

# Hadoop大数据处理行为痕迹记录的创新应用

负责人：杨秉学 13756697779

组员：刘俊龙 18072264663





## 项目成果

1

在服务器上成功部署了完全分布式Hadoop平台

2

根据实际需求优化部署

3

记录分析Hadoop的计算和I/O开销状况与文件输出时间

4

升级Hadoop计算框架算法



1

在服务器上成功部署了完全分布式Hadoop平台





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

1

## Master节点中NameNode进程检测

2

## Slaver中DataNode进程检测

3

## 命令行检测DataNode启动状况

4

## Web方式检测DataNode运行状况

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

[illegible]

[Reading](#)
[Overview](#)
[Statistics](#)
[Configure Website Features](#)
[Uninstall](#)
[Backup Progress](#)
[Others](#)

## Overview master:9000 (active)

**Status:**

Not Up

2018-07-20 01:48:43 UTC 2018

**Version:**

3.1.7

3.1.7 (c) 2016-2018 by the Net::Minion developers

**Compiled:**

2018-07-18T22:47:02

by tencent from branch:3.1.7

**Client ID:**

CD-7674718-4188-4610-ae0b-ae673220507a

**Block Peer ID:**

BP-07847248-202-204-61-135572309391

## Summary

Security is off

Subdomain is off

7 files and directories, 0 blocks = 7 total filesystem objects

Heap Memory used 73.94 MB of 218 MB total Heap Memory Max Heap Memory is 889 MB

Non Heap Memory used 47.71 MB of 48.63 MB Committed Non-Heap Memory Max Non-Heap Memory is 1.0.

Current Capacity:

248.00 GB

45.00 (%)

Non DPS Used:

36.00 GB

200.97 GB (87.32%)

DPS Remaining:

48.00 GB (%)

0.00% (0.00%) / 0.00% (0.00%)

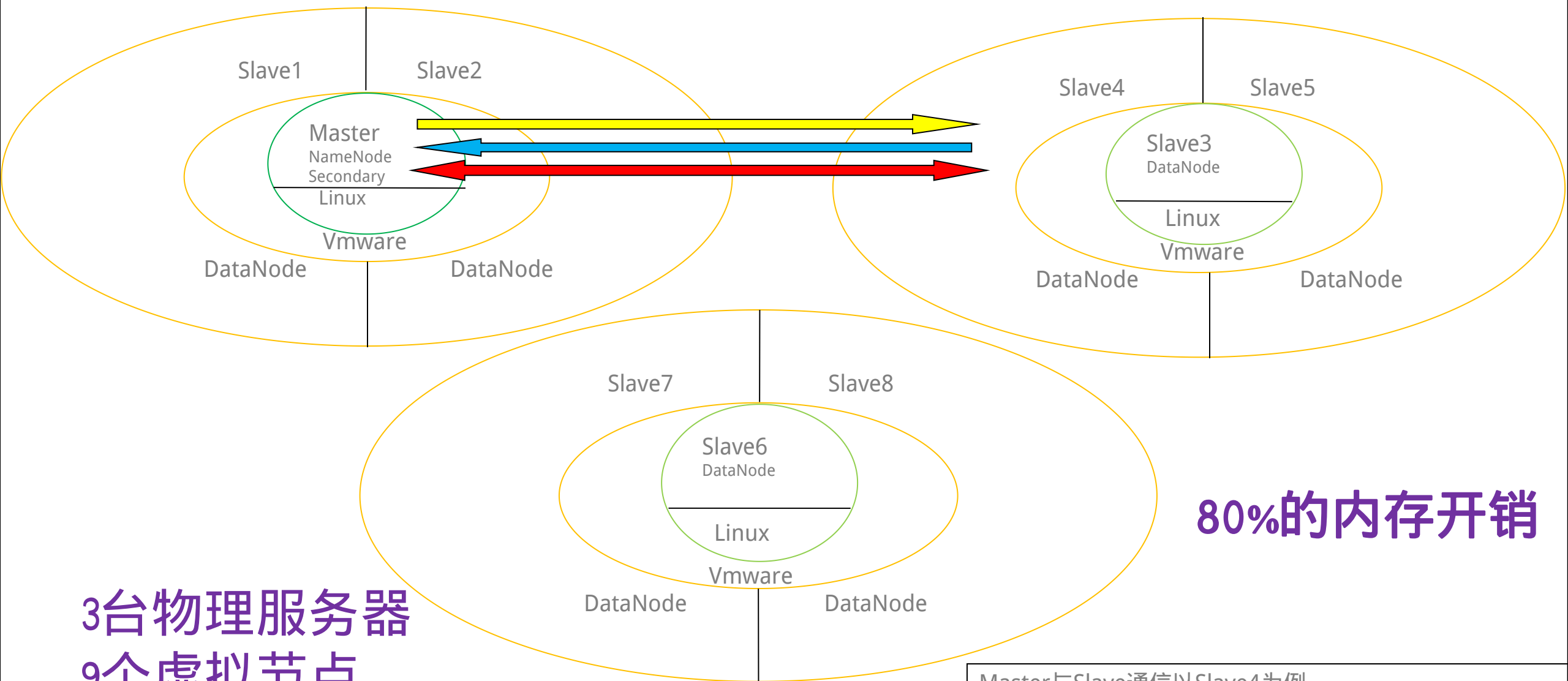
Blockfiles usage(s):

