DAY 6 **DATE:05/05/2025**

27739 – Annie John

TITLE: IMPLEMENTATION OF DATA DEDUPLICATION IN WINDOWS 2019, CLUSTER

- **What is Data Deduplication?**

Data Deduplication is a technique used to eliminate duplicate copies of repeating data. Instead of storing the same chunk of data multiple times, it stores it once and references it wherever needed.

Example: - If you have 100 identical copies of a file, deduplication stores one copy and uses pointers for the rest.

- **Necessity of Data Deduplication**

Reason	Why It Matters
Saves Disk Space	Reduces storage costs by storing only unique data chunks
Improves Backup Efficiency	Backup data often has lots of duplication; deduplication reduces size
Reduces Bandwidth Usage	Less data needs to be transferred over the network
Optimizes Storage Performance	Fewer writes = better performance in many workloads

✓ Used heavily in:

- File servers
- Virtual machine storage (VHD/VHDX)
- Backup targets
- Cloud storage

- **What is a Storage Pool?**

A Storage Pool is a group of physical disks combined to act as a single storage unit. You use it to create virtual disks.

Benefits:

- Combines multiple disks into one manageable unit
- Enables fault tolerance (mirror, parity)
- Allows for thin provisioning (allocate space as needed)
- Supports tiering (performance vs. capacity)

- **What is a Virtual Disk?**

A Virtual Disk is a logical disk created from a storage pool. You can think of it as a software-defined hard disk.

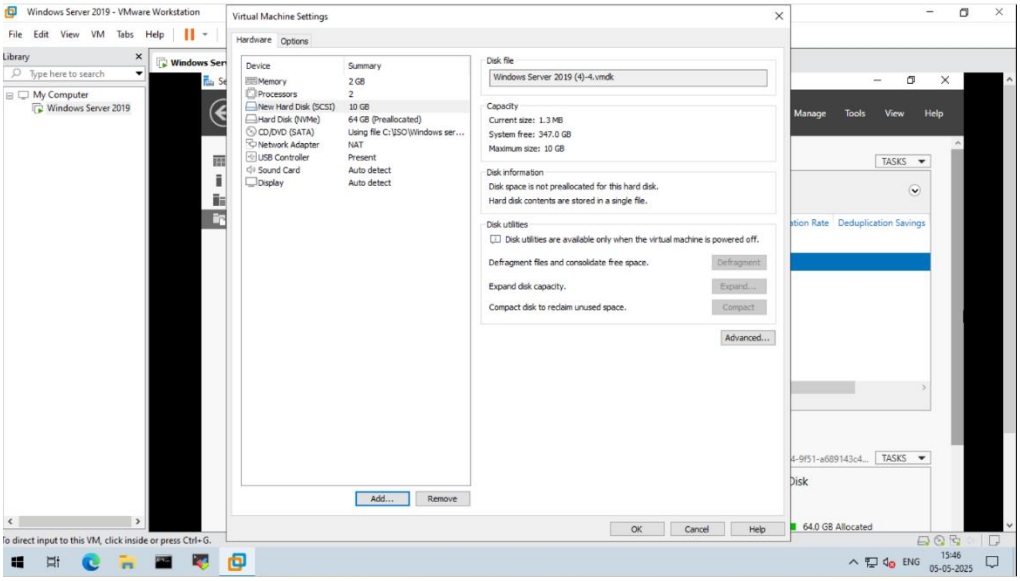
- Can be mirrored or parity-protected
- Appears like a regular disk in Disk Management
- On this virtual disk, you create volumes, format with NTFS, and store data

- **Implementation of data deduplication in Windows 2019 Server on VMWARE**

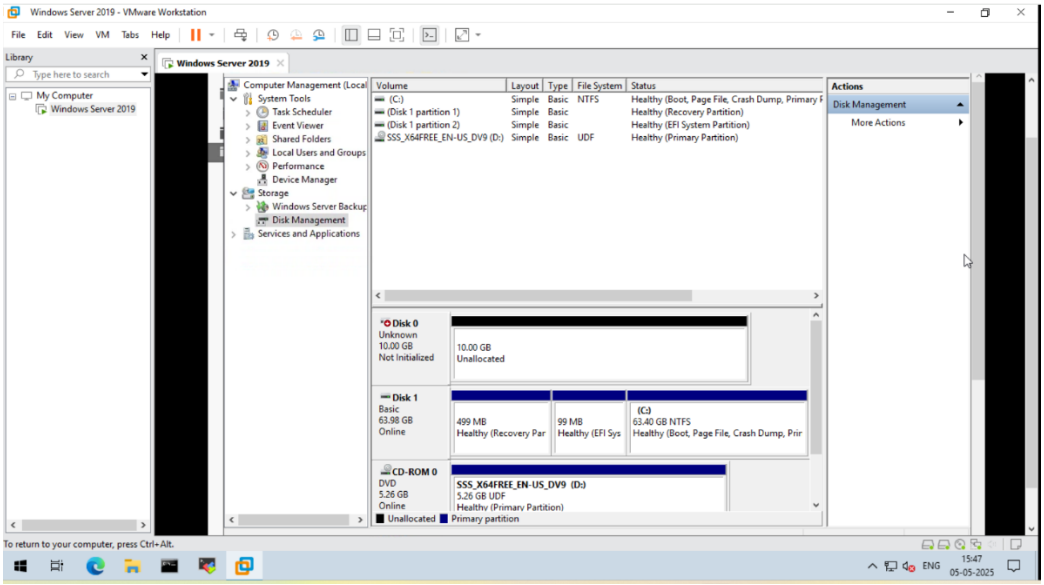
  Description:
  This guide presents a detailed walkthrough for implementing Data Deduplication within a Windows Server environment hosted on VMware. It systematically covers each stage—from enabling the Data Deduplication feature via Server Manager, to configuring storage pools, provisioning virtual disks, formatting volumes with NTFS, activating deduplication, and using PowerShell for performance monitoring. This guide supports effective storage management and optimization in virtualized infrastructures.

