MATES ED2MIT

Education and Training for Data Driven Maritime Industry

Tutorial DMG02

Data Management and Governance, DAMA Architecture, Data Governance and Management Best Practices Organisational Roles, Data Stewards

Yuri Demchenko MATES Project
University of Amsterdam

Maritime Alliance for fostering the European Blue economy through a Marine Technology Skilling Strategy

Co-funded by the Erasmus+ Programme of the European Union

Outline

- Importance of Data Management and Data Governance
- DAMA Body of Knowledge (DAMA-BOK)
- DAMA Data Architecture
- Data Governance, Data Management
- Data Modelling, Data formats
- Big Data Infrastructure and Data Workflow
- Data Management Maturity frameworks
- Summary and Discussion

- Practice: Best Practices in Data Governance and Management Plan
  - See supplementary documents folder

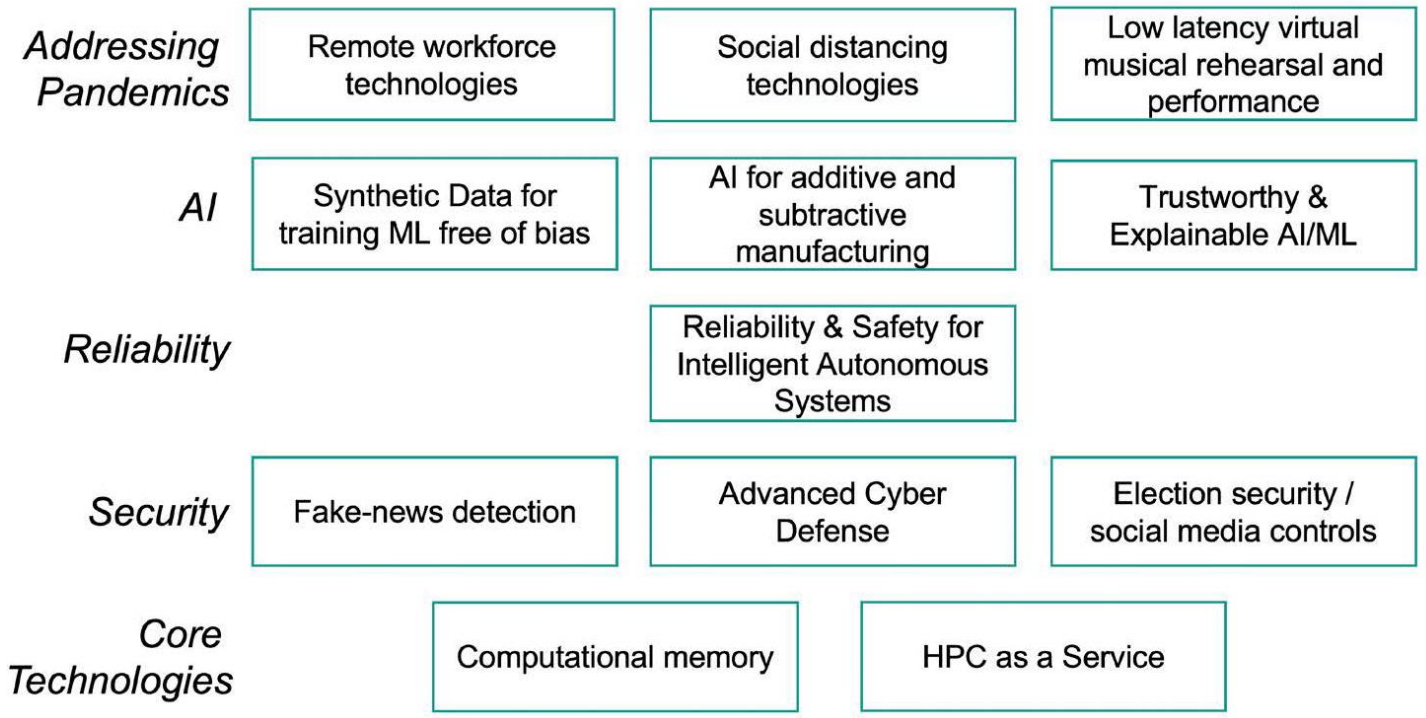
Newsboard: Top Predictions for Data Management in 2021

- Data Fabrics: platforms for heterogeneous data integration
  - Out-of-the-box approaches to data governance and data quality
  - Unified data platforms for better structure and control over data governance
- Organizations continue to embrace AI and ML to fuel data management strategies
  - Augmented data management (ADM)
- Rethinking data management for hybrid and multicloud strategy
- Data sharing and data exchange
- DataOps: Operationalizing Data Analytics and ML is a condition for getting value out of data
- The Use of Blockchain and Distributed Ledger Technology
  - Data lineage and data provenance
- CDO, Chief Data Office will become common role and position
- Democratisation of IT through low-code/no-code platforms
  - Self-Service Data Management And Analytics Tools

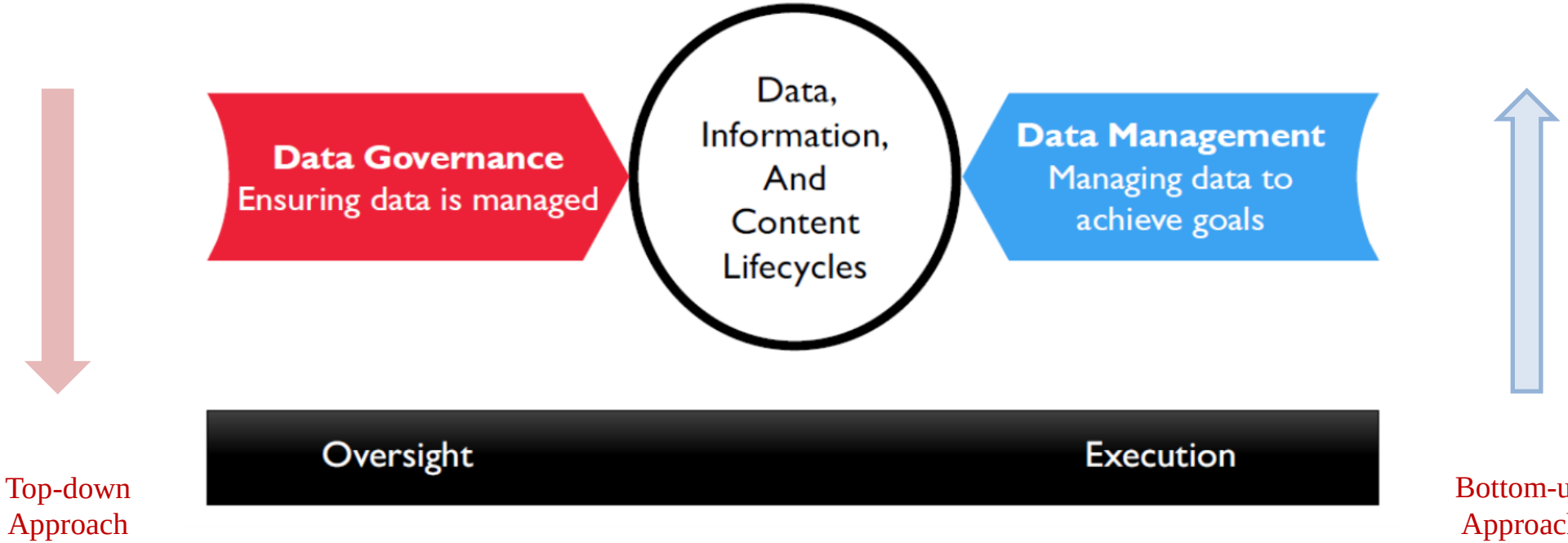
IEEE Computer Society Technology Predictions 2021

https://ieeecs-media.computer.org/media/tech-news/tech-predictions-report-2021.pdf

1. Remote workforce technologies
2. Social distancing technologies
3. Reliability/Safety for Intelligent Autonomous Systems
4. **Synthetic Data for training ML systems free of bias**
5. Fake-news detection
6. HPC as a Service
7. Election security / social media controls
8. Trustworthy & Explainable AI/ML
9. Low latency virtual musical rehearsal and performance
10. Computational memory
11. AI for additive & subtractive manufacturing
12. Advanced Cyber Weapons



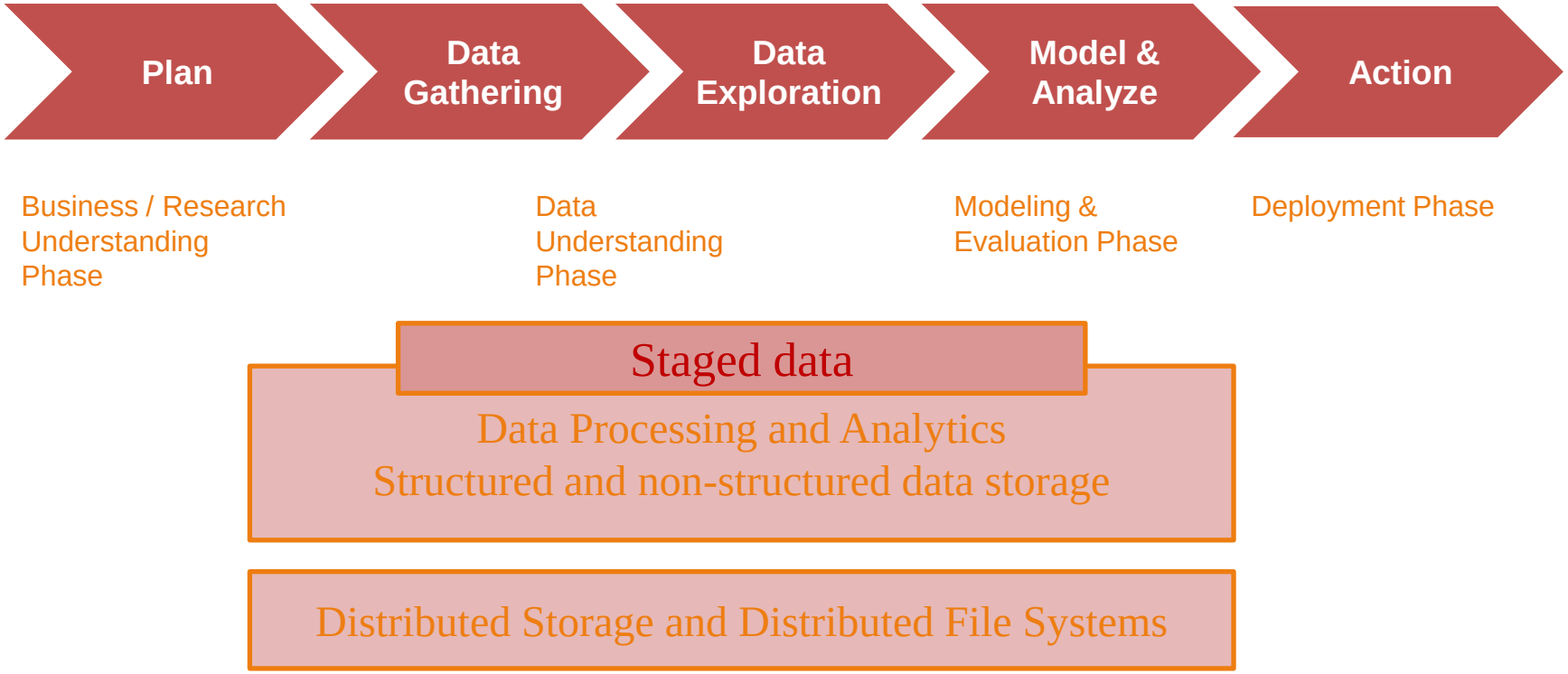
Data Governance and Data Management



- Data Governance – Top down
- Data Management – Bottom-up

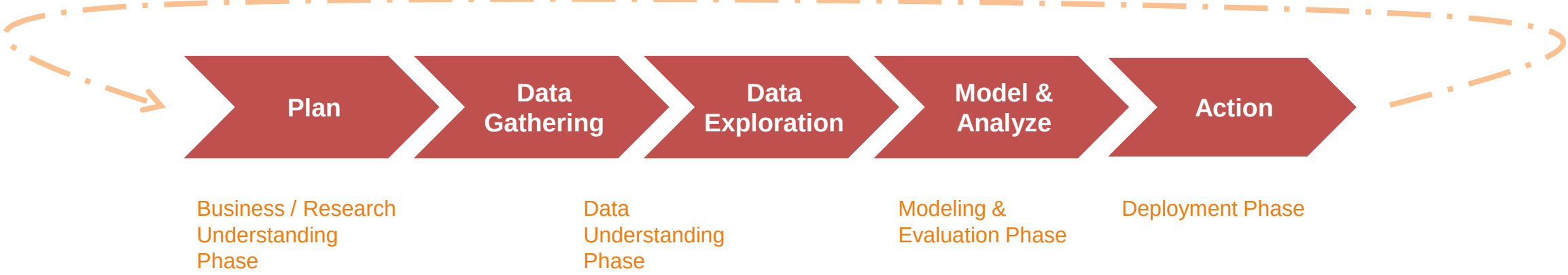
The Analytics Project Flow and Data Management

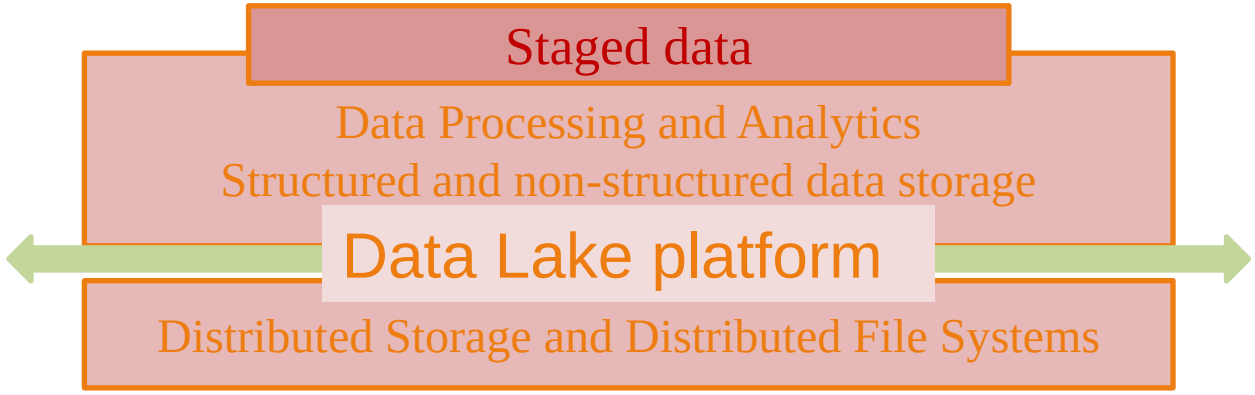
- From idea to action – Business View
- Consider iterative improvement process



The Analytics Project Flow and Data Management

- From idea to action – Business View
- Consider iterative improvement process



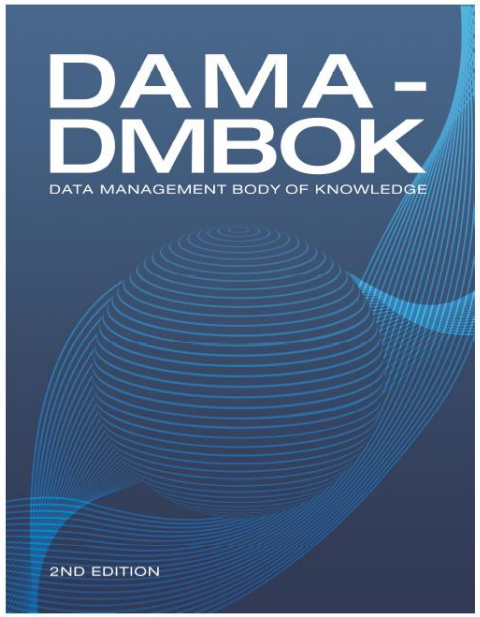


The DAMA-DMBOK Framework

The DAMA-DMBOK Framework goes into depth about the Knowledge Areas that make up the overall scope of data management.

- DAMA-DMBOK Guidelines describe DMBOK and provide re commendations for implementation

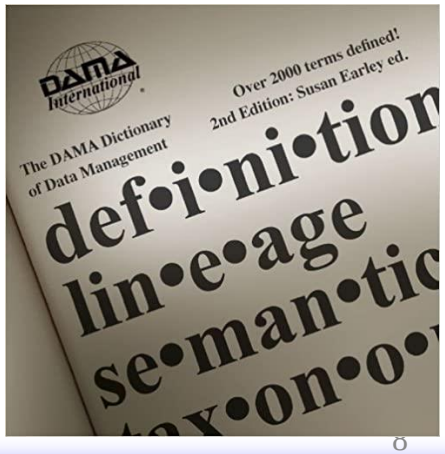
- The DAMA Wheel – 11 Knowledge Areas
- The Environmental Factors hexagon
- The Knowledge Area Context Diagram



626 pages

Technics Publications
BASKING RIDGE, NEW JERSEY

[ref] DAMA – DMBOK: Data Management Body of Knowledge, 2nd Edition, 2017. DAMA International, Technics Publications Llc



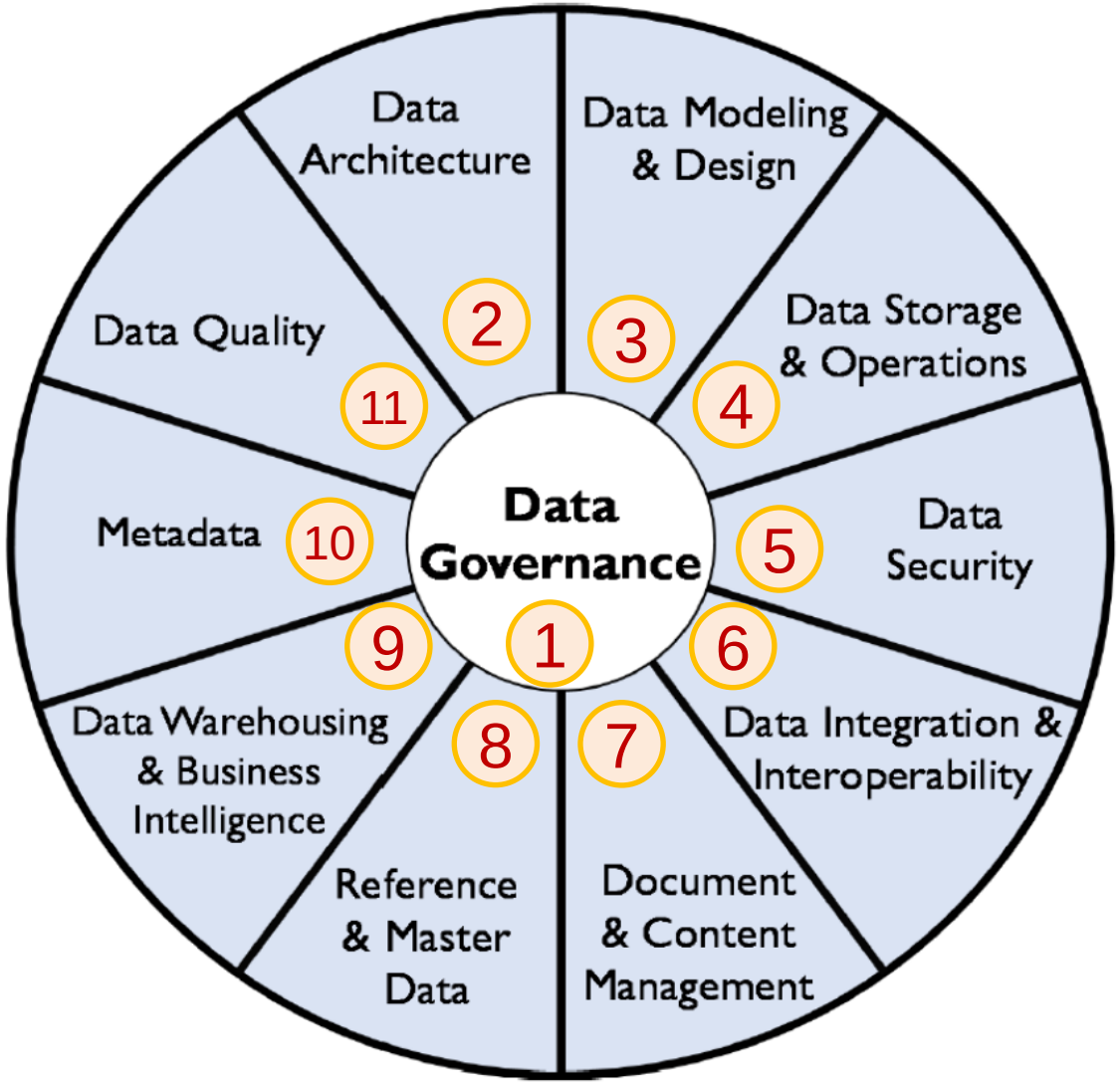
The DAMA Dictionary of Data Management, 2011, 260 pages

11 DMBOK Knowledge Areas

Knowledge Areas describe the scope and context of sets of data management activities.

