kettle入门(三) 之kettle连接hadoop&hdfs图文详解 - xiaohai798的专栏 - CSDN博客

1 引言：

项目最近要引入大数据技术，使用其处理加工日上网话单数据，需要kettle把源系统的文本数据load到hadoop环境中

2 准备工作：

1 首先

要了解**支持hadoop的Kettle版本情况**，由于kettle资料网上较少，所以最好去官网找,官网的url:

http://wiki.pentaho.com/display/BAD/Configuring+Pentaho+for+your+Hadoop+Distro+and+Version

打开这个url 到页面最下面的底端，如下图：

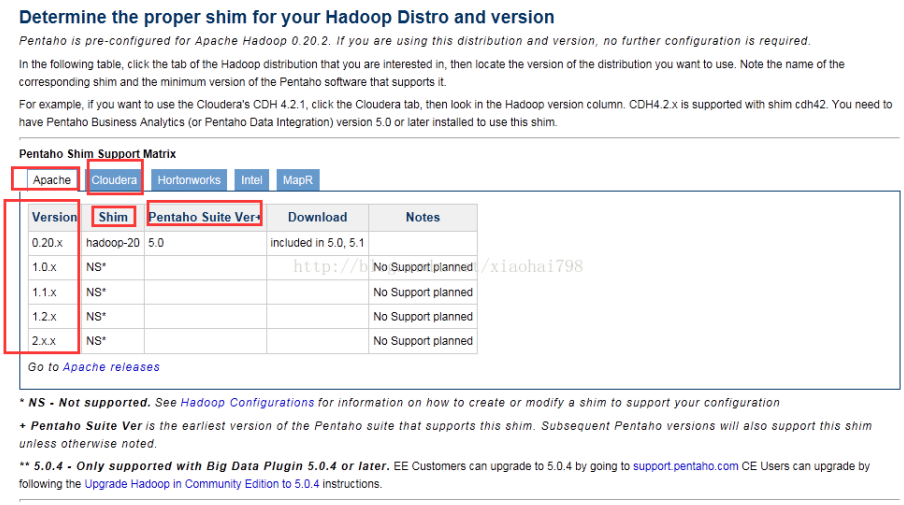
archive 下面的from PDI 4.3 、 from PDI 4.4 、 from PDI 5.0 即表示支持hadoop的pdi 版本。pdi即pentaho data integration 又称kettle。PDI 4.3 、 PDI 4.4 、 PDI 5.0 即是kettle 4.3 、4.4、 5.0 ,这个版本号 包括比其更高的版本（即kettle 5.0.X ， 5.1，5.2也支持hadoop）。

2 其次

不同的kettle版本支持的hadoop版本不一样，以5.1为例子，下面的链接是5.1的支持情况

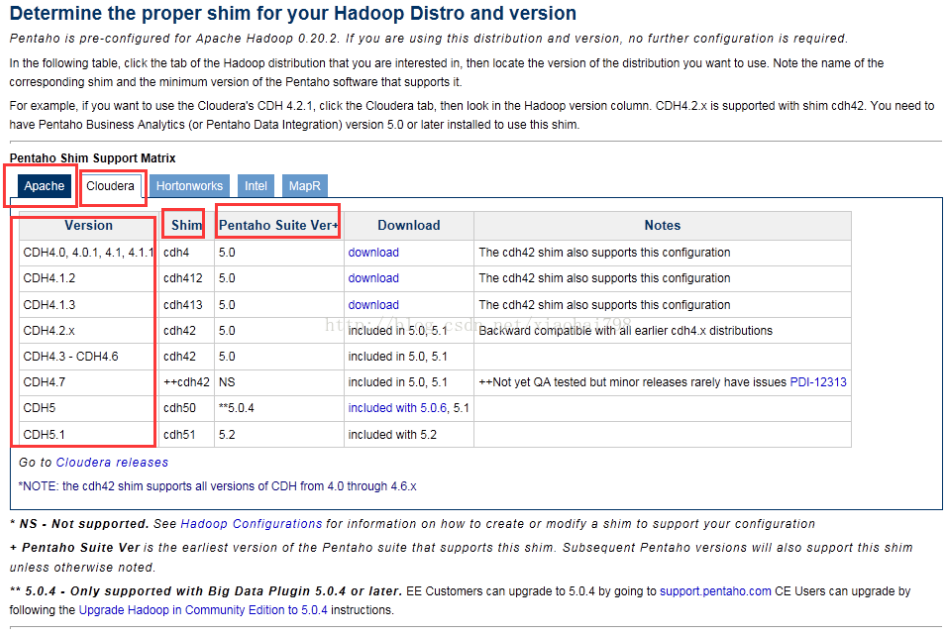
http://wiki.pentaho.com/display/BAD/Configuring+Pentaho+for+your+Hadoop+Distro+and+Version

