University of the Philippines
Open University

Fundamentals of Enterprise Data Management

A Business Analytics Course

Dr. Eugene Rex Jalao
Dr. Ria Mae Borromeo
Asst. Prof. Mari Anjeli Crisanto
Course Writers

University of the Philippines
OPEN UNIVERSITY

COMMISSION ON HIGHER EDUCATION

University of the Philippines
OPEN UNIVERSITY

UNIVERSITY OF THE PHILIPPINES OPEN UNIVERSITY

Fundamentals of Enterprise Data Management

A Business Analytics Course

The course is designed to introduce students to the fundamentals of database management systems, enterprise data management using data warehouse (DW or DWH), which can be used for further data mining, reporting and data analysis purposes. It describes various activities involved in data mining tasks like data anomaly detection (outlier/change/deviation detection), data association rule learning (dependency modeling), data clustering, data classification, data regression and data summarization. This course also introduces students to formalized means of organizing and storing structured and unstructured data in an organization. It describes the seriousness of information security and provides techniques to use predictive analytics for detection of fraudulent activities.

COURSE OBJECTIVES

At the end of the course, the students should be able to:

1. Understand database management systems;
2. Describe the process of data discovery and data patterns in large data sets;
3. Understand various methods related to intersection of artificial intelligence, machine learning, statistics, and database systems;
4. Understand various techniques related to data extraction and data preprocessing before using data modeling;
5. Understand the concept of master data management (MDM);
6. Describe data inference considerations, interestingness metrics, complexity considerations;
7. Understand various techniques used for post-processing of discovered structures and visualization;
8. Describe the importance of data warehouses (DW or DWH) for reporting and data analysis and understand the differences from operational data source (ODS);
9. Describe formalized means of organizing and storing of documents and other content in an organization related to the organization's processes;
10. Describe the need and policy around data security and privacy (information security) and techniques to restrict information from unauthorized access, use, disclosure, disruption, modification, perusal, inspection, recording or destruction;

11. Describe online fraud and their consequences and understand predictive analytics for detection of fraudulent activities; and
12. Develop an awareness of the ethical norms as required under policies and applicable laws governing confidentiality and non-disclosure of data/information/documents and proper conduct in the learning process and application of business analytics.

COURSE OUTLINE

Module 1 - Database Management Systems

A. Database Systems
B. Functions and Components of a Database Management Systems
C. Master Data Management
D. Databases and Normalization
E. Entity Relationship Diagram and Relational Modeling
F. Enterprise Content Management (ECM)

Module 2 - Data Warehousing

A. Data Warehouses, Data Marts, and Operational Data Source
B. Alternate Data Warehousing Architecture
C. The Kimball Lifecycle
D. ETL (Extraction, Transformation, Loading)
E. Using the Data Warehouse for Business Intelligence

Module 3 - Knowledge Discovery

A. Knowledge Discovery in Databases (KDD)
B. CRoss-Industry Standard Process for Data Mining (CRISP-DM)
C. Data Mining Methods
D. Programming Languages/Tools

Module 4 - Enterprise Data Management Issues

A. Data Security, Privacy, Online Frauds, and Ethical Norms

COURSE MATERIALS

The course materials shall also be posted on the virtual classroom or course site one week before the module schedule indicated in the study schedule below.

STUDY SCHEDULE

WEEK	TOPIC	ACTIVITY
1-4	Database Management Systems	WATCH: • Database Management Systems (Mari Anjeli Crisanto) SUBMIT: • Quiz 1 • Assignment 1
5-8	Data Warehousing	WATCH: • Data Warehousing (Mari Anjeli Crisanto) READ: • Getting Started with Data Warehousing Chapter 1 pp. 17-29 • Getting Started with Data Warehousing Chapter 4 pp. 55-61. • Data Warehousing: The Foundation of BI https://goo.gl/sErpvH SUBMIT: • Quiz 2 • Assignment 2
9-12	Knowledge Discovery	WATCH: • Knowledge Discovery (Ria Mae Borromeo) • CRISP-DM (Eugene Jalao) • Tools in Data Mining (Eugene Jalao) READ: • Predictive Analytics.pptx (Hsin-Chang Yang) pp. 6 – 67 SUBMIT: • Quiz 3 • Assignment 3
13-15	Enterprise Data Management Issues	WATCH: • Enterprise Data Management Issues (Mari Anjeli Crisanto) • Opportunities and Ethics in Data Warehousing (Mari Anjeli Crisanto) SUBMIT: • Quiz 4 and Assignment 4
16		**Final Exam**

COURSE REQUIREMENTS

To pass the course, you must accomplish the following and your final grade must be greater than or equal to 60.

