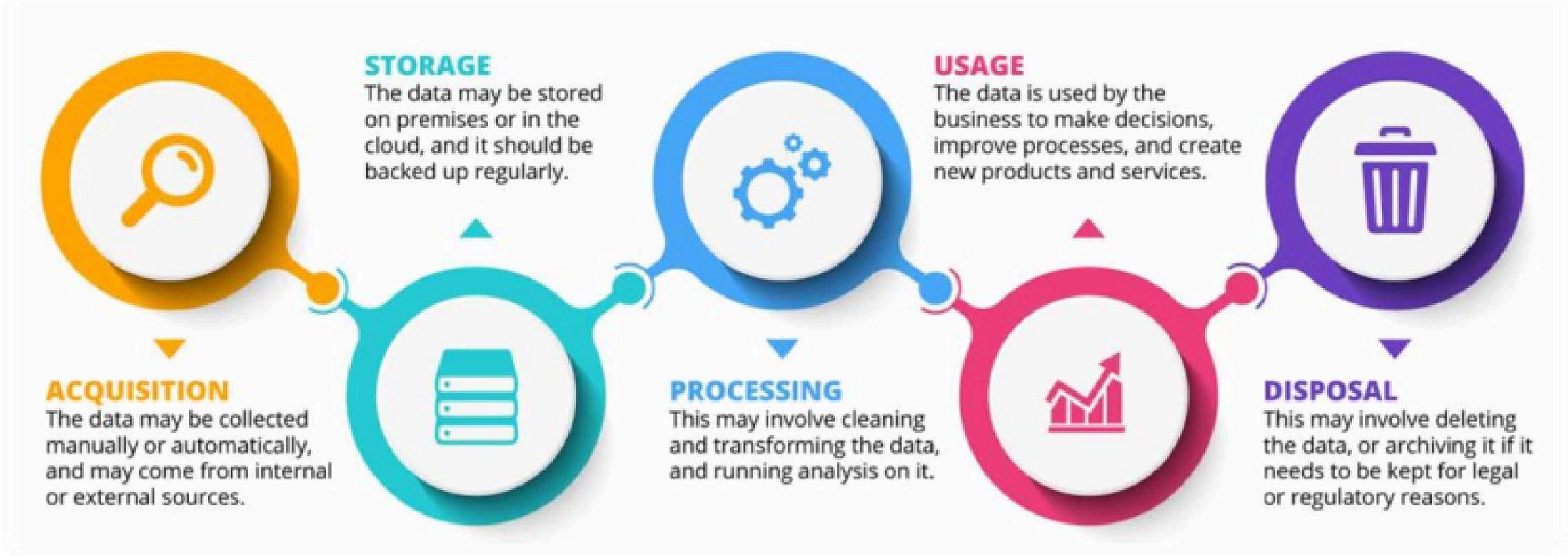
Modern Data Platforms

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

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

1 Data Lifecyle Managament


STORAGE
The data may be stored on premises or in the cloud, and it should be backed up regularly.
USAGE
The data is used by the business to make decisions, improve processes, and create new products and services.
ACQUISITION
The data may be collected manually or automatically, and may come from internal or external sources.
PROCESSING
This may involve cleaning and transforming the data, and running analysis on it.
DISPOSAL
This may involve deleting the data, or archiving it if it needs to be kept for legal or regulatory reasons.

1 Data Lifecyle Managament

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.

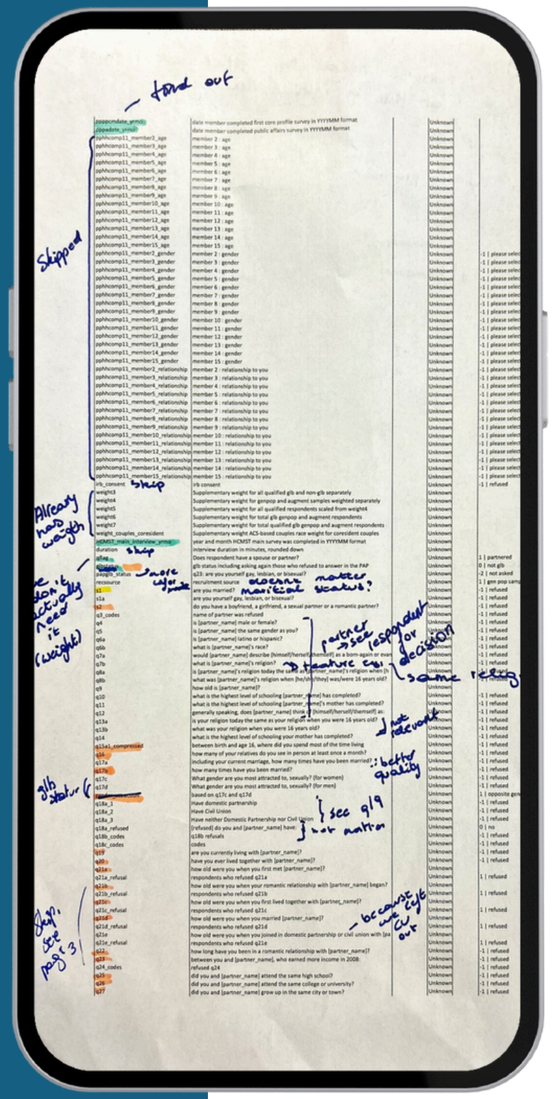
TOO big data?

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

With “common sense”: we went trough 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 Lifecyle Managament

