

Enabling a Data-informed Public Sector: An Introductory Course to BDTI Essentials

Session 3: Data Blending and Storage



Welcome!
The webinar will begin shortly.

DG CNECT

Directorate-General for Communications
Networks, Content and Technology

DG DIGIT

Directorate-General
for Digital Services



Some housekeeping



- The Chat and Q&A widgets – **Bottom right of the screen**
- Please post session-related questions in **Q&A**
- The session will be recorded and available shortly after the end of the session on our website

Do

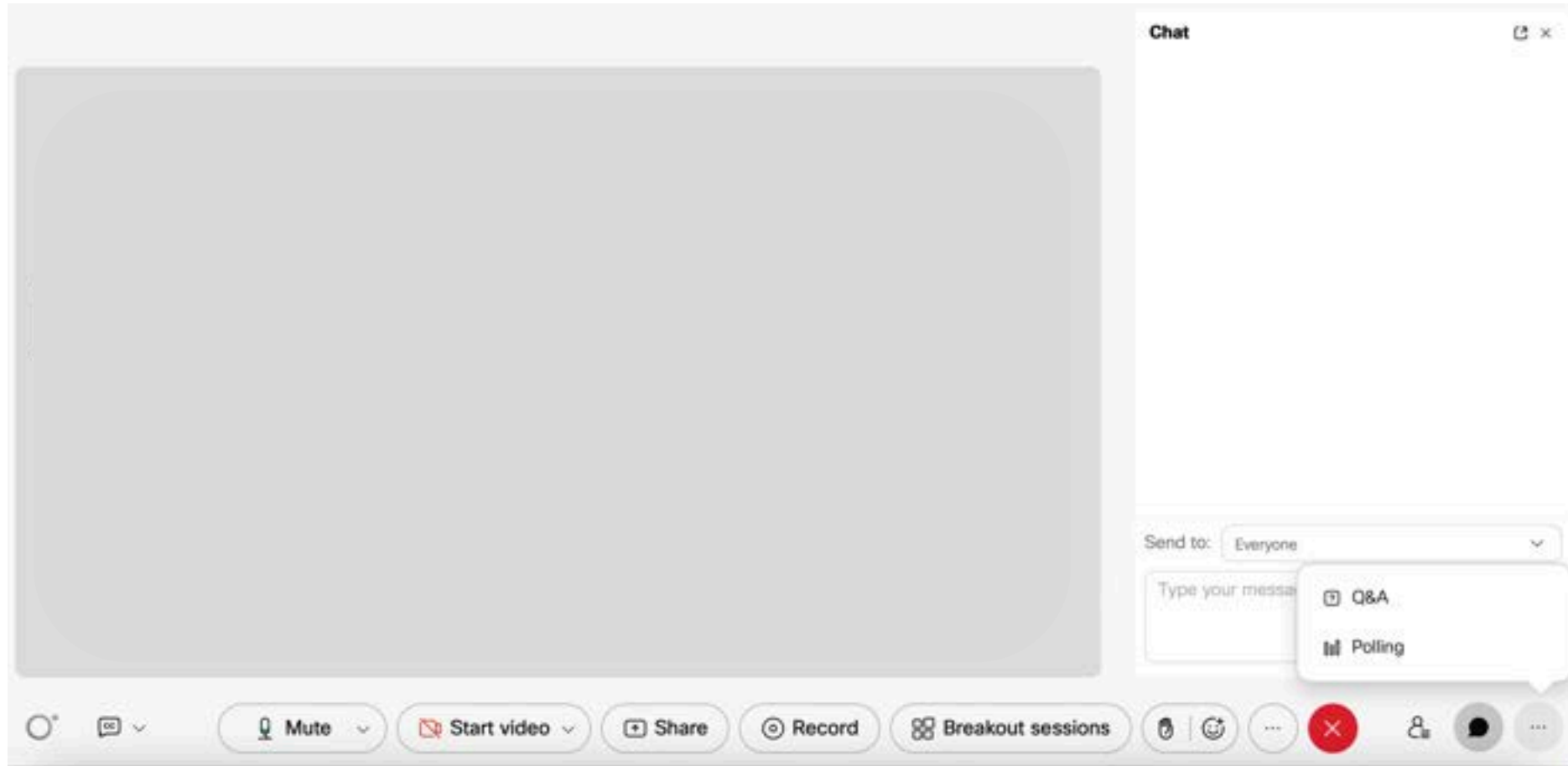
- Participate in the chat during the session
- Ask questions during the Q&A portion at the end
- Give feedback after the session
- Join the dedicated discussion board

Don't

- Self-promote
- Disrespect anyone



Locate Chat and Q&A - Webex



Enabling a Data-informed Public Sector: An Introductory Course to BDTI Essentials

Session 3: Data Blending and Storage



Agenda

- Welcome [11:00-11:05]
- Session 2 Recap [11.05-11.15]
- Data blending and storage [11:15-12:00]
- Q&A [12:00-12:10]
- Wrap up [12:10 – 12:15]

What is the Big Data Test Infrastructure (BDTI) ?



Not **only** for big data, for **public sector in general (open data)**

You have the key ingredients (datasets),
we equip you with the best **open-source tool**
to create amazing recipes for **public good**.



Six months free of charge service
for EU public administrations *



Ready-to-use
data analytics stack and
support



Cloud platform based on
open-source tools

* The cost of the pilot project must fit within the funding boundaries of the BDTI pilot budget



Who is the Big Data Test Infrastructure (BDTI) for?



European Public Administrations

All European Public Administrations at **local, regional and national level** can independently apply for a BDTI pilot project



Ecosystem with **academia** and **private sector**

Academia, spin-off, startups can apply for pilot projects once there is a **clear collaboration** with a Public Administration which will be the main point of contact for the project (**Master/PhD, GovTech startups**)



Are you working for a public administration in need of infrastructure for data analytics?

<https://big-data-test-infrastructure.ec.europa.eu/>



Open-source tools to support your data journey



100%

5. Decision-Making

1. Collection

4. Visualisation

Orchestration

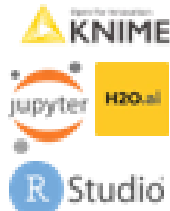
3. Analysis

2. Processing

Metabase

Apache Superset

Development Environments



MINIO

Data Lake

OPEN LINK VIRTUOSO UNIVERSAL SERVER

mongoDB

Database



elasticsearch

Advanced Processing Engines

APACHE Spark



Use case:

"Identify Innovative Green Energy Projects from EU Universities"

Zoi and her team, work in the **Educational Department of a highly polluted EU region.**

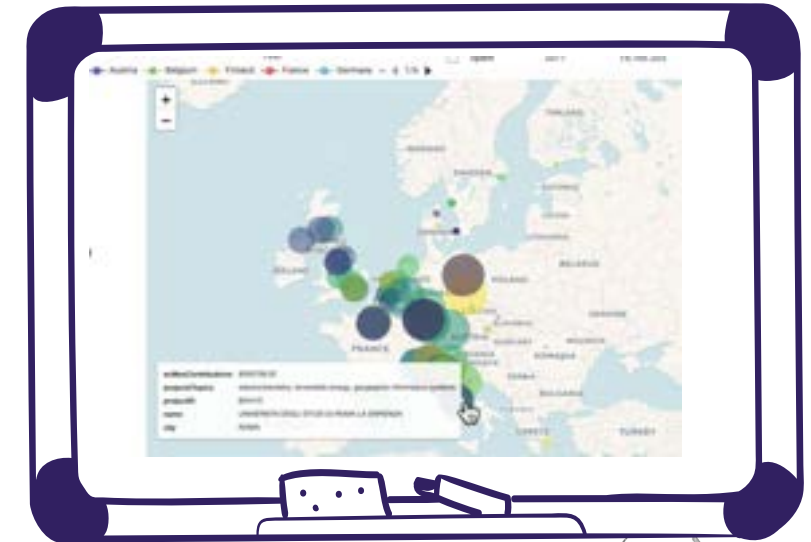
Their **purpose** is to identify potential **"green energy partnerships"** between EU universities to address pollution issues.

Zoi reached out the BDTI team to learn how to **connect the dots**, use the free BDTI playground to **create insights** and produce a **report**, which can be **easily updated** when new data is available.



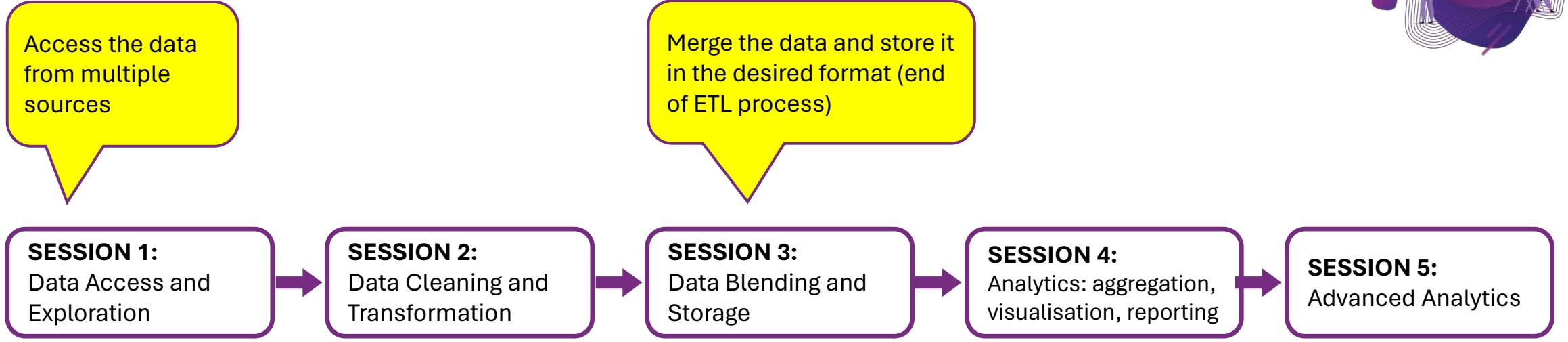
They found some interesting **open data**:

- **List of innovation green energy projects** across EU countries
- **CO2 emissions** of these countries



Recap: Zoi Data Journey

<https://code.europa.eu/bdti/bdti-essentials-course>



Access the data from multiple sources

Merge the data and store it in the desired format (end of ETL process)

EU Open data portal:

data.europa.eu

Formats:
csv, Excel, JSON

Tools:




Clean and transform the data to the correct shape





Open for Innovation




Make sense of data, extract insights and visualise them in a report

Enhance analytics by gathering data from the internet





Recap Session 2 – Data Cleaning and Transformation

Why Clean and Transform?

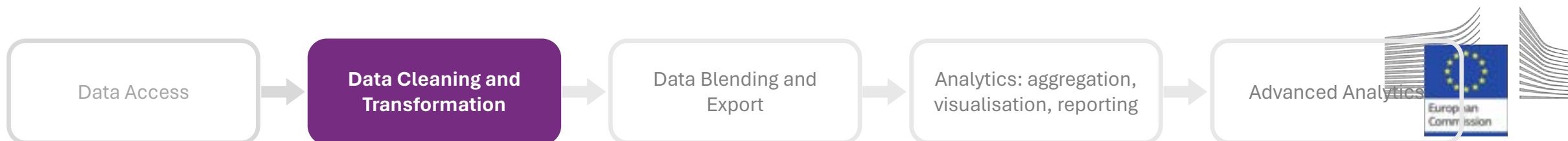
- The **core** of the data pipeline
- Data is rarely **clean**
 - Not useful data (columns/rows)
 - Duplicated data
 - Missing values
 - And more...
- Data not in the **correct shape**
 - **Extract** more information
 - Transform at cell or row **level**



Use Case

With an overview of the data, Zoi realises that it is not in the best shape to do an analysis, some columns are in the wrong format and she is also missing data.

In this session, she learned techniques **to clean and transform the data.**



How?



