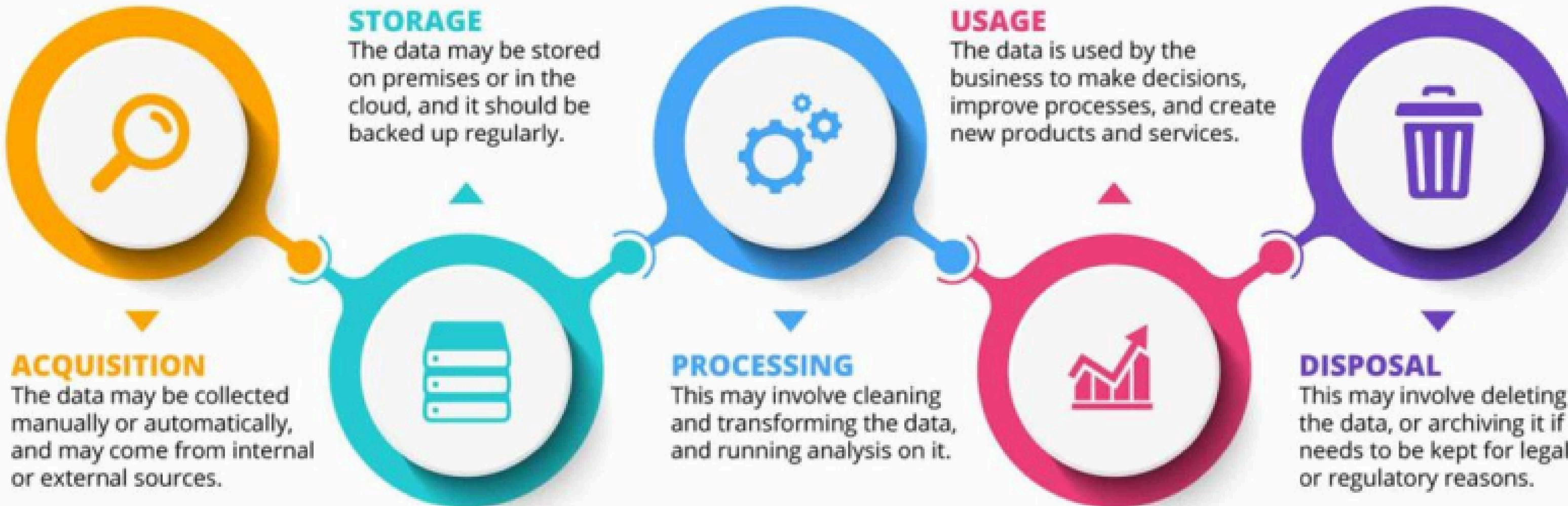


Modern Data Platforms

It is not you, it is me....

Group 1: Carolina Aparício, Trisha Chatterjee,
Luna Geens, Yash Ghai & Marthe Spriet

1 Data Lifecycle Management



1 Data Lifecycle Management

1.1 Acquisition

Project Scope



How Couples Meet and Stay Together (HCMST) is a study of how Americans meet their spouses and romantic partners.

- A first study spans from 2009 to 2015
- A second study spans from 2017 to 2022: Our focus.

Study follows respondents with a partner along waves.

You can only participate at wave x if you were still together with your partner at wave x-1.

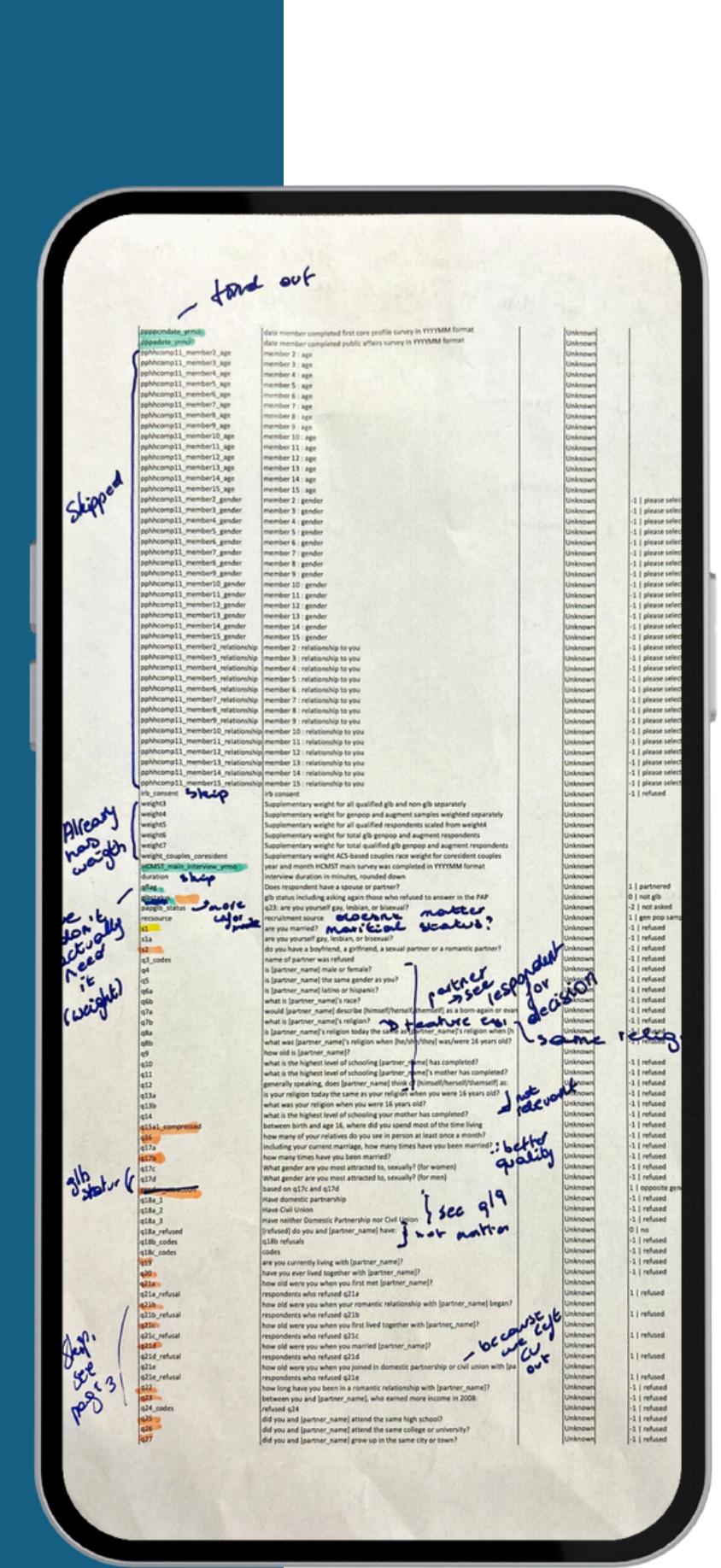
TOO big data?

Problem 1: We were so overwhelmed with the amount of data in these studies.

With “common sense”: we went through the dataset and ended up with 145 of 388 variables.

Next steps:

- Clean the dataset
 - With regression analysis: pick top 20 columns, on coefficient. rerun analysis with only these columns.
 - With predictive analysis: predict the likelihood of couple staying together



1 Data Lifecycle Management

1.2 Storage

Too big data?

Problem 2: We were so overwhelmed with the amount of services and complexity of Azure.

- We struggled with multiple Azure services due to a steep learning curve.
- Collaboration was challenging—permissions, service integrations, and differences between services were confusing.
- Just as we were close to finishing, our subscriptions expired.

