

BLDE Association's

S.B. Arts & K.C.P. Science College,

Vijayapur-586103

(Accredited at the B++ Grade with 2.99 CGPA in 4th cycle by NAAC)



BCA

Programme

Internl Assessment Test Booklet

V

Semester 2022-23

Name : Sneha . S. Bijjavagi

Roll No : 37 R.C.U. Exam Seat No. M2026837

Subject Code : BCA DCS S-2

Subject : Data warehouse & Data mining .

B.L.D.E. Association's

S. B. ARTS & K. C. P. SCIENCE COLLEGE,
VIJAYAPUR

Name of the Student : Sneha S. Bijjavagi

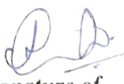
Roll No. : 37 R.C.U. Seat No. : M2016831

Subject Code No. : BCA DSC S.2.

Subject. : Datawarehouse & Mining.

Test	Maximum Marks	Marks Obtained
First Test	20	20
Second Test	20	18
Third Test	—	—
Total (Two best of three)	20	20


Signature of
Faculty


Signature of
Co-Ordinator


Signature of
Principal

20/20

PART-A

18

18 Data warehouse :

Data warehouse defines in many ways that refers to the operations which contains the organizational operational database.

22

2 Data warehouse is a collection of object oriented, integration, variant and non-volatile of data.

24 OLTP :

OLTP refers to online transaction processing

* OLTP is day to day operation and it is having read & write both accuracy.

* It is having highly availability and high performance. Ex: Online transaction (phonepay, etc)

30 ETL :

It refers to Extraction, Transformation and

loading.

42 Data model :

Data model is a simple and it is done by the three types -

1. Star model

2. Snowflake model

3. Fact constellation.

62 Attribute :

Attribute refers to the characteristics or features of data object.

Ex: State, Designation.

8* Visualization:

Visualization is the graphical representation of information and data.

Techniques of Data Visualization

- * pixel-oriented visualization.
- * Hierarchical oriented visualization.
- * Graphical oriented visualization.

5* Applications of Data Mining

- * webpage analysis
- * Biological & medical analysis
- * Data mining & software engineering
- * Past data in marketing

PART-B

11x

4x Attribute :

In Data Mining, attribute refers to the data field characteristics or features of data object.

There are four types of attribute.

1x Nominal attribute

2x Binary attribute

3x Ordinal attribute:

4x Numeric attribute.

1x Nominal attribute :

Nominal attribute refers to the names, and it represent with the symbols and name of the things.

* Nominal attribute is also called as co-ordinal attribute.

Ex: State, Designation.

2x Binary attribute :

Binary attribute is a nominal attribute but it is binary refers to the binary by zero and one where, zero represents the absent and one represent the present.

* Hence, Binary attribute is also called as Boolean attribute.

In Binary attribute there are two types they are:

* symmetric attribute

* asymmetric attribute

Ex:

It Symmetric attribute: Male or female.

It Asymmetric attribute: Medical report Mallesia is +ve and -ve.

It ordinal attribute :

ordinal attribute refers to the grades. it is used when attribute more than two

Ex:

Student grade is A, A+, D

Customer ^{Product} Salesman: Satisfied, Dissatisfied, not Satisfied.

It Numeric attribute :

- Numeric attribute is a quantitative attribute that represents the quantity.
- * Numeric attribute is used integer and floating point.

There are 2 types of numeric attribute:

* Inter-scale attribute.

* Ratio-scale attribute.

1 Inter-scale attributes

In, Inter-scale attribute there is no zero.

Ratio-scale attributes

It borrows the zero from the attribute.

2* components of data warehouse:

Datawarehouse is a collection of object oriented, integrated, non-volatile and variant of the organizational operational database.

In Datawarehouse there are some components they are-

- * Datawarehouse Database.
- * Acquisition, Extraction, Transformation and Load (ETL)
- * Metadata
- * Query tool
- * Database Bus Architecture.

1* Datawarehouse Database:

- * It is the central foundation of the Datawarehouse System.
- * In the third tier architecture, Datawarehouse Database is in the Bottom tier and it uses the Relational database (RDBMS).
- * It extracts the data by using API (Application program interface) usually known as gateway uses the RDBMS and DBMS.
- * In, Datawarehouse Database is having the OLAP Server and it uses theROLAP and MOLAP.

