Hadoop

一、课前准备

1. vmware虚拟机软件*1
2. centos7虚拟机*3
3. 3节点hadoop集群

二、课堂主题

本节课主要讲解大数据的背景，应用于哪些行业，hadoop是什么，hadoop生态圈，hadoop架构，hdfs分布式文件系统，hdfs的体系结构，hadoop常用命令

三、课堂目标

1.能够说出大数据的背景

2.了解hadoop是什么

3.了解hadoop的生态圈及架构

4.能够说出什么是分布式文件系统

5.理解hdfs的工作流程

6.了解mapreduce编程模式

7.可以使用常用命令操作hadoop

四、知识要点

1.大数据背景

1.1大数据起源与趋势

数据的爆炸式增长

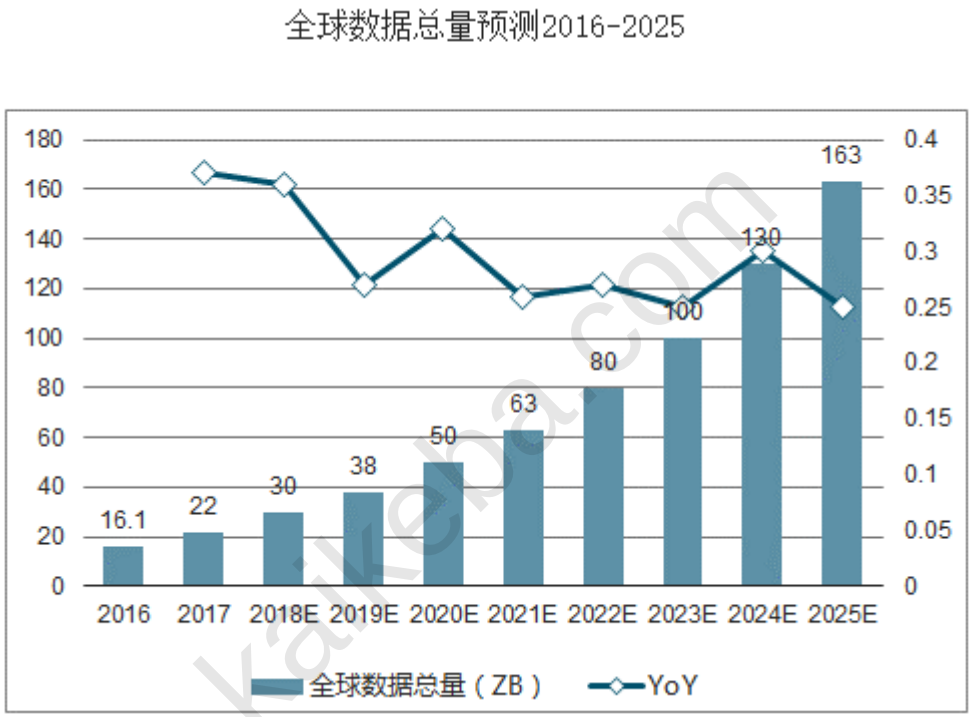
地球上至今总共的数据量：

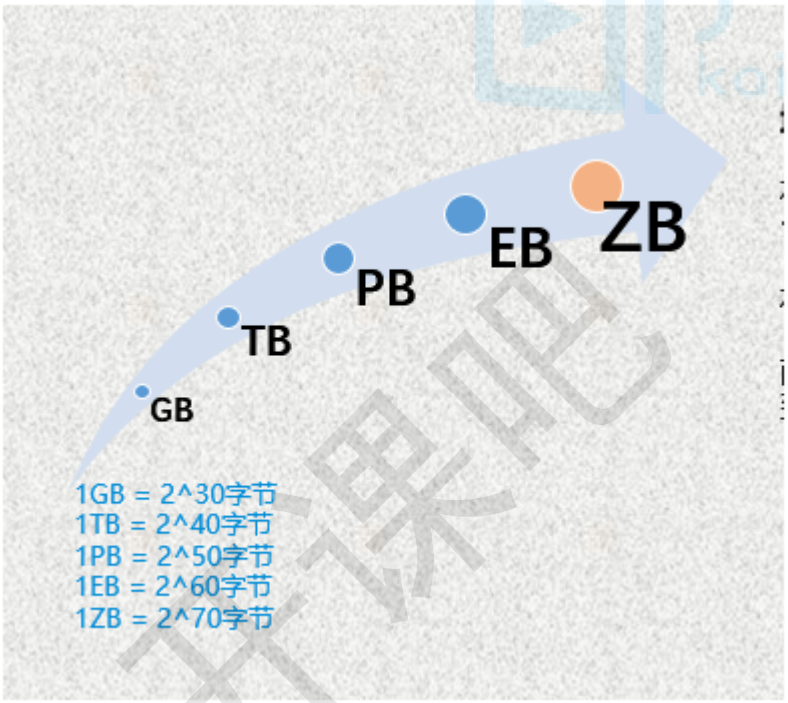
在2006 年，个人用户才刚刚迈进TB时代，全球一共新产生了约180EB的数据；

