Data Management

1 Introduction

Data Management is the development, execution and supervision of plans, policies, programs and practices that deliver, control, protect and enhance the value of data and information assets throughout their lifecycles.

A **Data Management Professional** is any person who works in any facet of data management, from highly technical to strategic business, to meet strategic organisational goals.

Data Management requires technical and non-technical skills, and business and IT people must collaborate to share responsibility for managing data. Data and information are vital to the operations of most organisations.

1.1 Business Drivers

- Competitive advantage of better decisions based on reliable, high quality data
- Failure to manage data is the same as the failure to manage capital
- The primary driver is to enable organisations to get value from their data assets

1.2 Goals

Data management goals within an organisation:

- Understanding and supporting the information needs of the enterprise including all stakeholders
- Capturing, storing, protecting and ensuring the integrity of data assets
- Ensuring quality of data and information
- Ensuring privacy and confidentiality of stakeholder data
- Preventing unauthorised or inappropriate access, manipulation or use of data
- Ensuring data can be used effectively to add value to the organisation

2 Essential Concepts

2.1 Data

Definitions of data:

- IT: Information that has been stored in digital form
- Facts which can be aggregated, analysed and used to make a profit, improve health or influence public policy
- Data is a means of representation which stands for things other than itself (Chisholm, 2010)
- Data is both an interpretation of the objects it represents and an object that must be interpreted (Sebastian-Coleman, 2013)
- To interpret data, context or Metadata is needed

2.2 Data and Information

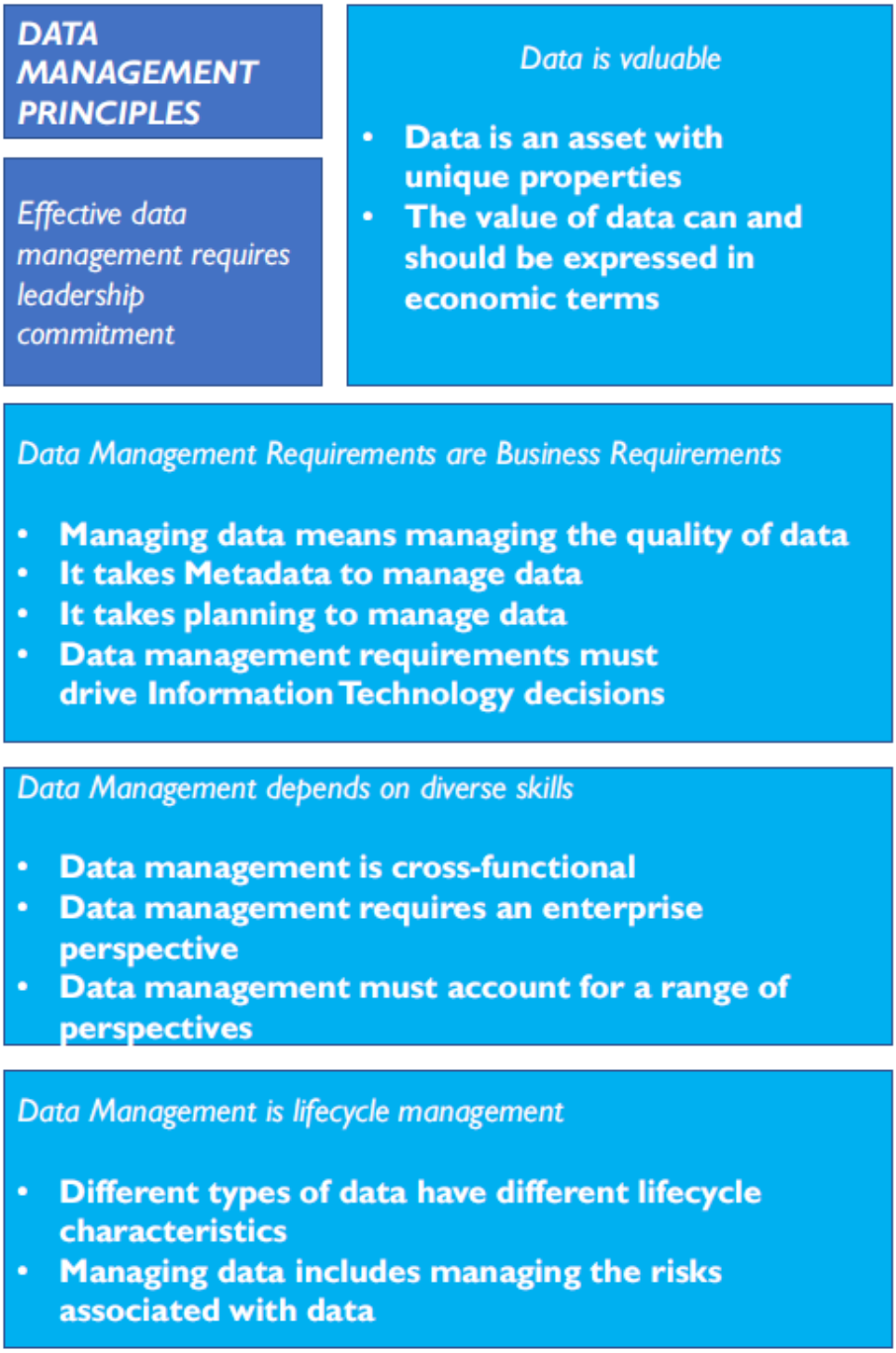
- Data does not simply exist; it has to be created.
- Knowledge is required to create data in the first place
- Data is a form of information and information is a form of data
- Organisations may differentiate between information and data for communication between different stakeholders

- The terms “data” and “information” are used interchangeably in the DMBOK

2.3 Data as an Organisational Asset

An asset is an economic resource, that can be owned or controlled, and that holds or produces value. As organisations become more reliant on data to make decisions the value of data assets is more clearly established.

2.4 Data Management Principles



- **Data is an asset with unique properties**: Not consumed as it is used unlike financial or physical assets.

- **The value of data can and should be expressed in economic terms**: No standards yet. Organisations should develop consistent ways to quantify value. Measure costs of low quality as well as benefits of high quality data.
- **Managing data means managing the quality of data**: Understand stakeholders' requirements for quality and ensure data is fit for their purpose.
- **It takes Metadata to manage data**: Metadata enables us to understand what intangible data is, and how to use it. Metadata originates from processes related to data creation, processing and use such as architecture, modelling, stewardship, governance, DQ management, SDLC, IT, business operations and analytics.
- **It takes planning to manage data**: Complex tech and business process landscapes, with moving data, coordinating work and keeping results aligned requires planning (architectural and business process)
- **Data management is cross functional; it requires a range of skills and expertise**: Collaboration between technical and business skills
- **Data management requires an enterprise perspective**: To be effective. That is why DM and DG are intertwined.
- **Data management must account for a range of perspectives**: DM must constantly evolve to keep up with the ways data is created and used
- **Data management is lifecycle management**: DM practices need to account for the data lifecycle
- **Different types of data have different lifecycle characteristics**: DM practices need to recognise this and be flexible enough to meet different kinds of lifecycle requirements
- **Managing data includes managing the risks associated with data**: Lost, stolen, misused and ethical implications. Managed as part of the lifecycle.
- **Data management requirements must drive Information Technology decisions**: Technology serves, rather than drives an organisation's strategic data needs.
- **Effective data management requires leadership commitment**: Complex processes of DM require the vision and purpose that comes from committed leadership.

2.5 Data Management Challenges

2.5.1 Data differs from other assets

Other assets exist in one place at one time. Data is different:

- Data is not tangible
- It is durable and does not wear out
- The value changes as it ages
- Easy to copy and transport
- Not easy to reproduce if lost or destroyed
- can be stolen but not gone
- Not consumed when used
- The same data can be used by multiple people at the same time

Data has a value but it is challenging to measure it. It must be managed with care.

2.5.2 Data valuation

Value is the difference between the cost of a thing and the benefit derived from that thing. The value of data is contextual and often temporal.

Sample cost categories:

- Obtaining and storing data
- Replacing lost data
- Impact of missing data to the organisation
- Risk mitigation and potential costs of risks associated with data
- cost of improving data
- Benefits of higher quality data
- What competitors would pay for data / what data could be sold for
- expected revenue from innovative uses of data.

2.5.3 Data Quality

Define business needs and the characteristics that make data high quality with business data consumers. People who use data need to assume data is reliable and trustworthy as they need to use the data to learn and create value, and make business decisions. Poor quality data is costly to the organisation.

Costs of poor quality data:

- Scrap and rework
- Work-arounds and hidden correction processes
- Organisational inefficiencies or low productivity
- Organisational conflict
- Low job satisfaction
- Opportunity costs, including inability to innovate
- Compliance costs or fines
- Reputational costs

Benefits of high quality data:

- Improved customer experience
- Higher productivity
- Reduced risk
- Ability to act on opportunities
- Increased revenue
- Competitive advantage gained from insights on customers, products, processes and opportunities

2.5.4 Planning for better data

View data as a product and plan for quality throughout its lifecycle. Planning is a collaboration between business and IT as it involves architecture, modelling and other design functions.

2.5.5 Metadata and Data Management

Includes:

- Business Metadata (Ch 12)
- Technical Metadata (Ch 12)
- Operational Metadata (Ch 12)
- Metadata embedded in: (Ch 4-11)
    - data architecture
    - data models
    - Data Security requirements

- Data integration standards
- Data operational processes

2.5.6 Data Management is Cross Functional

Data is managed in different places within the organisation by teams that have responsibility for different phases of the lifecycle.

2.5.7 Establishing an Enterprise Perspective

Data is a 'horizontal' as it moves across all the 'verticals' (sales, marketing, operations) of the organisation. Stakeholders assume the organisation's data should be coherent, and the goal of managing data is the make the disparate data originating from different places fit together in common sense ways so that it is usable by a wide range of consumers.

2.5.8 Accounting for other perspectives

People who create data must be aware others will use it later. Plan around other potential uses of the data to account for legal and compliance regulations and prevent future misuse.

2.5.9 The Data Lifecycle

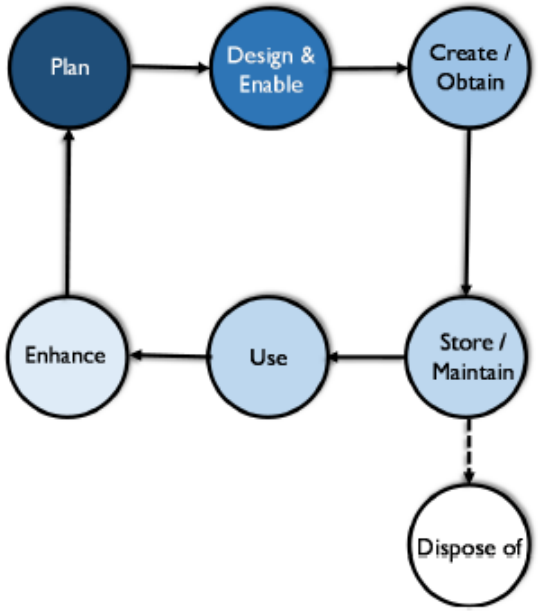


Figure 2 Data Lifecycle Key Activities

The data lifecycle is based on the product lifecycle. Throughout its lifecycle data may be cleansed, transformed, merged, enhanced or aggregate, and new data may be created. Managing data involves interconnected processes aligned with the lifecycle.

Data also has lineage, from point of origin to point of usage, which must be documented.

Implications of data lineage on the lifecycle:

- **Creation and usage are the most critical points in the data** lifecycle: it costs money to produce data and it only has value when it is used
- **Data Quality must be managed throughout the** lifecycle (see Chapter 13)
- **Metadata Quality must be managed through the data lifecycle** in the same way as the quality of other data (See Chapter 12)
- **Data Security must be managed throughout the data lifecycle:** Protection from creation to disposal (see Chapter 7)

- **Data management efforts should focus on the most critical data**

2.5.10 Different types of data

Classify the data to be managed as different types require different processes. Tools of data management focus on classification and control.

- By type:
  - Transactional data
  - Refence Data
  - Master Data
  - Metadata
  - Category data
  - Resource data
  - Detailed transaction data
- By content:
  - Data domains
  - Subject areas
- By format
- By the level of protection which the data requires

2.5.11 Data and Risk

- Low quality data represents risk as it is not right
- Data may be misunderstood and misused
- Information gaps
- Data privacy regulations
- Stakeholders requiring privacy may be broader than previously thought

2.5.12 Data Management and technology

Data requirements aligned with business should drive decisions about technology.

2.5.13 Effective data management requires Leadership and Commitment

A Chief Data Officer (CDO) leads data management initiatives and ensures data management is business driven instead of IT driven. The CDO leads cultural change required for the organisation to have a more strategic approach to its data.

2.6 Data Management strategy

A strategic plan is a high-level course of action to achieve high level goals. Should address all knowledge areas of the DAMA-DMBOK.

Components of a data management strategy:

- A compelling vision for data management
- A summary business case for data management with examples
- Guiding principles, values and management perspectives
- The mission and long term directional goals of data management
- Proposed measures of data management success
- Short term (12-24 months) Data Management program objectives that are SMART (Specific, Measurable, Actionable, Realistic, Time-bound)
- Description of Data Management roles, organisations and their responsibilities
- Descriptions of the Data Management program components and initiatives

- Prioritised program of work with scope boundaries
- Draft implementation roadmap with projects and action items

Deliverables:

- **Data Management Charter:** Overall vision, business case, goals, principles, measures of success, critical success factors, risks, operating model
- **Data Management Scope Statement:** Goals and objectives for usually 3 years, roles, organisations and individual leaders accountable
- **Data Management Implementation Roadmap:** Specific programs, projects, tasks and milestones

3 Data Management Frameworks

A framework helps to understand the data management comprehensively and see the relationships between component pieces. Frameworks are developed at different levels of abstraction and provide a range of perspectives.

Five models are presented in the DMBOK:

- The **Strategic Alignment Model** and the **Amsterdam Information Model** show high level relationships that influence how the organisation manages data
- The **DAMA DMBOK Framework** (DAMA Wheel, Hexagon and Context Diagram) describes Data Management Knowledge Areas and explains their visual representation within the DMBOK.
- The final two are rearrangements of the DAMA Wheel.

3.1 Strategic Alignment Model (Henderson and Venkatraman, 1999)

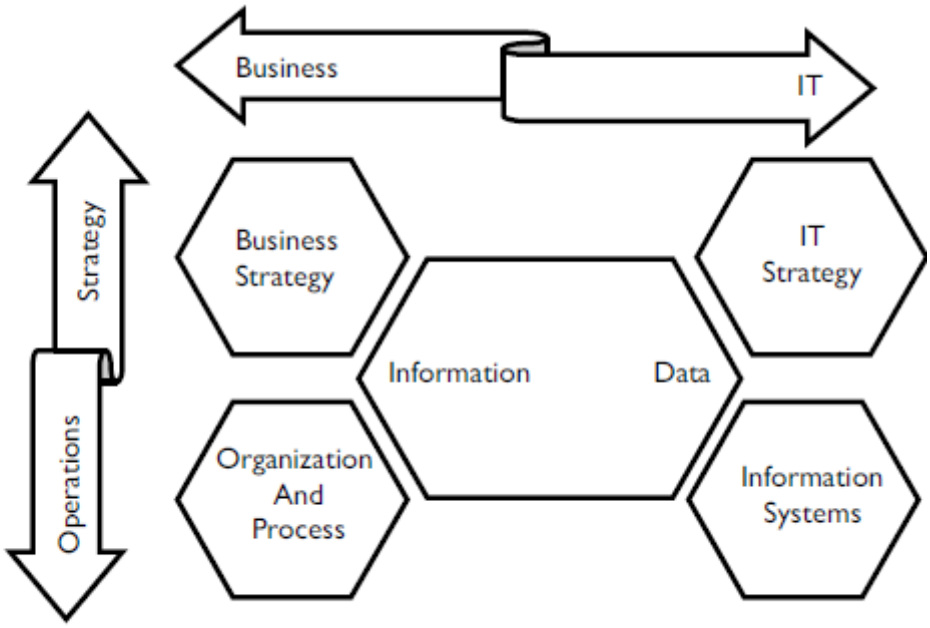
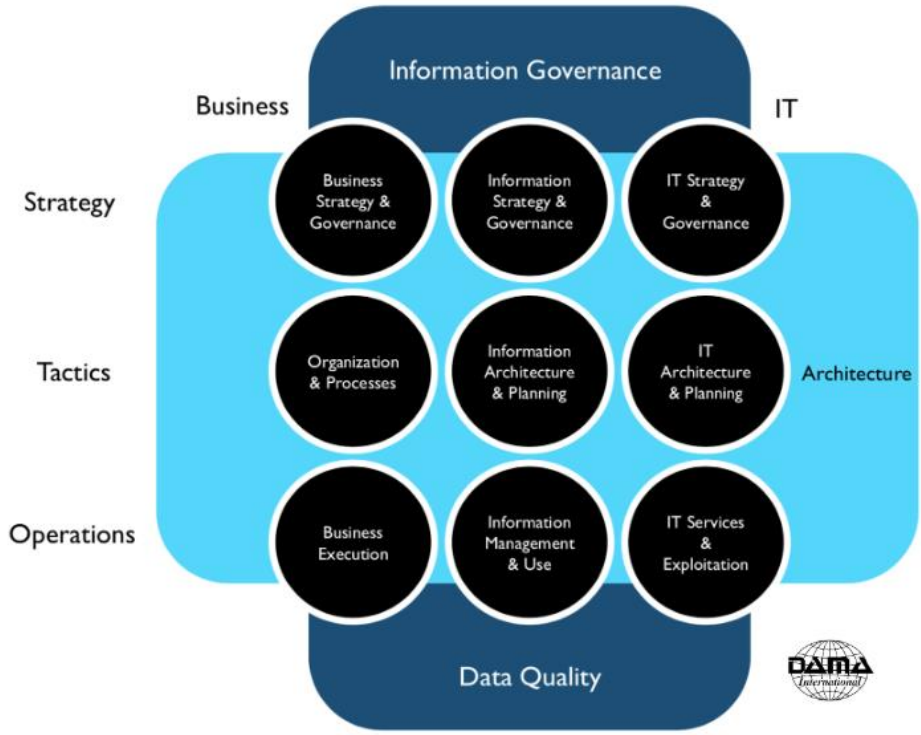


Figure 3 Strategic Alignment Model[12]

3.2 Amsterdam Information Model (Abcouwer, Maes and Truijens, 1997)



3.3 The DAMA-DMBOK Framework

Consists of three visuals:

- **The DAMA Wheel:** Governance in the centre for consistency within and balance between the Knowledge Areas

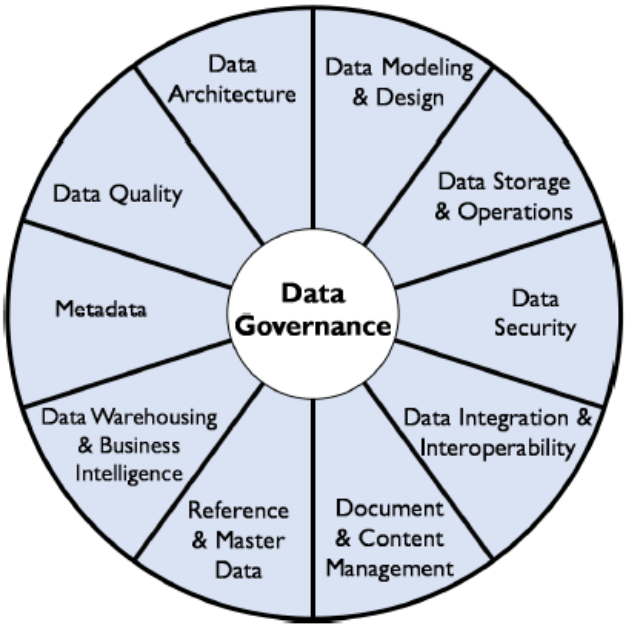


Figure 5 The DAMA-DMBOK2 Data Management Framework (The DAMA Wheel)

- **The Environmental Factors Hexagon:** Relationships between people, process and technology. Goals and principles in the centre.

Figure 6 DAMA Environmental Factors Hexagon

The Basic Environmental Elements are:

1. *Goals and Principles: The directional business goals of each function and the fundamental principles that guide performance of each function.*
2. *Activities: Each function is composed of lower level activities. Some activities are grouped into sub-activities. Activities are further decomposed into tasks and steps.*
3. *Deliverables: The information and physical databases and documents created as interim and final outputs of each function. Some deliverables are essential, some are generally recommended, and others are optional depending on circumstances.*
4. *Roles and Responsibilities: The business and IT roles involved in performing and supervising the function, and the specific responsibilities of each role in that function. Many roles will participate in multiple functions.*

The supporting Environmental Elements are:

5. *Techniques: Common and popular methods and procedures used to perform the processes and produce the deliverables. Practices and Techniques may also include common conventions, best practice recommendations, and alternative approaches without elaboration.*
6. *Tools: Categories of supporting technology (primarily software tools), standards and protocols, product selection criteria and common learning curves.*
7. *Organization and Culture: These issues might include:*
    - *Management Metrics measures of size, effort, time, cost, quality, effectiveness, productivity, success, and business value.*
    - *Critical Success Factors.*
    - *Reporting Structures.*
    - *Contracting Strategies.*
    - *Budgeting and Related Resource Allocation Issues.*
    - *Teamwork and Group Dynamics.*
    - *Authority and Empowerment.*
    - *Shared Values and Beliefs.*
    - *Expectations and Attitudes.*
    - *Personal Style and Preference Differences.*
    - *Cultural Rites, Rituals and Symbols.*

- *Organizational Heritage.*
- *Change Management Recommendations*.

Data Management Functions	Goals and Principles	Activities	Primary Deliverables	Roles and Responsibilities	Technology	Practices and Techniques	Organization and Culture
Data Governance							
Data Architecture Management							
Data Development							
Data Operations Management							
Data Security Management							
Reference and Master Data Management							
Data Warehousing and Business Intelligence Management							
Document and Content Management							
Meta-data Management							
Data Quality Management							

- **The Knowledge Area Context Diagram:**

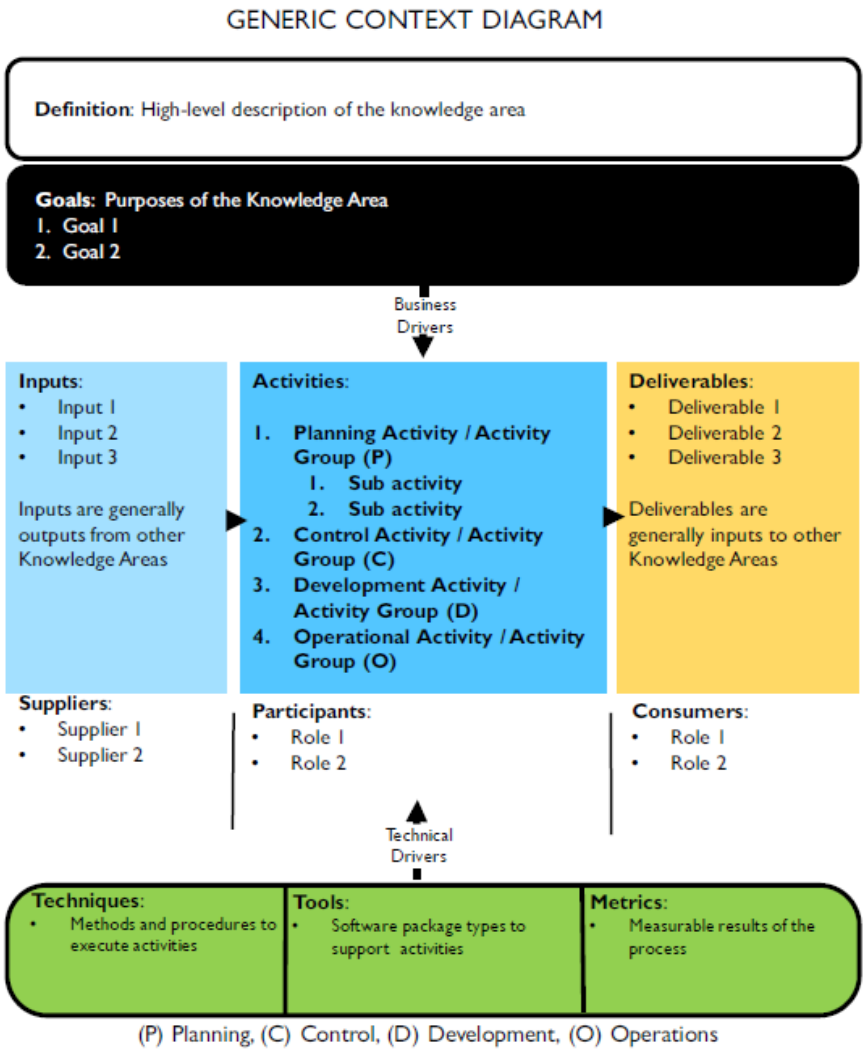


Figure 7 Knowledge Area Context Diagram

- Definition
- Goals
- Activities that drive goals classified in 4 phases:
  - (P) Planning – Strategic and tactical
  - (D) Development activities – organised around system lifecycle
  - (C) Control Activities – ensure quality, integrity, reliability, security
  - (O) Operational Activities – support systems and processes
- Inputs
- Deliverables
- Roles and Responsibilities
- Suppliers
- Consumers
- Participants
- Tools
- Techniques
- Metrics

3.4 DMBOK Pyramid (Aiken)

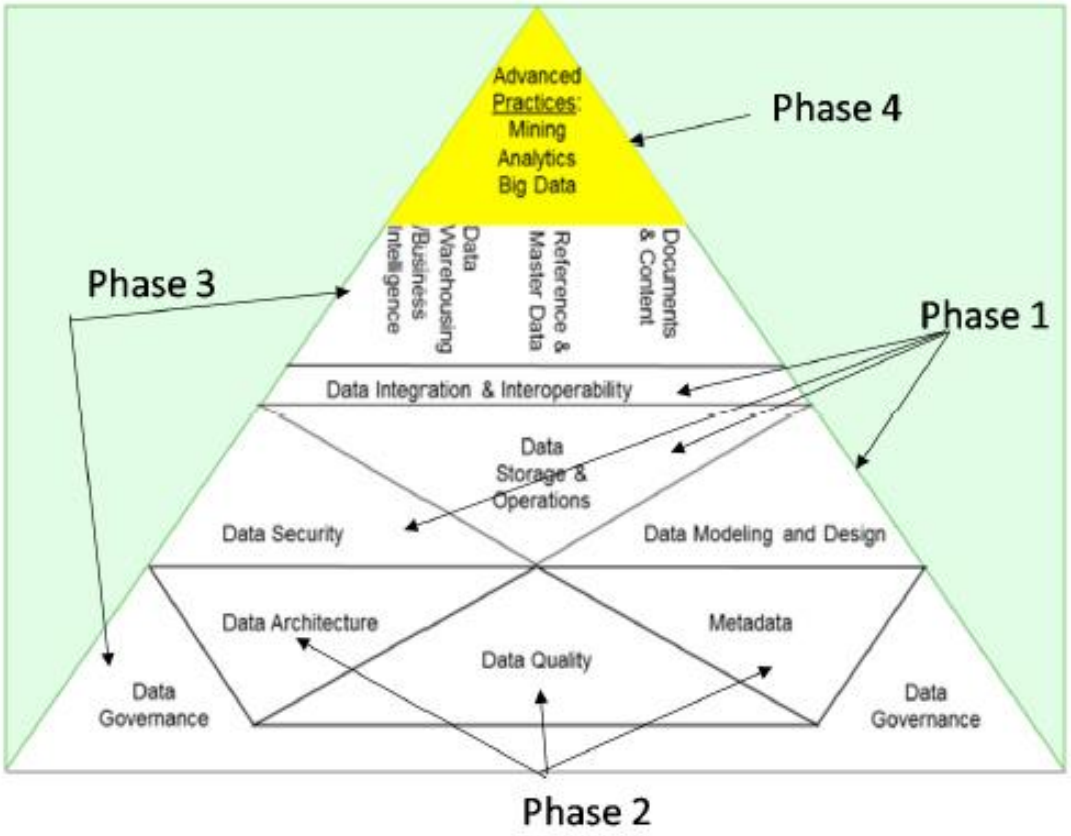


Figure 8 Purchased or Built Database Capability[17]

Describes how an organisation progresses to Data Management maturity:

3.5 DAMA Data Management Framework Evolved

Sue Geuens explored the dependencies between the knowledge areas, and developed fig 9, DAMA Functional Area Dependencies.

- BI/Analytics depend on all other functions
- DW and Master data depend on feeding systems/applications
- Depend on reliable data quality, design and integration
- Governance is the foundation on which all functions are dependent

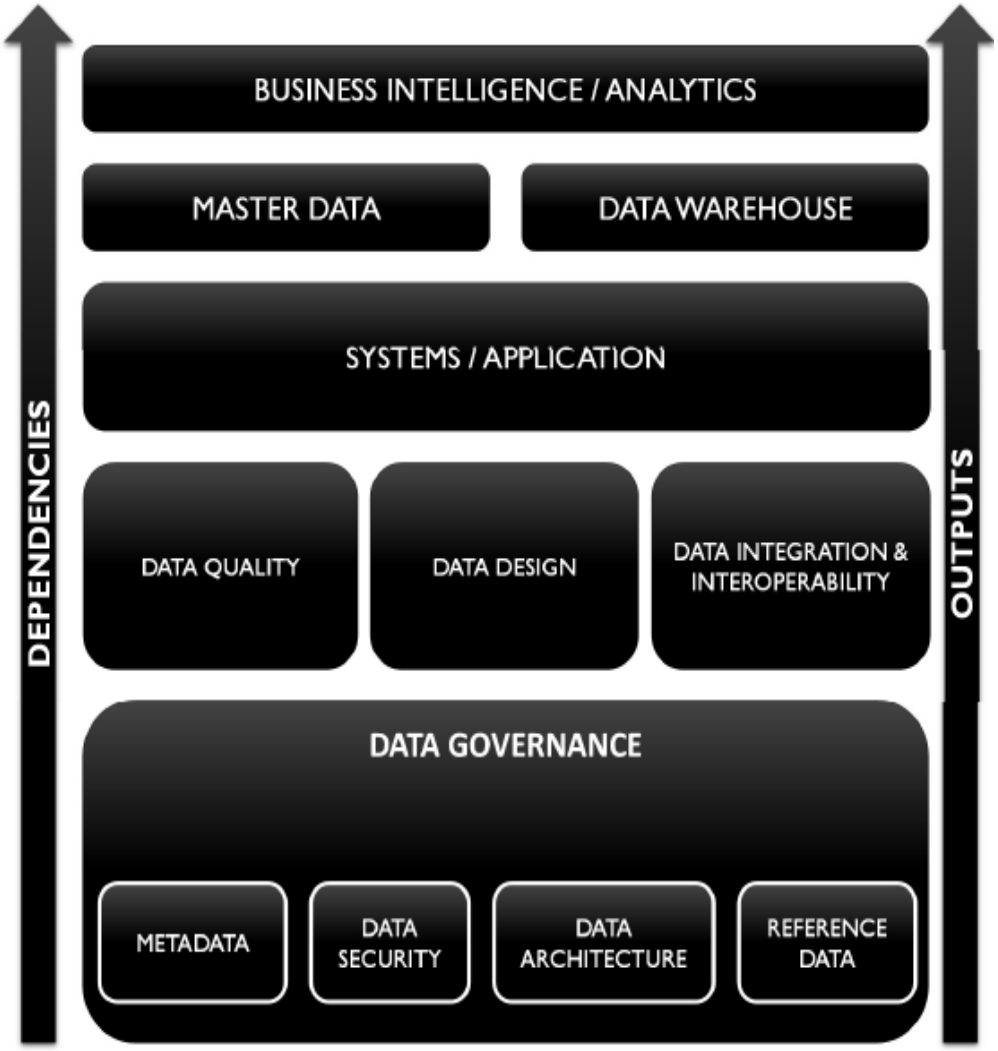


Figure 9 DAMA Functional Area Dependencies

DAMA Data Management Function Framework:

- Guiding purpose of Data Management
- Deriving value requires lifecycle management – centre of diagram
- Foundational activities support the lifecycle
- A data governance program enables the organisation to be data driven

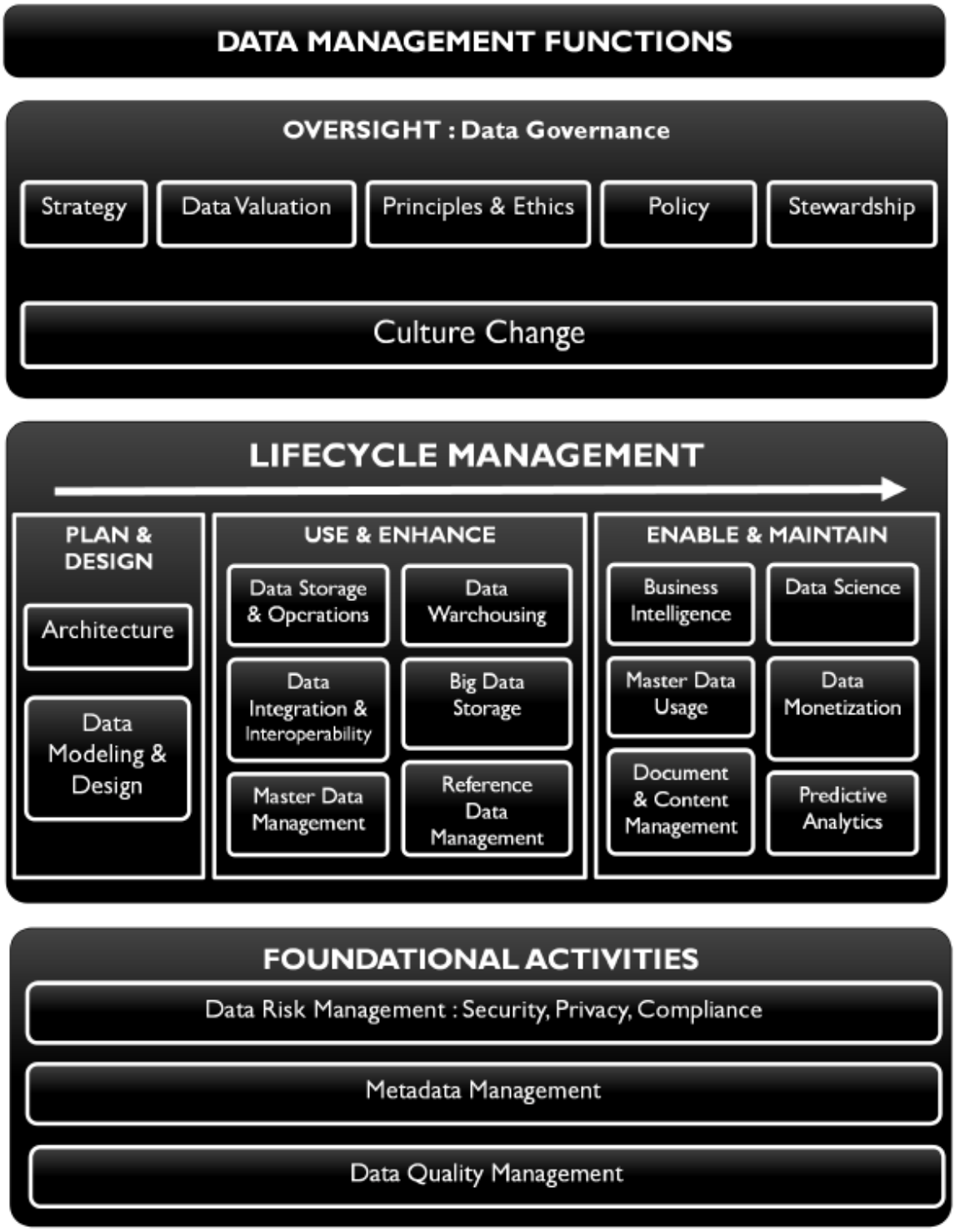


Figure 10 DAMA Data Management Function Framework

Evolution of the DAMA Wheel:

Core activities surrounded by lifecycle and usage activities, within the strictures of governance:

- Core activities in centre
- Surrounded by lifecycle and usage activities
- Lifecycle management: Plan and Enable
- Uses come from lifecycle activities
- Data Governance provides oversight and containment

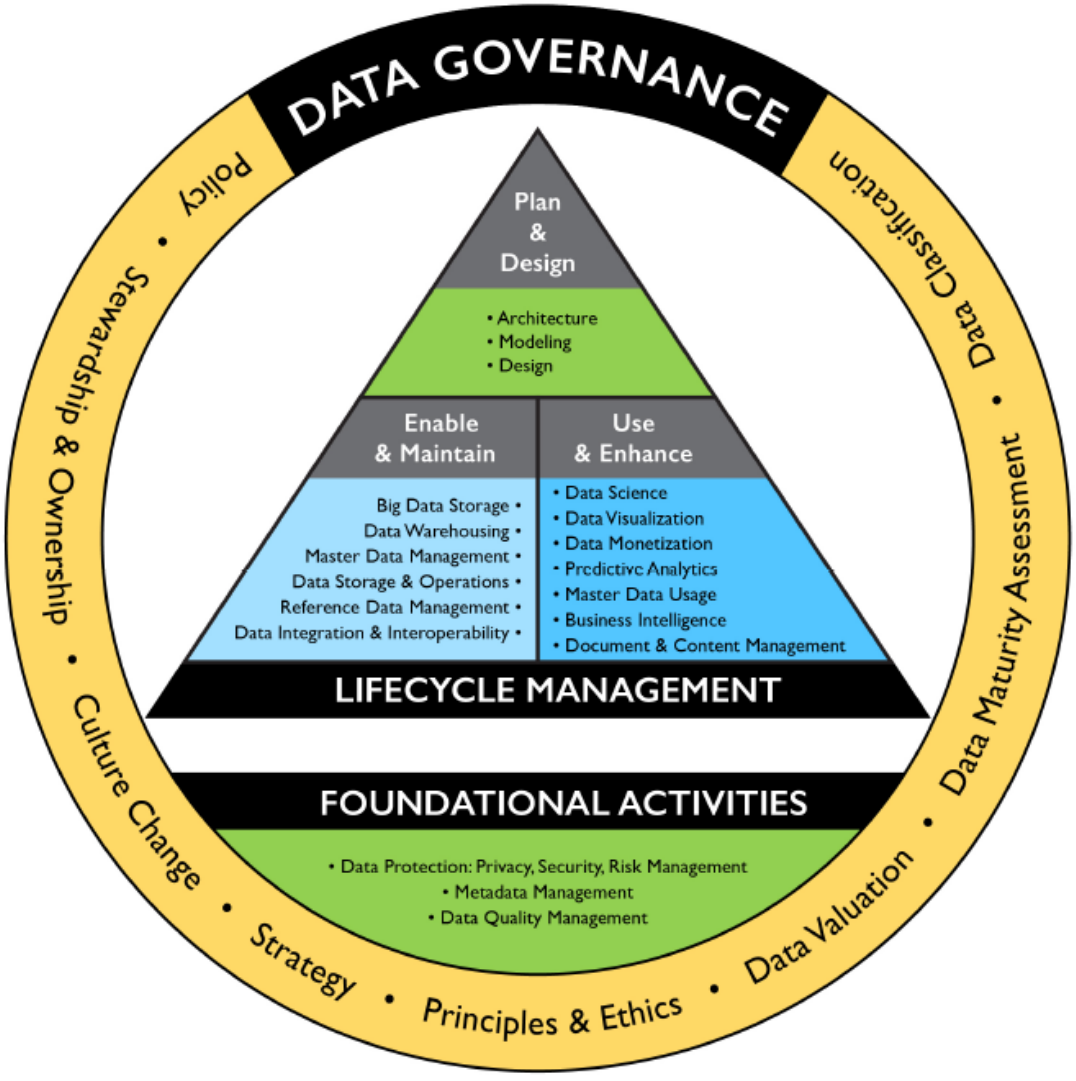


Figure 11 DAMA Wheel Evolved

DM Functions and Data Lifecycle Management

Relationships with additional content of the knowledge areas. Data management enables organisations to get value from their data. This require data lifecycle management, and these activities are in the centre of the diagram.

4 DAMA and the DMBOK

DAMA was founded to address the need for reliable data management practices.

The DMBOK supports DAMA's mission by:

- Providing a functional framework for the implementation of enterprise data management practices
- Enabling a common vocabulary for data management practices
- Serving as a fundamental guide for the CDMP exams

Knowledge Areas describe the scope and context of data management activities. They intersect with each other as data moves horizontally within an organisation.

1. **Data Governance** provides direction and oversight for data management by establishing a system of decision rights over data that accounts for the needs of the enterprise. (Ch 3 – **11%**)
2. **Data Architecture** defines the blueprint for managing data assets by aligning with organisational strategy to establish strategic data requirements and designs to meet these requirements. (Ch 4 – **6%**)
3. **Data Modelling and Design** is the process of discovering, analysing, representing and communicating data requirements in a precise form called a Data Model. (Ch 5 – **11%**)
4. **Data Storage and Operations** includes the design, implementation and support of stored data to maximise its value. Operations provide support throughout the data lifecycle from planning to disposal of data. (Ch 6 – **6%**)
5. **Data Security** ensures that data privacy and confidentiality are maintained, that data is not breached, and that data is accessed appropriately. (Ch 7 – **6%**)
6. **Data Integration and Interoperability** includes processes related to the movement and consolidation of data within and between data stores, applications and organisations. (Ch 8 – **6%**)
7. **Document and Content Management** includes planning, implementation and control activities used to manage the lifecycle of data and information found in a range of unstructured media, especially documents needed to support legal and regulatory compliance requirements. (Ch 9 – **6%**)
8. **Reference and Master Data** includes ongoing reconciliation and maintenance of core shared data to enable consistent use across systems of the most accurate, timely and relevant version of the truth about essential business entities. (Ch 10 – **10%**)
9. **Data Warehousing and Business Intelligence** includes the planning, implementation and control processes to manage decision support data to enable knowledge workers to get value from data via analysis and reporting. (Ch 11 – **10%**)
10. **Metadata** includes planning, implementation and control activities to enable access to high quality, integrated Metadata, including definitions, models, data flows and other information critical to understanding data and the systems through which it is created, maintained and accessed. (Ch 12 – **11%**)
11. **Data Quality** includes planning and implementation of quality management techniques to measure, assess and improve the fitness of data for use within the organisation. (Ch 13 – **11%**)
12. **Data Handling Ethics** describes the central role that data ethics plays in making informed, socially responsible decisions about data and its uses. Awareness of the ethics of data collection, analysis and use should guide all data management professionals. (Ch 2 – **2%**)

13. **Big Data and Data Science** describes the technologies and business processes that emerge as our ability to collect and analyse large and diverse data sets increases. (Ch 14 – **2%**)
14. **Data Management Maturity Assessment** outlines an approach to evaluating and improving an organisation's data management capabilities. (Ch 15)
15. **Data Management Organisation and Role Expectations** provide best practices and considerations for organising data management teams and enabling successful data management practices. (Ch 16)
16. **Data Management and Organisational Change Management** describes how to plan for and successfully move through the cultural changes that are necessary to embed effective data management practices within and organisation. (Ch 17)

Data Governance provides oversight and containment

DMBOK Quote

- Data governance activities provide oversight and containment, through strategy, principles, policy, and stewardship.
- They enable consistency through data classification and data valuation

Why is this important from a Governance point of view?

- To ensure that Data remains a strategic imperative for the business
- The Data Risks are identified & managed

How will Data Governance do this?

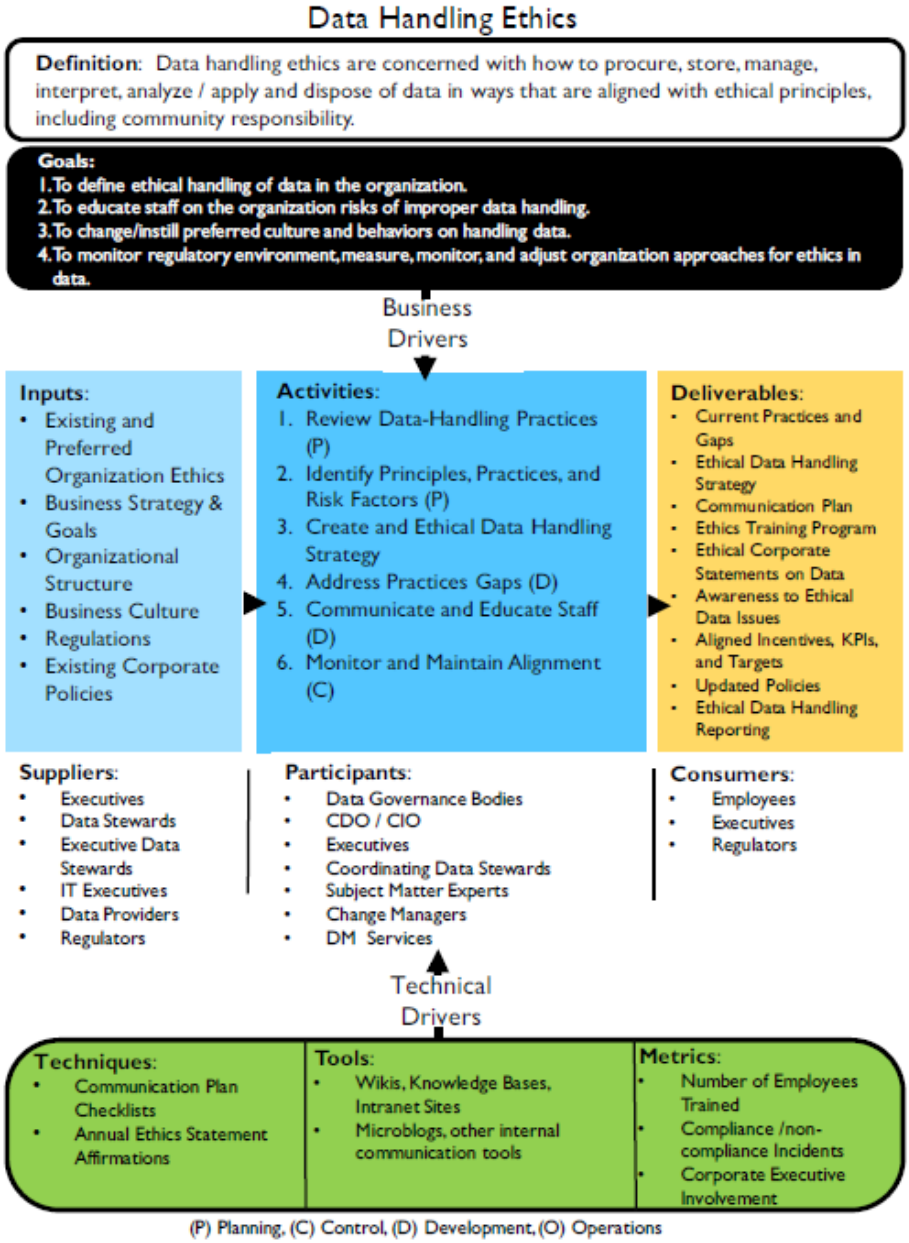
- Implement a Data Governance function and supporting deliverables

Data Handling Ethics

1 Introduction

Core Concepts:

- **Impact on people:**
  - Imperative to maintain the quality and reliability of individuals' data
  - Used to make decisions that impact people's lives
- **Potential for misuse:**
  - Negative effects on people and organisations
- **Economic value of data:**
  - Determine how and by whom that value of data can be accessed



2 Business Drivers

Ethics means “doing it right when no one is looking”.

Ethical handling of data is a competitive business advantage for an organisation:

- Increased trustworthiness and improved relationships – a competitive advantage
- Better relationships with stakeholders
- Reduced risk of misuse or a criminal breach
- Responsibilities when sharing data
- Organisation wide commitment to handling data ethically

Chief Data Officer, Chief Risk Officer, Chief Privacy Officer, Chief Analytics Officer focus on controlling risk by establishing acceptable practices for data handling.

3 Essential Concepts

3.1 Ethical Principles for Data

- **Respect for persons**: Treat people in a way that respects their dignity and autonomy as human individuals. Protect the dignity of those with “diminished Autonomy”.
- **Beneficence**: Firstly, do no harm. Secondly, maximise possible benefits and minimise possible harms.
- **No maleficence**: Minimise harm
- **Justice**: The fair and equitable treatment of people.

European Data Protection Supervisor, 2015 – the “engineering, philosophical, legal and moral implications” of developments in **data processing and Big Data**:

- Future-oriented regulation of data processing and respect for the rights to privacy and data protection.
- Accountable controllers who determine personal information processing
- Privacy conscious engineering and design of data processing products and devices
- Empowered individuals

3.2 Principles behind Data Privacy Law

Privacy regulations have been driven by human rights violations experienced during WW2 in Europe. The European Union’s Organisation for Economic Co-operation and Development (OECD) established 8 Guidelines and Principles for Fair Information Processing.

- Limitations on data collection
- An expectation that data will be of a high quality
- The requirement that when data is collected it will be for a specific purpose
- Limitations on data usage
- Security safeguards
- An expectation of openness and transparency
- The right of an individual to challenge the accuracy of data related to himself or herself
- Accountability of organisations to follow the guidelines

These underpin the GDPR:

3.2.1 General Data Protection Regulation of the EU

GDPR Principle	Description of Principle
Fairness, Lawfulness, Transparency	Personal data shall be processed lawfully, fairly, and in a transparent manner in relation to the data subject.
Purpose Limitation	Personal data must be collected for specified, explicit, and legitimate purposes, and not processed in a manner that is incompatible with those purposes.
Data Minimization	Personal data must be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed.
Accuracy	Personal data must be accurate, and where necessary, kept up-to-date. Every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purpose for which they are processed, are erased or rectified without delay.
Storage Limitation	Data must be kept in a form that permits identification of data subjects [individuals] for no longer than is necessary for the purposes for which the personal data are processed.
Integrity and Confidentiality	Data must be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures.
Accountability	Data Controllers shall be responsible for, and be able to demonstrate compliance with [these principles].

Other focusses for fair information practices include:

- Simplified consumer choice to reduce burden placed on consumers
- Recommendation to maintain comprehensive data management practices throughout the lifecycle
- A Do Not Track option
- Requirements for affirmative express consent
- Concerns regarding the data collection capabilities of large platform providers; transparency and clear privacy notices and policies
- Individuals' access to data
- Educating consumers about data privacy practices
- Privacy by Design

3.3 Online Data in an ethical context

- **Ownership of data:** Right to control own personal data
- **The right to be forgotten:** To have information about an individual erased from the web, particularly to adjust online reputation
- **Identity:** Identity must be correct, or one can opt for a private identity
- **Freedom of speech online:** Expression rather than online bullying or insulting

3.4 Risks of Unethical Data Handling Practices

Data should be measured against Data Quality dimensions such as accuracy and timeliness to ensure it is trustworthy. Unethical data practices include:

3.4.1 Timing

Omission or inclusion of data points based on timing, e.g., market manipulation of end-of-day stock price.

3.4.2 Misleading Visualisations

e.g. changing scale, comparing facts without clarifying their relationship

3.4.3 Unclear definitions or Invalid Comparison

- Provide context which informs meaning
- Statistical "smoothing" or over-training of a statistical model

3.4.4 Bias: an inclination of outlook

- Data collection for pre-defined results
- Biased use of data collected
- Hunch and search
- Biased sampling methodology
- Context and Culture

3.4.5 Transforming and Integrating Data

- Limited knowledge of data's origin and lineage
- Data of poor quality
- Unreliable metadata
- No documentation of data remediation history

3.4.6 Obfuscation / Redaction of Data

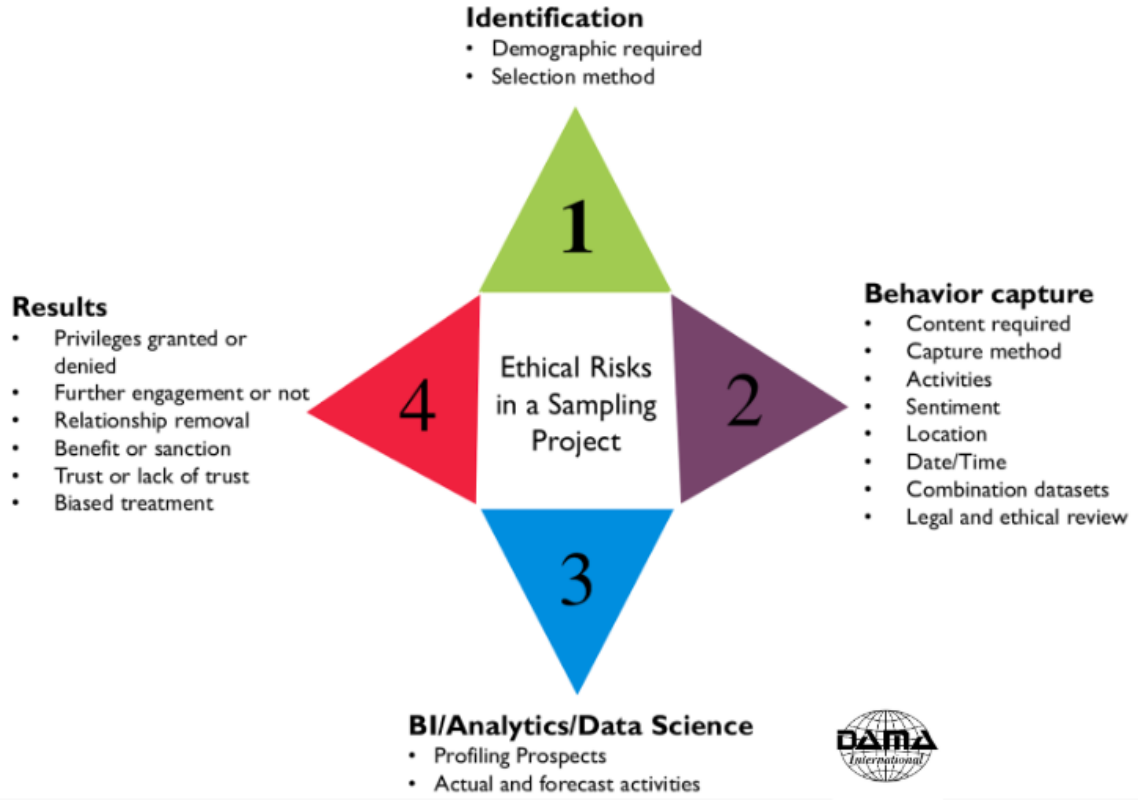
- Data aggregation
- Data Marking
- Data masking

3.5 Establishing an Ethical Data Culture

- **Review current state data handling practices:**
  - Understand degree of current practices to ethical and compliance drivers
  - Employees understanding of ethical implications of existing practices
- **Identify principles, practices and risk factors:**
- **Create an ethical data handling strategy and roadmap:** Ethical principles and expected behaviour related to data
  - **Values statements:** What the organisation believes in
  - **Ethical data handling principles:** How the organisation approaches challenges related to data
  - **Compliance framework:** Factors that drive organisational compliance obligations
  - **Risk assessment:** Likeliness of a specific problem occurring
  - **Training and communications:** Train on code of ethics. Communications must reach all employees
  - **Roadmap:** Timeline of activities that can be approved by management
  - **Approach to auditing and monitoring**
- **Adopt a socially responsible ethical risk model:** BI, analytics and data science responsible for data that describes:
  - Who people are (race, country or origin, religion)
  - What people do (Political, social, criminal activities)
  - How people live (Money, purchases, who they communicate with)
  - How people are treated (outcomes of analysis which privilege or prejudice them for future business)
- The risk model is used to determine:
  - Whether to execute the project
  - How to execute the project

- Actively identify potential risks
- Protect whistle-blowers
- Identify possible bias

3.5.1 Ethical Risk Model for Sampling Projects

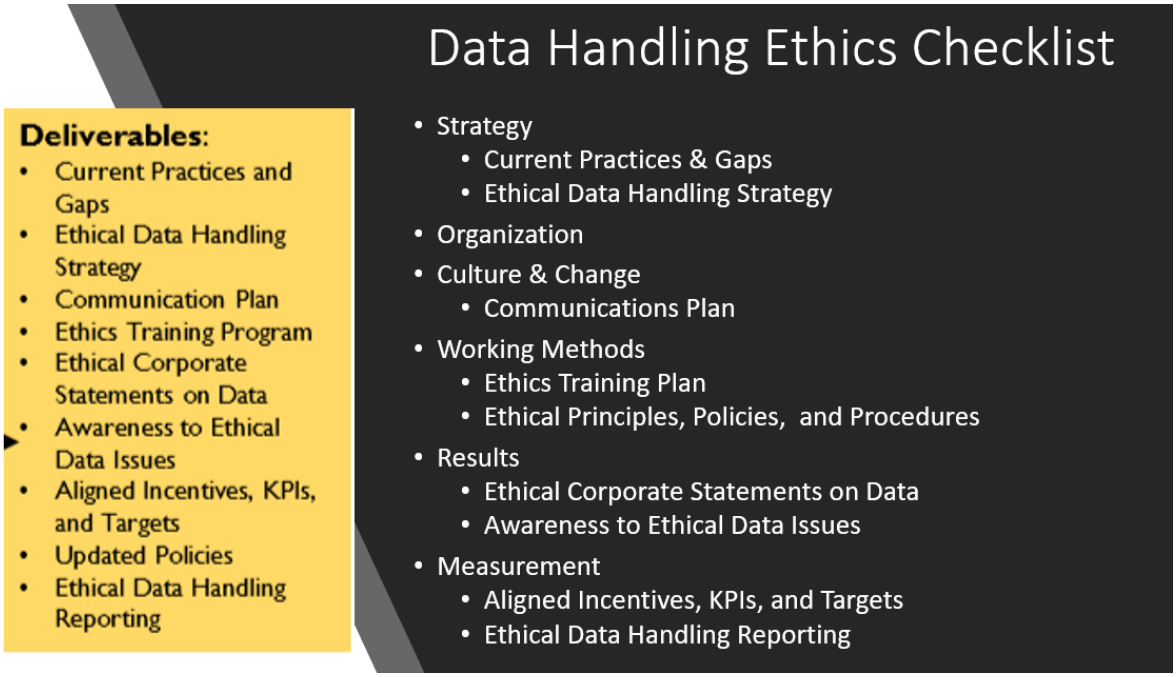


Address all potential ethical risks in the areas of consideration, with a particular focus on negative effects on customers or citizens

3.6 Data Ethics and Governance

Oversight falls under Data Governance and Legal counsel:

- Keep up-to-date on legal changes
- Ensure employees are aware of their obligations
- Data Governance must set standards and policies to provide oversight of data handling practices
- Particularly to review plans and decisions proposed by BI, analytics and Data Science
- CDMP Certification requires subscription to code of ethics


Data Handling Ethics Checklist
Deliverables:
Current Practices and Gaps
Ethical Data Handling Strategy
Communication Plan
Ethics Training Program
Ethical Corporate Statements on Data
Awareness to Ethical Data Issues
Aligned Incentives, KPIs, and Targets
Updated Policies
Ethical Data Handling Reporting
Strategy
Current Practices & Gaps
Ethical Data Handling Strategy
Organization
Culture & Change
Communications Plan
Working Methods
Ethics Training Plan
Ethical Principles, Policies, and Procedures
Results
Ethical Corporate Statements on Data
Awareness to Ethical Data Issues
Measurement
Aligned Incentives, KPIs, and Targets
Ethical Data Handling Reporting

Data Governance

1 Introduction

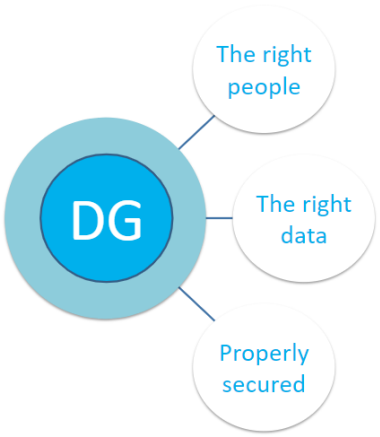
The purpose of Data Governance is to ensure that data is managed properly, according to policies and best practices:

- **Strategy:** Defining, communicating and driving execution of Data Strategy and Data Governance
- **Policy:** Setting and enforcing policies relating to data and Metadata management access, usage, security and quality
- **Standards and quality:** Setting and enforcing Data Quality and Data Architecture standards
- **Oversight:** Providing hands-on observation, audit and correction in key areas of quality, policy and data management **(Stewardship)**
- **Compliance:** Ensuring the organisation can meet the requirements of regulations
- **Issue management:** Identifying, defining, escalating and resolving issues related to data security, data access, data quality, regulatory compliance, data ownership, policy, standards, terminology or data governance procedures.
- **Data management projects:** Sponsoring efforts to improve data management practices.
- **Data asset valuation:** Setting standards and processes to define business value of assets

A Data Governance program develops policies and procedures, cultivates stewardship practices and organisational change management.

Organisational culture must be changed to value data and data management activities. Formal change management is required.

Three important things:



Data Governance provides oversight and containment

DMBOK Quote

- Data governance activities provide oversight and containment, through strategy, principles, policy, and stewardship.
- They enable consistency through data classification and data valuation

Why is this important from a Governance point of view?

- To ensure that Data remains a strategic imperative for the business
- The Data Risks are identified & managed

How will Data Governance do this?

- Implement a Data Governance function and supporting deliverables

1.1 Business Drivers

Reducing risk:

- **General risk management:** finance, reputation, legal e-discovery, regulatory issues
- **Data security:** Protection of data assets
- **Privacy:** Control of private, confidential and Personal Identifying Information (PII) through policy and compliance monitoring

Improving processes:

- **Regulatory compliance**
- **Data quality improvement**
- **Metadata management:** Business glossary and other Metadata
- **Efficiency in development projects:** SDLC improvements
- **Vendor management:** Contracts dealing with data

Data Governance is separate from IT governance, and is ongoing.

