# HDFS

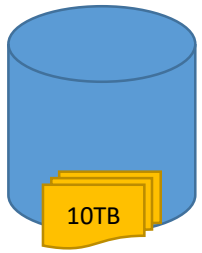
# Big Data Challenges

- Storage (Storing big data is a problem due to its massive volume and variety)
- Processing (Takes more time due to Huge Data)

# Solution : Hadoop HDFS

Storage Problem → 

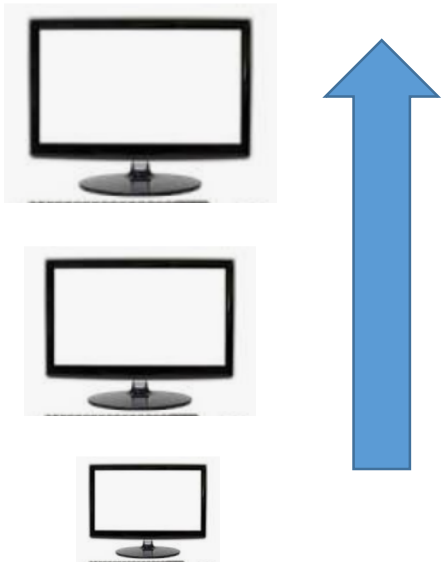

Processing Problem → 

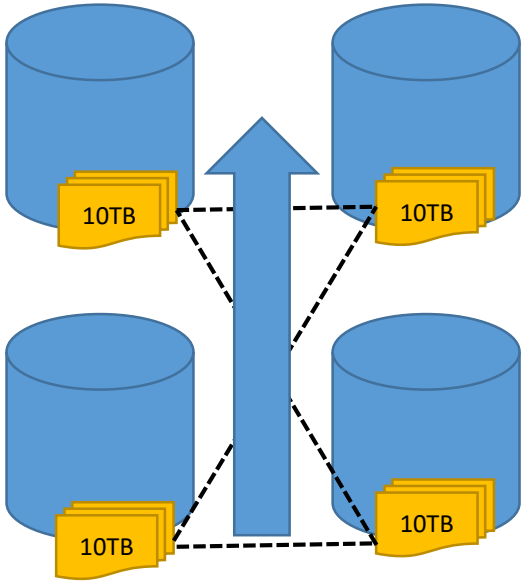



