HDFS(Hadoop Distributed File System)

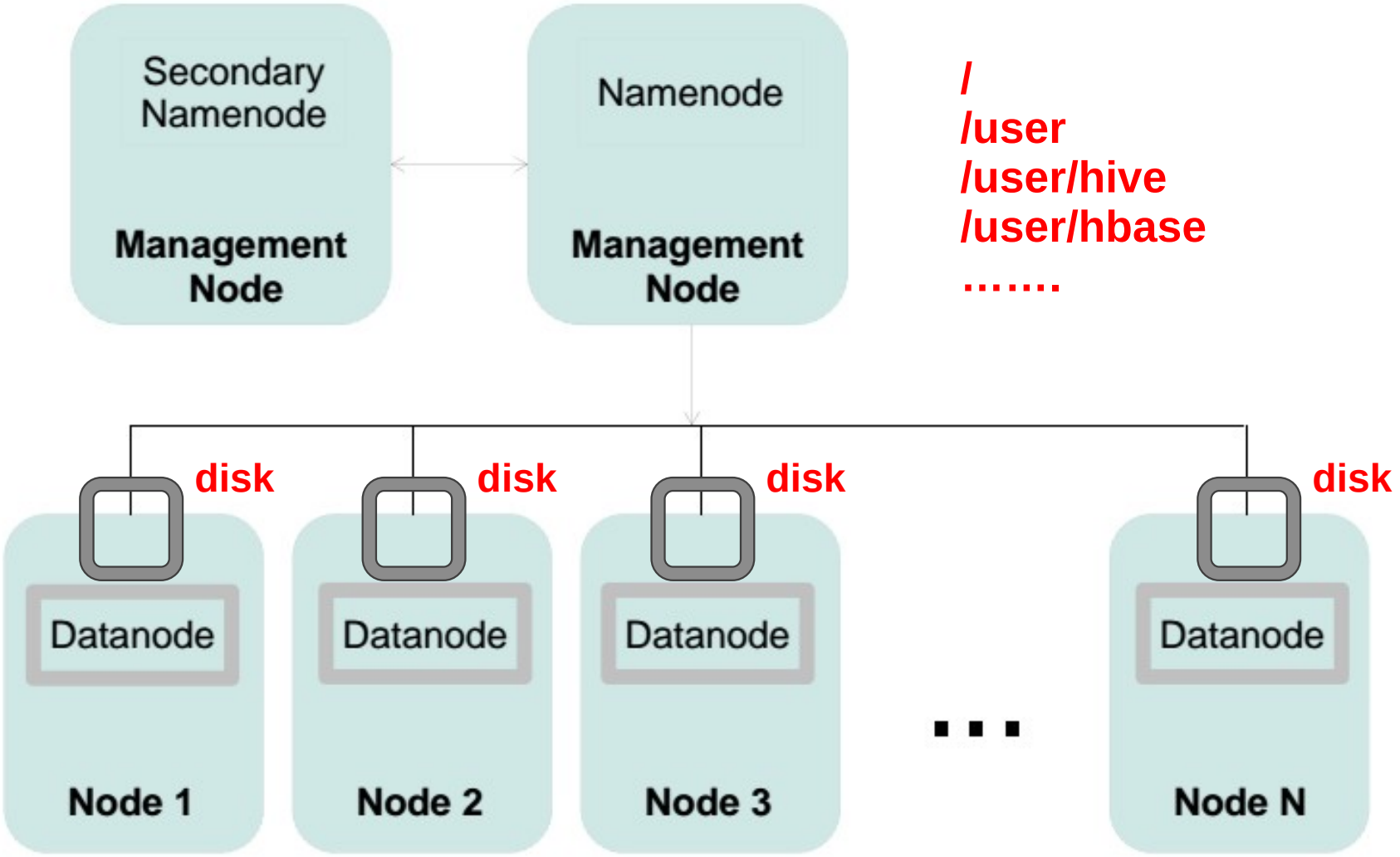
Agenda

- What is HDFS?
- HDFS Concepts
- Namenode & Datanode
- Files and Blocks
- Block Replication

What is HDFS?