0/0(s):Medium:Max(s):0/0(s)



根据实际需求优化部署

项目初期Hadoop架构图



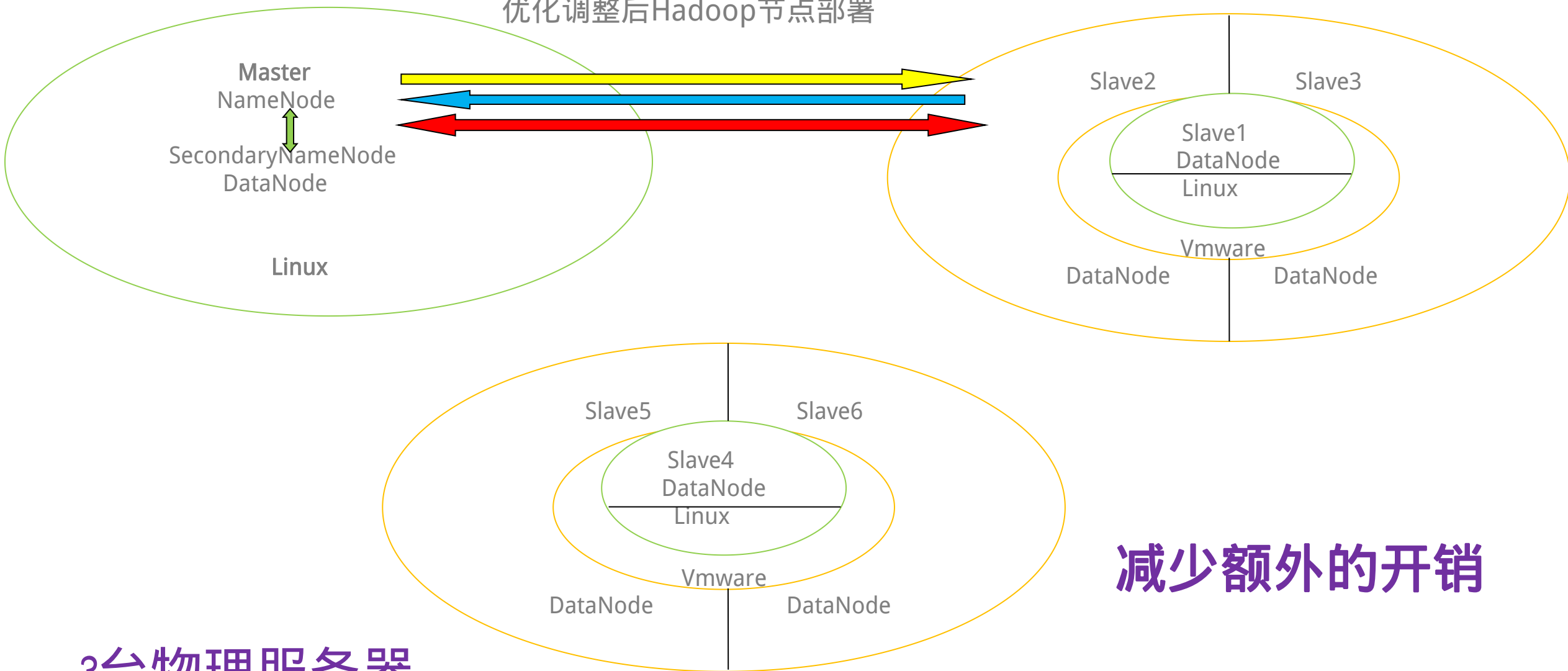
3台物理服务器  
9个虚拟节点

80%的内存开销

Master与Slave通信以Slave4为例

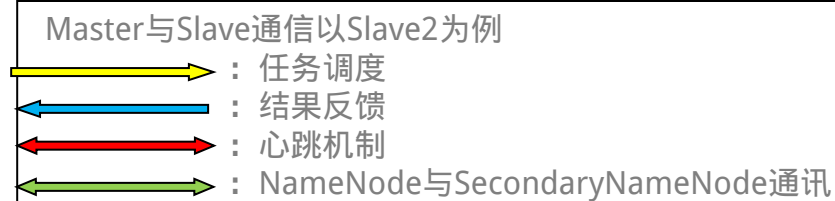
- : 任务调度
- : 结果反馈
- : 心跳机制
- : NameNodeSecondaryNameNode通讯

