

Základy forenzních databází

Tereza Uhlíková

verze y25

Kdo jsem

Tereza Uhlíková

Ústav analytické chemie

skupina teoretické spektroskopie

místnost A277

<https://web.vscht.cz/~uhlikovt/>

tereza.uhlikova@vscht.cz

1. lekce

- Co je to databáze
- Proč, kdo, kdy, jak ...
- Něco málo z historie
- CAP problém

Dobrodružství kriminalistiky - televizní seriál
Dějiny psané Římem - Vojtěch Zamarovský

2. lekce

- Základy informatiky
- Software
 - informace
 - záznam informace
 - číselné soustavy
 - písmenné kódy
- Hardware
 - Alan Turing, John von Neumann a počítač
 - historie vývoje počítače & super počítač
 - procesor a datová uložště

Alan Turing - Enigma
Blade Runner

3. lekce

velká uložště

Enigma

RSA šifra

- Algoritmus
 - vlastnosti (zápis, struktura)
 - Datové typy
 - strukturované programovací jazyky

Arthur C. Clarke - Devět miliard božích jmen

4. lekce

- Strukturované programovací jazyky
- Výroková logika
- Složitost

Aghata Christie - Hercules Poirot

O čem budeme mluvit dnes

- Databáze
 - DB a SŘBD
 - databázové modely
- Relační model
 - klíče
 - kardinalita vztahu

??

Úvod do databáze

Databáze neboli datová základna (Data Base) je místo, kam se ukládají **určitým** způsobem organizované a strukturované údaje.

= data i software

Úvod do databáze

Databáze neboli datová základna (Data Base) je místo, kam se ukládají **určitým** způsobem organizované a strukturované údaje.

= data i software

DBMS - DataBase Management System

SŘBD - Systém Řízení Báze Dat

je softwarové vybavení, které zajišťuje práci s databází.

Úvod do databáze

Databáze neboli datová základna (Data Base) je místo, kam se ukládají **určitým** způsobem organizované a strukturované údaje.

= data i software

DBMS - DataBase Management System

SŘBD - Systém Řízení Báze Dat

je softwarové vybavení, které zajišťuje práci s databází.

- Definovat bázi

Úvod do databáze

Databáze neboli datová základna (Data Base) je místo, kam se ukládají **určitým** způsobem organizované a strukturované údaje.

= data i software

DBMS - DataBase Management System

SŘBD - Systém Řízení Báze Dat

je softwarové vybavení, které zajišťuje práci s databází.

- Definovat bázi
- Konstruovat databázi

Úvod do databáze

Databáze neboli datová základna (Data Base) je místo, kam se ukládají **určitým** způsobem organizované a strukturované údaje.

= data i software

DBMS - DataBase Management System

SŘBD - Systém Řízení Báze Dat

je softwarové vybavení, které zajišťuje práci s databází.

- Definovat bázi
- Konstruovat databázi
- Manipulovat s databází

Úvod do databáze

Databáze neboli datová základna (Data Base) je místo, kam se ukládají **určitým** způsobem organizované a strukturované údaje.

= data i software

DBMS - DataBase Management System

SŘBD - Systém Řízení Báze Dat

je softwarové vybavení, které zajišťuje práci s databází.

- Definovat bázi
- Konstruovat databázi
- Manipulovat s databází

databázový systém x programovací jazyk

Příklady DBMS

Příklady DBMS

některých

- Microsoft Access
- MySQL
- SQLite
- PostgreSQL
- Oracle
- Informix
- DB2
- Sybase Adaptive Server Enterprise
- FileMaker
- Microsoft SQL Server
- Microsoft Visual FoxPro
- Progress
- CSQL
- OpenLink Virtuoso

Příklady programovacích jazyků

Příklady programovacích jazyků

některých

- Cobol
- SQL
- Python
- C
- Pascal
- C++
- PHP
- Baltík
- Basic
- Delphi
- PostScript
- Fortran
- Algol
- Matlab
- Perl

Data versus informace

Data versus informace

Data - konkrétní hodnoty (údaje), syntaktická složku

Informace - znalost, vyplývající z hodnot uložených v databázi, sémantická složka

Data versus informace

Data - konkrétní hodnoty (údaje), syntaktická složka

Informace - znalost, vyplývající z hodnot uložených v databázi, sémantická složka

Databázový model - způsob, jakým jsou data strukturována, propojena a spravována.

