SS3 Data Processing Lesson Note First Term

NAME:

FIRST TERM E-LEARNING NOTE

SUBJECT: DATA PROCESSING CLASS: SS 3

SCHEME OF WORK

WEEK TOPIC

THEME: DATA MANAGEMENT

1. Revision of last year's work
2. Database Security: (a) Introduction to database security (b) Access control
3. Database Security: (c) Roles of the database administrator in security (d) Encryption
4. Crash Recovery: (a) Introduction to Aries (analysis, redo and undo) (b) other recovery related data structure
5. Crash Recovery: (c) the write-Ahead log protocol (d) Check pointing (e) Media recovery
6. Parallel and Distributed databases: (a) Architecture for parallel databases (b) Introduction to distributed databases
7. Mid term break
8. Parallel and Distributed databases: (c) Distributed DBMS Architecture (d) Storing data in a distributed DBMS
9. Revision

11.13 Examination

Reference Book

A textbook of Data Processing for SSS 3 by Adedapo F O and Mitchell A. S

WEEK TWO

DATE:..

DATE:..

TOPIC: Database Security

CONTENT:

1. Introduction to database security
2. Types of Database Security Control

SUBTOPIC 1: Introduction to database Security

Database management systems are increasingly being used to store information about all aspects of an enterprise. The data stored in a DBMS is often vital to the business interests of the organization and is regarded as a corporate asset.

Database Security

Database security refers to the collective measures used to protect and secure a database or database management software from illegitimate use and malicious threats and attacks. OR

Is the means of ensuring that data is kept from corruption and that access to it is suitable controlled. Thus data security helps to ensure privacy. It also helps in protecting personal data. Data security is part of the larger practice of Information security.

Data is the raw form of information stored as columns and rows in our databases, network servers and personal computers.

Objectives to be considered

There are three main objectives to consider while designing a secure database application.

1. **Secrecy:**Information should not be disclosed to unauthorized users. E.g. a student should not be allowed to examine other students' grades.
2. **Integrity:**Only authorized users should be allowed to modify data. E.g. students may be allowed to see their grades, yet not allowed (obviously!) to modify them.
3. **Availability:** Authorized users should not be denied access. E.g. an instructor who wishes to change a grade should be allowed to do so.

To achieve these objectives, a clear and consistent *security policy* should be developed to described what security measures must be enforced. In particular, we must determine what part of the data is to be protected and which users get access to which portions of the data.

Next, the security mechanisms of the underlying DBMS (and OS, as well as external mechanisms such as securing access to buildings and so on) must be utilized to enforce the policy. We emphasize that security measures must be taken at several levels. Security leaks in the operating system or network connections can circumvent database security mechanisms.

Types of Database Security Control

- Access Control

- Database Audit

- Authentication

- Backup

- Password

- Encryption

Sub-topic 2

Access Control

An Access Control mechanism is a system of controlling data that is accessible to a given giver. This implies the use of authentication and authorization. A computer system that is meant to be used by those authorized to do so must attempt to detect and exclude the unauthorized users.

Approach to Access Control

A DBMS offers two main approaches to access control

1. **Discretionary Access Control:**Discretionary access control (DAC) is a type of security access control that grants or restricts object access via an access policy determined by an object's owner group and/or subjects. DAC mechanism controls are defined by user identification with supplied credentials during authentication, such as username and password. DACs are

discretionary because the subject (owner) can transfer authenticated objects or information access to other users. In other words, the owner determines object access privileges.

2. **Mandatory Access Control**: is a type of access controlin which only the administrator manages the access controls. The administrator defines the usage and access policy, which cannot be modified or changed by users, and the policy will indicate who has access to which programs and files. MAC is most often used in systems where priority is placed on confidentiality.

Database Audit

It involves observing a database so as to be aware of the actions of database users. Database administrators and consultants often set up auditing for security purposes, for example, to ensure that those without the permission to access information do not access it

Authentication

Is the act of confirming the truth of an attribute of a single piece of data claimed true by an entity. In contrast with identification, which refers to the act of stating or otherwise indicating a claim purportedly attesting to a person or thing's identity, authentication is the process of actually confirming that identity. It might involve confirming the identity of a person by validating their identity documents, verifying the authenticity of a website with a digital certificate, determining the age of an artifact by carbon dating, or ensuring that a product is what its packaging and labeling claim to be. In other words, authentication often involves verifying the validity of at least one form of identification.

