Module 2

Database:-

→ Stores data in tables
→ Deals with operational or transactional data (current data).
→ Can store MBs to GBs of data.

→ Used for OLTP (Online Transaction Processing)

Data Warehouse

→ Stores huge amounts of data
→ Data collected from multiple heterogeneous sources like files, DBMS, etc.

→ Stores historical data
→ Can store TBs of data

Eg How the placement of CS students have improved over the last 10 years, in terms of salaries, counts, etc.

Used for OLAP (Online Analytical Processing)

-> A data warehouse is a system that stores data from a company's operational databases as well as external sources

-> Data warehouse platforms are different from operational database because they store historical info making it easier for business leaders to analyze data over a specific period of time.

-> According to Bill Inmon - "A ~~data~~ warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process.

Data Warehouse Characteristics

i) Subject Oriented
ii) Integrated
iii) Time Variant
iv) Non Volatile

Subject Oriented :-

Data warehouse provides a consise view around a particular subject, such as customer, product or sales instead of

The global organization's ongoing operations

→ A DW is always a subject-oriented one, as it always provide info about a specific theme

ii Integrated:

→ A data warehouse integrates data from various heterogenous data source like RDBMS, flat file and online transaction records and combines it in a relational database

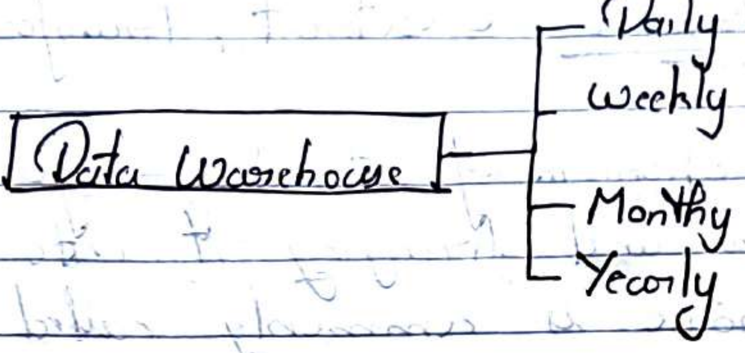
→ It requires data cleaning and integration during data warehousing to ensure consistency among different data sources

→ It must be consistent, readable and coded.

iii) Time Variant

→ Historical info is kept in DW.

→ One can retrieve files from 3 month, 6 months or 12 months or even previous data from a DW.



