# 第一讲3.16

# 什么是大数据？

数据：以编码形式存在的信息载体，是真实世界的碎片化反映。

数据的常见格式：文件、记录、报告、表格、图片、歌曲、视频

大数据是指无法在容许的时间内常规的软件工作对其内容进行抓取、管理和处理的数据集合，大数据规模的标准是持续变化的，当前泛指单一数据集的大小在十几TB和PB之间。

具有数量大、增长快、类型多、价值密度低等4V特征的数据集。

Volume规模大、Velocity速度快、Variety类型多、Value价值密度低

# 大数据能做什么？

数据采集、数据存储、计算分析

智慧交通、智慧城市

# 大数据关键科学问题

## 1.数学

## 2.统计学

## 3.计算机

## 4.可视化分析问题

# 第二讲3.23

# 复习

大数据时代思考

## 一、数据-人类建造文明的基石

1.数据

2.知识

3.数据的相关性

4.数据驱动方法，数据量足够，用若干简单模型取代一个复杂模型

计算量和数据量换取研究时间

## 二、思维的革命

世界的不确定性，用信息消除不确定的问题

大数据时代思维方法：从大量数据中直接找到答案，即使不知道原因。

## 三、大数据和智能革命的技术挑战

1.数据搜集问题

2.大数据并行计算

3.大数据实时处理问题

4.数据挖掘

5.数据安全技术面临问题

6.大数据时代保护隐私的技术

## 四、智能革命和未来社会

这是一个最好的时代,也是一个最坏的时代。

# GFS分布式文件系统

## 1.1计算机集群结构

分布式文件系统把文件分布存储到多个计算机节点上，成千上万的计算机节点构成计算机集群。

## 1.2分布式文件系统结构

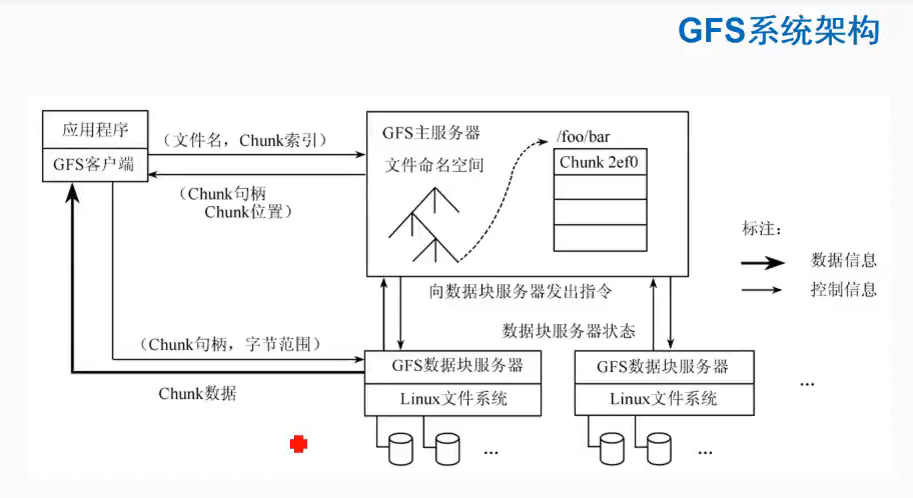
分布式文件系统在物理结构上是由计算机集群中的多个节点构成的，主要分为两类，一类是主节点（master node）或名称节点（name node），另一类是从节点（slave node）或者数据节点（data node）。

