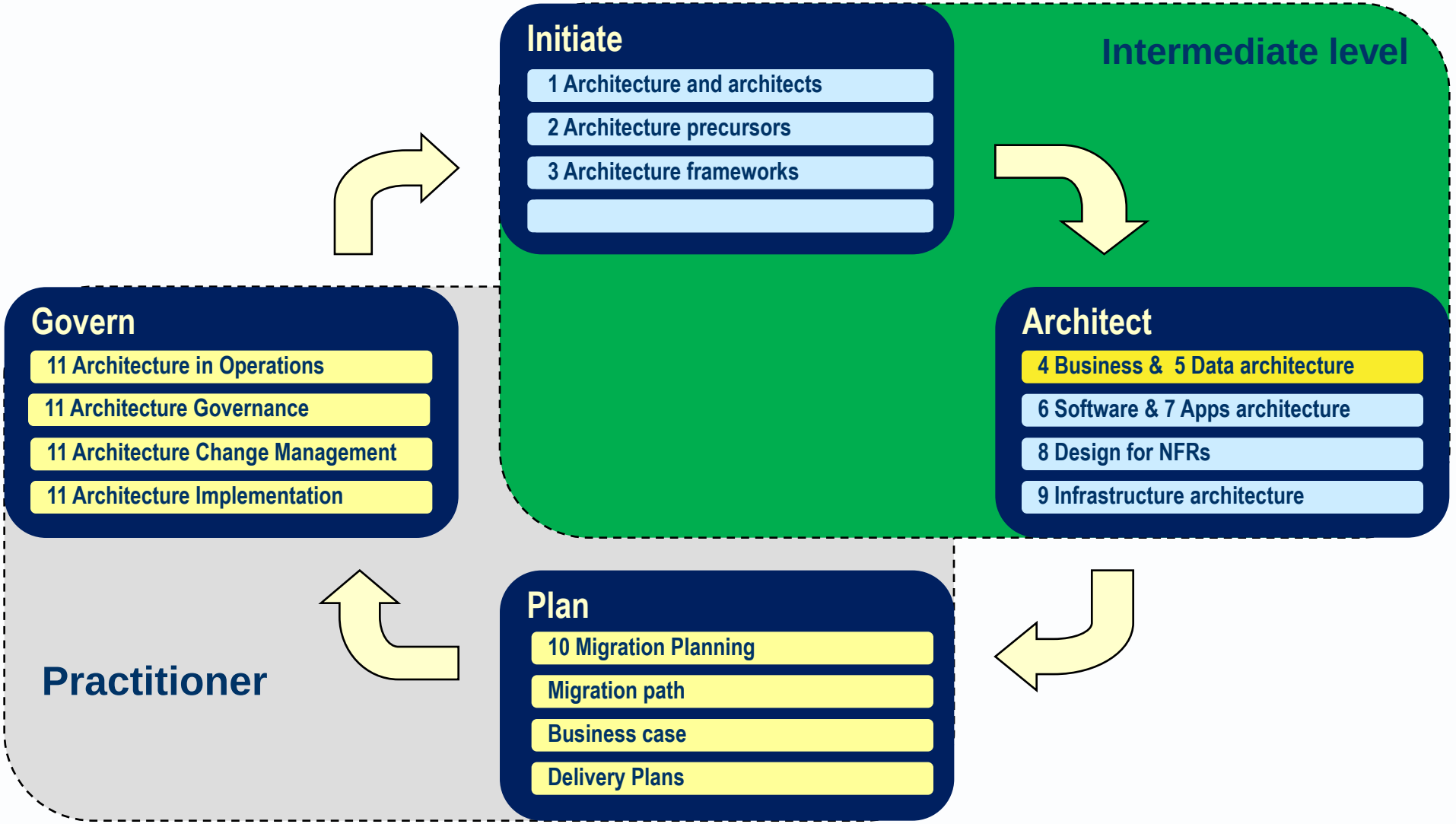
Avancier Reference Model

Data Architecture (ESA 5)

5. Data architecture



5.1 Foundation (not to be examined)

	Data in motion	Data at rest
Data object	Data event	Data entity
Data container	Data flow	Data store

Information

- Meaning created and found by actors in a data structure.
- The meanings can include descriptions, decisions, directions and opinions.

Data

- Representations of information in a structure or medium of any kind.
- It can be textual, numerical, graphical, pictorial, video, audio, biochemical etc..

Structured data

- [A data object] a data store or flow that fits a pre-defined data type or structure.
- It is composed of data items.
- It usually records real word entities or events.
- It may contain references to unstructured data.
- (All popular architecture frameworks focus on structured data.)

