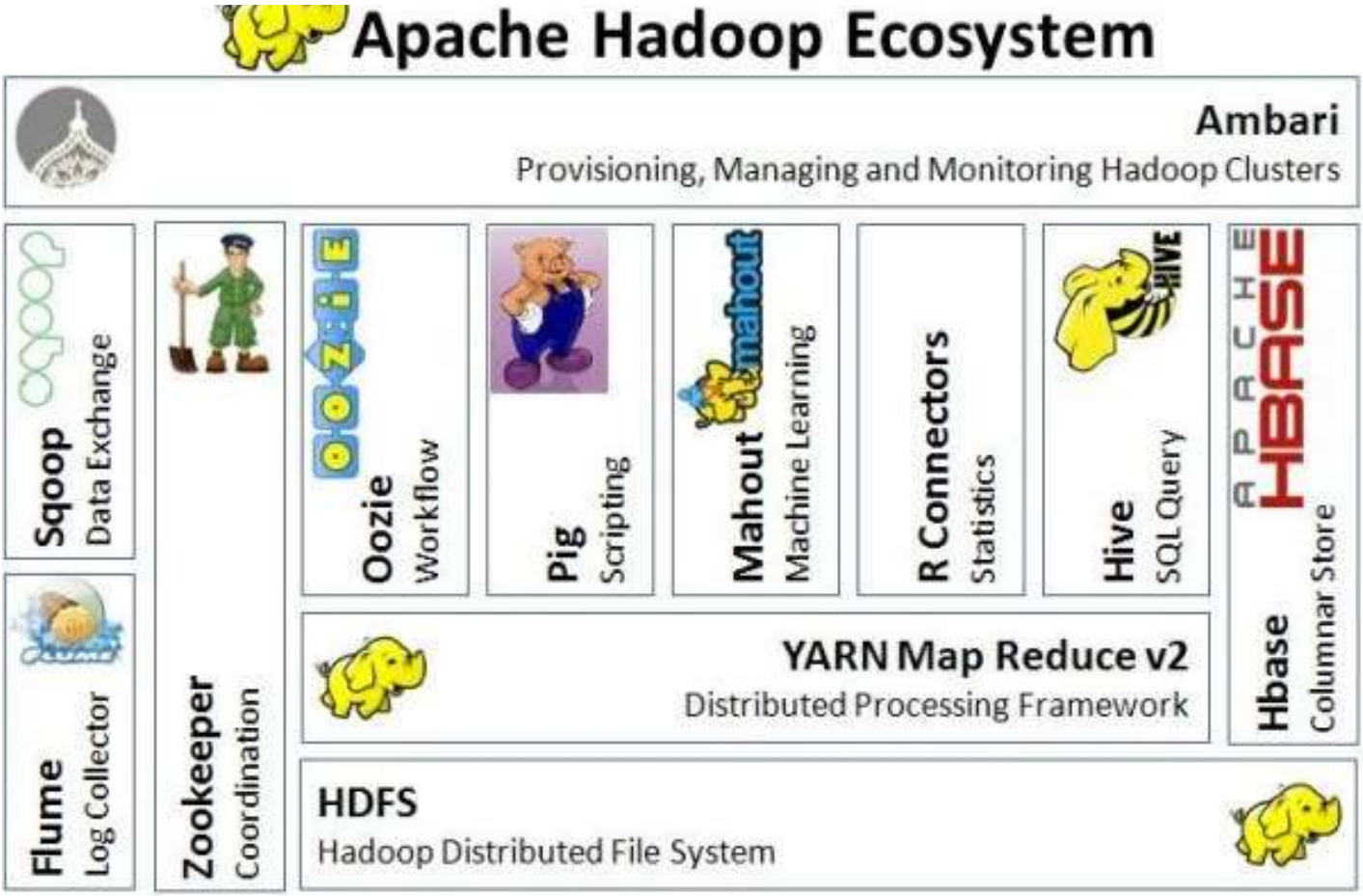
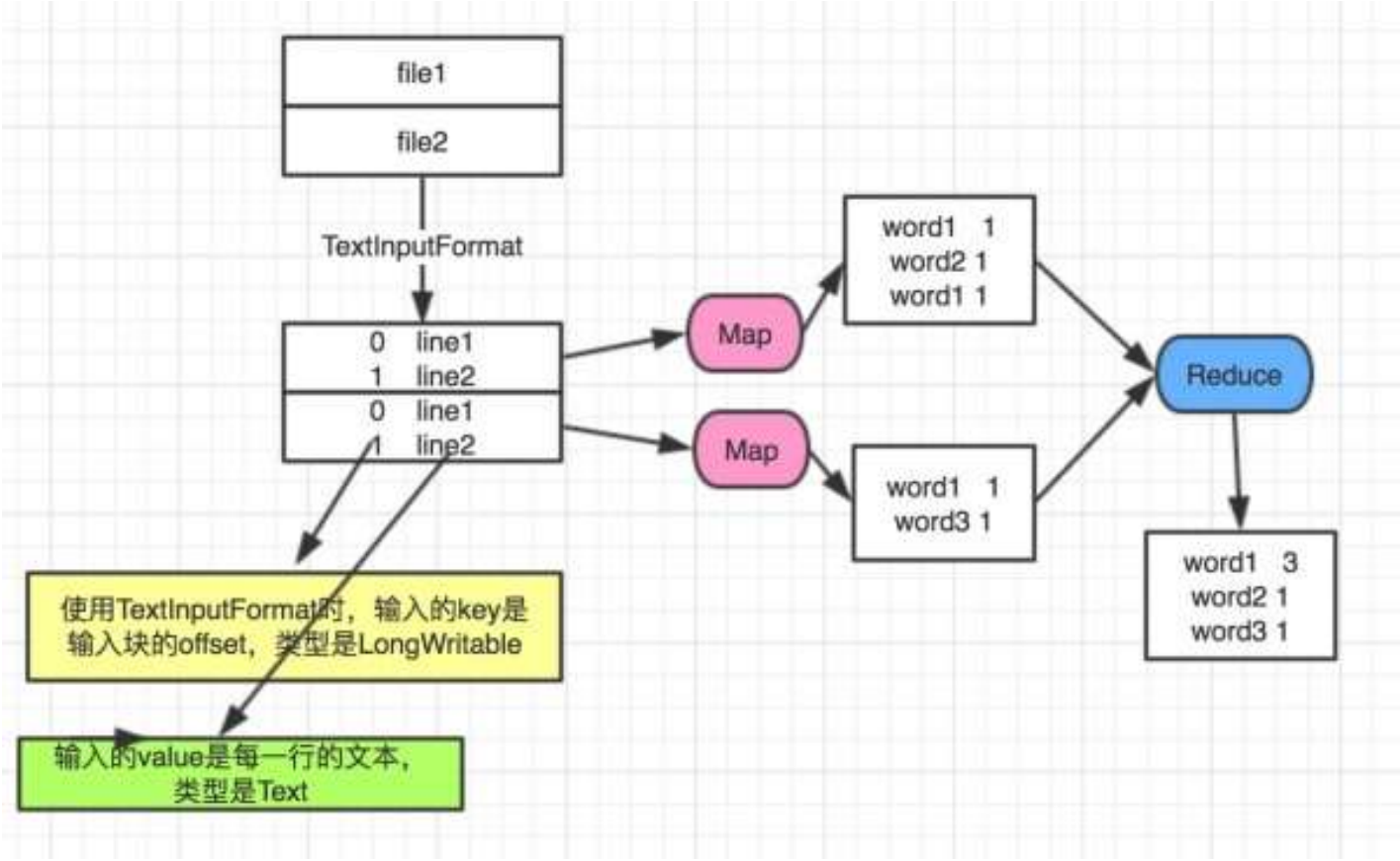


