STA 2102>IT FOR STATISTICS

1. INTRODUCTION TO IT AND COMPUTERS

WHAT IS IT?

Information >Refers to data that has been processed and organized so that It becomes useful and meaningful to the user

Data > Collection of raw facts, figures or instructions that do not have much meaning to the **user**

Technology > Refers to the tools, systems and methods used to process data and row facts into information

Information Technology is there for the use of computers, software and other digital tools to process, store and manage data and information

A computer is an electronic device that accepts data as input and transforms it under the set of special instructions called programs, to produce the desired output (information)

IINFORMATION VS TECHNOOLOGY

Data

1. Unprocessed (raw) facts or figures.
2. Not arranged.
3. Does not have much meaning to the user.
4. Cannot be used for decision-making.

Characteristics / Features of a Computer.

Information

1. It is the end-product of data processing
(processed data)
2. Arranged into a meaningful format.
3. More meaningful to the user.
4. Can be used

ROLE OF IT IN SOCIETY

1.Communication and connectivity

- Global reach >IT connects people globally through social media, messaging, and video conferencing, allowing for instant communication across distances
- Information access >The internet provides unparalleled access to information, new ideas, and different perspectives.

2.Economy and business

- Increased productivity >IT optimizes business operations through automation, data management and streamlining of tasks
- Market efficiency >It provides buyers and sellers with easier access to information, which can reduce costs and improve market operation

3.Education and healthcare

- Flexible learning >IT supports online learning, allowing people to study remotely and access educational resources at their own pace
- Improved healthcare >Technologies like telemedicine and electronic health records enhance patient care and expand access to medical professionals

4.Governance and public service- E-government >IT enables governments to provide services online, such as tax filing and access of public policy information

COMPUTING APPLICATION FINNANCIAL ENGINEERING

1.Derivatives pricing and valuation >Financial engineers use computational methods, such as Monte Carlo Simulations, to value complex financial instruments like options and futures

2.Risk management >This involves extensive data analysis and the use of specialized risk management software to help identify, measure and manage various financial risk

3.Portfolio Optimization > Complex algorithms and software are used to manage and optimize investment

4.Financial modeling and forecasting > Computational tools, including machine learning and deep learning modules are used to analyze historical data and predict future market trends

5.Fraud detection> Machine earning algorithms analyze large volumes of transaction data in real time to flag suspicious activities helping financial institutions protect assets

6.Algorithmic and high frequency trading > Computers are essential for executing trades automatically based on pre-programed instructions and market conditions, often at very high speed

Topic 2: Fundamentals of Computer Operations

1. Central Processing Unit (CPU)

- The CPU is the brain of the computer.
- Handles processing and controls all other components.
- Made up of ALU, Control Unit, and Registers.
- Executes all instructions through machine cycles.

2. Arithmetic and Logic Unit (ALU)

- Performs arithmetic operations (add, subtract, multiply, divide).
- Handles logic operations (AND, OR, NOT, comparisons).
- Supports decision-making processes in programs.

3. Control Unit (CU)

- Fetches instructions from memory.
- Interprets and decodes instructions.
- Controls execution and data flow within the CPU.
- Coordinates input, output, and memory operations.

4. Registers

- Small, fast memory inside the CPU.
- Store data, instructions, and memory addresses temporarily.
- Faster than RAM.
- Examples: Accumulator, Instruction Register, Program Counter, Memory Address Register.

5. System Clock

- Synchronizes all CPU operations.
- Measured in Hertz (Hz).
- Higher clock speed means more instructions processed per second.
- Controls timing of fetch–decode–execute cycle.

6. Machine Cycle

- Fetch – Retrieve instruction from memory.

- Decode – Interpret the instruction.
- Execute – Perform the required operation.
- Store – Save the output or result.

7. Parallel Processing Basics

- Uses multiple processors or cores at the same time.
- Improves speed and performance.
- Types include multicore processing, multiprocessing, and distributed computing.
- Useful for handling big data and simulations.

8. Performance Evaluation for Statistical Computation

- CPU speed affects how fast computations run.
- More cores enable parallel data processing.
- RAM capacity determines ability to handle large data sets.
- SSD storage improves data loading and execution speed.
- Throughput shows total work done in a time period.
- Latency measures delays in processing.

