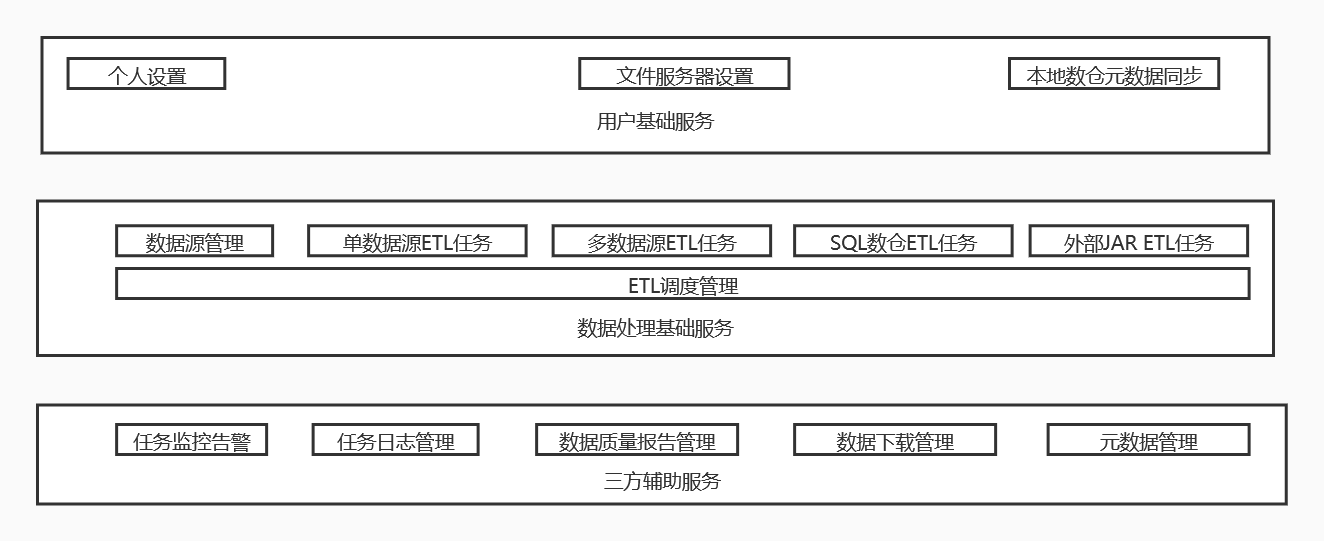
## 大数据采集平台

### 平台特点介绍

大数据采集平台是一个集调度,采集,处理,管理,并提供数据配置化的一个通用平台,在满足通用,易用的基础上并提供大数据量的高性能采集平台,此平台有以下提点降低大数据采集的入门门槛,易维护,高性能采集,数据可视化

