Hadoop大数据处理用户行为痕迹记录的

创新应用项目研究报告（3000字左右）

项目团队成员：120171080212、杨秉学、计算1702

120181090209、刘俊龙、水电1802

指 导 老 师 ：张至柔、网信办

吴 娟、网信办

项目起止时间：2019年4月-2020年6月

摘要：随着计算机和网络应用的广泛深入，网络教学成为教育领域一个重要的组成部分，针对当前网络视频存在老师与学生之间交互性、实时性、反馈客观性等方面的缺陷，提出一种基于Hadoop，实现了对学生观看视频全过程的行为监控、记录和反馈，客观地分析出该视频的重难点所在、学生的掌握情况等信息，辅助教学双方。

关键词：大数据、网络视频教学，时间轴，用户行为监测与反馈

一、项目研究背景及意义（黑体，四号）

（内容：宋体，小四，行距为1.25倍行距）

随着互联网的发展，互联网越来越贴近人们的生活，人们的衣食住行，娱乐与学习等方方面面都有着互联网的身影。依据国务院印发的《“十三五”国家信息化规划》，我国提出了大数据战略的重大决策，开启了信息化发展的新征程。教育部在2018年4月发布的《教育信息化行动计划2.0》也明确提出了促进信息技术在教育领域的广泛应用，推动教育的改革和发展，培养适应信息社会要求的创新人才以及促进教育现代化的目标。

但由于视频学习方式本身的局限性，传统网课需要等全部播放完才能获得用户的反馈意见，很多学生在通过视频自学的过程中并不了解课程的难点和重点，导致虽然看视频学了很长时间，效果却并不理想。而进行视频教学的老师，也无法了解学生们掌握的情况。因此，为了顺应时代的潮流、响应党和国家的号召、有效处理用户在进行视频学习过程中产生的大量行为数据，从中获取用户观看的教学视频的难点、重点，反馈给视频提供方和教师，促进他们对教学内容和方式的修改，从而提高网络教学质量。

每个用户在观看教学视频时都可能根据自己的需要和已掌握的相关知识重点看自己需要的部分，跳过不需要的部分，即对视频进度条进行向前、向后拖动或倍速播放，这就形成了观看视频时的用户行为，产生大数据分析的“滤镜效应”，定位观众的热点，通过后台的服务器自主进行计算获得用户在观看视频的用户行为数据，由于观看视频的用户数量巨大，这种行为数据的量也极大，因此数据的处理对计算、存储的要求很高。Hadoop是对大量数据进行分布式处理的软件架构，包含了当前主流的大数据处理技术，适合作为对用户行为数据进行计算、存储、管理的平台，因此我们在Hadoop平台上研发了教学视频的用户行为处理系统，该系统可将前台（视频播放器）提交的用户观看视频的行为数据通过计算转换为每个视频中每秒视频片段的播放次数统计，并存储的Hadoop文件系统HDFS中，作为该视频的播放情况记录，在前台需要时，这些数据将以曲线形式展示到播放界面上，作为新用户或视频提供方的参考，客观展示教学视频中的重点和难点。

二、研究内容及实施方案

1、研究内容：

随着计算机和网络应用的广泛深入，网络教学成为教育领域一个重要的组成部分，针对当前网络视频存在老师与学生之间交互性、实时性、反馈客观性等方面的缺陷，提出一种基于Hadoop，实现了对学生观看视频全过程的行为记录和反馈，客观地分析出该视频的重难点所在、学生的掌握情况等信息，辅助教学双方。

具体内容：

1）规划以及部署Hadoop，在实际应用需求、服务器的配置条件、Hadoop中各节点数量之间寻找平衡点；

2）根据Hadoop并行计算框架优缺点和应用程序的需求，编程实现；

3）模拟大量用户并发提交环境，使程序能在Hadoop平台上高效运行并得出所有用户观看视频时操作行为的汇总数据，最终获得所有用户对某个视频观看行为的汇总数据，写入文件系统HDFS的一个汇总文件中。