1.2 Goals and Principles

Goals:
1. Enable an organization to manage its data as an asset.
2. Define, approve, communicate, and implement principles, policies, procedures, metrics, tools, and responsibilities for data management.
3. Monitor and guide policy compliance, data usage, and management activities.

To achieve these goals a DG program must be:

- **Sustainable:** Ongoing, and depends on business leadership, sponsorship and ownership
- **Embedded:** DG activities incorporated into development methods, use of analytics, management of Master Data and risk management
- **Measured:** to show positive financial impact

Foundational principles:

- **Leadership and strategy:** Committed leadership, driven by enterprise business strategy
- **Business-driven:** DG is a business program and must govern IT decisions relating to data
- **Shared responsibility:** Across all Data Management Knowledge Areas
- **Multi-layered:** All levels from local to enterprise
- **Framework-based:** establish an operating framework
- **Principle-based:** Principles are the basis of policies. Principles can mitigate resistance.

1.3 Essential concepts

Ensuring data is managed without directly executing data management. **Separation of duties between oversight and execution.**

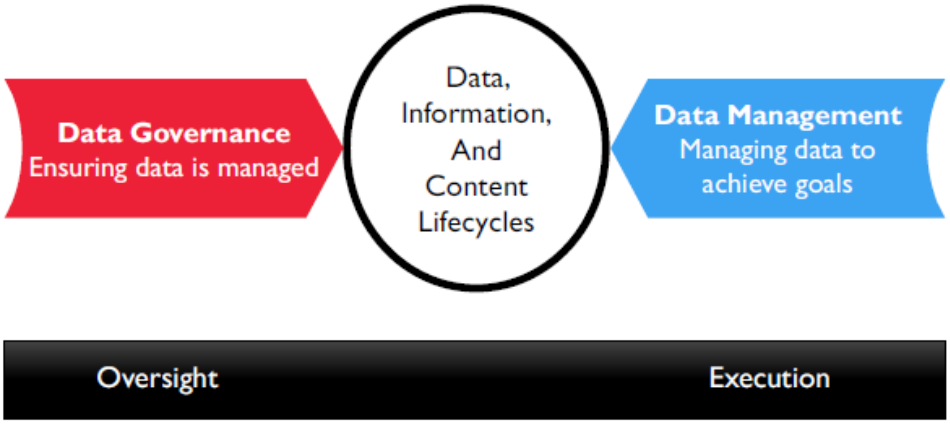


Figure 15 Data Governance and Data Management

1.3.1 Data-centric Organisation

A data-centric organisation values data as an asset and manages data through all phases of its lifecycle, including project development and ongoing operations.

Principles:

- Manage data as a corporate asset
- Incentivise Data management best practices across the organisation
- Enterprise data strategy is aligned to overall business strategy
- Continually improve data management processes.

1.3.2 Data Governance Organisation

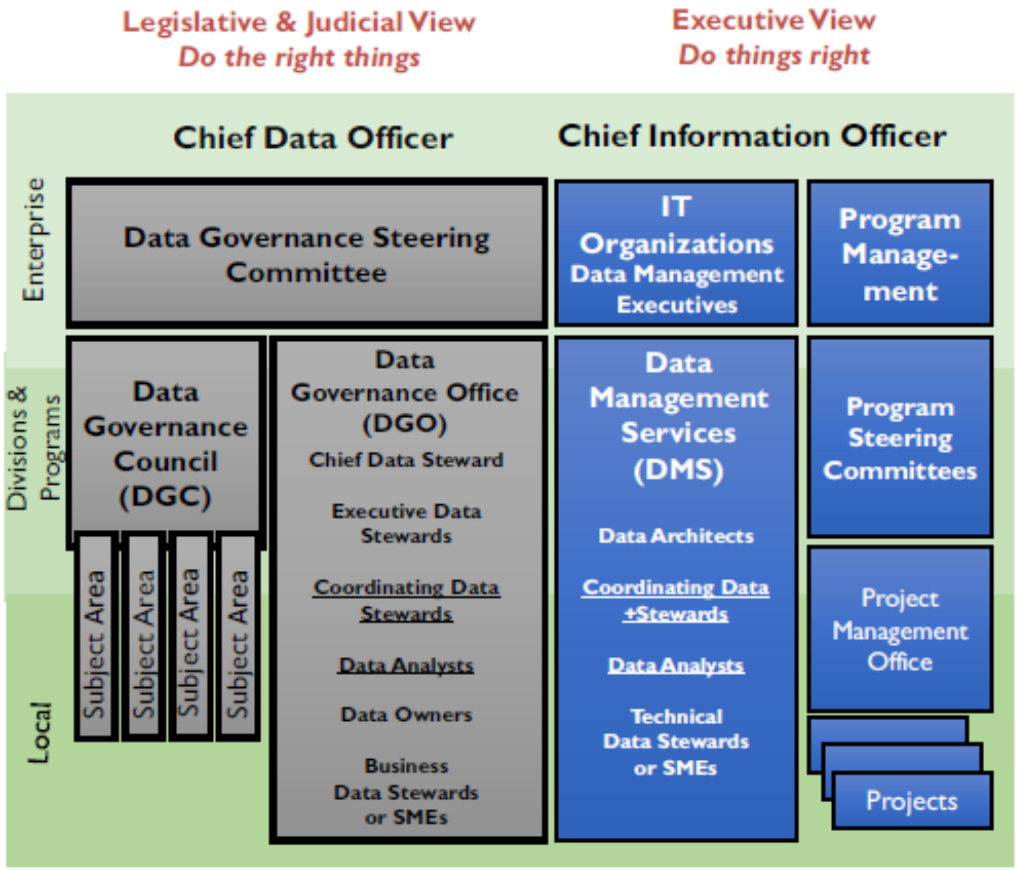


Figure 16 Data Governance Organization Parts

Data Governance Body	Description
Data Governance Steering Committee	The primary and highest authority organization for data governance in an organization, responsible for oversight, support, and funding of data governance activities. Consists of a cross-functional group of senior executives. Typically releases funding for data governance and data governance-sponsored activities as recommended by the DGC and CDO. This committee may in turn have oversight from higher-level funding or initiative-based steering committees.
Data Governance Council (DGC)	Manages data governance initiatives (e.g., development of policies or metrics), issues, and escalations. Consists of executive according to the operating model used. See Figure 17.
Data Governance Office (DGO)	Ongoing focus on enterprise-level data definitions and data management standards across all DAMA-DMBOK Knowledge Areas. Consists of coordinating roles that are labelled as *data stewards* or *custodians,* and *data owners*.
Data Stewardship Teams	Communities of interest focused on one or more specific subject-areas or projects, collaborating or consulting with project teams on data definitions and data management standards related to the focus. Consists of business and technical data stewards and data analysts.
Local Data Governance Committee	Large organizations may have divisional or departmental data governance councils working under the auspices of an Enterprise DGC. Smaller organizations should try to avoid such complexity.

1.3.3 Data Governance Operating Model Types

(More detail in Chapter 16)

- **Centralised**: One Central EIM / DG Team with Councils within the Business
- **Decentralised/Replicated**: One DG Team per Business Organization Structure
- **Hybrid**: One Central EIM and By Business Organization Structure
- **Federated**: Unite in a federation – individual autonomy. Centralised strategy, functions in the BUs
- **Self-organising/Network**: Non-Invasive Model based on RACI

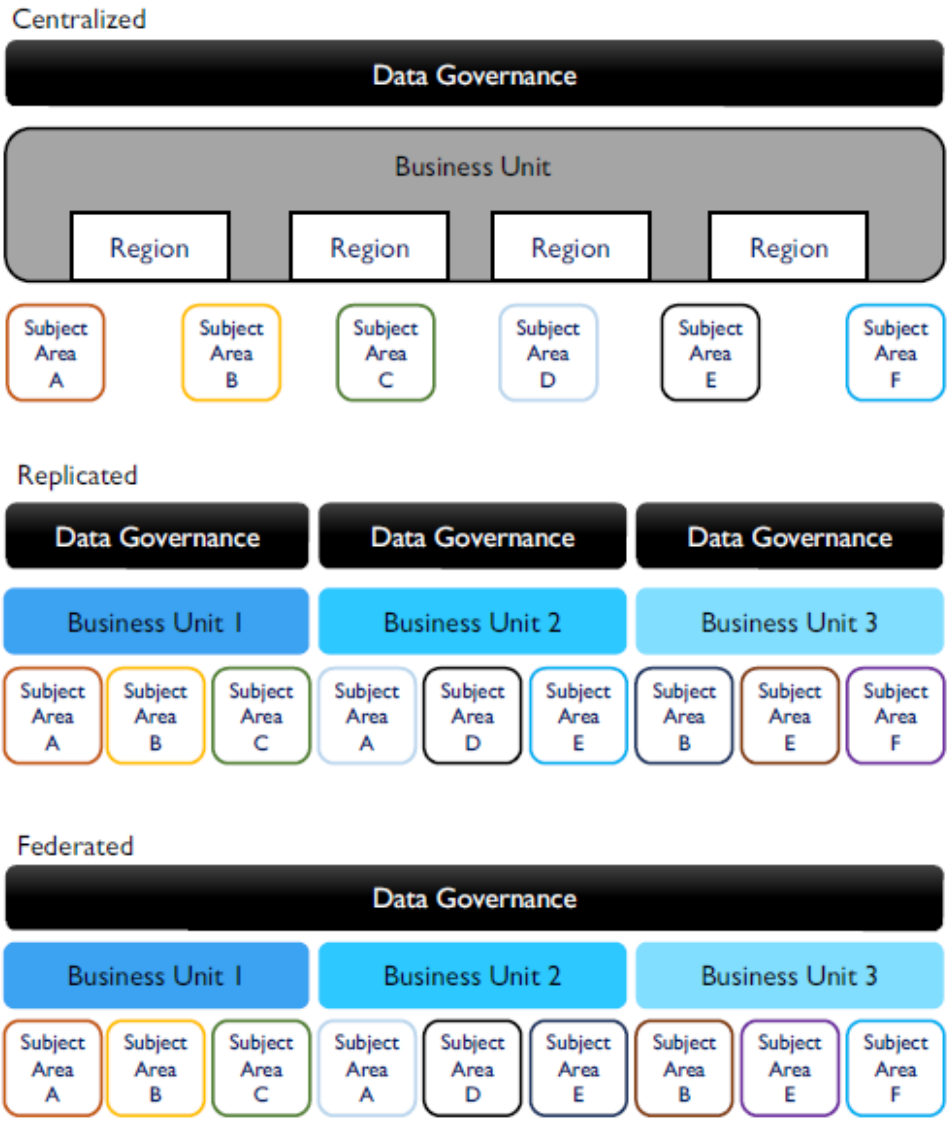


Figure 17 Enterprise DG Operating Framework Examples[27]

1.3.4 Data Stewardship

A **steward** is a person whose job it is to manage the property of another person. Effective data stewards are accountable and responsible for data processes that ensure effective control and use of data assets.

Activities include:

- Creating and managing core Metadata

- Documenting rules and standards
- Managing data quality issues
- Executing operational data governance issues

1.3.5 Types of Data Steward

A steward manages the property of another person. Data stewards manage the data assets on behalf of others and in the best interests of the organisation. (McGilvray, 2008). Data Stewards are accountable and responsible for data governance and have a portion of their time dedicated to these activities.

Types of Data Stewards:

- **Chief Data Stewards**: Chair DG bodies, act as CDO, be Executive Sponsors
- **Executive Data Stewards**: Senior managers who serve on the Data Governance Council
- **Enterprise Data Stewards**: Oversight of a data domain across business functions
- **Business Data Stewards**: business professionals, subject matter experts
- **A Data Owner**: a business data steward with approval authority for decisions within domain
- **Technical Data Stewards**: IT professionals operating in one of the knowledge areas e.g. DBAs, BI Specialists, Data Integration specialists, Data Quality analysts, metadata administrators
- **Coordinating Data Stewards**: Lead teams of business and technical data stewards in discussions across teams and with Executive Data Stewards.

Data Principles: Data Stewards:

- **Enterprise data must be modelled:**
  - modelled, named and defined according to standards across all business divisions
  - Effort made by management to share data – not maintain redundant data
  - Originating data stewards recognise needs of downstream processes
- **Enterprise data must be maintained close to the source:**
  - Create and maintain as close to the source as possible
  - Data Quality standards applied to approved reliability levels as defined by business units
- **Enterprise data must be safe and secured:**
  - In all electronic formats must be safeguarded on requirements and compliance guidelines
  - Guidelines are determined by business stewards of Enterprise data
  - Backup and recovery measures
- **Enterprise data must be accessible**:
  - Enterprise data and metadata shall be readily available and accessible to all except where restricted
  - Business stewards of enterprise data are responsible for defining the types of individuals, and levels of access privileges
  - Information Security will be responsible for the implementation of proper security controls
- **Metadata will be recorded and utilised:**
  - All projects utilise defined metadata for data naming, data modelling and database design.
  - Data Management is responsible for the metadata program

1.3.6 Data Policies

- Data Policies are directives that codify principles and management intent into fundamental rules governing the creation, acquisition, integrity, security, quality and use of data and information.
- Global
- Describe the "What" of data Governance (Standards and procedures describe "How")
- Should be few data polices, and should be stated briefly and directly

1.3.7 Data Asset Valuation

The process of understanding and calculating the economic value of data to the organisation.

Data sets are not interchangeable (fungible) between organisations. How an organisation gets value from customer data can be a competitive differentiator.

The following phases of the data lifecycle (acquiring, storing, administering and disposing) involve costs. Data only brings value when it is used but incurs risk management costs.

Ways to measure value:

- **Replacement cost**: of data lost in a breach or disaster
- **Market value:** as a business asset at the time of a merger or acquisition
- **Identified opportunities**: the income to be gained
- **Selling data**: as a package, or the insights to be gained
- **Risk cost**: potential penalties, remediation costs and litigation expenses

Table 5 Principles for Data Asset Accounting

Principle	Description
Accountability Principle	An organization must identify individuals who are ultimately accountable for data and content of all types.
Asset Principle	Data and content of all types are assets and have characteristics of other assets. They should be managed, secured, and accounted for as other material or financial assets.
Audit Principle	The accuracy of data and content is subject to periodic audit by an independent body.
Due Diligence Principle	If a risk is known, it must be reported. If a risk is possible, it must be confirmed. Data risks include risks related to poor data management practices.
Going Concern Principle	Data and content are critical to successful, ongoing business operations and management (i.e., they are not viewed as temporary means to achieve results or merely as a business by-product).
Level of Valuation Principle	Value the data as an asset at a level that makes the most sense, or is the easiest to measure.
Liability Principle	There is a financial liability connected to data or content based on regulatory and ethical misuse or mismanagement.
Quality Principle	The meaning, accuracy, and lifecycle of data and content can affect the financial status of the organization.
Risk Principle	There is risk associated with data and content. This risk must be formally recognized, either as a liability or through incurring costs to manage and reduce the inherent risk.
Value Principle	There is value in data and content, based on the ways these are used to meet an organization's objectives, their intrinsic marketability, and/or their contribution to the organization's goodwill (balance sheet) valuation. The value of information reflects its contribution to the organization offset by the cost of maintenance and movement.

2 Data Governance Activities

2.1 Define Data Governance for the Organisation

- DG must support business strategy and goals
- Enterprise Data Strategy is informed by business strategy and goals
- DG enables shared responsibility for data decisions
- Clear understanding of what and who are governed and who is governing
- DG is most effective when it is an enterprise effort

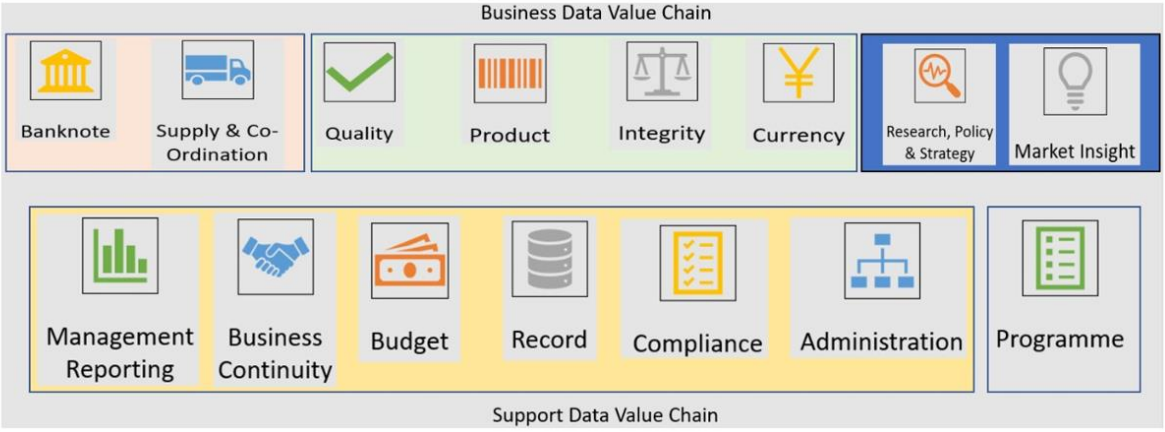
2.2 Perform Readiness Assessment

- **Data management maturity**: Understand what the organisation does with data, measure current management capabilities and capacity.
- **Capacity to change**: Measure capacity of organisation to change, and Identify possible resistance points
- **Collaborative readiness**: measures collaboration across functional areas
- **Business alignment**: how well data use aligns with business strategy

2.3 Perform Discovery and Business Alignment

- **Discovery Activity**
  - Identify and assess effectiveness of existing policies and guidelines (Risks, behaviours, implementation)
  - Identify opportunities for DG to improve usefulness of data
- **Business Alignment** attaches business benefits to DG program elements
- **Data Quality analysis**
  - Identify issues and risks associated with poor quality data
  - Identify business processes at risk from poor quality data
  - Financial and other benefits from creating a DQ program as part of DG
- Assessment of **data management practices**
- Derive a list of **DG requirements** which will drive **DG strategy** and tactics

Example of a Data Estate: a Central Bank



2.4 Develop Organisational Touch Points

Touch points that support alignment of an enterprise data governance and data management outside the direct authority of the CDO.

- **Procurement and contracts**: Enforce standard contract language
- **Budget and funding**: prevent duplicate acquisitions and ensure optimisation of data assets
- **Regulatory compliance**: CDO understands and works within regulatory requirements. Requires ongoing monitoring
- **SDLC/development framework**: identifies control points where enterprise policies, processes and standards can be developed

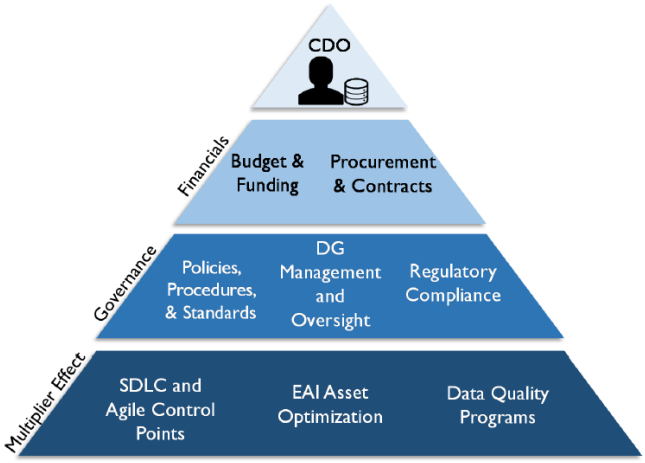


Figure 18 CDO Organizational Touch Points

2.5 Develop Data Governance Strategy

Defines the scope and approach to governance efforts. Deliverables:

- **Charter:** Business drivers, vision, mission and principles of data governance
- **Operating framework and accountabilities:** Structure and responsibilities
- **Implementation roadmap:** Timeframes
- **Plan for operational success:** Describe target state

2.6 Define the DG Operating Framework

Consider the following areas when constructing the organisations operating model:

- **Value of data to the organisation:** Important if the organisation sells data
- **Business model:** Decentralised, centralised, international
- **Cultural factors:** Acceptance of discipline and adaptability to change
- **Impact of regulation:** Level of regulation of the organisation

Layers of governance: determine where accountability resides for stewardship activities and data ownership.

The DG Operating framework defines:

- The interaction between Governance Organisation and people responsible for data management initiatives
- Engagement of change management initiatives to introduce DG
- Model for issue resolution pathways through governance

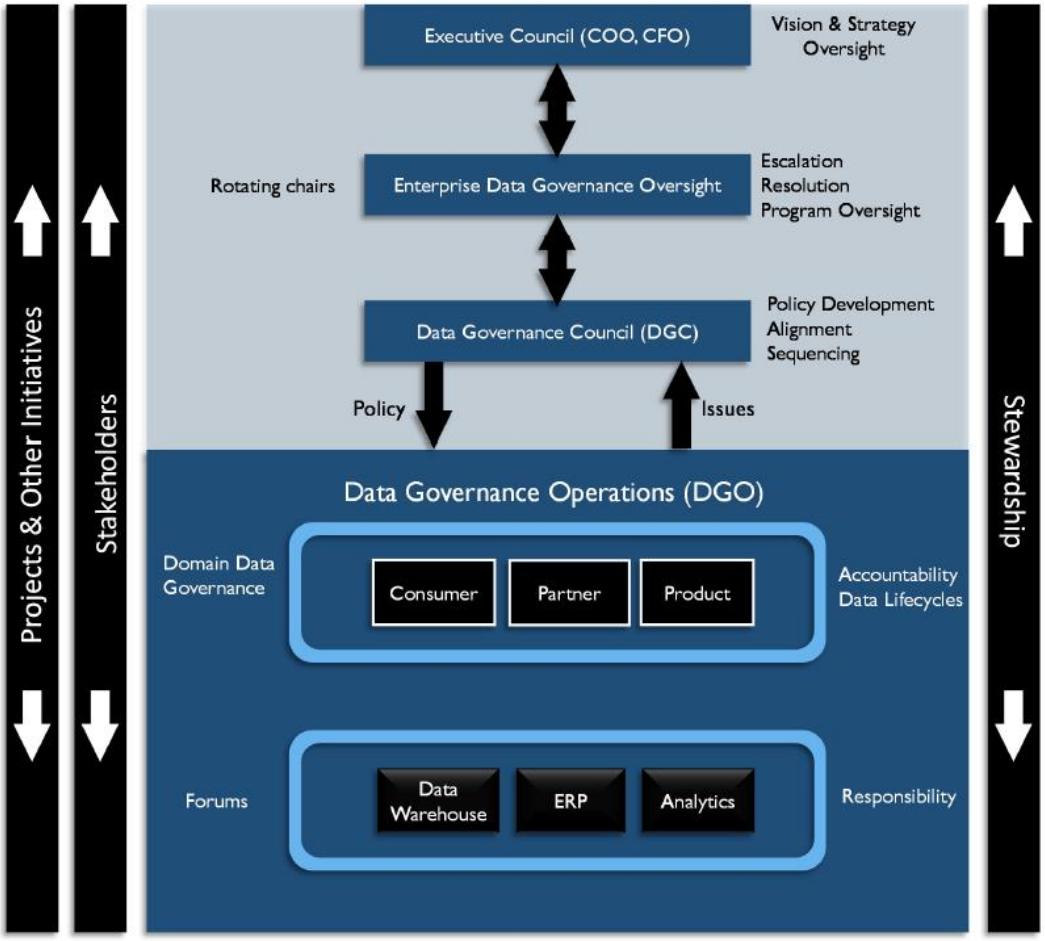


Figure 19 An Example of an Operating Framework

2.7 Develop Goals, Principles and Policies

Derived from the DG Strategy to the desired future state. They are drafted by data management professionals and/or business policy staff. Refined by data stewards and management. Data Governance Council conducts final review, revision and adoption.

Data policies must be effectively communicated, monitored, enforced and periodically re-evaluated. Data Governance Council may delegate this authority to the Data Stewardship Steering Committee.

2.8 Underwrite Data Management Projects

Promote enterprise wide data management improvements by articulating the ways they improve efficiency and reduce risk. Should be priority for organisations wanting more value from their data.

The DGC helps define the business case and oversees data management improvement project status and progress in coordination with the Project Management Office (PMO).

Data Management projects are part of the IT Project portfolio.

The DGC coordinates projects with enterprise wide scope such as Master Data Management, Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM).

Data management requirements must be captured in the planning and design phases of the SDLC

2.9 Engage Change Management

Organisational Change Management (OCM) - to bring about change in the organisation's systems and processes. A mature organisation in change management builds clear vision, leads and monitors from the top, and designs and manages small changes with feedback. Involves collaboration in whole organisation.

- **Planning:**
  - stakeholder analysis
  - gain sponsorship
  - Communications approach to resistance to change
- **Training:** Influencing systems development:
- **Policy implementation:** Communicate policies and the organisation's commitment to DM
- **Communications**:
  - Promoting the value of data Assets
  - Monitoring and acting on feedback about DG activities
  - Implementing data management training
  - Measuring the effects of change management in **5 key areas**:
    - Awareness of the need to change
    - Desire to participate and support the change
    - Knowledge about how to change
    - ability to implement new skills and behaviours
    - Reinforcement to keep change in place
- **Implementing new metrics and KPIs:** Employee incentives

2.10 Engage in Issue Management

The process of identifying, quantifying, prioritising and resolving DG issues:

- Authority
- Change management escalations
- Compliance
- Conflicts
- Conformance
- Contracts
- Data security and identity
- Data quality

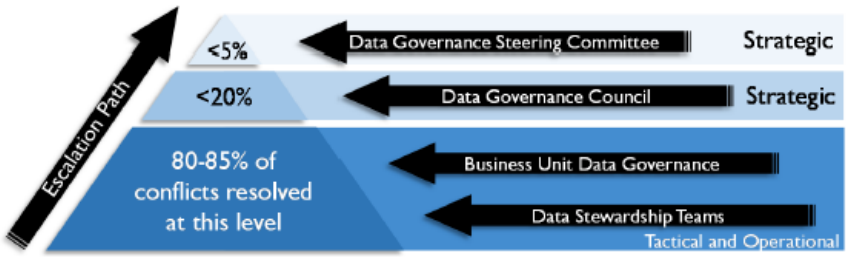


Figure 20 Data Issue Escalation Path

Develop control mechanisms and procedures for:

- Identifying, capturing, logging and updating issues

- Assignment and tracking of action items
- Documenting stakeholder viewpoints and resolution alternatives
- determining, documenting and communicating issue resolutions
- Facilitating objective, neutral discussions where all viewpoints are heard
- Escalating issues to higher levels of authority

2.11 Assess Regulatory Compliance Requirements

Part of the DG Function is to monitor and ensure regulatory compliance. DG guides the implementation of controls to monitor and document compliance with data-related regulations.

Examples of global regulations which impact data management practices.

- Accounting standards
- BCBS 239 (Basel Committee on Banking Supervision) and Basel II – for banks
- CPG 235 (Australian Prudential Regulation Authority APRA) – banks and insurance
- PCI-DSS – The Payment Card Industry Data Security Standards
- Solvency II - European Union Rules for insurance, similar to Basel II
- Privacy Laws

Evaluate the implications of the regulations

- Relevance of regulation to the organisation
- What constitutes compliance, and what policies and procedures are required?
- When is compliance achieved, and how and when is it monitored?
- Can industry standards achieve compliance?
- How is compliance demonstrated?
- Risks and penalties for non-compliance
- How is non-compliance identified and reported?
- How is non-compliance managed and rectified?

2.12 Implement DG

There are many complex activities which need to be coordinated using a roadmap with timeframes. Schedules may differ in a federated model with different business units based on differing levels of engagement, maturity and funding

Prioritised DG activities in the early stage:

- Define activities for high-priority goals
- Business glossary, document terminology and standards
- coordinate with Enterprise and data architecture to understand the data and systems
- Assign financial value to data assets to enable better decision making and understandingty654

2.13 Sponsor Data Standards and Procedures

Standard: Something set up and established by authority as a rule for measure of quantity, weight, extent, value or quality.

Standards help define DQ by providing a means of comparison.

Standards have the potential to simplify processes, as a decision is made once, and codified in a set of assertions (the standard).

Enforce standards to promote consistent results from the processes using them. Data can be measured against standards. The DGC or Data standards steering Committee should audit DM activities as part of the SDLC approval process or by schedule.

Standards difficulties in organisations: DG Standards should be mandatory

- Politicised
- Organisation not practiced at developing or enforcing DG standards
- Value of implementing standards is not recognised
- No knowledge how to implement standards

Different forms of data standards:

- How a field should be populated
- Rules governing relationships between fields
- Acceptable and unacceptable values
- Format

Process:

- Standards are drafted by data management professionals
- Reviewed, approved and adopted by the DGC or a Data Standards Steering Committee (A delegated workgroup)
- Document with capturing organisational knowledge in mind.

Data standards must be communicated, monitored, reviewed and updated. There must be a means to enforce them.

The DGC or DSSC should audit DM activities for standards compliance on a defined schedule or as part of the SDLC approval processes.

Data Management Procedures: The documented methods, techniques and steps followed to accomplish specific activities that produce certain outcomes and supporting artifacts.

Concepts can be standardised within all the Data Management Knowledge Areas.

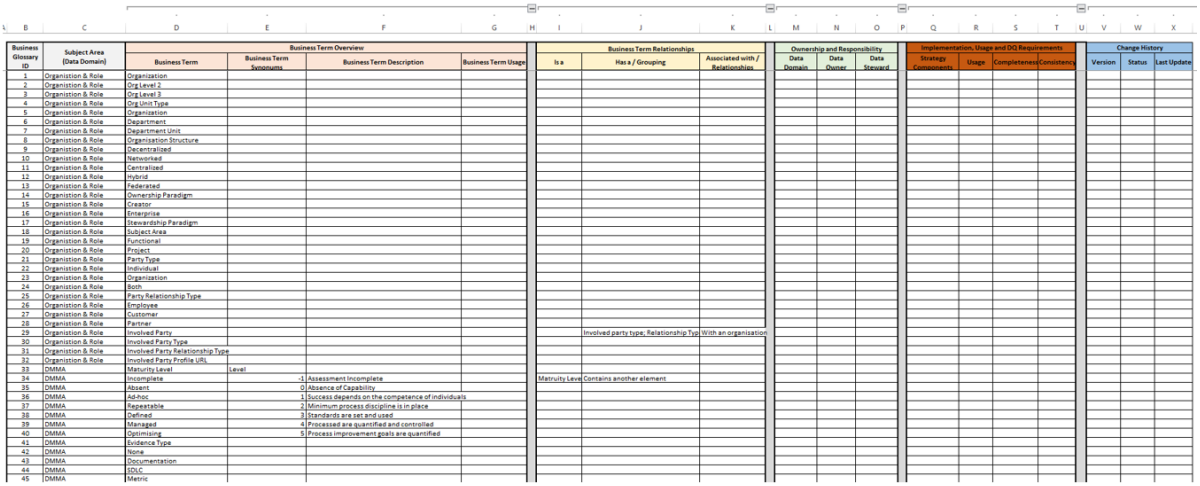
2.14 Develop a Business Glossary

A **Business Glossary** is a list of terms, definitions and other Metadata such as synonyms, metrics, lineage, business rules, the responsible steward for that term etc.

Objectives of business glossary:

- Enable common understanding of core business concepts and terminology
- Reduce the risk that data will be misused due to inconsistent understanding of the business concepts
- Improve alignment between technology assets and the business organisation
- Maximise search capability and enable access to documented institutional knowledge

Example of a Business Glossary



Business Glossary ID	Subject Area (Data Domain)	Business Term Overview				Business Term Relationships			Ownership and Responsibility			Implementation, Usage and DQ Requirements				Change History		
		Business Term	Business Term Synonyms	Business Term Description	Business Term Usage	Is a	Has a / Grouping	Associated with / Relationships	Data Domain	Data Owner	Data Steward	Strategy Component	Usage	Completeness	Consistency	Version	Status	Last Update
1	Organization & Role	Organization																
2	Organization & Role	Org Level 2																
3	Organization & Role	Org Level 3																
4	Organization & Role	Org Unit Type																
5	Organization & Role	Organization																
6	Organization & Role	Department																
7	Organization & Role	Department Unit																
8	Organization & Role	Organization Structure																
9	Organization & Role	Decentralized																
10	Organization & Role	Networked																
11	Organization & Role	Centralized																
12	Organization & Role	Hybrid																
13	Organization & Role	Federated																
14	Organization & Role	Ownership Paradigm																
15	Organization & Role	Creator																
16	Organization & Role	Enterprise																
17	Organization & Role	Stewardship Paradigm																
18	Organization & Role	Subject Area																
19	Organization & Role	Functional																
20	Organization & Role	Project																
21	Organization & Role	Party Type																
22	Organization & Role	Individual																
23	Organization & Role	Organization																
24	Organization & Role	Both																
25	Organization & Role	Party Relationship Type																
26	Organization & Role	Employee																
27	Organization & Role	Customer																
28	Organization & Role	Partner																
29	Organization & Role	Involved Party					Involved party type; Relationship Type	With an organization										
30	Organization & Role	Involved Party Type																
31	Organization & Role	Involved Party Relationship Type																
32	Organization & Role	Involved Party Profile URL																
33	DMMA	Maturity Level	Level															
34	DMMA	Incomplete		-1 Assessment Incomplete		Matruity Level	Contains another element											
35	DMMA	Absent		0 Absence of Capability														
36	DMMA	Ad-hoc		1 Success depends on the competence of individuals														
37	DMMA	Repeatable		2 Minimum process discipline is in place														
38	DMMA	Defined		3 Standards are set and used														
39	DMMA	Managed		4 Processed are quantified and controlled														
40	DMMA	Optimising		5 Process improvement goals are quantified														
41	DMMA	Evidence Type																
42	DMMA	None																
43	DMMA	Documentation																
44	DMMA	SDLC																
45	DMMA	Metric																

2.15 Coordinate with Architecture Groups

The DGC sponsors and approves data architecture artefacts.

Enterprise Data model is sponsored, reviewed and approved by DGC

Enterprise Data Architecture Steering Committee or Architecture Review Board appointed by DGC to oversee DA Program

- Enterprise Data model developed by Data Architects and Data Stewards in subject area teams
- Changes or extensions to EDM proposed and developed by Data Steward Teams.
- EDM must align with business strategy and data strategy

The Enterprise data model must be reviewed, approved and formally adopted by the DGC. It must align to key business strategies, processes and systems. Doing things right.

2.16 Sponsor Data Asset Valuation

DGC organises and standardises the effort to put monetary value to data

- Information gaps represent business liabilities
- Business value of the missing data can be the cost of closing the gaps
- Develop models to estimate value of information that does not exist
- Build value estimates into the data strategy road map
    - Justifies business cases for root cause DQ solutions
    - Business cases for other DG Initiatives

2.17 Embed DG

The organisation accepts the governance of data, and the processes and funding are in place to enable the continued performance of the DG framework.

Goal of the DGO:

- Embed behaviours related to managing data as an asset in processes
- Operations plan to implement and operate DG activities
- Activities, timing and techniques to ensure success

Sustainability of the DG organisational framework

- Processes and funding in place
- Organisation accepts the governance of data
- DG is measured, monitored and obstacles overcome

Create a Data Governance Community of Interest

- Deepens the understanding of DG
- Helpful in early years of governance

3 Tools and Techniques

DG is fundamentally about organisational behaviour, but tools can help with communication, metrics and business glossary development. Requirements must be clearly defined before purchasing a tool.

3.1 Online Presence / Websites

For collaboration, communication and sharing documents

3.2 Business Glossary

Core DG tool housing agreed-upon definitions and business terms and relates them to data.

3.3 Workflow tools

Connects processes to documents and is useful in policy admin and issue resolution.

3.4 Document management tools

3.5 Data Governance Scorecards

Collection of metrics to track DG activities. Can be automated.

4 Implementation Guidelines

4.1 Organisation and culture

The target of organisational change is **sustainability** (a quality of a process that measures how easy it is for the process to continue to add value).

4.2 Adjustment and Communication

- Business / DG strategy map
- DG Roadmap
- Ongoing business case for DG
- DG Metrics

5 Metrics

DG program must be able to measure progress and success and how DG participants have added value to business objectives.

- **Value**: Contributions to business objectives, reduction of risk, improved efficiency
- **Effectiveness**: Achievement of goals and objectives, Stewards using relevant tools, Effectiveness of communication, education and training
- **Sustainability**: Policies and processes working appropriately, Staff conforming to standards and procedures

Data Governance Checklist

Deliverables:

- Data Governance Strategy
- Data Strategy
- Business / Data Governance Strategy Roadmap
- Data Principles, Data Governance Policies, Processes
- Operating Framework
- Roadmap and Implementation Strategy
- Operations Plan
- Business Glossary
- Data Governance Scorecard
- Data Governance Website
- Communications Plan
- Recognized Data Value
- Maturing Data Management Practices

- Strategy
  - Data Governance Strategy
  - Data Strategy
  - Business / Data Governance Strategy Roadmap
  - Roadmap & Implementation Strategy
- Organization
  - Operating Framework
- Culture & Change
  - Communications Plan
- Working Methods
  - Data Principles, Policies & Processes
  - Maturing Data Management Practices
  - Operations
- Results
  - Business Glossary
  - Recognized Data Value
- Measurement
  - Data Governance Scorecard

Knowledge Area Data Governance

Data Governance	Measurement	Policy Framework	Change Management

• Data Sources to be integrated
• Data Quality rules to be enforced
• Conditions of use rules
• Activities to be monitored and the frequency of monitoring
• Priority and response levels of stewardship efforts
• How information is represented to meet stakeholder needs
• Standard approval gates, expectations in RDM and MDM deployment
	• Leading Indicators
	  • Knowledge Area Metric
	• Lagging Indicators
	  • Business SWOTS
	  • Insights
	  • Decision Making
	  • Revenue
	• Data Management SWOTS (Friction)
		• Principles
		• Policies
		• Procedures
		• Standards
		• Guidelines
		• Roles & Responsibilities
		• Policy Relationships
		  • Industry
		  • Organization
		  • Internal
			• Vision
			• Culture
			  • Decision Making
			  • Attitudes
			  • Behaviour
			• Communication
			• Collaboration Structures
			  • Agile

Data Architecture

1 Introduction

The essential components of Data Architecture:

- Data Architecture Outcomes: The Data Architecture Artefacts
- Data Architecture Activities: To fulfil Data Architecture intentions
- Data Architecture Behaviour: Collaboration among roles with an Enterprise view

Data Architecture Is fundamental to data Management. The vast data of an organisation must be represented at various levels of abstraction so that it can be understood for management to make decisions.

An organisation's Data Architecture consists of:

- Master design documents at different abstraction
- formal enterprise data model – containing data names, metadata definitions conceptual and logical entities and relationships and business rules.

Data Architecture enables consistent data standardisation and integration across the enterprise.

Data Architecture artefacts constitute metadata and should be stored and managed in and enterprise architecture artefact repository.

ISO/IEV 42010:2007 Definition:

The fundamental organization of a system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution.

1.1 Business Drivers

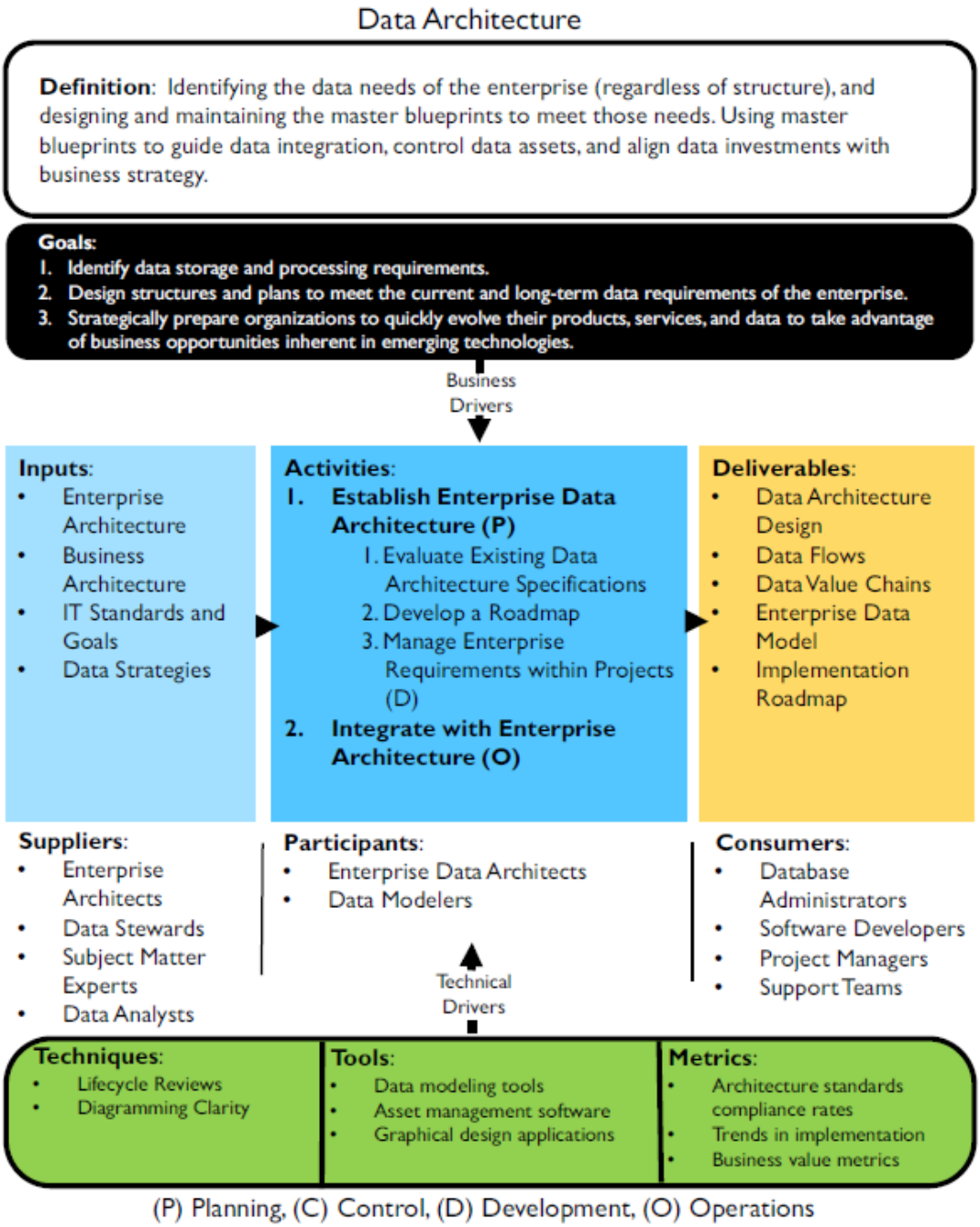
- Strategic preparation of evolution of products, services and data to take advantage of business opportunities in emerging technologies
- Translate business needs into data and system requirements
- Manage complex data delivery throughout the enterprise
- Facilitate alignment between business and IT
- Act as agents for transformation
- Influence measures of the value of data

Data Architecture is a bridge between business strategy and technology execution. Data architects create and maintain organisational knowledge about data and the systems through which it moves, which enables it to maintain data as an asset, and to increase the value it gets from data.

1.2 Data Architecture Outcomes and Practices.

Primary Data Architecture outcomes:

- Data storage and processing requirements
- Designs of structure and plans that meet current and long term data requirements of the enterprise



To reach goals Data Architects define and maintain specifications that:

- Define the current state of the organisation
- Provide standard business vocabulary for data and components
- Align Data Architecture with enterprise strategy and business architecture
- Express strategic data requirements
- Outline high level designs to meet these requirements
- Integrate with overall enterprise architecture roadmap

1.3 Essential Concepts

1.3.1 Enterprise Architecture Domains

Domain	Enterprise Business Architecture	Enterprise Data Architecture	Enterprise Applications Architecture	Enterprise Technology Architecture
Purpose	To identify how an enterprise creates value for customers and other stakeholders	To describe how data should be organized and managed	To describe the structure and functionality of applications in an enterprise	To describe the physical technology needed to enable systems to function and deliver value
Elements	Business models, processes, capabilities, services, events, strategies, vocabulary	Data models, data definitions, data mapping specifications, data flows, structured data APIs	Business systems, software packages, databases	Technical platforms, networks, security, integration tools
Dependencies	Establishes requirements for the other domains	Manages data created and required by business architecture	Acts on specified data according to business requirements	Hosts and executes the application architecture
Roles	Business architects and analysts, business data stewards	Data architects and modelers, data stewards	Applications architects	Infrastructure architects

1.3.2 Enterprise Architecture Frameworks

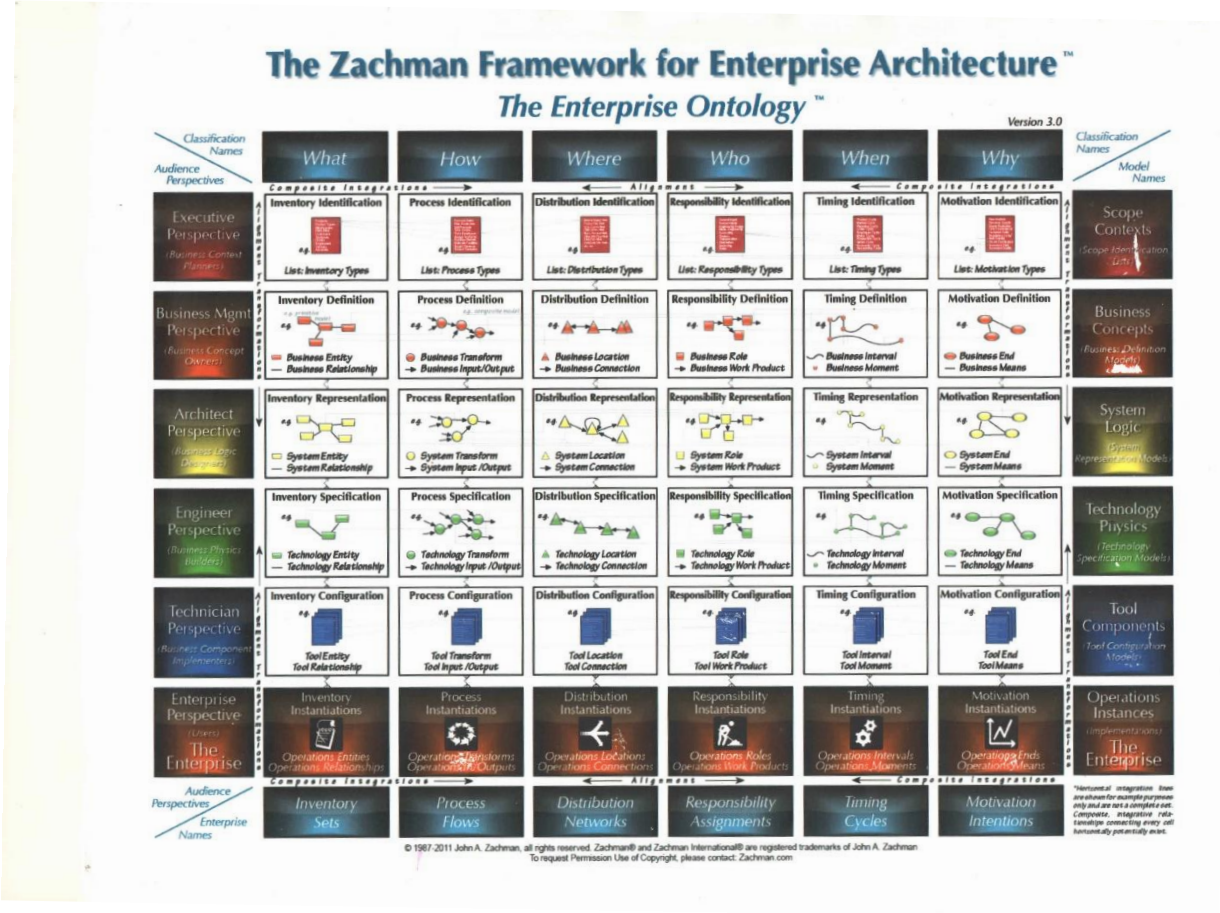
Provide a framework for thinking about and understanding architecture.

1.3.2.1 Zachman Framework for Enterprise Architecture

Developed by John A. Zachman in the 1980s. An ontology. The 6 x 6 matrix shows the models required to describe an enterprise, and the relationships between them. Each cell represents a unique type of design artefact defined by the intersection of row and column.

	What	How	Where	Who	When	Why	
Executive	Inventory Identification	Process Identification	Distribution Identification	Responsibility Identification	Timing Identification	Motivation Identification	Scope Context
Business Management	Inventory definition	Process Definition	Distribution Definition	Responsibility Definition	Timing Definition	Motivation Definition	Business Concepts
Architect	Inventory Representation	Process Representation	Distribution Representation	Responsibility Representation	Timing Representation	Motivation Representation	System Logic
Engineer	Inventory Specification	Process Specification	Distribution Specification	Responsibility Specification	Timing Specification	Motivation Specification	Technology Physics
Technician	Inventory Configuration	Process Configuration	Distribution Configuration	Responsibility Configuration	Timing Configuration	Motivation Configuration	Tool Components
Enterprise	Inventory Instantiations	Process Instantiations	Distribution Instantiations	Responsibility Instantiations	Timing Instantiations	Motivation Instantiations	Operational Instances
	Inventory Sets	Process Flows	Distribution Networks	Responsibility Assignments	Timing Cycles	Motivation Intentions	

Figure 22 Simplified Zachman Framework



Columns: Communication interrogatives

- What (the inventory column): Entities used to build the architecture
- How (the process column: Activities performed
- Where (the distribution column): Business location and technology location
- Who (the responsibility column): Roles and organisations

- When (the timing column): Intervals, events, cycles and schedules
- Why (the motivation column): Goals, strategies and means

Rows: Reification transformations – the steps necessary to translate an abstract idea into a concrete instance from different perspectives (planner, owner, designer, builder, implementer and user). Each perspective has a different relation to the What column:

- The executive perspective – business context
- The business management perspective – business concepts
- The architect perspective – business logic
- The engineer perspective – business physics
- The technician perspective – component assemblies
- The user perspective – operations classes

1.3.3 Enterprise Data Architecture

Enterprise Data Architecture descriptions include:

- **Enterprise Data Model (EDM**): A holistic, enterprise-level, implementation-independent conceptual or logical data model providing a common consistent view of data across the enterprise.
- **Data flow design:** Defines requirements and master blueprint for storage and processing across databases, applications, platforms and networks (the components).

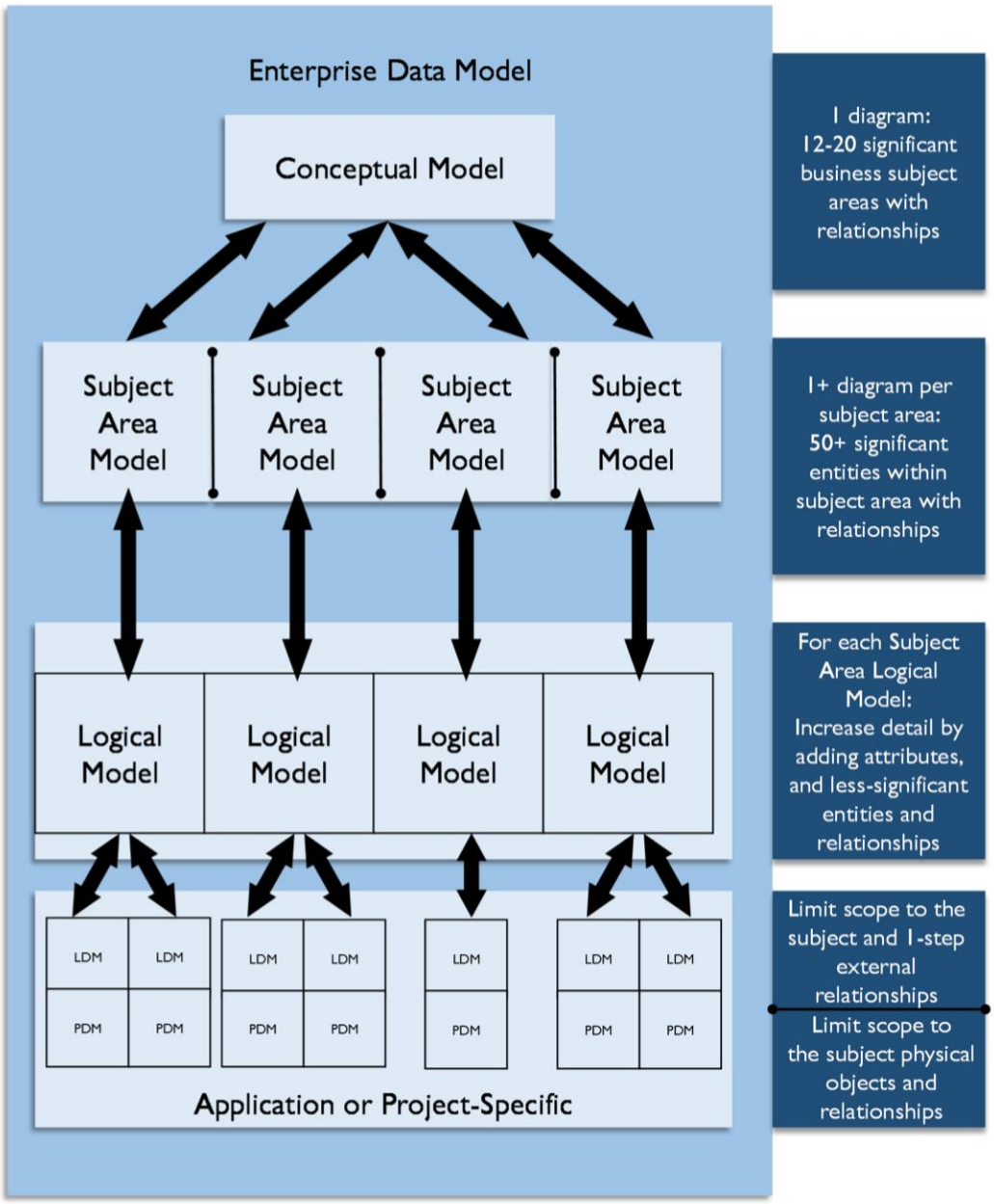
Both need to be reflected in current state (architecture perspective) and transition state (project perspective).

1.3.3.1 Enterprise Data Model

Can be a stand-alone artefact or may be composed of data models from different perspectives or levels of detail. An EDM includes both universal (Enterprise-wide Conceptual and Logical Models) and application or project specific data models, along with definitions, specifications, mappings and business rules.

Industry standard is a start but time and effort must also be invested to build and maintain the EDM.

- Conceptual overview over the enterprise's subject areas
- Views of entities and relationships for each subject area
- Detailed, partially attributed logical views for the same subject area
- Logical and physical models specific to an application or project



Different types of models are related vertically and horizontally.

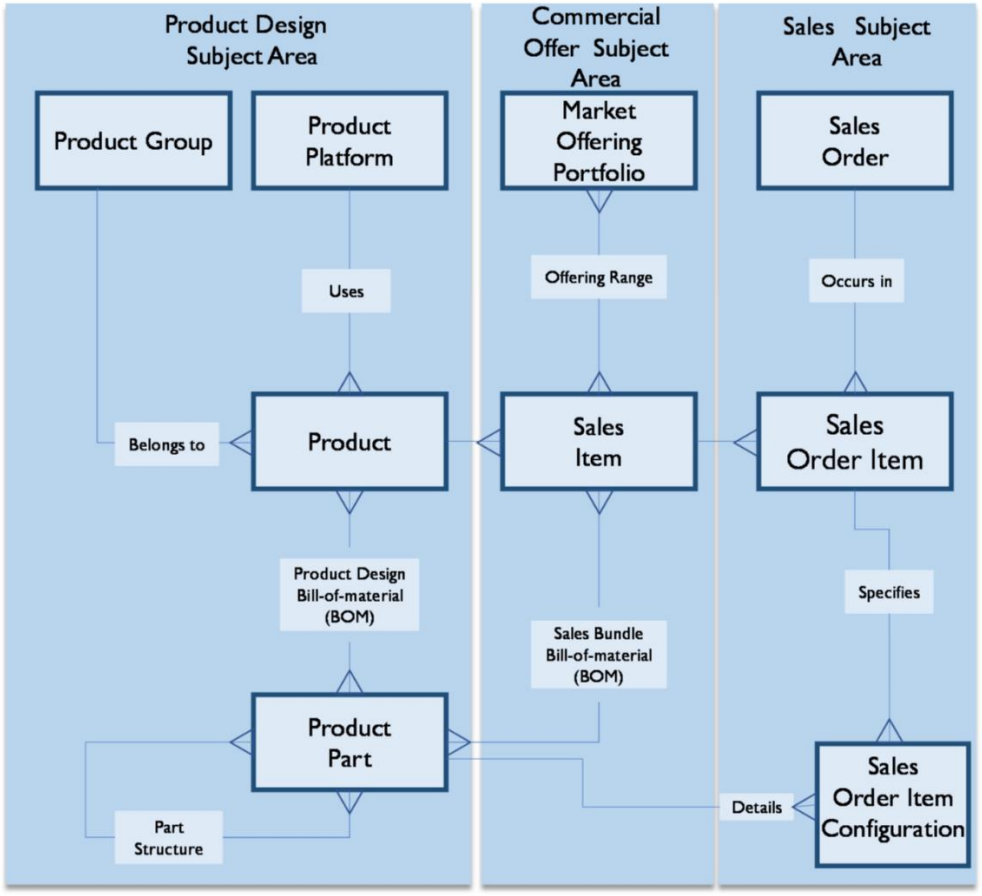
The Enterprise Data Model is built up by the combination of Subject Area Models:

- **Top Down approach**: Form Subject Areas then populate them with models
- **Bottom up Approach:** Subject Area structure is based on existing data models

The Subject Area Discriminator (the way subject areas are formed) must be consistent throughout the enterprise data model:

- **Funding:** Systems portfolios
- **Organisational**: Data governance structure and data ownership
- **Business value chains**: Top-level processes
- **Using business capabilities**

Subject Area Model Example

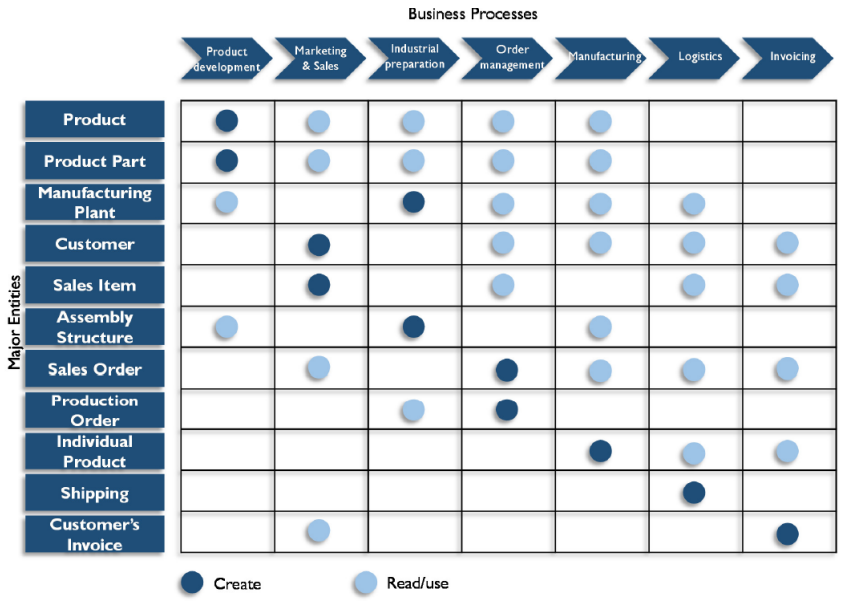


1.3.3.2 Data Flow Design

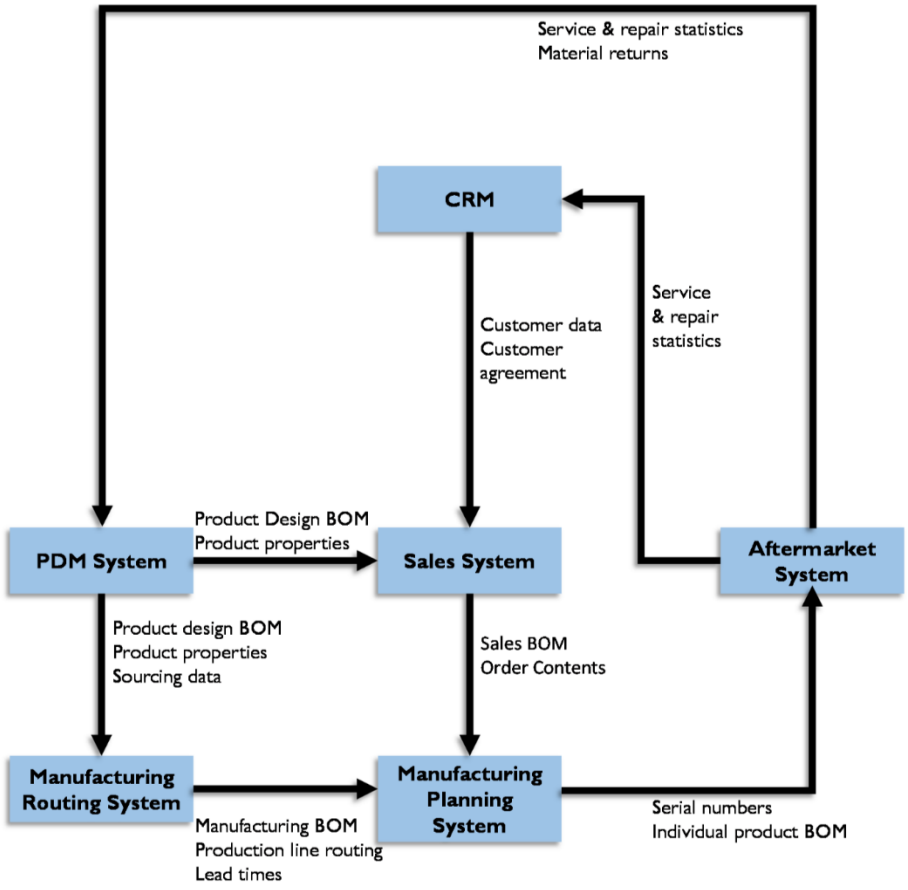
Data flows are data lineage documentation that depicts how data moves end-to-end through business processes and systems, where it originated, is stored and used and how it transforms.

