中期报告

一、项目阶段性成果：

项目组通过查阅书籍、网站博客、老师们的指导以及网络视频等资源获取关于本项目的相关信息并且不断试验，项目组取得了以下进展：

1. 在服务器上完成了Linux平台与Hadoop环境的搭建，并且安装了RPM与Hadoop生态系统中的yarn

本项目的Hadoop架构图

搭建过程如下：

1.准备工作

1)建立虚拟机并安装Linux系统

2)选定一台机器作为 Master，其余作为slave（原准备配置9个节点，后由于服务器内存不足便改为了3个节点）

2.网络配置(修改IP)

1)必要条件

(1)所有集群的节点都位于同一个局域网

(2)如果使用的是虚拟机安装的系统，那么需要更改网络连接方式为桥接（Bridge）模式，才能实现多个节点互连。

(3)虚拟机里面的CentOS是不会自动联网，手动配IP联网

(4)[CentOS7修改静态IP](http://mp.weixin.qq.com/s?__biz=MzU1OTM3ODQ0MA==&mid=2247483798&idx=7&sn=531d32452968855483687e2d139ecdbb&chksm=fc1978f1cb6ef1e730d1554efb101ee02e971a79168f0ba99f08525829d63f321fd487f2e255&scene=21#wechat_redirect)

3在 Master 节点上配置

1)关闭防火墙与SELinux

存在防火墙会导致ping的通，但是端口不通，从而导致DataNode启动，但是无法使用的问题。

2)安装 Java 环境

3)为了便于区分，可以修改各个节点的主机名

4)为了与其他机器方便通信，修改hosts

5)配置SSH免密

4.在其他 Slave 节点上配置 hadoop 用户、安装 SSH server、安装 Java 环境

5.在 Master 节点上安装 Hadoop，并完成配置

6.进行后续集群配置

1)关闭Hadoop(可以不执行)

2)集群/分布式模式需要修改 /usr/local/hadoop/etc/hadoop 中的5个配置文件

(1)**slaves**

(2)**core-site.xml**

(3).**hdfs-site.xml**

(4)**mapred-site.xml**

(5)**yarn-site.xml**

(6)/usr/local/hadoop/etc/hadoop/hadoop-env.sh

8)上述master配置成功以后，将 master 节点上的 /usr/local/hadoop 目录复制到其他 Slave 节点上9)slave1、slave2、slave3执行

