# Hadoop快速入门

## 目录

1	目的	. 2
2	先决条件	. 2
	2.1 支持平台	. 2
	2.2 所需软件	. 2
	2.3 安装软件	
3	下载	. 3
4	运行Hadoop集群的准备工作	. 3
5	单机模式的操作方法	. 3
6	伪分布式模式的操作方法	. 3
	6.1 配置	. 3
	6.2 免密码ssh设置	. 4
	6.3 执行	. 4
7	完全分布式模式的操作方法	

## 1. 目的

这篇文档的目的是帮助你快速完成单机上的Hadoop安装与使用以便你对<u>Hadoop分布式文件系统(HDFS)</u>和Map-Reduce框架有所体会,比如在HDFS上运行示例程序或简单作业等。

### 2. 先决条件

#### 2.1. 支持平台

- GNU/Linux是产品开发和运行的平台。 Hadoop已在有2000个节点的GNU/Linux主机 组成的集群系统上得到验证。
- Win32平台是作为开发平台支持的。由于分布式操作尚未在Win32平台上充分测试, 所以还不作为一个生产平台被支持。

#### 2.2. 所需软件

Linux和Windows所需软件包括:

- 1. JavaTM1.5.x,必须安装,建议选择Sun公司发行的Java版本。
- 2. ssh 必须安装并且保证 sshd一直运行,以便用Hadoop 脚本管理远端Hadoop守护进程。

Windows下的附加软件需求

1. Cygwin - 提供上述软件之外的shell支持。

### 2.3. 安装软件

如果你的集群尚未安装所需软件,你得首先安装它们。

以Ubuntu Linux为例:

\$ sudo apt-get install ssh
\$ sudo apt-get install rsync

在Windows平台上,如果安装cygwin时未安装全部所需软件,则需启动cyqwin安装管理器安装如下软件包:

• openssh - Net 类

### 3. 下载

为了获取Hadoop的发行版,从Apache的某个镜像服务器上下载最近的 稳定发行版。

## 4. 运行Hadoop集群的准备工作

解压所下载的Hadoop发行版。编辑 conf/hadoop-env.sh文件, 至少需要将JAVA\_HOME 设置为Java安装根路径。

尝试如下命令:

\$ bin/hadoop

将会显示hadoop 脚本的使用文档。

现在你可以用以下三种支持的模式中的一种启动Hadoop集群:

- 单机模式
- 伪分布式模式
- 完全分布式模式

## 5. 单机模式的操作方法

默认情况下,Hadoop被配置成以非分布式模式运行的一个独立Java进程。这对调试非常有帮助。

下面的实例将已解压的 conf 目录拷贝作为输入,查找并显示匹配给定正则表达式的条目。输出写入到指定的output目录。

- \$ mkdir input
- \$ cp conf/\*.xml input
- \$ bin/hadoop jar hadoop-\*-examples.jar grep input output 'dfs[a-z.]+'
- \$ cat output/\*

## 6. 伪分布式模式的操作方法

Hadoop可以在单节点上以所谓的伪分布式模式运行,此时每一个Hadoop守护进程都作为一个独立的Java进程运行。

#### 6.1. 配置

#### 使用如下的 conf/hadoop-site.xml:

## 6.2. 免密码ssh设置

现在确认能否不输入口令就用ssh登录1oca1host:

\$ ssh localhost

如果不输入口令就无法用ssh登陆localhost, 执行下面的命令:

\$ ssh-keygen -t dsa -P '' -f ~/.ssh/id dsa

\$ cat ~/.ssh/id\_dsa.pub >> ~/.ssh/authorized\_keys

#### 6.3. 执行

格式化一个新的分布式文件系统:

\$ bin/hadoop namenode -format

启动Hadoop守护进程:

\$ bin/start-all.sh

Hadoop守护进程的日志写入到 \${HADOOP\_LOG\_DIR} 目录 (默认是 \${HADOOP\_HOME}/logs).

浏览NameNode和JobTracker的网络接口,它们的地址默认为:

- NameNode http://localhost:50070/
- JobTracker http://localhost:50030/

将输入文件拷贝到分布式文件系统:

\$ bin/hadoop fs -put conf input

运行发行版提供的示例程序:

 $\$  bin/hadoop jar hadoop-\*-examples.jar grep input output 'dfs[a-z.]+'

#### 查看输出文件:

将输出文件从分布式文件系统拷贝到本地文件系统查看:

\$ bin/hadoop fs -get output output

\$ cat output/\*

#### 或者

在分布式文件系统上查看输出文件:

\$ bin/hadoop fs -cat output/\*

完成全部操作后, 停止守护进程:

\$ bin/stop-all.sh

## 7. 完全分布式模式的操作方法

关于搭建完全分布式模式的,有实际意义的集群的资料可以在这里找到。

Java与JNI是Sun Microsystems, Inc.在美国以及其他国家地区的商标或注册商标。