Big-data Storage

HDFS

HDFS Overview

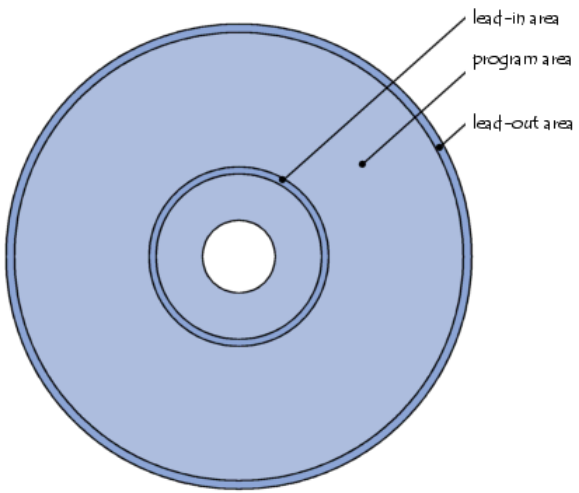
- A distributed file system
- Built on the architecture of Google File System (GFS)
- Shares a similar architecture to many other common distributed storage engines such as Amazon S3 and Microsoft Azure
- HDFS is a stand-alone storage engine and can be used in isolation of the query processing engine
- Even if you do not use Hadoop MapReduce, you will probably still use HDFS

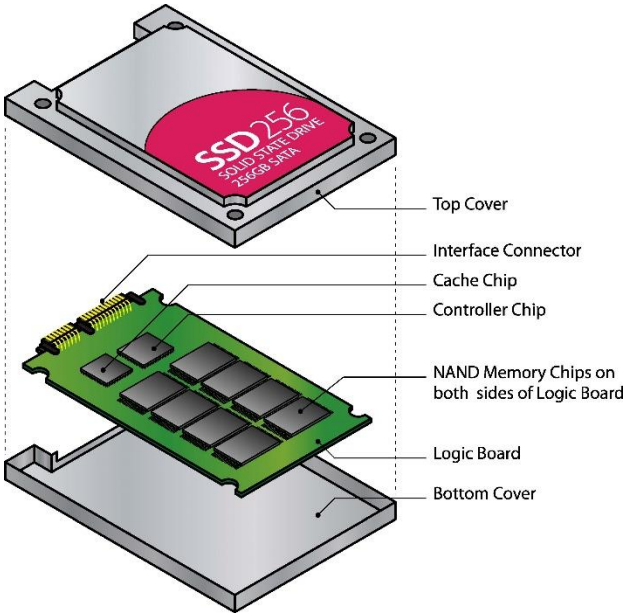
HDFS Topics

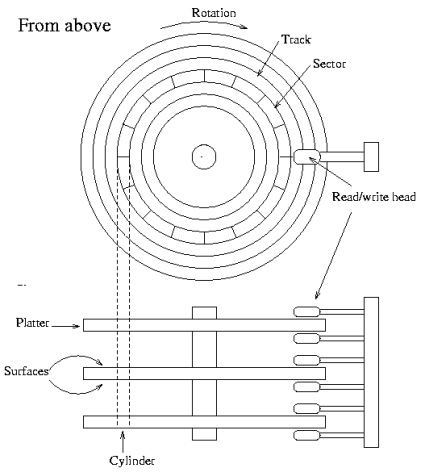
- HDFS design and architecture
- Fault tolerance in HDFS
- Command-line interface (HDFS Shell)
- Java API
- Create (Write) a file
- Stream reading
- Structured reading

Background on Disk Storage

- What are file systems and why do we need them?
- A file is a logical sequence of bits/bytes
- A physical disk stores data in sectors, tracks, tapes, blocks, … etc.







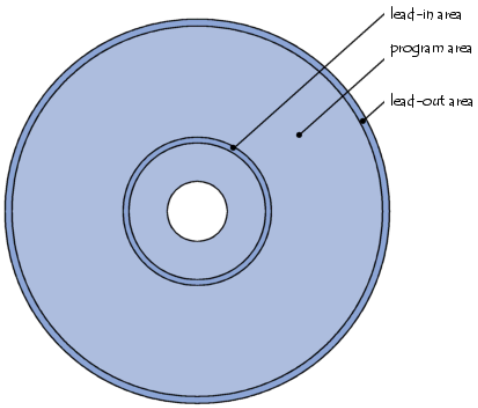
File System

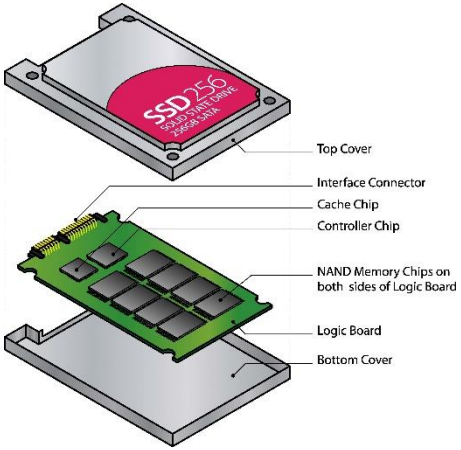
- Any file system, is a method to provide a high-level abstraction on physical disk to make it easier to store files

