**ANUBHAV BEJ**

**(1MS19MCA09)**

1. **Describe the different problems while using the traditional file management system.**
2. **Data Redundancy:** Since each application has its own data file, the same data may have to berecorded and stored in many files.

Example:  personal file and payroll file, both contain dataon employee name, designation etc. The result is unnecessary duplicate or redundant data items.This redundancy requires additional or higher storage space, costs extra time and money, andrequires additional efforts to keep all files upto-date.

1. **Data Inconsistency:** Data redundancy leads to data inconsistency especially when data is to beupdated. Data inconsistency occurs due to the same data items that appear in more than one filedo not get updated simultaneously in each and every file. For example, an employee is promoted from Clerk to Superintendent and the same is immediately updated in the payroll file may notnecessarily be updated in provident fund file. This results in two different designations of anemployee at the same time. Over the period of time, such discrepancies degrade the quality ofinformation contain in the data file that affects the accuracy of reports.
2. **Data Dependence:** The Applications/programs in file processing system are data dependenti.e., the file organization, its physical location and retrieval from the storage media are dictated by the requirements of the particular application. For example, in payroll application, the filemay be organized on employee records sorted on their last name, which implies that accessing ofany employee's record has to be through the last name only.
3. **Lack of Data Integration:** Since independent data file exists, users face difficulty in getting information on any ad hoc query that requires accessing the data stored in many files. In such cases complicated programs have to be developed to retrieve data from every file or the users have to manually collect the required information.
4. **Lack of flexibility:**A traditional file system can deliver routine scheduled reports after extensiveprogramming efforts, but it cannot deliver ad hoc reports or respond tounanticipated information requirements in a timely fashion. together the required data items in a new file.
5. **Limited Data Sharing:** There islimited data sharing possibilities with the traditional filesystem. Each application has its own private files and users have little choice to share the dataoutside their own applications. Complex programs required to be written to obtain data fromseveral incompatible files.
6. **Poor security:**Because there is little control or management of data, access to and disseminationof information may be out of control. Management may have no wayof knowing who is accessing or even making changes to the organizations.
7. **What are the problems of managing data resources in a traditional file environment and how are they solved by a database management system?**

Traditional files environment leads to data redundancy and inconsistency.

Since each application has its own data file, the same data may have to be recorded and stored in many files.The result is unnecessary duplicate or redundant data items. This redundancy requires additional or higher storage space, costs extra time and money, and requires additional efforts to keep all files up to-date.

Data redundancy leads to data inconsistency especially when data is to be updated. Data inconsistency occurs due to the same data items that appear in more than one file do not get updated simultaneously in each and every file.

* A DBMS reduces data redundancy and inconsistency by minimizing isolatedfiles in which the same data are repeated. The DBMS may not enable theorganization to eliminate data redundancy entirely, but it can help controlredundancy. Even if the organization maintains some redundant data, usinga DBMS eliminates data inconsistency because the DBMS can help theorganization ensure that every occurrence of redundant data has the samevalues.
* DBMS uncouples programs and data, enabling data to stand on
* their own. Access and availability of information will be increased and
* program development and maintenance costs reduced because users
* and programmers can perform ad hoc queries of data in the database.
* The DBMS enables the organization to centrally manage data.

1. **CAPABILITIES OF DATABASE MANAGEMENT SYSTEMS**

A DBMS includes capabilities and tools for organizing, managing, and accessingthe data in the database. The most important are its data definition language,data dictionary, and data manipulation language.

DBMS have a **data definition** capability to specify the structure of the contentof the database. It would be used to create database tables and to define thecharacteristics of the fields in each table. This information about the databasewould be documented in a data dictionary.A **data dictionary** is an automated ormanual file that stores definitions of data elements and their characteristics.Microsoft Access has a rudimentary data dictionary capability that displaysinformation about the name, description, size, type, format, and other propertiesof each field in a table.

Querying and ReportingDBMS includes tools for accessing and manipulating information in databases.Most DBMS have a specialized language called a **data manipulation language**that is used to add, change, delete, and retrieve the data in the database.This language contains commands that permit end users and programmingspecialists to extract data from the database to satisfy information requests anddevelop applications. The most prominent data manipulation language today is**Structured Query Language**, or **SQL**.

1. **What are some important database design principles?**
2. **Usability**: Any information which we are storing in any organization should be meaningful for that organization. If we are storing those factors which are actually not fit with organization’s requirement then this is just waste of resources.Primary objective of any information system should be to meet organization requirements. Properly get details about requirements.

* See how information can be fit with requirement.
* Trace requirement matrix to capture mapping of information architecture and requirements.
* Organize it simple.

1. **Extensibility:** information design should be extensible so that it can adopt new requirements without much efforts or without major breaking changes.If your initial design is too much complex or unorganized then it may create trouble for you to adopt new things effectively.

* Normalization and correct handling of optional data.
* Generalization of entities when designing the schema are necessary.

1. **Data Integrity:** Our system should be smart enough to handle incorrect, missing data attributes and based on that it should either take corrective actions or straightaway reject the data. Incorrect data should not be present in system or at least should not exposed to individuals creating misunderstanding.

Data integrity can be of many types:

* Entity Integrity
* Domain Integrity
* Referential Integrity
* Transactional Integrity
* User defined integrity.

