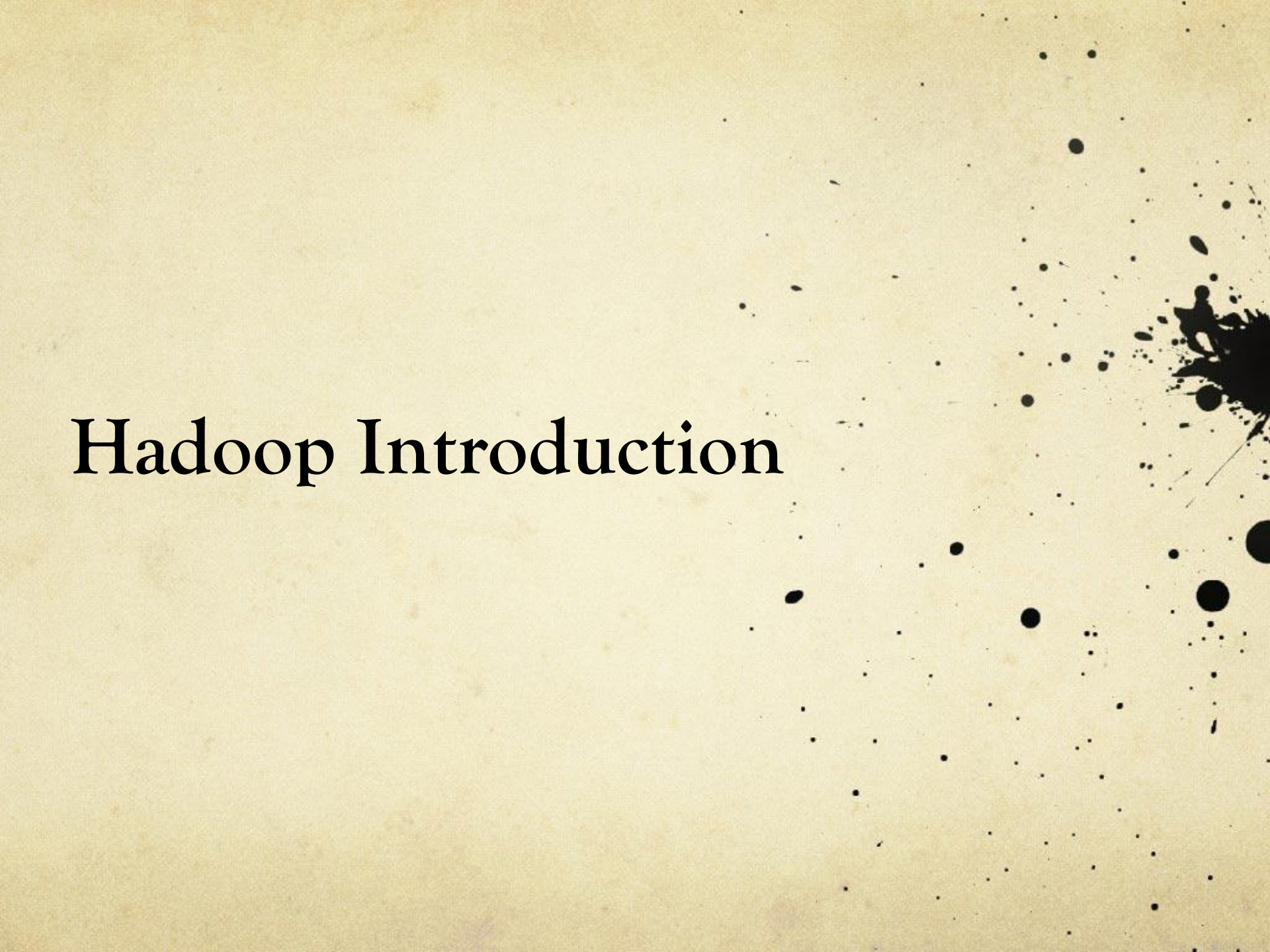
Hadoop Training - III

Architecture & Theory

Credit

- Show up: 50 points
- Asking questions: 10 points
- Answering questions: 15 points
- Exam: 20 points

Hadoop Introduction



History

- Created by Doug Cutting
- Named after his son's toy elephant
- Developed to support distribution for the Nutch search engine project

www.nytimes.com/imagepages/2009/03/16/business/17cl

The New York Times

March 16, 2009

Peter DaSilva for The New York Times

Doug Cutting with the stuffed elephant that inspired the name Hadoop, the software program he developed.

Close Window

http://www.nytimes.com/2009/03/17/technology/business-computing/17cloud.html?_r=1

