M.Tech Program

Advanced Industry Integrated Programs

Jointly offered by University and LTIMindTree

Data Engineering

Knowledge partner

LTIMindtree

Implementation partner

L&T EduTech

Course Objective

- Recognize data types and structures.
- Grasp big data fundamentals and analytics.
- Master data ingestion processes and tools.
- Understand exploratory data analysis techniques.
- Learn storage methods and data flow.

Modules

- Data Types & Formats
- Data Ingestion techniques
- Data Profiling & Visual Representation via various tools (Pandas)
- Storage and retrieval methods
- Data Lineage Analysis

Storage and Retrieval Methods - *Learning Outcomes..*

- Identify and evaluate various data storage methods (databases, data lakes, file systems, object storage, etc.) and their suitability for different data types.
- Compare and contrast local and distributed storage and retrieval methods, considering their benefits, challenges, and use cases.
- Gain knowledge of the hardware components involved in storage systems (HDD, SSD, RAM, network components) and their impact on performance.
- Understand the factors that influence the choice of storage methods, including data size, access patterns (read/write focus), and performance requirements.

Storage and Retrieval Methods

Types of data and methods of storage available

Types of data and methods of storage available

Introduction to Storage and Retrieval

- Storage and retrieval are core processes in managing data efficiently, supporting applications from small-scale software to large-scale enterprise systems.
- As data grows exponentially, effective storage solutions are critical for performance, scalability, and reliability.
- Storage systems ensure data integrity and accessibility throughout its lifecycle, from creation to archiving.

Types of data and methods of storage available

Introduction to Storage and Retrieval

Evolution of Storage Systems:

- **From Physical to Digital**: Transition from paper and analog systems to digital storage revolutionized how we handle data.
- **Modern Storage Technologies**: Introduction of SSDs, cloud storage, and distributed systems has redefined data access and speed.

Types of data and methods of storage available

Types of Data and Storage Methods

Types of Data

- Structured Data
- Semi-Structured Data
- Unstructured Data

Types of data and methods of storage available

Types of Data and Storage Methods

Types of Data

- **Structured Data**
- Semi-Structured Data
- Unstructured Data

Data that is organized into predefined formats (e.g., rows and columns) and easily searchable.

Types of data and methods of storage available

Types of Data and Storage Methods

Types of Data

- Structured Data
- **Semi-Structured Data**
- Unstructured Data

Data that does not follow a strict structure but contains tags or markers to separate data elements.

Types of data and methods of storage available

Types of Data and Storage Methods

Types of Data

- Structured Data
- Semi-Structured Data
- **Unstructured Data**

Data without a predefined format, often requiring more complex processing to analyze.

Types of data and methods of storage available

Types of Data and Storage Methods

Storage Methods

- The choice of storage method depends on the type of data, the volume of data, and specific application requirements.
- Familiarity with various storage systems, including file systems, databases, and cloud storage, to effectively manage data.

Types of data and methods of storage available

Types of Data and Storage Methods

Storage Methods

File Systems

- Hierarchical storage method where data is stored in files and folders.
- Best for unstructured data, such as documents, images, and videos.
- **Examples**: NTFS, ext4, HDFS.

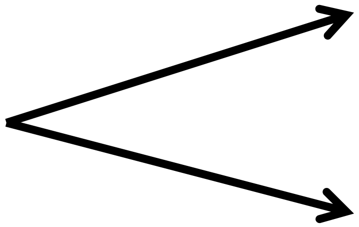
Types of data and methods of storage available

Types of Data and Storage Methods

Storage Methods

Databases

- Systems designed to store and manage structured data with capabilities for querying and reporting.
- **Types**



Relational (e.g., MySQL, PostgreSQL)

NoSQL (e.g., MongoDB, Cassandra)

Types of data and methods of storage available

Types of Data and Storage Methods

Storage Methods

Cloud Storage

- Storage services provided over the internet, offering scalable and flexible storage solutions.
- Advantages : Scalability, accessibility, and disaster recovery.
- Examples: Amazon S3, Google Cloud Storage, Azure Blob Storage.

Local vs Distributed (Storage & Retrieval)

Local vs Distributed (Storage & Retrieval)

Local Storage

- Local storage refers to storing data on physical devices that are directly accessible from a single computer system.
- Examples include internal hard drives, SSDs, external hard drives, USB flash drives, and optical disks.

