Data Life Cycle Model

December 2013

The Consortium for Advanced Management – International (CAM-I)

Intelligent Data Quality Management

Primary Team Authors (alphabetical by last name): Nandana Baruah (Boeing), Don Carlson (Bank of America), Nancy Lashbrook (Boeing), Rob Justus (Grant Thornton),Mike Lanouette (Definitive Logic), Jeff Lawton (Grant Thornton)

Primary Team Members (alphabetical by last name): Nandana Baruah (Boeing), Don Carlson (Bank of America), Jon Francois (USDA), Pamela Henderson (USAASC), Nancy Lashbrook (Boeing), Rob Justus (Grant Thornton), Mike Lanouette (Definitive Logic), Jeff Lawton (Grant Thornton)

6836 Bee Caves, Suite 256

Austin, Texas 78746USA

Ph 512-617-6428

www.cam-i.org

Introduction

Data is a corporate asset.It reduces uncertainty about decisions.It affects behavior.It can even have its own market value.When data is compromised by quality problems, not only does the value of it as an asset decrease, but there can be further reaching consequences as well (e.g., cost efficiency, risk management, regulatory compliance, agility, revenue growth, readiness).So if data is an asset, why is not always treated as such?For example, stores conduct regular audits of their inventory, but an organization will not inventory their data unless a data-related crisis has occurred that spurs the organization into action. Perhaps the answer lies in the unusual characteristics that define data:

- Data is easily created
- Data is easily duplicated
- Data is easily destroyed
- Data is easily altered
- Data is easily transported and transferred
- Data is easily misunderstood (the whole science of statistics has risen to describe data and what it means)
- Data is easily stolen
- Data can be combined with other data
- Data can lose value with age (updates required with time)
- Data is difficult to preserve (what can you do with a reel-to-reel tape now?)

We believe these characteristics defytraditional asset management techniques and are why organizations treat data differently from other common assets, such as property, plant, equipment, accounts receivable, fleet, knowledge, and branding.Furthermore, when equipment breaks down, a number of signals (e.g., sight, smell, sound) make the breakdown easily discoverable—thisis not true when data becomes corrupted.

CAM-I'sIntelligent Data Quality Management (IDQM) group seeks to apply greater rigor and structured, proactive approaches for managing data as an asset, with a particular focus on attaining a level of quality such that the data suits its intended purpose. IDQM's mission, therefore, is:

> "To create data quality and data management frameworks for better informed business decision making, improved investment analysis and allocation of appropriate funding/resources, reduced risk exposure, and controlled improvement."

To achieve its mission, IDQM ultimately will develop a framework that provides an intelligent (i.e., disciplined, consistent, informed) means of managing data quality by developing models

that assign dollar costs to data quality (thereby allowing data to be valued and, if corrupted, costed), analytics-based approaches, a data quality maturity model, and best practices research.

How the Data Life Cycle Model Was Created

For each of these elements of the framework to be integrated, a common understanding of datais needed (e.g., consistent terminology), particularly how it changes over time. The purpose of this document, therefore, is to provide such a model.To develop the model, IDQM first began by researching over 50 data life cycle models developed by academic, scientific, business, and public sector organizations to identify a model that could be used or repurposed without having to create one. However, upon examination of these models, we found that none were generic enough to support any industry, could support IDQM goals (e.g., creation of a valuation model), or aligned with IDQM members' best practices and knowledge. As an example of the latter, nearly two thirds of the models were expressed as circular, cyclical models. This implied that data, after it is deleted, necessarily leads to resurrection as new data. The IDQM perspective, however, is that the life cycle of data is linear: it is created, flows through a number of stages, and upon deletion, ends that data's life cycle—any new data that is created results from new requirements and is a separate flow.As another example, IDQM believes that a linear model better supports the variations in use cases by which certain stages of the lifecycle can be bypassed altogether or iterated. Lastly, none of the models examined were comprehensive enough to incorporate best practices(e.g., risk reduction, valuation) across industries, nor were they generic enough to accommodate virtually any type of data. With no existing data life cycle model that could be used or repurposed, IDQM drew from the strengths of several of the models reviewed to create a hybrid model, then added additional features to meet IDQM needs.

