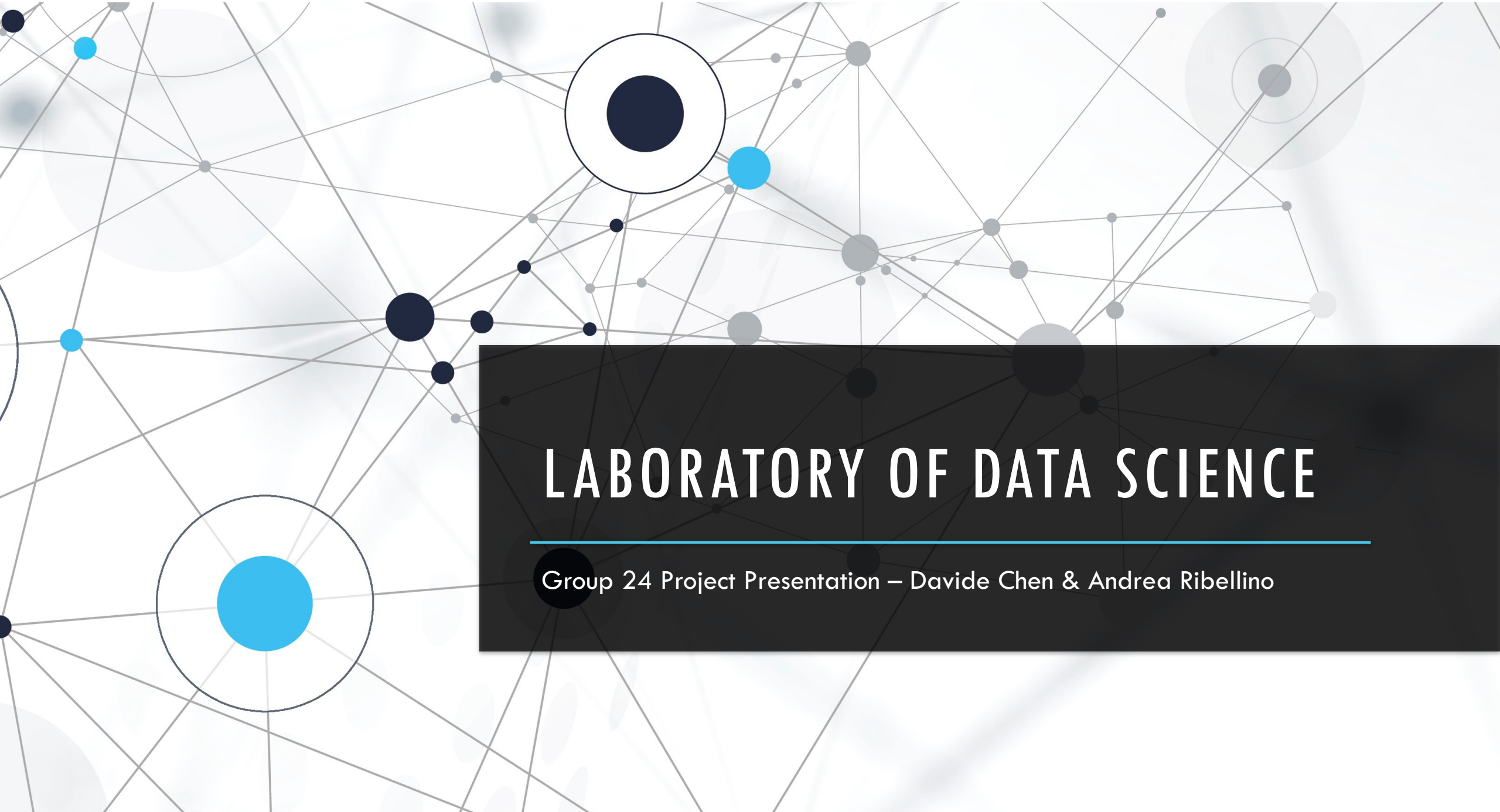
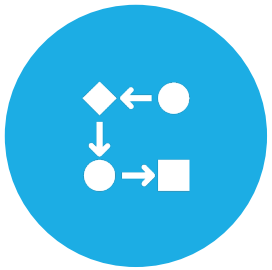


LABORATORY OF DATA SCIENCE

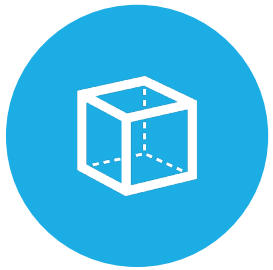
Group 24 Project Presentation – Davide Chen & Andrea Ribellino

PROJECT PIPELINE

DATA WAREHOUSE BUILDING



SSIS SOLUTION



MULTIDIMENSIONAL DATA ANALYSIS

REPORTING

DATA WAREHOUSE BUILDING

Group 24 Project Presentation –
Davide Chen & Andrea Ribellino

DATA UNDERSTANDING & PREPROCESSING

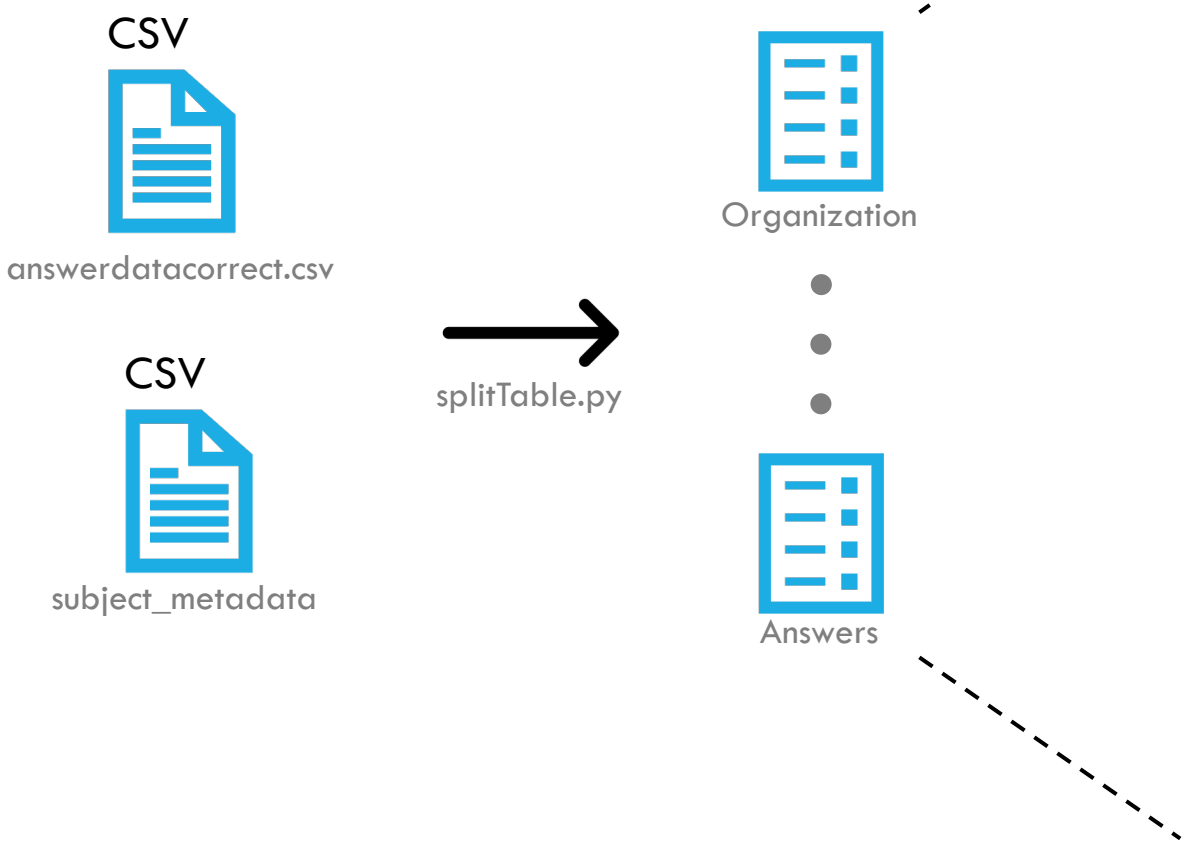
- The answerdatacorrect.csv data frame has 17 columns and 538835 rows with 0 missing value:
  - AnswerId has 538835 unique values and 0 duplicates
  - UsrId has 13630 unique values and 525205 duplicates
  - SubjectId has 412 unique values and 538423 duplicates
  - Organization table will have a total 24640 unique rows (given by [GroupId, QuizId, SchemeOfWorkId] unique rows)
  - Geography table will have a total 76 unique rows (given by Region unique rows)

- The CountyCode column has the following unique values:

[de, us, ie, it, nz, es, uk, fr, ca, be, au]

In order to match [ISO 3166-1 alpha-2] standard of country code, a transformation ‘uk’ to ‘gb’ was applied.

