

# 华北电力大学



## 大学生创新创业训练计划项目中期检查报告书

项目类别: 创新训练项目

项目名称: 基于大数据的用户播放行为监测 HTML5 播放器设计

立项时间: 2019 年 6 月

项目负责人: 赵鸿至

学院年级专业: 控制与计算机工程学院计算 17 级

联系电话: 18810807992

指导教师: 张至柔、吴娟

华北电力大学教务处

填表日期: 2019 年 11 月 10 日

## 一、项目研究进展情况

### 1、成员分工：

王一名：使用 Spring 框架做项目开发，并用 HTML5 开发播放器，用 JavaScript 实现用户行为的抓取，使用 ajax 方法将前端数据传输到服务器，并在 Hadoop 框架上实现 MapReduce 方法。

赵鸿至：辅助开发播放器界面，设计 Hadoop 数据处理程序的算法并实现接口。

### 2、项目研究方法：

使用 HTML5 制作网页播放器，使用 JavaScript 实现对用户视频观看行为的记录，将记录的数据转换为 JSON 格式，使用 ajax 方法发送请求将 JSON 文件写入 HDFS 中，之后使用 java 语言在 Hadoop 框架上，用 MapReduce 的方法对用户行为的数据进行计算、积累，获得每个视频以秒为单位的用户累计观看次数，并做成图表。当有用户观看该视频时，将最新的用户观看行为数据以曲线的形式显示在播放器进度条上，给当前用户作为参考。

### 3、项目研究内容：

1) 设计并开发 HTML5 播放器，使用 JavaScript 来记录用户对视频播放器的操作行为如播放、暂停、快进快退，调整倍速等。

2) 使用 ajax 方法将前端获取的数据以 JSON 格式发送至服务器 HDFS 中，并且解决短时间内传输数据并发量大导致可能溢出的问题。

3) 设计 Hadoop 数据处理程序的算法，实现 MapReduce 方法，将字符串数据转化为数组的形式，并将 key 值相同的数据累加，将结果输出到 HDFS 中。

4) 将视频播放情况数据响应给用户的网页端，并使用 JavaScript 将最终的数据以曲线显示在播放器页面中。

### 4、项目研究进度：

项目组已经完成了从播放器页面获取数据，将其传输到 HDFS，运行程序并输出结果到 HDFS 的整个业务逻辑。并且实现了 Hadoop 运算程序的分布式运行。

目前项目组正在研究如何使整个业务逻辑实现动态多线程并发处理以提升总体的运算速度，以及设置定时定量缓冲的机制。

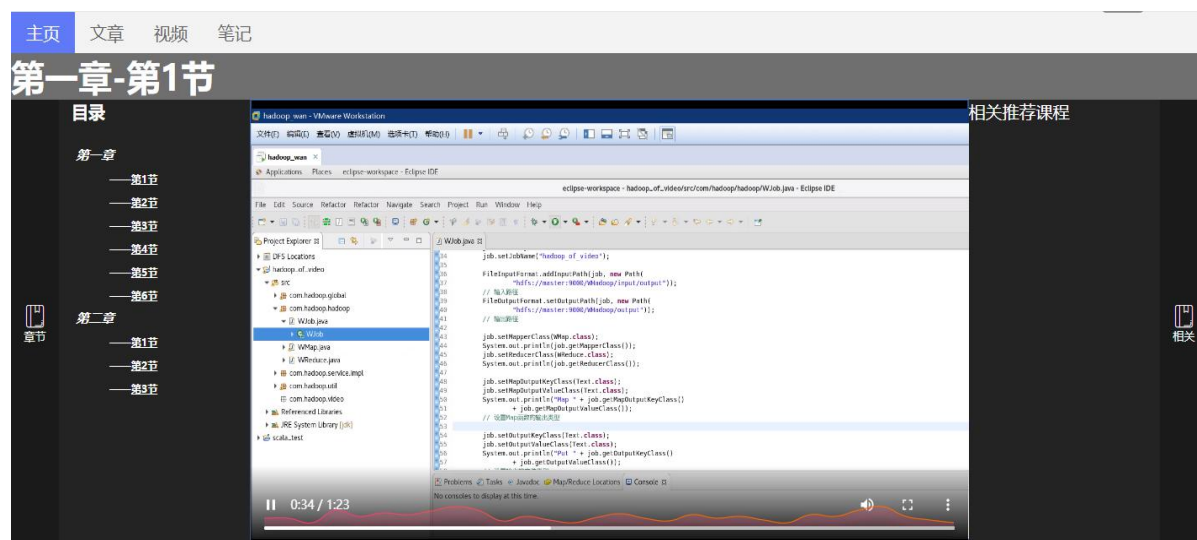
未来项目组将会研究前后端数据传输接口；完善播放器的网页制作，视频播放情况数据的展示；并模拟真实的应用环境，调整程序，使其更具有实用性。

