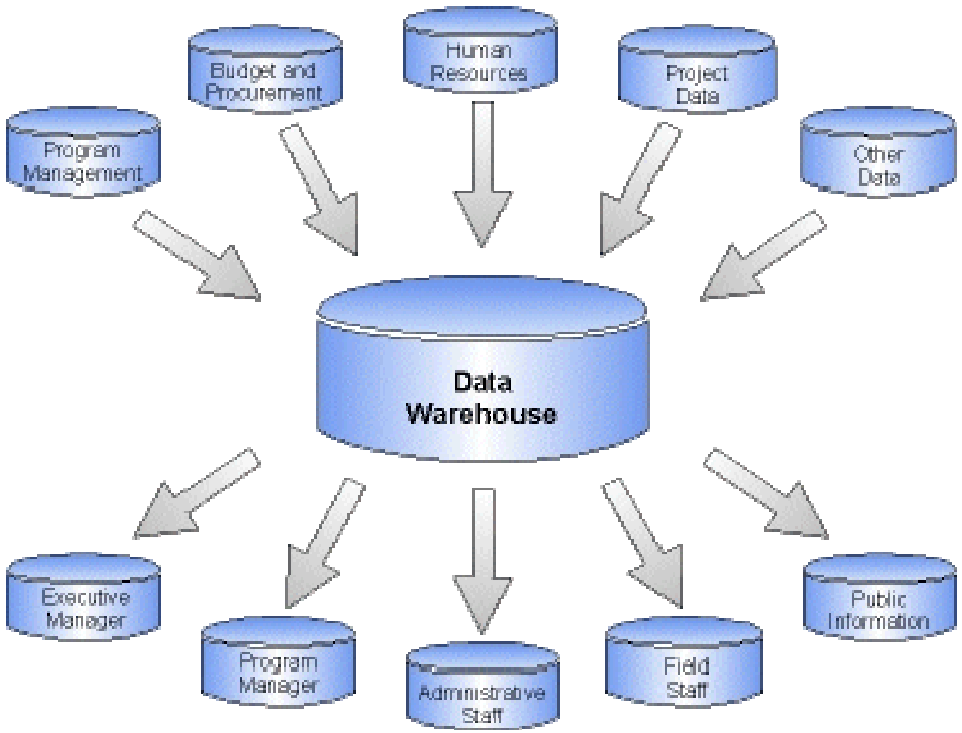
Organizace dat



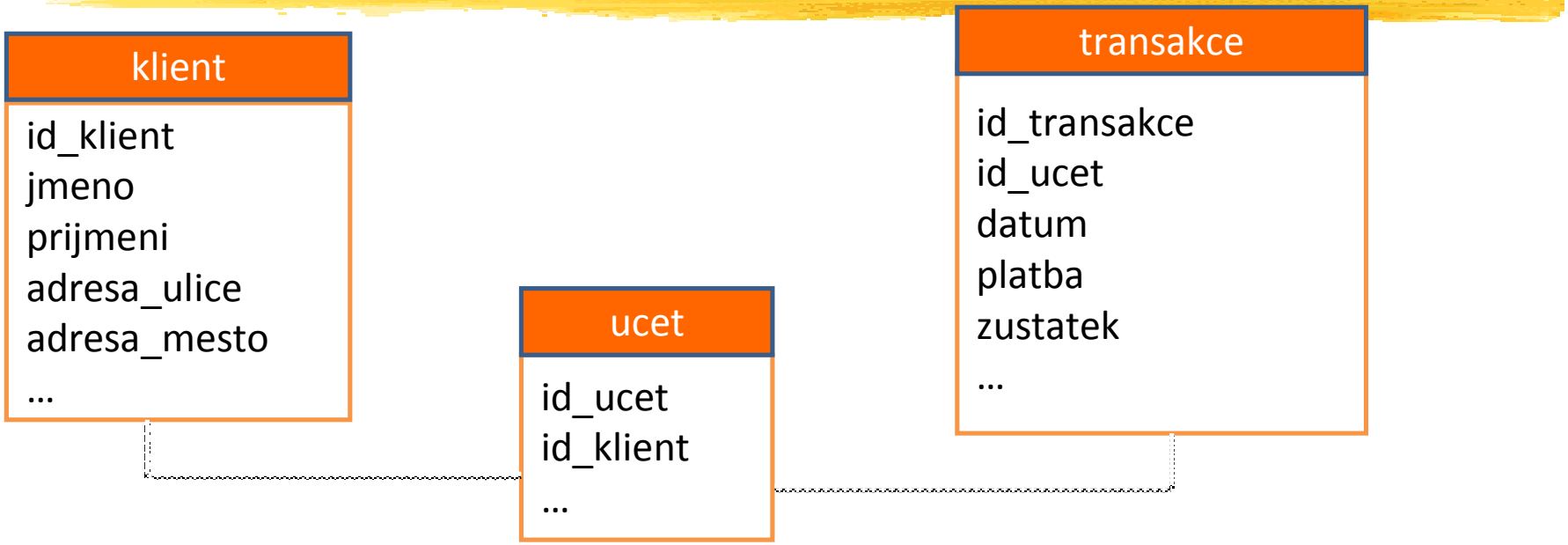
Historie skladování dat

V minulosti byla data ukládána v jednom velkém souboru, ke kterému se přistupovalo indexovanými sekvenčními metodami. Soubor byl indexován na základě předpokládaných způsobů dotazování. Velkou nevýhodou bylo to, že se informace v záznamech opakovaly a typy dotazů byly předurčeny.

Historie skladování dat

datum	jmeno	prijmeni	adresa_ulice	adresa_mesto	cislo_uctu	platba	zustatek
980103	Jan	Novak	Dlouha 5	Praha 1	9945371	100,00	100,00
980105	Jan	Novak	Dlouha 5	Praha 1	9945371	1500,00	1600,00
980106	Jan	Novak	Dlouha 5	Praha 1	9945371	-1500,00	50,00
980106	Karel	Nemec	Lucni 4	Praha 2	24867134	3000,00	6000,00
980107	Karel	Nemec	Lucni 4	Praha 2	24867134	-4000,00	2000,00
980108	Jan	Novak	Dlouha 5	Praha 1	9945371	-150,00	-100,00
980111	Karel	Nemec	Lucni 4	Praha 2	24867134	5000,00	7000,00

Relační databáze




```
SELECT klient.jmeno, klient.prijmeni, klient.adresa_ulice,
klient.adresa_mesto, ucet.cislo_uctu, transakce.zustatek
FROM klient, ucet, transakce
WHERE klient.id_klient = ucet.id_klient;
AND transakce.id_ucet = ucet.id_ucet;
AND transakce.zustatek < 100;
GROUP BY klient.adresa_mesto;
```

Relační databáze

- **Relační databáze** je databáze založená na relačním modelu. Často se tímto pojmem označuje nejen databáze samotná, ale i její konkrétní softwarové řešení.
- Relační databáze je založena na tabulkách, jejichž řádky obvykle chápeme jako záznamy a eventuelně některé sloupce v nich (tzv. cizí klíče) chápeme tak, že uchovávají informace o relacích mezi jednotlivými záznamy v matematickém slova smyslu.
- Termín *relační databáze* definoval Edgar F. Codd v roce 1970.