ELT in Azure: From storage to processing



Extract

Online



Stored as .xlsx in Azure blob storage



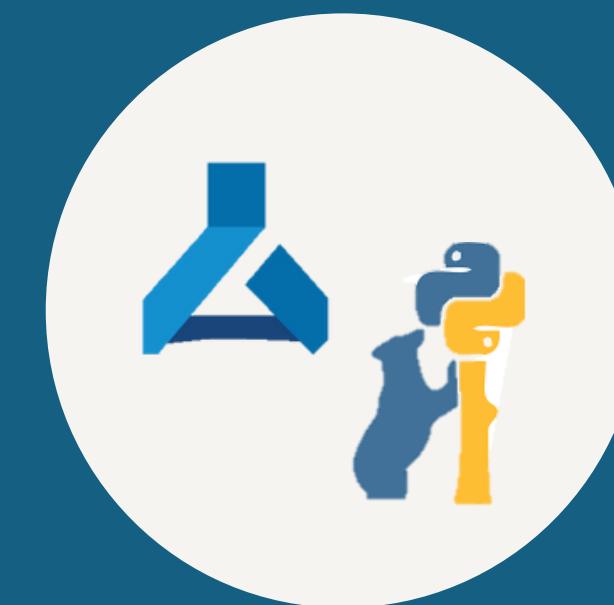
Load

Azure blob storage



Azure data factory to convert to SQL table

Azure SQL servers to store our database



Transform

Azure SQL server connection in python



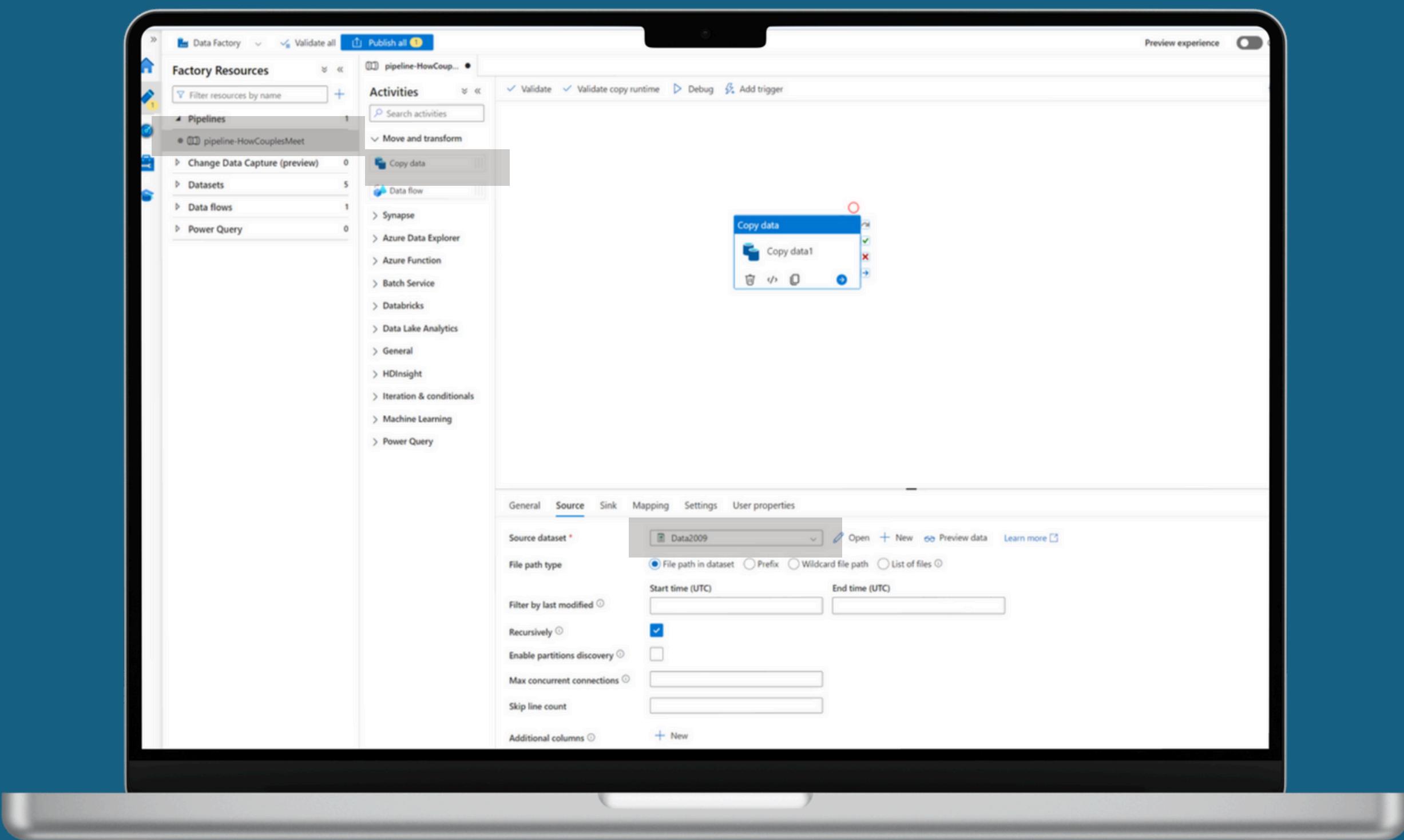
To load database in Azure ML (in a python notebook) as pandas dataframe

Extract: Azure Blob Storage

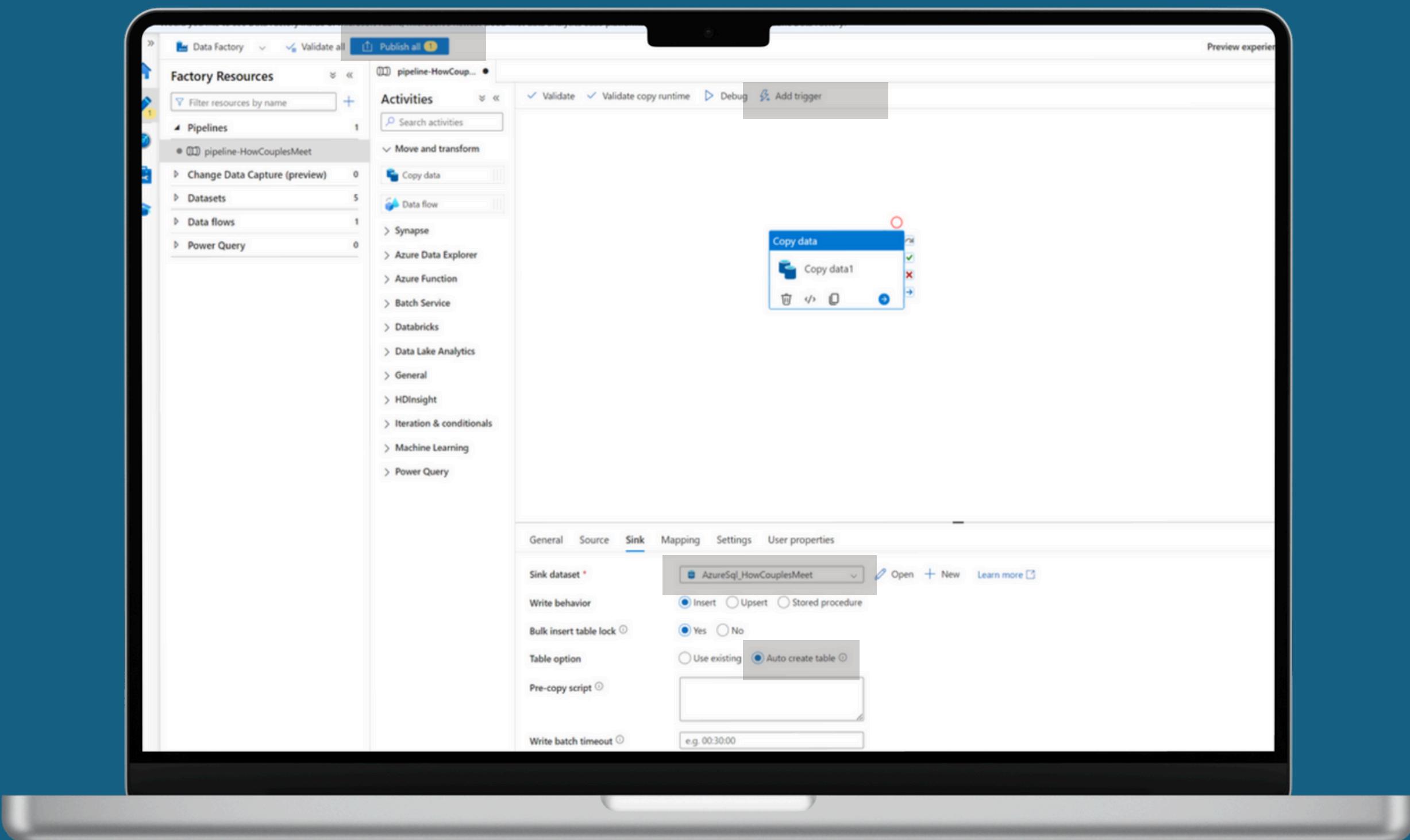
The screenshot shows the Azure Blob Storage interface for the container 'howcouplesmeet2009'. The left sidebar includes links for Overview, Diagnose and solve problems, Access Control (IAM), Settings (Shared access tokens, Access policy, Properties, Metadata), and a search bar. The main area displays a table of blobs with columns for Name, Modified, Access tier, and Archive status. The table lists numerous files, mostly CSV and PDF documents, with their last modified dates ranging from October 17, 2024, to October 24, 2024. Most files are in the 'Hot (Inferred)' access tier.

Name	Modified	Access tier	Archive status
SUCCESS	10/24/2024, 10:36:35 PM	Hot (Inferred)	
HCMST_2009_CleanedData.csv	10/24/2024, 9:56:23 PM	Hot (Inferred)	
HCMST_2009_CleanedData.xlsx	10/24/2024, 5:27:31 PM	Hot (Inferred)	
HCMST_2009_Codebook.pdf	10/17/2024, 9:27:24 PM	Hot (Inferred)	
HCMST_2009_CorrectData.csv	10/24/2024, 10:32:15 PM	Hot (Inferred)	
HCMST_2009_Data_ver_3.04.sav	10/23/2024, 11:57:49 PM	Hot (Inferred)	
HCMST_2009_Data_ver_3.04.xlsx	10/24/2024, 9:27:28 AM	Hot (Inferred)	
HCMST_2009_Data.xlsx	10/17/2024, 9:27:24 PM	Hot (Inferred)	
HCMST_2009_Questionnaire.pdf	10/23/2024, 11:57:48 PM	Hot (Inferred)	
HCMST_2017_Codebook_Covid.pdf	10/24/2024, 12:39:45 AM	Hot (Inferred)	
HCMST_2017_Codebook_Howmet.pdf	10/24/2024, 12:39:45 AM	Hot (Inferred)	
HCMST_2017_Codebook.pdf	10/24/2024, 12:39:46 AM	Hot (Inferred)	
HCMST_2017_UserGuide.pdf	10/24/2024, 12:39:45 AM	Hot (Inferred)	
part-00000-011d6787-fa06-4e30-bc66-77eef1934d8d-c000.csv	10/24/2024, 5:57:08 PM	Hot (Inferred)	
part-00000-0516d755-dcdd-4949-859a-11ab82825223-c000.csv	10/18/2024, 2:09:09 PM	Hot (Inferred)	
part-00000-175a64d6-7ad5-4e8e-af00-801e6a6f9ee8-c000.csv	10/24/2024, 10:02:31 PM	Hot (Inferred)	
part-00000-1cbf9501-0856-43cc-9aa5-5463400b46a4-c000.csv	10/24/2024, 10:10:00 PM	Hot (Inferred)	
part-00000-54babcc7-50f5-4701-b9c4-b3be3135a270-c000.csv	10/24/2024, 9:45:01 PM	Hot (Inferred)	
part-00000-75e9db15-e1b7-4295-a1f0-9c698970e97d-c000.csv	10/24/2024, 9:45:52 PM	Hot (Inferred)	
part-00000-db768cbb-1fa5-4f66-bf38-6d078e504674-c000.csv	10/24/2024, 10:36:35 PM	Hot (Inferred)	
part-00000-dff899127-b2ae-4ee6-a7e9-0ccfa50822b6-c000.csv	10/24/2024, 10:08:34 PM	Hot (Inferred)	
part-00000-e767684d-ea27-448b-a871-1068b98d56ad-c000.csv	10/24/2024, 5:46:16 PM	Hot (Inferred)	
part-00001-011d6787-fa06-4e30-bc66-77eef1934d8d-c000.csv	10/24/2024, 5:57:08 PM	Hot (Inferred)	
part-00001-175a64d6-7ad5-4e8e-af00-801e6a6f9ee8-c000.csv	10/24/2024, 10:02:31 PM	Hot (Inferred)	