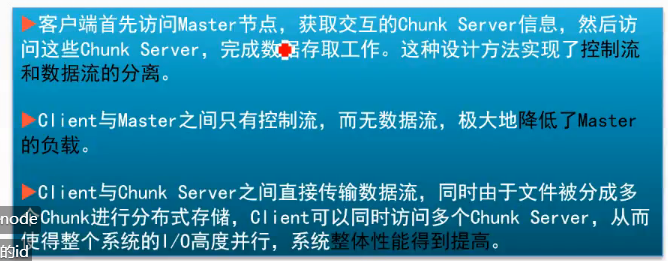
Google的核心技术

分布式基础设施：GFS和Protocol Buffer（独立数据交换的格式）

分布式大规模数据处理：MapReduce和Sawzall（类型安全的脚本语言）

分布式数据库技术：BigTable和数据库Sharding





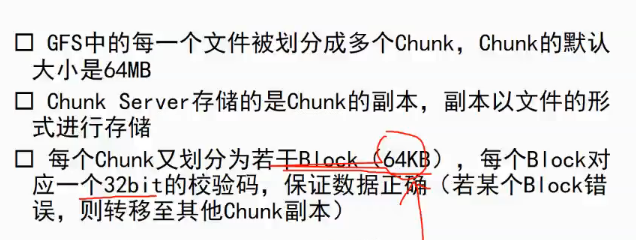
# Hadoop安装问题

## 1.JavaJDK1.8

## 2.Hadoop zip

# 第三讲3.30

## GFS文件系统



两个问题：

1、hadoop安装问题

2、HDFS编程问题——实现客户端调用

Shell，Java APP

3、MapReduce 编程

## 面向对象程序开发Java

## 1、JDK，JRE，JVM

JDK：Java开发工具包

JRE：Java运行环境

类库

JVM：Java虚拟机

两次编译

Hello.java🡪Hello.class🡪JVM解释执行（机器指令）

## 2、Eclipse+JDK8

完全面向对象

# 第四讲4.13

JDK8比较稳定，免费

CMM软件成熟度模型1-5

## 3.面向对象编程技术OOP

### 类：

现实世界🡪计算机世界🡪可以编程（完成系统，APP）

### 实体：

学生、教师、课程、……

=属性+操作

学生（学号、姓名、专业、入学年份、……、注册、选课、修改信息……）

类 具有相同属性和操作这样一类实体

学生类=学号、姓名、专业、入学年份、……、注册、选课、修改信息

### 如何在Java定义类？

Java类=属性变量+函数

Package cn.edu.shiep;

Public class Student{

Private int id;

Private string name, major;

Private int year;

Public void initInfo(nt sid, String sname, String smajor, int syear){

Id=sid;

Name=sname;

Major=smajor;

Year=syear;

}

Public void dispInfo(){

System.out.println(id+” ”+name+” ”+major+” ”+year)

}

### 类如何使用？

通过 类创建或定义对象，通过对象调动其中的函数或属性

Student s1,s2,s3; //三个对象，本质上就是变量，变量类型是自定义的

S1.initInfo(20190801,”崔荣成”,“电力信息技术”，2019);

类实例化，其实给类的属性和方法分配存储地址和函数入口地址，过程称为实例化。

S1=new Student(); //实例instance，对象s1引用了实例

S1.initInfo(20190801,”崔荣成”,“电力信息技术”，2019);

S2=s1；

S2.display();

### 2.类的构造函数

实例化为属性赋值

### 特殊函数：

特征特点？ 与类同名，没有任何返回值类型，可以有0到多个形参，一个类可以有多个构造函数

目的与作用？

什么时候调用？ 类在实例化的时候自动调用。

### 构造Point类

# 第五讲4.21

## 3.类继承

提高代码的复用性

父类（基类，超类）

子类

Eg：

Public class A{

Private int x;

Public int y;

Protected int z;

Public void fun1(){…}

Private void fun2(){…}

Protected void fun3(){…}

}

Public class B extends A{

…

}

父类中的public及protected成员（函数、属性变量）可以被子类继承，private成员是不可以继承。

缺省访问修饰符

个人、公众、家庭、组织内部（同一包类）

子类实例化；构造函数调用顺序？

## 继承本质

## 4.Final类

Public final class A{

…

}

不允许继承

Math，String——final类

Public class TestFinal{

Public static final doublePI=3.14;

}

## 5.函数重写（override）

Public class A{

Public void fun(){

…

}

}

Public class B extends A{

}

B obj=new B()；

Obj.fun();

## 6、抽象函数及抽象类

Public class Shape{

Public abstract double area();

Public void display(){

System.out.println();

