Q1 Define the following:
Data Mining, ETL, Transaction coordinator, Local Autonomy, Workload distribution

Q2 What are Data Mining Activities?

Q3 What are the basic ideas guide the creation of a data warehouse?

Q4 Answer by true or false

a. The file system is typically described as one file and numbers of different application programs are written to extract records from and add records to the appropriate files
b. Data Models is a collection of physical tools for describing data
c. Number of schema in DB is more than one
d. Fragment is not important for improve performance in DDBS.
e. Each transaction Coordinator is responsible for maintaining a log for recovery purposes

Q5 what is the recovery technique approaches?

Q6 A- Why database is damaged or lost?
B- What are the advantages and disadvantages of replication?

Q7 A- What is Type of DDBMS?
B- What is Disadvantages of Distribution Database?

Q8 The definition of distributed database includes four important aspects what it is?

Q9 Explain File System Disadvantage.

Q10 What are the main differences between DDB and centralized DB?

Q11 What we mean by: DDB, distributed processing?

.

The Answers:

Q1 Define the following:

Data Mining, ETL, Transaction coordinator, Local Autonomy, Workload distribution

Data Mining

There is no one single definition of data mining that would meet with universal approval, the following definitions are generally acceptable:

- Data Mining is the process of extracting previously unknown, valid and actionable information from large database and then using the information to make crucial business.
- Data Mining is the process of exploration and analysis by automatic or semi automatic means of large quantities of data in order to discover meaningful patterns and rules.
- Data Mining refers to extracting or mining knowledge from large amount of data.

Unknown, valid, actionable, business decisions lend insight into the essential nature of data mining and help to explain the fundamental differences between it and the traditional approaches to data analysis such as query and reporting and on-line analytical processing (OLAP) in essence, data mining is distinguished by the fact that it is aimed at the discovery of information, without a previously formulated hypothesis. Notes in all definitions of data mining emphasis is on large quantities of data and the patterns and rules to be found which ought to be meaningful. Clearly then data mining is more or less along of approaches to solving business problems through analysis of data to take crucial business decision. Data mining has alternative names like (knowledge discovery in database, knowledge extraction, data pattern analysis, data archeology, data dredging, information harvesting, business Intelligence, and exploratory science).

Extract Transform Load (ETL) software:

ETL software is used to move data into data warehouse tables. This process involves fetching data from source systems (extract), reorganizing it as required by the star schema design (transform), and inserting it into warehouse tables (load).

The main goal of maintaining an ETL process in an organization is to migrate and transform data from the source OLTP systems to feed a data warehouse and form data marts. ETL may be accomplished using specialized, packaged software, or by writing custom code. The ETL process may rely on a number of additional utilities and databases for staging data, cleansing it, automating the process, and so forth.

Transaction coordinator:

whose function is to coordinate the execution of the various transactions (both local and global) initiated at that site.

Each transaction manager is responsible for:

1- Maintaining a log for recovery purposes.

2- Participating in an appropriate Concurrency control scheme to coordinate the concurrency execution of the transaction execution at that site.

The transaction coordinator is not need in the centralized system since a transaction accesses data only at one single site.

Local Autonomy:

is the degree to which designer or administrator of one site may be independent of the remainder of the distributed system. we shall consider the issues of transparency and autonomy from the points of view of:

- Naming of data items.
- Replication of data items
- Fragmentation of data items
- Location of fragments and replicas.

Workload distribution :

distributing the workload over the sites is an important feature of distributed computer systems. Workload distribution is done in order to maximize the degree of parallelism of execution of applications. Storage cost and availability database distribution should reflect the cost and the availability of storage at the different sites . The cost of data storage is not relevant if compare with CPU, I /O and transmission cost of applications, but the limitation of available storage At each site must be considered. Using all the above criteria at the same time is extremely difficult, since this leads to complex optimization models. It is possible to consider some of the above features as constraints, rather than objective .

Q2 Data Mining Activities

The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis to engineering design and science exploration, now the term of DM is used for a specific set of activities, all of which involve extracting meaningful new information from the data:
- Classification, Estimation, Predication, association rules, Clustering, Description and Visualization.

The first three tasks - Classification, Estimation, and Predication – are all examples of directed data mining. In directed data mining, the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The next three tasks are examples of undirected data mining. In undirected

data mining, no variable is singled out as the target; the goal is to establish some relationship among all the variables.