- **HDFS is a filesystem designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware.**
- **Appears as a single disk**
- **Runs on top of a native filesystem**
  - Ext3,Ext4,XFS
- **Based on Google's Filesystem GFS**

HDFS Concepts



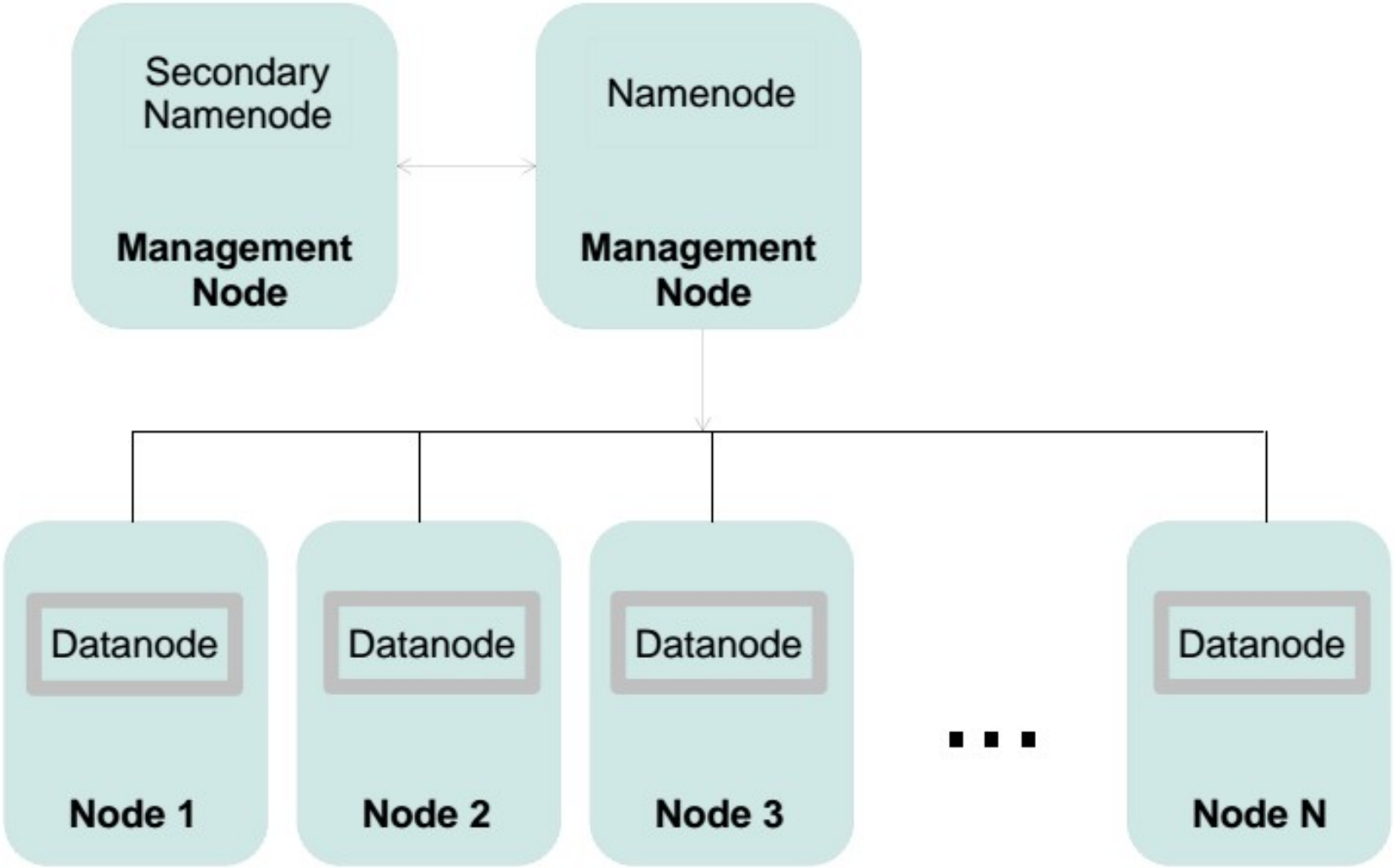
HDFS works well with

- **Very large Files**
  - Files that are hundreds of megabytes, gigabytes,or terabytes in size.
- **Streaming data access**
  - Write-once, read-many-times pattern.
  - The time to read the whole dataset is more important than the latency in reading the first record.
- **Commodity hardware**
  - It doesn’t require expensive, highly reliable hardware.

