Exercise II - SQL

Part 3 – Investigate Big Earnings with SQL

CEE412/CET522

Transportation Data Management and Visualization

WINTER 2020

Getting Started

In this exercise we will look at some data describing top earning businesses and CEOs, including some data about the countries in which they are located.

The tables we are using are described below, all are contained in the CEE412_CET522_W20 database:

- E2_Companies table describes the top 500 earning companies around the world in 2014:

  Companies(Company, Country, Sales, Profits, Assets, MarketValue)

- E2_CEOs table describes the top 200 highest payed CEOs in the USA for 2014:

  CEOs(Name, Company, OneYrPay, FiveYrPay, Shares, Age)

  One year payment

  Five year payment

  Shares held

- E2_Countries table describes nearly all countries and regions as of 2013:

  Countries(Country, GDPPC, Population, PercentWorld)

  GDP Per Capita

  Fraction of world population

Getting Started

As might be expected, the following relationships can be defined between the tables:

- Companies.Company = CEOs.Company
- Countries.Country = Companies.Country

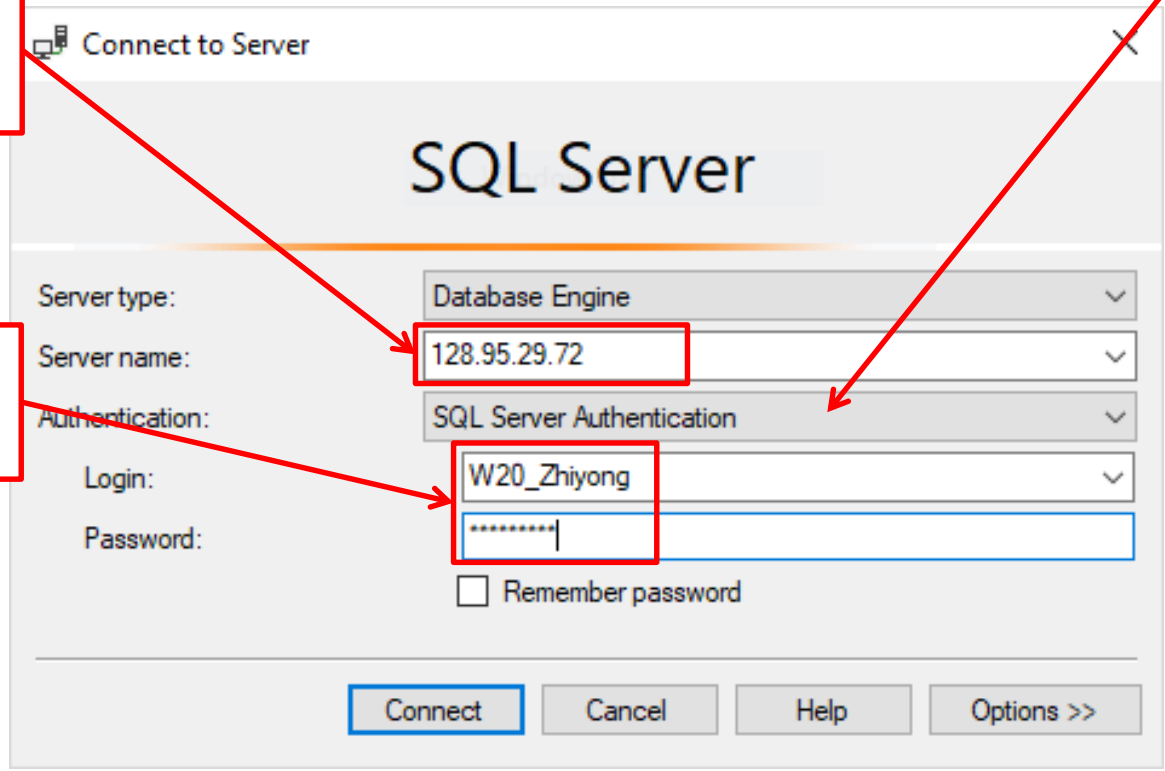
Note: the top 200 highest paid CEOs in the USA do not necessarily all work for any of the global highest earning companies. Thus, there will be many companies which do not have their CEOs recorded in the CEOs table, and likewise a number of CEOs which do not have their companies recorded in the Companies tables. Also, there will be a number of countries in which none of the top 500 companies are based.