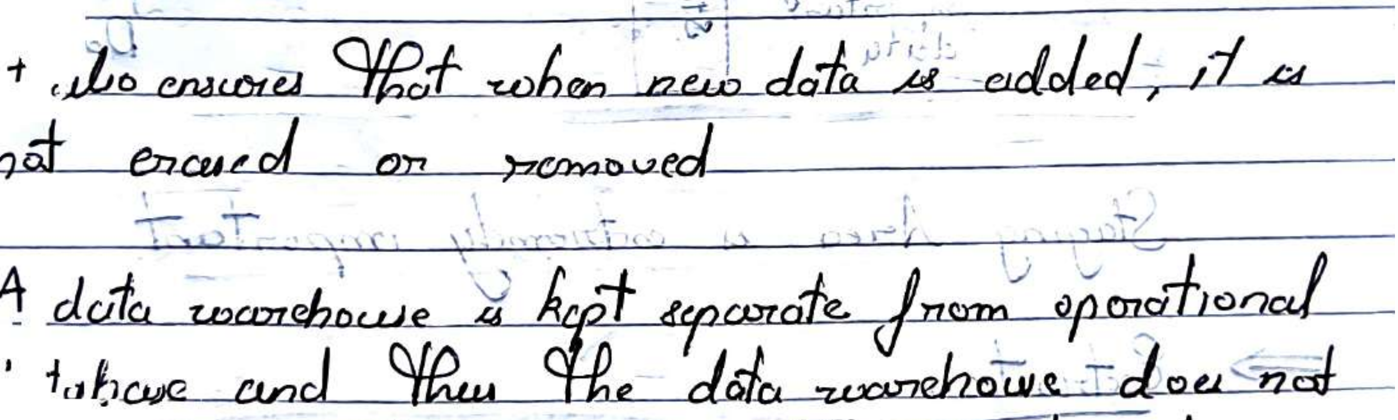
Non-Volatile :-

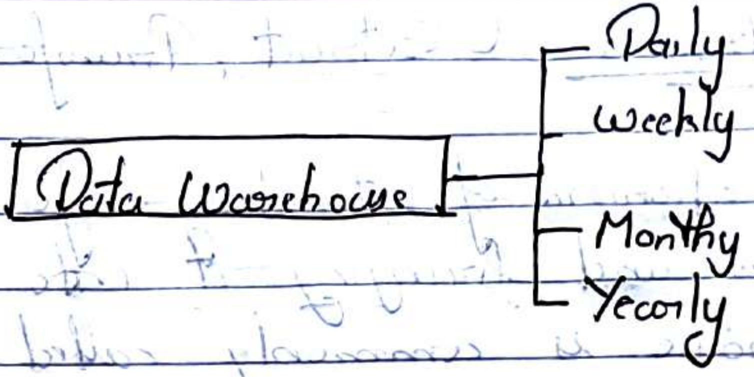
The data residing in data, warehouse is permanant

+ also ensures that when new data is added, it is not erased or removed

-> A data warehouse is kept separate from operational tabase and thus the data warehouse does not + regular change in the operational

[redacted]



