



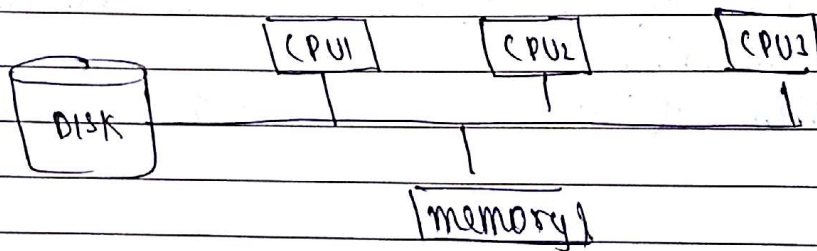
Unit: 2

Hardware & OS used in DW

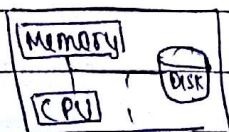
1. Parallel H/W technology
 - a) Symmetric Multiprocessing
 - b) Massively || processing system

clustered system

Symmetric Multiprocessing :- These sys consist of pair of about 64 processor that share memory & disk I/O resources equally under one copy of OS. Because " are shared hence, can be easily managed allows high speed interconn, high BW, & low commⁿ, latency is also achieved

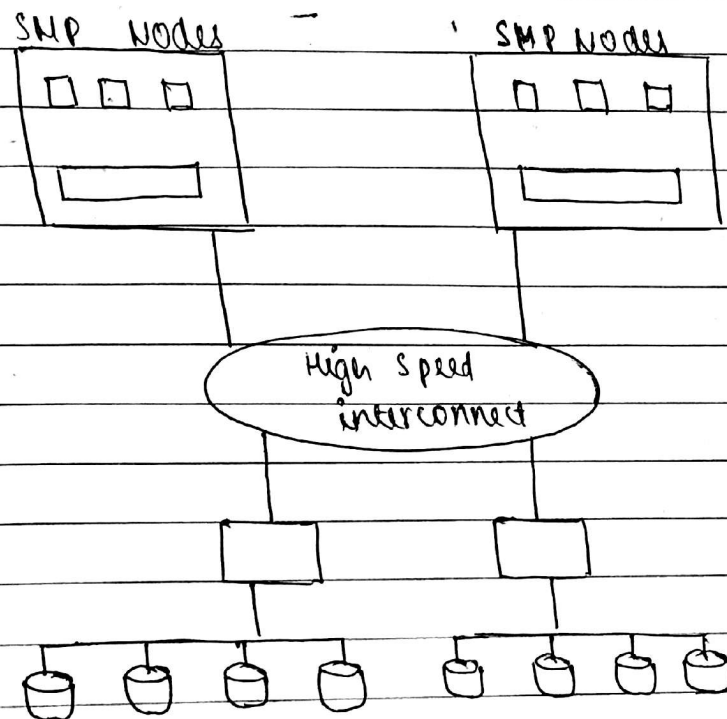


Massively || processing system :- It uses a large no. of nodes, each accessed using & interconnect mechanism that supports message based transfer typically of order 30x30 each node is self sufficient processor consisting of CPU, memory & disk sub system.



same as above

Clustered Sys: These sys are configured with multiported disk array so that nodes which have direct disk access ^{enjoy array} share same disk, I/O rates as stand alone, SMP. Node not having direct access to disk data must used high speed interconnect mechanism. DBMS or load balancing SW is responsible of distribution of threads across multiple node for max^m efficiency.



H/W selection Criteria :

1. scalability
2. financial stability
3. Price
4. low price.
5. High performance.

Delivery lead time: vendor can deliver H/W within time

6. Reference size: size that is using a similar unit for same purpose.

7. Availability of support.

Extraction &

transformation &

Warehouse
Tech.Data access &
retrieval

Source data

Middle ware

Extraction

Transformation

Quality assurance

Load image creation



OLAP

Report Lines

EIS/DSS

Data mining

Alert system

1. connectivity tool :- provides transparent access to source in heterogeneous computing.