⇒ efektivní ukládání, dotazování, sdílení a integrita dat.

Proč má DNA databáze jinou strukturu než databáze otisků prstů?

DNS vs. Otisk

Proč má DNA databáze jinou strukturu než databáze otisků prstů?



Historický vývoj databázových modelů

Éra	Model	Znak pro období	Charakteristika
1860	Lieární	Carl Lines	textový soubor
1960	Hierarchický	IBM IMS	Stromová struktura, pevné vazby
1970	Síťový	CODASYL COBOL	Komplexní vztahy, efektivní přístup
1980	Relační	E. F. Codd (IBM)	Tabulky, SQL, standardizace
1990	Objektový	Rozšíření relačního	multimediální data
2000	NoSQL / Big Data	MongoDB, Neo4j, Cassandra	Flexibilita, škálovatelnost, nestrukturovaná data

<https://bytebytego.com/guides/top-6-database-models/>

Hierarchický model

- stromová struktura
- vztahy podřízenosti a nadřízenosti
- každý nadřazený může mít více podřízených
- každý podřízený má právě jednoho nadřazeného
- vhodné pro modelování typu část/celek
- Vhodné pro data s jednoznačnou hierarchií (např. organigram, katalogy).

aplikace – rodokmen, evidence součástí v projektu Apollo, evidence vzorků → hlavní případ → podvzorky.

- Nevhodné pro data s mnoha vztahy (např. jeden pachatel ↔ více případů).

Síťový model

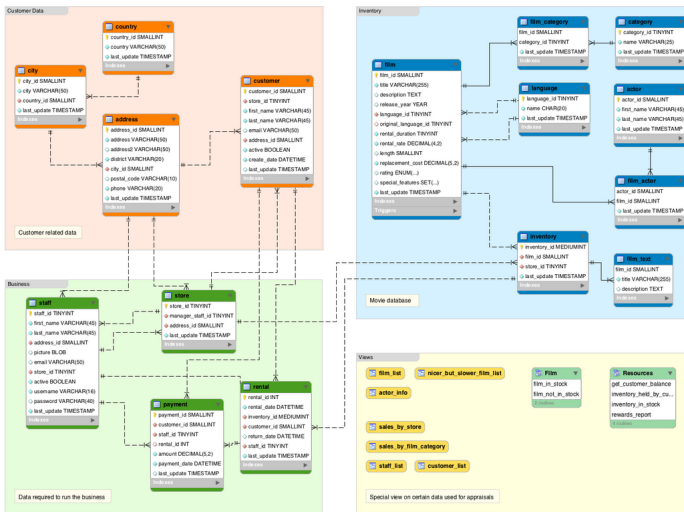
- orientovaný graf
- každý podřízený může mít více nadřízených
- rychlý přístup, ale složitá navigace.
- Forenzní využití: propojení pachatelů, stop a případů.

geografické informační systémy, propojení pachatelů, stop a případů.