SPLITING TABLES



Organization table
Computed attributes: OrganizationId

Date table
Computed attributes: DateId

Subject table
Computed attributes: Description

Geography table
Computed attributes: GeoId, CountryName and Continent

User table
Computed attributes : None

Answers table
Computed attributes : IsCorrect

UNIQUENESS & INDEX KEY MANAGEMENT

UNIQUENESS

For each row

dictionary

If not exist

Table with primary key

Then add row

IDX MANAGEMENT

Add index with
Key of another table

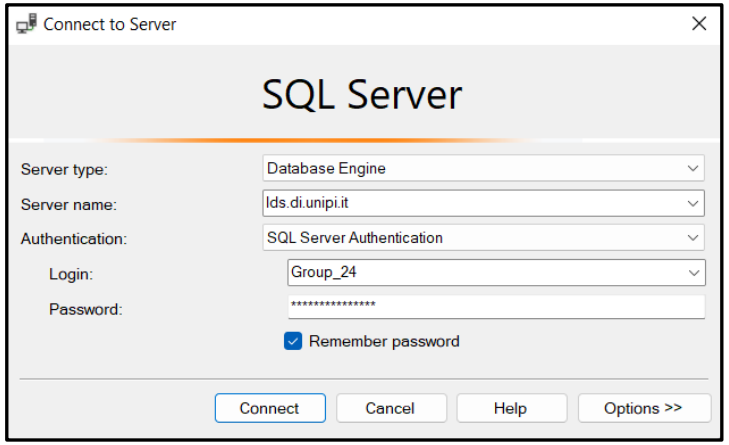
