

《数据采集与预处理》

教材官网: http://dblab.xmu.edu.cn/post/data-collection/

温馨提示:编辑幻灯片母版,可以修改每页PPT的厦大校徽和底部文字



第7章 ETL工具Kettle

(PPT版本号: 2022年1月版本)

林子雨 副教授

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页: http://dblab.xmu.edu.cn/linziyu





7.1 Kettle的基本概念
7.2 Kettle的基本功能
7.3 安装Kettle
7.4 数据抽取
7.5 数据清洗与转换
7.6 数据加载

本PPT是以下教材的配套讲义 林子雨编著《数据采集与预处理》 人民邮电出版社

教材官网: http://dblab.xmu.edu.cn/post/data-collection



厦门大学计算机科学与技术系

「「「「雨」

美人民邮电出版社 POSTS & TELECOM PRESS

美学生的**数据意识,数据思维和数据能力**

」中国工信出版集团

生形成对大数据专业的整体认知



7.1 Kettle的基本概念

一个数据抽取过程(如图7-1所示)主要包括创建一个作业(Job),每个 作业由一个或多个作业项(Job Entry)和连接作业项的作业跳(Job Hop) 组成。每个作业项可以是一个转换(Transformation)或是另一个作业。 一个转换由一个或多个步骤(Step)和连接步骤的跳(Hop)组成。



图7-1 一个数据抽取过程的构成要素



转换主要用于数据的抽取(Extraction)、转换(Transformation)以及 加载(Load),比如读取文件、过滤输出行、数据清洗或加载到数据库 等步骤。一个转换包含一个或多个步骤,每个步骤都是单独的线程,当 启动转换时,所有步骤的线程几乎并行执行。步骤之间的数据以数据流 方式传递。所有的步骤都会从它们的输入跳中读取数据,并把处理过的 数据写到输出跳,直到输入跳里不再有数据就终止步骤的运行;当所有 步骤都终止了,整个转换就终止了。由于转换里的步骤依赖前一个步骤 获取数据,因此转换里不能有循环。



相较于转换,作业是更加高级的操作。作业由一个或多个作业项(作业或转换)组成。所有的作业项是以某种自定义的顺序串行执行的。作业项之间可以传递一个包含了数据行的结果对象。当一个作业项执行完成后,再传递结果对象给下一个作业项。作业里可以有循环。

跳是步骤之间带箭头的连接线,它定义了一个单向通道,用于连接两个步骤,实现将数据从一个步骤(写入数据到行集)流向另一个步骤(从行集 中读取数据)。跳是两个步骤之间的被称为"行集"(Row Set)的数据 行缓存(可以在转换设置中定义行集大小)。若行集满了,则向行集写数 据的步骤将停止写入,直到行集里又有空间。若行集空了,则从行集读取 数据的步骤就会停止读取,直到行集里又有可读取的数据行。跳对于向行 集写入数据的步骤来说是输出跳,一个步骤可以拥有多个输出跳;跳对于 从行集中读取数据的步骤来说是输入跳。

作业跳是作业项之间带箭头的连接线,它定义了作业的执行路径。



7.2 Kettle的基本功能

Kettle的基本功能包括转换管理和作业管理。转换管理主要包括输入、输出、转换、应用、流程、脚本、查询、检验、作业、映射和批量加载等功能。作业管理主要包括通用、邮件、文件管理、条件、脚本、批量加载等功能。

林子雨



7.3 安装Kettle

在Windows系统中打开浏览器,访问Kettle官网 (https://sourceforge.net/projects/pentaho/),下载Kettle安装文件pdice-9.1.0.0-324.zip。或者,也可以直接到教材官网的"下载专区"的 "软件"目录中下载pdi-ce-9.1.0.0-324.zip文件。

把pdi-ce-9.1.0.0-324.zip解压缩到"D:\"目录下(或者也可以选择一个 其他目录,比如"C:\"),会生成一个"data-integration"目录,该目 录下就包含了Kettle。在data-integration目录里包含了Spoon的启动文件, 即spoon.bat,双击该文件就可以启动Spoon,启动界面如图7-2所示。



7.3 安装Kettle



图7-2 Spoon启动界面





启动成功以后的界面如图7-3所示。

<mark>麗</mark> Spoon - 欢迎! 文州(F) 編編 細图 执行 丁目 邦助					
					Connect
☑ 主对象树					
Search X 🖫 🗄 🗖					
□ 转换				- + ¹	
	vveice	ome to Pentar	no Data Integr	ation	
		·۲۲	–	00	
				670	
		_			
	WORK	LEARN	EXTEND	DISCOVER	
	Open files	Documentation	Marketplace	Pentaho News	
	New transformation	Watch tutorials	Build plugins	Pentaho Roadmap	
	New job	Release notes		Events & Webcasts	
	O Community		Forums	hitachiyantara.com	
	Join the discussion >	Join the o	discussion >	Visit us >	