The Data Life Cycle Model

The table below identifies our proposed data quality model (DQM) lifecycle phases.Each phase is listed and briefly defines in a separate column and are listed in chronological order from left to right. The table rows represent activities that arelikely to be performed within each lifecycle phase with specific activities identified for each lifecycle/activity combination.Our lifecycle activities include design and architecture, governance, communications and outreach, and QA/QC activities with specific detail-level activities identified within each of four the high-level activities for each lifecycle phase. For example, in the lifecycle Define phase and design and architecture activities, we recommend that users of our methodology:

- Describe data needed to accomplish goals
- Identify customers and uses for the data
- Define data structures
- Define a high-level architecture

	Life Cycle Stages								
	Define	**Appraise**	**Obtain**	**Transform**	**Store**	**Register**	**Consume**	**Archive**	**Dispose**
Definition of the Stage	Describe the data or information needed to support an objective	Determine the optimal source(s) of the data	Acquire the data from existing sources or create the data	Change the data to make it fit for purpose	Stockpile the data and make the data available for use	Publicize the existence of data and provide guidance on how to use the data	Utilize the data to create information to authorized users and authorized uses	Preserve the data for future uses	Purge the data such that it is no longer available for any use
Design and Architecture Activities	• Describe data content needs (e.g. what data do I need to accomplish my goals?) • Identify customer and uses (e.g., what is the data used	• Does data current exist (inventory by consulting register)? If so, then do we need to repurpose? • Reuse, repurpose,	• Design for data quality • Business rules in the forms of edits • Metadata (get, update or create) • Data (get or create) from	• Source to target mappings • Cleansing • Derived data • Repurposing • Lineage	• Build repository • Secure repository • Implement disaster recovery • Implement backup/restore • Load balancing	• Build & deploy communication mechanisms/registry • Create metadata (key search criteria) • Taxonomy	• Build models • Build access mechanisms (e.g., business intelligence) • Build feedback mechanism • Build interfaces to other systems for	• Design and implement archiving strategy • Defined retention period (e.g. ligation risk) and business rules	• Automate • Impact/depend encyanalysis • Build tool for doing the disposal • Last change review for possible

	Life Cycle Stages								
	Define	**Appraise**	**Obtain**	**Transform**	**Store**	**Register**	**Consume**	**Archive**	**Dispose**
	for?) • Define data structure needs (use taxonomy if needed) • High-level target architecture (include security assessment)	create, or buy/acquire decision • Low-level target architecture • Consult data management standards • Determine approved sources • Duplicate data	approved sources		and failover • Define data access needs and rules vs. costs		repurposing of the data • Security access that conforms to policy and compliance (access control) • Protection of PII and intellectual property, HIPAA • Identify actors (IT, regulators, system to system, end-users (e.g., analysts)	(e.g.,rolling process with daily deletes) • Automate • Build tool for doing archiving	repurpose
Governance Activities	• Identify stakeholders • Consult architecture board	• Impact analysis • Authorization • Strategic alignment • Investment review	• Assign stewardship for new (created) data	• Monitoring for changes	• Access • Define data access needs and rules	• Metadata	• Data usage governance (meaningful use, honor customer requests) • Data access governance • Regulatory requirements (e.g., telephone protection act)	• Execute per standards • SLAs (response time) • Retention policy	• Consult retention period • Consult business rules for purge • Create purge rules/standards • Approve ad-hoc purges
Communication and Outreach Activities	• Contact stakeholders	• Data owner/ stewards • System owners • Process owners • Vendors • What do we need to do to get the data?	• Data owners/ stewards • Systems owners • Vendors • Process owners • Upstream and downstream dependencies (other stakeholders)	• Verify business rules with business owners • If a change impacts, then have a dialog (understand the changes) • Alerts	• Dependent on type of change executed	• Publish metadata • User outreach • Training • Compliance activities by law or policy (SORN, system of record/origin/access) • Help desk • Register in corporate asset repository	• User outreach • Training • Compliance activities • Help desk • SMEs to provide guidance on proper use of data • User feedback on continued usability/correctness of data	• Notification of retention • Update registry of available data	• Publish retention schedule • Pre-notification sent to users • Remove tech metadata

	Life Cycle Stages								
	Define	**Appraise**	**Obtain**	**Transform**	**Store**	**Register**	**Consume**	**Archive**	**Dispose**
Quality Assurance and Quality Control Activities	• Characterize data quality requirements	• Assess level of data quality (data profiling) • Define business rules that govern selected data • Determine cleansing strategy or remediation (for new source of data)	• Reconciliation (did I get the data that I expected?) • Data movement controls	• Transformation and quality rules • Validate derived data • Document lineage • Identify outliers and researching • Perform reconciliation	• Security • Validation	• Data quality alerts • Register quality • Validate registry findings	• Feedback loop for errors found • Analyst performed QA/QC	• Reconciliation	• Verify that data was purged • Verify no archived copies anywhere

Use Cases

The life cycle has basic linear qualities, but data need not follow each phase in order from left to right (e.g., phases could be iterative or skipped). In this section, we provide some examples of use cases that describe how data could move through the cycle based on organization's data management decisions, practices, culture, and policies. The use cases into three categories: Managed,Unmanaged, and Iterative.

