UNIT- II

Types of digital data

Digital data

- repr other forms of data using specific language, then it can be interpreted by various technologies.
- Eg: binary system.

↓

It stores audio, video, doc & text info in the form of binary characters (0 1)

Digital data storage → hard drive

- Mostly offline storage [back up]
- Use Cloud storage. (online)
- server, contains all uploaded data.
- data stored in the form of code.

Different ways to store data

1. **RAM** → Temporary data storage.
   → allows to read data fast.
2. **hard disk** → Permanant data storage.
   → It can be external / Internal.

③ Cloud → we can store our data in cloud. we can access data from cloud anywhere and anytime.

④ Digital data center

- Save space, money & time.
- 24/7 security.
- quickly access the data.

⑤ Digital asset management

- servers.
- stores data in diff location.
- well suited for files & docs.
- Easily share the data to 3rd parties
- Organizes files efficiently.

- Digital data storage should not only help Storage space.
- A chance to quickly retrieve files.

Data

explanations

- set of facts, such as descriptions, observations & numbers.
- Three types.
  (1) Structured data
  (2) un " "
  (3) semi " "

Structured data

- Tabular data.
- repr by rows & columns.
- called relational database.
  ↳ formed set of columns.
- SQL used for structured data.

Eg

Student-id	name	age
1	Jim	20
2	Sam	19
3	Ram	18

sub id	subject	Teach
001	Maths	Joh
002	English	vict
003	Scienc	Bal

Stu-id	sub-id	grade
1	001	90
1	002	100
1	003	92
2	001	95
2	002	80
⋮	⋮	⋮

Semi-structured data

- It does not consist of structured data
- but still some structure on it.
- Contains both structured & unstructured information.
- Common in object-oriented databases.
- Does not follow any tabular data structure

eg
Email XML
CSV JSON → Javascript object notation.

Unstructured data

- Does not have any pre-defined data model.
- data does not organized in pre-defined manner.

eg video, audio, binary data., Image, audio, sensor data.

- Collection of various data types stored in native formats.

eg

Unstructured data

The university has 5600 students. John's ID is 1, he is 18 years old and holds B-sc degree. David's id is 2, he is 32 years old and holds a degree Ph.D. Robert's id is 3, he is 22 years old, he holds BE degree.

Semi_structured data

```
<university>
  <stud_ID="1">
    <NAME>John</NAME>
    <AGE>18</AGE>
    <DEGREE>B.SC</DEGREE>
  </stud>
  <stud_ID="2">
    <NAME>David</NAME>
    <AGE>32</AGE>
    <DEGREE>Ph.D</DEGREE>
  </stud>
  <stud_ID="3">
    <NAME>Robert</NAME>
    :
```

Structured data

ID	Name	age	Degree
1	John	18	B.sc
2	David	32	ph.D
3	Robert	22	BE

Characteristics of structured
- Relational → structured
- Non-relational - unstructured

Data quality

Def

The degree to which data meets a company's expectations of overall quality accuracy, validity, Completeness and Consistancy.

↓ follow same formats.
→ DOB entering → follow specific format
↓
valid.
→ If don't, not valid.

Completeness → Contains all relavent info.

Consistancy → same info stored in multiple sources, Matches equally.

Data quality characteristics / dimensions

- Accuracy.
- Completeness
- Consistancy
- Integrity → Accuracy + Completeness + Consistancy.
- Reasonability → Wheather data meets expected range, type & value
- Timeliness → speed of dissemination
- Deduplication. → eliminate redundant data
- Validity.
- Accessability. ⇒ spreading
  ↳ Make our data easy to get.

Data quality challenges

① Privacy and protection laws

- protection of private data which is accessible by people. (Individual)

↑ → provides some guidelines for

General data protection regulation (GDPR)

& California Consumer privacy act (CCPA)

→ organizations must store individual info securely and should not miss even a fraction of the Collected data.

(2) AI & ML

- Most of the companies uses AI & ML applns in their BI.
- Difficult to keep up the data secured
- Chance to happen mistakes & inaccuracies.
- Monitoring complicated tasks are more challenging.

(3) Data governance practices

↳ DBMS

- It ensures all data is consistant, validity & trustworthy.
- Should provide right data governance
- If not, confusions may occur.

6 ways to determine data quality

(same)

characteristics.

Data classification

(1) Data sensitivity levels.

↳ High sensitivity data
Medium " "
Low " "

High → ~~unauthorized~~ ~~to hackers.~~
→ Include personal info.
→ " Medical / health info

eg: Credit card / debit card / Military id no

→ Confidential.

Medium sensitivity data

- Internal business info & financial info.
- emails, documents

Low Sensitivity data

- Public info
- eg: Public Website Content

② Based on performance.

① Content-based classification

- Classify doc & files → based on reviews

② Content based classification

- based on meta data

③ User based classification

- based on manual judgement

Data states

- There are 4 states
  - * Rest
  - * Process
  - * Transit
  - * Confidential.

Takes less time to classify data

Data format

2 types < structured → DB
unstructured → email, video.
↳ take more time to classify the data.

Data discovery

- collect data from various sources.
- used to understand trends & patterns of data.

Data stored in
- DB
- cloud → Googledoc
- Bigdata platforms.
- files → PDF, email

- Before data classification, we must perform accurate data dictionary.

Create own classification policy

- Create our own classification policies.
- we should consider
  - which person/org/ → owns the info.
  - who is responsible for accuracy of info.
  - where the info stored.

structured	unstructured
1) displayed in rows/columns	cannot display data in table format.
2) Nos, dates, strings	Images, video, audio, emails
3) hold 20% enterprise data	80% enterprise data
4) Requires less storage	Requires more storage
5) Easy to manage	More difficult to manage.

Data warehouse

- RDBMS
- designed for query and analysis data.
- It derived data from various sources
- It cannot perform transaction.
- Mainly used for making decisions.

DW attributes

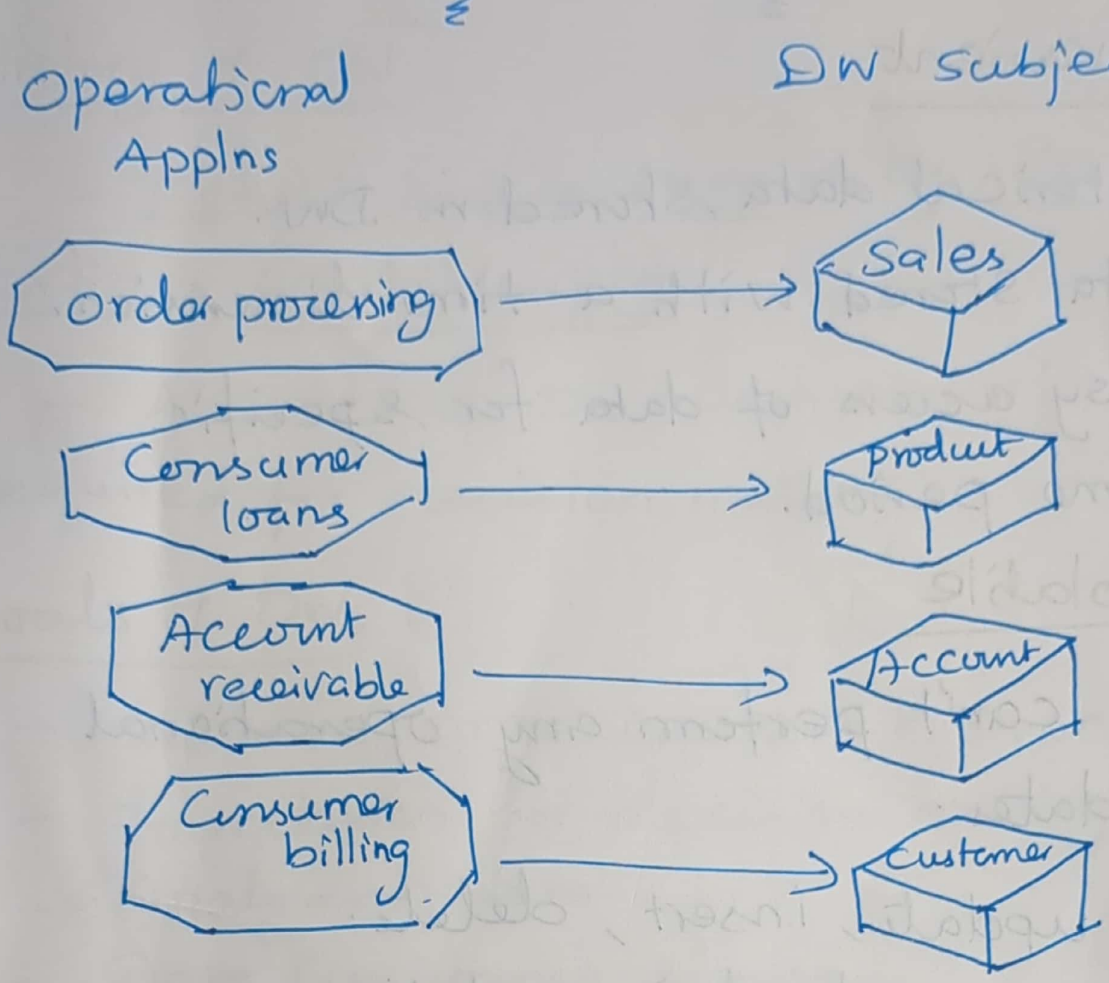
- DB, investigating data from various app
- Support relatively small no. of clients.
- Contains current & historical data.
- Tables.

