The Future of Data Governance: Data Governance in the Data Lake

Michael Davis
Data Governance Leader, Voya Financial

Practitioner Track Organized by

22nd Annual MIT International Conference on Information Quality
MIT ICIQ 2017
October 6 - 7, 2017. Hosted by the University of Arkansas at Little Rock

Future of Data Governance

Data Governance in the Data Lake

By Michael G. Davis

Pre-Historic Data Era

Topics

- Early forms of data
- Pre-historic Data Era
- Evolution of Data
- The Advent of Big Data
- Meta Data Management
- Data Governance

First Data recorded

Sumerian Limestone Kish Tablet

Early Days of Data

?

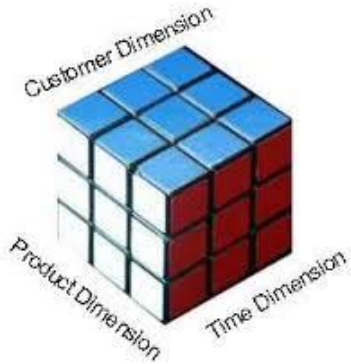
Evolution of Data

1960s to 2000s

Mainframe

Data Warehousing

Business Intelligence



The advent of Big Data

90%

2.5

What is the FAIR Principle?

Meta(data) is Findable

- (meta)data are assigned a globally unique and persistent identifier
- data are described with rich metadata
- metadata clearly and explicitly include the identifier of the data it describes
- (meta)data are registered or indexed in a searchable resource

Meta(Data) is Accessible

- (meta)data are retrievable by their identifier usin a standardized communications protocol
- the protocol is open, free, and universally implementable
- the protocol allows for an authentication and authorization procedure, where necessary
- metadata are accessible, even when the data ar no longer available

Meta(Data) is Interoperable

- (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (meta)data use vocabularies that follow FAIR principles
- (meta)data include qualified references to other (meta)data
- (meta)data are assigned a globally unique and persistent identifier

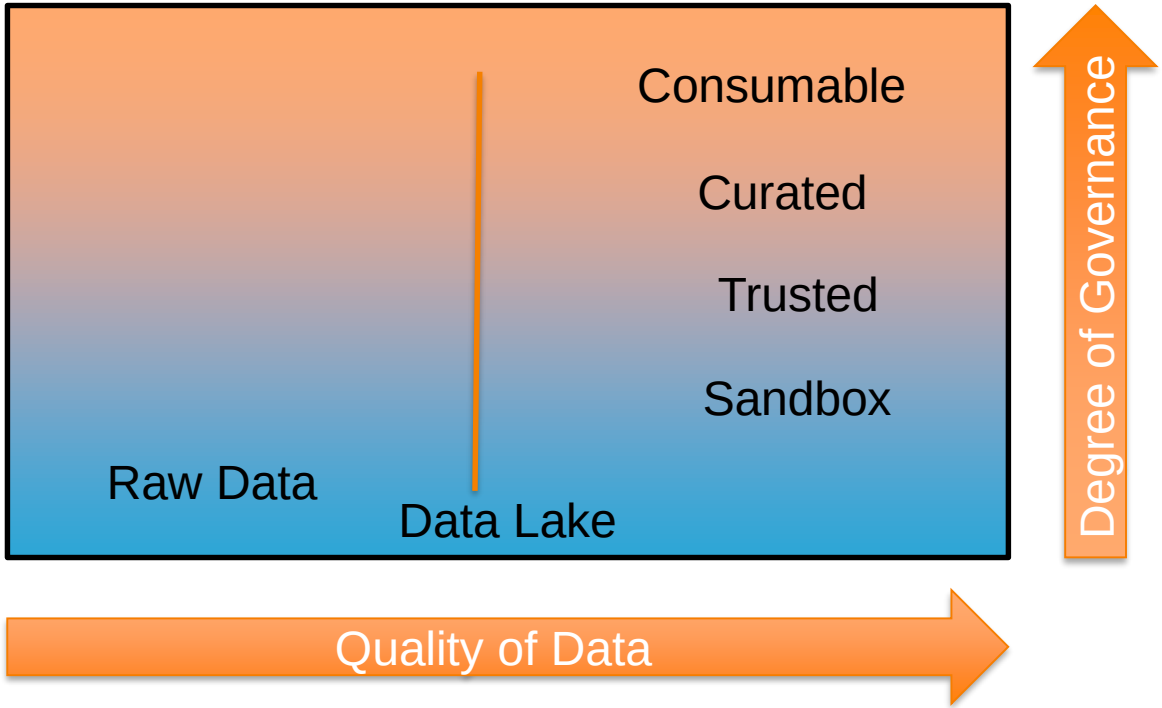
Meta(data) is Repeatable

- meta(data) are richly described with a plurality of accurate and relevant attributes
- (meta)data are released with a clear and accessible data usage license
- (meta)data are associated with detailed provenance (Data Lineage)
- (meta)data meet domain-relevant community standards