Load: Azure Data Factory



Load: Azure Data Factory



Load: Azure SQL Databases

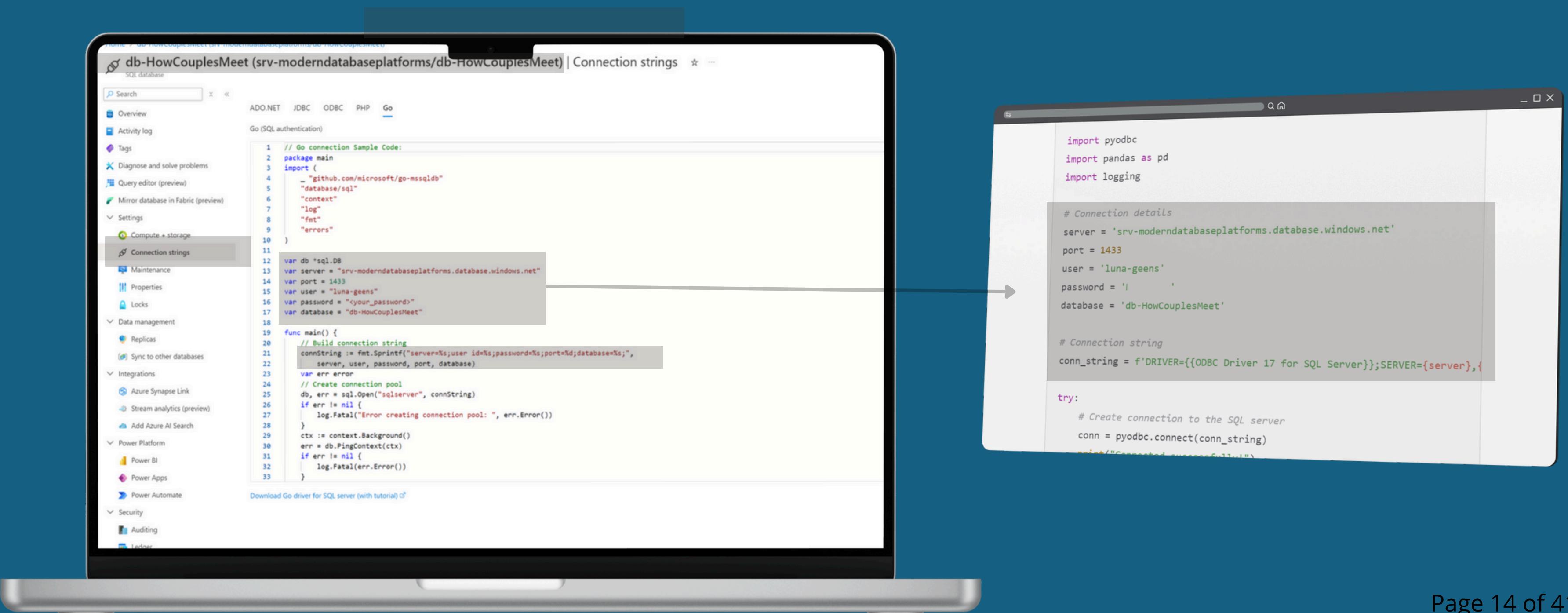
The screenshot shows the Azure portal interface for managing a database named 'db-HowCouplesMeet'. The left sidebar contains various management options like Overview, Activity log, Tags, Diagnose and solve problems, Query editor (preview), Mirror database in Fabric (preview), Settings, Data management, Integrations, Power Platform, Security, and more. The 'Query editor (preview)' section is active, displaying a query window with two tabs: 'Query 1' and 'Query 2'. The 'Query 1' tab contains the following SQL code:

```
1 SELECT TOP (1000) * FROM [dbo].[Data2009]
```

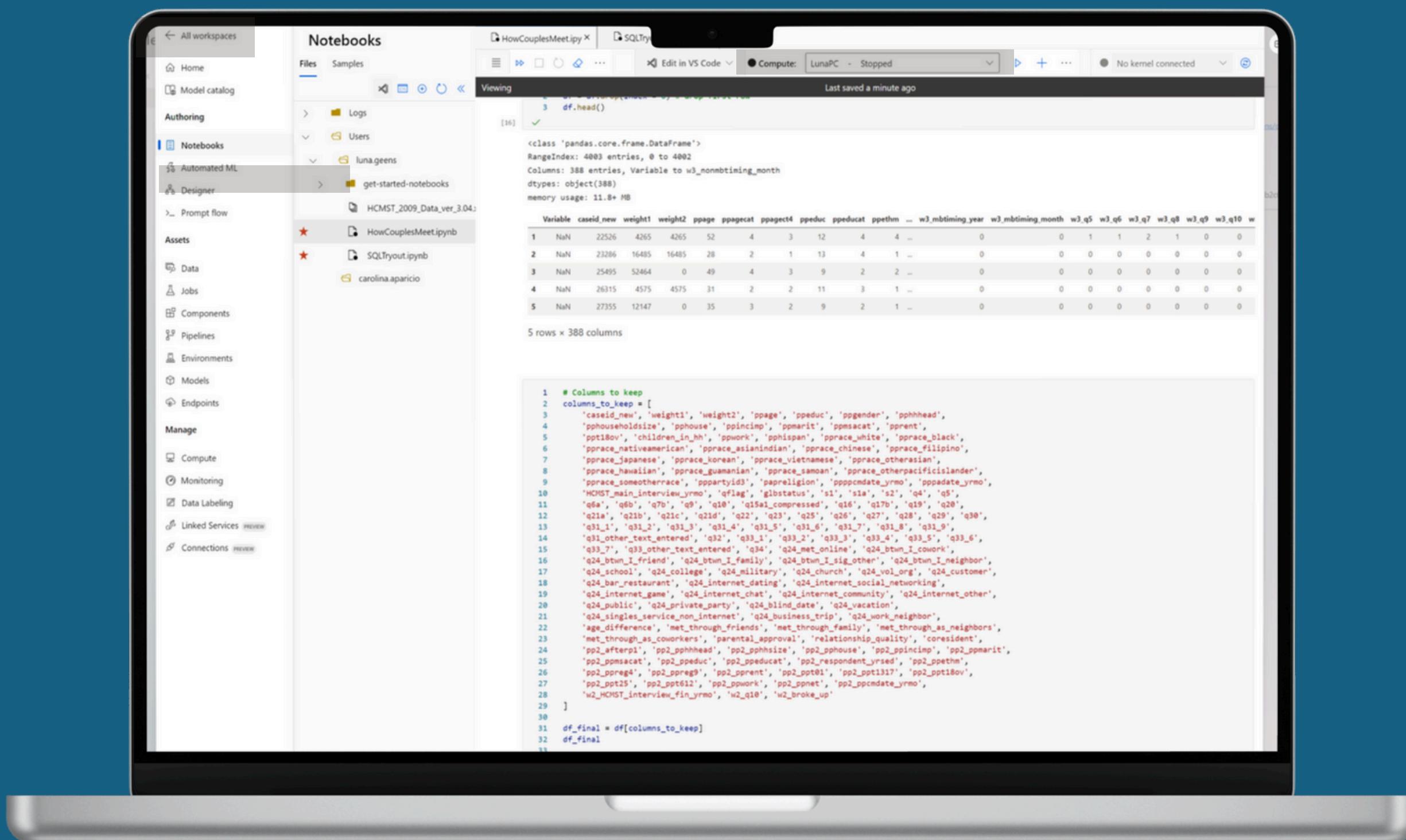
The results pane below shows a table with the following data:

caseid_new	weight1	weight2	ppage	ppduc	ppgender
22526	4265	4265	52	12	2
23286	16485	16485	28	13	2
25495	52464	0	49	9	2
26315	4575	4575	31	11	1
27355	12147	0	35	9	1
27695	1799	0	69	10	1
28536	1924	1924	53	12	1
29584	3173	3173	58	13	1
30393	68772	0	39	9	1
31456	1021	1021	45	10	1

Transform: Azure Machine Learning



Transform: Azure Machine Learning



1 Data Lifecycle Management

1.3 Processing

Some influential variables...

What is your marital status?

Category	Count	Percentage
3 Separated	57	1.5
1 Widowed	201	5.3
2 Divorced	461	12.1
5 Living with Partner	477	12.5
4 Never Married	988	25.9
0 Married	1627	42.7

How many times have you been married?

Category	Count	Percentage
0 Four or more times	5	0.5
3 Three times	14	1.4
4 Twice	83	8.2
2 Once	227	22.5
1 Never married	682	67.5

Have you ever lived together with a partner?