IBM - Data Miner Joiner

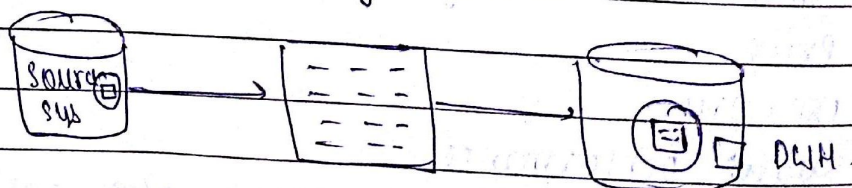
Oracle - Transparent Gateway

SAS - SAS/ connect

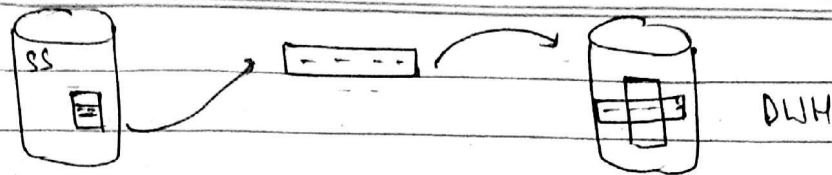
Sybase - Enterprise connect.

2. Extraction tool :-

- a) Bulk Extraction :- All applicable data are extracted from source sys for loading into WH. Entire WH is refreshed periodically.



- b) change Based Replication :- Only data that have been newly inserted or updated in source sys are extracted & loaded into WH.



SPW:-

Apertus

Carlton-tool available is passport.

Evolutionary Technology - tool is ETL extraction tool.
Platinum - Info pump.

4 Transformation Tool:-

Apertus Carlton - Enterprise integration

Data Mirror - Transformation Server

Informatica - Power mart Designer

5. Data Quality tool:- It helps to identify & correct data error ideally at source sys.

Data flux - Data Quality workbench.

Pine cone sys - Content tracker

Prim - Quality Manager

Validy tech - Integrity Data Reengg.

6. Data loaders tool:- Transform data into data WH.

Data Modelling tool - allow users to prepare & maintain an infoⁿ model of source & target database.



Cayenne SLW - Terrain

Relational matters - syntagma Designer

Sybase - Power designer warehouse architect

7. Warehouse Mgt tool : It assist data WH admin. in day to day mgt & " of

WH.

Pine cone system - usage tracker

Red brick " - Enterprise control & coordination

8 Data access & retrieval tool

a) OLAP tool : allows users to make adhoc queries or generate canned queries against WH db.

Arbor SLW : Essbase OLAP

Cognos : Power play

Intranet business system : R/OLAP XL.

b) Reporting tool : Allow user to produce graphic intensive & canned reports based on WH data.

IQ SLW : IQ / Smart servers.

Seagate SLW : crystal & Report

c) EIS : They provide diffⁿ executive reporting features including whatif or scenario based analysis capabilities for enterprise budgeting process.

Comshare - decision

Oracle - Oracle financial analyzer

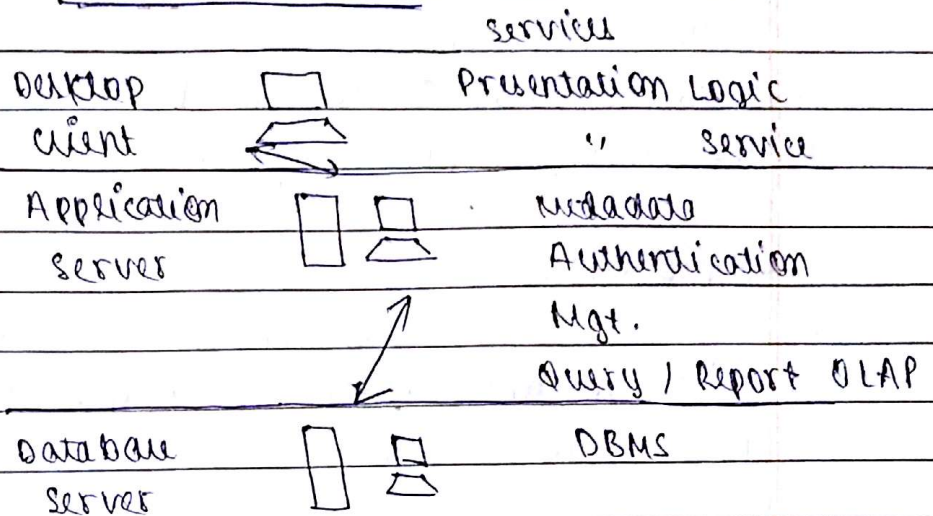
d) Alert sys :

Compulogic - dynamic query Manager

pine cone sys - Activator module.



