smart
casual
datadesign

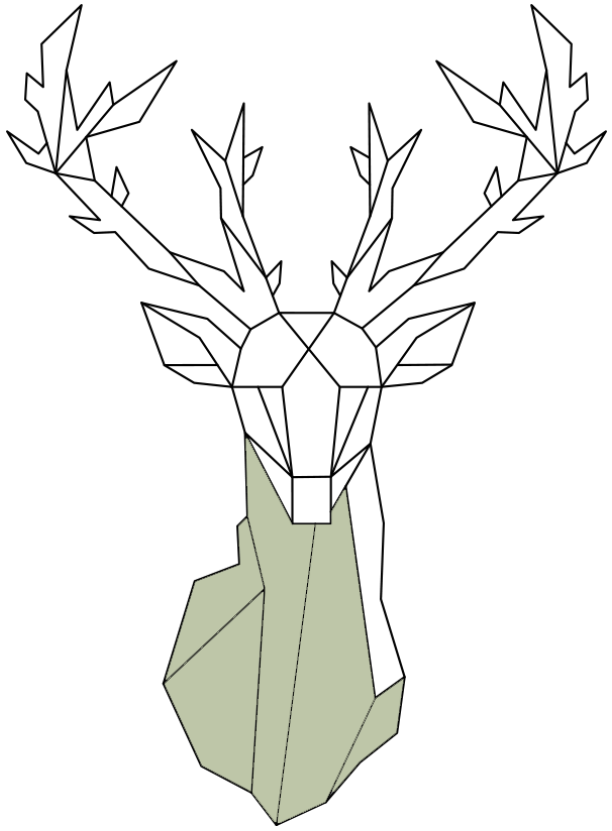
Data Quality Roundtrip 2025 in the Microsoft Data Platform

by Tillmann Eitelberg & Oliver Engels

Version 3.25 // 2025-01

© ppt by OH22

smart
casual
datadesign



WE CREATE SMART AND CASUAL DATADESIGN. EVERY DAY.

© ppt by OH22

THANK YOU

Data Community Austria Day 2025

Awesome Partner

Microsoft

Platinum

redgate

ACP IT for innovators.

Gold

b.telligent
smart data. smart decisions.

Lucient

Bronze

Tabular Editor

paiqo
The Platform & AI Company

Power BI Camp
www.linearis.at

smart
casual
datadesign





OLIVER ENGELS

CEO oh22data AG

/oengels @oengels @oengels.bsky.social



TILLMANN EITELBERG

CEO oh22information services GmbH

/tillmanneitelberg @teitelberg.bsky.social @teitelberg.mastodon.social

© ppt by OH22

smart
casual
datadesign

Our plan for today

- A bit on theory
- Map to lakehouse landscapes
- What has the Microsoft landscape to offer?
- Some samples

© ppt by OH22

smart
casual
datadesign

DATA QUALITY

A bit of theory for the roundtrip

© ppt by OH22

smart
casual
datadesign

What is Data Quality?

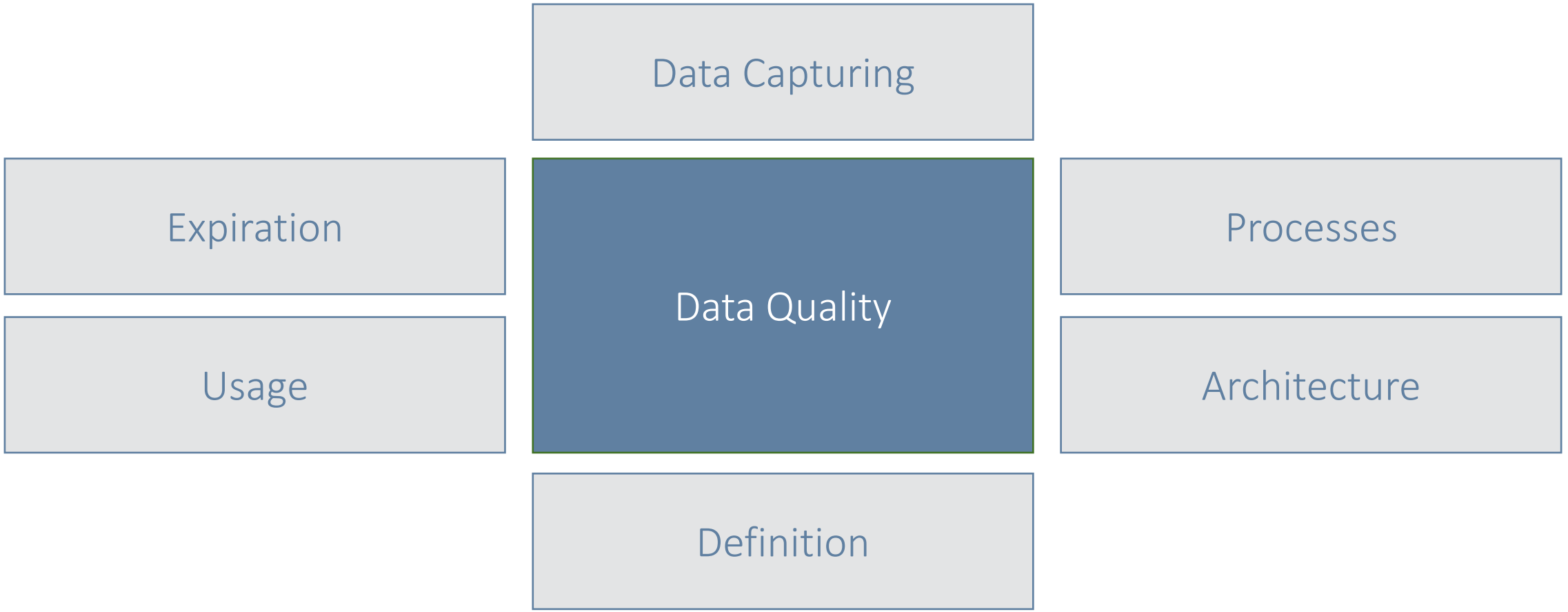
Data Quality represents the degree to which the data is suitable for business usages

Data Quality is built through:
People + Technology + Processes

Bad Data = Bad Business

© ppt by OH22

smart
casual
datadesign



smart casual datadesign

Dimensions of Data Quality

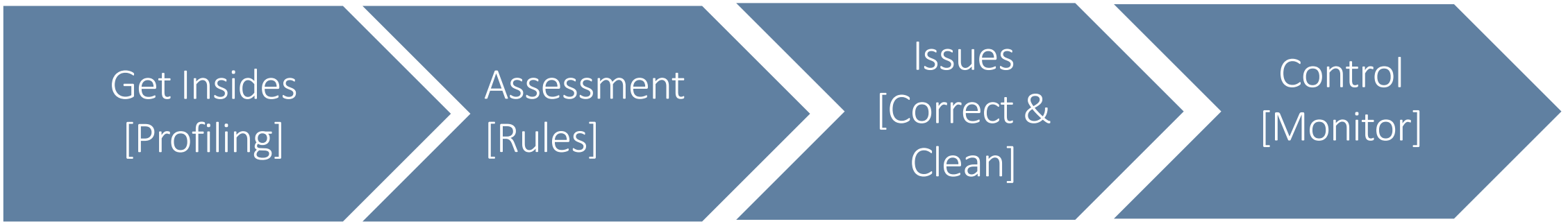
These dimension reflect the way to determine the quality of data

Accuracy	Consistency	Completeness	Timeliness	Validity	Uniqueness
Assesses the correctness of the data. Accurate data is free from errors and reflects the true values of the measured or observed attributes.	Examines the uniformity and agreement of data across different sources. Inconsistent data may have variations in format, values, or definitions.	Measures the extent to which data is whole, without missing values. It ensures that all required data elements are present	Assesses whether the data is up-to-date and relevant for the intended purpose. Outdated information may lose its value or relevance over time	Ensures that data conforms to predefined business rules and constraints. Valid data adheres to the specified formats and ranges	Measures the extent to which data is whole, without missing values. It ensures that all required data elements are present