dictionary

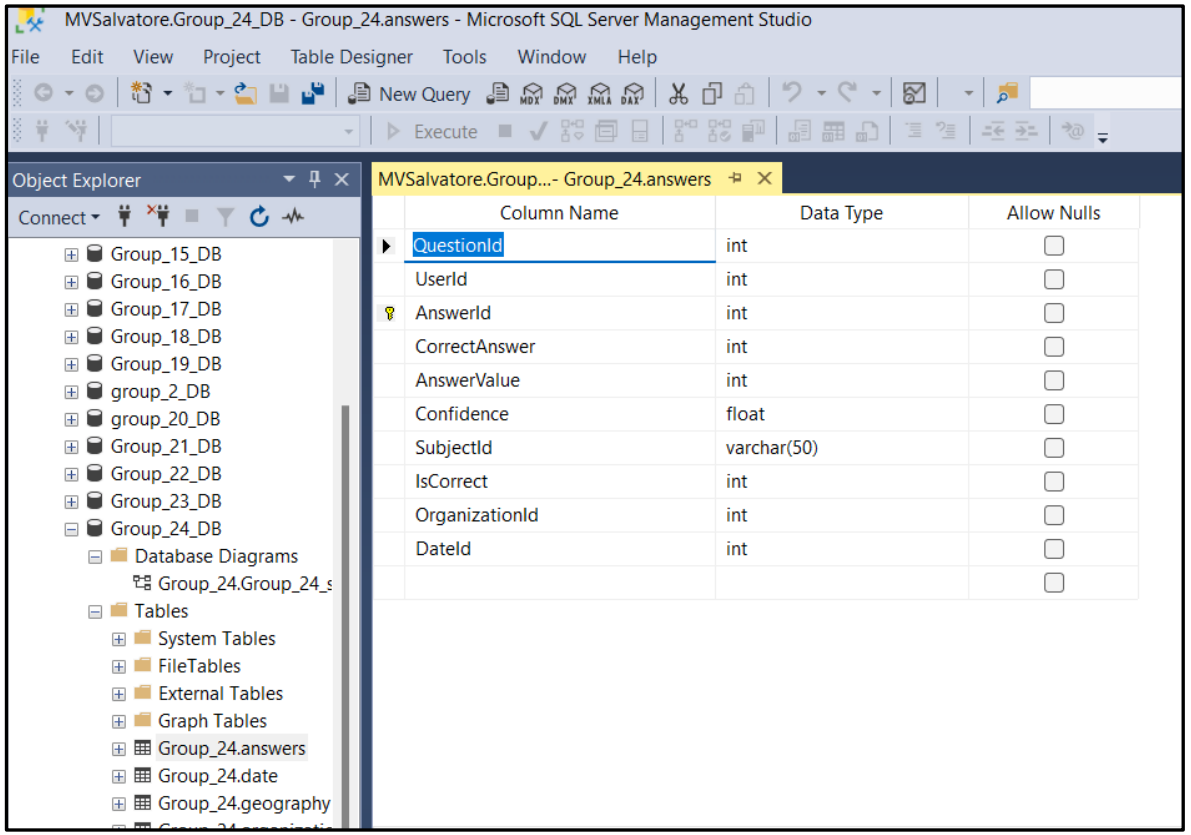
Retrieve the right
Index by the key

Table with external key

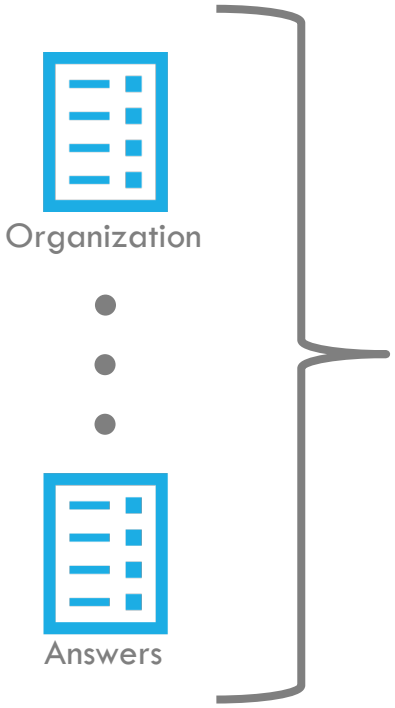
CREATION OF GROUP_24_DB

Microsoft SQL Server Management

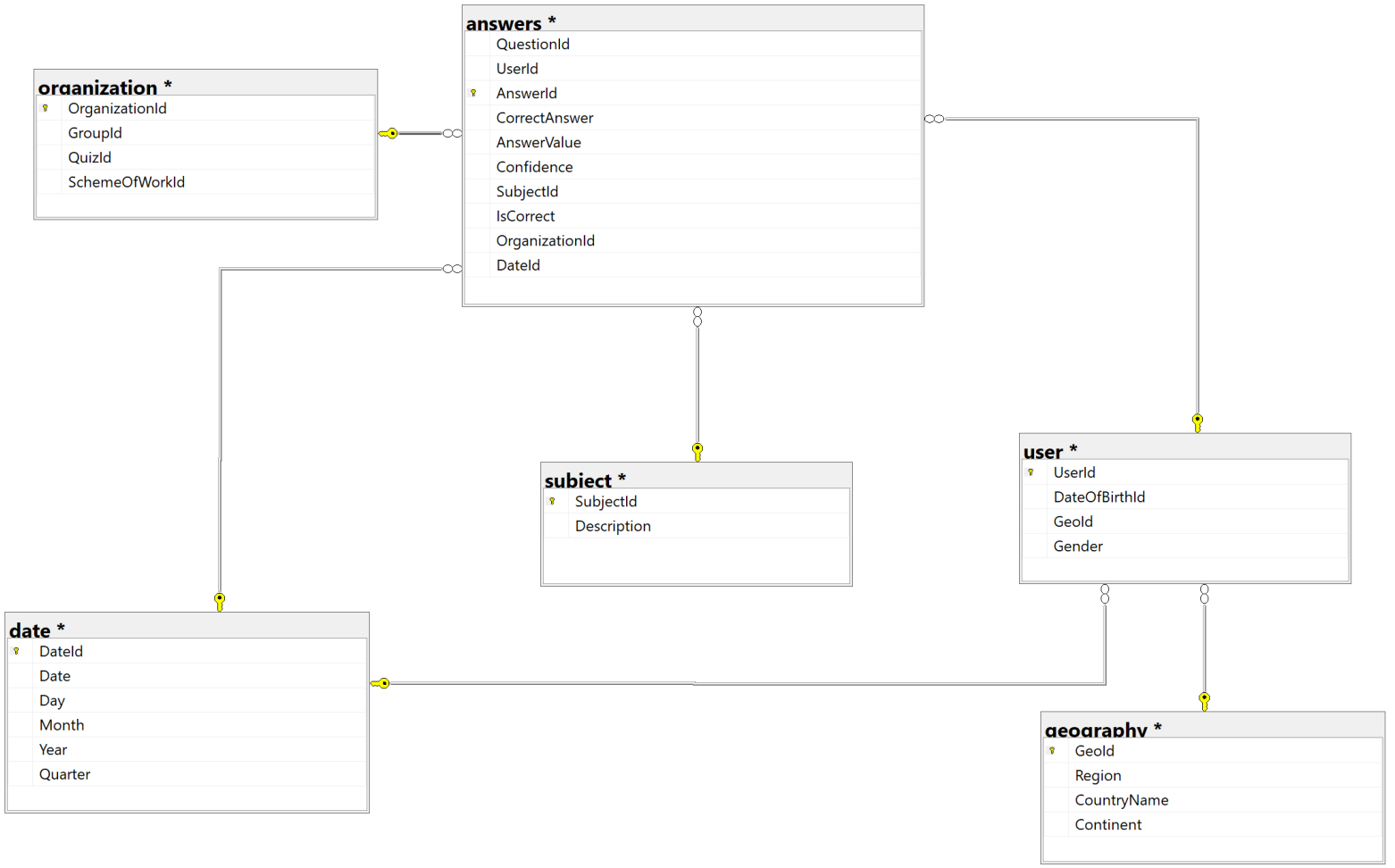




LOADING DATA



Insert row by row…
Commit every 100 rows



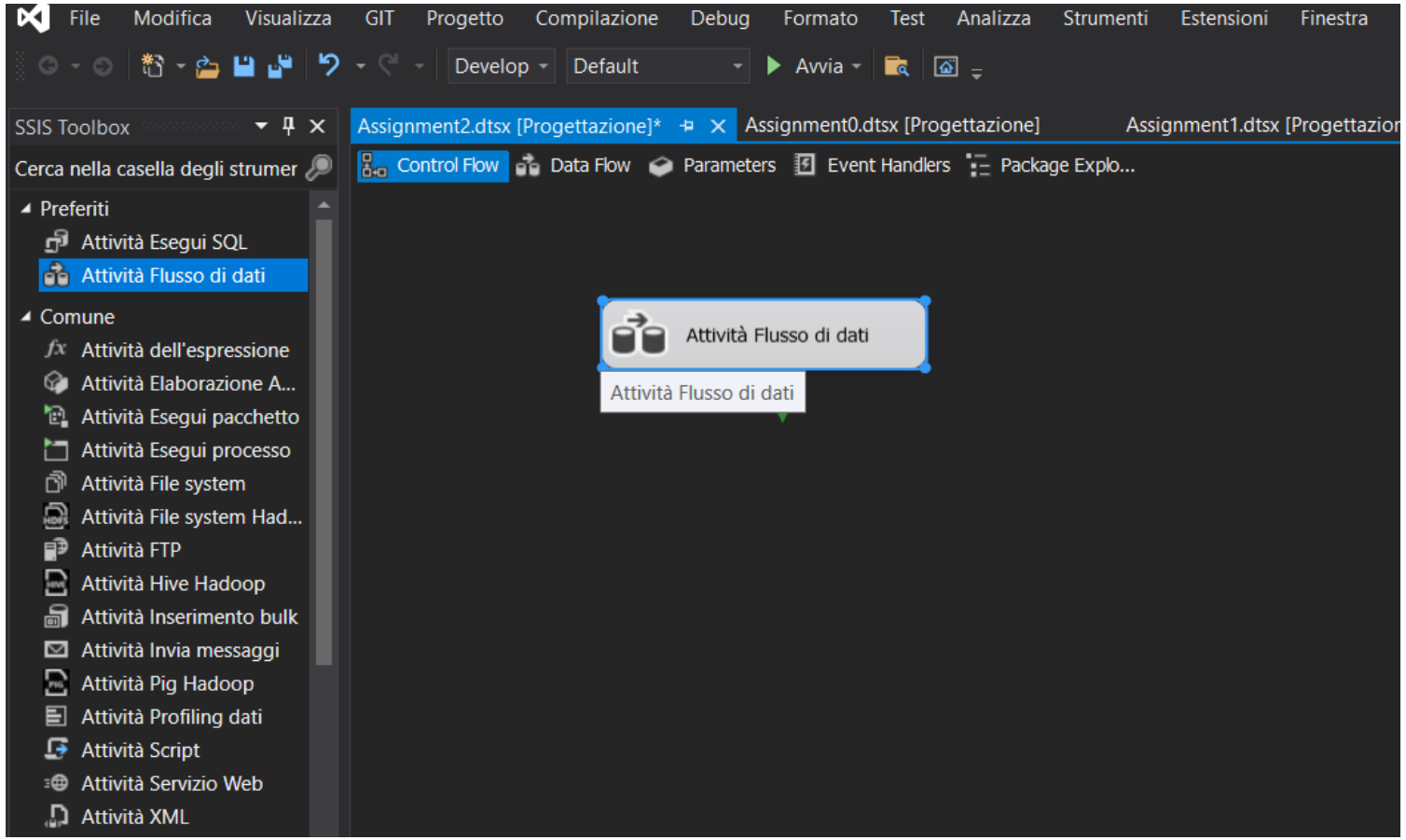
SISS – INTEGRATION SERVICE

Group 24 Project Presentation –
Davide Chen & Andrea Ribellino

VISUAL STUDIO 2019

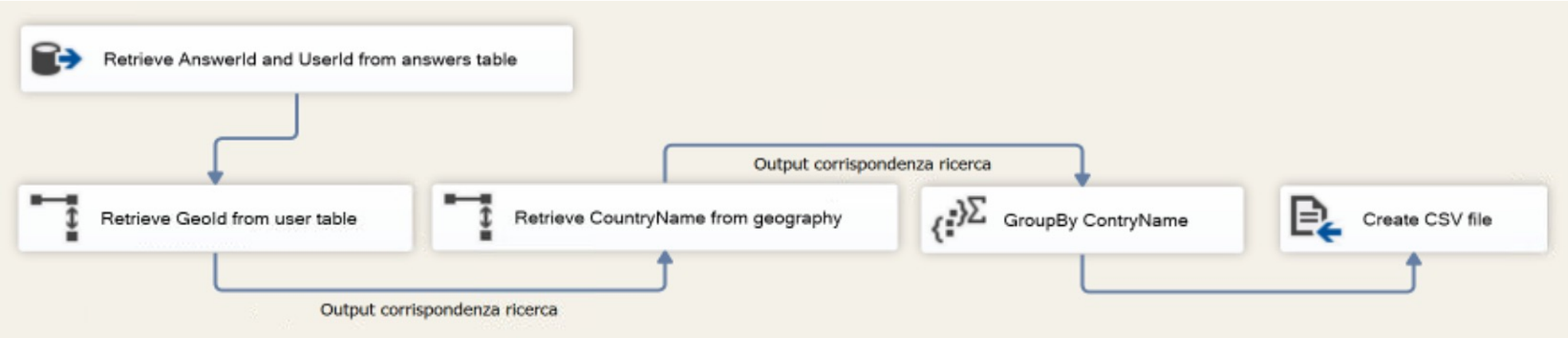
Visual Studio

Microsoft® SQL Server®