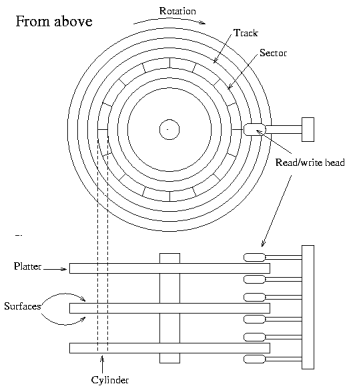
Files

Folders

File System





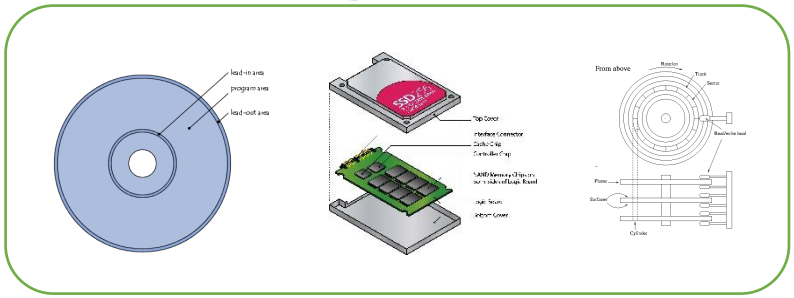


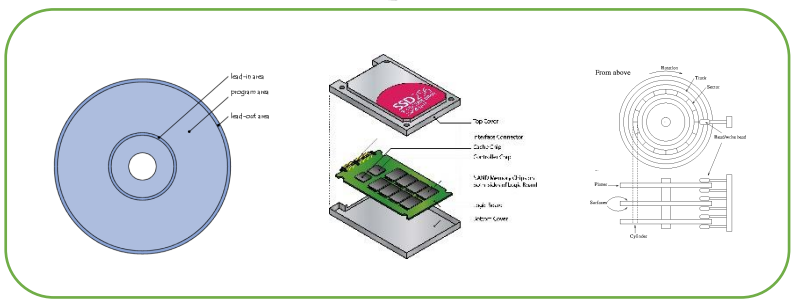
Distributed File System

Files

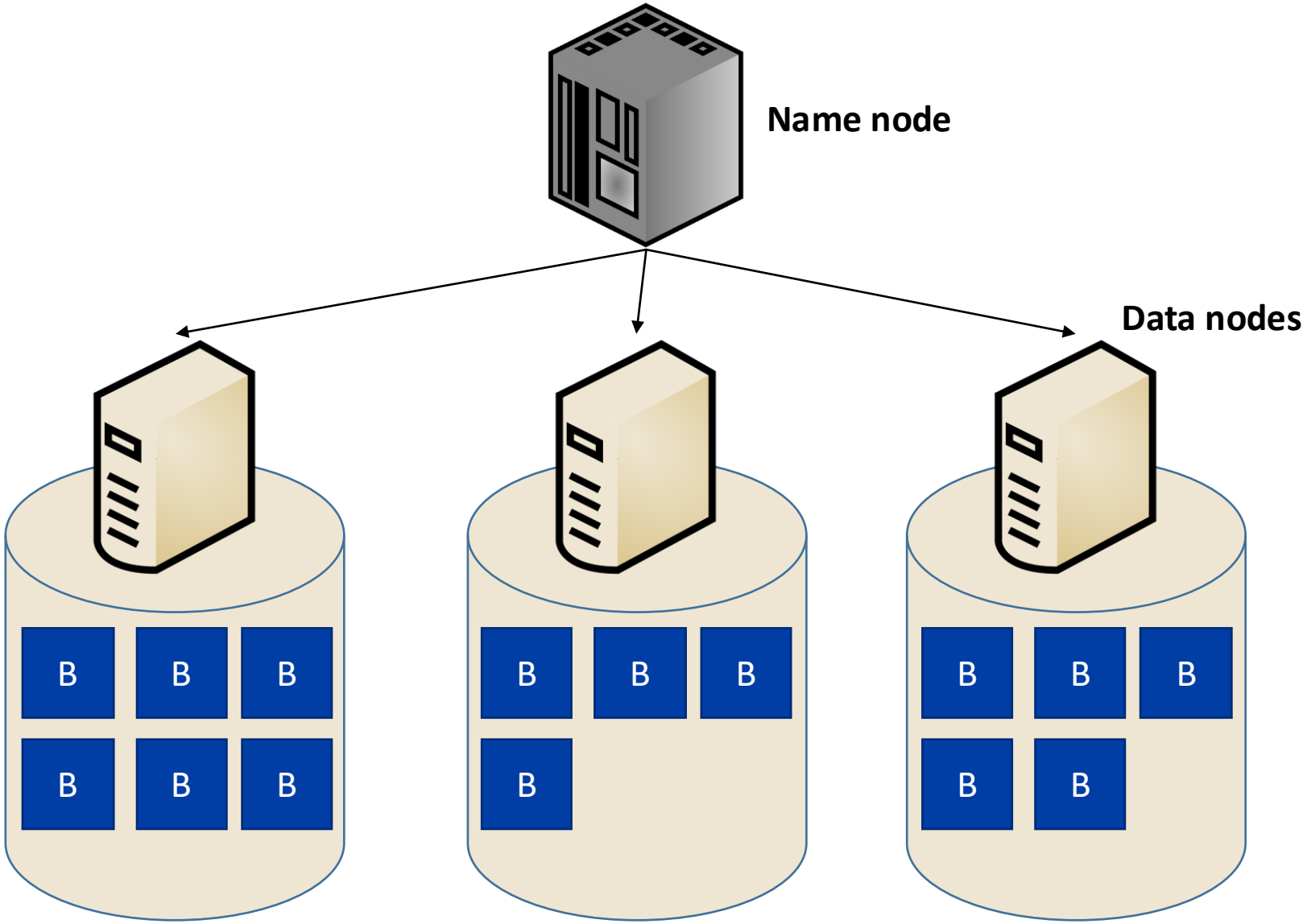
Folders

Distributed File System

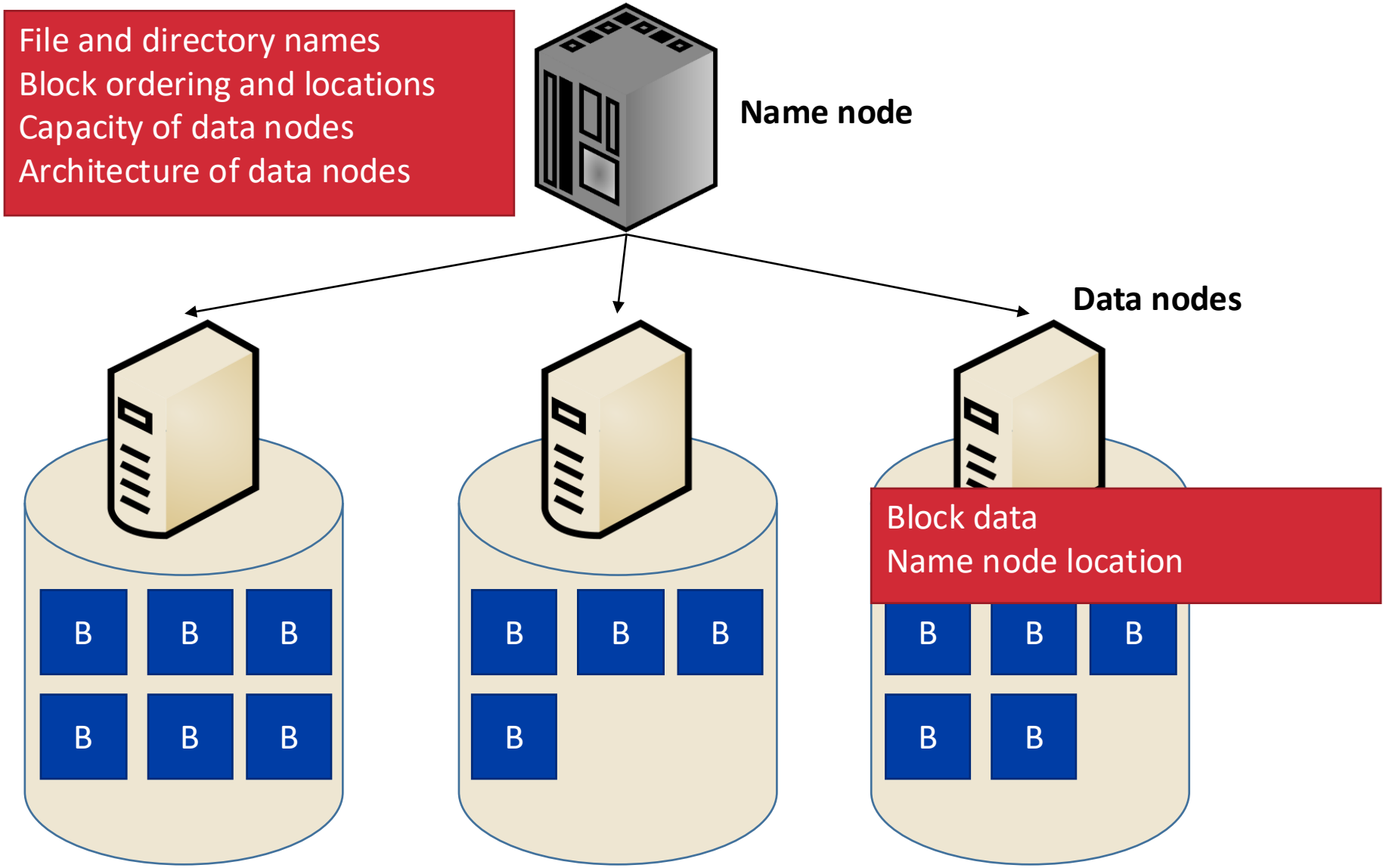




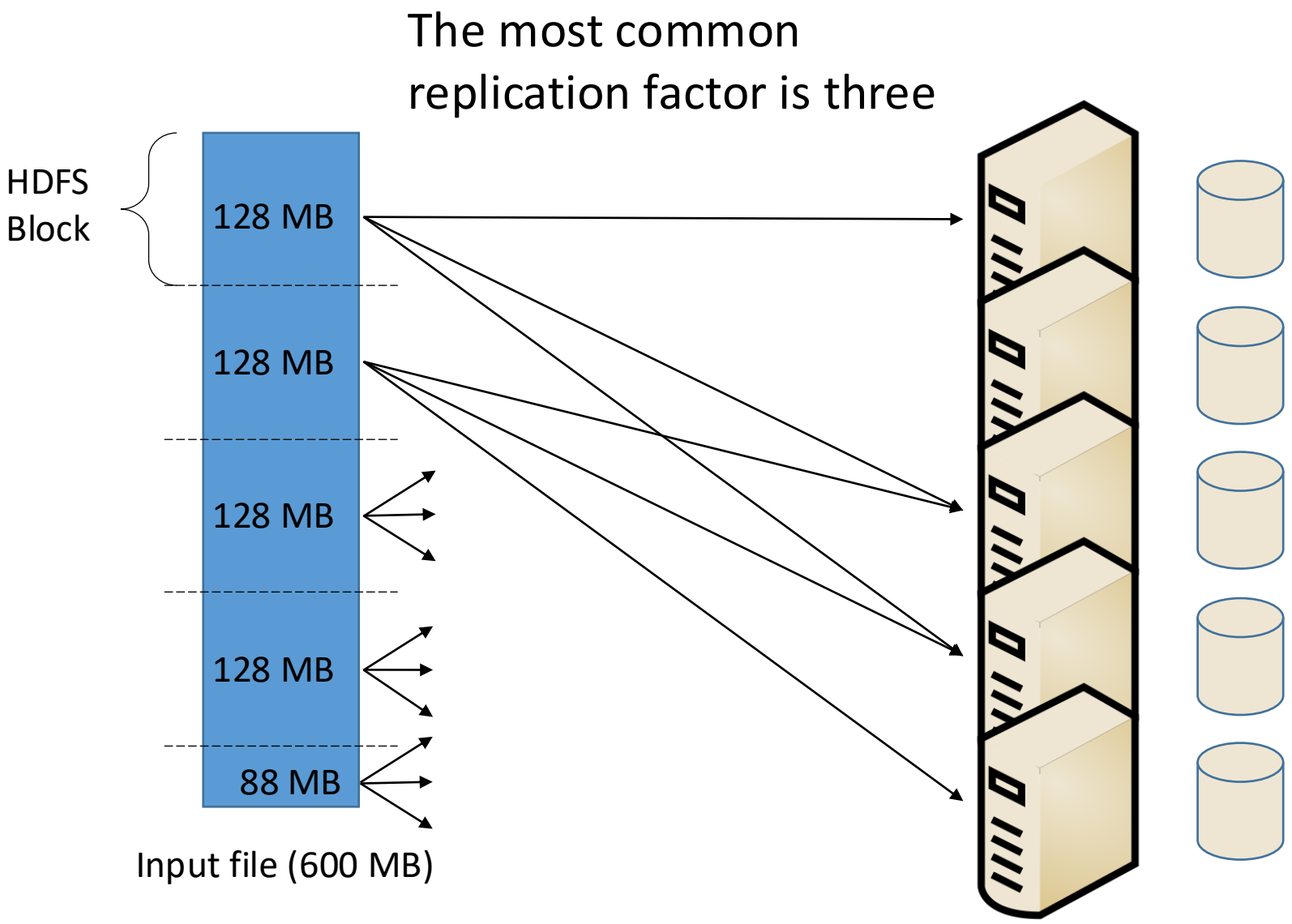
HDFS Architecture



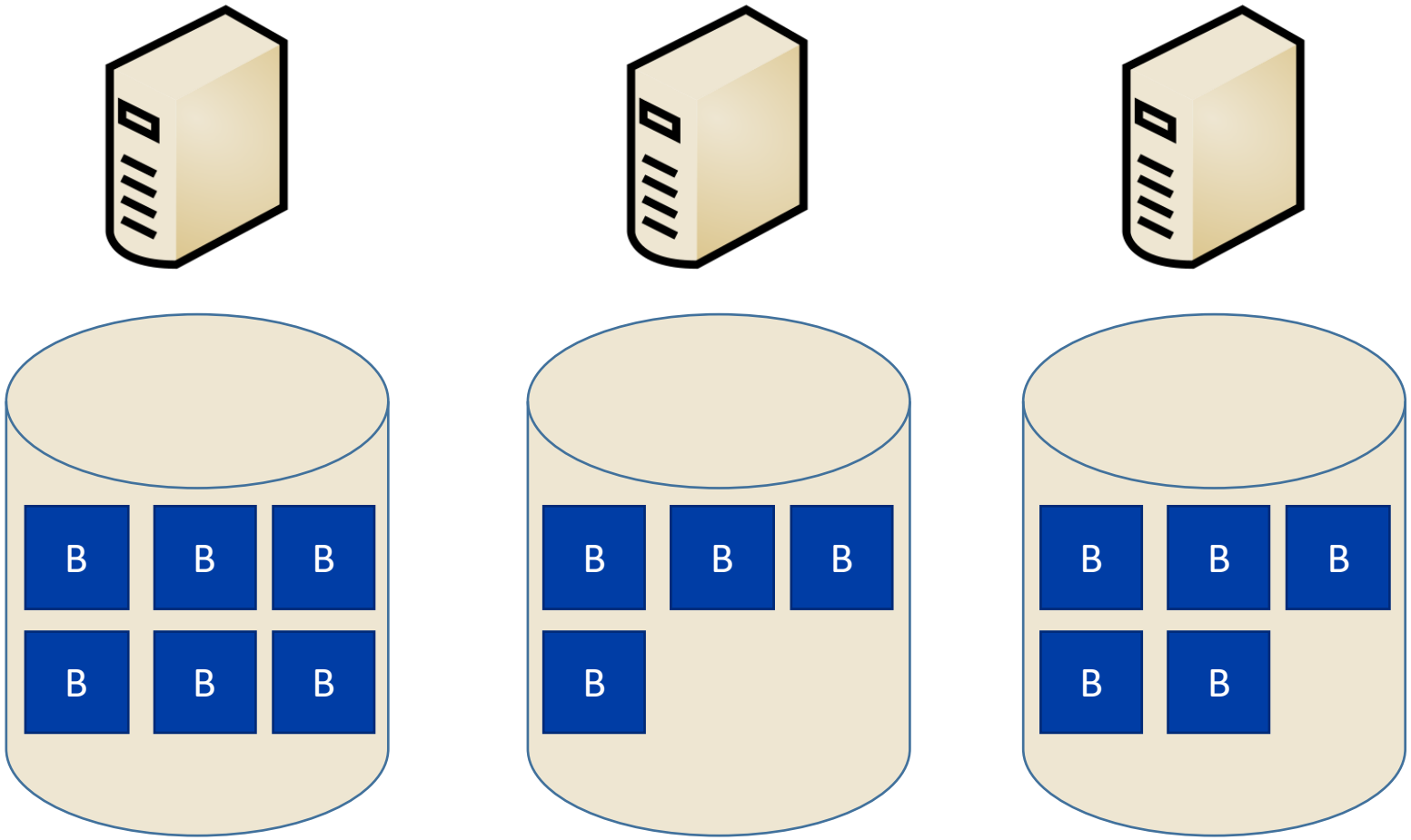
What is where?



Data Loading

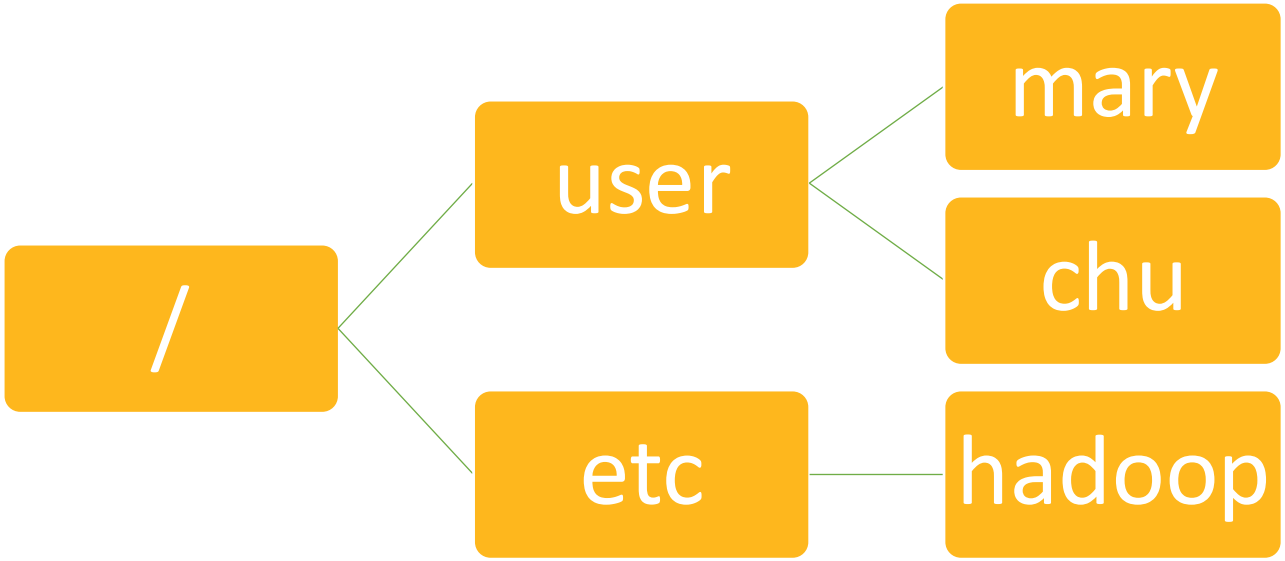


HDFS Storage



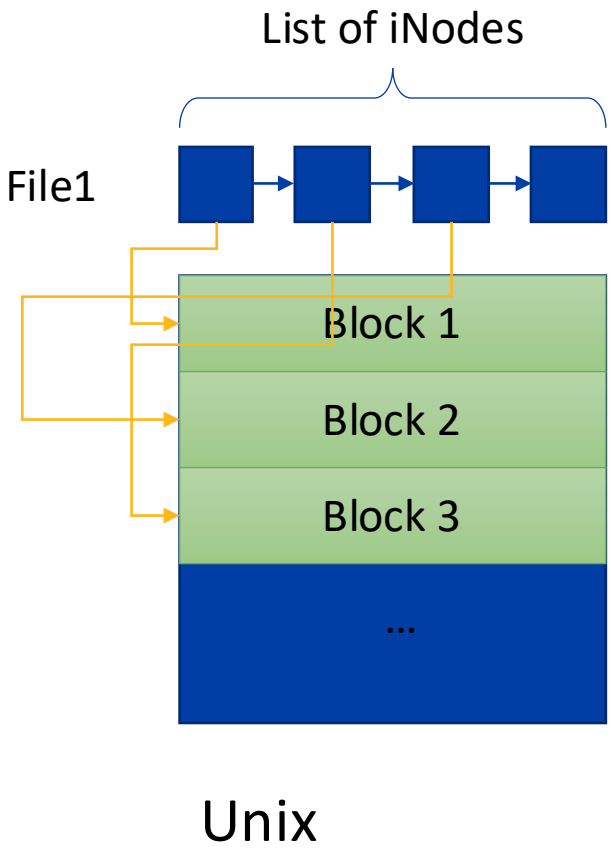
Analogy to Unix FS

The logical view is similar

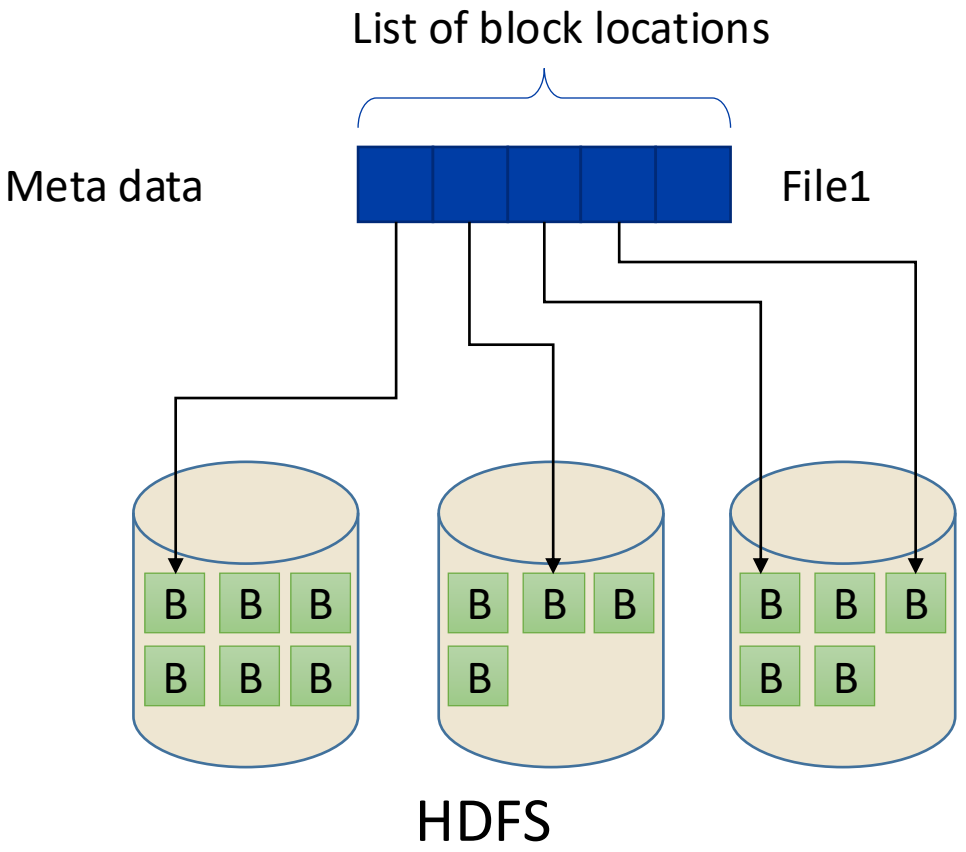


Analogy to Unix FS

The physical model is comparable



Unix



HDFS

Fault Tolerance in HDFS

Replication

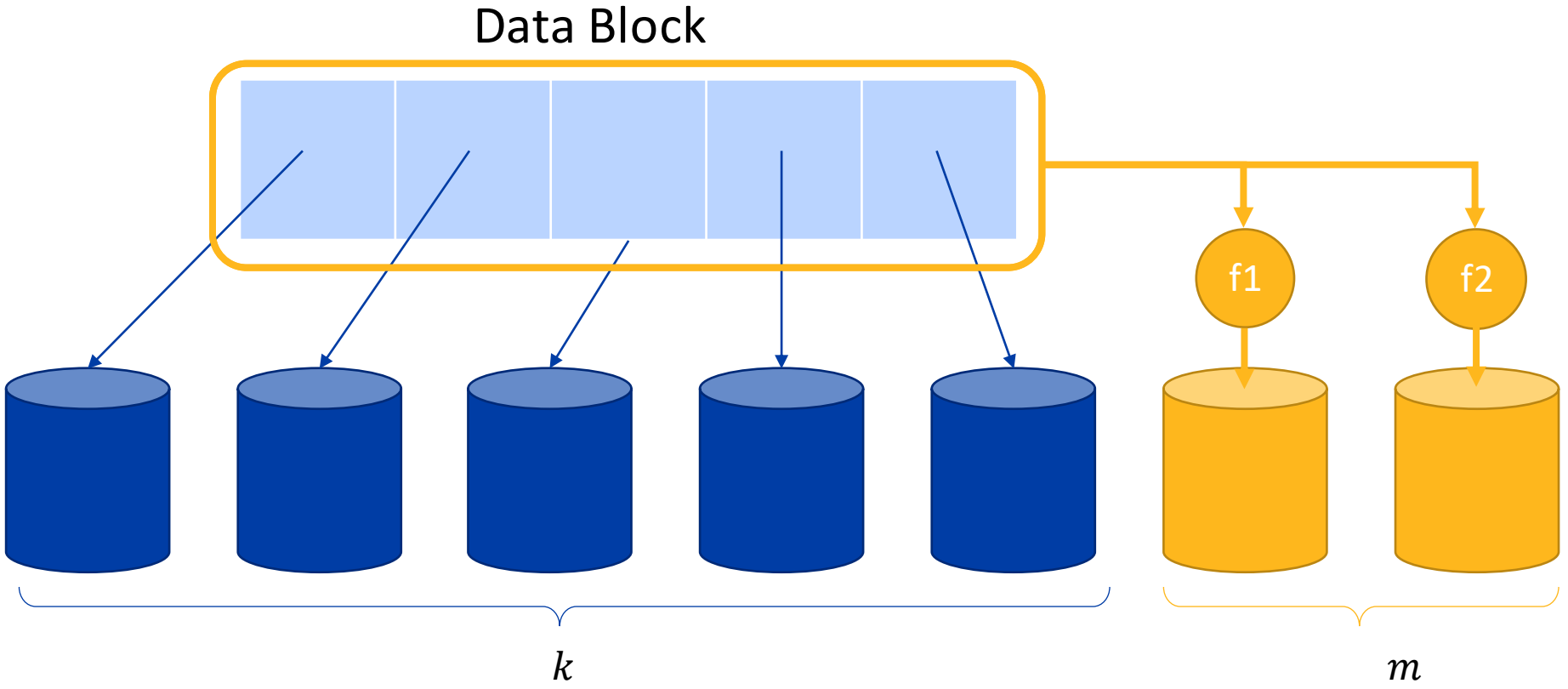
- The default fault tolerance mechanism in HDFS is replication
- The most common replication factor is three
- If one or two nodes are temporarily unavailable, the data is still accessible
- If one or two nodes permanently fail, the master node replicates the under-replicated blocks to reach the desired replication factor
- Drawback: reduced disk capacity

