04 Database Storage – Part II
**ADMINISTRIVIA**

**Project #1** will be released today (September 13\textsuperscript{th})

**Project #0** was due last night at 11:59pm

→ If you did not complete it with a score of 100%, you will not be able to continue in the course.
Vaccination Database Tech Talks
→ Mondays @ 4:30pm (starting today)
→ https://db.cs.cmu.edu/seminar2021-dose2
The DBMS assumes that the primary storage location of the database is on non-volatile disk.

The DBMS's components manage the movement of data between non-volatile and volatile storage.
The most common layout scheme is called **slotted pages**.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:

→ The # of used slots
→ The offset of the starting location of the last slot used.
The most common layout scheme is called **slotted pages**.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
- The # of used slots
- The offset of the starting location of the last slot used.
The most common layout scheme is called slotted pages.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
→ The # of used slots
→ The offset of the starting location of the last slot used.
The most common layout scheme is called slotted pages.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:

→ The # of used slots
→ The offset of the starting location of the last slot used.
The most common layout scheme is called slotted pages.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
→ The # of used slots
→ The offset of the starting location of the last slot used.
Instead of storing tuples in pages, the DBMS only stores log records.

The system appends log records to the file of how the database was modified:
→ Inserts store the entire tuple.
→ Deletes mark the tuple as deleted.
→ Updates contain the delta of just the attributes that were modified.
Instead of storing tuples in pages, the DBMS only stores log records.

The system appends log records to the file of how the database was modified:

→ Inserts store the entire tuple.
→ Deletes mark the tuple as deleted.
→ Updates contain the delta of just the attributes that were modified.
To read a record, the DBMS scans the log backwards and "recreates" the tuple to find what it needs.
To read a record, the DBMS scans the log backwards and "recreates" the tuple to find what it needs.
LOG-STRUCTURED FILE ORGANIZATION

To read a record, the DBMS scans the log backwards and "recreates" the tuple to find what it needs.

Build indexes to allow it to jump to locations in the log.

```
INSERT id=1,val=a
INSERT id=2,val=b
DELETE id=4
UPDATE val=X (id=3)
UPDATE val=Y (id=2)
```
To read a record, the DBMS scans the log backwards and "recreates" the tuple to find what it needs.

Build indexes to allow it to jump to locations in the log.
To read a record, the DBMS scans the log backwards and "recreates" the tuple to find what it needs.

Build indexes to allow it to jump to locations in the log.

Periodically compact the log.
LOG-STRUCTURED FILE ORGANIZATION

To read a record, the DBMS scans the log backwards and "recreates" the tuple to find what it needs.

Build indexes to allow it to jump to locations in the log.

Periodically compact the log.
LOG-STRUCTURED FILE ORGANIZATION

To read a record, the DBMS scans the log backwards and "recreates" the tuple to find what it needs.

Build indexes to allow it to jump to locations in the log.

Periodically compact the log.
To read a record, the DBMS scans the log backwards and "recreates" the tuple to find what it needs.

Build indexes to allow it to jump to locations in the log.

Periodically compact the log.

Page

id=1, val=a
id=2, val=Y
id=3, val=X
DELETE id=4
Log-structured compaction coalesces larger log files into smaller files by removing unnecessary records.
LOG-STRUCTURED COMPACTION

Compaction coalesces larger log files into smaller files by removing unnecessary records.

Level Compaction
Compaction coalesces larger log files into smaller files by removing unnecessary records.

**Level Compaction**

Level 0

Sorted Log File
Compaction coalesces larger log files into smaller files by removing unnecessary records.

**Level Compaction**

![Sorted Log File](Image)

![Sorted Log File](Image)
LOG-STRUCTURED COMPACTION

Compaction coalesces larger log files into smaller files by removing unnecessary records.

**Level Compaction**

- Level 0: Sorted Log File
  - Sorted Log File

Compa**
Compaction coalesces larger log files into smaller files by removing unnecessary records.

**Level Compaction**

Level 0

Sorted Log File

Compaction

Level 1
Compaction coalesces larger log files into smaller files by removing unnecessary records.

Level Compaction
Compaction coalesces larger log files into smaller files by removing unnecessary records.

**Level Compaction**
Compaction coalesces larger log files into smaller files by removing unnecessary records.

**Level Compaction**

- **Level 0**: Sorted Log File, Sorted Log File
- **Level 1**: Sorted Log File, Sorted Log File
- **Level 2**: Sorted Log File

Compaction
Compaction coalesces larger log files into smaller files by removing unnecessary records.