Cleaning

Explore Data: make decisions according to objectives

Filter columns-rows: Remove what is not needed

Missing Values or Duplicates: decide how to handle them based on needs

From: 25 columns

Rows: 177834 | Columns: 25

#	RowID	projectID	projectAc...	organisat...	vathNumber	name	shortName	SME
1	Row0	879926	EEN SACHSEN	999913831	⊖	INDUSTRIE-U...	BHK C	false
2	Row1	879926	EEN SACHSEN	994556036	DE176093942	ZTS-ZENTRU...	ZTS	true
3	Row2	879926	EEN SACHSEN	999797625	⊖	INDUSTRIE- U...	BHK DRESDEN	false
4	Row3	879926	EEN SACHSEN	999940409	⊖	HANDWERKS...	HWK DRESDEN	false
5	Row4	879926	EEN SACHSEN	999877844	DE140857609	TECHNISCHE...	⊖	false
6	Row5	879926	EEN SACHSEN	999914219	⊖	INDUSTRIE-U...	BHK-L	false
7	Row6	879926	EEN SACHSEN	969246408	⊖	HOCHSCHUL...	Hochschule Z...	false
8	Row7	879926	EEN SACHSEN	999917905	DE141484057	AGENTUR FU...	AGIL	false
9	Row8	740689	HEIMDALL	997710476	IT015032900...	CENTRO INTE...	FONDAZIONE...	false
10	Row9	740689	HEIMDALL	940057847	DK37222135	FREDERIKSB...	FBBR	false
11	Row10	740689	HEIMDALL	999703438	ESS0811001G	Departament ...	INT	false
12	Row11	740689	HEIMDALL	919049102	IT136697210...	ASSOCIAZIO...	⊖	false
13	Row12	740689	HEIMDALL	999778322	EL094149709	SPACE HELL...	SPACE HELL...	true
14	Row13	740689	HEIMDALL	999638739	ES062616586	CENTRE TEC...	CTTC	false



To: 7 columns

Rows: 177834 | Columns: 7

#	RowID	projectID	name	activity...	street	postCo...	city	country
1	Row0	879926	INDUSTRIE-UND H...	PUB	STRASSE DER ...	09111	CHEMNITZ	DE
2	Row1	879926	ZTS-ZENTRUM FUR...	OTH	INDUSTRIESTR...	01612	GLAUBITZ	DE
3	Row2	879926	INDUSTRIE- UND H...	PUB	LANGER WEG 4	01239	DRESDEN	DE
4	Row3	879926	HANDWERKSKAM...	PUB	AM LAGERPLA...	01099	DRESDEN	DE
5	Row4	879926	TECHNISCHE UNIV...	HES	STRASSE DER ...	09111	Chemnitz	DE
6	Row5	879926	INDUSTRIE-UND H...	PUB	GOERDELEERRI...	04109	LEIPZIG	DE
7	Row6	879926	HOCHSCHULE ZITT...	HES	THEODOR-KOE...	02763	Zittau	DE
8	Row7	879926	AGENTUR FUR INN...	PRC	LESSINGSTRA...	04109	LEIPZIG	DE
9	Row8	740689	CENTRO INTERNAZ...	REC	VIA ARMANDO...	17100	SAVONA	IT
10	Row9	740689	FREDERIKSBORG B...	PRC	LOGISMOSE 3	3600	FREDERIKSSU...	DK
11	Row10	740689	Departament d'Inter...	PUB	Carrer Diputaci...	08009	Barcelona	ES
12	Row11	740689	ASSOCIAZIONE DE...	OTH	VIA BERNARDI...	00151	Roma	IT
13	Row12	740689	SPACE HELLAS AN...	PRC	MESSOGION A...	153 41	Aghia Paraske...	EL
14	Row13	740689	CENTRE TECNOLO...	REC	AVINGUDA CA...	08860	Castelldefels ...	ES

How?



Transformation

Explore Data: check data types, missing information

Change shape: Extract data and add columns

Modify data at a cell level: correct strings misspellings, group multiple titles.

From

projectID String	euroSciVocCode String	euroSciVocPath String	euroSciVocTitle String
869855	/29/101/555/1359	/social sciences/sociolog...	automation
869855	/25/63/399	/engineering and technol...	textiles
869855	/25/73/453/58525161	/engineering and technol...	sensors
869855	/25/73/453/459	/engineering and technol...	robotics
869855	/29/93/47293307	/social sciences/psychol...	ergonomics
870148	/25/63/409	/engineering and technol...	colors
883285	/21/33/121/44109686/5...	/medical and health scien...	ebola
883285	/21/33/137/133/9678651	/medical and health scien...	pandemics
883285	/21/33/121/44109686/7...	/medical and health scien...	influenza
000000	/20/01/522/1212	/social sciences/econom...	employment



To

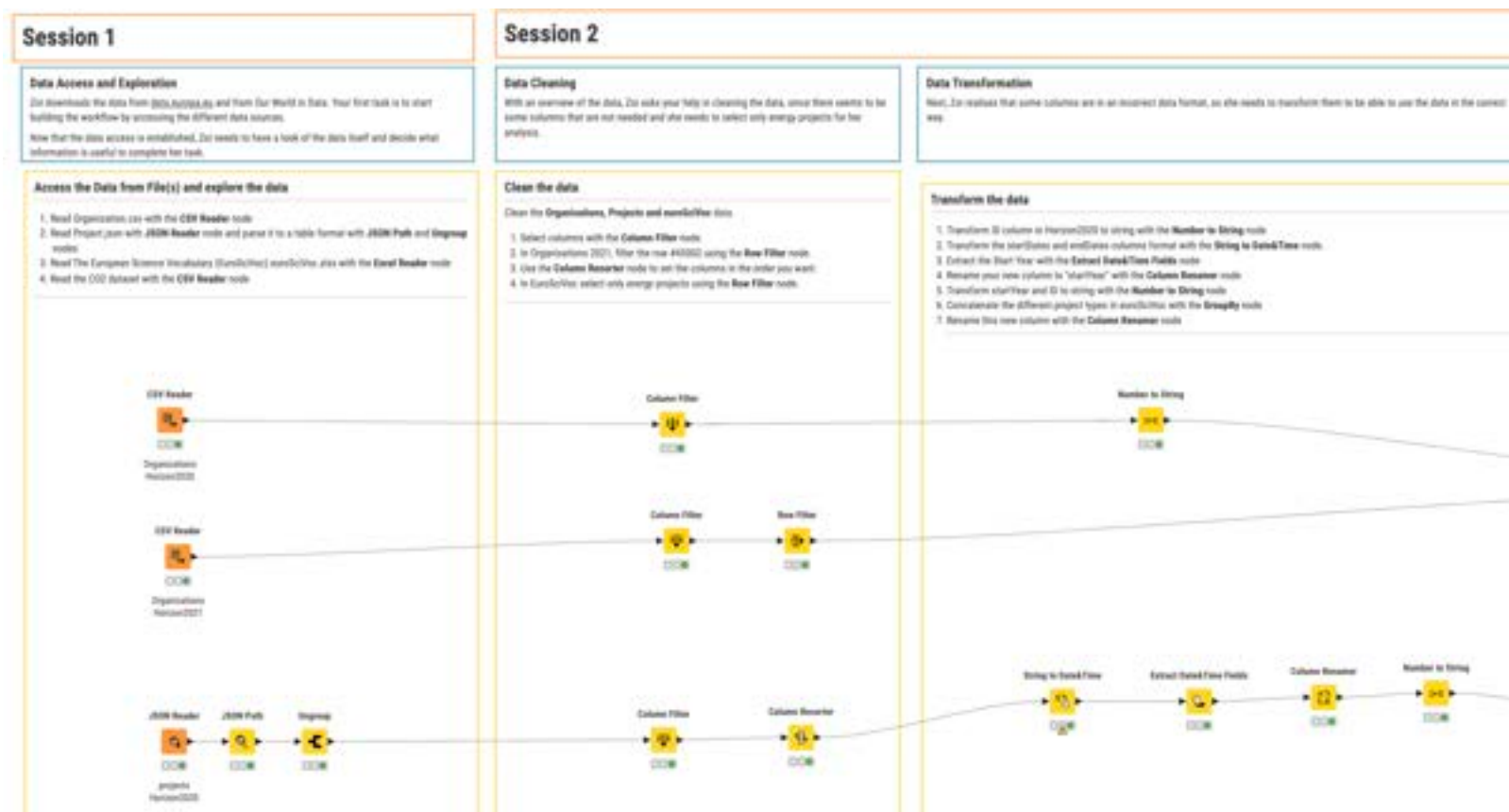
projectID String	projectsTopics String
101039110	energy conversion
101039446	energy and fuels
101039576	renewable energy
101039746	solar energy, energy conversion
101040341	geothermal energy
101040379	energy conversion
101040669	energy conversion
101040994	energy conversion
101041768	solar energy
101041809	solar energy
101042381	hydrogen energy



Session 2 Output: Workflow

- All these steps are in one single workflow that we can re-run and do it again!

<https://code.europa.eu/bdti/bdti-essentials-course>





Exercise Review

BDTI access refresh

To have access, public administrations need to ***apply on the official website***

- https://big-data-test-infrastructure.ec.europa.eu/apply-bdti_en

Once you have access follow these steps:

Step 1 > Access the portal via the website

Step 2 > Navigate to the "Service Catalog" and launch the tools

Step 3 > Go to "My Services" and open the tool you created

The screenshot shows the official website for the Big Data Test Infrastructure (BDTI). At the top, there is the European Commission logo and a search bar. The main navigation bar includes links for Home, About, Service Offering, Resources, Apply for BDTI, Success Stories, BDTI Portal (highlighted), What's new?, and FAQ. The page title is 'Big Data Test Infrastructure (BDTI)'. The main content area is titled 'BDTI Portal' and contains the following text: 'The BDTI portal is a web application which allows users to easily deploy and manage containerized data science workloads. In this section, you can access the portal and find documentation about the portal.' Below this, there is a section titled 'Access the BDTI Portal' with a disclaimer: 'Disclaimer: The BDTI portal is only available to users who have a BDTI pilot.' and a link to 'The user documentation for the BDTI portal can be found here'. The link 'Access the BDTI Portal' is highlighted with a purple oval. At the bottom, there is a 'Share this page' section with social media icons for Twitter, Facebook, LinkedIn, and Email, and a 'More share options' link. The footer contains the European Commission logo and the text 'Big Data Test Infrastructure (BDTI)'.

Exercise review



Use Case:
 An e-commerce store is experiencing a performance issue with its product recommendation engine. The product is a client portal using a recommendation engine to suggest products to customers based on their browsing history. The client portal is built using a microservices architecture. The client portal is built using a microservices architecture. The client portal is built using a microservices architecture.

Session 1

Data Access and Exploration:
 We need to explore the data to understand the performance issue. The first step is to identify the data sources and the data schema. The second step is to explore the data to understand the performance issue. The third step is to identify the data sources and the data schema. The fourth step is to explore the data to understand the performance issue.

Session 2

Data Cleaning:
 We need to clean the data to ensure it is accurate and consistent. The first step is to identify the data quality issues. The second step is to clean the data to ensure it is accurate and consistent. The third step is to identify the data quality issues. The fourth step is to clean the data to ensure it is accurate and consistent.

Data Transformation:
 We need to transform the data to make it suitable for analysis. The first step is to identify the data transformation requirements. The second step is to transform the data to make it suitable for analysis. The third step is to identify the data transformation requirements. The fourth step is to transform the data to make it suitable for analysis.

Prepare D3D Query Tables and Filter D3D Data:
 We need to prepare the data for D3D queries. The first step is to identify the D3D query requirements. The second step is to prepare the data for D3D queries. The third step is to identify the D3D query requirements. The fourth step is to prepare the data for D3D queries.

Filter only D3D Conditions:
 We need to filter the data to only include D3D conditions. The first step is to identify the D3D conditions. The second step is to filter the data to only include D3D conditions. The third step is to identify the D3D conditions. The fourth step is to filter the data to only include D3D conditions.