  - Step 1: Add a New Virtual Disk (if needed).
    - ✓ In VMware Workstation:

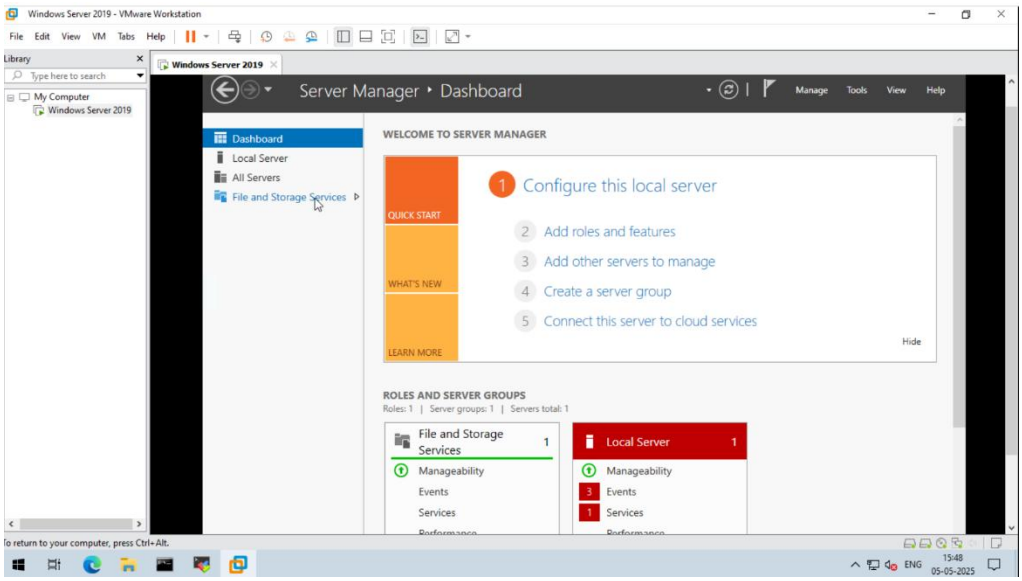
1. Go to VM -> Settings
2. Click Add -> Hard Disk
3. Select SCSI or NVMe
4. Choose Create a new virtual disk
5. Set size (e.g., 20 GB or more)
6. Complete the wizard and click OK



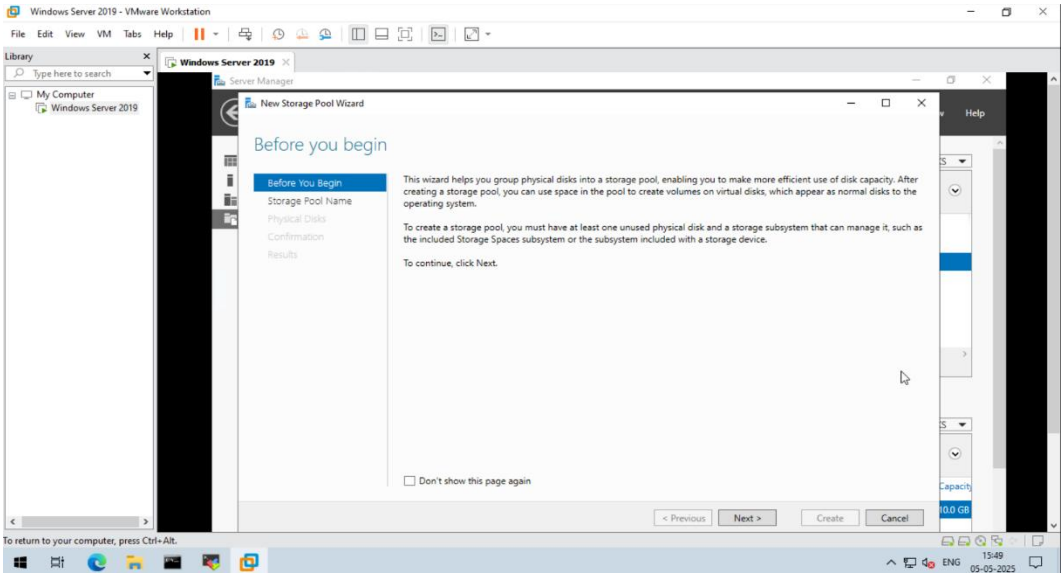
- Step 2: Go to disk management and make the disk online, don't initialize the disk.



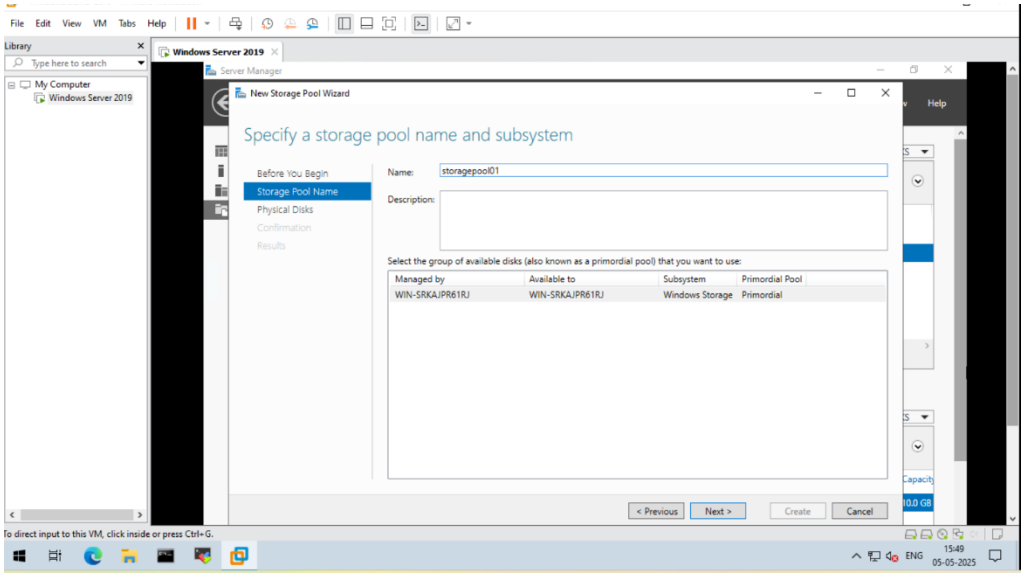
- Step 3: No w go to server manager->open file and storage services.



