



# MANAGING BIG DATA

PROF. DR. FLORIAN STAHL

# Managing Big Data



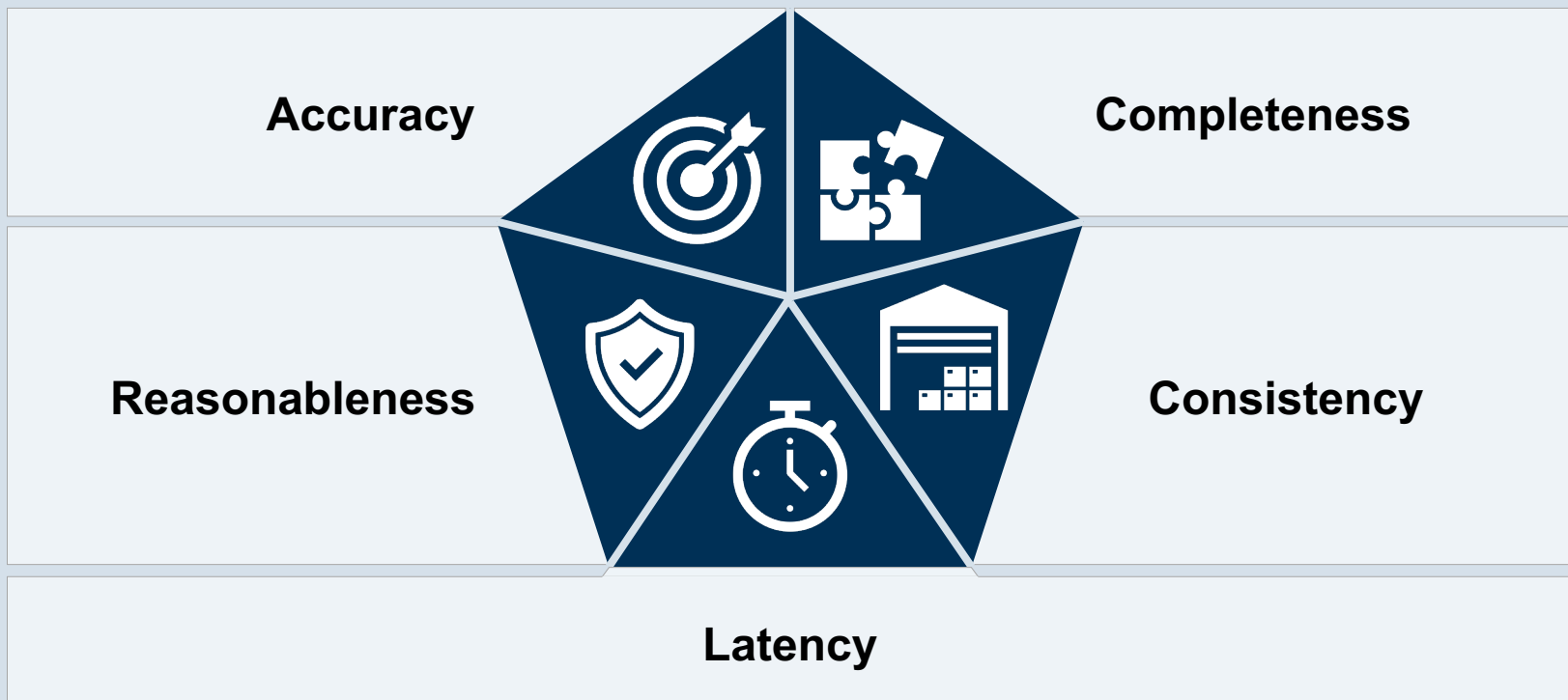
# What is Data Quality?



**Data quality** refers to the level of quality of data available to the business user. Data is considered high quality if they are fit for their intended uses in operations, decision making, and planning.



# Data Quality Dimensions



# Data Quality Dimensions

- Refers to the **correctness** of a data value in comparison to a reference source
- Establishing a system of record is necessary to determine accuracy

## Accuracy



## Completeness



- Refers to the **completeness** of the **information set** available to support a user's needs
- Reasons for incompleteness: missing datasets or metadata

## Consistency



- Refers to the **consistent representation** and **interpretation** of data across repositories, applications, tables, and fields across internal & external sources
- Data inconsistency creates unreliable information