Session 3: Data Blending and Storage

Session 3: Data Blending and Storage



Session	Topic	Duration
Webinar 1	Data Access and Exploration	45 min
Webinar 2	Data Cleaning and Transformation	75 min
Webinar 3	Data Blending and Storage	75 min
Webinar 4	Analytics	75 min
Webinar 5	Advanced Module: Gathering Data from the Web and Geo Visualisation	75 min

- Slides, hands-on exercises, and solutions will be provided on the webinar
- Gitlab repository <https://code.europa.eu/bdti/bdti-essentials-course>



Learning Outcomes



At the end of this session, you should be able to:

1. Sign in to the BDTI framework and initialise KNIME, PgAdmin, and PostgreSQL
2. Stack multiple tables
3. Join tables in multiple ways
4. Write data to various file formats
5. Store the data in a database





Data Blending



Data Blending

Combine data from different sources and create a single table

- 1) **Stack** new data at the bottom of a table
→ same columns, **add rows**
- 2) **Join** information from tables by defining matching criteria → **add new columns**

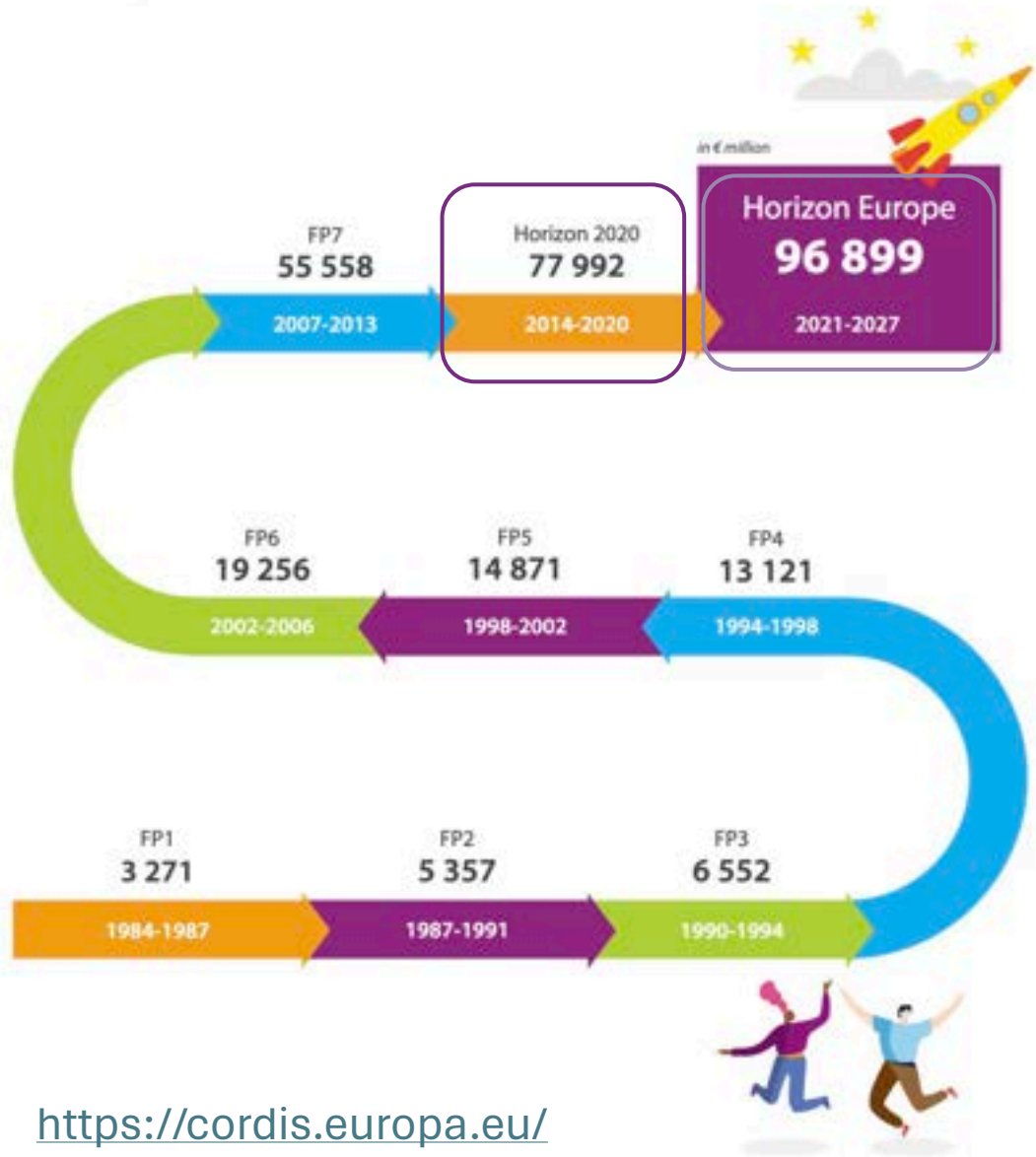


Use Case

With the data cleaned and in the correct shape, we will now help Zoi to **merge** the different data sources, by stacking tables and matching or joining others to add the necessary columns.

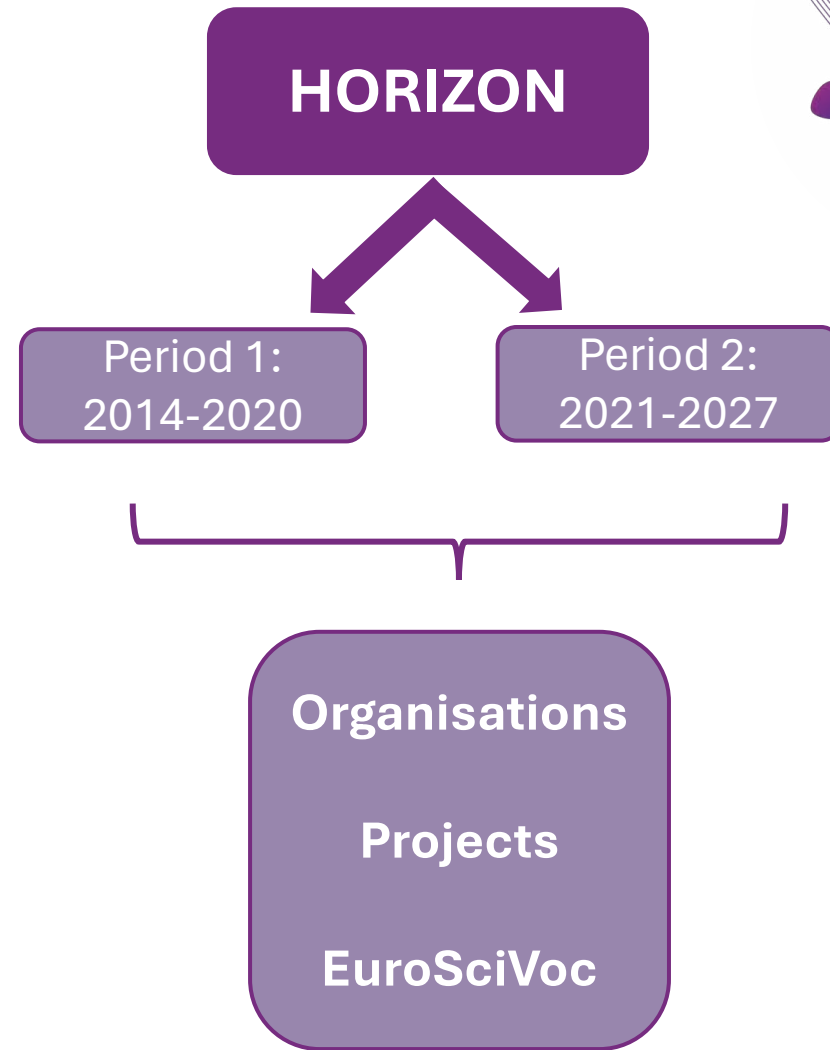
EVOLUTION OF THE R&I FRAMEWORK PROGRAMMES BUDGET

The EU budget has been significantly rising over time from EUR 3 271 million in FP1 to EUR 96 899 million today for Horizon Europe and Euratom. This illustrates the clear ambition of the EU for its main R&I funding instrument and increases relevance of science and technology in society to address our challenges.



<https://cordis.europa.eu/>

This budget compares favourably with the previous programme, Horizon 2020. In real terms, and excluding



Data Blending



Stack

Stack new data at the bottom of a table -> same columns, **add rows**

Join

Join information from tables by defining matching criteria -> **add new columns**

