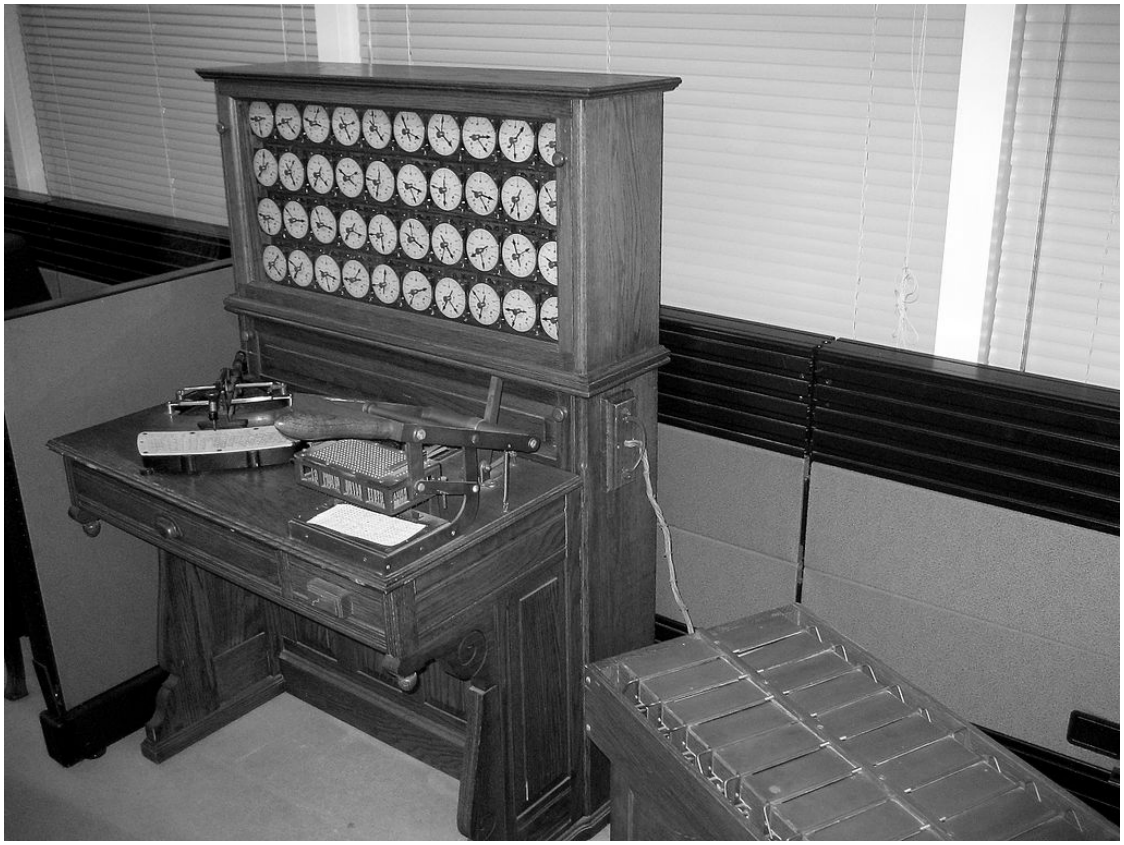
Introduction

History

- 1890: Herman Hollerith
  - punchcard machine
    - 1st automatic information processing equipment
    - census 1890 / 1900 were processed

History

- 1890: Herman Hollerith



– By Adam Schuster (Flickr: Proto IBM)

History

- 1911: IBM founded (Hollerith et. al.)
  - automatic data processing makes possible
    - control of industry production
    - tax calculation not possible without

- 1935: Social Security Act:
  - records of 26 million individuals

History

- The Census Bureau continued to use updated versions of Herman Hollerith's 1890 electric counting machine through the 1940 census.