Integration Services



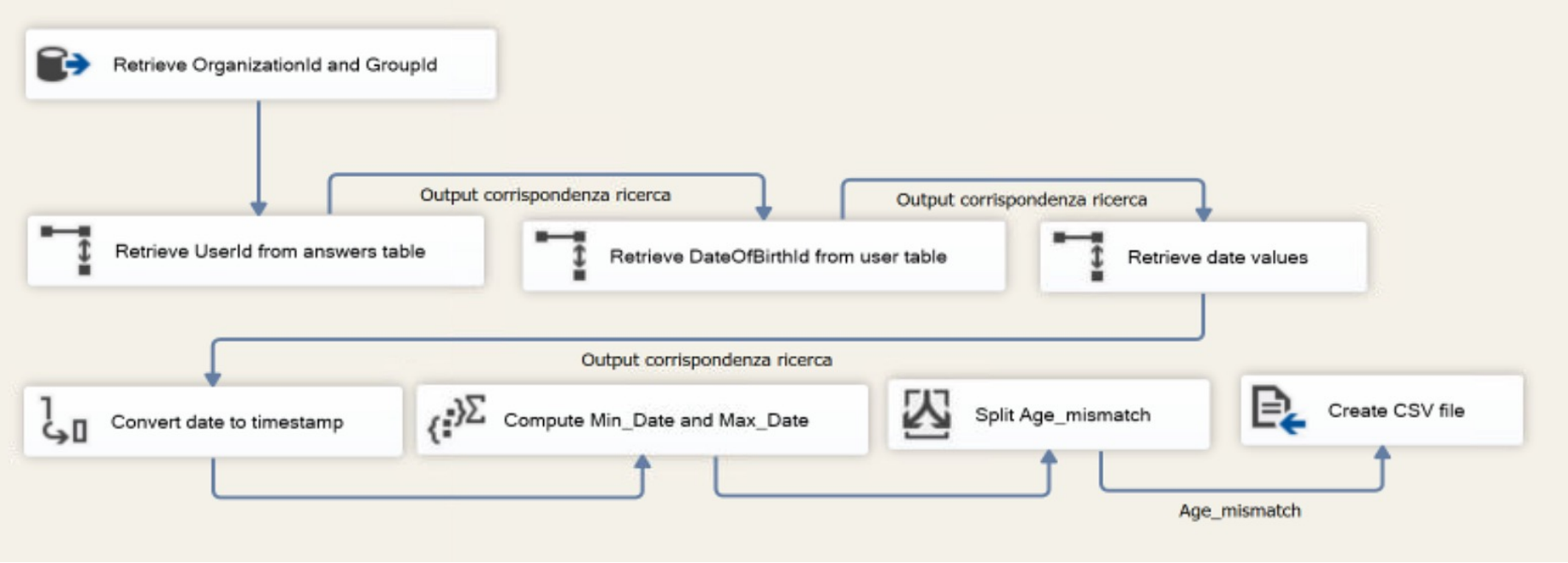
ASSIGNMENT 0

- For every country, the number of total answers.



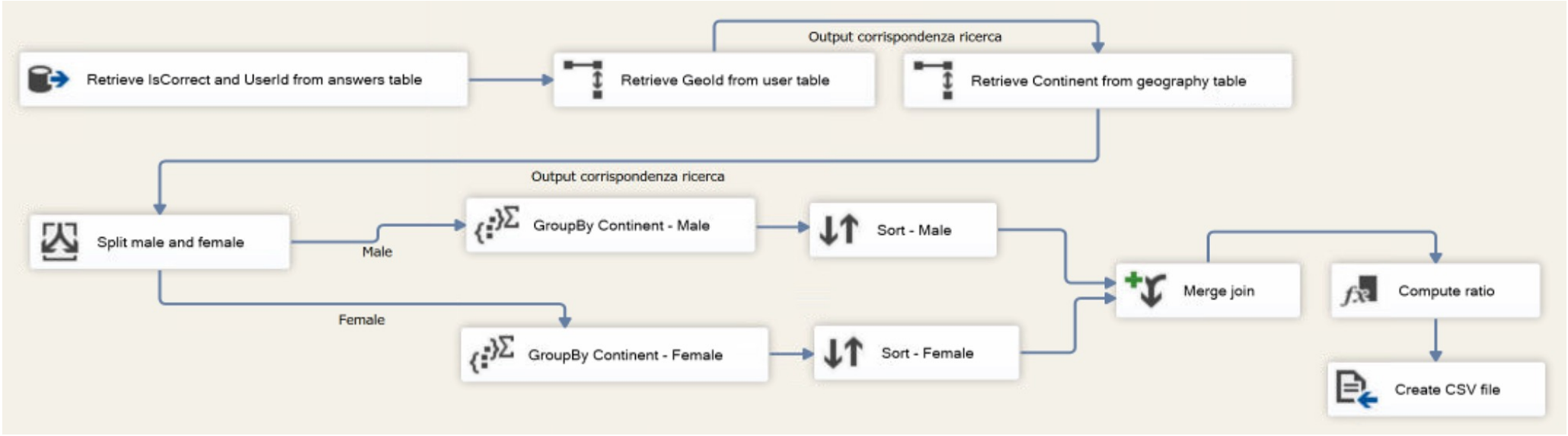
ASSIGNMENT 1

- A group (identified by GroupId) is said to have an age mismatch if the difference between the date of birth of the youngest participating student and the oldest is greater than 365 days. List all the groups with an age mismatch.



ASSIGNMENT 2

- For each continent the ratio between correct answers of males and correct answers of females



