In a large enterprise-wide database system, it is difficult to analyze the data since the data is stored in different formats and locations. To solve the integration problem of the enterprise systems, data warehouses are used as an integrated system. Educational institutions measures success very differently from business organizations therefore it is needed to develop a good data warehouse for academic institutions as well.

2.1. Definition of a Data Warehouse (DW)

Over the last few years, institutions have increasingly turned to data warehousing to improve information flow and decision support. A DW can be a valuable asset in providing easy access to data for analysis and reporting. Unfortunately, building and maintaining an effective DW has several challenges. Today many organizations are trying to develop a DW to meet their business requirements.

Han and Kamber, define a DW as "a repository of information collected from multiple sources, stored under a unified scheme, and which usually resides at a single site" [12]. In educational terms, all past information available in electronic format about a school or district such as budget, payroll, student achievement and demographics are stored in one location.

According to Inmon, "a DW is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process" [13].

Inmon and Hackathorn stated that:

"A DW is a separate store of data extracted from one or more production databases to produce an authoritative source for decision support. It can be considered a decision support system. Reporting data is another important aspect of data warehousing because the main output from DW systems are either queries with minimal formatting or formal reports."

We can summarize some features of DW as follows:

Subject-oriented: DWs are designed to help one analyze data. Usually, this type of data is not suitable for decision-makers to use. Decision-makers need subject-oriented data. For example, to learn more about university students success ratios, one can build a warehouse that concentrates on grades. Using this warehouse, one can answer questions like "How many Computer Engineering students passed from "Compe523" course last fall?" This ability to define a DW by subject matter, grades in this case makes the DW subject oriented.

Integrated: Integration is related to subject orientation. DW must put data from disparate sources (like excel sheets, relational database) into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Nonvolatile: Nonvolatile means that, once entered into the warehouse, data should not change. There is no update of data as in an Online Transaction Processing (OLTP) database, and does not require transaction processing, recovery and concurrency control mechanisms. DW requires only two operations in data accessing: loading and access of data. This is logical because the purpose of a warehouse is to enable one to analyze what has occurred. That is why it does not require transaction processing, recovery and control mechanisms.

Time-Variant: In order to discover trends in business, analysts need large amounts of data. This is in contrast with OLTP systems, where performance requirements demand that historical data be moved to an archive. A DW's focus on change over time is what is meant by the term time variant. Operational database contains current value data. The key structure of operational data may or may not contain some element of time.

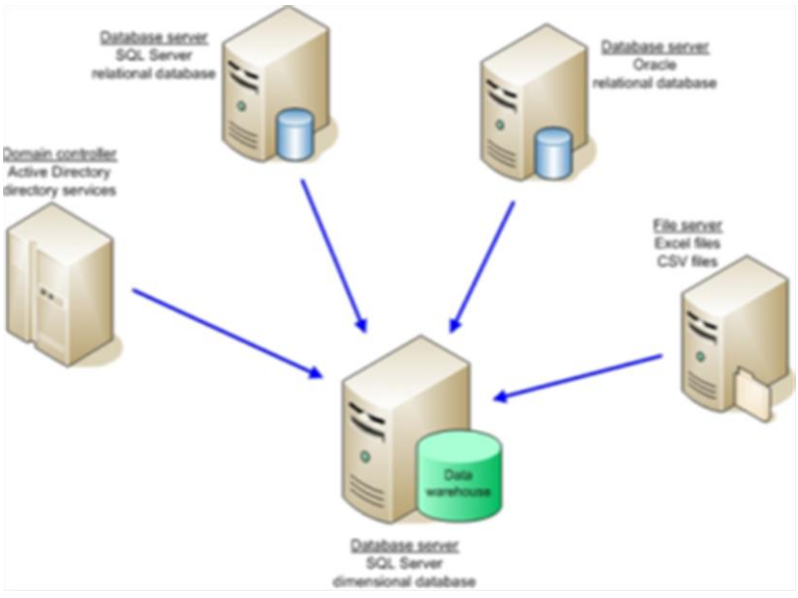

Figure 2.1: Using a Data Warehouse to Consolidate Heterogeneous Data

2.2. Transaction Processing versus Decision Support Databases

There are two main types of database systems: Operational systems and analytical systems that include DWs. They differ in using data. Operational systems are referred to as online transaction processing systems and designed for real-time data entry and editing. One account, one student record, one inventory item, or one order can be updated or deleted in any time. Transactions are generally predefined and require the database to provide fast access. Analytical systems, such as DWs, are designed to help an organization with decision-making. A DW is updated periodically and its solutions offer query capabilities, trend analysis, and reporting.