1 Classification

The classification task is to build a model that can be applied to unclassified data in order to classify it as in examples of :

- Classifying credit application as low, medium, or high risk.
- Determining, which home telephone lines are used for internet access.
- Assigning customers to predefined customer segment.

In all of these examples there are a limited number of already-known classes expected to be able to assign any record into one or another of them.

In above examples one notes that the classification is used to establish a specific class for each record in database. The class must be one from a finite set of possible, predetermined class values. There exist several methods of data classification :

- Statistical Algorithms:

  Statistical analysis system such as SAS and SPSS which have been used by analysts to detect unusual patterns and explain patterns using statistical models such as linear models.

- Neural Networks:

  Artificial neural networks mimic the pattern-finding capacity of the human brain and hence some researchers have suggested applying neural network algorithms to pattern mapping. Neural networks have been applied successfully in a few applications that involve classification.

- Genetic Algorithms:

  Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- Nearest Neighbor Method:

  A technique that classifies each record in a data based on a combination of the classes of the K record (s) most similar to it in a historical data set.
  It is sometimes called the K-nearest neighbor technique.

- Rule Induction:

  The extraction of useful if-then rules from data based on statistical significance.

- Data Visualization: The visual interpretation of complex relationships in multidimensional data is necessary to provide a clear classification of data mining system such a classification may help potential users distinguish data mining system and identify those that best match their needs.

Data mining systems can be categorized according to various criteria as follows:

- classification according to the kind of data base mined

- classification according to the kind of knowledge mined
- classification according to the kind of techniques utilized
- classification according to the application adapted

In general Classification deals with discrete outcomes: yes or no.

2 Estimation

It is a process of getting some unknown continuous variable by giving some input data. Estimations deal with continuously valued outcome [3] such as income, height, or credit card balance, a family total household income, the value of a piece of real estate.

Often, classification and estimation are used together, as when data mining is used to predict who is likely to respond to accredit card balance transfer and also to estimate the size of balance to be transferred.

3 Prediction

Arguably, there should not be a separate heading for prediction. Any prediction can be thought of as classification or estimation. The difference is one of emphasis predictive modeling akin to the human learning experience, where observation is used to form a model of the essential, underlying characteristics of some phenomenon.

In data mining, a predictive model is used to analyze an existing database to determine some essential characteristics about the data, of course, the data must include complete, valid observations from which the model can learn how to make accurate predictions [40].

The model must be told the correct answer to some already solved cases before it can start to make up its own mind about new observations. When an algorithm works in this way the approach is called supervised learning.

4 Association Rules

Affinity Grouping or Association Rules are the task of affinity grouping to determine which things go together. The prototypical example is determining what things go together in a shopping cart at the supermarket.

Retail chains can use affinity grouping to plan arrangement of items on store shelves or in catalog so that items often purchased together will be seen together and in designed attractive packages or grouping of products and services.

5 Clustering

Clustering is the task of segmenting a diverse group into a number of more similar subgroups or cluster, Data clustering has been a popular technique for grouping

unstructured data sharing of common or similar features. It facilitates recognition of patterns shared by some subsets of the data and identification of significant hidden signals.

Clustering is often done as a prelude to some other form of data mining or modeling. Data clustering is often used as the first step in mining a large quantity of data. Clustering in database are the process of separating a data set into component that reflect consistent patterns which have been established, they can then be used to reconstruct data into more understandable subset and also to provide subgroups of a population for further analysis or action. This extraction of sub-groups from a population is important when dealing with large database.

6 Description and Visualization

Data in database or data warehouse can be viewed at different levels of granularity or abstraction, or as different combinations of attributes, or dimensions. Data can be presented in various visual forms, such as box plots, 3-D cubes, data distribution charts, curves, surface, link graphs, and so on. Visual display can help give users a clear impression and overview of the data characteristic in database.

Sometime the purpose of data mining is simply to describe what is going on in complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the first place. A good enough description of a behavior will often suggest an explanation for it as well. Data visualization is one powerful form of descriptive data mining. It is not always easy to come up with meaningful visualization, but the right picture really can be worth thousand association rules since human beings are extremely practiced at extracting meaning visual scenes.

Q3 What are the basic ideas guide the creation of a data warehouse:

There are two basic ideas guide the creation of a data warehouse:

- Integration of data from distributed and differently structured databases, which facilitates a global overview and comprehensive analysis in the data warehouse.
- Separation of data used in daily operations from data used in the data warehouse for purposes of reporting, decision support, analysis and controlling.