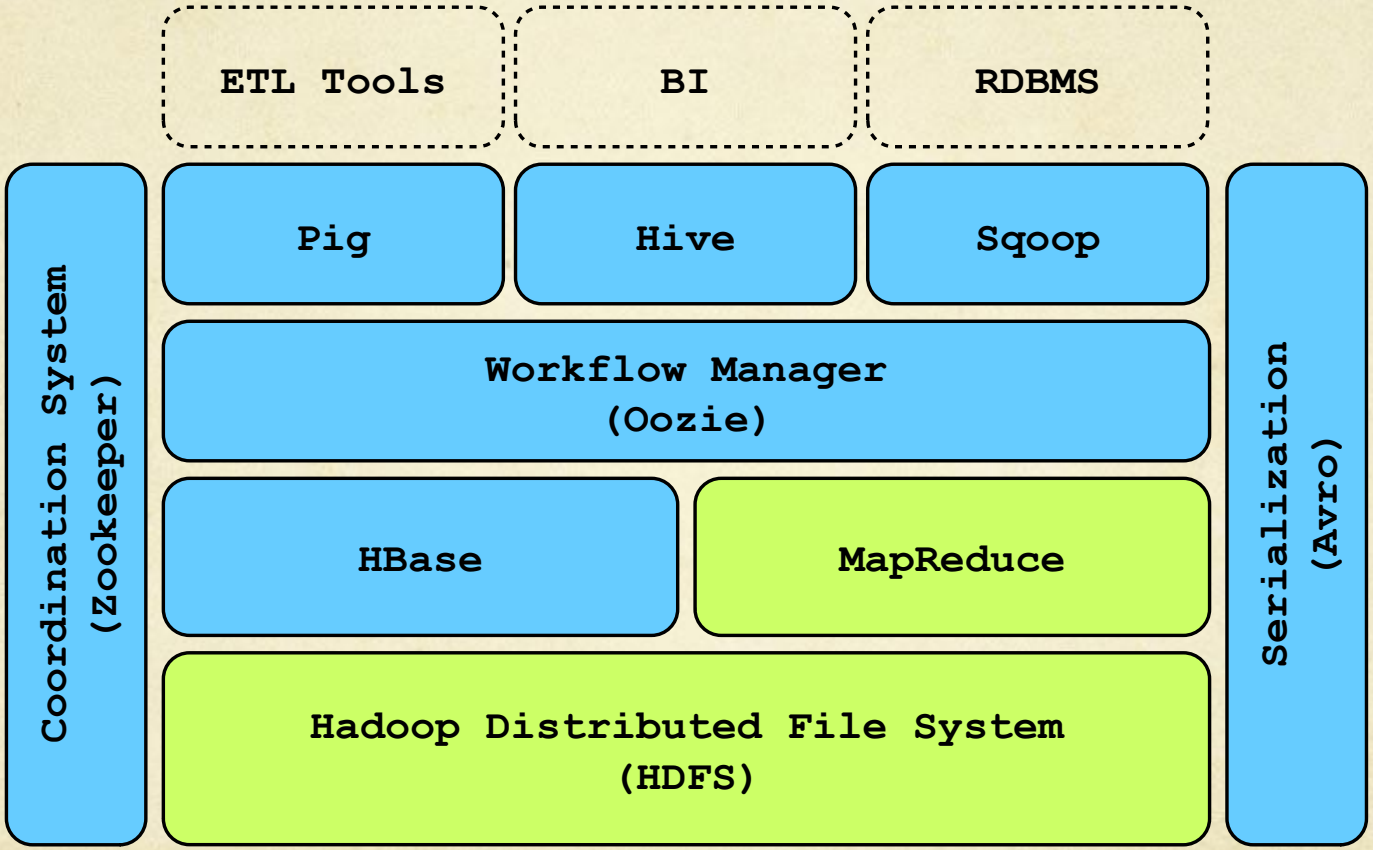
Assumptions and Goals

- Large Data Sets (GB, TB, PB, EB?, ZB?, YB?)
- Write-Once-Read-Many Access Model
- Streaming Data Access
- Moving Computation is Cheaper than Moving Data
- Hardware Failure
- Portability Across Heterogeneous Hardware and Software Platforms

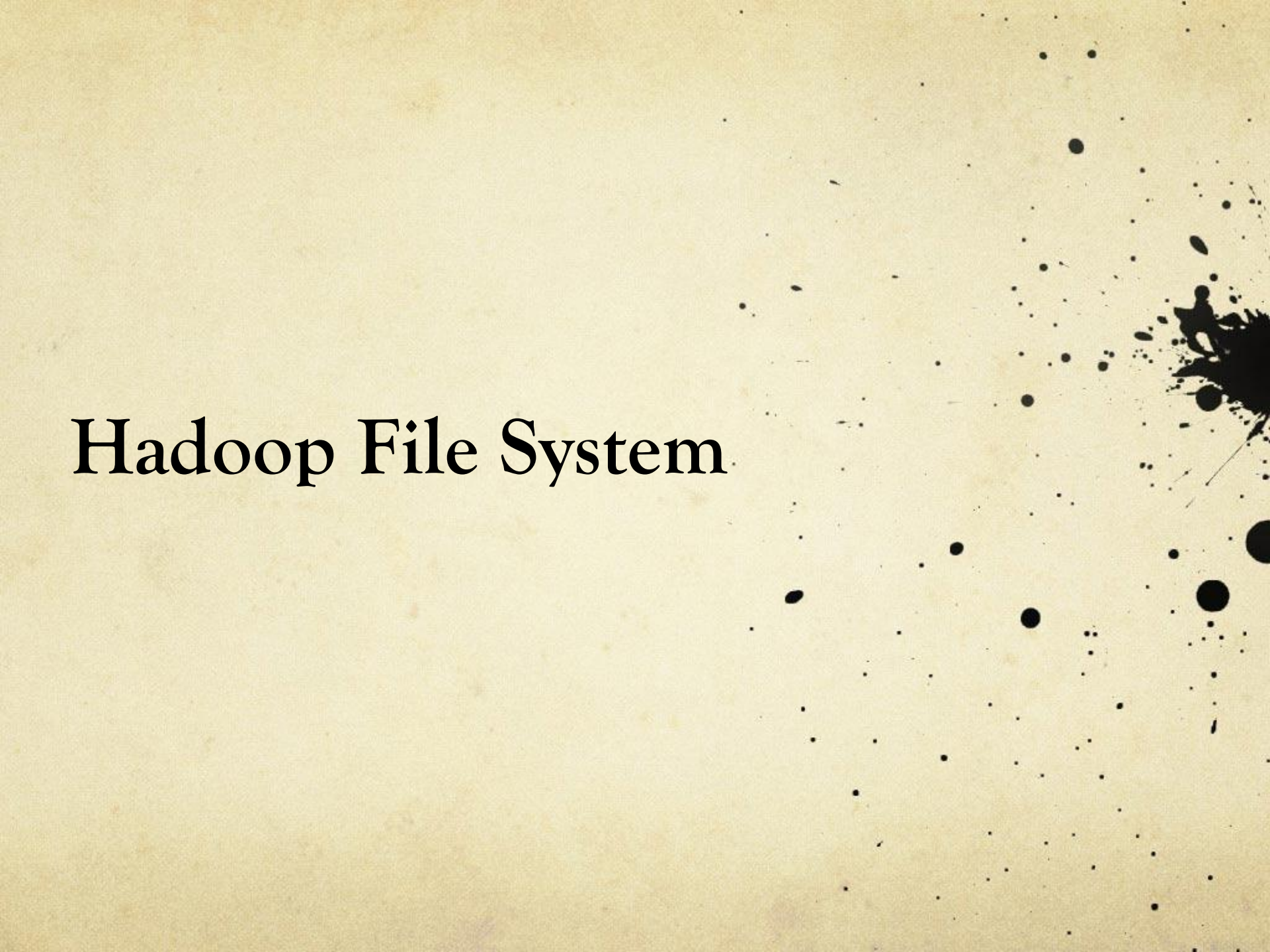
What Hadoop is Not

- Hadoop is not a substitute for a database
- MapReduce is not always the best algorithm
- HDFS is not a substitute for a High Availability SAN-hosted File System
- HDFS is not a POSIX File System

