Lalitha Priya Bijja  
101168225

**CSC 785 - INFORMATION STORAGE AND RETRIEVAL**

**ASSIGNMENT 01**

1. **50 Years of Data Science**

The historical development of data science is highlighted in this section. It analyses how data science developed into a field centred on "learning from data." It emphasizes:

* John Tukey's vision (1962), more than 50 years ago, of a field centred on data analysis.
* The separate appeals to the academic statistics community to adopt data science made by notable personalities including John Chambers (1993), Jeff Wu (2000), Bill Cleveland, and Leo Breiman (2001).
* The theoretical justification for data science, which seeks to enhance data analysis processes through evidence-based methods.
  1. **Data Science vs. Statistics**
* This section questions traditional criteria for differentiation and aims to set data science apart from traditional statistics.
* By demonstrating that statisticians have historically worked with huge datasets, "Big Data" is debunked as a sole criterion.
* A critical examination of the notion of "new skills" in data science contends that these abilities are sometimes just rebranded versions of previously developed statistical procedures.
* Discussion of how data science programs could not ensure better employment chances than statistics, and even if they do, real workplace restrictions may still be present.  
    
  **1.2. Review of the Idea**
* This section provides information on the historical background of data science and the significant contributions of prominent figures:
* John Tukey identified statistics, computing technology, the difficulty of processing greater datasets, and the requirement for quantification in many sectors as the four driving forces in data science.
* According to Tukey, data analysis is a complicated field that must change to meet the actual requirements of those who deal with it.
* A more comprehensive viewpoint that incorporates learning from data is advocated by John Chambers' concept of "greater" or "lesser" statistics.
* Data science is studied using William S. Cleveland's six foci, which are transdisciplinary investigations, models and methods for data, computing with data, pedagogy, tool assessment, and theory.
  1. **The Six Divisions**
* **Data Gathering, Preparation, and Exploration:** Involves collecting, cleaning, and exploring data to understand its characteristics.
* **Data Representation and Transformation:** Focuses on data formatting, making it suitable for analysis (e.g., converting acoustic data into the cepstrum or Fourier transform).
* **Computing with Data:** Stresses the importance of efficient programming and computational techniques.
* **Data Modelling:** Covers the process of building models to extract meaningful insights from data.
* **Data Visualization and Presentation:** Deals with the visual representation of data for effective communication.
* **Science About Data Science:** Explores the meta-analysis of data science practices, including studying analysis workflows and their impact on results.
  1. **Science About Data Science**
* The idea of "science about data science" is explained in more detail in this section:
* **Science-Wide Meta Analysis:** Involves analysing and researching the data analysis techniques utilized by various scientific fields.
* **Cross-Study Analysis:** Includes assessing the consistency of research done on the same dataset.
* **Cross-Workflow Analysis:** This technique focuses on examining how analysis processes affect the results.
* The ultimate objective is to increase the validity and dependability of scientific discoveries in a variety of fields.
  1. **Next 50 Years**
* This part speculates on the future of data science and envisions:
* A more open and reproducible research environment, characterized by increased sharing of code and data.
* The transformation of scientific publications into data that can be analysed, leading to more evidence-based research.
* The availability of abundant data for evaluating the performance of algorithms and models in diverse real-world scenarios.
* Anticipation of a significant expansion of data science's scope and impact in the coming decades.
  1. **Conclusion**
* The conclusion reiterates that data science is primarily driven by intellectual pursuits and scientific questions rather than purely industrial demands.
* It highlights how data science has the potential to be extremely important in improving the credibility of research done in academic contexts.