Unstructured data

- [A data object] containing text or images that do not fit a pre-defined data type or structure.
- E.g. emails, voice and video.
- However, it may contain recognisable structured items.

Meta data

- [A description] of data.
- A data structure, data type, constraint rule, derivation rule or other data quality.

Data structure

- [A type] a structure that arranges data items in one or more groups

Data type

- [A type] that defines the properties shared by instances of a data item or data entity.
- It defines the possible values for that type, the processes that can be performed on values of that type, the meaning of the data; and the way values of that type can be stored.

Data dictionary

- [An artifact] that catalogues data types and defines their meanings.
- It may include business rules in the form of constraints on data values and derivation rules.
- It may take the form of a canonical data model.

5.2 Data at rest (in store)

Data store

- A persistent data structure accessible by software.
- Traditionally stored on discs, increasingly on solid state drives.

- Hard Disc Drive (HDD)
  - Metal platters with a magnetic coating.
  - A read/write head on an arm accesses data while platters are spinning
- Solid State Drive (SSD)
  - Interconnected flash memory chips
  - Either in a computer or a box connected to one
  - Optimised for data written few times and read many times

- To find out what may replace flash memory try this:
- http://www.computerweekly.com/feature/Whats-wrong-with-flash-storage-And-what-will-come-after

Data server

- [A technology component] that hosts a database or file management system
- It enables a data store to be accessed by applications.
- It uses a particular physical data schema.

Physical data schema

- [A data structure] in the format required by a particular data server
- It may realise a whole logical data model, or part of one.
- It may be maintained by one or more application components.

- Transactional database
- Data warehouse
- Document store
- Big data store

} Sometimes called NoSQL databases, but do need some kind of SQL

OLTP v OLAP

- **Transactional (OLTP) data schema**
  - optimised for transaction processing, often a relational database.
  - It usually enables direct access to any data entity instance, using its primary key.
  - **Normalisation**
    - [A technique] commonly applied to the data structure of an OLTP data store.
    - It stores each fact once, to ensure data integrity, and speed up data update processes.

- **Data warehouse (OLAP) data schema**
  - optimised for the production of management information reports.
  - **De-normalisation**
    - [A technique] commonly applied to the data structure of an OLAP data store.
    - It duplicates some data storage to speed up data analysis and reporting processes

Document schema

- [A data schema] optimised for the storage of forms and documents.
- A document is stored in the structure it is created, and not modified thereafter.
- It is typically an aggregate data structure XML or JSON format.
- Changes to data values must be entered on new documents.

Big data schema

- [A data schema] that is characterised by a
  - high volume of data
  - high velocity of data capture
  - wide variety of data content.

- Store whatever (unstructured or structured) data is captured
- Data may be glued together later

Big data store features include

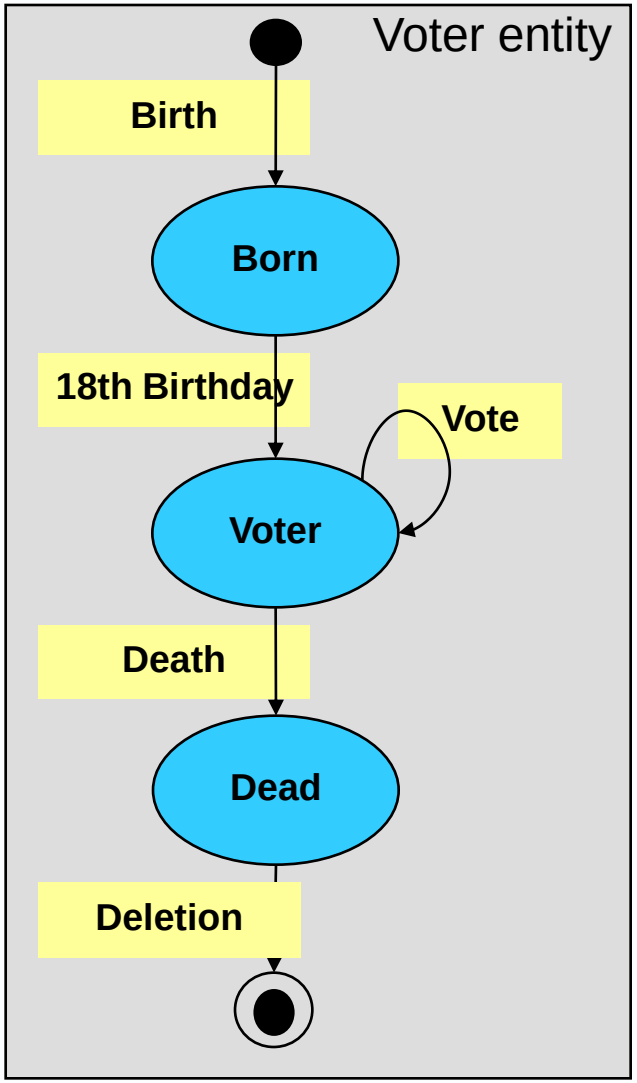
- **Search and discovery**
  - [a feature] that gathers information from diverse data sources.
- **Map and reduce**
  - [a feature] that compresses a data structure into a set of key/value pairs (aka tuples).
- **Sharding**
  - [a feature] that divides the population of a data entity type across (potentially thousands of) physically data servers.
- **Advanced analytics**
  - [a feature] that identifies patterns in data.

