L-1 ①

Charlie

Date

Page No.

Unit-1

→ • Overview

→ • Motivation [Data Mining]

• Data Mining - Definition and functionalities

• Data Warehousing

- overview
- definition
- delivery process
- difference between Database System & Data warehouse

• Multidimentional data Model.

- Data Cubes
- Snow flakes
- Fact constillation
- Concept Hierarchy
- Process Architecture
- Three Tier Architecture
- Data Marting
  - ROLAP
  - MOLAP
  - HOLAP

Unit-2

• Data Preprocessing

→ Data Cleaning : Missing Values : noisy data [binning]

→ clustering

→ Regression

→ Inconsistent data.

→ Data Integration and transformation.

Date Charlie
Page No.

→ Data Reduction:
- data cube aggregation
- dimentionality reduction
- data compression.
- numerosity reduction.
- CLUSTERING
- Descretization.
- Concept Hierarchy Generation.

Unit-3

Concept Description.
→ Definition
→ Data Generalisation
→ Analytical characterisation
→ Analysis of attribute relevance.
→ Mining class comparison:
→ Statistical measure in large database.
→ Measuring Central Tendency.
→ Measuring Dispersion of data.
→ Apriori Algorithm.
→ Mining Multilevel Association rule from transactional database.

Charlie

Date

Page No.

Unit-4:

Classification

→ What is classification.

→ Issues regarding classification.

→ Decission tree.

→ Bayesion Classification.

→ Classification by back propogation

Unit-5:

Cluster Analysis

→ Data types in cluster analysis.

→ Partitioning method.

→ Hierarichal Clustering

CURE and Chamelion

→ Density based methods.

→ DBSCAN → OPTICS.

→ Grid based methods

→ STING → CLIQUE

→ Outliner Analysis.

Charlie
Date
Page No.

→ Difference b/w database & data Warehouse

database
↓
collection of records

Definition: Data warehousing is a relational database management system. (RDBMS) designed specially, to meet the needs of transaction processing system.

RDBMS: It is a database, that stores information oriented, satisfy decision making request.

Father of Data Warehousing: Bill Inmon

He provides a definition as:
'It is a subject oriented, integrated, non volatile and time variant collection of data, in support of management's decision.'

Features of Data Warehouse:

→ It is seperate from operational database;
→ It integrates data from heterogeneous system.
→ It does not require data to be highly accurate.
→ Queries are generally complex.
→ Supports online analytical processing system [OLAP]

Charlie
Date
Page No.

: Influence decision making in favor of the Enterprise.

Goals of Data Warehousing

→ Data warehousing maintain organisations' historical information.
→ Data warehousing is said to be foundation for decision making.
→ Helps in reporting as well as in analysis

Need of Data Warehouse

(1) Business User

It requires data warehouse to view summarised data from the past. [for simpler representation].

(2) Stored Historical Data

Data warehouse is required to store the time variable data from the past.

(3) Make Strategic Decision

Some strategies may be depending upon the data in the data warehouse.

Date
Page No.

(d) High Response Time

Data warehouse has to be ready for freely unexpected loads and types of queries, which demand a high degree of flexibility and quick response time.

(e) For Data Consistency and Quality

Bringing the data from different sources, at a common place, user can effectively undertake to bring the uniformity and consistency in data.

Difference between operational Database System and Data Warehouse

Features	Operational DB (OLTP)	Data Warehouse (OLAP)
• Users	thousands	hundreds → specific analytics queries
• Workload	present transactions	
• Access	to hundred of records	to millions of records
• Goal	depends on application	decision making support
• Data Integration	Application based	subject based
• Quality	In terms of integrity	in terms of consistency

OLTP: Online Transaction Processing System.

Charlie (1)

Date

Page No.

- Time coverage → Current data only → Current and historical data
- Updates → continious → periodical
- Model → Normalised → denormalised
- Optimization → ~~for OLTP access~~ → for OLAP query to most of db. to db part

Operational Database: It may be categorised as OLTP system.

Data Warehouse: It may be categorised in OLAP system.

☆ The data warehouse and OLTP database are both relational database, but the objective of both these database are different.

DATA WAREHOUSE Db	OLTP Database
Designed for analysis of business measure by categories and attributes	Designed for real time business operations.
Optimized for bulk loads and large, complex, unpredictable queries	Optimised for a common set of transactions usually adding or retriving a

Charlie

Date

Page No.

that access many rows per table	single row at a time per table.
⑤ Loaded with consistent, valid data, requires no real time validation.	Optimized for validation of incoming data during transactions
⑥ Supports low concurrent users relative to OLTP	Supports thousands of concurrent users.

• Normalisation.
1NF 2NF 3NF BCNF

L-3 (9)

Charlie
Date 6-08-2018
Page No.

Differences between OLTP and OLAP

Features	OLTP	OLAP
1) Sources of Data	Operational Data	Consolidation Data
2) Purpose of Data	To control and run fundamental business class.	To help with planning, problem solving and decission support.
3) What the data is?	Reveils a snapshot of ongoing business process	Multidimentional ~~else~~ of various kind of business activities
4) Stored values.	Strictly with the coded data	Stores descriptive data.
5) Processing Speed	Fast	Depends on amount of data.
6) Queries	Standardise and simple query.	Complex queries.
7) Space Requirement	Small	Large
8) Database Design	Highly normalised with many tables	De normalised with few tables.
9) Normalisation	Fully normalised	Partial normalised
10) Modeling	Use entity relation model.	Dimentional model & Snow flake

Charlie
Date
Page No.

11) User	Clerk and IT professionals.	Knowledge worker.
12) Database Design	Application oriented.	Subject Oriented.

OLTP : Online Transaction Processing System.

The major task of Online Operational database system, is to perform online transaction and quick processing.

The cover day to day operations such as
→ purchasing, inventory, registration and accounting.

OLAP : Online Analytical Processing System.

Data warehouse on the other hand serve user or knowledge workers in the role of data analysis and decision making.

Such systems can organise and present data in various formats in order to accommodate the adverse ~~EGE~~ needs of the different users.

• clustering
• classification.

Charlie
Date
Page No.

Data Mining

It is a process of extracting hidden knowledge from large value of raw data, and using it to make crucial business decission.

Functions of Data Mining

1) Classification
→ Infers the defining characteristics of a certain group.

2) Clustering
→ Identifies groups of items that share a particular characteristics.

3) Association
→ {market strategy}
→ Identifies relationship between events that occur at one time,
Such as: shopping basket.
Bread + Jam

4) ~~Sequencing~~

Charlie

Date

Page No.

