UM-SJTU JOINT INSTITUTE

Introduction to Operating Systems

VE482

Laboratory Report

Lab 2

Name: Lu Mingxuan ID: 518021911166

Name: Shen Shiyu ID: 517370910132

Date: 26 May 2021

1 Tasks

In this lab, we set up Hadoop and constructed a cluster with two computers to execute a MapReduce program.

We also studied how the speed evolves as the size of the file increases in the case where the program runs on a single computer and a Hadoop cluster separately.

number of rows	cluster	single
100000	8.489	20.541
200000	19.246	23.202
300000	25.349	27.428
400000	32.385	35.672
500000	38.832	43.559
600000	44.989	49.406
700000	50.216	56.261
800000	55.234	62.593
900000	64.293	68.899
1000000	65.346	73.504

Table 1: Test Result with Hadoop

We can see that the performance on the multi-node cluster is generally better than on a single node cluster, and we can expect that the performance can be improved when there are more nodes connecting to the cluster.

2 TroubleShooting

There are several issues we met in the hadoop cluster setup.

- Only run $\mathrm{hdfs\ namenode\ -format}$ once. After successfully setting up a cluster once, information of the namenodes and datanodes, including $\mathrm{clusterID}$ is determined. If the format command is run again on the master node, the clusterID of the namenodes and datanodes of the master node will be assigned a new value, while those of any slave node won't, leading to potential failure when starting datanodes on slave nodes.
  - **Solution:** cd to your namenode path. Copy the clusterID in the $\mathrm{current/VERSION}$ file on the master node . On the slave node, cd to the datanode path. Replace the clusterID in the $\mathrm{current/VERSION}$ file. Make sure all clusterIDs are the same.

3 Codes

The codes are attached in the src, all the details are listed in README.md