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

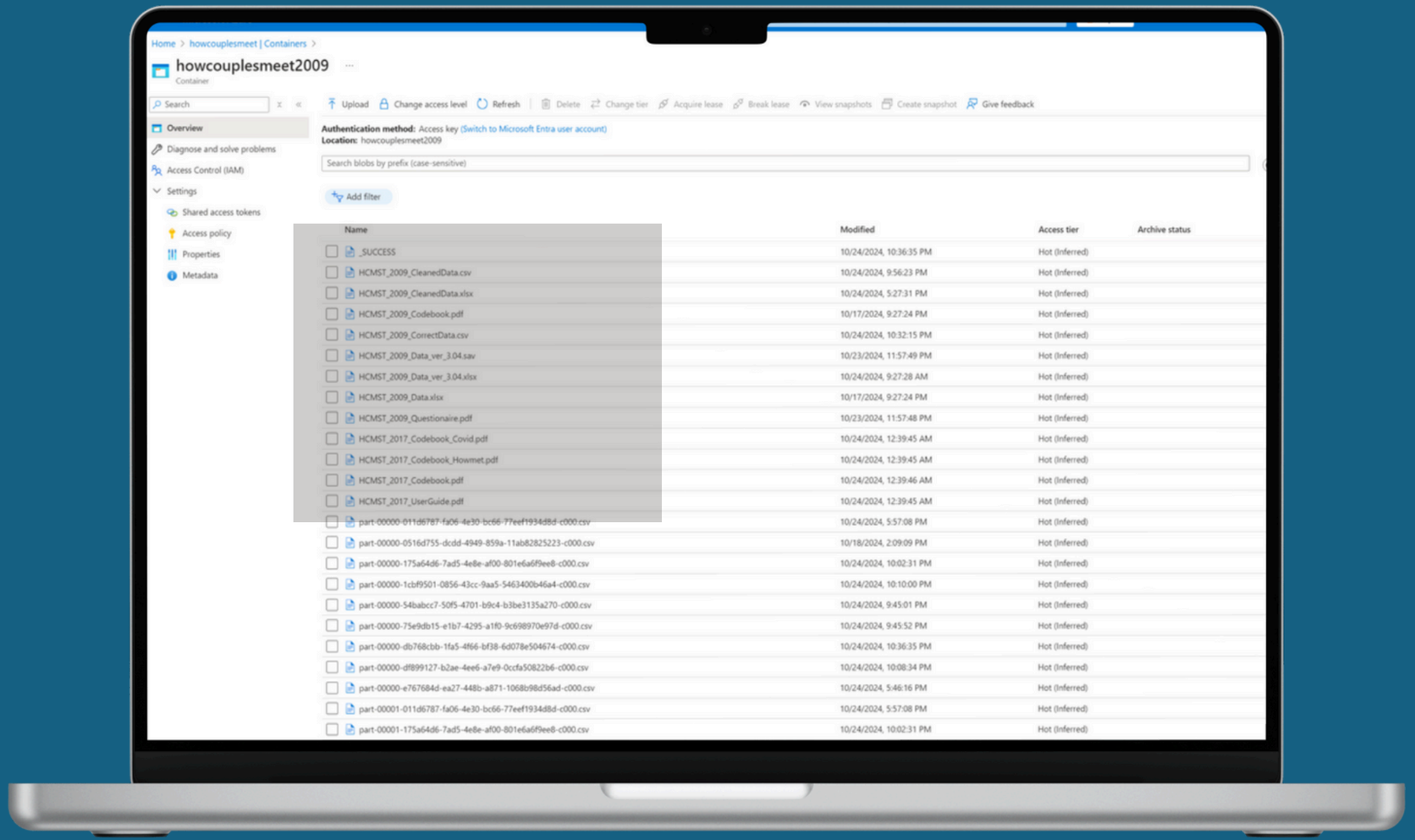
Transform

Azure SQL server connection in python

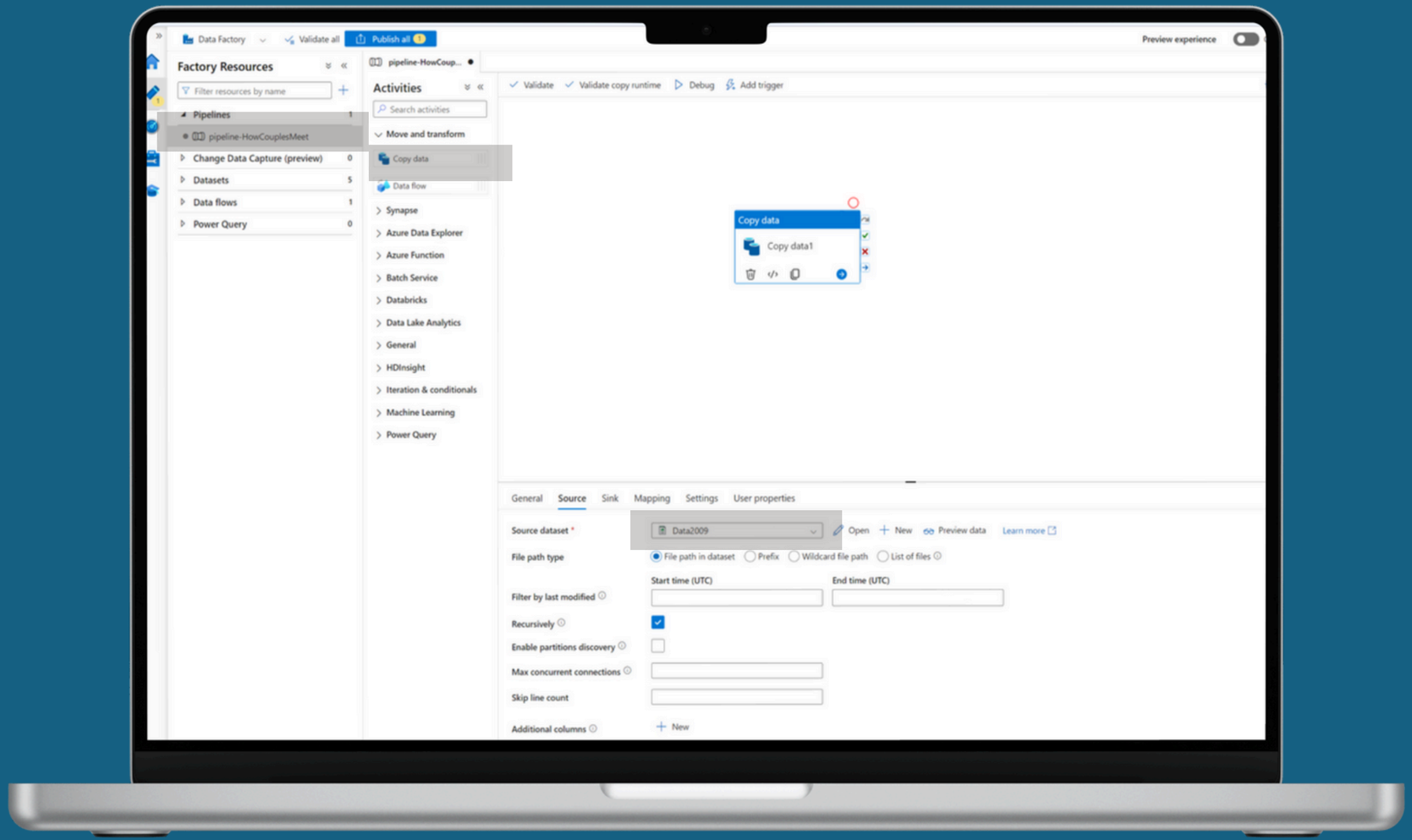
⬇

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

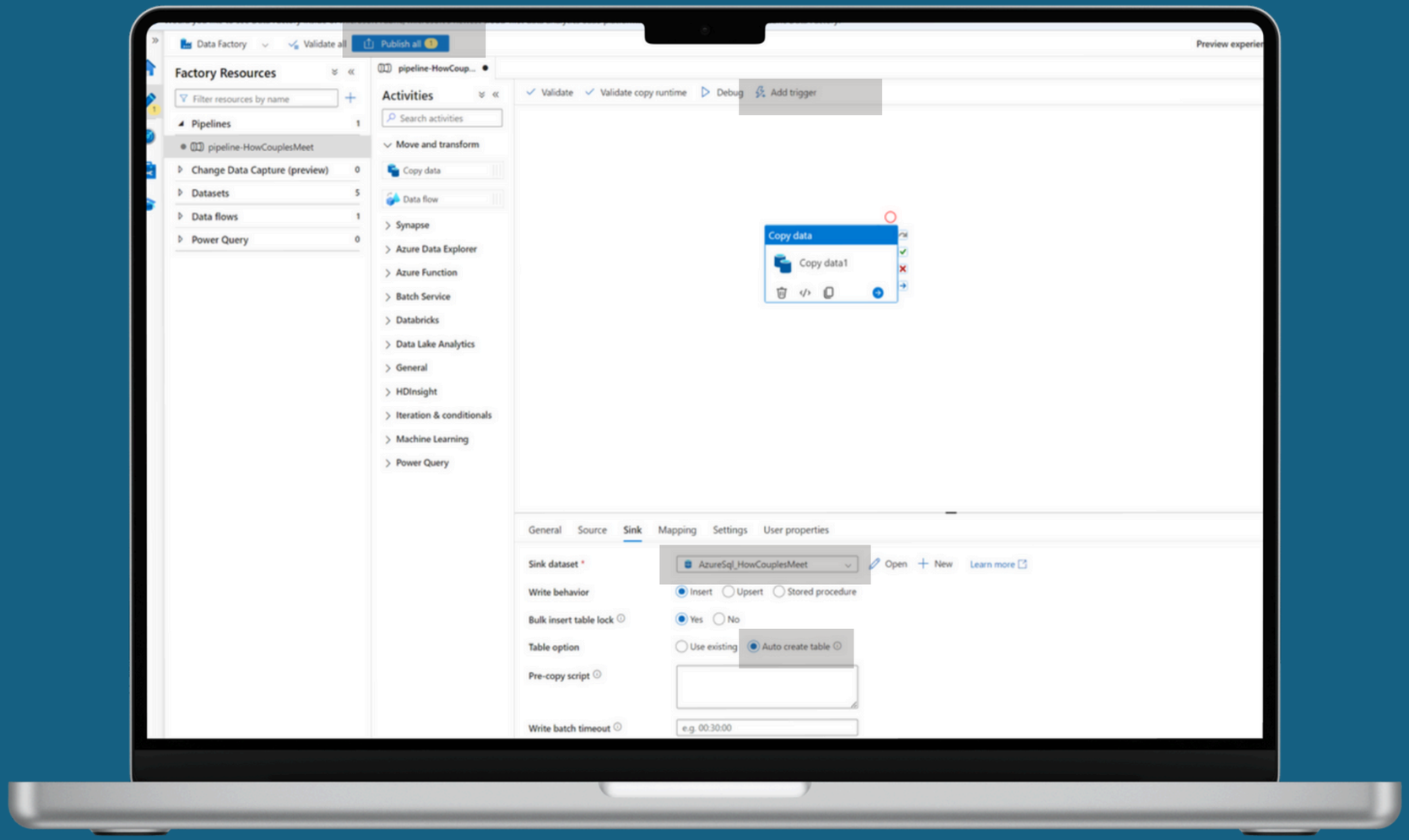
Extract: Azure Blob Storage



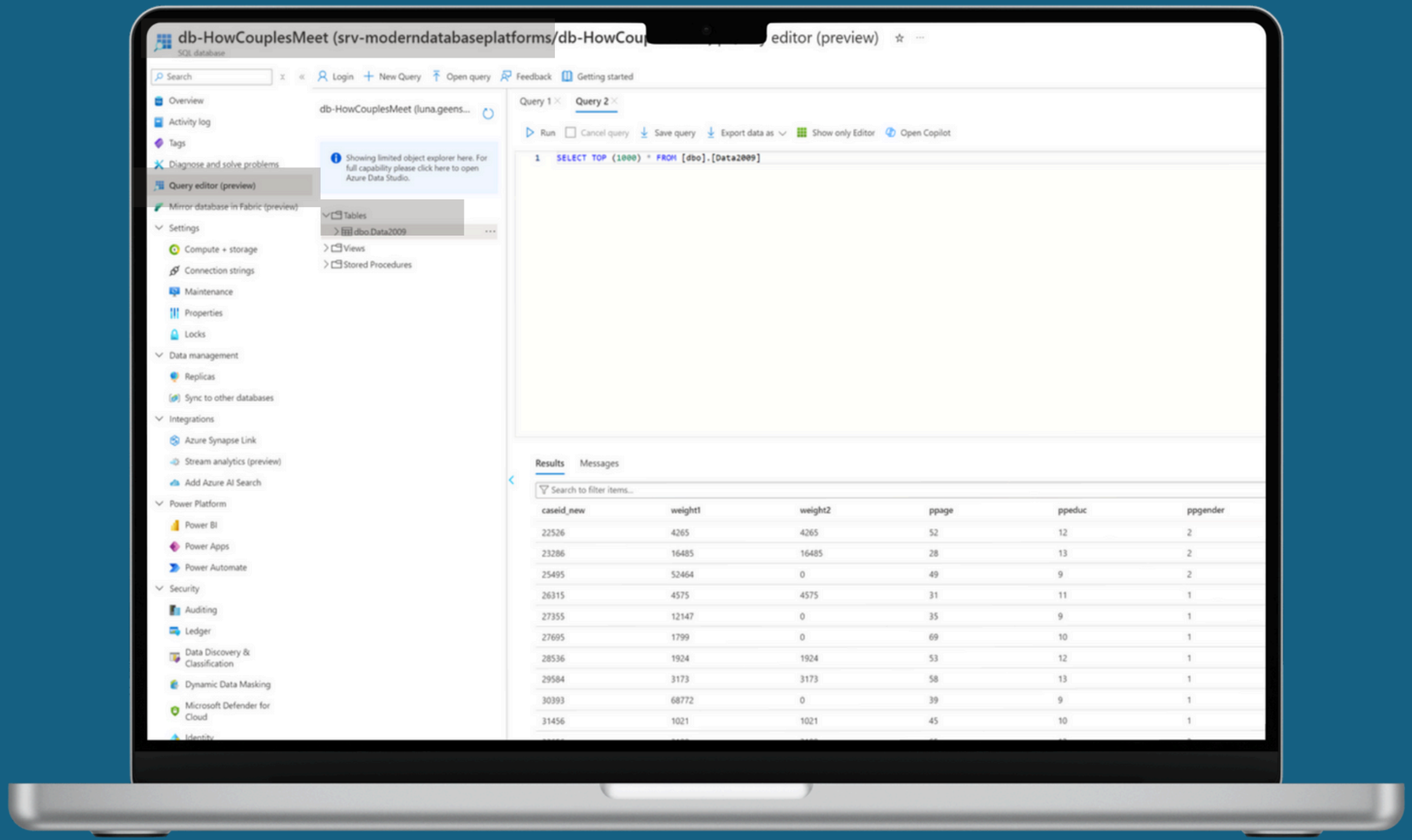