Hadoop入门-WordCount示例

Chown · 17 小时前

WordCount的过程如图，这里记录下入门的过程，虽然有很多地方理解的只是皮毛。



Hadoop的安装

安装比较简单，安装完成后进行单机环境的配置。

hadoop-env.sh:指定JAVA_HOME。

```
# The only required environment variable is JAVA_HOME.  All others are
# optional.  When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME="$(/usr/libexec/java_home)"
```

core-site.xml：设置Hadoop使用的临时目录，NameNode的地址。

```
<configuration>
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/usr/local/Cellar/hadoop/hdfs/tmp</value>
    </property>
    <property>
```

```
        <name>fs.default.name</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

hdfs-site.xml：一个节点，副本个数设为1。

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
```

mapred-site.xml:指定JobTracker的地址。

```
<configuration>
    <property>
        <name>mapred.job.tracker</name>
        <value>localhost:9010</value>
    </property>
</configuration>
```

启动Hadoop相关的所有进程。

```
➜  sbin git:(master) ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
16/12/03 19:32:18 WARN util.NativeCodeLoader: Unable to load native-hadoop librar
Starting namenodes on [localhost]
Password:
localhost: starting namenode, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/l
Password:
localhost: starting datanode, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/l
Starting secondary namenodes [0.0.0.0]
Password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cellar/hadoop/2.7.1/l
16/12/03 19:33:27 WARN util.NativeCodeLoader: Unable to load native-hadoop librar
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/
Password:
localhost: starting nodemanager, logging to /usr/local/Cellar/hadoop/2.7.1/libexe
```

（可以配置ssh无密码登录方式，否则启动hadoop的时候总是要密码。）

看看启动了哪些组件。

```
➜  sbin git:(master) jps -l
5713 org.apache.hadoop.hdfs.server.namenode.NameNode
6145 org.apache.hadoop.yarn.server.nodemanager.NodeManager
6044 org.apache.hadoop.yarn.server.resourcemanager.ResourceManager
5806 org.apache.hadoop.hdfs.server.datanode.DataNode
5918 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
```

访问localhost:50070/ 可以看到DFS的一些状态。

WordCount 单词计数

WordCount就是Hadoop学习的hello world，代码如下：

```
public class WordCount {

    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(LongWritable key, Text value, Context context)
                throws IOException, InterruptedException {
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            while (tokenizer.hasMoreTokens()) {
                word.set(tokenizer.nextToken());
                context.write(word, one);
            }
        }
    }

    public static class Reduce extends
            Reducer<Text, IntWritable, Text, IntWritable> {

        public void reduce(Text key, Iterable<IntWritable> values,
                           Context context) throws IOException, InterruptedExcep
            int sum = 0;
            for (IntWritable val : values) {
                sum += val.get();
```

```
            }
            context.write(key, new IntWritable(sum));
        }
    }

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();

        Job job = new Job(conf, "wordcount");
        job.setJarByClass(WordCount.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        /**
         * 设置一个本地combine,可以极大的消除本节点重复单词的计数,减小网络传输的
         */
        job.setCombinerClass(Reduce.class);

        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        job.waitForCompletion(true);
    }
}
```

构造两个文本文件, 把本地的两个文件拷贝到HDFS中：

```
➜  hadoop-examples git:(master) ✗ ln /usr/local/Cellar/hadoop/2.7.1/bin/hadoop
➜  hadoop-examples git:(master) ✗ ./hadoop dfs -put wordcount-input/file* input
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

16/12/03 23:17:10 WARN util.NativeCodeLoader: Unable to load native-hadoop librar
➜  hadoop-examples git:(master) ✗ ./hadoop dfs -ls input/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
```

```
16/12/03 23:21:08 WARN util.NativeCodeLoader: Unable to load native-hadoop librar
Found 2 items
-rw-r--r--   1 vonzhou supergroup         42 2016-12-03 23:17 input/file1
-rw-r--r--   1 vonzhou supergroup         43 2016-12-03 23:17 input/file2
```