4) Sequencing

. Similar to association rule, except that the relationship occurs exist over a period.

Some of the benefits

(1) Fraud Management

. Telecommunication, financial, insurance industries.

(2) Market Analysis

Customer, competetion

(3) Entertainment

Digital convergence

(4) Diagnosis and Monitoring

Medical

Q: Why do we use DataMining?

(1) Human analysis, scalar inadequate:
   i. Volume and dimentionality of data.
   ii. High data growth rate.

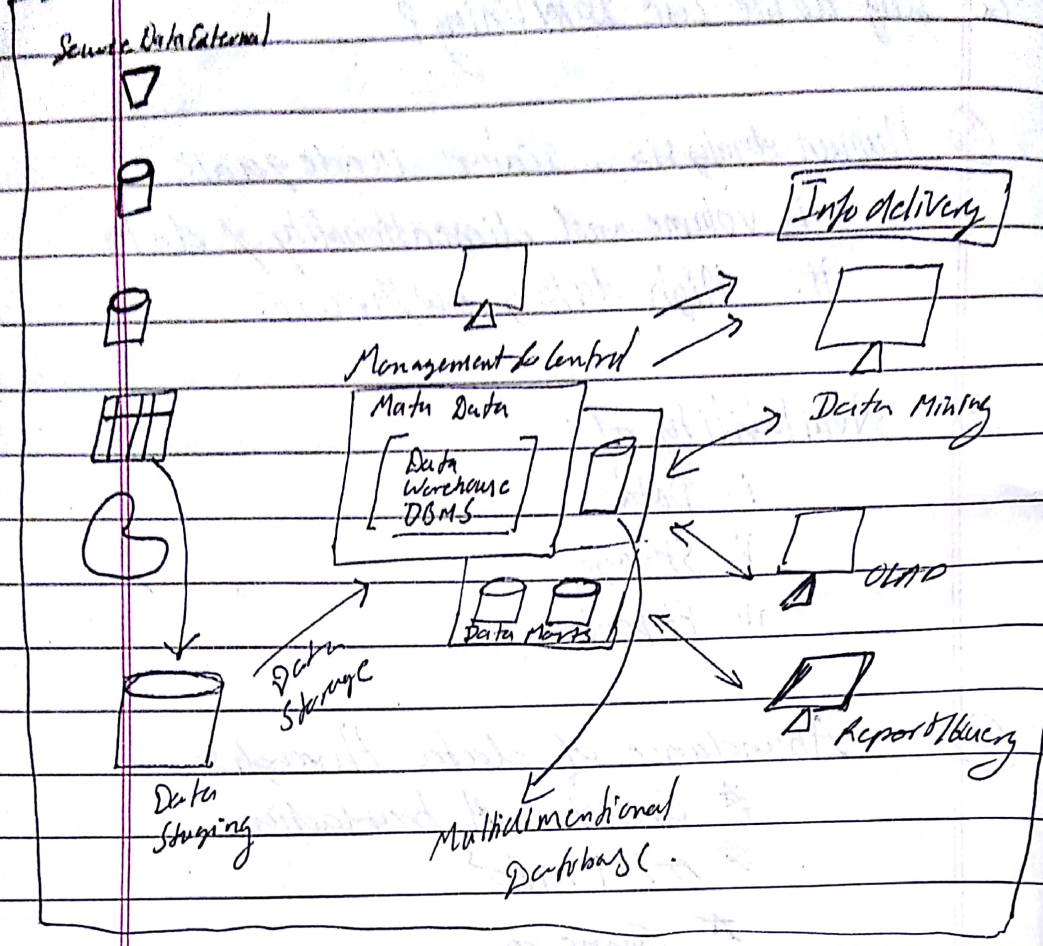
(2) Availability of:
   i. Data
   ii. Storage
   iii. Expertise

(3) Abundance of data through:
   ✯ Credit card transactions
   ✯ ATM M/C's
   ✯ Camera
   ✯ Direct Mail Response
   ✯ Call center records
   ✯ Consor Networks.

(4) Forecasting:
   ✯ Estimates future values, based on patterns within largeset of data.

-Snowflake
-Star Schema

Charlie
Date
Page No.



Data Warehouse Building Blocks or Components.

Data warehouse : collection of records

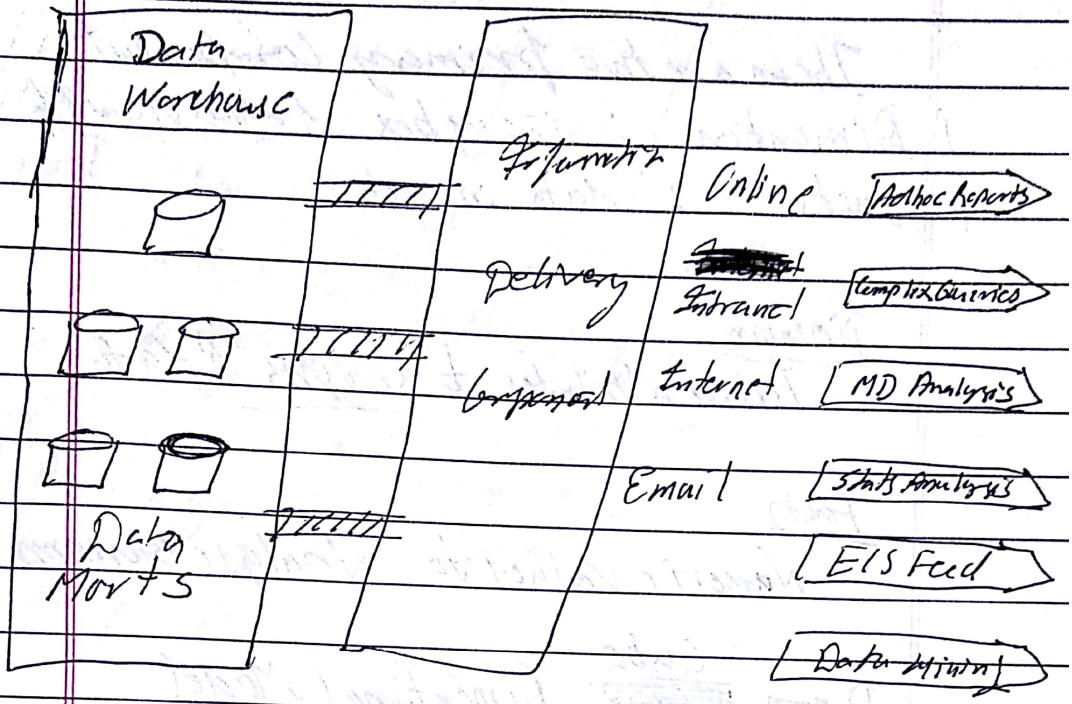
Data mart : catalogue.

