系统设计第五课

小憩一下，马上开课

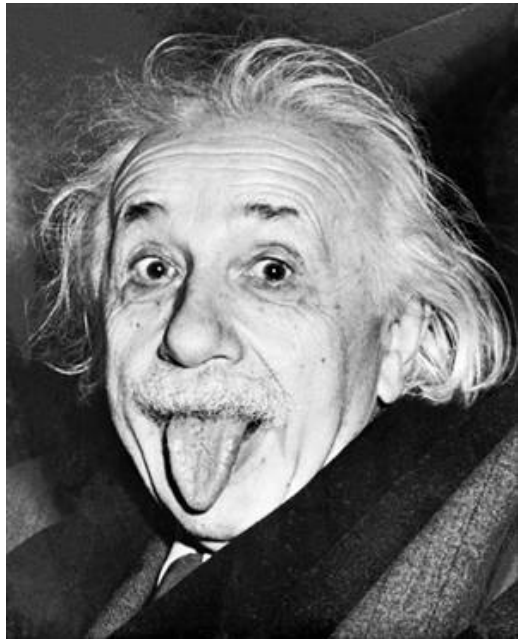
关注微信/微博，获取最新面试题及权威解答

微信: ninechapter

微博: http://www.weibo.com/ninechapter

官网: www.jiuzhang.com

Copyright © www.jiuzhang.com



System_Design_5

Distributed system

张无忌

2016-02-06

V4.00

九章算法

Copyright © www.jiuzhang.com

After class_5, you can answer

- Design distributed file system(GFS) and database(BigTable)?
  - Google, DrawBridge, Yelp
- Calculate word appearance/inverted index/anagrams with MapReduce
  - Google, Twitter, Drawbridge, Zenefits, Bloomberg, Genapsys, Liveramp, …

Copyright © www.jiuzhang.com

Interviewer: design search engine

Copyright © www.jiuzhang.com

Three Musketeers of Google

MapReduce
（名人）

BigTable
（？？）

GFS
（高富帅）

Read More

Novice, http://url.cn/SVDkio
Novice, http://url.cn/deN2uC
Novice, http://url.cn/Vf6ABZ
Expert/Master, http://url.cn/EZNxGe
Expert/Master, http://url.cn/dxcRxI

Copyright © www.jiuzhang.com

Do not recite the details of papers,

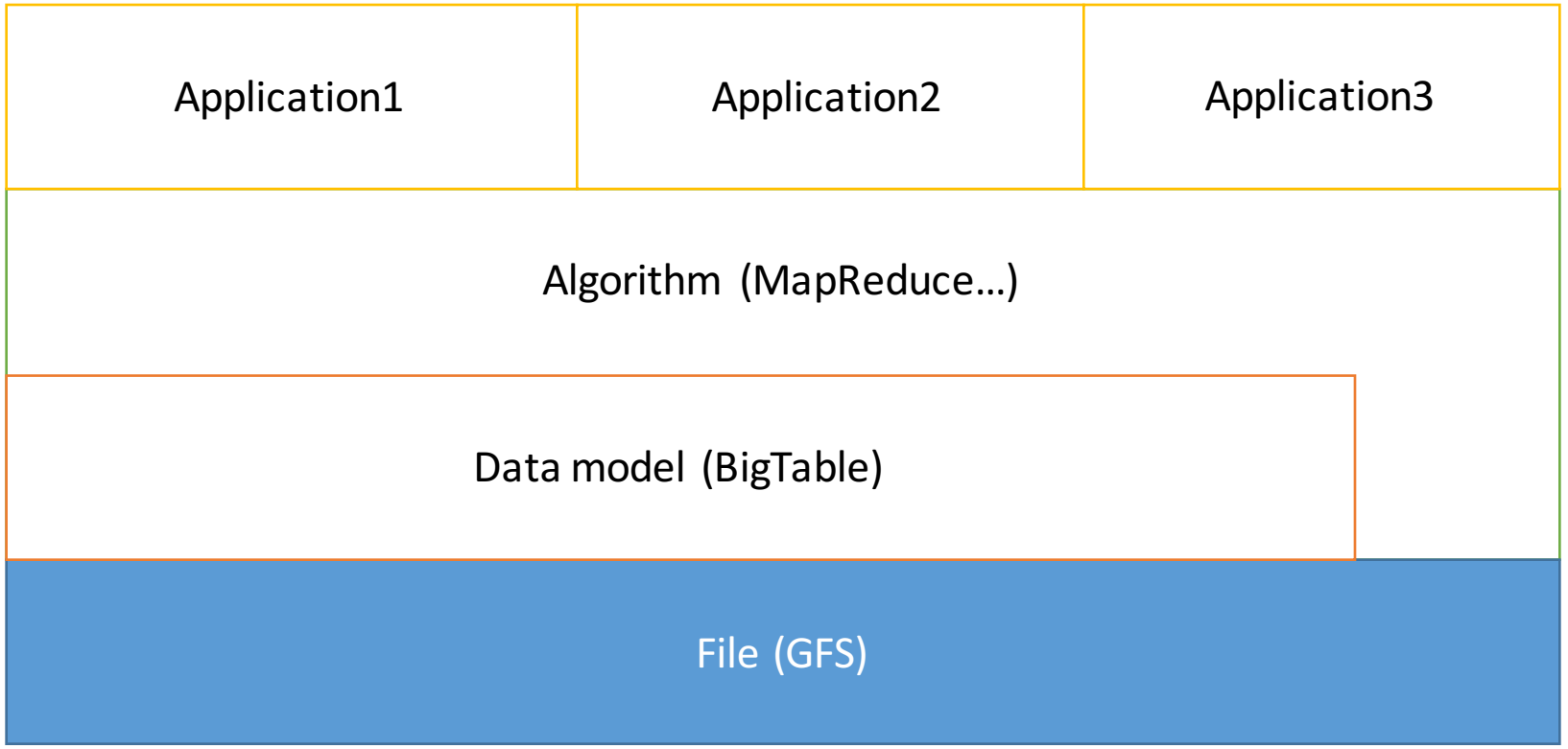
but find a suitable solution under your scenario.

Copyright © www.jiuzhang.com

Interviewer: What is the layers of your search engine?

Copyright © www.jiuzhang.com

Layers of system



Let’s design with a bottom-up method

Copyright © www.jiuzhang.com

Google File System

Read More

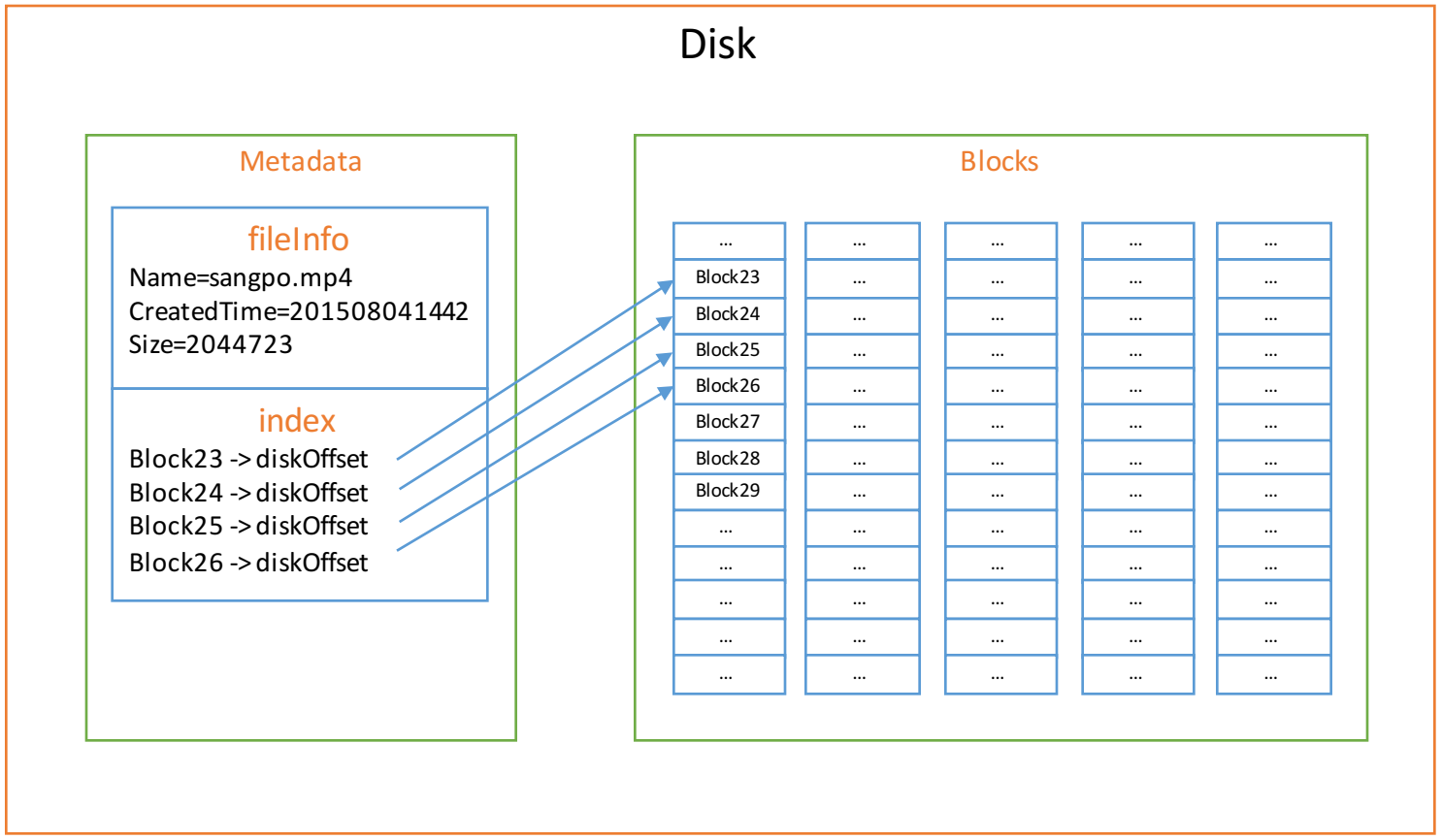
Expert/Master, http://url.cn/dOLFCs
Expert/Master, http://url.cn/eErkhm
Expert/Master, http://url.cn/LqTkoa

Copyright © www.jiuzhang.com

Interviewer: How to save a file?

Copyright © www.jiuzhang.com

How to save a file?



Key point

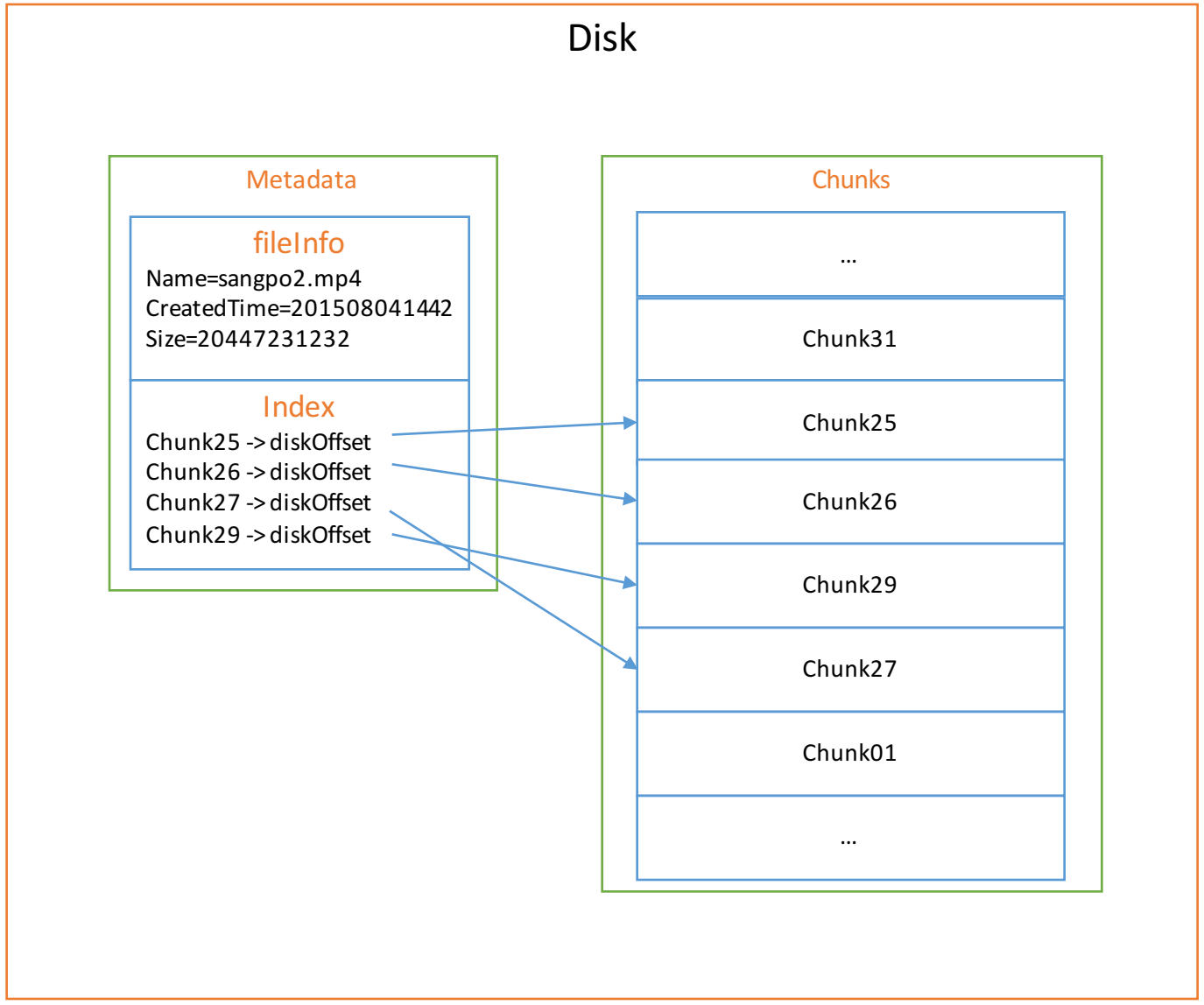
- 1 block = 1024Byte

Copyright © www.jiuzhang.com

Interviewer: How to save a large file?

Copyright © www.jiuzhang.com

How to save a large file?



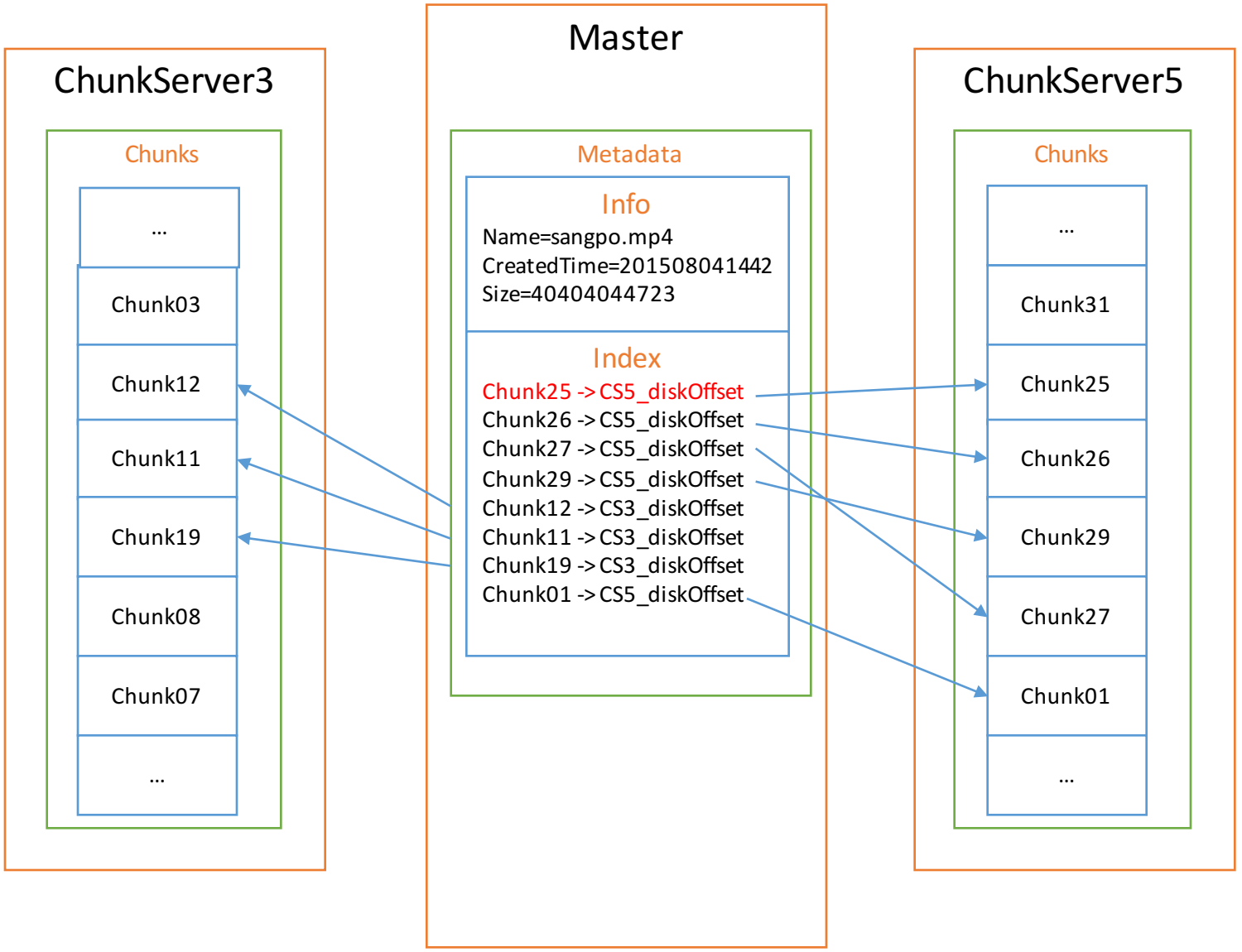
- Key point
  - 1 chunk = 64MB
    = 64*1024
    = 65,536 blocks
- Advantage
  - Reduce size of metadata
  - Reduce traffic
- Disadvantage
  - Waste space for small files

Copyright © www.jiuzhang.com

Interviewer: How to save an extra-large file?

