Chapter 2

Literature Review

This chapter aims to present literature survey in the area of multimedia data warehouse. Literature survey presents a study about various models, architectures and frameworks proposed for multimedia data warehouse, complex data warehouse and XML data warehouse. Along with the data warehouse systems, the face image data representation & retrieval, geographic image representation & retrieval and e-learning video data representation retrieval and analysis have also been studied. For the purpose of literature review a set of eighty seven research papers have been studied with rigorous study focus of forty five research paper and they are sited in literature survey reviews.

2.1 Literature Review

O. Zaane et al. [65] proposed a system prototype named MultiMediaMiner, for multimedia data mining. To facilitate the multidimensional analysis of large multimedia databases, multimedia data cubes can be designed and constructed in a manner similar to that for traditional data cubes from relational data. A multimedia data cube contains additional dimensions and measures for multimedia information, based on visual content such as color, texture, and shape. It provides the mining of multiple kinds of knowledge, including summarization, comparison, classification, association, and clustering.

J. Han et al. [38] proposed the DBMiner system applies multi-dimensional database structures, attribute-oriented induction, multi-level association analysis, statistical data analysis, and machine learning approaches for mining different kinds of rules in relational databases and data warehouses.

W Lee et al. [81] introduced multimedia data warehouse for EOD (Education On-Demand) systems that have meta-data descriptions independent of the EOD data contents. They propose this system against the database system. This system is implemented with two parts: the keyword extraction module, and the data retrieval
module. Video data is first segmented, after segmentation they process for keyword extraction. Keyword extraction is processed by deriving the importance of each word over a shot description. After the keyword extraction process shot is compressed and stored in physical data warehouse. They have used fact and dimension to store these data and have used the concept of indexing.

You et al. [87] proposed a data warehousing approach on hierarchical multimedia information retrieval. For multimedia data representation, storage, integration, indexing, similarity measures, searching methods and query processing, the proposed algorithm allows: (1) to extend the concepts of conventional data warehouse and multimedia databases to multimedia data warehouses for effective data representation and storage; (2) to develop a multimedia starflake schema to integrate multiple data streams for hierarchical data representation and indexing; (3) to apply data aggregation techniques for decision support to speed up query processing and searching. In addition, the new system architecture is compared with a conventional database structure. They presented a case study to illustrate the development of a content-based image retrieval system for cyclone pattern recognition.

X. Cao et al. [83] described a web-based data warehousing method for retrieving and analyzing neurological multimedia information. They build three-tier multimedia data warehouse. The data warehouse integrates clinical multimedia data related to epilepsy from disparate sources and archives them into a data model.

S. T. C. Wong et al. [75] proposed multi-tier image data warehouse framework and the corresponding software development process for supporting neuroimaging research and large-scale data analysis. This modular information framework is based on the OOAD and component-based development paradigm. They have implemented CORBA and web based architecture that separates the graphical user interface presentation, data warehouse business services, data staging area, and back end source systems into distinct software layers.

B. Milosavljevic et al. [16] proposed a design of a multimedia information retrieval system with extensibility providing integration of existing multimedia retrieval software modules and choice among different retrieval models. The system manages
multiple classes of documents, with each class being defined by an XML schema document. The documents themselves are stored in XML format according to the given schema. Design of the main subsystems - document storage, indexing, and retrieval - is presented, as well as an example of a retrieval model.

W Hummer et al. [80] proposed a family of XML document templates, called XCube, to describe a multidimensional structure, dimensions and fact data for integrating several data warehouses into a virtual or federated data warehouse. Aim to propose XCube is to reduce the amount of data transferred over the network.

H H Kim et al. [35] proposed a framework for multimedia data warehouse, which integrates diverse types of multimedia data from disparate sources. Because of the interoperability and flexibility of the XML, they use the XML technology for their framework. The multimedia data warehouse supports content-based integration and retrieval of multimedia data, and manages changes of data sources in a distributed environment.

J. Darmont et al. [37] proposed a modeling process for integrating heterogeneous data into a data warehouse. They have provided conceptual level, logical level and physical level models. At the conceptual level, a complex object is represented in UML. Their logical model is an XML schema that can be described with a DTD or the XML-Schema language. They have designed a Java prototype that transforms multiform input data into XML documents representing their physical model. The obtained XML documents are mapped into relational databases. They have re-modeled these data in a multidimensional way to allow their storage in a warehouse for analysis.