- způsoby kladení dotazů:
  - QBE (query by example)
  - SQL (structured query language)

Relační databáze

- Dle relační teorie lze pomocí základních operací (sjednocení, kartézský součin, rozdíl, selekce, projekce a spojení) uskutečnit veškeré operace s daty a ostatní operace jsou již jen kombinacemi těchto pěti.

Relační databáze

- Základem relačních databází jsou databázové tabulky. Jejich sloupce se nazývají atributy nebo pole, řádky tabulky jsou pak záznamy. Atributy mají určen svůj konkrétní datový typ - doménu. Řádek je řezem přes sloupce tabulky a slouží k vlastnímu uložení dat. Konkrétní tabulka pak realizuje podmnožinu kartézského součinu možných dat všech sloupců - relaci.
- **Primární klíč**
  - Primární klíč je jednoznačný identifikátor záznamu, řádku tabulky. Primárním klíčem může být jediný sloupec či kombinace více sloupců tak, aby byla zaručena jeho jednoznačnost. Pole klíče musí obsahovat hodnotu, tzn. nesmí se zde vyskytovat nedefinovaná prázdná hodnota NULL. V praxi se dnes často používají umělé klíče, což jsou číselné či písmenné identifikátory - každý nový záznam dostává identifikátor odlišný od identifikátorů všech předchozích záznamů (požadavek na unikátnost klíče), obvykle se jedná o celočíselné řady a každý nový záznam dostává číslo vždy o jednotku vyšší (zpravidla zcela automatizovaně) než je číslo u posledního vloženého záznamu (číselné označení záznamů s časem stoupá).
- **Cizí klíč**
  - Dalším důležitým pojmem jsou nevlastní/cizí klíče. Slouží pro vyjádření vztahů, relací, mezi databázovými tabulkami. Jedná se o pole či skupinu polí, která nám umožní identifikovat, které záznamy z různých tabulek spolu navzájem souvisí.

Relační databáze – vztahy mezi tabulkami

- Vztahy, neboli relace, slouží ke svázání dat, která spolu souvisejí a jsou umístěny v různých databázových tabulkách. V zásadě rozlišujeme čtyři typy vztahů.

- mezi daty v tabulkách není žádná spojitost, proto nedefinujeme žádný vztah.

- 1:1 používáme, pokud záznamu odpovídá právě jeden záznam v jiné databázové tabulce a naopak. Takovýto vztah je používán pouze ojediněle, protože většinou není pádný důvod, proč takovéto záznamy neumístit do jedné databázové tabulky. Jedno z mála využití je zpřehlednění rozsáhlých tabulek. Jako ilustraci je možné použít vztah řidič - automobil. V jednu chvíli (diskrétní časový okamžik) řídí jedno auto právě jeden řidič a zároveň jedno auto je řízeno právě jedním řidičem.

Relační databáze – vztahy mezi tabulkami

- 1:N přiřazuje jednomu záznamu více záznamů z jiné tabulky. Jedná se o nejpoužívanější typ relace, jelikož odpovídá mnoha situacím v reálném životě. Jako reálný příklad může posloužit vztah autobus - cestující. V jednu chvíli cestující jede právě jedním autobusem a v jednom autobuse může zároveň cestovat více cestujících.

- M:N je méně častým. Umožňuje několika záznamům z jedné tabulky přiřadit několik záznamů z tabulky druhé. V databázové praxi bývá tento vztah z praktických důvodů nejčastěji realizován kombinací dvou vztahů 1:N a 1:M, které ukazují do pomocné tabulky složené z kombinace obou použitých klíčů (třetí resp. tzv. vazební tabulka). Příkladem z reálného života by mohl být vztah výrobek - vlastnost. Výrobek může mít více vlastností a jednu vlastnost může mít více výrobků. V reálném životě nicméně existuje velké množství vztahů M : N, mimo jiné také proto, že často existuje praktická potřeba zachovávat i údaje o historii těchto vztahů z časového hlediska (jeden řidič v delším časovém období řídí více rozličných aut a jedno auto v delším časovém období může mít více různých řidičů).

Slovník pojmů

- ODS — Operational Data Store
- DWH — DataWareHouse
- DataMart
- Meta Data
- BI — Business Intelligence
- OLAP — On Line Analytical Processing
- OLTP — On Line Transaction Processing
- ETL — Extract, Transform, Load
- ELT — Extract, Load, Transform
- EAI — Enterprise Application Integration
- ERP — Enterprise Resource Planning

Slovník pojmů

ODS: Short for ***o****perational* ***d****ata* ***s****tore*, a type of database that serves as an interim area for a data warehouse in order to store time-sensitive operational data that can be accessed quickly and efficiently. In contrast to a data warehouse, which contains large amounts of static data, an ODS contains small amounts of information that is updated through the course of business transactions. An ODS will perform numerous quick and simple queries on small amounts of data, such as acquiring an account balance or finding the status of a customer order, whereas a data warehouse will perform complex queries on large amounts of data. An ODS contains only current operational data while a data warehouse contains both current and historical data.

DWH: Abbreviated *DW*, a collection of data designed to support management decision making. Data warehouses contain a wide variety of data that present a coherent picture of business conditions at a single point in time.
Development of a data warehouse includes development of systems to extract data from operating systems plus installation of a warehouse database system that provides managers flexible access to the data.
The term data warehousing generally refers to the combination of many different databases across an entire enterprise. Contrast with *data mart*.

DataMart: A database, or collection of databases, designed to help managers make strategic decisions about their business. Whereas a data warehouse combines databases across an entire enterprise, data marts are usually smaller and focus on a particular subject or department. Some data marts, called *dependent data marts,* are subsets of larger data warehouses.

Meta Data: Data about data. Metadata describes how and when and by whom a particular set of data was collected, and how the data is formatted. Metadata is essential for understanding information stored in data warehouses and has become increasingly important in XML-based Web applications.

BI: Most companies collect a large amount of data from their business operations. To keep track of that information, a business and would need to use a wide range of software programs , such as Excel, Access and different database applications for various departments throughout their organization. Using multiple software programs makes it difficult to retrieve information in a timely manner and to perform analysis of the data.
The term Business Intelligence (BI) represents the tools and systems that play a key role in the strategic planning process of the corporation. These systems allow a company to gather, store, access and analyze corporate data to aid in decision-making. Generally these systems will illustrate business intelligence in the areas of customer profiling, customer support, market research, market segmentation, product profitability, statistical analysis, and inventory and distribution analysis to name a few.

Slovník pojmů

OLAP: Short for ***O****nline* ***A****nalytical* ***P****rocessing,* a category of software tools that provides analysis of data stored in a database. OLAP tools enable users to analyze different dimensions of multidimensional data. For example, it provides time series and trend analysis views. OLAP often is used in data mining.
The chief component of OLAP is the OLAP server, which sits between a client and a database management systems (DBMS). The OLAP server understands how data is organized in the database and has special functions for analyzing the data. There are OLAP servers available for nearly all the major database systems.

ETL: Short for ***e****xtract,* ***t****ransform,* ***l****oad*, three database functions that are combined into one tool to pull data out of one database and place it into another database.

Extract -- the process of reading data from a database.

Transform -- the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.

