

**Literature Review**

This proposed research is inspired by the work of Mr Jagdish Sadhave 2009, who used the technology of Data Mining and Knowledge Discovery in Databases to build Examination Data Warehouse (EDW) [1]. This EDW can effectively be used for mining and answering some critical questions regarding examination results of University of Mumbai. This research work done can be very effectively extended by adding more dimensions to Examination Data Warehouse and making it capable for extracting more useful information regarding all Universities in India. This information can be very useful for Central Education Departments like Ministry of Human Resource and Development (MHRD) or equivalent Ministries at various state levels for decision making.

Alejandro Gutierrez and Adriana Marotta 2000, a data ware house is a structure that is optimized for distribution. It collects and integrates the sets of historical data from multiple operational systems and feed them to one or more data marts. It may also provide end-user access to support enterprise views of data. Knowledge specialists carefully design data warehouses, where data is extracted from these operational databases and centrally recorded. The key idea is to make available to the management the critical information that can be used for further analytical processing and decision-making.

Adviaans and Zantinge 1996, Data Warehouse is design for strategic support and is largely built up from databases that make up the operational database. Small local data warehouse are called data marts which are configured to generate reports for specific set of
users. The basic structure of data warehouse must be dependent, non-volatile, subject oriented and integrated.

Chaudhari and Dayal 1997, a separate data warehouse is needed for special data organisation access methods and separate implementation methods are needed to support multidimensional views and various queries. A data warehouse is designed especially as the Decision Supporting System (DSS). Therefore, only the data that is needed for decision support is extracted from the operational data and stored in warehouse is extracted from the operational data stored in the warehouse. The data warehouse is a central store of data that has been extracted from the operational data. The data store in a warehouse is subject oriented and of historic bearing. Thus, data warehouses tend to contain extremely large data sets. Separate implementation methods are needed to support multidimensional views and various queries.

Wiess and Indurkhya 1998, the data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. Matching relevant data and structuring the organization so that it can actually utilize the new knowledge transformation is a fascinating process. The key issue in KDD is to realize that there is more information hidden in our data than we are able to distinguish at first sight.

Westphal and Blaxton 2007, data mining is most useful in exploratory analysis scenarios where there are no predetermined notions about what will constitute an “interesting”
outcome. As knowledge is not measured by volume, but its content and experience, the timely human interaction is very much essential and in our experience has proved to be useful in KDD process. Though KDD process requires human interaction, it is very difficult to define the degree of level of human interaction.

Barquin, R., and H. Edelstein 1997, In this paper data mining is a very strong method for analyzing the large data to generate accurate future decisions. Data mining method prefer the classical sample and case model of data but have difficult reasoning with time and its increased dimensions.

Inmon 2001, many people treat data mining as a synonym for another popularly used term, "Knowledge Discovery in Databases", or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. The Data Mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. Note that according to this view, Data Mining is only one step in the entire process, albeit an essential one since it uncovers hidden patterns for evaluation. We agree that Data Mining is a knowledge discovery process. However, in industry, in media, and in the database research milieu, the term "Data Mining" is becoming more popular than the longer term of "knowledge discovery in databases".
Chuck Ballard, Dirk Herreman, Don Schau, Rhonda Bell 2009, This would typically require modification of the extract programs or development of new ones. This process is costly, inefficient, and very time consuming. Data warehousing offers a better approach.

Jiawei Han 1991, Data mining, or knowledge discovery in databases, has been popularly recognized as an important research issue with broad applications. We provide a comprehensive survey, in database perspective, on the data mining techniques developed recently. Several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization, and meta-rule guided mining, will be reviewed.

Paulraj Ponniah In the early 1970s, some major corporations created what were called information centers. The information center typically was a place where users could go to request ad hoc reports or view special information on screens. These were predetermined reports or screens. IT personnel were present at these information centers to help the users to obtain the desired information.

Micheline Kamber 2004 Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases.
A popular trend in the information industry is to perform data cleaning and data integration as a preprocessing step where the resulting data are stored in a data warehouse. Sometimes data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing.

A. Silberschatz, M. Stonebraker, and J. D. Ullman 1996, a set of records that refer to the same entity can be interpreted in two ways. One way is to view one of the records as correct and the other records as duplicates containing erroneous information. The task then is to cleanse the database of the duplicate records. Another interpretation is to consider each matching record as a partial source of information. The aim is then to merge the duplicate records, yielding one record with more complete information. The system described here gives a solution to the detection of approximately duplicate records only.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth 1996, across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD).

