SPL Base

集算器教案

集合间运算

目录

CONTENTS

01 集合成员为基本数据类型

1. 和
2. 交
3. 并
4. 差
5. 异或
6. 多个集合间的运算

02 集合成员为记录

1. 直接比对记录引用
2. 有序归并比对字段
3. 有序归并比对主键
4. 有序归并比对所有字段
5. 比对字段无序

03 大数据量下的集合间运算

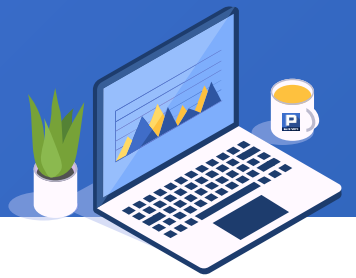
1. 简单和列
2. 有序归并比对列值
3. 隐含归并比对维字段

CONTENTS

1. 和
2. 交
3. 并
4. 差
5. 异或
6. 多个集合间的运算

集合成员为基本数据类型

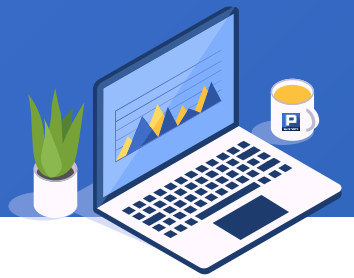
1. 和



2014年和2015年的销售记录分别存储在文件S2014.txt 和S2015.txt 中，求这两年内每个客户的销售次数。销售表结构相同，如下：

ID	Customer	Date	Amount
10400	EASTC	2014/01/01	3063.0
10401	RATTC	2014/01/01	3868.6
10402	ERNSH	2014/01/02	2713.5
…	…	…	…

1. 和

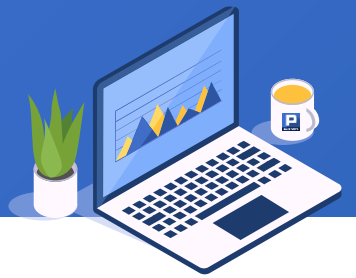


SPL如下，其中用到了符号“|” 求和列：

	A	B
1	=file("S2014.txt").import@t(Customer)	/导入2014年的客户
2	=file("S2015.txt").import@t(Customer)	/导入2015年的客户
3	=A1 | A2	/使用符号“|”将两年的客户合并。值得注意的是，因为要统计次数，重复的客户也要保留，所以要求和列。
4	=A3.groups(Customer; count(~):Count)	/统计每个客户的销售次数

A4	Product	Count
	ANATR	5
	ANTON	6
	…	…

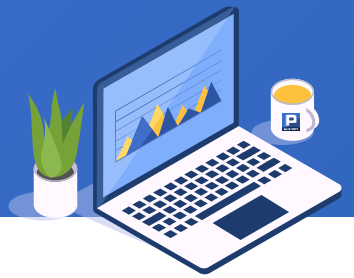
2. 交



统计有哪些同学同时报名了绘画班和舞蹈班。兴趣班报名表结构相同，如下：

ID	StudentID	Subject
1	2	Painting
2	4	Dance
3	3	Robot
4	2	Dance
5	5	Writing
…	…	…

2. 交

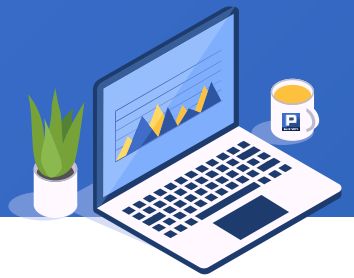


SPL如下，其中用到了符号“^” 求交列：

	A	B
1	=file("Interest.txt").import@t()	/从文件中导入兴趣班报名表
2	=A1.select(Subject:"Painting")	/选出报名绘画的记录
3	=A1.select(Subject:"Dance")	/选出报名舞蹈的记录
4	=A2.(StudentID) ^ A3.(StudentID)	/使用符号“^”求报名绘画和舞蹈的同学的交列

A4	Member
	2
	8
	11
	…

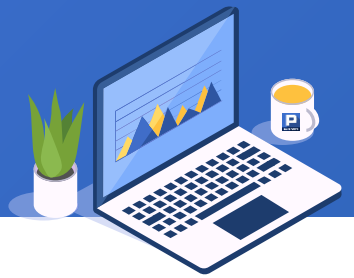
3. 并



统计绘画班和舞蹈班共有哪些同学。兴趣班报名表结构相同，如下：

ID	StudentID	Subject
1	2	Painting
2	4	Dance
3	3	Robot
4	2	Dance
5	5	Writing
…	…	…

3. 并

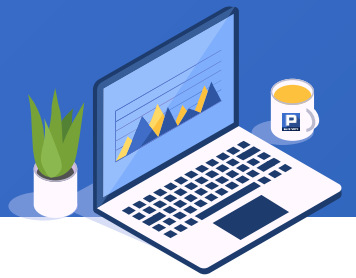


SPL如下，其中用到了符号“&” 求并列：

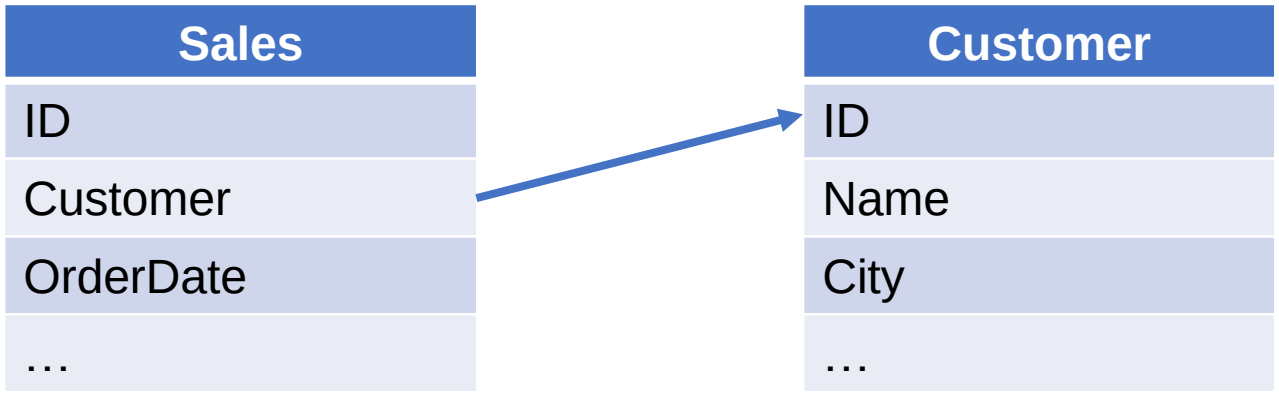
	A	B
1	=file("Interest.txt").import@t()	/从文件中导入兴趣班报名表
2	=A1.select(Subject:"Painting")	/选出报名绘画的记录
3	=A1.select(Subject:"Dance")	/选出报名舞蹈的记录
4	=A2.(StudentID) & A3.(StudentID)	/使用符号“&”求报名绘画和舞蹈的同学的并列