Data flows can be documented at different levels of detail: Subject Area, Business entity or attribute level.

Data Flow Matrix example:



High level data flow diagram:



2 Activities

Two approaches:

- **Quality-oriented**: Focus on execution within business and IT structures. Unless architecture is managed it will deteriorate. Architectural improvements are incremental.

- **Innovation-oriented:** Focus on transforming business and IT to address new expectations and opportunities. Requires interaction with business development representatives and designers

2.1 Establish Data Architecture Practice

A Data Architecture practice includes the following work streams:

- **Strategy:** Select frameworks, state approaches, develop roadmap
- **Acceptance and culture:** Inform and motivate changes in behaviour
- **Organisation:** Assign Data Architecture accountabilities and responsibilities
- **Working methods:** Define best practices and perform Data Architecture work within development projects, in coordination with Enterprise Architecture
- **Results:** Produce Data Architecture artefacts within an overall roadmap

Enterprise Architecture influences scope boundaries of projects/system releases:

- Defining project data requirements
- Reviewing project data designs
- Determining data lineage impact
- Data replication control
- Enforcing Data Architecture standards
- Guide data technology and renewal decisions

2.1.1 Evaluate Existing Data Architecture Specifications

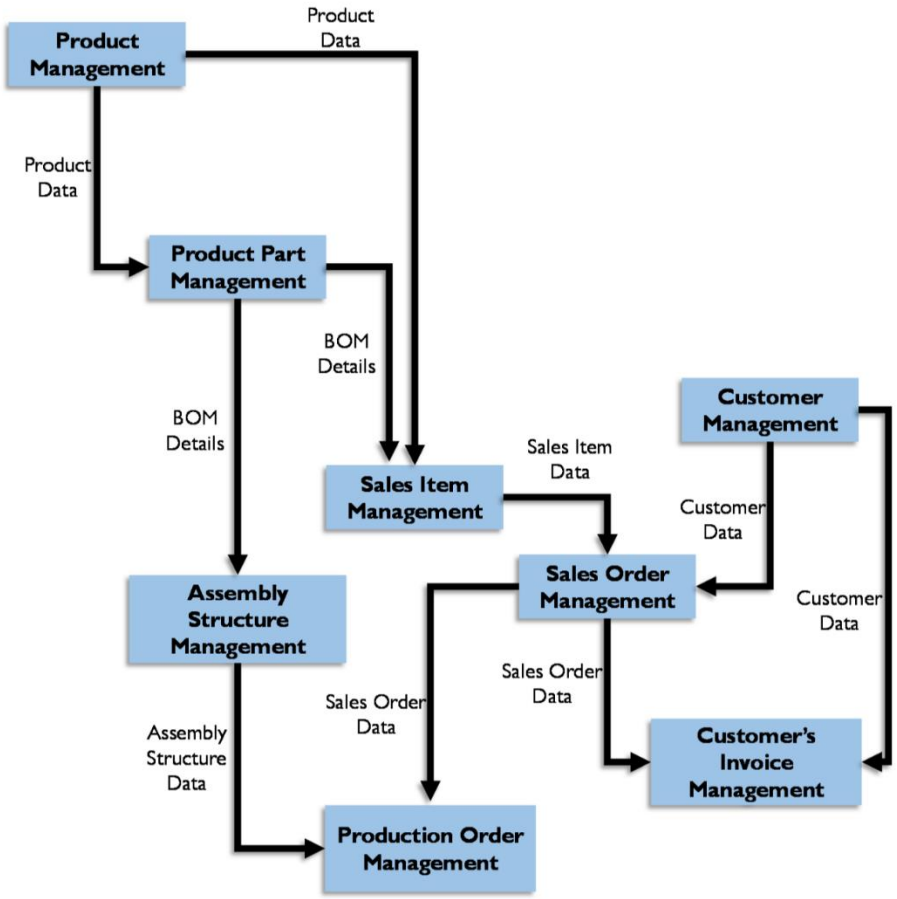
Evaluate existing documentation for accuracy, completeness and level of detail. Update if necessary.

2.1.2 Develop a Roadmap

Describes the architecture's 3 to 5-year development path, with business requirements, consideration of actual conditions and technical assessments. Must be integrated into the enterprise architecture roadmap. Include milestones, resources needed, cost estimations, and divided into work streams.

The Enterprise Data Architecture can be formed by resolving input and output data flows in the chain of dependencies between business capabilities.

Start with the most independent business capabilities and end with those most dependent on other activities. The following diagram has the lowest dependency on the top.



2.1.3 Manage Enterprise Requirements within Projects

Enterprise Data Architecture project-related activities include:

- **Define scope:** Ensure the scope and interface are aligned with Enterprise Data Model. Identify components that can be reused, and down-stream dependencies.
- **Understand business requirements**
- **Design:** Form detailed target specifications. Look for shareable constructs in the enterprise logical data model. Review and use technology standards.
- **Implement:**
  - **When buying:** Reverse engineer purchased applications and map against data structure
  - **When reusing data:** Map application data models against common data structures and new and existing processes to understand CRUD operations. Enforce use of authoritative data.
  - **When building:** Data storage according to data structure.

The role of Enterprise Data Architects and the process of building architectural activities into projects depends on the development methodologies:

- **Waterfall methods:** Construct systems in sequential phases as part of an overall design.
- **Incremental methods:** Learn and construct in gradual steps. creates prototypes based on vague overall requirements.
- **Agile, iterative methods:** Learn, construct and test in discrete delivery packages (Sprints).

2.2 Integrate with Enterprise Architecture

Integrate Enterprise Data Architecture matters with project portfolio management as funded projects drive architecture priorities and Data Architecture can influence the scope of projects.

3 Tools

- **Data Modelling Tools:** Include lineage and relation tracking to manage linkages between models
- **Asset Management Software:** Used to inventory systems, describe their content and track the relationships between them
- **Graphical Design Applications:** To create architectural design diagrams and other architectural artefacts

4 Techniques

4.1 Lifecycle Projections

Architecture designs can be:

- Aspirational and future-looking
- Implemented and active
- Plans for retirement

What architectural plans represent should be clearly documented:

- **Current:** Products supported and used
- **Deployment period:** Products deployed for use in 1-2 years
- **Strategic period:** Products available in the next 2+ years
- **Retirement:** Retired products, or retirement within 1 year
- **Preferred:** Products preferred for use by most applications
- **Containment:** Limited for use by certain applications
- **Emerging:** Researched and piloted for possible future deployment
- **Reviewed:** Evaluated products and their evaluation results

4.2 Diagramming clarity

Models and diagrams must conform to an established set of visual conventions:

- **Clear and consistent legend:** Identify all objects and lines and placed in the same spot in all diagrams.
- **Match between all diagram objects and the legend:** Not all legend objects need appear on diagram
- **Clear and consistent line direction:** Usually left to right. Backward lines must be clear
- **Consistent object attributes:** Differences in size, line thickness and colour should signify something
- **Linear symmetry:** Line up at least half of the objects to improve readability

5 Implementation Guidelines

As Data Architecture is about artefacts, activities and behaviour, Enterprise Data Architecture is about:

- Organising Enterprise Data Architecture teams and forums

- Producing initial versions of Data Architecture artefacts such as enterprise data model, enterprise wide data flow and road maps.
- Forming and establishing a data architectural way of working in development projects.
- Creating organisation wide awareness of the value of Data Architecture efforts.

A Data Architecture implementation should include at least 2 of the above.

Data models and other Data Architecture artefacts are captured within development projects and are then standardised and maintained by architects. There will be more architectural work in early projects which may need special architectural funding.

Enterprise Data Architecture evolves incrementally in a solution-oriented culture using agile development.

Enterprise Data Architecture starts with Master Data areas in need of improvement in planned development projects, and expands to include business and other data.

5.1 Readiness / Risk Assessment

More risks than other projects, especially during an organisation's first attempt:

- **Lack of management support**
- **No proven record of accomplishment**
- **Apprehensive sponsor**
- **Counter-productive executive decisions**
- **Culture shock**
- **Inexperienced project leader**
- **Dominance of a one-dimensional view**

5.2 Organisation and Cultural change

The ability of an organisation to adopt Data Architecture practices depends on several factors:

- Cultural receptivity to architectural approach
- Organisation recognises data as a business asset, not just an IT concern
- Ability to let go of a local perspective and adopt an enterprise perspective on data
- Ability to integrate architectural deliverables into project methodology
- Level of acceptance of formal data governance
- Ability to look holistically at the enterprise

6 Data Architecture Governance

Enterprise Data Architecture and the Data Governance organisation must be well aligned. A data steward and a data architect should be assigned to each subject area, even to each entity within, as Data Architecture activities support the alignment and control of data. Business event subject areas should be aligned with business processes governance as each event entity usually corresponds to a business process.

Data Architecture governance activities include:

- **Overseeing projects:** Projects comply with required Data Architecture activities, use architectural assets and are implemented according to data architectural standards.
- **Managing architectural designs, lifecycle and tools:** Designs must be defined, evaluated and maintained

- **Defining standards**
- **Creating data-related artefacts**

6.1 Metrics

Data Architecture metrics may be monitored annually for business customer satisfaction:

- **Architecture standard compliance rate:** How far projects comply with established data architectures.
- **Implementation trends:** The degree to which enterprise architecture has improved the organisation's ability to implement projects along at least two lines:
    - **Use/reuse/replace/retire measurements:** Proportion of new architectural artefacts to reused, replaced or retired artefacts.
    - **Project execution efficiency measurements:** Measure lead times for projects and their resource costs for delivery improvements with reusable artefacts and guiding artefacts.
- **Business value measurements:** Track progress towards expected business benefits
    - **Business agility improvements:** Account for the benefits of lifecycle improvements or the cost of delay
    - **Business quality:** Measure whether business cases are fulfilled and projects deliver changes leading to business improvements
    - **Business operation quality:** Measure of improved efficiency and accuracy
    - **Business environment improvements**

Data Modelling and Design

1 Introduction

Data Modelling is the process of discovering, analysing and scoping data requirements, and then representing and communicating these data requirements in a precise form called the data model.

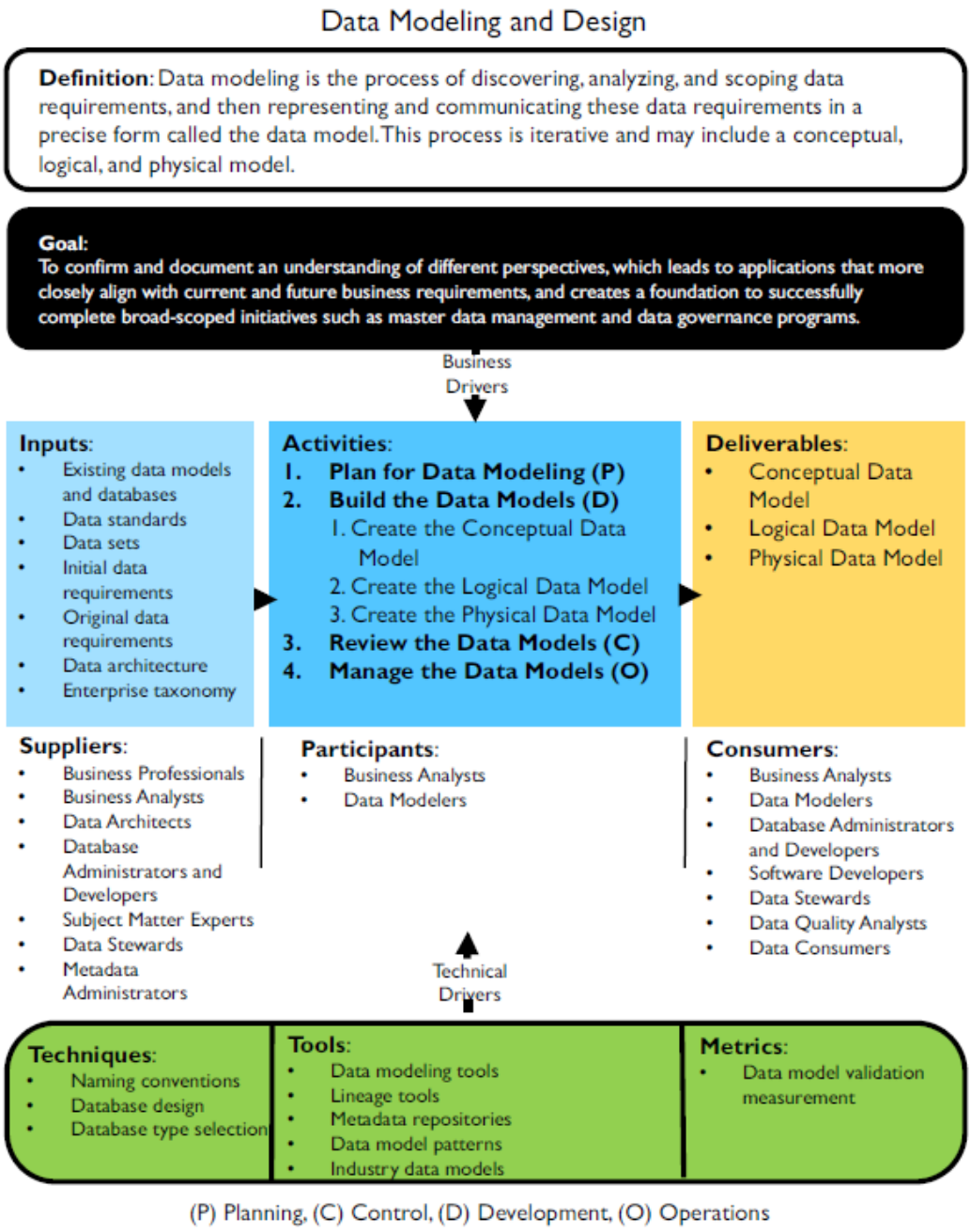
Six most commonly used schemes to represent data are:

- Relational
- Dimensional
- Object-Oriented
- Fact-Based
- Time-Based
- NoSQL

Models exist at three levels of detail:

- Conceptual
- Logical
- Physical

Each model contains a set of components such as entities, relationships attributes, facts and keys. Data Models contain Metadata uncovered during the modelling process which is essential to other data management functions.



1.1 Business Drivers

Data models:

- Provide common vocabulary around data
- Capture and document explicit knowledge about the organisation's data and systems
- Primary communications tool during projects
- Provide the starting point for customisation, integration or replacement of an application

1.2 Goals and Principles

Goal:
To confirm and document an understanding of different perspectives, which leads to applications that more closely align with current and future business requirements, and creates a foundation to successfully complete broad-scoped initiatives such as master data management and data governance programs.

Confirming and documenting understanding from different perspectives:

- **Formalisation**: A concise definition of data structures and relationships, how data is affected by implemented business rules. A disciplined structure reduces data anomalies occurring.
- **Scope definition**: Explain boundaries for data context in packages, projects or existing systems
- **Knowledge retention/documentation**: Preserves corporate memory regarding a system by capturing knowledge in an explicit form. Documentation for future projects. Understand the implications of modifications. Data models are reusable maps.

1.3 Essential Concepts

1.3.1 Data Modelling and Data Models

A model consists of diagrams of standard symbols which represent something that exists or something to be made. A data model describes the organisation's data as it is, or as how it wants it to be. Data models are the main medium to communicate data requirements from business to IT and within IT.

1.3.2 Types of data that are modelled

- **Category Information**: Data used to classify and define things (customers classified by market segment or products classified by colour)
- **Resource Information**: Master and Reference Data: objects needed to conduct operational processes such as Product, Customer, Supplier, Facility, Organization, and Account, Countries, Currencies
- **Business Event Information**: Transaction Data created while operational processes are in progress. Examples include Customer Orders, Supplier Invoices, Cash Withdrawal, and Business Meetings
- **Detail Transaction Information**: POS, Social Media, Clickstream, IoT event. Large volume and rapidly changing. Usually referred to as Big Data. Internet of Things – sensors etc.

1.3.3 Data Model Components

Basic building blocks of Data models:

- Entities
- Relationships
- Attributes
- domains

1.3.3.1 Entity

An entity is a thing about which and organisation collects information. A noun of the organisation. Questions to ask to identify entities:

Category	Definition	Examples
Who	Person or organization of interest. That is, *Who* is important to the business? Often a 'who' is associated with a party generalization, or role such as Customer or Vendor. Persons or organizations can have multiple roles or be included in multiple parties.	Employee, Patient, Player, Suspect, Customer, Vendor, Student, Passenger, Competitor, Author
What	Product or service of interest to the enterprise. It often refers to what the organization makes or what service it provides. That is, *What* is important to the business? Attributes for categories, types, etc. are very important here.	Product, Service, Raw Material, Finished Good, Course, Song, Photograph, Book
When	Calendar or time interval of interest to the enterprise. That is, *When* is the business in operation?	Time, Date, Month, Quarter, Year, Calendar, Semester, Fiscal Period, Minute, Departure Time
Where	Location of interest to the enterprise. Location can refer to actual places as well as electronic places. That is, *Where* is business conducted?	Mailing Address, Distribution Point, Website URL, IP Address
Why	Event or transaction of interest to the enterprise. These events keep the business afloat. That is, *Why* is the business in business?	Order, Return, Complaint, Withdrawal, Deposit, Compliment, Inquiry, Trade, Claim
How	Documentation of the event of interest to the enterprise. Documents provide the evidence that the events occurred, such as a Purchase Order recording an Order event. That is, *How* do we know that an event occurred?	Invoice, Contract, Agreement, Account, Purchase Order, Speeding Ticket, Packing Slip, Trade Confirmation
Measurement	Counts, sums, etc. of the other categories (what, where) at or over points in time (when).	Sales, Item Count, Payments, Balance

An **entity instance** is a particular occurrence of an **entity**.

Usage	Entity	Entity Type	Entity Instance
Common Use	Jane	Employee	
Recommended Use	Employee		Jane

Entites are represented graphically by **rectangles**

Definition of entites: (important as they are core Metadata). Definitions clarify the meaning of business vocabulary, and provide rigor to the business rules governing entity relationships.

- **Clarity:** Easy to read and grasp
- **Accuracy:** Precise and correct description of the entity
- **Completeness**: All parts of the definition are present. The scope of uniqueness is in the definition. e.g. examples of code values for a code entity.

1.3.3.2 Relationship

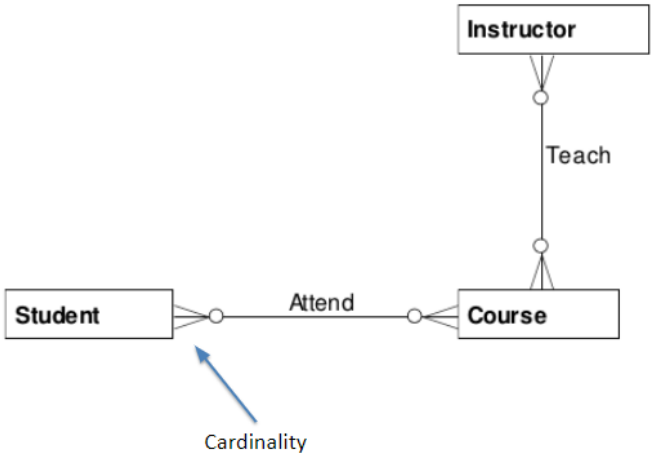
An association between entities.

Other aliases based on scheme:

- **Dimensional**: Navigation path

- **NoSQL**: Edge or Link
- **Relational on the physical level**: Constraint or Reference

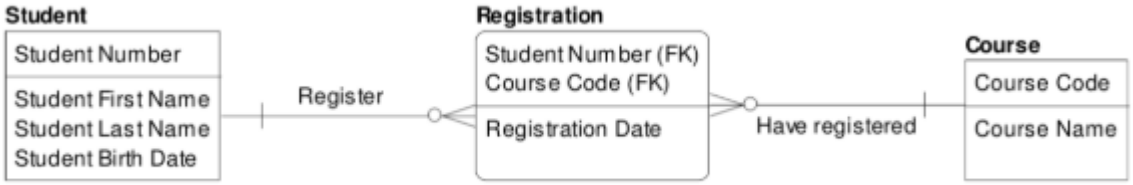
Graphically represented as lines on the diagram



Reading a Data Model: a 90 second video by Steve Hoberman
https://www.youtube.com/watch?v=adYohKb47f8

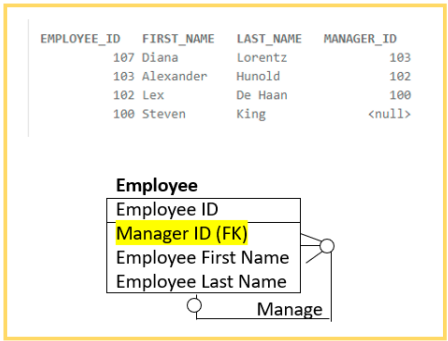
1.3.3.2.1 *Relationship cardinality*:

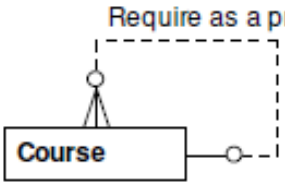
Data rules on how entities are connected are enforced. An instance of an entity may have a relationship with zero, one or many instances of another entity.



Unary (Recursive or self-referencing) relationship:

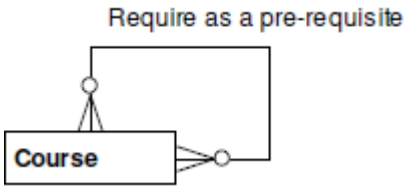
This is a nonidentifying, nonmandatory relationship in which the same entity is both the parent and the child. There is only one entity. A Foreign Key is used to distinguish between the roles or instances.





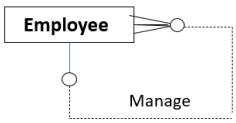
Hierarchy: One to many recursive relationship Child entity has at most 1 parent

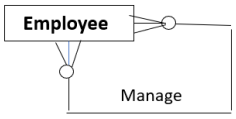
A course can be a prerequisite once

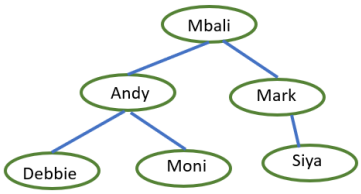


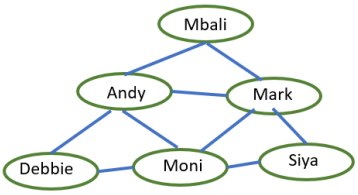
Network: Many to many recursive relationship. Entity instance can have more than one parent.

A course can be prerequisite for many courses

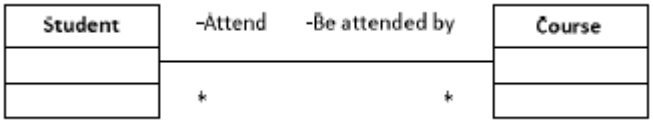




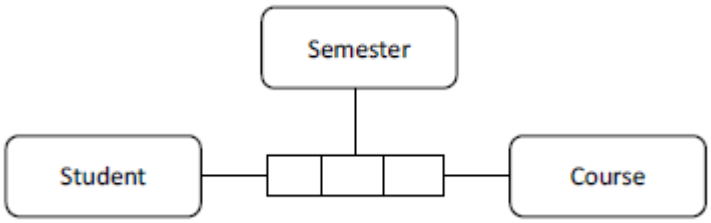




Binary Relationship: 2 Entities

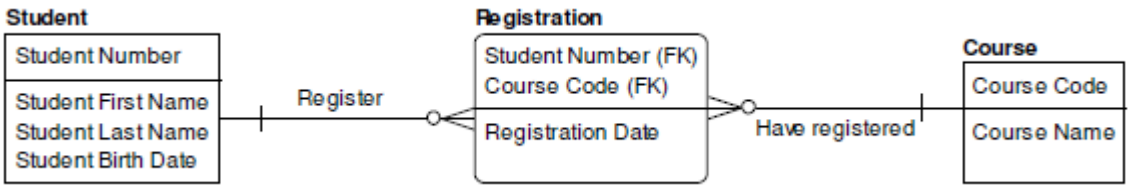


Ternary Relationship: 3 Entities.



1.3.3.2.2 Foreign Keys

Used in logical and physical data models to represent a relationship.



1.3.3.3 Attribute

A property that identifies, describes or measures an entity. A column in a table. May have domains. In a list as in Student above.

1.3.3.3.1 Identifiers (Key)

Set of one or more attributes that uniquely identifies and instance of an entity.

Types of keys:

- Construction type keys:
  - **Simple**: One attribute e.g. VIN number
  - **Surrogate:** Unique identifier, often system generated, without intelligence and not visible to end users
  - **Compound:** 2 or more attributes together e.g. *areacode+exchange+localnumber=phone number*
    IssuerID+AccountID+Check digit = Credit card number

    Each attribute is FK as for an associative entity or the PK on a Fact Table.

  - **Composite:** One compound key and at least one other simple or compound key or non-key attribute. e.g. a key on a multidimensional fact table. 2 or more attributes that identify an entity occurrence.
- Function type keys:
  - **Super Key**: any set of attributes that uniquely identify an instance
  - **Candidate key**: Minimal set of attributes that uniquely identifies an entity instance. Also called **Business** or **Natural** keys.
  - **Primary key**: The candidate key chosen to be the unique identifier. Often a surrogate key
  - **Alternate key**: Unique candidate key, but not chosen for primary. Usually a business key.

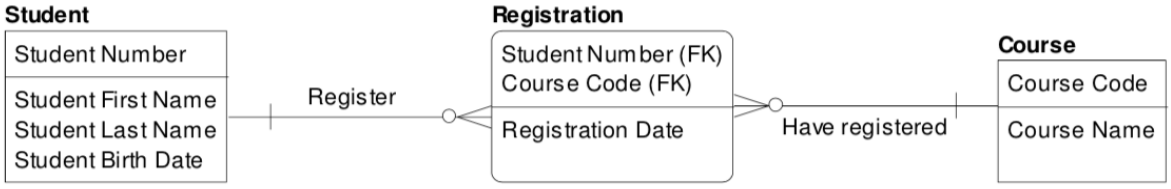
1.3.3.3.2 Identifying vs Non-Identifying Relationships

Independent Entity:

- Primary key contains only attributes belonging to that entity
- Non-identifying (weak) relationship ---------
- FK non-primary key attribute

Dependent Entity:

- The primary key contains at least one attribute from another entity
- Rectangles with rounded corners
- Identifying (strong) relationship _____



1.3.3.4 Domain

A domain is a complete set of possible values that an attribute can be assigned. It is used to standardise the characteristics of the attributes.

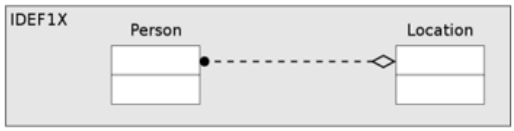
Domains can be restricted by adding **constraints** (additional rules). These can be format and/or logical.

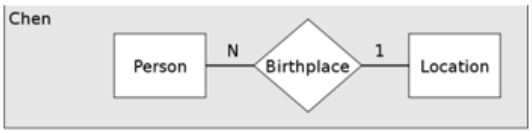
Defining domains:

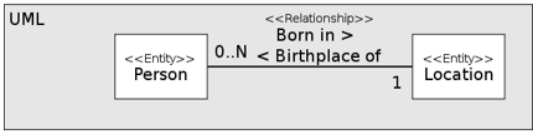
- **Data type:** standard type of data in an attribute assigned to that domain
- **Data Format:** domains using templates and character limitations
- **List:** Finite set of values
- **Range:** values of same data type between max and min values or can be open ended
- **Rule-based**: e.g. compare values to a calculated value

1.3.4 Data Modelling Schemes and Notations

Scheme	Sample Notations
Relational	Information Engineering (IE) Integration Definition for Information Modeling (IDEF1X) Barker Notation Chen
Dimensional	Dimensional
Object-Oriented	Unified Modeling Language (UML)
Fact-Based	Object Role Modeling (ORM or ORM2) Fully Communication Oriented Modeling (FCO-IM)
Time-Based	Data Vault Anchor Modeling
NoSQL	Document Column Graph Key-Value







Scheme	Relational Database Management System (RDBMS)	Multidimensional Database Management System (MDBMS)	Object Databases	Document	Column	Graph	Key-Value
Relational	CDM LDM PDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM
Dimensional	CDM LDM PDM	CDM LDM PDM					
Object-Oriented	CDM LDM PDM		CDM LDM PDM				
Fact-Based	CDM LDM PDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM
Time-Based	PDM						
NoSQL			PDM	PDM	PDM	PDM	PDM

The models which can be built for each scheme, based on the technology.

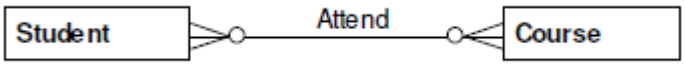
1.3.4.1 Relational

Dr Edward Codd (1970) proposed that data could be manages most effectively in terms of two dimensional relations. reducing redundancy and data storage. This approach was based on the mathematics of set theory.

Relationships capture business rules.

Exact expression of business data, and have one fact in one place (removal of redundancy). Ideal for operational systems requiring quick data entry, individual transaction processing and accurate storage.

Information Engineering (IE) syntax:

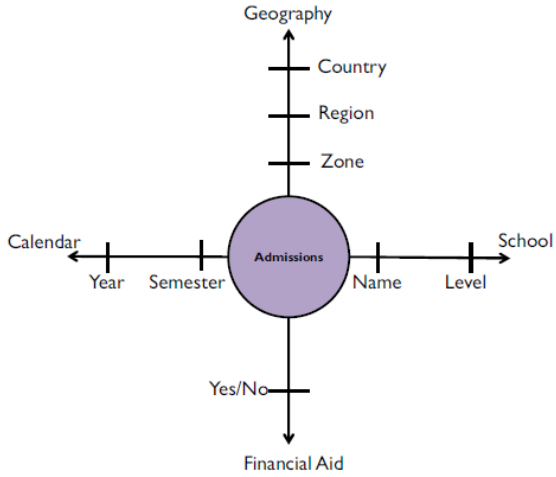


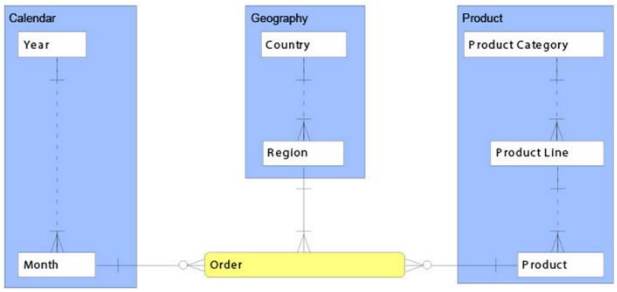
1.3.4.2 Dimensional

General Mills and Dartmouth College in the 1960s. Data is structured to optimise query and analysis of large amounts of data. Batch processing.

Dimensional models capture business questions focussed on a particular business process. Relationships capture navigation paths to answer the business question.

Axis (Peacock) notation diagram: Admissions

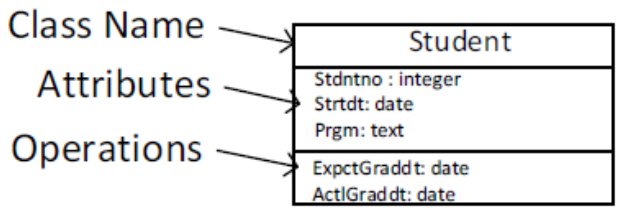


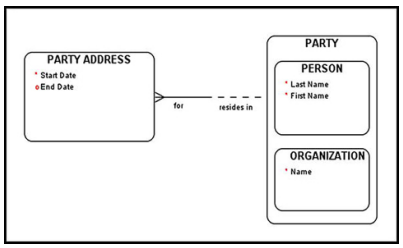


- **Fact Tables:** The measurement, result of calculations or algorithms. Is numeric. Metadata is important for understanding. Fact tables consist of a large number of rows. 90% of the data.
- **Dimension Tables:** Mostly textual. Used for querying. Highly denormalised. Must have a unique identifier for each row. Dimensions have attributes that change. Slowly Changing Dimensions (SCDs) manage changes based on the rate and type of change.
- **Snowflaking:** Normalising the star schema
- **Grain:** The most detail a row in the fact table can have
- **Conformed Dimensions:** Built with the entire organisation in mind and shared across dimensional models
- **Conformed Facts:** Use standardised definitions of terms across marts.

1.3.4.3 Object Oriented (UML)

Unified Modelling Language (UML) class model is used for databases. Specifies classes (entity types) and their relationship types.

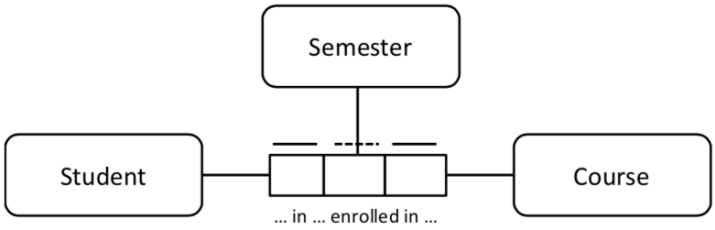




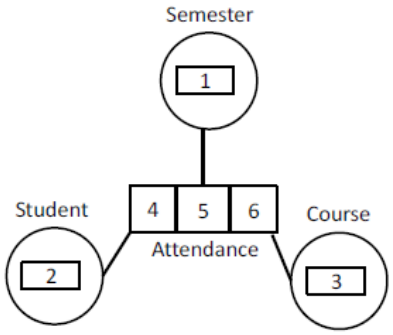
1.3.4.4 Fact based modelling

Based on natural verbalisation in the business domain:

- **Object role modelling (ORM):** Model driven engineering approach which verbalises required information at a conceptual level in a controlled natural language.



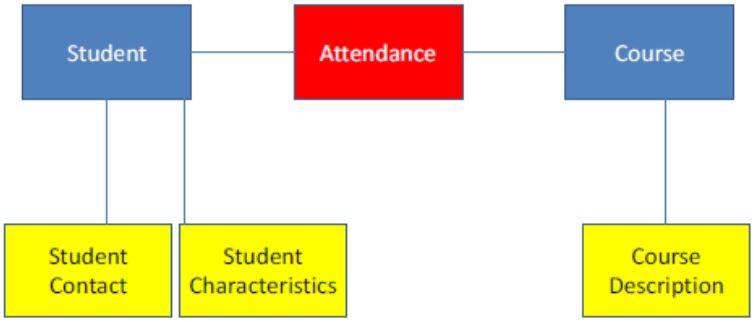
- **Fully communication oriented modelling (FCO-IM):**



1.3.4.5 Time-Based

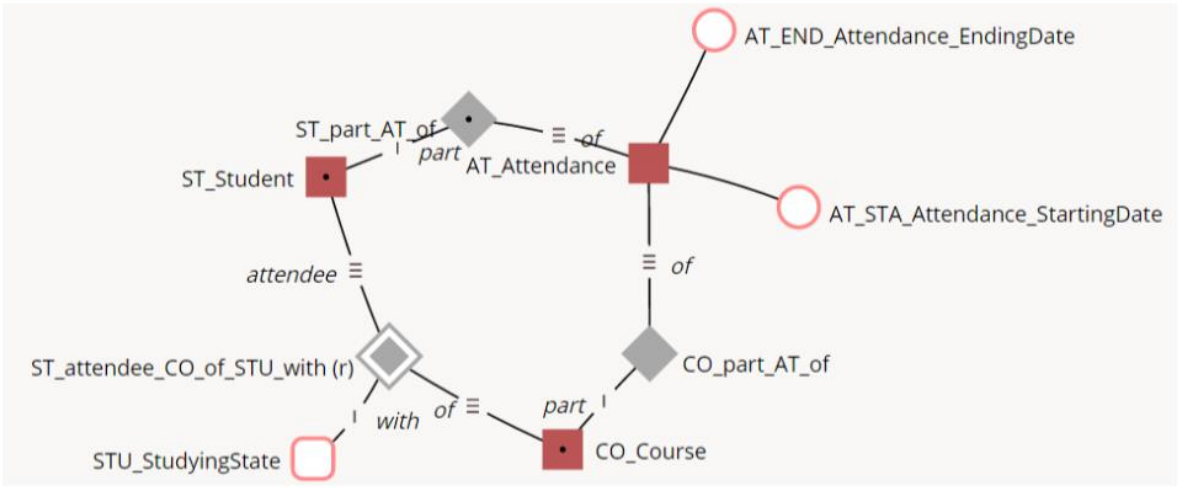
Time-based patterns are used when data values must be in chronological order and with time values. Used for data warehouses in a RDBMS environment.

- **Data Vault**: Detail oriented, time-based, normalised tables that support one or more functional areas of the business. Between 3NF and Star schema to meet needs of enterprise data warehouses.
  - **3 types of entities**:
    - **Hubs**: the primary key
    - **Links**: provide transaction integration between hubs
    - **Satellites**: provide the context of the hub primary key



- **Anchor Modelling**: Graphical notation for information that changes over time in both structure and content. Similar to traditional data modelling but with extensions for temporal data.
  - **4 basic concepts**:
    - **Anchors**: model entities and events
    - **Attributes**: model properties of anchors
    - **Ties**: model the relationship between anchors
    - **Knots**: model shared properties such as states

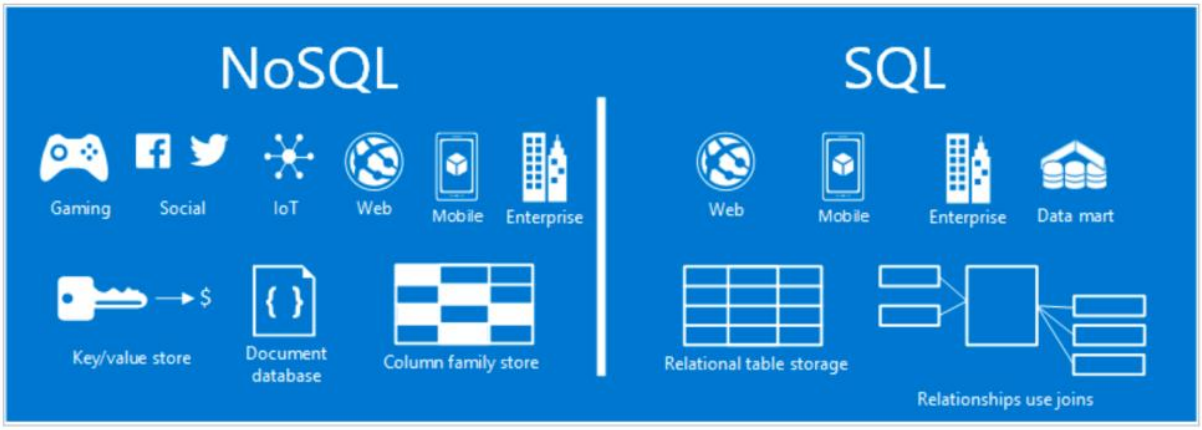
On the anchor model in Figure 45, **Student**, **Course**, and **Attendance** are anchors, the gray diamonds represent ties, and the circles represent attributes.



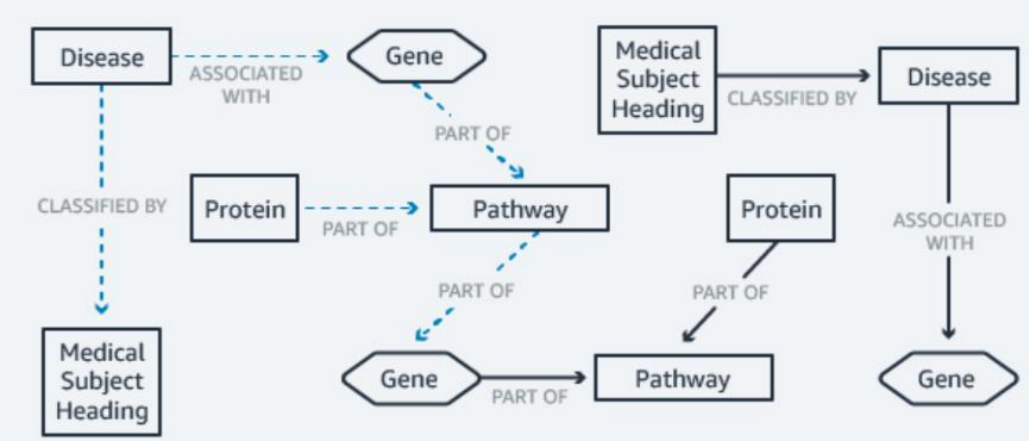
1.3.4.6 NoSQL

NoSQL is the name for databases built on non-relational technology. There are four types of NoSQL databases:

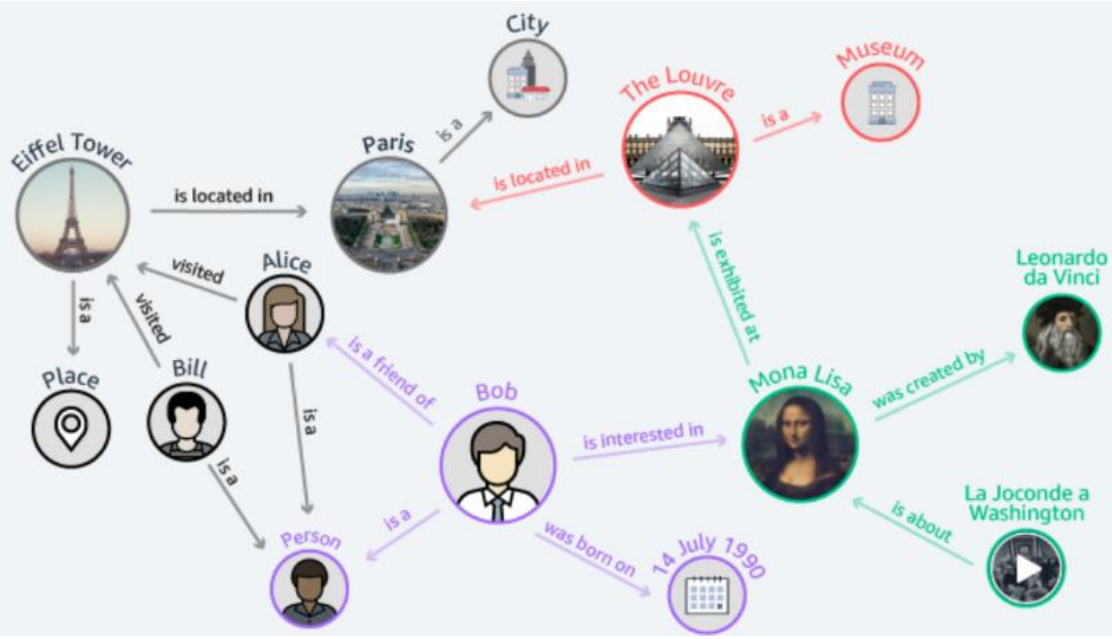
- **Document:**
  - Store the business subject in one structure called a document. For example, instead of storing Student, Course, and Registration information in three distinct relational structures, properties from all three will exist in a single document called Registration.
- **Key-value:**
  - Key-value databases allow an application to store its data in only two columns ('key' and 'value').
  - Value column can store anything (i.e. text OR video**)**
- **Column-oriented:**
  - RDBMSs work with a predefined structure and simple data types, such as amounts and dates, whereas column-oriented databases, such as Cassandra, can work with more complex data types including unformatted text and imagery
  - Store each column in its own structure
- **Graph**:
  - A graph database is designed for data whose relations are well represented as a set of nodes with an undetermined number of connections between these nodes



Knowledge Graph (Model)



Knowledge Graph (Data)



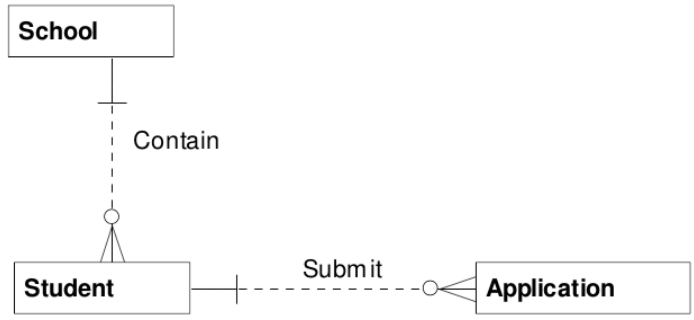
1.3.5 Data Model Levels of Detail

- **Conceptual** – Real world view of the enterprise modelled in the database

- **Logical** – Subsets of the enterprise model which represent the data requirements in a specific usage context, independent of technology.
- **Physical** – Internal or machine view. Describes the stored representation of the enterprise's data

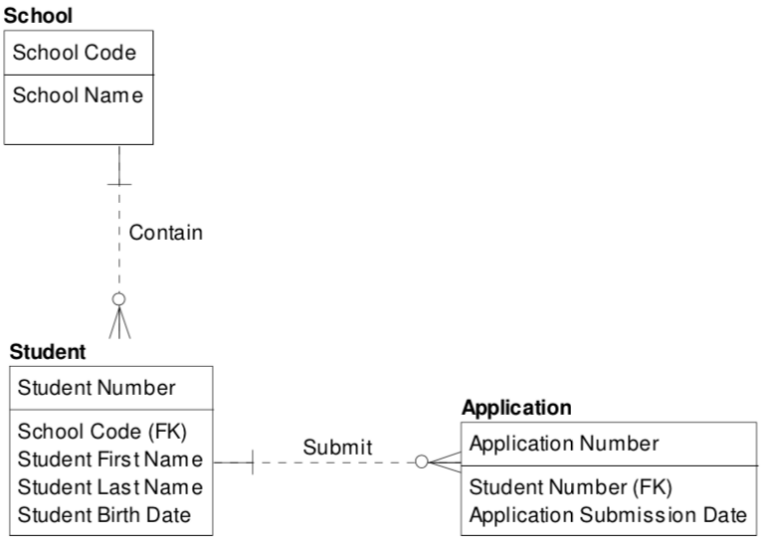
1.3.5.1 Conceptual (CDM)

The conceptual data model captures the high-level data requirements as a collection of related concepts. Basic business entities and the relationships between them (business rules) are described.



1.3.5.2 Logical (LDM)

A detailed, technology independent representation in a specific usage context. An extension of conceptual model. Attributes are assigned by applying normalisation.



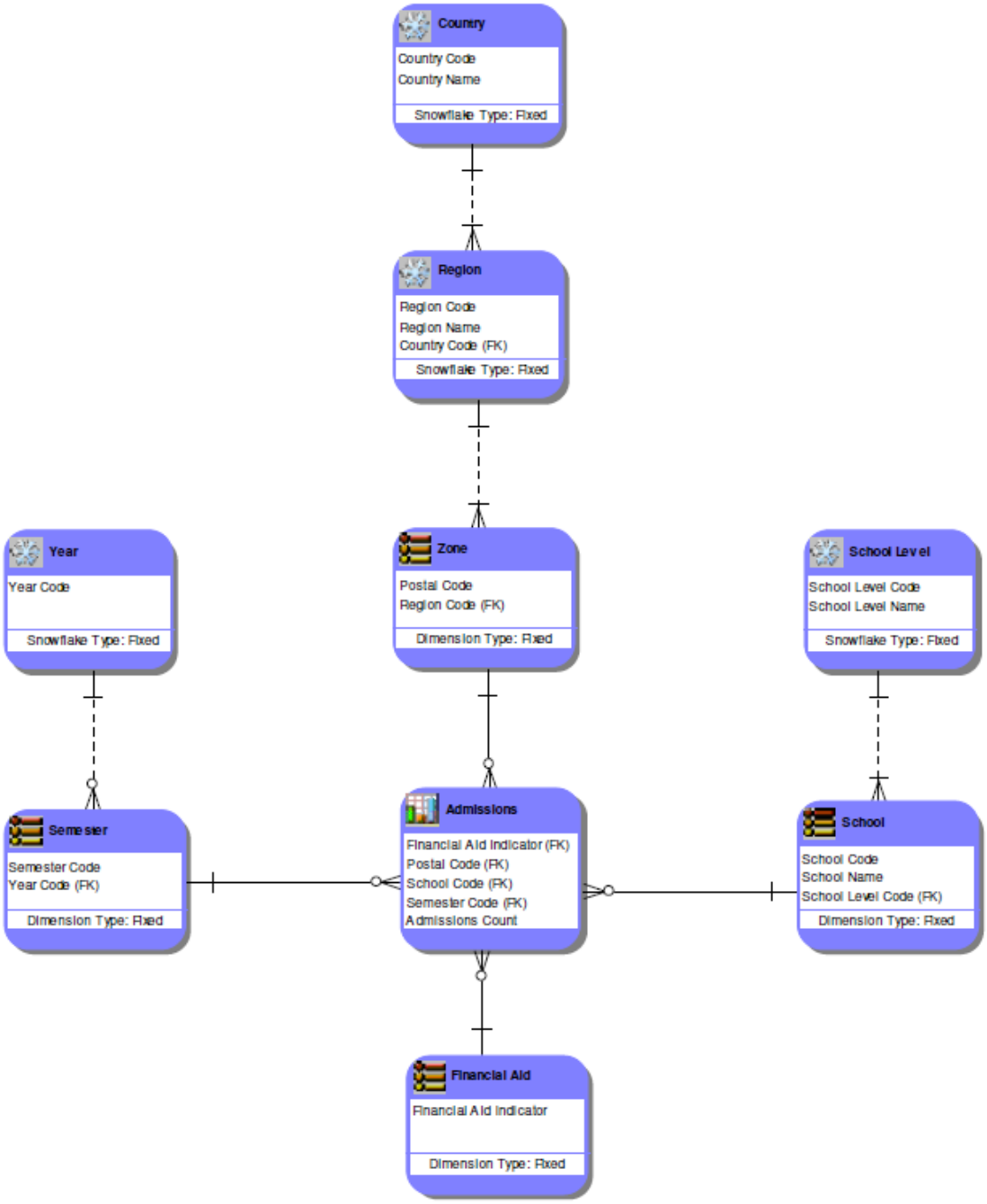
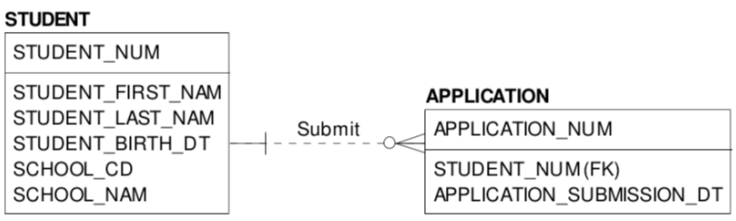


Figure 49 Dimensional Logical Data Model

1.3.5.3 Physical (PDM)

Detailed technical solution using the logical data model. Built for a particular technology of DBMS.



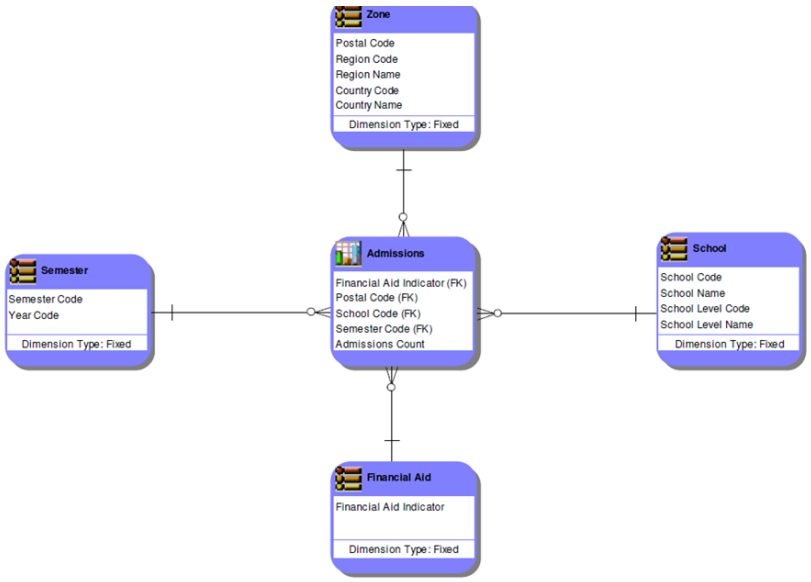


Figure 51 Dimensional Physical Data Model

A **Canonical Model** is a type of physical data model which describes the structure of data moving between systems. Structures should be generic and reusable. Used in the Enterprise Service Bus or Enterprise Application Integration (EAI).

Views are virtual tables used to simplify data access, control data access and rename columns without redundancy and loss of referential integrity due to denormalisation.

Partitioning: Process of splitting a table to improve performance

- **Vertically split**: subset tables contain subsets of columns
- **Horizontally split**: a value in a column is a delimiter to create subset tables

Denormalisation: Deliberate transformation of normalised logical model entities into physical tables with redundant data structures. Done to improve performance

1.3.6 Normalisation

Normalisation is the process of applying rules to organise business complexity into stable data structures. The basic goal is to eliminate redundancy by keeping each attribute in only one place. Requires understanding of attribute's relationship to its primary key.

Normalisation levels:

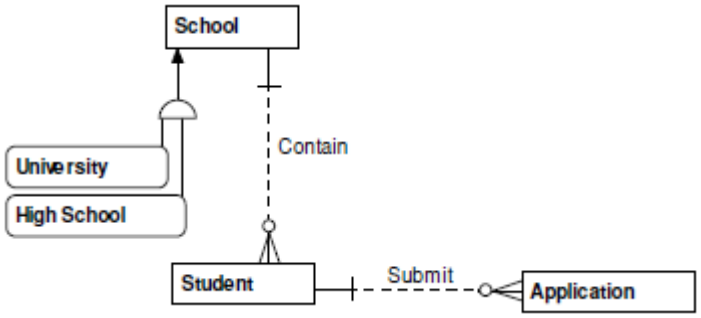
- **First normal form (1NF)**: Valid primary key. Each attribute depends on key. Eliminate repeating groups and ensure each attribute is atomic and not multi valued.
- **Second normal form (2NF)**: Each entity has the minimal primary key and each attribute depends on the complete primary key
- **Third normal form (3NF)**: no hidden primary keys. Each attribute depends on no attributes outside the key ("the key, the whole key and nothing but the key")
- **Boyce / Codd normal form (BCNF)**: Resolves overlapping composite candidate keys (hidden business rules)
- **Fourth normal form (4NF)**: Resolves all many to many relationships in pairs until they can't be broken down into smaller pieces.
- **Fifth normal form (5NF)**: Resolves inter-entity dependencies into basic pairs, and all join dependencies use parts of primary keys.

1.3.7 Abstraction

Abstraction is the removal of details in such a way as to broaden applicability to a wide class of situations while preserving the important properties, e.g. Party/Role structure

Includes:

- **Generalisation**: Groups common attributes and relationships into **supertype** entities
- **Specialisation**: Separates distinguishing attributes within an entity into **subtype** entities.
    - Subtypes can also be created using **roles** or **classification** e.g. Party with subtypes Individual and Organisation



2 Activities

2.1 Plan for Data Modelling

Tasks to plan for:

- Evaluating organisational requirement
- Creating standards
- determining data model storage

Deliverables of the modelling process:

- **Diagram:** The visual that captures the requirements in precise form. Depicts:
    - **Level of detail:** Conceptual, logical or physical
    - **Scheme:** Relational, Dimensional, Object-oriented, Fact-based, Time-based or NoSQL
    - **Notation:** information engineering, unified modelling language, object role modelling
- **Definitions**: for entities, attributes and relationships are essential for precision
- **Issues and outstanding questions**:
- **Lineage**: Where the data comes from.
    - A source/target mapping: Can see source system attributes and how they populate target system attributes
    - Trace components from conceptual to logical to physical
    - Data modeller obtains a good understanding of data requirements and can determine the source attributes
    - Source attributes can be used to check accuracy of the model and mapping

2.2 Build the data model

Modelling involves studying previous analysis, existing data models and databases etc.

Modelling is an iterative process: draft the model then go back to business analysts for clarification, update the model then ask more questions. This increases the precision of the model.

2.2.1 Forward Engineering

Build a new application from the requirements

- CDM: Understand the scope and key terminology
- LDM: Documents the business solution
- PDM: Documents the technical solution

2.2.1.1 Conceptual Data Modelling

Steps to create a CDM:

- **Select scheme:** Relational, dimensional, fact-based or NoSQL
- **Select notation:** Appropriate notation for selected scheme
- **Complete initial CDM:** Capture the viewpoint of the user group
    - Collect the highest-level concepts (nouns)
    - Collect the activities (verbs) that connect these concepts. Relationships can go both ways or involve more than two concepts.
- **Incorporate enterprise terminology:** ensure consistency with enterprise terminology and rules
- **Obtain signoff:** reviewed for data modelling best practices and that it meets requirements

2.2.1.2 Logical Data Modelling

LDM captures the data requirements within the scope of the CDM

- **Analyse information requirements:** elicitation, organisation, documentation, review, refinement approval and change control of business requirements.
- **Analyse existing documentation:** pre-existing artefacts provide a jump start for a new model
- **Add associative entities:** Used to describe the many-to-many relationships
- **Add attributes:** should be atomic
- **Assign domains:** allow for consistency of value sets and format
- **Assign keys:** Identify primary and alternate keys

2.2.1.3 Physical Data Modelling

LDM must be modified to perform well within storage applications

- **Resolve logical abstractions:** subtypes and supertypes become separate entities
- **Add attribute details:** Technical names, physical domain, physical data types and length of fields as well as constraints such as NOT NULL
- **Add reference data objects:** Small Reference Data value sets
- **Assign surrogate keys:** Unique key values not visible to business
- **Denormalise for performance:** dimensional structures
- **index for performance:** optimise query performance. Prevents every row being read (table scan)
- **Partition for performance:** Partition on a date key is usually recommended
- **Create views:** Requirements driven. Control access to certain data elements or embed joins or filters.

2.3 Review the Data Model

Apply a data quality verifier such as Steve Hoberman's Data Model Scorecard ® for quality control.

2.4 Maintain the Data Models

The data models need to be kept current. Update when business requirements or processes change. Compare physical with logical regularly.

3 Tools

- **Data Modelling Tools:** may support forward engineering from conceptual to logical to physical including DDL generation. Also reverse engineer. Some support naming standards and metadata storage.
- **Lineage Tools:** Capture and maintenance of source structures for each attribute in the data model. Excel is most commonly used.
- **Data Profiling Tools:** explores data content and validates it according to metadata and identifies data quality gaps/deficiencies
- **Metadata Repositories:** stores descriptive data about the model including diagram and definitions. Enables sharing.
- **Data Model Patterns:** Reusable modelling structures
- **Industry Data Models:** Pre-built for an entire industry. Needs to be customised.

4 Best Practices

4.1 Best Practices in Naming Conventions

ISO 11179 Metadata Registry is the international standard for representing metadata. Data architects, data analysts and database administrators jointly develop standards for an organisation, to complement related IT standards. Names should be unique and descriptive. Logical names must be meaningful to business users, whereas physical names must conform to DBMS restrictions.

4.2 Best Practices in Database Design

Design principles for the DBA (PRISM):

- **Performance and ease of use:** Quick and easy access
- **Reusability:** Multiple applications can use the data
- **Integrity:** Data should always have valid business meaning and value
- **Security:** Only available to authorised users
- **Maintainability:** ensure the cost of creating, storing, maintaining, using and disposing of data does not exceed its value to the organisation.

5 Data Model Governance

5.1 Data Model and Design Quality Management

Data models and database designs should be a reasonable balance between the short term needs and the long term needs of the enterprise.

5.1.1 Develop Data Modelling and Design Standards

Data modelling and database design standards help meet business data requirements, conform to Enterprise and Data Architecture and ensure data quality. Standards should include the following:

- List and description of standards data modelling and database design deliverables
- List of standard names, abbreviations and abbreviation rules
- List of standard naming formats for all data model objects
- List and description of standard methods of creating and maintaining these objects

- List and description of data modelling and database design roles and responsibilities.
- List and description of all Metadata properties captured in data modelling and database design
- Metadata quality expectations and requirements
- Too use guidelines
- guidelines for preparing and leading design reviews
- Guidelines for versioning models
- Practices that are discouraged

5.1.2 Review Data Model and Database Design Quality

- **Requirements Reviews:** Project team
  - Starting model
  - Changes made to the model
  - Rejected options
  - How well new model conforms to modelling and architectural standards
- **Design reviews:** Group of diverse subject matter experts
  - Chair meeting with an agenda to maintain order and move forward
  - Participants are given required documentation and chair solicits input
  - Summarise group's consensus finding
  - If there is no approval:
    - Modeller reworks
    - Final say should be given by the owner of the system

5.1.3 Manage Data Model Versioning and Integration

Preserve the lineage of changes:

- **Why** the project or situation required the change
- **What** and **how** the object changed
- **When** the change was approved and made to the model
- **Who** made the change
- **Where** the change was made (which model)

Modelling tool may have a repository which provides versioning and integration, else preserve models on DDL exports or XML files in a standard source code management system.

5.2 Data Modelling Metrics

Steve Hoberman's Data Model Scorecard® provides a way to measure the quality of a data model.

