TYPES OF PROBLEMS :-

① Decesion tree:- example:- 8.2.1, 8.2.2, 8.2.3, 8.2.4, 8.2.5, from Data warehousing & mining.

② Bayes model:- Example:- 8.2.9, 8.2.10, 8.2.11, 8.2.12, 8.2.13, 8.2.14, 8.2.15, 8.2.8, Page A-14.

③ frequent item sets

Example:- 10.4.1, 10.4.2, 10.4.3, 10.4.4, 10.4.5, 10.4.6, 10.4.7, 10.4.8, 10.4.9, Page A-9, page A-16

- Abbreviation you must know:- DW- Data warehouse
SCD- Slowly changing dimensions
MDS- multidimensional structure
DWI- Data warehouse Interfaces, BI- Business Intelligence
VSAM- virtual Storage access method, CMSM- Cross media storage Management
HSM- Hierarchical storage management
IDE- Immediate data extraction; CTDT- Capture through Database Triggers
CTTL- Capture through transactional log
CISA- Capture in source application; DDE- Deferred data extraction
CBDT- Capture based on date & Timestamp;
CBCF- Capture by Comparing files; ODS- operational data store
ODBC- open database connectivity
ERmodel- entity relationship model; OLTP- online Transaction Processing
OLAP- online Analytical processing;
ROLAP- Relational online Analytical processing; MOLAP- Multidimentional online analytical processing
etl or ETL- extract, tranformation & loading; DS- Data Science

Q) List the functions that a data quality tool is expected to perform.

— The two main functions that a data quality tool should posses are:-

(1) Error discovery features – It should:-

- ✓ Identify duplicate records.
- ✓ Identify values that are outside the domain range
- ✓ Find inconsistent data
- ✓ Check for range of allowable values
- ✓ Monitor trends in data quality over time
- ✓ Report to users on the quality of data

(2) Error Cossection features – It should:-

- ✓ Normalise inconsistent data
- ✓ Improve merging of data from dis-similar data sources
- ✓ provide measurements of data quality.
- ✓ prevent entry for data values whose values are outside the specified domain range.

Q) Define DW & enlist the benefits of data warehousing

→ DW were developed to meet the growing demand for information analysis that could not be met by operational systems. The DW is thus an informational environment which provides an integrated view of the enterprise and also provides enterprises current as well as historical data readily available for making strategic decisions.

→ Advantages of DW:-

① Potential high returns on investment & delivers enhanced business intelligence – Implementation of DW requires a huge investment, but it helps the organisation to take strategic decesion based on past historical data & organisation can improve the results of various processess like marketing segmentation, sales.

→② Competitive advantage – As previously unknown & unavailable data is available in DW, decesion makers can access that data to take decisions to gain competitive advantage.

→③ Saves time – As data from multiple sources is available in integrated foom, business users can access data from one place. There is no need to retrieve data from multiple sources.

→④ Better enterprise intelligence – It improves the customer service and productivity.

→⑤ High Quality data – Data in DW is cleaned & transferred into desired format. So data quality is high.

Q) explain s

→ In D
using

① User
team
role i
with con
database
at the le
an indiv
End-user
to the
granted
too. Ho
access a
then onl
Dimensio
cannot.

② Pass w
to get
Security
patterns
The D
record
access
a p
attem
temp
the

nt & delivers
mentation of DW
helps the
based on past
mprove the
arketing segmen-

dly unknown &
DW, decesion
ake decisions

ltiple sources
, business users
e is No need
ources.
improves the
y.
s cleaned &
o data quality

Q) explain Security mechanism in DW environment

→ In DW environment, security is provided using three mechanisms:-

(1) User privileges – For DW, the project team prefers a "role" based security. A role is nothing but grouping of users with common requirements for accessing the database. Access privileges may be granted at the level of a role or at the level of an individual. ex:- Consider John is an end-user. you have granted certain privileges to the users under this role. All privileges granted to end-users will be availed by John too. However, if some extra privilege to access a dimention table is also given to John, then only he can access that one extra Dimension table & the rest of the end users cannot.

(2) Password protection – Users need passwords to get into the DW. It is the duty of Security administrator to set up acceptable patterns & the expiry periods for the password. The DW security mechanism must make a record of unauthorised attempts to gain access using invalid passwords so that after a prescribed number of such ~~events~~ attempts, the user must be suspended temporarily from the DW, that is, until the DW administrator reinstates the user.

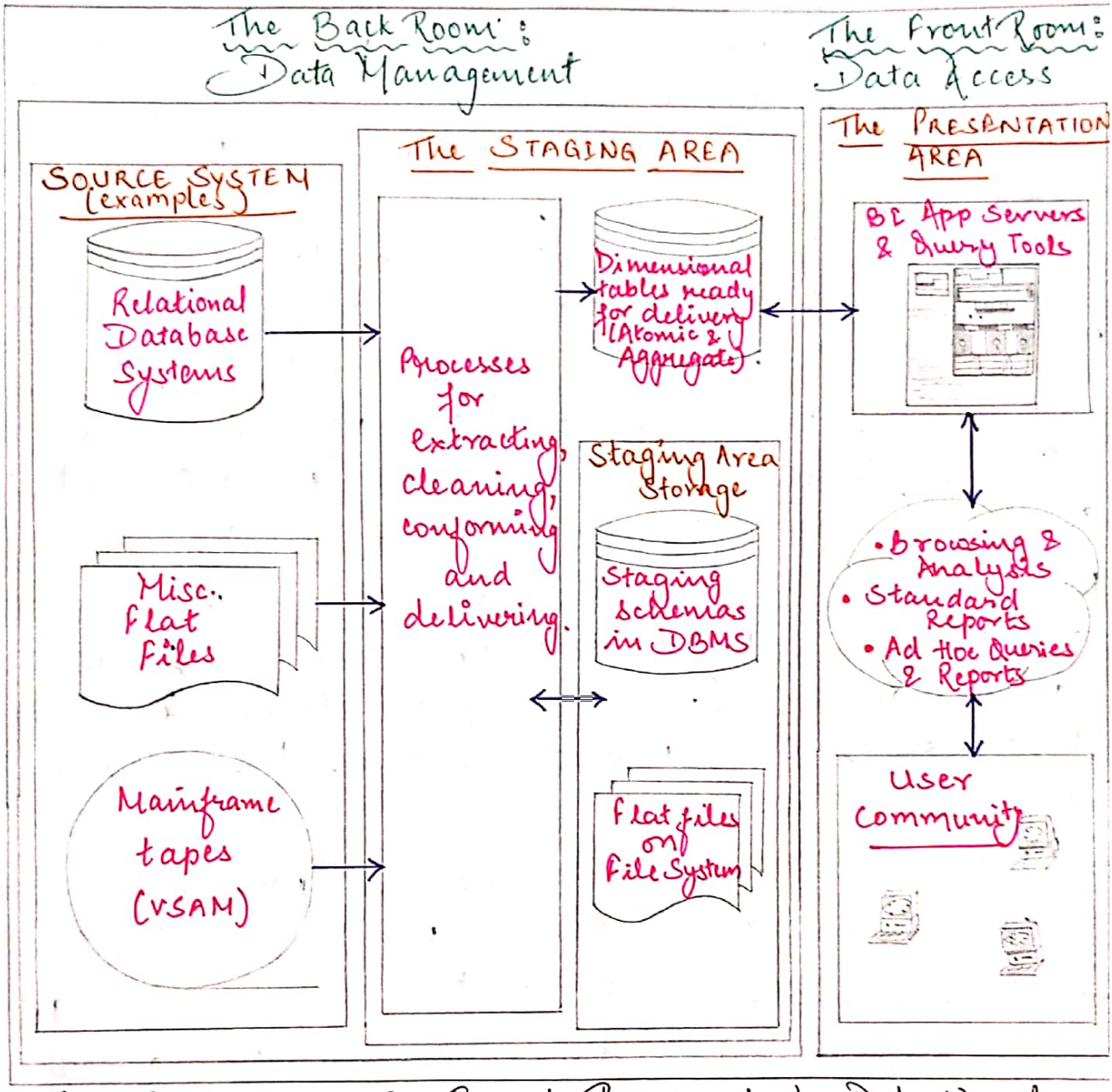
③ Security tools :- In
Security provided by
primary security te
also have third par
installed to govern t

Q) Difference between top-

TOPDOWN Approch
- Data is extracted from the operational system, trans-formed, cleaned and int-egrated to finally store it in DW
- Presents an enterprise view of data.
- Single central storage of data
- Implementation of Centralised rules & Control
- NO proof of concept with high risk to failure.

