data warehouses and OLAP

basic concepts

data warehouse = a large store of data accumulated from a wide range of sources within a company and used to guide management decisions

↳ subject-oriented, integrated (from multi sources), time variant (data info degraded), nonvolatile (persist without power), separate from company database

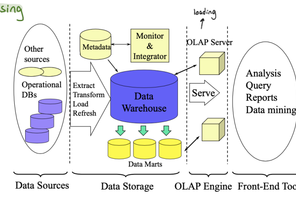
data warehousing = process of constructing and using data warehouses

- operational update of data does not occur → only initial loading of data + access of data
- does not require transaction processing, recovery, and concurrency control mechanisms

online transaction processing ↘ **online analytical processing** ↙

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

↳ DBMS ↳ warehouse



data warehouse models:

↳ **enterprise warehouse** = collects all the information about subjects spanning the entire organization

↳ **data mart** (market) = simple form, focused on a single subject • dependent → subset of enterprise / independent

↳ **virtual warehouse** = set of separate databases, which can be queried together

metadata = defines the warehouse objects, and structure of the data warehouse

↳ **operational meta data** = includes information about how and when data was created / transformed

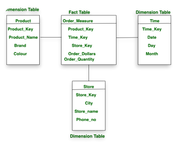
modeling: data cube and OLAP

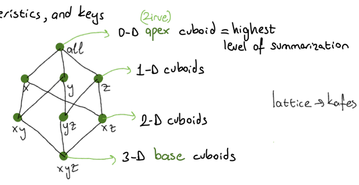
data cube = multidimensional structure used to store data

↳ represents the data in terms of dimensions and facts

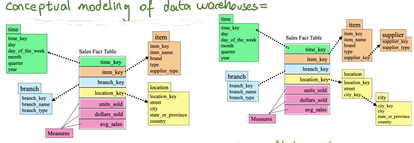
fact table = stores the measurements, metrics, or facts related to a operation

dimension table = contains descriptions of the objects in a fact table, and provide information about dimensions such as values, characteristics, and keys





conceptual modeling of data warehouses =



- **star schema**: fact table in the middle, connected to dimension tables
- **snowflake schema**: some dimensional hierarchy is normalized into a smaller dimension tables like snowflake
- **fact constellations**: (fakten yildiz) multiple fact tables share dimension tables, galaxy schema

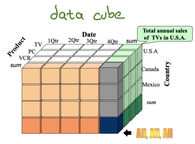
data cube measures =

↳ **distributive** = if you can apply a func to each partition and then, apply them to resulting values, and get the same answer → sum(), max(), min(), count()

↳ **algebraic** = (→ there is a bound of N integers) avg(), min_N(), standard_deviation() each obtained with distribute agg func

↳ **holistic** = if no constant bound on the storage size needed to describe a subaggregate median(), mode(), rank()

data cube



cuboids

olap operations =

↳ **roll up (drill up)** = summarize data by climbing up hierarchy or dimension reduction

↳ **roll down (drill down)** = from higher to lower level summary, detailed data, new dimension introducing

↳ **slice and dice** = (2D select / 3D select) project and select

↳ **pivot (rotate)** = (reorient) reorient the cube, visualization, 3D to 2D

↳ **drill across** = involving more than one table

↳ **drill through** = through the bottom level of the cube to its backend relation tables (using sql)

design and usage

- **top-down view** = selection of relevant information necessary
- **data source view** = reveals captured, stored and managed information by the operating system
- **data warehouse view** = fact and dimension tables
- **business query view** = sees the perspectives of data from the view of end-user

skipped

usage = information processing, analytical processing, data mining

OLAM (online analytical mining) = OLAP (online analytical processing) + data mining + multidim databases

implementation

materialization of data cube = computation of cuboids in a data cube lattice

↳ **full materialization** (every cuboid), **no materialization** (none), **partial materialization** (some)

DMQL = data mining query language

cube operator →
define cube sales [item, city, year]: sum (sales_in_dollars)
compute cube sales

bitmap index =

Base table

Cust	Region	Type
C1	Asia	Retail
C2	Europe	Dealer
C3	Asia	Dealer
C4	America	Retail
C5	Europe	Dealer

Index on Region

RecID	Asia	Europe	America
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	1	0

Index on Type

RecID	Retail	Dealer
1	1	0
2	0	1
3	0	1
4	1	0
5	0	1

```
SELECT item, city, year, SUM (amount)
FROM SALES
CUBE BY item, city, year
```

Server architectures =

↳ **relational olap (ROLAP)** = use relational DBMS to store and manage warehouse data

↳ **multidimensional olap (MOLAP)** = sparse array-based multidimensional storage engine

↳ **hybrid olap (HOLAP)** = low level: relational, high level: array → microsoft sqlserver

↳ **specialized sql servers** = over star / snowflake schemas

data generation by attribute oriented induction

- collect data → generalization with att removal or generalization → aggregation by merging → present to user

example = describing general characteristics of graduate students in the university database

Initial Relation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...						...	
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,...

```
Select * (i.e., name, gender, major, birth_place,
    birth_date, residence, phone#, gpa)
from student
where  student_status in {"MSc", "MBA", "PhD" }
```

Prime Generalized Relation

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
...	...	...	...	...	...	...

Gender \ Birth_Region	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

① **InitialRel** = query of task-relavant data

② **PreGen** = determine generalization plan for each attribute

③ **PrimeGen** = perform the generalization

④ **Presentation** = user interaction: adjust level by drilling, pivoting, mapping into rules, cross tabs, visualisation