Understanding RDBMS

IT4GIS

Keith T. Weber, GISP
GIS Director
ISU-GIS Training and Research Center

FUNDAMENTALS

Pocatello | Idaho Falls | Meridian | Twin Falls
Idaho State UNIVERSITY

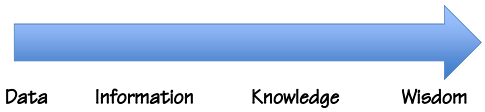
RDBMS

- Relational Database Management System
- The "I" in GIS (*Information*)

Pocatello | Idaho Falls | Meridian | Twin Falls
Idaho State UNIVERSITY

BTW

- The Data-to-Wisdom Pathway



cf., https://en.wikipedia.org/wiki/DIKW_pyramid

Database software...

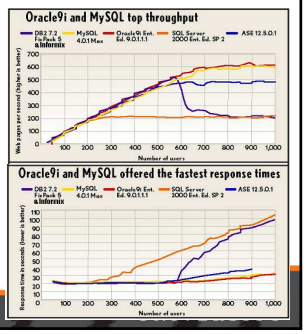
- Light Duty
- Medium Duty
- Heavy Duty

Database software...

- Light Duty
- Medium Duty
- Heavy Duty

Relational Databases

- Why are databases important?
  - Drive many business processes
  - Store large amounts of data
  - Retrieve data quickly



Spreadsheets vs. Databases

- **Integrity!**
- **Structure**

RDBMS CONCEPTS AND TERMS

Independence

- Physical
- Logical

Logical Consistency Example

- Character based database design
  - FirstName (1-4)
  - LastName (5-10)
  - Address (11-46)

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
P a u l B u n y u n 1 0 0 M a i n   S t r e e t     P o c a t e l l o   I D 8 3 2 0 1
J o h n H e n r y   1 5 0 M a i n   S t r e e t     P o c a t e l l o   I D 8 3 2 0 1
```

- Record #1: Paul Bunyun, 100 Main Street, Pocatello, ID 83201

NOTE: Record #2 starts at character #47

What Happens When We Add a New Field?

- New Field = ZIP+4 (47-50)
  - Example, Paul's ZIP+4 = 1234
- Scripts written and referring to the original design will fail
- Record #1: Paul Bunyun, 100 Main Street, Pocatello, ID 83201
- Record #2: 1234 Johnhe, nry 150 Main Street, Pocatello, ID 8

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
P a u l B u n y u n 1 0 0 M a i n   S t r e e t     P o c a t e l l o   I D 8 3 2 0 1
      J o h n H e n r y   1 5 0 M a i n   S t r e e t     P o c a t e l l o   I D 8
```

Integrity

- Important for consistency and transaction management.
- Types:
  - Domain: all values come from predefined domains or are null
  - Redundancy: problems can occur as a result of repetitive storage that is not consistently updated and from stored data that is derived from other stored data. Redundant data must be consistent.

Integrity Types (cont'd)

- Constraint: Business integrity. Stored data must not violate business rules.
- Entity: Every record must be uniquely identifiable (index field or ObjectID)
- Referential: Relationships must not be ambiguous. Two types…

  Cascading or non-cascading

Enforcing Integrity Rules

- Programmatic
- Systematic

Key Fields versus Index fields

- Unique Identifiers are Index fields
- Relate fields are Key fields
  - –Primary key
  - –Foreign key

RDBMS STRUCTURE

Database Tables

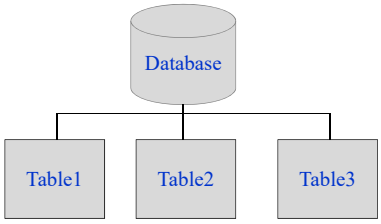


Table Structure

	COLUMN 1 (FIELD OR ATTRIBUTE)	COLUMN 2
ROW 1 (RECORD OR ENTITY)	VALUE	
ROW 2		

Data Value Types

Type Name	Storage Occupied/ data value	Valid Domain Range
Short Integer	2 bytes	-35768 to 32767
Long Integer	4 bytes	-2147483648 to 2147483647
Float	4 bytes	Any number from n^{-45} to n^{38}
Double	8 bytes	Any number from n^{-324} to n^{308}
Text (string)	10 + max. length = bytes	Any alphanumeric characters
Date	8 bytes	Jan 1, 100 to Dec. 31 9999
LOB (variant)	22 + max. length = bytes	Any alphanumeric characters