Q) With neat figure, explain front room and back room of a DW

- The back room & front room of DW are physically, logically & administratively seperate.
- The raw data coming from source systems is usually written directly to disc with some minimal restructuring. Data from structured source systems often is written to flat files or relational tables. This allows the original extract to be as simple and as fast as possible & allows greater flexibility to restart the extract if there is an interruption.
- The level of data quality acceptable for the source systems in most cases is different from quality required by DW. Data quality processing may involve many discrete steps including checking for valid values, ensuring consistency across values & checking whether complex business rules and procedures have been enforced.
- Data conformation is required whenever two or more data sources are merged in DW.
- The whole point of the backroom is to make the data ready for querying. The final and crucial backroom step is physically structuring the data into a set of schemas.



The Back Room & Front Room of A Data Warehouse

Q) Explain the role of Metadata in etl environment & describe the classification of metadata.

- Metadata in the data warehouse is similar to data dictionary in a database management system. The Metadata Component stores "data about the data" in DW.
- The metadata is often used for building, maintaining, managing & using the DW. It is the key to providing users & developers with a roadmap to the information in the warehouse. Thus, it forms an essential ingredient in the transformation of Raw data into Knowledge. The three main functions that metadata performs in a DW environment are that it :-

(1) Connects the different parts of DW thereby acting as a glue that connects all parts.
(2) Provides information about the contents of the data and its underlying structure to the DW administrator and other users.
(3) Enables end users to search for the desired data in their own business terms.

- Without a proper metadata in place, data stored in warehouse will be meaningless since the users will not be able to know - what, where, why & how the data exists within the organisation.

Data	
12233870	- This string of digits could either be sales of any product or population of city.
06/07/18	- This could be either date of birth or date when a particular sales transaction took place. It could either mean 6 July 2018 or 7 June 2018.
Metadata	Contains Data about data.
12233870	- Sales of T-Shirt in western zone
06/07/18	- Refers to sales transaction in format dd/mm/yy.

- Metadata can be classified in three main groups:-

(A) Operational metadata - Data for the DW comes from several operational systems of the organisation which contain different data structure that have varying field lengths & data-types. In selecting data from the source systems, you may either have to split certain records or may have to combine parts of record from different fields. Operational metadata solves this purpose by containing all of this information about the operational data source.

(B) Extraction & tranformational metadata
- It contains data about the data extraction from the source systems & the various transformation~~s~~ techniques that were applied to the data before storing it in DW.
- The primary reason for this metadata is to map every individual data element from its source system to the DW.

(C) End user metadata - It acts as a navigational map of DW by enabling end users to find information from DW.
- End user metadata translates a cryptic name code of a data element into a meaningful description of data element so that end users can understant & use the data.

ex:- metadata clarifies that the data element 'CName' represents 'Customer Name'.

Q) with a
of mult
- The m
① visua
in ter
it con
② Exec
agai
and
③ Spec
④ visu
The o
structu
represe
reposito
flow

MDS dat
cache

Da
ware
Admi

f

2) a) With a neat diagram, explain the overall framework of multidimensional structure.

- The multidimensional structure allows:-

① visualisation of data warehouse schemas in terms of multidimensional model so that it can be used ~~for~~ as reference for querying.

② Executing of textual & graphical queries against available multidimensional schema and views.

③ Specification of views.

④ Visualisation of result set of query execution

The overall framework of multidimensional structure is shown in fig 1. The Rectangles represent software modules, Cylinders represent repository of data & lines denote direction of flow of information.

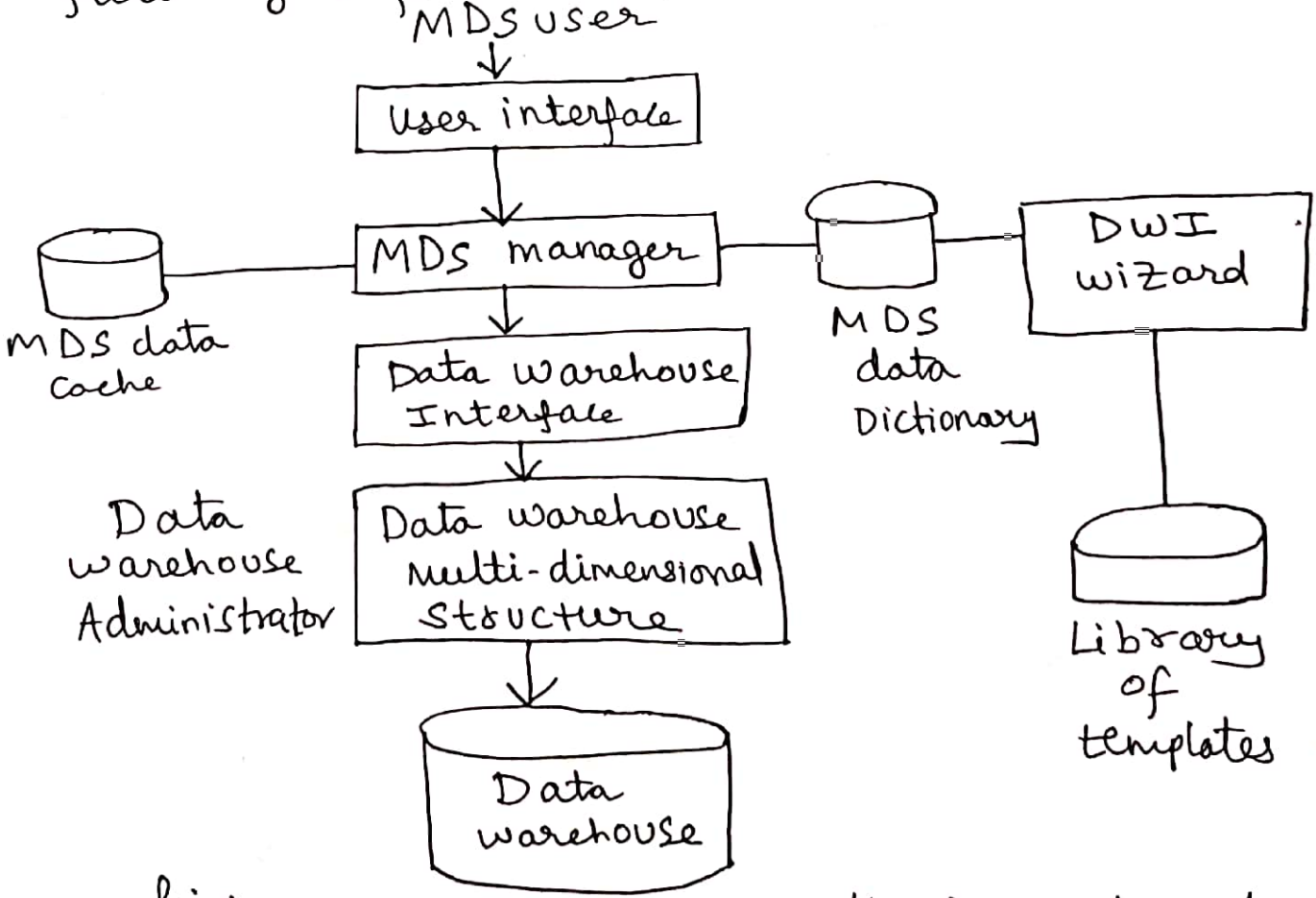


fig 1:- overall framework of multidimensional structure

Q a) Differentiate between ROLAP & MOLAP.