1. Complete four quizzes (35% of the final grade)
2. Submit four assignments (60% of the final grade)
3. Take a final examination (40% of the final grade)

Quizzes

Quizzes are meant to test the understanding of students about the concepts discussed. The questions may be of type multiple choice, true/false, or essay.

Assignments

The assignments are meant to enable students to practice or apply the concepts they have learned. Depending on the topic, assignments may be modelling exercises, position papers or programming assignments.

Final Exam

The final exam covers all topics studied in the course. Similar to the quizzes, the questions may be of type multiple choice, true/false, or essay.

MODULE 1: DATABASE MANAGEMENT SYSTEMS

Introduction

In this module, we will study the following:

1. Database Systems
2. Functions and Components of a Database Management Systems
3. Master Data Management
4. Databases and Normalization
5. Entity Relationship Diagram and Relational Modeling
6. Enterprise Content Management (ECM)

Learning Objectives

At the end of the module, the students should be able to:

1. Understand database management systems, their major DBMS components and their functions;
2. Understand the concept of master data management (MDM);
3. Be able to model an application's data requirements using conceptual modeling tools like ER diagrams and design database schemas based on the conceptual model.
4. Describe formalized means of organizing and storing of documents and other content in an organization related to the organization's processes;

1.1. Database Systems

Learning Resource

Watch: Database Management Systems video lecture 01:00 to 03:27

Data is typically stored in physical or digital files. However, many factors need to be considered as the size of the data increases. How do you find and retrieve data? How many files do you need? How many disks do you need? How do you operate on the data? How do you allow concurrent access and modifications to the data?

What is a database?

A database is a computerized system that makes it easy to search, select and store information. Databases are used in many different places. Your school might use a database to store information about attendance or to store pupils' and teachers' contact information. A database like this will probably be protected with a password to make sure that people's personal information is kept safe. Your library might also use a database to keep track of which books are available and which are on loan.

What is a database management system (DBMS)?

A DBMS refers to the technology for creating and managing databases. Basically, DBMS is a software tool to organize (create, retrieve, update and manage) data in a database.

The main aim of a DBMS is to supply a way to store up and retrieve database information that is both convenient and efficient. By data, we mean known facts that can be recorded and that have embedded meaning. Normally people use software such as Microsoft EXCEL to store data in the form of database. A datum is a unit of data. Meaningful data combined to form information. Hence, information is interpreted data - data provided with semantics. MS. ACCESS is one of the most common examples of database management software.

Study Questions

Give examples of data that can you store in a database.

1.2. Functions and Components of a Database Management System

Learning Resource

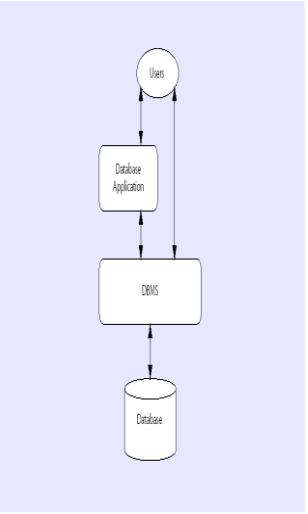
Watch: Database Management Systems video lecture 03:28 to 07:54

The DBMS should provide a convenient and efficient interface for storing, retrieving, updating data, and extracting useful information from the database. It should also

provide a clear and logical view of the process that manipulates the data. Aside from those main functions, a DBMS also has the following functionalities:

- Data Independence – it maintains the segregation between the program and the data.
- Concurrency Control – what happens when more than one person is accessing the data?
- Recovery Services – what if the system crashes and data is apparently lost?
- Utility Services – a DBMS also performs initialization and maintenance operations on a database

Components of a Database System



Users

- Database administrator (DBA) – the person responsible for all the data resources in an organization
- System developers – those who create the application programs that cater to the user requirements
- End-users – use the application program to accomplish their day-to-day tasks

Database Application – Computer programs that allow users to manipulate the data in a DBMS through a user-friendly interface

DBMS – decouples application programs from data. The database stores all its data in one location, thereby limiting data duplication**.** Examples include Access, Oracle, IBM's DB2, and SQL Server

Database – This is the space in the disk or computer where the data are actually stored.

Study Question

Give examples of database applications that you use. Hint: Social media applications use databases.

1.3. Master Data Management (MDM)

Learning Resource

Watch: Database Management Systems video lecture 07:55 to 10:13

- Master data is one single point of reference for every unique entity in the business.
- MDM establishes the rules, criteria, and details attached to the data so that everything in the business makes use of that single point of reference.
- The goal of master data management is to provide synchronization to the most critical pieces of data.

Core Functions of MDM

- Content – an MDM gives rules on how the data is represented as it must be consistent in all systems.
- Relationship – there must be a consistent definition of relationships and groupings.
- Security & Access – an MDM defines who has the authority to read or write to a data.
- Change Management – once data changes, how do we let others know that the data must be changed?
- Processing – This is what makes Master Data Management an MDM. It integrates all the first four functions together and executes them.