在2011 年，这个数字达到了1.8ZB。

而有市场研究机构预测：

到2020 年，整个世界的数据总量将会增长44 倍，达到35.2ZB（1ZB=10 亿TB）！





1.2大数据的特点

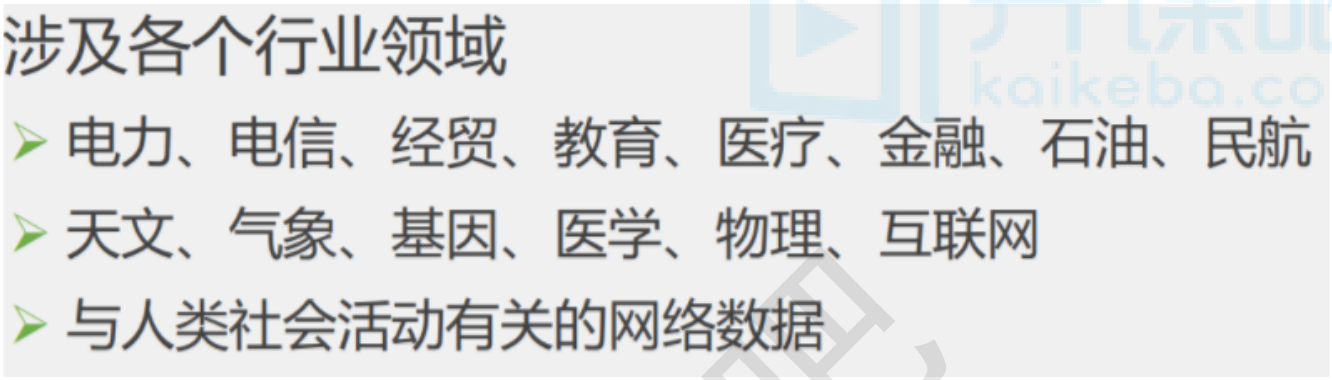
海量性

多样性

高速性

易变性

2.大数据落地行业



1.4大数据和我们生活息息相关

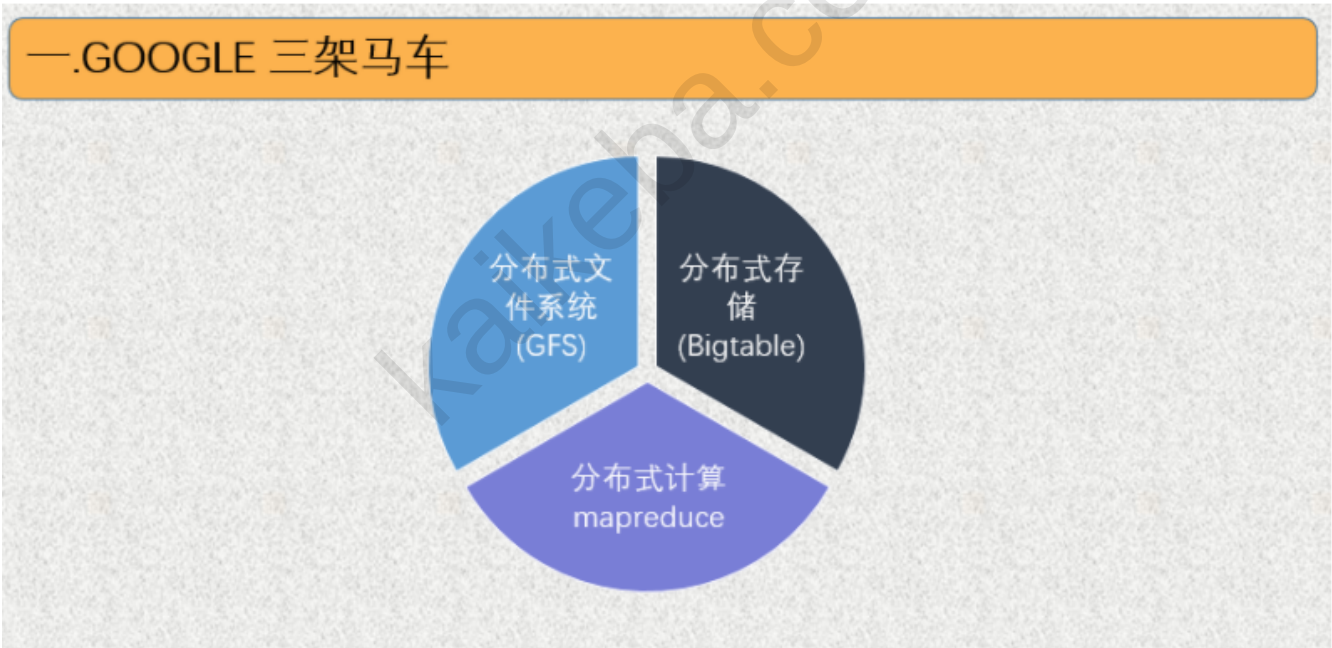
我们都是大数据的生产者

3.hadoop是什么

3.1hadoop是什么

Hadoop是一个开源的分布式系统基础框架，可编写和运行分布式应用处理大规模数据，是专为离线和大规模数据分析而设计的，并不适合那种对几个记录随机读写的在线事务处理模式