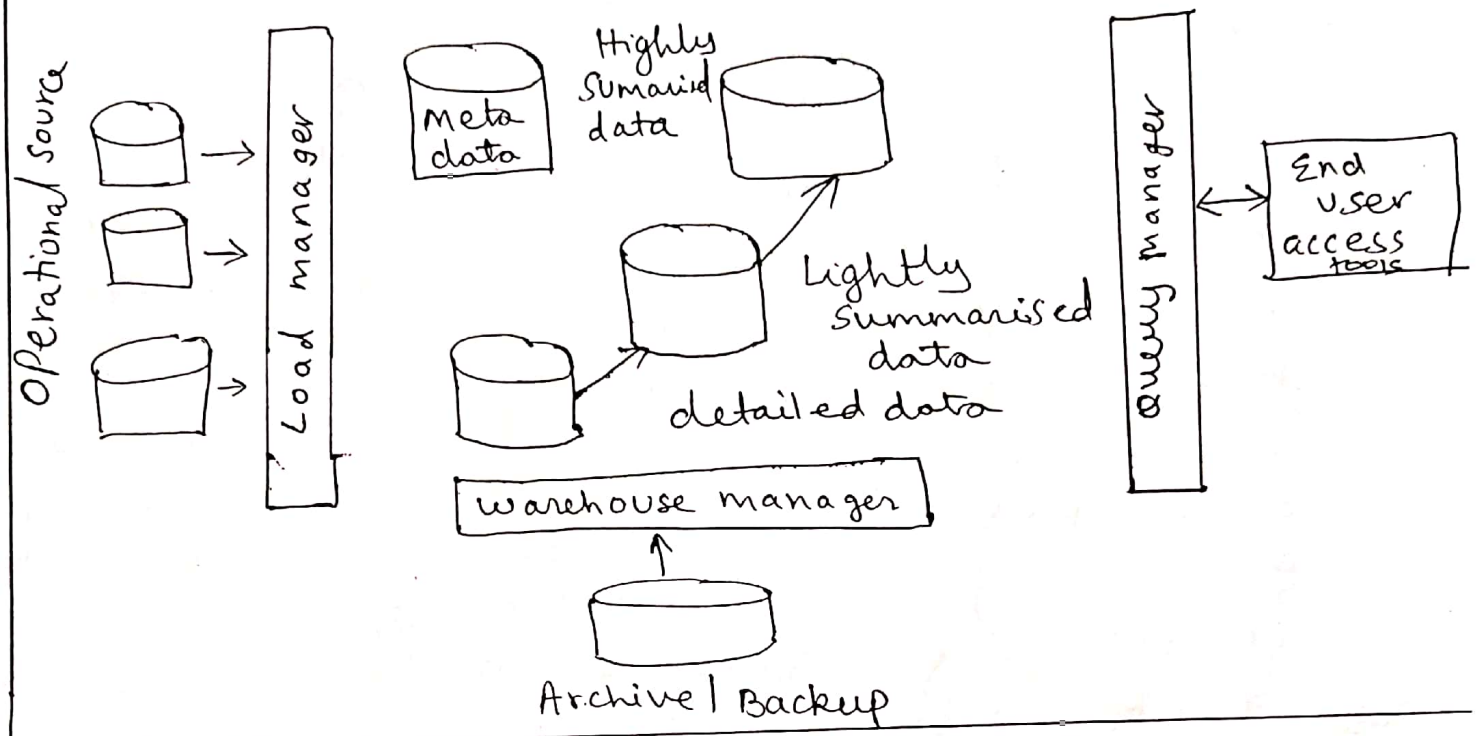
ROLAP	MOLAP
– ROLAP is relational online analytical processing where data is stored in the form of tables	MOLAP is multidimensional online analytical processing where data is stored in multidimensional array
– Data is stored & fetched from the main DW.	Data is stored & fetched from proprietary database
– very large data volumes can be handled by ROLAP	MOLAP can handle moderate data volumes
– Limitations on Complex analysis functions	Large library of functions for complex Calculations

Q) a) What is data velocity? Explain cyclicity of data.

- It is the speed with which data passes from initial capture to the point of use. It is calculated by finding out average time elapsed between the entry of data into the system & the usage of data by end-user.
It includes the time needed for editing the data, passing it to appropriate application then for Extraction, Tranformation & Loading [ETL] processing so that data can be stored in warehouse from where the end user will finally access it.
The main factor that affects the data velocity is the integration process - more the data to be integrated lesser is the velocity.
- Cyclicity of data means time elapsed between change of data in operational system & reflection of that change in the DW.
ex:- If a customer name was edited in operational system on 6th July 2017 at 12PM & this change was reflected in the DW on 7th July at 8AM, then cyclicity of data is 20 hours.

Q) With neat figure, explain the architecture of a data warehouse.

fig:- Data warehouse Architecture



- The data in data warehouse comes from operation systems of organisation as well as other extern sources. These are collectively referred to as source systems.
- The data extracted from source system is stored in an area called as data staging area where data is cleaned, transformed, combined and duplicated to prepare data in warehouse.
- The three different kinds of systems that are required for a warehouse are:-
  i) Source systems
  ii) Data Staging area
  iii) Presentation servers.

Operational source :- The source of data is supplied from mainframe systems in traditional network and hierarchical format. Data can also come from relational DBMS like Oracle, Informix.

Load manager :- It performs all operations associated with extraction and loading data into data warehouse. These operations include simple transformations of the data to prepare the data for entry into the warehouse.

Warehouse manager :- It performs all operations associated with the management of data in the warehouse. The operations include :-

1. Analysis of data to ensure consistency.
2. Create indexes & views on the base table.
3. Denormalisation.
4. Generation of Aggregation
5. Backing up & archiving of data.

Query manager :- It performs all operations with management of user queries. This component is usually constructed using vendor end-user access tools & custom built programs.

Detailed data :- This area of warehouse stores all detailed data in database schema. The detailed data is added regularly to the warehouse to supplement the aggregated data.

Lightly & Highly Summarized data :- This stores all predefined lightly & highly summarised data generated by warehouse manager. The main goal of the summarised information is to speed up the query performance.

Archive / Backup data :- The detailed & summarised data are stored for the purpose of archiving and backup. The data is transferred to storage archives such as magnetic tapes or optical disks

Metadata :- It is used for various purposes like:-

① Extraction & loading process
Metadata is used to map data sources to a common view of information within the warehouse.

② warehouse management process
Metadata is used to automate the production of summary tables.

③ Query management process.
Metadata is used to direct a query to the most appropriate data source.

End user ACCESS tools :- Examples of end user access tools are reporting and query tools, Application development tools, data mining tools etc.

Q) Compare dimentional

Dimentional model

(A) Support ad-hoc queries for business analyst & Complex analyses.
(Data warehouse & multi-dimentional database)

(B) Simplify the view of data model. You can rotate the data cube to see different views of data

(C) It is asymmetric.

(d) Permit redundancy

Q)(a) What is data cleaning. Mention reasons for dirty data. List the steps in data cleaning.

