Data warehouse and Knowledge mining

Assignment - 1

Name: B. R. Devaprasad

USN: ENG20DS0033

SEM: 5

SEC: H

Course code: 20DS3501

Module - 1

(i) Choose a real time domain and apply all the OLAP operations on the cube.

(ans) Dataset on sales of phone and earphones.

	Unnamed	month	year	location	Product	Value	quarter	country
0	18	1	2017	CA	PS5	420	Q_1	USA
1	39	1	2017	NY	PS5	510	Q_1	USA
2	77	1	2017	CA	VR	400	Q_1	USA
3	80	1	2017	WA	VR	560	Q_1	USA
4	114	1	2017	NY	PS5	600	Q_1	USA

(i) Dice:

each dimension to a certain range of values, while keeping the numbers of dimensions the same is the resulting cube.

We can slows in sales happening in [Jan/Feb/Mar]

year	Product	month	amount
2017	PS5	1	7100
		2	6890
		3	6622
	VR	1	5600
		2	4960
		3	6920

(ii) Roll up:

applying an aggregation function to collapse a number of dimension, we want to jows in the annual revenue for each product and collapse the location dimension.

year	amount
2018	200801
2019	199867

(iii) slice:

fixing certain dimensions to analyse the remaining dimensions. we can jows in the sales happening in 2019 or can jows on sales happening in 2019

	month	quarter	year	location	Product	amount
263	2	Q1	2019	CA	VR	2100
264	2	Q1	2019	QU	PS5	620
265	2	Q1	2019	WA	VR	2560
266	2	Q1	2019	NY	PS5	590
267	2	Q1	2019	CA	PS5	660

(iv) Drill down:

in the reverse of rollup and applying an aggregatio function to a fines level of granularity. we want to fows in the annual and monthly revenue for each product

year	amount
2018	12340
2019	210129

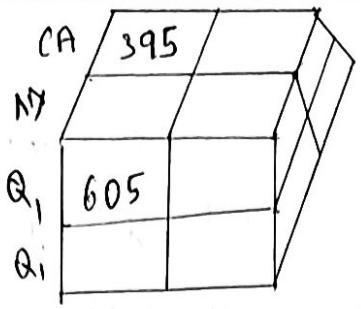
(v) Pivot:

analyzing the combination of a pair of selected dimensions. we want to analyze the revenue by year and month.

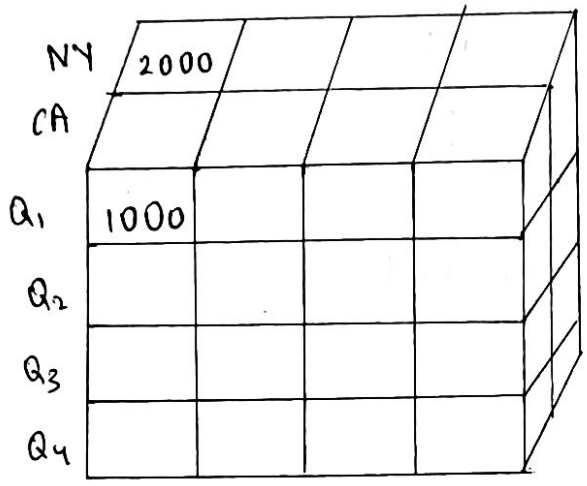
Product	CA	NY	QU	WA
VR	1219.1111	2291.3076	1410.000	1436.1020
PS5	629.4769	6612.419	2169.2120	1120.1211

→ applying this on code:

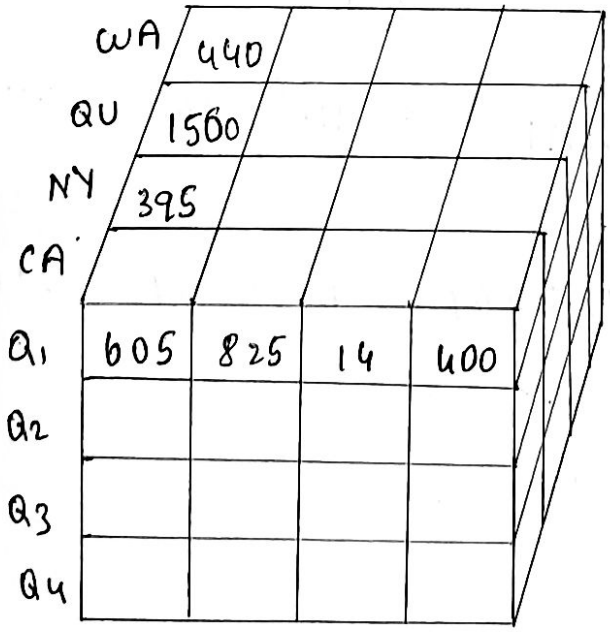
(i) Dice:



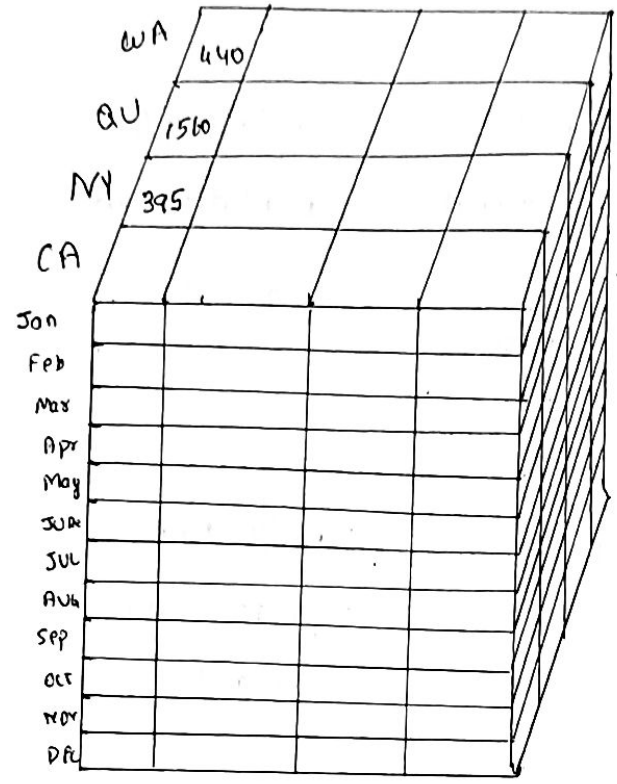
(ii) Rollup:



(iii) Slice:



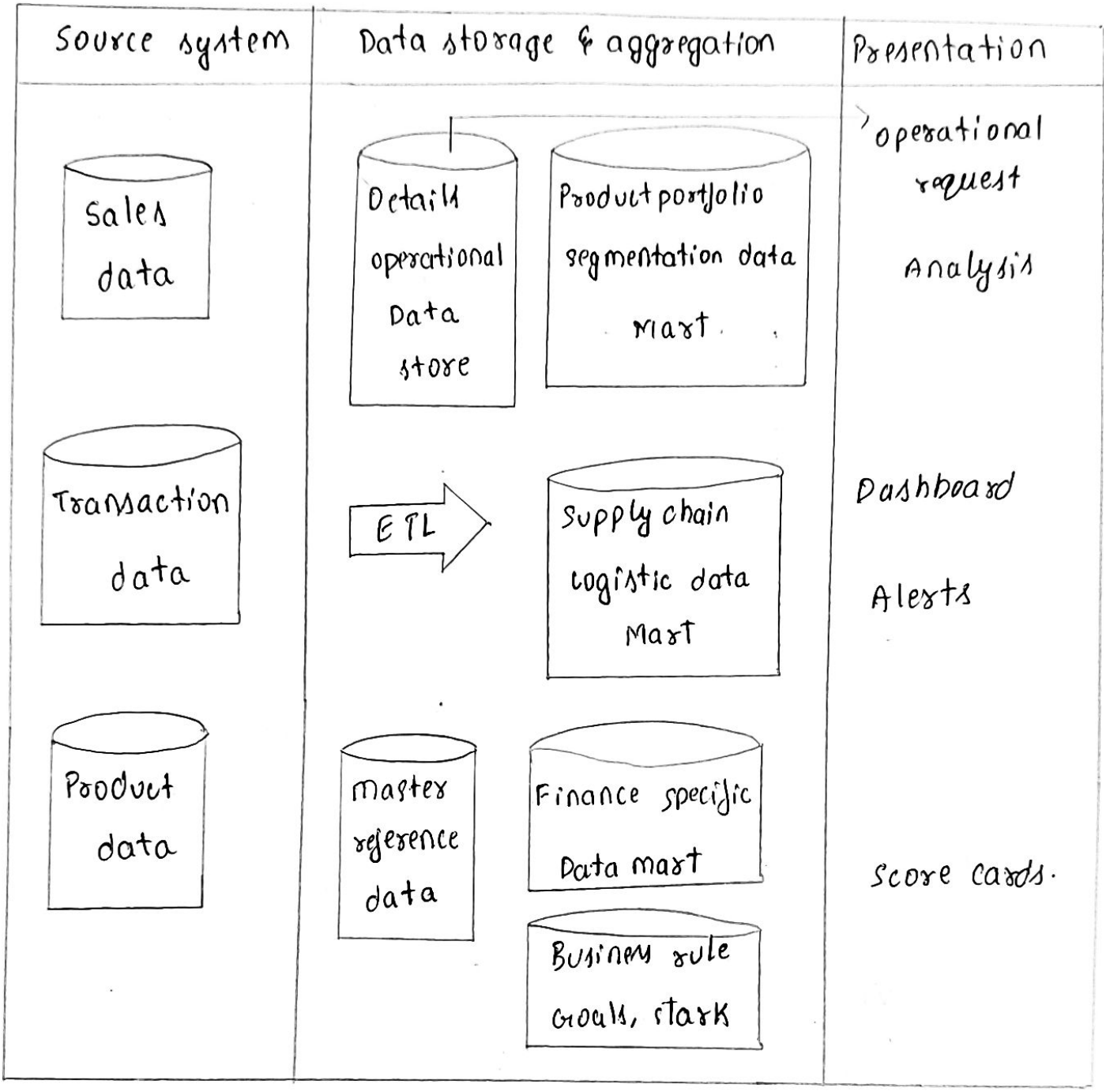
(iv) Drill down:



(v) Pivot:

YR				605
PS				825

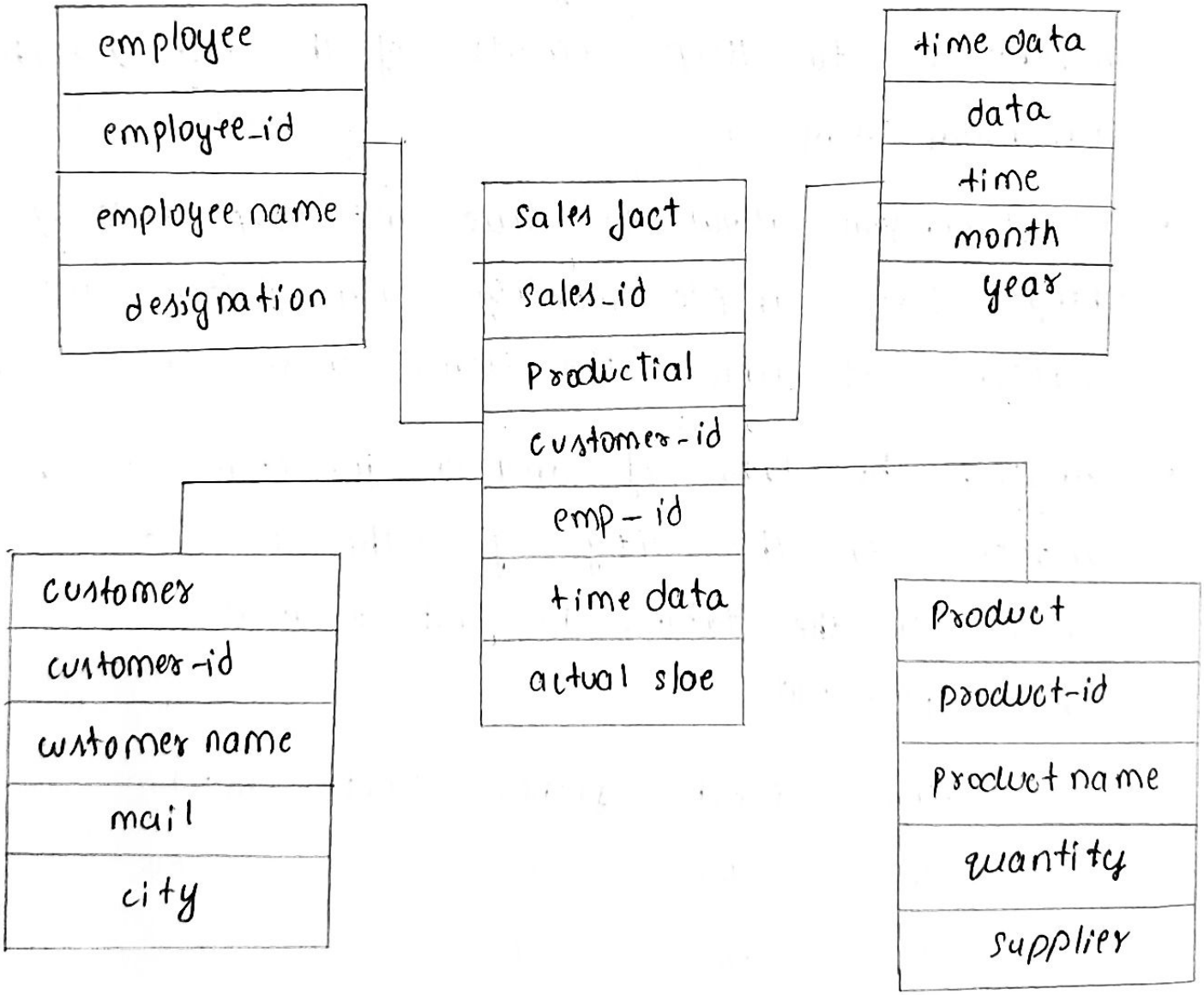
(2) Design a data warehouse for parallel processing. (Domain- Retail).