COMPUTER HARDWARE

Definition
Computer hardware consists of all the physical or tangible parts which together make up a computer system, including both mechanical and electronic parts. It includes:
- Peripheral devices
- Central Processing Unit (CPU)

- Storage and auxiliary equipment
Input Devices
Input devices allow users to input data and give instructions to the computer. Examples:
1. Textual Input Devices – Keyboard (used to enter characters and numbers)
2. Pointing/Navigation Devices – Mouse, Touchpad, Trackball (control cursor movement on the screen)
3. Audio Input Devices – Microcomputer (picks up sound)
4. Specialized Devices –
- Joystick for gaming
- Barcode Scanner: reads product codes in retail

Biometric Readers: authenticate users by their unique physical traits.
5. Visual Input Devices – Scanners, Webcams (convert images or live video to digital formats)
Output Devices
Output devices translate computer data into forms perceivable by humans—visual, audio, or printed. Examples:
1. Visual output devices: video card, projector, monitor - display images or videos 2. Print Output Devices – Printer, Plotter, Braille Reader ميزان hard copies of digital data 3. Sound Output Devices – Sound Card, Headphones, Speakers (provide audio output) 4. Data Output Devices – GPS provides readable information; access to underlying data usually is restricted.

COMPUTER HARDWARE II -STORAGE DEVICES AND MEMORY.

I. *Memory Hierarchy and Speed.*

- **Primary vs Secondary Memory:**
  - **Primary (e.g., RAM):** Directly accessible by the CPU, **fast**, **volatile** (data lost when power is off).
  - **Secondary (e.g., HDD, SSD):** Slower, **non-volatile** (data persists). Used for long term storage of files and programs.
- **Cache:** A very small, fast memory (SRAM) located close to the CPU. It stores copies of frequently used data from main memory (RAM) to minimize access latency and speed up processing.

II. Modern Storage Technologies

- **SSDs (Solid State Drives):** Use **NAND flash memory** to store data. They have **no moving parts**, offering dramatically **faster I/O performance**, lower power consumption, and better durability than traditional HDDs.
- **RAID (Redundant Array of Independent Disks):** A technology that combines multiple physical disk drives into a single logical unit. It's used to improve:
  - **Performance** (speed) through **striping** data across drives.
  - **Redundancy** (fault tolerance) through **mirroring** or **parity** checks.
- **Cloud Storage:** Data is stored digitally in pooled resources maintained by a third-party service provider and is accessible over the internet. It provides **scalability** and remote access.

III. Disk Storage Structure (HDD Terminology)

This describes the physical geometry of traditional spinning disk drives:

- **Surfaces:** The top and bottom sides of the platters where data is recorded magnetically.
- **Tracks:** Concentric rings on a surface, similar to grooves on a vinyl record, where data bits are aligned.
- **Sectors:** The smallest physical storage unit on a track (usually 512 bytes). Data is read and written in sectors.
- **Clusters:** A group of contiguous sectors. This is the **smallest unit of disk space** that the operating system (OS) allocates when storing a file.

IV. Implications for Large-Scale Datasets.

- **Context:** Processing large-scale **biostatistical and financial datasets** requires moving massive amounts of data quickly and reliably.
- **Critical Factor: I/O performance** (Input/Output speed) is paramount.
- **Solution Need:** Systems must utilize high speed storage (like **SSDs** or NVMe drives) and reliable, high throughput configurations (like **RAID 10** or optimized **cloud solutions**) to handle the constant reading and writing required for advanced data analysis without

causing bottlenecks.

COMPUTER SOFTWARE

A software is program that runs under a set of instructions in order to receive a certain service or mainten a computer system.

Types of System Software:

a) Operating System (OS)

Examples: Windows, macOS, Linux

Functions:

- Manage files and folders

- Allocate memory

- Control input/output operations

- Schedule programs

- Provide a user interface

b) Compiler

A compiler converts source code (written by humans) into machine code understood by the CPU.

Examples: C compiler, Java compiler, Python compiler

c) UtilitiesSmall programs that perform maintenance tasks.

Examples:

- File compression tools

- Disk cleanup/optimization tools

- Antivirus software

Application Software

These are the programs you use to perform specific tasks.

