BAIT507 - Data Management

MBAN - BAIT507

Data Management for Business Analytics

Simon Goring

About Me

Simon Goring

Assistant Scientist, University of Wisconsin

http://goring.org

Geoinformatics, paleogeosciences, natural language processing, data visualization, data management and cyberinfrastructure

Research Goals

Neotoma Paleoecology Database http://neotomadb.org

Research Goals

Neotoma Paleoecology Database http://neotomadb.org

Research Goals

Research Goals

Course Overview

- Course Topics
- Evaluation
- Group Project

Topic	Lectures	Concepts
Background	Lectures 1 – 3	What is a database; connecting to databases; working in R.
Building Queries	Lectures 3 – 6	Moving from raw data to data structures, and back again.
Advanced Database Concepts	Lectures 7 – 10	Database design, constraints and testing.

Background

3 Lectures

- What is a database?
- How do we use a database?
- Basic queries
- Preparing a database

Buildng Queries

4 Lectures

- Normalizing data to save space
- Joining data back together
- Complex queries

Advanced db Queries

4 Lectures

- Advanced data design
- Constraints & advanced indices

- Query optimization & hypothesis testing
- Structured data
- Big data tools

Grading

Assignments

- 3 x 10% each

Group Assignment

- 30% total
- Assigned Sept 5, Due Oct 3.
- Written & Oral components.

Individual Assignments

Assignment One:

Install Postgres, R and associated development tools locally and describe the process.

Individual Assignments

Assignment Two:

Connect to a remote database and perform basic queries using R in an RMarkdown notebook.

Individual Assignments

Assignment Three:

Build a simple database using assigned data. Perform basic analysis in an RMarkdown notebook.

Group Assignment

Fun Times!

Group Assignment

In assigned groups of three or four.

- Define a business analytics problem
- Discover relevant data
- Refine your question and model a database structure
- Write a group report

Group Assignment

Presentations

Four Pecha Kucha style presentations - Present one, submit all four

Shared with the class

Engagement

Grade: 10%

- Graded on in-class and online discussion

Expectations, Grading & Office Hours

Expectations

- **Be on time** (you and I!)
- No screens except during coding sessions
- Academic dishonesty is not tolerated

Grading

- Grading will be done ASAP
- Rubrics will be available with the assignments
- All grades will be provided before the final exam

Office Hours and Feedback

- Use Canvas discussions as much as possible please
- Office hours are Thursday 12 - 1:30pm ICICS/CS 187

Introduction to Data Management

Note: Data are fundamental to any analysis.

Key Points

- What are key data management concerns?
- How does ACID address these?
- What is a Database/DBMS?

Introduction to Data Management

The way data is used and managed is fundamental to any analysis.

Data Management is Crucial

- Data is highly variable but mission critical
- Poor data management makes everything harder
- Good data management makes everything easier
- **Good data management is not easy**

Good Data Management

- Data is well organized
- Versioning and updates are controlled
- Analysis is reproducible
- Results are free from artifacts

Data Organization

Good organization depends on: * Data types * Data applications

Data Organization

Data Types Involved in Analysis

- Transactions
- Spatial Data
- Assets/Objects
- Personal Data
- Events
- Organizational Data
- Temporal Data
- Files and Reports
- Relationships

Data Organization

Data Applications

- One-off Analysis
- Streaming data event detection
- Annual Reports
- Data Quality Assurance
- Machine Learning Applications
- Proprietary/Sensitive Data Management

Data Management Concerns

- Duplication
- Security
- Incomplete records
- Parallel Transactions

Analytic Transactions

Note: Some examples of an analytic transaction include the analysis of log files, but we can use banking as a good (and common example). Suggestion engines, that use streaming data, but then modify other tables.

ACID as a Central Concept

- Haerder and Reuter (1983) http://bit.ly/2C3lTLh

Data Transactions Should Be

Atomic, **C**onsistent, **I**solated, **D**urable

A Transaction

- Banking transaction

Atomicity

Atomicity

- The "transaction" is indivisible. It is the fundamental unit of operation.
- There can be multiple operations within a transaction.
- Any operation within a transaction will only succeed if all other operations succeed.

Consistency

Consistency

- The "transaction" can only return results that are "legal" in the context of the database.
- Data types must be preserved (no characters in integer fields).
- Data relationships must be preserved.

Isolation

Isolation

- When multiple transactions cannot affect one another.
- Concurrency is supported through isolation

Durability

Durability

- Results of a transaction must persist.
- Results can only be lost through subsequent transactions (e.g., DELETE)

Database Solutions

What is a Database?

- Most times Database Management System (DBMS)
- Postgres, MySQL, dbLite, Oracle
- Manages databases

What is a Database

- Databases store data in a structured format for storage and retreival
- They are self-describing
- DBMS manage the storage & retrieval of data, and interaction with other software systems through program interfaces
- A DBMS (Postgres) may contain multiple databases

What is a DBMS

- A DBMS manages interaction between the file system and the data queries.
- Postgres is a DBMS
- Manages how data is accessed

Key Points

- Why is good data management important?
- What are key data management concerns?
- How does ACID address these?
- What is a Database/DBMS?

Assignment One

- Files as Markdown (plain text - https://www.markdownguide.org/cheat-sheet/)
- Installing PostgreSQL, PGAdmin, R, RStudio
  - R libraries
    - `tidyverse`, `rmarkdown`, `DBI`, `RPostgreSQL`
    - `install.packages("tidyverse")`