# 20200704第2课

## 模板设计模式

## HDFS API 入口类 FileSystem

# 20200705第3课

## 引入MapReduce

## 引入Shuffle

相同的key分发到同一个reduce进行处理，就是shuffle

## MR中常用的数据序列化类型

## 词频统计、八股文编程

## Debug Hadoop源码

# 20200708第4课

## 序列化和反序列化

内存的对象转成字节数组，以便于存储或者网络的数据传输

## 自定义序列化类

## InputFormat

### getSplits()

### createRecordReader()

## 切片的源码讲解

## KeyValueTextInputFormat(不常用)

## 作业

Debug源码找到Hadoop在LocalFileSystem中的blockSize为32M的源码

# 20200711第5课

## NLineInputFormat(不常用)

## DBInputFormat

## 自定义Partitioner

## Combiner

## 自定义OutputFormat

## 作业

将MR读取MySQL的代码在服务器上跑起来，这里的MySQL驱动包是需要通过****-libjars****传入的

具体查看：

<https://github.com/apache/hadoop/blob/a89ca56a1b0eb949f56e7c6c5c25fdf87914a02f/hadoop-common-project/hadoop-common/src/main/java/org/apache/hadoop/util/GenericOptionsParser.java>

20200712第6课

1. 排序
   1. 全排序

保证输出是全局有序的;只有1个reduce; Hive中的order by就是全排序

* 1. 分区/局部排序

保证每个reducer的输出是有序的; Hive中的sort by

* 1. 二次排序

1. 备用