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

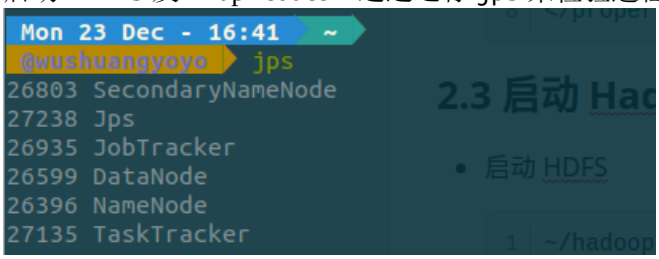
- 修改 ~/giraph-1.2.0-for-hadoop-1.2.1/bin/giraph-env，指定 Hadoop 安装路径：

```
# resolve links - $0 may be a softlink
sed -i '1i\export HADOOP_HOME=~/hadoop-1.2.1' ~/giraph-1.2.0-for-hadoop-1.2.1/bin/giraph-env
THIS="${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>
```

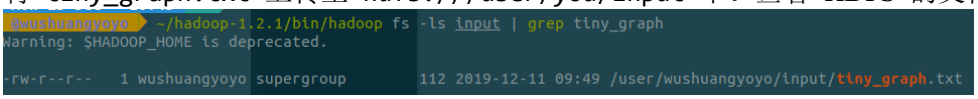
- 启动 HDFS 及 MapReduce。通过运行 `jps` 来检验进程状态：



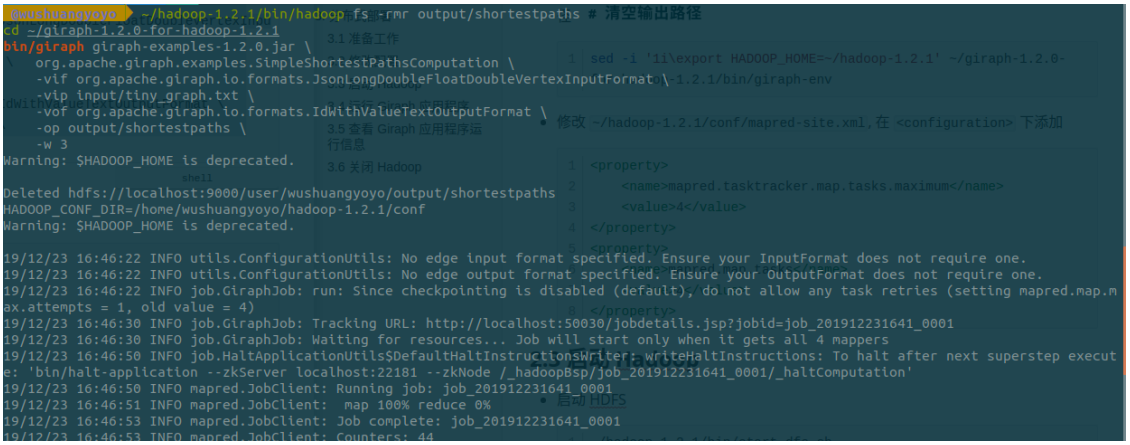
1.2 运行 Giraph 应用程序

Simple shortest paths computation 示例程序

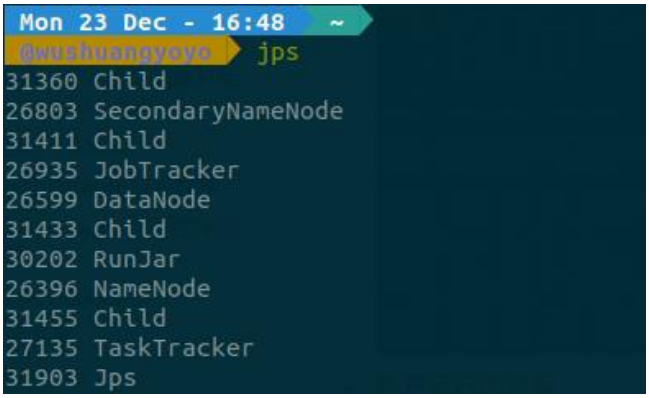
- 将 `tiny_graph.txt` 上传至 `hdfs:///user/you/input` 下。查看 HDFS 的文件信息：



