✓ data domain

Test Meetup

Non-Production Data Management

July 2024

Turning Your Data Vision Into Reality

"Vision Without Execution Is Hallucination" - Thomas Edison

Agenda

1. What is Non-Prod DM

2. Non-Prod Data Ecosystem

3. Data Governance & DM

4. Non-Prod Environment Risks

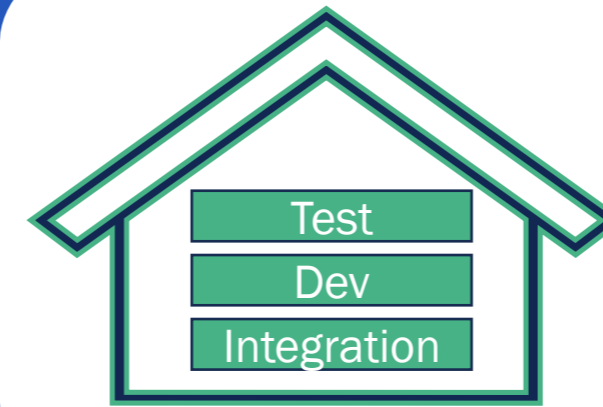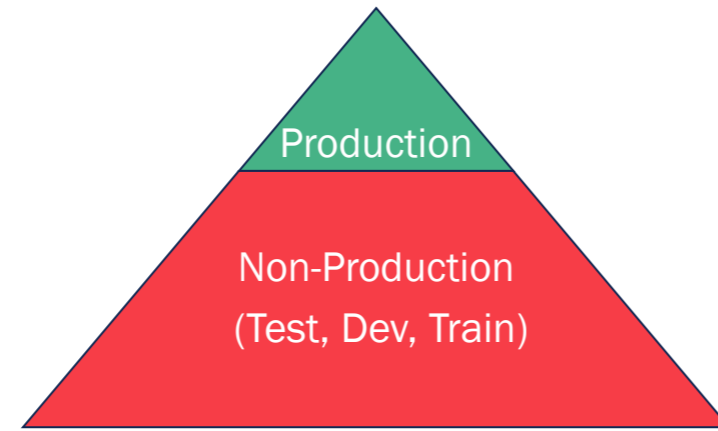5. Considerations for Non-Prod Env

6. Methodologies & Approaches

7. Summary

8. Discussions

✓data domain

What is Non-Production Data Management?

All non-production environments - not just for testing purposes.

Production

Non-Production (Test, Dev, Train)

Enterprise approach vs use case driven approach.

Test
Dev
Integration

Be of a standard where it complies with, and remains governed by, all internal and external security measures and policies.

