Date:

ETL Processing:-

There are number of Architecture to deploye the data warehouse.

Top-Down:-

Hub and Spoke:-

In this Data warehouse contain the historical or detail data.
In this dependent data Mart is present.
Data warehouse mostly is normalize we can build it in the dimension model like Snowflake or starflake.

Data Mart contain the summerize data if in the past frequent question is asked then we can add data mart.

Bottom up:-

In this independent data Mart are present we cannot apply across analysis. Because across analysis are not related to mart.

Date:

Data Mart Bus:-

We build data marts but its requirement are according to enterprise level.

=> ETL

We have number of data sources for warehouse or Mart.

We have to indentify which technique we use to access data and on which time and on what frequency.

first we create data extraction design

1) first identify the requirment
2) identify the data sources
3) identify the techniques to extract the data from sources which filter we have to apply.
4) first design is created then program is created on ETL tool and on the

Date:

We define if called or create programme.

There are three types of meta data
1) source meta data
2) Target meta data
3) intermidate meta data) data in the staging area

=> We cant directly convert data from source to target. We have to move it to the staging area.

=> In staging area we can apply different transformation according to the requirements of enterprise level.
like:- for male we chose 1, 2 because meaning of 0, 1 is in sense of true, false

Date:

1) In source system there are multiple domain for the gender but in target we have one domain In this case we do the scalar transformation This is one to one mapping.

2) Fist M L
A1, A2, A3
↓ Merge
A

1:1
M:1

is simple

3) (A) 202-1307
↓ split
A1, A2, A3
202 1307

This one is easy because splitting and ending is known

Some time we have to convert or split Address it takes some time weeks.

1:N
M:N

is difficult

That's why ETL takes cost and time.

Date:

⇒ Data Quality is a continous process if we deploy system with 50% accuracy then we better the quality or refine the programme.

⇒ Bussiness requirment must be clear is we have to move with the 100 percent quality or with less quality with multiple filter step. we verify it in the future. verifying process must take a less time.

⇒ Decision maker get data from sources and then give feedback to the operational system to take action

Date:

Loading Strategies:-

Option1 is that in the staging area we do aggregated construction then move the detail data to data warehouse and aggregated to mart.

Option2:- is that in the warehouse we do aggregated then move the aggregated data to mart. this increase load on the

In logical there are three types we get data from source and in the staging area transform and then move to target
In physical there are some time staging area is in the warehouse we do the decision of staging area seperate or not on the basis of load.

Date:

Data Extraction techniques:-

Immediate
Deferred.

if Data Extraction is Deffered and time frequency is on daily basis.

Current value vs periodic

operation system store only current value if we extract data of mid night then if the change is done three time in a single day. Then we only get the last value. for this we use Immediate extract

Every dimension has separate frequency technique and time.

if the operation system use the periodic means preserve the change. if the it is periodic then we can use the deferred exact method.

Date:

⇒ Immediate

1) Transaction log :-

if we perform any transaction in system then all the information is store in the log file. It is the auto function in the RDMS it store all the info of the transaction in the log file. if we change value then in log it store previous and new.

if the transaction failed then Database admin for use log file to recover the database.

if the log file full then it extend or create new log file.

if the transaction not failed then we not need log file.

Through log cache we can write in log file.

log file means history of change.

∴ In log file no activity perform for the warehouse, write down for the operation system and we read for the log. So performance high.

first option is to extract from the transaction log. if available

There is no performance issue or the extraction cost.

2) DB Trigger Programme:-

These trigger are call explicitly.

we write trigger for the tabel.

In warehouse if we want to exact data from trigger then write the programme in the source system.

if the log file is not available or we need data other then log. Then we write trigger in the source system from where we extract data. from source. when trigger can and it extract data from this activity and move it in the staging area.

So performance is high but low then

Date:

log and cost is high to create programe so there is addition cost.

Minor impact of performce of trigger on the source system. and development cost of the trigger.

3) Source Application

In this case the RDBMS is not available and we dont have log file and we are unable to trigger in the RDBMS in this case we write programme to extract files from the source Application and move it in the staging area. Write programme in the source Application

) There is a huge development cost of this programme and also huge performence issue.

