Hadoop学习之Hive编程指南学习(一)

1. Hive概述

Hive最适合数据仓库应用程序，使用该程序进行相关的静态数据分析，不需要快速响应给出结果，而且数据本身不会频繁变化。

* 1. 缺点：Hive不是完整的DB，受hadoop和hdfs影响，Hive不支持记录级别的增删改；查询延时比较严重(hadoop为批处理的)，即使数据集较小；hive不支持事务，不支持OTLP(联机事务处理)所需的关键功能，接近OLAP(联机分析技术)工具。
  2. 优点：用户可以通过查询生成新表或将查询结果导入文件；最适合数据仓库，可维护海量数据；可对数据进行挖掘，形成意见和报告；hive

使用sql语法，降低程序一直到hadoop的难度，便于用户使用；hiveQL和mysql的sql相近，便于mysql的DBA使用。

1. Hive 安装

因在伪分布模式和集群模式安装文件中有详述，此处不再重复。

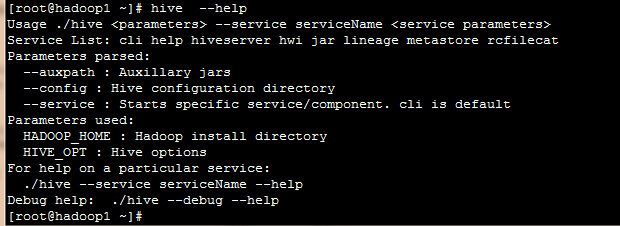
PS：如果频繁使用到hadoop dfs命令时，最好定义个简短的别名(如：alias hf=”hadoop fs”)!

PS：当处理小数据集时，使用本地模式执行可以使得hive执行的更快些，设置如下属性：set hive.exec.mode.local.auto=true;时，将会触发Hive主动使用本地模式。如果想默认使用这配置，可以将命令加到$HOME/.hiverc文件中

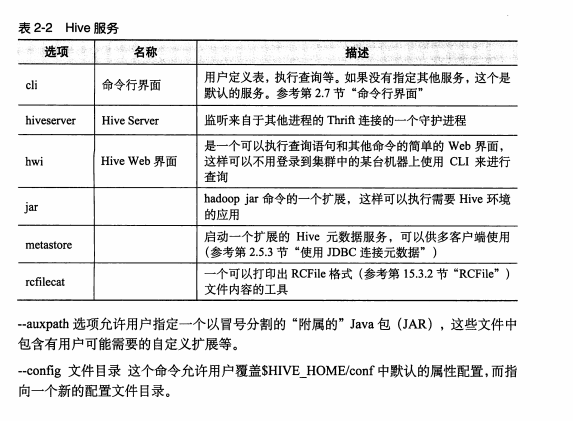
PS：对于非常大的文件，当指向查询开始或结尾部分信息时，没有提供-more，-head或者-tail子命令。替代方式是将-cat命令的输出通过管道传递给shell中的more，head，tail命令。

