on aggregation, what that means is that the system computes and issues a ranking system to the total sales column (for each country) for example. And this happens for each year. Other major operations include to compare any given two measurements, let it be like sales and budget which are further aggregated by a similar dimensions. OLAP can also be applied on time series to observe and make note of trends in various industries. It is also necessary to have a minimal knowledge on working of calendars to inspect dimensions of time. Previous a kind of analysis was purely being carried out on spreed sheets programs which then were moved to dimensional models design and usage. Today’s generation also has a minor focus on spreed sheets for carrying out OLAP operations. Sometimes, few industries also make use of excel sheets for handling giga bytes of data for more efficient operations and results. An example of this could be Arbor Corporation, its product named Essbase is developed using excel sheet for data processing of front end design.

OLAP is still equipped with other two major functions known as roll-up and drill-down. The basic working of these operations are like, roll-up makes a particular dimension to an upper level of abstraction and drill-down is something which tries to dig deeper into the given dimension. These operations are generalized as pivoting functions. Roll-up on sales is to aggregate on city by product. Its vice-versa becomes drill-down. Now, in order to have a projection of the data examined, we go with Slice() and Dice() functions, which helps in reducing the data dimension. A Slice() operations is to obtain a smaller piece of data from the original database. Suppose, we can obtain a sales information of a particular branch/ location for a particular day of business. We try to get a subset of the dimension selecting a minimal attribute values. Other functions to check with the weight of attributes on the data include ranking, sorting, selecting, defining and computing attributes as required by the end user.

### Utilities and Tools for Back-end

Data mining is an entirely different set of operational functioning on various set pieces of data from a hunk of information. Each function is basically abstracted with other functionalities. Warehouse systems use extraction tools, cleaning tools and then move on with ETL tools to popular the data warehouse. Gateways are used to call out data from various foreign resources

and are implemented at system interfaces using ODBC and Expertise Connect or Informix Enterprise Gateway. Cleaning of data present in the data warehouse is the first step to compute a proper decision from the data model for the desired business requirement. It should also be verified for the correctness of the data in the warehouse. It is also obvious that a data warehouse is complied with a lot of erroneous data as the data is collected from different sources and consists of large volumes in multiple areas.

Data extraction from “foreign” sources is usually implemented via gateways and standard interfaces (such as Information Builders EDA/SQL, ODBC, Oracle Open Connect, Sybase Enterprise Connect, Informix Enterprise Gateway).An example of where the data cleaning is required is, inconsistent field lengths, containing NULL values or missing data and if violated with constraints on integrity. Sources even provide data with inconsistency, designer will have to work on such data at times of cleaning. Followed by cleaning, data will have to be scrubbed for domain identifiable knowledge. Few softwares support for cleaning such as Integrity, Trillium.

#### Load

Data is now subject to load into the data warehouse, after going through the process of extract, clean and transform. Data still undergoes few steps of preprocessing such as sorting, and other summarization computation to construct a well structured data warehouse. Responsibilities also include to indexing and deriving access paths, partitioning for multiple storages before moving towards load operation. Load should also let the user to administrate the status and revoke access at times of failure.

#### Refresh

It is highly necessary to refresh the data warehouse to allow space for timely updates. Sometimes updates might be on both base data and derived data. Refresh has two different issues to be considered, how to refresh and when to refresh. Any warehouse is refreshed periodically, say every week or bi-weekly. A refreshing authority is taken care by warehouse administrator, and he makes changes according to user requirements and data traffic towards

the warehouse. For data warehousing, the most significant derived data are summary tables, single-table index and join index (IJISRT, 2021).

### Database Design Methodology

Each of the numeric measures depends on a set of *dimensions,* which provide the context for the measure (IJISRT, 2021). A MOLAP server implements the multi-dimensional data model. To make use of a ROLAP operations, relations has to be mapped. The techniques of OLTP environments is the ER Diagram and the Normalization. Querying is much important than ER Diagram for the purpose of decision making. Star schema is much used to represent the multi dimensional model. Dataset consists of one particular table for each dimension. Every tuple has a fact table associated to it.

