Hadoop Distributed File System(HDFS)

赵敏 老师

Outline

- What is HDFS & Why HDFS
- Demo of HDFS Commands
- Overview of HDFS Architecture
- HDFS System Design.

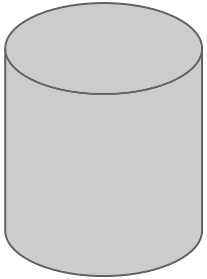
What is HDFS

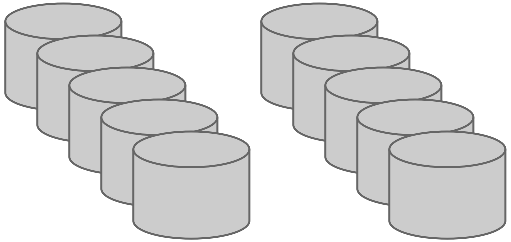
HDFS

- Hadoop Distributed File System
- A fault-tolerant file system designed to run on inexpensive hardware

Why HDFS

Challenge: Read 1TB of data





1 machine

10 machines

4 I/O channels
Each channel: 100MB/S

4 I/O channels
Each channel: 100MB/S

45 MIN

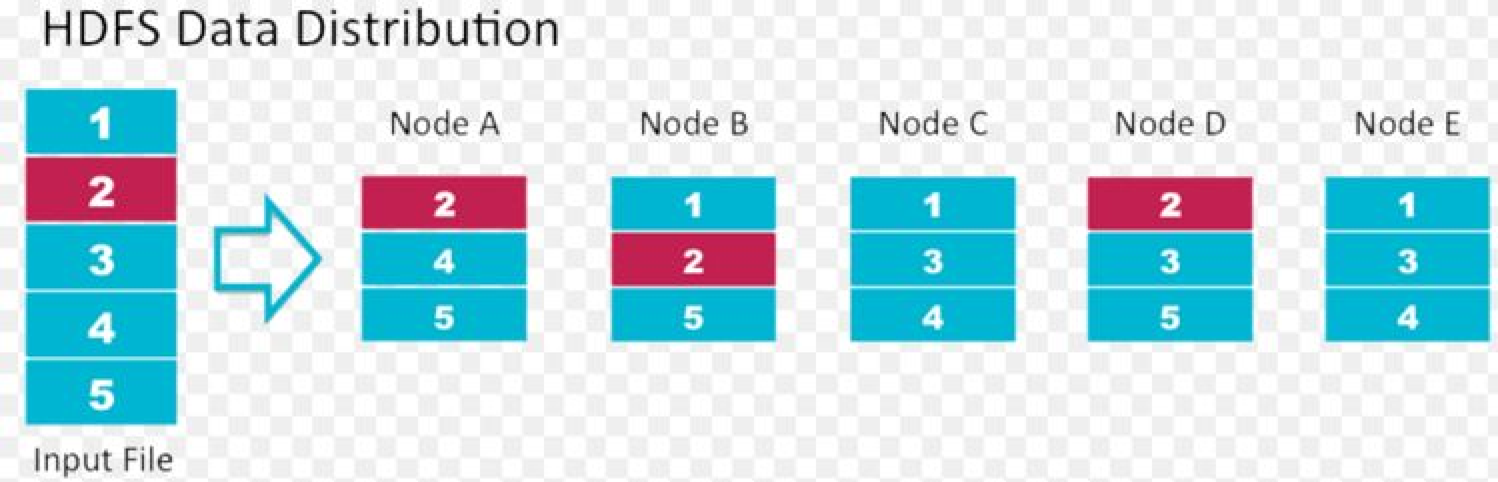
4.5 MIN

Why HDFS

Faster

Why HDFS

- Split
- Replication



Why HDFS

Easier

HDFS Commands

Demo

Overview of HDFS Architecture

Why do we learn HDFS architecture?

Overview of HDFS Architecture

- In interview: can be applied to all file system design
- Help you understand the whole hadoop ecosystem

Overview of HDFS Architecture

Since we have multiple machines, how could they communicate?

- Do they talk to each other?