Data entity

- [A data object] that represents a thing or event an expert recognises as important in their domain of knowledge.
- A data entity instance is identified using primary key composed of one or more attributes.
- It can be related to other data entities in a larger data structure.
- The data structure may be normalised, but does not have to be.

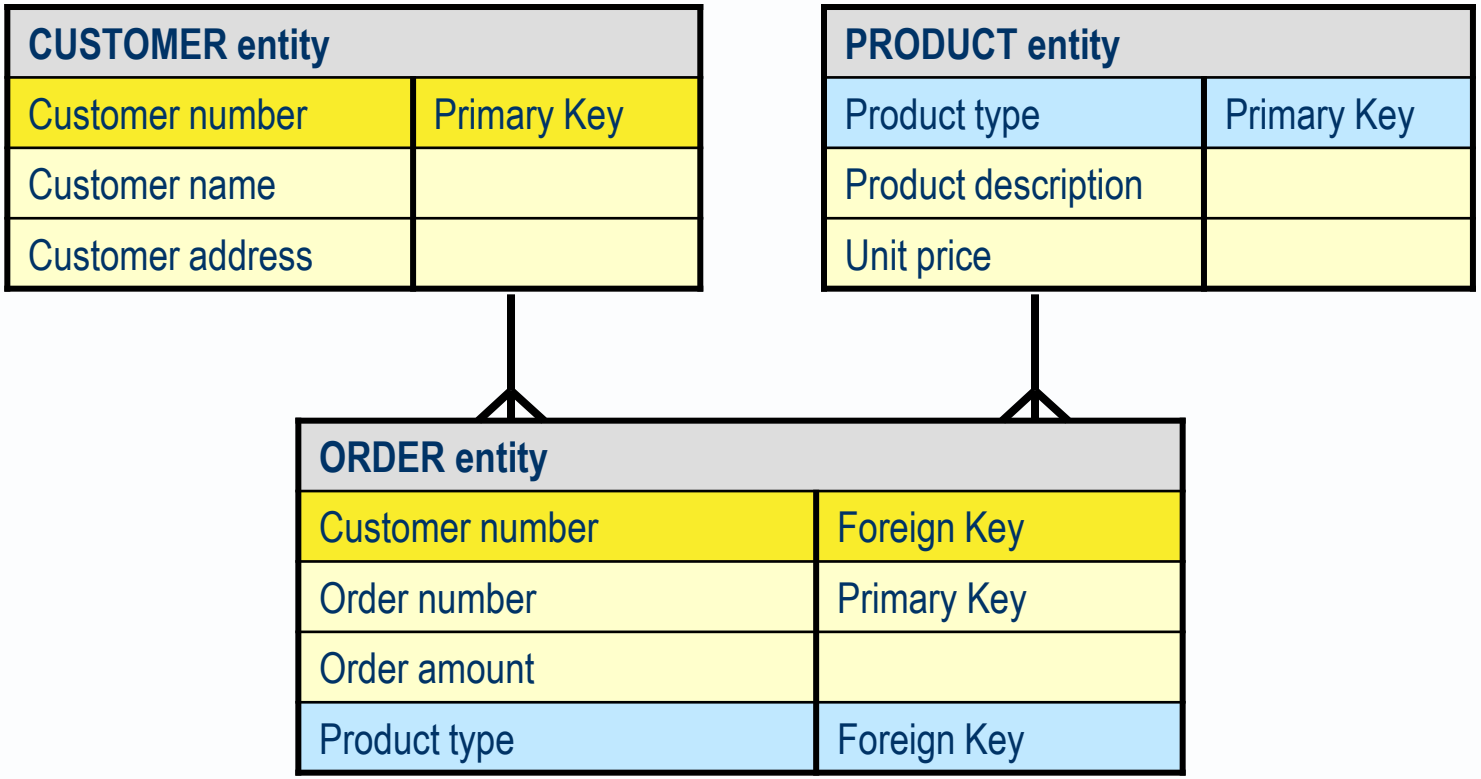
Data entity lifecycle diagram

- [An artifact] that shows life of a data entity in terms of
- states it passes (through from creation to deletion) and
- events that trigger state transitions.