Relační model

- data v tabulkách (relacích)
- každá tabulka je přímo přístupná
- každá tabulka může mít více “nadřazených i podřazených”
- všechny tabulky mohou být navzájem propojené - každá tabulka má klíč (Primary Key), který se dá spojit s jinými (Foreign Key)
- dotazovací jazyk SQL
- přehlednost, integrita, standardizace
- Nevhodné pro velmi nestrukturovaná data (např. obrazové záznamy z kamer)

Jak vypadá relační databáze



Objektový a relačně-objektový model

- třídídimensionální struktura
- objekty = data + metody
- mezi objekty existuje skládání, dědění, závislost, klasifikace podle tříd, ... strukturované informace není třeba rozdělovat jako v RDM
- protokol objektu je dán množinou přístupných zpráv (ne atributů jako v RMD)
- jedna množina (objektů) může s využitím polymorfismu obsahovat objekty s různou strukturou dat i metod
- je rozdíl mezi množinou objektů a třídou
- identita objektu je dána nejen vnitřními, ale i vnějšími vazbami, klíče jsou interní záležitostí

NoSQL modely

- pro velmi rozsáhlá, nestrukturovaná nebo rychle se měnící data
- Typy: dokumentové (MongoDB), grafové (Neo4j), klíč–hodnota (Redis), sloupcové (Cassandra)
- analýza kyberkriminality, korelace IP adres, logů, sociálních sítí, velká data
- výhoda - extrémní výkon a flexibilita
- nevýhoda - slabší datová integrita

Společná práce

Navrhněte DNA databázi a databázi otisků prstů? V čem by mohli být shodné a v čem se lišit. Jaký model by byl pro tu kterou databázi vhodnější a proč?

Společná práce

Navrhněte DNA databázi a databázi otisků prstů? V čem by mohli být shodné a v čem se lišit. Jaký model by byl pro tu kterou databázi vhodnější a proč?

DNA databáze a databáze otisků prstů mají odlišnou datovou povahu, způsob porovnávání i požadavky na uložení, jiný cíl...

Proto potřebují odlišnou strukturu databáze, optimalizovanou pro konkrétní typ informací.

DNA vs. Otisk

Parametr	DNA	Otisk
Data	Sekvence číselných markerů numerická a kódovatelná data	Obrazová data - digitální reprezentace linií, textur
Způsob porovnávání	Statistické srovnání shoda jednotlivých markerů	Geometrické a obrazové algoritmus hledá vzor a polohu linií
Velikost dat	desítky čísel → malý objem dat	několik MB → nutnost efektivního ukládání obrazů a metadat
Cíl	Vysoce přesné genetické identifikace indexace podle markerů.	Rychlé vizuální vyhledávání a porovnávání obrazových vzorů.
algoritmy	Matematické (pravděpodobnostní, statistické	Geometrické (vzorové rozpoznávání, AI / deep learning)
Datová struktura	relační model osoba – profil – marker	objektově orientovaný bitmapy, vektory, 3D skeny

NÁVRH TABULEK

NÁVRH TABULEK

Relace, Tabulka, Entita

Relace, Tabulka, Entita

Co je to relace?

- spojení, souvislost, vztah
 - v tabulce jsou data v relaci mezi sloupcem a řádkem
- => tabulka

Objekt reality, který je zapsán pomocí jedné tabulky => entita

Relace $\approx$ Tabulka $\approx$ Entita

relace (tabulka)	sloupec			
	Název sloupce 1	Název sloupce 2	...	Název sloupce N
řádek				
		pole		

Tabulka obsahuje

Atribut: sloupec

- určují vlastnosti objektů (příjmení, fakulta,...)
- jedinečný název a - nepoužívat diakritiku a mezery, tedy např.:
nazev_odberatele
- určený datový typ (číslo, text, logická hodnota, ...)

Záznam: řádek

- každý záznam je jednoznačně rozlišitelný - jeden řádek reprezentuje
např. jednoho zaměstnance s hodnotami daných atributů
- mít svůj jedinečný identifikátor

id_zamestnance	prijmeni	jmeno	plat	datum_nastupu
1	Novák	Adam	32000	2013-05-02
2	Nová	Jana	45000	2013-06-11

Společná práce

Navrhni tabulku pro lukostřelecké závody:

Dva lukostřelci Pavel a Michal byly na závodech. Pavel měl tři pokusy a stěfil se postupně na čísla 4, 9 a 2.