© ppt by OH22

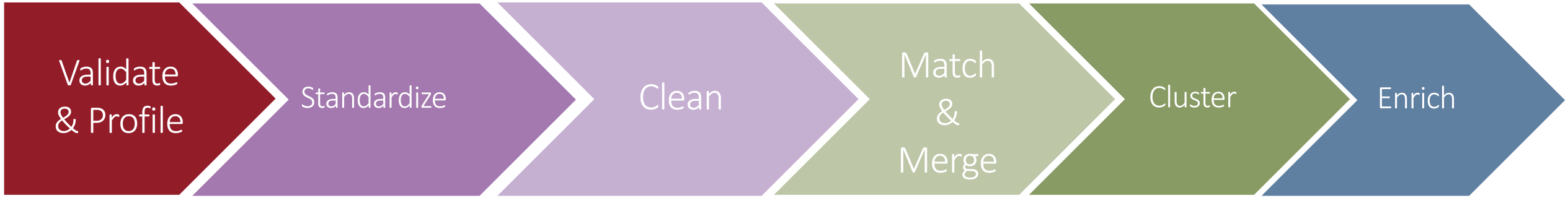
smart
casual
datadesign

Data Quality Process



smart
casual
datadesign

Data Quality Optimization Workflow



smart
casual
datadesign

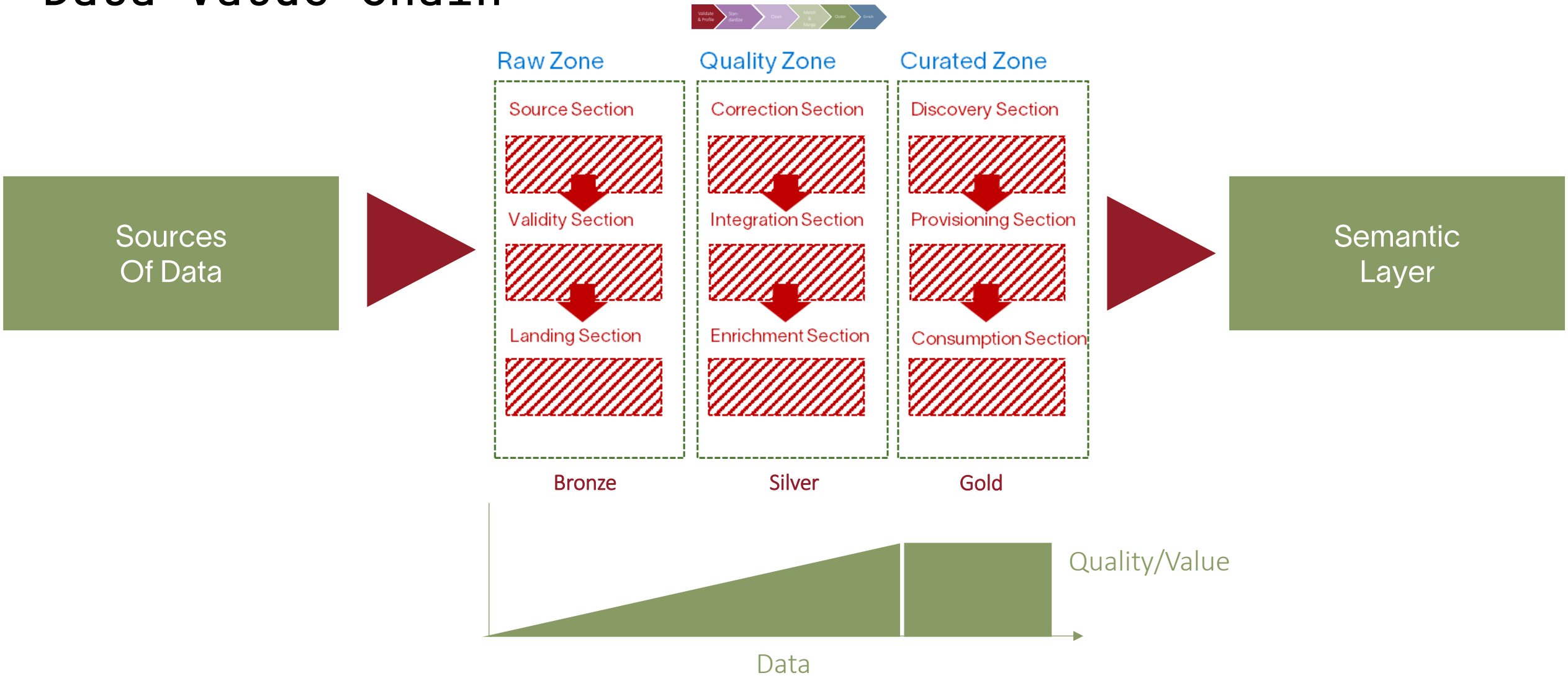
DATA QUALITY

In the age of Lakehouses

© ppt by OH22

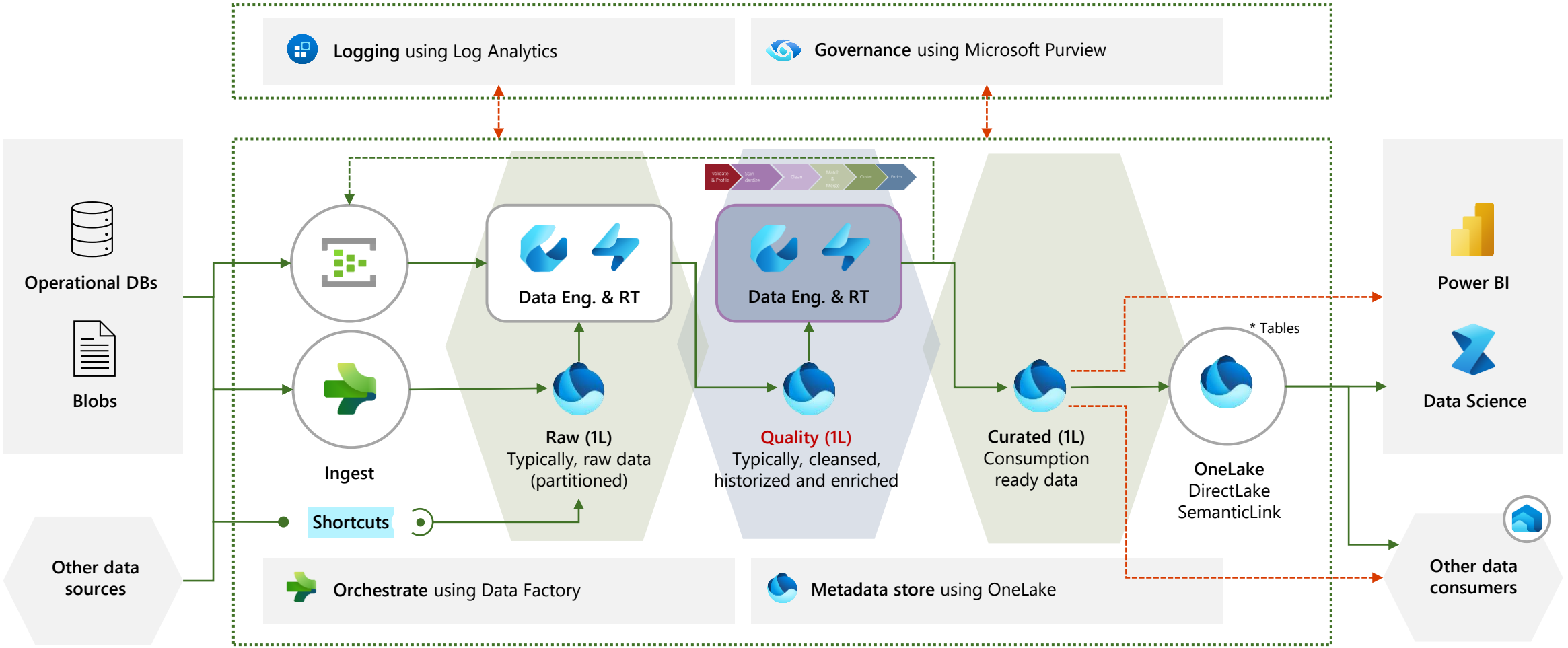
smart
casual
datadesign

Data Value Chain



→ DATA QUALITY IN THE FABRIC LANDSCAPE

smart casual datadesign



© ppt by OH22

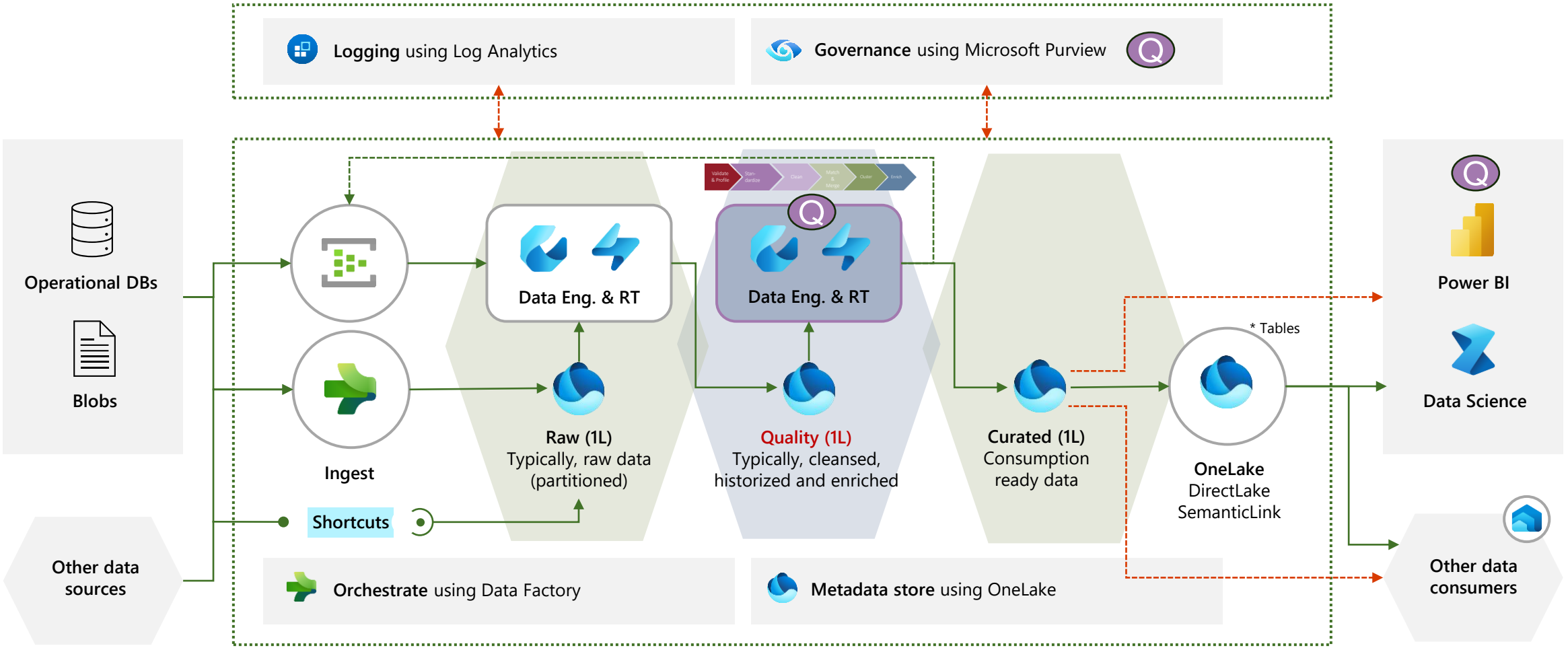
DATA QUALITY

Microsoft Solutions

© ppt by OH22

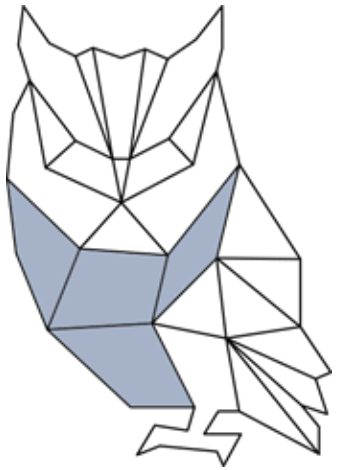
→ MICROSOFT DQ PORTFOLIO

smart casual datadesign



© ppt by OH22

smart
casual
datadesign



DQ POSSIBILITY DEMO'S

© ppt by OH22

smart casual datadesign

Microsoft & Data Quality

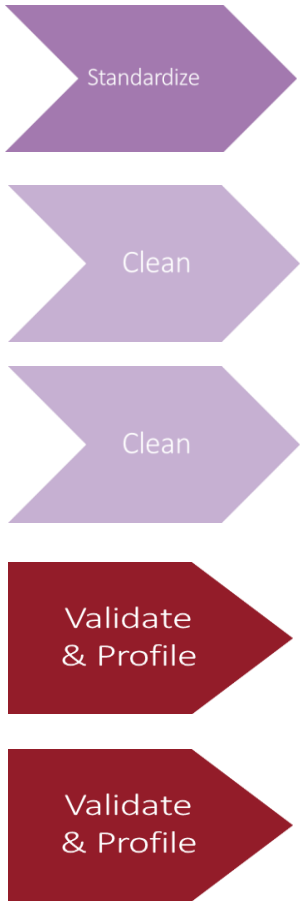
- Cloud Adoption Framework > Scenarios > Cloudscale Analytics
  - Govern your analytics estate
    - Data Quality Considerations, Recommendations and Solutions
      https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/govern-data-quality
    - Metrics

