

编写简单的Mapreduce程序并部署在Hadoop2.2.0上运行

经过几天的折腾，终于配置好了Hadoop2.2.0(如何配置在Linux平台部署Hadoop请参见本博客[《在Fedora上部署Hadoop2.2.0伪分布式平台》](#))，今天主要来说说怎么在Hadoop2.2.0伪分布式上面运行我们写好的Mapreduce程序。先给出这个程序所依赖的Maven包：

```
<dependencies>
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-mapreduce-client-core</artifactId>
    <version>2.1.1-beta</version>
  </dependency>
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-common</artifactId>
    <version>2.1.1-beta</version>
  </dependency>
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-mapreduce-client-common</artifactId>
    <version>2.1.1-beta</version>
  </dependency>
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-mapreduce-client-jobclient</artifactId>
    <version>2.1.1-beta</version>
  </dependency>
</dependencies>
```

记得加上

```
<dependency>
  <groupId>org.apache.hadoop</groupId>
  <artifactId>hadoop-mapreduce-client-common</artifactId>
  <version>2.1.1-beta</version>
</dependency>
<dependency>
  <groupId>org.apache.hadoop</groupId>
  <artifactId>hadoop-mapreduce-client-jobclient</artifactId>
  <version>2.1.1-beta</version>
```

</dependency>

否则运行程序的时候将会出现一下的异常：

```
Exception in thread "main" java.io.IOException: Cannot initialize Cluster.  
Please check your configuration for mapreduce.framework.name and the  
correspond server addresses.  
at org.apache.hadoop.mapreduce.Cluster.initialize(Cluster.java:120)  
at org.apache.hadoop.mapreduce.Cluster.<init>(Cluster.java:82)  
at org.apache.hadoop.mapreduce.Cluster.<init>(Cluster.java:75)  
at org.apache.hadoop.mapred.JobClient.init(JobClient.java:465)  
at org.apache.hadoop.mapred.JobClient.<init>(JobClient.java:444)  
at org.apache.hadoop.mapred.JobClient.runJob(JobClient.java:826)  
at com.wyp.hadoop.MaxTemperature.main(MaxTemperature.java:41)  
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  
at sun.reflect.NativeMethodAccessorImpl.invoke  
    (NativeMethodAccessorImpl.java:57)  
at sun.reflect.DelegatingMethodAccessorImpl.invoke  
    (DelegatingMethodAccessorImpl.java:43)  
at java.lang.reflect.Method.invoke(Method.java:606)  
at com.intellij.rt.execution.application.AppMain.main(AppMain.java:120)
```

好了，现在给出程序，代码如下：

```
package com.wyp.hadoop;  
  
import org.apache.hadoop.io.IntWritable;  
import org.apache.hadoop.io.LongWritable;  
import org.apache.hadoop.io.Text;  
import org.apache.hadoop.mapred.*;  
  
import java.io.IOException;  
  
/**  
 * User: wyp  
 * Date: 13-10-25  
 * Time: 下午3:26  
 * Email:wyphao.2007@163.com  
 */
```

```
public class MaxTemperatureMapper extends MapReduceBase
    implements Mapper<LongWritable, Text,
        Text,IntWritable>{
    private static final int MISSING = 9999;

    @Override
    public void map(LongWritable key, Text value,
        OutputCollector<Text, IntWritable> output,
        Reporter reporter) throws IOException {

        String line = value.toString();
        String year = line.substring(15, 19);
        int airTemperature;
        if(line.charAt(87) == '+'){
            airTemperature = Integer.parseInt(line.substring(88, 92));
        }else{
            airTemperature = Integer.parseInt(line.substring(87, 92));
        }

        String quality = line.substring(92, 93);
        if(airTemperature != MISSING && quality.matches("[01459]")){
            output.collect(new Text(year), new IntWritable(airTemperature));
        }
    }
}

package com.wyp.hadoop;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;

import java.io.IOException;
import java.util.Iterator;

/**
 * User: wyp
 * Date: 13-10-25
 * Time: 下午3:36
 * Email:wyphao.2007@163.com
 */
public class MaxTemperatureReducer extends MapReduceBase
    implements Reducer<Text, IntWritable,
```

```
        Text, IntWritable> {
    @Override
    public void reduce(Text key, Iterator<IntWritable> values,
        OutputCollector<Text, IntWritable> output,
        Reporter reporter) throws IOException {
        int maxValue = Integer.MIN_VALUE;
        while (values.hasNext()){
            maxValue = Math.max(maxValue, values.next().get());
        }

        output.collect(key, new IntWritable(maxValue));
    }
}

package com.wyp.hadoop;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;

import java.io.IOException;

/**
 * User: wyp
 * Date: 13-10-25
 * Time: 下午3:40
 * Email:wyphao.2007@163.com
 */
public class MaxTemperature {

    public static void main(String[] args) throws IOException {
        if(args.length != 2){
            System.err.println("Error!");
            System.exit(1);
        }

        JobConf conf = new JobConf(MaxTemperature.class);
        conf.setJobName("Max Temperature");

        FileInputFormat.addInputPath(conf, new Path(args[0]));
        FileOutputFormat.setOutputPath(conf, new Path(args[1]));
        conf.setMapperClass(MaxTemperatureMapper.class);
    }
}
```

```
conf.setReducerClass(MaxTemperatureReducer.class);
conf.setOutputKeyClass(Text.class);
conf.setOutputValueClass(IntWritable.class);

JobClient.runJob(conf);

}
}
```