编译程序得到jar:

```
mvn clean package
```

运行程序（指定main class的时候需要全包名限定）：

```
➜  hadoop-examples git:(master) ✗ ./hadoop jar target/hadoop-examples-1.0-SNAPS
16/12/03 23:31:19 WARN util.NativeCodeLoader: Unable to load native-hadoop librar
16/12/03 23:31:20 INFO Configuration.deprecation: session.id is deprecated. Inste
16/12/03 23:31:20 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=
16/12/03 23:33:21 WARN mapreduce.JobResourceUploader: Hadoop command-line option
16/12/03 23:33:21 INFO input.FileInputFormat: Total input paths to process : 2
16/12/03 23:33:21 INFO mapreduce.JobSubmitter: number of splits:2
16/12/03 23:33:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_loc
16/12/03 23:33:22 INFO mapreduce.Job: The url to track the job: http://localhost
16/12/03 23:33:22 INFO mapreduce.Job: Running job: job_local524341653_0001
16/12/03 23:33:22 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/12/03 23:33:22 INFO output.FileOutputCommitter: File Output Committer Algorith
16/12/03 23:33:22 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoo
16/12/03 23:33:22 INFO mapred.LocalJobRunner: Waiting for map tasks
16/12/03 23:33:22 INFO mapred.LocalJobRunner: Starting task: attempt_local524341
16/12/03 23:33:22 INFO output.FileOutputCommitter: File Output Committer Algorith
16/12/03 23:33:22 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree curren
16/12/03 23:33:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/12/03 23:33:22 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/us
16/12/03 23:33:22 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/12/03 23:33:22 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/12/03 23:33:22 INFO mapred.MapTask: soft limit at 83886080
16/12/03 23:33:22 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/12/03 23:33:22 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/12/03 23:33:22 INFO mapred.MapTask: Map output collector class = org.apache.ha
16/12/03 23:33:22 INFO mapred.LocalJobRunner:
16/12/03 23:33:22 INFO mapred.MapTask: Starting flush of map output
16/12/03 23:33:22 INFO mapred.MapTask: Spilling map output
16/12/03 23:33:22 INFO mapred.MapTask: bufstart = 0; bufend = 71; bufvoid = 1048
16/12/03 23:33:22 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 262
16/12/03 23:33:22 INFO mapred.MapTask: Finished spill 0
```

```
16/12/03 23:33:22 INFO mapred.Task: Task:attempt_local524341653_0001_m_000000_0 :
16/12/03 23:33:22 INFO mapred.LocalJobRunner: map
16/12/03 23:33:22 INFO mapred.Task: Task 'attempt_local524341653_0001_m_000000_0
16/12/03 23:33:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local52434
16/12/03 23:33:22 INFO mapred.LocalJobRunner: Starting task: attempt_local5243416
16/12/03 23:33:22 INFO output.FileOutputCommitter: File Output Committer Algorith
16/12/03 23:33:22 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree curren
16/12/03 23:33:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/12/03 23:33:22 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/us
16/12/03 23:33:22 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/12/03 23:33:22 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/12/03 23:33:22 INFO mapred.MapTask: soft limit at 83886080
16/12/03 23:33:22 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/12/03 23:33:22 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/12/03 23:33:22 INFO mapred.MapTask: Map output collector class = org.apache.ha
16/12/03 23:33:22 INFO mapred.LocalJobRunner:
16/12/03 23:33:22 INFO mapred.MapTask: Starting flush of map output
16/12/03 23:33:22 INFO mapred.MapTask: Spilling map output
16/12/03 23:33:22 INFO mapred.MapTask: bufstart = 0; bufend = 70; bufvoid = 1048
16/12/03 23:33:22 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 262
16/12/03 23:33:22 INFO mapred.MapTask: Finished spill 0
16/12/03 23:33:22 INFO mapred.Task: Task:attempt_local524341653_0001_m_000001_0 :
16/12/03 23:33:22 INFO mapred.LocalJobRunner: map
16/12/03 23:33:22 INFO mapred.Task: Task 'attempt_local524341653_0001_m_000001_0
16/12/03 23:33:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local52434
