5 UNIT

Data Visualization and Overall Perspective

CONTENTS

PART-1

Aggregation, Historical Information.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.1. **What do you mean by the term aggregation ?**

Answer

1. Data aggregation is a process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis.
2. The purpose is to get more information about particular groups based on specific variables such as age, profession, or income.
3. Data aggregation may be performed manually or through specialized software.

PART-2

Query Facility, OLAP Function and Tools, OLAP Servers, ROLAP, MOLAP, HOLAP

CONCEPT OUTLINE

- OLAP is an acronym for Online Analytical Processing and it performs multidimensional analysis of business data.
- Various OLAP servers are :
  1. ROLAP
  2. MOLAP
  3. HOLAP

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.2. **Explain OLAP in detail.**

Answer

1. OLAP (Online Analytical Processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view.
2. OLAP allows users to analyze database information from multiple database systems at one time.
3. OLAP data is stored in multidimensional databases. OLAP processing is often used for data mining.
4. There are the following key features of OLAP :
   i. Multidimensional views of data
   ii. Support for complex calculations
   iii. Time intelligence
5. Applications of OLAP are :
   i. OLE DB for OLAP
   ii. Marketing and sales analysis
   iii. Consumer goods industries
   iv. Financial services industry (insurance, banks etc.)
   v. Database marketing

Que 5.3. **Explain different types of OLAP operations.**

Answer

Different types of OLAP operations are :

1. **Roll-up :** Roll-up is also known as "consolidation" or "aggregation". In the roll-up process at least one or more dimensions need to be removed. The roll-up operation can be performed in two ways :
   i. Reducing dimensions
   ii. Climbing up concept hierarchy
2. **Drill-down :** In drill-down data is fragmented into smaller parts. It is the opposite of the roll-up process. It can be done via moving down the concept hierarchy and increasing a dimension.
3. **Slice :** One dimension is selected, and a new sub-cube is created.
4. **Dice :** This operation is similar to a slice. In Dice, we select two or more dimensions that result in the creation of a sub-cube.
5. **Pivot :** In pivot, we rotate the data axes to provide a substitute presentation of data.

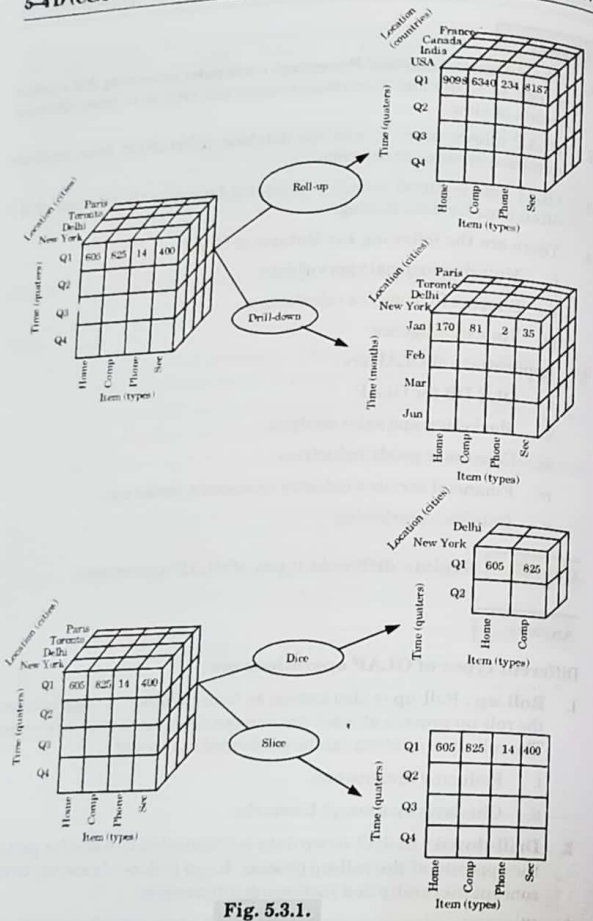

Fig. 5.3.1.

Que 5.4. **Explain the various types of OLAP servers. What are the steps for efficient processing of OLAP queries ?**

AKTU 2015-16, Marks 10

OR

Diagrammatically illustrate and discuss the architecture of MOLAP and ROLAP.

