UNIT-2

*

→ Measures for data quality:

accuracy, completeness, consistency, timeliness, interpretability.

→ Data pre-processing:

1. Data cleaning
2. Data integration
3. Data reduction → Dimensionality, Numerosity, Data compression.
4. Data transformation and data discretization.
   - → Normalization
   - → Concept hierarchy generation.

① DATA CLEANING:

* Incomplete / missing data (= "?") → handled using decision tree.
* Noise / error (-10)
* Inconsistent / discrepancy (Age = 42 DOB = "03/07/10")
* Intentional. (Jan 1 as everyone's birthday).

② → Binning, Regression, clustering

Process:

1. Data discrepancy detection → Scrubbing, auditing
2. Data migration & integration.

② Data Integration:

Redundancy detected after data integration is handled using correlation & covariance analysis.

$$\chi^2 = \sum \frac{(\text{observe} - \text{expect})^2}{\text{expected}} \Rightarrow \text{Nominal data.}$$

(Correlation coefficient)

Ex:

	Play chess	Not Playchess	Sum(row)
Likescience	250(90)	200(360)	450
Not like	50(210)	1000(840)	1050
Sum(col)	300	1200	1500

Expected = $\frac{\text{Row} \times \text{Col tot}}{\text{grand}}$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$\chi^2_{calc} = 507.93$. ~~dof = (columns-1) × (rows-1) = 1×1 = 1~~

Correlation coefficient (Pearson's product moment coefficient)
(Numeric data)

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{(n)\sigma_A \sigma_B}$$

$r_{A,B} > 0 \rightarrow$ +ve correlation.

$r_{A,B} = 0 \rightarrow$ independent.

$r_{A,B} < 0 \rightarrow$ −ve correlation.

* Co-variance (Numeric Data).

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B}))$$

$$= \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}$$

correlation coeff $\Rightarrow r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$

Ex: A → (6, 5, 4, 3, 2)
B → (20, 10, 14, 5, 5)

$\bar{A} = \frac{6+5+4+3+2}{5} = 4$

$\bar{B} = \frac{20+10+14+5+5}{5} = 10.8$

$Cov(A,B) = \frac{(6-4)(20-10.8) + \cdots + (2-4)(5-10.8)}{5}$

$= 7.$

* Data Reduction Strategies:

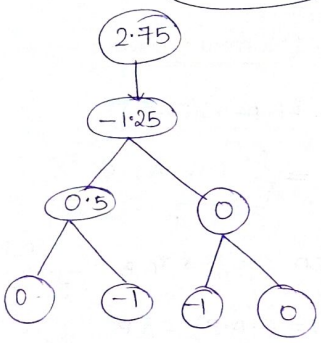
① Dimensionality: Wavelet transforms, PCA, Feature subset selection.

② Numerosity: Regression, Histograms, Data cube aggregation.

③ Data compression.

→ Wavelet Transform: decomposes a signal into diff. frequency signals.

* Wavelet decomposition: Slide 28



* Haar Wavelet Coefficients:

averages $\frac{a+b}{2}$

$S = [2,2,0,2,3,5,4,4] = 8 = 2^3$

Resolution	Averages	Detail coefficient	→ $\frac{a-b}{2}$
8	[2,2,0,2,3,5,4,4]		
4	[2,1,4,4]	[0,-1,-1,0]	$\frac{5.5}{2}$
2	[1.5,4]	[1/2, 0]	$= 2.75$.
1	2 3/4	-1 1/4	~~2 3/4~~ [2 3/4]

⇒ [2 3/4, -1 1/4, 1/2, 0, 0, -1, -1, 0]

* Principal Component Analysis (PCA) ⇒ Numeric data.

Steps:

① Normalize input data.
② Compute k orthonormal vectors.
③ Each input data (vector) is a linear combination of the k principal component vectors.
④ Principal components are sorted in order of decreasing strength.
⑤ Size of data can be reduced by eliminating weak components (low variance)

→ Remove redunt & irrelevant attributes to reduce dimensionality of data.

