07/04/21
Tuesday.

DATA MINING AND DATAWAREHOUSING
[CS 402]

Christy Varghese
S8 CSE A
Roll No. 34.

1 (a) Data warehouse provides architectures and tools for business executives to systematically organize, understand & use their data to make strategic decisions. Data warehouse is a subject-oriented, time variant, non-volatile & integrated collection of data in support of management's decision making process.

* Subject-oriented : A data warehouse is organized around major subjects. such as : customer.

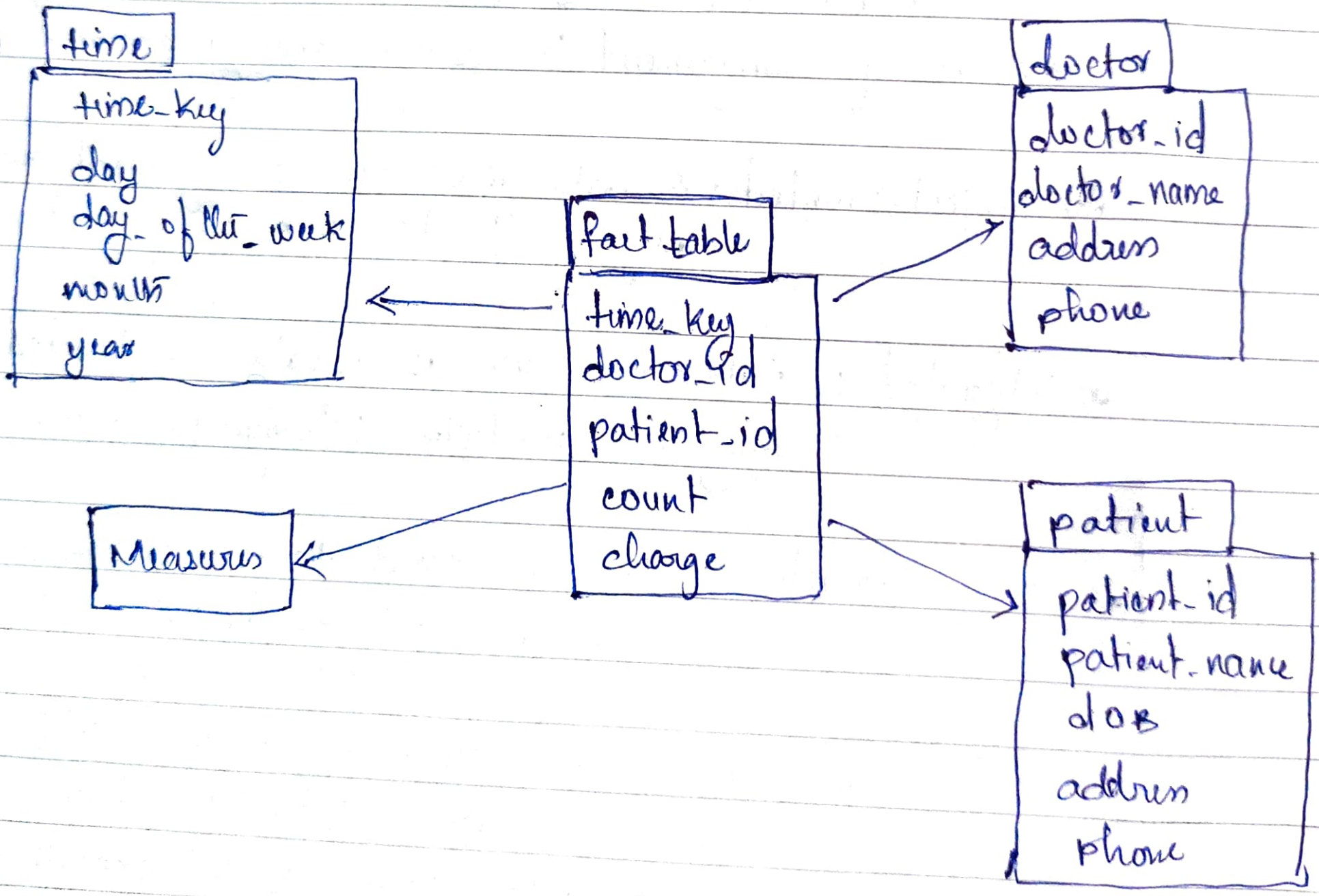
* Integrated : Data warehouse is usually constructed by integrating multiple heterogeneous data.