16/12/03 23:33:22 INFO mapred.LocalJobRunner: map task executor complete.
16/12/03 23:33:22 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/12/03 23:33:22 INFO mapred.LocalJobRunner: Starting task: attempt_local5243416
16/12/03 23:33:22 INFO output.FileOutputCommitter: File Output Committer Algorith
16/12/03 23:33:22 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree curren
16/12/03 23:33:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/12/03 23:33:22 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache
16/12/03 23:33:23 INFO mapreduce.Job: Job job_local524341653_0001 running in uber
16/12/03 23:33:23 INFO mapreduce.Job:  map 100% reduce 0%
16/12/03 23:33:53 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338
16/12/03 23:33:53 INFO reduce.EventFetcher: attempt_local524341653_0001_r_000000_
16/12/03 23:33:53 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle outpu
16/12/03 23:33:53 INFO reduce.InMemoryMapOutput: Read 86 bytes from map-output fo
16/12/03 23:33:53 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output o
16/12/03 23:33:53 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle outpu
16/12/03 23:33:53 INFO reduce.InMemoryMapOutput: Read 87 bytes from map-output fo
16/12/03 23:33:53 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output o
16/12/03 23:33:53 INFO reduce.EventFetcher: EventFetcher is interrupted.. Return:
16/12/03 23:33:53 INFO mapred.LocalJobRunner: 2 / 2 copied.
```

```
16/12/03 23:33:53 INFO reduce.MergeManagerImpl: finalMerge called with 2 in-memor
16/12/03 23:33:53 INFO mapred.Merger: Merging 2 sorted segments
16/12/03 23:33:53 INFO mapred.Merger: Down to the last merge-pass, with 2 segment
16/12/03 23:33:53 INFO reduce.MergeManagerImpl: Merged 2 segments, 173 bytes to (
16/12/03 23:33:53 INFO reduce.MergeManagerImpl: Merging 1 files, 175 bytes from (
16/12/03 23:33:53 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from
16/12/03 23:33:53 INFO mapred.Merger: Merging 1 sorted segments
16/12/03 23:33:53 INFO mapred.Merger: Down to the last merge-pass, with 1 segment
16/12/03 23:33:53 INFO mapred.LocalJobRunner: 2 / 2 copied.
16/12/03 23:33:53 INFO Configuration.deprecation: mapred.skip.on is deprecated. 
16/12/03 23:33:53 INFO mapred.Task: Task:attempt_local524341653_0001_r_000000_0 
16/12/03 23:33:53 INFO mapred.LocalJobRunner: 2 / 2 copied.
16/12/03 23:33:53 INFO mapred.Task: Task attempt_local524341653_0001_r_000000_0 
16/12/03 23:33:53 INFO output.FileOutputCommitter: Saved output of task 'attempt_
16/12/03 23:33:53 INFO mapred.LocalJobRunner: reduce > reduce
16/12/03 23:33:53 INFO mapred.Task: Task 'attempt_local524341653_0001_r_000000_0
16/12/03 23:33:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local52434
16/12/03 23:33:53 INFO mapred.LocalJobRunner: reduce task executor complete.
16/12/03 23:33:54 INFO mapreduce.Job:  map 100% reduce 100%
16/12/03 23:33:54 INFO mapreduce.Job: Job job_local524341653_0001 completed succe
16/12/03 23:33:54 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=54188
                FILE: Number of bytes written=917564
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=213
                HDFS: Number of bytes written=89
                HDFS: Number of read operations=22
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=5
        Map-Reduce Framework
                Map input records=5
                Map output records=14
                Map output bytes=141
                Map output materialized bytes=181
                Input split bytes=222
                Combine input records=0
                Combine output records=0
                Reduce input groups=11
                Reduce shuffle bytes=181
                Reduce input records=14
                Reduce output records=11
```

```
                Spilled Records=28
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=7
                Total committed heap usage (bytes)=946864128
        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=85
        File Output Format Counters
                Bytes Written=89
➜  hadoop-examples git:(master) ✗
```

