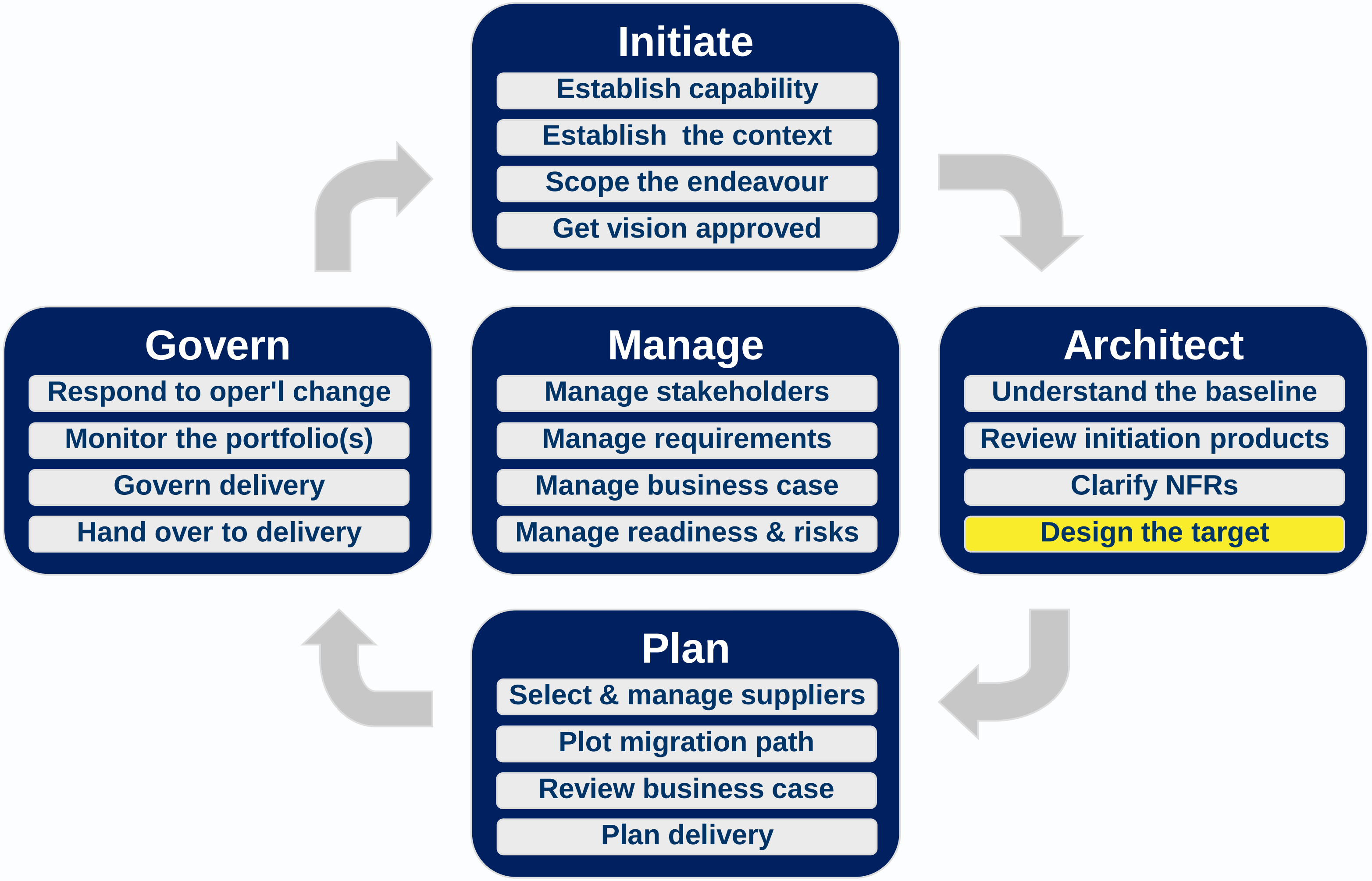
Avancier

Avancier Methods (AM)

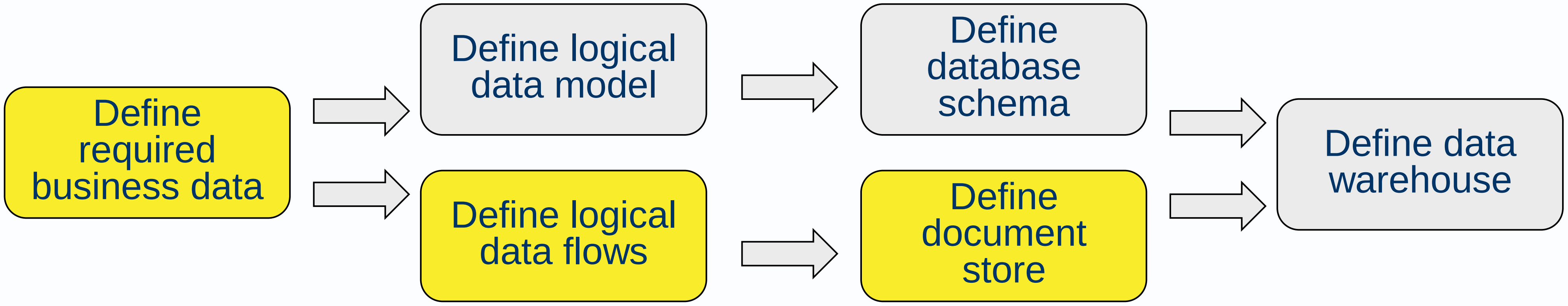
Data Architecture

Define data flows (logical and physical)

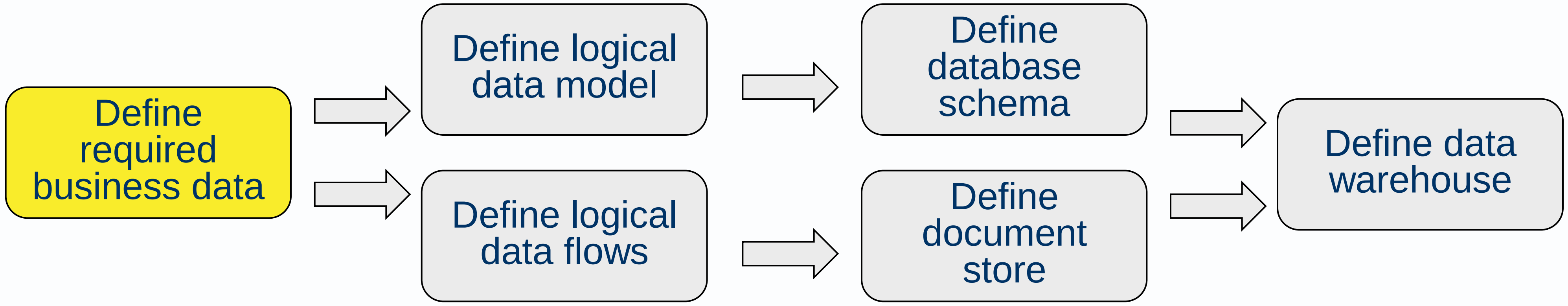
AM level 2 process



Define business data stores and flows


Define required business data
Define logical data model
Define logical data flows
Define database schema
Define document store
Define data warehouse

AM level 3 and 4 process: Define required business data

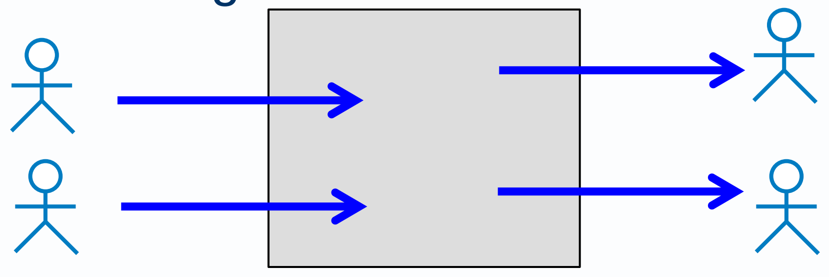

Define required business data
Define logical data model
Define logical data flows
Define database schema
Define document store
Define data warehouse