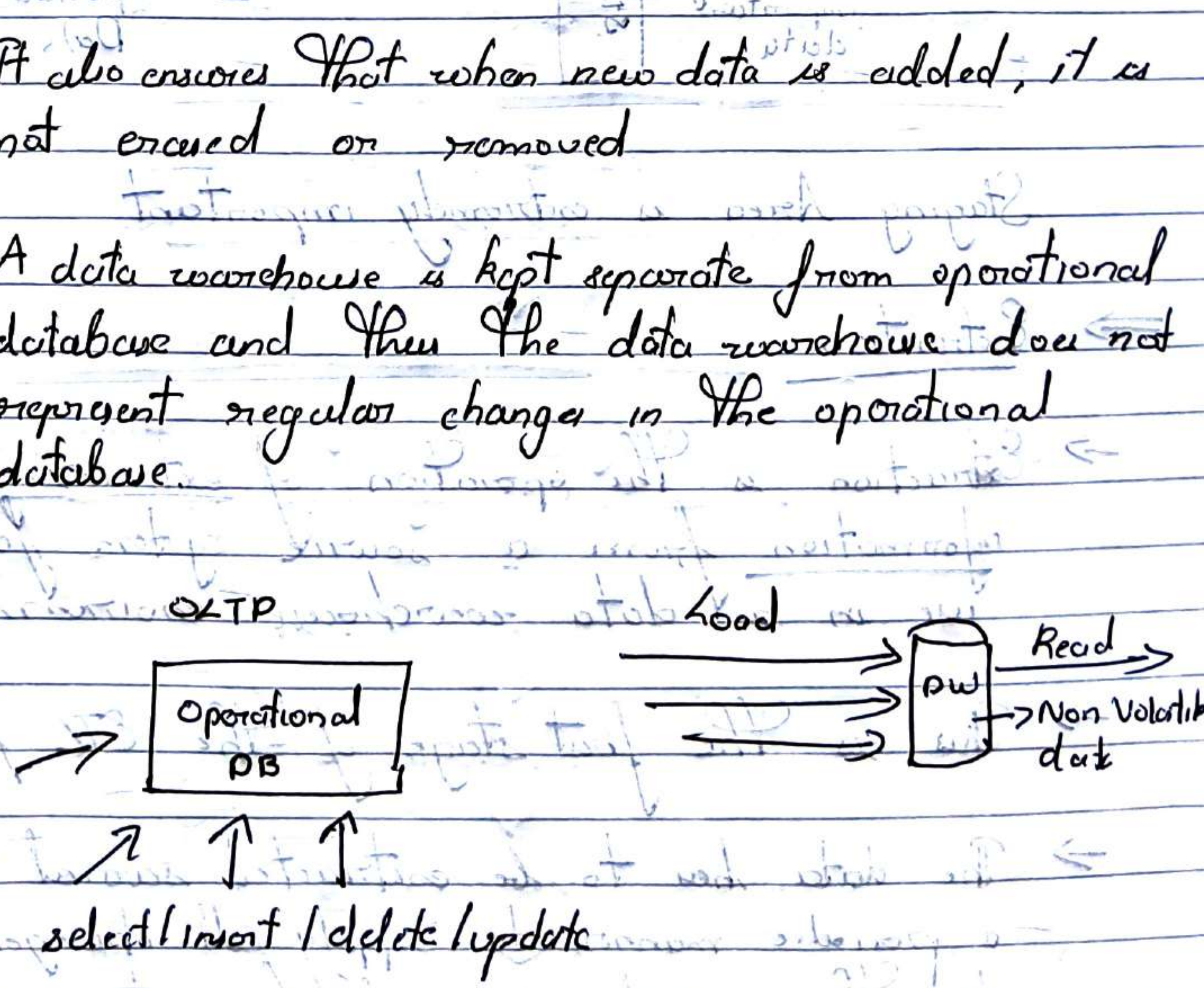


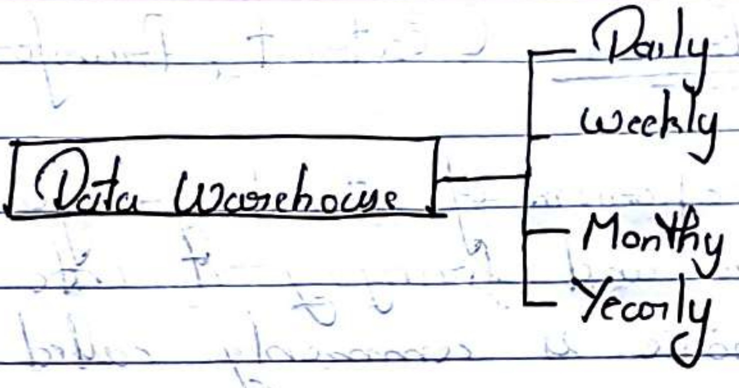
iv) Non-Volatile :-

→ The data residing in data warehouse is permanent

→ It also ensures that when new data is added, it is not erased or removed

→ A data warehouse is kept separate from operational database and thus the data warehouse does not represent regular change in the operational database.