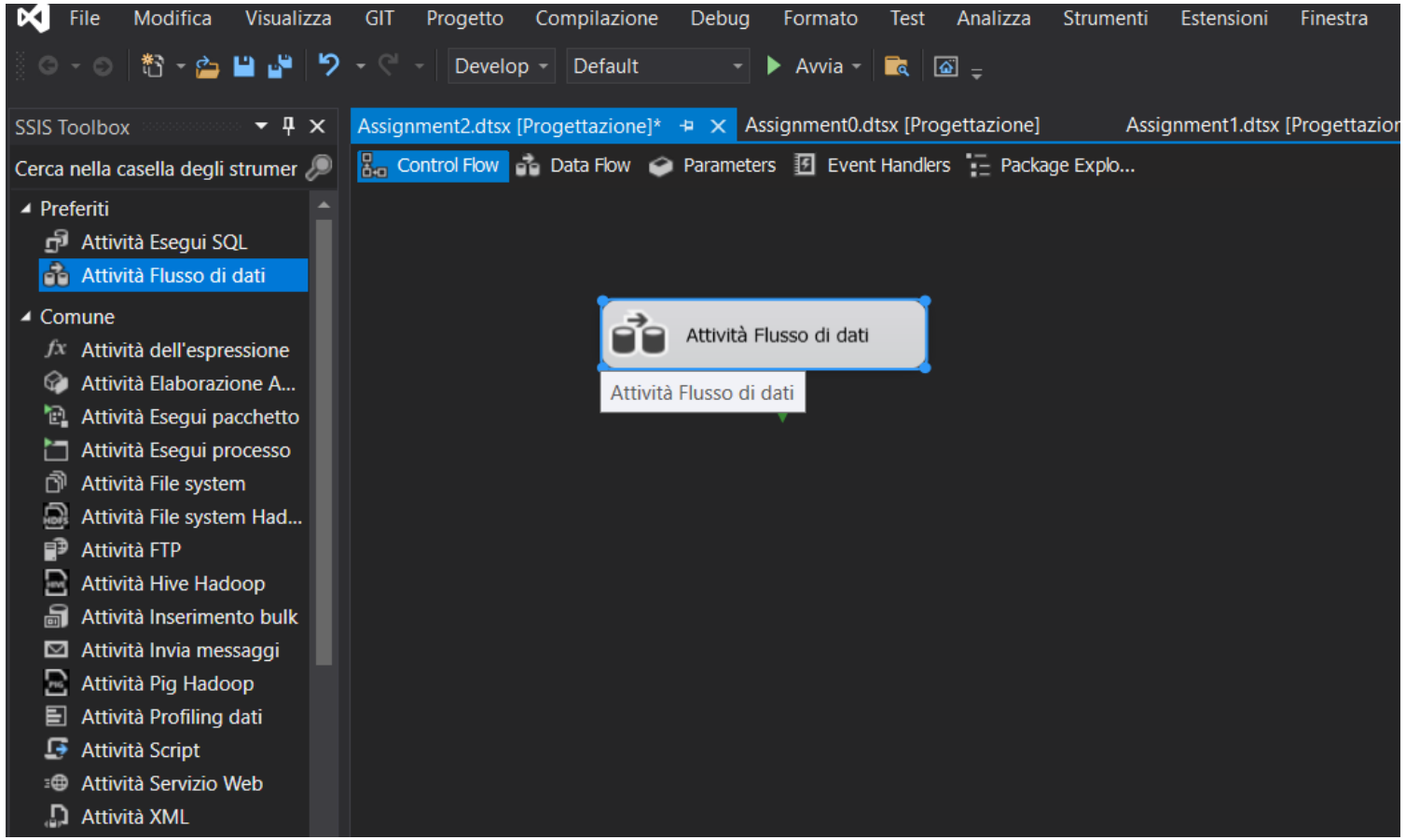
SASS – ANALYSIS SERVICE

Group 24 Project Presentation –
Davide Chen & Andrea Ribellino

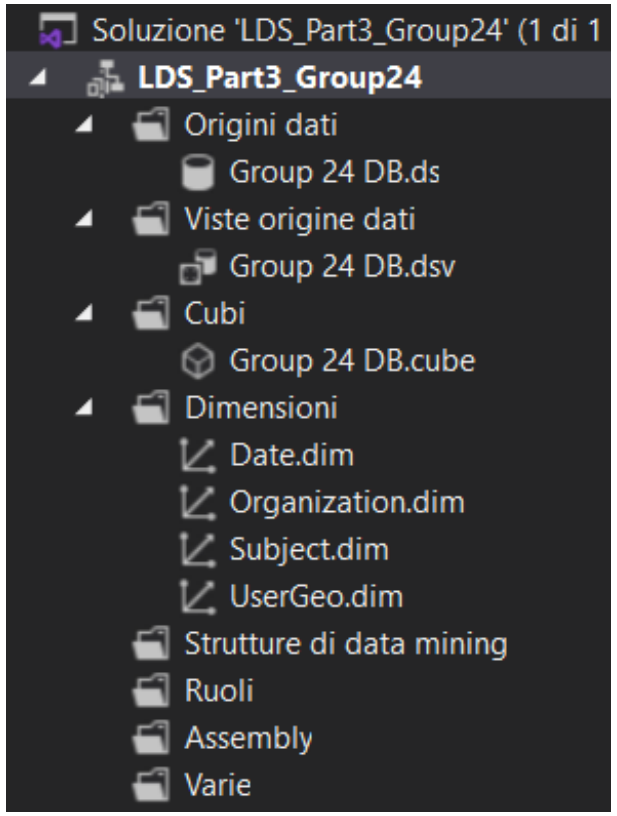
VISUAL STUDIO 2019 – CREATION OF OLAP CUBE

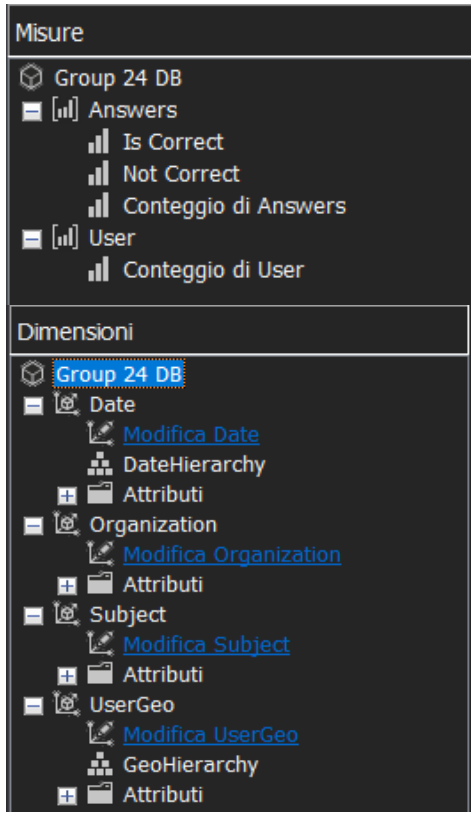
Visual Studio

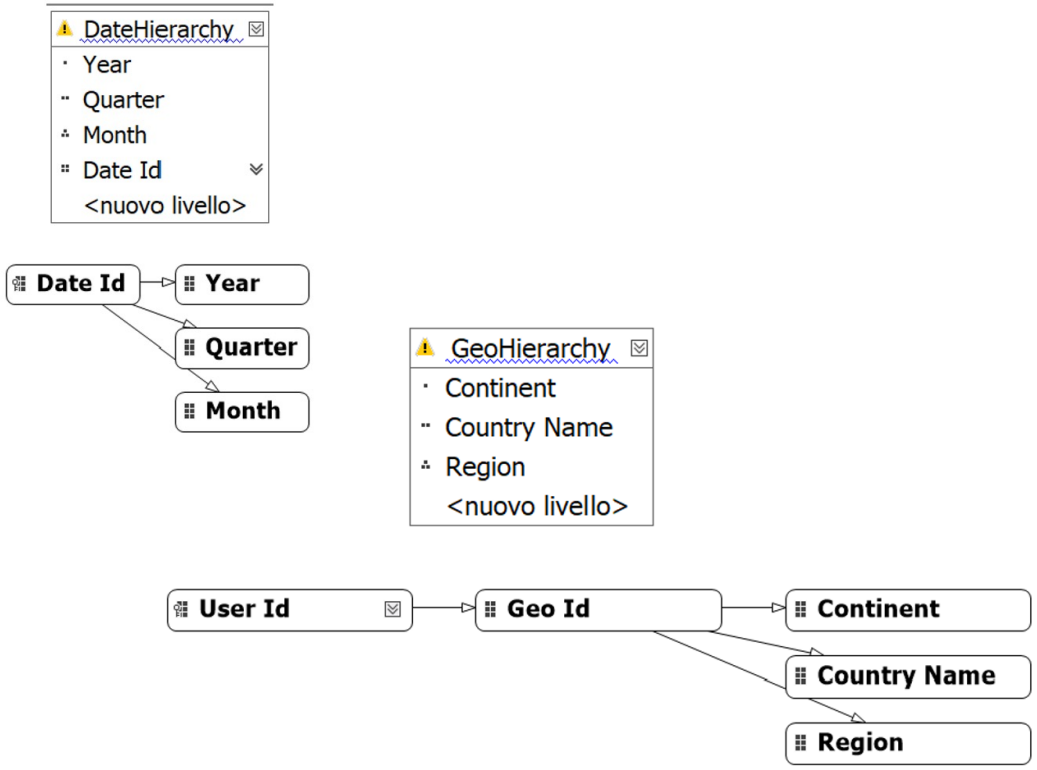
Microsoft® SQL Server®
Analysis Services