}

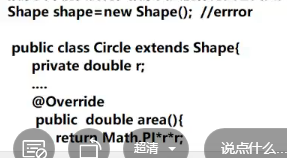
}

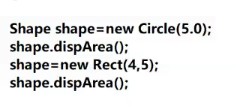
抽象函数：要去做什么，但没有告诉如何做；

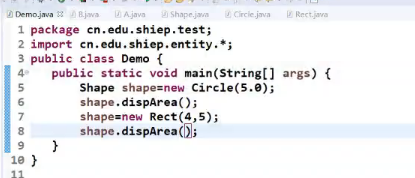
抽象函数功能由子类具体实现，不同子类抽象函数实现代码不同（方式不同），多态

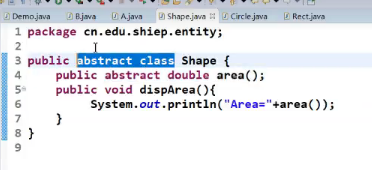
面向对象编程：封装、继承、多态

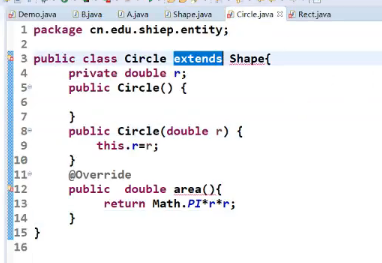
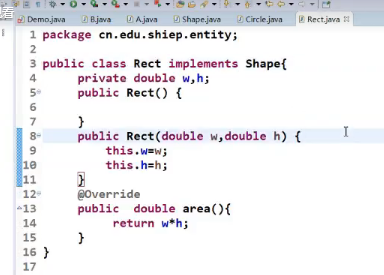
抽象类不能实例化；抽象类只能作为其他类父类，抽象基类







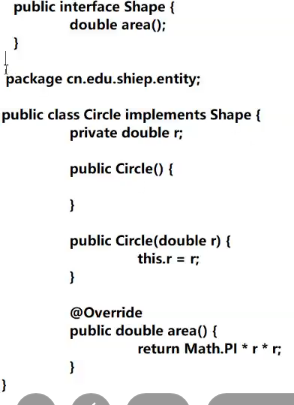


## 7.接口interface

特殊java类，接口中函数都是抽象

public interface Shape()