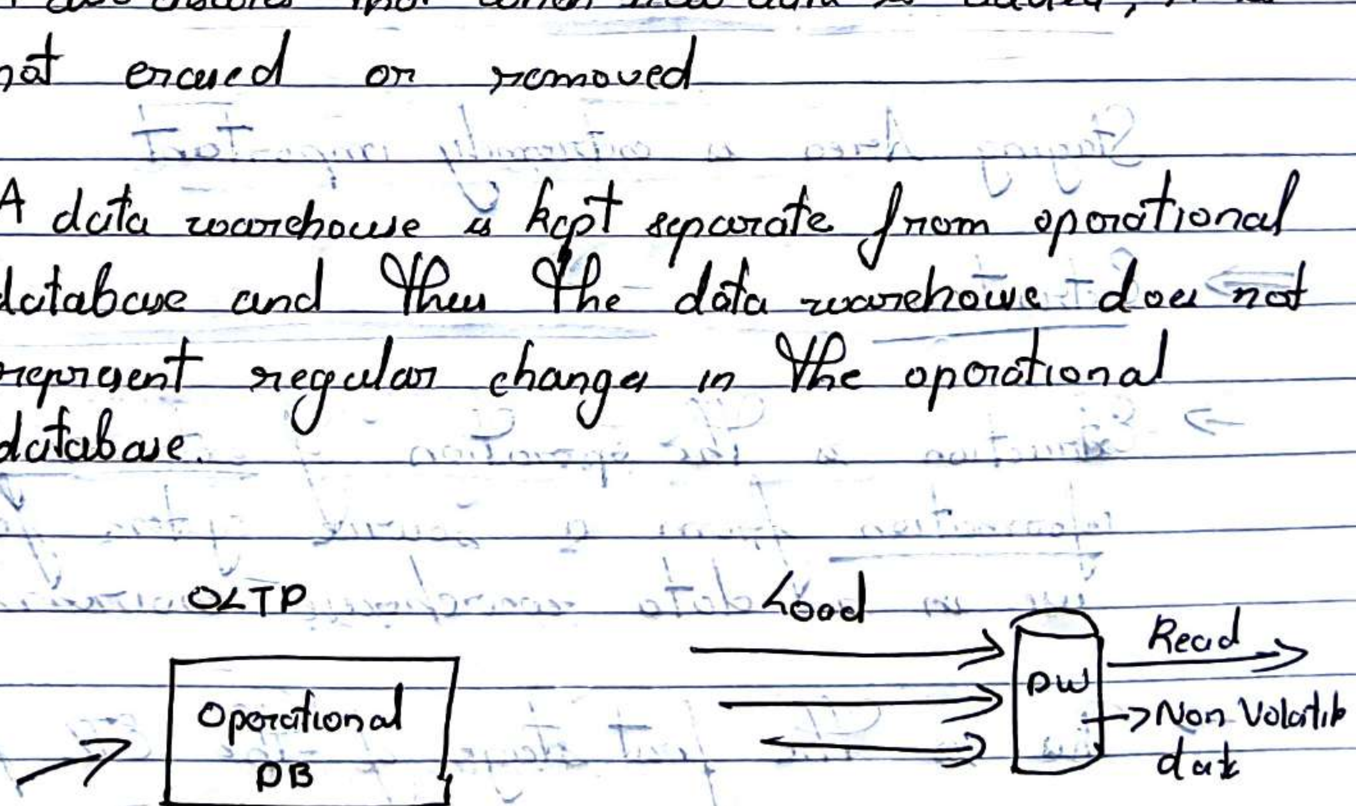


(iv) Non-Volatile :-

-> The data residing in data warehouse is permanant

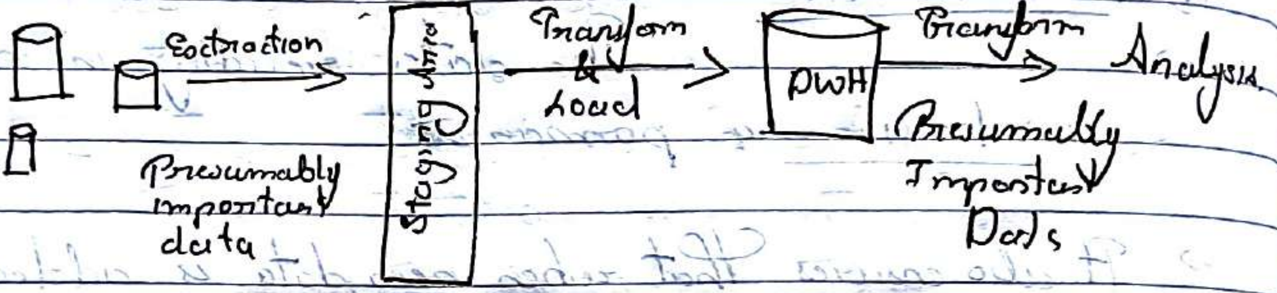
-> It also ensures that when new data is added, it is not erased or removed

-> A data warehouse is kept separate from operational database and thus the data warehouse does not represent regular changes in the operational database.



ETL Process (Extract, Transform and Load)

The mechanism of extracting info from source systems and bringing it into the data warehouse is commonly called ETL.



Staging Area is extremely important

⟹ Extraction

→ Extraction is the operation of extracting information from a source system for further use in a data warehouse environment.

→ This is the first stage of the ETL process

→ The data has to be extracted several times in a periodic manner to supply all changed data to the warehouse and keep it up-to-date

⇒ Cleansing:

→ The clensing stage is crucial in data warehouse technique because it is supposed to improve data quanlity.

→ Primary data cleansing features found in ETL tools are rectification and homogenization.

→ They use specific dictionaries to rectify typing mistakes and to recognize synonyms and defines approropriate associations between the values.

⇒ Transformation:-

→ Transformation converts records from its operational source format into a particular data warehouse format.

→ i.e data extracted from server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed.

→ It adds value and changes data such that insightful BI reports can be generated.

⇒ <u>Loading :-</u>

→ Loading data into the target datawarehouse is the last step of the ETL process.

→ Huge volume of data needs to be loaded in a relatively short period

→ Hence, load process should be <u>optimized</u> for performance.