Overview of HDFS Architecture

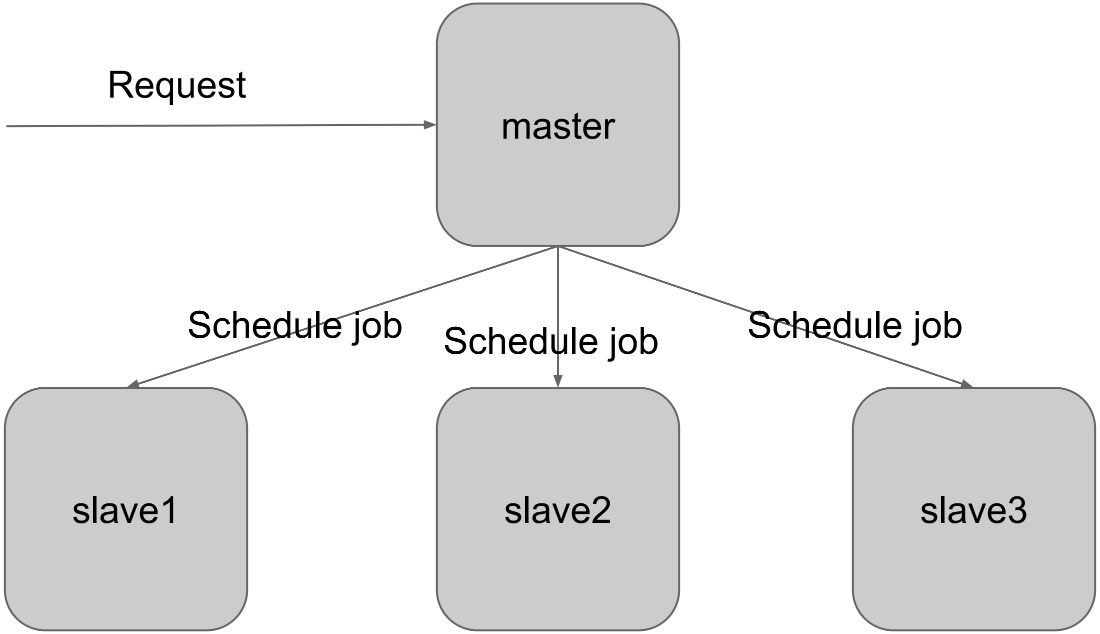
- Master slave model
  - High consistency
  - Simpler design
  - Single master node is not robust
- Peer peer model
  - Distributes read write load
  - One node down won’t affect others
  - Low consistency

http://stackoverflow.com/questions/3736969/master-master-vs-master-slave-database-architecture

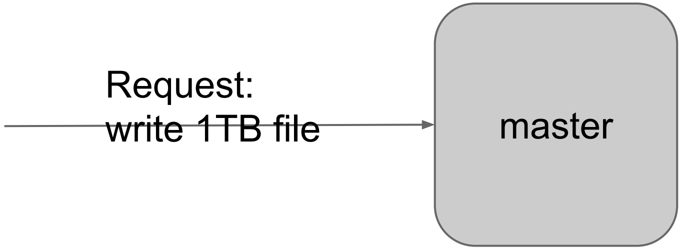
Overview of HDFS Architecture

HDFS uses master-slave model.

Overview of HDFS Architecture



Overview of HDFS Architecture



Will master transport data?

Overview of HDFS Architecture

- No, or it will be a bottleneck.
- Master will decide which slave node to read/write, then client will talk to slave node.

Overview of HDFS Architecture

How to store file?