Client Server arch



Today's DWH are built using client server arch. DWH executes on data ^{base} server component. Data repository of DWH sits on this m/c. Server component is a major component.

Adv 1. Work station independence.

2. Ease of effort & maintenance

3. Scalability

4. ^{Data} Integrity

5. System Independability

6. Centralized Mgt.

7. Resource Sharing

8. commⁿ cost are reduced.

9. Ease of effort & maintenance.

Disadv: Single pt. of error.

2. costly to maintain db server.

Working

1. Allows companies to take benefit of ll-computer techol
2. " processing to be perform close to source of



processed data thereby improving response time & reduce NW trafficking.

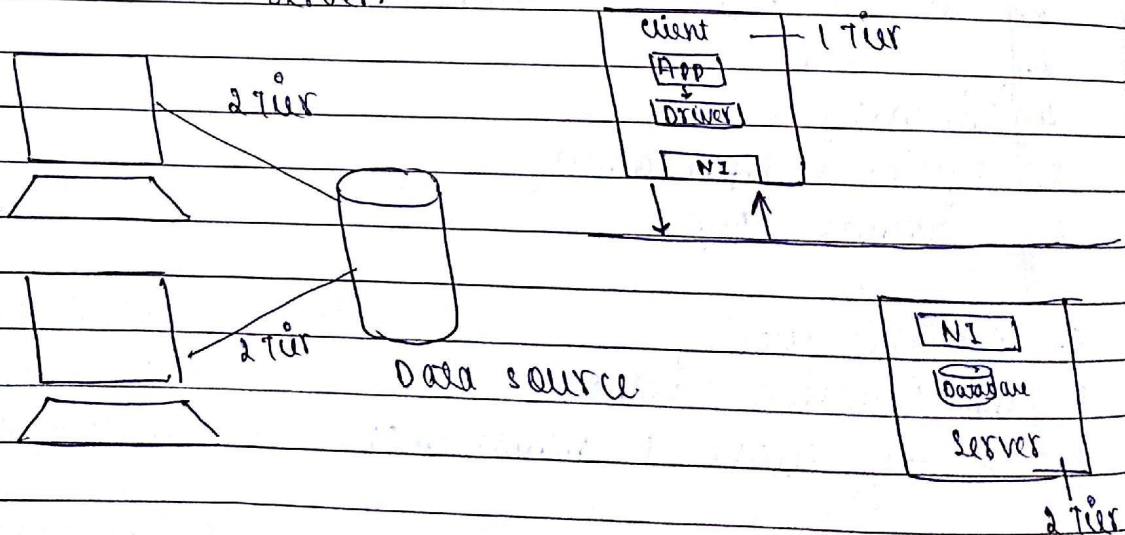
3. facilitate GUI & visual presentation technique.
4. Encourage Open system.

Components:

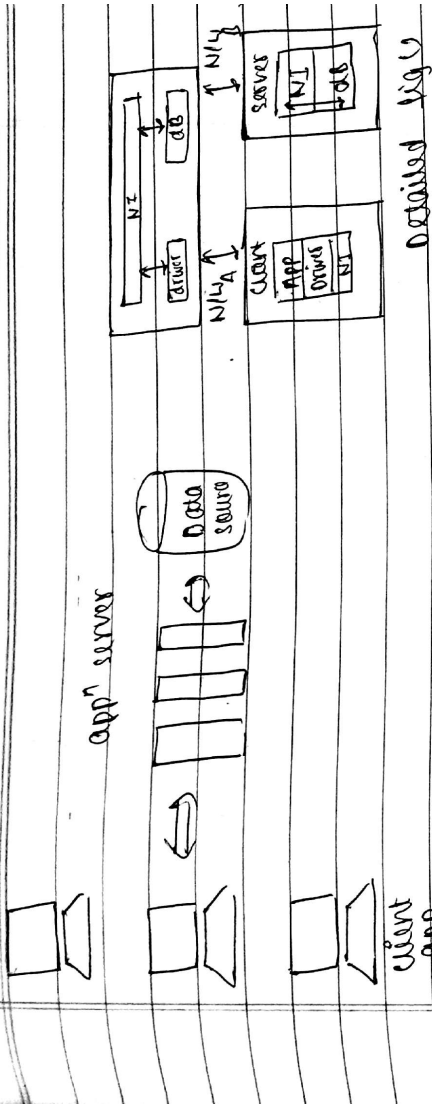
1. client 2. server 3. commⁿ Middle ware.
1. It is any computer which process the request service from the server.
2. server: Any computer providing services to client.
3. commⁿ MW: A " through which client & server communicate.