下图为链接打开的页面的中间部分：



determine the proper shim for hadoop Distro and version 大概意思是 为hadoop版本选择合适的套件。表格上面的一行：apache、cloudera、hortonworks、intel、mapr指的是发行方。点击他们来选择你 想连接的hadoop的发行方 。上图 以apache hadoop为例：

Version 指版hadoop版本号 ， shim 指kettle提供给该hadoop套件的名称，Download 里面的 included in 5.0,5.1 指kettle的5.0、5.1版本安装包里面已经有内置的插件，一句话来讲 就是kettle5.1及5.0版本已有插件提供支持apache hadoop版本0.20.x 。不需要额外下载。NS 是不支持的意思 图片下面也有解释。



Determine the proper shim for your Hadoop Distro and version

Pentaho is pre-configured for Apache Hadoop 0.20.2. If you are using this distribution and version, no further configuration is required.

In the following table, click the tab of the Hadoop distribution that you are interested in, then locate the version of the distribution you want to use. Note the name of the corresponding shim and the minimum version of the Pentaho software that supports it.

For example, if you want to use the Cloudera's CDH 4.2.1, click the Cloudera tab, then look in the Hadoop version column. CDH4.2.x is supported with shim cdh42. You need to have Pentaho Business Analytics (or Pentaho Data Integration) version 5.0 or later installed to use this shim.

Pentaho Shim Support Matrix

Apache | Cloudera | Hortonworks | Intel | MapR

Version	Shim	Pentaho Suite Ver+	Download	Notes
CDH4.0, 4.0.1, 4.1, 4.1.1	cdh4	5.0	download	The cdh42 shim also supports this configuration
CDH4.1.2	cdh412	5.0	download	The cdh42 shim also supports this configuration
CDH4.1.3	cdh413	5.0	download	The cdh42 shim also supports this configuration
CDH4.2.x	cdh42	5.0	included in 5.0, 5.1	Backward compatible with all earlier cdh4.x distributions
CDH4.3 - CDH4.6	cdh42	5.0	included in 5.0, 5.1	
CDH4.7	++cdh42	NS	included in 5.0, 5.1	++Not yet QA tested but minor releases rarely have issues PDI-12313
CDH5	cdh50	**5.0.4	included with 5.0.6, 5.1	
CDH5.1	cdh51	5.2	included with 5.2	

Go to Cloudera releases

*NOTE: the cdh42 shim supports all versions of CDH from 4.0 through 4.6.x

* ***NS - Not supported.*** *See Hadoop Configurations for information on how to create or modify a shim to support your configuration*

+ ***Pentaho Suite Ver*** *is the earliest version of the Pentaho suite that supports this shim. Subsequent Pentaho versions will also support this shim unless otherwise noted.*

** ***5.0.4 - Only supported with Big Data Plugin 5.0.4 or later.*** EE Customers can upgrade to 5.0.4 by going to support.pentaho.com CE Users can upgrade by following the Upgrade Hadoop in Community Edition to 5.0.4 instructions.