UAT
Dev
Sandpit
Training
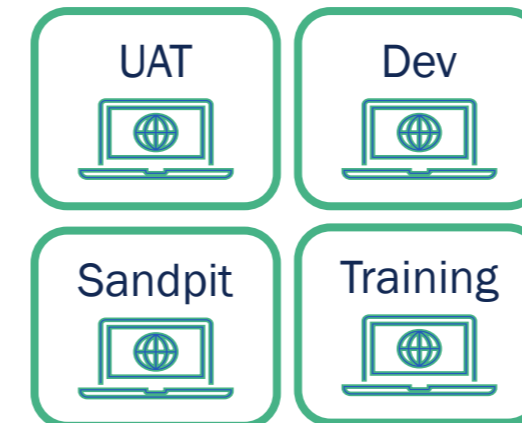
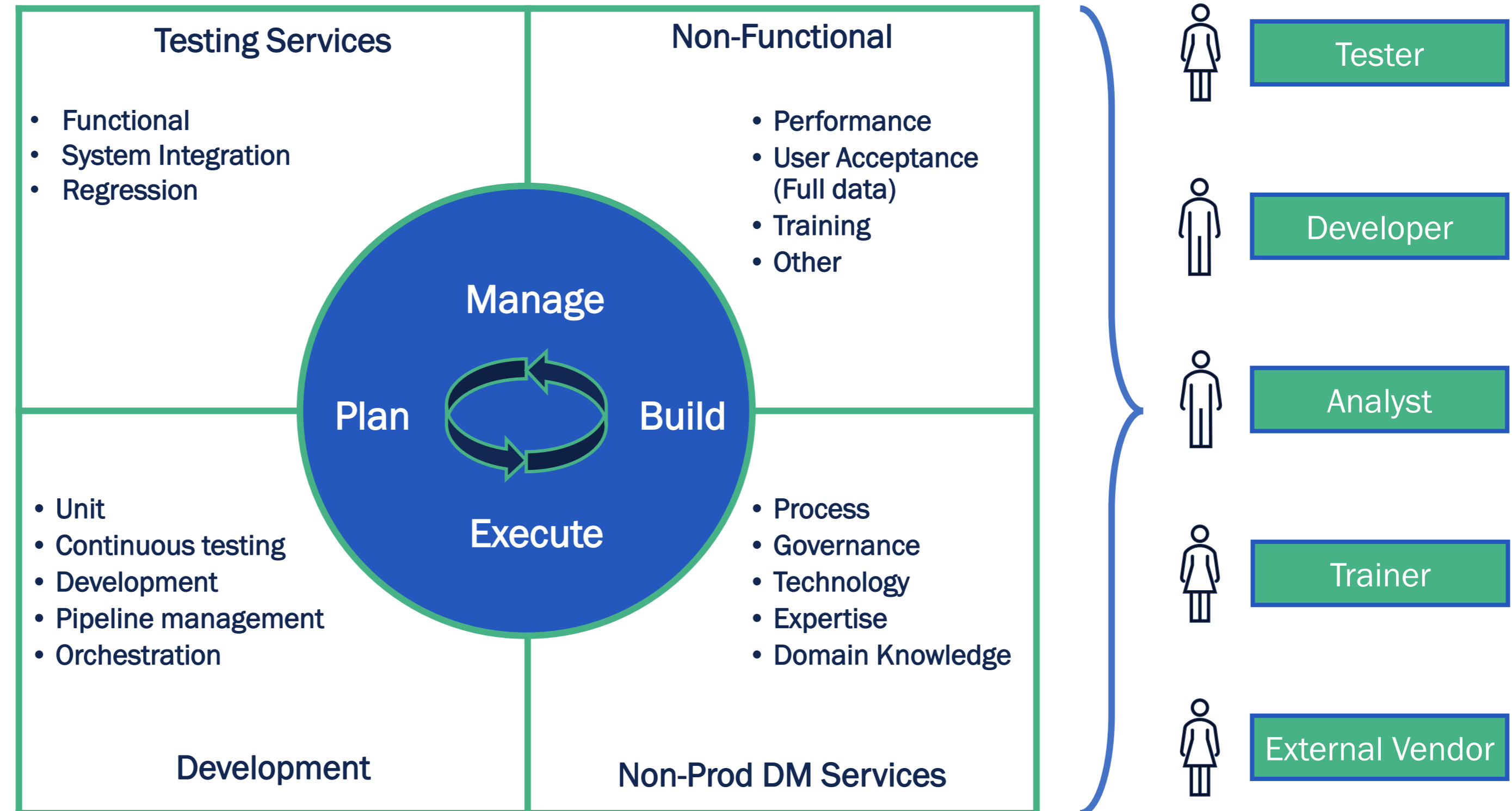Be of a standard that allows multiple use cases.