AKTU 2016-17, Marks 10

OR

Explain how query performance can be improved by cascading the operations.

AKTU 2015-16, Marks 10

Answer

Types of OLAP servers are :

1. **Relational OLAP :** ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP architecture : ROLAP includes the following components :

1. Database server
2. ROLAP server
3. Front-end tool

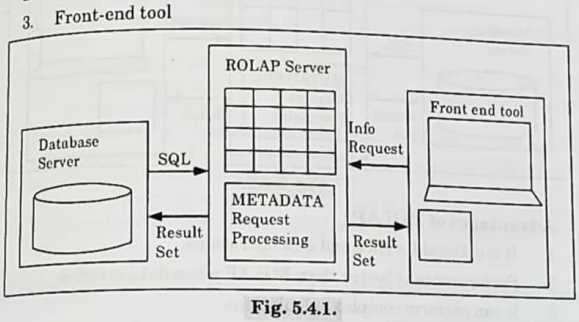

Fig. 5.4.1.

Advantages of ROLAP :

1. It can handle large amounts of data.
2. It can leverage functionalities inherent in the relational database.
3. ROLAP servers can be easily used with existing RDBMS.
4. Data can be stored efficiently, since no zero facts can be stored.
5. ROLAP tools do not use pre-calculated data cubes.
6. DSS server of micro-strategy adopts the ROLAP approach.

Disadvantages of ROLAP :

1. Performance can be slow
2. Limited by SQL functionalities
3. Hard to maintain aggregate tables

2. **Multidimensional OLAP :**
   i. MOLAP stores in optimized multidimensional array storage, rather than in a relational database.
   ii. With multidimensional data stores, the storage utilization may be low if the data set is sparse.
   iii. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

MOLAP architecture : MOLAP includes the following components :

1. Database server
2. MOLAP server
3. Front-end tool

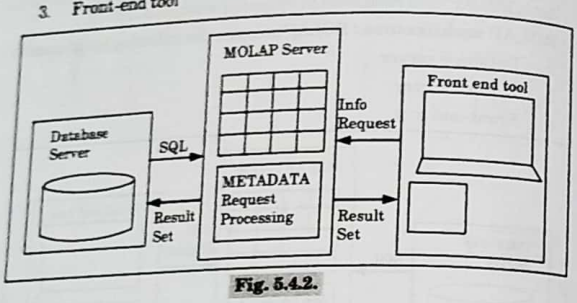

Fig. 5.4.2.

Advantages of MOLAP :

1. It is optimal for slice and dice operations.
2. Performance is better than ROLAP when data is dense.
3. It can perform complex calculations.
4. MOLAP allows fastest indexing to the pre-computed summarized data.
5. Helps the users connected to a network who need to analyze larger, less-defined data.
6. Easier to use, therefore MOLAP is suitable for inexperienced users.

Disadvantages of MOLAP :

1. Difficult to change dimension without re-aggregation.
2. MOLAP can handle limited amount of data.
3. Some MOLAP methodologies introduce data redundancy.
4. Requires additional investment.

3. **Hybrid OLAP :**
   i. Hybrid OLAP is a combination of both ROLAP and MOLAP.

   ii. It offers higher scalability of ROLAP and faster computation of MOLAP.
   iii. HOLAP server allows storing large data volumes of detailed information.
   iv. The aggregations are stored separately in MOLAP store.

Advantages of HOLAP :

1. HOLAP provides advantages of both MOLAP and ROLAP.
2. It provides fast access at all levels of aggregation.

Disadvantages of HOLAP : HOLAP architecture is very complex because it support both MOLAP and ROLAP servers.

Steps for efficient processing of OLAP queries :

Processing of OLAP queries :

To speed up the query processing in data cubes, the cuboids are materialized and OLAP index structures are constructed with following procedure :

1. **Determining which operation should be performed on the available cuboids :**
   a. This involves transformation of operations specified in the query into the corresponding SQL and/or OLAP operators.
   b. These operations include roll-up, drill-down, projection, selection, etc.
   c. For example, slicing and dicing operation on data cube can be transformed into selection and/or projection operations on materialized cuboids.
2. **Determining on which materialized cuboids(s) the relevant operations should be applied :** In this, all of the materialized cuboids are identified which may be useful for answering the query, pruning the relationships among the cuboids, estimating the cost of using the remaining materialized cuboids and selecting the cuboids with the least cost.