- Whole big file
- Blocks of small files

Overview of HDFS Architecture

- Blocks of small files

Overview of HDFS Architecture

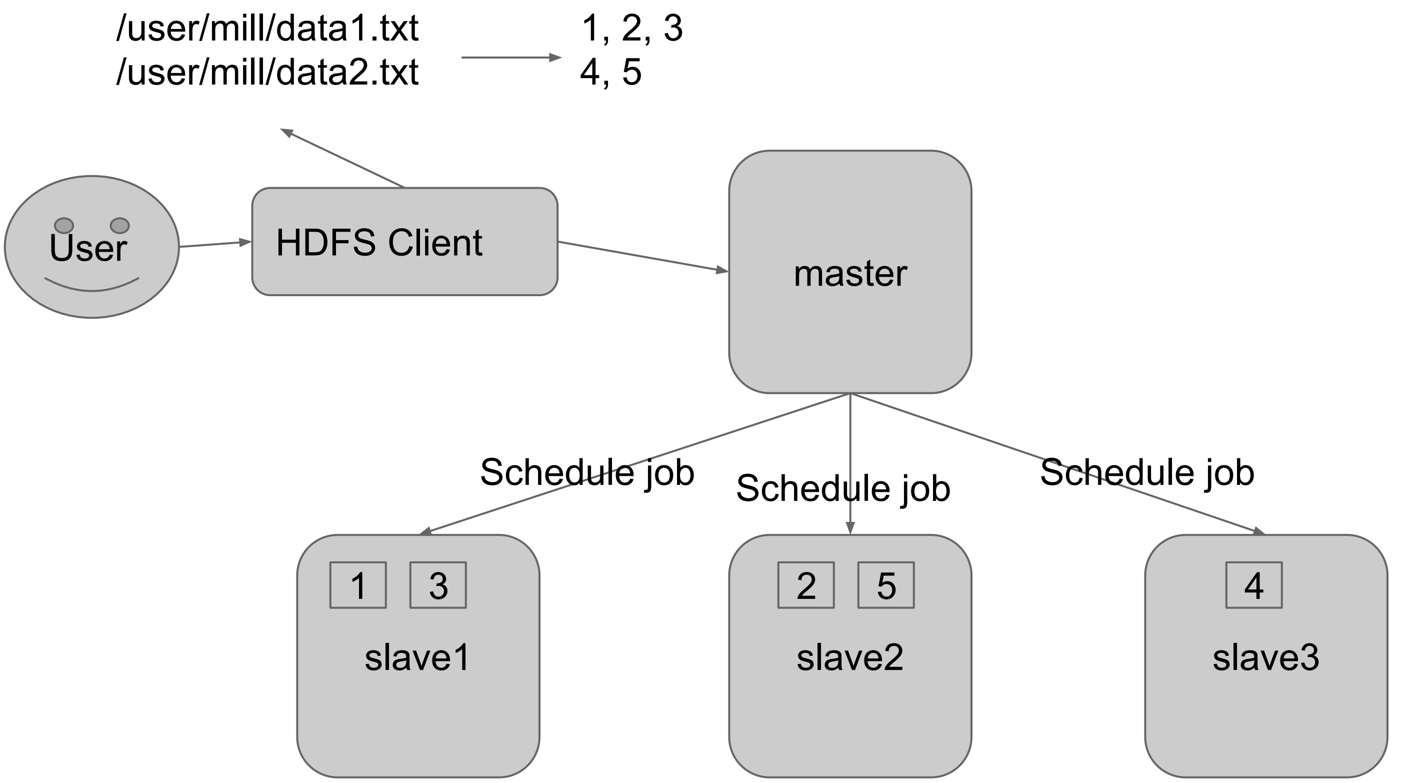
- Who will divide file into blocks?
    - Master node?
    - Slave node?

Overview of HDFS Architecture

Neither master nor slave….