Replication Overhead

- Example: A cluster with 12 nodes, each node has a 10TB disk, what is the actual HDFS capacity?
- Total physical disk available = $12 \times 10 = 120TB$
- Actual perceived capacity of HDFS = $\frac{120}{3} = 40$ TB
- The overhead is 200%

Advanced Feature: Erasure Coding

- Uses advanced algorithms for recovery, e.g., Reed-Solomon, XOR



How Redundancy Works

- Consider that we want to store the two values x and y
- Instead of repeating the raw values, we can store the following four values:
  - $x, y, x + y, x - y$
- If we lose one or two values, we can still recover the other two by solving an array of linear equations (two in this case)
- This idea can be generalized with more efficient formulas that lowers the computation

Storage Overhead with Erasure Coding

- HDFS Erasure Coding: (6, 3) RS Code
- $k = 6$ data blocks
- $m = 3$ parity blocks
- Storage overhead = $\frac{6+3}{6} - 1 = 0.5$

HDFS Shell

Command-line Interface (CLI)

HDFS Shell

- Used for common operations
- Its usage is similar to Unix shell commands
- Basic operations include
  - ls, cp, mv, mkdir, rm, …
- HDFS-specific operations include
  - copyToLocal, copyFromLocal, setrep, appendToFile, …

HDFS Shell

- General format

```
hdfs dfs -<cmd> <arguments>
```

- So, instead of

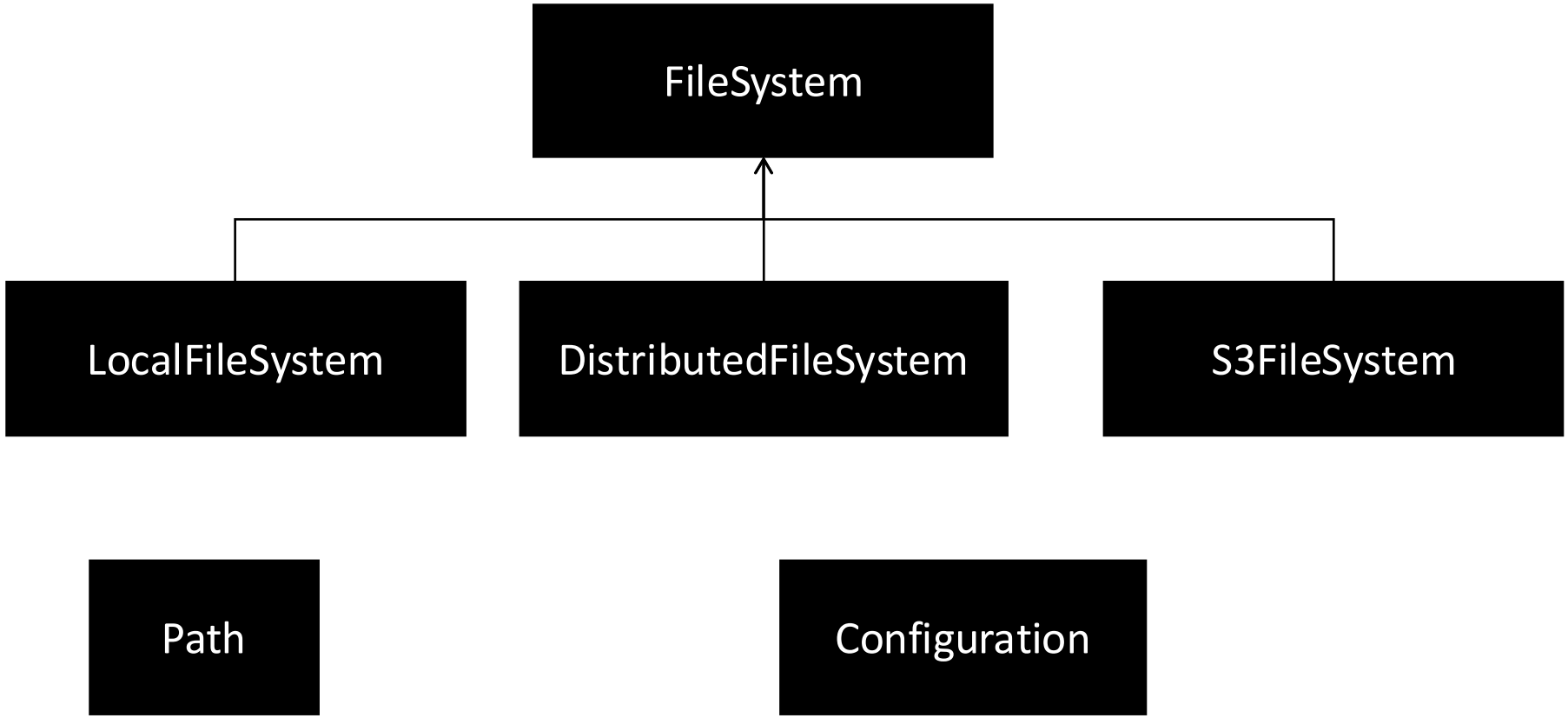
```
mkdir –p myproject/mydir
```

- You will write

```
hdfs dfs -mkdir –p myproject/mydir
```

- A list of shell commands with usage
  - https://hadoop.apache.org/docs/r3.3.0/hadoop-project-dist/hadoop-common/FileSystemShell.html

HDFS API


FileSystem
LocalFileSystem
DistributedFileSystem
S3FileSystem
Path
Configuration

HDFS API Classes

- Configuration: Holds system configuration such as where the master node is running and default system parameters, e.g., replication factor and block size
- Path: Stores a path to a file or directory
- FileSystem: An abstract class for file system operations

Fully Qualified Path

hdfs://namenode:9000/path/to/file

hdfs: the file system scheme. Other possible values are file, ftp, s3, …

namenode: the name or IP address of the node that hosts the the file system

9000: the port on which the master node is listening

/path/to/file: the absolute path of the file

Shorter Path Forms

- “file”: relative path to the current working directory in the default file system
- “/path/to/file”: Absolute path to a file in the default* file system (as configured)
- “hdfs:///path/to/file”: Use the default* values for the master node and port
- “hdfs://masternode/path/tofile”: Use the given masternode name or IP and the default* port
- “file:///home/user/file”: an absolute path in the local file system

*All the defaults are in the Configuration object

HDFS API (Java)

Create the file system

```
Configuration conf = new Configuration();
Path path = new Path("…");
FileSystem fs = path.getFileSystem(conf);

// To get the local FS
fs = FileSystem.getLocal (conf);

// To get the default FS
fs = FileSystem.get(conf);
```

