Data modeling (15) - Normalization

- Normalization – Overview
  - Why?
  - What is normalization?
  - Rules for normalizing data – Normalization Forms
    - First Normal Form (1NF)
    - Second Normal Form (2NF)
    - Third Normal Form (3NF)
    - Others:
    - Boyce-Codd Normal Form (BCNF)
    - Fourth Normal Form (4NF) – Isolate Independent Multiple Relationships
    - Fifth Normal Form (5NF) -. Optimal Normal Form
    - Domain-Key Normal Form (DKNF)

Data modeling (16) - Normalization

- Data normalization is a process in which data attributes within a data model are organized to increase the cohesion of entity types.
- In other words, the goal of data normalization is to reduce and even eliminate data redundancy, an important consideration for application developers because it is incredibly difficult to stores objects in a relational database that maintains the same information in several places.
- In this class we will deal only with the First 3 Forms of Normalization. Higher levels of data normalization are beyond the scope of this lesson.

Data modeling (16a) - Normalization

- Normalization – Why??? (in accordance to E.F. Codd)
  - To free the collection of relations from undesirable insertion, update and deletion dependencies.
  - To reduce the need for restructuring the collection of relations, as new types of data are introduced, and thus increase the life span of application programs.
  - To make the relational model more informative to users.
  - To make the collection of relations neutral to the query statistics, where these statistics are liable to change as time goes by.
- Normalization addresses potential anomalities of data management

Data modeling (17) - Normalization

- Insertion anomaly
  - No insert possible in a not-normalized data model

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201
424	Dr. Newsome	29-Mar-2007	?

  - no flexibility to add data

Data modeling (18) - Normalization

- Update anomaly
  - The same information can be expressed in multiple (repeating) rows in a non-normalized data model. Therefore, updates to data needs to be applied to all instances where this data is stored -> risk of data quality degrading

Employees' Skills

Employee ID	Employee Address	Skill
426	87 Sycamore Grove	Typing
426	87 Sycamore Grove	Shorthand
519	94 Chestnut Street	Public Speaking
519	96 Walnut Avenue	Carpentry

Data modeling (19) - Normalization

- Deletion anomaly
  - Requirement to delete data beyond the original intent

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201

DELETE

  - data needs to be deleted beyoned its intended scope (i.e. Course to be delted requires the whole record to be deleted)

Data modeling (20) - Normalization

- Minimize redesign efforts when extending the database structure
  - Maximize resilience of data model for changes (data structure change)
  - A fully normalized database allows its structure to be extended without changing existing structure too much
  - Important for application that are built on top of the data model (increases the autonomy between data tier and application tier)

Data modeling (21) - Normalization

Normalization – 3 Main Forms

Level	Rule
First normal form (1NF)	an entity type is in 1NF if **each attribute contains only a single value** (i.e. atomic values)
Second normal form (2NF)	an entity type is in 2NF when it is in 1NF and when **all** of its **non-key attributes** are **dependent** to the whole of **candidate keys**
Third normal form (3NF)	an entity type is in 3NF when it is in 2NF and when **all** of its **attributes** are **solely dependent** on the **primary key**

Data modeling (22) – Normalization – 1. NF

- an entity type is in 1NF if each attribute contains only a single value (i.e. atomic values)

Customer

Customer ID	First Name	Surname	Telephone Number
123	Pooja	Singh	555-861-2025, 192-122-1111
456	San	Zhang	(555) 403-1659 Ext. 53; 182-929-2929
789	John	Doe	555-808-9633

Not normalized

Customer

Customer ID	First Name	Surname	Telephone Number1	Telephone Number2
123	Pooja	Singh	555-861-2025	192-122-1111
456	San	Zhang	(555) 403-1659 Ext. 53	182-929-2929
789	John	Doe	555-808-9633	

1. NF

Data modeling (23) – Normalization – 2. NF

- an entity type is in 2NF when it is in 1NF and when all of its non-key attributes are dependent to the whole of candidate keys

Electric toothbrush models

Manufacturer	Model	Model full name	Manufacturer country
Forte	X-Prime	Forte X-Prime	Italy
Forte	Ultraclean	Forte Ultraclean	Italy
Dent-o-Fresh	EZbrush	Dent-o-Fresh EZbrush	USA
Brushmaster	SuperBrush	Brushmaster SuperBrush	USA
Kobayashi	ST-60	Kobayashi ST-60	Japan
Hoch	Toothmaster	Hoch Toothmaster	Germany
Hoch	X-Prime	Hoch X-Prime	Germany

1. NF

- Candidate key: composite of Manufacturer & Model
- Manufacturer country only partly dependent on candidate key (it is fully dependent on Manufacturer – but not on Manufacturer AND Model)

Electric toothbrush manufacturers

Manufacturer	Manufacturer country
Forte	Italy
Dent-o-Fresh	USA
Brushmaster	USA
Kobayashi	Japan
Hoch	Germany

Electric toothbrush models

Manufacturer	Model	Model full name
Forte	X-Prime	Forte X-Prime
Forte	Ultraclean	Forte Ultraclean
Dent-o-Fresh	EZbrush	Dent-o-Fresh EZbrush
Brushmaster	SuperBrush	Brushmaster SuperBrush
Kobayashi	ST-60	Kobayashi ST-60
Hoch	Toothmaster	Hoch Toothmaster
Hoch	X-Prime	Hoch X-Prime

2. NF

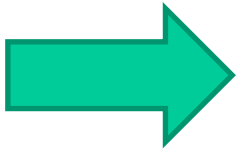
Data modeling (24) – Normalization – 3. NF

