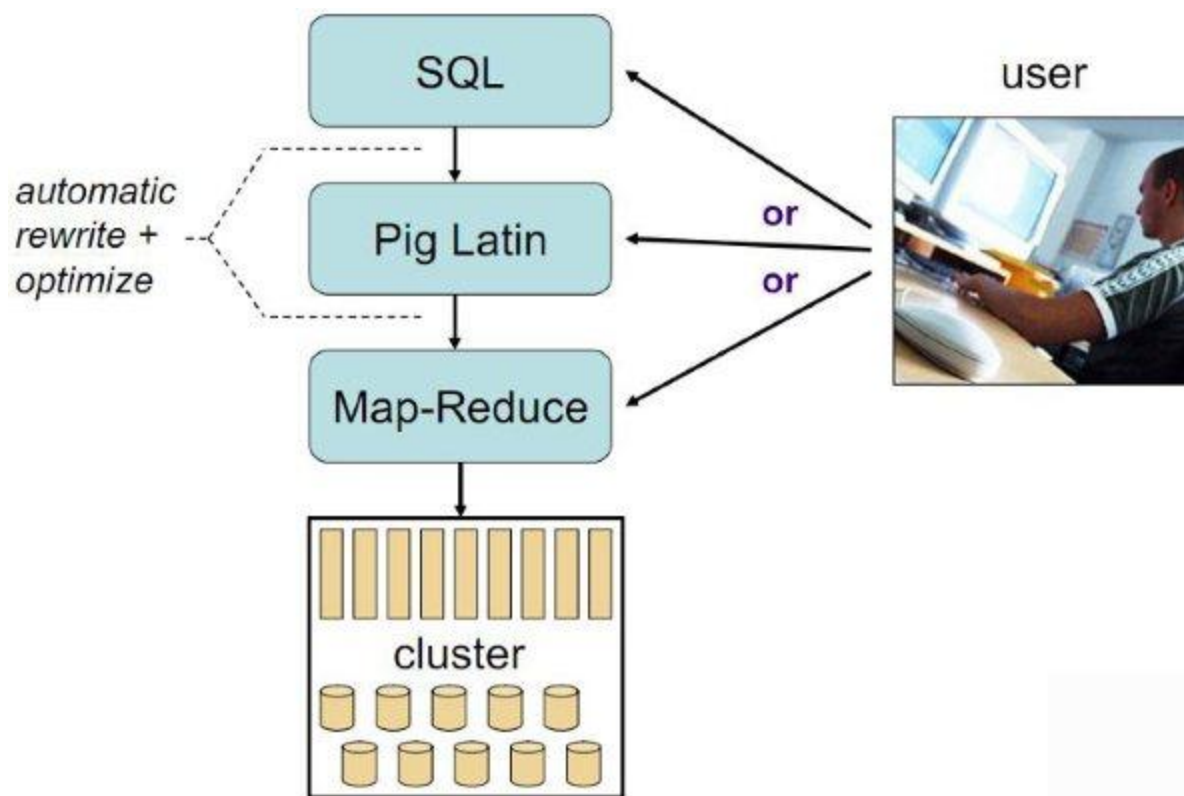




Hadoop数据分析平台 第8周

2012.10.23

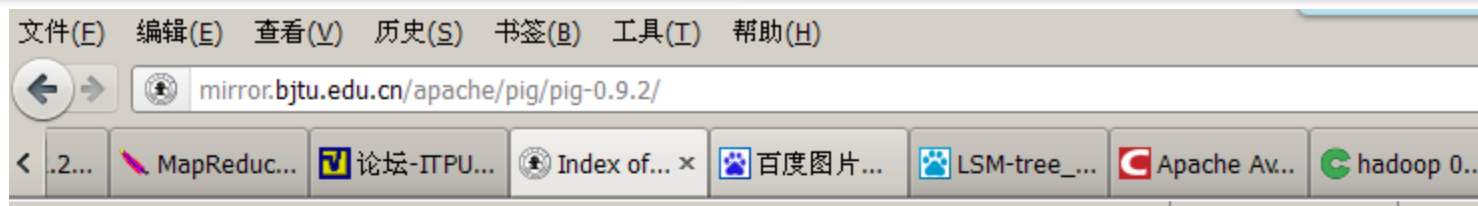
- Pig可以看做hadoop的客户端软件，可以连接到hadoop集群进行数据分析工作
- Pig方便不熟悉java的用户，使用一种较为简便的类似于SQL的面向数据流的语言pig latin进行数据处理
- Pig latin可以进行排序、过滤、求和、分组、关联等常用操作，还可以自定义函数，这是一种面向数据分析处理的轻量级脚本语言
- Pig可以看做是pig latin到map-reduce的映射器



