Pig 介绍、安装与应用案例

该系列课程是应邀实验楼整理编写的，这里需要赞一下实验楼提供了学习的新方式，可以边看博客边上机实验，课程地址为 https://www.shiyanlou.com/courses/237

【注】该系列所使用到安装包、测试数据和代码均可在百度网盘下载，具体地址为 http://pan.baidu.com/s/10PnDs，下载该 PDF 文件

1 搭建环境

部署节点操作系统为 CentOS，防火墙和 SElinux 禁用，创建了一个 shiyanlou 用户并在系统根目录下创建/app 目录，用于存放 Hadoop 等组件运行包。因为该目录用于安装 hadoop 等组件程序，用户对 shiyanlou 必须赋予 rwx 权限（一般做法是 root 用户在根目录下创建/app 目录，并修改该目录拥有者为 shiyanlou(chown –R shiyanlou:shiyanlou /app)。

Hadoop 搭建环境：

- 虚拟机操作系统： CentOS6.6 64 位，单核，1G 内存
- JDK：1.7.0_55 64 位
- Hadoop：1.1.2

2 Pig 介绍

Pig 是 yahoo 捐献给 apache 的一个项目，使用 SQL-like 语言，是在 MapReduce 上构建的一种高级查询语言，把一些运算编译进 MapReduce 模型的 Map 和 Reduce 中。Pig 有两种运行模式： Local 模式和 MapReduce 模式

- 本地模式：Pig 运行于本地模式，只涉及到单独的一台计算机
- MapReduce 模式：Pig 运行于 MapReduce 模式，需要能访问一个 Hadoop 集群，并且需要装上 HDFS

Pig 的调用方式：

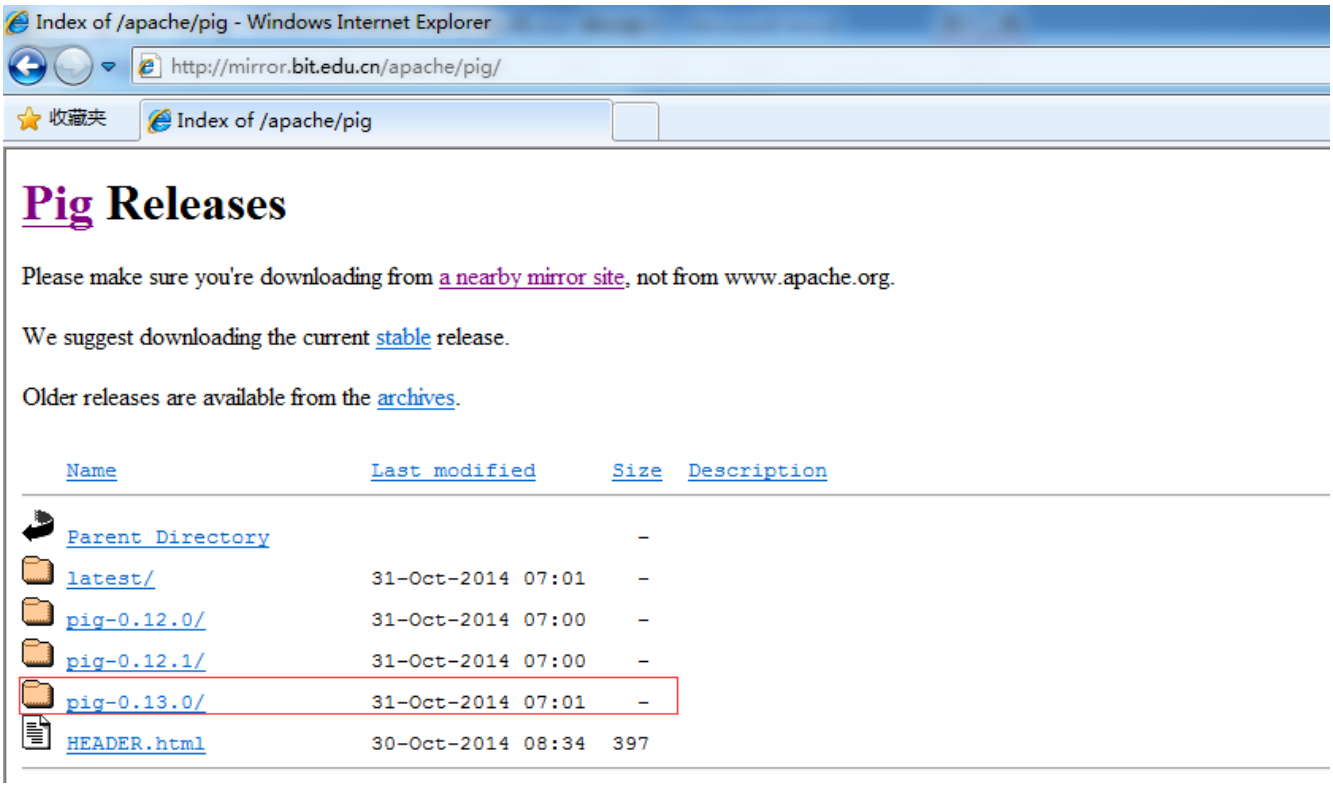
- Grunt shell 方式：通过交互的方式，输入命令执行任务；
- Pig script 方式：通过 script 脚本的方式来运行任务；