Charlie

Date

Page No.

Information Delivery Component



Multi Dimensional Data Model

(MDDM)

The dimensional model was developed for implementing data warehouse and data Marts.

It is an integral part of online analytical processing (OLAP)

It is designed to solve complex queries in real time

Charlie
Date
Page No.

Component of Multi Dimensional Data Model

There are two primary component.

i. Dimension : size of box (which view?)
ii. Facts : data in cells

Dimension

Feature attributes to do analysis on data.

Facts

Numeric values to Analyse Business

Data Cube ~~View~~ Dimensional Model

When data is grouped or combined together in Multidimensional Matrices, then it is called data Cube.

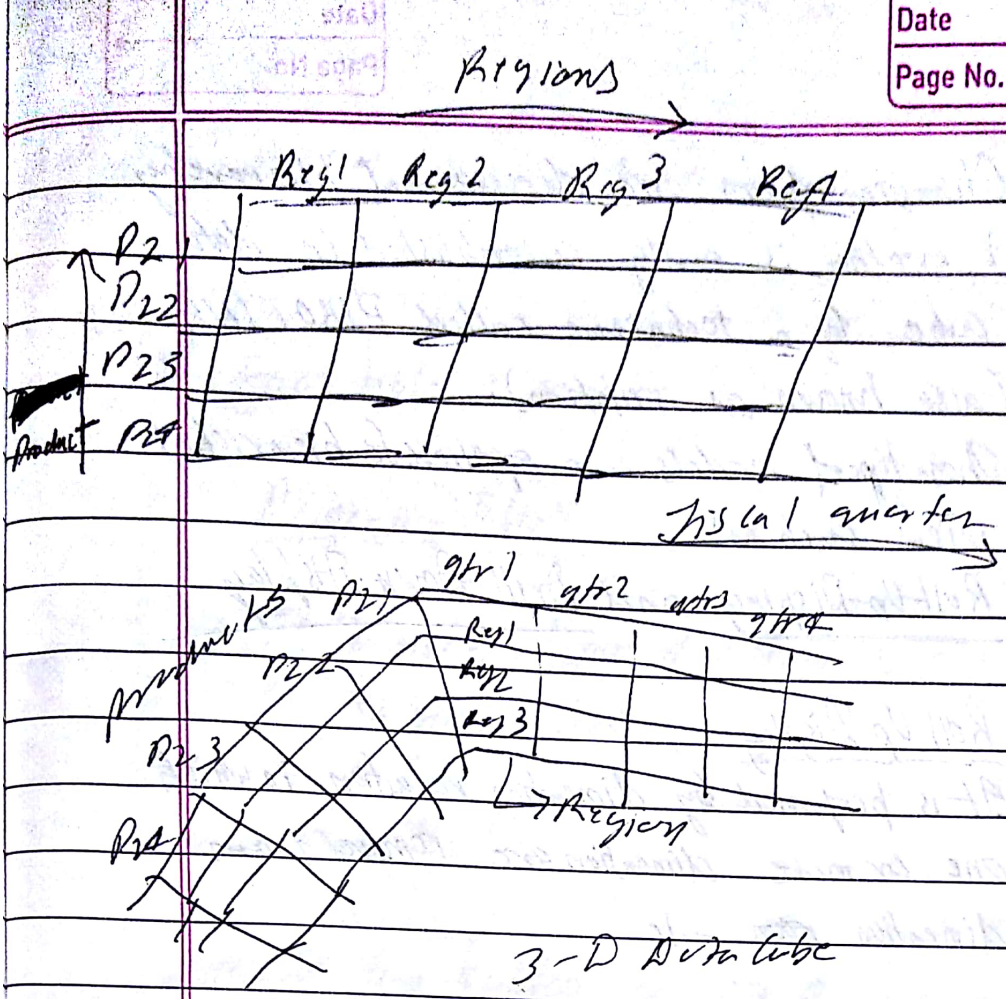
In two Dimension: There is row and column
or
Product and ~~[illegible]~~ fiscal quarters

In 3-D: one region, products and Fiscal Quarter

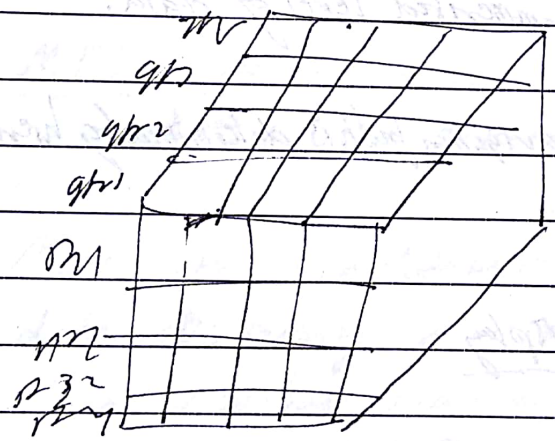
Charlie

Date

Page No.



3-D Data cube



Charlie

Date

Page No.

* Changing from one dimentional hierarchy to another is easily accomplished in data cube by a technique called PIROTING. (also known as rotation).
These type of models are applied to hierarchical view, such as:
Roll-Up-Display and Drill-Down-Display

Roll Up Display
It is performed by dimention reduction in which one or more dimention are removed from dimention ~~rule~~.

With the roleay capability user can zoom out to see a summerized level of data.

The ~~navi~~ navigation path is determined by hierarchy in dimention.

Drill-Down-Display

→ It is reverse of Roll Up Display.
→ It ~~na~~ navigates from less detailed data to more detailed data.
→ It can also be performed by adding new dimention to a cube

Charlie

Date

Page No.

MMDM involve two types of tables:

i. Dimension Table
ii. Fact table.

Dimension Table:

→ It consist of tuples of attributes of dimension.
→ It is a simple primary key.

Fact Table

→ It has tuples, one per recorded fact.
→ It is a compound primary key.

META DATA

→ Information about data.

Metadata in a data warehouse is similar to data dictionary or the data catalogue in a database management system.

Types of Meta data.

(1) Operational Meta Data

(2) Extraction and Transformation Meta Data.

(3) End User Meta Data.

Charlie

Date

Page No.

Operational

It contains all kind of information about the operational data sources.
(Split, Add etc)

Extraction and Transformation of Meta Data

It contains data about the extraction of data from the source system, namely:
Extraction Frequency,
Extraction Method

It also contains information about all the data transformation that take place in the data staging area

End User Meta Data

It is the navigational map of the Data Warehouse.
It enables End User to find information from the data warehouse.