### 平台功能图



### 平台数据功能

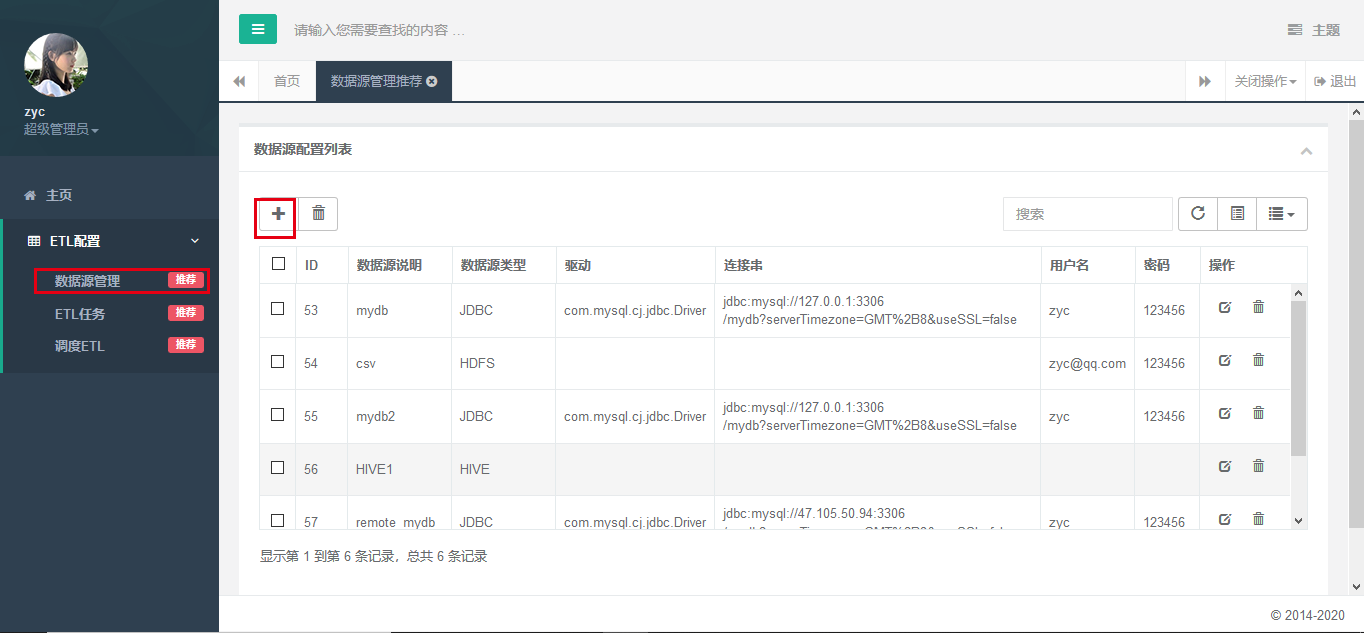
zdh 主要的作用 是从hdfs,hive,jdbc,http-json接口 等数据源拉取数据,并转存到hdfs,hive,jdbc等其他数据源 支持集群式部署

* 支持sql标准函数
* 支持界面选择配置
* 支持快速复制已有任务
* 支持外部调度工具(需要修改,新增特定接口)
* 弹性扩展(可单机,可集群)
* 支持客户级权限
* 简单易用支持二次开发
* 自带简单调度工具,可配置定时任务,时间序列任务,设定次数
* 调度依赖
* SQL数据仓库数据处理(单一数仓)
* 质量检测,及对应报告
* 支持SHELL 命令,SHELL 脚本,JDBC查询调度,HDFS查询调度
* 支持本地上传,下载文件
* 支持多源ETL
* 任务监控

### 数据源

#### 1 添加数据源

选择数据源管理-> 点击添加按钮



#### 2 配置数据源

##### 参数

数据源说明：简要说明数据源作用，此说明最好不要重复

数据源类型：选择数据源类型 目前支持 JDBC,HDFS,HBASE,HIVE,KAFKA

驱动类：当数据源选择JDBC 时 配置驱动程序

连接串：数据源选择JDBC,HDFS,HBASE,KAFKA 时填写

JDBC: 填写数据库连接的url

example:jdbc:mysql://127.0.0.1:3306/mydb?serverTimezone=GMT%2B8&useSSL=false

HDFS：填写hadoop 的连接url ，hdfs://ip:port

HBASE：填写hbase 对应zookeeper 的连接

KAFKA：填写对应zookeeper 的连接

MONGODB:填写对应的mongo连接串mongodb://[username:password@]host1[:port1][,host2[:port2],...[,hostN[:portN]]][/[database][?options]]

ES:

HTTP: 填写http://ip:port/[根路径]

