

集算器简易ETL工具

润乾软件出品



www.raqsoft.com.cn

目录 CONTENTS

1. 基本介绍
2. 数据源
3. 常见运算
4. 设置参数
5. 结果输出
6. 完整示例

集算器简易ETL工具



➤ 基本介绍

集算器ETL工具可以从多种数据源中抽取数据，进行过滤计算排序及关联运算等，再将结果导出到文本、数据库、Excel、集算器文件等。

简略来说，ETL过程就是三步：定义数据源→运算→结果写出。

ETL工具由数据源和视图关系两部分构成。前者处理数据来源；后者处理数据的关联与运算。

基本介绍



ETL工具的界面由菜单栏、工具栏、数据源、视图关系、选出列和别名、计算列等部分组成。

The screenshot shows the ETL tool interface with the following components:

- 菜单栏 (Menu Bar):** File, Edit, Tool, Window, Help
- 工具栏 (Toolbar):** Contains icons for file operations and execution.
- 数据源 (Data Sources):** A tree view on the left showing 'Data sources' with sub-items 'emp' and 'PERFORMANCE'.
- 视图关系 (View Relations):** A tree view at the bottom left showing 'View relations' with sub-items 'ExportDefine1' and 'ViewOut1'.
- 选出列和别名 (Selected Columns and Aliases):** A table titled 'The choices' with columns: Index, Alias, Select, Key, Reference name, emp, and PERFORMANCE. The 'EVALUATION' row is highlighted.
- 计算列 (Computed Columns):** A table titled 'Compute (Double click above reference name add to expression)' with columns: Index, Alias, and Expression.

Index	Alias	Select	Key	Reference name	emp	PERFORMANCE
1	EID	<input checked="" type="checkbox"/>	<input type="checkbox"/>	EID	EID	EMPLOYEEID
2	NAME	<input checked="" type="checkbox"/>	<input type="checkbox"/>	NAME	NAME	
3		<input type="checkbox"/>	<input type="checkbox"/>	SURNAME	SURNAME	
4		<input type="checkbox"/>	<input type="checkbox"/>	GENDER	GENDER	
5		<input type="checkbox"/>	<input type="checkbox"/>	STATE	STATE	
6		<input type="checkbox"/>	<input type="checkbox"/>	BIRTHDAY	BIRTHDAY	
7		<input type="checkbox"/>	<input type="checkbox"/>	HIREDATE	HIREDATE	
8		<input type="checkbox"/>	<input type="checkbox"/>	DEPTID	DEPTID	
9	SALARY	<input checked="" type="checkbox"/>	<input type="checkbox"/>	SALARY	SALARY	
10	EVALUATION	<input checked="" type="checkbox"/>	<input type="checkbox"/>	EVALUATION		EVALUATION
11		<input type="checkbox"/>	<input type="checkbox"/>	BONUS		BONUS

Index	Alias	Expression
-------	-------	------------

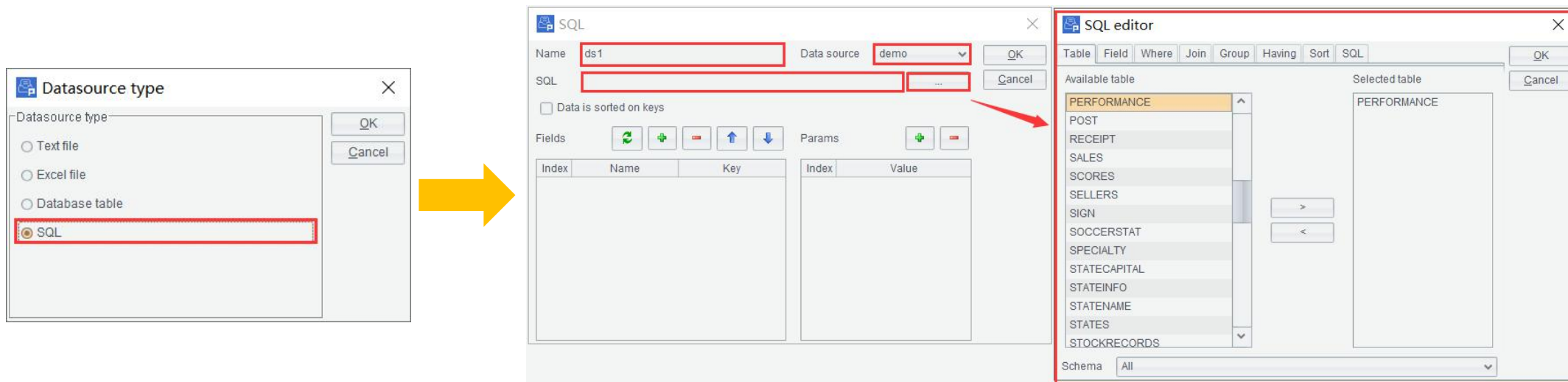
数据源

ETL工具支持各种数据源：文本文件、Excel文件、数据库表以及SQL。

以SQL数据源为例：

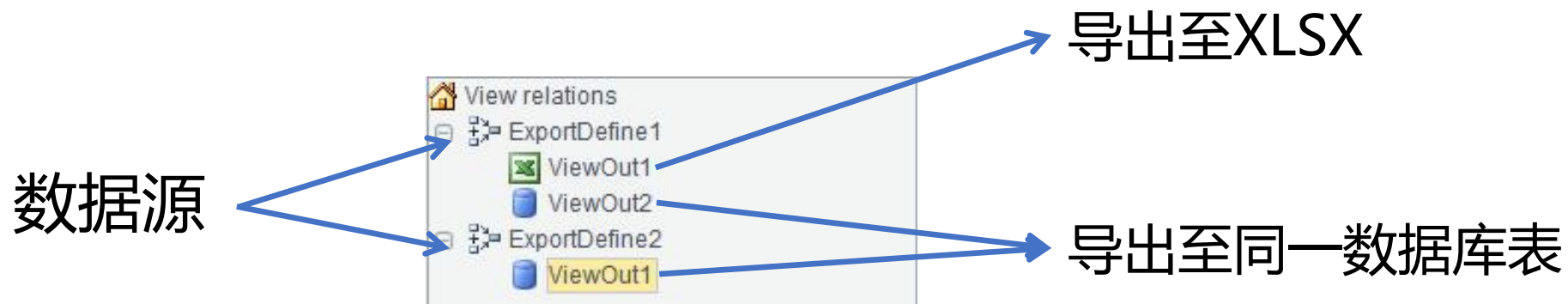
在数据源类型中选择SQL，编辑数据源名称ds1，选择数据源demo，并且输入SQL语句或者使用SQL编辑器编辑SQL语句。