→ Numerosity Reduction:

* Parametric methods: Regression; Log-Linear Models.

* Non-Parametric methods: Histograms, Clustering.

Parametric

→ Linear & Multiple Regression
→ Log linear model.

* Regression analysis

values of dependent (or) response variable.
and one (or) more independent (or) explanatory variables.

Linear: $y = wX + b$.

Multiple: $y = b_0 + b_1X_1 + b_2X_2$.

* Histogram Analysis:

→ Divide data into buckets & store avg sum.
  - → Equal width
  - → Equal frequency.

* Clustering based on centroid or diameter

* Sampling: obtaining small samples to represent the whole data set N.

Types: → Simple random sampling
→ Sampling without replacement
→ Sampling with replacement
→ Stratified sampling / Cluster.

* Data Cube Aggregation:

→ lowest level of data cube
→ Reduce the size of data further
→ use appropriate representation

→ Data compression:

① String compression
② Video/Audio Compression
③ Dimensionality & numerosity are a kind of data compression.

Original ⟶ Compressed
(lossless ⟵)
Compressed → Approximated (lossy)

* Data Transformation:

→ old values are mapped with new values for a given attribute.

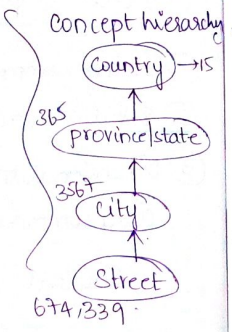
→ Methods:
  - → Smoothing
  - → Attribute/Feature construction.
  - → Aggregation
  - → Normalization, → discretization.

* Discretization:

- divide range of continuous attribute into intervals.
- → supervised v/s unsupervised.
- → split (top-down) v/s merge (bottom-up.

Methods:

- Binning
- Histogram analysis
- Clustering
- Decision tree analysis.
- Correlation.



* Binning

Ex: data: (sorted)

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34.

→ Equi depth:

Bin-1: 4, 8, 9, 15

Bin-2; 21, 21, 24, 25

Bin-3: 26, 28, 29, 34

→ By means:

Bin-1: 9, 9, 9, 9

Bin-2: 23, 23, 23, 23.

$\frac{4+8+9+15}{4}$

21
15
36

Bin-3: 29, 29, 29, 29.

→ By boundaries:

~~1 lowest value~~

Bin-1: 4, 4, 4, 15 · nearest value either min / max

Bin-2: 21, 21, 25, 25

Bin-3: 26, 26, 26, 34

8-4 = 4 ; 15-8 = 7

* UNIT-1

Proximity analysis for ordinal variables.

Name	Rank	r_{if}	z_{if}
Jack	Excellent	4	(4-1)/3 = 1
Mary	Better	3	0.67
Jim	Good	2	0.33
Pat	Average	1	0

$z_{if} = \frac{r_{if} - 1}{(M_f - 1)}$

$M_f = 4$

$z_{if} = \frac{r_{if} - 1}{3}$

dissimilarity

	J	M	Jim	P
J	0			
M	1-0.67	0		
Jim	1-0.33	0.67-0.33	0	
P	1-0	0.67-0	0.33-0	0

|1-0.67|

Manhattan distance

* Data warehouse:

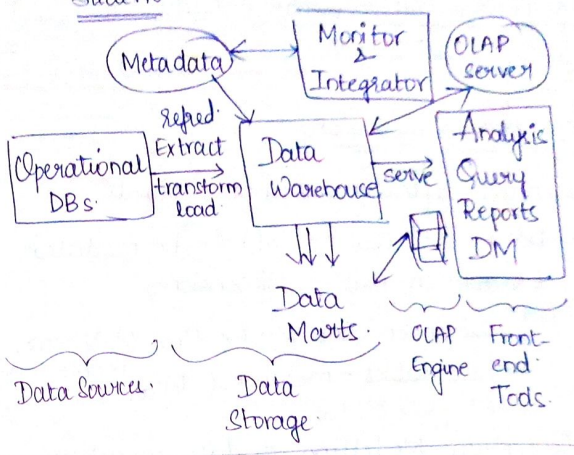
↳ subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

→ organized based on customer, sales, product
→ constructed using multiple heterogeneous data sources.
→ gives data over a no. of years.
→ operational updates are not seen data warehouse.
→ Requires only 2 operations in data accessing:
  ↳ initial loading of data
  ↳ access of data.