2* Acquisition, Extraction, Transform and Loading (ETL)

- * In Datawarehouse System The data uses the extraction, transformation, cleaning as well loading.
- * The data is entered from the Operational database and External Source.
- * In Datawarehouse the data is acquired, then extracted and the cleaned and it runs

the redundancy and it transformed to the external source and then data is loaded into the datawarehouse.

3b Meta data:

- * Meta data refers to the process of storing the data
- * Meta data is converted into datawarehouse and then it is divided into the small datamarts.

4b Query tool:

- * In the Datawarehouse architecture, the top tier is a front-end tool.
- * In front end tool uses the query tool, analysis tool and datamining tool.
- * Query tool is used to report the data.

5b Database Bus Architecture:

- * It is used to know the flow the data in the datawarehouse.

II- Internal Test

PART-A

1b

1b Frequent pattern :

Frequent patterns are the pattern frequently in the data set.

Ex: Buying a Bread, butter, biscuit is the itemset frequently purchased by the customer is called frequent pattern.

3b Association Rule :

Association rule is used to encapsulate both the minimum support threshold & minimum confidence threshold.

4b Applications of frequent pattern.

- * Subsequence
- * Substructure
- * Confidence
- * Support.

6b WEKA :

- * WEKA refers to the Waikato Environment for Knowledge Analysis.
- * WEKA is a machine learning tool developed by department of computer science in 1994. University of New Zealand.

2b Market Basket Analysis :

- * It is an best example of ItemSet.
- * Mining is market Basket Analysis This process analyse the method for finding the habbit of buying the customer.

3b *Explores

- * Experimental
- * Knowledge flow

PART-B

18 Apriori Algorithm:

Objective:

The main objective of apriori algorithm is for finding the ~~data~~ itemset in the candidate generation.

- * Apriori algorithm is ~~proposed~~ by R. Agrawal & R. Shreekanth in 1994.
- * It is used to explore the dataset.
- * K-itemset is used to explore the K+1 itemset.

Rules

- * Apriori algorithm is used to explore the data Set Transaction ID's and list of id's are represented and it divided to Subset.
- * candidate count is referred by Minimum Support 2.
 $\text{Min_supp} = 2$.
- * It compares the transaction with the candidate key.

Transaction	List of id
T100	I ₁ , I ₂ , I ₅
T200	I ₂ , I ₄
T300	I ₂ , I ₃
T400	I ₁ , I ₂ , I ₄
T500	I ₁ , I ₃
T600	I ₂ , I ₃
T700	I ₁ , I ₃
T800	I ₁ , I ₂ , I ₃ , I ₅
T900	I ₁ , I ₂ , I ₃

C_1

Count the candidate	item	count
	I_1	6
	I_2	7
	I_3	6
	I_4	2
	I_5	2

Count the
min. supp
out

 C_1

item	count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

 L_1

Generate L_1
from candid
-ate C_1

Set of item

 I_1, I_2 I_1, I_3 I_1, I_4 I_1, I_5 I_2, I_3 I_2, I_4 I_2, I_5 I_3, I_4 I_3, I_5 I_4, I_5

Count the
candidate

 C_2

items	count
I_1, I_2	4
I_1, I_3	4
I_1, I_4	1
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2
I_3, I_4	0
I_3, I_5	1
I_4, I_5	0

Count the
min. supp
out

item	count
I_1, I_2	4
I_1, I_3	4
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2

		L ₂	
		Set of items	
		I ₁ , I ₂ , I ₅	
Generate L ₂ from candid-ate C ₂		I ₁ , I ₂ , I ₃	count the candidate →
→			

		C ₃	
Set of items	Count	Count the min-support →	
I ₁ , I ₂ , I ₅	2		
I ₁ , I ₂ , I ₃	2		

		C ₃	
Set of item	Count	Set of item	Count
I ₁ , I ₂ , I ₅	2		
I ₁ , I ₂ , I ₃	2		

* Using Associate Rule, we can find the Support and confidence of the above dataset.

