润乾
RAQSOFT

集算器简易ETL工具

润乾软件出品

www.raqsoft.com.cn

目录 CONTENTS

01

集算器简易ETL工具

基本介绍

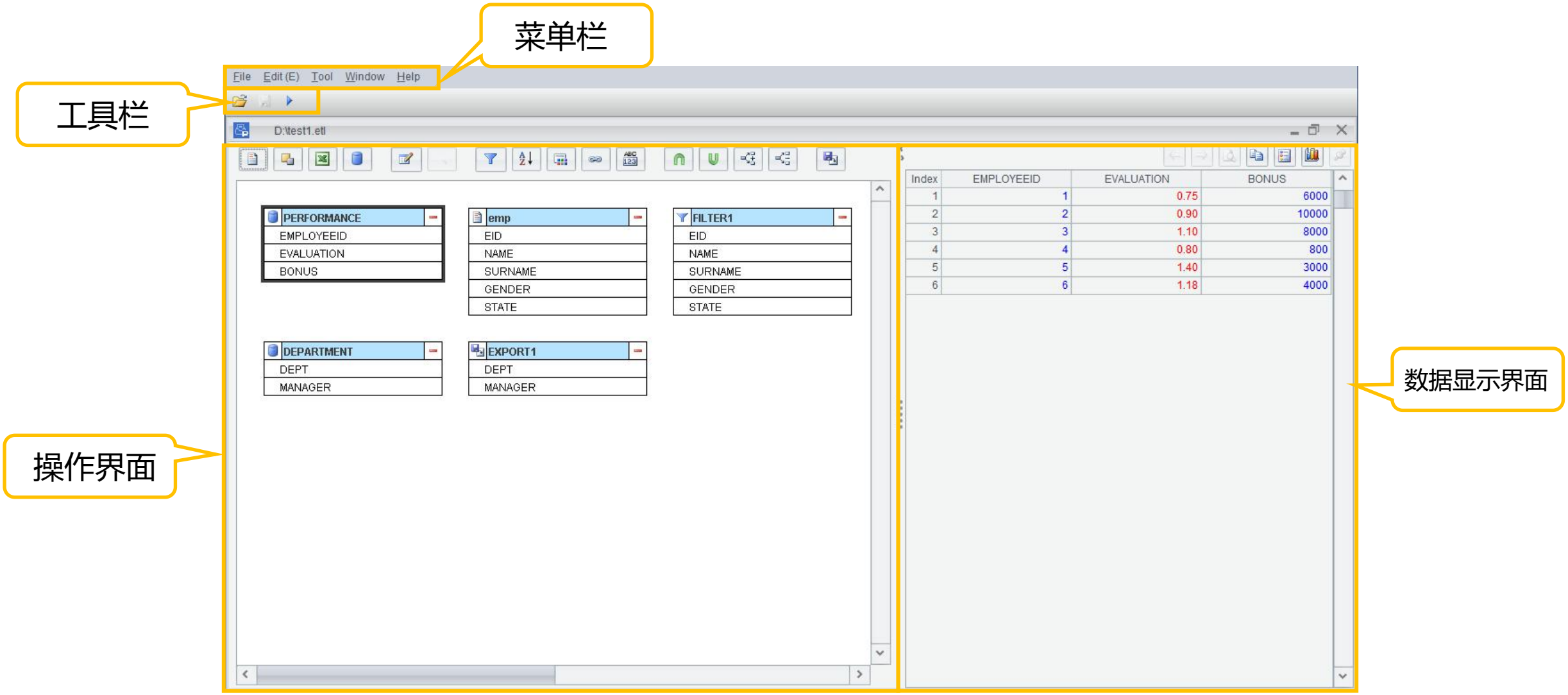
集算器ETL工具可以从多种数据源中抽取数据，进行过滤、排序、分组、连接、计算列及集合处理等，再将结果导出到文本、数据库、Excel、集算器文件等。

简略来说，ETL过程就是三步：

定义数据源 → 数据处理 → 结果写出

基本介绍

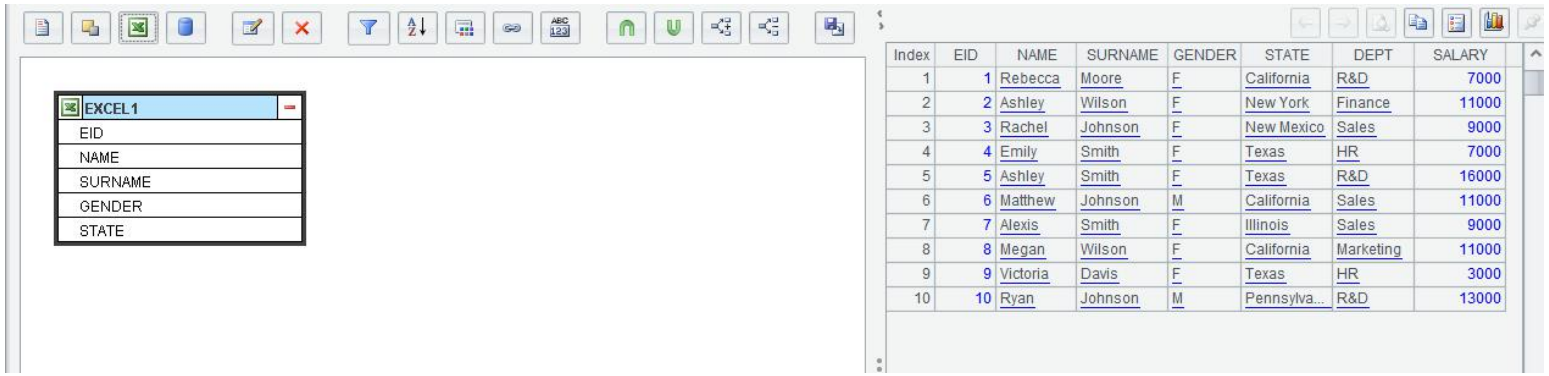
ETL工具的界面由菜单栏、工具栏、操作界面及数据显示界面组成。



数据源

ETL工具支持各种数据源：文本文件、Excel文件以及数据库表和BTX文件。

以Excel文件数据源为例：在操作界面中选择文本数据，编辑文本数据源的名称为EXCEL1，选择结果类型为序表，并在文件名称处打开需要作为数据源的文本文件emp.xlsx，选中第一行记录作为字段名，选择页面sheet2，点击确定，数据显示界面将会显示数据源中的所有数据。

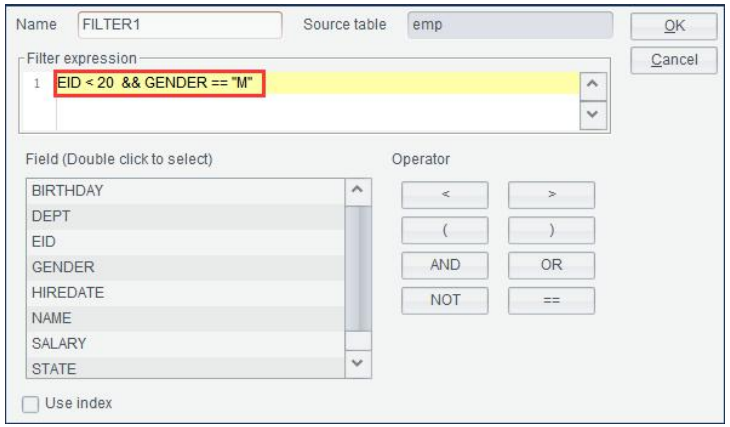


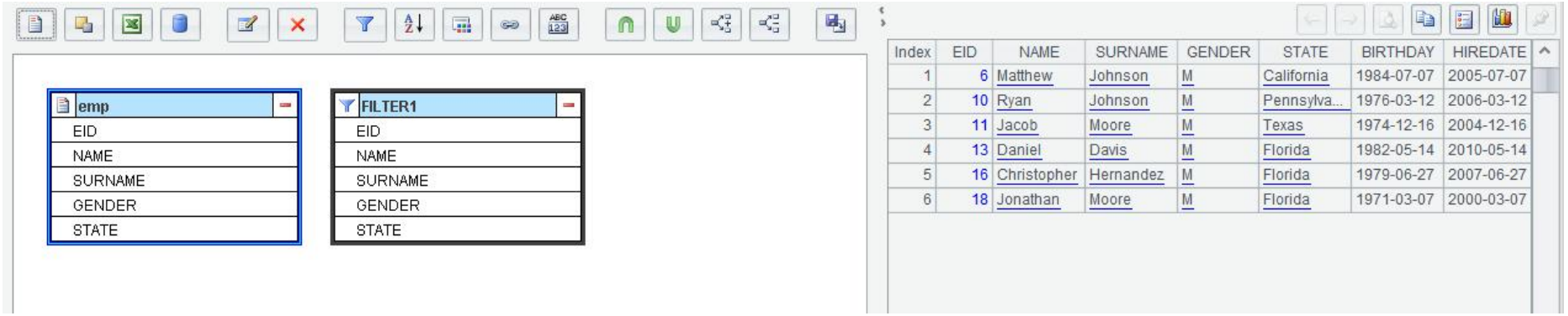
Index	EID	NAME	SURNAME	GENDER	STATE	DEPT	SALARY
1	1	Rebecca	Moore	F	California	R&D	7000
2	2	Ashley	Wilson	F	New York	Finance	11000
3	3	Rachel	Johnson	F	New Mexico	Sales	9000
4	4	Emily	Smith	F	Texas	HR	7000
5	5	Ashley	Smith	F	Texas	R&D	16000
6	6	Matthew	Johnson	M	California	Sales	11000
7	7	Alexis	Smith	F	Illinois	Sales	9000
8	8	Megan	Wilson	F	California	Marketing	11000
9	9	Victoria	Davis	F	Texas	HR	3000
10	10	Ryan	Johnson	M	Pennsylva...	R&D	13000