OLTP	OLAP
→ day to day operations	→ decision support: long term.
→ repetitive	→ ad-hoc
→ read/write, index/hash	→ lots of scans
→ DB size: 100MB-GB	→ 100GB-TB
→ transaction throughput	→ query throughput; response
→ current, detailed data	→ summarized historical data

Data warehouse = multi-tiered architecture

Slide: 10



* Data Warehouse Models:

→ Enterprise warehouse: collects all the info about entire organization.
→ Data Mart: a subset of corporate wide data specific group related.
→ Virtual warehouse: set of views over operational database.

* Extraction, Transformation, Loading: (ETL)

① Data Extraction
② Data cleaning

③ Data transformation - convert data from legacy to host

④ Load - sort, summarize, consolidate.

⑤ Refresh.

* Metadata Repository:

→ data defining warehouse objects.

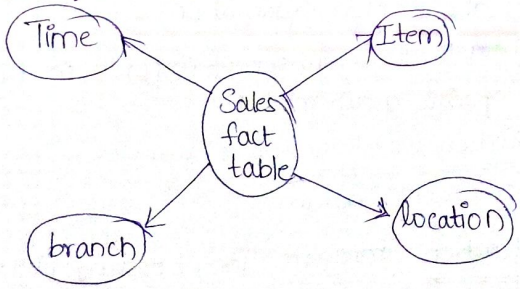
*→ Data cubes allows data to be modelled & viewed in multiple dimensions.

↳ dimension tables = Ex: time (day, week, month year)

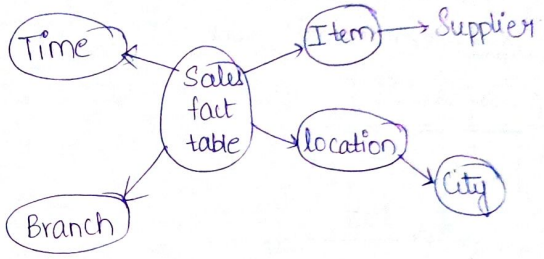
↳ Fact tables = measures (dollars_sold).

*→ Conceptual Modelling of data warehouses.

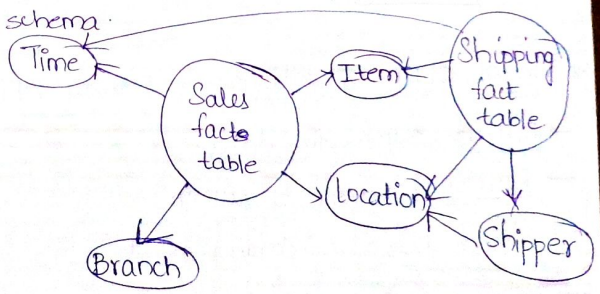
① Star Schema: fact table connected to set of dimension tables.



② Snowflake Schema: dimensional hierarchy is normalized into a set of smaller dimension tables.



③ Fact Constellations: fact tables share dimension tables also called galaxy schema.



Hierarchy →

all → region → country → city → office.

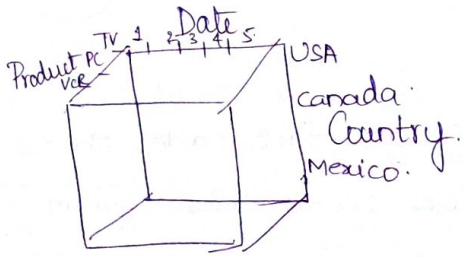
* Data Cube Measures:

→ Distributive: Eg: count, min, max.

→ Algebraic: avg, min, standard_deviation.

→ Holistic: Median, mode, rank.

* Sample data cube



* Typical OLAP operations:

→ Roll-up: summarize data.

→ Roll down: reverse of roll up.

