CSC 785: Information storage & retrieval
Assignment 01
Submitted By – Srijana Raut(101134199)

1. Discuss briefly on '50 years of Data Science'. Explain the difference between statistics and data science.

The title "50 years of Data Science" is taken from the book by David Donoho where he discusses the evolution of the field of data science and its relationship with traditional statistics. Here's a breakdown of the key points made in the passage:

1. Historical Perspective:

More than 50 years ago, John Tukey called for a reformation of academic statistics, emphasizing the need for a new science focused on learning from data, which he referred to as "data analysis."

2. Recent Developments:

In recent years, there has been a growing interest in data science as an academic discipline, with universities such as UC Berkeley, NYU, MIT, and the University of Michigan establishing Data Science programs.

3. Overlap with Statistics:

Data Science programs often cover topics that have significant overlap with traditional statistics courses. However, these programs tend to move away from close involvement with academic statistics departments.

4. Data Science as a Superset:

The paper suggests that Data Science is a superset of statistics and machine learning, with an additional focus on technology for handling big data. This choice is motivated by commercial rather than purely intellectual considerations.

5. The Intellectual Shift:

The paper argues that the true intellectual shift in the future will be the emergence of scientific studies of data analysis across all of science. This means moving beyond mere "scaling up" to a more comprehensive understanding of how data analysis impacts scientific research across various fields.

6. Predicting Impact:

The future of Data Science is envisioned as the ability to predict how changes in data analysis workflows would affect the validity of data analysis in various scientific domains. It aims to assess the impact of data analysis methodologies field-by-field.

7. Vision for Data Science:

The author proposes a vision of data science based on the activities of individuals engaged in "learning from data." This vision involves an evidence-based academic field dedicated to improving data analysis practices in a comprehensive and intellectually rigorous manner.

8. Enlargement of Statistics:

The paper suggests that this new academic field represents a more suitable enlargement of traditional statistics and machine learning than current Data Science Initiatives, aligning more closely with the intellectual goals of understanding and improving data analysis processes.

In summary, the passage highlights the historical context of data science, the emergence of data science, the relationship between data science and statistics, and the vision of data science as a field focused on improving data analysis practices across all scientific disciplines. The difference between Statistics and Data Science are given below;

Statistics	Data Science
Statistics" means the practice or science of collecting and analyzing numerical data in large quantities.	Data Scientist" means a professional who uses scientific methods to liberate and create meaning from raw data.
Traditional statistics often deals with smaller datasets, where mathematical techniques like hypothesis testing and confidence intervals are used to draw conclusions.	Data science often deals with big data, which involves handling and analyzing massive datasets
Statistics primarily uses classical statistical methods and tests like t-tests, ANOVA, and regression analysis.	Data science employs a broader array of tools and techniques, including machine learning algorithms, deep learning, data preprocessing, and data visualization libraries.

2. Enumerate and explain different functions of a DBMS?

A Database Management System (DBMS) is software that provides an interface for interacting with databases and manages various aspects of data storage, retrieval, security, and integrity.

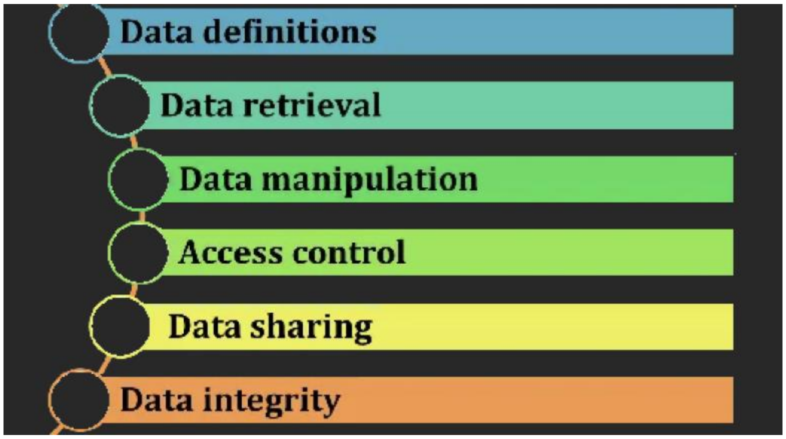


Figure – Various Functions of DBMS

 Data Definition: This includes.

- Schema Definition: DBMS allows users to define the structure of the database, including tables, relationships, constraints, and data types.
- Data Dictionary Management: It maintains metadata, such as table names, column names, and data types, which helps ensure data consistency and integrity.

 Data Retrieval and Querying: This includes.

- Query Language Support: DBMS provides a query language (e.g., SQL) for users to retrieve, filter, and manipulate data stored in the database.
- Query Optimization: It optimizes query execution by choosing the most efficient access paths and execution plans.

 Data Manipulation: It can, be various like;

- Insertion, Updating, Deletion: DBMS allows users to insert, update, and delete data records in the database while enforcing data integrity constraints.
- Transaction Management: It ensures the integrity and consistency of data by supporting transactions with features like ACID (Atomicity, Consistency, Isolation, Durability) properties.

 Access Control:

- DBMS manages simultaneous access to data by multiple users to ensure that data remains consistent despite concurrent updates.

