1. Introduction to Data Analysis and Modelling

a) Examples of database applications: Supermarket, Library, University, Health clinic, Travel agency

b) Basic Database Terminology:

i. **Data:** Basic facts, figures etc. Eg. item like "24" by itself doesn't indicate that it's a persons age
ii. **Information:** Prlocessed data in a useful structure enabling decitions to be made. Eg. Age of a person
iii. **Environment:** (World of reality) Organisation of place where the databse is designed and , developed. Eg. Library, University etc
iv. **Entity:** An object which is unique and which can be identified in a specific environment. Eg. Books, Authors, Publishers in a Library Environment.
v. **Attributes:** (Field/data items) Characteristics of an entity. ISBN No, Name, Author of a book entity.
vi. **Record:** A group of attributes needed by a particular entity. Eg. AA618829, History, Fajar Bakti is one record.
vii. **File:** A group of records required by an organisation. Eg. Book file, Publishers file etc.
viii. **Database:** A group of files required by an organisation. Eg. Catalogue Database, Environment is library, files are books, authors, publishers, suppliers.. etc

c) Ordinary File System: The system in storing the data in computers before the concept of database was introduced.

i. **In an Organisation:** Filing systems were used to store all internal and external informatin related to specific projects, products, tasks, customers or workers etc. There are many files involved and for security, each file is labelled and kept locked in different cabinets or safe locations.
ii. **At home:** we have our own filing system for such things as bank statements, bills, favourite recipes etc. when we want to find an info, we open our filing system and search until it is obtained. We may also use an index system to make fast searching.
iii. **When computers were first used:** the records were stored in electronic files using similar principles as a filing system. It was obviously faaster than the manual system, but still weaknesses were there in processing of the filing system

d) Weaknesses of Ordinary Filing System

i. **Duplication of data:** Uncentralised data causes uncontrolled duplication. This causes **Excess Data** as data is repeated**, Unorganised Data** as input data is not updated in other files and cause doubt regarding validity and authenticity, and lead to **Weak Data Control** as there will be departments with incomplete data.
ii. **Data Seperation:** Storing files seperately causes difficulty in acquiring information when combinations of two or more files are needed.
iii. **Data format Dependency:** Changes in data format requires a new program to be written to enter or process data**.**
iv. **File Incompatibility:** Files stored in different formats cause incompatibility and are difficult to update and process.
v. **Difficulty in Presenting Organisational Data:** Difficulty in making connections between records stored in each department in an organisation. So, it is difficult to portray the operations of an organisation in complete and accurate manner.

e) Concept of Database: was produced as a result of the need to construct a multi data processing system for organisations which need a data processing system to store data related to their daily activities and interactions.

i. **Database:** A collection of related data shared by various categories of users to fulfill the information needs and requrements of an organisation

ii. **Database Application:** An application program written in a specific programming language.
iii. **Database Management System (DBMS)**: A software system which enables users to define, create, organise, update, manage and administer databases, and also control access to data in databases.
iv. **Environmental componends of DBMS:**
    1. **Hardware**: All the physical instruments of the system. Eg Microcomputer, Computer apparatus etc
    2. **Software**: Collection of computer program applications in a database system, the OS, application program and any form of networking software is applied.
    3. **Data**: Raw data stored in files and the results of analysis related to the organisation consisting of entities, attributes and their relationships.
    4. **Proceedure**: Rules and instructions that govern the design, programming and the application of the database as well as DBMS.
    5. **Users**:
        a. **Common End Users:** Customers who use the DB system to obtain information or to carryout their duties. They also use the application programs to carryout routine organisational operations.
        b. **Upper End Users**: Users trained in the use of online query language. They know the DB structure and facilities provided by the DBMS
        c. **Database Designers**: Individuals with technical knowledge of DB. Responsibilities include designing of logical models and the development of the physical DB.
        d. **Application Programmer**: People whos duty is to write programs to implement specific DB functions according to the specifications provided by users and DB admins.
        e. **Database Administrator** (DBA): Individuals whos responsibilities are to plan, create, design, facilitate and control DB applications to meet user requirements and needs.
v. **Types of Database Management Systems:** Depends on the situation
    1. **Number of Users:** Solitary and Multi user
    2. **Location:** Centralised, Dispersed
    3. **Application:** Transaction/Production, Decition Support
vi. **Strengths and Weaknesses of DBMS**
    1. **Advantages:**
        a. **Control** over excess data, as data is stored in one location
        b. Better **organisaton** of data when there is control.
        c. More **information** can be produced from same amount of data.
        d. Better **sharing** of data amongst individuals or departments.
        e. Data **standardisation** can be carried out due to this sharing concept and management of centralised data.
        f. Better data **integrity**, **restoration**, **support**, **confidentiality** and **security** as password protection and **concurrency** control.
        g. **Economical** from the manpower, storage and cost point of view
        h. **Less conflict** among users as data is managed by DBMS
        i. Better **data readiness** as DBMS provides facilities for queries and ease in making reports
        j. Increase in **productivity** as there is less programming to be done compared to method in ordinary filing system
    2. **Disadvantages:**
        a. Difficulty in the development process will affect the **performance** of the system.
        b. Increase in the DBMS **size** involves a lot of storage
        c. High **cost** of **establishing** DBMS and preparation of other hardware
        d. High **cost** of **converting** and operating from a **manual system** or an ordinary filing system
        e. Higher **effects** of **failure** because the users and applications rely too much on DBMS