Load -- the process of writing the data into the target database.
ETL is used to migrate data from one database to another, to form data marts and data warehouses and also to convert databases from one format or type to another.

OLTP: Short for ***O****n-****L****ine* ***T****ransaction* ***P****rocessing*. Same as *transaction processing*.

Transaction processing: A type of computer processing in which the computer responds immediately to user requests. Each request is considered to be a *transaction.* Automatic teller machines for banks are an example of transaction processing.
The opposite of transaction processing is batch processing, in which a batch of requests is stored and then executed all at one time. Transaction processing requires interaction with a user, whereas batch processing can take place without a user being present.

EAI: Acronym for ***e****nterprise* ***a****pplication* ***i****ntegration*. EAI is the unrestricted sharing of data and business processes throughout the networked applications or data sources in an organization. Early software programs in areas such as inventory control, human resources, sales automation and database management were designed to run independently, with no interaction between the systems. They were custom built in the technology of the day for a specific need being addressed and were often proprietary systems. As enterprises grow and recognize the need for their information and applications to have the ability to be transferred across and shared between systems, companies are investing in EAI in order to streamline processes and keep all the elements of the enterprise interconnected.

ERP: Short for ***e****nterprise* ***r****esource* ***p****lanning,* a business management system that integrates all facets of the business, including planning, manufacturing, sales, and marketing. As the ERP methodology has become more popular, software applications have emerged to help business managers implement ERP in business activities such as inventory control, order tracking, customer service, finance and human resources.

Datový sklad (Data Warehouse)

- Definice (W.H. Inmon 1996):

Datový sklad je

- subjektově orientovaný
- integrovaný
- časově proměnný
- stálý

soubor dat, který slouží pro podporu rozhodování.

Datový sklad

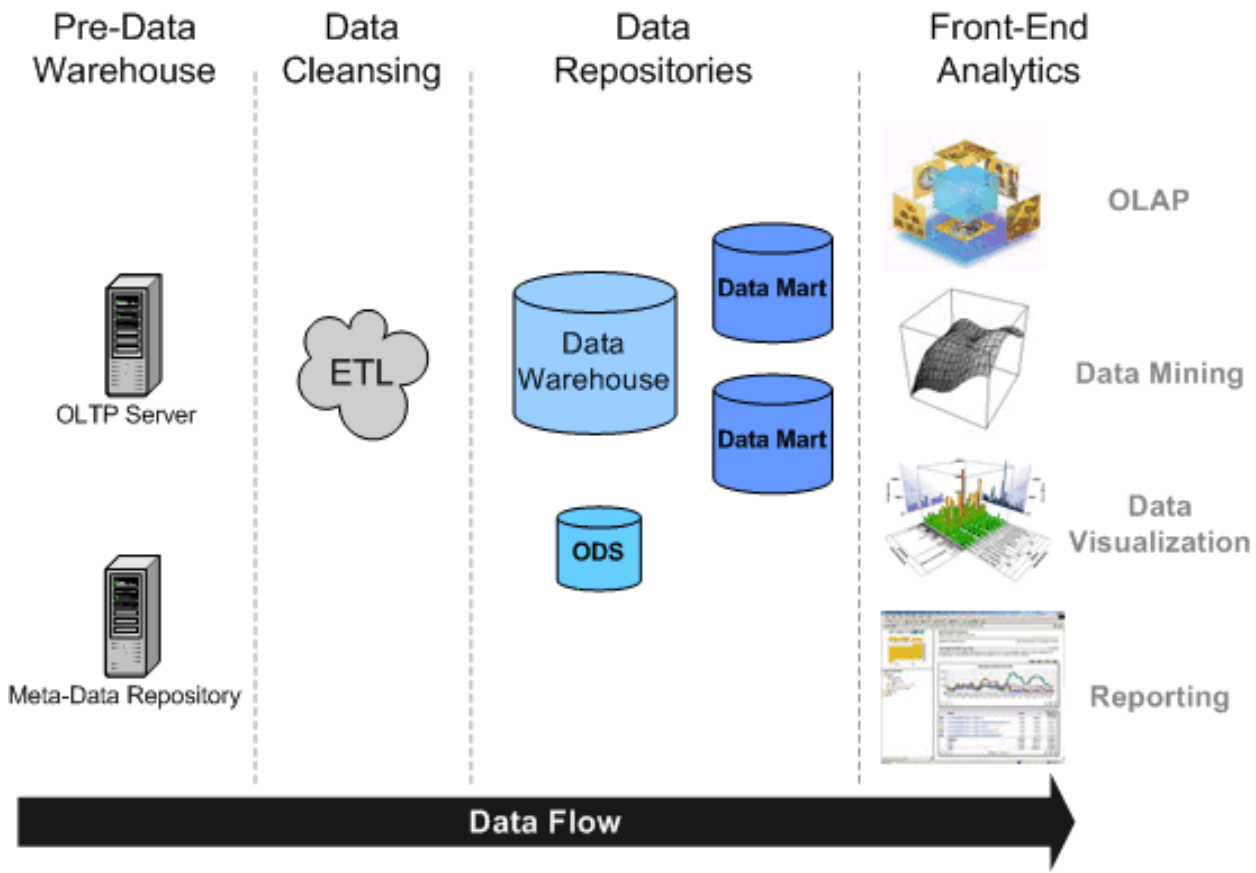
- prvotní koncepce datována počátkem 80.let
- vznik z potřeby jednoduchého přístupu ke strukturovanému úložišti kvalitních dat
- pomáhá získat odpovědi pro lepší rozhodování
- umožňuje použití dat pro dotazování, reportování a analýzu