Local vs Distributed (Storage & Retrieval)

Overview of Local Storage

Characteristics:

- Accessibility: Directly accessible by the computer it is attached to.
- Performance: Generally offers high read/write speeds.
- Capacity: Limited by the physical constraints of the storage device.
- Security: Data is more secure as it is not transmitted over networks.

Local vs Distributed (Storage & Retrieval)

Advantages and Disadvantages Local Storage

Aspect	Local Storage
Advantages	**High Performance**: Faster access speeds due to local data retrieval.
	Simplicity: Easier to set up and manage for small-scale applications.
	Low Cost: Lower initial investment compared to distributed systems.
Disadvantages	**Limited Scalability**: Hard to scale beyond the capacity of a single machine.
	Single Point of Failure: Risk of data loss if the local machine fails.
	Limited Access: Access is restricted to the local network or machine.

Local vs Distributed (Storage & Retrieval)

Overview of Distributed Storage

Distributed storage refers to storing data across multiple physical locations, often using a network of servers.

Characteristics:

- Scalability: Can easily scale out by adding more nodes to the network.
- Redundancy: Data is often replicated across multiple nodes to ensure reliability.
- Accessibility: Data can be accessed from multiple locations.

Local vs Distributed (Storage & Retrieval)

Advantages and Disadvantages of Distributed Storage

Aspect	Distributed Storage
Advantages	**Scalability:** Can handle large volumes of data by adding more nodes to the network.
	Fault Tolerance: Redundant storage ensures data is not lost if a node fails.
	High Availability: Data is accessible from multiple locations, enhancing availability and performance.
Disadvantages	**Complexity:** More complex to set up and manage compared to local storage.
	Latency: Potentially higher latency due to network communication.
	Cost: Higher initial and ongoing costs for infrastructure and maintenance.

Local vs Distributed (Storage & Retrieval)

Local Retrieval

- Local retrieval involves accessing data stored on local storage devices.
- Direct file access, database queries, and application-specific data retrieval methods.
- Local retrieval is typically fast and straightforward due to the proximity of the storage device.

Local vs Distributed (Storage & Retrieval)

Techniques and Tools

Techniques:

- File System Operations: Using OS-level commands to access files (e.g., Windows Explorer, Linux shell commands).
- Database Access: Querying local databases using SQL or other database management tools.
- Application APIs: Utilizing application-specific APIs for data retrieval.

Local vs Distributed (Storage & Retrieval)

Techniques and Tools

Tools:

- File Managers: Windows File Explorer, Finder on macOS.
- Database Management Systems: MySQL, SQLite.
- Programming Languages: Python, Java, C++ for building custom retrieval solutions

Local vs Distributed (Storage & Retrieval)

Performance Considerations

Factors Affecting Performance:

- Storage Device Speed: SSDs vs. HDDs.
- Data Organization: Efficient data structuring and indexing.
- System Resources: CPU, memory, and I/O capabilities of the host system.

Local vs Distributed (Storage & Retrieval)

Trade-offs Between Local and Distributed Storage

Aspect	Local Storage	Distributed Storage
Data Size	Best for small to medium-sized datasets.	Ideal for large datasets requiring scalability.
Access Frequency	Suitable for high-frequency access with minimal latency.	Handles varying access patterns but may have higher latency.
Scalability	Limited scalability; constrained by physical hardware.	Scalable by adding more nodes or resources.
Cost	Lower initial cost; higher cost for scaling.	Higher initial investment but cost-effective at scale.

Hardware Aspects of Storage & Retrieval

Hardware Aspects of Storage & Retrieval

Introduction to Hardware Aspects of Storage & Retrieval

- Hardware components play a critical role in determining the efficiency and performance of data storage and retrieval processes.
- The choice of hardware influences data access speeds, storage capacity, and overall system responsiveness.

Hardware Aspects of Storage & Retrieval

Types of data storage devices

Hard Disk Drives (HDD):

- HDDs use spinning magnetic disks to read and write data.
- **Components**: Platter, read/write head, spindle, actuator arm, controller board.
- **Mechanism**: Data is written by magnetizing the thin film of ferromagnetic material on the disk, and read by sensing the magnetization of the disk.