**Local File System**

## Vertical Scaling

*(Increase of size of Instance on the same Machine – RAM, CPU etc)*

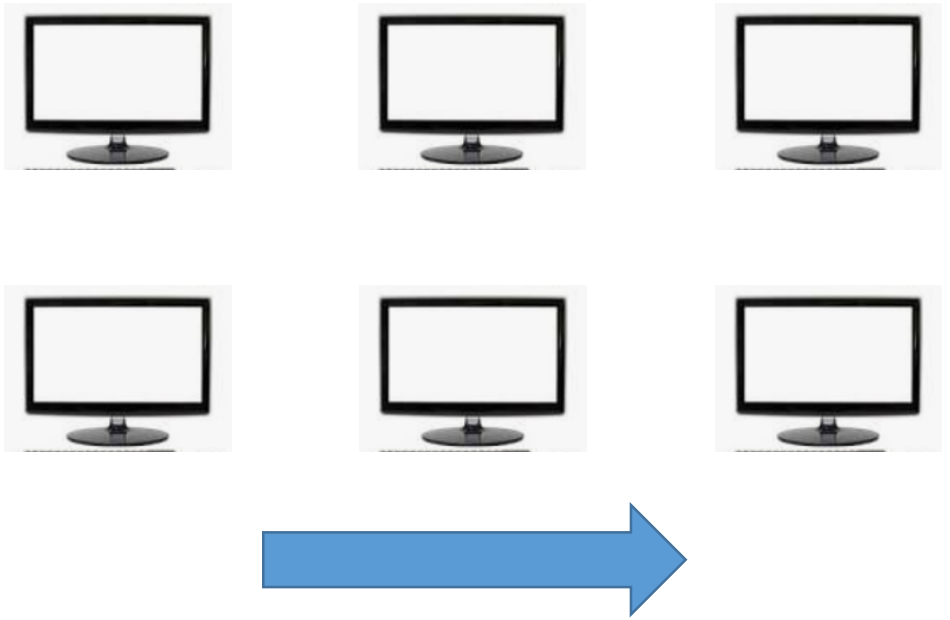





**Distributed File System(HDFS)**

## Horizontal Scaling

*(Add more instances)*



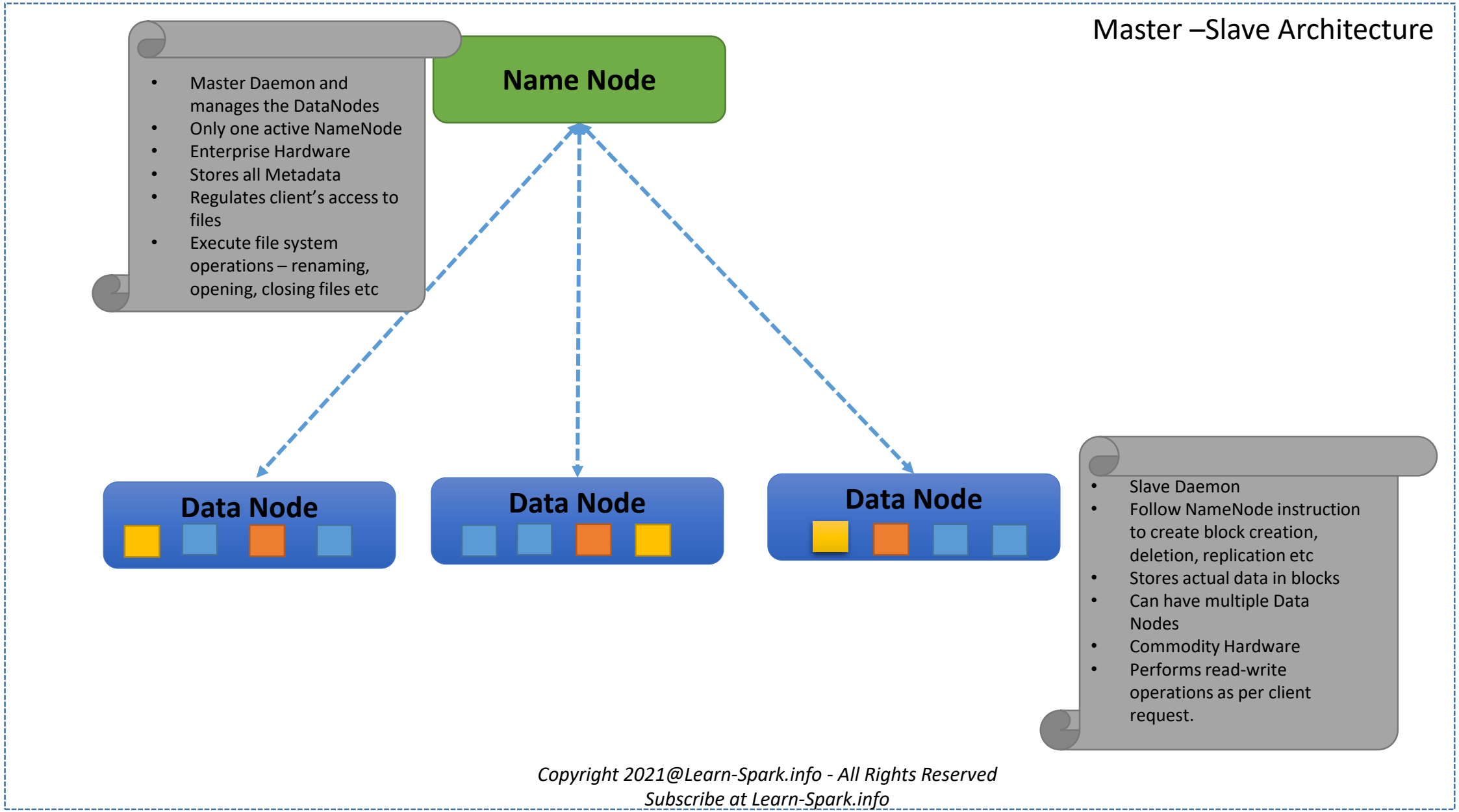
# What is Hadoop HDFS

HDFS is specially designed file system for storing massive amount of data in commodity hardware.