Load: Azure Data Factory



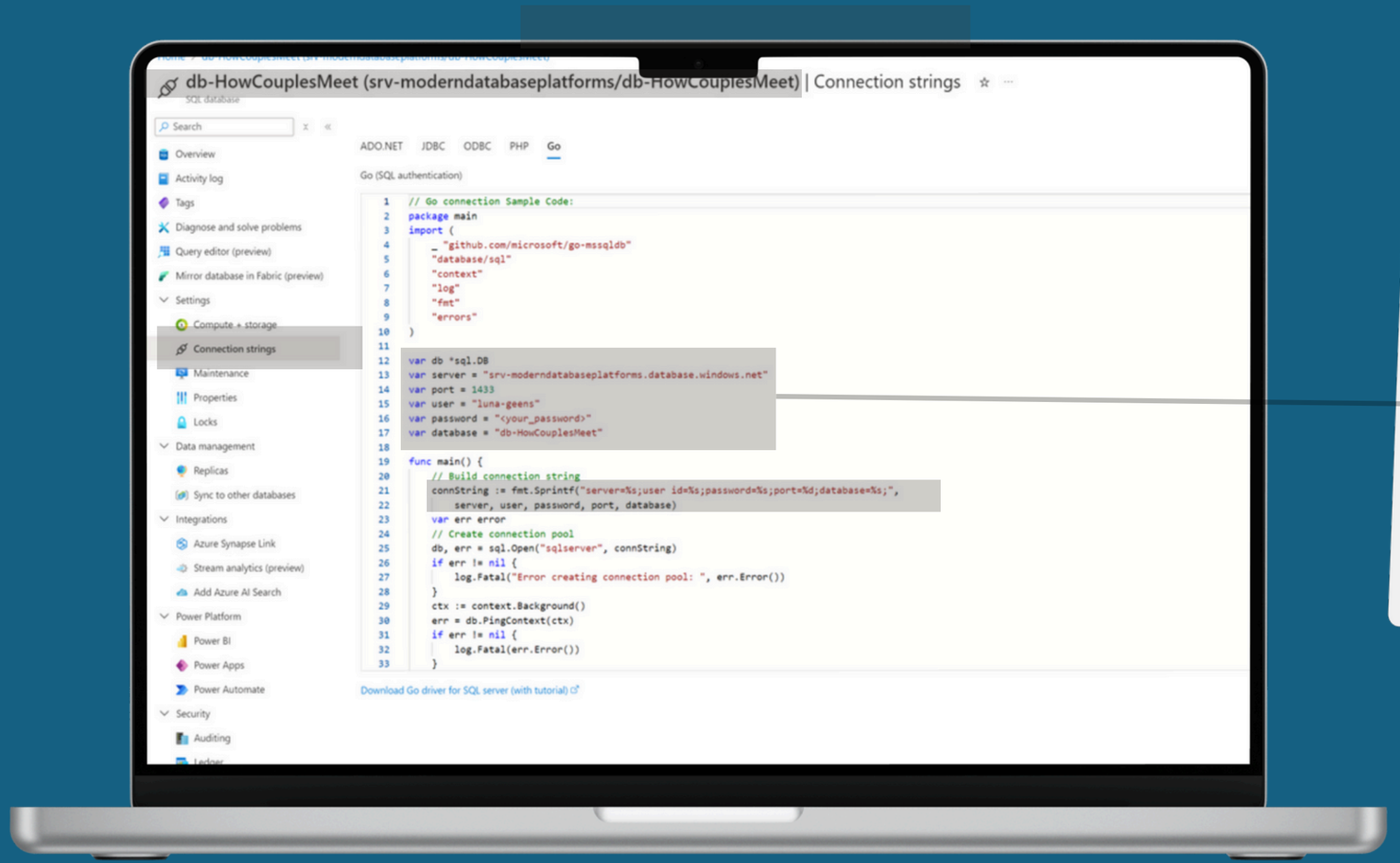
Load: Azure Data Factory

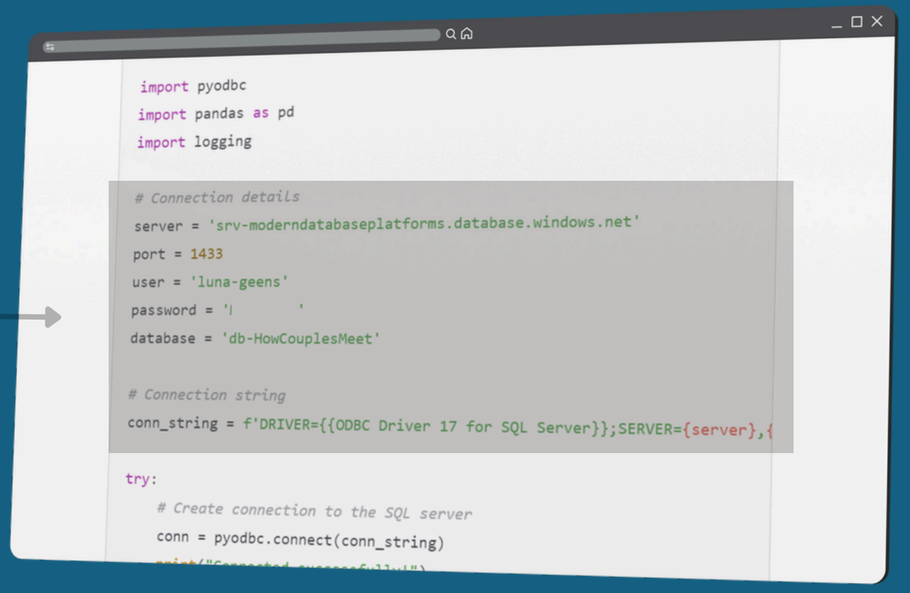


Load: Azure SQL Databases

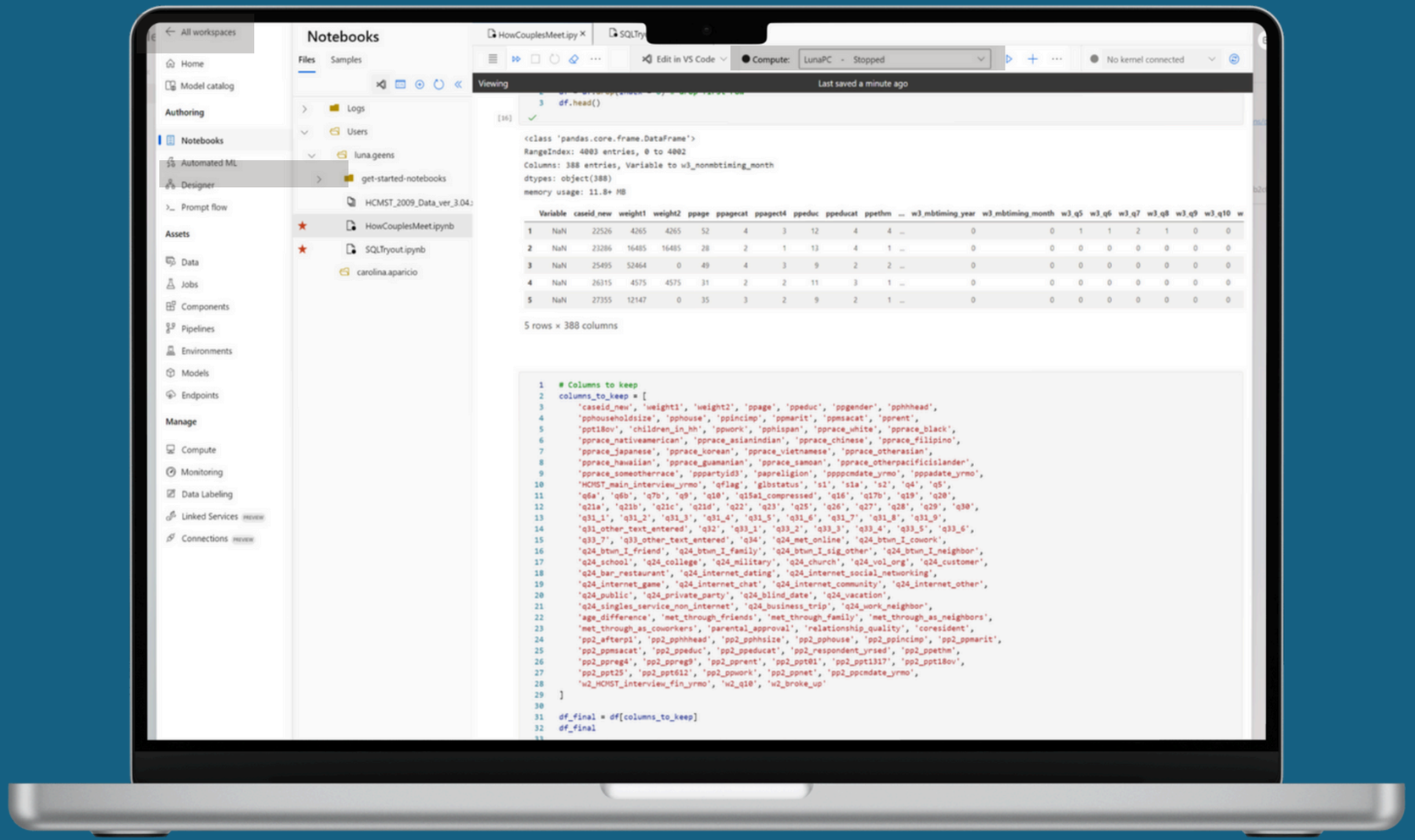


Transform: Azure Machine Learning





Transform: Azure Machine Learning



1 Data Lifecyle Managament

1.3 Processing