- an entity type is in 3NF when it is in 2NF and when all of its attributes are **solely** dependent on the primary key (i.e. no transitive dependency which means, they are only dependent to the identified primary key)

Tournament winners

Tournament	Year	Winner	Winner's date of birth
Indiana Invitational	1998	Al Fredrickson	21 July 1975
Cleveland Open	1999	Bob Albertson	28 September 1968
Des Moines Masters	1999	Al Fredrickson	21 July 1975
Indiana Invitational	1999	Chip Masterson	14 March 1977

2. NF



Tournament winners

Tournament	Year	Winner
Indiana Invitational	1998	Al Fredrickson
Cleveland Open	1999	Bob Albertson
Des Moines Masters	1999	Al Fredrickson
Indiana Invitational	1999	Chip Masterson

Winner's dates of birth

Winner	Date of birth
Chip Masterson	14 March 1977
Al Fredrickson	21 July 1975
Bob Albertson	28 September 1968

3. NF

- Primary key: composite of Tournament and Year
- Winner's date of birth: not dependent on primary key but dependent on Winner

Data modeling (25) - Denormalization

- Denormalization
  - To increase read performance of a database
  - Introduce redundant copies of data by grouping data in accordance to performance requirements for the subset of data
  - Only applicable for very complex queries (complex join relationships to be established) applied to huge quantities of data

Data modeling (26) – Star Schema

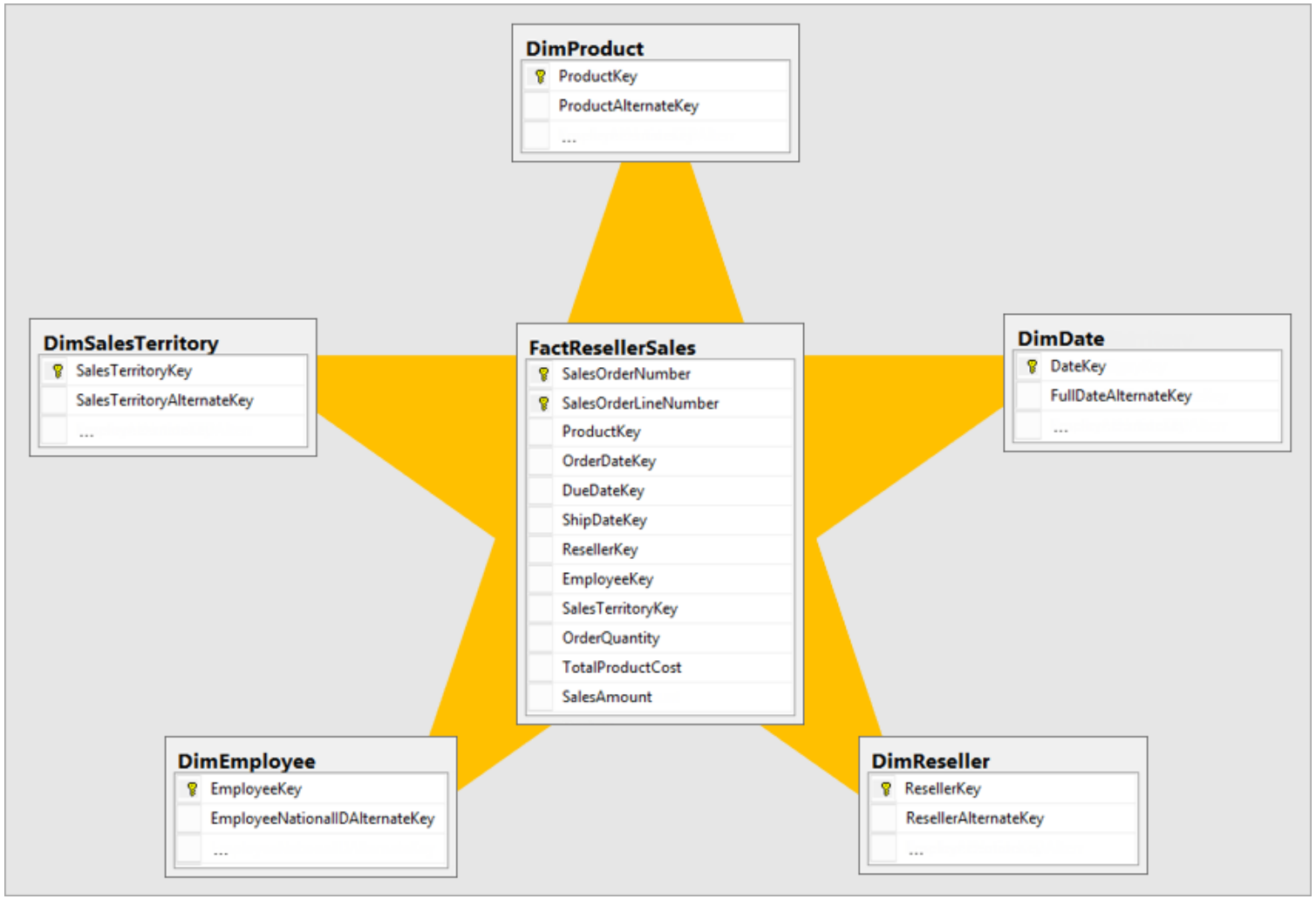
- **Star schema** is a mature modeling approach widely adopted by *relational data warehouses*. It requires modelers to classify their model tables as either *dimension* or *fact*.
- **Dimension tables** describe business entities—the *things* you model. Entities can include products, people, places, location, time and other concepts. The most consistent table you'll find in a star schema is a date dimension table. A dimension table contains:
- dimensions tables are related to the fact tables table
  - a key column (or columns) that acts as a unique identifier
  - foreign keys from the fact table to connect facts with the dimensions
  - and the descriptive columns depicting the dimensions

