Databases in Seismology

Seismology dataset overview

Robert Casey, IRIS Data Management Center

Purpose

- Cover relevant areas pertaining to seismic data.
- Lead in to discussing the PDCC schema.
- MySQL hands-on will follow.

Main Themes of Seismic Data Storage

- Station information
- Waveform data recordings
- Event Hypocenters
- Logs, comments, data problem reports

Station information

- Station name and location
- Channels available at that station
- Instrument makeup of each channel
- Response coefficients
- Site visits, log entries, etc.

Metadata

- Such information is typically referred to as metadata.
- Metadata = data about data
- Specifically, we are taking about data of our source of data, the recording instruments.

Fairly static information

- Unlike continuous waveform data
- Station entries grow very little over time
- Growth of station entries come about through a change of state of a station installation.

State change

- Repositioning of instruments (lat/lon, depth/altitude, dip/azimuth)
- Calibration changes
- Instrument changes

Effective time

- …The beginning time of a new state for a station up until the beginning of the next state change
- Start time and end time
- Final end time is a very large future value when it is still in operation

Effective time

- Both stations and channels have effective time entries
- Stations are geographic locations of the general instrument siting
- Channels specifically indicate the locations and state of the sensing instruments

Channels within Stations

- Typically, changes to channel only cause a change of state for the channel entry
- Station entries need only encapsulate all the channel changes in its effective time window
- …Until the moment that the properties of the entire station changes!

Adding channels

- When adding new channels to a station, this could be perceived as a change of effective time for the station itself
- The number of channels for the station have changed, so the channel count for the station entry gets updated

Example

- Station ABC 2001,194,01:50:00 2003,201,14:34:18
- Station ABC 2003,201,14:34:18 2599,365,23:59:59
- No time gap between states

Time Gap

- We could have a time gap introduced between effective times if the station were down and not recording data for a length of time
- Station ABC 2001,194,01:50:00 2003,201,14:34:18
- Station ABC 2003,201,15:01:56 (27:36) 2599,365,23:59:59

Channel encapsulation

- The important thing is that the channel times are always within a particular station effective time.

Example

- Station ABC 2003,201,15:01:56 2599,365,23:59:59
- Channel BHE 2003,201,15:01:56 2004,050,12:22:03
- Channel BHE 2004,050,12:22:03 2599,365,23:59:59

Example of an Incorrect Entry

- Station ABC 2003,201,15:01:56 2599,365,23:59:59
- Channel BHE 2003,201,14:34:18 2004,050,12:22:03
- Channel BHE 2004,050,12:22:03 2599,365,23:59:59

Channels are about sensor systems

- Specific sampling frequency
- Specific gain
- Specific orientation
- Instrument response
- Not all are seismometers!

Response values

- A channel represents a cascade of systems
- Each system has its own transfer function that contributes to the overall response of the channel

Response values

- Each of these system elements is referred to as a stage.
- The first stage is typically the sensor itself, followed by filters, and finally the recording instrument.

Multiple stages per channel

- As a result, when storing channel information, there is typically more than one record of stage response information referring to that channel.
- Stage responses are tied directly to the channel effective time
- They do not have their own effective time

Overview

- Station ABC Eff Time 1
  - Channel BHE Eff Time 1
    - Response stage 1
    - Response stage 2
    - …
  - Channel BHN Eff Time 1
  - Channel BHZ Eff Time 1

Overview

- Station ABC Eff Time 1
  - Channel BHE Eff Time 1
  - Channel BHE Eff Time 2
  - Channel BHE Eff Time 3
- Station ABC Eff Time 2
  - Channel BHE Eff Time 1
  - Channel BHE Eff Time 2

Database tables

- Heirarchy of Station, Channel, and Response
- 3 basic tables in a database for normalization

Responses in a database

- The format for describing stations and channels is nearly always the same.
- Responses can be specified in different ways!
- Therefore, first normal form is difficult to attain with a single response table format

Multiple response tables

- FIR response coefficients
- IIR response coefficients
- Amplitude and Phase specification
- Gain and Sensitivity values
- Decimation

Points to the same channel

- Need different tables for responses, but all should point to the same channel.
- Each response table needs the same foreign key field
  - station/channel/eff-time name pair
  - A single id number

Response ordering

- Need to track the responses and their ordering
  - Stage number in channel
  - Their placement in the sequence of responses within a single stage

Response ordering approach

- Specify the number of stages in the channel table
- Response table lists its stage number as well as a sequence number

Result example

- Channel BHE - id = 200
  - FIR response stage 1 - sequence 1
  - Decimation stage 1 - sequence 2
  - Gain stage 1 - sequence 3
  - IIR response stage 2 - sequence 1
  - Gain stage 2 - sequence 2

Station Metadata Conclusion

- This is the basic technique for storing station metadata
- Some institutions may want more detailed information on instrument specifications, logs, serial numbers, etc.
- The schema used is tailored to the required use of the data and the intent of the dataset.