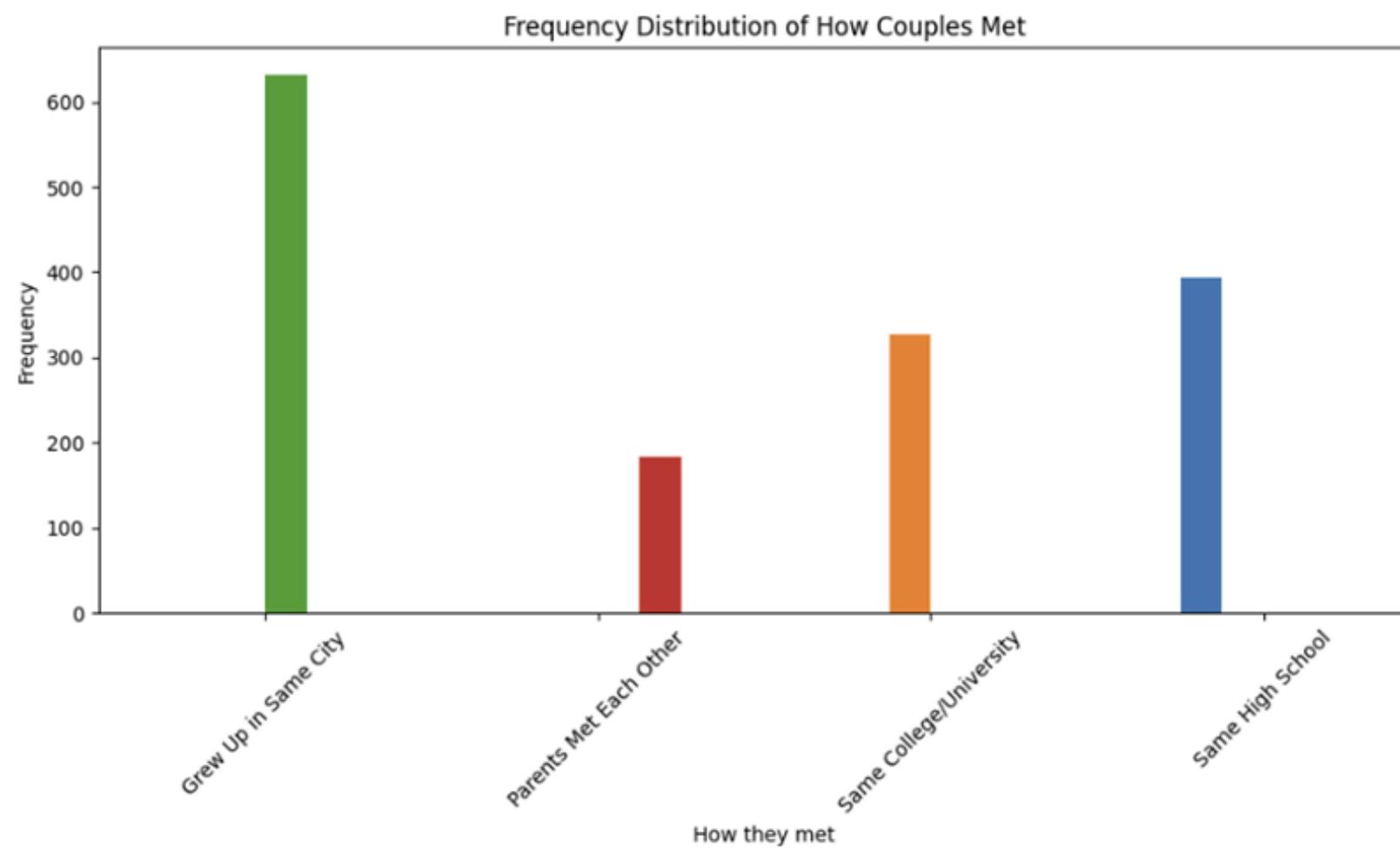
Category	Count	Percentage
1 Yes	161	26.7
0 No	442	73.3

Do you have the same gender as your partner?

Category	Count	Percentage
0 No, we are an opposite-sex couple	214	31.6
1 Yes, we are a same-sex couple	464	68.4

Some influential variables...

How did you meet?



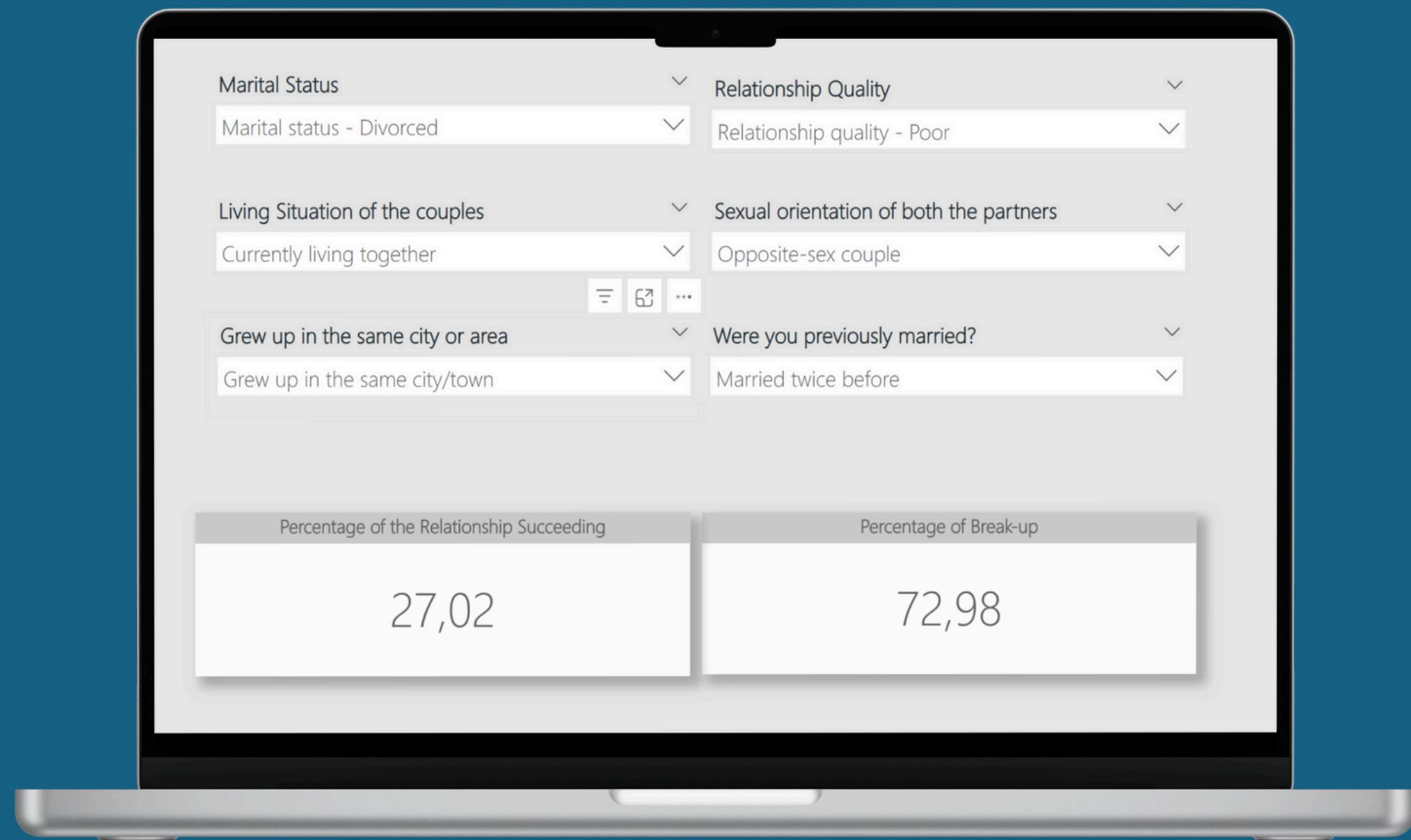
How would u rate the quality of your relationship?

Category	Count	Percentage
4 Very Poor	19	0.7
3 Poor	37	1.3
1 Fair	232	8.3
2 Good	851	30.3
0 Excellent	1666	59.4