Log on Your Database Account

Use **SQL Server Authentication**

Input the class server IP address: 128.95.29.72

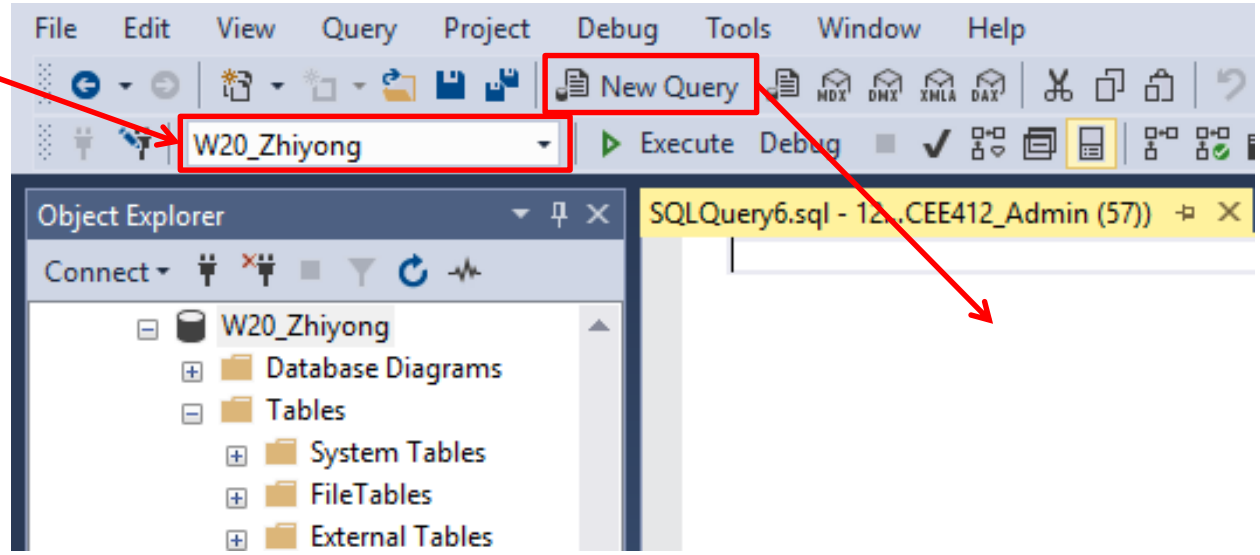
Input your account name and your password



Make Copies of the Tables

- Note that the tables are stored in the class database (CEE412_CET522_W20), where you only have the permission of the datareader.
- It is good to copy those tables into your own database, where you can have full control permissions of your tables.
- To copy the tables, first create a new query by clicking the **New Query** button.

Make sure you are working in your own database!



Make Copies of the Tables

- Remember that we can use the SELECT INTO function to create a new table and store the results from a query.
- Run the following SQL queries to copy tables into your own database.

```
SELECT *
INTO CEOs
FROM CEE412_CET522_W20.dbo.E2_CEOs

SELECT *
INTO Companies
FROM CEE412_CET522_W20.dbo.E2_Companies

SELECT *
INTO Countries
FROM CEE412_CET522_W20.dbo.E2_Countries
```

- Note that you are working in your own database, but you want to query data from another database (CEE412_CET522_W20).
- You can query a table in another DB using the following form:

[DB Name].[Schema Name].[Table Name]

E.g., CEE412_CET522_W20.dbo.E2_CEOs

- CEE412_CET522_W20: database name
- dbo: schema name, short for “database owner”
- E2_CEOs: table name

Simple Queries

Question: Find the gross domestic product per capita (GDPPC) of all of the countries in which the top 500 companies are based.

Tips*:

- Use an inner join between the tables.
- In the SELECT clause, list only the attributes we are interested in (e.g. country name, and country GDPPC).

* In this exercise, I will give you some tips that may help you develop your queries. Note that there are usually multiple queries that can answer the same question. You may choose whether you want to follow my tips.

Develop your query, and then compare with my solution on the next slide.

Simple Queries

Possible solution:

```
SELECT DISTINCT Companies.Country, Countries.GDPPC
  FROM Countries JOIN Companies
    ON Countries.Country = Companies.Country
```

What would happen in this case if we used a RIGHT outer join?

