Cornell University
Center for Advanced Computing

Relational Databases
Not your Father's Flat Files

Linda Woodard

Why use Flat Files instead of Databases?

- Flat files are familiar and conceptually easy to use

Station Name	Latitude	Longitude	Elevation	County	State	In NE?	Weather Date	Radar
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/15/2008	0.1
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/16/2008	9.9
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/17/2008	2.4
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/15/2008	0
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/16/2008	0
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/17/2008	6.4

- Flat files are portable--readable by most statistical and graphics software
- No additional software layer to master
- No database software currently available on Ranger

Why use Databases instead of Flat Files?

- Extensive self-documentation via lookup tables
- Data integrity checks:
  no row duplication
  enforce allowable data ranges
  a piece of data appears once; easy to correct errors
- Easy to share portions of data without extensive programming
- Easy to use in a Web environment

What is a Relational Database?

- Popular misconception is a flat file dropped into a table

Station Name	Latitude	Longitude	Elevation	County	State	In NE?	Weather Date	Radar
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/15/2008	0.1
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/16/2008	9.9
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/17/2008	2.4
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/15/2008	0
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/16/2008	0
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/17/2008	6.4

What is a Relational Database?

- Popular misconception is a flat file dropped into a table

Station Name	Latitude	Longitude	Elevation	County	State	In NE?	Weather Date	Radar
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/15/2008	0.1
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/16/2008	9.9
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/17/2008	2.4
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/15/2008	0
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/16/2008	0
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/17/2008	6.4

- Set of tables that “relate” to one another

StationID	StationName	GeographicID	Latitude	Longitude	Elevation
316	ITHACA CORNELL UNIV	151	42.4674	-76.4399	960
636	HARTFORD	58	44.3718	-70.3171	700

stationID	weatherdate	radar
316	2008-10-15 00:00:00	0.10
316	2008-10-16 00:00:00	9.90
316	2008-10-17 00:00:00	2.40
636	2008-10-17 00:00:00	6.40
636	2008-10-16 00:00:00	0.40
636	2008-10-15 00:00:00	0.00

GeographicID	County	State	InNE
151	Tompkins	NY	1
58	Oxford	ME	1

How do you choose between Flat Files & Databases?

- Flat files are a no brainer:
  Small number of rows and/or columns
  6 rows x 9 columns

- Databases are a no brainer:
  “Large” number of rows or columns, or files or hierarchical data
  860 weather stations x 1388 days and counting

- Murky middle ground:
  When does small become large?
  When does data complexity make flat files too cumbersome?
  When does the need to share datasets or parts of datasets become important to your research collaborations?

When does small become large?

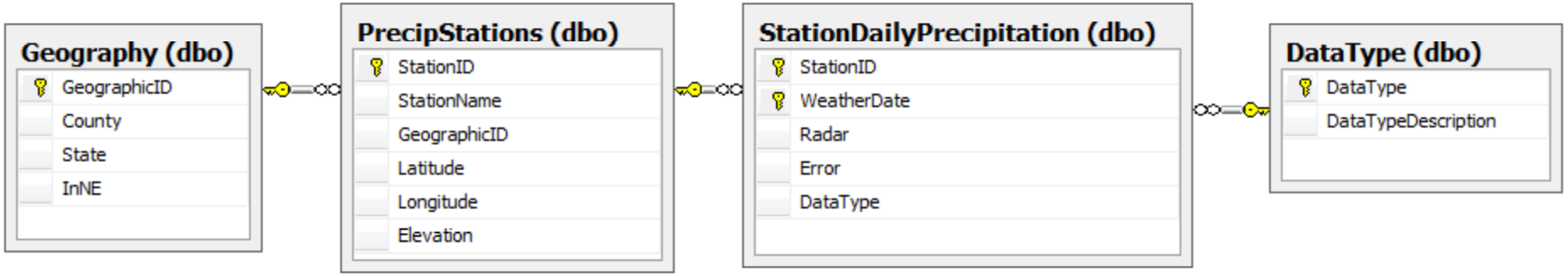
- When data management takes time away from data analysis
- Depends on number of files rather than number of rows or columns 100 files perhaps, 1000 for sure
- Depends on number of people sharing the data; the more people involved in generating and analyzing the data, the more difficult data management and documentation become
- Depends on how homogeneous the data is; even small changes or additions between data files need to be documented

When does data become too complex?