$$P(A|B) \Rightarrow P(B)/P(A) \Rightarrow \frac{P(A \cup B)}{P(B)}$$

So, the support and confidence of I₁, I₂, I₅ of dataset is,

(I ₁ , I ₅): I ₂	confidence: $2/4 = 50\%$
(I ₂ , I ₅): I ₁	confidence: $2/2 = 100\%$
(I ₁ , I ₅): I ₂	confidence: $2/4 = 50\%$
I ₁ ⇒ (I ₂ , I ₅)	confidence: $6/2 = 30\%$
I ₂ ⇒ (I ₁ , I ₅)	confidence: $7/2 = 35\%$
I ₅ ⇒ (I ₁ , I ₂)	confidence: $2/6 = 33\%$

28 FP-Growth:

- * FP growth refers to the frequent pattern growth.
- * Apriori algorithm bases FP-growth algorithm is used to encapsulate the pattern items and pattern tree.
- * FP-Growth is used to high the transaction data.

Rules

- * In this Algorithm, it consists the transaction id's and data item sets.
- * The count value is arranged in ascending order.

	Transaction ID	Set of ID's
	T100	I ₁ , I ₂ , I ₅
	T200	I ₂ , I ₄
	T300	I ₂ , I ₃
	T400	I ₁ , I ₂ , I ₄
	T500	I ₁ , I ₃
	T600	I ₂ , I ₃
/	T700	I ₁ , I ₃
X	T800	I ₁ , I ₂ , I ₃ , I ₅
/	T900	I ₁ , I ₂ , I ₃

Step 2: This algorithm count the value of candidate.

Items	count
I ₁	6
I ₂	7
I ₃	6
I ₄	2
I ₅	2

T100

T200

T300

T400

T500

T600

T700

T800

T900

Wrong

Step 3 :

Arrange the candidate values into Ascending order.

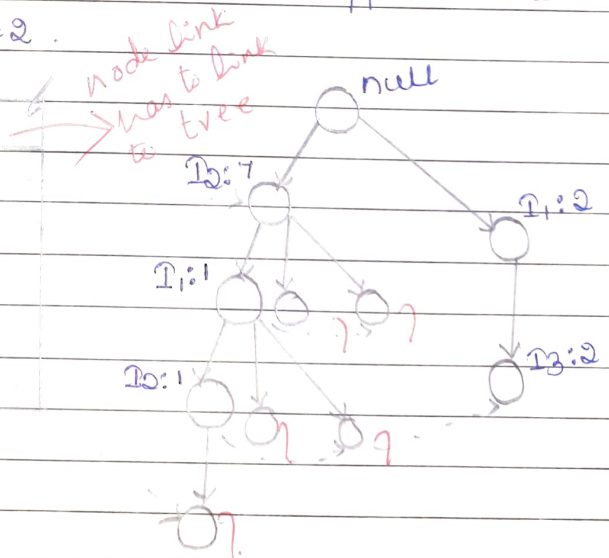
itemset	Supp. count
T₂	7
T ₁	6
T ₃	6
T ₄	2
T ₅	2

Step 3 :

count the minimum support is 2.

min-Supp = 2

item	Supp. Count
T ₂	7
T ₁	6
T ₃	6
T ₄	2
T ₅	2



Step 4:

itemset	count
I_1, I_2, I_5	2
I_1, I_2, I_8	2

BLDE Association's
S.B. Arts & K.C.P. Science College,
Vijayapur-586103

(Accredited at the B++ Grade with 2.99 CGPA in 4th cycle by NAAC)



Bca Programme

Internl Assessment Test Booklet

I Semester 2022-23

Name : Mashai kavari

Roll No : 19 R.C.U. Exam Seat No. M2016819

Subject Code : BCADSL502

Subject : Data warehouse & Mining

BLDE Association's

S.B. Arts & K.C.P. Science College,

Vijayapur-586103

(Accredited at the B++ Grade with 2.99 CGPA in 4th cycle by NAAC)



BCA

Programme

Internl Assessment Test Booklet

I

Semester 2022-23

Name : Mashal kaveri

Roll No : 19 R.C.U. Exam Seat No. M2016819

Subject Code : BCADSL502

Subject : Data warehouse & Mining

20/2/2024

1) Data warehouse.