2. Database Environment

a) Database Architecture (ANSI/SPARC): Proposes 3 level database architecture with the objective of seperating each user's view of the physical database presentation.

i. **External Level:** Represents the **user's view** of the database where only a **part** of the database is suitable for the requirement of each user. End users interact with this level and **cannot view the other levels**. Some views are:
   1. **Accesibility** to certain data in specific view are restricted to users. Eg confidential files are allowed access to only cetian individuals.
   2. Same data **presented** differently. Eg change of date format etc
   3. Some views only show **virtual data**, only when needed. Eg Age can be virtually derived when needed from Date of Birth stored in DB.

ii. **Conceptual Level:** The middle level between the enternal and internal levels. **Represents the logical structure of the entire DB** as seen by a DBA. It is a complete view of the data requirements of the organisation yet it is free from any storage consideration. This level contains:
   1. All **entities**, **attributes** and **relationships** between data in the database
   2. The **constrains** of the data
   3. **Semantic** information of the data
   4. **Security** and **integrity** information

iii. **Internal (Physical) Level:** Lowest level, closest to the DB, this explains how data is stored in the DB. ie:
   1. **Storage** space required for data and indexes
   2. **Description** of record storage (with storage sizes)
   3. Record **positions**
   4. Data **compression** and **encryption** techniques

iv. **Scheming and Mapping:** Scheme is an overall description of a DB. There are **3 types of schemes** which depend on the level of the ANSI/SPARC architecture. **External** Scheme, **Conceptual** Scheme and **Internal** Scheme. If stored structure of scheme is changed, the mapping process must also be changed.
   There are 2 types of mapping:
   1. **External/Conceptual Scheme Mapping:** Object mapping at conceptual and external levels
   2. **Conceptual/Internal Scheme Mapping:** Object mapping at conceptual and internal levels

v. **Data Independence:** Major objective of 3-level architecture is to provide data independence. The uppermost level is unaffected by chamges in the lowest level. There are 2 types of data indeppendence:
   1. **Logical Data independence:** The immunity of the external scheme towards changes in the conceptual scheme. Eg Changes to the conceptual scheme such as addition of a new entity need to be clear without changing the existing external scheme. Users involved need to be informed of the changes.
   2. **Physical Data Independence:** The immunity of the conceptual scheme towards changes in the internal scheme. Eg changes in the internal scheme like use of different storage structures should be possible without changing conceptual or exteral schemes.

b) Database Language:

i. **Data Definition Language (DDL):** Used to determine the DB schemes, Changes data produces by an application program or by the terminal from the conceptual level (programmer view) to the physical level (manner of data storage). The result is a table stored in a special dictionary called the **data dictionary** which stores the detailed information (metadata) on data in a DB.

ii. **Benefits of Data Dictionary are:**
   1. Easier to control data as information are stored centrally.
   2. Meaning of the data can be defined clearly
   3. Because of this, communication and undestanding between users is easier
   4. Excess lack of coordination, hononyms and synonyms can be identified and avoided.
   5. Each change to the DB can be easily recorded

6. Effect of any change being made can be evaluated before actually made as DM manager has complete record of data.

iii. **Data Manipulation Language (DML):** Used to read and update the DB, Provides a set of operations to support data manipulation operations on data in the database. They include; addition of new data, updating, access, deletion. Classes of DML are:
   1. **Procedural language:** allows users to inform the data system what data is needed and how to access
   2. **Non-Procedural Language:** allows users to explain what data is needed without determining how the data can be accessed.