Characteristics of DW

(1) Subject-oriented.
(2) Integrated
(3) Time-variant
(4) Non-Volatile.

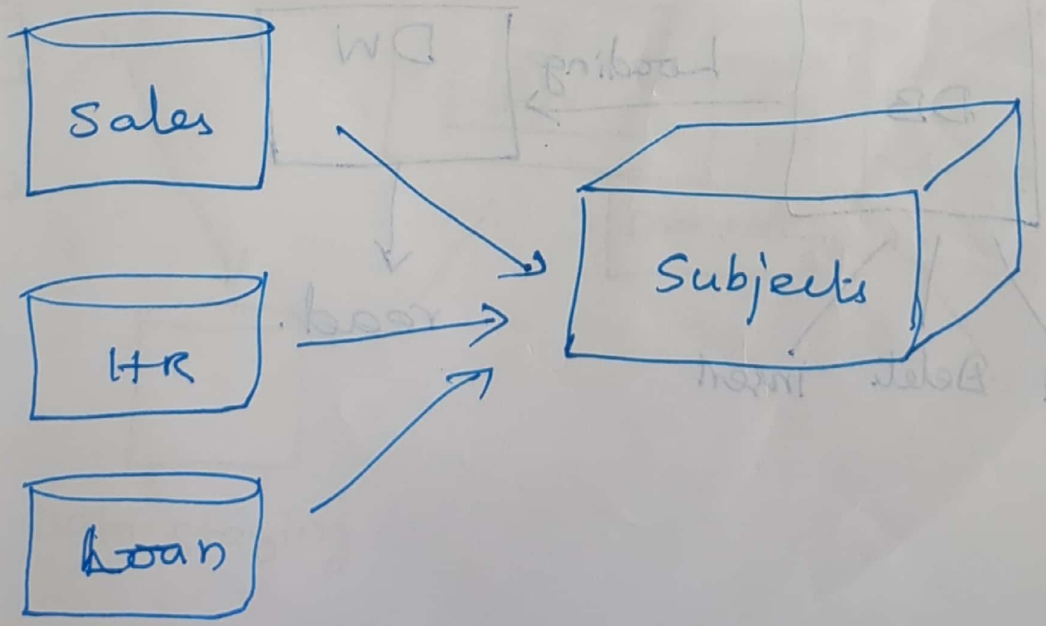
Subject-oriented

- The data is organized in specific subject.
- So, we can access and track the data easily.



Integrated

- DW integrates various data sources like RDBMS, flat files, & online transaction records
- Done data cleaning.



Time variant

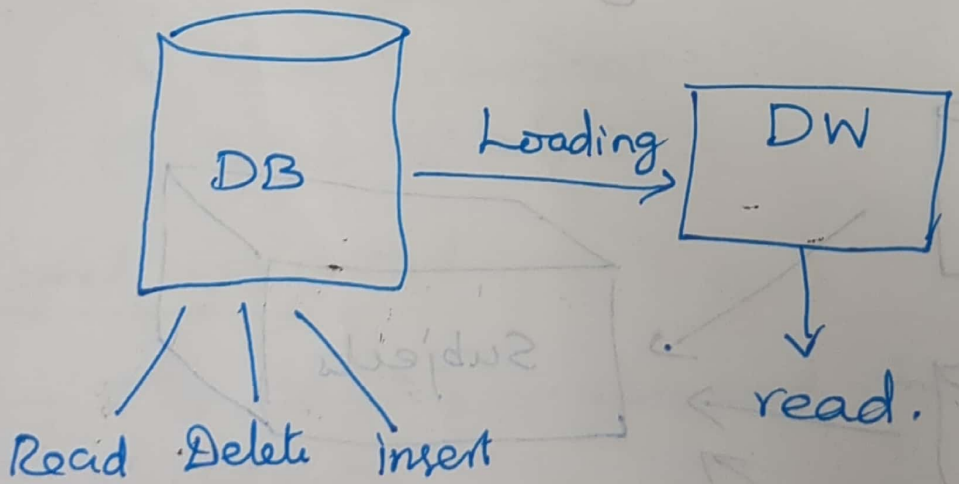
- Historical data stored in DW.
- Data stored with a time dimension.
- Easy acces of data for specific time period.

Non- Volatile

- We can't perform any operational updates.

  update, insert, delete.

- 2 fns < loading of data
  Accessing of data.
- Once the data entered in to the Datawarehouse, data should not change



History of DW

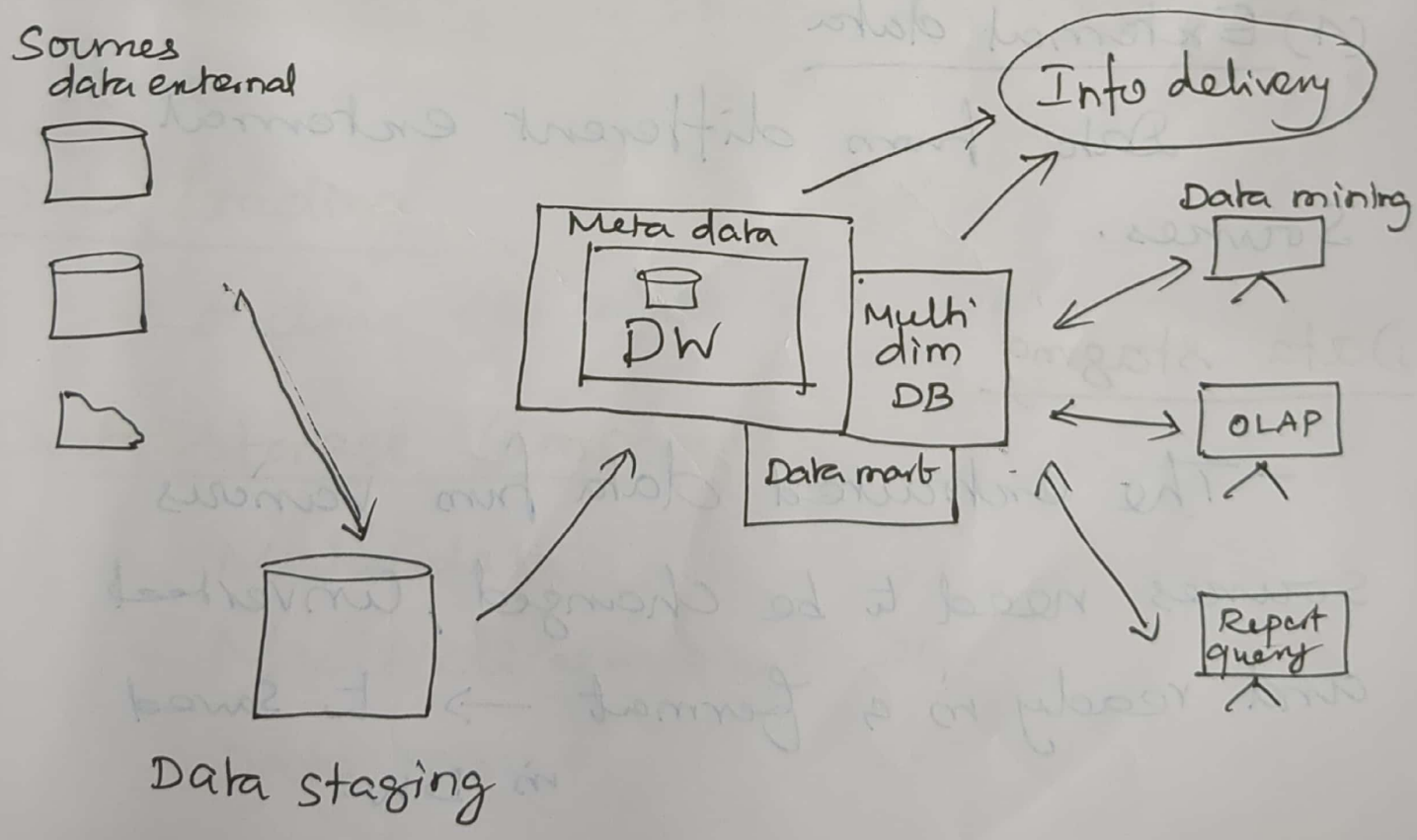
- Established in 1980's by Barry Devlin & paul Murphy