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.
2. **Data Architecture** defines the blueprint for managing data assets by aligning with organizational strategy to establish strategic data requirements and designs to meet these requirements.
3. **Data Modeling and Design** is the process of discovering, analyzing, representing, and communicating data requirements in a precise form called the *data model*.
4. **Data Storage and Operations** includes the design, implementation, and support of stored data to maximize its value. Operations provide support throughout the data lifecycle from planning for to disposal of data.
5. **Data Security** ensures that data privacy and confidentiality are maintained, that data is not breached, and that data is accessed appropriately.
6. **Data Integration and Interoperability**
7. **Document and Content Management**
8. **Reference and Master Data**
9. **Data Warehousing and Business Intelligence**
10. **Metadata**
11. **Data Quality**

The DAMA2 Wheel: Knowledge Areas



- The DAMA Wheel defines the Data Management Knowledge Areas.
- Data Governance is placed at the center of data management activities, since governance is required for consistency within and balance between the functions.
- The other Knowledge Areas are all necessary parts of a mature data management function, but they may be implemented at different times, depending on the requirements of the organization.

Environmental Factors for DAMA implementation

- The Environmental Factors hexagon shows the relationship between people, process, and technology, and provide a key for reading the DMBOK context diagrams.
- Goals and principles at the center, since these provide guidance for how people should execute activities and effectively use the tools required for successful data management.

Additional Aspects in DMBOK Guidelines

- **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.
- **Big Data and Data Science** describes the technologies and business processes that emerge as our ability to collect and analyze large and diverse data sets increases.
- **Data Management Maturity Assessment** outlines an approach to evaluating and improving an organization's data management capabilities. (Chapter 15)
- **Data Management Organization and Role Expectations** provide best practices and considerations for organizing data management teams and enabling successful data management practices.
- **Data Management and Organizational Change Management** describes how to plan for and successfully move through the cultural changes that are necessary to embed effective data management practices within an organization.

DATA MANAGEMENT PRINCIPLES

Effective data management requires leadership commitment

Data is valuable

- **Data is an asset with unique properties**
- **The value of data can and should be expressed in economic terms**

Data Management Requirements are Business Requirements

- **Managing data means managing the quality of data**
- **It takes Metadata to manage data**
- **It takes planning to manage data**
- **Data management requirements must drive Information Technology decisions**

Data Management depends on diverse skills

- **Data management is cross-functional**
- **Data management requires an enterprise perspective**
- **Data management must account for a range of perspectives**

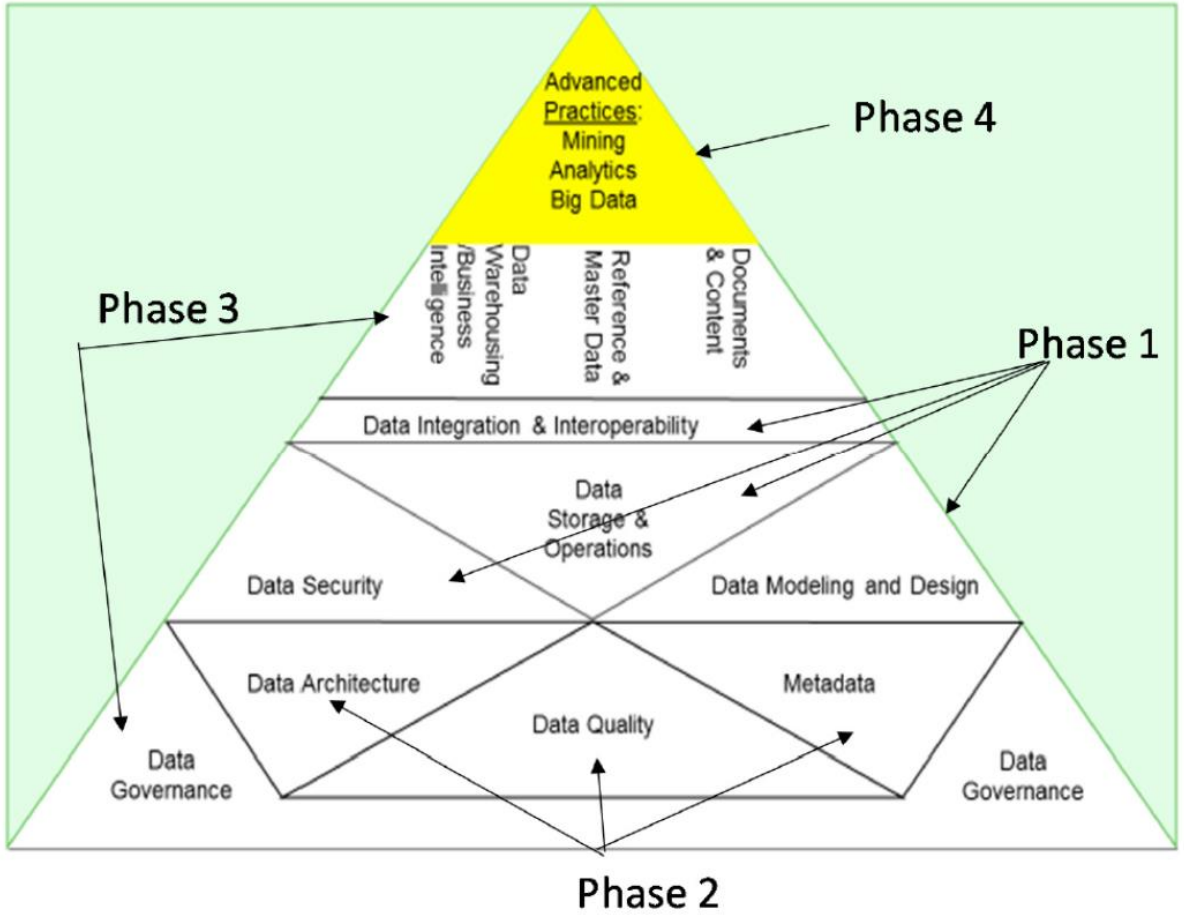
Data Management is lifecycle management

- **Different types of data have different lifecycle characteristics**
- **Managing data includes managing the risks associated with data**

Data Management Principles

- Data is an asset with unique properties
- The value of data can and should be expressed in economic terms
- Managing data means managing the quality of data
- It takes Metadata to manage data
- It takes planning to manage data
- Data management requirements must drive Information Technology decisions
- Data management is cross-functional; it requires a range of skills and expertise
- Data management requires an enterprise perspective
- Data management must account for a range of perspectives
- Data management is lifecycle management
- Different types of data have different lifecycle characteristics
- Managing data includes managing the risks associated with data
- **Effective data management requires leadership commitment**

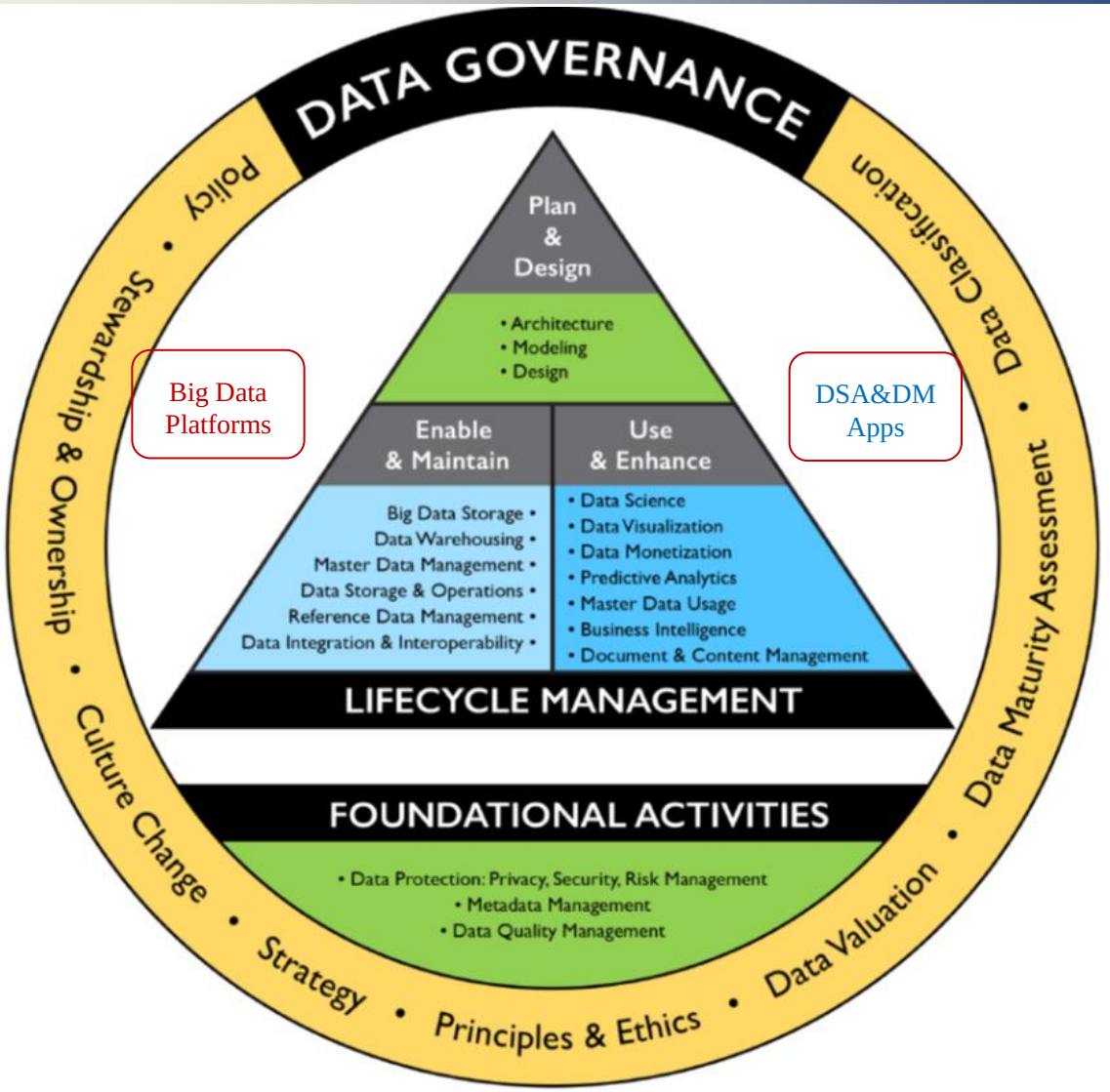
DM-BOK Pyramid (Aiken)



Peter Aiken's framework uses the DMBOK functional areas to describe the situation in which many organizations find themselves.

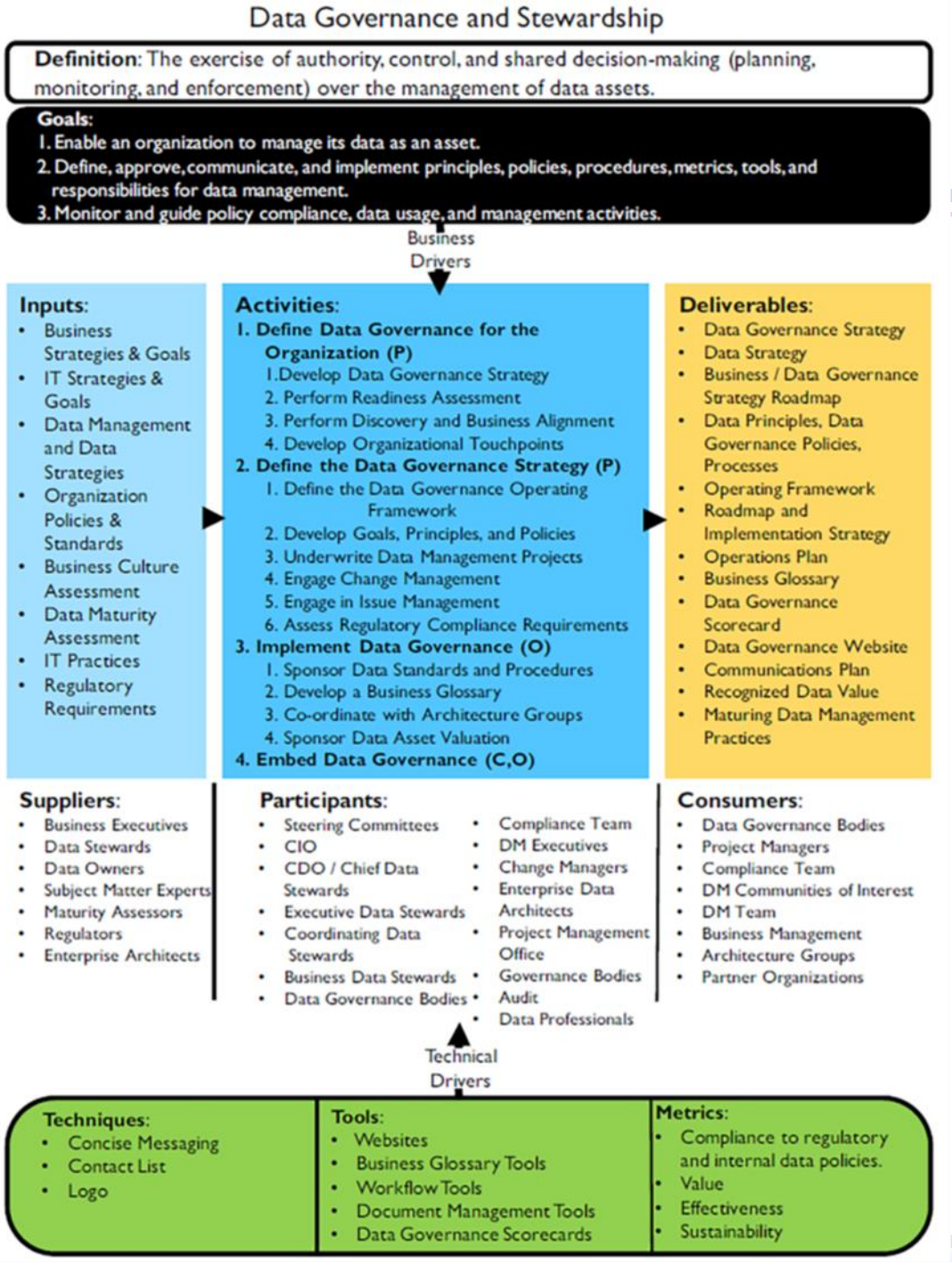
- **Phase 1**: The organization purchases an application that includes database capabilities.
- **Phase 2**: Once they start using the application, they will find challenges with the quality of their data.
- **Phase 3**: Disciplined practices for managing Data Quality, Metadata, and architecture require Data Governance that provides structural support for data management activities.
- **Phase 4**: The organization leverages the benefits of well-managed data and advances its analytic capabilities.

DAMA Wheel Evolved



Data governance activities provide oversight and containment, through strategy, principles, policy, and stewardship.

- Core activities, including Metadata Management, Data Quality Management, and data structure definition (architecture) are at the center of the framework.
- Lifecycle management activities may be defined from a planning perspective and an enablement perspective
- Usages emerge from the lifecycle management activities
- Business Intelligence, Data Science, predictive analytics, data visualization..



