

# 《信息系统集成课程实习》

## 实习一报告

学 院: 遥感信息工程学院

班 级: 20F10

学 号: 2020302131249

姓 名: 马文卓

实习地点: 教学实验大楼 101 机房

指导教师: 王华敏

2022 年 12 月 6 日

# 目录

一、 实验概述 .....	3
1. 实验代号 .....	3
2. 实验目的 .....	3
二、 实验准备 .....	3
1. 实验要求 .....	3
2. 环境搭建 .....	3
(1) 系统环境 .....	3
(2) 服务器 .....	4
(3) MySQL .....	4
(4) Kettle .....	5
(5) MySQL Connector java .....	6
3. 数据准备 .....	6
三、 实验内容 .....	7
1. 创建数据库表 .....	7
2. 连接数据库 .....	8
3. 定义转换规则 .....	8
(1) ftp 下载 .....	9
(2) 清空数据表 .....	9
(3) 格式转换 .....	10
4. 执行转换过程 .....	12
5. 导出 sql 文件并检查 .....	13
四、 实验分析 .....	13
1. 方法分析 .....	13
2. 中文乱码 .....	14
五、 实验心得 .....	14

# 一、实验概述

## 1. 实验代号

本次实验的代号为 halibut。

## 2. 实验目的

本实验考查我们对 ETL 知识的理解情况。Kettle 是一款开源的、元数据驱动的 ETL 工具集，是开源 ETL 工具里功能比较强大的一个。本实验利用 Kettle 提供的开源免费版本，完成一个具体的数据 ETL 过程。通过该过程，我们可以理解数据的抽取、转换、装载流程，从而加深对 ETL 知识的理解。

# 二、实验准备

## 1. 实验要求

本次实验的为：使用 Kettle 中的 Spoon 组件对 ftp serve 中的 halibut.log 日志文件发起请求，并且通过在 Spoon 中定义的转换规则，将其加载到数据库表中，并导出相应文件。

最后需要提交两个文件：最终导出的 sql 文件 halibut.sql 和实验描述文档 halibut.doc。

- halibut.sql 可以导入数据库当中，并且其内容应该与原始的 halibut.log 一致
- 实验描述文档应当排版美观、内容详实、观点正确

## 2. 环境搭建

### (1) 系统环境

- 本次实验的操作系统为 Win10

Windows 规格	
版本	Windows 10 家庭中文版
版本号	21H2
安装日期	2021/6/5
操作系统内部版本	19044.2251
序列号	PF2D4K13
体验	Windows Feature Experience Pa

Figure 1: 系统环境

## (2) 服务器

本实验所使用的服务器信息如下。

- host: 119.36.242.188
- 协议: FTP
- 端口: 21
- user: MaWenZhuo
- password: mwz021248
- 连接工具: Xftp 7



Figure 2: 服务器连接信息

## (3) MySQL

- 数据库: MySQL
- 版本: 8.0.29
- 由于之前安装过 MySQL, 在此就不过多赘述安装过程。通过 `net start mysql` 启动服务, `mysql -h localhost -u root -p` 输入密码后进入 MySQL, 使用 `select version()` 查看版本信息如下:

```
C:\WINDOWS\system32>net start mysql
mysql 服务正在启动。
mysql 服务已经启动成功。

C:\WINDOWS\system32>mysql -h localhost -u root -p
Enter password: *****
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 8
Server version: 8.0.29 MySQL Community Server - GPL

Copyright (c) 2000, 2022, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>select version();
+-----+
| version() |
+-----+
| 8.0.29    |
+-----+
1 row in set (0.00 sec)
```

Figure 3: MySQL 版本信息

#### (4) Kettle

##### 1>搭建 java 环境

- 下载安装 JDK
- java version: 1.8.0\_201

```
C:\Users\mwz>java -version
java version "1.8.0_201"
Java(TM) SE Runtime Environment (build 1.8.0_201-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.201-b09, mixed mode)
```

