**Hadoop大数据处理用户行为痕迹记录的创新应用**

**1.0**

**技术说明书**

华北电力大学-版权所有

**目录**

[一、总体介绍 1](#_Toc5815213)

[二、环境总体介绍 2](#_Toc5815214)

[2.1 开发环境 2](#_Toc5815215)

[2.1.1 开发平台 2](#_Toc5815216)

[2.1.2 开发语言 2](#_Toc5815217)

[2.2 硬件环境 2](#_Toc5815218)

[三、程序总体设计 2](#_Toc5815221)

[3.1 程序算法设计 2](#_Toc5815222)

[3.2 Hadoop平台框架 3](#_Toc5815223)

[3.3 数据记录设计](#_Toc5815226) 5

[四、使用结果 5](#_Toc5815224)

[4.1模拟前端用户数据提交 5](#_Toc5815225)

[4.2 数据分布式处理结果展示](#_Toc5815226) 6

[4.3 Hadoop监控平台](#_Toc5815226) 8

# 

# 一、总体介绍

随着互联网的发展，互联网越来越贴近人们的生活。根据相关数据表明，人们每月花在上网的时间高达350亿小时。因此，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。

在互联网迅猛发展的背景之下，授课打破了传统教学在时空上的限制。以慕课、网络课件等为代表，他们把教学资源及其相关的内容放在服务器上，方便学习者根据自己时间情况随时进行学习。就目前网络教学的状况而言，在师生交流方面仍有极大的发展空间，尽管有弹幕，提问区域等方式促进教学互动，但这种交互方式学生的参与度并不高，也无完全地将问题表达，并且回复也难以做到全面与系统。教师对于讲义中的重点、难点以及学生的困惑难以准确地把握，使得传统教学中教师那种言传身教在网络教学平台中无法很好的实现，这是网络教学相对于传统教学的严重缺陷。

因此，为了顺应时代的潮流、响应党和国家的号召，有效处理在线学习用户产生的大量行为痕迹数据，以反映学习者在视频学习中的重点、难点，反馈给视频发布者和教师，作为他们改进教学视频的参考，对弥补上述网络教学的缺陷具有重要作用。

由于在线学习的资源和用户量极大，抓取用户行为痕迹将产生海量行为数据，因此这些数据的处理、处理结果的存储都对应用系统具有很高的要求。Hadoop是当前主流的大数据技术，对海量数据的处理、存储具有优势，因此文中选用Hadoop平台作为用户行为痕迹处理和记录的平台，基于Hadoop技术，将从网络教学网站抓取的用户观看视频的行为痕迹处理、记录，并进行分析，例如处理用户的倍速播放的片段与次数以及前进、回退播放的片段起始、结束位置与次数等，将这些数据转换为视频每帧的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总结。这样能客观并且直观地发现教学中的难度较大与有待改进的环节，将学习视频中存在的问题全面且客观地反馈给网络教学网站和教师，便于教师改进优化教学方案，提升教学质量。其次，它能提醒学生在重点、难点处集中注意力认真研习，更了解自己知识的薄弱环节，便于洞察知识中的漏洞，清楚学习情况，做出科学合理的学习计划。

此系统将全新的Hadoop大数据技术应用于网络教学，不仅能推进网络教学中的教与学的改进与优化，对于网络教学技术的变革也有着重要的影响。

# 二、环境总体介绍

## 2.1 开发环境

### 2.1.1 开发平台

基于Linux系统系列中的CentOS搭建完全分布式Hadoop环境，其内核Shell为bash 4.2.46，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开发平台。

### 2.1.2 开发语言

运用Java语言进行研究与开发，编写Job函数与MapReduce函数分布式处理用户的倍速播放的片段与次数以及前进、回退播放的片段起始、结束位置与次数等行为痕迹数据。

## 2.2 硬件环境

(1) CPU：Intel Xeon E5-2620 v2 @ 24x 2.6GHz

(2) 硬盘:56G以上

(3) 内存：15833MiB

# 三、程序总体设计

## 3.1 程序算法设计

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。

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 1 程序算法设计图

## 3.2 Hadoop平台框架

系统节点的分布架构图如下所示。

Slave5 Slave6

Vmware

DataNode DataNode

Slave2 Slave3

Vmware

DataNode DataNode

Master

NameNode

SecondaryNameNode

DataNode

Linux

Slave1

DataNode

Linux

Slave4

DataNode

Linux

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

图 2系统节点分布架构图

## 3.3 数据记录设计

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型  （精度范围） | 主键 | 外键 | 允许为空Y/N | 唯一Y/N | 默认值 |
| URL | string | 是 | 否 | N | Y |  |
| ID | string | 否 | 否 | N | Y |  |
| start | float | 否 | 否 | N | N | 0.0 |
| end | float | 否 | 否 | N | N | 0.0 |
| speed | float | 否 | 否 | N | N | 1.0 |