Data Governance and Stewardship

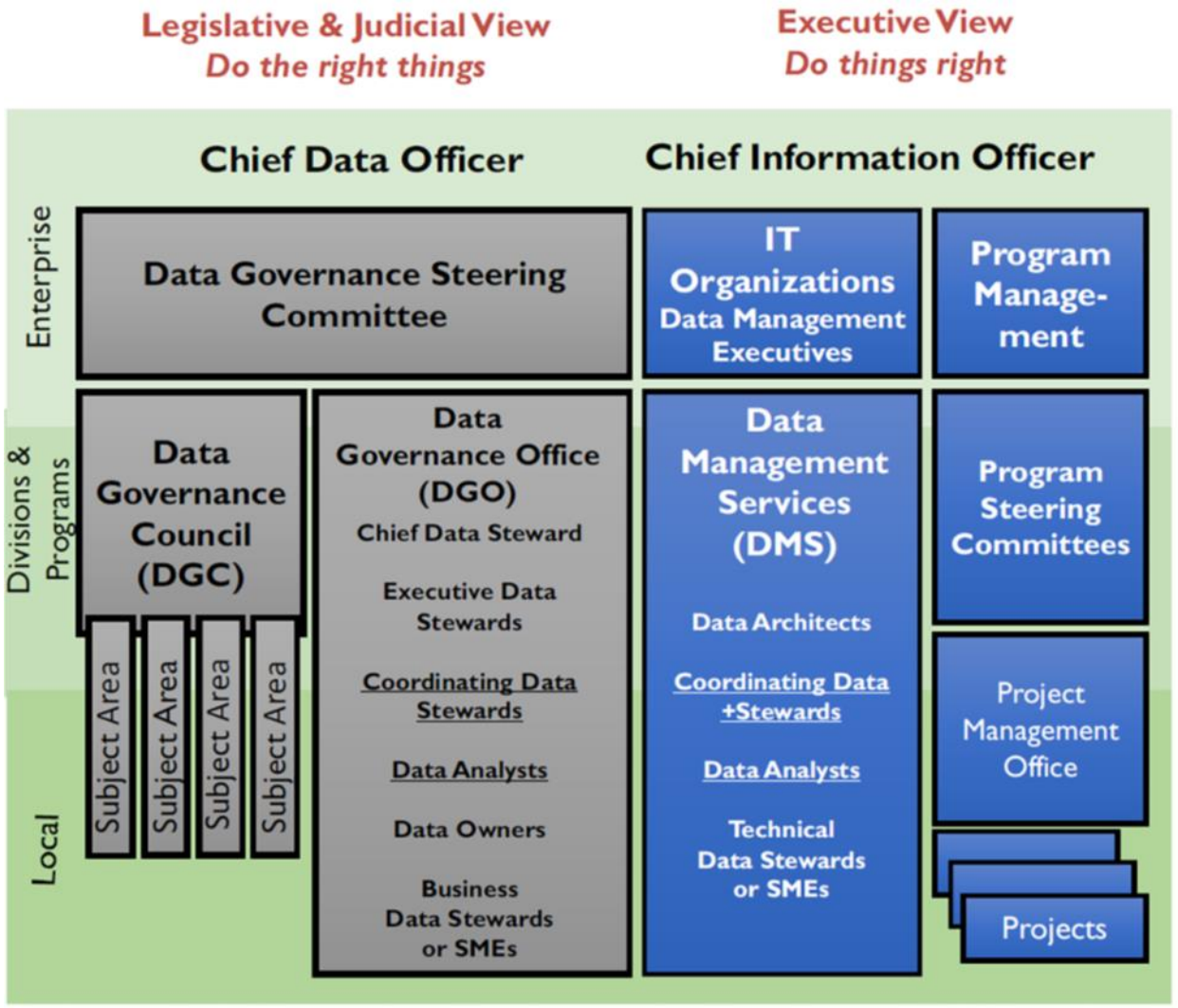
Goals

- Enable an organisation to manage it data as an assets
- Define, approve, communicate, and implement principles, procedures, metrics, tools and responsibilities for data management
- Monitor and guide policy compliance, data usage, and management activities

Scope of a Data Governance Programme

- Strategy
- Policy
- Standards and quality
- Oversight
- Compliance
- Issue management
- Data management projects
- Data asset valuation

Data Governance Organisation Parts



- Separation or governance responsibilities
- Multi-layer
- CDO
- CIO
- Councils

Data Governance Office (DGO)

- Chief Data Steward
- Executive Data Steward
- Business Data Steward or SME

Data Governance

- Data Governance is a collection of practices and processes which help to ensure the formal management of data assets within an organization.
- Data Governance often includes other concepts such as Data Stewardship, Data Quality, and others to help an enterprise gain better control over its data assets, including methods, technologies, and behaviors around the proper management of data.
- It also deals with security and privacy, integrity, usability, integration, compliance, availability, roles and responsibilities, and overall management of the internal and external data flows within an organization.

Data Assets Valuation

- **Replacement cost**: The replacement or recovery cost of data lost in a disaster or data breach, including the transactions, domains, catalogs, documents and metrics within an organization.
- **Market value**: The value as a business asset at the time of a merger or acquisition.
- **Identified opportunities**: The value of income that can be gained from opportunities identified in the data (in Business Intelligence), by using the data for transactions, or by selling the data.
- **Selling data**: Some organizations package data as a product or sell insights gained from their data.
- **Risk cost**: A valuation based on potential penalties, remediation costs, and litigation expenses, derived from legal or regulatory risk from:
  - The absence of data that is required to be present.
  - The presence of data that should not be present (e.g., unexpected data found during legal discovery; data that is required to be purged but has not been purged).
  - Data that is incorrect, causing damage to customers, company finances, and reputation in addition to the above costs.
  - Reduction in risk and risk cost is offset by the operational intervention costs to improve and certify data

Quality of data

Poor quality data is simply costly to any organization.

- Organizations spend between 10-30% of revenue handling data quality issues.
- IBM estimated the cost of poor quality data in the US in 2016 was $3.1 Trillion.

Costs come from:

- Scrap and rework
- Work-arounds and hidden correction processes
- Organizational inefficiencies or low productivity
  - Organizational conflict
  - Low job satisfaction
  - Customer dissatisfaction
  - Lost opportunity costs, including inability to innovate
  - Compliance costs or fines
  - Reputational costs

Benefits of high quality data include:

- 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

[ref] Reported in Redman, Thomas. “Bad Data Costs U.S. $3 Trillion per Year.” Harvard Business Review. 22 September 2016.

https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year.

Regulatory compliance requirements

Assess Regulatory Compliance Requirements:

- In what ways is a regulation relevant to the organization?
- What constitutes compliance? What policies and procedures will be required to achieve compliance?
- When is compliance required? How and when is compliance monitored?
- Can the organization adopt industry standards to achieve compliance?
- How is compliance demonstrated?
- What is the risk of and penalty for non-compliance?
- How is non-compliance identified and reported? How is non-compliance managed and rectified?

Several global regulations have significant implications on data management practices. For example List of standards for compliance,

- **Accounting Standards**: The Government Accounting Standards Board (GASB) and the Financial Accounting Standards Board (FASB) accounting standards also have significant implications on how information assets are managed (in the US).
- **BCBS 239** (Basel Committee on Banking Supervision) and **Basel II** refer to Principles for Effective Risk Data Aggregation and risk reporting.
- **PCI-DSS**: The Payment Card Industry Data Security Standards (PCI-DSS).
- **Solvency II**: European Union regulations, similar to Basel II, for the insurance industry.
- **Privacy laws**: Local, sovereign, and international laws all apply; GDPR for Europe

RDM Focus: FAIR Data Principles

Findable:

- F1 (meta)data are assigned a globally unique and persistent identifier;
- F2 data are described with rich metadata;
- F3 metadata clearly and explicitly include the identifier of the data it describes;
- F4 (meta)data are registered or indexed in a searchable resource;

Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles;
- I3. (meta)data include qualified references to other (meta)data;

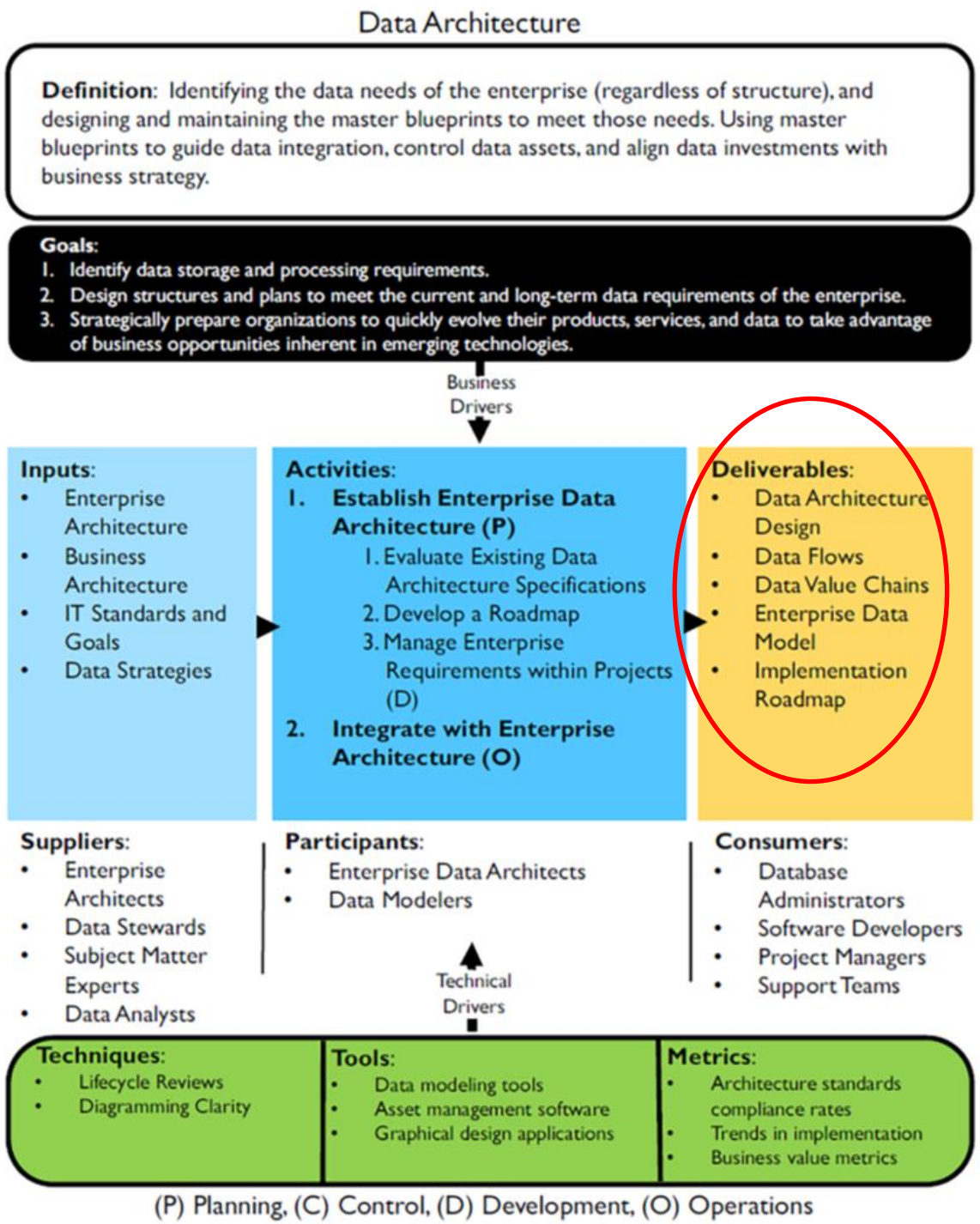
- https://fairdataforum.org/
- Cost of not having FAIR research data http://publications.europa.eu/resource/cellar/d375368c-1a0a-11e9-8d04-01aa75ed71a1.0001.01/DOC_1

Accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol;
  - A1.1 the protocol is open, free, and universally implementable;
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary;
- A2 metadata are accessible, even when the data are no longer available;

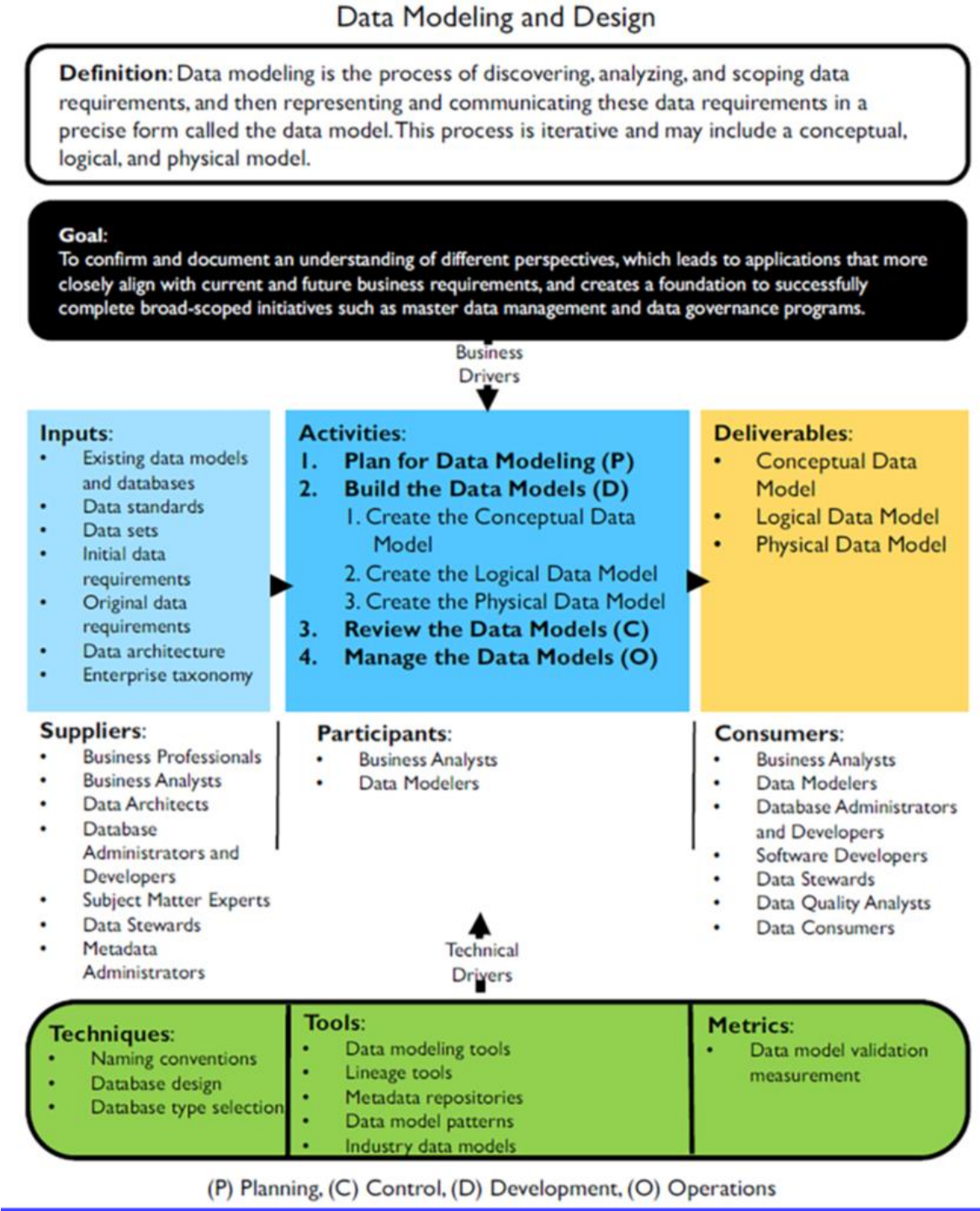
Reusable:

- R1 meta(data) are richly described with a plurality of accurate and relevant attributes;
- R1.1 (meta)data are released with a clear and accessible data usage license;
- R1.2 (meta)data are associated with detailed provenance;
- R1.3 (meta)data meet domain-relevant community standards;



Data Architecture

- Blueprint for organisation on Data Governance and Data Management
- Enterprise Data Architecture, liaised with the Business Architecture
- Data Strategies
- Defines entities, roles and relations
- Link to other organisational model and industry best practices



Data Modeling and Design

- Data structures
- Data Standards
- Data Modeling
- Data Manipulation

Data Modeling Schemes

Scheme	Sample Notations
Relational	Information Engineering (IE) Integration Definition for Information Modeling (IDEF1X) Barker Notation
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

Data structures and data models

Cloud based IaaS/PaaS/SaaS platforms need to provide storage and processing environment for different types of data generated and used by enterprise and user applications.

- Different stages of the data transformation will use or produce data of different structures, models and formats.

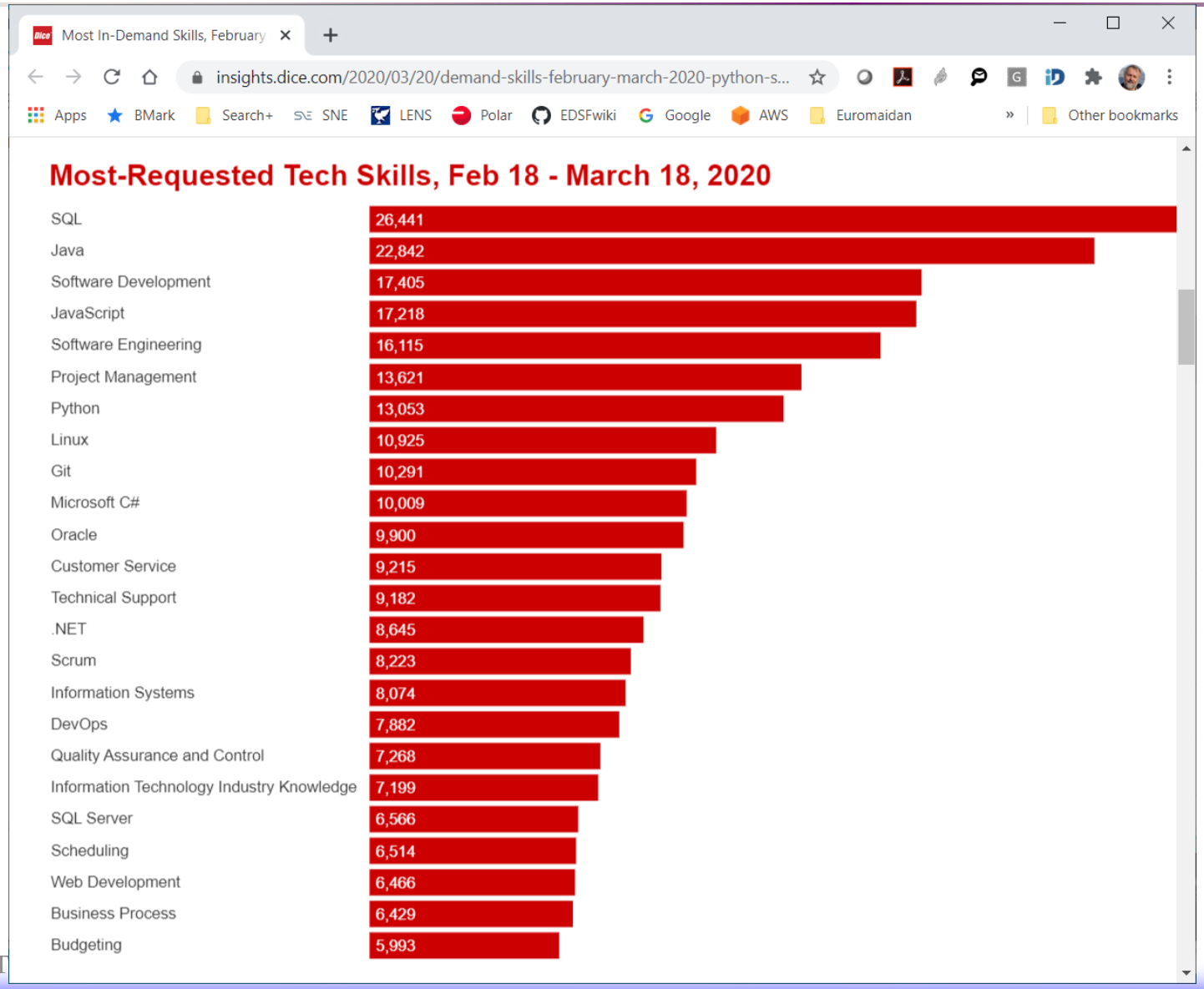
Data types:

- Data described via a formal data model, which are the majority of structured data, data stored in databases, archives, etc.
- Data described via a formalized grammar (e.g. machine generated textual data or forms)
- Data described via a standard format (e.g. digital images, audio or video files)
- Arbitrary textual or binary data

Data models

- Structured data (e.g. relational)
- Unstructured data (e.g. text or HTML pages)
- Semi-structured Data (e.g. tables)
- Key-value pairs
- XML: Hierarchical data (e.g. document)
- RDF: Semantic data (e.g. RDF, triple store)