## 二、已取得的阶段性成果（已取得的阶段性成果或收获等）

### 1、播放器页面

HTML5 播放器已经制作完成，已经使用 JavaScript 实现了抓取用户行为数据的功能。

播放器界面如下：



### 2、播放器抓取数据并传输的代码

```
my_video.addEventListener(“mousedown”, mouseDown);
my_video.addEventListener(“mouseup”, mouseUp);
function mouseDown() {
    调用 getSeekStart() 方法
}
function mouseUp() {
    调用 getSeekEnd() 方法
}
function getSeekStart() {
    获取当前视频帧数
}
function getSeekStart() {
    利用 ajax 向服务器上传请求
}
```

### 3、ajax 通过 post 方法发送请求代码

```
video_request.open( ‘POST’ ,window.location.href,true);
video_request.setRequestHeader( “Content-Type”, ” application/x-www-fo
```

```
rm-urlencoded" );  
    video_request.send(param); //发送 request, 并传输 JOSN 文件 param
```

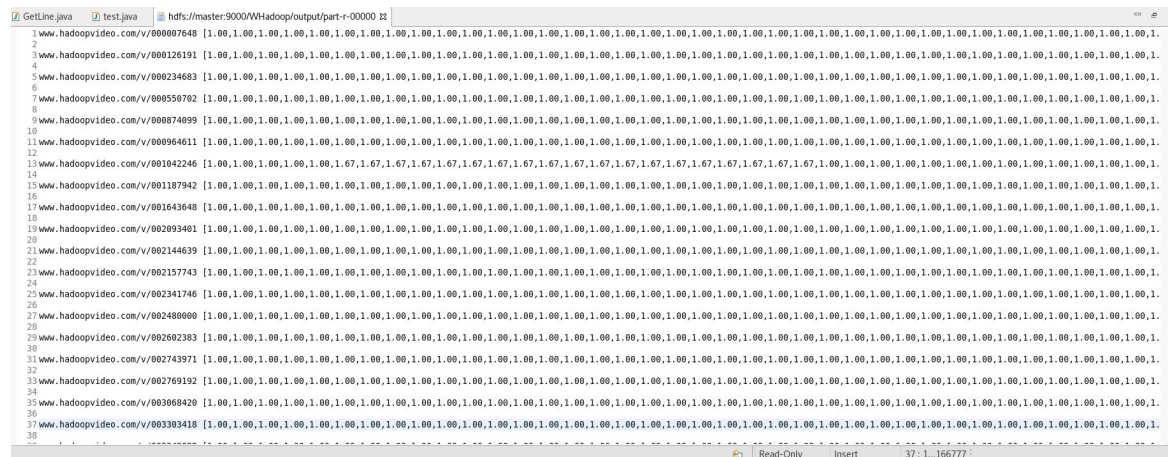
#### 4、在客户端成功使前端抓取 JSON 格式的数据再写入到 HDFS 中

在当前页面获取 request 并调用 getLine(video)

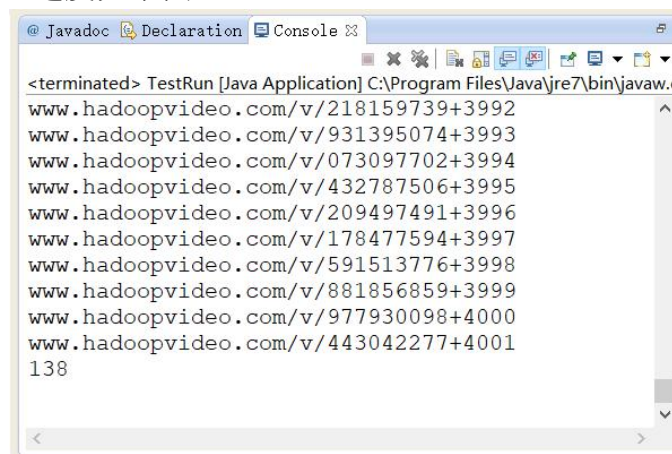
```
getLine(Video video) {  
    将 video 放入本地  
    if(file.length()>16777216) {  
        将 video 放入 hdfs 中  
        并删除本地文件  
    }  
}
```