# Data Quality Dimensions



## Latency

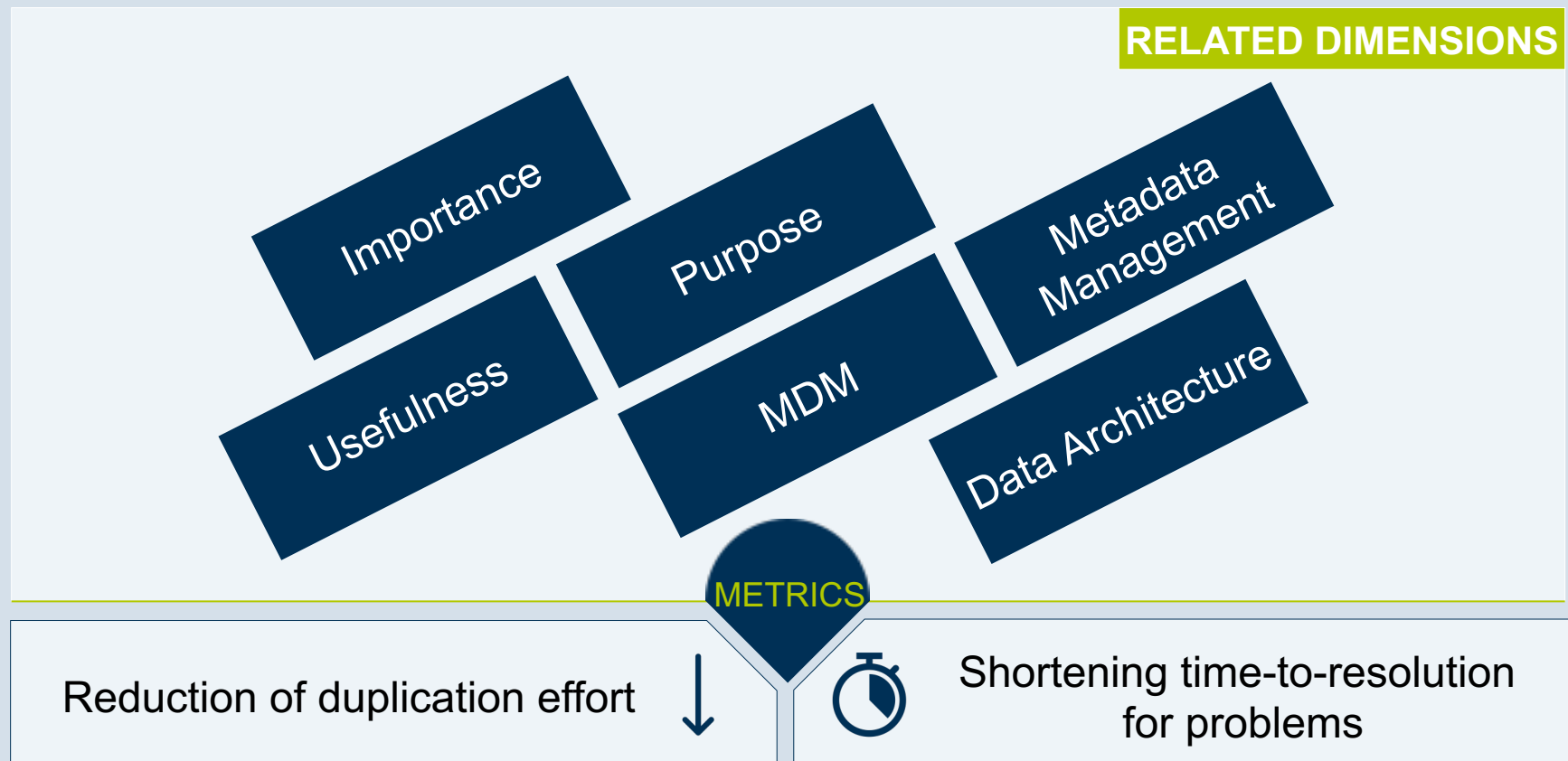
- Issues: **Currency of data, timeliness of data**
- Issues of latency are primarily addressed by technologies such as replication, real-time streaming, batch loads, and virtualization
- Significant data quality **costs to reduce latency** to near real time

- Refers to the **overall credibility** of a dataset
- **Quality** in aggregate terms
- **Accuracy** to the domain it models