Logical data model

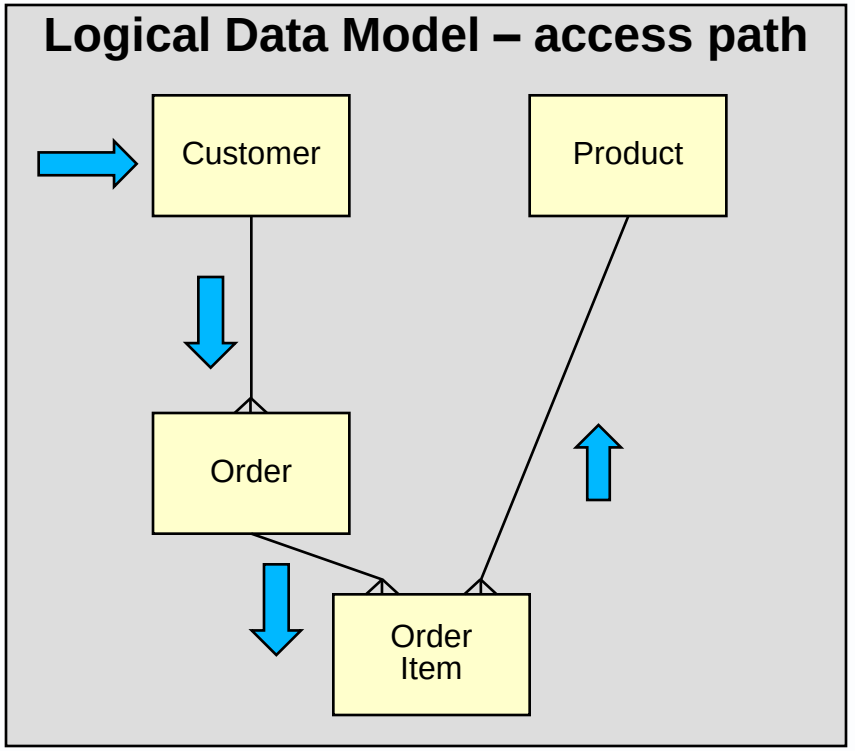
- [A data structure] composed of logically-related data entities.
- It may define data in one data store, or in several coordinated data stores.
- It may be normalised, but does not have to be.



Data access path diagram

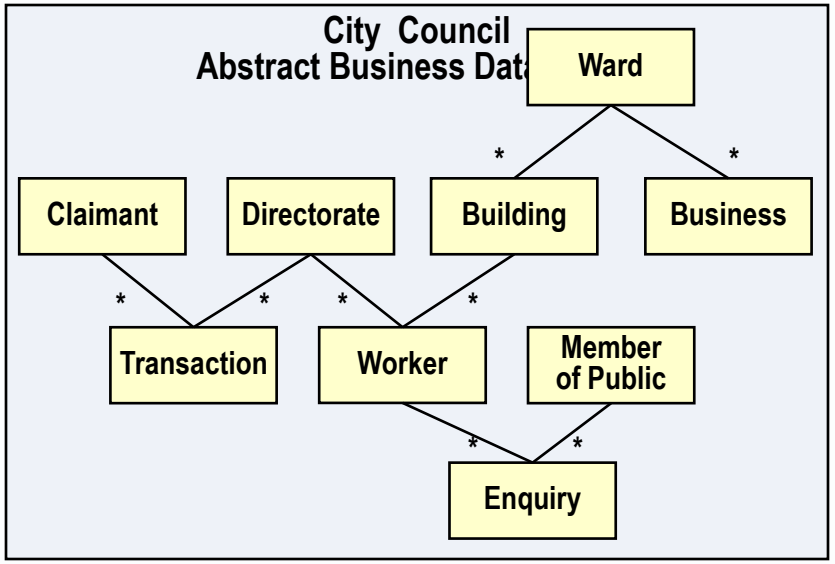
- [An artifact] that shows the route that a process takes through a data structure
- It is used to validate the data structure and study performance issues.