Copyright © www.jiuzhang.com

How to save an extra-large file?



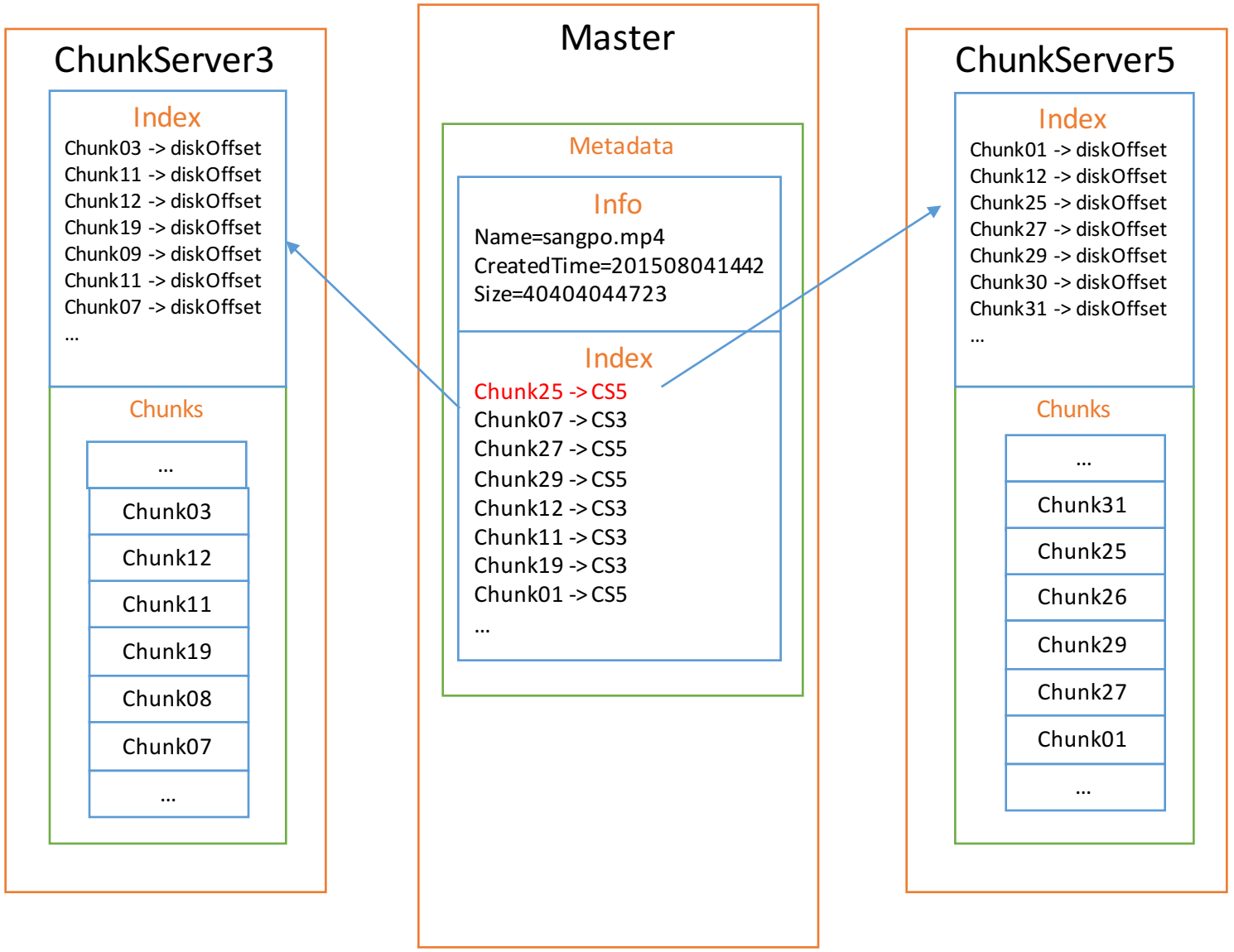
- Key point
  - One master + many ChunkServers
- Disadvantage
  - Any change of the diskOffset of a chunk in a ChunkServer need to notify the Master

Copyright © www.jiuzhang.com

Interviewer: How to reduce traffic and storage of master?

Copyright © www.jiuzhang.com

How to reduce traffic and storage of master?



- Key point
  - The master don't record the diskOffset of a chunk

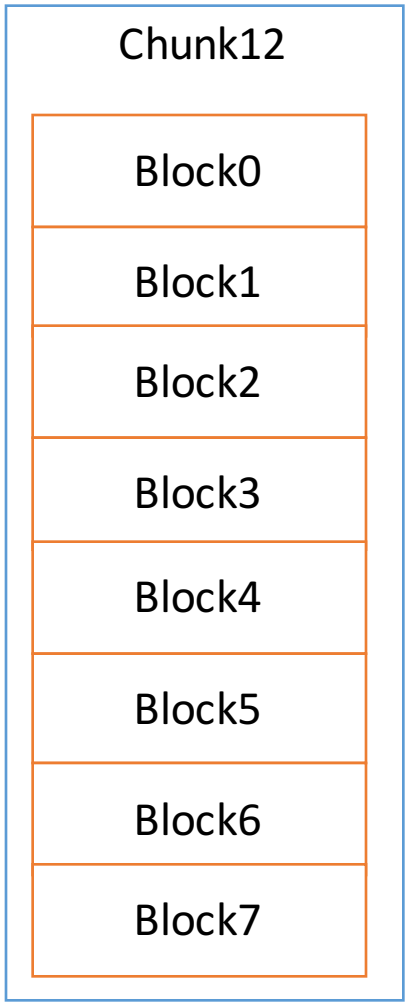
- Advantage
  - Reduce the size of metadata in master
  - Reduce the traffic between master and ChunkServer

Copyright © www.jiuzhang.com

Interviewer: How to identify whether a chunk on the disk is broken?

Copyright © www.jiuzhang.com

How to identify whether a chunk on the disk is broken?



metadata
Block0 -> checksum
Block1 -> checksum
Block2 -> checksum
...

- Key point
  - 1 chunk = a list of blocks
  - 1 block = 64KB
  - Each block has a checksum
  - 1 checksum = 32bit
  - The size of checksum of 1T file
    - 1T/64KB*32bit = 64MB

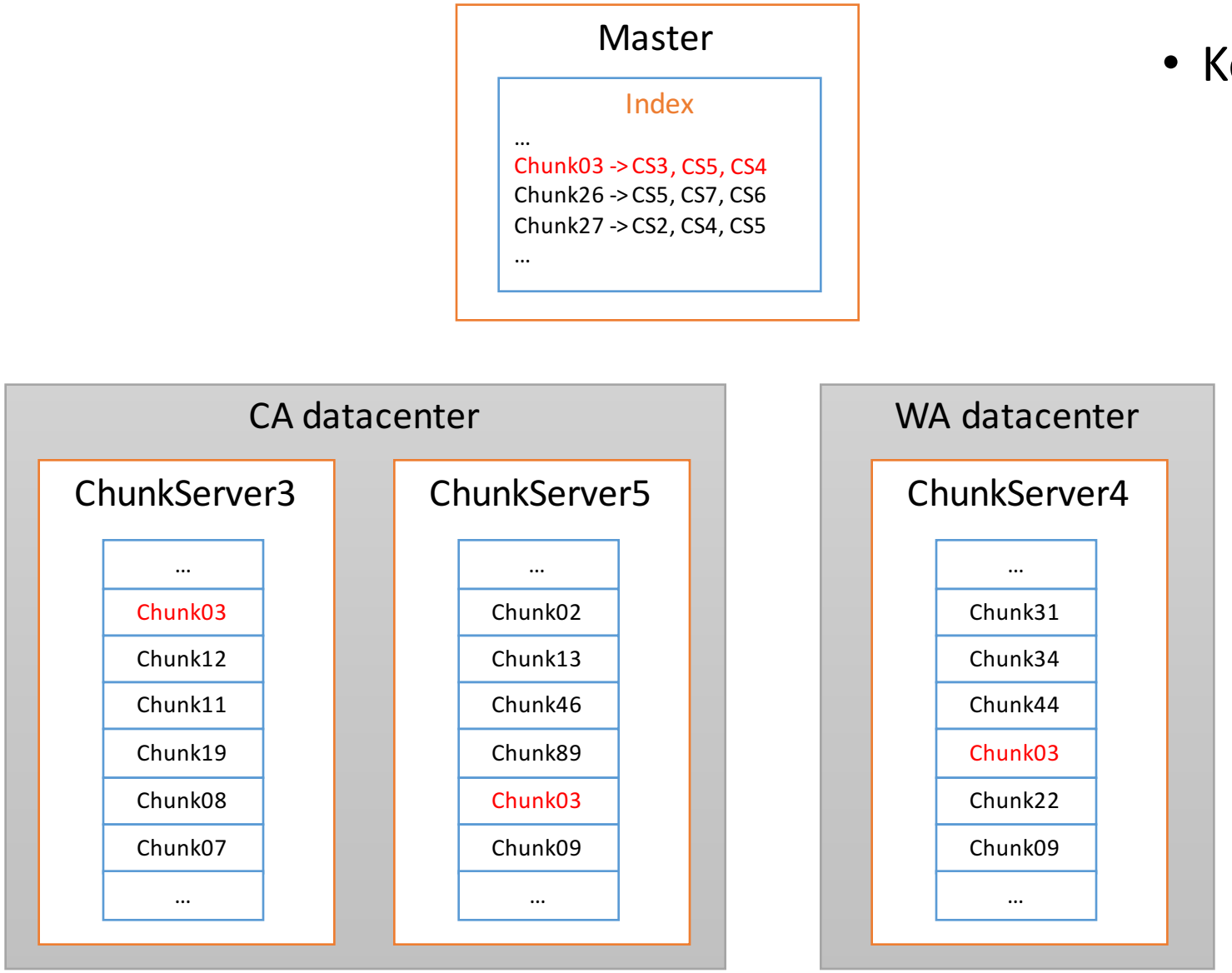
- It will compare its checksum when it reads a block

Copyright © www.jiuzhang.com

Interviewer: How to avoid data loss when a ChunkServer is down/fail?

Copyright © www.jiuzhang.com

How to avoid the loss of data when a ChunkServer is down/fail?



- Key point
  - Replicate the Chunks
  - How many replications?
    - 3
  - How to select ChunkServers of a chunk?
    - Server with below-average disk space utilization
    - Limited number of “recent” creation
    - Across racks: 2+1

Read More

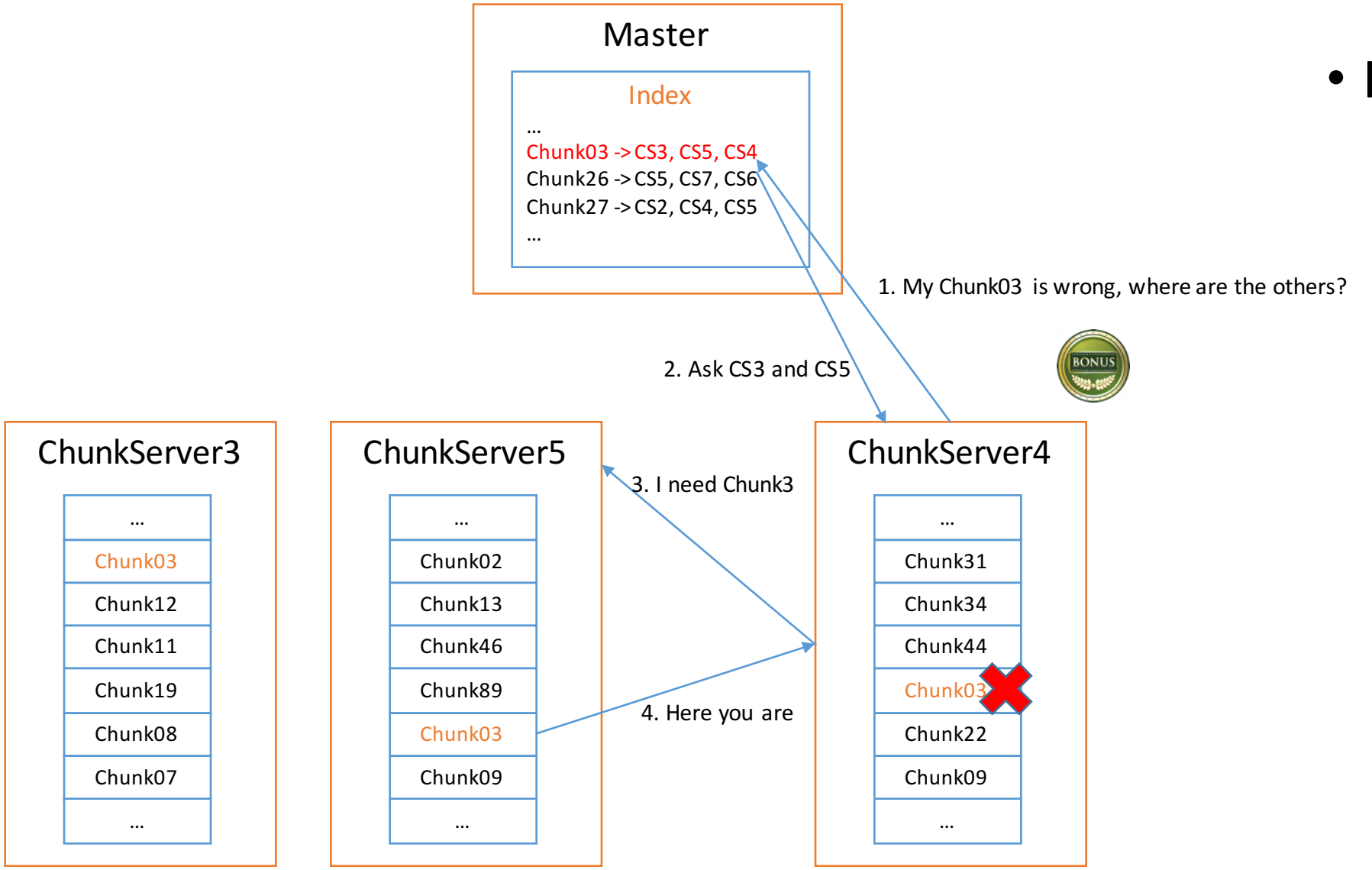
Expert/Master, http://url.cn/d1BrZU

Copyright © www.jiuzhang.com

Interviewer: How to recover when a chunk is broken?

Copyright © www.jiuzhang.com

How to recover when a chunk is broken?



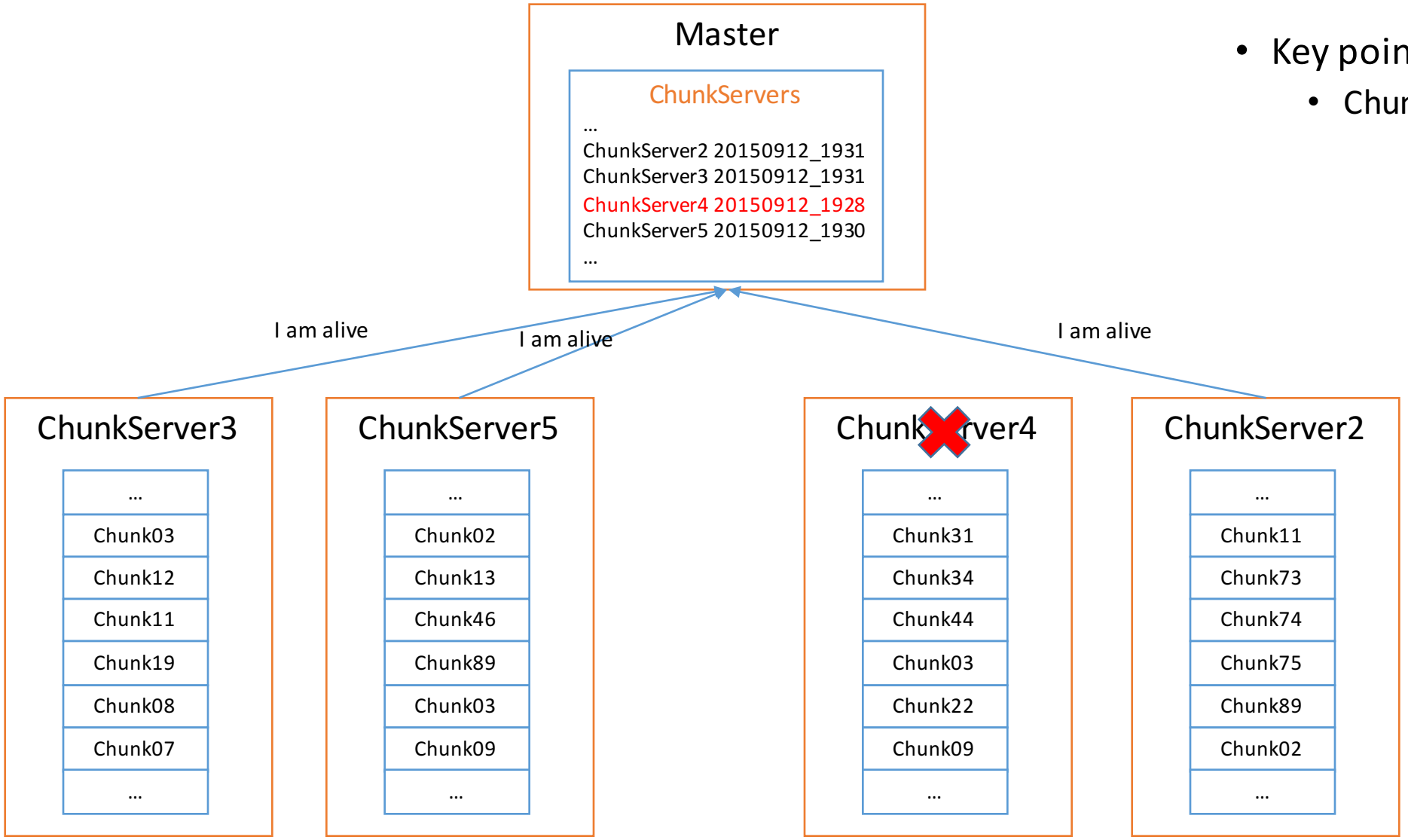
- Key point
  - Ask master for help

Copyright © www.jiuzhang.com

Interviewer: How to find whether a ChunkServer is down?

