Big Data Overview
Big data refers to datasets too large for traditional databases to handle in terms of capture, storage, and analysis. It is characterized by:

- **Volume**: Large data quantities.
- **Velocity**: High speed of data generation/processing.
- **Variety**: Diverse data types (structured, semi-structured, unstructured).

Introduction to Hadoop and HDFS

- A **file system** indexes the physical location of data on storage devices.
- A **distributed file system (DFS)** stores files across multiple servers, accessible by authorized clients in a network.
- **Hadoop Distributed File System (HDFS)** is Hadoop's distributed file system, designed for:
    - Storing **very large files**.
    - Supporting **streaming data access**.
- HDFS is **not suitable** for:
    - Low-latency data access.
    - Handling many small files.

HDFS Blocks

- Like traditional file systems, HDFS divides files into **blocks**, but these are much larger:
    - 64 MB in Hadoop 1.x, 128 MB in Hadoop 2.x (compared to typical disk block sizes of 512 bytes).
- Each block is **replicated** (default: 3 copies) across different nodes for fault tolerance.

HDFS Components (Hadoop 1.x)

1. **NameNode**:
    - Master node managing the file namespace and metadata.
    - Tracks which **DataNodes** store blocks for each file.
    - Single point of failure (SPOF).
2. **Secondary NameNode**:
    - Not a backup NameNode but maintains a merged copy of the namespace image to mitigate SPOF risks.
3. **DataNodes**:
    - Workhorse nodes that store and manage blocks.
    - Report block information to the NameNode.
4. **Job Tracker and Task Trackers**:
    - **Job Tracker**: Coordinates job execution.
    - **Task Trackers**: Execute smaller tasks (job splits) assigned by the Job Tracker.

HDFS Storage Architecture

- The **NameNode** oversees the file system metadata and block locations.
- **DataNodes** (e.g., Datanode1, Datanode2, etc.) physically store the data blocks, with replication ensuring data reliability and availability.