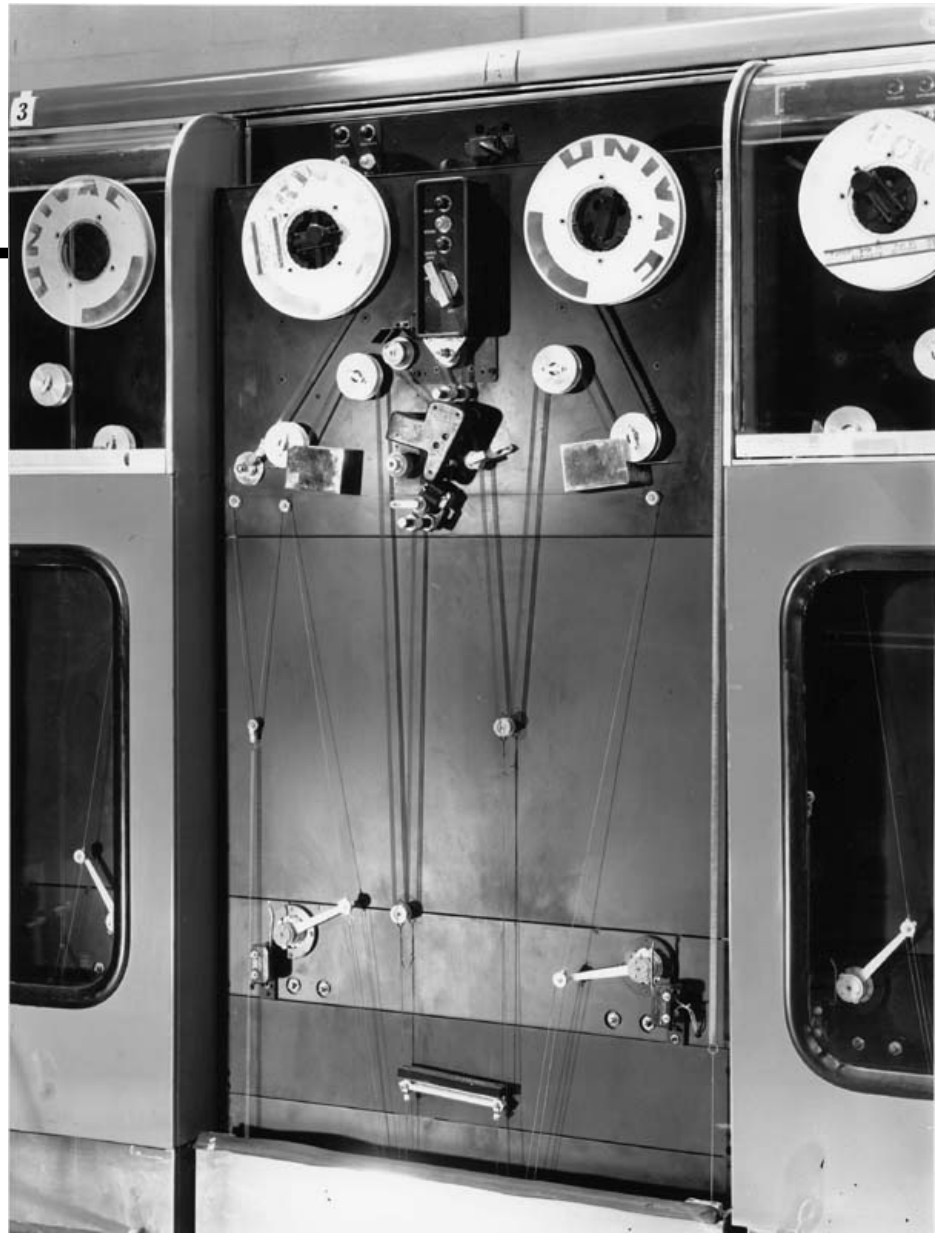
- Processing and tabulation technology took a great leap forward during World War II, when the War Department began to explore the use of electronic digital computers.

– https://www.census.gov/history/www/innovations/technology/univac_i.html

History

- 1951 Census Bureau begins using Universal Automatic Computer (**UNIVAC**).
  - Data could be input using magnetic computer tape.
  - It was tabulated using vacuum tubes and state-of-the-art circuits then either printed out or stored on more magnetic tape.
    - https://www.census.gov/history/www/innovations/technology/univac_i.html

History



– http://www.computer-history.info/Page4.dir/pages/Univac.dir/

History

– http://www.computer-history.info/Page4.dir/pages/Univac.dir/

History

Standardization efforts

- computing for commercial market
- techniques for:
  - Data access
  - Data quality
  - Security
  - Control

- tapes => hard disks (serial => random)

– Matthias Mann for TechnoCircle HVB Information Services

History

Standardization efforts

- 1971: formal standard:
  - Codasyl approach to database management
    - hierarchical database model
    - network database model
    - navigational databases
  - The programmer as navigator –Bachmann 1973