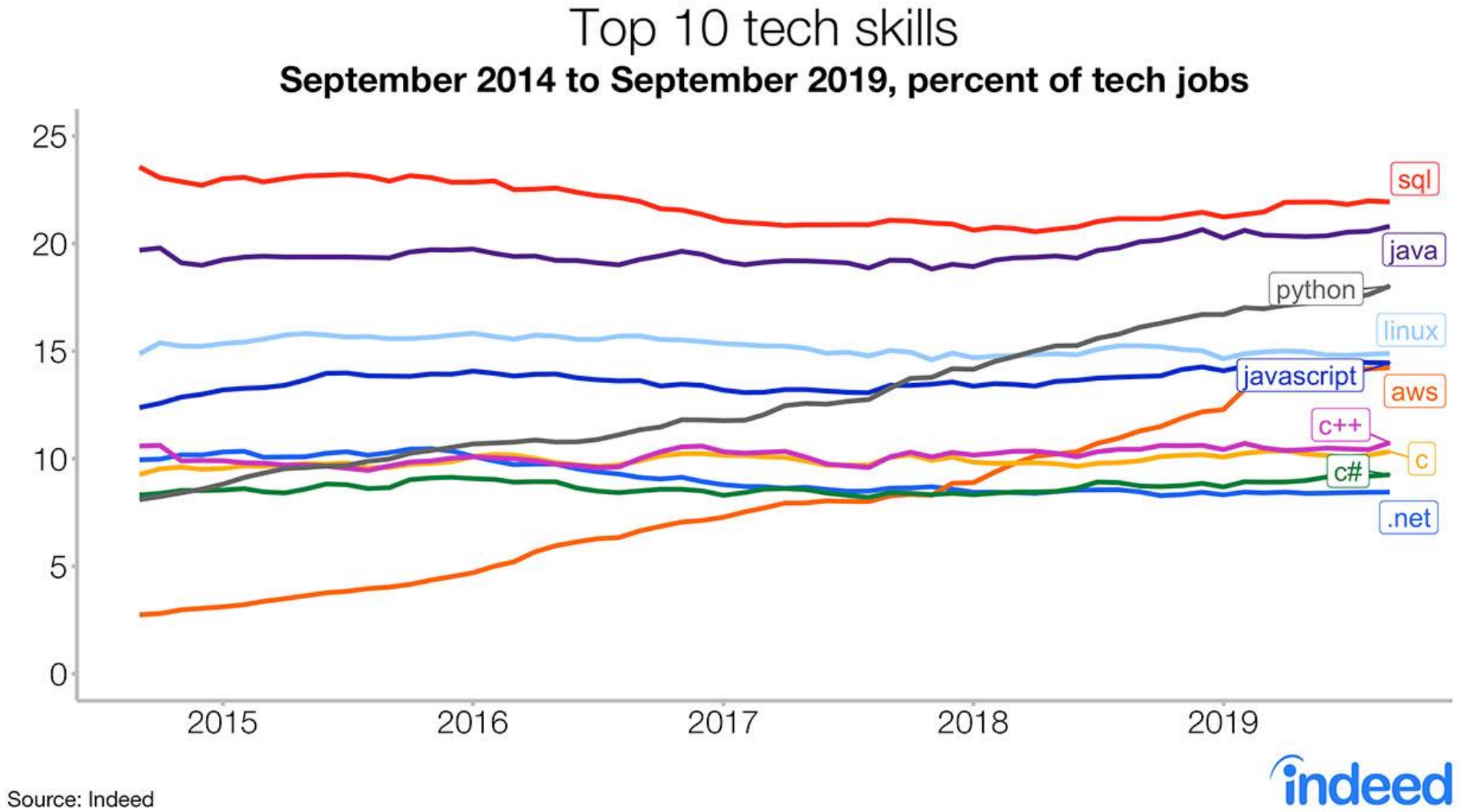
Skills demand on USA Job market

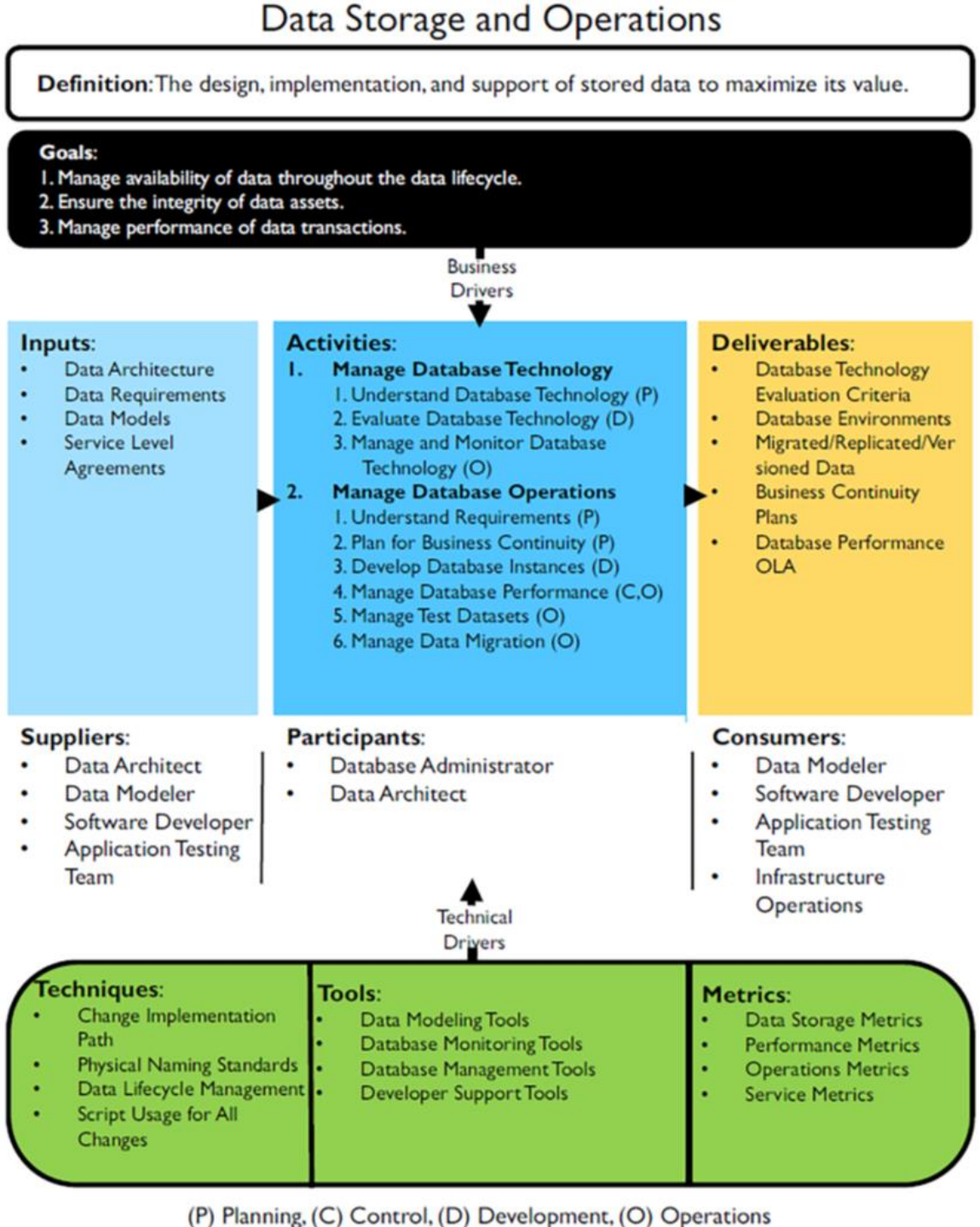


- SQL
- Java
- JavaScript
- Python
- Linux
- Git

SQL, Java Top List of Most In-Demand Tech Skills

https://spectrum.ieee.org/view-from-the-valley/at-work/tech-careers/sql-java-top-list-of-most-indemand-tech-skills





Data Storage and Operations

- Data storage types
- Database management
- Database processing types
- Consistency levels

Big *Database Processing Types*

There are two basic types of database processing in achieving database properties.

- ACID and BASE define alternative model
  - Atomicity – Consistency – Isolation – Durability
  - Basically Available – Soft State - Eventual Consistency
- CAP Theorem is used to define how closely a distributed system may match either ACID or BASE.

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

Database Processing Types: ACID and BASE

ACID

- **Atomicity**: All operations are performed, or none of them is, so that if one part f the transaction fails, then the entire transaction fails.
- **Consistency**: The transaction must meet all rules defined by the system at all times and must void half-completed transactions.
- **Isolation**: Each transaction is independent unto itself.
- **Durability**: Once complete, the transaction cannot be undone.

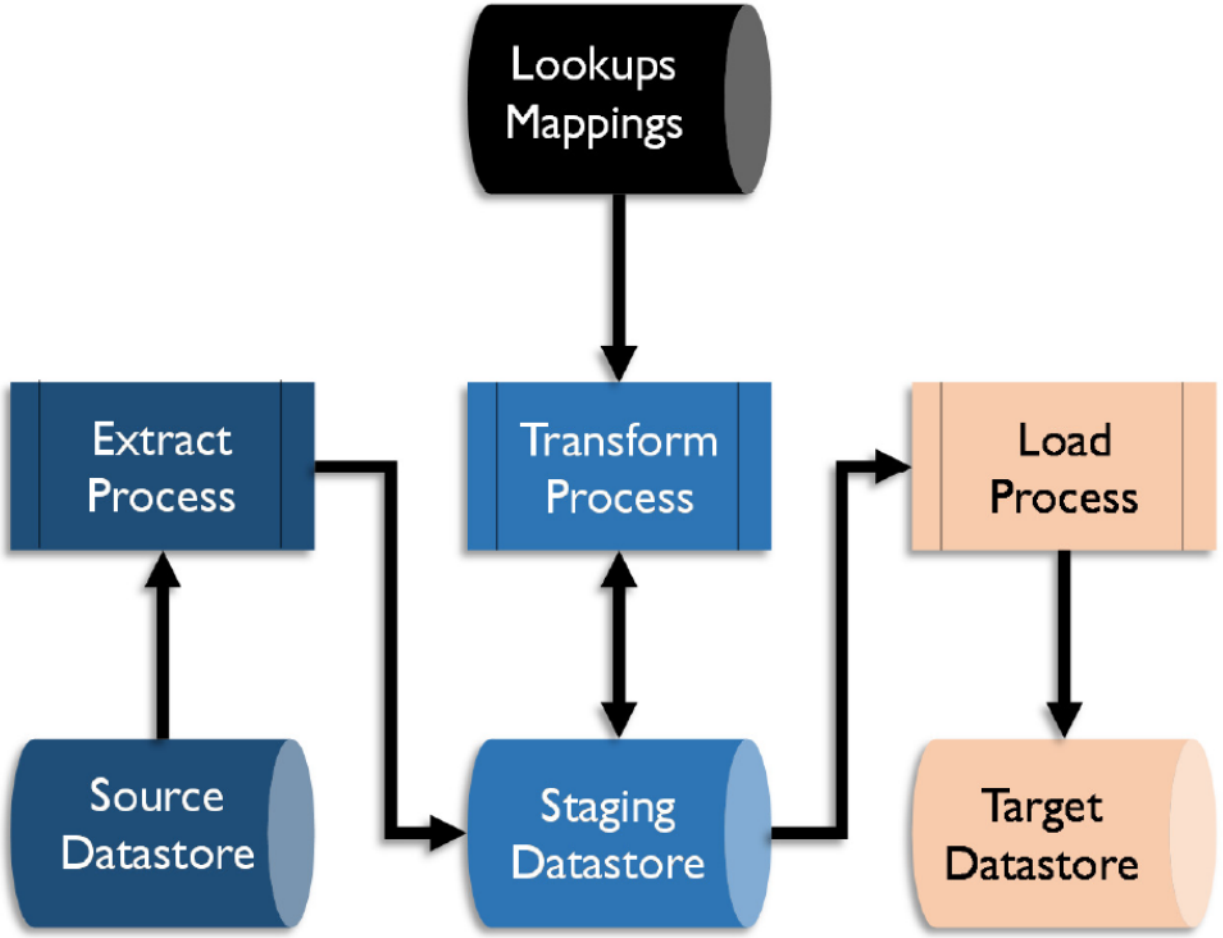
Relational ACID technologies are the dominant tools in relational database storage; most use SQL as the interface.

BASE

- **Basically Available**: The system guarantees some level of availability to the data even when there are node failures. The data may be stale, but the system will still give and accept responses.
- **Soft State**: The data is in a constant state of flux; while a response may be given, the data is not guaranteed to be current.
- **Eventual Consistency**: The data will eventually be consistent through all nodes and in all databases, but not every transaction will be consistent at every moment.

BASE-type systems are common in Big Data environments. Large online organizations and social media companies commonly use BASE implementations, as immediate accuracy of all data elements at all times is not necessary.

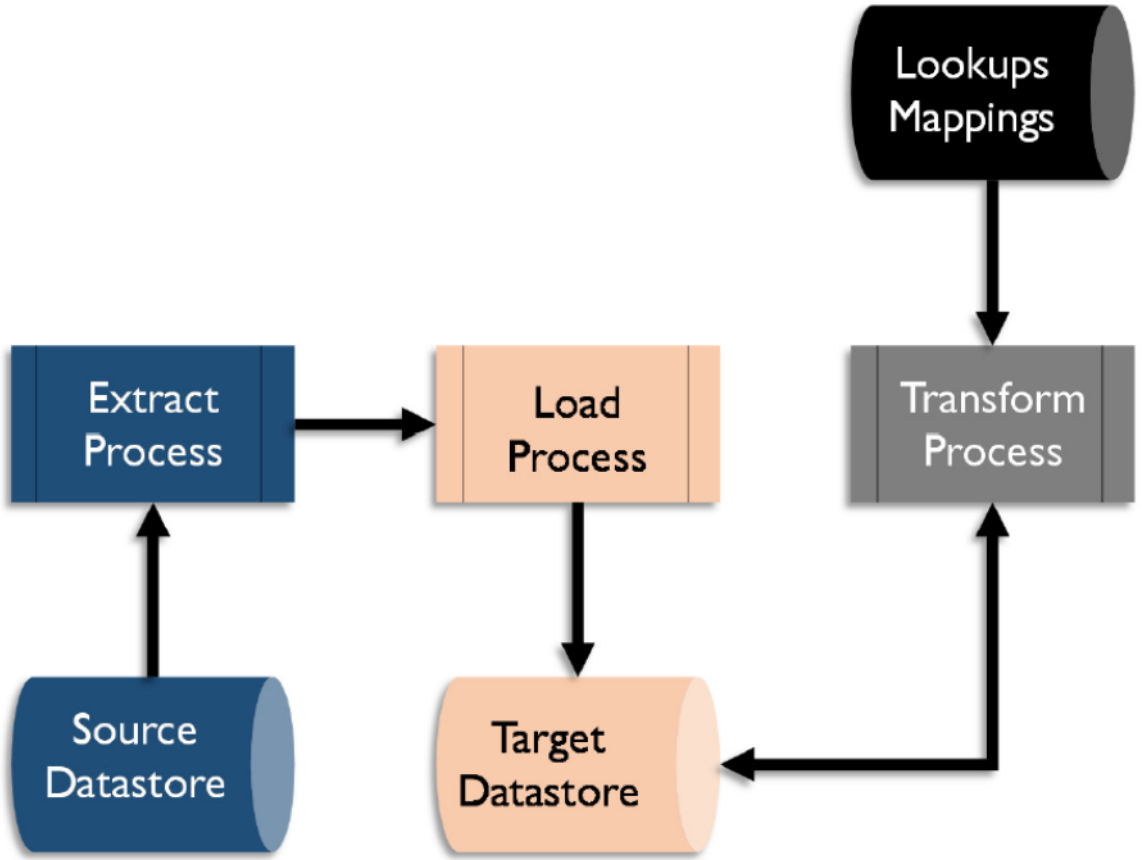
ETL Processes



Extract - Transfer – Load

- Primary model for RDBMS and SQL
- Enforces schema

ELT Processes



Extract – Load - Transfer

- Modern Big Data infrastructure and NoSQL databases
- Load in native data format and schema apply on read
- CEP Complex Events Processing
- Data Exchange Standards
- ESB – Enterprise Service Bus

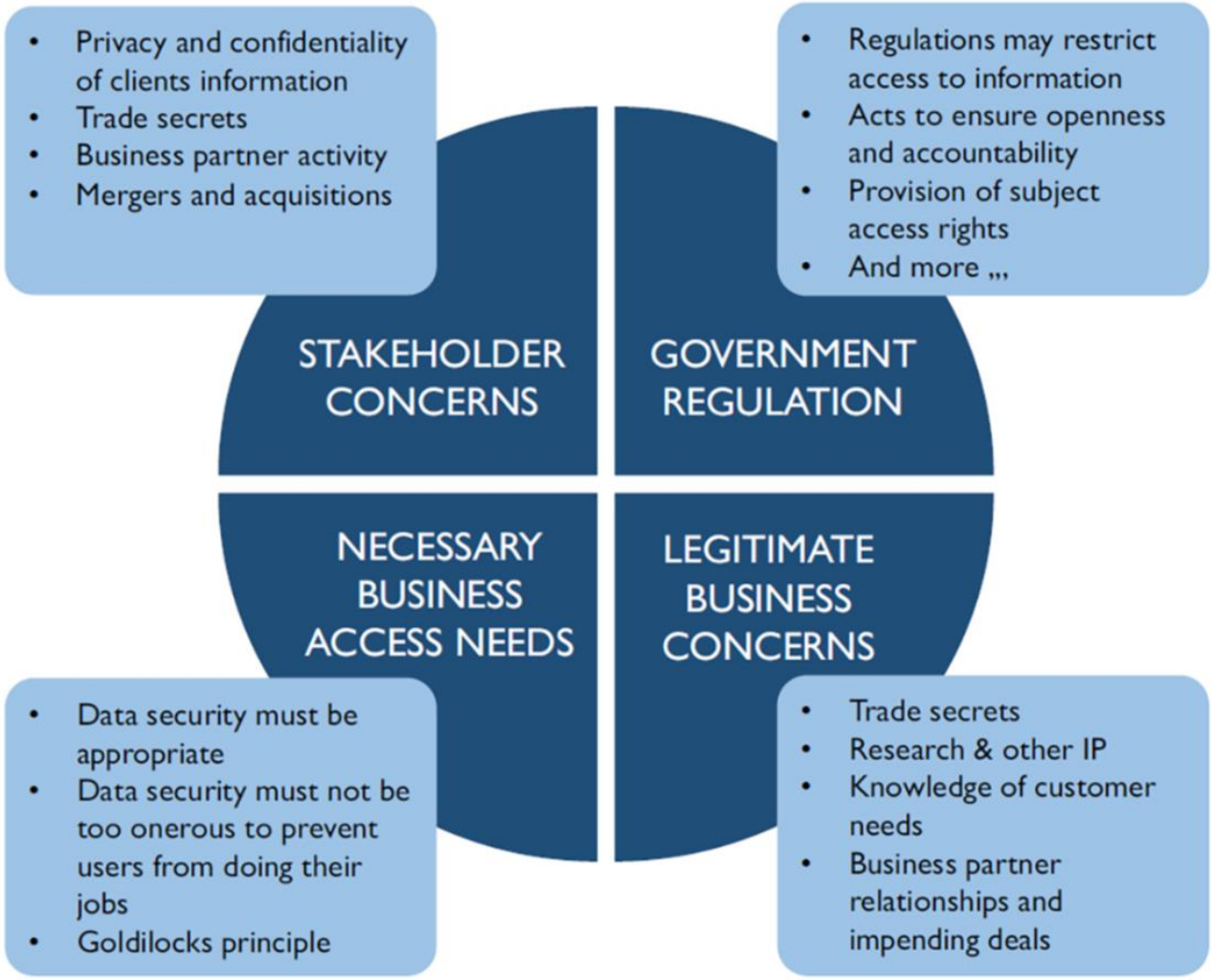
Other DMBOK Knowledge Areas

- Data Security
- Document and Content Management
- Reference and Master Data
- Metadata Management
- Data Quality Management
- Data Warehousing and Business Intelligence
- Data Integration and Interoperability (DII)