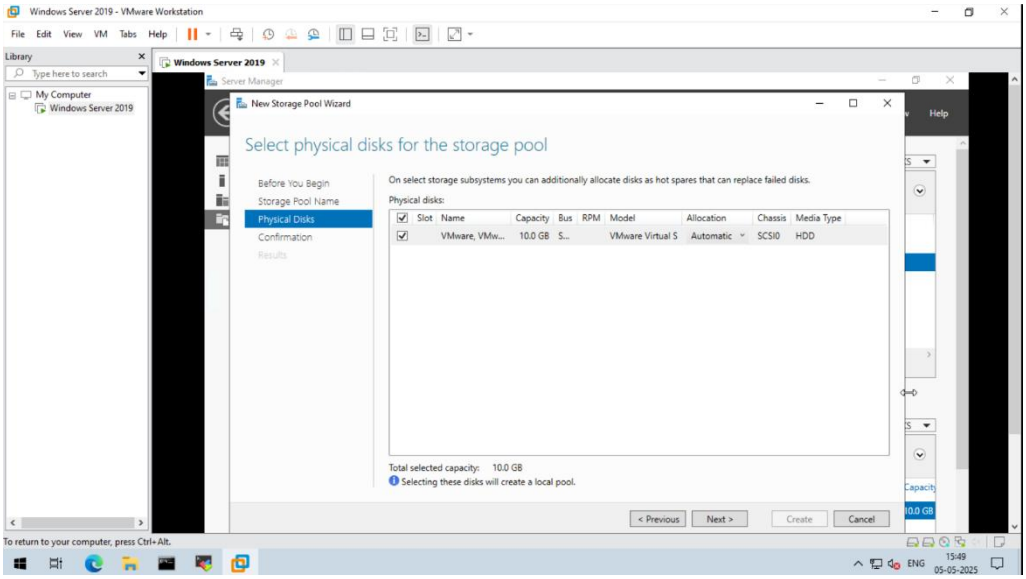
- Step 4: Under volume select the option storage pool->right click select new storage pool, it will get directed to a window new storage pool wizard.



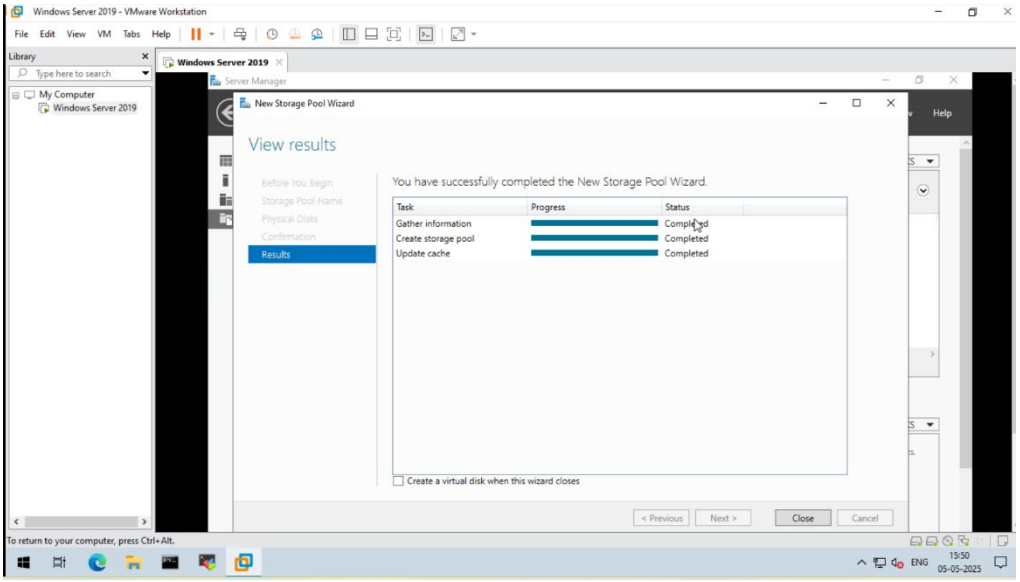
i. Click Next->specify a storage pool name and subsystem->Next



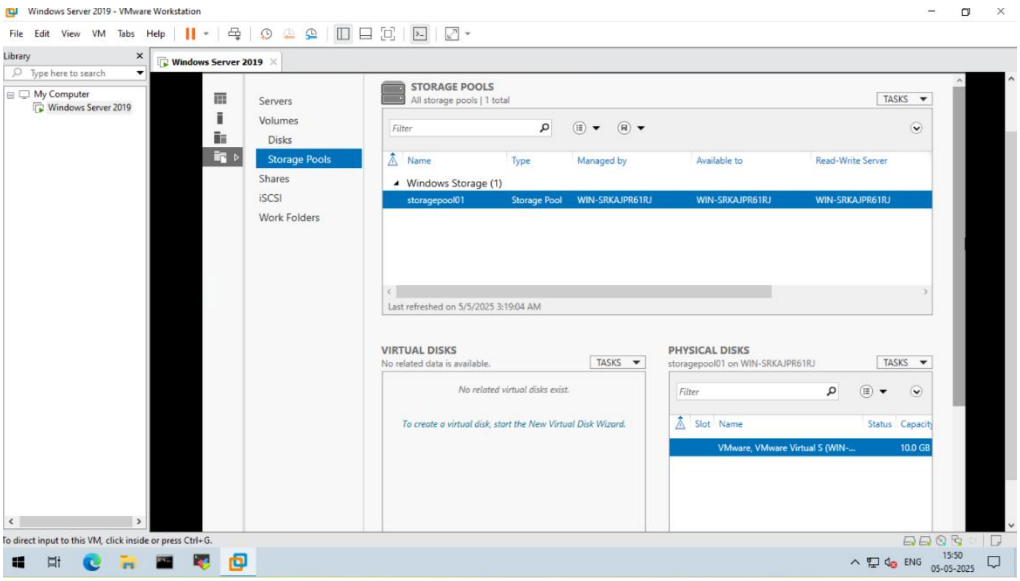
ii. Select checkbox of the virtual disk we created physical disks for storage pool->Click Next.



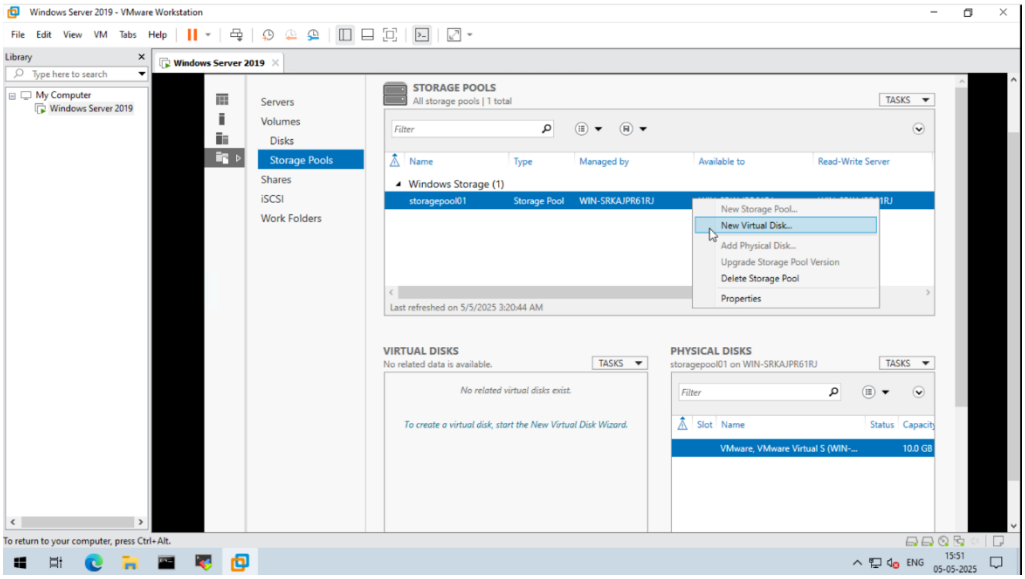
iii. Click on create.