Q5 What are the Architecture components of data warehouse

Architecture components

There are four major components of the data warehouses that have already been discussed: the operational systems and the data warehouse. In addition to these databases, every data warehouse requires two additional components (software programs that move data from the operational systems to the data warehouse, and software that is used to develop queries and reports). These major components are illustrated as:

1 Operational systems:

An operational system is an application that supports the execution of a business *the major components of the data warehouse* Process, recording business activity and serving as the system of record. Operational systems may be packaged or custom-built applications. Their databases may reside on a variety of platforms, including relational database systems, mainframe based systems, or proprietary data stores.

2 Extract Transform Load (ETL) software:

ETL software is used to move data into data warehouse tables. This process involves fetching data from source systems (extract), reorganizing it as required by the star schema design (transform), and inserting it into warehouse tables (load).

The main goal of maintaining an ETL process in an organization is to migrate and transform data from the source OLTP systems to feed a data warehouse and form data marts.ETL may be accomplished using specialized, packaged software, or by writing custom code. The ETL process may rely on a number of additional utilities and databases for staging data, cleansing it, automating the process, and so forth.

3 Dimensional data warehouse:

The dimensional data warehouse is a database that supports the measurement of enterprise business processes. It stores a copy of operational data that has been organized for analytic purposes according to the principles of dimensional modeling. Information is organized around a set of conformed dimensions, supporting enterprise-wide cross-process analysis. A subject area within the data warehouse is referred to as a data mart. The dimensional data warehouse is usually implemented on a relational database management system (RDBMS).

4 Front-end software:

It is any tool that consumes information from the data warehouse, typically by issuing a SQL query to the data warehouse and presenting results in a number of

different formats. Most architectures incorporate more than one front-end product. Common front-end tools include business intelligence (BI) software, enterprise reporting software, ad hoc query tools, data mining tools, and basic SQL execution tools. The architecture of every data warehouse includes each of these fundamental components. Each component may comprise one or more products or physical servers.

Q4 Answer by true or false

f. The file system is typically described as one file and numbers of different application programs ***false***
g. Data Models is a collection of physical tools for describing data ***false***
h. Number of schema in DB is more than one ***false***
i. Fragment is not important for improve performance in DDBS. ***false***
j. Each transaction Coordinator is responsible for maintaining a log for recovery purposes ***false***

Q5 A recovery technique approaches:

1-Switch

In order to be switch to an existing copy of the database, at least two copies of database must be kept and update simultaneously. When failures occur processing is switching to the duplicate copy of the database this approach is faster recovery.

2-Restore/Rerun

This technique involves reprocessing the day's transaction (up to point failure) against the backup copy of the database. **The advantage of (restore/rerun):** The DBMS does not need to create a database change and no special restart procedures are required. **The disadvantage of (restore/rerun):**

- Several hours of reprocessing may be required.
- Processing new transaction will have to be deferred until recovery is complete

3-Transaction integrity

A database is update by processing transaction that result in changes to one or more database records if an error occurs during the processing of transaction the database may be compromised and some of database recovery is required.

4-Backup Recovery (rollback)

Used to back out undo unwanted changes to the database before images of the records that have been changed are applied to the database and the data base is returned to an earlier state used to reverse the changes made by transaction that have been a borated or terminated abnormally.

5-Forward recovery (roll forward)

Starts with an earlier copy of database after images (the image of good transaction) are applied to the database and the database is quickly moved forward to later state.

Q6

A- ***Database are damaged or lost because***:

- Some system problems that may be caused by human error.
- Hardware failure.
- Invalid or incorrect data.
- Program error.
- Computer viruses
- Natural catastrophes.

B- There are many number of a advantage and disadvantage to replication :

Advantage:

1- Availability

If one of the sites containing relation **r** fails, then the relation **r** many be found in another site .thus the system may continue in process queries involving r despite the failure of one site.

2- Increased parallelism

Several sites can process queries involving **r** in parallel.

3- Data replications minimize movement of data between sites.

4- Replication enhances the performance of read operation and increase the availability of data to read transaction.

Disadvantage :

1- Increased overhead on update since the update must be propagated to all sites.
2- Problem of controlling concurrent updates by several transactions.
3- Management of replicas of relation r by choosing one of them as the primary copy of r.

Q7 A **Type of DDBMS:**

There are two type of DDBMS

- **Homogeneous DDBMS**: it mean that the DDBMS with the same DBMS at each site even if the computers (and/or) the operating system are not the same.
- **Heterogeneous DDBMS**: it mean the DDBMS with at least two different DBMS so it cause the problem of translating between the different data models of different local DBMS to the complexity of homogenous DDBMS.