Copyright © www.jiuzhang.com

How to find whether a ChunkServer is down?



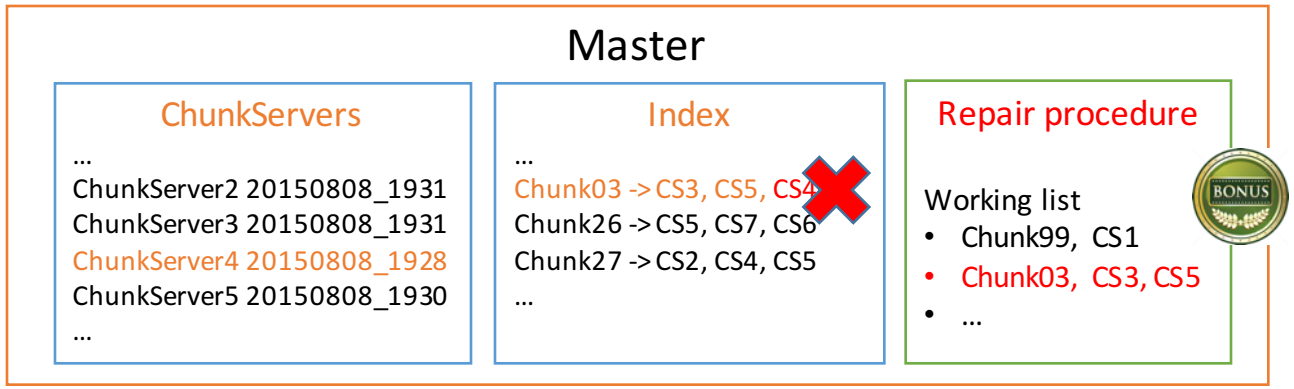
- Key point
  - ChunkServer sends heartbeat

Copyright © www.jiuzhang.com

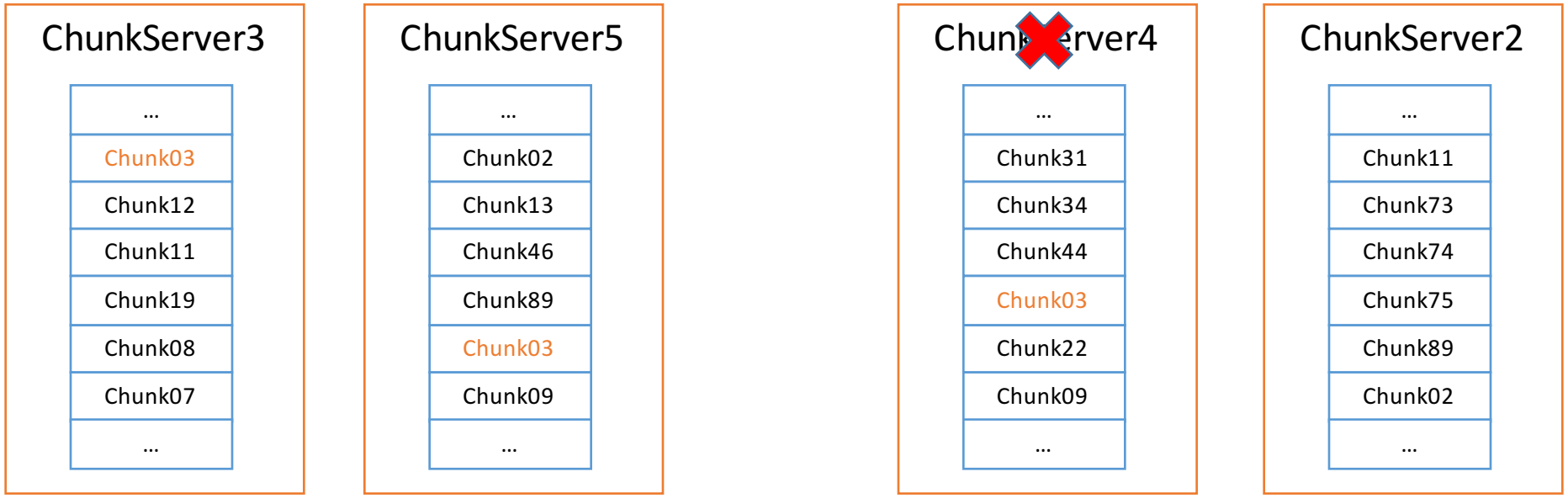
Interviewer: How to recover the data when a ChunkServer is down?

Copyright © www.jiuzhang.com

How to recover data when a ChunkServer is down?



- Key point
  - Repair priority is based on the number of replications



Copyright © www.jiuzhang.com

Interviewer: How to avoid hot spot?

Copyright © www.jiuzhang.com

How to avoid hot spot?

Master

Rebalance procedure

Chunk stats
- Chunk00, access frequency = 100
- Chunk01, access frequency = 50
- ...

Server stats
- ChunkServer01, free space = 21%, free bandwidth = 15%
- ChunkServer02, free space = 50%, free bandwidth = 21%
- ...

- Key point
  - Replicate a chunk into more replications when it is busy
  - Fill the ChunkServer with more space and more bandwidth

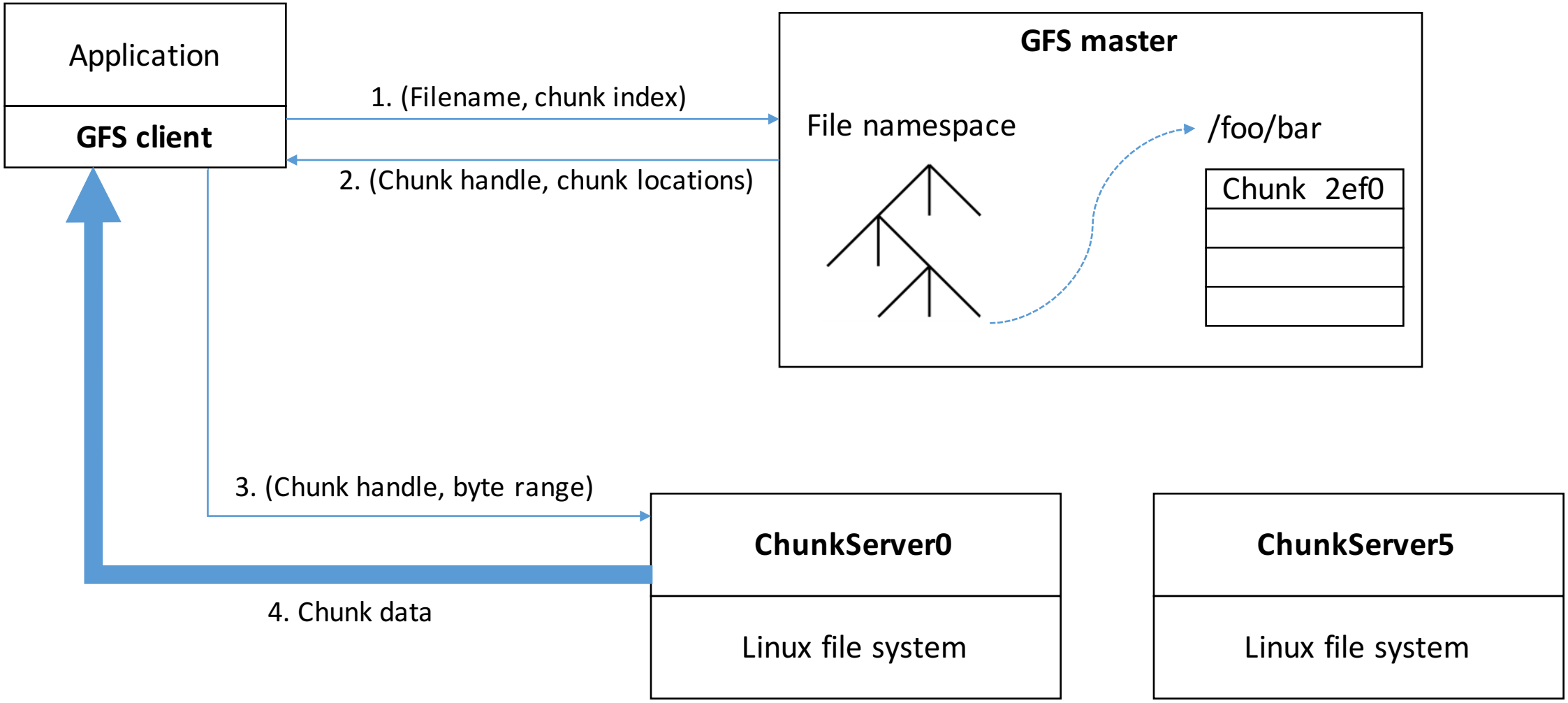
ChunkServer3	ChunkServer5	ChunkServer7	ChunkServer2
...	...	...	...
Chunk03	Chunk02	Chunk31	Chunk11
Chunk12	Chunk13	Chunk34	Chunk73
Chunk11	Chunk46	Chunk44	Chunk74
Chunk19	Chunk89	Chunk03	Chunk75
Chunk08	Chunk03	Chunk22	Chunk89
Chunk07	Chunk09	Chunk09	Chunk02
...	...	...	...

Copyright © www.jiuzhang.com

Interviewer: How to read from a file?

Copyright © www.jiuzhang.com

How to read from a file?

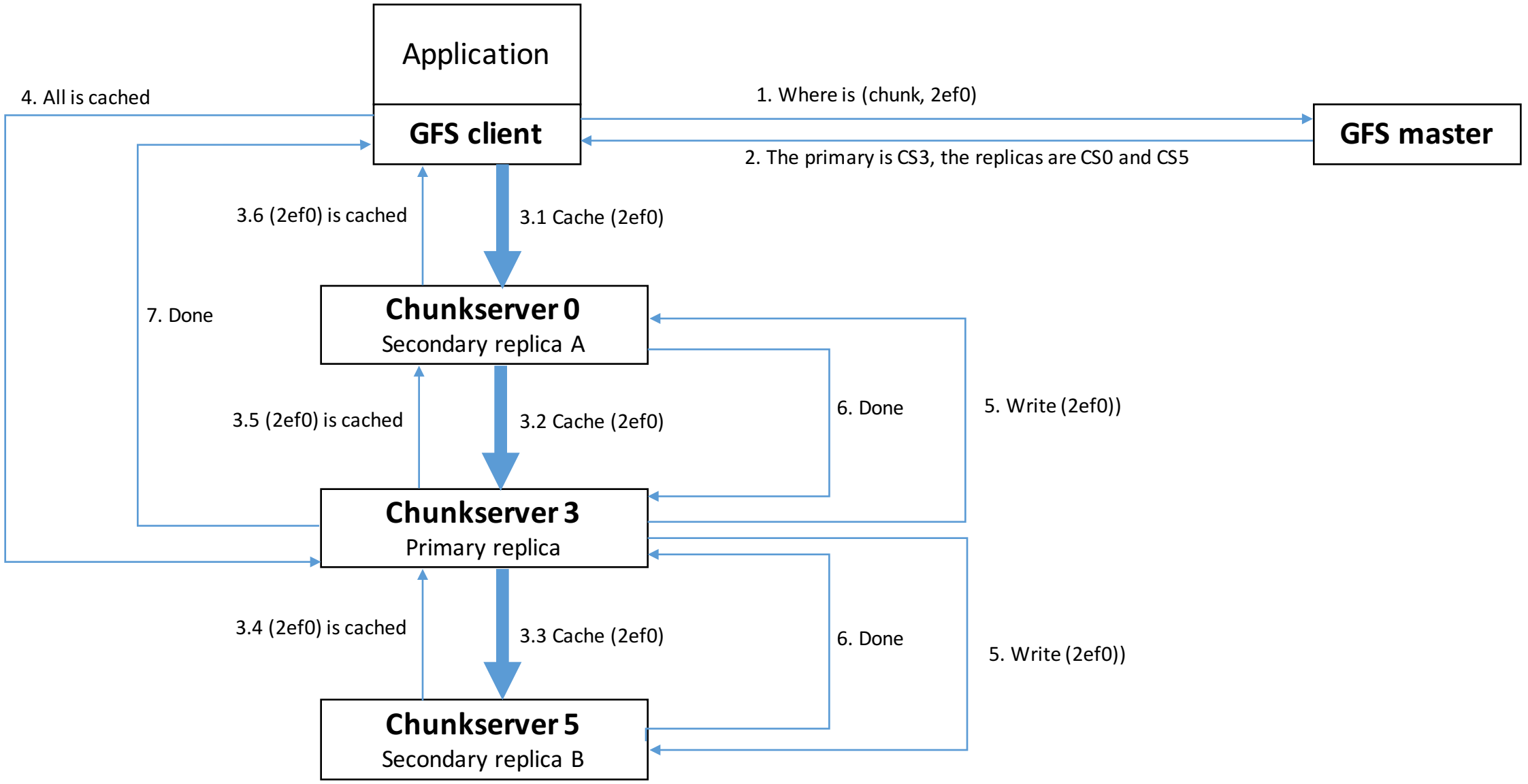


九章算法
Copyright © www.jiuzhang.com

Interviewer: How to write into a file?

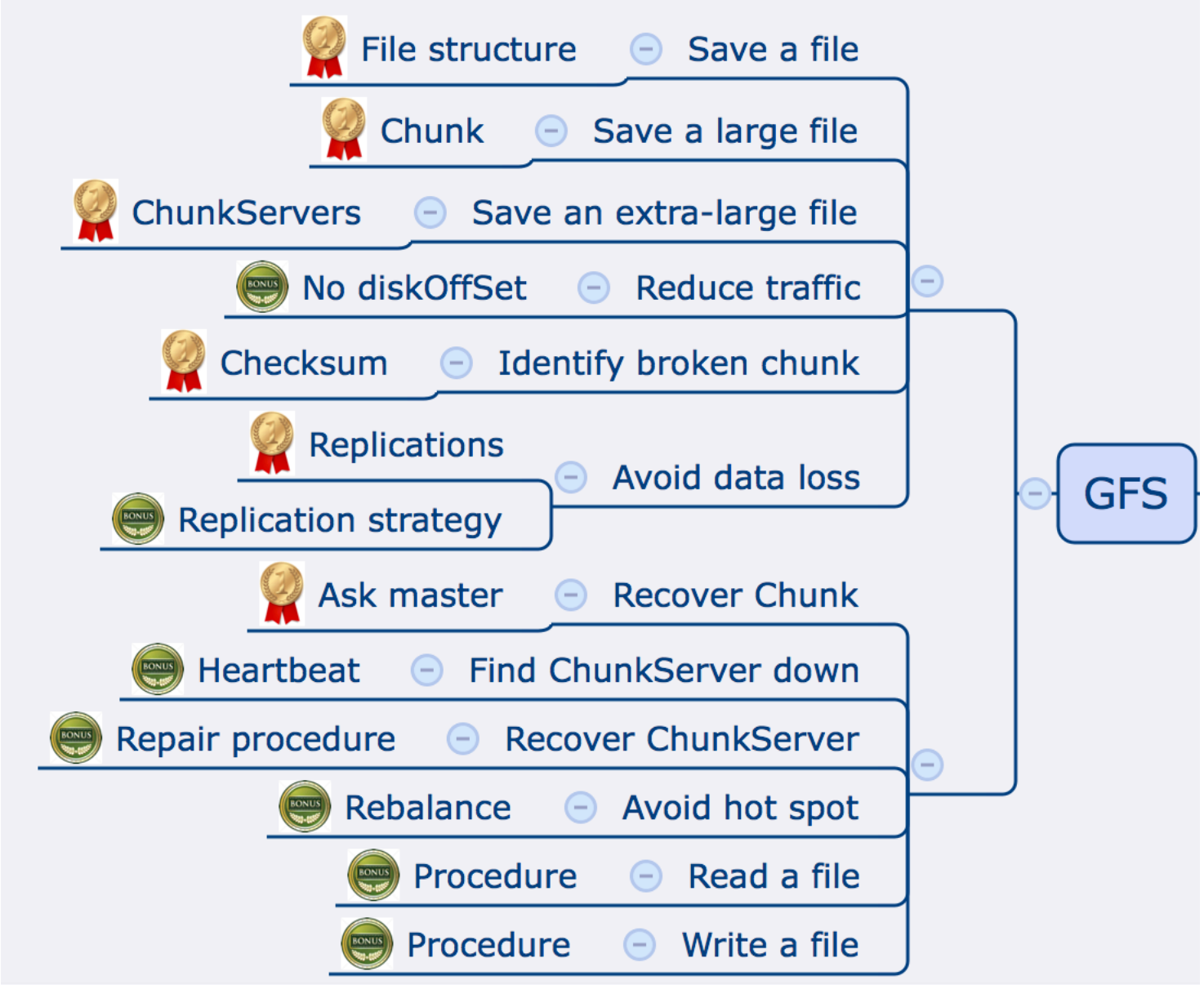
Copyright © www.jiuzhang.com

How to write into a file?