安装pig

- 下载并解压pig安装包 (<http://pig.apache.org/>)
- 设置环境变量
- 进入grunt shell验证

下载并解压pig安装包



Index of /apache/pig/pig-0.9.2/

../		
RELEASE NOTES.txt	18-Jan-2012 16:14	2224
pig-0.9.2-1.i386.rpm	18-Jan-2012 16:13	33419952
pig-0.9.2-1.i386.rpm.asc	18-Jan-2012 16:13	189
pig-0.9.2.tar.gz	18-Jan-2012 16:14	47875717
pig-0.9.2.tar.gz.asc	18-Jan-2012 16:14	189
pig_0.9.2-1_i386.deb	18-Jan-2012 16:13	33157082
pig_0.9.2-1_i386.deb.asc	18-Jan-2012 16:13	189

```
checkpoint data hadoop-0.20.2 hadoop-0.20.2.tar.gz input name pig-0.9.2.tar.gz  
[grid@h1 ~]$ tar xzvf pig-0.9.2.tar.gz
```

2012.10.23

```
# .bash_profile

# Get the aliases and functions
if [ -f ~/.bashrc ]; then
    . ~/.bashrc
fi

# User specific environment and startup programs

PATH=$PATH:/home/grid/pig-0.9.2/bin:$HOME/bin
JAVA_HOME=/usr

export JAVA_HOME
export PATH
```

重新登录使环境变量生效

■ 用set命令检查环境变量

```
IFS=$'\n'
INPUTRC=/etc/inputrc
JAVA_HOME=/usr
LANG=en_US.UTF-8
LESSOPEN='|/usr/bin/lesspipe.sh %s'
LINES=32
LOGNAME=grid
LS_COLORS='no=00:fi=00:di=01:34:ln=01:36:pi=40:33:so=01:35:bd=40:33:01:cd=40:33:01:or=01:05:37:41:mi=01:05:37:41:ex=01:32:*.c
md=01:32:*.exe=01:32:*.com=01:32:*.btm=01:32:*.bat=01:32:*.sh=01:32:*.csh=01:32:*.tar=01:31:*.tgz=01:31:*.arj=01:31:*.taz=01:
31:*.lzh=01:31:*.zip=01:31:*.z=01:31:*.Z=01:31:*.gz=01:31:*.bz2=01:31:*.bz=01:31:*.tz=01:31:*.rpm=01:31:*.cpio=01:31:*.jpg=01:
:35:*.gif=01:35:*.bmp=01:35:*.xbm=01:35:*.xpm=01:35:*.png=01:35:*.tif=01:35:'
MACHTYPE=i686-redhat-linux-gnu
MAIL=/var/spool/mail/grid
MAILCHECK=60
OPTERR=1
OPTIND=1
OSTYPE=linux-gnu
PATH=/usr/kerberos/bin:/usr/java/jdk1.6.0_26/bin:/usr/java/jdk1.6.0_26/jre/bin:/usr/local/bin:/bin:/usr/bin:/home/grid/pig-0.
9.2/bin:/home/grid/bin
PIPESTATUS=([0]="0")
PPID=4435
```

2012.10.23

进入grunt shell

```
[grid@h1 ~]$ pig -x local
2012-06-24 11:43:07.760 [main] INFO org.apache.pig.Main - Apache Pig version 0.10.0 (r1328203) compiled Apr 19 2012, 22:54:
2
2012-06-24 11:43:07.761 [main] INFO org.apache.pig.Main - Logging error messages to: /home/grid/pig_1340552587755.log
2012-06-24 11:43:07.992 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop f
le system at: file:///
grunt>
```


Pig工作模式

- 本地模式：所有文件和执行过程都在本地，一般用于测试程序
- Mapreduce模式：实际工作模式

配置pig的map-reduce模式

- 设置PATH，增加指向hadoop/bin
- 设置PIG_CLASSPATH环境变量
- 修改hosts文件
- 启动pig

设置PIG_CLASSPATH环境变量

- 设置完成后重新登录使环境变量生效

```
# .bash_profile

# Get the aliases and functions
if [ -f ~/.bashrc ]; then
    . ~/.bashrc
fi

# User specific environment and startup programs

PATH=$PATH:/home/grid/hadoop-0.20.2/bin:/home/grid/pig-0.10.0/bin:$HOME/bin
JAVA_HOME=/usr
PIG_CLASSPATH=/home/grid/hadoop-0.20.2/conf/

export PIG_CLASSPATH
export JAVA_HOME
export PATH
~
~
```

