

Giraph 基于 MapReduce v1 部署

1 单机伪分布式部署

1.1 准备工作&修改配置并启动Hadoop和Giraph

- 修改 `~/giraph-1.2.0-for-hadoop-1.2.1/bin/giraph-env`, 指定 Hadoop 安装路径：

```
# resolve links - $0 may be a softlink
sed -i '1i\export HADOOP_HOME=~/.hadoop-1.2.1' ~/giraph-1.2.0-for-hadoop-1.2.1/bin/giraph-env
THIS="${BASH_SOURCE:-0}"
while [ -h "$THIS" ]; do
    ls=`ls -ld "$THIS"`
```

- 修改 `~/hadoop-1.2.1/conf/mapred-site.xml`, 结果如下：

```
<configuration>
  <property>
    <name>mapred.job.tracker</name>
    <value>219.228.135.32:9001</value>
  </property>
  <property>
    <name>mapred.tasktracker.map.tasks.maximum</name>
    <value>4</value>
  </property>
  <property>
    <name>mapred.map.tasks</name>
    <value>3</value>
  </property>
</configuration>
```

- 启动 HDFS及 MapReduce

```
1 | ~/hadoop-1.2.1/bin/start-dfs.sh
2 | ~/hadoop-1.2.1/bin/start-mapred.sh
```

- 通过运行 `jps` 来检验进程状态：

```
Mon 23 Dec - 16:41 ~
@wushuangyoyo jps
26803 SecondaryNameNode
27238 Jps
26935 JobTracker
26599 DataNode
26396 NameNode
27135 TaskTracker
```

1.2 运行 Giraph 应用程序

Simple shortest paths computation 示例程序

- 将 `tiny_graph.txt` 上传至 `hdfs:///user/you/input` 下

```
1 ~/hadoop-1.2.1/bin/hadoop fs -mkdir input
2 ~/hadoop-1.2.1/bin/hadoop fs -put ~/tiny_graph.txt input/
```

- 查看 HDFS 的文件信息：

```
@wushuangyoyo ~ -/hadoop-1.2.1/bin/hadoop fs -ls input | grep tiny_graph
Warning: $HADOOP_HOME is deprecated.
-rw-r--r-- 1 wushuangyoyo supergroup 112 2019-12-11 09:49 /user/wushuangyoyo/input/tiny_graph.txt
```

- 执行程序

按照如下代码执行：

```
@wushuangyoyo ~ -/hadoop-1.2.1/bin/hadoop fs -rmr output/shortestpaths # 清空输出路径
cd ~/giraph-1.2.0-for-hadoop-1.2.1
bin/giraph giraph-examples-1.2.0.jar \ 3.1 准备工作
  -v org.apache.giraph.examples.SimpleShortestPathsComputation \ 1 sed -i 'i\export HADOOP_HOME=~/.hadoop-1.2.1' ~/.giraph-1.2.0-
  -vfp input/tiny_graph.txt \
  -vof org.apache.giraph.io.formats.IdWithValueTextOutputFormat \ 修改 ~/hadoop-1.2.1/conf/mapred-site.xml, 在 <configuration> 下添加
  -op output/shortestpaths \ 3.5 查看 Graph 应用程序运
  -w 3 行信息
Warning: $HADOOP_HOME is deprecated. 3.6 关闭 Hadoop
Deleted hdfs://localhost:9000/user/wushuangyoyo/output/shortestpaths
HADOOP_CONF_DIR=/home/wushuangyoyo/hadoop-1.2.1/conf
Warning: $HADOOP_HOME is deprecated.
19/12/23 16:46:22 INFO utils.ConfigurationUtils: No edge input format specified. Ensure your InputFormat does not require one.
19/12/23 16:46:22 INFO utils.ConfigurationUtils: No edge output format specified. Ensure your OutputFormat does not require one.
19/12/23 16:46:22 INFO job.GiraphJob: run: Since checkpointing is disabled (default), do not allow any task retries (Setting mapred.map.m
ax.attempts = 1, old value = 4)
19/12/23 16:46:30 INFO job.GiraphJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=job_201912231641_0001
19/12/23 16:46:30 INFO job.GiraphJob: Waiting for resources... Job will start only when it gets all 4 mappers
19/12/23 16:46:50 INFO job.HaltApplicationUtils$DefaultHaltInstructionsWriter: writeHaltInstructions: To halt after next superstep execut
e: 'bin/halt-application --zkServer localhost:22181 --zkNode /_hadoopBsp/job_201912231641_0001/_haltComputation'
19/12/23 16:46:50 INFO mapred.JobClient: Running job: job_201912231641_0001
19/12/23 16:46:51 INFO mapred.JobClient: map 100% reduce 0%
19/12/23 16:46:53 INFO mapred.JobClient: Job complete: job_201912231641_0001
19/12/23 16:46:53 INFO mapred.JobClient: Counters: 44
```