编辑完成SQL语句后，点击刷新即可显示出对应表的字段。



数据源与视图关系

一套数据源可以创建多个视图关系，每个视图关系也可以创建多个视图结果，这些视图结果可以同时输出。若输出目标相同，即为实现UNION运算。



数据源与视图关系

例如现有两个结构相同但数据不同的表STU1与STU2，要将两表中的ID、NAME和GENDER导出至数据库表STUDENT中。

新建视图关系stu1，选出STU1表中的字段；再次新建视图关系stu2，选出STU2表中的字段，将两个视图关系导出至STUDENT表中。



➤ 常见运算：过滤

选择数据源后，可以在数据源界面添加过滤表达式，对数据进行过滤。

下图以文本文件为例，对emp.txt中的EID设置过滤条件：

Text file

Name: emp Options: t

File name: D:\emp.txt

Charset: Default Seperator: TAB

Filter expression: EID < 50

Data is sorted on keys

Index	Name	Key
1	EID	<input checked="" type="checkbox"/>
2	NAME	<input type="checkbox"/>
3	SURNAME	<input type="checkbox"/>
4	GENDER	<input type="checkbox"/>
5	STATE	<input type="checkbox"/>
6	BIRTHDAY	<input type="checkbox"/>
7	HIREDATE	<input type="checkbox"/>
8	DEPTID	<input type="checkbox"/>
9	SALARY	<input type="checkbox"/>



Preview data: emp

EID	NAME	SURNAME	GENDER	STATE	BIRTHDAY	HIREDATE	DEPT	SALARY
30	Joseph	Robinson	M	California	1973-11-23	2003-11-23	Sales	9000
31	Sarah	Miller	F	California	1977-04-27	2007-04-27	Marketing	10000
32	Andrew	Williams	M	Texas	1971-08-27	2001-08-27	Finance	11000
33	Matthew	Martinez	M	Pennsylva...	1980-07-19	2000-07-19	R&D	11000
34	Ryan	Johnson	M	Texas	1983-06-15	2003-06-15	Sales	5000
35	Justin	Smith	M	Texas	1978-08-20	2008-08-20	R&D	7000
36	Emily	Smith	F	Illinois	1983-12-27	2003-12-27	Production	12000
37	Hannah	Taylor	F	Pennsylva...	1984-07-20	2004-07-20	Marketing	5000
38	Matthew	Johnson	M	New York	1972-11-20	2002-11-20	R&D	6000
39	Andrew	Williams	M	California	1980-07-19	2000-07-19	Sales	3000
40	Madeline	Johnson	F	California	1971-12-27	2001-12-27	Sales	15000
41	Emily	Davis	F	Illinois	1984-07-07	2004-07-07	Production	9000
42	Michael	Jones	M	Pennsylva...	1978-08-20	2008-08-20	Administra...	12000
43	Joshua	Williams	M	California	1980-11-25	2000-11-25	Sales	12000
44	Emily	Jones	F	California	1977-12-24	2007-12-24	Sales	8000
45	Kayla	Miller	F	Florida	1984-08-25	2004-08-25	Production	3000
46	Alexander	Johnson	M	New York	1978-08-20	2008-08-20	Marketing	10000
47	Elizabeth	Brown	F	Pennsylva...	1971-08-27	2001-08-27	Marketing	16000
48	Emma	Smith	F	Florida	1976-11-25	2006-11-25	Sales	5000
49	Austin	Martinez	M	New York	1983-12-27	2003-12-27	Production	8000

常见运算：关联

在关系编辑中设定关联源表与目标表之间的关联字段，即可完成两表之间的关联；支持左连接和内连接两种可选；并且可以选择去掉目标字段的重复键值。

下图文本文件emp.txt和数据库表PERFORMANCE通过职员ID相关联，选择关联字段并选择内连接的连接方式，视图关系就创建好了。

The screenshot illustrates the process of creating a relationship between two data sources. On the left, the 'Relation setting' dialog box is shown with the following configuration:

- Source table: emp
- Target table: PERFORMANCE
- Join type: Inner join (selected)
- Remove duplicate key values: unchecked
- Join fields table:

Index	Source table field	Target table field
1	EID	EMPLOYEEID

A yellow arrow points from the dialog box to the right, where the 'Data sources' pane shows the 'emp' table and the 'SQL PERFORMANCE' table. Below this, the 'View relations' pane shows the relationship 'ExportDefine1'. To the right of the dialog box, a diagram shows the 'emp' table and the 'PERFORMANCE' table connected by a line between the 'EID' field of 'emp' and the 'EMPLOYEEID' field of 'PERFORMANCE'.