常用功能：数据过滤

选择数据源后，可以通过添加过滤表达式，对数据进行过滤。

下图以文本数据为例，对emp.xlsx设置过滤条件EID小于20并且GENDER为M，此时界面中会显示一个新的数据表FILTER1：

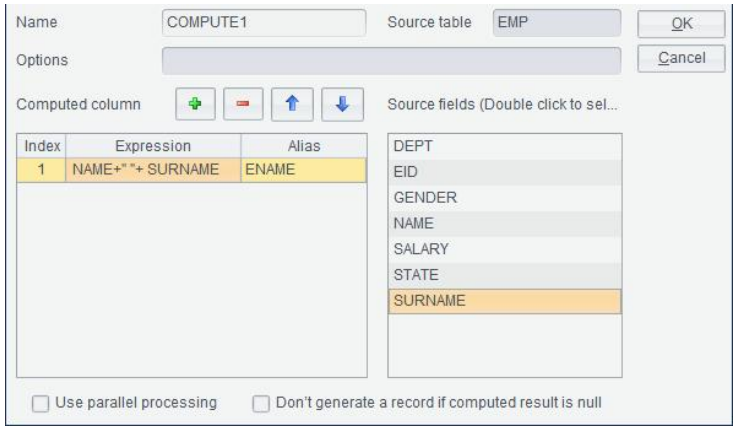


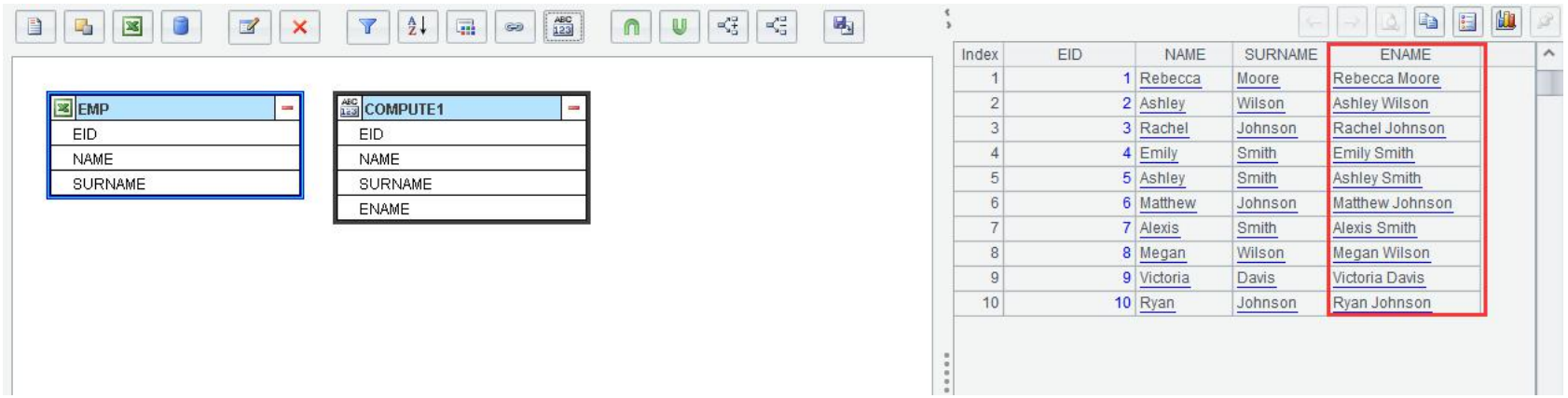


常用功能：计算列

计算列功能可以通过编写合适的表达式，生成新的计算列。

例如在EMP表中，增加一列ENAME，由NAME及SURNAME拼接而成，计算结果显示在新表COMPUTE1中：

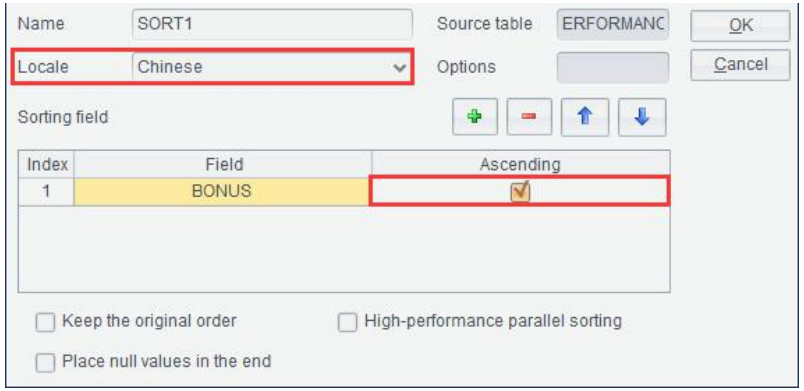


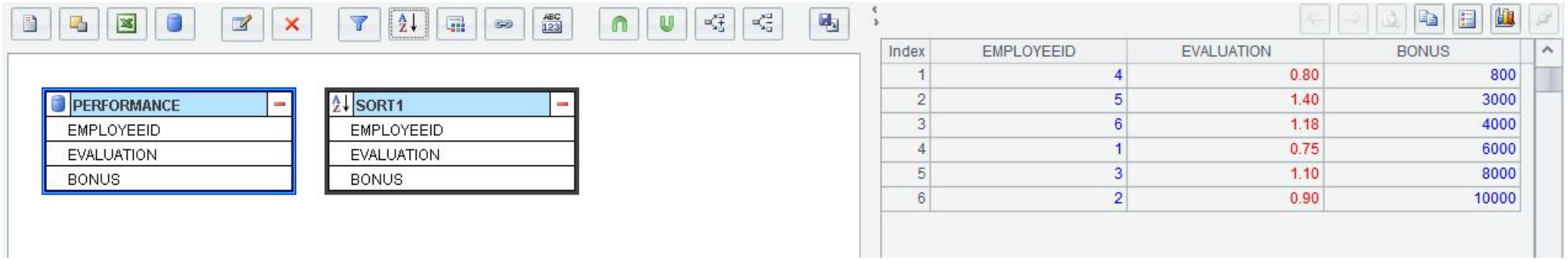


常用功能：数据排序

排序功能将数据按字段进行升降序排序。

下图以数据库demo中的表PERFORMANCE为例，对表中的BONUS字段选择根据中文进行升序排序，此时界面中会显示一个新的数据表SORT1：

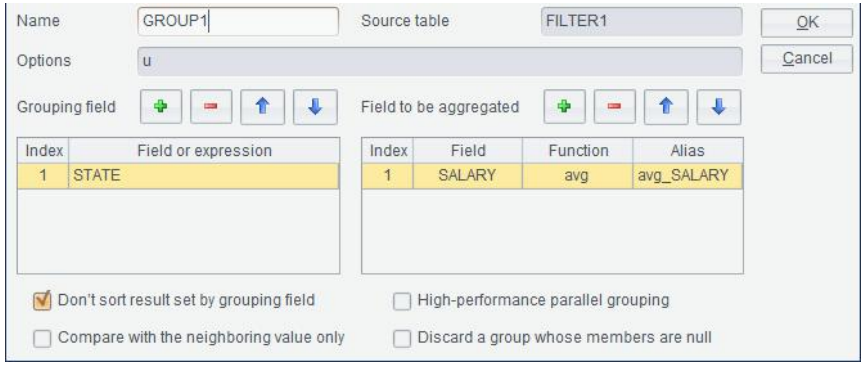


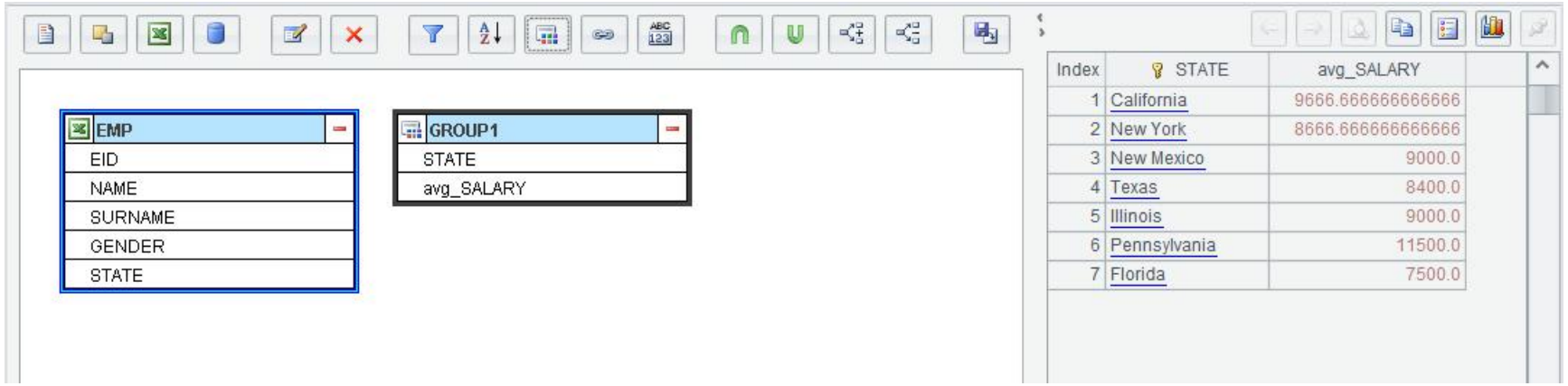


常用功能：数据分组

当需要对数据进行聚合处理时，我们可以选择数据分组功能。

例如对EMP表中的SALARY计算平均值，并设置分组表达式为STATE，选中结果集不再按分组字段排序，继而生成一个新表GROUP1，用以显示EMP表中各STATE的平均SALARY：

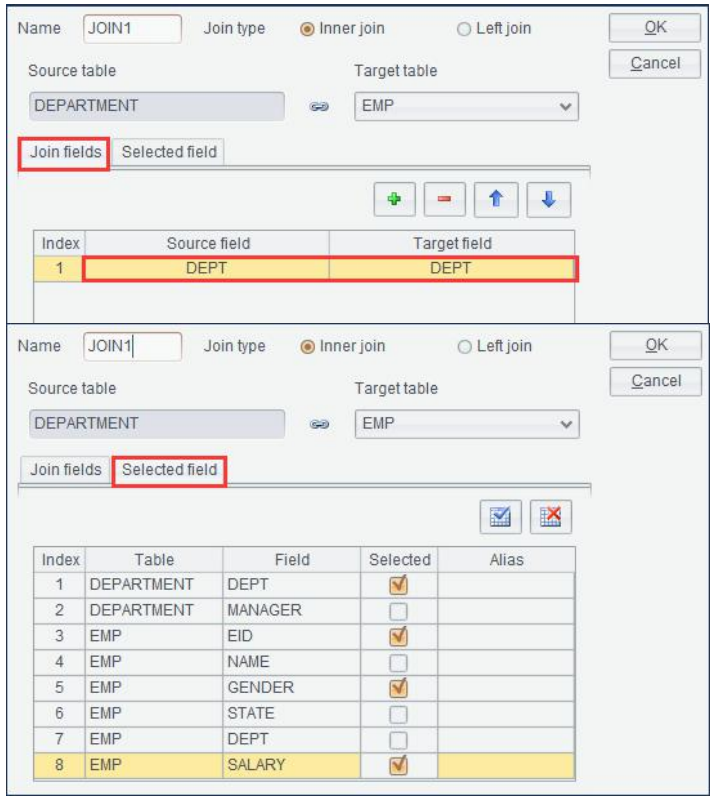


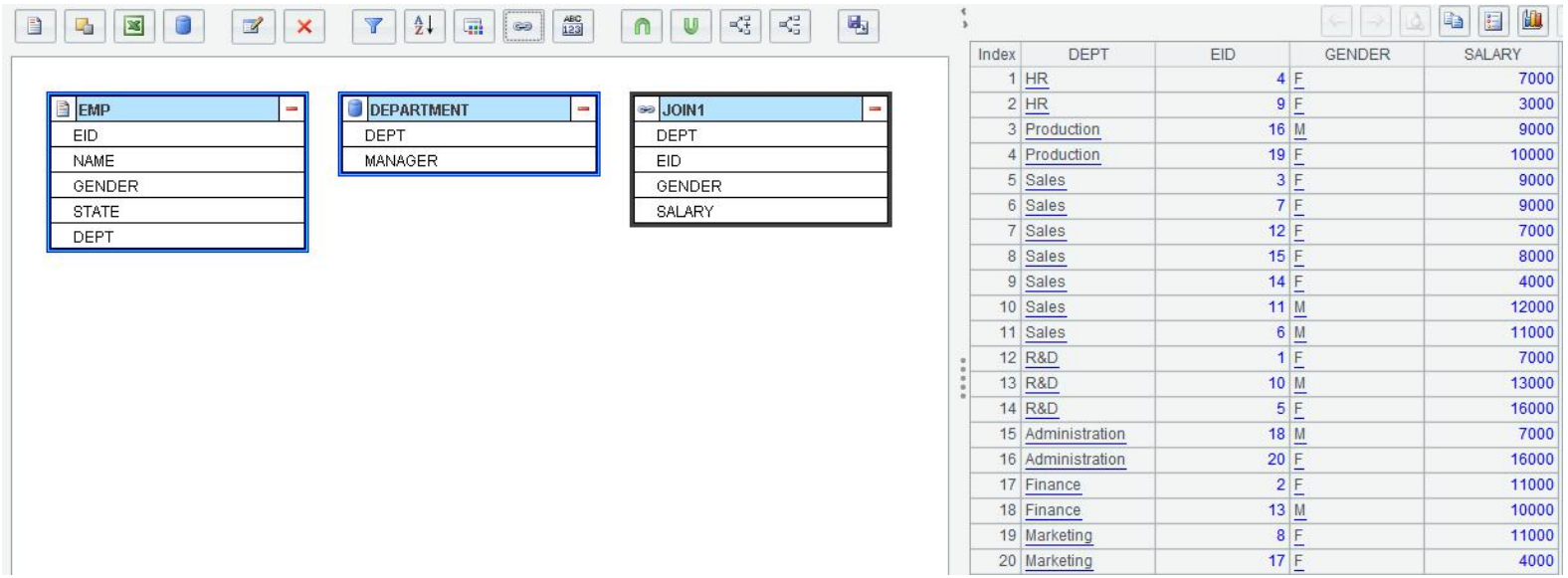


