第11章 关于pig

很多时候数据处理需要多个MapReduce过程才能实现，而使数据处理过程与该模式匹配可能很困难。有了pig就能使用更为丰富的数据结构了。

Pig包括两个部分：

1. 用于描述数据流的语言 Pig Latin
2. 用于运行Pig latin 程序的执行环境，当前有两个环境：单JVM中的本地执行环境和Hadoop集群上的分布式执行环境。

Pig是一种探索大规模数据集的脚本语言..MapReduce的一个缺点就是开发周期太长。

Pig的诱人之处在于仅用控制台上的五六行Pig Latin 代码就能出来TB级的数据。

Pig的一个很有用的特性：它支持在输入数据的一个有代表性的子集上运行。

Pig被设计为可扩展的。

Pig并不适合所有饿数据处理任务，和MapReduce一样，它是为数据批处理而设计的。

除非你愿意花大量的时间优化java MapReduce ,否则使用pig Latin来写查询的确能够帮助你解决时间。

本章只介绍pig的基础内容，要获得更详细的指南，可以参考AlanGates p《pig程序设计》

## 11.1 安装与运行Pig

下载一个稳定版本的pig，将其解压到工作台的合适路径下。再将pig的二进制文件路径加入到命令行路径也很方便。

### 11.1.1 执行类型

Pig有两种执行类型：本地模式，MapReduce集群模式。

1. **本地模式**

该模式只适合于处理小规模数据集和试用pig时。

执行类型可用 -x或-exectype 选项进行设置，如果要使用本地模式运行，应把该选项设置为local：

% pig -x local

grunt>

grunt是与pig进行交互的外壳程序（shell）

1. **MapReduce模式**

如果要使用MapReduce模式，你首先需要检查下载的Pig版本是否与正在使用Hadoop版本兼容。

设置好pig和Hadoop集群的连接。 -x mapreduce 或者忽略该选项，因为MapReduce模式是pig默认执行模式

### 11.1.2 运行pig程序

有三种执行pig程序的方法，他们在本地和MapReduce模式下都适用。

1. 脚本

% pig script.pig

1. Grunt

在shell中可以使用run和exec运行脚本

1. 嵌入式方法

在java中使用PigServer类运行Pig程序，就像在java中使用JDBC运行SQL程序。

### 11.1.3 Grunt

可以创建一个autocomplete的文件把他放到Pig的类路径或者启动Grunt的目录中，以此定制自动补全的单词。文件中列出常用的文件路径也特别有用。

使用help命令列出命令列表

使用quit结束grunt会话

### 11.1.4 Pig Latin 编辑器

Pigpen是一个提供pig程序开发环境的Eclipse插件。它包含Pig脚本编辑器，示例生成器以用于hadoop集群上运行脚本的按钮。它还提供了一个操作窗口，能够以图形的方式显示脚本，将数据流可视化。

## 11.2 示例