**LOG-STRUCTURED COMPACTION**

**Level Compaction**

- **Level 0**
  - Sorted Log File
  - Sorted Log File

- **Level 1**
  - Sorted Log File
  - Sorted Log File

- **Level 2**
  - Sorted Log File

**Universal Compaction**

- Sorted Log File
- Sorted Log File
- Sorted Log File
- Sorted Log File
Compaction coalesces larger log files into smaller files by removing unnecessary records.

**Level Compaction**

- Level 0
  - Sorted Log File
  - Sorted Log File

- Level 1
  - Sorted Log File
  - Sorted Log File

- Level 2
  - Sorted Log File

**Universal Compaction**

- Sorted Log File
- Sorted Log File
- Sorted Log File
- Sorted Log File

LOG-STRUCTURED COMPACTION
TODAY'S AGENDA

Data Representation
System Catalogs
Storage Models
TUPLE STORAGE

A tuple is essentially a sequence of bytes.
It's the job of the DBMS to interpret those bytes into attribute types and values.

The DBMS's catalogs contain the schema information about tables that the system uses to figure out the tuple's layout.
DATA REPRESENTATION

INTEGER / BIGINT / SMALLINT / TINYINT
→ C/C++ Representation

FLOAT / REAL vs. NUMERIC / DECIMAL
→ IEEE-754 Standard / Fixed-point Decimals

VARCHAR / VARBINARY / TEXT / BLOB
→ Header with length, followed by data bytes.

TIME / DATE / TIMESTAMP
→ 32/64-bit integer of (micro)seconds since Unix epoch
DATA REPRESENTATION

INTEGER/BIGINT/SMALLINT/TINYINT
→ C/C++ Representation

FLOAT/REAL vs. NUMERIC/DECIMAL
→ IEEE-754 Standard / Fixed-point Decimals

VARCHAR/VARBINARY/TEXT/BLOB
→ Header with length, followed by data bytes.

TIME/DATE/TIMESTAMP
→ 32/64-bit integer of (micro)seconds since Unix epoch
Inexact, variable-precision numeric type that uses the "native" C/C++ types.
→ Examples: FLOAT, REAL/Doubles

Store directly as specified by IEEE-754.

Typically faster than arbitrary precision numbers but can have rounding errors…
VARIABLE PRECISION NUMBERS

Rounding Example

```c
#include <stdio.h>

int main(int argc, char* argv[]) {
    float x = 0.1;
    float y = 0.2;
    printf("x+y = %f\n", x+y);
    printf("0.3 = %f\n", 0.3);
}
```
VARIABLE PRECISION NUMBERS

Rounding Example

```c
#include <stdio.h>

int main(int argc, char* argv[]) {
    float x = 0.1;
    float y = 0.2;
    printf("x+y = %f\n", x+y);
    printf("0.3 = %f\n", 0.3);
}
```

Output

```
x+y = 0.300000
0.3 = 0.300000
```
VARIABLE PRECISION NUMBERS

Rounding Example

```c
#include <stdio.h>

int main(int argc, char* argv[]) {
    float x = 0.1;
    float y = 0.2;
    printf("x+y = %.20f\n", x+y);
    printf("0.3 = %.20f\n", 0.3);
}
```

Output

```
x+y = 0.30000000000000000000
0.3 = 0.30000000000000000000
```
```c
#include <stdio.h>

int main(int argc, char* argv[]) {
    float x = 0.1;
    float y = 0.2;
    printf("x+y = %.20f\n", x+y);
    printf("0.3 = %.20f\n", 0.3);
}
```

**Output**

```
x+y = 0.30000001192092895508
0.3 = 0.29999999999999998890
```
Numeric data types with (potentially) arbitrary precision and scale. Used when rounding errors are unacceptable.
→ Example: NUMERIC, DECIMAL

Many different implementations.
→ Example: Store in an exact, variable-length binary representation with additional meta-data.
→ Can be less expensive if you give up arbitrary precision.
typedef unsigned char NumericDigit;
typedef struct {
    int ndigits;
    int weight;
    int scale;
    int sign;
    NumericDigit *digits;
} numeric;
typedef unsigned char NumericDigit;

typedef struct {
    int ndigits;
    int weight;
    int scale;
    int sign;
    NumericDigit *digits;
} numeric;
### POSTGRES: NUMERIC