2.2.1. Online Analytical Processing (OLAP)

OLAP is a category of software technology that provides fast, consistent and interactive access to information. OLAP is designed for planning, forecasting, managing - decision making and creating business structures and combine them in such a way as to allow users to quickly answer business questions. "How many blue sweaters were sold via mail-order in Ankara so far this year?" or in a school database "How many students pass from Compe523 lesson this year?" are kinds of questions that OLAP is very good at answering.

There are several advantages of OLAP systems. Its simple design is easy to understand. Queries return consistent results. Data can be separated and combined by means of every possible measure in the business and users have access to business descriptions. In terms of the technology, an OLAP database can be implemented on top of an existing relational database (called Relational OLAP, ROLAP) or it can be implemented via a specialized data store (called MOLAP). In ROLAP, the data request is translated into SQL and the relational database is queried for the answer. In MOLAP, the specialized data store is preloaded with the answers to all possible queries so that any request for data can be returned quickly [14].

2.2.2. Online Transaction Processing (OLTP)

OLTP systems are useful for addressing the operational data needs of an organization and designed for day-to-day operations like payroll and accounting systems. They are not suitable for supporting decision-support queries. The data is updated continuously [15]. These systems contain vast amount of data which can be difficult to access. However, to get valuable information from this data is crucial for the decision-maker. OLTP systems do not hold the data constantly. Hence the formalization of DW and data marts occurs.

There are several benefits of OLTP systems. It reduces performance cost of redundant data, supports high transaction rates and captures primary business production information. Besides its advantages, there are also some problems. These can be itemized as follows;

- Only "pre-canned" reports could be generated on a scheduled basis
- ad-hoc real-time querying was virtually impossible
- Virtually no usage of historical data
- With a normalized structure, we have to query from multiple tables to get the detailed information.

Today's online transaction processing increasingly requires support for transactions that span a network and may include more than one company. For this reason, new OLTP software uses client/server processing and brokering software that allows transactions to run on different computer platforms in a network.

2.3. Data Mart

Data warehousing is a large, expensive solution and generally used for big organizations. However a data mart is similar to a DW according to structure and purpose but it is limited to a departmental or specific subject need. Thomas Connolly and Carolyn Begg [16] defined a data mart as a subset of the data in a DW and it summarized the data relating to a department or a specific function. A schema showing the data marts and DW is shown in **Figure 2.2** Data marts contains less data than DWs as they serve to a specific purpose.

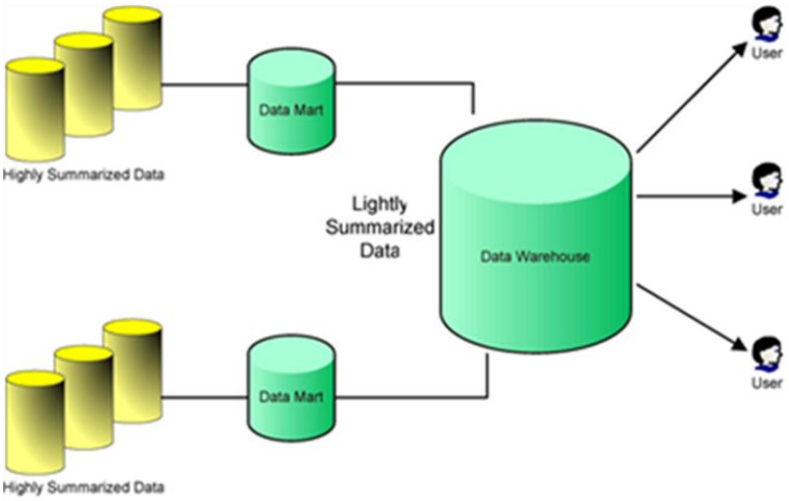

Figure 2.2: Data Marts and data warehouse

The followings are advantages of using a data mart rather than a DW.