Managed Use Cases

Managed processes follow the life cycle phases in order from left to right without skipping phases. There are three flavors of managed process used cases: Managed, Risk Averse, and Pack Rat.

Managed

Data follows each life cycle step. When data no longer serves a purpose, it is disposed—maintaining data costs resources, and with no value, it is removed.

Pack Rat

Data is archived when it reaches the end of its useful life. By archiving to cheaper, offline storage, the data is kept in perpetuity in case it is ever needed attain. Data is never disposed, just in case.

Impatient

Data is never archived or disposed. It is constantly accumulated over time. Any data ever collected is instantly available with no restoration lag. New storage must be purchased and brought online frequently to accommodate the rising volumes of data. Maintenance costs will be on an ever rising curve.

Unmanaged Use Cases

Unmanaged processes follow the life cycle phases in order from left to right, but skips phases. It is important to note that unmanaged use cases are not meant to be negative in nature, just that not all phases are needed or not all best practices are being brought to bear. There are many flavors of unmanaged processes. Three interesting types are provided here: Exploration, Waste, and Mystery.

Exploration

Users obtain data sets without much knowledge, if any, about the data sets. Understanding of the data is achieved through consumption and then disposed when the target insight is achieved.

Waste

Data is properly managed and stored, but there is no use for it. Users do not consume any of the data. Waste is not intended to have negative connotations, just that effort and resources are expended with no perceived value. An exception may be legal compliance or risk mitigation.

Mystery

Data is not registered, so only the data manager knows of its existence.

IterativeUse Cases

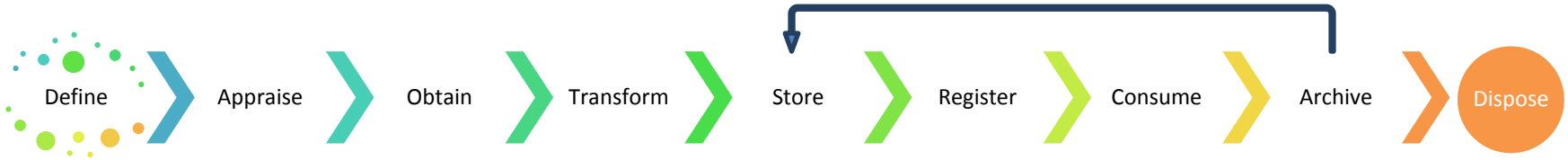
Iterative use cases follow a managed or unmanaged process from left to right, but for various reasons, may move from right to left.

Oops

Data that was disposed either should not have been disposed or has developed new value. However, the data is known and thus the Define and Appraise phases are not needed.

Need It Again

Data that has been moved to offline archives is once again needed and is restored to online storage.



Found a Problem

In consuming data, an error is found (i.e., data quality issue) and the organization takes steps to resolve the error.

Need More of What We Have

In consuming data, it is determined that more data of an existing type is required (e.g., additional customers, more years of data).

Need More of Something Else

In consuming data, it is determined that an expansion of existing data is required (e.g., new descriptive elements of an entity).

Need Something New

In consuming data, it is determined that either the data that exists is inadequate in meeting an evolved data need, thus revisitation of the data requirements for new or expanded data is required.

Enablers

To further demonstrate the use of the life cycle model, the table below describes how various enablers (e.g., tools, standards, technical and business metadata, change management techniques) are applied during each of the stages. Enablers can be applied in a number of ways, such as the creation, use, consultation, and update of these enablers. The value the table below brings is the identification of when and how enablers can be expected to be involved and how they change as the data they support moves through its life cycle.