嵌入式方式：嵌入 java 源代码中，通过 java 调用来运行任务。

3 搭建 Pig 环境

3.1 下载并解压安装包

在 Apache 下载最新的 Pig 软件包，点击下载会推荐最快的镜像站点，以下为下载地址：http://mirror.bit.edu.cn/apache/pig/



也可以在/home/shiyanlou/install-pack 目录中找到该安装包，解压该安装包并把该安装包复制到/app 目录中

cd /home/shiyanlou/install-pack

tar -xzf pig-0.13.0.tar.gz

mv pig-0.13.0 /app

```
[shiyanlou@b393a04554e1 ~]$ cd /home/shiyanlou/install-pack
[shiyanlou@b393a04554e1 install-pack]$ tar -xzf pig-0.13.0.tar.gz
[shiyanlou@b393a04554e1 install-pack]$ ls
apache-maven-3.0.5-bin.tar.gz                hive-0.13.0-bin.tar.gz
chukwa-0.6.0.tar.gz                          jdk-7u55-linux-x64.tar.gz
class4                                       mahout-distribution-0.6.tar.gz
class5                                       MySQL-client-5.6.21-1.el6.x86_64.rpm
class6                                       mysql-connector-java-5.1.22-bin.jar
eclipse-jee-luna-SR1-linux-gtk-x86_64.tar.gz MySQL-devel-5.6.21-1.el6.x86_64.rpm
flume-1.5.2-bin.tar.gz                       MySQL-server-5.6.21-1.el6.x86_64.rpm
hadoop-1.1.2-bin.tar.gz                      pig-0.13.0
hadoop-2.2.0.tar.gz                          pig-0.13.0.tar.gz
hadoop-eclipse-plugin-1.1.2.jar              protobuf-2.5.0.tar.gz
hbase-0.96.2-hadoop1-bin.tar.gz              sqoop-1.4.5.bin__hadoop-1.0.0.tar.gz
[shiyanlou@b393a04554e1 install-pack]$ mv pig-0.13.0 /app
[shiyanlou@b393a04554e1 install-pack]$ ls /app
compile  hadoop-1.1.2  hadoop-2.2.0  lib  pig-0.13.0
[shiyanlou@b393a04554e1 install-pack]$
```

3.2 设置环境变量

使用如下命令编辑/etc/profile 文件：

sudo vi /etc/profile

```
hadoop1 | hadoop2 | hadoop3
[hadoop@hadoop1 /]$
[hadoop@hadoop1 /]$ sudo vi /etc/profile
```