#	Category	Total score	Model score	%	Comments
1	How well does the model capture the requirements?	15			
2	How complete is the model?	15			
3	How well does the model match its scheme?	10			
4	How structurally sound is the model?	15			
5	How well does the model leverage generic structures?	10			
6	How well does the model follow naming standards?	5			
7	How well has the model been arranged for readability?	5			
8	How good are the definitions?	10			
9	How consistent is the model with the enterprise?	5			
10	How well does the metadata match the data?	10			
	TOTAL SCORE	100			

Description of each category:

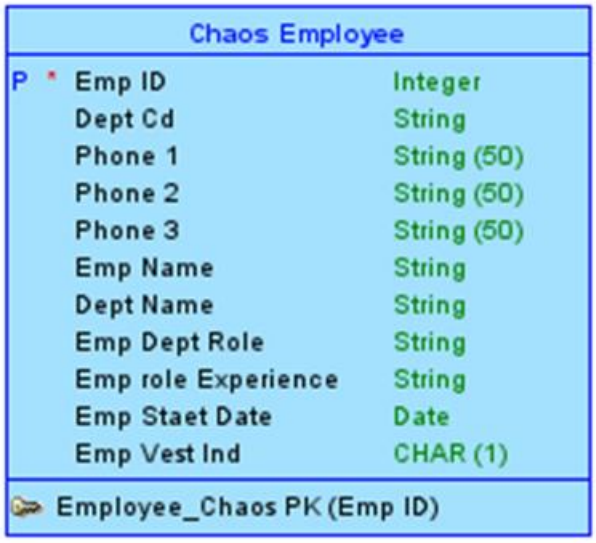
1. **How well does the model capture the requirements?** The model supports all required queries
2. **How complete is the model?** Completeness of requirements and completeness of Metadata, nothing extra
3. **How well does the model match its scheme?** Level of detail (conceptual, logical or physical) and the scheme (Relational, dimensional, NoSQL etc.) matches the definition of the type of model being reviewed.
4. **How structurally sound is the model?** Validate the design practices to ensure a database can be built
5. **How well does the model leverage generic structures?** Appropriate level of abstraction
6. **How well does the model follow naming standards?** Ensure correct and consistent naming standards have been applied to the model
7. **How well has the model been arranged for readability?** Parent entities above child, groupings and shorter relationship lines improve readability
8. **How good are the definitions?** Clear, complete and accurate
9. **How consistent is the model with the enterprise?** If one exists
10. **How well does the Metadata match the data?** Confirm the actual data to be stored fits in the model.

6 Normalisation Example (Steve Hoberman's Mastering Data Modelling Masterclass)

Normalisation

- **First normal form (1NF)**: Valid primary key. Each attribute depends on key. Eliminate repeating groups and ensure each attribute is atomic and not multi valued.
- **Second normal form (2NF)**: Each entity has the minimal primary key and each attribute depends on the complete primary key
- **Third normal form (3NF)**: no hidden primary keys. Each attribute depends on no attributes outside the key ("the key, the whole key and nothing but the key")
- **Boyce / Codd normal form (BCNF)** Resolves overlapping composite candidate keys (hidden business rules)
- **Fourth normal form (4NF)** Resolves all many to many relationships in pairs until they can't be broken down into smaller pieces.
- **Fifth normal form (5NF)**: Resolves inter-entity dependencies into basic pairs, and all join dependencies use parts of primary keys.

Chaos Logical Model



Chaos Employee Data

Emp ID	Dept Cd	Phone 1	Phone 2	Phone 3	Emp Name	Dept Name	Emp Dept Role	Emp Role Experience	Emp Start Date	Emp Vest Ind
1	DG, DM	083-676-9948	083-111-1112, 083-111-1113	083-111-1111	Howard Diesel	Data Governance, Data Modelling	EIM Manager, Data Modeller	5, 10	01-Jan-18	N
2	DM	082-875-5674	083-111-1113	083-111-1111	Veronica Diesel	Data Modelling	Data Modeller	6	01-Jan-18	N
3	MD	083-408-2593	083-111-1114	083-111-1111	Paul Grobler	Master Data	Master Data Manager	8	01-Jan-17	Y

1. Attribute Names are UNCLEAR & NOT single-valued
    1. Phone 1, Phone 2, Phone 3
2. Data IS NOT single-valued
    1. Dept Cd
    2. Phone 2
    3. Dept-Name
    4. Emp Dept Role
    5. Emp Role Experience

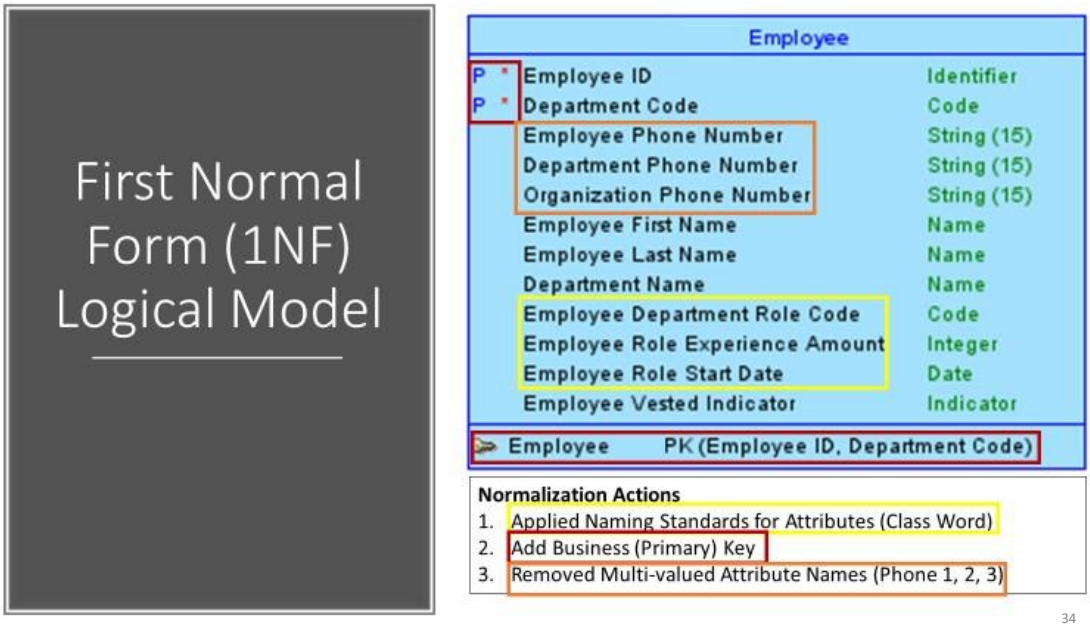
Normalization in a nutshell

Every attribute is single-valued (1NF) and depends completely (2NF) and only on its primary key (3NF).

Steve Hoberman

It is ALL about the Key, the WHOLE Key and NOTHING but the Key

So Help me Codd



1NF Data Observations

	Employee ID	Department Code	Employee Phone Number	Department Phone Number	Organization Phone Number	Employee First Name	Employee Last Name	Department Name	Employee Role Experience Amount	Employee Start Date	Employee Vested Indicator
1	1	DG	083-676-9948	083-111-1112	083-111-1111	Howard	Diesel	Data Governance	5	2018-01-01	N
2	1	DM	083-676-9948	083-111-1113	083-111-1111	Howard	Diesel	Data Modelling	10	2018-01-01	N
3	2	DM	082-875-5674	083-111-1113	083-111-1111	Veronica	Diesel	Data Modelling	6	2018-01-01	N
4	3	MD	083-408-2593	083-111-1114	083-111-1111	Paul	Grobler	Master Data	8	2017-01-01	Y

1. Attribute Names are CLEAR
    1. Department Code
    2. Role Experience Amount (Class Word)
2. Attribute Names are SINGLE-VALUED
    1. Phone 1 – Employee Phone Number
    2. Phone 2 - Department Phone Number
    3. Phone 3 – Organization Phone Number
3. Data IS SINGLE-VALUED
    1. No Data Separators

Second Normal Form (2NF)

Ensure minimal set of attributes that uniquely identify each entity instance

- Are all of the attributes in the primary key needed to retrieve a single instance of [[insert attribute name here]]?

Employee

Employee Identifier Department Code
Employee Phone Number Department Phone Number Organization Phone Number Employee First Name Employee Last Name Department Name Employee Start Date Employee Vested Indicator

Employee		
P *	Employee ID	Identifier
P *	Department Code	Code
	Employee Phone Number	String (15)
	Department Phone Number	String (15)
	Organization Phone Number	String (15)
	Employee First Name	Name
	Employee Last Name	Name
	Department Name	Name
	Employee Department Role Code	Code
	Employee Role Experience Amount	Integer
	Employee Role Start Date	Date
	Employee Vested Indicator	Indicator
Employee	PK (Employee ID, Department Code)	

Questions and Answers for 2NF

- Are all the attributes dependent on the WHOLE key (Employee ID and Department Code)
  - No
    - Employee Name is not dependent on Department Code
    - Department Code is not dependent on Employee Identifier
    - Organization Phone is not dependent on the Employee Identifier
    - Employee Role Experience not dependent on Department Code
    - Employee Role Start Date not dependent on Department Code
  - Yes (the WHOLE key: Employee ID & Department Code)
    - Employee Department Role Code

Second Normal Form (2NF) Logical Model – Step 1

Normalization Actions

1. Separated Employee & Department
2. Created a JOIN table to resolve MANY-TO-MANY

2NF Data

	Employee_ID	Employee_Phone_Number	Employee_First_Name	Employee_Last_Name	Employee_Start_Date	Employee_Vested_Indicator
1	1	083-676-9948	Howard	Diesel	2018-01-01	N
2	2	082-875-5674	Veronica	Diesel	2018-01-01	N
3	3	083-408-2593	Paul	Grobler	2017-01-01	Y

	Employee ID	Department Code	Employee Department Role Name	Employee Role Experience Amount	Employee Start Date
1	1	DG	EIM Manager	5	2018-01-01
2	1	DM	Data Modeller	10	2018-01-01
3	2	DM	Data Modeller	6	2018-01-01
4	3	MD	Master Data Manager	8	2017-01-01

	Department_Code	Department_Name	Department_Phone_Number	Organization_Phone_Number
1	DG	Data Governance	083-111-1112	083-111-1111
2	DM	Data Modelling	083-111-1113	083-111-1111
3	MD	Master Data	083-111-1114	083-111-1111

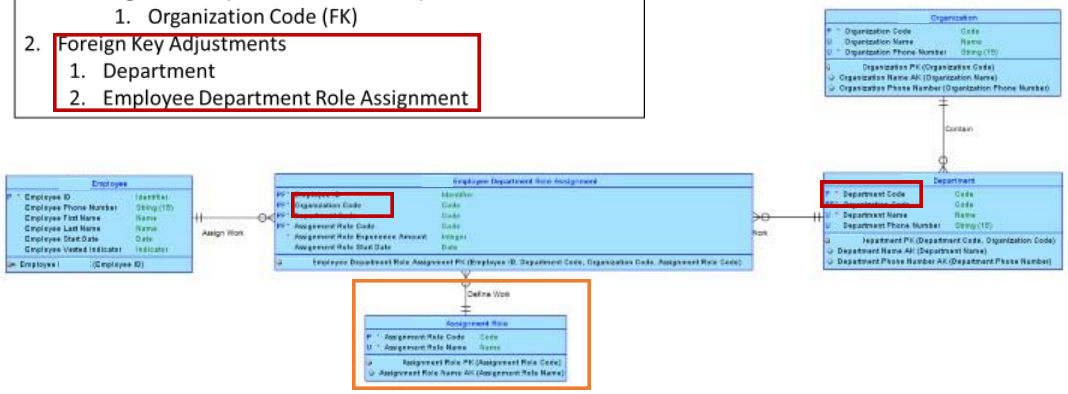
Questions and Answers for 2NF – Step 2

- Are all the attributes dependent on the WHOLE key (Employee ID and Department Code)
  - No (Employee Department Assignment)
    - Employee Department Role Name is not dependent on the Department Code
  - No (Department)
    - Organization Phone is not dependent on the Department Code
  - Ask Business:
    - Is "Role Experience" related the the Employee and Role or the Assignment Experience of the Employee to the Department

Second Normal Form- Step 2

Normalization Actions

1. WHOLE KEY
   1. Assignment Role
   2. Organization (No MANY-TO-MANY)
      1. Organization Code (FK)
2. Foreign Key Adjustments
   1. Department
   2. Employee Department Role Assignment



2NF Data: Step 2

	Employee_ID	Employee_Phone_Number	Employee_First_Name	Employee_Last_Name	Employee_Start_Date	Employee_Vested_Indicator
1	1	083-676-9948	Howard	Diesel	2018-01-01	N
2	2	082-875-5674	Veronica	Diesel	2018-01-01	N
3	3	083-408-2593	Paul	Grobler	2017-01-01	Y

	Employee ID	Organization Code	Department Code	Assignemnt Role Experience Amount	Assignment Role Start Date
1	1	MDS	DG	5	2019-05-09
2	1	MDS	DM	10	2019-05-09
3	2	MDS	DM	6	2019-05-09
4	3	MDS	MD	8	2019-05-09

	Department_Code	Organization_Code	Department_Name	Department_Phone_Number
1	DG	MDS	Data Governance	083-111-1112
2	DM	MDS	Data Modelling	083-111-1113
3	MD	MDS	Master Data	083-111-1114

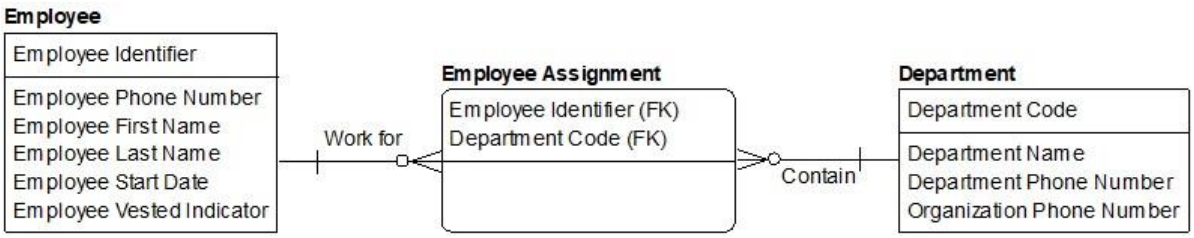
	Organization_Code	Organization_Name	Organization_Phone_Number
1	MDS	Modelware Systems	083-111-1111

	Assignment_Role_Code	Assignment_Role_Name
1	DM	Data Modeller
2	DG	EIM Manager
3	MD	Master Data Manager

Third Normal Form (3NF)

Remove hidden dependencies

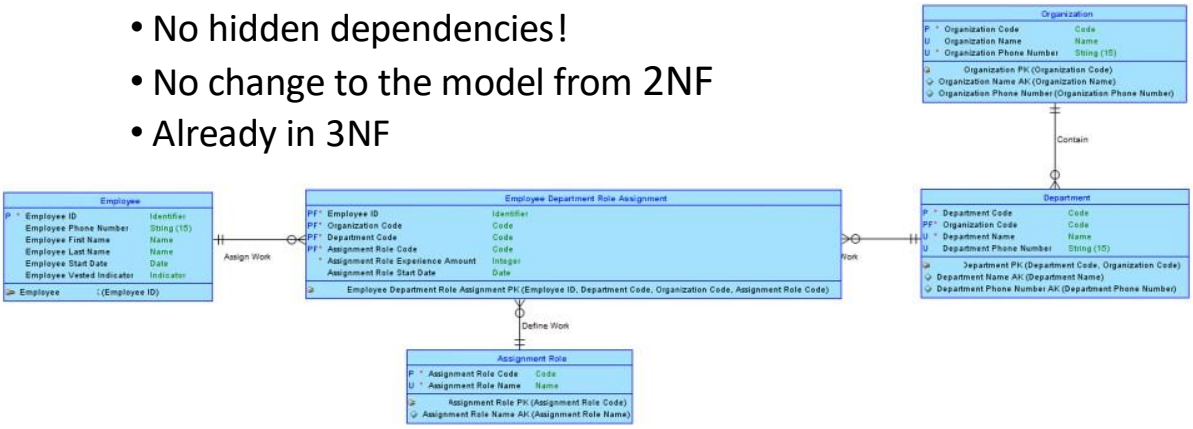
- Is [[insert attribute name here]] a fact about any other attribute in this same entity?



Third Normal Form (3NF)

Remove hidden dependencies

- No hidden dependencies!
- No change to the model from 2NF
- Already in 3NF



Third Normal Form Data

	Employee_ID	Employee_Phone_Number	Employee_First_Name	Employee_Last_Name	Employee_Start_Date	Employee_Vested_Indicator
1	1	083-676-9948	Howard	Diesel	2018-01-01	N
2	2	082-875-5674	Veronica	Diesel	2018-01-01	N
3	3	083-408-2593	Paul	Grobler	2017-01-01	Y

	Employee ID	Organization Code	Department Code	Assignment Role Name	Assignemnt Role Experience Amount	Assignment Role Start Date
1	1	MDS	DG	EIM Manager	5	2019-05-09
2	1	MDS	DM	Data Modeller	10	2019-05-09
3	2	MDS	DM	Data Modeller	6	2019-05-09
4	3	MDS	MD	Master Data Manager	8	2019-05-09

	Department_Code	Organization_Code	Department_Name	Department_Phone_Number
1	DG	MDS	Data Governance	083-111-1112
2	DM	MDS	Data Modelling	083-111-1113
3	MD	MDS	Master Data	083-111-1114

	Organization_Code	Organization_Name	Organization_Phone_Number
1	MDS	Modelware Systems	083-111-1111

	Assignment_Role_Code	Assignment_Role_Name
1	DM	Data Modeller
2	DG	EIM Manager
3	MD	Master Data Manager

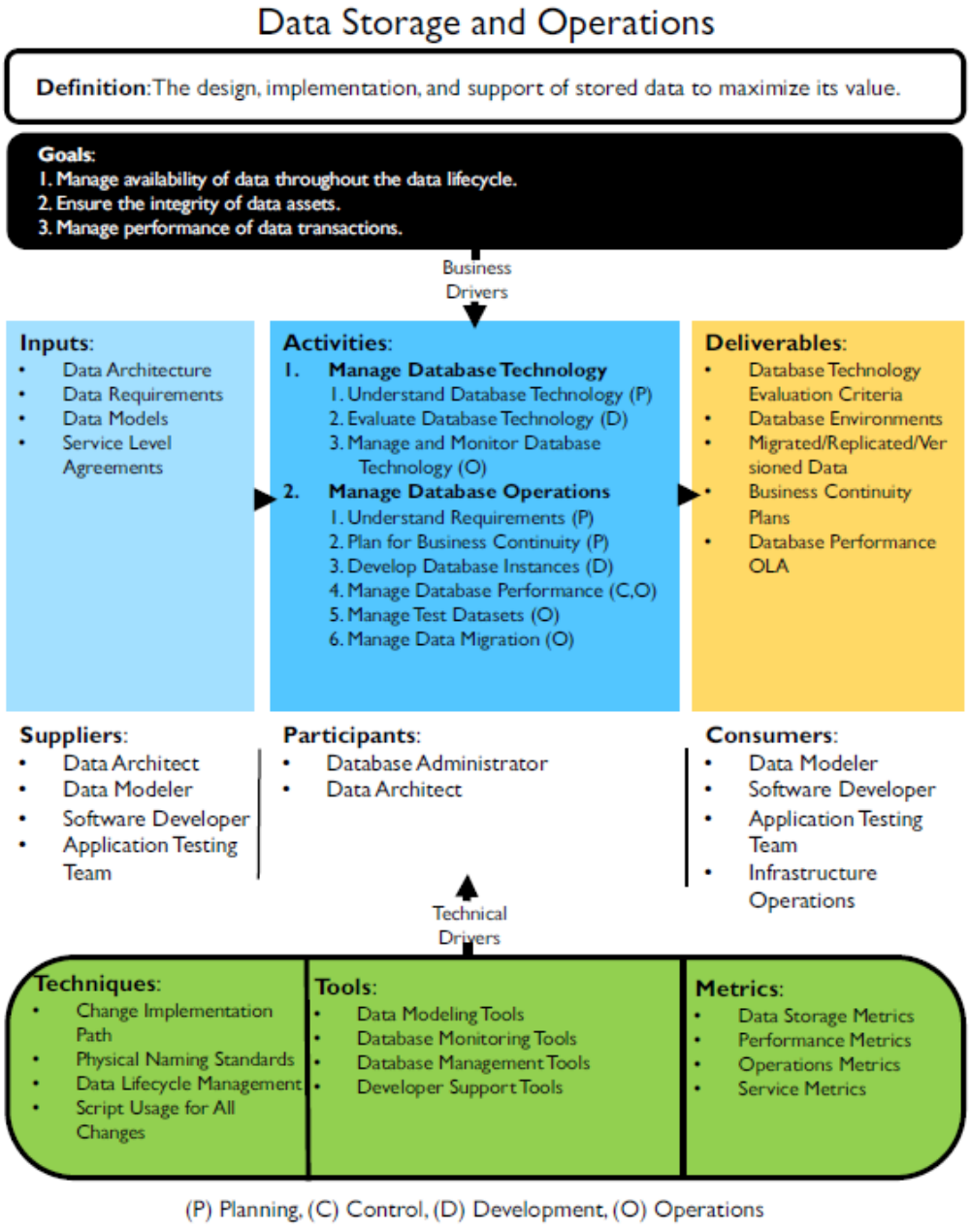
Data Storage and Operations

1 Introduction

True custodians of the data – data at rest.

Data Storage and Operations includes the design, implementation and support of stored data, to maximise its value throughout the lifecycle, from creation to disposal. Two sub-activities:

- **Database support:** activities related to the data lifecycle from implementation of a database environment, through obtaining, backing up and purging data. It includes ensuring the database performs well. Monitoring and tuning.
- **Database technology support:** Defining technical requirements that meet organisational needs, defining technical architecture, installing, maintaining technology, and resolving related issues.



1.1 Business Drivers

Business continuity

1.2 Goals and principles

Goals:
1. Manage availability of data throughout the data lifecycle.
2. Ensure the integrity of data assets.
3. Manage performance of data transactions.

Highly technical side of data management. Guiding principles:

- **Identify and act on automation opportunities:** Automate database processes, develop tools and processes that shorten cycles, reduce errors and development rework. Do in collaboration with data modelling and Data Architecture.
- **Build with reuse in mind:** Develop abstracted and reusable data objects
- **Understand and appropriately apply best practices:**
- **Connect database standards to support requirements:** SLAs reflect DBA-recommended and developer-accepted methods of ensuring integrity and data security.
- **Set expectation for the DBA role in project work:** Include the DBA in all SDLC phases of a project

1.3 Essential Concepts

1.3.1 Database Terms

- **Database:** Any collection of stored data
- **Instance:** An execution of database software controlling access to a certain area of storage.
- **Schema:** A subset of database objects contained within a database or instance, usually with something in common. Used to isolate sensitive data from general user base.
- **Node:** An individual computer part of a distributes database
- **Database abstraction:** An API (common Application Interface) is used to call database functions.

1.3.2 Data Lifecycle Management

The DBA is the custodian of all database changes. The DBA defines the precise changes, implements and controls the changes. The DBA must have a plan to back out the changes.

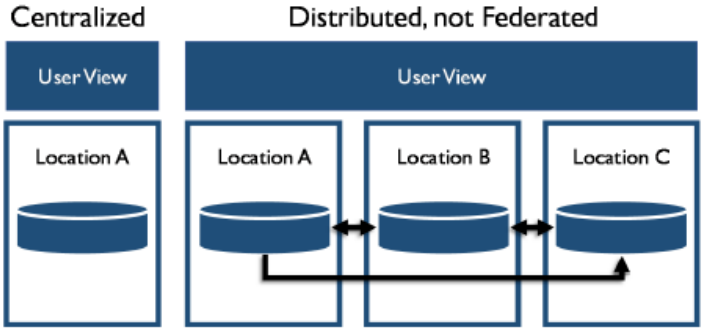
1.3.3 Administrators

- **Production DBA:** Responsible for data operations management
    - Ensures performance and reliability of the database – performance tuning
    - Backup and recovery mechanisms
    - Implement clustering and failover of the database if continual data availability is required
    - Other database maintenance activities such as archiving.
    - Deliverables of Production DBAs:
        - A production database environment
        - Mechanisms and processes for controlled implementation of changes to databases in the production environment
        - Mechanisms for ensuring availability, integrity and recoverability of data in response to all circumstances that could result in loss or corruption of data.
        - Mechanisms for detecting and reporting database errors

- Data availability, recovery and performance in accordance with SLAs
- Mechanisms and processes for monitoring database performance

- **Application DBA:** Responsible for databases in all environments (Development/test, QA and production). Part of an application support team
- **Procedural and Development DBAs:**
  - Administration of procedural database objects (stored procedures, triggers and user-defined functions)
  - Development DBAs focus on creating and managing special use databases such as "sandbox".
- **NSA:** Network Storage Administrators support data storage arrays.

1.3.4 Database Architecture Types



1.3.4.1 Centralised Databases

All the data in one place. All users come to the one system, no alternatives if it is unavailable.

1.3.4.2 Distributed Databases

Quick access to data over many of nodes, each offering local computation and storage. DBMS replicates across nodes, and can detect and handle failures. MapReduce divides a data request into small fragments over the nodes.

1.3.4.2.1 Federated Databases

A federated database system maps multiple autonomous database systems into a single federated database, connected via a network. Remain autonomous but allow sharing of data.

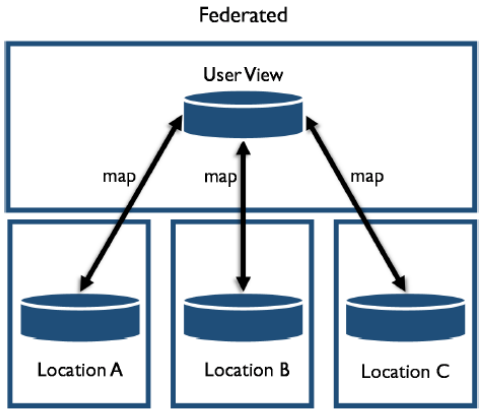
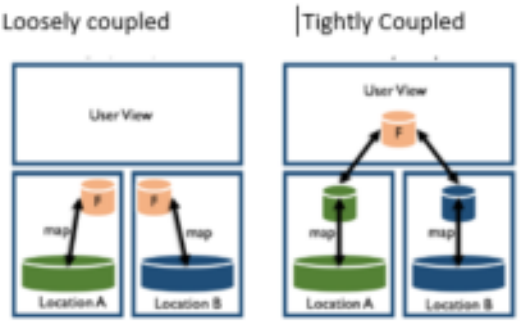


Figure 56 Federated Databases

A FDBMS can be loosely or tightly coupled:

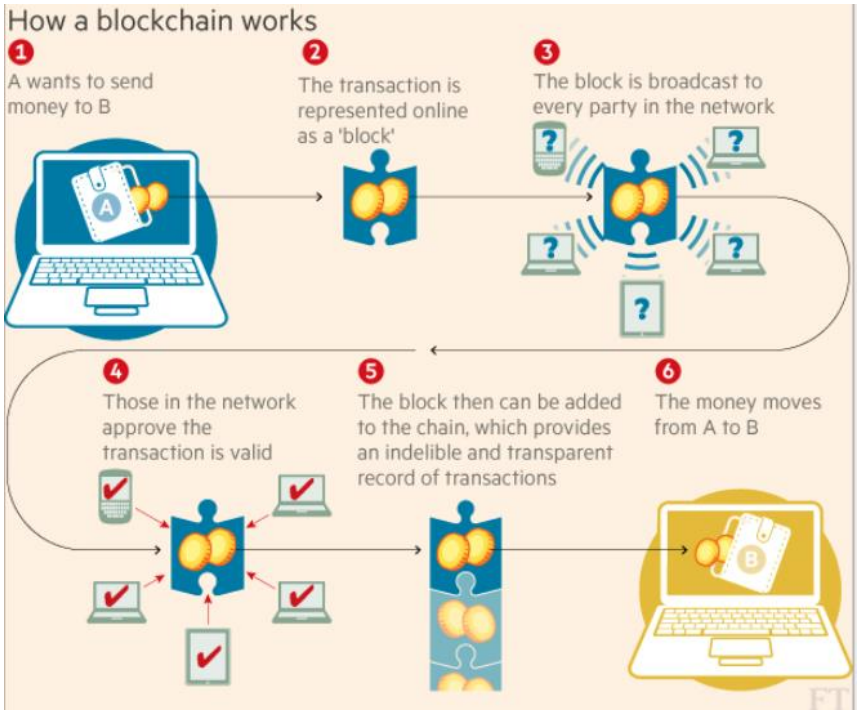
- **Loosely coupled**: Component databases construct their own federated schema.

- **Tightly coupled**: Component systems use independent processes to construct and publish an integrated federated schema



1.3.4.2.2 Blockchain database

A type of federated database used to securely manage financial transactions. Each transaction has a record. The database creates chains of time bound groups (blocks) that also contain information from the previous block in the chain. Hash algorithms create information on the transactions which can never change. Tampering is evident if the hash values no longer match.



1.3.4.3 Virtualisation / Cloud platforms

Virtualisation (Cloud computing) provides computation, software, data access and storage without end-user knowledge of the physical location or configuration of the system.

- **Virtual machine image:** Users purchase virtual machine instances for a limited time. Either load own database or use an optimised installation of the database.
- **Database-as-a-service (DaaS):** Database service provider installs and maintains the database and application owners pay according to their usage.
- **Managed database hosting in the cloud:** The cloud provider hosts the database and manages it on the application owner's behalf.

DBAs and Network and System Administrators need to establish a systematic project approach to the following functions (as well as the security aspects):

- **Standardisation/consolidation:** Based on Data Governance policy, consolidation reduces the number of data storage locations and standard procedures are developed by DBAs and Data Architects.
- **Server virtualisation:** Enables reduction of cost in infrastructure management.
- **Automation:** of data tasks
- **Security:** Integrate security of virtual systems with security of physical infrastructures

1.3.5 Database Processing Types

ACID: 1980s – reliability within database transactions. Relational – SQL Server.

- **Atomicity:** All operations are performed, or none are. If one fails, the entire transaction fails
- **Consistency:** the transaction must meet all rules defined by the system at all times and must void half-completed transaction
- **Isolation:** Each transaction is independent
- **Durability:** Once complete, the transaction cannot be undone

BASE: Increase in volumes and variability of data, the need to document and store **l**ess structured data, read-optimised data workloads, greater flexibility of scaling and design

- **Basically available:** System guarantees some level of data availability even if there are node failures
- **Soft state:** System in constant state of flux. Data may be available but not current
- **Eventual consistency:** Data eventually becomes consistent through nodes.

Item	ACID	BASE
Casting (data structure)	Schema must exist	Dynamic
	Table structure exists	Adjust on the fly
	Columns data typed	Store dissimilar data
Consistency	Strong Consistency Available	Strong, Eventual, or None
Processing Focus	Transactional	Key-value stores
Processing Focus	Row/Column	Wide-column stores
History	1970s application storage	2000s unstructured storage
Scaling	Product Dependent	Automatically spreads data across commodity servers
Origin	Mixture	Open-source
Transaction	Yes	Possible

- **CAP:** Brewers theorem – the larger the distributed system, the lower the compliance to ACID. At most two of the following properties can exist in a shared data system:
  - **Consistency:** The system must operate as designed and expected at all times
  - **Availability:** The system must be available and respond to each request
  - **Partition Tolerance:** The system must be able to continue operations during occasions of data loss or partial system failure.

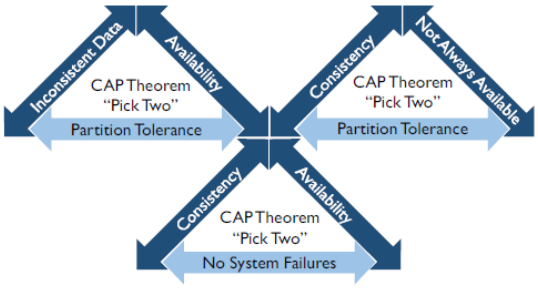


Figure 58 CAP Theorem

Used to drive Lambda Architecture for big data, which uses two paths for data: (refers to ch14 where it appears as 14.1.3.7 Services-Based Architecture (SBA))

- Speed path: Availability and Partition Tolerance
- Batch path: Consistency and Availability

1.3.6 Data Storage Media

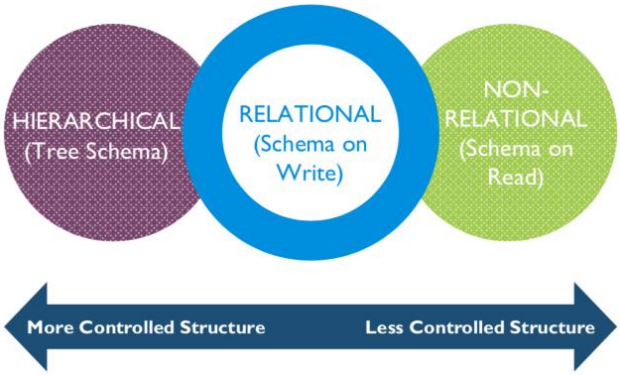
- **Disk and Storage Area Networks (SAN):** Disk arrays and SAN. Persistent storage.
- **In-Memory:** In-Memory Databases (IMDB) are loaded into volatile memory where all processing takes place. Faster response than disk.
- **Columnar Compression Solution:** For data sets where data values are repeated to a large extent and may be compressed. Reduces I/O time
- **Flash Memory:** Solid state

1.3.7 Database Environments

There are various environments used during the system development lifecycle.

- **Production Environments:** The technical environment where all business processes occur. The DBA team should be the only team implementing changes, adhering strictly to standards and procedures.
- **Pre-Production Environments:**
    - **Development:** Slimmer version of production for developers to test new code or patches
    - **Test:** Used for:
        - **Quality Assurance testing (QA)**
        - **Integration testing**
        - **User Acceptance Testing (UAT)**
        - **Performance Testing**
    - **Sandboxes or Experimental Environments:** Allows read only access to production data for experimentation.

1.3.8 Database Organisation



1.3.8.1 Hierarchical

Oldest database model. Tree structure with one to many parent-child relationships.

1.3.8.2 Relational

Based on set theory and relational algebra. Set operations (union, intersection, minus) in the form of SQL (Structured Query Language) are used to retrieve data. To write data the schema must be known (schema on write). Relational databases are row-oriented. The database management system is called RDBMS

- **Multidimensional:** Allows searching using several data element filters simultaneously. Uses a type of SQL called MDX (Multidimensional eXpression)
- **Temporal:** Built in support for handling time.

1.3.8.3 Non-relational (NoSQL – Not only SQL)

Schema-on-read allows data to be read in different ways. NoSQL means the storage structure is not bound by tabular design. NoSQL databases are used in Big Data and real-time web applications as they are optimised for retrieval and appending operations.

- **Column-oriented:** Used in BI applications as redundant data can be compressed. Difference between column-oriented and row-oriented (usually relational):
    - **Column-oriented is more efficient when:**
        - aggregating over many rows
        - new values for that column are applied at once as there is no need to touch columns for the other rows
        - Online Analytical Processing (OLAP)
    - **Row-oriented is more efficient when:**
        - Many columns of a single row required at once
        - Writing a whole row of new data
        - Online Transaction Processing (OLTP)
- **Spatial:** Store and query data that represents objects defined in geometrical space. Use special indexes to perform database operations. Open Geospatial Consortium standard.
- **Object/Multi-media:** Hierarchical storage Management System manages the media objects
- **Flat File Database:** Data in rows and columns as a single file. Used by Hadoop databases.
- **Key-Value Pair:** A key identifier and a value:
    - **Document Databases:** Each document has a key. Use XML or JSON (Java Script Object Notation) structures.
    - **Graph Databases:** Key-value pairs where the focus is on the relationship between nodes rather than the nodes themselves.

- **Triplestore:** A data entity composed of subject-predicate-object. Best for taxonomy and thesaurus management. Resource Description Framework (RDF):
    - **subject:** a resource
    - **predicate:** relationship between the subject and object
    - **object:** the object

1.3.9 Specialised Databases

- Computer Assisted Design and Manufacturing (CAD/CAM): Object database
- Geographical Information Systems (GIS): Geospatial databases
- Shopping-cart applications: XML databases to store customer order data

1.3.10 Common Database Practices

1.3.10.1 Archiving

The process of moving data off immediately accessible media onto less expensive media with lower retrieval performance. Schedule regular restoration tests.

Archival processes must be aligned with the partitioning strategy to ensure optimal availability and retention:

- Create a secondary area on a secondary database server
- Partition database tables into archival blocks
- Replicate to the separate database
- Create backups (tape or disk)
- Create jobs that periodically purge unneeded data

1.3.10.2 Capacity and growth projections

Decide how much storage is needed, and work out expansion needs

1.3.10.3 Change Data Capture (CDC)

Detecting that data has changed and ensure information relevant to the change is stored appropriately. Log-based replication.

1.3.10.4 Purging

Purging is the process of completely removing data from media so that is cannot be retrieved.

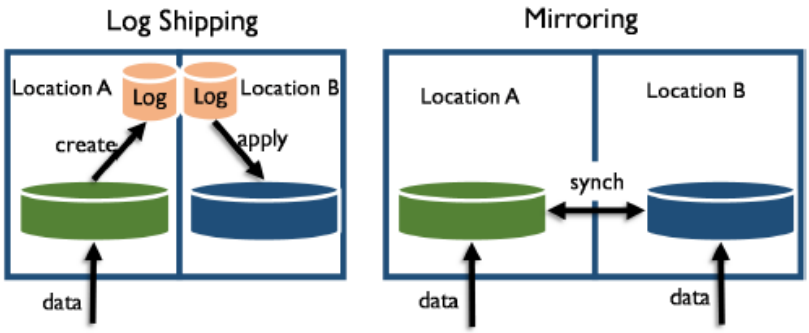
1.3.10.5 Replication

Replication transparency: The same data is stored on multiple storage devices and is consistent throughout the database system, so that users cannot tell which database copy they are using.

- **Active replication:** Create and store the same data at every replica from every other replica
- **Passive replication:** recreating and storing data on the primary, then transferring it to secondary replicas

Two primary replication patterns:

- **Mirroring:** Updates to the primary are replicated immediately to the secondary as part of a two-phase commit**.**
- **Log Shipping:** The secondary receives and applies copies of the primary database's transaction log at regular intervals



1.3.10.6 Resiliency and Recovery

Resiliency is a measure of how tolerant a system is to error conditions, and how well it recovers. Three recovery types:

- **Immediate recovery:** Predicting and automatically resolving issues
- **Critical recovery:** Plan to restore the system quickly to minimise delays to business processes
- **Non-critical recovery**: Restoration can be delayed until critical systems have been restored

1.3.10.7 Retention

How long data is kept available. Affects capacity planning. Data retention plans are also affected by Data Security, as some data has legal requirements to be retained a specific time.

1.3.10.8 Sharding

Small chunks of the database are isolated so replication is merely a file copy.

2 Activities

Data Technology Support (selecting and maintaining the software that stores and manages the data) and Data Operations Support (the data and processes that the software manages).

2.1 Manage Database Technology

The Information Technology Infrastructure Library (ITIL) is the leading reference model.

2.1.1 Understand Database Technology Characteristics

DBAs and Data Architects combine their knowledge of available tools with business requirements to suggest the best technology.

2.1.2 Evaluate Database Technology

Some of the factors to consider when selecting a DBMS:

- Product architecture and complexity
- Volume and velocity limits
- application profile such as transaction processing and BI
- Hardware platform and operating system support
- Availability of supporting software tools
- Performance benchmarks
- Scalability
- Software, memory and storage requirements
- Resiliency, including error handling and reporting

Price and the possible necessity to employ extra staff.

2.1.3 Manage and Monitor Database Technology

DBAs need to be trained to function as Level 2 technical support. They need to have a working knowledge of application development skills. DBAs work with business users and application developers to ensure the most effective use of technology.

2.2 Manage Databases

Database support is provided by DBAs and NSAs (Network Storage Administrators).

2.2.1 Understand Requirements

- **Define storage requirements:** Initial capacity estimate for first year and growth projection for next few years. Take data storage, indexes, logs and mirrors into account.
- **Identify usage patterns:** Databases have predictable usage patterns:
  - Transaction based
  - Large data set write or retrieval
  - Time based – month-end, lighter on weekends
  - Location based – more populated areas have more transactions
  - Priority based – some departments or batch IDs have higher priority
- **Define access requirements:** The authorisation to access different data files, and the standard languages and methods to do it.

2.2.2 Plan for Business Continuity

Recovery plan for all databases and database servers in the event of a disaster which could result in loss or corruption of data. Identify critical databases which need to be restored first.

This plan should be reviewed by the business continuity group.

- **Make backups:**
  - Back up databases and transaction logs
  - Backup frequency determined by SLA.
  - Incremental and complete backups
  - Keep backups on a separate file system
- **Recover data:** DBA executes restoration of data. Test recovery periodically

2.2.3 Develop Database Instances

- Installing and updating DBMS software
- Maintaining multiple environment installations, including different DBMS versions
- Installing and administering related data technology

2.2.3.1 Manage the Physical Storage Environment

Storage environment management needs to follow Software Configuration Management (SCM) processes or ITIL methods to record modification to the database configuration. Four processes:

- **Configuration Identification:** DBAs with Data Stewards, Data Architects and Data Modellers to identify attributes that define end-user configuration. These must be baselined, recorded and only changed with formal change control.
- **Configuration change control:** Processes and approval stages to change the above attributes
- **Configuration status accounting:** Report on the configuration at any point in time
- **Configuration audits:**
  - Physical configuration audit: an item is installed in accordance with design documentation

- Functional configuration audit: Performance attributes of an item are achieved

DBAs must communicate any changes to the physical attributes to modellers, developers and Metadata managers.

DBAs also maintain metrics on data volume, capacity projections, query performance and statistics on the physical objects.

2.2.3.2 Manage Database Access Controls

DBAs oversee the following functions to protect data assets:

- **Controlled Environment:** DBAs and NSAs. Network roles and permissions, 24/7 monitoring, firewall management, patch management
- **Physical security:** Simple Network Management Protocol (SNMP)-based monitoring, data audit logging, disaster management and database backup plans.
- **Monitoring:** Continuous monitoring of servers
- **Controls:** Access controls, database auditing, intrusion detection and vulnerability assessment tools

2.2.3.3 Create Storage Containers

All data must be stored on the physical drive and organised for ease of load, search and retrieval.

2.2.3.4 Implement Physical Data Models

DBAs implement the physical layout of the data model in storage. The physical data model includes storage objects, indexing objects, and any encapsulated code objects required to enforce quality rules, connect database objects and achieve database performance.

2.2.3.5 Load Data

DBMS should have a bulk load facility to load data into a new database. Data must be in the right format for the target object.

Third party data can be updated regularly by the service. DBAs must be aware of any legal restrictions before loading third party data

DBAs may load data manually, or automate and schedule loading.

Responsibility for data acquisition services in a managed environment is centralised with data analysts who document it in the logical data model. Developers create scripts if necessary. DBA implements the processes to load the data into the database.

2.2.3.6 Manage Data Replication

DBAs advise data replication decisions on:

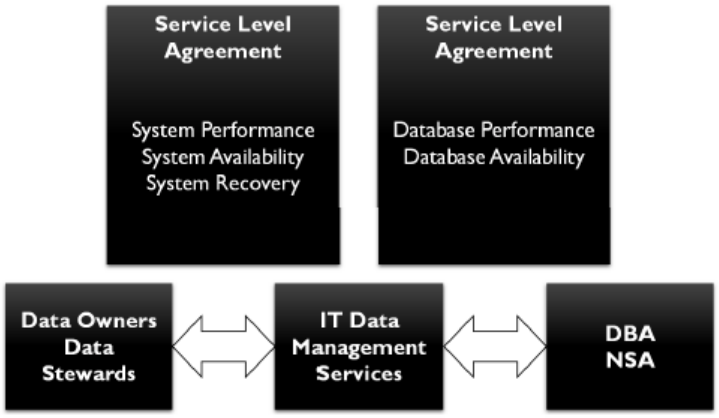
- Active or passive replicating
- Distributed concurrency control from distributed data systems
- The appropriate methods to identify updates to data under the Change Data Control process:
  - Timestamp
  - Version numbers

2.2.4 Manage Database Performance

Database performance depends on availability and speed. Availability of space and query optimisation are part of performance.

2.2.4.1 Set Database performance service levels

System performance, data availability and recovery expectations, and expectations for teams to respond are governed by Service Level Agreements (SLAs) between IT data management services and data owners.



2.2.4.2 Manage Database Availability

Availability is the percentage of time that a system or database is available for productive work. Four factors affect availability:

- **Manageability:** The ability to create and maintain an environment
- **Recoverability:** Establish service after interruption, and correct the errors caused
- **Reliability:** Ability to deliver service at specified levels for a stated period
- **Serviceability:** Identify and diagnose problems and solve them

Some things that prevent databases from being available:

- Planned outages
- Unplanned outages
- Application problems
- Data problems
- Human error

DBAs are responsible for doing everything possible to ensure databases stay online and operational:

- Run database backup utilities
- Run database reorganisation utilities
- Run statistics gathering utilities
- Run integrity gathering utilities
- Automating the above utilities
- Exploiting table clustering and partitioning
- Replicating data across mirror databases to ensure high availability

2.2.4.3 Manage Database Execution

DBAs manage database execution, logging and log sizes and synchronisation.

2.2.4.4 Maintain database performance service levels

DBAs generate performance analysis reports regularly and compare them with previous reports to identify negative trends and analyse problems over time.

- **Transaction performance vs Batch performance**: Batch jobs must complete within a batch window
- **Issue remediation:** Common reasons for poor database performance are:
    - Memory allocation or contention:
    - Locking and blocking
    - Inaccurate database statistics
    - Poor coding
    - Inefficient complex table joins
    - Insufficient indexing
    - Application activity
    - Overloaded servers
    - Database volatility
    - Runaway queries

2.2.4.5 Maintain alternate environments

Types of alternate environments:

- **Development:** Test changes that will be implemented in production
- **Test:** QA, Integration testing, UAT and performance testing
- **Sandboxes:** Experimental environments
- **Alternate production environments:** support failover, offline backups and resiliency support systems

2.2.5 Manage Test Data Sets

Efficient testing requires high quality test data to be generated and managed

2.2.6 Manage data Migration

Data Migration is the process of transferring data between storage types, formats or computer systems with as little change as possible. Usually performed programmatically, automated based on rules.

3 Tools

- Data modelling tools
- Database monitoring tools
- Database Management tools
- Developer support tools

4 Techniques

4.1 Test in lower environments

Test upgrades and patches on the lowest level first, development. Then install and test on higher levels, production last.

4.2 Physical naming standards

ISO/IEC 11179 – Metadata Registries (MDR)

4.3 Script usage for all changes

Test any change scripts in non-production before applying.

5 Implementation Guidelines

5.1 Readiness assessment/Risk assessment

Two central ideas:

- **Data Loss**: SLAs specify general requirements for protection. Ongoing assessment to ensure robust technical responses to cyber threat are in place.
- **Technology readiness**: Does the organisation have the skill set to implement newer technology.

5.2 Organisation and Cultural Change

- **Proactively communicate**: DBAs should be in close communication with project teams at all stages of the project, to detect and resolve issues as early as possible
- **Communicate with people on their level and in their terms**
- **Stay business focussed:** objective is to meet business requirements and derive maximum value
- **Be helpful:** Not helping may force people to ignore standards and find another way
- **Learn continually:** Setbacks and problems are lessons which can be applied later.

Understand stakeholders and their needs. Develop clear, concise, practical, business focussed standards for doing the best possible work in the best possible way. Teach and implement those standards in a way that provides maximum value to stakeholders and earns their respect.

6 Data Storage and Operations Governance

6.1 Metrics

Storage metrics may include:

- Count of databases by type
- aggregated transaction stats
- Capacity metrics
- Storage service usage
- requests made against storage services
- Performance improvements of applications using service

Performance metrics:

- Transaction frequency and quantity
- Query performance
- API service performance

Operational metrics:

- Aggregated statistics about data's retrieval times
- Backup size
- Data quality measurement
- Availability

Service metrics:

- Issue submission, resolution and escalation count by type
- Issue resolution time

6.2 Information asset Tracking

Ensure organisation complies with software licensing and annual support agreements.

Determine TCO (total cost to ownership) for each technology product.

6.3 Data Audits and Data validation

A data audit is the evaluation of data based on defined criteria, typically performed to investigate specific concerns about the data. A data audit includes:

- Project specific checklist
- comprehensive checklist
- Required deliverables
- Quality control criteria

Data validation is the process of evaluating stored data against established acceptance criteria (from the Data Quality team or customer specifications) to determine its quality and usability.

DBAs support data audits and validation by:

- Help develop and review the approach
- Perform preliminary data screening and review
- Develop data monitoring methods
- Applying statistical, geo-statistical and bio-statistical techniques to optimise the data
- Support sampling and analysis
- Review data
- Provide support for data discovery
- Act as the SME for questions related to database administration

Data Security

1 Introduction

Effective data security policies and procedures ensure the right people can use and update data the right way, and all inappropriate access and update is restricted (Ray, 2012).

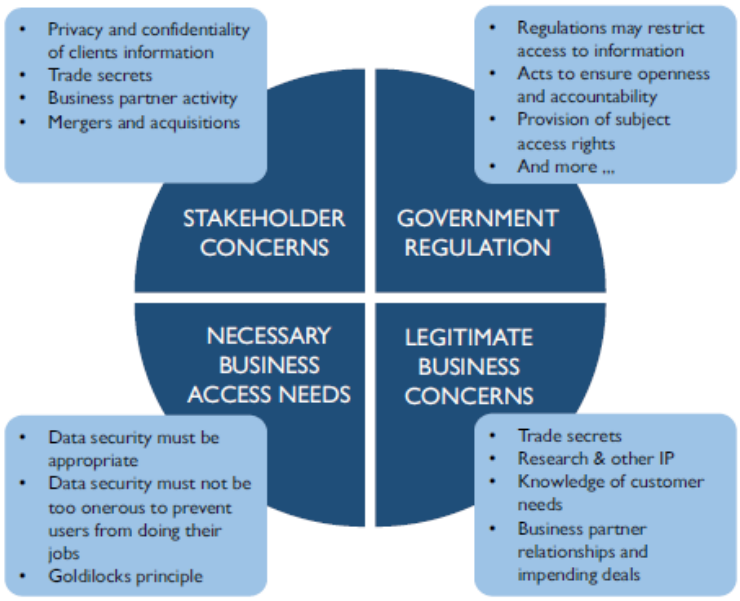
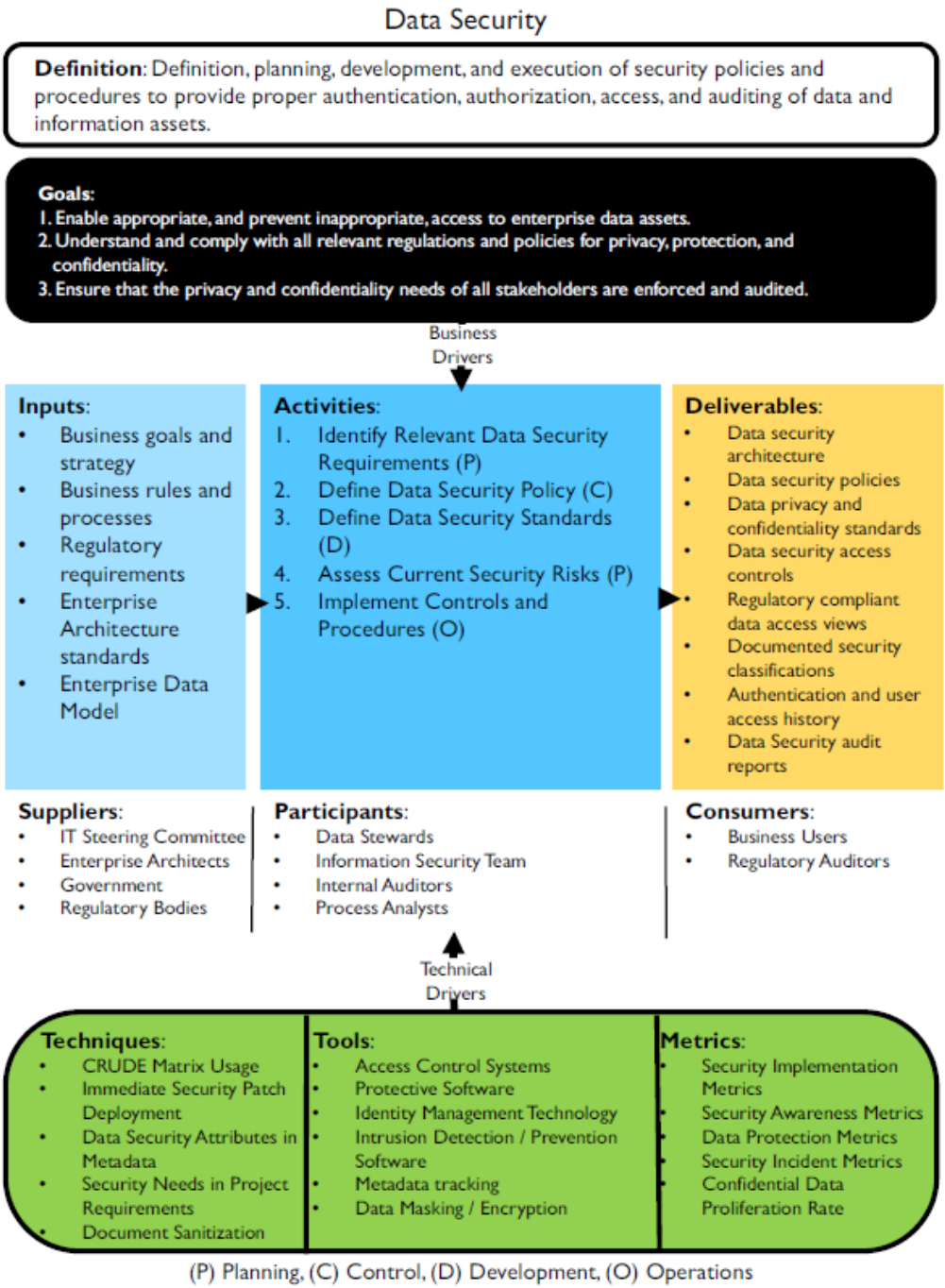


Figure 62 Sources of Data Security Requirements

- **Stakeholders:** Everyone in an organisation must be a responsible trustee of stakeholders' data
- **Government regulations:** Some restrict access and others ensure openness, transparency and accountability
- **Proprietary business concerns:** Data which provides competitive advantage must be protected.
- **Legitimate access needs:**
- **Contractual obligations:** Contractual and non-disclosure agreements



1.1 Business Drivers

Data security risks are associated with regulatory compliance, fiduciary responsibility, reputation and a legal and moral responsibility to protect private and sensitive data. Mitigating risks and growing the business can be complimentary when integrated into an information protection strategy.

- **Risk reduction:** Should be enterprise wide. Classify organisation's data:
  - Identify and classify sensitive data assets
  - Locate sensitive data throughout the enterprise
  - Determine how each asset needs to be protected
  - Identify how the information interacts with business processes
  - Identify external (hackers) and internal (employees and processes) threats.

- **Business Growth:** Data security breaches can impact business growth. Robust data security inspires consumer confidence. Trusted e-commerce drives profit and growth.
- **Security as an asset:** Tag data with security metadata

1.2 Goals and Principles

Goals:

- Enabling appropriate access and preventing inappropriate access to enterprise data assets
- Enabling compliance with regulations and policies for privacy, protection and confidentiality
- Enabling that stakeholder requirements for privacy and confidentiality are met.

Principles:

- **Collaboration**: Data Security is a collaborative effort involving IT security administrators, data stewards/data governance, internal and external audit teams, and the legal department.
- **Enterprise approach**: Data Security standards and policies must be applied consistently across the entire organization.
- **Proactive management**: Success in data security management depends on being proactive and dynamic, engaging all stakeholders, managing change, and overcoming organizational or cultural bottlenecks such as traditional separation of responsibilities between information security, information technology, data administration, and business stakeholders.
- **Clear accountability**: Roles and responsibilities must be clearly defined, including the 'chain of custody' for data across organizations and roles.
- **Metadata-driven**: Security classification for data elements is an essential part of data definitions.
- **Reduce risk by reducing exposure**: Minimize sensitive/confidential data proliferation, especially to non-production environments.

1.3 Essential Concepts

1.3.1 Vulnerability

A **vulnerability** is a weaknesses or defect in a system that allows it to be successfully attacked and compromised – essentially a hole in an organization's defences. Some vulnerabilities are called exploits.

1.3.2 Threat

A **threat** is a potential offensive action that could be taken against an organization. Threats can be internal or external. They are not always malicious. A uninformed insider can take offensive actions again the organization without even knowing it.

1.3.3 Risk

The term **risk** refers both to the possibility of loss and to the thing or condition that poses the potential loss. Risk calculations:

- Probability that the threat will occur and its likely frequency
- The type and amount of damage created each occurrence might cause, including damage to reputation
- The effect damage will have on revenue or business operations
- The cost to fix the damage after an occurrence
- The cost to prevent the threat, including by remediation of vulnerabilities

- The goal or intent of the probable attacker

1.3.4 Risk classifications:

Describe the sensitivity of the data and the likelihood it may be sought after for malicious purposes.

- **Critical Risk Data (CRD)**: Personal information aggressively sought for unauthorized use by both internal and external parties due to its high direct financial value. Compromise of CRD would not only harm individuals, but would result in financial harm to the company from significant penalties, costs to retain customers and employees, as well as harm to brand and reputation.
- **High Risk Data (HRD)**: HRD is actively sought for unauthorized use due to its potential direct financial value. HRD provides the company with a competitive edge. If compromised, it could expose the company to financial harm through loss of opportunity. Loss of HRD can cause mistrust leading to the loss of business and may result in legal exposure, regulatory fines and penalties, as well as damage to brand and reputation.
- **Moderate Risk Data (MRD)**: Company information that has little tangible value to unauthorized parties; however, the unauthorized use of this non-public information would likely have a negative effect on the company.

1.3.5 Data Security Organisation

The Information Security Function depends on the size of the enterprise. May be:

- Dedicated Information Security group within IT
- Chief Information Security officer (CISO) reporting to the CIO or CEO
- Smaller organisations data security is the responsibility of data managers

Dedicated Information Security personnel are most concerned with the technical aspects, such as combating malicious software attacks. Data Management are concerned with regulatory aspects. A standard sharing process should be in place where both groups are kept informed of data regulations, data loss threats and data protection requirements.

NIST (National Institute of Standards and Technology) Risk Management Framework:

- All enterprise information must be categorised.
- The location of all sensitive information must be known.
- Enterprise data model is essential

Data managers, IT Developers and cyber security professionals work together to:

- identify regulated data so that
- sensitive systems are protected
- User access controls designed to
- Enforce confidentiality, integrity and data regulatory compliance

1.3.6 Security Processes

the Four As and an E

- **Access**: Enable individuals with authorization to access systems in a timely manner.
- **Audit**: Review security actions and user activity to ensure compliance with regulations and conformance with company policy and standards.
- **Authentication**: Validate users' access.

- **Authorization**: Grant individuals privileges to access specific views of data, appropriate to their roles.
- **Entitlement**: An Entitlement is the sum-total of all the data elements that are exposed to a user by a single access authorization decision

Systems should include monitoring controls that detect unexpected events. Real time active monitoring for confidential information. System interruption if an event that does not follow procedure occurs. Passive monitoring takes snapshots at regular intervals.

1.3.7 Data Integrity

In security, data security is the state of being whole, protected from improper deletion, alteration or addition.

1.3.8 Encryption

The process of translating plain text into complex codes to hide privileged information, verify complete transmission or verify the sender's identity. Cannot be read without the decryption key. Four main methods of encryption:

- **Hash:** Uses algorithms
- **Symmetric**
- **Private-key:** Both sender and receiver have the same key.
- **Public-key:** Sender uses a public key that is freely available and receiver uses a private key

1.3.9 Obfuscation or masking

The appearance of the data is changed. two types of data masking, Persistent and Dynamic:

- **Persistent data masking:** Permanently and irreversibly alters the data. Used for test environments
    - **In-flight persistent masking:** Data is masked when it is moving from source (production) to destination (non-production). Secure as there is no intermediate file
    - **In-place persistent masking:** Source and destination are the same**.** Unmasked data is read, masked then written over the unmasked data
- **Dynamic data masking:** Makes changes to appearance of data to the end user system without changing the underlying data
- **Masking methods:**
    - **Substitution**: Replace characters or whole values with those in a lookup or as a standard pattern. For example, first names can be replaced with random values from a list.
    - **Shuffling**: Swap data elements of the same type within a record, or swap data elements of one attribute between rows. For example, mixing vendor names among supplier invoices such that the original supplier is replaced with a different valid supplier on an invoice.
    - **Temporal variance**: Move dates +/– a number of days – small enough to preserve trends, but significant enough to render them non-identifiable.
    - **Value variance**: Apply a random factor +/– a percent, again small enough to preserve trends, but significant enough to be non-identifiable.
    - **Nulling or deleting**: Remove data that should not be present in a test system.
    - **Randomization**: Replace part or all of data elements with either random characters or a series of a single character.

- **Encryption**: Convert a recognizably meaningful character stream to an unrecognizable character stream by means of a cipher code. An extreme version of obfuscation in-place.
- **Expression masking**: Change all values to the result of an expression. For example, a simple expression would just hard code all values in a large free form database field (that could potentially contain confidential data) to be 'This is a comment field'.
- **Key masking**: Designate that the result of the masking algorithm/process must be unique and repeatable because it is being used mask a database key field (or similar). This type of masking is extremely important for testing to maintain integrity around the organization.

1.3.10 Network Security Terms: Data-in-motion

- **Backdoor:** A hidden entry to a computer system bypassing password requirements. Usually left by developers for maintenance.
- **Bot or Zombie:** A workstation taken over by a Trojan, Virus, Phish or download of an infected file. Bots are remotely controlled to perform malicious tasks.
- **Cookie:** Small data file an internet commerce website installs on a computer's hard drive to identify returning visitors and their preferences. Could be used by spyware.
- **Firewall**: Software and/or hardware that filters network traffic to protect against unauthorised access or attack.
- **Perimeter**: Boundary between organisation's systems and outside. Firewall sits here.
- **DMZ**: De-militarised Zone. Located between the perimeter firewall and a firewall between it and the internet. Used to pass and temporarily store information moving between organisations
- **Super User Account**: Administrator access to be used in an emergency. Credentials are highly secured and controlled by time, location and user ID.
- **Key Logger**: Attack software that captures keystrokes
- **Penetration testing**: An ethical hacker tries to expose vulnerabilities.
- **Virtual Private Network (VPN)**: Use the unsecured internet to create an encrypted tunnel

1.3.11 Types of Data Security

Data security involves not just preventing inappropriate access, but also enabling appropriate access to data. Access to sensitive data is controlled by granting permissions (opt-in).