(See: https://learn.microsoft.com/en-us/power-bi/guidance/star-schema, last visited Dec,2023 - adopted)

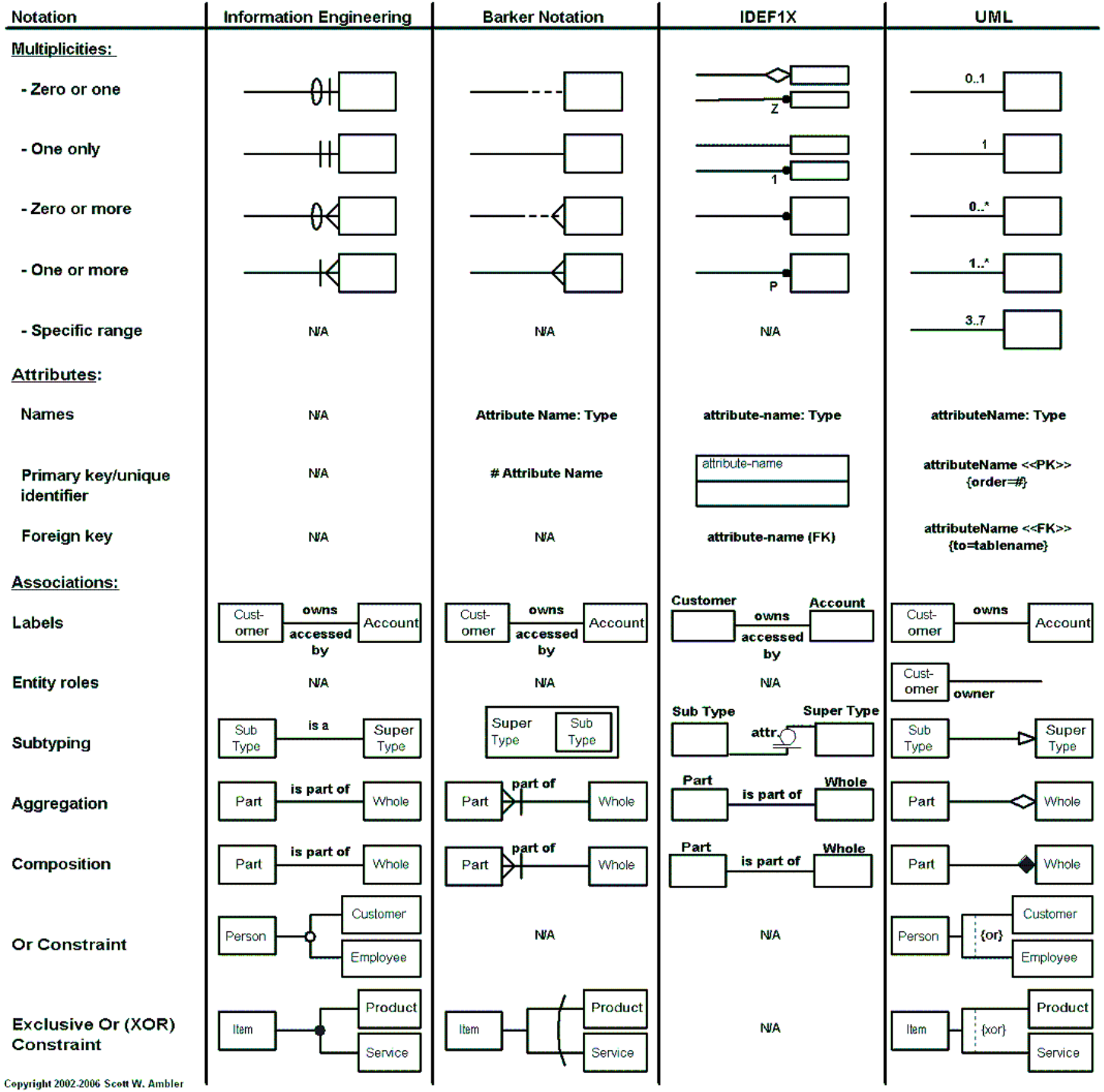
Data modeling (26a) – Star Schema

- **Fact tables** store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc. A fact table contains dimension key columns that relate to dimension tables, and numeric measure columns. The dimension key columns determine the *dimensionality* of a fact table, while the dimension key values determine the *granularity* of a fact table. For example, consider a fact table designed to store sale targets that has two dimension key columns **Date** and **ProductKey**. It's easy to understand that the table has two dimensions. The granularity, however, can't be determined without considering the dimension key values. In this example, consider that the values stored in the **Date** column are the first day of each month. In this case, the granularity is at month-product level.
- Generally, dimension tables contain a relatively small number of rows. Fact tables, on the other hand, can contain a very large number of rows and continue to grow over time.