Charlie

Date

Page No.

Goal of Multidimensional Data Model

→ It support analysis in a simple and faster way executives, managers and business professionals.

→ Application:

→ Used in building online analytical processing engines.

Data Cube

When data is grouped, or combined together in Multidimensional Matrices, then it is called Data Cube.

It is constructed from a subset of attributes in the database.

Each dimention may have a table associated with it, called as a dimentional table.

The lattice of cuboids form a data cube.

Multi Dimentional Model, can exist in the form of star schema, snow flake schema or fact constellation schema.

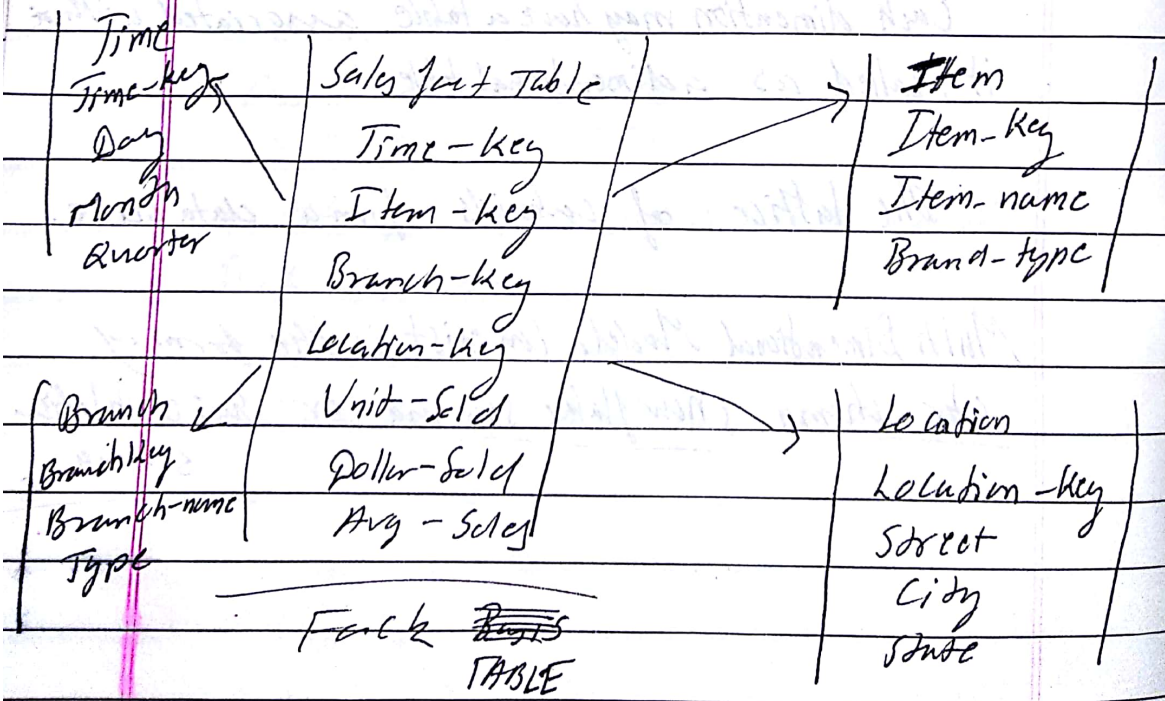
Charlie
Date
Page No.

STAR Schema Model

→ It is also known as STAR join Schema
→ It is the simplest style of Data Warehouse schema.
→ It is called 'star schema' because its entity relationship diagram of this schema resembles a STAR with points radiating from the central table.

→ A star query is a join between a fact table and a number of dimentional table.

→ Each dimention table is joined to the fact table using primary key to foreign key join, but dimention table are not joined to each other.



Charlie

Date

Page No.

Characteristics

→ It creates a denormalised database, that can quickly provide query responses.

→ It reduces the complexity of Meta Data for both developers and End User.

→ Advantage

→ It has simple structure.

→ It has smaller no. of tables and clear join paths.

→ Easy for end users and applications to understand and navigate.

Charlie
Date
Page No.

Snow Flake Schema

→ It is an extention of Star Schema
→ It is different from Star schema, in which the dimention tables from a star schema are ~~are~~ organised into a hierarchy by Normalising them.
→ It is represented by centralised fact table which are connected to multiple dimentions
→ The snow flaking effecting ~~only~~ only the dimention table not the fact table.

★ Definition
Snow flaking is a method of ~~in~~ normalising the dimention table in a star schema.

It improves the performance of certain queries.