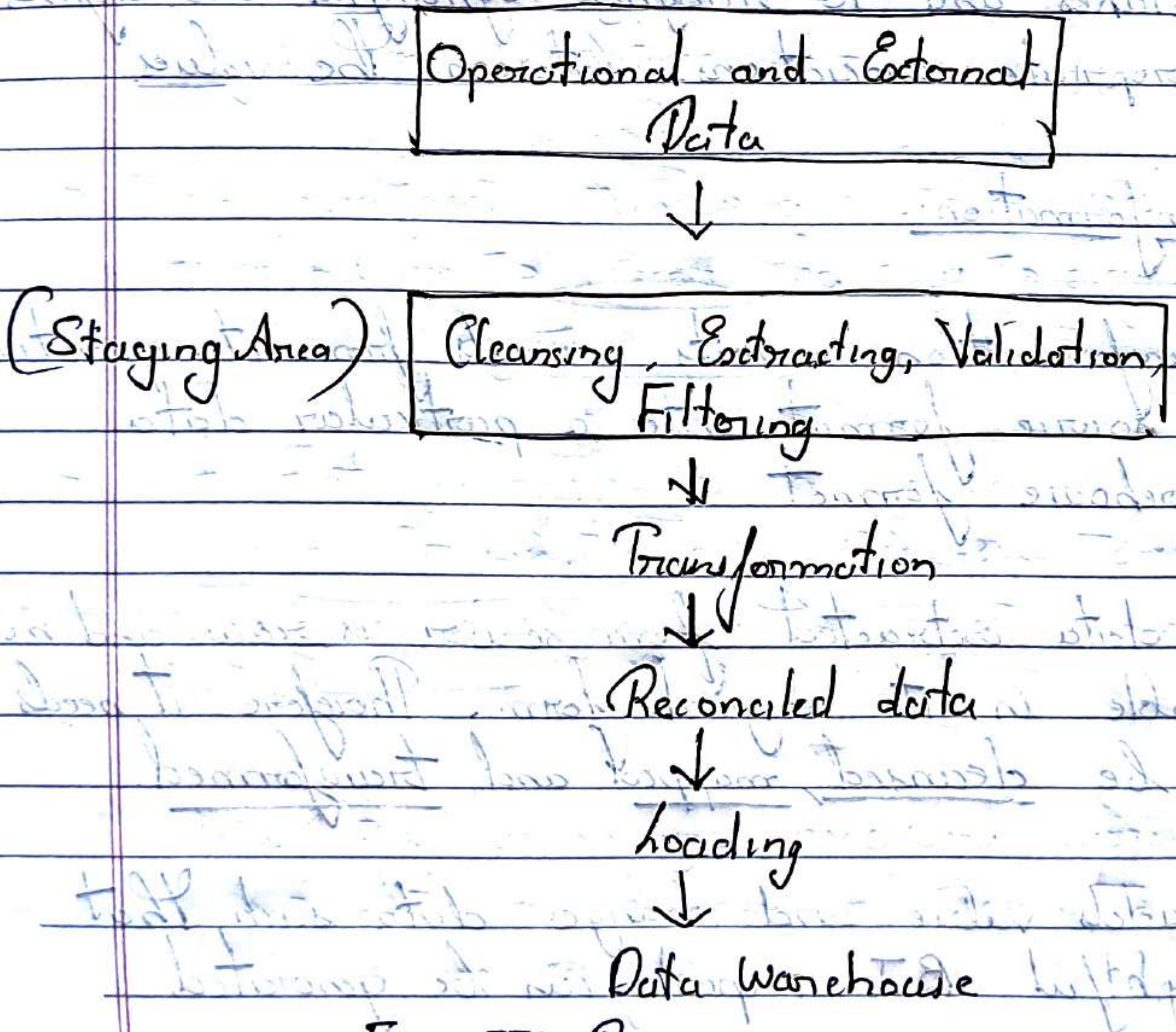

Operational and External Data
(Staging Area)
Cleansing, Extracting, Validation, Filtering
Transformation
Reconciled data
Loading
Data Warehouse

Fig: ETL Process

Components or Building Blocks of Data Warehouse

1 Data Source Component

↳ Internal Data // Refer and Expand
↳ Archived Data
↳ External Data.

2 Data Staging Component:-

↳ After we extract data we have to prepare the files for storing in data warehouse.

↳ Data Extraction // Refer and Expand.
↳ Data Transformation
↳ Data Loading

3 Data Storage Component:-

↳ Data storage component of the data warehouse store the data. It's advantage includes the ability to store large amounts of data in a single location, fast and efficient data retrieval, and improved data quality due to data cleansing and standardization.

Essay Data Warehouse Architecture:-

→ Data warehouse architecture defines the arrangement of the data in different databases.

Different Architecture

→ Single-tier architecture
→ Two-tier architecture
→ Three tier architecture.

Single Tier Architecture :-

→ An operational system is a method used in data warehousing to process the day to day transaction of an organization.