Metrics	Metrics Definitions
Completeness = % total of non-nulls + nonblanks	Measures data availability, fields in the dataset that aren't empty, and default values that were changed. For example, if a record includes 01/01/1900 as a date of birth, it's highly likely that the field was never populated.
Uniqueness = % of nonduplicate values	Measures distinct values in a given column compared to the number of rows in the table. For example, given four distinct color values (red, blue, yellow, and green) in a table with five rows, that field is 80% (or 4/5) unique.
Consistency = % of data having patterns	Measures compliance within a given column to its expected data type or format. For example, an email field containing formatted email addresses, or a name field with numeric values.
Validity = % of reference matching	Measures successful data matching to its domain reference set. For example, given a *country/region* field (complying with taxonomy values) in a transactional records system, the value of "US of A" isn't valid.
Accuracy = % of unaltered values	Measures successful reproduction of the intended values across multiple systems. For example, if an invoice itemizes a SKU and extended price that differs from the original order, the invoice line item is inaccurate.
Linkage = % of well-integrated data	Measures successful association to its companion reference details in another system. For example, if an invoice itemizes an incorrect SKU or product description, the invoice line item isn't linkable.

© ppt by OH22

smart
casual
datadesign

Possibilities in 2025?



Standardize, Versionize, Control with LakeFS

Clean with Data Wrangler

Anonymize with Presidio

Data Quality – Microsoft Purview

Data Quality – Databricks & Unity Catalog

© ppt by OH22

smart
casual
datadesign

DATA QUALITY

With LakeFS

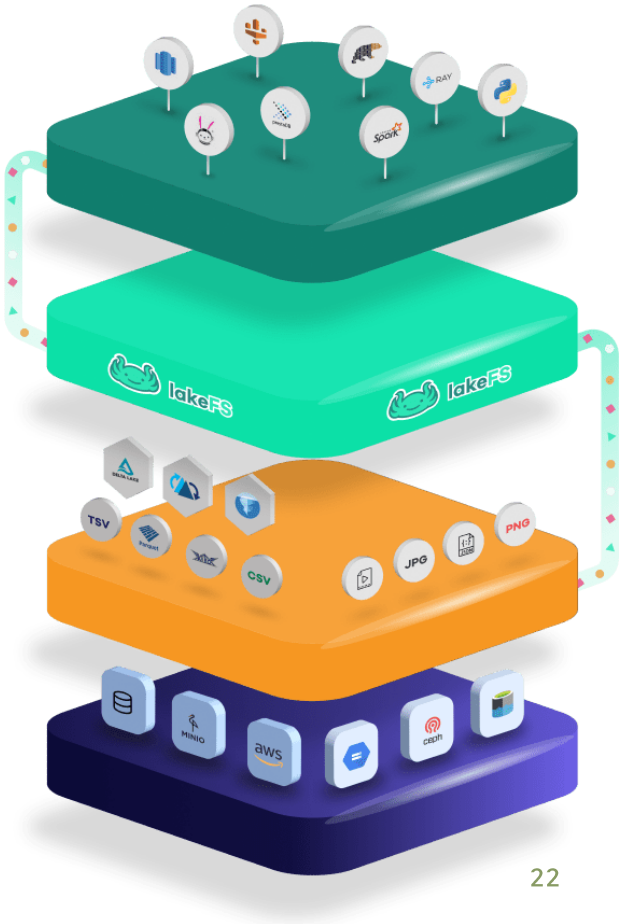
© ppt by OH22

smart casual datadesign

lakeFS - Git-like semantics on top of data lakes

lakeFS

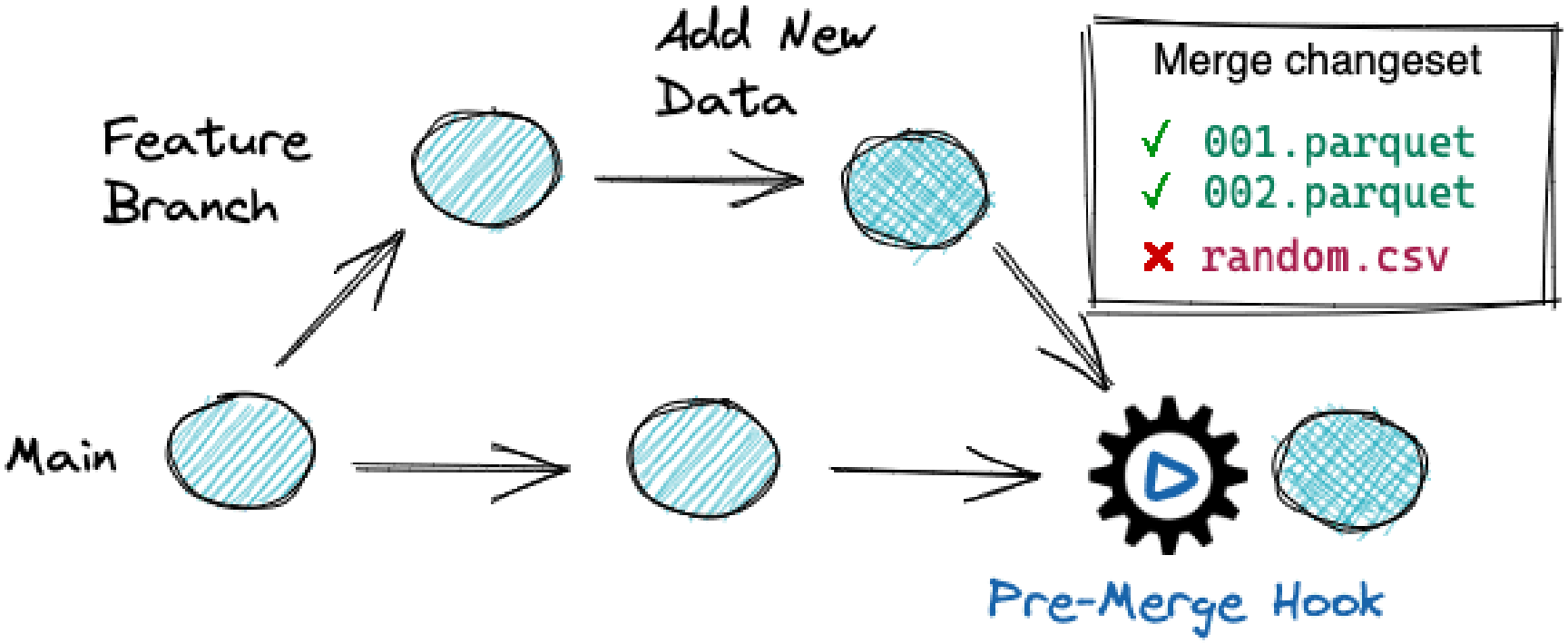
- lakeFS is an open-source data versioning and management tool for data lakes
- It provides Git-like capabilities for data, enabling version control, branch management, and atomic change
- Supports large-scale, distributed data lakes, including S3, ADLS, Google Cloud Storage, and on-premise object stores (e.g., MinIO)
- Integrates seamlessly with data processing tools like Apache Spark, Databricks, Azure Synapse, Presto, Hive or Snowflake
- Key features include branching, rollback, reproducibility of data pipelines, and automated CI/CD workflows for data