Advantages

→ No ~~de~~ redundency
→ Greater flexibility

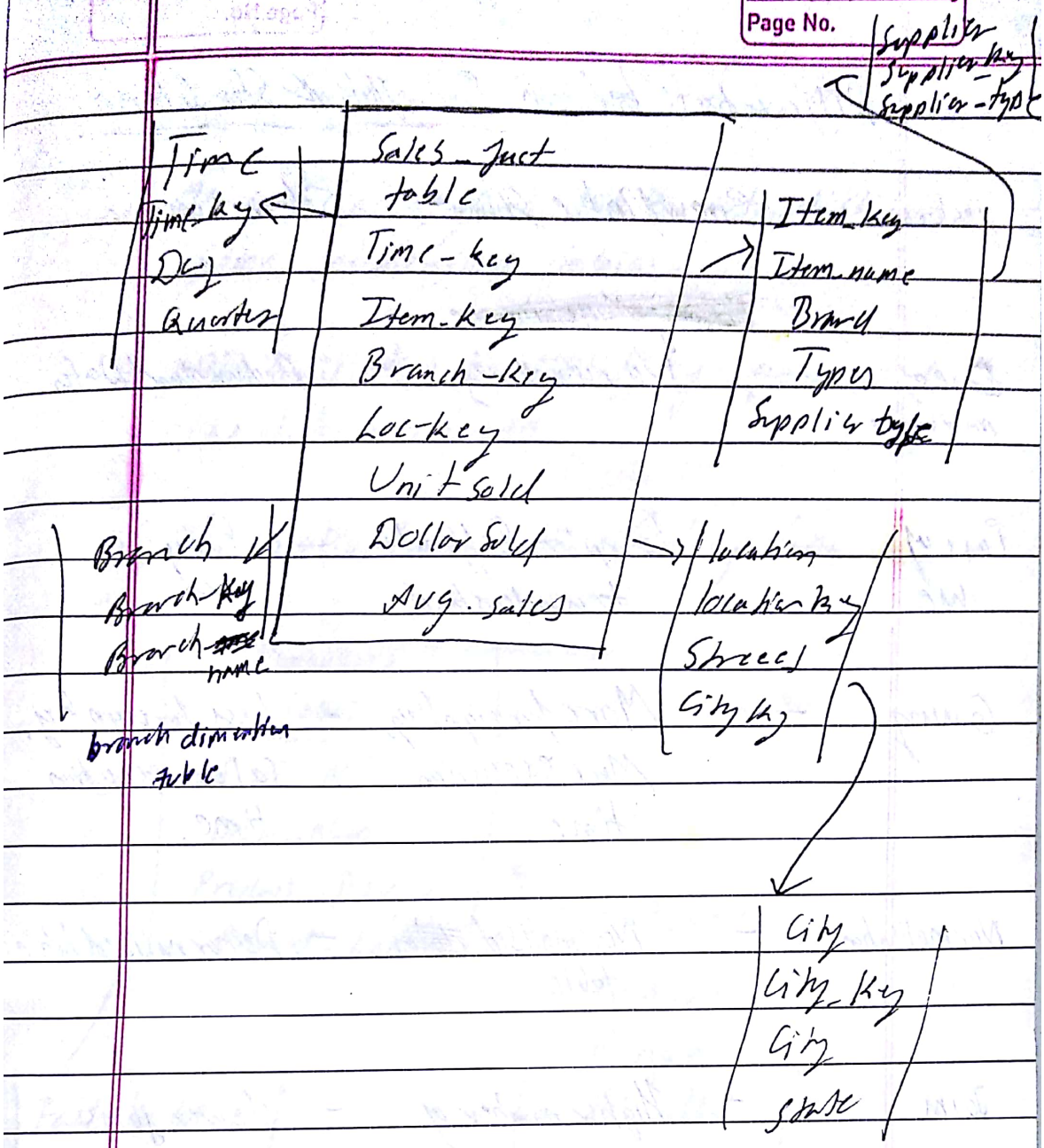
Disadvantage

→ More complex queries
→ Difficult to understand.
→ More tables, more query execution time.

Charlie

Date

Page No.



Charlie

Date

Page No.

Differentiate between Snowflake & Star Schema

Features	Snowflake Schema	Star Schema
Ease of maintainance	No redundancy	Redundant Data.
Ease of use	Complex, Difficult to understand	Easy
Query	More foreign key, More execution time	Less foreign key, Solves execution time.
Normalisation:	Normalised table	Denormalised table
Joins	Higher number of joins	Lower joins
Data Warehouse Systems	Better for small data warehouse, data mart	Works best in any data warehouse data mart

Charlie

Date

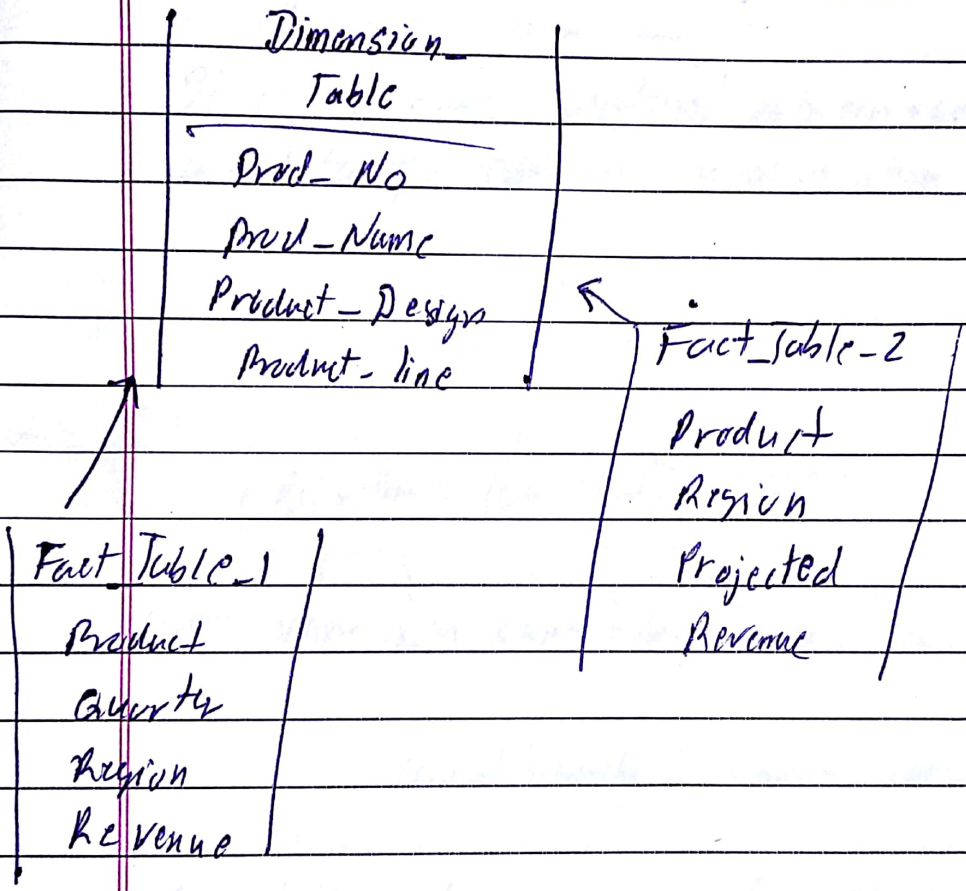
Page No.

Fact Constellation

i. It is a set of fact tables, that share some dimention tables.

ii. It limits the possibility of quering for the data warehouse

iii. It is also known as Galaxy Schema.



Assignment-1

① → Discuss the delivery process in Data warehouse

② → Discuss ROLAP, MOLAP and HOLAP

③ Difference between data warehouse and Data Mart.

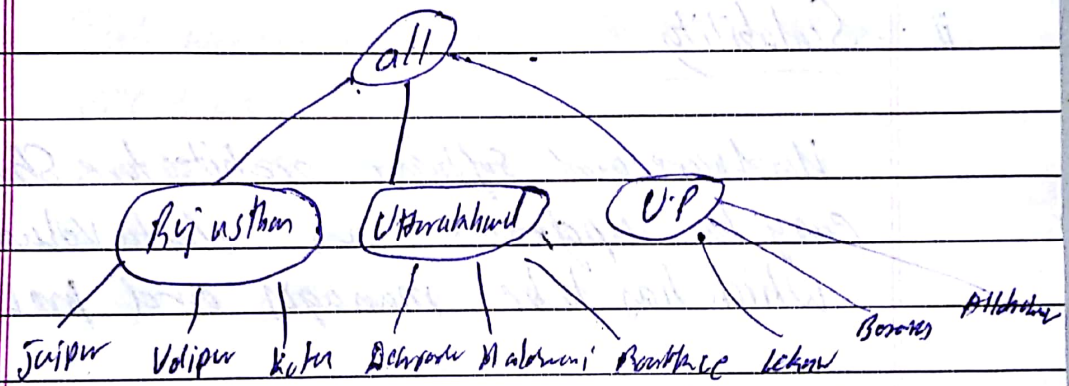
Charlie
Date
Page No.