Other aspects of Data Management

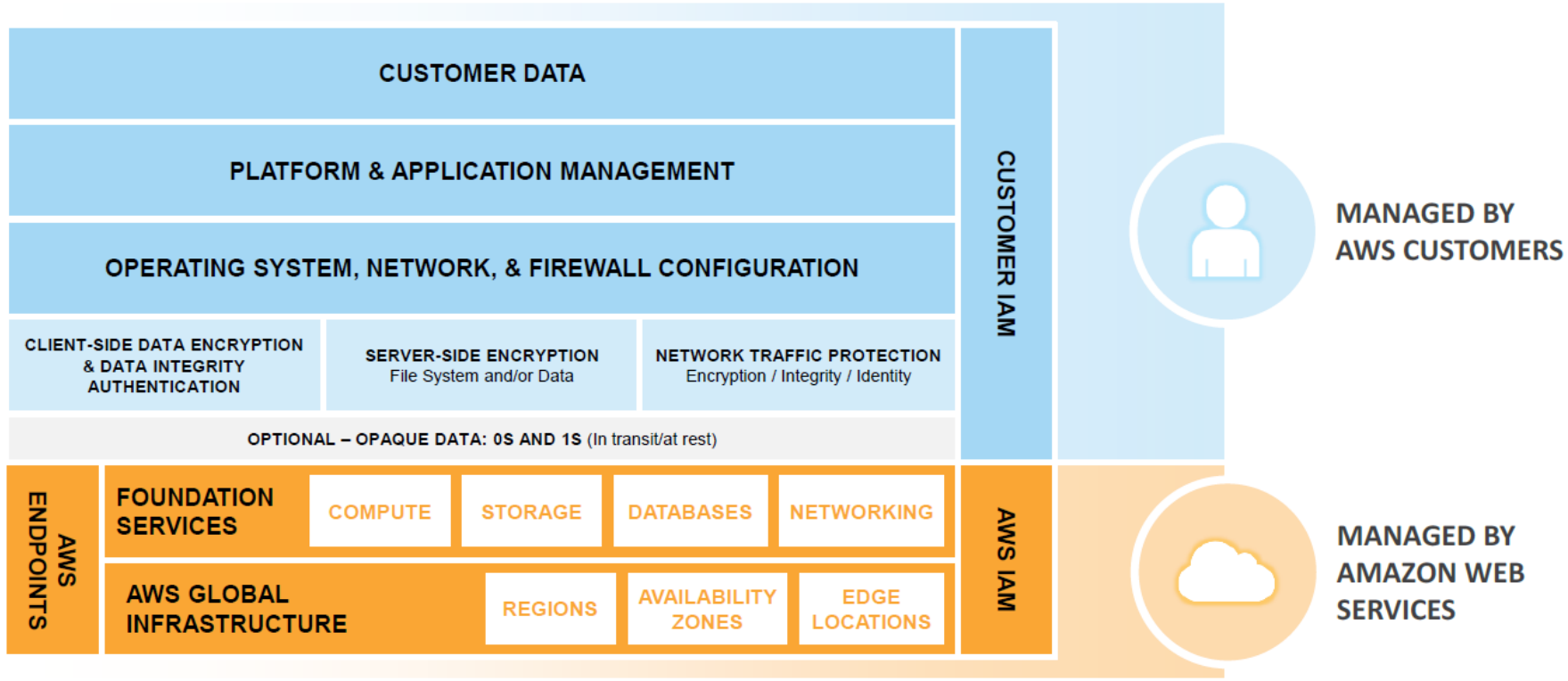
- Data Maturity Management Assessment
- Big Data Infrastructure and Data Science

DMBOK Data Security: Multi-stakeholder Shared Responsibility



- DMBOK accepts the shared responsibility security model where responsibility is shared between stakeholder, in particular users and service providers
- Similarly to shared responsibility in cloud security

Example: Security responsibility sharing in AWS IaaS infrastructure services



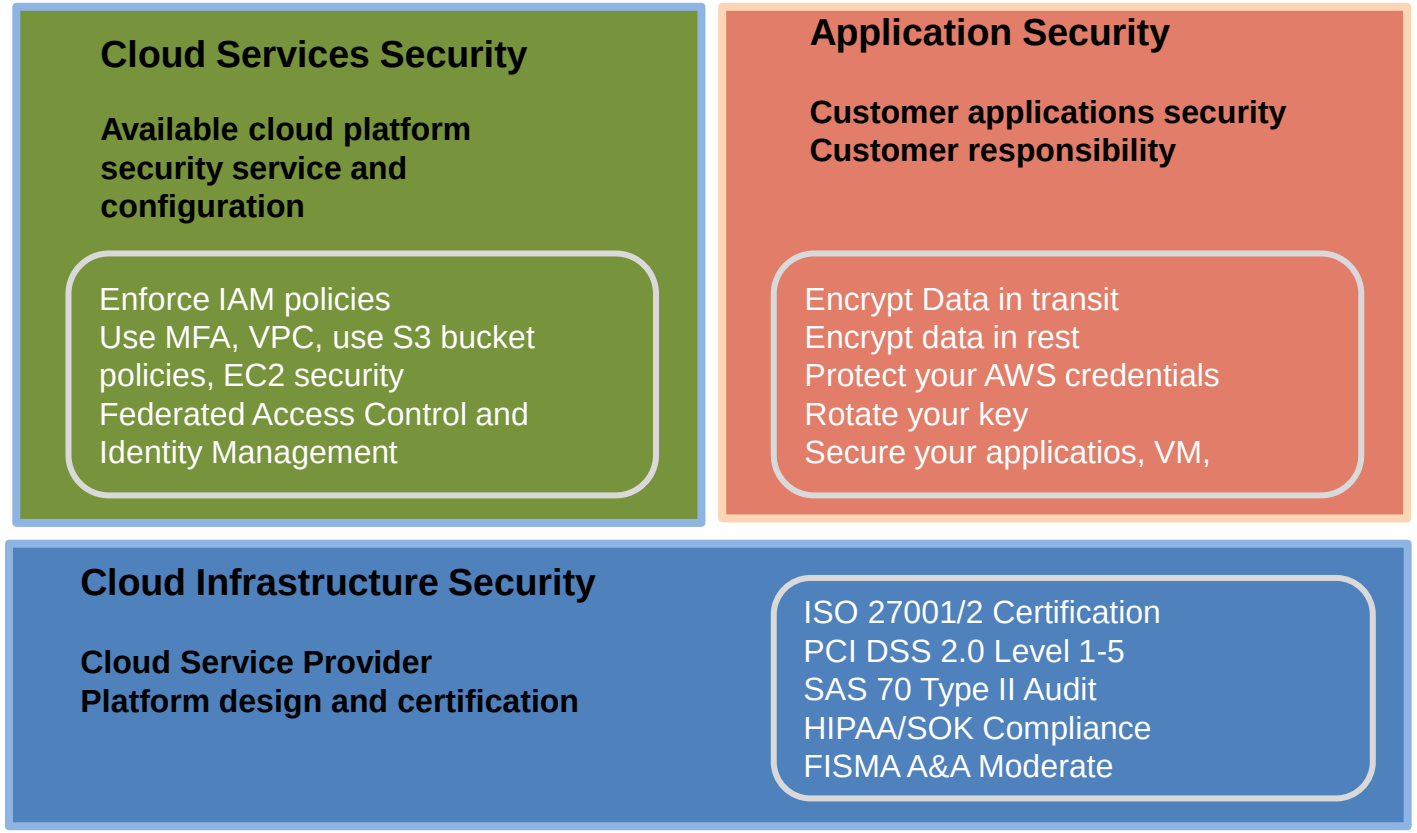
User access and security is ensured by the cloud based applications providers

Using cloud based services such as AWS Identity and Access Management (IAM) over Amazon EC2 and S3 infrastructure services

- For other cloud service models PaaS and SaaS the responsibility of AWS goes up to OS, network and firewall for PaaS, and also includes the application platform and container for SaaS.
  - However, the responsibility for data remains with the customer.

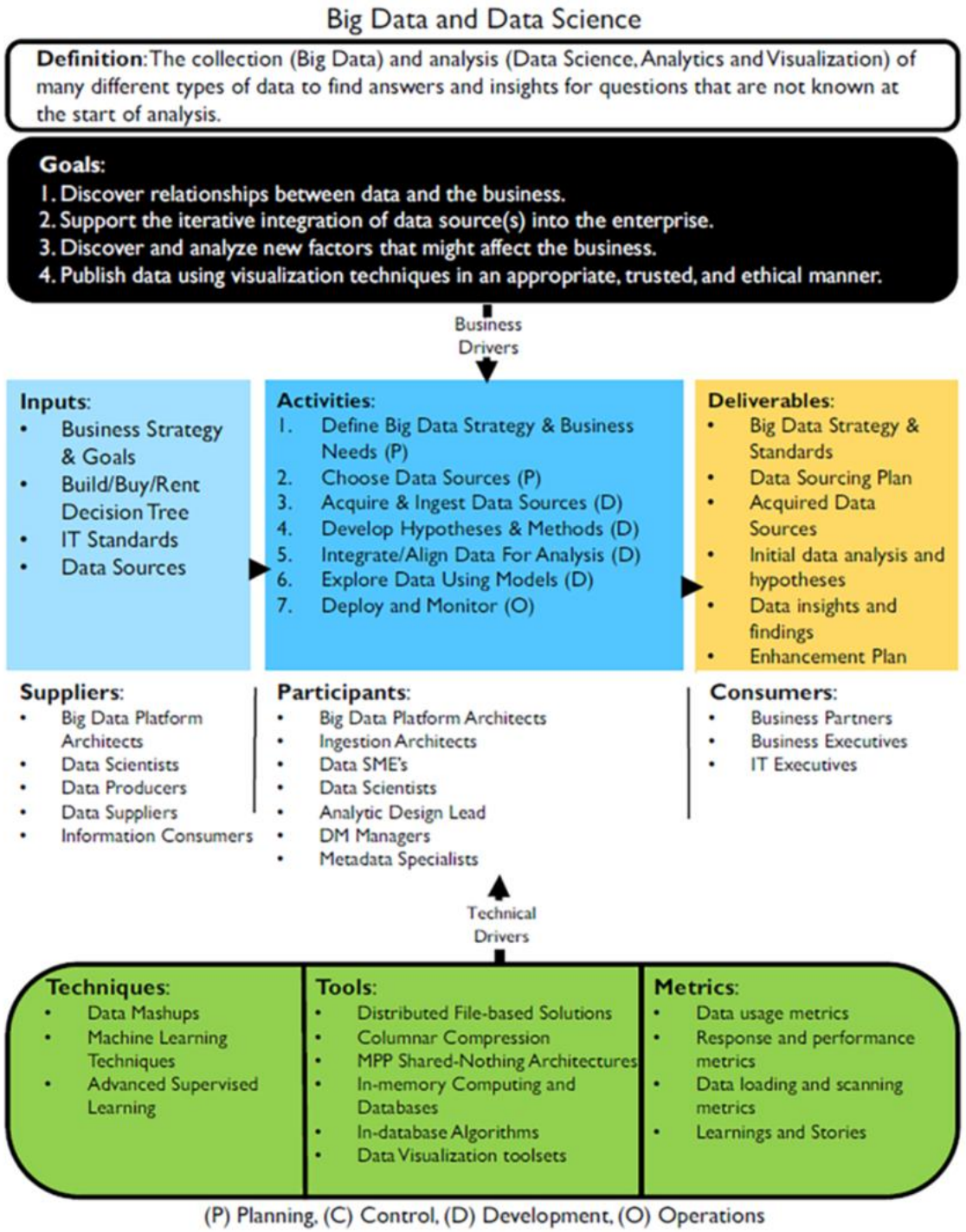
[ref] Todorov, D. & Ozkan, Y. (November 2013) 'AWS security best practices', Amazon Web Services [Online]. Available from: http://media.amazonwebservices.com/AWS_Security_Best_Practices.pdf

Amazon Web Services Security Model



Security is declared as one of critical importance to AWS cloud that is targeted to protect customer information and data from integrity compromise, leakage, accidental or deliberate theft, and deletion.

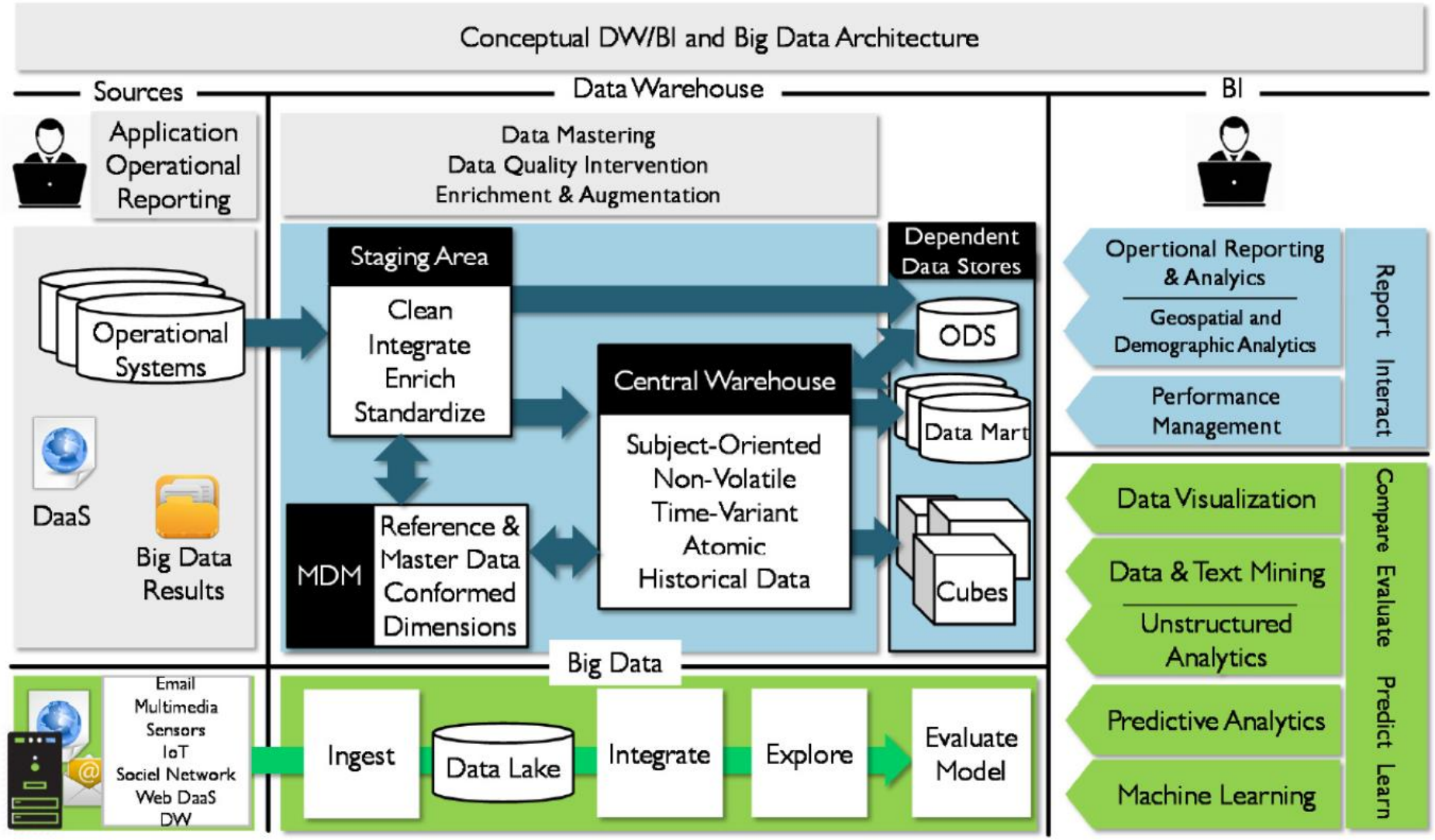
- The AWS infrastructure is designed with the high availability and sufficient redundancy to ensure reliable services operation.