1. Identify where data is created and used
2. Define data created and used in business activities
3. Define data dictionary

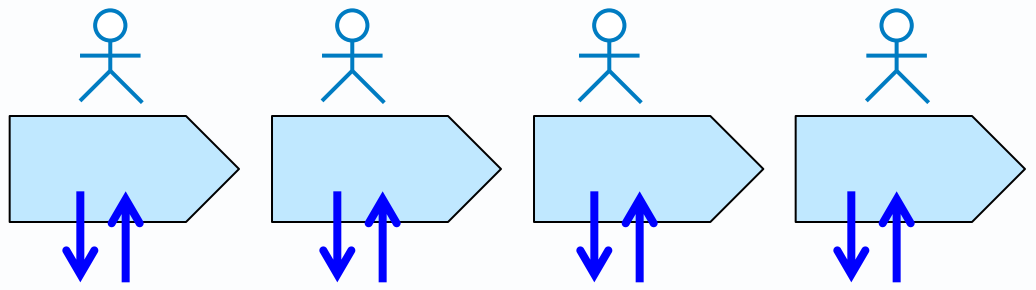
Skip to slide 15

Identify where data is created and used

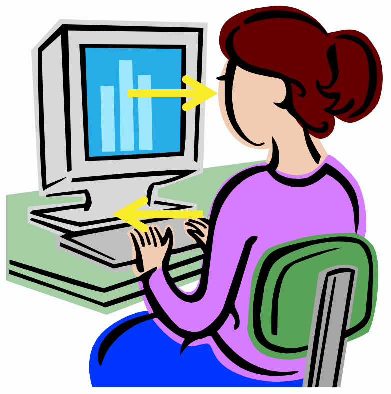
- Context diagram



- Value stream / scenario diagrams (showing OPOPOT activities)



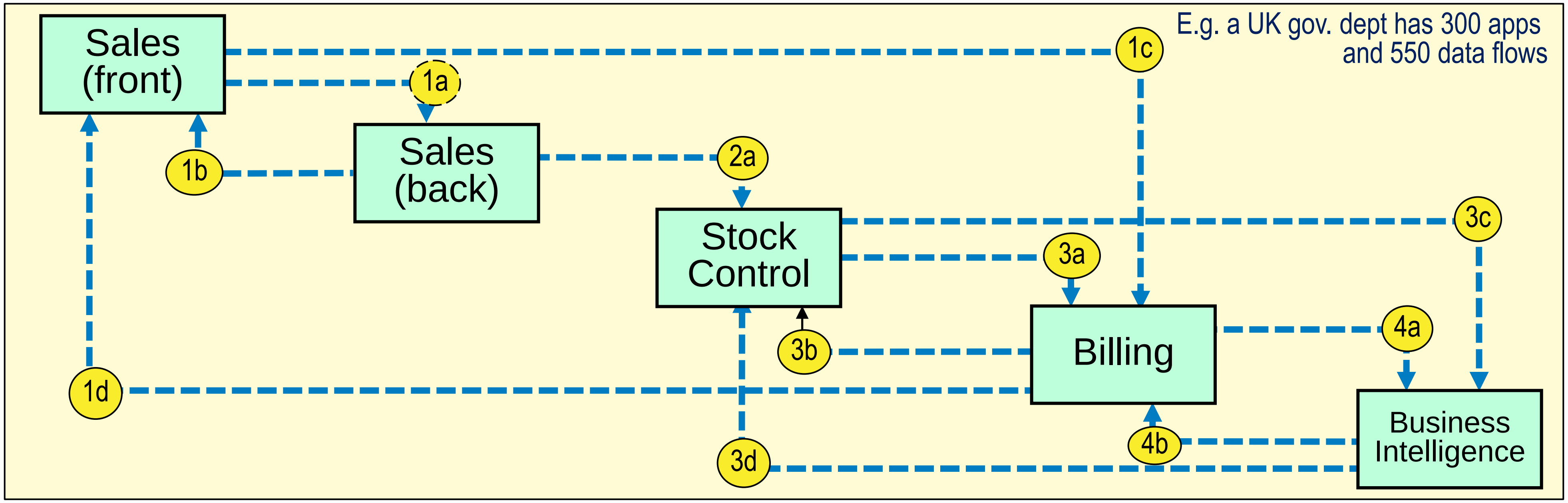
- Client devices and user interfaces



Avancier

Data flow documentation (solution level)

Data Flow (aka Application Communication) Diagram



Data Flow (aka Interface) Catalogue

Id	Flow Name	Source	Destination	Content
1a	Order entry	Sale (front)	Sale (back)	Ref. 999
1b	Order accepted	Sale (back)	Sale (front)	Ref. 999
2a	Notification	Sale (back)	Stock Control	Ref. 999

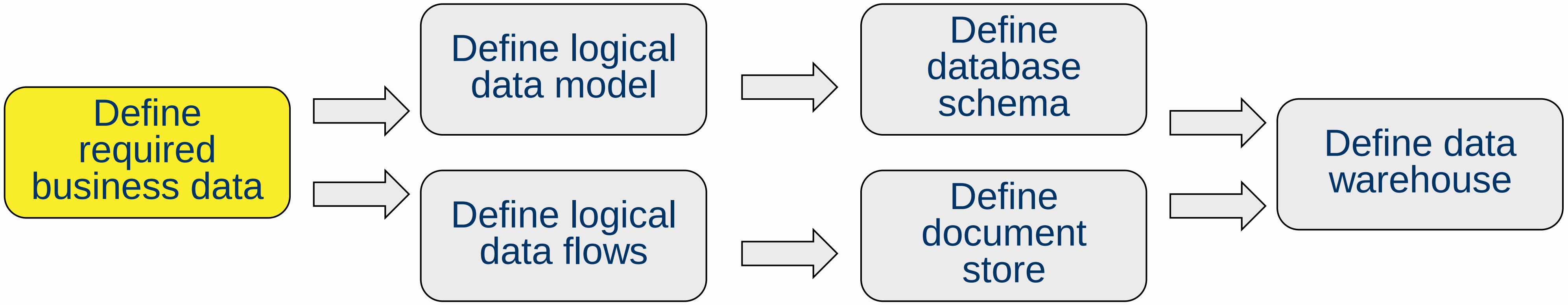
Data flow (aka Interface) catalogue

- Catalog the key data flows

FLOW	Functional			Non-functional					Media	
Name	Source	Destination	Content	Frequency	Vol.	Confident.	Integrity	Avaialbility	Technology	Protocol
Order entry	CRM	Sales	Ref. 999	1K per day		High	Medium	24*7	Web	http
Order accepted	Sales	CRM	Ref. 999	1K per day		Medium	Medium	0900-1800	Web	http
Notification	Sales	Stock	Ref. 999	100 per day		High	Medium	24*7	Web	http

- Like many such illustrations, this shows what *could* be documented
- Understanding what is possible in theory is a precursor to deciding what to do in practice.