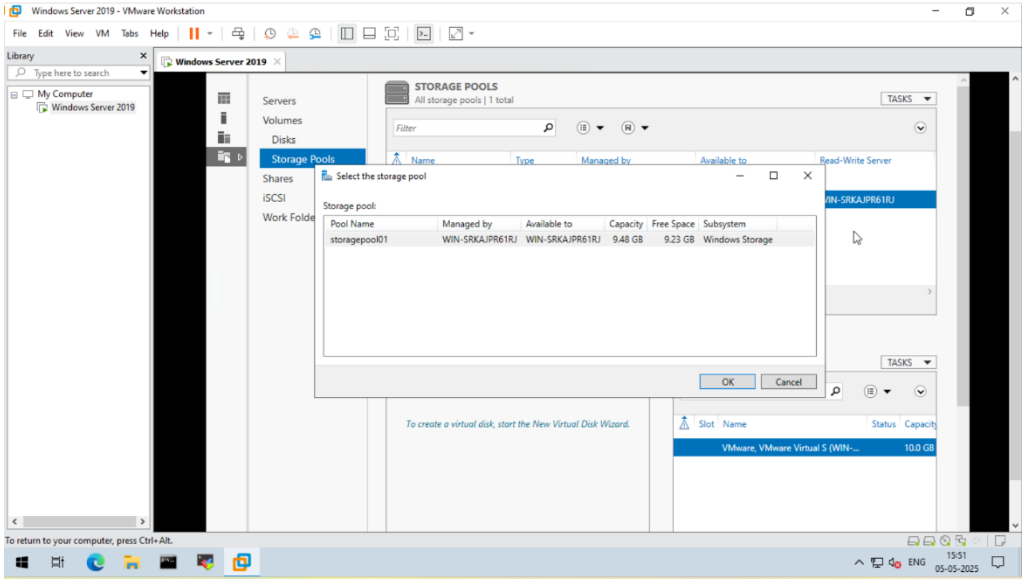
iv. Completed close the tab.



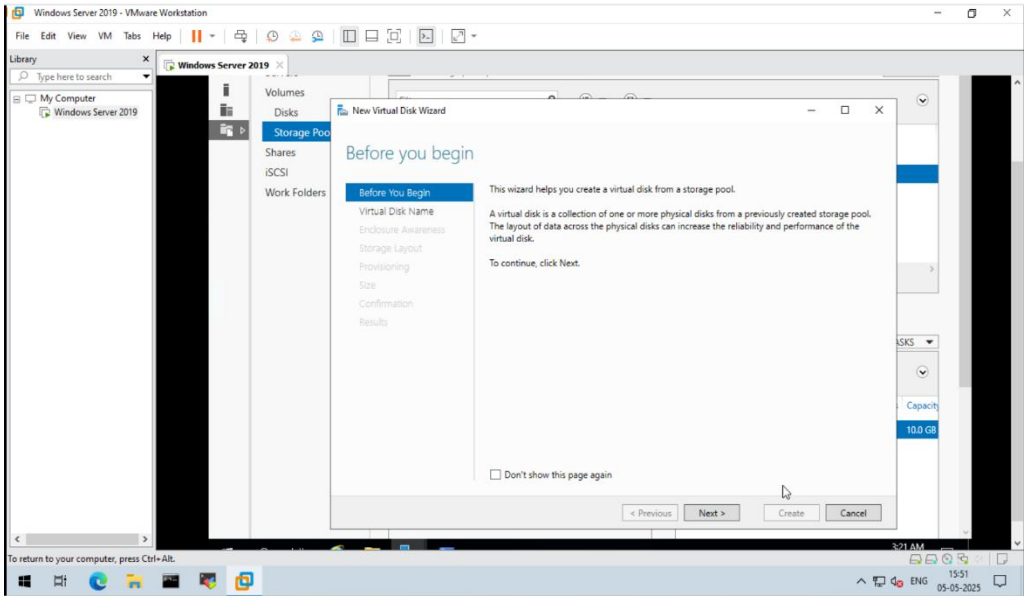
- Step 5: Now right click on the storage pool created->Click New Virtual Disk.