图7-3 Spoon启动以后的欢迎界面

《数据采集与预处理》

厦门大学计算机科学与技术系

林子雨



7.4 数据抽取

7.4.1 把文本文件导入到Excel文件中 7.4.2 把文本文件导入MySQL数据库中(请直接参考教材) 7.4.3 把Excel文件导入到MySQL数据库中(请直接参考教材)



- 这里给出一个实例,演示如何使用Kettle把 文本文件导入到Excel文件中,具体包括如 下步骤:
- •创建文本文件;
- •建立转换;
- •设计转换;
- •执行转换。



1.创建文本文件

在"D:\"目录下新建一个文本文件studentinfo.txt,其内容如图7-4所示, 文件的第1行是字段名称,包括sno、name、sex和age,字段之间用"|" 隔开,其余行都是记录,字段之间也是用"|"隔开。



图7-4 studentinfo.txt文件内容



2.建立转换

在Spoon主界面的"主对象树"栏目中,在"转换"上面(如图7-5所示)单击鼠标右键,在弹出的菜单中点击"新建"。点击Spoon主界面左上角的"保存"图标,把这个转换保存到某个路径下并且名称为"text_to_excel"。

×	Spoo	n - 欢ì	<u>íp</u> i				
文	件(F)	编辑	视图	执行	I具	帮助	
Ð	ß		- 6	\otimes	•		
G] 主对:	象树	ℓ核	心对象			
	Searc	h		>	<	ዲ 83	=
]转换					
		」作业	<u>.</u>				

图7-5 新建"转换"

《数据采集与预处理》



3.设计转换

在"核心对象"栏目中,在"输入"控件里把"文本文件输入"拖到右侧 设计区域,然后在"输出"控件里把"Excel输出"拖到右侧设计区域,然 后为这两个控件建立连线(如图7-6所示),这里的连线就是前文介绍过的 "跳"。为这两个控件建立连线的方法是,按住键盘上的Shift键,然后用 鼠标左键单击"文本文件输入"控件图标,再用鼠标左键单击"Excel输出" 控件图标,最后在其他空白区域单击鼠标左键,这样就建立了一条从"文 本文件输入"到"Excel输出"的连线。



图7-6 放置文本文件输入和Excel输出两个控件



双击设计区域的"文本文件输入"控件,打开设置界面,点击"文件" 选项卡,点击"文件或目录"右侧的"浏览"按钮(如图7-7所示), 把studentinfo.txt文件添加进来,然后点击"增加"按钮, studentinfo.txt文件就会被增加到"选中的文件"中,增加后的效果如 图7-8所示。

D, 文本文件输入		
步骤名称	文本文件输入	
文件 内容 错误处理 过滤 字段 其他輸出字段		
文件或目录		登加 浏览(B)…
规则表达式		♦
正则表达式(排除)		•
选中的文件:		通配符号(
	1	
		删除
		编辑
	<	4
- 从上一步掠获载义件者		
从以前的步骤接受字段名		
步骤读取的文件名来自		Ŧ
在输入里的字段被当作文件名		
	显示文件名」显示文件内容 显示来自第一条数据行的内容	
() Help	确定(O) 预览记录 取消(C)	

图7-7 添加文件



E、文本文件输入							
步骤名称	文本文件输入						
文件 内容 错误处理 过滤 字段 其他輸出字段)						
文件或目录		登					
规则表达式		•					
正则表达式(排除)		\$					
选中的文件:	<i>*</i> 文件/目录	通配符 通配符号(
	1 \${Internal.Entry.Current.Directory}studentinfo.txt	miles					
		编辑					
	<	•					
○从上一步骤获取文件名							
从以前的步骤接受文件名							
从以前的步骤接受字段名							
步骤读取的文件名来自		Ψ					
在输入里的字段被当作文件名							
显示文件名 显示文件内容 显示来自第一条数据行的内容							
(Help	确定(O) 预览记录 取消(C)						

图7-8 添加文件以后的效果



E、文本文件输入	
步骤名称	文本文件输入
文件 内容 错误处理 过滤 字段 其他输出字段	
文件类型	CSV
分隔符	Insert TAB
文本限定符	•
逃逸字符	
头部	☑ 头部行数量 1
尾部	■ 尾部行数量 1
包装行?	🔲 以时间包装的行数 1
分页布局 (printout)?	■ 每页记录行数 80
	文档头部行 0
压缩	None
没有空行	
在输出包括字段名?	■ 包含文件名的字段名称
輸出包含行数?	□ 行数字段名称
	按文件取行号 🗌
格式	DOS
编码方式	GB2312
Length	Characters 🔹
记录数量限制	0
解析日期时候是否严格要求?	
本地日期格式	zh_CN 🔫
6 结果文件名	
添加文件名	5 🗸
L	
	确定(O) 预览记录 取消(C)

图7-9设置"内容"选项卡



在"字段"选项卡中(如图7-10所示),点击"获取字段"按钮,会弹出 如图7-11所示的样本数据行数设置界面,直接点击"确定"按钮,会得到 如图7-12所示结果。这时,点击界面底部的"预览记录",就可以看到如 图7-13所示的数据。最后,点击界面底部的"确定"按钮,完成文本文件 输入控件的设置。