AM level 3 and 4 process: Define required business data


Define required business data
Define logical data model
Define logical data flows
Define database schema
Define document store
Define data warehouse

1. Identify where data is created and used
2. Define data created and used in business activities
3. Define data dictionary

Define data used

Salesman wants

- Customer Order History
  - **Customer id**
  - **Customer name and address**
  - Orders Placed
    - **Order id**
    - **Order value**
    - Products Ordered
      - **Product type**
      - **Product amount**
    - Products Ordered End
  - Order Placed End
- Customer Order History END

Product manager wants

- Product Demand Report
  - **Product type**
  - **Amount on hand**
  - Products ordered
    - **Product amount**
    - **Order id**
  - Products Ordered End
- Product Demand Report End

Define data created and maintained

- Customer creates

Shopping Basket
    Customer id
    Order id
    Order value
    Products Ordered
        Product type
        Product amount
    Products Ordered End
Shopping Basket

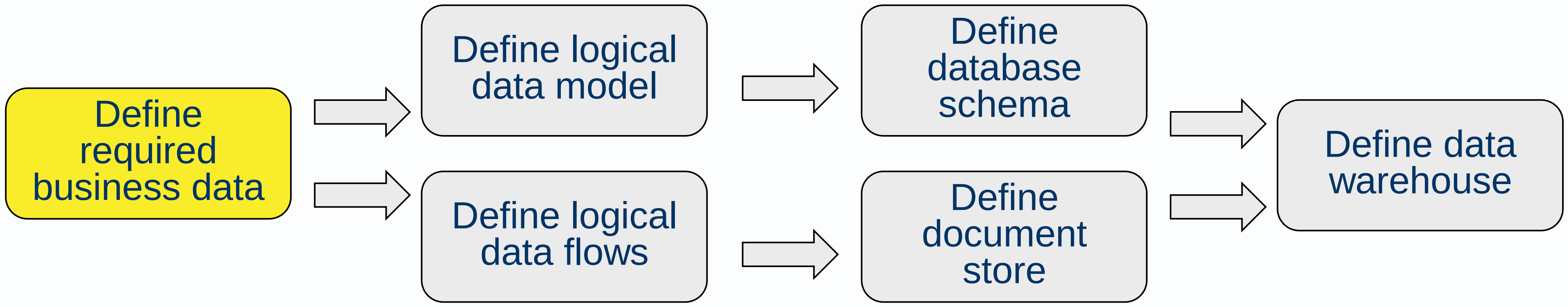
- The HR department maintains a spreadsheet of all employees

Human resources
Employee Number, Name, Role, Grade

- The sales manager has a card file with all salesmen in it

Salesman card file
Employee Number, Name, Commission Rate, Sales Area

AM level 3 and 4 process: Define required business data


Define required business data
Define logical data model
Define logical data flows
Define database schema
Define document store
Define data warehouse

1. Identify where data is created and used
2. Define data created and used in business activities
3. Define data dictionary

Avancier

Data dictionary (solution level)

1. Define entities and items in I/O data flows
2. Define data that must be remembered for future activity
3. Define business rules associated with data items

Name	Facts	Constraints	Derivation rule
Currency Code	abbreviates **Currency**	is a three letter **String** in the range defined at ref. 999…	
Currency	denominates a **Value**		
Item Value	is an attribute of an **Order Item** is associated with] **Currency**	is a **Number** in the range 0 to 999	= **Product Amount Ordered** * **Unit Price**
Order Value	is an attribute of **Order** is calculated from **Item Values**	is a **Number** in the rang 0 to 9999	= sum of **Item Values** for an **Order** - **Discount**

Assign items to primitive data types (e.g. as in Java)

Avancier

Primitive data types

- Boolean
- Character
- Integer
  - Byte
  - Short
  - Integer
  - Long integer
- Floating point
- Double floating point

User defined data types

- Name (Character)
- City (Character)
- Order value (Integer)

Type	Contains	Default	Size	Range
boolean	true or false	false	1 bit	NA
char	Unicode character	\u0000	16 bits	\u0000 to \uFFFF
byte	Signed integer	0	8 bits	-128 to 127
short	Signed integer	0	16 bits	-32768 to 32767
int	Signed integer	0	32 bits	-2147483648 to 2147483647
long	Signed integer	0	64 bits	-9223372036854775808 to 9223372036854775807
float	IEEE 754 floating point	0.0	32 bits	±1.4E-45 to ±3.4028235E+38
double	IEEE 754 floating point	0.0	64 bits	±4.9E-324 to ±1.7976931348623157E+308

Complex data types

A source of many system integration problems

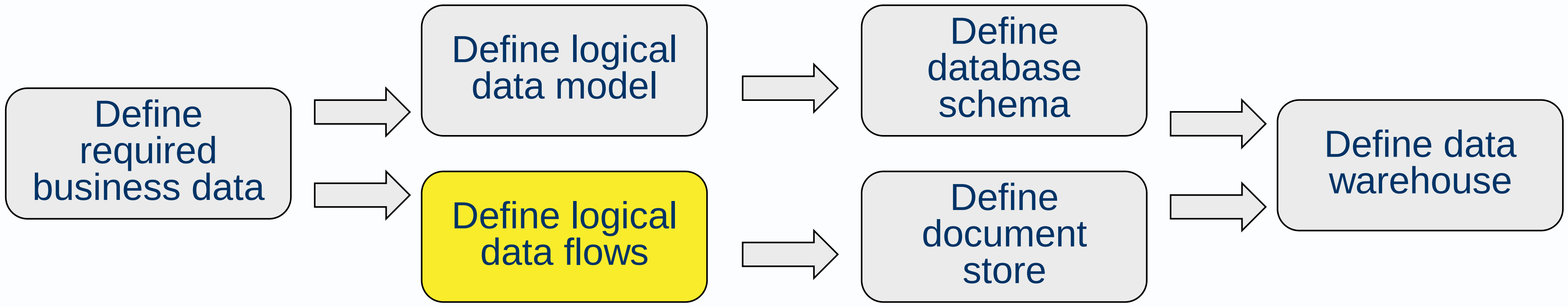
Avancier

Complex data types (simple data structures)

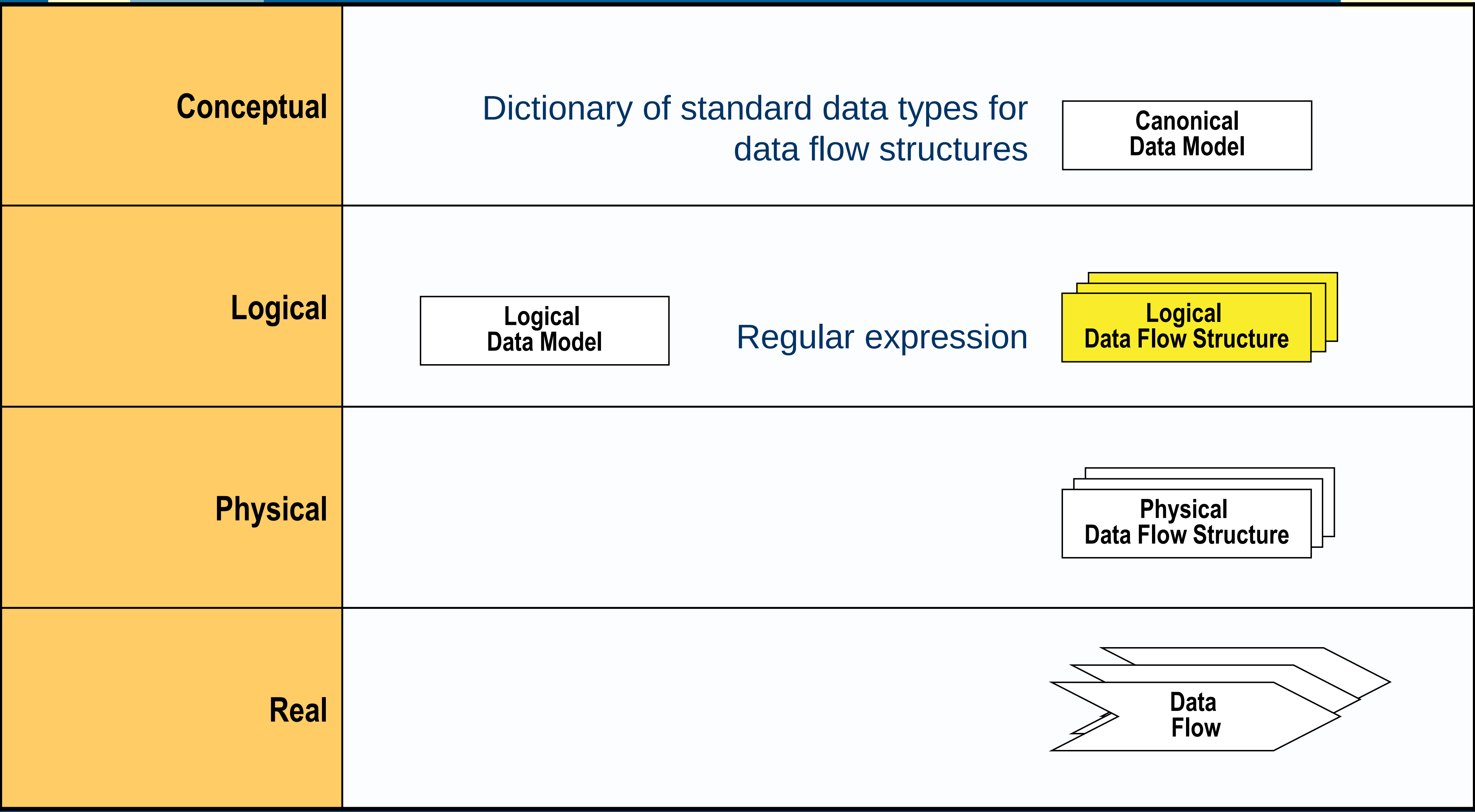
- Date
  - DD
  - MM
  - YYYY
- Person
  - Title
  - First name
  - Last name
- Address
  - Address Line 1
  - Address Line 2
  - Address Line 3
  - City
  - County/State
  - Postcode