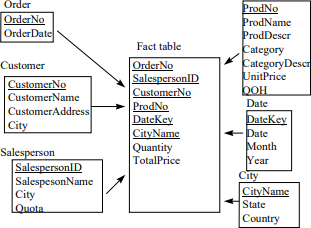


Figure 4. A Star Schema (Chegg.com)

Hierarchy attributes are not provided by star schema. A refinement to star schema is the snowflake schema. It would be appropriate to browse denormalized structure for a dimensional star schema. A sharing of dimensional tables occurs in a fact constellations which are an instance of more complex structures. As an example, actual vs estimated expense could form a fact constellation as they share many dimensions (IJISRT, 2021).

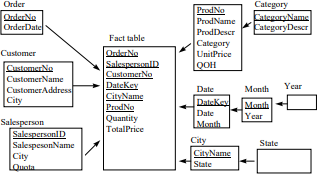


Figure 5. A Snowflake Schema (Chegg.com)

Data warehouses also hold a pre computed tables which hold the data of dimension tables and selected summary. This pre computed data refers to the aggregate fact table with 1 or more selected dimensions. This pre computed data can be presented in at least 2 ways. As an alternative, summary table can be encoded to aggregate tables into fact tables without manipulating other tables (IJISRT, 2021).

### 3 DISCUSSIONS

OLTP is short for Online Transaction Processing. The system provides data to a [dedicated storage](https://phoenixnap.com/servers/storage) [server](https://phoenixnap.com/servers/storage) directly from the source (Phoenixnap Solutions, 2021 ).

The main characteristics of OLTP are

* Frequent query processing
* Fast transactions
* Data integrity

#### OLTP Use Cases

OLTP is being put into use by most of the consumer based systems. Few examples could be,

* + 1. Online Appointment booking : A reservation or ticket system uses OLTP and its methodology.
    2. Process a pay : Any debit or credit card transactions are taken care by OLTP processes.
    3. ATM and Banking : Simple daily transactions are being supported by OLTP resources.
    4. Generating a record : OLTP best suites for record generating and storing.

OLAP is short for Online Analytical Processing. The method takes data collected by an OLTP system and prepares it for analytical purposes (Phoenixnap, 2021).

The main characteristics of an OLAP system are:

* Smaller query volume.
* Complex transactions.
* Query speed.

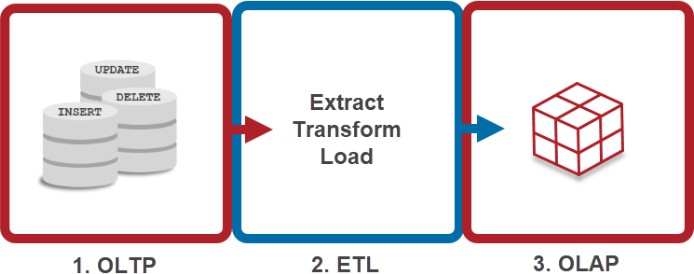


Figure 6. OLTP and OLAP comparison (Phoenixnap, 2021)

#### OLAP Use Cases

OLAP works for Data Analytics and data processing. Some applications can be,

* + 1. Behavior of customer : Identifies the working of decisions by customers at different geographical locations and supplies a business rise to e-commerce organizations.
    2. Rural business development : Recent study shows that OLAP with its edge computing process can develop agricultural practices in rural areas.
    3. Trend study : Brings out the statistical data from a pre existing data at the warehouse.

#### Advantages and Disadvantages of OLTP

OLTP can be described as a system working towards huge transaction count with a quick response. OLTP is a transaction process which has its own pros and cons.

#### Some advantages of using OLTP are

* + 1. Atomicity : This is a feature in which entire transaction occurs or doesn’t occur at all.
    2. Speed : Transactions are put simple. Each transaction occurs with the same velocity with no disturbance to other.
    3. Concurrency : Two or more different transactions take place at the same time frame.