3.2hadoop来源



Hadoop之父Doug Cutting

3.1能干什么

举例：一个1T的硬盘，传输速度100M每秒，扫描全盘2个半小时，读取全部数据的话。。。。。不知道多长时间

4.Hadoop 生态系统

hadoop：分布式系统框架

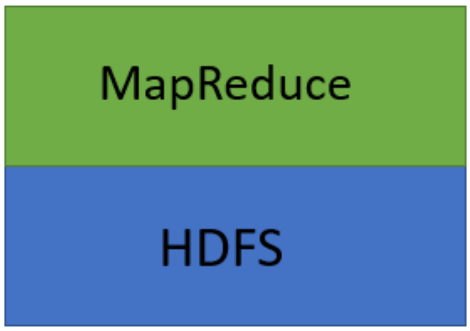
hive:数据仓库

mahout：算法库

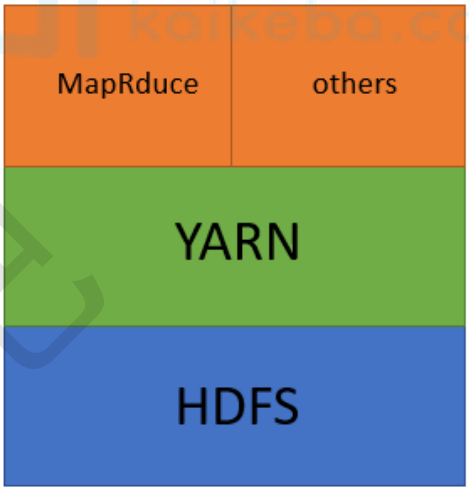
storm：实时计算框架

hbase：实时列式存储数据库

5.Hadoop架构和组件



Hadoop1.x



Hadoop2.x

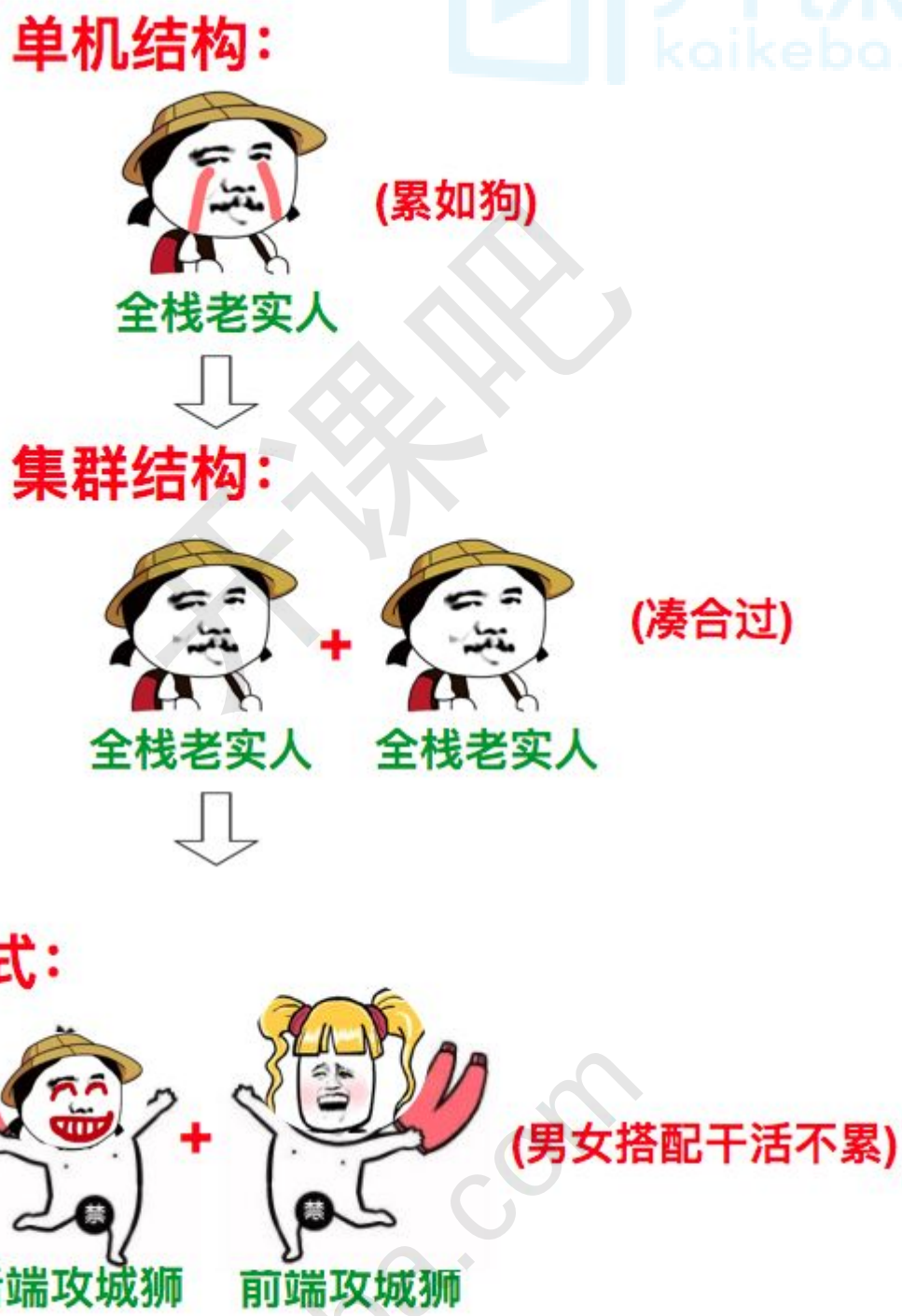
HDFS：分布式文件系统

YARN：资源调度器

MapReduce：分布式计算框架

6.HDFS 分布式文件系统

1.什么是分布式



2.什么是文件系统

文件系统是操作系统用于明确存储设备（常见的是磁盘，也有基于NAND Flash的固态硬盘）或分区上的文件的方法和数据结构，即在存储设备上组织文件的方法

我们要拿着一个小本本，上面记着，文件名，文件所在扇区以及文件大小。每次要读写文件，我们要人工查询这个账本，知道我们要的文件在哪里。如果文件A所在的扇区M已经写满了，随后的一个扇区M+1被文件B占用了，我们还想接着写文件A，怎么办呢？只能从其他地方找一个空闲扇区N，然后在账本上把N记录到文件A占用的扇区项中。

我们如何知道硬盘上还有哪些空间可以用呢？难道每次都从前往后把扇区使用情况计算一遍吗吗？可能还需要另起一个账本记录扇区使用情况，删除文件，我们把对应的扇区标记为空闲，如果创建文件，把对应的扇区标记为不能使用。

对于操作系统而言呢？我觉得，没有文件系统就不会有操作系统，这样的操作系统充其量就是一个硬盘驱动。为什么？可以设想一下创建文件的过程:

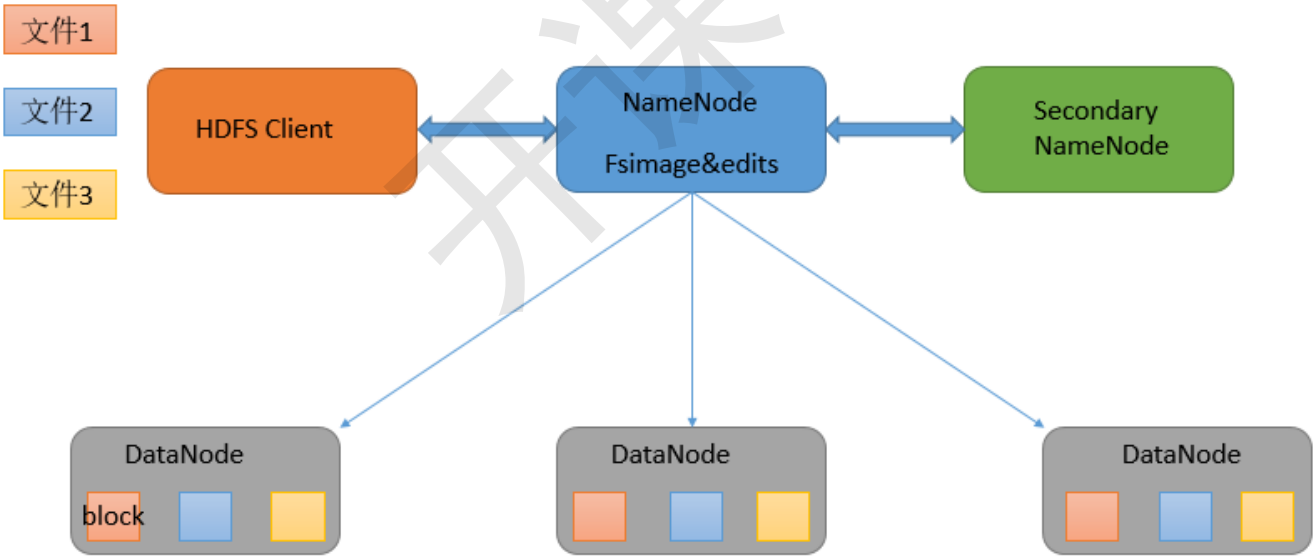
1. 用户告诉这样的操作系统，说要创建一个文件A
2. 计算机输出，请你自己记录好文件名，并告诉我要在哪个扇区创建。并且记录好这个文件你占用了哪些扇区