* Time-variant : Data are stored to provide information from a historical perspective.

* Non-volatile : A datawarehouse is always a physically separate store of data transformed from the application data found in the operational env.

(1)

(b)


time
time_key
day
day_of_the_week
month
year
fact table
time_key
doctor_id
patient_id
count
charge
doctor
doctor_id
doctor_name
address
phone
Measures
patient
patient_id
patient_name
dob
address
phone

(2)

OLTP	OLAP
✱ Operational data, OLTPs are the original source of data	✱ Consolidation data, OLAP data comes from various OLTP databases.
✱ Short & fast inserts & updates by end users.	✱ Periodic long running batch jobs refresh the data.
✱ A snapshot of on-going business processes.	✱ Multidimensional views of various kinds of business activities

⊛ To control and run fundamental business tasks	⊛ To help with planning, problem solving decision support.
⊛ Relatively standarized and simple queries returning relatively few records.	⊛ often complex queries involving aggregations.

Time

(3) (i) Stepwise forward selection : The procedure starts with an empty set of attributes as the reduced set. The best of the original attributes is determined and added to the reduced set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set. elimination.

(ii) Stepwise backward ~~selection~~ : The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set.

(3) (iii) A combination of forward selection & backward elimination:

The stepwise forward selection & backward elimination methods can be combined so that, at each step, the procedure selects the best attribute & removes the worst from among the remaining attributes.

(4) (b) (*) The entropy of S is defined as:

$$\text{Entropy}(S) = \sum_{i=1}^{C} -p_i \log_2(p_i)$$

where S is a segment of dataset having a number of class labels.

p_i is the proportion of examples in S having the i^{th} class label.

(4) (a) 5, 9, 11, 13, 15, 9, 8, 12, 11, 13, 18, 19, 18, 19, 13.

$$\text{width} = \frac{\text{max} - \text{min}}{n}.$$

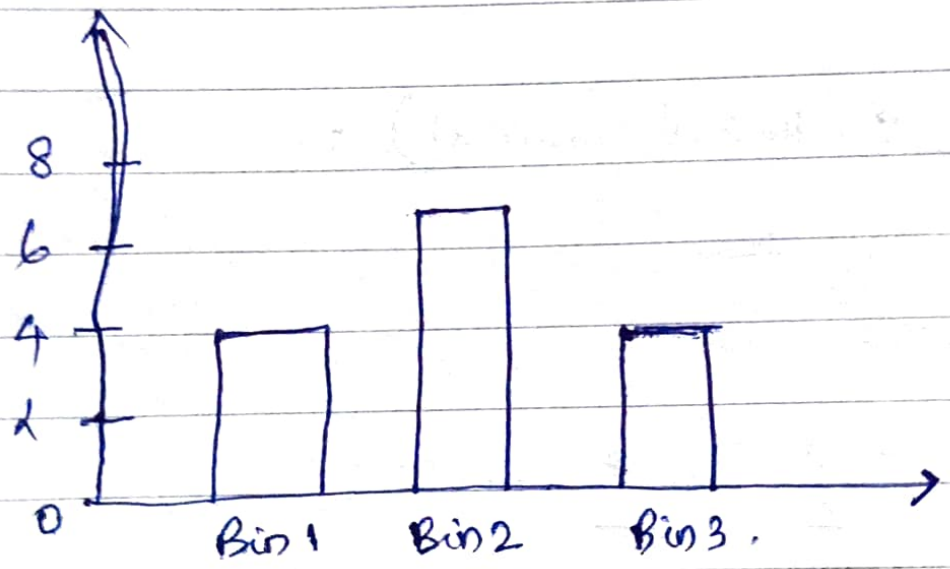
$$\text{width} = \frac{19 - 5}{3} = \frac{14}{3} = 4.666 = 5.$$