T Cerquitelli et al. [76] provided mechanisms for analyzing multimedia data from heterogeneous and distributed sources in a single repository. They proposed a mediated query service for distributed multimedia data, adapted to build multimedia data warehouses. They have configured individual wrappers according to the source needs to map between relational model used in the source and XML-exported schema. They have configured adapter that implements associated rules, based in the first order logic, to solve the mapping between the data warehouse schema and the
application schema. The proposed approach provides materialized views to use in the analysis of multimedia data.

L I Rusu et al. [48] established a framework for building an XML data warehouse. They focused on identifying a systematic approach for building a data warehouse of XML documents, specifically for transferring data from an underlying XML database into a defined XML data warehouse. The proposed methodology on building XML data warehouses covers processes such as data cleaning and integration, more detailed analysis of data, summarization, intermediate XML documents, and updating/linking existing documents and creating fact tables. They utilize the XQuery technology for all the processes.

A M Arigon et al. [7] define a conceptual model and proposed multidimensional model in which they provide content based and descriptor based descriptors in separate dimension. They implement a multimedia data warehouse in the medical field by integrating the multimedia data into a multidimensional model. They present prototype for this study.

L I Rusu et al. [47] proposed methodology on building XML data warehouse. They include data cleaning and integration, summarization, intermediate XML documents and updating & linking existing documents and creating fact tables. For all the processes they use XQuery.

B K Park et al. [17] proposed a framework for multidimensional analysis of XML documents, called XML-OLAP. In XML-OLAP data for fact and dimension are stored as XML documents. They build XML cube named XQ-Cube for XML warehouses. An XQ-Cube is constructed for the measure data specified by an XQuery expression. They used text mining operations for the aggregation of text measure data. They proposed XML-MDX as a new multidimensional expression language for XQ-Cubes.

Wen-Yang Lin and Chin-Ang Wu [78] proposed a data model for the heterogeneous data warehouse. This model resembles to the star schema to inherit multidimensional aspects and incorporates features of object relational database to meet the
requirements of integrating heterogeneous types of data. They present an object-relational data model for clinical data warehouse.

Wen-Yang Lin et al. [79] proposed a data model for the complex data warehouse. This model incorporates star schema with object-relational concepts, inheriting multidimensional characteristic and providing mechanisms to integrate complex data in a data warehouse model. In fact tables of the proposed model, measures are provided for complex data which users are interested in analyzing with. Many-to-many relationships between a fact table and some dimensions can be modeled as degenerated dimensions via complex objects, and simple aggregation of whole-part relationships are modeled as traditional relational tables while generalization-specialization are drawn with type inheritance from super-class to sub-class.

A Arigon et al. [8] proposed a new multidimensional model that integrates functional dimension versions allowing the descriptors of the multidimensional data to be computed by different functions. These descriptors are stored in dimensions. This model is used to develop an OLAP application for navigation into a hypercube integrating various functional dimension versions for the calculation of descriptors in a medical use case.

A Vanea and Rodica P [5] proposed an architecture for a multimedia data warehouse. They acquire data to prepare it be stored in XML files. They have used semantics concept. For semantic data, they use two repositories of metadata which describes data in hierarchical manner in XML files. One repository describes the terms specific to the business domain and the other describes the technical terms that the system can extract and process. They have presented a data model for representing facts and dimensions according to the hierarchical structure of entities captured in multimedia objects. For aggregation of data and to manipulate metadata they have provided separate processing and maintenance block. The query processor is connected to data warehouse, metadata and processing and maintenance block to resolves the semantics of the query that the user inputs. For each fact level of each data mart is stored in a different XML file; this creates hierarchical structure of the data marts, in which a fact table may become a dimension for another fact table. Such a fact file contains the result of the aggregation and at least one set of references for each dimension used in
the aggregation process. To speed up the process they have developed a procedure to automatically creating new queries, an XML file is created, containing technical terms associated with every existing fact table.