将上面的程序编译和打包成jar文件，然后开始在Hadoop2.2.0（本文假定用户都部署好了Hadoop2.2.0）上面部署了。下面主要讲讲如何去部署：

首先，启动Hadoop2.2.0，命令如下：

```
[wyp@wyp hadoop]$ sbin/start-dfs.sh
[wyp@wyp hadoop]$ sbin/start-yarn.sh
```

如果你想看看Hadoop2.2.0是否运行成功，运行下面的命令去查看

```
[wyp@wyp hadoop]$ jps
9582 Main
9684 RemoteMavenServer
16082 Jps
7011 DataNode
7412 ResourceManager
7528 NodeManager
7222 SecondaryNameNode
6832 NameNode
```

其中jps是jdk自带的一个命令，在jdk/bin目录下。如果你电脑上面出现了以上的几个进程（NameNode、SecondaryNameNode、NodeManager、ResourceManager、DataNode这五个进程必须出现！）说明你的Hadoop服务器启动成功了！现在来运行上面打包好的jar文件（这里为Hadoop.jar，其中/home/wyp/IdeaProjects/Hadoop/out/artifacts/Hadoop_jar/Hadoop.jar是它的绝对路径，不知道绝对路径是什么？那你好好去学学吧！），运行下面的命令：

```
[wyp@wyp Hadoop_jar]$ /home/wyp/Downloads/hadoop/bin/hadoop jar \W
/home/wyp/IdeaProjects/Hadoop/out/artifacts/Hadoop_jar/Hadoop.jar \W
com/wyp/hadoop/MaxTemperature \W
```

```
/user/wyp/data.txt W  
/user/wyp/result
```

(上面是一条命令,由于太长了,所以我分行写,在实际情况中,请写一行!)其中,/home/wyp/Downloads/hadoop/bin/hadoop是hadoop的绝对路径,如果你在环境变量中配置好hadoop命令的路径就不需要这样写;com/wyp/hadoop/MaxTemperature是上面程序的main函数的入口;/user/wyp/data.txt是Hadoop文件系统(HDFS)中的绝对路径(注意:这里不是你Linux系统中的绝对路径!),为需要分析文件的路径(也就是input);/user/wyp/result是分析结果输出的绝对路径(注意:这里不是你Linux系统中的绝对路径!而是HDFS上面的路径!而且/user/wyp/result一定不能存在,否则会抛出异常!这是Hadoop的保护机制,你总不想你以前运行好几天的程序突然被你不小心给覆盖掉了吧?所以,如果/user/wyp/result存在,程序会抛出异常,很不错啊)。好了。输入上面的命令,应该会得到下面类似的输出:

```
13/10/28 15:20:44 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032  
13/10/28 15:20:44 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032  
13/10/28 15:20:45 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not  
performed. Implement the Tool interface and execute your application with ToolRunner to re  
medy this.  
13/10/28 15:20:45 WARN mapreduce.JobSubmitter: No job jar file set. User classes may not b  
e found. See Job or Job#setJar(String).  
13/10/28 15:20:45 INFO mapred.FileInputFormat: Total input paths to process : 1  
13/10/28 15:20:46 INFO mapreduce.JobSubmitter: number of splits:2  
13/10/28 15:20:46 INFO Configuration.deprecation: user.name is deprecated. Instead, use ma  
preduce.job.user.name  
13/10/28 15:20:46 INFO Configuration.deprecation: mapred.output.value.class is deprecated. I  
nstead, use mapreduce.job.output.value.class  
13/10/28 15:20:46 INFO Configuration.deprecation: mapred.job.name is deprecated. Instead,  
use mapreduce.job.name  
13/10/28 15:20:46 INFO Configuration.deprecation: mapred.input.dir is deprecated. Instead, u  
se mapreduce.input.fileinputformat.inputdir  
13/10/28 15:20:46 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead,  
use mapreduce.output.fileoutputformat.outputdir  
13/10/28 15:20:46 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead,  
use mapreduce.job.maps  
13/10/28 15:20:46 INFO Configuration.deprecation: mapred.output.key.class is deprecated. In  
stead, use mapreduce.job.output.key.class  
13/10/28 15:20:46 INFO Configuration.deprecation: mapred.working.dir is deprecated. Instead  
, use mapreduce.job.working.dir  
13/10/28 15:20:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1382942307  
976_0008  
13/10/28 15:20:47 INFO mapred.YARNRunner: Job jar is not present. Not adding any jar to the  
list of resources.  
13/10/28 15:20:49 INFO impl.YarnClientImpl: Submitted application application_13829423079
```