HDFS API

Create a new file

```
FSDataOutputStream out =
fs.create(path, …);
```

Delete a file

```
fs.delete(path, recursive);
fs.deleteOnExit(path);
```

Rename/Move a file

```
fs.rename(oldPath, newPath);
```

HDFS API

Open a file

```
FSDataInputStream in = fs.open(path, …);
```

Seek to a different location

```
in.seek(pos);
in.seekToNewSource(pos);
```

HDFS API

Concatenate

```
fs.concat(destination, src[]);
```

Get file metadata

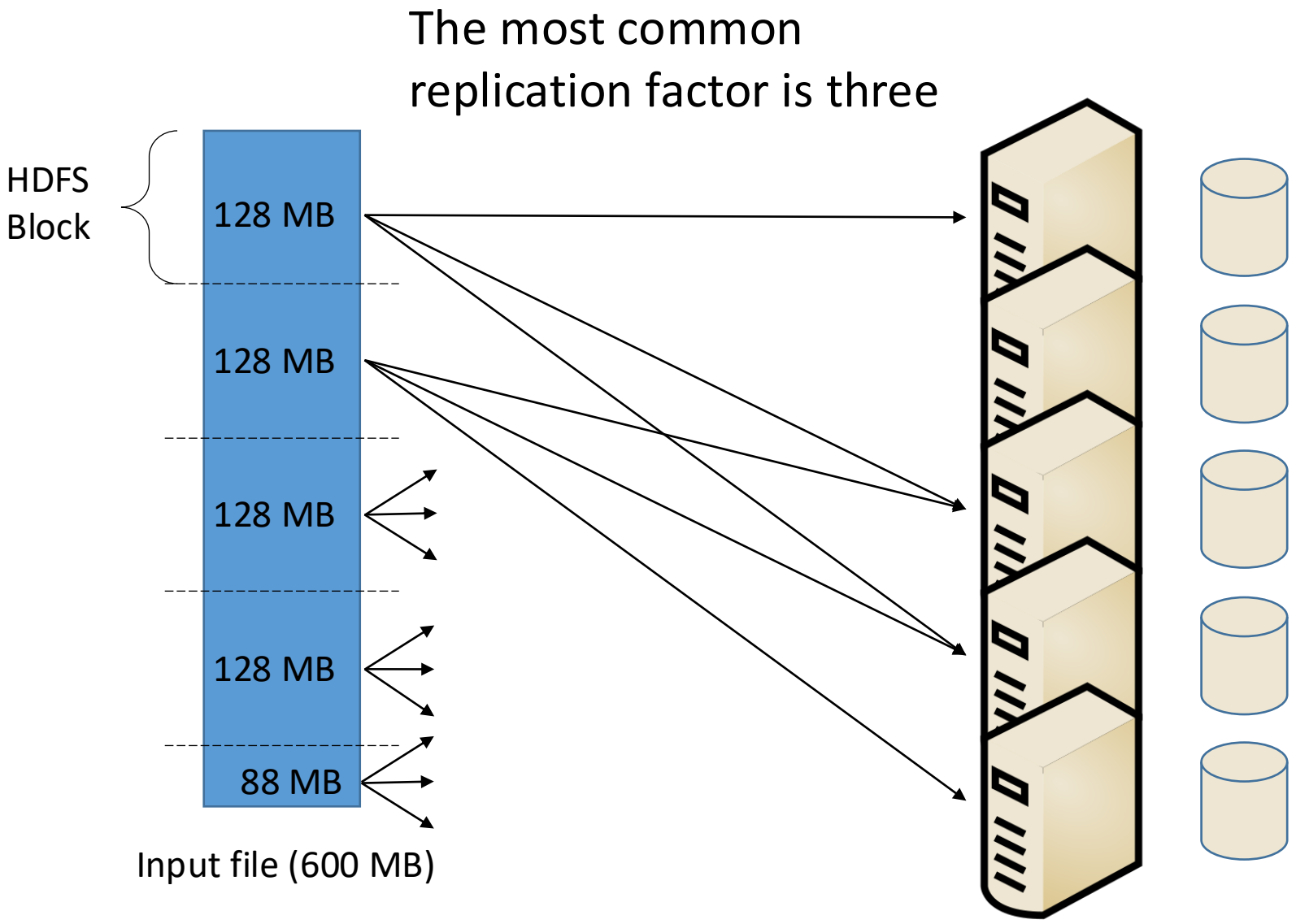
```
fs.getFileStatus(path);
```

Get block locations

```
fs.getFileBlockLocations(path, from, to);
```

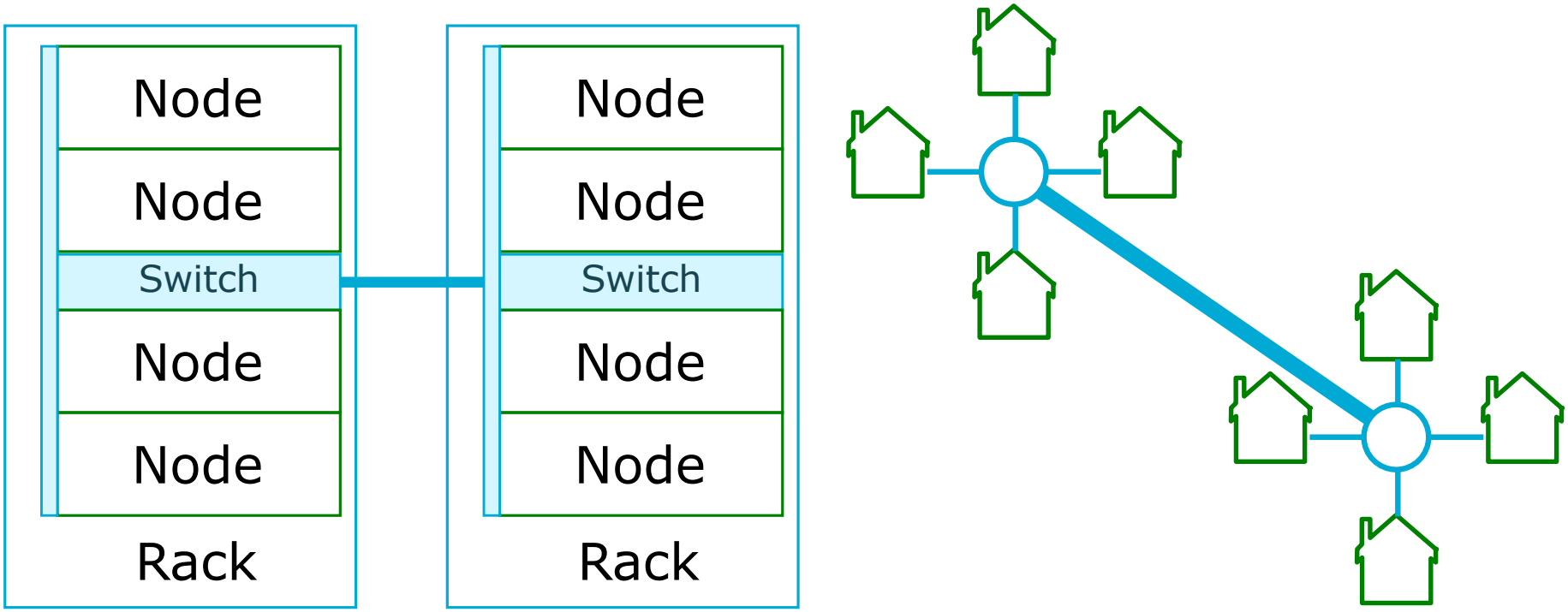
Internals of HDFS Writing

Data Loading



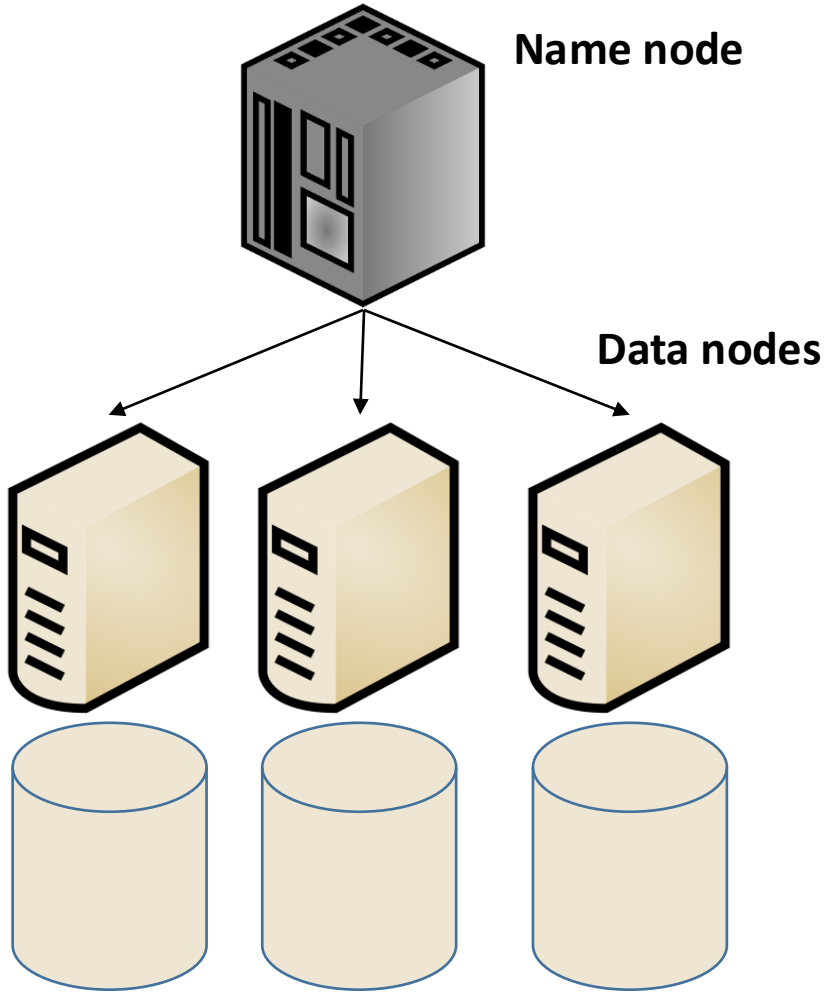
Physical Cluster Layout

Analogy of racks

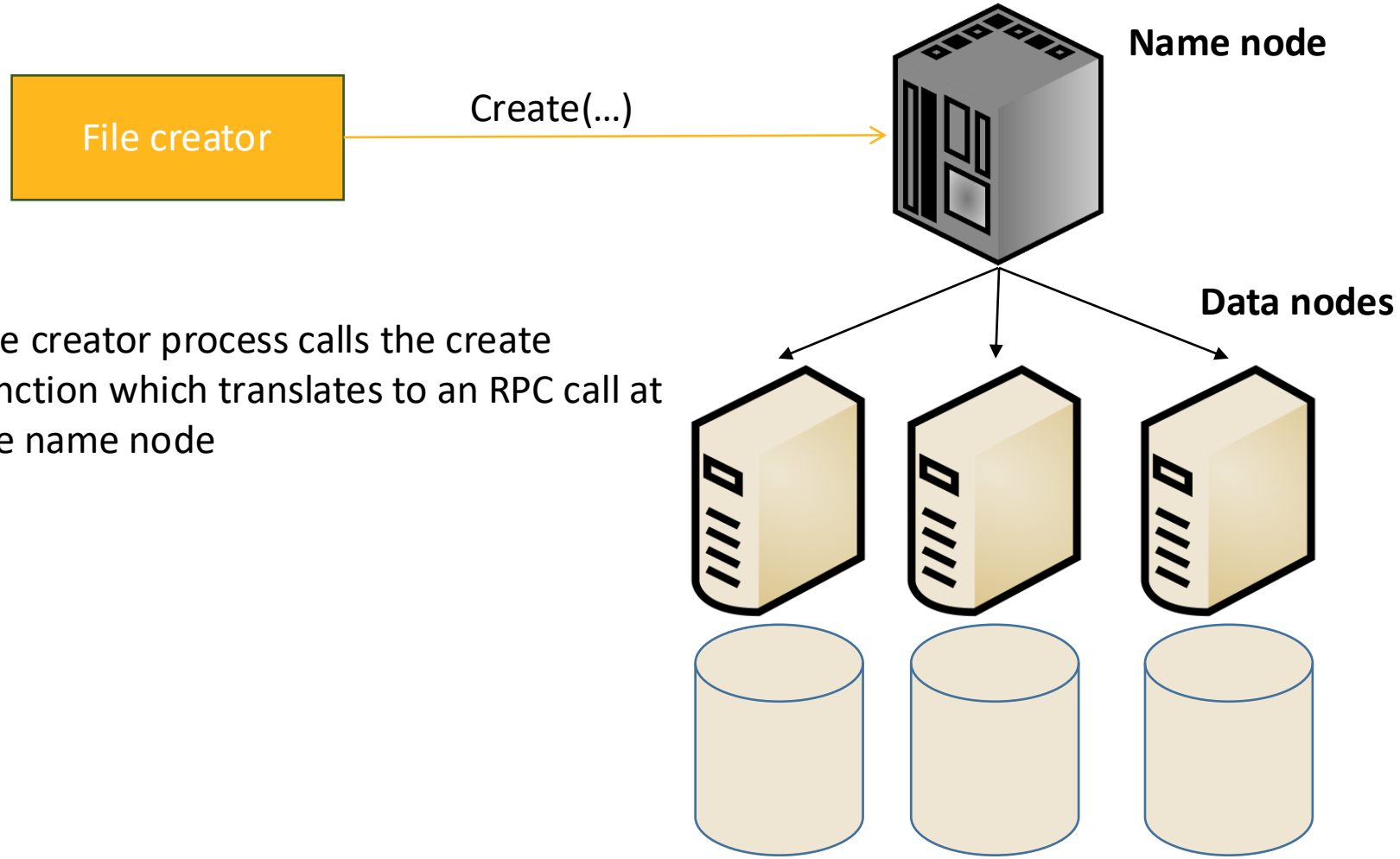


HDFS Writing Process

File creator

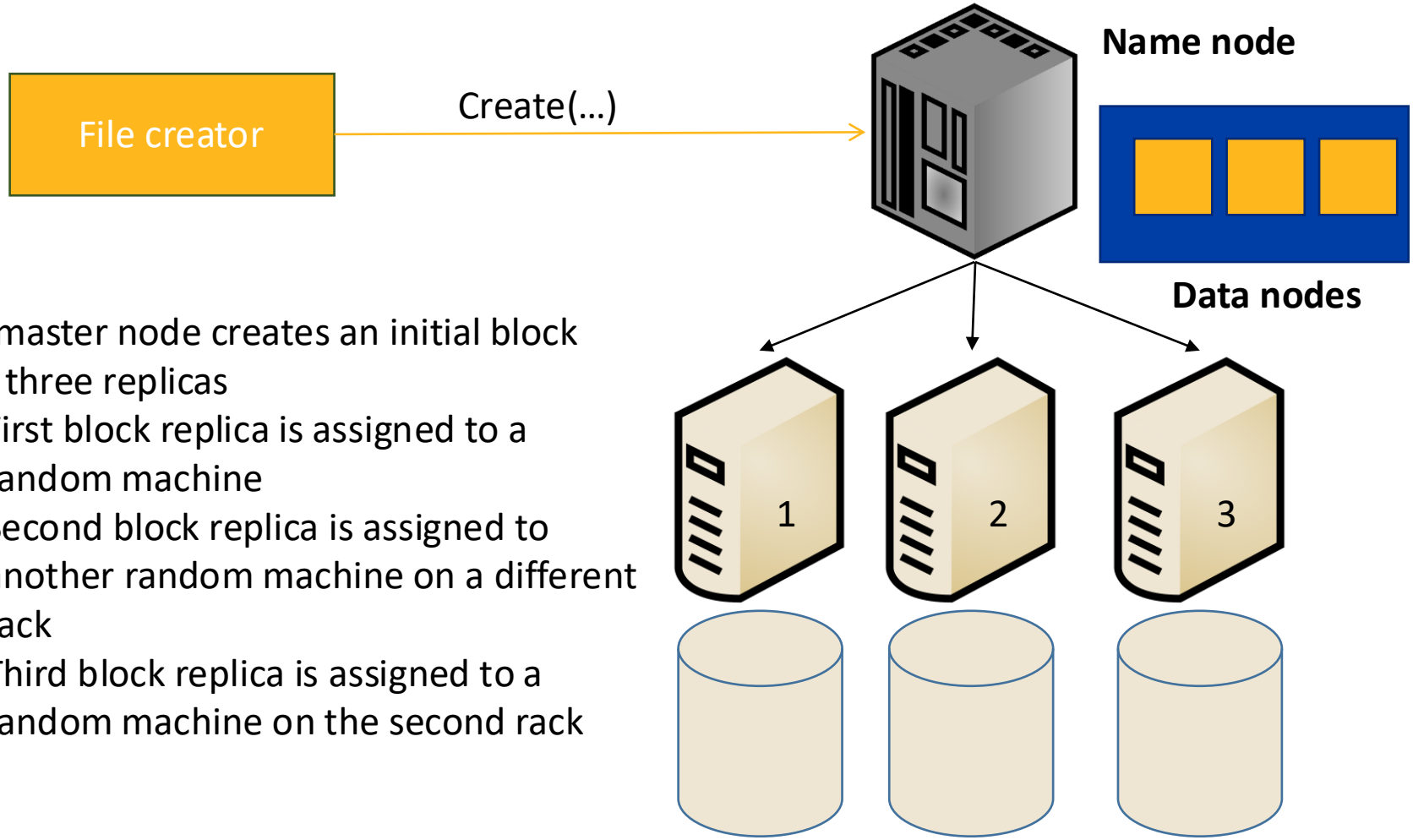


HDFS Writing Process



The creator process calls the create function which translates to an RPC call at the name node

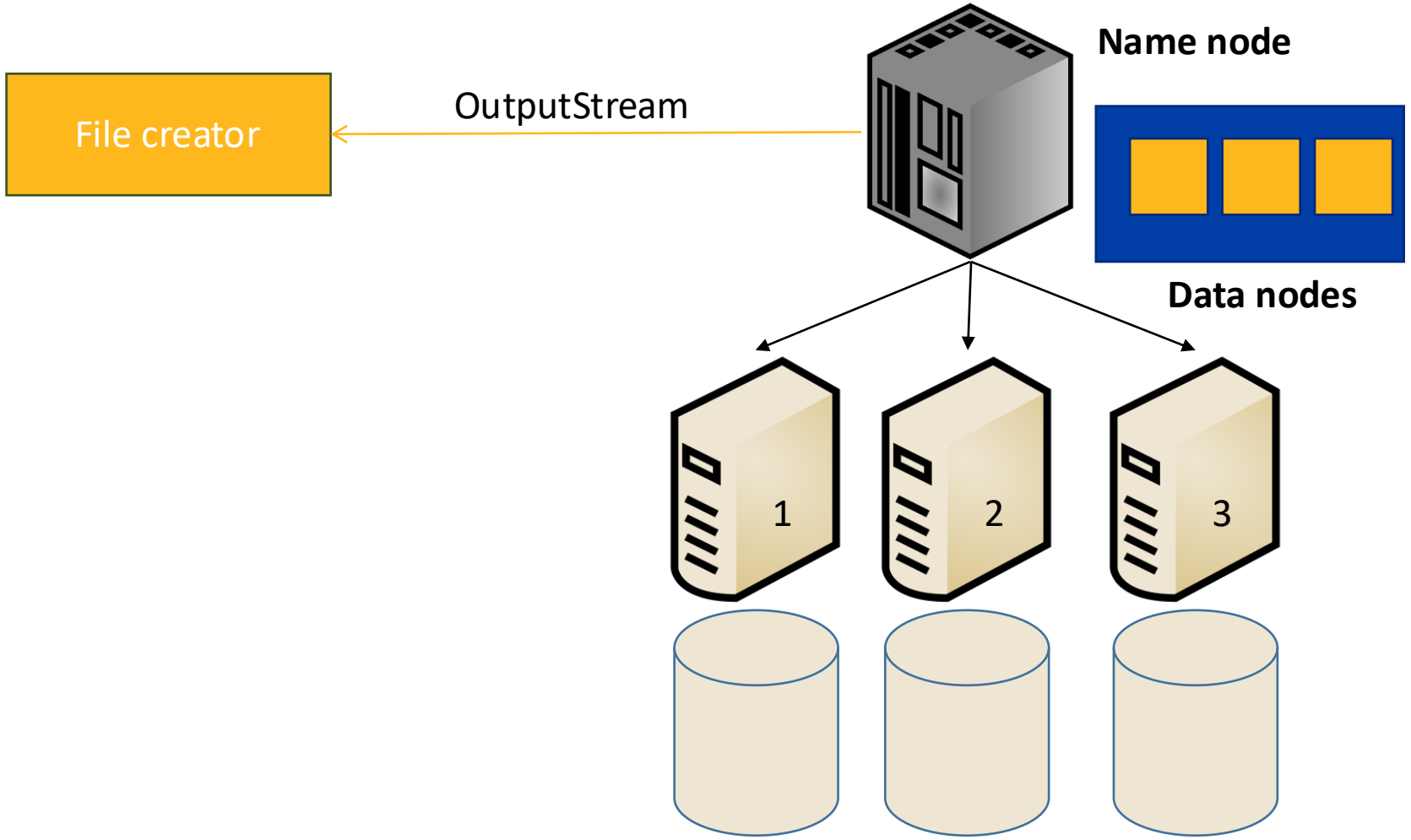
HDFS Writing Process



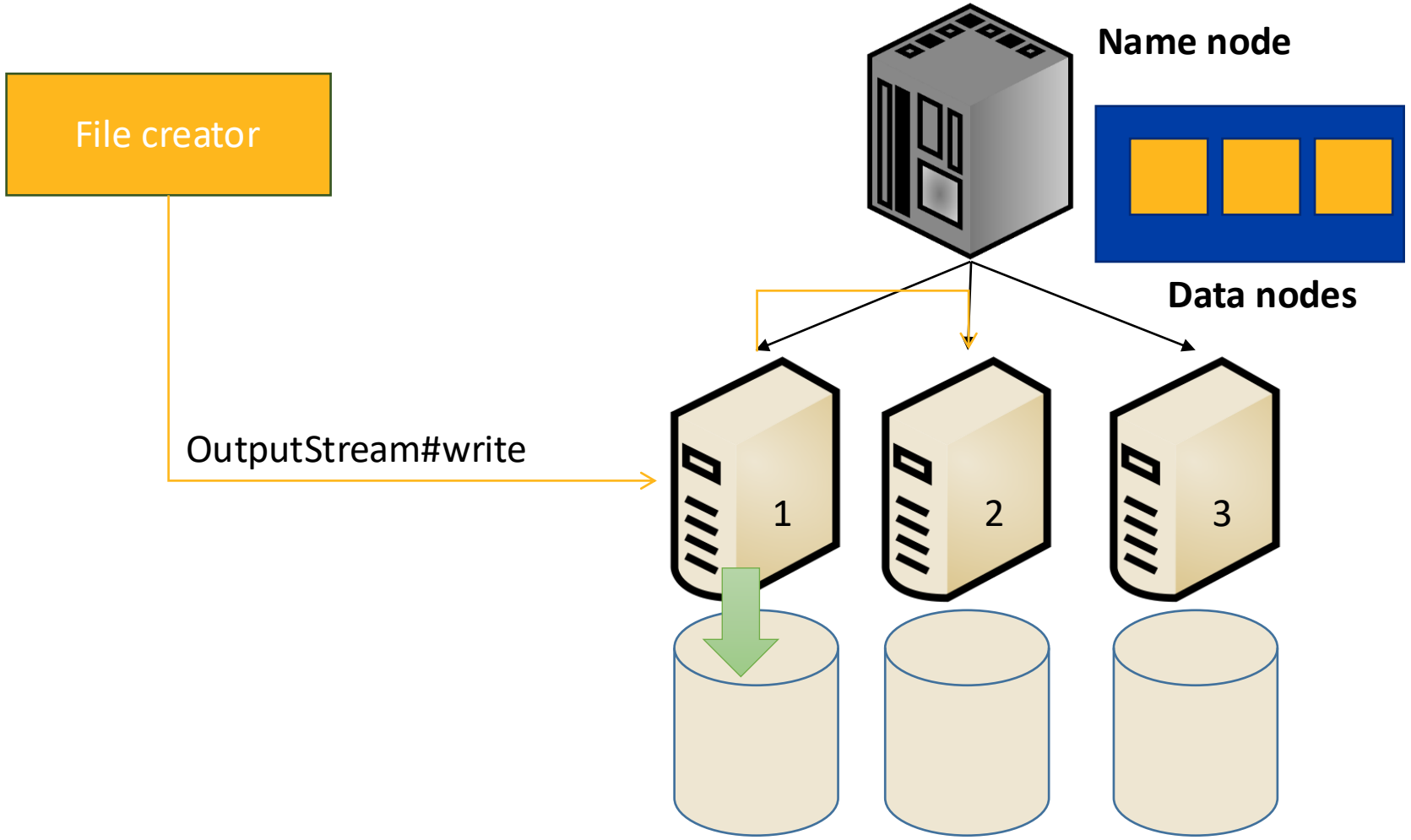
The master node creates an initial block with three replicas