Figure 4: Java 版本信息

- 配置环境变量

JAVA_HOME	D:\Users\mwz\bigdata\jdk
JRE_HOME	D:\Users\mwz\bigdata\jdk\jre

Figure 5: Java 环境变量

##### 2>下载安装 Kettle

- 版本: 7.1.0.0-12
- [官网](#)下载 Kettle7.1.0.0-12: 在下载时出现了下载速度非常缓慢的问题, 解决方法为换源


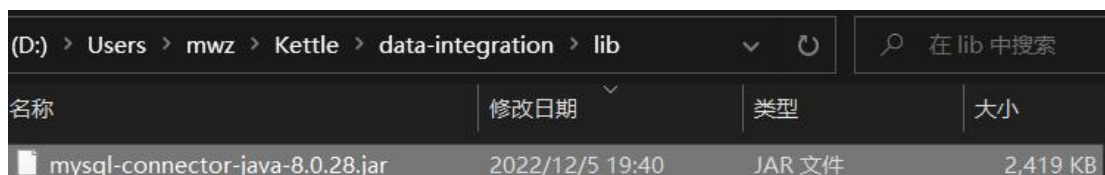
pdj-ce-7.1.0.0-12.zip	2017-05-17	903.9 MB	621		
-----------------------	------------	----------	-----	---	---

Figure 6: Kettle 版本信息

- 通过 Spoon.bat 打开, 安装完成

## (5) MySQL Connector java

- 驱动: MySQL Connector java
- 在[官网](#)MySQL8.0 对应的驱动 MySQL Connector java, 并将下好的 jar 文件放在 Kettle 安装目录的 lib 文件夹下。



(D:) > Users > mwz > Kettle > data-integration > lib				在 lib 中搜索
名称	修改日期	类型	大小	
mysql-connector-java-8.0.28.jar	2022/12/5 19:40	JAR 文件	2,419 KB	

Figure 7: 驱动位置

## 3. 数据准备

本次实验的原始数据为本人访问 fpt 数据库的日志文件 halibut.log(存放在 fpt serve 上)。其共包含 14 项字段, 54 行数据。具体字段含义如下图所示。



1	Tue Nov 16 15:41:44 2021	# 记录的时间
2	1	# 传输文件花费的时间 (秒)
3	113.57.80.253	# 客户端ip地址
4	437423	# 传输的文件大小 (byte)
5	/zhuguobin/Maze.zip	# 传输的文件
6	b	# 传输类型, b表示二进制传输, a表示 ascii码传输
7	_	# 特殊动作标记
8	i	# 传输方法, o表示从服务器下载, i表示 向服务器上传
9	g	# 访问模式, g表示虚拟用户, a表示匿名 用户
10	profzhu	# 用户名
11	ftp	# 服务名
12	0	# 授权方式
13	*	# 表示授权用户已认证IP
14	c	# 完成状态, c表示complete, i表示 incomplete

Figure 8: 字段含义表

### 三、实验内容

#### 1. 创建数据库表

- 命令行中 `net start mysql` 启动 MySQL 服务

```
C:\WINDOWS\system32>net start mysql
mysql 服务正在启动 .
mysql 服务已经启动成功。
```

Figure 9: 启动服务

- `mysql -h localhost -u root -p` 输入密码进入 MySQL

```
D:\本科\信息集成与管理\实习\mwz\halibut>mysql -h localhost -u root -p
Enter password: *****
```

Figure 10: 进入 mysql

- `create database kettle;` 创建名为 kettle 的数据库

```
mysql> create database kettle;
Query OK, 1 row affected (0.00 sec)
```

Figure 11: 创建数据库

- `create table log(...);` 创建包含如下 14 个字段的表

```
mysql> create table log(date varchar(40),timeconsuming int,ip varchar(40),filesize varchar(20),
-> filename varchar(100),transtyp varchar(2),specialopt varchar(2),transmethod varchar(2),
-> accesspattern varchar(2),username varchar(20),servicename varchar(10),authorize varchar(2),
-> identified varchar(2),state varchar(2));
Query OK, 0 rows affected (0.03 sec)
```

Figure 12: 创建表

- `desc log;` 检查表结构

```
mysql> desc log
-> ;
```

Field	Type	Null	Key	Default	Extra
date	varchar(40)	YES		NULL	
timeconsuming	int	YES		NULL	
ip	varchar(40)	YES		NULL	
filesize	varchar(20)	YES		NULL	
filename	varchar(100)	YES		NULL	
transtype	varchar(2)	YES		NULL	
specialopt	varchar(2)	YES		NULL	
transmethod	varchar(2)	YES		NULL	
accesspattern	varchar(2)	YES		NULL	
username	varchar(20)	YES		NULL	
servicename	varchar(10)	YES		NULL	
authorize	varchar(2)	YES		NULL	
identified	varchar(2)	YES		NULL	
state	varchar(2)	YES		NULL	