上图说明的是对 cloudera的 hadoop支持的情况 ， Download 里面 download的蓝色字体超链接的说明 是要除了下kettle的安装包外另外下载的 ， 带 included in 5.0,5.1 说明 kettle 5.0,5.1版本的本身就支持（内置有插件）。

由上面两图得到的结论是 kettle 5.1 支持 apache hadoop 0.20.x版本 及cloudera hadoop CDH4.0 到CDH5。

3 试验运行：

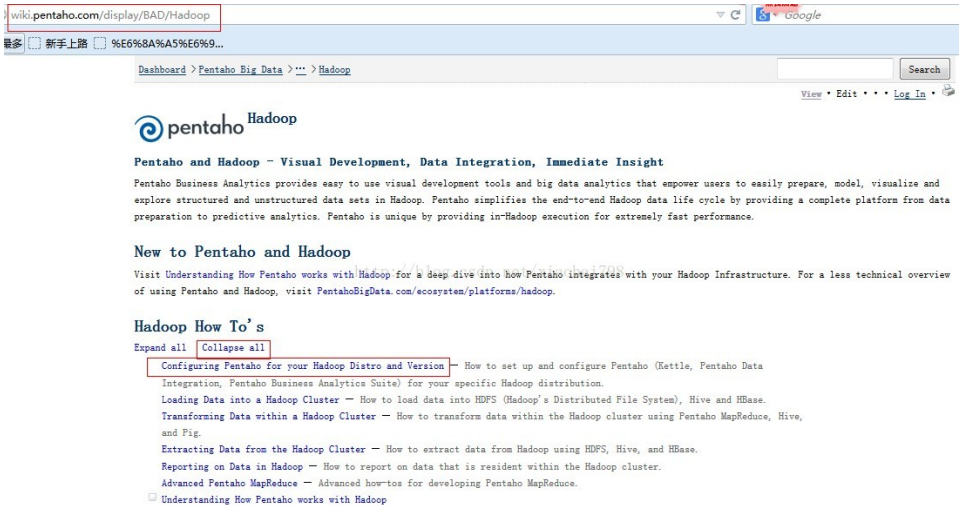
1 首先配置工作

当前我用的hadoop 版本是hadoop-2.2.0-cdh5.0 所以用kettle 5.1 且其内置有hadoop插件。去kettle官网下载：

解压之后 就是：

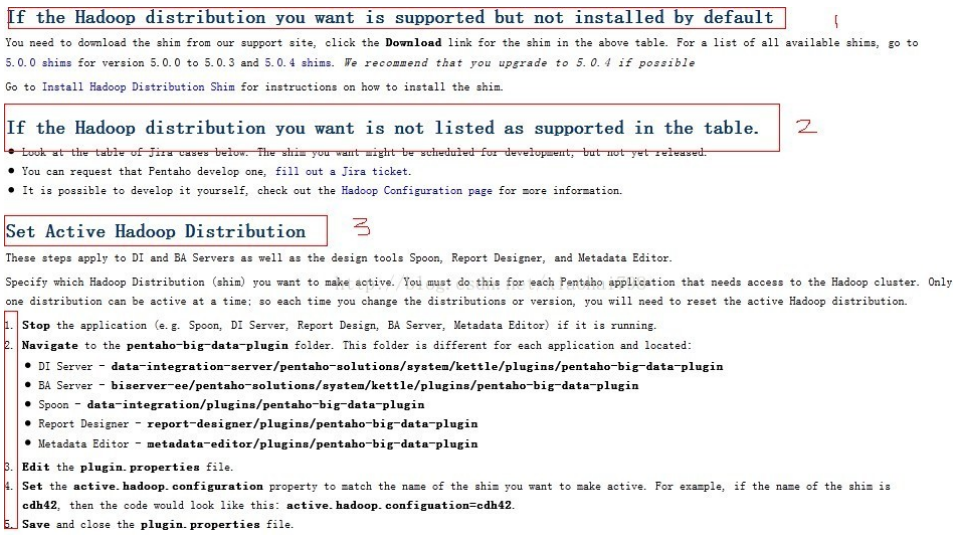
下载好之后，现在就需要做配置的工作了，配置的工作在kettle安装文件里面做：

配置办法参考：http://wiki.pentaho.com/display/BAD/Hadoop



进页面之后 先点击collapse 收缩所有的菜单树 如上图。 Configuring Pentaho for your Hadoop Distro and Version 意思是为hadoop 版本做配置 点击进去：页面的上面 就是上面说过的kettle对hadoop的支持情况。