Struktura datového skladu

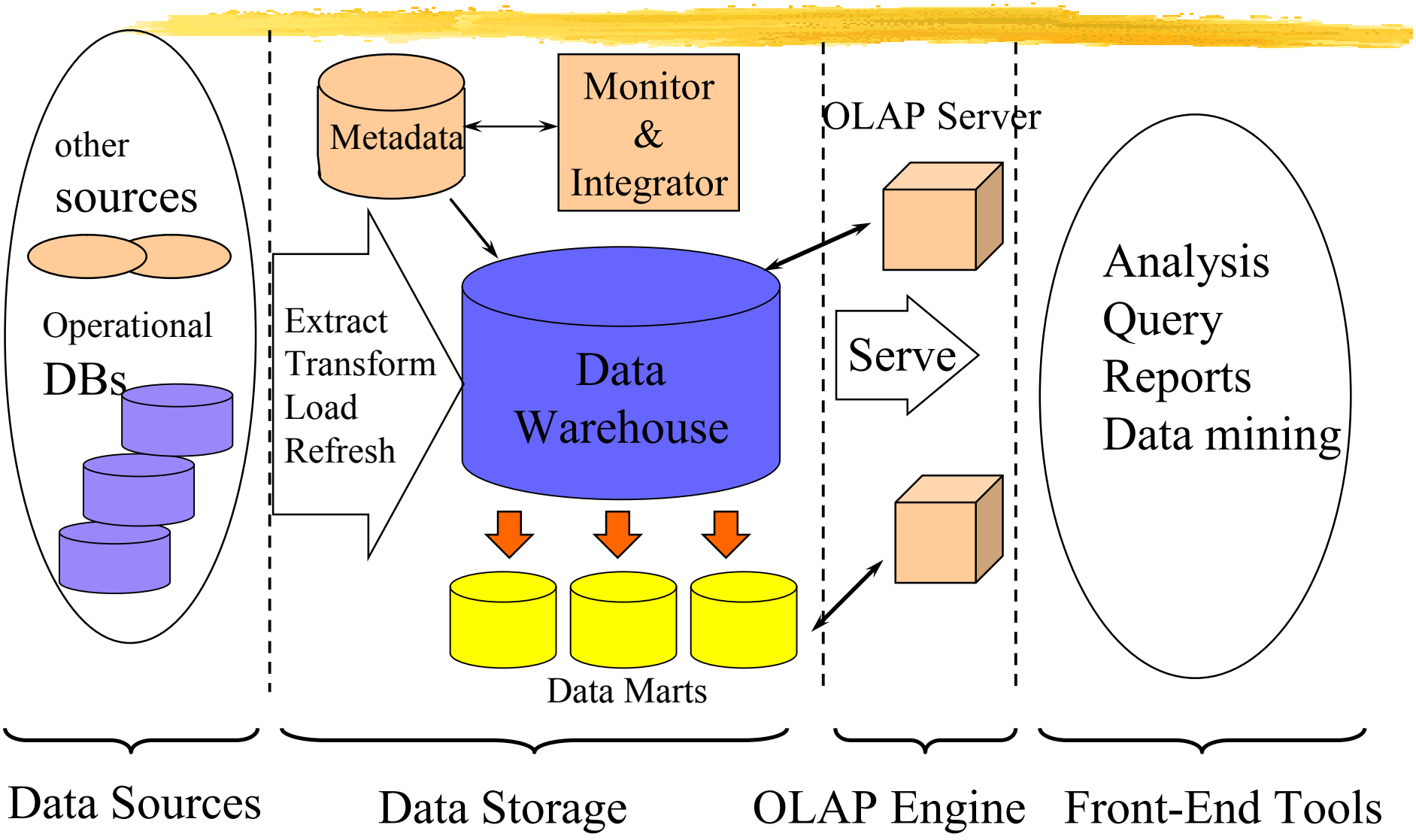
- třívrstvá architektura:
  - datový sklad
  - aplikační vrstva
  - prezentační vrstva