---> HDFS client

Overview of HDFS Architecture



Overview of HDFS Architecture

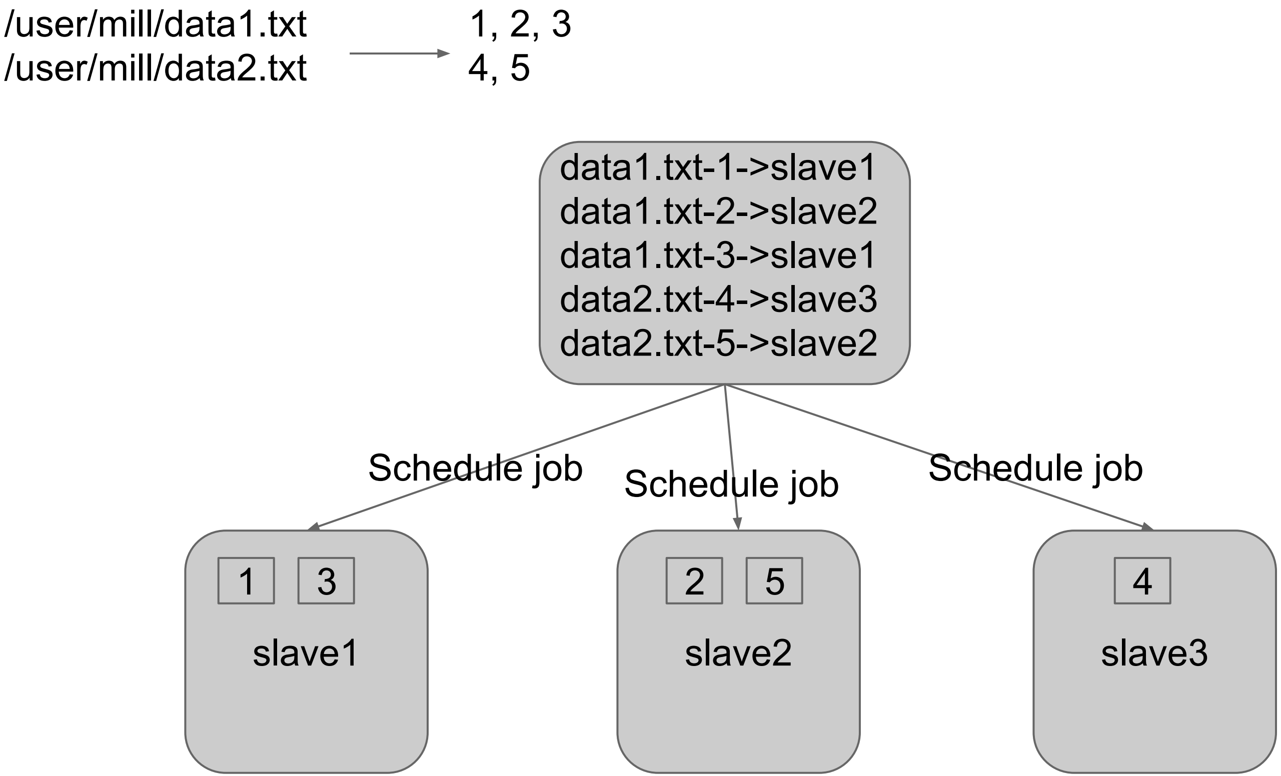
What do we store on master machines?

