1 UNIT

Data Warehousing

CONTENTS

PART-1

Overview, Definition.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.1. **What do you mean by data warehouse ? Discuss its key features with suitable example.**

Answer

1. A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources.
2. A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels.

Key features of data warehouse are :

a. **Subject-oriented :** A data warehouse can be used to analyze a particular subject area. For example, sales, marketing, etc can be a particular subject.

b. **Integrated :** A data warehouse integrates data from multiple data sources. For example, application *A* and *B* stores information in different way, but in a data warehouse, all this information is stored in common format.

c. **Time-variant :** Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse *i.e.*, a data warehouse can hold all addresses associated with a customer.

d. **Non-volatile :** Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it. Data is read-only and periodically refreshed. For example, changes in transaction status.

e. **Data granularity :** Data granularity can be defined as the level of details of data. In OLTP, the data granularity is the number of units for each unique product. For example, user will look the sale of products across all the stores, which region have recorded the maximum sale, which store in that region has given maximum sale etc.

PART-2

Data Warehousing Components.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.2. **Describe the components of data warehouse.**

Answer

Following are the various components of data warehouse :

a. **Data warehouse database :** The database is implemented on the RDBMS technology. Due to some constraints, different approaches to database are used :

1. RDBMS are deployed in parallel to allow scalability.
2. New index structures are used to bypass relational table scans and improve speed.
3. Multidimensional databases are used to overcome any limitations due to relational data model.

b. ***ETL tools :*** *The functionality of sourcing, acquisition, cleanup and transformation tools also called as ETL tools includes :*

1. *Removing unwanted data from operational databases.*
2. *Converting to common data names and definitions.*
3. *Establishing defaults for missing data.*

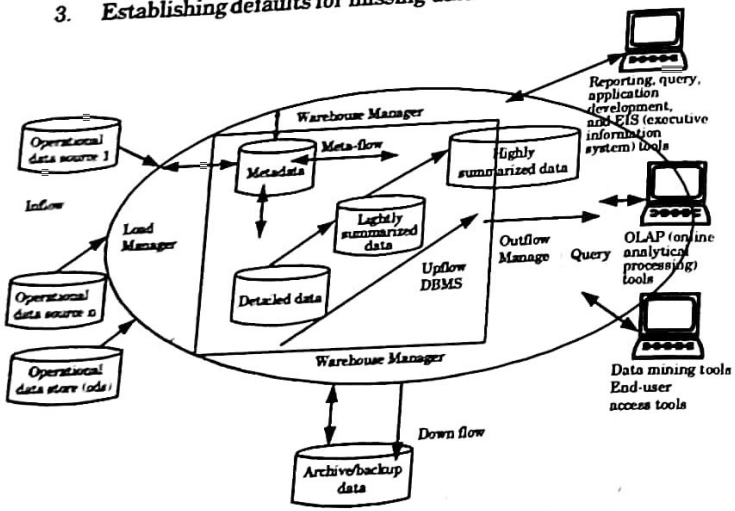


Fig. 1.3.1. Information flow of a data warehouse.

c. **Metadata :** Metadata is data about data that describes the data warehouse. It is used for building, maintaining, managing and using the data warehouse.

d. **Access tools :** Access tools are divided into four main categories :

1. Query and reporting tools
2. Application development tools
3. Online analytical processing tools
4. Data mining tools

e. **Data warehouse bus architecture :** Data warehouse bus determines the flow of data in our warehouse. The data flow in a data warehouse can be categorized as Inflow, Upflow, Downflow, Outflow and Meta flow.

PART-3

Building a Data Warehouse, Warehouse Database.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.3. **Explain the concept of building a data warehouse.**

Answer

Following steps should be adopted to build a successful data warehouse :

1. **Business considerations :**
   a. **Approach :** For data warehouse development, one of the two approaches is used :
      i. **Top-down approach :** In the top-down approach, data warehouse is built first. The data marts are then created from the data warehouse.
      ii. **Bottom-up approach :** In the bottom-up approach, data marts are created first and then data warehouse is built.
   b. **Organizational issues :** Most IS organizations have expertise in developing operational systems.
2. **Design considerations :** There are several points related to data warehouse design :
   a. **Data content :** The data warehouse system should not contain as much detail-level data as the operational system used to source this data in.

b. **Metadata :** Metadata is data about data. It means it is a description and context of the data. It helps to organize, find and understand data.

c. **Data distribution :** It become necessary to know how the data should be divided across multiple servers and which users should get access to which type of data.

d. **Tools :** The tools provide the facilities for defining the transformation and cleanup rules, data movement, user query, reporting and data analysis.

e. **Performance considerations :** An ideal data warehouse system should support interactive query processing.

3. **Technical considerations :** A number of technical issues are to be considered when implementing and building a data warehouse system.
   a. **Hardware platform :** The data warehouse server has to be able to support large data volumes and complex queries.
   b. The database management system that supports the warehouse database.
   c. **Communication infrastructure :** A data warehouse user requires a large bandwidth to interact with the data warehouse and retrieve a large amount of data for analysis.
   d. The hardware platform and software to support the metadata repository.
   e. The systems management framework that enables centralized management and administration of the entire environment.

4. **Implementation consideration :** The implementation of data warehouse requires the integration of many products.
   a. **Access tools :** Ranking, statistical analysis, time series analysis, artificial intelligence, information mapping are some of the examples of access tools types.
   b. Data extraction, cleanup and transformation and migration.
   c. **Data placement strategies :** As a data warehouse grows, there should be a way to store the data in a storage media and distribute the data in the data warehouse across multiple servers.
   d. **Metadata :** Metadata is data about data. It means it is a description and context of the data. It helps to organize, find and understand data.
   e. **User sophisticated levels :** A certain degree of sophistication is required to effectively use the warehouse.

PART-4

Mapping the Data Warehouse to a Multiprocessor Architecture.

