|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 《大数据技术原理与应用》实验报告 | | | | |
| 题目： | MapReduce求ncdc每年最低温度 | 姓名： |  | 日期： |
| 实验环境：  操作系统：  PRETTY\_NAME="Ubuntu 22.04.1 LTS"  NAME="Ubuntu"  VERSION\_ID="22.04"  VERSION="22.04.1 LTS (Jammy Jellyfish)"  VERSION\_CODENAME=jammy  ID=ubuntu  ID\_LIKE=debian  HOME\_URL="https://www.ubuntu.com/"  SUPPORT\_URL="https://help.ubuntu.com/"  BUG\_REPORT\_URL="https://bugs.launchpad.net/ubuntu/"  PRIVACY\_POLICY\_URL="https://www.ubuntu.com/legal/terms-and-policies/privacy-policy"  UBUNTU\_CODENAME=jammy  Hadoop 版本：3.3.6。  JDK 版本：1.8。  Java IDE：IntelliJ Idea Umltimate 远程开发） | | | | |
| 实验内容与完成情况：  1、下载数据：  下载地址：<ftp://ftp.ncdc.noaa.gov/pub/data/>  通过wget下载：wget -r -c <ftp://ftp.ncdc.noaa.gov/pub/data/gsod/2017>    2、解压合并数据：  下载后的数据是多个gz包，需要处理成我们能用的文本数据。  $ zcat \*.gz > ../ncdc.txt    查看12行数据：  $ head -12 ncdc.txt    数据格式说明：wget -r -c ftp://ftp.ncdc.noaa.gov/pub/data/gsod/readme.txt  删除标题行，使用sed命令删除匹配STN的行  $ sed -i '/STN/d' ncdc.txt    3、上传准备好的数据至hdfs：  hadoop fs -mkdir /in  hadoop fs -copyFromLocal ncdc.txt /data    4、编写求最低温度的mapreduce代码：  MinTemperature.java  import org.apache.hadoop.fs.Path;  import org.apache.hadoop.io.IntWritable;  import org.apache.hadoop.io.Text;  import org.apache.hadoop.mapreduce.Job;  import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;  import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;    public class MinTemperature {    public static void main(String[] args) throws Exception {  if(args.length != 2) {  System.err.println("Usage: MinTemperature<input path> <output path>");  System.exit(-1);  }    Job job = new Job();  job.setJarByClass(MinTemperature.class);  job.setJobName("Min temperature");  FileInputFormat.addInputPath(job, new Path(args[0]));  FileOutputFormat.setOutputPath(job, new Path(args[1]));  job.setMapperClass(MinTemperatureMapper.class);  job.setReducerClass(MinTemperatureReducer.class);  job.setOutputKeyClass(Text.class);  job.setOutputValueClass(IntWritable.class);  System.exit(job.waitForCompletion(true) ? 0 : 1);  }  }  MinTemperatureMapper.java  import java.io.IOException;  import org.apache.hadoop.io.IntWritable;  import org.apache.hadoop.io.LongWritable;  import org.apache.hadoop.io.Text;  import org.apache.hadoop.mapreduce.Mapper;    public class MinTemperatureMapper extends Mapper<LongWritable, Text, Text, IntWritable>{    private static final int MISSING = 9999;    @Override  public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {    String line = value.toString();  String year = line.substring(14, 18);    int airTemperature;    airTemperature =(int)Math.floor(Double.valueOf(line.substring(24, 30).trim()));    if(airTemperature != MISSING )  {  context.write(new Text(year), new IntWritable(airTemperature));  }  }  }  MinTemperatureReducer.java  import java.io.IOException;  import org.apache.hadoop.io.IntWritable;  import org.apache.hadoop.io.Text;  import org.apache.hadoop.mapreduce.Reducer;    public class MinTemperatureReducer extends Reducer<Text, IntWritable, Text, IntWritable> {    @Override  public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {    int minValue = Integer.MAX\_VALUE;  for(IntWritable value : values) {  minValue = Math.min(minValue, value.get());  }  context.write(key, new IntWritable(minValue));  }  }  5、编译代码：  2.7.\* 版本没有core这个包，需要在~/.bashrc这个文件中增加以下两列  $ gedit ~/.bashrc  export HADOOP\_HOME=/soft/hadoop-2.7.2  export CLASSPATH=$($HADOOP\_HOME/bin/hadoop classpath):$CLASSPATH  使其生效：  $ source ~/.bashrc  编译，这事没有错误  $javac \*.java  打包，移动到上级目录，并删除所有class：  $jar cvf ./MinTemperature.jar ./\*.class  mv \*.jar ..  rm \*.class  6、执行mapreduce最低气温例子  hadoop jar MinTemperature.jar MinTemperature /data/ncdc.txt /ncdc      hadoop fs -cat /ncdc/part-r-00000 | | | | |
| 出现的问题：  1.直接访问速度过慢，下载数据过慢。 | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）：   1. 使用代理方式，学术加速解决。 | | | | |