History

Edgar F. Codd (IBM)

- dissatisfied with "taking the old line view, that the burden of finding information should be placed on users …"

- 1970:
  - "A relational Model of Data for large shared Data Banks"

History

Edgar F. Codd (IBM)

- Independence of Data from the Hardware- and Storage Implementation

- automatic navigation to the data set
  - high level nonprocedural language

- “Keys” replace pointers

History

Edgar F. Codd (IBM)

- theoretical proposal
  - no practical design or implementation

– Matthias Mann for TechnoCircle HVB Information Services

Definitions

Database

- collection of interrelated data
  - represents some aspects of the real world
  - is logically coherent with some inherent meaning
  - has an intended group of users and applications

– Bart Baesens

Definitions

Database

- a single large repository of data... which is used simultaneously by many departments and users.

- ...some collection of persistent data that is used by the application systems of some given enterprise.

– C.J. Date

Definitions

Database

- data that is stored more-or-less permanently in a computer –Ullman

- A collection of files under the control of a database management system. –Borys

Definitions

Database Management System (DBMS)

- Between the physical database itself (i.e. the data as actually stored) and the users of the system is the DBMS. All requests for access to data from users are handled by the DBMS... One general function of the DBMS is the shielding of database users from hardware-level details. In other words, the DBMS provides a view of the data that is elevated above the hardware level, and supports user operations… that are expressed in terms of that higher-level view.

–C. J. Date (Introduction to Database Systems, 6th edition, pg 7)

Definitions

Database Management System (DBMS)

- A DBMS is a Software System that provides **efficient**, **convenient**, and **safe** **multi-user** storage of and access to **massive** amounts of **persistent** data

– Jennifer Widom

DBMS Benefits

Efficient:

- Don't search all files in order to get price of one book, get all customers from northern California, get bestselling books from last week

– Jennifer Widom

DBMS Benefits

Convenient

- Simple commands to
  - find a specific book, list all books in a certain category and price range, generate an order history, produce sales figures grouped by state, etc.
- Also unpredicted queries should be easy

– Jennifer Widom

DBMS Benefits

Massive data:

- DBMS's carefully tuned for performance

– Jennifer Widom

DBMS Benefits

Multi-user

- Problems may appear similar to concurrent programming problems (synchronization, semaphores, etc.)

– Jennifer Widom

DBMS Benefits

Multi-user

- Problems may appear similar to concurrent programming problems (synchronization, semaphores, etc.)
- BUT: data not main-memory variables

– Jennifer Widom

DBMS Benefits

Multi-user:

- Problems may appear similar to file system concurrent access

– Jennifer Widom

DBMS Benefits:

Multi-user:

- Problems may appear similar to file system concurrent access
- BUT: want to control at smaller granularity

– Jennifer Widom

DBMS Benefits:

Multi-user

- Multiple writers via locking mechanism
- Detection and resolution of deadlocks

DBMS Benefits

Safe

- From system failures
- From malicious users

– Jennifer Widom

DBMS Benefits

Rollback / recovery mechanism

- Ensure physical integrity after failure
- Logical integrity dependent upon application