Que 5.5. **Define and describe the basic similarities and differences among ROLAP, MOLAP and HOLAP.** **AKTU 2014-15, Marks 10** **AKTU 2015-16, Marks 7.5**

OR

Compare MOLAP vs HOLAP. **AKTU 2013-14, Marks 05**

OR

Write a short note on ROLAP vs MOLAP. **AKTU 2017-18, Marks 2.5**

Answer

Similarities between ROLAP, MOLAP and HOLAP : These three OLAP servers are used to implement data warehouses, and they are related to the logical model used to represent data.

Differences between ROLAP, MOLAP and HOLAP :

S. No.	Basis	ROLAP	MOLAP	HOLAP
1.	Storage location for detail data	Relational database	Multidimensional database	Relational database
2.	Storage location for summary aggregations	Relational database	Multidimensional database	Multidimensional database
3.	Storage space requirement	Large	Medium	Small
4.	Query-response time	Slow	Fast	Medium
5.	Processing time	Slow	Fast	Fast
6.	Latency	Low	High	Medium

Que 5.6. **Give E.F. Codd's 12 guidelines for OLAP.**

AKTU 2013-14, Marks 10

Answer

Dr. E.F. Codd the father of the relational model, created a list of rules to deal with the OLAP systems.

1. **Multidimensional conceptual view :** The OLAP should provide an appropriate multidimensional business model that suits the business problems and requirements.
2. **Transparency :** The OLAP tool should provide transparency to the input data for the users.
3. **Accessibility :** The OLAP tool should only access the data required only to the analysis needed.
4. **Consistent reporting performance :** The size of the database should not affect in any way the performance.
5. **Client/server architecture :** The OLAP tool should use the client server architecture to ensure better performance and flexibility.
6. **Generic dimensionality :** Data entered should be equivalent to the structure and operation requirements.
7. **Dynamic sparse matrix handling :** The OLAP too should be able to manage the sparse matrix and so maintain the level of performance.
8. **Multi-user support :** The OLAP should allow several users working concurrently to work together.

9. **Unrestricted cross-dimensional operations :** The OLAP tool should be able to perform operations across the dimensions of the cube.
10. **Intuitive data manipulation :** Data manipulation inherent in the consolidation path, such as drilling down or zooming out, should be accomplished via direct action on the analytical model's cells, and not require use of a menu or multiple trips across the user interface.
11. **Flexible reporting :** It is the ability of the tool to present the rows and column in a manner suitable to be analyzed.
12. **Unlimited dimensions and aggregation levels :** This depends on the kind of business, where multiple dimensions and defining hierarchies can be made.

PART-3

Data Mining Interface, Security, Backup and Recovery.

CONCEPT OUTLINE

- Data Mining Interface (DMI) is a web-based, interactive, dynamic report building module.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.7. **Write a short note on data mining interface.**

AKTU 2013-14, Marks 05

OR

Describe data mining interface in details. **AKTU 2014-15, Marks 05**

Answer

Data mining interface provides the medium that allows users to communicate with data mining processes. It is difficult to use data mining query languages. A graphical user interface (GUI) can be used to communicate with data mining systems.

A data mining interface may consist of following functional components :

i. **Data collection and data mining query composition :** It allows user to specify task relevant data sets and to compose data mining queries.

ii. **Presentation of discovered patterns :** It allows the display of discovered patterns in various forms like tables, graphs, charts, and other visualization techniques.

iii. **Hierarchy specification and manipulation :** It allows to do the specification of concept hierarchy, either manually or automatically.

iv. **Manipulation of data mining primitives :** It allows the dynamic adjustment of data mining operations like selection, display, and modification of concept hierarchies.

v. **Interactive multilevel mining :** It allows the roll-up or drill-down operations on discovered patterns.

The design of data mining interface should also consider the different classes of users. Users of data mining system can be classified into two categories : business analysts and business executives.

Que 5.8. **Write short note on backup and recovery.**

AKTU 2013-14, Marks 05

OR

Explain different backup and recovery models in data warehousing.