Hadoop Ecosystem



Hadoop File System

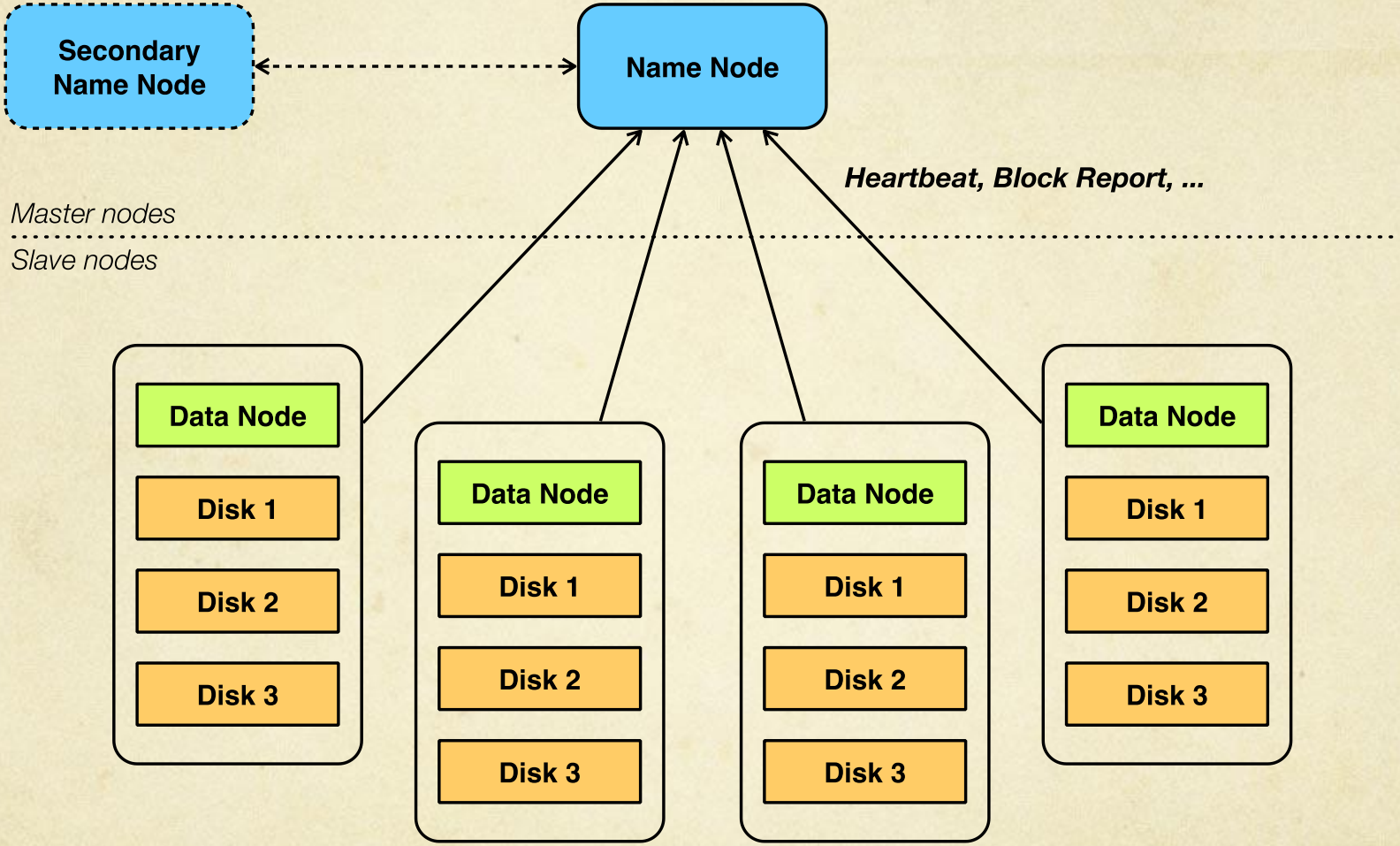


HDFS Concepts

- Files are broken into block-sized chunks
  - **dfs.block.size**, default is 64 MB
- File Replication
  - **dfs.replication**, default is 3
  - **dfs.replication.max**, default is 512
  - **dfs.replication.min**, default is 1

HDFS Architecture

- Master-Slave Architecture
- Name Node (Master)
- Data Node (Slave)
- Secondary Name Node


Secondary Name Node
Name Node
Heartbeat, Block Report, ...
Master nodes
Slave nodes
Data Node
Disk 1
Disk 2
Disk 3
Data Node
Disk 1
Disk 2
Disk 3
Data Node
Disk 1
Disk 2
Disk 3
Data Node
Disk 1
Disk 2
Disk 3

Name Node

- Manage File System Namespace
  - Maintain two critical tables
    1. File Name ➔ Block Sequence
    2. Block ID ➔ Machine List
- Single Point of Failure

File Name	Replicas	Block Sequence	Others
/data/part-0	2	B1, B2, B3	user, group, ...
/data/part-1	3	B4, B5	foo, bar, ...

Memory

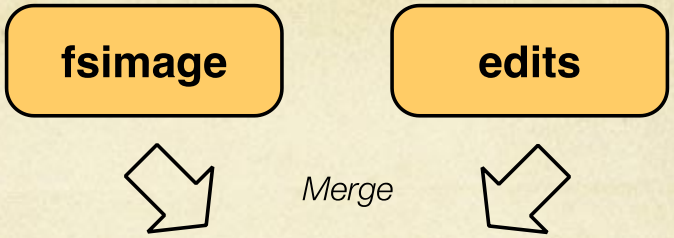
Disk

fsimage

File Name	Replicas	Block Sequence	Others
/data/part-0	3	B1, B2, B3	user, group, ...
/data/part-1	3	B4, B5	user, group, ...

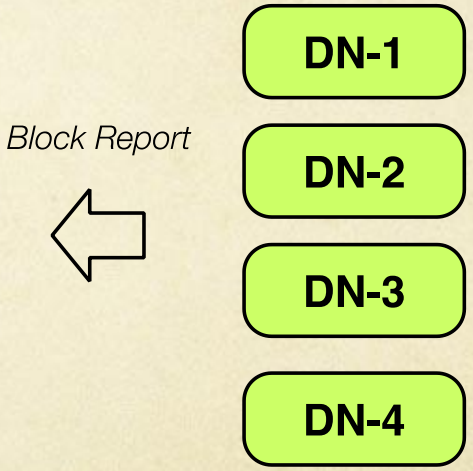
edits

OP Code	Operands
OP_SET_REPLICATION	"/data/part-0", 2
OP_SET_OWNER	"/data/part-1", "foo", "bar"



File Name	Replicas	Block Sequence	Others
/data/part-0	2	B1, B2, B3	user, group, ...
/data/part-1	3	B4, B5	foo, bar, ...

Block ID	Machine (Data Node) List
B1	DN-1, DN-2
B2	DN-2, DN-3
B3	DN-3, DN-4
B4	DN-4, DN-1, DN-2
B5	DN-2, DN-3, DN-1