76_0008 to ResourceManager at /0.0.0.0:8032

13/10/28 15:20:49 INFO mapreduce.Job: The url to track the job: http://wyp:8088/proxy/application_1382942307976_0008/

13/10/28 15:20:49 INFO mapreduce.Job: Running job: job_1382942307976_0008

13/10/28 15:20:59 INFO mapreduce.Job: Job job_1382942307976_0008 running in uber mode : false

13/10/28 15:20:59 INFO mapreduce.Job: map 0% reduce 0%

13/10/28 15:21:35 INFO mapreduce.Job: map 100% reduce 0%

13/10/28 15:21:38 INFO mapreduce.Job: map 0% reduce 0%

13/10/28 15:21:38 INFO mapreduce.Job: Task Id : attempt_1382942307976_0008_m_000000_0, Status : FAILED

Error: java.lang.RuntimeException: Error in configuring object

at org.apache.hadoop.util.ReflectionUtils.setJobConf(ReflectionUtils.java:109)

at org.apache.hadoop.util.ReflectionUtils.setConf(ReflectionUtils.java:75)

at org.apache.hadoop.util.ReflectionUtils.newInstance(ReflectionUtils.java:133)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:425)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)

at org.apache.hadoop.mapred.YarnChild\$2.run(YarnChild.java:162)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:415)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491)

)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)

Caused by: java.lang.reflect.InvocationTargetException

at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)

at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)

at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)

at java.lang.reflect.Method.invoke(Method.java:606)

at org.apache.hadoop.util.ReflectionUtils.setJobConf(ReflectionUtils.java:106)

... 9 more

Caused by: java.lang.RuntimeException: java.lang.RuntimeException: java.lang.ClassNotFoundException:

Exception: Class com.wyp.hadoop.MaxTemperatureMapper1 not found

at org.apache.hadoop.conf.Configuration.getClass(Configuration.java:1752)

at org.apache.hadoop.mapred.JobConf.getMapperClass(JobConf.java:1058)

at org.apache.hadoop.mapred.MapRunner.configure(MapRunner.java:38)

... 14 more

Caused by: java.lang.RuntimeException: java.lang.ClassNotFoundException: Class com.wyp.hadoop.MaxTemperatureMapper1 not found

at org.apache.hadoop.conf.Configuration.getClass(Configuration.java:1720)

at org.apache.hadoop.conf.Configuration.getClass(Configuration.java:1744)

... 16 more

Caused by: java.lang.ClassNotFoundException: Class com.wyp.hadoop.MaxTemperatureMapper1 not found

at org.apache.hadoop.conf.Configuration.getClassByName(Configuration.java:1626)

at org.apache.hadoop.conf.Configuration.getClass(Configuration.java:1718)

... 17 more

Container killed by the ApplicationMaster.
Container killed on request. Exit code is 143

程序居然抛出异常（ClassNotFoundException）！这是什么回事？其实我也不太明白！！

在网上Google了一下，找到别人的观点：

经个人总结，这通常是由于以下几种原因造成的：

- （1）你编写了一个java lib，封装成了jar，然后再写了一个Hadoop程序，调用这个jar完成mapper和reducer的编写
 - （2）你编写了一个Hadoop程序，期间调用了一个第三方java lib。
- 之后，你将自己的jar包或者第三方java包分发到各个TaskTracker的HADOOP_HOME目录下，运行你的JAVA程序，报了以上错误。

那怎么解决呢？一个笨重的方法是，在运行Hadoop作业的时候，先运行下面的命令：

```
[wyp@wyp Hadoop_jar]$ export W  
HADOOP_CLASSPATH=/home/wyp/IdeaProjects/Hadoop/out/artifacts/Hadoop_jar/
```

其中，/home/wyp/IdeaProjects/Hadoop/out/artifacts/Hadoop_jar/是上面Hadoop.jar文件所在的目录。好了，现在再运行一下Hadoop作业命令：