- **Facility Security:** Locked data centre
- **Device Security:** Standards for portable devices
  - Access policies regarding connection using mobile devices
  - Storage of data on portable devices
  - Data wiping and disposal of devices in compliance with records management processes
  - Installation of anti-malware and encryption software
  - Awareness of security vulnerabilities
- **Credential Security:** Each user is assigned User ID and Password to access system
  - **Identity management Systems:** Single sign-on gets user onto many systems
  - **User ID Standards for email systems:** Unique within the system. Usually use the user's name in some way.
  - **Password standards:** First line of defence. Should be "strong" and changed every 45-180 days.

    - **Multiple Factor Identification:** Additional identification – code to phone, hardware, biometric.
- **Electronic Communication Security:** Train users not to send confidential information over email or other insecure applications

1.3.12 Types of Data Security Restrictions

Security restrictions are driven by Confidentiality and Regulation:

- **Confidential Data**: Confidential means secret or private and is shared on a "need-to-know" basis. Internally defined. Confidentiality level of a data set depends on the most sensitive item. Typical classification schema:
    - **For General Audiences**: available to anyone
    - **Internal use only**: employees or members. May be discussed but not copied outside the organisation
    - **Confidential**: cannot be shared outside the organisation without a non-disclosure agreement
    - **Restricted Confidential**: Need-to-know
    - **Registered Confidential**: Legal agreement to assume responsibility for the data's secrecy must be signed.
- **Regulated Data:** Regulatory categories are assigned according to laws. Shared on an "allowed-to-know' basis. Externally defined. A single data set may have multiple regulatory categories. It is a good idea to collect many nations personal data privacy laws into a single standard to enforce, achieving international compliance. Sample regulatory families:
    - **Personal Identification Information (PII):** Personally Private Information (PPI). Any information that can identify an individual. EU Privacy Directives
    - **Financially Sensitive Data:** In the US covered by Insider Trading Laws
    - **Medically Sensitive Data / Personal Health Information (PHI):** US covered by HIPAA (Health Information Portability and Accountability Act)
    - **Educational Records:** US covered by FERPA (Family Educational Rights and Privacy Act)
- **Industry or Contract-based Regulations**:
    - **Payment Card Industry Data Security Standard (PCI-DSS):** Any information that can identify an individual with an account at a financial institution.
    - **Competitive advantage or trade secrets**: Protected by industry regulations / Intellectual property laws
    - **Contractual restrictions**: An organisation may restrict how information may be shared in their contracts with others.

1.3.13 System Security Risks

Identify risks inherent in systems:

- **Abuse of excessive privilege:** Principle of least privilege should be applied. Query-level access control is better, but time consuming to set up.
- **Abuse of legitimate privilege:** For unauthorised purposes. Enforce policies for end-point machines using time of day, location and amount of data downloadable.
- **Unauthorised Privilege Elevation:** Convert from ordinary user to Administrator using software vulnerabilities. Prevent with Intrusion Prevention Systems (IPS) and query-level access control.
- **Service account or shared account abuse:**

    - **Service Accounts**: Batch IDs. Untraceable to a particular user.
    - **Shared Accounts:** Generic IDs with one password. Provide ungoverned access.
- **Platform intrusion attacks:** Protect databases with Intrusion Protection and Intrusion Detection Systems. Any update patches should be installed immediately.
- **SQL Injection vulnerability:** A SQL command is inserted in a web input space**.** Sanitise all inputs before passing to server.
- **Default passwords:** Usually supplied by software vendor. Eliminate them
- **Backup data abuse**: Encrypt all database backups and securely manage decryption keys

1.3.14 Hacking/Hacker

A hacker finds unknown pathways in complex computer systems. Can be good or bad:

- White Hat hacker (Western movies the hero always wore a white hat) finds vulnerabilities which are fixed in the patches.
- Malicious hackers intentionally breach systems to steal information or do damage.

1.3.15 Social Threats to Security / Phishing

Involves direct communication to trick people to provide confidential information - Social engineering. Phishing is the call or message.

1.3.16 Malware

Any malicious software created to damage, change or improperly access a computer or network.

- **Adware**: Spyware that slips into the computer from an internet download. It monitors browsing and buying habits. Not illegal.
- **Spyware**: Any program that slips in without consent
- **Trojan Horse**: A malicious program that enters the system embedded in legitimate software.
- **Virus:** A program that attaches itself to an executable file, and delivers a destructive payload
- **Worm**: A program built to reproduce and spread across a network by itself. Usually harms networks by consuming bandwidth.
- **Malware Sources:**
    - Instant Messaging (IM):
    - Social Networking Sites:
    - Spam

2 Activities

2.1 Identify Data Security Requirements

There is no one prescribed way to implement data security. Organisations should design their own security controls.

- **Business Requirements:** Analyse business rules and processes to identify security touch points.
- **Regulatory requirements:** Create a central inventory of all data regulations and the data subject areas affected by each regulation.

Table 13 Sample Regulation Inventory Table

Regulation	Subject Area Affected	Security Policy Links	Controls Implemented

2.2 Define Data Security Policy

Data security policies describe behaviours determined to be in the best interests of the organisation protecting its data. Must be auditable and audited. Requires collaboration between IT Security administrators, Security Architects, Data Governance committees, Data Stewards, internal and external audit teams and the legal department.

Security Policy contents:

- **Enterprise Security Policy:** Global policies
- **IT Security Policy:** Directory structures standards, password policies and identity management framework
- **Data Security Policy:** Categories for individual applications, database roles, user groups and information sensitivity

2.3 Define Data Security Standards

Standards supplement policies and provide detail on how to meet the intention of the policies.

- **Define Confidentiality Levels:** Confidentiality classification is important metadata, guides how users are granted access privileges
- **Define Regulatory Categories:** Regulated Information. Regulations Imply a goal - compliance
- **Define Security Roles:** Two ways to organise
  - **Role Assessment grid:** starting from the data

Table 14 Role Assignment Grid Example

	Confidentiality Level		
	General Audience	Client Confidential	Restricted Confidential
Not Regulated	Public User Role	Client Manager Role	Restricted Access Role
PII	Marketing Role	Client Marketing Role	HR Role
PCI	Financial Role	Client Financial Role	Restricted Financial Role

  - **Role Assessment Hierarchy:** Starting from the User

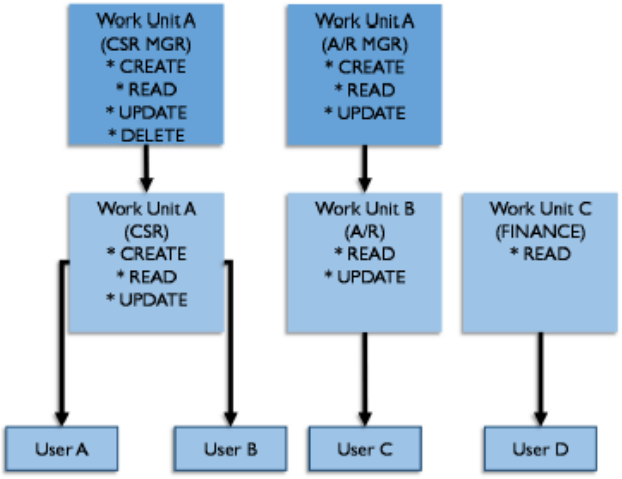


Figure 65 Security Role Hierarchy Example Diagram

- **Assess Current Security Risks:** Identify where sensitive data is stored. Evaluate each system for the following:
  - The sensitivity of data stored of in transit
  - The requirements to protect that data
  - The current security protections in place

  Document the findings as they provide a baseline and may be a requirement of compliance.

- **Implement Controls and Procedures:** Responsibility of security administrators in coordination with data stewards and technical teams. Controls and procedures should at minimum cover:
  - How users gain and lose access to systems
  - How users are assigned to and removed from roles
  - How privilege levels are monitored
  - How requests for access changes are handled and monitored
  - How data is classified according to confidentiality and regulations
  - How data breaches are handled
- **Assign confidentiality levels:**
- **Assign regulatory Categories:**
- **Manage and maintain data security**:
  - Control Data Availability / Data-centric Security
  - Monitor User Authentication and Access Behaviour:
    - Required for compliance audits
    - Lack of automated recording of sensitive and unusual data base transactions represents serious risks
    - Implement a network based audit appliance to mitigate the risks.
- **Manage security policy compliance**
  - Manage regulatory compliance
  - Audit data security and compliance activities

3 Tools

- **Anti-Virus Software / Security Software**
- **HTTPS:** The web address begins https:// a security layer is present

- **Identity Management Technology:**
- **Intrusion Detection (IDS) and Prevention Software (IPS):**
- **Firewalls (Prevention):**
- **Metadata tracking:**
- **Data Masking/Encryption**

4 Techniques

- **CRUD Matrix usage:** Create and use data-to-process and data-to-role relationships
- **Immediate Security Patch Deployment:** No one should be able to delay this update
- **Data Security Attributes in Metadata:** Metadata repository is essential
- **Metrics:** Frame as positive value percentages
    - **Security implementation Metrics:** maintain a reasonable number of actionable metrics in appropriate categories over time to assure compliance
    - **Security Awareness metrics:**
        - Risk assessment findings
        - Risk events and profiles
        - Formal feedback surveys and interviews
        - Incident post-mortems, lessons learned and victim interviews
        - Patching effectiveness audits
    - **Data Protection Metrics:**
        - Criticality ranking
        - Annualised loss expectancy
        - Risk of specific data losses
        - Risk mapping of data to specific business processes
        - Threat assessments
        - Vulnerability assessments
    - **Security Incident Metrics:**
        - Intrusion attempts detected and prevented
        - Return on investment for security costs using savings from prevented intrusions
    - **Confidential Data Proliferation**:
- **Security Needs in Project Requirements:** Identify in the analysis phase
- **Efficient Search for Encrypted Data:**
- **Document Sanitisation:** Clean the Metadata preventing embedded confidential data being shared

5 Implementation Guidelines

- **Readiness Assessment / Risk Assessment:**
    - **Training:** Training on security initiatives at all levels of the organisation
    - **Consistent policies:** Data security policies should align with enterprise policies
    - **Measure the benefits of security:** Link to organisational activities
    - **Set security requirements for vendors:** Include in SLAs and contracts
    - **Build a sense of urgency:** Emphasise legal, regulatory and contractual requirements to build a sense of urgency
    - **Ongoing Communications:**
- **Organisation and Cultural Change:**

- **Visibility into User Data Entitlement:** Requires Metadata of classification and the authorisations themselves
- **Data Security in an Outsourced World:** Anything can be outsourced except liability.
    - Tighter management of control mechanisms
- **Data Security in Cloud Environments:** Data security policies should account for data distributed over these platforms, and should be the same as the rest of the enterprise

6 Data Security Governance – Data Security and Enterprise Architecture

Requires cooperation between IT and business stakeholders, and strong policies and procedures.

Data security Architecture is a component of enterprise architecture that describes how data security is implemented. Architecture influences:

- Tools used to manage data security
- Data encryption standards and mechanisms
- Access guidelines to external vendors and contractors
- Data transmission protocols over the internet
- Documentation requirements
- Remote access standards
- Security breach access reporting

Security architecture is particularly important for integration of data between:

- Internal systems and business units
- an organisation and its external business partners
- An organisation and regulatory agencies

Data Integration and Interoperability

1 Introduction

DII describes processes relating to the movement and consolidation of data within and between data stores, applications and organisations. **Data Integration** consolidates data into consistent forms. **Data Interoperability** is the ability for multiple systems to communicate.

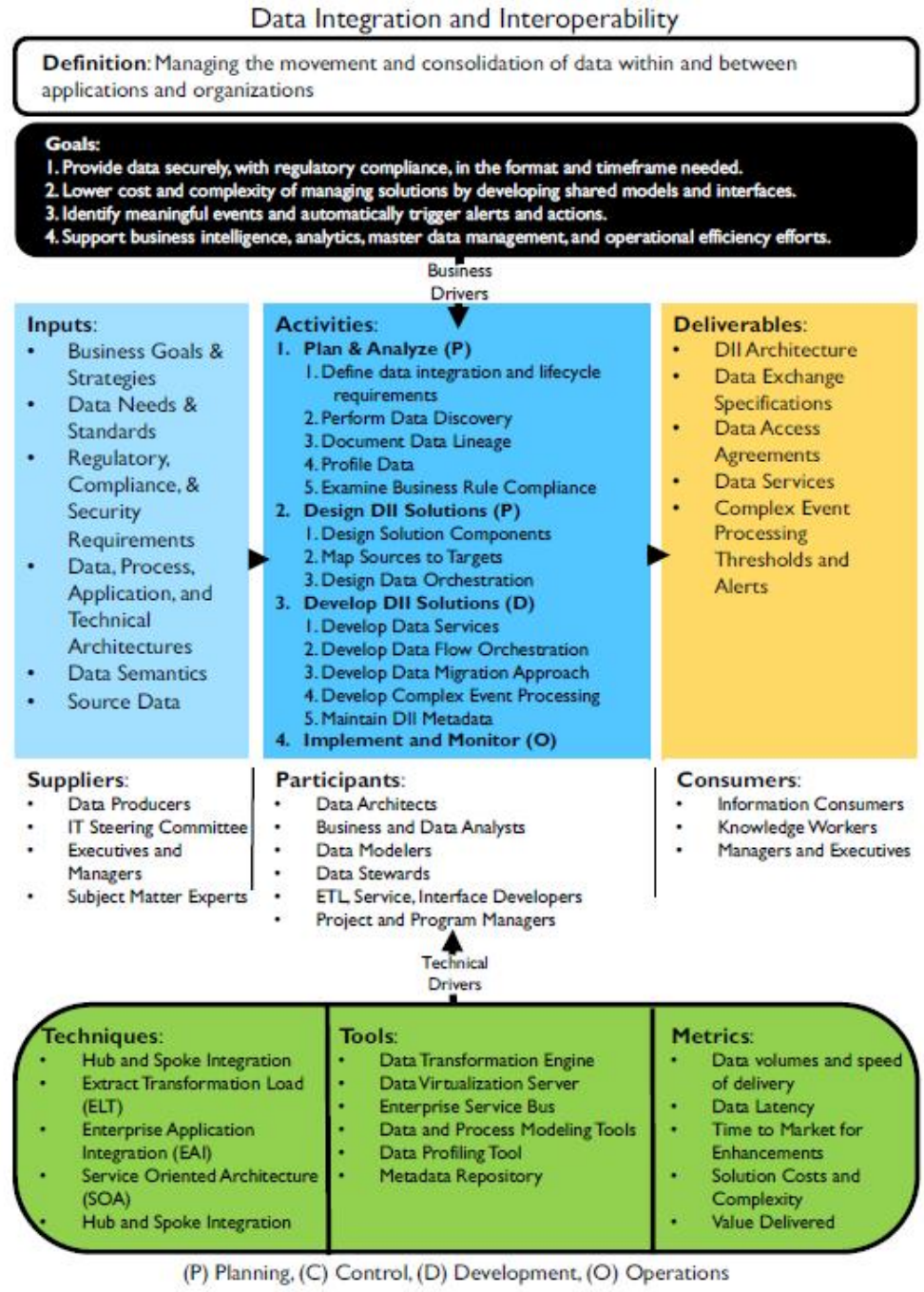
DII provides:

- Data migration and conversion
- Data consolidation into hubs or marts
- Integration of vendor packages into organisation
- Data sharing across applications or organisations
- Distributing data across data stores and data centres
- Archiving data
- Managing data interfaces
- Obtaining and ingesting external data
- Integrating structured and unstructured data
- Providing operational intelligence and management decision support

DII is dependent on these other areas of data management:

- **Data Governance:** The transformation rules and message structures
- **Data Architecture**: Designing solutions
- **Data Security**: ensuring solutions protect the security of data
- **Metadata:**
    - Tracking technical inventory of data (persistent, virtual and in motion)
    - Business meaning of data
    - Business rules for transforming data
    - Operational history and lineage of the data
- **Data Storage and Operations:** Managing the physical instantiation of the solutions
- **Data Modelling and Design:** Designing the physical and virtual data structures, and messages passing information between applications and organisations

DII is critically important **to Data Warehousing and Business Intelligence** as well as **Reference Data and Master Data Management**.



1.1 Business Drivers

- The need to manage data movement efficiently
- Purchased applications come with its own data stores that must integrate with the other data store in the organisation.
- An enterprise view of data integration is more cost effective than point to point solutions
- Data hubs such as data warehouses and Master Data solutions
- Managing the cost of support by using standard tools and reducing the complexity of interface management
- DII supports the organisation's ability to comply with data handling standards and regulations.

1.2 Goals and Principles

Goals:
1. Provide data securely, with regulatory compliance, in the format and timeframe needed.
2. Lower cost and complexity of managing solutions by developing shared models and interfaces.
3. Identify meaningful events and automatically trigger alerts and actions.
4. Support business intelligence, analytics, master data management, and operational efficiency efforts.

When implementing DII follow these principles:

- Design should take an enterprise perspective (for future extensibility), but implement iteratively and incrementally
- Balance local data needs with enterprise data needs, including support and maintenance
- Ensure business accountability for DII design and activity.

1.3 Essential concepts

1.3.1 Extract, Transform and Load (ETL)

The essential steps in moving data around

- **Extract:** Select required data and extract it from its source, and stage it physically or in memory
- **Transform:** Make the data compatible with the structure of the target store. May be done in batch or real-time:
  - **Format changes:** Technical format e.g. EBCDIC to ASCII
  - **Structure changes:** e.g. denormalised to normalised
  - **Semantic conversion:** Conversion of values to maintain consistent semantic representation
  - **De-duping:** If rules require unique values, scan target and remove duplicate rows
  - **Re-ordering:** Change to order of the file data elements to fit a pattern
- **Load:** Physically store the result of the transformation in the target system

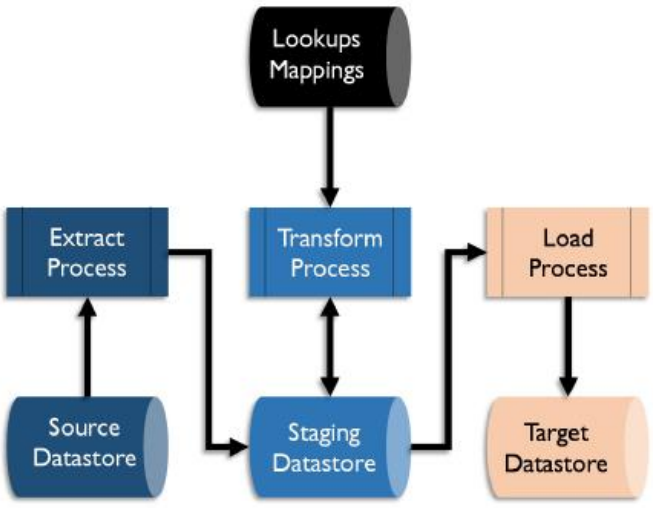


Figure 67 ETL Process Flow

- **ELT (Extract, Load and Transform):** Used if the target system has more transformation capability. also allows data to be instantiated on the target as raw data.

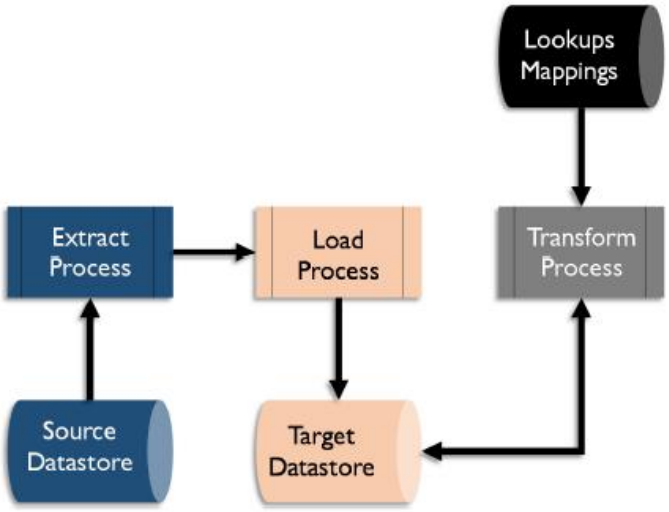


Figure 68 ELT Process Flow

- **Mapping:** A synonym for transformation, the process of developing a lookup matrix from source to target structures, and the result of that process.

1.3.2 Latency

The time difference between when data is generated in the source system and when it is available in the target system. Can be high (batch), low (event driven) to very low (real-time synchronous).

- **Batch:** Data moving in clumps or files periodically. Called a **batch** or **ETL**.
  - **Snapshot:** Full set of data at a point in time
  - **Delta:** Data that has changed values since the last time data was sent.
  - High latency
  - Used for data conversions, migrations, archiving and extracting from and loading data warehouses and marts
- **Change data Capture:** Filtering data to include only the data that has changed in a given timeframe (the delta) and passes it on to data consumers.
- **Near-real-time and Event Driven:** Data is processed in smaller sets spread across the day, or when and event such as an update occurs. Lower latency than batch.
- **Asynchronous:** The source does not wait for the target to acknowledge before continuing processing. The target need not be available.
- **Real-time, Synchronous:** No time delay or other differences between source and target are acceptable. The executing process waits for confirmation before executing its next transaction.
- **Low Latency or Streaming:** Extra hardware costs. Need extremely fast transfer of large amounts of data over large distances.

1.3.3 Replication

Applications maintain exact copies of data sets on multiple physical locations. Better response times for international users. Use DBMS replication utilities. Not recommended if changes to the data occur at more than one site.

1.3.4 Archiving

ETL functions can transport and possibly transform infrequently used data to a cheaper storage solution.

1.3.5 Enterprise Message Format/ Canonical Model

A canonical model is the common model used by an organisation to standardise the format in which data will be shared. Transformations need only be done to and from the canonical model.

- Hub-and-spoke
- All systems interact with central information hub
- Data is transformed based on the enterprise message format of the organisation
- Reduces transformations as each system only needs to transform data to and from the central canonical model
- Reduces complexity of DII in the enterprise
- Lowers cost of support
- Complex to develop
- Justified in managing more than 3 systems and critical for more than 100

1.3.6 Interaction Models

Describe ways to make connections between systems:

- **Point-to-point:**
  - Systems pass data directly to each other.
  - Suitable in small systems.
  - Can impact processing
  - Many interfaces to manage
  - Multiple interfaces can lead to inconsistent data
- **Hub-and-spoke:**
  - Consolidates shared data in a central hub.
  - Examples are Data Warehouses, Data Marts. Operational Data Stores and Master Data Management Hubs.
  - Easy to add more systems.
  - Enterprise Service Buses (ESB) – near real-time sharing of data where the hub is a virtual canonical model.
- **Publish-subscribe:** Systems push out (publish) data and other systems pull data in (subscribe).

1.3.7 DII Architecture Concepts

- **Application coupling:**
  - **Tight coupling**: Synchronous interface, one waits for the other to respond
  - **Loose coupling**: Preferred interface design as data is passed without waiting for a response and without causing both systems to be unavailable if one is unavailable. e.g. Service Oriented Architecture using an Enterprise Service Bus

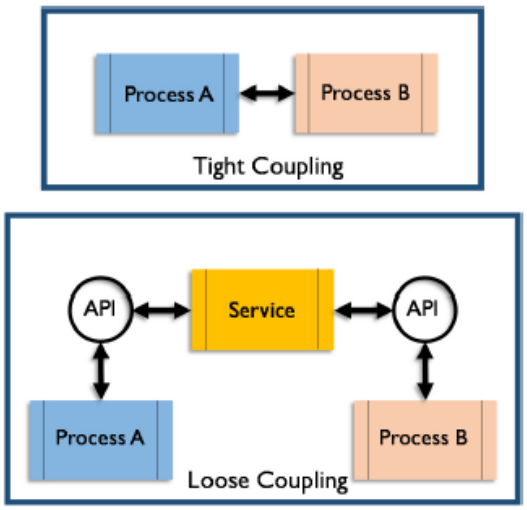


Figure 69 Application Coupling

- **Orchestration and Process Controls:**
  - **Orchestration:** How multiple processes are organised and executed in a system. All systems handling messages must be able to manage the order of those processes.
  - **Process Controls:** The components that ensure shipment, delivery, extraction and loading of data is complete. Include:
    - Database activity logs
    - Batch job logs
    - Alerts
    - Exception logs
    - Job dependence charts with remediation options, standard responses
    - Job clock information – length of jobs and computing window time.
- **Enterprise Application Integration (EAI):** Software modules only interact with each other through APIs.
- **Enterprise Service Bus (ESB):** An intermediary system, passing messages between other systems.

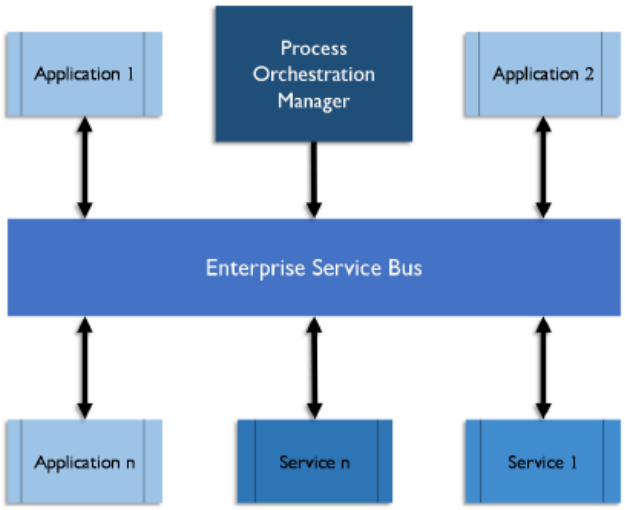


Figure 70 Enterprise Service Bus

- **Service-Oriented Architecture (SOA):** The functionality of providing data or updating data is provided through well-defined service calls between applications enabling application independence. Usually implemented through APIs.
- **Complex Event Processing (CEP):** Tracks and analyses data about events from multiple sources to predict meaningful events (threats, opportunities) to predict behaviour or automatically trigger a response.
- **Data Federation and Virtualisation:** Provides access to a combination of individual data stores. Virtualisation enables distributed databases to be accessed as a single database.
- **Data-as-a-Service (DaaS):** Data licensed from a vendor and provided on demand.
- **Cloud-based Integration:** Also called Integration Platform-as-a-service (IPaaS) and are usually run as DaaS applications at the data centres of vendors

1.3.8 Data Exchange Standards

Data Exchange Standards are formal rules for the structure of data elements. ISO or any common model used by an organisation or data exchange group. It simplifies data interoperability, lowers support costs and increases understanding of the data.

2 Data Integration Activities

Data Integration activities follow a development lifecycle (plan, design, development, testing and implementation)

2.1 Plan and Analyse

- Define data integration and lifecycle requirements:
  - Data and technology required to meet business objectives
  - Laws and restrictions on data contents
  - Defined by Business Analysts, data stewards and various architects
  - Creates and uncovers valuable Metadata – manage throughout lifecycle
- Perform Data Discovery:
  - Identify potential sources of data for the integration effort
  - High level assessment of data quality
  - Maintain the inventory of organisational data in a Metadata repository
  - Plan for acquiring and integrating external data
- Document Data Lineage:
  - How data under analysis is acquired or created by the organisation, where it moves, how it is changed, how it is used by the organisation for analytics, decision making or triggering.
- Profile Data:
  - Data profiles reveals actual data structure and contents
  - Assess the quality of the data
  - Can reveal differences from what is assumed leading to early intervention
  - Basic profiling involves analysis of:
    - Data format (data structures and inferred from actual data)
    - Data population (NULLS, blanks and defaults)
    - How data values correspond to a set of valid values
    - Patterns and relationships internal to the data set
    - Relationships to other data sets
  - The requirement to profile data must be balanced with an organisation's privacy and security rules (see Chapter 13)

- Collect Business Rules: rules harvesting or mining
    - A Business Rule is a statement that defines or constrains and aspect of business processing.
    - Four categories:
        - Definition of business terms
        - facts relating terms to each other
        - Constraints or action assertions
        - Derivations
    - Use business rules to support DII to:
        - Assess data in potential source and target data sets
        - Direct the flow of data in the organisation
        - Direct when to automatically trigger events and alerts

2.2 Design Data Integration Solutions

2.2.1 Design Data Integration Architecture

DI solutions should be specified at both enterprise and individual solution level. Enterprise standards save time as planning has been done and group licences and resource sharing and reusing DII components saves money.

- **Solution architecture indicates**:
    - Techniques and technologies to be used
    - Inventory of involved data structures
    - Orchestration and frequency of data flow
    - Regulatory and security concerns and remediation
    - Operating concerns around backup and recovery, archiving and retention
- **Select Interaction model:** hub-and-spoke, point-to-point or publish-subscribe
- **Design data services or exchange patterns:** Start with industry standards or existing patterns.

2.2.2 Model Data Hubs, Interfaces, Messages and Data Services

Model persistent and transient datatypes.

2.2.3 Map Data Sources to Targets

Specify the rules for transforming data from one location and format to another. For each attribute mapped specify:

- Technical format of source and target
- Transformations required for intermediate staging points
- How each attribute in the final or intermediate data store will be populated
- Whether data values need to be transformed
- calculations required

2.2.4 Design Data Orchestration

Pattern of data flows from start to finish, including intermediate steps.

- **Batch**: Frequency of data movement and transformation is usually coded into a scheduler
- **Real-time**: Usually triggered by an event such as an update.

2.3 Develop Data Integration Solutions

- **Develop Data Services:** Can be tools or vendor suites

- **Develop data flows:**
  - ETL flows developed in a tool such as a scheduler for batch which manages the order, frequency and dependency of executing the data pieces
  - Real-time: Monitor for events that trigger services to acquire, transform or publish data
- **Develop data migration approach:** Moving data needs proper analysis and testing
- **Develop a publication approach:** Push changed data using common message definitions (canonical model) and notify the recipients
- **Develop complex event processing flows:**
  - Preparing historical data needed for the predictive model
  - Pre-populate predictive model and identify meaningful events
  - Executing triggered action in response to a prediction
- **Maintain DII Metadata:**
  - Manage Metadata uncovered during the development of DII solutions
  - Document all data structures involved
  - ETL vendors have Metadata repositories
  - Service-Oriented Architecture registry

2.4 Implement and Monitor

- Activate the data services that have been developed and tested.
- Real-time data processing needs real-time monitoring for issues, human or automated.
- Monitor and service at the level of the most demanding target application or consumer.

3 Tools

- **Data Transformation Engine/ETL Tool:** Primary tool that supports operation and design.
- **Data Virtualisation Server:** Perform data extract, transform and integrate virtually. A data warehouse is often input to a data virtualisation server
- **Enterprise Service Bus:** Middleware to support near real-time messaging between heterogeneous sources in the same enterprise
- **Business Rules Engine:** Allows non-technical users to manage business rules implemented by software
- **Data and Process Modelling Tools:** Used to design target and intermediate data structures
- **Data Profiling tool:** Use a tool to profile large amounts of data
- **Metadata Repository:** Tools have Metadata repositories which store the Rules for transformation, lineage and processing, as well as the instructions for scheduled processes and triggers.

4 Techniques

Described in Essential Concepts

5 Implementation Guidelines

5.1 Readiness Assessment / Risk Assessment

As all organisations already have DII in place assess for readiness/risk around tool implementation or interoperability. An enterprise data integration solution supports the movement of data between many applications and organisations.

Working DII solutions shouldn't be replaced. Focus on where none exists. Additional use of data integration adds to the investment in a data warehouse to Master Data Management hub.

It is necessary to sponsor the implementation of an enterprise data integration program by a high level of authority over solution design and technology purchase, to prevent local data integration solutions from developing. It may be perceived the cost of these is less than the enterprise wide solution.

Don't' become too focussed on the tool and lose focus on the business needs.

5.2 Organisation and Cultural change

- Local teams understand data in their applications.
- Central teams know tools and techniques.
- Enterprise solutions should be overseen by a Centre of Excellence
- Although technical, data integration solutions must be based on deep business knowledge to successfully deliver value.
- Modelling and data analysis should be done by business resources
- Canonical message model (consistent standard for how data is shared in the organisation) development requires technical and business resources
- Business SMEs review transformation mapping design and changes

6 DII Governance

Business is responsible for:

- Decisions about the design of data messages, data models and data transformation rules
- Defining the rules for loading and transforming data
- approve changes to these rules (which are captured as Metadata for cross enterprise analysis)
- Identifying and verifying predictive models and defining the actions they trigger

Governance controls to support trust that the DII will perform as promised:

- Determine what events trigger governance reviews (exceptions or critical events)
- Map each trigger to reviews that engage with governance bodies
- Event triggers may be part of SDLC at Stage Gates or part of User Stories

Controls come from governance-driven management routines such as mandated review of models, Metadata audits, gating of deliverables or required approval of changes to the transformation rules.

Include real-time operational data integration solutions in SLAs and Business Continuity/Disaster Recovery plans on the same tier as the most critical system to which they provide data.

Policies need to be established to ensure the organisation benefits from an enterprise approach to DII.

6.1 Data Sharing agreements

A data sharing agreement or memorandum of understanding (MOU) must be put in place before developing interfaces, and be approved by business data stewards, which stipulates:

- Responsibilities
- Acceptable use of data to be exchanged

- Anticipated use and access to the data
- Restriction on use
- Expected service levels
- Required system up times and response times

6.2 DII and Data Lineage

Governance to ensure that knowledge of data origins and movement is documented. Data lineage must be managed as it is critical Metadata.

Compliance standards require an organisation be able to describe where its data originated, and how it has changed as it moves through systems.

Impact analysis when making changes to data structures, data flows or data processing requires forward and backward data lineage.

6.3 Data Integration metrics

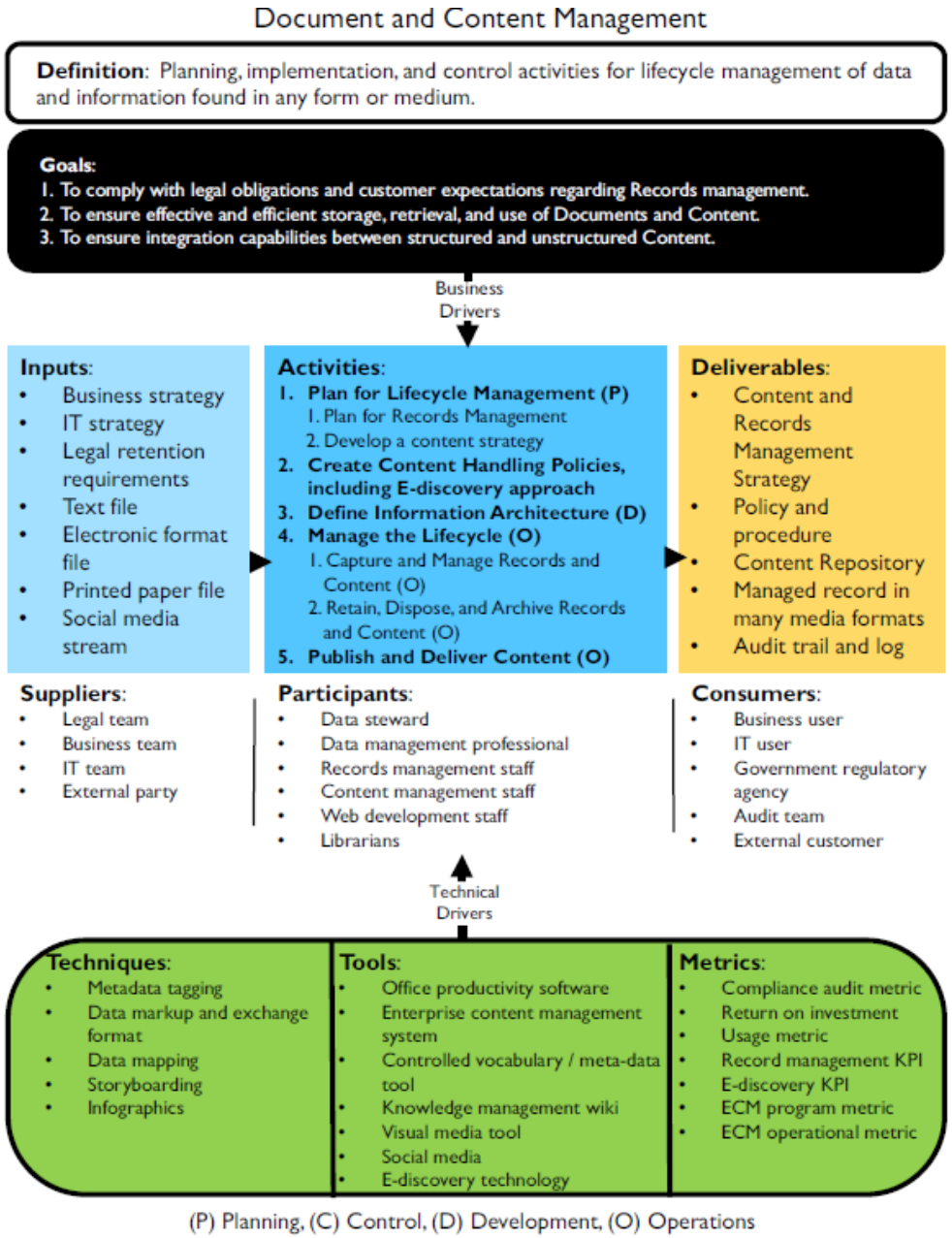
To measure scale and benefits of DII solutions:

- Data Availability (of data requested)
- Data volumes and speed
    - Volumes of data transported and transformed
    - Volumes of data analysed
    - Speed of transmission
    - Latency between update and availability
    - Latency between event and triggered action
    - Time to availability of new data sources
- Solution costs and complexity
    - Cost of developing and managing solutions
    - Ease of acquiring new data
    - Complexity of solutions and operations
    - Number of systems using data integration solutions

Document and Content Management

1 Introduction

Document and Content Management entails controlling the capture, storage, access and use of data and information stored outside relational databases. Records include both paper documents and ESI (Electronically Stored Information).



1.1 Business Drivers

- **Regulatory compliance:** Respond efficiently and consistently
- **The ability to respond to litigation**
- **e-discovery:** The process of finding electronic records that might serve as evidence in litigation. Includes email, chats, websites, raw application data and Metadata. Huge volume of ESI.
- **Business continuity requirements**

- **Efficiency**: Use technology to streamline processes, manage workflow, eliminate repetitive manual tasks and enable collaboration
- **Customer satisfaction**: Well organised and searchable websites

1.2 Goals and Principles

Goals:
1. To comply with legal obligations and customer expectations regarding Records management.
2. To ensure effective and efficient storage, retrieval, and use of Documents and Content.
3. To ensure integration capabilities between structured and unstructured Content.

Goals of implementing best practices around Document and Content Management:

- Ensuring effective and efficient retrieval and use of data in unstructured formats
- Ensuring integration capability between structured and unstructured data
- Complying with legal obligations and customer expectations

Management of Document and Content follow these guiding principles:

- Everyone in the organisation must create, use, retrieve and dispose of records in accordance with the established policies and procedures. Train everyone.
- Experts in the handling of records and content should be fully engaged in policy and planning

In 2009 ARMA International published GARP (Generally Acceptable Recordkeeping Principles® how business records must be maintained: Details on ARMA website

- **Principle of Accountability:**
  - Senior executive appropriates individuals, adopts policies and processes and ensures auditability
- **Principle of Integrity:**
  - Guarantee of authenticity and reliability
- **Principle of Protection:**
  - Personal information is protected
- **Principle of Compliance:**
  - Program should comply with laws, binding authorities and the policies of the organisation
- **Principle of Availability:**
  - Information is managed in a manner that ensures timely, efficient and accurate retrieval of information
- **Principle of Retention:**
  - Retain information for an appropriate time taking legal and regulatory requirements into consideration
- **Principle of Disposition:**
  - Provide secure and appropriate disposition of information in accordance with operational, legal, regulatory and fiscal requirements
- **Principle of Transparency:**
  - Policies, processes, activities, governance program documented and understood by all staff and appropriate interested parties

1.3 Essential Concepts

1.3.1 Content

Document is to content what a bucket is to water: a container. **Content** refers to the data and information inside the file, document or website. Content has a lifecycle and can become an official **record**. Official records are treated differently.

- **Content Management:** the processes, techniques and technologies for organising, categorising and structuring information resources so that they can be stored, published and reused.
- **Content Metadata:** Essential for managing unstructured data and "Big Data":
  - **Format:** often dictates the method to access
  - **Search-ability:** search tools already exist
  - **Self-documentation:** Metadata self-documenting as in file systems
  - **Existing patterns:** can they be adopted
  - **Content subjects:** The things people are likely to be looking for
  - **Requirements:** Is a content tagging tool necessary?
- **Content Modelling:** The process of converting logical content concepts into content types, attributes and data types with relationships using Metadata Management and Data Modelling techniques.
  - Information product level e.g. website
  - Component level
- **Content Delivery Methods:** Convert to XML to enable reuse over many channels. Types of delivery systems:
  - **Push:** Users choose type of content to be delivered on a schedule.
  - **Pull:** Users pull content through the Internet
  - **Interactive:** e.g. electronic point of sale. High volumes of real-time data

1.3.2 Controlled Vocabularies

A defined list of explicitly allowed terms to index, categorise, tag, sort and retrieve content through browsing and sorting. Controlled vocabularies should be aligned with the entity names and definitions in the enterprise conceptual model. May be considered Reference Data and Metadata.

- **Vocabulary Management:** The function of defining, sourcing, importing and maintaining any given vocabulary.
  - ANSI/NISO Z39.19-2005 definition: Way to improve the effectiveness of information storage and retrieval systems, web navigation systems and other environments that seek to both identify and locate desired content via a description using language.
  - Purpose of vocabulary control is consistency in the description of content objects to facilitate retrieval
- **Vocabulary Views and Micro-Controlled Vocabulary:**
  - a vocabulary view is a subset of a vocabulary relevant to a group of users
  - A micro-controlled vocabulary is a vocabulary view containing highly specialised terms
- **Term and Pick Lists:** lists in applications that do not contain relationships
- **Term Management:** One or more words designating a concept (ANSI/NISO Z39.19-2005). Terms should be specified and managed through governance processes. Three types of term relationships:
  - **Equivalent term relationship:** Leads to one or more terms to use

    - **Hierarchical relationship:** depicts broader (general) to narrower (specific) or whole-part relationships.
    - **Related term relationship:** a term that is associatively but not hierarchically linked to another term in a controlled vocabulary.
- **Synonym rings and Authority lists:**
    - **Synonym ring:** a set of terms with roughly equivalent meaning. A search on one reveals content on the others
    - **Authority list:** a controlled vocabulary of descriptive terms designed to facilitate retrieval of information within a specific domain
- **Taxonomies:** a naming structure containing a controlled vocabulary used for outlining topics and enabling navigation and search systems. Can be a data model.
    - **Flat taxonomy:** no relationships among the set of controlled categories
    - **Hierarchical taxonomy:** Tree structure where nodes are related by a rule. Going up expands the category and going down refines.
    - **Polyhierarchy:** Tree structure with more than one rule. Multiple parents. Paths are complex
    - **Facet taxonomy:** Star structure. each node is associated to a central node e.g. Metadata where each attribute is a facet of a content object
    - **Network taxonomy:** Both hierarchical and facet structures. e.g. a thesaurus
- **Classification Schemes and Tagging:** Codes that represent controlled vocabulary. Folksonomy: names are obtained though social tagging
- **Thesauri:** A Thesaurus is a controlled vocabulary used for content retrieval. Provides information about each term and its relationship to other terms.
- **Ontology:**
    - a type of taxonomy that represents a set of concepts and their relationships within a domain.
    - Semantic Web.
    - Developed using ontology languages such as RDFS (Resource Description Framework Schema)
    - Describe classes (concepts), Individuals (instances) attributes, relations and events.
    - two **differences** between taxonomy and ontology:
        - **Taxonomy:** data content classifications for a given concept area.
        - **Ontology:** Content concepts are mixed and identified through Metadata.
        - **Taxonomy:** closed-world assumption, only what is known is defined.
        - **Ontology:** open-world assumption, possible relationships are inferred

1.3.3 Documents and Records

Documents are electronic or paper objects which communicate information and knowledge.

Records are a subset of documents which provide evidence of actions taken and decisions made in accordance with procedures. Can be created manually or automatically by equipment.

- **Document Management:** Processes, techniques and procedures for controlling and organising documents and records throughout their lifecycle.
    - **Inventory:** Identification of existing and newly created documents and records
    - **Policy:** Creation, approval and enforcement of documents and records policies
    - **Classification**
    - **Storage:** Short and long term physical storage

- **Retrieval and circulation:** allow access on accordance with policies, security and legal standards
- **Preservation and disposal:** Archiving and destroying documents and records according to organisational needs, statutes and regulations.

Figure 72 Document Hierarchy based on ISO 9001-4.2

- **Records Management:**
  - Managing records through the full lifecycle:
    - Record creation or receipt
    - processing
    - distribution
    - organisation
    - retrieval
    - disposition
  - Records can be physical, electronic, web content, documents on all kinds of media, data in any kind of database.
  - A **Vital Record**: required by an organisation to resume after a disaster
  - Trustworthy records have signatures for regulatory compliance and integrity.
  - Characteristics of well-prepared records:
    - **Content**: Accurate, complete, truthful
    - **Context:** Collect Metadata
    - **Timeliness:** created promptly after the event
    - **Permanency:** Records may never be changed
    - **Structure:** Content should be clear and legible

- **Digital Asset Management:** Document management for rich media documents (videos, logos, photographs etc.)

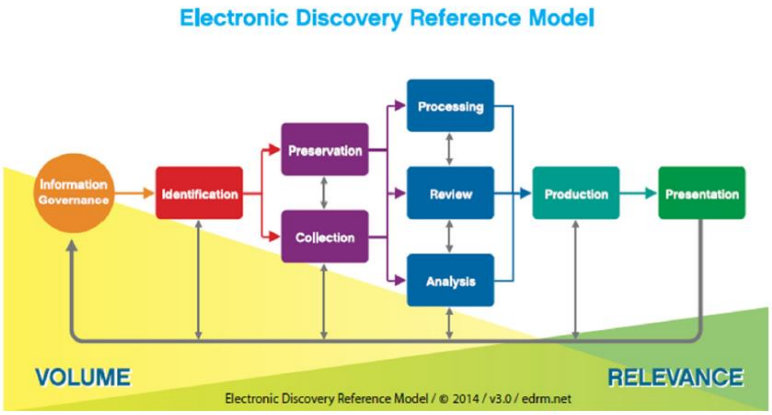
1.3.4 Data Map

A data map is an inventory of all ESI (Electronically Stored Information) data sources, applications and IT environments.

1.3.5 E-*D*iscovery

Discovery is the pre-trial phase of a lawsuit where both parties request information from each other to find the facts of the case. Rules were amended in 2006 to include ESI (Electronically Stored Information)

- Electronic documents have Metadata which plays a part in evidence.
- Legal requirements:
  - e-discovery
  - data and records retention practices
  - Legal hold notification (LHN) process
  - Legally defensible disposition practices



EDRM is a standards and guidelines organisation for e-discovery. The EDRM model assumes information governance is in place. Eight e-discovery phases:

- **Identification:**
  - **Early Case Assessment:** Legal case assessed for descriptive information or Metadata
  - **Early Data Assessment:** Type and location of data relevant to the case is assessed.
    - Identify policies for retention or destruction of relevant data so ESI can be preserved
    - Interviews with involved personnel to get pertinent information
    - Involved personnel need to understand the case background, legal hold and their roles
- **Preservation:** Identified potentially relevant data is put on hold to avoid destruction
- **Collection:** Acquisition and transfer information to legal counsel in a legally defensible manner
- **Processing:** Data de-duplicated, searched and analysed to determine which moves to review
- **Review:** Documents identified to be presented on request
- **Analysis:** Understand the content to formulate a legal strategy
- **Production:** Information turned over to opposing counsel based on agreed to specifications
- **Presentation:** Displaying the ESI at depositions, hearings and trials

1.3.6 Information Architecture

Information architecture is the process of creating structure for a body of information or content. Identifies links and relationships between documents and content. Components:

- Controlled vocabularies
- Taxonomies and ontologies
- Navigation maps
- Metadata maps
- Search functionality specifications
- Use cases
- User flows

The information architecture and the content strategy together describe WHAT content will be managed in the system. Design phases describe HOW the strategy will be implemented.

1.3.7 Search Engine

A search engine is software that searches for information based on terms and retrieves websites that have those terms within their content. e.g. Google.

1.3.8 Semantic Model

Semantic modelling is a type of modelling that describes a network of concepts and their relationships, and allows users to ask questions of the system in a non-technical way.

1.3.9 Semantic Search

A semantic search engine can use artificial intelligence to identify query matches based on keywords and their context. BI and analytics tool users often have semantics search requirements.

1.3.10 Unstructured Data (rather call it non-tabular or semi-structured data)

About 80% of all stored data is outside relational databases. Include word processing documents, emails, social media, chats, flat files, spreadsheets, XML files, transactional messages, reports, graphics, digital images, microfiche and video and audio recordings.

There is no data model to aid understanding. No tags. Not organised in rows and columns.

The fundamental principles of Data Management apply. Unstructured data requires data governance, architecture, security, Metadata and data quality.

1.3.11 Workflow

Manage content management through a workflow that ensures content is created on schedule and receives proper approvals. Should be automated through use of a content management system (CMS) to provide version control.

2 Activities

2.1 Plan for Lifecycle Management

Planning for document management:

- For the document's lifecycle
- Developing classification/indexing systems and taxonomies
- Creating policy specifically for records

Identify an organisational unit responsible for managing documents and records.

2.1.1 Plan for records Management

Define what constitutes a record. The defining team for a functional area should include SMEs and records system people. Decisions:

- where to store and archive
- account for paper records as well as electronic unstructured and structured data

2.1.2 Develop a content Strategy

Defines how content will be prioritised, organised and accessed. Identify Content drivers (the reasons content is needed).

- inventory of current state
- gap assessment

2.1.3 Create Content Handling Policies

Policies help employees understand and comply with the requirements for document and records management.

Most document management programs have policies relating to:

- Scope and compliance with audits
- Identification and protection of vital records
- Purpose and schedule for retaining records
- How to respond to information hold orders (in the event of a lawsuit holding records beyond expiry of retention)
- Requirements for onsite and offsite storage of records
- Use and maintenance of hard drive and shared network drives
- email management, addressed from content management perspective
- Proper destruction methods
- **Social Media policies:** Does a social media post from an employee conducting business become a record?
- **Device Access Policies:** Content and records management functions need to work with user driven IT (BYOD bring-your-own-devices)
- **Handling Sensitive data:** Data Security and Data Governance establish confidentiality schemes and identify confidential and restricted assets. People who assemble content must apply these classifications.
- **Responding to litigation:** Prepare for the possibility of litigation through proactive e-discovery. Create an inventory of data sources and their associated risks.

2.1.4 Define Content Information Architecture

Users need to submit search to a system which contains unstructured data in a form understandable to that system. The inventory of structured and unstructured documents needs to be described or indexed for quick retrieval.

2.2 Manage the lifecycle

2.2.1 Capture records and content

Capturing data:

- Electronic content is in the form to be stored in electronic repositories
- Paper: scan and upload using an electronic signature

- Tag/index with the appropriate Metadata necessary for retrieval:
  - Document identifier
  - Date and time of capture
  - Title and author(s)

2.2.2 Manage versioning and control

ANSI Standard 859 has three levels based on criticality of the data:

- **Formal control:** Formal change control processes through change control authority
- **Revision control:** Notify stakeholders and increment versions when a change is required.
- **Custody control:** Least formal, requires safe storage and means of retrieval.

Table 15 Levels of Control for Documents per ANSI-859

Data Asset	Formal	Revision	Custody
Action item lists		x	
Agendas			x
Audit findings		x	x
Budgets	x		
DD 250s			x
Final Proposal			x
Financial data and reports	x	x	x
Human Resources data		x	
Meeting minutes			x
Meeting notices and attendance lists		x	x
Project plans (including data management and configuration management plans)	x		
Proposal (in process)		x	
Schedules	x		
Statements of Work	x		
Trade studies		x	
Training material	x	x	
Working papers			x

ANSI 859 recommends taking the following into account when determining control level for an asset:

- Cost of providing and updating the asset
- Project impact of changes in terms of cost
- Other consequences of change to the enterprise or project
- Need to reuse or earlier versions of the asset
- Maintenance of a history of change

2.2.3 Backup and recovery

Include the documents/record management system in the organisation's **corporate backup and recovery plan** and **Business Continuity Plan**. A **vital records program** allows the business to conduct business during a disaster and resume normal business afterwards. Vital records must be identified and plans for their protection and recovery must be developed and maintained.

2.2.4 Manage retention and disposal

A retentions and disposition policy defines (in accordance with legal and regulatory requirements):

- The timeframes during which documents for operational, legal, fiscal or historical value must be maintained

- When inactive documents can be transferred to a secondary storage facility
- Processes for compliance
- Methods and schedules for disposition of documents.

Records managers and information asset owners ensure privacy and data protections requirements and prevent identity theft.

Software considerations: may require a certain version to read. An upgrade can make documents unreadable or inaccessible.

Non-value-added information must be identified and removed as it wastes resources and can be discoverable in the event of litigation.

2.2.5 Audit Documents/Records

Table 16 Sample Audit Measures

Document / Records Management Component	Sample Audit Measure
Inventory	Each location in the inventory is uniquely identified.
Storage	Storage areas for physical documents / records have adequate space to accommodate growth.
Reliability and Accuracy	Spot checks are executed to confirm that the documents / records are an adequate reflection of what has been created or received.
Classification and Indexing Schemes	Metadata and document file plans are well described.
Access and Retrieval	End users find and retrieve critical information easily.
Retention Processes	The retention schedule is structured in a logical way either by department, functional or major organizational functions.
Disposition Methods	Documents / records are disposed of as recommended.
Security and Confidentiality	Breaches of document / record confidentiality and loss of documents / records are recorded as security incidents and managed appropriately.
Organizational understanding of documents / records management	Appropriate training is provided to stakeholders and staff as to the roles and responsibilities related to document / records management.

Steps of an audit:

- Define organisational drivers and stakeholders – the why
- Gather data on the process (the how), determine what to measure and what tools to use.
- Report the outcomes
- Develop an action plan of next steps and timeframes

2.3 Publish and deliver Content

- **Provide access, search and retrieval** once it is described by Metadata/key word tagging and classified within the appropriate information content architecture
- **Deliver through acceptable channels**: Users want to consume data on the device of their choice meaning the format may change. It may be difficult to bring it back to the original format e.g. HTML to originally structured.

3 Tools

3.1 Enterprise Content Management Systems

An ECM consists of a platform of core components or applications which can be in-house or in the cloud.

Reports can be delivered through common tools or a document management system interface allowing search, view, download and print. Report management is

the ability to add, change or delete reports organised in folders. Reports retention can be set for automatic purge or archive.

3.1.1 Document Management

A **document management system** is an application used to track and store electronic documents and electronic images of paper documents. Provides storage, versioning, security, Metadata management, content indexing and retrieval.

Once a document is created within the document management system it is indexed with keywords so that it can be found. Metadata (Extracted automatically or added manually) is stored for each document. Bibliographic records of documents are descriptive structured data in MARC (Machine-Readable Cataloging) standard format.

A document repository enables check-in and check-out features, versioning, collaboration, comparison, archiving, status state, migration from one storage media to another and disposition.

May have a module that supports different types of workflows: Manual, rules-based or dynamic rules based on content.

Document management systems have a rights module to allocate permissions.

Specialised document management:

- **Digital asset management:** Document management systems may include management of audio, video and digital photographs
- **Image processing:** Captures, transforms and manages images of paper and electronic documents. Uses:
    - **Scanning:** digitising
    - **Optical Character Recognition (OCR):** Mechanical or electronic conversion of digitised print or handwriting to form recognised by software
    - **Intelligence character recognition (ICR):** more intelligent, convert cursive handwriting
    - **Form processing:** capture of printed forms. System recognises the format
    - **Other digitised images:** binary image file formats:
        - **vector**: mathematical formulae to create graphics. Can be resized without compromising resolution. File formats: .eps, .ai and .pdf
        - **raster**: bitmap of coloured pixels. .jpeg, .gif, .png, .tiff
- **Records management system:**
    - Automation of retention and disposition
    - e-discovery support
    - long term archiving to comply with legal and regulatory requirements
    - Vital records program for critical records

3.1.2 Content Management System

A **content management system** is used to collect, index and retrieve content, storing it either as components or whole documents while maintaining links between the components. It maintains content throughout the lifecycle

3.1.3 Content and Document workflow

Workflow tools support business processes, route content and documents, assign work tasks, track status and create audit trails.

3.2 Collaboration tools

For teams

3.3 Controlled vocabulary and Metadata tools

Office productivity software, Metadata repositories, BI tools and document and content management systems

3.4 Standard Markup and exchange formats

To allow computer systems to process unstructured data it must be reformatted.

- **XML:** Extensible Markup Language
  - Uses Metadata to describe content, structure and business rules of a document.
  - Translates structure of the document for data exchange
  - XML tags data elements to identify the meaning of data
  - Older markup methods are HTML and SGML
  - Need for XML-capable content management:
    - Integrating structured and unstructured data in a relational database
    - XML can build enterprise or corporate portals
    - XMP provides labelling and identification of unstructured data
- **JSON:** JavaScript Object Notation
  - Text format is language independent and easy to parse
  - two structures:
    - Objects: a collection of unordered name/value pairs
    - An array: an ordered list of values
  - Preferred format for web and NoSQL databases
- **RDF and related W3C specifications:**
  - **Resource Description Framework** is the standard for web data exchange
  - Stored in a **triplestore:** Subject (Resource) – predicate (property name) – Object (property value).
  - URI (Uniform Resource Identifier) describes subject-predicate-object.
  - URL (Uniform Resource Locator) is a form of URI
  - Semantic Web needs access to both data and relationships between data sets (Linked Data). This is provided by URIs
  - RDF uses XML as its encoding syntax
  - SKOS (Simple Knowledge Organisation System) is an RDF vocabulary
  - OWL (W3C Web Ontology Language) is a vocabulary extension of RDF used for publishing and sharing OWL documents on the web.
  - Big Data: If data is accessible using the RDF Triples model, SPARQL query language can be used to find patterns without predefining a schema.
- **Schema.org:** Labelling content with semantic markup

3.5 E-Discovery technology

Large volumes of data. Technology Assisted Review (TAR) is a process where a team can review selected documents and mark them relevant or not. This is input to a predictive coding engine that reviews and sorts the remaining documents according to relevancy.

4 Techniques

- Litigation response playbook:

- Litigation Response data map

5 Implementation Guidelines

5.1 Readiness Assessment / Risk Assessment

Identify areas where content management improvement is needed. A Data Management Maturity Assessment Model can help.

Specific ECM critical success factors:

- Content audit and classification for existing content
- Appropriate information architecture
- Support of the content lifecycle
- Definitions of appropriate Metadata tags
- The ability to customise functions of the ECM solution

5.1.1 Records Management Maturity

ARMA Generally Accepted Recordkeeping Principles guides policy and practice for Records Management. ARMA Information Governance Maturity Model assesses the organisation's recordkeeping program:

- **Level 1 Sub-Standard:** Information governance and recordkeeping concerns not addressed or just minimally
- **Level 2 In Development:** Developing recognition information governance and recordkeeping can have an impact.
- **Level 3 Essential:** Minimum requirements to meet legal and regulatory requirements
- **Level 4 Proactive:** A proactive information governance
- **Level 5 Transformational**: Information governance is integrated into the corporate infrastructure and business processes

Analyse gaps and risks and their impact on the organisation. If an organisation does not inventory records it is already at risk.

5.1.2 E-Discovery Assessment

A readiness assessment examines and identifies improvements to the litigation process. A mature process will be documented, defensible and auditable and will specify:

- Clear roles and responsibilities
- preservation protocols
- data collection methodologies
- disclosure processes

5.2 Organisation and Cultural Change

ECM can lead to more tasks, such as scanning and defining Metadata.

Management of content and records may be siloed. Training may be necessary to enforce policies across the enterprise.

Content and records management should be elevated to a high priority activity organisationally, especially in highly regulated industries where the RIM function needs to be closely aligned with Corporate legal and e-discovery functions.