- Data warehouse architecture of retail industry helps to understand the basic overview which can be optimised for retail stores
- Data is obtained from multiple sources for example sales data, transaction data.
- The data is extracted from the sources and data cleaning operations are implemented on this data

- This transformed data is fed to data storage and aggregation layer which consists of data warehouse and several data mode
- The data is then presented to the decision making in the form of operational reports, analysis, dashboard.

Schema diagram



⟨3⟩ Multidimentional data model for amazon

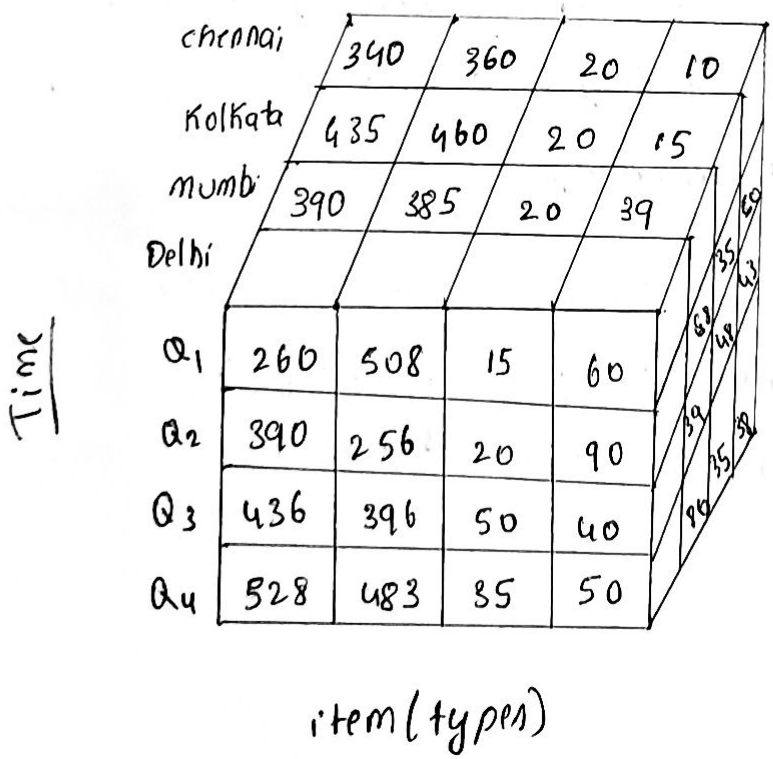
⟨ans⟩

- a multi dimensional model views data in the form of data cube. A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and fact
- The dimensions are the perspective or entities concerning which an organisation keeps record. For example Amazon may create a sales data warehouse to keep records of the sales for the dimension time.
- The dimension allow to save and keep track of things for example monthly sales of items and location at which the items were sold.
- consider the data of amazon for items sold per quarter in the city of delhi. The data is shown in the table. The fact or measures displayed in rupee-sold.