1. First block replica is assigned to a random machine
2. Second block replica is assigned to another random machine on a different rack
3. Third block replica is assigned to a random machine on the second rack

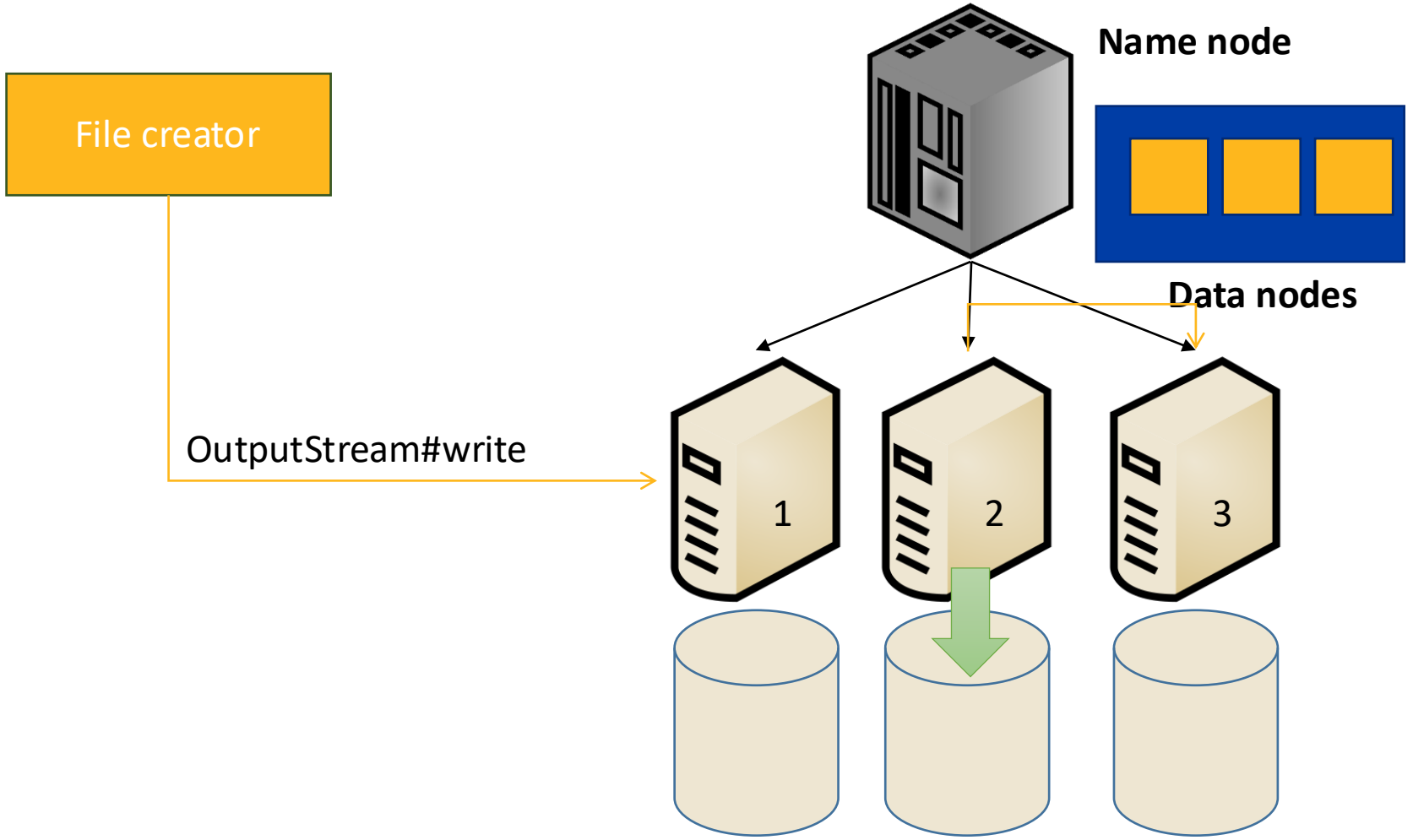
HDFS Writing Process



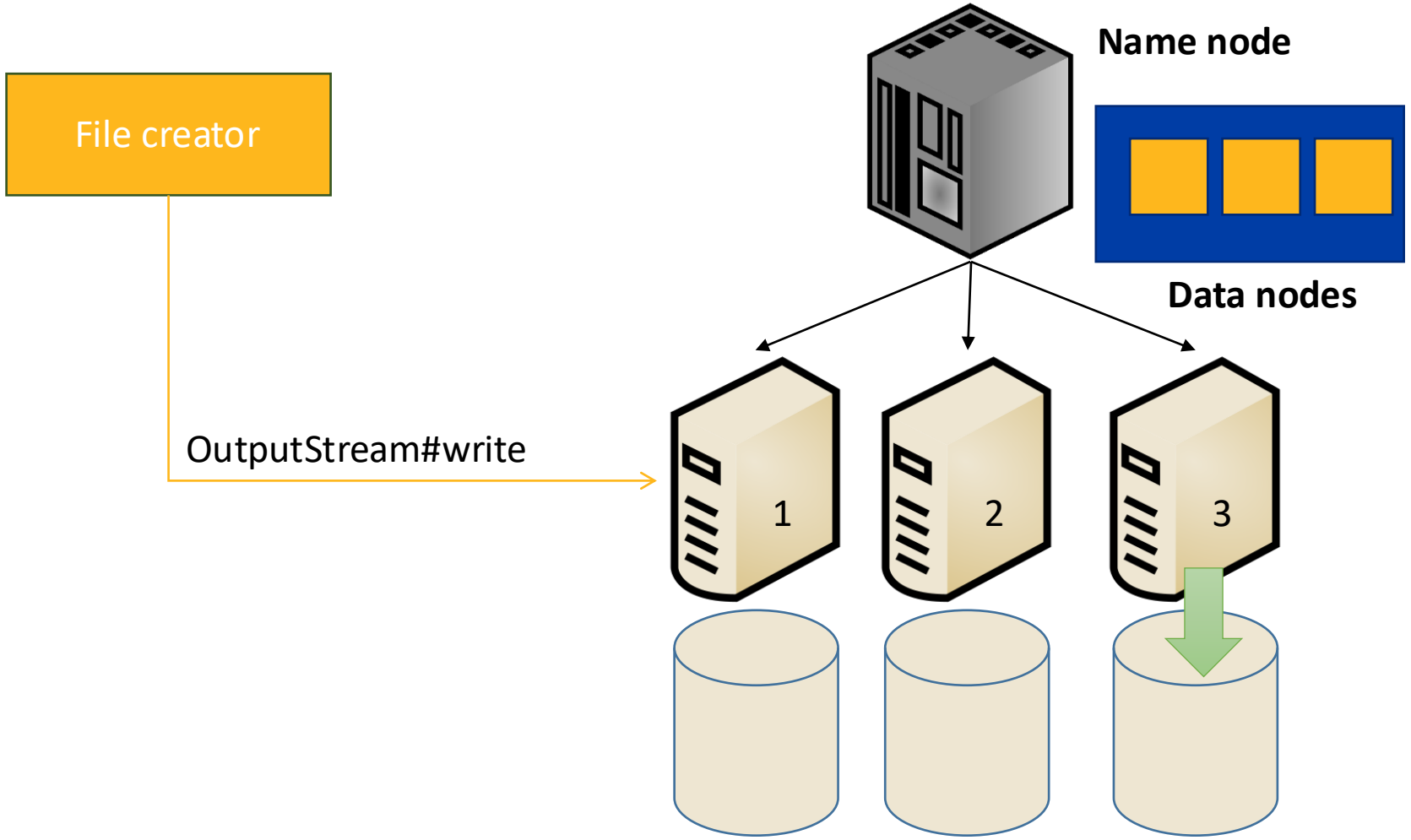
HDFS Writing Process



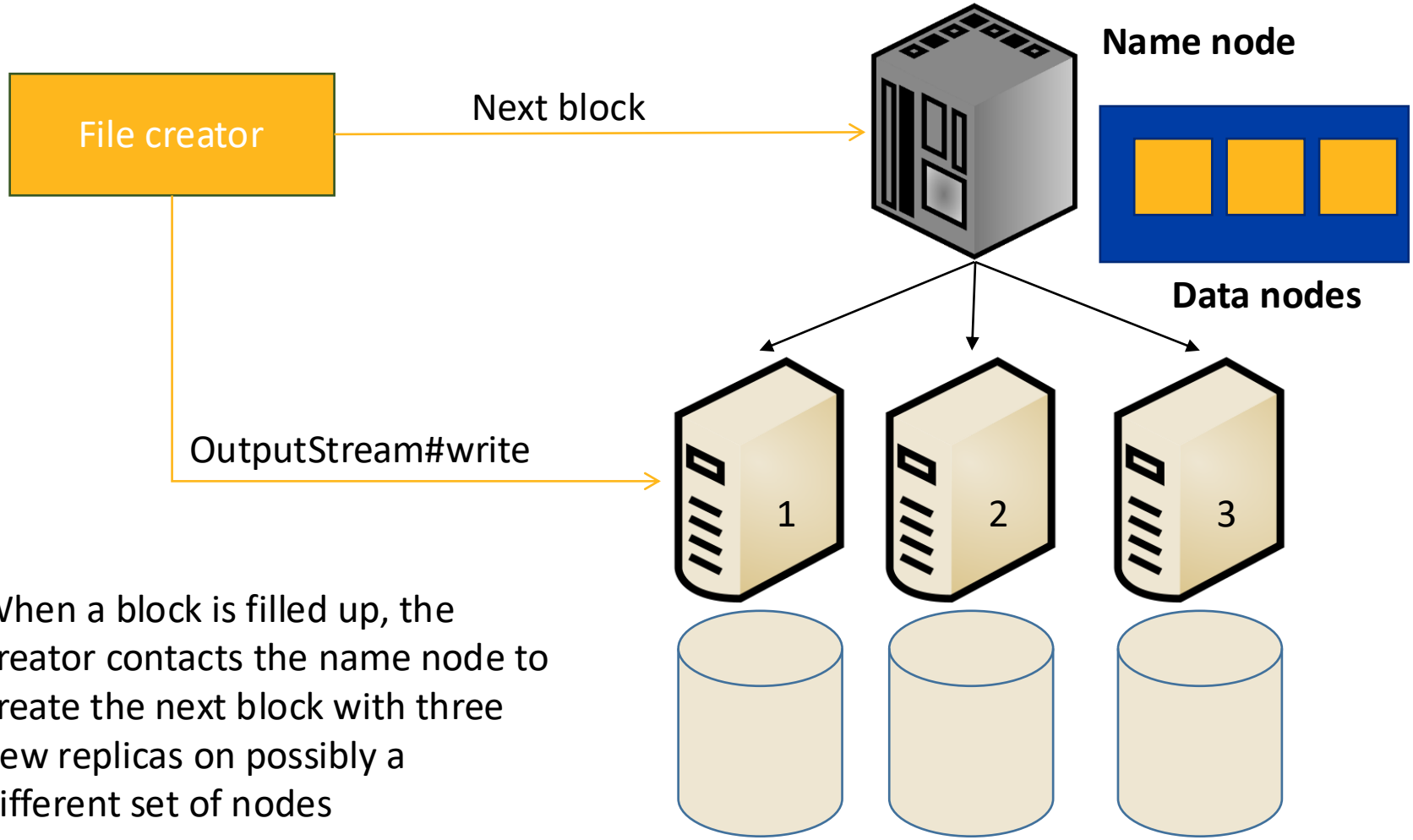
HDFS Writing Process



HDFS Writing Process



HDFS Writing Process



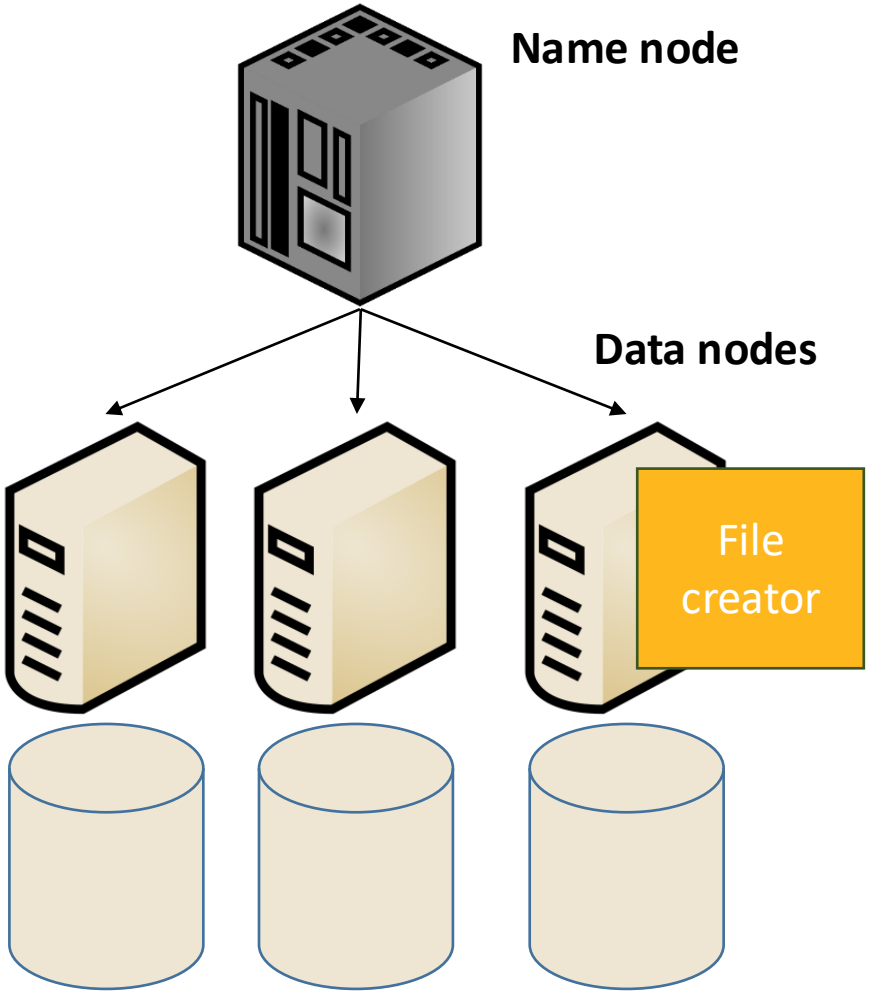
When a block is filled up, the creator contacts the name node to create the next block with three new replicas on possibly a different set of nodes