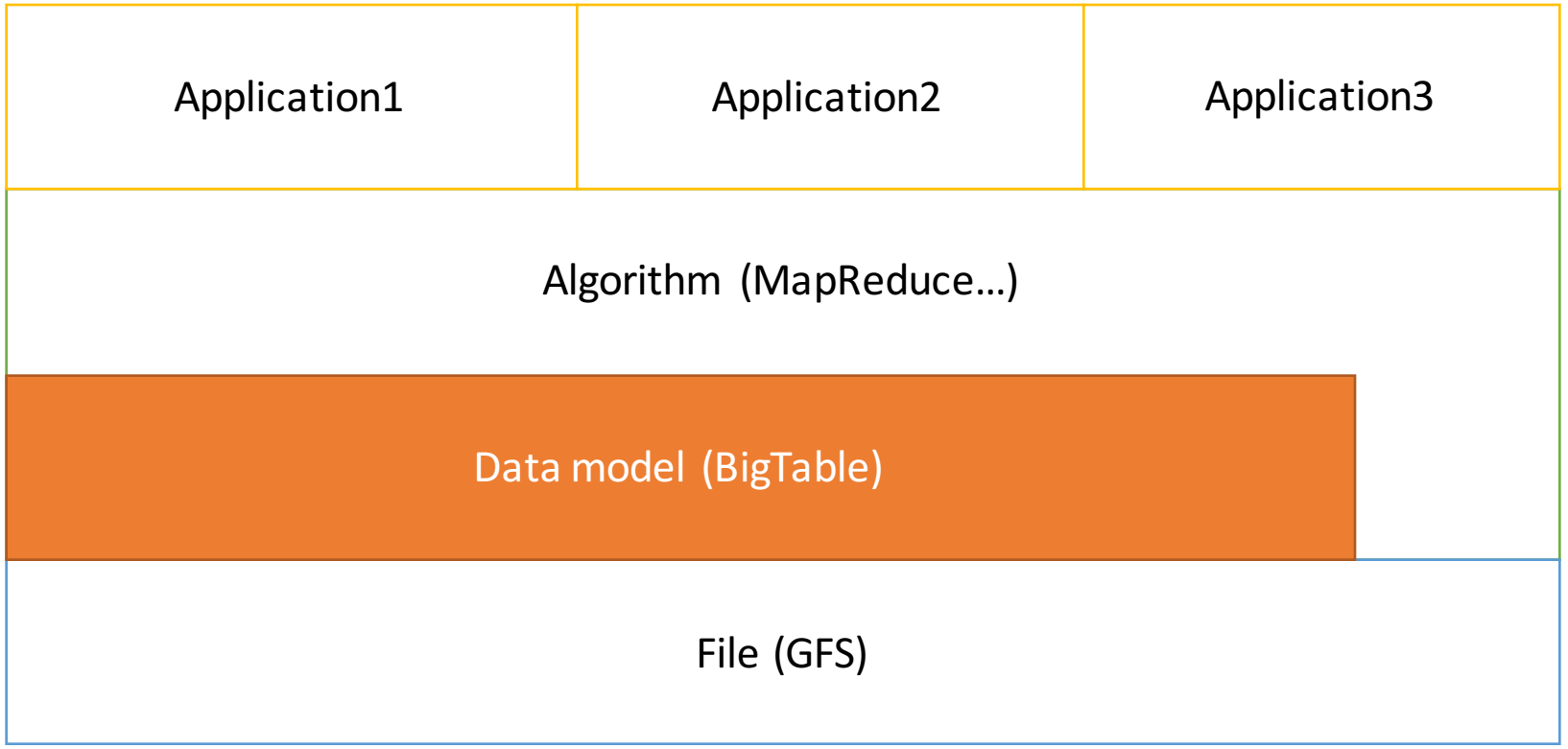
九章算法
Copyright © www.jiuzhang.com

GFS Summary (6+7)



Copyright © www.jiuzhang.com

Layers of system



Copyright © www.jiuzhang.com


我又不是因為想
得到你們的承認
才當英雄的
是因為
我想當才當的

Copyright © www.jiuzhang.com

BigTable

Read More

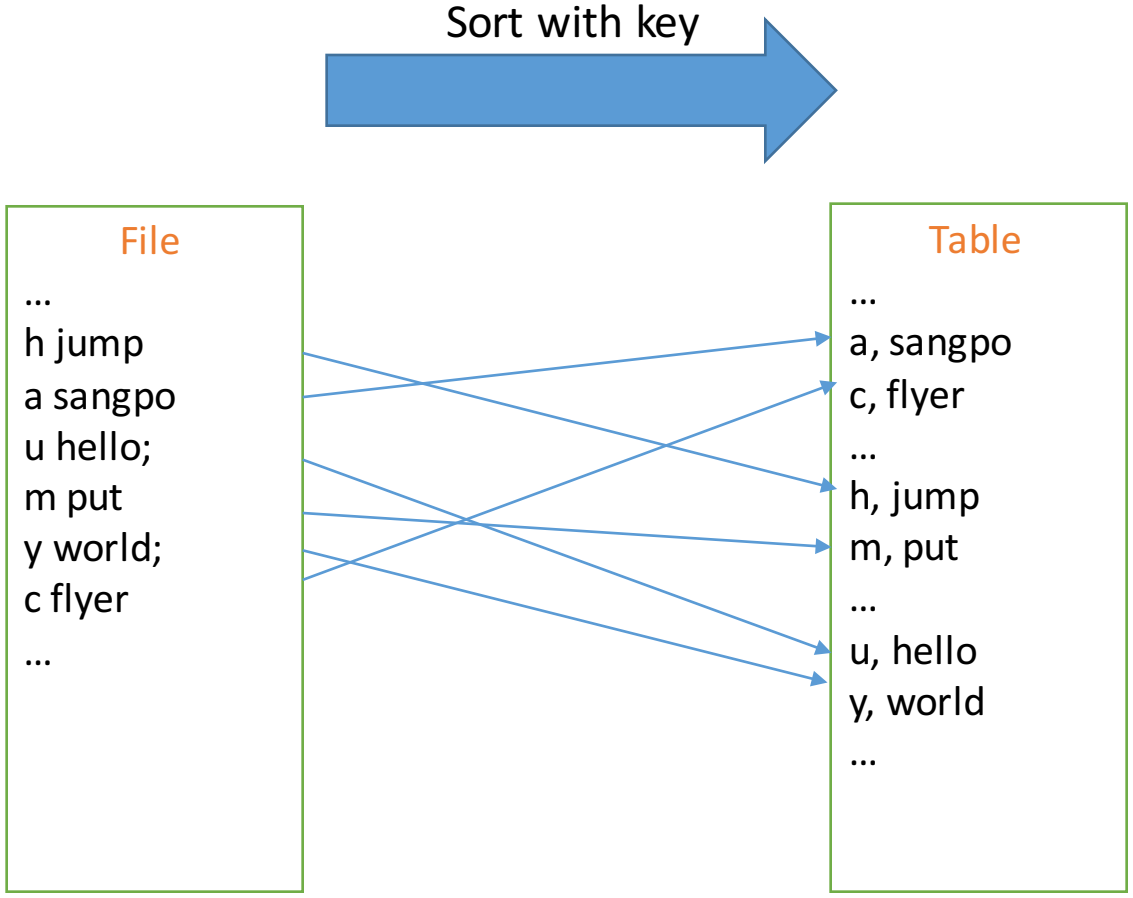
Expert/Master, http://url.cn/eJjmEp
Expert/Master, http://url.cn/VRwOSW
Expert/Master, http://url.cn/W6Xua6
Expert/Master, http://url.cn/2aslLW

Copyright © www.jiuzhang.com

Interviewer: How to support lookup and range query on a file?

Copyright © www.jiuzhang.com

How to support lookup and range query on a file?



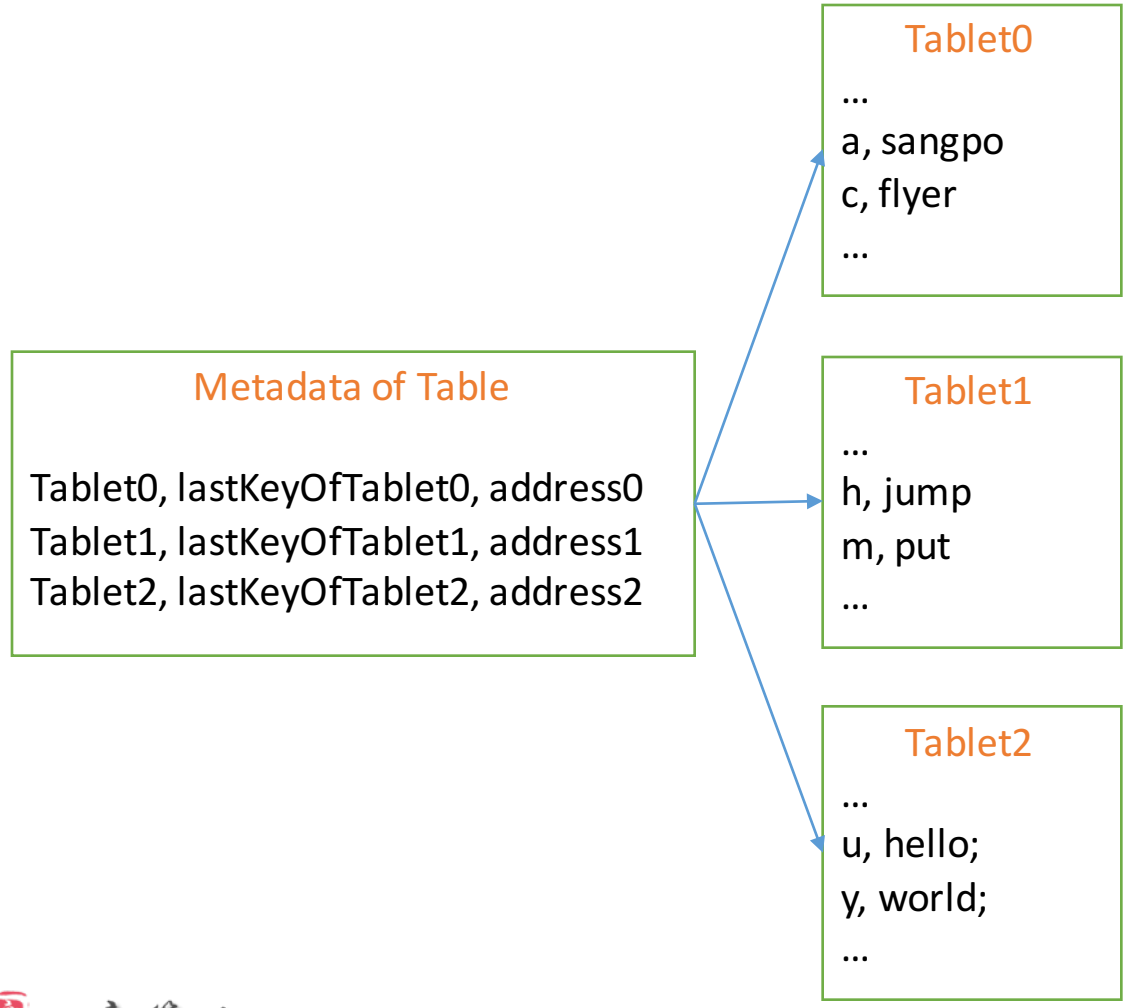
- Key point
  - Table = A list of sorted <key, value>

九章算法
Copyright © www.jiuzhang.com

Interviewer: How to save a large table?

Copyright © www.jiuzhang.com

How to save a large table?



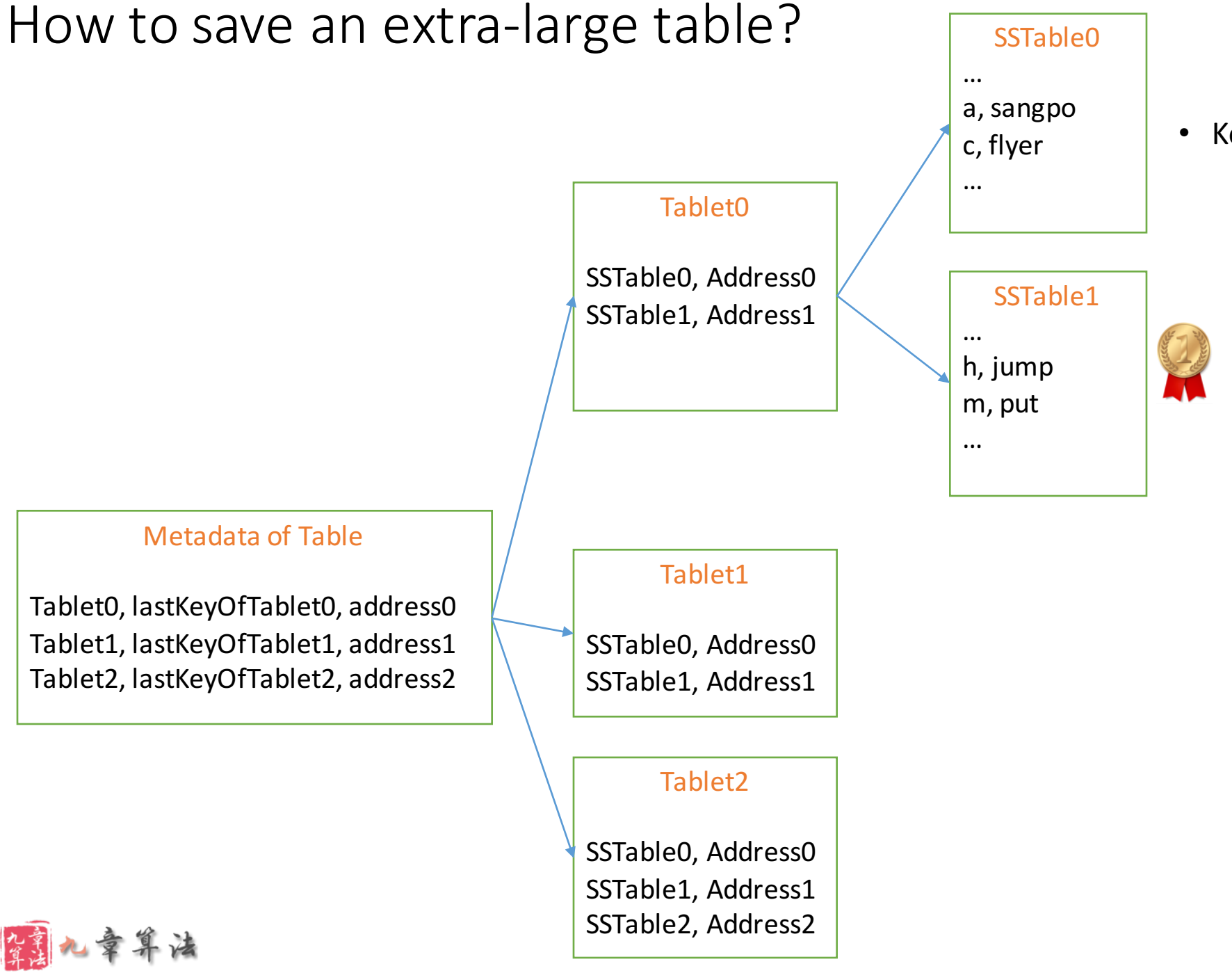
- Key point
  - A table = a list of tablets
  - A tablet = a list of sorted <key value>

Copyright © www.jiuzhang.com

Interviewer: How to save an extra-large table?

Copyright © www.jiuzhang.com

How to save an extra-large table?