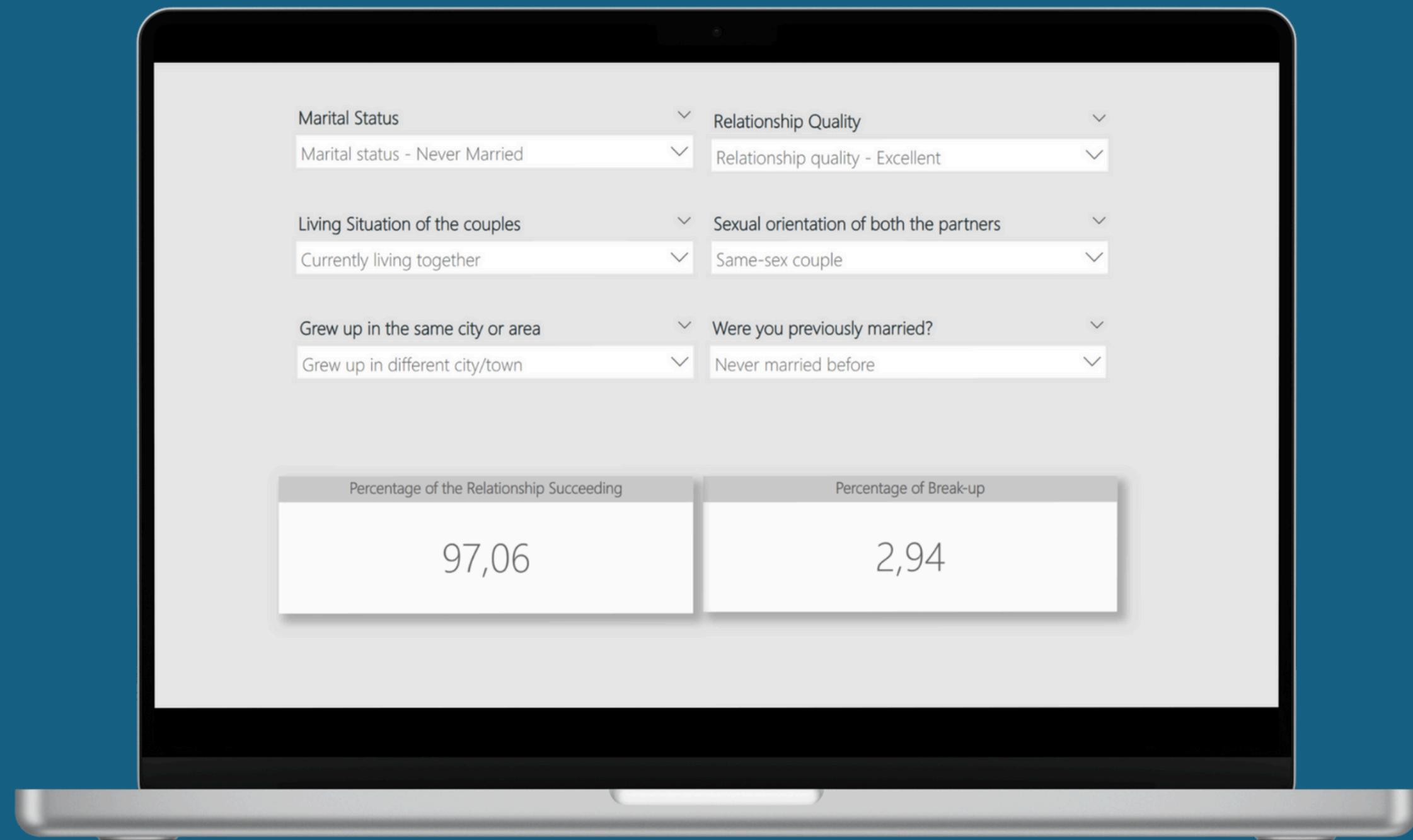
1 Data Lifecycle Management

1.4 Usage

Power BI



Power BI



1 Data Lifecycle Management

1.5 Disposal

Problem 5: Time

By the time we were at this phase....

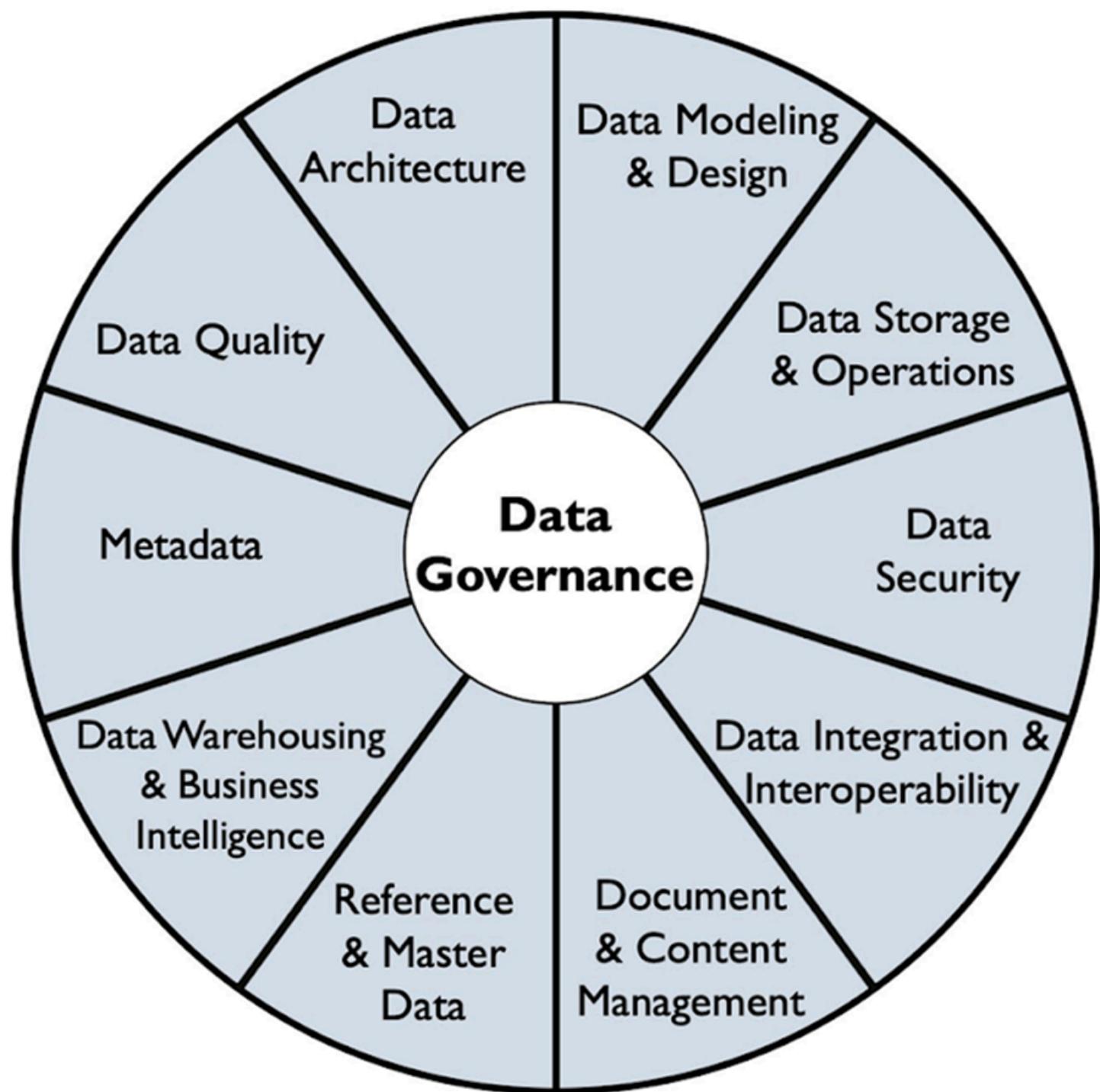
All jokes aside,

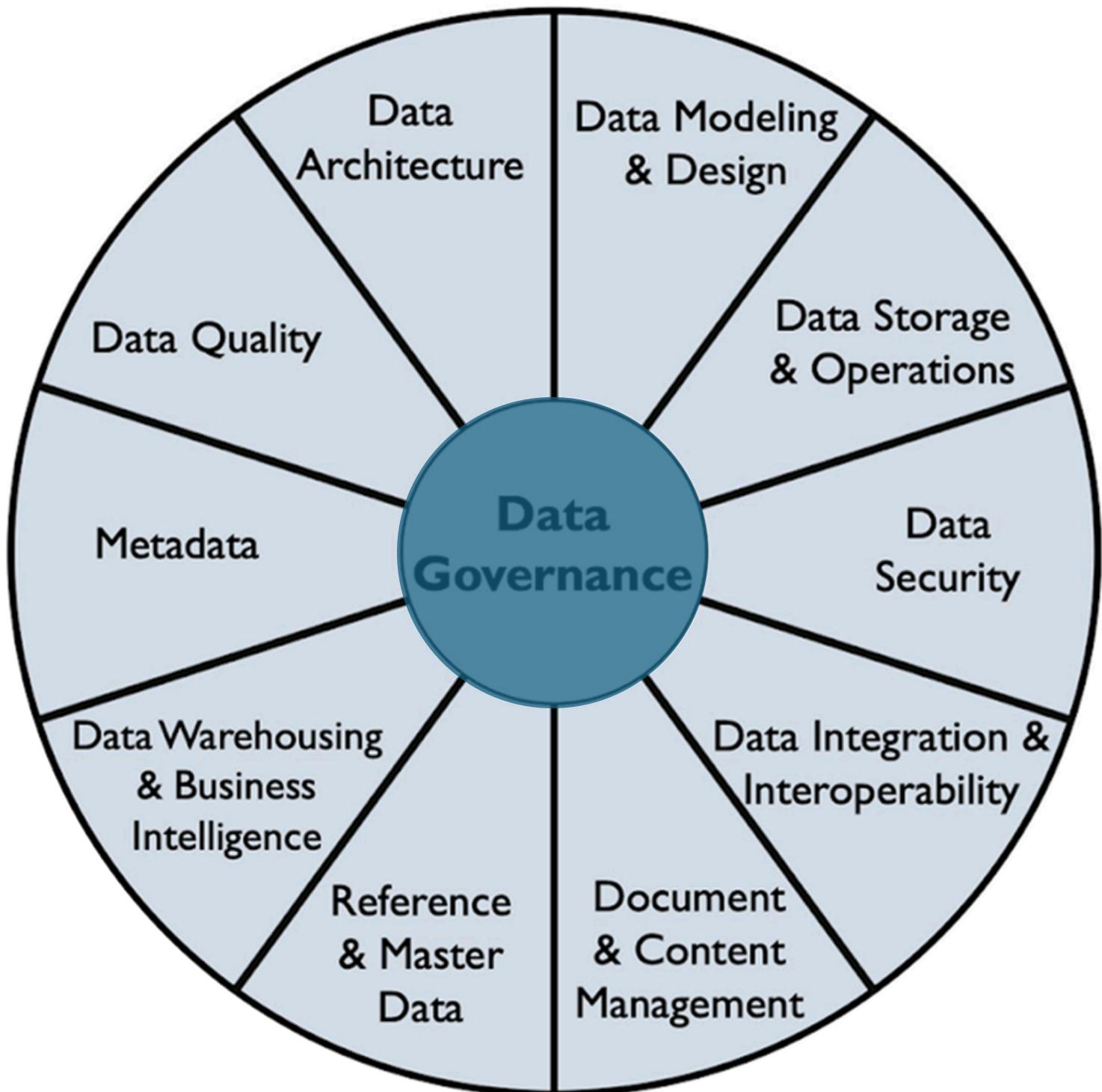
- GDPR: No identifiers, but using anonymization techniques could be beneficial.
- Security: Not needed with a public dataset.



2 Data Management Knowledge Areas

DAMA DMBoK







Data Governance

Data governance defines the policies, processes, roles, and responsibilities required to ensure that data is managed as a valuable resource.

