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：Hadoop 大数据处理行为痕迹记录的创新应用

立 项 时 间：2019 年 6 月

项目负责人：杨秉学

学院年级专业：控制与计算机工程学院计算机科学与技术

联 系 电 话：13756697779、18072264663

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 12 日

一、项目研究进展情况（包含小组成员的分工合作，项目研究方法、内容、进度等）

1、成员分工：

杨秉学：负责完全分布式 Hadoop 环境的搭建以及配置、管控，设计技术路线和算法，编写并优化并行式处理用户行为痕迹记录程序；

刘俊龙：辅助搭建 Hadoop 环境，程序运行过程监控，算法设计与优化。

2、项目研究方法：

基于 Linux 系统虚拟化搭建完全分布式 Hadoop 环境，为运用大数据处理方式处理用户观看视频行为痕迹记录提供平台；根据实际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 MapReduce 函数分布式处理用户行为痕迹记录数据；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 MapReduce 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、项目研究内容：

1）在服务器上部署完全分布式 Hadoop，为并行式处理分析海量的观看视频时用户行为痕迹记录数据提供平台，根据实际应用需求以及服务器的属性配置，优化调整 Hadoop 中计算节点、数据节点、HDFS 的部署和配置。

2）将前端播放器用户传来的用户行为痕迹记录数据置于 Hadoop 平台上，测试运行用户观看视频行为痕迹记录程序，监控 Hadoop 平台运行过程并记录分析 Hadoop 的计算任务和文件 IO 开销，以优化程序算法。

3）根据 Hadoop 并行计算框架优缺点和应用程序的需求，分析现有技术方案和程序实际应用环境，不断测试以调整 Mapper 和 Reducer 的数量，来改善 DataNode 的运算能力，提高应用程序的效率。

4）模拟大量用户并发提交环境，并使程序能在 Hadoop 平台上快速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，最终将所得汇总数据结果以不同深浅的颜色曲线反馈到视频播放页面上。

4、项目研究进度：

目前已成功在服务器上搭建完全分布式 Hadoop 平台，并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，由起初 1 个 master、8 个 slave 节点优化调整为 1 个 master、6 个 slave 节点。完成了并行式处理用户行为痕迹记录数据程序，并将前端传来的用户观看视频时操作行为数据置于 Hadoop 上进行程序测试运行并得出了所有用户观看视频时操作行为的汇总数据。

下一步，项目组将重点优化并行式处理用户观看视频时操作行为算法，通过不断测试来根据待处理数据数量动态调整 Mapper 和 Reducer 的数量和运算能力，优化算法，以提升总体效率。

二、已取得的阶段性成果（已取得的阶段性成果或收获等）

1、在服务器上成功部署了完全分布式 Hadoop 平台

于 master 节点上成功启动了完全分布式 Hadoop，同时通过 jps 命令监测各节点启动进程并分别运

用命令行方式与 Web 方法检测 DataNode 运行状况，以此验证完全分布式 Hadoop 集群的成功启动。

1）Master 节点中 NameNode、ResourceManager、SecenderyNameNode、JobHistoryServer 进程

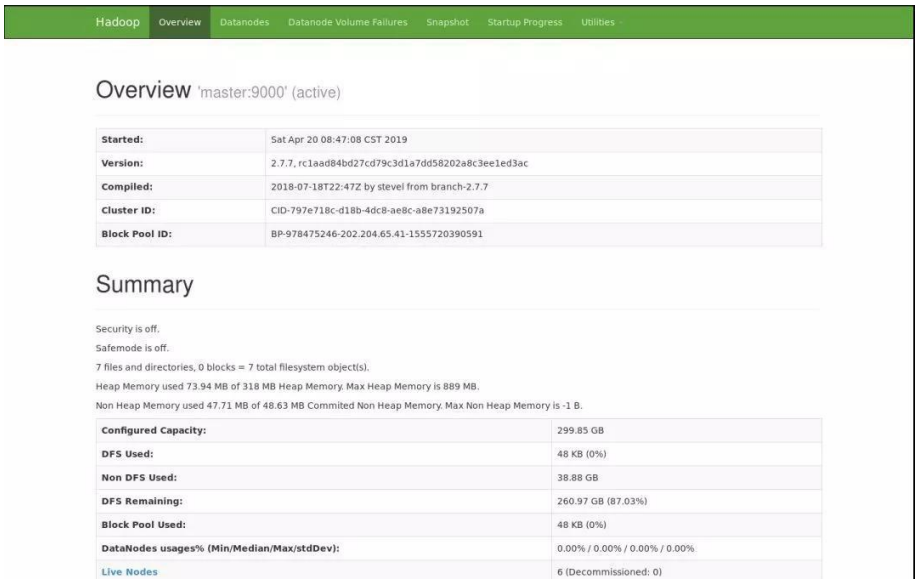
```
[root@slave8 ~]# jps
13618 Jps
129352 NodeManager
127530 DataNode
[root@slave8 ~]#
```