(BTW) Raster Data Types Worth Knowing

- **1_BIT**—A 1-bit unsigned integer. The values can be 0 or 1.
- **2_BIT**—A 2-bit unsigned integer. The values supported can be from 0 to 3.
- **4_BIT**—A 4-bit unsigned integer. The values supported can be from 0 to 15.
- **8_BIT_UNSIGNED**—An unsigned 8-bit data type. The values supported can be from 0 to 255.
- **8_BIT_SIGNED**—A signed 8-bit data type. The values supported can be from -128 to 127.
- **16_BIT_UNSIGNED**—A 16-bit unsigned data type. The values can range from 0 to 65,535.
- **16_BIT_SIGNED**—A 16-bit signed data type. The values can range from -32,768 to 32,767.
- **32_BIT_UNSIGNED**—A 32-bit unsigned data type. The values can range from 0 to 4,294,967,295.
- **32_BIT_SIGNED**—A 32-bit signed data type. The values can range from -2,147,483,648 to 2,147,483,647.
- **32_BIT_FLOAT**—A 32-bit data type supporting decimals.
- **64_BIT**—A 64-bit data type supporting decimals.

Making Sense of all this…

- Recall, there are 8 bits in 1 byte
- Cross-reference
  - 8-bit is byte data
  - 16-bit is short integer (2 bytes)
  - 32-bit (signed or unsigned) is long integer (4 bytes)
  - 32-bit (float) is single-precision floating point (4 bytes)
  - 64-bit is double-precision floating point (8 bytes)

DATABASE DESIGN

Basic Steps in Database Design

- Understand and document the business' needs.
  - Problem statement
  - Business object types
  - Business relationships
  - Business constraints
- Create an ERM
- Data and process inventory
- Develop tuple types
- Tuple types to tables
- Integrity
- Populate the database

READING A BUSINESS STATEMENT

Identify Candidate Classes

- A *candidate* class may or may not remain a class throughout the design process
  - A *candidate* class may or may not become a table
- Try not to think about tables when reading the business statement at this point

Think Object-Oriented

- Classes are nouns
- A noun is a "person, places, and things"

SCHOOLHOUSE ROCK

Methods

- Identifying *candidate* methods allows us to better understand how the business operates and how the Enterprise uses GIS data.
- A method is a behavior…a relationship between classes
- Ultimately, a connection between two tables
- The candidate methods will describe an inheritance, aggregation, or dependency relationship

And now…Verbs

- *Candidate* methods are verbs
  - They show action
  - They are behaviors

Create an Entity Relationship Model (ERM)

- Symbolized.
  - Standard Representation
  - Attribute Representation
  - Entity Instance Representation

BUILDINGS
K BldgNum: 126 Name: Graveley Hall Type: Education

Relationships

- Determine the **Relationship** between **Entity Types**.
- Add these to the **ERM**

(more about database relationship classes later in the semester)

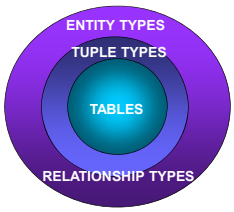
Data and Process Inventory

- Database Dictionary
  - BldgName
    - The name of the building
  - Type
    - Primary use of the building (e.g., 0 = Unknown or n/a; 1 = Education, 2 = Offices, etc.)
  - Floors
    - The total number of floors

Develop Tuple Types

- Use your ERM with relationships
- Perform a "Walk-through" exercise
  - Simulate data is being added/retrieved in your database.
- Create another, more matjure ERM using Attribute Representation

Tuple Types to Tables



Normalization

- First-Fifth Form Normal (**1FN , 2FN ,…5FN**)
- Academic
- Applied

1FN

- "All values are atomic"
  - Each cell in the table contains only a single data value
- Eliminate repeating groups
  - Puppy_Trick1, Puppy_Trick2, …

.

