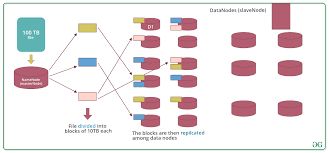
**HDFS**

**Hadoop Distributed File System** (HDFS) is an open-source file system that stores and manages large amounts of data for Hadoop applications. HDFS is a key component of Hadoop systems and is used by many companies to handle and store big data. 

**Here are some features of HDFS:**

* Fault-tolerant: HDFS is designed to be highly fault-tolerant and can scale to hundreds or thousands of nodes.
* Runs on commodity hardware: HDFS is designed to run on low-cost hardware.
* High throughput: HDFS provides high throughput access to application data.
* Streaming access: HDFS relaxes some POSIX requirements to enable streaming access to file system data.
* File permissions: HDFS file permissions are similar to other platforms like Linux.
* fsck command: HDFS supports the fsck command to check for inconsistencies in files.
* DFSAdmin command: The bin/hadoop dfsadmin command supports several HDFS administration operations.

**Benefits of HDFS**

### Fault tolerance and fast recovery from hardware failures

Because one HDFS instance might consist of thousands of servers, failure of at least one server is always a possibility.

HDFS has been built to detect faults and automatically recover quickly.

[Data replication](https://www.ibm.com/topics/data-replication) with multiple copies across many nodes helps protect against data loss. HDFS keeps at least one copy on a different rack from all other copies.

This [data storage](https://www.ibm.com/topics/data-storage) in a large cluster across nodes increases reliability.

In addition, HDFS can take storage snapshots to save point-in-time (PIT) information.

**Cost-efficiency**

Because the data is stored virtually, the costs for file system metadata and file system namespace data storage can be reduced.

**HDFS architecture and how it works**



**HDFS has a director/worker architecture**.

* An HDFS cluster includes one **NameNode**, which is the *director* server.
* The NameNode tracks the status of all files, the file permissions and location of every block. The NameNode software manages the file system namespace which in turn tracks and controls client access to the files and performs operations such as file opening, closing and renaming directories and files.  
    
  The file system namespace also divides files into blocks and maps the blocks to the DataNodes, which is the *worker* portion of the system.
* **By configuring with only a single NameNode per cluster, the system architecture simplifies**[**data management**](https://www.ibm.com/topics/data-management)**and storage of the HDFS metadata**.
* In addition, greater security is built in by keeping user data from flowing through the NameNode.