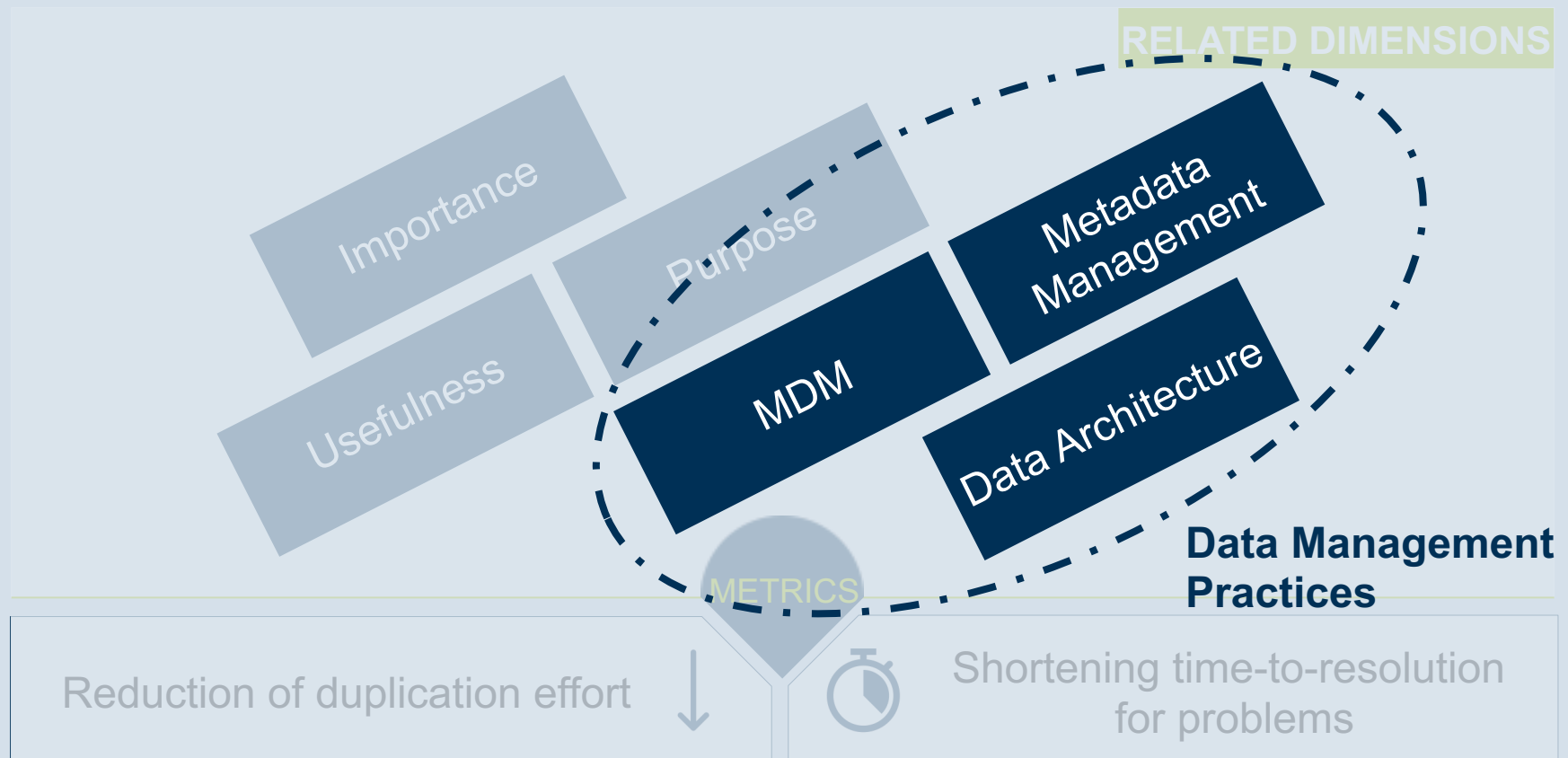
## Reasonableness



# Data Quality Dimensions



# Data Quality Dimensions





# Data Quality Activities



## Data Profiling

- Examines **data sources** using statistical methods to establish summaries and a snapshot of data structure, content, rules, and relationships
- Foundational for data quality metrics



## Data Quality Monitoring

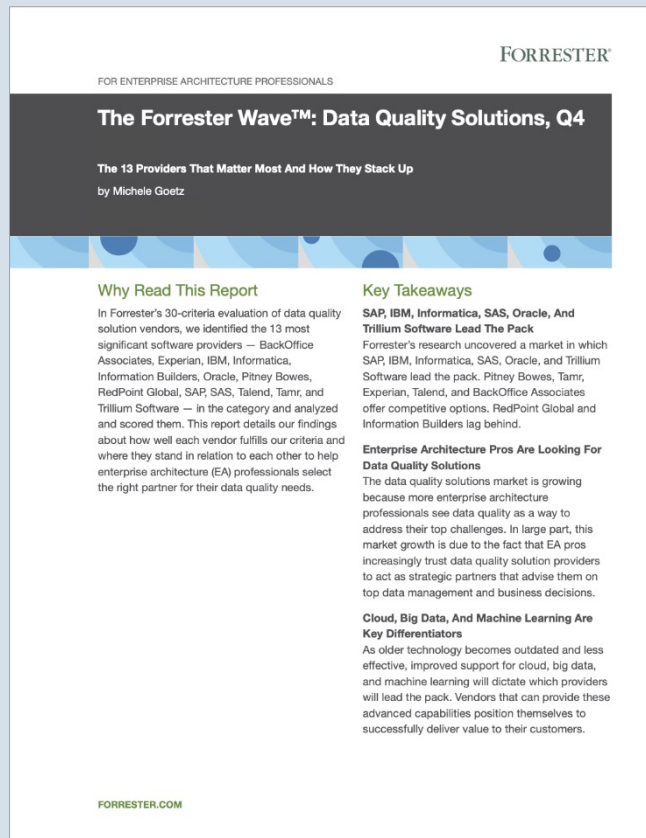
- Ongoing practice of **data profiling** and **reporting metrics** to stakeholders
- Basis for defining improvement initiatives and improving trust of data among users



## Data Cleansing

- **Detecting** and **correcting** corrupt or **inaccurate records** from a database
- May be performed interactively or as a batch
- To maintain data consistency, data shall be cleansed as close to the point of origin as possible

# Trusting Your Data



**Q4 2014**

## Global Online Data Quality and Trust Survey



43% of business and technology decision makers are somewhat confident and 25% are concerned about the quality of data.



Nearly 1/3 of data quality professionals spend more than 40% of their time validating data before their organization can use it for analysis and decision making.

# Creation of Trusted Data Environments



# Data Quality Challenges



**Inadequate  
Controls** at  
the Point of  
Origin



**Volume,  
Variety,  
Velocity**



**Environment  
Complexity**



**Too much  
proliferation  
and  
duplication**



**Poor  
metadata,  
unclear  
definitions,  
and multiple  
interpre-  
tations**



Data quality challenges point to the **lack** of a **coherent data strategy** and data quality governance approach across the organization.

# Data Quality Challenges



## Inadequate Controls at the Point of Origin

### External Data

- Inadequate acquisition controls
- Poorly defined SLAs
- Lack of governed process for sourcing data

### Internal Data

- Inadequate definition and/or application of business rules within transactional systems
- Errors during manual data entry processes

# Data Quality Challenges



**Volume,  
Variety,  
Velocity**

- **Increased pace** of data creation
- Need to keep pace in identifying data quality issues
- **Storage** in **nonstandard structures**
- **Unpredictability** in the structure and content of data
- Problems of automating data quality checks

# Data Quality Challenges



## Environment Complexity

- **Complexity** from data **distribution environments**
- **Physical** distribution (clouds, platforms, disparate technologies)
- **Global** distribution
- Across **internal** and **external sources**
- Difficulty to ensure **coherence** of data

# Data Quality Challenges



**Too much  
proliferation  
and  
duplication**

- **Technical ease** of duplicating and sharing data
- **Low cost** of **storage**
- **Ever-growing** number of sets of **data available**
- "Spreadmarts" proliferated on individual desktops
- Collective **tendencies** to duplicate and proliferate data