6 Document and Content Governance

6.1 Information Governance Frameworks

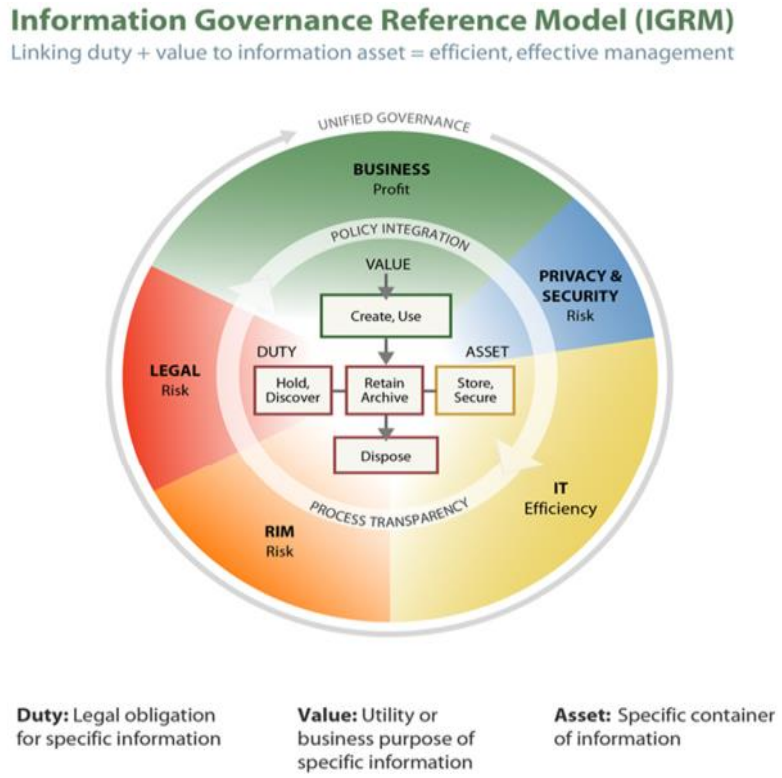
Documents, records and other unstructured content poses risk to an organisation. Drivers to managing the risk include:

- Legal and regulatory compliance
- Defensible disposition on records
- pro-active preparation for e-discovery
- Security of sensitive information
- Management of risks associated with email and Big Data

Principles of successful Information Governance programs:

- ARMA GARP® principles
- Assign executive responsibility for accountability
- Educate employees on information governance responsibilities
- Classify information correctly
- Ensure authenticity and integrity of information
- Determine that the official record is electronic unless specified directly
- Develop policies for alignment of business systems and third-parties to information governance standards
- Store, manage and make accessible, monitor and audit approve enterprise repositories and systems for records and content
- Secure confidential or personally identifiable information
- Control unnecessary growth of information
- Dispose information at the end of its lifecycle
- Comply with requests for information.
- Improve continuously

The Information Governance Reference model (IGRM) shows the relationship of Information Governance with other business functions. The outer ring shows the stakeholders who put policies, standards, processes, tools and infrastructure in place. The centre is a lifecycle diagram.



6.2 Proliferation of Information

The challenge of governance of unstructured data:

- Unstructured data grows much faster than structured data.
- Ownership can be difficult to ascertain
- Difficult to classify as business purpose cannot always be inferred
- Unstructured data without Metadata is a risk as it can be misinterpreted and may be mishandled

6.3 Govern for Quality Content

Requires effective partnership between data stewards and other data management professionals and records managers. Business data stewards can help define web portals, enterprise taxonomies, search engine indexes and content management issues.

Defining quality content requires understanding the context of its production and use:

- **Producers:** Who creates the content and why
- **Consumers:** Who uses the information and for what purpose
- **Timing:** When and how frequently the information is needed
- **Format:** Is a consistent format required? Are there unacceptable formats?
- **Delivery:** How the information is delivered, how it will be accessed and security to prevent inappropriate access

6.4 Metrics

KPIs at strategic and operational levels, especially if they measure lifecycle functions or risks.

6.4.1 Records Management

Strategic level:

- Within regulatory requirements e.g. time taken to meet a requirement
- Governance e.g. compliance with policies

Operational level:

- Within areas of records management resources e.g. operational and capital costs
- Training
- Delivery of daily records management services and operations
- Integration of records management functions with other business functions

Criteria to measure the success of a records management system:

- Percentage of total documents and email per user identified as corporate records
- Percentage of these records put under records control
- Percentage of the total stored records that have proper retention rules applied.

Use ARMA's GARP principles and maturity model to guide definition of KPIs.

6.4.2 E-Discovery

KPIs of e-discovery:

- cost reduction
- efficiency of collecting documents ahead of time rather than reactively
- How quickly an organisation can implement a legal hold notification process (LHN)

6.4.3 ECM (Electronic Content Management)

KPIs measure tangible and intangible benefits to the organisation:

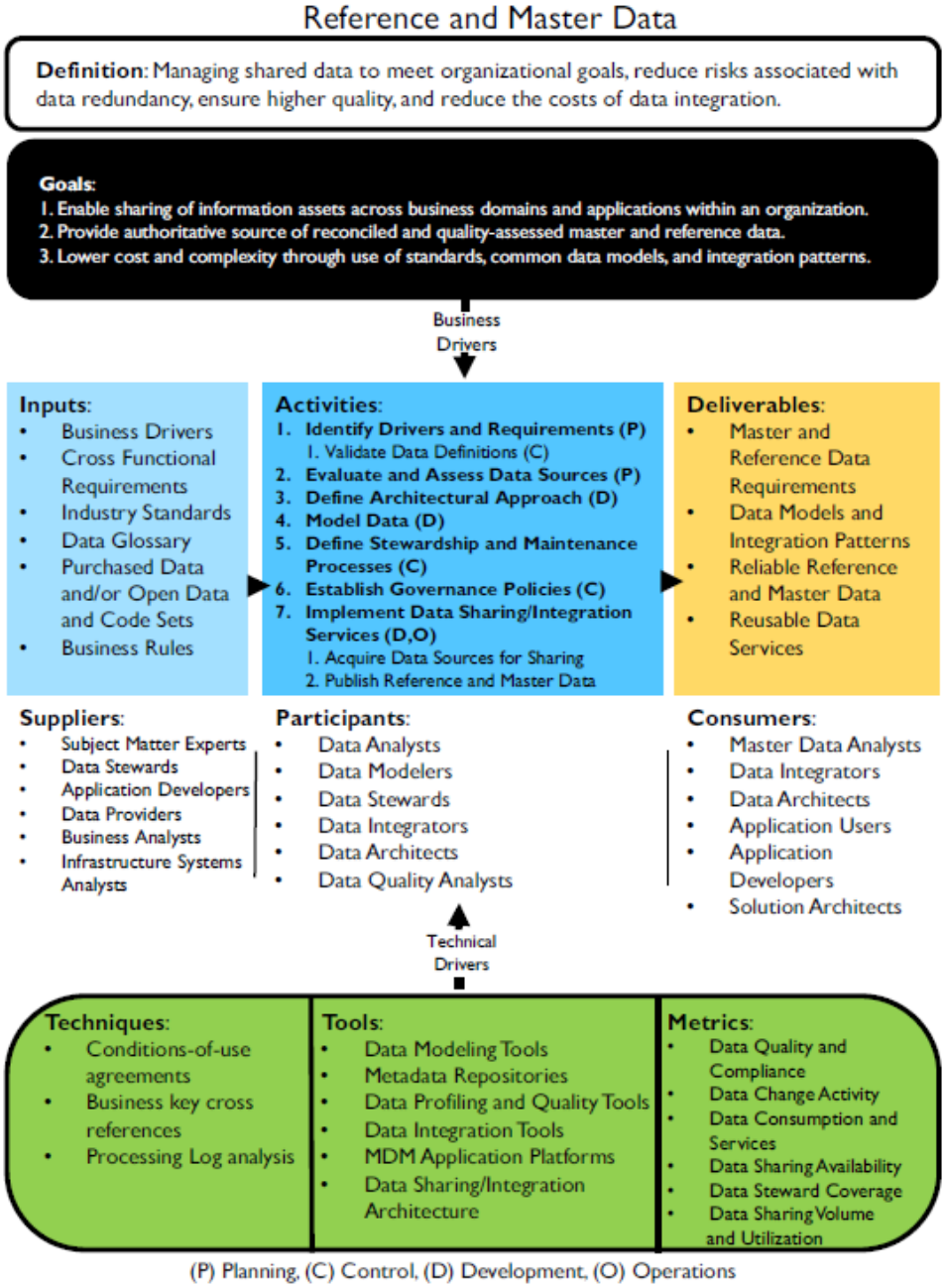
- Tangible:
    - Increased productivity
    - Cost reduction
    - Improved information quality
    - Improved compliance
- Intangible:
    - Improved collaboration
    - simplification of job routines and workflow

Reference and Master Data

1 Introduction

An organisation and its customers benefit if the data required across business areas, processes and systems is shared, allowing the same customer lists, geographic codes, parts codes etc. to be accessed, to produce a level of consistency.

Systems and data evolve organically resulting in multiple systems executing similar functions isolated from each other, leading to inconsistencies in data structure and values, and increased costs and risks. Both can be reduced through the management of reference and master data.



1.1 Business Drivers

For Master Data management program:

- **Meeting organisational data requirements**: Multiple areas need to access same data sets, which confidence that they are complete and consistent
- **Managing data quality**: MDM enables a consistent representation of critical entities
- **Managing the costs of data integration:**
- **Reducing risk:** Simplifies the data sharing environment

Centrally managed Reference Data enables the organisation to:

- Meet data requirements for multiple initiatives, reduce costs and risks of data integration
- Manage quality of reference data

1.2 Goals and Principles

Goals:
1. Enable sharing of information assets across business domains and applications within an organization.
2. Provide authoritative source of reconciled and quality-assessed master and reference data.
3. Lower cost and complexity through use of standards, common data models, and integration patterns.

R and MDM guiding principles:

- **Shared data:** managed to be sharable across organisation
- **Ownership**: Belong to the organisation. Require high level of stewardship
- **Quality**: Require ongoing monitoring and governance
- **Stewardship**: Business data stewards responsible
- **Controlled Change**:
    - **Master Data**: Represents the best view of currency and accuracy at any point in time. Caution when matching rules that change values. Should be reversible.
    - **Reference Data**: Change follows defined process. Approve and communicate before implementing
- **Authority**: Master data values should be replicated only from the system of record.

1.3 Essential concepts

1.3.1 Differences between Master and Reference Data

Malcolm Chisholm proposed a six-layer taxonomy of data (2008):

- Metadata
- Reference Data
- Enterprise structure data
- Transaction structure data
- Transaction activity data
- transaction audit data

Chisholm's definition of Master Data: An aggregation of Reference Data, enterprise structure data and transaction structure data.

- **Reference data:** code and description tables used to categorise other data in the organisation. Relates data in the database to information outside the organisation
- **Enterprise Structure Data:** Business activity data e.g. chart of accounts

- **Transaction Structure data:** Describes things that must be present for a transaction to occur e.g. products, customers, vendors

DAMA Dictionary Definition (2009): Master Data is the data that provides the context for business activity data in the form of common and abstract concepts that relate to the activity. It includes the details (definition and identifiers) of internal and external objects involved in business transactions, such as customers, products, employees, vendors and controlled domains (code values).

David Loshin describes Master Data objects as core business objects used in different applications across the organisation along with their associated Metadata, attributes, definitions, roles, connections and taxonomies. Master Data objects represent those things that matter most to the organisation, that are logged in transaction, reported on, measured and analysed.

Master Data requires identifying and developing a trusted version of the truth for each instance of conceptual entities, and maintaining the currency of that version. Master Data Management works to resolve the differences in associations between data in different systems and processes to consistently identify individual entity instances in different contexts. This process must be managed over time so that the identifiers for these Master Data entity instances remain consistent.

- Shared purposes of Reference and Master Data:
- Provide context to the creation and use of transactional data
- Reference data provides context for Master Data
- Enable data to be meaningfully understood
- Shared resources managed at the enterprise level
- Reference Data compared to Master Data sets:
- Less volatile
- Fewer columns and rows than transactional or master data sets
- No entity resolution challenges

Different focus of data management:

- **Master Data Management (MDM):** Control over Master Data values and identifiers that enable consistent use of the most accurate and timely data about essential business entities. Ensure availability of accurate current values while reducing risks of ambiguity.
- **Reference Data Management (RDM):** Control over defined domain values and their definitions. Ensure the organisation has access to a complete set of accurate and current values for each concept represented.

RDM is responsible for obtaining data and managing updates, as reference data can originate inside or outside the organisation.

1.3.2 Reference Data

Reference data is any data used to characterise or classify other data, or to relate data to information external to an organisation (Chisholm, 2001). Can be codes and description or more complex hierarchies and mappings.

Common storage techniques:

- Code tables in relational databases linked by foreign keys
- Reference Data Management systems
- Object attribute specific Metadata to specify permissible values for APIs

Reference Data Management entails control and maintenance of defined domain values, definitions and the relationships with and across domain values. The goal is to ensure values are consistent, current and accessible to the organisation

1.3.2.1 Reference data structure:

Depends on the granularity and complexity.

1.3.2.2 Lists

Code value and a description which can be used in a drop down.

Table 17 Simple Reference List

Code Value	Description
US	United States of America
GB	United Kingdom (Great Britain)

Definitions added for a Help function.

Table 18 Simple Reference List Expanded

Code	Description	Definition
1	New	Indicates a newly created ticket without an assigned resource
2	Assigned	Indicates a ticket that has a named resource assigned
3	Work In Progress	Indicates the assigned resource started working on the ticket
4	Resolved	Indicates request is assumed to be fulfilled per the assigned resource
5	Cancelled	Indicates request was cancelled based on requester interaction
6	Pending	Indicates request cannot proceed without additional information
7	Fulfilled	Indicates request was fulfilled and verified by the requester

1.3.2.3 Cross-reference lists

Used to translate code values of the same concept. May be at different granularities or same granularity with different values.

Table 19 Cross-Reference List

USPS State Code	ISO State Code	FIPS Numeric State Code	State Abbreviation	State Name	Formal State Name
CA	US-CA	06	Calif.	California	State of California
KY	US-KY	21	Ky.	Kentucky	Commonwealth of Kentucky
WI	US-WI	55	Wis.	Wisconsin	State of Wisconsin

Table 20 Multi-Language Reference List

ISO 3166-1 Alpha 2 Country Code	English Name	Local Name	Local Name Local Alphabet	French Name	…
CN	China	Zhong Guo	中国/中國	Chine	

1.3.2.4 Taxonomies

Taxonomic Reference Data structures capture information at different levels of specificity to support multifaceted navigation required by Business Intelligence.

Table 21 UNSPSC (Universal Standard Products and Services Classification)[57]

Code Value	Description	Parent Code
10161600	Floral plants	10160000
10161601	Rose plants	10161600
10161602	Poinsettias plants	10161600
10161603	Orchid plants	10161600
10161700	Cut flowers	10160000
10161705	Cut roses	10161700

1.3.2.5 Ontologies

Ontologies can be part of Reference Data as they are used to characterise other data or relate organisational data to information beyond the boundaries of the organisation.

1.3.2.6 Proprietary or internal reference data

Reference data created within the organisation to support internal systems. RDM consists of managing them and ensuring consistency.

1.3.2.7 Industry reference data

Industry Reference Data describes data sets which are created and maintained by industry associations and government bodies in order to provide a standard for codifying important concepts.

1.3.2.8 Geographic or geo-statistical data

Enables classification or analysis based on geography.

1.3.2.9 Computational reference data

Used for common, consistent calculations

1.3.2.10 Standard reference data set metadata

Maintain key Metadata about Reference Data sets to ensure their lineage and currency are understood and maintained.

Table 23 Critical Reference Data Metadata Attributes

Reference Data Set Key Information	Description
Formal Name	Official, especially if external name of the Reference Data set (e.g., ISO 3166-1991 Country Code List)
Internal Name	Name associated with the data set within the organization (e.g., Country Codes – ISO)
Data Provider	The party that provides and maintains the Reference Data set. This can be external (ISO), internal (a specific department), or external – extended (obtained from an external party but then extended and modified internally).
Data Provider Data Set Source	Description of where data provider's data sets can be obtained. This is likely a Universal Resource Identifier (URI) within or outside of the enterprise network.
Data Provider Latest Version Number	If available and maintained, this describes the latest version of the external data provider's data set where information may be added or deprecated from the version in the organization
Data Provider Latest Version Date	If available and maintained, this describes when the standard list was last updated
Internal Version Number	Version number of the current Reference Data set or version number of the last update that was applied against the data set
Internal Version Reconciliation Date	Date when data set was last updated based on the external source
Internal Version Last Update Date	Date data set was last changed. This does not mean reconciliation with an external version.

1.3.3 Master Data

Master Data is about the key business entities and should represent the authoritative most accurate data available, which can be trusted and used with confidence.

Business rules dictate the format and allowable values. Common organisational Master Data is data about:

- **Parties:** Individuals, organisations and their roles
- **Products and services:** Internal and external
- **Financial structures:** e.g. contracts, general ledger accounts
- **Locations:** e.g. addresses and GPS coordinates

1.3.3.1 System of Record, System of Reference

Where there are potentially different versions of the truth, we need to know more about the data to distinguish between them:

- **A System of Record** is an authoritative system where data is created/captured and maintained through a defined set of rules and expectations.
- **A System of Reference** is an authoritative system where data consumers can obtain reliable data to support transactions and analysis. Examples are MDM applications, Data Sharing Hubs and Data Warehouses.

1.3.3.2 Trusted Source, Golden Record

A **Trusted Source** is recognised as the "best version of the truth" based on a combination of automated rules and manual stewardship. (any MDM system)

A **Golden Record** represents the most accurate data about an entity, also referred to as a "single version of the truth". Not always possible for multiple systems to have one version of the truth.

1.3.3.3 Master Data Management

Gartner's Definition: A technology-enabled discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, semantic consistency and accountability of the enterprise's official shared Master Data assets. Master Data is the consistent and uniform set of identifiers and extended attributes that describes the core entities of the enterprise, including customers, prospects, citizens, suppliers, sites, hierarchies and chart of accounts.

MDM is a discipline made up of people, processes and technology.

Criteria to assess MDM requirements:

- Which roles, organisations, places and things are referenced repeatedly?
- What data is used to describe people, places and things
- How the data is defined, structured, and the granularity
- Where data is created/sourced, stored, made available and accessed
- How it changes as it moves through systems
- Who uses the data, and for what purposes?
- Criteria used to understand the quality and reliability of the data and its sources

Planning for Master Data Management within a domain:

- Identify candidate sources that will provide a comprehensive view of Master Data entities
- Develop rules for accurately matching and merging entity instances

- Establish an approach to identify and restore inappropriately matched and merged data
- Establish an approach to distribute trusted data systems across the enterprise

1.3.3.4 Master Data Management Key Processing steps

Figure 76 Key Processing Steps for MDM

- **Data Model Management:** Clear and consistent definitions make sense to business at the enterprise level
- **Data Acquisition:** Data representing the same entity can look different. Plan for acquiring new data as a reliable repeatable process. High level cleansing tools, and matching rules, then perform data quality on the new data.

Table 24 Source Data as Received by the MDM System

Source ID	Name	Address	Telephone
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, DA	2345678900
345	Jane Smith	123 Main, Dataland, DA	234-567-8900

- **Data Validation, Standardisation and Enrichment:** Reduce variation in format and reconcile values:
  - **Validation:** Identify clearly incorrect or defaulted data
  - **Standardisation:** Data conforms to standard Reference Data values and formats
  - **Enrichment:** Add attributes that improve entity resolution services

Table 25 Standardized and Enriched Input Data

Source ID	Name	Address (Cleansed)	Telephone (Cleansed)
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900
345	Jane Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900

- **Entity Resolution and Identifier Management:**

  Entity resolution is the process of determining whether two references to real world objects refer to the same object or different objects.

  Activities:

  - **Matching:** or candidate identification is the process of identifying how different records relate to a single entity. Use similarity analysis to avoid false positives or negatives.
  - **Identity Resolution:** Keep a history of matches so that those less confident matches due to conflicting values can be undone if found to be incorrect.

Table 26 Candidate Identification and Identity Resolution

Source ID	Name	Address (Cleansed)	Telephone (Cleansed)	Candidate ID	Party ID
123	John Smith	123 Main, Dataland, SQ 98765		XYZ	1
234	J. Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900	XYZ, ABC	2
345	Jane Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900	ABC	2

  - **Matching Workflows / Reconciliation Types:** Different scenarios require different workflows:
    - **Duplicate identification match rules:** Specific set of data elements that uniquely identify an entity and identify merge opportunities
    - **Match-link rules:** Identify and cross-reference records that appear to relate to the master record without updating the content of the cross-referenced record
    - **Match-merge rules:** Match records and merge data into a single unified and comprehensive record. Complex.
  - **Master Data ID Management:** Two types of identifiers managed in a MDM environment:
    - **Global ID** is the MDM solution assigned unique identifier attached to reconciled records. Should be automatically generated.
    - **X-Ref Management** is management of the relationship between source IDs and the Global ID.
  - **Affiliation Management:** Establishing and maintaining relationships between Master Data records of entities that have real-world relationships. Data Architecture design of the MDM system which kind of relationship between entities:
    - **Affiliation relationships** are programmed and are the most flexible
    - **Parent-child relationships** have implied hierarchical navigation structure
- **Data Sharing and Stewardship:** Data Stewards resolve incorrectly matched situations and improve matching algorithms.

1.3.3.5 Party Master Data

Data about individuals, organisations and the role they play in business relationships. Examples from different environments:

- **Commercial:** customers, employees, vendors, partners, competitors
- **Public sector:** citizens
- **Law enforcement:** suspects, witnesses, victims
- **not for profit:** members, donors
- **healthcare:** patients, providers
- **Education:** students, faculty

Customer relationship Management (CRM) systems manage Master Data about customers

Master Data is challenging for parties playing more than one role in an organisation.

1.3.3.6 Financial Master Data

Data about business units, cost centres, profit centres, general ledger accounts, budgets, projections and projects. The central hub of financial Master Data is an Enterprise Resource Planning (ERP) system.

1.3.3.7 Legal Master Data

Data about contracts, regulations and other legal matters.

1.3.3.8 Product Master Data

Can focus on the organisation's products and services or on industry-wide products and services. Different types of product Master Data solutions:

- **Product Lifecycle Management (PLM):** managing the lifecycle of a product/service from conception to disposal
- **Product Data Management (PDM):** engineering and manufacturing. Enables secure sharing of product information such as design drawings (CAD)
- **Product data in Enterprise Resource Planning (ERP):** SKUs to support order entry to inventory
- **Product data in Manufacturing Execution Systems (MES):** Raw inventory to finished goods
- **Product data in Customer Relationship Management (CRM):** Marketing, sales and support interactions

1.3.3.9 Location Master Data

The ability to track and share geographic information and to create hierarchical relationships based on geographic information.

- **Location Reference Data** is usually geopolitical data handled by external organisations
- **Location Master Data** are address related to parties and businesses

1.3.3.10 Industry Master Data – Reference Dictionaries

Authoritative listings of Master Data entries that can be purchased. They can provide a starting point for matching and linking new records.

1.3.4 Data Sharing Architecture

Each Master Data subject area usually has its own system of record e.g. CRM or ERP systems. Hub-and-spoke model for sharing Maser Data:

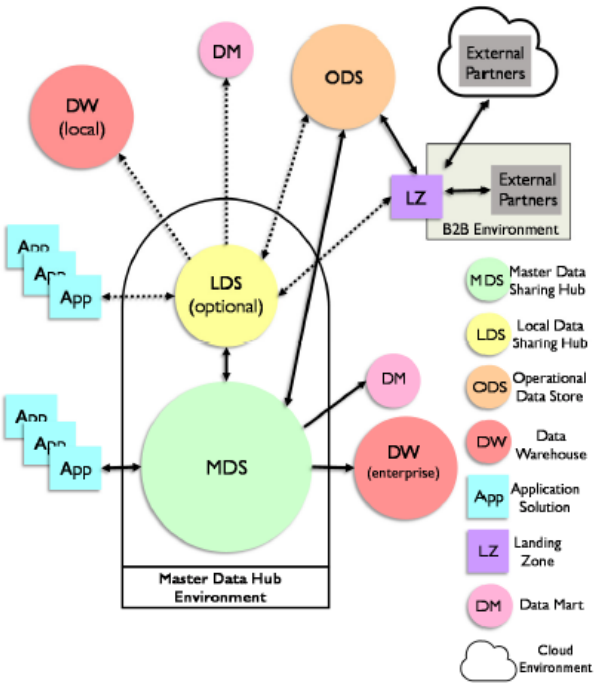


Figure 77 Master Data Sharing Architecture Example

Three approaches to implementing Master Data hub environment:

- **Registry:** An index that points to Master Data in various systems of record which manage Master Data local to their applications. Easy to implement, but complex queries are challenging. Multiple business rules need to be implemented.
- **Transaction Hub:** Applications interface with the hub to access and update Master Data which exists only within the Transaction Hub and not in the applications. The Hub is the system of record for Master Data. Better governance, but business rules reside in the Hub.
- **Consolidated approach**: Systems of record manage Master Data for their applications, and it is also consolidated within a common repository (replication). There is no need to access directly from the systems of record.

2 Activities

2.1 MDM Activities

- **Define MDM Drivers and Requirements:**
  - Easier to define requirements for an application than the whole enterprise
  - Prioritise Master Data efforts on cost benefit of the proposed improvements
  - Start with simplest category to learn from the process
- **Evaluate and Assess Data Sources**
  - Understand the structure of existing data and how it is collected or created
  - Understand the quality of the data, as poor quality data complicates a Master Data project
  - Assess disparity between sources
  - May be able to purchase standardised data such as Reference Directories
- **Define Architectural Approach:** Depends on business strategy, the platforms for existing sources and the lineage and volatility of the data.
- **Model Master Data:** As Master Data is an integration process, model data within subject areas. A logical or canonical model
- **Define Stewardship and Maintenance Processes:** Technical solutions still require the oversight of Data Stewards to address records that fall out of the process and why.
- **Establish Governance Policies to enforce use of Master Data:** The benefits come once people start using the Master Data

2.2 Reference Data Activities

- **Define Drivers and Requirements:**
  - **Drivers:** Operational efficiency and higher data quality
  - **Requirements:** Driven by the most important reference data sets
- **Evaluate and Assess Data Sources:**
  - **External:** vendor that guarantees updates on a schedule and ensures quality data
  - **Internal:** Owners should understand the benefits of central management of their data sets
- **Define Architectural Approach:** Tool should allow for manual updates
- **Model Reference Data sets:** Models help consumers understand the relationships within the reference data sets, and the data quality rules
- **Define Stewardship and Maintenance Processes:** Capture Metadata about reference data sets
- **Establish Reference Data Governance Policies:**

3 Tools and Techniques

MDM requires identity management enabled tools:

- Data integration tools
- Data remediation tools
- Operational data stores (ODS)
- Data sharing hubs (DSH)
- specialised MDM applications

4 Implementation Guidelines

As Master and Reference Data Management are forms of data integration, the same implementation principles that apply to Data Integration and Interoperability.

Implement incrementally through a series of projects defined in an implementation roadmap, prioritised on business needs and guided by an overall architecture.

It is vital to have data governance professionals who understand the challenges of RDM and MDM and can assess the maturity and ability of the organisation to meet them.

4.1 Adhere to Master Data Architecture

The integration process should take into account:

- The organisational structure of the business
- the number of distinct systems of record
- the data governance implementation
- The importance of access and latency of data values
- The number of consuming systems and applications

4.2 Monitor data movement

Monitor data as it flows within the Reference or Master Data sharing environment to:

- Show how data is used across the organisation
- Identify data lineage from/to administrative systems and applications
- Assist root cause analysis of issues
- Show effectiveness of data ingestion and consumption integration techniques
- Denote latency of data values from source systems through consumption
- Determine validity of business rules and transformations executed within integration components

4.3 Manage Reference Data change

Reference data is a shared resource, therefore it should not be locally controlled, but channels to receive and respond to change requests must be provided according to policies and procedures put in place by the Governance Council.

Planned/scheduled changes such as periodic updates to industry codes require less governance than ad hoc changes.

Figure 78 Reference Data Change Request Process

4.4 Data sharing agreements

Data sharing agreements stipulate what data can be shared and under what conditions. Helps when issues regarding quality of data brought in or availability arise. Driven by the Data Governance program and involves Data Architects, Data Providers, Data Stewards, Application Developers, Business Analysts, Compliance/Privacy Officers and Security Officers.

SLAs should be in place so that the quality data can be provided to downstream consumers.

5 Organisation and Cultural Change

It is not easy for people to relinquish control of their data to create shared resources. People may perceive MDM and RDM efforts as adding complications to their processes.

The most challenging cultural change is determining which individuals are accountable for which decisions.

6 Reference and Master Data Governance

Because they are shared resources, Reference and Master Data require governance and stewardship. Governance processes will determine:

- The data sources to be integrated
- Data quality rules to be enforced
- conditions of use rules
- Activities to be monitored and the frequency of monitoring
- Priority and response levels of stewardship efforts
- How information is represented to meet stakeholder needs
- Standard approval gates, expectations in RDM and MDM deployment

Governance processes bring compliance and legal stakeholders together with information consumers to ensure risks are mitigated through definition and incorporation of privacy, security and retention policies.

Data Governance must have the ability to review, receive and consider new requirements and changes. Make principles, rules and guidelines available to all using Reference and Master Data.

6.1 Metrics

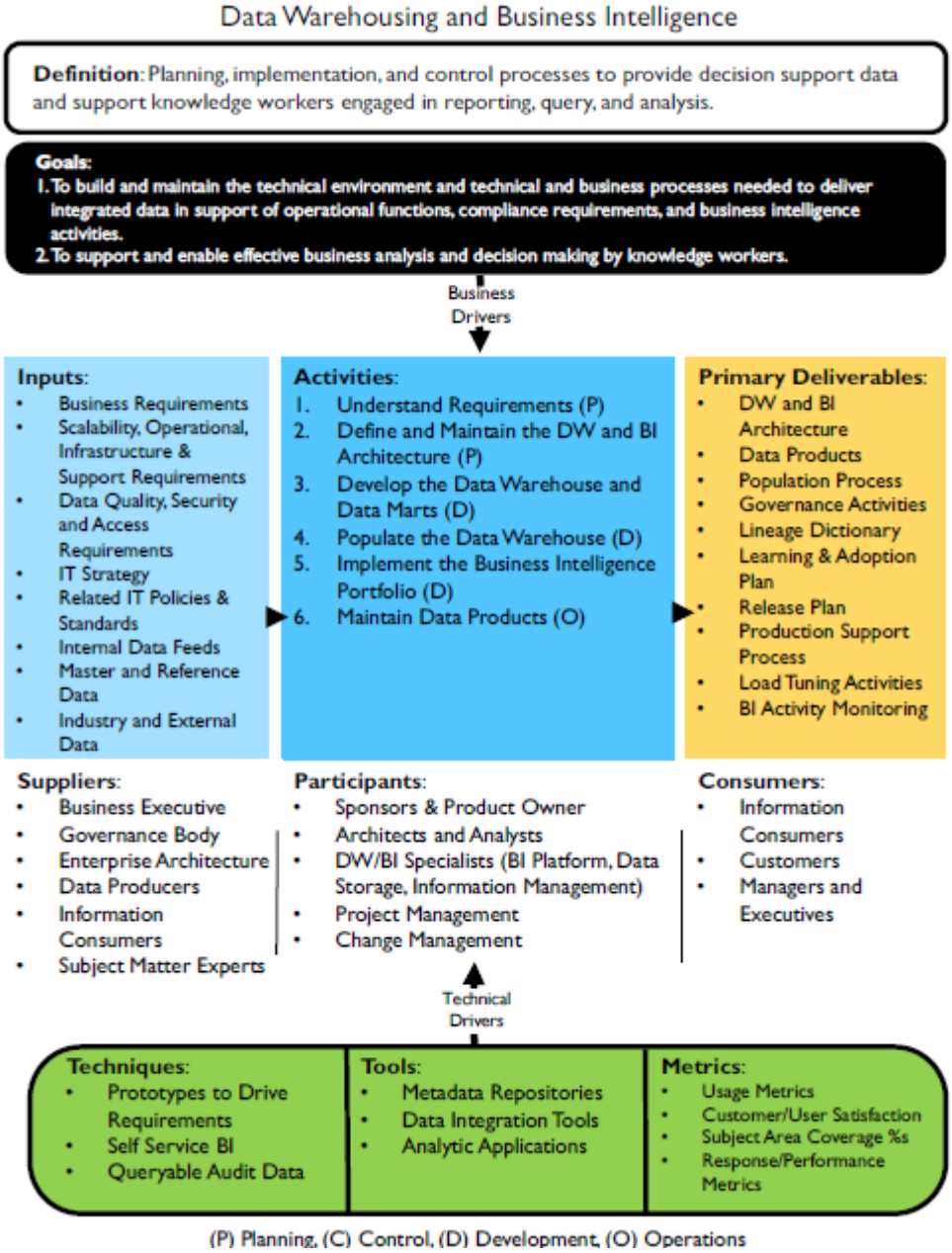
- **Data quality and compliance:** DQ dashboards
- **Data change activity:**
  - Metrics denote rate of change of data values
  - Provide insight to systems providing data to sharing environment
  - Used to tune algorithms in MDM process
- **Data ingestion and consumption:** Denote and track data contributing systems and what business areas are subscribing to shared data

- **Service Level Agreements:** Level of adherence to SLAs provides insight to data and technical problems
- **Data Steward coverage:** Used to identify gaps in support
- **Total cost of Ownership:** Must be consistently applied across the organisation to be effective
- **Data sharing volume and usage:** The volume and velocity of data defined, ingested and subscribed to and from the data sharing environment.

Data Warehousing and Business Intelligence

The data warehouse is meant to enable decision support systems which could share core enterprise data from a common data model. The enterprise warehouse reduces data redundancy, improves the consistency on information, and enables the enterprise to make better decisions.

1 Introduction



1.1 Business Drivers

- Business Intelligence support
- Compliance requirements
- Support operational activities

- Enables effective business analysis and decision-making
- Find ways to innovate based on insights from data

1.2 Goals and Principles

Goals:
1. To build and maintain the technical environment and technical and business processes needed to deliver integrated data in support of operational functions, compliance requirements, and business intelligence activities.
2. To support and enable effective business analysis and decision making by knowledge workers.

Organisations implement data warehouses to:

- Support BI activity
- Enable effective business analysis and decision making
- Find ways to innovate based on insights from data

Principles to implement a Data Warehouse:

- **Focus on business goals:** DW solves business problems
- **Start with the end in mind:** Business priority and scope end-data-delivery in BI space drives creation of DW content
- **Think and design globally; act and build locally:** Architecture guided by end-vision, but build and deliver in sprints to enable return on investment.
- **Summarise and optimise last, not first:** Build on atomic data
- **Promote transparency and self-service:** Provide more context (Metadata) and keep users informed of updates and changes
- **Build Metadata with the warehouse:** To be able to explain the data, capture as part of the development cycle and manage as an ongoing activity
- **Collaborate:** With other data initiatives especially DG, DQ and Metadata
- **One size does not fit all:** Different groups of data consumers need different tools

1.3 Essential Concepts

1.3.1 Business Intelligence

Two meanings:

- **Type of data analysis** aimed at understanding organisational activities and opportunities
- **A set of technologies** that enable decision support analysis (querying, data mining, statistical analysis, reporting, scenario modelling, data visualisation and dashboarding).

1.3.2 Data Warehouse

A Data Warehouse (DW) is a combination of two components:

- Integrated decision support database
- Software programs used to collect, cleanse, transform and store data from a variety of internal and external sources

An **Enterprise Data Warehouse** (EDW) is as centralised data warehouse designed to service the BI needs of the entire organisation. Adheres to the enterprise data model.

1.3.3 Data Warehousing

Data warehousing describes the operational extract, cleansing, transformation, control and load processes that maintain the data in a data warehouse. Integrated, historical, enforces business rules and maintains business relationships. Processes interact with Metadata repositories.

- **Structured data:** elements in defined fields as documented in data models
- **Semi-structured data:** Electronic elements organised as semantic entities with no required attribute affinity
- **Unstructured data:** Not predefined through a data model

1.3.4 Approaches to Data Warehousing

Two thought leaders – Bill Inmon and Ralph Kimball:

- **Inmon**:
  - A DW is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making processes.
  - Normalised, relational model
- **Kimball**:
  - A copy of transaction data specifically structured for query and analysis.
  - Dimensional model

Core ideas:

- Warehouses store data from other systems
- Storage includes organising data in ways that increases its value
- Warehouses make data accessible and useable for analysis
- Organisations build warehouses to make reliable, integrated data available to authorised stakeholders
- Warehouse data serves many purposes

1.3.5 Corporate Information Factory (Inmon)

CIF Illustrates the differences between warehouses and operational systems

- **Subject oriented:** Based on major business entities, not function or application
- **Integrated:** The warehouse becomes a system of record for the data.
- **Time variant:** Stores data as it exists at a set point in time
- **Non-volatile:** New records are appended not updated
- **Aggregate and detail data:** Details of atomic level transactions as well as summarised data
- **Historical:** The focus of operational systems is current data. Warehouses contain lots of historical data as well.

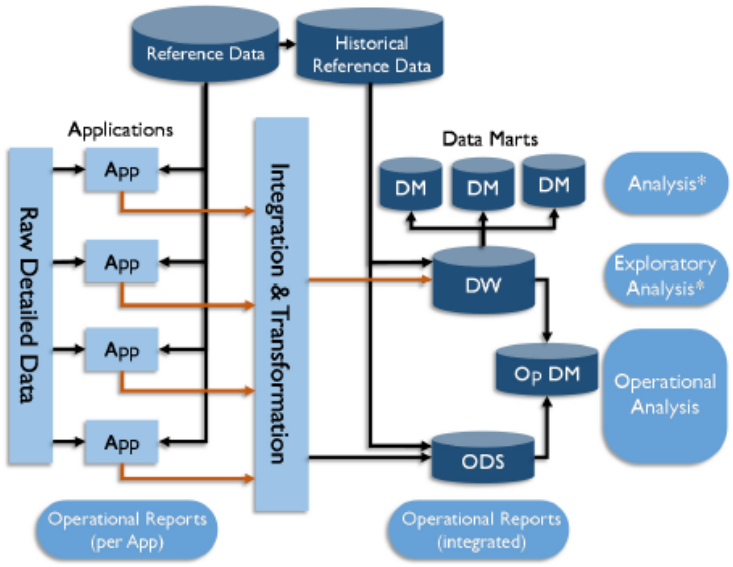


Figure 80 The Corporate Information Factory

CIF Components:

- **Applications:** Applications perform operational processes and bring data into the DW and operational data store (ODS) where it can be analysed
- **Staging area:** A database between the operational and target where the ETL takes place. Data is transient.
- **Integration and transformation:** Data from disparate sources is transformed in the integration layer into the standard corporate representation model in the DW and ODS
- **Operational data store (ODS):** an integrated relational database of operational data, containing current or near-term data, volatile. Meet the need for low latency data. Can be the primary source for the data warehouse.
- **Data Marts:** Provide data prepared for analysis. Sub-set of the warehouse designed for the needs of specific consumers. Data marts are designed using dimensional modelling and denormalisation.
- **Operational Data Mart (OpDM):** Sourced from ODS, volatile and focused on tactical decision support
- **Data Warehouse:** A single integration point for corporate data to support management decision making and strategic analysis and planning.
- **Operational reports:** The output from the data stores.
- **Reference, master and external data:** Data required to understand data from applications, and simplifies integration to DW.

Comparison between data stored in DW and Marts and Data on operational systems:

Data in DW and Marts	Data in Operational Systems
• Organised by subject	• Organised by function
• Integrated	• Siloed
• Data is time-variant	• Current value only
• High latency	• Lower latency
• Significantly more historical data	• Less historical data

1.3.6 Dimensional DW (Kimball)

Kimball: “a copy of transaction data specifically structured for query and analysis”.

Warehouse data is stored in a dimensional model, not normalised, enabling consumers to use the data and optimise query performance.

Star Schema or **dimensional** models:

- **Facts** contain quantitative data about business processes. (The fact table is also called a meter which contains measures – Hoberman)
- **Dimensions** store descriptive attributes related to fact data and allow consumers to answer questions about the facts

One Fact table is joined to many dimensions – looks like a star.

Conformed dimensions are common dimensions and are shared by multiple fact tables via a bus. Multiple data marts can also be integrated at enterprise level by plugging into the bus of conformed dimensions.

Table 27 DW-Bus Matrix Example

	Subject Areas				
Business Processes	**Date**	**Product**	**Store**	**Vendor**	**Warehouse**
Sales	X	X	X		
Inventory	X	X	X	X	X
Orders	X	X		X	
Conformed Dimension Candidate	*Yes*	*Yes*	*Yes*	*Yes*	*No*

Kimbal’s Data Warehouse Chess Pieces is more expansive than Inmon’s:

- **Operational source systems**: Equivalent to the applications in CIF
- **Data Staging area:** Processes needed to integrate and transform data for presentation. Kimball’s EDW
- **Data presentation area:** Like data marts in the CIF. The DW Bus conformed dimensions
- **Data access tools:** End users’ needs drive adoption of data access tools

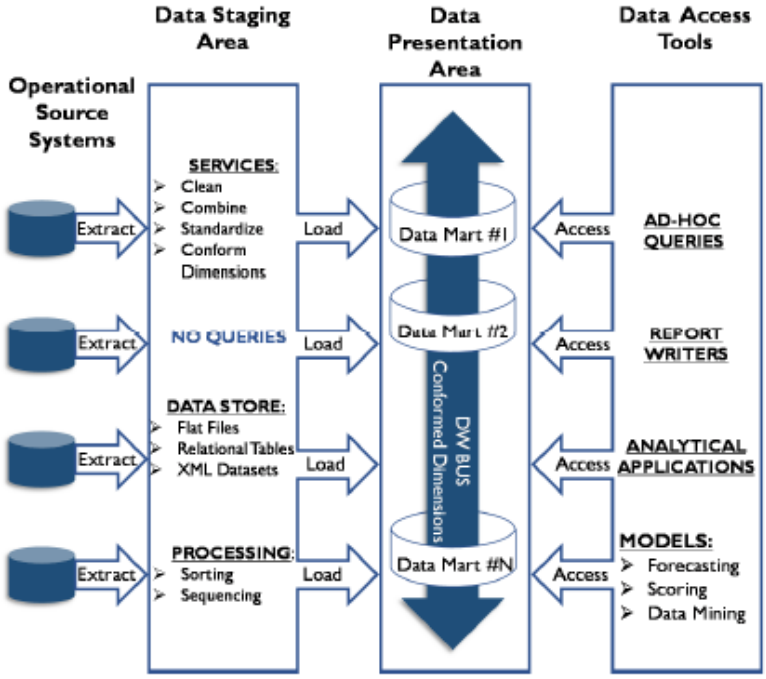
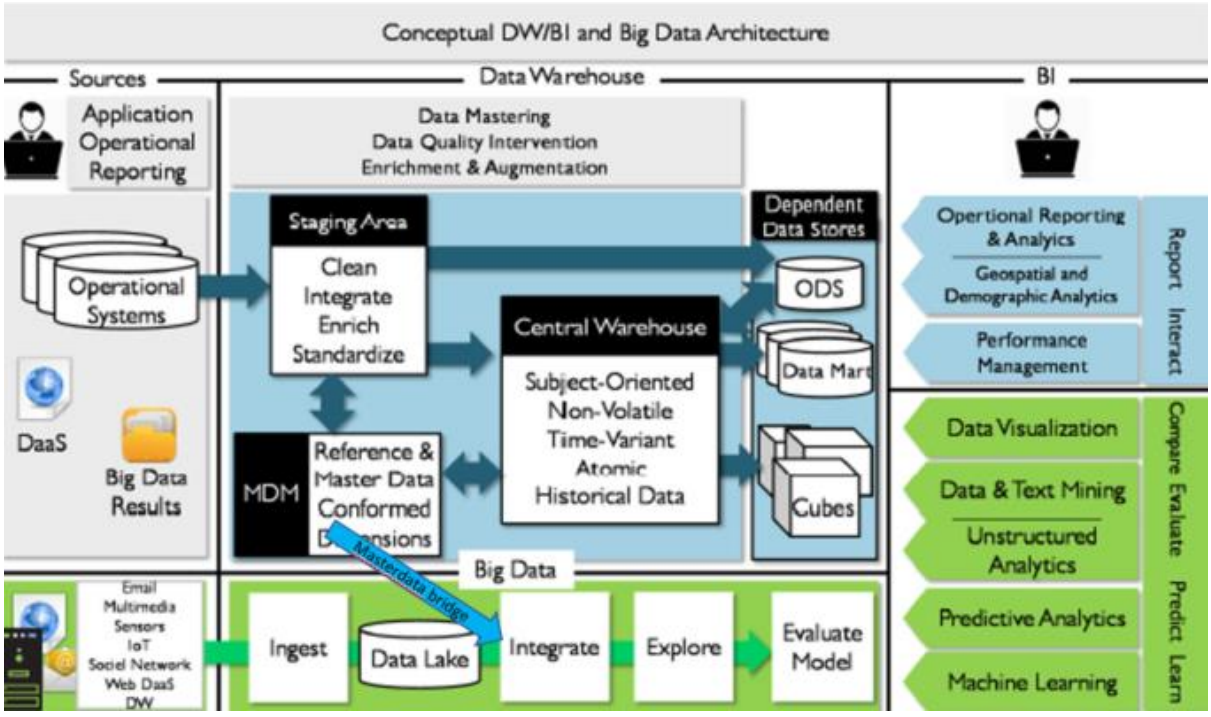


Figure 81 Kimball's Data Warehouse Chess Pieces[67]

1.3.7 DW Architecture Components



The evolution of Big Data has added to the DW/BI landscape by providing another data entry path. Lifecycle is depicted:

- **Source Systems:** Operational systems and external data to be brought into the DW/BI environment
- **Data Integration:** ETL, data virtualisation and other techniques to get data into a common form and location. (Arrows represent the integration process)
- **Data Storage Areas:** The warehouse has a set of storage areas

- **Staging area:** Intermediate storage for ETL and integration
- **Reference and Master data conformed dimensions:** May be separate repositories
- **Central warehouse:** DW Data usually persists in the central or atomic layer, which maintains all historic atomic data as well as the latest batch run. Data structure based on performance needs and use patterns:
  - Relationship between Business and surrogate keys for performance
  - Indexes and foreign keys to support dimensions
  - Change data capture (CDC) techniques that detect, maintain and store history
- **Operational data store (ODS):** Version of the central persisted store that supports operational use
- **Data Marts:** Type of data store often used to support presentation layers of the data warehouse environment
- **Cubes:** Three classic implementation approaches support Online Analytical Processing (OLAP). Their names relate to underlying database types, such as Relational (ROLAP), Multi-dimensional (MOLAP) and Hybrid (HOLAP).

1.3.8 Types of Load Processing

Historical (loaded once) and ongoing updates (consistently scheduled)

- **Historical Data:** Data warehouse captures detailed history of the data it stores
  - **Inmon:** All data stored in single DW layer with common integration and transformation layer. (Need Enterprise Model for success)
  - **Kimball:** DW is a combination of departmental data marts which store history at atomic level
  - **Data Vault:** Hybrid between 3NF and Star Schema (See DM & D Chapter 5). Hubs (PK), Links and Satellites. Facts persist as atomic structures.
- **Batch Change Data Capture:** Different change capture techniques:

Table 28 CDC Technique Comparison

Method	Source System Requirement	Complexity	Fact Load	Dimension Load	Overlap	Deletes
Time stamped Delta Load	Changes in the source system are stamped with the system date and time.	Low	Fast	Fast	Yes	No
Log Table Delta Load	Source system changes are captured and stored in log tables	Medium	Nominal	Nominal	Yes	Yes
Database Transaction Log	Database captures changes in the transaction log	High	Nominal	Nominal	No	Yes
Message Delta	Source system changes are published as [near] real-time messages	Extreme	Slow	Slow	No	Yes
Full Load	No change indicator, tables extracted in full and compared to identify change	Simple	Slow	Nominal	Yes	Yes

- **Near-real-time and Real-time:** Operational BI requires lower latency and the inclusion of volatile data in the warehouse.
  - **Trickle feeds (source accumulation):** Batch loads on a more frequent schedule, or when a threshold is reached

- **Messaging (Bus accumulation):** Small messages are published to a bus when they occur. Used by Data-as-a-Service (DaaS)
- **Streaming (Target accumulation):** A target system collects data as it is received into a buffer or queue.

2 Activities

2.1 Understand Requirements

Operational systems depend on precise, specific requirements. A data warehouse brings together data that will be used in many systems, used to explore and analyse.

2.2 Define and maintain the DW/BI Architecture

DW/BI architecture describes where the data comes from, where it goes, when it goes, why and how it goes into the warehouse. The technical requirements include performance, availability and timing needs.

2.2.1 Define DW/BI Technical Architecture

DW/BI architectures should have a mechanism to connect back to the transactional and operational reports in an atomic DW.

Conceptual model aligns with business needs. Test by prototyping. Validate with the Enterprise Data Model.

2.2.2 Define DW/BI Management Processes

Address production management with a coordinates and integrated maintenance process, delivering regular releases to the business community. Establish a release schedule to manage each update to the deployed data product as a software release.

2.3 Develop the Data Warehouse and Data Marts

Three concurrent development tracks:

- **Data:** The data business requires for the analysis it wants to do:
  - Identify best sources
  - Remediation, transformation, integration, storage and availability rules
  - How to handle data that does not fit
- **Technology**: Back end systems and processes supporting data storage and movement
- **Business Intelligence tools**

2.3.1.1 Map Sources to Targets

Source to target mapping establishes transformation rules for data elements from individual sources to the target system. A solid taxonomy is necessary. It is often the logical model.

2.3.1.2 Remediate and transform data

Cleansing activities enforce standards to correct and enhance the domain values of individual data elements. Should be done in source systems.

2.4 Populate the data warehouse

The DW/BI team must publish clear rules for what data detail the DW contains, and what will be available via only operational reporting.

Key factors to consider when defining the population approach:

- required latency
- availability of sources
- batch windows or upload intervals
- target databases
- dimensional aspects
- timeframe consistency of the data warehouse and data mart.
- Data quality processing
- Time to perform transformations and late-arriving dimensions and data rejects
- Change data capture processes

2.5 Implement the Business Intelligence Portfolio

Identify the right tools for business communities – find similarities through alignment of common business processes, analyses, management styles and requirements.

- **Group users according to needs:** Know the user groups, then match the tool.
    - IT Developers extracting data
    - Information consumers
    - Users' needs may change according to roles and skills
- Match tools to user requirements:
    - Imbedded analytics
    - Virtualisation
    - BI Suites

2.6 Maintain Data Products

An implemented data warehouse and its customer-facing BI tools is a data product. Enhancements and extensions should be implemented incrementally.

- **Release Management:** Release management is critical to an incremental development process that grow new capabilities. This process keeps the warehouse up-to-date, clean and performing at its best. Below a quarterly schedule is illustrated.

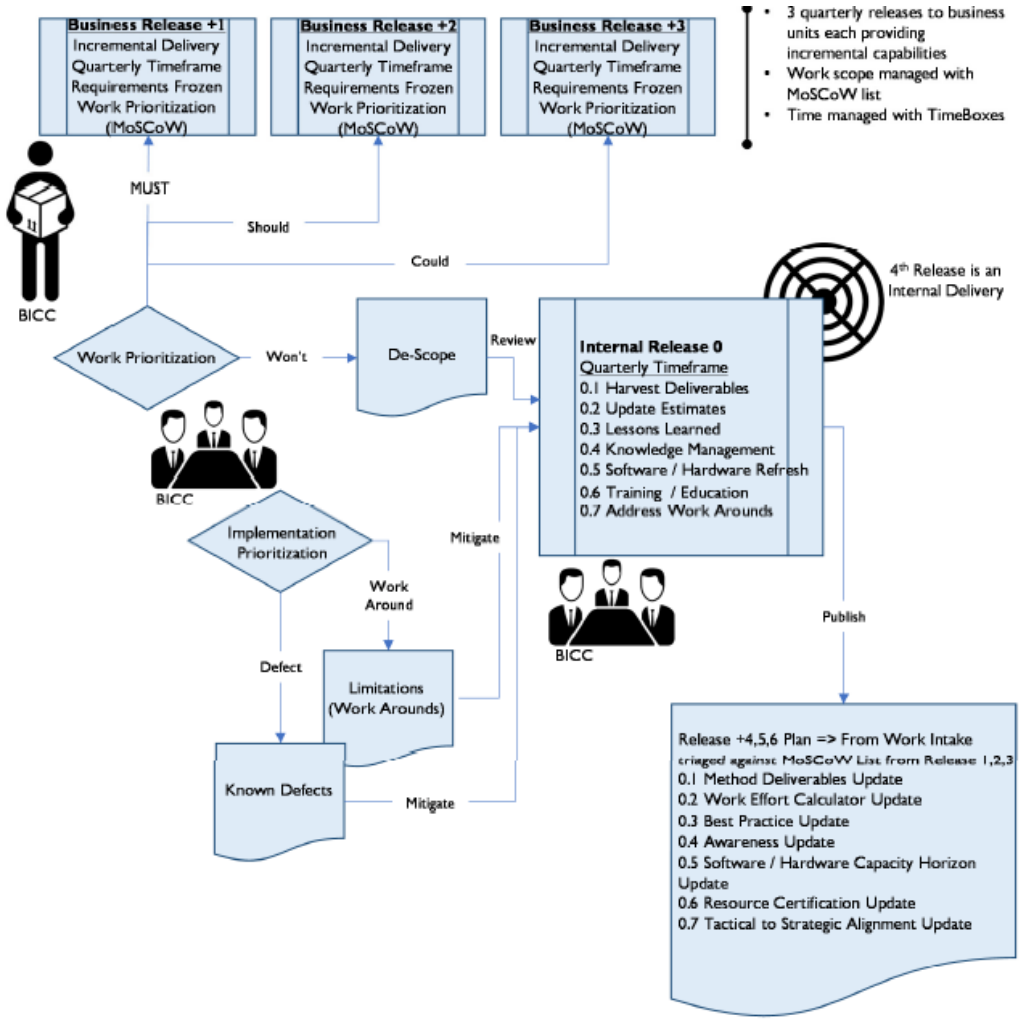


Figure 83 Release Process Example

- **Manage Data Product Development Lifecycle:** While the consumers are using the DW, the DW team is preparing the next iteration, which can extend an existing increment or add new functionality by onboarding a new business unit.
- **Monitor and tune load processes:** Monitor for bottlenecks and dependencies. Tune the database. Users consider the warehouse an active archive.
- **Monitor and tune BI activity and performance:** Customer-facing satisfaction metrics

3 Tools

- **Metadata repository:** Automate and integrate population
  - **Data dictionary / glossary:** Necessary for use of DW. Contains business terms and other information needed to use data.
  - **Data and Data Model Lineage:** Integration tools offer lineage analysis for the population code and the physical model and database. Uses for documented data lineage:
    - investigation of root causes of issues
    - impact analysis of changes
    - determine reliability of data based on origin
- **Data Integration Tools:** Used to populate the warehouse. Also schedule jobs that account for complex data delivery from many sources.
- **Business Intelligence Tools Types:**

- **Operational Reporting:** Analyse business trends to discover patterns. Tactical BI to support short term decisions
- **Business Performance Management:** Formal assessment of metrics aligned with organisational goals. Strategic BI to support long term corporate goals and objectives.
- **Descriptive, self-service analytics:** BI to the front line of business. Service-oriented Architecture (SOA) and Big Data.
- **Operational Analytic Applications:** Differ from OLAP and BI tools as they include processes to extract data from source systems, a data mart and pre-built reports and dashboards.
- **Multi-dimensional Analysis – OLAP:** Online Analytical processing
  - **Slice:** A subset of a multidimensional array corresponding to a single value for one or more members not in the subset
  - **Dice:** A slice on more than two dimensions of a data cube or more than two consecutive slices
  - **Drill down/up:** User navigates among levels of data from most summarised (up) to most detailed (down)
  - **Roll-up:** Compute all the data relationships for one or more dimensions.
  - **Pivot:** Changes the dimensional orientation of a report or page display
- **3 implementation processes:**
  - **Relational Online Analytical Processing (ROLAP):** Supports OLAP using multidimensionality in the two-dimensional tables of RDBMS
  - **Multi-dimensional Online Analytical Processing (MOLAP):** Supports OLAP by using specialised multi-dimensional database technology
  - **Hybrid Online Analytical Processing (HOLAP):** Combination of ROLAP and MOLAP

4 Techniques

- Prototypes to drive requirements
  - Profile the data:
    - Reduce risk of unexpected data
    - Disclose differences in sources that may be obstacles in integration
- Self-service BI
- Audit data that can be queried

5 Implementation Guidelines

- **Readiness assessment / Risk Assessment:** May be a gap between embracing venture and ability to sustain it
  - Pre-requisite checklist:
    - Has business support
    - Aligned with strategy
    - has defined architectural approach
  - Define data sensitivity and security constraints
  - Select tools
  - Secure resources
  - Create source data ingestion process
  - Identify sensitive or restricted data that may need masking
  - Ensure data governance processes for review and approval have been followed

- **Release Roadmap:** Suggested approach:
  - Incremental leveraging of DW Bus matrix as a marketing and communication tool
  - Use business-determined priorities tethered to exposure metrics to determine how much rigor and exposure to apply to each increment
  - Apply consistent needs and abilities processes to determine next business unit to onboard.
  - Maintain a back-order work item list to identify outstanding capabilities
  - Determine technical difficulties which may alter the order of delivery
  - Package work into a software release
  - Agree on pace of release (quarterly, monthly, weekly or faster if appropriate)
  - Roadmap of release dates to manage with business partners
- **Configuration Management:**
  - Aligns with release roadmap
  - Scripts to automate development, testing and transport to production
  - Provides version control as it brands the release at the database level
  - Automated and manually generated programs are tied to that brand
- **Organisation and Cultural Change:** Keep business focus throughout the DW/BI lifecycle. Align projects behind real business needs and assess necessary business support. Critical success factors are:
  - **Business sponsorship:** DW/BI projects require strong executive sponsorship, an engaged steering committee and commensurate funding
  - **Business goals and scope:** Clearly identified business need
  - **Business resources:** Management commitment to resource
  - **Business readiness:**
    - Business prepared for long term incremental delivery
    - Business committed to establishing centres of excellence to sustain product
    - Breadth of knowledge or skill gap within target community that can be crossed within a single increment
  - **Vision alignment:** How well does the IT strategy support business vision?
  - **Dedicated Team:** To manage ongoing operations of the production environment:
    - **Front office group** to notify maintenance team of deficiencies to be addressed
    - **Back office** support team ensures production configuration has executed as required.

6 DW/BI Governance

Governance activities should be completed and addressed during implementation. Specific governance deliverables can be added to the Software Development Lifecycle. Warehouse governance processes should be:

- Aligned with risk management
- Business driven as different business units have different needs
- Mitigate risks (not curtail execution)

Most critical functions:

- Those that govern the business operated discovery or refinement area
  - Handshaking
  - Instantiate data

    - Transfer data
    - Discard data
    - Data archival and time horizons for boundaries to prevent sprawl
- Those that ensure quality within the warehouse
    - Assign time, resources and programs to remediate data
- One-off events: part of lifecycle but curtail them in the pilot area
- Policies required for procedures in real-time environment
- Risk exposure mitigation matrix. Curtail risk with governance functions:

6.1 Enabling Business Acceptance

Sign-off by business is part of User Acceptance Testing, and is paramount for every DW/BI implementation. Critically important architectural sub-components:

- **Conceptual Data Model:** Key business concepts and how they are related to each other
- **Data quality feedback loop:**
    - How data issues are identified and remediated
    - Owners of the systems where they originated informed and held accountable for fixing them
    - How are issues caused by the DW data integration process remediated?
- **End-to-end Metadata:** How does the architecture support integrated end-to-end flow of Metadata? Is access to meaning and context part of the architecture?
- **End-to-end verifiable data lineage:** Is a system of record identified for all data?

6.2 Customer / User Satisfaction

Collecting, understanding and acting on customer feedback can be facilitated through regular meetings with user representatives.

6.3 Service Level Agreements

Specify business and technical requirements.

6.4 Reporting Strategy

Reporting strategy across the BI portfolio must address:

- Standards, processes, guidelines, best practices and procedures
- Security access to sensitive data elements to only entitled users
- Access mechanisms to describe how users want to interact, report, examine or view data
- User community type and appropriate tool to consume it with
- Nature of reports (summary, detailed, exception) and frequency, timing, distribution, storage formats
- Use of visualisation capabilities
- Trade-offs between timeliness and performance

Evaluate regularly to see if they are still providing value.

Data Source governance, monitoring and control are vital.

Centre of Excellence can empower business users to the self-service model.