A4	Member
	2
	4
	8
	…

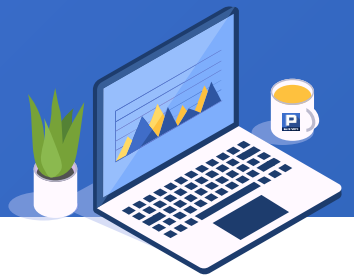
4. 差



有销售表和客户表，查询2014年的新增客户，即销售客户不在客户表中的。



4. 差



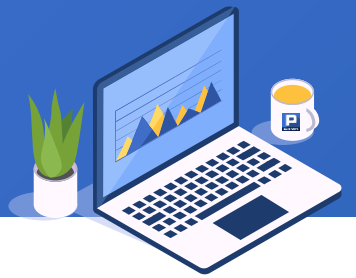
SPL如下，其中用到了符号“\” 求差列：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from Sales where year(OrderDate)=2014")	/查询2014年的销售记录
3	=A1.query("select * from Customer")	/查询客户表
4	=A2.id(Customer)	/使用id函数去重，取客户的唯一值序列
5	=A3.(ID)	/取出客户表中的客户ID序列
6	=A4\A5	/使用符号“\” 求差列

A6	Members
	DOS
	HUN
	URL

注意：本例只是为了介绍差列，使用函数 A.switch@d() / A.join@d() 进行连接过滤更加简便。

5. 异或



成绩表按学期保存在不同的文件中，要查询上下学期只有一次进入总分前十名的学生ID。

CLASS	STUDENTID	SUBJECT	SCORE
Class one	1	English	84
Class one	1	Math	77
Class one	1	PE	69
Class one	2	English	81
Class one	2	Math	80
…	…	…	…

5. 异或

SPL如下，其中用到了符号“%” 求异或列：

	A	B
1	=file("Scores1.csv").import@ct()	/导入学生上学期成绩
2	=file("Scores2.csv").import@ct()	/导入学生下学期成绩
3	=A1.groups(STUDENTID; sum(SCORE):Score)	/分组汇总上学期学生总成绩
4	=A2.groups(STUDENTID; sum(SCORE):Score)	/分组汇总下学期学生总成绩
5	=A3.top(-10;Score).(STUDENTID)	/选出上学期总分前十名的学生ID
6	=A4.top(-10;Score).(STUDENTID)	/选出下学期总分前十名的学生ID
7	=A5%A6	/选出上下学期的学生ID不重复的记录。

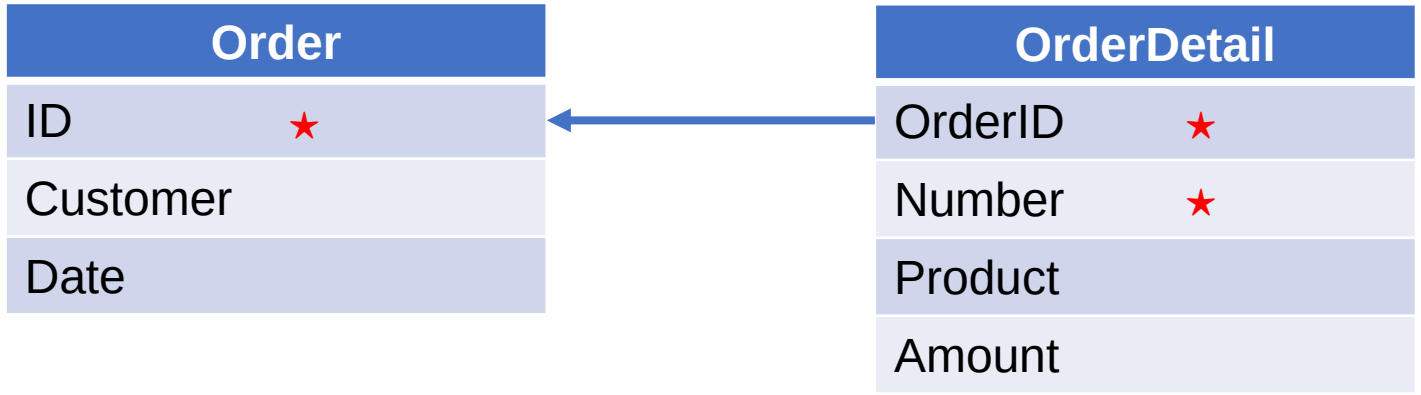
A5	Member
	2
	9
	4
	10
	…

A6	Member
	12
	1
	8
	4
	…

A7	Member
	2
	9
	10
	7
	…

6. 多个集合间的运算：和列

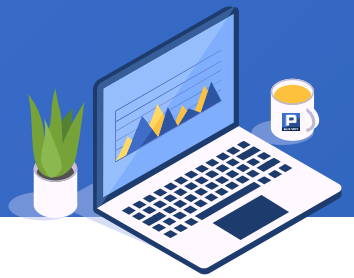
订单表和订单明细表是主子表关系，每个订单有多条明细数据。如下图：



订单明细表中每个订单的明细数据是不定长的。想要查询出如下表格：

ID	Customer	Date	Product1	Amount1	Product2	Amount2	Product3	Amount3
1	3	20190101	Apple	5	Milk	3	Salt	1
2	5	20190102	Beef	2	Pork	4		
3	2	20190102	Pizza	3				

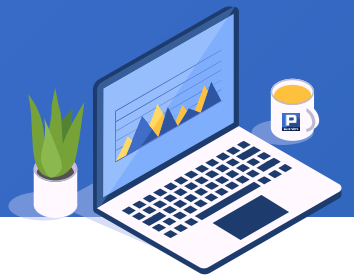
6. 多个集合间的运算：和列



SPL如下，其中用到了A.conj()函数合并序列成员：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from OrderDetail left join Order on Order.ID=OrderDetail.OrderID")	/导入订单明细表和订单表，并按订单ID连接订单表
3	=A2.group(ID)	/将取出的数据按订单ID分组
4	=A3.max(~.count()).("Product"+string(~)+","+"Amount"+string(~)).concat@c()	/找到分组后成员最多的一组确定目标表格数据结构
5	=create(ID,Customer,Date,${A4})	/根据A4确定的数据结构创建序表
6	>A3.run(A5.record([ID,Customer,Date]|~.([Product,Amount]).conj()))	/循环分组数据，每个分组内将成员拼到一个序列，这里用到了conj函数取每组各个产品和数量的和列。最后把生成的记录插入到A5创建的序表中。