Michal se trefil na čísla 1, 5 a 7.

Možné řešení

vícehodnotná položka

strelec	zasahy
Pavel	4, 9, 2
Michal	1, 5, 7

Identifikujte pole - při kolikátém zásahu se Pavel trefil nejbližše 10?

Upravené lepší řešení

vícehodnotná položka

špatně

správně

strelec	zasahy
Pavel	4, 9, 2
Michal	1, 5, 7

strelec	zasah1	zasah2	zasah3
Pavel	4	9	2
Michal	1	5	7

Návrh tabulek

vícehodnotná položka

špatně

správně

strelec	zasahy
Pavel	4, 9, 2
Michal	1, 5, 7

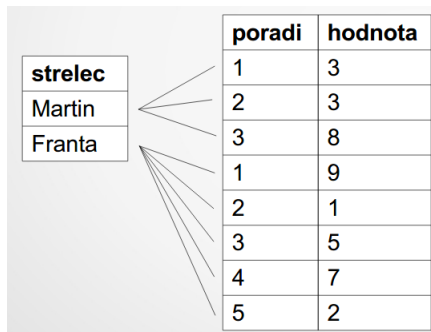
strelec	zasah1	zasah2	zasah3
Pavel	4	9	2
Michal	1	5	7

různý počet pokusů

strelec	z1	z2	z3	z4	z5	z6	z7
Martin	3	3	8				
Franta	9	1	5	7	2		

Dělení tabulek

Lze však tabulky rozdělit.



Opakování jedné hodnoty

Lze tato tabulka zlepšit?

jmeno	mesto
Martin	Praha
Franta	Plzeň
Alena	Plzeň
Jana	Plzeň

Rozdělení tabulek

opakující se hodnoty
plýtvání místem
provede-li se změna, musí se opravit všude
špatně
správně



Určitelné informace

vypočítaná položka

- sloupec, který se dá vytvořit výpočtem z ostatních položek v záznamu, by neměl být součástí tabulky
- výpočet se provede v dotazu

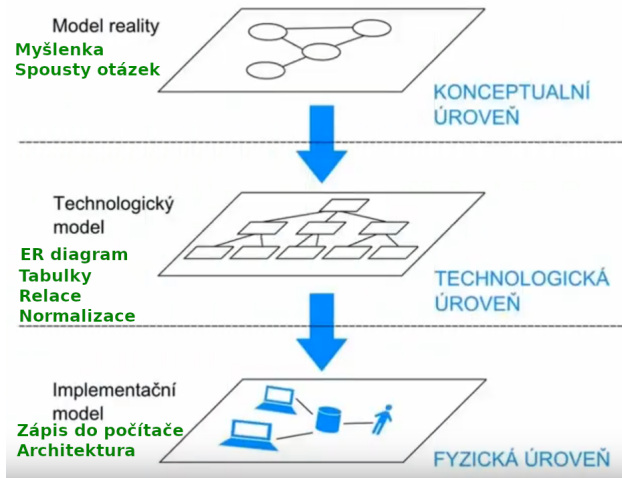
špatně

obdelnik	sirka	vyska	obvod	obsah
malý	2	4	10	8
velký	30	15	90	450

ERA MODEL

ERA MODEL

Návrh relační databáze



ERA model

- **Entity** (objekt z reálného světa) → tabulka
- **Relationship** (vztah mezi entitami) → vztahy mezi tabulkami
- **Attribute** (vlastnosti entit) → sloupeček tabulky