CONCEPT OUTLINE

- Mapping the relational database to the multiprocessing hardware architectures allows successful implementation of data warehouse.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.4. **Enumerate the steps involved in mapping the data warehouse to a multiprocessor architecture.**

AKTU 2016-17, Marks 10

OR

What is the architecture of data warehouse operations ?

AKTU 2013-14, Marks 05

Answer

Steps involved in mapping the data warehouse to a multiprocessor architecture are :

1. **Relational database technology for data warehouse :**
   a. **Linear speed up :** The ability to increase the number of processor to reduce response time.
   b. **Linear scale up :** The ability to provide same performance on the same requests as the database size increases.

Types of parallelism :

i. **Horizontal parallelism :** In this, different server threads or processes handle multiple requests at the same time.

ii. **Vertical parallelism :** This form of parallelism decomposes the serial SQL query into lower level operations such as scan, join, sort etc.

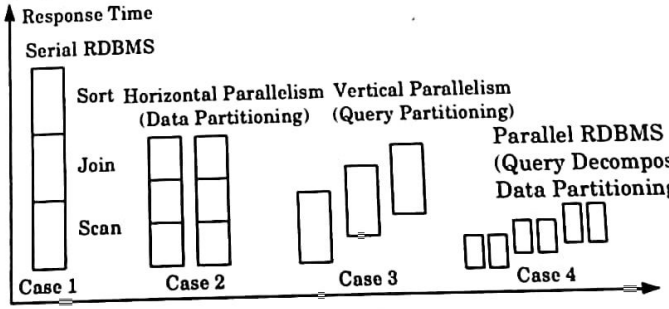


Fig. 1.4.1.

iii. **Data partitioning :** Data partitioning is the key component for effective parallel execution of database operations. Partition can be done randomly or intelligently :

a. **Random partitioning :** Includes random data striping across multiple disks on a single server.

b. **Intelligent partitioning :** Assumes that DBMS knows where a specific record is located and does not waste time searching for it across all disks.

c. **Hash partitioning :** A hash algorithm is used to calculate the partition number based on the value of the partitioning key for each row.

d. **Key range partitioning :** Rows are placed and located in the partitions according to the value of the partitioning key.

e. **Schema partitioning :** An entire table is placed on one disk; another table is placed on different disk etc. This is useful for small reference tables.

f. **User defined partitioning :** It allows a table to be partitioned on the basis of a user defined expression.

2. **Database architectures of parallel processing :** There are three DBMS software architecture styles for parallel processing :

a. **Shared memory or shared-everything architecture :** It has the following characteristics :

i. Multiple Processing Units (PU) share memory.

ii. It is simple to implement and provide a single system image, implementing an RDBMS on SMP (Symmetric Multiprocessor).

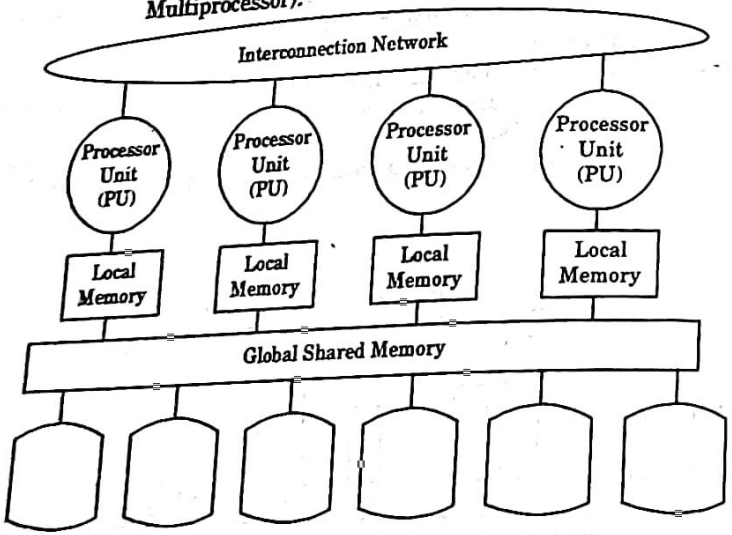


Fig. 1.4.3. Shared memory architecture.

b. **Shared disk architecture :** Shared disk architecture implements a concept of shared ownership of the entire database between RDBMS servers, each of which is running on a node of a distributed memory system.

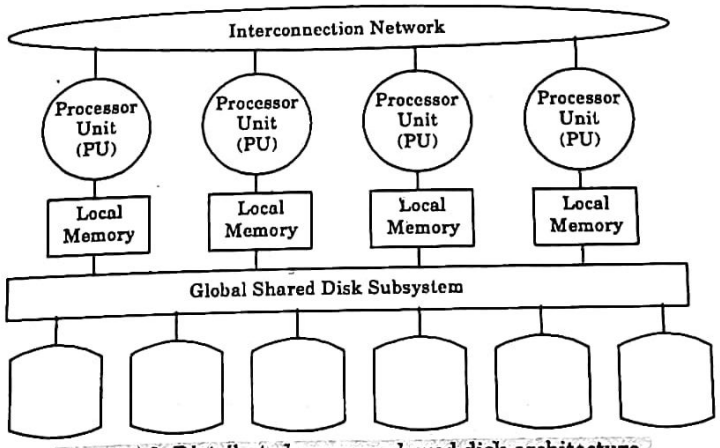


Fig. 1.4.2. Distributed memory shared disk architecture.

c. **Shared nothing architecture :** In shared architecture systems, only one CPU is connected to a given disk. If a table or database is located on that disk shared nothing systems are concerned with access to disks, not with access to memory.