Hardware Aspects of Storage & Retrieval

Types of data storage devices

Solid-State Drives (SSD)

- SSDs use flash memory for data storage, providing faster access times.
- **Components**: NAND flash memory, controller, cache.
- **Mechanism**: Data is stored in integrated circuits and accessed electronically, eliminating the need for moving parts.

Hardware Aspects of Storage & Retrieval

Memory (RAM)

Dynamic RAM (DRAM)

- DRAM is a type of volatile memory used for temporary data storage.
- **Characteristics**: Requires periodic refreshing to maintain data.
- **Mechanism**: Stores each bit of data in a separate capacitor within an integrated circuit.

Hardware Aspects of Storage & Retrieval

Memory (RAM)

Static RAM (SRAM)

- SRAM is a type of volatile memory that does not require periodic refreshing.
- **Characteristics**: Uses bistable latching circuitry to store data.
- **Mechanism**: Each bit is stored in a flip-flop, providing faster access times than DRAM.

Hardware Aspects of Storage & Retrieval

Storage Interfaces

Serial ATA (SATA)

- SATA is an interface to connect storage devices to the motherboard.
- **Characteristics**: Supports hot-swapping and improved data transfer rates.
- **Versions**: SATA I (1.5 Gb/s), SATA II (3 Gb/s), SATA III (6 Gb/s).

Hardware Aspects of Storage & Retrieval

Storage Interfaces

Peripheral Component Interconnect Express (PCIe)

- PCIe is a high-speed interface used to connect components like GPUs and high-performance SSDs.
- **Characteristics**: Higher bandwidth and lower latency than SATA.
- **Versions**: PCIe 3.0 (8 GT/s), PCIe 4.0 (16 GT/s), PCIe 5.0 (32 GT/s).

Hardware Aspects of Storage & Retrieval

Optimizing Storage and Retrieval Performance

Configuring RAID

- RAID (Redundant Array of Independent Disks) combines multiple disk drives into a single unit.
- **Levels**: RAID 0 (striping), RAID 1 (mirroring), RAID 5 (striping with parity), RAID 10 (striping and mirroring).
- **Benefits**: Improves performance, provides redundancy, and increases storage capacity.

Hardware Aspects of Storage & Retrieval

Optimizing Storage and Retrieval Performance

Caching Mechanisms

- Caching stores frequently accessed data in a temporary storage area for quick access.
- **Types**: CPU cache, disk cache, web cache.
- **Levels**: L1, L2, and L3 caches in CPU, each with increasing size and decreasing speed.

How to choose storage methods

How to choose storage methods

Choosing Storage Methods

Key Considerations

- Size of Data
- Read-Focused vs. Write-Focused
- Performance Expectations
- RTO & RPO
- Cost

How to choose storage methods

Choosing Storage Methods

- Data size and access patterns significantly influence the choice of storage methods.
- For instance, big data systems like Hadoop require distributed file systems, whereas transactional databases might use SSDs for quick writes.
- Cost considerations include not just the initial investment but also ongoing maintenance, scalability, and potential migration costs.
- RTO & RPO are critical for business continuity, particularly in sectors with stringent data integrity requirements.

How to choose storage methods

Data Partitioning and Sharding

Data Partitioning

- Dividing large datasets into smaller, manageable partitions for efficient storage and retrieval.
- **Types:** Horizontal (row-based) and vertical (column-based) partitioning.
- **Benefits:** Improves performance, scalability, and maintenance (e.g., load balancing, faster query processing).

How to choose storage methods

Data Partitioning and Sharding

Data Sharding

- Distributing data across multiple servers or nodes to enhance scalability and performance.
- **Shard Key**: Selecting an appropriate key to ensure even distribution of data across shards.
- **Benefits**: Supports large-scale distributed systems, reduces single points of failure, and enables parallel processing.

How to choose storage methods

Data Replication and Redundancy

Data Replication

- Creating multiple copies of data across different locations or systems to ensure availability and durability.
- **Types**: Synchronous (real-time replication) and asynchronous (delayed replication).
- **Use Cases**: High-availability systems, disaster recovery, load balancing.