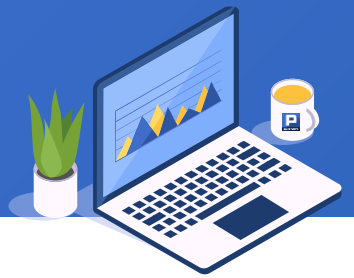
6. 多个集合间的运算：和列



下面是某时刻，新冠状病毒世界各地确诊人数的JSON数据，要统计世界确诊人数。

```
[
{Region:"China",Confirmed:[
        {Region:"Hubei",Confirmed:[
                {Region:"Wuhan",Confirmed:51986},
                {Region:"Xiaogan",Confirmed:3009},
                {Region:"Huanggang",Confirmed:3791},
                ...]
        },
        {Region:"Taiwan",Confirmed:18},
        ...]
},
{Region:"Thailand",Confirmed:33},
...]
```

6. 多个集合间的运算：和列



SPL如下，其中用到了A.conj@r()函数来递归合并序列成员：

	A	B
1	=json(file("COVID-19.json").read())	/导入JSON数据文件
2	=A1.field@r("Confirmed")	/使用A.field()函数的@r选项递归获取所有确诊字段
3	=A2.conj@r()	/使用A.conj()函数的@r选项递归合并
4	=A3.sum()	/确诊人数求和

A2

Member
[[51986,3009,3791],[58,…]
251
33
28
19
16
…

A3

Member
51986
3009
3791
1447
1206
1125
…

A4

Member
64438

6. 多个集合间的运算：并列和差列

有课程表和选课表，查询有哪些课没有学生选修。其中选课表可以多选，用逗号分隔，部分数据如下：

Course		
ID	NAME	TEACHERID
1	Environmental protection and …	5
2	Mental health of College Students	1
3	Computer language Matlab	8
4	Electromechanical basic practice	7
5	Introduction to modern life science	3
6	Modern wireless communication system	14
…	…	…

SelectCourse		
ID	STUDENTID	COURSE
1	59	2,7
2	43	1,8
3	52	2,7,10
4	44	1,10
5	37	5,6
6	57	3
…	…	…

6. 多个集合间的运算：并列和差列

使用了A.union()函数求序列的序列成员的并列，使用了A.diff()函数求序列的序列成员的差列， SPL如下：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from Course")	/读取课程表
3	=A1.query("select * from SelectCourse")	/读取学生选课表
4	=A3.union(COURSE.split@cp())	/将选课表中的课程按逗号拆分后，使用union()函数将课程序列求交列
5	=A2.(ID)	/所有课程的序号
6	=A2(A5.pos([A5,A4].diff()))	/使用diff()函数求课程表和选课表的课程序号的差列，即没有学生选择的课程，在A5中定位后，从A2中选出。

A6

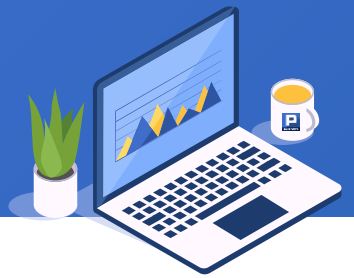
ID	NAME	TEACHERID
1	Fundamentals of economic management	21

6. 多个集合间的运算：交列

销售表部分数据如下，统计出2014年每个月销售金额均排在前20名的客户名称。

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

6. 多个集合间的运算：交列



SPL如下，其中用到了A.isect()函数求成员交集：

	A	B
1	=connect("db").query("select * from sales")	/连接数据源，读取销售表
2	=A1.select(year(OrderDate)==2014)	/选出2014年数据
3	=A2.group(month(OrderDate))	/使用group函数，将2014年的数据按照月份分组
4	=A3.(~.group(Customer))	/分组后的成员按照客户分组
5	=A4.(~.top(-20;sum(Amount)))	/循环每个月的数据，计算每月销售额前20的客户
6	=A5.(~.(Customer))	/列出了销售额前20名客户名称
7	=A6.isect()	/使用isect()函数求每组之间的交集

A7	Member
	HANAR
	SAVEA

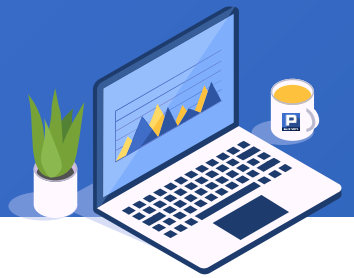
6. 多个集合间的运算：交列

还可以使用A.isect(x)函数求经过 x 运算后的成员交集：

	A	B
1	=connect("db").query("select * from sales")	/连接数据源，读取销售表
2	=A1.select(year(OrderDate)==2014)	/选出2014年数据
3	=A2.group(month(OrderDate))	/使用group函数，将2014年的数据按照月份分组
4	=A3.(~.group(Customer))	/分组后的成员按照客户分组
5	=A4.(~.top(-20;sum(Amount)))	/循环每个月的数据，计算每月销售额前20的客户
6	=A5.isect(~.(Customer))	/每组取出客户名称，使用isect()函数求每组之间的交集

A6	Member
	HANAR
	SAVEA

6. 多个集合间的运算：异或列



查询2014年只有一次进入月销售额前三的客户。销售表如下：

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

6. 多个集合间的运算：异或列

SPL如下，其中用到了A.xunion() 函数将序列中的序列成员之间不重复的成员选出：

	A	B
1	=file("Sales.csv").import@ct()	/导入销售表
2	=A1.select(year(OrderDate)==2014).group(month(OrderDate))	/选出2014年的记录，并按月分组
3	=A2.(~.groups(Customer; sum(Amount):Amount))	/分组汇总每个客户的总销售额
4	=A3.(~.top(-3;Amount).(Customer))	/选出每个月销售额前三的客户
5	=A4.xunion()	/使用xunion函数，选出每个月只出现过一次的客户