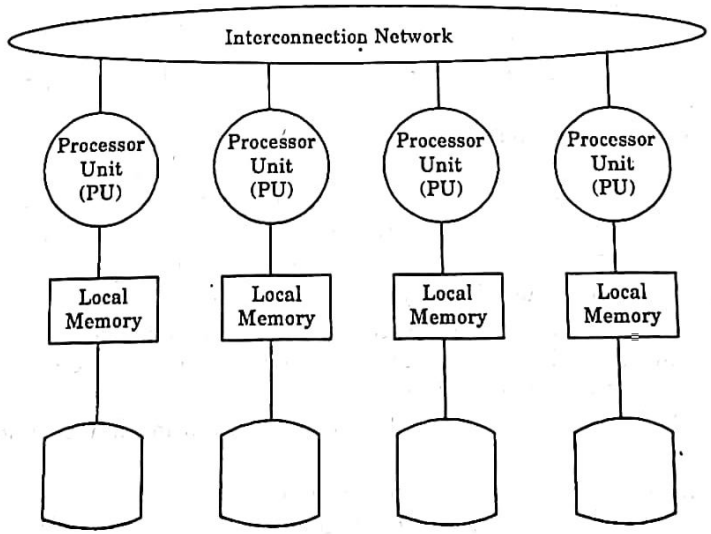


Fig. 1.4.4. Shared nothing architecture.

3. **Parallel DBMS features :**
   a. Parallel environment
   b. DBMS management tools
   c. Price / Performance
   d. Scope and techniques of parallel DBMS operations
   e. Optimized implementations
   f. Application transparency
4. **Alternative technologies :** For improving performance in data warehouse environment include following :
   a. Advanced database indexing products
   b. Multidimensional databases
   c. Specialized RDBMS
5. **Parallel DBMS Vendors :**
   a. **Oracle :** Support parallel database processing.
   b. **Informix :** It supports full parallelism.
   c. **IBM :** It is a parallel client/server database product-DB2-E (parallel edition).
   d. **SYSBASE :** It implemented its parallel DBMS functionality in a product called SYSBASE MPP (SYSBASE+NCR).

Que 1.5. **Define data warehouse. What strategies should be taken care while designing a warehouse ?** **AKTU 2017-18, Marks 10**

Answer

Data warehouse : Refer Q 1.1, Page 1–2D, Unit-1.

The strategies that should be taken care while designing a warehouse are :

1. ***Educate yourself :*** We must understand what users want because the purpose of a data warehouse system is to provide decision-makers the accurate, timely information they need to make the right choices.
2. **Determine business requirements :** To determine business requirements are should understand the following :
   a. Why the requestor needs a data warehouse.
   b. What are they trying to accomplish – saving time in collecting data, higher quality of data, supporting certain applications etc., we need to tie these business objectives to data sources.
   c. What business rules to follow and what users and/or applications to support.
3. ***Make a timeline :*** Break up business objectives mentioned above into two to three month incremental deliverables.
4. *Choosing architecture, methodology and technology and building a team.*

PART-5

Difference between Database System and Data Warehouse, Multidimensional Data Model.

CONCEPT OUTLINE

- A database system describes processing at operational sites whereas a data warehouse describes processing at warehouse.
- A multidimensional data model is used for the design of corporate data warehouses.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.6. **What is data warehouse ? How does it differ from a database ?** **AKTU 2013-14, Marks 05**

Answer

Data warehouse : Refer Q. 1.1, Page 1–2D, Unit-1.

Difference :

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.

Que 1.7. **Differentiate between OLTP and OLAP with example.** **AKTU 2013-14, Marks 10**

OR

Differentiate between OLAP and OLTP. **AKTU 2017-18, Marks 10**

Answer

S. No.	Basis	OLAP	OLTP
1.	Abbreviation	It stands for 'Online Analytical Processing'.	It stands for 'Online Transaction Processing'.
2.	Use	It is used for Query Processing.	It is used for transaction Processing
3.	Data	It holds historical data. It stores only relevant data.	It holds current data. It stores all data.
4.	Type	It is analysis driven.	It is application driven.
5.	Source	The data comes from various OLTP sources	It is the original source of data.
6.	Purpose	To help with planning problem solving and decision support.	To control and run fundamental business tasks.
7.	Business	It reveals the multidimensional view of all types of business activities.	It reveals the ongoing business process.
8.	Speed	It is slow depending on the data.	It is very fast.
9.	Market	It is customer oriented.	It is market oriented.
10.	Database design	It is de-normalized with fewer tables and makes use of star or snow flake schemas.	It is highly normalized with many tables.
11.	View	It represents managerial view.	It represents clerical or operator view.
12.	Users	It has few concurrent users.	It has many concurrent users.

Que 1.8. **What is metadata and why is it important ? Discuss the multidimensional data.** **AKTU 2013-14, Marks 10**

Answer

Metadata :

1. Metadata is data about data. It means it is a description and context of the data. It helps to organize, find and understand data.
2. In data warehouse, metadata are the data that defines warehouse objects.
3. Metadata can be classified into two types : Technical metadata and Business metadata.

Importance of metadata :

1. Metadata drives data warehouse processes.
2. Metadata gives user the meaning of each data element.
3. Metadata establishes the context for data elements.

Multidimensional data :

1. Multidimensional data model stores data in the form of data cube. A data cube allows data to be viewed in multiple dimensions.
2. Data warehouses and Online Analytical Processing (OLAP) tools are based on a multidimensional data model.
3. Multidimensional data model provide both a mechanism to store data and a way for business analysis.
4. The two primary component of multidimensional data model are dimensions and facts. Dimensions are the entities with respect to which an organization wants to keep records and facts are the numerical measures.
5. There are three types of multidimensional data model :
   a. Star schema model
   b. Snow flake schema model
   c. Fact constellations

PART-6

Data Cubes.

CONCEPT OUTLINE

- A data cube is a multidimensional array of values which is used to view the data.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.9. **Describe data cubes with suitable example.**

Answer

a. A data cube is a method of storing data in a multidimensional form. It is a structure that enables OLAP to achieve multidimensional functionality.