- Data marts contain less data, therefore query response times are shorter compared to DW.
- Data integration and transformation process of data takes shorter time.
- Data marts are more simplified than DWs, so building takes less time.
- Besides the advantages, there are some disadvantages of using a data mart.
- Size of a data mart is smaller than a DW as it answers small organization needs. If the size will be increased, its performance will decrease.

2.4. Structure of the DW

Data contained in a DW contains five types of data: data currency, existing data, data summarization (lightly and highly summarized data), and Metadata. The following **Figure 2.3** shows the structure of the DW [17].

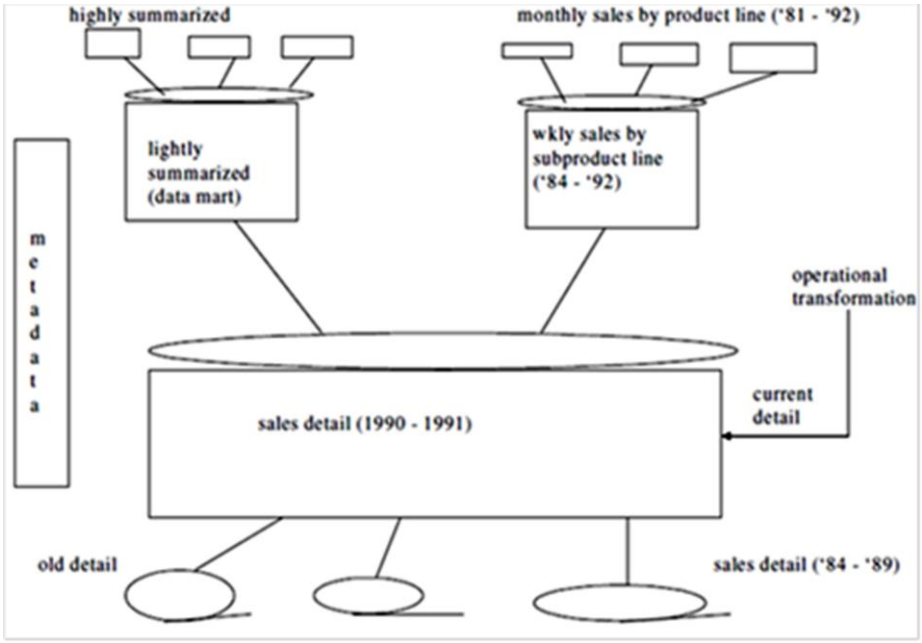

Figure 2.3: the structure of the DW

Metadata

Tannenbaum [18] defined metadata as; the detailed description of the instance data; the format and characteristics of populated instance data; instances and values depend on the role of the metadata recipient. Examples of metadata include filenames, data element definitions, data element names, lengths, and program names and so on.

An important component of the DW environment is metadata. Metadata or data about data is a critical component in a warehouse environment. Metadata has many uses. Users need to know what data is available, how it is organized and what it means. It allows the end user to navigate through the possibilities. In other words, when a user approaches a DW where there is no metadata, the user does not know where to begin the analysis. Metadata typically contains:

- Structure of data as known to the programmer and to the DSS analyst
- Source data
- Transformation of data
- Data model
- DW
- History of extracts

Data Currency

Current detailed data is at the heart of the DW and reflects the most recent happenings. This data is stored according to a data model. Older detailed data accumulates over time as fresh data enters the application. Generally, in most DSS, data obtained from detailed records on transactions are cleaned and then loaded [19].

Data Summarization

Data is lightly summarized and stored in the DW. The data might be summarized over different units of time and different attributes. Data could also be found in a highly summarized form for quick retrieval. The highly summarized data could be physically stored within or outside the DW but is logically considered a part of the DW [20].

2.5. DW Modeling Techniques

DW modeling is the process of building a model for the data in order to store in the DW. This chapter provides a basic understanding of data modeling, specifically for the purpose of implementing a DW. There are two data modeling techniques that are relevant in a data warehousing environment are Entity Relationship (ER) modeling and dimensional modeling.