AM level 3 and 4 process: Define required business data

Avancier


Define required business data
Define logical data model
Define logical data flows
Define database schema
Define document store
Define data warehouse

Defining data flows at a logical level


Conceptual
Dictionary of standard data types for data flow structures
Canonical Data Model
Logical
Logical Data Model
Regular expression
Logical Data Flow Structure
Physical
Physical Data Flow Structure
Real
Data Flow

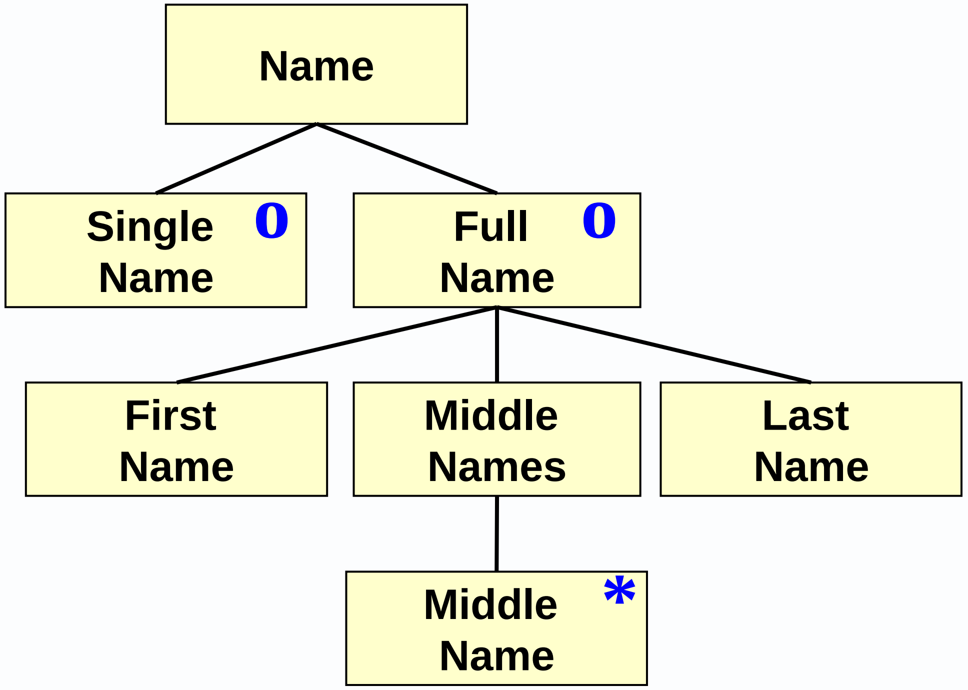
Data flow or message catalogue

- **Data flow:** the passage of data structure in a message, file, form, report, display from a sender to a receiver.
- (Message, Form, Report, Keyboard input, User interface display, Serial file sent via ETL)

Id	Flow Name	Source	Destination	Content
1a	Order entry	Sale (front)	Sale (back)	Ref. 999
1b	Order accepted	Sale (back)	Sale (front)	Ref. 999
2a	Notification	Sale (back)	Stock Control	Ref. 999

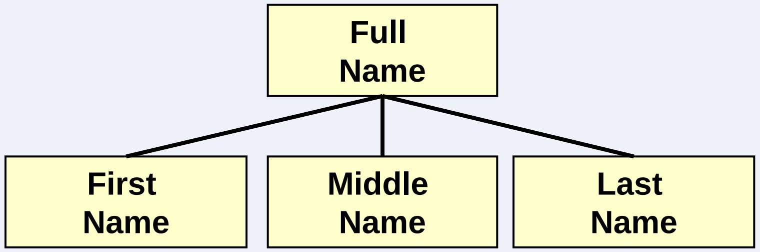
Logical data flow structure (or regular expression)

- [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.



Document as a simple sequence of fields

Jackson structure

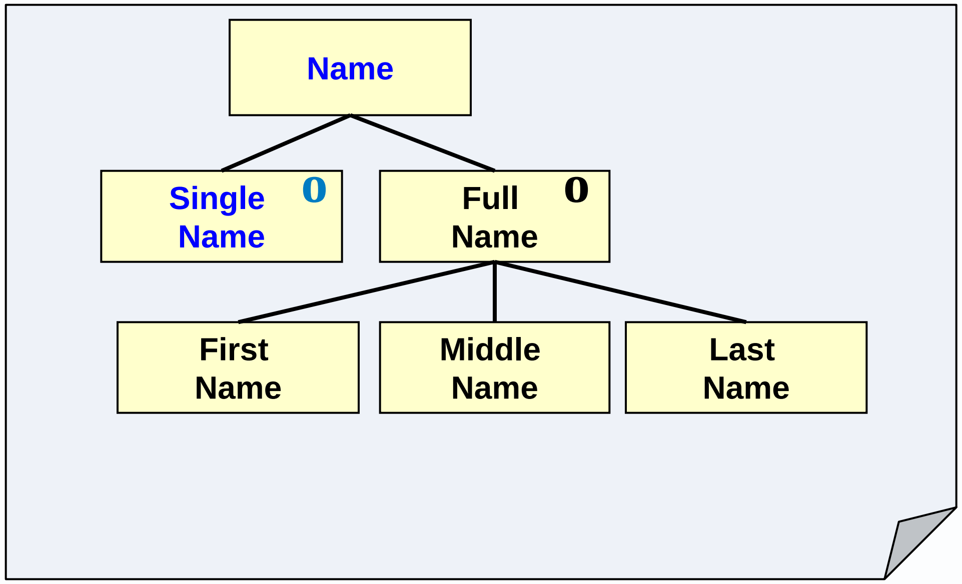


Schematic logic

```
Full name SEQUENCE
        First Name
        Middle Name
        Last Name
Full name ENDS
```