- When large numbers of columns need to be duplicated for each row

Station Name	Latitude	Longitude	Elevation	County	State	In NE?	Weather Date	Radar
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/15/2008	0.1
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/16/2008	9.9
ITHACA	42.47	-76.44	960	Tompkins	NY	1	10/17/2008	2.4
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/15/2008	0
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/16/2008	0
HARTFORD	44.37	-70.32	213	Oxford	ME	1	10/17/2008	6.4

- The hierarchal extent of the data determines when the amount of data that needs to be associated with many rows becomes unwieldy



How do you share data with a wider research audience?

- Science Gateways (http://www.teragrid.org/programs/sci_gateways/)

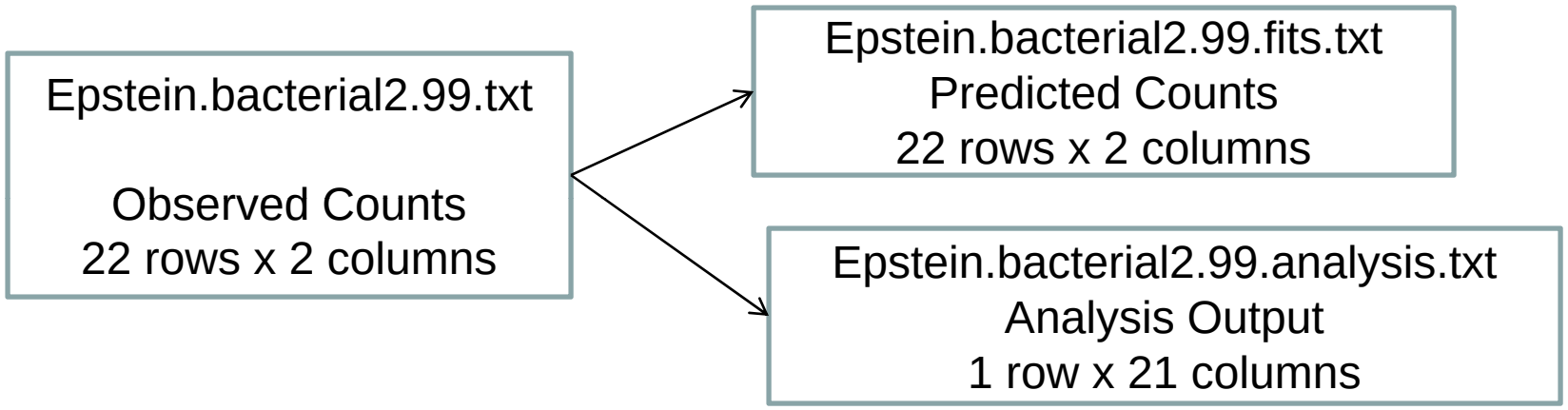
  The CAC database housing high resolution daily temperature and precipitation data for the NE US is accessible via web services on the Science Gateway.

- Web Sites that display data chosen by the user with a database backend

  http://www.cac.cornell.edu/maps/zoom.aspx

  http://sonofstager.tc.cornell.edu/RightWhales/

Case Study – Microbial Diversity



All documentation about the data in the file name

(7 models) x (2 output file types) x (# of unique cutoffs--5 in this case)