Figure 13: 表结构

## 2. 连接数据库

在 Spoon 中新建作业（命名为：实验 1），然后在其中新建 DB 连接。

值得注意的是，由于我使用的是 MySQL8.0，尝试发现这个版本的 MySQL 不能通过 mysql 类型直接连接，解决办法是：使用 Generic database 类型的连接，输入 URL 和驱动，即可连接到 kettle 数据库。

- URL: jdbc:mysql://localhost:3306/kettle?useUnicode=true&characterEncoding=UTF-8&useSSL=false&serverTimezone=Asia/Shanghai&zeroDateTimeBehavior=CONVERT\_TO\_NULL（其中的 kettle 为数据库名称）
- 驱动: com.mysql.cj.jdbc.Driver

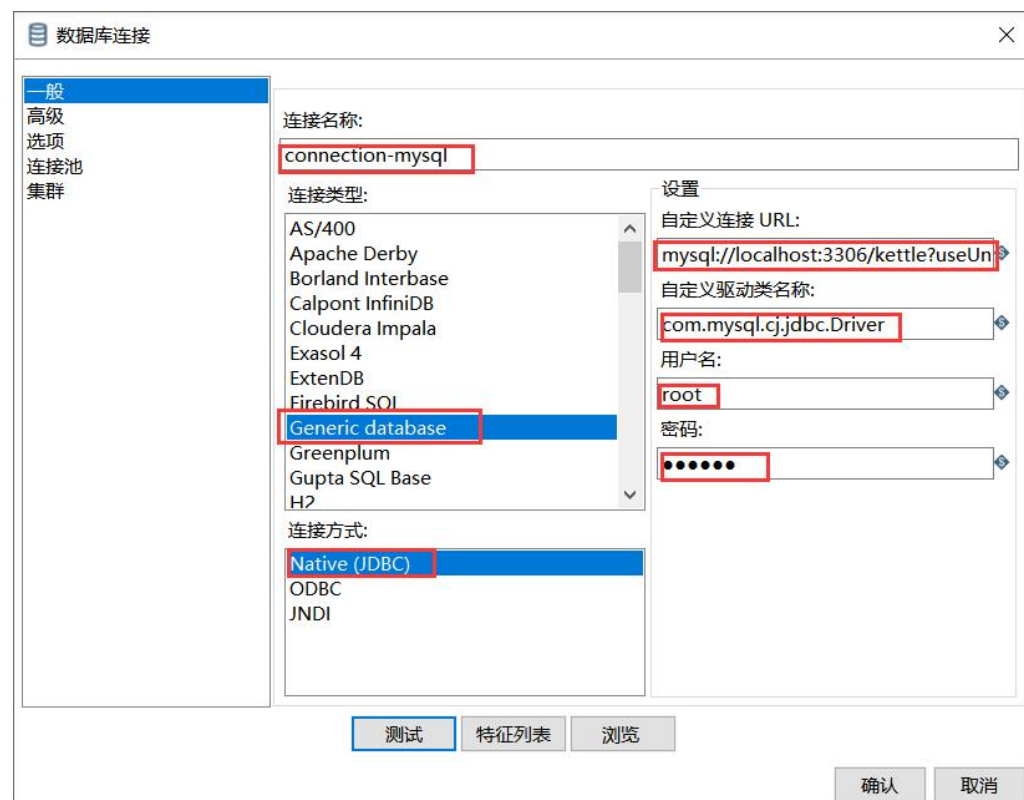


Figure 14: 连接数据库

测试连接，连接成功。



Figure 15: 连接测试结果

## 3. 定义转换规则

本步骤主要在 Spoon 中实现，总体的作业流程如下图所示，主要包含三个板块：FTP 下载、清空数据表、格式转换。（由于使用 Kettle7.0 版本，没有 SQL 文件导出组件，所以单独使用命令行导出，见第五小节）。