- Answer: (give it a try) there should be no difference because there are no unmatched rows in the Companies table.

What would happen in this case if we used a LEFT outer join?

- Answer: (give it a try) there should be a number of null values in the left column of the result, because there are a number of countries with no top 500 companies.

Simple Queries

Question: Find out the names and populations for countries in which the companies employ their CEOs in the CEOs table.

- Note: its mostly USA because the CEO table describes USA CEO's only.

Tips:

- Use the distinct keyword to make sure you list each country once.
- Join three tables together to get the result.

Develop your query, and then compare with my solution on the next slide.

Simple Queries

Possible solution:

```
SELECT DISTINCT Companies.Country, Population
  FROM CEOs, Countries, Companies
 WHERE CEOs.Company = Companies.Company
       AND Countries.Country = Companies.Country
```

Result:

Country	Population
Ireland	4609600
United Kingdom	64105654
United States	320314000

Logical Operators

Question: Find the CEOs for which at least one of the following is true:

1. Runs a company in China AND either over the age of 60 or under the age of 50.
2. Make between $20 million and $60 million in one year.
3. Work for a company that has over $100,000,000,000 in sales.
4. Hold no less than 100 shares in their company.

Note: the CEO pay is in millions, so some conversion will be necessary

Develop your query, and then compare with my solution on the next slide.

Logical Operators

Possible solution:

```
SELECT Name, Age, OneYrPay, Shares,
       Companies.Company, Sales
  FROM CEOs JOIN Companies
    ON CEOs.Company = Companies.Company
 WHERE Companies.Country = 'China' AND (Age > 60 OR Age < 50)
       OR OneYrPay BETWEEN 20 AND 60
       OR Sales > 100000000000
       OR Shares >= 100
```

Aggregation

Question: Find the number of top 500 companies that each country holds.

Tips:

- Use the COUNT() aggregation function, and give the result a column name.
- Group by Country.
- Order your result in a way you like.

Develop your query, and then compare with my solution on the next slide.

Aggregation

Possible solution:

```
SELECT Country, COUNT(*) AS CountOfCompanies
  FROM Companies
 GROUP BY Country
 ORDER BY CountOfCompanies DESC
```

Follow up question: How would this query be changed to return only the countries that have more than 10 top companies?

Aggregation

Follow up solution (using HAVING clause):

```
SELECT Country, COUNT(*) AS CountOfCompanies
  FROM Companies
 GROUP BY Country
HAVING COUNT(*) > 20
 ORDER BY CountOfCompanies DESC
```

Aggregation

Question: Find out how many top 500 companies are located in each country that has over 80,000,000 people.

Tips:

- Try using a subquery to return a list of the countries with over 80 million people.

Develop your query, and then compare with my solution on the next slide.

Aggregation

Possible solution:

```
SELECT Country, COUNT(*) AS CountOfCompanies
  FROM Companies
 WHERE Country IN (SELECT Country
                     FROM Countries
                    WHERE Population > 80000000)
 GROUP BY Country
```

Country	CountOfCompanies
Brazil	5
China	30
Germany	20
India	10
Indonesia	2
Japan	51
Mexico	3
Russia	9
United States	169

Follow up question: There are 16 countries that have over 80 million people, yet only 9 have a top 500 company. Which countries with over 80 million people do not have a top 500 company?

Aggregation

Follow up solution:

```
SELECT Country
  FROM Countries
 WHERE Population > 80000000
       AND Country NOT IN (SELECT Country
                           FROM Companies)
```

Country
Egypt
Philippines
Nigeria
Vietnam
Pakistan
Bangladesh
Ethiopia

Union

Question: What is the average GDP per capita for the countries with top 500 companies, and how does it compare with countries that do not have top 500 companies?

Tips:

- Write two queries, one to answer each half of the question.
- Use a union operator to combine the results.

Develop your query, and then compare with my solution on the next slide.

Union

Possible solution:

Here I have put named values in the place of attributes. You can regard it as a column with the text 'Countries w/ Companies' in all rows.

```
(SELECT 'Countries w/ Companies' AS CountryGroup,
        ROUND(AVG(GDPPC),2) AS AvgGDPPC
   FROM Countries
  WHERE Country IN (SELECT Country
     FROM Companies))
  UNION
(SELECT 'Countries w/o Companies' AS CountryGroup,
        ROUND(AVG(GDPPC),2) AS AvgGDPPC
   FROM Countries
  WHERE Country NOT IN (SELECT Country
         FROM Companies))
```

Round the data to keep 2 decimal places.

