

华东师范大学数据科学与工程学院实验报告

课程名称：分布式系统与编程

年级：2017 级

上机实践成绩：

指导教师：徐辰

姓名：吴双

上机实践名称：**Giraph 部署与编程**

学号：10164102141

上机实践日期：2019/12/20

上机实践编号：05

组号：01

上机实践时间：Week 15-16

一、实验目的

- 学习 Giraph 的部署，理解 Giraph 与 MapReduce 之间的关系
- 练习以顶点为中心的图算法编程方式，体会与基于 MapReduce/Spark/Flink 进行图算法编程的区别

二、实验任务

- Giraph 部署【第 15 周】：单机集中式、单机伪分布式（在个人用户下独立完成）、分布式（多位同学新建一个相同的用户，例如 ecnu，协作完成（实在无法完成））
- Giraph 编程【第 16 周】

三、使用环境

- Hadoop 2.9.2
- Ubuntu LTS 18.04
- Flink 1.7.2

四、实验过程

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 - $0 may be a softlink
sed -i 'i\export HADOOP_HOME=~/.hadoop-1.2.1' ~/.giraph-1.2.0-for-hadoop-1.2.1/bin/giraph-env
THIS="${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。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。查看 HDFS 的文件信息：

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

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
  -vlf org.apache.giraph.examples.SimpleShortestPathsComputation \ 1 sed -i '11\export HADOOP_HOME=~ /hadoop-1.2.1' ~/giraph-1.2.0-
  -vlf org.apache.giraph.to.formats.JsonLongDoubleFloatDoubleVertexInputFormat \ 1.2.1/bin/giraph-env
  -vlf input/tiny_graph.txt \
  -vof org.apache.giraph.to.formats.IdWithValueTextOutputFormat \ 修改 ~/hadoop-1.2.1/conf/mapred-site.xml, 在 <configuration> 下添加
  -op output/shortestpaths \ 3.5 查看 Giraph 应用程序运
  -w 3 行信息
Warning: SHADOOP_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: SHADOOP_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 /_hadoop8sp/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: SHADOOP_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 Hadoop Map/Reduce Administration

State: RUNNING
Started: Mon Dec 23 16:41:02 CST 2019
Version: 1.2.1-1503152
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 &id=1' will filter by 'smith' only in the user field and '1' only in all fields

Running 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

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_0002	Mon Dec 23 16:48:29 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

- 查看程序日志

- JobHistory 日志默认位置: ~/hadoop-1.2.1/logs

[job_201912111815_0002_conf.xml](#)
[job_201912231641_0001_conf.xml](#)
[job_201912231641_0002_conf.xml](#)
[userlogs/](#)

98508 bytes Dec 11, 2019 6:20:25 PM
98508 bytes Dec 23, 2019 4:46:30 PM
98508 bytes Dec 23, 2019 4:48:29 PM
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>

← → ↺ ↻

localhost:50030/logs/userlogs/job_201912231641_0001/

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 4:46:30 PM

[job-acls.xml](#) 507 bytes Dec 23, 2019 4:46:36 PM

3.6 关闭 Hadoop

略

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. 编写并调试 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:

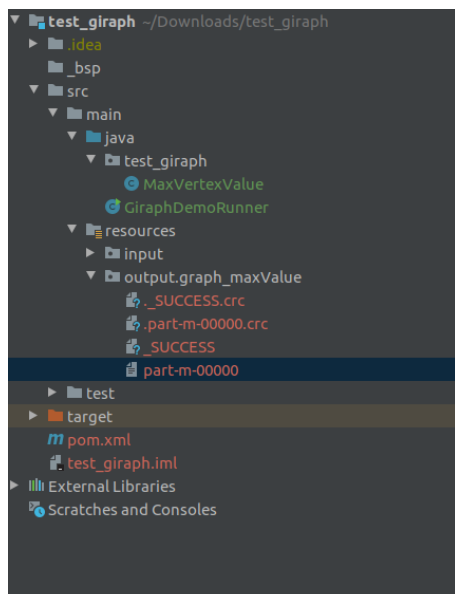
```
Mon 23 Dec - 18:25 ~/Downloads/test_giraph/src/main/resources/input origi
n @master 76*6 14-10+
@wushuangyoyo cat graph-data1.txt
[0,100,[[[1,1],[3,3]]]]
[1,20,[[[0,1],[2,2],[3,1]]]]
[2,90,[[[1,2],[4,4]]]]
[3,50,[[[0,3],[1,1],[4,4]]]]
[4,80,[[[3,4],[2,4]]]]
```

• `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

• 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
```

2. 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，具体的文件情况如下：

```
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/wushuangyoyo/input/graph-data1.txt
-rw-r--r-- 1 wushuangyoyo supergroup 112 2019-12-11 09:49 /user/wushuangyoyo/input/tiny_graph.txt
```

- 运行 giraph 程序
- 在终端中运行此 jar 包程序，结果如下：

```
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.io.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 InputFormat does not require one.
19/12/23 18:47:19 INFO utils.ConfigurationUtils: No edge output format specified. Ensure your OutputFormat 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_201912231842_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 /_hadoopBsp/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: Succeeded: 0
```

- 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/maxVertexValue/_SUCCESS
drwxr-xr-x 1 wushuangyoyo supergroup 0 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertexValue/_logs
-rw-r--r-- 1 wushuangyoyo supergroup 16 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertexValue/part-m-00001
-rw-r--r-- 1 wushuangyoyo supergroup 16 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertexValue/part-m-00002
-rw-r--r-- 1 wushuangyoyo supergroup 8 2019-12-23 18:47 /user/wushuangyoyo/output/maxVertexValue/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
```

分布式和上一部分一致，无法完成。

五、总结

1. Giraph 在分布式的支持上还是有一些不太会的地方，而且对外界的依赖比较高；
2. **论文要好好看，好多问题的出现其实就是没仔细看论文，没理解本质的含义造成的。