→ Slice & dice: project and select.

→ Pivot (rotate)

→ drill across: involving more than one fact table.

→ drill through: bottom level of cube to relational tables.

* Design of data warehouse:

→ 4 views:

① Top-down view: allows selection of relevant information necessary for data warehouse.

② data source view: exposes into being captured, stored, and managed by OS.

③ data warehouse view: consists of facts tables and dimension tables.

④ Business query view: perspectives of data in warehouse from the view of end-user.

→ Design Process:

① Top-down, bottom-up approaches (or) a combination of both:

→ Top-down: starts with overall design & planning.

→ Bottom-up: starts with experiments & prototypes.

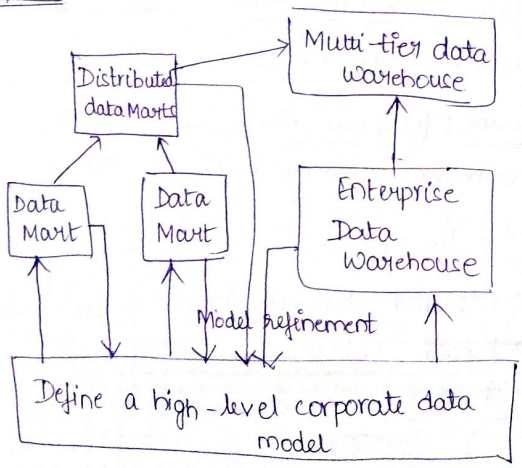
② SE pov:

→ Waterfall, Spiral.

→ Process:

1. Choose a business process.
2. Choose the grain
3. Choose the dimensions
4. Choose the measure that will generat each fact table record.

Approach:



Applications of data warehouse:

→ Information processing
→ Analytical processing
→ data mining

→ from OLAP to OLAM

* high quality of data in data warehouses.
* available info processing structure surrounding data warehouses.
* OLAP based exploratory data analysis
* On-line selection of DM fns.

* Data warehouse implementation:

→ Efficient cube computation

*→ Data cube can be viewed as a lattice of cuboids.

No. of cuboids in an n-dimensional cube with L levels: $T = \prod_{i=1}^{n} (L_i + 1)$

* Compute Cube Operator:

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

compute cube sales.

→ Select item, city, year, SUM(amount) from SALES cube by item, city, year.

* Efficient processing OLAP queries:

→ Determine which operations shd be performed on the available ~~cleed~~ cuboids.

→ Determined which materialized cuboids shd be selected for OLAP OP.

→ Exploring index structures & compressed v/s dense array structs in MOLAP.

* OLAP server architectures:

ROLAP (Relational)	MOLAP (Multidimensional)	HOLAP (Hybrid)
→ Relational DBMS. Optimization of DBMS backend.	→ Sparse array based multidimensional storage engine	Low-level: Relational High-level: array.
→ Greater Scalability	→ Fast indexing	→ Flexibility Ex: MS SQL server.
→ Slow query processing	→ Fast	→ medium.
→ Very slow.	→ fast processing	→ fast
→ large storage space.	→ medium storage space	→ small storage space.
→ Low latency	→ High latency	→ Medium latency.

* Attribute oriented Induction: (AOI)

→ perform generalization by attribute removal or attribute generalization.

Ex: ① Select * from student where student_status in {"MSc", "MBA", "PhD"}.

② perform attribute-oriented induction

③ present results in generalized relation, cross-tab, rule forms.

* Principles:

① Data focusing: task-relevant data.
② Attribute removal:
③ Attribute generalization
④ Attribute threshold control.
⑤ Generalized relation threshold control.

* Algorithm:

→ InitialRel: derive initial relation.
→ PreGen: removal (or) generalization.
→ PrimeGen: perform particular action from PreGen.
→ Presentation user interaction

* Presentation of Generalized results:

1. Generalized relation
2. Cross tabulation
3. Quantitative Characteristic rules.

Concept Description (AOI)	Cube-Based OLAP
→ automated desired level allocation	→ systematic preprocessing
→ data is not in relational forms.	→ works on query independent.