Notes about writing to HDFS

- Data transfers of replicas are pipelined
- The data does **not** go through the name node
- Random writing is **not** supported
- Appending to a file is supported but it creates a new block

Writing from a datanode

If the file creator is running on one of the data nodes, the first replica is always assigned to that node
The second and third replicas are assigned as before



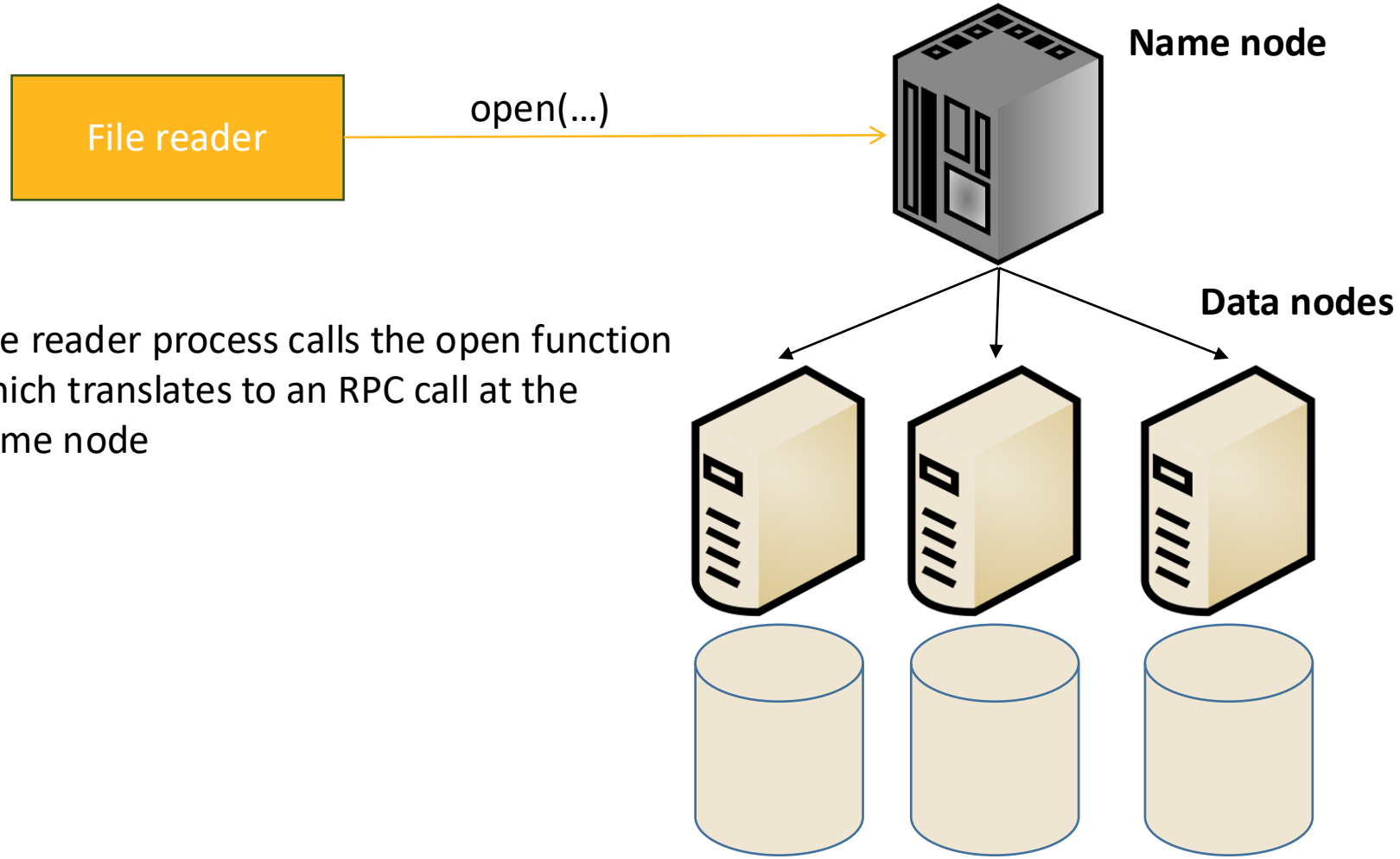
Internals of HDFS

Stream reading

Reading from HDFS

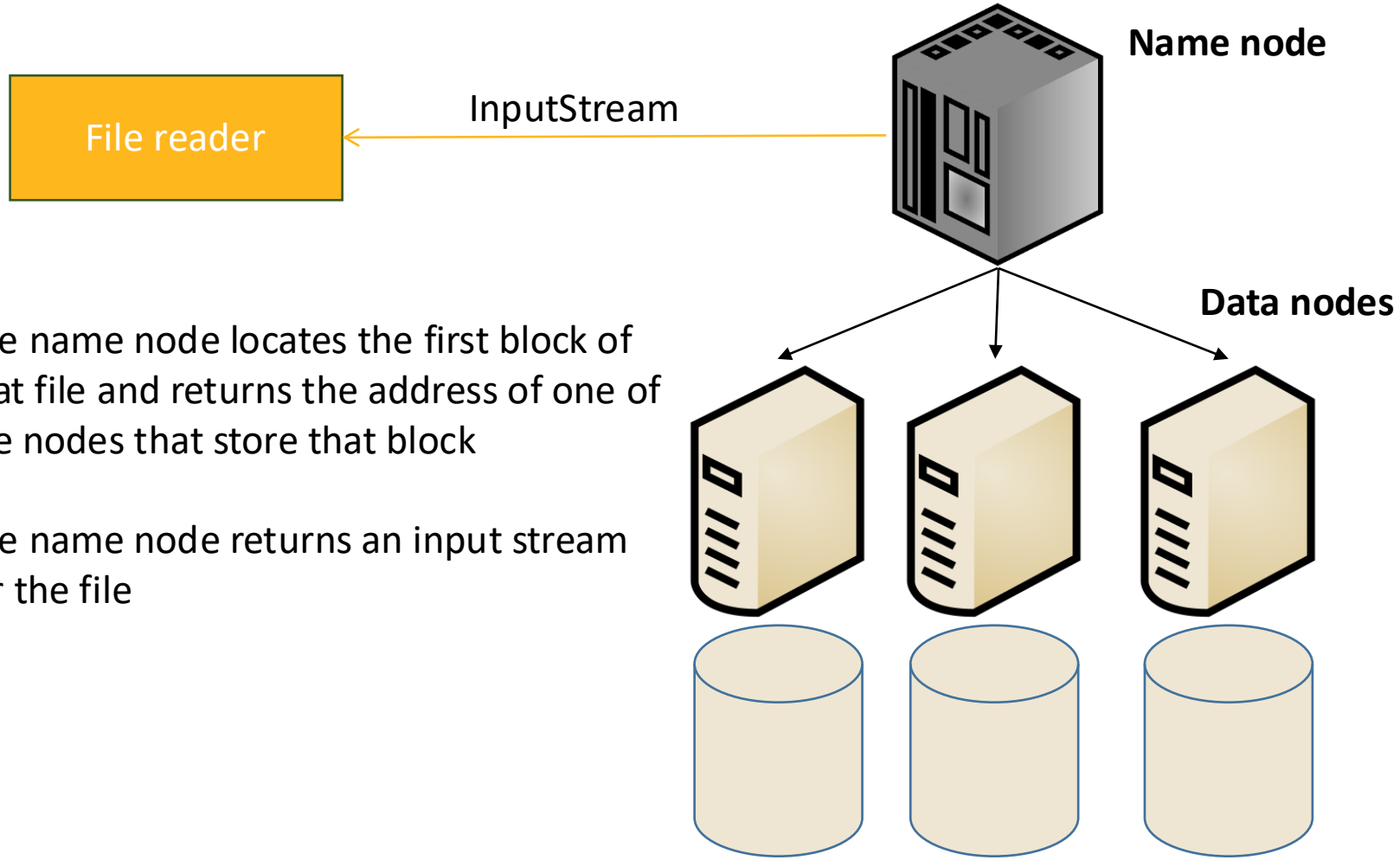
- Reading is relatively easier
- No replication is needed
- Replication can be exploited
- Random reading *is* allowed

HDFS Reading Process



The reader process calls the open function which translates to an RPC call at the name node

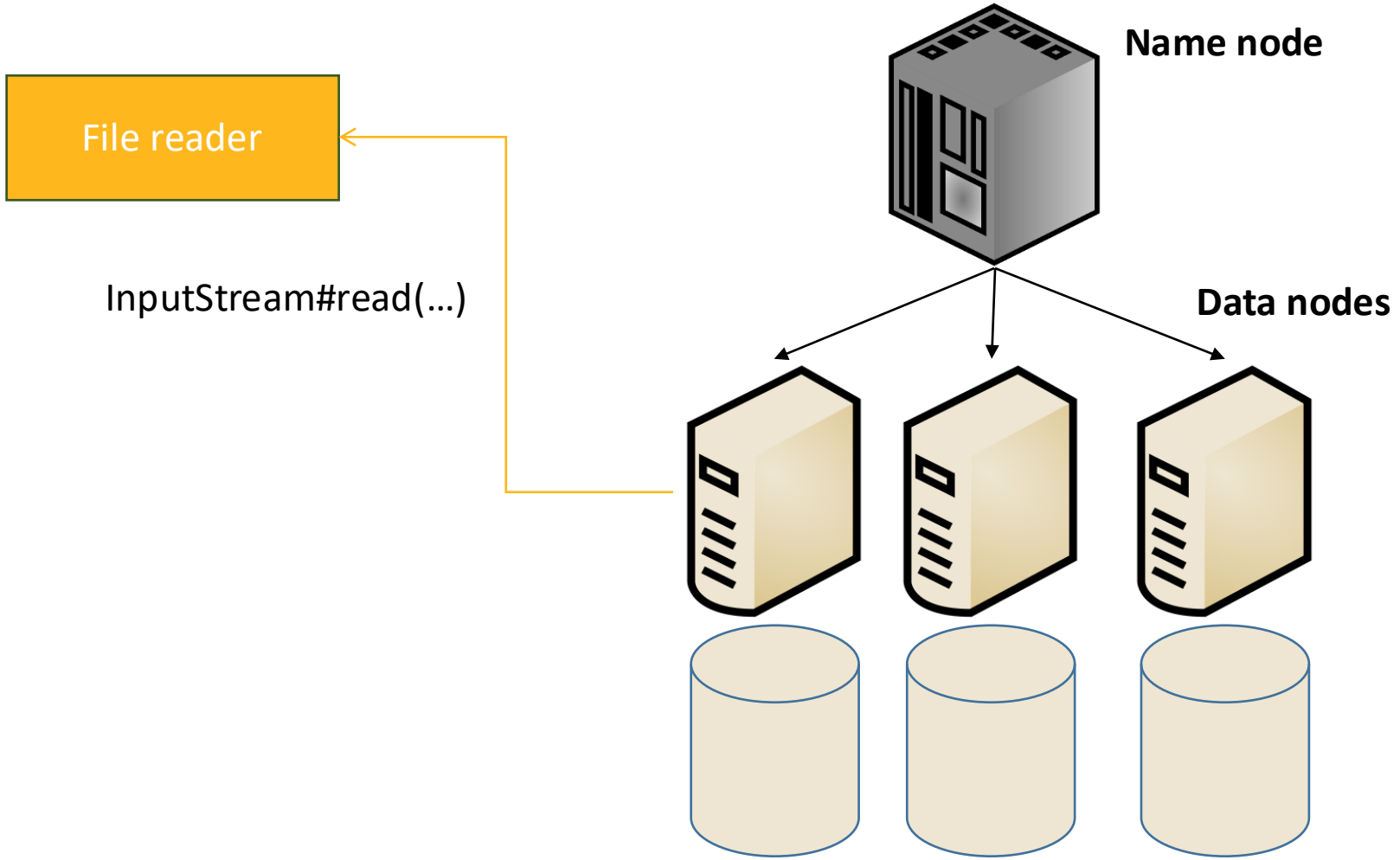
HDFS Reading Process



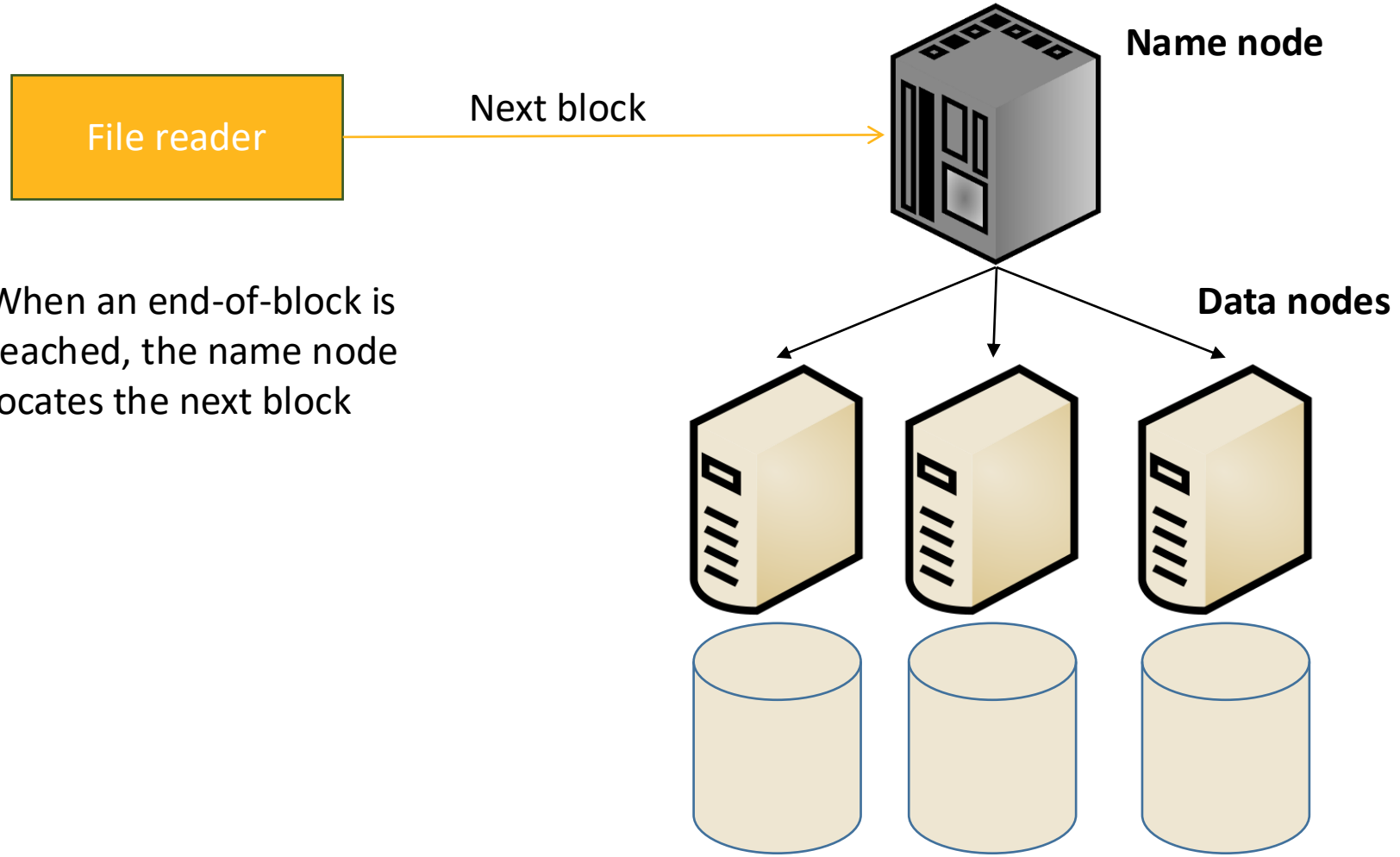
The name node locates the first block of that file and returns the address of one of the nodes that store that block