Study Question

Give an example of where Master Data Management can be used in a business.

1.4. Normalization

Learning Resource

Watch: Database Management Systems video lecture 10:14 to 15:40

Normalization is the process of reorganizing data in a database so that it meets two basic requirements:

1. There is no redundancy of data (all data is stored in only one place), and
2. data dependencies are logical (all related data items are stored together).

Normalization is important for many reasons, but chiefly because it allows databases to take up as little disk space as possible, resulting in increased performance.

Study Question

What are the consequences of non-normalized data?

1.5. Entity Relationship Model and Rational Modeling

Learning Resource

Watch: Database Management Systems video lecture 15:41 to 22:49

Entity Relationship Model

An entity-relationship model (ERM) is a theoretical and conceptual way of showing data relationships in software development. ERM is a database modeling technique that generates an abstract diagram or visual representation of a system's data that can be helpful in designing a relational database. These diagrams are known as entity-relationship diagrams, ER diagrams or ERDs.

Entity-relationship patterns were first proposed by Peter Pin-Shan Chen of the Massachusetts Institute of Technology (MIT) in 1976.

Entity Relationship Diagram

An entity-relationship diagram (ERD) is a data modeling technique that graphically illustrates an information system's entities and the relationships between those entities. An ERD is a conceptual and representational model of data used to represent the entity framework infrastructure.

The elements of an ERD are:

- Entities

- Relationships
- Attributes

Steps involved in creating an ERD include:

1. Identifying and defining the entities
2. Determining all interactions between the entities
3. Analyzing the nature of interactions/determining the cardinality of the relationships
4. Creating the ERD

Study Question

Identify the entities and the relationships in an online classroom system.

1.6. Enterprise Content Management

Learning Resource

Watch: Database Management Systems video lecture 22:50 to 25:18

Enterprise Content Management (ECM) is an organizational process methodology designed for complete content life cycle management. ECM content includes documents, graphics, email and video.

ECM is derived from electronic document management systems (EDMS) used during the late 1980s to early 1990s for smaller scale imaging and work flow. Today, ECM solutions employ a single software package that encompasses multiple enterprise divisions, including accounting, customer service and human resources (HR).

Study Question

Why is ECM important for business?

Learning Activity

LEARNING ACTIVITY

Quiz 1

1. Which of the following does not involve a Database Management System (DBMS)?
a. Calculator
b. Store Inventory System
c. Package Tracking System
d. all of the above
e. none of the above

2. Which of the following is true about DBMS?
a. It should provide convenient and efficient interface only for storing data.
b. All software are DBMS.
c. It is a collection of interrelated data including the software and hardware required to access the data.
d. all of the above
e. none of the above

3. True or False
End-users of a database system are more concerned on how to get data in a usable form rather than knowing how data is stored.

4. Which of the following is true about abstraction in a DBMS?
a. Security details may be considered in the View level.
b. The database administrator uses the Conceptual level.
c. In the Physical level data structures and file formats used by the system are described.
d. all of the above
e. none of the above

5. True or False
Each level is more specific in its description than the one before it.

6. Which level can the following description of a database be found?

```
type item = record
        name : string[80];
        price : real;
        quantity : integer;
end;
```

a. Physical Level
b. Conceptual Level
c. View Level
d. Model Level
e. none of the above

7. True or False
All types of users in a DBMS has the same View level.

8. True or False
Physical level is for programmers. Conceptual level is for database administrators. View level is for end users.

9. True or False
The database administrator knows what data goes in the system and the relationships that exist among those data.

10. True or False
A standard file-processing system is a very good substitute for a DBMS.

11. Suppose we want to make a database of students. Which of the following is false?
a. Each student in the database is referred to as a record.
b. The name of the student is another record.
c. The student number is a field in the student record.
d. all of the above
e. none of the above

12. Which of the following describes an entity.
a. An entity is a thing in the modeled world.
b. An entity pertains to a real-world object, person, place such as a student in a school DBMS.
c. An entity is something which is perceived to exist
d. all of the above
e. none of the above

13. True or False
Data models are used to organize a database structure.

14. True or False
Attributes may be referred to as details of a particular entity.

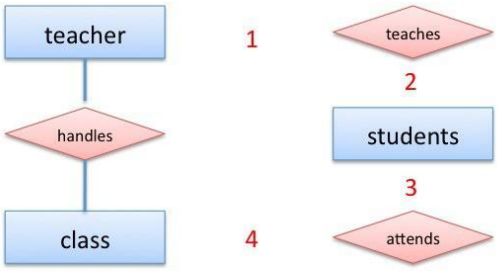
15. Which of the following describes a Superkey
a. Attributes that distinguish entities.
b. A set of attributes that uniquely identify an entity.
c. It is also known as the primary key.
d. a and b
e. b and c

16. True or False
Suppose that student and teacher are in a one-to-many relationship. This means that one teacher can have many students but a student can only have one teacher.