AKTU 2014-15, Marks 10

Answer

1. Backup and recovery refers to the process of backing up data in case of a loss and setting up systems that allow data recovery due to data loss.
2. A data warehouse is a complex system and it contains a huge volume of data.
3. Therefore, it is important to backup all the data so that it becomes available for recovery in future as per requirement.
4. Some of the backup terminologies are :
   a. **Complete backup :** It backup the entire database at the same time.
   b. **Partial backup :** Partial backup is very useful because various parts of the database are backed up in a round-robin fashion on a day-to-day basis.
   c. **Cold backup :** Cold backup is taken when the database is completely shut down.
   d. **Hot backup :** Hot backup is taken when the database engine is up and running.
   e. **Online backup :** It is quite similar to hot backup.

Following are different backup and recovery models :

1. **Full recovery model :** It provides the most flexibility for recovering database to an earlier point in time.

2. **Bulk-logged recovery model :** Bulk-logged recovery provides higher performance than lower log space consumption for certain large scale operations.
3. **Simple recovery model :** Simple recovery provides the highest performance and lowest log space consumption but with the significant exposure to data loss in the event of a system failure.

Que 5.9. **How data backup and data recovery is managed in data warehouse ?**

AKTU 2017-18, Marks 10

Answer

1. Managing the recovery of a large data warehouse is a difficult task and traditional OLTP backup and recovery strategies may not meet the needs of a data warehouse.
2. We should plan a backup strategy as part of our system design and consider what to backup and how frequently to backup.
3. The most important variables in our backup design are the amount of resources we have to perform a backup or recovery and the recovery time objective.
   a. NOLOGGING operations must be taken into account when planning a backup and recovery strategy. Traditional recovery, restoring a backup and applying the changes from the archive log, does not apply for NOLOGGING operations.
   b. Never make a backup when a NOLOGGING operation is taking place.
   c. Plan for one of the following or a combination of the following strategies :
      i. **The ETL strategy :** Recover a backup that does not contain non-recoverable transactions and replay the ETL that has taken place between the backup and the failure.
      ii. **The incremental backup strategy :** Perform a backup immediately after a non-recoverable transaction has taken place.

Strategies and best practices for backup and recovery : The following best practices can help us to implement our warehouse's backup and recovery strategy :

1. Use ARCHIVELOG mode
2. Use RMAN mode
3. Use read-only tablespaces
4. Plan for NOLOGGING operations
5. Not all tablespaces are equally important

PART-4

Tuning Data Warehouse and Testing Data Warehouse.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.10. Explain tuning in data warehouse.

Answer

1. Tuning in data warehouses are the processes of selecting adequate optimization techniques in order to make queries and updates run faster.
2. A data warehouse is usually accessed by complex queries for key business operations.
3. Therefore it becomes more difficult to tune a data warehouse system. The tuning of data warehouse can be done to improve the performance.
4. **Difficulties in data warehouse tuning are :**
   a. Data warehouse is dynamic; it never remains constant.
   b. It is very difficult to predict what query the user is going to post in the future.
   c. Business requirements change with time.
   d. Users and their profiles keep changing.

Que 5.11. Write a short note on testing data warehouse.

AKTU 2013-14, 2014-15; Marks 05

Answer

Testing is very important for data warehouse systems to make them work correctly and efficiently. There are three basic levels of testing performed on a data warehouse :

1. **Unit testing :** In unit testing, each component is separately tested. Each module, *i.e.*, procedure, program, SQL Script, Unix shell is tested. This test is performed by the developer.

2. **Integration testing :** In integration testing, the various modules of the application are brought together and then tested against the number of inputs. It is performed to test whether the various components do well after integration.
3. **System testing :** In system testing, the whole data warehouse application is tested together. The purpose of system testing is to check whether the entire system works correctly together or not. System testing is performed by the testing team.

Challenges of data warehouse testing are :

1. Data selection from multiple source and analysis that follows pose great challenge.
2. Volume and complexity of the data.
3. Redundant data in a data warehouse.
4. Inconsistent and inaccurate reports.