Stack data to add rows

Stack

Stack new data at the bottom of a table -> same columns, **add rows**

Table 1		
ID	Column 1	Column 2
Id-1		
Id-2		

Table 2		
ID	Column 1	Column 2
Id-3		
Id-4		

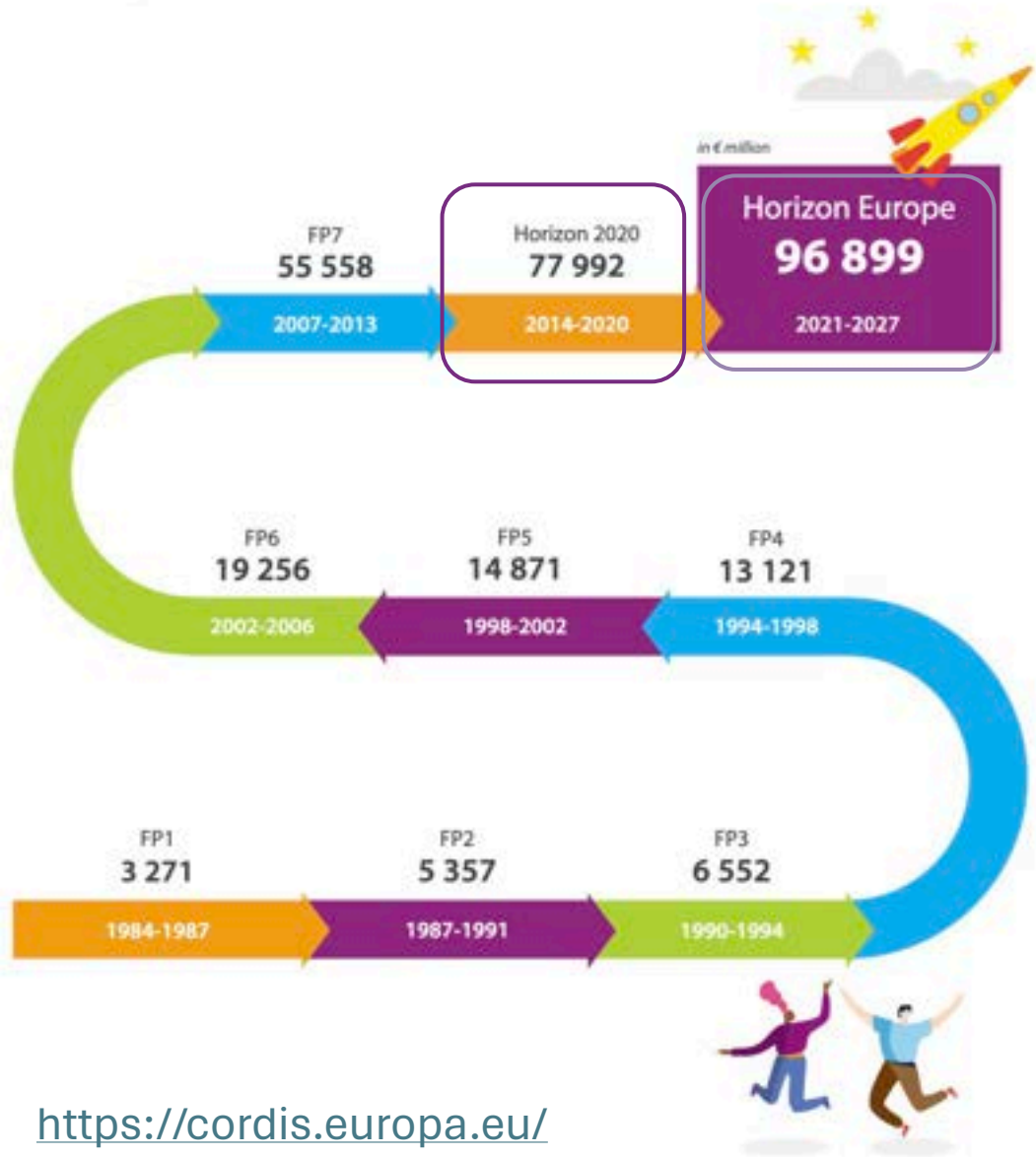


Result Table		
ID	Column 1	Column 2
Id-1		
Id-2		
Id-3		
Id-4		



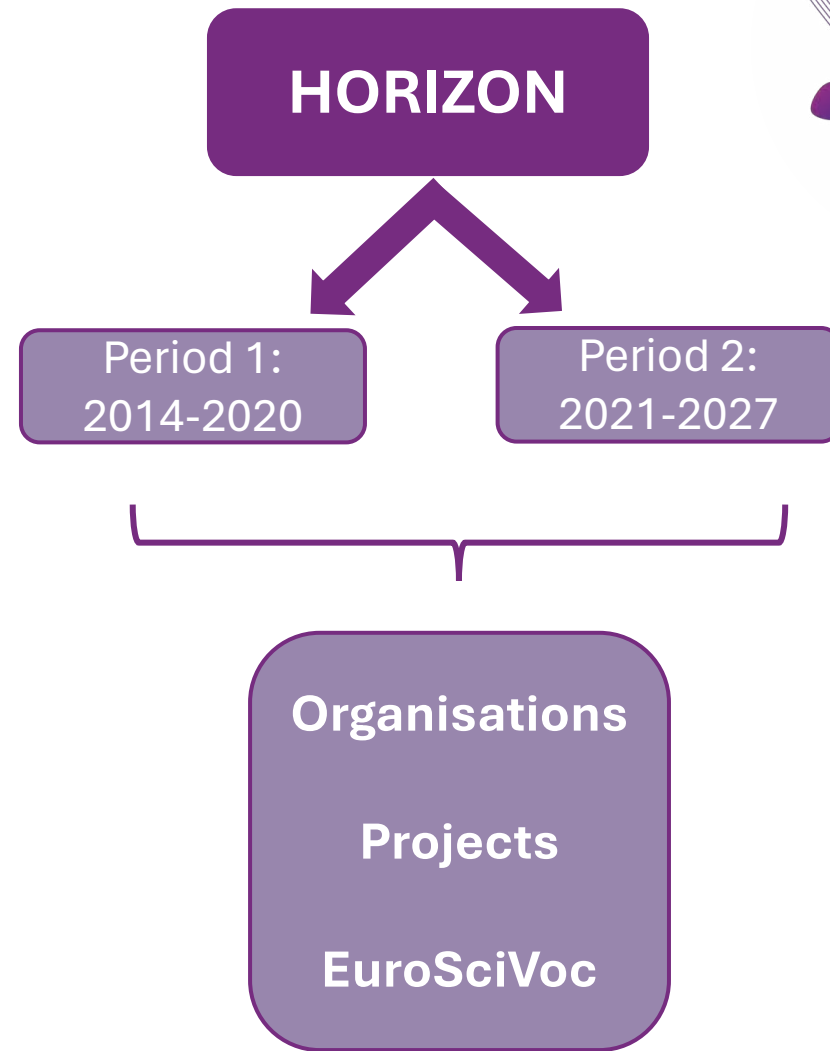
EVOLUTION OF THE R&I FRAMEWORK PROGRAMMES BUDGET

The EU budget has been significantly rising over time from EUR 3 271 million in FP1 to EUR 96 899 million today for Horizon Europe and Euratom. This illustrates the clear ambition of the EU for its main R&I funding instrument and increases relevance of science and technology in society to address our challenges.



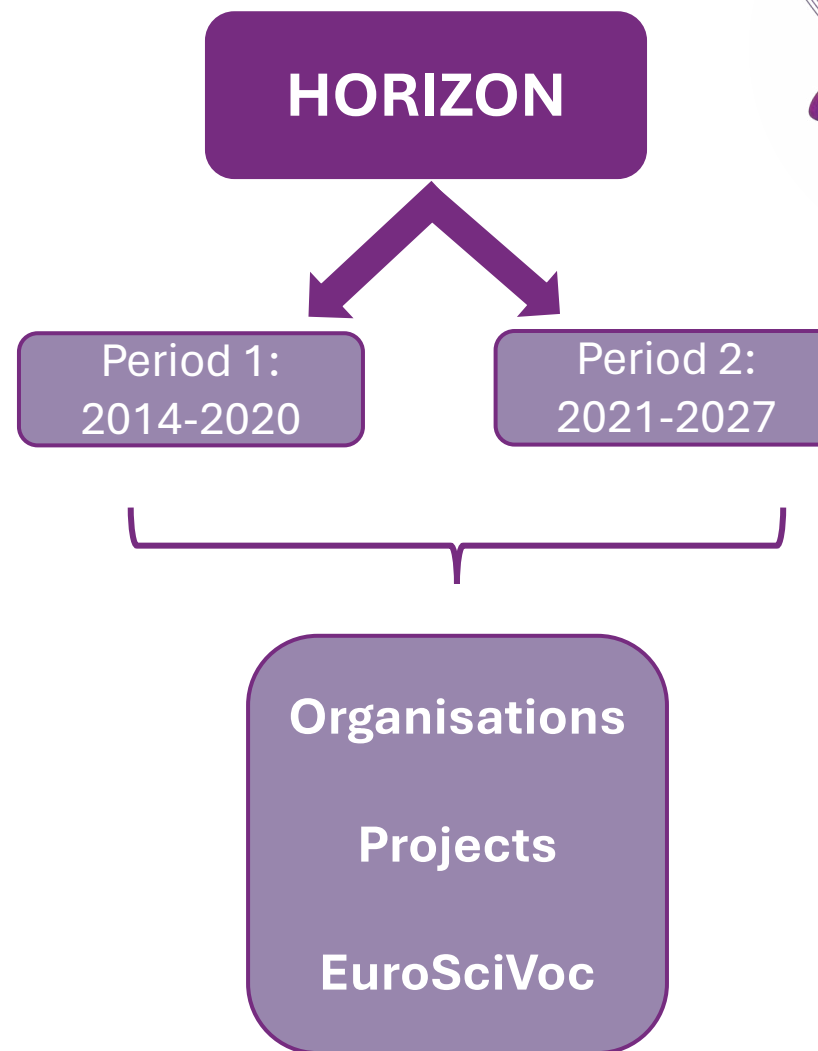
<https://cordis.europa.eu/>

This budget compares favourably with the previous programme, Horizon 2020. In real terms, and excluding



1) Why stack data?

- To combine two or more tables/datasets with the **same structure** but different data, thus, **add rows**.
- Example: Horizon from 2 periods that contain data for different years



1) Stack data: add rows



Horizon data from different periods:

Organisations

Rows: 177834 | Columns: 7

#	RowID	projectID	name	activityTy...	street	postCode	city	country
1	Row0	879926	INDUSTRIE-U...	PUB	STRASSE DE...	09111	CHEMNITZ	DE
2	Row1	879926	ZTS-ZENTRU...	OTH	INDUSTRIEST...	01612	GLAUBITZ	DE
3	Row2	879926	INDUSTRIE-U...	PUB	LANGER WEG...	01239	DRESDEN	DE
4	Row3	879926	HANDWERKS...	PUB	AM LAGERPL...	01099	DRESDEN	DE
5	Row4	879926	TECHNISCHE...	HES	STRASSE DE...	09111	Chemnitz	DE
6	Row5	879926	INDUSTRIE-U...	PUB	GOERDELERR...	04109	LEIPZIG	DE
7	Row6	879926	HOCHSCHUL...	HES	THEODOR-KO...	02763	Zittau	DE
8	Row7	879926	AGENTUR FU...	PRC	LESSINGSTR...	04109	LEIPZIG	DE
9	Row8	740689	CENTRO INT...	REC	VIA ARMAND...	17100	SAVONA	IT
10	Row9	740689	FREDERIKSB...	PRC	LOGISMOSE 3	3600	FREDERIKSS...	DK
11	Row...	740689	Departament...	PUB	Carrer Diputa...	08009	Barcelona	ES
12	Row...	740689	ASSOCIAZIO...	OTH	VIA BERNAR...	00151	Roma	IT
13	Row...	740689	SPACE HELL...	PRC	MESSOGION ...	153 41	Aghia Parask...	EL

Rows: 55994 | Columns: 7

#	RowID	projectID	name	activityType	street	postCode	city	country
1	Row0	101091483	LA PALMA RES...	PRC	CALLE EL CAST...	38787	Santa Cruz De L...	ES
2	Row1	101091483	GEOLOGIAN TU...	REC	VUORIMIEHENT...	02151	Espoo	FI
3	Row2	101091483	INTERNATIONA...	OTH	RUE VAUTIER 54	1000	Bruxelles / Brus...	BE
4	Row3	101091483	PROMAN MAN...	PRC	WEINGARTENS...	2214	Auersthal	AT
5	Row4	101091483	ALDA - ASSOCI...	OTH	PLACE DES OR...	67000	Strasbourg	FR
6	Row5	101091483	REGIONE EMILI...	PUB	VIALE ALDO MO...	40127	Bologna	IT
7	Row6	101091483	TELOS ALEFF L...	PRC	36 ERITH HIGH ...	DAB 1QY	ERITH	UK
8	Row7	101091483	PETRONAVIT AS	PRC	STOKKAHAGEN...	4022	Stavanger	NO
9	Row8	101091483	MACCABE DUR...	PRC	20 FITZWILLIA...	D02YV58	DUBLIN 2	IE
10	Row9	101091483	GENERATOR BE...	PRC	QUINTA DA GA...	2560629	LISBOA	PT
11	Row...	101091483	ISTITUTO SUPE...	REC	VIA VITALIANO ...	00144	Roma	IT
12	Row...	101091483	GEOLOGIA E GE...	PRC	RUA CUNHA JU...	4250-186	Porto	PT

Period 1: 2014- 2020

Rows: 177834

Number of Columns: 7

Period 2: 2021-2027

Rows: 55994

Number of Columns: 7



1) Stack data: add rows



Horizon data from different periods: Projects

Rows: 35386 | Columns: 7

#	RowID	ids	startDates	endDates	ecMaxContrl...	totalCosts	objectives	startYear
		String	Local Date	Local Date	Number (Double)	Number (Double)	String	String
1	Row...	879926	2020-01-01	2021-12-31	125,559	125,560	The aim of the pr...	2020
2	Row...	740689	2017-05-01	2021-01-31	7,836,370.63	8,585,414.19	The project aims ...	2017
3	Row...	730082	2016-11-01	2019-04-30	1,500,000	1,850,000	E2mC aims at de...	2016
4	Row...	101023342	2021-06-01	2024-11-30	3,755,470.75	5,214,286.25	Sustainability in c...	2021
5	Row...	831993	2019-01-01	2021-12-31	796,056.25	796,056.25	New technologie...	2019
6	Row...	641451	2015-01-01	2016-12-31	999,653.25	1,323,315	GEO-VISION is an...	2015
7	Row...	731665	2017-01-01	2018-12-31	1,498,857.5	1,498,857.5	The objective of t...	2017
8	Row...	653866	2015-06-01	2017-11-30	1,990,114.25	2,018,034.75	This proposal se...	2015
9	Row...	770037	2017-11-01	2019-10-31	2,003,918.75	2,003,918.75	Since 2015, migr...	2017
10	Row...	636906	2015-02-01	2017-01-31	610,013.75	792,938.75	The worldwide ec...	2015
11	Row...	825132	2019-01-01	2022-10-31	5,121,692.5	5,121,692.5	Current industrial...	2019
12	Row...	780785	2018-01-01	2021-12-31	2,940,920	3,877,451.43	Micro-RDS will be...	2018
13	Row...	952633	2020-11-01	2023-10-31	4,480,308.75	4,480,308.75	Despite the evide...	2020