- The Average GDPPC of countries with top 500 companies is nearly four times that of other countries.

CountryGroup	AvgGDPPC
Countries w/ Companies	40069.45
Countries w/o Companies	11824.74

Getting Creative

Question: Find the companies in each country which had the highest profits.

How to do this?

- Start by writing a query to find the highest profits in each country:

```
SELECT Country, MAX(Profits) AS MaxProfits
  FROM Companies
 GROUP BY Country
```

Country	MaxProfits
Australia	14800000000
Austria	1500000000
Belgium	14500000000
…	…

- We can now join the output of this query with the companies table. The only problem is that “Profits” is a float value, which is stored as approximate values in your database. Thus, it’s not a good idea to evaluate whether two float values are equal.

Getting Creative

- Note that although profits are in float type, all the decimal part in your values are zeros. A possible solution may be converting profits to integers and then compare the values.
- Possible solution is given below:

Convert Profits from float type to big integer.

```
SELECT c.Country, Company, Profits
  FROM Companies AS c JOIN (SELECT Country,
                                   CAST(MAX(Profits) AS BIGINT) AS MaxProfits
                              FROM Companies
                             GROUP BY Country) AS a
    ON c.Country = a.Country
       AND CAST(c.Profits AS BIGINT) = a.MaxProfits
```

Don’t forget to name your subquery when using it as a relation.

Getting Creative

Result:

Country	Company	Profits
Australia	BHP Billiton	14800000000
Austria	OMV Group	1500000000
Belgium	Anheuser-Busch InBev	14500000000
Brazil	Petrobras	10900000000
…	…	…

Getting too many zeros?

Let's make the table look better:

- Order the result by Profits
- Make the format of Profits more readable (e.g., $14.8 Billion)

Getting Creative

- You can use cast and string concatenation tools to change the format of the Profits column.
- One this to keep in mind is that, in order to concatenate strings or character variable types, you should have everything in one of the several available character variable types.
- For example, the following will work:

```
SELECT 'CEE' + '412'
```

→ The result is `'CEE412'`

- But the following will not, because I am trying to concatenate an integer with a string, and there is confusion about whether "+" means "add" or "concatenate".

```
SELECT 'CEE' + 412
```