O. Boussaid et al. [62] provided an approach to warehouse complex data. They presented a generic UML model that helps to model low level as well as semantic information concerning the complex data to be analyzed. Complex data are then stored in XML documents generated by their software. For complex data, dimensional modeling has been used. They use data mining techniques to enrich the knowledge of designers about the stored complex data and help them build better adapted data warehouses. The limitation of their approach is that they have used decision trees and rules they obtained are not all relevant. Even external techniques are necessary to complement the description of complex data significantly.

O Boussaid et al. [64] proposed an XML based methodology, named X-Warehousing. X - Warehousing designs a warehouse at a logical level, and populates them with XML documents at a physical level. They uses XML schema which is merged with XML data sources. The resulted XML Schema represents the logical model of a data warehouse. Whereas, XML documents validated against the analysis objectives populate the physical model of the data warehouse, called the XML cube. They defined a general formalization for modeling star and snowflake schemas within XML. They also use the concept of attribute trees in order to help the creation and the warehousing of homogeneous XML documents by merging initial XML sources with a reference multidimensional model. They demonstrated their approach for designing and warehousing complex data using XML by using case study on breast cancer domain.

J Darmont et al. [44] proposed a general architecture framework for warehousing complex data. This architecture relies on the XML language, which helps storing data, metadata and domain specific knowledge, and facilitates communication between the various warehousing processes. This proposal takes into account the virtual and centralized XML warehousing. They presented the main issues in complex data warehousing, regarding data integration, the modeling of complex data cubes, and performance.
J. Darmont et al. [43] proposed a modelling process for integrating heterogeneous data. They have represented complex objects at conceptual level using UML based logical model. They have transformed heterogeneous data into XML documents and represents physical model. XML documents are mapped into a relational database in ODS to be remodelled in a multidimensional schema for analysis.

A M Arigon et al. [9] has addressed the issue of storage of multimedia data in various formats in a data warehouse describing content based or description based descriptors. They have considered dynamic aspects in multidimensional models that support temporal evolution and data analysis and there by suggested 3 layer models. They have implemented prototype in cardiology ECG data supporting OLAP, data modeling and visualization. The limitation of their work is efficiency of storage and optimization of processes.

O. Boussaid et al. [63] proposed generic data warehousing and on-line analysis process for complex data. They present generic UML model that allows to model low-level and semantic information for complex data. They integrated complex data as XML document into an ODS. This complex data integration approach is based on both the data warehouse technology and multi-agent systems. They also proposed X-Warehousing approach based on the XML to warehouse complex data. They exploited the concept of attribute trees to help in the creation and the warehousing of homogeneous XML documents, by merging initial XML sources with a reference multidimensional model. X-Warehousing method highlighted performance problems when storing warehouses in XML-native DBMSs. Therefore, they proposed a new join index that is adapted to XML, multidimensional data warehouses. This data structure allows optimizing the access time to several XML documents by eliminating join costs, while preserving the information contained in the initial warehouse. They have also presented a view selection algorithm, which they combine with the index structure to improve the overall efficiency of their strategy.

M Chen et al. [56] presents a new database structure model which organizes the multimedia data in a multi-dimensional data cube. Clustered data are stored in instance table in the data cube. A corresponding leading data are stored in dimension
When querying, a leading data is gained from dimension table firstly, followed by a k-nearest neighbor query from the corresponding category, the retrieval results are returned.

Y Sun et al. [88] presents a new database structure model which organizes the multimedia data in a multi-dimensional data cube based on XML database. In this data cube of XML, clustered data are stored in instance table. Corresponding data are stored in dimension tables. When querying, a leading data is gained from dimension table of XML, then receiving the multimedia data through XQuery.

Hadj Mahboubi et al. [32] proposed XML warehouse reference model to analyze complex data named X-WACoDa. Proposed model consists of three components for data integration, multidimensional modeling and analysis. For data integration they use ETL component that extract complex data in homogeneous XML format and integrate them into the XML data warehouse. To extract useful knowledge for warehousing they use data mining techniques. Warehouse component manages XML data with respect to XML data warehouse reference model. Their analysis component permits various analyses such as explanation and prediction over the data warehouse.

Jane you et al. [42] presents content based image retrieval using dynamic indexing and guided search combined with data mining and data warehousing techniques. They have developed wavelet based scheme for multiple feature extraction and developed multimedia starflake schema for image data warehouse, which support multiple feature integration and dynamic image indexing.