- Absence of DW, vast amt of space required for decision making.

Goals of DW

- Maintain organisation's historical info
- foundation for decision making.
- Analyzing data.
- Data Consistancy & quality.
- High response time.

Components of DW (Building Blocks)



Source data component

4 categories

(1) **Production data**

-data comes from the different OS of the enterprise.

(2) **Internal data**

- Client keeps their private info like spreadsheet, reports, profile, dept info.

(3) **Archieved data**

- Periodically take the old data and store in archived files.

(4) **External data**

Data from different external sources.

Data staging

- The extracted data from various sources need to be changed, converted and ready in a format → to saved in DW.

2 processes

(1) Data Extraction

- appropriate techniques

(2) Data transformation

- Cleaning the data
  - Error data
  - unwanted "
  } → remove
- Standardization
  - Combining pieces of data from various sources.
- Transformation
  - String & merging of data.

Data loading

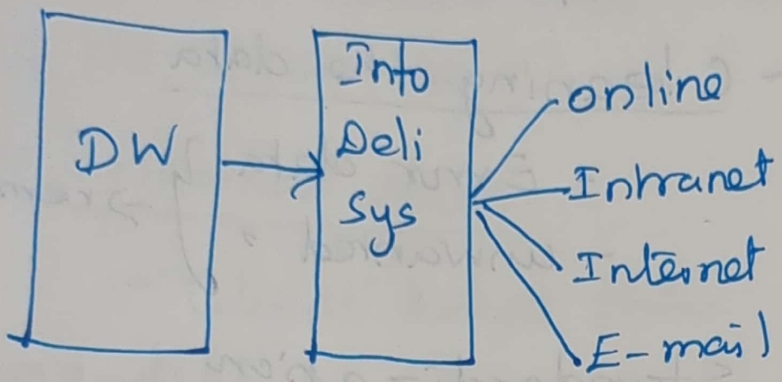
Loading the data into DW.

Data storage Components

- Meta data
- Multidimensional DB.
- Data mart.

Information Delivery Component

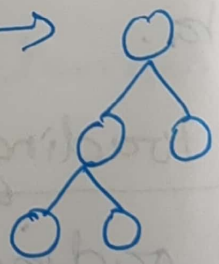
- Process of transferring DW files to one or more destinations.



- private n/w ↓ securely share company info among employees

Meta data

- Data about data.
  - logical structure of data → records addresses.



Data mart

- Contains particular subject.
- Several data marts available in organization.

Management

- Coordinate the services & functions in DW.
- Control data transformation.
- ensure the data correctly saved in repositories.

Diff b/w DB & DW

Database	Data warehouse
① Used for OLTP	Used for OLAP.
② Tables & joins are complicated. Response time – Max ~~Stand~~	Response time → Min
③ Data is dynamic ↓ changes made on data.	data is largely static ↓ fixed data
④ Read operation, Write, update, Modify	~~Write~~ Read operation
⑤ Performance is low	High performance.

Data lake

- When we want to storing big data. 2 options are available.

(1) Data lake

(2) DW.

(1) Types of data

- Structured data
- un " "
- semi " "

Purpose

- Cost effective storage of large amt of data from various sources.
- Allow any structure of data.
- Doesn't need to fit a specific schema

Users

- Data analysts
- Business "
- Data Engineers → Maintain data
- Data Scientist

Task performed by Data lake

- Data engineers use data lakes to store incoming data.
- Perform bigdata analytics.
  ↳ using Apache Spark & Hadoop

Diff b/w data lake & DW

	Data lake	DW
Type of data	unstructured & structured data	Historical data.
purpose	Cost effective big data storage	Analytics for business purpose.
Users	Data scientist's & engineers	Data analyst & business analyst.
Task	- Storing large amt of data. - analyze big data. - deep learning, real-time analytics	Aggregating & Summarizing data.
Size	stores all data &	only stores relavant data.

Business reporting

- Shows business related info like facts, figures & analysis.
- Used to make decisions & plan for the future.

Reporting process involves

(1) Compiling relavant data.
(2) Reviewing info with specific areas.
(3) Analysis.
(4) Drawing conclusions.
(5) Recommendations

Types

(1) Analytical reports
- Used for taking decisions on a business pbm.

(2) Informational reports
- Gives the info abt no. of employees, in which dept they belong, role of employee
- Show report data in diff ways graph, chart, table.

(3) Research report

- plan to launch new product in the market
- plan to expand office area.

Benefits

(1) Adhere to regulation & Competence

- Obey the laws, regulation, guidelines and speitications of the busines organisation.

(2) Transparency

- Provide clear picture for everyone involved in the organization.

(3) Improving efficiency

- Business Managers & owners need to share their experiance to their employees in all the way of taking decision & improving efficiency.

(4) Troubleshooting

- Identify pbm in-time.
- Escalating pbms to higher levels.

(5) Customer ~~focus~~ focus

- Improve the organization on every level and provide better Customer experiance.

Visual Analytics

Data visualization

- showing data in visual format.
- Makes easier to understand the concepts.
- In the form of

  chart, graph, list, maps, dashboards, scure card.

Visual analytics

It does the following jobs

* - Identify & reveal patterns and trends.
  - It prepares data for data visualizati
    * Examine data
    * Align
    * Identify the patterns
    * Gain meaningful insights from complex data sets.

Role of visualization in Analytics

- Data visualization is either Static or interactive

Static → single view of dataset.

Interactive → Various views of same dataset.

Interactive visualization

- Data transformation
- Data understanding
- Business understanding

Types of Analytics

3 types.

(1) Descriptive → simplest type.

- process of using current and historical data to identify trends.