Using Microsoft Purview

- Overall Data Governance
 - Consent Form
 - Policies

STANFORD UNIVERSITY
Stanford, California 94305 - 5401
(650) 723-2480
(650) 725-8013

Certification of Human Subjects Approvals

Penelope D Eckert, Ph.D.
CHAIR, PANEL ON MEDICAL HUMAN SUBJECTS

Date: December 17, 2010
To: Michael J Rosenfeld, Sociology
From: Christina A Stimmel
Protocol: Penelope D Eckert, Ph.D., Administrative Panel on Human Subjects in Medical Research
Protocol ID: 8303

How Couples Meet: Interviews and Surveys

IRB Number: 349 (Panel: 2)

The IRB approved human subjects involvement in your research project on 12/17/2010. Prior to sub
recruitment and enrollment, if this is: a Cancer-related study, you must obtain Cancer Center Sci
Review Committee (SRC) approval; a GCRC study, you must obtain GCRC approval; a VA study, you must ob
VA R and D Committee approval; and if a contract is involved, it must be signed.

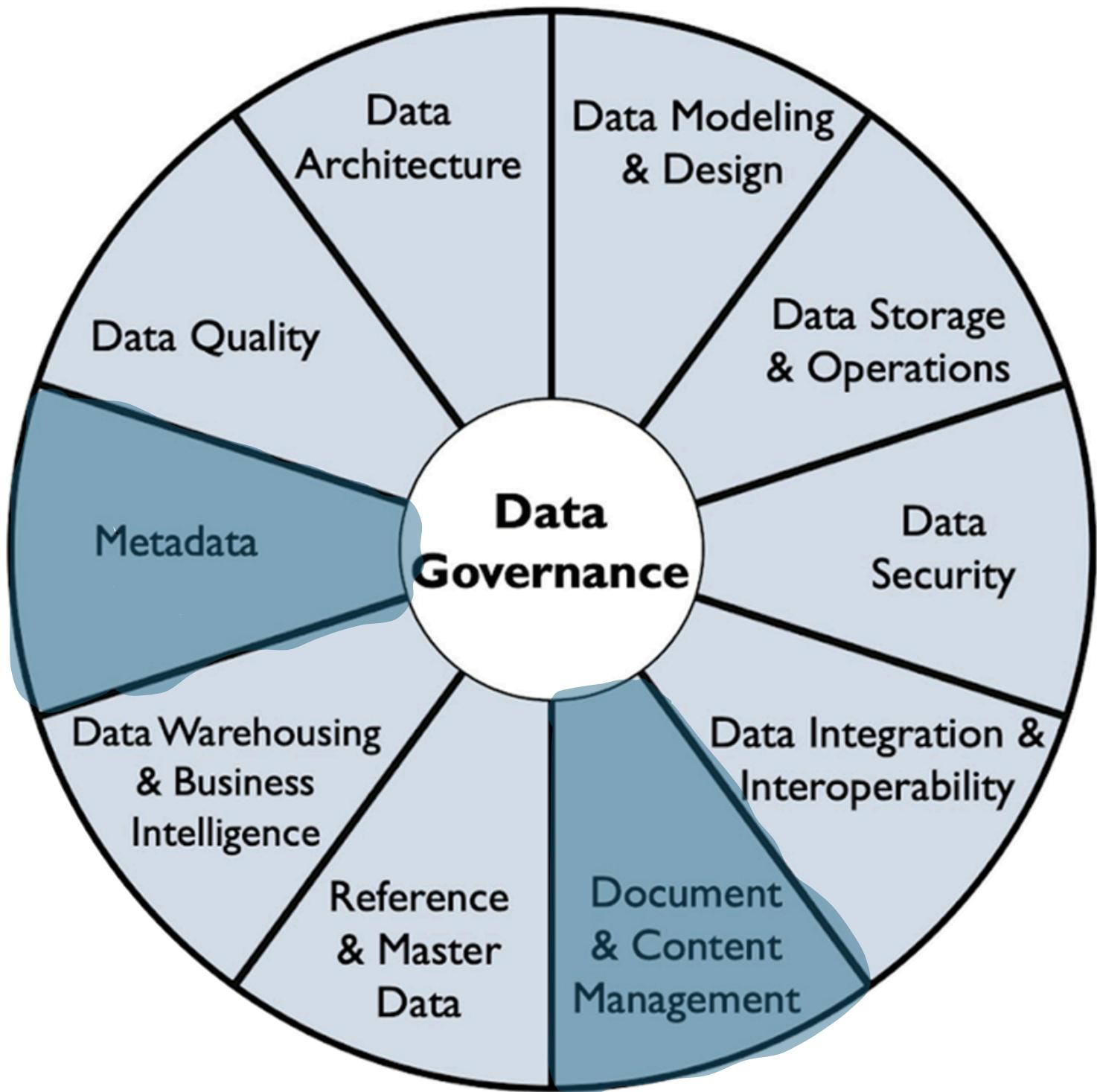
The expiration date of this approval is 12/16/2011 at Midnight. If this project is to continue beyond that date, y
must submit an updated protocol in advance for the IRBs re-approval. If this protocol is used in conjunction with a
other human use it must be re-approved. Proposed changes to approved research must be reviewed and approve
prospectively by the IRB. No changes may be initiated without prior approval by the IRB, except where necessary
eliminate apparent immediate hazards to subjects. (Any such exceptions must be reported to the IRB within 1
working days.) Unanticipated problems involving risks to participants or others and other events or information tha
defined and listed in the Report Form, must be submitted promptly to the IRB. (See Events and Information tha
Require Prompt Reporting to the IRB at <http://humansubjects.stanford.edu>.)

All continuing projects and activities must be reviewed and re-approved on or before Midnight of the expiration date.
The approval period will be less than one year if so determined by the IRB. It is your responsibility to resubmit the
project to the IRB for continuing review and to report the completion of the project to the IRB within 30 days.

Please remember that all data, including all signed consent form documents, must be retained for a minimum of
three years past the completion of this research. Additional requirements may be imposed by your funding agency,
<http://stanford.edu/dept/DoR/rph/2-10.html>.

This institution is in compliance with requirements for protection of human subjects, including 45 CFR 46, 21 CFR 50
and 56, and 38 CFR 16.

Page 27 of 41





Data Documentation

Using Microsoft Purview

- Data Catalogue

The screenshot shows the Microsoft Purview Data Catalogue interface. At the top, it displays the dataset name "howcouplesmeet2009" with a "Certified" badge, the source "Azure Blob Resource Set", and the last update date "Updated on October 12". Below the header, there are tabs for "Edit", "Select for bulk edit", "Request access", "Refresh", "Delete", and "Edit columns". The main area is divided into sections: "Overview" (showing 725 of 725 items) and "Properties". The "Schema" section lists various columns with their properties:

Schema	Column name	Glossary terms	Data type	Column description	Sensitivity label
Lineage	caseid_new		int	Case Identifier: A unique identifier ...	
Contacts	w3_Weight	Wave 3, General Population	float	The weight factor used for general ...	
Related	w3_Weight_LGB	Wave 3, LGB-Specific Weight	float	Weight applied specifically to LGB r...	
Privacy	w3_combo_weight	Combined Weights, Wave 3	float	Combined weight factor accountin...	
Risk Responses	w3_attrition_adj_weight	Attrition-Adjusted Weights,	float	A weight factor accounting for part...	
History	w2_weight_genpop	Wave 2, General Population	float	Adjusts for representativeness in th...	
	w2_weight_LGB	Wave 2, LGB-Specific Weight	float	Corrects for any under-sampling or...	
	w2_combo_weight	Combined Weights, Wave 2	float	These weights combine both gener...	
	w2_attrition_adj_weights	Attrition-Adjusted Weights,	float	These weights adjust for participan...	