- Key point
  - A table = a list of tablets
  - A tablet = a list of SSTables
  - A SSTable = a list of sorted <key, value>

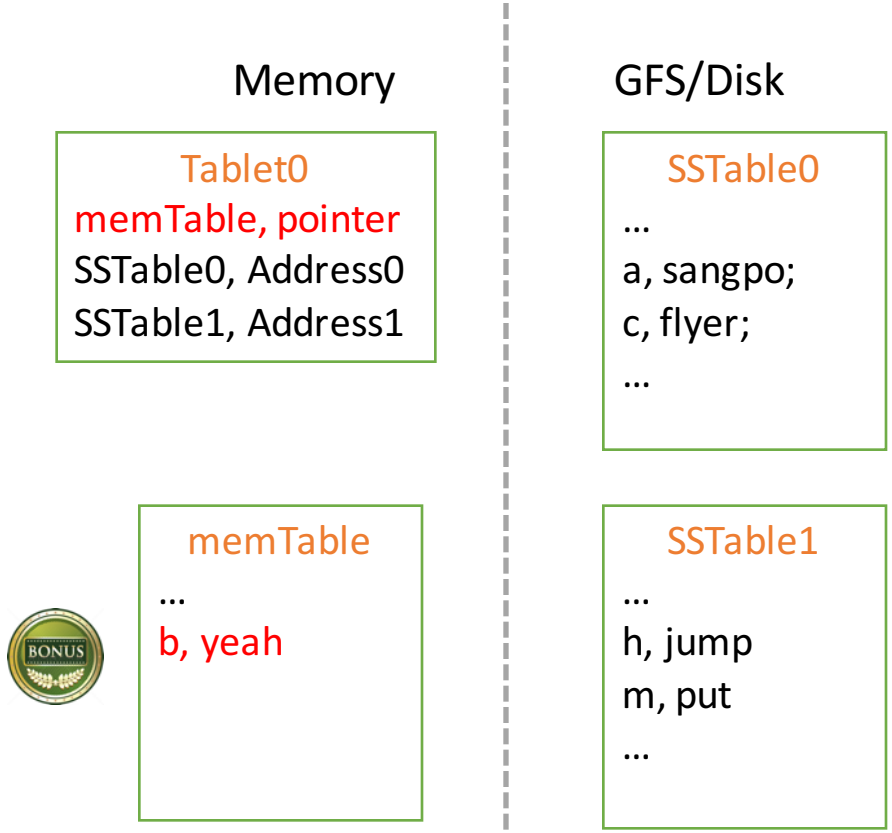
Copyright © www.jiuzhang.com

Interviewer: How to write into a table?

Copyright © www.jiuzhang.com

How to write into a table?

Task: add <b, yeah>



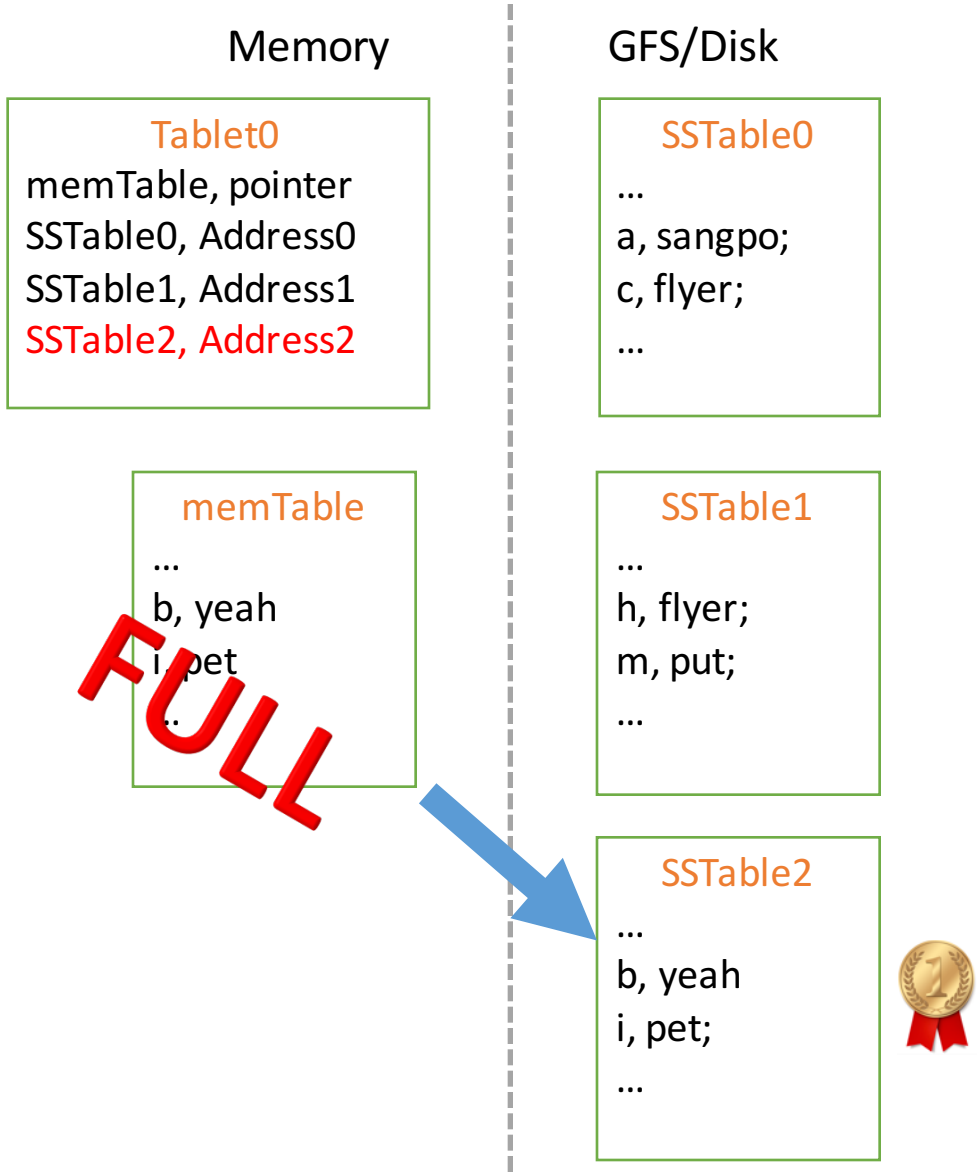
- Key point
  - Write into a memTable to increase speed
  - A tablet = memTable + a list of SSTables

Copyright © www.jiuzhang.com

Interviewer: What if memTable is too large?

Copyright © www.jiuzhang.com

What if memTable is too large?



- Key point
  - Copy into a new SSTable in GFS/Disk when memTable is full

Copyright © www.jiuzhang.com

Interviewer: How to avoid system failure?

Copyright © www.jiuzhang.com

How to avoid system failure? Task: add <b, yeah>

Memory

Tablet0
memTable, pointer
SSTable0, Address0
SSTable1, Address1

memTable
...
b, yeah

GFS/Disk

SSTable0
...
a, sangpo;
c, flyer;
...

SSTable1
...
h, flyer;
m, put;
...
Index1 of blocks

tabletLog
...
Add <b,yeah>

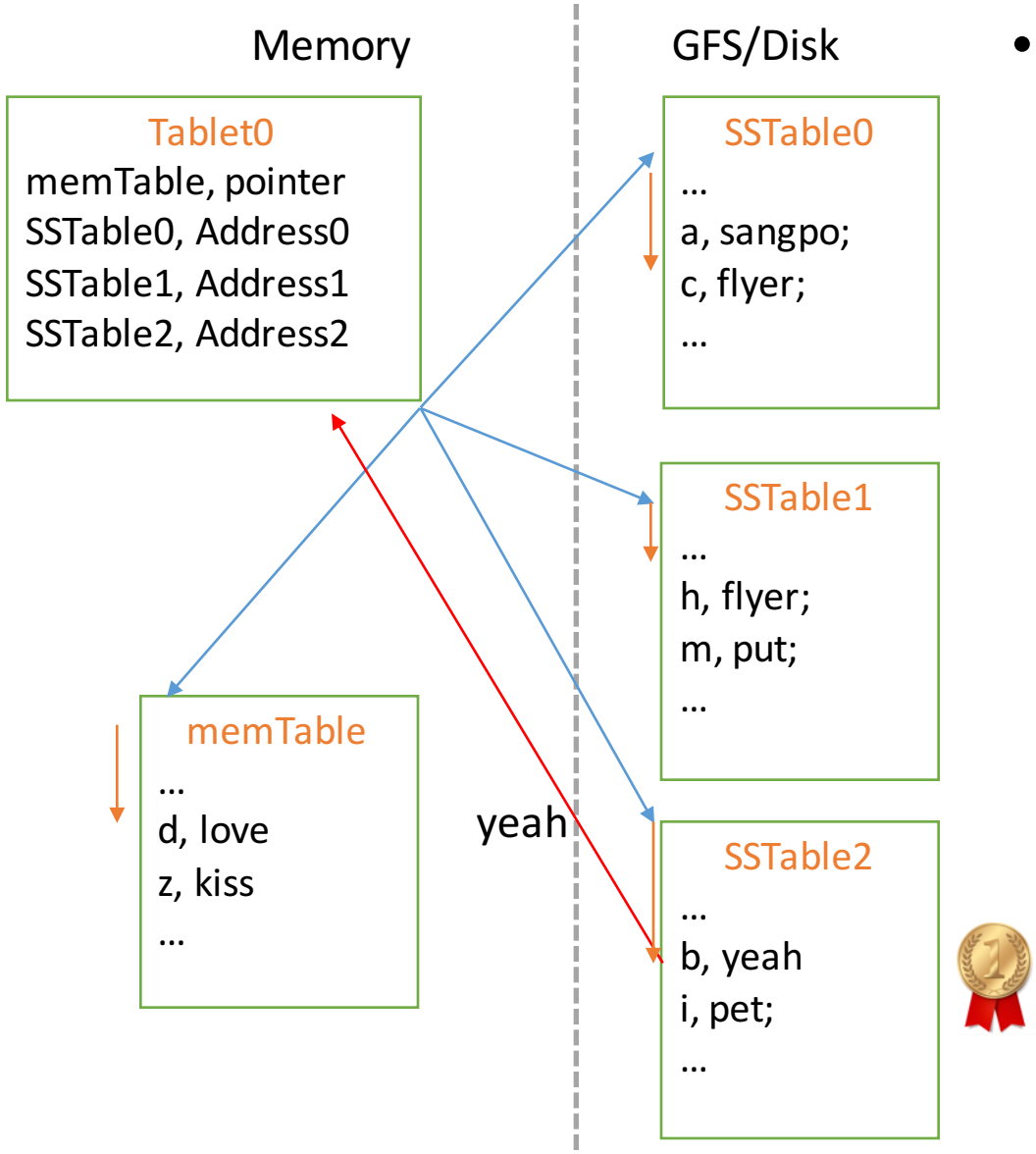
- Key point
  - A tablet = memTable + a list of SSTables + log

Copyright © www.jiuzhang.com

Interviewer: How to read data?

Copyright © www.jiuzhang.com

How to read data? Task: lookup(b)



- Key point
  - Due to our write strategy
    - Data is ordered inside a SSTable
    - Data is not ordered among SSTables
  - A read need to ask all SSTables and memTable
  - Each SSTable needs to use Disk search to find the element

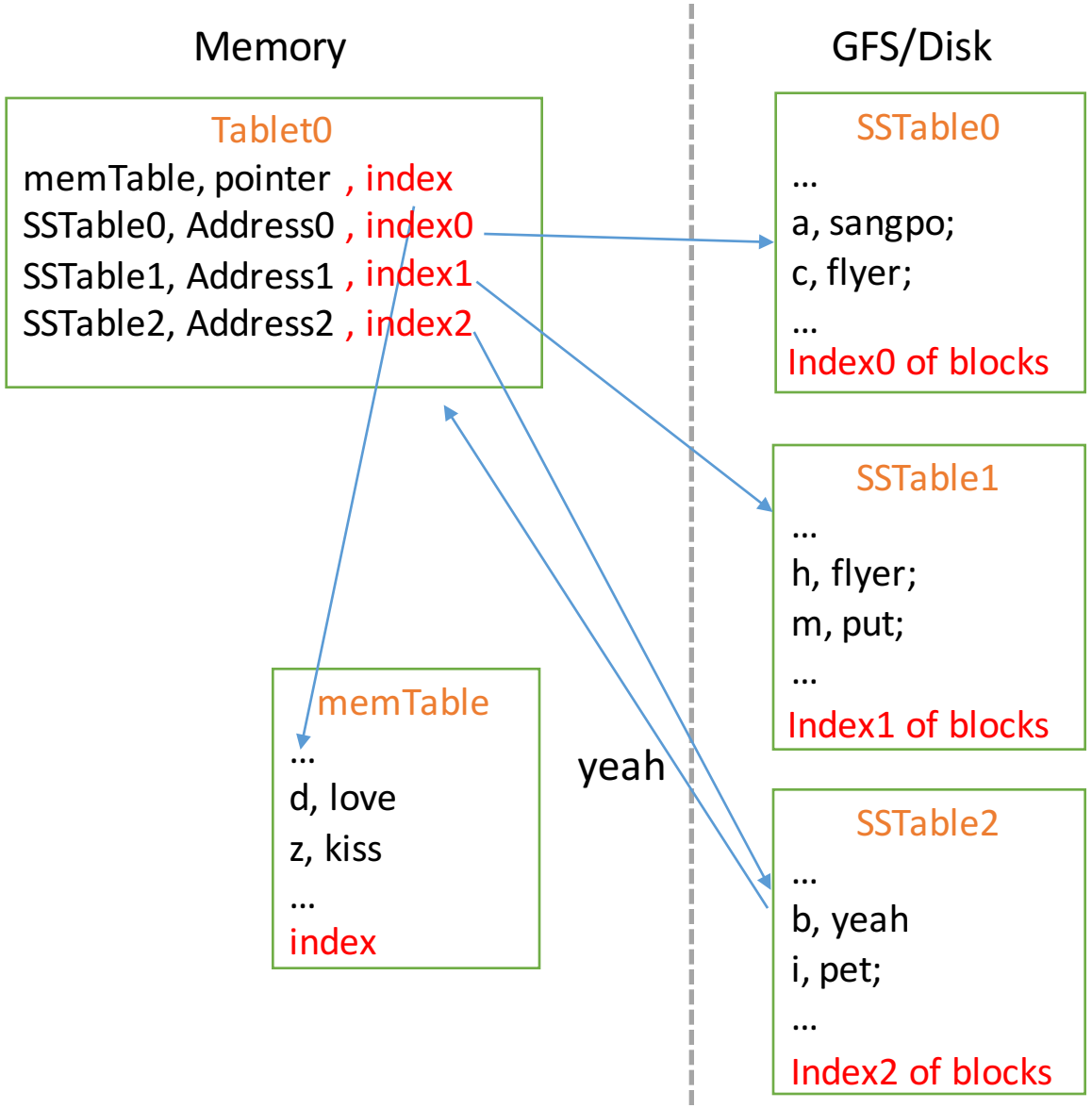
Copyright © www.jiuzhang.com

Interviewer: How to read data fast?

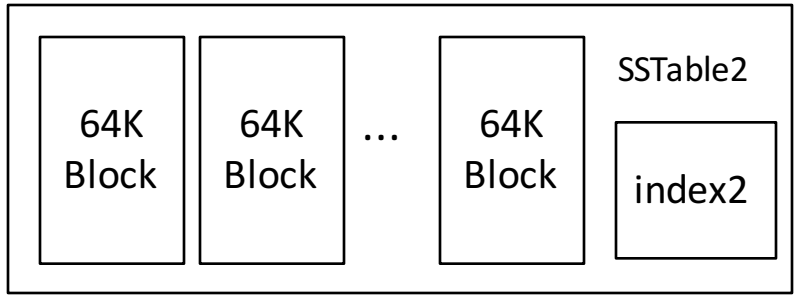
Copyright © www.jiuzhang.com

How to read data fast?

Task: lookup(b)?



- Key point
  - A SSTable = a list of sorted <key, value>
    = a list of 64K blocks + inde
  - Indices are pre-loaded into memory
  - Read fast by
    - Use indices to identify the place in disk

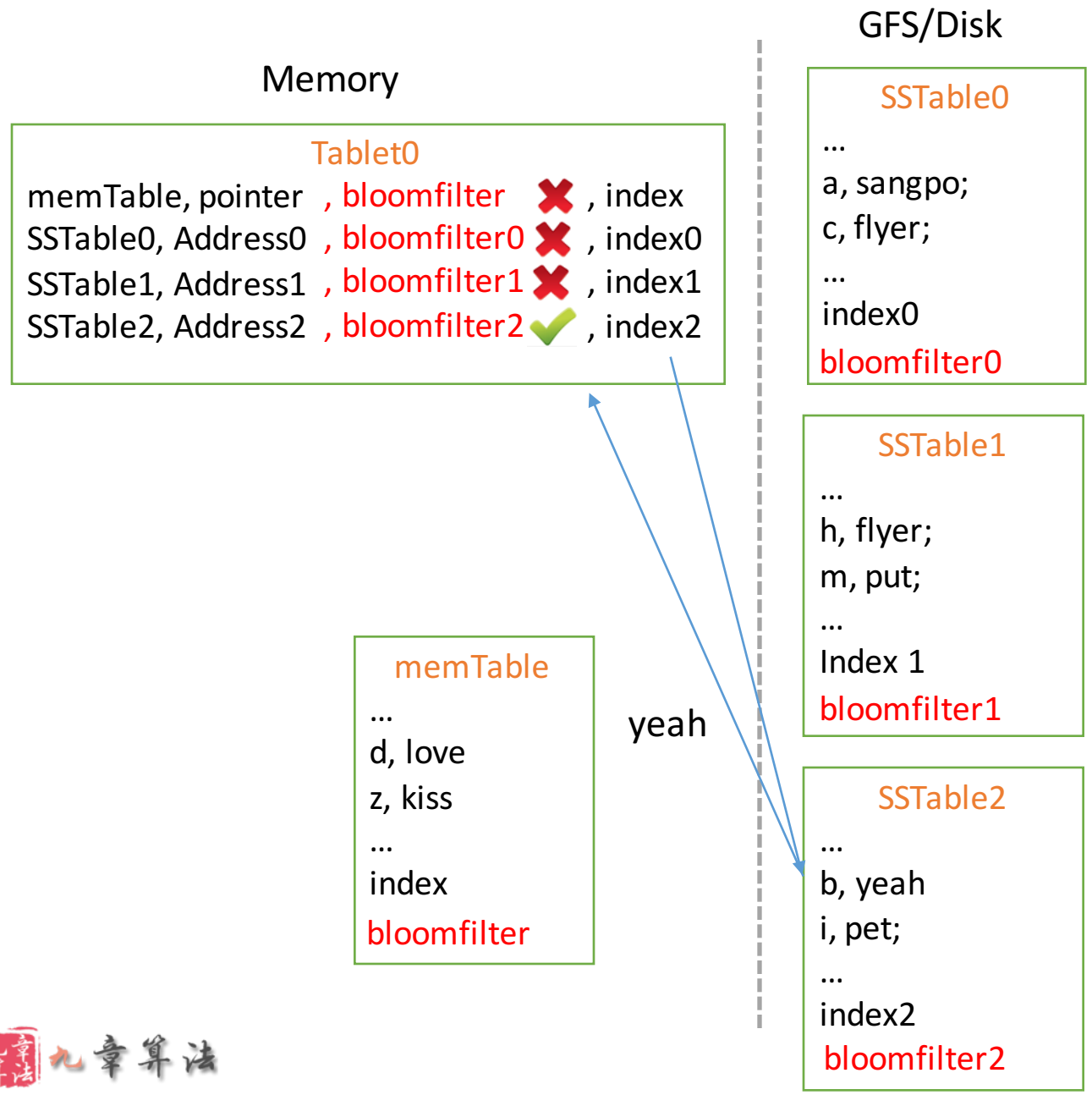


Copyright © www.jiuzhang.com

Interviewer: Read faster?