A5	Member
	KOENE
	HANAR
	RATTC
	BOTTM
	…

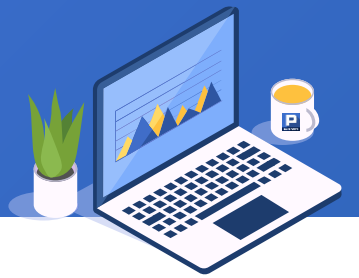
CONTENTS

1. 直接比对记录引用
2. 有序归并比对字段
3. 有序归并比对主键
4. 有序归并比对所有字段
5. 比对字段无序

02

集合成员为记录

1. 直接比对记录引用



销售表部分数据如下，查看2014年每个月单笔销售额排在前3名的销售记录。

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

1. 直接比对记录引用

使用 A.conj() 函数将各个序表中的记录合并成一个序表。SPL如下：

	A	B
1	=connect("db")	/连接数据源
2	=A1.query("select * from Sales")	/读取销售表
3	=A2.select(year(OrderDate)==2014)	/选出2014年的记录
4	=A3.groups(month(OrderDate):Month; top(-3;Amount):Top3)	/按月份分组，选出每月销售额前三的记录
5	=A4.conj(Top3)	/使用conj函数将前三的记录拼成序表返回

A5	OrderID	Customer	SellerId	OrderDate	Amount
	10424	MEREP	7	2014/01/23	11493.2
	10417	SIMOB	4	2014/01/16	11283.2
	10430	ERNSH	4	2014/01/30	5796.0
	…	…	…	…	…

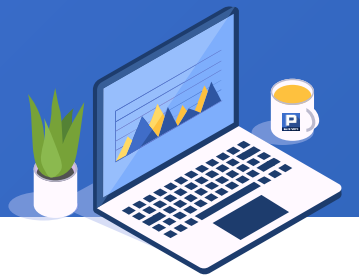
1. 直接比对记录引用

还可以使用 A.merge() 函数的@o选项，无@u/@i/@d/@x选项时，与A.conj() 函数作用相同，将各个序表中的记录合并成一个序表。 SPL如下：

	A	B
1	=connect("db")	/连接数据源
2	=A1.query("select * from Sales")	/读取销售表
3	=A2.select(year(OrderDate)==2014)	/选出2014年的记录
4	=A3.groups(month(OrderDate):Month; top(-3;Amount):Top3)	/按月份分组，选出每月销售额前三的记录
5	=A4.merge@o(Top3)	/使用A.merge@o()函数将前三的记录拼成序表返回

A5	OrderID	Customer	SellerId	OrderDate	Amount
	10424	MEREP	7	2014/01/23	11493.2
	10417	SIMOB	4	2014/01/16	11283.2
	10430	ERNSH	4	2014/01/30	5796.0
	…	…	…	…	…

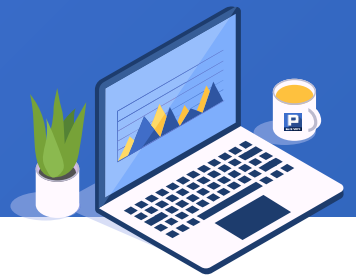
1. 直接比对记录引用



某公司要对年龄低于30岁或者入职年限小于3年的新员工进行培训，请选出这些员工记录。员工表部分数据如下：

ID	NAME	BIRTHDAY	HIREDATE	DEPT
1	Rebecca	1974/11/20	2005/03/11	R&D
2	Ashley	1980/07/19	2008/03/16	Finance
3	Rachel	1970/12/17	2010/12/01	Sales
4	Emily	1985/03/07	2006/08/15	HR
…	…	…	…	…

1. 直接比对记录引用

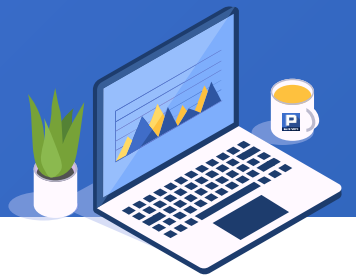


SPL如下，其中用到了 A.union() 函数将各个序表中的记录取并集，返回一个排列：

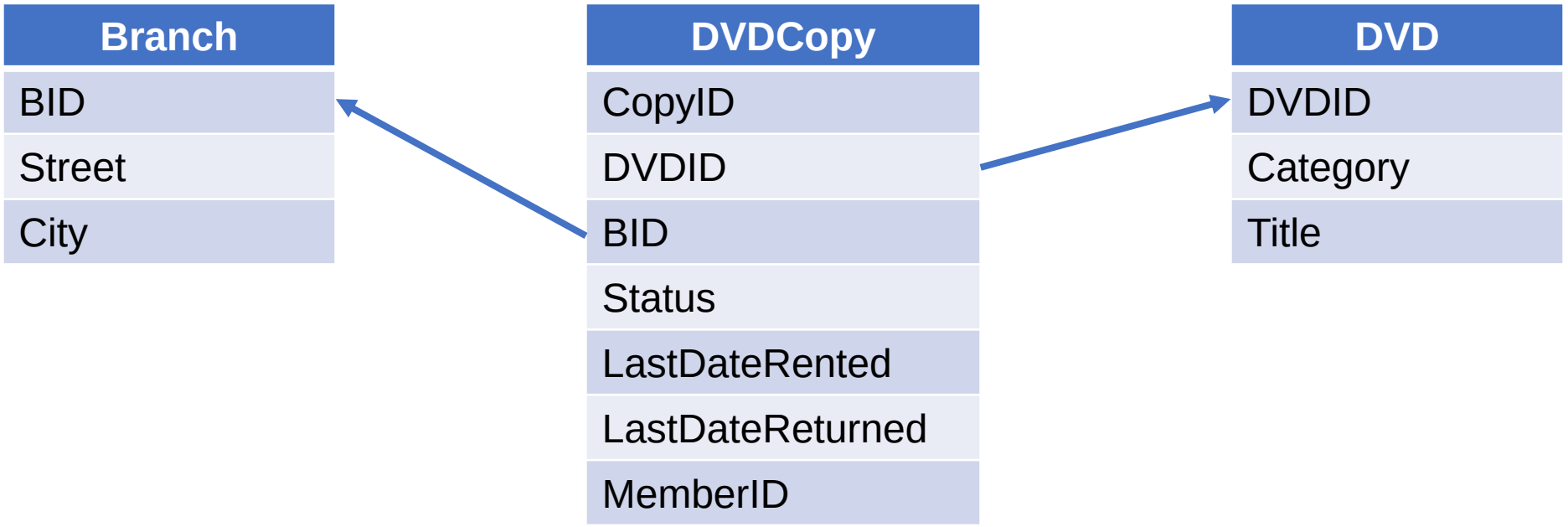
	A	B
1	=connect("db")	/连接数据源
2	=A1.query("select * from Employee")	/读取员工表
3	=A2.select(age(BIRTHDAY) < 30)	/选出年龄小于30的员工
4	=A2.select(age(HIREDATE) < 3)	/选出入职不满3年的员工
5	=[A3,A4].union()	/使用union函数将员工取并集拼成序表返回