© ppt by OH22

smart casual datadesign

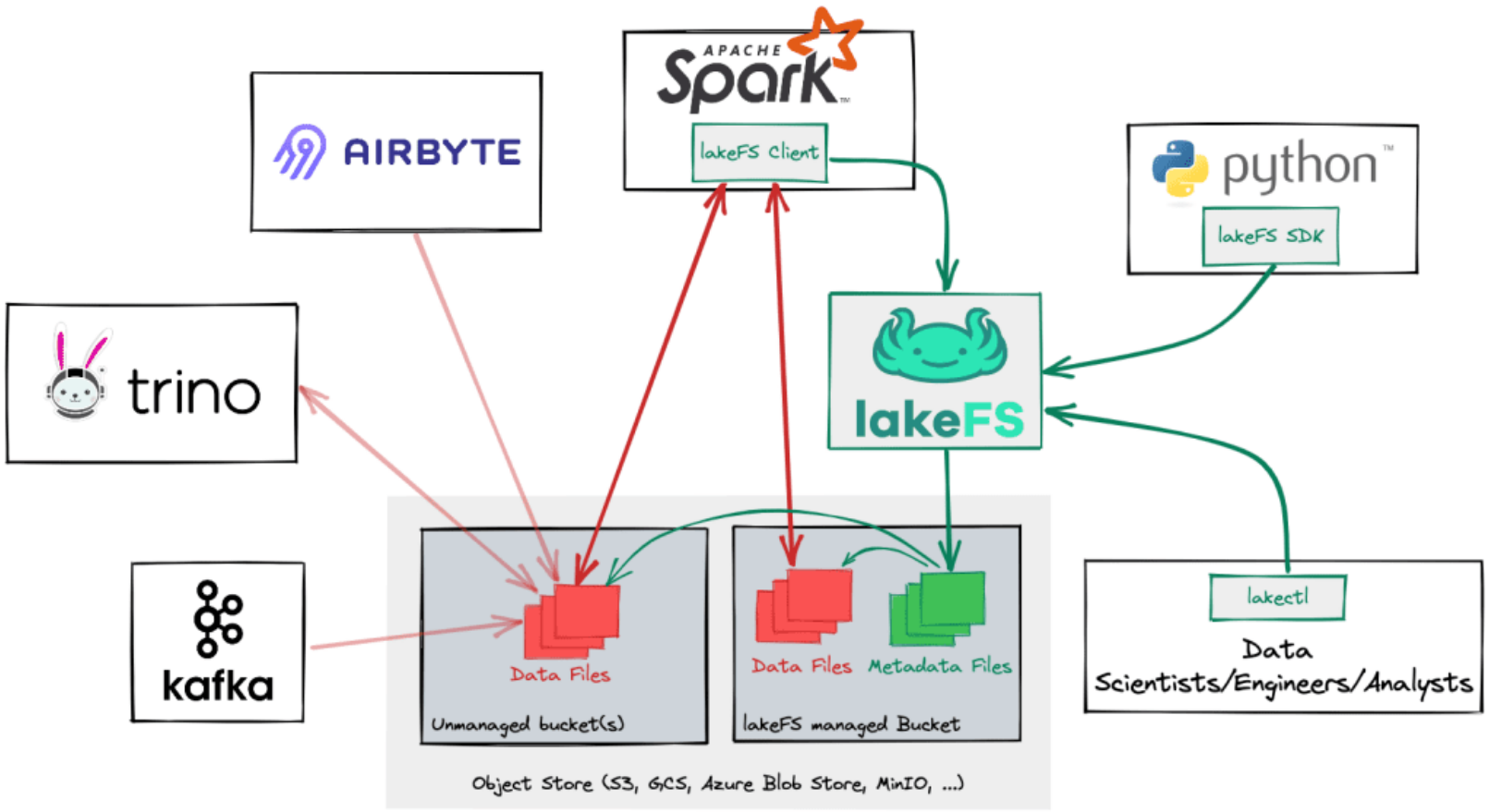
LakeFS



© ppt by OH22

smart
casual
datadesign

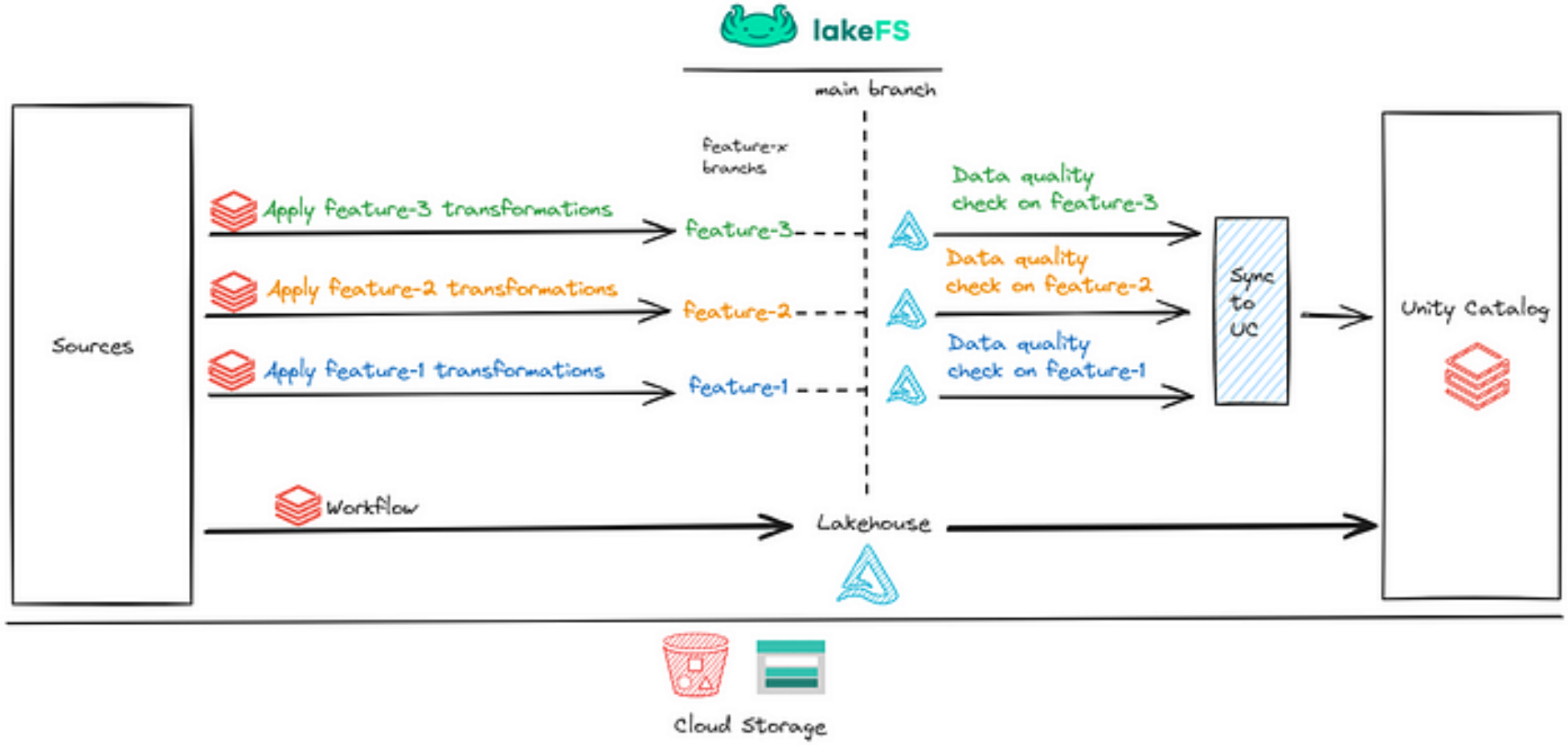
LakeFS



© ppt by OH22

smart
casual
datadesign

LakeFS



© ppt by OH22

smart casual datadesign

How can lakeFS help to improve data qua



- Data branching and versioning – track changes to repositories or datasets
- Design and test modifications in isolation before they become part of production data
- Guarantees consistent datasets for debugging and audits
- Quick recovery from errors by rollback to a previous version
- Prevents quality issues from affecting production by working in separate branches
- Commit changes to get a complete snapshot of a repository at a specific moment in time
- Hooks – actions to be triggered when particular events occour

© ppt by OH22

smart
casual
datadesign

DATA QUALITY

With Data Wrangler

© ppt by OH22

smart
casual
datadesign

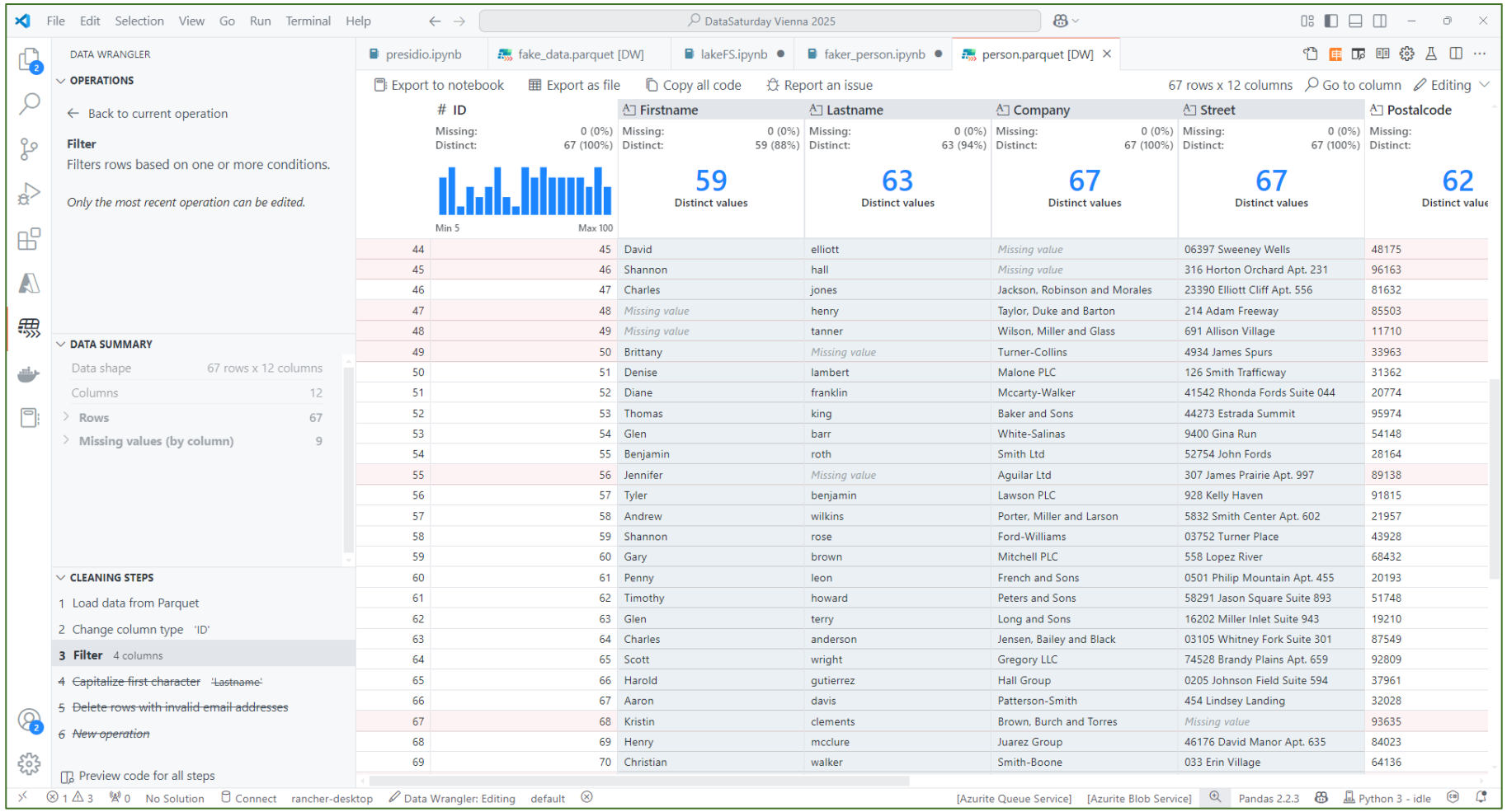
Data Wrangler 

- interactively inspect and analyze DataFrames
- providing a preview of the data and insights into its structure and statistics
- user-friendly graphical interface for applying common transformations (e.g., filtering, grouping, merging)
- automatically generates Python code
- Fully integrates with Python scripts and Jupyter Notebooks in VSCode