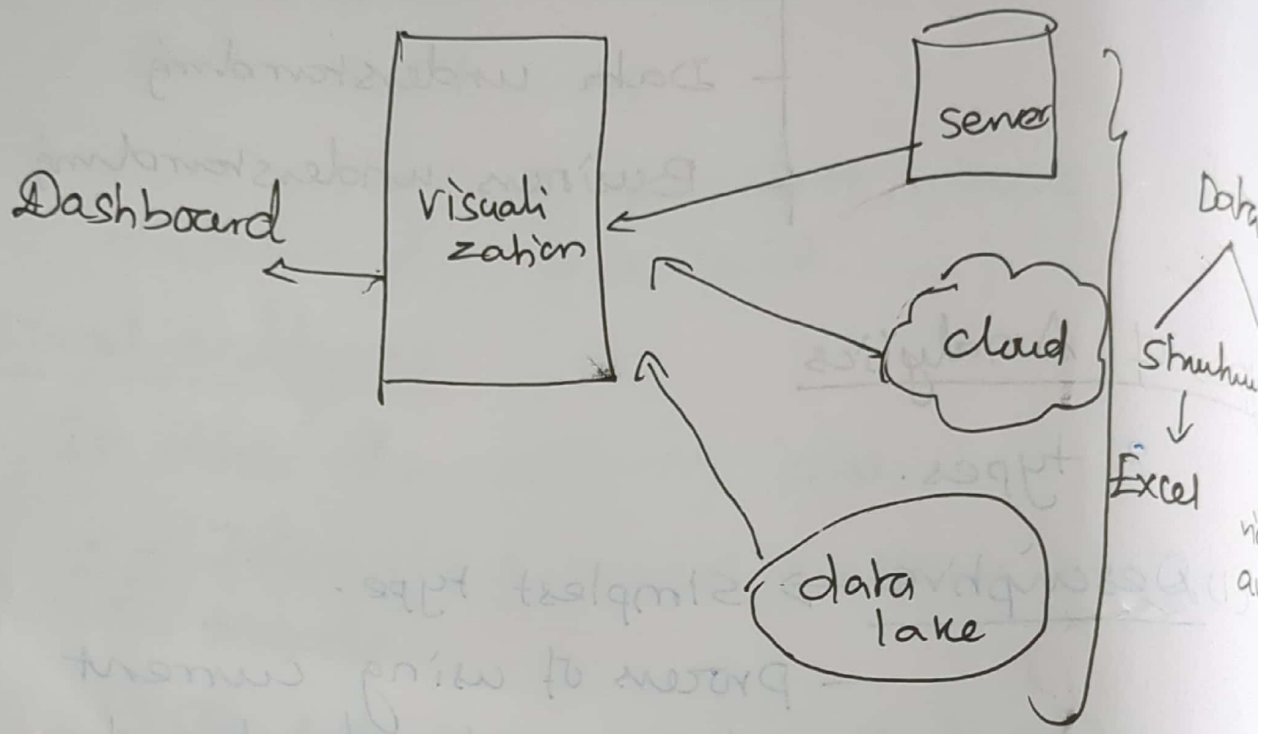
(2) Prescriptive analytics

- Yields recommendations from customers for next step.
- ML algms are used here.

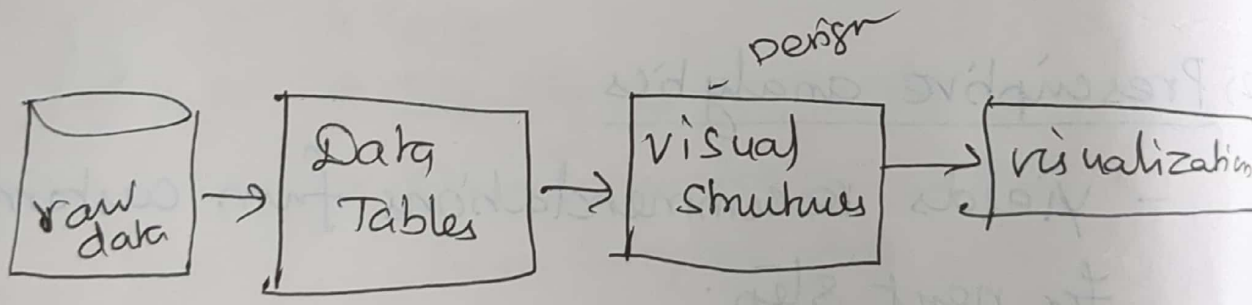
(3) Predictive analytics

- Complex
- analyze
  - * forthcoming scenarios
  - * New trends
  - * Make them efficient & effective

Process



Different types of charts & graphs



Types

* Chart
  - bar chart
  - pie chart
* Table
* Graphs
* Map
* Dashboard

Uses of Chart & graph

- Easy to understand abt particular process.
- Impress stake holders.
- Build trust

How to choose right chart or graph?

1. Identify the goal for presenting the data.
2. Gather data.
3. Based on select graph / chart

Comparing values

* Column chart
* Bar "
* Pie "
* Line "
* Scatter plot

} easily show low & high values in data set.

mixed Composition values

* Pie, * Area * Waterfall, * Mekko

Distributing data

Scatter, pie, bar, Line

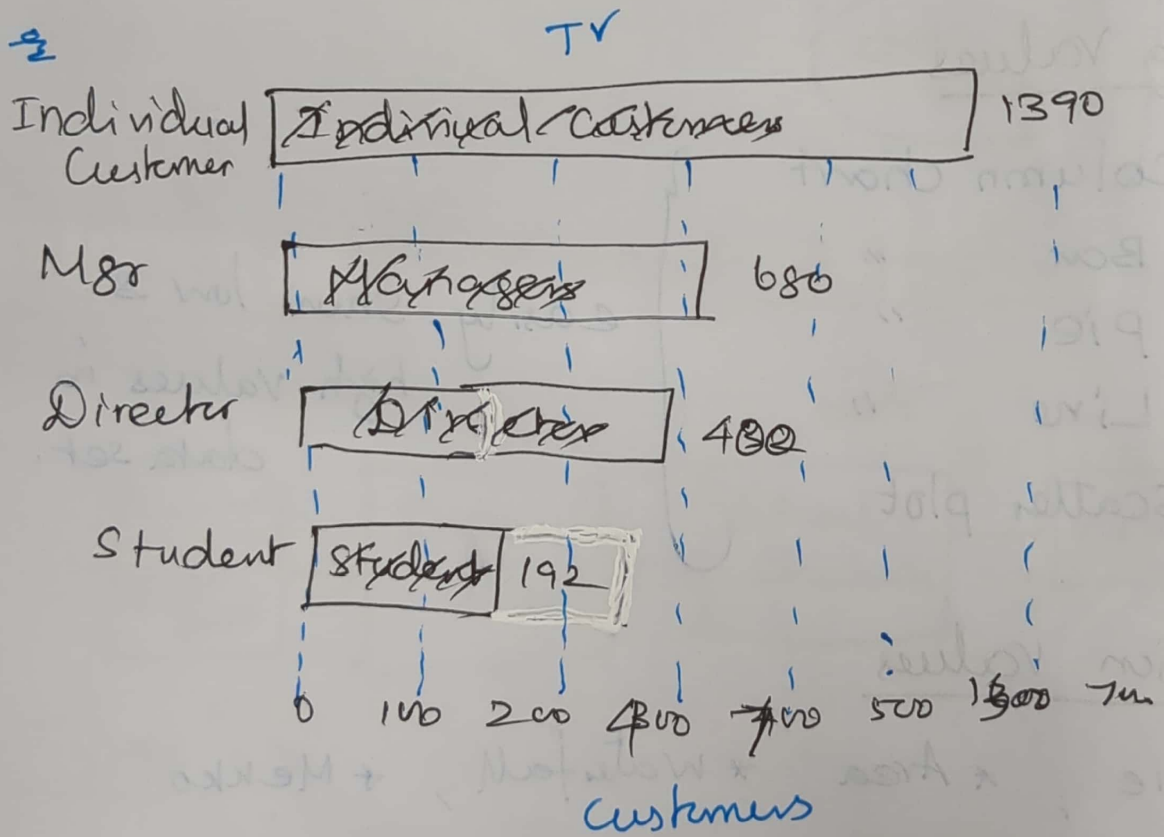
Analyzing trends → Line, column

Relationship b/w data sets → Line, Bubble, scatter plot

① Bar graph

- Helps to compare data b/w diff groups.
  - * Product comparisons
  - * Price "
  - * category "
- Attributes