- 查看运行中进程

```
Mon 23 Dec - 16:48 ~
@wushuangyoyo jps
31360 Child
26803 SecondaryNameNode
31411 Child
26935 JobTracker
26599 DataNode
31433 Child
30202 RunJar
26396 NameNode
31455 Child
27135 TaskTracker
31903 Jps
```

- 运行完成后查看输出

```
@wushuangyoyo ~ -/hadoop-1.2.1/bin/hadoop fs -cat output/shortestpaths/p*
Warning: $HADOOP_HOME is deprecated.
0 1.0
3 1.0
1 0.0
4 5.0
2 2.0
```

1.3 查看 Giraph 应用程序运行信息

- 访问 JobTracker 网页 (<http://localhost:50030>),

localhost:50030/jobtracker.jsp

Quick Links

localhost Hadoop Map/Reduce Administration

State: RUNNING
Started: Mon Dec 23 16:41:02 CST 2019
Version: 1.2.1, r1503152
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf
Identifier: 201912231641
SafeMode: OFF

Cluster Summary (Heap Size is 140 MB/1.74 GB)

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Occupied Reduce Slots	Reserved Map Slots	Reserved Reduce Slots	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes	Graylisted Nodes	Excluded Nodes
0	0	2	1	0	0	0	0	4	2	6.00	0	0	0

Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)
Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

none

Completed Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201912231641_0001	Mon Dec 23 16:46:30 CST 2019	NORMAL	wushuangyoyo	Giraph: org.apache.giraph.examples.SimpleShortestPathsComputation	100.00%	4	4	100.00%	0	0	NA	NA

点击正在运行或已完成的 Giraph 应用程序, 可看到 Giraph 应用程序的统计信息

Hadoop job_201912231641_0002 on localhost

User: wushuangyoyo
Job Name: Giraph: org.apache.giraph.examples.SimpleShortestPathsComputation
Job File: hdfs://localhost:9000/home/wushuangyoyo/pdos-tmp-1.2.1/mapred/staging/wushuangyoyo/staging/job_201912231641_0002/job.xml
Submit Host: Master-yoyo
Submit Host Address: 127.0.0.1
Job-ACLs: All users are allowed
Job Setup: [Successful](#)
Status: Succeeded
Started at: Mon Dec 23 16:48:29 CST 2019
Finished at: Mon Dec 23 16:48:52 CST 2019
Finished in: 22sec
Job Cleanup: [Successful](#)

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	4	0	0	4	0	0 / 0
reduce	100.00%	0	0	0	0	0	0 / 0

	Counter	Map	Reduce	Total
Map-Reduce Framework	Spilled Records	0	0	0
	Virtual memory (bytes) snapshot	0	0	8,750,948,352
	Map input records	0	0	4
	SPLIT_RAW_BYTES	176	0	176
	Map output records	0	0	0
	Physical memory (bytes) snapshot	0	0	893,489,152
	CPU time spent (ms)	0	0	6,690
	Total committed heap usage (bytes)	0	0	1,237,319,680
Zookeeper halt node	/_hadoopBsp/job_201912231641_0002/_haltComputation	0	0	0
Zookeeper server:port	localhost:22181	0	0	0
	Superstep 1 SimpleShortestPathsComputation (ms)	58	0	58
	Initialize (ms)	1,070	0	1,070
	Superstep 0 SimpleShortestPathsComputation (ms)	63	0	63

localhost:50030/jobconf.jsp?jobid=job_201912231641_0002

- 查看程序日志
 - JobHistory 日志默认位置: ~/hadoop-1.2.1/logs
 - [job_201912111815_0002_conf.xml](#) 98508 bytes Dec 11, 2019 6:20:25 PM
 - [job_201912231641_0001_conf.xml](#) 98508 bytes Dec 23, 2019 4:46:30 PM
 - [job_201912231641_0002_conf.xml](#) 98508 bytes Dec 23, 2019 4:48:29 PM
 - [userlogs/](#) 4096 bytes Dec 23, 2019 4:48:30 PM
 - Task 日志默认位置: ~/hadoop-1.2.1/logs/userlogs/<jobid>/<attempt-id>