常用功能：数据连接

数据连接功能使得两表可以通过有关联的字段进行连接。

例如DEPARTMENT表通过DEPT字段内连接EMP表的DEPT字段，并且选择需要显示的字段，连接后的表显示为JOIN1：





常用功能：数据交集、数据并集

选择数据交集或数据并集功能，将数据结构相同的两表进行交集或并集。

例如两表EMP1及EMP2，进行数据交集和数据并集，分别生成新表INTERSECTION1和UNION1：

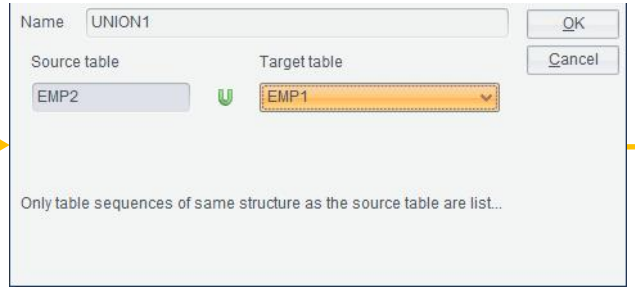
EMP1

Index	EID	NAME	DEPT	SALARY
1	1	Rebecca	R&D	7000
2	2	Ashley	Finance	11000
3	3	Rachel	Sales	9000
4	4	Emily	HR	7000
5	5	Ashley	R&D	16000
6	6	Matthew	Sales	11000
7	7	Alexis	Sales	9000
8	8	Megan	Marketing	11000
9	9	Victoria	HR	3000
10	10	Ryan	R&D	13000
11	11	Jacob	Sales	12000
12	12	Jessica	Sales	7000
13	13	Daniel	Finance	10000
14	14	Alyssa	Sales	4000
15	15	Alexis	Sales	8000
16	16	Christopher	Production	9000
17	17	Hannah	Marketing	4000
18	18	Jonathan	Administration	7000
19	19	Samantha	Production	10000