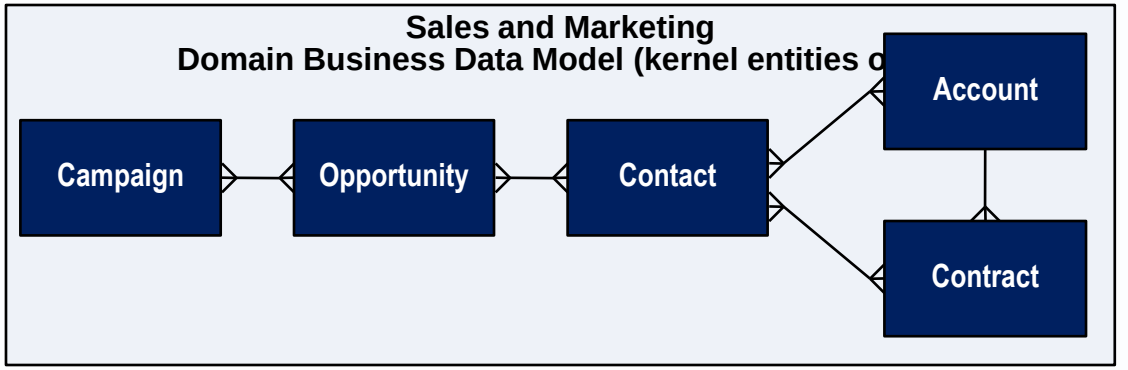
List all the products that a customer has ordered



Business data model

- [A data structure] composed of data entities recognised by business people.
- It may be a very abstract model of data stored across a whole enterprise.
- Or a more detailed model of data in several data stores in one business domain.
- Some maintain a business data entity catalogue instead of a data model.





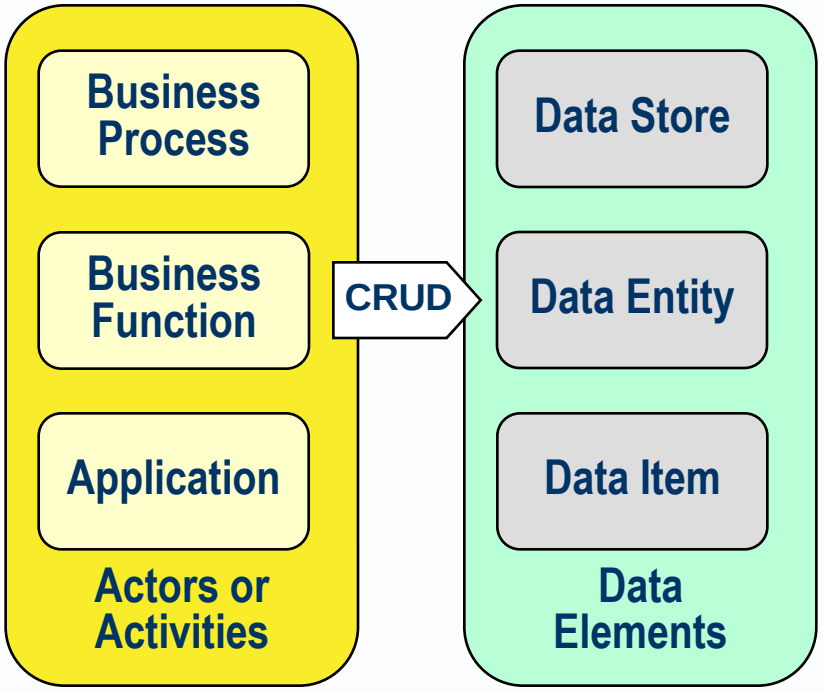
Data Dissemination matrix

- [An artifact] that tabulates data entities against data stores.
- It shows duplication of data between data stores.
- It is useful in analysis of change impacts, data mastering and security vulnerabilities.
- It may be used to define the master and copies of a data entity.

Common Entities / Data stores	*Customer*	*Product*	*Asset*	*Employee*
CRM system.	**Master**			Copy
Call-center system.	Copy			
Contact-management system	Copy			
ERP system.		**Master**		
Order-processing system		Copy		
GL tracking			Copy	
Asset database			**Master**	
Timesheet				Copy
Expense Claim				Copy
Contract DB				Copy
Company Directory				**Master**

Data entity / business function matrix

- [An artifact] that maps data entities to the business functions that create and use them.
- Cluster analysis can be used to cluster data that is created by the same functions, and functions that create the same data.



Function / Data Entity		
	Create	Read
	Update	Create

Application / Data Entity		
	Create	Read
	Update	Create

5.3 Data in motion (in flow)

Data flow

- [A motion] the passage of data structure in a message, file, form, report, display from a sender to a receiver.