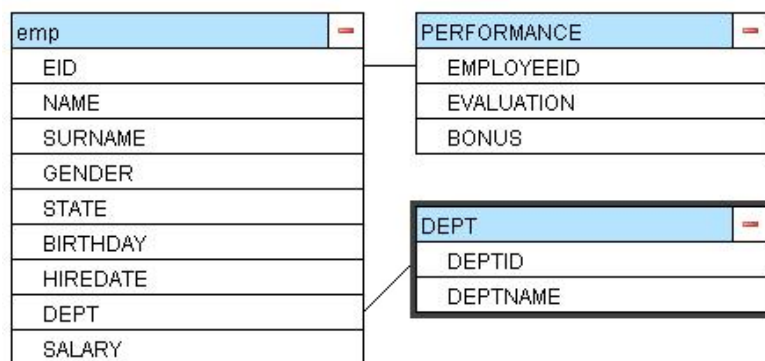
emp	PERFORMANCE
EID	EMPLOYEEID
NAME	EVALUATION
SURNAME	BONUS
GENDER	
STATE	
BIRTHDAY	
HIREDATE	
DEPT	
SALARY	



常见运算：关联

表关联还允许同一表与多表关联。当源表关联两个甚至两个以上子表时，在关系编辑中，选定相同的源表，不同的子表选为不同的目标表即可。

例如emp表可以同时关联PERFORMANCE表和DEPT表。



Relation setting

Source table: emp Target table: PERFORMANCE

Join type: Inner join Left join Remove duplicate key values

Join fields: + -

Index	Source table field	Target table field
1	EID	EMPLOYEEID

Relation setting

Source table: emp Target table: DEPT

Join type: Inner join Left join Remove duplicate key values

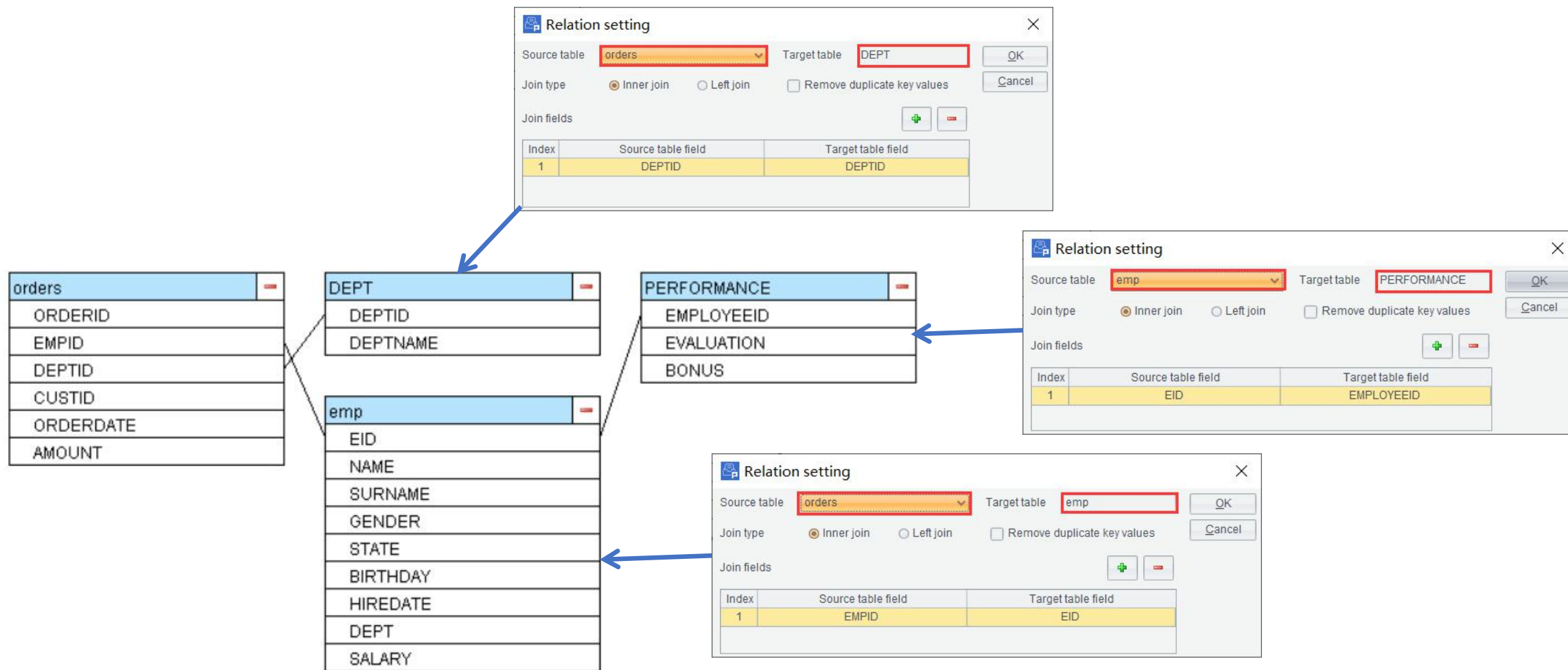
Join fields: + -

Index	Source table field	Target table field
1	DEPT	DEPTID



常见运算：关联

多表关联时要注意选对所关联的表和键；若需要修改关联的表，可以在表上右键选择编辑或删除。



➤ 常见运算：追加计算列

在视图结果中可以追加计算列、选出字段及调整字段次序。

例如在emp和PERFORMANCE的关联表中，选出所要导出的字段EID、GENDER、STATE和SALARY，并且增加一列ENAME，由NAME及SURNAME拼接而成。

Index	Alias	Select	Key	Reference name	emp	PERFORMANCE
1	EID	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	EID	EID	EMPLOYEEID
2		<input type="checkbox"/>	<input type="checkbox"/>	NAME	NAME	
3		<input type="checkbox"/>	<input type="checkbox"/>	SURNAME	SURNAME	
4	GENDER	<input checked="" type="checkbox"/>	<input type="checkbox"/>	GENDER	GENDER	
5	STATE	<input checked="" type="checkbox"/>	<input type="checkbox"/>	STATE	STATE	
6		<input type="checkbox"/>	<input type="checkbox"/>	BIRTHDAY	BIRTHDAY	
7		<input type="checkbox"/>	<input type="checkbox"/>	HIREDATE	HIREDATE	
8		<input type="checkbox"/>	<input type="checkbox"/>	DEPTID	DEPTID	
9	SALARY	<input checked="" type="checkbox"/>	<input type="checkbox"/>	SALARY	SALARY	
10		<input type="checkbox"/>	<input type="checkbox"/>	EVALUATION		EVALUATION
11		<input type="checkbox"/>	<input type="checkbox"/>	BONUS		BONUS