Range

Bin 1 : (5 to 10) : 5, 8, 9, 9

Bin 2 : (10-15) : 11, 11, 12, 13, 13, 13, 15

Bin 3 : (15-20) : 18, 18, 19, 19.



(5) **OLAP** (Online Analytical Processing).

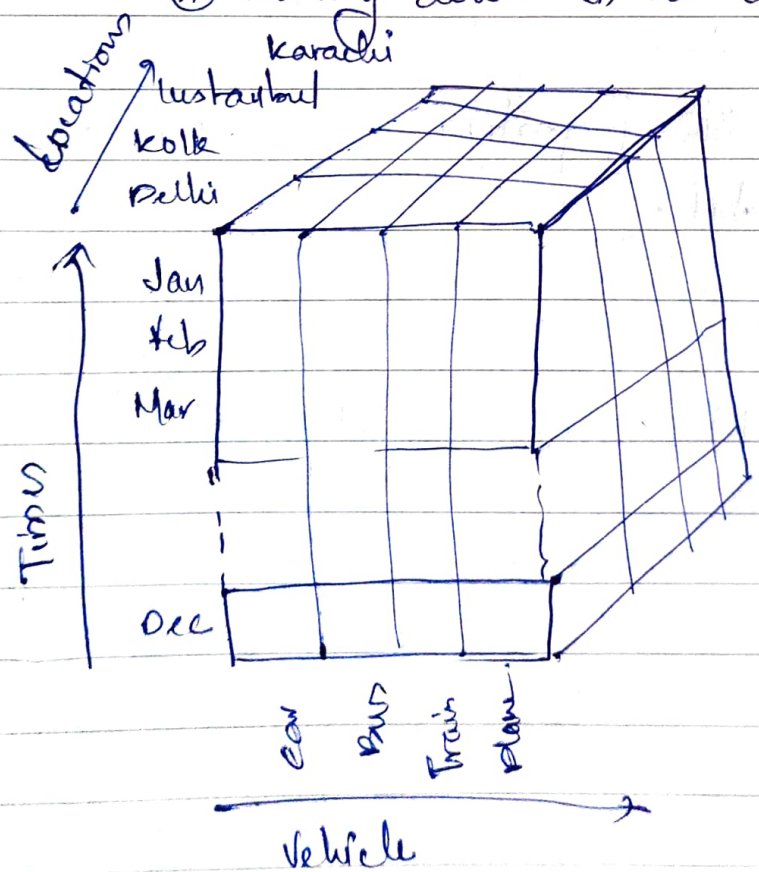
It is a s/w technology that allows users to analyze information from multiple database systems at the same time.

OLAP Operations

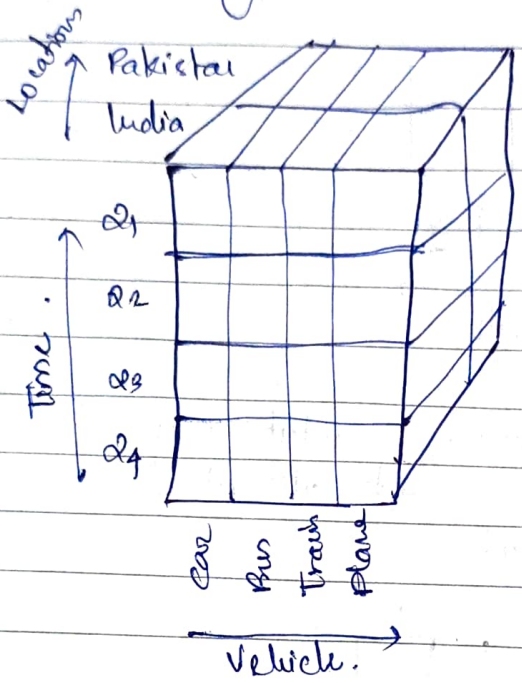
(i) **Drill Down**

In this, the less detailed data is converted into highly detailed data.