2）slave 节点中 DataNode 与 NodeManager 进程

```
[root@master ~]# jps
16080 Jps
12099 SecondaryNameNode
14724 ResourceManager
15707 JobHistoryServer
11404 NameNode
```

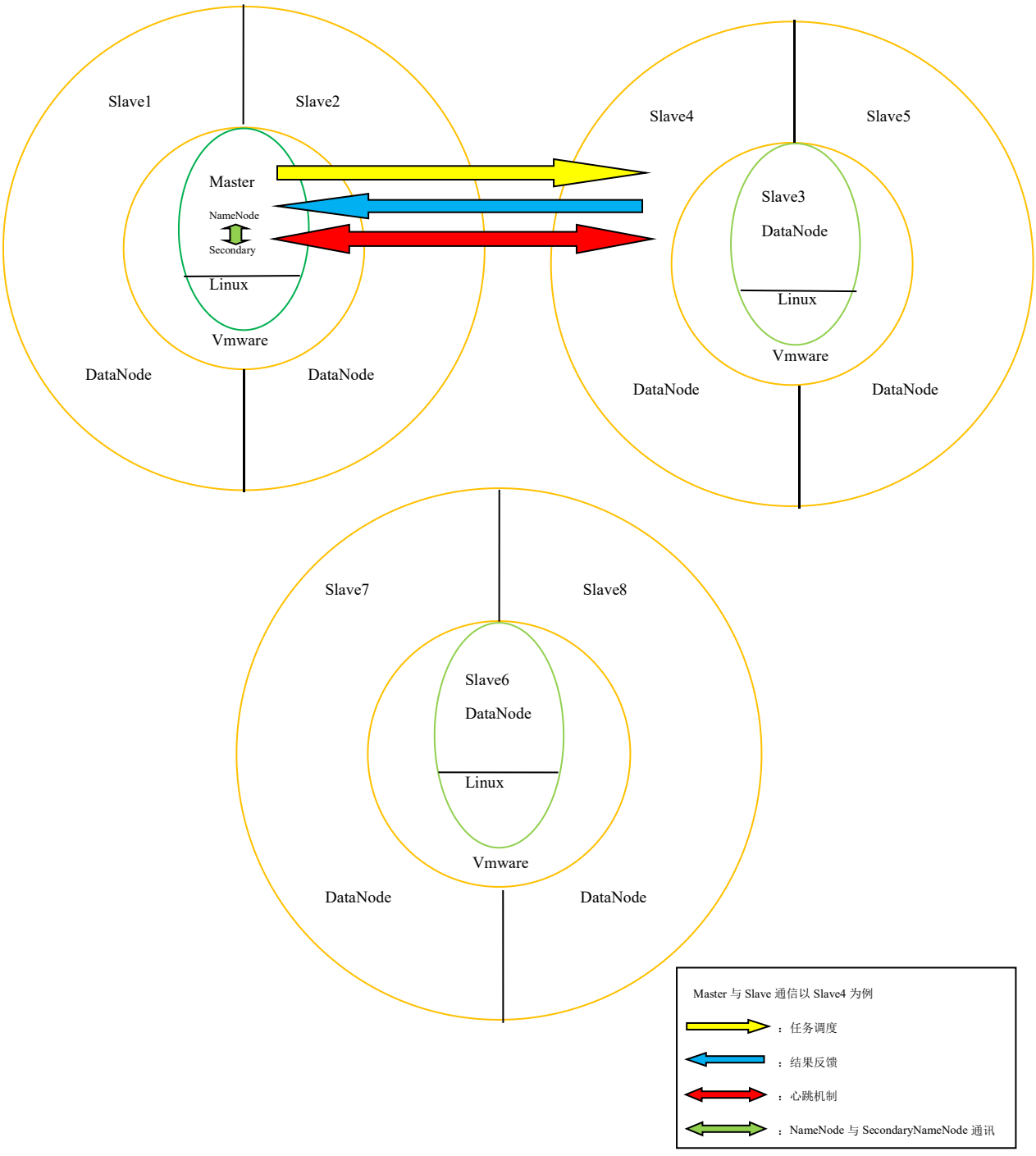
3）命令行方式与 Web 方式检测 DataNode 启动状况

```
11404 NameNode
[root@master ~]# hdfs dfsadmin -report
Configured Capacity: 321965260800 (299.85 GB)
Present Capacity: 280215257088 (260.97 GB)
DFS Remaining: 280215207936 (260.97 GB)
DFS Used: 49152 (48 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-------------------------------------------------
Live datanodes (6):

Name: 202.204.65.23:50010 (slave4)
Hostname: slave4
Decommission Status : Normal
Configured Capacity: 53660876800 (49.98 GB)
DFS Used: 8192 (8 KB)
Non DFS Used: 6636486656 (6.18 GB)
DFS Remaining: 47024381952 (43.79 GB)
DFS Used%: 0.00%
DFS Remaining%: 87.63%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sat Apr 20 08:57:57 CST 2019


Name: 202.204.65.119:50010 (slave8)
Hostname: slave8
Decommission Status : Normal
Configured Capacity: 53660876800 (49.98 GB)
DFS Used: 8192 (8 KB)
Non DFS Used: 6453903360 (6.01 GB)
DFS Remaining: 47206965248 (43.96 GB)
DFS Used%: 0.00%
DFS Remaining%: 87.97%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
```