Backup

Is the process of backing up, refers to the copying and archiving of computer data so it may be used to restore the original after a data loss event. The verb form is to **back up** in two words, whereas the noun is backup.

Password

This is an un-spaced sequence of secret characters used to enable access to a file, program, computer system and other resources.

Encryption

Encryption is the most effective way to achieve data security. To read an encrypted file, you must have access to a secret key or password that enables you to decrypt it. Unencrypted data is called plain text ; encrypted data is referred to as cipher text.

EVALUATION:

(i) What do you mean by data security?

1. How is date secured?

iii. Why is mandatory access control better than discretionary access control?

READING ASSIGNMENT:

Study the topic ‘Database Security’ using students’ textbook

WEEKEND ASSIGNMENT:

OBJECTIVE TEST:

1. DBMS can use ___ to protect information in certain situations where the normal security mechanisms of the DBMS are not adequate. (a) access control (b) encryption (c) data mining (d) security guard
2. ____access control is based on system wide policies that cannot be changed by individual users. (a) discretionary (b) secure (c) mandatory (d) insecure

WEEK THREE

DATE:.......................................

TOPIC: Database Security (Cont.)

CONTENT:

1. Roles of the database administrator in security

2. Encryption

SUBTOPIC 1: Roles of the database administrator in security

The database administrator (DBA) plays an important role in enforcing the security-related aspects of a database design. In conjunction with the owners of the data, DBA will probably also contribute to developing a security policy. The DBA has a special account, which we will call the system account, and is responsible for the overall security of the system.

DBA deals with the following:

- Back up and recover the database.
- Install and configure Oracle software.
- Create new databases.
- Design the database schema and create any necessary database objects.
- Formulate optimal application SQL.
- Ensure database security is implemented to safeguard the data.
- Work closely with application developers and system administrators to ensure all database needs are being met.
- Apply patches or upgrades to the database as needed.
- Maintaining archiveddata
- Contacting database vendorfor technical support
- Generating various reports by querying from database as per need
- Managing and monitoring data replication

Sub-topic 2

Encryption

The basic idea behind encryption is to apply an *encryption algorithm*, which may be accessible to the intruder, to the original data and a user-specified or DBA-specified *encryption key* which is kept secret.

Another approach to encryption, called public-key encryption, has become increasingly popular in recent years. The encryption scheme proposed by Rivest, Shamir and Adleman, called RSA, is a well-known example of public-key encryption. Each authorized user has a public encryption key, known to everyone, and a private description key (used by the decryption algorithm), chosen by the user and known only to him or her.

EVALUATION:

(i) Explain Encryption.

1. What is the role of a database administrator?

READING ASSIGNMENT:

Study the topic 'Crash Recovery' using students' textbook

WEEKEND ASSIGNMENT:

OBJECTIVE TEST:

1. DBMS can use ___ to protect information in certain situations where the normal security mechanisms of the DBMS are not adequate. (a) access control (b) encryption (c) data mining (d) security guard
2. ____access control is based on system wide policies that cannot be changed by individual users. (a) discretionary (b) secure (c) mandatory (d) insecure

WEEK FOUR

DATE:.......................................

TOPIC: CRASH RECOVERY

CONTENT:

1. Introduction to Aries (analysis, redo and undo)
2. Other recovery related data structure

Sub-topic 1

Crash Recovery: Is the process by which the database is moved back to a consistent and usable state. This is done by rolling back incomplete transactions and completing committed transactions that were still in memory when the **crash** occurred.

To be able to do this, the DBMS maintains a record, called the Log, of all write to the database.

What is Log?

This is the history of actions executed by the DBMS. Physically, Log is a file of records stored in stable storage, which is assumed to survive crashes.

