ST. XAVIER’S COLLEGE

**(Affiliated to Tribhuvan University)**

Maitighar, Kathmandu



DATABASE MANAGEMENT SYSTEM

THEORY ASSIGNMENT #9

**Submitted by:**

Lokendra Puri  
013BSCCSIT023

**Submitted to:**

|  |  |
| --- | --- |
| Er. Sanjay Kumar Yadav  Lecturer |  |

Department of Computer Science

Date of Submission: October 1, 2015

**DATA RECOVERY**

**Data recovery** is a process of salvaging inaccessible data from corrupted or damaged [secondary storage](https://en.wikipedia.org/wiki/Secondary_storage), [removable media](https://en.wikipedia.org/wiki/Removable_media) or [files](https://en.wikipedia.org/wiki/Computer_file), when the data they store cannot be accessed in a normal way. The data is most often salvaged from storage media such as internal or external [hard disk drives](https://en.wikipedia.org/wiki/Hard_disk_drive) (HDDs), [solid-state drives](https://en.wikipedia.org/wiki/Solid-state_drive) (SSDs), [USB flash drives](https://en.wikipedia.org/wiki/USB_flash_drive), [magnetic tapes](https://en.wikipedia.org/wiki/Magnetic_tape_data_storage), [CDs](https://en.wikipedia.org/wiki/CD), [DVDs](https://en.wikipedia.org/wiki/DVD), [RAID](https://en.wikipedia.org/wiki/RAID) subsystems, and other electronic devices. Recovery may be required due to physical damage to the storage device or logical damage to the [file system](https://en.wikipedia.org/wiki/File_system) that prevents it from being [mounted](https://en.wikipedia.org/wiki/Mount_%28computing%29) by the host [operating system](https://en.wikipedia.org/wiki/Operating_system) (OS).

1. **PURPOSE OF DATA RECOVERY**
2. **TYPES OF FAILURE**

**Transaction failure:**

A transaction has to abort when it fails to execute or when it reaches a point from where it can’t go any further. This is called transaction failure where only a few transactions or processes are hurt.

**Logical errors**: A transaction cannot complete due to some internal error condition

**System errors:** The database system must terminate an active transaction due to an error condition (e.g., deadlock)

**System crash:** A power failure or other hardware or software failure causes the system to crash.

**Fail-stop assumption:** Non-volatile storage contents are assumed to not be corrupted by system crash. Database systems have numerous integrity checks to prevent corruption of disk data

**Disk failure:** A head crash or similar disk failure destroys all or part of disk storage.

1. **THE STORAGE HIERARCHY**

Databases are stored in file formats, which contain records. At physical level, the actual data is stored in electromagnetic format on some device. These storage devices can be broadly categorized into three types:



* **Primary Storage** − The memory storage that is directly accessible to the CPU comes under this category. CPU's internal memory (registers), fast memory (cache), and main memory (RAM) are directly accessible to the CPU, as they are all placed on the motherboard or CPU chipset. This storage is typically very small, ultra-fast, and volatile. Primary storage requires continuous power supply in order to maintain its state. In case of a power failure, all its data is lost.
* **Secondary Storage** − Secondary storage devices are used to store data for future use or as backup. Secondary storage includes memory devices that are not a part of the CPU chipset or motherboard, for example, magnetic disks, optical disks (DVD, CD, etc.), hard disks, flash drives, and magnetic tapes.
* **Tertiary Storage** − Tertiary storage is used to store huge volumes of data. Since such storage devices are external to the computer system, they are the slowest in speed. These storage devices are mostly used to take the back up of an entire system. Optical disks and magnetic tapes are widely used as tertiary storage.

1. **BUFFER MANAGEMENT**

Database maintains an in-memory buffer of data blocks. When a new block is needed, if buffer is full an existing block needs to be removed from buffer. If the block chosen for removal has been updated, it must be output to disk.

As a result of the write-ahead logging rule, if a block with uncommitted updates is output to disk, log records with undo information for the updates are output to the log on stable storage first.

No updates should be in progress on a block when it is output to disk can be ensured as follows.

Before writing a data item, transaction acquires exclusive lock on block containing the data item

Lock can be released once the write is completed.

* Such locks held for short duration are called **latches**.

Before a block is output to disk, the system acquires an exclusive latch on the block

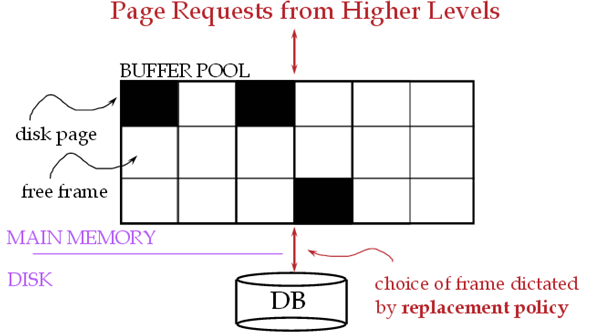
* Ensures no update can be in progress on the block

Database buffer can be implemented either

* in an area of real main-memory reserved for the database, or
* in virtual memory

Implementing buffer in reserved main-memory has drawbacks:

* Memory is partitioned before-hand between database buffer and applications, limiting flexibility.
* Needs may change, and although operating system knows best how memory should be divided up at any time, it cannot change the partitioning of memory.