1. **# of Digits**
2. **Weight of 1st Digit**
3. **Scale Factor**
4. **Positive/Negative/NaN**
5. **Digit Storage**

```c
typedef unsigned char NumericDigit;

typedef struct {
    int ndigits;
    int weight,
    int scale;
    int sign;
    NumericDigit *digits;
} numeric;
```
typedef unsigned char NumericDigit;

typedef struct {
    int ndigits;
    int weight;
    int scale;
    int sign;
    NumericDigit* digits;
} numeric;

/* add_var() -
 * Full version of add functionality on variable level (handling signs).
 * result might point to one of the operands too without danger.
 */

int PGTYPESnumeric_add(numeric *var1, numeric *var2, numeric *result)
{
    /* Decide on the signs of the two variables what to do */
    if (var1->sign == NUMERIC_POS)
    {
        if (var2->sign == NUMERIC_POS)
        {
            /* Both are positive result = +(ABS(var1) + ABS(var2)) */
            if (add_abs(var1, var2, result) != 0)
            {
                return -1;
                result->sign = NUMERIC_POS;
            }
        }
        else
        {
            /* var1 is positive, var2 is negative. Must compare absolute values */
            switch (cmp_abs(var1, var2))
            {
                case 0:
                { /* zero */
                    /* ABS(var1) == ABS(var2)
                     * result = ZERO */
                    zero_var(result);
                    result->rscale = Max(var1->rscale, var2->rscale);
                    result->dscale = Max(var1->dscale, var2->dscale);
                    break;
                }
                case 1:
                { /* ABS(var1) > ABS(var2) */
                    /* result = +(ABS(var1) - ABS(var2)) */
                    /* */
                    if (sub_abs(var1, var2, result) != 0)
                    {
                        return -1;
                        result->sign = NUMERIC_POS;
                    }
                    break;
                }
                case -1:
                { /* ABS(var1) < ABS(var2) */
                    /* result = -(ABS(var2) - ABS(var1)) */
                    break;
                }
            }
        }
    }
}
MYSQL: NUMERIC

typedef int32 decimal_digit_t;
struct decimal_t {
    int intg, frac, len;
    bool sign;
    decimal_digit_t *buf;
};
typedef int32 decimal_digit_t;

struct decimal_t {
    int intg, frac, len;
    bool sign;
    decimal_digit_t *buf;
};

# of Digits Before Point
# of Digits After Point
Length (Bytes)
Positive/Negative
Digit Storage
typedef int32 decimal_digit_t;

struct decimal_t {
    int intg, frac, len;
    bool sign;
    decimal_digit_t *buf;
};
typedef int32 decimal_digit_t;

struct decimal_t {
    int intg, frac, len;
    bool sign;
    decimal_digit_t *buf;
};

int11 = ROUND_UP(from1->intg, from1->frac),
int2 = ROUND_UP(from2->intg, from2->frac),
intg1 = std::max(intg1, intg2),
intg2 = std::max(intg1, intg2),
error:

static int do_add(const decimal_t *from1, const decimal_t *from2, decimal_t *to) {
    to->sign = from1->sign ^ from2->sign;
    to->frac = from1->frac + from2->frac;
    to->buf[0] = 0; /* safety */
    int to->len = from1->len + from2->len + 1;
    if (unlikely(to->len > DIG_MAX + 1)) /* there is a need for extra word because of carry? */
        fix_int_frac_error(&to->len, &intg1, &frac1, &error);
    to->len = max_decimal(to->len, DIG_PER_DEC1, 0, to);
    int max = from1->max + from2->max;
    if (unlikely(error == E_DEC_OVERFLOW))
        return error;
    buf0 = to->buf + intg0 + frac0;
    buf0[0] = 0; /* safety */
    if (unlikely(buf0[0] != 0)) /* is there a need for extra word because of carry? */
        fix_int_frac_error(&buf0[0], &intg2, &frac2, &error);
    buf1 = to->buf + intg1 + frac1;
    buf1[0] = 0; /* safety */
    if (unlikely(buf1[0] != 0)) /* is there a need for extra word because of carry? */
        fix_int_frac_error(&buf1[0], &intg2, &frac2, &error);

    # of Digits Before Point  # of Digits After Point  Length (Bytes)  Positive/Negative  Digit Storage

Most DBMSs don't allow a tuple to exceed the size of a single page.