X Jin et al. [82] proposed algorithm for visual cube construction. They have provided responses to user requests with summarized statistics of image information and handles semantics related to image visual features. For visual cube construction they built two types of dimensions and four categories of measure. One type of dimension is meta information dimension which stores meta information regarding image and other is visual dimension which stores data based on image visual features. A clustering structure measure is proposed to help users freely navigate and explore images. They have proposed the idea of dynamic aggregation selection and cell overlapping issue in visual cube and OLAP.
H. Mahboubi et al. [31] introduced XML Warehouse Benchmark (XWeb). XWeB derives from the relational decision support benchmark. It is mainly composed of a test data warehouse that is based on a unified reference model for XML warehouses and that features XML-specific structures and its associate XQuery decision support workload. XWeB’s usage is illustrated by experiments on several XML database management systems.

M Srivastava et al. [55] proposed multimedia warehouse architecture by incorporating the concept of Content Server. The proposed Content Server will maintain the indexes for metadata of the stored multimedia content as well as will have a repository of multimedia content. The Content server stores the domain specific multimedia metadata repository. They have implemented content server as a custom application deployed on an Application Server. ETL tool extracts the data, transform them and store them in XML file. The data are stored in the warehouse in XML file format. The physical records will be retrieved by comparing their indexes with the domain specific indexes stored in content server. The Content Server manages both the retrieval of multimedia data and their metadata internally. OLAP Engine is a combination of OLAP server, Content Server and Knowledge Server. OLAP server is used for analytical operations over the warehouse.

Meenakshi Sharma and Sonia Dora [54] proposed approach for compression in Data warehouse. The proposed technique is used at attribute level by applying compression on string, integer and date type of attribute. The intelligent compression technique automatically identifies the type of attribute in a relation and performs the appropriate compression on it. Their results show that they achieve 79.09% compression ratio.

Mohd. Fraz et al. [59] proposed technique to improve compression efficiency of data warehouse. They have applied compression technique on relational databases. The proposed technique is used at attribute level on data warehouse by applying lossless compression on string, integer and float type of data and lossy compression on image attribute. They have used JPEG coding algorithm to compress image. They have got 13.5% compression ratio on image.
P Singh et al. [68] discussed the compression for image and video sequence. They concluded that Lossy compression is commonly used in multimedia data like audio, video and still images. Although the compression efficiency is the most important feature in any compression scheme and the compression depends on scalability, complexity and delay factors.

Z. Nebic et al. [90] offer tools to analyze and interpret results of usage statistics, course activity in e-learning through data warehouse and OLAP technologies for decision making. They build dimensional schema built upon requirements and generate reports built upon data warehouse for e-learning system. They analyze system workload by analyzing user activity by time of day and year. For usage of e-learning tools they have also analyzed the type of activity by hour and change in user activity by years.

M. Hassan et al. [58] investigates the use of business intelligence and OLAP tools in e-learning environments and presents a case study of how to apply these technologies in the database of an e-learning system. They use BI and OLAP technology to monitor and analyze the learner’s behaviour and performance in e-learning. In their proposed architecture they are integrating data from multiple sources of educational system in star schema structure. Further, data can be accessed and analyzed through the usage of API functions. They analyze number of activities per week and per course and types of activity per month to analyze the behaviour of learner.

Ankush Mittal et al. [6] presents a personalized delivery system for e-learning. Presented system uses a domain ontology and pedagogical model to arrange course materials in response to a user’s query. They have classified contents and used context based retrieval system. They have designed a system for indexing videos using audio, video and powerPoint slides and segmenting them into various lecture components.

Ahmed Abdu Alattab and Sameem Abdul Kareem [3] developed a facial image retrieval computational model. They have selected the semantic features for face. For face retrieval process, system uses the specific semantic features, of the face that is user looking for, to narrow down the search space. Eigenfaces is then projected onto
the narrowed down human faces search space to identify and retrieve the similar faces to the query face from the database. They have used Euclidean distance which is used for classification purpose.

Bor-Chun Chen et al. [15] proposed two orthogonal methods named attribute-enhanced sparse coding and attribute embedded inverted indexing to automatically detected human attributes for content based image retrieval. They have combined low-level features and automatically detected human attributes for content-based face image retrieval. Attribute-enhanced sparse coding uses the global structure and uses several human attributes to construct semantic-aware codewords in the offline stage. Attribute-embedded inverted indexing further considers the local attribute signature of the query image and provides efficient retrieval in the online stage.