Overview of HDFS Architecture

Master should know where to write and read

- Metadata
  - File blocks
  - Where are the blocks

Overview of HDFS Architecture

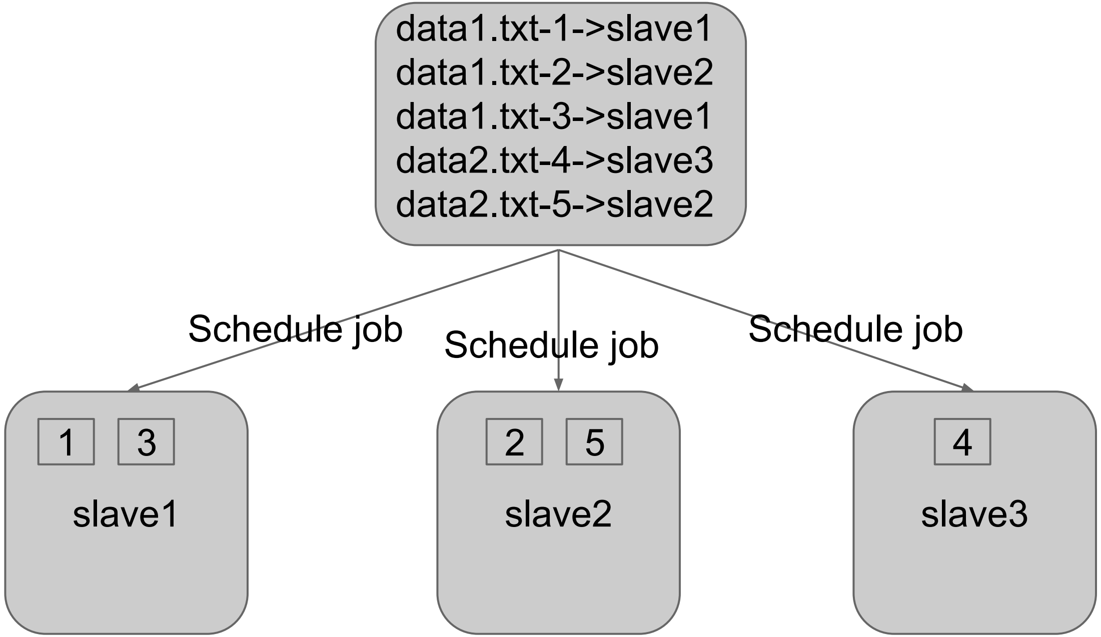


Overview of HDFS Architecture

What if one slave node fail?

Overview of HDFS Architecture

Lose data!



data1.txt-1->slave1
data1.txt-2->slave2
data1.txt-3->slave1
data2.txt-4->slave3
data2.txt-5->slave2

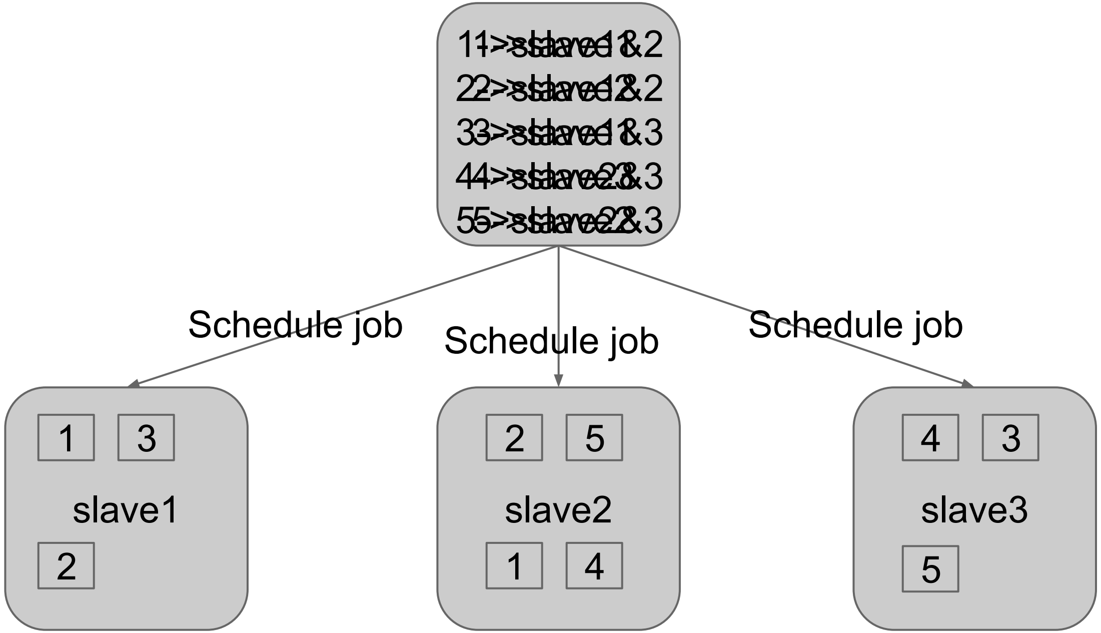
Overview of HDFS Architecture

How to avoid data loss?

Overview of HDFS Architecture

Data replication

Overview of HDFS Architecture



1->slave1&2
2->slave1&2
3->slave1&3
4->slave2&3
5->slave2&3

Overview of HDFS Architecture

What if one master node fail?

Overview of HDFS Architecture

We have a checkpoint node to copy all the data from master node per hour.

Overview of HDFS Architecture

Master-slave architecture

- *Single NameNode* - a master server that manages the file system namespace and regulates access to files by clients.