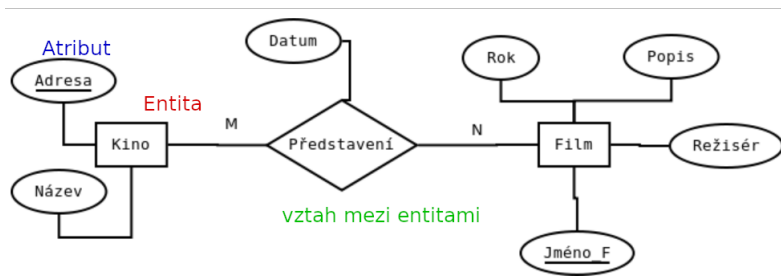
Modeluje „oblast zájmu“ (např. podnik) jako kolekci entit a vztahů (relationships) mezi nimi

- grafická reprezentace struktury databáze

dělá se před vytvářením databáze!! a slouží k rozmyšlení, jak bude databáze vypadat

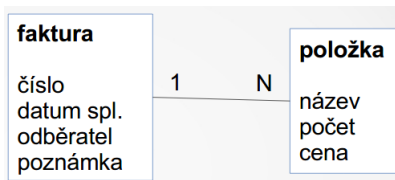
ERA model

Klasický formát

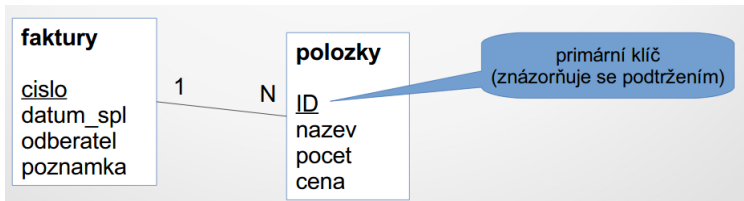


ERA model

tabulková varianta



implementační



Společná práce

Navrhněte ERA model pro databázi DNA vzorků.

Společná práce

Navrhněte ERA model pro databázi DNA vzorků.
Model by měl umožnit uchovávat informace o:

Společná práce

Navrhněte ERA model pro databázi DNA vzorků.

Model by měl umožnit uchovávat informace o:

Osobách (ID, jméno, pohlaví, datum narození, stát původu),

Vzorcích DNA (ID vzorku, datum odběru, místo odběru, stav analýzy),

Laboratořích (název, adresa, kontaktní osoba),

Analýzách (metoda, datum, výsledek, identifikátor osoby),

Shodách (matches) mezi vzorky z různých případů.

Uřčete entity, jejich atributy a vazby („má vzorek“, „analyzuje“, „obsahuje shodu“).

Co by se stalo, kdyby se změnila struktura pro uložení STR markerů?

Klíče a integrita

KLÍČE A INTEGRITA

Udržuje logickou správnost dat

Primární klíč

Entitní – jedinečná identifikace každé entity - pomocí primárního klíče (PK)

Primární klíč

- zajistí jednoznačnost záznamů v tabulce,
- AUTO INCREMENT - zajistí automatické číslování => hodnota primárního klíče se v tabulce nesmí opakovat
- data nesmí obsahovat duplicitu nebo hodnotu Null.
- teoreticky může být libovolného typu
- může být složený z více sloupců – unikátnost je pak brána ze všech najednou, v jednotlivých sloupcích se tedy mohou opakovat stejné hodnoty

Referenční klíč

Referenční – pomocí cizího klíče (FK)

Cizí klíč

- odkazuje se na primární klíč v jiné tabulce – odkazuje se na právě 1 tabulku
- dvojice primární a cizí klíč tvoří vazbu (relaci) mezi tabulkami
- cizí klíč musí být stejného typu a velikosti jako primární