→ Concept Hierarchy
→ Process Arch.
→ 3-Tier Arch

Concept Hierarchy

It is defined as a sequence of Mapping from a set of low level concepts to higher level concepts to a more generalised concept.

It is an order relation between a set of attributes of a concept or dimension.



Concept hierarchy for dimension location

Concept hierarchy may also be defined as grouping values for a given dimension or attribute, resulting in a set of grouping hierarchy.
It can be provided manually by system user, domain expert or knowledge Engineer.

Charlie

Date

Page No.

Data Warehouse Architecture:

There are some properties for a data warehouse system:

i. Operation [OLAP, OLTP]
   - Insertion, Deletion, Updation
   - For stored in database

Analytical and transactional processing should be kept apart as much as possible.

ii. Scalability

Hardware and software architecture should be easy to upgrade as data volume which has to be managed and processed.

iii. Extensibility

The architecture should be able to host new application and technology without redesigning the whole system.

Charlie
Date
Page No.

iv. Security

Monitoring access is essential, because of the strategic data stored in data warehouse.

v. Administrability

Data warehouse management, should not be overall difficult.

3-Tier Data Warehouse System

Data warehouse adopts 3-tier architecture: There are:

1. Bottom Tier [Data Warehouse Server]
2. Middle Tier [OLAP Server]
3. Top Tier [Front End Tools]

Bottom Tier

→ It acts as a relational database system.
→ Data from operational databases and external sources are extracted using application program interface known as Gateways. [such as: Customer Profiles, information provided by external consultants.
→ A gateway is supported by the underlying

Charlie
Date
Page No.

DBMS and allow client program to generate SQL code to be executed at the server.

Middle Tier

It is an OLAP server that is typically implemented by using either of the following.

→ ~~An external~~ A relational OLAP [ROLAP]
- It is an ~~extended~~ external relational database management system, that maps operations on multidimensional data to Standard Relational Operation.

→ A Multidimensional OLAP. [MOLAP].
- It is a model that is a special purpose server, that directly implements multidimensional data and operations.

Top Tier

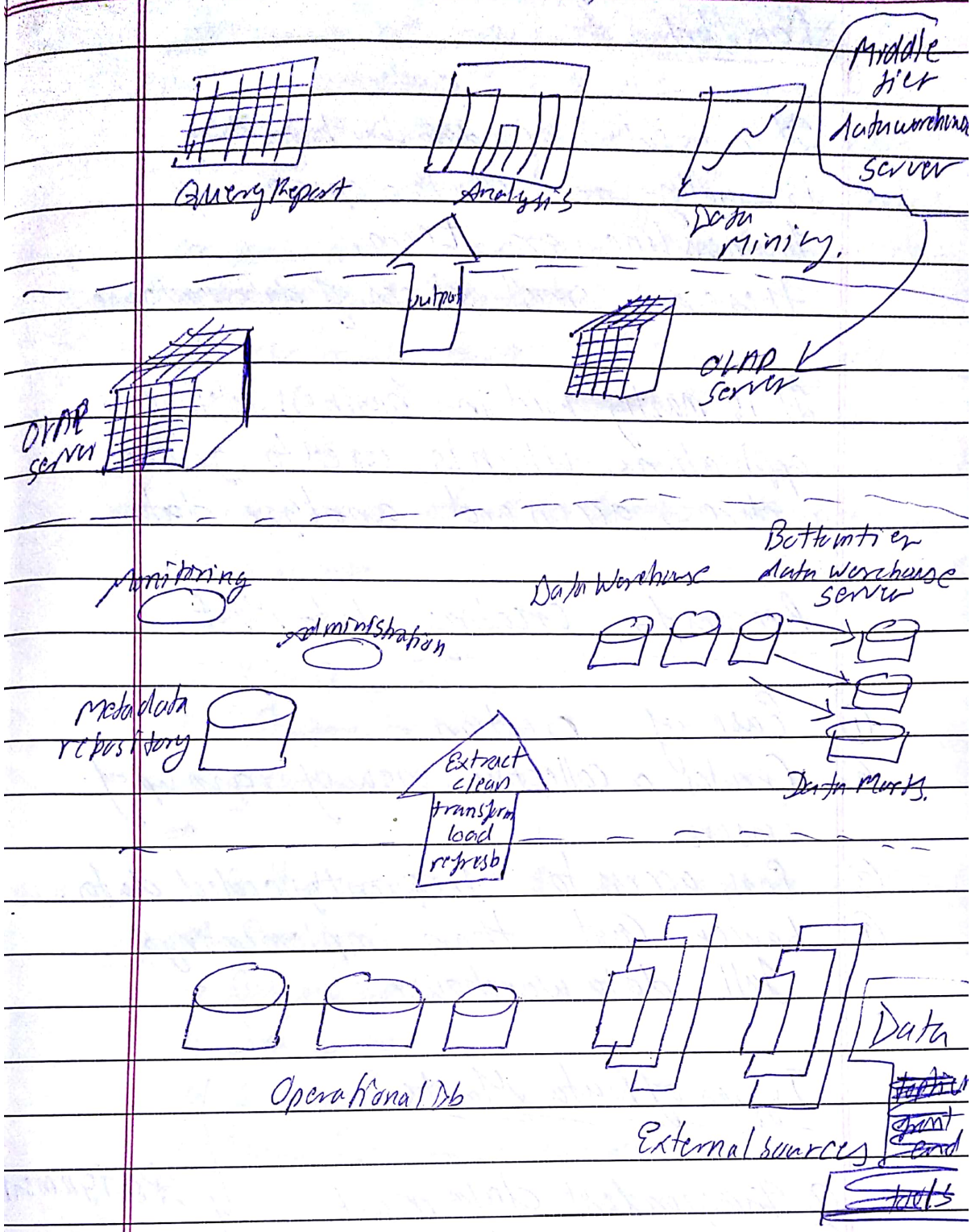
- It is a client which contains query and reporting tools, analysis tool or data mining tools.
  eg: Trend Analysis, prediction and so on.

Charlie

Date

Page No.

Top Tier Front end tools



3 tier Data Warehouse Architecture

Date
Page No.