Compute (Double click above reference name add to expression)

Index	Alias	Expression
1	ENAME	NAME+'-'+SURNAME





Preview view					Close
EID	GENDER	STATE	SALARY	ENAME	
1	F	California	7000	Rebecca Moore	
2	F	New York	11000	Ashley Wilson	
3	F	New Mexico	9000	Rachel Johnson	
4	F	Texas	7000	Emily Smith	
5	F	Texas	16000	Ashley Smith	
6	M	California	11000	Matthew Johnson	
7	F	Illinois	9000	Alexis Smith	
8	F	California	11000	Megan Wilson	
9	F	Texas	3000	Victoria Davis	
10	M	Pennsylva...	13000	Ryan Johnson	
11	M	Texas	12000	Jacob Moore	
12	F	New York	7000	Jessica Davis	
13	M	Florida	10000	Daniel Davis	
14	F	Florida	4000	Alyssa Wilson	
15	F	New York	8000	Alexis Smith	
16	M	Florida	9000	Christopher Hernandez	
17	F	Texas	4000	Hannah Johnson	
18	M	Florida		Jonathan Moore	
19	F	Pennsylva...	10000	Samantha Williams	
20	F	Florida	16000	Alexis Allen	
21	M	Pennsylva...	10000	Jacob Moore	
22	M	Texas	16000	Jacob Davis	







常见运算：追加计算列

调整字段次序只需要选中需要调整的字段，点击右上角的上移或下移即可。选出列和别名以及计算列都可以对字段进行次序调整。

The choices

Select all Select none  

Index	Alias	Select	Key	Reference name	emp	PERFORMANCE
1	EID	<input checked="" type="checkbox"/>	<input type="checkbox"/>	EID	EID	EMPLOYEEID
2		<input type="checkbox"/>	<input type="checkbox"/>	NAME	NAME	
3		<input type="checkbox"/>	<input type="checkbox"/>	SURNAME	SURNAME	
4	GENDER	<input checked="" type="checkbox"/>	<input type="checkbox"/>	GENDER	GENDER	
5	STATE	<input checked="" type="checkbox"/>	<input type="checkbox"/>	STATE	STATE	
6		<input type="checkbox"/>	<input type="checkbox"/>	BIRTHDAY	BIRTHDAY	
7		<input type="checkbox"/>	<input type="checkbox"/>	HIREDATE	HIREDATE	
8		<input type="checkbox"/>	<input type="checkbox"/>	DEPT	DEPT	
9	SALARY	<input checked="" type="checkbox"/>	<input type="checkbox"/>	SALARY	SALARY	
10		<input type="checkbox"/>	<input type="checkbox"/>	EVALUATION		EVALUATION
11		<input type="checkbox"/>	<input type="checkbox"/>	BONUS		BONUS

Compute (Double click above reference name add to expression)    

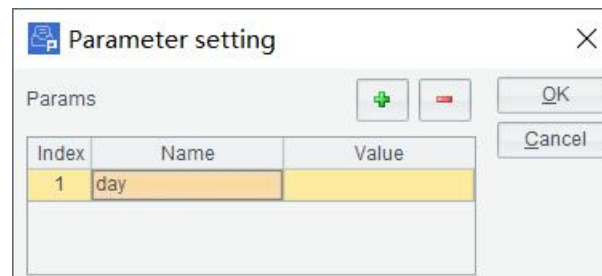
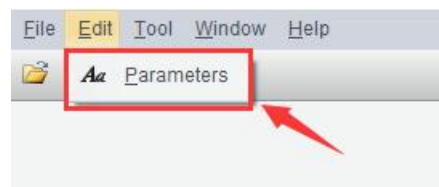
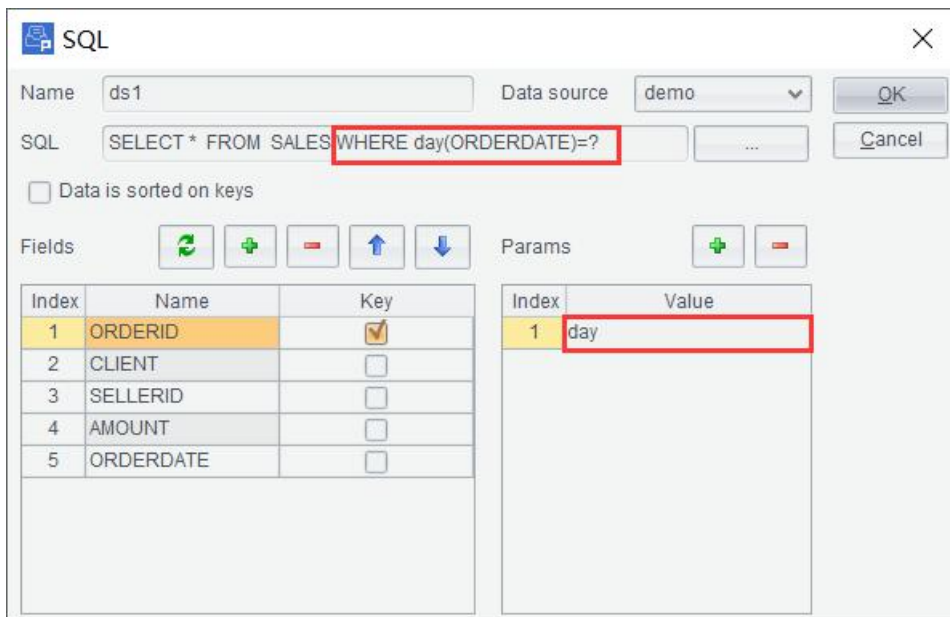
Index	Alias	Expression
1	ENAME	NAME+" "+SURNAME