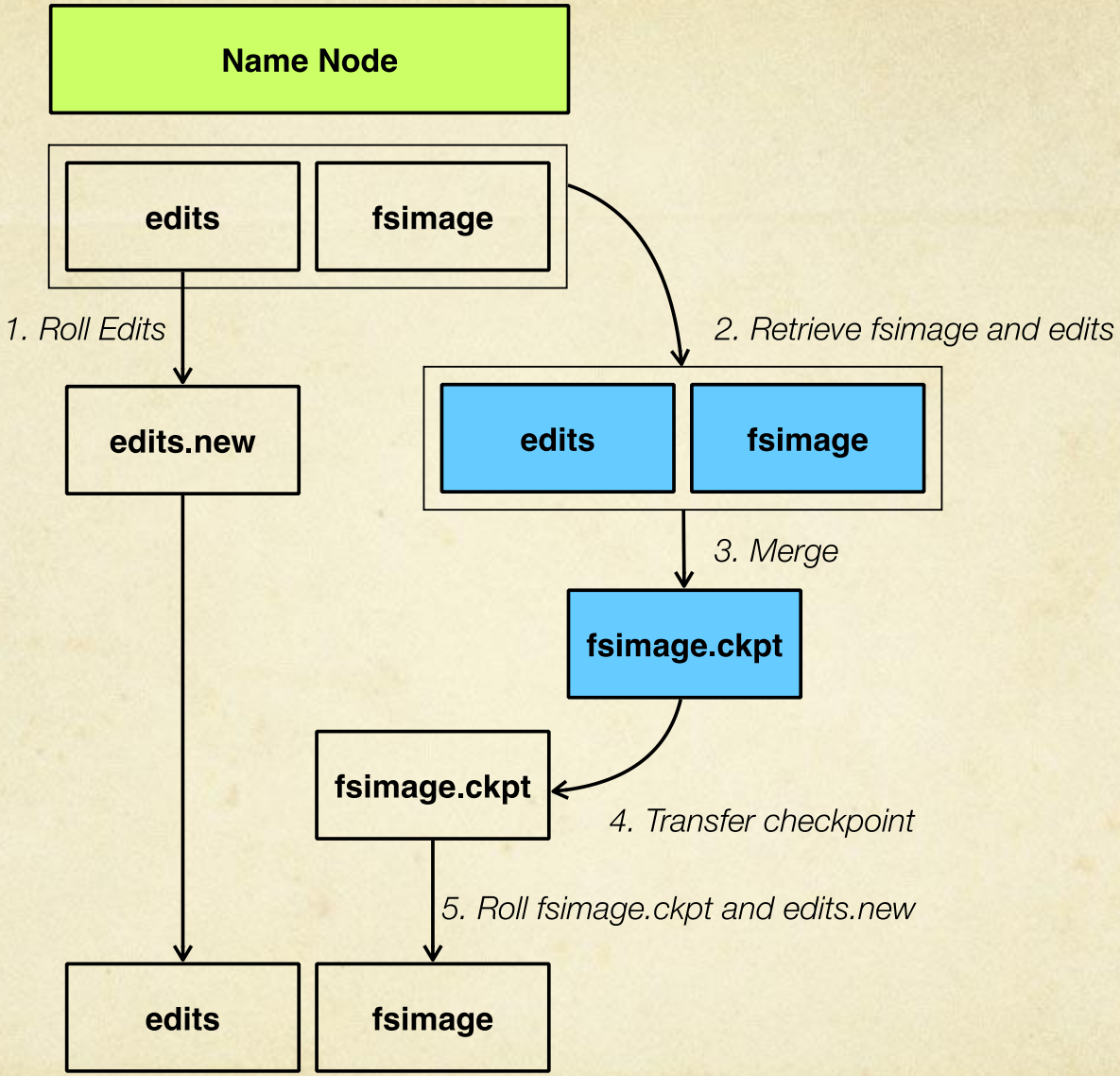


Data Node

- Manage Storage
- Replicate Blocks
- Block Report
- Send Heartbeat to Name Node
  - **dfs.heartbeat.interval**, default 3 seconds

Secondary Name Node

- Not a backup of Primary Name Node
- Merge fsimage and edits log


Name Node
edits
fsimage
1. Roll Edits
2. Retrieve fsimage and edits
edits.new
edits
fsimage
3. Merge
fsimage.ckpt
fsimage.ckpt
4. Transfer checkpoint
5. Roll fsimage.ckpt and edits.new
edits
fsimage