Q7 B Disadvantages of distribution Database

The primary disadvantage of distributed database systems is the added complexity among the sites. This increased complexity takes the form of:

1. software development cost It is more difficult to implement a distributed database system and, thus, more costly.
2. Greater potential for bugs
   Since the sites that comprise the distributed system operate in parallel, it is harder to ensure the correctness of algorithms. The potential exist for extremely subtle bugs .
3. Increased processing overhead.
   The exchange of messages and the additional computation required to achieve
   interstice coordination are a form of overhead that does not arise in centralized systems.

Q8 The definition of distributed database includes four important aspects what it is?

1. Distribution :

The fact that the data are not resident at the same (processor) so that we can distinguish a distributed database from a single centralized database.

2. Logical correlation :
   the fact that the data have some properties which use them together, so that we can distinguish a distributed database from a set of local
   database or files which are resident at different sites of a computer network.
   The problem with above definition is that both properties, distributions and logical correlation, are too vaguely defined to always discriminate between those cases.
3. Local application :
   The database must be local database as well as its new properties of distribution .
4. Global application :
   Any database to distributed database must have at least one global database

In order to develop a more specific definition , let us consider a few examples:

Q9. File System Disadvantage:-

1- Data redundancy and inconsistency existed.
2- Difficulty in accessing data.
3- Data isolation.
4- Concurrent access anomalies.
5- Security problems existed.
6- Integrity problems existed.

Q10 What are the main differences between DDB and centralized DB?

A Distributed databases are present different features from traditional (centralized system) so that it is useful to look at the typical features of traditional database and compare them with the features of distributed database. The features which characterize the traditional database approach are:

1. Centralized control

The possibility of providing centralized control over the information resources of whole enterprise or organization was considered of the strongest motivations for introducing database, they were developed as the

evolution of information systems in which each application had its own private files. The fundamental of a database administrator(DBA) was to guarantee the safety of data itself was recognized to be an important investment of the enterprise which required a centralized responsibility. In distributed database, the idea of centralized control is much less emphasized. In general, in DDB it is possible to identify a hierarchical control structure based on a global database administrator, who has the central responsibility of the whole database and on local database administrators, who have the responsibility of their respective local database.

2. Data independence

Means that the actual organization of data is transparent to the application program, having conceptual view of data called conceptual schema. These programs are unaffected by changes in the physical organization of data. In distributed database , data independence has the same importance as in centralized database , however , a new aspect is added to the usual notion of data independence , namely, distribution transparency . By distribution transparency we mean that programs can be written as if the database were not distributed. Thus the correctness of programs is unaffected by the movement of data from one site to another, however , their speed of execution is affected.

3. Reduction of redundancy

In traditional database, redundancy was reduced as far as possible for two reasons:

first, inconsistencies among several copies of the same logical data are automatically avoided by having only one copy, and second, storage space is saved by eliminating redundancy. Reduction of redundancy was obtained by data sharing , by allowing several applications to access the same files and records. In distributed database, however, there are several reasons for considering data redundancy as a desirable feature: first, the locality of

applications can be increased if the data is replicated at all sites where applications need it, and second, the availability of the system can be increased, because a site failure does not stop execution of applications at other sites if the data is replicated.

4. Complex physical structures and efficient access

The reason for providing complex accessing structures is to obtain efficient access to the data. In distributed database, complex accessing structures are not the right tool for efficient access because cannot be provided by using physical structures and it is very difficult to build and maintain such structures.

5. Integrity, recovery, and concurrency control

The database integrity is the means of atomic transactions, it is a sequence of operations which either are performed completely or are not performed at all. In DDB transaction atomicity has a particular flavor: when the system have two sites, the first site is operational and the second site is not operational and when we must transfer application from first site to the second , the transaction should

be aborted. The atomicity property is ensured by the various recovery and concurrency control schemes . When we are dealing with a distributed database system , since several sites may be participating in its execution . The failure of one of these sites, or the failure of a communication link connecting these sites, may result in erroneous computations.

6. Privacy and security

In traditional database, the database administrator, having centralized control, can ensure that only authorized access to the data is performed . But in DDB ,

local administrators are faced with the same problem as database administrator in traditional database.

Q11 What we mean by: DDB, distributed processing?

- DDB: a set of database in distributed system that can appear to application as a single data source.
- distributed processing: the operation that occurs when an application distributed its tasks among different computers in a network.