Data matching, the fast way

Finlay Thompson

finlay@dragonfly.co.nz

Wellington Perlmongers
12 June 2012

DRAGONFLY

What is data matching?

Linking two different sources of data together

What is data matching?

Linking two different sources of data together

For example:

- Linking addresses together
- Linking observer reported fishing to fisher reported fishing

Why is it hard ?

Problem has order n^2. It grows fast as the number of records increases.

Why is it hard ?

Problem has order n^2. It grows fast as the number of records increases.

Typically its not too difficult to link manually a single record. But this is slow.

Why is it hard ?

Problem has order n^2. It grows fast as the number of records increases.

Typically its not too difficult to link manually a single record. But this is slow.

Context matters. Linking sets of records together is often necessary.

Strategies

Iterate through records:

The obvious approach is to work through the records one at a time.

Initial algorithm used this, but typically takes a long time.

Indexes help here, but still takes a while just to work through each record.

Strategies

Iterate through the match types:

Start with a list of match criteria.

Iterate through the criteria, and match everything you can at each level.

Problems come from the ordering of the criteria. Also each query can take a while, essentially passing over the data multiple times.

The right way!

Combination of both:

Break the data into sets of similar records.

For each one set, iterate through the match criteria.

Parallelise the processing.

Details

Use a database, for example, Postgres

Match criteria are nicely expressed in SQL, and can take into consideration context.

Each query can be optimised, in particular, ensuring that enough, and the right indexes are available.

Each match criteria is separated into its own file.

Details

Use a temporary table each set of data.

The data being matched can be separated into sets. The main algorithm starts by putting selecting the relevant data into a temporary table.

In my case this was an observer trip. It should be a coherent unit of data. It should be small enough that indexes are not appropriate.

The temporary table is only visible in one transaction. Providing an opportunity to parallelise the process.

Details

Parallelise the process on another machine.

The database machine needs to be dedicated to database, no context switching.

Multiple threads can run, each taking tasks/sets from a queue and running them in parallel.

Make sure the database can get everything in memory.