if primární klíč AUTO INCREMENT => cizí klíč dlouhé celé číslo

if primární klíč složený => struktura cizího klíče stejná

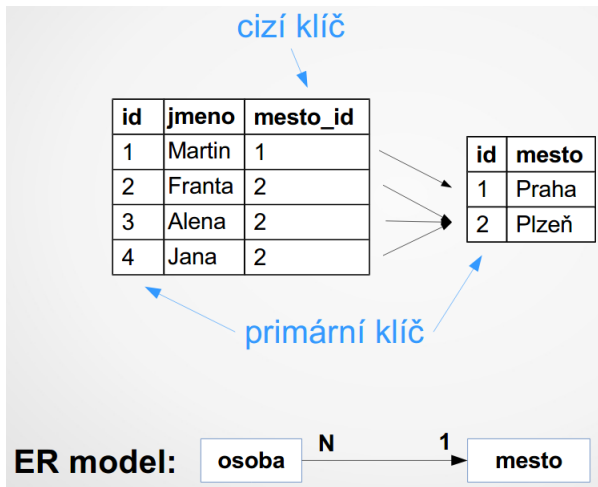
Do sloupce s definovaným cizím klíčem nelze vložit hodnotu, která není vložena v nadřazeném sloupci primární tabulky.

Restrict – z nadřazené nelze smazat, je-li odkaz

Cascade – po výmazu z nadřazené smaže i odkazy v podřazené

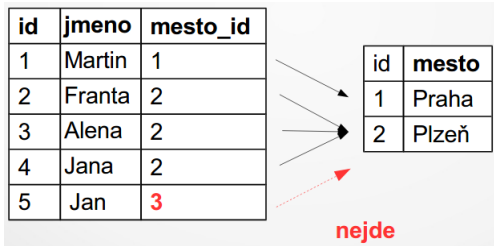
Set null – po výmazu v nadřazené dostanou odkazy v podřazené hodnotu NULL

Klíče



Vztahy klíčů

databáze nedovolí do cizího klíče vložit hodnotu, která neexistuje v primárním klíči v cílové tabulce



Kardinalita vztahu

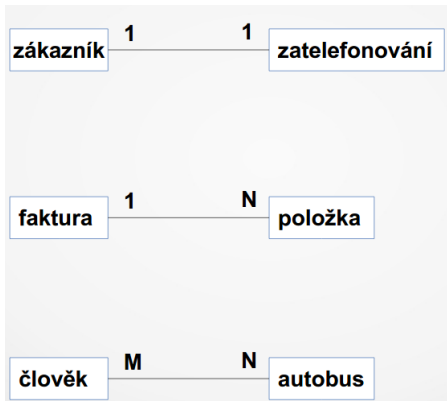
mocnost vztahu mezi entitami
neboli

Kolik položek z tabulky A se váže na kolik položek z tabulky B?

3 typy

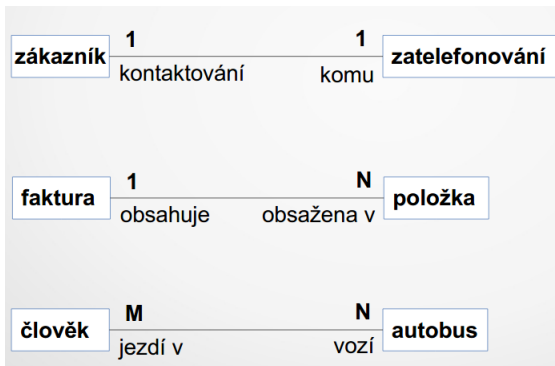
- 1:1 - na 1 položku z tabulky A připadá maximálně 1 položka z tabulky B, ta patří jen té jediné z A, žena-muž (ČR)
- 1:N, N:1 - na 1 položku z tabulky A připadá libovolné množství položek z tabulky B, na 1 položku z tabulky B připadá maximálně 1 položka z tabulky A
- M:N - bez omezení množství vazeb v obou směrech, žena-muž (ČR) v čase