Check this (1FN)…

Field Name	Data Type	Length	Description	Examples
OWNER1	Text	100	Owner of Parcel	John Smith
OWNER2	Text	100	Additional owner of parcel	Mary Smith
MAIL_ADD1	Text	100	Mailing address of owner	1234 S Paper Rd
MAIL_ADD2	Text	100	Additional mailing address of owner	Apt 5C
MAIL_CITY	Text	100	Mailing city of owner	Anaktuvuk Pass
MAIL_STATE	Text	2	Mailing state of Owner	AK
MAIL_ZIP	Text	10	Mailing U.S. zip code of owner	99721-0000
MAIL_CNTRY	Text	100	Mailing country of owner	USA
SITE_ADD	Text	100	Site address of property	6769 W Stapley Ave
SITL_CITY	Text	100	City of property	Nampa
SITL_ZIP	Text	10	Zip code of property	83653-0000
CATEGORY1	Text	2	Assessed land use	01
CATEGORY2	Text	2	Assessed land use	02
CATEGORY3	Text	2	Assessed land use	03
CATEGORY4	Text	2	Assessed land use	04
CATEGORY5	Text	2	Assessed land use	05
CATEGORY6	Text	2	Assessed land use	06
CATEGORY7	Text	2	Assessed land use	07
IRR_ACRES	Double	TBD	Irrigated acreage by land use category	7.560
DRY_ACRES	Double	TBD	Dryland agricultural acreage by land use category	1.648
ZONING	Text	TBD	Zoning category	RSW
DESC1	Text	TBD	Property description	PAR #9500 of S2SE4
DESC2	Text	TBD	Property description	SEC26 3N 1E
DESC3	Text	TBD	Property description	#449100-S
DESC4	Text	TBD	Property description	
DESC5	Text	TBD	Property description	
SUBDIV	Text	TBD	Subdivision name	Happy Valley
VALUATION	Integer	TBD	Net assessed value of property	100.000

Pocatello | Idaho Falls | M

Idaho State UNIVERSITY

2FN

- Satisfy 1FN and…
- Redundant **data** must be eliminated
  - How?
  - Example: Puppy_ID, Trick_ID, Trick_Name

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State UNIVERSITY

Check this (2FN)…

Field Name	Data Type	Length	Description	Examples
OWNER1	Text	100	Owner of Parcel	John Smith
OWNER2	Text	100	Additional owner of parcel	Mary Smith
MAIL_ADD1	Text	100	Mailing address of owner	1234 S Paper Rd
MAIL_ADD2	Text	100	Additional mailing address of owner	Apt 5C
MAIL_CITY	Text	100	Mailing city of owner	Anaktuvuk Pass
MAIL_STATE	Text	2	Mailing state of Owner	AK
MAIL_ZIP	Text	10	Mailing U.S. zip code of owner	99721-0000
MAIL_CNTRY	Text	100	Mailing country of owner	USA
SITE_ADD	Text	100	Site address of property	6769 W Stapley Ave
SITL_CITY	Text	100	City of property	Nampa
SITL_ZIP	Text	10	Zip code of property	83653-0000
CATEGORY1	Text	2	Assessed land use	01
CATEGORY2	Text	2	Assessed land use	02
CATEGORY3	Text	2	Assessed land use	03
CATEGORY4	Text	2	Assessed land use	04
CATEGORY5	Text	2	Assessed land use	05
CATEGORY6	Text	2	Assessed land use	06
CATEGORY7	Text	2	Assessed land use	07
IRR_ACRES	Double	TBD	Irrigated acreage by land use category	7.560
DRY_ACRES	Double	TBD	Dryland agricultural acreage by land use category	1.648
ZONING	Text	TBD	Zoning category	RSW
DESC1	Text	TBD	Property description	PAR #9500 of S2SE4
DESC2	Text	TBD	Property description	SEC26 3N 1E
DESC3	Text	TBD	Property description	#449100-S
DESC4	Text	TBD	Property description	
DESC5	Text	TBD	Property description	
SUBDIV	Text	TBD	Subdivision name	Happy Valley
VALUATION	Integer	TBD	Net assessed value of property	100.000

Pocatello | Idaho Falls | M

Idaho State UNIVERSITY

3FN

- Satisfy 1NF and 2FN and…
- No non-key attributes are dependent on other non-key attributes.
  - Example: Appointment_ID, Name, Date, Time, Species

After Normalization

- New **tuple types** will be created.
- New **tables** will be planned.
- Many-many relationships will be handled using associative tables (bridge tables).

De-Normalization

- What? Is this heresy?

Designing the Actual RDBMS

- Visual modeling based upon your **ERM** and **Tuple type** model.
- Implementation of integrity rules based upon your business constraints.

Populate the Database

- Questions and concerns to revisit
  - –Null data
  - –Reporting discrepancies and variations
  - –Measuring or estimating methods
  - –Client utility/efficiency

The Last Step?

Professional Hints and Tips

- Using Google drive, sharing files/folders and communicating this in email

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State UNIVERSITY

Questions?

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State UNIVERSITY