For recovery purposes, every page in the database contains the LSN of the most recent log record that describe a change to this page. This LSN is called the pageLSN. Every log record is given a unique ID called the Log Sequence Number (LSN).

Introduction to Aries (analysis, redo and undo)

ARIES (Algorithm for Recovery and Isolation Exploiting Semantics) is a recovery algorithm that is designed to work with a steal, no-force approach.

When the recovery manager is invoked after a crash, restart proceeds in three phases:

1. **Analysis:**Identifies dirty pages in the buffer pool (i.e. changes that have not been written to disk) and active transactions at the time of the crash.
2. **Redo:**Repeats all actions, starting from an appropriate point in the log, and restores the database state to what it was at the time of the crash.
3. **Undo:**Undoes the actions of transactions that did not commit, so that the database reflects only the actions of committed transactions.

There three main principles behind the ARIES recovery algorithm:

1. **Write-ahead logging:**Any change to a database object is first recorded in the log; the record in the log must be written to stable storage before the change to the database object is written to disk.

1. **Repeating history during Redo**: Upon restart following a crash, ARIES retraces all actions of the DBMS before the crash and brings the system back to the exact state that it was in at the time of the crash. Then, it undoes the action of transactions that were still active at the time of the crash (effectively aborting them)

iii. Logging changes during Undo: Database while undoing a transaction are logged in order to ensure that such an action is not repeated in the event of repeated (failures causing) restarts.

Sub-topic 2

Other Recovery related data structures

In addition to the log, the following two tables contain important recovery-related information:

Transaction table: This table contains one entry for each active transaction. The entry contains (among other things) the transaction ID, the status, and a field called *lastLSN*, which is the LSN of the most recent log record for this transaction. The status of a transaction can be that it is in progress, is committed, ort is aborted.

Dirty page table: This table contains one entry for each dirty page in the buffer pool, that is, each page with changes that are not yet reflected on disk. The entry contains a field *reeLSN*, which is the LSN of the first log record that caused the page to become dirty.

pageID recLSN

prevLSN transID type pageID length offset before after

image image

DIRTY PAGE TABLE

transID lastLSN

LOG

TRANSACTION TABLE

EVALUATION:

(i) What do you mean by the log?.

1. What is full form of ARIES?

READING ASSIGNMENT:

Study the topic ‘Crash Recovery’ using students’ textbook

WEEKEND ASSIGNMENT:

OBJECTIVE TEST:

1. This process identifies dirty pages in the buffer pool. (a) Redo (b) Undo

(c) Analysis (d) none of the above

1. CLRb means (a) Combined lateral Register (b) Compensation Log Record

(c) Compromise Log Record (d) Complete Loggers record

WEEK FIVE

DATE:.......................................

TOPIC: CRASH RECOVERY

CONTENT:

1. The write-ahead protocol
2. Check pointing
3. Media recovery

Sub-topic 1

The write-ahead protocol

Before writing a page to disk, every update log record that describes a change to this page must be forced to stable storage. This is accomplished by forcing all log records up and including the one with LSN equal to the pageLSN to stable storage before writing the page to disk.

The important of the WAL protocol cannot be overemphasized – WAL is the fundamental rule that ensures that a record of every change to the database is available while attempting to recover from a crash. If a transaction made a change and committed, the no-force approach means that some of these changes may not have been written to disk at the time of a subsequent crash. Without a record of these changes, there would be no way to ensure that the changes of a committed transaction survive crashes. Note that the definition of a committed transaction is effectively ä transaction whose log records, including a commit record, have all been written to stable storage".

Check pointing

A checkpoint is like a snapshot of the DBMS state, and by taking checkpoints periodically, as we will see, the DBMS can reduce the amount of work to be done during restart in the event of a subsequent crash.

Check pointing in ARIES has three steps.

1. **Begin-checkpoint:**this is written to indicate when the checkpoint starts.
2. **End-checkpoint:**record is constructed, including in it the current contents of the transaction table and the dirty page table, and appended to the log.
3. **Fuzzy Checkpoint:** It is written after end checkpoint is forced to the disk.