6.5 Metrics

- **Usage Metrics:**
    - Number of registered, connected or concurrent users

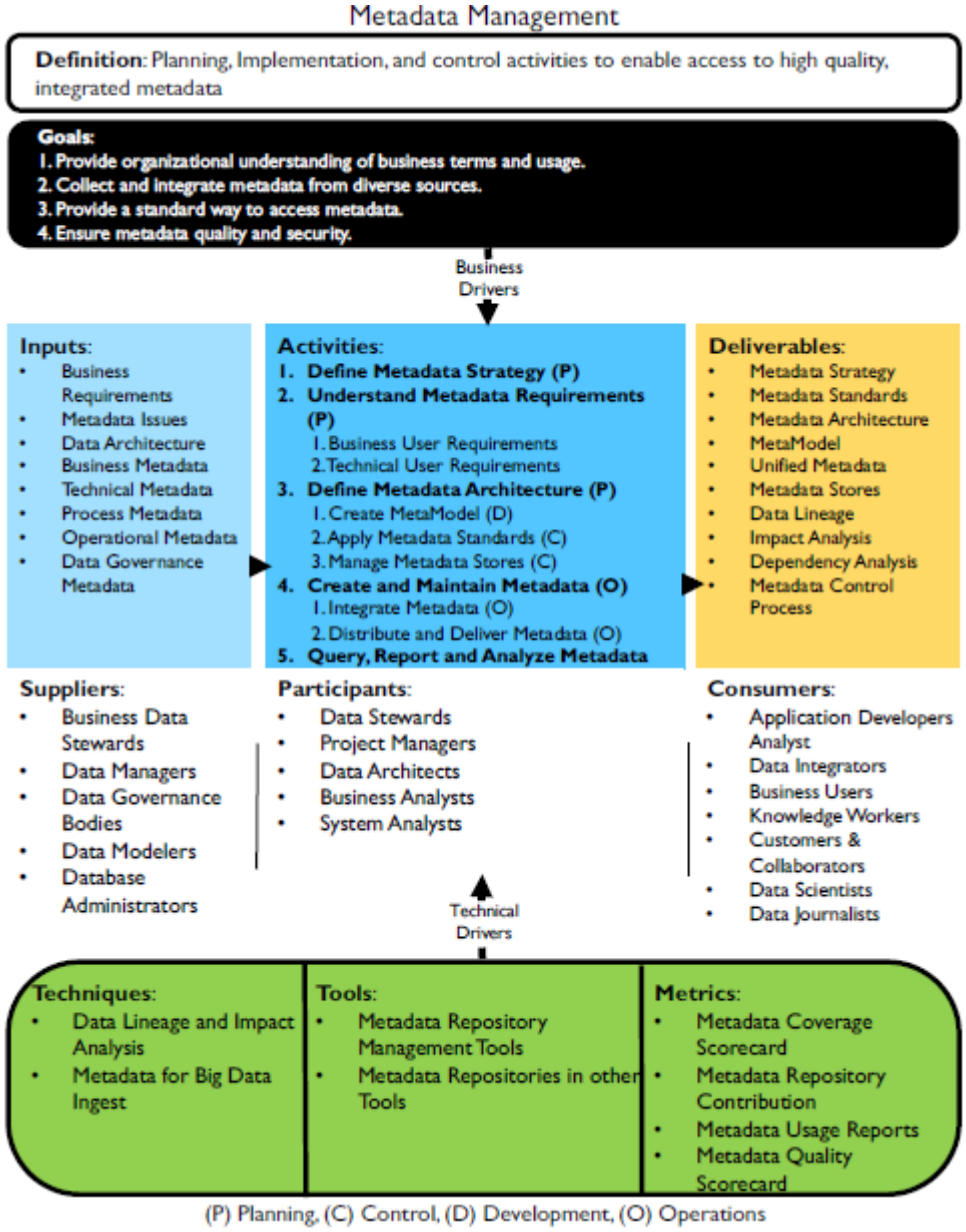
    - Number of user accounts for each tool
    - Number of queries per user community per timeframe
- **Subject area coverage percentages**
    - How much of the warehouse is being accessed by each department?
    - Mapping operational sources to targets
- **Response and performance metrics**
    - Response times can be retrieved from query tools
    - Load times for each data product in raw format
    - Query records, data refresh and data extract times for objects provided to users
    - Use the metrics to validate or adjust service levels

Metadata Management

1 Introduction

Metadata helps an organisation understand its data, systems and workflows. Metadata includes information about:

- Technical and business processes
- Data rules and constraints
- Logical and physical data structures
- Describes the data (databases, data elements, data models)
- Concepts the data represents
- Relationships between data and concepts



1.1 Business Drivers

Data cannot be managed without Metadata, which must also be managed. Metadata helps:

- Increase confidence in data by providing context, and measurement of data quality
- Increase value of strategic information (Master data) by enabling multiple uses
- Operational efficiency by identifying redundant data and processes
- Prevent the use of out of date or incorrect data
- Reduce data-oriented research time
- Improve communication between business and IT
- Create accurate impact analysis
- Reduce system development lifecycle time
- Support regulatory compliance

1.2 Goals and Principles

Goals:
1. Provide organizational understanding of business terms and usage.
2. Collect and integrate metadata from diverse sources.
3. Provide a standard way to access metadata.
4. Ensure metadata quality and security.

Implementation of a Metadata solution depends on the following principles:

- **Organisational commitment:** senior manager support. Metadata management is an enterprise program
- **Strategy:** How Metadata will be created, maintained, integrated and accessed. Metadata strategy must align with business
- **Enterprise perspective:** To ensure future extensibility but implement iteratively
- **Socialisation:** Communicate the necessity and purpose of each type of Metadata
- **Access:** Ensure staff members know how to access and use Metadata
- **Quality:** Process owners accountable
- **Audit:** Set, enforce and audit standards for Metadata
- **Improvement:** Feedback mechanism

1.3 Essential Concepts

1.3.1 Metadata vs. Data

Organisations should define Metadata requirements and what they need it for. What is data and what is Metadata depends on the organisation

Metadata is the "Who, What, Where, Why, When & How" of Data

Who	What	Where	Why	When	How
Who created this data?	What is the business definition of this data element?	Where is this data stored?	Why are we storing this data?	When was this data created?	How is this data formatted? (character, numeric, etc.)
Who is the Steward of this data?	What are the business rules for this data?	Where did this data come from?	What is its usage & purpose?	When was this data last updated?	How many databases or data sources store this data?
Who is using this data?	What is the security level or privacy level of this data?	Where is this data used & shared?	What are the business drivers for using this data?	How long should it be stored?	
Who "owns" this data?	What is the abbreviation or acronym for this data element?	Where is the backup for this data?		When does it need to be purged/deleted?	
Who is regulating or auditing this data?	What are the technical naming standards for database implementation?	Are there regional privacy or security policies that regulate this data?			

Global Data Strategy, Ltd. 2017 7

1.3.2 Types of Metadata

(NB If an exam question has any other type, refer to the DMBOK V1 notes, and please notify me)

Types:

- **Business Metadata:** Focus on the content and condition of data, and includes details related to Data Governance. Examples:
  - Definitions and descriptions of data sets
  - Business rules, transformation rules, calculations
  - Data models
  - Data quality rules
  - Update schedules
  - Provenance and data lineage
  - Data standards
  - Stakeholder contact details
  - Security/privacy level
  - Data usage notes
- **Technical Metadata:** Provides information about technical details of data, systems that store it and the processes that move it. Examples:
  - Physical database table and column names
  - Column properties
  - Database object properties
  - Access permissions
  - Data CRUD rules
  - Physical data models
  - Relationships between data models and physical assets
  - ETL job details
- **Operational Metadata:** Describes details of the processing and accessing of data. Examples:
  - Logs of job execution of batch jobs
  - History of extracts and results
  - Error logs

- Schedule anomalies
- Results of audit, balance, control measurements
- Reports and query access patterns, frequency, and execution time
- Patches and Version maintenance plan and execution, current patching level
- Backup, retention, date created, disaster recovery provisions
- SLA requirements and provisions
- Volumetric and usage patterns
- Data archiving and retention rules, related archives
- Data sharing rules and agreements
- Purge criteria
- Technical roles and responsibilities

1.3.3 ISO / IEC Metadata Registry Standard

ISO /IEC 11179 is structured in 6 parts:

- Part 1: Framework for the generation and standardisation of Data Elements
- Part 3: Basic Attributes of Data Elements
- Part 4: Rules and Guidelines for the Formulation of Data Elements
- Part 5: Naming and Identification Principles for Data Elements
- Part 6: Registration of Data Elements

1.3.4 Metadata for Unstructured data

Types of Metadata for unstructured data (Stored with the document):

- **Descriptive:** Catalogue information and thesauri keywords
- **Structural:** tags, field structures, format, terms on the Business Glossary
- **Administrative:** Sources, update schedules, access rights, navigation information
- **Bibliographic:** Library catalogue entries
- **Record keeping Metadata:** Retention policies
- **Preservation Metadata:** Storage, archival condition, rules for conservation, e-discovery

1.3.5 Sources of Metadata

- **Application Metadata Repositories:** The physical tables where Metadata is stored
- **Business Glossary:**
  - A document of the organisation's business concepts and terminology, definitions and the relationships between those terms. Accounts for hardware, software, database, processes and different user roles and responsibilities:
    - **Business users:** Data analysts, research analysts, management to understand terminology
    - **Data Stewards:** Manage the lifecycle of terms and definitions
    - **Technical users:** Use glossary terms to make design decisions
  - Business glossary should capture business terms attributes such as:
    - Term name, definition
    - Ownership and stewardship
- **Business Intelligence (BI) Tools:** Overview information, classes, objects, derived and calculated items
- **Configuration Management Tools:** CMDB manage Metadata related to IT assets.
- **Data Dictionaries:** Defines the contents and structure of data sets. This Metadata is embedded in database/modelling tools, and must be extracted to use

- **Data Integration Tools:** Lineage Metadata, movement of data
- **Database Management and System Catalogues:** Describe the content, sizing information, software version, deployment status, network and infrastructure uptime, availability etc.
- **Data Mapping Management Tools:** Mapping tools and data integration tools can exchange data with Metadata repositories
- **Data Quality Tools:** Can share quality scores and patterns with Metadata repositories
- **Directories and Catalogues:** Contains information about systems, sources and locations of data in the organisation. It is useful for developers and super users.
- **Event Messaging Tools:** Require a lot of Metadata to move messages between diverse systems.
- **Modelling Tools and Repositories:** Produce Metadata relevant to the design of the application or system model.
- **Service Registries:** Technical information about services and service end points e.g. APIs
- **Other Metadata Stores:** Specialised lists, repositories of repositories and business rules.

1.3.6 Types of Metadata Architecture

All architectural layers should point to the Metadata lifecycle:

- Metadata creation and sourcing
- Metadata storage in one or more repositories
- Metadata integration: Bring together Metadata from various repositories
- Metadata delivery: Formats for delivery to another system
- Metadata usage: Searching
- Metadata control and management

All require standards to be in place and enforced.

1.3.6.1 Centralised Metadata Architecture

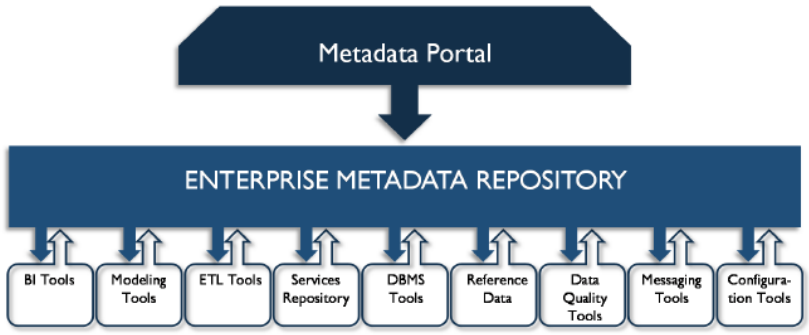


Figure 85 Centralized Metadata Architecture

A single Metadata repository containing copies of Metadata from the various sources.

Advantages of a centralised repository:

- High availability – independent of source systems
- Quick retrieval as repository and query reside together
- Resolved database structure not affected by proprietary nature of third party systems
- Quality is improved as extracted Metadata may be enhanced with Metadata from elsewhere

Limitations of a centralised repository:

- Complex processes ensure changes to source Metadata are quickly replicated into the repository
- Maintenance can be costly
- Extraction may require custom modules or middleware
- Increased demands on IT and vendor staff to maintain customised code

1.3.6.2 Distributed Metadata Architecture

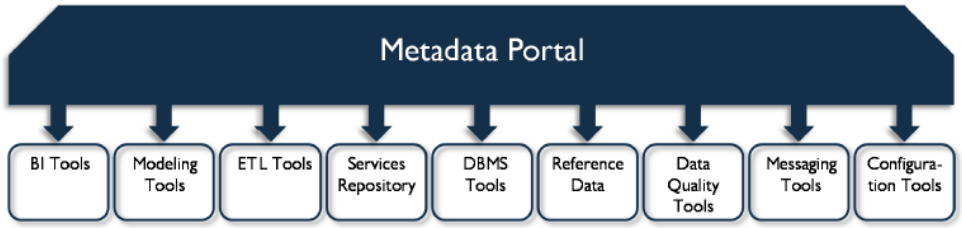


Figure 86 Distributed Metadata Architecture

No persistent repository, the portal passes user requests to the appropriate tool to execute. The Metadata retrieval engine retrieves data from source systems in real time. The Metadata management environment maintains source system catalogs and lookup information.

Advantages of the distributed Metadata architecture:

- Metadata is current and valid as it is retrieved from the source
- Queries are distributed – improved process and response time
- Implementation and maintenance effort minimised as Metadata requests from proprietary systems are queries, so no understanding of proprietary data structures is required
- Automated Metadata query processing is simpler
- Reduced batch processing

Limitations of distributed architecture:

- Not able to support user-added Metadata entries as there is no repository to put them
- Standardisation of presenting Metadata from various systems
- Query capabilities depend on the availability of the source systems
- Quality of Metadata depends on the source systems

1.3.6.3 Hybrid Metadata Architecture

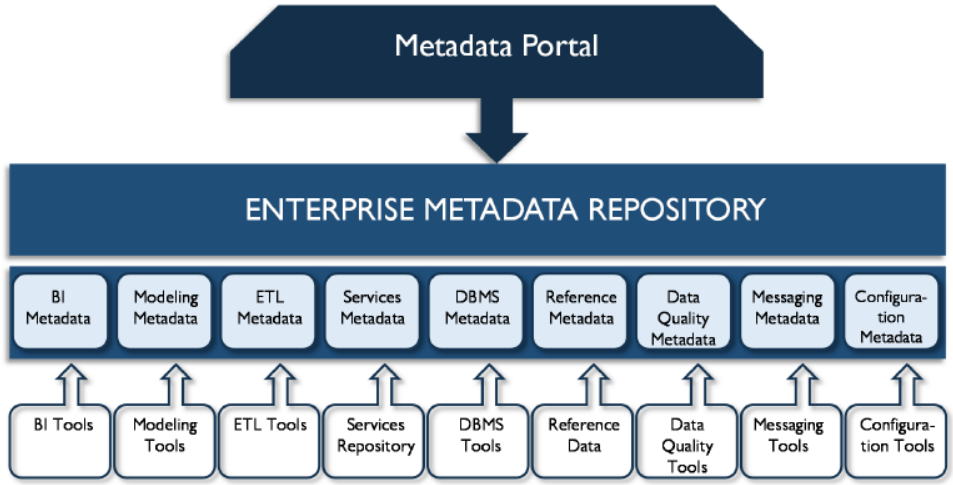


Figure 87 Hybrid Metadata Architecture

The repository design only accounts for the user-added Metadata, critical standardised items and the additions from manual sources. Near real-time retrieval of Metadata from source and enhanced Metadata when needed.

1.3.6.4 Bi-Directional Metadata Architecture

Allows Metadata to change in any part of the architecture and then feedback is coordinated from the repository to the original source. The repository is forced to contain the latest version of the Metadata source, and must manage changes to the source as well.

2 Activities

2.1 Define Metadata Strategy

Define the future state enterprise Metadata architecture and the implementation phases:

- **Initiate Metadata strategy planning:** Key stakeholders involved in planning
  - short- and long-term goals
  - Charter, scope, objectives and communications plan
- **Conduct key stakeholder interviews:** Foundation knowledge
- **Assess existing Metadata sources and information architecture:** Assess difficulty of project
  - Interview IT staff
  - Review system architecture and model documentation
- **Develop future Metadata architecture:** Develop long term architecture for the managed Metadata environment
- **Develop a phased implementation plan:** Prioritise findings from the interviews and data analyses to define a phased implementation plan.

2.2 Understand Metadata Requirements

What Metadata is needed and at what level. Functionality-focussed requirements:

- **Volatility:** Update frequency
- **Synchronisation:** Timing of updates in relation to source changes
- **History:** Do historical versions need to be retained?
- **Access rights:** Who, how and interface functionality

- **Structure:** Metadata model for storage
- **Integration:** Degree of and rules for integration
- **Maintenance:** Processes and rules for updating
- **Management:** Roles and responsibilities
- **Quality:** Metadata quality requirement
- **Security:** Some Metadata cannot be exposed as it will reveal sensitive data

2.3 Define Metadata Architecture

Metadata Management System must be able extract Metadata from many sources by scanning the sources and updating the repository, while supporting manual updates, searches and lookups by various user groups. Single access point for the Metadata repository which is transparent to users.

2.3.1 Create MetaModel

Data model for the Metadata repository:

- High-level conceptual model – relationships between systems
- Lower level metamodel that describes the elements and processes
- The metamodel is a planning tool, and Metadata itself

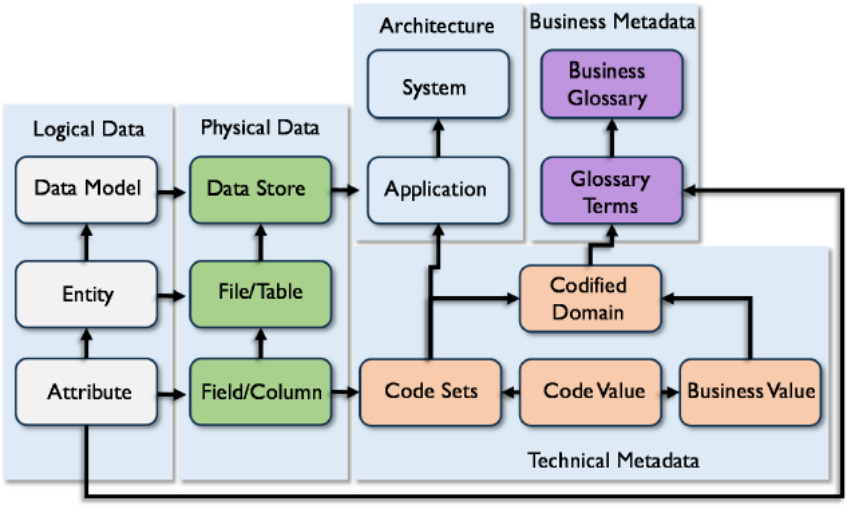
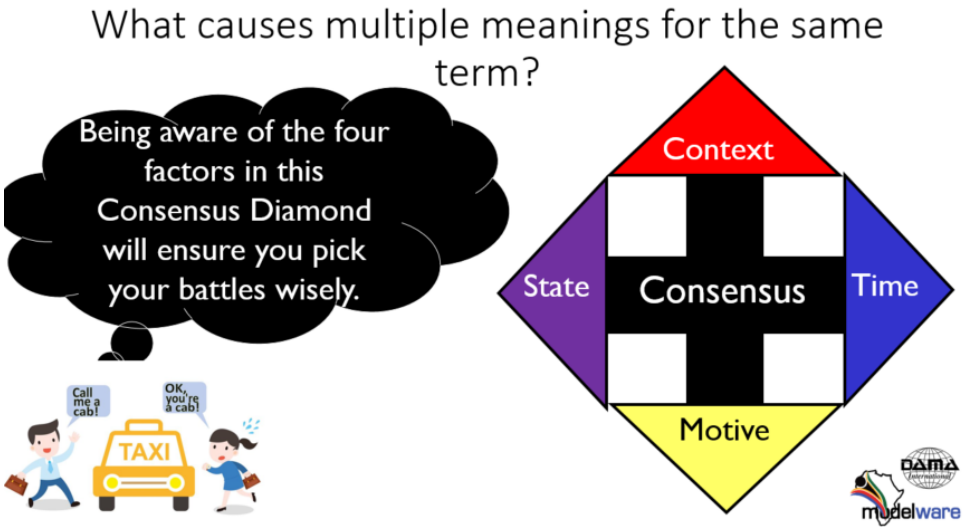


Figure 88 Example Metadata Repository Metamodel

2.3.2 Apply Metadata Standards

Governance monitors for compliance:

- Internal standards such as naming conventions (ISO 11179 for naming conventions)
- External standards such as data exchange formats



2.3.3 Manage Metadata Stores

Control activities should have governance oversight and include:

- Job scheduling and monitoring
- Load statistical analysis
- Backup, recovery, archive, purging
- Configuration modifications
- Performance tuning
- Query statistics analysis
- Query and report generation
- Security management

Quality control activities:

- QA, quality control
- Matching update sets to timeframes
- Missing Metadata reports

Metadata management activities include:

- Loading, scanning, importing and tagging assets
- Source mapping and movement
- Versioning
- User interface management
- Linking data sets Metadata maintenance – for NoSQL provisioning
- Linking data sets to internal data acquisition – custom links and job Metadata
- Licensing for external data sources and feeds
- Data enhancement Metadata e.g. Link to GIS

Training:

- Education and training of users and data stewards
- Management metrics generation and analysis
- Training on the control activities and query and reporting

2.4 Create and Maintain Metadata

Metadata should be planned and created as a product. Profile and inspect for quality. Schedule maintenance.

General principles of Metadata management:

Accountability: Metadata is often produced through existing processes – hold those process owners accountable for quality of Metadata.

Standards: Set, enforce and audit Metadata standards

Improvement: Create a consumer feedback mechanism

2.4.1 Integrate Metadata

As Metadata is gathered from many sources, and integrated into the Metadata repository, challenges arise which require governance.

Two approaches for repository scanning:

- **Proprietary interface:** Collection and loading of Metadata occurs in single step. No format specific file output.
- **Semi-proprietary interface:** Two step process where scanner collects Metadata from a source and outputs it to a format specific data file

Files used during the scanning process:

- **Control file:** Contains the source structure of the data model
- **Reuse file:** Contains the rules for managing reuse of process tools
- **Log file:** Produced during each phase of the process
- **Temporary and backup files:** Used for traceability

2.4.2 Distribute and Deliver Metadata

Metadata is delivered to consumers/applications/tools requiring Metadata feeds:

- Metadata intranet websites
- Reports, glossaries and other documents
- Data warehouses, data marts and BI tools
- Modelling and software development tools
- Messaging and transactions
- Web services and APIs
- External organisation interface solutions

Metadata is exchanged with external organisations using files (flat, XML or JSON) or web services.

2.5 Query, Report and Analyse Metadata

A Metadata repository has a front-end application for search and retrieval as Metadata guides the use of data assets

3 Tools

The Metadata repository is the primary tool used to manage Metadata. It has an integration layer and an interface for manual updates. Metadata repository management tools are also a source of Metadata

4 Techniques

4.1 Data Lineage and Impact Analysis

Metadata about the physical assets provides information about how the data is transformed as it moves between systems. Limited to the scope of the Metadata management system.

- **'As Implemented Lineage'**:
  - Current version of lineage based on programming code.
  - Imported from various tools
- **'As Designed Lineage'**:
  - Lineage described in mapping specification documents.
  - Not extractable by Metadata management system
- **Stitching**: The process whereby the Metadata management system augments the 'As Implemented' data lineage with the 'As Designed'.

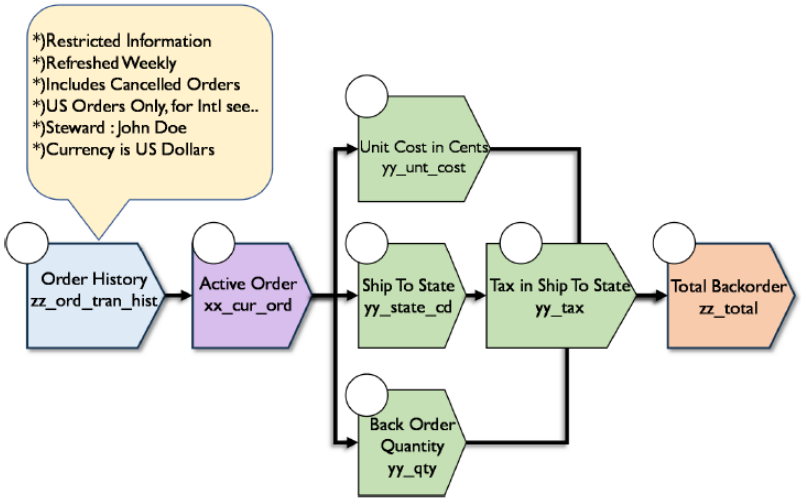


Figure 89 Sample Data Element Lineage Flow Diagram

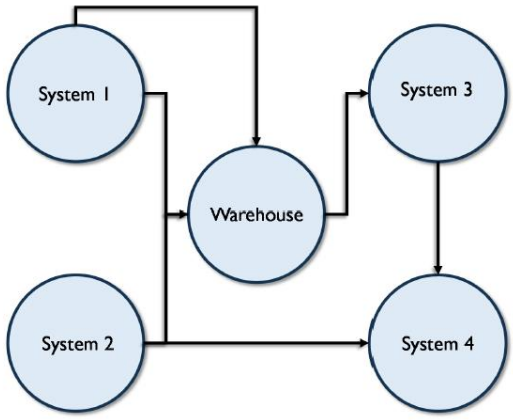


Figure 90 Sample System Lineage Flow Diagram

Successful lineage discovery needs to account for both business and technical focus:

- **Business focus:** Data elements prioritised by business
  - Start at target and trace back to source
  - Gives business understanding what happens to a data element as it moves
  - DQ measurements with lineage can pinpoint where system design impacts quality.
- **Technical focus:**

- Start at source systems
- Identify immediate consumers, then next sets of consumers until all systems are identified.

4.2 Metadata for Big Data Ingest

Metadata tags should be applied to data on ingestion to the data lake. Ingestion engines can profile data as well.

5 Implementation Guidelines

Implement the Metadata environment in incremental steps to minimise risks and facilitate acceptance. Use a relational database platform. Contents should be generic in design and should be integrated so that consumers can see across different data sources. Should house current, planned and historical versions of Metadata.

5.1 Readiness assessment

People should be aware of the risks of not managing Metadata:

- Errors in judgement due to lack of knowledge of the context of data
- Exposure of sensitive data
- Risk that SMEs will leave and take their knowledge of the data with them

A formal assessment of the current maturity of Metadata activities includes:

- Critical data elements
- Available Metadata glossaries
- Lineage
- Data profiling and data quality processes
- MDM maturity

5.2 Organisational and Cultural Change

Metadata efforts often meet with resistance. Needs senior management support and engagement. Business and technical staff work closely in a cross-functional manner.

6 Metadata Governance

Determine the specific requirements for the management of the Metadata lifecycle, and establish governance processes. Formal roles and responsibilities need to be assigned to dedicated resources.

6.1 Process Controls

Governance team responsible for:

- Defining standards
- Managing status changes for Metadata
- Promotion of Metadata
- Training
- Management of business terms

Metadata strategy should be integrated into the SDLC to ensure Metadata is collected and remains current.

6.2 Documentation of Metadata Solutions

A master catalogue of Metadata of the sources and targets currently on scope. It is a 'what-is-where' guide for the user community and includes:

- Metadata implementation status
- Source and target Metadata store
- Schedule information for updates
- Retention and versions kept
- Contents
- Quality statements or warnings
- System of record or other data source statuses
- Tools, architecture and people involved
- Sensitive information and removal or masking for the source

6.3 Metadata Standards and Guidelines

Metadata standards are required in the exchange of data with operational trading partners.

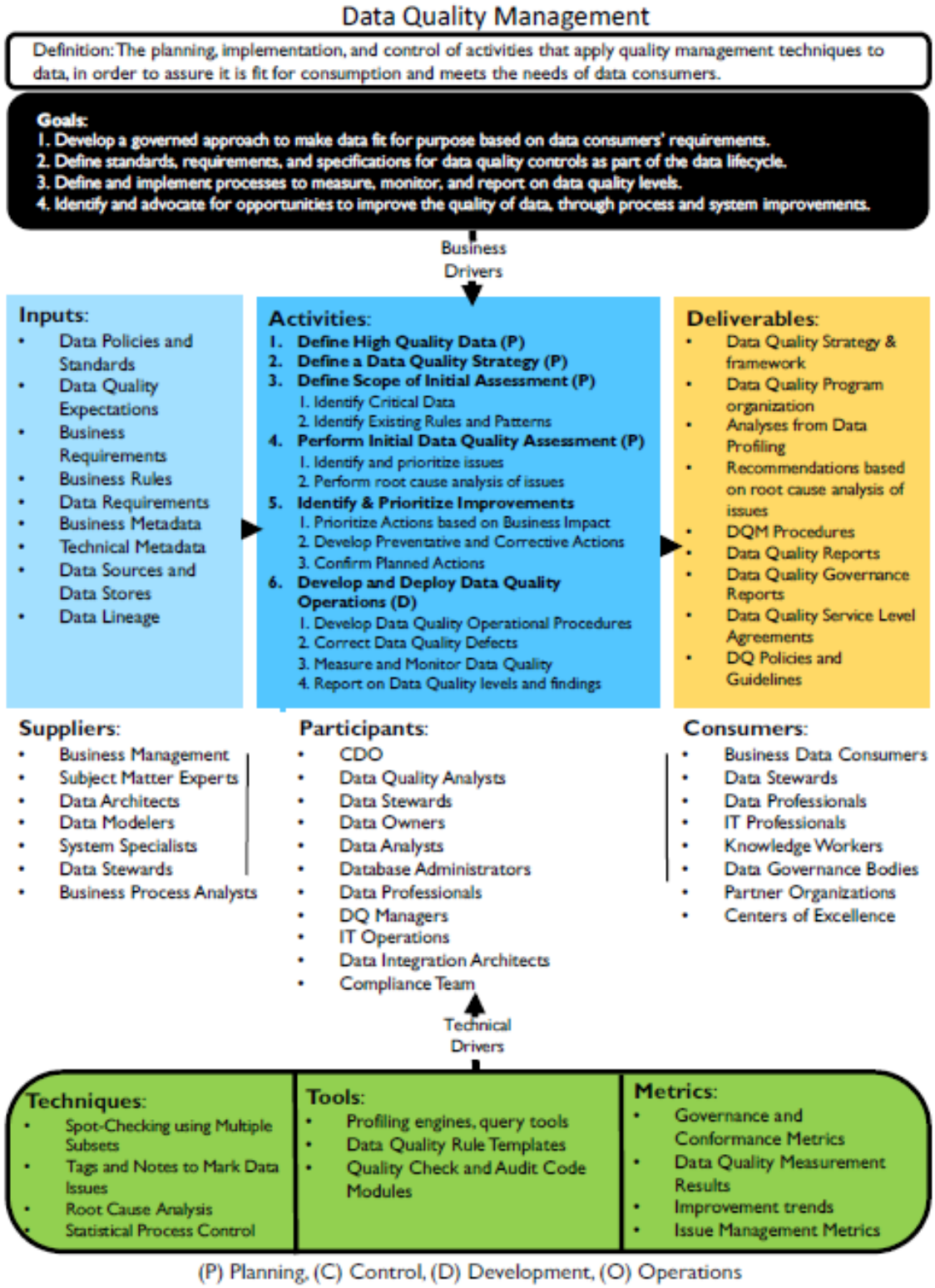
- Use industry-based Metadata standards early
- Tool vendors provide XML, JSON or REST support to exchange their data
- Tools offer import/export capabilities using XML
- Templates and examples
- ISO standards

6.4 Metrics

- **Metadata repository completeness:** Ideal coverage compared to actual coverage
- **Metadata Management Maturity:** Based on the Capability Maturity Model (CMM-DMM) assessment approach
- **Steward representation:** Coverage across enterprise for stewardship
- **Metadata usage:** User uptake measured in logins
- **Business Glossary activity:** Usage, update, resolution of definitions, coverage
- **Metadata documentation quality:** Assess automatically and manually
- **Metadata repository availability:** Uptime, processing time (batch and query)

Data Quality

1 Introduction



The value of data is that data is reliable and trustworthy i.e. of high quality.

Factors that contribute to poor quality data:

- Lack of understanding of the effects of poor quality data on the decision-making process
- Bad Planning
- Siloed system design
- Inconsistent development processes
- Incomplete documentation
- A lack of standards
- A lack of governance
- Failure to define what makes data fit for purpose

High quality data should be the goal of all data management disciplines and they all contribute to the quality of data. Data Quality should be managed by a Data Quality Program team as Data Quality is an enterprise program like Data Governance.

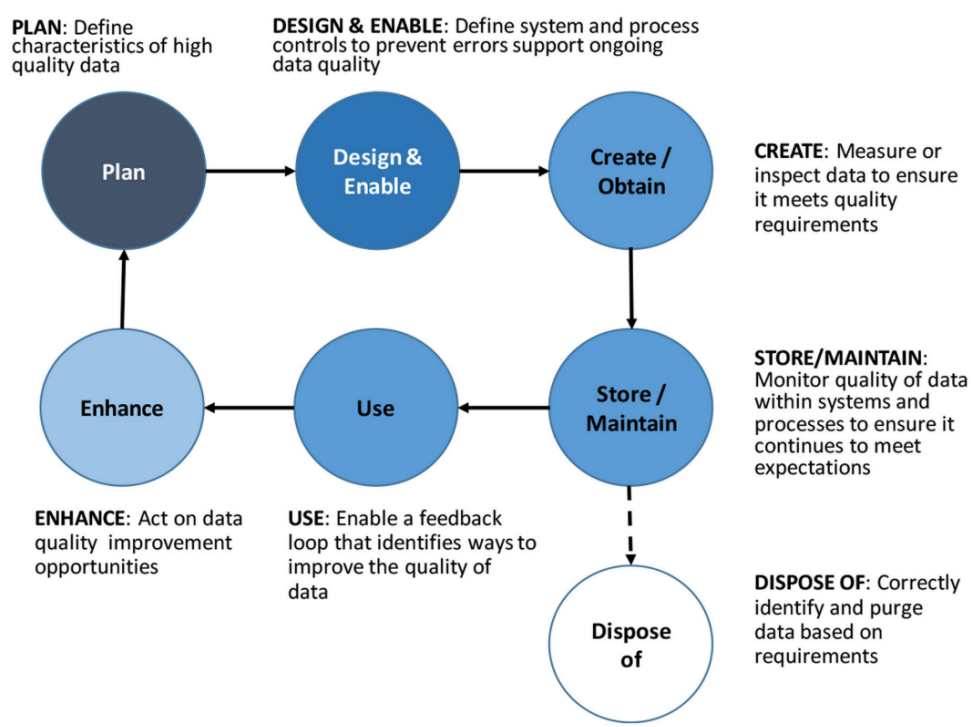


Figure 27: Data Quality Management and the Data Lifecycle (Adapted from DMBOK2, p. 29)

From Navigating the Labrynth

1.1 Business Drivers

Business drivers for establishing a Data Quality Program:

- Increase value of business data and opportunities to use it
- Reduce risks associated with poor quality data
- Improve organisational efficiency and productivity
- Protect and enhance the organisation's reputation

Direct costs associated with poor quality data include:

- Inability to invoice correctly

- Increased customer service calls and decreased ability to resolve them
- Revenue lost due to missed opportunities
- Delay if integration during mergers or acquisitions
- increased exposure to fraud
- loss due to bad business decisions driven by bad data
- Loss of business due to poor credit standing

1.2 Goals and Principles

General goals:

Goals:
1. Develop a governed approach to make data fit for purpose based on data consumers' requirements.
2. Define standards, requirements, and specifications for data quality controls as part of the data lifecycle.
3. Define and implement processes to measure, monitor, and report on data quality levels.
4. Identify and advocate for opportunities to improve the quality of data, through process and system improvements.

A Data Quality Program should be guided by the following principles:

- **Criticality:** Focus on data most critical to the enterprise and its customers
- **Lifecycle management:** Manage data across data lifecycle from creation through disposal.
- **Prevention:** Focus should be on prevention of data errors
- **Root cause remediation:** Don't just correct errors, address the root cause
- **Governance:** Support from Data Governance activities
- **Standards-driven**: Define requirements as measurable standards
- **Objective measurement and transparency:** Measurement and methodology must be communicated to stakeholders
- **Embedded in business processes:** Business process owners must enforce data quality standards
- **Systematically enforced:** System owners must enforce data quality requirements
- **Connected to service levels:** Data quality reporting and issues management must be incorporated into SLAs.

1.3 Essential Concepts

1.3.1 Data Quality

Data is of high quality when it meets the expectations and needs of data consumers, i.e. it is fit for the purpose to which they want to apply it.

1.3.2 Critical Data

How to identify critical data. It is usually required by:

- Regulatory reporting
- Financial reporting
- Business policy
- Ongoing operations
- Business strategy, especially efforts at competitive differentiation
- MASTERDATA is always critical

1.3.3 Data Quality Dimensions

A measurable characteristic of data which form the basis for measurable rules.

DAMA UK white paper (2013) – six core dimensions:

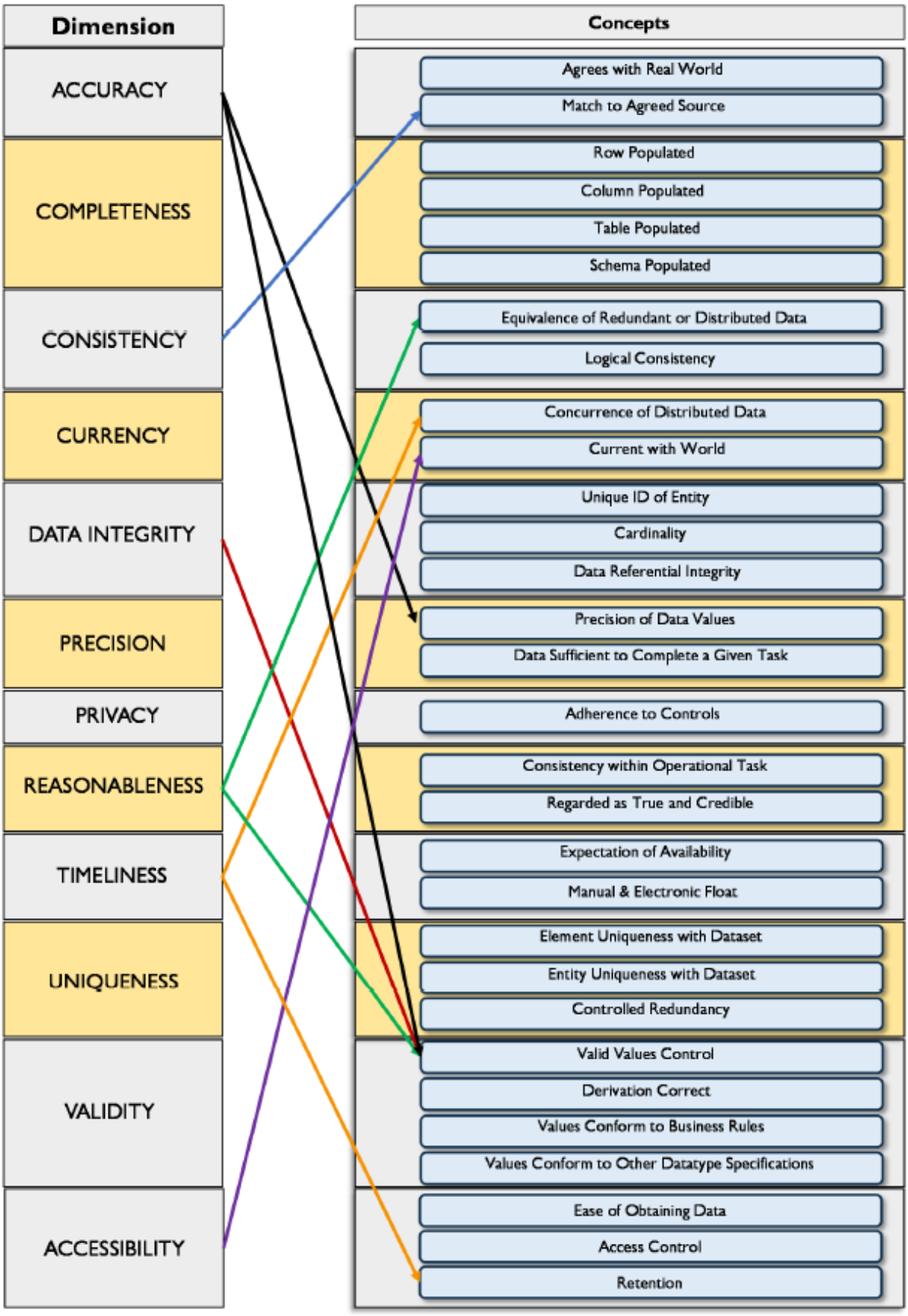
- **Completeness:** Proportion of data stored against 100%
- **Uniqueness:** no entity instance recorded more than once
- **Timeliness:** Degree to which data represents reality at any point in time
- **Validity:** Data conforms to the syntax of its definition
- **Accuracy:** degree to which data describes the real world
- **Consistency:** no difference found when comparing two or more representations of a thing to definitions

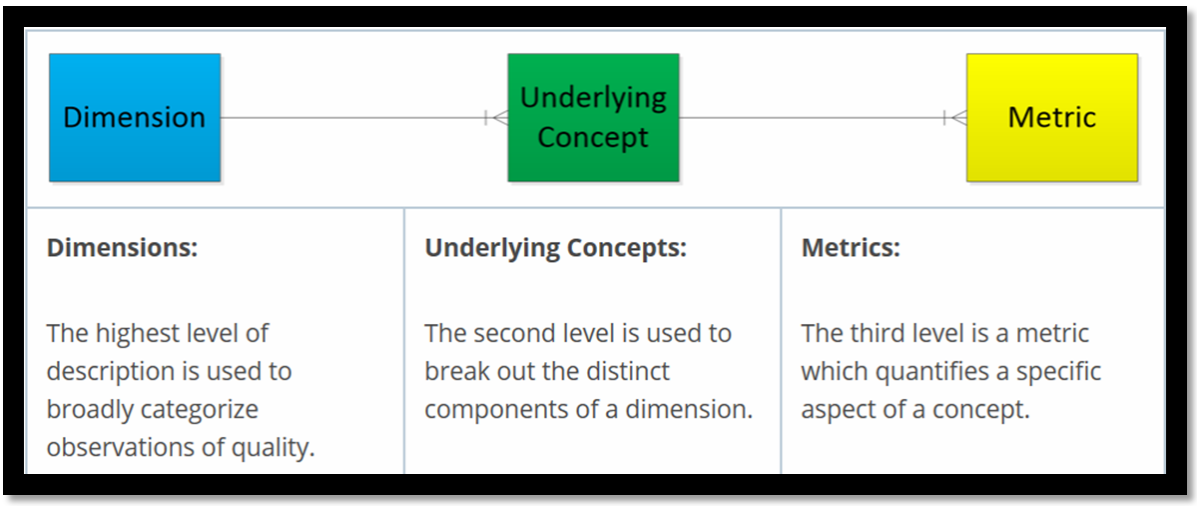
Other useful measurable characteristic (not listed as dimensions):

- **Usability:** Is the data understandable, simple, relevant, accessible, maintainable, and at the right level of precision?
- **Timing issues:** Is it stable yet responsive to legitimate change requests?
- **Flexibility:** Can it be repurposed? Is it easy to manipulate?
- **Confidence:** Are Data Governance processes in place? Is the data verified?
- **Value:** Good benefit case for the data? Is it being optimally used?

Dimension of Quality	Description
Accuracy	Accuracy refers to the degree that data correctly represents 'real-life' entities. Accuracy is difficult to measure, unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy rely on comparison to a data source that has been verified as accurate, such as a system of record or data from a reliable source (e.g., Dun and Bradstreet Reference Data).
Completeness	Completeness refers to whether all required data is present. Completeness can be measured at the data set, record, or column level. Does the data set contain all the records expected? Are records populated correctly? (Records with different statuses may have different expectations for completeness.) Are columns/attributes populated to the level expected? (Some columns are mandatory. Optional columns are populated only under specific conditions.) Assign completeness rules to a data set with varying levels of constraint: Mandatory attributes that require a value, data elements with conditional and optional values, and inapplicable attribute values. Data set level measurements may require comparison to a source of record or may be based on historical levels of population.
Consistency	Consistency can refer to ensuring that data values are consistently represented within a data set and between data sets, and consistently associated across data sets. It can also refer to the size and composition of data sets between systems or across time. Consistency may be defined between one set of attribute values and another attribute set within the same record (record-level consistency), between one set of attribute values and another attribute set in different records (cross-record consistency), or between one set of attribute values and the same attribute set within the same record at different points in time (temporal consistency). Consistency can also be used to refer to consistency of format. Take care not to confuse consistency with accuracy or correctness. Characteristics that are expected to be consistent within and across data sets can be used as the basis for standardizing data. Data Standardization refers to the conditioning of input data to ensure that data meets rules for content and format. Standardizing data enables more effective matching and facilitates consistent output. Encapsulate consistency constraints as a set of rules that specify consistent relationships between values of attributes, either across a record or message, or along all values of a single attribute (such as a range or list of valid values). For example, one might expect that the number of transactions each day does not exceed 105% of the running average number of transactions for the previous 30 days.
Integrity	Data Integrity (or Coherence) includes ideas associated with completeness, accuracy, and consistency. In data, integrity usually refers to either referential integrity (consistency between data objects via a reference key contained in both objects) or internal consistency within a data set such that there are no holes or missing parts. Data sets without integrity are seen as corrupted, or have data loss. Data sets without *referential* integrity have 'orphans' – invalid reference keys, or 'duplicates' – identical rows which may negatively affect aggregation functions. The level of orphan records can be measured as a raw count or as a percentage of the data set.
Reasonability	Reasonability asks whether a data pattern meets expectations. For example, whether a distribution of sales across a geographic area makes sense based on what is known about the customers in that area. Measurement of reasonability can take different forms. For example, reasonability may be based on comparison to benchmark data, or past instances of a similar data set (e.g., sales from the previous quarter). Some ideas about reasonability may be perceived as subjective. If this is the case, work with data consumers to articulate the basis of their expectations of data to formulate objective comparisons. Once benchmark measurements of reasonability are established, these can be used to objectively compare new instances of the same data set in order to detect change. (See Section 4.5.)
Timeliness	The concept of data Timeliness refers to several characteristics of data. Measures of timeliness need to be understood in terms of expected volatility – how frequently data is likely to change and for what reasons. Data currency is the measure of whether data values are the most up-to-date version of the information. Relatively static data, for example some Reference Data values like country codes, may remain current for a long period. Volatile data remains current for a short period. Some data, for example, stock prices on financial web pages, will often be shown with an as-of-time, so that data consumers understand the risk that the data has changed since it was recorded. During the day, while the markets are open, such data will be updated frequently. Once markets close, the data will remain unchanged, but will still be current, since the market itself is inactive. Latency measures the time between when the data was created and when it was made available for use. For example, overnight batch processing can give a latency of 1 day at 8am for data entered into the system during the prior day, but only one hour for data generated during the batch processing. (See Chapter 8.)
Uniqueness / Deduplication	Uniqueness states that no entity exists more than once within the data set. Asserting uniqueness of the entities within a data set implies that a key value relates to each unique entity, and only that specific entity, within the data set. Measure uniqueness by testing against key structure. (See Chapter 5.)
Validity	Validity refers to whether data values are consistent with a defined domain of values. A domain of values may be a defined set of valid values (such as in a reference table), a range of values, or value that can be determined via rules. The data type, format, and precision of expected values must be accounted for in defining the domain. Data may also only be valid for a specific length of time, for example data that is generated from RFID (radio frequency ID) or some scientific data sets. Validate data by comparing it to domain constraints. Keep in mind that data may be valid (i.e., it may meet domain requirements) and still not be accurate or correctly associated with particular records.

Relationship between Data Quality Dimensions and Data Quality Concepts:





1.4 Data Quality and Metadata

Metadata defines what data represents. A Metadata repository can house the results of data quality measurements so that they can be shared, and expectations clarified.

1.5 Data Quality ISO Standard

ISO 8000 is the international standard for data quality. ISO 8000 defines the characteristics that can be tested by any organisation in the data supply chain to objectively determine conformance.

ISO 8000 defines quality data as "portable data that meets stated requirements". Portable means that the data can be separated from a software application. Stated requirements must be clearly defined.

1.6 Data Quality Improvement Lifecycle

Approach data quality improvement based on the technique of quality improvement in physical products. Outputs from one process become inputs to other processes and can impact data quality. Use "plan-do-check-act" from problem solving model, the Shewhart/Deming Cycle.

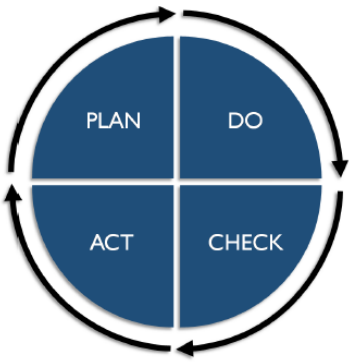


Figure 93 The Shewhart Chart

- **Plan Stage**: DQ team assesses scope, impact and priority of known issues, and evaluates alternatives to address them at the root cause.
- **Do stage**: DQ team leads efforts to address issues at the root cause and plans for ongoing monitoring of data
- **Check stage**: Actively monitor data against standards. If data falls below, act to bring it to acceptable levels
- **Act stage**: Activities to address emerging data issues. The cycle restarts.

1.7 Data Quality Business Rule Types

Data Quality Business Rules describe how data should exist in order to be useful within the organisation. Implemented in software or data entry templates.

Common business rule types:

- **Definitional conformance:** Data definitions used properly across organisation
- **Value presence and record completeness:** Rules for acceptability of missing values
- **Format compliance:** Values have a pattern e.g. phone numbers
- **Value domain membership:** exists in a defined data domain. (Master and reference data)
- **Range conformance:** within a defined range of values
- **Mapping conformance:** Maps to a domain (Reference data)
- **Consistency rules:** Maintain a relationship between two attributes based on the values
- **Accuracy verification:** Compare value to trusted source
- **Uniqueness verification:** Specify which entities must have unique representation. Primary key.
- **Timeliness validation:** Characteristics associated with expectations for accessibility and availability of data, and also when it was last updated.

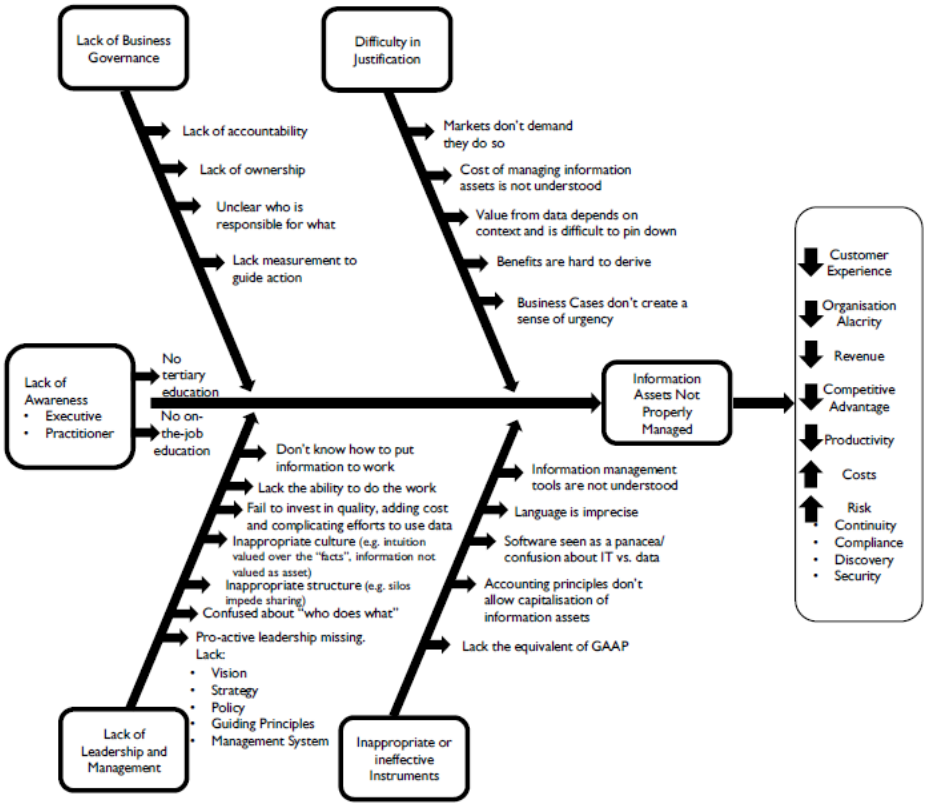
1.8 Common Causes of Data Quality Issues

Data quality issues can arise at any time in the lifecycle, and may have multiple causes: Data entry, data processing, system design and manual intervention in automated processes.

1.8.1 Issues caused by lack of leadership

Barriers to effective management of data quality:

- Lack of awareness on the part of leadership and staff
- Lack of business governance
- Lack of leadership and management
- Difficulty in justification of improvements
- Inappropriate or ineffective instruments to measure value



Barriers that slow/hinder/prevent companies from managing their information as a business asset
Most commonly observed root causes
Danette McGilvray / James Price / Tom Redman
October 2016

Work based on research by Dr. Nina Evans and James Price, see "Barriers to the Effective Deployment of Information Assets" at www.dataleaders.org

1.8.2 Issues caused by data entry process

- **Data entry interface issues**: Needs to have edits and controls
- **List entry placement**: Order of values in a dropdown list
- **Field overloading**: Reusing fields for different business purposes over time creates confusion
- **Training Issues**: Awareness of the impact of incorrect data. Incentivise for accuracy, not speed.
- **Changes to business processes**: New business rules and data quality requirements need to be incorporated into systems.
- **Inconsistent business process execution**: will produce inconsistent data

1.8.3 Issues caused by data processing functions

- **Incorrect assumptions about data sources**: Not enough known about data sources, details are missed
- **Stale business rules:** Business rules may change over time, and should be periodically reviewed
- **Changed data structures:** Source changes without informing downstream

1.8.4 Issues caused by system design

- **Failure to enforce referential integrity**:

    - Duplicate data that breaks uniqueness rules
    - Orphan rows, in some reports and not others, leading to different values for the same calculation
    - Inability to upgrade due to restored or changed referential integrity requirements
    - Inaccurate data due to missing data being assigned default values
- **Failure to enforce uniqueness constraints:** multiple copies result in overstated aggregation
- **Coding inaccuracies and gaps:** Data mapping or rules for processing incorrect
- **Data model inaccuracies:** assumptions in data model must be supported by actual data
- **Field overloading**
- **Temporal data mismatches:** Need a consolidated data dictionary
- **Weak Master Data Management:** Choose unreliable data sources
- **Data duplication:** Two main types:
    - **Single source – Multiple Local instances:** e.g. instances of the same customer in multiple tables
    - **Multiple sources – Single Instance:** Data instances with multiple authoritative sources

1.8.5 Issues caused by fixing issues

Manual data patches directly to database. Most change the data in place, and can only be undone by a database restore.

1.9 Data Profiling

Data profiling is a form of data analysis used to inspect the data and assess data quality. A profiling engine produces statistics that can be analysed to identify patterns in data content and structure.

- **Counts of NULLS:** Inspect to see if they are allowed
- **Max/Min value:** Identifies outliers
- **Max/Min Length:** Outliers in fields for specific length values
- **Frequency distribution** of values for individual columns
- **Data Type and Format:** Non-conformance or unexpected formats

1.10 Data Quality and Data Processing

1.10.1 Data Cleansing

Data cleansing or Scrubbing transforms data to conform to data standards or domain rules. Detecting and correcting errors. The need for data cleansing can be addressed by:

- Implementing controls to prevent data entry errors
- Correcting the data in the source system
- Improving the business processes that create data

1.10.2 Data enhancement

Data enhancement or enrichment is the process of adding attributes to data to increase its quality or usability.

- **Time/Date Stamp:**
- **Audit data:** adds lineage
- **Reference vocabularies**
- **Contextual information:**
- **Geographic information:** Address standardisation and geocoding

- **Demographic information:** customer data enhanced through demographic information
- **Psychographic information:** segment target populations by behaviours
- **Valuation information:** Assets

1.10.3 Data parsing and formatting

Data Parsing is the process of analysing data using pre-determined rules to define its content or value.

1.10.4 Data Transformation and Standardisation

Rule based transformations map data values in their original formats and patterns into target representation. Standardisation employs rules that capture context, linguistics and idioms recognised as common over time.

2 Activities

2.1 Define High Quality Data

High quality data is fit for the purposes of data consumers. Understand business needs, define terms, identify pain points and start to find consensus about drivers and data quality priorities. Ask questions of stakeholders to determine the understanding of the business benefits of high-quality data and the impact of poor quality data.

Understand the current state of data quality:

- Understand business strategy and goals
- Pain points, risks and business drivers
- Direct assessment of data (profiling)
- Documentation of data dependencies in business processes

Prioritise opportunities based on benefits to the organisation.

2.2 Define a Data Quality Strategy

Data quality priorities must align with business strategy. Develop a framework which includes methods to:

- Understand and prioritise business needs
- Identify data critical to business needs
- Define business rules and data quality standards based on business requirements
- Assess data against expectations
- Share findings and get feedback from shareholders
- Prioritise and manage issues
- Identify and prioritise opportunities for improvement
- Measure, monitor and report on data quality
- Manage Metadata produced through quality processes
- Integrate data quality controls into business and technical processes

2.3 Identify Critical Data and Business Rules

- Critical data
  - If it were higher quality would provide greater value
  - Regulatory requirements
  - Financial value

    - Direct impact to customers
    - Start with Master Data
    - Ranked list of data for DQ team
- Business Rules
    - Describe expectations about the quality of the data
    - Often not documented – need to reverse engineer
    - How data is collected or created
    - Measurements describe if data is fit for use
    - Discovery and refinement of rules is an ongoing process

2.4 Perform an Initial Data Quality Assessment

Query the data to understand content and relationships. Data stewards, SMEs, data consumers and DQ analysts prioritise findings.

The goal of the initial assessment is to learn about the data to make an actionable plan for improvement:

- Define the goals to drive the work
- Identify data to be assessed. Start small
- Identify uses of the data and consumers of the data
- Identify known risks of the data
- Inspect data based on known and proposed rules
- Document levels of non-conformance and types of issues
- Perform in depth analysis to prioritise issues and develop root cause hypotheses
- Confirm issues and priorities with stakeholders
- Planning

2.5 Identify and Prioritise Potential Improvements

- Prioritise actions based on business impact
- Develop preventative and corrective actions
- Confirm planned actions with stakeholders
- Large-scale profiling efforts should focus on the most critical data
- Profiling identifies issues, but not root causes

2.6 Define Goals for Data Quality Improvement

- Quick hits as well as long term strategic changes
- Address root causes
- Set achievable goals based on quantification of the business value of DQ improvements
- Determine ROI of fixes of issues based on:
    - Criticality of the data
    - Amount of affected data
    - Age of the data
    - Number and type of business processes impacted
    - No of stakeholders affected
    - Associated risks
    - Cost of root cause remediation
    - Cost of work arounds

2.7 Develop and Deploy Data Quality Operations

2.7.1 Manage Data Quality rules

Define rules up front to:

- Set clear expectations
- Provide requirements for edits and controls to prevent data issues from being introduced
- Provide DQ requirements to vendors
- Foundation for DQ measuring

DQ rules should be managed as Metadata and should be:

- **Documented consistently:** Templates
- **Defined in terms of DQ Dimensions:** Help people understand what is being measured
- **Tied to business impact:** Connect standards and rules to organisational success
- **Backed by data analysis:** Test rules on actual data
- **Confirmed by SMEs**
- **Accessible to all data consumers**

2.7.2 Measure and monitor Data Quality

Two reasons to implement operational data quality measurements:

- To inform consumers about levels of quality
- Manage risk that may be introduced by technical or business changes

Express as a percentage where (r) is the rule being tested.

$$ValidDQI(r) = \frac{(TestExecutions(r) - ExceptionsFound(r))}{TestExecutions(r)}$$

$$InvalidDQI(r) = \frac{(ExceptionsFound(r))}{TestExecutions(r)}$$

Example:

10000 tests of business rule (r) found 560 exceptions

Therefore: Valid DQ = (10000-560)/10000 = 94.4%

Invalid DQ = 560/10000 = 5.6%

Organise results in a table as shown below:

Table 30 DQ Metric Examples

Dimension and Business Rule	Measure	Metrics	Status Indicator
Completeness Business Rule 1: Population of field is mandatory	Count the number of records where data is populated, compare to the total number of records	Divide the obtained number of records where data is populated by the total number of records in the table or database and multiply it by 100 to get to percentage complete	Unacceptable: Below 80% populated Above 20% not populated
Example 1: Postal Code must be populated in the address table	Count populated: 700,000 Count not populated: 300,000 Total count: 1,000,000	Positive measure: 700,000/1,000,000*100 = 70% populated Negative measure: 300,000/1,000,000 *100 = 30% not populated	Example result: Unacceptable
Uniqueness Business Rule 2: There should be only one record per entity instance in a table	Count the number of duplicate records identified; report on the percentage of records that represent duplicates	Divide the number of duplicate records by the total number of records in the table or database and multiply it by 100	Unacceptable: Above 0%
Example 2: There should be one and only one current row per postal code on the Postal Codes master list	Count of duplicates: 1,000 Total Count: 1,000,000	10,000/1,000,000*100 = 1.0% of postal codes are present on more than one current row	Example result: Unacceptable
Timeliness Business Rule 3: Records must arrive within a scheduled timeframe	Count the number of records failing to arrive on time from a data service for business transactions to be completed	Divide the number of incomplete transactions by the total number of attempted transactions in a time period and multiply by 100	Unacceptable: Below 99% completed on time Above 1% not completed on time
Example 3: Equity market record should arrive within 5 minutes of being transacted	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000	Positive: (1,000,000 – 2000) / 1,000,000*100 = 99.8% of transaction records arrived within defined timeframe Negative: 2000/1,000,000*100 = 0.20% of transactions did not arrive within defined timeframe	Example Result: Acceptable
Validity Business Rule 4: If field X = value 1, then field Y must = value 1-prime	Count the number of records where the rule is met	Divide the number of records that meet the condition by the total number of records	Unacceptable : Below 100% adherence to the rule
Example 4: Only shipped orders should be billed	Count of records where status for shipping = Shipped and status for billing = Billed: 999,000 Count of total records: 1,000,000	Positive: 999,000/1,000,000*100 = 99.9% of records conform to the rule Negative: (1,000,000-999,000) / 1,000,000 *100 = 0.10% do not conform to the rule	Example Result: Unacceptable

Measurements can be taken at three levels of granularity:

Table 31 Data Quality Monitoring Techniques

Granularity	In-stream (In-Process Flow) Treatment	Batch Treatment
Data Element	Edit checks in application Data element validation services Specially programmed applications	Direct queries Data profiling or analyzer tool
Data Record	Edit checks in application Data record validation services Specially programmed applications	Direct queries Data profiling or analyzer tool
Data set	Inspection inserted between processing stages	Direct queries Data profiling or analyzer tool

2.7.3 Develop Operational Procedures for Managing Data Issues

The DQ Team must respond to issues timeously and effectively. Design and implement operational procedures for:

- **Diagnosing issues:** Root cause analysis requires input from technical and business SMEs
    - Review issues in context of processing flows to isolate point where flaw is introduced
    - Evaluate whether there been environment changes that could have caused errors
    - Evaluate whether other process issues contributed
    - Determine whether there are issues with external data that could have affected this data
- **Formulating options for remediation:** Based on diagnosis, evaluate alternatives for addressing the issue:
    - Non-technical such as lack of training, leadership support, accountability
    - Modification of systems to eliminate root causes
    - Developing controls to prevent issue
    - Additional inspecting and monitoring
    - Directly correcting flawed data
    - Take no action based on cost and impact of correction versus the value of data correction.
- **Resolving issues:** DQ team and business data owners determine the best way to solve issues

An incident tracking system should be used:

- Standardise data quality issues and activities
- Provide an assignment process for data issues
- Manage issue escalation procedures
- Manage data quality resolution workflow

2.7.4 Establish Data Quality Service Level Agreements

A DQ SLA specifies the organisation's expectation for response and remediation for DQ issues in each system. The SLA defines roles and responsibilities associated with performance DQ procedures.