ER Modeling and Dimensional Fact Modeling

ER modeling produces a data model of the specific area of interest, using two basic concepts: entities and the relationships between those entities. Detailed ER models also contain attributes, which can be properties of either the entities or the relationships. The ER model is an abstraction tool because it can be used to understand and simplify the ambiguous data relationships in the business world and complex systems environments.

Dimensional modeling uses three basic concepts: measures, facts, and dimensions. Dimensional modeling is powerful in representing the requirements of the business user in the context of database tables. Measures are numeric values that are can be added and calculated. ER modeling uses the following concepts: entities, attributes and the relationships between entities. The ER model can be used to understand and simplify the ambiguous data relationships in the business world and complex systems environments. An entity is represented on the diagram as a rectangular box. The name of the entity appears in the top section of the rectangle, such as Instructor and Department in the below Figure 2.4. Attributes are listed in the lower part of the rectangle that represents the entity which they belong to. A relationship is a line between the boxes that represents the entities that it links. The relationship may be identified in different ways like; one-to-one, many-to many, and one to many.

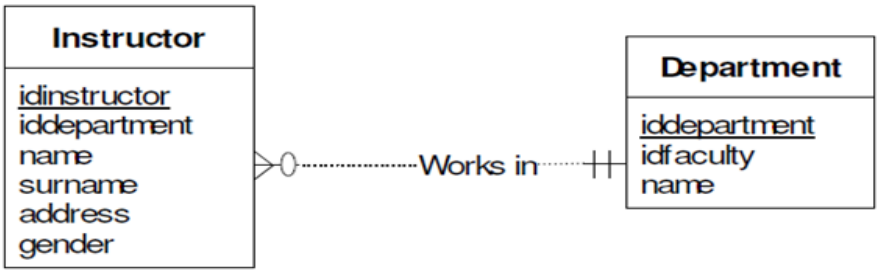

Figure 2.4: Entity Relationship Diagrams

Some respects, dimensional modeling is simpler, more expressive, and easier to understand than ER modeling. Dimensional modeling uses three basic concepts: measures, facts, and dimensions. Dimensional modeling is powerful in representing the requirements of the business user in the context of database tables [21]. The dimensional model is a logical design technique that seeks to make data available to the user in an intuitive framework that is intended to facilitate querying. Dimensional model is composed of fact tables and dimensional tables where fact tables are normalized tables that represent the process being tracked. It is a logical design technique, used in Data Warehousing databases, which is an alternative to Entity-Relationship (ER) modeling. **Figure 2.5** represents the dimension tables and fact table with primary keys and foreign keys.

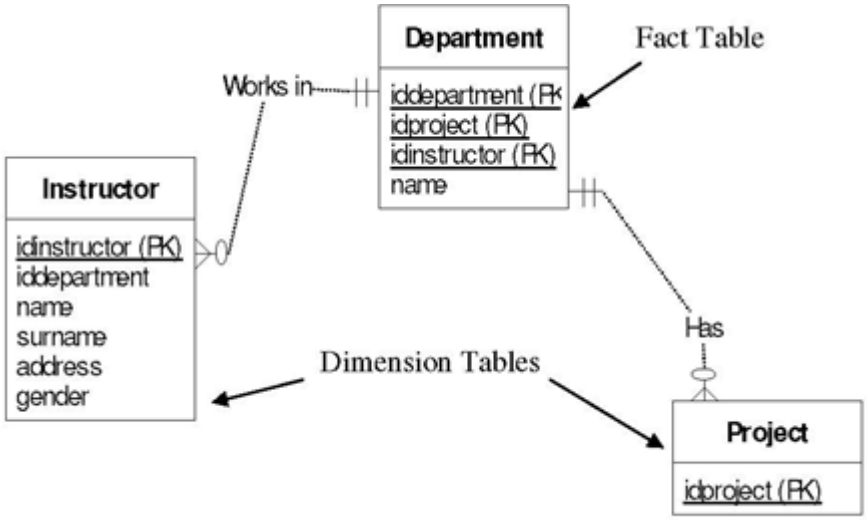

Figure 2.5: the dimension tables and fact table with primary keys and foreign keys.