- fyzicky centralizovaný nebo distribuovaný

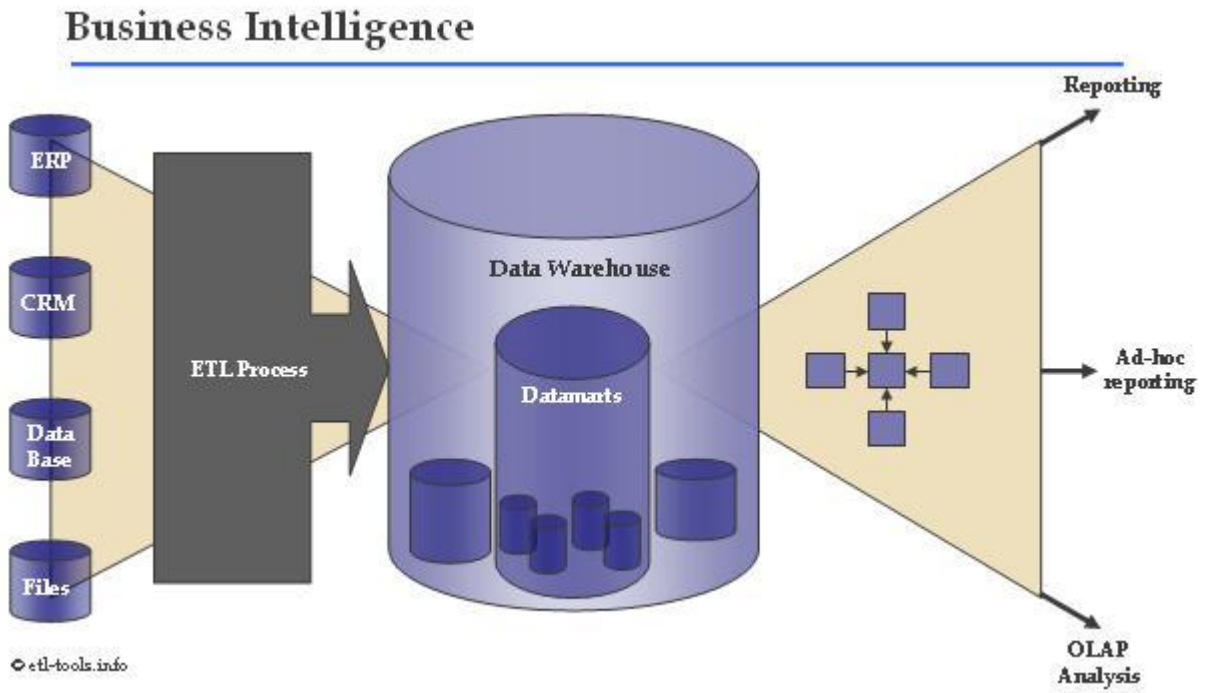
Data Warehouse



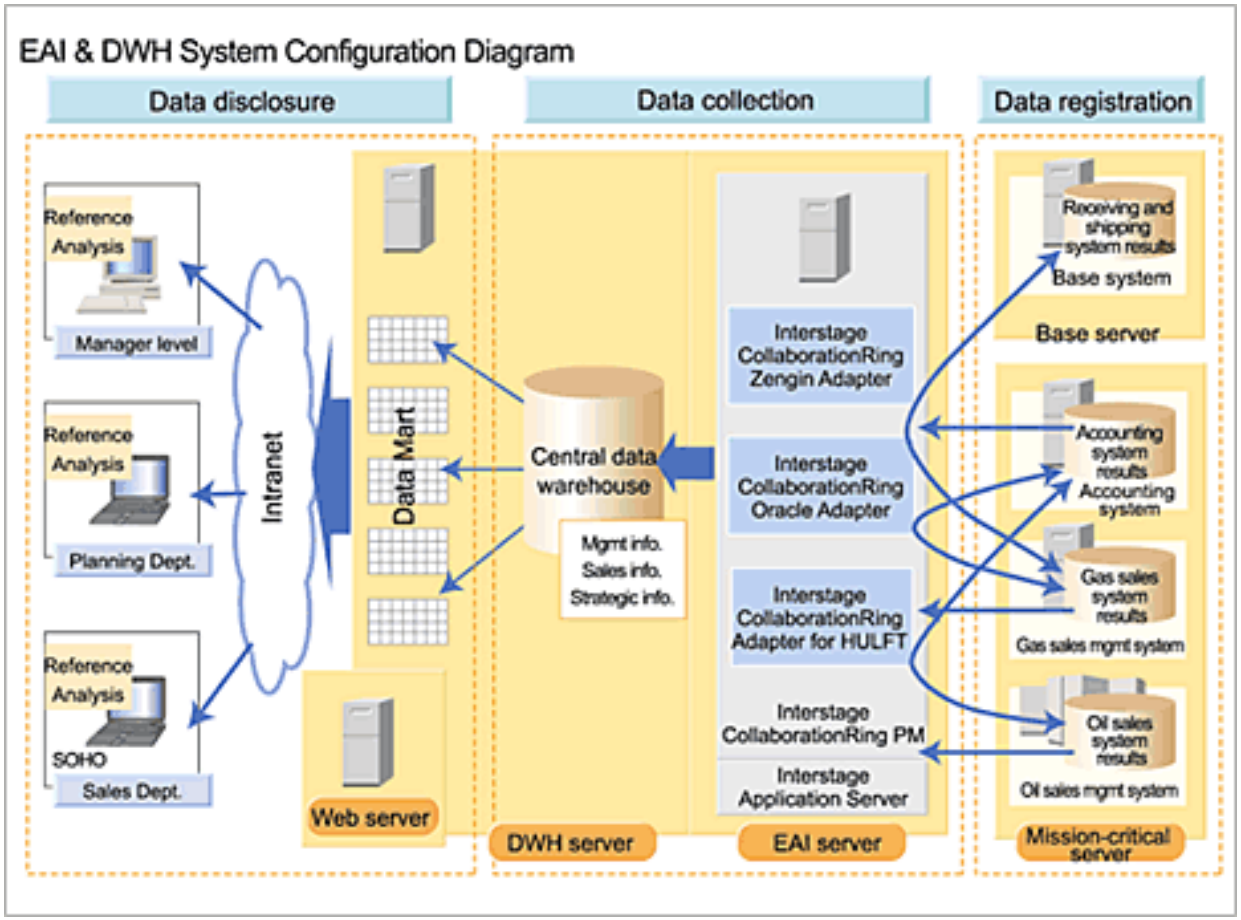
Data Warehouse



Data Warehause



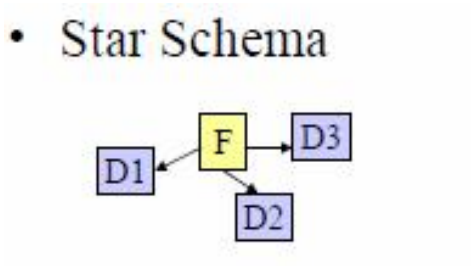
Data Warehouse



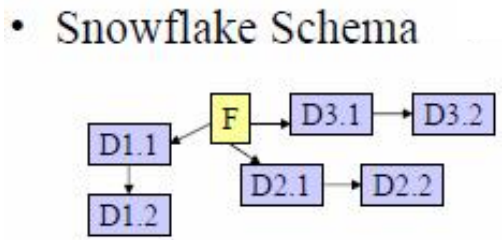
SOHO: Short for ***s**mall **o**ffice/**h**ome **o**ffice*, a term that refers to the small or home office environment and the business culture that surrounds it. A SOHO is also called a *virtual office*.

Data Models

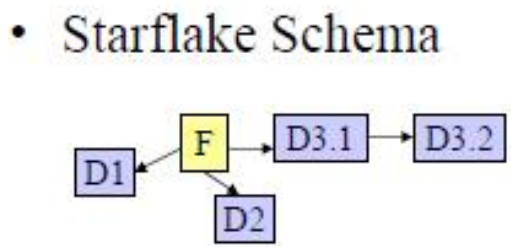
- Star (hvězda)



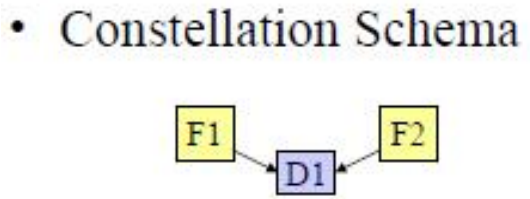
- Snowflake (vločka)