  Color, Label, axis

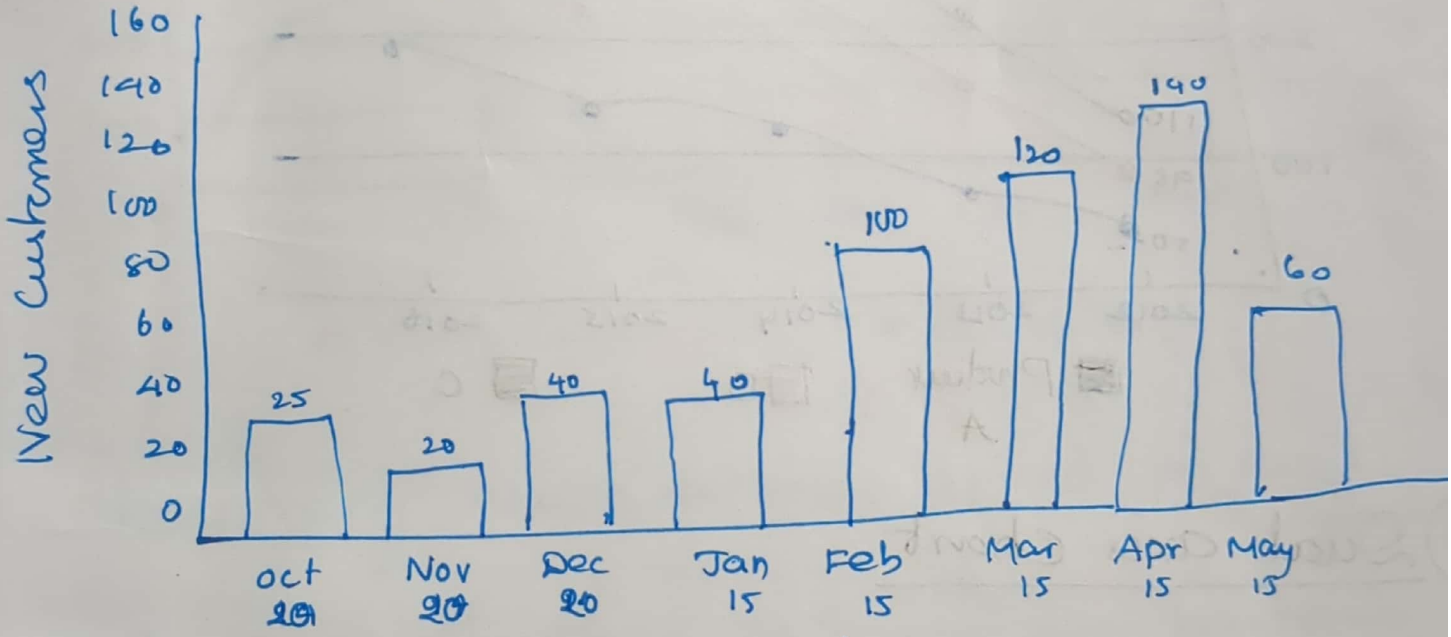


② Column Chart

- Comparison among diff items.
- Show info vertically.

* Customer survey data
* Sales
* Profit & loss

attributes → color, Labels, axis

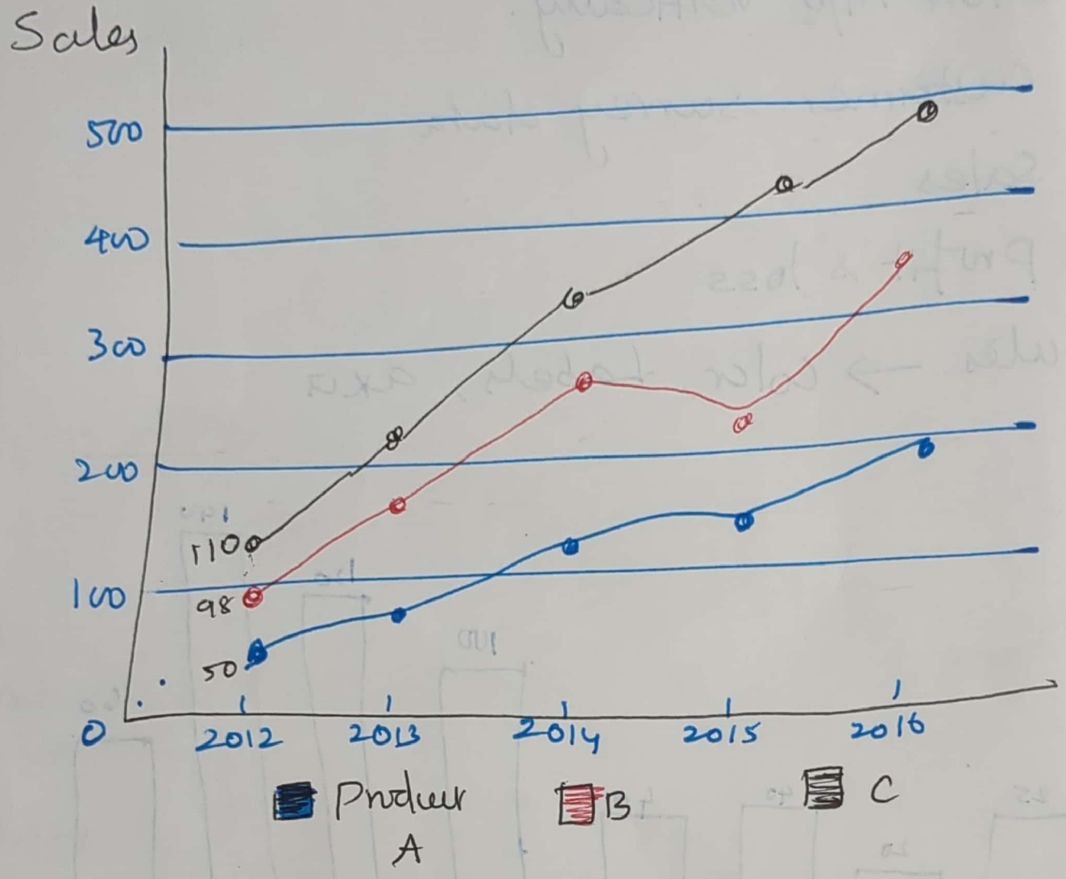


③ Line graph

- Comparison of diff categories of data
- Track changes over short & long period of time.
- Good for seeing small changes.

eg:- Track how many chats / emails → your team responds per month

- Use solid lines
- Don't plot more than 4 lines.