REDIS: 填写ip:port 格式，如果是redis cluster 模式 填写其中一个ip 和port 即可

CASSANDRA:填写ip:port 格式，如果多个ip 则写ip1,ip2,ip3:port 格式，待优化

SFTP:填写ip:port 格式

KUDU:填写ip:port 格式，如果多个ip 则写ip1,ip2,ip3:port 格式

外部上传:需要在用户设置中设置sftp 登陆的用户名密码

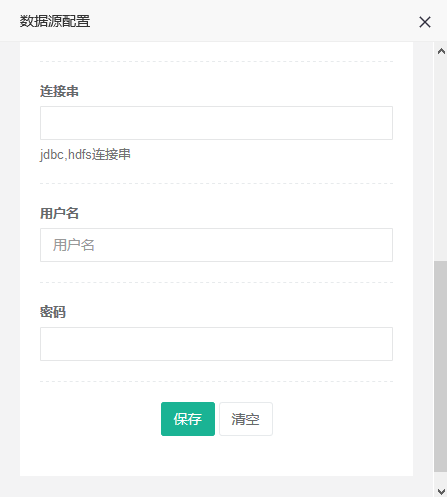
FLUME:填写ip:port 格式，如果多个ip 则写ip1,ip2,ip3:port 格式

外部下载:需要在用户设置中设置sftp 登陆的用户名密码

用户名：JDBC 数据源 用户名

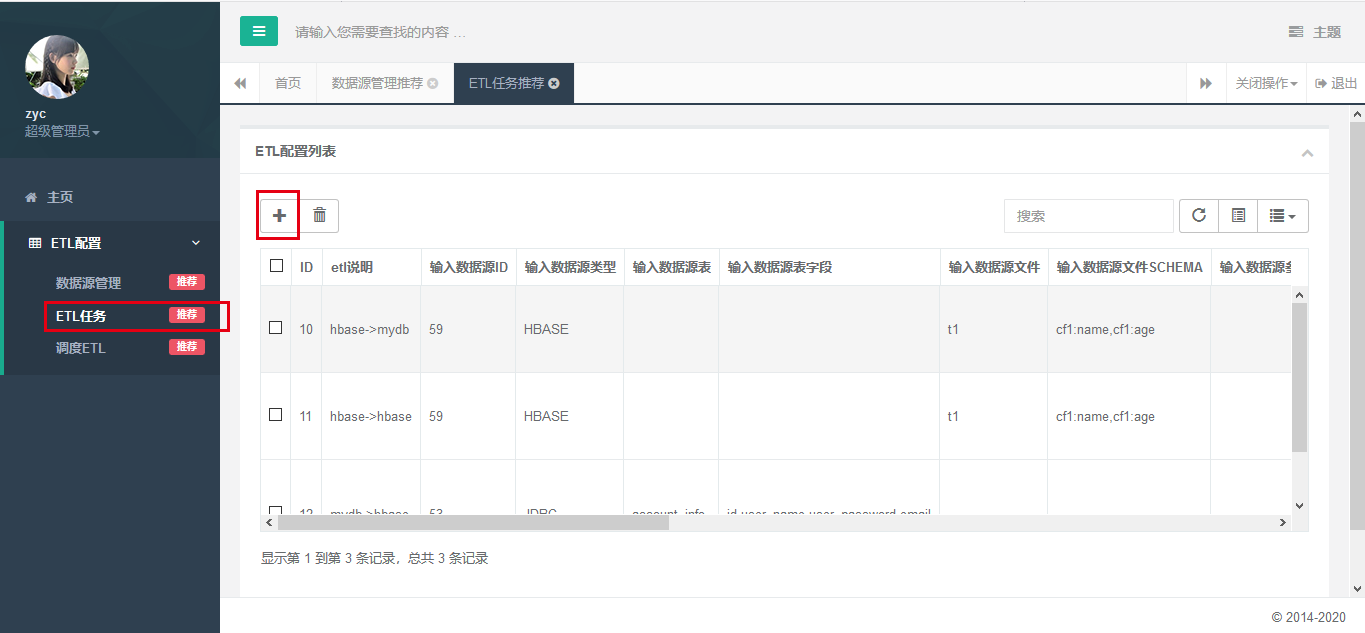
密码：JDBC数据源 密码





### ETL任务

#### 1 添加任务



#### 2 配置任务

##### 输入参数

1 ETL说明：ETL 任务的的一个简要说明，比如说从数据库mydb:account到mydb:accoun\_tmp

2 输入数据源：选择要读取的数据源

3 数据源类型：只读模式，选择数据源后会自动填充

4 其他参数：表示对输入数据源的一些配置，具体如下：

* JDBC: 暂时没有参数
* HDFS:
* sep:分割符,特殊字符需要转义,example:{"sep":"\\|"},{"sep","="}
* fileType:文件类型,可选择parquet,orc,csv,json...,example {"fileType":"csv"}
* partitionBy:分区字段,example {"partitionBy":"ETL\_DATE"}
* model:文件写入模式 默认append,overwrite,append,errorifexists,ignore,example {"model":"overwrite"}
* encoding:文件编码 默认UTF-8,example {"encoding":"gbk"}
* merge: 合并小文件 手动填写分区个数example {"merge":"50"}
* HBASE:暂时没有参数
* HIVE:
* format:orc,parquet
* partitionBy:分区字段