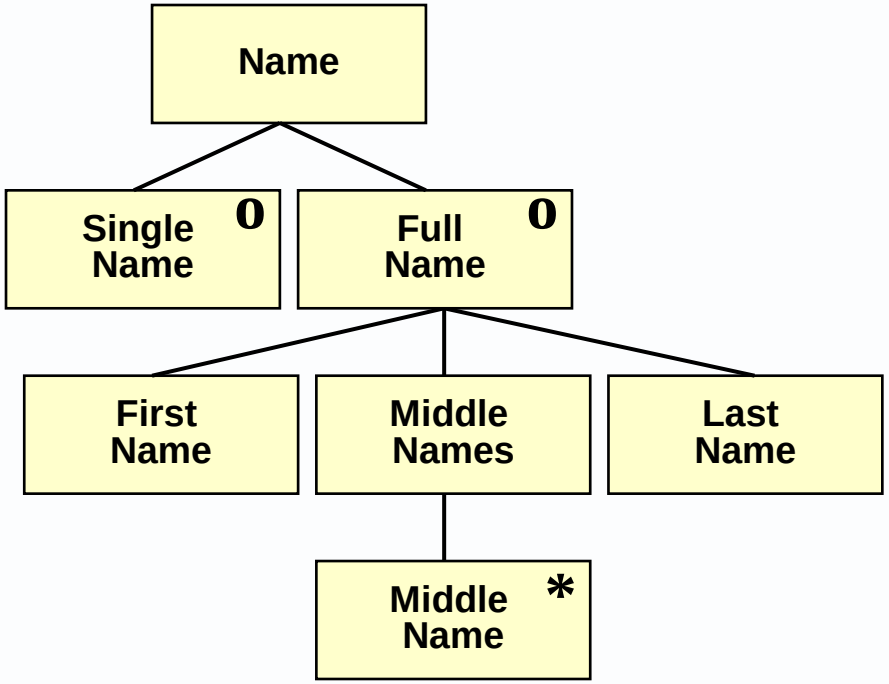
- Message
- Form
- Report
- Keyboard data entry
- User interface display
- Serial file sent via ETL

Data flow or message catalogue

- [An artefact] that lists the data flows transported from senders to receivers, with attributes such as trigger, sender, receiver, transport technology and non-functional measures

Data flow structure

- [A data object] conveyed in a data flow
- **Logical data flow structure (or regular expression)**
- [A data flow structure] that is a hierarchy in which every element is part of a sequence, or an option of a selection or an occurrence of an iteration.



Physical data flow structure

Avancier

- [A data flow structure] represented in the format or schema required by a particular data format or technology

- **Data format**
- [A standard] for the definition and organisation of a data flow structure.
- E.g.
- Comma Separated Values (CSV),
- JSON,
- Extensible Mark Up Language (XML).

Data format standard

- [A standard] for the content of a data structure.
- E.g.
- EDIFACT
- Domain-specific XML Schema Definition (XSD).

Order, Invoice, Payment ,etc.

Flat text file
Fixed position fields
Fixed length fields?

Canonical data model

- [A standard] that provides the “one true definition” of data types and structures used in data flow structures.
- It defines what data can appear in messages between applications, and in the signatures of automated services.
- It may be defined at a physical level using a data format standard such as XML.

5.4 Data qualities and integration

Data quality

- [A property] of a data item, data structure or data store.
- Meta data such as Confidentiality, Integrity and Availability (CIA).

Data quality	Aims are to ensure that
Confidentiality	Enterprise data is protected Private data remains private, accessible only to authorized readers.
Integrity	Business decisions (especially if safety-critical) are right, because a data item value is: •Consistent •Conformant to rules. •Correct •Controlled
Availability	The data (or systems) are available when needed.

Scoring data qualities (Tom Peltier)

Avancier

- Score each data item/group/store H/M/L thus

Confidentiality	Integrity	Availability
Impact of unauthorized use or disclosure	Impact of data inaccuracy, incompleteness or unauthorized modification	Impact of unavailable information
Severely impairs business operations, make a segment of the company unable to function or cause high monetary loss.	Causes failures of operations, revenue loss, wrong decisions to be made, loss in productivity or loss of customer confidence or market share.	Impairs business operations, affects customer service or makes it impossible to process revenues.
Does not severely affect operations or does not result in high monetary loss.	Makes it impossible to make some decisions, but the problem is not difficult to detect and correct, and does not severely impact business operations.	Causes productivity loss, but does not interrupt customer service or revenue generation.
Does not affect operations or result in significant monetary loss.	Does not disable business operations, since alternative validations of the information make it possible to continue	Does not severely impact business operations.

