Introducing Big Data & Hadoop Ecosystem

课程版本 v2.0　主讲 赵敏

版权声明

九章的所有课程均为直播课程，受法律版权保护
禁止录像与传播录像，否则将追究法律责任和经济赔偿

My Background

- 毕业于北美TOP5 CS专业
- 就职于FLAG之一的企业，曾在FLAG中的2家企业实习
- 从事cloud, big data 相关工作
- 拥有1年的面试官经验

What is big data?

大！数！据！

What do you want to do with big data?

Teenage Sex

Job Classification

- Data Warehousing
- Data Analysis
- Data Transformation
- Data Collection
- ...

Is that required to have big data background?

YES

- Try to learn big data from different perspective
- Do something to build you profile
- Keep energetic and keep learning

Let’s learn something new!

Outline

- Introduction to Hadoop Ecosystem
- MapReduce Project
- Course Outline

Question

- Have you heard Spark?
- Have you heard Hadoop?

Hadoop vs. Spark

- Interview
- Hadoop是工具包，Spark是工具
- MapReduce is widely used
- Java vs. Python

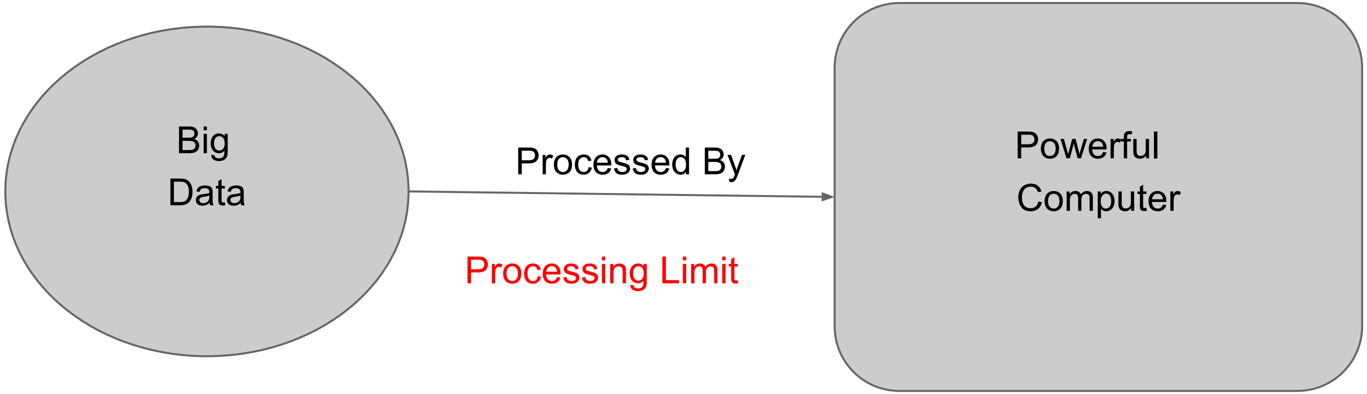
What is Hadoop?

- Application?