* path:如果存储打算存储外部表 需要指定路径example 外部表t1 path:/dir1/dir2/t1
* merge: 合并小文件 手动填写分区个数example {"merge":"50"}
* KAFKA:
* msgType:消息类型,支持csv,json 2种格式
* sep:分割符 当消息类型选择csv 时使用
* groupId:消费者组，默认g1
* HTTP:
* sep:分割符，可为空
* fileType:可选csv,json 默认json
* REDIS:
* data\_type:写入数据类型，可选，string,hash,list,set,table
* key.column:当选择写入数据类型为table 时配置此选项可以生成特定规则key
* CASSANDRA:
* KUDU:
* primary\_key:创建表的主键字段,如果不输入此参数会创建一个默认字段zdh\_auto\_md5 作为主键
* replicas:副本个数
* SFTP:
* sep:分割符，可为空
* fileType:可选csv,json,orc 默认csv
* 通用参数:
* repartition\_num:加载数据后数据进行重洗分区数
* repartition\_cols:加载数据后数据进行重洗分区字段，多个字段使用逗号分隔

5 输入表名/文件名：当选择JDBC 数据源时，可以下拉选择对应的表，其他数据源都需要手动输入表名或者文件名,

KAFKA 数据源输入对应得topic

HTTP数据源输入要查询的路径名

REDIS数据源 输入要查询的表名称，key 或者key 表达式

CASSANDRA数据源 输入要查询的keyspace.table

KUDU 数据源 输入查询的表名

MONGODB 数据源 输入查询collection(可以认为是数据库中的表名)

HBASE 数据源 输入查询的表名

SFTP 数据源 输入相对SFTP服务器的绝对路径

6 过滤条件：表示对选择的数据过滤条件 直接写sql 表达式 example: name=zyc,如果式hbase数据源只需输入开始的rowkey 和结束的rowkey 通过逗号分割

7 配置选择的列：如果选的是JDBC 数据源 那么直接点击[获取] 按钮即可，否则只能手动输入列名，多个列名直接通过逗号分割（特殊情况，如果是hbase 数据源 列名格式必须是列族:列名，这种格式）