Figure 16: 作业流程图

## (1) ftp 下载

首先我们需要对 ftp serve 中的 halibut.log 文件发起 ftp 请求，将其下载以作为后续输入。具体的信息如下图所示。

- 其中服务器相关信息与上述服务器账户信息一致，测试连接成功。
- 源路径为根目录下的 halibut.log 文件，下载路径为本地桌面。

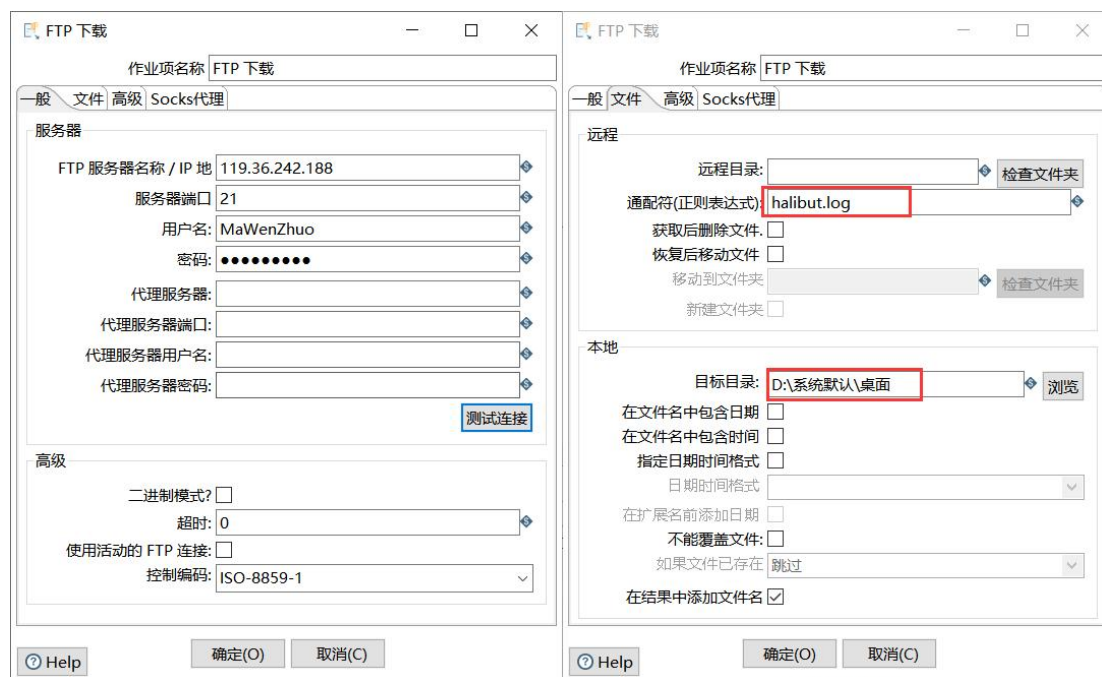


Figure 17: ftp 下载请求信息

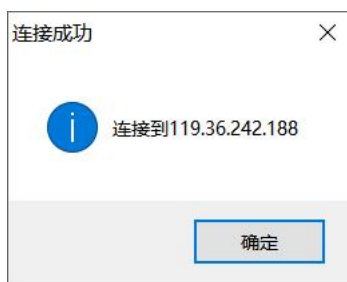


Figure 18: ftp 下载请求连接测试

## (2) 清空数据表

为了防止重复执行时数据库相应表中不为空，导致执行出错，因此在每次转

换、加载数据前都需要清空数据库当中的相应表格。

- SQL 语句: delete from log;
- 选择之前创建的数据库连接 connection，编辑上述 SQL 语句，即可完成对 kettle 数据库中 log 表的清空操作。

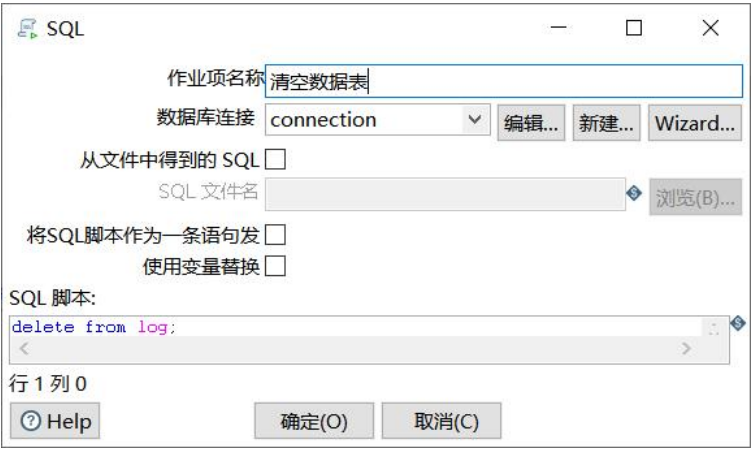


Figure 19: 清空数据表

### (3) 格式转换

这个部分主要是将 halibut.log 文件中的数据格式转换为与 MySQL 数据库 log 表相匹配的格式，以便于将其数据导入数据表中。其总体流程如下：



Figure 20: 格式转换

#### 1>文本文件输入

这个部分主要是将 halibut.log 作为文本文件，按照一定的内容格式输出为字段，以作为后续操作的输入。

主要处理思路：将 halibut.log 中的每一行数据分为两个字段（part1 和 part2）。其中 part1 为每行的前 24 个字符（日期时间），part2 为每行的第 25 个字符到结尾（除开日期时间的其他部分）。