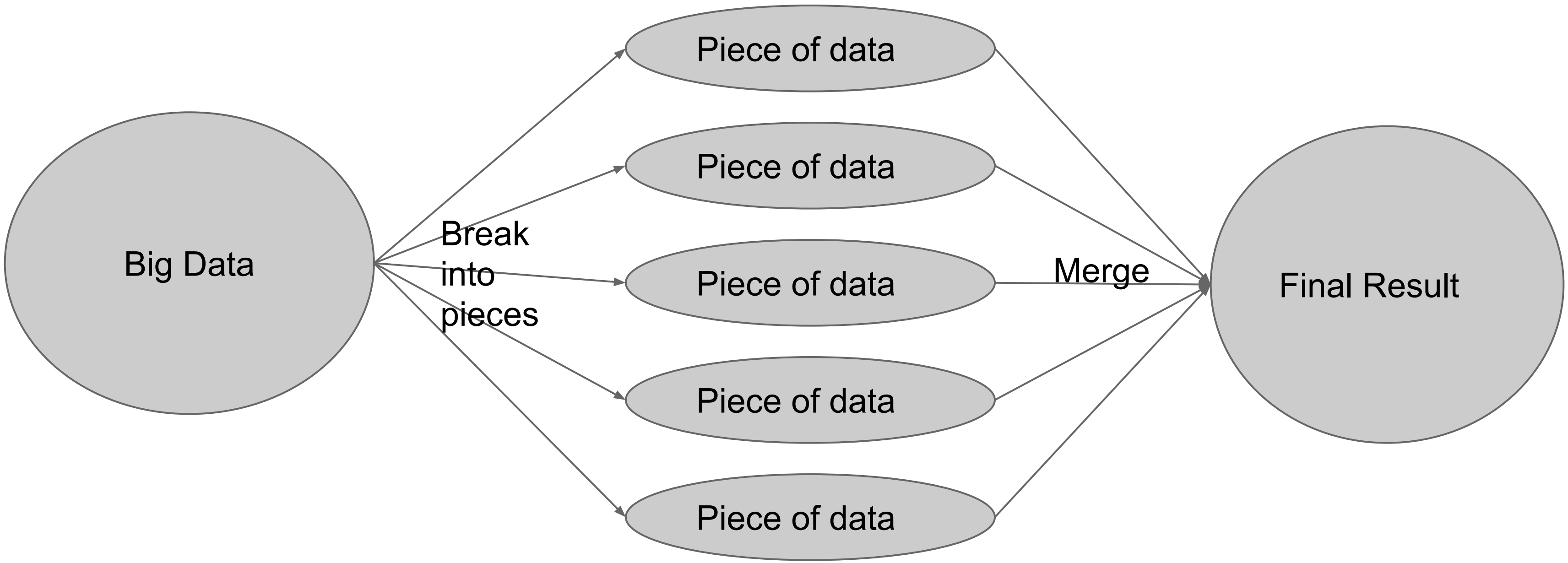
What is hadoop?

Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware.