To store values that are larger than a page, the DBMS uses separate overflow storage pages.
→ Postgres: TOAST (>2KB)
→ MySQL: Overflow (>½ size of page)
→ SQL Server: Overflow (>size of page)
Most DBMSs don't allow a tuple to exceed the size of a single page.

To store values that are larger than a page, the DBMS uses separate overflow storage pages.

→ Postgres: TOAST (>2KB)
→ MySQL: Overflow (>½ size of page)
→ SQL Server: Overflow (>size of page)
Most DBMSs don't allow a tuple to exceed the size of a single page.

To store values that are larger than a page, the DBMS uses separate overflow storage pages.
→ Postgres: TOAST (>2KB)
→ MySQL: Overflow (>½ size of page)
→ SQL Server: Overflow (>size of page)
Some systems allow you to store a really large value in an external file. Treated as a **BLOB** type.

→ Oracle: **BFILE** data type
→ Microsoft: **FILESTREAM** data type

The DBMS **cannot** manipulate the contents of an external file.

→ No durability protections.
→ No transaction protections.
Some systems allow you to store a really large value in an external file. Treated as a **BLOB** type.

→ Oracle: **BFILE** data type
→ Microsoft: **FILESTREAM** data type

The DBMS **cannot** manipulate the contents of an external file.
→ No durability protections.
→ No transaction protections.
Some systems allow you to store a really large value in an external file. Treated as a **BLOB** type.

→ Oracle: **BFILE** data type

→ Microsoft: **FILESTREAM** data type

The DBMS **cannot** manipulate the contents of an external file.

→ No durability protections.

→ No transaction protections.
A DBMS stores meta-data about databases in its internal catalogs.

→ Tables, columns, indexes, views
→ Users, permissions
→ Internal statistics

Almost every DBMS stores the database's catalog inside itself (i.e., as tables).

→ Wrap object abstraction around tuples.
→ Specialized code for "bootstrapping" catalog tables.
You can query the DBMS’s internal **INFORMATION_SCHEMA** catalog to get info about the database.

→ ANSI standard set of read-only views that provide info about all the tables, views, columns, and procedures in a database

DBMSs also have non-standard shortcuts to retrieve this information.
**ACCESSING TABLE SCHEMA**

List all the tables in the current database:

```sql
SELECT * 
FROM INFORMATION_SCHEMA.TABLES 
WHERE table_catalog = '<db name>'; 
```

<table>
<thead>
<tr>
<th>Database</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postgres</td>
<td>\d;</td>
</tr>
<tr>
<td>MySQL</td>
<td>SHOW TABLES;</td>
</tr>
<tr>
<td>SQLite</td>
<td>.tables</td>
</tr>
</tbody>
</table>

**SQL-92**
ACCESSING TABLE SCHEMA

List all the tables in the student table:

```
SELECT * FROM INFORMATION_SCHEMA.TABLES
WHERE table_name = 'student'
```

SQL-92

```
\d student;
```

Postgres

```
DESCRIBE student;
```

MySQL

```
.schema student
```

SQLite
On-Line Transaction Processing (OLTP)
→ Fast operations that only read/update a small amount of data each time.

On-Line Analytical Processing (OLAP)
→ Complex queries that read a lot of data to compute aggregates.

Hybrid Transaction + Analytical Processing
→ OLTP + OLAP together on the same database instance
DATABASE WORKLOADS

Operation Complexity

Complex

OLAP

Simple

OLTP

Workload Focus
BIFURCATED ENVIRONMENT

OLTP Data Silos

OLAP Data Warehouse
BIFURCATED ENVIRONMENT

Transactions

OLTP Data Silos

OLAP Data Warehouse
BIFURCATED ENVIRONMENT

Transactions

OLTP Data Silos

Extract
Transform
Load

OLAP Data Warehouse
BIFURCATED ENVIRONMENT

Transactions

OLTP Data Silos

Extract
Transform
Load

OLAP Data Warehouse
BIFURCATED ENVIRONMENT

Transactions

OLTP Data Silos

Extract
Transform
Load

OLAP Data Warehouse
BIFURCATED ENVIRONMENT

Transactions

Extract Transform Load

OLTP Data Silos

Analytical Queries

OLAP Data Warehouse
BIFURCATED ENVIRONMENT

Transactions

OLTP Data Silos

Extract Transform Load

Analytical Queries

OLAP Data Warehouse
**BIFURCATED ENVIRONMENT**

- Transactions
- Analytical Queries

**HTAP Database**

**Extract**

**Transform**

**Load**

**OLAP Data Warehouse**
The relational model does not specify that we have to store all of a tuple's attributes together in a single page.