Copyright © www.jiuzhang.com

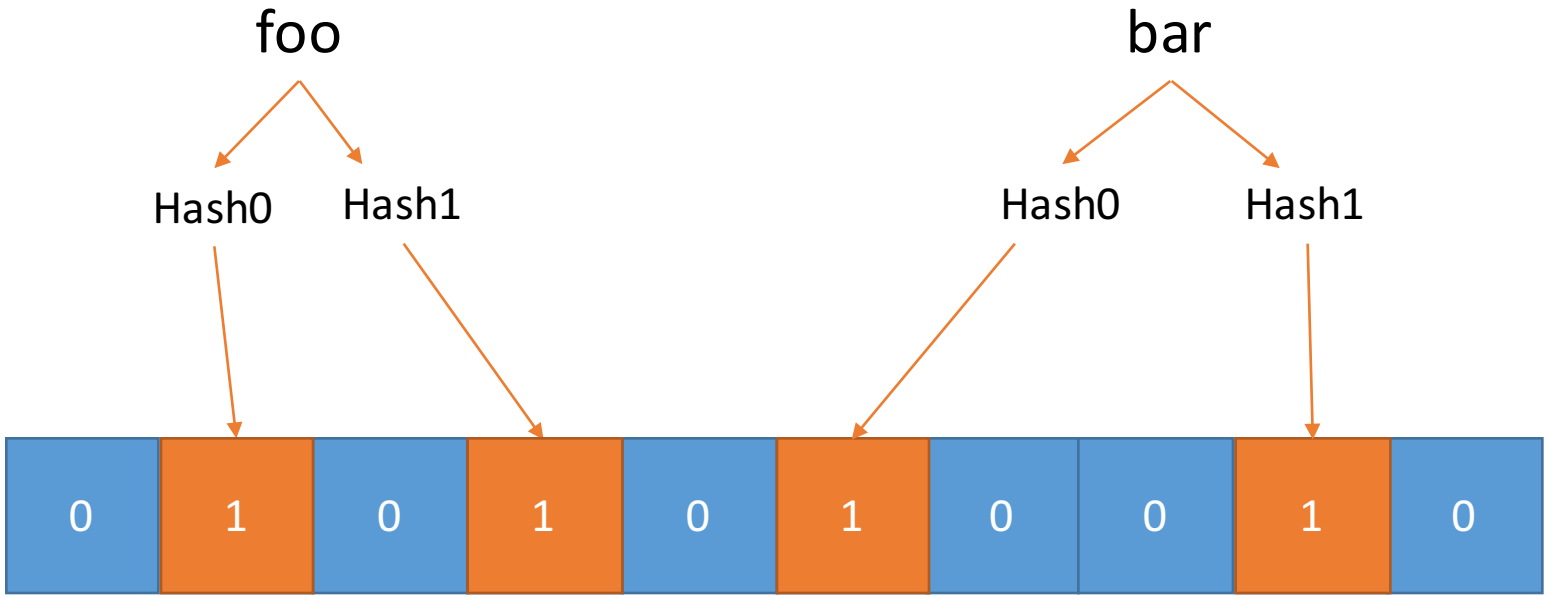
How to read data faster? Task: lookup(b)?



- Key point
  - A SSTable = a list of sorted <key, value>
    = a list of 64K blocks + index + bloomfilter
  - Bloomfilters are pre-loaded into memory
  - Read faster by
    - Check the existence of the element with bloomfilters

Copyright © www.jiuzhang.com

How to build bloom filter?

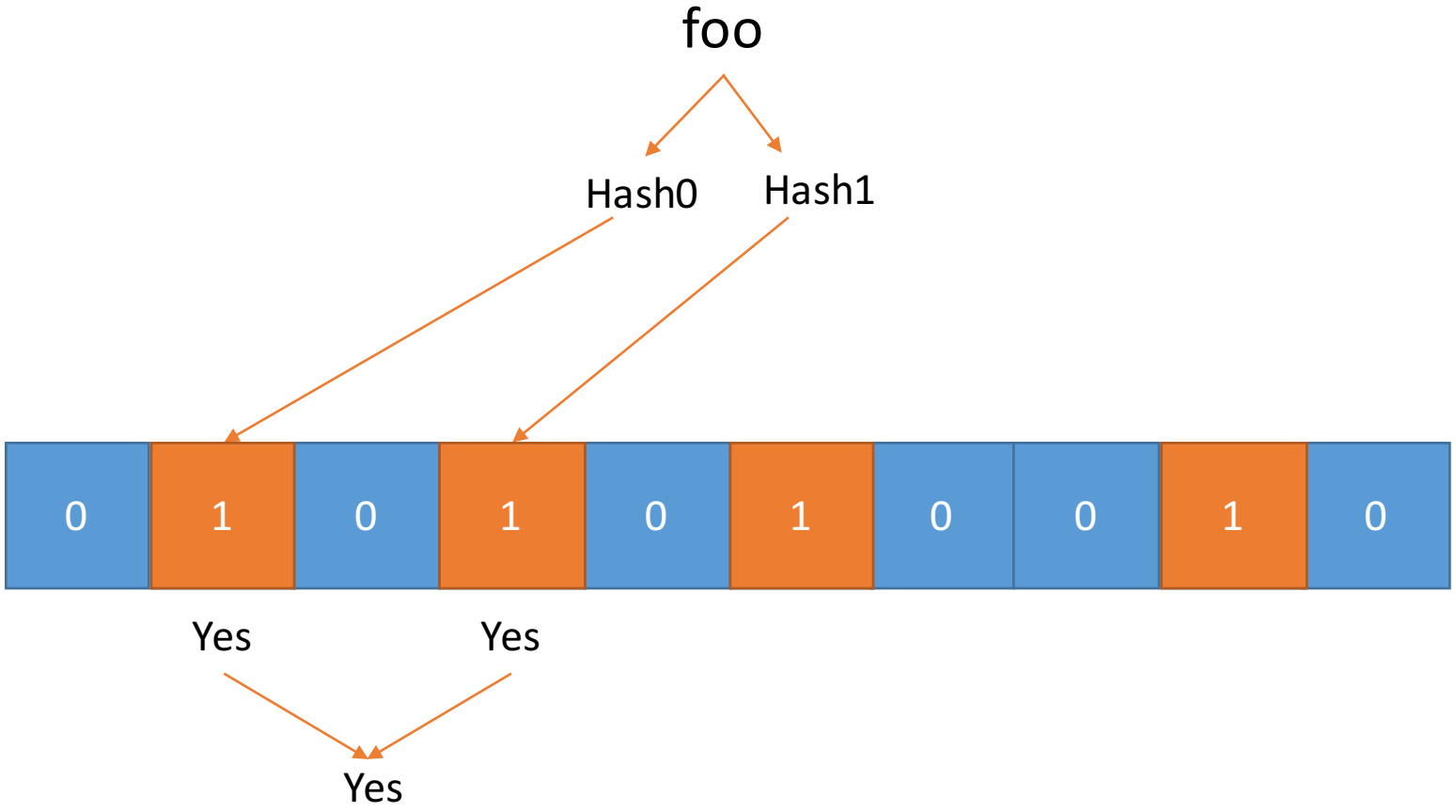


Read More

Novice, http://url.cn/Qu6dab
Expert/Master, http://url.cn/aPCYvo
Expert/Master, http://url.cn/TkF35M

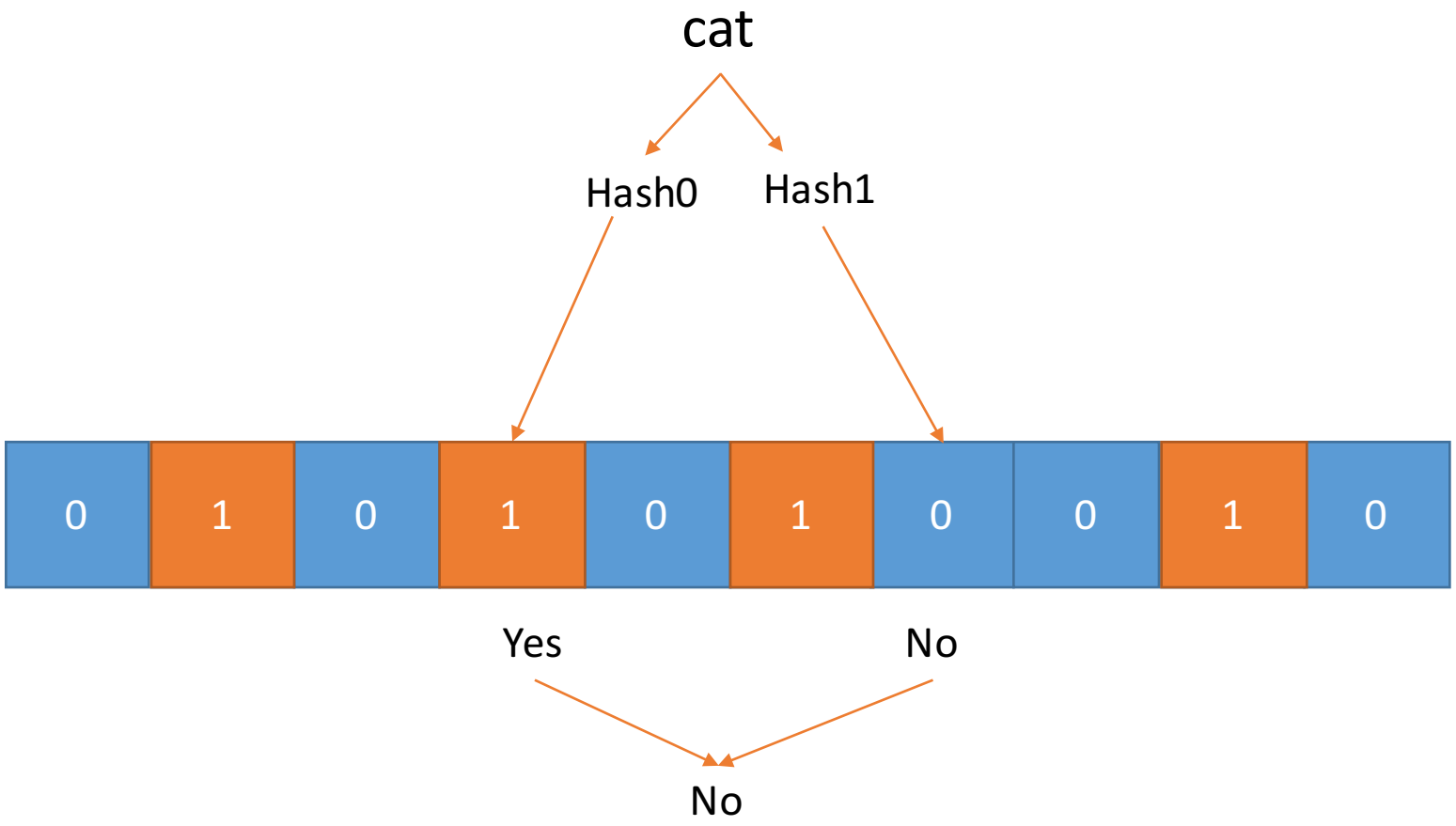
Copyright © www.jiuzhang.com

Lookup “foo”



九章算法
Copyright © www.jiuzhang.com

Lookup “cat”

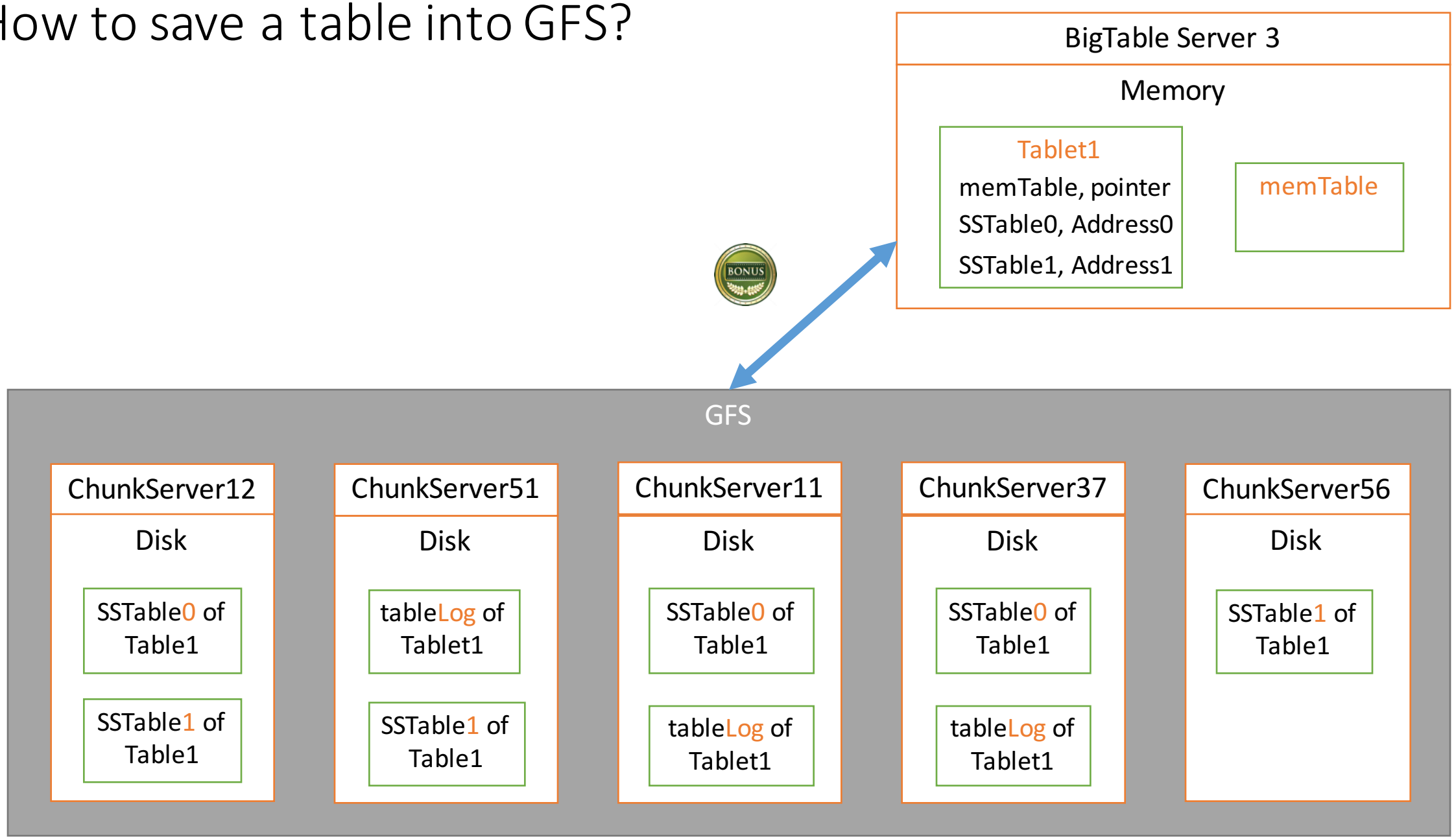


九章算法
Copyright © www.jiuzhang.com

Interviewer: How to save a table into GFS?

Copyright © www.jiuzhang.com

How to save a table into GFS?