1 input file generates 70 output files

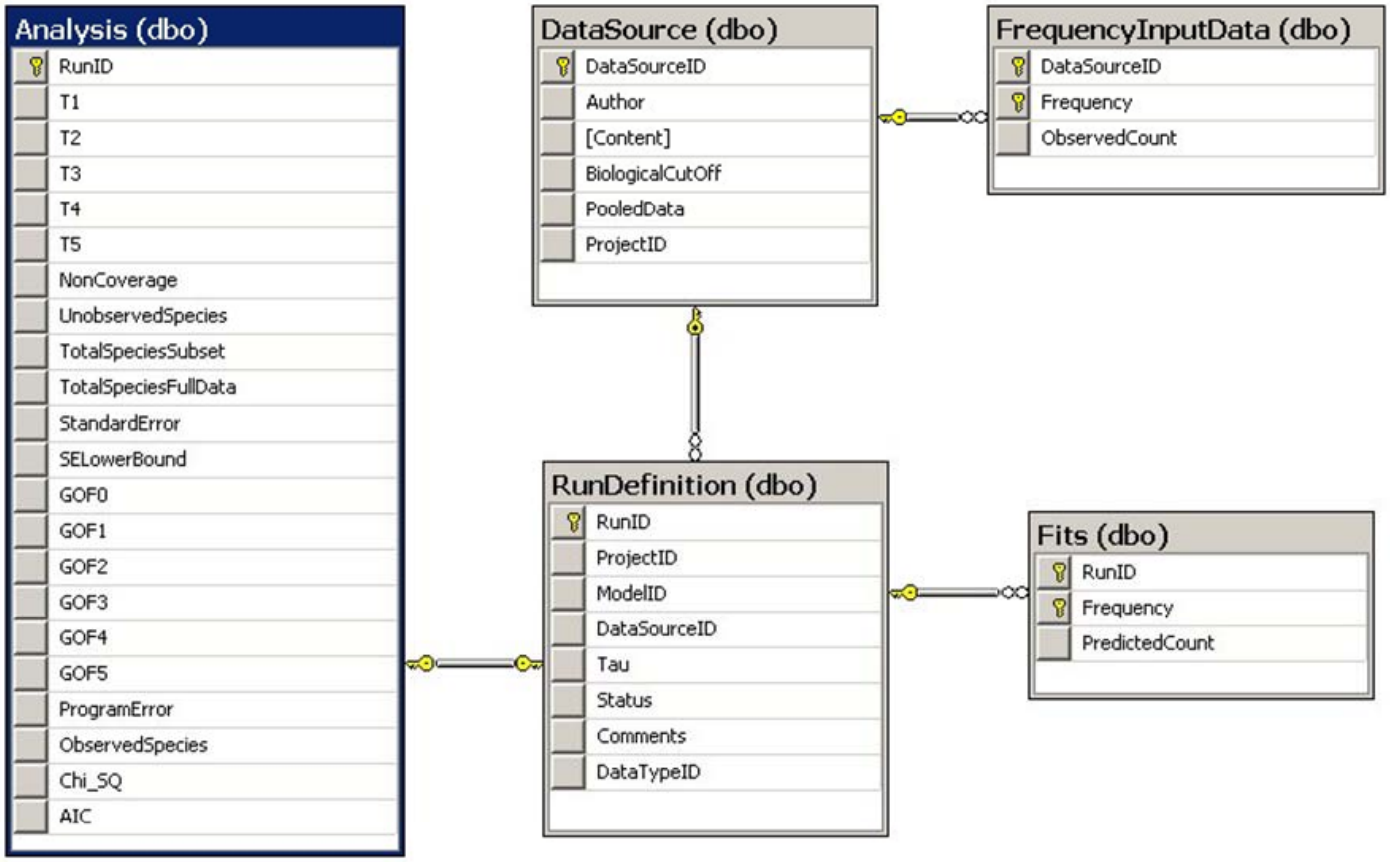
Case Study – Microbial Diversity

- Fast forward 2 years:
  11 projects
  1078 input files
  1,931,776 output files

- Data that took months to analyze now takes days

"This multifaceted, comparative statistical approach has caused a major re-evaluation of statistical methodology within microbiology, where such analyses are crucial. This work was first reported in the Proceedings of the National Academy of Sciences and led to the keynote presentation at an international conference last summer. A complete meta-analysis of the entire GenBank database was published in a special issue of Biometrical Journal.

Case Study – Microbial Diversity



Summary

- Flat files are good when the amount of data is “small” and simple

- Databases should be considered
  when data management is getting in the way of data analysis
  when sharing data with collaborators is important
  when display of data on the Web is integral to your research

- Good database design is critical
  a badly designed database is worse than no database