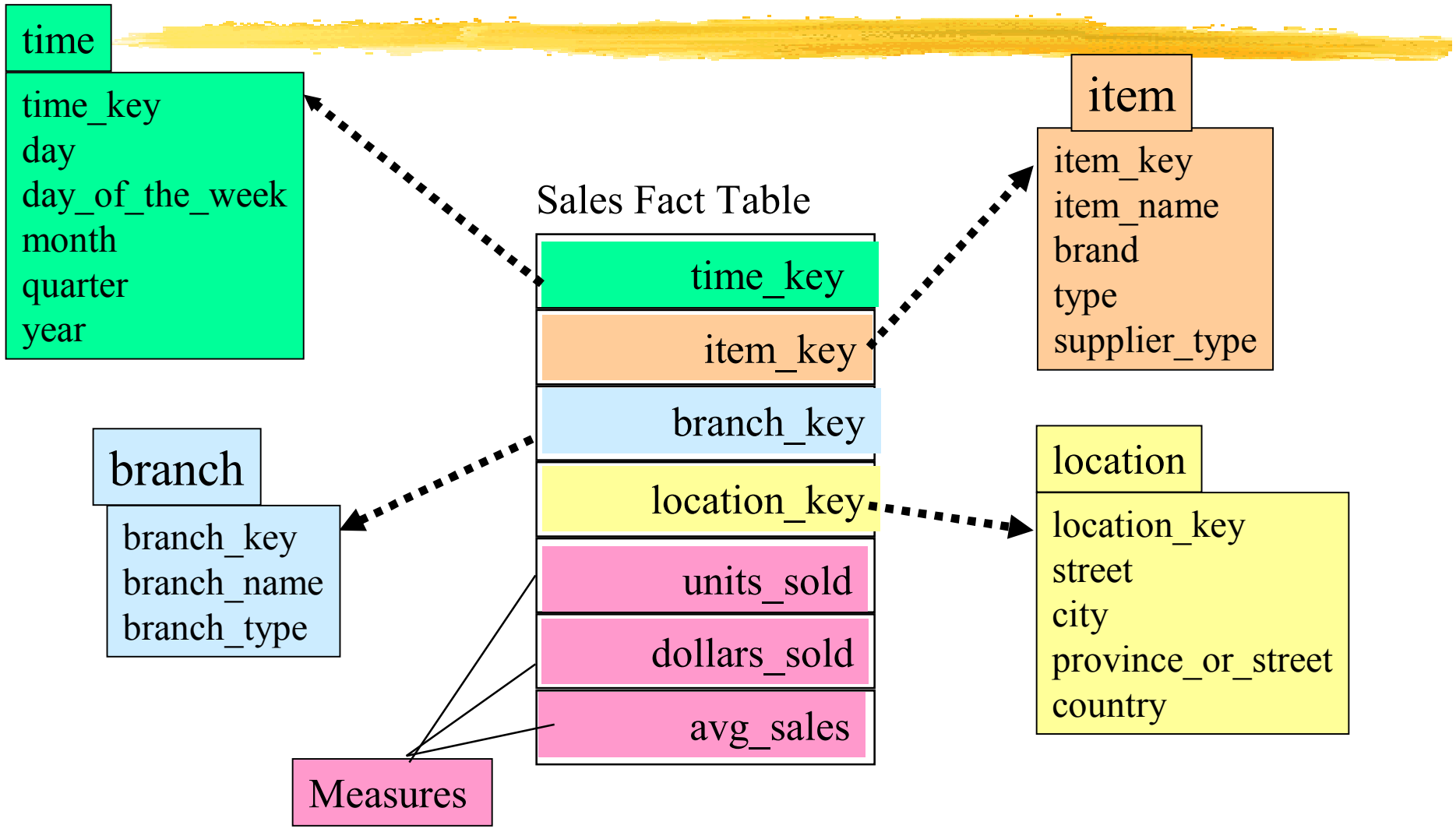
- Starflake



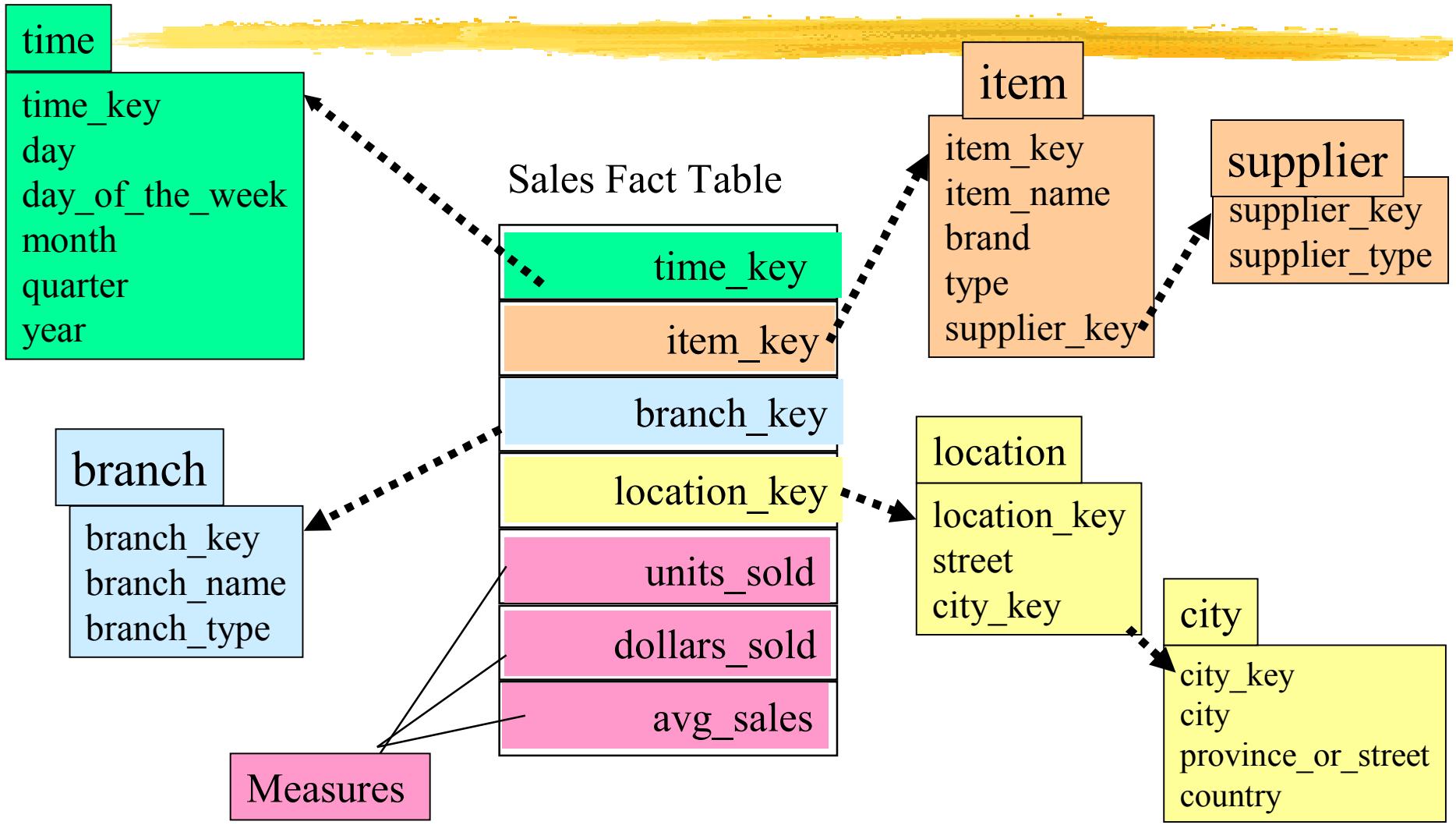
- Constellation (souhvězdí)