****

1. **TRANSACTION LOG**

A DBMS uses a **transaction log** to keep track of all transactions that update the database. The information stored in this log is used by the DBMS for a recovery requirement triggered by a ROLLBACK statement, a program’s abnormal termination, or a system failure such as a network discrepancy or a disk crash. After a server failure, for example, Oracle automatically rolls back uncommitted transactions and rolls forward transactions that were committed but not yet written to the physical database.

While the DBMS executes transactions that modify the database, it also automatically updates the transaction log. The transaction log stores record for the beginning of the transaction.

For each transaction component (SQL statement):

* The type of operation being performed (update, delete, insert).
* The names of the objects affected by the transaction (the name of the table).
* The “before” and “after” values for the fields being updated.
* Pointers to the previous and next transaction log entries for the same transaction.
* The ending (COMMIT) of the transaction.

1. **DATA UPDATES**

**Immediate Update:** As soon as a data item is modified in cache, the disk copy is updated.

**Deferred Update:**  All modified data items in the cache is written either after a transaction ends its execution or after a fixed number of transactions have completed their execution.

**Shadow update:** The modified version of a data item does not overwrite its disk copy but is written at a separate disk location.

**In-place update:** The disk version of the data item is overwritten by the cache version.

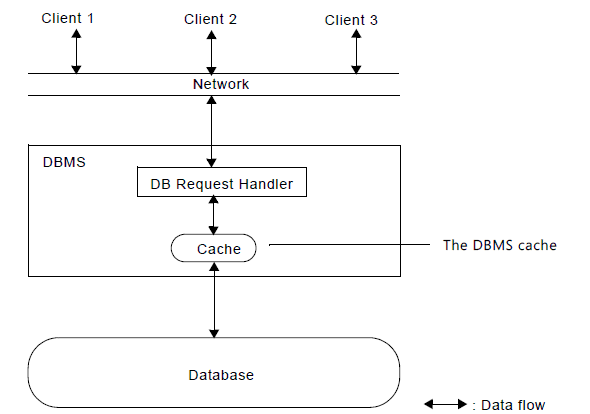
1. **DATA CACHING**

The Database Management System (DBMS) is a memory buffer which stores copies of portions of the database that the DBMS is currently using. Reading from memory is much faster than reading from the disk. The DBMS therefore returns a record more quickly if it is already stored in cache. As long as the required data is stored in cache, the data is immediately available. When the required data is not stored in cache, it must be copied from the disk and then stored in cache.

**DBMS Cache Transparency**

The DBMS cache is transparent to the user. For example, when a user requests data, the data is automatically copied into the cache and stored there. If the data is modified, it is automatically copied back to the physical disk. These data transfers take place automatically. The user does not need to know about the cache.

For example, three users send requests to the DBMS. When user 2 sends a request to read data from the database, the request handler determines whether the desired data can be fetched directly from the cache or whether it must be fetched from a disk.



At the same time, another user can modify a record in a table in the database. The modified data will be written to the DBMS cache, and not to the disk. When this user completes the write transaction (that is, commits the changes), the data in the cache that was modified during the transaction is written to the disk. The cache is then said to be flushed.

The DBMS cache always contains the most recently used data. The cache is continually updated with the relevant data from the database.

The size of the cache greatly affects performance. When you set the size of the cache, you must remember two simple rules:

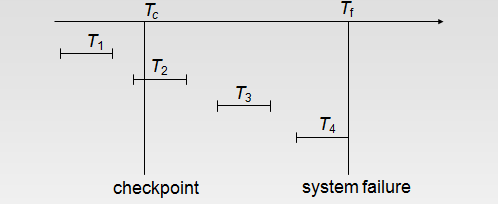
* The more memory you assign to the cache, the more efficient it becomes. (Of course, there is no reason to assign more memory to the cache than the total size of your database.)
* The size of the cache must not exceed the amount of physical memory available on your system. This is because the operating system may swap the cache memory in and out of the disk. This will considerably slow down overall performance.

1. **TRANSACTION ROLL BACK(UNDO) & ROLL FORWARD**

* In the Automated Recovery approach, we introduce a **Log** file – this is a file separate from the data that records all of the changes made to the database by transactions.
* This *transaction log* Includes information helpful to the recovery process such as: A transaction identifier, the date and time, the user running the transaction, *before images* and *after images*
* **Before Image**: A copy of the table record (or data item)  before it was changed by the transaction.
* **After Image**: A copy of the table record (or data item)  after it was changed by the transaction.
* **Rollback**: Undo any partially completed transactions (ones in progress when the crash occurred) by applying the *before images* to the database.
* **Rollforward**: Redo the transactions by applying the *after images* to the database. This is done for transactions that were committed before the crash.
* The Automated Recovery process uses both rollback and rollforward to restore the database.
* In the worst case, we would need to rollback to the last database backup point and then rollforward to the point just before the crash.
* **Checkpoints** can also be taken (less time consuming) in between database saves.
* The DBMS flushes all pending transactions and writes all data to disk and transaction log.
* Database can be recovered from the last checkpoint in much less time.