HDFS is not a good fit for-

- **Low-latency data access**
  - HDFS is optimized for delivering a high throughput of data.
  - Not good for applications that require low-latency access to data(ms response).
  - HBase is currently a better choice for low-latency access.
- **Lots of small files**
  - Namenode holds filesystem metadata in memory, governing the limit to the number of files.
- **Multiple writers, arbitrary file modifications**
  - No support for multiple writers or for modifications at arbitrary offsets in the file.

Namenodes and Datanodes(Master-Worker)



HDFS Daemons

- **Namenode**
  - manages the File System's namespace/meta-data/file blocks
  - Runs on 1 machine to several machines
- **Datanode**
  - Stores and retrieves data blocks
  - Reports to Namenode
  - Runs on many machines
- **Secondary Namenode**
  - It periodically merges the namespace image with the edit log to prevent the edit log from becoming too large.
  - Requires similar hardware as Namenode machine
  - Not used for high-availability – not a backup for Namenode

Namenode

- **It maintains the filesystem tree and the metadata for all the files and directories in the tree.**
- **This information is stored persistently on the local disk in the form of two files:**
  - namespace image
  - edit log.
- **It also knows the datanodes on which all the blocks for a given file are located.**

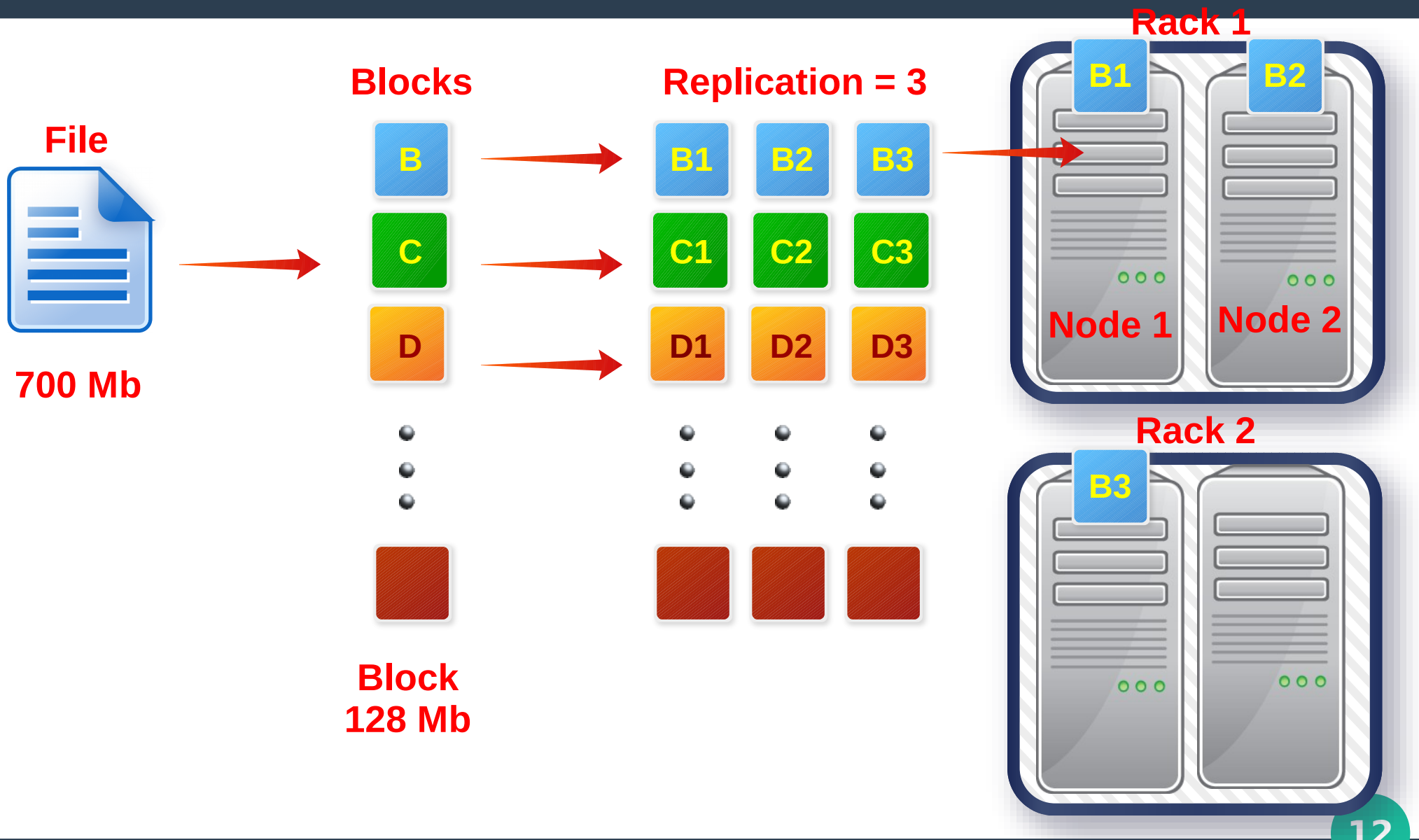
Datanode

- Datanodes are the workhorses of the filesystem.
- They store and retrieve blocks when they are told to (by clients or the namenode).
- They report back to the namenode periodically with lists of blocks that they are storing.

Files and Blocks

- **Files are broken into blocks**
- **Block- the minimum amount of data that it can read or write.**
  - 128 MB by default

Files and Blocks



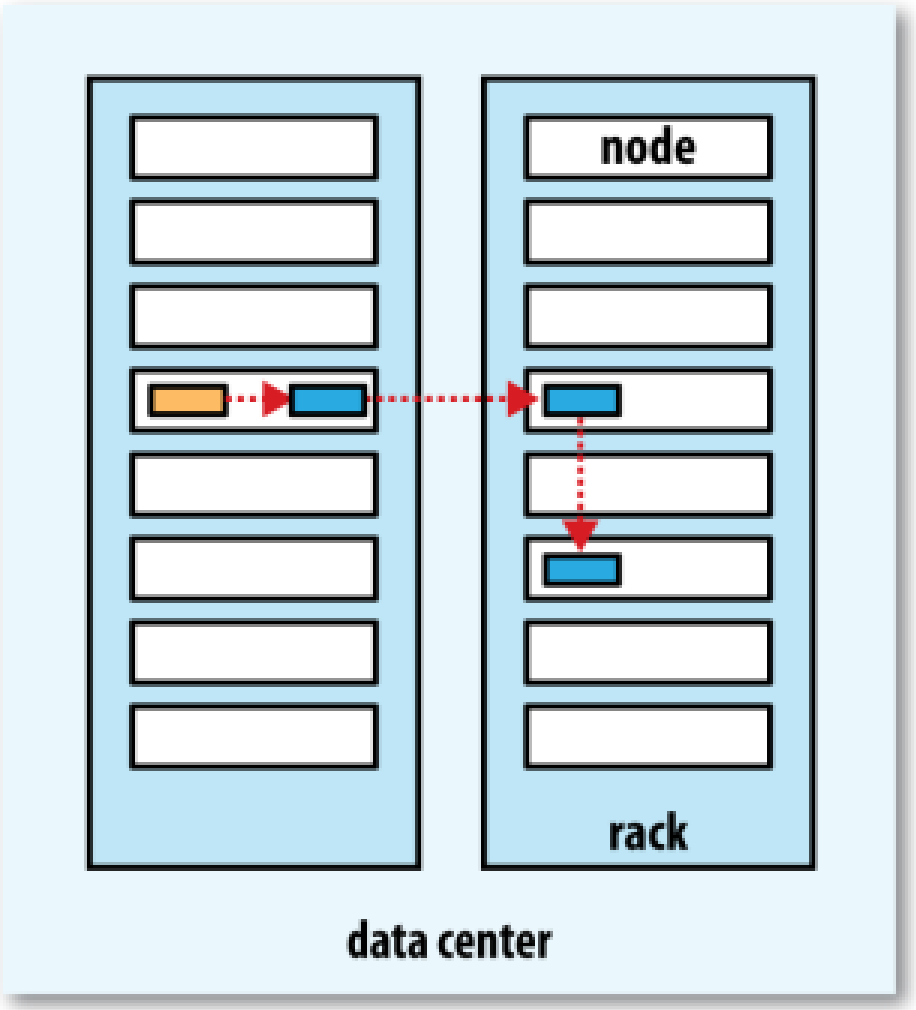
Files and Blocks

- File smaller than single block does not occupy a full block’s worth of underlying storage.
- Uses replication for providing fault tolerance and availability.

- **Block Replication**

- **Namenode determines replica placementReplica placements are rack aware**
  - 1st replica on the local rack
  - 2nd replica on the local rack but different machine
  - 3rd replica on the different rack

Block Replication



Resources

- **Hadoop: The Definitive Guide**
  - Tom White (Author)
  - O'Reilly Media; 4th Edition.