Kardinalita vztahu



Kardinalita vztahu

pomůže lepšímu porozumění ER(A) modelu



Kardinalita 1:N

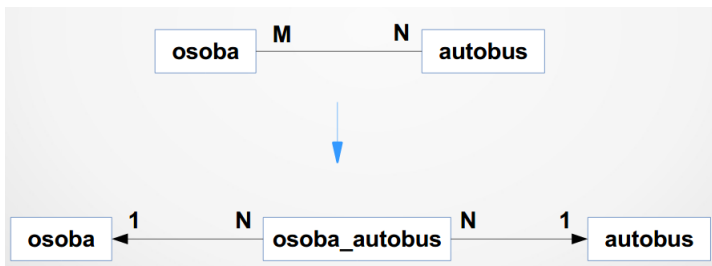
- jedné entitě odpovídá několik (N) jiných entit
- např. vztah třída - žák
 - jeden žák navštěvuje jednu třídu
 - jedna třída se skládá z více žáků

Kardinalita M:N

- několika (M) entitám odpovídá několik (N) jiných entit
- např. vztah student - předmět

- jeden student studuje více předmětů
 - jeden předmět navštěvuje více studentů
 - nejsložitější případ vazby

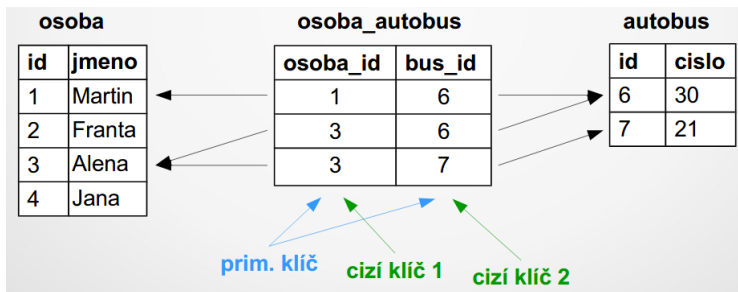
vazbu M:N nelze vytvořit přímo mezi tabulkami
je potřeba rozložit pomocí nové tabulky na dvě vazby 1:N



Kardinalita M:N

rozkladová tabulka často neobsahuje další informace

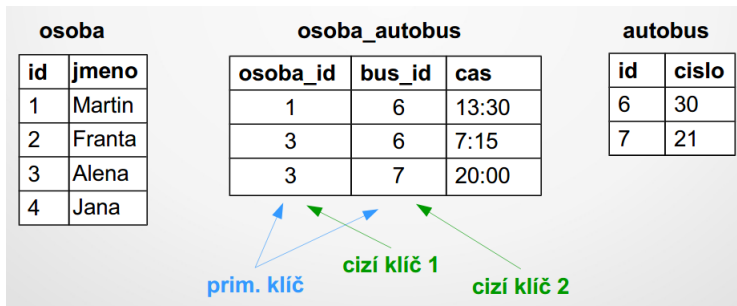
primární klíč v rozkladové tabulce se často skládá z obou cizích klíčů (ale může být i vlastní sloupec id)



Kardinalita M:N

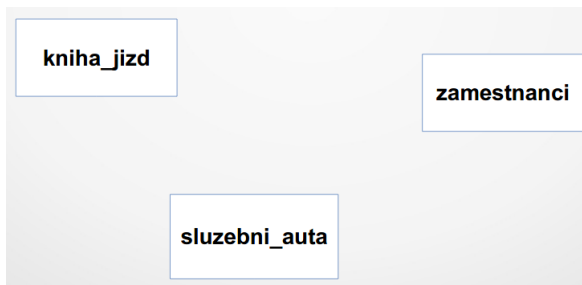
spojovací tabulka může obsahovat další informace

- např. v kolik kdo jakým autobusem jezdí
- není nutné kvůli tomu vytvářet další tabulku



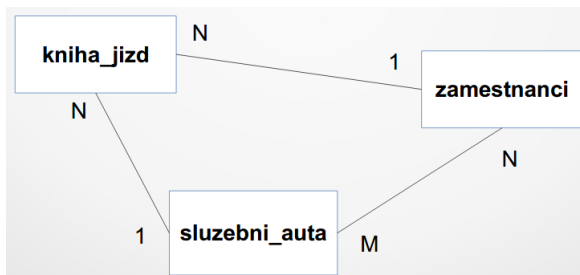
Příklad

doplňte vazby a jejich kardinality do ER modelu – zaměstnanci používají služební auta a zapisují do knihy jízd (každý zaměstnanec pro každé auto svou knihu jízd)



Kardinalita vztahu

výsledek



Společná práce

Do vašeho ERA modelu pro databázi DNA vzorků přidejte primární klíče, cizí klíče.