- Data cleaning also known as data scrubbing deals with detecting and removing errors and inconsistencies from data in order to improve quality of data. Data quality problems arise due to misspellings during data entry, missing values or any other invalid data.
- Reasons for dirty data:-
  - Dummy values.
  - Absence of data.
  - Multipurpose fields.
  - Contradicting data.
  - Inappropriate use of Address lines.
  - Reused primary keys.
  - Data integration problems.
- Steps in Data Cleaning are:-

① Parsing - It is a process in which individual data elements are located & identified in source systems and then these elements are isolated in target files. ex- parsing of name into first name, middle name, last name or parsing address into Street name, city, state, country.

② Correcting - This is the next phase after parsing in which individual data elements are corrected using data algorithms & secondary data sources.

③ Standardizing – In this process, Conversion routines are used to transform data into a Consistent format using both Standard & Custom business rules.

④ Matching – Matching process involves eliminating duplications by Searching and matching records with parsed, Corrected & Standardised data using some standard business rule. ex- identifie of Similar names & addresses.

⑤ Consolidating – It involves merging of the records into one representation by analysing & identifying relationship between matched records.

⑥ Data Staging – It is an interim step between data extraction & remaining steps. Using different processes like native interfaces, flat files; FTP sessions, data is accumulated from asynchronous sources. After a certain predefined interval, data is loaded into warehouse after the tranformation process. NO end user access is available to the staging file. For data staging, operational data store may be used.

Q) Explain format revision, decoding of fields, splitting of fields, character set conversion, conversion of units & de-duplication in ETL.

- These include changes in data-types & lengths of individual data fields.
- When data comes from multiple source systems, the same data items may have been described by different field values.
- Earlier legacy systems stored names & addressess in large text field. All the components of Name- first, middle & last were stored in one large field called "Names". But the need today is to split the individual components in separate field so that operating performance can be improved by indexing on the individual components.
- This type of data tranformation is done to textual data to convert its character set to an agreed standard character set.
- Many companies may have global branches. So the sales amount may be represented in different curriencies in different source systems. Before moving the data into DW, you need to convert the figure into a common unit of measurement.
- Sometimes the record for the same customer may be stored in many files. When you

extract data from source systems, you have to pay special attention to find such duplicate records & remove duplicates while storing record in DW, ensuring that the information about the customer is stored only once forming a single record.

Q Q) Write a note on Loading in a temporal & non-temporal data-mart.

- Loading a temporal data mart:-

(A) Complete refresh - The data marts are loaded by reloading the entire tables every time. This is done by truncating the tables and then loading the data again. The key benefit with this technique is that it captures everything that is in the transaction repository & reduces the back-up requirements.

(B) Cumulative refresh - In this technique you just need to append to the data mart all facts from transaction repository that has appeared since the last load program was executed. This technique speeds up the load process considerably for large data marts.

- Loading a non-temporal data mart:-

(C) Loading from transaction repository
When data is loaded from transaction

repositories, then we are actually going directly to the source systems. In this case we build the current data mart without relying on the temporal data mart having been loaded first.

(3) Loading from temporal data mart

This method seems to be most attractive as it takes full advantage of the work already done by the load of the other data mart. As long as the dependency on the load time for the temporal data mart is not an issue, this technique is always recommended for best performance.

Q 9) Explain the different mechanism to manage the flow of data from DW to near line storage.

— The three different mechanisms are:-

(1) Manual transfer

In manual tranfer the data warehouse administrator himself moves the data from one medium to another. The administrator places a DW monitor that continiously monitors the usage of data. The data that is not being used frequently is moved from DW environment to near line storage. This option provides flexibility & uses minimal technology.

(2) Hierarchical Storage Management (HSM)
This mechanism is fully automated & is free from any human intervention. It moves entire set of data between the DW and the near line storage.

(3) Cross Media storage Management (CMSM)
It is a fully automated procedure. The main difference between HSM & CMSM is that CMSM operates at the row level of granularity so that rows can be migrated to and from the DW and near line storage.

- When the user poses a query that needs data stored in the DW, the system fetches the data and proceeds with the execution of the query. However, if data is present in any near line storage, then first the system collects the data from near line storage media & then proceeds with the query.
- In CMSM approach, it is the DW monitor that continiously monitors what data is being used by the queries posed by the end-users.
- It identifies the data that is not being used frequently by queries & monitors the usage of data at the row level so that data can be more finely tuned with the warehouse.

Q) Explain roles & responsibility with the data quality framework.

Roles	Responsibilities
Data Consumer	They are the users who use the DW for queries, reports & analysis
data producer	These people are charged with maintaining the quality of data input from source systems
Data expert	They are responsible for identifying pollution in source system
ata policy dministrator	They are responsible for resolving data corruption as data is transfered & moved into DW
ata integrity Specialist	They are responsible for ensuring that data in source systems conform to the business rules
Data correction authority	They are responsible for applying the data cleansing techniques.
Data Consistency expert	These experts are responsible for synchronising the data within the DW repository.

Q Q) Explain the different levels of testing a DW.

– The different levels of testing are :-

1. Unit testing
2. Integration testing
3. System & acceptance testing
4. Performance testing

(1) In unit testing, also called white-box testing, each development unit is tested on its own by the developer of that particular module.

(2) In Integration testing, different modules make up a component of data warehouse application are tested to ensure that they work together.

(3) In System testing, the entire DW application is tested as a single unit.

Acceptance testing also tests the Complete DW application but the basic difference is that during acceptance tests the users will conduct their own tests on the system.

(4) – It is the most important aspect after data validation as many DW systems might satisfy all the tests above & may fail at performance level test at the end.

Performance testing should :-

(A) Test whether ETL process Completes within the load window. Also check the ETL process for the time taken for updating & processing of reject records.

(B) Test the time taken to refresh standard reports.

(C) Test the time taken for refreshing Complex reports.

Q 9) Explain parsing, correcting, standardising, matching, consolidating & Data staging in etl.

- It is a process in which individual data elements are located and identified in Source Systems & then these elements are isolated in target files. ex:- parsing of name into first name, middle name, last name or parsing address into street name, city, state, country.

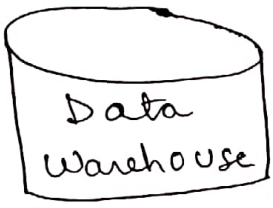
This is the next phase after parsing in which individual data elements are corrected using data algorithm & Secondary data Sources.

In this process Conversion routines are used to transform data into a consistent format using both standard and custom business rules.

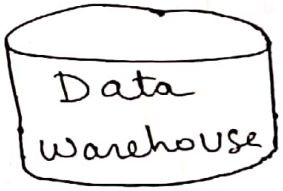
- It involves eliminating duplications by searching & matching records with the parsed, corrected & standardised data using some standard business rules.
  ex:- Identification of similar names and addressess.
- It involves merging the records into one representation by analysing & identifying relationship between matched records.
- It is an interim step between data extraction & remaining steps. using different processes like native interfaces, flat files, FTP sessions, data is accumulated from asynchronous sources. After a certain predefined interval, data is ~~located~~ loaded into warehouse after the transformation process. No end user access is available to the staging file. For data staging, Operational data store may be used.

Q) a) Explain the Concept of data Granularity

- Granularity refers to the level of detail or the level of summarisation of data in DW. In technical terms, data granularity is inversely proportional to the level of data. More detailed data means lower level of granularity. Similarly, less detail indicates higher level of granularity.
- ex:- A simple sales transaction would be at a low level of granularity whereas a summary of all the sales transactions for the entire month would be at a high level of granularity.



High level of detail - Low level of granularity.
ex:- details of phone call made by a customer in a month.