How to choose storage methods

Data Replication and Redundancy

Data Redundancy

- Storing redundant copies of data to prevent data loss in case of hardware or software failures.
- **Techniques**: RAID, data mirroring, and geographically dispersed storage.
- **Benefits**: Enhances data integrity, ensures business continuity, and minimizes downtime.

How to choose storage methods

Data Compression and Encoding

Data Compression

- Reducing the size of data to save storage space and improve transfer speeds.
- **Types**: Lossless (e.g., ZIP, GZIP) and lossy (e.g., JPEG, MP3).
- **Benefits**: Lower storage costs, faster data transmission, improved performance.

How to choose storage methods

Data Compression and Encoding

Data Encoding

- Converting data into a specific format for efficient processing, transmission, and storage.
- **Techniques**: Base64, UTF-8, ASCII encoding.
- **Benefits**: Ensures data integrity, compatibility across systems, and enhanced security.

How to choose storage methods

Data Archiving and Retrieval

Data Archiving

- Long-term storage of data that is no longer actively used but must be preserved for future reference.
- **Strategies**: Hierarchical storage management (HSM), cloud-based archiving.
- **Benefits**: Frees up primary storage, ensures compliance with legal and regulatory requirements.

How to choose storage methods

Data Archiving and Retrieval

Data Retrieval

- Efficiently accessing and retrieving archived data when needed.
- **Challenges**: Retrieval speed, data integrity, searchability.
- **Solutions**: Metadata indexing, automated retrieval systems, data catalogs.

How to choose storage methods

Backup and Disaster Recovery

Backup

- Regularly creating copies of data to restore in case of accidental deletion, corruption, or disaster.
- **Types**: Full, incremental, differential.
- **Best Practices**: Automated backups, off-site storage, encryption.

How to choose storage methods

Backup and Disaster Recovery

Disaster Recovery

- Strategies to restore data and systems after a catastrophic event.
- **Plans**: RTO, RPO, business continuity planning.
- **Technologies**: Cloud-based recovery, hot and cold sites, DRaaS (Disaster Recovery as a Service).

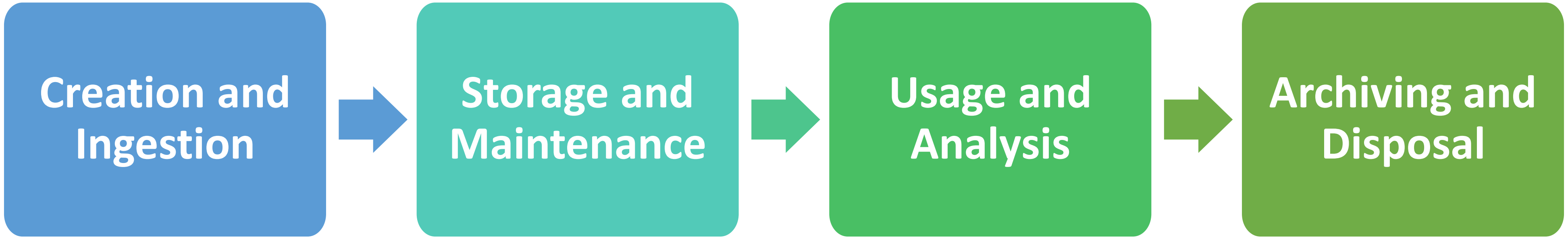
How to choose storage methods

Data Lifecycle Management

Managing Data Lifecycle

The process of managing data from creation through to its eventual archival and deletion.

Key Stages



Creation and Ingestion

Storage and Maintenance

Usage and Analysis

Archiving and Disposal

How to choose storage methods

Data Lifecycle Management

- Lifecycle management is critical for organizations dealing with large volumes of data, ensuring that data remains accurate, secure, and compliant throughout its existence.
- Data governance plays a key role in lifecycle management, ensuring that data policies and procedures are followed consistently.

Summary

Data Profiling & Visual Representation via various tools (Pandas)

Introduction to Storage and Retrieval

Understanding the fundamental concepts of how data is stored, accessed, and managed in computing systems.

Types of Data and Storage Methods

Exploring different data types and the various methods used to store them, including local and distributed storage solutions.

Local Storage

Directly connected storage devices that are part of a single system.