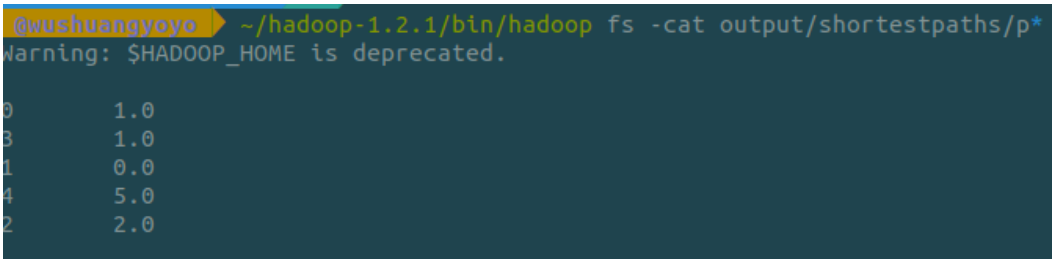
按照如下代码执行：



- 查看运行中进程

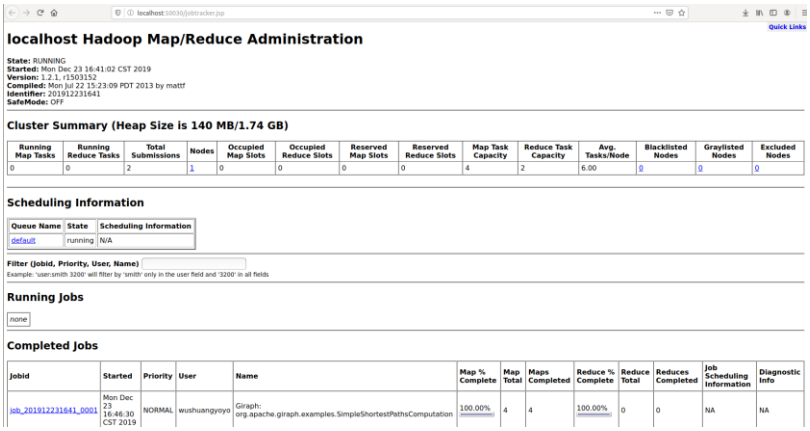


- 运行完成后查看输出



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, 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

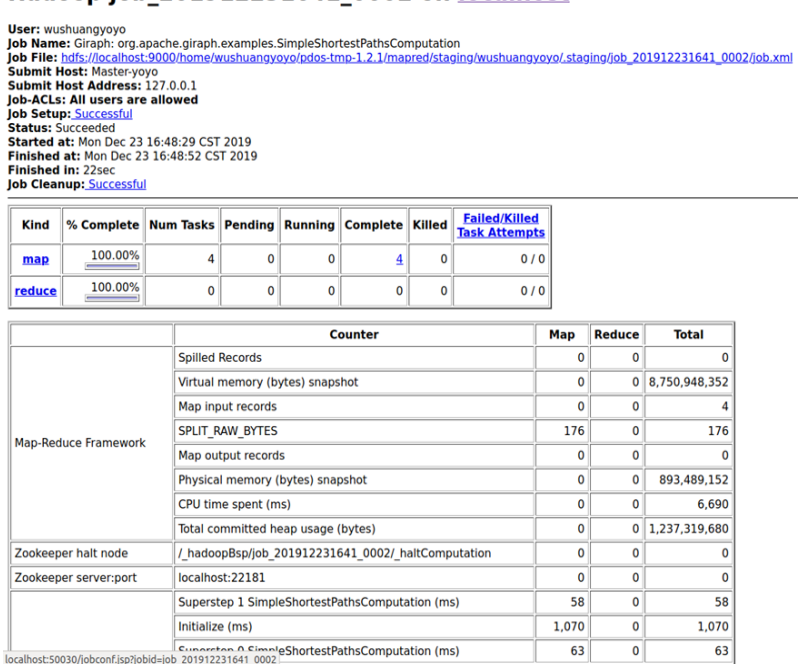
Running Jobs

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>

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 4:46:38 PM
job-acls.xml 507 bytes Dec 23, 2019 4:46:36 PM