1. Hive命令
   1. 帮助命令

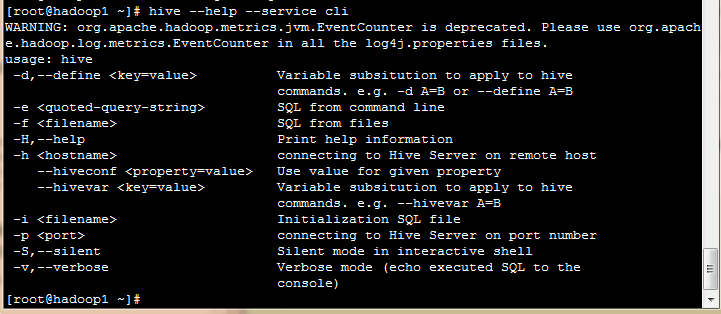
root# hive –help #显示hive的帮助命令



最有用的服务：



* 1. 显示hive Cli的帮助

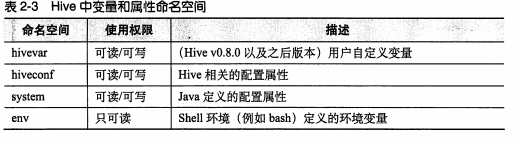


* 1. 变量和属性

--define key=value 和—hivevar key=value是等价的，在命令行自定义变量，只在hive 0.8及以后版本中支持。

Hive会将这些键值对放在hivevar命名空间，可以和内置命名空间（hiveconf、system、env）区分。

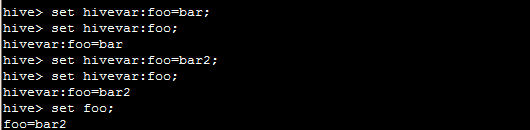
变量或属性是在不同上下文中使用的术语，但在大多数情况下功能相同。



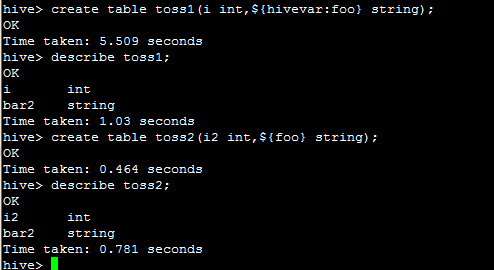
* 1. 定义和改变变量

root#hive –define foo=bar;

hive > set hivevar:foo=bar;#两者都是定义hivevar命名空间的foo变量的值为bar。



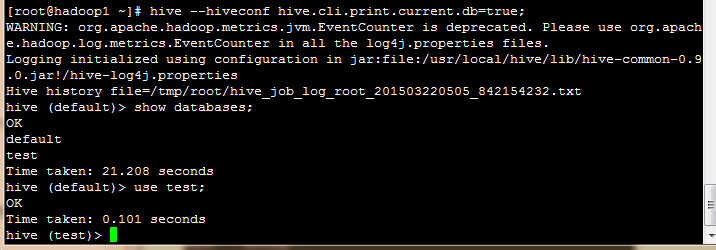
#如果变量在命名空间中唯一也可以只使用变量名，即前缀hivevar:是可选的。



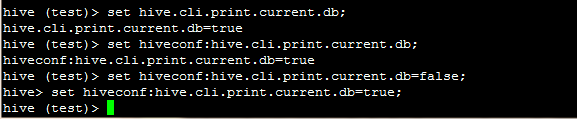
在查询语句中便令引用会先被值替换掉然后提交给查询处理器。

* 1. 显示数据库名称

[root@hadoop1 ~]# hive --hiveconf hive.cli.print.current.db=true;



在hive客户端设置DB名称是否显示

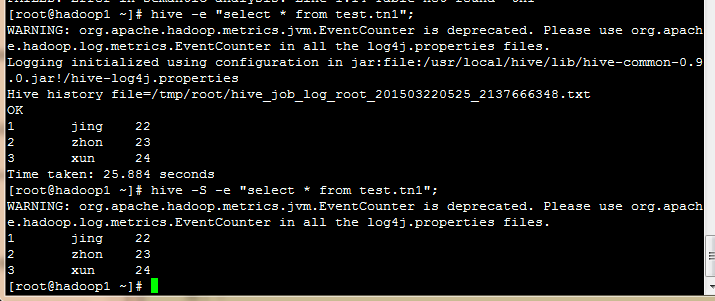


* 1. “一次使用”命令

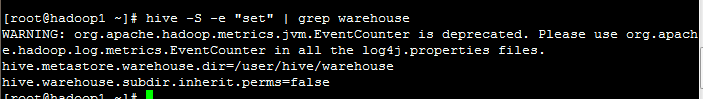
#如果使用的不是默认的数据库，就需要使用dbName.tableName代替tableName

root#hive –e “select \* from DB\_name.table\_name limit 3”;