The name node returns an input stream for the file

HDFS Reading Process

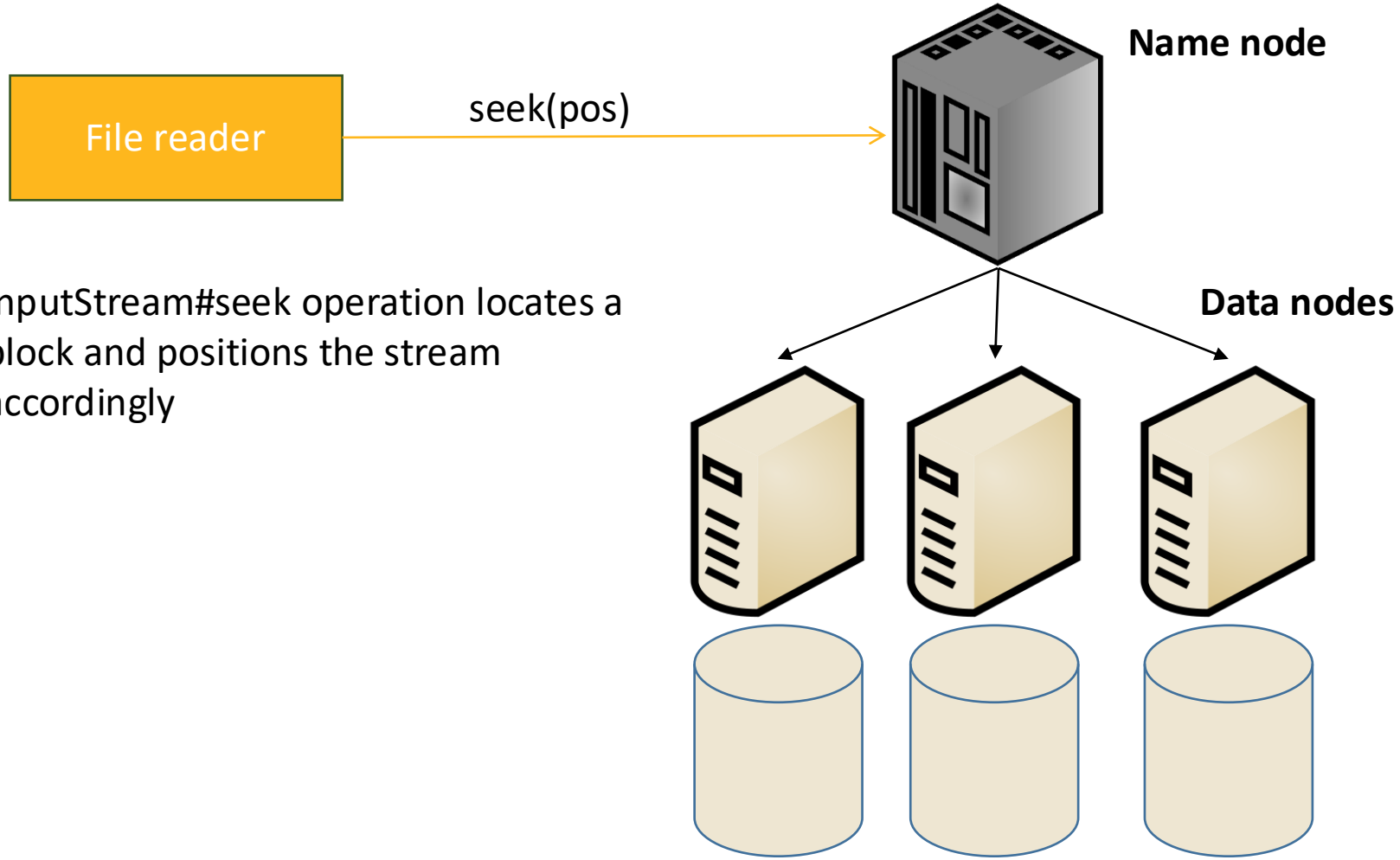


HDFS Reading Process



When an end-of-block is reached, the name node locates the next block

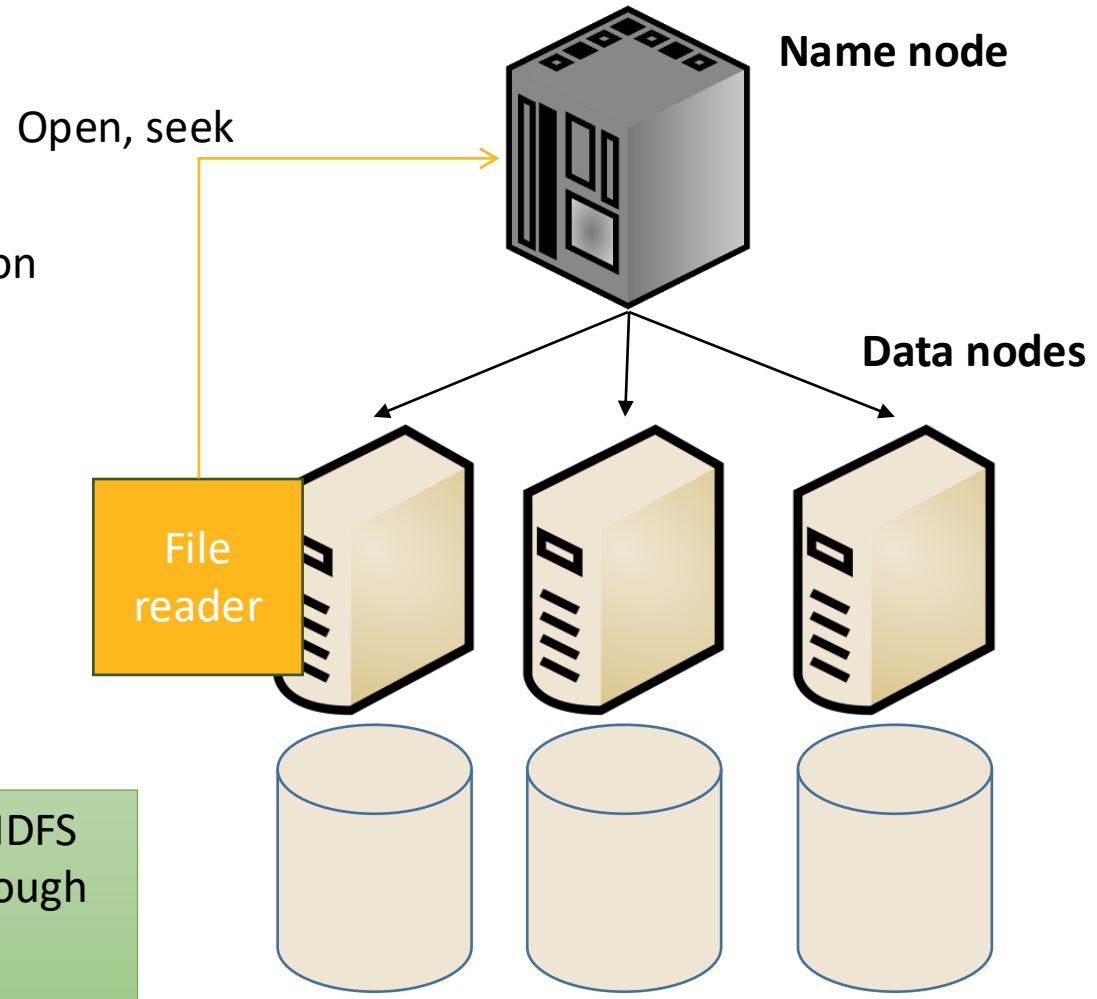
HDFS Reading Process



InputStream#seek operation locates a block and positions the stream accordingly

Reading from a datanode

1. If the block is locally stored on the reader, this replica is chosen to read
2. If not, a replica on another machine in the same rack is chosen
3. Any other random block replica is chosen



When self-reading occurs, HDFS can make it much faster through a feature called short-circuit

Notes About Reading

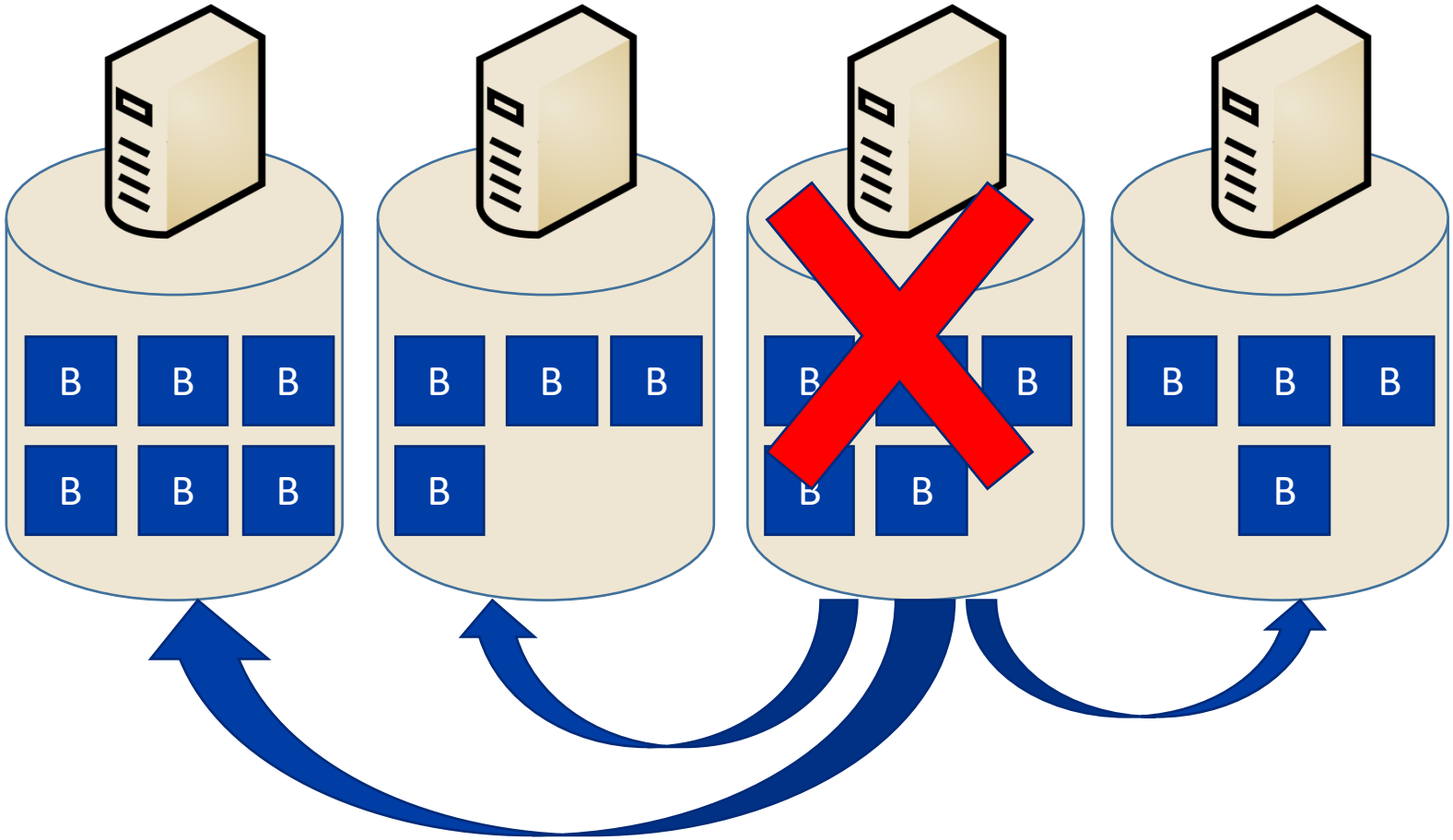
- The API is much richer than the simple open/seek/close API
  - You can retrieve block locations
  - You can choose a specific replica to read
- The same API is generalized to other file systems including the local FS and S3
- Review question: Compare random access read in local file systems to HDFS