	Life Cycle Stages								
	Define	**Appraise**	**Obtain**	**Transform**	**Store**	**Register**	**Consume**	**Archive**	**Dispose**
Tools	• Use (modeling tools)	• Use • Consult	• Use (ETL) • Create (SQL, SAS)	• Use • Create	• Use	• Create • Use	• Create • Use	• Create • Use	• Automate • Use ad-hoc process
Standards	• Create/consult	• Consult	• Create (e.g. interface control document) • Update • Consult	• Create • Update • Consult	• Use • Create • Consult	• Use • Consult	• Create • Consult	• Create • Consult	• Create • Consult
Technical Metadata	• None	• Consult	• Consult • Create	• Create • Update	• Update	• Create • Publish	• Consult	• Consult • Update	• Delete • Consult
Business Metadata	• Create	• Consult	• Consult • Create	• Create	• None	• Publish	• Consult	• Consult	• Consult
Change Management	• Consult	• Determine impact • Change assessment • Consult	• Consult	• Consult	• Consult	• Consult	• Consult	• Consult	• Consult

Risk and Valuation

To support future IDQM deliverables, the table below identifies various valuation and risk considerations for each stage of the data life cycle.

	Life Cycle Stages								
	Define	**Appraise**	**Obtain**	**Transform**	**Store**	**Register**	**Consume**	**Archive**	**Dispose**
Valuation	• Target costing • Cost benefit • Expected value • ROI • Define success criteria	• Identify costs (extraction, purchase, opportunity, design) • Regulatory compliance and risk • Single source for data • Revenue generation opportunities	• Vendor cost • Architectural costs (multiple point to point, off the grid, using non supported data structures) • Not following SDLC • Timing • Quality (fits the purpose/intended need) • Active management of obtaining to standard • Cost of data movement	• Manage costs efficiently • Lost data cost • Data quality cost • Data gains benefits • Early warnings/alerts • Better change management • Costs associated with transformation rules embedded in code rather than metadata driven rules • Transparency of rules add value	• Optimize storage costs • Maintenance costs • Disaster recovery costs • Upgrade costs	• Ease by which it can be located or found • Costs (various) • Share data with others (if allowable)	• Number of users (people and applications) • Output (reports, analyses) • Value of data driven decisions • Data as product • Data innovation (analysts discover new uses for existing data)	• How often is the data used? (usage metrics) • Regulatory compliance • Reduce maintenance costs • Data medium usability risk • Accessibility costs	• Reduce legal risk through discovery • Regulatory compliance • Reduce maintenance costs • Future use risk
Risks Involved in Each Stage	• Inaccurate definition of proper data uses • Poor project definition • Granularity of data improperly defined at too high a level	• Creation of redundant and/or conflicting data stores • Misunderstanding of data sources leads to incorrect buy/make decision • Lack of funding to get the data that is really needed to address the problem, use data not	• Not getting from the authoritative/correct source • Poor or misunderstood definitions from providers • Incorrect extraction (change data capture wrong) • Data is insecure during transport	• Complexity of transformations • Decentralization of transformations obfuscates data lineage • Incorrect transformation/bug • Cleansing divorces the data from its source—no longer the same	• Cloud vs. internally hosted (security) • Insufficient backups result in loss • Corruption of data	• Incomplete, incorrect descriptions of data leads to misuse • Missing descriptions leads to non-use • Brain drain reduces knowledge retention if not captured	• Insufficient security to protect the data from unauthorized consumption • Siloed nature of data prevents making the necessary connections • At this stage, there is an aggregation of all errors made in each stage prior • User error and	• Insufficient archiving leads to inadvertent loss or failure to recover • Takes too long to recover from archive, beyond timeframe where data is needed – not responsive enough • Archiving policy incorrectly defined (e.g., disposal automatic)	• Data you really wanted is gone

	Life Cycle Stages								
	Define	**Appraise**	**Obtain**	**Transform**	**Store**	**Register**	**Consume**	**Archive**	**Dispose**
		designed to answer instead • Data sensitivity is misclassified					user training issues such that the correct data is not queried/ aggregated/ disaggregated		

Acronyms

The table below defines acronyms found within this document.

Acronym	Definition
BI	Business Intelligence
ETL	Extract, Transform, and Load
HIPAA	Health Insurance Portability and Accountability Act
IDQM	Intelligent Data Quality Management
IT	Information Technology
PII	Personally Identifiable Information
QA/QC	Quality Assurance/Quality Control
ROI	Return on Investment
SDLC	System Development Life Cycle
SLA	Service Level Agreement
SME	Subject Matter Expert

About CAM-I

The Consortium for Advanced Management –International (CAM-I) is a research organization consisting of sponsoring companies and academia who work in collaboration to study and solve management problems and critical business issues common to the group in the areas of cost, process and performance management. More information here:http://www.cam-i.org/