Each dimension table is assigned a primary key that are not related to one another. The primary key is unique key for the dimension table, which is replicated in a fact table where it is referred to as a foreign key. A fact table is a table that contains the data (factual history) like Student Id, Course Id, Semester Id, etc.

Connolly and Begg derived the way a DW or a data mart structure in dimensional modeling into several ways [22]; Flat schema, Terraced Schema, Star Schema, Fact Constellation Schema, Galaxy Schema, Snowflake Schema, Star Cluster Schema, and Starflake Schema. However there are two basic models that widely used in dimensional modeling: Star model and Snowflake model.

Star Schema

The schema associated with a dimensional model is referred to as a star schema. It is widely used because of its simplicity. It is called a star schema as its shape resembles a star. A star schema consists of a central fact table joined with several dimension tables. The fact table is large and has all the transaction data and numbers, the dimension tables are smaller and have

descriptive information. In Figure 2.6, you can see fact and dimension tables. Sales table is the fact table Period ,Market, Product tables are dimension tables.

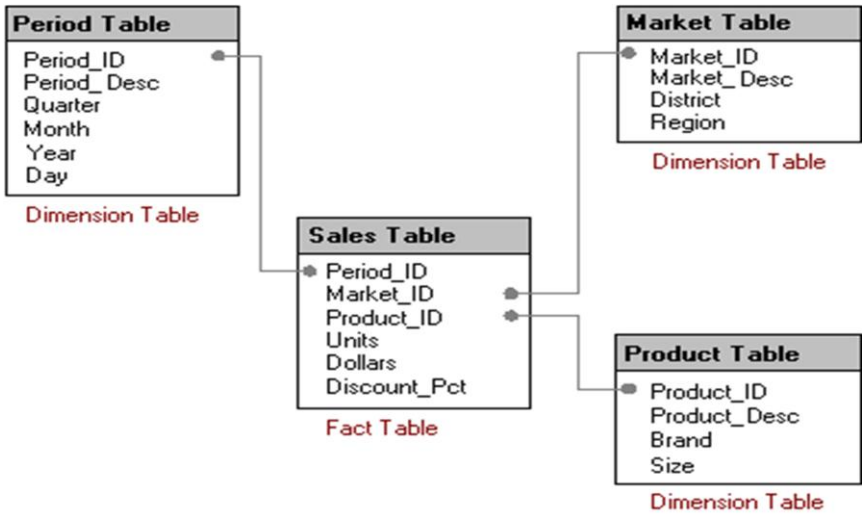

Figure 2.6: the dimension tables and fact table in star schema.

Snowflake Schema

Snowflake schema model is derived from the star schema and, as can be seen, looks like a snowflake. The snowflake model is the result of decomposing one or more of the dimensions, which generally have hierarchies between themselves. Many-to-one relationships among members within a dimension table can be defined as a separate dimension table, forming a hierarchy. For example in the Figure 2.7, a starflake schema is given.

