数据分析编程从 SQL 到 SPL：人口和语言分析

数据结构

国家表 world.country

Index	Code	Name	Population
1	ABW	Aruba	103000
2	AFG	Afghanistan	22720000
3	AGO	Angola	12878000
4	AIA	Anguilla	8000
5	ALB	Albania	3401200
6	AND	Andorra	78000
7	ANT	Netherlands Antilles	217000
8	ARE	United Arab Emirates	2441000
9	ARG	Argentina	37032000
10	ARM	Armenia	3520000

Code 是国家编码，Name 是国家名称，Population 是国家人口。

国家语言表 world.countrylanguage

Index	CountryCode	Language	IsOfficial	Percentage
1	ABW	Dutch	T	5.3
2	ABW	English	F	9.5
3	ABW	Papiamento	F	76.7
4	ABW	Spanish	F	7.4
5	AFG	Balochi	F	0.9
6	AFG	Dari	T	32.1
7	AFG	Pashto	T	52.4
8	AFG	Turkmenian	F	1.9
9	AFG	Uzbek	F	8.8
10	AGO	Ambo	F	2.4

CountryCode 是国家编码，Language 是语言，IsOfficial 是否官方语言（T 代表是官方语言），Percentage 是使用该语言的人口百分比。

1. 查询官方语言最多的国家、人口及官方语言数

通过示例数据可以看到，每个国家使用的语言可能有多种，其中有官方语言也有非官方语言。

SQL 语句如下:

```
with t1 as (
    select CountryCode, count(*) Num
    from world.countrylanguage
    where isOfficial='T'
    group by CountryCode),
t2 as (
    select *
    from t1
    where Num=(select Max(Num) from t1))
select Name,Population,Num
    from world.country c join t2
    on c.Code=t2.CountryCode;
```

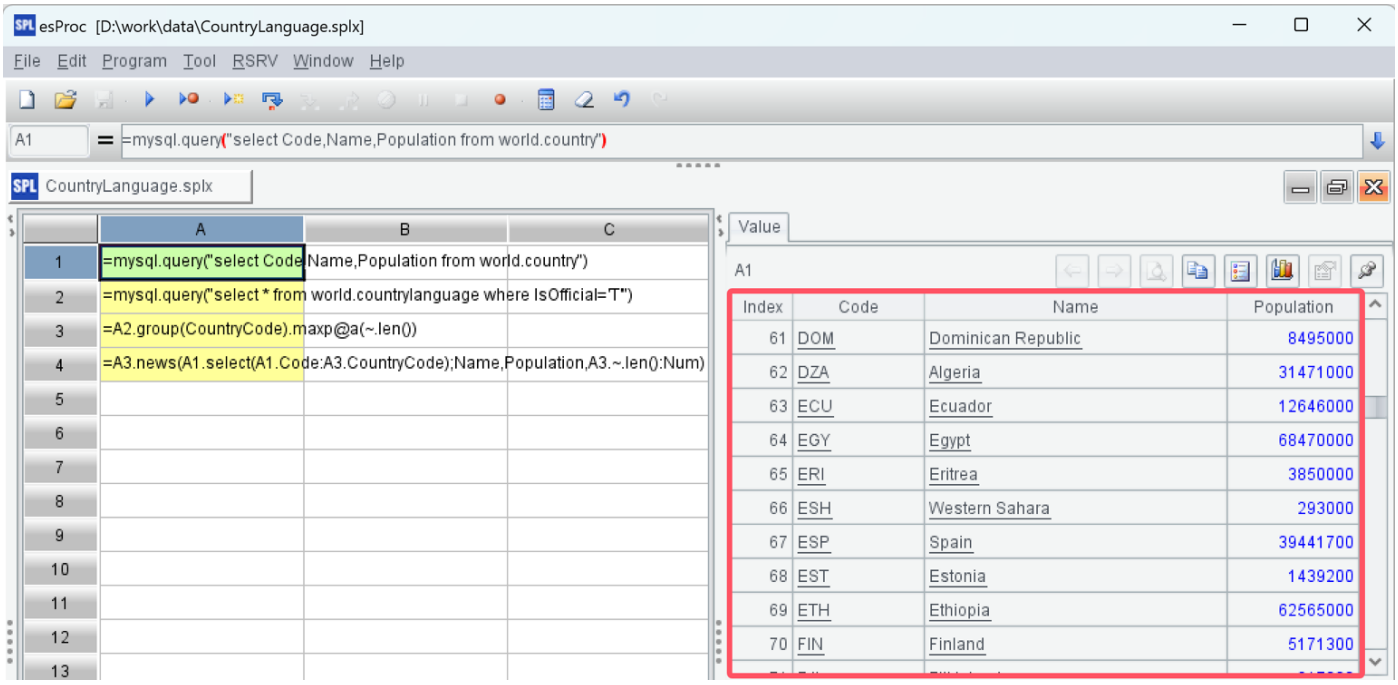
SQL 语句在使用了窗口函数后仍要嵌套多层，比较繁琐。主要是因为 SQL 的分组功能总是和汇总同时出现，而且支持的聚合函数有限。

这个问题其实很简单，将国家语言表按照国家分组，选出语言数量最多的组的所有记录，再从国家表中找到对应的国家名称和人口数量即可。

SPL 支持独立的分组运算，这样更方便分步运算，还有更多的聚合方式。SPL 脚本如下:

A1: 读取国家表。

SPL 的 IDE 有很好的交互性，可以执行后在右边的值面板中直观地查看到每一步的结果：



A2: 读取国家语言表，选出官方语言。

A3: 按国家编号分组，并取出成员最多的所有组的记录。官方语言数最多的国家可能有多个，这里使用了函数 maxp 的选项 @a，用于返回所有使计算表达式的值最大的成员。

在 IDE 中点击 A3 时可以看到选出了两个结果：

Index	CountryCode	Num
1	CHE	4
2	ZAF	4

A4: 将 A3 的国家编号外键对象化，并构造出需要的目标结构。

首先把国家编号转换为对应的国家记录：

Index	CountryCode	Num
1	[CHE,Switzerland,7160400]	4
2	[ZAF,South Africa,40377000]	4

双击国家编号，可以看到对应的国家记录：

Code	Name	Population
CHE	Switzerland	7160400

然后构造出目标结构：

Index	Name	Population	Num
1	Switzerland	7160400	4
2	South Africa	40377000	4

2. 查询官方语言最多的国家名称、人口、使用最多的官方语言及人口比例

这个问题比问题 1 多了一步，需要选出使用最多的官方语言。

SQL 语句如下：

```
with t1 as (
    select CountryCode, count(*) Num,
    RANK()OVER (ORDER BY COUNT(*) DESC) AS rk
    from world.countrylanguage
    where isOfficial='T'
    group by CountryCode),
t2 as (
    select *
    from t1
    where rk=1),
t3 as (
    select cl.CountryCode, Language, Percentage
    from world.countrylanguage cl join t2
    on cl.CountryCode=t2.CountryCode),
t4 as (
    select CountryCode, max(Percentage) MaxP
    from t3
    group by CountryCode),
t5 as (
    select t3.CountryCode, t3.Language, t3.Percentage
    from t3 join t4
    on t3.CountryCode=t4.CountryCode and t3.Percentage=t4.MaxP)
select c.Name,c.Population,t5.Language,t5.Percentage
    from world.country c join t5
    on c.Code=t5.CountryCode;
```

虽然只是增加了一步选出最大值，但是 SQL 语言的复杂程度却增加了很多。主要是因为 SQL 的分组功能总是和汇总同时出现，无法保留分组子集再运算。

对于 SPL，还是按自然逻辑，只要在问题 1 的基础上选出使用最多的官方语言即可：

A1/A2: 读取国家表和国家语言表并选出官方语言。

A3: 按国家编码分组，并取出元素最多的组，然后在每组中查找比例最大的记录。

Index	CountryCode	Language	IsOfficial	Percentage
1	CHE	German	T	63.6
2	ZAF	Zulu	T	22.7

A4: 将 A3 的国家编号外键对象化，并构造出需要的目标结构。

Index	Name	Population	Language	Percentage
1	Switzerland	7160400	German	63.6
2	South Africa	40377000	Zulu	22.7

这个 SPL 脚本相比问题 1 并没有复杂很多，这主要得益于 SPL 支持独立的分组运算，在分组后保留了分组子集。后续想要选出种类最多、使用最多的语言时，可以继续使用分组子集进行各种运算。

3. 将官方语言最多的国家的官方语言及使用比例在同一行按从大到小排列

期待的目标结果集是这样的：

Index	CountryCode	Language1	Percentage1	Language2	Percentage2	Language3	Percentage3	Language4	Percentage4
1	CHE	German	63.6	French	19.2	Italian	7.7	Romansh	0.6
2	ZAF	Zulu	22.7	Xhosa	17.7	Afrikaans	14.3	English	8.5

可以看到，目标数据结构是由数据计算而来的，不是固定不变的。普通的 SQL 不能解决这样的问题，需要使用动态 SQL，实现起来很繁琐，这里就不再给出 SQL 的解决方案了。

SPL 解决这个问题仍不复杂，只要在问题 1 基础上，将官方语言最多的国家按照使用比例排序，再构造出目标数据结构即可：

A1: 读取国家语言表并选出官方语言。

A2: 按国家编号分组，并取出成员最多的所有组的记录。再将这些记录拼成一个序列，并按照国家编号和使用比例排序，其中使用比例降序排列。

Index	CountryCode	Language	IsOfficial	Percentage
1	CHE	German	T	63.6
2	CHE	French	T	19.2
3	CHE	Italian	T	7.7
4	CHE	Romansh	T	0.6
5	ZAF	Zulu	T	22.7
6	ZAF	Xhosa	T	17.7
7	ZAF	Afrikaans	T	14.3
8	ZAF	English	T	8.5

A3：将相同国家编号及所有的语言和使用比例转到同一行。这里使用函数 groupc，对序列的序列执行行列转换计算。

Index	CountryCode	_2	_3	_4	_5	_6	_7	_8	_9
1	CHE	German	63.6	French	19.2	Italian	7.7	Romansh	0.6
2	ZAF	Zulu	22.7	Xhosa	17.7	Afrikaans	14.3	English	8.5

A4：将 A3 第 2 列及以后的列名依次改成 Language**i**和 Percentage**i**:

Index	CountryCode	Language1	Percentage1	Language2	Percentage2	Language3	Percentage3	Language4	Percentage4
1	CHE	German	63.6	French	19.2	Italian	7.7	Romansh	0.6
2	ZAF	Zulu	22.7	Xhosa	17.7	Afrikaans	14.3	English	8.5

数据文件附件：

country.txt

countrylanguage.txt