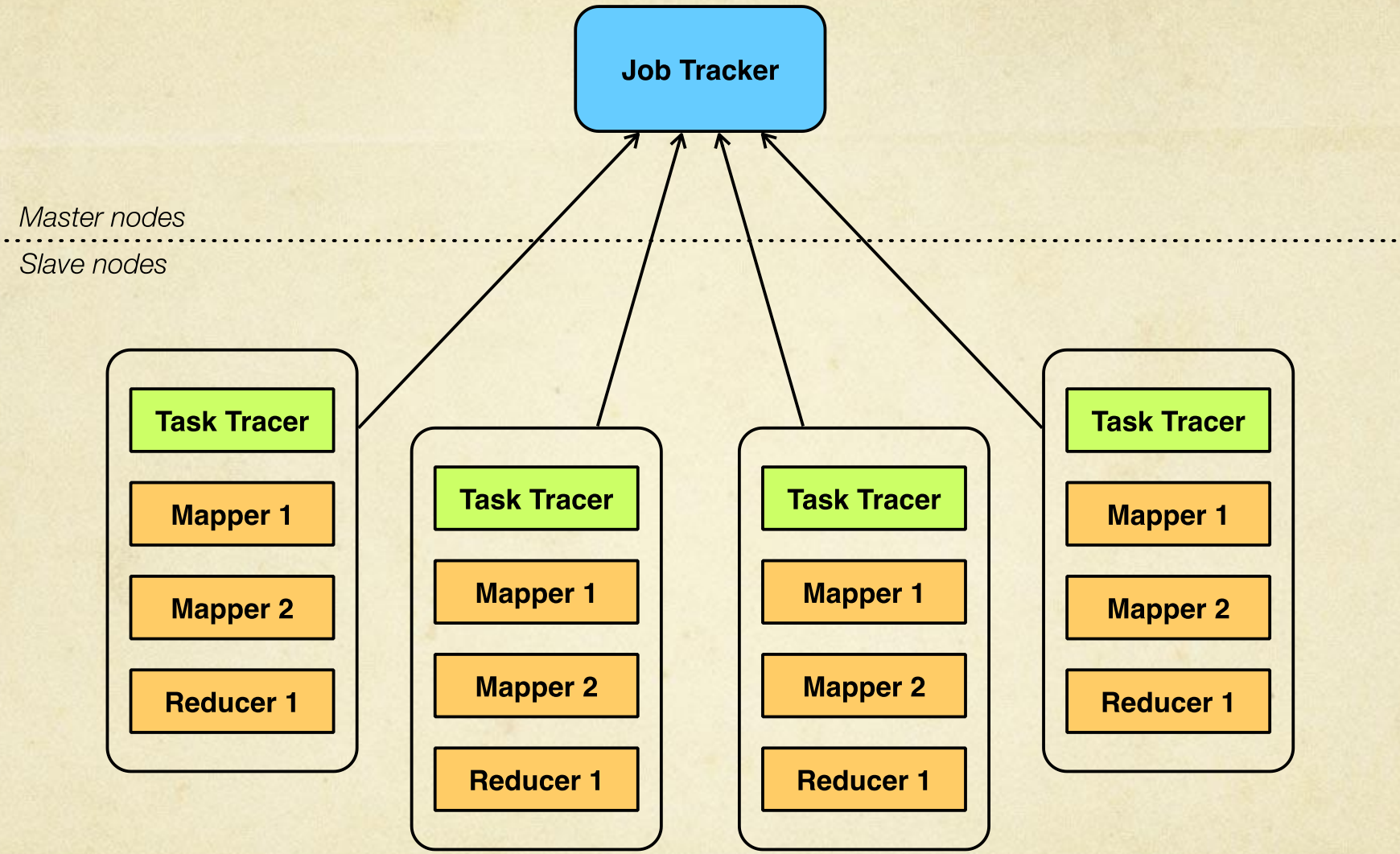
Hadoop MapReduce

MapReduce Concepts

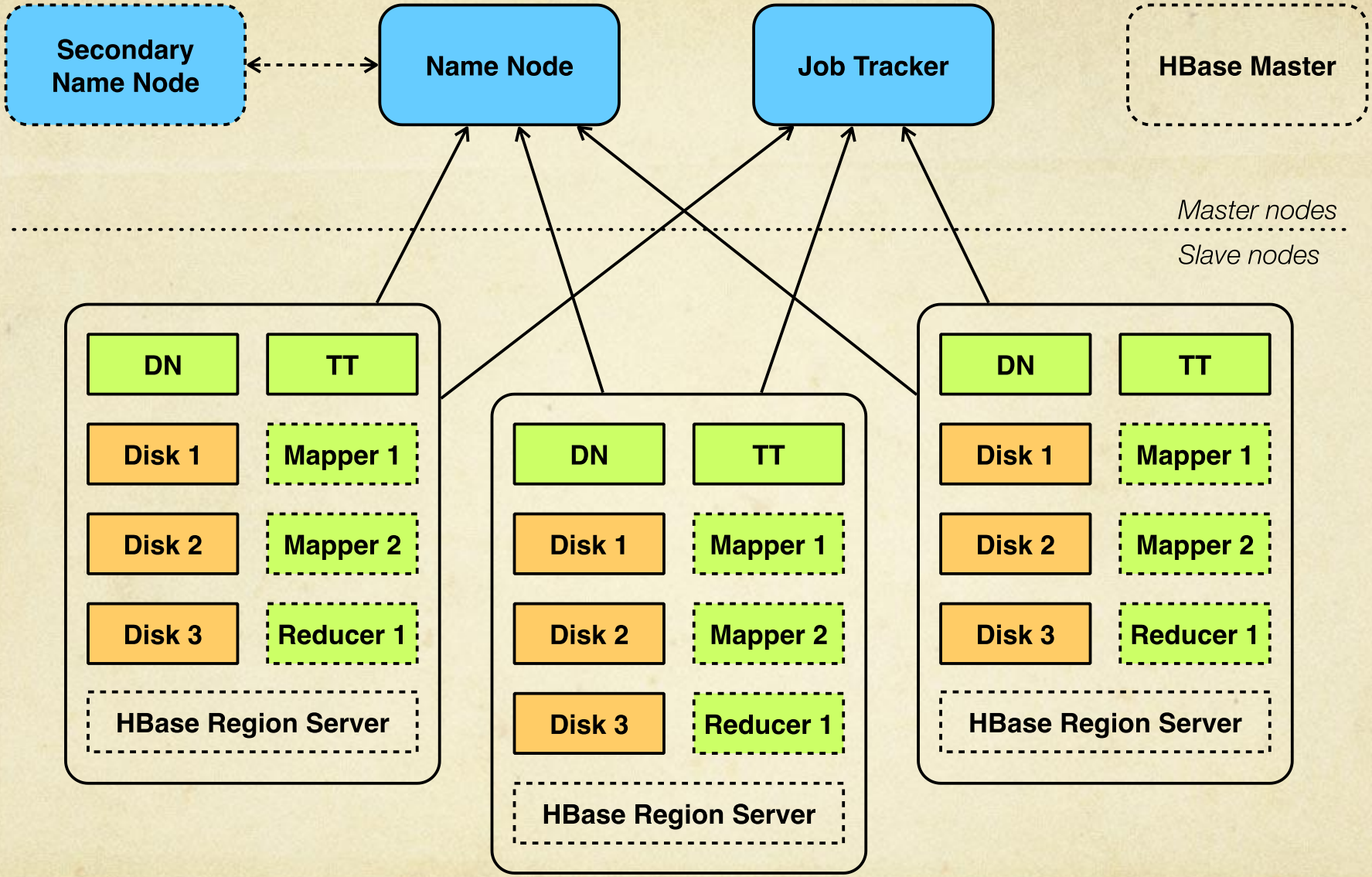
- Jobs are broken into tasks
- Parallelize
  - Computation
  - Disk IO

MapReduce Architecture

- Master-Slave Architecture
- Job Tracker (Master)
- Task Tracker (Slave)
  - Mapper
  - Reducer


Job Tracker
Master nodes
Slave nodes
Task Tracer
Mapper 1
Mapper 2
Reducer 1
Task Tracer
Mapper 1
Mapper 2
Reducer 1
Task Tracer
Mapper 1
Mapper 2
Reducer 1
Task Tracer
Mapper 1
Mapper 2
Reducer 1

Hadoop Cluster


Secondary Name Node
Name Node
Job Tracker
HBase Master
Master nodes
Slave nodes
DN
TT
Disk 1
Mapper 1
Disk 2
Mapper 2
Disk 3
Reducer 1
HBase Region Server
DN
TT
Disk 1
Mapper 1
Disk 2
Mapper 2
Disk 3
Reducer 1
HBase Region Server
DN
TT
Disk 1
Mapper 1
Disk 2
Mapper 2
Disk 3
Reducer 1
HBase Region Server

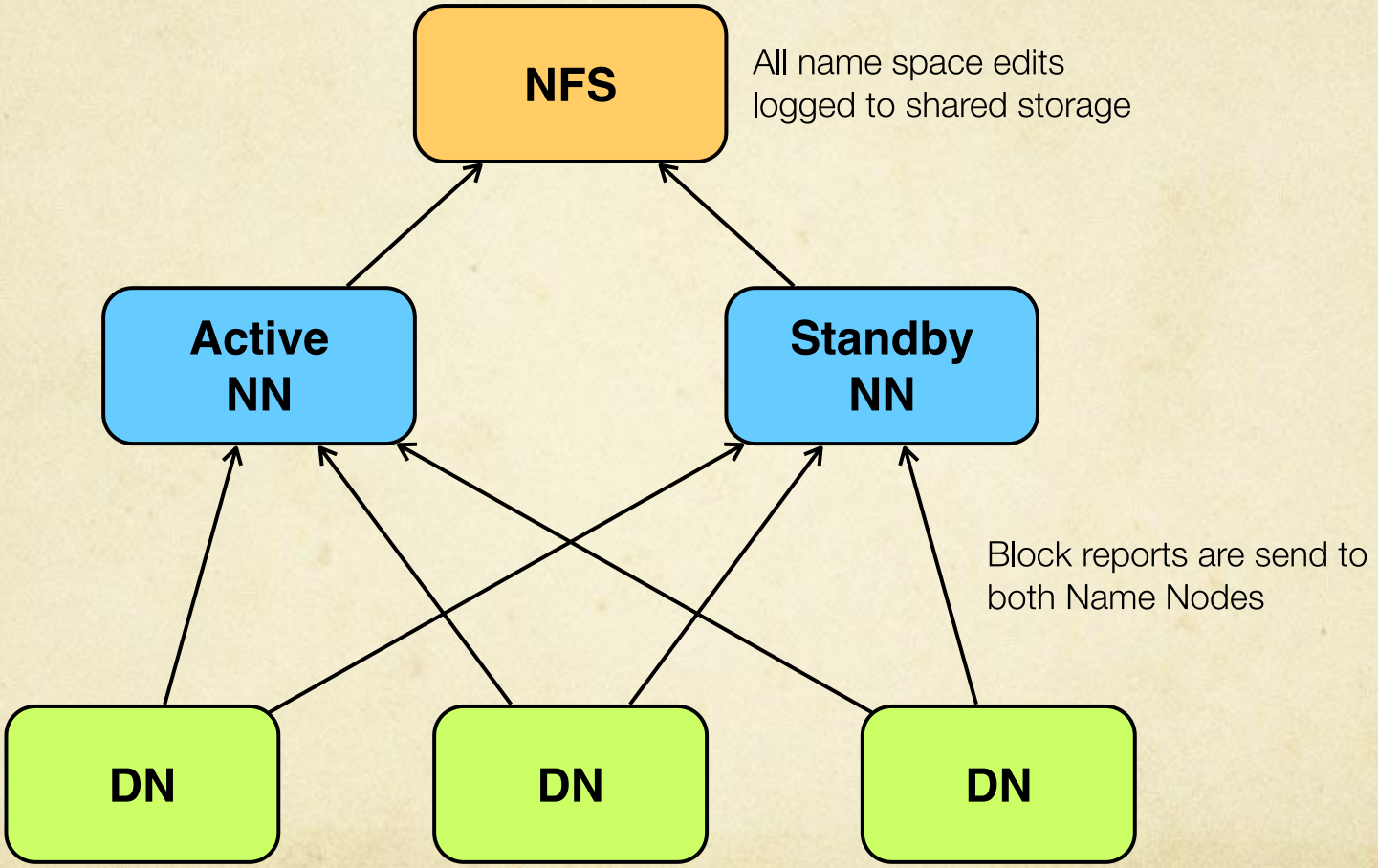
Slave Node

- Data Node
- Task Tracker
  - Mapper
  - Reducer
- HBase Region Server

High Availability for the HDFS (Name Node)

- Storage
  - RAID 1
  - Shared Storage
- Host
  - DRDB-LinuxHA
  - HDFS HA
  - Avatar Node