Data modeling (26c) – Star Schema

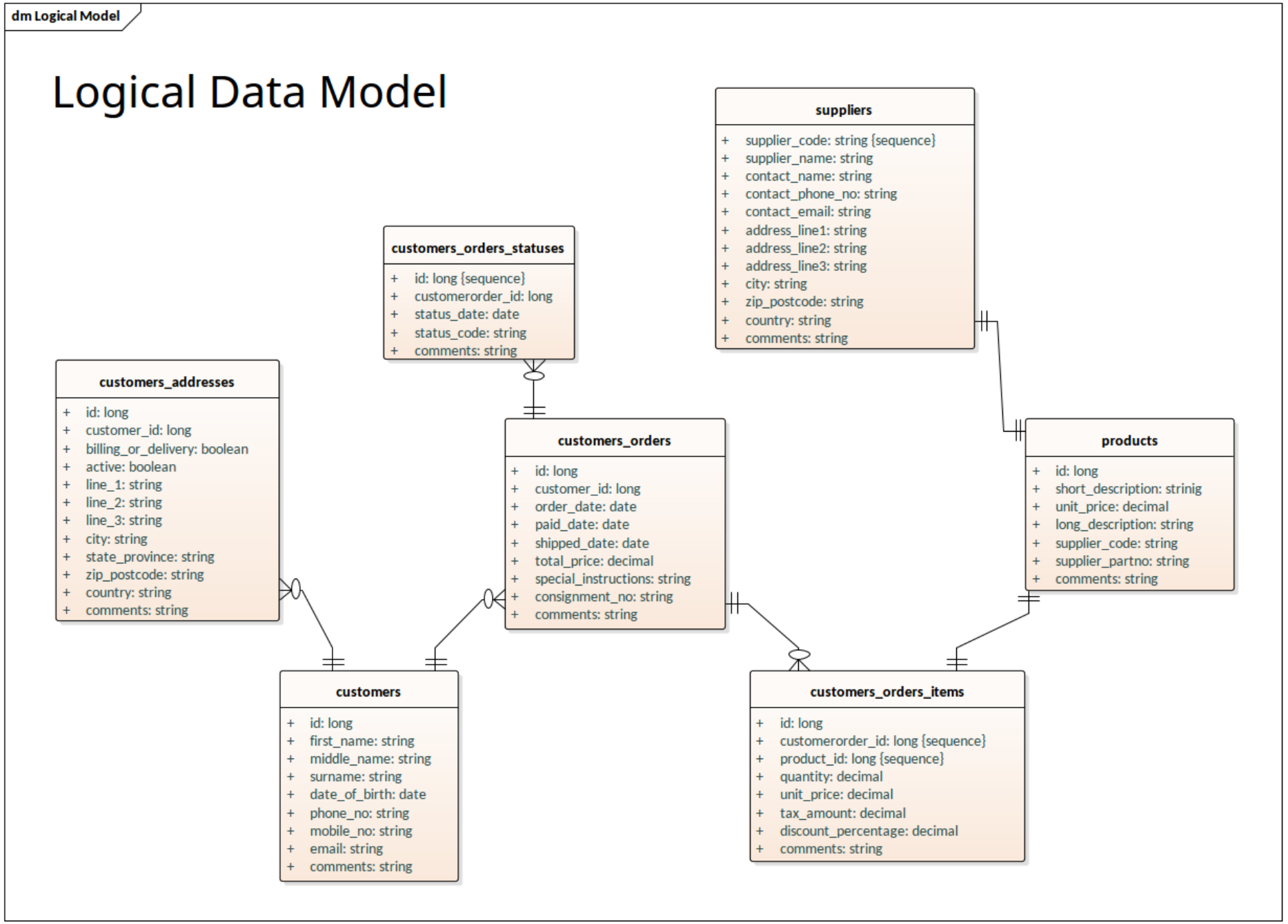


Data modeling - notations



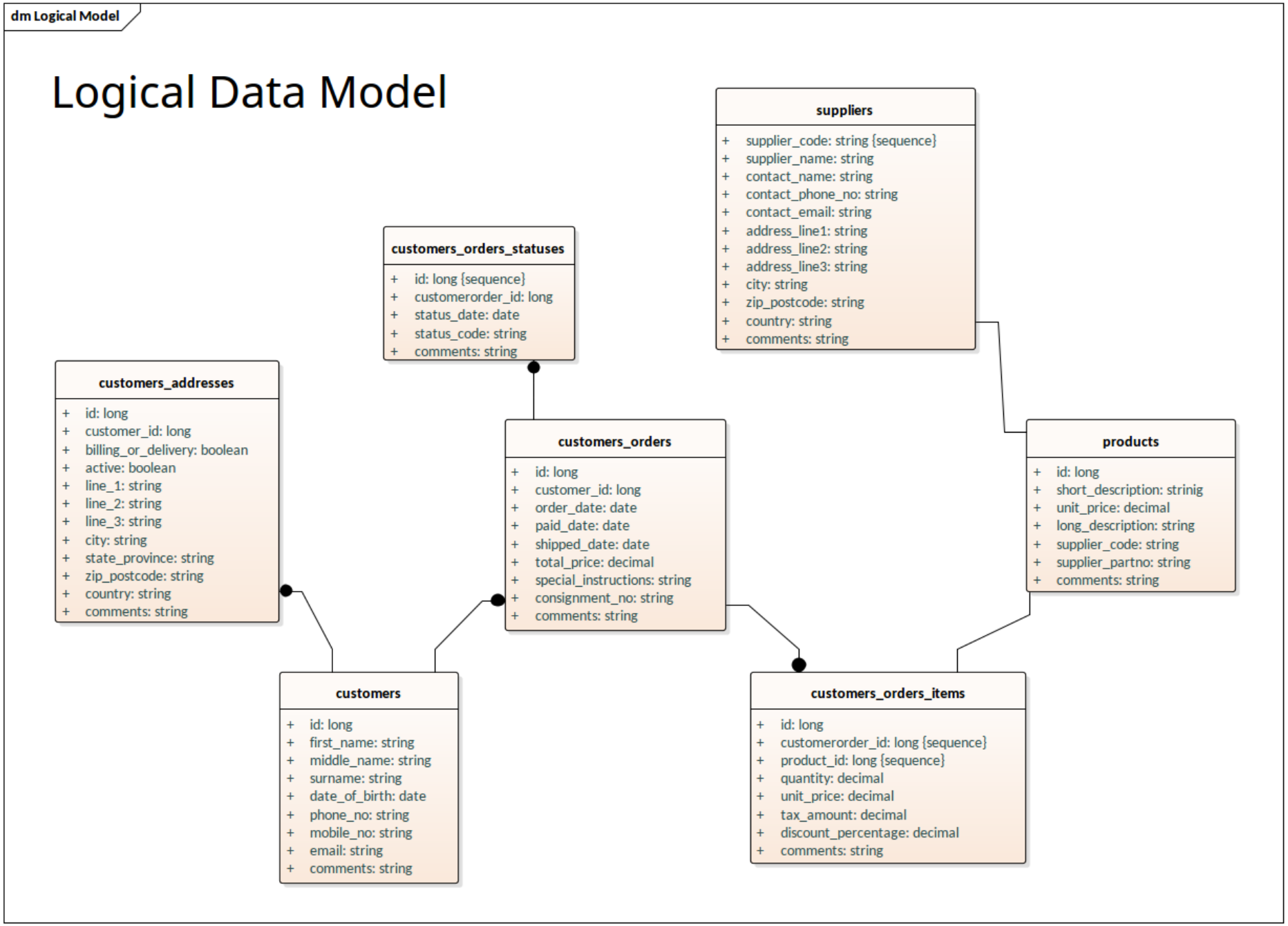
(taken from: http://www.cems.uwe.ac.uk/~pchatter/resources/html/common_data_modelling_notations.htmlhttp://www.cems.uwe.ac.uk/~pchatter/resources/html/common_data_modelling_notations.html, last visited Dec-21)