Directory: /logs/userlogs/job_201912231641_0001/

[Parent Directory](#)

attempt_201912231641_0001_m_000000_0/	4096 bytes	Dec 23, 2019 4:46:53 PM
attempt_201912231641_0001_m_000001_0/	4096 bytes	Dec 23, 2019 4:46:50 PM
attempt_201912231641_0001_m_000002_0/	4096 bytes	Dec 23, 2019 4:46:50 PM
attempt_201912231641_0001_m_000003_0/	4096 bytes	Dec 23, 2019 4:46:50 PM
attempt_201912231641_0001_m_000004_0/	4096 bytes	Dec 23, 2019 4:46:53 PM
attempt_201912231641_0001_m_000005_0/	4096 bytes	Dec 23, 2019 4:46:38 PM
job-acls.xml	507 bytes	Dec 23, 2019 4:46:36 PM

3.6 关闭 Hadoop

- 关闭 HDFS

```
1 ~/hadoop-1.2.1/bin/stop-dfs.sh
```

- 关闭 MapReduce

```
1 ~/hadoop-1.2.1/bin/stop-mapred.sh
```

分布式在构建过程中有很多错误，实在无法进行，故放弃。

Giraph 应用编程实践

1. 编写Giraph程序

- 创建maven项目

参考文档：[create maven.md](#)

- 添加pom依赖

在pom.xml文件中添加以下依赖：`giraph-core`、`giraph-examples`、`hadoop-common` 和 `hadoop-client`。

```
1 <dependencies>
2   <!-- https://mvnrepository.com/artifact/org.apache.giraph/giraph-core -->
3   <dependency>
4     <groupId>org.apache.giraph</groupId>
5     <artifactId>giraph-core</artifactId>
6     <version>1.2.0</version>
7   </dependency>
8   <!-- https://mvnrepository.com/artifact/org.apache.giraph/giraph-examples -->
9   <dependency>
10    <groupId>org.apache.giraph</groupId>
11    <artifactId>giraph-examples</artifactId>
12    <version>1.2.0</version>
13  </dependency>
14  <!-- https://mvnrepository.com/artifact/org.apache.hadoop/hadoop-core -->
15  <dependency>
16    <groupId>org.apache.hadoop</groupId>
17    <artifactId>hadoop-core</artifactId>
18    <version>1.2.1</version>
```

```

19     </dependency>
20     <!-- https://mvnrepository.com/artifact/org.apache.hadoop/hadoop-client -->
21     <dependency>
22         <groupId>org.apache.hadoop</groupId>
23         <artifactId>hadoop-client</artifactId>
24         <version>1.2.1</version>
25     </dependency>
26 </dependencies>

```

- 编写Giraph应用程序代码

- 新建 `src/main/java/example/MaxVertexValue.java` 类

```

1  package example;
2
3  import org.apache.giraph.graph.BasicComputation;
4  import org.apache.giraph.graph.Vertex;
5  import org.apache.hadoop.io.DoubleWritable;
6  import org.apache.hadoop.io.FloatWritable;
7  import org.apache.hadoop.io.LongWritable;
8
9  import java.io.IOException;
10
11  /**
12   * Vertex ID: LongWritable
13   * Vertex value: DoubleWritable
14   * Edge value: FloatWritable
15   * Message: DoubleWritable
16   * <p>
17   * Assumption:
18   * 1. The graph is strongly connected
19   */
20  public class MaxVertexValue extends BasicComputation<
21      LongWritable, DoubleWritable, FloatWritable, DoubleWritable> {
22      public void compute(Vertex<LongWritable, DoubleWritable,
23          FloatWritable> vertex, Iterable<DoubleWritable> messages) throws
24          IOException {
25          boolean changed = false;
26
27          for (DoubleWritable msg : messages) {
28              /* Collect messages from in-neighbours and update if necessary
29              */
30              if (vertex.getValue().get() < msg.get()) {
31                  vertex.setValue(new DoubleWritable(msg.get()));
32                  changed = true;
33              }
34          }
35          /* Send the message to out-neighbours at Superstep 0 or Vertex value
36          is changed */
37          if (getSuperstep() == 0 || changed) {
38              sendMessageToAllEdges(vertex, vertex.getValue());
39          }
40          vertex.voteToHalt();
41      }
42  }