17. In the UPOU system, what is the relationship that exists between students and teachers?
a. one-to-one
b. one-to-many
c. many-to-one
d. many-to-many
e. none of the above

18. In the _____, relationships between records are organized into tables wherein each record has relationships with other records which is stored in a table containing the primary keys of each entity.
a. Hierarchical Model
b. Network Model
c. Conceptual Model
d. Tree Model
e. none of the above

19. Which relationships are appropriate for the following diagram. (C)



a) 1 ← 2 — 3 ← 4 ←
b) 1 — 2 → 3 — 4 —
c) 1 — 2 — 3 — 4 —
e) 1 — 2 — 3 → 4 ←

e) None of the above

20. True or False
In the hierarchical data model, tree structures are not used to model data.

21. True or False
The hierarchical can effectively model many-to-one and many-to-many relationships, one-to-one and one-to-many relationships result in redundant records.

22. True or False
The network model addresses the problem of redundancy in the hierarchical model of representing many-to-one and many-to-many relationships.

23. True or False
Links in the network and hierarchical model denote file positions of the next record.

24). Which of the following does not describe an entity-relationship-diagram
a. They represent relational databases
b. Each entity is represented by boxes and relations by diamonds
c. Lines show the existence and types of relationships
d. all of the above
e. none of the above

25. True or False.
In the relational entities, attributes, keys are stored in tables.

Assignment 1

Pick a task which could be done faster or more efficiently using a DBMS then create an Entity-Relationship Diagram for that task.

References

https://www.techopedia.com/definition/1015/enterprise-content-management-ecm
https://www.techopedia.com/definition/7057/entity-relationship-model-er-model
https://www.techopedia.com/definition/1200/entity-relationship-diagram-erd
https://www.techopedia.com/definition/840/master-data-management-mdm
https://www.techopedia.com/definition/1221/normalization

MODULE 2: DATA WAREHOUSING

Introduction

In this module, we will study the following:

1. Data Warehouses, Data Marts, and Operational Data Source
2. Alternate Data Warehousing Architecture
3. The Kimball Lifecycle
4. ETL (Extraction, Transformation, Loading)
5. Using the Data Warehouse for Business Intelligence

Learning Objectives

At the end of the module, the students should be able to:

1. Describe the importance of data warehouses (DW or DWH) for reporting and data analysis and understand the differences from operational data source (ODS);
2. Understand the relationship of data warehouses and business intelligence

Self Study

Watch: Data Warehousing video 00:35 to 08:29
Study: Getting Started with Data Warehousing Chapter 1 pp. 17-29

2.1. Data Warehouses, Data Marts and Operational Data Source

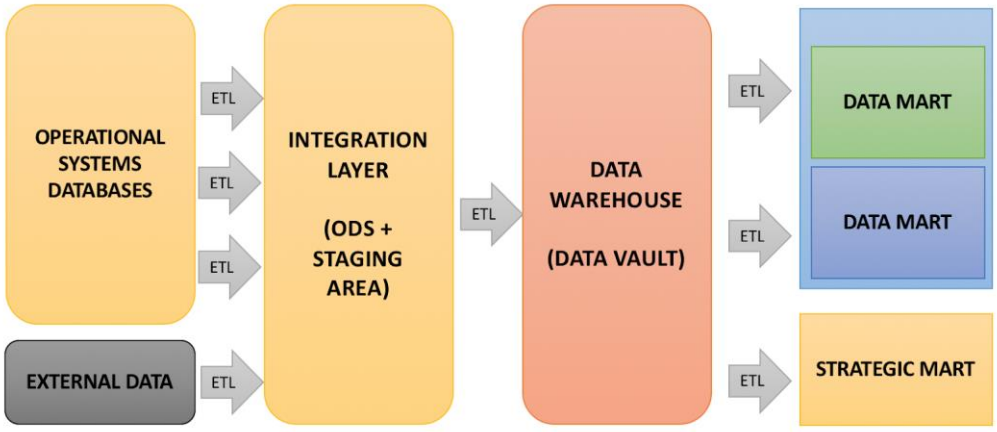
Data Warehouse

- A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format
- A collection of integrated, subject-oriented databases where each unit of data is non-volatile and relevant to some moment in time.
- It is meant to provide aggregate data for decision making.

Study Question

Where do you think can data warehouses be used?

Common Data Warehouse Architecture



Components

- **Operational Systems Databases** – data from different databases within an organization
- **Integration Layer** – staging area where data are cleaned up and formatted
- **Data Warehouse** – data vault where data are stored in 2nd normal form
- **Data Mart** – a subset of the data warehouse that support the requirements of a particular department or business function; It is used to provide specialized and strategic answers for specific people.