这样处理的原因是因为日期时间中包含了空格，如果和其他数据一起处理，就无法使用空格作为分隔符了，加大了处理难度。所以这里我们将其分为两部分，把日期时间单独分离出来成为一个字段，其他部分作为另外一个字段。

当然，另外一种思路为：首先按照空格拆分数据为 18 个字段，然后通过脚本将前面五个字段合并成日期时间字段，最后输出为 14 个字段。

本次实验采用第一种思路，单独分离日期时间，具体信息如下。



Figure 21: 本文件输入（文件）

这里需要注意，一定要把第一部分的位置和长度，以及第二部分的起始位置设置正确。

文件 内容 错误处理 过滤 字段 其他输出字段													
#	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	part1	String		0	24							不去掉空格	否
2	part2	String		25	255							不去掉空格	否

Figure 22: 本文件输入（字段）

2>拆分字段

根据上述思路，这个步骤需要做的就是将之前的 part2 以空格为分隔符拆分为 13 个字段，这样加上 part1 中的日期时间字段，正好一共 14 个字段和数据库表中的字段对应。具体拆分信息如下。

拆分字段

步骤名称 拆分字段

需要拆分的字段 part2

分隔符 空格

Enclosure

字段

#	新的字段	ID	移除ID?	类型	长度	精度	格式	分组符号	小数点符号	货币符号	Nullif	缺省	去除空格类型
1	timeconsuming		N	Integer									不去掉空格
2	ip		N	String	40								不去掉空格
3	filesize		N	String	20								不去掉空格
4	filename		N	String	100								不去掉空格
5	transtype		N	String	2								不去掉空格
6	specialopt		N	String	2								不去掉空格
7	transmethod		N	String	2								不去掉空格
8	accesspattern		N	String	2								不去掉空格
9	username		N	String	20								不去掉空格
1..	servicename		N	String	10								不去掉空格
1..	authorize		N	String	2								不去掉空格
1..	identified		N	String	2								不去掉空格
1..	state		N	String	2								不去掉空格

Figure 23: 拆分字段

3>插入更新

上述操作已经将 halibut.log 中的数据拆分为了 14 个字段，现在本步骤所需要做的就是数据库 log 表中查询相对应的字段，并进行更新。

- 首先按照之前同样的方法建立一个到 “kettle” 数据库的连接 connetion2
- 选择需要插入更新的目标表 log
- 给定提交记录的数量
- 在查询映射中，左边填写 log 中的字段，右边填写之前拆分的对应字段，中间的比较符为=
- 在下面的更新映射中，左边填写 log 中的字段，右边填写之前拆分的对应字段，建立对应的映射关系，以更新数据

具体信息见下图。

步骤名称 插入 / 更新

数据库连接 connection2

编辑... 新建... Wizard...

目标模式

浏览(B)...

目标表 log

浏览...

提交记录数量 54

不执行任何更新: ☐

Figure 24: 插入更新信息

用来查询的关键词:

#	表字段	比较符	流里的字段1	流里的字段2
1	date	=	part1	
2	timeconsuming	=	timeconsuming	
3	ip	=	ip	
4	filesize	=	filesize	
5	filename	=	filename	
6	transtype	=	transtype	

更新字段:

#	表字段	流字段	更新
1	timeconsuming	timeconsuming	Y
2	ip	ip	Y
3	filesize	filesize	Y
4	filename	filename	Y
5	transtype	transtype	Y
6	specialopt	specialopt	Y
7	transmethod	transmethod	Y
8	accesspattern	accesspattern	Y
9	username	username	Y
1..	servicename	servicename	Y
1..	authorize	authorize	Y
1..	identified	identified	Y
1..	state	state	Y
1..	date	part1	Y

Figure 25: 插入更新映射

4. 执行转换过程

上述操作完成之后，就可以执行作业了。点击 Spoon 执行按钮，执行实验 1，执行成功。

成功	任务执行完毕	成功	3 2022/12/06 1...
任务: 实验1	任务执行完毕	成功 完成	3 2022/12/06 1...