P. Petchimuthu et al. [67] automatically detects the human attributes using attribute enhanced sparse coding and attribute embedded inverted indexing method. In the proposed method the images are converted into local patches and the patches will be converted into sparse coding and indexing is performed. For image retrieval system uses appearances and high level human attributes.

T. Li and T.Bretschneider [84] have retrieved satellite images based on semantic concept. They have classified satellite images into predefined semantic concepts as cloud, water, forest, urban area, farmland, bate soil, rock using grayscale images.

M. Ferecatu and N. Boujemaa [51] retrieved six predefined classes as city, cloud, desert, field, forest, and sea from isolated images and ground truth database. Image retrieval is done on the basis of these predefined classes.

L. Niu and L. Ni [46] have used multi-band isolated JPEG2000 codec images to retrieve area of interest depending on query image using hue, saturation, and value color model conversion.

N. Laban et al. [61] introduced an approach to define image retrieval process using the semantic class properties for satellite images. They have determined the different semantic classes using the polygon shapes namely, water, sand dunes, desert,
vegetation, rock, and urban. They have used Query by polygon (QBP) paradigm for the content of interest. They first extract features from the satellite images using multiple tiling sizes. Accordingly the system uses these multilevel features within a multilevel retrieval system that refines the retrieval process.

2.2 Literature Review Analysis and Findings

In the late 1990s, [65, 38] proposed a system prototype, for multimedia data mining. They have created multimedia data cubes that uses low level features. W Lee et al [81] developed multimedia data warehouse for EoD systems. The system is implemented for video data, the relevant shot is pre processed for its physical and semantic structure by providing appropriate indexing and retrieval characteristics. For data storage they use star schema model. You et al [87] presents a data warehousing approach on hierarchical multimedia information retrieval. Develop a multimedia starflake schema to integrate multiple data streams for hierarchical data representation and indexing. [75] proposed multi-tier image data warehouse framework based on the OOAD and component based development and have not described modeling technique much.

Researchers have built multimedia data warehouse which can analyze data coming from heterogeneous and distributed sources [76, 35]. Because of the interoperability and flexibility of the XML, researchers also proposed model which is used to build XML based data warehouse [35]. [76] provides materialized views to use in the analysis of multimedia data.

Multimedia data can be represented by different levels of features. Researchers have suggested multiversion multidimensional model [7, 8, 9] which describes and stores multimedia data with content based or description based descriptors. They also use aggregation functions for multimedia data that are integrated into the data warehouse and in the OLAP engine. The limitation of their work is efficiency of storage and optimization of processes.

[5, 29, 82] presented the model that uses semantic based data. [5] proposed hierarchical way to store semantic data, they use two repositories of metadata which describes data in hierarchical manner in XML files. The authors of [29] have built a
data warehouse which has two ontologies, one for the specific business terms and one for the technical terms, specific to the aggregation and knowledge extraction tools. This requires a one-time collaboration between the business experts and data warehouse designers, to produce a mapping between two ontologies. For multimedia data storage and representation [82] designed visual cube which uses two types of dimension. One type of dimension is meta information dimension which stores meta information regarding image and other is visual dimension which stores data based on image visual features. They also propose the idea of dynamic aggregation selection.

Y Sun et al [88] presents a data cube of XML, clustered data are stored in instance table and receiving the multimedia data through XQuery. The concept of content server is also proposed [55], which stores content of multimedia data in content server. Proposed Content Server will maintain the indexes for metadata of the stored multimedia content as well as will have a repository of multimedia content.

Image, audio, video, text, documents are known as complex data. [78, 79, 44, 63] proposed architecture and data model for complex data. For warehousing complex data for web architectures are suggested in [43, 62].

Because of the interoperability and flexibility of the XML, researchers also proposed model which is used to build XML based data warehouse [17, 29, 31, 32, 47, 48, 80].

To achieve and enhance query performance indexing [81, 87, 42], compression techniques [81, 55] and materialized view [76] are suggested. For traditional data warehouse, [13] provides approach for selection of materialized view, and selection of Materialized view by Query Clustering techniques is suggested in [33] for XML warehouse. [34, 63] suggested join indexing for XML data warehouse.