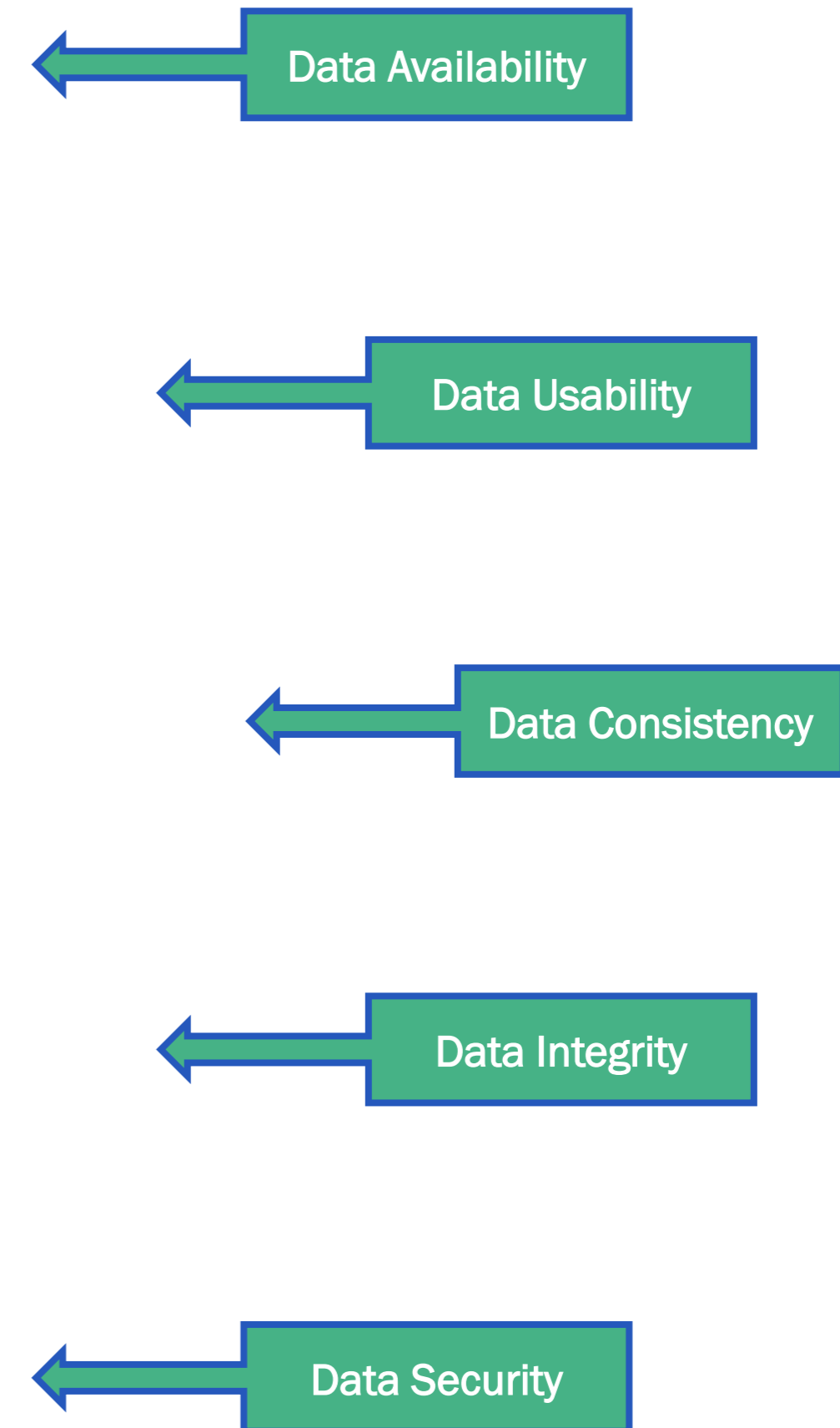Data and Analytics – Data *is* part of the solution.

Methods, techniques, practices for data anonymisation and masking.

Non-Production Data Ecosystem

- ▶ Non-Production environments differ.

- ▶ Vastly bigger than Production and grows at a faster pace.

- ▶ Often sub-standard security compared to Production.

- ▶ Data management of Non-Production environments typically means Production is catered for.