```

- 新建 `src/main/java/GiraphDemoRunner.java` 类

```
1  import example.MaxVertexValue;
2  import org.apache.giraph.conf.GiraphConfiguration;
3  //import org.apache.giraph.examples.SimpleShortestPathsComputation;
4  import org.apache.giraph.io.formats.*;
5  import org.apache.giraph.job.GiraphJob;
6  import org.apache.hadoop.conf.Configuration;
7  import org.apache.hadoop.fs.Path;
8  import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
9  import org.apache.hadoop.util.Tool;
10 import org.apache.hadoop.util.ToolRunner;
11
12 public class GiraphDemoRunner implements Tool{
13
14     private Configuration conf;
15     public Configuration getConf() {
16         return conf;
17     }
18     public void setConf(Configuration conf) {
19         this.conf = conf;
20     }
21
22     public int run(String[] args) throws Exception {
23         /**
24          * 设置输入输出路径
25          * */
26         // String inputPath="src/main/resources/input/tiny_graph.txt";
27         // String
28         outputPath="src/main/resources/output/graph_shortestPaths";
29         String inputPath="src/main/resources/input/graph_data1.txt";
30         String outputPath="src/main/resources/output/graph_maxValue";
31
32         GiraphConfiguration giraphConf = new
33         GiraphConfiguration(getConf());
34
35         /**
36          * 配置具体用户自定义应用计算类
37          * */
38         //
39         giraphConf.setComputationClass(SimpleShortestPathsComputation.class);
40         giraphConf.setComputationClass(MaxVertexValue.class);
41
42         giraphConf.setVertexInputFormatClass(JsonLongDoubleFloatDoubleVertexInputFormat.class);
43         GiraphFileInputFormat.addVertexInputPath(giraphConf, new
44         Path(inputPath));
45
46         giraphConf.setVertexOutputFormatClass(IdWithValueTextOutputFormat.class);
47
48         giraphConf.setLocalTestMode(true);
49         giraphConf.setWorkerConfiguration(1, 1, 100);
50         giraphConf.SPLIT_MASTER_WORKER.set(giraphConf, false);
51         InMemoryVertexOutputFormat.initializeOutputGraph(giraphConf);
52         GiraphJob giraphJob = new GiraphJob(giraphConf, "GiraphDemo");
```

```

47     FileOutputStream.setOutputPath(giraphJob.getInternalJob(), new
Path(outputPath));
48     giraphJob.run(true);
49     return 0;
50 }
51
52 public static void main(String[] args) throws Exception{
53     ToolRunner.run(new GiraphDemoRunner(), args);
54 }
55 }

```

2. 调试Giraph程序

- 配置程序输入

在 `src/main/resources/input/` 路径下添加输入文件 `graph-data1.txt` 和 `tiny_graph.txt`：

```

Mon 23 Dec - 18:24 ~/Downloads/test_giraph/src/main/resources/input origi
n master 76* 6 14- 10+
@wushuangyoyo ls
graph-data1.txt tiny_graph.txt/main/resources/output/graph_maxValue 文件夹，文

```

- `graph-data1.txt` 文件内容：

```

Mon 23 Dec - 18:25 ~/Downloads/test_giraph/src/main/resources/input origi
n master 76* 6 14- 10+
@wushuangyoyo cat graph-data1.txt
[0,100,[[[1,1],[3,3]]]
[1,20,[[[0,1],[2,2],[3,1]]]
[2,90,[[[1,2],[4,4]]]
[3,50,[[[0,3],[1,1],[4,4]]]
[4,80,[[[3,4],[2,4]]]