windows：NTFS

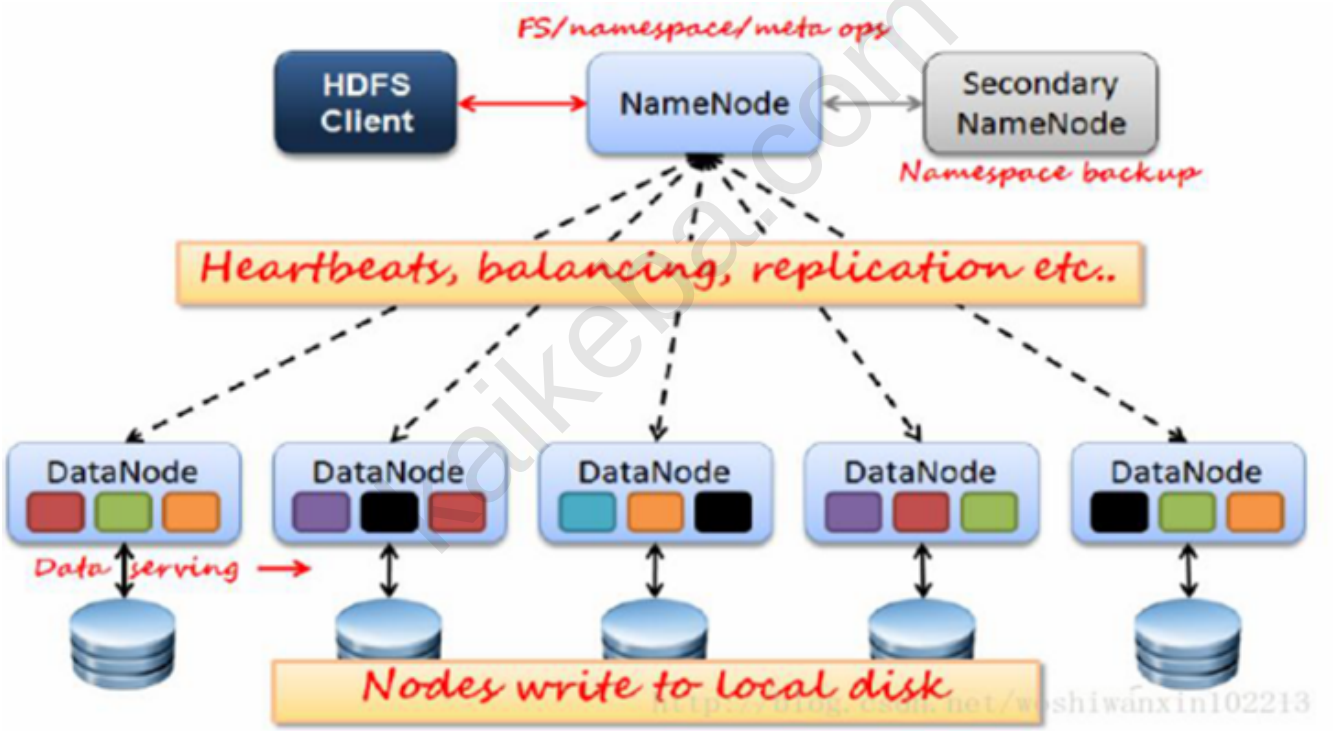
linux:EXT3

hadoop:hdfs

7.HDFS核心设计



8.HDFS体系结构



1.namenode：

接收客户端的读写请求 存储元数据信息 接收datanode的心跳报告 负载均衡 分配数据块的存储节点

2.datanode：

真正处理客户端的读写请求 向namenode发送心跳 向namenode发送块报告 真正的数据存储 副本之间的相互复制

3.secondarynamenode：

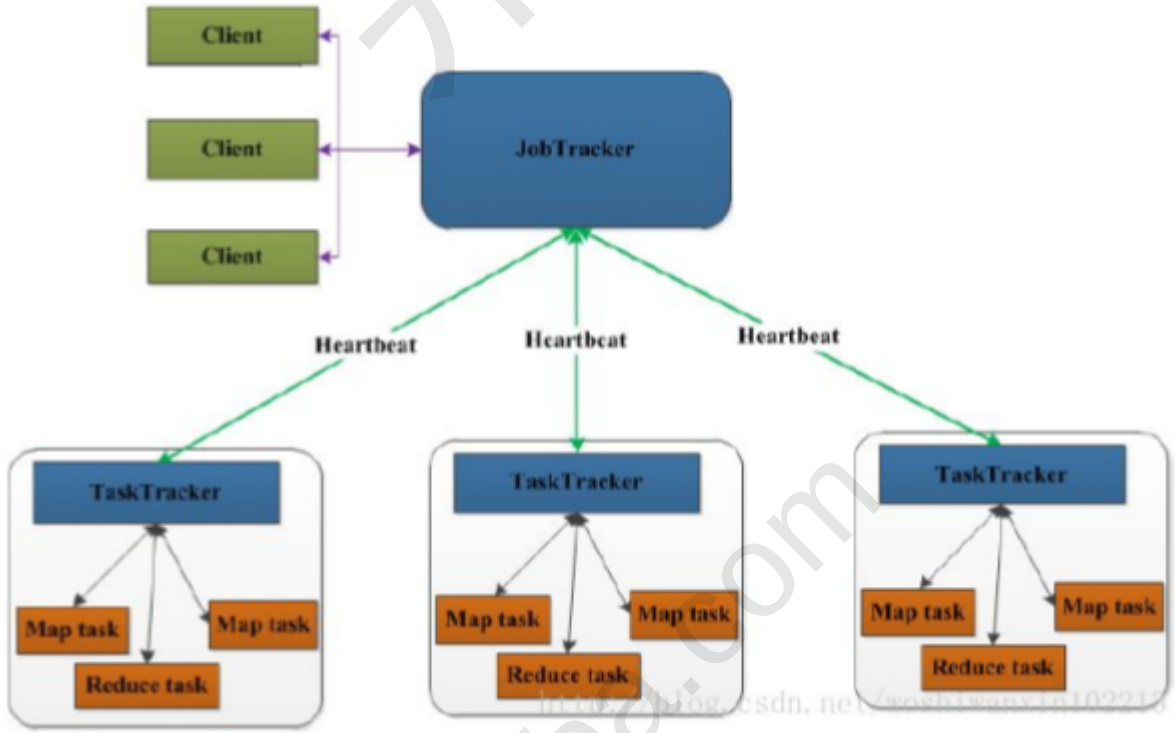
备份元数据信息 帮助namenode进行元数据合并 减轻namenode的压力

4.客户端

进行数据块的物理切分 向namenode发送读写请求 向namenode发送读写响应

9.MapReduce分布式计算框架

工作原理



resourcemanager（JobTracker）

1、处理客户端请求

2、启动或监控 MRAppMaster