D, 文	本文件输入	\											
	步骤名称 文本文件輸入												
文件	内容(错)	吴处理 过滤 字段	其他輸出字段										
-	名称	类型	格式	位置	长度	精度	货币类型	小数	分组	Null if	默认	去除空字符串方式	重复
1													
					- \								
	×												
0	Help				确定(0)	预览记录	取消	i(C)				

图7-10 设置"字段"选项卡

厦门大学计算机科学与技术系

林子雨



📽 Sample data 🛛 💽
Number of lines to sample:
100
Show sample summary
确定(O) 取消(C)

图7-11 设置样本数据行数



B, 文													
	步骤名称 文本文件输入												
文件	内容(错误	設理 过 違 字段	其他輸出字段	e)									
ŵ	名称	类型	格式	位置	长度	精度	货币类型	小数	分组	Null if	默认	去除空字符串方式	重复
1	sno	Integer	#		15	0	¥		1.00	-		不去掉空格	否
2	name	String			5		¥		1.00	-		不去掉空格	否
3	sex	String			2		¥		1.00	-		不去掉空格	否
4	age	Integer	#		15	0	¥		1.00	-		不去掉空格	否
					_								
	获取字段 Minimal width												
L													
6	确定(O) 预览记录 取消(C)												
0	Help												

图7-12 获取字段以后的效果

《数据采集与预处理》

厦门大学计算机科学与技术系

林子雨



🍍 预览数据

步骤 文本文件输入 的数据 (6 rows)

-					
#	sno	name	sex	age	
1	1	王小明	男	24	
2	2	张璐	女	23	
3	3	<null></null>	<null></null>	<null></null>	
4	4	马琼	女	25	
5	5	<null></null>	<null></null>	<null></null>	
6	6	候杰	男	23	

图7-13 预览记录



双击设计区域的"Excel输出"控件图标,打开设置界面(如图7-14所示),在"文件"选项卡中,设置"文件名"为"D:\file"。

ì Excel输出	- • •
步骤名称 Excel输出	
文件 内容 格式 字段	
→ 文件名 D:\file	
创建父目录 🔲	
启动时不创建文件 📃	
扩展名 xls	•
在文件名里包含步骤数? 📃	
在文件名里包含日期? 📃	
在文件名里包含时间?	
指定时间格式	
时间相武	.
显示字段名称	
结果中添加文件名 🔽	
④ Help 确定(O) 取消(C)	

图7-14 设置文件名



在"字段"选项卡中(如图7-15所示),点击界面底部的"获取字段"按钮,成功获取字段以后的效果如图7-16所示,然后把"sno"和 "age"字段的"格式"设置为"#"。最后,点击"确定"按钮完成 "Excel输出"控件的设置。全部设置完成以后,需要保存设计文件。



图7-15 "字段"选项卡





图7-16 获取字段后的效果



4. 执行转换

在转换设计界面中(如图7-17所示),点击三角形按钮开始执行转换, 会弹出如图7-18所示界面,在界面中点击"启动",如果转换执行成功, 会显示如图7-19所示的效果,在两个控件图标上都会显示绿色的勾号。 这时,到D盘根目录下就可以看到新生成的文件file.xls,可以使用Excel 软件打开file.xls查看内容(如图7-20所示)。



图7-17 运行转换

《数据采集与预处理》



💥 执行转换								
Run configuration: <mark>Pentaho local</mark> _ 细节								
🔽 Clear log bef	ore running	日志级别	基本日志 🔻					
🔲 启用安全模式								
Gather perfo	rmance metrics							
命名参数 变量								
命名参数	默认值	值	Description					
位置参数								
Always show di	alog on run							
() Help			(启动(L) 取消(C)				

图7-18 转换启动界面





图7-19 转换执行成功的效果

《数据采集与预处理》

厦门大学计算机科学与技术系

林子雨



		模板专区	x∎ file.xls	; [兼容模式]	\times +
4	Α	В	С	D	E
1	sno	name	sex	age	
2	1	王小明	男	24	
3	2	张璐	女	23	
4	3				
5	4	马琼	女	25	
6	5				
7	6	侯杰	男	23	
8					

图7-20 file.xls文件内容

林子雨



7.5 数据清洗与转换

- 7.5.1 使用Kettle实现数据排序
- 7.5.2 在Kettle中用正则表达式清洗数据(请直接参考教材)
- 7.5.3 使用Kettle去除缺失值(请直接参考教材)
- 7.5.4 使用Kettle转化MySQL数据库中的数据(请直接参考教材)



- 这里给出一个实例,演示如何使用Kettle实 现数据排序,具体包括如下步骤:
- •创建文本文件;
- •建立转换;
- •设计转换;
- •执行转换。



1.创建文本文件

在"D:\"目录下新建一个文本文件score.txt,其内容如图7-51所示, 文件的第1行是字段名称,包括name和score,字段之间用分号隔开, 其余行都是记录,字段之间也是用分号隔开。