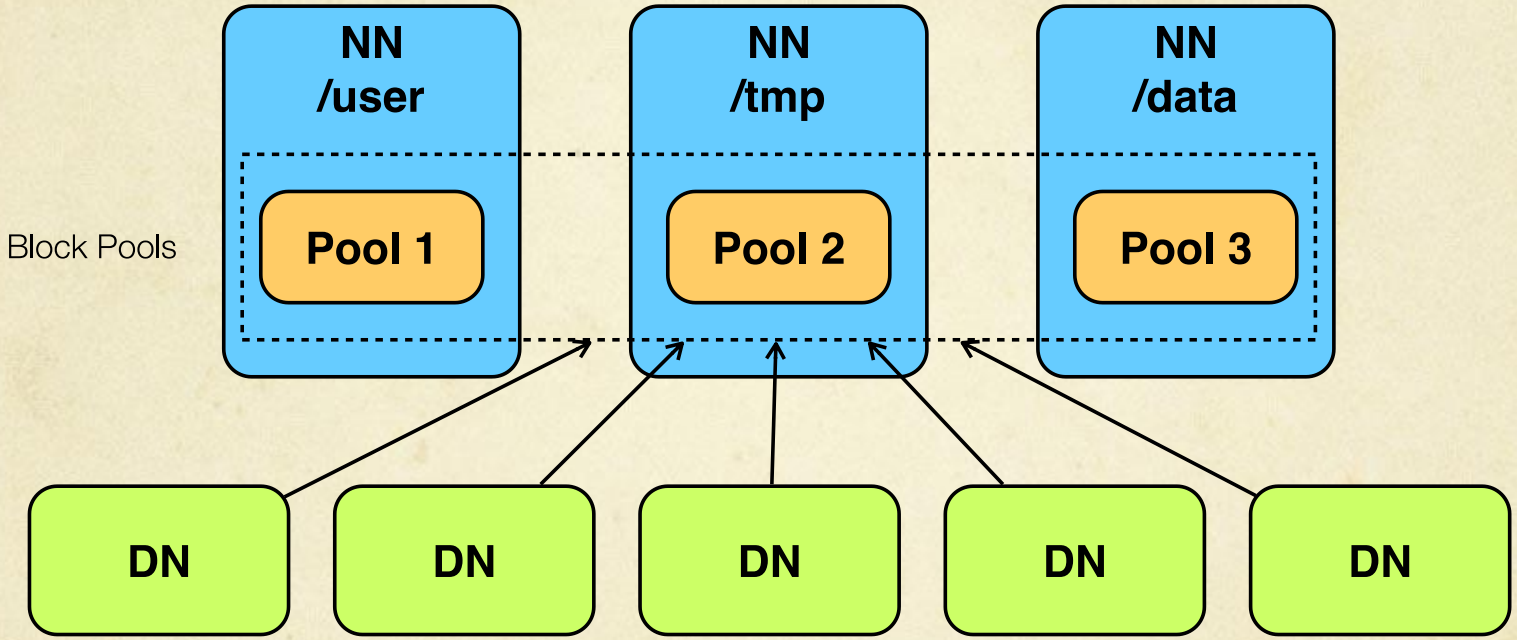
HDFS HA


NFS
All name space edits
logged to shared storage
Active
NN
Standby
NN
Block reports are send to
both Name Nodes
DN
DN
DN

Avatar Node

- Developed by Facebook
- Roles
  - Primary Avatar Node
  - Standby Avatar Node (Hot Standby)
  - Avatar Data Node