我们到页面的中间部分去，如下图：



1 意思是 你想要连接的hadoop发行版 已经被kettle支持了，但是没有内置插件，需要下载，这种情况最好 看下：Install Hadoop Distribution Shim

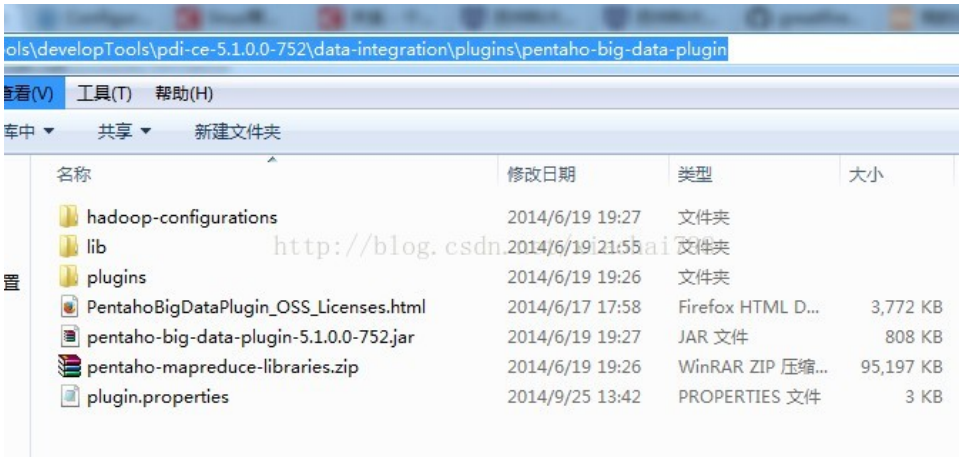
2 意思是你想连接的hadoop发行版 还有没有被kettle支持，可以自己填写相应的信息 要求pentaho 开发一个。

还有1种情况 就是上面说的hadoop发行版 已经被kettle支持了 且有内置的插件。

3 就是配置了。

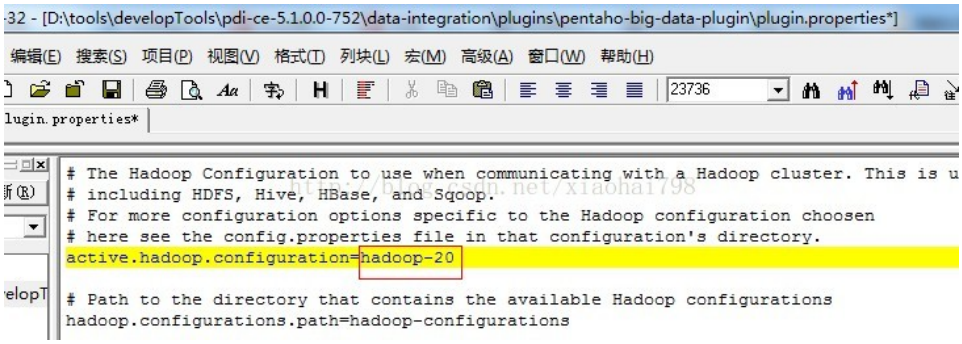
3.1 stop application 就是如果kettle在运行 先停掉他。

3.2 打开安装文件夹 我们这边是kettle 所以就是spoon那个的文件路径：



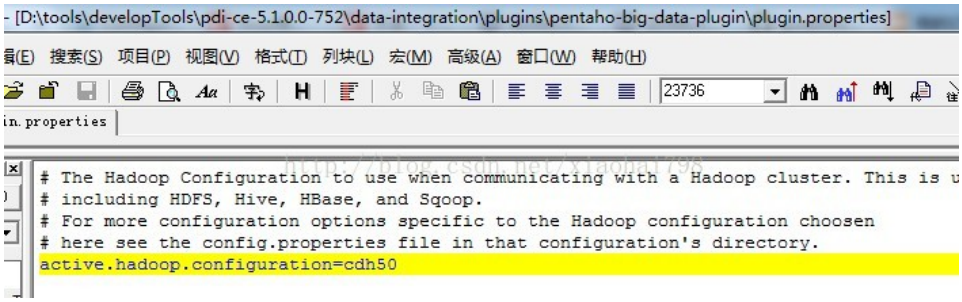
3.3 编辑 plugin.properties文件

3.4 改一个配置值 下图画圈的地方



改成 对应你hadoop的shim值 （上图的表格里面的shim） 我这边是cdh50:

改之后保存：



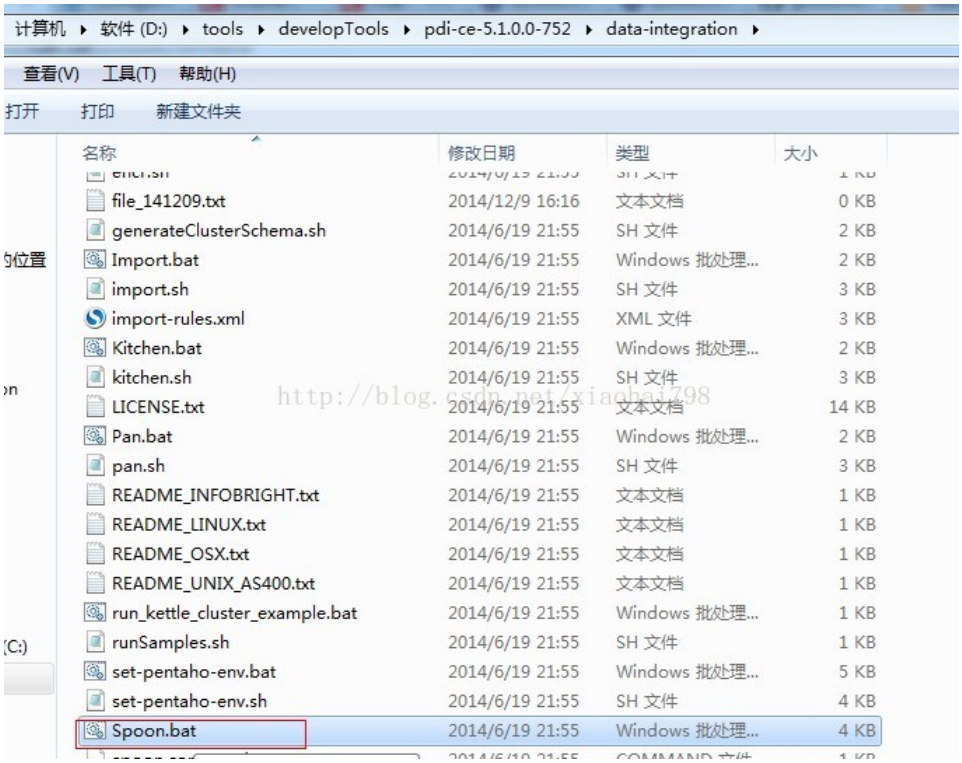
至此 配置工作做完。

2 然后开发脚本工作

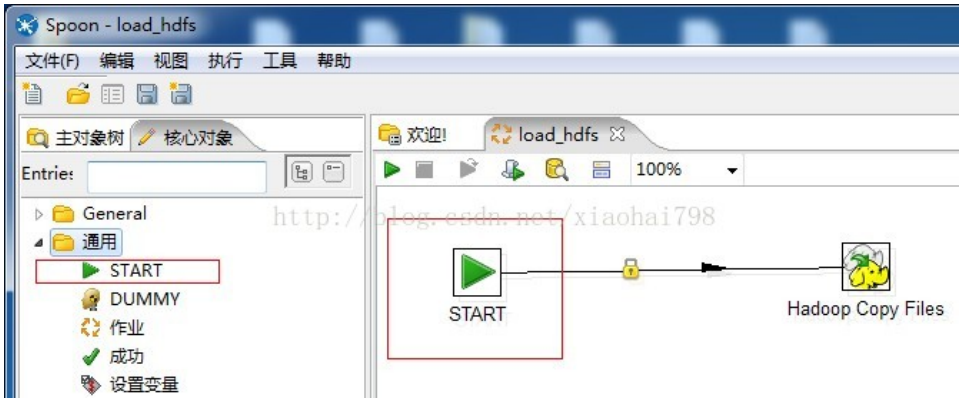
下面开始开发脚本 官方参考：http://wiki.pentaho.com/display/BAD/Loading+Data+into+HDFS