Big Data and Data Science

- Big Data infrastructure
- Big Data technologies and platforms
- Data processing workflow management

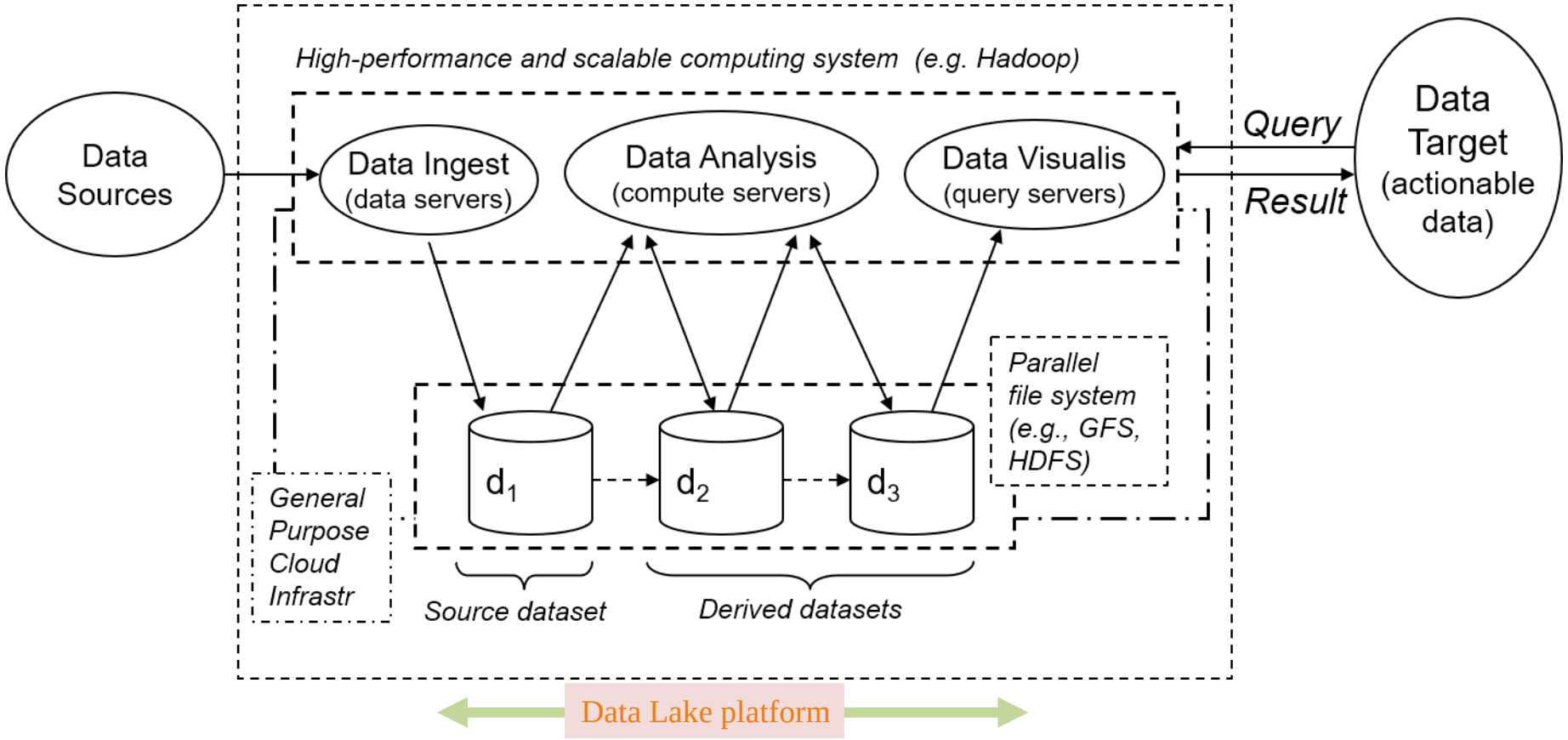
Big Data Infrastructure and Data Workflow (according to DMBOK)



Cloud Platform Benefits for Big Data

- Cloud deployment on virtual machines or containers
  - Applications portability and platform independence, on-demand provisioning
  - Dynamic resource allocation, load balancing and elasticity for tasks and processes with variable load
- Availability of rich cloud based monitoring tools for collecting performance information and applications optimisation
- Network traffic segregated and isolation
  - Big Data applications benefit from cloud based clusters: dynamic cluster resizing, load balancing, and other scale-out operations
  - Internal clouds network separates networks traffic for data and for management
  - Layer 2 and Layer 3 virtual networks inside user/application VPC
- Cloud tools for large scale applications deployment and automation
  - Provide basis for agile services development and Zero-touch services provisioning
  - Applications deployment in cloud is supported by major Integrated Development Environment (IDE)
  - Built-in interfaces for content distribution and mobile access

Cloud Based Big Data Services



Characteristics:

- Massive data and computation on cloud, queries from applications, store results
- **Full data lifecycle support: Data staging and format transformation**

Big Data Stack components and technologies

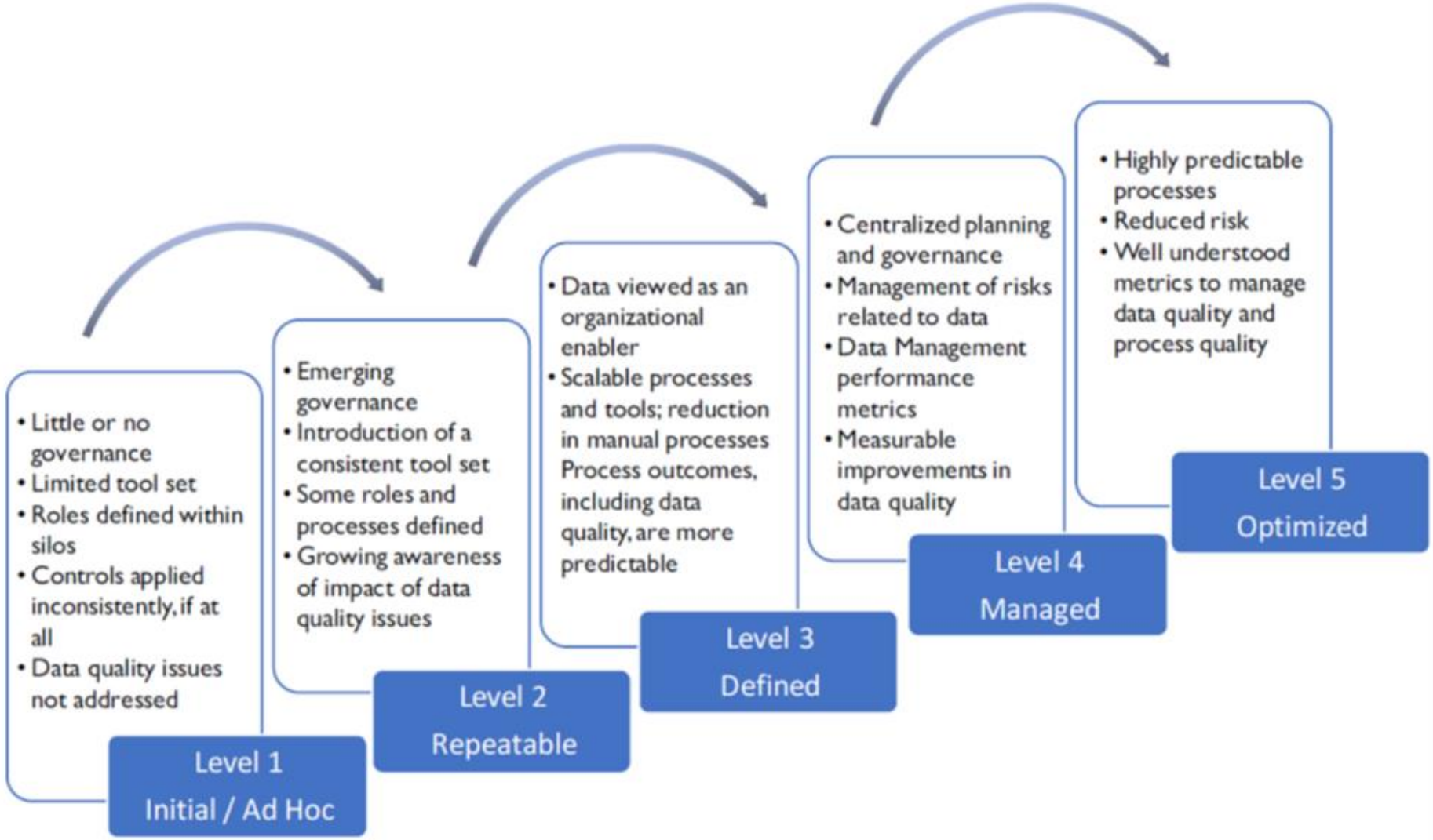
The major structural components of the Big Data stack are grouped around the main stages of data transformation

- **Data ingest:** Ingestion will transform, normalize, distribute and integrate to one or more of the Analytic or Decision Support engines; ingest can be done via ingest API or connecting existing queues that can be effectively used for handles partitioning, replication, prioritisation and ordering of data
- **Data processing:** Use one or more analytics or decision support engines to accomplish specific task related to data processing workflow; using batch data processing, streaming analytics, or real-time decision support
- **Data Export:** Export will transform, normalize, distribute and integrate output data to one or more Data Warehouse or Storage platforms;
- **Back-end data management, reporting, visualization:** will support data storage and historical analysis; OLAP platforms/engines will support data acquisition and further use for Business Intelligence and historical analysis.

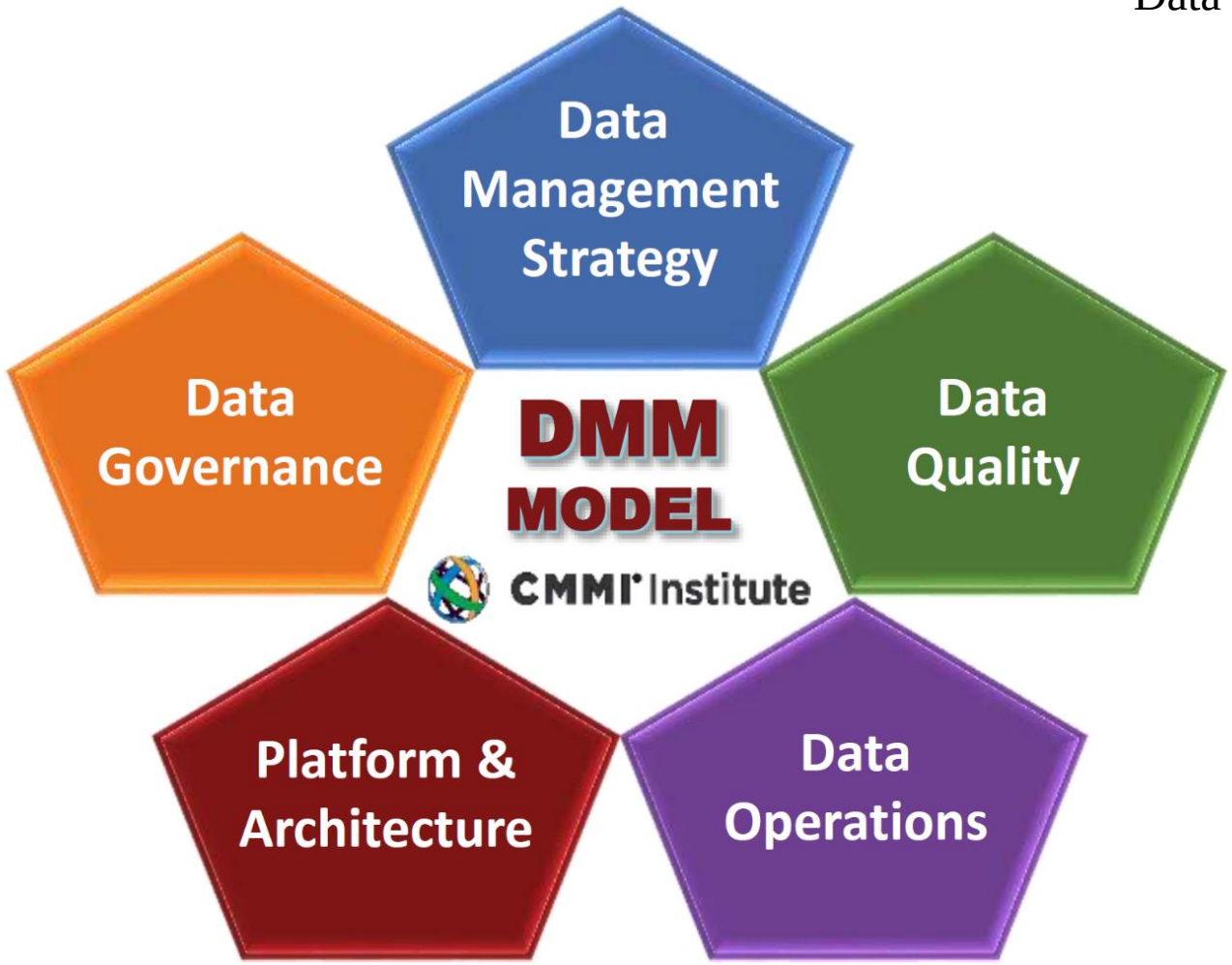
Organisational Big Data readiness and adoption levels (according to NIST SP1500, 2015)

Readiness Level	Adoption Level
1. No Big Data • No awareness or efforts around Big Data exist in the organization	**1. No Adoption** • No current adoption of Big Data technologies within the organization
2. Ad Hoc • Awareness of Big Data exists • Some groups are building solutions • No Big Data plan is being followed	**2. Project** • Individual projects implement Big Data technologies as they are appropriate
3. Opportunistic • An approach to building Big Data solutions is being determined • The approach is opportunistically applied, but is not widely accepted or adopted within the organization	**3. Program** • A small group of projects share an implementation of Big Data technologies • The group of projects share a single management structure and are smaller than a business unit
4. Systematic • The organizational approach to Big Data has been reviewed and accepted by multiple affected parties. • The approach is repeatable throughout the organization and nearly-always followed.	**4. Divisional** • Big Data technologies are implemented consistently across a business unit
5. Managed • Metrics have been defined and are routinely collected for Big Data projects • Defined metrics are routinely assessed and provide insight into the effectiveness of Big Data projects	**5. Cross-Divisional** • Big Data technologies are consistently implemented by multiple divisions with a common approach • Big Data technologies across divisions are at an organizational readiness level of Systematic or higher
6. Optimized • Metrics are always gathered and assessed to incrementally improve Big Data capabilities within the organization. • Guidelines and assets are maintained to ensure relevancy and correctness	**6. Enterprise** • Big Data technologies are implemented consistently across the enterprise • Organizational readiness is at level of Systematic or higher

DAMA Data Management Maturity Levels



CIMI Data Management Maturity Model

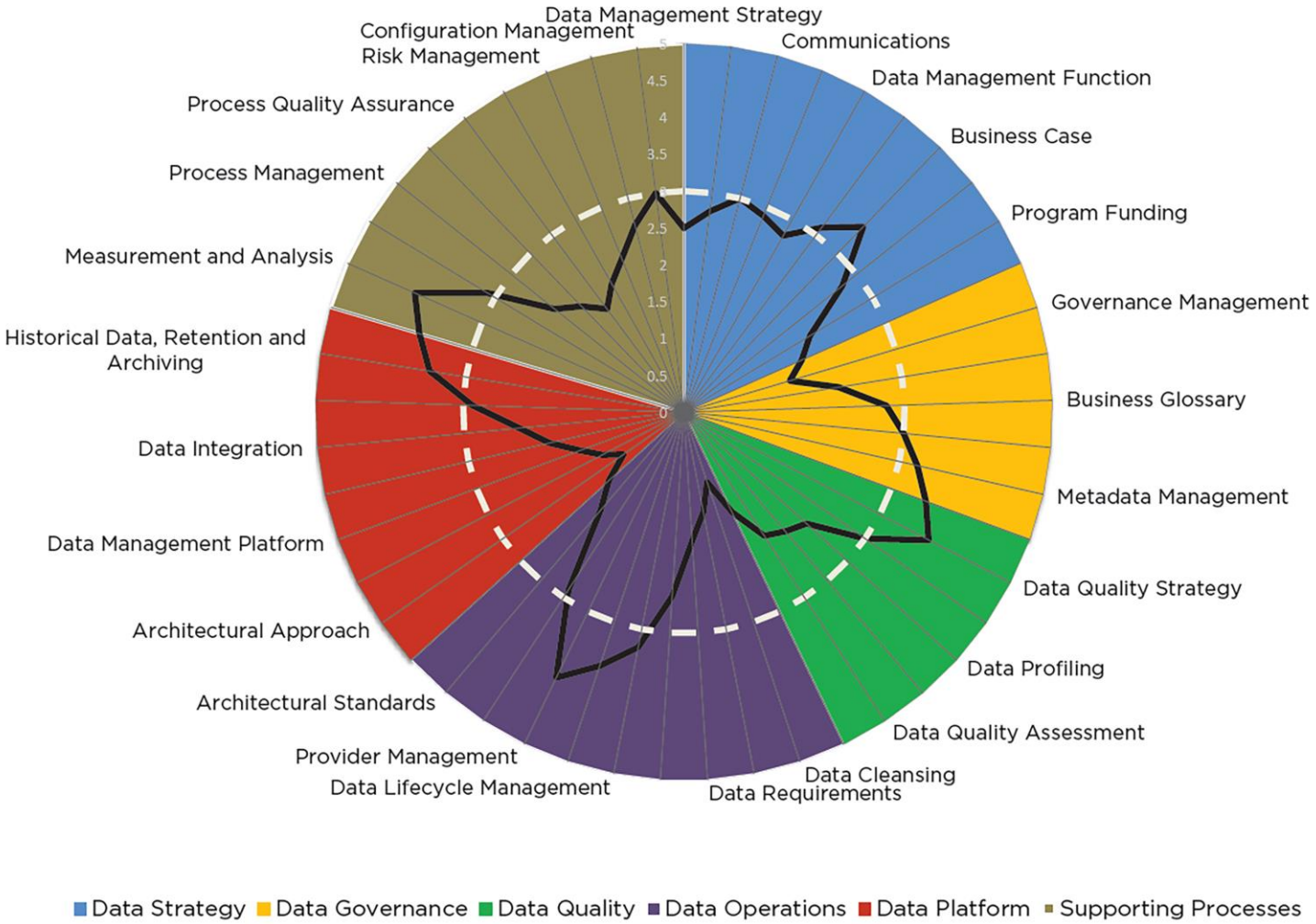


Data Management Maturity Dimensions

DATA MANAGEMENT STRATEGY	Data Management Strategy Communications Data Management Function Business Case Funding
DATA GOVERNANCE	Governance Management Business Glossary Metadata Management
DATA QUALITY	Data Quality Strategy Data Profiling Data Quality Assessment Data Cleansing
DATA OPERATIONS	Data Requirements Definition Data Lifecyle Management Provider Management
PLATFORM & ARCHITECTURE	Architectural Approach Architectural Standards Data Management Platform Data Integration Historical Data & Archiving
SUPPORTING PROCESSES	Measurement & Analysis Process Management Process Quality Assurance Risk Management Configuration Management

- https://cmmiinstitute.com/data-management-maturity (paid)

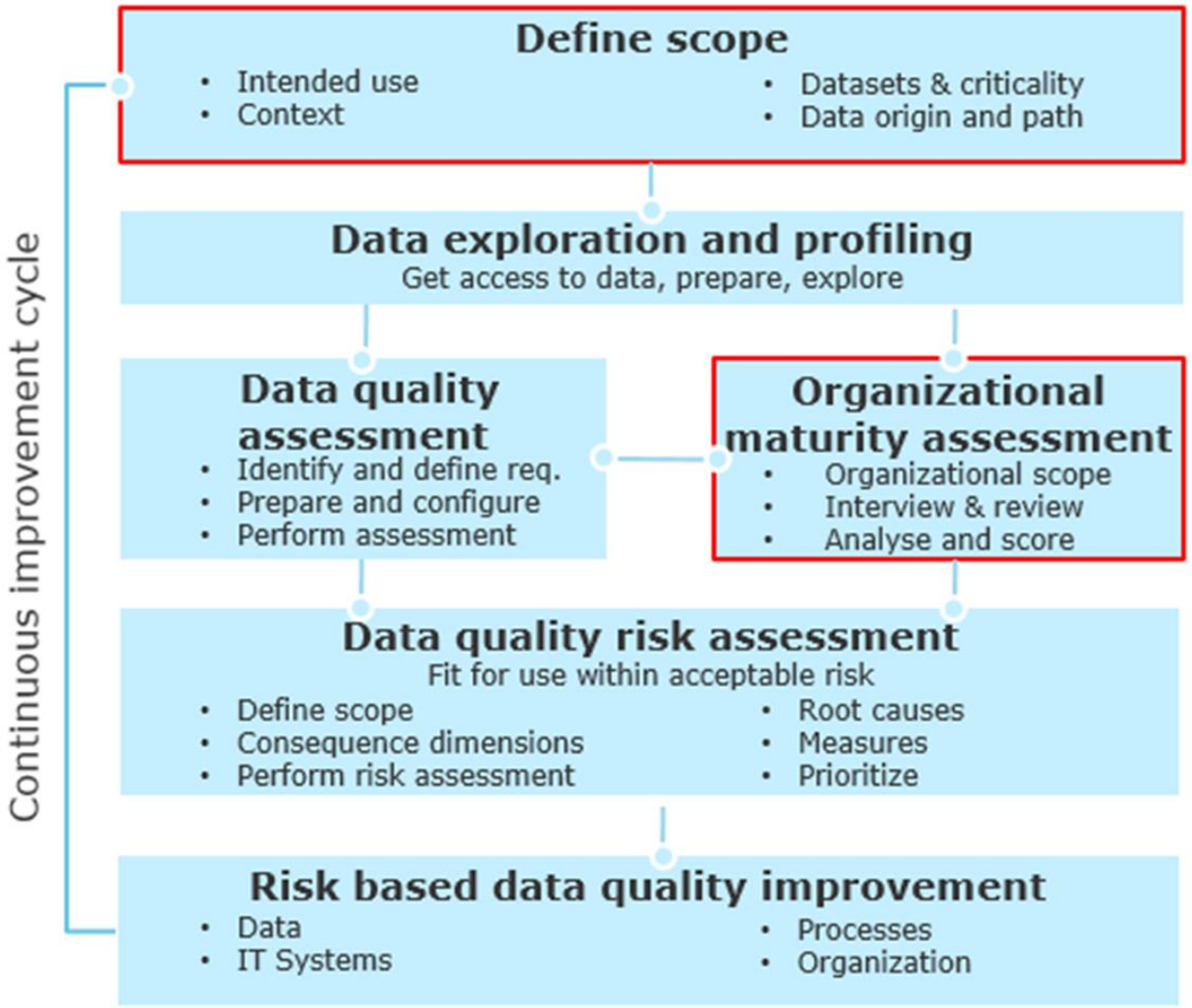
CMMI DMM Maturity Assessment



CMMI DMM Maturity Levels:

1. Performed: Processes are defined ad-hoc, primarily at the project level.
2. Managed: Processes are planned and executed according to policy
3. Defined: Set of standard processes is employed and consistently followed
4. Measured: Process metrics have been defined and used for data management
5. Optimised: Process performance is optimised based on measured metrics

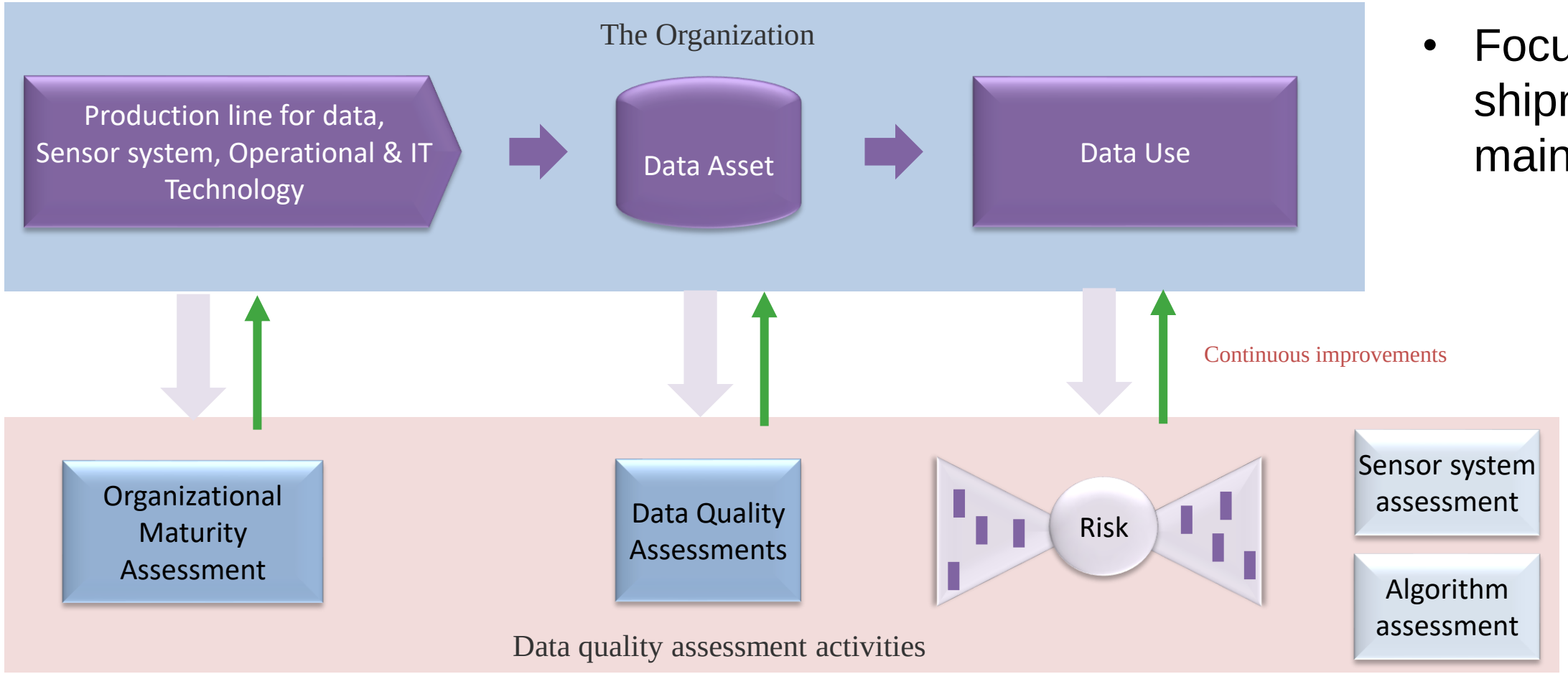
DNV-GL Data Quality Assessment Framework



- Links Data Quality and Risk assessment
  - Assessing incidents related to data quality
- Links to organisational Data Management Maturity

[ref] DNVGL-RP-0497 Data Quality Assessment Framework
http://rules.dnvgl.com/docs/pdf/dnvgl/rp/2017-01/dnvgl-rp-0497.pdf

Data Quality assessment Framework in use by DNV-GL



- Focus: Sea shipment, maintenance

[ref] DNVGL-RP-0497 Data Quality Assessment Framework http://rules.dnvgl.com/docs/pdf/dnvgl/rp/2017-01/dnvgl-rp-0497.pdf

Mind the gap, manage the TO-DO:

Maturity level	Governance	Organization and people	Processes	Process Efficiency	Requirement definition	Metrics and dimensions	Architecture, tools and technologies	Data standards
LEVEL 5 - Optimized	Data management policies governs and drives improvements	Data management board oversees improvement activities	Processes for continuous improvement in place	Processes provides feed-back and feed-forward to support continuous improvement	Baseline established and improvements measured according to requirements	Metrics defines baseline to support continuous improvement	Tools support policy driven continuous improvement cycle	Standard compliance and domain models are subject to continuous improvement
LEVEL 4 – Managed	Policies defined in relation to business objectives	Skillset extended to include risk analysis of quality issues aligned with business objectives	Processes for impact analysis and risk mgmt. in place	Monitoring is performed across enterprise and published as KPI's and trends	Requirements are linked to business impacts	Metrics are linked to business impacts and risk analysis	Tools are driven by business objectives and include support for root cause analysis and risk mgmt.	Standards are used actively to reduce risk for critical business operations
LEVEL 3 – Defined	Policies defined at enterprise level	Roles and required skills defined at enterprise level	Processes are defined and implemented consistently across enterprise	Defined metrics are monitored in advance of business impact	Requirements defined and communicated at enterprise level	Framework for metrics and dimensions defined at enterprise level	Architecture in place at enterprise level supporting full stack data management	Standards, domain models and semantics used at enterprise level
LEVEL 2 – Repeatable	Local initiatives address the requirement for policies	Locally defined roles and some basic skills	Best practices in place but not used consistently	Generic metrics are monitored at point of impact	Local initiatives define requirements	Metrics are reused locally in projects	Tools and technologies used consistently in selected projects	Industry standards and domain models used selectively across projects
LEVEL 1 - Initial	Only ad-hoc or temporal policies in place	No formally defined roles or skillset	Ad-hoc or reactive responses to quality issues	No baseline and no monitoring of quality issues	Re-engineering used to derive requirements	Project specific metrics	Tools are used ad-hoc per project	Ad-hoc and inconsistent use of standards
Objectives,	Policy, Culture, Awareness, Risks, Capabilities to handle DQ issues	Organization, roles, responsibilities, authority, skillsets	Structured and vetted ways of handling and preventing DQ issues	Measure, monitor and use metrics to mitigate DQ issues	DQ Requirements defined, communicated and acted upon	DQ metrics defined, setup, measured and monitored	DQ Tools for processing, analysing and correcting DQ issues with data assets	Use available standards, models, ontologies and taxonomies – a corporate «DQ language»

[ref] DNVGL-RP-0497 Data Quality Assessment Framework http://rules.dnvgl.com/docs/pdf/dnvgl/rp/2017-01/dnvgl-rp-0497.pdf

Mind the gap, manage the TO-DO:

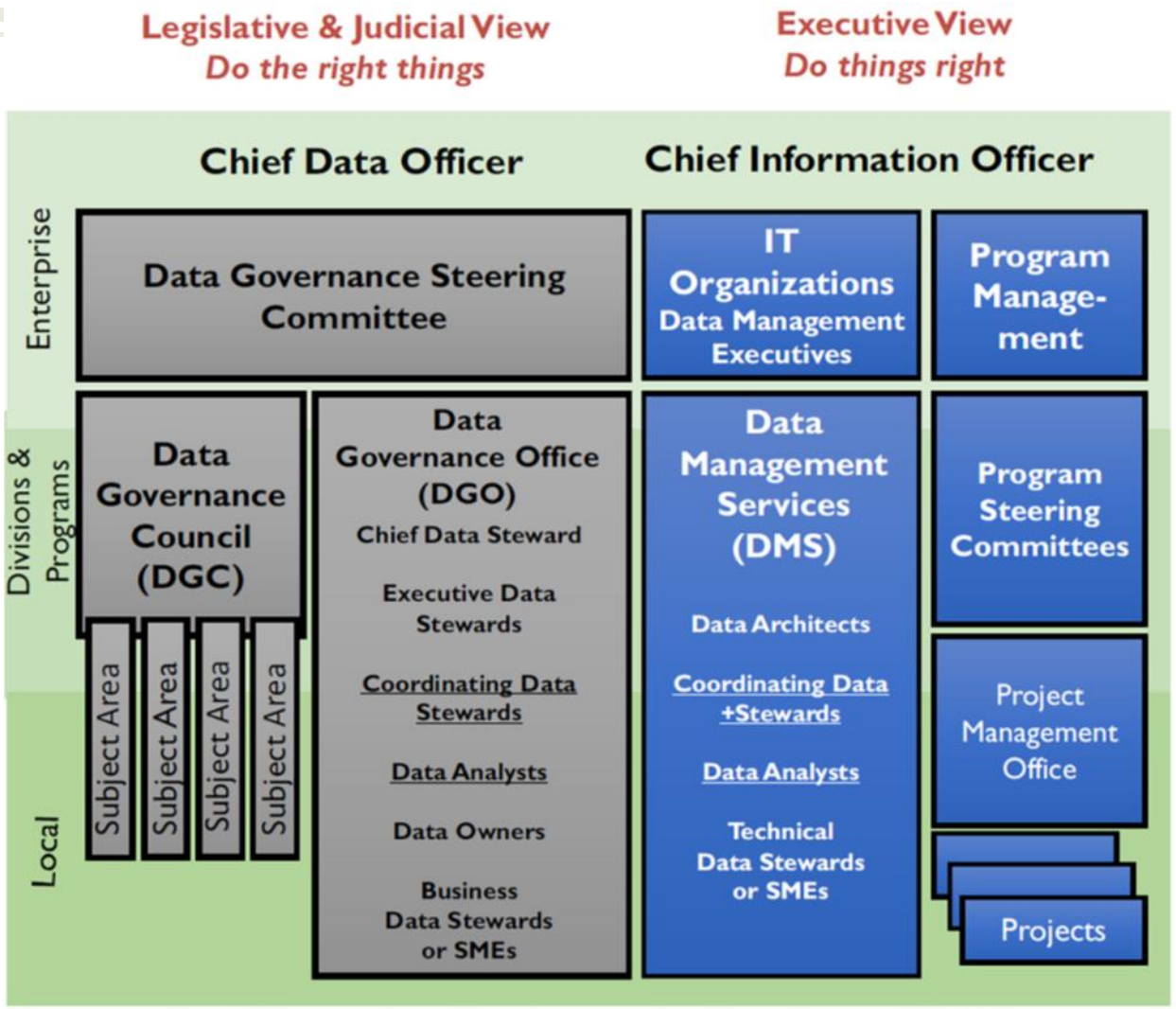
Maturity level	Governance	Organization and people	Processes	Process Efficiency	Requirement definition	Metrics and dimensions	Architecture, tools and technologies	Data standards
LEVEL 5 - Optimized	Data management policies governs and drives improvements	Data management board oversees improvement activities	Processes for continuous improvement in place	Processes provides feed-back and feed-forward to support continuous improvement	Baseline established and improvements measured according to requirements	Metrics defines baseline to support continuous improvement	Tools support policy driven continuous improvement cycle	Standard compliance and domain models are subject to continuous improvement
LEVEL 4 – Managed	Policies defined in relation to business objectives	Skillset extended to include risk analysis of quality issues aligned with business objectives	Processes for impact analysis and risk mgmt. in place	Monitoring is performed across enterprise and published as KPI's and trends	Requirements are linked to business impacts	Metrics are linked to business impacts and risk analysis	Tools are driven by business objectives and include support for root cause analysis and risk mgmt.	Standards are used actively to reduce risk for critical business operations
LEVEL 3 – Defined	Policies defined at enterprise level	Roles and required skills defined at enterprise level	Processes are defined and implemented consistently across enterprise	Defined metrics are monitored in advance of business impact	Requirements defined and communicated at enterprise level	Framework for metrics and dimensions defined at enterprise level	Architecture in place at enterprise level supporting full stack data management	Standards, domain models and semantics used at enterprise level
LEVEL 2 – Repeatable	Local initiatives address the requirement for policies	Locally defined roles and some basic skills	Best practices in place but not used consistently	Generic metrics are monitored at point of impact	Local initiatives define requirements	Metrics are reused locally in projects	Tools and technologies used consistently in selected projects	Industry standards and domain models used selectively across projects
LEVEL 1 - Initial	Only ad-hoc or temporal policies in place	No formally defined roles or skillset	Ad-hoc or reactive responses to quality issues	No baseline and no monitoring of quality issues	Re-engineering used to derive requirements	Project specific metrics	Tools are used ad-hoc per project	Ad-hoc and inconsistent use of standards
Objectives,	Policy, Culture, Awareness, Risks, Capabilities to handle DQ issues	Organization, roles, responsibilities, authority, skillsets	Structured and vetted ways of handling and preventing DQ issues	Measure, monitor and use metrics to mitigate DQ issues	DQ Requirements defined, communicated and acted upon	DQ metrics defined, setup, measured and monitored	DQ Tools for processing, analysing and correcting DQ issues with data assets	Use available standards, models, ontologies and taxonomies – a corporate «DQ language»

[ref] DNVGL-RP-0497 Data Quality Assessment Framework http://rules.dnvgl.com/docs/pdf/dnvgl/rp/2017-01/dnvgl-rp-0497.pdf

Data Management and Organisational Roles

- Data Governance Office
- Data Stewards
- Data Science/Data Management team

DMBOK: Data Governance Organisation Parts



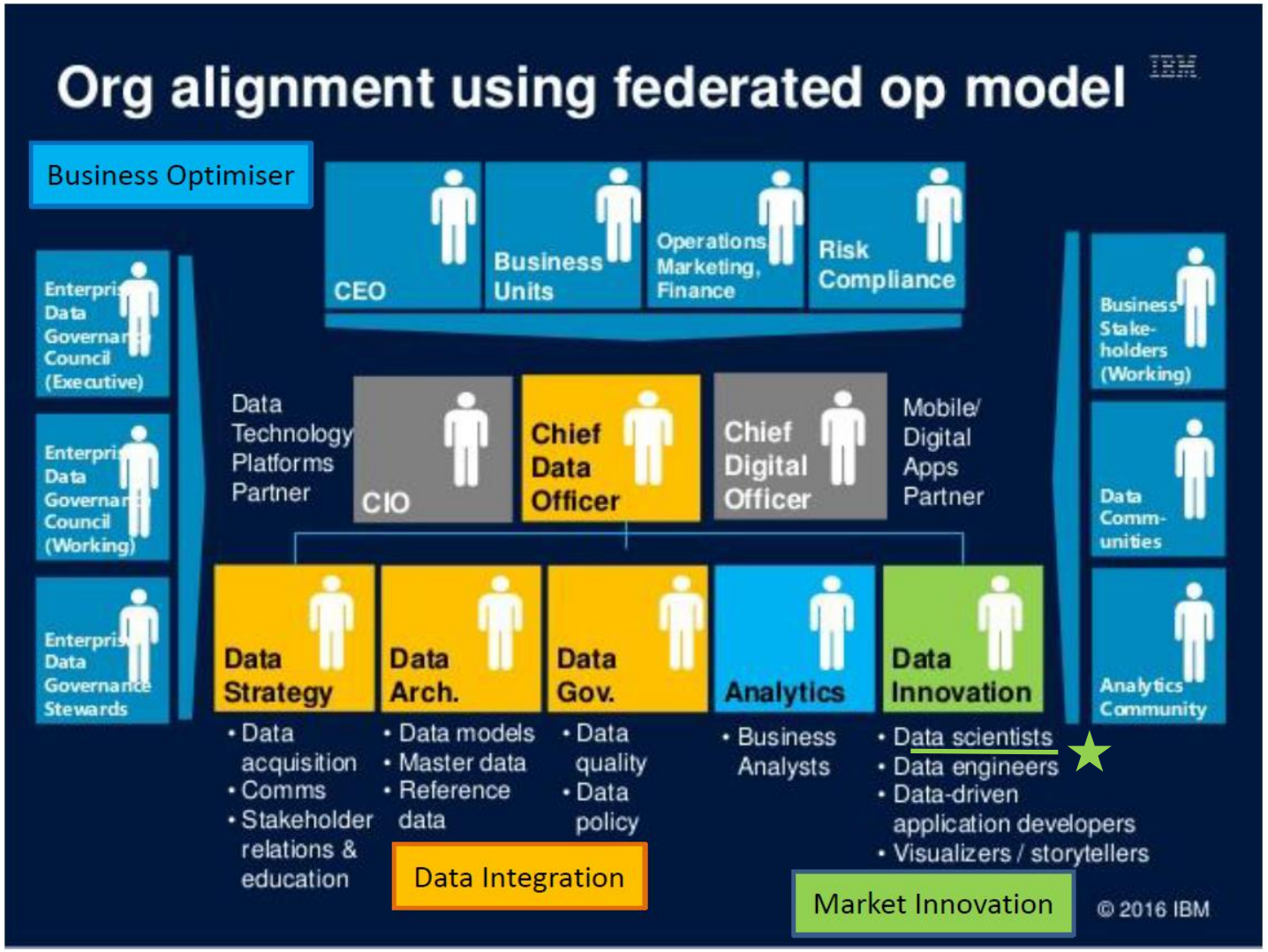
- Separation of governance responsibilities
- Multi-layer
- CDO
- CIO
- Councils

Data Governance Office (DGO)

- Chief Data Steward
- Executive Data Steward
- Business Data Steward or SME

EXAMPLE of IBM definition of new organisational roles

[ref] Cortnie Abercrombie, What CEOs want from CDOs and how to deliver on it (2016) [online]
http://www.slideshare.net/IBMBDA/what-ceos-want-from-cdos-and-how-to-deliver-on-it



Data Governance Office (DGO)

- Chief Data Steward
- Executive Data Steward
- Business Data Steward or SME

Data Stewardship (according to DM-BOK)

- **Creating and managing core Metadata**: Definition and management of business terminology, valid data values, and other critical Metadata.
- **Documenting rules and standards**: Definition/documentation of business rules, data standards, and data quality rules.
  - High quality data are often formulated in terms of rules rooted in the business processes that create or consume data.
  - Stewards help surface these rules and ensure their consistent use.
- **Managing data quality issues**: Stewards are often involved with the identification and resolution of data related issues or in facilitating the process of resolution.
- **Executing operational data governance activities**: Stewards are responsible for ensuring that, day-today and project-by-project, data governance policies and initiatives are adhered to. They should influence decisions to ensure that data is managed in ways that support the overall goals of the organization.