→ A Flat file system is a system of files in which transactional data is stored and every file in the system must have different names.

→ Meta Data summarizes necessary information about data which can be used to access data more easily.

→ End User access tools provide information to the business managers for strategic decision making.

→ The various end user tools include
- i Reporting and Query Tools
- ii Application Development Tools
- iii Executive Information Tools
- iv Online Anlytical processing Tools (OLAP Tools)
- v Data Mining Tools

→ A single tier architecture helps to minimize the amount of data stored to reach the goal i.e it removes data redundancy.

Two Tier Architecture

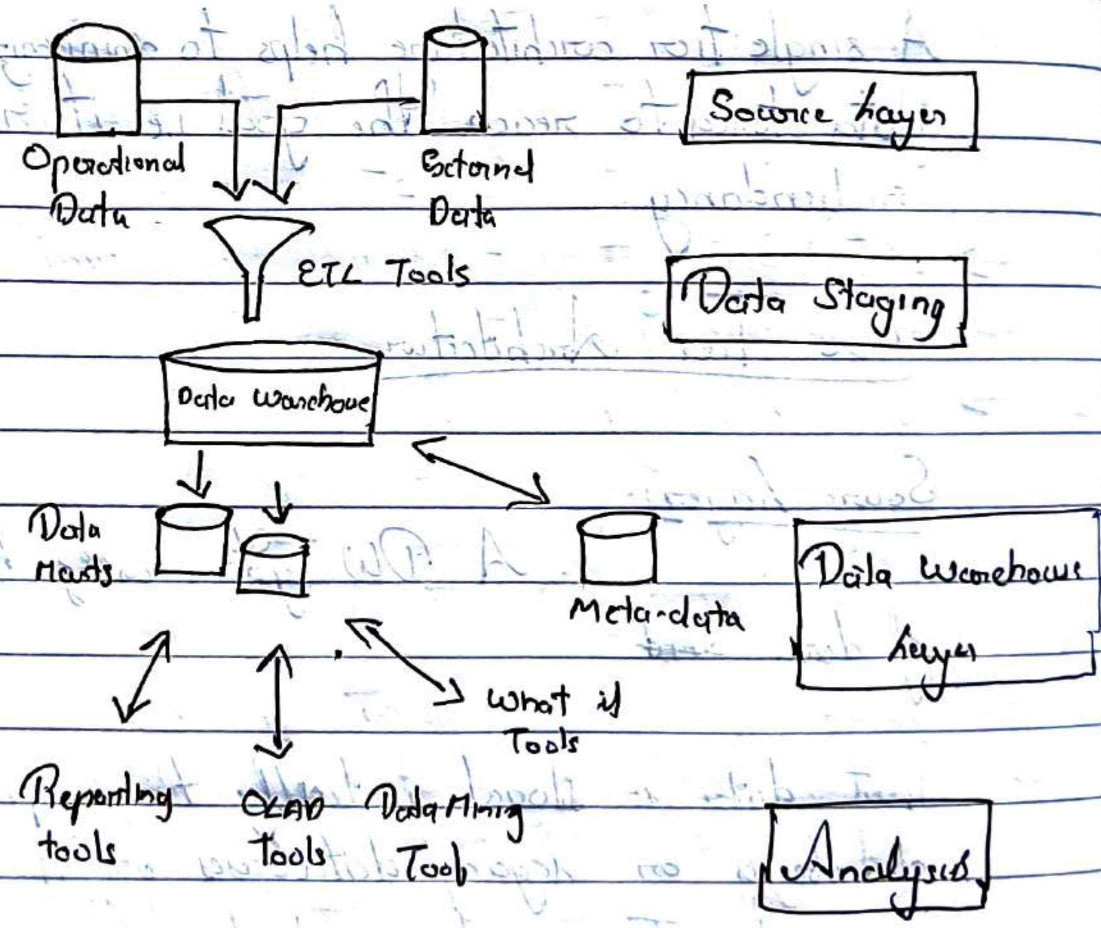
Source Layer:- A DW system uses heterogenous source of data ~~and~~

> That data is stored initially to corporate relational databases or legacy databases or it may come from an information system outside the corporate walls