8 配置字段映射：配置好第7步之后点击获取按钮，会自动生成默认的字段映射

如果想对字段做操作，比如字段截取，数学运算 都是在此处输入的如下图

此处支持的函数运算都是基于标准sql 的 sql 的函数可以通过

<http://spark.apache.org/docs/2.4.4/api/sql/index.html> 网站获取



至此输入数据源参数已经配置完成

9 配置字段高级用法

(1)新增一列和跑批日期相关 只需 字段表达式写 $zdh\_etl\_date 即可,日期默认格式yyyy-MM-dd HH:mm:ss

(2) to\_date($zdh\_etl\_date) 结果是yyyy-MM-dd 格式

参考函数http://spark.apache.org/docs/2.4.4/api/sql/index.html

##### 输出参数

1 输出数据选择

2 数据源类型：只读，自动填充

3 其他参数：参考输入数据源其他参数

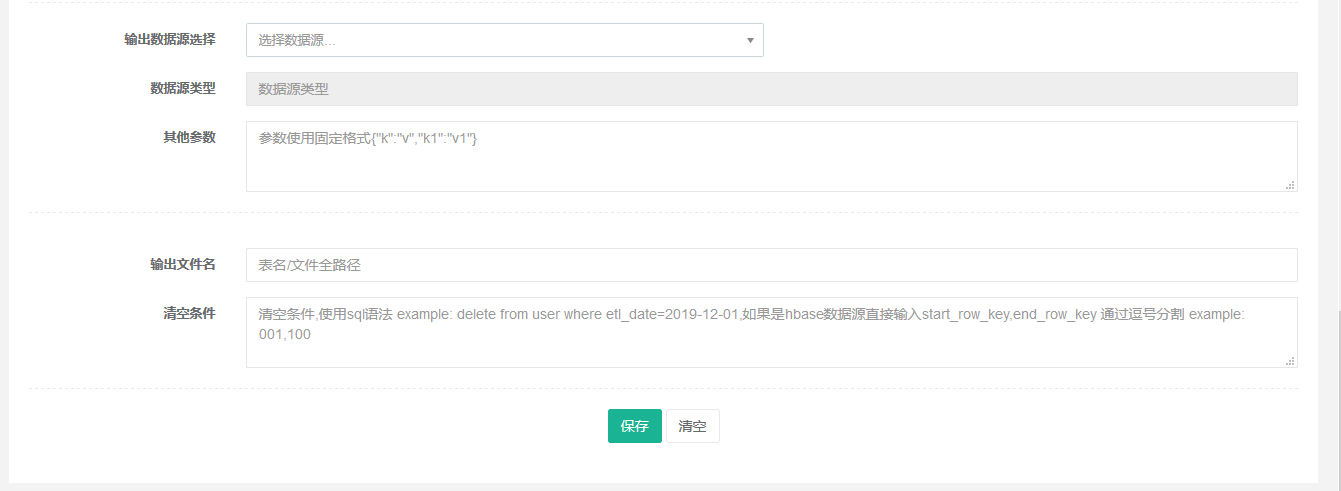
4 输出表名/文件名：当选择jdbc 数据源时可以从下拉框中选择，如果想创建新表，可以在文件输入框中手动输入