Rows: 9000 | Columns: 7

#	RowID	ids	startDates	endDates	ecMaxContrl...	totalCosts	objectives	startYear
		String	Local Date	Local Date	Number (Double)	Number (Double)	String	String
1	Row...	101091483	2023-01-01	2025-12-31	2,291,975	2,291,975	CBAN will develo...	2023
2	Row...	101058526	2022-04-01	2025-03-31	3,998,440	3,998,440	The introduction ...	2022
3	Row...	101058362	2022-12-01	2026-05-31	4,645,639	6,762,123.75	Each year the EU ...	2022
4	Row...	101060904	2022-09-01	2025-08-31	2,499,951	2,499,951	Europe's regional ...	2022
5	Row...	101080025	2022-10-01	2025-09-30	2,486,190	2,507,255	Housing policy in...	2022
6	Row...	101121171	2023-08-01	2024-07-31	910,966.65	0	SHIELD4CROWD ...	2023
7	Row...	101070085	2022-09-01	2024-08-31	4,293,263.5	4,293,263.75	The BAMBAM pr...	2022
8	Row...	101059425	2022-10-01	2026-09-30	2,999,605	2,999,609.38	Secure attachme...	2022
9	Row...	101120323	2023-10-01	2026-09-30	2,908,625	3,550,250	Procedural Knowl...	2023
10	Row...	101082630	2022-11-01	2024-10-31	2,334,143.75	2,958,812.5	"On B2B and B2C...	2022
11	Row...	101091967	2023-01-01	2026-12-31	6,584,313.5	8,829,675	Buildings are res...	2023
12	Row...	101094391	2023-01-01	2025-12-31	999,235	999,235	Building Gender+	2023
13	Row...	101111996	2023-09-01	2027-02-28	4,683,365.49	4,683,365.49	The general obje...	2023

Period 1: 2014- 2020

Rows: 35386

Columns: 7

Period 2: 2021-2027

Rows: 9000

Columns: 7



1) Stack data: add rows



Horizon data from different periods:

euroSciVoc

Rows: 2461 | Columns: 2

<input type="checkbox"/>	#	RowID	projectID String	projectsTopics String
<input type="checkbox"/>	1	Row0	101000158	renewable energy
<input type="checkbox"/>	2	Row1	101000470	renewable energy
<input type="checkbox"/>	3	Row2	101000496	energy and fuels
<input type="checkbox"/>	4	Row3	101000612	energy and fuels
<input type="checkbox"/>	5	Row4	101000785	geothermal energy
<input type="checkbox"/>	6	Row5	101000828	hydrogen energy
<input type="checkbox"/>	7	Row6	101000832	renewable energy
<input type="checkbox"/>	8	Row7	101000987	energy and fuels
<input type="checkbox"/>	9	Row8	101001078	energy and fuels
<input type="checkbox"/>	10	Row9	101001081	renewable energy
<input type="checkbox"/>	11	Row...	101001514	energy conversion
<input type="checkbox"/>	12	Row...	101001626	energy conversion

Period 1: 2014- 2020

Rows: 2461

Columns: 2

Rows: 470 | Columns: 2

<input type="checkbox"/>	#	RowID	projectID String	projectsTopics String
<input type="checkbox"/>	1	Row0	101039110	energy conversion
<input type="checkbox"/>	2	Row1	101039446	energy and fuels
<input type="checkbox"/>	3	Row2	101039576	renewable energy
<input type="checkbox"/>	4	Row3	101039746	solar energy, energy conversion
<input type="checkbox"/>	5	Row4	101040341	geothermal energy
<input type="checkbox"/>	6	Row5	101040379	energy conversion
<input type="checkbox"/>	7	Row6	101040669	energy conversion
<input type="checkbox"/>	8	Row7	101040994	energy conversion
<input type="checkbox"/>	9	Row8	101041768	solar energy
<input type="checkbox"/>	10	Row9	101041809	solar energy
<input type="checkbox"/>	11	Row...	101042781	hydrogen energy

Period 2: 2021-2027

Rows: 470

Columns: 2



Stack in the different BDTI tools (add rows)



SQL

```
SELECT * FROM table1  
UNION ALL  
SELECT * FROM table2;
```

R

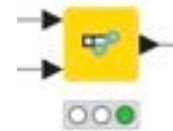
```
library(dplyr)  
df1 <- read.csv('table1.csv')  
df2 <- read.csv('table2.csv')  
combined_df <- bind_rows(df1, df2)
```

PYTHON

```
import pandas as pd  
df1 = pd.read_csv('table1.csv')  
df2 = pd.read_csv('table2.csv')  
combined_df = pd.concat([df1, df2], ignore_index=True)
```

KNIME

Concatenate



Intersection
of columns

1) How to stack data?

Steps:

1. Identify matching columns
2. Decide how to combine columns
3. Decide on how to handle duplicate row ids





1) Stack data: add rows

Rows: 177834 | Columns: 7

#	RowID	projectID	name	activityType
1	Row0	879926	INDUSTRIE-U...	PUB
2	Row1	879926	ZTS-ZENTRU...	OTH
3	Row2	879926	INDUSTRIE- U...	PUB
4	Row3	879926	HANDWERKS...	PUB
5	Row4	879926	TECHNISCHE...	HES
6	Row5	879926	INDUSTRIE-U...	PUB

Period 1

Rows: 55994 | Columns: 7

#	RowID	projectID	name	activityType
1	Row0	101091483	LA PALMA RES...	PRC
2	Row1	101091483	GEOLOGIAN TU...	REC
3	Row2	101091483	INTERNATIONA...	OTH
4	Row3	101091483	PROMAN MAN...	PRC
5	Row4	101091483	ALDA - ASSOCI...	OTH

Period 2

Rows: 233828 | Columns: 7

#	RowID	projectID	name	activityType
1	Row0	879926	INDUSTRIE-UND HANDELSKAMME...	PUB
2	Row1	879926	ZTS-ZENTRUM FUR TECHNOLOGIE...	OTH
3	Row2	879926	INDUSTRIE- UND HANDELSKAMME...	PUB
4	Row3	879926	HANDWERKSKAMMER DRESDEN	PUB
5	Row4	879926	TECHNISCHE UNIVERSITAET CHEM...	HES
6	Row5	879926	INDUSTRIE-UND HANDELSKAMME...	PUB
7	Row6	879926	HOCHSCHULE ZITTAU/GOERLITZ	HES
8	Row7	879926	AGENTUR FUR INNOVATIONSFORD...	PRC
9	Row8	740689	CENTRO INTERNAZIONALE IN MON...	REC
10	Row9	740689	FREDERIKSBORG BRAND OG REDNI...	PRC
11	Row10	740689	Department d'Interior - Generalitat ...	PUB

One table with shared columns and more rows



1) Summary: Why stack data?



- To combine two or more tables/ datasets with the same structure but different data, thus, add rows.
- Some tools call it UNION, others Concatenate. In the end, they all add rows, remember to be mindful of the **operation**.



Data Blending



Stack

Stack new data at the bottom of a table -> same columns, **add rows**

Join

Join information from tables by defining matching criteria -> **add new columns**

2) Why join?



- Situation: When we have information from different tables with different structures, but they share the **main category** (Projects in our case).
- Action: **add columns** from different tables/ data sets together.
- Goal: To **enrich** the information about the **main category**.



Join data to add columns



Table 1		
ID	Column 1	Column 2
Id-1		
Id-2		



Table 2		
ID	Column 3	Column 4
Id-1		
Id-2		

Result Table				
ID	Column 1	Column 2	Column 3	Column 4
Id-1				
Id-2				

2) What do we join?



Two tables: we join a "left table" and a "right table" on a Key

- Left table: first dataset in the join operation.
- Right table: second dataset in the join operation.

Key:

- Key: The **common** column(s) used to match rows between the left and right tables.



2) Joining Columns of Data

- We want to have all the information **together**
- We need a **common** column to join the data – normally “Key” or “ID”

ID	Item	Cost
Id-1	Table	50
Id-2	Couch	100

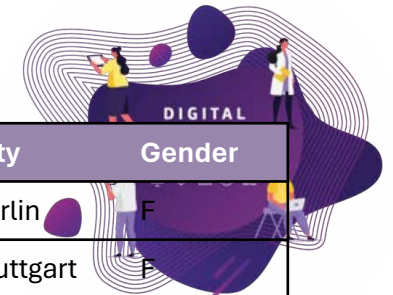
ID	Store Type	City
Id-1	Onsite	NY
Id-2	Online	NY

Diagram illustrating the joining of two tables. A bracket connects the ID columns of both tables, indicating they are the common key used for joining.

ID	Item	Cost	Store Type	City
Id-1	Table	50	Onsite	NY
Id-2	Couch	100	Online	NY



2) Types of Joins



Key	OrderDate	OrderID
22	2019-09-23	#23444
24	2019-09-30	#23457
15	2019-10-07	#28985
10	2019-10-13	#29999

Key	DoB	City	Gender
17	1974-02-23	Berlin	F
65	2001-05-25	Stuttgart	F
35	1988-08-05	Cologne	M
15	1983-07-20	Hamburg	M
10	1993-01-13	Berlin	M

Join by Key

Inner Join

Left Table

Right Table

Key	OrderDate	OrderID	DoB	City	Gender
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2019-10-13	#29999	1993-01-13	Berlin	M

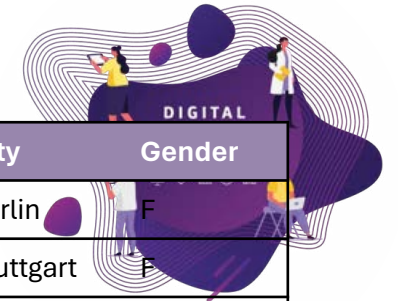
Left Outer Join

Right Outer Join

Key	OrderDate	OrderID	DoB	City	Gender
22	2019-09-23	#23444	?	?	?
24	2019-09-30	#23457	?	?	?
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2019-10-13	#29999	1993-01-13	Berlin	M

Key	OrderDate	OrderID	DoB	City	Gender
17	?	?	1974-02-23	Berlin	F
65	?	?	2001-05-25	Stuttgart	F
35	?	?	1988-08-05	Cologne	M
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2019-10-13	#29999	1993-01-13	Berlin	M

2) Joining Columns of Data



Key	OrderDate	OrderID
22	2019-09-23	#23444
24	2019-09-30	#23457
15	2019-10-07	#28985
10	2019-10-13	#29999

Left Table

Key	DoB	City	Gender
17	1974-02-23	Berlin	F
65	2001-05-25	Stuttgart	F
35	1988-08-05	Cologne	M
15	1983-07-20	Hamburg	M
10	1993-01-13	Berlin	M