Testing Services

- Functional
- System Integration
- Regression

Non-Functional

- Performance
- User Acceptance (Full data)
- Training
- Other

Manage

Plan

Build

Execute

Development

- Unit
- Continuous testing
- Development
- Pipeline management
- Orchestration

Non-Prod DM Services

- Process
- Governance
- Technology
- Expertise
- Domain Knowledge

Tester

Developer

Analyst

Trainer

External Vendor

✓data domain

Data Governance & Data Management

data domain

Data Domain Confidential – Do Not Distribute
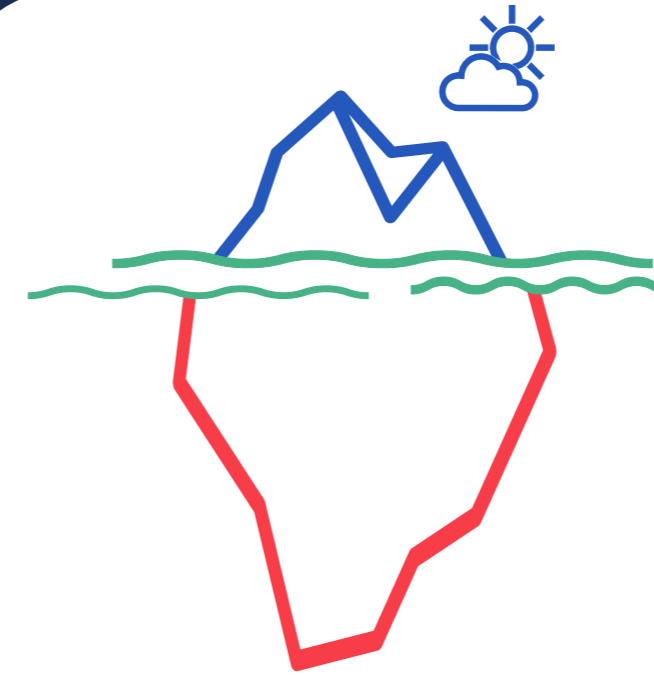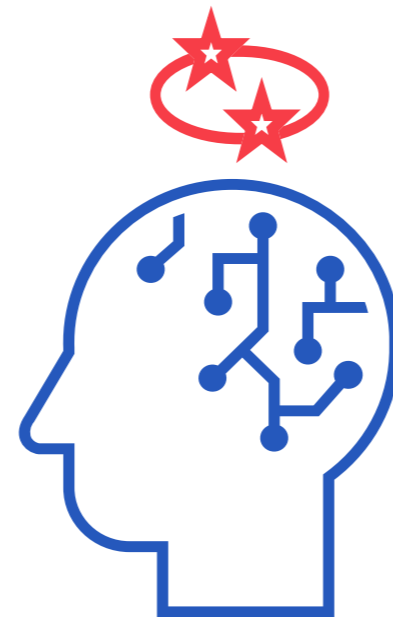
Non-Production Environment Risks

What happens if we get security wrong? Reputational damage, commercial damage, lawsuits, liability i.e. OCP compliance notices.
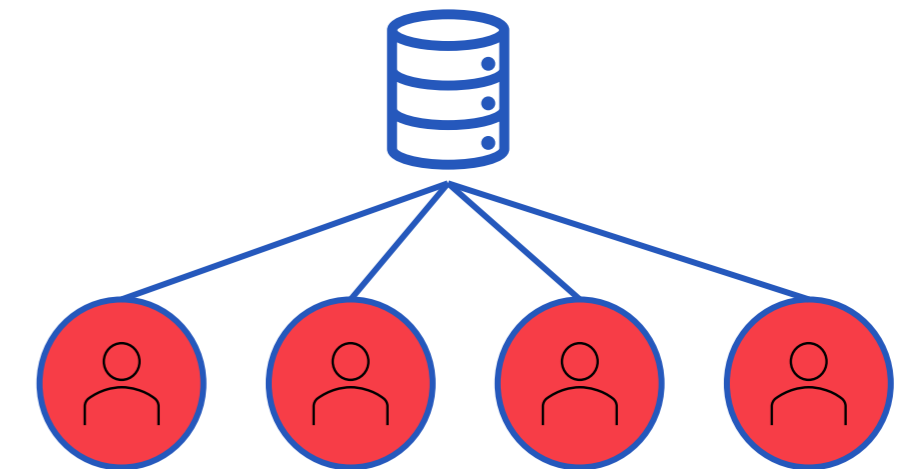
Non-Production data is a major hidden risk.