注：URL与ID无默认值，其值视具体视频信息与用户信息而定。

程序以每秒为一个时间段划分视频，每秒均有一组（start，end，speed）标记记录，分别代表每个时间段的开始时间，结束时间与倍速播放的速值，用户在观看视频时将鼠标抬起将开启一个新时间段并赋值给start；按下将结束此时间段并赋值给end；speed的初值为1.0，观看视频时将每各时间段current\_speed的值赋值给speed，以此方式记录用户观看视频时的行为痕迹数据。

**四、系统使用结果**

## 4.1模拟前端用户数据提交

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。在客户端成功使前端抓取JSON格式的数据再写入到HDFS中。

下图为24M的文件的并发提交数据环境，



图 3 前端用户提交数据模拟

## 

## 4.2 数据分布式处理结果展示

将这些数据转换为视频每帧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。

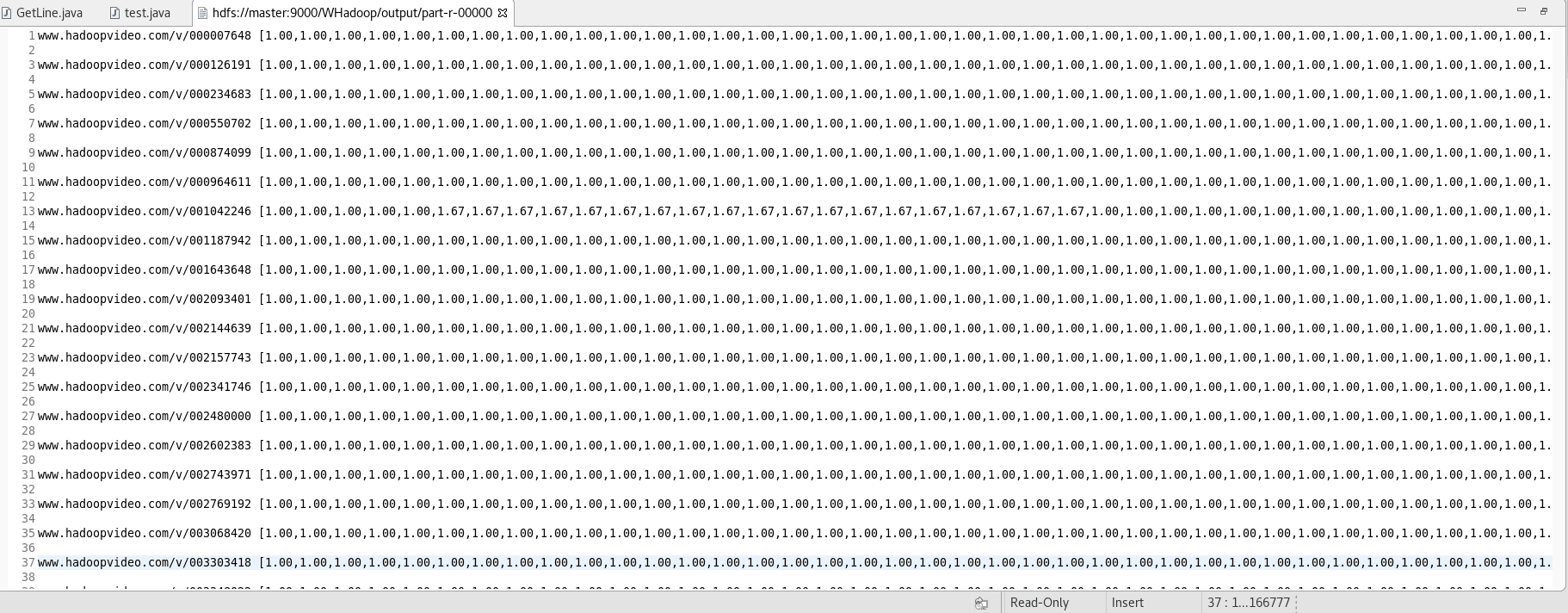