ETL testing is performed in five stages :

1. Identifying data sources and requirements
2. Data acquisition
3. Implement business logics and dimensional modeling
4. Build and populate data
5. Build reports

PART-5

Warehousing Applications and Recent Trends : Types of Warehousing Applications.

CONCEPT OUTLINE

- Applications of data warehouse are :
  i. Airline
  ii. Banking
  iii. Healthcare
  iv. Public sector

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.12. **What are the applications of data warehousing ?**

AKTU 2016-17, Marks 05

Answer

Applications of data warehousing are :

1. **Airline :** In the Airline system, it is used for operation purpose like crew assignment, analysis of route profitability, frequent flyer program promotions, etc.
2. **Banking :** It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for the market research, performance analysis of the product and operations.
3. **Healthcare :** Healthcare sector also used data warehouse to strategize and predict outcomes, generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.
4. **Public sector :** In the public sector, data warehouse is used for intelligence gathering. It helps government agencies to maintain and analyze tax records, health policy records, for every individual.
5. **Investment and insurance sector :** In this sector, the warehouses are primarily used to analyze data patterns, customer trends, and to track market movements.
6. **Retain chain :** In retail chains, data warehouse is widely used for distribution and marketing. It also helps to track items, customer buying pattern, promotions and also used for determining pricing policy.
7. **Telecommunication :** A data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.
8. **Hospitality industry :** This industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

PART-6

Web Mining Spatial Mining and Temporal Mining.

CONCEPT OUTLINE

- Web mining is of three types :
  - i. Web content mining
  - ii. Web usage mining

  - iii. Web structure mining
- Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases.
- Temporal data mining refers to the extraction of implicit, non-trivial, and potentially useful abstract information from large collections of temporal data.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.13. **What is web mining ? Differentiate between web content mining, web structure mining and web usage mining.**

AKTU 2016-17, Marks 10

Answer

Web mining :

1. Web mining is an application of data mining techniques to find information patterns from the web data.
2. Web mining helps to improve the power of web search engine by identifying the web pages and classifying the web documents.
3. There are three types of web mining :
   a. **Web content mining :** Web content mining can be used for mining of useful data, information and knowledge from web page content. Web content mining performs scanning and mining of the text, images and groups of web pages according to the content of the input (query), by displaying the list in search engines.
   b. **Web usage mining :** Web usage mining is used for mining the web log records (access information of web pages) and helps to discover the user access patterns of web pages.
   c. **Web structure mining :** The web structure mining can be used to discover the link structure of hyperlink. The purpose of structure mining is to produce the structural summary of website and similar web pages.

Que 5.12. What are the applications of data warehousing ?

AKTU 2016-17, Marks 05

Answer

Applications of data warehousing are :

1. **Airline :** In the Airline system, it is used for operation purpose like crew assignment, analysis of route profitability, frequent flyer program promotions, etc.
2. **Banking :** It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for the market research, performance analysis of the product and operations.
3. **Healthcare :** Healthcare sector also used data warehouse to strategize and predict outcomes, generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.
4. **Public sector :** In the public sector, data warehouse is used for intelligence gathering. It helps government agencies to maintain and analyze tax records, health policy records, for every individual.
5. **Investment and insurance sector :** In this sector, the warehouses are primarily used to analyze data patterns, customer trends, and to track market movements.
6. **Retain chain :** In retail chains, data warehouse is widely used for distribution and marketing. It also helps to track items, customer buying pattern, promotions and also used for determining pricing policy.
7. **Telecommunication :** A data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.
8. **Hospitality industry :** This industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

PART-6

Web Mining Spatial Mining and Temporal Mining.

CONCEPT OUTLINE

- Web mining is of three types :
  - i. Web content mining
  - ii. Web usage mining
  - iii. Web structure mining
- Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases.
- Temporal data mining refers to the extraction of implicit, non-trivial, and potentially useful abstract information from large collections of temporal data.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.13. What is web mining ? Differentiate between web content mining, web structure mining and web usage mining.