Testing becomes invalid, production fails, invalid decisions, AI Hallucinations.

Pushing down security responsibilities to individual users and user groups.

Considerations for Non-Production Environments

What
- What data security policy is already in place to provide guidance on the treatment of sensitive data?
- What technology is already in place and in what area i.e. data encryption, masking approaches and tools and at what layer (disk, database, data)?

Where
- Where can the data reside, is it limited to only on-premise? Can it exist in the cloud, hybrid cloud, data lake?
- Where is the data when being moved? Is it secured whilst being transferred?

Who
- Who executes the data security policies? Internal teams, external parties, DBAs, combination? Who owns and governs the data and environments? Is there a Data Steward, Security Administrator to provide guidance?
- Who needs to be involved with Non-Prod data management discussions? Do developers and development teams need to be taken into consideration?

How
- How is sensitive data currently being treated in all environments? Is it consistent? Does it need to be consistent (generally, it should)? If not, then why not?
- How do new technologies, tools, and AI make things easier (or harder)?

When
- Compliance by Design.
- Before the start of any project, otherwise development deadlines will encourage poor practices.

✓data domain

Methodologies & Approaches - Key Capabilities

Data Provisioning

Able to generate non-production data by using cloning, data virtualisation, synthetic generation, or sub-setting methods to produce and refresh large volumes of data that meet requirements.

Data Classification & Security

Capable of automatically identifying sensitive data and provide masking and obfuscation functionality.

Self-Service

Allows separation of the non-production data repository and controls access and privileges to the environment. Data in the non-production repository is managed, shared, and reused as required.

Monitoring & Observability

Able to ensure compliance with data governance norms by monitoring the usage and providing reports.

✓data domain

Data Provisioning

Copying / Cloning

Pros:
- Relatively simple to implement.
- Provides a complete set of data.

Cons:
- Copying - Expensive, time-consuming, not agile or scalable.
- Copying / Cloning - Completeness of the data can lead to miss-use of the data.
- Requires detailed data classification and security.

Synthetic Data Generation

Pros:
- Provides complete data anonymity and security.
- Generates scenarios not available in the data (e.g. new solutions).

Cons:
- Can be complicated to create and data usability can be compromised if the system is complex and there are many properties, patterns, and permutations within the data.
- Data integrity can still be lost as it no longer reflects the original data and retain proper context.
- Can "miss" real life scenarios.

Sub-Setting

Pros:
- Easier to create and generally less complex than a synthetic method.
- Aligns better with Agile principals and a popular and most accepted method of provisioning non-production data.
- Reduced processing time and deployable in multiple environments.

Cons:
- Referential integrity can be hard to manage in federated environment scenarios and across multiple sources and platforms.
- Limited breadth of conditions normally available within a full Production set thus risks missing test scenarios.

Data Virtualisation

Pros:
- Fast to setup, rapid data refresh, less support required
- No physical storage means less expensive to implement and maintain.

Cons:
- The virtualised environment can contain vast volumes of data and can be time consuming to prepare and execute.
- Tied to Production sources.

✓data domain

# Data Classification & Security: Simple Masking Approaches
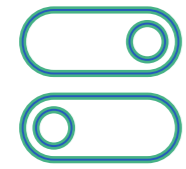
Substitution

Complete replacement of value.