图 4 数据分布式处理结果

在Hadoop框架中应用分布式计算处理速度截图为：

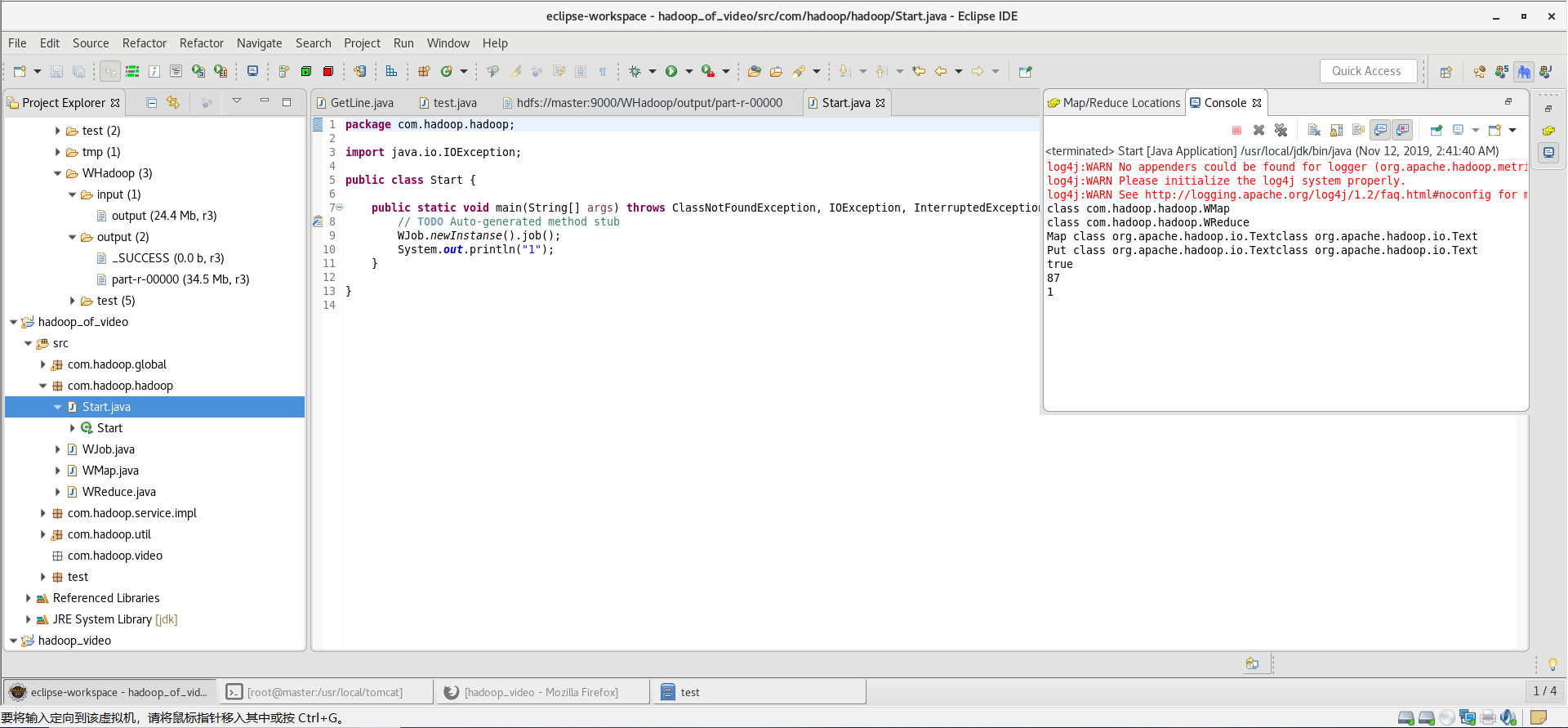


图 5 数据分布式处理结果

文中提出了选用当前主流的大数据技术Hadoop平台作为用户行为痕迹处理和记录的平台对海量数据进行处理、存储，其具有处理数据的高效性与运行的稳定性。基于Hadoop技术，将从网络教学网站抓取的用户观看视频的行为痕迹处理、记录，并进行分析，这样能客观并且直观地发现教学中的难度较大与有待改进的环节，将学习视频中存在的问题全面且客观地反馈给网络教学网站和教师，便于教师改进优化教学方案，提升教学质量。其次，它能提醒学生在重点、难点处集中注意力认真研习，更了解自己知识的薄弱环节，便于洞察知识中的漏洞，清楚学习情况，做出科学合理的学习计划。

## 4.3 Hadoop监控平台

于master节点上启动完全分布式Hadoop，同时可通过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，并且可以检测集群所提供的服务是否符合期望。

