

7、hive

- 1、安装hive

hive一般在工作站上运行。它把sql查询转换为一系列在Hadoop集群上运行的作业。hive把数据组织为表，通过这种方式为存储在hdfs上的数据赋予结构。元数据（如表模式）存储在metastore数据库中。

- 2、hive shell

hive的shell环境是我们和hive交互、发出hiveQL命令的主要方式。第一次启动hive时，我们可以通过列出hive的表（show tables;）来检查hive是否正常工作，此时应该没有任何表。命令必须以分号结束，告诉hive立刻执行该命令。

- 非交互式hive shell

使用-f选项可以运行指定文件中的命令。比如要运行脚本文件script.q: `hive -f script.q`

- 运行较短的脚本

可以使用-e选项在行内嵌入命令。此时不需要表示结束的分号，比如hive -e "select * from dummy"

- 运行hive

- 配置hive

hive的配置文件为hive-site.xml，它在hive的conf目录下。通过这个文件，可以设置每次运行hive时希望hive使用的选项。该目录下还包括hive-default.xml（其中记录hive使用的选项及其默认值）

- --config参数

使用--config选项参数给hive命令，可以通过这种方式重新定义hive查找hive-site.xml文件的目录（注意是目录）：`hive --config /Users/tom/dev/hive-conf`

- hive-site.xml

适合存放集群连接信息，因为可以使用Hadoop属性fs.defaultFS和yarn.resourcemanager.address来指定文件系统和资源管理器，如果没有指定这两个参数，它们就像在Hadoop中一样，被设为默认值，也就是使用本地文件系统和本地作业运行期。metastore的配置选项一般也能在hive-site.xml中找到。

- 执行引擎

hive的原始设计是以mapreduce作为执行引擎，目前hive的执行引擎还包括apache tez，另外hive对spark的支持也在开发中，具体使用那种引擎由属性hive.execution.engine来控制，其默认值为mr