Pros: Values remain realistic and behave in the same way as real data.

Cons: Someone with knowledge of the data may still be able to identify the attribute when used in conjunction with other values.

Example: John Smith is used for the customer's name instead of the real name.
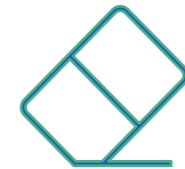
Nulling/Deleting

Nulls or deletes the value in the field.

Pros: Complete masking.

Cons: Data integrity is completely lost.

Example: Customer first name is a null field.

Fixed Masking

Partial or complete replacement of values.

Pros: Retains data integrity.

Cons: Limits aspects around data usability.

Example: 5595 XXXX XXXX 1234, X is used to mask the card number.
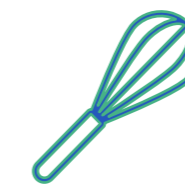
Scrambling

Data within the field is randomly scrambled.

Pros: Retains data integrity.

Cons: More effective on numeric fields.

Example: Customer number 12345 becomes 42315.

Variance

Sets a range variance to numeric or date fields.

Pros: Retains data integrity.

Cons: Only usable on date and numeric fields.

Example: 1000 with a +/-100 range is converted to a number between 900 – 1100.

✓data domain

Data Classification & Security: Advanced Masking Approaches
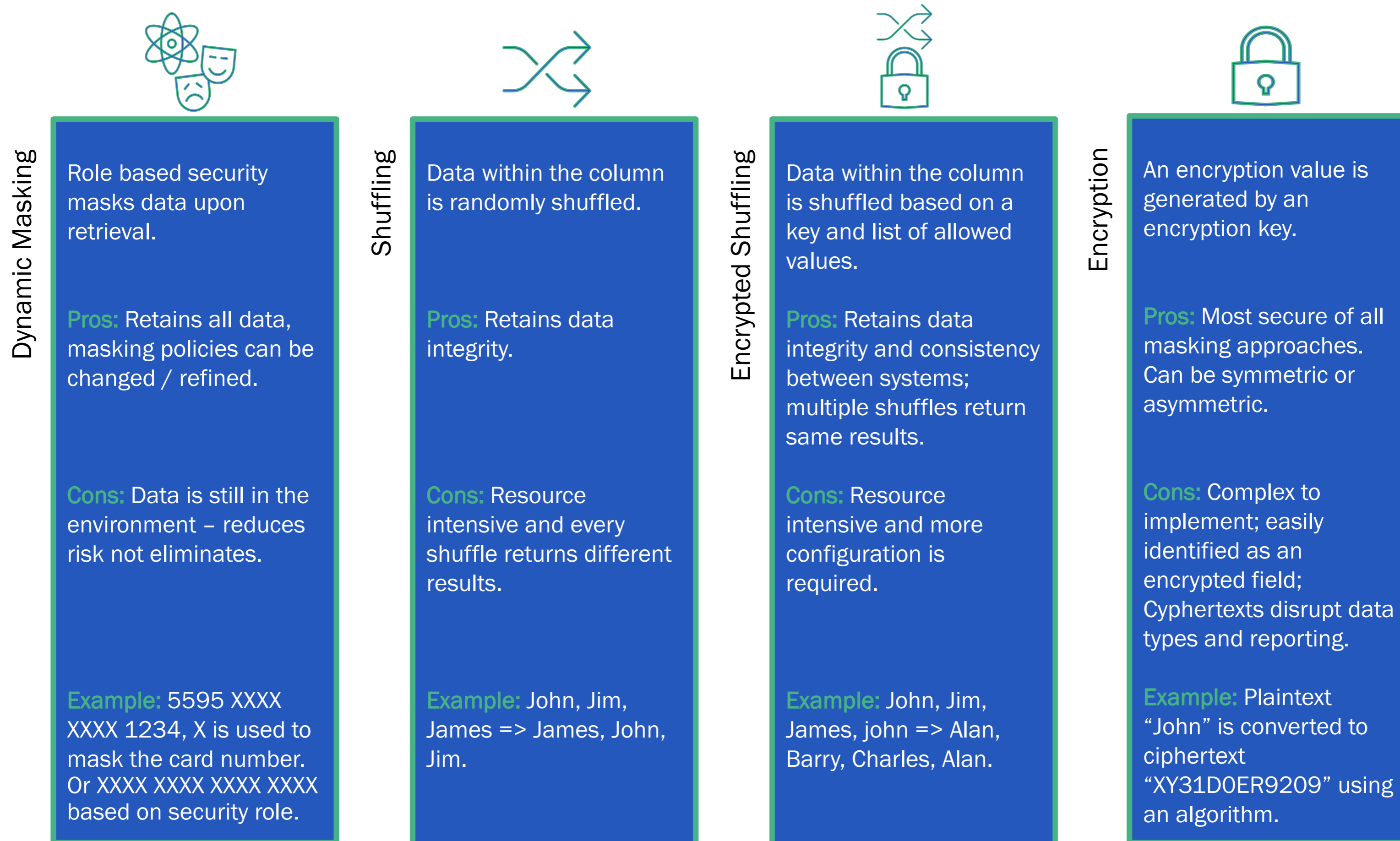
Dynamic Masking