设置参数

可以设置参数来取出精准或规定范围内的数据：

数据源的过滤条件中设置参数，然后在菜单栏选择编辑 → 参数中添加参数。

下图在SQL中查询SALES表时添加过滤条件 “WHERE day(ORDERDATE)=?” ，用来选出指定日期的记录，并设置参数为day；然后在ETL工具中添加参数day。



设置参数

参数设置完后，在执行操作时即可填入参数，并导出相应数据。
这里参数为21，则导出的数据为ORDERDATE日期为21的部分。

Title	Value
day	21



Index	ORDERID	CLIENT	SELLERID	AMOUNT	ORDERDATE
1	13	HL	12	21400.0	2012-11-21 15:28:05
2	16	AYWYN	4	6566.0	2012-11-21 15:28:05
3	20	EGU	8	11700.0	2012-11-21 15:28:05

结果输出：导出结果

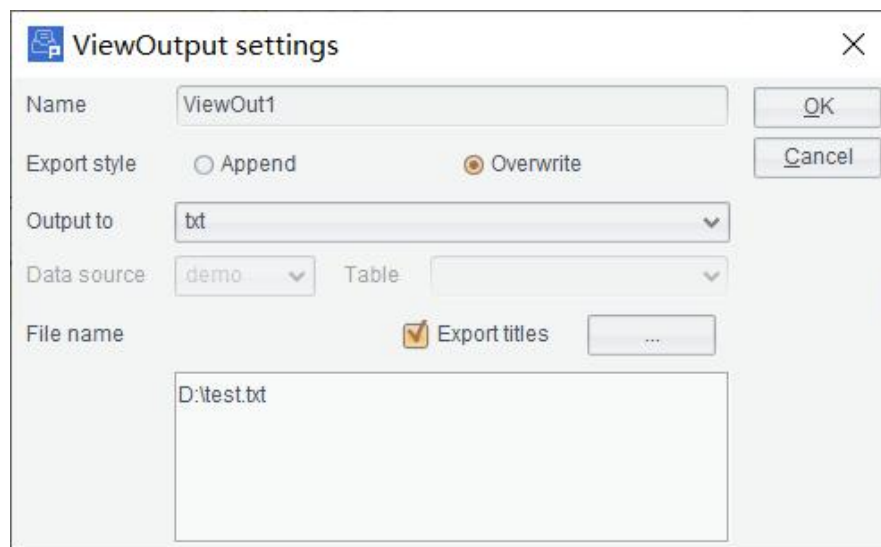
结果集可以导出至CTX、BTX、TXT、CSV、XLSX和数据库表中。

导出前可选择追加或覆盖，追加则会在原导出数据中追加，覆盖则会替换原导出数据。

在导出为TXT、CSV及XLSX时，可以选择是否导出标题。

若导出为数据库表，无论是追加或是覆盖，所要导出的表中列名都要与数据库表中的列名对应。

下图将结果集导出为txt格式，命名为test并且导出文件第一行为列名。



EID	NAME	SURNAME	GENDER
1	Rebecca	Moore	F
2	Ashley	Wilson	F
3	Rachel	Johnson	F
4	Emily	Smith	F
5	Ashley	Smith	F
6	Matthew	Johnson	M
7	Alexis	Smith	F
8	Megan	Wilson	F



结果输出：调度执行

将ETL过程存为.ept格式的文件，可以使用命令直接调度执行。执行语句语法如下：

```
Esprocx [eptFile] [argN=value] [ExportDefine=name] [ViewOut=name]...
```

[eptFile]: 相对于寻址路径或者主路径的ept文件名，也可以是绝对路径；

[argN=value]: eptFile有参数，参数按照“参数名=参数值”指定；

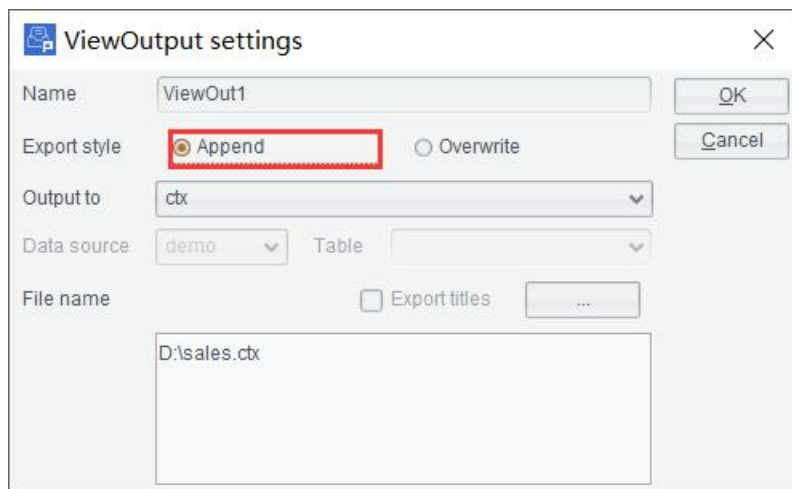
注意：分节串等号两边不能有空格；分节串之间用空格分隔！

[ExportDefine=name]: 执行eptFile中名为name的视图关系，缺省表示执行全部导出；

[ViewOut=name]: 执行eptFile中名为name的视图结果，缺省表示执行全部导出。

结果输出：调度执行

在视图结果配置中选择导出风格为追加；使用设置参数中所举的例子将.ept存在本地，并且执行命令语句：Esprocx SALES.ept day=21
21日的所有数据将会追加到sales.ctx中。