4）使用压缩算法，提高数据传输效率。

2、实施方案：

1）根据实际的应用需求与服务器属性配置调整优化Hadoop中节点的部署，通过Java编写Job函数与Map函数和Reduce函数分布式处理用户行为痕迹记录数据；

2）模拟前端不同速度提交用户行为数据，尤其是大并发量提交的情况下，动态调整Map和Reduce的数量，提高计算节点的适应性和运行效率，从而使应用程序既能在此环境下快速有效运行得出所有用户观看视频时操作行为的汇总数据，又不会过于频繁启动，造成资源浪费；

3）测试计算节点的用户行为数据处理速度，调整计算节点处理数据分片大小，增强程序适应性。

三、研究成果及结论

1、系统研究结果

测试运行时使用的系统环境的Hadoop节点部署为1个master节点、2个slave节点，其中slave1,slave2均为DataNode节点，形成分布式存储数据的HDFS文件系统。前端播放器提交的JSON格式的用户行为数据由系统发送至HDFS中，之后运用Java语言编写Job函数与Map函数和Reduce函数分布式处理用户的倍速播放的片段起始、结束位置和播放速度，及前进、回退播放的片段起始、结束位置行为痕迹数据，获得每个视频以秒为单位的播放次数结果，将计算结果形成文件并进行压缩，存储到HDFS中，成为某视频播放情况记录文件，并在前端提出请求时反馈给前端。

2、算法研究结果

**反馈给前端**

**模拟前端数据**

**HDFS**

**缓冲区**

**Shuffle阶段**

**Map阶段**

**Reduce阶段**

图1 程序总体流程设计图

(1)、模拟前端发送的数据

这里模拟前端发送的数据传输格式，直接写入到生成文件，一条URL数据的格式“视频URL-用户id-视频片段长度-片段开始时间-片段结束时间-观看倍速”