- *Multiple DataNodes* – typically one per node in the cluster.
  - Manage storage
  - Serving read/write requests from clients
  - Block creation, deletion, replication based on instructions from NameNode

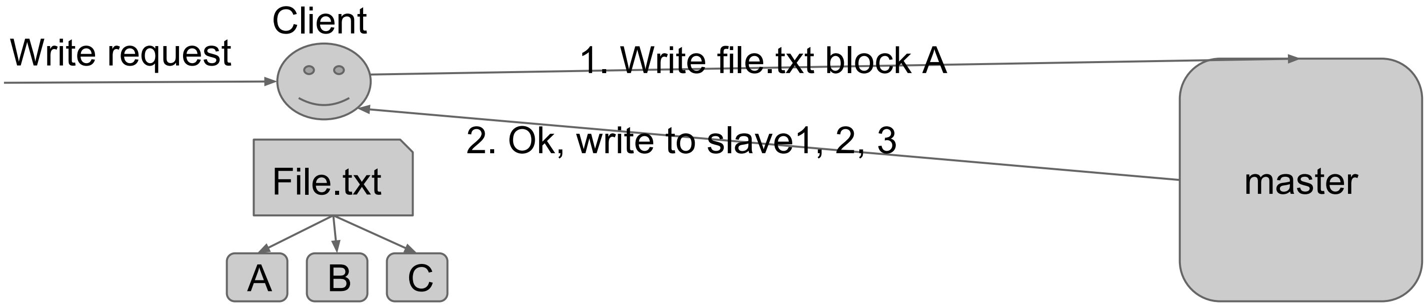
Overview of HDFS Architecture

Read & Write

Overview of HDFS Architecture

How to do the write operation?

Overview of HDFS Architecture: Write



Overview of HDFS Architecture

Who should client write to?

Overview of HDFS Architecture

Slave nodes

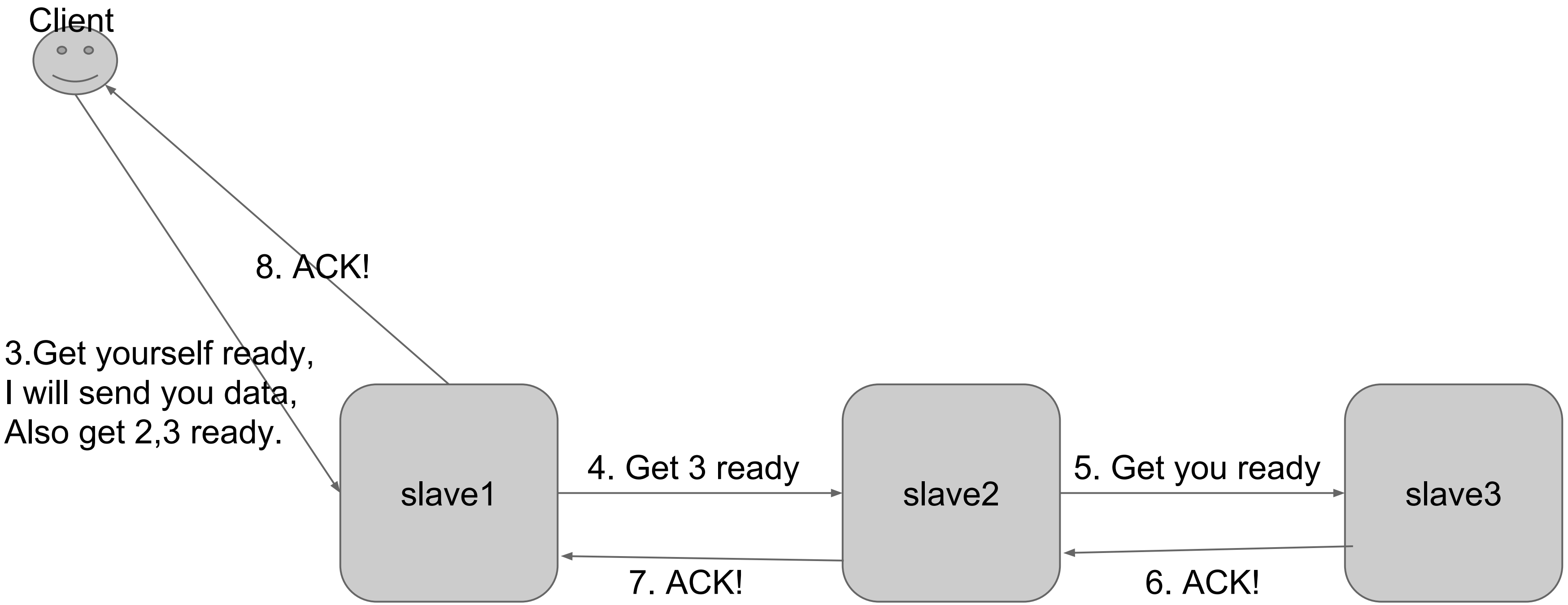
Overview of HDFS Architecture

Since we have three slave nodes to write, should client writes to three nodes?