选择导出风格为追加

Index	ORDERID	CLIENT	SELLERID	AMOUNT	ORDERDATE
1	13	HL	12	21400.0	2012-11-21 15:28:05
2	16	AYWYN	4	6566.0	2012-11-21 15:28:05
3	20	EGU	8	11700.0	2012-11-21 15:28:05
4	38	DILRT	12	18300.0	2012-12-15 15:28:05
5	39	GLH	14	22200.0	2012-12-15 15:28:05
6	43	HANAR	5	196.0	2012-12-15 15:28:05

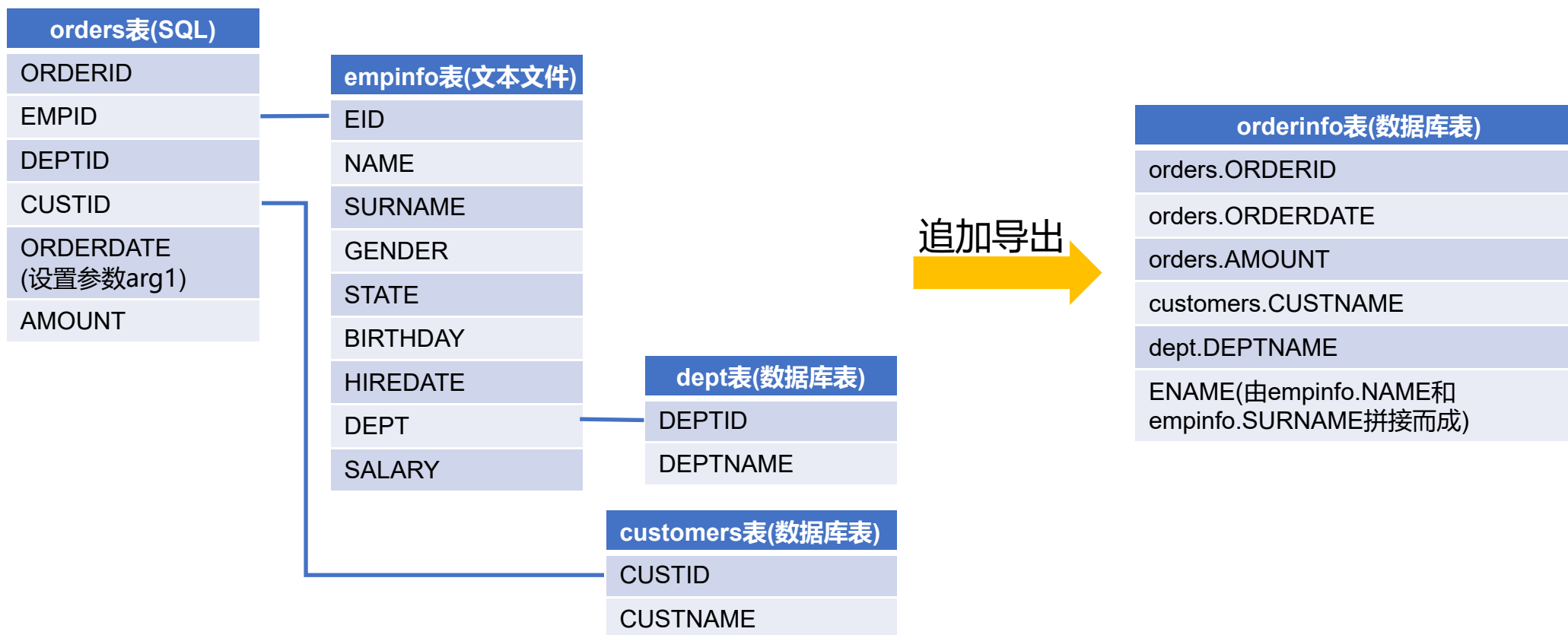
执行语句后的SALES.ctx



完整示例：描述示例

现有四个源表如下图，ETL任务为：将它们关联后导出至数据库表orderinfo中。

orders表中数据每天更新，使用ORDERDATE作为参数可以查询出当天数据，关联后追加进orderinfo表追加。ETL过程将被每天调度使用。



完整示例：添加文本类型数据源

首先添加数据源。

添加一个文本类型数据源，empinfo.txt中的数据如下，命名数据源为empinfo。

EID	NAME	SURNAME	GENDER	STATE	BIRTHDAY	HIREDATE	DEPT	SALARY
1001	Rebecca	Moore	F	California	1974-11-20	2005-03-11	1	7000
1002	Ashley	Wilson	F	New York	1980-07-19	2008-03-16	12	11000
1003	Rachel	Johnson	F	New Mexico	1970-12-17	2010-12-01	7	9000
1004	Emily	Smith	F	Texas	1985-03-07	2006-08-15	4	11000
1005	Ashley	Smith	F	Texas	1975-05-13	2004-07-30	8	8000
1006	Matthew	Johnson	M	California	1984-07-07	2005-07-07	5	11000
1007	Alexis	Smith	F	Illinois	1972-08-16	2002-08-16	8	9000
1008	Megan	Wilson	F	California	1979-04-19	1984-04-19	7	11000
1009	Victoria	Davis	F	Texas	1983-12-07	2009-12-07	3	3000
1010	Ryan	Johnson	M	Pennsylva...	1976-03-12	2006-03-12	9	13000

Text file dialog box configuration:

- Name: empinfo
- File name: D:\empinfo.txt
- Charset: Default
- Seperator: TAB
- Fields list:

Index	Name	Key
1	EID	<input type="checkbox"/>
2	NAME	<input type="checkbox"/>
3	SURNAME	<input type="checkbox"/>
4	GENDER	<input type="checkbox"/>
5	STATE	<input type="checkbox"/>
6	BIRTHDAY	<input type="checkbox"/>
7	HIREDATE	<input type="checkbox"/>
8	DEPT	<input type="checkbox"/>
9	SALARY	<input type="checkbox"/>