DBMS Benefits

Centralized control of data

- Reduce redundancy
- Avoid inconsistency
- Share and secure data

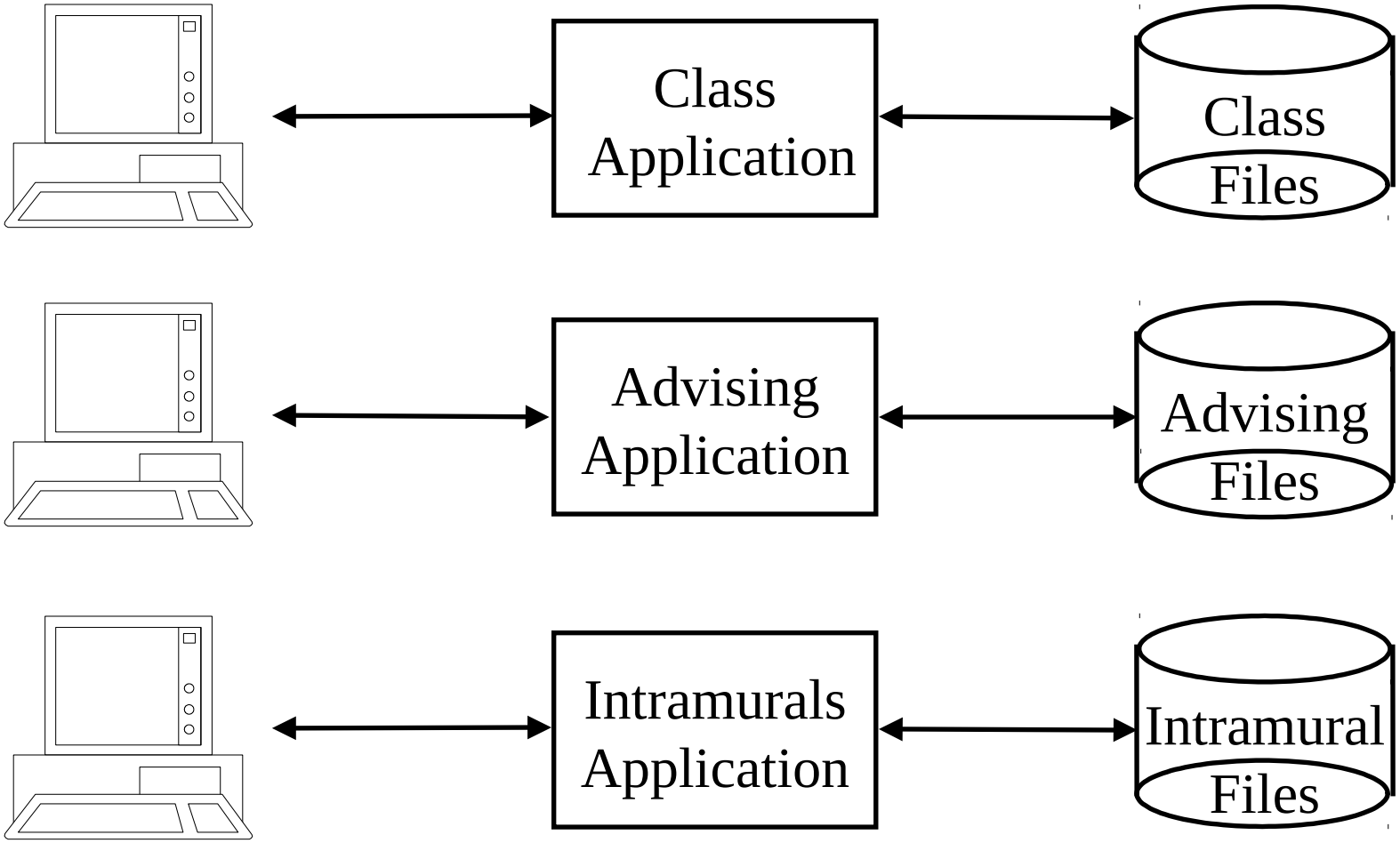
– Borys

DBMS Benefits

Data independence

- Programs unaware of physical structures
- Application independent structures
- Separation of programs and data