A5	ID	NAME	BIRTHDAY	HIREDATE	DEPT
	89	Emily	1990/12/09	2017/02/01	Technology
	241	Samantha	1991/12/04	2016/01/01	Finance
	393	Hannah	1990/09/06	2016/01/01	Sales
	…	…	…	…	…

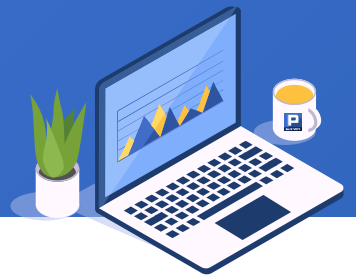
1. 直接比对记录引用



查询出缺货的DVD分店，即现存的DVD拷贝不到4类的分店。其中Branch表，存储DVD分店信息；DVD表，存储DVD的标题及分类信息；DVDCopy表，存储DVD的多张拷贝，DVD拷贝是真正的光盘，以实体形式存放于各个分店。



1. 直接比对记录引用

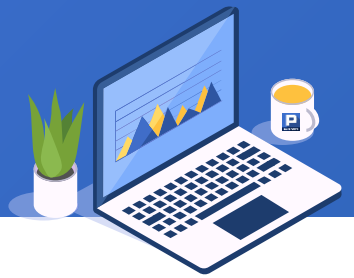


SPL如下，其中用到了符号“|” 求和列，用到了符号“\” 求差列：

	A	B
1	=connect("db")	/连接数据源
2	=Branch=A1.query("select * from Branch")	/读取分店信息，并定义为Branch变量
3	=DVD=A1.query("select * from DVD")	/读取DVD信息，并定义为DVD变量
4	=DVDCopy=A1.query("select * from DVDCopy")	/读取DVDCopy信息，并定义为DVDCopy变量
5	=DVDCopy.switch(DVDID,DVD:DVDID; BID,Branch:BID)	/将DVDCopy的DVDID字段切换成DVD中对应的记录
6	=DVDCopy.select(STATUS!="Miss" && LASTDATERETURNED!=null)	/过滤丢失的和未归还的DVD拷贝
7	=A6.group(BID)	/对过滤后的数据按照BID分组
8	=A7.select(~.icount(DVDID.CATEGORY)<4)	/选出DVD拷贝小于4类的门店
9	=A8.(BID) | (Branch \ A7.(BID))	/缺货的门店。其中A8.(BID)表示DVD拷贝小于4类的门店，Branch \ A7.(BID)表示DVDCopy未出现过的门店。

A9	BID	STREET	CITY
	B002	Street2	Houston
	B003	Street3	LA
	B004	Street4	Lincoln

2. 有序归并比对字段



学生的数学成绩和英语成绩分别存放在 Math.txt 和 English.txt 两个文件中。统计每位学生的总分。成绩表结构相同，如下：

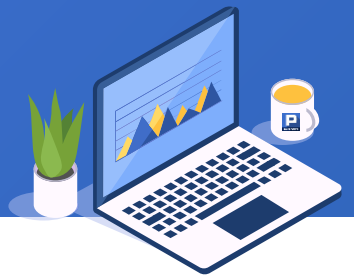
数学：

CLASS	STUDENTID	SUBJECT	SCORE
1	1	Math	77
1	2	Math	80
…	…	…	…

英语：

CLASS	STUDENTID	SUBJECT	SCORE
1	1	English	84
1	2	English	81
…	…	…	…

2. 有序归并比对字段

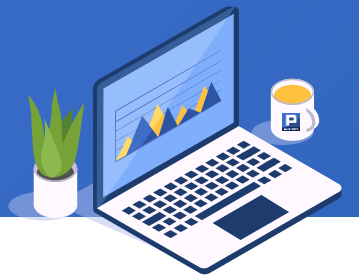


SPL如下，其中用到了 A.merge(xi, …) 函数将多个序表按照 xi, … 有序合并：

	A	B
1	=file("Math.txt").import@t()	/读取数学成绩表
2	=file("English.txt").import@t()	/读取英语成绩表
3	=A1.sort(CLASS,STUDENTID)	/数学成绩表按班级和学生排序
4	=A2.sort(CLASS,STUDENTID)	/英语成绩表按班级和学生排序
5	=[A3,A4].merge(CLASS,STUDENTID)	/使用merge函数，按班级和学生字段有序归并
6	=A5.groups@o(CLASS,STUDENTID; ~.sum(SCORE):TOTALSCORE)	/使用groups函数的@o选项，相邻值变化时重新分组。并统计每个学生的总分。

A6	CLASS	STUDENTID	TOTALSCORE
	1	1	161
	1	2	161
	1	3	159
	…	…	…

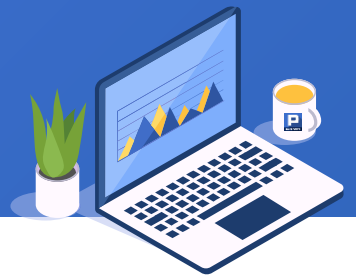
2. 有序归并比对字段



某商家根据销售渠道不同，销售记录分别存储在线上和实体店两个表。有时线上线下同时搞活动，部分销售记录被同时存储在两个表中。求商家实际的总销售额。销售表结构相同，如下：

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

2. 有序归并比对字段

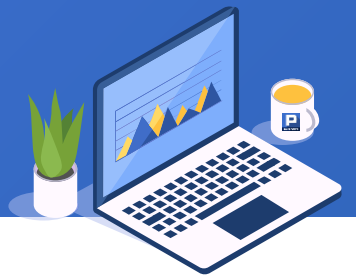


SPL如下，其中用到了 A.merge(xi, …) 函数的@u选项，有序归并时去掉重复的成员：

	A	B
1	=file("Online.txt").import@t()	/读取线上销售表
2	=file("Store.txt").import@t()	/读取实体店销售表
3	=A1.sort(OrderID)	/线上销售表按订单ID排序
4	=A2.sort(OrderID)	/实体店销售表按订单ID排序
5	=[A3,A4].merge@u(OrderID)	/使用merge函数的@u选项，两表按订单ID有序归并，删除重复的记录
6	=A5.sum(Amount)	/汇总销售额