查看执行的结果：

```
➜  hadoop-examples git:(master) ✗ ./hadoop dfs -ls output
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

16/12/03 23:36:42 WARN util.NativeCodeLoader: Unable to load native-hadoop librar
Found 2 items
-rw-r--r--   1 vonzhou supergroup          0 2016-12-03 23:33 output/_SUCCESS
-rw-r--r--   1 vonzhou supergroup         89 2016-12-03 23:33 output/part-r-00000
➜  hadoop-examples git:(master) ✗ ./hadoop dfs -cat output/part-r-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

16/12/03 23:37:03 WARN util.NativeCodeLoader: Unable to load native-hadoop librar
big     1
by      1
data    1
google  1
hadoop  2
hello   2
learning        1
papers  1
step    2
```

```
vonzhou 1
world   1
```

「真诚赞赏，手留余香」

赞赏

还没有人赞赏，快来当第一个赞赏的人吧！

Hadoop MapReduce

1

分享 举报

文章被以下专栏收录

编程之路

进入专栏

1 条评论

写下你的评论

winner0715

大神只在这上面写文章吗，博客呢？看了你的文章受益匪浅，收下我的膝盖

7 小时前

推荐阅读

Maven profile入门实践

Maven profile入门实践前言Java后端开发经常需要面对需要管理多套环境，多种不同配置的情况，有效的管理不同的配置，保持项目结构清晰是非常… 查看全文 >

Chown · 4 天前　　发表于 编程之路

微博炫出2016年的数据，感觉金光四射

导语：付费问答、众筹、一元夺宝、Youtube式的博主管理模式以及你能想到的变现方法，微博可能都要做了。 微博如何走到今天，明天又将如何发展… 查看全文 >

闫浩 · 1 个月前 · 编辑精选　　发表于 杂文漫谈

为什么日系车在北美和欧洲销量差异巨大？

不知道你们了不了解，在北美大红大紫的日系车在欧洲从来都是不温不火。数据不会说谎，在过去二十年，欧洲外来品牌的市场占有率从31%上升到了3… 查看全文 >

Fazioli · 1 个月前 · 编辑精选

把粥熬稠，有什么小窍门？

粥，看起来是最简单不过的一道餐食，大米加水煮一煮便是。但是，如何才能煮出绵、软、滑的粥：米粒开了花，与水交融为一体，晶透莹亮，舀一勺… 查看全文 >

下厨房 · 25 天前 · 编辑精选