iv. **Fourth Generation Language (4GL):** A data sub-language. Easier to learn and spplied compated to the 3GL. Commands are brief and non-procedural. Characteristics are:
   1. Presentation language such as query language
   2. Special language such as DB language
   3. Production of applications which define, add, update, access data from the DB to build applications. Eg Application Producer, Report producer, Form Producer, Graphic Producer and Query language
   4. High level language to produce application codes

c) Data Model and Conceptual Modelling:

i. **Data Model:** A collection of concepts, constrains and integrity rules that describe the situation, relationship and constrains of data in a particular organisation. 3 parts of data model:
   1. **Scheme**: Explains the contents of data structure
   2. **Manipulation**: Explains the type of data administration and how it's executed.
   3. **Data Control and Validation**: States the rules of integrity and constrains related to data.

ii. **Types of Data Models:**
   1. **Object Based Data Model:** Basic Concepts are Entity, Attribute and Relationship. 3 Data models are:
      a. **Entity Relationship Model:** Most often used model, based on special symbols to label entity, attribute and relationships
      b. **Semantic Object Model:** A collection of attribute names which can explain a specific thing with certainty. 3 types of attributes, Brief Attribute, Mixed attribute and Semantic Object Attribute.
      c. **Object Oriented Model:** Extends entity and attribute definitions to include the actions associated with it.
   2. **Record Based Data Model:** Closely resembles ordinary filing concept. 3 Record based models are:
      a. **Hierarchy Data Model:** Father and Son (One to Many) Tree structure consisting of nodes with the root node at the upper most level and the leaf nodes at the lowest level. (Each node is a record in DB)
      b. **Network Data Model:** Improvements to the hierarchy model, 3 main components; data elements, records, set of one-to-many relationships.
      c. **Relationship Data Model:** Represented as a table with rows (record) and columns (attribute).
   3. **Functions of DBMS:** Codd (1982) listed these functions:
      a. Data Storage, Access & Update
      b. User Accessible Catalogue
      c. Transaction Support
      d. Concurrency Control Service
      e. Repair Service
      f. Authorisation Service
      g. Support for Data Communication
      h. Integrity Services
      i. Services to promote data independence
      j. Utility Service

iii. **Multi-User DBMS Architecture**

1. **Teleprocessing:** is the standard architecture for a multi-user system. It consists of one computer with a single processing unit (where all processing takes place), and a number of terminals (which are joined to the central computer). Terminals send messages via communication control sub-system of the OS to the users programs according to the order through utilizing DBMS services.
2. **File Server:** DBMS in each workstation send request for the file server for data, which acts as a shared hard disk drive. Its disadvantages are:
    a. Large amounts of network traffic
    b. Full copy of DBMS needed for each work station
    c. Co-operation, retoration and integrity control are complex as several DBMS can access a file at the same time.
3. **Client Server:** In this architecture, the DB and DBMS are located in a server (computer with high processing capabilities). Apart from DB, aother resources can be shared as well, like printer, scanner. Client request DB access, server provides services for DB management and communication. Suitable for small and medium workgroups

3. Database Life Cycle

a) Database Development Life Cycle (DDLC): The life cycle that governs the database system which is a component of information system. It contains 6 phases:

i. **Database Planning:**
1. Analyze current situation of the organisation
2. Defining problems and constrains
3. Defining Objective
4. Defining scope and Boundary
5. Feasibility Research

ii. **Database Design:**
1. Conceptual Design: Data Modeling, Normalization, Data Model Validation.
2. Logical Design: Translating the model into a DBMS compatible data representation form.
3. Physical Design: Physical aspects of data in storage and performance of the system.

iii. **Implementation and Downloading:**
1. Setting up a DBMS
2. Creating a Database
3. Data downloading/changing

iv. **Testing and Evaluation:**
1. Database Testing and Application Proceedures: integration test, unit test, system test.
2. Database Evaluation on performance and system security

v. **Operation:**
1. Direct Transition
2. Parallel Transition
3. Pioneer Transition (start using complete system in one section or department)
4. Staggered Transition (phased/staged)

vi. **Maintenance and Evolution:** Maintenance is the process of monitoring and maintaining the performance of the system. Maintaining is performed to:
1. Correct errors in the system, system competence and other factors previously not addressed
2. Changes of organisational rules or policy will require adjustments to be made to the system
3. System needs to be upgraded to the latest version to support the requirements of users and the organisation.
(However, these changes must not affect other users)