Right Table

Join by Key

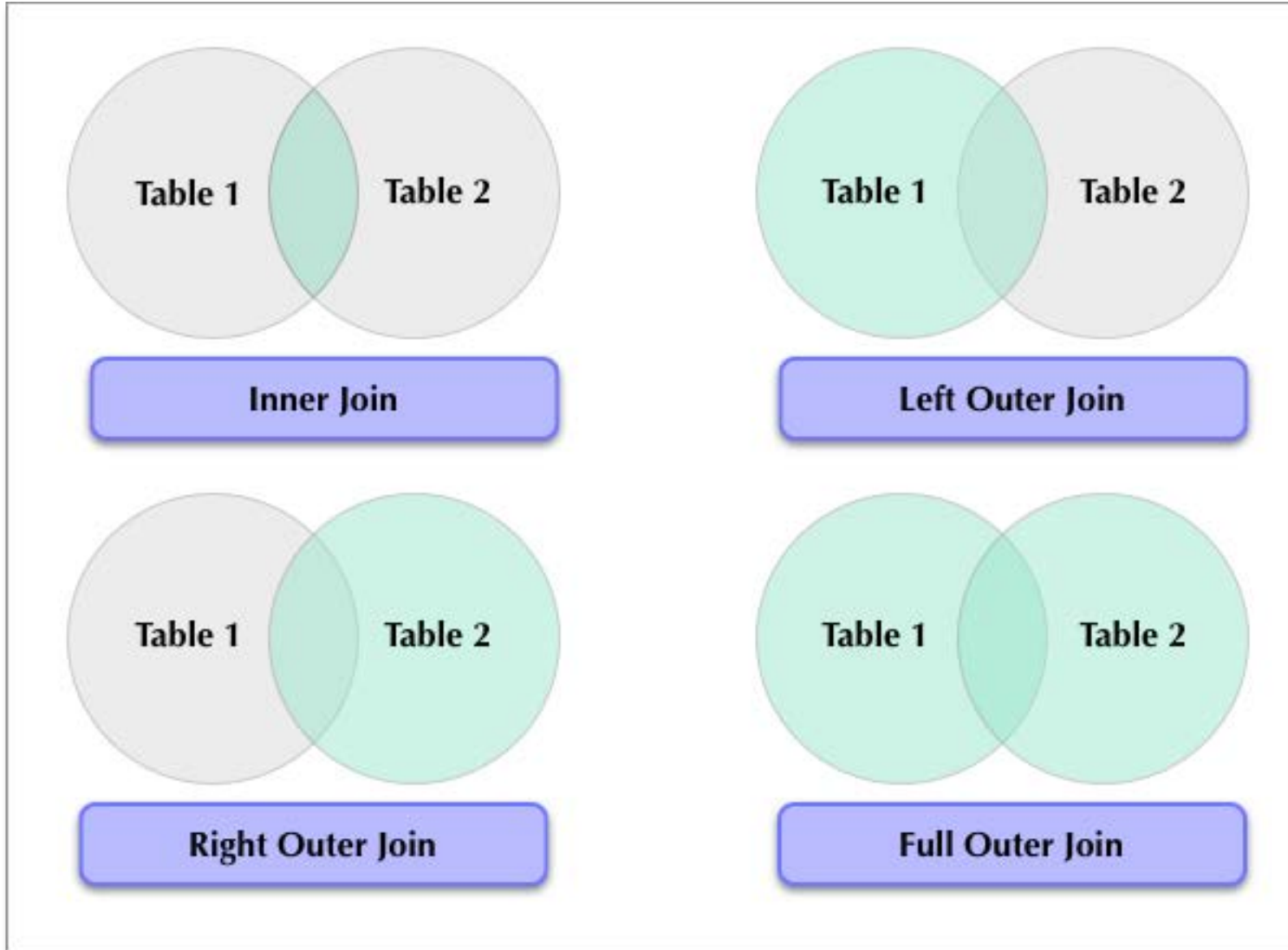
Full outer Join

Key	OrderDate	OrderID	DoB	City	Gender
17	?	?	1974-02-23	Berlin	F
65	?	?	2001-05-25	Stuttgart	F
35	?	?	1988-08-05	Cologne	M
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2019-10-13	#29999	1993-01-13	Berlin	M
22	2019-09-23	#23444	?	?	?
24	2019-09-30	#23457	?	?	?

Missing values in the left table

Missing values in the right table

What types of join can we implement?



2) How to join?

Steps:

1. Choose **matching** column (key) from left and right table
2. **Decide** which information to keep (type of join)



2) What do we join? Horizon Data



Rows: 44386 | Columns: 7

#	RowID	ids ↓ String	startDates Local Date	endDates Local Date	ecMaxContri... Number (double)	totalCosts Number (double)	objectives String	startYear String
1	Row1	879926	2020-01-01	2021-12-31	125,559	125,560	The aim of the pr...	2020
2	Row2	740689	2017-05-01	2021-01-31	7,836,370.63	8,585,414.19	The project aims ...	2017
3	Row3	730082	2016-11-01	2019-04-30	1,500,000	1,850,000	E2mC aims at de...	2016
4	Row4	101023342	2021-06-01	2024-11-30	3,755,470.75	5,214,286.25	Sustainability in c...	2021
5	Row5	831993	2019-01-01	2021-12-31	796,056.25	796,056.25	New technologie...	2019
6	Row6	641451	2015-01-01	2016-12-31	999,653.25	1,323,315	GEO-VISION is an...	2015
7	Row7	731665	2017-01-01	2018-12-31	1,498,857.5	1,498,857.5	The objective of t...	2017
8	Row8	653866	2015-06-01	2017-11-30	1,990,114.25	2,018,034.75	This proposal se...	2015
9	Row9	770037	2017-11-01	2019-10-31	2,003,918.75	2,003,918.75	Since 2015, migr...	2017
10	Row10	636906	2015-02-01	2017-01-31	610,013.75	792,938.75	The worldwide ec...	2015

Rows: 233828 | Columns: 7

#	RowID	projectID String	name String	activityType String	street String	postCode String	city String	country String
1	Row1	879926	INDUSTRIE-UND ...	PUB	STRASSE DER N...	09111	CHEMNITZ	DE
2	Row2	879926	ZTS-ZENTRUM F...	OTH	INDUSTRIESTRA...	01612	GLAUBITZ	DE
3	Row3	879926	INDUSTRIE-UND ...	PUB	LANGER WEG 4	01239	DRESDEN	DE
4	Row4	879926	HANDWERKSKA...	PUB	AM LAGERPLATZ...	01099	DRESDEN	DE
5	Row5	879926	TECHNISCHE UN...	HES	STRASSE DER N...	09111	Chemnitz	DE
6	Row6	879926	INDUSTRIE-UND ...	PUB	GOERDELERRING...	04109	LEIPZIG	DE
7	Row7	879926	HOCHSCHULE ZI...	HES	THEODOR-KOER...	02763	Zittau	DE
8	Row8	879926	AGENTUR FUR IN...	PRC	LESSINGSTRASS...	04109	LEIPZIG	DE
9	Row9	740689	CENTRO INTERN...	REC	VIA ARMANDO M...	17100	SAVONA	IT

Left Table: Projects

Right Table: Organisations

Join by ID

Rows: 233828 | Columns: 13

#	RowID	pro... String	name String	activit... String	street String	post... String	city String	coun... String	startDates Local Date	endDates Local Date	ecMaxCon... Number (double)	totalCosts Number (double)	objectives String	startYear String
1	Row1	879926	INDUSTRIE-UN...	PUB	STRASSE DER ...	09111	CHEMNITZ	DE	2020-01-01	2021-12-31	125,559	125,560	The aim of the ...	2020
2	Row2	879926	ZTS-ZENTRUM...	OTH	INDUSTRIESTR...	01612	GLAUBITZ	DE	2020-01-01	2021-12-31	125,559	125,560	The aim of the ...	2020
3	Row3	879926	INDUSTRIE- U...	PUB	LANGER WEG 4	01239	DRESDEN	DE	2020-01-01	2021-12-31	125,559	125,560	The aim of the ...	2020
4	Row4	879926	HANDWERKSK...	PUB	AM LAGERPLA...	01099	DRESDEN	DE	2020-01-01	2021-12-31	125,559	125,560	The aim of the ...	2020
5	Row5	879926	TECHNISCHE ...	HES	STRASSE DER ...	09111	Chemnitz	DE	2020-01-01	2021-12-31	125,559	125,560	The aim of the ...	2020
6	Row6	879926	INDUSTRIE-UN...	PUB	GOERDELERRI...	04109	LEIPZIG	DE	2020-01-01	2021-12-31	125,559	125,560	The aim of the ...	2020
7	Row7	879926	HOCHSCHULE ...	HES	THEODOR-KOE...	02763	Zittau	DE	2020-01-01	2021-12-31	125,559	125,560	The aim of the ...	2020
8	Row8	879926	AGENTUR FUR...	PRC	LESSINGSTRA...	04109	LEIPZIG	DE	2020-01-01	2021-12-31	125,559	125,560	The aim of the ...	2020
9	Row9	740689	CENTRO INTE...	REC	VIA ARMANDO...	17100	SAVONA	IT	2017-05-01	2021-01-31	7,836,370.63	8,585,414.19	The project ai...	2017
10	Row10	740689	FREDERIKSBO...	PRC	LOGISOSE 3	3600	FREDERIKSSU...	DK	2017-05-01	2021-01-31	7,836,370.63	8,585,414.19	The project ai...	2017

Inner Join



Join in the different BDTI tools (add columns)



SQL

```
SELECT table1.*, table2.column_name  
FROM table1  
JOIN table2 ON table1.id = table2.id;
```

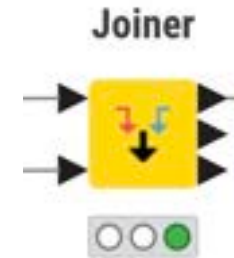
R

```
library(dplyr)  
df1 <- read.csv('table1.csv')  
df2 <- read.csv('table2.csv')  
merged_df <- inner_join(df1, df2, by = "id")
```

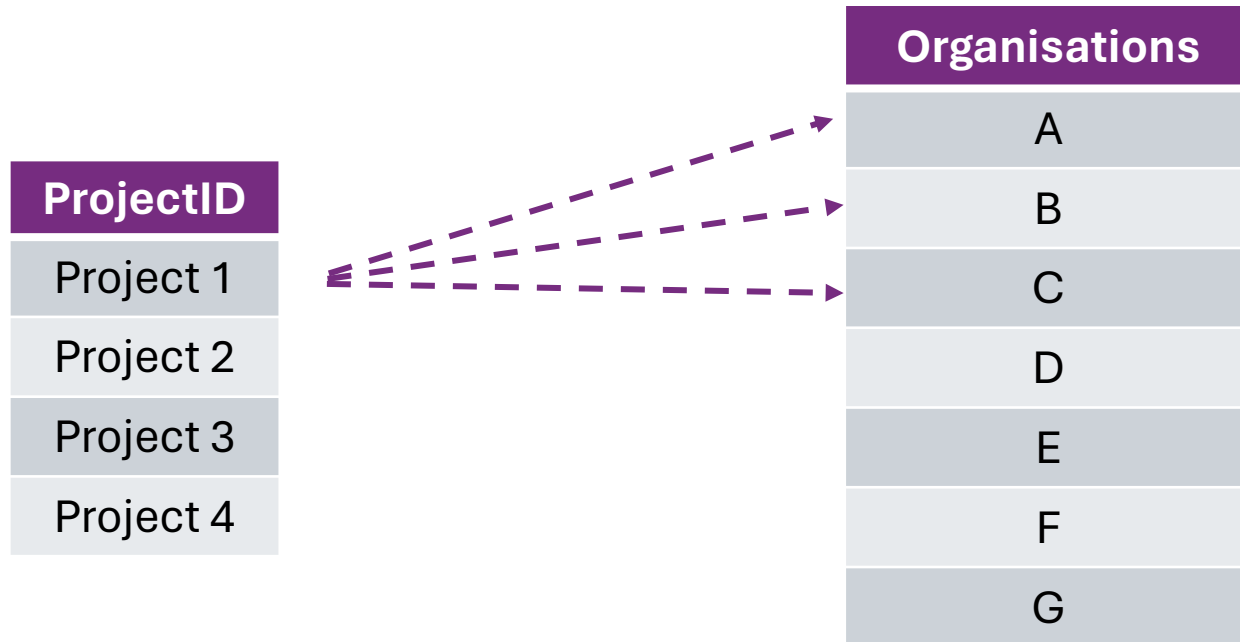
PYTHON

```
import pandas as pd  
df1 = pd.read_csv('table1.csv')  
df2 = pd.read_csv('table2.csv')  
merged_df = pd.merge(df1, df2, on='id', how='inner')
```