Example – LDM Information Engineering Notation



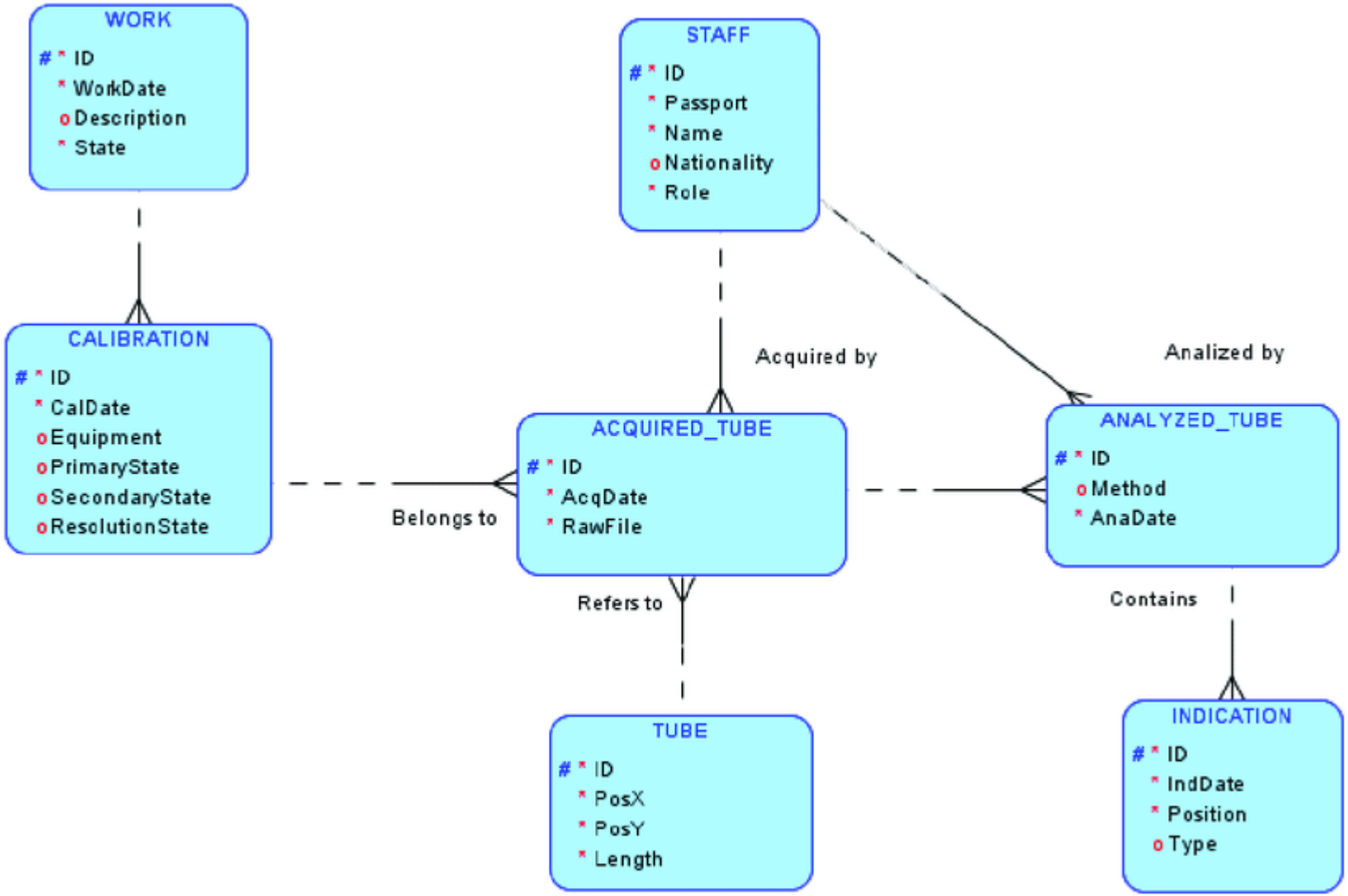
Taken from https://sparxsystems.com/resources/gallery/diagrams/software/sw-logical-data-model-info-eng-notation.html , last visited Dec. 21)