Java 继承都是单一继承，一个直接父类最多最能一个

Java 类可以实现多个接口

Public class Loginfrm extends JFrame implements ActionListener{

}

Swing界面设计

## 8.其他JAVA核心内容

1）Java界面Swing 了解java设计图像界面；事件处理流程等

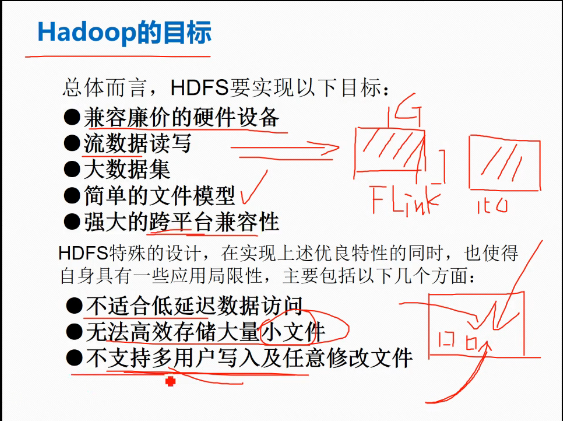
2）Java数据库JDBC开发

3）Java集合框架及集合Stream运算

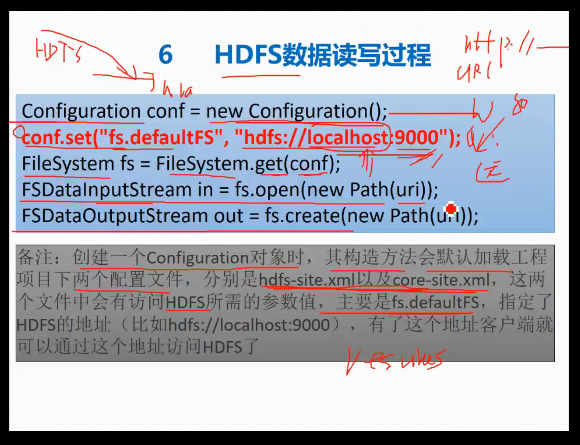