1. **Performance: p**erformance of the system should be up to the mark. As data in increasing day by day so at some time there will be impact on performance if database design is poor or we’ll not take any actions to improve performance.

Following could be few strategies which we can implement when there is need as data increases:

* + A well-designed schema with normalization and generalization
  + fast transactions that reduce locking and blocking
  + Partitioning, which is useful for advanced scalability.

1. **Availability:** The availability of information refers to the information’s accessibility when required regarding uptime, locations, and the availability of the data for future analysis. Disaster recovery, redundancy, archiving, and network delivery all affect availability
2. **Security:** For any organizational asset, the level of security must be secured depending on its value and sensitivity. Sometime organizations have suffered a lot because of data leaks which results in loss of faith and tends to business risk. So, security is one of the most important aspect of good database design.

We can enhance Security by the following:

* Physical security and restricted access of the data centre.
* Appropriate operating system security.
* Identifying and documenting ownership of the data.
* Granting access according to the principle of least privilege, which is the concept that users should have only the minimum access rights required to perform necessary functions within the database.

1. **Illustrate with the help of a block diagram the components of a data warehouse.**

A **data warehouse** is a database that stores current and historical data of

potential interest to decision makers throughout the company. The data

originate in many core operational transaction systems, such as systems

for sales, customer accounts, and manufacturing, and may include datafrom Web site transactions.

The components are:

The data warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key components to make the entire environment functional, manageable and accessible

There are mainly five components of Data Warehouse:

**DATA WAREHOUSE DATABASE**

The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology. Although, this kind of implementation is constrained by the fact that traditional RDBMS system is optimized for transactional database processing and not for data warehousing. For instance, ad-hoc query, multi-table joins, aggregates are resource intensive and slow down performance.Hence, alternative approaches to Database are used :

* In a Datawarehouse, relational databases are deployed in parallel to allow for scalability. Parallel relational databases also allow shared memory or shared nothing model on various multiprocessor configurations or massively parallel processors.
* Use of multidimensional database (MDDBs) to overcome any limitations which are placed because of the relational data model. Example: Essbase from Oracle.

**SOURCING,ACQUISITION,CLEAN-UPAND TRANSFORMATION TOOLS**

The data sourcing, transformation, and migration tools are used for performing all the conversions, summarizations, and all the changes needed to transform data into a unified format in the Datawarehouse. They are also called Extract, Transform and Load (ETL) Tools.These Extract, Transform, and Load tools may generate jobs, background jobs, Cobol programs, shell scripts, etc. that regularly update data in Datawarehouse. These tools are also helpful to maintain the Metadata. These ETL Tools have to deal with challenges of Database & Data heterogeneity.

**METADATA**

The name Meta Data suggests some high- level technological concept. However, it is quite simple. Metadata is data about data which defines the data warehouse. It is used for building, maintaining and managing the data warehouse.In the Data Warehouse Architecture, meta-data plays an important role as it specifies the source, usage, values, and features of data warehouse data. It also defines how data can be changed and processed. It is closely connected to the data warehouse.

**QUERY TOOLS**

One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions. Query tools allow users to interact with the data warehouse system.

These tools fall into four different categories:

1. Query and reporting tools
2. Application Development tools
3. Data mining tools
4. OLAP tools



1. **What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?**
   1. D**ata warehouse:** Itis a database that stores current and historical data ofpotential interest to decision makers throughout the company. The dataoriginate in many core operational transaction systems, such as systemsfor sales, customer accounts, and manufacturing, and may include datafrom Web site transactions. The data warehouse consolidates and standardizesinformation from different operational databases so that the information can beused across the enterprise for management analysis and decision making.
   2. **OLAP**: It supports multidimensional data analysis, enabling users toview the same data in different ways using multiple dimensions. Each aspectof information—product, pricing, cost, region, or time period—represents adifferent dimension. So, a product manager could use a multidimensional dataanalysis tool to learn how many washers were sold in the East in June, how thatcompares with the previous month and the previous June, and how it compareswith the sales forecast. OLAP enables users to obtain online answers to ad hocquestions such as these in a fairly rapid amount of time, even when the data arestored in very large databases, such as sales figures for multiple years.
   3. **Data mining:** Itis more discovery-driven. Data mining provides insights intocorporate data that cannot be obtained with OLAP by finding hidden patterns andrelationships in large databases and inferring rules from them to predict future behaviour. The patterns and rules are used to guide decision making and forecastthe effect of those decisions.

The types of information obtainable from datamining include associations, sequences, classifications, clusters, and forecasts.

* Associations are occurrences linked to a single event.
* In sequences, events are linked over time.
* Classification recognizes patterns that describe the group to which an itembelongs by examining existing items that have been classified and byinferring a set of rules.
* Clustering works in a manner similar to classification when no groups have yetbeen defined.
  1. **Text mining and web mining**: Text miningtools are now available to help businesses analyse these data. These tools areable to extract key elements from large unstructured data sets, discoverpatterns and relationships, and summarize the information. Businesses mightturn to text mining to analyse transcripts of calls to customer service centres toidentify major service and repair issues.The Web is another rich source of valuable information, some of which cannow be mined for patterns, trends, and insights into customer behaviour.The discovery and analysis of useful patterns and information from the WorldWide Web is called **Web mining**. Businesses might turn to Web mining to helpthem understand customer behaviour, evaluate the effectiveness of a particularWeb site, or quantify the success of a marketing campaign.