完整示例：添加数据源dept

添加一个数据库表类型数据源，命名数据源为dept。

添加另一个数据库表类型数据源，命名数据源为customers。

DEPTID	DEPTNAME
1	Sales Department
2	Comprehensive Department
3	Marketing Department
4	Finance Department
5	Office of the General Manager
6	Office of the President
7	Technical support department
8	Sell One
9	Sell Two
10	Research and Development Department
11	Technical Consultancy Center
12	Run Qian Company

Database table

Name: dept

Data source: mysql1

Table: dept

Where:

Quote type: Double quote

Fields:

Index	Name	Key
1	DEPTID	<input type="checkbox"/>
2	DEPTNAME	<input type="checkbox"/>

CUSTID	CUSTNAME
1	ALFKI
2	ANATR
3	ANTON
4	AROUT
5	BERGS
6	BLAUS
7	BLONP
8	BOLID
9	BONAP
10	BOTTM
11	BSBEV
12	CACTU

Database table

Name: customers

Data source: mysql1

Table: customers

Where:

Quote type: Double quote

Fields:

Index	Name	Key
1	CUSTID	<input type="checkbox"/>
2	CUSTNAME	<input type="checkbox"/>



完整示例：添加SQL类型数据源及设置参数

再添加一个SQL数据源，orders表中数据如下，命名数据源为orders。

因为orders表中的数据每天都在新增，在导出数据的时候我们根据ORDERDATE进行筛选导出是最方便的。我们在SQL语句中添加WHERE过滤条件，设置通过ORDERDATE过滤数据，并且设置参数arg1。

ORDERID	EMPID	CUSTID	ORDERDATE	AMOUNT
1	1007	33	2020-01-15	1081
2	1004	12	2020-02-06	414
3	1007	52	2020-01-21	228
4	1001	40	2020-02-19	1299
5	1005	46	2020-01-08	1210
6	1003	17	2020-01-04	555
7	1009	36	2020-02-26	544
8	1004	38	2020-01-30	958
9	1001	80	2020-02-05	1556
10	1005	16	2020-01-29	307
11	1003	19	2020-02-05	1268
12	1006	48	2020-01-30	818
13	1003	89	2020-02-09	189
14	1007	29	2020-02-18	1388
15	1003	61	2020-01-03	277
16	1003	36	2020-02-16	120
17	1008	88	2020-02-13	1171
18	1008	58	2020-01-17	895
19	1006	26	2020-01-10	860

SQL

Name: orders Data source: mysql1

SQL: SELECT * FROM orders where DATE_FORMAT(ORDERDATE, '%Y%m%d') = ?

Data is sorted on keys

Index	Name	Key
1	ORDERID	<input checked="" type="checkbox"/>
2	EMPID	<input type="checkbox"/>
3	ORDERDATE	<input type="checkbox"/>
4	AMOUNT	<input type="checkbox"/>

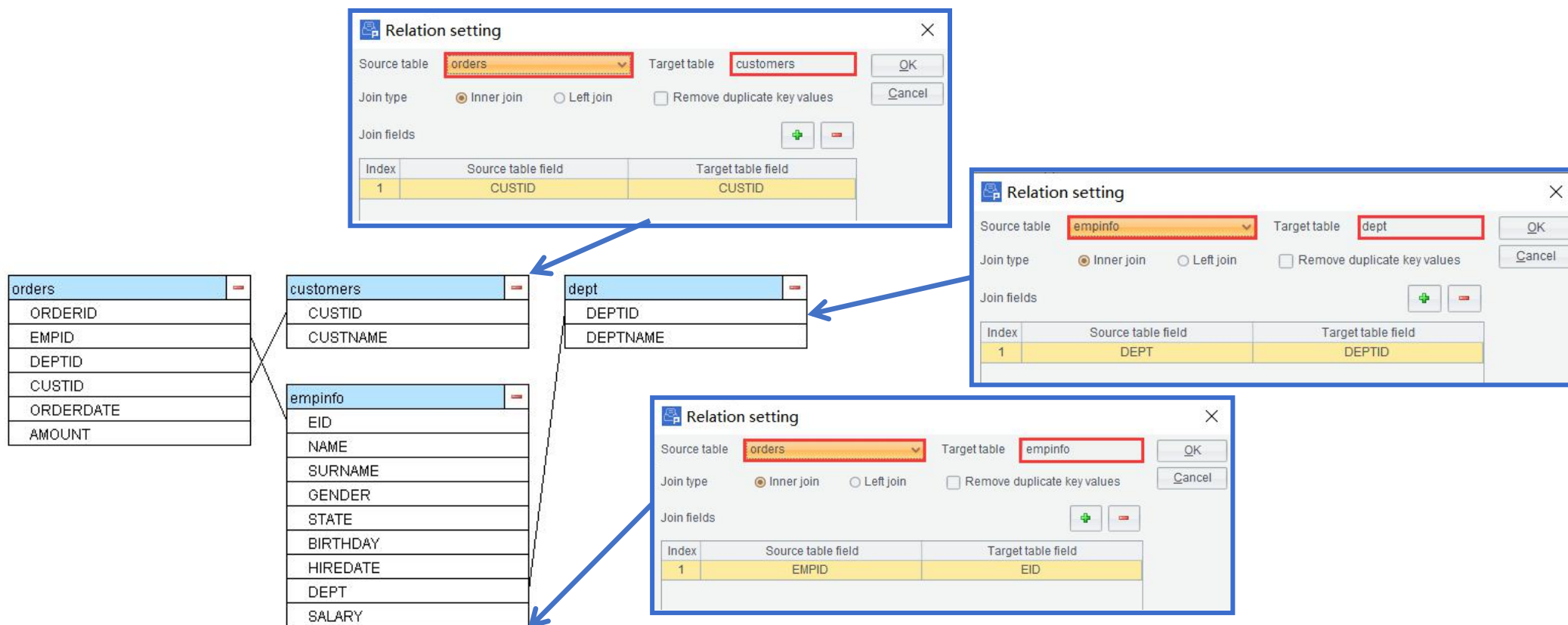
Index	Value
1	arg1



完整示例：新建视图关系

接下来新建视图关系。依次拖入拖拽orders表、customers表、empinfo表以及dept表。

子表根据不同的外键与主表进行连接——customers表中CUSTID与orders表中CUSTID相关联、empinfo表中EID与orders表中EMPID相关联、dept表中DEPTID与empinfo表中DEPT相关联。