🧾 score.txt - 记事本					
文件(F)	编辑(E)	格式(O)			
name;s m 建 剑 东 燕 梅 形 、 の (で で が が に 朝 に 朝 で い の で の で の の で の の で の の の の の の の の	core ;87 8 ;92 ;77 ;89				

图7-51 score.txt文件内容



2.建立转换

在Spoon主界面的"主对象树"栏目中,在"转换"上面(如图7-52所示)单击鼠标右键,在弹出的菜单中点击"新建"。点击Spoon主界面 左上角的"保存"图标,把这个转换保存到某个路径下并且名称为 "sort_data"。

🦉 Spoon - 欢迎	ļ				
文件(F) 编辑	视图	执行	工具	帮助	
0 🗗 🗉	6	\otimes	•		
主对象树	ℓ 核	心对象			
Search		>	<	% 83	:
□ 转换					
Co Meyle					

图7-52 新建"转换"



3.设计转换

在"核心对象"栏目中,在"输入"控件里把"文本文件输入"拖到右侧设计区域,然后在"转换"控件里把"排序记录"拖到右侧设计区域,然后为这两个控件建立连线(如图7-53所示)。



图7-53 放置文本文件输入和排序记录两个控件



双击设计区域的"文本文件输入"控件图标,打开设置界面(如图7-54 所示),点击"文件或目录"右侧的"浏览"按钮,添加文件 "D:\score.txt",然后,点击"增加"按钮,执行效果如图7-55所示。

已, 文本文件输入						
步骤名称	文本	本文件输入				
文件 内容 错误处理 过渡 字段 其他输出字段						
文件或目录						
规则表达式						 ♦
正则表达式(排除)						•
选中的文件:	ŵ	文件/目录	通配符	通配符号(排除)	要求	包含子目录
	1					
						一川除
						编辑
从以前的步骤接受文件名						
从以前的步骤接受字段名	í 🗌					
步骤读取的文件名来自						Ŧ
在输入里的字段被当作文件名	í 📃					
				L. L		
	THEY.		或時行的内容			
	ſ	确定(O) 预览记录 取消(C)				
() Help						

图7-54 添加文件



□、文本文件输入							
步骤名称	文本	文本文件输入					
文件 内容 错误处理 过滤 字段 其他输出字段							
文件或目录							加 浏览(B)
规则表达式						•	
正则表达式(排除)						\$	
选中的文件:		文件/目录	通配符	通配符号(排除)	要求	包含子目录	
	1	\${Internal.Entry.Current.Directory}score.txt			否	否	110 RA
							编辑
1							
从门前的步骤接受文件名							
从以前的步骤接受字段名							
步骤读取的文件名来自							~
在输入里的字段被当作文件名							
				ר			
	T ALEX		四时口的内容]			
() Help	[确定(O) 预览记录 取消(C)					

图7-55 添加文件以后的效果



在"内容"选项卡中,设置分隔符为分号";"(如图7-56所示)。

2、文本文件输入	
步骤名称	文本文件输入
文件 内容 错误处理 过滤 字段 其他输出字段	
文件类型	
分隔符	S Insert TA
文本限定符	: •
逃逸字符	
头部	◎ 头部行数量 1
尾部	
包装行行	? 🔲 以时间包装的行数 1
分页布局 (printout)?	? 🔲 每页记录行数 80
	文档头部行 0
压缩	None ·
没有空行	
在输出包括字段名?	■ 包含文件名的字段名称
輸出包含行数?	行数字段名称
	按文件取行号 🗌
格式	DOS
编码方式	
Length	Characters
记录数量限制	0
解析日期时候是否严格要求?	
本地日期格式	t zh_CN
结果文件名 添加文件4	R √
? Нер	确定(O) 预览记录 取消(C)

图7-56 设置"内容"选项卡


在"字段"选项卡中(如图7-57所示),点击"获取字段"按钮,成功获 取字段以后的效果如图7-58所示。

Đ,														
	步骤名称 文本文件输入													
3	文件 内容 错误处理 过滤 字段 其他输出字段													
	ŵ.	名称	类型	格式	位置	长度	精度	货币类型	小数	分组	Null if	默认	去除空字符串方式	重复
	1													
							获取字段	Mini	mal width					
	∂ ⊦	lelp				(明元三)	.0)	Maka		(C)				

图7-57 获取字段

B, 文	本文件輸)	\												
文件	内容(错)	影神理 讨渡 字段	\ 其他輸出字目	e)	步骤名称	文本文件	输入							
	名称	类型	格式	位置	长度	精度	货币类型	小数	分组	Null if	默认	去除空字符串方式	重复	
1	name	String			3		¥			-		不去掉空格	否	
2	score	Integer	#		15	0	¥		1.00	-		不去掉空格	否	
0	Debit And													
				冬	7-5	58 	丧取字	的	成功	以后	f的	效果		

厦门大学计算机科学与技术系



这时,点击界面(如图7-58所示)底部的"预览记录"按钮,就可以预 览数据(如图7-59所示)。最后,点击界面底部的"确定"按钮,完成 "文本文件输入"控件的设置。