Low level of detail - high level of granularity
ex:- Summary of phone calls made by a customer in a month

- Generally in DW, data is kept at different levels of summarisation. Depending on the query the user can go to particular level of detail & satisfy the query. In a DW, if data has to be kept at lowest level of detail, then a lot of data has to be stored. Thus the choice of granularity is an important design issue because:-

(A) Lower the level of granularity, the larger is the amount of data stored in warehouse.

(B) Higher the level of granularity, the lesser is the level of detail for which queries can be answered.

- The choice of granularity, calls for a trade-off between volume of data and the level of query detail.

Q) List the features of a data warehouse (DW)

- The features of data warehouse are as follows:-

① Subject-oriented

Data warehouse are designed to help analyze data. ex:- To learn more about banking data a warehouse can be built that concentrates on transactions, loans etc. This warehouse can be used to answer questions like which customer has taken maximum loan amount for last year. The ability to define data warehouse by subject matter, in this case "loan", makes the data warehouse subject oriented.

② Integrated

A data warehouse is constructed by integrating multiple heterogeneous data sources like relational database, flat files, online transactional records. The data collected is cleaned & then data integration techniques are applied which ensures consistency.

③ Non-volatile

Non volatile means that once data entered into the warehouse, it cannot be removed or changed because the purpose of warehouse is to analyze the data.

④ Time Variant

A data warehouse maintains historical data. ex:- A customer record has details of his job, all his previous jobs (historical information) All data in DW is identified with a particular time period.

Q) a) Explain the data extraction process in etl with figures. (1b)

- The data flows from the data sources and pauses at staging area. After transformation and integration, the data is made ready for loading into data warehouse repository. For majority of DW, the primary data source consists of enterprises operational systems. Effective data extraction strategies include

a) Identifying the application & systems from which the data will be extracted.

b) For each identified data source, determine the method for data extraction; i.e whether it will be done manually or by using tools. (1c)

c) Determine the extraction frequency.

d) Estimate the acceptable time window for the process from each data source. The (DE) data extraction technique can be broadly classified

(1) Immediate data extraction technique (2)

In this technique, the data extraction is real-time. It occurs as transaction happen at source databases & files. The three options for Immediate data extraction technique is:- (2a)

(1a) Capture through transactional logs.

It makes use of transaction logs of DBMS. It reads the transaction log and selects all committed transactions. Since logging is already done as a part of transactional processing in all modern DBMS, there is no extra overhead incurred in

Operational system.

(1b) Capture through database triggers

Triggers are stored ~~procedures~~ procedures that are stored on databases & fired when certain predefined events occur. Triggers can be created for all events for which data needs to be captured. The output of the trigger program is written on a separate file that will be used to extract data.

(1c) Capture in source application

In this method, source application is used to capture data for DW. All relevant applications that write to the source files are modified to write all adds, updates and deletes to both the source files and database tables.

(2) Deferred data extraction

In this technique, data capture does not take place in real time. The capture is done at a later point of time.

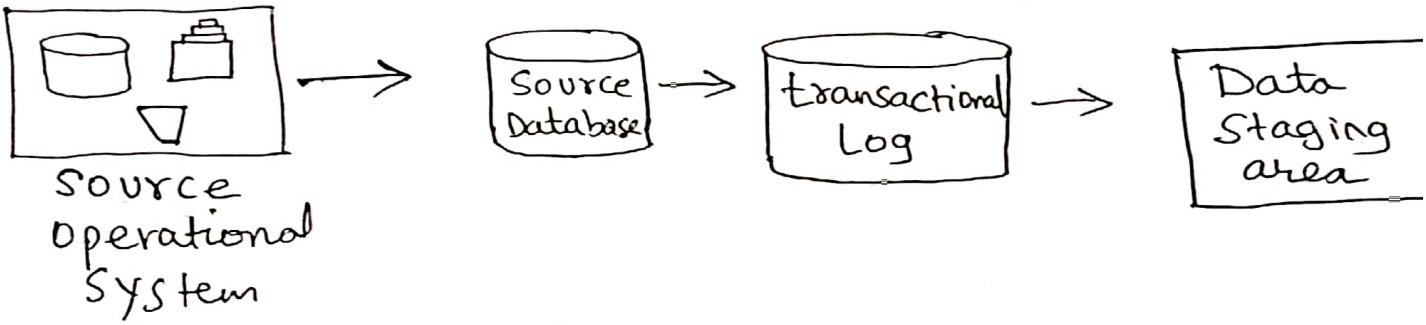
(2a) Capture based on date & timestamp

Everytime a record in the source system is created or updated it will be marked with a timestamp that will be used for selecting the record for data extraction. The timestamp shows the date and time at which the source record was created or updated.

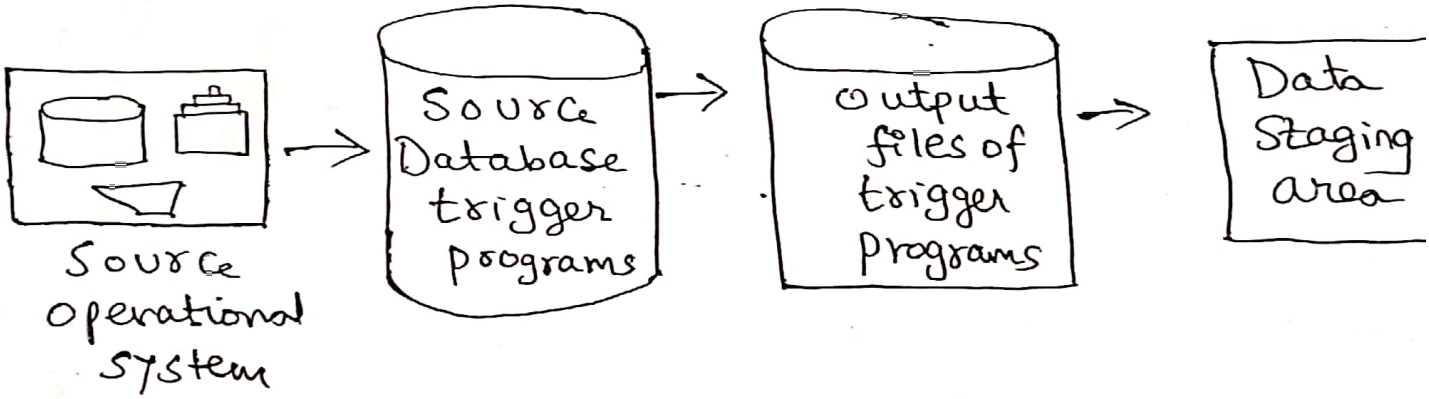
(2b) Capture by Comparing files

It is also known as Snapshot differentia technique because it Compares two Snapshots of the Source data. It necessitates the keeping of prior Copies of all relevant Sou data.