This may not actually be the best layout for some workloads...
**CREATE TABLE** useracct (  
  userID INT PRIMARY KEY,  
  userName VARCHAR UNIQUE,  
);  

**CREATE TABLE** pages (  
  pageID INT PRIMARY KEY,  
  title VARCHAR UNIQUE,  
  latest INT  
  # REFERENCES revisions (revID),  
);  

**CREATE TABLE** revisions (  
  revID INT PRIMARY KEY,  
  userID INT REFERENCES useracct (userID),  
  pageID INT REFERENCES pages (pageID),  
  content TEXT,  
  updated DATETIME  
);
CREATE TABLE useracct (  
    userID INT PRIMARY KEY,  
    userName VARCHAR UNIQUE,  
);
CREATE TABLE useracct ( 
userID INT PRIMARY KEY, 
userName VARCHAR UNIQUE, 
);

CREATE TABLE pages ( 
pageID INT PRIMARY KEY, 
title VARCHAR UNIQUE, 
latest INT 
REFERENCES revisions (revID), 
);

CREATE TABLE revisions ( 
revID INT PRIMARY KEY, 
userID INT REFERENCES useracct (userID), 
pageID INT REFERENCES pages (pageID), 
content TEXT, 
updated DATETIME 
);
On-line Transaction Processing:
→ Simple queries that read/update a small amount of data that is related to a single entity in the database.

This is usually the kind of application that people build first.

SELECT P.*, R.*
FROM pages AS P
INNER JOIN revisions AS R
ON P.latest = R.revID
WHERE P.pageID = ?

UPDATE useracct
SET lastLogin = NOW(), hostname = ?
WHERE userID = ?

INSERT INTO revisions
VALUES (?, ?, ..., ?)
On-line Analytical Processing:
→ Complex queries that read large portions of the database spanning multiple entities.

You execute these workloads on the data you have collected from your OLTP application(s).

```
SELECT COUNT(U.lastLogin),
       EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY
       EXTRACT(month FROM U.lastLogin)
```
The DBMS can store tuples in different ways that are better for either OLTP or OLAP workloads.

We have been assuming the **n-ary storage model** (aka "row storage") so far this semester.
The DBMS stores all attributes for a single tuple contiguously in a page.

Ideal for OLTP workloads where queries tend to operate only on an individual entity and insert-heavy workloads.
The DBMS stores all attributes for a single tuple contiguously in a page.
The DBMS stores all attributes for a single tuple contiguously in a page.
The DBMS stores all attributes for a single tuple contiguously in a page.
The DBMS stores all attributes for a single tuple contiguously in a page.
N-ARY STORAGE MODEL (NSM)

```sql
SELECT * FROM useracct
WHERE userName = ?
AND userPass = ?
```
**N-ARY STORAGE MODEL (NSM)**

```
SELECT * FROM useracct
WHERE userName = ?
AND userPass = ?
```
**N-ARY STORAGE MODEL (NSM)**

```
SELECT * FROM useracct
WHERE userName = ?
AND userPass = ?
```
SELECT * FROM useracct
WHERE userName = ?
AND userPass = ?
**N-ARY STORAGE MODEL (NSM)**

```
SELECT * FROM useracct
WHERE userName = ?
AND userPass = ?
```
**N-ARY STORAGE MODEL (NSM)**

```
SELECT * FROM useracct
WHERE userName = ?
AND userPass = ?

INSERT INTO useracct
VALUES (?, ?, ..., ?)
```
N-ARY STORAGE MODEL (NSM)

```
SELECT * FROM useracct
WHERE userName = ?
AND userPass = ?

INSERT INTO useracct
VALUES (?, ?, ..., ?)
```
SELECT COUNT(U.lastLogin),
    EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
N-ARY STORAGE MODEL (NSM)

```
SELECT COUNT(U.lastLogin),
    EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```
N-ARY STORAGE MODEL (NSM)

```
SELECT COUNT(U.lastLogin),
       EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```
**N-ARY STORAGE MODEL (NSM)**

```sql
SELECT COUNT(U.lastLogin),
       EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```
N-ARY STORAGE MODEL (NSM)

SELECT COUNT(U.lastLogin),
    EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
N-ARY STORAGE MODEL (NSM)