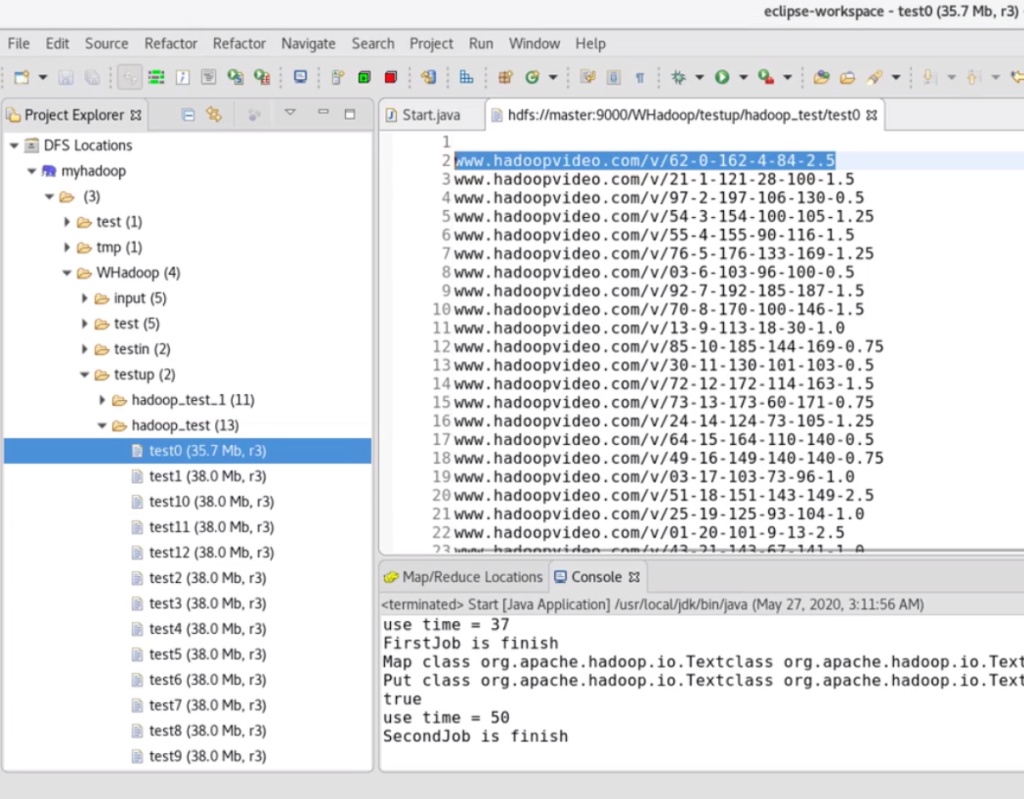


图2 数据文件图

数据文件如图，共870M大小的总数据，共19800792条如图3所示,采用内存溢写的方式进行划分，分成了24个数据文件，这些数据文件被分别存放在了图中左侧部分内的hadoop\_test\_1（11块数据文件）与hadoop\_test（13块）两个文件夹内。其中每个数据文件的格式内容如图2中右侧大框内所示。

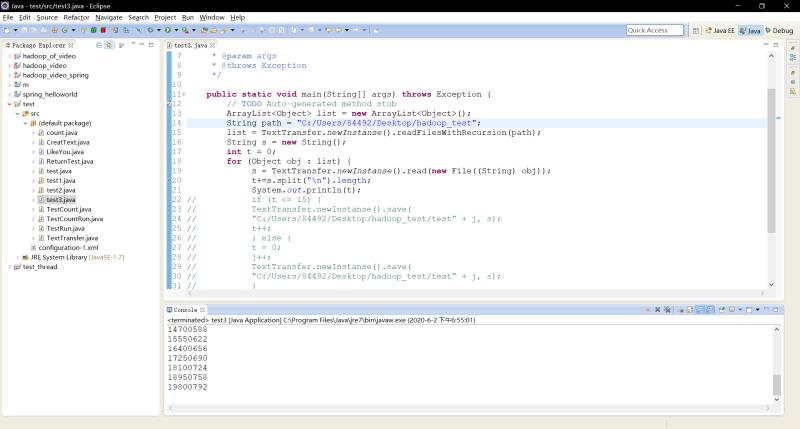


图3 模拟的总数据条数

(2)、MapReduce数据处理算法

在Hadoop平台对大量用户行为数据进行分析与处理，设计MapReduce算法将不同用户观看不同视频的用户行为数据装入多个MapTasks里，本项目设置了2个Map，由Map将每一条用户行为数据转化为该用户观看某个视频中以秒为单位的视频片段次数，形成一条记录，这些记录再根据视频的URL分配给多个Reduce，通过视频的数量来决定partion数量，100个视频对应100个partion,由partion设置了100个ReduceTasks,由这100个ReduceTasks将这些记录合并计算，得到各视频以秒为单位的总体播放情况，并用这些数据与HDFS中存储的该视频原总体播放情况数据累加，更新该视频总体播放情况数据。