10)首次启动需要先在 Master 节点执行 NameNode 的格式化

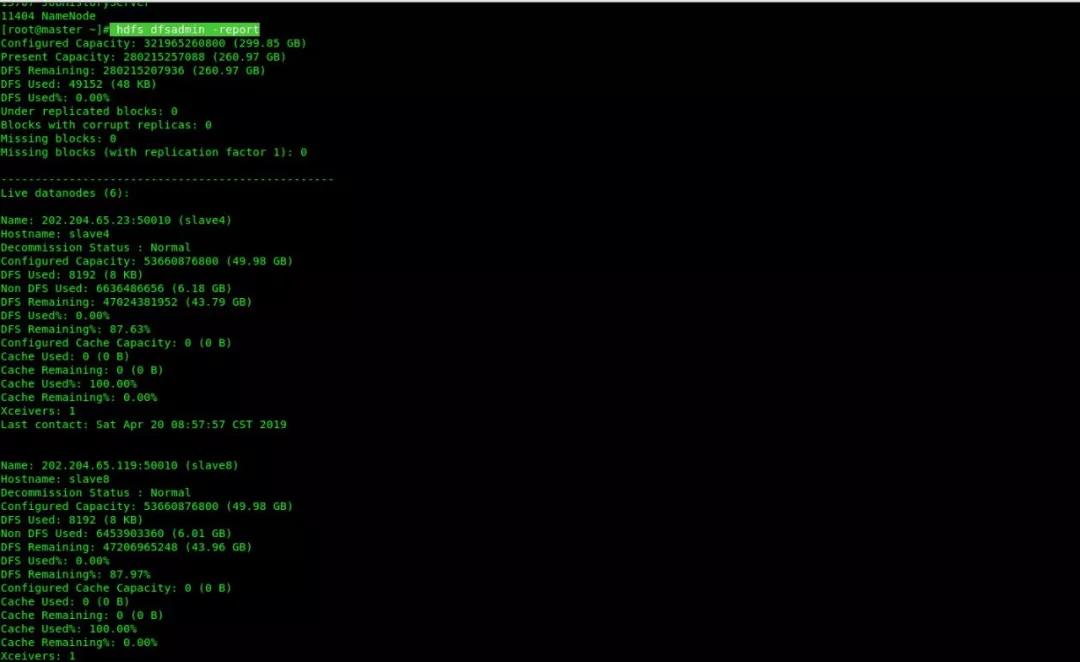
11)接着可以启动 hadoop 了，启动需要在 Master 节点上进行

12)通过jps命令查看各节点启动进程，

13)查看 DataNode 是否正常启动

(1)命令行方式

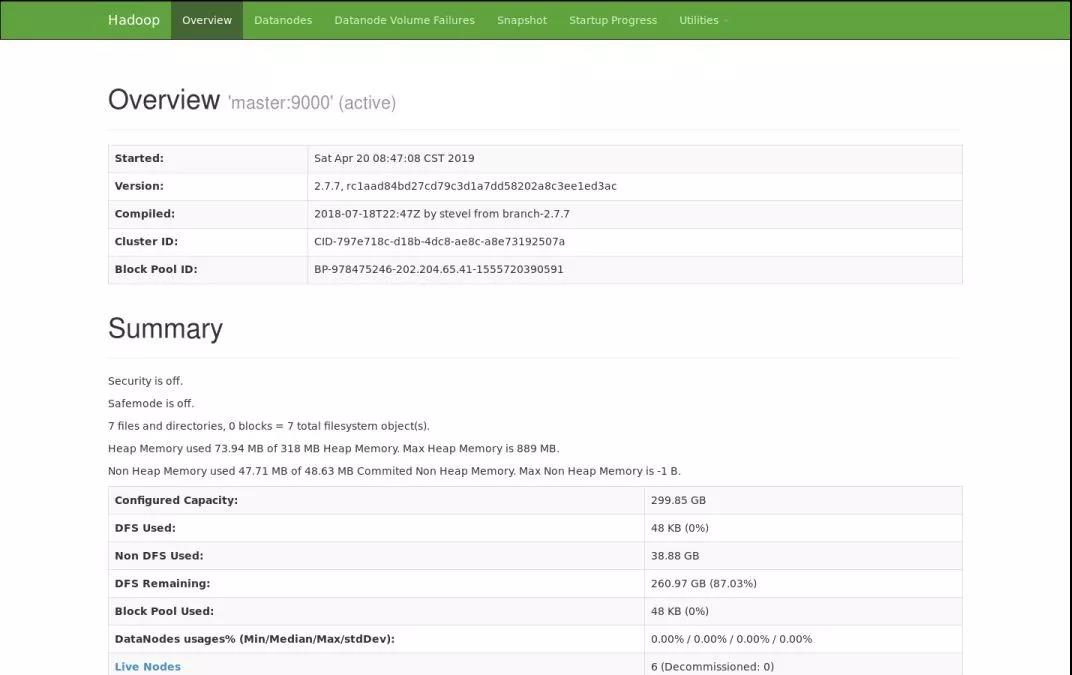
$> hdfs dfsadmin –report



如果 Live datanodes 不为 0 ，则说明集群启动成功

(2)Web方法

查询<http://master:50070/>



7.感受Hadoop的运行

1)命令行方法

2)图形方法：访问http://master:8088/cluster

8.yarn的安装

1）网络yum源

2）本地光盘yum

1. Hadoop环境升级为spark计算环境

本项目的Spark架构图

Hadoop

HDFS

APACHE

Spark

SparkMLlib

可扩展的机器学习库，一个专门用于大数据处理的通用快速的引擎

SparkStreaming

构建在Spark上处理Stream数据的框架

SparkSQL

用于在spark平台上处理结构化数据

升级原因：

* + 1. HDFS文件不能追加
    2. 硬盘读写IO开销太大，不适合项目组的算法
    3. Spark具有能追加文件，IO开销小的优势

因此项目组将Hadoop环境升级为了spark计算环境

升级过程：

1. 首先需要配置好Hadoop的分布式环境
2. 在Master节点上下载spark
3. 安装spark，解压，配置环境变量
4. 配置spark
5. 配置slaves
6. 配置spark-env.sh
7. 传到其他slaver节点上
8. 解压其他节点
9. 启动spark
10. 启动hadoop
11. 启动spark
12. 启动master
13. 启动slave
14. 启动spark-shell

4、可在master主机上打开浏览器，访问http://master:8080查看是否启动成功

（实际由于端口8080与其他端口共用产生了冲突，因此将spark配置文件中的访问端口改为了8001）

1. Linux安装JVM运行Scala程序
2. 安装Scala

* 1. 未来工作安排与展望：

1、 由于Scala语言与Spark能够紧密集成，其中的Scala可以像操作本地集合对象一样轻松地操作分布式数据集。因此项目组决定运用Scala语言将处理用户观看视频时的行为痕迹记录进行修改优化，使其能在spark计算环境中正常运行

2、 思考论文的撰写，确定论文框架，对前阶段所获取的相关信息进行系统的梳理以及筛选并融入论文的撰写当中