3、监控 NodeManager的健康状况 nodemanager定期的向resourcemanager进行发送心跳报告

4、资源的分配与调度

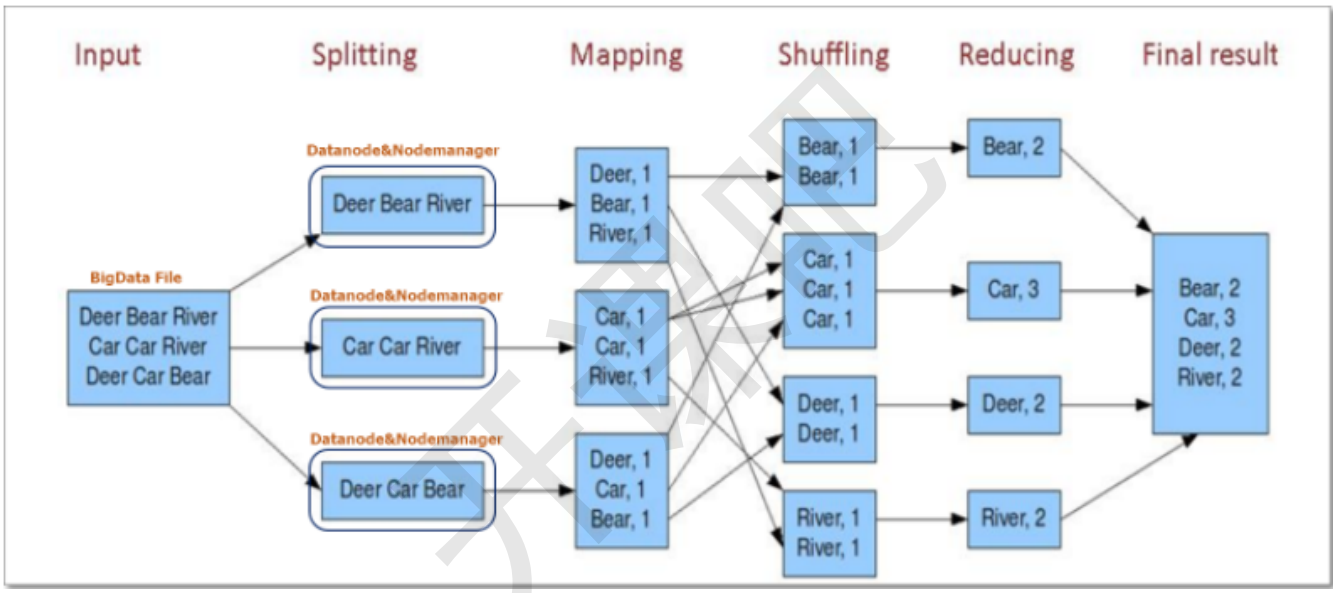
nodemanager（TaskTracker ）

1、管理单个节点上的资源

2、处理来自 ResourceManager 的命令（启动mrappmaster的时候）

3、处理来自 MRAppMaster 的命令（启动maptask reducetask任务的时候）

mapreduce编程模型



使用以下命令执行mr

先创建一个文件

vi words

```
hello world

i like java

i like java too
```

上传至文件系统

hdfs dfs -put /home/hadoop/words /

hadoop jar /software/hadoop/hadoop-2.7.3/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.3.jar wordcount /test/words /test/output