V. Ganti, J. Gehrke, R. Ramakrishnan 1999, Data mining, also known as knowledge discovery in databases, gives organizations the tools to sift through these vast data stores to find the trends, patterns, and correlations that can guide strategic decision making. Traditionally, algorithms for data analysis assume that the input data contains relatively few
records. Current databases however, are much too large to be held in main memory. To be efficient, the data mining techniques applied to very large databases must be highly scalable.

M. S. Chen, J. Han, and P. S. Yu 1996, Mining information and knowledge from large databases has been recognized by many researchers as a key research topic in database systems and machine learning and by many industrial companies as an important area with an opportunity of major revenues.

K. Han 1996, mining data streams with concept drifts for actionable insights has become an important and challenging task for a wide range of applications including credit card fraud protection, target marketing, network intrusion detection, etc. Conventional knowledge discovery tools are facing two challenges, the overwhelming volume of the streaming data, and the concept drifts.

U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy 1999, understanding data mining and model induction at this component level clarifies the behavior of any data-mining algorithm and makes it easier for the user to understand its overall contribution and applicability to the KDD process.

J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh 1999, to build the d-dimensional data cube, for on-line analytical processing, in the relational algebra, the database programming language must support a loop of d steps. Each step of the loop involves a different attribute of the data relation being cubed, so the
language must support attribute metadata. A set of attribute names is a relation on the new data type, attribute. It can be used in projection lists and in other syntactical positions requiring sets of attributes. It can also be used in nested relations, and the transpose operator is a handy way to create such nested metadata.

E. Thomsen 1997, OLAP enables users to access information from multidimensional data warehouses almost instantly, to view information in any way they like, and to cleanly specify and carry out sophisticated calculations. Although many commercial OLAP tools and products are now available, OLAP is still a difficult and complex technology to master.

R. Kimball 1997, Dimensional modeling has become the most widely accepted approach for data warehouse design. A DWH is a data structure that is optimized for distribution. It collects and integrated sets of historical data from multiple operational systems and feed them to one or more data marts. It may also provide end-user access to support enterprise views of data. Knowledge specialists carefully design data warehouses, where data is extracted from these operational databases and centrally recorded. The key idea is to make available to the management the critical information that can be used for further analytical processing and decision-making.

V. Harinarayan, A. Rajaraman, and J. D. Ullman 1996, decision support applications involve complex queries on very large databases. Since response times should be small, query optimization is critical. Users typically view the data as multidimensional data cubes. Each cell of the data cube is a view consisting of an aggregation of interest, like total sales.
The values of many of these cells are dependent on the values of other cells in the data cube. A common and powerful query optimization technique is to materialize some or all of these cells rather than compute them from raw data each time. Commercial systems differ mainly in their approach to materializing the data cube. In this paper, we investigate the issue of which cells (views) to materialize when it is too expensive to materialize all views. A lattice framework is used to express dependencies among views. We present greedy algorithms that work off this lattice and determine a good set of views to materialize. The greedy algorithm performs within a small constant factor of optimal under a variety of models. We then consider the most common cause of the hypercube lattice and examine the choice of materialized views for hypercube in detail, giving some good tradeoffs between the space used and the average time to answer a query.

S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi 2004, at the heart of OLAP or multidimensional data analysis applications is the ability to simultaneously aggregate across many sets of dimensions. Computing multidimensional aggregates is a performance bottleneck for these applications.

R. Agrawal, A. Gupta, and S. Sarawagi in 1996 have proposed a data model and a few algebraic operations that provide semantic foundation to multidimensional databases. The distinguishing feature of the proposed model is the symmetric treatment not only of all dimensions but also measures. The model provides support for multiple hierarchies along each dimension and support for adhoc aggregates. The proposed operators are composable, re-orderable, and closed in application. These operators are also minimal in the sense that
none can be expressed in terms of others nor can anyone be dropped without sacrificing functionality. They make possible the declarative specification and optimization of multidimensional database queries that are currently specified operationally. The operators have been designed to be translated to SQL and can be implemented either on top of a relational database system or within a special purpose multidimensional database engine. In effect, they provide an algebraic application programming interface (API) that allows the separation of the frontend from the backend. Finally, the proposed model provides a framework in which to study multidimensional databases and opens several new research problems.