<mark>洋</mark> 预览数据											
步骤 文本文件输入 的数据 (6 rows)											
#	name	score									
1	陈建好	87									
2	郝剑	68									
3	沈东风	92									
4	韩燕	69									
5	张梅花	77									
6	林彤文	89									

图7-59 预览数据

《数据采集与预处理》

厦门大学计算机科学与技术系



双击设计区域的"排序记录"控件图标,打开设置界面(如图7-60所示),在"字段名称"下拉列表中选择"score",在"升序"下拉列表中选择"是",然后点击"确定"按钮完成设置。全部设置完成以后,需要保存设计文件。

□ 排序记录						
步骤名称	排序记录					
排序目录	%%java.io.tmpdir%%		� 浏览(B)			
临时文件前缀	out					
排序缓存大小(内存里存放的记录数)	1000000		\$			
未使用内存限值 (%)						
压缩临时文件?	\$		۵			
仅仅传递非重复的记录? (仅仅校验关键字)						
字段:						
# 字段名称 升序 大小写敏感	Sort based on current locale?	Collator Strength	Presorted?			
1 score 🗏 ▼						
1 Help	确定(O) 取消(C) 获取字	段				

图7-60 排序记录设置界面



4.执行转换

在转换设计界面中(如图7-61所示),点击三角形按钮开始执行转换,在 弹出的界面中点击"启动",如果转换执行成功,会显示如图7-62所示的 效果,在两个控件图标上都会显示绿色的勾号。这时,在"执行结果"的 "Preview data"选项卡中就可以预览排序后的数据(如图7-63所示)。





执行	执行结果										
	🗐 日志 🕗 执行历史 🔚 步骤度量 🛃 性能图 🖹 Metrics 👁 Preview data										
0											
÷	name	score									
1	郝剑	68									
2	韩燕	69									
3	张梅花	77									
4	陈建好	87									
5	林彤文	89									
6	沈东风	92									

图7-63 排序后的数据

厦门大学计算机科学与技术系





7.6.1把本地文件加载到HDFS中(请直接参考教材) 7.6.2把HDFS文件加载到MySQL数据库中



这里给出一个实例,演示如何使用Kettle把HDFS文件导入到MySQL数 据库中,具体包括如下步骤:

- •新建HDFS文件;
- •创建数据库;
- •建立转换;
- •创建MySQL连接和Hadoop连接;
- •设计转换;
- •执行转换。



1. 新建HDFS文件

在Windows系统中打开一个cmd窗口,启动Hadoop。在"D:\"目录下新建一个文本文件student.txt,其内容如图7-140所示,文件的第1行是字段名称,包括no、name、sex和age,字段之间用"|"隔开,其余行都是记录,字段之间也是用"|"隔开。



图7-140 student.txt文件内容



在cmd窗口中执行如下命令,把本地文件student.txt上传到HDFS系 统的根目录下:

- > cd c:\hadoop-3.1.3\bin
- > hadoop fs -put D:\book.txt hdfs://localhost:9000/
- 可以继续执行如下命令查看HDFS中student.txt的内容:
- > hadoop fs -cat hdfs://localhost:9000/student.txt
- 或者,也可以打开浏览器,访问"http://localhost:9870",使用 HDFS的WEB管理界面查看文件内容。



2.创建数据库

在Windows系统中启动MySQL服务,打开MySQL命令行客户端,执行 如下SQL语句创建数据库: CREATE DATABASE kettle: 继续执行如下SQL语句创建student_table表: USE kettle: #------创建表student_table DROP TABLE IF EXISTS student_table; CREATE TABLE student_table (no int, name VARCHAR(10), sex VARCHAR(2), age int);



3.建立转换

在Spoon主界面的"主对象树"栏目中,在"转换"上面(如图7-141所示)单击鼠标右键,在弹出的菜单中点击"新建"。点击Spoon主界面 左上角的"保存"图标,把这个转换保存到某个路径下并且名称为 "hdfs_to_mysql"。

- 灣 Spoon - 欢迎! 文件(D) / / / / / / / / / / / / / / / / / / /	
	_
 ○ 主对象树 ② 核心对象 	_
Search X L B=	
□ 转换	

图7-141 新建"转换"



3. 创建MySQL连接和Hadoop连接

参照本章7.4.2节的内容,建立一个名称为"mysql"的数据库连接(如图7-142所示)。

目 数据库连接			— ×
● 数据库连接 一般 高级 边坝 连接池 集計	注接名称: mysql 连接线型: Ingres VectorWise Ingres VectorWise Intersystems Cache KingbaseES LuidDB MS ACcess MS SQL Server MS SQL Server MS SQL Server MS SQL Server MS SQL Server MS SQL Server MariaDB MariaDB MariaDB MaxDB (SAP DB) MonetDB MonetDB MonetDB MonetDB MonetDB MonetDB MonetDB MonetDB MonetDB MonetDB MonetDB MariaDB	设置 主机名称: localhost 数编库名称: kettle 靖口号: 3306 用户名: root 密码: e●●●● ♥ Use Result Streaming Cursor	• • •
	连接方式: Native (JDBC) ODBC JNDI 测试 特征列表 浏览	減	取消