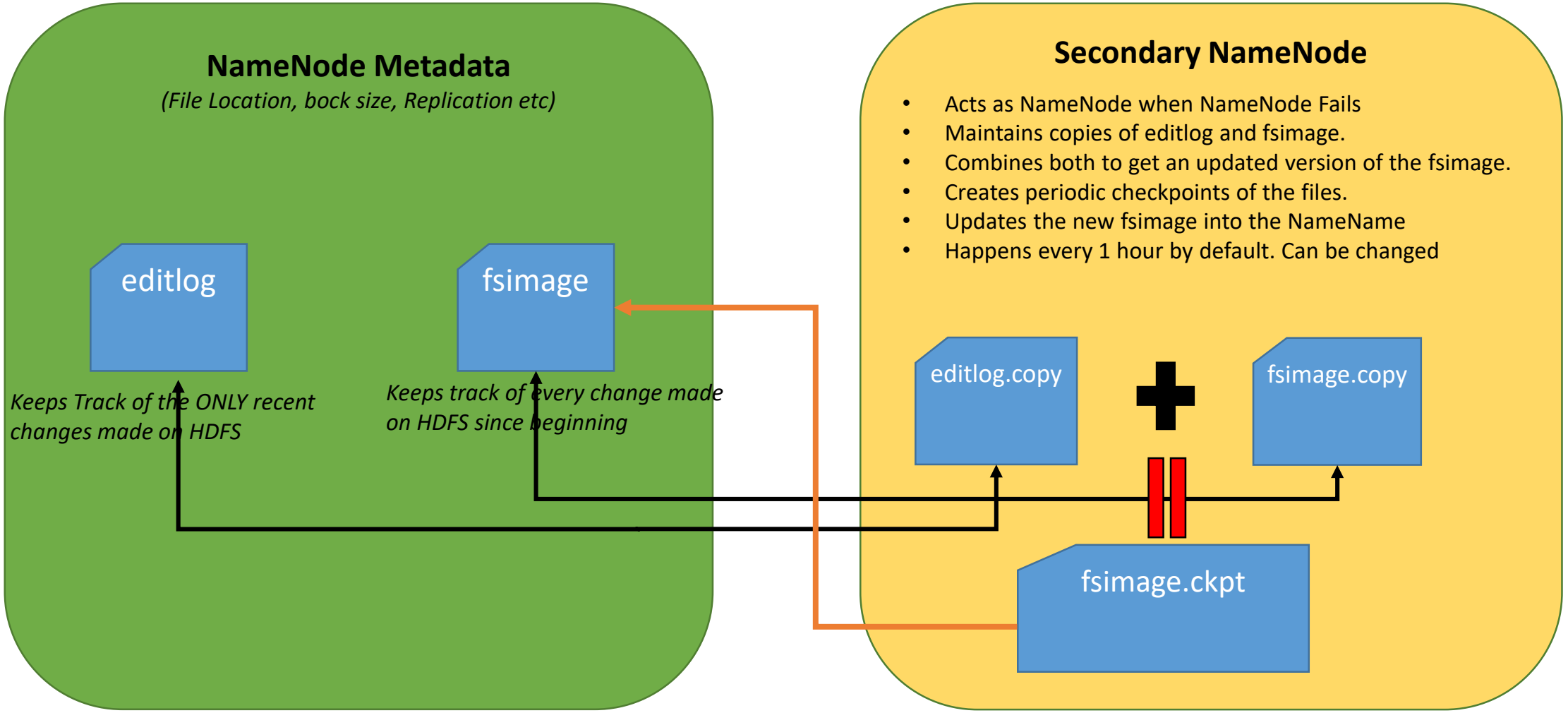
Advantages of HDFS:

1. Distributed storage for Big Data
2. Cost effective as uses commodity hardware
3. Fault-tolerant as data copies available – Replication.
4. Data is secure as provides data security

# HDFS Components



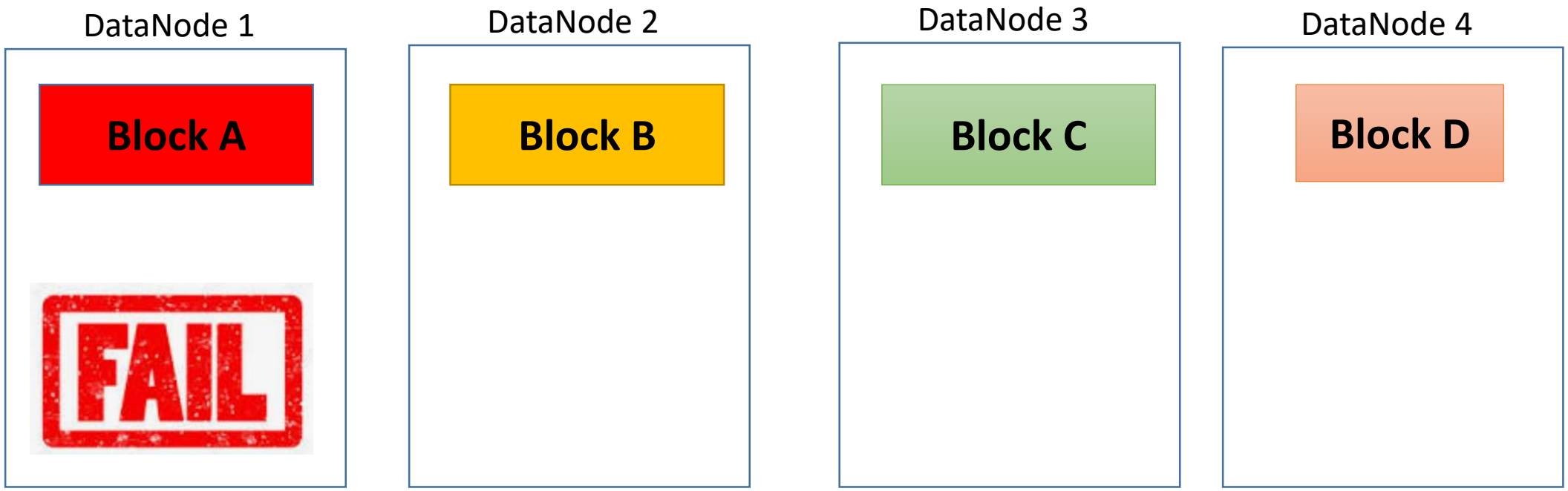
# MetaData



# HDFS Data Blocks

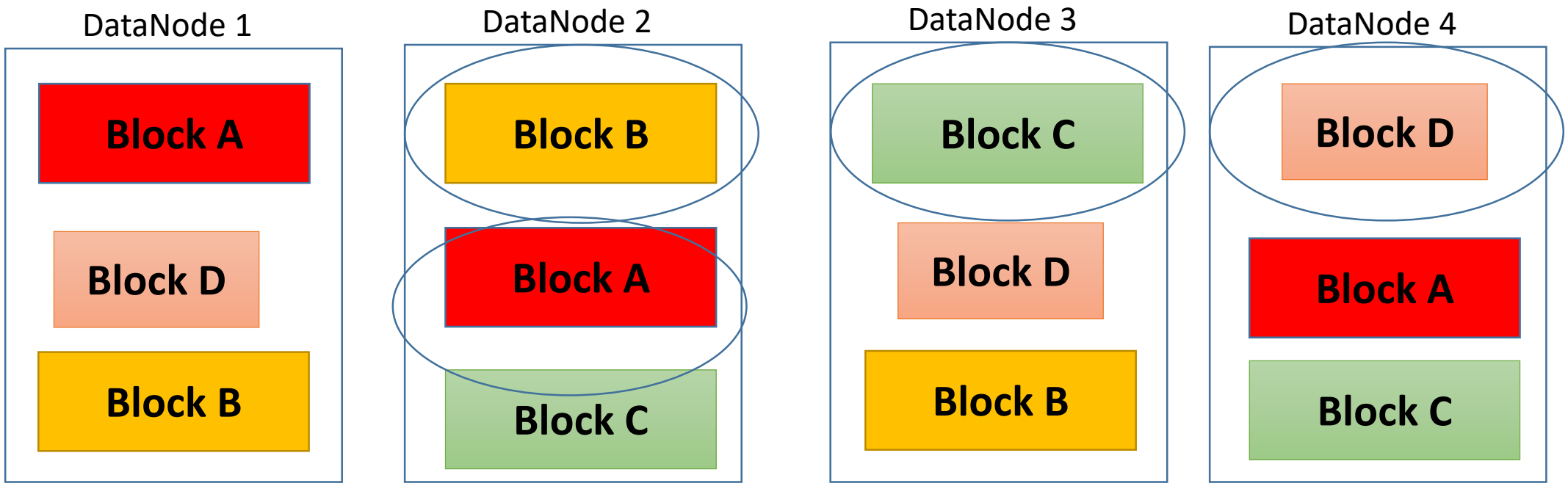
- HDFS Splits massive files into small split files – called Data Blocks.
- Default size 128MB in Hadoop 2.x and 64MB for Hadoop 1.x.
- Can be configured - /opt/hadoop/etc/hadoop/hdfs-site.xml
- Block size small – There will be too many data blocks and so lots of metadata which causes lot of overhead.
- Block Size Large – Processing size of each block increases.