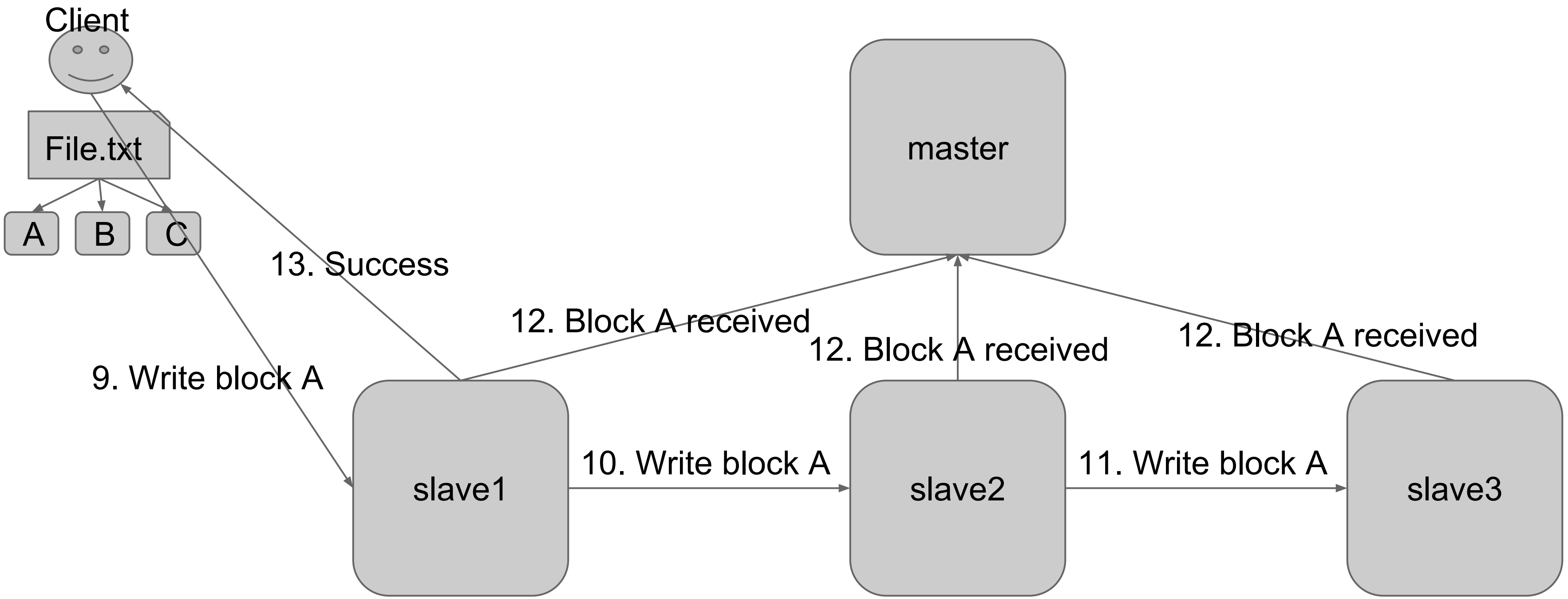
Overview of HDFS Architecture

No, will result in bottleneck.