Data Profiling & Visual Representation via various tools (Pandas)

Overview of Local Storage	Insight into data storage directly connected to a computer or server.
Advantages and Disadvantages Local Storage	Evaluating the benefits and drawbacks of using local storage solutions.
Overview of Distributed Storage	Understanding storage systems that distribute data across multiple physical locations.

Data Profiling & Visual Representation via various tools (Pandas)

Advantages and Disadvantages

Assessing the pros and cons of using distributed storage systems.

Local Retrieval

Methods like file systems and software utilities used to access and manage data stored locally.

Techniques and Tools

Exploring methods and tools for retrieving data from local storage.

Data Profiling & Visual Representation via various tools (Pandas)

Performance Considerations	Understanding factors that affect the performance of data retrieval from local storage.
Types of data storage devices	Understanding the use and function of HDDs for data storage. Exploring the role and benefits of SSDs in data storage.
Memory (RAM)	Understanding the characteristics and use cases of DRAM. Exploring the features and applications of SRAM.

Data Profiling & Visual Representation via various tools (Pandas)

Storage Interfaces

Overview of SATA as a standard interface for connecting storage devices.

Understanding PCIe as a high-speed interface for connecting storage devices.

Optimizing Storage and Retrieval Performance

Exploring RAID configurations to improve storage performance and redundancy.

Understanding how caching improves data retrieval speed and efficiency.

Choosing Storage Methods

Guidelines for selecting appropriate storage methods based on data requirements and system constraints.

Data Profiling & Visual Representation via various tools (Pandas)

Data Partitioning and Sharding

Techniques for dividing data into manageable parts to improve performance and scalability.

Data Replication and Redundancy

Strategies for replicating data to ensure availability and fault tolerance.

Data Compression and Encoding

Methods to reduce data size and enhance storage efficiency.

Data Profiling & Visual Representation via various tools (Pandas)

Data Archiving and Retrieval	Processes for storing and retrieving archived data for long-term retention.
Backup and Disaster Recovery	Planning and implementing backup and recovery strategies to protect data from loss.
Data Lifecycle Management	Managing data through its lifecycle from creation to deletion to ensure compliance and efficiency.

Knowledge Check

Storage and Retrieval Methods

Q1 : Which of the following is a primary goal of data storage and retrieval systems?

A. Minimizing data redundancy

B. Ensuring data integrity

C. Optimizing data retrieval speed

D. All of the above

Storage and Retrieval Methods

Q2 : Which storage method is most suitable for structured data?

A. File systems

B. Relational databases

C. NoSQL databases

D. Object storage

Storage and Retrieval Methods

Q3 : Which of the following is a key advantage of distributed storage over local storage?

A. Lower initial setup cost

B. Enhanced scalability and fault tolerance

C. Easier data management

D. Faster data retrieval for small datasets

Storage and Retrieval Methods

Q4 : Which of the following is a common challenge associated with distributed storage systems?

A. Limited scalability

B. Increased risk of data loss due to hardware failure

C. Data consistency and synchronization across nodes

D. Higher initial setup costs compared to local storage

Storage and Retrieval Methods

Q5 : Which of the following storage devices offers the highest data transfer speed?

A. Hard Disk Drive (HDD)

B. Solid State Drive (SSD)

C. Optical Disk

D. Magnetic Tape

Storage and Retrieval Methods

Q6 : Which type of memory requires periodic refreshing to maintain data?

A. Static RAM (SRAM)

B. Dynamic RAM (DRAM)

C. Read-Only Memory (ROM)

D. Flash Memory

Storage and Retrieval Methods

Q7 : Which of the following RAID levels offers a balance between improved performance and redundancy by using striping with parity?

A. RAID 0

B. RAID 1

C. RAID 5

D. RAID 10

Storage and Retrieval Methods

Q8 : What is the main benefit of data sharding in a distributed database system?

A. Improved data compression

B. Enhanced data redundancy

C. Better scalability and load balancing

D. Faster backup processes

Storage and Retrieval Methods

Q9 : Which of the following is a primary challenge in data retrieval from archived data?

A. Ensuring data redundancy

B. Maintaining data integrity

C. Achieving high retrieval speed

D. Balancing cost with performance

Storage and Retrieval Methods

Q10 : What is the first stage in the data lifecycle management process?

A. Data archiving

B. Data storage

C. Data creation and ingestion

D. Data deletion

?

Thank You !!!