#### The disadvantages of OLTP include

1. Security : OLTP has to become vulnerable in terms of treatment towards data while there occurs a high inflow of transactions.
2. Volume : A number of transaction requests make the system overwhelming. This needs to be handled by a qualified team.
3. Downtime : A time where system is down causing a bottleneck, and high available solutions must be in usage.

#### Advantages - Disadvantages of OLAP

OLAP is a procedure followed on data knowledge discovery and works much on multi- dimensional data cubicles. Analysis comes with both advantages and disadvantages.

#### Overall advantages of OLAP system are

* + 1. Decision-making support : Star and snowflake schemas help OLAP system in providing flexibility to come up with a best decision over the data.
    2. Comprehension Problem : If the query is found complex, a multidimensional data provides a complete overview of data from myriad databases.
    3. Flat learning curve : There is no must necessity on learning technical aspects on OLAP systems for the end users.

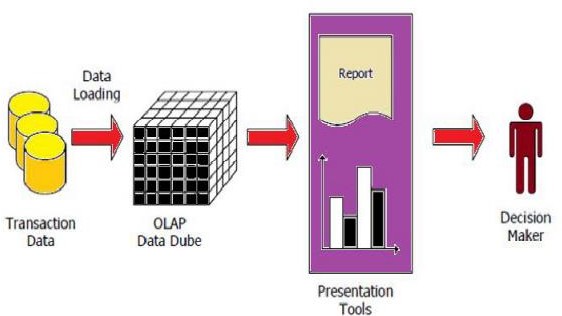


Figure 7. OLAP Architecture (AJER, 2016 )

#### Weaknesses in OLAP systems are

1. Redundancy : The process of denormalization bring a lot a redundant copies.
2. Computation : There comes a requirement to use a third party applications and tools to compute few complex transactions using OLAP. As non technical employee work, it becomes difficult for understanding the functionality.
3. Scalability : OLAP systems fail to give access to scalability function while the data size proceeds to increase.

### CONCLUSION and IMPLICATIONS

We have discussed the technical aspects of developing and designing the decision making systems. Though we are here with myriad services and deployable products, we still have scope for research in this field. This paper has emphasis on the inconsistency of the data rather than the schema itself. Data cleaning is one step to overcome this. To design a physical model, interest should be expressed on selection of index, partitioning data and its selection appropriately. It is also necessary to check on the aggregation at such times. The management of data warehouses also presents new challenges. Detecting runaway queries, and managing and scheduling resources are problems that are important but have not been well solved (Qassim, 2012 ).

A support system for decision making provides optimization of queries with an increase in challenge of traditional estimations and costs without exploding the search space. We have a lot of options for transformations, a reliable technique is to be chosen. ROLAP layer, that is the layer between middleware and partitioning function at the back end server operations is also an important area of research interest.

### REFERENCES

1. Codd F, Codd S.B, Salley C.T. *Technical Report.* San Jose, Calif: Codd EF & Associates; 1993. *Providing OLAP (Online Analytical Processing) to User-Analysts: An IT Mandate*.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1184053/

1. International Journal of Managing Information Technology (IJMIT) Vol.4, No.1, February 2012 DOI : 10.5121/ijmit.2012.4103 25 Apply On-Line Analytical Processing (OLAP)With Data Mining For Clinical Decision Support Dr Walid Qassim Qwaider Majmaah University College of Science and Humanities in Ghat Management Information Systems Dept. Saudi Arabia

https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3673889

1. Business Intelligence : How Amazon Uses Data Warehouses and Business Intelligence by Troy, January 31, 2022

https:/[/www.ictsd.org/how-amazon-uses-data-warehouse-and-business-intelligence/](http://www.ictsd.org/how-amazon-uses-data-warehouse-and-business-intelligence/)

1. International Journal of Innovative Science and Research Technology, IJISRT A digital library, ISSN No: 2456-2165 Paper Submission on – January 2022.