Time	Fresh	fashion	Electro	Kitchen
Q_1	260	508	15	60
Q_2	390	256	20	90
Q_3	436	396	50	40
Q_4	528	483	35	50

- If we want to view the sales data with a third dimension. Suppose the data according to time and item, as well as the location is considered for the cities chennai. The 3D data are shown below.

	chennai			Kolkata		
Time	Fresh	Jachion	electronic	Fresh	Jashion	otecronic
Q_1	340	360	20	435	460	20
Q_2	490	490	16	389	385	45
Q_3	680	583	46	684	490	39
Q_4	585	694	39	335	365	83



(4) Design a data cube for market basket analysis.

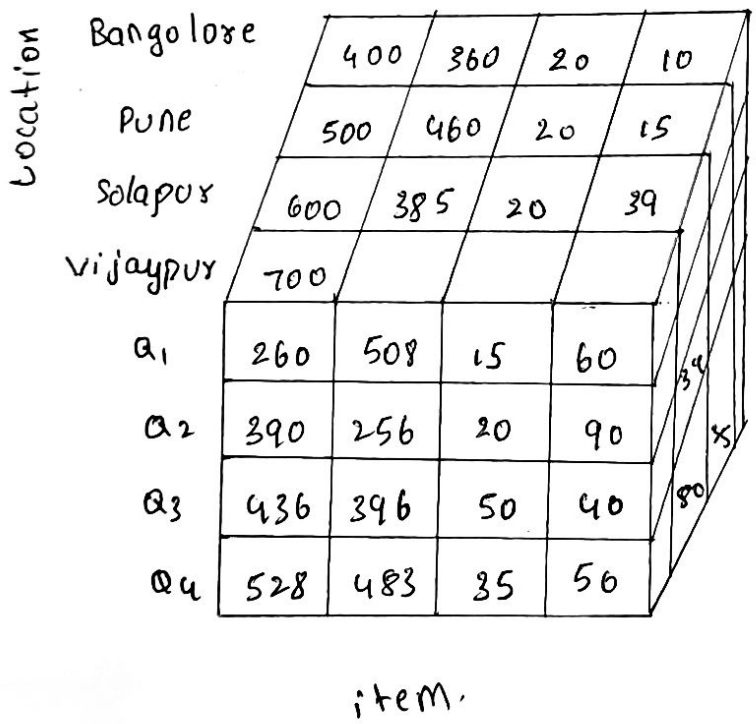
Ans)

- First lest understand market basket analysis
- a data mining technique that gives the careful study of purchase done by a customer is a supermarket.
- This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offer by the company and data mining techinique.
- example:

  Data mining concept are in use for sales and marketing to provide betler customer service, to improue cross-selling opportunity to increase direct mail response rate
- customer retention in the form of paltern identificat and prediction to likely detection is possible is possible by data mining.
- Risk assumed and Fraud area also use the data mining concept for identifying innoppropriate or unusual behaviours etc.
- 2 D data:

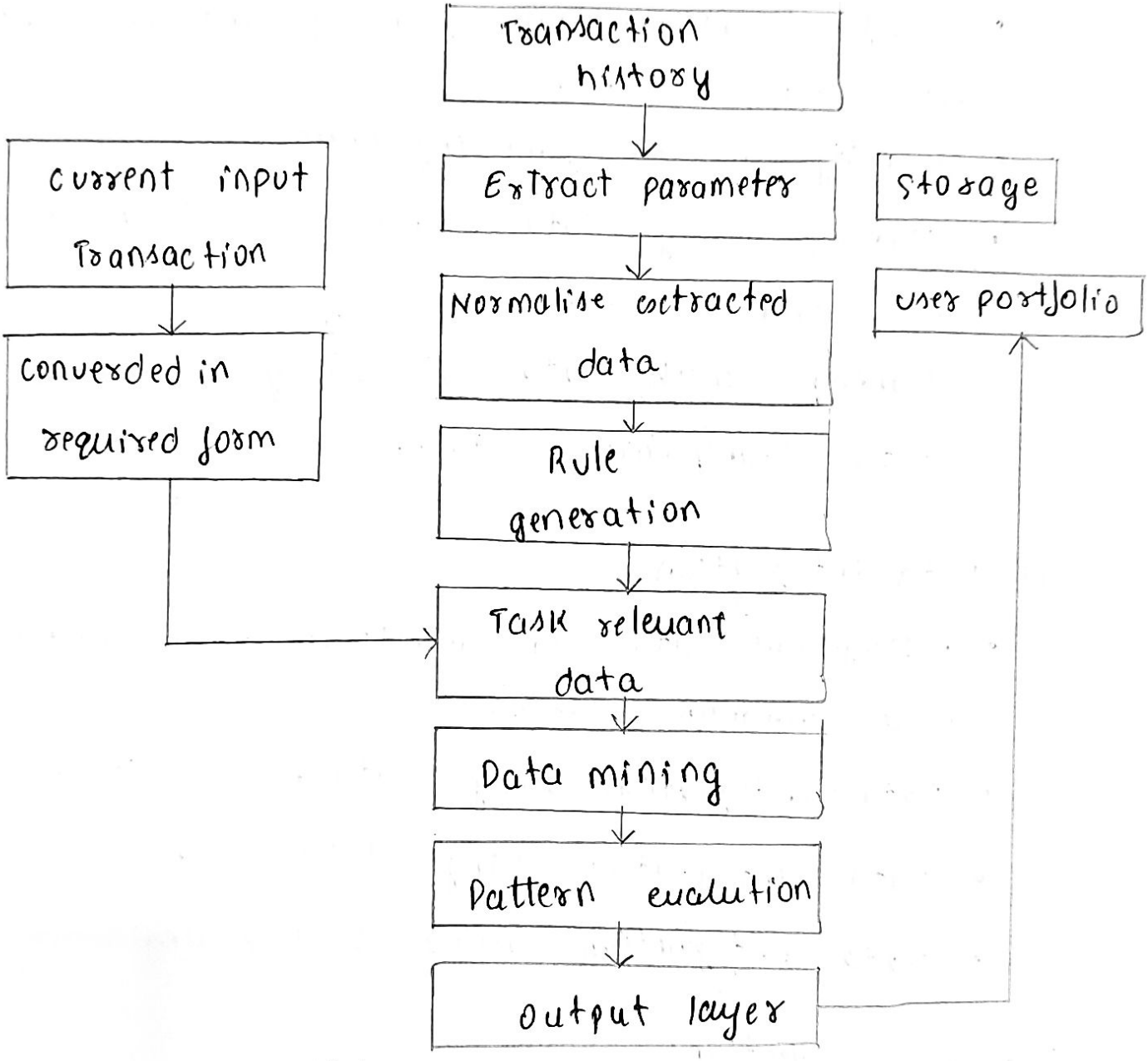
	Bangalore				Pune			
Time	egg	milk	bread	Biscuit	egg	milk	bread	Biscuit
Q_1	400	360	20	10	500	460	20	15
Q_2	300	490	16	50	200	385	45	35
Q_3	200	583	46	43	100	490	39	48
Q_4	100	699	89	38	600	365	83	35

- 3D data:



Module - 2

(1) Design a KDD model for credit card fraud detection



(2) Explain the data preprocessing for information retrival application.

Ans)

(i) Data cleaning:

- is defined as removal of noisy and irrelevant data from collection
- cleaning in case of missing value
- cleaning noisy data, where noise is a random or variance error.
- cleaning with data discrepancy detection and data transformation tools.

(ii) Data integration:

- hetrogenous data from multiple source combined in a common source.
- Data integration using data mining tools
- Data integration using data synchronization
- Data integration using ETL process

(iii) Data selection:

- process where data relevant to the analysis is decided and retained from the data collection
- Data collection using Neural network

- Data selection using decision tree
- Data selection using naive bayes
- Data selection using clustering, regression.

(iv) Data Transformation

- process of transforming data into appropriate required form by mining
- data mapping
- code generation

(v) Data mining:

clever techniques that are applied to extract pattern potentially useful

- Transforms task relevant data into pattern
- Decides purpose of model using classification or characteristics.

(vi) Pattern evaluation:

- identifying strictly increasing pattern representing Knowledge based measures
- Find interesting score of each pattern

(vii) Knowledge representation:

- utilizes visualization tools to represent data mining results.
- generate report
- generate table

(3) Evaluate the statistical description for stock market analysis with data visualization.

(ans) GitHub.