修改hosts文件

```
[root@h1 grid]# vi /etc/hosts

# Do not remove the following line, or various programs
# that require network functionality will fail.
127.0.0.1          localhost.localdomain localhost
::1               localhost6.localdomain6 localhost6
192.168.1.102      h1-master
192.168.1.102      h1
192.168.1.103      h2
192.168.1.104      h3
192.168.1.163      dog
192.168.1.162      cat
192.168.1.161      gangster
~
```

启动grunt shell

```
Last login: Sun Jun 24 11:38:53 2012 from 192.168.1.100
```

```
[grid@h1 ~]$
```

```
[grid@h1 ~]$ pig
```

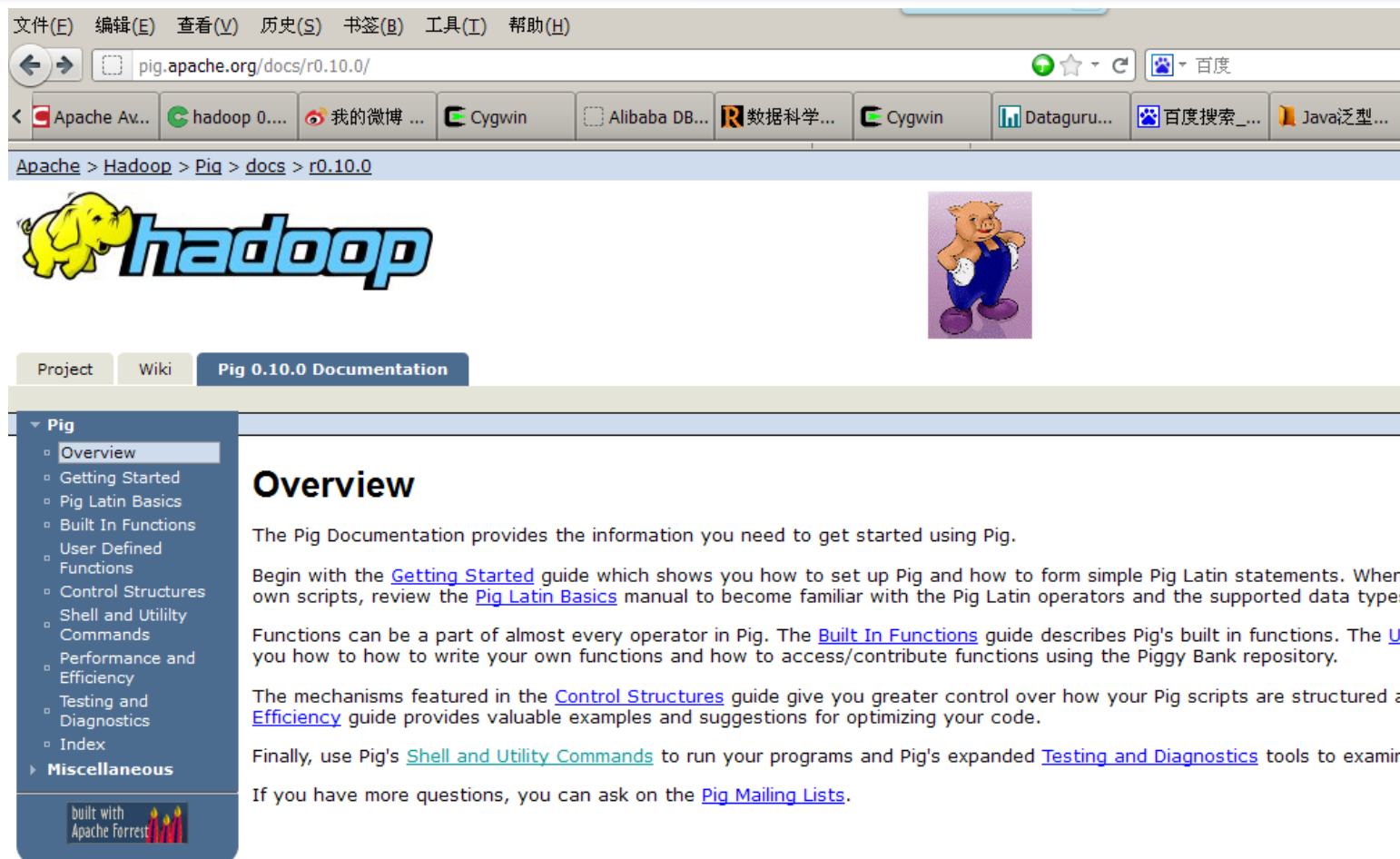
```
2012-06-24 11:40:31,782 [main] INFO org.apache.pig.Main - Apache Pig version 0.10.0 (r1328202)
```

```
2012-06-24 11:40:31,783 [main] INFO org.apache.pig.Main - Logging error messages to: /home/g
```

```
2012-06-24 11:40:32,633 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecution  
le system at: hdfs://h1:9000
```

```
2012-06-24 11:40:32,858 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecution  
e job tracker at: h1:9001
```

```
grunt>
```