© ppt by OH22

smart casual datadesign

Data Wrangler

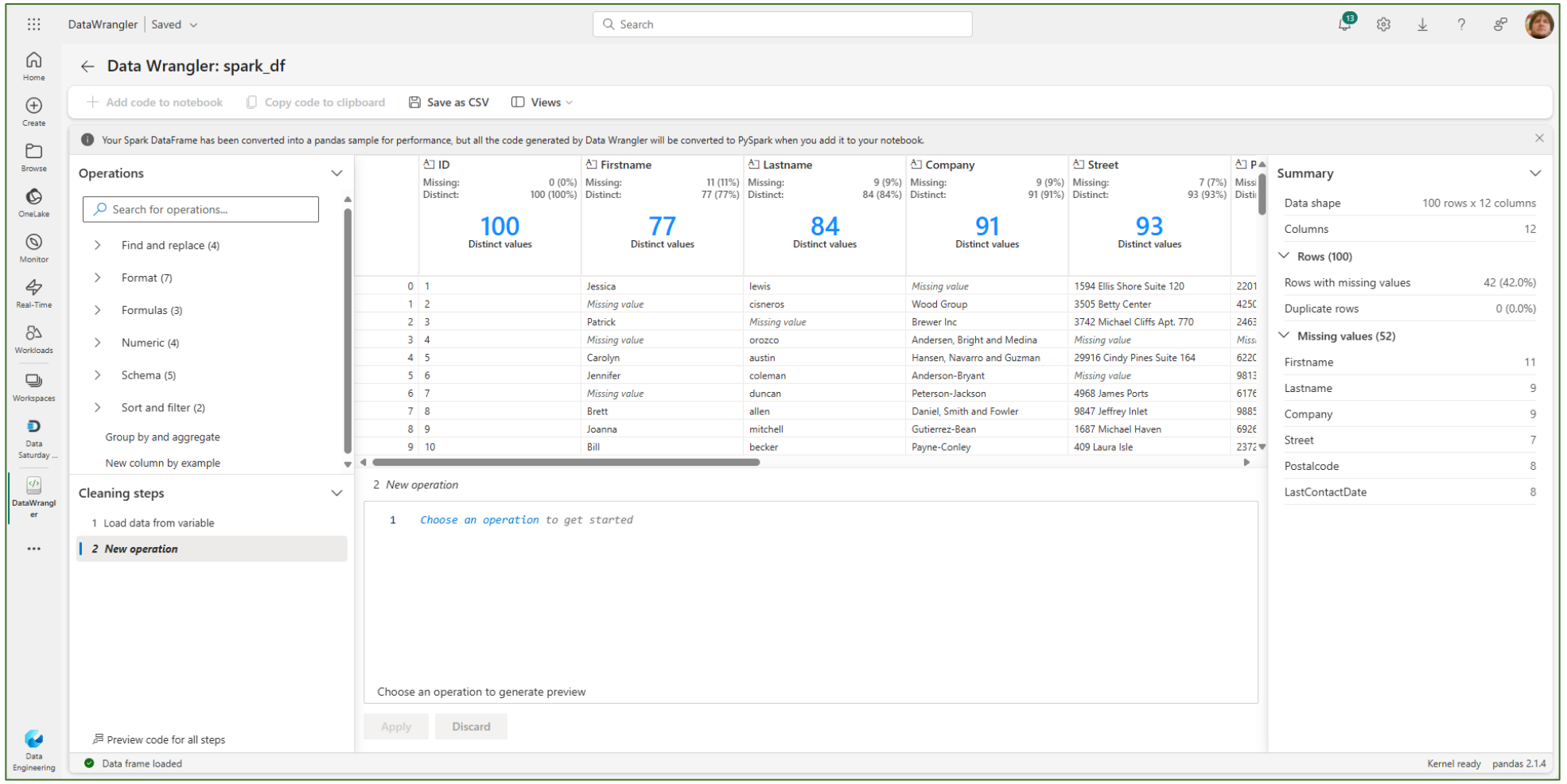


#	ID	Firstname	Lastname	Company	Street	Postalcode
44	45	David	elliott	*Missing value*	06397 Sweeney Wells	48175
45	46	Shannon	hall	*Missing value*	316 Horton Orchard Apt. 231	96163
46	47	Charles	jones	Jackson, Robinson and Morales	23390 Elliott Cliff Apt. 556	81632
47	48	*Missing value*	henry	Taylor, Duke and Barton	214 Adam Freeway	85503
48	49	*Missing value*	tanner	Wilson, Miller and Glass	691 Allison Village	11710
49	50	Brittany	*Missing value*	Turner-Collins	4934 James Spurs	33963
50	51	Denise	lambert	Malone PLC	126 Smith Trafficway	31362
51	52	Diane	franklin	Mccarty-Walker	41542 Rhonda Fords Suite 044	20774
52	53	Thomas	king	Baker and Sons	44273 Estrada Summit	95974
53	54	Glen	barr	White-Salinas	9400 Gina Run	54148
54	55	Benjamin	roth	Smith Ltd	52754 John Fords	28164
55	56	Jennifer	*Missing value*	Aguilar Ltd	307 James Prairie Apt. 997	89138
56	57	Tyler	benjamin	Lawson PLC	928 Kelly Haven	91815
57	58	Andrew	wilkins	Porter, Miller and Larson	5832 Smith Center Apt. 602	21957
58	59	Shannon	rose	Ford-Williams	03752 Turner Place	43928
59	60	Gary	brown	Mitchell PLC	558 Lopez River	68432
60	61	Penny	leon	French and Sons	0501 Philip Mountain Apt. 455	20193
61	62	Timothy	howard	Peters and Sons	58291 Jason Square Suite 893	51748
62	63	Glen	terry	Long and Sons	16202 Miller Inlet Suite 943	19210
63	64	Charles	anderson	Jensen, Bailey and Black	03105 Whitney Fork Suite 301	87549
64	65	Scott	wright	Gregory LLC	74528 Brandy Plains Apt. 659	92809
65	66	Harold	gutierrez	Hall Group	0205 Johnson Field Suite 594	37961
66	67	Aaron	davis	Patterson-Smith	454 Lindsey Landing	32028
67	68	Kristin	clements	Brown, Burch and Torres	*Missing value*	93635
68	69	Henry	mcclure	Juarez Group	46176 David Manor Apt. 635	84023
69	70	Christian	walker	Smith-Boone	033 Erin Village	64136

© ppt by OH22

smart casual datadesign

Data Wrangler



© ppt by OH22

smart
casual
datadesign

Data Wrangler

PROBLEMS 4 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER AZURE POLYGLOT NOTEBOOK DATA WRANGLER

6 *Custom operation*

Please delete all rows with an invalid country. Use a library for the validation

```
1  import pycountry
2  # Function to validate country
3  def is_valid_country(country_name):
4      try:
5          return pycountry.countries.lookup(country_name) is not None
6      except LookupError:
7          return False
8  # Delete rows with invalid country
9  df = df[df['Country'].apply(is_valid_country)]
```

Previewing

Update Cancel

© ppt by OH22

smart
casual
datadesign

DATA QUALITY

With Presidio

© ppt by OH22

smart
casual
datadesign

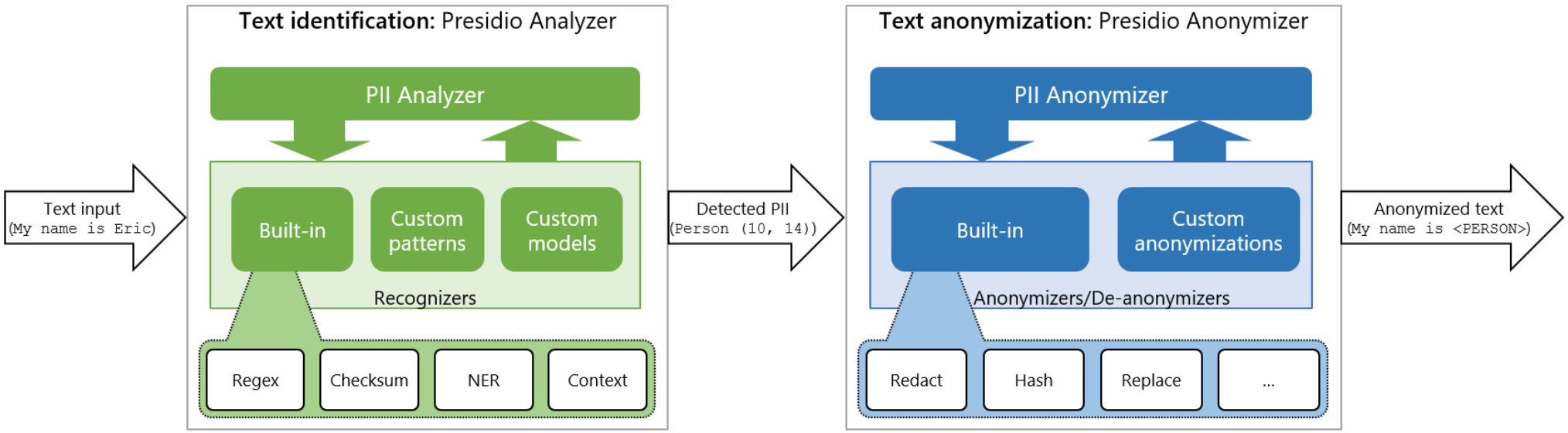
Presidio

- Open-source tool for detecting and anonymizing sensitive information (PII) in text and structured data
- Uses NLP to identify PII entities like names, emails, phone numbers, and allows adding custom recognizers
- Supports multiple anonymization techniques, such as masking, redaction, and replacement
- Enables large-scale processing of data via batch analysis and anonymization
- Supports multiple languages and can be integrated into data pipelines and APIs
- Presidio enhances the DQ cleaning process by detecting and anonymizing sensitive data (PII)