CREATION OF OLAP CUBE

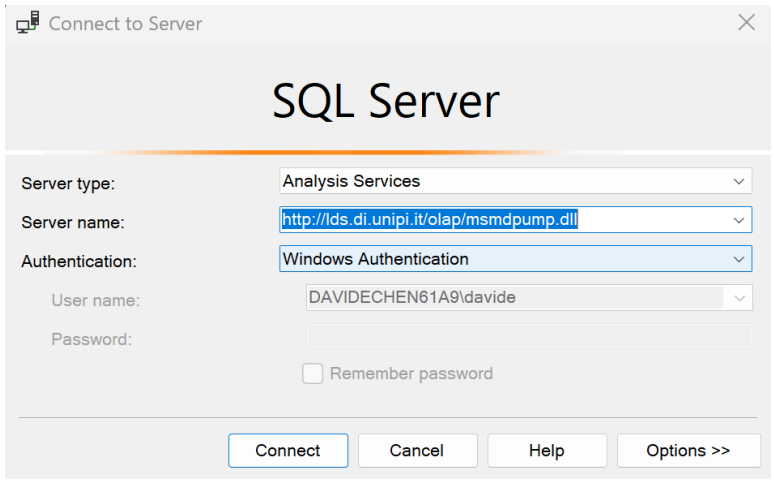


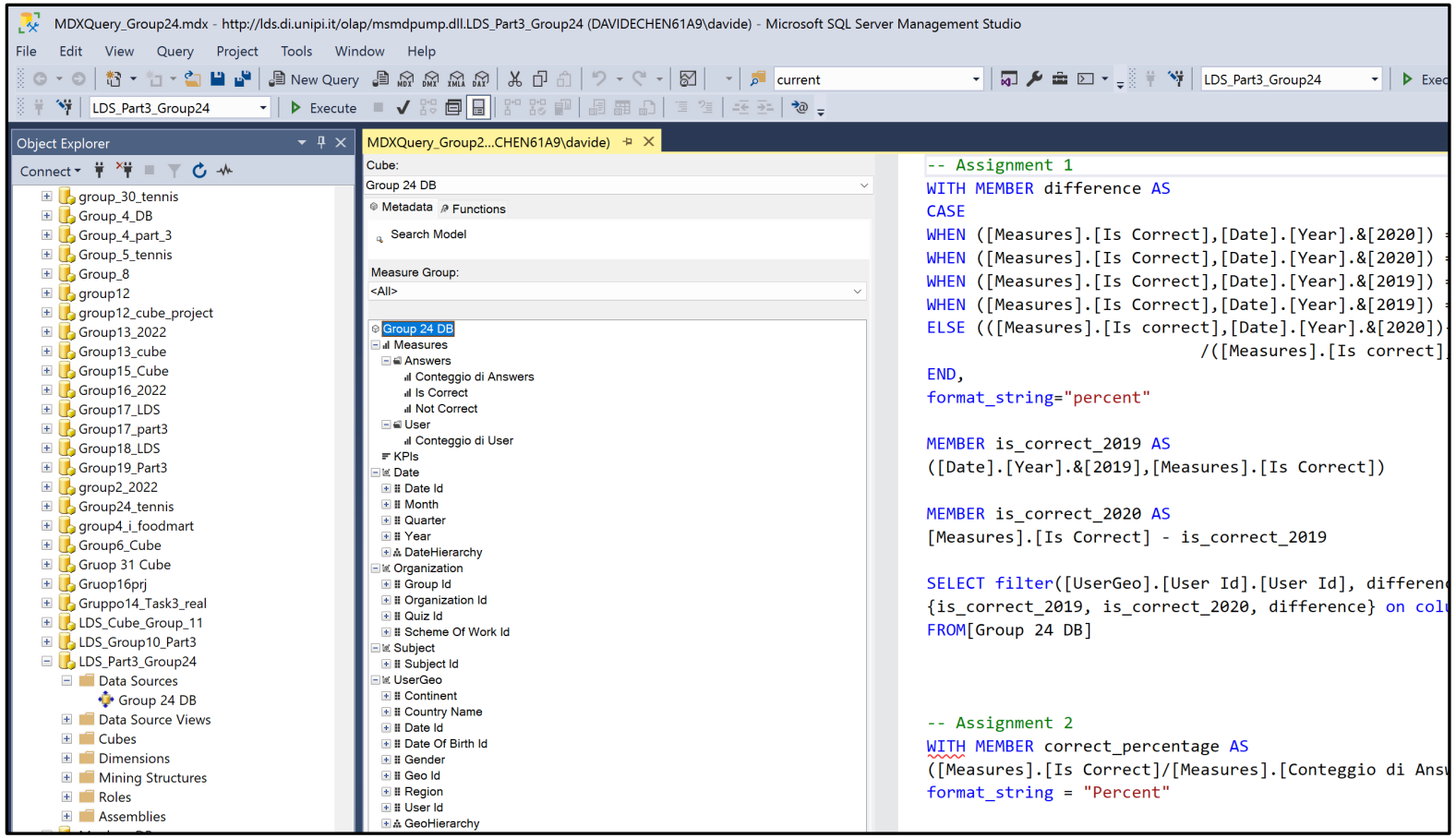




SQL MANAGEMENT SERVICE - ANALYSIS SERVICE

Microsoft SQL Server Management





MDX QUERY - 1

- Show the percentage increase or decrease in correct answers with respect to the previous year for each student

```
-- Assignment 1
WITH MEMBER difference AS
CASE
WHEN ([Measures].[Is Correct],[Date].[Year].&[2020]) = 0 then null
WHEN ([Measures].[Is Correct],[Date].[Year].&[2020]) = null then null
WHEN ([Measures].[Is Correct],[Date].[Year].&[2019]) = 0 then null
WHEN ([Measures].[Is Correct],[Date].[Year].&[2019]) = null then null
ELSE (([Measures].[Is correct],[Date].[Year].&[2020])-([Measures].[Is correct],[
    Date].[Year].&[2019]))
            /([Measures].[Is correct],[Date].[Year].&[2019])
END,
format_string="percent"

MEMBER is_correct_2019 AS
([Date].[Year].&[2019],[Measures].[Is Correct])

MEMBER is_correct_2020 AS
[Measures].[Is Correct] - is_correct_2019

SELECT filter([UserGeo].[User Id].[User Id], difference <> null) on rows,
{is_correct_2019, is_correct_2020, difference} on columns
FROM[Group 24 DB]
```

Idea di output

	is_correct_2019	is_correct_2020	variation
userID			
userID			
userID			

$$variation = \frac{correct\_answer\_of\_2020 - correct\_answer\_of\_2019}{correct\_answer\_of\_2019}$$

	is_correct_2019	is_correct_2020	difference
38	75.00	45.00	-40.00%
99	4.00	63.00	1475.00%
132	21.00	7.00	-66.67%
164	4.00	1.00	-75.00%
182	27.00	1.00	-96.30%
199	6.00	11.00	83.33%
218	185.00	56.00	-69.73%
274	5.00	32.00	540.00%
323	112.00	87.00	-22.32%

MDX QUERY - 2

- For each subject show the total correct answers in percentage with respect to the total answers of that subject.

```
-- Assignment 2
WITH MEMBER correct_percentage AS
([Measures].[Is Correct]/[Measures].[Conteggio di Answers]),
format_string = "Percent"

SELECT correct_percentage on columns,
[Subject].[Subject Id].[Subject Id] on rows
FROM [Group 24 DB]
```

OUTPUT

	correct_percentage
Maths, Advanced Pure, Functions, Composite Functions	37.64%
Maths, Advanced Pure, Functions, Function Notation	42.21%
Maths, Advanced Pure, Functions, Inverse Functions	42.36%
Maths, Algebra, Algebraic Fractions, Adding and Subtracting Algebra...	57.86%
Maths, Algebra, Algebraic Fractions, Multiplying and Dividing Algebra...	39.59%
Maths, Algebra, Algebraic Fractions, Simplifying Algebraic Fractions	55.85%
Maths, Algebra, Algebraic Fractions, Simplifying Algebraic Fractions	55.13%

$$correct\_percentage = \frac{Number\_of\_correct\_answer\_of\_given\_subject}{Number\_of\_answer\_of\_that\_subject}$$

MDX QUERY - 3

- Show the students having a total incorrect answers greater or equal than the average incorrect answers in each continent.

```
-- Assignment 3
WITH MEMBER avg_incorrect AS
(([UserGeo].[Continent], [UserGeo].[User Id].[ALL],
[Measures].[Not Correct])/([UserGeo].[Continent], [UserGeo].[User Id].[ALL], [
    Measures].[Conteggio di User])),
format_string = "standard"

SELECT {[Measures].[Not Correct], avg_incorrect} on columns,
filter (([UserGeo].[Continent].[Continent], [UserGeo].[User Id].[User Id]),
[Measures].[Not Correct] >= avg_incorrect) on rows

FROM [Group 24 DB];
```

OUTPUT

		Not Correct	avg_incorrect
EU	35	26.00	14.31
EU	38	69.00	14.31
EU	99	91.00	14.31
EU	218	24.00	14.31
EU	257	61.00	14.31

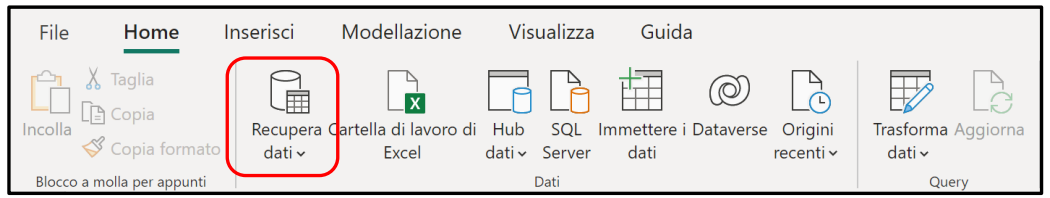
$$avg\_incorrect = \frac{sum\_of\_incorret\_answer\_of\_students\_by\_continent}{Number\_of\_student\_in\_that\_continent}$$

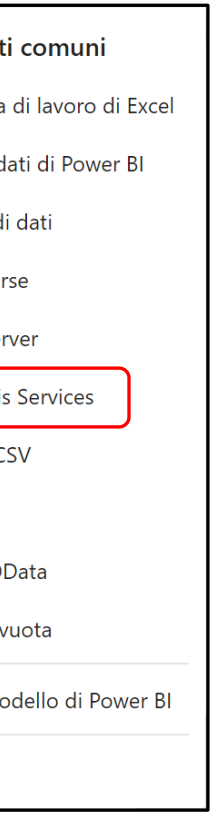
REPORTING

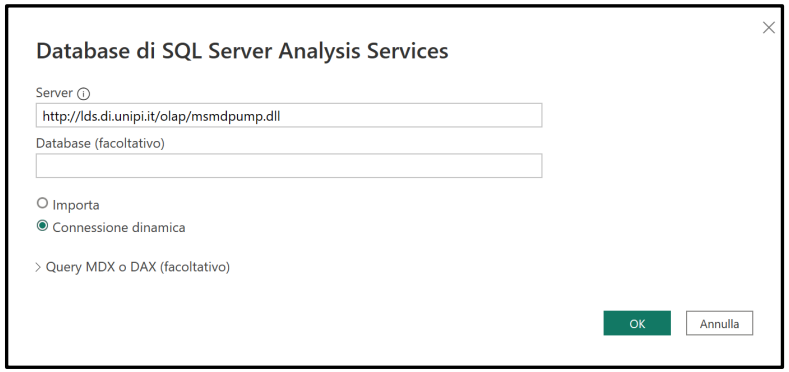
Group 24 Project Presentation –
Davide Chen & Andrea Ribellino