AKTU 2016-17, Marks 10

Answer

Web mining :

1. Web mining is an application of data mining techniques to find information patterns from the web data.
2. Web mining helps to improve the power of web search engine by identifying the web pages and classifying the web documents.
3. There are three types of web mining :
   a. **Web content mining :** Web content mining can be used for mining of useful data, information and knowledge from web page content. Web content mining performs scanning and mining of the text, images and groups of web pages according to the content of the input (query), by displaying the list in search engines.
   b. **Web usage mining :** Web usage mining is used for mining the web log records (access information of web pages) and helps to discover the user access patterns of web pages.
   c. **Web structure mining :** The web structure mining can be used to discover the link structure of hyperlink. The purpose of structure mining is to produce the structural summary of website and similar web pages.

Difference :

Criterion	Web content mining		Web structure mining	Web usage mining
	IR view	DB view		
View of data	a. Unstructured b. Structured	a. Semi-structured b. Website as DB	a. Link structure	a. Interactivity
Main data	a. Text documents b. Hypertext documents	a. Hypertext documents	a. Link structure	a. Server logs b. Browser logs
Representation	a. Bag of words, *n*-gram terms b. Phrases, concepts or ontology c. Relational	a. Edge labeled graph b. Relational	a. Graph	a. Relational table b. Graph
Method	a. Machine learning b. Statistical (including NLP)	a. Proprietary algorithms b. Association rules	a. Proprietary algorithms	a. Machine learning b. Statistical c. Association rules
Applications categories	a. Categorization b. Clustering c. Finding extract rules d. Finding patterns in text	a. Finding frequent sub structures b. Web site schema discovery	a. Categorization b. Clustering	a. Site construction b. Adaptation and management

Que 5.14. Write a short note on spatial and temporal data mining.

Answer

Spatial mining :

1. Spatial data mining is the application of data mining to spatial models.
2. In spatial data mining, analysts use geographical or spatial information to produce business intelligence or other results.
3. Challenges involved in spatial data mining include identifying patterns or finding objects that are relevant to the research project.

Temporal mining :

1. Temporal data mining is a single step in the process of knowledge discovery in temporal databases that enumerates structures over the temporal data.
2. Temporal data mining is concerned with the analysis of temporal data and for finding temporal patterns and regularities in sets of temporal data tasks of temporal data mining are :
   a. Data characterization and comparison
   b. Cluster analysis
   c. Classification
   d. Association rules
   e. Prediction and trend analysis
   f. Pattern analysis

Que 5.15. Compare and contrast spatial, temporal mining with relevant examples. **AKTU 2016-17, Marks 15**

Answer

S.No.	Spatial mining	Temporal mining
1.	Spatial mining is the extraction of knowledge/ spatial relationships and interesting measures that are not explicitly stored in spatial database.	Temporal mining is the extraction of knowledge about occurrence of an event or values whether they follow cyclic, random, seasonal variations etc.
2.	It deals with spatial (location, geo-referenced) data.	It deals with implicit or explicit temporal content, from large quantities of data.

3.	It includes finding characteristic rules, discriminant rules, association rules and evaluation rules etc.	It aims at mining new and unknown knowledge, which takes into account the temporal aspects of the data.
4.	**For example :** Determining hotspots, unusual locations.	**For example :** An association rule which looks like – "Any person who buys a car also buys steering lock". By temporal aspect, this rule would be "Any person who buys a car also buys a steering lock after that".

☺☺☺

1 UNIT

Data Warehousing (2 Marks Questions)

1.1. Briefly explain important approaches to build the data warehouse. **AKTU 2015-16, Marks 02**

Ans. Two approaches to build a data warehouse are :

1. **Top-down approach :** In the top-down approach, data warehouse is built first. The data marts are then created from the data warehouse.
2. **Bottom-up approach :** In the bottom-up approach, data marts are created first and then data warehouse is built.

1.2. Why data warehouse is maintained separately from database ? **AKTU 2015-16, Marks 02**

Ans.

1. An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contract, data warehouse queries are often complex and they present a general form of data.
2. An operational database query allows to read and modify operations, while an OLAP query needs only read only access of stored data.

1.3. How is the data warehouse different from a database ? **AKTU 2016-17, Marks 02**

Ans.

S.No.	Data warehouse	Database
1	It involves historical processing of information.	It involves day-to-day processing.
2.	It is used to analyze the business.	It is used to run the business.
3.	It focuses on information out.	It focuses on data in.
4.	It contains historical data.	It contains current data.