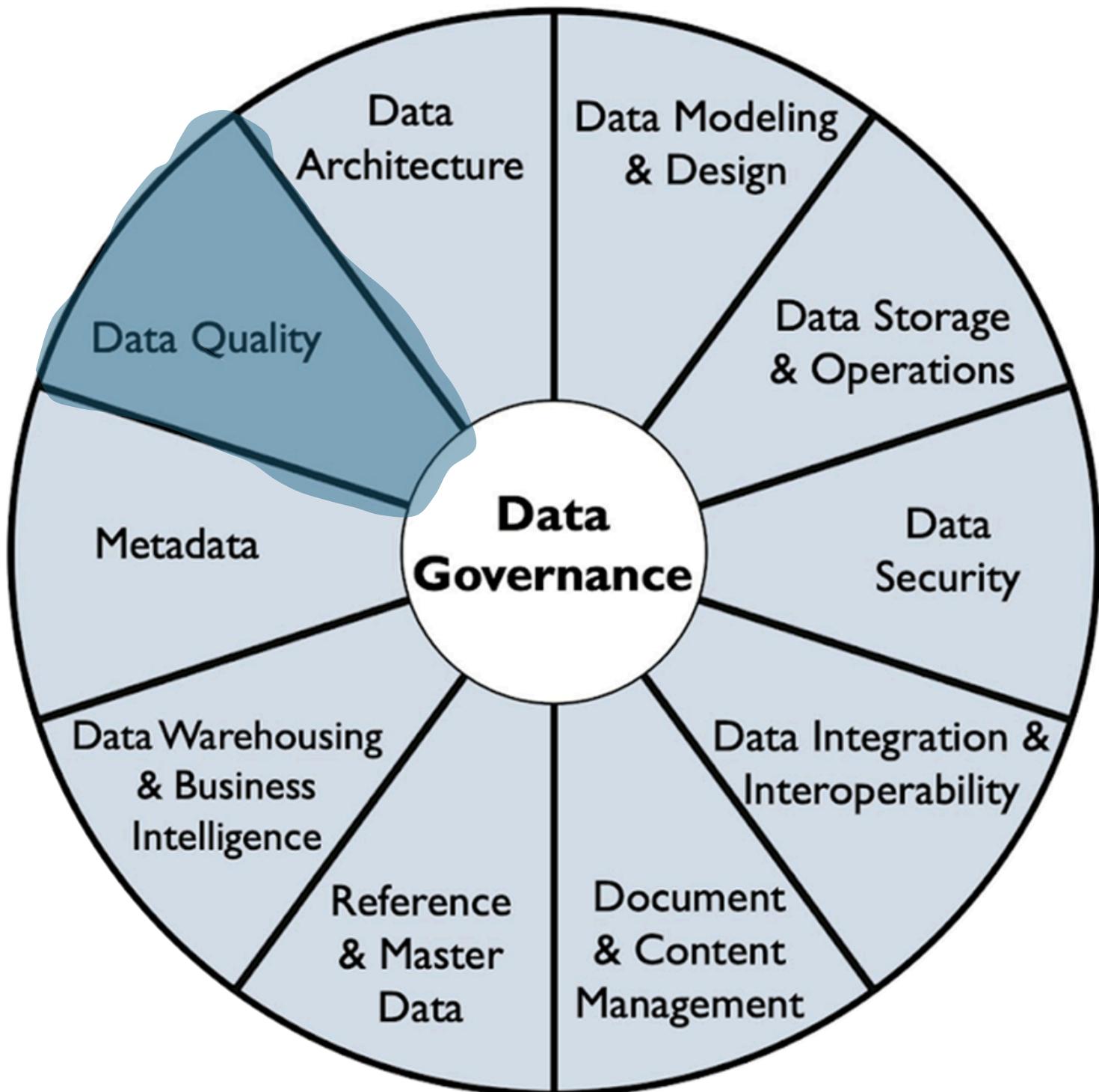
Data Documentation

Using Microsoft Purview

- Data Catalogue
- Data Glossary

The screenshot shows the Microsoft Purview Business glossary interface. The title bar reads "Business glossary" and "HowCouplesMeet". The main area is titled "HowCouplesMeet" and shows a list of 41 terms. The columns are: Name, Term template, Status, Definition, Stewards, and Experts. The list includes categories like Demographic, Age, Education, Family, Household head, Household size, and Number of Children. Most terms are marked as "Approved".

Name	Term template	Status	Definition	Stewards	Experts
Demographic	System default	Approved	These are all va...	-	-
Age	System default	Approved	Respondent ag...	-	-
Age - 4 categories	System default	Approved	Variable name: ...	-	-
Age - 7 Categories	System default	Approved	Variable name: ...	-	-
Education (highest degree received)	System default	Approved	Variable name:...	-	-
Education (categorical)	System default	Approved	Variable name:...	-	-
Family	System default	Approved	These variables...	-	-
Household head	System default	Approved	Variable name: ...	-	-
Household size	System default	Approved	Including yours...	-	-
Number of Children	System default	Approved	Not a column L...	-	-
Number of adult children	System default	Approved	number of adul...	-	-

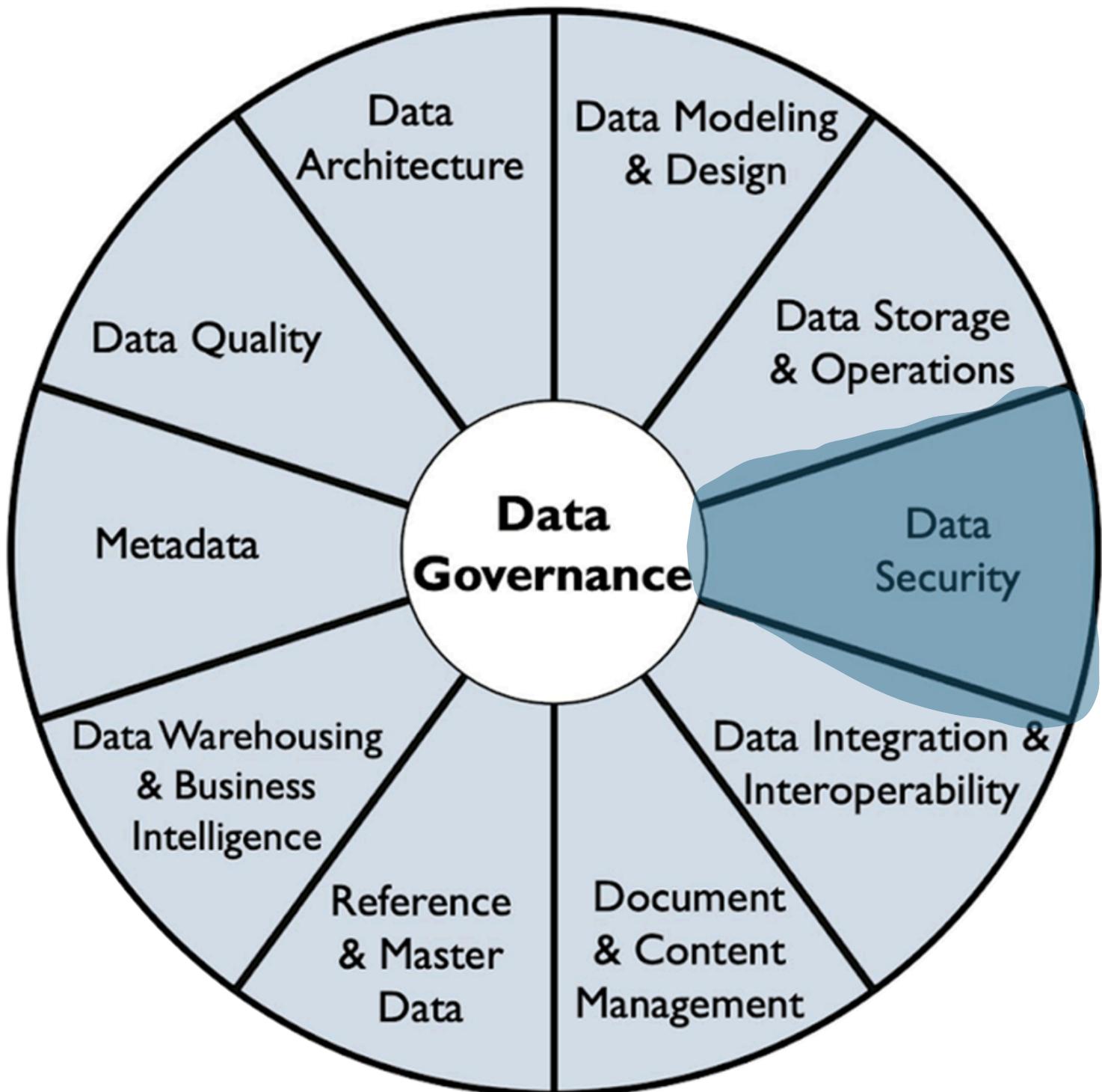




Data Quality

- Dataset
 - Up-to-date (2009-2015 & 2017-2022)
 - Consistency
- Accuracy
- Completeness

Value	Label	Unweighted Frequency	%
0	still together	2231	55.7 %
1	broke up	248	6.2 %
2	partner passed away	41	1.0 %
	Missing Data		
.	-	1482	37.0 %
	Total	4,002	100%





Data Security

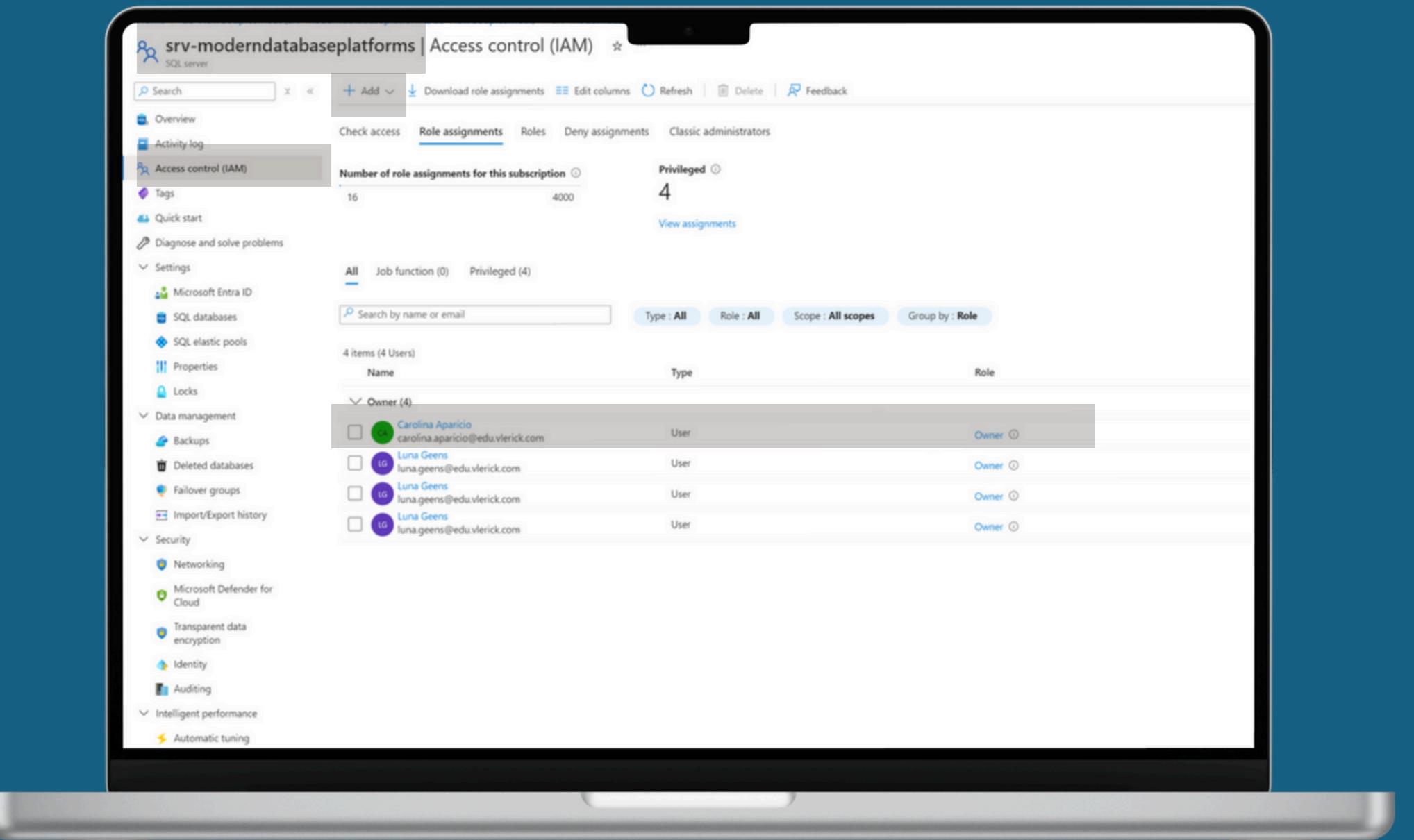
Data security is where we ensure confidentiality, accountability, authentication and authorization.