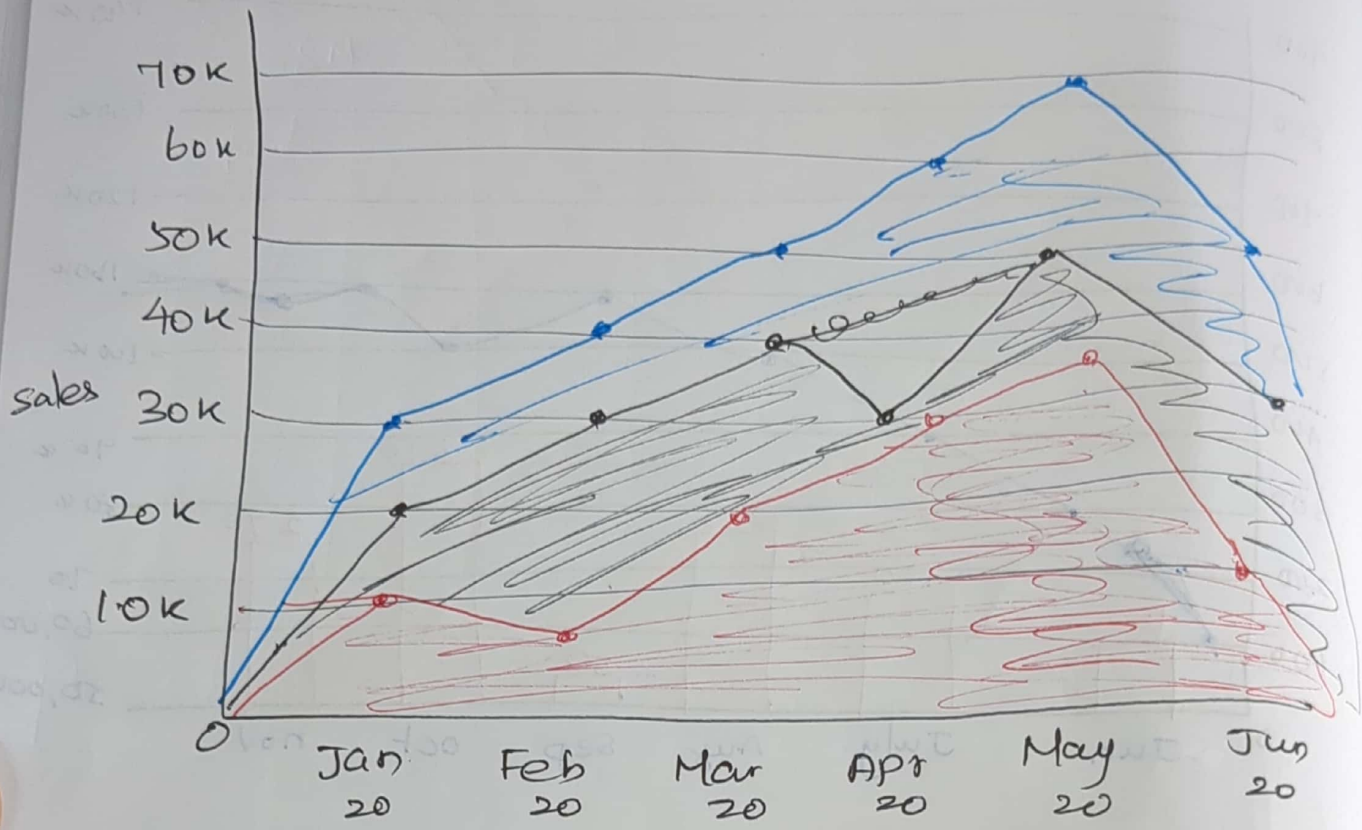
④ Dual axis chart

- 2 y axis & one x axis
- Contains 3 data sets.
- 2 (y) data shared one x axis data
- Uses
  - * Price & volume of products
  - * Revenue & units.
  - * sales & profit.

- how many new customers joined in each month.
- Revenue of that customers are bringing.

⑤ Area chart

- combines line chart & bar chart
- Data points are plotted and connected by line segments.
- The area below the line is coloured.
- Spot & analyze trends.
- Don't display more than 4 categories.



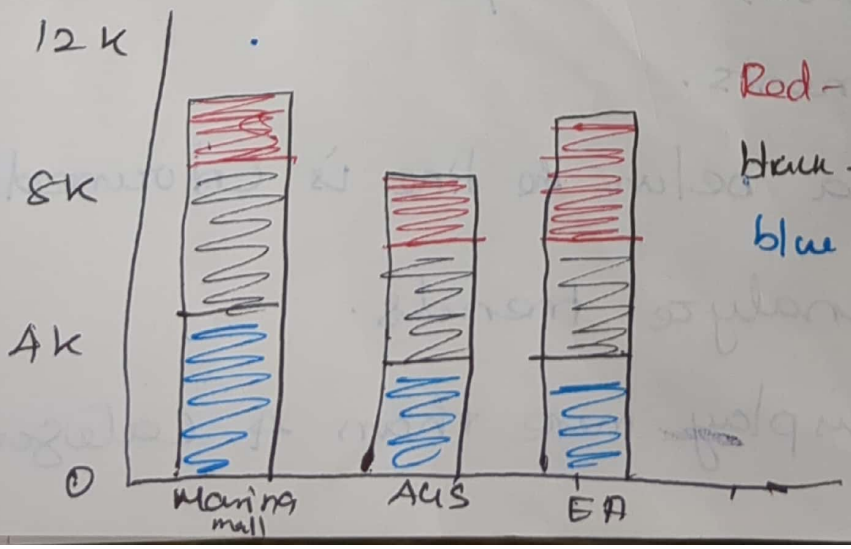
● → Store 1

● → Store 2

● → Store 3

Stacked bar chart

- Used to compare many items.
- Extension of bar chart



Red - clothing

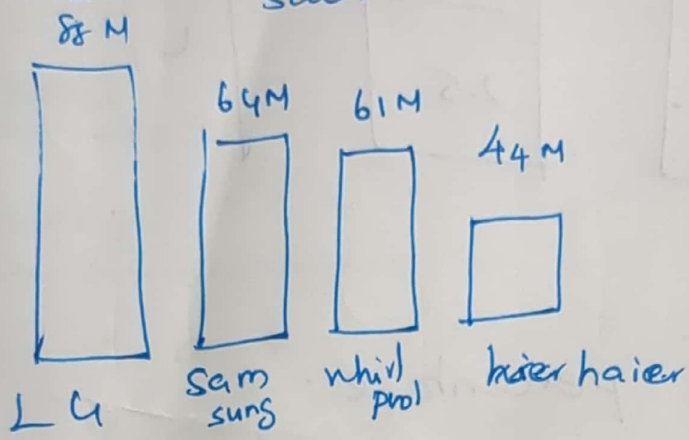
Black - Accessories

blue - Snacks

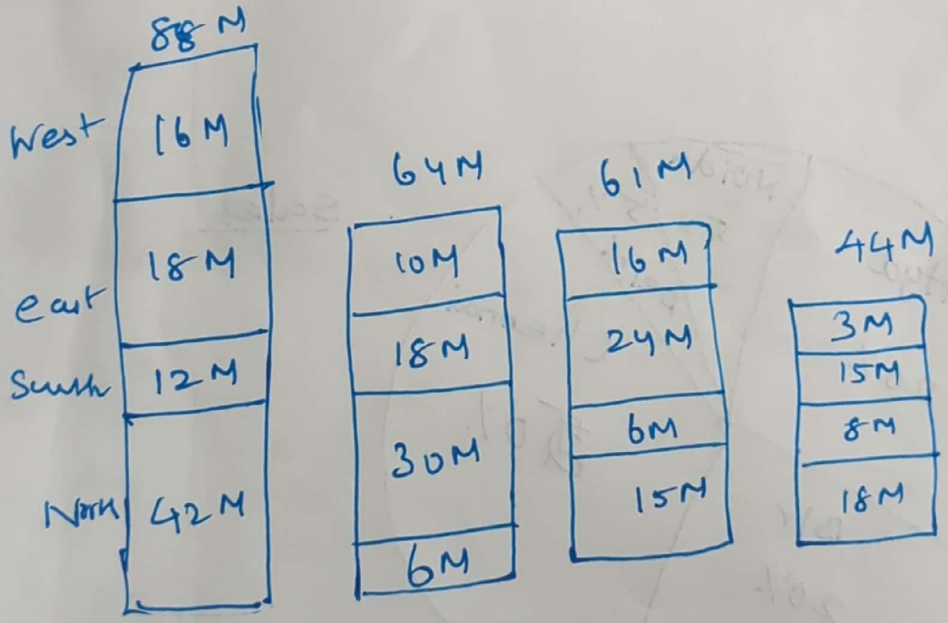
Mekko Chart

- also known as Mari mekko chart.
- Used for growth, market, sales analysis.
- Complex.