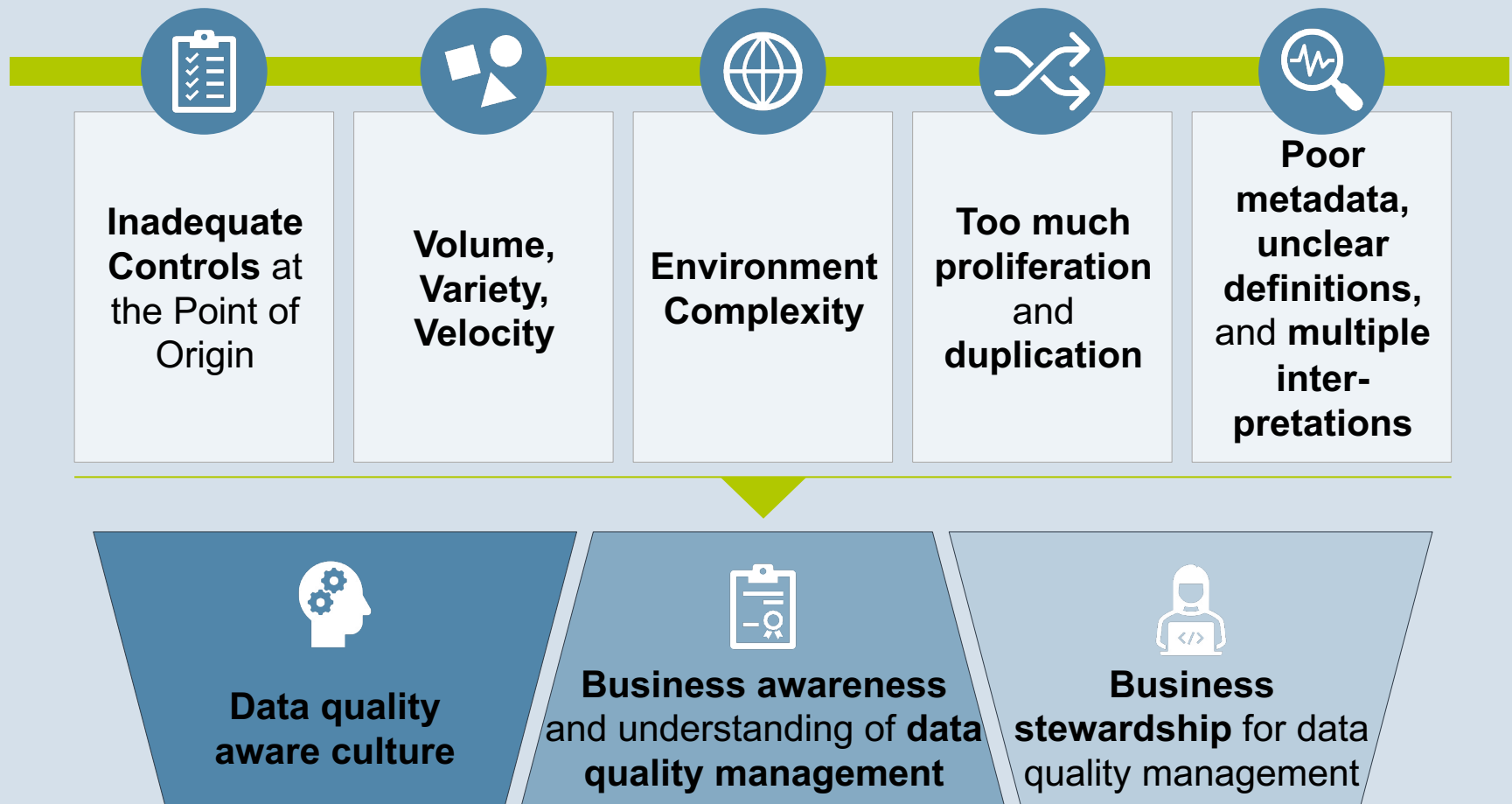
# Data Quality Challenges



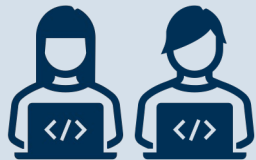
**Poor  
metadata,  
unclear  
definitions,  
and multiple  
inter-  
pretations**

- Organizations historically **have not prioritized** developing and **documenting** the **data** they create, acquire and share
- **No common, consistent understanding** of data
- **Metadata** as the single, most common interpretation of an organization's data assets  
→ Critical to **data quality**

# Data Quality Challenges



# Key Roles in Data Quality?



**Data  
Producers**

- **Introduce data** into the organization's environment
- **Establishing high quality data** at origin
- Often, producer is identified as **data steward/owner**
- Responsible for the definition of **preventive controls** that determine the quality of data at inception

## Data Quality

- Drive the **reason** for **data quality**
- May be many and varied across the organization
  - Data needs for **different purposes** depend on their **business role**
  - Understanding **data quality expectations** of different consumers as a driver for **data architecture/governance**



**Data  
Consumers**

# Key IT Roles in Managing Data Quality



## Data architects

- Define and maintain the **data environment blueprint**
- Managing redundancy, deploying appropriate technology and controls



## Data modelers

- **Represent** the **domain** and its **content**, the constraints, and business rules that bound data and ensure alignment to its intent
- Use of referential integrity, data types, “valid values”, ...