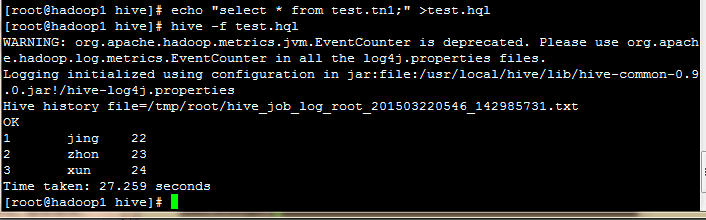
#添加 –S参数可以去掉输出中的ok和time taken两行，但不会去掉警告信息。



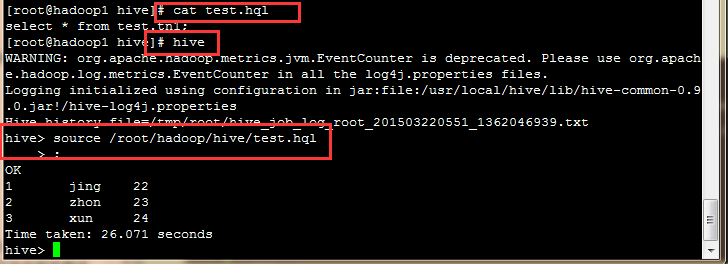
* 1. 属性名模糊查询

[root@hadoop1 ~]# hive -S -e "set" | grep warehouse 

* 1. 从文件中执行hive查询(两种方式)
     1. 直接从本机命令执行文件中的语句



* + 1. 从hive shell中执行脚本文件



* 1. Tab键自动补全

如果在提示符后直接敲击tab键，会看到如下回复：



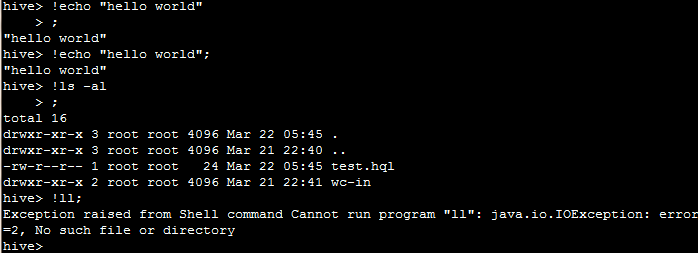
所以在输入查询语句，尤其是文件中语句前面一般不要放Tab字符，否则会导致命令执行失败！！

* 1. 查看操作命令历史

使用上下箭头滚动查看命令，使用左右箭头移动光标位置，修改命令，执行命令直接enter键，不需要光标移动到命令尾部。

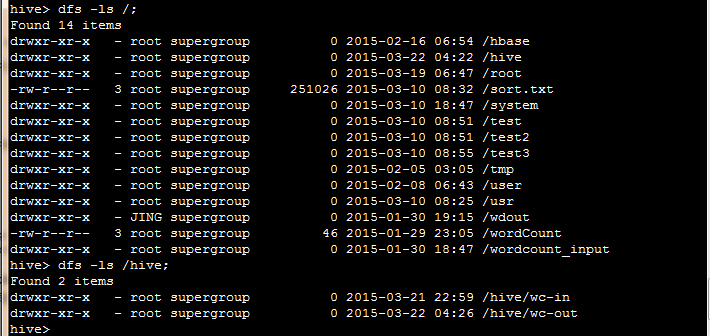
* 1. 执行shell命令

不需要退出hive CLI就可以执行bash shell命令，在命令前加上”!”,以分号结尾即可。可以使用shell命令，但不能使用别名。只能是shell的命令。



* 1. 在hive中使用hadoop的dfs命令

直接将hadoop命令的关键字hadoop去掉然后以分号结尾即可。



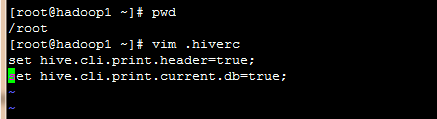
在hive中执行hadoop命令比在bash shell中执行的hadoop命令高效，因为后者会启动新的JVM实例，而hive会在同一个进程中执行这些命令。

* 1. Hive脚本如何注释

在hive中，以—开头的字符串表示注释

* 1. 显示字段名称

--显示字段名称，可以将这句话放置在$HOME/.hiverc文件中



hive (default)> set hive.cli.print.header=true;