图7-142 建立数据库连接



参照本章7.6.1节的内容,建立一个名称为"Hadoop3"的Hadoop连接(如图7-143所示)。

	Edit	cluster	
	Cluster name		
	Hadoop3		
	Driver	Version	
	Hortonworks ~	3.0 ~	
	Site XML files		
	Browse to add file(s)		
	hdfs-site.xml	<u>remove</u>	
	core-site.xml	remove	
	HDFS		
	Hostname	Port	
	localhost	9000	-
	< [m	•
🕜 Help		Cance	Next

图7-143 建立Hadoop连接

厦门大学计算机科学与技术系



4. 设计转换

在Spoon主界面的"核心对象"的"Big Data"里面,找到"Hadoop file input"控件,放置到设计区域,在"核心对象"的"输出"里面,找到"表输出"控件,放置到设计区域,为两个控件建立连线(如图7-144所示)。



图7-144 放置"Hadoop file input"和"表输出"控件



在设计区域双击"Hadoop file input" 控件图标,打开设置界面(如图7-145所示),用鼠标点击"Environment"下面的空白单元格,会出现如图7-146所示的下拉列表,选中"Hadoop3"。

B, Ha	adoop file input					
			步骤	詺称	Hadoop file inp	ıt
文件	内容错误处理] 过滤 字段				
洗中	的文件:					×
-	Environment	File/Folder	通配符	要求	包含子目录	
1						
الل	上一步骤获取文件:	名			_	
		M.	以前的步骤接受 21前的步骤接受	文件名		
		ince ż	少骤读取的文件	:3来自		Ŧ
		在輸入目	里的字段被当作	文件名		
					显示又件名	显示又件內容」 显示来自第一杀数据行的内容
0	Help					确定(O) 预览记录 取消(C)

图7-145 "Hadoop file input" 设置界面





图7-146 设置Environment



点击"File/Folder"下面的空白单元格,会出现如图7-147所示的效果, 点击省略号按钮,会弹出如图7-148所示的界面,选中HDFS中的 student.txt文件,点击"OK"按钮,返回到"Hadoop file input"设置 界面。



图7-147 设置File/Folder



🖳 Open File 📉 💌
Open from Folder: hdfs://localhost:9000/ 👻 🎦 🗙 🕂
Name
4 🗀 /
Lenovo
administrator
Image: Provide the second s
input_kettle
Ienovo
b ist
Iser
⊳ 🚞 weblog
📄 student.txt
Filter: 所有文件 🔹
OK Cancel

图7-148 选中HDFS中的student.txt文件

厦门大学计算机科学与技术系



点击 "Hadoop file input" 设置界面的"内容"选项卡,会出现如图7-149所示的界面,"文件类型"选择"CSV",把"分隔符"设置为 "|","头部"后面的勾号选中,设置"头部行数量"为1。

🖳 Hadoop file input	
步骤名称	Hadoop file input
文件 内容 错误处理 过滤 字段	
文件类型	CSV +
→ 分隔符	Insert TAB
文本限定符	•
在文本限定符里允许换行	2
逃逸字符	
市 ション	☑ 头部行数量 1
尾部	同 尾部行数量 1
包装行	2 □ 以时间包装的行数 1
分贝布局 (printout)	2 ● 每页记录行数 80
	文档头部行 0
压頻	None 🔻
没有空行	
在輸出包括字段名?	■ 包含文件名的字段名称
輸出包含行数?	□ 行数字段名称
	按文件取行号 📃
格式	DOS
编码方式	•
记录数量限制	0
解析日期时候是否严格要求?	
本地日期格式	t zh_CN
──────────────────────────────────────	
添加文件:	S 🔽
() Help	确定(O) 预览记录 取消(C)

图7-149 设置分隔符



点击 "Hadoop file input" 设置界面的"字段"选项卡,会出现如图7-150 所示的界面,点击界面底部的"获取字段"按钮,会弹出如图7-151所示的界面,直接点击"确定"按钮返回字段设置界面,最后,点击该界面上的"确定"按钮。

E, Ha	R, Hadoop file input												
文件	步骤名称 Hadoop file input												
-	名称	类型	格式	位置	长度	精度	货币类型	小数	分组	Null if	默认	去除空字符串方式	重复
1	no	Integer	#		15	0	¥			-		不去掉空格	否
2	name	String			5		¥		1.	-		不去掉空格	否
3	sex	String			1		¥		1.	-		不去掉空格	否
4	age	Integer	#		15	0	¥			-		不去掉空格	否
0	4 age Integer # 15 0 ¥ . , - 小去掉空格 合												

图7-150 设置"字段"选项卡

厦门大学计算机科学与技术系



🐮 Nr of lines to sample. 0 mea 💌						
Number of sample lines (0=all lines)						
100						
确定(O) 取消(C)						