ArrayList，HashSet，TreeSet

# 第六讲4.26

## Hadoop简介

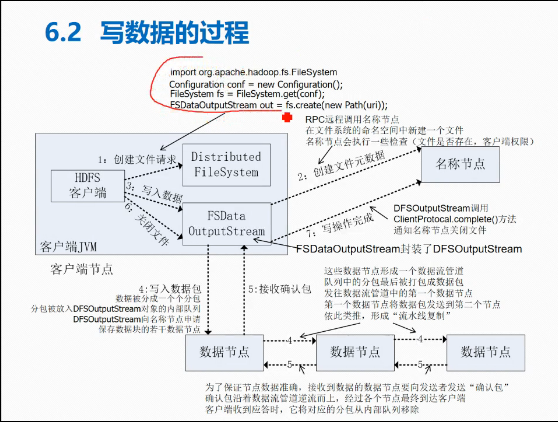


# 第七讲4.27



## 读数据

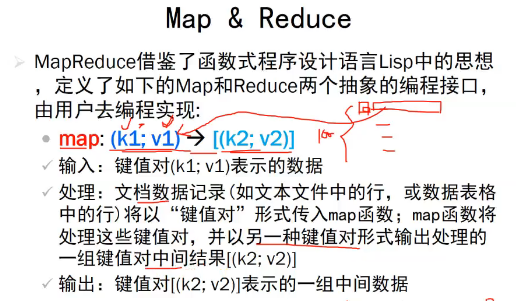
## 写数据

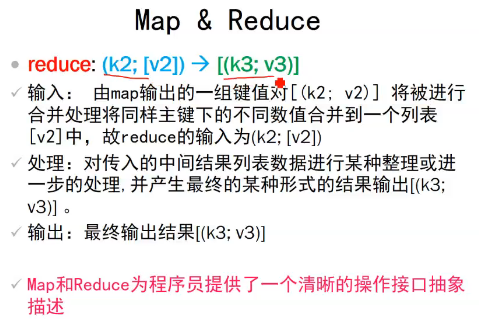


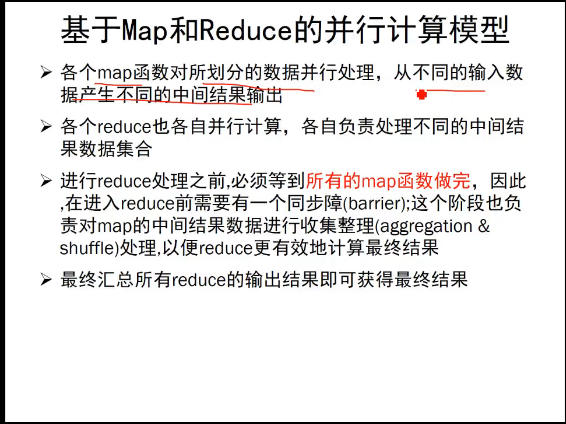
# 第八讲5.11

## HDFS常用Java API及应用实例

## MapReduce











# 第九讲5.18

MapReduce并行处理的基本过程