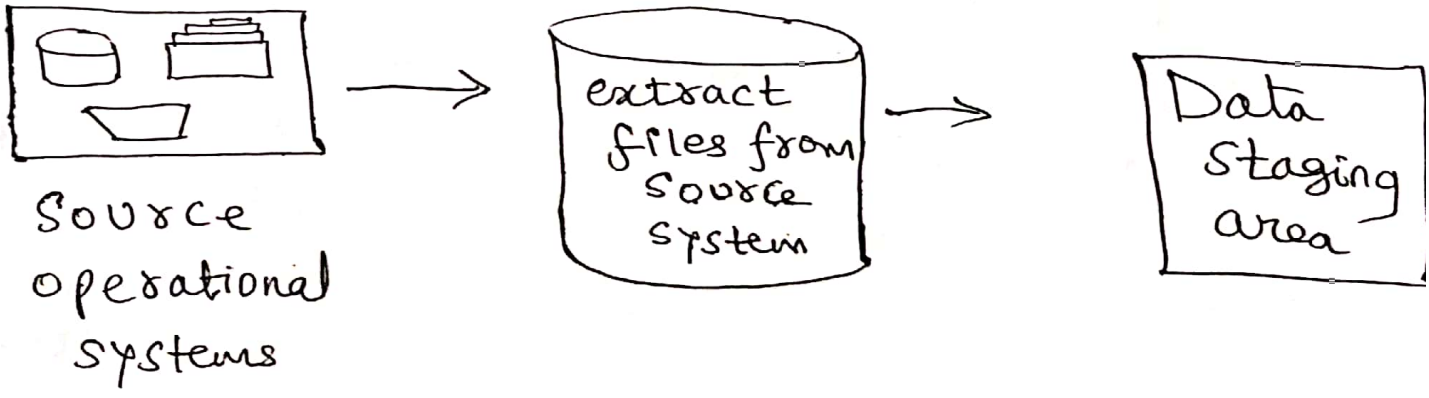
Capture through transactional logs:-



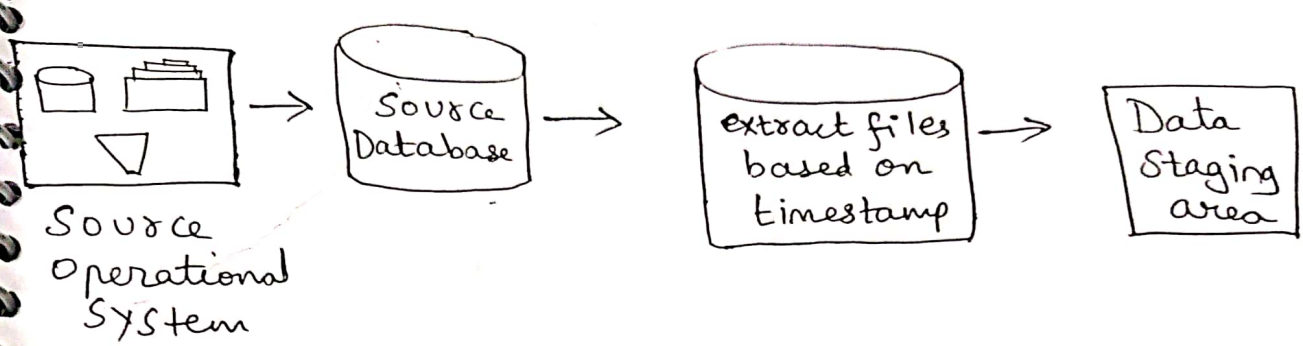
Capture through Database triggers



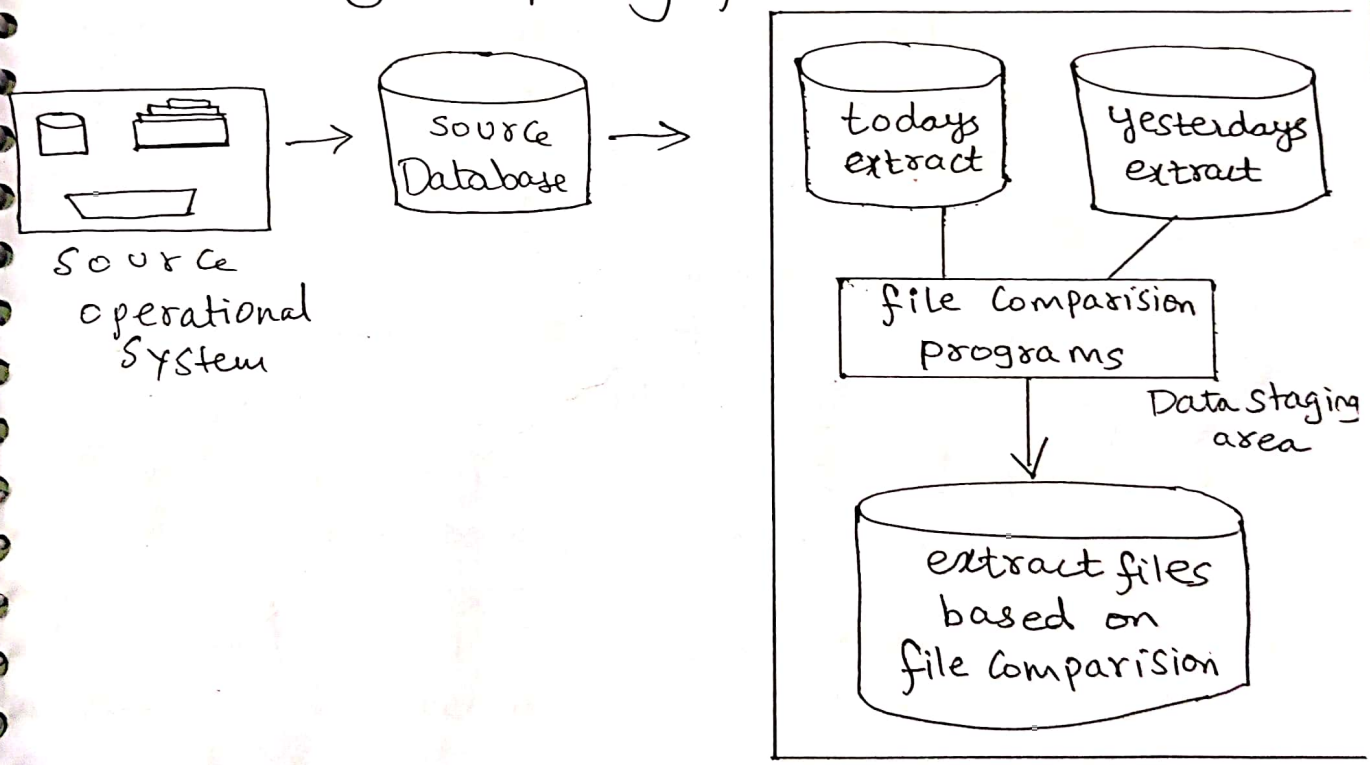
Capture in Source Application



Capture based on Date & timestamp



Capture by Comparing files



Parameter	OLTP
Basic	It is online transactional system & manages database modification
Focus	insert, update, delete information from the database
Data	OLTP and its transactions are the original source of data
Transaction	OLTP has short transactions
Time	The processing time of a transaction is comparatively less in OLTP.
Example	ATM

Q) Explain the data tranformation tasks in etL

- The tranformation process deals with rectifyi-ng the inconsistency. Improving the quality of data forms a major task within the data tranformation process. It takes the following covi

① Map the input data from source system to data for DW repository.

② Clean the data, fill all the missing values by some default value.

③ Remove duplicate records. perform splitting and merging of fields. Sort the records

④ De-normalise the extracted data accordi-ng to dimentional model of DW.

⑤ Convert to appropriate data types. Perform aggregations & Summarisations.

The transformation tasks which are commonly performed on extracted data are:-

Ⓐ Format Revision - These include changes to data types and lengths of individual data fields.

Ⓑ Decoding of fields:- when data comes from multiple source systems, the same data items may have been described by different field values.

Ⓒ Splitting of fields:- Earlier legacy syste stored names & addressess in large text field. All components of Name-

- first, middle & last were stored in one large field called "Names". But the need today is to split the individual component into separate field so that operating performance can be improved by indexing on the individual components.

(D) Character set Conversion:- This type of data tranformation is done to textual data to convert its character set to an agreed standard character set.

(E) Conversion of units:- Many companies may have global branches. So the Sales amount may be represented in different currencies in different source Systems. Before moving the data in DW you need to convert the figure into a common unit of measurement.

(F) Date & time conversion:- The date and time format values also needs to be represented in a standard format.