## Database administrators (DBAs)

- **Oversee care** and feeding of **physical data stores**
- Monitoring data for integrity issues and performance
- Provide input and guidance to managing data quality issues

# Data Quality Controls



Preventive



Detective



Corrective

- Ensuring quality at transition
- Most critical quality gate in the data quality cycle
- Achieved through **controls** within data origination systems, at manual data entry points, and via controls on data imported from external sources
- **Business users** as a significant component to manage preventive controls
- **Goal:** Data has a **clear definition, system of record, managed duplication, alignment to business rules** and constraints

# Data Quality Controls



Preventive



Detective



Corrective

- Executed to **ensure data quality after data has been created**
- May be executed across all parts of the data environment
- Key roles: Business and technical analysts, IT
- Examples: *Integrity checks, profiling, sampling, reporting, etc.*

# Data Quality Controls



Preventive



Detective



Corrective

- Address data quality issues **once they occur**, pointing to issues in preventive or detective controls
- Typically **procedural**
- Should be fully **auditable** to ensure that modifications to data are clearly explained and understood
- Recommended pattern: **Correct data at source** and **propagate the correction** to all points of consumption
- Key roles: Producers, consumers, IT

# Implementing Data Quality

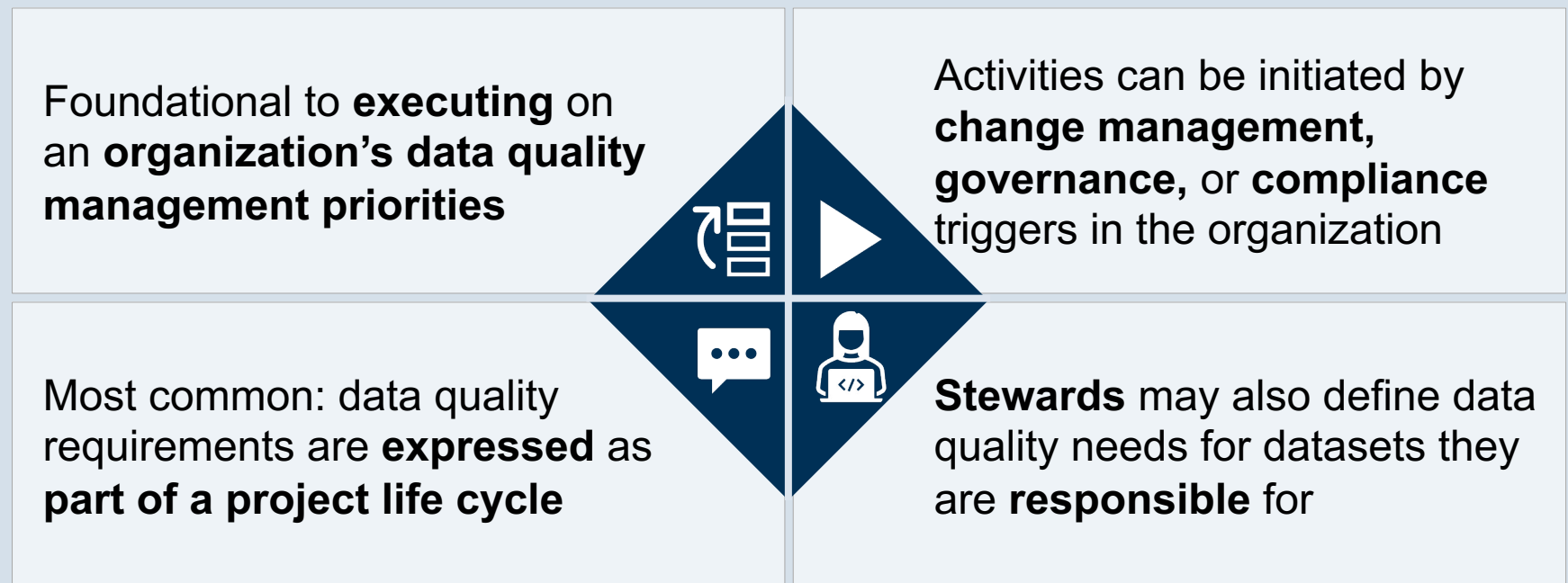
1		Defining data quality	5		Measuring data quality
2		Deploying data quality	6		Data classification
3		Monitoring data quality	7		Data certification
4		Resolving data quality issues	8		Data quality - trends and challenges



# Implementing Data Quality

1		Defining data quality	5		Measuring data quality
2		Deploying data quality	6		Data classification
3		Monitoring data quality	7		Data certification
4		Resolving data quality issues	8		Data quality - trends and challenges

# Defining Data Quality



# Defining Data Quality – Key Considerations



**Coordinated, collaborative data governance function**

**Holistic view** of data quality; important supporting process



Suitable **rigor** and **documentation** in defining data quality needs

**Clarity** and aligning to a **common understanding** of organizational **quality goals**



**Educating users** on data quality

**Articulation of quality needs**, matures DQ management



**Documenting and disseminating data quality requirements**

Consistent **understanding** and **transparency** to data quality expectations within an organization

# Implementing Data Quality

1		Defining data quality	5		Measuring data quality
2		<b>Deploying</b> data quality	6		Data classification
3		Monitoring data quality	7		Data certification
4		Resolving data quality issues	8		Data quality - <b>trends and challenges</b>

# Deploying Data Quality



This process is usually executed within **system development lifecycles** as part of the development and deployment process. It can also be enforced via **manual processes** and procedures as well.