Some influencial 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?

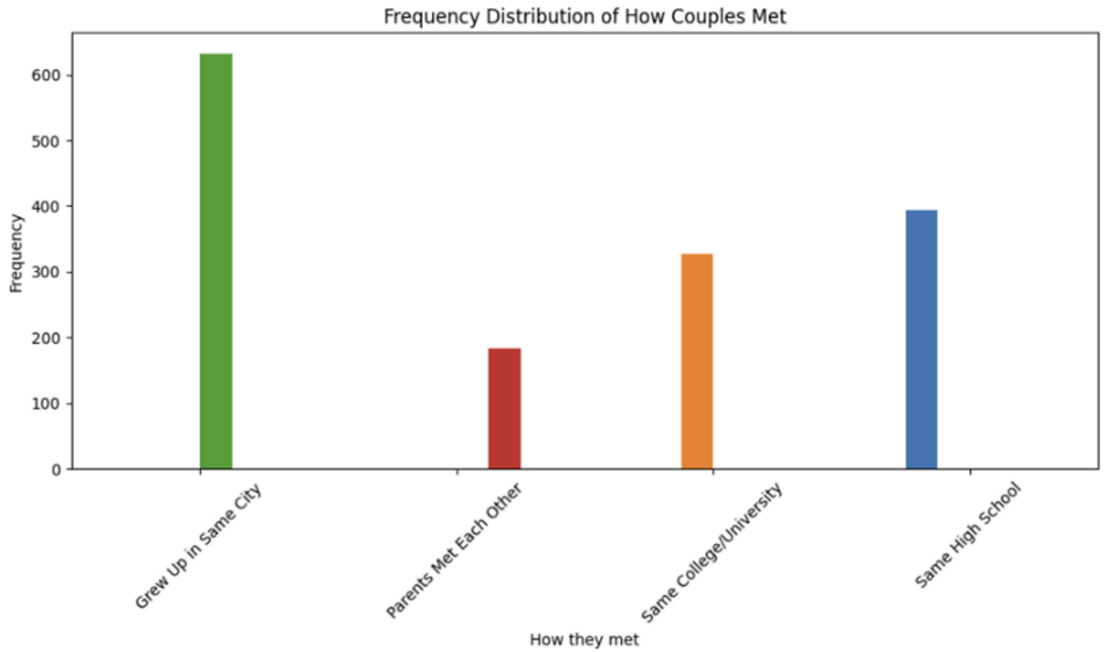
	Category	Count	Percentage
1	Yes	161	26.7
0	No	442	73.3

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 influencial variables...

How did you meet?



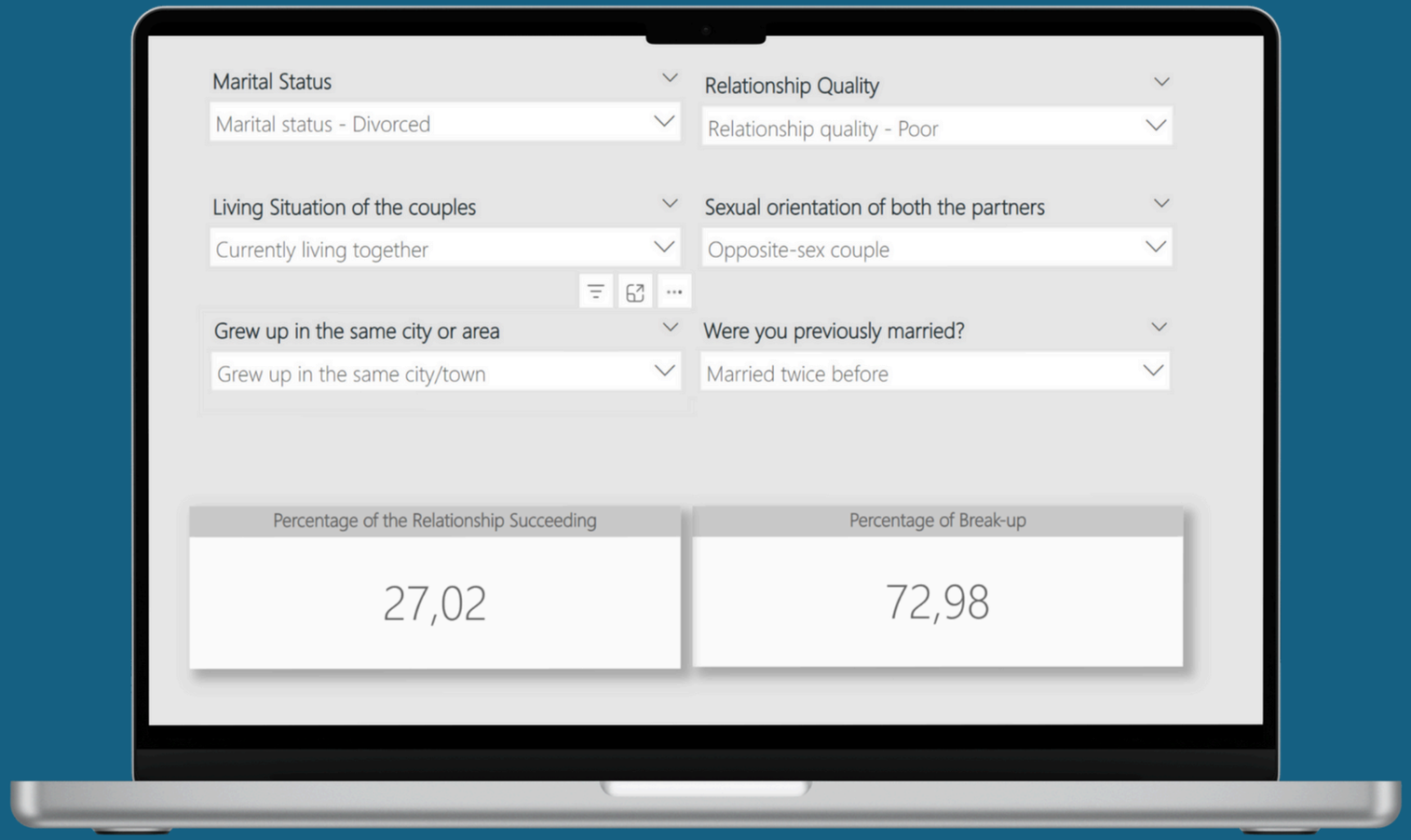
How would u rate the quality of your relationship?