```
SELECT COUNT(U.lastLogin),
       EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```
**N-ARY STORAGE MODEL (NSM)**

```sql
SELECT COUNT(U.lastLogin),
       EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```
### N-ARY STORAGE MODEL (NSM)

```sql
SELECT COUNT(U.lastLogin),
       EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```

#### NSM Disk Page

```
<table>
<thead>
<tr>
<th></th>
<th>userID</th>
<th>userName</th>
<th>userPass</th>
<th>hostname</th>
<th>lastLogin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Header</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Header</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Header</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Header</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
N-ARY STORAGE MODEL (NSM)

```sql
SELECT COUNT(U.lastLogin),
       EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```
N-ARY STORAGE MODEL

Advantages
→ Fast inserts, updates, and deletes.
→ Good for queries that need the entire tuple.

Disadvantages
→ Not good for scanning large portions of the table and/or a subset of the attributes.
The DBMS stores the values of a single attribute for all tuples contiguously in a page. → Also known as a "column store"

Ideal for OLAP workloads where read-only queries perform large scans over a subset of the table’s attributes.
The DBMS stores the values of a single attribute for all tuples contiguously in a page.
→ Also known as a "column store".
The DBMS stores the values of a single attribute for all tuples contiguously in a page. → Also known as a "column store".
The DBMS stores the values of a single attribute for all tuples contiguously in a page.
→ Also known as a "column store".

**DECOMPOSITION STORAGE MODEL (DSM)**
The DBMS stores the values of a single attribute for all tuples contiguously in a page. → Also known as a "column store".
SELECT COUNT(U.lastLogin),
    EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
DECOMPOSITION STORAGE MODEL (DSM)

```
SELECT COUNT(U.lastLogin),
       EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```
DECOMPOSITION STORAGE MODEL (DSM)

```
SELECT COUNT(U.lastLogin),
   EXTRACT(month FROM U.lastLogin) AS month
FROM useracct AS U
WHERE U.hostname LIKE '%.gov'
GROUP BY EXTRACT(month FROM U.lastLogin)
```
TUPLE IDENTIFICATION

Choice #1: Fixed-length Offsets
→ Each value is the same length for an attribute.

Choice #2: Embedded Tuple Ids
→ Each value is stored with its tuple id in a column.

Offsets

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Embedded Ids

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
DECOMPOSITION STORAGE MODEL (DSM)

Advantages
→ Reduces the amount wasted I/O because the DBMS only reads the data that it needs.
→ Better query processing and data compression (more on this later).

Disadvantages
→ Slow for point queries, inserts, updates, and deletes because of tuple splitting/stitching.
**DSM SYSTEM HISTORY**

1970s: Cantor DBMS

1980s: DSM Proposal

1990s: SybaseIQ (in-memory only)

2000s: Vertica, VectorWise, MonetDB

2010s: Everyone
### DSM System History

<table>
<thead>
<tr>
<th>Decade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970s</td>
<td>Cantor DBMS</td>
</tr>
<tr>
<td>1980s</td>
<td>DSM Proposal</td>
</tr>
<tr>
<td>1990s</td>
<td>SybaseIQ (in-memory only)</td>
</tr>
<tr>
<td>2000s</td>
<td>Vertica, VectorWise, MonetDB</td>
</tr>
<tr>
<td>2010s</td>
<td>Everyone</td>
</tr>
</tbody>
</table>
DSM SYSTEM HISTORY

1970s: Cantor DBMS
1980s: DSM Proposal
1990s: SybaseIQ (in-memory only)
2000s: Vertica, VectorWise, MonetDB
2010s: Everyone
DSM SYSTEM HISTORY

1970s: Cantor DBMS
1980s: DSM Proposal
1990s: SybaseIQ (in-memory only)
2000s: Vertica, VectorWise, MonetDB
2010s: Everyone
The storage manager is not entirely independent from the rest of the DBMS.

It is important to choose the right storage model for the target workload:
→ OLTP = Row Store
→ OLAP = Column Store
DATABASE STORAGE

Problem #1: How the DBMS represents the database in files on disk.

Problem #2: How the DBMS manages its memory and move data back-and-forth from disk.
Problem #1: How the DBMS represents the database in files on disk.

Problem #2: How the DBMS manages its memory and move data back-and-forth from disk.