Using Microsoft Purview

- Security
 - Access via Key Vault Azure
 - Access via Access Control (IAM)



Data Security

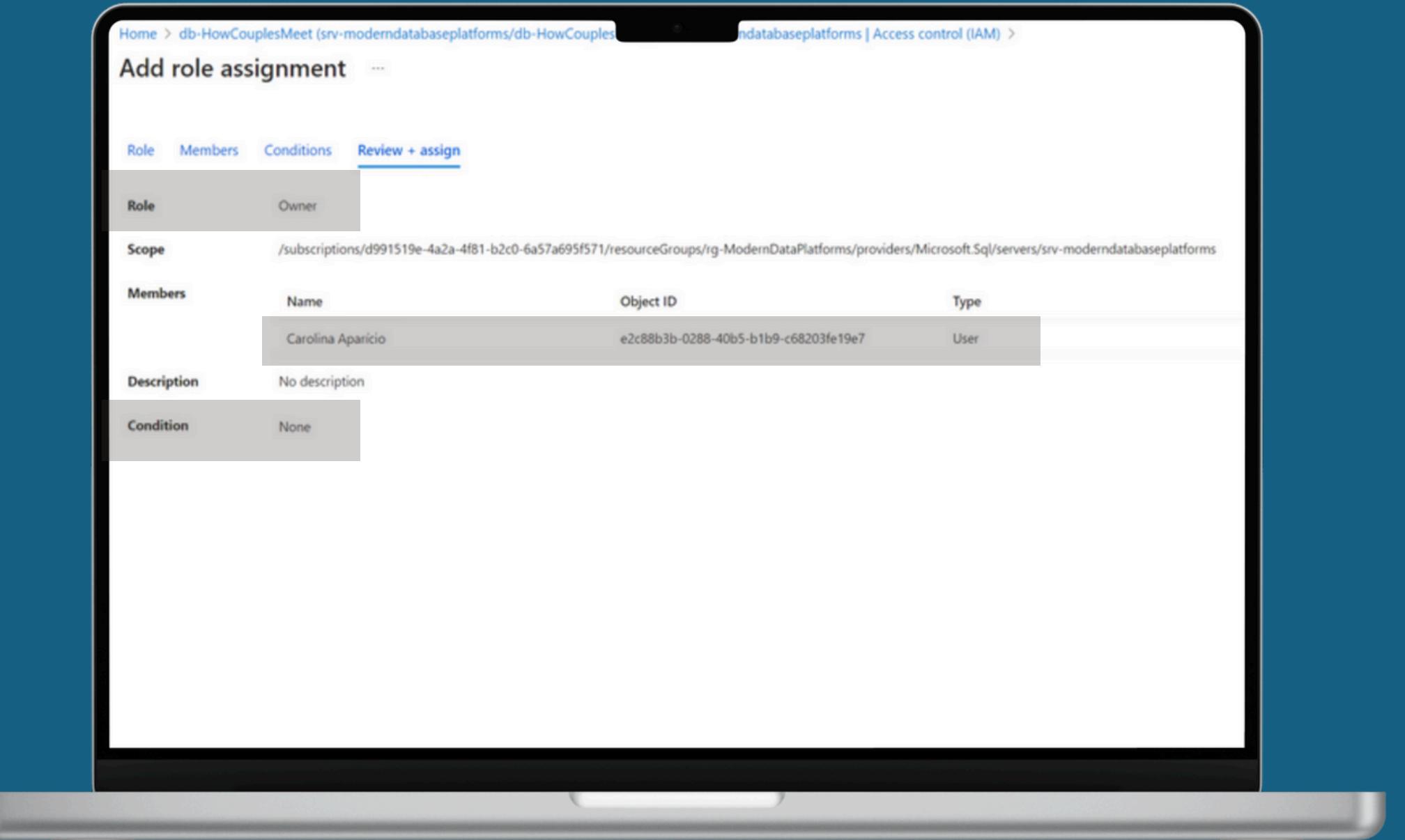


How to access the server, when you did not make it?

Access for owner role assignment.



Data Security



How to
access the server,
when you did not
make it?

Access for owner
role assignment.



Data Security

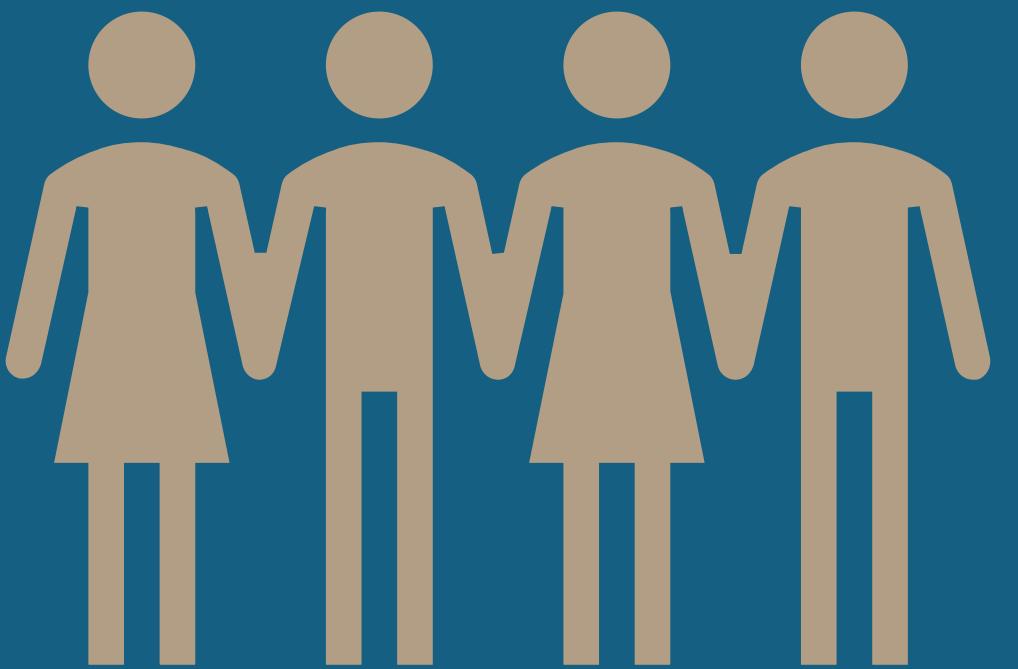
How to
access the notebooks,
when you did not
make it?

Access for owner
role assignment.

3 Conclusion

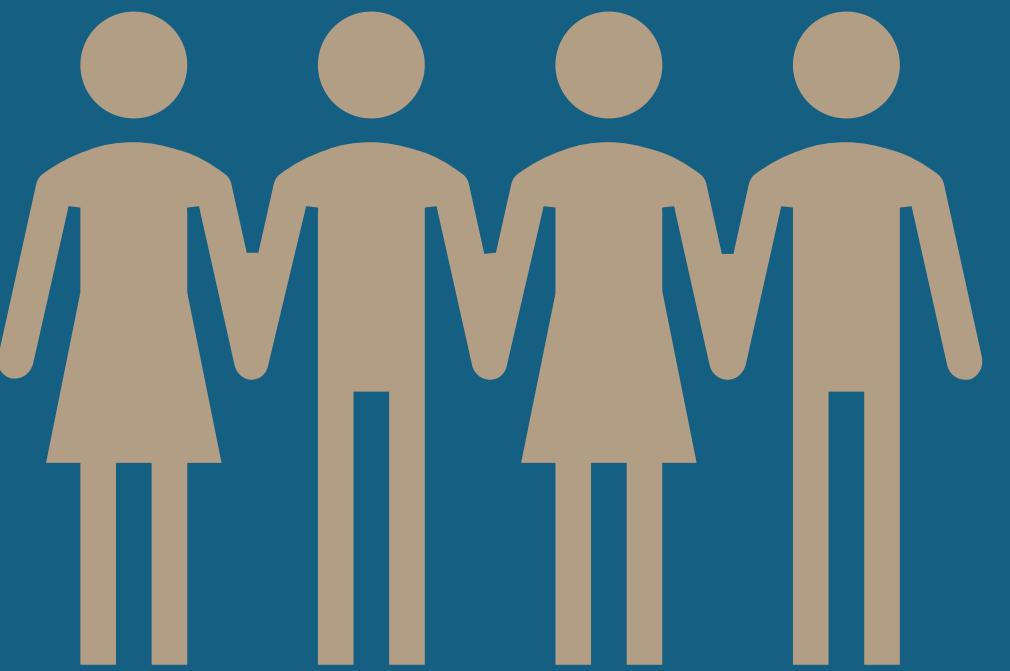
Conclusion

- What did we learn?
 - Less is more when it comes to variables.
 - How to clean a dataset?
 - How to use Azure Purview?
 - Tools for every purposes



Conclusion

- Struggles
 - How to teach non-technical people these skills?
 - How to contribute work more equal?
 - How can you work more efficiently in terms of time, stress, and finances?
 - Integration between services.



Conclusion

- What could we do next time?
 - Don't let one person struggle on their own.
 - Let one non-tech and one tech person work together.
 - Time management.