HDFS Federation

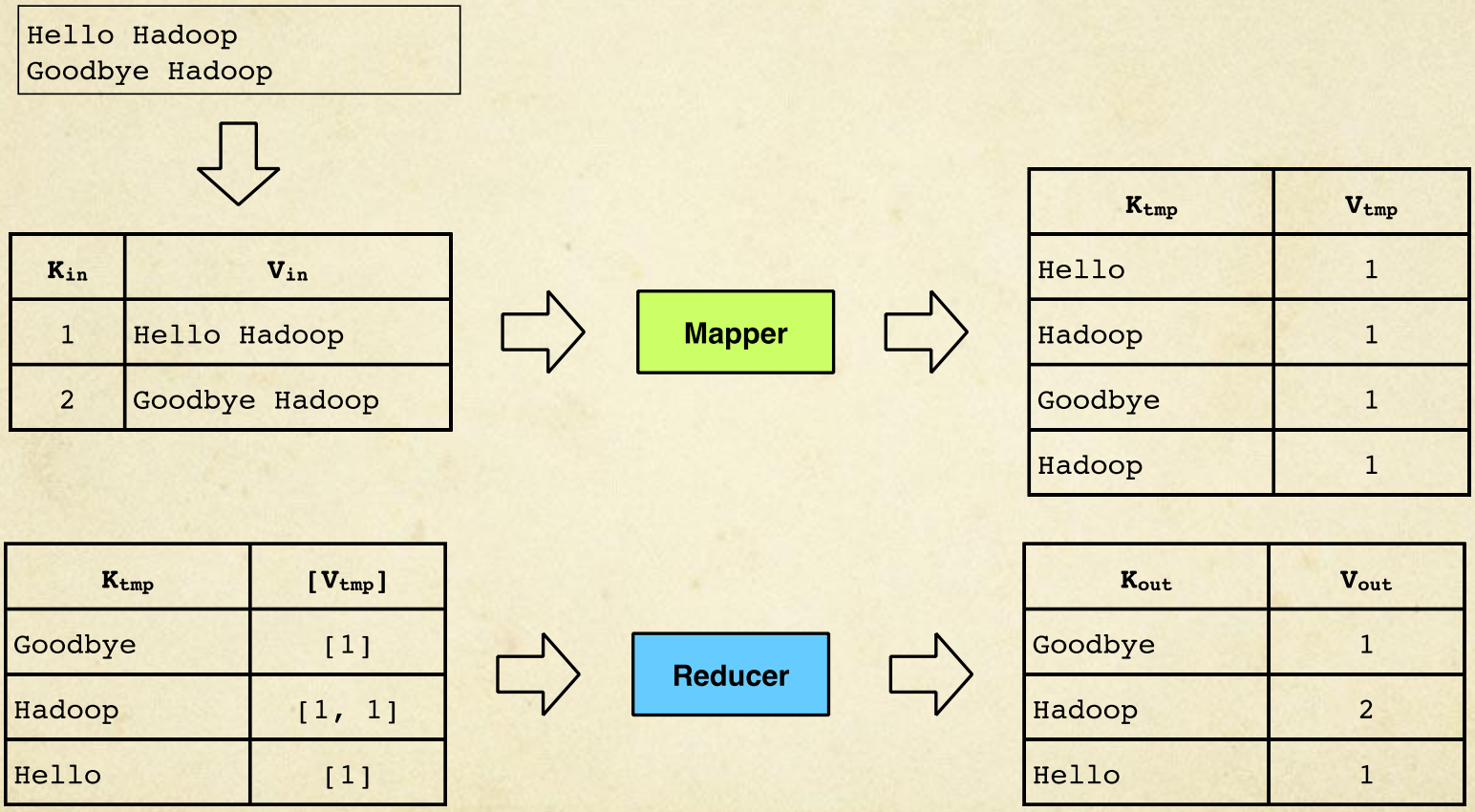


MapReduce

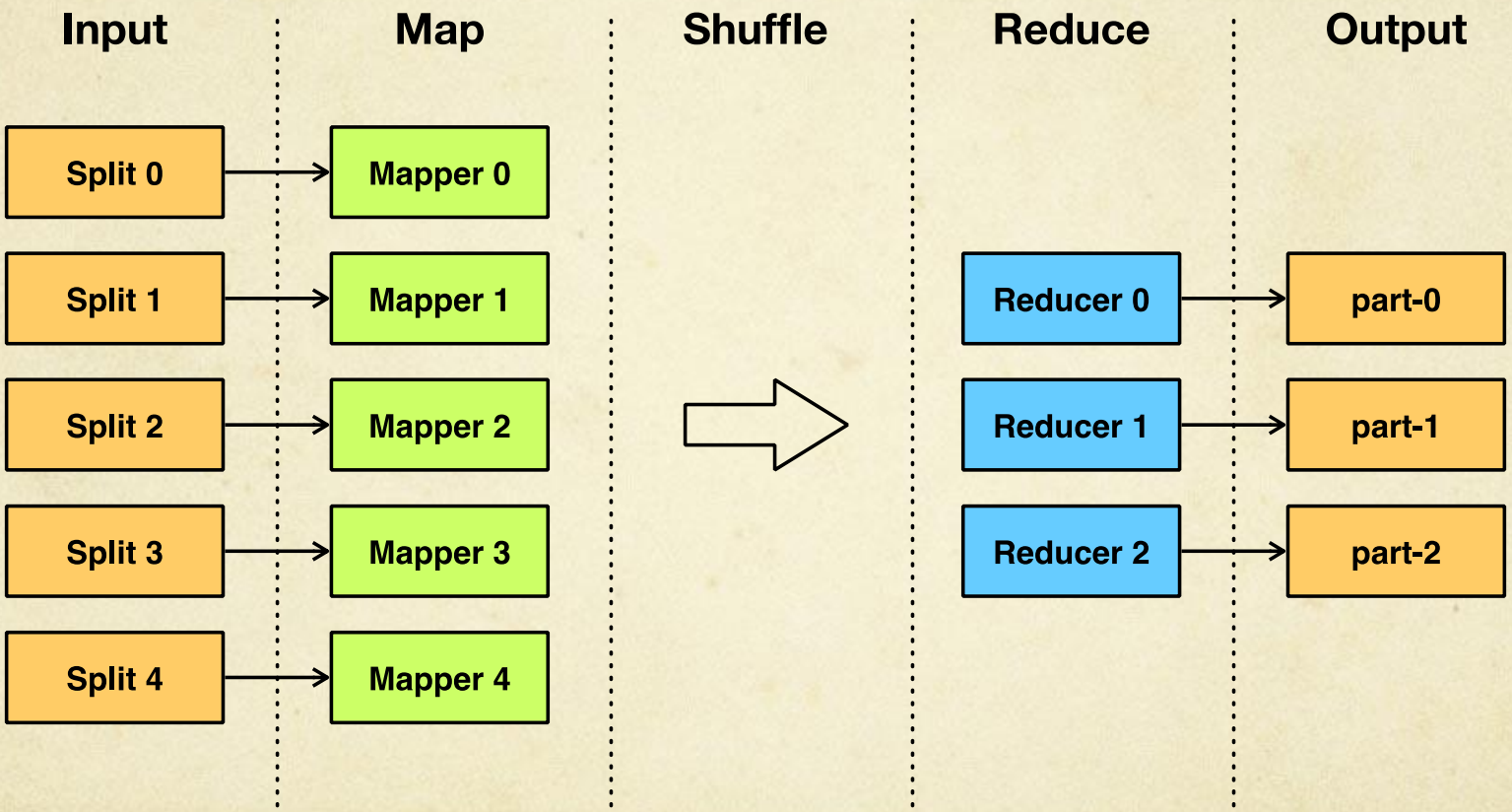
Program Model

- Mapper
  - $M(K_{in}, V_{in}) \rightarrow (K_{tmp}, V_{tmp})$
- Reducer
  - $R(K_{tmp}, [V_{tmp1}, V_{tmp2}, \ldots]) \rightarrow (K_{out}, V_{out})$

Example: Word Count

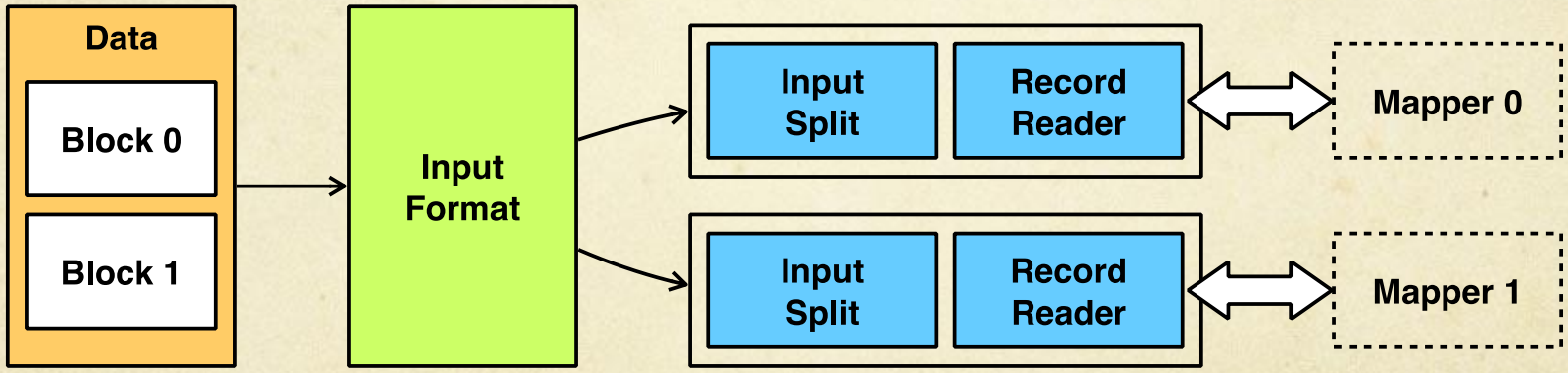


Phases


Input
Map
Shuffle
Reduce
Output
Split 0
Split 1
Split 2
Split 3
Split 4
Mapper 0
Mapper 1
Mapper 2
Mapper 3
Mapper 4
Reducer 0
Reducer 1
Reducer 2
part-0
part-1
part-2