(G) De-duplication - Sometimes, the record for the same customer may be stored in many files. When you extract data from source systems, you have to pay special attention to find such duplicate records & remove duplicates while storing record in DW, ensuring that the information about the customers is stored only once forming a single record.

Q) Q) With neat diagram, write a note on ODBC in etl process.

- Open database connectivity (ODBC) was created to enable users to access Databases from their windows applications. The original intention for ODBC was to make applications portable, that is if an application's underlying Database changed- Say from DB2 to oracle - the application layer did not need to be recorded & Compiled to accomodate the change. Instead you simply change the ODBC driver which is transparent to the application. You can obtain ODBC drivers for practically every DBMS in existence on virtually any platform. you can also use ODBC to access flat files.

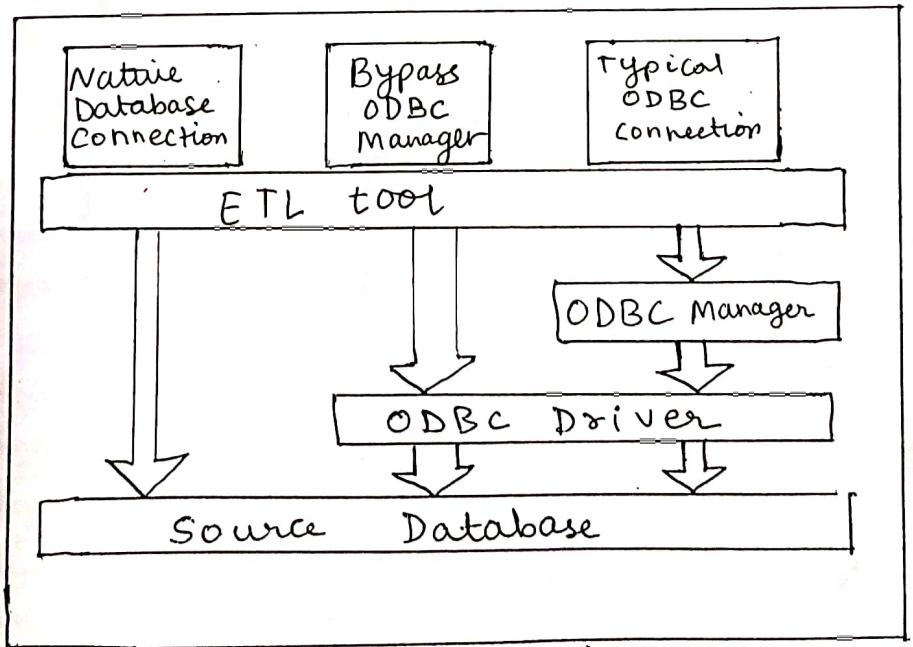


fig:- topology of ODBC in ETL process

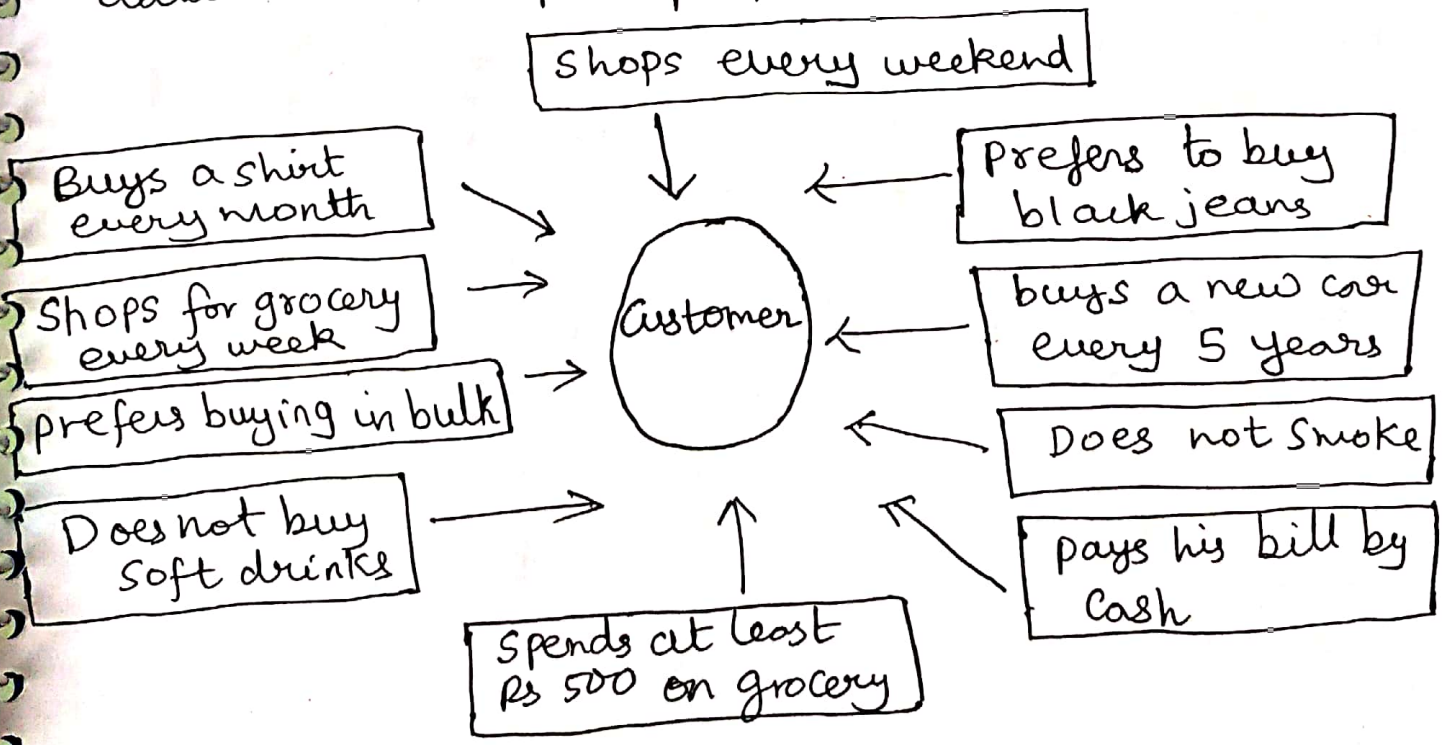
ODBC manager :- The ODBC manager is a program that accepts SQL from the ETL application and routes it to the appropriate ODBC driver. It also maintains connection between application and the ODBC driver.

ODBC driver :- The ODBC driver is the real workhouse in the ODBC environment. The ODBC driver translates ODBC SQL to the native SQL of the underlying database.

The drawback to ODBC's flexibility is that it comes at a performance cost. ODBC adds several layers of processing and passing of data to the data-manipulation process. For the ETL process to utilise data via ODBC, two layers are added between the ETL system and the underlying database.

Q) Explain the operational data store (ODS) with respect to nature of data, underlying technology, profile records & classes.

- **Nature of Data:-** In striking Constrast with the DW, an ODS contains very limited amount of historical data. while the data warehouse stores 5-10 years of historical data, the ODS stores only a months old data.
- **Underlying Technology:-** The ODS is designed using a hybrid approach since a part of it is designed using relational technology and the rest is designed using multi-dimensional technology.
- **Profile records:-** A profile record is one that is formed from many observations about an entity. A profile record creates a synopsis from multiple occurences of data. A sample profile is shown below:-



- profile record captures massive amount of data very concisely. Once the information is captured in profile records, it can be easily and quickly accessed as when the need arises.

Classes of ODS :- An ODS is categorised into 4 classes depending upon how fast the data arrives into it.

Class I :- It takes a few milliseconds for the data to arrive in ODS. The time elapsed is transparent to the users. Application - Applied in airline reservation system.