## MapReduce案例：

### 数据去重

### 数据排序

# 第十讲5.25

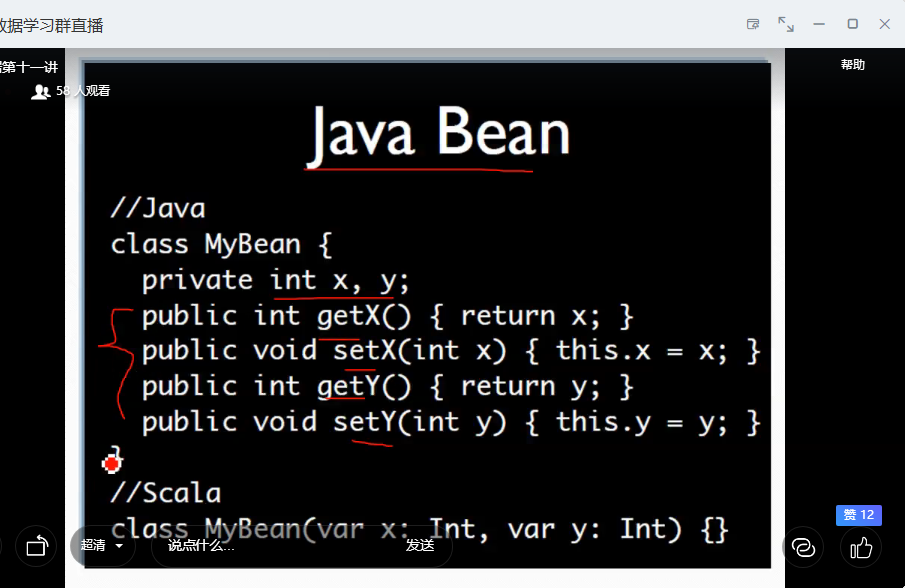
## Spark

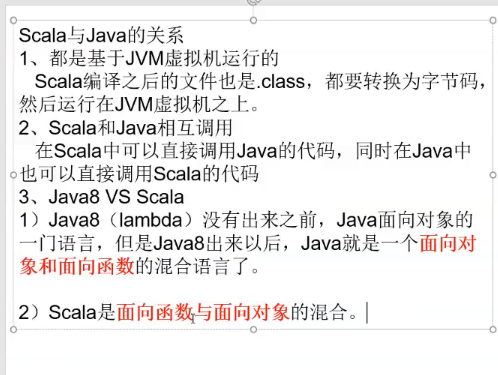


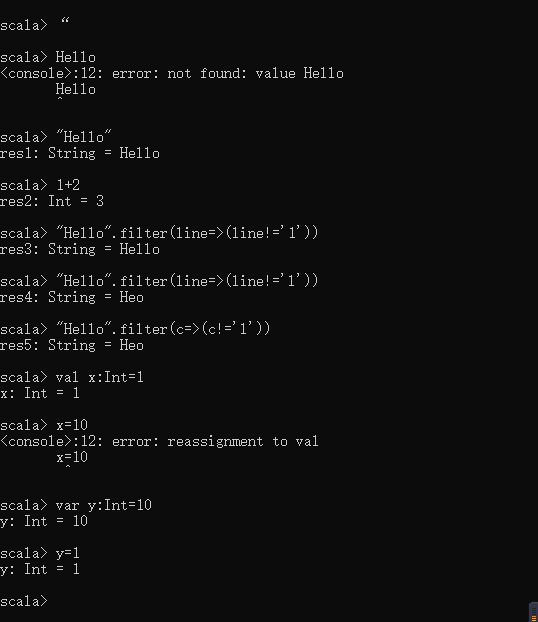
# 第十一讲6.1

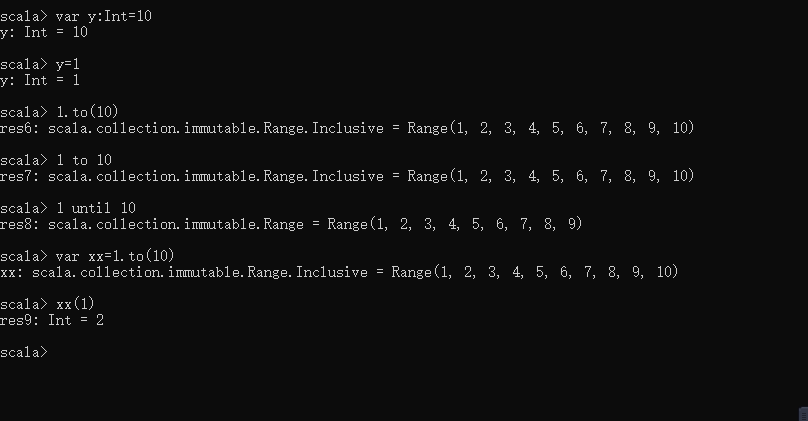
# Scala

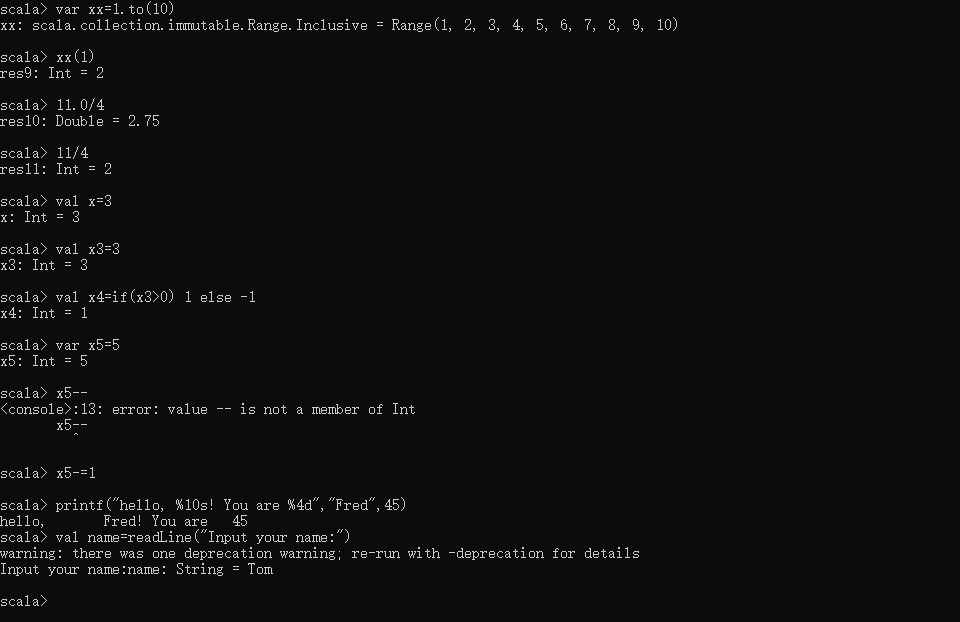