EMP2

Index	EID	NAME	DEPT	SALARY
1	13	Daniel	Finance	10000
2	14	Alyssa	Sales	4000
3	15	Alexis	Sales	8000
4	16	Christopher	Production	9000
5	17	Hannah	Marketing	4000
6	18	Jonathan	Administration	7000
7	19	Samantha	Production	10000
8	20	Alexis	Administration	16000
9	21	Jacob	Marketing	10000
10	22	Jacob	R&D	16000
11	23	Joseph	Finance	6000
12	24	Chloe	Finance	10000

数据交集



数据并集

INTERSECTION1

Index	EID	NAME	DEPT	SALARY
1	13	Daniel	Finance	10000
2	14	Alyssa	Sales	4000
3	15	Alexis	Sales	8000
4	16	Christopher	Production	9000
5	17	Hannah	Marketing	4000
6	18	Jonathan	Administration	7000
7	19	Samantha	Production	10000

UNION1

Index	EID	NAME	DEPT	SALARY
1	13	Daniel	Finance	10000
2	14	Alyssa	Sales	4000
3	15	Alexis	Sales	8000
4	16	Christopher	Production	9000
5	17	Hannah	Marketing	4000
6	18	Jonathan	Administration	7000
7	19	Samantha	Production	10000
8	20	Alexis	Administration	16000
9	21	Jacob	Marketing	10000
10	22	Jacob	R&D	16000
11	23	Joseph	Finance	6000
12	24	Chloe	Finance	10000
13	1	Rebecca	R&D	7000
14	2	Ashley	Finance	11000
15	3	Rachel	Sales	9000
16	4	Emily	HR	7000
17	5	Ashley	R&D	16000
18	6	Matthew	Sales	11000
19	7	Alexis	Sales	9000
20	8	Megan	Marketing	11000
21	9	Victoria	HR	3000
22	10	Ryan	R&D	13000
23	11	Jacob	Sales	12000
24	12	Jessica	Sales	7000

常用功能：数据和集、数据差集

选择数据和集或数据差集功能，将数据结构相同的两表进行和集或差集计算。

例如两表EMP1及EMP2，进行数据和集和数据差集，分别生成新表CONCATENATION1和DIFFERENCE1：

EMP1

Index	EID	NAME	DEPT	SALARY
1	1	Rebecca	R&D	7000
2	2	Ashley	Finance	11000
3	3	Rachel	Sales	9000
4	4	Emily	HR	7000
5	5	Ashley	R&D	16000
6	6	Matthew	Sales	11000
7	7	Alexis	Sales	9000
8	8	Megan	Marketing	11000
9	9	Victoria	HR	3000
10	10	Ryan	R&D	13000
11	11	Jacob	Sales	12000
12	12	Jessica	Sales	7000
13	13	Daniel	Finance	10000
14	14	Alyssa	Sales	4000
15	15	Alexis	Sales	8000
16	16	Christopher	Production	9000
17	17	Hannah	Marketing	4000
18	18	Jonathan	Administration	7000
19	19	Samantha	Production	10000

EMP2

Index	EID	NAME	DEPT	SALARY
1	13	Daniel	Finance	10000
2	14	Alyssa	Sales	4000
3	15	Alexis	Sales	8000
4	16	Christopher	Production	9000
5	17	Hannah	Marketing	4000
6	18	Jonathan	Administration	7000
7	19	Samantha	Production	10000
8	20	Alexis	Administration	16000
9	21	Jacob	Marketing	10000
10	22	Jacob	R&D	16000
11	23	Joseph	Finance	6000
12	24	Chloe	Finance	10000

数据和集

数据差集

CONCATENATION1

Index	EID	NAME	DEPT	SALARY
1	13	Daniel	Finance	10000
2	14	Alyssa	Sales	4000
3	15	Alexis	Sales	8000
4	16	Christopher	Production	9000
5	17	Hannah	Marketing	4000
6	18	Jonathan	Administration	7000
7	19	Samantha	Production	10000
8	20	Alexis	Administration	16000
9	21	Jacob	Marketing	10000
10	22	Jacob	R&D	16000
11	23	Joseph	Finance	6000
12	24	Chloe	Finance	10000
13	1	Rebecca	R&D	7000
14	2	Ashley	Finance	11000
15	3	Rachel	Sales	9000
16	4	Emily	HR	7000
17	5	Ashley	R&D	16000
18	6	Matthew	Sales	11000
19	7	Alexis	Sales	9000
20	8	Megan	Marketing	11000
21	9	Victoria	HR	3000
22	10	Ryan	R&D	13000
23	11	Jacob	Sales	12000
24	12	Jessica	Sales	7000
25	13	Daniel	Finance	10000
26	14	Alyssa	Sales	4000
27	15	Alexis	Sales	8000
28	16	Christopher	Production	9000
29	17	Hannah	Marketing	4000
30	18	Jonathan	Administration	7000
31	19	Samantha	Production	10000

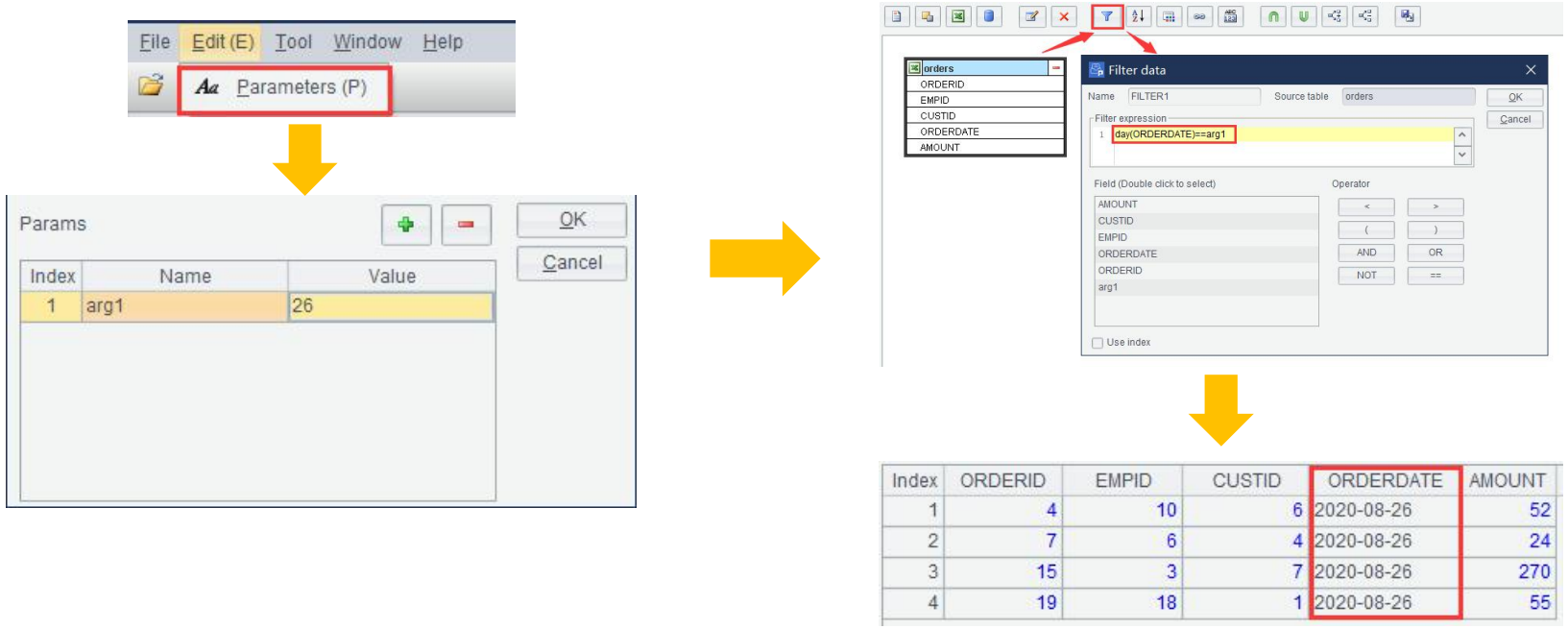
DIFFERENCE1

Index	EID	NAME	DEPT	SALARY
1	20	Alexis	Administration	16000
2	21	Jacob	Marketing	10000
3	22	Jacob	R&D	16000
4	23	Joseph	Finance	6000
5	24	Chloe	Finance	10000

设置参数

可以设置参数来取出精准或规定范围内的数据，在菜单栏选择编辑 → 参数中添加参数，然后在过滤表达式中设置参数。

下图对orders表进行数据过滤，首先添加参数，设置参数为arg1，参数值为26；在数据过滤时添加过滤表达式“day(ORDERDATE)==arg1”，用来选出指定日期的记录，点击确定后，生成FILTER1用以显示orders表中日期为26的ORDERDATE。