打开 kettle 运行spoon.bat

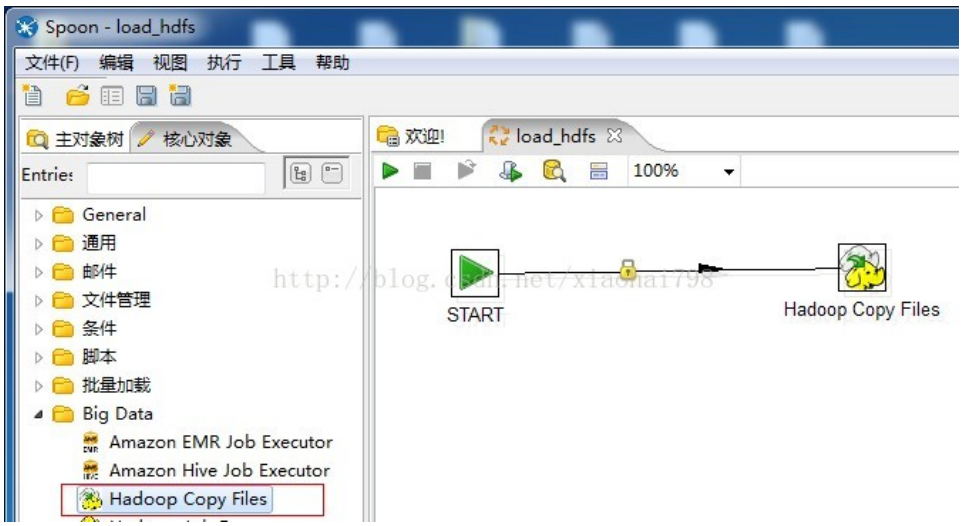
：



新建一个kjb文件 拖一个开始图元

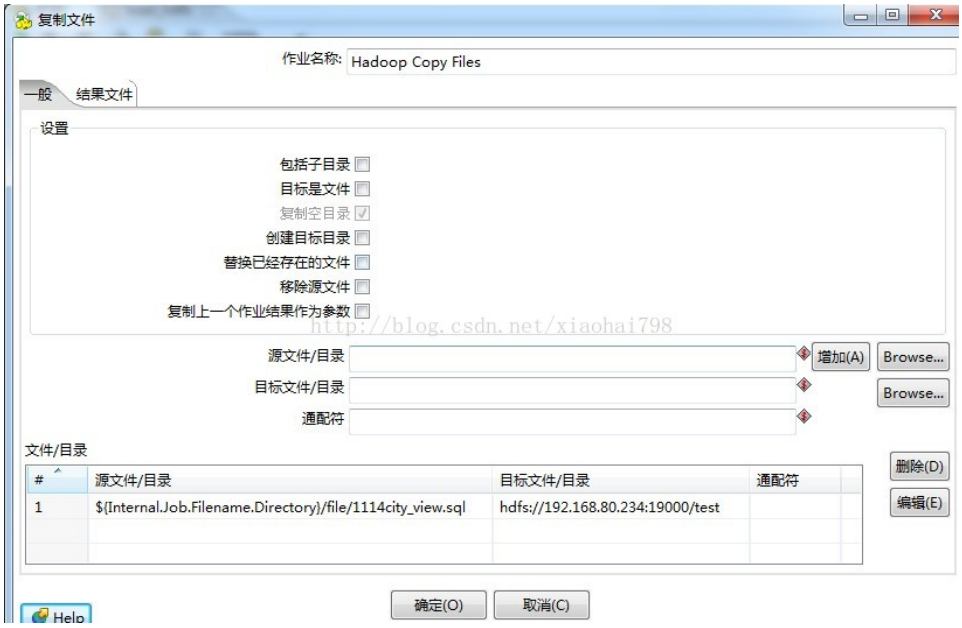


再拖一个

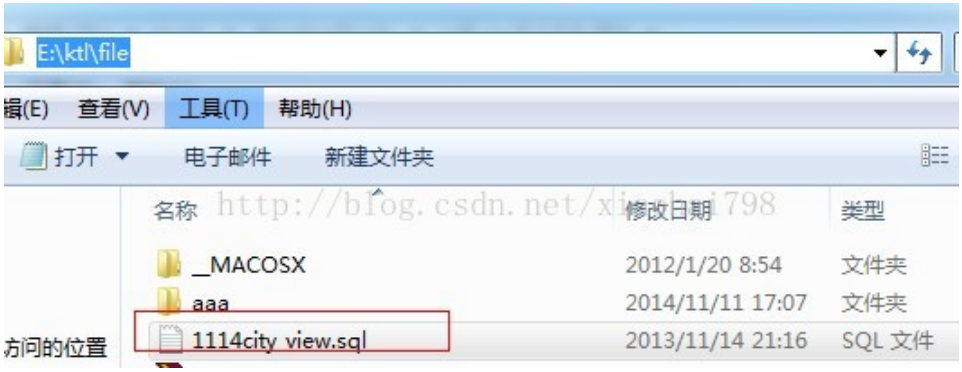


hadoop copy files即是 load数据到 hdfs里面。

copy files里面的配置：



意思是当前kjb脚本所在路径 在我这边文件夹是：