In option we extract file from the source Application according to target data requirment.

Date:

- we use option 3 in the worst case.

The best option is the log file if the data that we need is not available in the log then use option 2.

These are the option if we need data extraction immediately. We check what is the strategy of the operation either for
1) current value 2) periodic value

if current then we need to preserve them.

preserve history

⇒ 2) Defferred :-

option 1 :- Date Time stamp

we extract data on daily basis or on monthly or on hourly then how can we now is it is the today data or not.
for this we use date time stamp then it is very easy to extract data

Date:

Option 2 :- Comparison

We have the copy of the yesterday and of mid night we extract today and subtract the Today from ~~then~~ yesterday and get the today

performence issue and we have to maintain large file cost high and file comparison programme create performence issue.

Data loading Strategies :-

-) Full Data Refresh
-) Incremental Data Refresh
-) Trickle feed - continous update.

1) full Data Refresh

In this we empty the loaded data and then in the staging area we merge the previous and

Date:

new data and then load if in the load label.
In this we reload the previous data again.

if the existing/previous data is 95% and new data is 5%. then if the new come we reload this 95% by merging it in the staging area.
Only insertion operation perform in the warehouse because we do merge in the staging we dont need to compare. It take less time if the data is high.

2) Incremental Data Refresh.

In this we only load the new data in the loading staging. By using ~~disruptive~~ destructive merge or constructive merge or append.
Destructive merge replace the previous.
Constructive merge store previous and new
Append only add the new.
In this we need to do the comparison and

if the data is small.

Date:

if the ratio of new and previous is 10% then the full Data Refresh is high if the ratio increase then the full Data Refresh is better.

if the ratio is low then we use Incremental.
But this is not hard and fast rule.

3) if we use active warehouse then we use Trickle feed - because in this continuous update is need.

Date:

In case of small dataset we use the Nested Loop join. Nested loop join perform efficient.

In case of large dataset or very large dataset we use the hash join, sort Merge join, Merge join.

→ DSS vs OLTP.

DSS is for the different flavour we do analysis in DSS and there is a read operation only so there is no conflict. In OLTP there is a conflict of read and write. we must have concurrency control protocol.

→ Optimizer is the component of DBMS software. Main role of query optimizer is to get the request and give the execution plan.

query optimizer have number of option then we select the best one from this.

if the optimizer select the Top 4 option for the query

Date:

then we can say it as good if it select the average option then average optimizer.

Optimizer get the request or the input and generate the execution plan other task of this is to select the best path to execute this query.

Path select on the base of temporary space communication channel in case of remote access it is important check the I/O operation.

1) I/O operation: almost every optimizer consider if I/O operation are most costly. In I/O operation we read data from disk to cache and write from cache to disk.