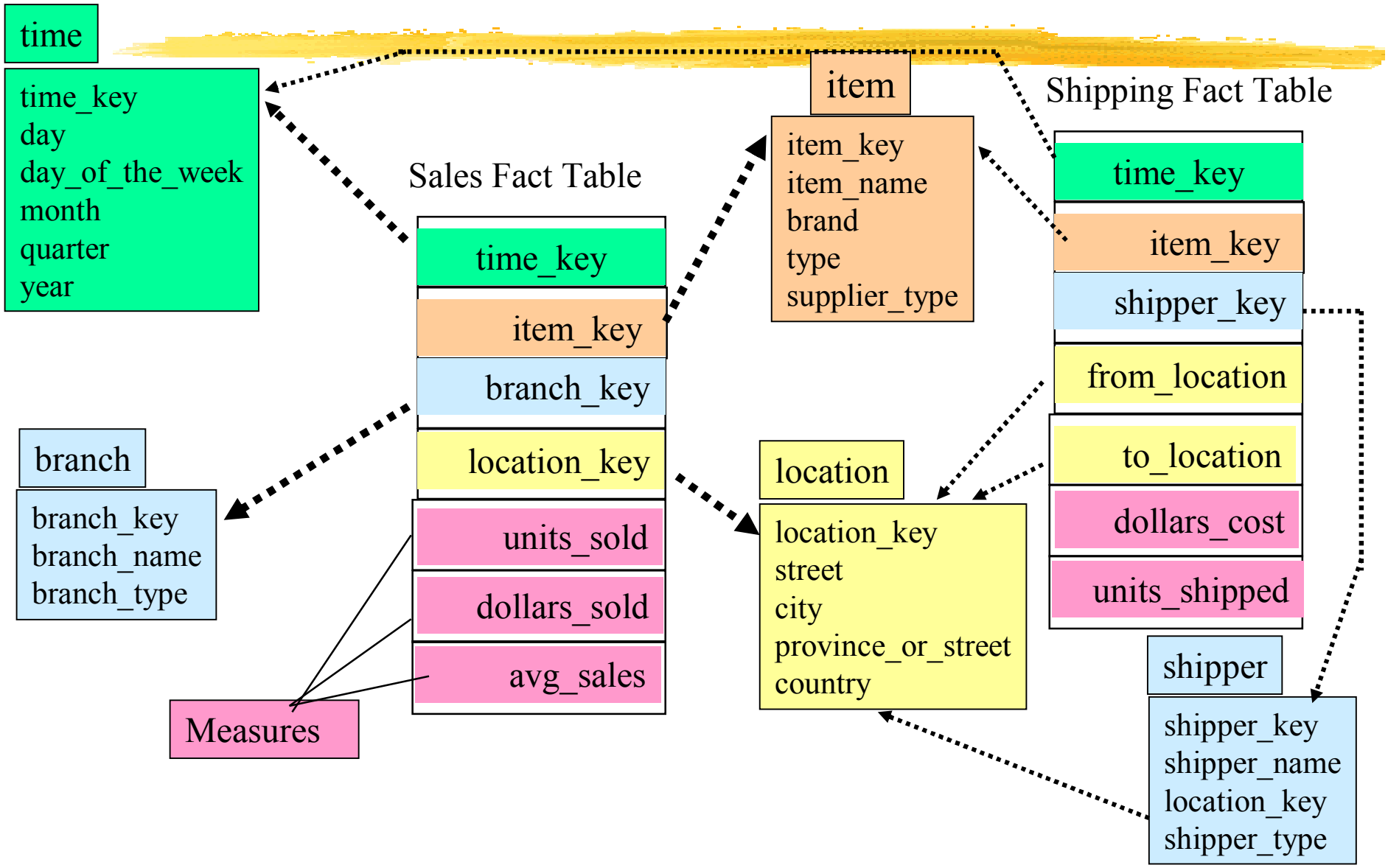
Example of Star Schema



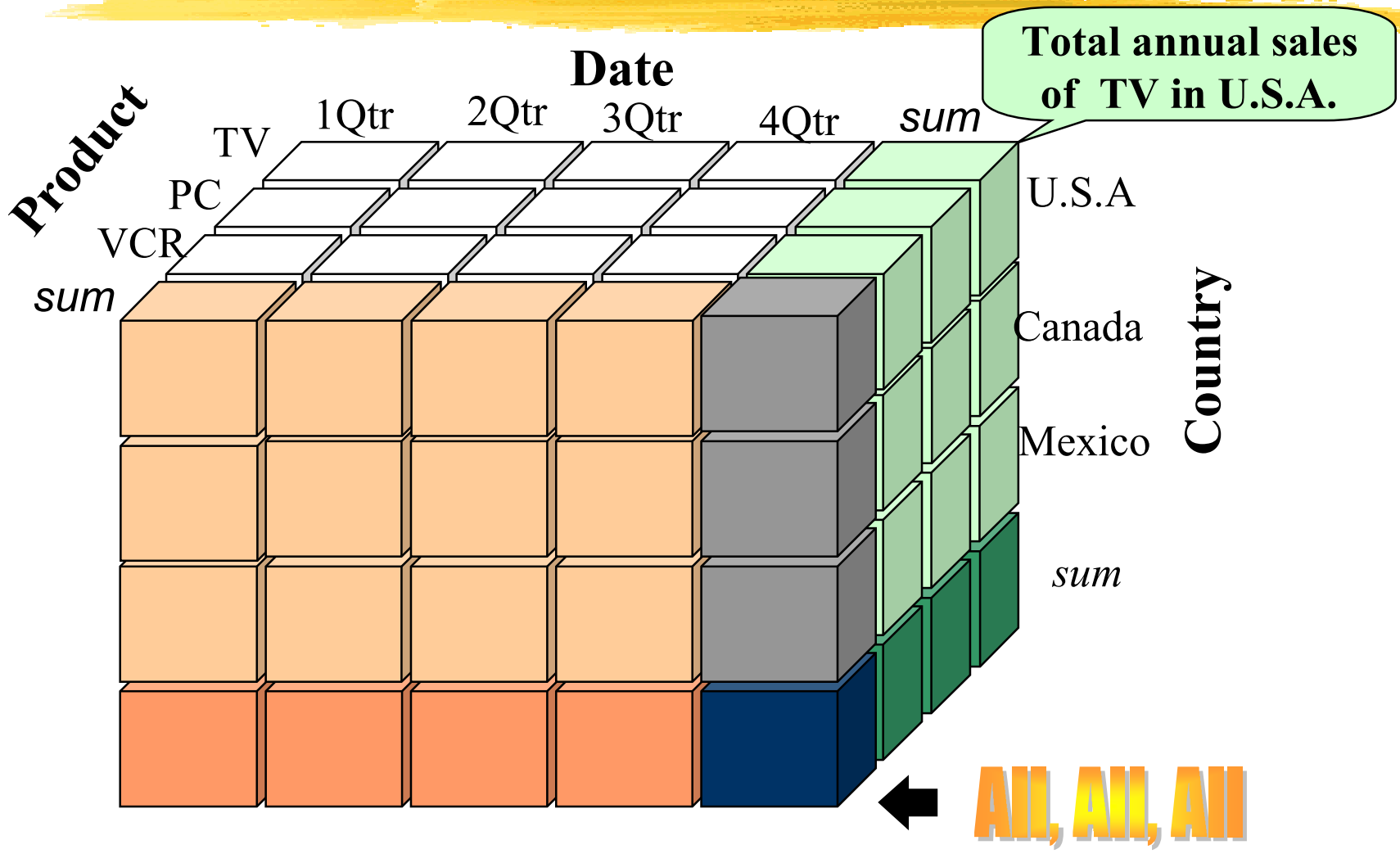
Example of Snowflake Schema



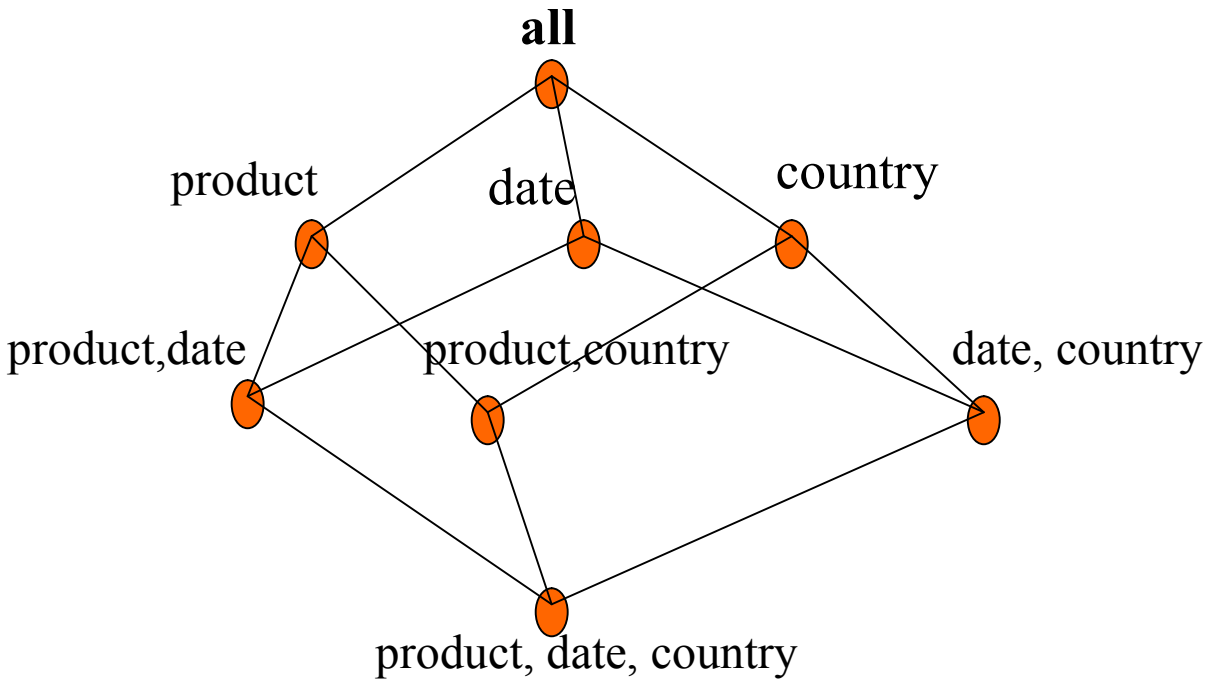
Example of Fact Constellation



A Sample Data Cube



Cuboids Corresponding to the Cube



0-D(apex) cuboid

1-D cuboids

2-D cuboids

3-D(base) cuboid

Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice (krájet a kostkovat):
  - *project and select*
- Other operations
  - *drill across: involving (across) more than one fact table*
  - *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*