4. Entity Relationship Model

a) **ER-Model:** Used to construct conceptual data model, representing the structure and constraints of a database which is not dependent on a software (like DBMS) or any data model to be used to implement a database. Its basic elements are Entities, Attributes and Relationships.

b) **Entity:** An object of the realworld which can store data and can be defined clearly. Eg Customer, Furniture etc.

c) **Attribute:** Description or characteristics of an entity (what differentiates one entity from another). There are several types of attributes:

i. **Simple Attributes:** Simple attributes has only one component, is independent and cannot be broken up.
ii. **Composite Attributes:** Composite Attributes has many components each one existing independently.
iii. **Solitary Value Attributes:** Contains one value
iv. **Multiple Value Attributes**: Contains many values. Eg Phone Numbers
v. **Derived Attribute:** Where its value is derived from another attribute or set of attributes, Eg age.

d) **Attribute Domain:** A set of values for an attribute. Eg Attribute Domain for Staff_Number is integer (1-30). Integer types are; Character, Numeric, Date

e) **Null Value:** An attribute value which does not exist, unknown at times or not related. (It is not zero)

f) **Key:** is one or several attributes which can differentiate the entities they describe. Its value must be unique and must not be null. Eg ID_Number

g) **Relationship:** The link between entities. There are 3 types of relationships:

i. **Unary Relationship** (Recursive): Is a relationship involving only one entity. Eg Staff manages Staff
ii. **Binary Relationship:** Relationship between two entities. Eg Student registered_to course
iii. **Ternary Relationship:** Simultaneous relationship between three entities. Eg Sponsor offer Scholarship to Student

h) **Relationship Attribute:** Like entity attributes, relations also can have attributes which describes the relationship. Eg Treatment relationship between Patient and Doctors can contain Type_of_treatment, type_of_ailment and medicine.

i) **Relationship Cardinality:** describes the number of relationships between one entity to other entities. There are 3 types of relationship cardinality:

i. **One to One** Relationships (1:1)
ii. **One to Many** Relationships (1:M)
iii. **Many to Many** Relationships (M:N)

j) **Relationship Participation:** Entity participation in a relationship can be either compulsory or optional.

k) **Guidelines and Steps in Constructing an ER Model:**

i. Do not insert the **System Environment** as an entity or attribute etc
ii. **Attribute** and **key** for a particular environment are not necessarily same for others. Eg In a library, Burrower may be student or lecturer.
iii. An entity must contain **description**. Objects with only one characteristic are attributes, not entities.
iv. **Convert** multiple value attributes to entities.
v. Two entities can have more than one **relationship**
vi. Apply **the top down approach** in modelling entity and main relationship with sets of limited attributes.

Follow the steps below for top down approach:

1. Determine **entities** and **relationships** between them. Start with the **main entity** followed by others
2. Determine the **attributes** related to the **entities**
3. Determine the **attributes** related to the **relationships** (if any)
4. Choose the **keys** for the **entities**
5. Determine the **domain** for each **attribute**
6. **Combine** the diagrams of entity, relationship and attribute to develop a complete ER model. (No hanging entities)
7. Thoroughly **check** and **refine** the ER model (If necessary, discuss with users)

5. Enhanced Entity Relationship Model

a) EER-Model: The ER model is enhanced so that data in a complex business environment can bbe represented more accurately. It has additional concepts such as :

i. **Weak Entity:** Has no significance and its existance depends on another strong entity, without which it doesn't have any meaning. It doesn't even have its own key.
ii. **Composite Entity:** is formed when a M:N relationship is transformed to an entity
iii. **Super Class and Sub Class Entity:** Entity type used to represent a group of entities having same characteristics.
    1. **Super Class Entity**: Type of entity which is more general and has relationship with one or more sub class.
    2. **Sub Class Entity**: is one or several entities with different attributes from one another but share the same attributes as its super class
iv. **Generalization:** Process of creating a type of entity that is more general than a set of special entities
v. **Specialization:** (Opposite to Generalisation) Process of determining one or several subclasses from an entity (which later becomes a superclass)
vi. **Aggregation**
vii. **Disjoint Rule:** States that for the same superclass, entity occurance of a subclass may not become a member of another subclass at the same time.
viii. **Overlap rule:** (Opposite of Disjoint Rule) States that in the same super class, an entity occurance of one subclass can be a member of one or more than one of the other subclasses at the same time.