Charlie

Data Mart

It is a subset of data warehouse that is usually oriented to a specific business line or team.
These are small slices of data warehouse.

It is mostly used in business intelligence applications, which is used to gather, store, access and analyse data.

Reasons for creating data Mart

1. Ease of creation.
2. Creates a collective view of, group of users.
3. Easy access to frequently needed data.
4. Lower cost than implementing a full data warehouse.

Types of Data Marts

1. Independent data mart
2. Dependent data mart.

Assignment

→ Architecture
→ Advantages Disadvantages.
→ Differentiation.

Charlie

Date

Page No.

The data we wish to analyse by data mining technique, are in incomplete form, (lacking attribute values or only certain attribute of interest), noisy (containing errors or outliners value which deviate from the expected and inconsistent data).

Discrepencies error in departmental code, used to category data items

Why we need preprocess the data!

→ Redundency

→ Aggregation

→ Removal of Noisy Data

→ Inconsistent Data

→ Generalization

Why we need the transformation of data!

→ Durability

→ Data Analysis

→ Data scaling

→ Consistency

→ To make the dataview in a simple form

Different types of Data Preprocessing techniques.

→ Data Cleaning

→ Data Integration

→ Data transformation

→ Data Reduction.

Charlie

Date

Page No.

Data Cleaning

It can be applied to remove noise and correct inconsistent data.

Data Integration

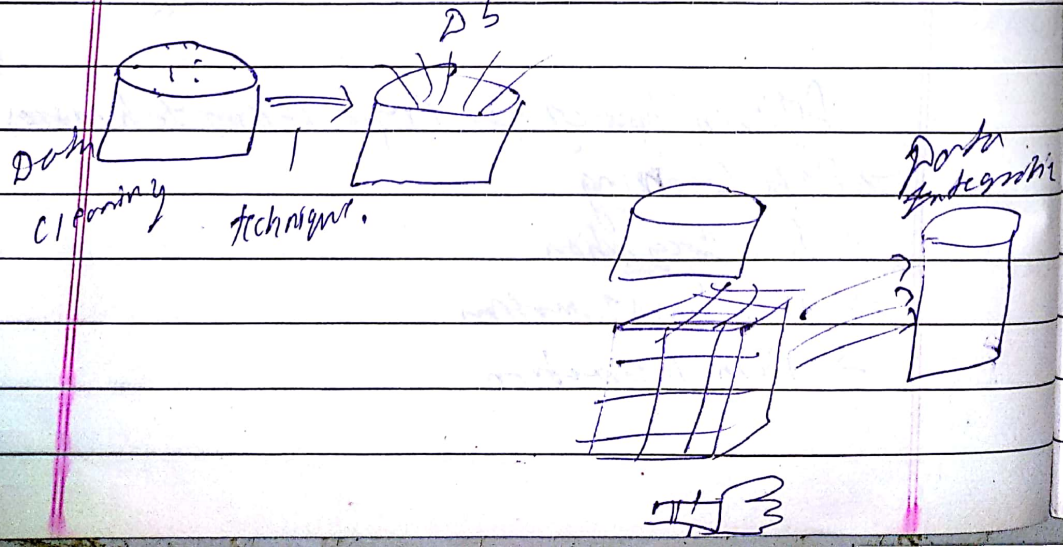
It merges data from multiple resources, into coherent data.

Data transformation

In this technique normalization, can be applied to improve the accuracy, efficiency of Data Mining Algo.

Data Reduction

It reduces the data size, by aggregating, eliminating redundant data or clustering.



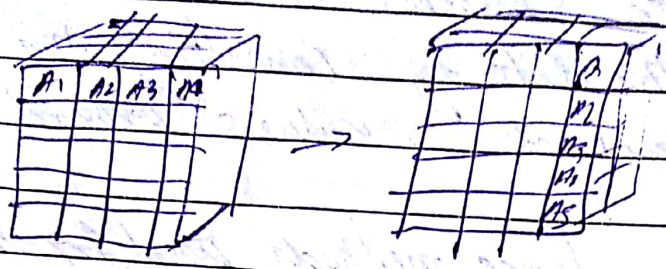
Charlie

Date

Page No.

front view → top view → side view

Data Reduction



Noisy Data

It is a Random Error, or a variance in a measured variable, due to randomness, it is very difficult to follow the strategy for noise removal from the data.

Common methods for removing noise from data, are as follows:

1. Binning Method.
2. Regression
3. Clustering
4. Combine computer & Human Inspection.
5.

Date
Page No.

Binning method

It smoothes the data or sorts the data by consulting the neighbourhood or values around it.

There are three methods used for binning.

i. Partition into equal frequency bins.
ii. Smoothing by bin means.
iii. Smoothing by bin boundaries.

Numerical

→ Let there by the data price in $,
4, 8, 15, 21, 21, 24, 25, 28, 34

Solve the question by binning method:

i. Partition into equal frequency bins.
ii. Smoothing by bin means.
iii. Smoothing by bin boundaries.

(1) Bin 1 (4, 8, 15)
Bin 2 (21, 21, 24)
Bin 3 (25, 28, 34)

(2) Mean of bin1 was

Bin 1 (9, 9, 9)
Bin 2 (22, 22, 22)
Bin 3 (29, 29, 29)

Charlie

Date

Page No.

(3) Replace by minimum.

Bin1 (4, 4, 15)
Bin2 (21, 21, 24)
Bin3 (25, 25, 34)

REGRESSION

- regression data can be smoothed by fitting the data to a function
- linear regression involves:
  finding the best line to fit the value/ two variables so that one variable can be used for predicting the other.
- multiple linear regression.
  it is an extention of linear regression where more than two variables are involved, and the data are fit to a multidimentional surface.

regression is used for -
- classification, prediction of data.

co-relation formula.

Charlie

Date

Page No.

CLUSTERING

~~Def~~ In clustering, grouping of data into different groups are performed, so that data in each group shares similar trends and patterns.

Data Integration and transformation

Data integration means, merging of data from multiple data ~~source~~ stores.

This data may also be needed to transform into forms (4) appropriate for mining.

The source may include multiple databases. The data integration problem: i.e Redundency, can be removed by co-relation.

Numerical

Find the Co-relation!

(a) Charlie

Date

Page No.

No. of study hours (x)	No. of sleeping hours (y)
2	10
4	9
6	8
8	7
10	6
Σ 30	40

$$r_{xy} = \frac{\Sigma(x-\bar{x})\cdot(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \cdot \Sigma(y-\bar{y})^2}}$$

$$= \frac{-20}{20} = -1$$

$\bar{x} = 6$

$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
-4	16	-8
-2	4	-2
0	0	0
2	4	-2
4	16	-8
	40	Σ -20

$\bar{y} = 8$

$y-\bar{y}$	$(y-\bar{y})^2$
2	4
1	1
0	0
-1	1
-2	4
	10

Charlie
Date
Page No.