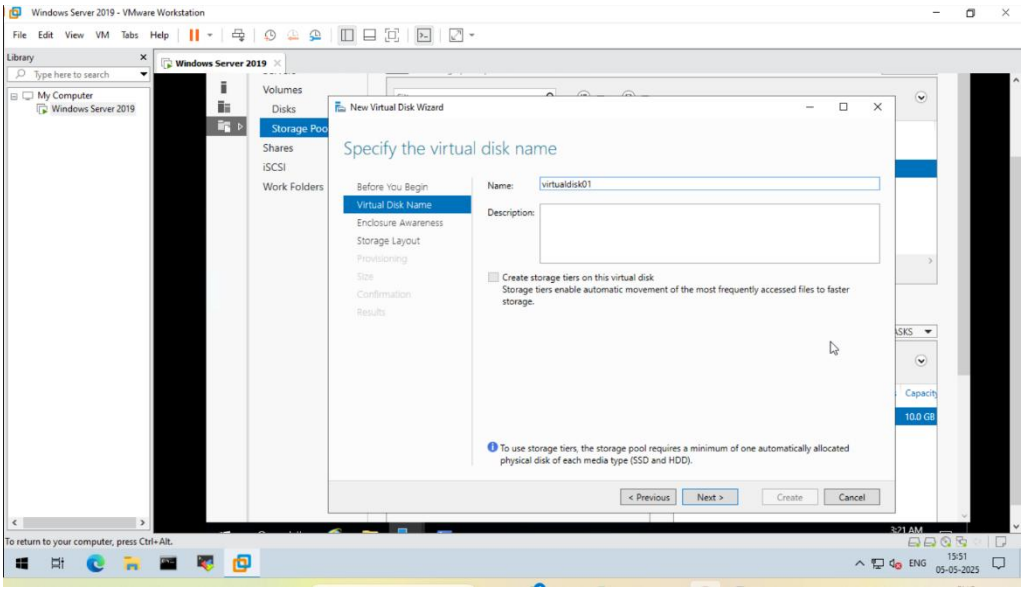
1. Select the storage pool ->ok



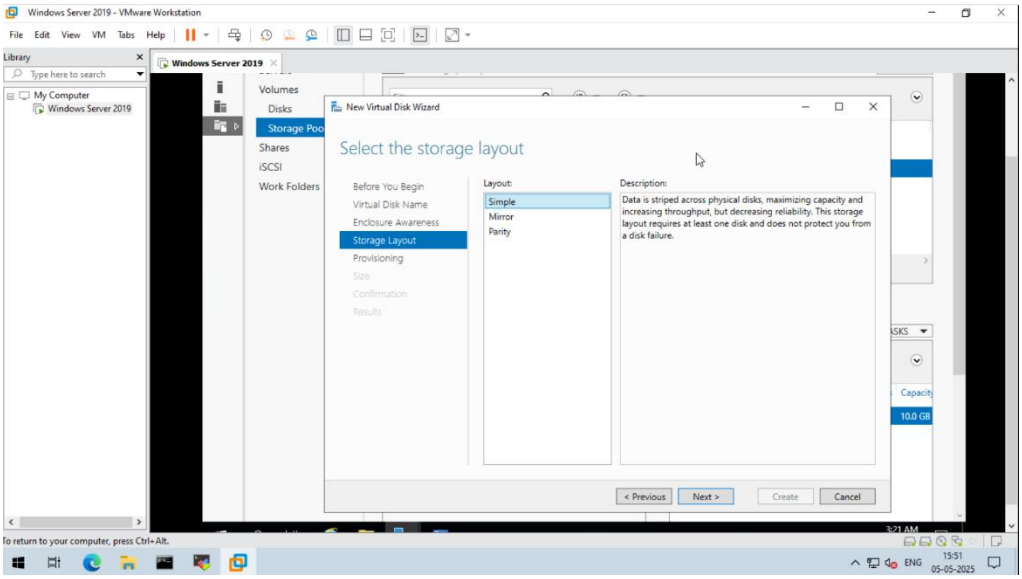
2. The it gets directed to the tab New Virtual disk wizard-> Click Next.



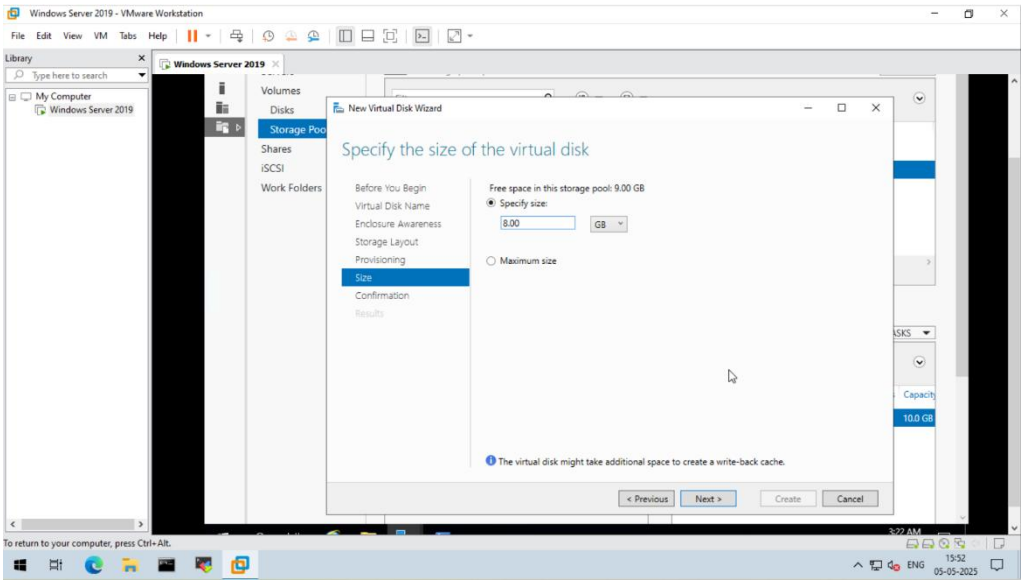
3. Give Virtual disk a name->then click Next.



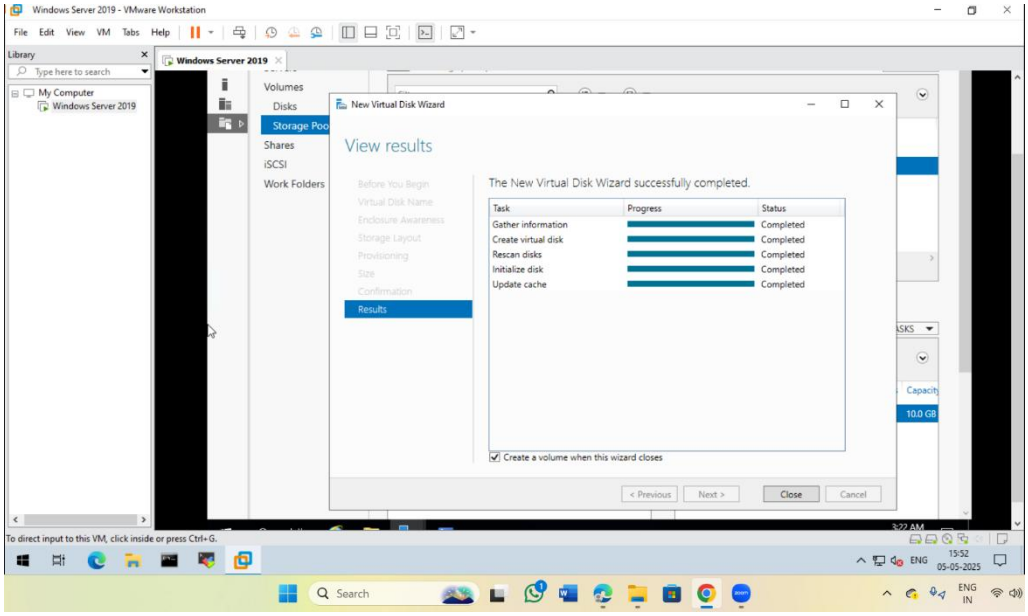
4. Click Next until you see this page->Storage layout->select simple(since only 1 disk is here).