#### 5、完全分布式后的性能提升

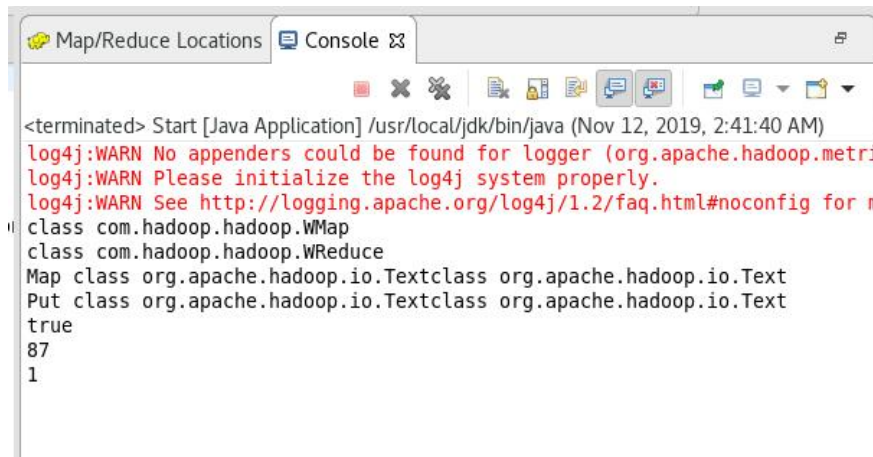
以处理一个 24M 的文件为例：在我们的项目中输出的结果如下图



在串行中的处理速度如下图

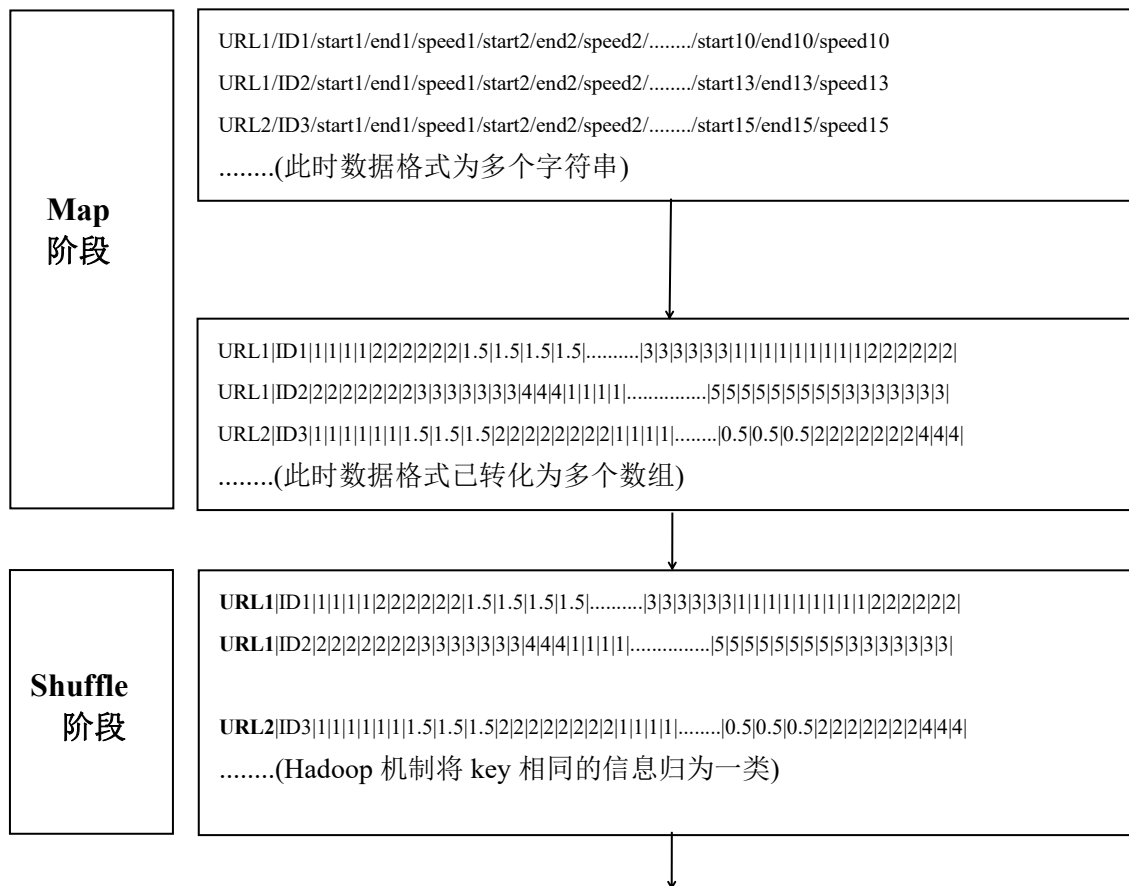


在 hadoop 框架中应用分布式计算处理速度如下图
----------------------------



可见处理速度有了质的提升

## 5、算法设计



## Reduce 阶段

```
URL1|ID1|1|1|1|1|2|2|2|2|2|1.5|1.5|1.5|1.5|.....|3|3|3|3|3|1|1|1|1|1|1|1|1|2|2|2|2|2|
+
URL1|ID2|2|2|2|2|2|2|3|3|3|3|3|3|4|4|4|1|1|1|1|.....|5|5|5|5|5|5|5|5|3|3|3|3|3|3|
=
URL1|3|3|3|3|3|4|4|4|4|5|5|4.5|4.5|4.5|5.5|5.5|4|4|4|4|.....|8|8|8|8|8|6|6|6|6|5|5|5|5|5|

URL2|ID3|1|1|1|1|1|1|1.5|1.5|1.5|2|2|2|2|2|2|1|1|1|1|.....|0.5|0.5|0.5|2|2|2|2|2|4|4|4|
+.....(将 key 值相同的 value 累加在一起)
```

Map 阶段：各个节点将前端提交的每一条用户行为数据，按照多段起始时间及对应倍速，对数组进行加权计算，从而得到多个观看次数数组，数组的 key 值为视频 URL，value 值为某个用户观看该视频的过程中，对每秒视频的根据观看倍速进行加权的数值。

Shuffle 阶段：MapReduce 的关键环节，将 key 值相同的信息分到一类，并交给同一个 Reducer 处理，从而对信息进行整理和分类。

Reduce 阶段：将 key 值相同的信息中的 value 值累加在一起，更新每个视频所有用户观看行为的总记录。

主要成果					
成果名称	形式	参与者	发表（出版）情况		
