HDFS

1. **Big Bookshelf (Cluster):**
   - HDFS is like a giant bookshelf made up of many smaller shelves. Each shelf is actually a computer (called a node) in a cluster, and together they store all the data.
2. **Book Storage (Files and Blocks):**
   - Instead of storing one big book on a single shelf, you break it into smaller chunks (like pages) and store these chunks on different shelves. These chunks are called blocks. In HDFS, each file is split into blocks, and each block is stored across multiple computers in the cluster.
3. **Copying for Safety (Replication):**
   - To make sure you don't lose a book if one shelf breaks, you keep copies of each chunk of the book on different shelves. HDFS does this by replicating each block multiple times (usually three). So, if one computer fails, you can still get the block from another computer.
4. **Index Card System (NameNode):**
   - To find a particular book or chunk of a book, you need a catalog that tells you where each chunk is stored. This catalog is managed by a special computer called the NameNode. It keeps track of which chunk is on which computer but doesn't actually store the chunks itself.
5. **Library Helpers (DataNodes):**
   - The actual chunks of the books are stored on the computers called DataNodes. These are like the helpers who manage the books on the shelves. They handle the reading and writing of data.
6. **Finding the Book (Data Retrieval):**
   - When you need to read a book, you ask the NameNode where the chunks are located. The NameNode tells you which DataNodes have the chunks, and then you can get the chunks from those DataNodes and put them together to read the book.

In summary, HDFS is designed to store large amounts of data across many computers, ensure that the data is reliably available even if some computers fail, and make it easy to find and access the data when needed.

YARN

Imagine your library not only stores books but also hosts reading events, book signings, and other activities. To manage all these different events efficiently, you need a system to organize who gets which space and resources (like tables, chairs, or staff assistance) at what times. YARN does something similar for managing computing resources in a big data system.

Here's how it works:

1. **Library Manager (Resource Manager):**

- Think of the Resource Manager as the head librarian or the main manager of your library. This person decides who gets which space and resources. In YARN, the Resource Manager keeps track of all the available computing resources across the entire system and allocates them to various tasks or applications.

2. **Event Organizers (Application Masters):**
   - Each event in the library needs its own organizer. For instance, a book signing might have its own team, and a reading event might have another. In YARN, these organizers are called Application Masters. Each Application Master is responsible for managing its own application's resources and tasks. They request resources from the Resource Manager and then oversee the execution of their application.
3. **Library Staff (Node Managers):**
   - The Node Managers are like the library staff who help set up the space and resources for each event. They handle the day-to-day operations on each individual shelf or room. In YARN, Node Managers are responsible for managing resources on their specific computers (nodes) and reporting their status back to the Resource Manager. They ensure that the required resources are available and report on how the resources are being used.
4. **Event Spaces (Containers):**
   - The physical spaces in the library (like meeting rooms or reading corners) are similar to containers in YARN. Containers are isolated spaces where specific tasks or processes run. Just as different events might use different spaces in the library, different tasks run in different containers on the computing nodes.
5. **Scheduling Events (Resource Allocation):**
   - When a new event needs to be scheduled, the Resource Manager checks availability and allocates the required space and resources. Similarly, when a new application or job needs resources, the Resource Manager decides which Node Managers will provide those resources and how they will be allocated.
6. **Monitoring and Reporting (Resource Monitoring):**
   - The Resource Manager keeps an eye on the overall usage of resources and ensures everything is running smoothly. It gets regular updates from the Node Managers about resource usage and availability. If something goes wrong, the Resource Manager can adjust resource allocation to address any issues.

HDFS ARCHITECTURE:

Imagine you have a huge collection of books, and you want to keep them safe and easily accessible. HDFS, or Hadoop Distributed File System, is like a special library system designed to handle massive amounts of data across many computers. Here's a simple breakdown of how it works:

1. **Big Library with Many Shelves**: HDFS treats data like books in a library. Instead of storing everything in one big bookshelf (or computer), it spreads the data across many shelves (or computers). This helps manage large amounts of data more efficiently.
2. **Master Librarian**: There's one main librarian called the NameNode. This librarian knows where every book is located in the library but doesn't actually hold any books. Instead,

the NameNode keeps a catalog of where each book (or piece of data) is stored across the shelves.

3. **Shelf Managers**: Each shelf in the library has a Shelf Manager called a DataNode. These Shelf Managers actually store the books and manage the day-to-day borrowing and returning of books.
4. **Copies for Safety**: To prevent losing books if a shelf is damaged, each book is copied and stored on several different shelves. This way, even if one shelf gets damaged or loses books, there are copies available on other shelves.
5. **Borrowing Books**: When you want to read a book, you tell the main librarian (NameNode) what book you want. The librarian then tells you which shelf (DataNode) has your book. If the book is copied across several shelves, the system can quickly find and provide the book from one of those shelves.

In summary, HDFS is like a huge, well-organized library system for data, where a central catalog keeps track of where everything is stored, and multiple shelves ensure that books are safely and efficiently managed and accessible.