```

- `tiny_graph.txt` 文件内容：

```

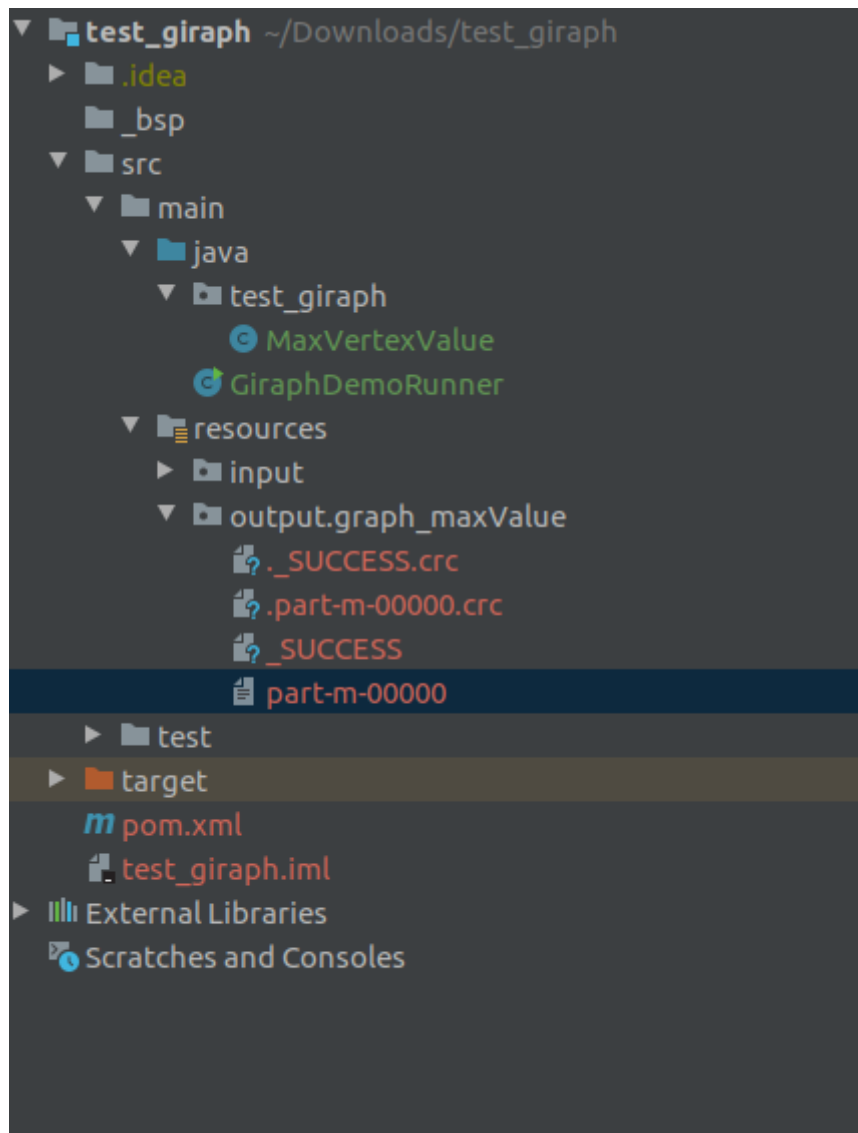
Mon 23 Dec - 18:24 ~/Downloads/test_giraph/src/main/resources/input origi
n master 76* 6 14- 10+
@wushuangyoyo cat tiny_graph.txt
[0,0,[[[1,1],[3,3]]]
[1,0,[[[0,1],[2,2],[3,1]]]
[2,0,[[[1,2],[4,4]]]
[3,0,[[[0,3],[1,1],[4,4]]]
[4,0,[[[3,4],[2,4]]]

```

- IDE中直接运行Giraph MaxVertexValue 应用程序

直接在 IDEA 中运行 `src/main/java/GiraphDemoRunner.java` 类，并查看输出结果。正常执行下，程序会产生 `src/main/resources/output/graph_maxValue` 文件夹，文件夹内包含程序输出内容。

- 正常执行情况下，项目结构：



- 程序输出内容：

```
om.xml x  MaxVertexValue.java x  GiraphDemoRunner.java x  part-m-00000 x
0    100.0
1    100.0
2    100.0
3    100.0
4    100.0
```

3. 运行Giraph程序

- 利用IDE打包jar文件

- 伪分布式模式下运行Giraph `MaxVertexValue` 程序

- 上传输入文件至 HDFS

```
Mon 23 Dec - 18:43 ~
@wushuangyoyo ./hadoop-1.2.1/bin/hadoop fs -ls input | grep graph
Warning: $HADOOP_HOME is deprecated.

-rw-r--r--  1 wushuangyoyo supergroup      118 2019-12-23 18:43 /user/wushuan
gyoyo/input/graph-data1.txt
-rw-r--r--  1 wushuangyoyo supergroup      112 2019-12-11 09:49 /user/wushuan
gyoyo/input/tiny_graph.txt
```