2.2. Alternate Data Warehousing Architectures

Learning Resource

Watch: Data Warehousing video 08:30 to 10:09

1. **Independent Data Marts** are the simplest and least costly architecture alternative. The data marts are developed independently of each other. Each then serves the needs of the individual units.
2. **Data Mart Bus Architecture** involves marts that are linked together using a middleware.
3. **Hub-and-Spoke Architecture** focuses on building a scalable and maintainable infrastructure. It allows easy customization of user instances and reports.
4. **Centralized Data Warehouse** is similar to hub and spoke except that there are no dependent data marts. Instead, it has a very large data warehouse that serves all the needs of the organizational units.
5. **Federated Data Warehouse** uses all the possible ways to integrate analytical resources from multiple sources to meet changing needs or business conditions. It involves integrating disparate

Study Questions

Which data warehousing architecture do you think is most suitable for your organization?

2.3. The Kimball Lifecycle

Learning Resource

Watch: Data Warehousing video 10:10 to 12:40

The Kimball Lifecycle

- Formerly known as Business Dimensional (BI) Lifecycle but was renamed Kimball Lifecycle in 2008.
- Focus on adding business value across the enterprise
- Dimensionally structures the data that's delivered to the business
- Uses iterations and increments in a manageable lifecycle
- Steps:
    1. Program/Project Planning and Management – In this phase, plans for three streams (technology track, data track, and application track) are created simultaneously.
    2. Deployment – After planning, the plans are deployed.
    3. Maintenance – While being deployed, it must be maintained as well.
    4. Growth – In order for the iteration to grow, we go back to the planning stage and the process continues.

Study Question

Which part of the Kimball Lifecycle could you be potentially involved in?

2.4. ETL (Extraction, Transformation, Loading)

Learning Resource

Study: Getting Started with Data Warehousing Chapter 4 pp. 55-61

Extract transform load (ETL) is the process of extraction, transformation and loading during database use, but particularly during data storage use. It includes the following sub-processes:

- Data Extraction – pulling data from different source systems
- Data Transformation - transforming data into an understandable format, where data is typically stored together with an error detection and correction code to meet operational needs
- Data Loading - transmitting and loading data to the receiving end

Study Questions

What are the different data sources you have in your business?

2.5. Using the Data Warehouse for Business Intelligence

Learning Resource

What are the different data sources you have in your business?

Business Intelligence – refers to the applications and technologies used to gather, provide access to, and analyze data and information about a company's operation.

Data Warehouse – a repository for a company's historical data.

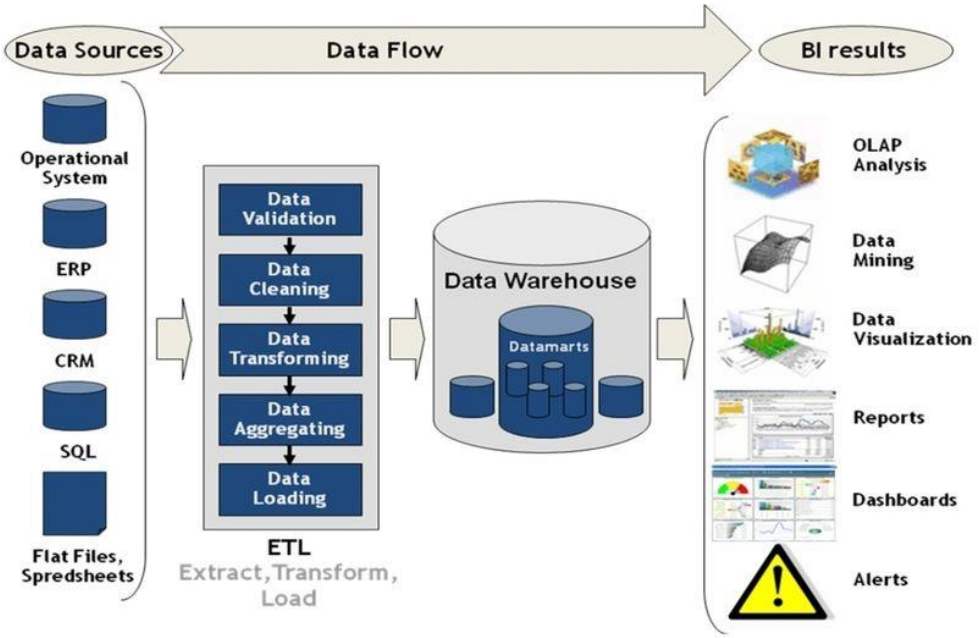


Figure 3. The Business Intelligence Data Flow (source: http://tommytoy.typepad.com/.a/6a0133f3a4072c970b01b7c7e03225970b-popup)

In Figure 3, we can see that a data warehouse makes a BI solution possible.