Class II :- It takes several hours for the data to arrive in ODS, once a transaction takes place. The end users can visualise that there is a time gap between arrival & occurence of transaction. Application - Applied to update name and address change of a customer.

Class III :- There is an overnight gap or longer between occurence of transaction and the arrival of data in ODS. Application - used for applying Sales transaction.

Class IV :- The time gap between occurence of transaction and its arrival into ODS is much longer, often in some months or years. The source of data can be a DW or some other. May be created from output of special reports or projects. Application - A Survey of customer buying habits.

Q) Write a detailed note on SCD

- Slowly Changing dimensions (SCD) is the term used for managing issues associated with the impact of changes to attributes of a dimension table.
- Design approaches that deal with the issue of SCD are categorised into three change types:-
  - TYPE1: overwrite the dimension record
  - TYPE2: Add a new dimension record
  - TYPE3: Create new fields in dimension record

Type1 Changes :- correction of errors

- It relates to correction of errors in source system. ex:- Suppose a spelling error in a Customer name is corrected to read as "John Michael" from an incorrect entry of "John Michel".
- There is no need to preserve old values. Since the old name is incorrect, it must be discarded.
  - Figure 1A&B: shows the method for applying "Type1" Change.
    - ✓ Overwrite the value of attribute with new value in dimension table row.
    - ✓ The old value is discarded, that is not preserved.
    - ✓ No other changes are made in the dimension table row
    - ✓ The Key of this dimension table row is NOT affected.

Customer Key	Customer ID	Customer name	Marital Status	Address
2252134	C-123	John Michael	Single	XYZ

fig 1A Type 1 Change

TYPE 2 changes:- preservation of history

-These changes relate to true changes in source system. The history must be preserve in DW. ex:- A customer Jenny David whose marital status has changed from 16 Jan 2006 all orders from Jenny David before that date must be included under marital status "single" and all orders after 16 Jan 2006 should be included under marital status "married".

- Figure 2 shows method of applying TYPE 2 changes.
- ✓ A new dimension table row with new value of changed attribute is added.
- ✓ A new column called "effective date" is added in dimension table & the original row is not changed. The key of original row remains same.
- ✓ A new row is inserted with a new key in dimension table.

Customer key	Customer ID	Customer name	marital status	Address
2252134	C-123	John michel	Single	XYZ

fig 1B: Type 1 change

fig 2: Type 2 change

Customer key	Customer ID	Customer name	marital status	Address
2252134	C-123	Jenny David	Single	XYZ

Customer key	Customer ID	Customer name	Marital status	Address	EFFECT DATE
2252134	C-123	Jenny David	Single	XYZ	1 mar 2003
111234	C-123	Jenny David	Married	ABC	16 Jan 200

TYPE 3:- Tentative Soft revisions ✓

- ✓ They are used to compare performance accross transition.
- ✓ They are used when there is a need to track history with both old & new values of the same attribute.
- ✓ An "old" field is added in dimension table for affected attribute.

✓ The existing value of attribute is "pushed down" from "current field" to "old field".

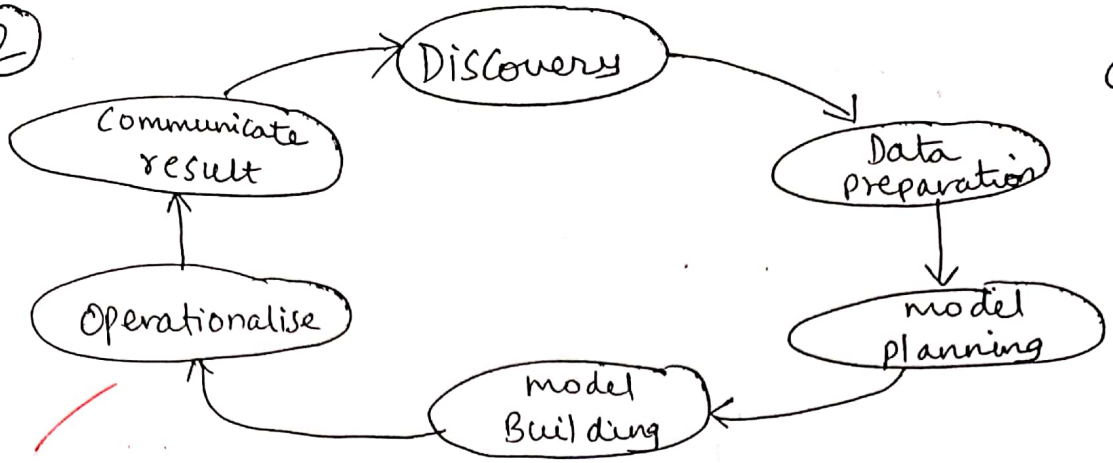
✓ The existing queries will automatically switch to current values.

Salesperson Key	Sales-person ID	Sales-person name	Old location	Current location	Effective date
2252134	S-001	Jenny David	—	Delhi	1 mar 2003

After applying TYPE 3 change

Salesperson Key	Sales-person ID	Sales-person name	Old location	Current location	Effective date
2252134	S-001	Jenny David	Delhi	Mumbai	16 Jan 200

27 (2)



(9) Lifecycle of DS problem

Discovery :- Involves acquiring data from all internal and external sources that can answer the business question. Data can be in any form like sheets, files et

Data preparation :- Data can have a lot of inconsistencies like missing values, blank columns, incorrect data, abrupt values which needs to be cleaned. It is required to explore & process data prior to modelling.

Model planning :- Here we determine the method or technique to draw relationship between variables. We use visualisation techniques like histogram, box plots, graphs to get a fair idea of distribution of data.

Model building :- Develop data sets for training & testing purpose. Analyse various techniques like classification, clustering, association to built model.

Operationalise :- Make the final report, briefing code and technical document.

Communicate result :- T

Q) a) Explain the various Sources of pollution of data.

- System Conversion - System conversion and migrations are prominent reasons for data pollution. ex:- Conversion from flat files to hierarchical database & finally to relational database.
- Data ageing - Older values loose their significar with the passage of time. ex:- product code as a part of product tables primary key may not be relevant and newer application thus want to remove it.
- Heterogeneous System Integration - Heterogeneous and disparate source systems may lead to corrupt or inconsistent data. ex:- if the sour for one table involve several systems like flat files, network databases & hierarchical databases.
- Poor database design - DBMS not providing support for verification of conformance of business rules can pollute the data.
- Incomplete information at data entry Some fields may either have no values for the data items or have N/A or other generic values that may be of no use.
- Fraud - Deliberate attempts to enter incorrect data. ex:-

ex:- The users may not have filled the data about their telephone numbe

or may have filled 1's or 2's in all ten digits

- Fraud - Deliberate attempts to enter incorrect data. ex:- fields containing the units of prod Sold may represent an incorrect value.

- Lack of policies - If there are no prevention rules followed by a company to cater for incorrect and corrupted data, then it would lead to creeping of data quality problems. ex:- If there are no checks to see if the user is filling the value within the specified domain range, then the user may intentionally or unintentionally fill incorrect values.

Draw Phasor diagram to represent conditions in a single phase transformer supplying load at unity pf load.

(A) - DC motor characteristics.

(A) Sketch & explain torque, armature current, Speed & armature current & Speed torque characteristics of DC Series motor.

→ Construction & working principle of 3 phase Induction motor.

→ Draw neat diagram of 3 phase star & delta connection explain line & phase qualities for these connections.

→ Two wattmeter method of 3phase ckt power measurement.

→ (A) construction & working principle of DC motor

→ Compare series & parallel resonance ckt

→ Compare, Series & parallel magnetic ckt

(A) Explain electrical characteristics of DC series motor with derivation to justify their shape.