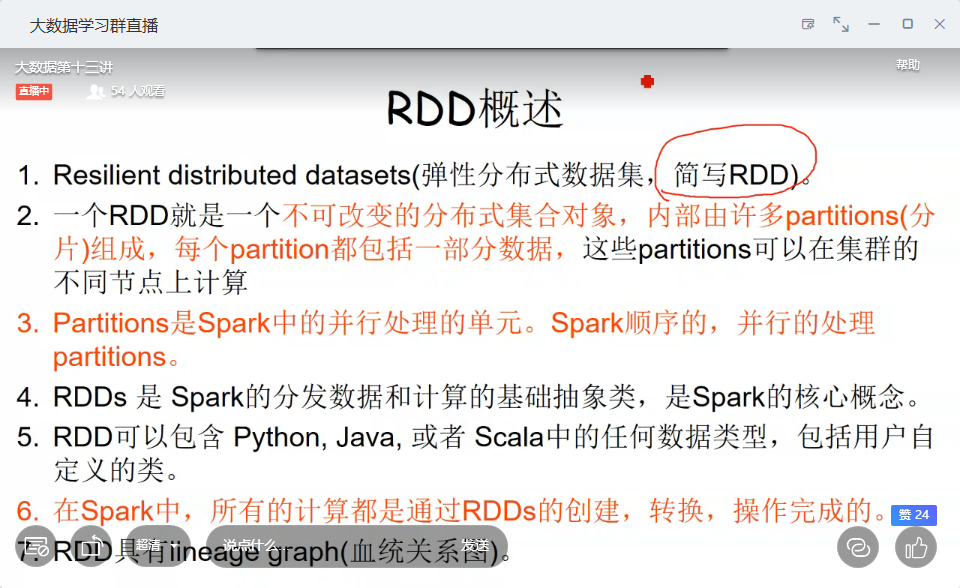


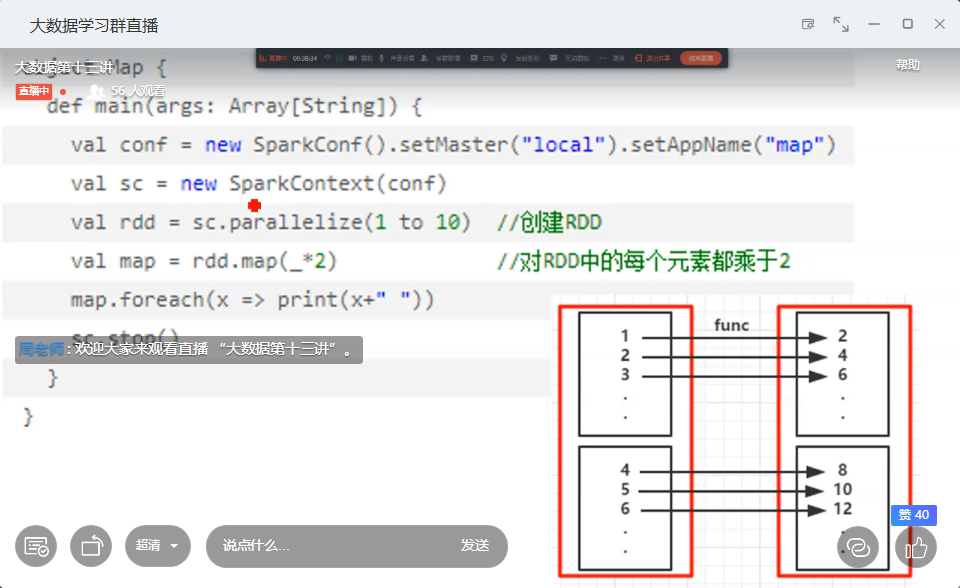
# 第十二讲 6.8

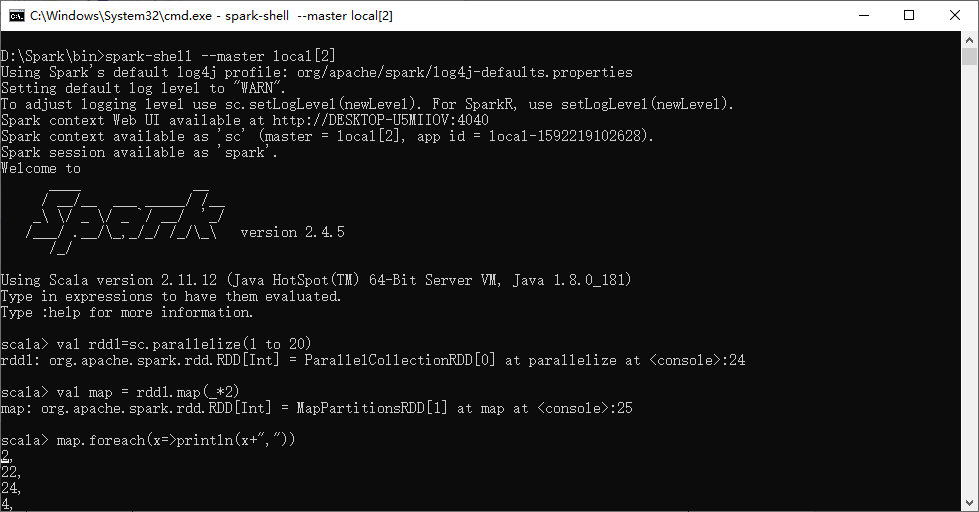


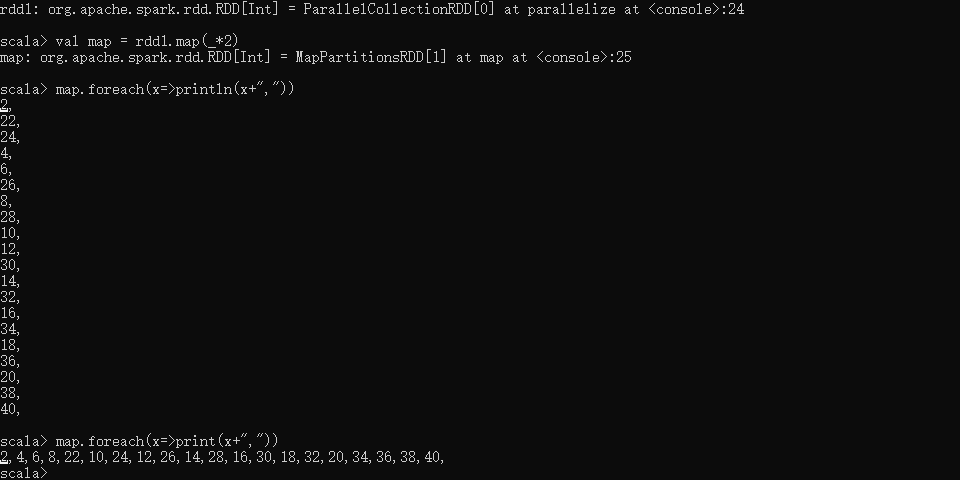
# 第十三讲6.15

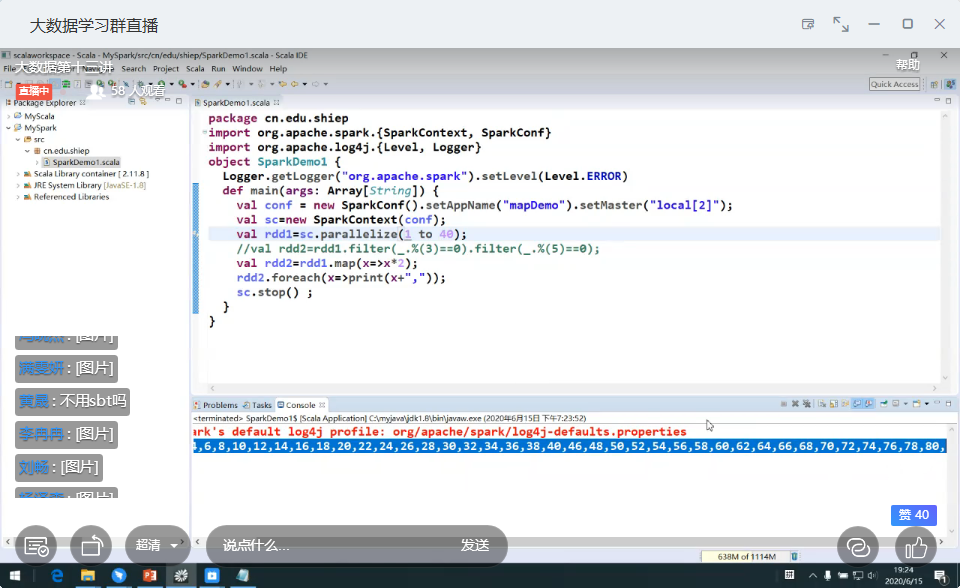
## RDD

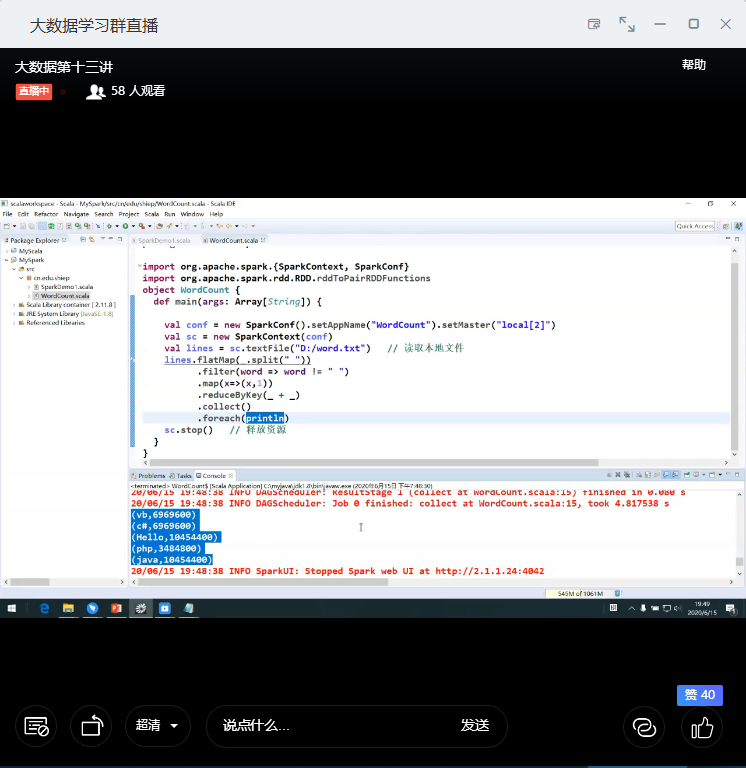