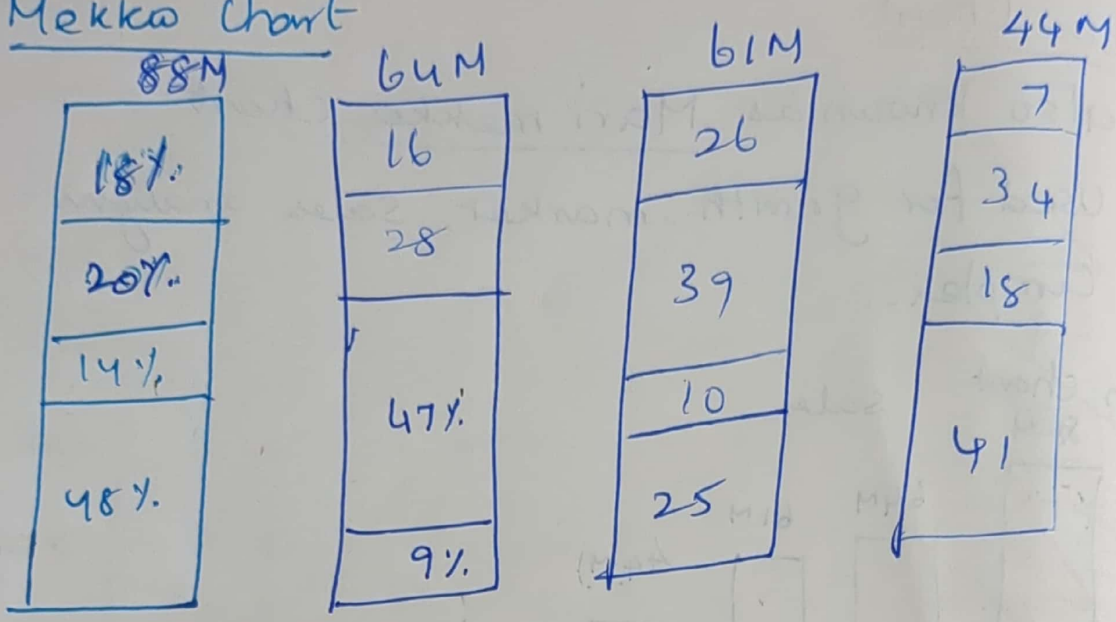
Column chart



Stacked Column chart

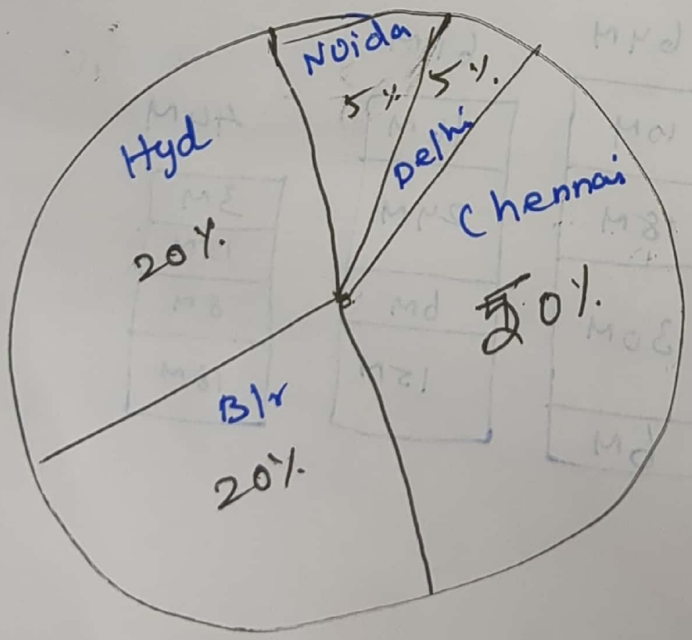


Mekka Chart



Pie Chart

- repr² ng in percentage.
- Total sum of segments → 100%.



Sales

Scatter plot chart

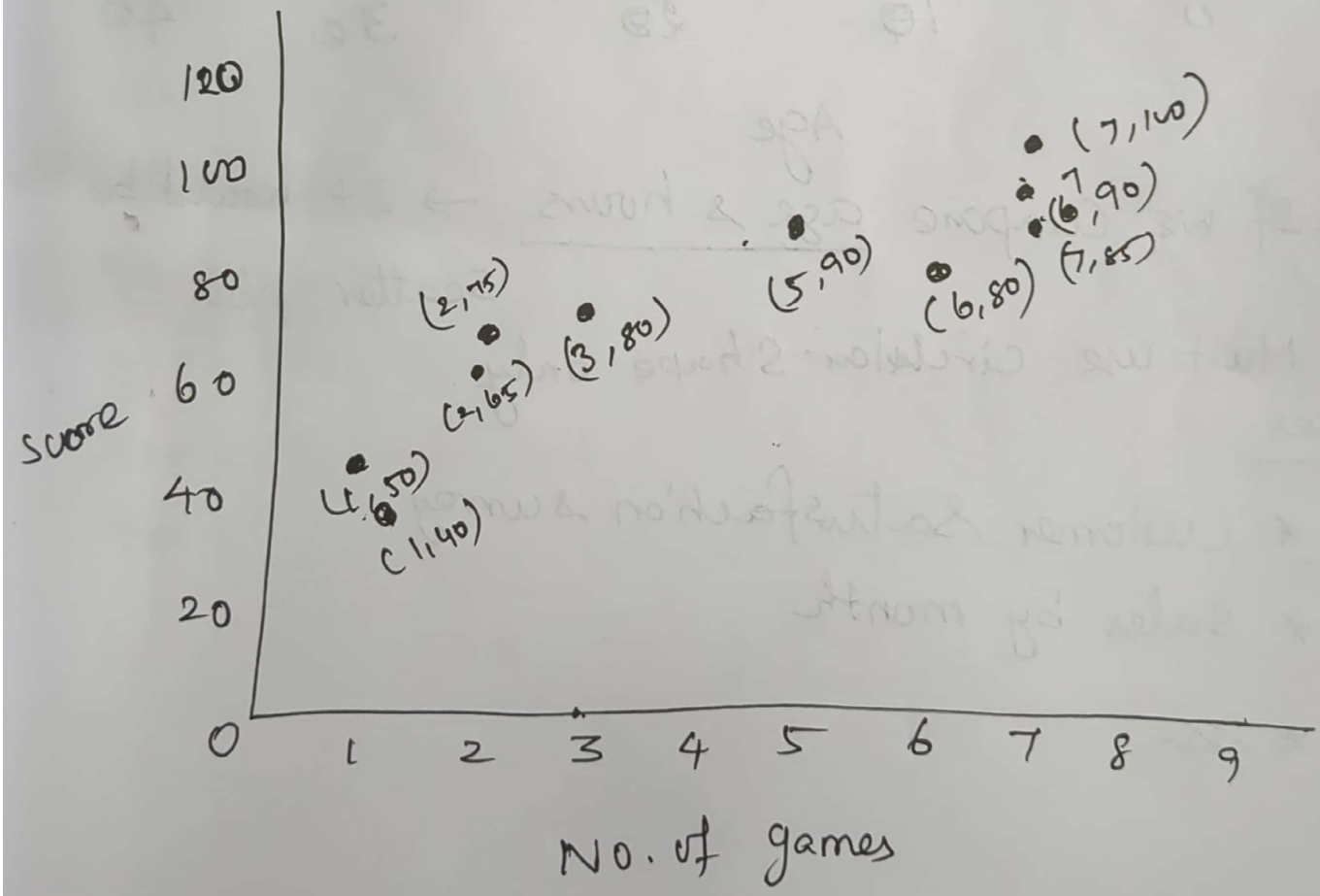
- shows relationship b/w 2 variables.
- Use scatter plot chart, when we have to compare 2 large set of numerical values.

Eg.

No. of games	3	5	2	6	7	1	2	7	1	7
scores	80	90	75	80	90	50	65	85	40	100

X_axis → No. of games.

Y_axis → scores.



Bubble Chart

- Similar to scatter plot, but there is a third dataset shown by bubble.

- If we compare age & hours → It would be scatter plot.
- Must use circular shape only.