1. **CHECK POINTING**

* Problem - Prevent Restart from scanning back to the start of the log
* A checkpoint is a procedure to limit the amount of work for Restart
* Commit-consistent checkpointing
  + Stop accepting new update, commit, and abort operations
  + make list of [active transaction, pointer to last log record]
  + flush all dirty pages
  + append a checkpoint record to log, which includes the list
  + resume normal processing
* Database and log are now mutually consistent

****

T1 can be ignored (updates already output to disk due to checkpoint)

T2 and T3 redone

T4 undone

**SHADOW PAGING**

Shadow paging is an alternative to log-based recovery; this scheme is useful if transactions execute serially

**Idea:** maintain *two* page tables during the lifetime of a transaction –the current page table, and the shadow page table.

Store the shadow page table in nonvolatile storage, such that state of the database prior to transaction execution may be recovered.

Shadow page table is never modified during execution.

To start with, both the page tables are identical. Only current page table is used for data item accesses during execution of the transaction.

Whenever any page is about to be written for the first time

* A copy of this page is made onto an unused page. [to be a new current page]
* The current page table is then made to point to the copy
* The update is performed on the copy

1. **RECOVERY SCHEMES (WAL: WRITE AHEAD LOGGING PROTOCOLS)**

When **in-place** update (immediate or deferred) is used then log is necessary for recovery and it must be available to recovery manager. This is achieved by **Write-Ahead Logging (WAL)** protocol. WAL states that

* + **For Undo**: Before a data item’s AFIM is flushed to the database disk (overwriting the BFIM) its BFIM must be written to the log and the log must be saved on a stable store (log disk).
  + **For Redo**: Before a transaction executes its commit operation, all its AFIMs must be written to the log and the log must be saved on a stable store.

1. **FAILURE WITH LOSS OF NON VOLATILE STORAGE(GENERAL CONCEPTS)**

A volatile storage like RAM stores all the active logs, disk buffers, and related data. In addition, it stores all the transactions that are being currently executed. What happens if such a volatile storage crashes abruptly? It would obviously take away all the logs and active copies of the database. It makes recovery almost impossible, as everything that is required to recover the data is lost.

Following techniques may be adopted in case of loss of volatile storage −

* + We can have checkpoints at multiple stages so as to save the contents of the database periodically.
  + A state of active database in the volatile memory can be periodically dumped onto a stable storage, which may also contain logs and active transactions and buffer blocks.
  + <dump> can be marked on a log file, whenever the database contents are dumped from a non-volatile memory to a stable one.

1. **RECOVERY IN MULTI-DATABASE SYSTEM**

To maintain the atomicity of a multidatabase transaction, it is necessary to have a two-level recovery mechanism. A global recovery manager, or coordinator, is needed to maintain information needed for recovery, in addition to the local recovery managers and the information they maintain (log, tables).

The coordinator usually follows a protocol called the two-phase commit protocol, whose two phases can be stated as follows:

**Phase 1:** When all participating databases signal the coordinator that the part of the multidatabase transaction involving each has concluded, the coordinator sends a message "prepare for commit" to each participant to get ready for committing the transaction. Each participating database receiving that message will force-write all log records and needed information for local recovery to disk and then send a "ready to commit" or "OK" signal to the coordinator. If the force-writing to disk fails or the local transaction cannot commit for some reason, the participating database sends a "cannot commit" or "not OK" signal to the coordinator. If the coordinator does not receive a reply from a database within a certain time out interval, it assumes a "not OK" response.

**Phase 2:** If all participating databases reply "OK," and the coordinator’s vote is also "OK," the transaction is successful, and the coordinator sends a "commit" signal for the transaction to the participating databases. Because all the local effects of the transaction and information needed for local recovery have been recorded in the logs of the participating databases, recovery from failure is now possible. Each participating database completes transaction commit by writing a [commit] entry for the transaction in the log and permanently updating the database if needed. On the other hand, if one or more of the participating databases or the coordinator have a "not OK" response, the transaction has failed, and the coordinator sends a message to "roll back" or UNDO the local effect of the transaction to each participating database. This is done by undoing the transaction operations, using the log.

The net effect of the two-phase commit protocol is that either all participating databases commit the effect of the transaction or none of them do. In case any of the participants—or the coordinator—fails, it is always possible to recover to a state where either the transaction is committed or it is rolled back. A failure during or before Phase 1 usually requires the transaction to be rolled back, whereas a failure during Phase 2 means that a successful transaction can recover and commit.