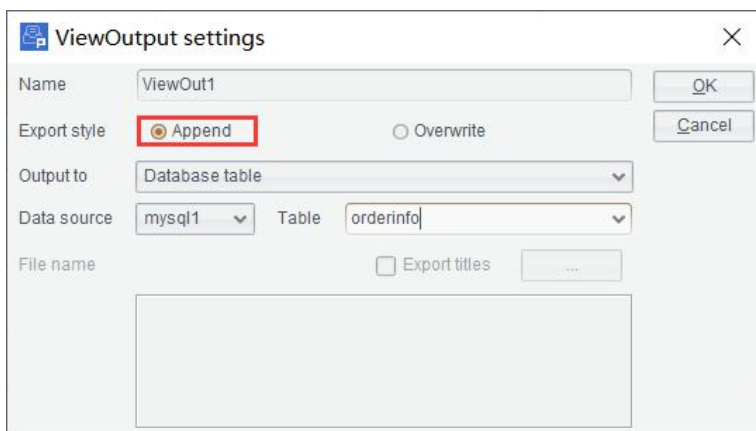


完整示例：新建视图结果及追加计算列

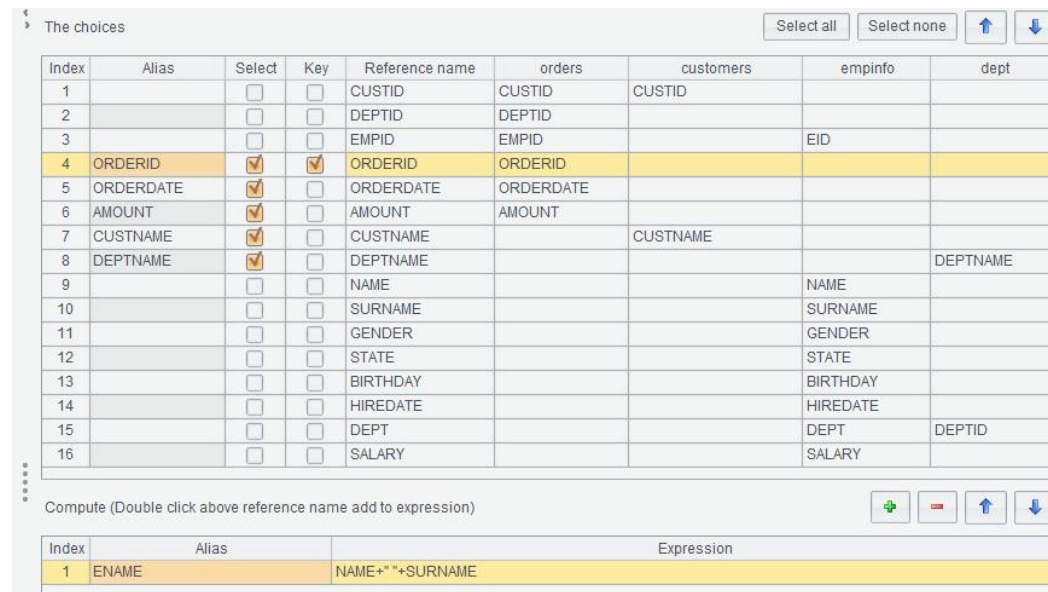
创建好视图关系，继续创建视图结果。

导出风格选择“追加”，导出为“数据库表”，并选择所要导出的数据源及数据表。

设置好导出设置后，我们对需要导出的列进行选择，此时我们添加一列计算列，计算列为职员完整姓名ENAME，ENAME由NAME及SURNAME拼接而成。



进行导出设置



选择列以及追加列



完整示例：导出及调度执行

将ETL过程存为orderinfo.ept。

使用命令行调度执行，每天更新时只需要更换不同的日期参数即可。

例如执行命令行 `Esprocx D:\orderinfo.ept arg1=20200219`

orderinfo.txt中追加ORDERDATE为2020年2月19日的数据。

ORDERID	ORDERDA...	AMOUNT	CUSTNAME	DEPTNAME	ENAME
8	2020-01-30	958	KOENE	Finance Department	Emily Smith
12	2020-01-30	818	MAGAA	Office of the General Manager	Matthew Johnson
38	2020-01-30	690	SPLIR	Marketing Department	Victoria Davis
100	2020-01-30	395	TOMSP	Office of the General Manager	Matthew Johnson
255	2020-01-30	495	WHITC	Sell Two	Ryan Johnson
286	2020-01-30	837	LAZYK	Technical support department	Rachel Johnson
301	2020-01-30	750	SANTG	Sell One	Alexis Smith
359	2020-01-30	322	MAGAA	Sell One	Ashley Smith
448	2020-01-30	386	WARTH	Sales Department	Rebecca Moore
469	2020-01-30	136	FOLKO	Sell One	Ashley Smith
4	2020-02-19	1299	LAMAI	Sales Department	Rebecca Moore
40	2020-02-19	229	SANTG	Sell Two	Ryan Johnson
43	2020-02-19	228	TORTU	Technical support department	Rachel Johnson
87	2020-02-19	1313	TRAIH	Sell One	Alexis Smith
163	2020-02-19	364	MAGAA	Technical support department	Megan Wilson
249	2020-02-19	1226	MAGAA	Technical support department	Megan Wilson

THANKS

—— 创新技术 推动应用进步 ——