Special Features in HDFS

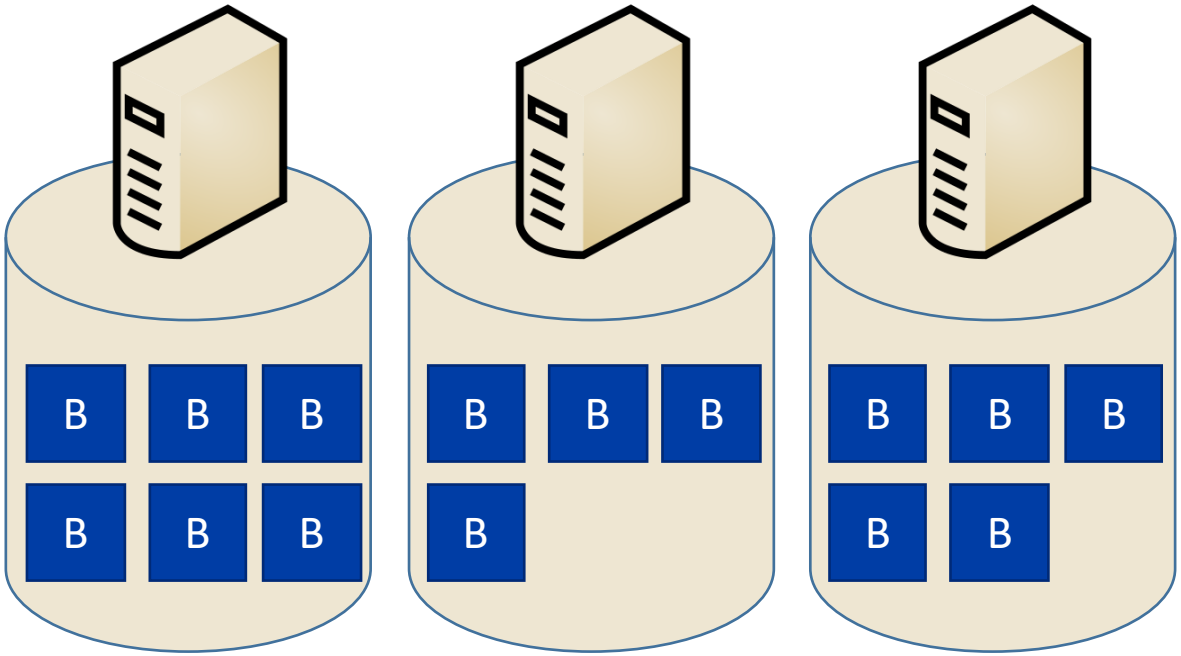
HDFS Special Features

- Node decommission
- Load balancer
- Cheap concatenation

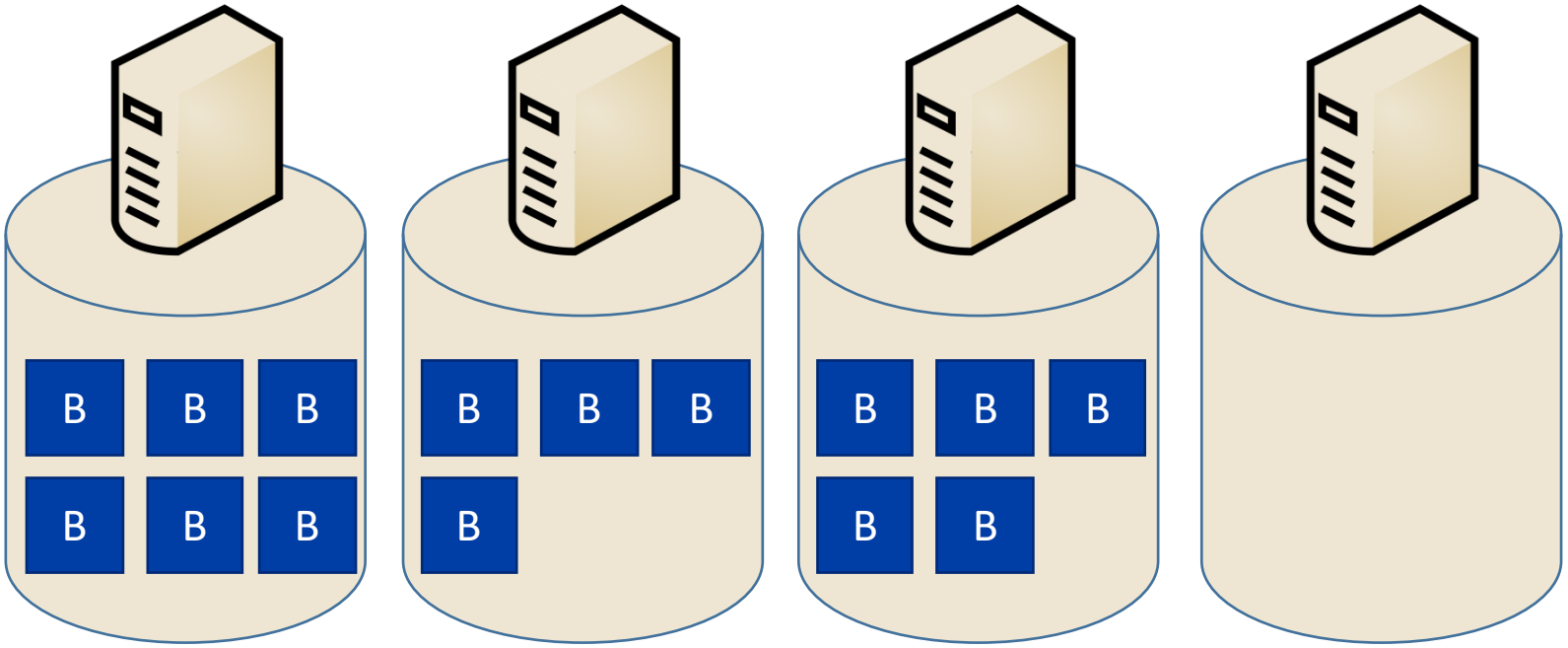
Node Decommission



Load Balancing

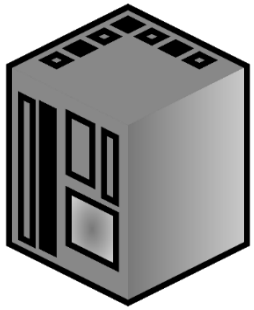


Load Balancing

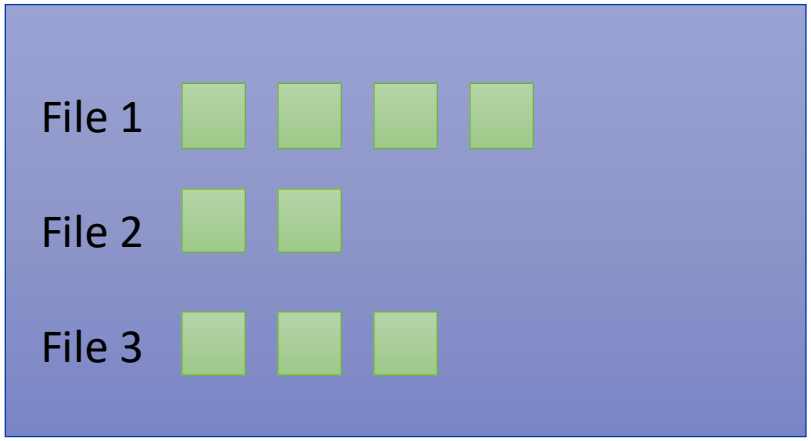


Start the load balancer

Cheap Concatenation



Name node



Concatenate File 1 + File 2 + File 3 ➔ File 4

Rather than creating new blocks, HDFS can just change the metadata in the name node to delete File 1, File 2, and File 3, and assign their blocks to a new File 4 in the right order.

Internals of HDFS
Structured Reading

If you start something, finish it

Data files

- In distributed big data processing, input files contain records not just raw bytes
- We need to a way to read records from files in HDFS
- For efficiency, we should split the file and read it in parallel

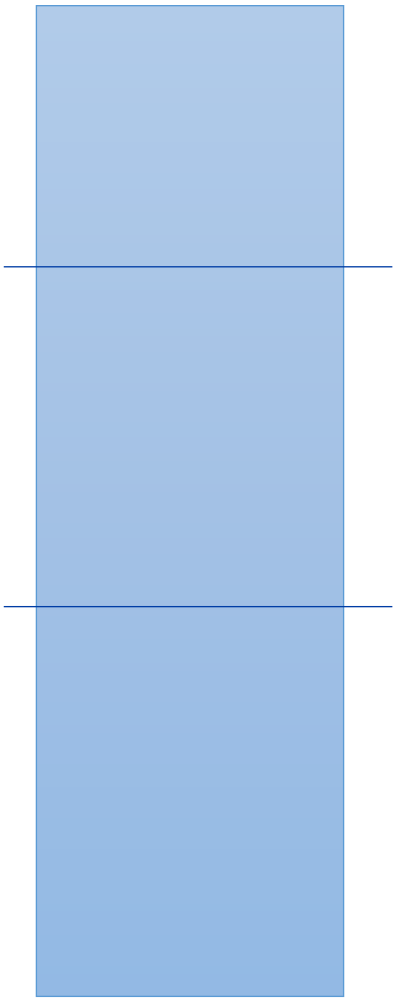
Block

- Part of the file that is **physically** stored in one or more data nodes

Split

- A **logical** part of the file defined by an offset and length
- Might or might not align with blocks

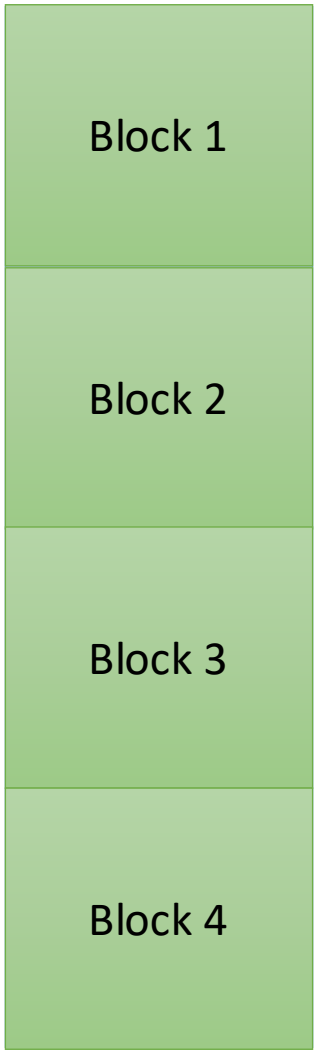
File Splitting



Input File

- How to split the file?
  - By record: For fixed-size records
  - By size: For variable-size records
- Considerations, splitting the file should be fast
- Should not need to read the entire file to split it

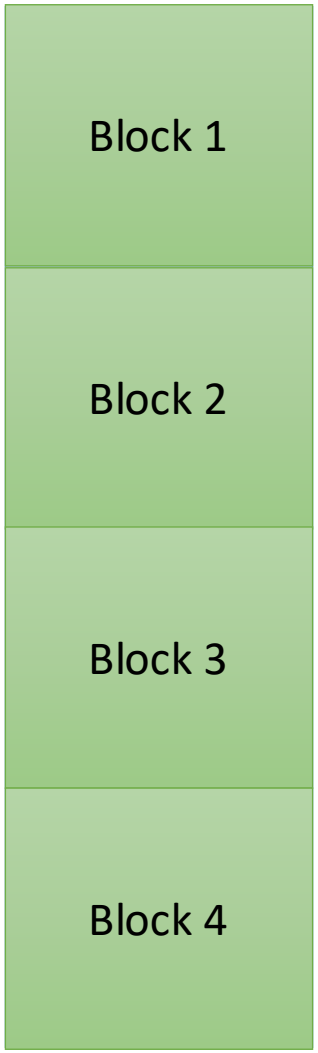
Simple File Splitting



Input File

- A split is created for each block
- Advantages
  - Data locality
  - Efficiency
- Drawback
  - A record might span more than one split

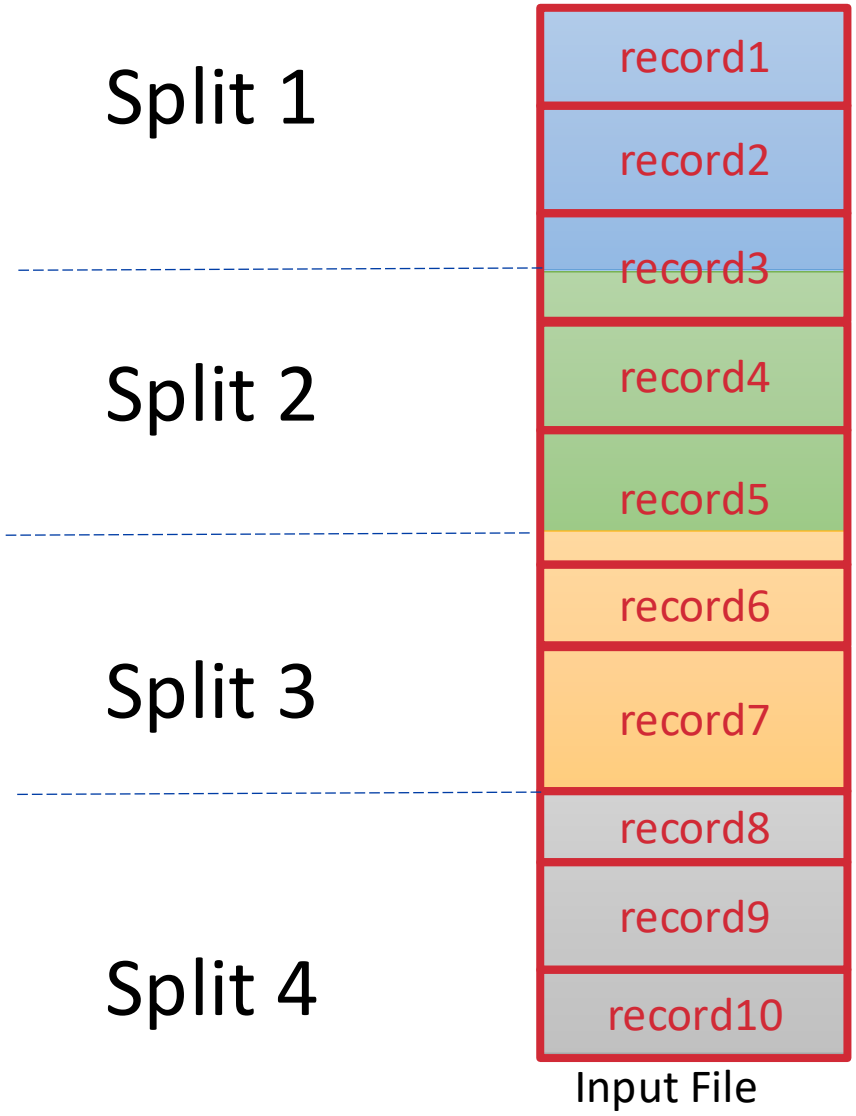
Simple File Splitting



Input File

- A split is created for each block
- Advantages
  - Data locality
  - Efficiency
- Drawback
  - A record might span more than one split

Read data in every split

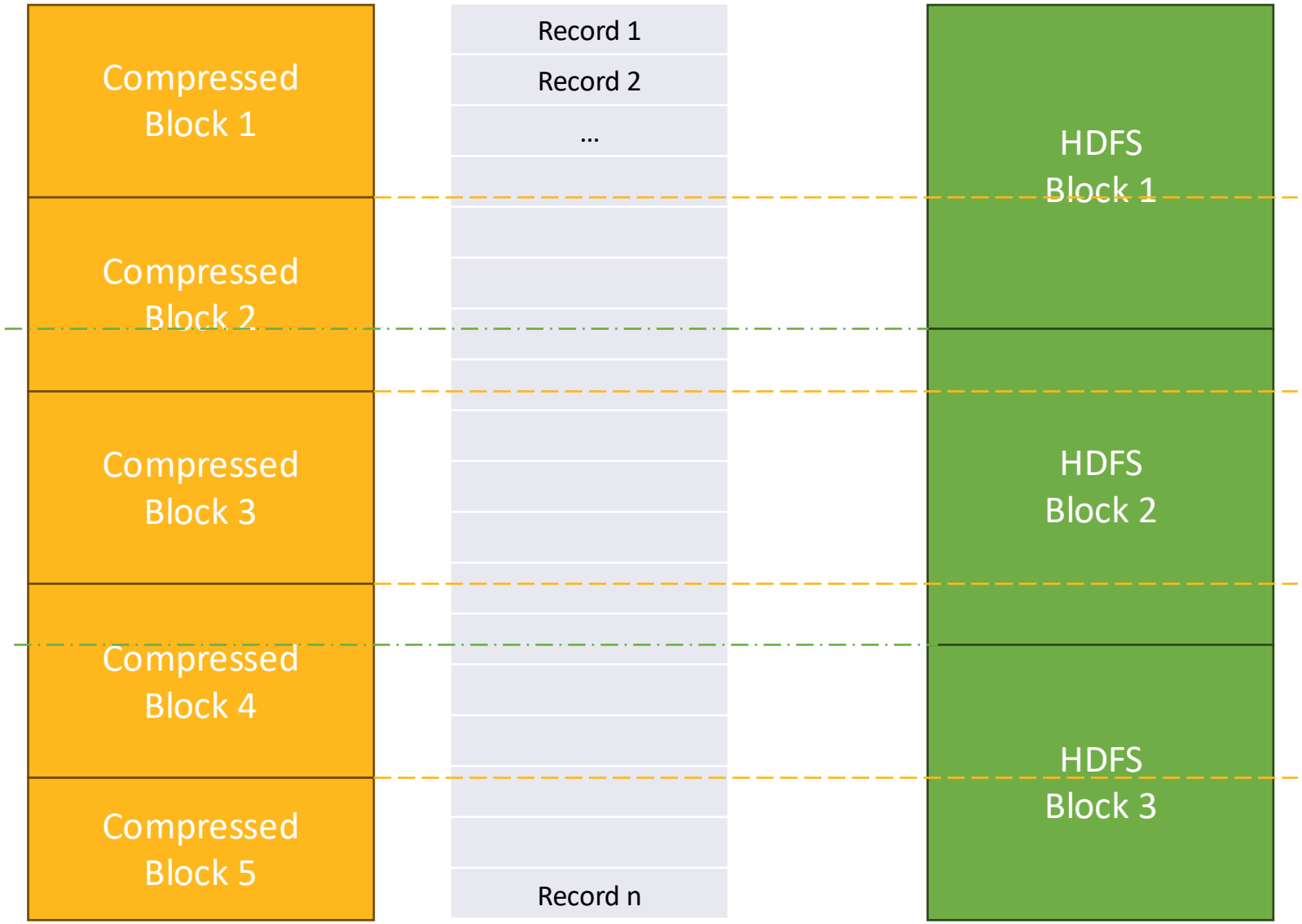


- Which records will be read for each of the four splits?

Reading process

- Split the file based on the file metadata
  - File size, block sizes, # of nodes
- Each split is defined by:
  - File name, Start offset, Length
- For each split:
  - Seek to the start offset
  - Skip the first record (except for the first split)
  - Read until the beginning of the record goes beyond the start + length

Reading block-compressed files



Conclusion

- HDFS is a general-purpose distributed file system
- It provides a write-once ready-many access interface
- Supports random reading which can be used for stream reading and structured reading
- Provides a Unix-like shell
- Provides a Java API for programming

Further Readings

- HDFS Architecture
  - https://hadoop.apache.org/docs/r3.2.2/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
- Shell commands
  - https://hadoop.apache.org/docs/r3.2.2/hadoop-project-dist/hadoop-common/FileSystemShell.html
- FileSystem API
  - https://hadoop.apache.org/docs/r3.2.2/api/org/apache/hadoop/fs/FileSystem.html