Figure 26: 执行状态

在 MySQL 中查看 log 表中的数据: select \* from log;共 54 行数据（无数数据丢失），作业成功。

count(\*)  
54  
1 row in set (0.01 sec)

mysql> select \* from log;

date	servicename	timeconsuming	ip	filesize	filename	transtype	specialopt	transmethod	accesspattern	u
Sat Dec 3 21:42:34 2022	allenZhuo   ftp	0	1   113.57.80.53	85	/tuna.txt	b	_	i	g	M
Wed Nov 30 17:16:46 2022	allenZhuo   ftp	0	1   113.57.80.53	437423	/textbook/refs/Maze.zip	b	_	o	g	M
Wed Nov 30 17:16:19 2022	allenZhuo   ftp	0	1   113.57.80.53	51184	/textbook/refs/type.zip	b	_	o	g	M
Wed Nov 30 09:10:22 2022	allenZhuo   ftp	0	27   113.57.80.53	54423256	/textbook/ppts/L9-步进式项目管理过程.pptx	b	_	o	g	M
Wed Nov 30 09:10:06 2022	allenZhuo   ftp	0	14   113.57.80.53	19773351	/textbook/ppts/LB-集成系统管理概述.pptx	b	_	o	g	M
Wed Nov 30 09:09:40 2022	allenZhuo   ftp	0	1   113.57.80.53	171370	/textbook/works/作业8-tuna.pdf	b	_	o	g	M
Wed Nov 30 09:09:11 2022	allenZhuo   ftp	0	1   113.57.80.53	272	/buffalofish.score.txt	b	_	o	g	M
Sun Nov 27 14:17:13 2022	allenZhuo   ftp	0	1   113.57.80.53	73	/buffalofish.txt	b	_	i	g	M
Wed Nov 23 10:24:13 2022	allenZhuo   ftp	0	1   113.57.80.64	159679	/textbook/works/作业7-buffalofish.pdf	b	_	o	g	M
Wed Nov 23 10:24:02 2022	allenZhuo   ftp	0	1   113.57.80.64	40	/textbook/works/swordfish_answer.txt	b	_	o	g	M
Mon Sep 26 22:51:39 2022	allenZhuo   ftp	0	10   113.57.80.52	27590088	/textbook/ppts/L0-关于本课程.pptx	b	_	o	g	M
Mon Sep 26 22:52:08 2022	allenZhuo   ftp	0	6   113.57.80.52	18469696	/textbook/ppts/L1-信息系统集成的概念.pptx	b	_	o	g	M
Mon Sep 26 22:52:19 2022	allenZhuo   ftp	0	1   113.57.80.52	134	/textbook/works/contact.txt	b	_	o	g	M
Mon Sep 26 22:52:11 2022	allenZhuo   ftp	0	1   113.57.80.52	134	/textbook/works/contact.txt	b	_	o	g	M
Wed Sep 28 17:33:35 2022	allenZhuo   ftp	0	1   113.57.80.89	94289	/textbook/works/作业1-shark.pdf	b	_	o	g	M
Wed Sep 28 17:47:21 2022	allenZhuo   ftp	0	9   113.57.80.89	26497693	/textbook/refs/RESTful_Web_APIs中文版.pdf	b	_	o	g	M
Wed Sep 28 17:56:58 2022	allenZhuo   ftp	0	5   113.57.80.89	12514015	/textbook/ppts/L2-信息系统集成体系框架.pptx	b	_	o	g	M

Figure 27: 结果检查



## 5. 导出 sql 文件并检查

由于使用的 Kettle7.1 版本, 没有 SQL 文件导出的工具, 因此使用命令行导出 halibut.sql 文件。

- 命令: `mysqldump -u root -p kettle log > halibut.sql`

```
D:\本科\信息集成与管理\实习\mwz\halibut>mysqldump -u root -p kettle log > halibut.sql
Enter password: *****
```