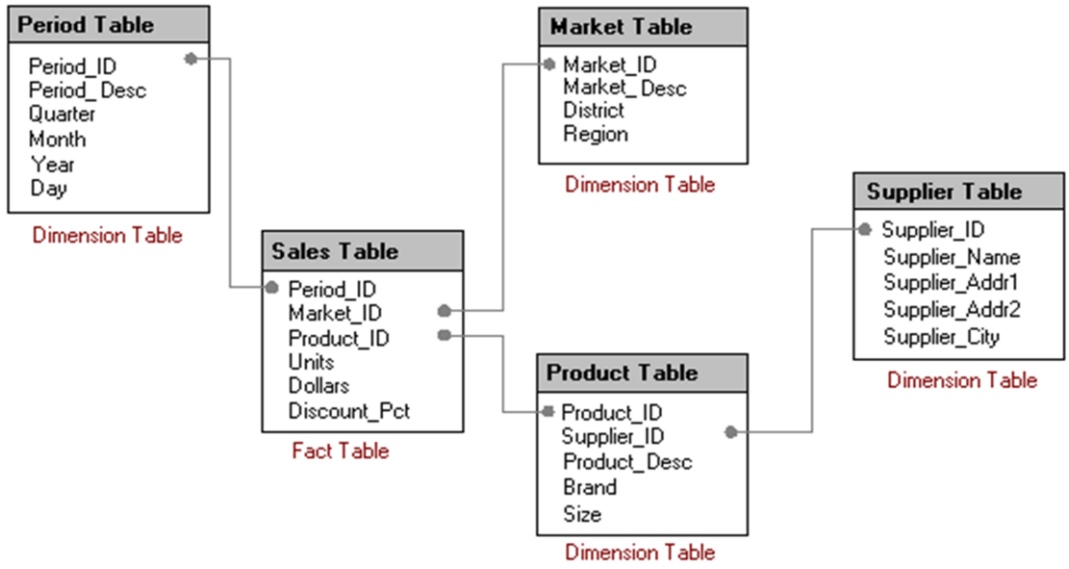

Figure 2.7: An Example of a Snowflake Schema

The Star Schema vs. the Snowflake Schema

There is much debate over which schema model is better. Some suggest—in fact, insist—that the star schema with its denormalized structure should always be used because of its support for simpler and better performing queries. However, such a schema makes updating the data a more complex process. The snowflake schema, on the other hand, is easier to update, but queries are more complex. At the same time, there's the argument that SSAS queries actually perform better against a snowflake schema than a star schema. And the debate goes on.

In many cases, data modellers choose to do a hybrid of both, normalizing those dimensions that support the user the best.

2.5.1. Benefits of Data Warehousing

There are several benefits of data warehousing. The most important ones are listed below.

- DW improves access to administrative information for decision makers.
- It can get data quickly and easily perform analysis. One can work with better information, make decisions based on data. DW increases productivity of corporate

decision-makers.

- Data extraction from its original data sources into the central area resolves the performance problem, which arises from performing complex analyses on operational data.
- Data in the warehouse is stored in specialized form, called a multidimensional database. This form makes data querying efficient and fast.
- A huge amount of data is usually collected in the DW. Compared with relational databases that are still very popular today, data in the warehouse does not need to be in normalized form. In fact, it is usually de-normalized to support faster data retrieval.

2.5.2. DW Modeling Tools

Building a DW from independent data sources is a difficult process. This process involves extracting, converting, cleaning, integration and transformation of the data. In order to do these operations, an ETL (extract, transform, and load) tool is required. The key steps that need to be undertaken to transform raw operational data to a form that can be stored in a DW for analysis are: [23]

- **Extraction** The goal of the data extraction step is to bring data from different sources into a database before modification.
- **Converting** the data into a format that is suitable to the DW.
- **Cleaning** of the data. Data entry errors and differences in schema formation can cause for example student dimension table to have several corresponding entries for a single student.
- **Integration** of the different datasets to suit the data model of the DW.
- **Transformation** of the data through summarization and creation of new attributes. It is a set of rules and scripts that typically handles the transformation of data from an input schema to the destination schema.

2.5.3. Data Analysis and Assessment Tools

When these steps have been completed the data is ready for analysis. There are several options to carry out these steps. The following products allow schools and districts to gather, analyze, and make the most effectiveness from their data [24].

Quella Business Intelligence Solution is an easy-to-use tool covering the Business Intelligence Process from Data Extraction to Analysis. As a total solution, Quella extracts transactional data from heterogeneous data sources, consolidates the extracted data and provides necessary business information to decision makers through List Reports, Charts and OLAP-Reports.

Cognos allows users to create individual education plans. Its web-based system lets multiple people access the forms simultaneously.

Oracle Warehouse Builder (OWB) is Oracle's comprehensive tool for ETL, fully integrated relational and dimensional modeling, data quality, data auditing, and full life cycle management of data and metadata.

SQL Manager Lite for Oracle Simplify and automate database development process, design, explore and maintain existing databases, build compound SQL query statements, manage database user rights and manipulate data in different ways.

Apple's PowerSchool is a web-based SIS that delivers real time information across the Internet. This means you can easily access your data anywhere, anytime via your web browser using your current Internet connection. PowerSchool stores all data in a single centralized relational database, and allows reporting.