2) The data warehouse is defined as the collection of Subject oriented, integrated, time variant and non-volatile to collection of data support the managing of decision making process, known as data warehouse.

3) The data source, transformation and migration which is used to perform the conversions, Summarisation and all cleanup. whenever the changed needed data transfer to unified format called as ETL.

4) 1) webpage Analysis
2) collaborative analysis and recommended system
3) Basket data analysis to targeted market.
4) Biological and medical data analysis.

5) Attribute is defined as the data field can be represents by characteristics or features of data object called as attribute.

Ex: 1) dimensions.

It is used in datawarehouse.

2) variable - used in Statistics.

8

Types of attributes.

- 1) Nominal attribute
- 2) Binary attribute
- 3) ordinal attribute
- 4) Numeric attribute

7) mean -

It is set most common and effective center set of data and the Arithmetic mean. let $x_1, x_2, x_3, \dots, x_n$ be the n set value. Sometime the numeric attribute x like. ~~var~~ mean of set of value is $\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$.

2) mode -

It is set of data. the value can occur frequently in the dataset. mode which represents quantitative and qualitative attribute. the mode class known as mode. the mode class can be ^{access} ~~modify~~ with model set or tri mode set.

4) Datamodel

The datamodel is computer in which represents in multidimensional datamodels like Star, Snowflake, or ~~to~~ fact-costillation.

- 1) Star datamodel
- 2) Snowflake datamodel
- 3) fact-costillation datamodel.

2) OLTP — online transaction processing

→ OLTP defines it is customer oriented which used in transactional query processing like client, Ecommerce and IT professionals.

→ OLTP can manage only the current data is too detailed.

Ex: For Googlepay or Amazon.

8) data visualization

Aims to communicate the clearly and effectively in a graphical representation. The data visualization includes the managing business operation, and reporting etc.

It includes

1) pixel-orientation

2) Geometric presentation

3) Icon Based

4) hierarchy

Part - B

1) Users and System orientation.

- * The OLTP is customer oriented which used in transactional query processing like client, clerk and IT professionals.
- * OLAP is a subject oriented which used in data analysis ^{with} knowledge workers like managers and analyst.

2) Data content

- * The OLTP which manage the current data in a too detailed.
- * The OLAP which manage the historical data.

3) Database design.

- * The OLTP system which adapts the entity relationship datamodel.
- * It is application oriented.
- * The OLAP system which adapts the multidimensional model.
- * It is ~~is~~ Subject oriented.

4) View

- * The OLTP ~~wh~~ system focuses on the current data within the department.
- * The OLAP System focus on Summarization and multidimensional.

5) Access pattern.

• The OLTP system which consist the automatically transaction like concurrency control, etc.

• The OLAP system which consist only read and write.

O'	OLTP	OLAP
1) orientation users	client, clerk and IT professional	knowledge workers like managers and analyst.
2) features	day-to-day transaction	long term information.
3) orientation	transaction	analysis.
4) Data Content	current data manage	manage historical data.
5) Database design	PR- ^{based} data model it is application oriented.	multidimensional model • Subject oriented
6) Unit of work	Short, Small amount of transaction	• complex query.
7) Access operation	read / write.	lots of scans
8) No. of units	ten	million
9) Priority	highly probability and highly availability	highly flexible and end users autonomy.
10) metrics	transaction throughput	query throughput.

4) Types of attribute.

→ The attribute is defined as it is data field which represents the characteristics or features of data object.

1) Nominal attribute.

2) Binary "

3) ordinal

4) Numerical attribute

1) Nominal attribute.

The nominal attribute is related to the name the value of nominal attribute is symbols and name of the thing, the each value is ~~related~~ in kind of category is called as category attribute.

Ex: state

2) Binary attribute

The binary attribute is nominal attribute which contains the two states like 0 and 1. 0 can represent absent and 1 can represent present. is known as boolean attribute. i.e. true or false.

It has ~~symmetric~~ and asymmetric attribute.

1) Symmetric attribute

Gender state like male or female.