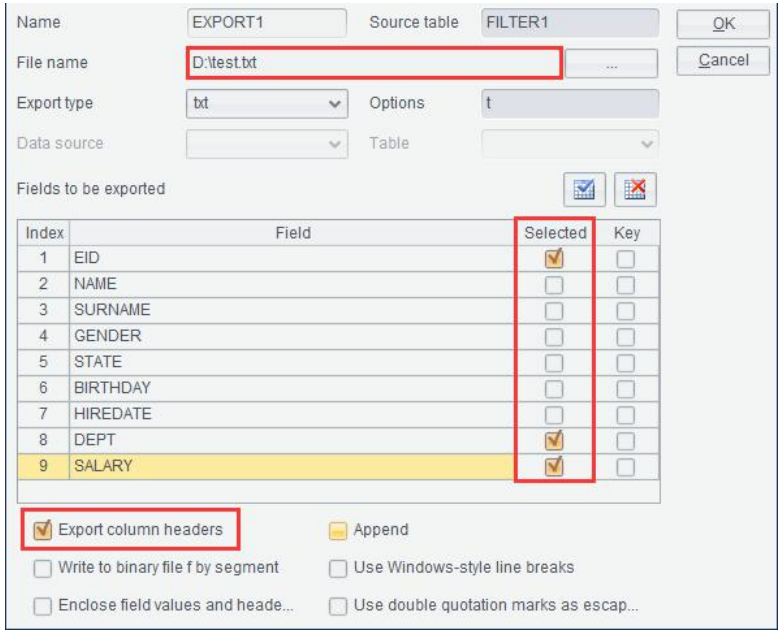
Index	ORDERID	EMPID	CUSTID	ORDERDATE	AMOUNT
1	4	10	6	2020-08-26	52
2	7	6	4	2020-08-26	24
3	15	3	7	2020-08-26	270
4	19	18	1	2020-08-26	55

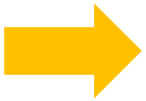
结果输出：数据导出

数据导出可以导出至CTX、BTX、TXT、CSV、XLSX和数据库表中。

在数据导出界面，我们可以根据自己的需求选择相应的选项，导出为不同格式的文件会存在一些异同，比如：在导出为TXT、CSV及XLSX时，可以选择是否导出标题，其他格式则无法选择；而导出为数据库表，无论是追加或是覆盖写入，所要导出的表中列名都要与数据库表中的列名对应。

下例将结果集导出为txt格式，选择需要导出的字段，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.etl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 [etlFile] [argN]...

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

[argN]：etlFile有参数时，参数按照参数顺序指定；

示例：

esprocx demo.etl 1

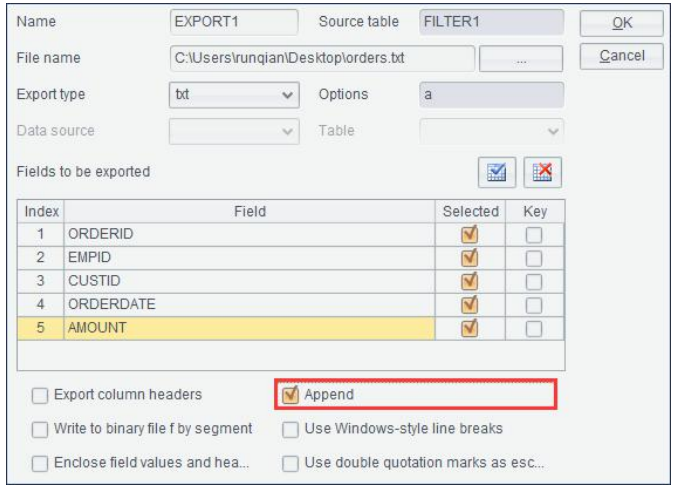
对应参数month为1月，执行寻址路径上的demo.etl。

结果输出：调度执行

在数据导出中选择导出风格为追加写入；使用设置参数中的例子

将.etl存在本地，并且执行命令语句：esprocx ORDERS.etl 27

27日的所有数据将会追加到orders.txt中。



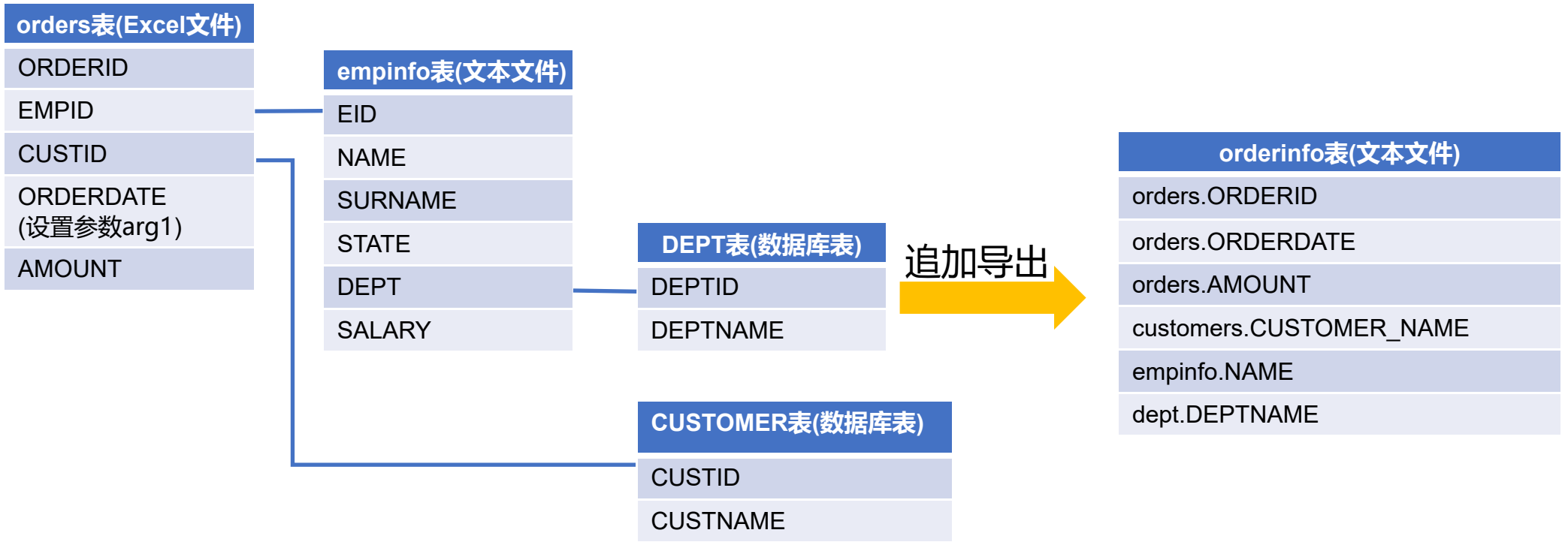
选择导出风格为追加

ORDERID	EMPID	CUSTID	ORDERDATE	AMOUNT
4	10	6	2020-08-26	52
7	6	4	2020-08-26	24
15	3	7	2020-08-26	270
19	18	1	2020-08-26	55
16	5	7	2020-08-27	694

执行语句后的orders.txt

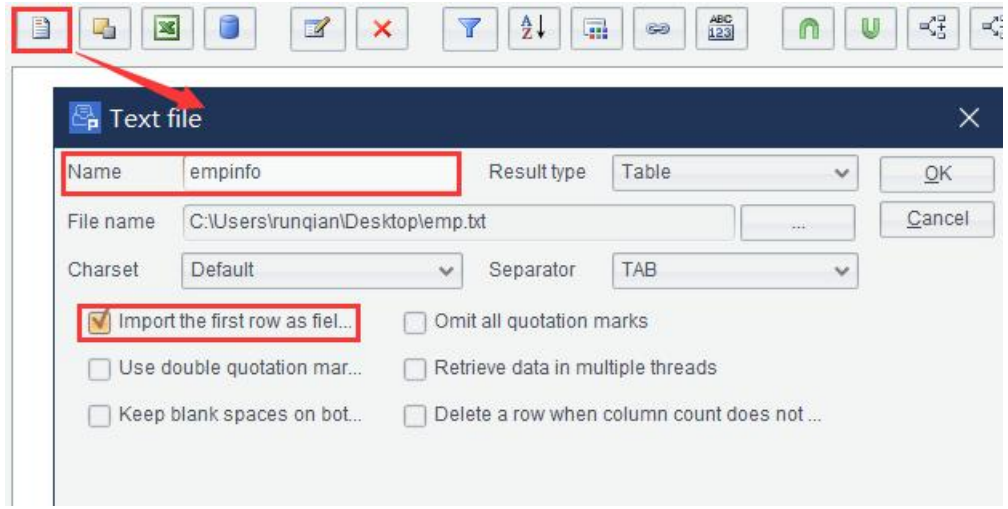
完整示例：描述示例

现有四个源表如下图，ETL任务为：将它们关联后导出至文本文件orderinfo.txt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.txt追加。ETL过程将被每天调度使用。