设置 pig 的 class 路径和在 path 加入 pig 的路径，其中 PIG_CLASSPATH 参数是设置 pig 在 MapReduce 工作模式：

export PIG_HOME=/app/pig-0.13.0

export PIG_CLASSPATH=/app/hadoop-1.1.2/conf

export PATH=$PATH:$PIG_HOME/bin

```
export JAVA_HOME=/app/lib/jdk1.7.0_55
export PATH=$JAVA_HOME/bin:$PATH:/app/hadoop-1.1.2/bin
export CLASSPATH=.:$JAVA_HOME/lib/dt.jar:$JAVA_HOME/lib/tools.jar

export MAVEN_HOME=/app/lib/apache-maven-3.0.5
export PATH=$PATH:$MAVEN_HOME/bin

export PIG_HOME=/app/pig-0.13.0
export PIG_CLASSPATH=/app/hadoop-1.1.2/conf
export PATH=$PATH:$PIG_HOME/bin
```

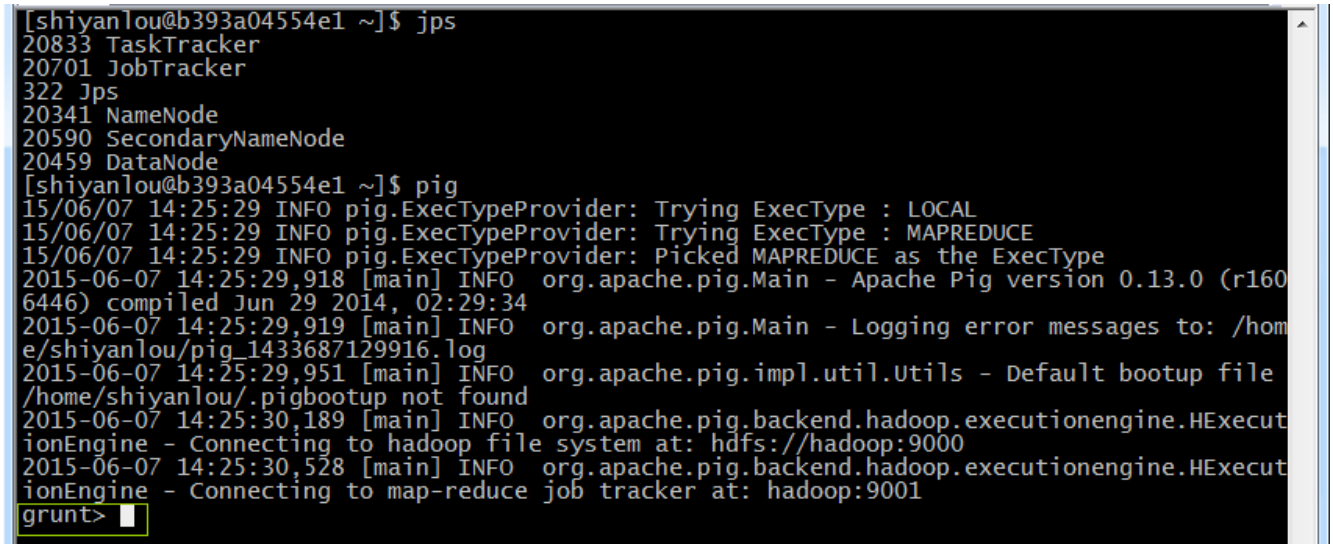
编译配置文件/etc/profile，并确认生效

source /etc/profile

echo $PATH

3.3 验证安装完成

重新登录终端，确保 hadoop 集群启动，键入 pig 命令，应该能看到 pig 连接到 hadoop 集群的信息并且进入了 grunt shell 命令行模式：



4 测试例子

4.1 测试例子内容

在/home/shiyanlou/install-pack/class7 中有 website_log.zip 测试数据文件，该文件是某网站访问日志，请大家使用 pig 计算出每个 ip 的点击次数，例如 123.24.56.57 13 24.53.23.123 7 34.56.78.120 20 .... 等等