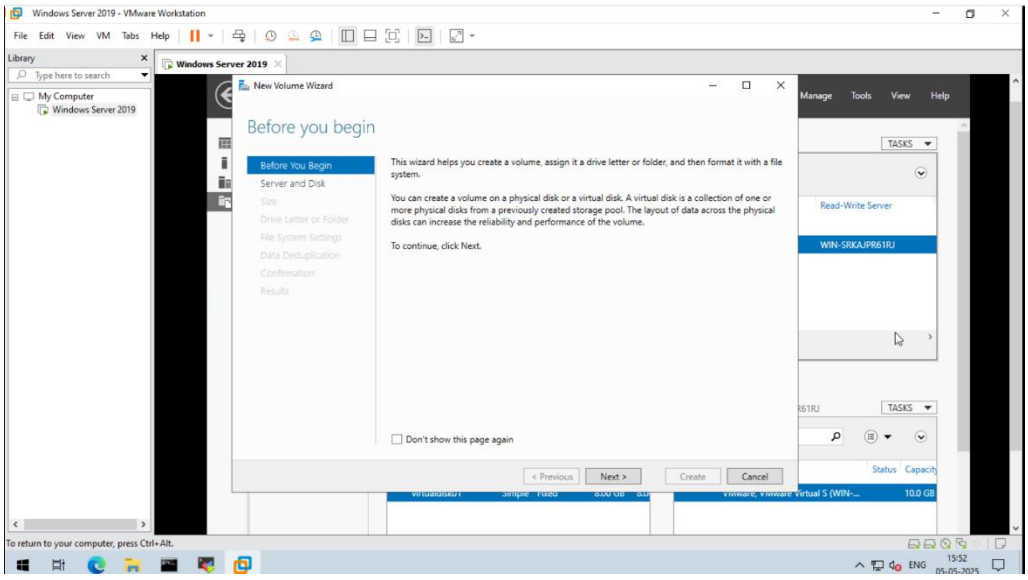
5. Click Next until you see this tab->specify the size of virtual disk.



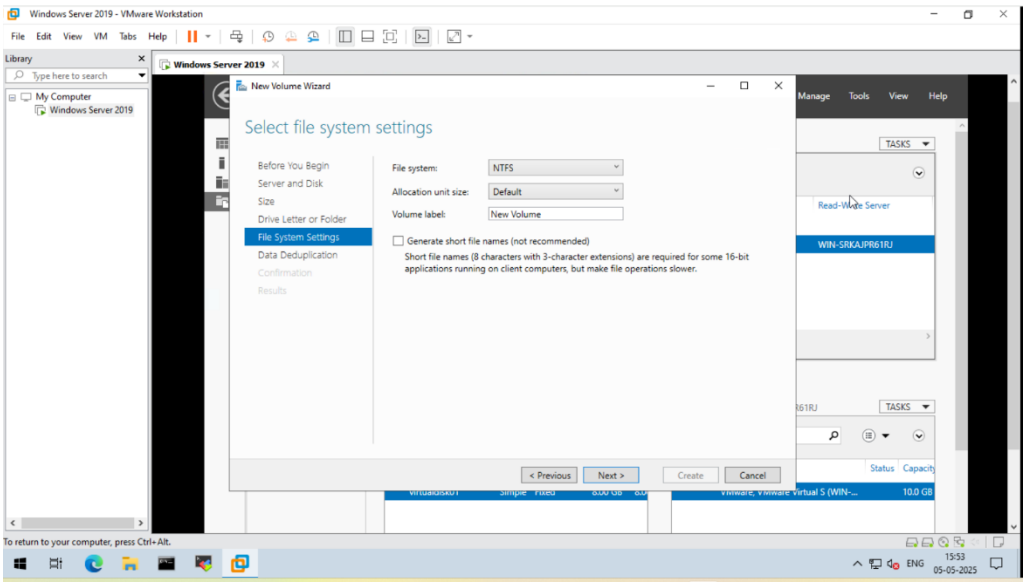
6. Click Next and create.



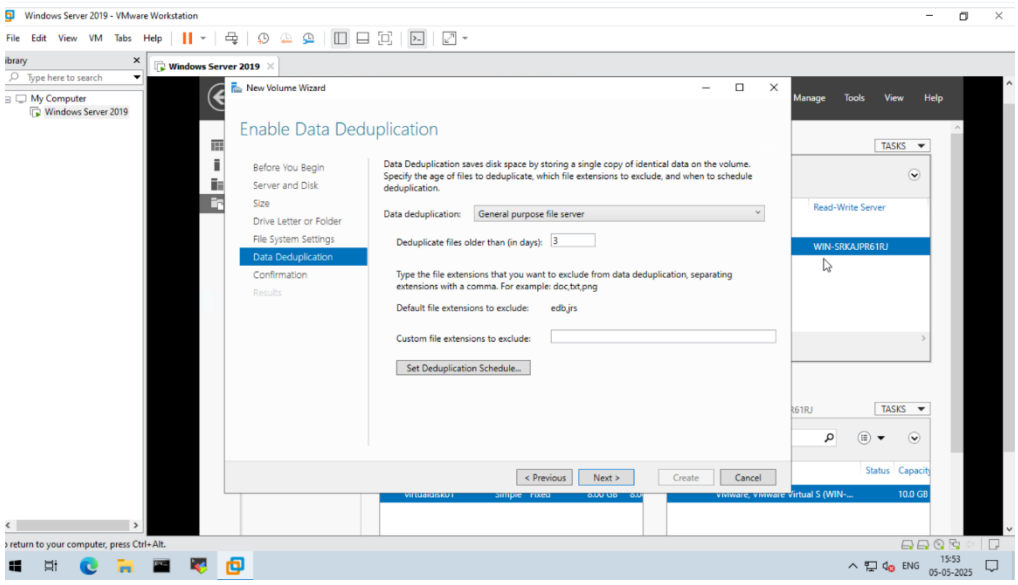
- Step 6: When the virtual disk is created it automatically redirects to new volume wizard tab to create volume.



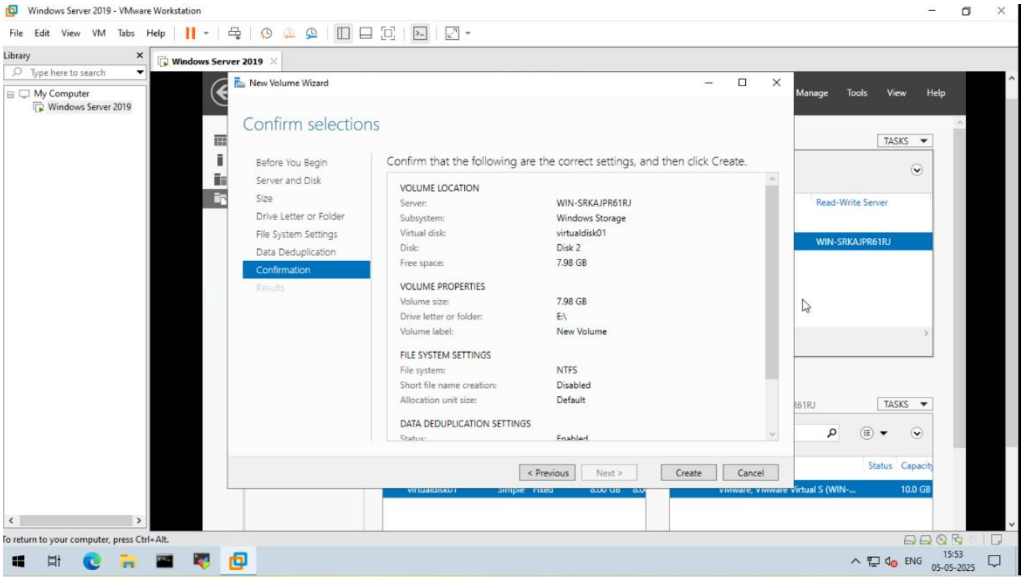
1. Click Next until you see this page->select NTFS in the file system->Next.