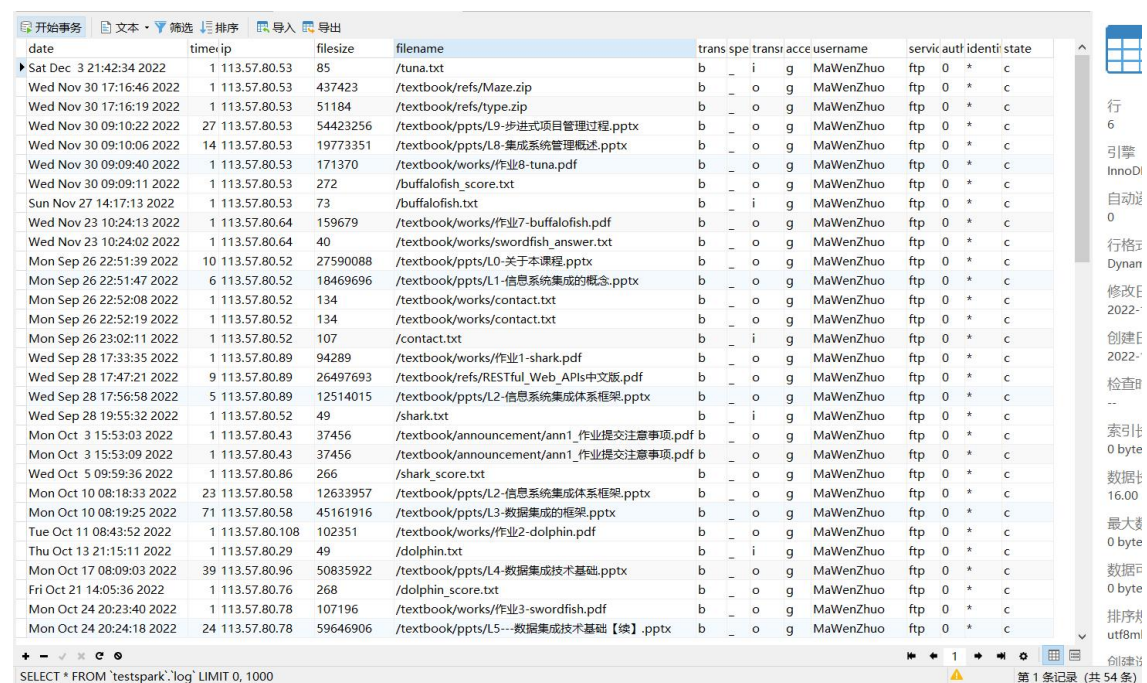
Figure 28: sql 文件导出

将导出的 sql 文件导入到新的数据库中检查是否内容和原始的 halibut.log 文件一致。我将其导入新建的 test 数据库当中, 检查后内容一致, 作业成功。

```
mysql> create database test;
Query OK, 1 row affected (0.00 sec)

mysql> use test;
Database changed
mysql> source halibut.sql;
```

Figure 29: sql 文件导入检查



date	time	ip	filesize	filename	trans	spe	trans	acce	username	servic	autl	identit	state
Sat Dec 3 21:42:34 2022	1	113.57.80.53	85	/tuna.txt	b	-	i	g	MaWenZhao	ftp	0	*	c
Wed Nov 30 17:16:46 2022	1	113.57.80.53	437423	/textbook/refs/Maze.zip	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Nov 30 17:16:19 2022	1	113.57.80.53	51184	/textbook/refs/type.zip	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Nov 30 09:10:22 2022	27	113.57.80.53	54423256	/textbook/ppts/L9-步进式项目管理过程.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Nov 30 09:10:06 2022	14	113.57.80.53	19773351	/textbook/ppts/L8-集成系统管理概述.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Nov 30 09:09:40 2022	1	113.57.80.53	171370	/textbook/works/作业8-tuna.pdf	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Nov 30 09:09:11 2022	1	113.57.80.53	272	/buffalofish_score.txt	b	-	o	g	MaWenZhao	ftp	0	*	c
Sun Nov 27 14:17:13 2022	1	113.57.80.53	73	/buffalofish.txt	b	-	i	g	MaWenZhao	ftp	0	*	c
Wed Nov 23 10:24:13 2022	1	113.57.80.64	159679	/textbook/works/作业7-buffalofish.pdf	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Nov 23 10:24:02 2022	1	113.57.80.64	40	/textbook/works/swordfish_answer.txt	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Sep 26 22:51:39 2022	10	113.57.80.52	27590088	/textbook/ppts/L0-关于本课程.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Sep 26 22:51:47 2022	6	113.57.80.52	18469696	/textbook/ppts/L1-信息系统集成的概念.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Sep 26 22:52:08 2022	1	113.57.80.52	134	/textbook/works/contact.txt	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Sep 26 22:52:19 2022	1	113.57.80.52	134	/textbook/works/contact.txt	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Sep 26 23:02:11 2022	1	113.57.80.52	107	/contact.txt	b	-	i	g	MaWenZhao	ftp	0	*	c
Wed Sep 28 17:33:35 2022	1	113.57.80.89	94289	/textbook/works/作业1-shark.pdf	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Sep 28 17:47:21 2022	9	113.57.80.89	26497693	/textbook/refs/RESTful_Web_APIs中文版.pdf	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Sep 28 17:56:58 2022	5	113.57.80.89	12514015	/textbook/ppts/L2-信息系统集成体系框架.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Sep 28 19:55:32 2022	1	113.57.80.52	49	/shark.txt	b	-	i	g	MaWenZhao	ftp	0	*	c
Mon Oct 3 15:53:03 2022	1	113.57.80.43	37456	/textbook/announcement/ann1_作业提交注意事项.pdf	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Oct 3 15:53:09 2022	1	113.57.80.43	37456	/textbook/announcement/ann1_作业提交注意事项.pdf	b	-	o	g	MaWenZhao	ftp	0	*	c
Wed Oct 5 09:59:36 2022	1	113.57.80.86	266	/shark_score.txt	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Oct 10 08:18:33 2022	23	113.57.80.58	12633957	/textbook/ppts/L2-信息系统集成体系框架.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Oct 10 08:19:25 2022	71	113.57.80.58	45161916	/textbook/ppts/L3-数据集成框架.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c
Tue Oct 11 08:43:52 2022	1	113.57.80.108	102351	/textbook/works/作业2-dolphin.pdf	b	-	o	g	MaWenZhao	ftp	0	*	c
Thu Oct 13 21:15:11 2022	1	113.57.80.29	49	/dolphin.txt	b	-	i	g	MaWenZhao	ftp	0	*	c
Mon Oct 17 08:09:03 2022	39	113.57.80.96	50835922	/textbook/ppts/L4-数据集成技术基础.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c
Fri Oct 21 14:05:36 2022	1	113.57.80.76	268	/dolphin_score.txt	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Oct 24 20:23:40 2022	1	113.57.80.78	107196	/textbook/works/作业3-swordfish.pdf	b	-	o	g	MaWenZhao	ftp	0	*	c
Mon Oct 24 20:24:18 2022	24	113.57.80.78	59646906	/textbook/ppts/L5-数据集成技术基础【续】.pptx	b	-	o	g	MaWenZhao	ftp	0	*	c