4.2 程序代码

```
//加载 HDFS 中访问日志，使用空格进行分割，只加载 ip 列
records = LOAD 'hdfs://hadoop:9000/class7/input/website_log.txt' USING
PigStorage(' ') AS (ip:chararray);

// 按照 ip 进行分组，统计每个 ip 点击数
records_b = GROUP records BY ip;
records_c = FOREACH records_b GENERATE group,COUNT(records) AS click;

// 按照点击数排序，保留点击数前 10 个的 ip 数据
records_d = ORDER records_c by click DESC;
top10 = LIMIT records_d 10;

// 把生成的数据保存到 HDFS 的 class7 目录中
STORE top10 INTO 'hdfs://hadoop:9000/class7/out';
```

4.3 准备数据

可以在/home/shiyanlou/install-pack/class7 中找到本节使用的测试数据 website_log.zip 文

件，使用 unzip 文件解压缩，然后调用 hadoop 上传本地文件命令把该文件传到 HDFS 中的 /class7 目录，如下图所示：

cd /home/shiyanlou/install-pack/class7

unzip website_log.zip

ll

hadoop fs -mkdir /class7/input

hadoop fs -copyFromLocal website_log.txt /class7/input

hadoop fs -cat /class7/input/website_log.txt | less

```
[shiyanlou@b393a04554e1 ~]$ cd /home/shiyanlou/install-pack/class7
[shiyanlou@b393a04554e1 class7]$ unzip website_log.zip
Archive:  website_log.zip
  inflating: website_log.txt
[shiyanlou@b393a04554e1 class7]$ ll
total 7264
-rw-rw-r-- 1 shiyanlou shiyanlou 7118627 Feb  1  2012 website_log.txt
-rw-r--r-- 1 shiyanlou shiyanlou  316603 Jun  7 14:25 website_log.zip
[shiyanlou@b393a04554e1 class7]$
[shiyanlou@b393a04554e1 class7]$ hadoop fs -mkdir /class7/input
[shiyanlou@b393a04554e1 class7]$ hadoop fs -copyFromLocal website_log.txt /class7/input
[shiyanlou@b393a04554e1 class7]$ hadoop fs -cat /class7/input/website_log.txt | less
220.181.108.151 - - [31/Jan/2012:00:02:32 +0800] "GET /home.php?mod=space&uid=158&do=album
&view=me&from=space HTTP/1.1" 200 8784 "-" "Mozilla/5.0 (compatible; Baiduspider/2.0; +htt
p://www.baidu.com/search/spider.html)"
208.115.113.82 - - [31/Jan/2012:00:07:54 +0800] "GET /robots.txt HTTP/1.1" 200 582 "-" "Mo
zilla/5.0 (compatible; Ezooms/1.0; ezooms.bot@gmail.com)"
220.181.94.221 - - [31/Jan/2012:00:09:24 +0800] "GET /home.php?mod=spacecp&ac=pm&op=showms
g&handlekey=showmsg_3&touid=3&pmid=0&daterange=2&pid=398&tid=66 HTTP/1.1" 200 10070 "-" "S
ogou web spider/4.0(+http://www.sogou.com/docs/help/webmasters.htm#07)"
```

4.4 实现过程

4.4.1 输入代码

进入 pig shell 命令行模式：

```
[shiyanlou@b393a04554e1 ~]$ pig
15/06/07 14:43:00 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
15/06/07 14:43:00 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
15/06/07 14:43:00 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2015-06-07 14:43:00,363 [main] INFO  org.apache.pig.Main - Apache Pig version 0.13.0 (r160
6446) compiled Jun 29 2014, 02:29:34
2015-06-07 14:43:00,363 [main] INFO  org.apache.pig.Main - Logging error messages to: /hom
e/shiyanlou/pig_1433688180360.log
2015-06-07 14:43:00,395 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file
/home/shiyanlou/.pigbootup not found
2015-06-07 14:43:00,646 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecut
ionEngine - Connecting to hadoop file system at: hdfs://hadoop:9000
2015-06-07 14:43:00,994 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecut
ionEngine - Connecting to map-reduce job tracker at: hadoop:9001
grunt>
```

