Hive 入门

翻译作品

屈庆磊

quqinglei@pwrd.com

2013 年 *9* 月 *11* 日

目录

1 安装配置

1.1 安装要求

首先要已经安装了 *Hadoop*，本文档不讲述 *Hadoop* 的安装方式和使用方法。

- Java 1.6 以及更高版本
- Hadoop 0.20.x 以及更高版本

1.2 安装稳定版本的 HIVE

用户可以从 *Apache* 的官方站点下载 *Hive* 的稳定版本：*http://hive.apache.org/releases.html*

```
# 1. 解压安装包，hive 安装包一般命名为 hive-x.y.z
$ tar -xzvf hive-x.y.z.tar.gz

```

```
# 2. 设置环境变量
$ cd hive-x.y.z

$ export HIVE_HOME={{pwd}} # 临时的环境变量，一般我们要写到脚本里或者 bashrc 文件里
# 比如我们把 hive 解压到了/home/hadoop/hive 路径，那么我们可以如下面一样导出环境变量
$ export HIVE_HOME=/home/hadoop/hive-x.y.z
$ export PATH=$PATH:$HIVE_HOME/bin

# 其实以上步骤执行完就算是安装完了，它就是个解压包，无需配置过多
# 至于编译安装，不再介绍，看官网的吧，此文档为自己学习使用，为加深印象
```

2 运行 HIVE

Hive 运行需要以下两个条件:

- 你必须在你的路径里安装了 Hadoop
- export HADOOP_HOME=<hadoop-instsall-dir> 或许你已经在安装 Hadoop 时导出了环境变量可以使用 `echo $HADOOP_HOME` 进行测试。

现在就可以直接使用了，直接在 Terminal 里面输入：`hive` 就可以直接进入 hive 的命令行了。你会得到如下所示的命令行输入界面:

```
hive>
```

2.1 配置管理概述

- Hive 的的默认配置位置为`<install-dir>/conf/hive-default.xml`
- 配置文件另是可以通过环境变量`HIVGE\_CONF\_DIR` 去改变的
- Log4j[1] 的配置文档在`<install-dir>/hive-log4j.properties`
- 配置文件建立在 Hadoop 的配置之上，并默认继承 Hadoop 的配置
- Hive 的配置可以通过以下途径更改:
  - 更改`hive-site.xml` 文件
  - 通过命令行更改，如：`hive -hiveconf x1=y1 -hiveconf x2=y2` 此命令分别把 y1 和 y2 赋值给变量 x1 和 x2
  - 通过设置环境变量 `HIVE_OPTS="hiveconf x1=y1 hiveconf x2=y2"` 可以达到同样的效果。
- 运行时配置
  - Hive 的查询使用了 map-reduce，因此查询行为可以通过更改 hadoop 配置变量来控制

[1]Log4j 简单的说就是 Log for java，即 java 的日志组件

– 命令行中的'SET' 可以用来设置 hadoop 或者 hive 的配置变量，如下：

```
# SET -v 会打印出现在的所有变量信息，不加-v 只会打印 hive 的变量信息
hive> SET mapred.job.tracker=myhost.mycompany.com:50030;
hive> SET -v;
```

2.2 Hive, 本地和 mapreduce

Hive 编译器为大部分的查询生成 map-reduce 作业。这些作业是否提交为 Map-Reduce 集群取决于变量：`mapred.job.tracker`

Hadoop 也提供一个小巧的参数，用来指定作业是运行在集群上还是运行在用户的工作站上，虽然一般来说一个 map-reduce 集群就意味着多个节点。当数据比较小的时候，在单台机器上的运行速度反而要比在集群上运行速度快。数据可以直接从 HDFS 上透明访问。当然数据量大的时候集群处理肯定比本地运行速度快，必经本地只有一个 reducer。

Hive 完全支持本地模式运行，下面的参数可以设置其本地运行：

```
hive> SET mapred.job.tracker=local;
```

另外变量 `mapred.local.dir` 应该指向一个路径，否则会收到一个分配磁盘空间的异常，一般定义为：`/tmp/<username>/mapred/local`

我们还可以通过变量 `hive.exec.mode.local.auto` 来控制 hive 是否能自动执行本地模式。此模式默认是关闭的，如果打开此模式 Hive 会分析每一个 map-reduce 作业的大小，并根据以下参数确定是否执行本地模式：

- 输入数据量小于参数：`hive.exec.mode.local.auto.inputbytes.max` 的设置值，默认为 128MB
- Map-task 数量少于参数：`hive.exec.mode.local.auto.tasks.max` 的值，默认为 4
- reduce-task 应该是 1 或者 0