图7-151 设置取样行数

厦门大学计算机科学与技术系



双击设计区域的"表输出"控件图标,打开"表输出"设置界面(如图 7-152所示),在"数据库连接"右边的下拉列表中选择"mysql",点 击"目标表"右侧的"浏览"按钮,会弹出如图7-153所示界面,在界 面中选中"student_table"表,点击"确定"按钮,返回到"表输出" 设置界面,再点击"确定"按钮,完成设置。全部设置完成以后,需要

保存设计文件。

→ 表輸出	
步骤名称	表输出
数据库连接	mysql 新建) Wizard
目标模式	◆ 浏览(B)
目标表	
提交记录数量	1000
裁剪表	
忽略插入错误	
指定数据库字段	
主选项数据库字段	
表分 区 数据	
	¥
每天分区数据	
	Ŭ
使用批量插入	
表名定义在一个字段里?	
包含表名的字段:	
存储表名字段	√
返回一个日初广生的天键字 自动产生的关键之的方色之称	
	■
⑦ Help 确定(C)) 取消(C) SQL

图7-152 表输出设置界面



目 数据库浏览器		×
	Q۴	动作 🗸
🔺 🛅 mysql		*
⊿ 🗋 表		
table01		
📰 book		
orders 📰		
=== product		=
=== student_info		
📰 student_table 🦰		
user		
=== student		
student1		
sys_config		
📰 wlslog		-
<		•
确	定	取消

图7-153 选择student_table表

《数据采集与预处理》

厦门大学计算机科学与技术系



5. 执行转换

在转换设计界面中(如图7-154所示),点击三角形按钮开始执行转换, 在弹出的界面中点击"启动",如果转换执行成功,会显示如图7-155 所示的效果,在两个控件图标上都会显示绿色的勾号。



图7-154 转换设计界面



图7-155 执行成功的效果



这时,到MySQL数据库命令客户端执行如下SQL语句查看数据库中的数据: mysql> USE kettle; mysql> SELECT * FROM student_table;

上面SQL语句的执行结果如图7-156所示。

m	mysql> select * from student_table;										
+-					-+		+		-+		
H	no			name		se	× I	age			
+-					-+		+		-+		
H		1		Mike		Μ		21			
H		2		John		Μ		22			
H		3		Kate		\mathbf{F}		21			
H		4		Jenny		\mathbf{F}		21			
+-					-+		+		-+		
4	row	IS	i	ı set	(Ø	.00	sec	>			

图7-156 数据库查询执行结果





大数据应用系统的构建过程中,数据清洗是一个非常重要的环节。通过使用ETL工具,可以大幅提高数据清洗的效率,本案例采用开源工具Kettle实现数据的ETL操作。本章介绍了Kettle的基本概念、基本功能和安装方法,并通过实例演示了使用Kettle进行数据抽取、数据清洗与转换、数据加载的具体方法。本章介绍的内容属于比较基础的Kettle使用方法,如果要学习更加高级复杂的Kettle使用方法,可以参考相关书籍或网络资料。



附录A: 主讲教师林子雨简介



主讲教师:林子雨

单位: 厦门大学计算机科学与技术系 E-mail: ziyulin@xmu.edu.cn 个人网页: http://dblab.xmu.edu.cn/linziyu 数据库实验室网站: http://dblab.xmu.edu.cn



林子雨,男,1978年出生,博士(毕业于北京大学),全国高校知名大数据教师,现为厦门大学计算机 科学系副教授,厦门大学信息学院实验教学中心主任,曾任厦门大学信息科学与技术学院院长助理、晋江 市发展和改革局副局长。中国计算机学会数据库专业委员会委员,中国计算机学会信息系统专业委员会委 员。国内高校首个"数字教师"提出者和建设者,厦门大学数据库实验室负责人,厦门大学云计算与大数 据研究中心主要建设者和骨干成员,2013年度、2017年度和2020年度厦门大学教学类奖教金获得者,荣 获2019年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学 成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、 云计算和物联网,并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点 期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年 基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业 务费项目(No.2011121049), 主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作 育人项目,同时,作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程 区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个"数字教师"提出者和建 设者,2009年至今,"数字教师"大平台累计向网络免费发布超过1000万字高价值的研究和教学资料, 累计网络访问量超过1000万次。打造了中国高校大数据教学知名品牌,编著出版了中国高校第一本系统 介绍大数据知识的专业教材《大数据技术原理与应用》,并成为京东、当当网等网店畅销书籍;建设了国 内高校首个大数据课程公共服务平台,为教师教学和学生学习大数据课程提供全方位、一站式服务,年访 问量超过400万次,累计访问量超过1500万次。



附录B: 大数据学习路线图



大数据学习路线图访问地址: http://dblab.xmu.edu.cn/post/10164/

《数据采集与预处理》

厦门大学计算机科学与技术系



附录C:林子雨大数据系列教材



用于导论课、专业课、实训课、公共课

了解全部教材信息: http://dblab.xmu.edu.cn/post/bigdatabook/





开设全校公共选修课的优质教材



本课程旨在实现以下几个培养目标: □引导学生步入大数据时代,积极 投身大数据的变革浪潮之中 □了解大数据概念,培养大数据思 维,养成数据安全意识 □认识大数据伦理,努力使自己的 行为符合大数据伦理规范要求 □熟悉大数据应用,探寻大数据与 自己专业的应用结合点 □激发学生基于大数据的创新创业 热情