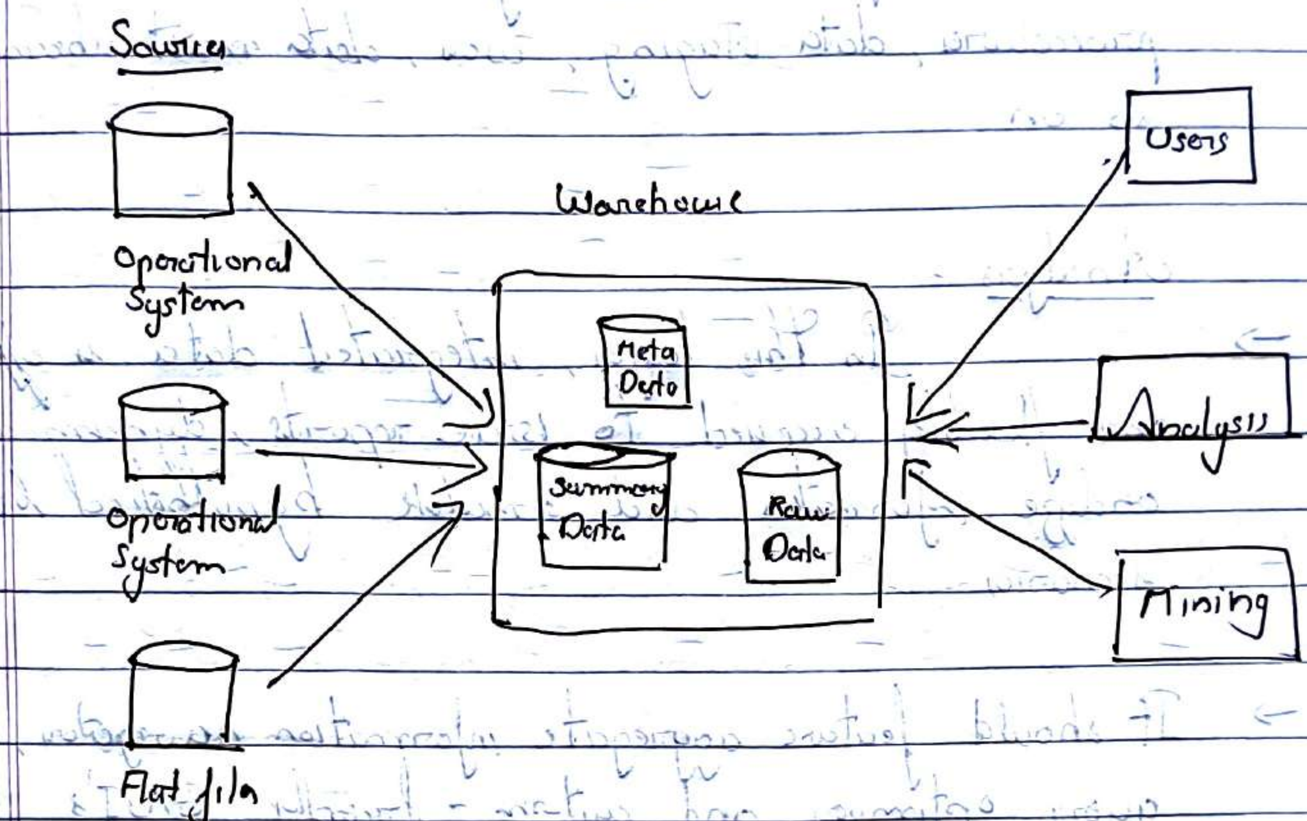
: Data Staging :-

→ The data stored to the source should be extracted, cleaned to remove inconsistencies and fill gaps and integrated to merge heterogenous sources into one standard schema.

→ Extraction, Transformation and Loading Tools (ETL) can combine heterogenous schema, extract and transform, cleanse, validate, filter and load source data into a data warehouse



Single Tier Diagram :-



Two ~~Three Tier Architecture~~ :- Two Tier (Contd...)

Data Warehouse Layer:- Information is saved to centralized individual repository i.e the data warehouse.

-> The data warehouse can be directly accessed but it can be also used as a source for creating data marts, which partially replicate data warehouse contents and are designed for specific enterprise departments.

→ Meta-data repositories store information on sources, access procedures, data staging, users, data mart schema, and so on.

Analysis :-

→ In this layer, integrated data is efficiently and flexibly accessed to issue reports, dynamically analyze information and simulate hypothetical business scenarios

→ It should feature aggregate information navigators, complex query optimizers and customer-friendly GUI's