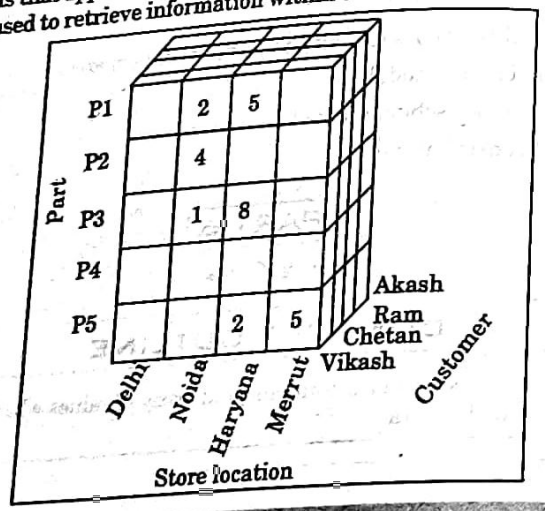
b. Data cubes are mainly categorized into two categories :
   i. Multidimensional data cube
   ii. Relational data cube

c. The data cube is used to represent data along some measure of interest.

d. Three important concepts associated with data cubes are :
   i. Slicing
   ii. Rotating
   iii. Dicing

Example :

1. We have a database that contains transaction information relating company sales of a part to a customer at a store location.
2. The data cube formed from this database is a three dimensional representation, with each cell (p, c, s) of the cube representing a combination of values from part, customer and store-location.
3. The content of each cell is the count of the number of times that specific combination of values occurs together in the database.
4. Cells that appear blank in fact have a value of zero. The cube can then be used to retrieve information within the database.



PART-7

Stars, Snow Flakes, Fact Constellations.

CONCEPT OUTLINE

- A multidimensional model can exists in the following three forms :
  i. Star schema
  ii. Snow flake schema
  iii. Fact constellations

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.10. **Explain with diagram the star, snow flake and fact constellation schemas for multidimensional databases.**

AKTU 2014-15, Marks 10

OR

Give the difference between the star and fact constellation multidimensional data model. **AKTU 2015-16, Marks 10**

OR

"A data warehouse can be modeled by either a star schema or a snow flake schema". With relevant examples discuss the two types of schema. **AKTU 2016-17, Marks 10**

OR

What are different database schemas shows with an example ?

AKTU 2017-18, Marks 05

Answer

Star schema :

1. The simplest data warehouse schema is star schema because its structure resembles a star.
2. Star schema consists of data in the form of facts and dimensions. The fact table present in the center of star and points of the star are the dimension tables.
3. In star schema fact, table contains a large amount of data, with no redundancy. Each dimension table is joined with the fact table using a primary or foreign key.
4. The main characteristics of star schema are that it is easy to understand and small number of tables can join.
5. The advantage of star schema is that it provides highly optimized performance for typical star users.

Example : Let us consider the "Employment" data warehouse. We have three dimension tables and one fact table. The star schema is shown in Fig. 1.10.1.

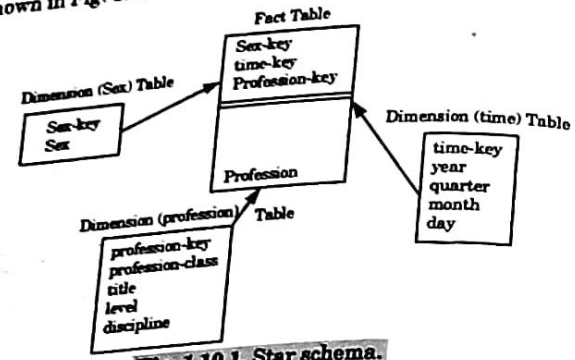


Fig. 1.10.1. Star schema.

Snow flake schema :

1. The snow flake schema is different from star schema because dimension tables of the snow flake are normalized.
2. The snow flake schema is represented by centralized fact table which is connected to multiple dimension table.
3. The major difference between the snow flake and star schema models is that the dimension tables of the snow flake model are normalized to reduce redundancies.

Example : Snow flake schema for a company XYZ electronics. Dimension table is normalized resulting in two tables.

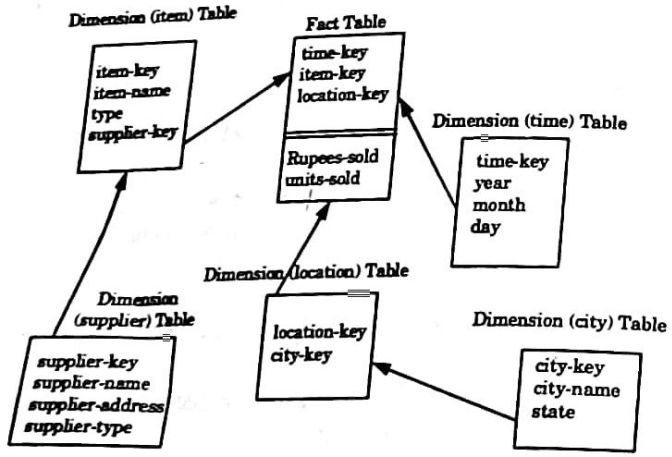


Fig. 1.10.2. Snow flake schema.

Fact constellations :

1. A fact constellation can have multiple fact tables that share many dimension tables.
2. This type of schema can be viewed as a collection of stars, snow flake and hence is called a galaxy schema or a fact constellation.
3. The main disadvantage of fact constellation schemas is its more complicated design.

Example : Let us assume that Deccan Electronics would like to have another fact table for supply and delivery. It may contain five dimensions, or keys : time, item, delivery-agent, origin, destination along with the numeric measure as the number of units supplied and the cost of delivery. It can be seen that both fact tables can share the same item-dimension table as well as time-dimension table. A fact constellation schema is shown in Fig. 1.10.3.