**Data Science Vs. Statistics**

|  |  |
| --- | --- |
| **Data Science** | **Statistics** |
| Data science is an interdisciplinary field that combines statistics, computer science, and technology to extract valuable insights from large volumes of data. It involves converting real-life problems into research projects and using statistical analysis, machine learning algorithms, and computational tools to make data-driven decisions. | Statistics is a mathematical science that involves analysing existing data to solve specific problems. It focuses on applying statistical tools and techniques to data, interpreting the results, and presenting them in a way that aids decision-making. Statistics is often used in data science as a modified application to analyse smaller sampled data and determine cause-and-effect relationships. |
| Data science aims to identify underlying trends and patterns in data to drive decision-making. It can handle both quantitative and qualitative data and can work with datasets of any size, including big data. Key steps include data mining, data pre-processing, exploratory data analysis, and model building and optimization. Techniques such as regression and classification are commonly used. | Statistics focuses on determining cause-and-effect relationships in analysed data through a purely mathematical approach. It mostly applies to quantitative data and typically analyses smaller sampling data. Mean, median, mode, standard deviation, and variance are important concepts. Techniques like probability distribution, acceptance sampling, and statistical quality control are commonly used. |
| Data science finds application in specialized areas such as computer vision, natural language processing, disaster management, recommender systems, search engines, and more. It is used in sectors like engineering, finance, marketing, e-commerce, and healthcare. | Statistics is applied in various industries where random variations are observed in sampled data. It has applications in fields including business, economics, engineering, finance, marketing, and information technology. It is frequently used for projects like market analysis, financial forecasts, and quality control. |
| Data science has applications in various domains, including healthcare, computer vision, retail and e-commerce, banking and finance for fraud detection, aviation for flight planning, manufacturing for predictive maintenance, transportation and logistics for fleet management, and chatbots. | Statistics has real-life applications in areas such as the stock market, weather forecasting, sports and sporting events, research, public administration, business, consumer goods, insurance, and disaster prevention, among others. |

1. **Functions of DBMS**
2. Data Storage Administration

* The creation and management of complex data databases is one of the most crucial DBMS responsibilities.
* It makes simple for the user by giving complex data sets a framework so that users may access and alter them with ease,
* Modern database systems store and handle metadata (data of data) such as data procedural rules, validation rules, etc. in addition to providing storage for the data itself.
* Additionally, DBMS offers performance tunning, which helps speed up data access.

1. Security Administration

* Another issue that the database management systems solve is security.
* Utilizing a range of security algorithms, database systems offer a high level of security safeguards to protect data and secure data privacy.
* What information can be accessed from the database and by which users are controlled by a set of security rules.
* Additionally, it ensures that specified data can be subjected to read, write, and delete actions.
* It is crucial for organizations where multi-user databases are present.

1. Backup and Recovery Management

* The database system offers tools for backup and recovery management in order to safeguard the data and guarantee its integrity.
* In the event that the system malfunctions for any reason, the data is recovered and kept secure.

1. Database Access Language and Application Programming Interface

DBMS provides a database access language which is also called a query language. Query languages are non-procedural languages used to access the database and manipulate the data. SQL is an example of a query language. The majority of DBMS vendors provide the support of various query languages to access the data.

1. Data Transformation and Presentation

* DBMS provides the functionality of data transformation, which means programmers need not worry about the logical and physical representation of the data.
* DBMS stores the data in the determined data structure.
* For example, if a user asks for the date from a database and he receives it as 14 December 2022, but in the database, it is stored in different columns of month, date and year.

1. Data Integrity Management

* Database systems provide data integrity management by maximizing the data consistency and minimizing the data redundancy.
* The data dictionary is the feature database system used to store the relationships of the data to keep the data integrity.
* Data integrity is needed where a transaction-based database system is present.

1. Database Communication Interface

When a user requests data from the database, it uses some environments like browsers (Chrome or Firefox etc.) to get the data.

An end user can access data in the following ways:

* If he asks for the data through any form and sends the request.
* He can get the data if DBMS publishes the data on any website without asking him.
* He can get the data through a third-party distribution network.