Tier

1. 2 tier arch: Direct commⁿ takes place b/w client & server.

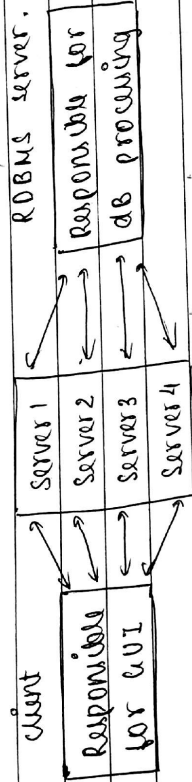


2. 3 tier arch: It comprises of presentation tier, business or data access tier & a data tier



Abstract fig (1)

3. N-tier Arch



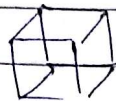
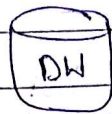
Thin & Fat client model

- Difference b/w thin & fat client
- Warehouse strategy.

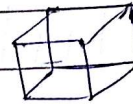
1. Preliminary
2. Preliminary architecture :- 30 hum architecture sochte hai vo strategy ka 2nd part wo jata hai. Jaise ki Data Mart jo hai involve honge, OLAP, EIS/ESS.

Alert system jo hai hai vo ek data se link honge yhi hai Preliminary architecture

- No of users.
- Location.



Data Mart



Report service.