Cancery's Student Management System is designed as a web-based Student Information System built on Microsoft .NET. This release gives districts and schools an open, centralized system for managing student information. System requires lower hardware, software, infrastructure, and personnel costs than decentralized solutions. It allows user to visualize data through graphs and charts and generate basic reports.

ERStudio a model-driven data architecture and database design solution, helps companies discover, document, and reuse data assets. With round-trip database support, data architects have the power to analyze existing data sources as well as design and implement high quality databases. The highly-readable visual format enhances communication across functions.

Sunguars's Performance PLUS provides data extraction, aggregation, and analytic tools.

Octopus is a simple Java-based Extraction, Transform, and Loading (ETL) tool. It may connect to any JDBC data sources and perform transformations defined in an XML file. A load job-generator is provided to generate Octopus load job skeletons from an existing database. Many different types of databases can be mixed (MSSQL, Oracle, DB2, QED, JDBC-ODBC with Excel and Access, MySQL, CSV-files, XML-files,...) Three special JDBC drivers come with Octopus to support JDBC access to CSV-files (CSV-JDBC), MS-SQL (FreeTDS) and XML. Octopus supports Ant and JUnit to create a database / tables and extract/load data during a build or test process.

BrioQuery is a sophisticated ad hoc query and reporting application for client/server database environments. BrioQuery provides non-technical users with easy access to the valuable data available in the university's Sybase data extracts.

2.6. Data Warehouse in Education

To stay competitive in today's rapidly changing environment, an organization needs an effective database management system. Business enterprises can benefit from data warehousing. The concept of collecting data into separate, multidimensional repositories to handle complex decision making-activities can be used for educational purposes as well. Educational institutions need this capability to ensure quality data management for strategic decision-making.

A review of the literature reveals that data warehousing is becoming an increasingly popular way to store and retrieve data. Many educational institutions are creating a DW to provide integrated administrative information for planning and reporting purposes. Inmon stated that; data warehousing is becoming "a solid business strategy" for institutional

research. He also declares that, it helps higher education enterprises understand who their customer base is, what they do, and what types of courses and services do they offer affect their learning and satisfaction.

Data warehousing collects and organizes data from multiple sources so that it can then be easily analyzed, extracted and used. Because of this, the DW is a core component for enabling decision making for educational institutions. The use of a DW presents several potential advantages for educational institutions, including timely access and to better evaluate the data. Institutions can easily and quickly identify troublesome trends in its enrollment and evaluate why they are occurring. They can determine strengths and weaknesses within themselves. The use of a DW allows an educational institution to use this information in making appropriate decisions, and that is the desired end goal of the data warehousing process [25]. By using a DW, educational institutions could overcome several problems. Not only do they set operational data free from complex retrieval, but also they benefit from many other issues, concerning the management. For example, consider these few questions, which could be answered easily, using a DW in a higher education institution:

- Total number of academic staff in all programs according to their titles in a range of years
- Total number of students in faculty of engineering according to their programs in a range of years
- An estimate of the total number of students that will take the course.
- Letter grades of a specific course according to different years
- Success ratios of faculties and departments in a range of years
- The number of female/male students registered in the university according to program and year.
- Total number of students in a faculty or program and total number of academic staff according to their titles in a range of years
- Total number of part-time and full-time instructor list according to their titles.
- Total number of credits an academic staff gives in a range of years.

Note that these questions above are not very easy to answer just by using a relational database. The main reason behind this problem is a simple fact namely, "relational databases are good at retrieving small number of records quickly, but they do not retrieve a large number of records and summarize them on the fly" .

Nowadays, many higher education institutions are starting to see the value of the integrated, standardized, clean and easy access to data for better decisionmaking. With the usage of data warehousing, analysis, management decision making and other reports can be done in a simpler way.

Breiter and Light stated that [26]; school information systems constitute a clear sub-group of management information systems that are used in educational organizations. In school, distinct information systems support different types of decisions: administrative information systems, learning management systems and assessment information systems. Figure 2.8 illustrates the school information systems topology.