	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

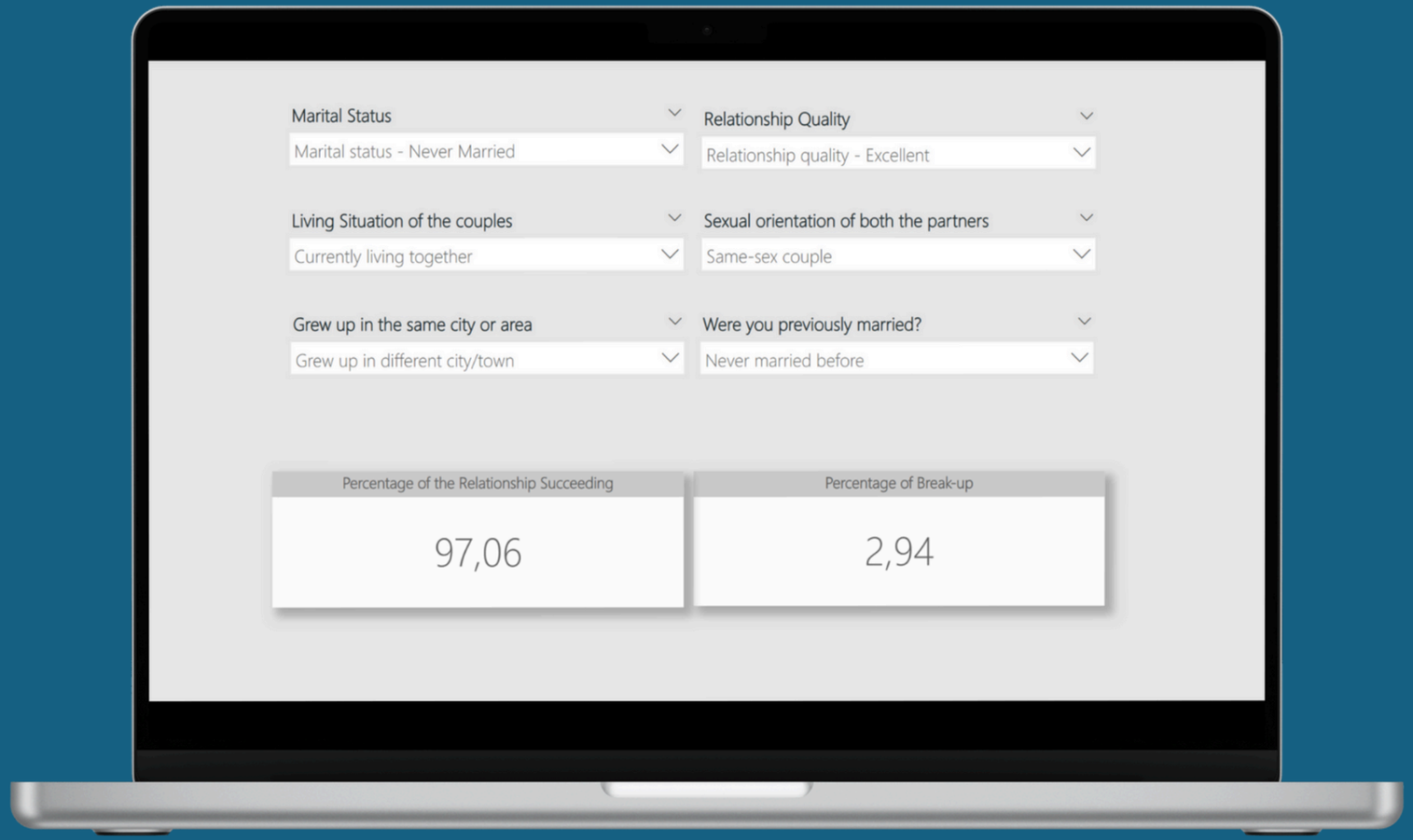
1 Data Lifecyle Managament

1.4 Usage

Power BI


Marital Status
Marital status - Divorced
Relationship Quality
Relationship quality - Poor
Living Situation of the couples
Currently living together
Sexual orientation of both the partners
Opposite-sex couple
Grew up in the same city or area
Grew up in the same city/town
Were you previously married?
Married twice before
Percentage of the Relationship Succeeding
27,02
Percentage of Break-up
72,98

Power BI


Marital Status
Marital status - Never Married
Relationship Quality
Relationship quality - Excellent
Living Situation of the couples
Currently living together
Sexual orientation of both the partners
Same-sex couple
Grew up in the same city or area
Grew up in different city/town
Were you previously married?
Never married before
Percentage of the Relationship Succeeding
97,06
Percentage of Break-up
2,94

1 Data Lifecyle Managament

1.5 Disposal

Problem 5: Time

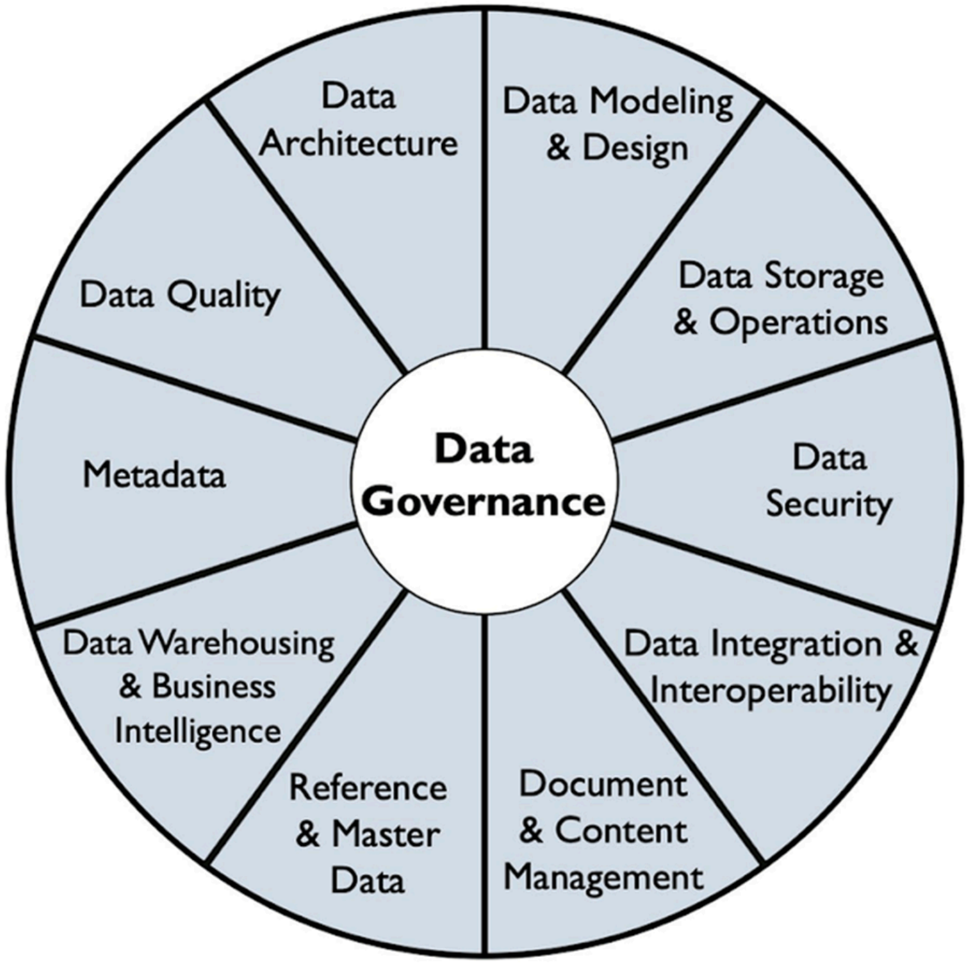
By the time we were at this phase....

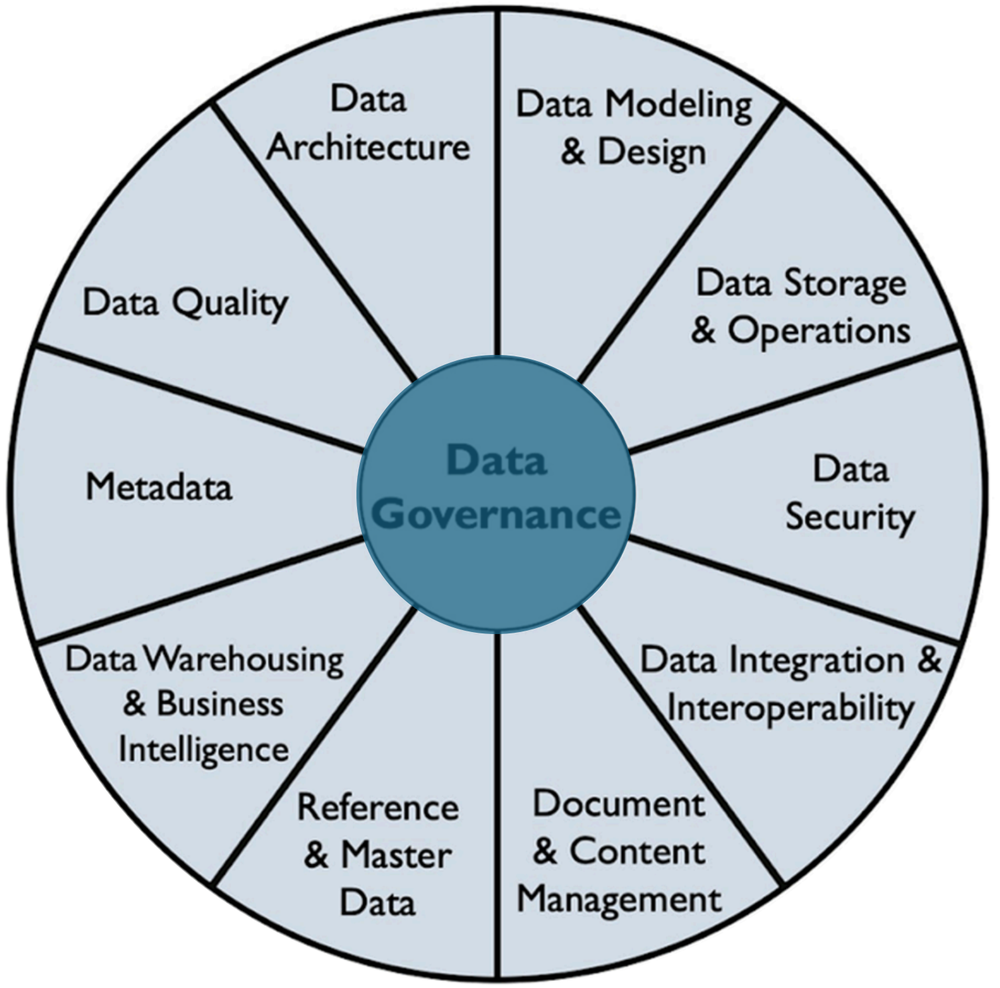
All jokes aside,

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

2 Data Management Knowledge Areas

DAMA DMBoK


Data Governance
Data Architecture
Data Modeling & Design
Data Storage & Operations
Data Security
Data Integration & Interoperability
Document & Content Management
Reference & Master Data
Data Warehousing & Business Intelligence
Metadata
Data Quality