例如本项目中一个文件经过Map后得到的一组视频id为00的视频的用户数据，在Reduce之前的数据如下图所示：

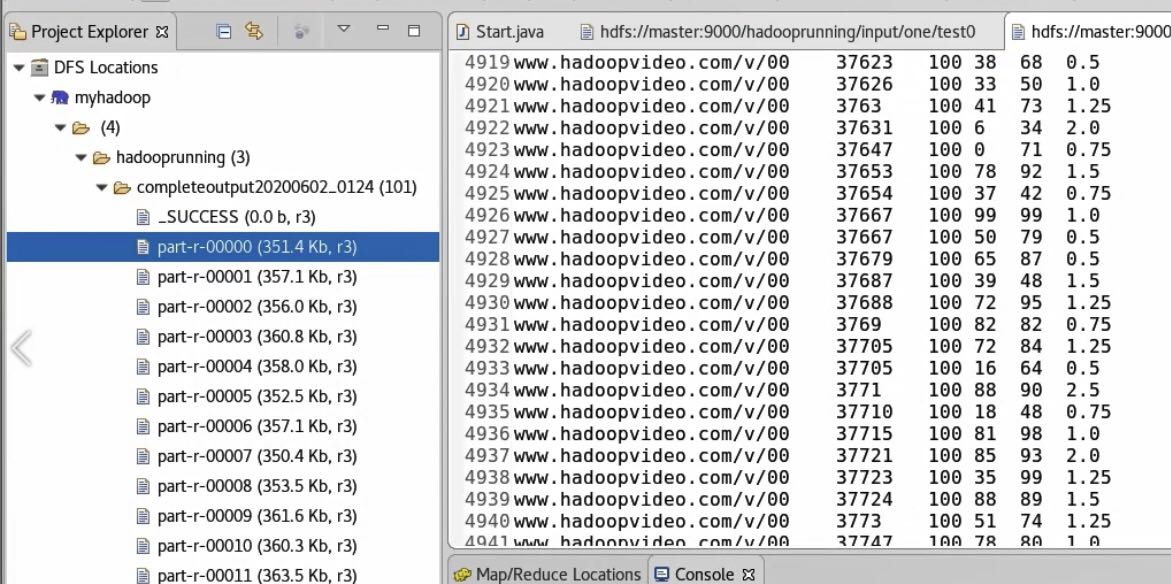


图4 Map之后、Reduce之前的数据图

经过Reduce累加处理后得到结果为下图所示：

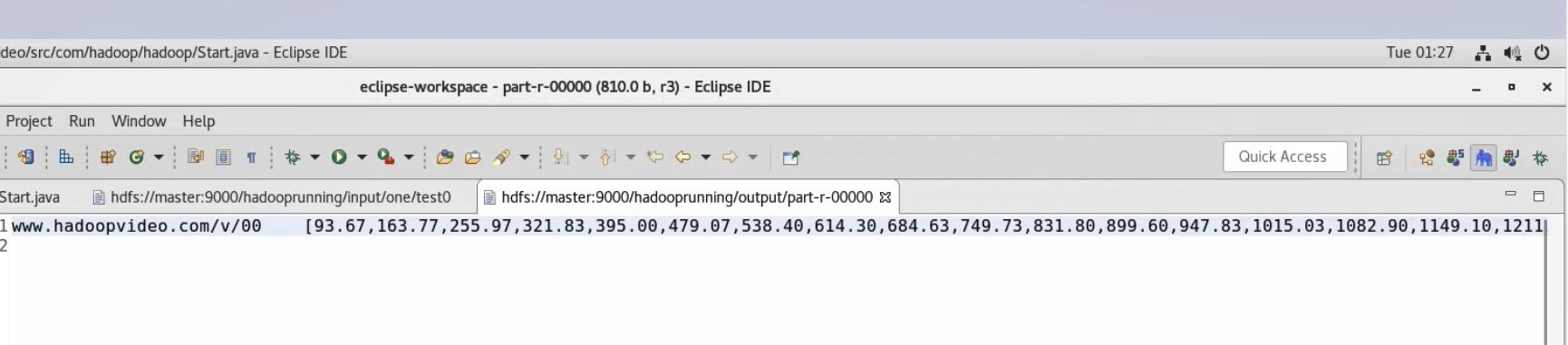


图5 Reduce累加数据后得到的数据图

具体算法设计如下：

Map阶段：每个节点运行一个Map，每个Map将前端提交的每一条用户行为数据，按照每个播放片段的起始时间、结束时间及播放倍速，对整个视频以秒为单位形成的数组进行加权计算，多个Map是并行的。例如这里以一个Map为例，一条用户行为数据为该用户从视频的第1秒到第30秒以2倍速进行播放，则该视频数组第1秒，第2秒，直至第30秒的数据都加0.5（1/2）。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，其余的Map是与这个Map工作相同采用并行的方式进行的。