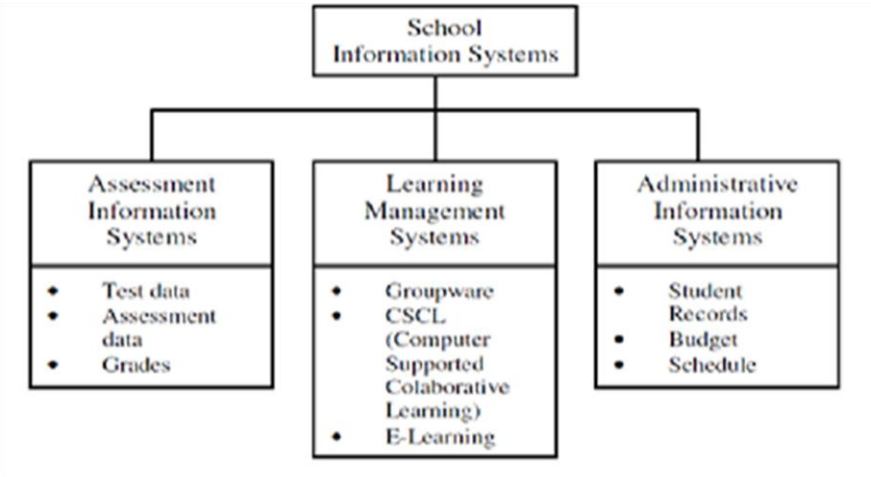

Figure 2.8: the school information systems topology

Administrative information systems consist of financial, human resources, student grades, course, student fees, research, student enrollment, faculty workload and administrative areas. Learning Management System (LMS) supports learning process. LMS is software that automates the administration of training events.

According to Remes [27], all LMS's manage the log-in and registration of users, manage course catalogs, record data from learners, and provide reports to management. Assessment information systems get data about the student online test performance. In an effort to contribute to the growth of the higher education Data Warehousing industry and to

see the different usages of data warehousing, different higher education institutions are examined from literature. From the literature survey, we have understood that, higher educational institutions are using DW solutions for administrative and assessment information purposes only. Research about learning management systems were not encountered. The following sections summarize these studies.

2.6.1. Administrative DW Solutions

In order to allow university decision-makers to evaluate various programs and academic departments, Baylor University started using a DW and decision support system. The main goal was to bring data together for administrative process that let administrators and decision makers answer questions and make better decisions. Other goals included giving university decision makers the ability to create ad-hoc queries, generate reports, and perform analyses without having to request work from the university's IT department. To do this, they used SAS System as a database and SAS MDDB Report viewer as a Query tool. They developed an internet-based system. They have reported that, by giving user direct access to information the organization saved the time of their staff and helped to make quick and effective decisions.

Manitoba University [28] has a number of legacy systems that are used to support the day to day operations of the university. The disadvantages were numerous to the university in relying on operational systems for data in support of management decision making. Accessing different systems to get different kinds of information and then trying to rationalize them is both difficult and time consuming. They thought that data warehousing would increase their ability to respond and their productivity. In this university, the clients using the system from outside could not make their own analyses because the hierarchical storage structure of the data did not allow for ad-hoc access. In order to extract the data, high level programming was needed, and there was no easy way to get information into desktop tools. These situations made the development of a warehouse a priority for the university. Their data model defines subject areas, such as **Students**, **Employees**, **Courses**, **Space**, and **Alumni**. They designed the DW on a Sybase Server.

MIT'S DW has been in existence for about 5 years and they have Personnel, Financial, Purchasing, Telephone, Student, and Award information in the warehouse [33].

The main purpose of the warehouse was a reporting and data distribution environment for departments, labs and centers at the MIT. Besides, their DW is acting as a hub, to facilitate the exchange of information between systems. They are using BrioQuery tool as an end user query and reporting tool. Developers designed the DW in a way that users can easily understand and use. They have organized the information in star schemas and they gave direct SQL access to the DW in order to give flexibility to the users.

2.6.2. Assessment DW Solutions

Ingham reports That.Polytechnic University, using an integrated student information system as a main source of data developed a DW because of the difficulty of retrieving information in such enterprise system. They developed their DW for assessment process in order to get student test scores and alumni data as a report. They stored the institutional data in a relational database format and used ten year data purged from existing system. The system serves to the department heads and administrators.