Input Phase

- Input Format
- Input Split
- Record Reader



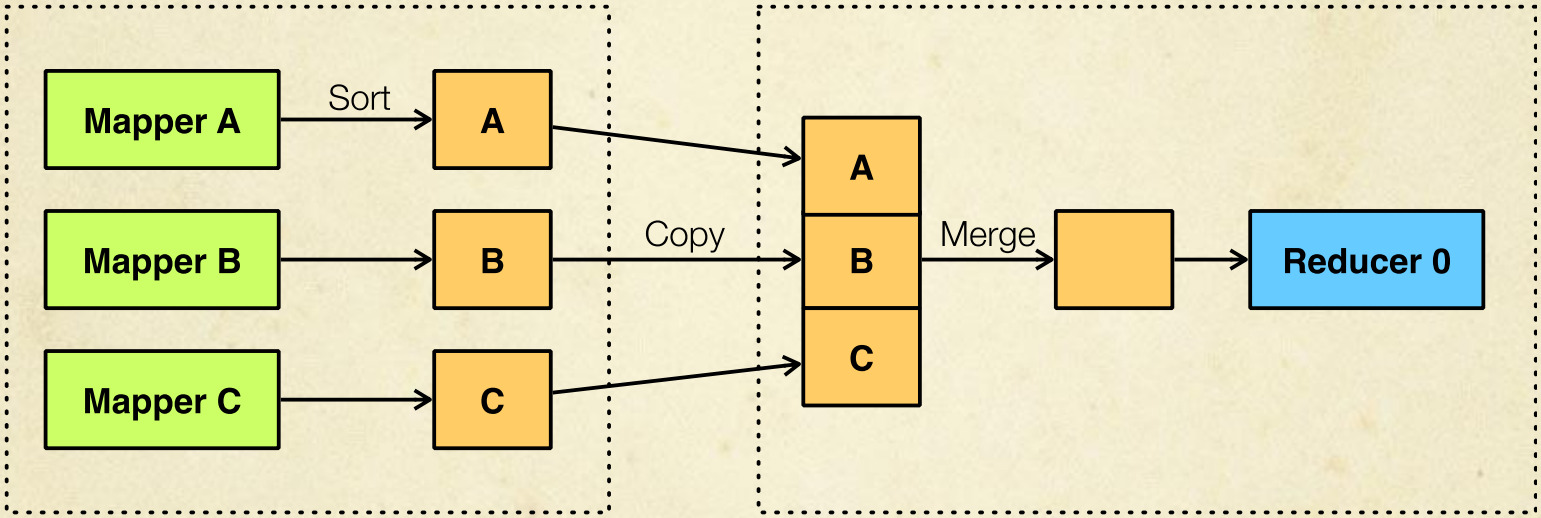
File Formats

- Text File
- Binary File
  - Sequence File
  - Map File
  - TFile
  - HAR

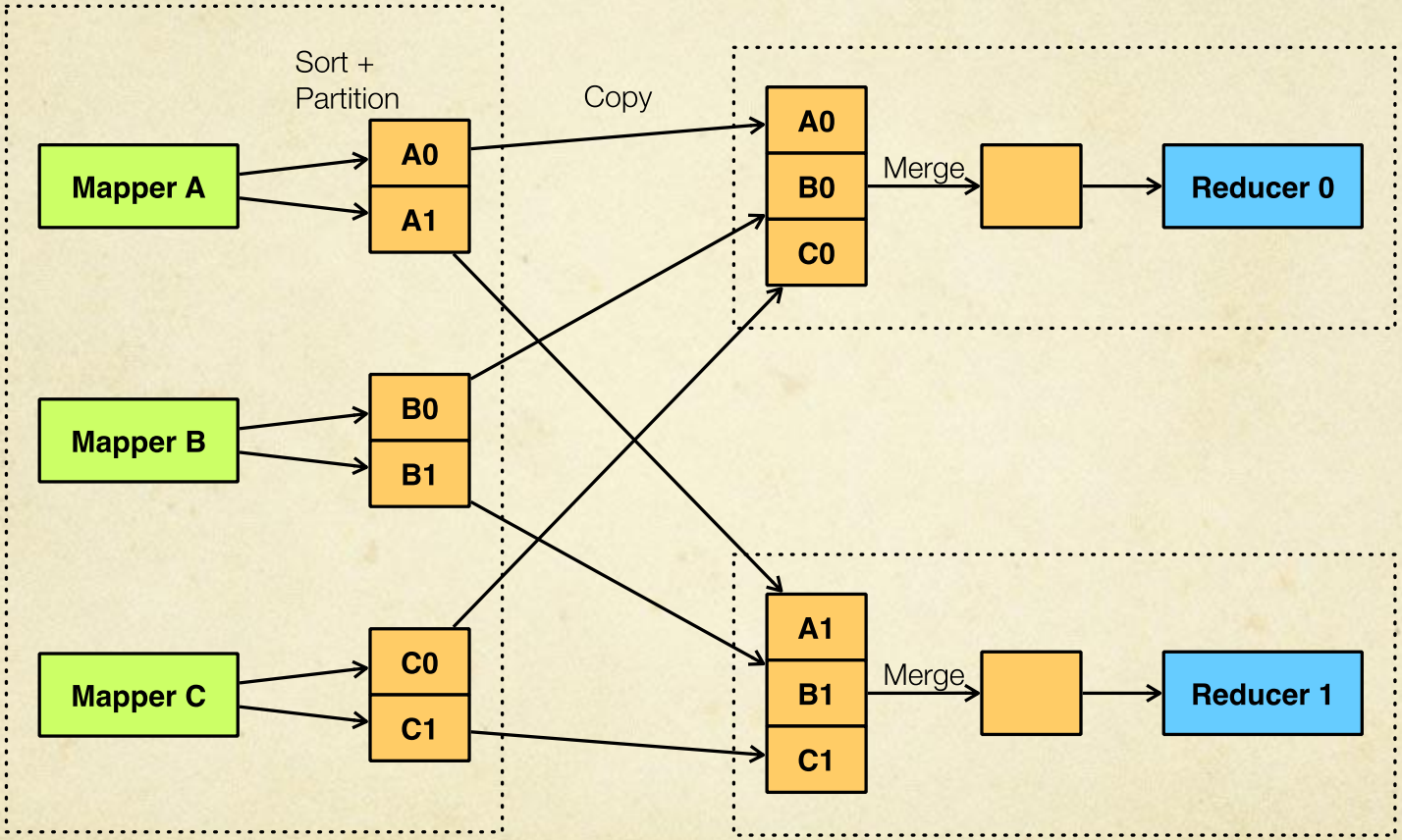
Compression and Split

- Why Compression?
  - Reduce Data Transfer Time (Disk IO, Network)
  - Save Storage
- Not all compression formats support splitting

Map, Shuffle and Reduce Phase

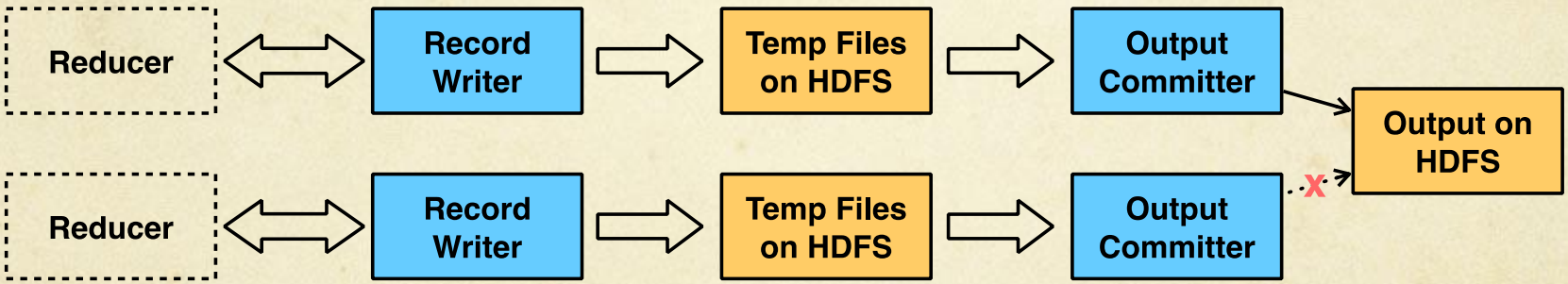


Multiple Reducers



Output Phase

- OutputFormat
- RecordWriter
- OutputCommitter



Attempts and Speculative Execution

- Attempts
  - **mapred.map.max.attempts**, default is 4
  - **mapred.reduce.max.attempts**, default 4
- Speculative Execution
  - **mapred.map.tasks.speculative.execution**, default is true
  - **mapred.reduce.tasks.speculative.execution**, default is true

- Combiner
  - Not Map-Side Reducer
  - No guarantee of how many times will be invoked
- Counters
- Number of Reducers
- Hadoop Streaming
- Hadoop Pipes

- Job Scheduler
  - Job Queue Task Scheduler (default)
  - Fair Scheduler
  - Capacity Scheduler
- MRv2 (YARN)
  - Apache Hadoop NextGen MapReduce
  - Included in hadoop-0.23, CDH4
  - Not stable yet