(*) Moving down in the concept hierarchy.

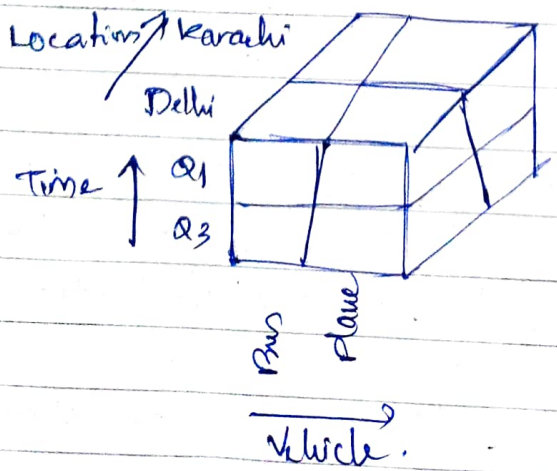


(ii) **Roll up** : It is just opposite of the drill down op. It performs aggregations on the OLAP cube.
- (i) Climbing up the concept hierarchy
- (ii) Reducing the dimensions.

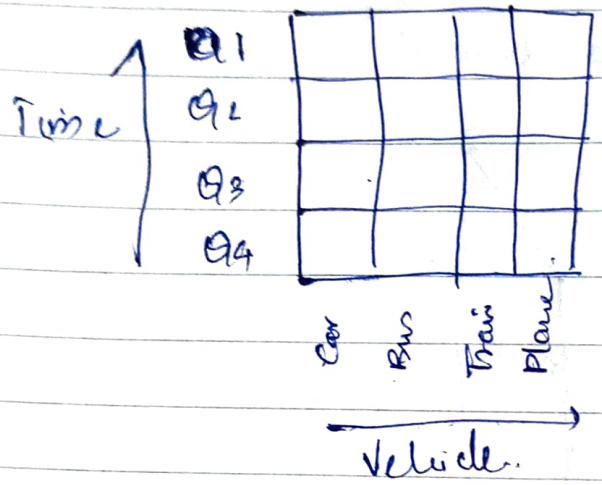


(iii) **Dice**: It selects a sub_cube from the OLAP cube by selecting 2 or more dimensions

- (i) Location : Karachi, Delhi
- (ii) Time : Q1, Q3
- (iii) Vehicle : Bus, Plane.

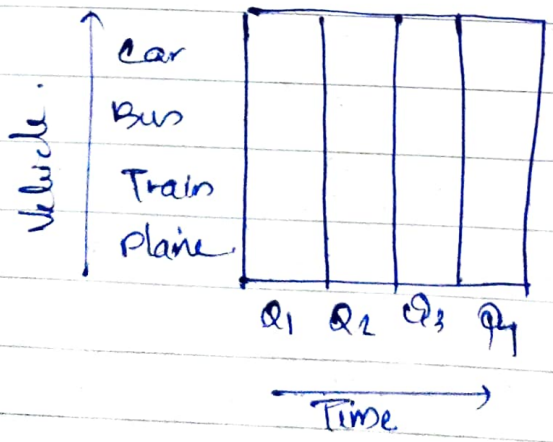


(iv) Slice: it selects a single dimension from OLAP cube which selects new sub-cube creation.



Location = Delhi.

(v) Pivot: It is also known as rotation operation as it rotates the current view to get a new view of the representation.