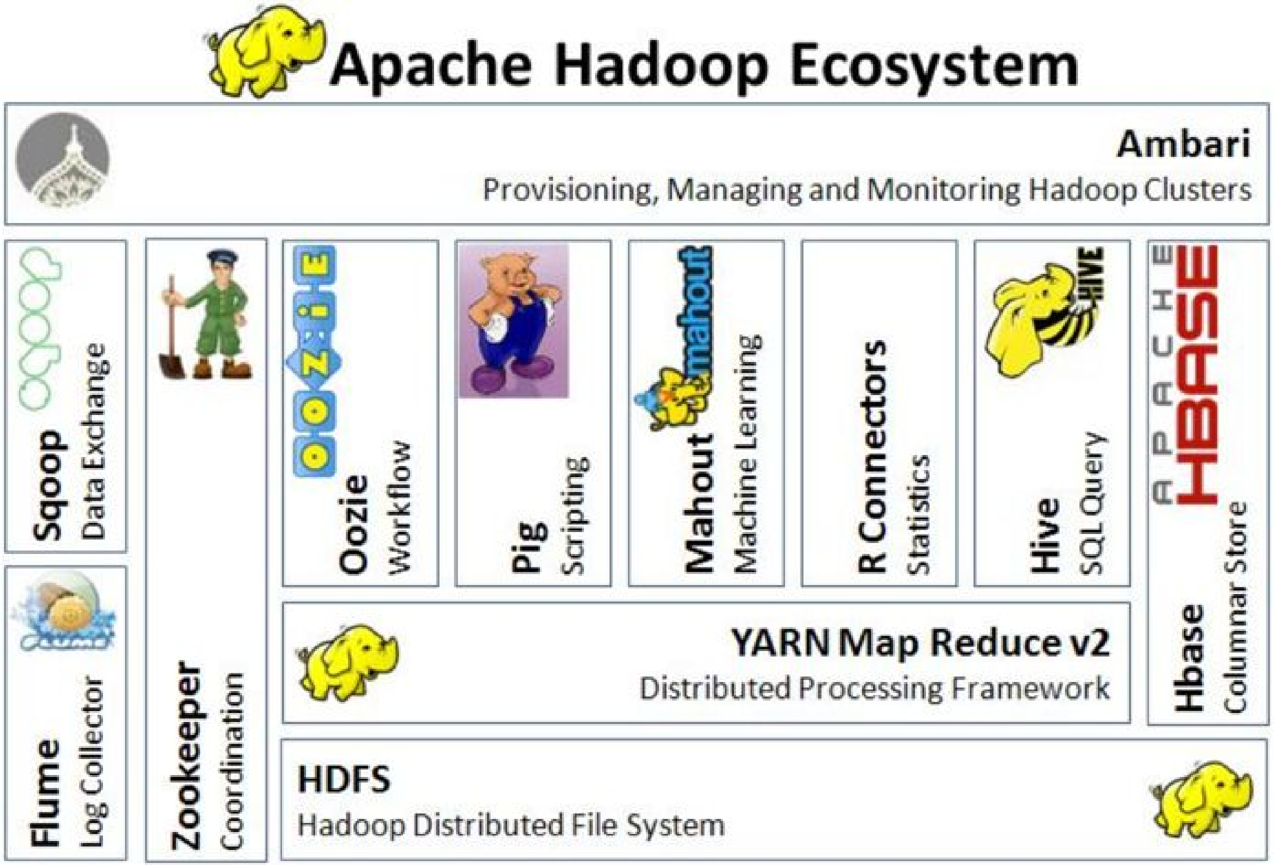
Why is Hadoop



Why is Hadoop



Hadoop Ecosystem



Apache Hadoop Ecosystem

Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

Sqoop
Data Exchange

Flume
Log Collector

Zookeeper
Coordination

Oozie
Workflow

Pig
Scripting

Mahout
Machine Learning

R Connectors
Statistics

Hive
SQL Query

Hbase
Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System

Hadoop Ecosystem

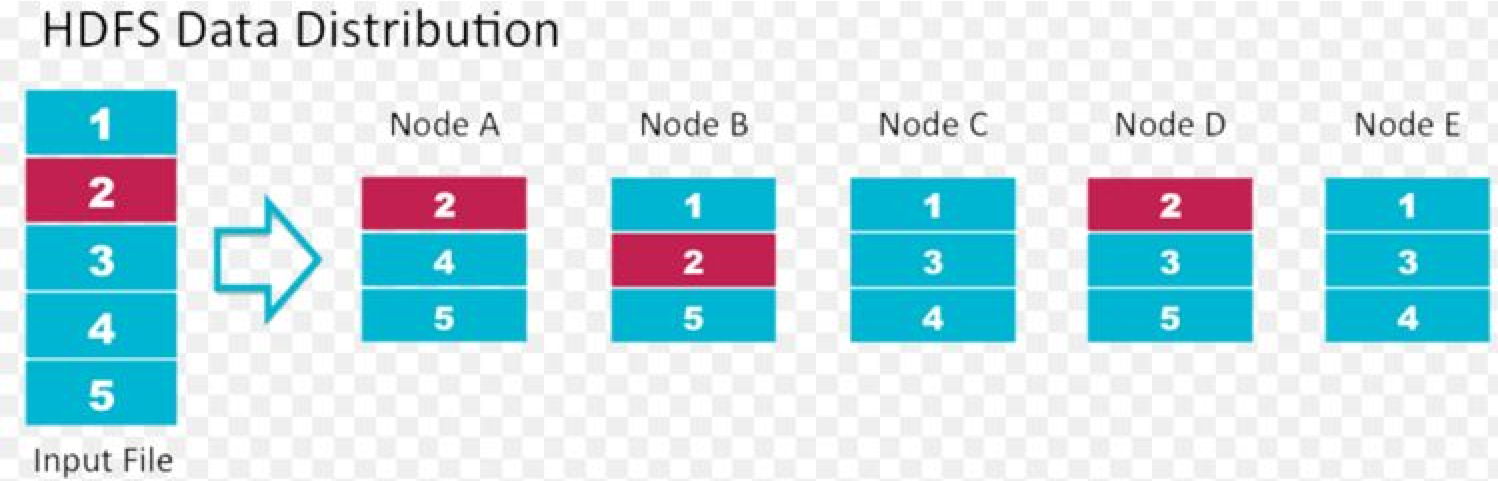
Now we have big data, what is the first thing to do?