Two category of Optimizer
1) Ruled base OLTP
2) Cost base Data warehouse

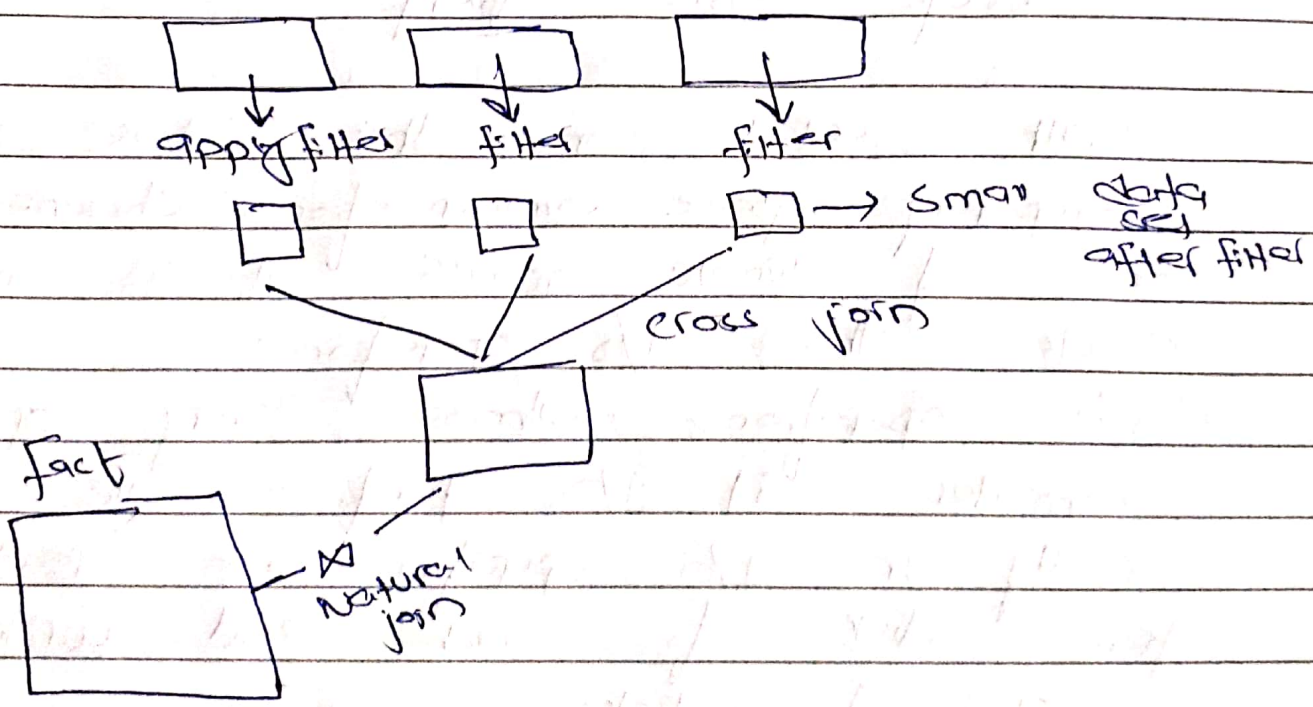
In OLTP Ruled base is efficient. In this according to set

Date:

of define rule plan is executed.

In Data ware house cost base is efficient. Cost base consider the other option like cost and plan execute on the base of cost.

Note:- When we apply star join



This star join is the cost base optimizer. In Rule base we join the first with fact.

The work of optimizer is to get the query rewrite it

Date:

according to define paradim. This functionality is build in the optimizer.

It is the work of designer to create the path and optimizer has to select the best path.

DSS vs OLTP

DSS contain mostly the large data set and it may consist of both OLTP and DSS. In this Merge or hash perform best.

OLTP contain the small dataset In this mostly nested join perform better.

As the administrator or the user we can give hint to the optimizer to force optimizer to consider it and execute it.

join with high selectivity is select first. Means when we join that join have the minimum No. of rows.

Date:

join with high selectivity is (select (means low rows) join with low selectivity is not select first (means high rows)

=> No. of rows and the data of label distributed on the block and the No. of duplicated and Null value in the column is predefine. All the statistic are precomputed in ETL we recollect the statistic when the major change done

when we do the first join with high selectivity then second and third if poor then it does not much matter.

Block NLJ :-

for each qualifying block of outer label we have to read the inner label one time. O(R x S)
100 x 1000

Basic NLJ :-

for each row we read

if we have index NLJ :-

R : S
1 : 100

$3 * (1 + 100)$

(Index cost + Base cost)

↳ average no. of rows

if we have Clustered NLJ :-

$3 * (1 + 1)$

if you place the file in order then there is a cost of mainfaince when we remove or add file we have to mainfain it.
if you place the file randomly then there is not a cost of mainfaince but difficulty in search.

In Block NLJ we cannot access or use index access path

Block factor is the number of rows in one block. It is represented by bfr.

Notes we divide the

In clusterd the index tabel give the index of the anchor block these block are corelated but we have to scan the all blocks.

if the index is on column and the tabel is odered on the base of clustered N2J. this column then it is

It is not naccessary that all the blocks are colocated. All blocks are coreloted

Examples:- index tabel is less-then-tabel of record.

DHW block size is greater then the OLTP block size

Combine selectivity:-

3% of (6% of (15-1. of (120,000)

K = 100 Blocks

B = 16K

rs = 128,000 (No. of Record)

Rs = 256 bytes (Rec size)

rA = 250,000

RA = 256 (Rec size)

Ris = 16

RiA = 16 bytes