Shuffle阶段：MapReduce算法的关键环节，Map的计算结果进行“洗牌”，将key值相同的数据分到一类，并交给同一个Reduce处理。

Reduce阶段：此次测试通过视频的数量来决定partion数量，100个视频对应100个partion,由partion设置了100个ReduceTasks,从而得到了100个Reduce。这些Reduce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，并且Reduce也是设置多Ruduce来完成工作加快执行速度。

系统的程序算法设计图如下所示:

**Map**

**阶段**

URL1/ID1/start1/end1/speed1/start2/end2/speed2/......../start10/end10/speed10

URL1/ID2/start1/end1/speed1/start2/end2/speed2/......../start13/end13/speed13

URL2/ID3/start1/end1/speed1/start2/end2/speed2/......../start15/end15/speed15

........(此时数据格式为多个字符串)

URL1|ID1|1|1|1|1|2|2|2|2|2|2|1.5|1.5|1.5|1.5|..........|3|3|3|3|3|3|1|1|1|1|1|1|1|1|1|2|2|2|2|2|2|

URL1|ID2|2|2|2|2|2|2|2|3|3|3|3|3|3|3|4|4|4|1|1|1|1|...............|5|5|5|5|5|5|5|5|5|3|3|3|3|3|3|3|

URL2|ID3|1|1|1|1|1|1|1.5|1.5|1.5|2|2|2|2|2|2|2|2|1|1|1|1|........|0.5|0.5|0.5|2|2|2|2|2|2|2|4|4|4|

........(此时数据格式已转化为多个数组)

**Shuffle**

**阶段**

**URL1**|ID1|1|1|1|1|2|2|2|2|2|2|1.5|1.5|1.5|1.5|..........|3|3|3|3|3|3|1|1|1|1|1|1|1|1|1|2|2|2|2|2|2|

**URL1**|ID2|2|2|2|2|2|2|2|3|3|3|3|3|3|3|4|4|4|1|1|1|1|...............|5|5|5|5|5|5|5|5|5|3|3|3|3|3|3|3|

**URL2**|ID3|1|1|1|1|1|1|1.5|1.5|1.5|2|2|2|2|2|2|2|2|1|1|1|1|........|0.5|0.5|0.5|2|2|2|2|2|2|2|4|4|4|

........(Hadoop机制将key相同的信息归为一类)

**Reduce**

**阶段**

**URL1**|ID1|1|1|1|1|2|2|2|2|2|2|1.5|1.5|1.5|1.5|..........|3|3|3|3|3|3|1|1|1|1|1|1|1|1|1|2|2|2|2|2|2|

**+**

**URL1**|ID2|2|2|2|2|2|2|2|3|3|3|3|3|3|3|4|4|4|1|1|1|1|...............|5|5|5|5|5|5|5|5|5|3|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|4|...............|8|8|8|8|8|8|6|6|6|6|5|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|2|2|1|1|1|1|........|0.5|0.5|0.5|2|2|2|2|2|2|2|4|4|4|

+........(将key值相同的value累加在一起)

图6 Hadoop程序算法设计图

(3)、处理结果压缩算法

由于前述MapReduce算法计算出的视频总体播放行为数据量比较大，在输出到HDFS文件系统存储时I/O开销比较大，各主机节点之间交互频繁，为了提高网络利用率，以及处理速度的提升，我们将计算结果进行压缩后再输出到HDFS中，这样既节约了存储空间，又节约了网络带宽。