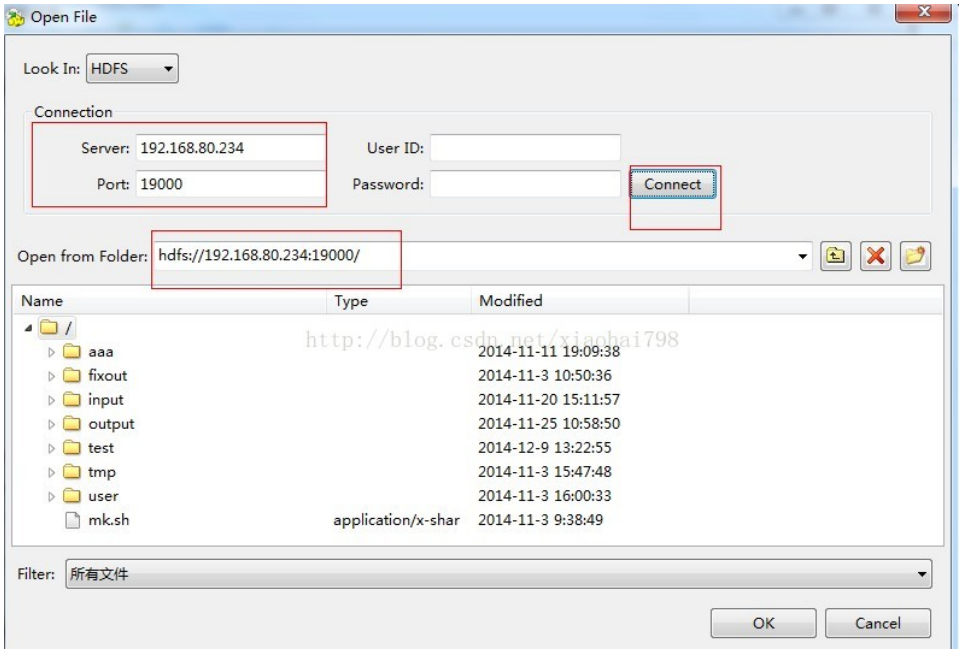


目标文件 是 hdfs://ip:hdfs端口/路径

填之前可以点击

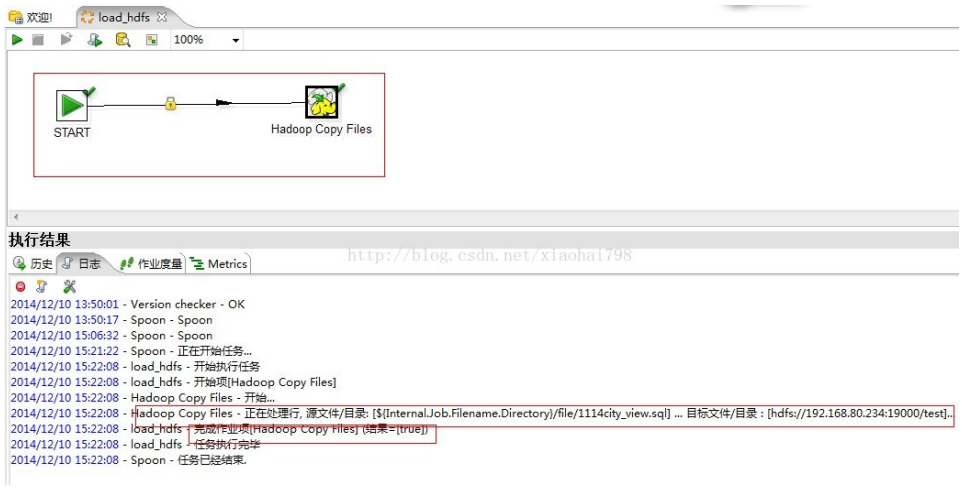
browse 按钮 测试

如下图 ：填好server 和port后 点击connect 如果没有报错 出现红框里面的hdfs://..... 就说明连接成功了（如下图）。



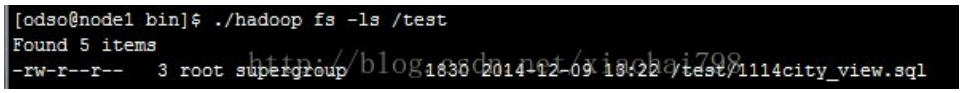
注意只要连接成功，说明kettle对hadoop的配置就没有问题。