## 优化调整后Hadoop节点部署



减少额外的开销

3台物理服务器  
7个虚拟节点





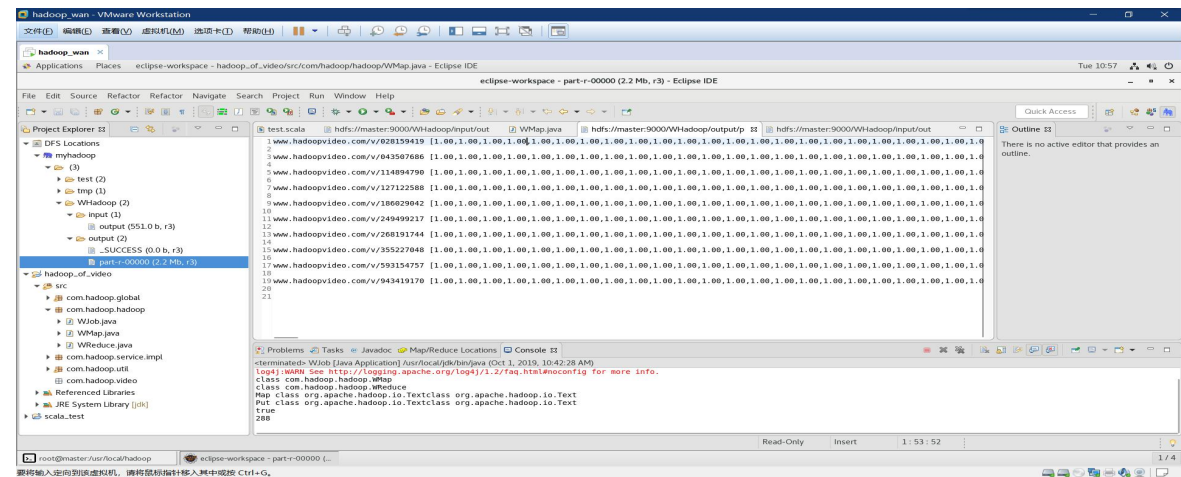
# 3


记录分析Hadoop的计算和I/O开销状况与文件输出时间。





物理服务器单节点处理 **24M**  
用户行为数据用时 **255s**





# 4

优化Hadoop计算框架算法



# 项目目前存在需要解决的主要问题



**问题1：** 程序运行时间与空间复杂度比较高

**问题2：** 现在采用的算法是简单的for循环，没有经过优化