5 清空条件：当选择jdbc 数据源时 直接输入清空语句 example:delete from t1

当选择输出数据源时hbase 时 输入开始rowkey,结束rowkey 通过逗号分割







### 任务调度

#### 1 添加调度任务

1调度说明：简要概述此调度任务

2 ETL任务：选择创建好的ETL任务

3 任务类型：选择调度任务的类型，现支持SHELL，JDBC,FTP,CLASS 调度

4 表达式：表示触发任务的时间段，可以是特定时间间隔，10s,10m,10h ,也支持corn 表达式

5 开始时间，结束时间：表示任务执行时间，调度器会记录执行日期，每成功执行一次日期就加一天

6 步长:表示下次执行时间 在本次执行时间的距离

7 任务执行模式：

顺时间执行，日期限制，当执行到结束日期 则停止

执行一次

重复执行 ，只有次数限制，不做日期限制，任务运行的永远都是设置好的那天

8 执行任务次数：设置为-1 表示无限次

9 重置执行次数：当前执行的任务次数设置为0

10 任务命令：

SHELL 命令 ，最终结果输出 success 表示成功

JDBC 命令，select \* from t1 查询语句

11 其他参数配置：

参数支持日期，select \* from t1 where etl\_date=zdh.date ，加入当前任务调度日期是2020-02-29此语句会转换成