1. **Traditional File Based approach VS. Database Management Approach**

|  |  |  |
| --- | --- | --- |
| **SNO** | **Database Management Approach** | **File Based Approach** |
| 1. | A database management system, or DBMS, is a software application that allows you to access, create, and manage databases. | A file system is a software application that organizes and maintains files on a storage device. It manages the storage and retrieval of data. |
| 2. | DBMS is a collection of data. In DBMS, the user is not required to write the procedures. | The file system is a collection of data. In this system, the user has to write the procedures for managing the database. |
| 3. | Due to the centralized approach, data sharing is easy. | Data is distributed in many files, and it may be of different formats, so it isn't easy to share data. |
| 4. | DBMS gives an abstract view of data that hides the details. | The file system provides the detail of the data representation and storage of data. |
| 5. | DBMS provides a crash recovery mechanism, i.e., DBMS protects the user from system failure. | The file system doesn't have a crash mechanism, i.e., if the system crashes while entering some data, then the content of the file will be lost. |
| 6. | In the database approach, 3 types of data models exist:   1. Hierarchal data models 2. Network data models 3. Relational data models | In the file system approach, there is no concept of data models exists. |
| 7. | Changes are often a necessity to the content of the data stored in any system, and these changes are more easily with a database approach. | Changes are often a necessity to the content of the data stored in any system, and these changes are more easily with a database approach. |
| 8. | Due to the centralization of the database, the problems of data redundancy and inconsistency are controlled. | In this, the files and application programs are created by different programmers so that there exists a lot of duplication of data which may lead to inconsistency. |
| 9. | Examples: Oracle RDBMS, SQL Server, Sybase, IBM DB2 etc. | Examples: Cobol, C++ etc. |

**Limitations of File Based Approach**

1. **Data Redundancy**:  
   It’s possible that the same data is duplicated in many files. This results in data redundancy and memory waste.
2. **Data Inconsistency**:  
   Due to data redundancy, it’s probable that data won’t be consistent.
3. **Accessing Data Is Difficult**:  
   It’s not simple to get data out of a file management system. It is not as convenient as it should be. When a user needs to access information, they must run specialized software.   
   For example, accessing a record of an employee in the traditional file-based system among huge number of files (data).
4. **Limited Data Sharing**:  
   Data is dispersed across multiple files. Moreover, separate files may have different formats, and these files may be housed in several directories belonging to various departments. As a result of this data segregation, data sharing between applications is challenging.
5. **DDL Vs. DML Vs. DCL**

**DDL**

* DDL is an abbreviation for **Data Definition Language**.
* It is concerned with database schemas and descriptions of how data should be stored in the database.
* DDL statements are auto-committed, meaning the changes are immediately made to the database and cannot be rolled back. These commands enable database administrators and developers to manage and optimize MySQL databases effectively.
* DDL is used for defining and managing the structure of database objects, such as tables, indexes, and constraints. It does not deal with the actual data within tables.

**Examples:**

**CREATE TABLE:**

CREATE TABLE students (

student\_id INT PRIMARY KEY,

first\_name VARCHAR(50),

last\_name VARCHAR(50),

admission\_date DATE

);

**ALTER TABLE:** Modifies the structure of an existing table, such as adding or dropping columns

ALTER TABLE workers

ADD email VARCHAR(100);

**DROP TABLE:** Deletes an entire table

DROP TABLE customers;

**DML**

* DML stands for **Data Manipulation Language**.
* It deals with data manipulation and includes the most common SQL statements such as SELECT, INSERT, UPDATE, DELETE, etc.
* DML statements are not auto-committed, meaning the changes can be rolled back if necessary.

**Examples:**

**SELECT**

SELECT first\_name, last\_name FROM students WHERE department = 'Compsci';

**INSERT**

INSERT INTO students (student\_id, first\_name, last\_name) VALUES (101, 'John', 'Doe');

**UPDATE**

UPDATE products SET price = 9.99 WHERE product\_id = 1011;

**DELETE**

DELETE FROM orders WHERE order\_id = 1001;

**DCL**

* DCL stands for **Data Control Language**.
* It includes commands such as GRANT and is primarily concerned with rights, permissions, and other controls of the database system.
* DCL statements are also auto-committed.

**Examples:**  
**GRANT**

GRANT SELECT, INSERT ON employees TO hr\_user;

**REVOKE**

REVOKE INSERT ON customers FROM banking\_user;

**COMMIT:** Confirms and saves all changes made during the current transaction to the database.

**ROLLBACK:** Undoes any changes made during the current transaction and returns the database to its previous state.