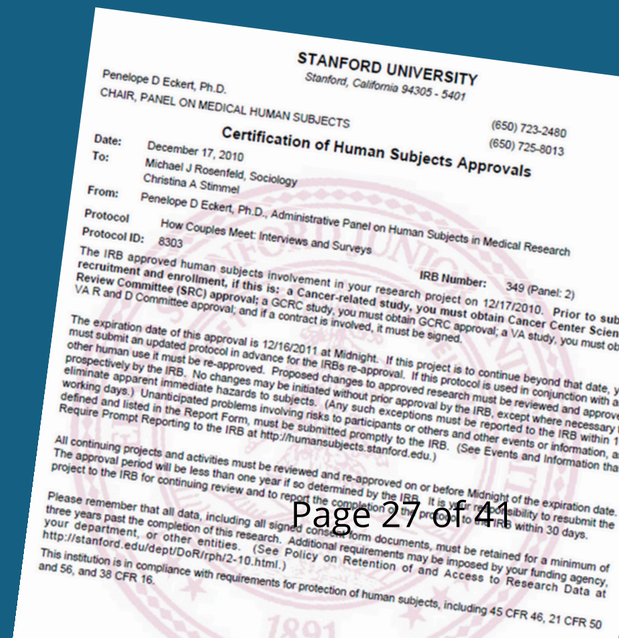

Data Governance
Data Architecture
Data Modeling & Design
Data Storage & Operations
Data Security
Data Integration & Interoperability
Document & Content Management
Reference & Master Data
Data Warehousing & Business Intelligence
Metadata
Data Quality

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

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

(650) 723-2480
(650) 725-8013

Certification of Human Subjects Approvals

Date: December 17, 2010
To: Michael J Rosenfeld, Sociology
Christina A Stimmel
From: Penelope D Eckert, Ph.D., Administrative Panel on Human Subjects in Medical Research
Protocol How Couples Meet: Interviews and Surveys
Protocol ID: 8303
IRB Number: 349 (Panel: 2)

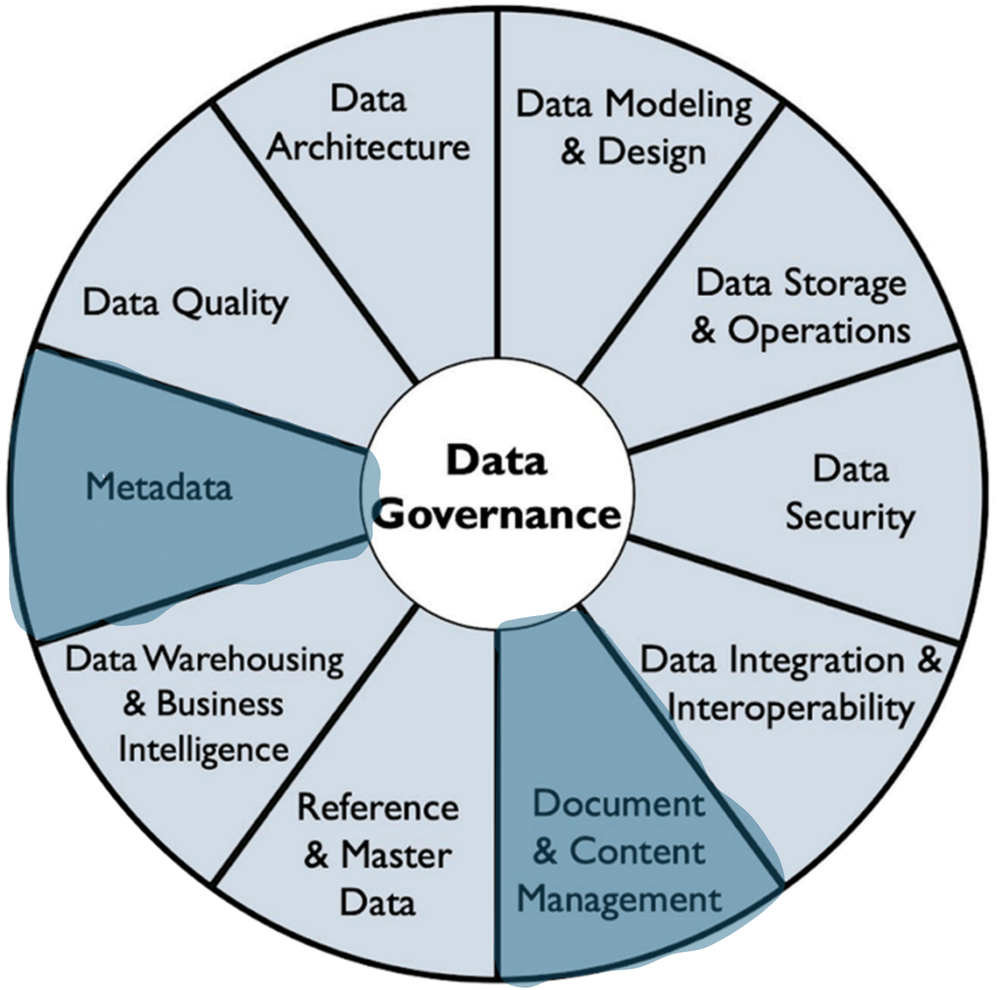
The IRB approved human subjects involvement in your research project on 12/17/2010. Prior to sub
recruitment and enrollment, if this is: a Cancer-related study, you must obtain Cancer Center Scien
Review Committee (SRC) approval; a GCRC study, you must obtain GCRC approval; a VA study, you must ob
VA 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 approv
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, a
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 r sibility 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,
your department, or other entities. (See Policy on Retention of and Access to Research Data at
http://stanford.edu/dept/DoR/rph/2-10.html.)

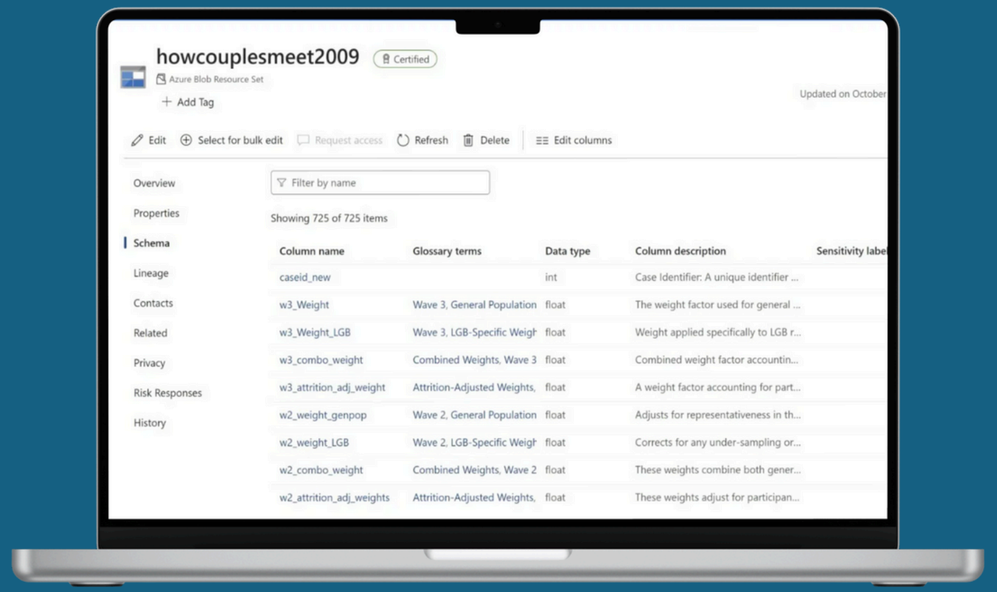
This institution is in compliance with requirements for protection of human subjects, including 45 CFR 46, 21 CFR 50
and 56, and 38 CFR 16.


Data Architecture
Data Modeling & Design
Data Storage & Operations
Data Security
Data Integration & Interoperability
Document & Content Management
Reference & Master Data
Data Warehousing & Business Intelligence
Metadata
Data Quality
Data Governance