完整示例：添加文本文件类型数据源empinfo

首先添加一个文本文件数据源，打开文件emp.txt，选择第一行记录作为字段名，并且命名数据源为empinfo。

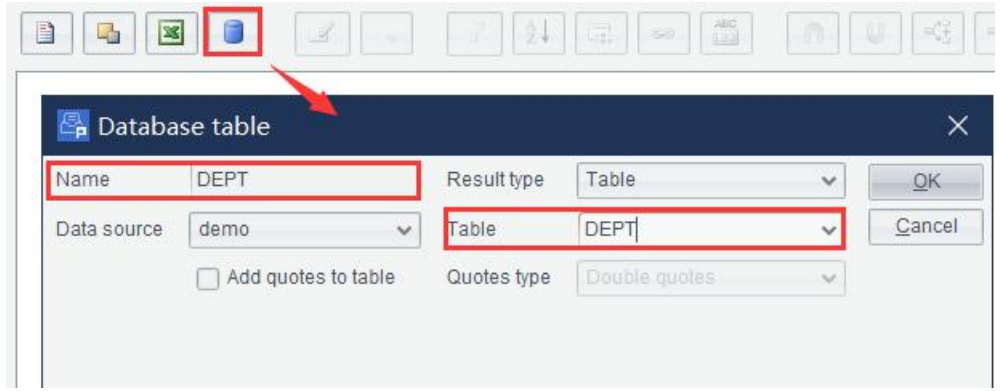


Index	EID	NAME	SURNAME	STATE	DEPT	SALARY
1	1	Rebecca	Moore	California	2	7000
2	2	Ashley	Wilson	New York	1	11000
3	3	Rachel	Johnson	New Mexico	6	9000
4	4	Emily	Smith	Texas	2	7000
5	5	Ashley	Smith	Texas	10	16000
6	6	Matthew	Johnson	California	9	11000
7	7	Alexis	Smith	Illinois	4	9000
8	8	Megan	Wilson	California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

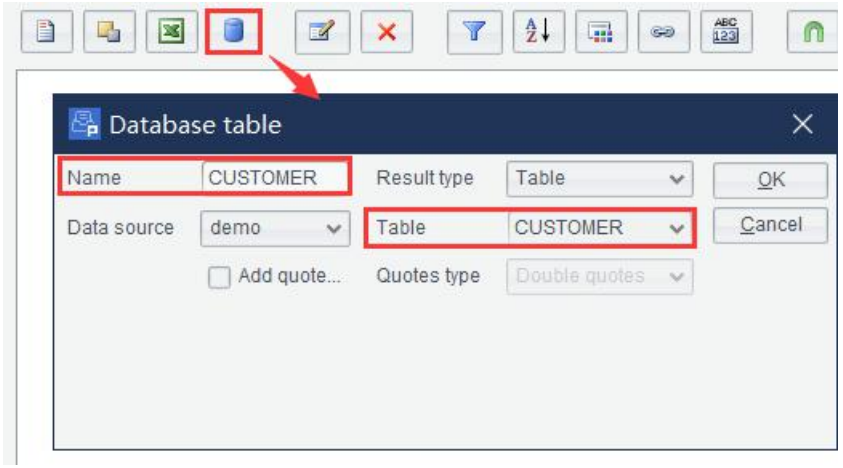
完整示例：添加数据源dept和customers

添加一个数据库表类型数据源，选择DEPT表，命名数据源为DEPT。

添加另一个数据库表类型数据源，选择CUSTOMER表，命名数据源为CUSTOMER。



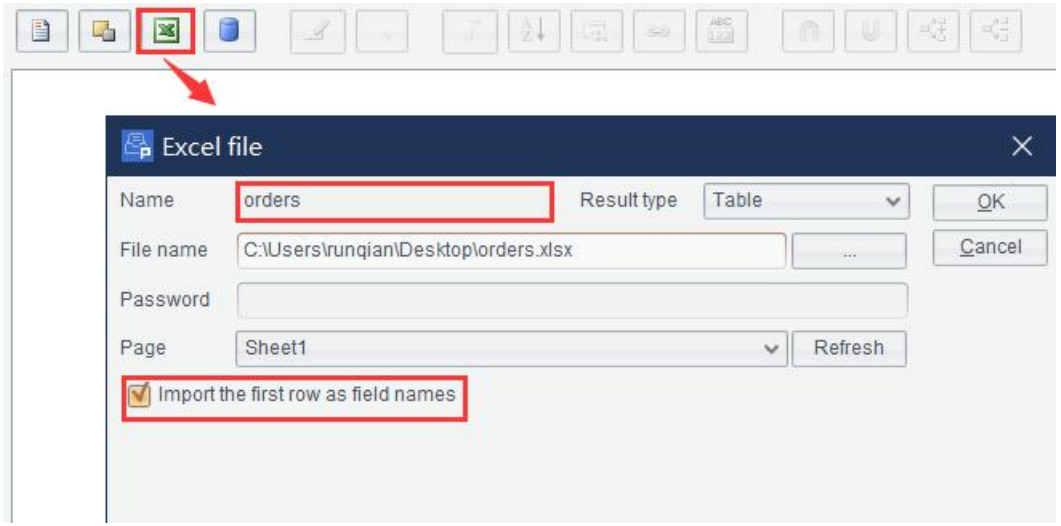
Index	DEPTID	DEPTNAME	FATHER
1	1	Sales Department	12
2	10	Research and Development Department	12
3	12	Run Qian Company	(null)
4	11	Technical Consultancy Center	12
5	2	Comprehensive Department	12
6	3	Marketing Department	12
7	4	Finance Department	2
8	5	Office of the General Manager	12
9	6	Office of the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



Index	CUSTOMER_ID	CUSTOMER_NAME
1	5	Crystal
2	6	Chemist
3	7	Murli
4	8	Chemical Mumai
5	1	WuXi food processing plant
6	2	TianJin the Great Wall che...
7	3	JiangSu Tianlong food group
8	4	China chemical raw materi...

完整示例：添加Excel类型数据源orders

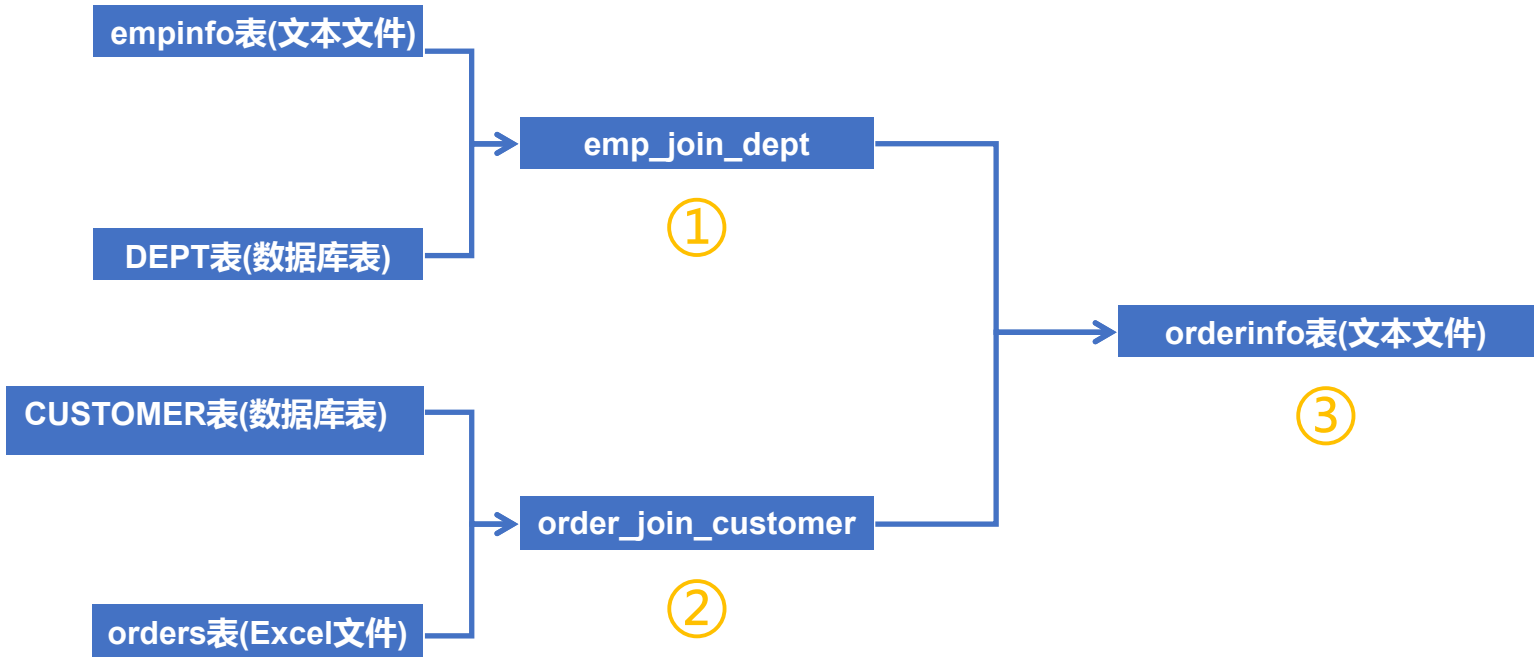
添加Excel文件，打开文件orders.xlsx，选择第一行记录作为字段名，并且命名数据源为orders。