为了保证传输过程是无损传输，后续数据处理是正确的，发送方压缩之后得出校验码，接收方接收数据之后进行校验，以保证数据在传输过程中是准确无误的。

我们采用文件字符里面重复字用“数字（即重复次数）+字符”代替原来重复字符的方式进行压缩。压缩前后数据结构如下。由于视频文件每秒片段被播放的次数相同的概率很大，这样压缩出来的文件很小，压缩率很高，可以大大减少读写HDFS的开销。

**URL1**|ID1|1|1|1|1|2|2|2|2|2|2|1.5|1.5|1.5|1.5|..........|3|3|3|3|3|3|1|1|1|1|1|1|1|1|1|2|2|2|2|2|2|

**URL2**|ID2|2|2|2|2|2|2|2|3|3|3|3|3|3|3|4|4|4|1|1|1|1|...............|5|5|5|5|5|5|5|5|5|3|3|3|3|3|3|3|

……

**URL1**|ID1|4 1|6 2|4 1.5|..........|6 3|9 1|6 2|

**URL2**|ID2|7 2|7 3|3 4|4 1|...............|9 5|7 3|

……

**压缩算法**

图7 压缩算法设计图

校验采用java.util.zip.CheckedInputStream里面的getChecksum()方法进行校验。

3、撰写了一篇技术说明书。

4、申请了一项软件著作权《基于Hadoop大数据的用户行为记录处理系统1.0》。

5、在《河南科技》期刊上发表了论文《Hadoop 大数据处理用户行为记录的创新应用》一篇。

6、结论

用Hadoop大数据将单个用户的视频观看行为痕迹进行计算、分析、存储，转化为各教学视频的总体播放情况数据，这种方法以数字化的手段直观地记录用户的视频观看信息，从大数据的角度，监测、存储和分析用户观看行为数据，使得教学信息的反馈方式更加实时、客观、可靠，对提高网络教学的效果具有重要意义，会成为促进网络教学水平提高的有力工具，未来我们会进一步丰富获取用户行为数据的内容和方式，采用更丰富、灵活的方式反馈教学情况，帮助师生在网络上高效获取知识。

大数据分析技术为快餐式观看视频提供技术手段，虽然大数据的作用很大，但是他仅仅是一种手段，不能完全替代认真观看的地位。

四、研究计划与研究情况对比

1、系统搭建计划与情况对比：在服务器上搭建完全分布式Hadoop平台时，起初原计划搭建1个master、8个slave节点，后根据实际应用需求与服务器配置属性，进行了Hadoop节点部署的优化调整，并最终优化调整为了1个master、2个slave节点。优化后的架构图

如下图所示：

图8 优化调整后的Hadoop架构图

Master与Slave通信以Slave2为例

：任务调度

：结果反馈

：心跳机制

：NameNode与SecondaryNameNode通讯

2、算法研究计划与情况对比：计划通过Java编写Job函数、多个Map函数与多个Reduce函数，分布式处理用户行为痕迹记录数据。例如记录用户的倍速播放的片段长度与次数以及前进、回退播放的片段起始、结束位置与次数等。利用Hadoop将这些数据转换为视频每秒的点击播放次数进行颗粒化存储，结合用户数据进行统计分析，可获得不同用户、不同视频的学习情况分析结果。

将前端接口提交的870M大小的数据格式的用户行为痕迹记录数据转换为视频每秒的播放次数并颗粒化存储到HDFS中，结合用户数据进行统计分析，获得不同用户、不同视频的学习情况总结。