有一个比较推荐的方法，就是在提交作业的时候加上-libjars参数，后面跟着需要的类库的绝对路径。

```
[wyp@wyp Hadoop_jar]$ hadoop jar /home/wyp/IdeaProjects/Hadoop/out/artifacts/Hadoop_jar/Hadoop.jar com/wyp/hadoop/MaxTemperature /user/wyp/data.txt /user/wyp/result
13/10/28 15:34:16 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
13/10/28 15:34:16 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
13/10/28 15:34:17 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
13/10/28 15:34:17 INFO mapred.FileInputFormat: Total input paths to process : 1
13/10/28 15:34:17 INFO mapreduce.JobSubmitter: number of splits:2
13/10/28 15:34:17 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
13/10/28 15:34:17 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
```

```
13/10/28 15:34:17 INFO Configuration.deprecation: mapred.output.value.class is deprecated. Instead, use mapreduce.job.output.value.class
13/10/28 15:34:17 INFO Configuration.deprecation: mapred.job.name is deprecated. Instead, use mapreduce.job.name
13/10/28 15:34:17 INFO Configuration.deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
13/10/28 15:34:17 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
13/10/28 15:34:17 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
13/10/28 15:34:17 INFO Configuration.deprecation: mapred.output.key.class is deprecated. Instead, use mapreduce.job.output.key.class
13/10/28 15:34:17 INFO Configuration.deprecation: mapred.working.dir is deprecated. Instead, use mapreduce.job.working.dir
13/10/28 15:34:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1382942307976_0009
13/10/28 15:34:18 INFO impl.YarnClientImpl: Submitted application application_1382942307976_0009 to ResourceManager at /0.0.0.0:8032
13/10/28 15:34:18 INFO mapreduce.Job: The url to track the job: http://wyp:8088/proxy/application_1382942307976_0009/
13/10/28 15:34:18 INFO mapreduce.Job: Running job: job_1382942307976_0009
13/10/28 15:34:26 INFO mapreduce.Job: Job job_1382942307976_0009 running in uber mode : false
13/10/28 15:34:26 INFO mapreduce.Job: map 0% reduce 0%
13/10/28 15:34:41 INFO mapreduce.Job: map 50% reduce 0%
13/10/28 15:34:53 INFO mapreduce.Job: map 100% reduce 0%
13/10/28 15:35:17 INFO mapreduce.Job: map 100% reduce 100%
13/10/28 15:35:18 INFO mapreduce.Job: Job job_1382942307976_0009 completed successfully
13/10/28 15:35:18 INFO mapreduce.Job: Counters: 43
File System Counters
  FILE: Number of bytes read=144425
  FILE: Number of bytes written=524725
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=1777598
  HDFS: Number of bytes written=18
  HDFS: Number of read operations=9
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=38057
  Total time spent by all reduces in occupied slots (ms)=24800
```

```
Map-Reduce Framework
Map input records=13130
Map output records=13129
Map output bytes=118161
Map output materialized bytes=144431
Input split bytes=182
Combine input records=0
Combine output records=0
Reduce input groups=2
Reduce shuffle bytes=144431
Reduce input records=13129
Reduce output records=2
Spilled Records=26258
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=321
CPU time spent (ms)=5110
Physical memory (bytes) snapshot=552824832
Virtual memory (bytes) snapshot=1228738560
Total committed heap usage (bytes)=459800576

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=1777416
File Output Format Counters
Bytes Written=18
```

到这里，程序就成功运行了！很高兴吧？那么怎么查看刚刚程序运行的结果呢？很简单，运行下面命令：

```
[wyp@wyp Hadoop_jar]$ hadoop fs -ls /user/wyp
Found 2 items
-rw-r--r-- 1 wyp supergroup 1777168 2013-10-25 17:44 /user/wyp/data.txt
drwxr-xr-x - wyp supergroup 0 2013-10-28 15:35 /user/wyp/result
[wyp@wyp Hadoop_jar]$ hadoop fs -ls /user/wyp/result
Found 2 items
-rw-r--r-- 1 wyp supergroup 0 2013-10-28 15:35 /user/wyp/result/_SUCCESS
```

```
-rw-r--r-- 1 wyp supergroup 18 2013-10-28 15:35 /user/wyp/result/part-00000  
[wyp@wyp Hadoop_jar]$ hadoop fs -cat /user/wyp/result/part-00000  
1901 317  
1902 244
```

到此，你自己写好的一个Mapreduce程序终于成功运行了！
附程序测试的数据的下载地址：<http://pan.baidu.com/s/1iSacM>

本博客文章除特别声明，全部都是原创！
原创文章版权归过往记忆大数据（[过往记忆](#)）所有，未经许可不得转载。
本文链接: **【】**（ ）