Společná práce

Do vašeho ERA modelu pro databázi DNA vzorků přidejte primární klíče, cizí klíče.

Osoba (ID osoby),

Vzorek DNA (ID vzorku, FK = ID osoby, FK = analytik),

Laboratoř (ID lab),

Analytik (ID analytik, FK = ID lab)

Analýza (ID analiza, FK = ID vzorku, FK = ID analytik),

Shoda (ID shoda, FK = ID vzorku).

Co se stane, když smažeme osobu, která má přidělené vzorky?

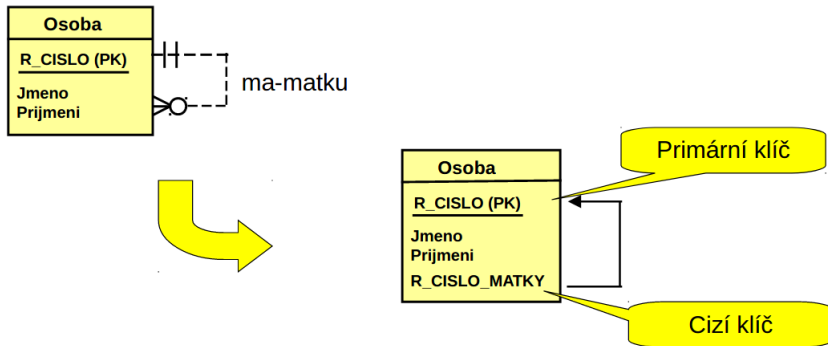
Jak byste zajistili, aby každý vzorek měl přiřazenou laboratorní analýzu?

Co by se stalo, kdyby dva záznamy měly stejný primární klíč?

Rekurze

vztah mezi osobou a matkou dané osoby

Matka osoby je rovněž osobou, nemůže být tudíž v jiné tabulce!



Skripta, stránky a materiály

https://www.fi.muni.cz/~popel/lectures/bak_logika/

<https://www.siliconrepublic.com/science/logic-maths-puzzle-george-boole>

<http://www.di.unito.it/~anselma/pdf/preprintBoole.pdf>

Simultaneous Equations and Boolean Algebra in the Analysis of Judicial Decisions

Fred Kort Law and Contemporary Problems Vol. 28, No. 1, Jurimetrics (Winter, 1963), pp. 143-163 (21 pages)

https://en.wikipedia.org/wiki/Truth_table

https://sites.ff.cuni.cz/uisk/wp-content/uploads/sites/62/2016/01/Datov%C3%A9-modely-a-n%C3%A1vrhy-rela%C4%8Dn%C3%ADch-sch%C3%A9mat_Sk%C5%99ivan.pdf

<http://www.databaze.chytrak.cz/modely.htm>

<http://books.fs.vsb.cz/dbacc20/Welcome.htm>

<http://books.fs.vsb.cz/>

<http://projekty.fs.vsb.cz/463/edubase/>

[VY_01_042/Z%C3%A1klady%20informatiky.pdf](http://projekty.fs.vsb.cz/463/edubase/VY_01_042/Z%C3%A1klady%20informatiky.pdf)

http://projekty.fs.vsb.cz/463/edubase/VY_01_044/Datab%C3%A1zov%C3%A9%20syst%C3%A9my%202005.%20Normalizace%20dat.pdf

<https://bytebytego.com/guides/top-6-database-models/>