3) Shortlisting options:-

" the best of the given tools & technology

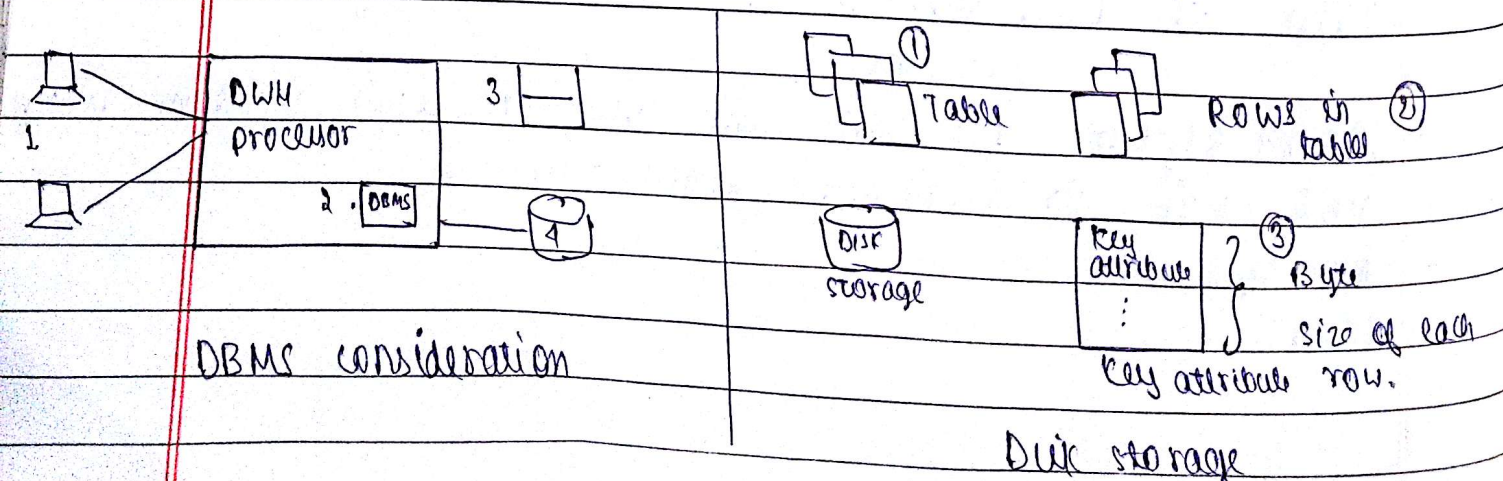
⇒ WH Mgt & Support process

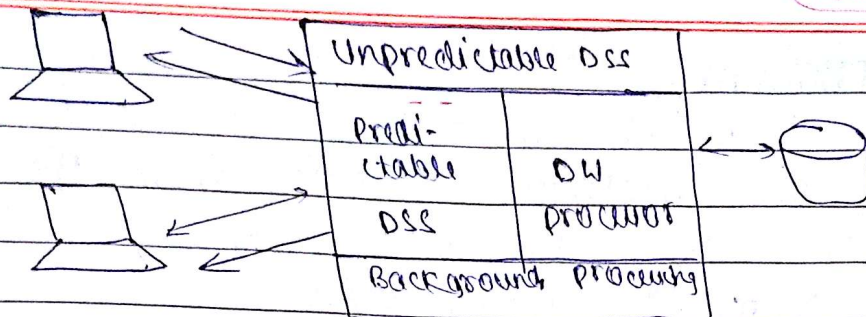
They are designed to address the aspects of the planning & managing DW project, subject to successful implementation & extension of SLW.

1. Define issue tracking & resolution process.

No.	Issue Description	Urgency	Raised by	Date opened	Date close	Resolved by	Resolution describes

2. Perform capacity Planning:- space required, m/c processing power, n/w BW. no. of concurrent users.





2) Define WH purging rule:

- Define mechanism for removing older data from data WH check illegal, regulatory or auditing req.

4) Define security mgt:-

- Secure to prevent loss of competitive info either to disaster or unexpressed users.
- Asset classification practice.
- Risk assessment & acceptance.
- Asset ownership.
- Policies regarding mishandling of security assets.
- How security violation are reported & responded to?
- Security awareness practice.
- " audit.

7) Define Backup & recovery strategy:-

- Data to be backed up.
- Batch window of DW.
- Max acceptance time for recovery.
- Acceptable cost for Backup & recovery.

6) Set a colleⁿ of WH user statistic:-

- Define mechanism for collecting these statistics.



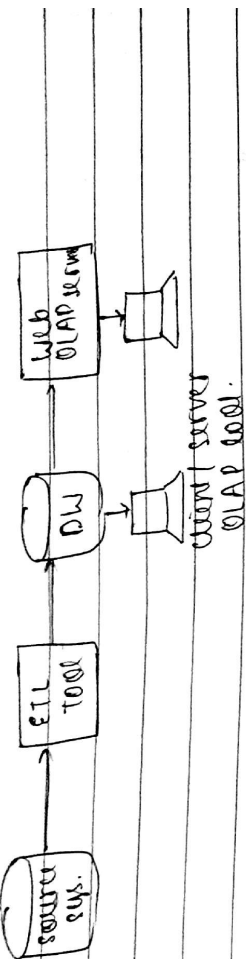
DWH Planning:-

1. Assemble & orient the team.
 - i) Identify the employees & brief them the project.
 - ii) Distribute copies of DW strategy.
 - iii) Set up team & specify roles.
 - iv) Give training if req.
 - v) Set up main milestones & check point.
2. Conduct decisional req. analysis.
 - Gain through understanding info needs of decision makers.
3. Conduct decisional source sys. audit.
 - a) Survey current source of data for DW.
4. Design logical & physical DW schema.
5. Produce source to target field mapping.
6. Select development & prodⁿ env. tool.
7. Create prototype for this Roll out.
8. " implementation plan for this rollout.

★ ETL

ETL takes place in data staging to reshape the relevant data from source system into useful info to be stored in DWH.

Steps 1. Extraction : Relevant data is extracted from source on the basis of its quality. It depends on how compressive & accurate the constraints implement it in source are, how suitable data formats, how the clear the schema are?



2. Cleaning: It improves data quality. There are specific dictionaries to rectify typing mistakes & recognize synonyms. Rule based cleaning enforces domain specific rules & define appropriate association b/w values.