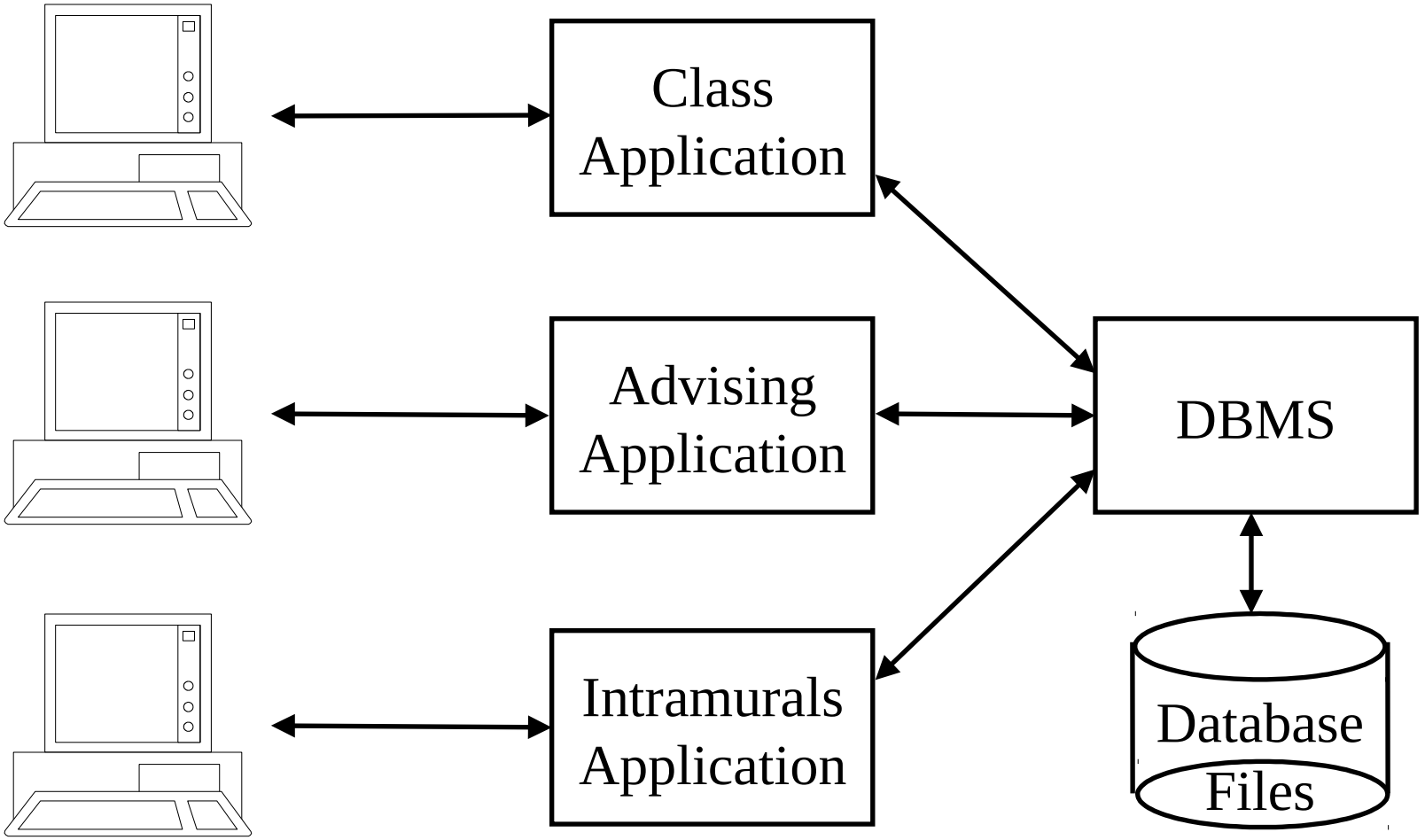
Traditional File Approach



Application-Centric

– Borys

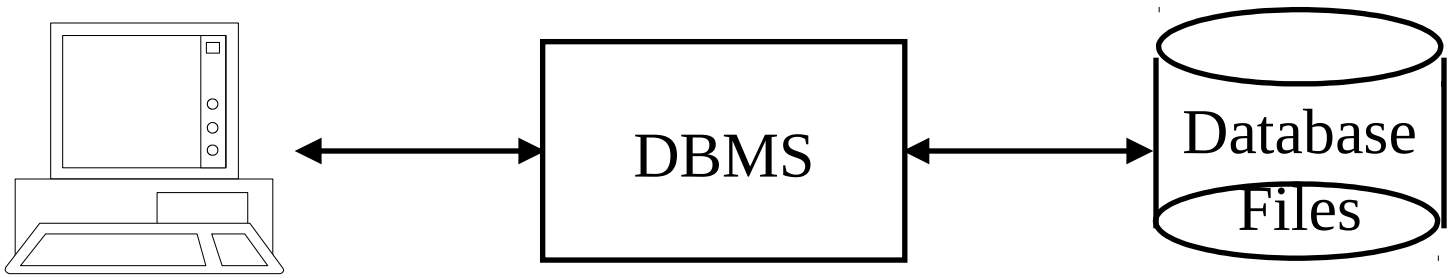
Database Approach



Data-Centric

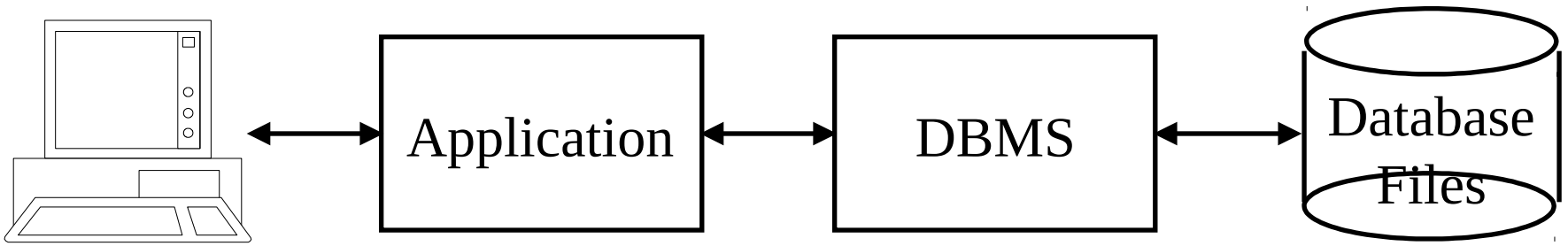