2) asymmetric attribute.

medical test material like positive or negative.

3) Ordinal attribute.

The ordinal attribute is attribute in which represents any possible values with meaningful order in ranking.

Ex: Grade A A like A, A+, A++, O

4) Numerical attribute.

Numerical attribute is quantitative which the value can measure in quantities, which it represents in integers and Real numbers. Numerical attribute have Interval Scaled and Ratio Scaled attribute.

1) Interval Scaled attribute

It is a numeric attributes have there zero is no zero point exist. It can be compute the mean values. an additional mode and median to measures central tendency.

2) Ratio scaled attribute.

It is a numeric attribute inherit the zero point. It can speak only the value of other type multiple value. we can compute the value and also difference between mode, mean and median.

19/20

II - Internal.

1) Frequent patterns are the patterns appears frequently in dataset called as frequent pattern.

Ex: Itemset milk, bread, butter are the frequent together of data set.

2) Mining is a market Basket analysis. This process is analysing for customer buying habit. with association with different itemset frequently itemset purchased together by customer.

The discovery of these association which help the relating develop of market strategies by gaining insight into which items shows

Ex: Customer who ^{purchased} buys the computer and also tend to buy the anti viruses at time.

Computer $\longrightarrow$ Antivirus
 $\downarrow$
implies.

Support 2

confidence 60%.

4) classification

clustering

Recommender

Market Basket analysis

Software debug detection.

6) multilevel association rule can be mined by efficiency of support and confidence of itemset.

6) Waikato Environment of Knowledge Analysis.

→ Weka is a data mining or machine learning tool which developed by department of Computer Science University of Waikato New Zealand.

→ wai weka is a bird.

7) 1) Explorer

2) Experimenter

3) Knowledge flow

4) Workbench

5) Simple GUI

Part-B

- The basic objective of algorithm for finding frequent itemset, combined candidate generate.
- It is seminal algorithm which proposed by R. Agrawal and R. in 1994.
- This algorithm is based on the fact and use the apriori knowledge frequent pattern, too.
- Apriori Employee is iterative approach is called level wise Search.
- k item Set explore the $k+1$ itemset.
- To improve the efficiency of level-wise-generation of frequent itemsets, the Apriori pattern is used to reduce the space.
- It has two properties prun and join.

Algorithm:

- each item is a member of Set of candidate itemset then we can call it c_1 . Scan the all transaction and count the number of items.
- Suppose the minimum Support count m then term is $\text{min-support} = 2$. The set of one item is L_1 and for Set of frequent two we can use join operation.

$$L_1 \bowtie L_2$$

Tid	Item Set
T100	I ₁ , I ₂ , I ₃ , I ₅
T200	I ₂ , I ₄
T300	I ₂ , I ₃
T400	I ₁ , I ₂ , I ₄
T500	I ₁ , I ₃
T600	I ₂ , I ₃
T700	I ₁ , I ₃
T800	I ₁ , I ₂ , I ₃ , I ₅
T900	I ₁ , I ₂ , I ₃

C1

L1

Scan D of count for each candidate	Itemset	sup-count	candidate support count associated with min-support count	Itemset	sup-count
	{I ₁ }	6		{I ₁ }	6
	{I ₂ }	7		{I ₂ }	7
	{I ₃ }	6		{I ₃ }	6
	{I ₄ }	2		{I ₄ }	2
	{I ₅ }	2		{I ₅ }	2

C2

C2

Generate C2 from L1	Itemset	sup-count	Generate C2 from L1	Itemset	sup-count
	{I ₁ , I ₂ }			{I ₁ , I ₂ }	4
	{I ₁ , I ₃ }			{I ₁ , I ₃ }	4
	{I ₁ , I ₄ }			{I ₁ , I ₄ }	1
	{I ₁ , I ₅ }			{I ₁ , I ₅ }	2
	{I ₂ , I ₃ }			{I ₂ , I ₃ }	4
	{I ₂ , I ₄ }			{I ₂ , I ₄ }	2
	{I ₂ , I ₅ }			{I ₂ , I ₅ }	2
	{I ₃ , I ₄ }			{I ₃ , I ₄ }	0
	{I ₃ , I ₅ }			{I ₃ , I ₅ }	1
	{I ₄ , I ₅ }			{I ₄ , I ₅ }	0