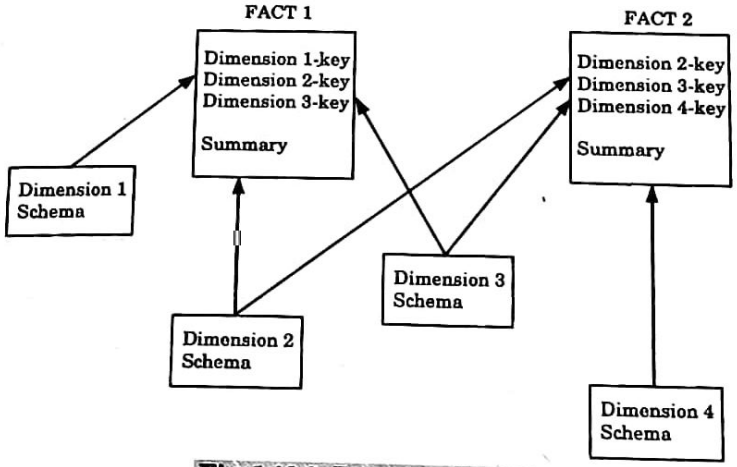


Fig. 1.10.3. Fact constellation.

Difference :

S. No.	Star schema	Fact constellation
1	In star schema, each dimension is represented by only one table.	In fact constellation, each dimension is represented by multiple fact tables.
2.	It is simple to understand and easily designed.	It is more complex and hard to design.
3.	It does not use normalization.	It uses normalization.
4.	It saves the space due to single fact table.	It does not save space due to multiple fact table.

Que 1.11. Suppose that a data warehouse for a University consists of the following four dimensions : student, course, semester and instructor, and two measures such as count and avg_grade. When at the lowest conceptual level (for example, for a given student, course, semester and instructor combination) the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination.

i. Draw a snow flake schema diagram for the data warehouse.

ii. Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (for example, roll-up from semester to year) should one perform in order to list the average grade of CS courses for each student of the University.

AKTU 2016-17, Marks 15

Answer

i.

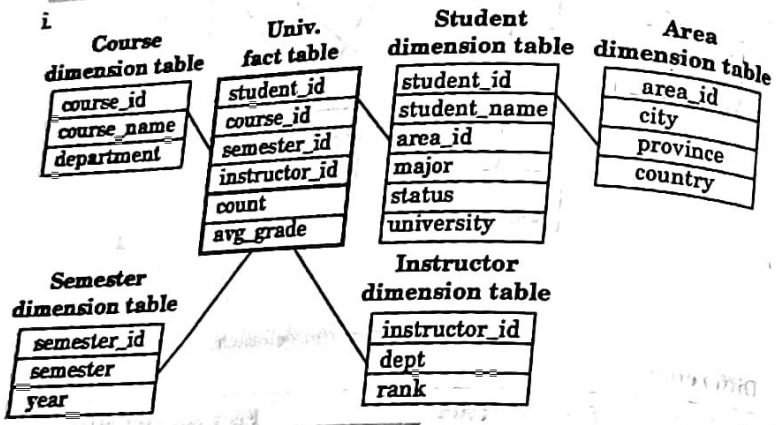


Fig. 1.11.1.

ii. Starting with the base cuboid [student, course, semester, instructor] :

1. Roll-up on course from (course_key) to major.
2. Roll-up on student from (student_key) to University.
3. Dice on course, student with department = "CS" and University = "Big University".
4. Drill-down on student from University to student name.

PART-8

Concept Hierarchy.

CONCEPT OUTLINE

- Concept hierarchy is a directed acyclic graph of concepts, where each of the concepts is identified by a unique name.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

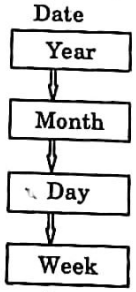
Que 1.12. Describe concept hierarchy with example.

AKTU 2013-14, Marks 05

AKTU 2017-18, Marks 2.5

Answer

1. Concept hierarchy represents the relationship between data elements in such a way that they can relate to each other as one above another, one below another.
2. Example of concept hierarchy is Date hierarchy which forms a relationship as Year → Month → Day → Week etc.



3. There are three main types of hierarchies in data warehouse design :
   a. Balanced hierarchy
   b. Unbalanced hierarchy
   c. Ragged hierarchy
4. Concept hierarchy reduces the data by collecting and replacing low level concepts by higher level concepts.

Que 1.13. What do you mean by granularity ? What is partitioning ?

AKTU 2013-14, Marks 05

Answer

Granularity :

1. Granularity means the level of the data stored in any table in the data warehouse.
2. High granularity refers to data that is at or near the transaction level.
3. Low granularity refers to data that is summarized or aggregated, usually from the atomic level data.
4. A fact table is usually designed at a low level of granularity.
5. The process consists of the following two steps :
   a. Determining the dimensions that are to be included.
   b. Determining the location to place the hierarchy of each dimension of information.

Partitioning : Refer Q 1.4, Page 1–6D, Unit-1.

2 UNIT

Data Warehouse and Process Technology

CONTENTS

PART-1

Warehousing Strategy, Warehouse Management and Support Processes.

CONCEPT OUTLINE

- A warehouse strategy involves many important decisions such as the investment and operation costs that make up the logistics overhead.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.1. **List the elements of warehouse strategy.**

Answer

The data warehouse strategy should include the following elements :

1. **Preliminary data warehouse rollout plan :** All the user requirements cannot be met in one data warehouse project as the project would necessarily be large.
2. **Preliminary data warehouse architecture :** It defines the overall and initial architecture of each rollout.
3. **Shortlisted data warehouse environment and tools :** Create a shortlist for the tools and environment that appear to meet warehousing needs.

Que 2.2. **Explain warehouse management and support processes.**

Answer

Warehouse management and support processes are designed to address the aspects of the planning and managing DW project, subject to successful implementation and extension of software.