Data Documentation

Using Microsoft Purview

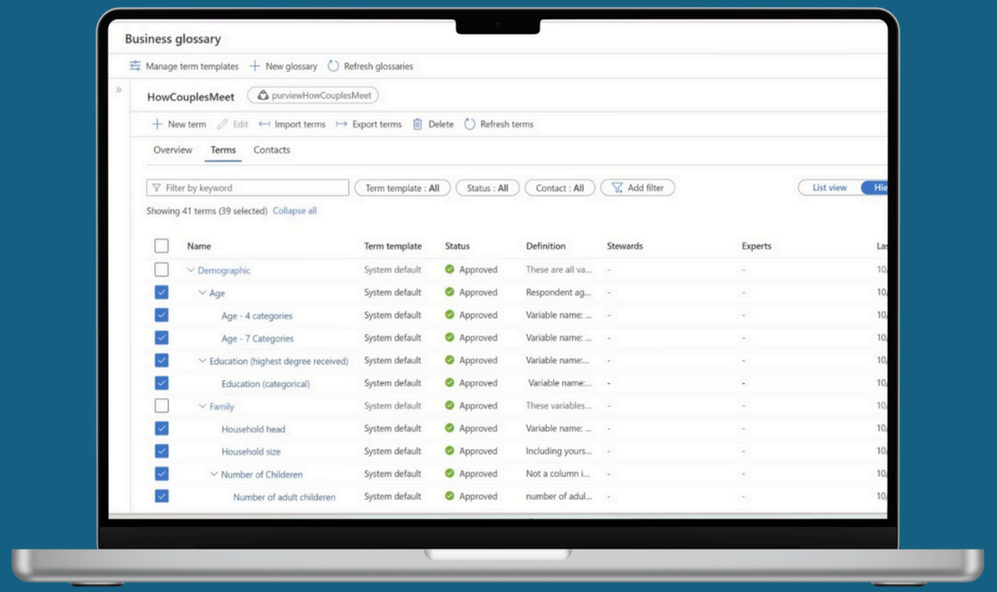
- Data Catalogue

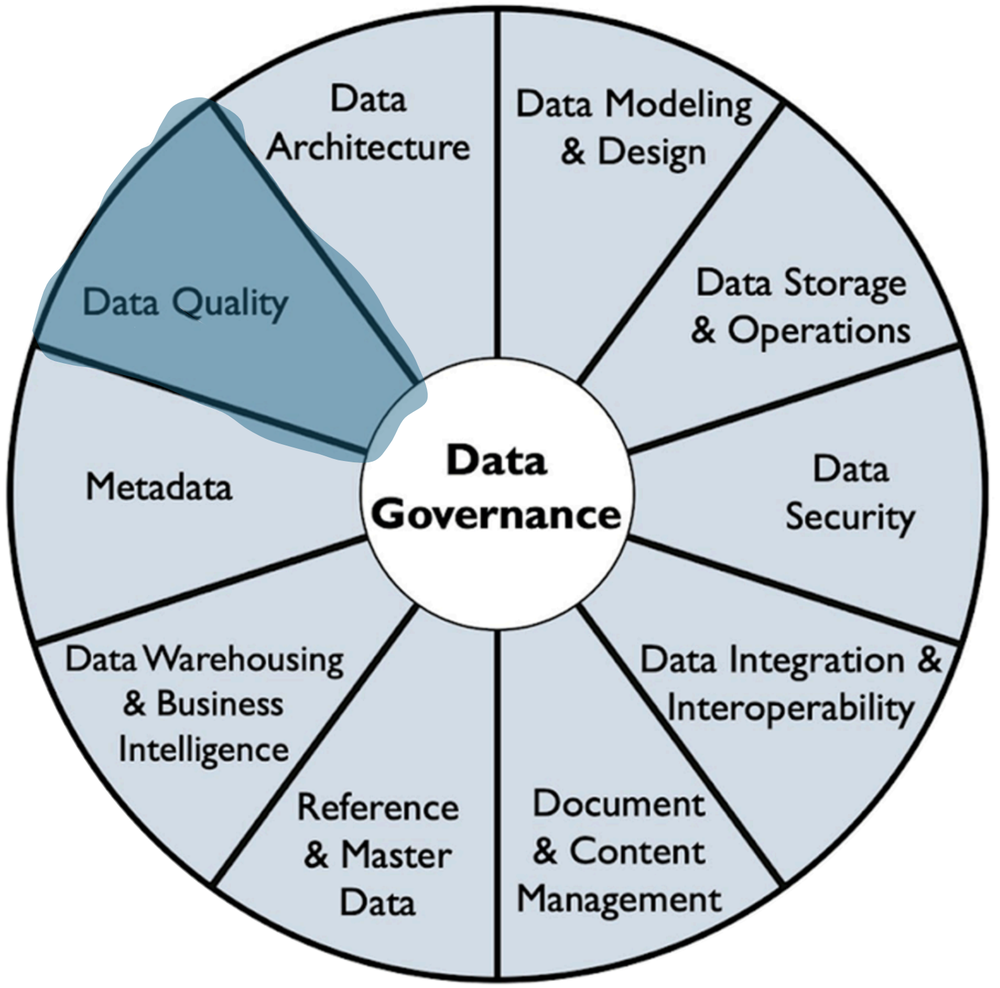


Data Documentation

Using Microsoft Purview

- Data Catalogue
- Data Glossary



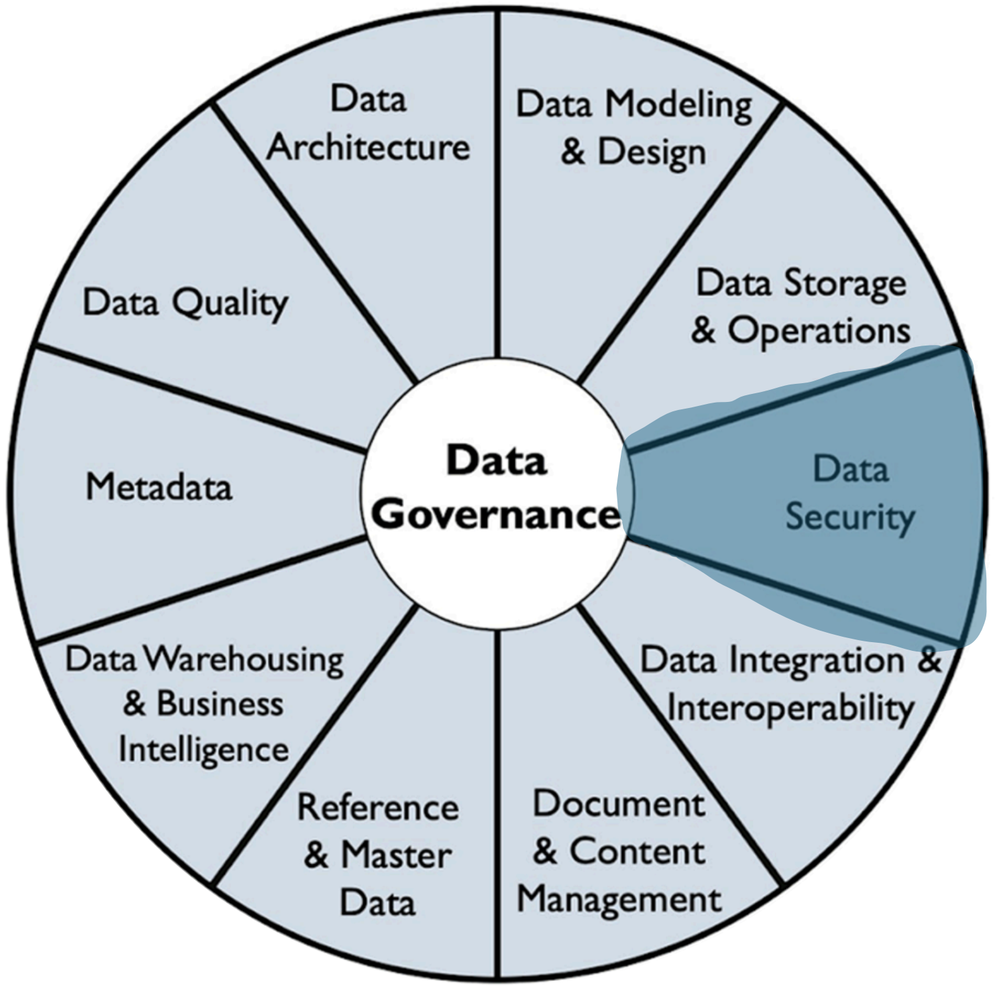

Data Architecture
Data Modeling & Design
Data Storage & Operations
Data Security
Data Integration & Interoperability
Document & Content Management
Reference & Master Data
Data Warehousing & Business Intelligence
Metadata
Data Quality
Data Governance

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 Governance
Data Architecture
Data Modeling & Design
Data Storage & Operations
Data Security
Data Integration & Interoperability
Document & Content Management
Reference & Master Data
Data Warehousing & Business Intelligence
Metadata
Data Quality