1. Master节点中NameNode、ResourceManager、SeconderyNameNode、JobHistoryServer进程：



图 6 NameNode、ResourceManager、SeconderyNameNode、JobHistoryServer进程监控图

1. slave节点中DataNode与NodeManager进程：



图 7 DataNode、NodeManager进程监控图

3）命令行方式与Web方式检测DataNode启动状况：

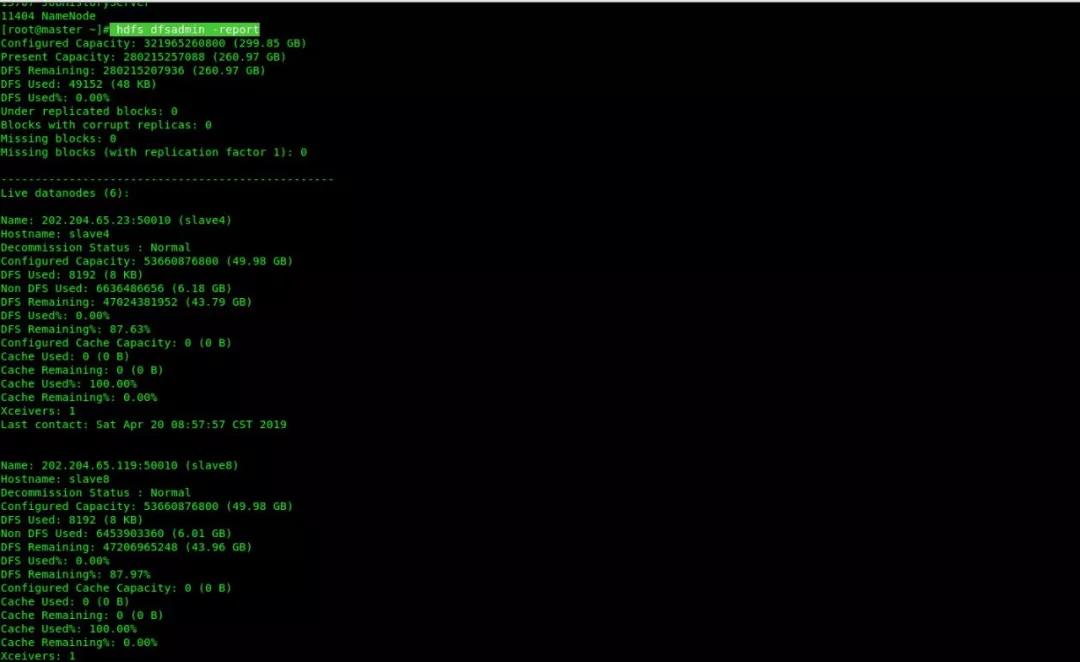


图 8 命令行方式的DataNode启动监控图

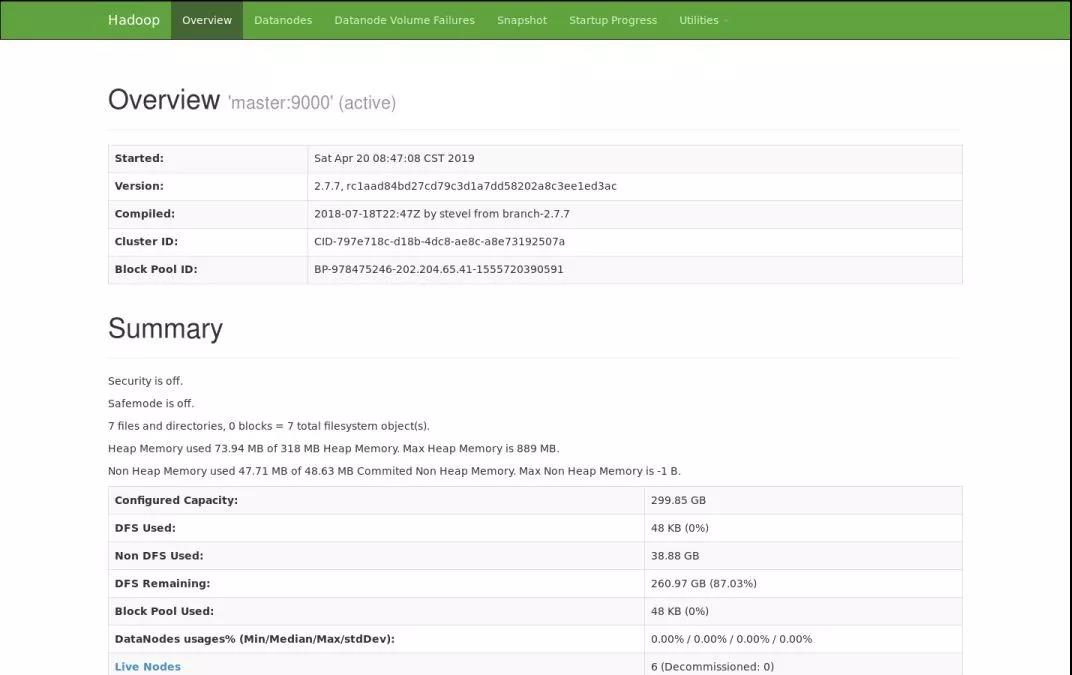


图 9 Web方式的DataNode启动监控图