# Deploying Data Quality – Key Considerations



Rules should be applied  
at the **point of creation**

- Designating data originating systems
- Data changes only at the SOR
- Distribution as read-only copies



**SLAs for externally sourced data**

- Should address DQ expectations
- Enforce them at the point of ingest



**Auditability** of data changes

- Transparency on changes
- Lets users know what they are getting



**Automating DQ enforcement**









- Improving predictability of outcomes, being more efficient



**Documentation of DQ rules;**

- Traceability to requirements across the organization

# Implementing Data Quality

1		Defining data quality	5		Measuring data quality
2		Deploying data quality	6		Data classification
3		<b>Monitoring</b> data quality	7		Data certification
4		Resolving data quality issues	8		Data quality - <b>trends</b> and <b>challenges</b>

# Monitoring Data Quality

## DQ HEALTH-CHECK

Guiding **understanding**, **awareness**, and **prioritization** of DQ investments

Understanding data quality over time builds **confidence** among users and drives impactful quick-wins to organizational objectives



DQ profiling can be initiated by a variety of triggers

## PROFILING PERSPECTIVES

- Data analysis within stores
- Providing statistics, summaries, aggregates that are informative to stakeholders

Top  
Down

Up  
Bottom

- Assessing health of data against domain rules and requirements provided by business SMEs
- Ability to verify alignment of data quality to needs



# Monitoring Data Quality – Key Considerations



**Profiling** is very important to **informing other processes**









You cannot improve something that is not measured



Data quality has **many dimensions to consider**

Careful consideration should be given to all of them in developing profiling characteristics

# Implementing Data Quality

1		Defining data quality	5		Measuring data quality
2		Deploying data quality	6		Data classification
3		Monitoring data quality	7		Data certification
4		<b>Resolving</b> data quality issues	8		Data quality - <b>trends</b> and <b>challenges</b>

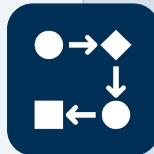
# Resolving Data Quality Issues



**Data quality issues management** includes processes to identify, evaluate issues, execute corrective activities, and report on DQ issues to concerned stakeholders.



An organization **requires processes to track, address, and resolve data quality issues** as they are encountered in order to meet the data quality expectations of the organization.



**Embedded processes**, particularly **automated** ones, are valuable in validating, identifying, and resolving issues in data quality in a timely and repeatable manner and preventing data degradation over time.

# Resolving DQ Issues – Key Considerations



Collecting **data quality information** over time



Reporting to stakeholders on **trends and metrics**



Embedding **preventive control processes**



Identifying **underlying or systemic issues**



Ensuring **auditability** of changes made in DQ issues management



**Automating** the issues resolution process with appropriate controls

- Building **awareness** of the state and progress of data quality issues

- **Preempting slowdowns** arising from data issues during mission critical periods

# Implementing Data Quality

1		Defining data quality	5		<b>Measuring data quality</b>
2		Deploying data quality	6		Data classification
3		Monitoring data quality	7		Data certification
4		Resolving data quality issues	8		Data quality - <b>trends and challenges</b>

# Measuring Data Quality



# Implementing Data Quality

1		Defining data quality	5		Measuring data quality
2		Deploying data quality	6		<b>Data classification</b>
3		Monitoring data quality	7		Data certification
4		Resolving data quality issues	8		Data quality - trends and challenges

# Data Classification



**Data classification** is the process of sorting and categorizing data into various types, forms, or any other distinct class. It enables the separation and classification of data according to data requirements for various business objectives.



**Classifying data**  
from the perspective  
of **business impact**



**Prioritizing data**  
**quality investments**  
based on that



**Critical data elements identification** describes the process of identifying the critical data elements as a data governance practice to prioritize funding, improve revenue and product quality.



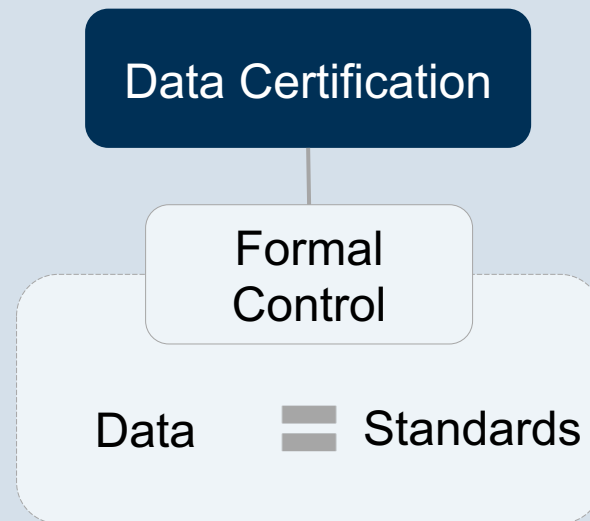
# Implementing Data Quality

1		Defining data quality	5		Measuring data quality
2		Deploying data quality	6		Data classification
3		Monitoring data quality	7		<b>Data certification</b>
4		Resolving data quality issues	8		Data quality - trends and challenges