Note

If the resulting value = 0, then we say that x and y are independent.

If the resulting value is less than 0,

then we say that x and y are negatively co-related, where the values of one attribute, increases as the values of other attribute decreases.

Data Transformation

In data transformation, the data are transformed into appropriate forms for mining.

The data transformation can involve, the following:

i. Smoothing
ii Aggregation
iii Generalisation

Smoothing

Its function is to remove noise from the data.

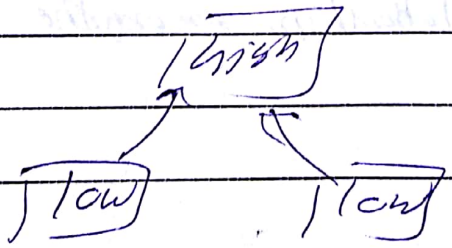
These technique involve:
- binning - regression & clustering

Aggregation

In this, the summary or aggregation operation are applied for the data, eg: the daily sales data maybe aggregated, so as to compute monthly and annual total amount.

Generalisation

Generalisation of data, where low level or primitive data are replaced by high level concept, through the use of concept hierarchies, eg: 'categorial attributes' like street can be categorical, to a higher level concept, like City or Country.



Charlie

Date

Page No.

Normalisation

In this the attributes are scaled, so as to fall within a small specified range; there are many methods for normalisation:

i. Min-Max Normalisation

→ it performs linear transformation on the original data.

Suppose that:

Min A and max A are the minimum and maximum values of an attribute then, min-max can be computed as:

$$v' = \frac{v - \min A}{[\max_A - \min_A]} \cdot [new\_max A - new\_min A] + new\_min$$

where v to v' is the range and new-max A, new-min A represent the relationship among the original data values.

06 Charlie

Date
Page No.

Numerical

Suppose that the minimum and maximum values for the attribute income are ₹12,000 and ₹98,000 respectively.

We would like to map income to the range [0, 1.0] by Min-Max normalisation. a value of ₹73,600 for the income, will be transformed to ?

$$v' = \left[\frac{73600 - 12000}{98000 - 12000}\right] \cdot [1.0 - 0] + 0$$

$$= \frac{61600}{86000} \cdot 1$$

$$= \frac{616}{860} \; \frac{308}{430} \; \frac{154}{215}$$

$$= 0.716$$

Charlie

Date
Page No.

Z-Score Normalisation

In this the values for an attribute A are normalised based on the mean and standard deviation of A.
A value 'v' of A is normalised to 'v' by is given by $\frac{v - \bar{A}}{\sigma_A}$, where $\bar{A}$ and σ_A are mean and standard deviation.

This method of normalisation is used when the actual minima and maximum of attribute A are known.

Numerical

Q. Suppose that the mean and the standard deviation of the values for the attribute income are $54,000 and 16,000 respectively, with Z is for normalisation, a value of $73,600 for a income is transformed by?

$$\frac{73600 - 54000}{16,000}$$

$$= \frac{19600}{16000} = 1.225$$

Charlie

Date

Page No.

Normalisation by Decimal Scaling

It is normalised by moving the decimal point of values of attribute A.

The no. of decimal point move, depends on the maximum absolute value of A.

A value V of A, is normalised to V' by computing

$$V' = \frac{V}{10^j}$$, where j is the smallest integer, such that maximum of V' is less than 1.

Date Charlie
Page No.

Data Integration

For a categorial data (discrete), a correlation relationship b/w two attributes A and B can be discovered by 'Chi-Square' χ^2.

Suppose that, a 2×2 contingency table for the data is given below:

	male	female	Total
fiction	250	200	950
non-fiction	50	1000	1050
Total	300	1200	1500

for numeric data

find the co-relation value of the given table.

$$r_{AB} = \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{N \sigma_A \sigma_B}$$

$$= \frac{\sum (a_i b_i) - N\bar{A}\bar{B}}{N \sigma_A \sigma_B}$$

where, N = no. of tuples

a_i b_i = represent values of a and b in the tuples

$\bar{A}, \bar{B}$ = represent mean value of A and B.

σ_A σ_B = Std. deviation of A & B

$$-1 \leq r_{AB} \leq +1$$

① if $r_{AB} > 0$ then, A & B are truly co-related.

② if values = 0, then A and B are independent, and there is no-correlation between them.

③ if $r_{AB} < 0$ the A & B are ~~truly~~ negatively co-related.

For categorical data

Note

Suppose A has c distinct values, $a_1, a_2 \cdots a_c$ and B has $b_1, b_2 \cdots b_c$

The data tuples described by A & B can be shown as a contingency table as:

Let (A_i, B_j) denote the event that attribute A takes on the value a_i & attribute B takes on value b_j

where,

$A = a_i$ and $B = b_j$

Charlie

Date

Page No.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where,

o_{ij} = ~~Observed~~ observed frequency

and

e_{ij} = expected frequency.

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B)}{N}$$

Charlie

Date

Page No.

Numerical on Contingency

	M	F	Total
Smokers	250 C_{11}	200 C_{12}	450
non-Smokers	50 C_{21}	100 C_{22}	1050
total	300	300	1500

$$E_{11} = \frac{\text{Count (male)} \times \text{Count (Smokers)}}{N}$$

$\chi(11)$

$$= \frac{300 \times 450}{1500}$$

$$= 90$$

$\chi(12)$

$$C_{female} \frac{1200 \times 450}{1500}$$

$$= 360$$

Non Smokers

$$e_{male} \frac{300 \times 1050}{1500}$$

$\chi(21)$ $= 210$

$$C_{female} = \frac{1200 \times 1050}{1500}$$

$\chi(22)$ $= 840$

90	360
210	840

Charlie
Date
Page No.

$$x^2 = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360}$$

$$+ \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$

$$= \frac{(160)^2}{90} + \frac{(-160)^2}{360} + \frac{(160)^2}{210} + \frac{(160)^2}{840}$$

$$= \frac{25600}{90} + \frac{25600}{360} + \frac{25600}{210} + \frac{25600}{840}$$

$$= 284.44 + 71.11 + 121.90 + 30.47$$

$$= 507.92$$

$$= 22.537$$

15-oct-19

53 Charlie

Date

Page No.

Assignment-2

Write short note on

Data cube Aggregation.

Data Compression

Numerosity Reduction

Clustering

Concept hierarchy generation.