Steps in warehouse management and support processes :

1. **Define issue tracking and resolution process :** It includes following guidelines : Issue description, urgency, raised by, assigned to, date opened, date closed, resolved by and resolution description.
2. **Perform capacity planning :** It can be done is following forms :

   a. **Space required :** Space requirements are determined by schema design, backup and recovery strategy, indexing strategy, aggregation, metadata etc.
   b. **Machine processing power :** It chooses a configuration that is scalable and meets the processing requirements.
   c. **Network bandwidth :** It verify all assumptions about the network bandwidth before proceeding with each rollout.
3. **Define warehouse purging rules :** It define the mechanism for archiving or removing older data from the data warehouse and check for any legal, regulatory or auditing requirements.
4. **Define security management :** It keeps the data warehouse secure to prevent the loss of information either due to disaster or due to an unauthorized user. Various steps involved in security management are :
   i. Determine and evaluate IT assets
   ii. Analyze risk
   iii. Define security practices
   iv. Implement practices
   v. Monitor violations and take corresponding actions
   vi. Re-evaluate IT assets and risk
5. **Define backup and recovery strategy :**
   i. **Data to be backed up :** Identify the data that must be backed up on a regular basis.This gives us an indication of the regular backup size.
   ii. **Batch window of the warehouse :** It determines the maximum allowable down time for the warehouse.
   iii. **Maximum acceptable time for recovery :** It determines the maximum acceptable time for the warehouse data and metadata to be restored.
   iv. **Acceptable costs for backup and recovery :** Different backup mechanisms imply different backup costs.
6. **Set up collection of warehouse usage statistics :** Warehouse usage statistics are collected to provide the data warehouse designer with inputs for further refining the data warehouse design and to track the general usage and acceptance of warehouse.

PART-2

Warehouse Planning and Implementation.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.3. **Write a short note on data warehouse planning.**

Answer

The data warehouse planning describes the activities related to planning on rollout of the data warehouse. Different approaches for data warehouse planning are :

1. **Assemble and orient the team :**
   i. Identify the employees and brief them about the project.
   ii. Distribute copies of DW strategy.
   iii. Set up teams and specify roles.
   iv. Give training if required.
   v. Set up milestones and check points.
2. **Conduct decisional requirements analysis :** It means gain a thorough understanding of the information needs of decision-makers.
3. **Conduct decisional source system audit :** Survey current source of data for data warehouse.
4. **Design logical and physical warehouse schema :** It includes two schema design techniques :
   i. **Normalization :** Normalize scaled attribute data so as to fall within a small specified range, such as 0.0 to 1.0.
   ii. **Dimensional modeling :** This technique produces denormalized, star schema designs consisting of fact and dimension tables. A variation of the dimensional star schema also exists (*i.e.*, snowflake schema).
5. **Produce source-to-target field mapping :** The source-to-target field mapping documents how fields in the operational systems are transformed into the data warehouse fields.
6. **Select development and production environment and tools :** It finalizes the computing environment and tool set for rollout based on the results of development and production environment.
7. **Create prototype for this rollout :** It creates a prototype of the data warehouse using the final tools and production environment.
8. **Create implementation plan for this rollout :** It drafts an implementation plan for the rollout.

Que 2.4. **Explain all steps and guidelines for data warehouse implementation.** **AKTU 2014-15, Marks 10**

Answer

Steps for data warehouse implementation :

1. ***Requirements analysis and capacity planning :*** The first step in data warehousing involves defining enterprise needs, defining architecture, carrying out capacity planning and selecting the hardware and software tools.

2. **Hardware integration :** Once the hardware and software have been selected, they need to be put together by integrating the servers, the storage devices and the client software tools.
3. **Modeling :** Modeling is a major step that involves designing the warehouse schema and views. This may involve using the modeling tool if the data warehouse is complex.
4. **Physical modeling :** This involves designing the physical data warehouse organization, data placement, data partitioning, deciding on access methods and indexing.
5. **Sources :** The data for the data warehouse is likely to come from a number of data sources. This step involves identifying and connecting the sources using gateways, ODBC drives or other wrappers.
6. **ETL :** The data from the source systems will need to go through an ETL process. The step of designing and implementing the ETL process may involve identifying a suitable ETL tool vendor and purchasing and implementing the tool.
7. **Populate the data warehouse :** Once the ETL tools have been agreed upon, testing the tools will be required, perhaps using a staging area.
8. **User applications :** For the data warehouse to be useful there must be end-user applications. This step involves designing and implementing applications required by the end users.
9. **Roll-out the warehouse and applications :** Once the data warehouse has been populated and the end-user applications are tested, the warehouse system and the applications may be rolled out for the user community to use.

Guidelines for data warehouse implementation :

1. **Build incrementally :** Data warehouses must be built incrementally. It is recommended that a data part may first be built with one particular project in mind and then data warehouse can be implemented in an iterative manner allowing all data parts to extract information from the data warehouse.
2. **Need a champion :** A data warehouse project must have a champion who is willing to carry out considerable research into expected costs and benefits of the project.
3. **Senior management support :** A data warehouse project must be fully supported by the senior management. Give the resource intensive nature of such projects and the time they take to implement, a warehouse project calls for a sustained commitment from senior management.
4. **Ensure quality :** Only data that has been cleaned should be loaded in the data warehouse.

5. **Corporate strategy :** A data warehouse project must fit with corporate strategy and business objectives.
6. **Business plan :** The financial costs (hardware, software, and peopleware), expected benefits and a project plan (including an ETL plan) for a data warehouse project must be clearly outlined and understood by all stakeholders.
7. **Training :** A data warehouse project must not overlook data warehouse training requirements.
8. **Adaptability :** The project should build in adaptability so that changes can be made to the data warehouse when required. Like any system, a data warehouse will need to change, as needs of an enterprise change.
9. The project must be managed by both IT and business professionals in the enterprise.

PART-3

Hardware and Operating Systems for Data Warehousing.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.5. **Explain hardware and operating systems used in data warehouse.**

Answer

Hardware and operating system refers to the server platforms and operating system that serve as the computing environment of the data warehouse. Hardware and operating system used in data warehouse are :

1. ***Parallel hardware technology :***
   a. **Symmetric multiprocessing :** These system consists of pair of about 64 processor that share a common memory and operating system. Since, system resources are shared hence, they can be easily managed and make use of high speed interconnections, high bandwidth and low latency is also achieved.