- 运行 giraph 程序
在终端中运行如下命令：


```

1 jps    #查看和确保 NameNode,DataNode,JobTracker以及TaskTracker服务正常启动
2 cd giraph-1.2.0-for-hadoop-1.2.1/
3 ./bin/giraph giraph.jar example.MaxVertexValue -vif
org.apache.giraph.io.formats.JsonLongDoubleFloatDoubleVertexInputFormat -
vip input/graph-data1.txt -vof
org.apache.giraph.io.formats.IdWithValueTextOutputFormat -op
output/maxVertexValue -w 3

```

结果如下：

```

X Mon 23 Dec - 18:46 ~/giraph-1.2.0-for-hadoop-1.2.1
@wushuangyoyo ./bin/giraph ./test_giraph.jar test_giraph.MaxVertexValue -vif org.apache.giraph.io.
formats.JsonLongDoubleFloatDoubleVertexInputFormat -vip input/graph-data1.txt -vof org.apache.giraph.i
o.formats.IdWithValueTextOutputFormat -op output/maxVertexValue -w 3
HADOOP_CONF_DIR=/home/wushuangyoyo/hadoop-1.2.1/conf
Warning: $HADOOP_HOME is deprecated.

19/12/23 18:47:19 INFO utils.ConfigurationUtils: No edge input format specified. Ensure your InputForm
at does not require one.
19/12/23 18:47:19 INFO utils.ConfigurationUtils: No edge output format specified. Ensure your OutputFo
rmat does not require one.
19/12/23 18:47:20 INFO job.GiraphJob: run: Since checkpointing is disabled (default), do not allow any
task retries (setting mapred.map.max.attempts = 1, old value = 4)
19/12/23 18:47:24 INFO job.GiraphJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=job_20
1912231842_0001
19/12/23 18:47:24 INFO job.GiraphJob: Waiting for resources... Job will start only when it gets all 4
mappers
19/12/23 18:47:45 INFO job.HaltApplicationUtils$DefaultHaltInstructionsWriter: writeHaltInstructions:
To halt after next superstep execute: ./bin/halt_application --zkServer localhost:22181 --zkNode /_hado
opBsp/job_201912231842_0001/_haltComputation'
19/12/23 18:47:45 INFO mapred.JobClient: Running job: job_201912231842_0001
19/12/23 18:47:46 INFO mapred.JobClient: map 100% reduce 0%
19/12/23 18:47:48 INFO mapred.JobClient: Job complete: job_201912231842_0001
19/12/23 18:47:48 INFO mapred.JobClient: Counters: 44
19/12/23 18:47:48 INFO mapred.JobClient: Map-Reduce Framework
19/12/23 18:47:48 INFO mapred.JobClient: Spilled Records: 8

```

- 查看输出结果

执行命令：

```

Mon 23 Dec - 18:47 ~/giraph-1.2.0-for-hadoop-1.2.1
@wushuangyoyo ~/hadoop-1.2.1/bin/hadoop fs -ls output/maxVertexValue
Warning: $HADOOP_HOME is deprecated.

Found 5 items
-rw-r--r- 1 wushuangyoyo supergroup 0 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertex
Value/_SUCCESS
drwxr-xr-x 1 wushuangyoyo supergroup 0 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertex
Value/_logs
-rw-r--r- 1 wushuangyoyo supergroup 16 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertex
Value/part-m-00001
-rw-r--r- 1 wushuangyoyo supergroup 16 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertex
Value/part-m-00002
-rw-r--r- 1 wushuangyoyo supergroup 8 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertex
Value/part-m-00003

Mon 23 Dec - 19:29 ~/giraph-1.2.0-for-hadoop-1.2.1
@wushuangyoyo ~/hadoop-1.2.1/bin/hadoop fs -cat output/maxVertexValue/p*
Warning: $HADOOP_HOME is deprecated.

0      100.0
3      100.0
1      100.0
4      100.0
2      100.0

```

- 分布式模式下运行Giraph MaxVertexValue 程序

和上一部分一致，无法完成。