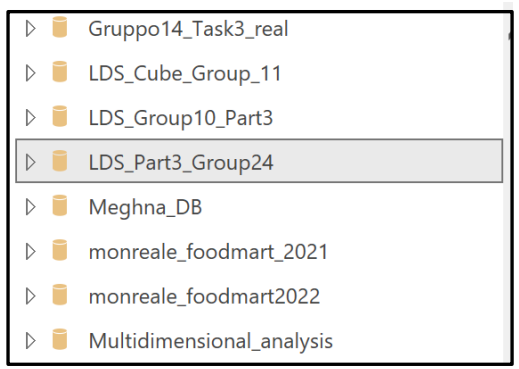
MICROSOFT POWER BI - CONNECTION

Power BI

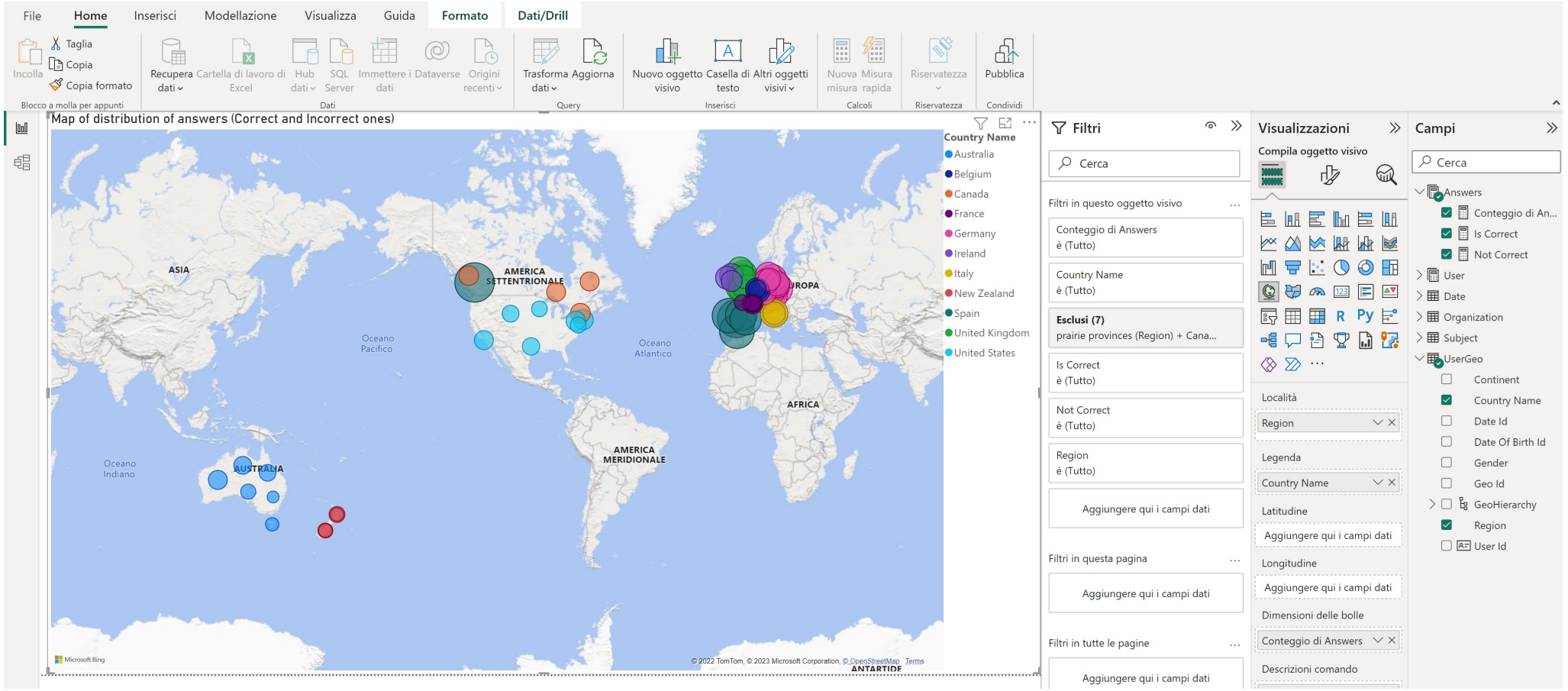




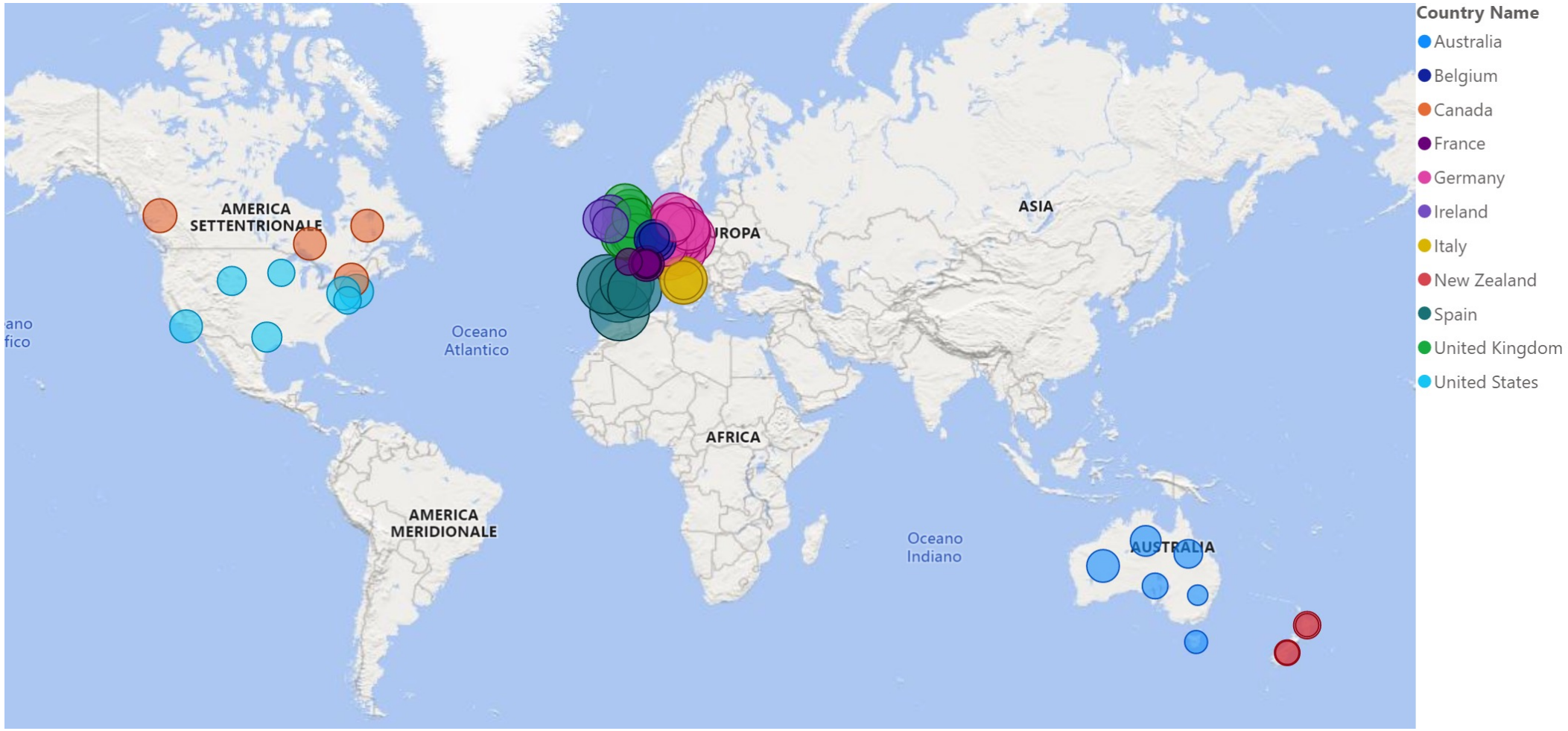




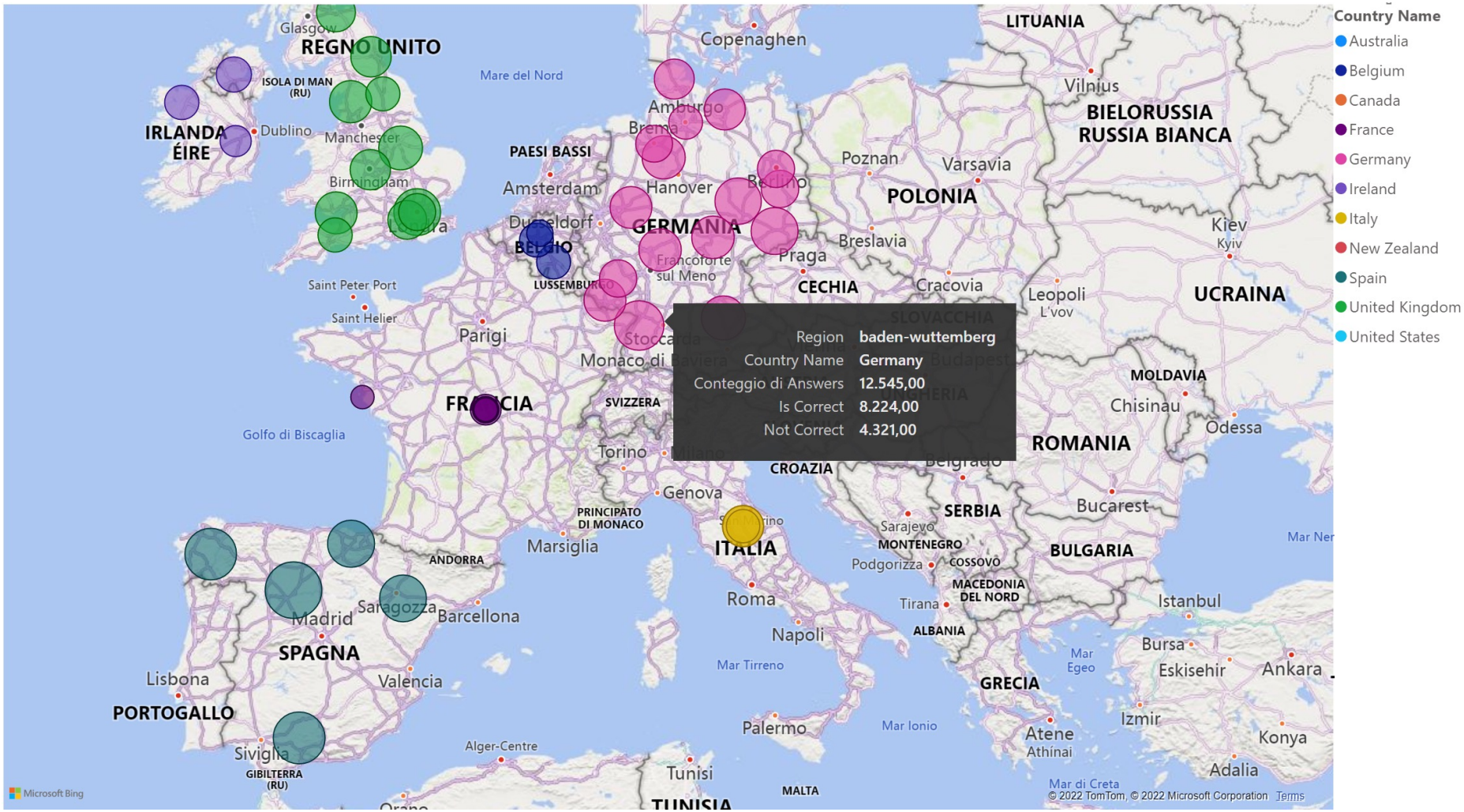
MICROSOFT POWER BI - USAGE