Example – LDM IDEF1X Notation



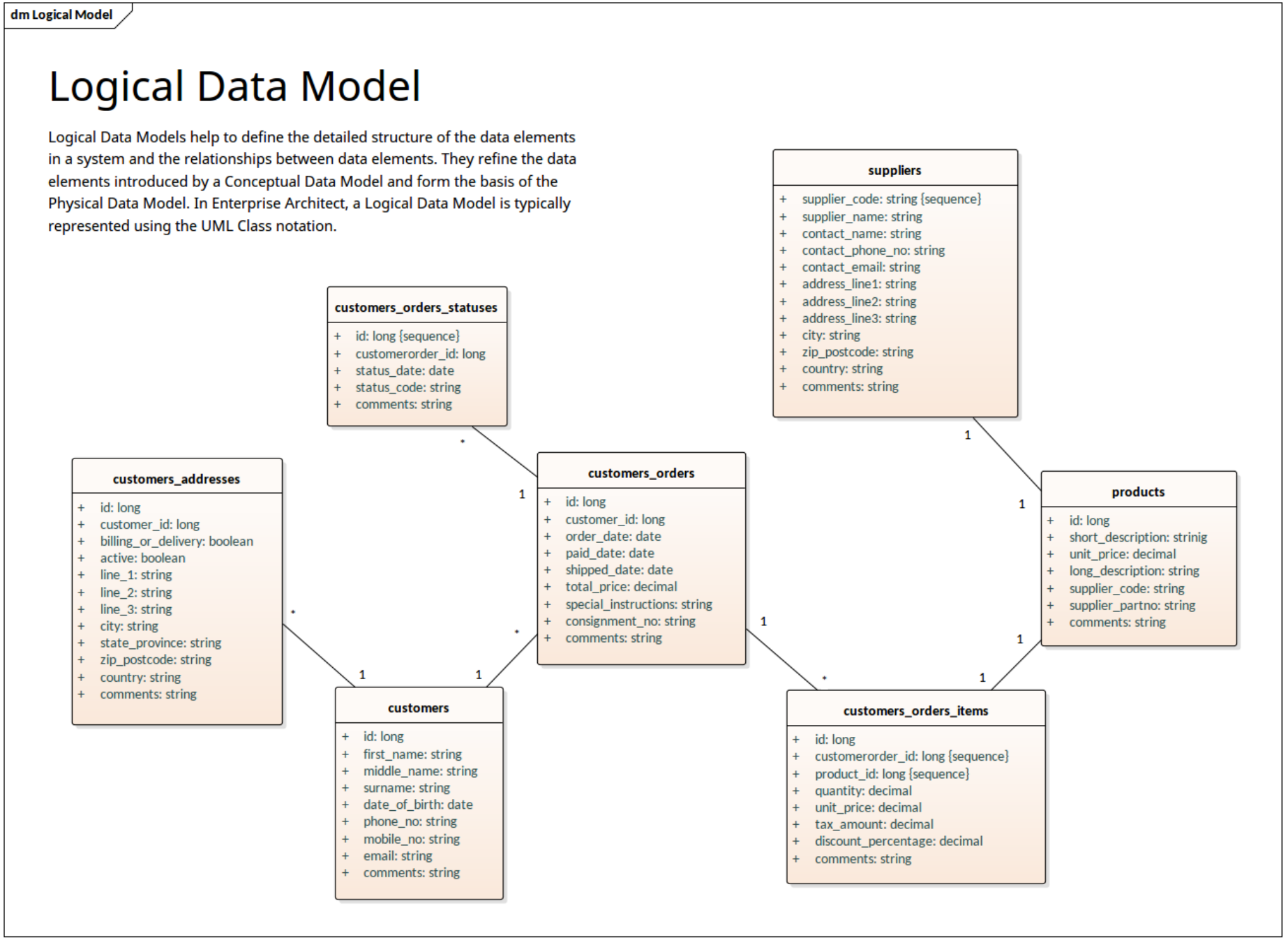
Taken from https://sparxsystems.com/resources/gallery/diagrams/software/sw-logical_data_model-idef1x.html , last visited Dec. 22)

Example – LDM Barker Notation



Taken from https://www.researchgate.net/figure/Simplified-entity-relationship-diagram-barker-notation-of-the-data-needed-in-a-typical_fig2_347449294 , last visited Dec. 22)