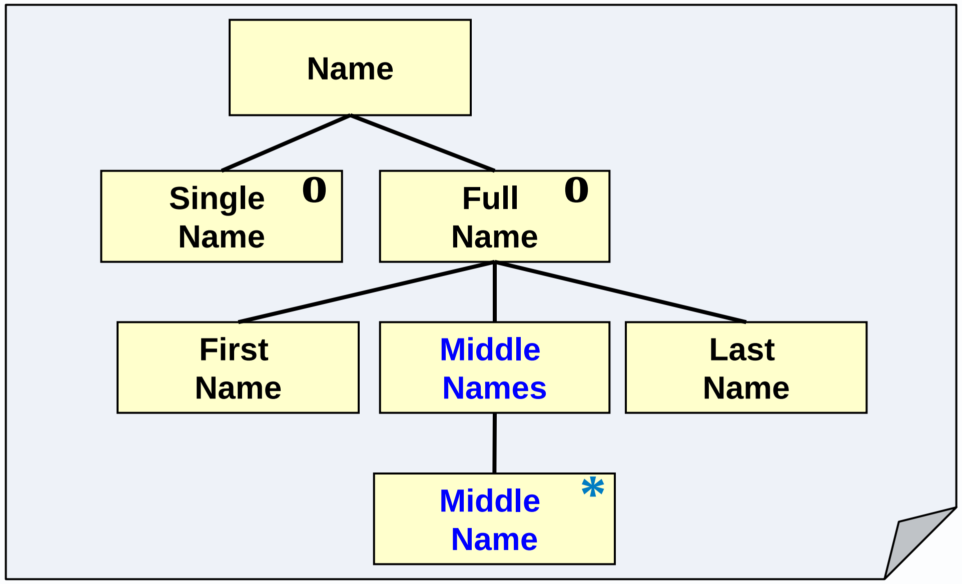
Document with optional elements

Avancier



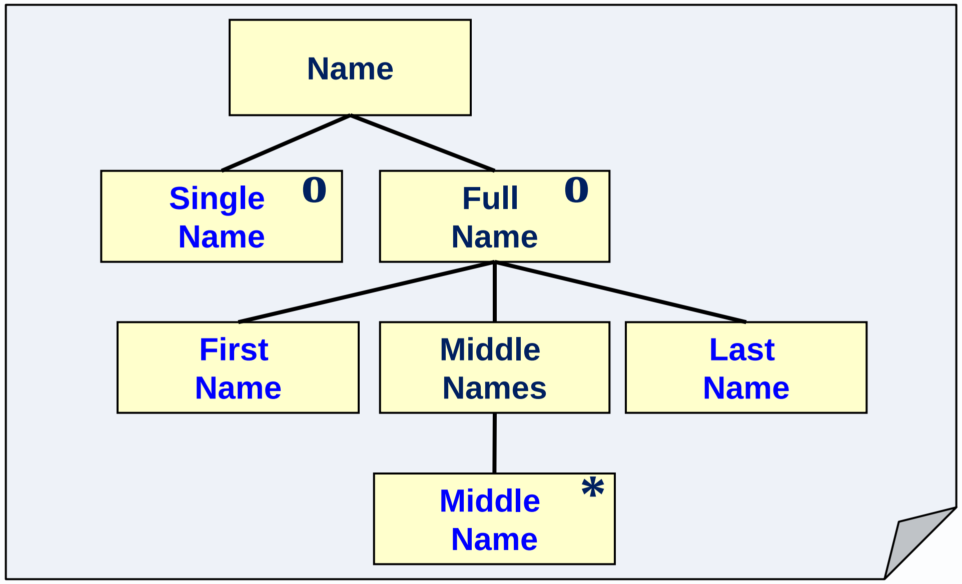
```
Name SELECT
        Single Name
Name OR
        Full name SEQUENCE
                First Name
                Middle Name
                Last Name
        Full name END
Name END
```

Document with iterated element



```
Name SELECT
        Single Name
Name OR
        Full name SEQUENCE
                First Name
                Middle Names ITERATION
                        Middle Name
                Middle Names END
                Last Name
        Full name END
Name END
```

Document as XML schema



```
Name SELECT
        Single Name
Name OR
        Full name SEQUENCE
                First Name
                Middle Names ITERATION
                        Middle Name
                Middle Names END
                Last Name
        Full name END
Name END
```

```
xsd:choice
        xsd:element name="SingleName" type="Text" minOccurs="1" maxOccurs="1" /
        xsd:sequence
          xsd:element name="FirstName" type="Text" minOccurs="1" maxOccurs="1" /
          xsd:element name="MiddleName" type="Text" minOccurs="0" maxOccurs="unbounded"/
          xsd:element name="LastName" type="Text" minOccurs="1" maxOccurs="1" /
        /xsd:sequence
/xsd:choice
```

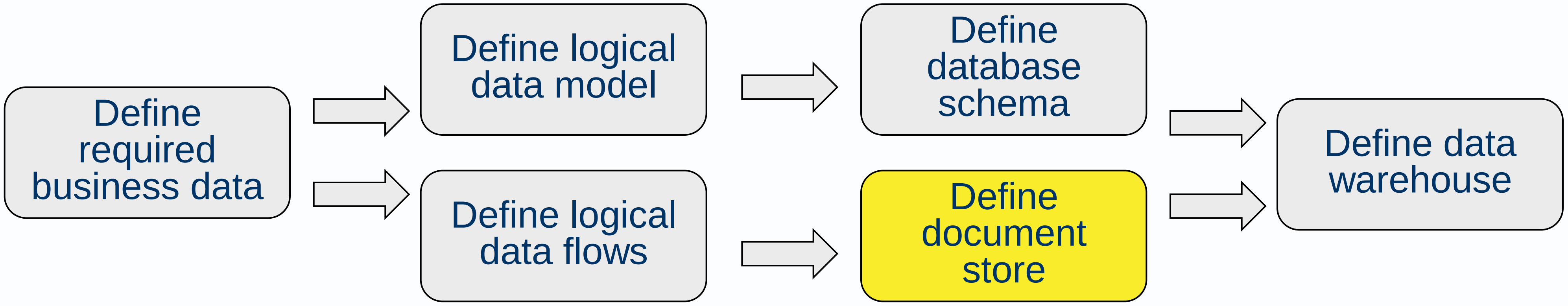
Avancier

Avancier Methods (AM)

Data Architecture

Define data flows (physical level)

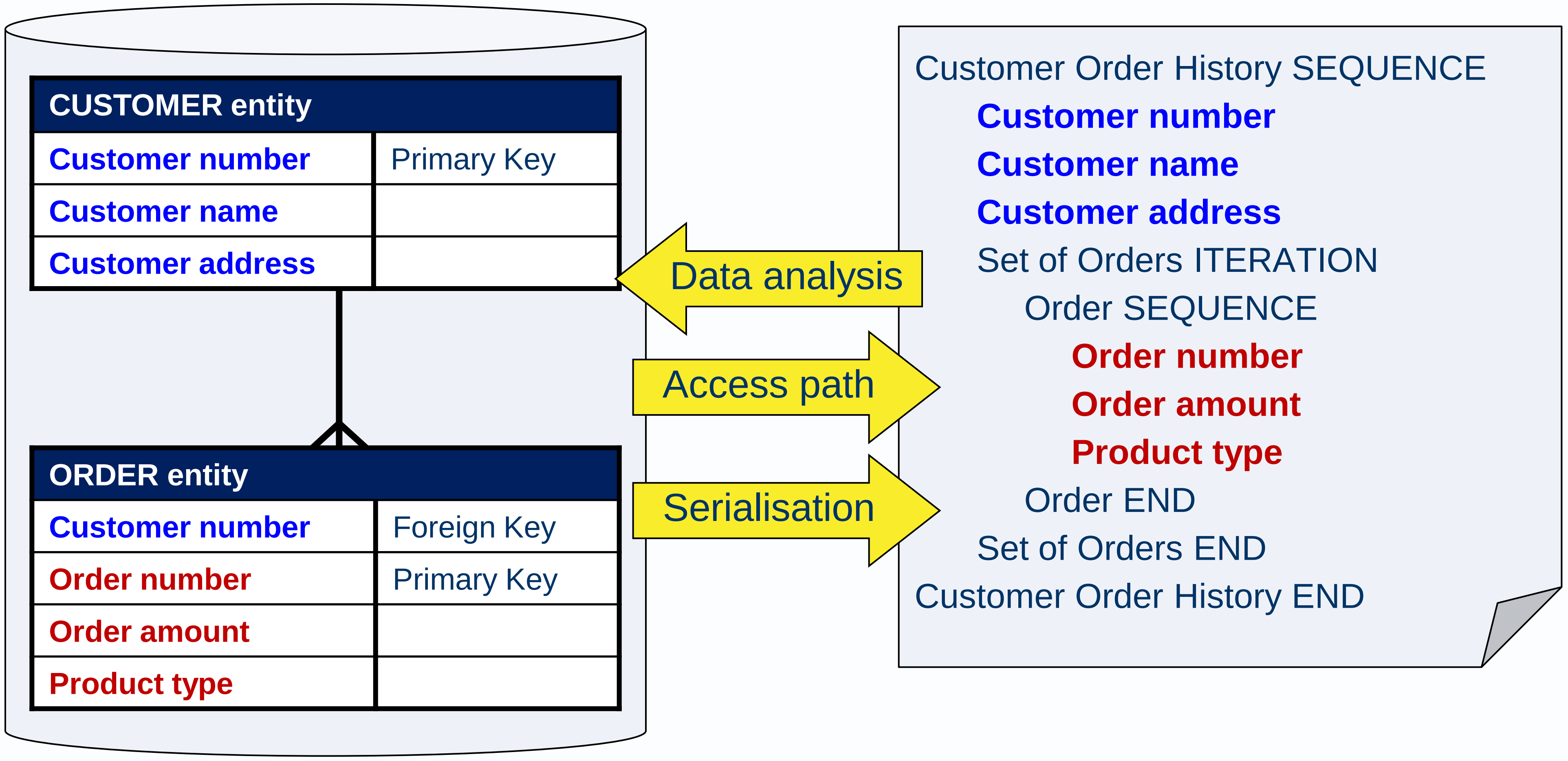
AM level 3 and 4 process: Define required business data