– Borys

Interact Directly With DBMS



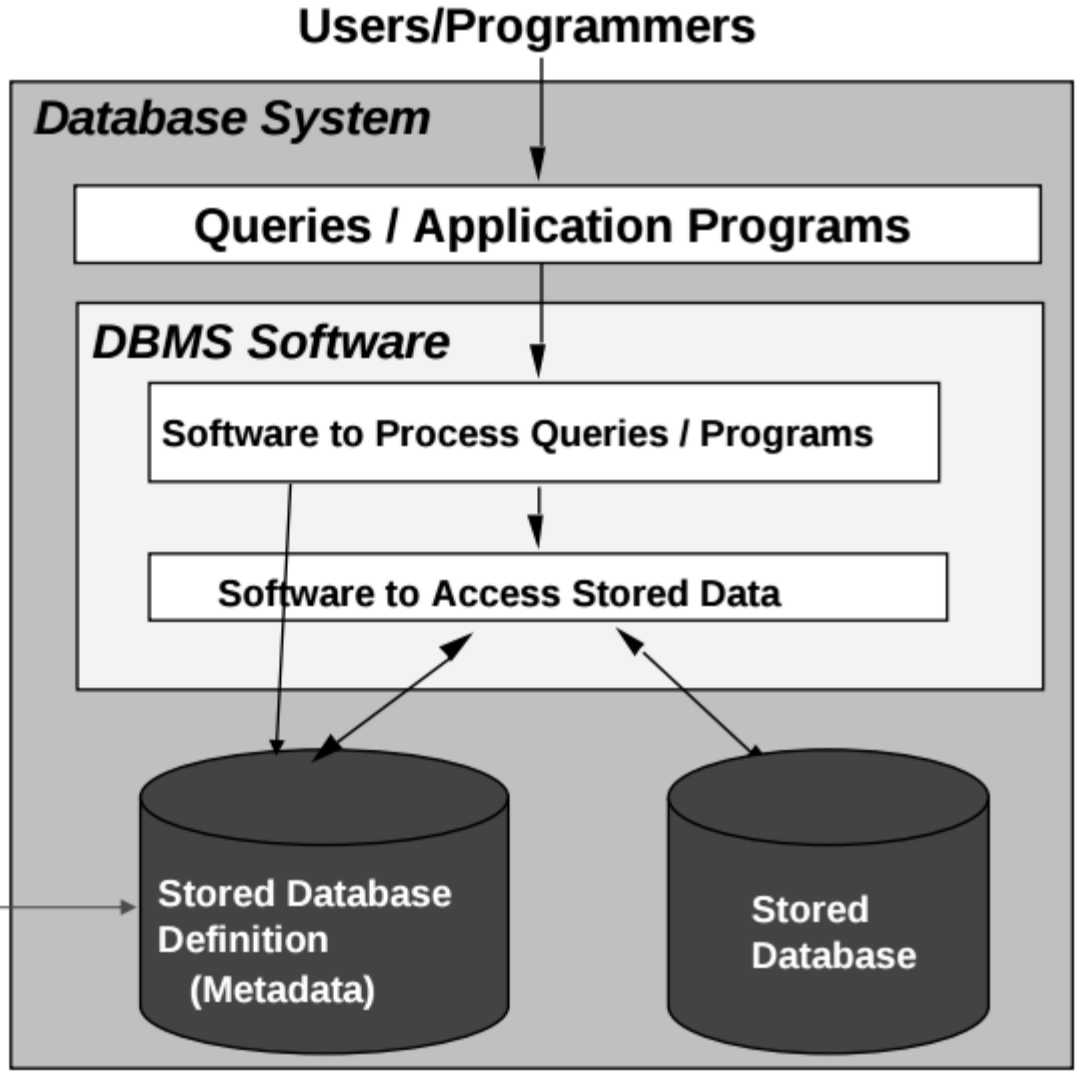
– Borys

Interact Via Application



– Borys

Interact via abstractions



"Catalogue",
"Data dictionary"