Uses

* Customer satisfaction survey.
* sales by month
* ~~xxx~~

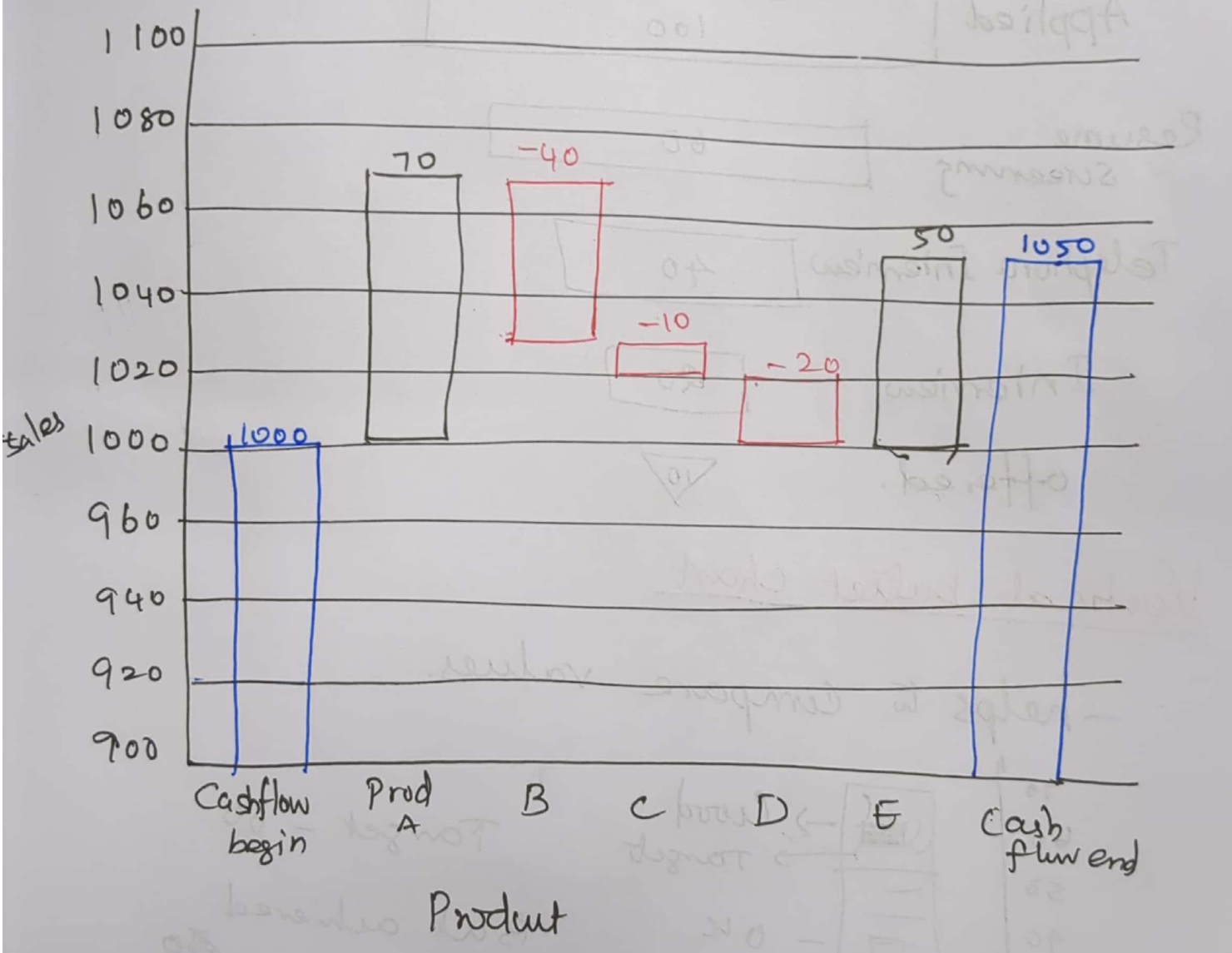
Waterfall chart

- shows how an initial value changes with intermediate values.
- It shows the impact of overall result.

eg

Cash flow begin month → 1000
" " end " → 1050

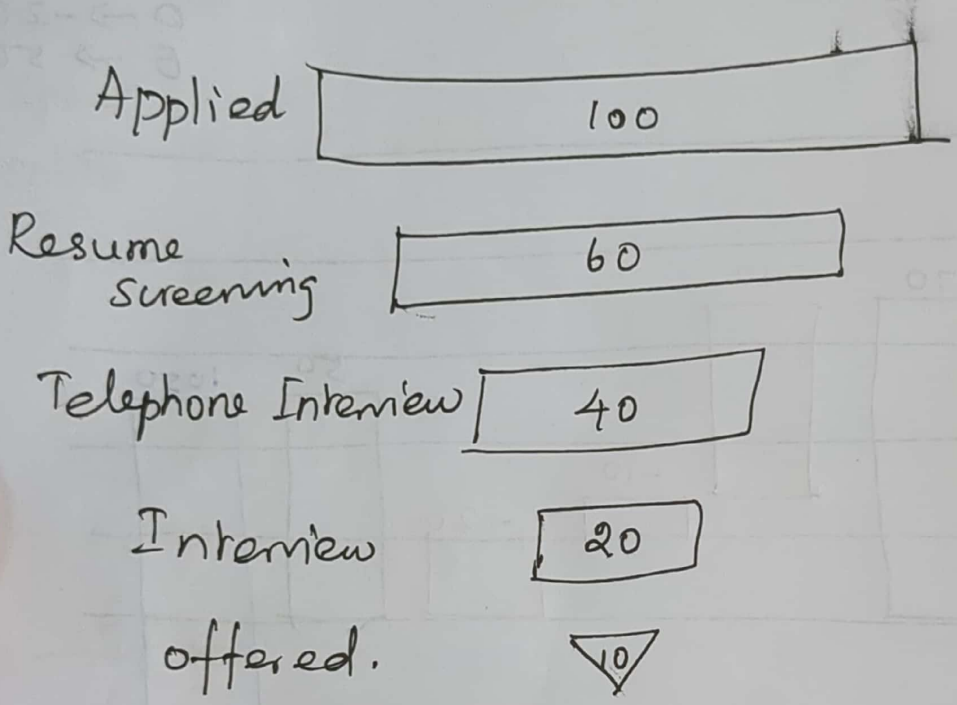
Prod A → 70
B → -40
C → -10
D → -20
E → 50



Funnel chart

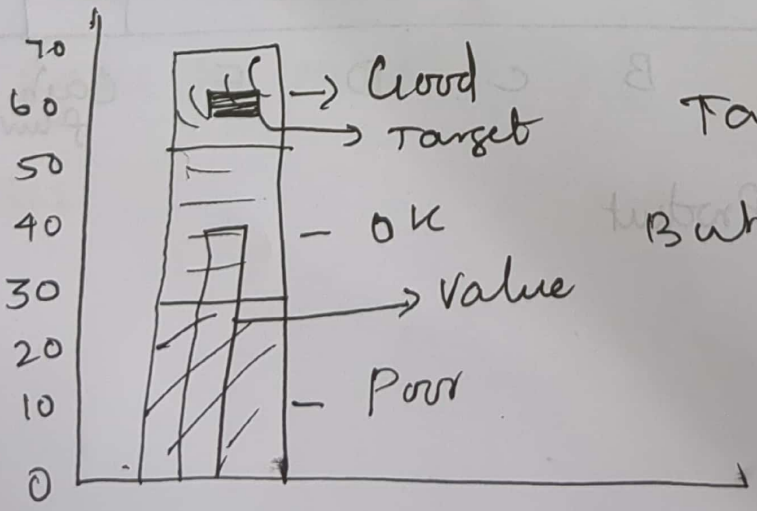
- how data moves through a process.
- used in
  sales, recruitment, real estate.

Recruitment process



Vertical bullet chart

- helps to compare values.



Target – 60
But achieved 40