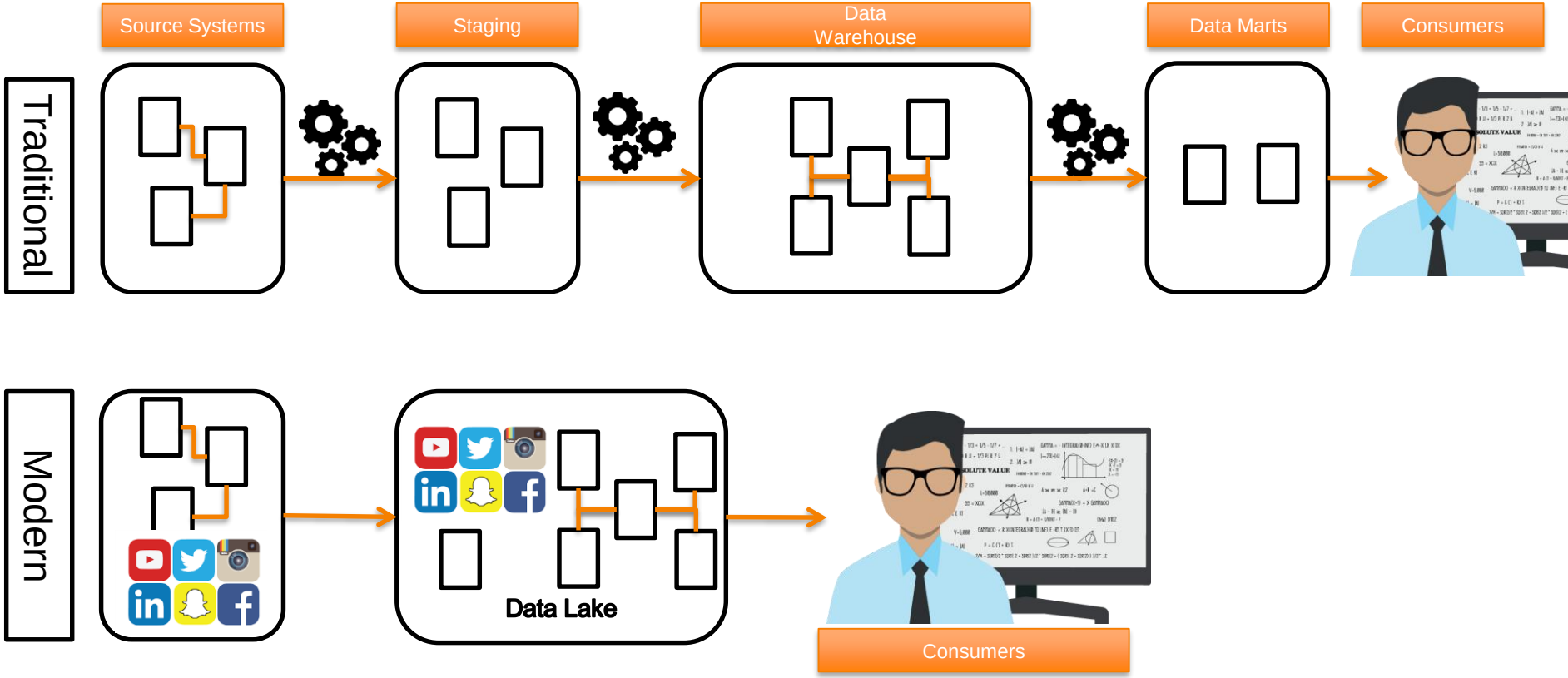
Data Lake Success Rate

According to Gartner, “through 2018, 80% of data lakes will not include effective metadata management capabilities, making them inefficient”

Data Lake


Consumable
Curated
Trusted
Sandbox
Raw Data
Data Lake
Degree of Governance
Quality of Data

Traditional vs Modern Data Architecture



Data Lake Governance