10.Hadoop 常用命令

hadoop fs所有文件系统都可以使用

hdfs dfs仅针对于hdfs文件系统

1.查看所有目录及文件

hdfs dfs -ls /

2.在hdfs文件系统中创建目录

hdfs dfs -mkdir /shell

3.在hdfs文件系统中创建文件

hdfs dfs -touchz /kkb.txt

4.在hdfs文件系统中删除文件

hdfs dfs -rm /kkb.txt

5.向hdfs上传文件

hdfs dfs -put /kkb1.txt /

6.查看hdfs上的文件内容

hdfs dfs -cat /kkb1.txt

7.从hdfs下载文件

hdfs dfs -get /kkb.txt ./

8.递归删除目录

hdfs dfs -rmr /shell

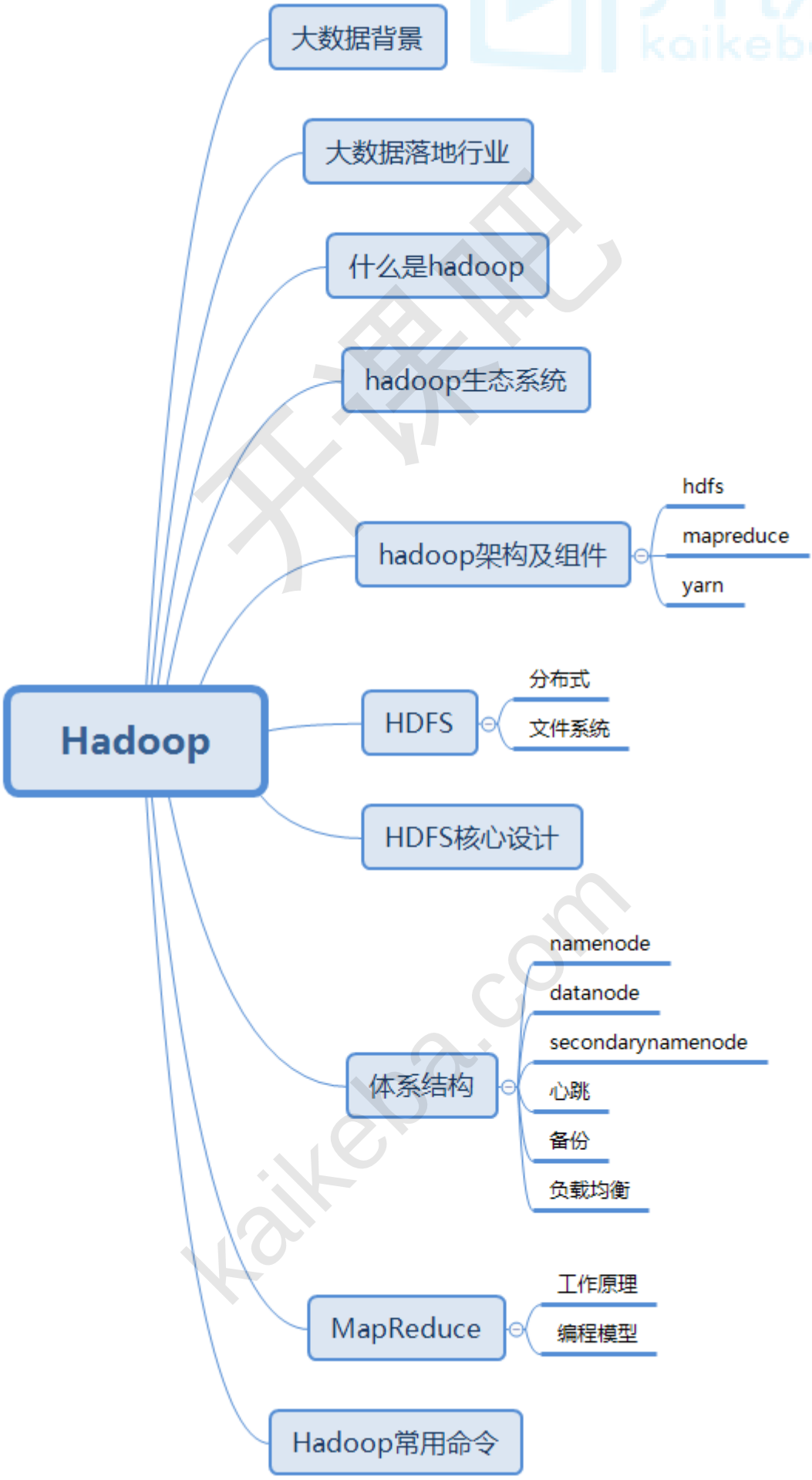
五、拓展点、未来计划、行业趋势（5分钟）

大数据学习方法：

https://github.com/apache/hadoop

http://hadoop.apache.org

六、总结（5分钟）


Hadoop
大数据背景
大数据落地行业
什么是hadoop
hadoop生态系统
hadoop架构及组件
hdfs
mapreduce
yarn
HDFS
分布式
文件系统
HDFS核心设计
体系结构
namenode
datanode
secondarynamenode
心跳
备份
负载均衡
MapReduce
工作原理
编程模型
Hadoop常用命令

七、作业

1.安装好自己的hadoop集群

2.在hdfs上直接创建文件并下载到本地

3.查看创建的文件

4.删除此文件

5.在本地创建文件并上传至hdfs

八、互动问答

九、题库 - 本堂课知识点

1.HDFS 中的 block 默认保存几份（ ）

A.3份

B.2份

C.1份

D.不确定

2.HDFS（2.0以后版本）默认 BlockSize 是（ ）

A.32MB

B.64MB

C.128MB

3.Client 端上传文件的时候下列哪项正确（ ）

A.数据经过NameNode传递DataNode

B.Client端将文件切分为Block，依次上传

C.Client只上传数据到一台DataNode，然后由NameNode负责Block复制工作

4.下面哪个程序负责 HDFS 数据存储（ ）

A.NameNode

B.JobTracker

C.DataNode

D.SecondaryNameNode

E.tasktracker

5.关于SecondaryNameNode 哪项是正确的？（ ）

A.它是NameNode的热备

B.它对内存没有要求

C.他的目的使帮助NameNode合并编辑日志，减少NameNode 启动时间

D.SecondaryNameNode应与NameNode 部署到一个节点