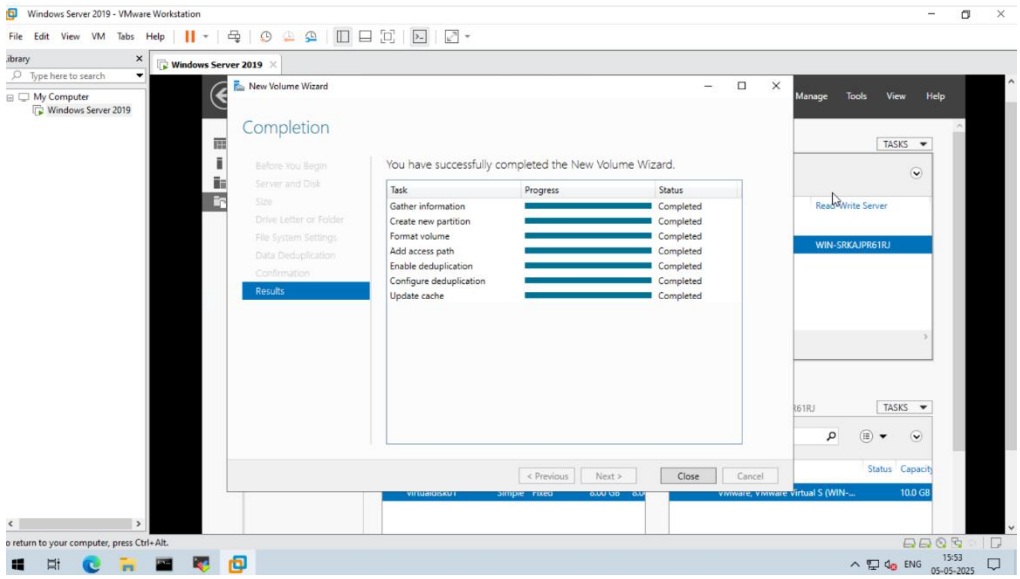
2. In data deduplication->select general purpose file server ->Next.



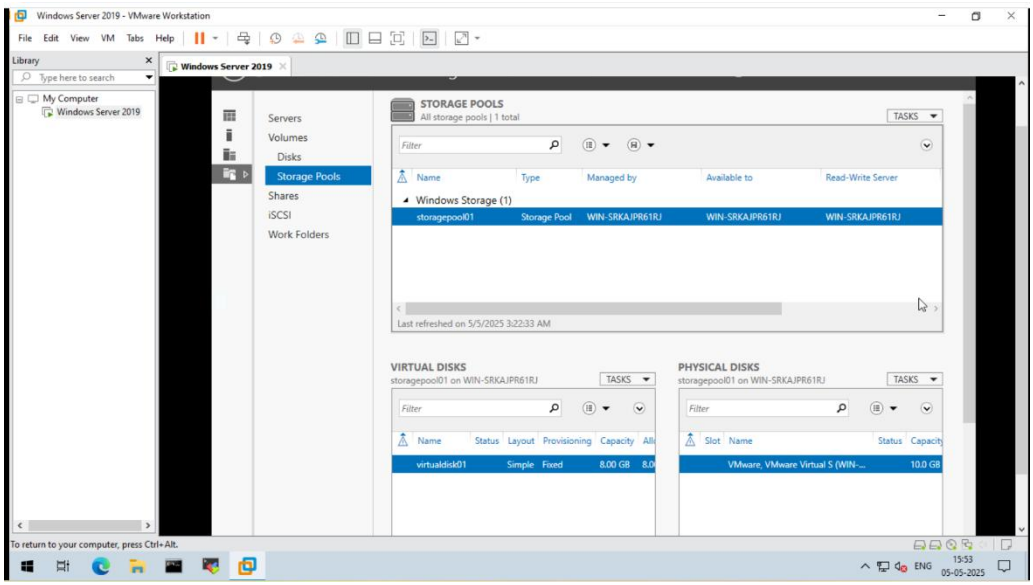
3. Confirm selection->create.



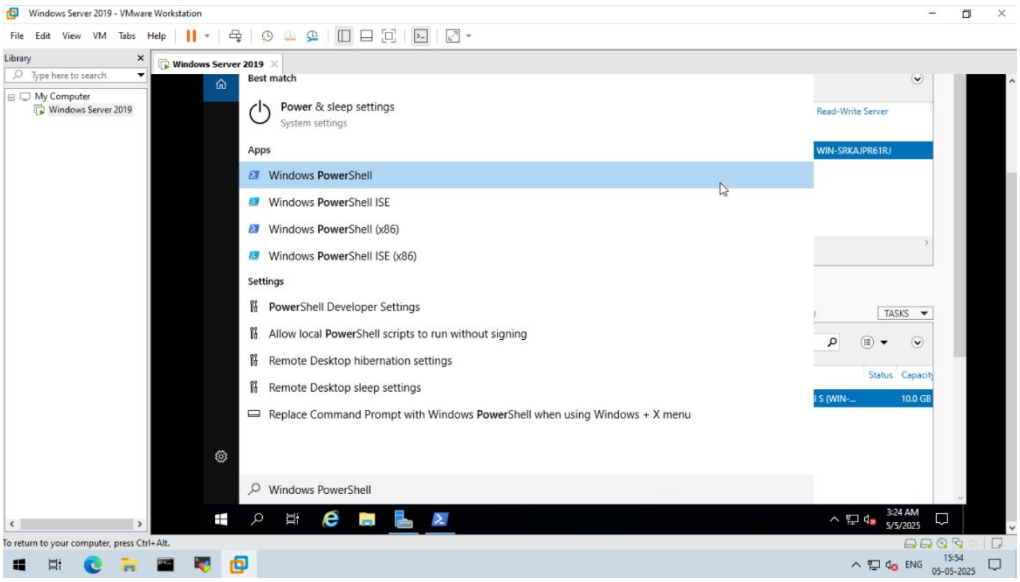
4. Complete->close.



