L1 E2 - 3 - Grouping Sets

July 1, 2021

1 Exercise 02 - OLAP Cubes - Grouping Sets

All the databases table in this demo are based on public database samples and transformations - `Sakila` is a sample database created by `MySql` Link - The postgresql version of it is called `Pagila` Link - The facts and dimension tables design is based on O'Reilly's public dimensional modelling tutorial schema Link

Start by connecting to the database by running the cells below. If you are coming back to this exercise, then uncomment and run the first cell to recreate the database. If you recently completed the slicing and dicing exercise, then skip to the second cell.

```
In [1]: !PGPASSWORD=student createdb -h 127.0.0.1 -U student pagila_star
        !PGPASSWORD=student psql -q -h 127.0.0.1 -U student -d pagila_star -f Data/pagila-star.s
```

```
 set_config
------------

(1 row)

 setval
--------
    200
(1 row)

 setval
--------
    605
(1 row)

 setval
--------
     16
(1 row)

 setval
--------
    600
(1 row)
```

```
 setval
--------
    109
(1 row)

 setval
--------
    599
(1 row)

 setval
--------
      1
(1 row)

 setval
--------
      1
(1 row)

 setval
--------
      1
(1 row)

 setval
--------
      1
(1 row)

 setval
--------
  16049
(1 row)

 setval
--------
   1000
(1 row)

 setval
--------
   4581
(1 row)

 setval
--------
```

```
      6
(1 row)

 setval
--------
  32098
(1 row)

 setval
--------
  16049
(1 row)

 setval
--------
      2
(1 row)

 setval
--------
      2
(1 row)
```

1.0.1 Connect to the local database where Pagila is loaded

```
In [2]: import sql
        %load_ext sql

        DB_ENDPOINT = "127.0.0.1"
        DB = 'pagila_star'
        DB_USER = 'student'
        DB_PASSWORD = 'student'
        DB_PORT = '5432'

        # postgresql://username:password@host:port/database
        conn_string = "postgresql://{}:{}@{}:{}/{}" \
                                .format(DB_USER, DB_PASSWORD, DB_ENDPOINT, DB_PORT, DB)

        print(conn_string)
        %sql $conn_string
```

```
postgresql://student:student@127.0.0.1:5432/pagila_star
```

```
Out[2]: 'Connected: student@pagila_star'
```

1.0.2 Star Schema

2 Grouping Sets

- It happens often that for 3 dimensions, you want to aggregate a fact:
  - by nothing (total)
  - then by the 1st dimension
  - then by the 2nd
  - then by the 3rd
  - then by the 1st and 2nd
  - then by the 2nd and 3rd
  - then by the 1st and 3rd
  - then by the 1st and 2nd and 3rd
- Since this is very common, and in all cases, we are iterating through all the fact table anyhow, there is a more clever way to do that using the SQL grouping statement "GROUPING SETS"

2.1 Total Revenue

TODO: Write a query that calculates total revenue (sales_amount)

```
In [3]: %%sql
        select sum(sales_amount) as revenue
        from factSales;

 * postgresql://student:***@127.0.0.1:5432/pagila_star
1 rows affected.


Out[3]: [(Decimal('67416.51'),)]
```

2.2 Revenue by Country

TODO: Write a query that calculates total revenue (sales_amount) by country

```
In [5]: %%sql
        select dimStore.country, sum(sales_amount) as revenue
        from factSales
        join dimStore on (dimStore.store_key = factSales.store_key)
        group by dimStore.country
        order by dimStore.country, revenue desc

 * postgresql://student:***@127.0.0.1:5432/pagila_star
2 rows affected.


Out[5]: [('Australia', Decimal('33726.77')), ('Canada', Decimal('33689.74'))]
```

2.3 Revenue by Month

TODO: Write a query that calculates total revenue (sales_amount) by month

```
In [7]: %%sql
        select dimDate.month, sum(sales_amount) as revenue
        from factSales
        join dimDate on (dimDate.date_key = factSales.date_key)
        group by dimDate.month
        order by dimDate.month, revenue desc

 * postgresql://student:***@127.0.0.1:5432/pagila_star
5 rows affected.


Out[7]: [(1, Decimal('4824.43')),
         (2, Decimal('9631.88')),
         (3, Decimal('23886.56')),
         (4, Decimal('28559.46')),
         (5, Decimal('514.18'))]
```