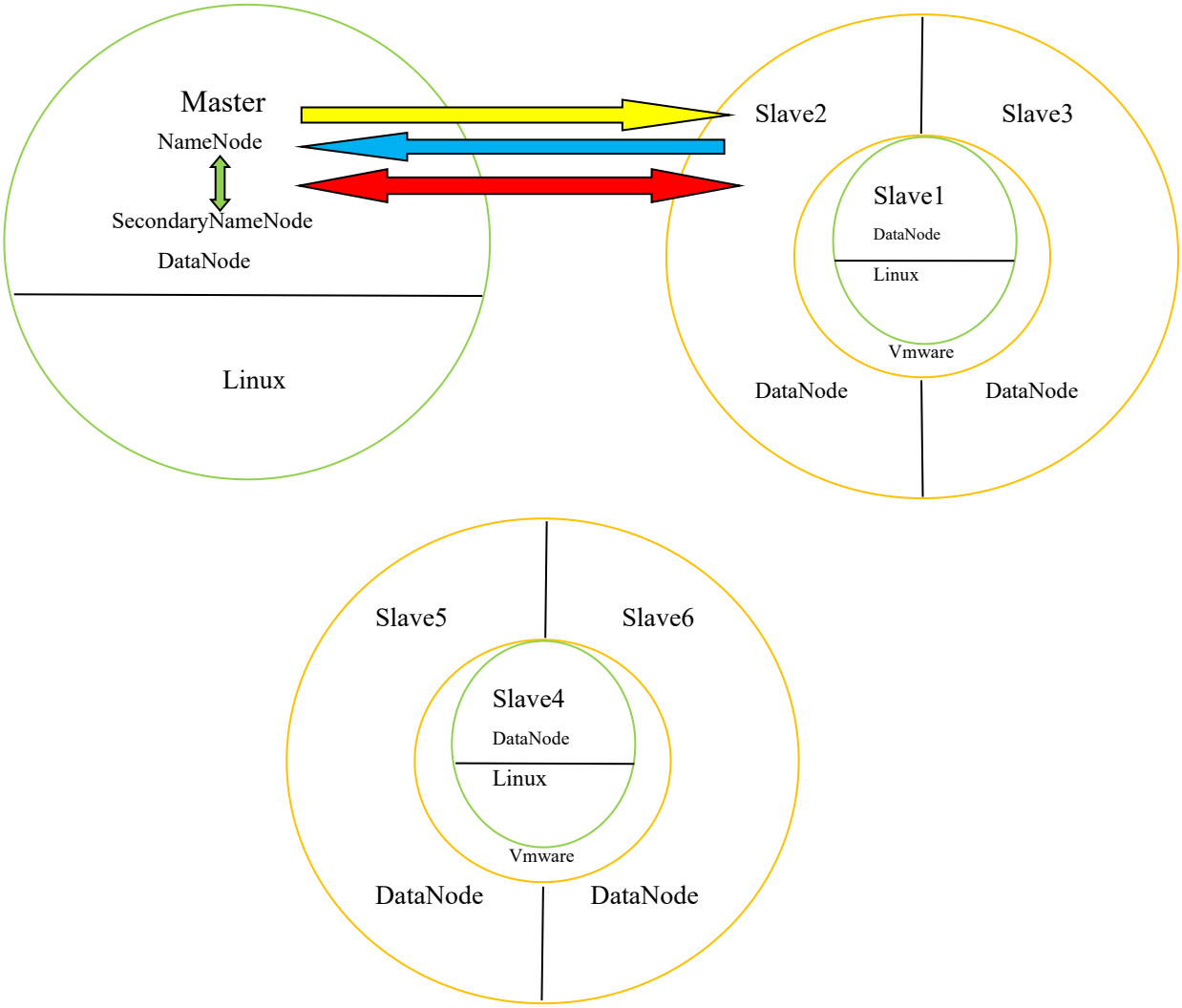
2、根据实际应用需求与服务器属性配置优化调整了 Hadoop 的节点部署

项目初期 Hadoop 架构：



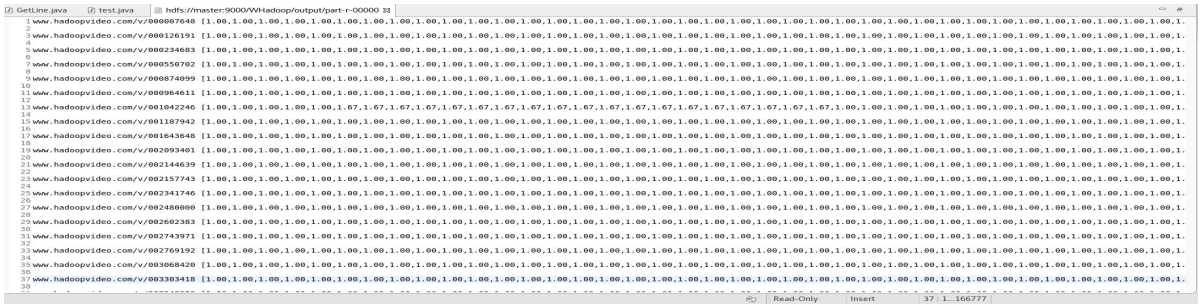
由于虚拟化搭建完全分布式 Hadoop 集群对服务器内存资源开销过大，无法满足项目实际应用需求，因此将 Hadoop 节点部署优化调整为 1 个 master、6 个 slave 节点