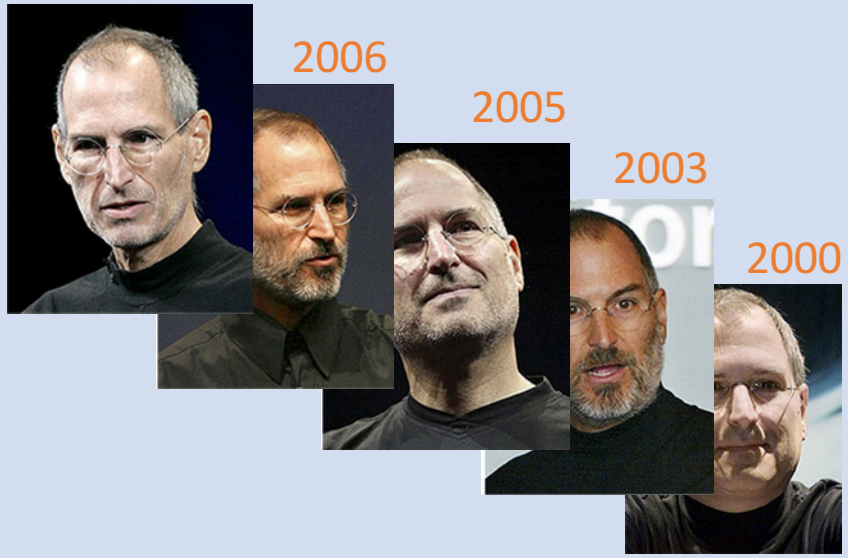


Copyright © www.jiuzhang.com

Interviewer: What is the logical view of table?

Copyright © www.jiuzhang.com

What is the logical view of table?

		Body	
Name	Photo	Weight	Height
Sangpo			
Steve		110lb, 2009 120lb, 2008 150lb, 1995	6'2", 2011 5'7", 1987
Tiger			

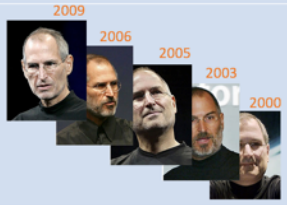
Copyright © www.jiuzhang.com

Interviewer: How to transform logic view into physical storage?

Copyright © www.jiuzhang.com

How to transform logic view into physical storage?

Logic view of Table

Name	Photo	Body	
		Weight	Height
Sangpo			
Steve		110lb, 2009 120lb, 2008 150lb, 1995	6'2", 2011 5'7", 1987
Tiger			

Key = string(row, column, time)

Physical storage of <Key, value>

Table

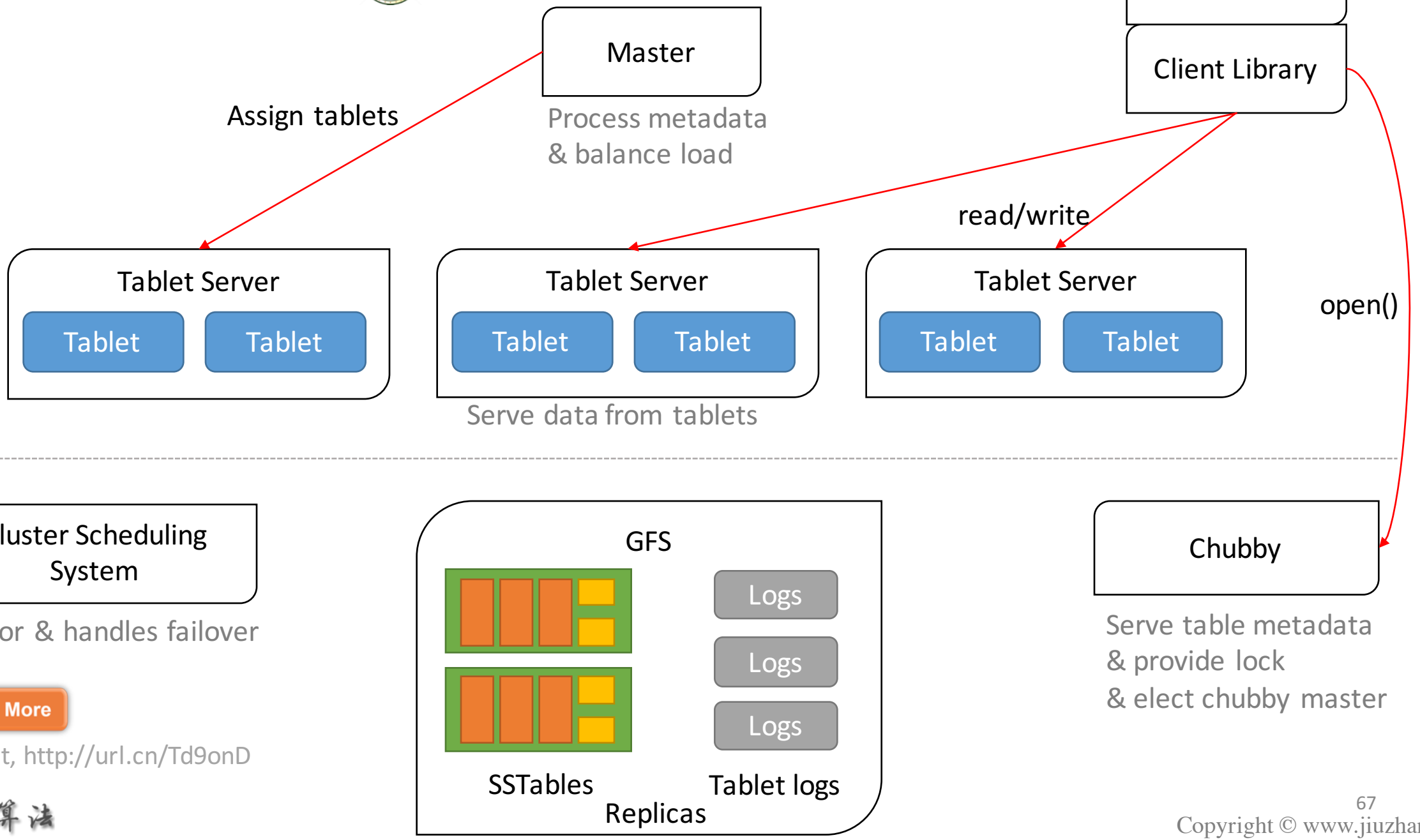
...
Steve;Body:Height;2011 -> 6'2"
Steve;Body:Height;1987 -> 5'7"
Steve;Photo;2009 -> 2009.JPG
Steve;Photo;2006 -> 2006.JPG
Steve;Photo;2005 -> 2005.JPG
...

Copyright © www.jiuzhang.com

Interviewer: What is the architecture of BigTable?

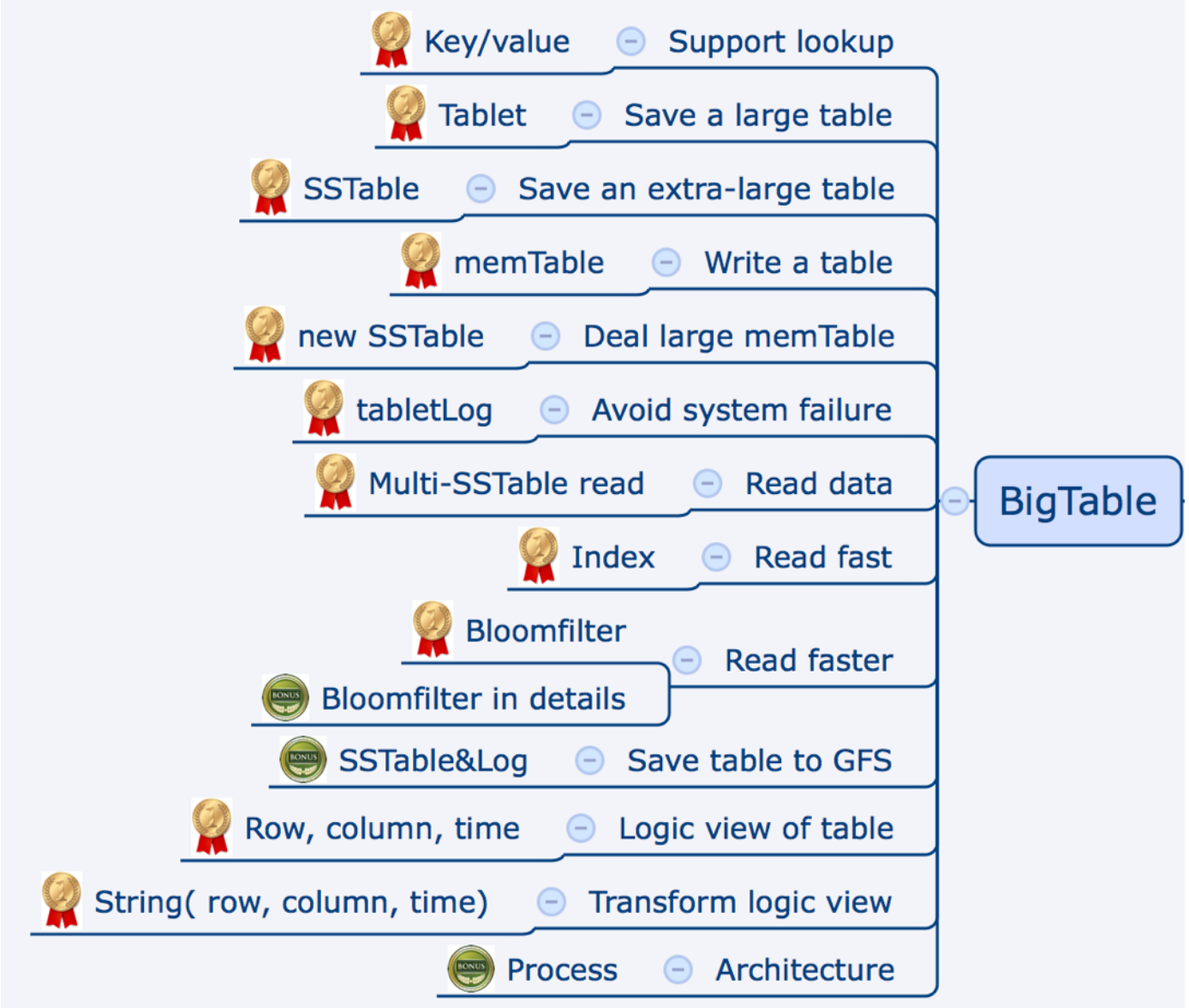
Copyright © www.jiuzhang.com

BigTable architecture



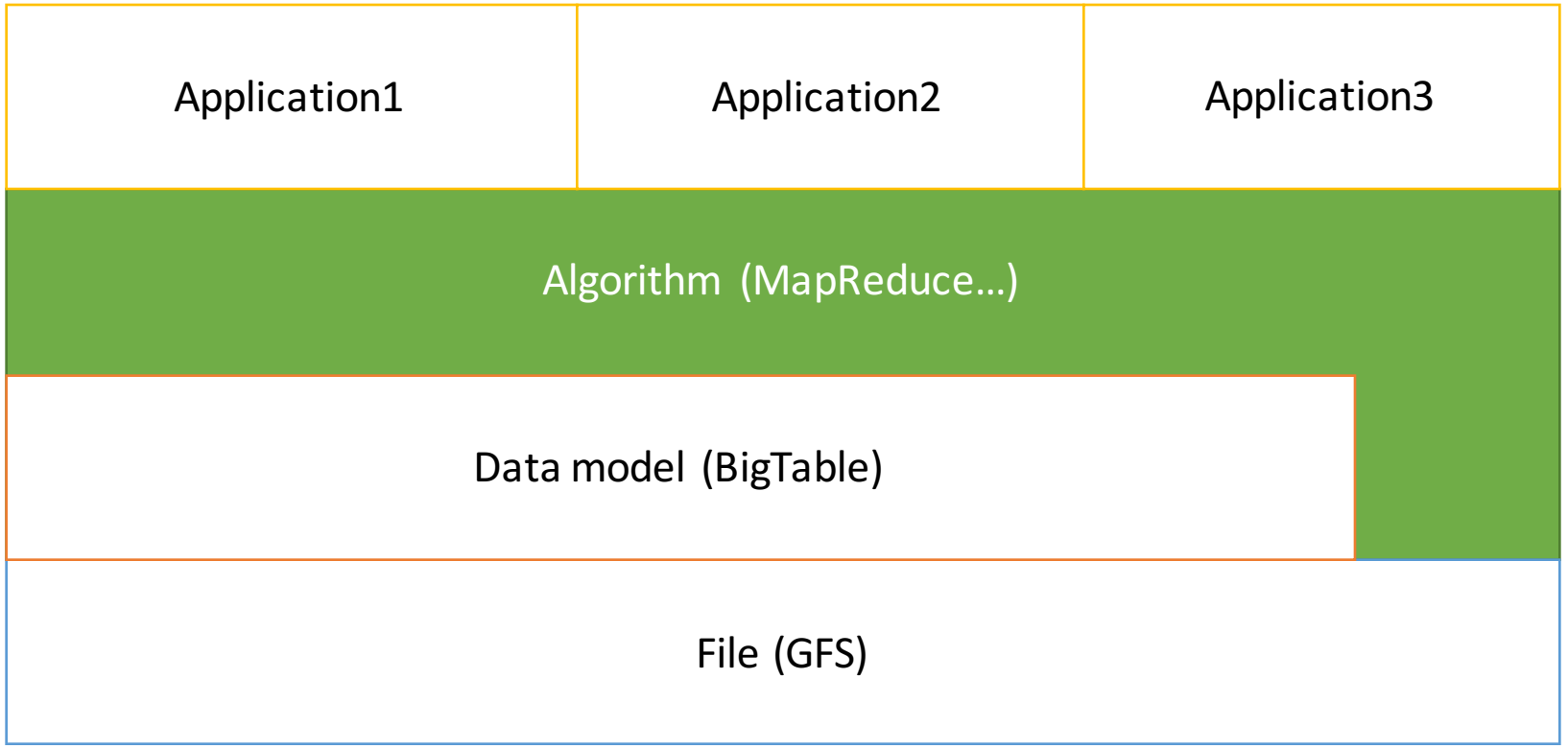
九章算法
Copyright © www.jiuzhang.com

BigTable summary (11+3)



Copyright © www.jiuzhang.com

Layers of system



Copyright © www.jiuzhang.com

MapReduce

Read More

Novice, http://url.cn/YM1tHI
Expert, http://url.cn/b41Qzf
Expert, http://url.cn/1VO6Qa
Expert, http://url.cn/ccvLOr
Expert/Master, http://url.cn/SuVoAP
Expert/Master, http://url.cn/SJCoso
Master, http://url.cn/Z3OOVZ

Copyright © www.jiuzhang.com

For 1TB <key, value> data, how to compute

- appearances of words
- inverted index
- anagrams

Copyright © www.jiuzhang.com

Interviewer: What is Map and Reduce?

Copyright © www.jiuzhang.com

What is map?

Input

Map

Output

Copyright © www.jiuzhang.com

What is reduce?

Input

Reduce

Output

九章算法
Copyright © www.jiuzhang.com

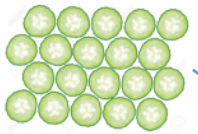
MapReduce

Input | Split | Map | Shuffle | Reduce | Finalize














Ticket 13

Ticket 14

Ticket 15

Read More

Novice, http://url.cn/Sd0btu

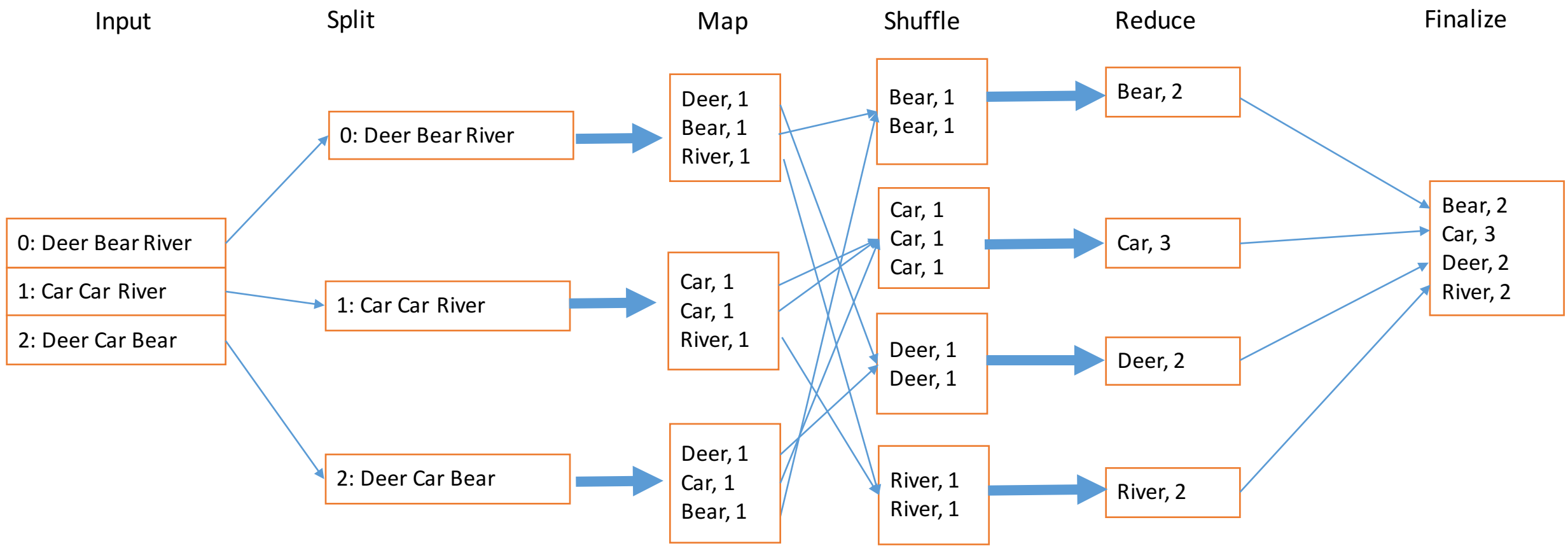
九章算法

Copyright © www.jiuzhang.com

Interviewer: Calculate word appearance with MapReduce

Copyright © www.jiuzhang.com

Calculate each word's number of appearance




```
Map( string key, string value)
    //key: the id of a line
    //value: the content of the line
    for each word in value:
        OutputTemp( word, 1);
```