高等教育出版社 ISBN:978-7-04-053577-8 定价: 32元 版次: 2020年2月第1版 教材官网: http://dblab.xmu.edu.cn/post/bigdataintroduction/

《数据采集与预处理》

厦门大学计算机科学与技术系



附录E: 《大数据导论》教材

- 林子雨 编著《大数据导论》
- 人民邮电出版社,2020年9月第1版
- ISBN:978-7-115-54446-9 定价: 49.80元 教材官网: http://dblab.xmu.edu.cn/post/bigdata-introduction/



开设大数据专业导论课的优质教材



扫一扫访问教材官网

《数据采集与预处理》

厦门大学计算机科学与技术系

卜子雨

ziyulin@xmu.edu.cn



附录F: 《大数据技术原理与应用(第3版)》教材

《大数据技术原理与应用——概念、存储、处理、分析与应用(第3版)》,由厦门大学计算机科学系林子雨博士编著, 是国内高校第一本系统介绍大数据知识的专业教材。人民邮 电出版社 ISBN:978-7-115-54405-6 定价: 59.80元

全书共有17章,系统地论述了大数据的基本概念、大数据处 理架构Hadoop、分布式文件系统HDFS、分布式数据 库 HBase、NoSQL数据库、云数据库、分布式并行编程模型 MapReduce、Spark、流计算、Flink、图计算、数据可视化 以及大数据在互联网、生物医学和物流等各个领域的应用。 在Hadoop、HDFS、HBase、MapReduce、Spark和Flink 等重要章节,安排了入门级的实践操作,让读者更好地学习 和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的 大数据课程教材,也可供相关技术人员参考、学习、培训之 用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、 分析与应用》教材官方网站: http://dblab.xmu.edu.cn/post/bigdata3



扫一扫访问教材官网





本书是与《大数据技术原理与应用(第3版)》教材配套的唯一指定实验指导书



•步步引导,循序渐进,详尽的安装指南为顺利搭建大数据实验环境铺平道路
•深入浅出,去粗取精,丰富的代码实例帮助快速掌握大数据基础编程方法
•精心设计,巧妙融合,八套大数据实验题目促进理论与编程知识的消化和吸收
•结合理论,联系实际,大数据课程综合实验案例精彩呈现大数据分析全流程

林子雨编著《大数据基础编程、实验和案例教程(第2版)》 清华大学出版社 ISBN:978-7-302-55977-1 定价: 69元 2020年10月第2版



附录H:《Spark编程基础(Scala版)》



《Spark编程基础(Scala版)》

厦门大学林子雨, 赖永炫, 陶继平 编著

披荆斩棘,在大数据丛林中开辟学习捷径 填沟削坎,为快速学习Spark技术铺平道路 深入浅出,有效降低Spark技术学习门槛 资源全面,构建全方位一站式在线服务体系



人民邮电出版社出版发行, ISBN:978-7-115-48816-9 教材官网: http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言,系统介绍了Spark编程的基础知识。 全书共8章,内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、 Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作,以便读者更好地学习和掌 握Spark编程方法。本书官网免费提供了全套的在线教学资源,包括讲义PPT、习题、 源代码、软件、数据集、授课视频、上机实验指南等。



附录I: 《Spark编程基础(Python版)》



《Spark编程基础(Python版)》

厦门大学林子雨,郑海山,赖永炫 编著

披荆斩棘,在大数据丛林中开辟学习捷径 填沟削坎,为快速学习Spark技术铺平道路 深入浅出,有效降低Spark技术学习门槛 资源全面,构建全方位一站式在线服务体系



人民邮电出版社出版发行, ISBN:978-7-115-52439-3 教材官网: http://dblab.xmu.edu.cn/post/spark-python/

本书以Python作为开发Spark应用程序的编程语言,系统介绍了Spark编程的基础知识。全书共8章,内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作,以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源,包括讲义PPT、习题、源代码、软件、数据集、上机实验指南等。





http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/



1一扫观看3分钟FLASH动画宣传片


附录K: 高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求,厦门大学数据库实验室林子雨 老师团队联合企业共同开发了《高校大数据实训课程系列案例》,目前已经完成开 发的系列案例包括:

- 《电影推荐系统》(已经于2019年5月出版)
- 《电信用户行为分析》 (已经于2019年5月出版)
- 《实时日志流处理分析》
- 《微博用户情感分析》
- 《互联网广告预测分析》
- 《网站日志处理分析》

系列案例教材将于2019年陆续出版发行,教材相关信息,敬请关注网页后续更新! http://dblab.xmu.edu.cn/post/shixunkecheng/





扫一扫访问大数据实训课程系列案例教材主页

厦门大学计算机科学与技术系

林子雨





Department of Computer Science, Xiamen University, 2022