Define required business data
Define logical data model
Define logical data flows
Define database schema
Define document store
Define data warehouse

1. Refine the document store to facilitate reports?
2. Define integrity constraints?
3. Define physical data format

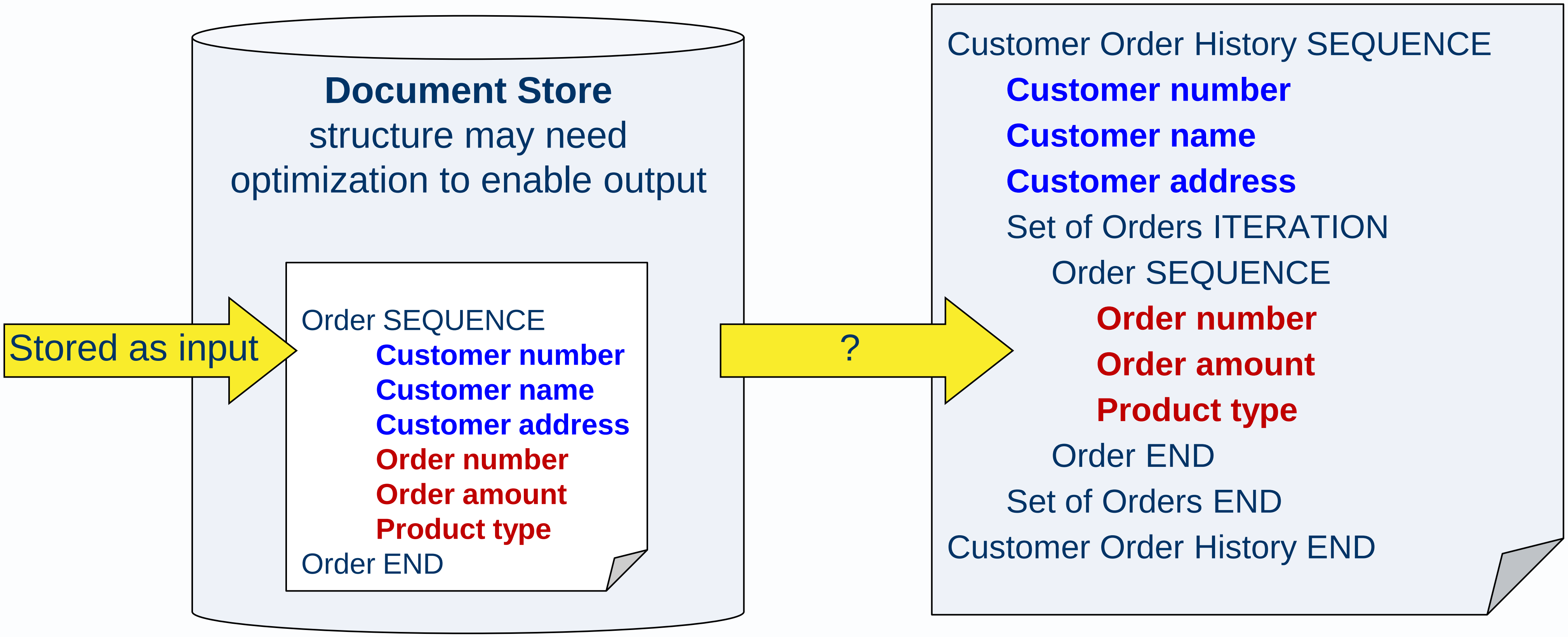
Design time v run time

- Relational data store
- Output data flow



Refine the document store to facilitate reports?

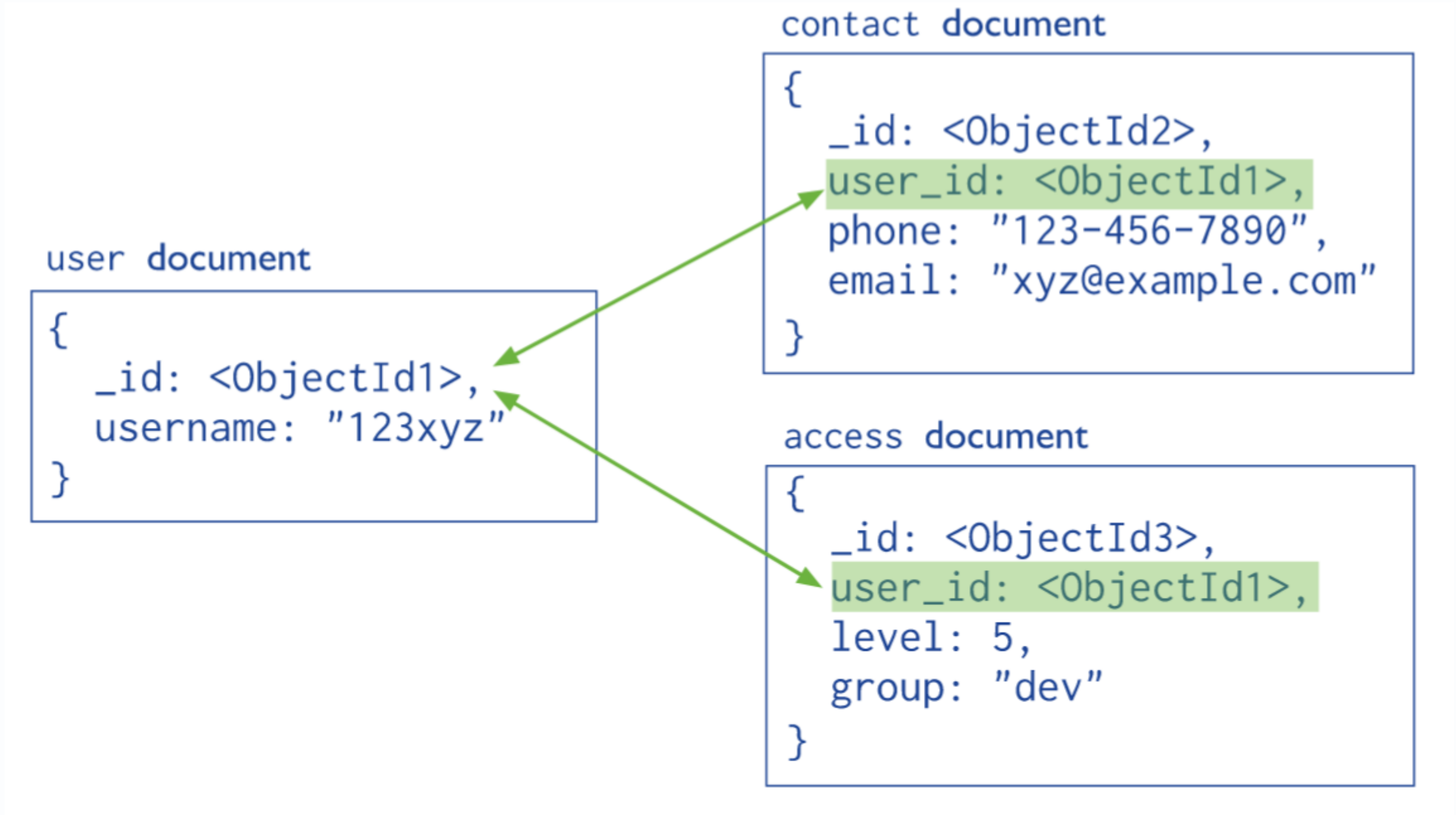
- Document store
- Output data flow



Define integrity constraints?