# Replication



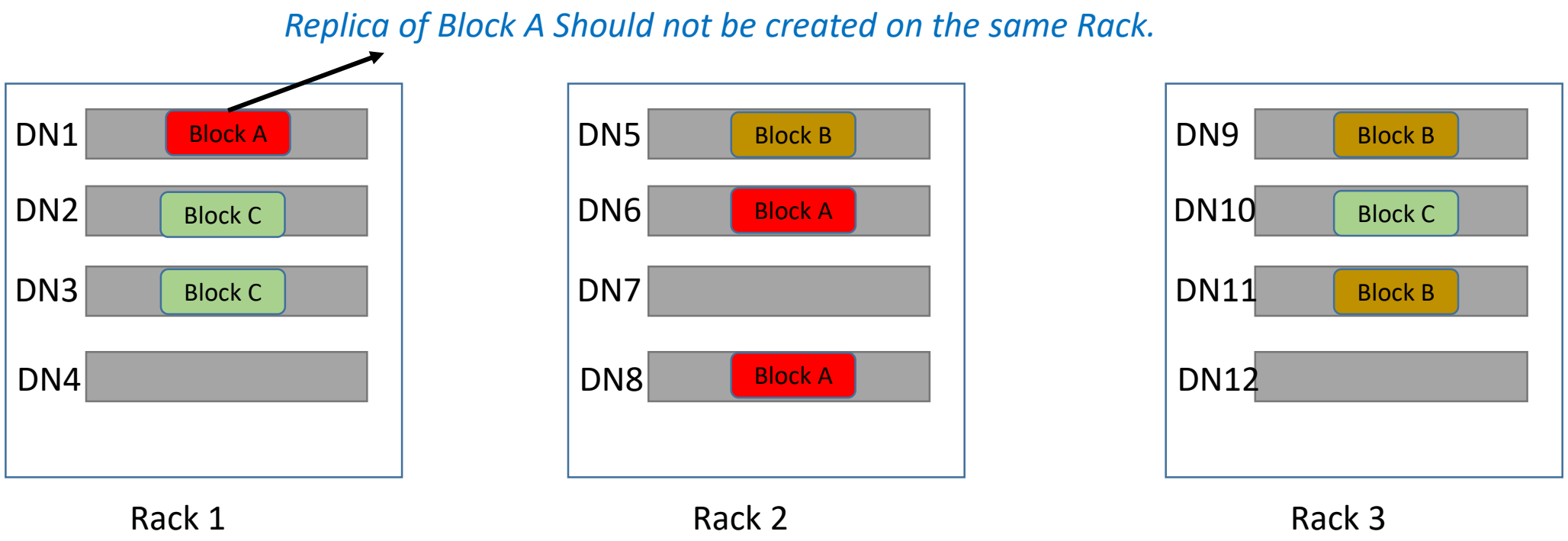
# Replication

Default Replication = 3. Can be set in hdfs-site.xml

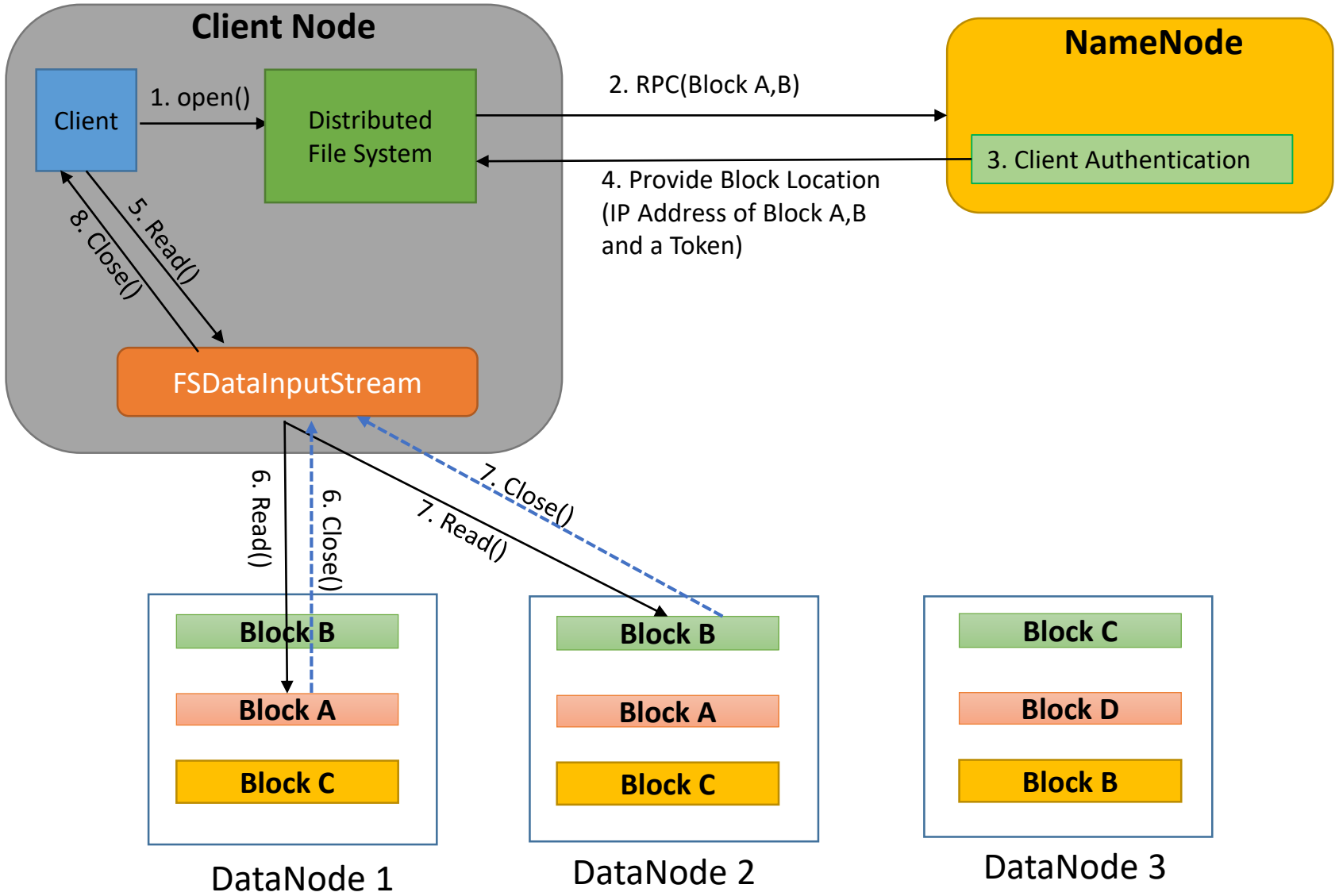


# Rack Awareness

- It is a concept that helps to decide where a copy of the data block should be stored.
- Rack is a collection of 30-40 DataNodes