Index	ORDERID	EMPID	CUSTID	ORDERDATE	AMOUNT
1	1	5	1	2020-08-12	32
2	2	3	7	2020-08-13	78
3	3	1	8	2020-08-14	456
4	4	10	6	2020-08-26	52
5	5	7	5	2020-08-16	422
6	6	12	2	2020-08-29	78
7	7	6	4	2020-08-26	24
8	8	4	1	2020-08-19	454
9	9	18	2	2020-08-20	120
10	10	2	3	2020-08-29	24
11	11	20	4	2020-08-12	201
12	12	15	8	2020-08-23	820
13	13	13	5	2020-08-24	12
14	14	7	7	2020-08-25	25
15	15	3	7	2020-08-26	270
16	16	5	7	2020-08-27	694
17	17	4	2	2020-08-12	24
18	18	8	3	2020-08-29	89
19	19	18	1	2020-08-26	55
20	20	11	6	2020-08-29	34

完整示例：多表数据连接

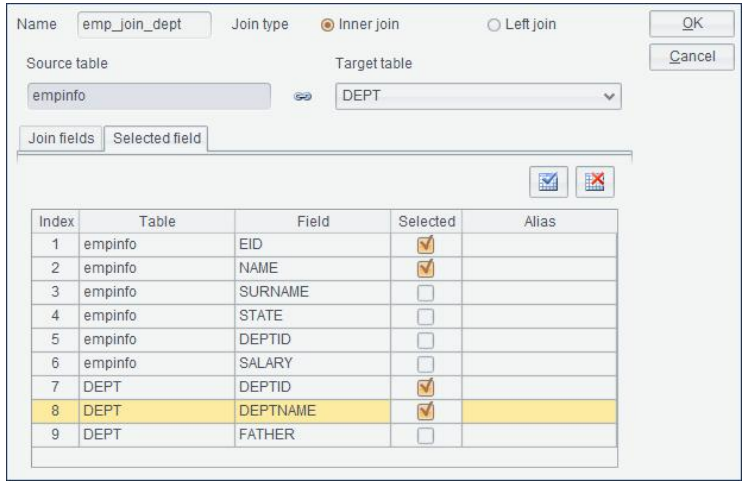
接下来进行多表数据连接。分析需求中的表关系，我们可以分为三步进行多表连接。



完整示例：多表数据连接

首先通过empinfo表的DEPTID与DEPT表的DEPTID进行内连接，选择所需的字段，并将连接后生成的新表命名为emp_join_dept。

两表连接



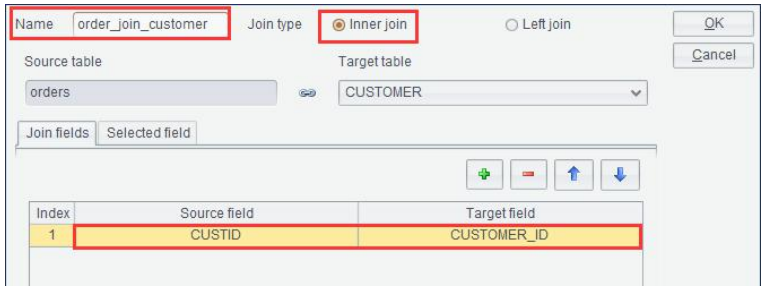
选择字段

Index	EID	NAME	DEPTID	DEPTNAME
1	2	Ashley	1	Sales Department
2	14	Alyssa	1	Sales Department
3	1	Rebecca	2	Comprehensive Department
4	10	Ryan	2	Comprehensive Department
5	12	Jessica	2	Comprehensive Department
6	4	Emily	2	Comprehensive Department
7	9	Victoria	3	Marketing Department
8	13	Daniel	3	Marketing Department
9	7	Alexis	4	Finance Department
10	11	Jacob	5	Office of the General Manager
11	16	Christopher	5	Office of the General Manager
12	3	Rachel	6	Office of the President
13	18	Jonathan	6	Office of the President
14	15	Alexis	7	Technical support department
15	20	Alexis	7	Technical support department
16	19	Samantha	8	Sell One
17	6	Matthew	9	Sell Two
18	5	Ashley	10	Research and Development Department
19	8	Megan	11	Technical Consultancy Center
20	17	Hannah	12	Run Qian Company

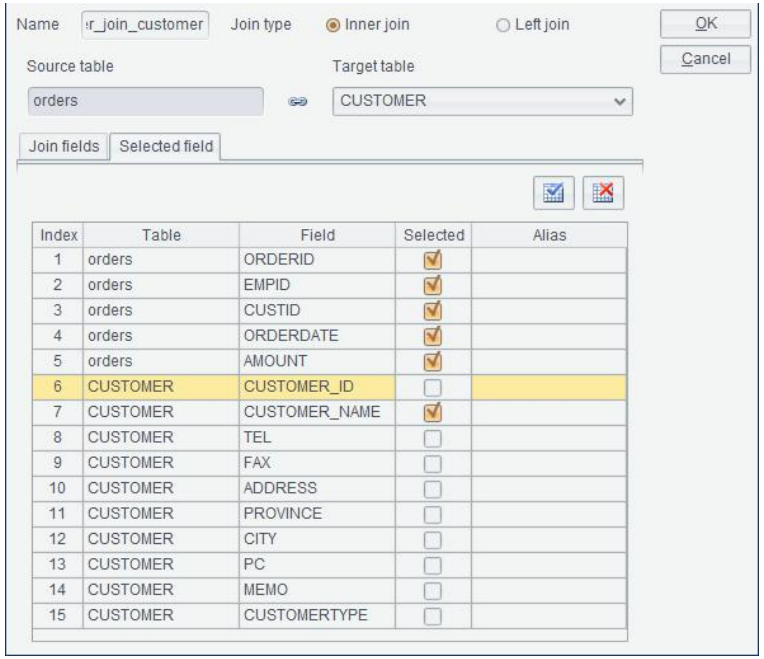
emp_join_dept

完整示例：多表数据连接

然后将CUSTOMER表与orders表关联起来，选择内连接，将新表命名为order_join_customer。



两表连接



选择字段

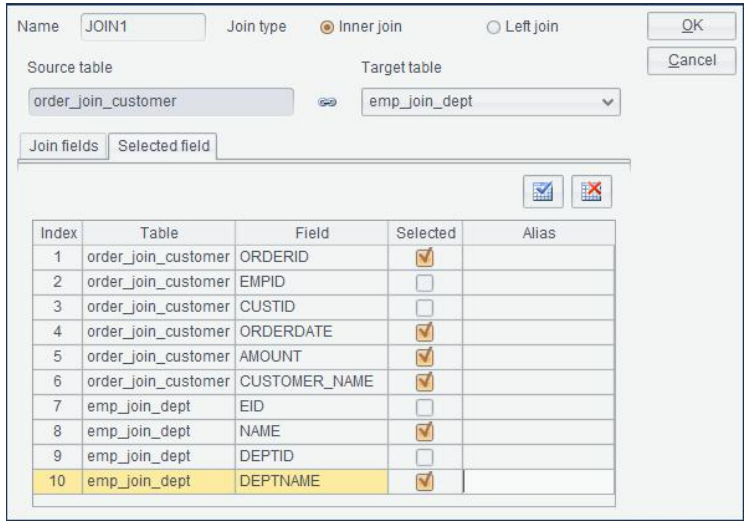
Index	ORDERID	EMPID	CUSTID	ORDERDATE	AMOUNT	CUSTOMER_NAME
1	1	5	1	2020-08-12	32	WuXi food processing plant
2	19	18	1	2020-08-26	55	WuXi food processing plant
3	8	4	1	2020-08-19	454	WuXi food processing plant
4	6	12	2	2020-08-29	78	TianJin the Great Wall chemical plant
5	17	4	2	2020-08-12	24	TianJin the Great Wall chemical plant
6	9	18	2	2020-08-20	120	TianJin the Great Wall chemical plant
7	10	2	3	2020-08-29	24	JiangSu Tianlong food group
8	18	8	3	2020-08-29	89	JiangSu Tianlong food group
9	7	6	4	2020-08-26	24	China chemical raw material plant
10	11	20	4	2020-08-12	201	China chemical raw material plant
11	5	7	5	2020-08-16	422	Crystal
12	13	13	5	2020-08-24	12	Crystal
13	4	10	6	2020-08-26	52	Chemist
14	20	11	6	2020-08-29	34	Chemist
15	2	3	7	2020-08-13	78	Murli
16	15	3	7	2020-08-26	270	Murli
17	16	5	7	2020-08-27	694	Murli
18	14	7	7	2020-08-25	25	Murli
19	3	1	8	2020-08-14	456	Chemical Mumai
20	12	15	8	2020-08-23	820	Chemical Mumai