Hadoop Ecosystem

Storage on multiple machines

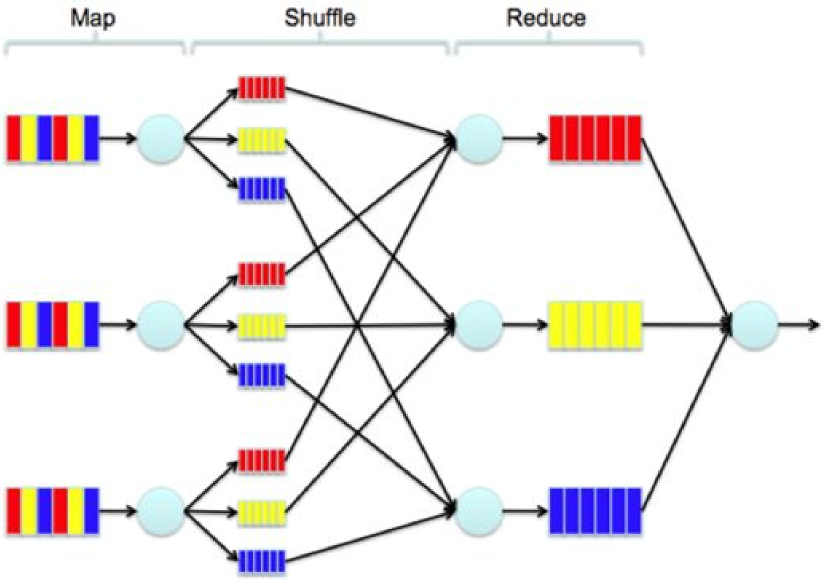
HDFS

- Data Storage
  - Data Split
  - Data replication
- All you care about is the path of file: /hdfs/input/doc1.txt



MapReduce

- Data Processing
  - How to leverage job
  - How do nodes communicate
  - How to deal with node failure
- All you care about is input and output



HBase

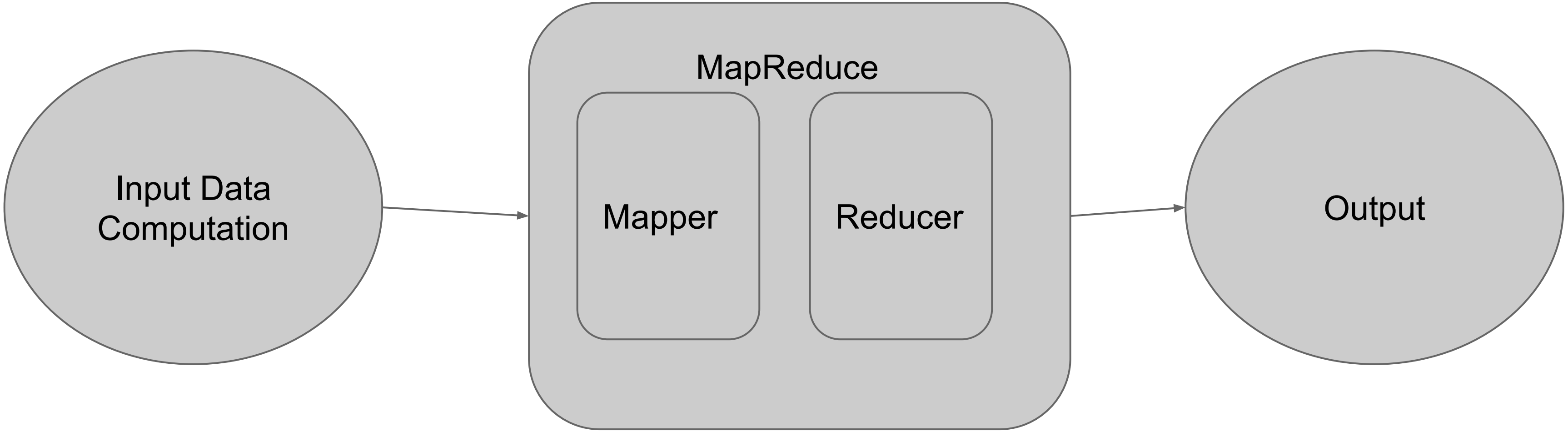
- HBase is an open source, non-relational, distributed database
- Key-Value store

MapReduce

What is MapReduce?

MapReduce

- A programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster.