可以运行脚本试试了：



如上图，脚本运行成功。

在hadoop home bin下面查看：



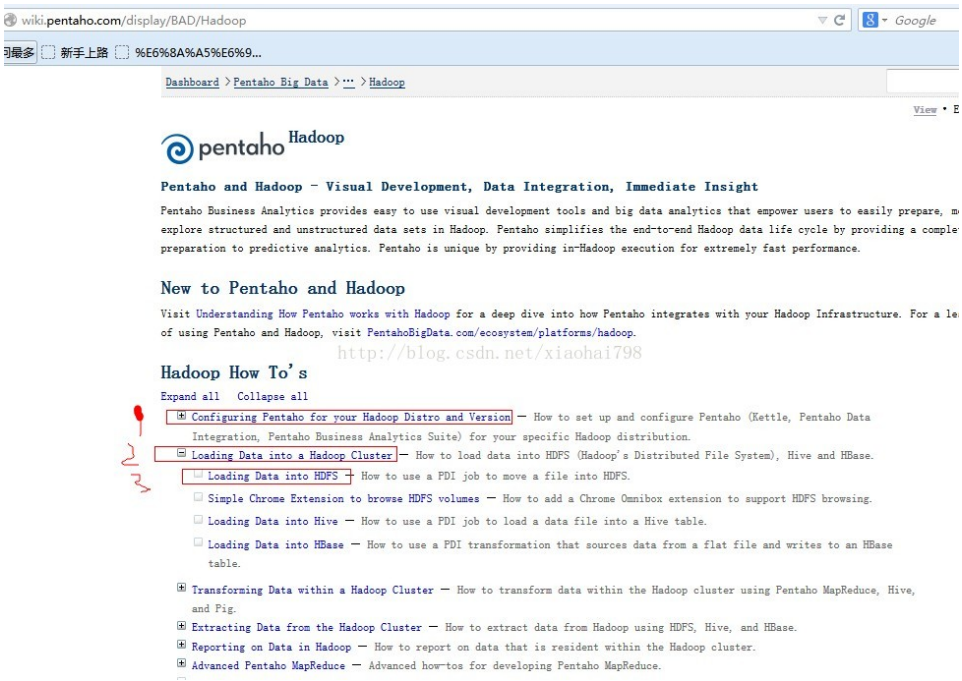
文件成功load.

至此，kettle load文本数据到hdfs成功！

4 备注：

所有的步骤都可以参考官网：

http://wiki.pentaho.com/display/BAD/Hadoop



上图 1 是配置 2 是加载数据到hadoop 集群 3 是加载数据到hdfs 还有其他到 hive 到hbase等。

ps：写一段长的博客真累，感觉比干活还累