– Werner Nutt

Data Model

1. An abstract, self-contained, logical definition of the data structures, data operators, and so forth, that together make up the abstract machine with which users interact (contrast implementation).

2. A model of the persistent data of some particular enterprise (in other words, a conceptual or logical database design).

–The Relational Database Dictionary pg 24, C.J. Date

Data Model

- Defined to consist of three components:
  1. A collection of **data object types**, which form the basic building blocks for any database that conforms to the model.
  2. A collection of **general integrity rules**, which constrain the set of occurrences of those object types that can legally appear in any such database.
  3. A collection of **operators**, which can be applied to such object occurrences for retrieval and other purposes.

– Date (paraphrasing Codd) pg. 103

Data Model

data object types

- Two small but important distinctions are needed:
  - An object type is the base concept or idea of an object; for example, the concept of a table or index.
  - An object instance is an example of an object type. For example, a table called CUSTOMER is an instance of the object type TABLE.

– https://www.techopedia.com/definition/24081/database-object

Data Model

data object types

- Most of the major database engines offer the same set of major database object types:
  - *Tables*
  - *Indexes*
  - *Sequences*
  - *Views*
  - *Synonyms*

– https://www.techopedia.com/definition/24081/database-object

Data Model

Integrity

- The term “integrity” is used to refer to the accuracy or correctness of the data in the database.
- We regard the database as “correct” iff it satisfies the logical AND of all known rules.

A system that does not support much in the way of integrity support will only have very weak sense of what it means for the database to be “correct.”

– Date pgs 441-444

Data Model

Integrity rules in general have three components:

- A name.
- The constraint that must be satisfied
  - specified by means of a truth-valued expression.
  - The integrity rule is said to be **satisfied** iff the constraint evaluates to true
  - The integrity rule is said to be **violated** iff the constraint evaluates to false.
- A violation response.

– Date pgs 441-444

Data Model

Integrity rule specification (in Postgres)

```
ALTER TABLE distributors
    ADD CONSTRAINT zipchk
    CHECK (char_length(zipcode) = 5);
```

Data Model

Operator

- Either a read-only operator or an update operator

– The New Relational Database Dictionary by C.J. Date

Data Model

Read-only operator

- Generally, a function that, when invoked, updates nothing but returns a value, of a type declared when the operator in question is defined.
- A read-only operator invocation thus denotes a value; i.e., it’s an expression.
- It can be nested inside other expressions.

– The New Relational Database Dictionary by C.J. Date

Data Model

Update operator

- An operator that, when invoked, returns no value but updates a variable.
- An update operator invocation doesn't denote a value—loosely speaking, it's a statement, not an expression.
- It can't be nested inside an expression.
- Every update operator invocation is logically equivalent to some assignment (possibly a multiple assignment).

– The New Relational Database Dictionary by C.J. Date

Data Models

Models:

- Hierarchical Model
- Network Model
- The Relational Model

Data Models

Models:

- Hierarchical Model
- Network Model
- The Relational Model

More on the first two next class...

Relational Model

The Relational Model

- Used by most commercial DBMSs
- Very simple model
- Enables simple, clean, declarative query languages

Relational Model

The Relational Model

- Database = set of relations (or tables)
  - each with a distinct name
  - Example: Student, Course

Relational Model

The Relational Model

- Each relation has a set of attributes (or columns), with a distinct name within its relation.

Relational Model

The Relational Model

- Each attribute has a type (or domain).

Relational Model

The Relational Model

- schema = complete description of structure of relations in database: relation names, attribute names, types, etc.
- instance = actual contents (tuples) of relations

Relational Model

The Relational Model

- Keys: A key for a relation is a set of attributes such that no two tuples can have the same values for all of their key attributes.
- Key values identify specific tuples.
- Other tuples may use key values as logical "pointers".