# Exercise - HDFS Read Mechanism



# Exercise – (HDFS CLI Help Commands)

Hadoop provides a command Line Interface to interact with HDFS.

✓ *List all HDFS Commands.*
hadoop fs or hdfs dfs

✓ *Basic usage for any command.*
hadoop fs -usage ls

✓ *Full detailed information for any commands.*
hadoop fs -help ls

# Exercise (Get Data from GitHub to HDFS)

### Get data files from GitHub to our Unix System
git clone https://github.com/sibaramKumar/dataFiles

#Rename the Folder
cd dataFiles

### Unzip the Files
sudo apt install unzip
unzip SalesData.zip
ls –lrt
rm SalesData.zip

### Create a Folder at HDFS
hadoop fs -mkdir -p practice/retail_db/

### Copy the Files from Local to HDFS
hadoop fs -put dataFiles/* practice/retail_db/

# Exercise (List/Sort Files in HDFS)

- ✓ **Command –** ls
- ✓ Use a pattern to select.
- ✓ -R : Recursively list the contents of directories.
- ✓ -C : Display the paths of files and directories only.

By default it sorts in ascending order by name.

- ✓ ls –r : Reverse the order of the sort.
- ✓ ls –S : Sort files by size.
- ✓ ls –t : Sort files by modification time (most recent first).

# Exercise (Creating/Deleting Directories in HDFS)

- ✓ **Command : rmdir**
  - Remove Directory if it is empty.
  - --ignore-fail-on-non-empty : Suppress Error messages if the Folder you are trying to remove is non empty.
- ✓ **Command : rm**
  - rm remove files.
  - With option –r , recursively deletes directories
  - With option –skipTrash bypasses trash, if enabled.
  - With option –f, no error message even if file does not exist. Check with Unix command $?. It returns the status of last ran job. 0 means successful and 1 means not successful.
- ✓ **Command: mkdir**
  - Create Directory
  - –p : Do not fail if the directory already exists. Also we can create multiple folders recursively.

# Exercise - Copy from HDFS to Local

✓ Command - copyToLocal or get

```
hadoop fs -get practice/retail_db/orders .
```

✓ Error if the destination path already exists. To overwrite use –f flag.

```
hadoop fs -get practice/retail_db/orders .
hadoop fs -get –f practice/retail_db/orders .
```

✓ -p flag to preserves access and modification times, ownership and the mode.

```
hadoop fs -get -p practice/retail_db/orders .
```

✓ To Only copy the files with out folder use a pattern.

```
hadoop fs -get practice/retail_db/orders/* .
```

✓ When copying multiple files, the destination must be a directory.

```
mkdir copyHere
hadoop fs -get practice/retail_db/orders/* practice/sample.txt copyHere
```

# Exercise - Copy data from Local to HDFS

✓ Command - copyFromLocal or put

```
hadoop fs -mkdir –p practice/retail_db
hadoop fs -put dataFiles/* practice/retail_db/

hadoop fs -mkdir –p practice/retail_db1
hadoop fs -put dataFiles practice/retail_db1/  #Creates a subfolder dataFiles under retail_db
```

✓ Error if the destination path already exists. To overwrite use –f flag.

```
hadoop fs -put -f dataFiles/* practice/retail_db/
```

✓ -p flag to Preserves timestamps, ownership and the mode.

```
hadoop fs -put -p dataFiles/* practice/retail_db/
```

✓ We can also copy multiple files.

```
hadoop fs -put -f dataFiles/* sample.txt practice/retail_db/
```