Waveform data storage

- Waveform data tends to be continuous and growing in size
- Large amounts of data - gigabytes to terabytes
- Data is continuously time-indexed
- Cannot store directly in a database

Make use of a card-catalog concept

- Take cues from your public library
- Books are stored away on shelves
- Not easy to find a specific title or author from the shelves
- Card catalog is a compact reference to find where in the library the book is

We apply this in the database

- Waveform data is written to files in some archival format - or the original format
- Sometimes files are on a RAID, sometimes on a tape system
- We track where this data is in the database and how to get to it

Waveform table in the database

- Reference the station, channel, and timestamp of the data
- Specifics of the station are left to the metadata tables!
- Also, where the file is located, where in the file the data is, and how large

How many entries in database?

- For broadband data, we might have time references every couple of minutes
- For hours and days and years of data, this can be a lot of db records!
- Waveform data tends to be continuous between time references

Continuous data

- For time-continuous data, the end time of one data record is nearly or exactly equal to the start time of the next
- Therefore, it is redundant to create a new database entry for each data record
- We can treat the group of records as a single continuous stream

Data Trace

- We refer to this time-continuous set of records as a data trace.
- The data trace starts with the time index of the first record and ends on either a predefined boundary (a day) or when the data stream is broken
- Sometimes, the end time must be calculated

Calculating the end time

- The simple way to calculate the end time is to find the total number of samples in the data trace and divide by the sampling frequency

  #samples / freq = number of seconds

- Then add to the start time

Point of contention

- There are actually two schools of thought regarding end time calculation
- The end time of one record should be equal to the start time of the next if continuous, OR
- …the end time should represent one sample period before the start of the next record

Alternative end time calculation

- (#samples - 1) / frequency + start time
- In this way, the end time of one record does not equal the start time of the next
- Difference is approximately (1 / freq) in seconds, or a single sample period

Either way is fine

- Though there are good arguments over which technique may be ‘better’, the decision is really a matter of preference of the network data center
- Important: do not assume that your data users know which technique you use -- tell them!

Storing waveforms

- Waveform data is read in
- Waveform data is scanned and analyzed
- Write card-catalog index to database
- Write waveform to a disk or tape file for storage

Reading waveforms

- Data is requested for stations, channels, and a time window
- A database catalog makes it very easy to look up what is available
- You can perform sorting and filtering using the database before you extract the data files!

Reading waveforms

- Need routines to read from the data files
- Database catalog will indicate the source waveform file
  - Byte offset within the file
  - Number of bytes to read
  - This represents the data trace
- Additional filtering can be performed by other routines after extraction of the data stream

Overview

- We have information about our stations
- We have information about the data we have collected
- Remember, we can join these tables, data and metadata, to show many complex representations to users

Event information

- Independent of sensing stations and waveform data
- Refers to actual physical phenomena detected somewhere on the Earth
- The location, magnitude, depth, and time of the event is referred to as a hypocenter

Information details may vary

- For a single earthquake event, we can get hypocenter reports from many sources
- Catalogs come in over time, and the results generally are different with each
- Magnitude intensity and type vary based on the contributor and extent of analysis

Details, details

- Some data centers may just want the basic information
- Others may want to add phase picking analysis and moment descriptions
- May also want local witness reports of damage

All depends on your goals

- What you decide to include for event information in a database is determined by your mission goals
- The minimum data is usually a magnitude value, magnitude type, lat/lon, depth, and time of onset

Finding stations

- With hypocenter information, you can join to the station and channel tables to determine which sensing stations were a certain distance and azimuth from the event

Getting waveforms

- Knowing the event time and the station distance
  - Make use of travel time tables, such as IASPEI
  - Estimate time delay for wave arrival at the station
  - Request waveform from that station using a travel-time shifted time window

Many catalogs

- If you choose to store many different hypocenter catalogs, it can be difficult to determine which one to use
- This is a matter of preference
- Typically, choose a catalog that is typically slower to publish, but more thorough in its analysis

Preferred hypocenter

- This may be referred to as the ‘preferred’ hypocenter selection
- Be sure to display the catalog source so others know how you arrived at the values you display

Other seismic data

- Networks that maintain instruments may need separate tables to indicate site maintenance
  - Site visits
  - Calibration details
  - Logs
  - Repair work

Link back to station and channel

- Typically these maintenance-oriented tables are for internal purposes only
- Still, these tables should be foreign-key linked to the station and/or channel metadata tables for later reference

End of Presentation