RDBMS Definition

- A relational database is a database that is perceived by the user as a collection of normalized relations. –Codd

– Date pg. 98

RDBMS Definition

- A relational database is a database that is perceived by the user as a collection of normalized relations. –Codd

- The phrase "*perceived by the user*" is crucial.

– Date pg. 98

RDBMS Definition

- A relational database is a database that is perceived by the user as a collection of normalized relations. –Codd

- The phrase "*perceived by the user*" is crucial.

  - The ideas of the relational model apply at the external and conceptual levels of the system, not the internal level.

– Date pg. 98

RDBMS Definition

- A relational database is a database that is perceived by the user as a collection of normalized relations. –Codd

- The phrase "*perceived by the user*" is crucial.

  - The relational model represents a database system at a level of abstraction that is somewhat removed from the details of the underlying machine.

– Date pg. 98

RDBMS Definition

- A relational database is a database that is perceived by the user as a collection of normalized relations. –Codd

- The phrase "*perceived by the user*" is crucial.

  - The relational model can be regarded as a rather abstract programming language that is oriented specifically towards database applications.

– Date pg. 98

RDBMS Benefits

- **Integration**
  - Many organizations integrate between applications by sharing a database.
  - Different apps talk to the same tables.
  - Essential question for anyone considering a non-relational technology.

– Fowler

RDBMS Costs

- **Impedance Mismatch**
  - What you have in memory and what you have in the database is different.

– Fowler

RDBMS Costs

- **Impedance Mismatch**
  - Tabular data is what databases are all about.
    - (Relations are the mathematical term.)
  - Everything needs to be partitioned out into individual rows and tables.
  - Doesn't necessarily fit how you want to
    - present the data
    - manipulate the data in memory.

– Fowler

RDBMS Costs

- **Impedance Mismatch**
  - Tabular data is what databases are all about.
    - (Relations are the mathematical term.)
  - You often have richer structures
    - Hierarchical structures.
    - Not easily represented in the database.

– Fowler

RDBMS Costs

- **Impedence Mismatch**
  - Particularly talked about in terms of
  - Object-Oriented systems.
    - "The Object-Relational Mismatch Problem"
  - Happens also in
    - Functional programming,
    - C
    - elsewhere

– Fowler

RDBMS Components

- Data Definition Language (DDL)
  - Declare data structures
- Data Manipulation Language (DML)
  - Usually embedded in application programs
- Database Control System (DBCS)
  - Runtime system to service DB requests

RDBMS Utilities

- Load / unload data
- Disk file space analysis
- I/O response time statistics
- Report on additional data, e.g.
  - High water mark of concurrent users, resources
  - Database access strategy for a specific statement
- File repair

– Borys

RDBMS Pain

- Initial purchase / lease costs
- Maintenance costs
  - Annual fees
  - Personnel time
- Consumes disk space and main memory
- Complex product to learn and use properly

– Borys

RDBMS Gain

- A highly integrated database yields:
- Reduced redundancy
  - Elimination of code to copy and maintain duplicated data values
  - Saves disk space
- Ability to (relatively) easily add new data structures and applications especially after “core” applications and data structures exist
  - This is database nirvana

– Borys

Data Management Terms

- Record
  - Physical grouping of data fields
- Page
  - Unit of physical I/O from the DBMS's and OS's perspective
- Record / page relationship
  - Common for an entire record to fit on one page
  - A record could be fragmented across ≥ 2 pages

– Borys

More Terminology

- Database management system
  - Manipulates record occurrences via page requests
- File manager
  - Inventories files on disk
- Disk manager
  - Initiates and controls physical disk I/Os

– Borys

Deep Thoughts

- A database management system is a tool, not a solution. (1987)
- Databases are created to model a piece of the real world. (1993)
- Plenty of technology available, just need to figure out how to apply it wisely. (1996)

– Borys

More Deep Thoughts...

- Users are not an annoyance; they are the reason you have a job. (2003)
- Power tools + amateurs = poor results (2005)

– Borys

And A Few More…

- If it initially seems easy, it'll be damn hard. If it initially appears hard, you're doomed! (2008)
- Be careful what you measure. Bad metrics are worse than no metrics. (2011)
- You can make a lot of money fixing other people's mistakes. (2014)

– Borys