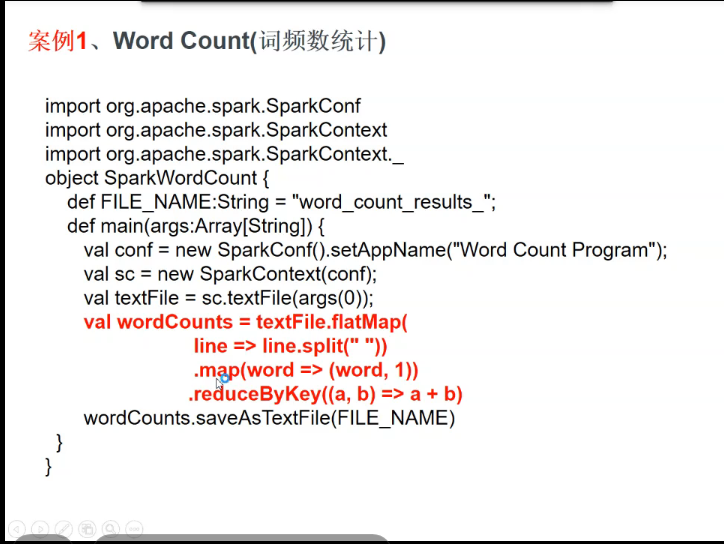








# 第十四讲6.22



# 大数据技术原理大作业

## 1、总评成绩计算来源

平时成绩（直播观看次数及时长，网上视频观看次数及时长，实验完成情况）60%

期末成绩（课程报告，大数据期刊论文翻译（技术、理论））40%

## 2、需要提交文档

1）实验报告4篇（带有封面，命名：学号\_姓名\_实验一.docx）

2）课程报告（带有封面，命名：学号\_姓名\_大数据课程报告.docx）

3）论文（带有封面，命名：学号\_姓名\_论文（中文）.docx；学号\_姓名\_论文（英文）.docx）

## 3、课程报告内容