2.4 Revenue by Month & Country

TODO: Write a query that calculates total revenue (sales_amount) by month and country. Sort the data by month, country, and revenue in descending order. The first few rows of your output should match the table below.

```
In [9]: %%sql
        select dimDate.month, dimStore.country, sum(sales_amount) as revenue
        from factSales
        join dimDate on (dimDate.date_key = factSales.date_key)
        join dimStore on (dimStore.store_key = factSales.store_key)
        group by dimDate.month, dimStore.country
        order by dimDate.month, dimStore.country, revenue desc

 * postgresql://student:***@127.0.0.1:5432/pagila_star
10 rows affected.


Out[9]: [(1, 'Australia', Decimal('2364.19')),
         (1, 'Canada', Decimal('2460.24')),
         (2, 'Australia', Decimal('4895.10')),
         (2, 'Canada', Decimal('4736.78')),
         (3, 'Australia', Decimal('12060.33')),
         (3, 'Canada', Decimal('11826.23')),
         (4, 'Australia', Decimal('14136.07')),
         (4, 'Canada', Decimal('14423.39')),
         (5, 'Australia', Decimal('271.08')),
         (5, 'Canada', Decimal('243.10'))]
```

```
<tbody><tr>
    <th>month</th>
    <th>country</th>
    <th>revenue</th>
</tr>
<tr>
    <td>1</td>
    <td>Australia</td>
    <td>2364.19</td>
</tr>
<tr>
    <td>1</td>
    <td>Canada</td>
    <td>2460.24</td>
</tr>
<tr>
    <td>2</td>
    <td>Australia</td>
    <td>4895.10</td>
</tr>
<tr>
    <td>2</td>
    <td>Canada</td>
    <td>4736.78</td>
</tr>
<tr>
    <td>3</td>
    <td>Australia</td>
    <td>12060.33</td>
</tr>
```

2.5 Revenue Total, by Month, by Country, by Month & Country All in one shot

TODO: Write a query that calculates total revenue at the various grouping levels done above (total, by month, by country, by month & country) all at once using the grouping sets function. Your output should match the table below.

```
In [10]: %%sql
         select dimDate.month, dimStore.country, sum(sales_amount) as revenue
         from factSales
         join dimDate on (dimDate.date_key = factSales.date_key)
         join dimStore on (dimStore.store_key = factSales.store_key)
         group by grouping sets ((), dimDate.month, dimStore.country, (dimDate.month, dimStore.c
         order by dimDate.month, dimStore.country, revenue desc

 * postgresql://student:***@127.0.0.1:5432/pagila_star
18 rows affected.
```

```
Out[10]: [(1, 'Australia', Decimal('2364.19')),
          (1, 'Canada', Decimal('2460.24')),
          (1, None, Decimal('4824.43')),
          (2, 'Australia', Decimal('4895.10')),
          (2, 'Canada', Decimal('4736.78')),
          (2, None, Decimal('9631.88')),
          (3, 'Australia', Decimal('12060.33')),
          (3, 'Canada', Decimal('11826.23')),
          (3, None, Decimal('23886.56')),
          (4, 'Australia', Decimal('14136.07')),
          (4, 'Canada', Decimal('14423.39')),
          (4, None, Decimal('28559.46')),
          (5, 'Australia', Decimal('271.08')),
          (5, 'Canada', Decimal('243.10')),
          (5, None, Decimal('514.18')),
          (None, 'Australia', Decimal('33726.77')),
          (None, 'Canada', Decimal('33689.74')),
          (None, None, Decimal('67416.51'))]
```

```
<tbody><tr>
    <th>month</th>
    <th>country</th>
    <th>revenue</th>
</tr>
<tr>
    <td>1</td>
    <td>Australia</td>
    <td>2364.19</td>
</tr>
<tr>
    <td>1</td>
    <td>Canada</td>
    <td>2460.24</td>
</tr>
<tr>
    <td>1</td>
    <td>None</td>
    <td>4824.43</td>
</tr>
<tr>
    <td>2</td>
    <td>Australia</td>
    <td>4895.10</td>
</tr>
<tr>
    <td>2</td>
    <td>Canada</td>
    <td>4736.78</td>
```

```
</tr>
<tr>
    <td>2</td>
    <td>None</td>
    <td>9631.88</td>
</tr>
<tr>
    <td>3</td>
    <td>Australia</td>
    <td>12060.33</td>
</tr>
<tr>
    <td>3</td>
    <td>Canada</td>
    <td>11826.23</td>
</tr>
<tr>
    <td>3</td>
    <td>None</td>
    <td>23886.56</td>
</tr>
<tr>
    <td>4</td>
    <td>Australia</td>
    <td>14136.07</td>
</tr>
<tr>
    <td>4</td>
    <td>Canada</td>
    <td>14423.39</td>
</tr>
<tr>
    <td>4</td>
    <td>None</td>
    <td>28559.46</td>
</tr>
<tr>
    <td>5</td>
    <td>Australia</td>
    <td>271.08</td>
</tr>
<tr>
    <td>5</td>
    <td>Canada</td>
    <td>243.10</td>
</tr>
<tr>
    <td>5</td>
```

```
    <td>None</td>
    <td>514.18</td>
</tr>
<tr>
    <td>None</td>
    <td>None</td>
    <td>67416.51</td>
</tr>
<tr>
    <td>None</td>
    <td>Australia</td>
    <td>33726.77</td>
</tr>
<tr>
    <td>None</td>
    <td>Canada</td>
    <td>33689.74</td>
</tr>
```