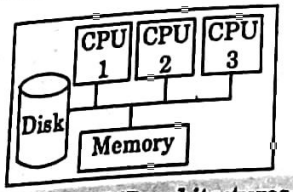


Fig. 2.5.1. SMP architectures.

   b. **Massively parallel processor systems :** It uses a large number of processors which communicate using some message interface. Each processor has its own CPU, memory and disk subsystem.

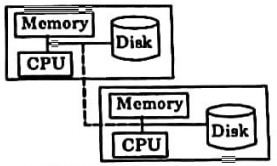


Fig. 2.5.2. MPP architecture.

2. **Clustered system :** These systems are configured with multiported array so that nodes which have direct disk access enjoy same disk I/O rates as standalone SMP systems. Nodes which not have direct disk access must use the high-speed cluster interconnect mechanism.

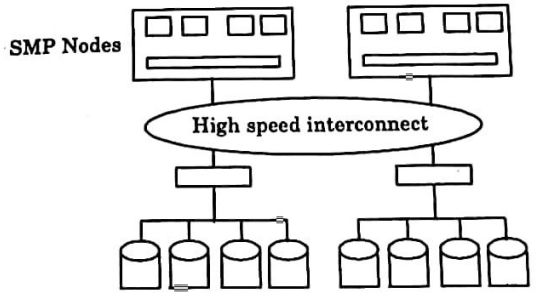


Fig. 2.5.3. Cluster of four SMP systems.

Que 2.6. **Mention criteria for hardware selection.**

Answer

The following selections are recommended for hardware selection :

1. Scalability
2. Financial stability
3. Price/Performance
4. Delivery lead time
5. Reference sites
6. Availability of support

PART-4

Client/Server Computing Model and Data Warehousing, Parallel Processors and Cluster Systems.

CONCEPT OUTLINE

- Client/server is a program client which requests a service or resource from the server.
- Parallel processing is a method of dividing the program into multiple fragments to speed up the execution of programs.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

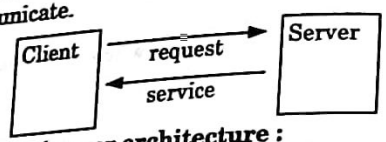
Que 2.7. Explain client/server architecture.

Answer

1. Client/server architecture is a network architecture in which each computer on the network is either a client or a server.
2. Client/server architecture works when the client sends a request to the server over the network connection, which is then processed and delivered to the client.

Components of client/server architecture :

1. **Client :** It is a computer which processes the request service from the server.
2. **Server :** Any computer can provide services to the client.
3. **Communication middleware :** A computer through which client and server communicate.



Advantages of client/server architecture :

1. Scalability
2. Resource sharing
3. Data integrity
4. Communication cost is reduced.
5. Ease of effort and maintenance.

Disadvantages of client/server architecture :

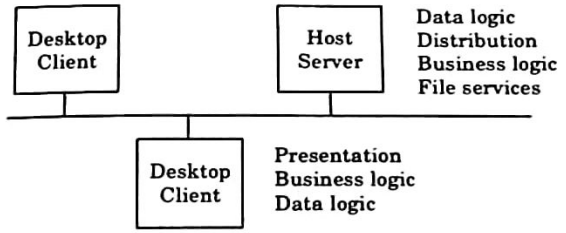
1. Single point of failure
2. Costly to maintain DB server

Que 2.8. What are the types of client/server architecture ?

Answer

Types of client/server architecture are :

1. **Two-tier architecture :** A two-tier architecture is where a client talks directly to a server. It is typically used in small environments.



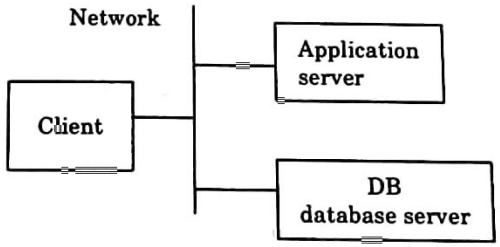
Advantages of two-tier architecture :

1. Interoperability
2. Portability
3. Integration
4. Transparency
5. Security

Disadvantages of two-tier architecture :

1. Network traffic is handled less efficiently.
2. The client and server are tightly coupled.

2. **Three-tier architecture :** In the three-tier architecture, a middleware is used between the client environment and the database management server environment. It is used in large environment.



Advantages of three-tier architecture :

1. Improve performance
2. Improve flexibility

Disadvantages of three-tier architecture :

1. The development environment is more difficult to use.

Que 2.9. Describe distributed memory architecture. What are its types ?

Answer

In distributed memory architecture all processors in the system are directly connected to own memory and caches. Any processor cannot directly access another processor's memory.

Two types of distributed memory architecture are :

1. **Shared nothing architecture :**
   a. Shared nothing architecture is used in distributing computing in which each node have their own memory, storage and independent input/output interfaces.
   b. Each node do not shares any resources with other nodes and communicate with each other by passing messages.

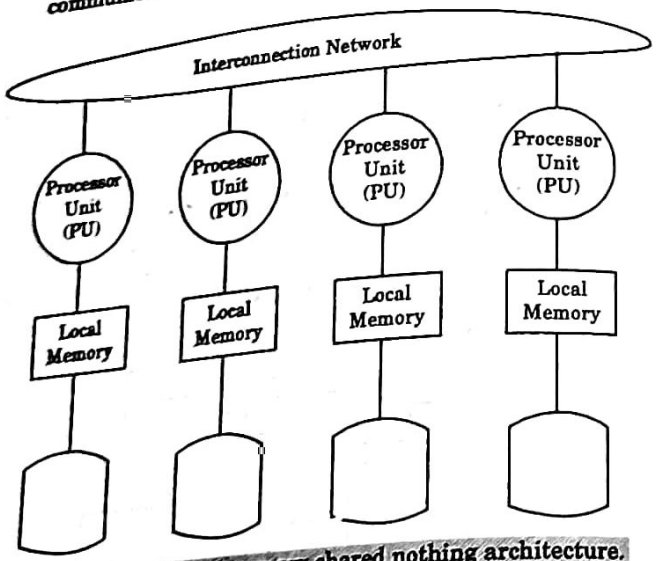


Fig. 2.9.1. Distributed memory shared nothing architecture.