L2

candidate count also associated with min-Support count →

Item Set	Sup-count
{I ₁ , I ₂ }	4
{I ₁ , I ₃ }	4
{I ₁ , I ₅ }	2
{I ₂ , I ₃ }	4
{I ₂ , I ₄ }	2
{I ₂ , I ₅ }	2

C3

C3

Generate C3 From L2	Item Set	Scan D of count for each candidate	Item set	Sup-count
	{I ₁ , I ₂ , I ₅ }		{I ₁ , I ₂ , I ₅ }	2
	{I ₁ , I ₂ , I ₃ }		{I ₁ , I ₂ , I ₃ }	2

L3

candidate count associated with min support count →

Item set	Sup-count
{I ₁ , I ₂ , I ₅ }	2
{I ₁ , I ₂ , I ₃ }	2

> Fp growth algorithm.

→ Frequent pattern growth algorithm is a method for finding the frequent pattern without candidate generation.

→ It supports the divide and conquer method.

procedure FP growth.

- 1) First scan the database is similar to Apriori algorithm for minimum support count is 2.
- 2) Resulting Sort the list in descending order denoted $L = \{I_2:7, I_1:6, I_3:6, I_4:2, I_5:2\}$
- 3) Create a root tree labelled with null.
- 4) Scan the database D in second time, then transaction procedure in order of list and bench is created.

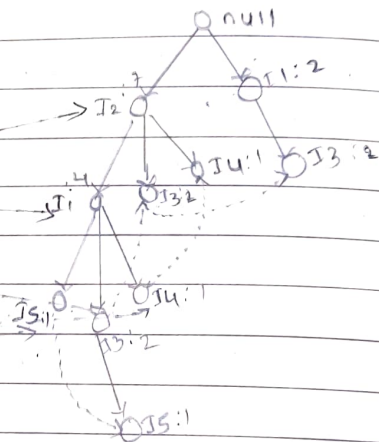
TID	Item Set
T100	I_2, I_1, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	I_2, I_1, I_4
T500	I_1, I_3
T600	I_2, I_3
T700	I_1, I_3
T800	I_2, I_1, I_3, I_5
T900	I_2, I_1, I_3

8) Scan the D table with minimum support count.

Itemset	Sup. Count
$\{I_2\}$	7
$\{I_1\}$	6
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	2

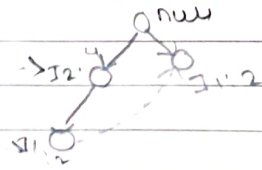
9) Create a fp-tree with null

I-Set	S-count	
$\{I_2\}$	7	
$\{I_1\}$	6	
$\{I_3\}$	6	
$\{I_4\}$	2	
$\{I_5\}$	2	



Item	conditional pattern base	condition FP-tree	frequent pattern generate
I_5	$\{I_2, I_1:1\} \{I_2, I_1, I_3:1\}$	$\{I_2:2\} \{I_1:2\}$	$\{I_2, I_5\} \{I_2, I_1, I_5\}$
I_4	$\{I_2, I_1:1\} \{I_2:1\}$	$\{I_2:2\}$	$\{I_2, I_4\}$
I_3	$\{I_2, I_1:2\} \{I_1:2\} \{I_2:2\}$	$\{I_2:4\} \{I_2:2\} \{I_1:2\} \{I_1:2\}$	
I_1	$\{I_2:4\}$	$\{I_2:4\}$	

q1 (ms)	min	max
T2	4	
T1	4	



4) mining process can uncover the 1000 of rules of given data set.

→ often ~~reeds~~ the users have a good sense for ~~diff~~ dimension of mining get interesting of pattern form want they find.

→ users have good sense of conditions of rule

→ Thus, The interest of users specify Expected Such as constraints. it is confined search sp

Knowledge:

association, correlation, classification.

data constraints

task relevant method

multidimensional.

multilevel, ~~at~~ abstract level.

Interestness

correlation. Support, confidence.

Rule - min, max, attribute equality