KNIME



2) Joining relationships

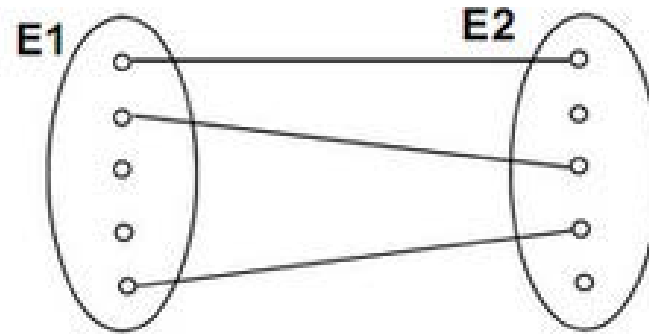


1 to many

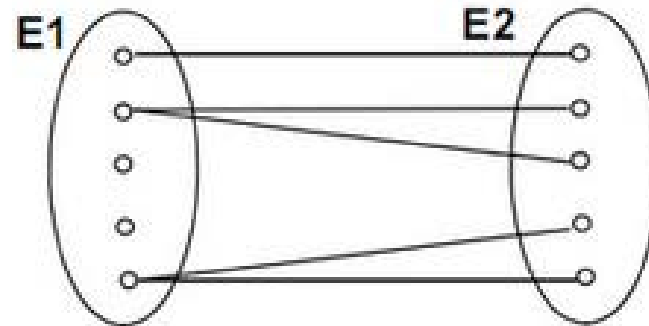


2) Joining relationships

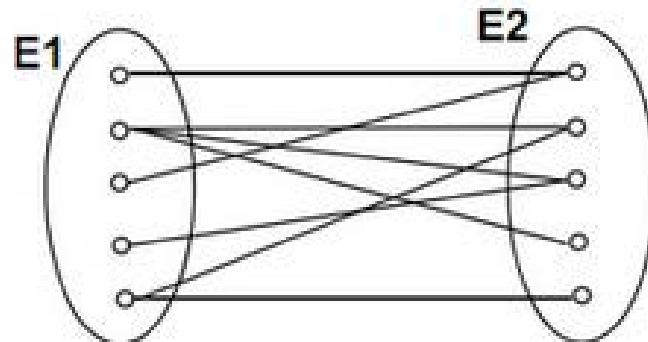
**One to one
1:1**



**One to many
1:M**



**Many to many
M:N**

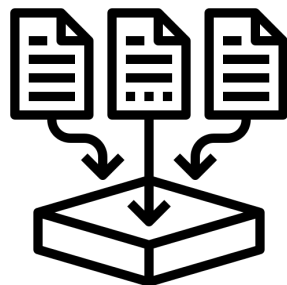


Summary: Data Blending



From multiple data files (8)

- Organisations 20-21
- Projects 20-21
- EuroSciVoc 20-21
- CO2
- EU Country Names



To a **single** data file with all the information!



Data Export and Storage

Data Export and Storage



1. Export the data to store it in different file formats: CSV, Excel, JSON.
2. **Scalability**, when things gets **bigger**, and we have too much data is better to use Databases

1) Export in different file formats

- End of ETL process, we can export the final files in the format that is more convenient for us.
- Export the data to store it or continue the computation somewhere else
- In Zoi's case, she will export all the Horizon data into a CSV file



CSV Writer



2) Export and storage in databases

- **Scalability**, when things gets bigger, and we have too much data is better to use Databases
- Facilitates seamless transfer of data across platforms and applications.
- Ensures compatibility and accessibility of data for various analytical and storage needs.



2) Data Export and Storage in BDTI



Where can we store the data in **BDTI**?

- *PostgreSQL: object-relational database*
- **PGAdmin:** is the administration and development platform for PostgreSQL



- MongoDB: relational NoSQL database (JSON objects)



- MinIO: S3-compatible object storage



2) Data Export and Storage in BDTI



In the BDTI framework, we can connect the tools!

- KNIME for Analytics process



- PgAdmin-SQL to Store and manage the data



Working with Databases in BDTI

How to initialise a SQL database in BDTI?

- Step 1 > Access the portal via the website
- Step 2 > Navigate to the "Service Catalog" and launch
 - a) PostgreSQL instance
 - b) PgAdmin instance
- Step 3 > Go to "My Services" and open **PgAdmin**
 - a) Input your credentials (password in "My Data")

A screenshot of the PgAdmin login interface. The form is set against a dark blue background. At the top, it says "PgAdmin" and "Login". Below this are three input fields: "Email Address / Username" with a user icon, "Password", and a language dropdown menu currently set to "English". A "Forgotten your password?" link is located to the right of the password field. At the bottom of the form is a blue "Login" button.

Configure your database



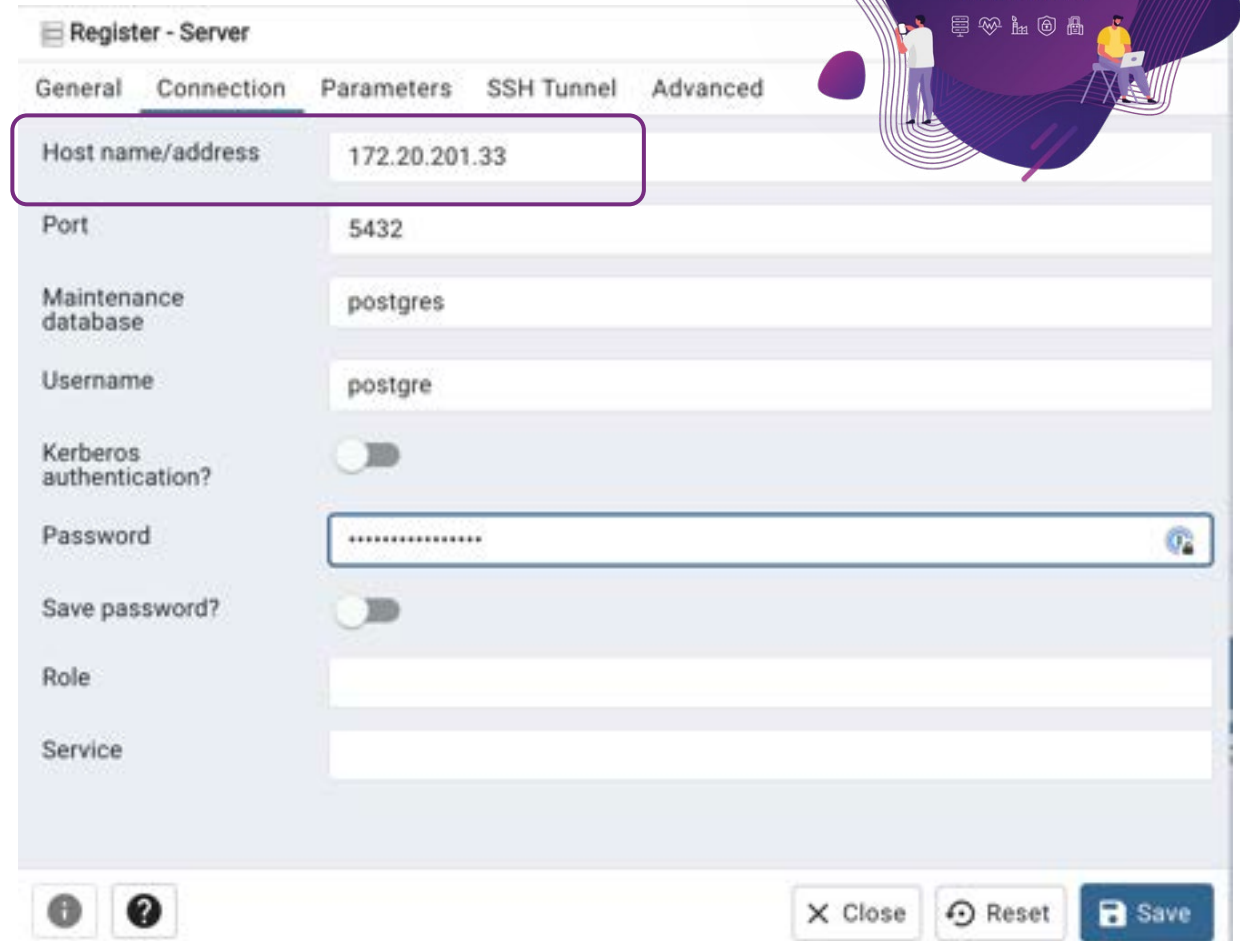
1) Click on
“Add New Server”



Configure your database

2) A configuration window will open, go to “Connection”:

- Host name/address → this information is in *My Services* under your “**PostgreSQL**” instance. Copy the host address by clicking on the related *Copy* button.
- This information is the same to use in “General” under *Name*
- Password: the postgresSQL password saved in *My Services*

A screenshot of a web-based configuration window titled 'Register - Server'. The 'Connection' tab is selected. The form contains the following fields: 'Host name/address' (172.20.201.33), 'Port' (5432), 'Maintenance database' (postgres), 'Username' (postgre), 'Kerberos authentication?' (disabled), 'Password' (masked with dots), 'Save password?' (disabled), 'Role' (empty), and 'Service' (empty). At the bottom, there are buttons for 'Close', 'Reset', and 'Save'.

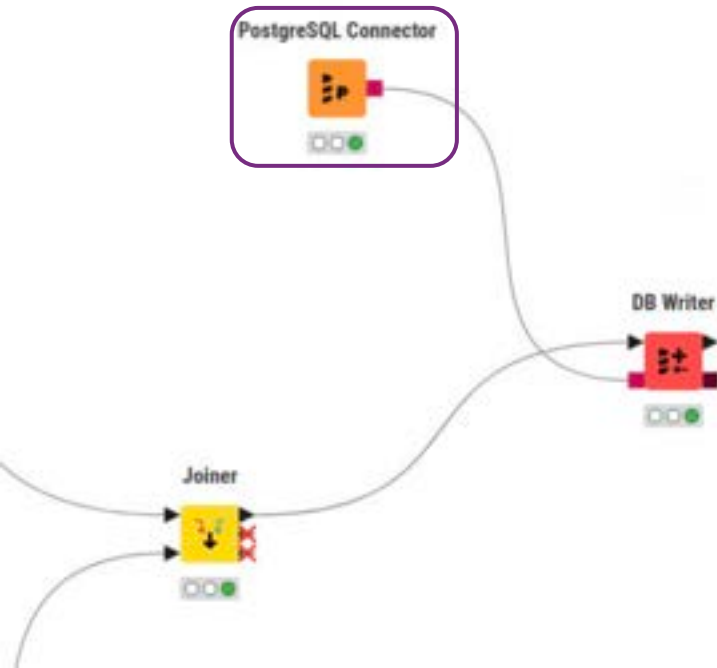
Field	Value
Host name/address	172.20.201.33
Port	5432
Maintenance database	postgres
Username	postgre
Kerberos authentication?	<input type="checkbox"/>
Password
Save password?	<input type="checkbox"/>
Role	
Service	



Connect the tools

In your KNIME instance, add the following nodes after the last join operation:

- **“PostgreSQL Connector”**
 - Add your database credentials



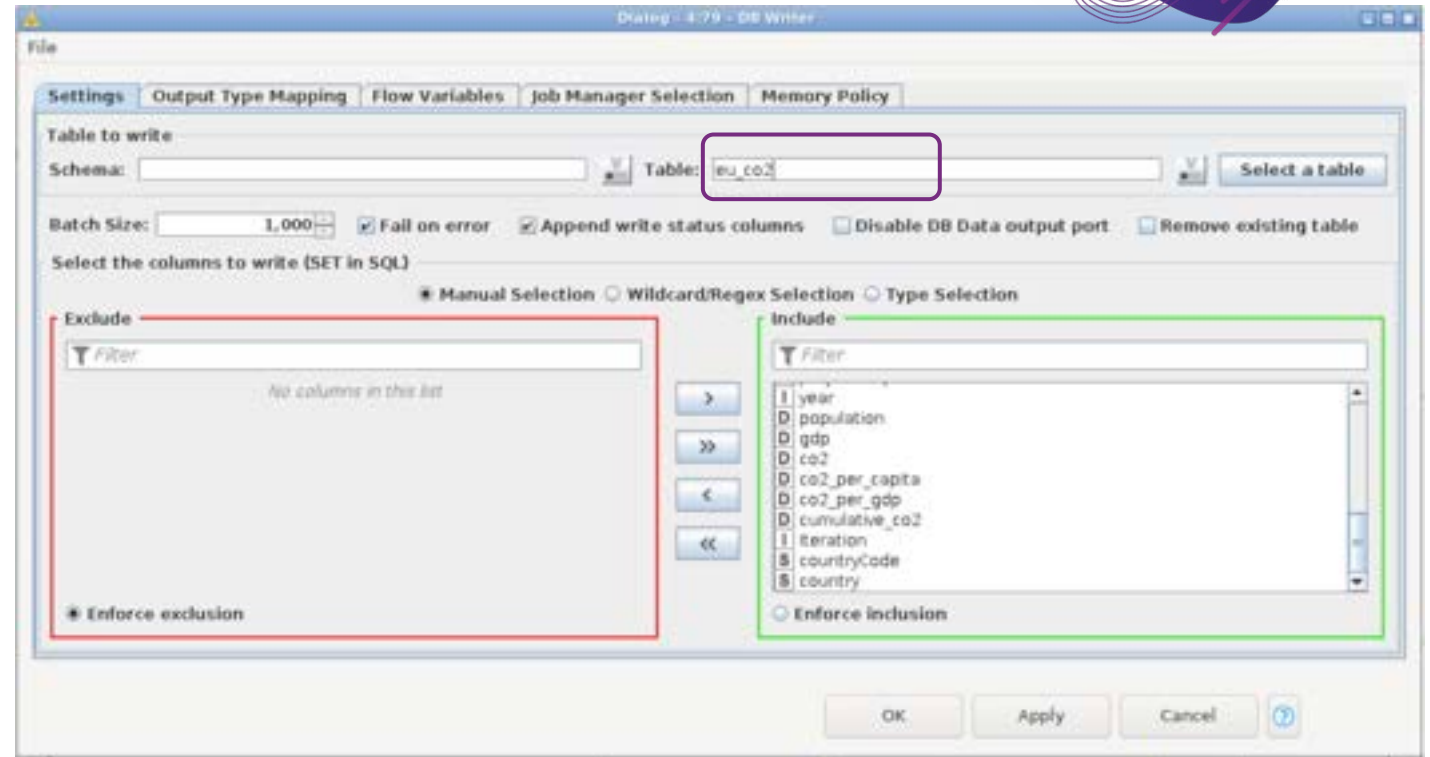
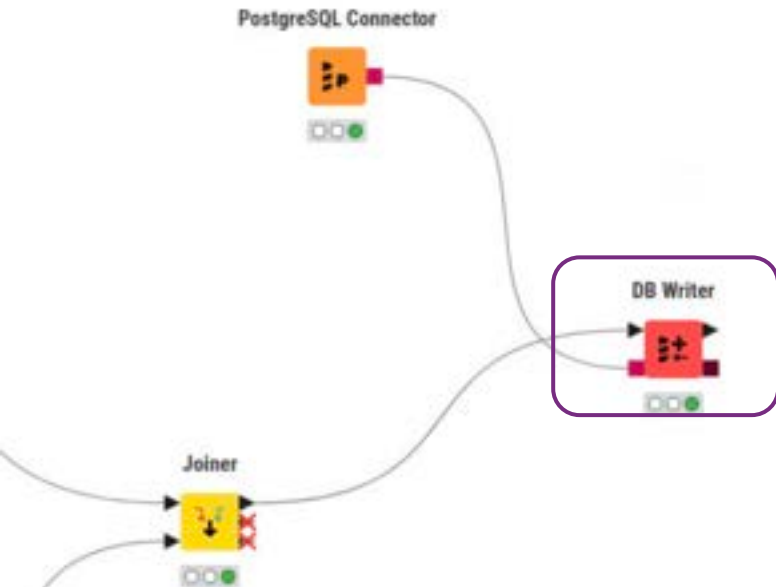
The screenshot shows the 'PostgreSQL Connector' configuration dialog box. The 'Connection Settings' tab is active. The 'Configuration' section shows 'Database Dialect' set to 'PostgreSQL' and 'Driver Name' set to 'Driver for PostgreSQL v. 42.3.5 [ID: built-in-postgres-42.3.5]'. The 'Location' section shows 'Hostname' set to '172.20.201.33', 'Port' set to '5,432', and 'Database name' set to 'postgres'. The 'Authentication' section shows 'Username & password' selected, with 'Username' set to 'postgres' and 'Password' masked with dots. The 'Kerberos' option is unselected. The dialog box has 'OK', 'Apply', and 'Cancel' buttons at the bottom.



Connect the tools

In your KNIME instance add the following nodes after the last join operation:

- **“DB Writer”**
 - Give your data a name
 - Choose columns to include



Connect the tools

Now, you have created a table in your database with all the data you have manipulated!

The screenshot shows the PgAdmin interface. The left sidebar shows the database structure, with 'public.eu_co2' selected under 'Tables (1)'. The main window shows a SQL query: `SELECT * FROM public.eu_co2`. The results are displayed in a table with the following columns: projectID, organisationID, name, activityType, street, postCode, and city. The data is as follows:

projectID	organisationID	name	activityType	street	postCode	city
673504	934522736	COIN CONSULTING GMBH	PRC	FRONDENBERGER STRASSE 75	58706	MENDEN
832274	965252424	CORSO MAGENTA	PRC	22/24 RUE LAVOISIER	92000	NANTERRE
644859	967294461	HOLONIX SRL	PRC	CORSO ITALIA 8	20821	Meda (MI)
644859	942487891	MANDELLI LAURA	PRC	VIA BIANCHI EDUARDO 30	20090	SERGATE MI
101035820	999421653	TALLINN UNIVERSITY	HES	Narva Road 25	10120	Tallin
644859	999977463	TECHNISCHE UNIVERSITÄT MÜNCHEN	HES	Ancistrasse 21	80333	Muenchen
644859	999976202	UNIVERSITAT POLITÈCNICA DE CATALUNYA	HES	CALLE JORDI GIRONA 31	08034	Barcelona
644859	951359123	IDEAL S R L C R	PRC	Via Gustava di Valdengo 1	13900	Biella
644859	952402746	NOT JUST A LABEL LIMITED	PRC	94 LEONARD STREET	EC2A 4AD	LONDON
832042	999610609	FERNUNIVERSITÄT IN HAGEN	HES	UNIVERSITÄTSSTRASSE 47	58097	Hagen
644859	956450450	SPARSITY SL	PRC	CARRER COMTE GUELL 40	08028	BARCELONA
644859	940565836	WEBLOGS SL	PRC	PLAZA DE ALONSO MARTINEZ 7 PLANTA 1 IZD PUERTA DC	28004	Madrid
650176	952464438	VASCULAR FLOW TECHNOLOGIES LIMITED	PRC	UNIT 1 PROSPECT BUSINESS CENTRE	DD2 1TY	DUNDEE
827113	946428225	NUTRIPHARMA HUNGARIA TAPES GYOGYSZER ELOALL	PRC	TARGATO LEJTO 8 II EM 5	1021	BUDAPEST
676786	999848938	PHILIPPS UNIVERSITÄT MARBURG	HES	BIEGENSTRASSE 10	35037	Marburg
739677	920095961	CERCARE MEDICAL APS	PRC	INSE LEDMANNS GADE 10	8000	Aarhus C

Apply queries to validate your data

A screenshot of a PostgreSQL query editor interface. The query editor shows a SQL query with line numbers 1 through 12. The query is as follows:

```
1 /*
2 SELECT * FROM public.eu_co2;
3 */
4
5 /*
6 SELECT COUNT(DISTINCT country) AS unique_country_count
7 FROM public.eu_co2;
8 */
9
10 SELECT country, COUNT(*) AS country_count
11 FROM public.eu_co2
12 GROUP BY country;
```

The 'Data Output' tab is active, showing a table with two columns: 'country' (character varying) and 'country_count' (bigint). The table contains 15 rows of data for various European countries. A purple callout box points to the 'Data Output' tab with the text: 'Here we have the count of countries in our data'.

	country	country_count
1	Austria	6012
2	Belgium	10330
3	Bulgaria	968
4	Croatia	857
5	Cyprus	1193
6	Czech Republic	2283
7	Denmark	4987
8	Estonia	1011
9	Finland	4296
10	France	20542
11	Germany	24994
12	Greece	7110
13	Hungary	1657
14	Ireland	3568
15	Italy	20483
16	Latvia	586
17	Lithuania	702

Summary



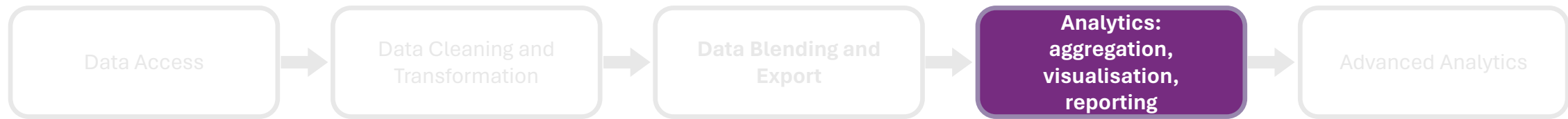
Now you should be able to:



1. Sign in to the BDTI framework and initialise KNIME, PgAdmin, and PostgreSQL
2. Stack multiple tables
3. Join tables in multiple ways
4. Write data to various file formats
5. Store the data in a database



Next session:



Practice, practice and practice!



- Slides, hands-on exercises, and solutions will be provided on the webinar Gitlab repository

<https://code.europa.eu/bdti/bdti-essentials-course>

A screenshot of the GitLab repository page for 'BDTI Essentials Course'. The page shows the repository name, project ID (035), and a 'Request Access' button. It indicates 12 commits, 1 branch, 0 tags, and 2 MB of project storage. A recent commit by Maria Claudia BODINO is highlighted, titled 'Upload banner session 1'. Below this, a table lists the repository's files and their commit history. The 'README.md' file is expanded, showing copyright information for the European Union and a license notice. At the bottom of the README, there is a banner image with the text 'From hype to action:' and the Digital Europe Programme logo.

Name	Last commit	Last update
Session 1: Data Access and Exploration	Upload banner session 1	4 days ago
BDTI_Banner_generic.png	Uploaded New File	5 days ago
LICENSE	Add LICENSE	5 days ago
README.md	Update README.md	5 days ago



Practice, practice and practice!



- Slides, hands-on exercises, and solutions will be provided on the webinar Gitlab repository

<https://code.europa.eu/bdti/bdti-essentials-course>

Session 3 - Exercise		
Data Blending Now that Zoi has the columns she needs, he wants to blend both data sources into one single table and organize the columns in a way that is easier for her to work with it.		Data Storage-Export: Finally, Zoi will export the data to a database and other file formats.
Concatenate datasets 1. Concatenate the different versions of Horizon 2020-2021 with the Concatenate node by intersection of columns	Step 5. Merge all datasets into the final one 1. Join Organization and Project on the ProjectID and its columns with the Joiner node. 2. Join the resulting table with euroSciVoc on ProjectID columns with the Joiner node 3. Use the Joiner node to join the Horizon joined table with the CO2 EU with the countryCode and country columns	Store the data in a DB 1. Export the preprocessed data into a DB 2. SQL file for exercise



Q&A

Your BDTI journey starts here



Congrats! You are on the first step to data-driven innovation.



•
**BDTI
Essentials
Course**



•
**Brainstorm your
data project**



•
**Apply for BDTI Pilot-
light process**

https://big-data-test-infrastructure.ec.europa.eu/apply-bdti_en



Course discussion board



joinup Interoperable Europe Interoperability Solutions Support Centre

Big Data Test Infrastructure (BDTI)

Leave this solution

eGovernment Topics: Open Source Software

About Members Discussions BDTI Newsletter Overview

BDTI Essentials Course Discussion Board

Kim Gillick Published on: 24/01/2024 Last update: 26/01/2024 Discussion

Unlike (3) Translate

Welcome!

Here, you can ask questions and discuss topics related to the [BDTI Essentials Online Course](#). The board is moderated by the BDTI team, so your questions will be answered quickly. We also encourage members to discuss and help each other where possible to build a collaborative space and rewarding community.

Helpful links



Get in touch and follow the BDTI activities



Are you working for a public administration in need of infrastructure for data analytics?



EC-BDTI-PILOTS@ec.europa.eu



**Visit
BDTI's
website**



**Subscribe to
BDTI's
newsletter**



**Subscribe to
BDTI's
Joinup**



**Register for BDTI
Essentials online
course**

<https://big-data-test-infrastructure.ec.europa.eu/>