Sub-topic 2

Media Recovery

Media recovery is based on periodically making a copy of the database. Because copying a large database object such as a file can take a long time, and the DBMS must be allowed to continue with its operations in the meantime, creating a copy is handled in a manner similar to taking a fussy checkpoint.

When a database object such as a file or a page is corrupted, the copy of that object is brought up-to-date by using the log to identify and reapply the changes of committed transactions and undo the changes of uncommitted transactions (as of the time of the media recovery operation).

EVALUATION:

(i) What are the three phases of restart after crash?

1. What is write ahead logging?

READING ASSIGNMENT:

Study the topic ‘Parallel and Distributed Databases’ using students’ textbook

WEEKEND ASSIGNMENT:

OBJECTIVE TEST:

1. This table contains one entry for each active transaction (a) dirty page table (b) write ahead table (c) LSN table (d) Transaction table

2. Any changes to a database object is first recorded in the log. (a) write ahead logging (b) repeated history (c) logging changes during undo (d) all of the above

WEEK SIX

DATE:.......................................

TOPIC: PARALLEL AND DISTRIBUTED DATABASE

CONTENT:

1. Architecture for parallel database
2. Introduction to distributed databases

Sub-topic 1

Architecture for parallel database

Parallel Database

A parallel database system, is one that seeks to improve performance through parallel implementation of various operations such as loading data, building indexes, and evaluating queries.

Architecture for parallel database

The basic idea behind parallel database is to carry out evaluation steps in parallel whenever possible in order to improve performance.

Three main architectures have been proposed for building DBMSs.

1. **In a shared-memory system,**multiple CPU are attached to an interconnection network and can access a common region of main memory.
2. **In a shared-disk system,**each CPU has a private memory and direct access to all disks through an interconnection network.

iii. **In a shared-nothing system**, each CPU has local main memory and disk space, but no two CPUs can access the same storage area; all communication between CPUs is through a network connection.

SHARED DISK

Advantages of Parallel Databases

1. **Higher Performance:**with more CPUs available to an application, higher speedup and scaleup can be attained.
2. **High Availability:**Nodes are isolated from each other, so failure at one node does not bring the entire system down.
3. **Greater Flexibility:**An OPS environment is extremely flexible. You can allocate or deal-locate instances as necessary.
4. **More Users:**Parallel database technology can make it possible to overcome memory limits, enabling a single system to serve thousands of users.

Disadvantages of Parallel Databases

1. Cost is increased considerably
2. Hug Number of resources are required to support parallelism
3. Managing such system simultaneously becomes difficult.

Sub-topic 2

Introduction to distributed databases

Distributed Database, this is when data is physically stored across several sites, and each site is typically managed by a DBMS that is capable of running independently of the other sites. The location of data items and the degree of autonomy of individual sites have a significant impact on all aspects of the system, including query optimization and processing, concurrency control and recovery. In contrast to parallel database, the distribution of data is governed by factors such as local ownership and increased availability in addition to performance issues.

The classical view of a distributed database system is that the system should make the impact of data distribution transparent.

Below are the properties to be considered:

Distributed Data Independence

Users should be able to ask queries without specifying where the referenced relations, or copies or fragments of the relations, are located. This principle is a natural extension of physical and logical data independence.

Distributed Transaction Atomicity

Users should be able to write transactions that access and update data at several sites just as they would write transactions over purely local data.

Types of Distributed Databases

1. Homogeneous distributed database system
2. Heterogeneous distributed database system

iii. Multi-database system

EVALUATION:

(i) What is a parallel database?

1. What is distributed database?

READING ASSIGNMENT:

Study the topic ‘Parallel and Distributed Databases’ using students’ textbook

WEEKEND ASSIGNMENT:

OBJECTIVE TEST:

1. Breaking a relation into smaller relations (a) Replication (b) server

(c) Fragmentation (d) client

1. ____, each CPU has a private memory and direct access to all disks through an interconnection network. (a) shared memory (b) shared disk (c) shared nothing (d) none of the above.

WEEK 7 MID TERM BREAK

WEEK EIGHT

DATE:.......................................