[https://ijisrt.com/wp-content/uploads/2017/07/A-Comprehensive-Review-on-OLAP- Models-and-operations-1.pdf](https://ijisrt.com/wp-content/uploads/2017/07/A-Comprehensive-Review-on-OLAP-%20Models-and-operations-1.pdf)

1. OLTP vs OLAP : A comprehensive Comparison, June 8, 2021

<https://phoenixnap.com/kb/oltp-vs-olap>

1. Qwaider, Walid Qassim, Apply On-Line Analytical Processing (OLAP) With Data Mining For Clinical Decision Support (2012). International Journal of Managing Information Technology (IJMIT) Vol.4, No.1, February 2012, Available at SSRN: <https://ssrn.com/abstract=3673889>

<https://deliverypdf.ssrn.com/delivery.php?ID=394013103113028031123030103122078>

[10905006305006807906900711500500907102110702012200903410209802611005906](https://deliverypdf.ssrn.com/delivery.php?ID=394013103113028031123030103122078109050063050068079069007115005009071021107020122009034102098026110059062007088122092008014072045007060077040003020097086127115118070000035046003082001008029004080001084095098123083072028006012074027125029064026022066102&EXT=pdf&INDEX=TRUE)

[20070881220920080140720450070600770400030200970861271151180700000350460](https://deliverypdf.ssrn.com/delivery.php?ID=394013103113028031123030103122078109050063050068079069007115005009071021107020122009034102098026110059062007088122092008014072045007060077040003020097086127115118070000035046003082001008029004080001084095098123083072028006012074027125029064026022066102&EXT=pdf&INDEX=TRUE)

[03082001008029004080001084095098123083072028006012074027125029064026022](https://deliverypdf.ssrn.com/delivery.php?ID=394013103113028031123030103122078109050063050068079069007115005009071021107020122009034102098026110059062007088122092008014072045007060077040003020097086127115118070000035046003082001008029004080001084095098123083072028006012074027125029064026022066102&EXT=pdf&INDEX=TRUE)

[066102&EXT=pdf&INDEX=TRUE](https://deliverypdf.ssrn.com/delivery.php?ID=394013103113028031123030103122078109050063050068079069007115005009071021107020122009034102098026110059062007088122092008014072045007060077040003020097086127115118070000035046003082001008029004080001084095098123083072028006012074027125029064026022066102&EXT=pdf&INDEX=TRUE)

1. International Journal of Innovations in Engineering and Technology (IJIET) Data Warehousing Methods and Techniques: A Survey of Literature to Identify Current Research Needs Deepak Asrani, Renu Jain Department of Computer Science Engineering, University Institute of Engineering & Technology, Kanpur, U.P, India

<http://ijiet.com/wp-content/uploads/2015/12/35.pdf?cv=1>

1. Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 978-0-471-22852-3. OCLC 50055336.

<https://en.wikipedia.org/wiki/Data_mining>

## My DB Research Paper (1).docx

ORIGINALITY REPORT

13

## %

SIMILARITY INDEX

11%

INTERNET SOURCES

5%

PUBLICATIONS

7%

STUDENT PAPERS

PRIMARY SOURCES

# Submitted to CSU, Fullerton

1

Student Paper

# ijiet.com

2

Internet Source

# ijirt.org

3

Internet Source

# phoenixnap.com

4

Internet Source

# en.m.wikipedia.org

5

Internet Source

# [www.rimtengg.com](http://www.rimtengg.com/)

6

Internet Source

# [www.ictsd.org](http://www.ictsd.org/)

7

Internet Source

# Submitted to Lamar High School

8

Student Paper

# Submitted to Intercollege

9

Student Paper

2%

2%

2%

2%

1%

1%

1%

1%

<1%

Exclude quotes On Exclude bibliography On

Exclude matches < 10 words