A6	Member
	678756.41

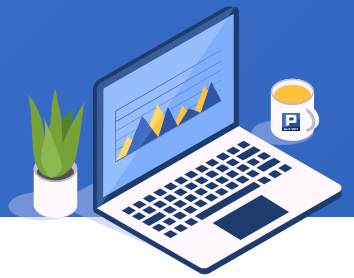
2. 有序归并比对字段



同样的例子，现在想要查询线上线下重复保存的销售记录有多少条。销售表结构相同，如下：

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

2. 有序归并比对字段

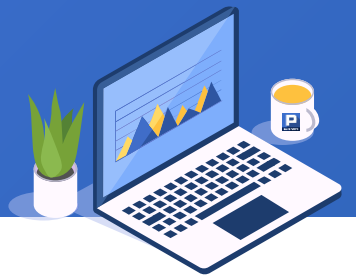


SPL如下，其中用到了 A.merge(xi, ...) 函数的@i选项，返回A(i)共同的成员组成的序表：

	A	B
1	=file("Online.txt").import@t()	/读取线上销售表
2	=file("Store.txt").import@t()	/读取实体店销售表
3	=A1.sort(OrderID)	/线上销售表按订单ID排序
4	=A2.sort(OrderID)	/实体店销售表按订单ID排序
5	=[A3,A4].merge@i(OrderID)	/使用merge函数的@i选项，两表按订单ID有序归并，返回共同的成员
6	=A5.count()	/统计共同订单的数量

A6	Member
	70

2. 有序归并比对字段



2015年3月的交易信息存储文件中，早一点的是old.csv中，晚一点的是new.csv。文件中的逻辑主键是UserName和Date，需要分别找出新增的、删除的、修改的记录。

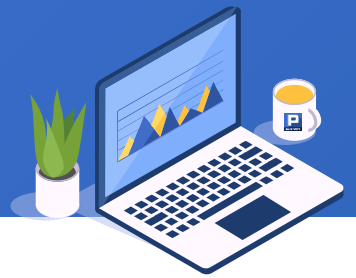
old.csv

UserName	Date	SaleValue	SaleCount
Rachel	2015-03-01	4500	9
Rachel	2015-03-03	8700	4
Tom	2015-03-02	3000	8
Tom	2015-03-03	5000	7
Tom	2015-03-04	6000	12
John	2015-03-02	4000	3
John	2015-03-02	4300	9
John	2015-03-04	4800	4

new.csv

UserName	Date	SaleValue	SaleCount
Rachel	2015-03-01	4500	9
Rachel	2015-03-02	5000	5
Ashley	2015-03-01	6000	5
Rachel	2015-03-03	11700	4
Tom	2015-03-03	5000	7
Tom	2015-03-04	6000	12
John	2015-03-02	4000	3
John	2015-03-02	4300	9
John	2015-03-04	4800	4

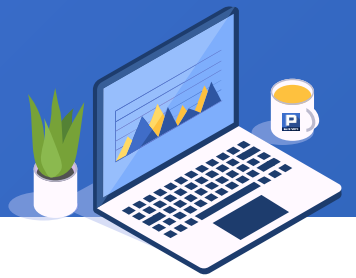
2. 有序归并比对字段



SPL如下，其中用到了 A.merge(xi, ...) 函数的@d选项，从A(1)中去掉A(2) &...A(n)中的成员后形成新的序表：

	A	B
1	=file("old.csv").import@ct()	/读取早一点的数据表
2	=file("new.csv").import@ct()	/读取晚一点的数据表
3	=A1.sort(UserName,Date)	/数据表按用户名和日期排序
4	=A2.sort(UserName,Date)	/数据表按用户名和日期排序
5	=new=[A4,A3].merge@d(UserName,Date)	/使用merge函数的@d选项，有序归并时从A4中去掉A3中包含的记录，剩下的是新增记录组成的序表
6	=delete=[A3,A4].merge@d(UserName,Date)	/使用merge函数的@d选项，有序归并时从A3中去掉A4中包含的记录，剩下的是删除记录组成的序表
7	=diff=[A4,A3].merge@d(UserName,Date,SaleValue,SaleCount)	/使用merge函数的@d选项，有序归并时从A4中去掉A3中发生变化（有字段值不同）的记录
8	=update=[diff,new].merge@d(UserName,Date)	/使用merge函数的@d选项，有序归并时从发生变化的记录中去掉新增记录，剩下的是更新记录组成的序表
9	return [new, delete, update]	/返回序列，成员分别是新增、删除和更新记录组成的序表

2. 有序归并比对字段



A9
Members
[[Ashley,2015-03-01,6000,5], …]
[[Tom,2015-03-02,3000,8]]
[[Rachel,2015-03-03,11700,4]]

new

UserName	Date	SaleValue	SaleCount
Ashley	2015-03-01	6000	5
Rachel	2015-03-02	5000	5

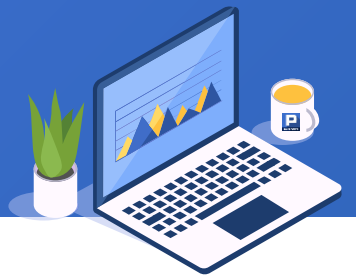
delete

UserName	Date	SaleValue	SaleCount
Tom	2015-03-02	3000	8

update

UserName	Date	SaleValue	SaleCount
Rachel	2015-03-03	11700	4

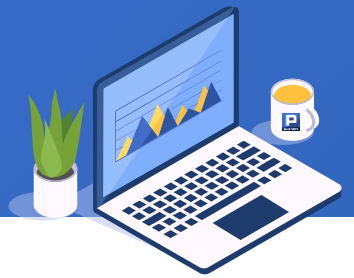
2. 有序归并比对字段



下面是随机抽样后生成的文件，比较两次抽样后的文件选出了多少不同的序号。文件数据结构相同，如下：

ID	Predicted_Y	Original_Y
10	0.012388464367608093	0.0
11	0.0151989912397898 8	0.0
13	0.0007920238885061248	0.0
19	0.0012656367468159102	0.0
21	0.009460545997473379	0.0
23	0.024176791871681664	0.0
…	…	…