Learning Activity
Quiz 2

LEARNING ACTIVITY

Quiz 2

1. What does OLTP stand for?
A. Online Travel Processing
B. Online Travel Planning
C. Online Transactional Processing
D. Offline Transactional Processing

2. What does OLAP stand for?
A. Online Analytical Processing
B. Online Analytical Programming
C. Offline Analysis and Programming
D. Online Arithmetic Programming

3. Which one of the following is/are true for an OLTP system?
A. Tuned for quick reads
B. Tuned for quick inserts, updates and deletes
C. Need not require backup and recovery
D. Contain historical data in huge tables

4. Which one of the following is/are true for an OLAP system?
A. Tuned for quick reads
B. Tuned for quick inserts, updates and deletes
C. Need not require backup and recovery
D. Contain historical data in huge tables

5. True or False
OLAP systems are used as data stores for real-time applications.

6. True or False
In a transactional system, data is normalized to the highest form possible.

7. True or False
Data is generally extracted from an OLAP system, transformed and then loaded into an OLTP system

8. True or False
An OLAP structure is the basic architectural block of a Data Warehouse.

9. True or False
In an analytic system, data is normalized to the fastest form possible.

10. What does ETL stand for?

11. Yes or No. Can all ETL operations be achieved using SQL statement only?

12. Yes or No. To connect to multiple data sources, are multiple ETL tools required?

13. Yes or No. Can a flat file (comma delimited) also be a valid data source and destination?

14. Yes or No. Can data be loaded into a single table via multiple parallel loads?

15. Yes or No. Can ETL tools read and write streaming data?

16. What is the storage space used during data extraction called?
A. Temporary area
B. Temporary storage
C. Staging area
D. Staging storage

17. Which following operation does data transformation include?
A. Format conversion
B. Data size conversion
C. Data type conversion
D. NULL value handling
E. All of above

18. For high performance data loading in partitioned environment, it should be executed in
A. Sequential mode
B. Parallel mode
C. Mixed mode

Assignment 2

Design a guidance counseling appointment system/architecture and depict it using a block diagram, which consists of the following components:
A Web interface that accepts online data from the user in form of new reservation request, reservation modification request and reservation cancellation requests.
A real-time transactional system, which stores, retrieves and updates the live reservation data.
C. An ETL system that extracts data from a data source in real-time, does some transformation and loads into target data warehouse.
D. A Data Warehouse where the transformed data is loaded.
E. A reporting interface, which is used to present the analytic reports to the end-user.

REFERENCES

1. Data Warehousing video by Mari Anjeli Crisanto
2. Getting Started with Data Warehousing (eBook by IBM) https://goo.gl/83u4Nz
3. https://www.techopedia.com/definition/345/business-intelligence-bi
4. https://www.techopedia.com/definition/24170/extract-transform-load-etl
5. http://www.itprotoday.com/microsoft-sql-server/data-warehousing-foundation-bi

MODULE 3: KNOWLEDGE DISCOVERY
Introduction

In this module, we will study the following:

1. Knowledge Discovery in Databases (KDD)
2. Cross-Industry Standard Process for Data Mining (CRISP-DM)
3. Data Mining Methods
4. Programming Languages/Tools

Learning Objectives

At the end of the module, the students should be able to:

1. Describe the process of data discovery and data patterns in large data sets;
2. Understand various methods related to intersection of artificial intelligence, machine learning, statistics, and database systems;
3. Describe data inference considerations, interestingness metrics, complexity considerations; and
4. Understand various techniques used for post-processing of discovered structures and visualization.

Self Study

Watch: Knowledge Discovery video lecture

3.1. Knowledge Discovery in Databases (KDD)

Learning Resource

Study: Predictive Analytics.pptx (Hsin-Chang Yang) pp. 6 – 34.

There is currently a deluge of data but it is a challenge to make sense of them as they are from different forms and are in different formats. Knowledge Discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. It is tagged by different terms such as data mining, knowledge extraction, information harvesting, data archaeology, and data pattern processing.

Goals of Knowledge Discovery

1. Verification – verify a specific hypothesis

2. Discovery – discovery knowledge which is new, correct, and potentially useful

KDD Process

Figure 4. The KDD Process

1. Domain Understanding – develop an understanding of the application domain and the relevant prior knowledge; Identify the goal of the KDD process from the customer's viewpoint.
2. Data Selection – select a data set on which discovery is to be performed.
3. Data Pre-processing – remove noise if appropriate; collect necessary information to model or account for noise; decide on strategies for handling missing data fields; account for time-sequence information and known changes.
4. Transformation – convert data into a form that is appropriate for data mining tasks.
5. Data Mining – match the goals of the project to a data mining method; Choose data mining algorithms and methods (i.e. clustering, classification, regression); Search for patterns of interest.
6. Evaluation – interpret mined patters through visualization and reporting; Evaluate results.