**问题3：** 目前程序读写HDFS文件系统开销较大

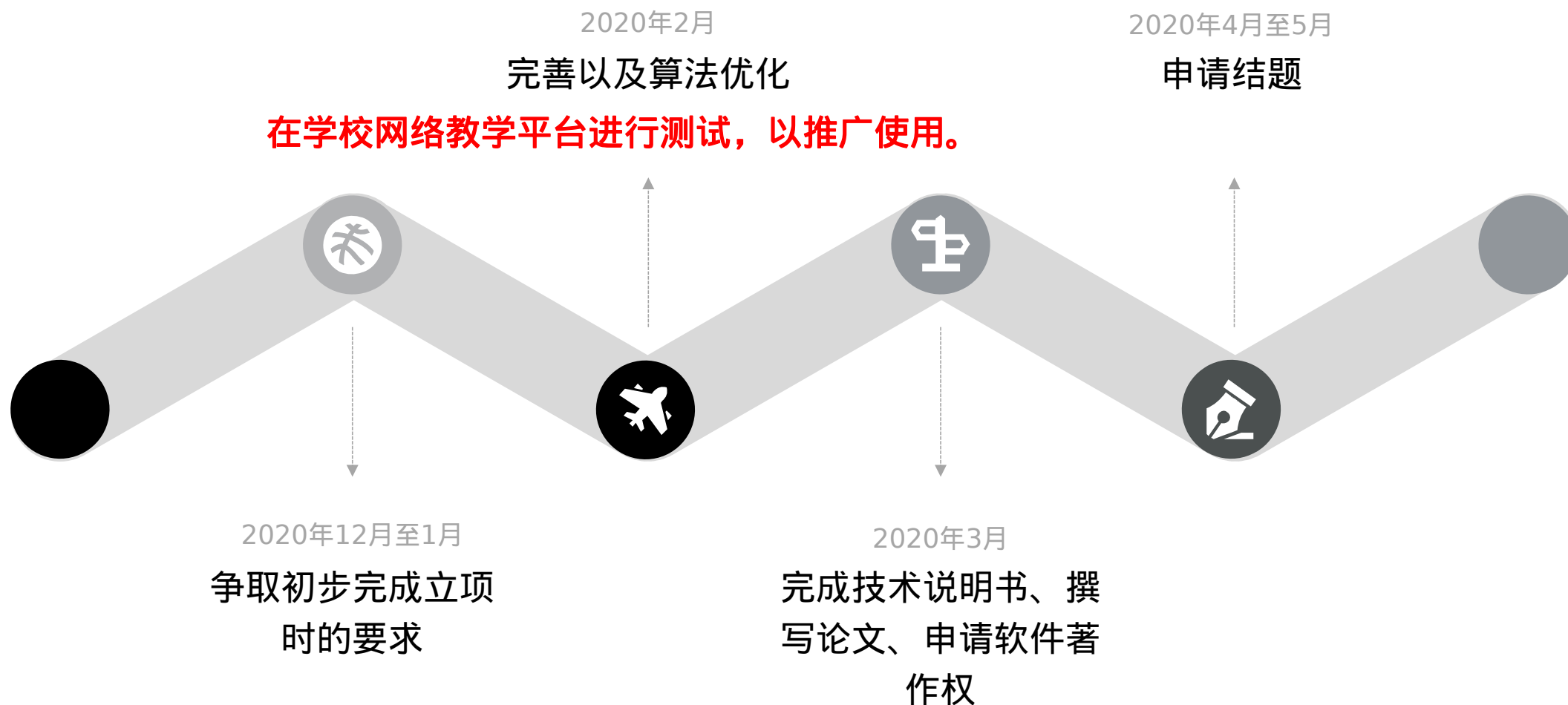
## 项目主要问题解决方案

**措施1：** 算法优化，测试以获得在前端提交速度不同情况下,动态调整Mapper和Reducer数量来增强处理能力的合适算法，以增强程序的适应性,满足我们在实际应用场景的需要。

**措施2：** 对用户行为的分片大小进行优化

**措施3：** 研究视频播放情况结果文件的压缩方法，减少程序读写文件尺寸，从而减少I/O开销

# 下阶段主要计划及时间安排





**谢谢聆听！**