HDFS Practice Problem (Replication & Storage)

Arjun Vankani

08-08-2025

Problem 1: HDFS & Job Scheduling Example

Given:

- File size: 700 MB
- HDFS block size: 128 MB
- Replication factor: 3
- Cluster: 4 DataNodes

Step 1 – Splitting into Blocks

$$\text{No. of blocks} = \lceil \frac{700}{128} \rceil = 6$$

Blocks: B1, B2, B3, B4, B5, B6.

Step 2 – Replication

$$\text{Total stored blocks} = 6 \times 3 = 18$$

Step 3 – Example Block Distribution

Block	Replica 1	Replica 2	Replica 3
$B1$	$DN1$	$DN2$	$DN3$
$B2$	$DN2$	$DN3$	$DN4$
$B3$	$DN3$	$DN4$	$DN1$
$B4$	$DN4$	$DN1$	$DN2$
$B5$	$DN1$	$DN3$	$DN4$
$B6$	$DN2$	$DN4$	$DN1$

Step 4 – Storage Calculation Each block = 128 MB, replicated 3 times:

$$\text{Storage} = 6 \times 128 \times 3 = 2304 \text{ MB } (\approx 2.25 \text{ GB})$$

Step 5 – Job Scheduling Example

1. Client submits job to JobTracker.

2. JobTracker splits into map tasks (1 per block).
3. TaskTrackers run tasks on DataNodes storing the block (data locality).
4. Shuffle & sort intermediate output.
5. Reduce phase aggregates results to HDFS.

—

Problem 2 (Easy) – Block Calculation

A file of size 1.2 GB is stored in HDFS with:

- Block size = 256 MB
- Replication factor = 2

Calculate:

1. Number of blocks
2. Total storage space used with replication

—

Problem 3 (Easy) – Block Calculation

A 2 GB file is stored in HDFS with:

- Block size = 512 MB
- Replication factor = 3

Find the number of blocks and total storage required.
—

Problem 4 (Difficult) – Mixed File Sizes

A dataset consists of three files:

- File 1: 5 GB
- File 2: 2.5 GB
- File 3: 1.7 GB

Block size = 256 MB, replication factor = 3. Calculate total number of blocks and total storage requirement.

—

Problem 5 (Difficult) – Large File with High Replication

A single file of 12 GB is stored in HDFS with:

- Block size = 128 MB
- Replication factor = 4

Find the number of blocks and storage used.

—

Problem 6 (Difficult) – Different Block Sizes in Cluster

You upload a 10 GB file twice:

1. First time with block size = 128 MB, replication factor = 3
2. Second time with block size = 256 MB, replication factor = 2

Calculate blocks and storage for both cases.