Study Questions

Think of examples where you can apply knowledge discovery.

3.2. CRoss-Industry Standard Process for Data Mining (CRISP-DM)

Learning Resource

Watch: CRISP-DM video lecture

CRISP-DM is another data mining process model which is the standard model used in the industry. It involves the following processes.

1. Business Understanding – identify the project objectives, requirements, and the definition of the data mining problem

2. Data Understanding – perform initial data collection and familiarization and as well as identify data quality problems because the data are the ones that the your problem
3. Data Preparation – pre-process data and fit them into the model that will be used to solve the problem
4. Modeling – select modeling technique, generate test design, build model and assess model
5. Evaluation – identify whether or not the output of that the model generated solves the problem. If it does not, the redefine the model.
6. Deployment – deploy the model such that end users within the organization will be able to reap the benefits of this data mining solution

Study Question

Which process model do you think is more appropriate for your organization? CRISP-DM or KDD?

3.3. Data Mining Methods

Learning Resource

Study: Predictive Analytics.pptx (Hsin-Chang Yang) pp. 35 – 67.

Data Mining

- Discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
- Involves repeated iterative application of particular data-mining methods.
- Involves fitting models to, or determining patterns from, observed data.

Some Methods

1. Clustering – identifying groups/classes in data which are similar to each other
2. Classification – identifying which category an observation belongs to
3. Regression – identifying the extent of relationships among variables
4. Anomaly Detection – identification of unusual patterns, outliers, which help us in understanding the variation in data
5. Association Rule Mining – discovering association patterns among variables

Study Questions

Which method do you think is most appropriate for your data mining task?

3.3. Programming Languages/Tools

What are the programming languages or tools you need to know to perform knowledge discovery or data mining tasks? Here are some examples.

1. R – a language and environment for statistical computing and graphics. It is the most popular tool used for data mining.
2. Python – a general purpose programming language with many libraries for data analysis.
3. Microsoft Excel – an office tool, which can be easily used by anyone. It has add-ons for data analysis.
4. SQL – a language used in programming and managing a database management system

Learning Resource

Which tool(s) would you most likely use for data mining?

LEARNING ACTIVITY

Quiz 3

1. True or False
Data mining is the application of specific algorithms for extracting patterns from data.

2. True or False
Data warehousing helps set the stage for KDD in through data cleaning and data access.

3. What are the steps in KDD?

4. Which of the following is not a data mining method?
A. Calculation
B. Classification
C. Regression
D. Clustering
E. Summarization

5. Why do we need KDD?

6. Which of the following terms is an alternate name for Data Mining and KDD?
A. Knowledge extraction
B. Information discovery
C. Data archaeology
D. All of the above
E. None of the above

7. Give an application of KDD

8. True or False
The basic problem addressed by the KDD process is
one of mapping low-level data into other forms that might be more compact, more abstract, or more useful.

9. True or False
KDD is interdisciplinary in nature.

10. True or False
The two major goals of KDD are verification and discovery.

Assignment 3

Based on the KDD process create a KDD plan for an ice cream business.
Guide questions:
1. What do you want to discover from the data?
2. What are the information that you need to discover those?
3. Where will you get your information?
4. What are the methods you will use?
5. How will you put the discovered knowledge into action?

REFERENCES

- Knowledge Discovery video lecture by Ria Mae Borromeo
- CRISP-DM video lecture by Eugene Rex Jalao
- Tools in Data Mining video lecture by Eugene Rex Jalao
- Predictive Analytics lecture slides by Hsin-Chang Yang

MODULE 4: ENTERPRISE DATA MANAGEMENT ISSUES

Introduction

In this module, we will talk about data security privacy, online frauds, and ethical norms in data management..

Learning Objectives

At the end of the module, the students should be able to:

1. Describe the need and policy around data security and privacy (information security) and techniques to restrict information from unauthorized access, use, disclosure, disruption, modification, perusal, inspection, recording or destruction;
2. Describe online fraud and their consequences and understand predictive analytics for detection of fraudulent activities; and
3. Develop an awareness of the ethical norms as required under policies and applicable laws governing confidentiality and non-disclosure of data/information/documents and proper conduct in the learning process and application of business analytics.

Self Study

Watch: Enterprise Data Management Issues video lecture

4.1. Data Security

Data security refers to protective digital privacy measures that are applied to prevent unauthorized access to computers, databases and websites. Data security also protects data from corruption. Data security is an essential aspect of IT for organizations of every size and type.

Study Questions

What are your current practices do to secure your data and prevent unauthorized access?