2. 有序归并比对字段

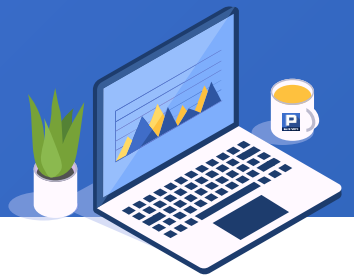


SPL如下，其中用到了 A.merge(xi, …) 函数的@x选项，返回A(i)去掉共同的成员组成的序表：

	A	B
1	=file("p1.txt").import@t()	/读取第一个抽样文件
2	=file("p2.txt").import@t()	/读取第二个抽样文件
3	=A1.sort(ID)	/第一个文件按ID排序
4	=A2.sort(ID)	/第二个文件按ID排序
5	=[A3,A4].merge@x(ID)	/使用merge函数的@x选项，按ID有序归并，返回序号不同的记录。
6	=A5.len()	/返回不同序号的个数

A6	Member
	458

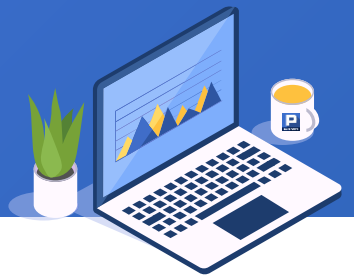
3. 有序归并比对主键



疫情期间，学生每天需要上报体温。查询6月1日到20日期间，连续发烧3日及以上的同学。文件名是日期，例如6月1日的文件是601.txt，文件数据结构相同，如下：

StudentID	Name	Fever
10	Ryan	0
5	Ashley	0
13	Daniel	1
19	Samantha	0
1	Rebecca	0
…	…	…

3. 有序归并比对主键

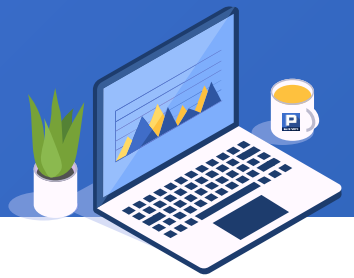


SPL如下，其中用到了 A.merge() 函数， 当 A(i) 有主键时，有序归并比对主键：

	A	B
1	=to(601, 620)	/创建表名序列
2	=A1.(file(string(~)+".txt").import@t())	/循环导入6月1日到20日的文件
3	=A2.(~.keys(StudentID).sort(StudentID))	/设置学生ID为主键，并按学生ID排序
4	=A3.merge()	/使用merge函数有序归并比对主键。
5	=A4.group@o(StudentID,Fever)	/使用group函数的o选项，字段值发生变化时重新分组
6	=A5.select(~.Fever==1 && ~.len()>=3).id(Name)	/选出连续发烧3天的学生姓名

A6	Name
	Ashley
	Rachel

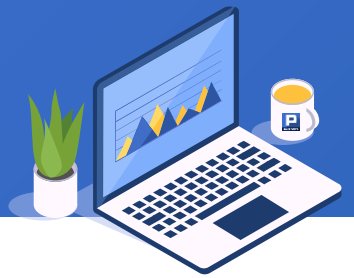
4. 有序归并比对所有字段



比较两个文件 p1.csv 和 p2.csv 有多少行数据有差异。文件数据结构相同，如下：

ID	Predicted_Y	Original_Y
10	0.012388464367608093	0.0
11	0.01519899123978988	0.0
13	0.0007920238885061248	0.0
19	0.0012656367468159102	0.0
21	0.00946054599747337 9	0.0
23	0.024176791871681664	0.0
…	…	…

4. 有序归并比对所有字段

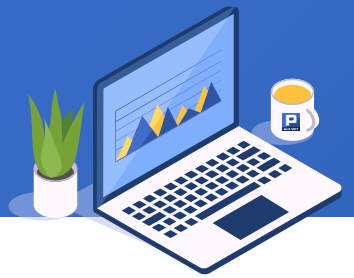


SPL如下，其中用到了 A.merge() 函数，当 A(i) 无主键时，有序归并比对所有字段：

	A	B
1	=file("p1.txt").import@t()	/读取第一个抽样文件
2	=file("p2.txt").import@t()	/读取第二个抽样文件
3	=[A1,A2].merge@x()	/使用merge函数有序归并比对主键。这里使用了@x选项，返回不同的主键，即序号不同的行。
4	=A3.len()	/返回不同的行数

A4	Member
	458

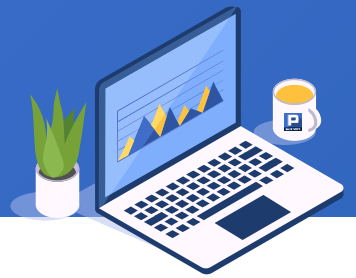
5. 比对字段无序



某公司的销售数据，部分存储在旧数据库db1中，部分存储在新数据库db2中。求2014年的总销售额。销售表结构相同，如下：

OrderID	Customer	SellerId	OrderDate	Amount
10426	GALED	4	2014/01/27	338.2
10676	TORTU	2	2014/09/22	534.85
10390	ERNSH	6	2013/12/23	2275.2
10400	EASTC	1	2014/01/01	3063.0
10464	FURIB	4	2014/03/04	1848.0
…	…	…	…	…

5. 比对字段无序



SPL如下，其中用到了 A.merge(xi, …) 函数的@o选项，不假定A(i)对[xi,…]有序：

	A	B
1	=connect("db1").query("select * from Sales")	/从db1中读取销售表
2	=connect("db2").query("select * from Sales")	/从db2中读取销售表
3	=[A1,A2].merge@ou(OrderID)	/使用merge函数按订单ID有序归并。使用了@o选项，销售表不保证按订单ID有序。使用了@u选项，去掉ID重复的记录。
4	=A3.select(year(OrderDate)==2014)	/选出2014年的记录
5	=A4.sum(Amount)	/统计2014年的总销售额