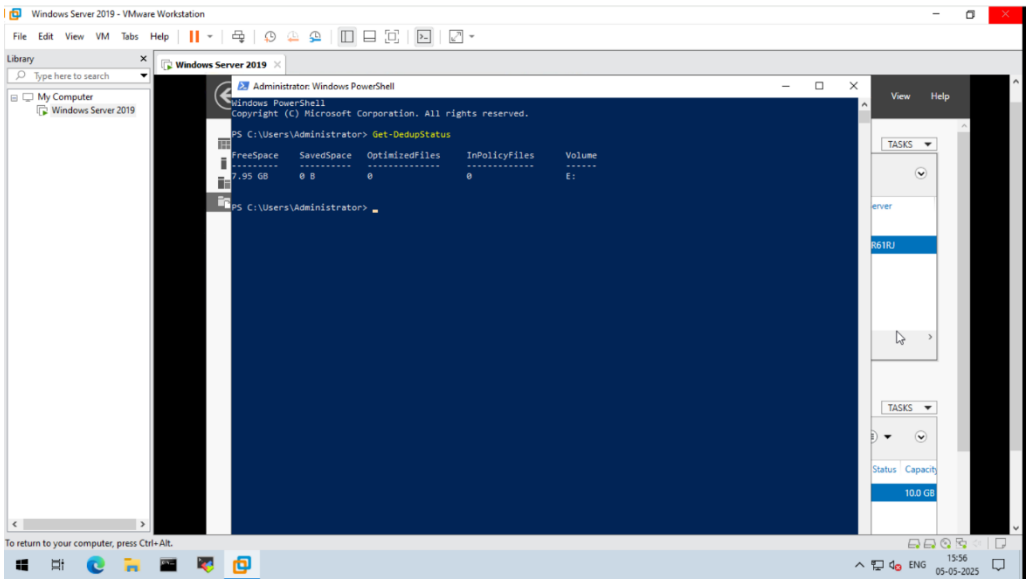
5. Now we have created storage pool, virtual disk and physical disk.



- Step 7: Now go to windows->search Microsoft PowerShell and open it.

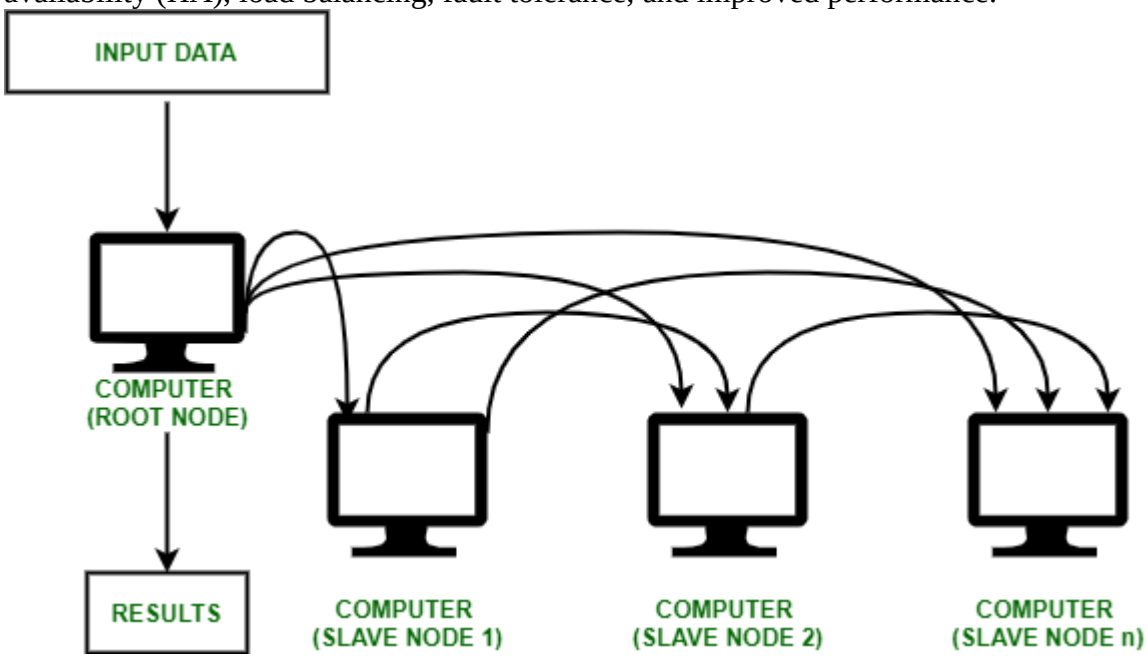


- Step 8: verify deduplication by typing command Get-DedupStatus.
    1. Get-DedupStatus - Is a PowerShell command used to check the status and savings of data deduplication on a volume in Windows Server 2019.



➢ What is a Cluster?

A cluster in computing refers to a group of connected computers (nodes) that work together so they can be viewed as a single system. Clusters are used for high availability (HA), load balancing, fault tolerance, and improved performance.



➢ Types of Clusters

- High Availability (HA) Cluster

Purpose: Ensures services keep running if one node fails.

Used for: Databases, web servers, file servers.

How it works: If Node A fails, Node B takes over (failover).

Example: Microsoft Failover Clustering.

- Load Balancing Cluster

Purpose: Distributes workloads evenly across multiple nodes.

Used for: Web servers, application servers.

How it works: Incoming traffic is shared among available nodes to prevent overload.

Example: Network Load Balancing (NLB) in Windows Server.

- High Performance Computing (HPC) Cluster

Purpose: Solves complex calculations using multiple computers working together.

Used for: Scientific simulations, research, engineering.

How it works: Splits a large task into smaller parts and processes them in parallel.

Example: Linux Beowulf cluster.

- **What is Fault Tolerance?**

  Fault tolerance is the ability of a system—especially a server, network, or application—to continue operating even if some of its components fail.

- **Comparison Between Load Balancer and Cluster.**

  Load Balancer-A Load Balancer is like a traffic cop—it directs incoming requests to the best available server to balance the load and avoid congestion.

  Cluster-A Cluster is like a team—a group of servers working together to provide continuous service, so if one fails, the others take over without downtime.

- A **normal cluster** typically refers to a group of servers or machines working together to perform tasks, but it may not be optimized for high availability, fault tolerance, or scalability. It often lacks features like load balancing, which can make it less resilient and flexible in handling high traffic or failure scenarios.

  On the other hand, a **native cluster**, often designed for cloud-native environments, integrates features like **load balancing**, ensuring even distribution of traffic across nodes, **high availability**, ensuring that if one node fails, others can take over without downtime, **fault tolerance**, which ensures the system continues to function despite failures, and **scalability**, enabling the system to expand or contract based on demand. Native clusters are designed with these principles in mind to optimize performance and resilience.