			发表时间	发表刊物（出版部门）	字数

### 三、目前存在的主要问题及应对措施

问题 1：实际应用中，网页很可能在短时间内收到巨大的点击量，数据并发量过大，从而导致数据的溢出和丢失。

措施：在服务器端处理程序中，准备采用 Spring 框架，用以处理大并发量的问题，以提高效率和保证数据的完整性。

问题 2：在整个程序流程中，数据到文件系统 HDFS 的 I/O 传输占用时间的比率较大。

措施：在文件传输的时候，准备利用 Hadoop 框架，分布式传输文件。这样可以减少 IO 传输的时间。

### 四、下阶段主要计划及时间安排

- 1：将单线程的运行逻辑全面转换到完全分布式的多线程并发式运行逻辑。
- 2：解决前端短时间内传输的数据并发量大导致可能溢出的问题。
- 3：将视频观看行为记录数据以 JSON 格式传输到网页端，并用 JavaScript 编程，将数据以曲线的形式，形象地显示在播放器进度条上。
- 4：撰写论文
- 5：申请软件著作权

经费使用情况		
报销经费包括：实验费、材料费、加工测试费、资料费、打（复）印费、交通费等支出		
序号	支出项目	金额（元）

合计			
项目负责人签字：		年 月 日	
<b>指导教师鉴定意见</b> （从研究内容和进展、阶段性成果、存在问题和建议等加以评价） <p>项目研究进展正常，目前已完成播放器界面设计，前端用户行为抓取、MapReduce 数据处理程序的初步编写，进度达到预期，下一步将实现完全分布式计算程序，前端播放器界面以及视频播放情况数据展示程序的编写，并在程序完成后，撰写论文，申请软件著作权。</p> <p style="text-align: right;">签章：2019 年 11 月 19 日</p>			
<b>学院意见</b> （请给出评审意见及评定成绩） <p style="text-align: right;">签章：年 月 日</p>			
<b>学校专家组意见 1</b> <p style="text-align: right;">签章：年 月 日</p>			
中期检查成绩评定		优      良      通过      未通过	