OLAP Server Architectures

- Relational OLAP (ROLAP)
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware to support missing pieces
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
- Multidimensional OLAP (MOLAP)
  - Array-based multidimensional storage engine (sparse matrix techniques)
  - fast indexing to pre-computed summarized data
- Hybrid OLAP (HOLAP)
  - User flexibility, e.g., low level: relational, high-level: array
- Specialized SQL servers
  - specialized support for SQL queries over star/snowflake schemas

ROLAP

- Data uložená v relační databázi – nejsou duplikována, ovšem není k nim možný přístup bez připojení k zdrojové databázi.
- dotazy OLAP se převádějí do klasických dotazů SQL – může být nevýhodou (limitované možnosti SQL, pomalejší odezva).
- Vhodný jen pro omezené množství dat.

MOLAP

- „tradiční“ OLAP.
- Data uložena v multidimenzionálních kostkách mimo relační databázi. Jsou tudíž duplikována a je možný přístup i bez spojení s původním zdrojem dat.
- Hlavní výhodou je rychlá odezva na dotazy. Vše je předpočítáno a uloženo při tvorbě kostek.

HOLAP

- ponechává původní data v relačních tabulkách, agregace ukládá v multidimenzionálním formátu
- poskytuje propojení mezi rozsáhlými objemy dat v relačních tabulkách
- výhoda rychlejšího výkonu multidimenzionálně uložených agregací

Budování datového skladu

- metoda „velkého třesku“:
  - analýza požadavků podniku
  - vytvoření podnikového datového skladu
  - vytvoření datových tržišť

- přírůstková (evoluční) metoda

Plnění datového skladu

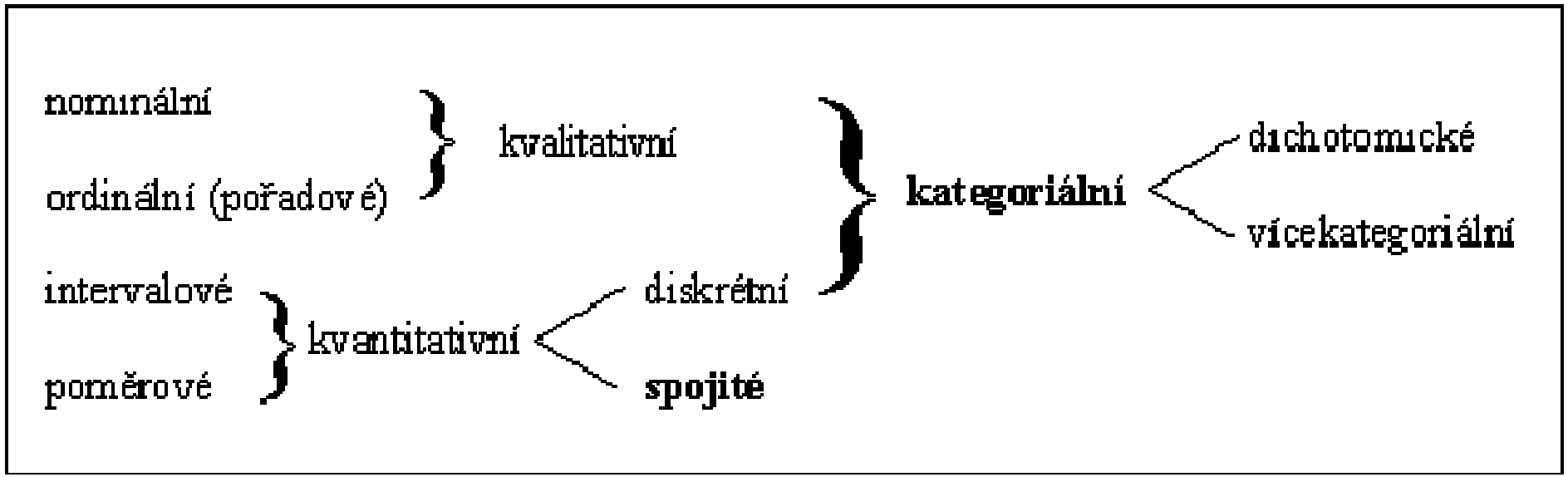
- počáteční plnění + pravidelná aktualizace

- plnění pomocí datových pump
- postupy ETL:
  - extrakce
  - transformace
  - loading

Klasifikace dat

- Kvalitativní (kategoriální)
  - nominální — rodinný stav, region
    - Alternativní — pohlaví
  - ordinální — vzdělání
- Kvantitativní (numerické)
  - diskrétní — počet dětí
    - binární (dichotomické) — indikátor dobrého klienta
  - spojité — věk, příjem, výše úvěru

Klasifikace dat



Klasifikace dat II

- Demografické znaky:
  - **Klienta (věk, pohlaví, rodinný stav, počet dětí, druh bydlení, kraj/okres trvalého bydliště…)**
  - **Prodejního místa (kraj/okres, typ, prodejní plocha,…)**
  - **Prodejce (věk, pohlaví, kraj/okres trvalého bydliště…)**
- Behaviourální znaky:
  - **Klienta („stáří“ klienta, doposud splacená jistina, dlužná jistina, počet dní po splatnosti,…)**
  - **Prodejního místa („stáří“ prodejny, počet uzavřených smluv, objem uzavřených smluv, podíl nesplácených úvěrů,…)**
  - **Prodejce (počet uzavřených smluv, objem uzavřených smluv, podíl nesplácených úvěrů …)**
- Produktové znaky:
  - **Výše úvěru, délka smlouvy, akontace, RPSN,…**

Datová matice

- Nutný formát dat pro modelování.
- 2-rozměrná matice *n* x *p*.
- Řádky reprezentují *n* statistických jednotek (klientů)
- Sloupce reprezentují *p* statistických proměnných.

f:\apps\spsswin\bank.sav

1:id 628

	id	salbeg	sex	time	age	salnow	edlevel	work	jobcat	minority	sexr
1	628	8400	0	81	28.50	16080	16	.25	4	0	
2	630	24000	0	73	40.33	41400	16	12.50	5	0	
3	632	10200	0	83	31.08	21960	15	4.08	5	0	
4	633	8700	0	93	31.17	19200	16	1.83	4	0	
5	635	17400	0	83	41.92	28350	19	13.00	5	0	
6	637	12996	0	80	29.50	27250	18	2.42	4	0	
7	641	6900	0	79	28.00	16080	15	3.17	1	0	
8	649	5400	0	67	28.75	14100	15	.50	1	0	

Zásady tvorby datové matice

- Replikovatelnost tvorby dat. matice
  - žádné manuální úpravy dat
- Srozumitelnost tvorby dat. matice
  - podrobné komentáře
- Zpětná konektivita dat. matice
  - primární klíče (id) všech podkladových datových tabulek