select \* from t1 where etl\_date=2020-02-29

日期参数zdh.date => yyyy-MM-dd 模式   
日期参数zdh.date.nodash=> yyyyMMdd 模式

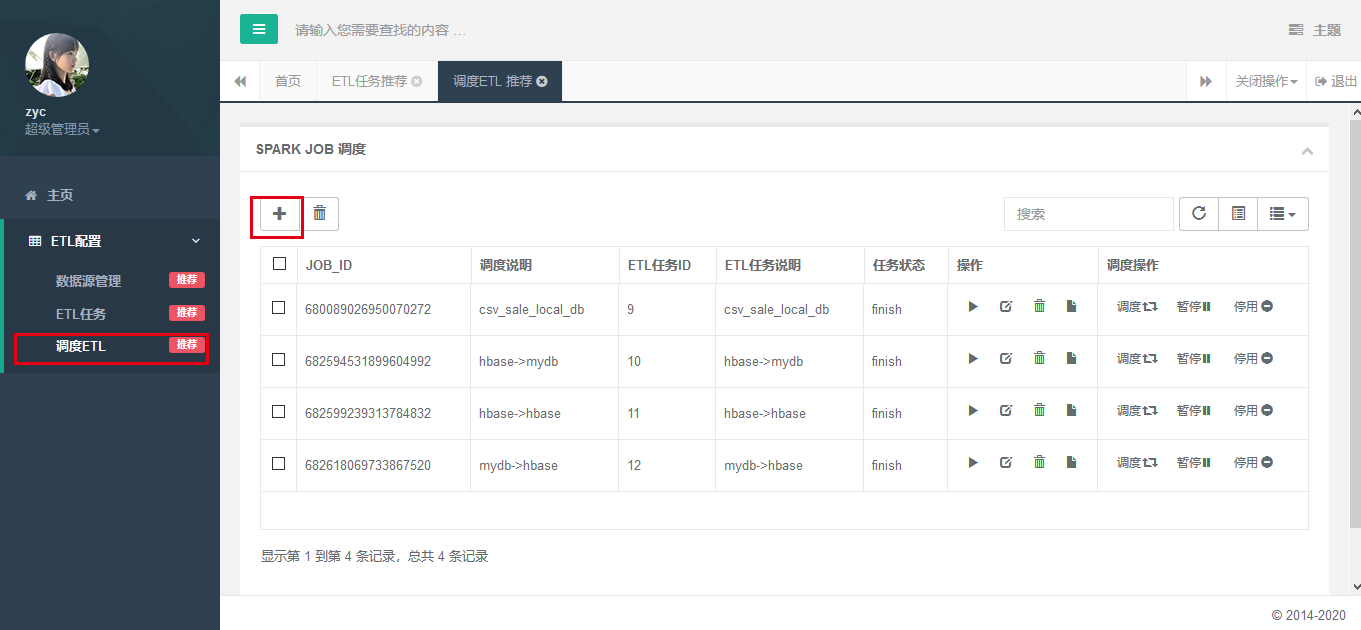
当使用JDBC 调度模式时需要配置基本连接串

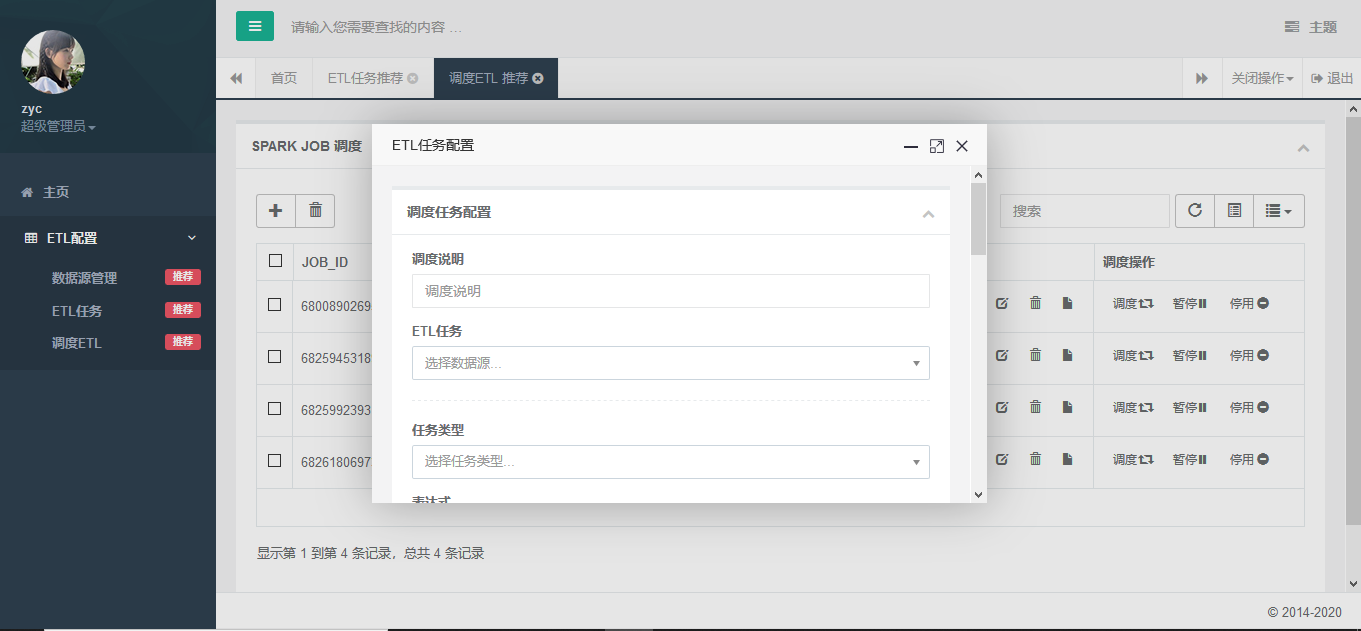
{"zdh.jdbc.url":"jdbc:mysql://127.0.0.1:3306/mydb?serverTimezone=GMT%2B8&useSSL=false",

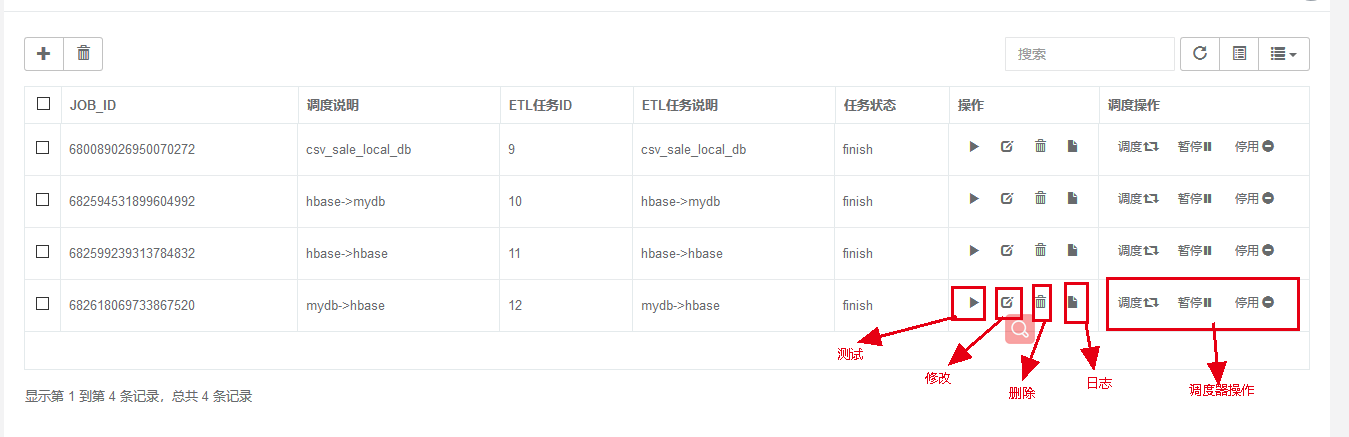
"zdh.jdbc.driver":"com.mysql.cj.jdbc.Driver",

"zdh.jdbc.username":"zyc",

"zdh.jdbc.password":"123456"}







### 未来与展望

大数据平台降低的入门的开发门槛,提高了数据管理，数据使用的效率

1方便快捷的增加数据的使用率和准确率

2监控模块使用运维人员更加方便

3数据使用易于管理

未来会从数据可视化,管理,统一等方向优化此平台,提升用户的使用体验