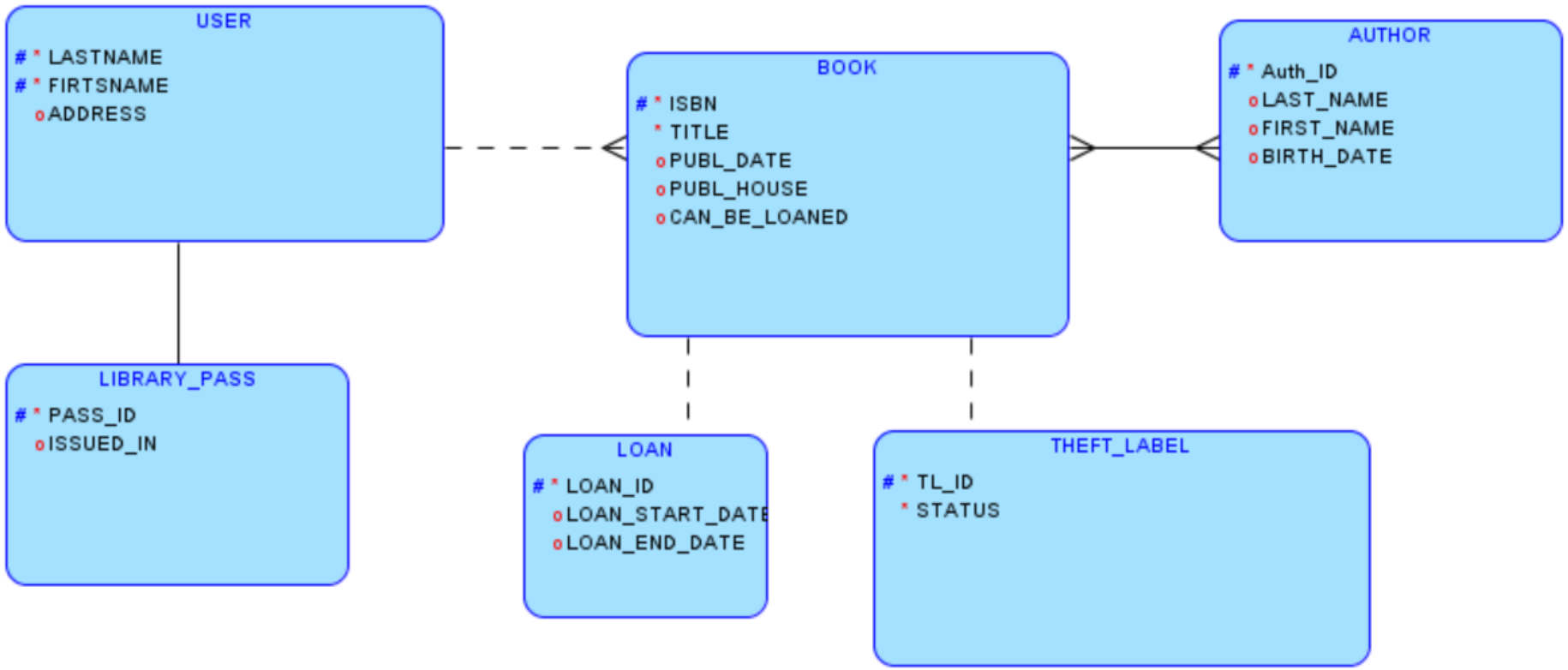
Example – LDM UML Notation



Taken from https://sparxsystems.com/resources/gallery/diagrams/software/sw-logical_data_model-idef1x.html , last visited Dec. 22)

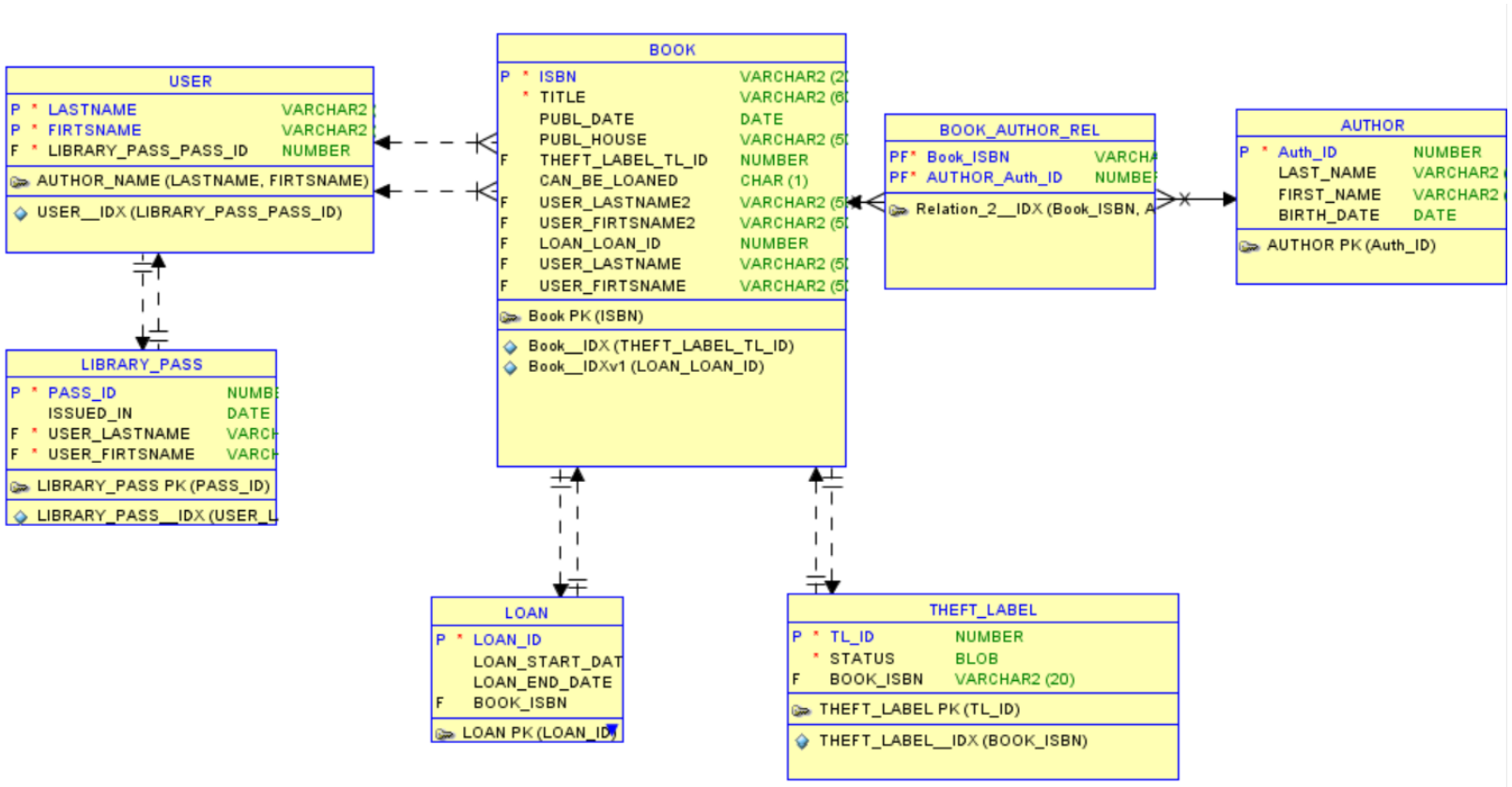
Possible Solutions of the Library Example