The screenshot shows a web browser window displaying the Apache Pig 0.10.0 documentation. The browser's address bar shows the URL `pig.apache.org/docs/r0.10.0/`. The page features the Hadoop logo on the left and a cartoon pig character on the right. The navigation menu on the left includes links to Project, Wiki, and Pig 0.10.0 Documentation. The main content area is titled "Overview" and provides information about getting started with Pig, including links to the Getting Started guide, Pig Latin Basics manual, Built In Functions guide, User Defined Functions, Control Structures, Shell and Utility Commands, Performance and Efficiency, Testing and Diagnostics, and Index. A small "built with Apache Forrest" logo is visible in the bottom left corner of the page.

文件(E) 编辑(E) 查看(V) 历史(S) 书签(B) 工具(T) 帮助(H)

← → + ☆ ↻ 百度

< Apache Av... hadoop 0... 我的微博 ... Cygwin Alibaba DB... 数据科学... Cygwin Dataguru... 百度搜索... Java泛型...

Apache > Hadoop > Pig > docs > r0.10.0

Project Wiki **Pig 0.10.0 Documentation**

▼ Pig

- Overview
- Getting Started
- Pig Latin Basics
- Built In Functions
- User Defined Functions
- Control Structures
- Shell and Utility Commands
- Performance and Efficiency
- Testing and Diagnostics
- Index

► Miscellaneous

built with Apache Forrest

Overview

The Pig Documentation provides the information you need to get started using Pig.

Begin with the [Getting Started](#) guide which shows you how to set up Pig and how to form simple Pig Latin statements. When you have written your own scripts, review the [Pig Latin Basics](#) manual to become familiar with the Pig Latin operators and the supported data types.

Functions can be a part of almost every operator in Pig. The [Built In Functions](#) guide describes Pig's built in functions. The [User Defined Functions](#) guide describes how to write your own functions and how to access/contribute functions using the Piggy Bank repository.

The mechanisms featured in the [Control Structures](#) guide give you greater control over how your Pig scripts are structured and the [Performance and Efficiency](#) guide provides valuable examples and suggestions for optimizing your code.

Finally, use Pig's [Shell and Utility Commands](#) to run your programs and Pig's expanded [Testing and Diagnostics](#) tools to examine the execution of your scripts.

If you have more questions, you can ask on the [Pig Mailing Lists](#).

2012.10.23

Pig的运行方法

- 脚本
- Grunt
- 嵌入式

- 自动补全机制
- Autocomplete文件
- Eclipse插件PigPen

Grunt shell命令

```
"cat" ...  
"fs" ...  
"sh" ...  
"cd" ...  
"cp" ...  
"copyFromLocal" ..  
"copyToLocal" ...  
"dump" ...  
"describe" ...  
"aliases" ...  
"explain" ...  
"help" ...  
"kill" ...  
"ls" ...  
"mv" ...  
"mkdir" ...  
"pwd" ...  
"quit" ...  
"register" ...  
"rm" ...  
"rmf" ...  
"set" ...  
"illustrate" ...  
"run" ...  
"exec" ...  
"scriptDone" ...
```

2012.10.23

ls、cd、cat

```
grunt>
grunt> ls
hdfs://h1:9000/user/grid/in      <dir>
hdfs://h1:9000/user/grid/in1    <dir>
hdfs://h1:9000/user/grid/out    <dir>
hdfs://h1:9000/user/grid/out1   <dir>
grunt> cd in
grunt> ls
hdfs://h1:9000/user/grid/in/test1.txt<r 3>      12
hdfs://h1:9000/user/grid/in/test2.txt<r 3>      13
grunt> cat test1.txt
hello world
grunt>
```

copyToLocal

```
grunt>  
grunt> ls  
hdfs://h1:9000/user/grid/in/test1.txt<r 3>  
hdfs://h1:9000/user/grid/in/test2.txt<r 3>  
grunt> copyToLocal test1.txt ttt  
grunt> █
```

```
[grid@h1 ~]$  
[grid@h1 ~]$ ls -l ttt  
-rwxrwxrwx 1 grid grid 12 Jun 24 12:14 ttt  
[grid@h1 ~]$  
[grid@h1 ~]$
```