# Showing Data in HDFS

- ✓ **Command – head**
  - Show the first 1KB of the file
- ✓ **Command – tail**
  - Show the last 1KB of the file
  - Option –f shows appended data as the file grows.
- ✓ **Command – cat**
  - Fetch the whole File
  - Not Recommended for Big Files.
- ✓ **First 10**

hadoop fs -cat practice/retail_db/orders/part-00000 | head -10

- ✓ **Last 10**

hadoop fs -cat practice/retail_db/orders/part-00000 | tail -10

# Knowing Statistics in HDFS

- ✓ Command – stat
  - Print statistics related to any file/directory
- ✓ default or %y - Modification Time
- ✓ %b - File Size in Bytes
- ✓ %F - Type of object.
- ✓ %o - Block Size
- ✓ %r - Replication
- ✓ %u - User Name
- ✓ %a - File Permission in Octal
- ✓ %A - File Permission in Symbolic

# Knowing Storage in HDFS

- ✓ **Command – df**
  - Shows the capacity, free and used space of the HDFS file system.
  - hadoop fs -df
  - -h →Readable Format

- ✓ **Command – du**
  - Show the amount of space, in bytes, used by the files that match the specified file pattern.
  - hadoop fs -du practice/retail_db
  - -h :Readable Format
  - -v : Displays with Header
  - -s : Summary of total size

# File Metadata

- ✓ Command – fsck
- ✓ Even if a Size of a file has less than 128MB , it will still occupy 1 Block.
- ✓ Help: hadoop fsck –help

### Print the fsck High Level Report
hadoop fsck practice/retail_db

### -files → Print a detailed file level report.
hadoop fsck practice/retail_db –files

### -files -blocks →Print a detailed file and block report.
hadoop fsck practice/retail_db –files -blocks

### -files -blocks -locations → Print out locations for every block
hadoop fsck practice/retail_db –files –blocks –locations

### -files -blocks -racks → Print out network topology for data-node locations

# HDFS File Permission

- HDFS File Permission is similar to Linux File Permission.

| Owner/user | Group | Others |
|---|---|---|
| rwx | rwx | rwx |

- For HDFS File,
  *r → Read Permission*
  *w → Write or Append*
  *x → No Meaning in HDFS*

```
easylearnspark1@training:~$ hadoop fs -ls practice/retail_db/order_items/part-00000
-rw-r--r--   1 easylearnspark1 supergroup    5408880 2021-12-25 22:11 practice/retail_db/order_items/part-00000
```

User Group Others

- For HDFS Directory,
  r → Able to List content
  w → Able to Create or Delete
  x → Able to access a child

Directory

```
easylearnspark1@training:~$ hadoop fs -ls practice/retail_db/
Found 6 items
drwxr-xr-x   - easylearnspark1 supergroup          0 2021-12-25 22:11 practice/retail_db/categories
drwxr-xr-x   - easylearnspark1 supergroup          0 2021-12-25 22:11 practice/retail_db/customers
drwxr-xr-x   - easylearnspark1 supergroup          0 2021-12-25 22:11 practice/retail_db/departments
drwxr-xr-x   - easylearnspark1 supergroup          0 2021-12-25 22:11 practice/retail_db/order_items
```

# HDFS File Permission

- Change Permission using –chmod
- Octal Mode: –chmod 755

hadoop fs -chmod 755 practice/retail_db/orders/part-00000

| Numeric Value | Permission |
|---|---|
| 0 | --- |
| 1 | --x |
| 2 | -w- |
| 3 | -wx |
| 4 | r-- |
| 5 | r-x |
| 6 | rw- |
| 7 | rwx |

- Symbolic Mode: -chmod g+x

hadoop fs -chmod g+w practice/retail_db/orders/part-00000

```
u   -   user
g   -   group
o   -   other
```

# HDFS Override Properties

- ✓ 1. Change Properties in hdfs-site.xml or core-site.xml
- ✓ 2. Override the default properties while Copying the Files into HDFS (-D or –conf)
- ✓ 3. Change after copying the Files in HDFS (-setRep for changing replication.)