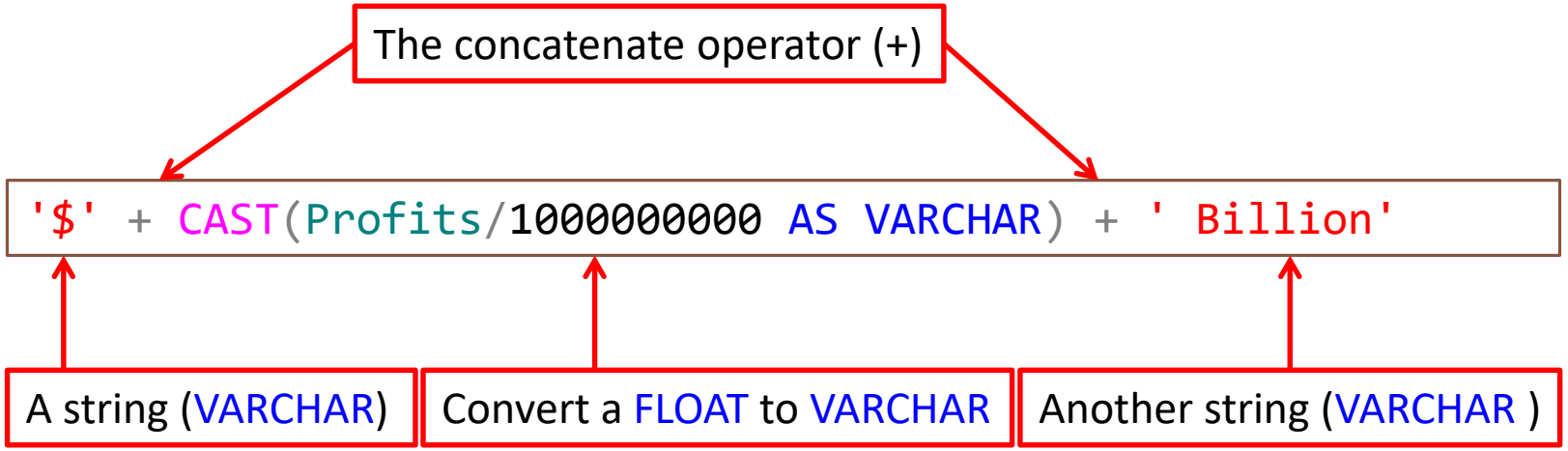
Getting Creative

- On the other hand, if I use CAST() to convert the integer to character, it will work:

```
SELECT 'CEE' + CAST(412 AS VARCHAR)
```

→ The result is `'CEE412'`

- Thus, to format the Profits column in my query, I need to have something like this in my SELECT clause:



Getting Creative

My final query:

```
SELECT c.Country, Company,
       '$' + CAST(Profits/1000000000 AS VARCHAR) + ' Billion' AS Profits
  FROM Companies AS c JOIN (SELECT Country,
                                   CAST(MAX(Profits) AS BIGINT) AS MaxProfits
                             FROM Companies
                            GROUP BY Country) AS a
    ON c.Country = a.Country
       AND CAST(c.Profits AS BIGINT) = a.MaxProfits
 ORDER BY c.Profits DESC
```

I want to order by the original numeric values of Profits without data type conversion.

Country	Company	Profits
United States	Fannie Mae	$84 Billion
China	ICBC	$42.7 Billion
Russia	Gazprom	$39 Billion
United Kingdom	Vodafone	$31.8 Billion
South Korea	Samsung Electronics	$27.2 Billion
…	…	…

Getting Creative

Answer the following questions:

1. After classifying age in 10 year bins (i.e. 10-19, 20 – 29, etc.), how many top CEOs in each age group?
2. What's the average one year pay for CEOs in each age group?

Tips:

- Create an age group column to help answer these questions.
- You may create a view or temporary table that can be accessed in later steps.
- Divide age by 10 and use FLOOR and CEILING functions to round to the nearest integer.
- Note: FLOOR and CEILING functions round down or up to the nearest integers respectively.

Develop your query, and then compare with my solution on the next slide.

Getting Creative

- I want to create an attribute “AgeGroup” for the ease of my further analysis.
- Consider the data format: values in the AgeGroup column should be something like “10-19”, “20-29”, “30-39”, etc.
- If the age is 54, the following expression will give me 50:

```
FLOOR(AGE/10)*10
```

- And the following expression can give me 59:

```
FLOOR(AGE/10+1)*10-1
```

- Why I don’t use `CEILING(AGE/10)*10-1`? When age is exactly 50, `FLOOR(AGE/10)` and `CEILING(AGE/10)` will both give me 50, thus creating a age group of “50-49”.
- After I get the bin numbers, I can convert them into the character type and then concatenate into a single string.

Getting Creative

Create the AgeGroup attribute:

```
SELECT *, CAST(FLOOR(AGE/10)*10 AS VARCHAR(2)) + '-'
          + CAST(FLOOR(AGE/10+1)*10-1 AS VARCHAR(2)) AS AgeGroup
  INTO #CEOs
  FROM CEOs
```

Save the result into a temporary table.

Convert the data type and concatenate the characters.

With the AgeGroup column in my temporary table, it becomes very easy to answer these questions:

1. How many top CEOs in each age group?
2. What's the average one year pay for CEOs in each age group?

Getting Creative

Possible solutions:

1. How many top CEOs in each age group?

```
SELECT AgeGroup, COUNT(*) AS CEOCount
  FROM #CEOs
 GROUP BY AgeGroup
 ORDER BY AgeGroup
```

AgeGroup	CEOCount
40-49	13
50-59	108
60-69	73
70-79	5
80-89	1

2. What's the average one year pay for CEOs in each age group?

```
SELECT AgeGroup, ROUND(AVG(OneYrPay),2) AS AvgPay
  FROM #CEOs
 GROUP BY AgeGroup
 ORDER BY AgeGroup
```

AgeGroup	AvgPay
40-49	15.3
50-59	19.63
60-69	19.47
70-79	25.9
80-89	24.79