To represent face image data, semantic based data [3, 15, 67] are used. Geographic image data is represented by low level features [46, 61] and by semantic features [84, 51, 61]. Satellite images also retrieved using automatic semantic categorization and content-based description [12]. [41, 90, 58] proposed e-learning data warehouse system. For e-learning video data [6] used semantic based data. [27] classified the
color images and retrieved the images using data mining algorithm. [39] proposed approach for color image retrieval.

**Challenges:**

- The storage mechanism for targeted access is challenging in case of multimedia data. Multimedia data are usually specified by low level content based features (Color, Shape, Texture, etc.) or semantic features. The continued challenge is how to model effectively the levels of features that facilitates the access mechanism to support further analysis of data.
- The consequent challenge that surfaces up is the functionality performance of system because of bulky storage and its access. The existing models have given partial solution to this challenge. These limitations are generating subsequent challenges of manageability of data volume and response time when accessed.
- The existing models and architectures of multimedia data warehouse has led to the establishment of multimedia data warehouse which are facing challenges of enhanced query response. To enhance the query response time with storage efficiency an improved multimedia data warehouse model is needed.

### 2.3 Problem Formulation

The multimedia data are bulky, semi structured or unstructured in nature, the performance level of storage, retrieval and analysis of multimedia data is a critical issue of research as indicated by literature review and its findings. This undertaken research has considered these issues to be taken care to evolve optimal solution. To achieve optimum levels of performance, the proposed research is targeted to develop a suitable model and subsequent framework to achieve the targets of research. The model/framework is to be mapped to the physical design of multimedia data warehouse leading to implementable prototype. This prototype is to be implemented to justify the targeted research outcomes and validate the proposed work to the acceptable extent.

The storage efficiency will be enhanced by widely used methodologies of compression techniques. For efficient retrieval indexing techniques, partitioning techniques and materialized view are used. By using such techniques performance of storage and retrieval of multimedia data can be achieved.
2.4 Research Problem Statement

The purpose of the undertaken research is to address the issues of managing and analyzing large volume of multimedia data and for the same, the work is targeted to propose a multimedia data warehousing model and a prototype based on dimensional modeling.

In order to accomplish this goal, the following objectives are identified:
1. Identify the best way to represent multimedia data, applicable models and technologies and develop a data warehouse for multimedia data.
2. Demonstrate the effectiveness and efficiency in terms of warehouse performance by:
   1. Representing multimedia data by maximum possible feature abstraction level
   2. For effective storage and retrieval existing compression technique is applied, indexing has been used and partitioning approach is used.
3. To prove proposed model, prototyping model is implemented which is validated against biometric image data, geographic image data and e-learning video data.

2.5 Proposed work

The proposed research work is aimed at modeling the prototype of Data Warehouse which houses multimedia data. The proposed model is developed to provide higher level of efficiency in data storage and access mechanism. The implementation of the proposed model is to house multimedia data. The multimedia data sources provide large data sets related to biometric data, geographic data and e-learning video data.

Proposed model based prototype implements the methodology that extracts multimedia data from targeted data sources. After extracting multimedia data, the relevant characteristic of multimedia data i.e. low level feature based data, high level feature based data and calculated features are scheduled to be extracted according to the need and goal of the analysis. The fine tuned extraction provides extracted data that is taken at staging level. At the data staging level, the extracted multimedia data is compressed using existing techniques. The data generated in extraction phase, is then ready to get loaded in data warehouse. In the data warehouse, relational data is to
be modeled in multidimensional structure. The multidimensional model is modeled in such a way that low-level, high level and calculated features are stored at dimension level and measures are stored at fact level. The proposed prototype uses different techniques like compression, indexing, partitioning and materialized view by which performance of storage and retrieval of multimedia data can be achieved. Compression is applied only on multimedia data to reduce the size of data without losing much quality. Indexing is applied on feature data and data that uniquely identify the dimension. Partitioning is applied at the cube level which slices a cube. Materialized view is created for frequently executed queries against the multimedia data warehouse. For data analysis, a tool has been created that accesses and analyzes multimedia data from the multimedia data warehouse.

This prototype is to be implemented to justify the targeted research outcomes and validate the proposed work to the acceptable extent.