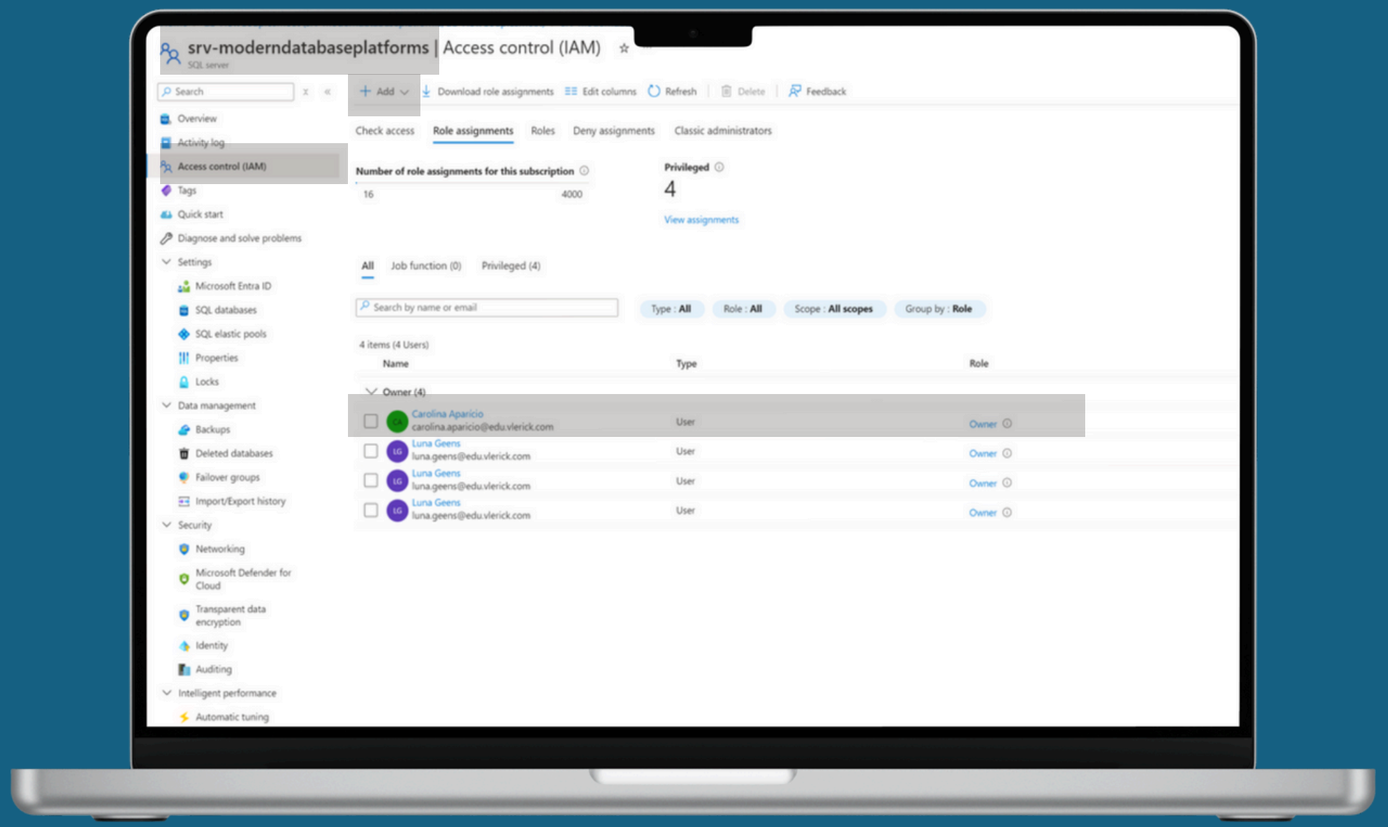
Data Security

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

Using Microsoft Purview

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

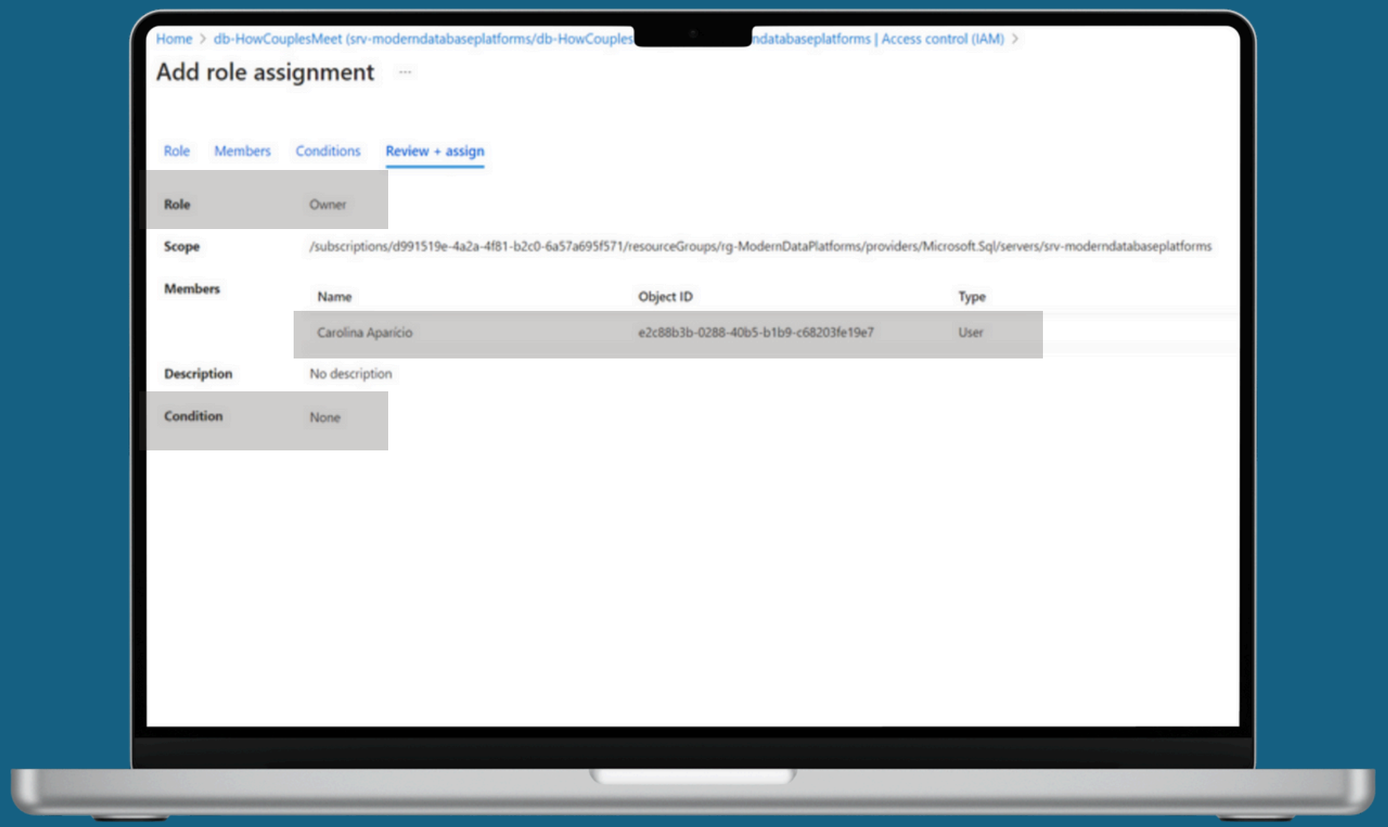
Data Security



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

Access for owner
role assignment.

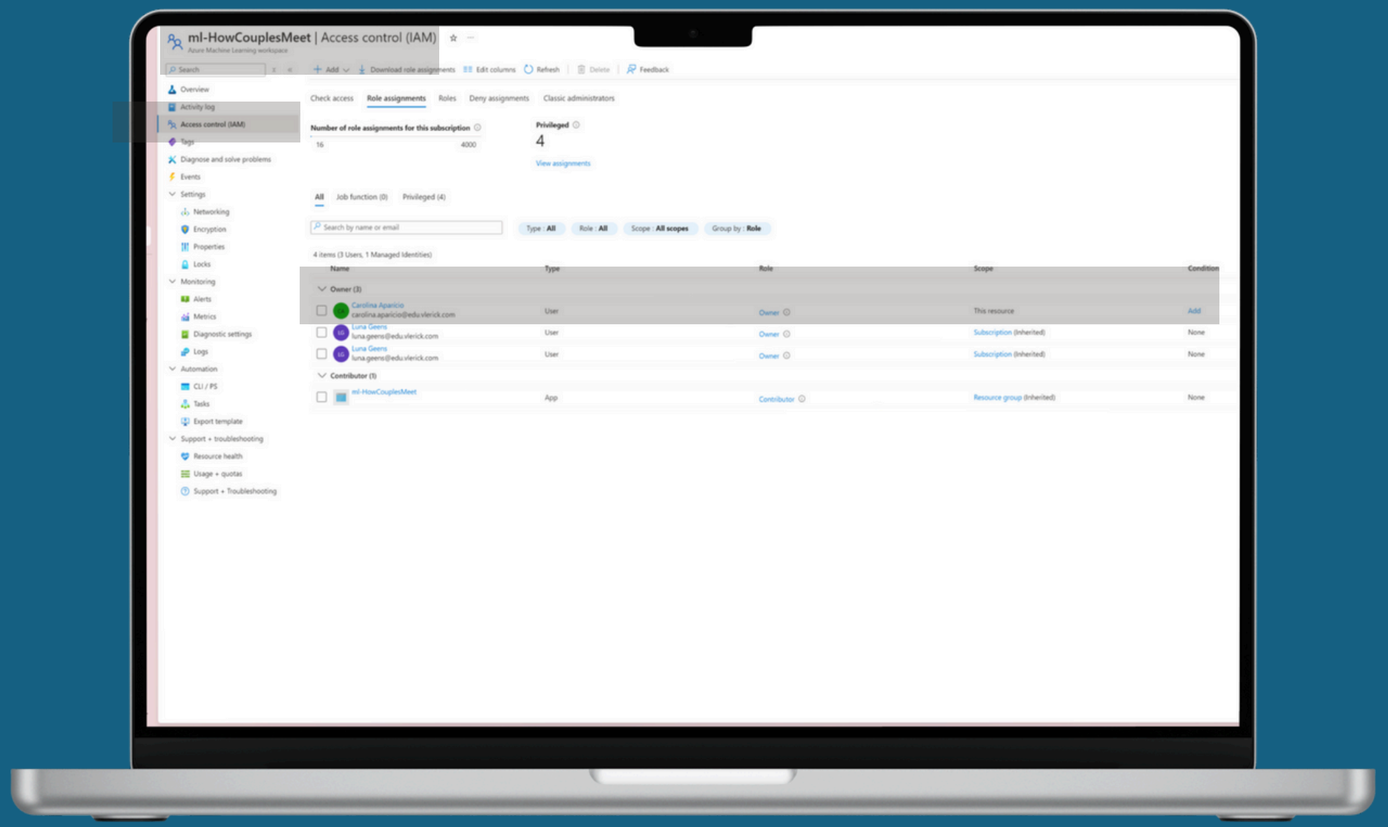
Data Security



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.