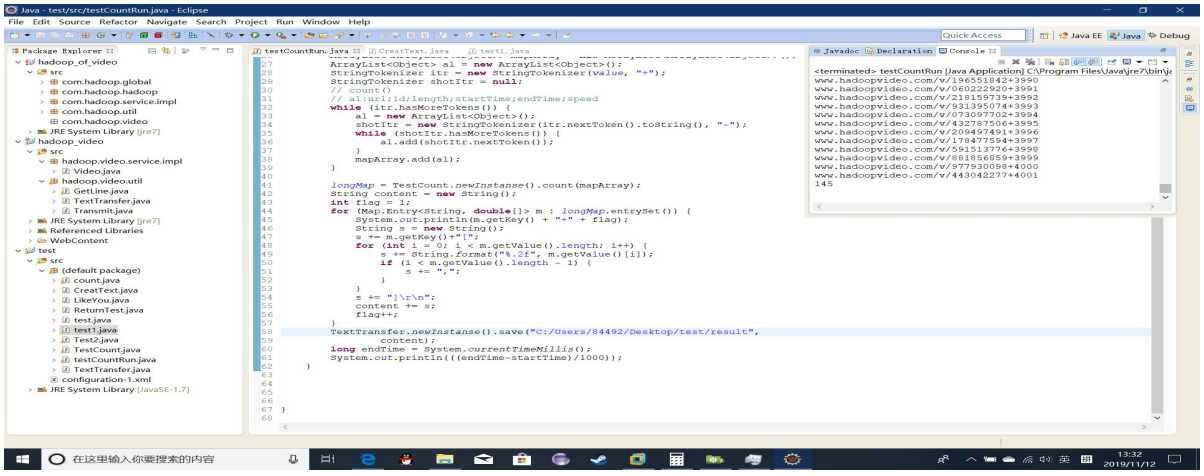
优化调整后的 Hadoop 架构：

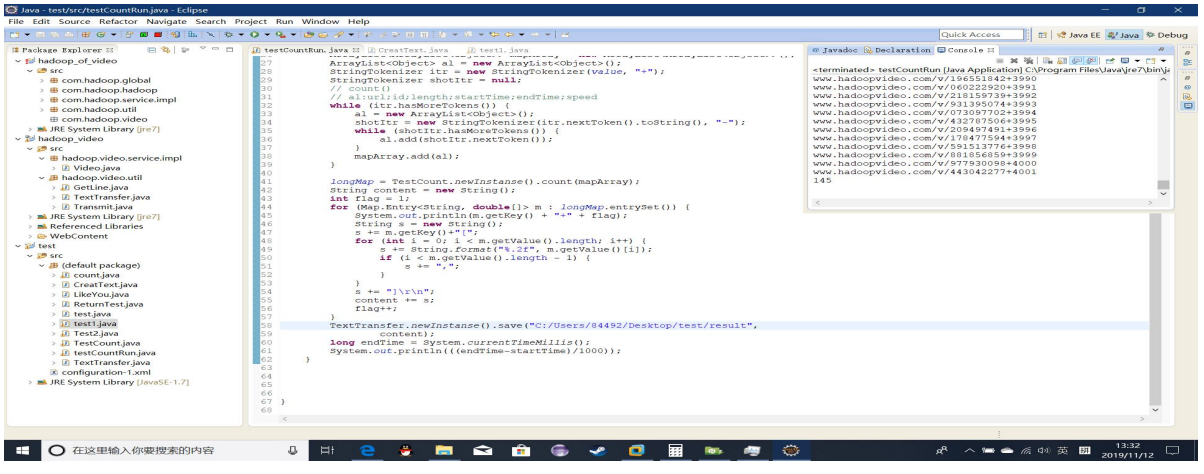


3、将前端播放器用户传来的用户行为痕迹记录数据置于 Hadoop 平台上，测试运行用户观看视频行为痕迹记录程序，监控 Hadoop 平台运行过程并记录分析 Hadoop 的计算和 IO 开销状况与文件输出的时间。

前端所传用户行为痕迹记录数据、并行式处理分析用户行为痕迹记录获取视频播放总体情况数据结果以及关键代码截图如下图：







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，从而导致前端提交到服务器的数据的溢出和丢失。

措施：在服务器端处理程序中，编程实现监控数据流量的功能，并以此动态调整线程数量。

问题 2：每一个线程中的 for 循环增加了时间与空间的复杂度，无法完全满足项目算法要求。

措施：优化 Hadoop 中任务调度算法，不断测试以获得满足前端提交速度不同情况下的动态调整 Mapper 和 Reducer 数量和处理能力的合适算法，来改善 DataNode 的运算能力，使处理用户行为痕迹记录程序在 Hadoop 平台上高效快速地并行式归纳分析用户观看视频时的操作行为痕迹，最终将所得汇总数据结果以不同深浅的颜色曲线反馈到视频播放页面上。

四、下阶段主要计划及时间安排