A Simple Question

Input:

I love big data course since big data is interesting.

```
String[] words = line.split(“ ”);
for(String word: words) {

}
```

```
HashMap<String, int> counts = new HashMap();
```

Key -- unique word
value -- count

Follow up

What if the input is a big file = 1T?

Follow up

Can we process it with one machine?

Follow up

No!

Out of memory

Too slow

Follow up

What should we do?

Follow up

Use multiple machines

MapReduce

- Mapper
- Reducer

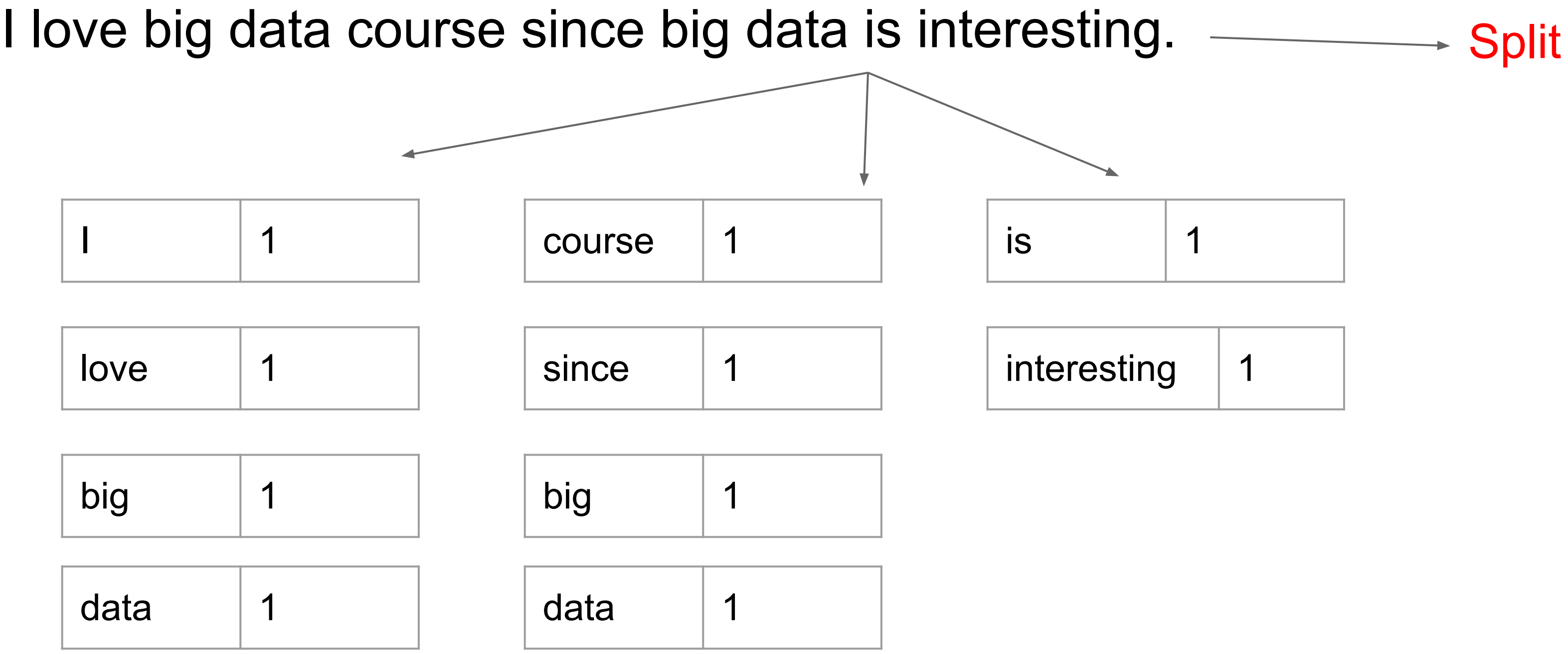
Mapper

- A stage
    - Splits into small chunks
- A class
    - https://hadoop.apache.org/docs/r2.6.2/api/org/apache/hadoop/mapreduce/Mapper.html
    - Simplify your work
        - Read/Write

Reducer

- A stage
  - Combines data from Mapper
- A class
  - https://hadoop.apache.org/docs/r2.6.2/api/org/apache/hadoop/mapreduce/Reducer.html
  - Simplify your work
    - Read/Write
    - Many to Many

Mapper



I love big data course since big data is interesting.