2. **Shared disk architecture :**
   a. A shared disk architecture is a distributed computing architecture in which all disks are accessible from all cluster nodes.

   b. Multiple processors can access all disks directly via intercommunication network and every processor has local memory.

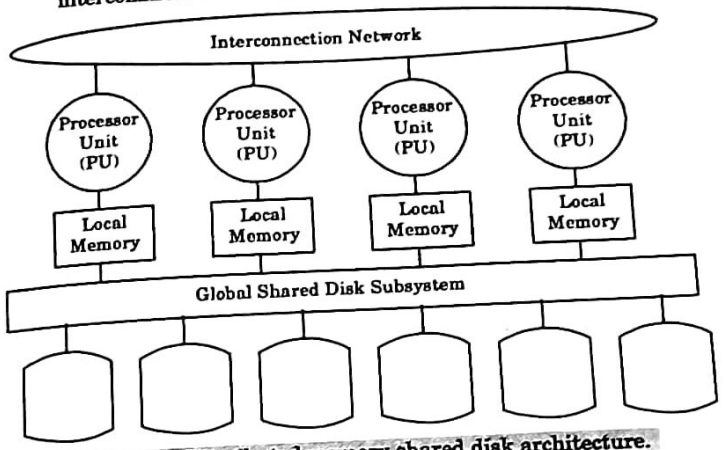


Fig. 2.9.2. Distributed memory shared disk architecture.

Que 2.10. Write a short note on cluster system.

Answer

1. In a cluster system, every processor unit (PU) executes a copy of operating system, and the inter-PU communications are performed over an open-systems-based interconnection.
2. Cluster system is designed for high availability by providing shared access to disks.
3. Cluster system describes many characteristics of MPP system, including a very high-speed scalable interconnection mechanism and support for hundreds of PUs.

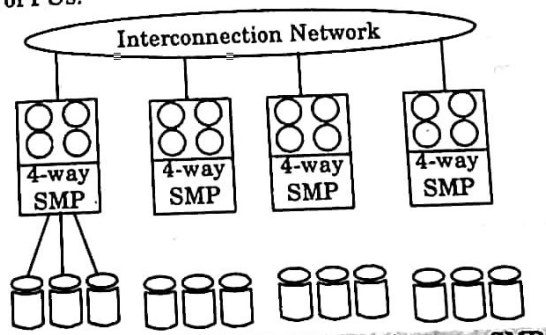


Fig. 2.10.1. Distributed memory cluster of four-way SMP nodes.

PART-5

Distributed DBMS Implementations, Warehousing Software and Warehouse Schema Design.

CONCEPT OUTLINE

- Schema is a logical description of the entire database.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.11. **Explain warehousing software tools in detail.**

Answer

Warehousing software tools are :

1. **Connectivity tools :** These tools provide transparent access to source system in heterogeneous environment.

   For example :

   i. **IBM :** Data joiner

   ii. **Oracle :** Transparent gateway

   iii. **SAS :** SAS/connect

   iv. **Sybase :** Enterprise connect

2. **Extraction tools :** There are two primary methods to use extraction tools *i.e.*, bulk extraction and change-based replication.

   For example :

   i. **Apertus carleton :** Passport

   ii. **Platinum :** InfoPump

3. **Transformation tools :** These tools has following features :

   i. Field splitting and consolidation

   ii. Standardization

   iii. De-duplication

   For example :

   i. **Apertus carleton :** Enterprise/Integrator

   ii. **Data mirror :** Transformation server

4. **Data quality tools :** It helps to identify and correct data error at source systems.

   For example :

   i. **Data flux :** Data quality workbench

   ii. **Prism :** Quality manager

   iii. **Pine cone systems :** Content tracker

5. **Data loaders :** It transforms data into data warehouse.

6. **Data access and retrieval tools :** These tools are classified into two categories :

   i. **OLAP tools :** These allow users to make ad hoc queries or generate queries against warehouse database.

   ii. **Reporting tools :** These allow users to produce scanned and sophisticated reports based on warehouse data.

7. **Data modeling tools :** These tools allow users to prepare and maintain an information model of both source and target database.

   For example :

   i. Cayenne software, Terrain

   ii. Relational matters, Syntagma designer

   iii. Sybase, PowerDesigner WarehouseArchitect

8. **Warehouse management tools :** These tools assist warehouse admin in the day-to-day management and administration of the warehouse.

   For example :

   i. Pine cone systems, usage tracker, refreshment tracker.

   ii. Red brick systems, enterprise control and coordination.

Que 2.12. **Discuss various warehouse schema design techniques.**

Answer

Various warehouse schema design techniques are :

1. OLTP systems use normalized data structures.
2. **Dimensional modeling for decisional systems :** It provides a number of techniques for denormalizing database to create schema.
3. **Star schema :** Refer Q.1.10, Page 1–14D, Unit-1.
4. **Dimensional hierarchies :** Each dimension will have hierarchies that imply grouping and structure.
5. **Granularity of the fact table :** The first step in designing a fact table is to determine the granularity of the fact table. By granularity, we mean the lowest level of information that will be stored in the fact table. This constitutes two steps :

   a. Determine which dimensions will be included.

   b. Determine where along the hierarchy of each dimension the information will be kept.

6. ***Aggregates or summaries :*** Aggregates are the summarization of fact related data for the purpose of improved performance. Aggregates are to be considered for use when the number of detailed records to be processed is large and/or the processing of the customer queries begins to impact the performance.
7. ***Dimensional attributes :*** The attribute values are used to establish the context of the facts.
8. ***Multiple star schemas :*** A data warehouse will have multiple star schemas, *i.e.*, many fact tables.

Data Mining

CONTENTS