输入代码：

```
grunt> records = LOAD 'hdfs://hadoop:9000/class7/input/website_log.txt' USING PigStorage('
') AS (ip:chararray);
grunt> records_b = GROUP records BY ip;
grunt> records_c = FOREACH records_b GENERATE group,COUNT(records) AS click;
grunt> records_d = ORDER records_c by click DESC;
grunt> top10 = LIMIT records_d 10;
grunt> STORE top10 INTO 'hdfs://hadoop:9000/class7/out';
2015-06-07 14:48:43,647 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig featu
res used in the script: GROUP_BY,ORDER_BY,LIMIT
2015-06-07 14:48:43,704 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanO
ptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, GroupByConstParallelSetter, Limi
tOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, Pus
hDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[F
ilterLogicExpressionSimplifier]}
2015-06-07 14:48:43,875 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapRedu
ceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2015-06-07 14:48:43,923 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapRedu
ceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner
```

4.4.2 运行过程

在执行过程中在 JobTracker 页面观察运行情况，链接地址为：http://**.***.**.***:50030/jobtracker.jsp

Running Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total
job_201411201452_0014	Thu Nov 20 15:58:16 CST 2014	NORMAL	hadoop	PigLatin:DefaultJobName	0.00%	1	0	0.00%	1

点击查看具体作业信息

Hadoop job_201411201452_0014 on hadoop1

User: hadoop
Job Name: PigLatin:DefaultJobName
Job File: hdfs://hadoop1:9000/usr/local/hadoop-1.1.2/tmp/mapred/staging/hadoop/.staging/job_201411201452_0014/job.xml
Submit Host: hadoop1
Submit Host Address: 10.88.147.221
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Running
Started at: Thu Nov 20 15:58:16 CST 2014
Running for: 11sec
Job Cleanup: Pending

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	1	0	0	1	0	0 / 0
reduce	0.00%	1	0	1	0	0	0 / 0

可以观察到本次任务分为 4 个作业，每个作业一次在上一次作业的结果上进行计算

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
1.1.2   0.13.0  shiyanlou       2015-06-07 14:52:32     2015-06-07 14:54:21     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducet
ime     Alias   Feature Outputs
job_201506040132_0019   1       1       5       5       5       5       9       9       9       9       records,records_b,records_c     GROUP_BY,COMB
INER
job_201506040132_0020   1       1       2       2       2       2       9       9       9       9       records_d       SAMPLER
job_201506040132_0021   1       1       3       3       3       3       9       9       9       9       records_d       ORDER_BY,COMBINER
job_201506040132_0022   1       1       2       2       2       2       9       9       9       9       records_d               hdfs://hadoop:9000/cl
ass7/out,

Input(s):
Successfully read 28134 records (7118993 bytes) from: "hdfs://hadoop:9000/class7/input/website_log.txt"

Output(s):
Successfully stored 10 records (191 bytes) in: "hdfs://hadoop:9000/class7/out"
```

```
Counters:
Total records written : 10
Total bytes written : 191
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201506040132_0019   ->      job_201506040132_0020,
job_201506040132_0020   ->      job_201506040132_0021,
job_201506040132_0021   ->      job_201506040132_0022,
job_201506040132_0022
```

4.4.3 运行结果

通过以下命令查看最后的结果：

hadoop fs -ls /class7/out

hadoop fs -cat /class7/out/part-r-00000

```
[shiyanlou@b393a04554e1 ~]$ hadoop fs -ls /class7/out
Found 3 items
-rw-r--r--   1 shiyanlou supergroup          0 2015-06-07 14:54 /class7/out/_SUCCESS
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-07 14:54 /class7/out/_logs
-rw-r--r--   1 shiyanlou supergroup        191 2015-06-07 14:54 /class7/out/part-r-00000
[shiyanlou@b393a04554e1 ~]$
[shiyanlou@b393a04554e1 ~]$ hadoop fs -cat /class7/out/part-r-00000
218.20.24.203   4597
221.194.180.166 4576
119.146.220.12  1850
117.136.31.144  1647
121.28.95.48    1597
113.109.183.126 1596
182.48.112.2    870
120.84.24.200   773
61.144.125.162  750
27.115.124.75   470
[shiyanlou@b393a04554e1 ~]$
```