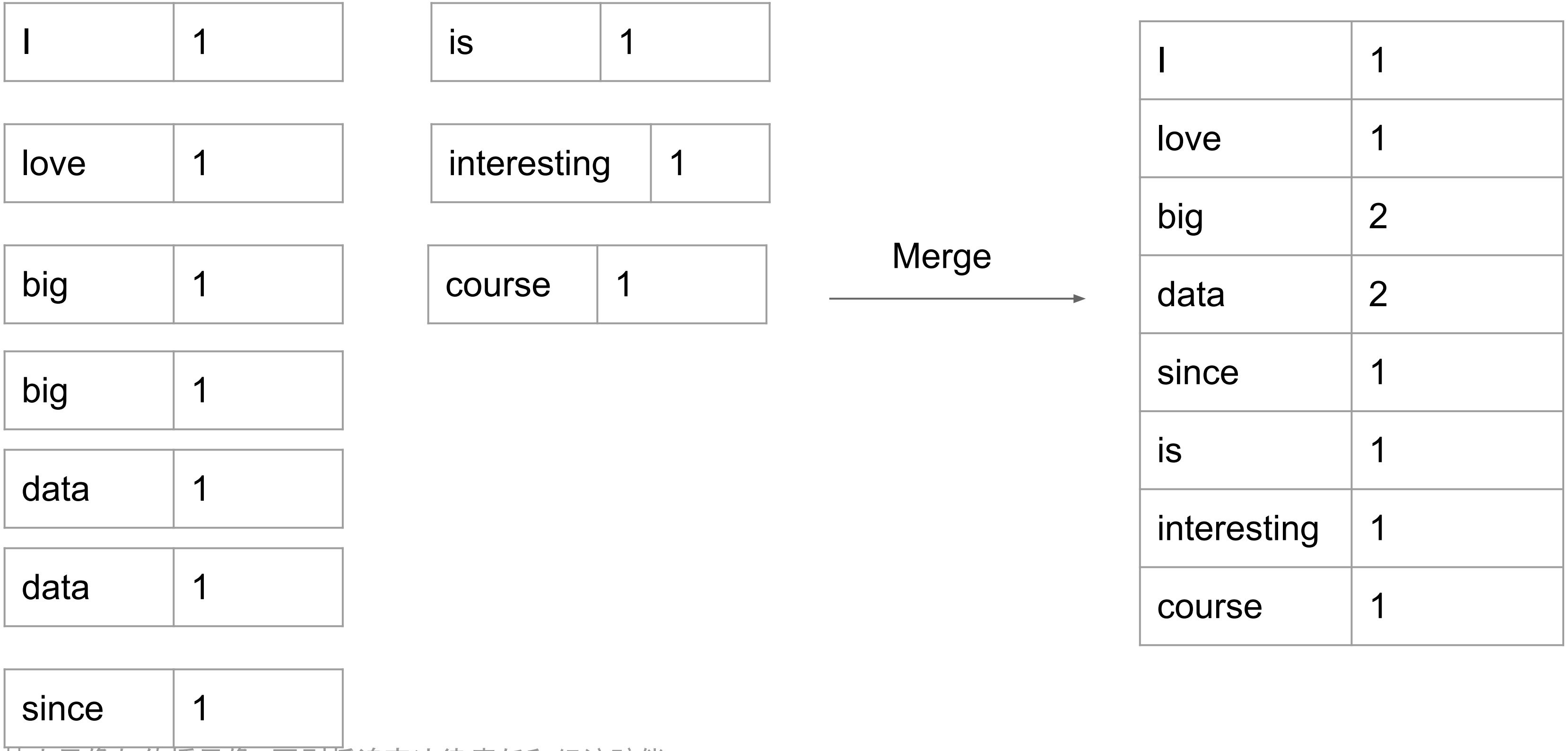
Split

I	1
love	1
big	1
data	1

course	1
since	1
big	1
data	1

is	1
interesting	1

Reducer



Input

I	1
love	1
big	1
big	1
data	1
data	1
since	1
is	1
interesting	1
course	1

Merge

I	1
love	1
big	2
data	2
since	1
is	1
interesting	1
course	1

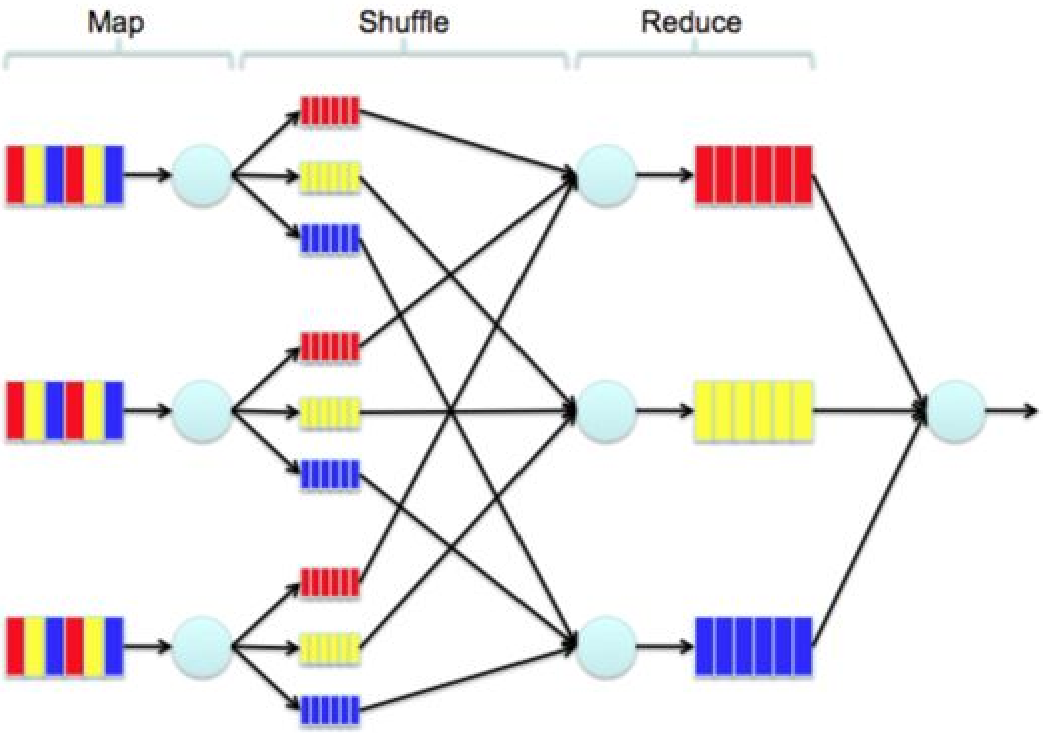
MapReduce

Let’s write code!

Shuffle

- Transport
- Sort

MapReduce



Follow up2

Find out the hottest five words on Twitter

Why is Hadoop

Advantages of Hadoop?

Performance

- Throughput
  - Run computation in parallel
- Scalability
  - Store and distribute very large data sets across clusters of hundreds of inexpensive servers operating in parallel
- Reliability
  - Even if individual nodes experience high rates of failure when running jobs on a large cluster, data is replicated across a cluster so that it can be recovered easily in the face of disk, node or rack failures
- Cheap
  - Cluster of inexpensive servers

What I have learned

- Hadoop Ecosystem
  - HDFS
  - MR
  - Hbase
- MapReduce
- MapReduce Project

Course Outline

- MapReduce Project - Google Search Auto Complete I
- MapReduce Project - Google Search Auto Complete II
- Hadoop Distributed File System
- MapReduce Project - PageRank
- 面试大数据常见问题之TopK - 微博今日热门话题
- MapReduce Project - Recommender System I
- MapReduce Project - Recommender System II

FAQ

- 大家做相同的project，是否会导致重复性高？
- 用什么平台做Mapreduce project？
- Project主要用什么语言？
- 会带着我们写代码吗？
- 如何合理的做项目？
- 课程价格是否偏贵？
- 用Windows可以吗？

版权声明

九章的所有课程均为直播课程，受法律版权保护
禁止录像与传播录像，否则将追究法律责任和经济赔偿

Q&A

常见问题 http://www.jiuzhang.com/qa/3/