 Data Sharing:

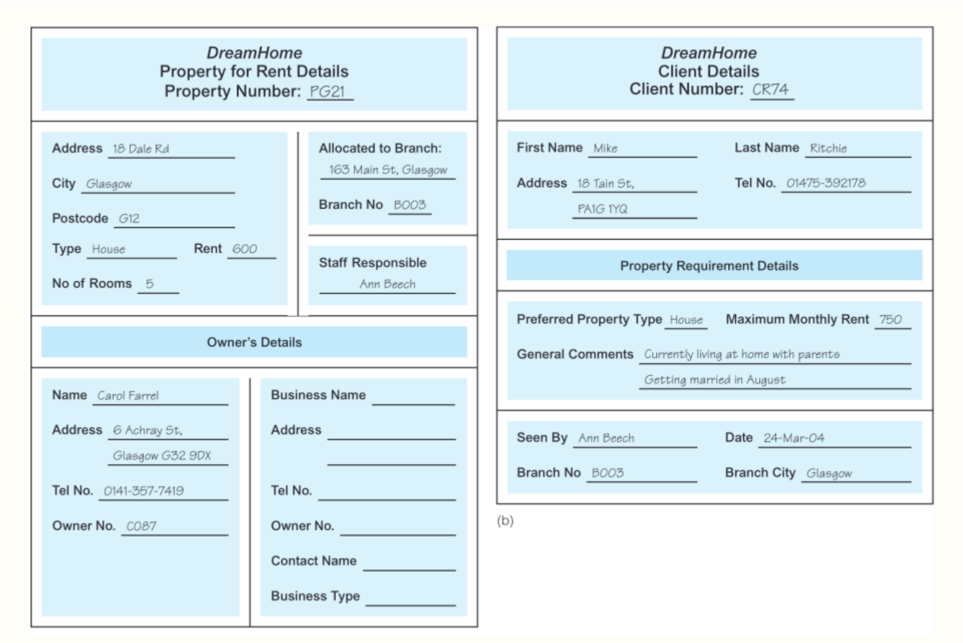
- Data can be shared within various application of database which can be easy.

 Data Integrity:

- Referential Integrity: DBMS enforces referential integrity constraints, ensuring that relationships between tables are maintained.
- Data Validation: It checks data for accuracy and consistency based on defined rules and constraints.

3. Describe traditional file-based system and compare it with database approach. Don’t forget to mention the limitations of file-based approach.

   Traditional file-based system is Collection of application programs that perform services for the end users. For example, reports. In this approach, each program defines and manages its own data.



DreamHome
Property for Rent Details
Property Number: PG21

Address 18 Dale Rd
City Glasgow
Postcode G12
Type House Rent 600
No of Rooms 5

Allocated to Branch: 163 Main St, Glasgow
Branch No B003

Staff Responsible Ann Beech

Owner's Details

Name Carol Farrel
Address 6 Achray St,
Glasgow G32 9DX
Tel No. 0141-357-7419
Owner No. CO87

Business Name
Address
Tel No.
Owner No.
Contact Name
Business Type

DreamHome
Client Details
Client Number: CR74

First Name Mike
Last Name Ritchie
Address 18 Tain St,
PA1G 1YQ
Tel No. 01475-392178

Property Requirement Details

Preferred Property Type House
Maximum Monthly Rent 750
General Comments Currently living at home with parents
Getting married in August

Seen By Ann Beech
Date 24-Mar-04
Branch No B003
Branch City Glasgow

(b)

Figure a) Property and b) Clients

The difference between the traditional file-based system with the database approach are.

Traditional File Based System	**Database approach**
A traditional file-based system is a data management approach where data is stored in separate files.	The database approach is an organized method of managing data where Database Management System (DBMS) is used to store, retrieve, and manage the data.
In file-based systems, data redundancy is prevalent because the same data may be stored in multiple files.	Data is stored centrally in a database, rather than being scattered across multiple files. Thus, no chance of Data redundancy.
In a file-based system it has its own data files, making it difficult to share data across different purpose. Thus, results on data integration and sharing.	A database system allows data to be shared and integrated across different applications.
File-based systems often lack robust security mechanisms.	Database systems provide robust security features, including user authentication, authorization, and encryption.
Retrieving specific data from a file-based system can be tedious.	Retrieving data from a database is more straightforward and efficient due to the structured query language (SQL) and indexing mechanisms.

The limitation of File based approach are.

- Separation and isolation of data

    - Each program maintains its own set of data.
    - Users of one program may be unaware of potentially useful data held by other programs.
  - Duplication of data
    - Same data is held by different programs.
    - Wasted space and potentially different values and/or different formats for the same item.
  - Data dependence
    - File structure is defined in the program code.
  - Incompatible file formats
    - Programs are written in different languages, and so cannot easily access each other's files
  - Fixed Queries/Proliferation of application programs
    - Programs are written to satisfy functions.
    - Any new requirement needs a new program.

4. How do you differentiate data definition language (DDL), data manipulation language (DML) and data control language (DCL). Provide examples (mandatory).

   The difference between DDL, DML and DCL are given below;

SN	Data Definition Language (DDL)	Data Manipulation Language (DML)	Data Control Language (DML)
1.	DDL Statements are used to define the database structure or schema.	DML are used for managing data within schema object.	DCL is used to control like granting and revoking databases.
2.	For examples. CREATE – to create objects in the database. ALTER-alters the structure of the database. DROP- delete objects from the databases.	For examples. SELECT – retrieve data from the database. INSERT- insert data into the table. UPDATE – updates existing data within a table.	For examples. **GRANT –** gives users privileges to the database. **REVOKE-** withdraw access priviliges given with the GRANT command.
3.	Proper SQL query to create the table. CREATE TABLE Employees (EmployeeID INT PRIMARY KEY, FirstName VARCHAR(50),	Proper SQL query to Select data from the table. SELECT FirstName, LastName FROM Employees	Proper SQL query to provide specific privileges to users or roles. GRANT SELECT ON Employees TO User1;

	LastName VARCHAR(50), Salary DECIMAL(10, 2));	WHERE Salary > 50000;	