1、2019 年 10 月至 2020 年 1 月优化处理用户行为痕迹记录算法，不断测试以获得动态调整 Mapper 和 Reducer 的数量和处理能力的优化算法，来改善 DataNode 的处理能力，提高应用程序的效率；同时模拟大量用户并发提交环境，并使处理用户观看视频时的操作行为数据程序在此环境下快速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，最终将所得结果以不同深浅的颜色曲线反馈到视频播放页面上；

2、2020 年 2 月至 5 月将进行算法优化、完成技术说明书、撰写论文、申请软件著作权。

3、2020 年 5 月系统完成后放到学校网络教学平台上进行测试，推广使用。

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）
1	印刷品 Spark 编程基础（Scala 版）	39.68
合计		

项目负责人签字：　　2019 年 11 月 15 日
指导教师鉴定意见（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） 项目研究进展正常，目前已完成环境构建，确定技术路线和系统架构及部分程序的编写，进度达到预期，下一步将实现模拟环境测试、算法优化、前端提交数据到服务器的接口程序编写，撰写论文，申请软件著作权。 签章：　　2019 年 11 月 15 日
学院意见（请给出评审意见及评定成绩） 签章：　　年　　月　　日
学校专家组意见 签章：　　年　　月　　日

中期检查成绩评定	优　　良　　通过　　未通过