3. Transformation: It converts data from operational source format into specific DWH format. Conversion & normalization both operate on storage format & units of measure to make data uniform. Matching associated equivalent fields in diff^r sources. Selection reduces no. of source fields & records.

4. Loading: Refresh-in this DWH data is completely rewritten. Older data is replaced. Update - only those changes applied to source data are added to DWH. Update is typically carried out without deleting or modifying pre-existing data.

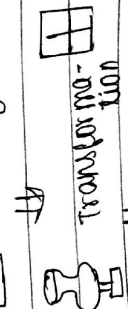
operational & external data



extraction



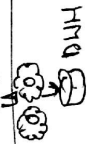
filtering



transformation



reconciled layer



DWH



Unit - 4

1. Frequent Pattern
2. Market Based Analysis (support), confidence
3. ~~Apriori~~ Algo → Numerical.

statistical measure in large DB :-

1. measuring the central tendency.
 2. " " dispersion of data.
- i) Mean :- $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$

vi)

- ii) Distributive measure :- It is measure that can be computed for a given dataset by partitioning data into smaller subset, computing measure of each subset & then merging the results in order to give a measured values for original data set.
 eg:- sum(), count() etc.

- iii) Algebraic Measure :- can be computed by applying an algebraic funcⁿ to one or more distributive measures

eg:- $\text{Average}() = \frac{\text{sum}()}{\text{count}()}$

- iv) Weighted Average :- Weight reflects the occurring frequency attach to their respective values.

vii)

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n}$$

- v) Median :- It is the middle value of the ordered set, when 'n' is even the median is average of middle two values.

viii)

$$\text{Median} = L_1 + \left(\frac{N/2 - (\sum \text{freq})_1}{\text{freq}_{\text{median}}} \right) \text{width}$$

L_1 = lower boundary of median interval

N = no. of values in entire data set.

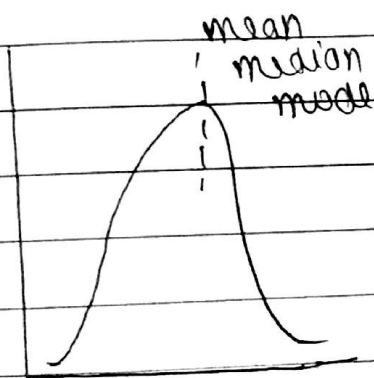
$\sum (\text{freq})_1$ = sum of freq. of all the interval that are lower than median interval.

$\text{freq}_{\text{median}}$ = freq. of median interval.

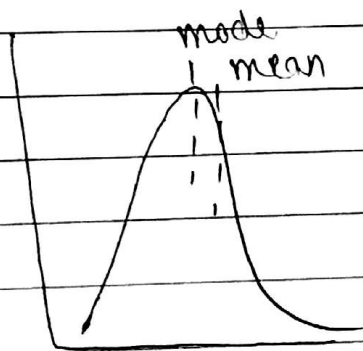
width = width of median "

vi) Holistic Measure : It is the measure that must be computed on entire data set as a whole.

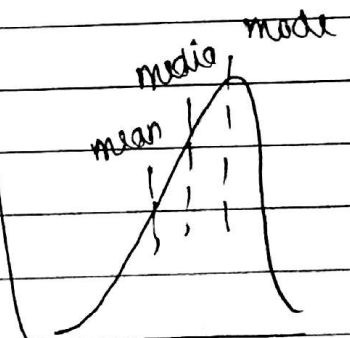
It cannot be computed by partitioning the given data into subsets & merging the value obtained from for the measure in each subset.



symmetric data



+ve skewed data



-ve skewed data

vii) Mode : It is set of data, value that occurs most frequently in the set is called mode. Data set with one mode is unimodal, 2 mode - bimodal, 3 - mode - trimodal.

viii) Mid-range : used to access the central tendency of data set. It is the largest & smaller value in the set.



age	f.
0-5	200
5-15	450
15-20	300
(20)-50	1500
50-80	700
80-110	44

$$L_1 = 20, \quad \text{width} = 50 - 20 = 30.$$

$$N = 3194. \quad \Rightarrow \quad \frac{N}{2} = 1597.$$

$$\text{freq med} = 1500.$$

$$= 20 + \left(\frac{1597 - 950}{1500} \right) \times 30 = 7.06$$