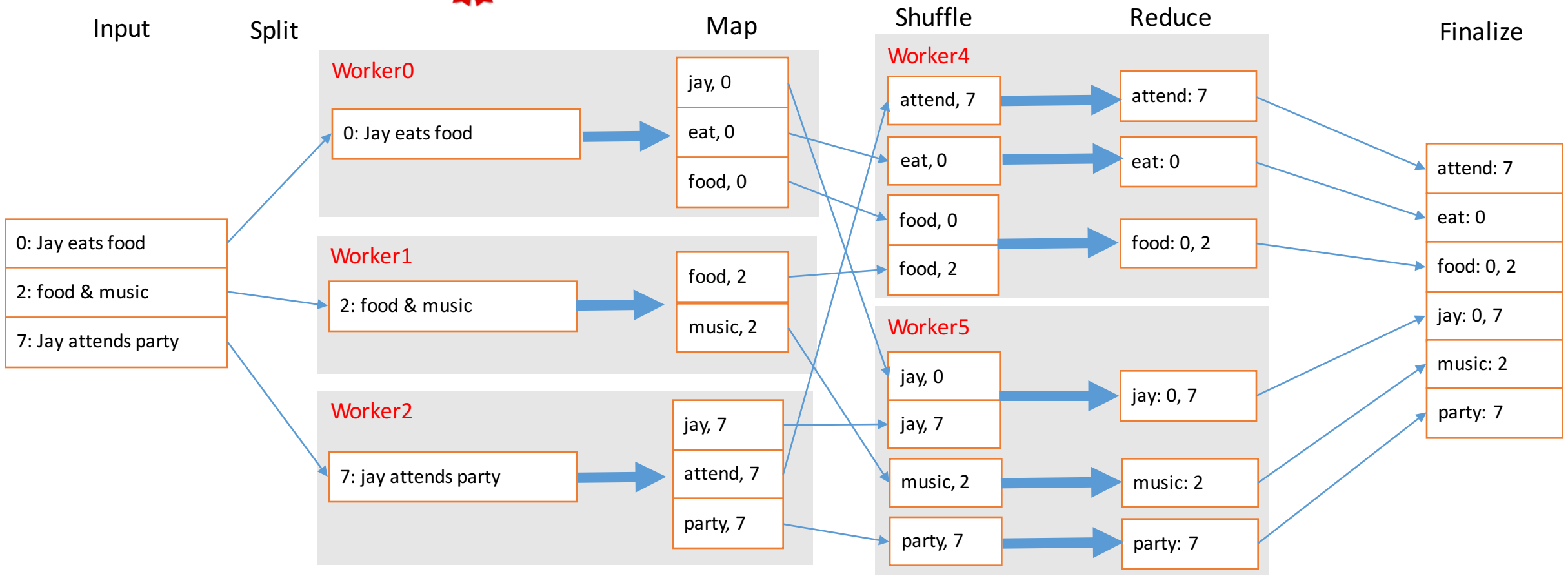
```
Reduce( string key, list valueList)
    //key: the name of a word
    //valueList: the appearances of this word
    int sum = 0;
    for value in valueList:
        sum+=value;
    OutputFinal( key,sum);
```


Copyright © www.jiuzhang.com

Interviewer: Build inverted index with MapReduce?

Copyright © www.jiuzhang.com

Build inverted index




```
Map( string key, string value )
    //key: the id of a document
    //value: the content of the document
    for each word in value:
        OutputTemp( word, key );
```

```
Reduce( string key, list valueList )
    //key: the name of a word
    //valueList: the appearances of this word in documents
    List sumList;
    for value in valueList:
        sumList.append(value);
    OutputFinal( key, sumList );
```

Read More

Novice/Expert, http://url.cn/fsZ927

九章算法 Copyright © www.jiuzhang.com

Interviewer: Calculate anagrams with MapReduce?

Copyright © www.jiuzhang.com

LintCode *Coding interview questions online training system*

Medium Anagrams

Given an array of strings, return all groups of strings that are anagrams.

Example

Given `["lint", "intl", "inlt", "code"]`, return `["lint", "inlt", "intl"]`.

Given `["ab", "ba", "cd", "dc", "e"]`, return `["ab", "ba", "cd", "dc"]`.

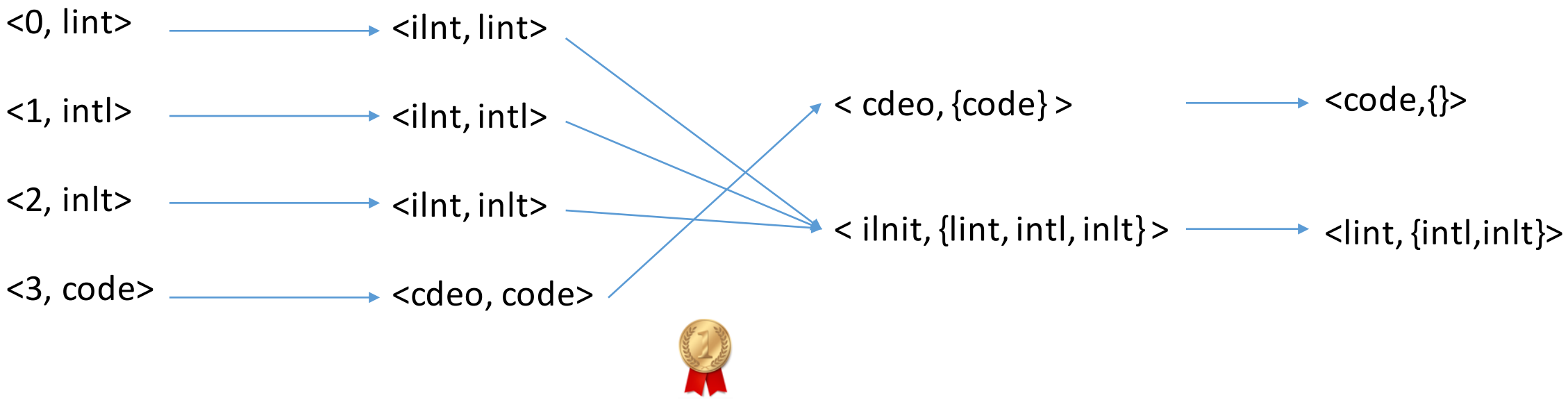
Anagram(lint) = { <i,1>, <l,1>, <n,1>, <t,1> }
Anagram(intl) = { <i,1>, <l,1>, <n,1>, <t,1> }
Anagram(inlt) = { <i,1>, <l,1>, <n,1>, <t,1> }

Anagram(code) = { <c,1>, <d,1>, <e,1>, <o,1>}

Read More

Novice, http://url.cn/WGXjtQ
Novice/Expert, http://url.cn/cGRS1l

Copyright © www.jiuzhang.com



Map

Input
k = lineID
v = word

Output
k = sort(word)
v = word

Shuffle

Reduce

Input
k = sort(word)
v = {anagram0, anagram1, ...}

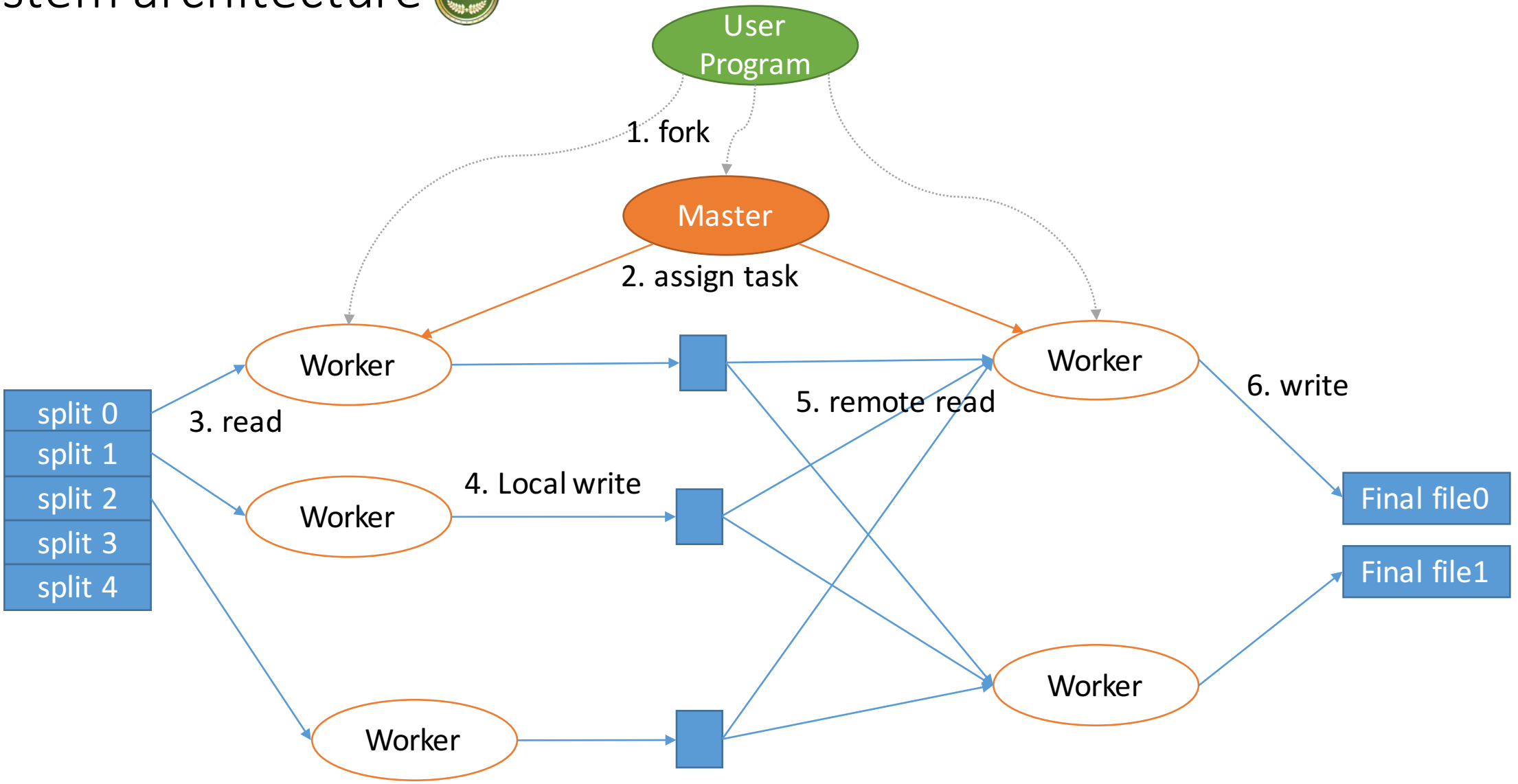
Output
k = anagram0
v = {anagram1, anagram2,...}

Copyright © www.jiuzhang.com

Interviewer: What is the architecture of MapReduce?

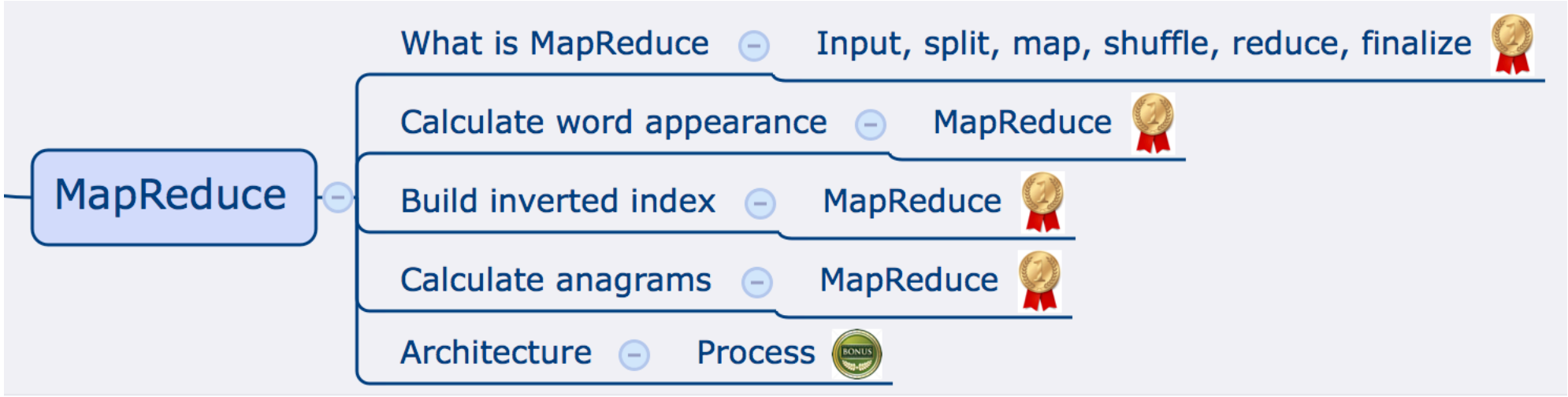
Copyright © www.jiuzhang.com

System architecture



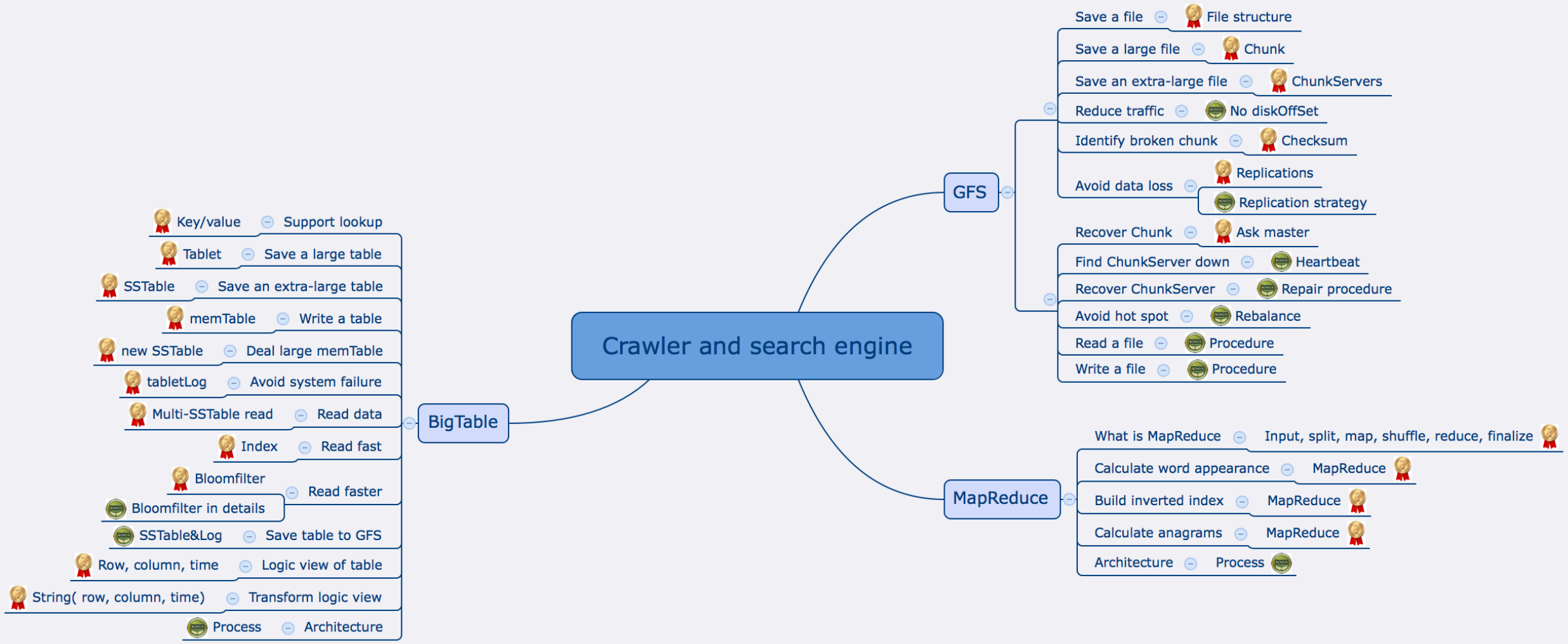
九章算法
Copyright © www.jiuzhang.com

MapReduce summary (4+1)



Copyright © www.jiuzhang.com

Class5 summary



Copyright © www.jiuzhang.com

Keyword

- GFS
- BigTable
- MapReduce

- Bloom filter

- Log analysis

Copyright © www.jiuzhang.com

Homework

- Output top10 IPs in the latest hour in real-time

http://www.jiuzhang.com/qa/109/

Copyright © www.jiuzhang.com

QA

关注微信/微博，获取最新面试题及权威解答

微信: ninechapter

微博: http://www.weibo.com/ninechapter

官网: www.jiuzhang.com

Copyright © www.jiuzhang.com

MR. Problem

Divide & Conquer

Map is disassemble;
Reduce is assemble.
Our life is a sequence of
assemble and disassemble.

You may wonder why!
Is it true?
But how to solve my job interview,
such a real problem?

Disassemble MR. problem again and again,
Until you can assemble the courage to face it.
Wish you, my friend,
Assemble all the pieces to win your job, hohoho!

Copyright © www.jiuzhang.com