4.2. Data Privacy

Data privacy is the privacy of personal information and usually relates to personal data stored on computer systems. The need to maintain information privacy is applicable to collected personal information, such as medical records, financial data, criminal records, political records, business related information or website data.

Data Privacy Act (Republic Act. No. 10173, Ch. 1, Sec. 2)

- Implemented to protect the fundamental human right of privacy and communication while ensuring free flow of information to promote innovation and growth.
- Specifies that consent is needed before the collection of all personal data. The data subject must also be informed of the extent to which their personal information will be processed.

Study Questions

How does the data privacy act affect enterprise database systems?

4.3. Online Fraud

Online fraud is the use of Internet services or software with Internet access to defraud victims or to otherwise take advantage of them.
Examples:

- Business Email Compromise (BEC) – Using legitimate business email accounts through social engineering or intrusion techniques to conduct unauthorized transfer of funds.
- Data Breach – Sensitive, protected or confidential information is released to an untrusted malicious environment.
- Denial of Service – An authorized user is denied access to a system or network.
- Business fraud, credit card fraud, internet auction fraud, non-delivery of merchandise, etc.

Study Question

How can an enterprise database system reduce the likelihood of malicious users from stealing information from their system?

4.4. **Ethical Norms**

Learning Resource

Watch: Opportunities and Ethics in Data Warehousing video lecture

There are practices in data warehousing that are useful and done with good intent but may be unethical. For example, you have missing data for a person and you want to complete your database. You know that you can find that missing data in another database that you are not authorized to access. However, you have a colleague who is authorized to access it. Is it ethical to ask your colleague? As a colleague, is it ethical to divulge the data?

The following guidelines were created to ensure ethical practices in data warehousing.

- Develop service level agreements with end users that define who has access to what levels of information.
- Have end-users involved in defining the ethical standards of use for the data that will be delivered.
- Define the bounds around the integration efforts of public data, where it will be integrated and where it will not – so as to avoid conflicts of interest.
- Do not use “live” or real data for testing purposes – or lock down the test environment; too often test environments are left wide-open and accessible to too many individuals.
- Define where, how, and who will be using Data Mining – restrict the mining efforts to specific sets of information. Build a notification system to monitor data mining usage.

- Allow customers to “block” the integration of their own information (this one is questionable) depending on if the customer information after integration will be made available on the web.
- Remember that any efforts made are still subject to governmental laws. What laws do we have right now concerned with data privacy? Note that future laws could also be developed and we must be aware of those.

Study Question

What other issues could businesses around you face in terms of security, privacy, online frauds, and ethical norms?

LEARNING ACTIVITY

Quiz 4

1. True or False
Data security refers to protective digital privacy measures that are applied to prevent unauthorized access to computers, databases and websites

2. True or False
The need to maintain information privacy is applicable to collected personal information, such as medical records, financial data, criminal records, political records, business related information or website data.

3. True or False
The Republic Act. No. 10173, also known as the Data Privacy Act was implemented to protect the fundamental human right of privacy and communication while ensuring free flow of information to promote innovation and growth.

4. True or False
Business email compromise, Data breach and denial of service are examples of online fraud.

1. How does the data privacy act affect knowledge discovery?
2.

Assignment 4

List down at least 3 security threats in a university's learning management system.
For each security threat, design an action plan to prevent such threat.

REFERENCES

- Enterprise Database Management Issues video lecture by Mari Anjeli Crisanto
- Fundamentals of Data Warehousing (Opportunities and Ethics) video lecture by Mari Anjeli Crisanto
- https://www.techopedia.com/definition/26464/data-security
- https://www.techopedia.com/definition/10380/information-privacy
- https://www.fbi.gov/scams-and-safety/common-fraud-schemes/internet-fraud

UNIVERSITY OF THE PHILIPPINES OPEN UNIVERSITY

BAFEDM 2 – Fundamentals of Enterprise Data Management

FINAL EXAM

1. Enumerate 5 examples of data that can you store in a database.

2. Enumerate 5 examples of database applications that you use. Hint: Social media applications use databases.

3. Give an example of where Master Data Management can be used in a business.

4. List down 3 consequences of non-normalized data?

5. Identify the entities and the relationships in an online classroom system.

6. Why is ECM important for business?

7. Enumerate 5 possible applications of data warehouses.

8. Enumerate 5 different data sources you have in your business.

9. Enumerate 5 BI results does your company need.

10. Enumerate 5 examples where you can apply knowledge discovery.

11. Which process model do you think is more appropriate for your organization? CRISP-DM or KDD? Why?

12. List at least 3 practices you could do to secure your data and prevent unauthorized access?

13. How does the data privacy act affect enterprise database systems?

14. How can an enterprise database system reduce the likelihood of malicious users from stealing information from their system?

15. What other issues could businesses around you face in terms of security, privacy, online frauds, and ethical norms?