"Best Data Steward is not made but found" DMBOK1 (2009)

Data Steward

https://www.dataversity.net/what-is-data-stewardship/

- DAMA DMBOK
- “The most common label to describe accountability and responsibility for data and processes that ensure effective control and use of data assets. Stewardship can be formalized through job titles and descriptions, or it can be a less formal function driven by people trying to help an organization get value from its data.”

- According to the Data Governance Institute:
- “Data Stewardship is concerned with taking care of data assets that do not belong to the stewards themselves. Data Stewards represent the concerns of others. Some may represent the needs of the entire organization. Others may be tasked with representing a smaller constituency: a business unit, department, or even a set of data themselves.”

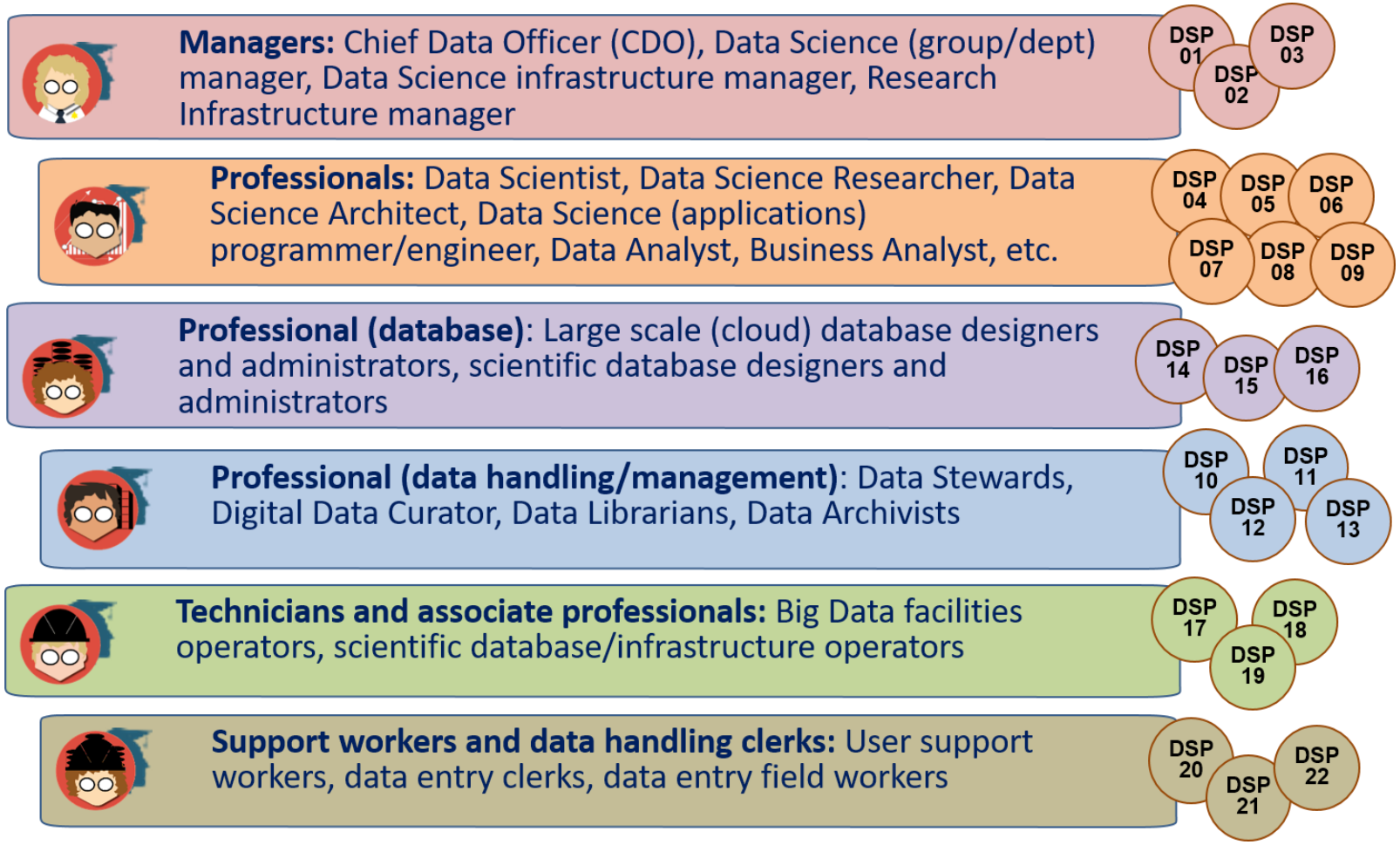
Data Stewardship in Research and FAIR Principles

- FAIR Initiative by Dutch Techcentre for Life Science (DTLS) – Prof. Barend Mons
  - Supported by Germany and France
  - Part of Horizon 2020 Programme
- FAIR Principles for research data:
  Findable – Accessible – Interoperable - Reusable
- Data Stewards as a key bridging role between Data Scientists as (hard)core data experts and scientific domain researchers (HLEG EOSC report)
- Current definition of the Data Steward (part of Data Science Professional profiles)
  - Data Steward is a data handling and management professional whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation.
  - Data Steward creates data model for domain specific data, support and advice domain scientists/ researchers during the whole research cycle and data management lifecycle.

HLEG report on European Open Science Cloud

(October 2016)

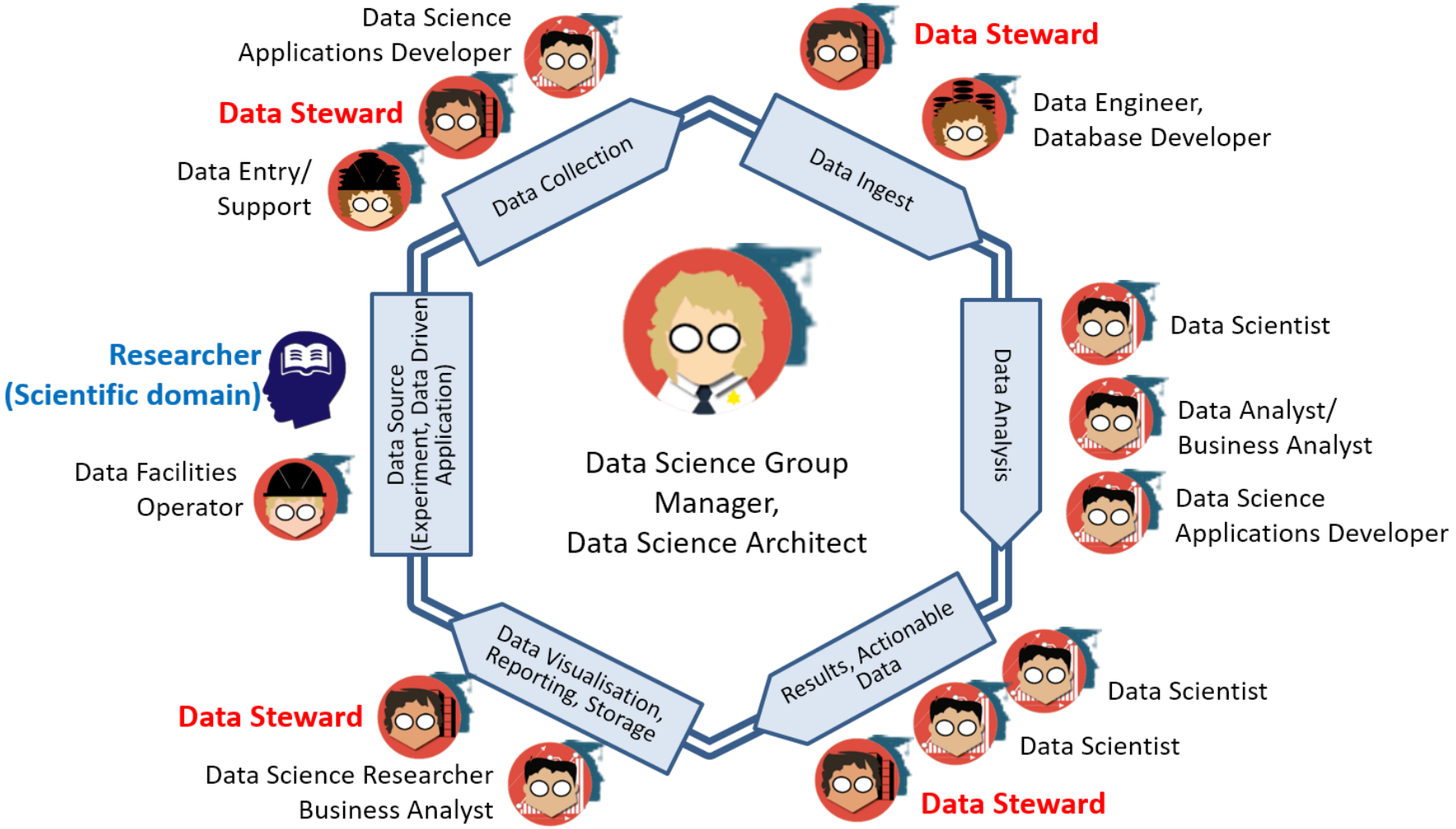
Why we need to view the whole Data Science Professional Family?



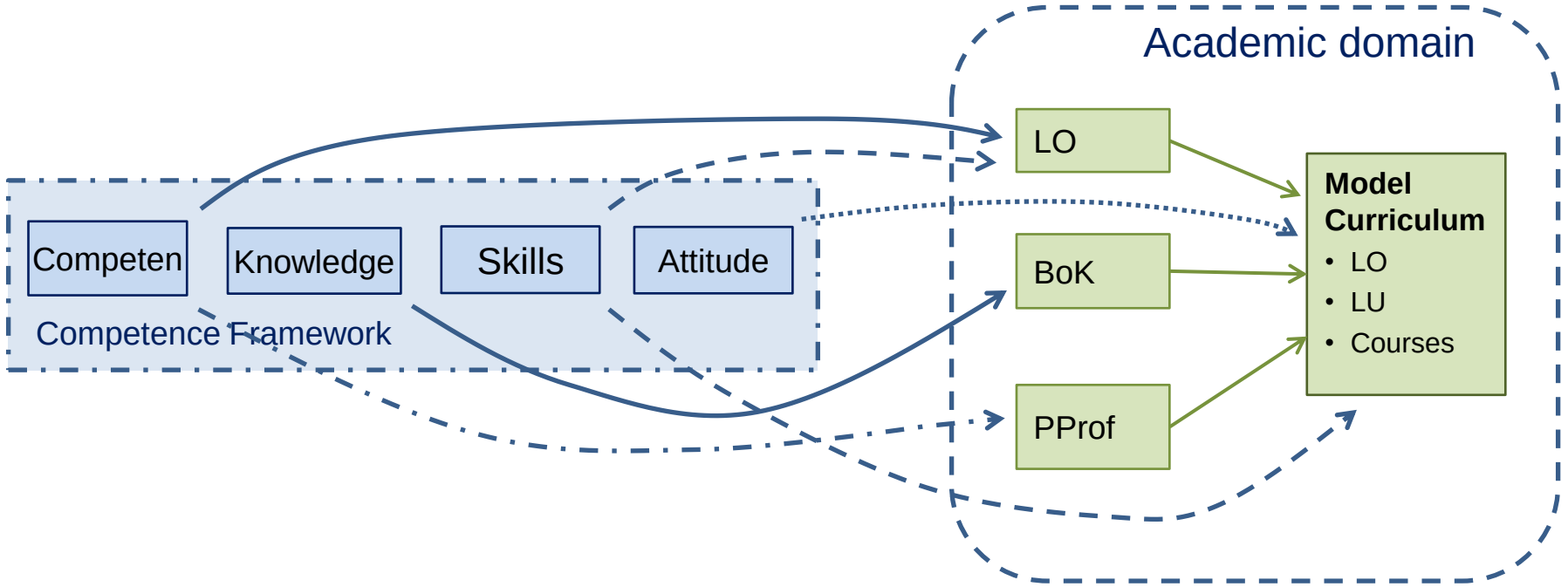
EDISON Data Science Professional Profiles family EDSF, Part 4

- Career path
- Team composition
- Agile Team working
- Education programs alignment
- HR capacity building and management
- Employability/ mobility for job seekers
- Easy to market

Research Cycle and Data Science Team



Relations between Competences, Skills, Knowledge/BoK, Professional Profiles

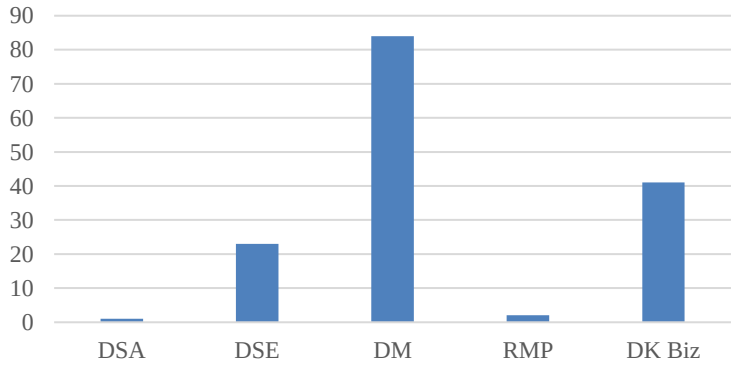


- Must be implemented in a competence framework

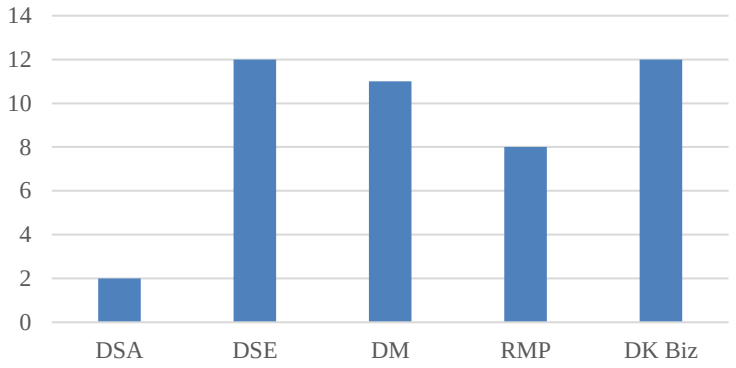
Vacancies profile – By Data Science Competence Groups

Wide range of Competences: Responsibility, Functions, activities

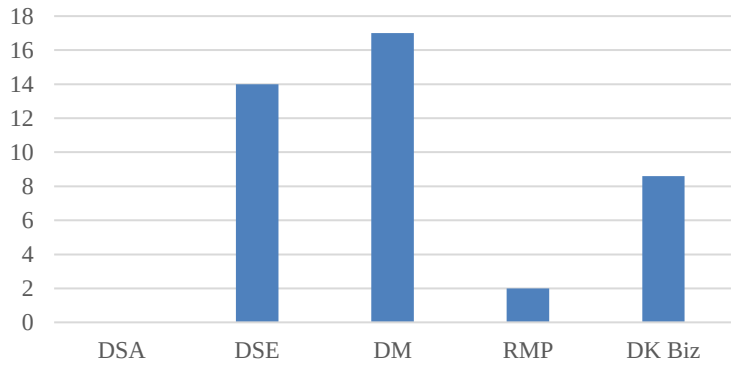
Functions/Abilities - Competences



Knowledge topics



Required Experience/skills



DSA – Data Science and Analytics
DSE – Data Science Engineering
DM – Data Management and Governance
RMP – Research Methods and Project Management
DK Biz – Domain Knowledge, particular Business domain

Important Knowledge Items extracted from Job vacancies (indeed.com – NL, DE, UK, US, Sept 2020)

- Data Management techniques
- FAIR data principles
- Data Management and Data Governance principles
- Data integrity
- Metadata, PID and linked data
- Ontology and Semantics
- FAIR metrics and Maturity framework, FAIR certification
- Data compliance regulations and standards
- Data privacy law
- GDPR
- Ethics
- Research methods
- Project management
- Business process management
- Marketing
- Banking financial services and data management
- Multilevel Bill of Materials
- Data Warehouses
- Version control system
- Master Data Management (MDM) and Reference Data
- Data analysis and visualisation tools
- Data lifecycle, lineage, provenance
- Visual Basic for Applications (VBA) and interface design
- WebAPI use for data access, collection and publishing
- DevOps, Agile, Scrum methods and technologies
- Data formats, standards
- Data modeling (SQL and EDBMS, NoSQL)
- Modern data infrastructure: Data registries, Data Factories, Semantic storage, SQL/NoSQL

Summary and Takeaway

- DAMA DMBOK provide conceptual and methodological base for establishing consistent Data Governance and Data Management Framework based on industry standards and best practices
- Data are becoming important asset for organisations and source of the future growth and operations optimisation
- Consistent and mature Data Governance require agile and manageable Enterprise Data Infrastructure capable of handling modern (big) data
- Data Quality Management is essential component and stage in Digital Transformation

Discussion Questions

- Data Governance and Data Management processes look so complex and formal. Do we really need to put so much efforts to it? How does it help our company?
- Do we need to follow all industry standards?
- What is the purpose of all these admin roles in DGov&DMngnt?
- DGov&DMngnt are horizontal activities in an organisation. How can they be done in hierarchical organisation model?
- How DGov&DMngnt are related to IT management? How to achieve active cooperation.

Mentimeter Question 11.01

- What aspects of data management are addressed in your company (to best of your knowledge)?

Mentimeter Question 11.02

- What problems with data management you are experiencing in your company (based on your experience)?

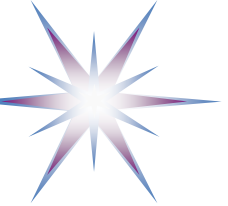
Acknowledgement

- This work is supported by the ERASMUS+ MATES project

- The work is committed to the Open Source under Creative Commons 4.0 CC BY License

Co-funded by the
Erasmus+ Programme
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



SQL

1. It’s the Industry Standard.
2. It’s Semantically Mature.
3. It’s a Living Language.
4. It’s Simple and Effective.
5. It’s Open Source.
6. It’s Secure.
7. It’s Extensible.
8. It’s Logically Founded Upon Relational Theory.
9. It’s Useful Everywhere.
10. Nobody Has Come up With a Better Database Language.