Overview of HDFS Architecture: Write



Overview of HDFS Architecture: Write



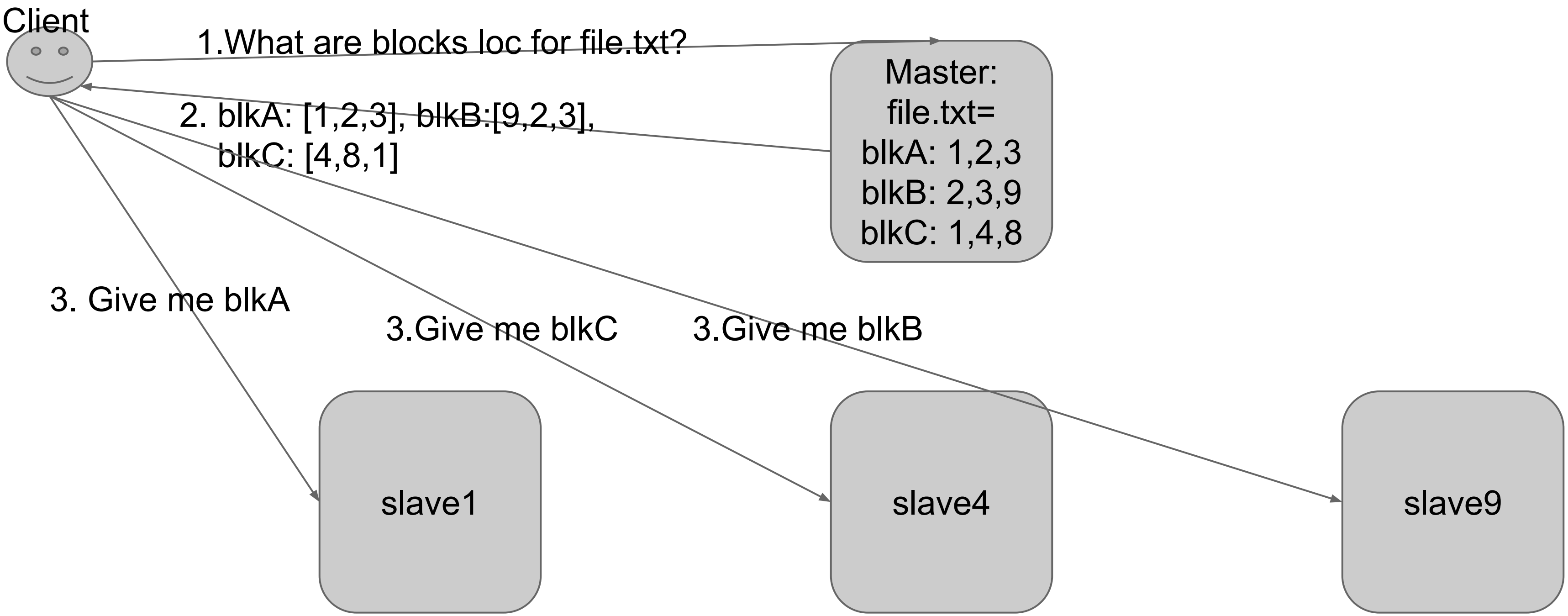
Overview of HDFS Architecture: Write

- Data is kept in different racks. To ensure if one rack fails, we still have another rack to hold the data.
- Keep two blocks in same rack to achieve high throughput while reading data because two machines in same rack have more bandwidth and lower latency.
- Client does not send blocks to all 3 data nodes identified by Name node. The reason is Client will be choked by data transmission at a time.
- Name Node creates metadata from block reports received from data nodes

Overview of HDFS Architecture

How to do the read operation?

Overview of HDFS Architecture: Read



Client

1.What are blocks loc for file.txt?

Master:
file.txt=
blkA: 1,2,3
blkB: 2,3,9
blkC: 1,4,8

2. blkA: [1,2,3], blkB:[9,2,3],
blkC: [4,8,1]

3. Give me blkA

3.Give me blkC

3.Give me blkB

slave1

slave4

slave9

Overview of HDFS Architecture: Read

blkB:[9,2,3]

Since we have 3 replica, which to read?

Overview of HDFS Architecture: Read

blkB:[9,2,3]

To visit the closest one.

Overview of HDFS Architecture

blkB:[9,2,3]

Why does the master provide this slave nodes order?

Overview of HDFS Architecture

- The slave nodes order are decided by the distance between slave node and the client.
- The closer, the faster.

What you have learned

- Concept about HDFS
- HDFS Commands
- How to design a file system
- How to design the communication in file system