order_join_customer

完整示例：多表数据连接

最后将order_join_customer与emp_join_dept连接起来，多表数据连接结束。

两表连接



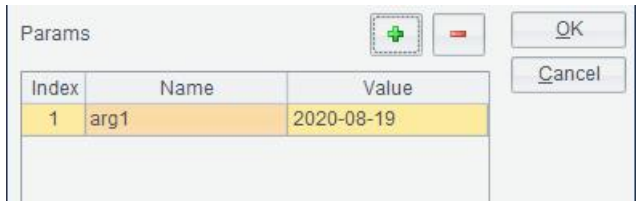
选择字段

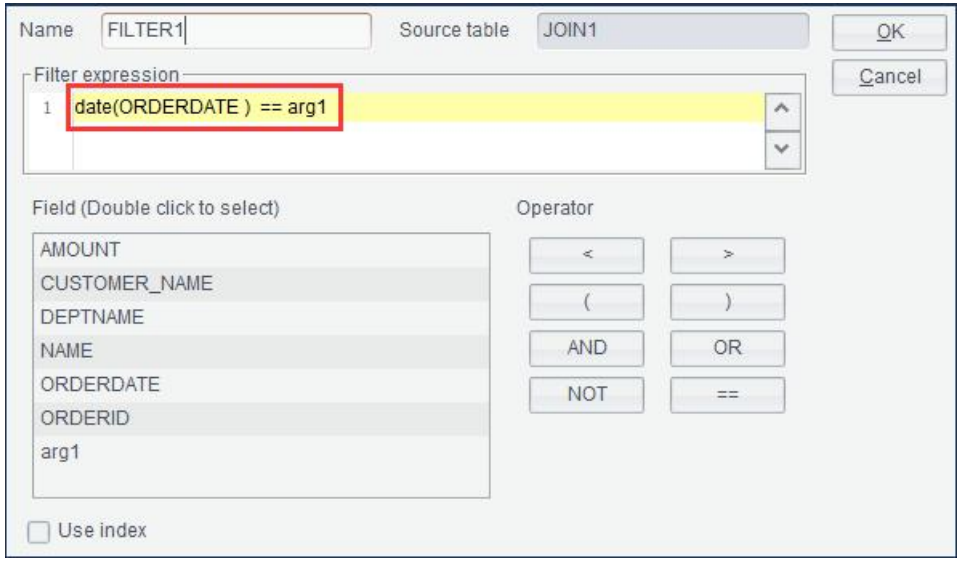
Index	ORDERID	ORDERDATE	AMOUNT	CUSTOMER_NAME	NAME	DEPTNAME
1	3	2020-08-14	456	Chemical Mumai	Rebecca	Comprehensive ...
2	10	2020-08-29	24	JiangSu Tianlong f...	Ashley	Sales Department
3	2	2020-08-13	78	Murli	Rachel	Office of the Presi...
4	15	2020-08-26	270	Murli	Rachel	Office of the Presi...
5	8	2020-08-19	454	WuXi food processi...	Emily	Comprehensive ...
6	17	2020-08-12	24	TianJin the Great W...	Emily	Comprehensive ...
7	1	2020-08-12	32	WuXi food processi...	Ashley	Research and De...
8	16	2020-08-27	694	Murli	Ashley	Research and De...
9	7	2020-08-26	24	China chemical ra...	Matthew	Sell Two
10	5	2020-08-16	422	Crystal	Alexis	Finance Departm...
11	14	2020-08-25	25	Murli	Alexis	Finance Departm...
12	18	2020-08-29	89	JiangSu Tianlong f...	Megan	Technical Consult...
13	4	2020-08-26	52	Chemist	Ryan	Comprehensive ...
14	20	2020-08-29	34	Chemist	Jacob	Office of the Gene...
15	6	2020-08-29	78	TianJin the Great W...	Jessica	Comprehensive ...
16	13	2020-08-24	12	Crystal	Daniel	Marketing Depart...
17	12	2020-08-23	820	Chemical Mumai	Alexis	Technical support...
18	19	2020-08-26	55	WuXi food processi...	Jonathan	Office of the Presi...
19	9	2020-08-20	120	TianJin the Great W...	Jonathan	Office of the Presi...
20	11	2020-08-12	201	China chemical ra...	Alexis	Technical support...

得到所需的表JOIN1

完整示例：设置参数

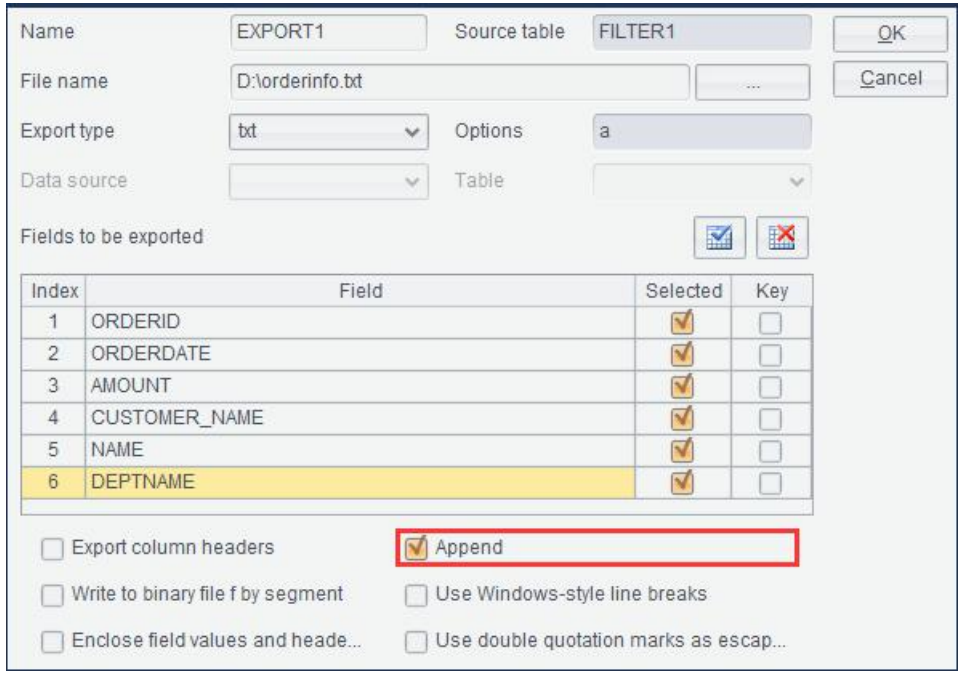
因为orders表中的数据每天都在新增，在导出数据的时候我们根据ORDERDATE进行筛选导出是最方便的。我们对JOIN1表进行数据过滤，设置参数arg1，设置过滤条件，表示通过字段ORDERDATE过滤数据，过滤后的表名为FILTER1。





完整示例：数据导出并设置追加写入

设置对FILTER1进行数据导出，选择需要导出的列，并且选中追加写入选项，每次执行时，根据不同的参数筛选出的数据将会追加写入到D盘下的orderinfo.txt中。



完整示例：调度执行

将ETL过程存为orderinfo.etl。

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.etl 2020-08-13

orderinfo.txt中追加ORDERDATE为2020年8月13日的数据。

ORDERID	ORDERDATE	AMOUNT	CUSTOMER_NAME	NAME	DEPTNAME
17	2020-08-12	24	TianJin the Great Wall chemical plant	Emily	Comprehensive Department
1	2020-08-12	32	WuXi food processing plant	Ashley	Research and Development Department
11	2020-08-12	201	China chemical raw material plant	Alexis	Technical support department
2	2020-08-13	78	Murli	Rachel	Office of the President

润乾
RAQSOFT

THANKS

创新技术 推动应用进步

www.raqsoft.com.cn