Object analysis
Possible Solution (LDM) – Step 3 (DM draft)



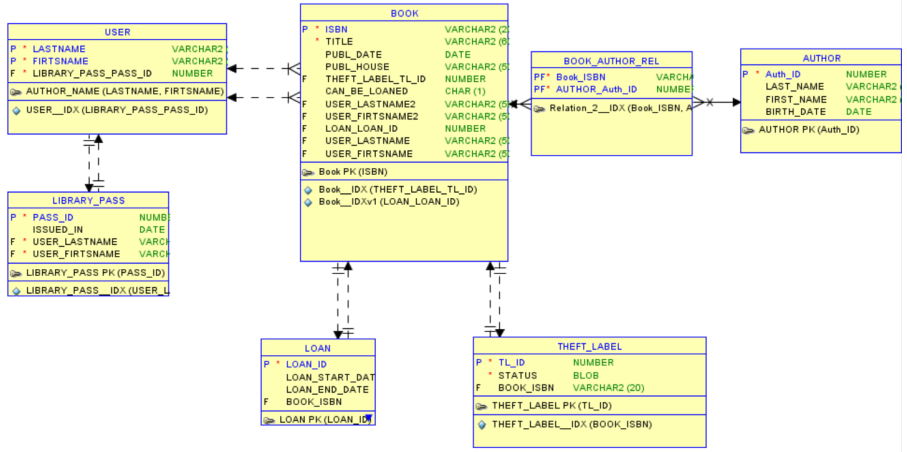
Barker Notation (designed in: Oracle SQL Developer Data Modeler)

Object analysis
Possible Solution (PDM) – Step 4 (DM draft)



designed in: Oracle SQL Developer Data Modeler

Object analysis
Possible Solution (PDM) – Step 4 (DM draft)



From a PDM you can automatically generate the Data Definition SQL Syntax to generate the Data Base Schema (Oracle SQL Developer Data Modeler)

```
CREATE TABLE author (
    auth_id     NUMBER NOT NULL,
    last_name   VARCHAR2(50),
    first_name  VARCHAR2(50),
    birth_date  DATE
);

ALTER TABLE author ADD CONSTRAINT "AUTHOR PK" PRIMARY KEY ( auth_id );

CREATE TABLE book (
    isbn              VARCHAR2(20) NOT NULL,
    title             VARCHAR2(60) NOT NULL,
    publ_date         DATE,
    publ_house        VARCHAR2(50),
    theft_label_tl_id NUMBER,
    can_be_loaned     NUMBER,
    user_lastname2    VARCHAR2(50),
    user_firtsname2   VARCHAR2(50),
    loan_loan_id      NUMBER,
    user_lastname     VARCHAR2(50),
    user_firtsname    VARCHAR2(50)
);

CREATE UNIQUE INDEX book__idx ON
    book (
        theft_label_tl_id
    ASC );

CREATE UNIQUE INDEX book__idxv1 ON
    book (
        loan_loan_id
    ASC );

ALTER TABLE book ADD CONSTRAINT "Book PK" PRIMARY KEY ( isbn );

CREATE TABLE book_author_rel (
    book_isbn       VARCHAR2(20) NOT NULL,
    author_auth_id  NUMBER NOT NULL
);
```

Summary Questions (pot. exam questions)

- Terminology
  - Model, modeling, data model, data modeling, business process, ….
- What is a CDM, LDM and a PDM and how do they differ from each other?
- What are keys, what kind of keys can you name and what are their meaning?
- What types of relationships do you know
- Can you explain normalization. Explain1NF, 2NF and 3NF?
- What is the star schema? What is the setup of a star schema?
- Can you explain the steps required to create a data model?

Good reading resources

Make sure that you have the required background knowledge:

- Agile/Evolutionary Data modeling: From Domain modeling to Physical modeling
  http://www.agiledata.org/essays/agileDataModeling.html
- Agile Data modeling 101
  http://www.agiledata.org/essays/dataModeling101.html
- In addition: Normalization explained
  https://docs.microsoft.com/en-US/office/troubleshoot/access/database-normalization-description
  http://en.wikipedia.org/wiki/Database_normalization

Hungry Mind assignment

- Define a logical model of the business process: “selling books online”
  - Identify all entity attributes / keys (primary and foreign) / relationships (cardinality)
  - Use – if possible – a modelling notation (I would recommend the Barker notation – see previous slide)
- If you really want the extra challenge
  - Download from the Oracle Technology network the tool: SQL Developer Data modeler (it’s a great tool for data modelling – free of charge)

  https://www.oracle.com/tools/downloads/sql-data-modeler-downloads.html

  You can use this tool independent of an Oracle database to design a data model