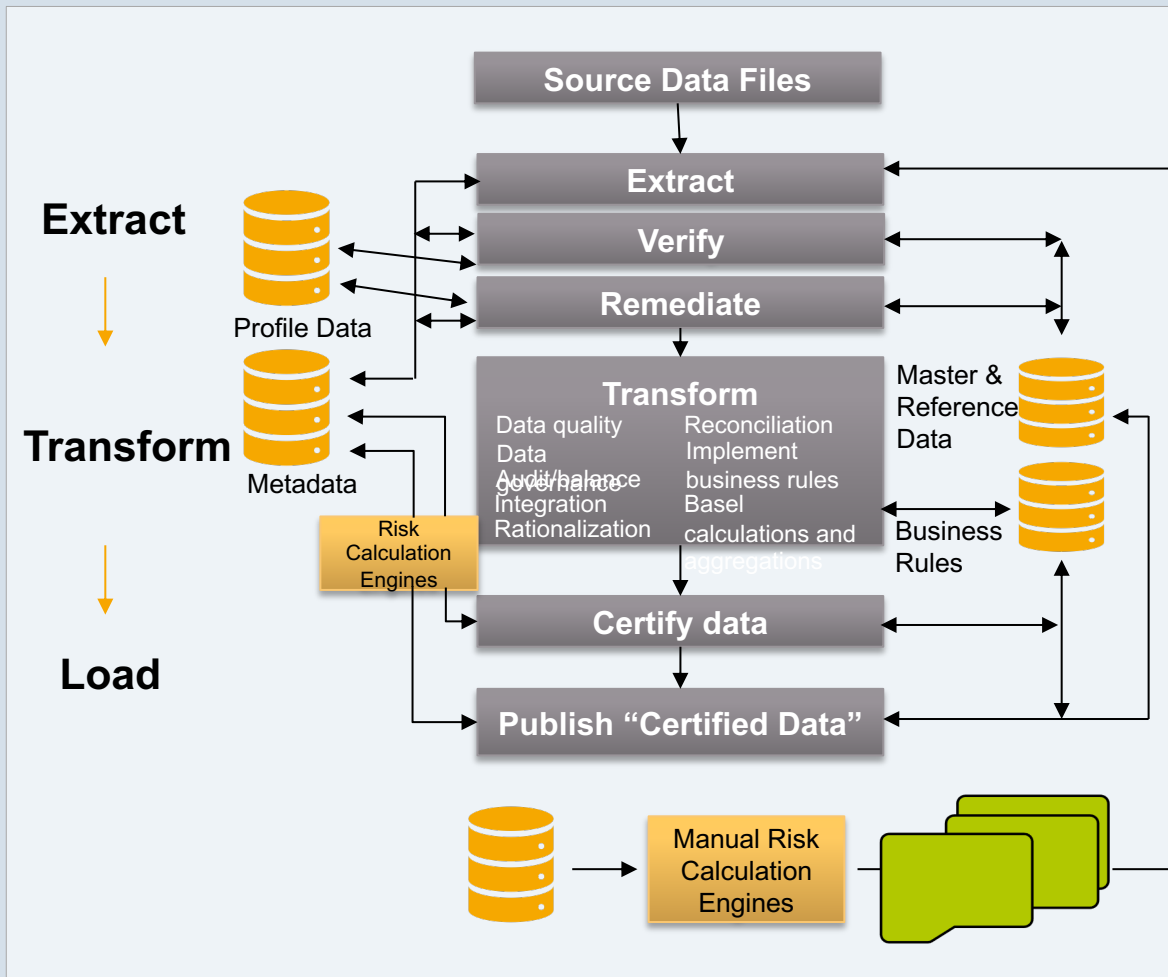
# Data Certification



**Certified data** can be defined as data that has been **subjected to a structured quality process to ensure** that it meets or exceeds the **standards** established by its intended consumers. Such standards are typically documented via service level agreements and administered by an **organized data governance structure**.