© ppt by OH22

smart
casual
datadesign

Presidio



© ppt by OH22

smart
casual
datadesign

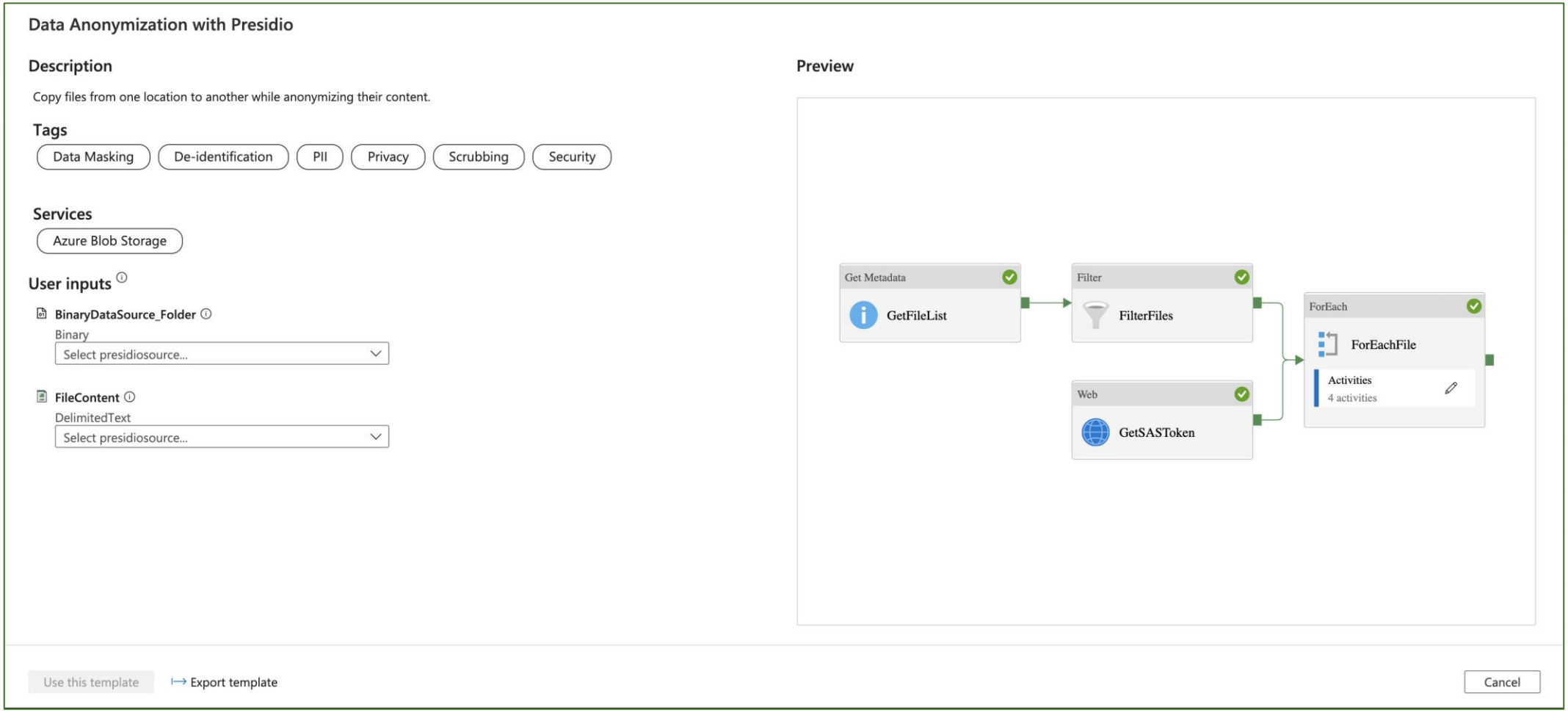
Presidio

- Use directly in Python scripts or applications to process text or structured data programmatically
- RESTful service using Docker containers, enabling remote processing of sensitive data
- Specific batch engines to process large datasets
- Expandability with own custom recognizers for domain specific PII detection
- Embed Presidio into data engineering workflows (e.g., Apache Airflow, Azure Data Factory)
  - ETL using AppService/Databricks - Microsoft Presidio
  - presidio-data-factory-template-gallery · microsoft/presidio

© ppt by OH22

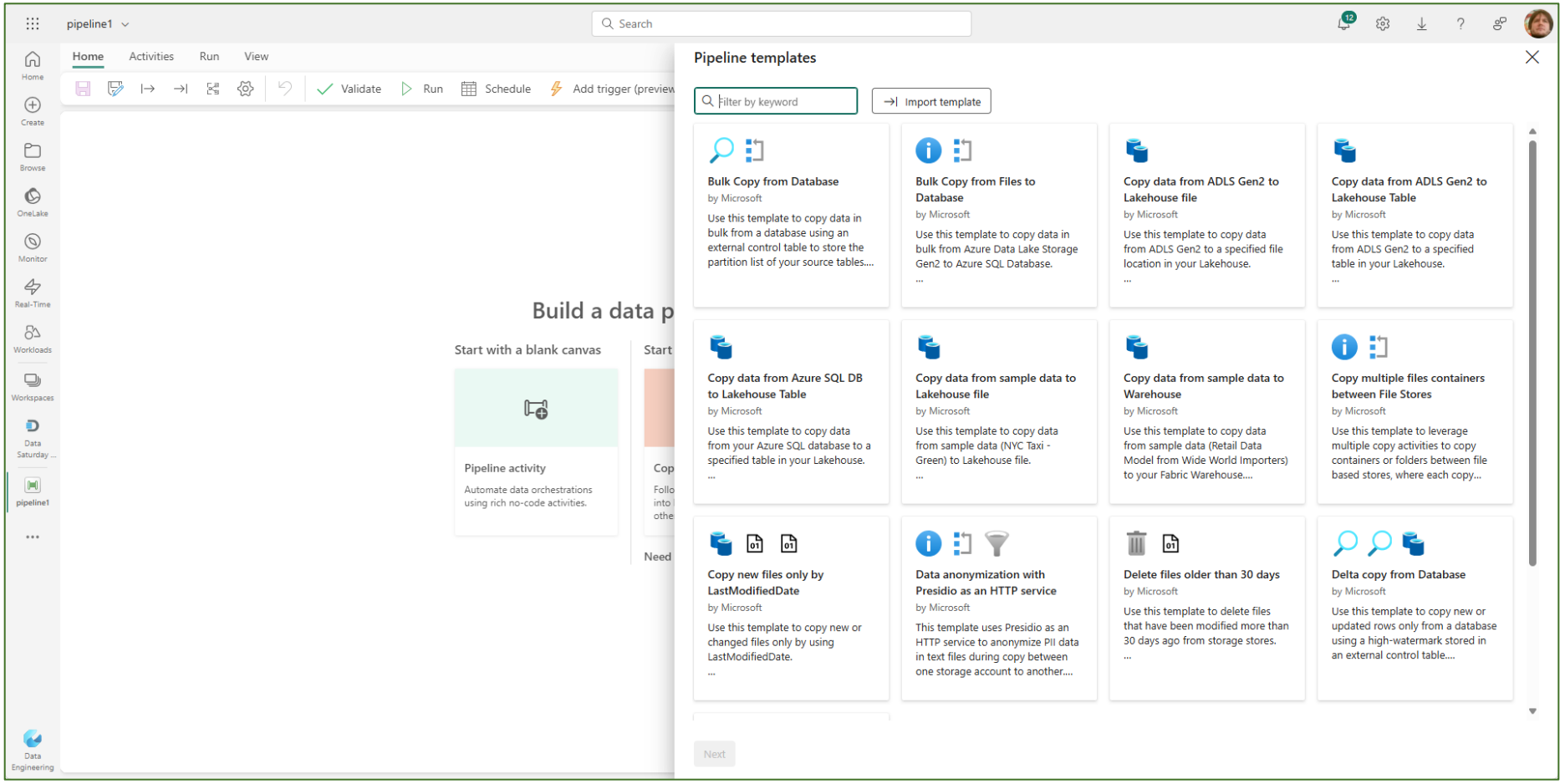
smart
casual
datadesign

Presidio



smart casual datadesign

Presidio



© ppt by OH22

smart
casual
datadesign

Presidio

Firstname	Lastname	Street	Postalcode	Email
<Person>	<Person>	3505 <LOCATION>	22013	<EMAIL_ADDRESS>
<Person>	<Person>	52 <LOCATION>	24632	<EMAIL_ADDRESS>
<Person>	<Person>	654 <LOCATION>	62204	<EMAIL_ADDRESS>
<Person>	<Person>	882 <LOCATION>	52362	<EMAIL_ADDRESS>
<Person>	<Person>	123 <LOCATION>	55587	<EMAIL_ADDRESS>