A5	Member
	723388.75

注意：前面提到过，A.merge()函数的o选项单独使用时类似于A.conj()函数。

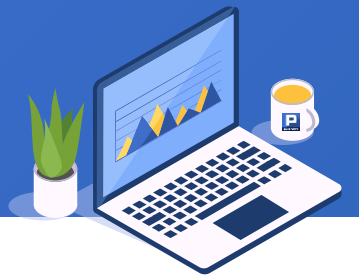
更常见的做法是本例这样，@o选项与@u/@i/@d/@x选项搭配使用。

CONTENTS

1. 简单和列
2. 有序归并比对列值
3. 隐含归并比对维字段

大数据量下的集合间运算

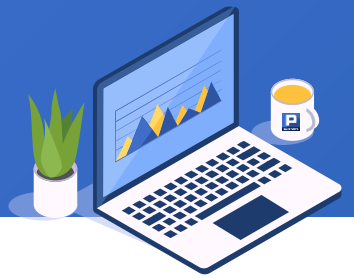
1. 简单和列



查询每个月销售额最高的记录。销售表数据量大，无法加载到内存，如下：

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

1. 简单和列

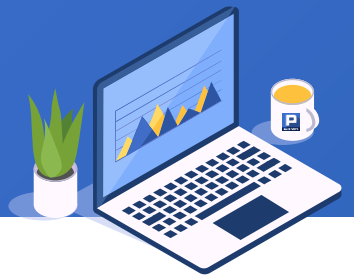


SPL如下，其中用到了 cs.group(x, ...) 针对游标记录做相邻值分组，返回原游标：

	A	B
1	=connect("db").query("select * from Sales order by OrderDate")	/从数据库中读取销售表，按销售日期排序
2	=A1.group(month(OrderDate))	/使用cs.group()函数比较相邻月份分组
3	=A2.(~.maxp(Amount))	/选出每月销售额最高的记录
4	=A3.conj()	/返回每月销售额最高的记录的和列
5	=A4.fetch()	/从游标中取数，此时数据集较小

A5	OrderID	Customer	SellerId	OrderDate	Amount
	10267	FRANK	4	2013/07/29	4031.0
	10286	QUICK	8	2013/08/21	3016.0
	...	...	...	...	...

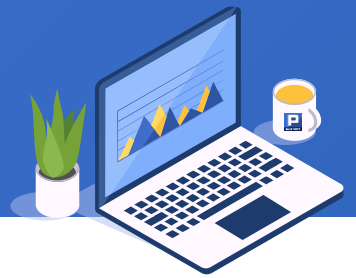
1. 简单和列



2014年和2015年的销售记录，分别存储在表S2014和S2015中。求这两年来总销售额前三的客户。销售表结构相同，数据量大无法加载到内存，如下：

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

1. 简单和列

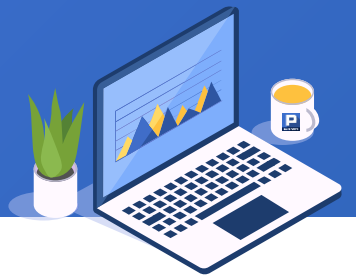


SPL如下，其中用到了 CS.conjx() 函数将多个游标纵向连接，相当于合并游标中的数据：

	A	B
1	=connect("db")	/连接数据库
2	=A1. cursor("select * from S2014")	/获取2014年销售表游标
3	=A1. cursor("select * from S2015")	/获取2015年销售表游标
4	=[A2,A3].conjx()	/使用CS.joinx()函数，将多个游标合并
5	=A4.groups(Customer; sum(Amount):Amount)	/对合并后的游标分组汇总，统计每个客户的总销售额
6	=A5.top(-3;Amount)	/选出两年来总销售额前三的客户

A6	Customer	Amount
	SAVEA	177478.89
	QUICK	102764.99
	ERNSH	94066.28

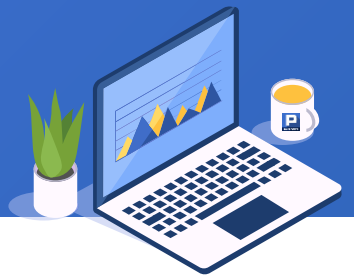
2. 有序归并比对列值



某公司的销售数据，部分存储在旧数据库db1中，部分存储在新数据库db2中。统计2014年每个月的销售数量。销售表结构相同，数据量大无法加载到内存，如下：

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

2. 有序归并比对列值

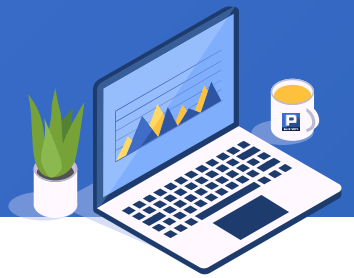


使用 CS.mergex(xi, …) 函数，针对游标序列成员做归并运算。SPL如下：

	A	B
1	=connect("db1").cursor("select * from Sales order by OrderDate")	/从db1中读取销售表，按订单日期排序
2	=connect("db2").cursor("select * from Sales order by OrderDate")	/从db2中读取销售表，按订单日期排序
3	=[A1,A2].mergex(OrderDate)	/使用mergex函数将游标按订单日期归并
4	=A3.select(year(OrderDate)==2014)	/选出2014年的记录
5	=A4.groups@o(month(OrderDate):Month; count(~):Count)	/使用groups函数分组汇总统计每个月的销售数量。使用了@o选项，月份发生变化时重新分组

A5	Month	Count
	1	33
	2	29
	…	…

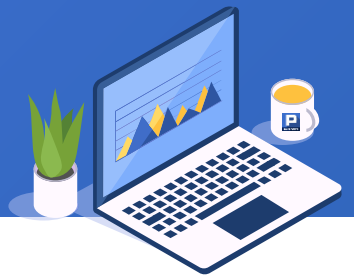
2. 有序归并比对列值



同样的例子，假设两个表中可能有重复数据。统计2014年每个客户的销售额。销售表如下：

OrderID	Customer	SellerId	OrderDate	Amount
10400	EASTC	1	2014/01/01	3063.0
10401	HANAR	1	2014/01/01	3868.6
10402	ERNSH	8	2014/01/02	2713.5
10403	ERNSH	4	2014/01/03	1005.9
10404	MAGAA	2	2014/01/03	1675.0
…	…	…	…	…

2. 有序归并比对列值



函数 CS.mergex(xi, ...) 也支持@u/@i/@d/@x等选项，与 A.merge() 函数的选项用法类似。SPL如下：

	A	B
1	=connect("db1").cursor("select * from Sales order by OrderID")	/从db1中读取销售表，按订单ID排序
2	=connect("db2").cursor("select * from Sales order by OrderID")	/从db2中读取销售表，按订单ID排序
3	=[A1,A2].mergex@u(OrderID)	/使用mergex函数将游标按订单ID归并，使用了@u选项去掉重复记录
4	=A3.select(year(OrderDate)==2014)	/选出2014年的记录
5	=A4.groups(Customer; sum(Amount):Amount)	/使用groups函数分组汇总统计每个客户的销售额

A5	Customer	Amount
	ANATR	1129.75
	ANTON	6452.15
	…	…

THANKS

感谢观看