# Data Certification



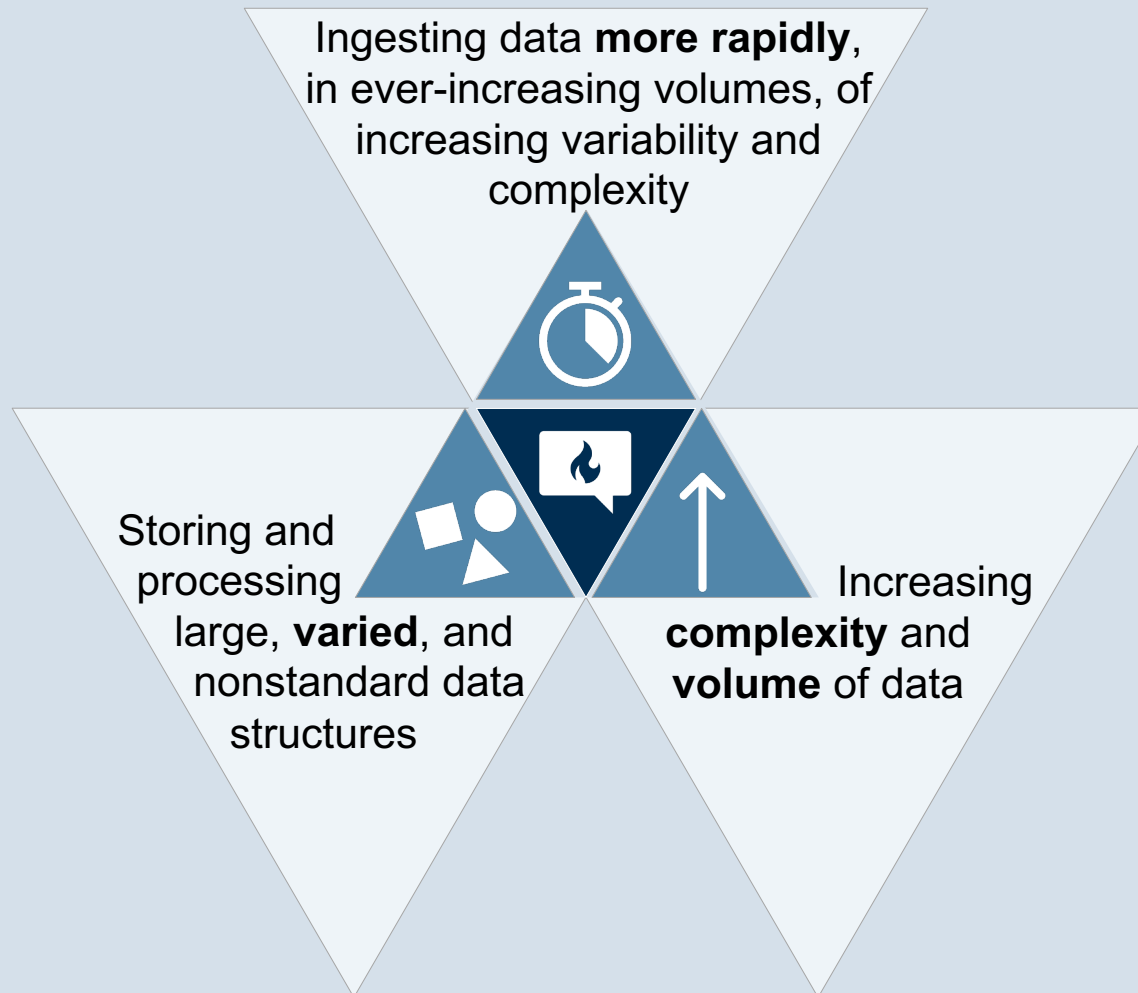
## Strengths of Data Hubs

- Data quality compliance Basel
- Sarbanes-Oxley compliant
- Enables data governance standards
- Enables LOB ownership and responsibility for data at an enterprise level
- Enables metadata management
- Enables business rules management
- Single version of the data truth
- Ability to source multiple BI platforms

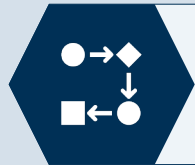
# Implementing Data Quality

1		Defining data quality	5		Measuring data quality
2		Deploying data quality	6		Data classification
3		Monitoring data quality	7		Data certification
4		Resolving data quality issues	8		Data quality - <b>trends and challenges</b>

# Data Quality – Trends



# Data Quality – Challenges



**Processes** to verify, control, correct, and disseminate high-quality data



Quality and provenance of data from **external sources**



**Preventing** proliferation of **poor-quality data** despite pace, complexity, and volume of data at entry points



Balancing **governance** and **agility**

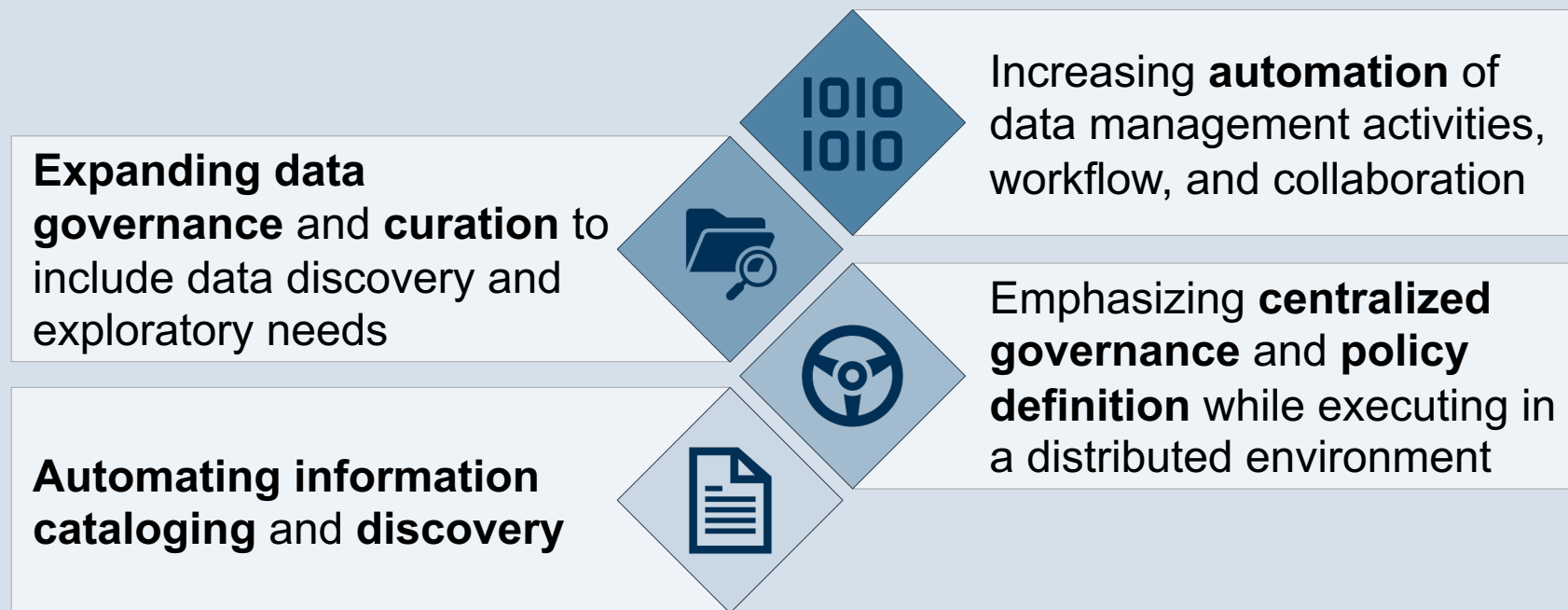


**Rapid data ingestion:** more agile, automated, control mechanisms and intervention technologies



**Documenting** and **making visible** the **quality of the data** using robust metadata capabilities

# Data Quality – Recent Techniques









# Data Quality Tools

Increasing complexity  
of data environment

Increasing data quality  
management needs

Substantial growth in the data  
quality tools market

## KEY FUNCTIONS

	<b>Profiling and metrics</b>		<b>Issue resolution and workflow</b>
	<b>Standardization and cleansing</b>		<b>Cleaning and enhancement</b>
	<b>Monitoring</b>		<b>Identity resolution</b>



# Key Take Aways



## Data Quality



Definition,  
dimensions, and  
activities



Challenges of data  
quality



Implementation of  
data quality



Emerging trends,  
challenges, and  
tools