Avancier

- You can maintain integrity between documents in MongoDB
- But I think you have to do it manually?



Defining data flows at a physical level

Conceptual	Dictionary of standard data types for data flow structures	**Canonical Data Model**
Logical	Regular expression	**Logical Data Flow Structure**
Physical	CSV, JSON, XML	**Physical Data Flow Structure**
Real	The physical medium of wires, microwaves, sound waves	**Data Flow**

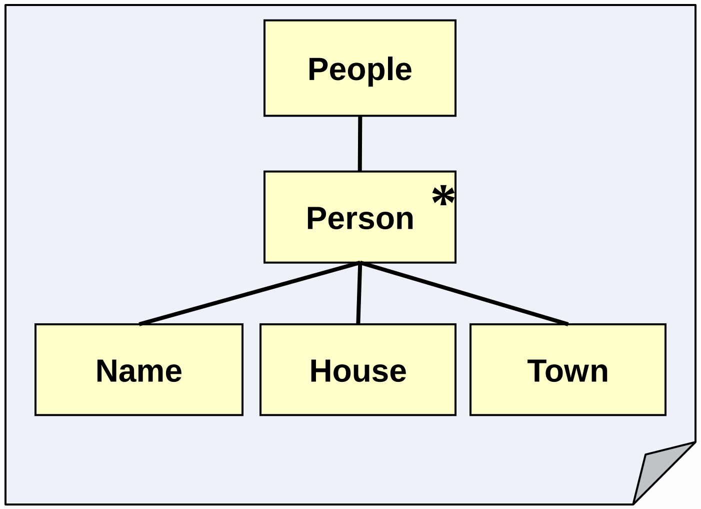
Data format

- [A standard] for the definition and organisation of a data flow structure.

- E.g.
  - Comma Separated Values (CSV),
  - JSON,
  - Extensible Mark Up Language (XML).

Comma Separated Values (CSV)

- Simple data flow structures
  - John, 3 South Street, Big Town
  - Mary, 44 North Street, Small Town
- Used (e.g.) in spreadsheet import/export



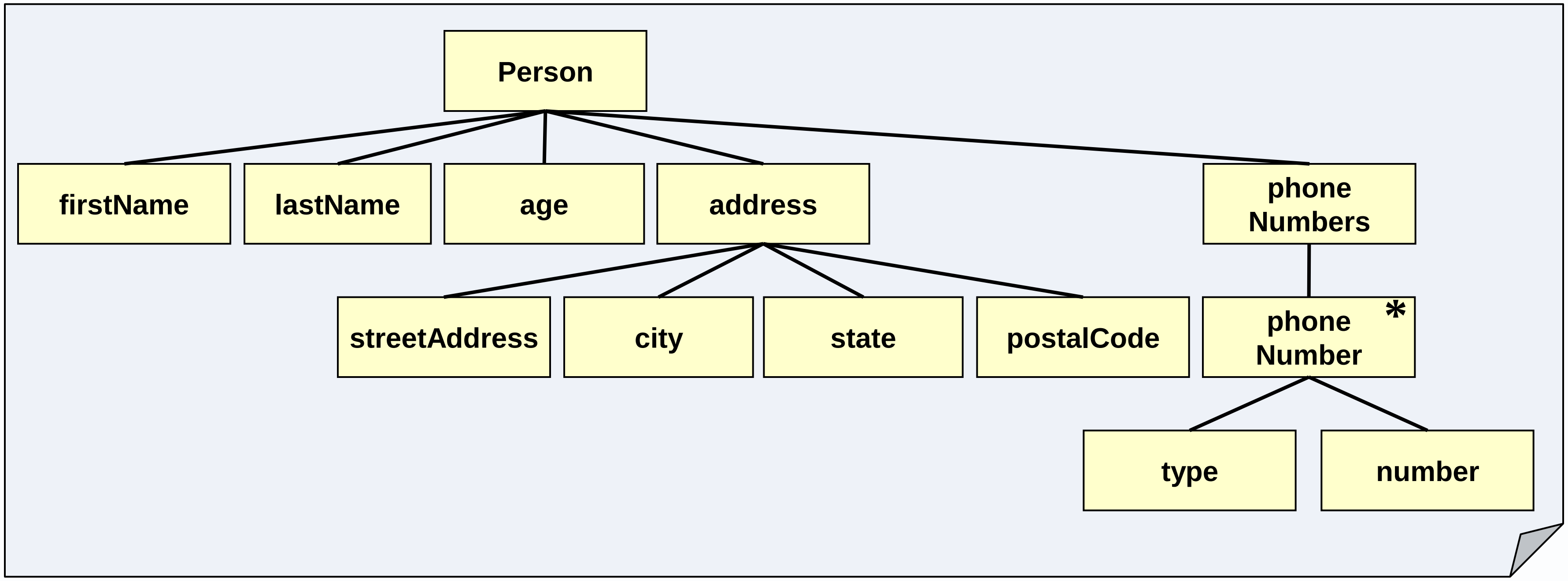
- The raw data can appear meaningless
  - A, 0001, 23
  - B, 9888, 10
- So, sender and receiver must know the meaning of each data item

JSON – self-describing data flow format

- A JSON data flow contains not only
  - Data (data values) but also
  - Data descriptors (data types)
- So it can be sent to recipients who do not already know the data types, but find them in the flow itself.
- JSON data types
  - **Number** (double precision floating-point format in JavaScript, generally depends on implementation)
  - **String** (double-quoted Unicode, with backslash escaping)
  - **Boolean** (true or false)
  - **Array** (an ordered sequence of values, comma-separated and enclosed in square brackets; the values do not need to be of the same type)
  - **Object** (an unordered collection of key:value pairs with the ':' character separating the key and the value, comma-separated and enclosed in curly braces; the keys must be strings and should be distinct from each other)
  - **Null** (empty)
  - **“Structural characters" (**i.e. brackets "{ } [ ]", colons ":" and commas ",").

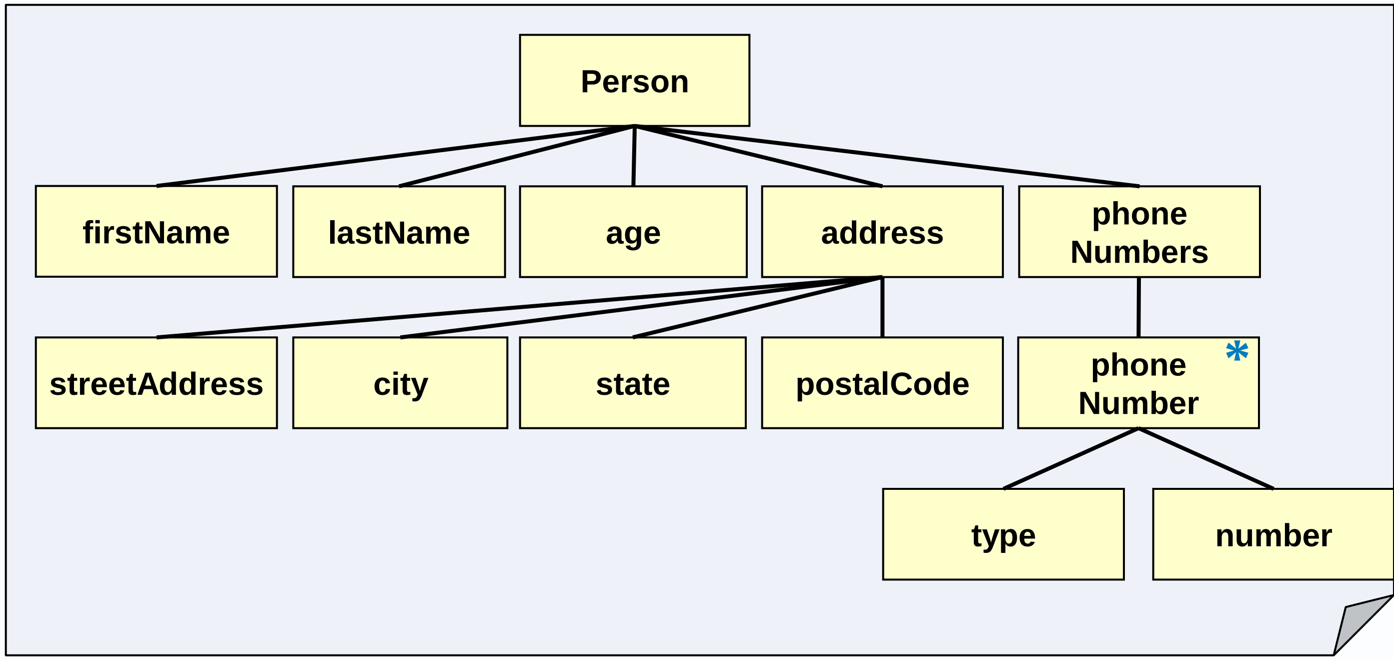
Data flow structure diagram (regular expression)