执行操作系统命令：sh

```
grunt>  
grunt>  
grunt> sh /usr/java/jdk1.6.0_26/bin/jps  
5022 RunJar  
4152 SecondaryNameNode  
5745 Jps  
4221 JobTracker  
4027 NameNode  
grunt> █
```

- Bag : 表
- Tuple : 行, 记录
- Field : 属性
- Pig不要求同一个bag里面的各个tuple有相同数量或相同类型的field

Pig latin常用语句

- LOAD : 指出载入数据的方法
- FOREACH : 逐行扫描进行某种处理
- FILTER : 过滤行
- DUMP : 把结果显示到屏幕
- STORE : 把结果保存到文件

LOAD、FOREACH、STORE三部曲

```
[grid@h1 ~]$ pig -x local
2012-06-24 15:33:18,714 [main] INFO org.apache.pig.Main - Apache Pig version 0.10.0
2
2012-06-24 15:33:18,715 [main] INFO org.apache.pig.Main - Logging error messages to
2012-06-24 15:33:18,947 [main] INFO org.apache.pig.backend.hadoop.executionengine.L
le system at: file:///
grunt> A =LOAD '/home/grid/csdn.txt'
>> USING PigStorage('#')
>> AS (id,pw,em);
2012-06-24 15:33:28,287 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Ini
Tracker, sessionId=
grunt> B =FOREACH A
>> GENERATE em;
2012-06-24 15:34:27,004 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Can
e=JobTracker, sessionId= - already initialized
2012-06-24 15:34:27,060 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Can
e=JobTracker, sessionId= - already initialized
grunt> STORE B INTO '/home/grid/email.txt'
>> USING PigStorage();
2012-06-24 15:36:01,661 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Can
e=JobTracker, sessionId= - already initialized
2012-06-24 15:36:01,682 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Can
e=JobTracker, sessionId= - already initialized
2012-06-24 15:36:01,694 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig
2012-06-24 15:36:01,698 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Can
e=JobTracker, sessionId= - already initialized
```

2012.10.23

```
[grid@h1 ~]$ ls
checkpoint  email.txt          input              pig-0.10.0.tar.gz  pig_1340551252887.log  pig_134056596739
csdn.txt    hadoop-0.20.2      name              pig-0.9.2          pig_1340552343347.log  ttt
data        hadoop-0.20.2.tar.gz pig-0.10.0        pig-0.9.2.tar.gz   pig_1340552631218.log

[grid@h1 ~]$ cd email.txt
[grid@h1 email.txt]$ ls
part-m-00000  part-m-00002  part-m-00004  part-m-00006  part-m-00008
part-m-00001  part-m-00003  part-m-00005  part-m-00007  _temporary
[grid@h1 email.txt]$ ls -l
total 123948
-rwxrwxrwx 1 grid grid 14935615 Jun 24 15:36 part-m-00000
-rwxrwxrwx 1 grid grid 14954292 Jun 24 15:36 part-m-00001
-rwxrwxrwx 1 grid grid 14831079 Jun 24 15:36 part-m-00002
-rwxrwxrwx 1 grid grid 14802578 Jun 24 15:36 part-m-00003
-rwxrwxrwx 1 grid grid 14600189 Jun 24 15:36 part-m-00004
-rwxrwxrwx 1 grid grid 14591448 Jun 24 15:36 part-m-00005
-rwxrwxrwx 1 grid grid 14573905 Jun 24 15:36 part-m-00006
-rwxrwxrwx 1 grid grid 14750540 Jun 24 15:36 part-m-00007
-rwxrwxrwx 1 grid grid 86822256 Jun 24 15:36 part-m-00008
drwxrwxr-x 2 grid grid 4096 Jun 24 15:36 _temporary
[grid@h1 email.txt]$ head part-m-00000
zdg@csdn.net
chengming_zheng@163.com
fstao@tom.com
hujiye@263.net
ccedcjl@21cn.com
songmail@21cn.com
appollp@netease.com
junlu@peoplemail.com.cn
```


- 支持使用Java、Python、Javascript三种语言编写UDF
- Java自定义函数较为成熟，其它两种功能还有限



Thanks

FAQ时间