Threats to data qualities

Avancier

Data quality	Security and data architects consider
Confidentiality	Deliberate theft. Identity theft is a common goal of criminal attacks against systems. Accidental revelation through loose identity management (including loose roles and authorities).
Integrity	Unauthorised creates, updates, deletes of data in data stores Tampering with data being transported in data flows. Duplication of data storage Duplication of data entry Low quality data entry
Availability	Attacks that disable access to systems. Denial-of-service attacks (can cost as much) Inadequate design for reliability and disaster recovery

Data architects especially interested

Data integrity

Avancier

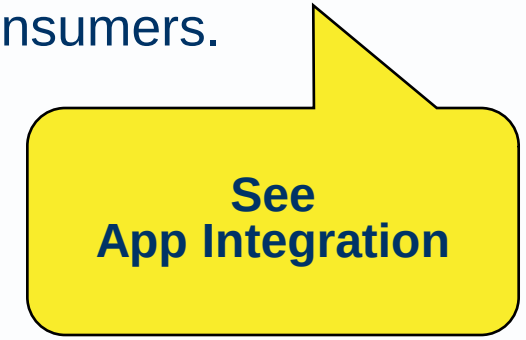
- [A property] that may embrace any or all of four qualities:
- Consistent: a data item (e.g. customer name) has the same value in every part of a distributed system, in all locations that data item is stored.
- Conformant: a data item obeys relevant business rules, sometimes in relation to another data item. E.g. an order must be for a known customer.
- Correct: a data item accurately represents a fact about an entity or event. The value of a data item is consistent with a fact in the real world.
- Controlled: a data flow has the same data content when it reaches its destination as it did when it left its source. OR data in a data store is not changed without authorisation.

Data disintegrity

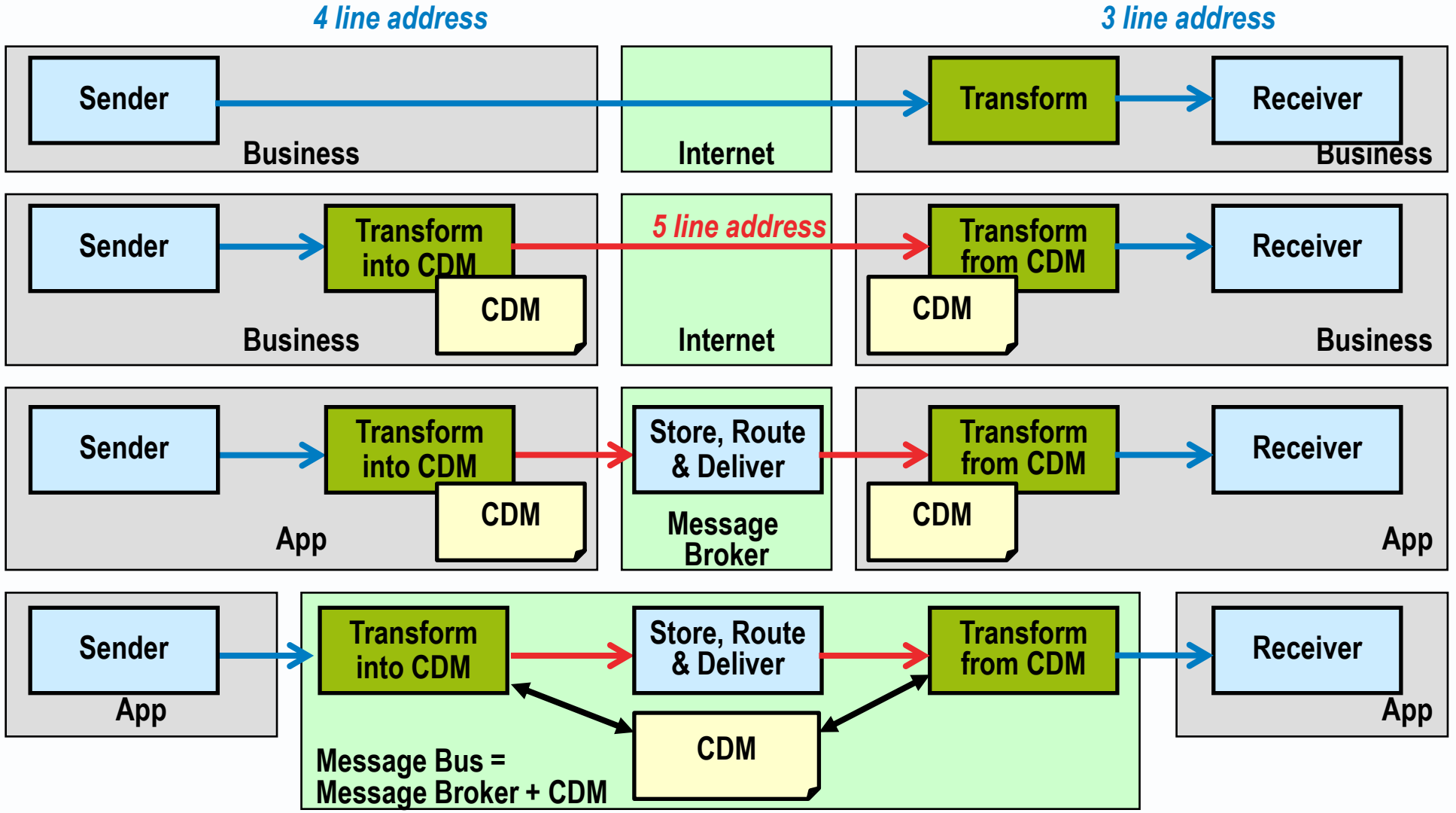
- [A property] an issue that may be redressed by one-off data quality improvement exercises, and by a variety of application integration patterns.