© ppt by OH22

smart
casual
datadesign

DATA QUALITY

With Microsoft Purview

© ppt by OH22

smart casual datadesign

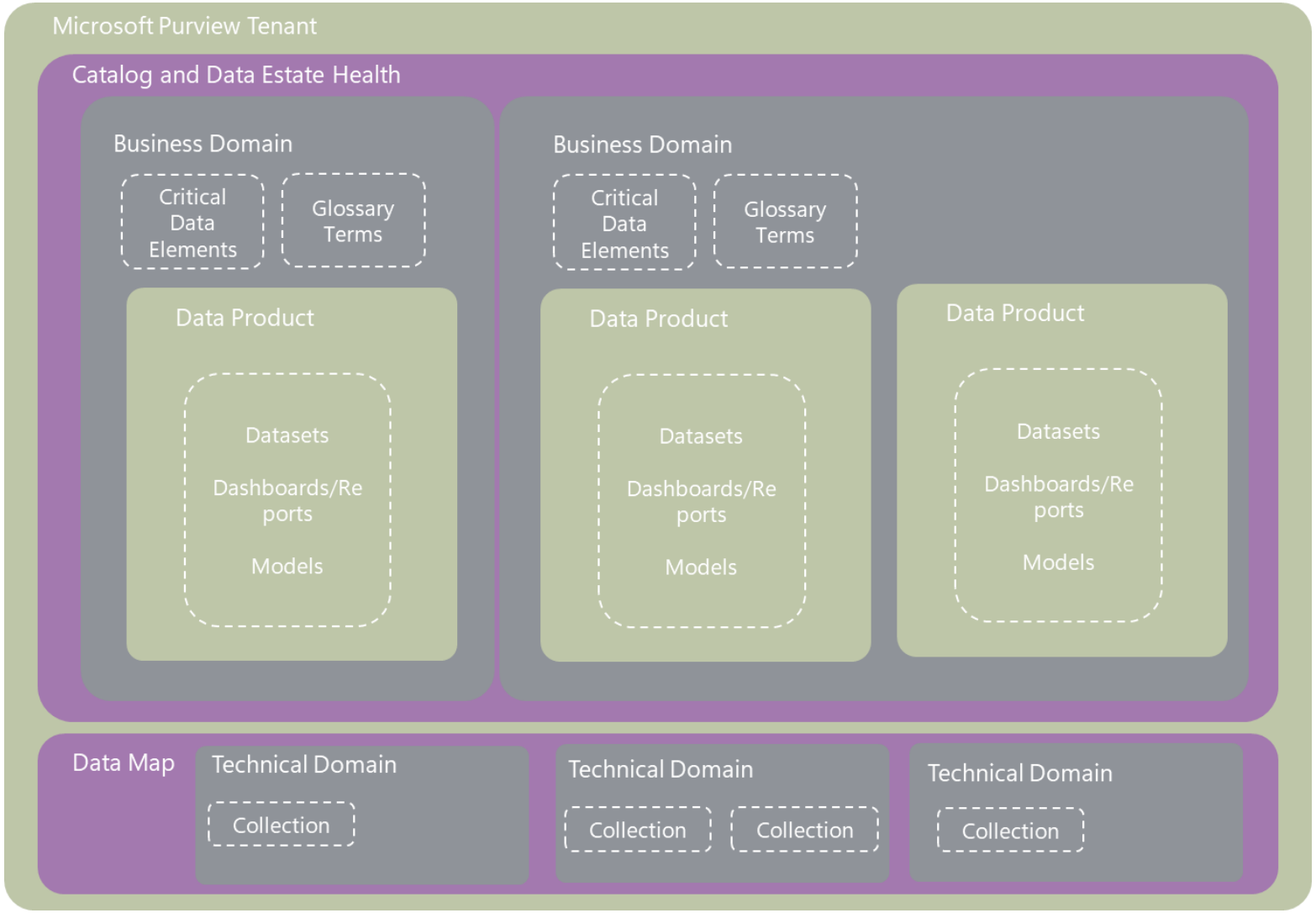
Microsoft Purview

Business domains are a new way of organizing your data estate. You want and need to access your entire data estate, but a single, uncategorized list is overwhelming. A business domain is a boundary that aligns your data estate to your organization; think of it as a mini catalog inside your data catalog.

Critical data elements are a logical grouping of important pieces of information across your data estate. For example: A "Customer ID" critical data element can map "CustID" from one table and "CID" from another table into the same logical container. These groupings can make data easier to understand as well as promoting standardization. Data quality rules and access policies can be attached to these elements to further secure sensitive information across your data estate.

Glossary terms provide a vocabulary for business users. It consists of business terms that can be related to each other and allows them to be categorized so that they can be understood in different contexts. This helps in abstracting the technical jargon associated with the data repositories and allows the business user to discover and work with data in the vocabulary that is more familiar to them.

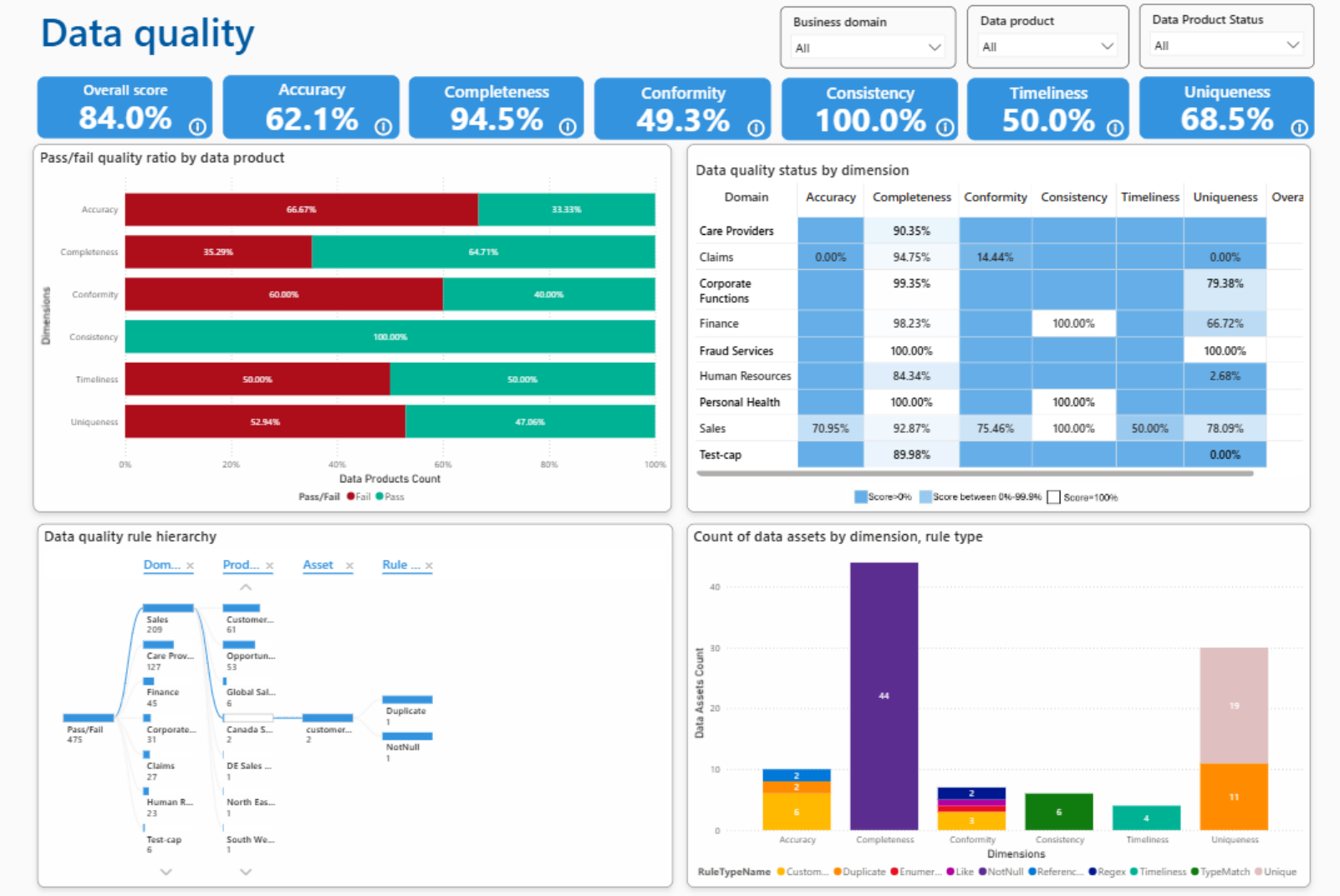
A **data product** is a group of data assets (tables, files, PBI reports, etc.), a packaged offering to an enterprise that provides assets with a use case to be shared to data consumers. A business domain can house many data products, but a data product is managed by a single business domain and can be discovered across many domains.



© ppt by OH22

smart
casual
datadesign

Microsoft Purview DQ Healthmonitor

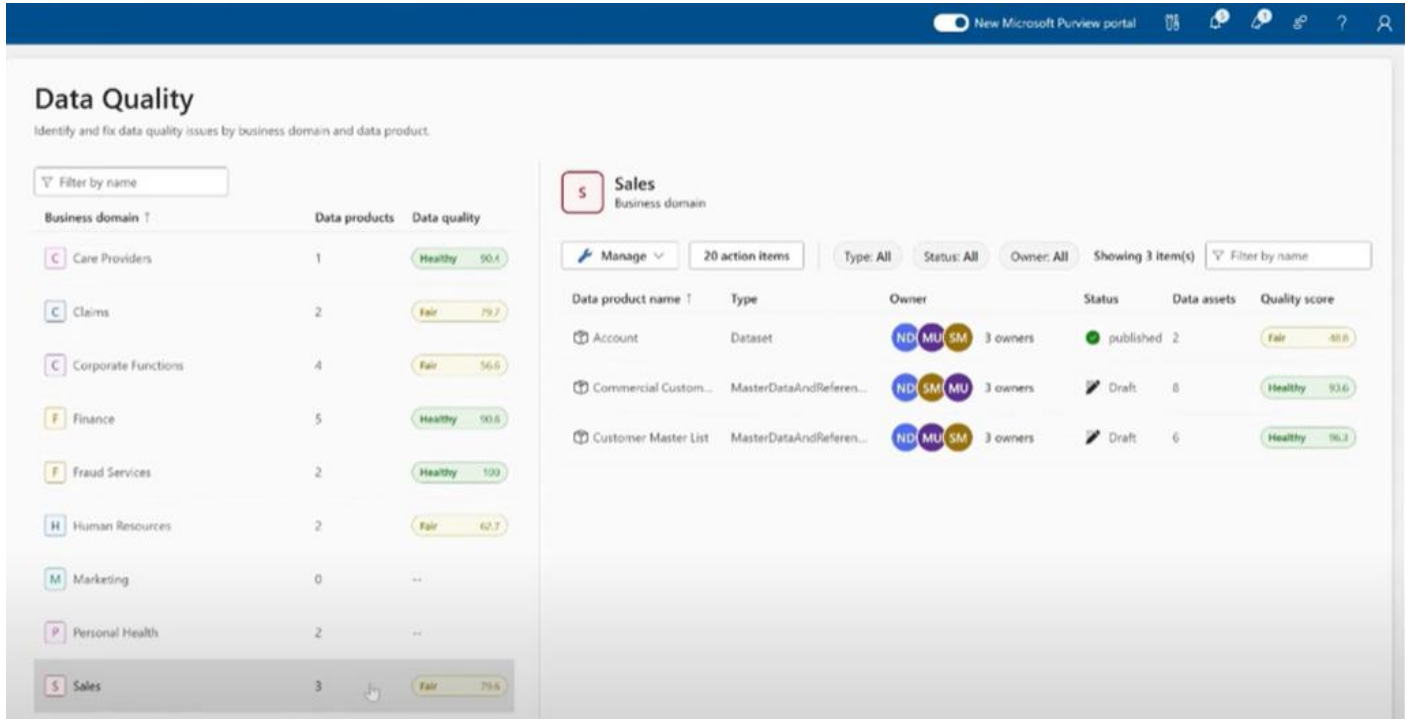


© ppt by OH22

smart
casual
datadesign

Microsoft Purview

- Rules are applied at the column level and aggregated to provide scores at the levels of data assets, data products, and business domains
- Ensures end-to-end visibility of data quality within each domain
- Preconfigured with meta model for data estate
- Provides scores and trends for data quality
- Allows drill down into rules and scans