- A document in a data flow structure can be defined as
  - A hierarchical structure – a regular expression
  - A “Jackson structure” like this, or in the form of an XML schema



JSON data structure

```
{
    "firstName":
    "lastName":
    "age":
    "address":
        "streetAddress":
        "city":
        "state":
        "postalCode":
     "phoneNumbers": [
        {
            "type":
            "number":
        },
        {
            "type":
            "number":
        }
     ]
 }
```



http://en.wikipedia.org/wiki/JSON

JSON representation of an object that describes a person.

```
{
    "firstName": "John",
    "lastName": "Smith",
    "age": 25,
    "address": {
        "streetAddress": "21 2nd Street",
        "city": "New York",
        "state": "NY",
        "postalCode": 10021 },
     "phoneNumbers": [
        {
            "type": "home",
            "number": "212 555-1234"
        },
        {
            "type": "fax",
            "number": "646 555-4567"
        }
    ]
}
```

The object has
string fields for first and last name,
a **number** field for age,
an **object** composes of address fields

an **array** of phone number objects.

XML (eXtensible Mark up Language)

Avancier

- Another delf-describing data flow format

- Like JSON, an XML data flow contains
  - not only the data (values) but also
  - descriptors of the data (types)

- How does XML differ from JSON?
  - Con: Clunkier
  - Pro: Can be supported by a separate XML Schema Definition (XSD)

JSON Schema.org?

XSD (XML Schema Definition language)

Avancier

- used to
  - define the data structure of an XML document
  - enables verification of a document’s data integrity
  - loosens coupling between sender and receiver
- E.g. a simple type - a subtype of string – with a range of values

```
?xml version="1.0"?
xsd:schema xmlns:xsd=http://www.w3.org/1999/XMLSchema

    xsd:simpleType name="PersonTitle" base="xsd:string"
        xsd:enumeration value="Mr." /
        xsd:enumeration value="Ms." /
        xsd:enumeration value="Dr." /
        xsd:enumeration value="Rev." /
    /xsd:simpleType

    xsd:complexType name="Text" content="textOnly" base="xsd:string" derivedBy="restriction" /
```

There are many other standard data flow formats,

Digital image data

- TIFF version 6 uncompressed (.tif)
- JPEG (.jpeg, .jpg)
- PDF (.pdf)

Digital video data:

- MPEG-4 High Profile (.mp4)
- JPEG 2000 (.mj2)

Digital audio data

- Free Lossless Audio Codec (FLAC) (.flac)
- Waveform Audio Format (WAV) (.wav)
- MPEG-1 Audio Layer 3 (.mp3)

Documentation and scripts

- Open Document Text (.odt)
- Rich Text Format (.rtf)
- HTML (.htm, .html)
- plain text (.txt)
- widely-used proprietary formats
  - e.g. MS Word (.doc/.docx) or MS Excel (.xls/ .xlsx)
- XML marked-up text (.xml) to a DTD or schema, e.g. XHMTL 1.0
- PDF (.pdf)

Geospatial data; vector and raster data

- ESRI Shapefile (essential -- .shp,.shx, .dbf;
- optional -- .prj, .sbx, .sbn)
- geo-referenced TIFF (.tif, .tfw)
- CAD data (.dwg)
- tabular GIS attribute data

Qualitative data, textual

- **eXtensible Mark-up Language (XML) text according to a Document Type Definition (DTD) or schema (.xml)**
- Rich Text Format (.rtf)
- plain text data, ASCII (.txt)
- Hypertext Mark-up Language (HTML) (.html)
- widely-used proprietary formats, e.g. MS Word (.doc/.docx)

Quantitative tabular data with extensive metadata

- SPSS portable format (.por)
- delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information
- structured text or mark-up file containing metadata information, e.g. DDI XML file
- MS Access (.mdb/.accdb)

Quantitative tabular data with minimal metadata:

- **comma-separated values (CSV) file (.csv)**
- tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate
- delimited text of given character set -- only characters not present in the data should be used as delimiters (.txt)
- widely-used formats, e.g. MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf) and OpenDocument Spreadsheet (.ods)

Data format standard

- [A standard] for the content of a data structure.

- E.g. EDIFACT (>> GS1? UNCFACT?)
  - Electronic Data Interchange For Administration Commerce and Transport
    - Order, Invoice, Payment etc.

- Any domain-specific XML Schema Definition (XSD).

  FPML (financial products)
  FIXML (financial instruments)
  OASIS (Names? Addresses?)
  Open Travel Alliance (OTA) Cars, hotels, insurance, airports, currencies, countries
  GTFS (General Transit Feed Specification)
  TransXchange SIRI (Schedules)
  TRANSMODEL (EU public transport info.)
  Air travel – PNR passenger name record
  ARTS (association of retail, textile…)
  Open Geospatial Consortium (OGC)
  JISC – universities - HEDIIP

Major data format standards defined using XML

Avancier

- A short tour of the non-technical industry efforts to create a common XML-based vocabulary for specified purposes and industries. PETE O'DELL "Silver Bullets"
- **1. Astronomy.** See http://fits.gsfc.nasa.gov.
- **2. Built environment, and infrastructure systems integration.** See www.obix.org.
- **3. Distribution/Commerce.** See www.rosettanet.org.
- **4. Education.** See schools interoperability framework.
- **5. Financial reporting. See www.xbrl.org.**
- **6. Financial research.** See www.rixml.org.
- **7. Food.** See www.mpxml.org.
- **8. Healthcare. See www.hl7.org.**
- **9. Information technology architecture.** (opengroup.org)
- **10. Instruments.** See www.nasa.gov
- **11. Insurance. See www. acord.org.**
- **12. Legal.** See www.legalxml.org.
- **13. Manufacturing.** See www.pslx.org. (no longer available)
- **14. News.** See www.iptc.org.
- **15. Oil and gas.** See www.pidx.org.
- **16. Publishing.** See www.oasis.org.
- **17. Real Estate.** See www.RETS.org.
- **18. Research.** See www.casrai.org.
- **19. Telecommunications.** See www.atis.org. + TMF

- FPML (financial products)
- FIXML (financial instruments)
- OASIS
  - Names? Addresses?
- Open Travel Alliance (OTA)
  - Cars, hotels, insurance, airports, currencies, countries
- GTFS (General Transit Feed Specification)
- TransXchange SIRI (Schedules)
- TRANSMODEL (EU public transport info.)
- Air travel – PNR passenger name record
- ARTS (association of retail, textile…)
- Open Geospatial Consortium (OGC)
- JISC – universities - HEDIIP

Run-time transmission of data

Conceptual	Dictionary of standard data types for data flow structures	**Canonical Data Model**
Logical	Regular expression	**Logical Data Flow Structure**
Physical	CSV, JSON, XML	**Physical Data Flow Structure**
Real	The physical medium of wires, microwaves, sound waves	**Data Flow**