Figure 30: 新数据库中的 sql 文件

## 四、实验分析

### 1. 方法分析

上文提到了格式转换的两种方法:

- 在文本输入时单独分理出日期时间, 后续只要对其他部分进行拆分操作

- 在拆分字段时按照空格作为分隔符进行拆分，得到 18 个字段，然后加入 js 脚本进行前 5 个字段的合并，以达到最终的 14 个字段的输出

两种方法各有利弊，可以分情况选择。

第一种方法优点是不需要额外的增加脚本，在文本输入当中直接分为两个部分，以达到对日期时间的特殊处理；缺点为如果第一部分的长度不确定，就不能根据长度来确定第一个部分的内容，也就不能使用这种方法了。

第二种方法的优点是无论需要特殊处理的部分长度确不确定，都可以先拆分后通过脚本合并的方式进行处理。缺点是需要增加额外的处理步骤。

## 2. 中文乱码

在实验中会遇到数据中的中文显示为乱码的情况，这是由于编码格式的问题导致的。需要注意的是在文本输出字段的时候将编码格式设置为 utf-8 即可解决。

在检查 sql 文件是否正确时，我还遇到过由于编码方式与 mysql 不符导致无法导入 mysql 的问题，解决方式是在导出文件时就限定编码方式，导入时按照相对应的编码方式进行导入即可。

```
C:\Users\mwz>mysqldump -u root -p kettle log --default-character-set=utf8 > halibut.sql
```

```
mysql> source halibut.sql --default-character-set=utf8;
```

Figure 31: 改变编码方式

## 五、实验心得

通过本次实验我对 ETL 工具 Kettle 及其组件 Spoon 有了更加深入的了解认识和掌握熟练度。虽然 ETL 的定义 Extract、Transform、Load 看起来比较抽象，其实总结而来【ETL 就是一个从数据源获取原始数据，并且定义转换规则将数据转换为符合存储数据规范的格式，最后将转换好的数据加载到存储的位置。】事实上，ETL 就是做了一个桥梁，连接了数据源和存储数据的位置，将数据按照我们想要的格式从数据源引流过来，供我们使用。

通过实际的操作，也会发现很多理论课中无法遇见的问题，例如中文乱码等等。也是解决这些问题的过程促进了动手能力的提高，我想这也是实验课程的核心目标之一。

本次实验我学到最有价值的处理问题的思想是：分而治之。即遇到需要特殊处理的数据（例如本次实验中的日期时间），我们可以考虑将其分离出来单独处理它。

最后，谢谢老师的悉心教导，本次实验收获颇丰。