SLA establishes time limits for notification generation, the names in the management chain and when escalation should occur.

SLA reporting can be scheduled or driven by business.

2.7.5 Develop Data Quality Reporting

Reporting should focus around:

- DQ Scorecard
- DQ trends
- SLA metrics
- DQ Issue management focussing on the status of issues and resolutions
- Conformance of the DQ team to governance policies
- Conformance of IT and business teams to DQ policies
- Positive effects of improvement projects

3 Tools

- **Data Profiling Tools:**
    - Produce high level statistics to identify patterns in the data
    - Perform initial assessment of DQ

    - Enable assessment of large data sets
- **Data Querying Tools:**
    - Query more deeply to answer questions raised by profiling
- **Modelling and ETL Tools:**
    - Can be detrimental to data if used without knowledge of the data
    - Can improve quality if used with the data in mind
- **Data Quality Rule Templates**
- **Metadata Repositories:** Definitions of high quality data is Metadata

4 Techniques

4.1 Preventative Actions

Ways to prevent poor quality data entering an organisation:

- **Establish data entry controls:** Data entry rules
- **Train data producers:** Value accuracy rather than speed
- **Define and enforce rules:** Create a 'data firewall', a table with all the business rules, to check quality before the data is used
- **Demand high quality data from data suppliers:** Examine processes to check structures, data sources and provenance
- **Implement data governance and stewardship:**
- **Institute formal change control:** Ensure changes are tested before implementing

4.2 Corrective Actions

Perform data correction in three general ways:

- **Automated correction:**
    - Rule based standardisation, normalisation and correction
    - No manual intervention
    - Requires environment with well-defined standards, rules and known error patterns.
- **Manually-driven correction:**
    - Use automated tools with manual review before committing modified values to persistent storage.
    - Environments where data sets require human oversight e.g. MDM
- **Manual correction:**
    - Best done through an interface with controls and edits which produces an audit trail.
    - Avoid making manual corrections directly to the production environment

4.3 Quality Check and Audit Code Modules

4.4 Effective Data Quality Metrics

Characteristics of informative metrics:

- Measurability
- Quantifiable within a discrete range
- Business relevance
- Correlate with the influence of the data on the key business expectations
- Acceptability
- Data meets business expectation based on acceptability thresholds
- Accountability/Stewardship

- Understood and approved by key stakeholders
- Controllability
- Should trigger an action to improve data
- Trending
- Measure DQ over time

4.5 Statistical Process Control

SPC is a method to manage processes by analysing measurements of variation in process inputs, outputs or steps.

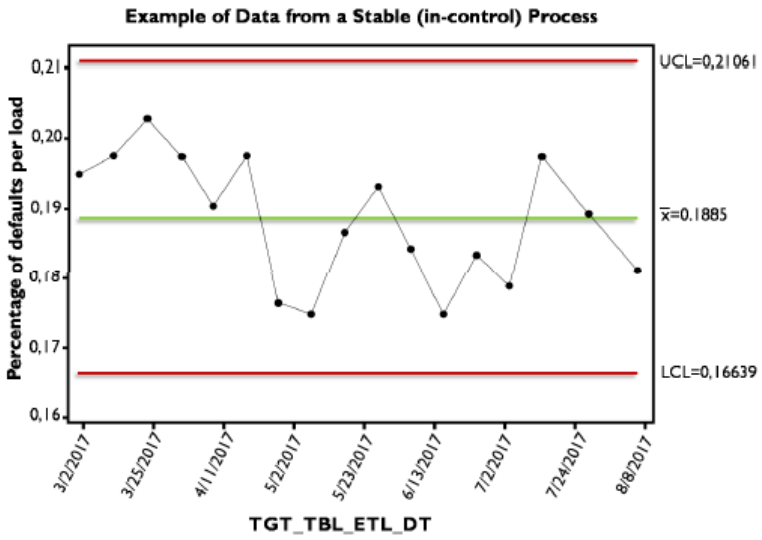
A process is a series of steps to turn inputs to outputs. SPC is based on the assumption that when a process with consistent inputs is executed consistently it will produce consistent outputs.

Uses measures of central tendency (mean, median, mode) and variability around a central value (range, variance, standard deviation) to establish tolerances for variances within a process

Two types of variance:

- Common causes:
  - Inherent in the process
  - Process is in statistical control when only common causes, and range of variation (baseline) is established.
- Special causes: Unpredictable or intermittent

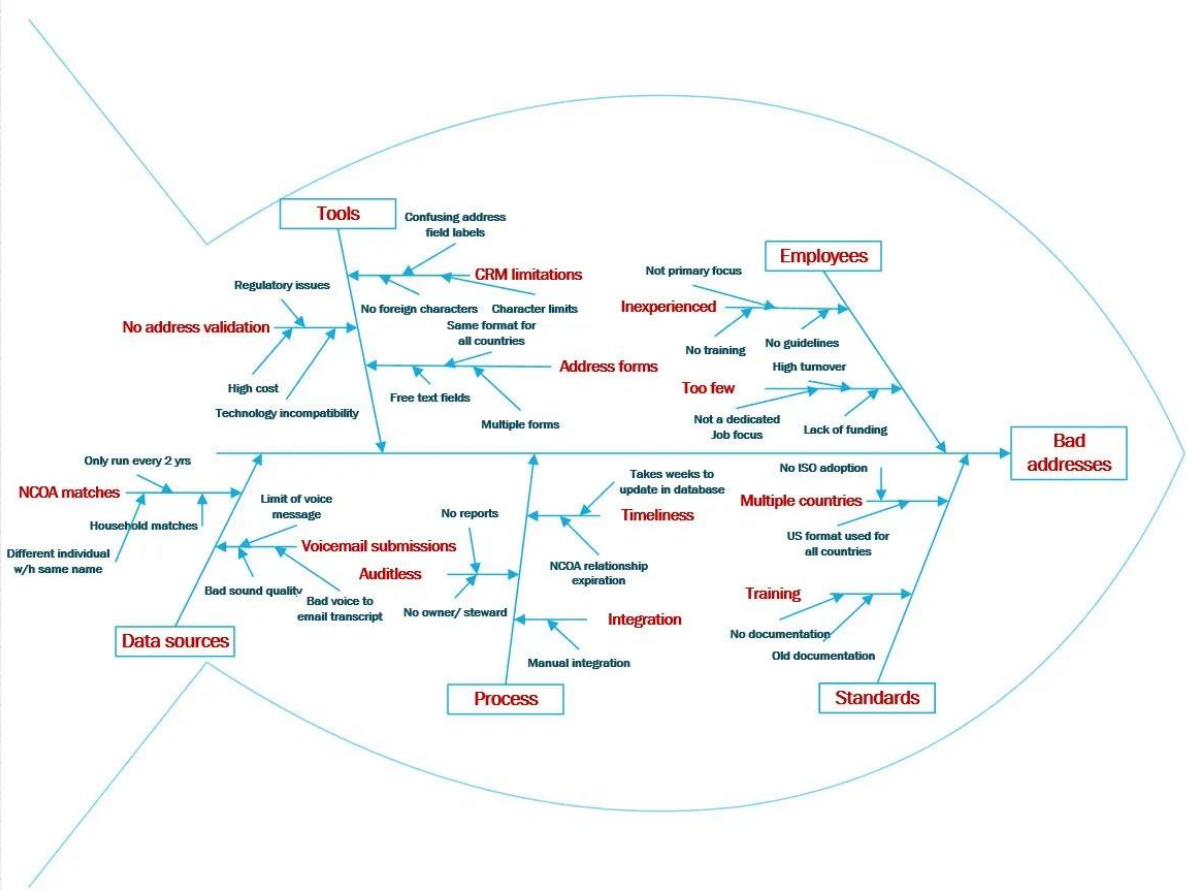
SPC is used for control, detection and improvement. Early detection of unexpected variation simplifies root cause investigation.



4.6 Root cause Analysis

Common techniques:

- Pareto analysis (the 80/20 rule)
- Fishbone diagram analysis
- Track and trace
- Process analysis
- Five Whys (McGilvray, 2008)



The Ishikawa / Fish Bone Diagram

5 Implementation Guidelines

Improving DQ in an organisation requires changing how the people think and behave towards data. A DQ implementation needs to plan for:

- **Metrics on the value of data and the cost of poor data:** To raise organisational awareness and funding for improvements
- **Operating mode for IT/Business interactions:** Business people know how important the data is and IT can translate definitions of data quality into queries that identify records which don't comply
- **Changes in how projects are executed:** Identify issues early and build data quality expectations into projects
- **Changes to business processes:** DQ team assesses processes and recommends changes
- **Funding for remediation and improvement projects:** Document costs and benefits of fixing data so that it can be prioritised
- **Funding for DQ operations:** Ongoing monitoring, reporting and fixing DQ issues

5.1 Readiness Assessment/Risk Assessment

Consider the following to assess the organisational readiness to accept DQ practices:

- **Management commitment to managing data as a strategic asset:**
  - How much do they know about data as an asset, risks of poor-quality data, importance of data governance?
- **The organisation's current understanding of the quality of its data:**

    - Pain points – helps identify and prioritise improvement projects
- **The actual state of the data:**
    - Profiling, analysis and quantification of known pain points
- **Risks associated with data creation, processing or use:**
    - Identify what can go wrong, and the potential damage to the organisation
- **Cultural and technical readiness for scalable data quality monitoring:**
    - Requires a good collaborative relationship between IT and business teams

5.2 Organisation and Cultural Change

Promote awareness of the role and importance of data in the organisation.

All employees raise DQ issues, ask for good quality data as consumers, and provide good quality data to others. Every person who touches the data can impact its quality.

Employees need to think and act differently to produce and manage better quality data. This requires training focussing on:

- Common causes of data problems
- Relationships and why DQ requires an enterprise approach
- Consequences of poor quality data
- Necessity for ongoing improvement
- Becoming data lingual
- Introduce process changes

6 Data Quality and Data Governance

A data quality program is more effective when part of a data governance program.

6.1 Data Quality Policy

All Data Management Knowledge Areas require a data policy, especially if they touch on regulatory areas:

- Purpose, scope and applicability of the policy
- Definition of terms
- Responsibilities of the Data Quality program
- Responsibilities of other stakeholders
- Reporting
- Implementation of the policy, including links to risk, preventative measures, compliance, data protection and data security.

6.2 Metrics

High level categories of DQ metrics:

- **Return on investment:**
- **Levels of quality:**
- **Data Quality trends:**
- **Data issue management metrics:**
- **Conformance to service levels:**
- **Data Quality plan rollout:**

Big Data and Data Science

1 Introduction

Generate, store and analyse larger amounts of data:

- **Big Data:** Volume, variety and velocity it is produced
- **Data Scientists:** People who mine and develop predictive, machine learning and prescriptive models and analytics
- **Data Science:** Applies data mining, statistical analysis and machine learning with data integration and data modelling to predict behaviours. Forward Looking. BI describes past trends – rear-window view.

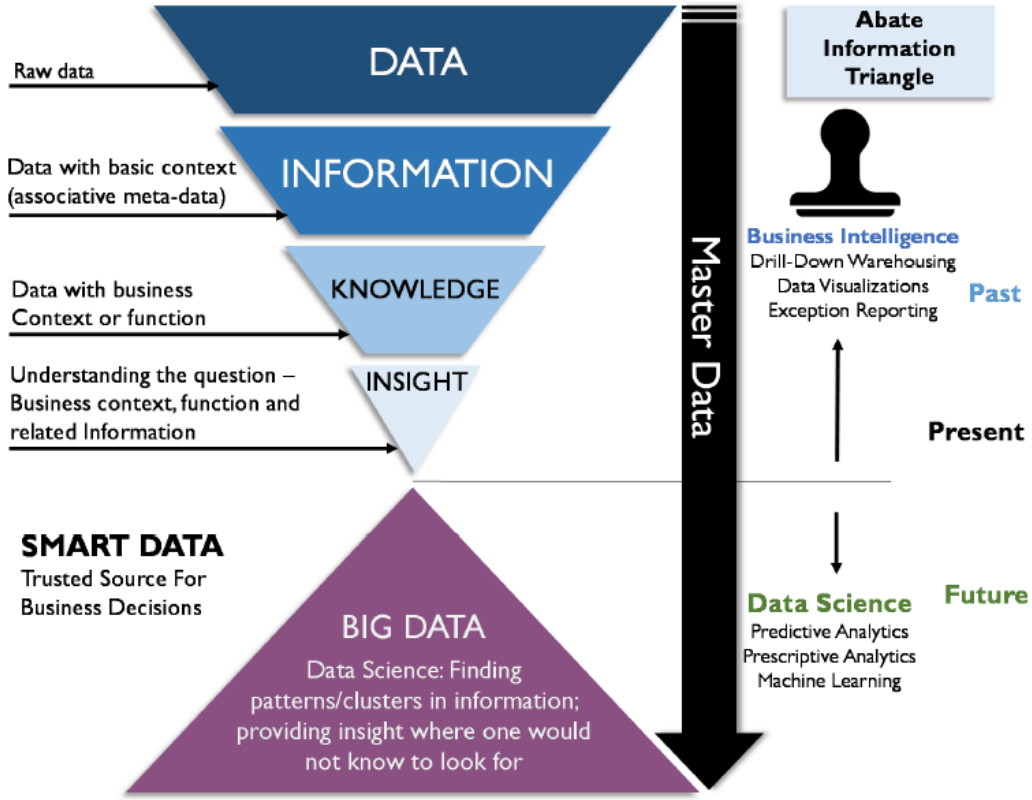


Figure 96 Abate Information Triangle

Different management and storage. Big Data relies on ELT – Loading and then transforming – and is not stored in the relational model. Speed and volume of data requires different approaches to Integration, Metadata and Data Quality management.

1.1 Business Drivers

- The desire to discover and act on business opportunities through applying techniques on diversely generated data.
- Data Science can improve operations
- Machine learning for automation of complex time-consuming activities

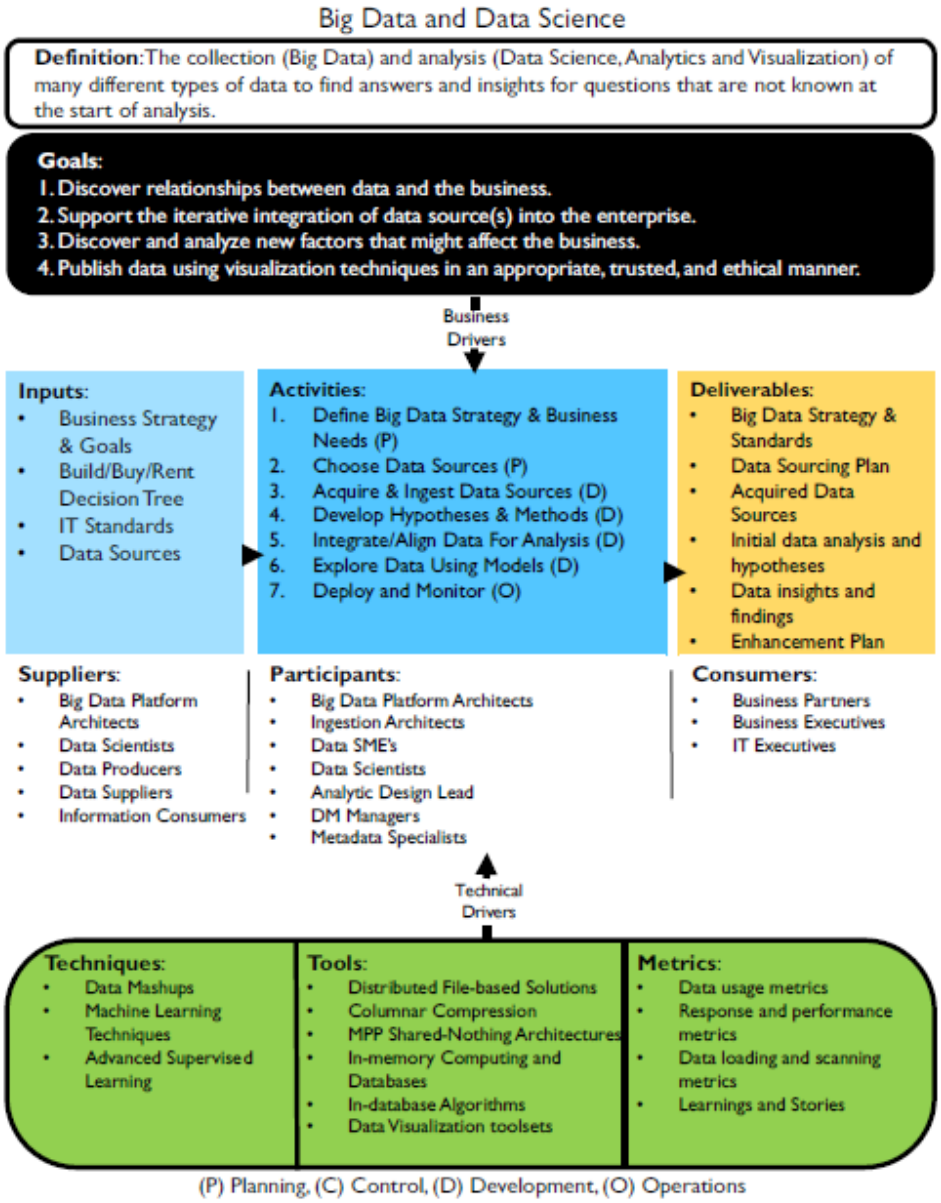


Figure 97 Context Diagram: Big Data and Data Science

1.2 Principles

Important to manage Metadata related to Big Data sources for inventory, their origins and value.

1.3 Essential Concepts

1.3.1 Data Science

Developing predictive models that explore data content patterns. Based on hypothesis which may be statistically confirmed by historical data. Iterative inclusion of data sources into models that develop insights.

Data science depends on:

- **Rich data sources:** Potential to show otherwise invisible patterns in organisational or customer behaviour.
- **Information alignment and analysis:** Techniques to understand data content and combine data sets for meaningful patterns

- **Information delivery:** Visualisations of results of insight gained from mathematical models
- **Presentation of findings and data insights:** Presentation and sharing of insights

Table 32 Analytics Progression

DW / Traditional BI	Data Science	
Descriptive	Predictive	Prescriptive
Hindsight	Insight	Foresight
Based on history: What happened? Why did it happen?	Based on predictive models: What is likely to happen?	Based on scenarios: What should we do to make things happen?

1.3.2 The Data Science Process

Data Science follows the scientific method of refining knowledge by making observations and testing hypotheses, observing results and formulating theories to explain results. The output of each phase in the diagram is input to the next.

Figure 98 Data Science Process

- **Define Big Data strategy and business needs:** Requirements that identify desired outcomes with measurable benefits
- **Choose data sources:** Identify gaps in current data base and find sources
- **Acquire and ingest data stores:**
- **Develop Data Science hypotheses and methods:**
- **Integrate and align data for analysis:** Data integration and cleansing for quality (Wrangling and Munging)
- **Explore data using models:**
  - Apply statistical analysis and machine learning algorithms
  - Validate, train and evolve model
  - New hypotheses may be introduced
- **Deploy and monitor:** often becomes warehouse

1.3.3 Big Data

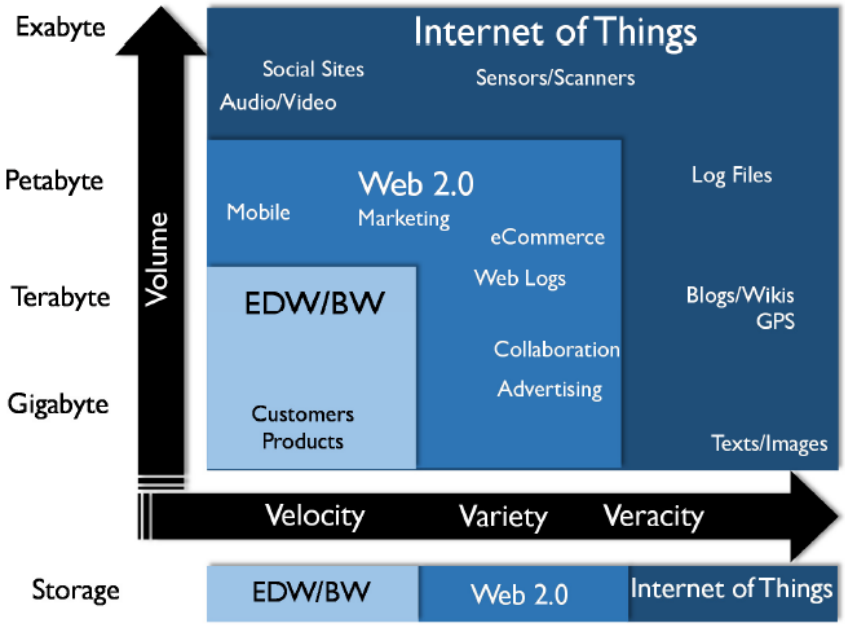


Figure 99 Data Storage Challenges[88]

- **Volume:** Amount – billions of records - >100 Terabytes
- **Velocity:** Speed at which data is captured, generated or shared – often analysed at real-time
- **Variety / Variability:** Forms data is captured or delivered
- **Viscosity:** How difficult the data is to use and integrate
- **Volatility:** How often the data changes
- **Veracity:** How trustworthy the data is

1.3.4 Big Data Architecture Components

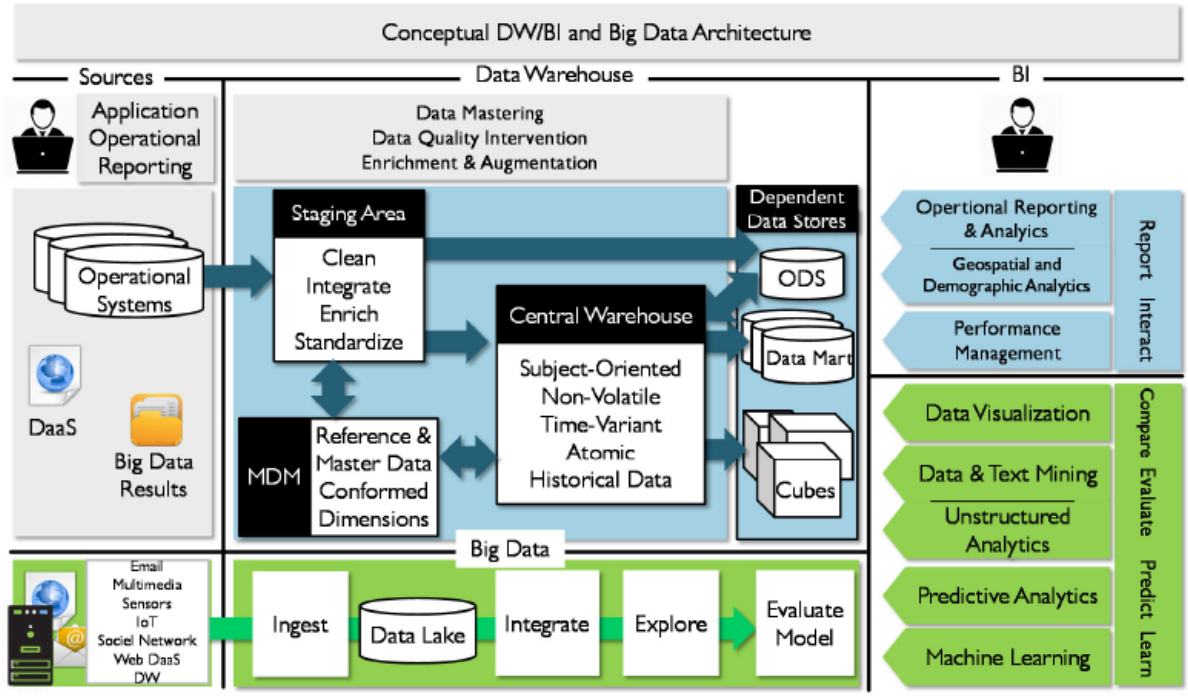


Figure 100 Conceptual DW/BI and Big Data Architecture

1.3.5 Sources of Big Data

Internet of Things, email, social media, online orders, video games, phones, POS devices, surveillance systems, sensors, medical monitoring, satellites, military etc.

1.3.6 Data Lake

An environment where a vast amount of data of various types and structures can be ingested, stored, assessed and analysed:

- Environment for Data Scientists to mine and analyse data
- Central storage for raw data with minimal transformation
- Alternate storage for detailed historical data warehouse data
- Online archive for records
- Environment to ingest streaming data with automated pattern identification

Manage Metadata to prevent it becoming a data swamp.

1.3.7 Services-Based Architecture

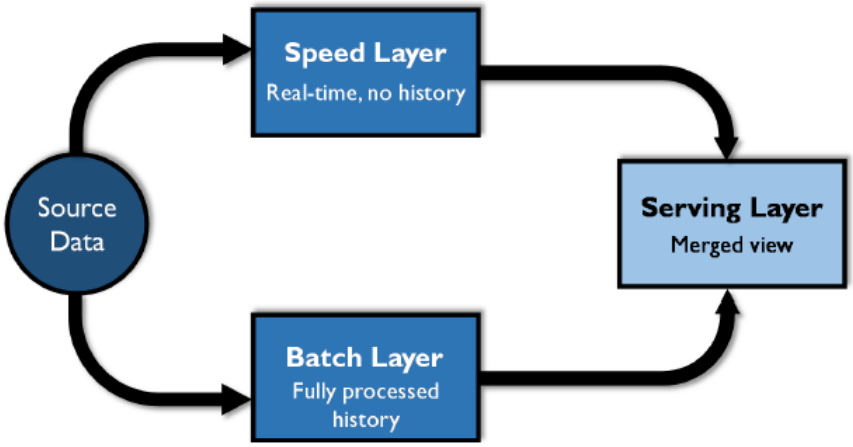


Figure 101 Services-based Architecture

Also called Lambda Architecture, referred to in Ch 5 p181, CAP / Brewers Theorem.

SBA provides immediate (maybe not complete or accurate) data. Updates complete, accurate historical data set using same source

SBA has 3 main components:

- **Batch layer:** A data lake, historical, structure-over-time component, every transaction is an *insert*.
- **Speed layer:** Only real-time data, Operational Data Store (ODS), all transactions are *updates* or inserts
- **Serving layer:** Interface to join batch and speed layers, uses Metadata to determine where to "serve" the data

Data is loaded into both Batch and Speed layers simultaneously, creating a current state, and a history layer.

1.3.8 Machine Learning

The construction and study of learning algorithms. Subfield of Artificial Intelligence. Three types:

- **Supervised learning:** Based on generalised rules e.g. separating SPAM emails

- **Unsupervised Learning:** Identifying hidden patterns i.e. data mining
- **Reinforcement learning:** Based on achieving a goal e.g. beating a chess opponent

Ethical implications:

- Transparency:
  - It is not clear how Deep Learning Neural Networks (DLNN) learn
  - Need to see how decisions are made

1.3.9 Sentiment Analysis

Looking for key words in semi-structured data using Natural Language Processing (NLP) to look for sentiment. Requires understanding of the meaning of a post.

1.3.10 Data and Text Mining

Data Mining is an offshoot of machine learning that reveals patterns in data using algorithms. Unsupervised learning where algorithms are applied to a data set without knowledge of outcomes, to reveal patterns and relationships.

Text mining analyses documents with text analysis and data mining techniques to classify content automatically into workflow guided and SME-directed ontologies.

Data and text mining techniques:

- **Profiling:** Characterise behaviour norms of an individual, group or population for anomaly detection (e.g. fraud)
- **Data reduction:** Make a similar smaller data set from a large one for easier analysis.
- **Association:** Unsupervised learning process to find relationships based on transactions
- **Clustering:** Group elements by shared characteristics
- **Self-organising maps:** Neural network method of cluster analysis by reducing dimensionality in the evaluation space. Also called Kohonen Maps or topologically ordered maps.

1.3.11 Predictive Analytics

Subfield of supervised learning. The development of probability models based on variables related to possible events. When it receives some information, the model triggers a response by the organisation. A forecast is a simple predictive model.

1.3.12 Prescriptive Analytics

Prescriptive analytics anticipates what will happen, when it will happen and implies why it will happen. Shows implications of various decisions, and can suggest how to take advantage of an opportunity or avoid a risk.

1.3.13 Unstructured Data Analytics

An iterative process of scanning and tagging to add hooks to unstructured data, to allow filtering and linking to related structured data.

1.3.14 Operational Analytics

BI or streaming analytics.

1.3.15 Data Visualisation

The process of interpreting concepts, ideas and facts by using pictures of graphical representations.

1.3.16 Data Mashups

Combine data and internet-based services to create visualisation for insight and analysis

2 Activities

2.1 Define Big Data Strategy and Business Needs

Must be aligned with overall business strategy:

- What problems does the organisation need analytics to solve?
- What data sources to use or acquire?
- The timeliness and scope of the data to provision
- The impact on and relation to other data structures
- Influences to existing modelled data

2.2 Choose Data Sources

Big Data comes from many internal and external sources. Evaluate the Quality and reliability, and know its origin (provenance), format, what elements represent, how it connects to other data, and how frequently it will be updated.

Review available data sources, processes that create the data, and manage the plan for new sources.

- **Foundational data:** e.g. POS in sales analysis
- **Granularity:** Most granular form is ideal
- **Consistency:** Select data that appears appropriately and consistently across visualisations
- **Reliability:** Use trusted, authoritative sources
- **Inspect / profile new sources:** Test changes before adding new data sets

Risks:

- Privacy
- Filters which may introduce bias

2.3 Acquire and Ingest Data Sources

Capture critical Metadata about the source (origin, size, currency and additional knowledge about the content) when ingesting data sources into the Big Data environment. Ingestion engines may profile the data. Provides information on how to integrate with other data sets, such as Master Data or historical warehouse data, also how to train and validate models.

2.4 Develop Data Hypotheses and Methods

Data science entails building statistical models that find correlations and trends within and between data sets to find insights within the data. Models should be tested for a range of outcomes. Models depend on the quality of input data.

2.5 Integrate / Align Data for Analysis

Preparing data for analysis involves understanding what is in the data and links between data from various sources.

- Daily data would have to be aggregated to link to monthly data.
- Common key
- Similarity and record linking algorithms

- Clustering used to determine groupings

2.6 Explore Data using Models

- **Populate the Predictive Model**
- **Train the Model:** Repeated runs of the model against the data resulting in changes to the model
- **Evaluate the Model:**
  - Evaluated and validated against training sets
  - Ethical training to remove the biases of the creators
- **Create Data Visualisations:** Ensure the visualisation addresses the audience.

2.7 Deploy and Monitor

- Expose Insights and Findings
- Integrate with additional Data Sources

3 Tools

3.1 MPP (Massively Parallel Processing) Shared-nothing Technologies and Architecture

A system that automatically distributes data and parallelises query workload across all available hardware.

Data is partitioned across multiple processing nodes each processing data locally. Communication is controlled by a master host and occurs over a network. There is no disk sharing or memory contention. Linearly scalable. Distribution of workload to processor level. Speeds up analytical functions.

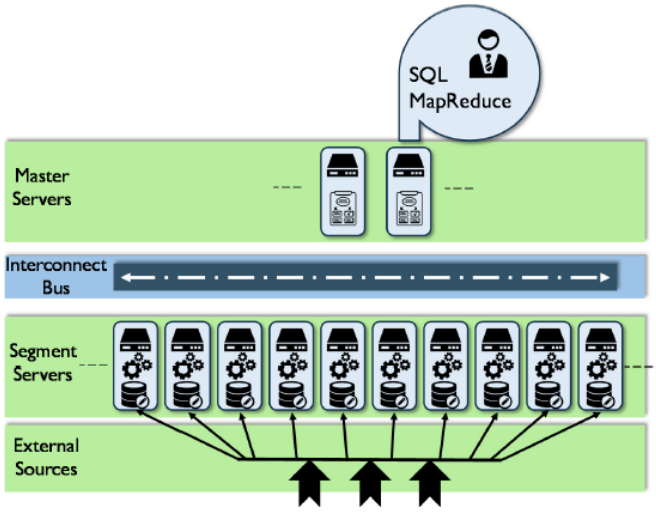


Figure 102 Columnar Appliance Architecture[95]

3.2 Distributed File-based Databases:

- Open source Hadoop stores files of any type
- Language used is called MapReduce:
  - **Map:** Identify and obtain data to be analysed
  - **Shuffle:** Combine data according to desired analytical patterns
  - **Reduce:** Remove duplication and perform aggregation to reduce the size of the data set to only what is required.

3.3 In-database Algorithms

Each processor in an MPP Shared-nothing platform can run queries independently. Computation close to data reduces time for complex algorithms.

3.4 Big Data Cloud Solutions

Vendors provide cloud storage and enhancements

3.5 Statistical computing and graphical languages

R – scripting language with statistical computing and graphics.

3.6 Data visualisation tools

Traditional visualisation tools: both data and graphical

Information graphics / Infographics: Insights, changes in data over time. Interactive, sophisticated analysis, adherence to visualisation best practices

4 Techniques

- **Analytic Modelling:** Different depths of analysis:
  - **Descriptive modelling:** Summarises data structures in a compact manner
  - **Explanatory modelling:** statistical models for testing a hypothesis
- **Big Data Modelling:** Data needs to be integrated, specified and managed by applying traditional Enterprise Architecture principles.

5 Implementation Guidelines

5.1 Strategy Alignment

Strategically aligned with business objectives. Documents goals, approach and governance principles. Strategy deliverables should account for managing:

- Information lifecycle
- Metadata
- Data Quality
- Data acquisition
- Data access and security
- Data governance
- Data privacy
- Learning and adoption
- Operations

5.2 Readiness Assessment / Risk Assessment

Critical success factors:

- **Business relevance:** Big Data/Data Science initiatives must enforce business function
- **Business readiness:**
  - Prepared for long term delivery?
  - Committed to establishing centres of excellence to sustain the product?
  - How broad is the skill gap?
- **Economic viability:** Assessment of ownership costs – buying vs leasing
- **Prototype:** Can the solution be prototyped?

5.3 Organisation and Cultural Change

Need a communications program to engage stakeholders and a centre of excellence to provide training, best practices, knowledge management and communication across developer, designer, analyst and data consumer communities.

Cross functional roles:

- **Big Data platform architect:** Hardware, OS, file systems and services
- **Ingestion architect:** Data analysis, systems of record, modelling
- **Metadata specialist:** Metadata interfaces, architecture and contents
- **Analytic Design Lead:** End user analytic design
- **Data Scientist:** Architectural and model design based on statistical knowledge

6 Big Data and Data Science Governance

The enterprise view of data should drive decisions on sourcing, sharing, Metadata, enrichment and access.

6.1 Visualisation Channels management

Alignment of the appropriate visualisation tools at the right level of complexity for the user community.

6.2 Data Science and Visualisation Standards

Vital for customer-facing and regulatory-facing content:

- Tools standards by analytic paradigm, user community, subject area
- Requests for new data
- Data set process standard
- Processes to avoid biased results:
    - Data inclusion and exclusion
    - Assumptions in the models
    - Statistical validity of results
    - Validity of interpretation of results
    - Appropriate methods applied

6.3 Data Security

- Agree upon levels of access for authorised personnel
- Mask data for those not authorised
- Use encryption for highly sensitive data
- Recombination measures the ability to reconstitute sensitive or private data and must be managed
- Outcomes of analysis may violate privacy

6.4 Metadata

Managed as part of data ingestion else data lake becomes a swamp.

6.5 Data Quality

Data Quality is a measure of deviation from expected result, the smaller the difference the higher the quality. Mature Big Data organisations scan data inputs using data quality tools to understand the information within:

- **Discovery:** Where information resided in the data set
- **Classification:** What type of information based upon standardised patterns
- **Profiling:** How data is populated and structured
- **Mapping:** What other data sets can be mapped to these values

6.6 Metrics

- **Technical Usage Metrics:**
  - Look for data hot spots to manage distribution and performance
  - Growth rates for capacity planning
- **Loading and scanning metrics:**
  - Ingestion rate
  - Interaction with the community
  - Provided by execution logs
- **Learnings and stories:**
  - Counts and accuracy of models and patterns
  - Revenue realised for identified opportunities
  - Cost reduction from avoiding identified threats

Data Management Maturity Assessment

1 Introduction

Capability Maturity Assessment (CMA) is an approach to process improvement based on the Capability Maturity Model (CMM) that describes how processes evolve from ad hoc to optimal.

Progression happens in a set order:

- **Level 0:** Absence of capability. No formal process for managing data. Very few organisations exist at Level 0.
- **Level 1:** Initial or ad hoc: success depends on the capability of individuals
- **Level 2:** Repeatable: Minimum process discipline is in place
- **Level 3:** Defined: Standards are set and used
- **Level 4:** Managed: Processes are quantified and controlled
- **Level 5:** Optimised: Process improvement goals are quantified

Criteria are described across process features related to execution, level of automation, policies, controls and/or process details.

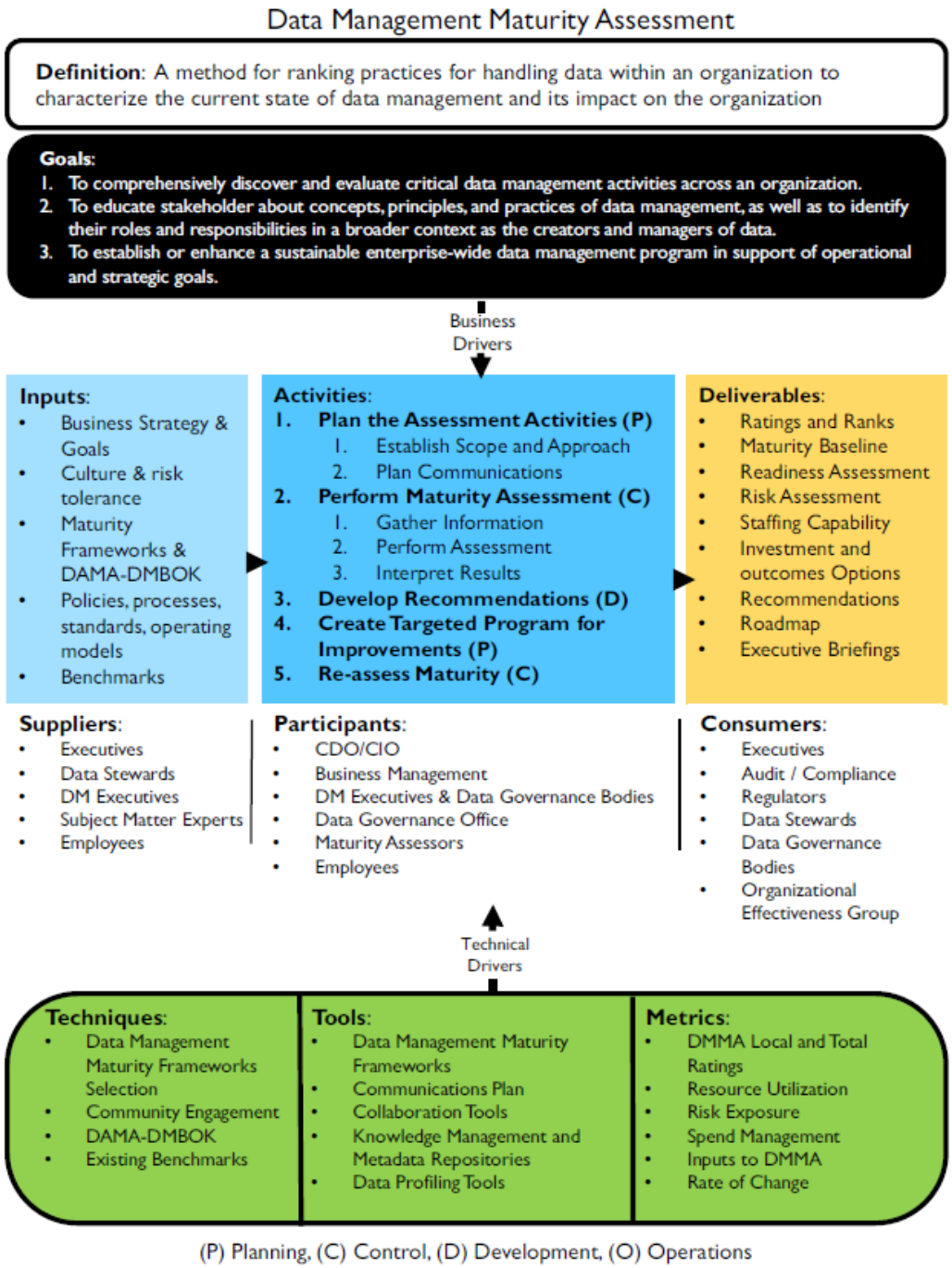
A Data Management Maturity Assessment (DMMA) evaluates DM overall, within a knowledge area of even a process. Used as a bridge between business and IT.

DMMA provides a common vocabulary and a stage-based roadmap to improvement across knowledge areas.

1.1 Business Drivers

Reasons organisations conduct capability maturity assessments:

- **Regulation:** Minimum level of maturity required
- **Data Governance:** The data Governance function requires a maturity assessment for planning and compliance
- **Organisational readiness for process improvement:** Begin by assessing current state
- **Organisational change:** e.g., challenge posed by a merger. DMMA provides input
- **New technology:** To understand the likelihood of success in adopting new technology
- **Data management issues:** Baseline current state in order to make better decisions on change



1.2 Goals and Principles

Goals:
1. To comprehensively discover and evaluate critical data management activities across an organization.
2. To educate stakeholder about concepts, principles, and practices of data management, as well as to identify their roles and responsibilities in a broader context as the creators and managers of data.
3. To establish or enhance a sustainable enterprise-wide data management program in support of operational and strategic goals.

Primary goal is to evaluate the current state of maturity to plan for improvement. The organisation is placed on the maturity scale by clarifying specific strengths and weaknesses.

Benefits:

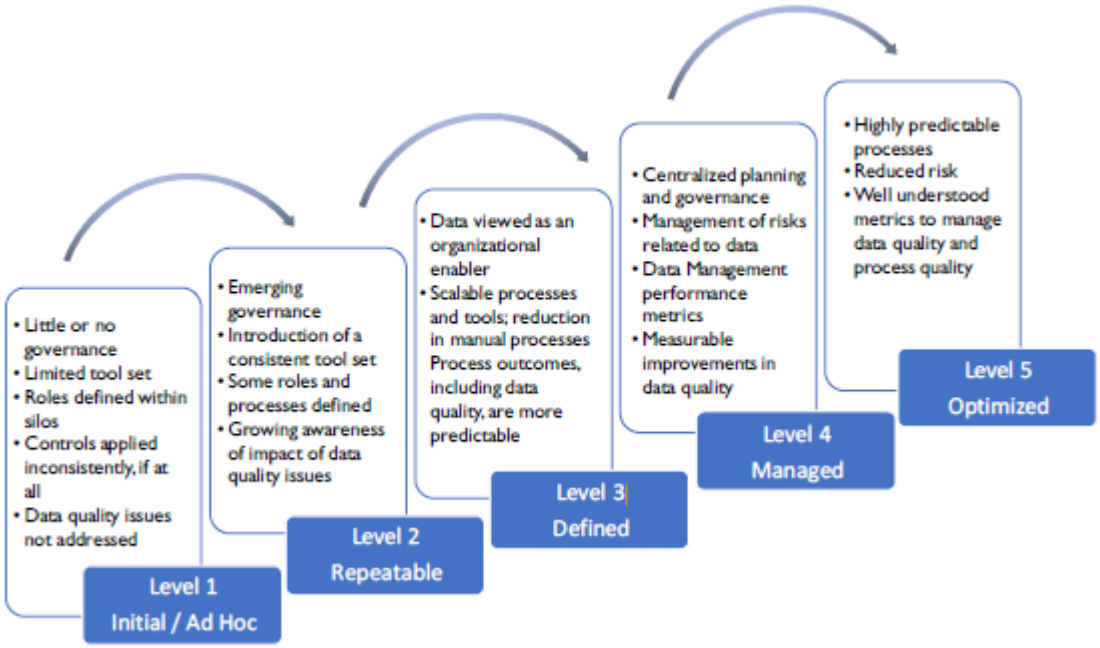
- Educates stakeholders about data management concepts, principles and practices
- Clarify stakeholder roles and responsibilities in relation to organisational data
- Highlight the need to manage data as a critical asset
- Broaden recognition of data management activities across the organisation
- Improve the collaboration necessary for effective data governance

1.3 Essential concepts

1.3.1 Assessment levels and characteristics

CMMs define 5 or 6 levels of maturity:

Level 0: No Capability

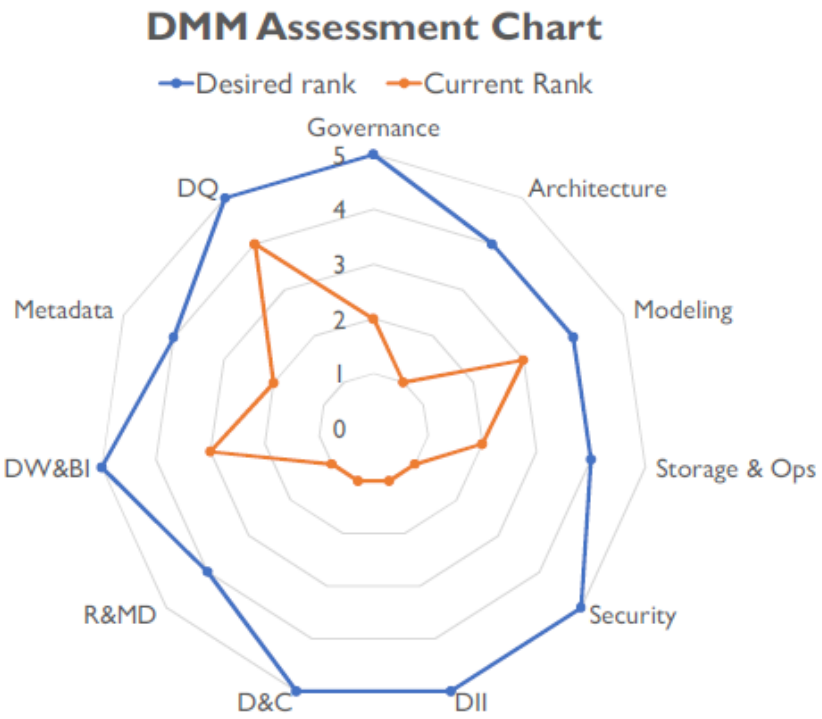


1.3.2 Assessment criteria

Assessment criteria are dependent on the process being evaluated. Will be evaluated along a scale such as 1 -Not started, 2 – in process, 3 – functional, 4 – effective, showing progress and movement to the next level.

When assessing using a model that can be mapped to a DAMA-DMBOK Data Management Knowledge Area, formulate criteria according to the relevant Context Diagram:

- **Activity:** To what degree is the activity in place? Are execution criteria defined? How well defined and executed is the activity? Are best practice outputs produced?
- **Tools:** To what degree is automation and supported by tools? Is tool training provided? Are tools available where needed? Optimally configured? To what extent is long term technology planning in place?
- **Standards:** To what degree is the activity supported by a common set of standards? Are standards enforced and supported by governance and management?
- **People and Resources**: To what degree is the organisation staffed to carry out the activity? What specific skills, training and knowledge are required? How well are roles and responsibilities defined?



1.3.3 Existing DMMA Frameworks

- **CMMI (Capability Maturity Model Institute) Data Management Maturity Model (DMM):** Provides assessment criteria for the following data management areas:
  - Data Management Strategy
  - Data Governance
  - Data Quality
  - Platform and Architecture
  - Data Operations
  - Supporting Procedures

  Identifies sub-processes within the areas and accounts for relation between the areas.

- **EDM Council DCAM:** Enterprise Data Management Council – financial services in USA – Data Management Capability Assessment Model
- **IBM Data Governance Council Maturity Model:** Organised around four key categories:
  - **Outcomes:** Data risk management and compliance, value creation
  - **Enablers:** Organisational structure awareness, policy, stewardship
  - **Core disciplines:** Data Quality Management, information lifecycle management, information security and privacy
  - **Supporting Disciplines:** Data Architecture, classification and Metadata, audit information, logging and reporting.
- **Stanford Data Governance Maturity Model:** Focusses on Data Governance
- **Gartner's Enterprise Information Management Maturity Model**

2 Activities

2.1 Plan Assessment activities

Plan for assessments to be conducted in a short, defined timeframe to expose current strengths and opportunities for improvement.

The goal is to reach a consensus view of current capabilities by interviewing business, data management and Information technology participants, supported by evidence.

2.1.1 Define objectives

Must describe the focus and influence the scope of the assessment. Must be understood by business who help align with the organisation's strategic direction. Assessment objectives help decide which model to adopt and which business areas to prioritise.

2.1.2 Choose a framework

In the context of assumptions about current state, and which will inform the organisation

2.1.3 Define organisational scope

Keep scope manageable and phase into a larger enterprise assessment: Trade-offs between local and enterprise assessments:

- **Localised assessments:** Deeper into details, and is done more quickly. Can be aggregated and weighted to form an enterprise assessment
- **Enterprise Assessments:** Broad. parts of the organisation are sometimes disconnected. Can evaluate different functions based on the same criteria.

2.1.4 Define Interaction approach

Follow recommendations on the selected model. Information gathering should work well within the organisation's culture and minimise participants' time. Includes inspection and evaluation of artefacts and other evidence.

2.1.5 Plan communications

Communications contributes to success. Ensure participants understand the assessment model, as well as how the findings will be used. Describe:

- The purpose of the DMMA
- How it will be conducted
- Their involvement
- Schedule of assessment activities

Include report on findings to all levels.

2.2 Perform Maturity assessment

- **Gather Information:** Formal ratings of assessment criteria.
- **Perform the Assessment:**
  - Review results against a rating method and assign a preliminary rating
  - Document supporting evidence
  - Review with participants to get consensus on a final rating
  - Document finding using model criteria statements
  - Visualisations to illustrate findings

2.3 Interpret results

Identify improvement opportunities aligned with organisational strategy. Recommend actions, the steps towards the target state.

Present the meaning of the ratings to the organisation first. Link current capabilities with the business processes and strategies they support, and the benefits of improving these capabilities.

- **Report assessment results:**
  - Business drivers for assessment
  - Overall results of assessment
  - Ratings by topic with gaps indicated
  - Recommended approach to close gaps
  - Strengths of organisation observed
  - Risks to progress
  - Investment and outcomes options
  - Governance and metrics to measure progress
  - Resource analysis and potential future utilisation
  - Artefacts that can be reused
- **Develop executive briefings**
  - Prepare executive briefings that summarise findings – strengths, gaps and recommendations.

2.4 Create a Targeted Program for Improvements

Recommendations from the DMMA should have a direct impact on data strategy and IT governance, and should be actionable.

Identify Actions and create a Roadmap:

- Sequenced activities for improvements in different DM functions
- Timeline
- Expected improvements in DMMA ratings
- Oversight activities

Should be accompanied by an approach for measuring progress.

2.5 Re-assess Maturity

Reassessments are part of a lifecycle of continuous improvement.

- First assessment establishes a baseline
- Define re-assessment parameters (including organisational scope)
- Repeat DMMA on a published schedule
- Track trends relative to initial baseline
- Develop recommendations based on findings

3 Tools

- **Data Management Maturity Framework:** Primary tool
- **Communication plan:** engagement model for stakeholders, type of information required and schedule
- **Collaboration tools:** Allow findings to be shared

- **Knowledge Management and Metadata Repositories:** Data standards, policies, methods, agendas, minutes of meetings or decisions and business and technical artefacts are managed here.

4 Techniques

4.1 Selecting a DMM Framework

- **Accessibility:** Non-technical terms that convey functional essence of the activity
- **Comprehensiveness:** Framework address a broad scope of data management activities.
- **Extensible and flexible:** Able to enhance model. Model can be used in part or whole:
- **Future progress path built-in:** Logical way forward within each of the DMM functions
- **Industry-agnostic vs. industry-specific:** Depends on the organisation, but best practices should be adhered to.
- **Level of abstraction or detail:** To be related to the organisation and the work it performs
- **Non-prescriptive:** Describes what must be performed, not how
- **Organised by topic:** DM activities placed in context
- **Repeatable:** Consistent over time
- **Supported by a neutral, independent organisation:** Vendor neutral to avoid conflicts of interest
- **Technology neutral:** Based on practices rather than tools
- **Training support included:** Model is supported by comprehensive training.

4.2 DAMA-DMBOK Framework use

DAMA-DMBOK can be used to establish criteria for a DMMA

5 Guidelines for a DMMA

5.1 Readiness Assessment / Risk Assessment

Risk	Mitigation
Lack of organizational buy-in	Socialize the concepts related to the assessment. Establish benefit statements before conducting the assessment. Share articles and success stories. Engage an executive sponsor to champion the effort and review the results.
Lack of DMMA expertise Lack of time or in-house expertise Lack of communication planning or standards	Use third party resources or specialists. Require knowledge transfer and training as part of the engagement.
Lack of 'Data Speak' in the organization; Conversations on data quickly devolve into discussions about systems	Relate the DMMA to specific business problems or scenarios. Address in the communications plan. The DMMA will educate all participants regardless of background and technical experience. Orient participants to key concepts prior to the DMMA.
Incomplete or out-of-date assets for analysis	Flag 'as of' or balance the rating accordingly. For example, give a -1 to everything that is over 1 year out-of-date.
Narrow focus	Reduce the investigation depth to a simple DMMA and go to other areas for a quick assessment to establish ratings for a later comparative baseline. Conduct the first DMMA as a pilot, then apply lessons learned to address a broader scope. Present in-scope focus of proposed assessment in context of DAMA-DMBOK Knowledge Areas. Illustrate what is being left out of scope and discuss the need to include.
Inaccessible staff or systems	Reduce the horizontal scope of the DMMA by focusing only on available Knowledge Areas and staff
Surprises arise such as a regulation changes	Add flexibility into the assessment work stream and focus.

5.2 Organisational and Cultural change

DMMA locates the organisation on a maturity scale and provides a roadmap for improvement, pointing the organisation through change.

6 Maturity Management Governance

DMMA is part of an overall set of data governance activities, each having a lifecycle. DMMA Lifecycle:

- Initial planning and assessment
- Recommendations
- Action plan
- Periodic re-evaluation

6.1 DMMA Process Oversight

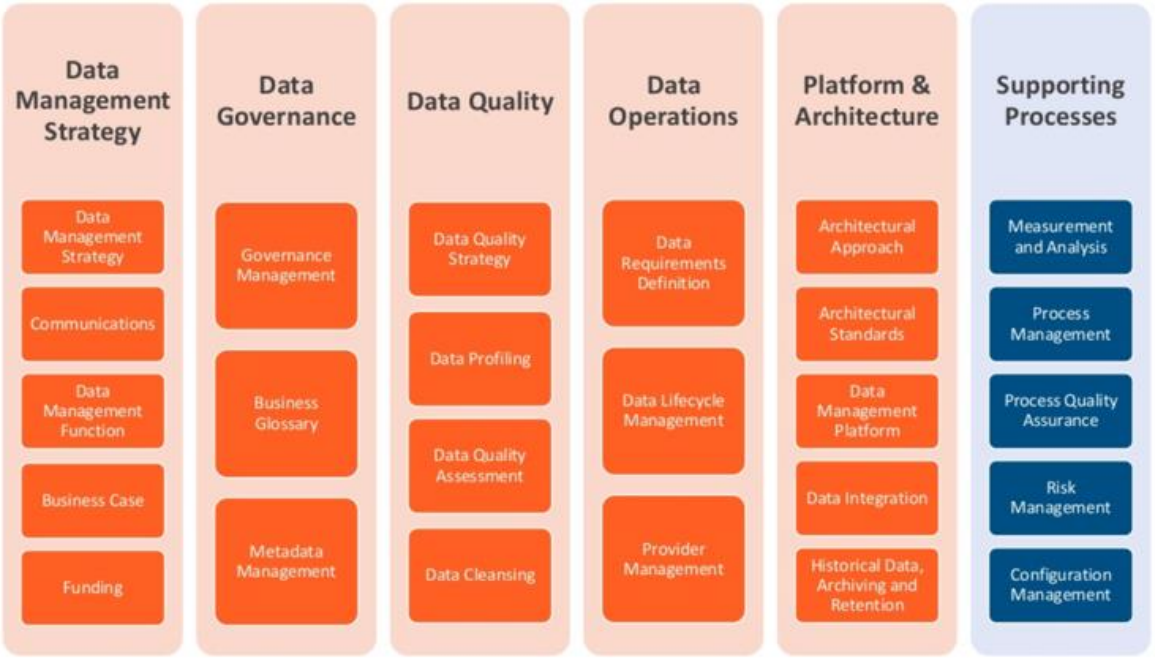
Data Governance team

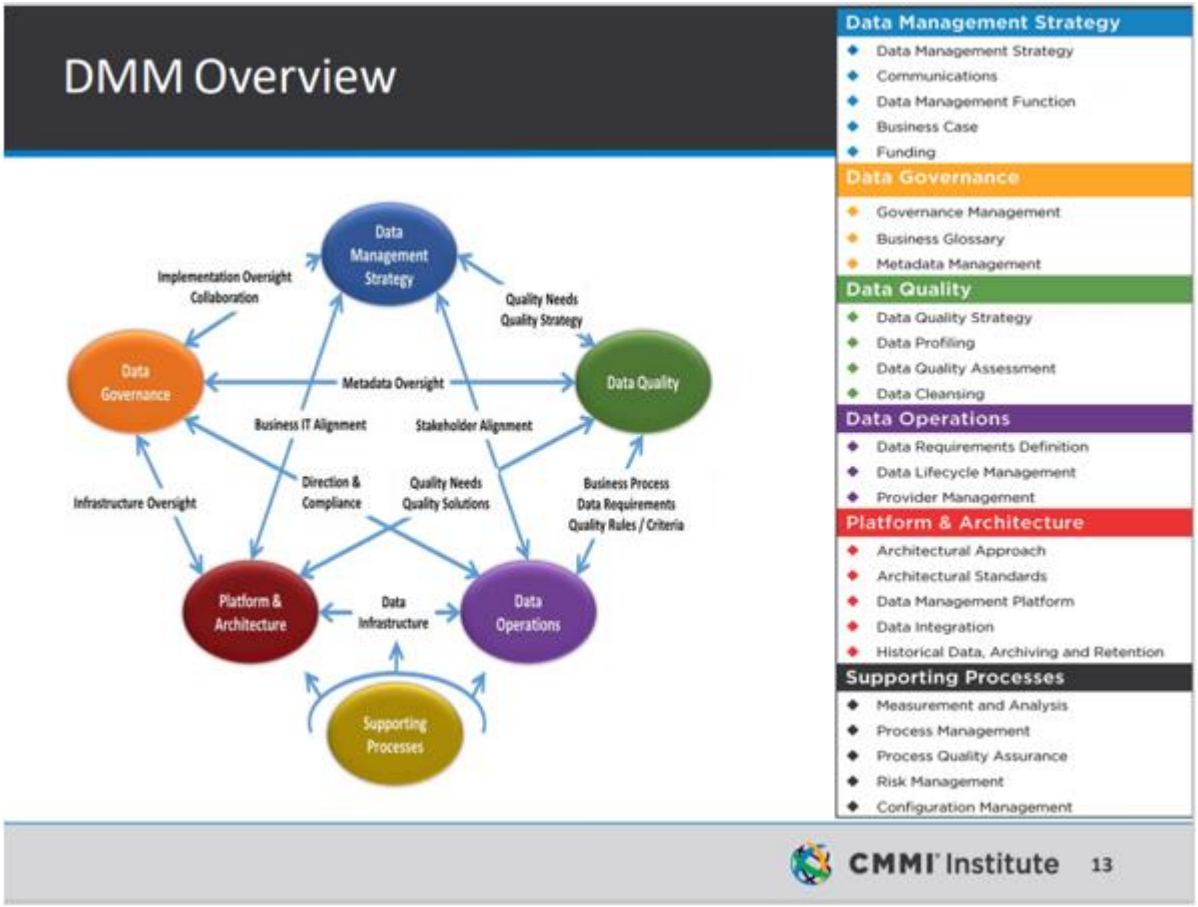
6.2 Metrics

Represent the current state:

- **DMMA Ratings**: Snapshot of the organisations capability level
- **Resource utilisation rates:** cost management in the form of a head count
- **Risk exposure:** or ability to respond to risk
- **Spend management:** How the cost of data management is spread over the organisation. Overlap with Data Governance metrics
    - Data management sustainability
    - Achievement of initiative goals and objectives
    - Effectiveness of communication
    - Effectiveness of education and training
    - speed of change adoption
    - data management value
    - contributions to business objectives
    - reductions in risks
    - improved efficiency in operations
- **Inputs to the DMMA**: Indicate completeness of coverage
- **Rate of change**: Baseline is established. Periodic reassessment to trend improvement.

The DMM^{SM} has been organized into 5 categories with 20 process areas and 5 supporting processes.





Recognizing Business Data Capability Levels

Disabled	Enabled	Led	Driven
• DISTRUST • People Story Telling • Undefined Sources for Decision Making	• AWARENESS • Augmented Story Telling • Curated Data Sources for Decision Making	• INSIGHTS • Data Story Telling • Data Estate Sources for Decision Making	• ACTION • Feedback Story Telling • Data Value Chain Source of Decision Making