© ppt by OH22

smart
casual
datadesign

DATA QUALITY

With Azure Databricks

© ppt by OH22

Data Quality @ Azure Databricks

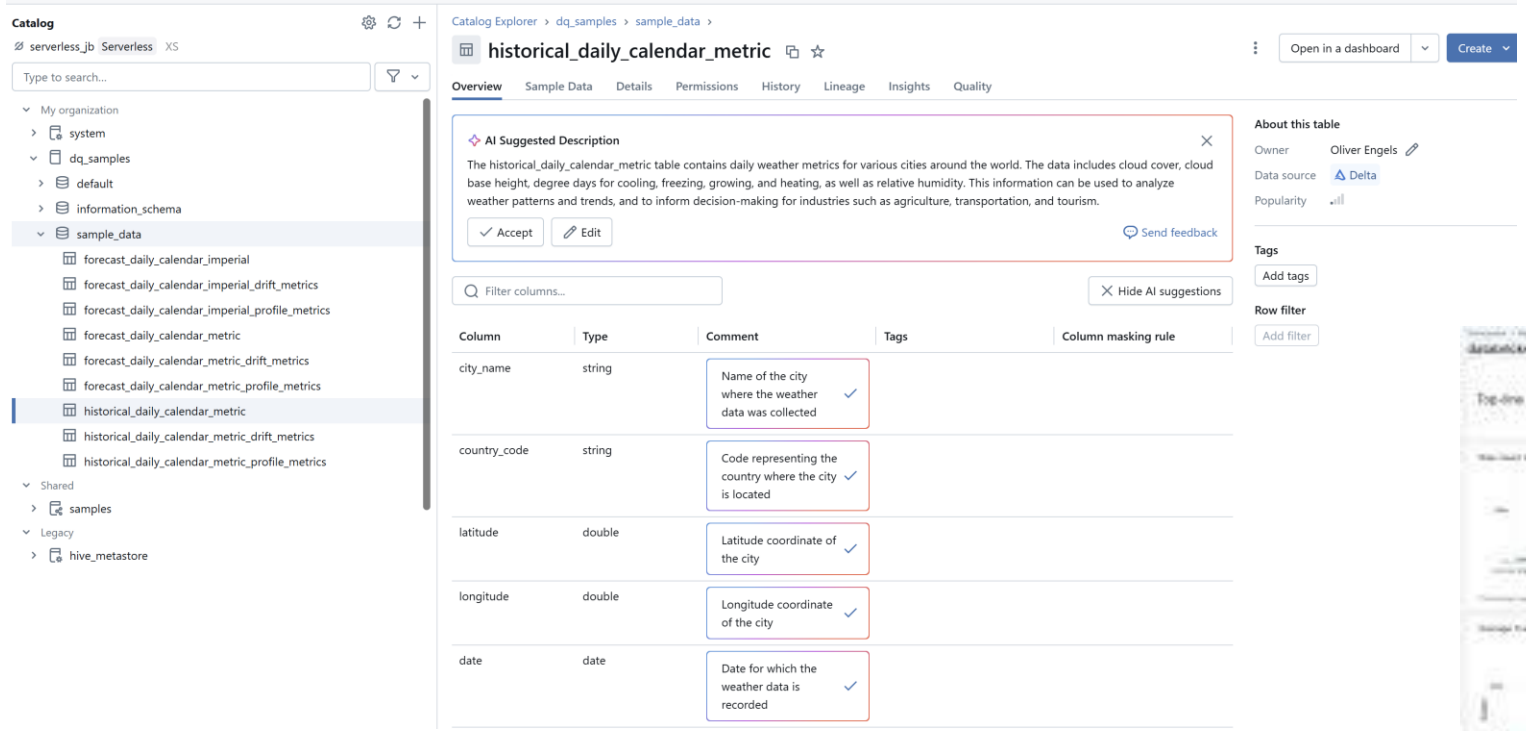
- Databricks providing best practise blogs on Data Quality:
- https://www.databricks.com/discover/pages/data-quality-management



© ppt by OH22

smart casual datadesign

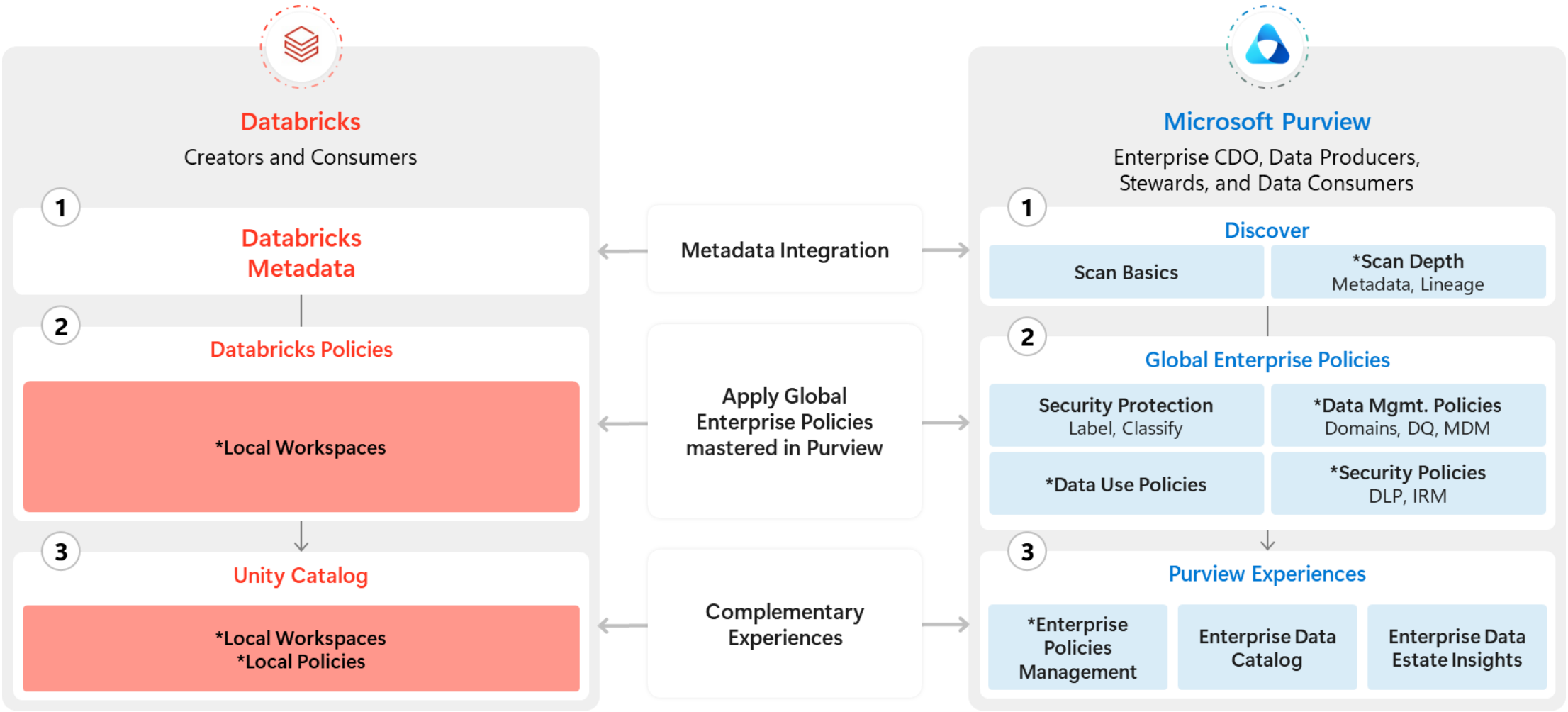
Data Quality @ Azure Databricks Unity Catalog





smart casual datadesign

Data Quality @ DBX & Purview



© ppt by OH22

smart
casual
datadesign

Future DQ Landscape

- Still not a consistent approach to DQ from Microsoft
- Will Microsoft Purview help? → Roadmap?
- Strategy One:
  - Build your own DQ functionality
  - Flexible, but can be very complex
- Strategy Two:
  - Buy a third party tool
  - More or less flexible, can be very expensive
- Some samples

© ppt by OH22

smart
casual
datadesign

Evaluations, evaluations...



https://evals.datagrillen.com/evals_vienna.aspx

© ppt by OH22

→ THX for your audience

smart
casual
datadesign

Oliver Engels

CEO
oh22data AG

o.engels@oh22.net

Otto-Hahn-Str. 22
65520 Bad Camberg
Deutschland

↗ data.oh22.net

Tillmann Eitelberg

CEO
oh22information services GmbH

t.eitelberg@oh22.net

Kellerstr. 3
53772 Königswinter
Deutschland

↗ www.oh22.is