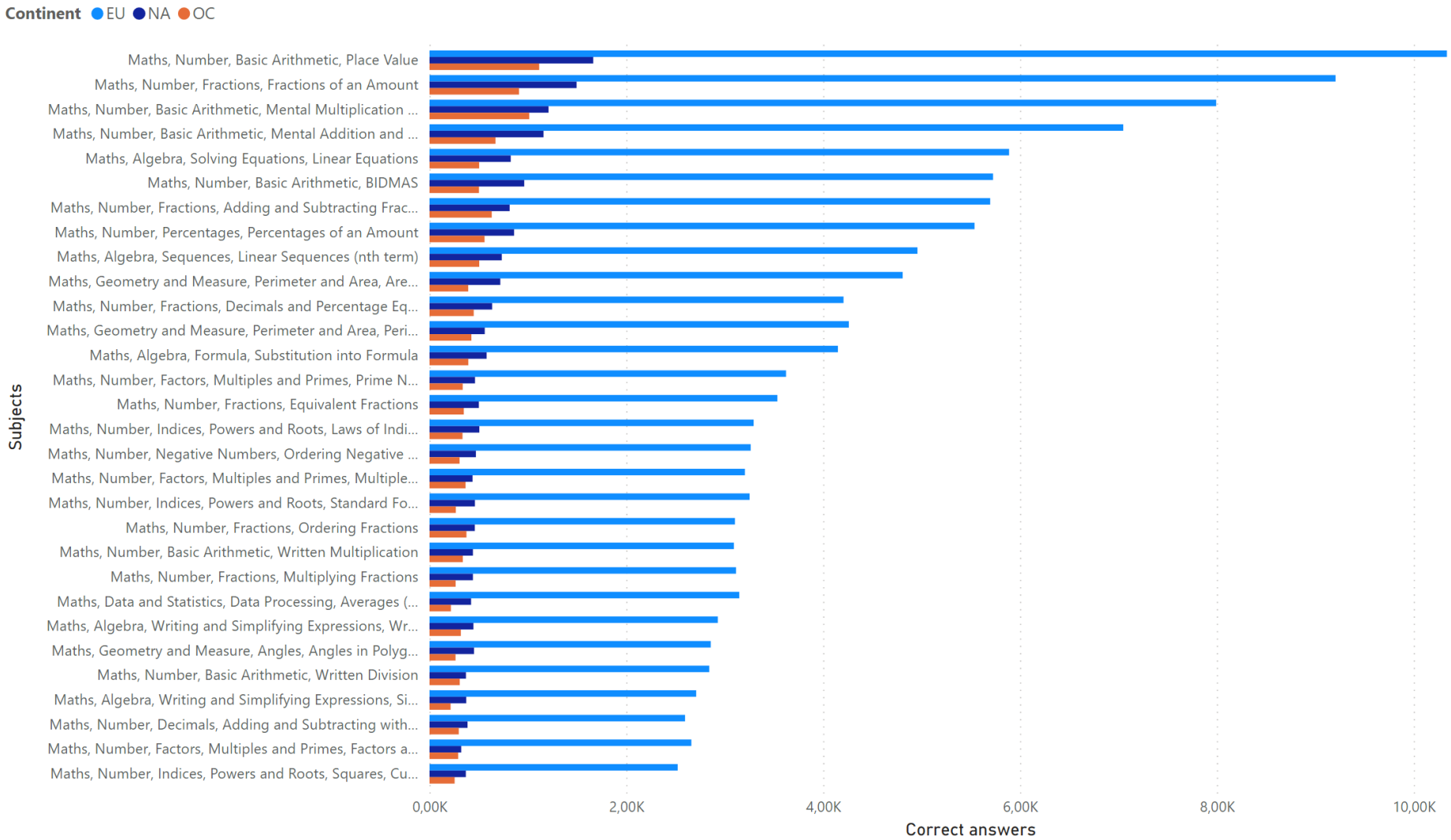
MAP OF ANSWERS DISTIBUTIONS



ZOOM-IN TO EUROPE



CORRECT ANSWERS BY SUBJECT AND CONTINENT



THANKS FOR THE ATTENTION

Group 24

Davide Chen

Email: d.chen@studenti.unipi.it

Andrea Ribellino

Email: a.ribellino@studenti.unipi.it