(6) (i) min_max normalization.

newMax = 1 newMin = 0

data = 200, 400, 500, 700, 1000

$$v' = \frac{v - \min A}{\max A - \min A} (\text{newMax}_A - \text{newMin}_A) + \text{newMax}_A.$$

$$v' = \frac{200 - 200}{1000 - 200} (1 - 0) + 1 = 0$$

$$V' = \frac{400-200}{1000-200}(1-0)+1 = \frac{200}{800}(1) = 0.25$$

$$V' = \frac{500-200}{1000-200}(1-0)+1 = \frac{300}{800}(1) = \frac{300}{800} = 0.375$$

$$V' = \frac{700-200}{1000-200}(1) = \frac{500}{800}(1) = 0.625$$

$$V' = \frac{1000-200}{1000-200} = 1$$

(6) (ii) z-score normilization

$$\Rightarrow v' = \frac{v_i - \bar{A}}{\sigma_A}$$

$$A' = \frac{200 + 400 + 500 + 700 + 1000}{5} = \frac{560}{}$$

$$\sigma^2 = \frac{1}{5}\left[(200^2 - 560^2) + [400^2 - 560^2] + (500^2 - 560^2) + (700^2 - 560^2) + (1000^2 - 560^2) \right]$$

$$\Rightarrow \frac{1}{5}\left[-273600 + -153600 + -63600 + 176400 + 686400 \right]$$

$$= \frac{1}{5}\left[372000 \right] = \frac{74400}{} = 272.764$$

z-score

$$200 \rightarrow \frac{200 - 560}{272.764} = -1.319$$

$$400 \rightarrow \frac{400 - 560}{272.764} = -0.795$$

$$500 \rightarrow \frac{500 - 560}{272.764} = -.298$$

$$700 \rightarrow \frac{700 - 560}{272.764} = .5132$$

$$1000 \rightarrow \frac{1000 - 560}{272.764} = 2.186$$

Ans $\Rightarrow$ $-1.319, -0.795, -0.298, 0.5132, 2.186$.

(6) (iii) Normalization of decimal scoring.

$\Rightarrow v' = \frac{v}{10^j}$ $\quad j = 4$

$\Rightarrow v' = \frac{200}{10000} = 0.02$

$\Rightarrow v' = \frac{400}{10000} = 0.04$

$\Rightarrow v' = \frac{500}{10^4} = 0.05$

$v' = \frac{700}{10^4} = 0.07$

$v' = \frac{1000}{10000} = 0.1$

(7) (i) I.G (S, aerial, animal)

A ⇒ 3 M - 2 R - 1

B - 2 f - 2

$$\text{Entropy}(S) = -\frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{1}{10}\log_2\frac{1}{10}$$

$$= 2.246 \Rightarrow \underline{2.25}$$

$$\text{Entropy}(S_{AM}) = -\frac{2}{2}\log_2\frac{2}{2} = 0$$

$$\text{Entropy}(S_{NO}) \Rightarrow -\frac{2}{8}\log_2\frac{2}{8} - \frac{1}{8}\log_2\frac{1}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{3}{8}\log_2\frac{3}{8}$$

$$\Rightarrow \underline{1.905}$$

$$I.G(S, \text{aerial animal}) \Rightarrow 2.246 - \left(\frac{2}{10}\times 0\right) - \left(\frac{8}{10}\times 1.905\right)$$

$$\Rightarrow \underline{0.722}$$

(7) (ii) IG (S, has legs)

$$\text{Entropy (Eyes)} = -\frac{2}{7}\log_2\frac{2}{7} - \frac{3}{7}\log_2\frac{3}{7} - \frac{2}{7}\log_2\frac{2}{7}$$

$$= 1.0327 + 0.5238 = 1.5567.$$

$$\text{Entropy (SNO)} = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3})$$

$$= 0.3899 + 0.52832$$

$$= 0.9183$$

$$\text{IG (S, has legs)} \Rightarrow 2.246 - \left(\frac{7}{10} \times 1.5567\right) - \left(\frac{3}{10} \times 0.9183\right)$$

$$\Rightarrow .88082$$