Role based security masks data upon retrieval.

Pros: Retains all data, masking policies can be changed / refined.

Cons: Data is still in the environment – reduces risk not eliminates.

Example: 5595 XXXX XXXX 1234, X is used to mask the card number. Or XXXX XXXX XXXX XXXX based on security role.

Shuffling

Data within the column is randomly shuffled.

Pros: Retains data integrity.

Cons: Resource intensive and every shuffle returns different results.

Example: John, Jim, James => James, John, Jim.

Encrypted Shuffling

Data within the column is shuffled based on a key and list of allowed values.

Pros: Retains data integrity and consistency between systems; multiple shuffles return same results.

Cons: Resource intensive and more configuration is required.

Example: John, Jim, James, john => Alan, Barry, Charles, Alan.

Encryption

An encryption value is generated by an encryption key.

Pros: Most secure of all masking approaches. Can be symmetric or asymmetric.

Cons: Complex to implement; easily identified as an encrypted field; Cyphertexts disrupt data types and reporting.

Example: Plaintext "John" is converted to ciphertext "XY31D0ER9209" using an algorithm.
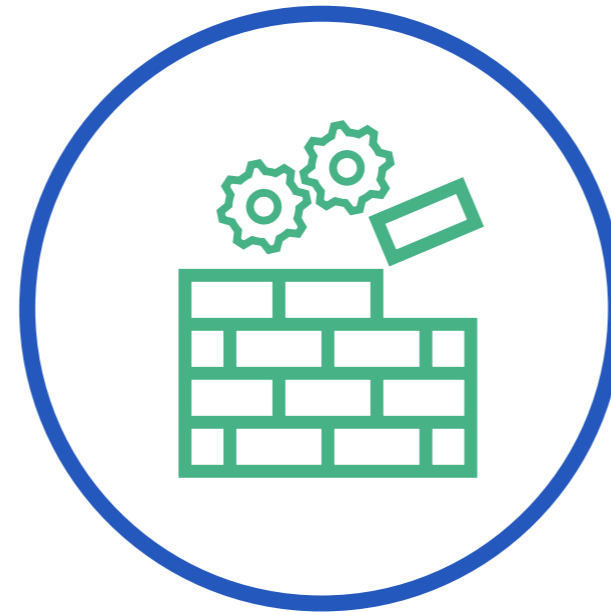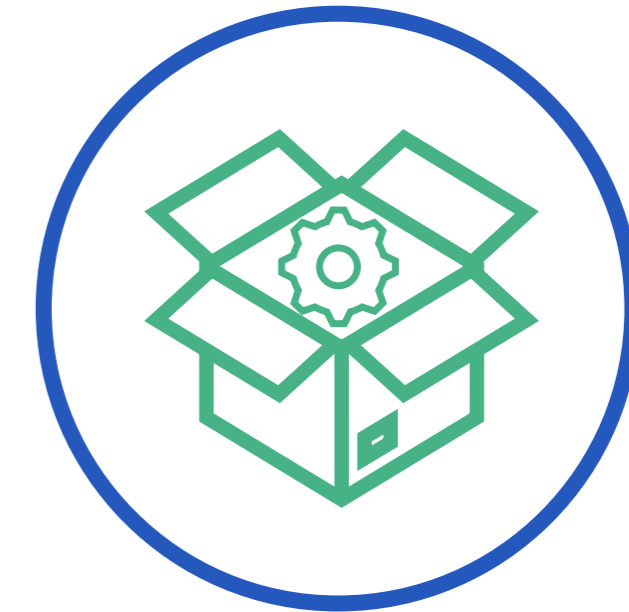
✓data domain