本项目设定了从00到99的100个视频id值，每个视频id的汇总数据会分门别类在output里存在单独的文件里，如下图所示：

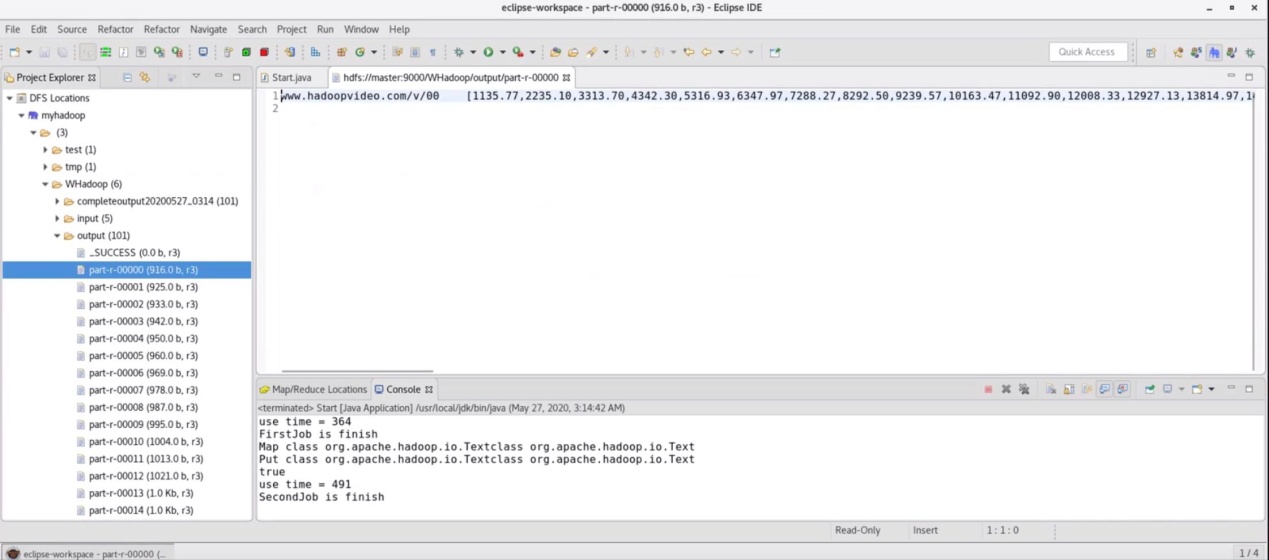


图 9 数据处理结果

图中左侧output文件夹内存放着100个视频的汇总数据，每个汇总数据文件的内容为图中右侧大框内的一行数据，这里以id为00的视频为例。

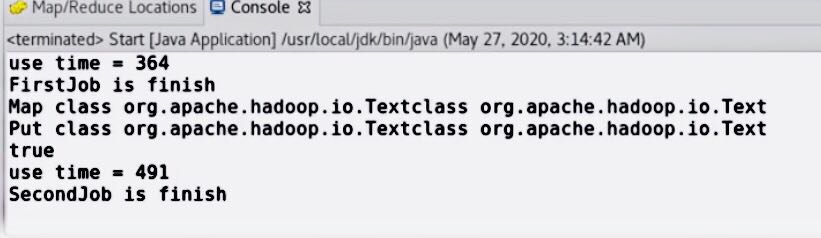


图 10 数据处理时间(单位：s)

数据处理时间如上图，图中有FirstJob和SecondJob两过程运行所耗时间。其中FirstJob程序所花时间为364s，此过程对原始数据按视频id进行分类排序并保存；SecondJob程序所花时间为491s，此过程将各个不同视频的用户行为数据进行累加计算。整个数据处理流程所花时间为两个时间相加：855s。

五、收获与体会

1、学习掌握了Java语言的知识与自主深入学习计算机语言的能力；

2、熟练掌握Linux常用命令与网络配置，用户以及权限管理操作，熟悉软件包及系统命令管理，对shell编程具有足够的学习能力；

3、对于项目所需的相关数据库基础知识，虚拟机的搭建，安装虚拟机以及配置虚拟机网络已有了较深入的了解；

4、拥有了学习掌握大数据中Hadoop搭建、Hadoop的核心内容MapReduce、HDFS、Yarn以及旁系知识的能力；

5、在不断攻克项目一个个难题的同时，也提高了我们制定计划、自主学习、查阅文献、分工合作、表达沟通以及创新实践的能力，并且培养了坚持与不惧压力的精神。