**Basic Definition & Examples**

**1. What is a Cluster?**

* A **cluster** is a group of interconnected computers (or servers) that work together as a single system to provide higher availability, scalability, and fault tolerance.
* **Example**: A database cluster can handle more users or data because multiple machines share the load.

**Key Features**:

* If one machine fails, others can take over (high availability).
* Tasks can be distributed among machines (load balancing).

**2. What is a Node?**

* A **node** is an individual machine (physical or virtual) within a cluster.
* Each node can perform specific tasks and communicate with other nodes in the cluster.

**Example**:

* In a Hadoop cluster, each server (or node) processes part of the data.

**3. What is Replication?**

* **Replication** is the process of copying data from one node or system to another to ensure **redundancy** and **high availability**.
* It ensures data is not lost if one node fails.

**Example**:

* In a database cluster, data on Node A is replicated to Node B and Node C, so if Node A crashes, the system can continue using data from Node B or C.

**Use Cases**:

* Disaster recovery.
* Load balancing (users can read data from multiple replicas).

**4. What is Sharding?**

* **Sharding** is the process of splitting a large dataset into smaller, more manageable pieces (called **shards**) that are stored on different nodes.
* Each shard contains only a portion of the data.

**Example**:

* A large user database can be sharded so:
  + User data with IDs 1–1000 is stored in Node A.
  + IDs 1001–2000 are in Node B, and so on.

**Advantages**:

* Improves performance by spreading the load across multiple nodes.
* Makes it easier to scale as the dataset grows.

**5. What is Parallel Processing?**

* **Parallel processing** refers to the ability to divide a task into smaller subtasks and process them simultaneously across multiple processors or nodes.
* It’s used to handle large tasks more quickly.

**Example**:

* Processing a large dataset by dividing it into chunks and analyzing them simultaneously on multiple CPUs or nodes in a cluster.

**Benefits**:

* Speeds up task execution.
* Utilizes resources more effectively.

**Summary Table**

| **Term** | **Description** | **Example** |
| --- | --- | --- |
| **Cluster** | Group of interconnected machines working together. | A Kubernetes cluster with multiple servers. |
| **Node** | A single machine in the cluster. | A server in a Hadoop cluster. |
| **Replication** | Copying data to multiple nodes for redundancy. | Database backups on multiple servers. |
| **Sharding** | Dividing data into smaller pieces stored separately. | User database split by user ID ranges. |
| **Parallel Processing** | Executing multiple tasks simultaneously. | Running data analysis on multiple CPUs. |