Word Processor

- Microsoft Word: writing reports

Spreadsheet

- Excel: calculation, charts, basic data analysis

Database

- Access: storing and querying structured data

Presentation Tools

- PowerPoint: presenting findings

Workflow Software

- Used to streamline analysis tasks, automate reports, save steps, or help manage data/sheets

DATA AND DATA FILES

Types of files

1.Sequential file

A file where records are stored in order

Used in making payrolls and reports

Advantages

*simple to create and maintain

* use less storage

Disadvantage

*slow for searching

*updating requires rewriting entire file

2.Random/direct files

Records are stored randomly and accessed directly using a key and a hashing algorithm

It is efficient for large files

Advantage

*fast retrieval

*Efficient for large data sets

Disadvantage

*Complex to create

*Require more storage

3.Structured files

Files organized in a predefined format with fields and records

-Organized and predictable

Advantage

*Easy to validate

*Support indexing

*Efficient processing

Disadvantage

*Change in structure require redesign

*Not suitable for multimedia content

4.Unstructured files

Files that don't follow a predefined data model

Require AI for analysis and is hard to process automatically

Advantage

*Flexible

*Easy storage

Disadvantage

*Hard to search

*Require more processing power

STEPS IN FILE CREATION

1.Design file structure (field and data types)
2.Choose file organization (sequential or random)
3.Create empty file
4.Insert records
5.Save and name the file
6.Store metadata

File indexing

Types of indexing
Primary Index
Secondary index
Dense index
Sparse index

Benefits
Fast searching and efficient sorting

File retrieval methods

*Sequential retrieval(record by record)
*Direct/Random retrieval (instant access by key)
*Indexed retrieval (to locate records faster)

File optimization techniques

*Indexing
*Hashing
*Clustering
*Compression
*Using efficient file structure
*Cleaning and defragmentation

Relational database concepts

Stores data in tables made up of rows and columns
*Tables _collection of related data
*Records(rows) _one item
*Field(column) _attributes
*Primary key_unique identifier
*Foreign key_linking one table to another
*Relationship types_one to one, one to many, many to many

Normalization

The process of organizing database tables to reduce redundancy and improve data integrity.

Normal forms
1.First normal form(1NF) _no repeating groups
2.Second Normal Form(2NF) _no partial dependency
3.Third Normal Form(3NF) _no transitive dependency

Objective of normalization

*Remove data duplication
*Ensure consistent data
* Improve efficiency

DISK STORAGE :**Summary notes**

1. Core Concept

· Non-volatile storage: Data persists after power loss.
· Purpose: Long-term storage of operating systems, applications, and files.

2. Types of Disk Storage

Feature HDD (Hard Disk Drive) SSD (Solid-State Drive)
Technology Magnetic platters & read/write heads Flash memory (NAND chips)
Moving Parts Yes (spinning platters, actuator arm) No
Speed Slower (mechanical latency: seek time, rotation) Extremely Fast (instant electronic access)
Durability Less durable (sensitive to shock/vibration) More durable (no moving parts)
Cost/Capacity Low cost per GB, High capacity (multi-TB) Higher cost per GB, lower capacity for price
Power/Noise Higher power, audible noise Lower power, silent
Use Case Bulk storage, archives, budget builds OS, applications, gaming, high-performance systems

Hybrid Drives (SSHD): A small SSD cache paired with a large HDD for a balance of speed and capacity.

3. Key Interfaces & Form Factors

· SATA (Serial ATA): Common standard for HDDs and 2.5" SSDs. Max speed ~600 MB/s.
· NVMe (Non-Volatile Memory Express): Modern protocol for SSDs over the PCIe bus. Much faster than SATA.
· M.2: A physical form factor (a small card). Critical: M.2 slots can support both SATA (slower) and NVMe (faster) SSDs.

4. Organizing Disks: RAID (Redundant Array of Independent Disks)

Combines multiple physical disks for performance and/or redundancy.

RAID Level Name Description Key Benefit
RAID 0 Striping Splits data across disks. Speed (No Redundancy)
RAID 1 Mirroring Duplicates data on two disks. Redundancy (Survives 1 disk failure)
RAID 5 Striping with Parity Data & parity spread across 3+ disks. Balance of speed & redundancy (Survives 1 disk failure)
RAID 10 Mirroring + Striping Creates a striped set from mirrored pairs. High Speed & Redundancy

5. Key Terminology

· Platter: The physical disk inside an HDD where data is stored magnetically.
· Sector: The smallest addressable unit of storage on a disk (traditionally 512 bytes, now commonly 4096 bytes or "4K").
· NAND Flash: The type of memory used in SSDs. More bits per cell (SLC, MLC, TLC, QLC) increases capacity but reduces speed/endurance.