TOPIC: PARALLEL AND DISTRIBUTED DATABASE

CONTENT:

1. Distributed DBMS Architecture
2. Storing data in a distributed DBMS

Sub-topic 1

Architectures of Distributed Database Systems

The three major distributed DBMS architectures are:

1. Client Server
2. Collaborating Server

iii. Middleware

Client Server

- A client server architecture has a number of clients and a few servers connected in a network.
- A client sends a query to one of the servers. The earliest available server solves it and replies.
- A Client-server architecture is simple to implement and execute due to centralized server system.

Collaborating Server

- Collaborating server architecture is designed to run a single query on multiple servers.
- Servers break single query into multiple small queries and the result is sent to the client.
- Collaborating server architecture has a collection of database servers. Each server is capable for executing the current transactions across the databases.

Middleware

- Middleware architectures are designed in such a way that single query is executed on multiple servers.
- This system needs only one server which is capable of managing queries and transactions from multiple servers.
- Middleware architecture uses local servers to handle local queries and transactions.

Sub-topic 2

STORING DATA IN A DISTRIBUTED DBMS

In a distributed DBMS, relations are stored across several sites. Accessing a relation that is stored at a remote site incurs message-passing costs, and to reduce this overhead, a single relation may be partitioned, or *fragmented* across several sites, with fragments stored at the sites where they are most often accessed or *replicated* at each site where the relation is in high demand.

Fragmentation

This consists of breaking a relation into smaller relations or fragments, and storing the fragments (instead of the relation itself), possibly at different sites.

In *horizontal fragmentation*, each fragment consists of a subset of rows of the original relation.

In *vertical fragmentation*, each fragment consists of a subset of column s of the original relation.

TID

T1

T2

T3

T4

T5

Vertical Fragment Horizontal Fragment

Typically, the tuples that belong to a given horizontal fragment are identified by a selection query; for example, employee tuples might be organized into fragments by city, with all employees in a given city assigned to the same fragment.

Replication

This means that we store several copies of a relation or relation fragment. An entire relation can be replicated at one or more sites. Similarly, one or more fragments of a relation can be replicated at other sites. E.g. if a relation R is fragmented into R1, R2 and R3, there might be just one copy of R1, whereas R2 is replicated at two other sites and R3 is replicated at all sites.

Advantages of distributed databases

1. Management of distributed data with different levels of transparency.
2. Increase reliability and availability

iii. Easier expansion

1. Reflects organizational structure – database fragments are located in the departments they relate to

2. Local autonomy – a department can control the data about them
3. Protection of valuable data

vii. Improved performance

viii. Economics – it costs less to create a network of smaller computers with the power of a single large computer

1. Modularity – system can be modified, added and removed from the distributed database without affecting other modules (system)
2. Reliable transaction
3. Continuous operation

xii. Distributed Query processing

Disadvantages of distributed databases

1. Complexity – extra work must be done by the DBA to ensure that the distributed nature of the system is transparent.
2. Economics – increased complexity and a more extensive infrastructure means extra labour costs.

iii. Security – remote database fragment must be secured, and they are not centralized so the remote sites must be secured as well.

1. Difficult to maintain integrity
2. Inexperience – distributed database are difficult to work with.
3. Lack of standards – there are no tools or methodologies yet to help users convert a centralized DBMS into a distributed DBMS

vii. Database design more complex

viii. Additional software is required

1. Operating system should support distributed environment
2. Concurrency control: it is a major issue. It is solved by locking and time stamping.

EVALUATION:

(i) Compare horizontal and vertical fragmentation

1. What do you mean by fragmentation?

iii. What are the disadvantages of distributed databases?

READING ASSIGNMENT:

Study the topic ‘all the topics for this term’ using students’ textbook

WEEKEND ASSIGNMENT:

OBJECTIVE TEST:

1. This architecture does not allow a single query to span multiple servers (a) client-server (b) collaborating server (c) Synchronous server (d) Heterogeneous server
2. Performance is sustained if the number of CPU and disks are increased in proportion to the amount of data. (a) Linear speed up (b) Linear scale up (c) collaborating server (d) Middleware

Week 10: Revision

Week 11-13 Examination