Implementation Options

Database Tool

- ➤ Database Management Tools.
- ➤ Can generate volumes easily.
- ➤ May need to be highly customised and unsophisticated.
- ➤ May provide limited self—service capabilities.

Build

- ➤ Scripted approach
- ➤ Technical know-how
- ➤ Human resource costs (OPEX)
- ➤ High maintenance

Specialised TDM

- ➤ Off-the-shelf, built for purpose.
- ➤ Fast to implement and execute.
- ➤ Vendor assistance.
- ➤ Can be costly to purchase (CAPEX).
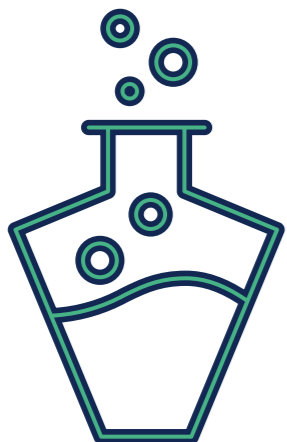
✓data domain

Summary

Layers of Security

- UAT, Pre-Prod, and Prod should have the same security – NO developer or tester admin access.
- Non-Production i.e. Dev and Test have a wider set of use cases, therefore access with controls are required.

Compliance by Design

- Necessary and *before* development and testing commences.
- Development and test environments should be ephemeral, regular hygiene practices should be followed – a chance for remediation.

Combination of Tools & Approaches

- Production – Fixed & Dynamic Masking, Encryption.
- Non-Production – Like Production *plus* scrambling, or synthetic data (for simple cases).

✓data domain

Discussions

✓data domain

✓data domain | Contact Us

New Zealand

Shed 19 Princes Wharf
137 Quay Street
Auckland 1010, New Zealand
+64 9 948 2225
info@datadomain.co

United States

25211 Grogans Mill Rd #450
The Woodlands,
Texas 77380
(832) 813-8063
sales.us@ttcglobal.com

Europe | UK

10 John Street
London WC1N 2EB
United Kingdom
+44 7384 719098
uk@ttcglobal.com

Singapore

Hong Leong Building
6 Raffles Quay, #33-03
Singapore 048581
+65 9822 6679
singapore@ttcglobal.com

United Arab Emirates

14th Floor, Al Khatem Tower
Wework Hub 71 Abu Dhabi
Global Market Square, Al
Maryah Island
Abu Dhabi, UAE
+971 58 5233912
UAE@ttcglobal.com

India

6 Floor Westport S. No.
32/1A/1/30 to 38 & 54 Pan
Card Club Rd, Baner, Pune,
Maharashtra 411045
india@ttcglobal.com

Australia

Level 4, 50 Miller St
North Sydney
NSW 2060
+61 2 8999 1965
australia@ttcglobal.com