3.6 关闭 Hadoop

略

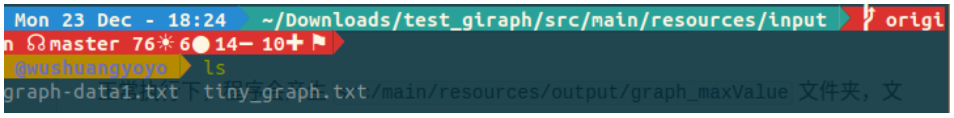
分布式在构建过程中有很多错误，实在无法进行，故放弃。

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：

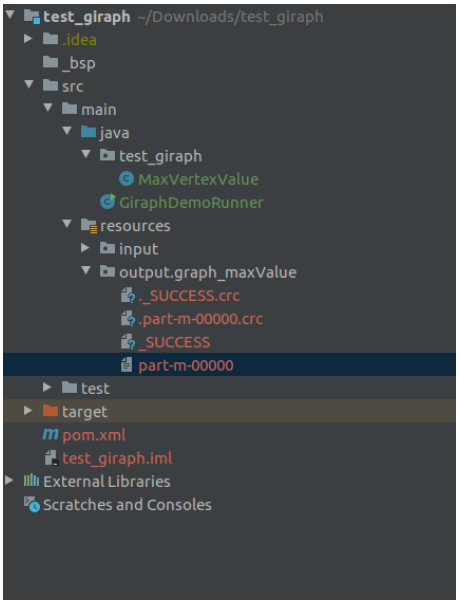


- graph-data1.txt 文件内容：

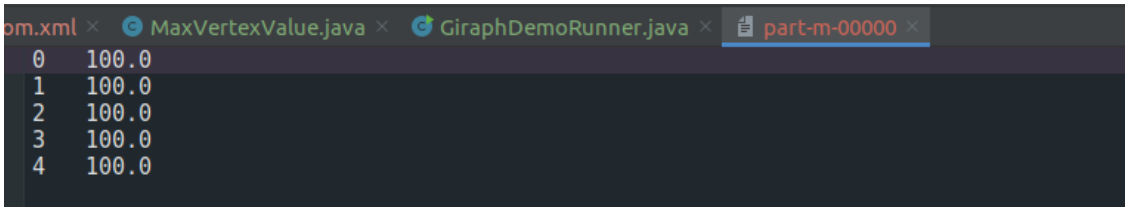
- tiny_graph.txt 文件内容：

- *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：

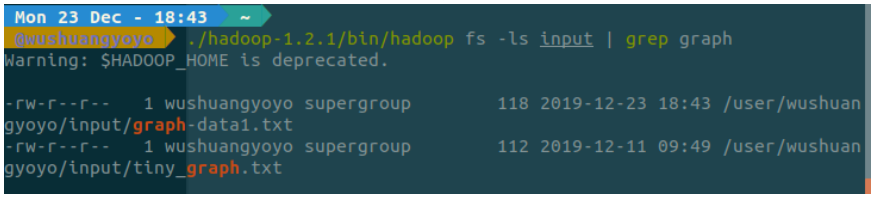


2. 运行 Giraph 程序

- *利用 IDE 打包 jar 文件*

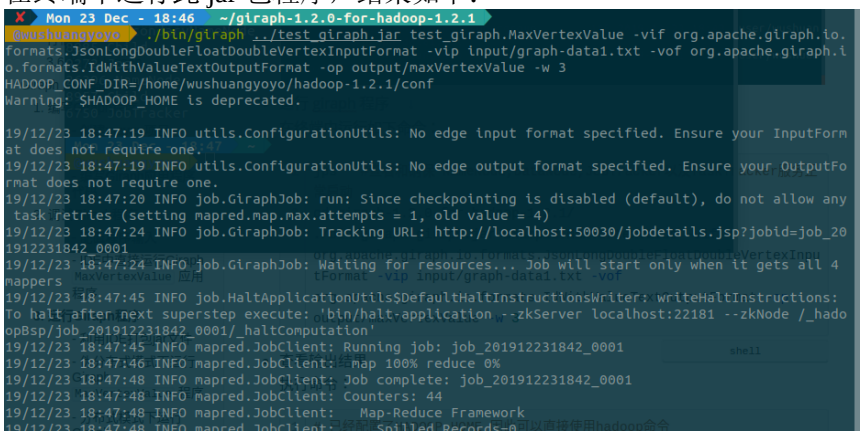
- *伪分布式模式下运行 Giraph MaxVertexValue 程序*

- 上传输入文件至 HDFS，具体的文件情况如下：



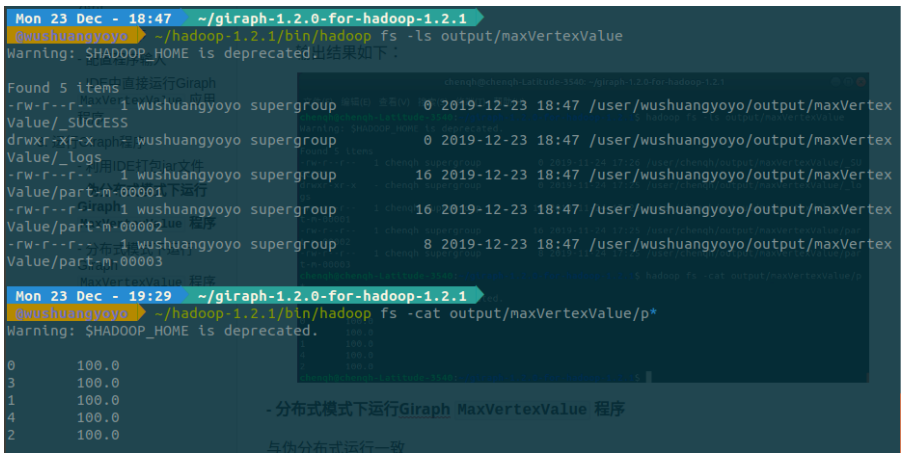
- 运行 giraph 程序

在终端中运行此 jar 包程序，结果如下：



- 查看输出结果

执行命令：



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。