Master data management

- [A technique] that enables an enterprise to maintain and/or find one “master” version of a data item or data structure, such as a customer or product data record.
- It is supported by a range of application integration patterns and technologies, including some that hide the reality of disparate data sources from data consumers.



Application integration using a Canonical Data Model

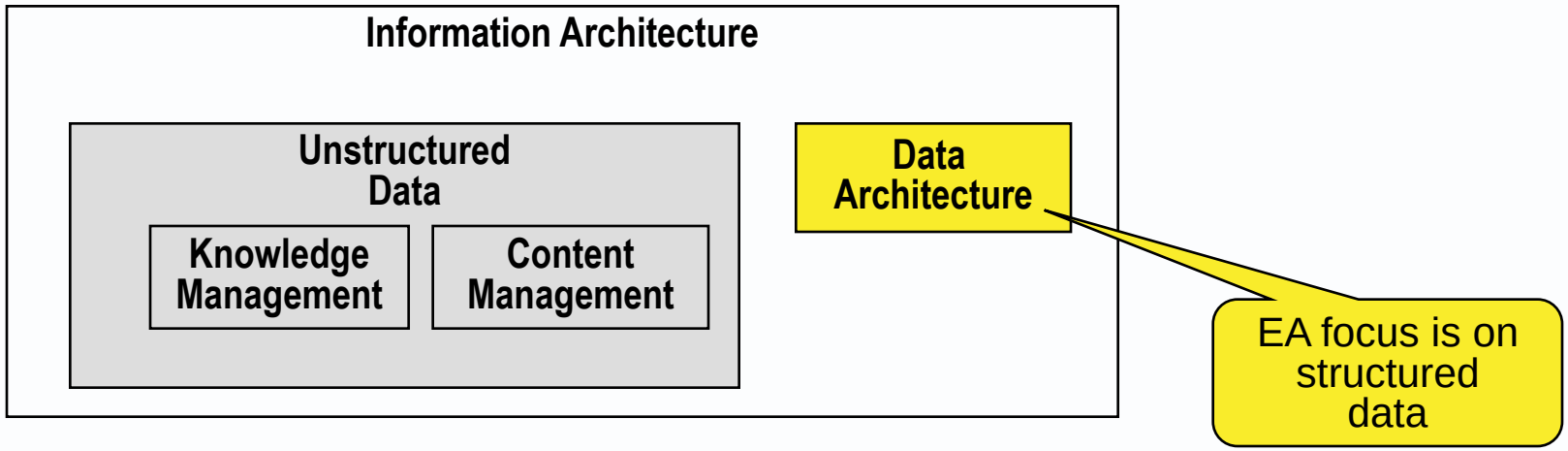


Unstructured data

Avancier

Knowledge and/or content management	The organisation, systems and processes for producing, storing, editing, sharing and searching unstructured data. Roles can include creator, editor, publisher, administrator (managing access permissions etc.) and consumer, viewer or guest. Knowledge and/or content management is regarded in this reference model as a matter for systems analysts and applications architects to address, rather than data architects.

Unstructured data is out of scope for us here



- Twitter say they have four fundamental data types and query patterns:
  - tweets,
  - timelines,
  - social graphs
  - search indices.
- For each, Twitter implemented custom data stores because existing solutions were insufficient.

Brands can be identified by various elements

- When the levers of control are strongly centralized, content management systems are capable of delivering an exceptionally clear and unified brand message.

- All following may be trademarked as “brands”

  - **name:** word(s) used to identify a company, product, service, or concept
  - **logo:** a visual trademark
  - **tagline or catchphrase**: e.g. “Never-knowingly undersold"
  - **graphics:** e.g. the "dynamic ribbon" is a trademarked part of Coca-Cola's brand
  - **shapes:** e.g. the Coca-Cola bottle and Volkswagen Beetle shapes
  - **colors:** e.g. Owens-Corning is the only fiberglass insulation that can be pink.
  - **sounds:** e.g. a unique tune or chord: e.g. Windows
  - **scents:** e.g. the rose-jasmine-musk scent of Chanel No. 5 is trademarked
  - **tastes:** e.g. a trademarked recipe of herbs and spices for fried chicken
  - **movements:** e.g. the upward motion of Lamborghini car doors