Query optimization
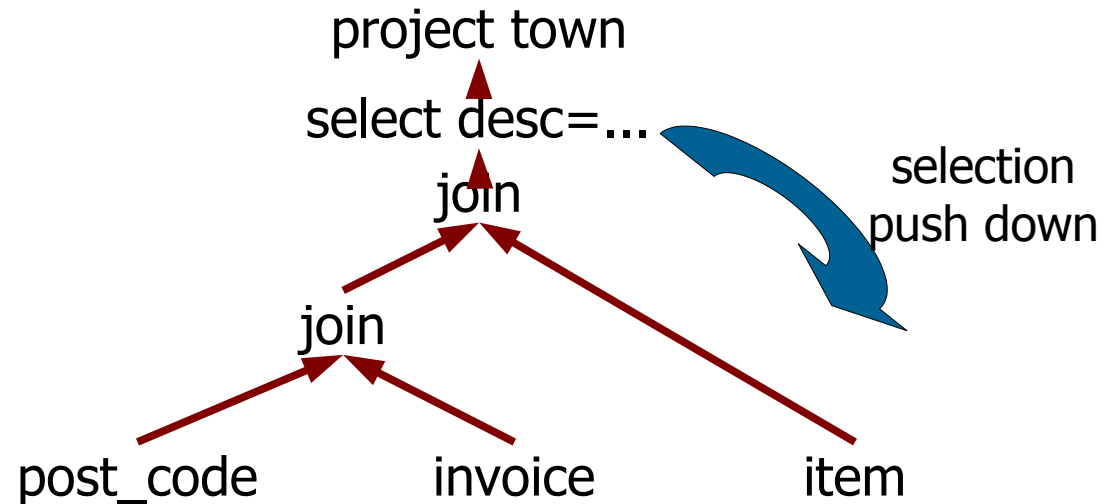
- Two independent logical stages:
    - Generate multiple plans
    - Select the least costly plan
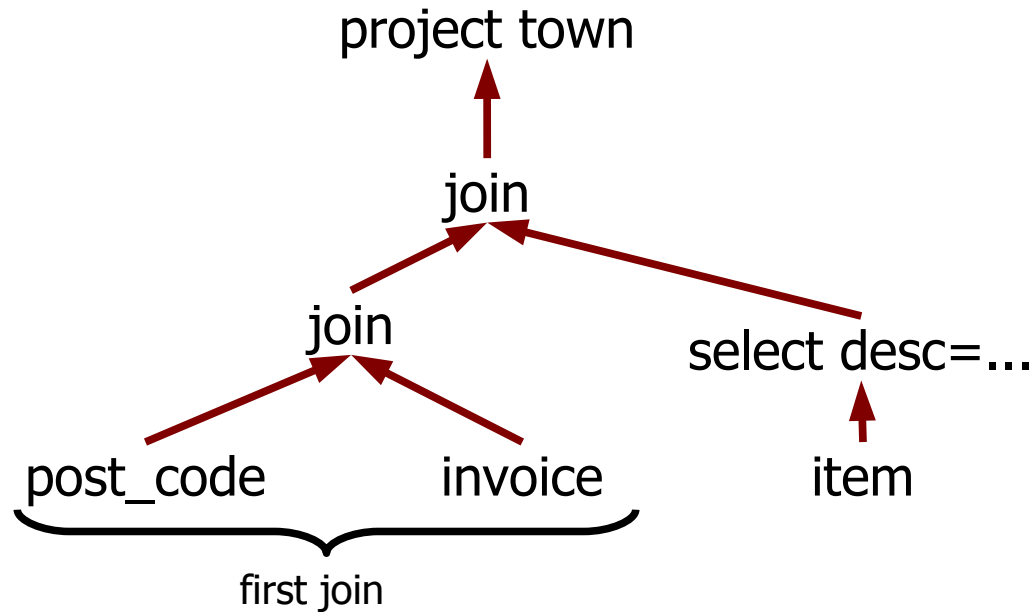
Database Administration

Example

select town from
    post_code natural join invoice
    natural join item where desc=...

project town

select desc=...

join

selection
push down

join

post_code   invoice   item

Example

select town from
post_code natural join invoice
natural join item where desc=...

project town

↑

join

join   select desc=...

post_code   invoice   item

first join

Example

select town from
post_code natural join invoice
natural join item where desc=...
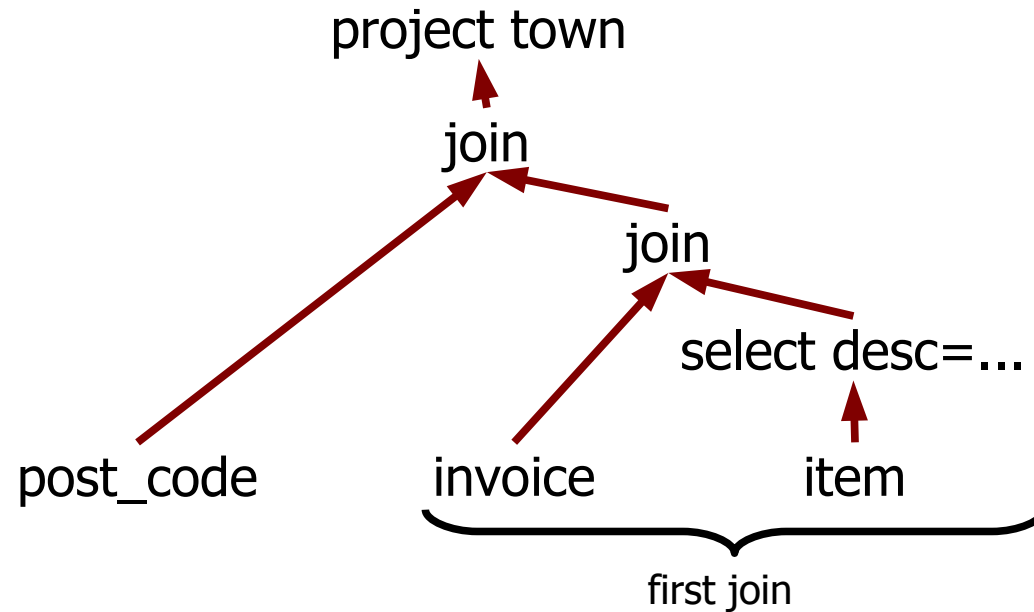
Alternative plans

- Reasons for alternative plans:
  - Join order (permutations)
  - Indexes
 - Alternative physical operators

Plan selection

- Predict resource usage:
  - CPU =~ <u>#records</u> processed
  - Disk I/O =~ #blocks visited
    - =~ <u>#records</u> used (with index)
    - =~ <u>#records</u> in relation (with scan)

- Memory is limiting when physical operator materializes results:
  - =~ <u>#records</u>

#Records?

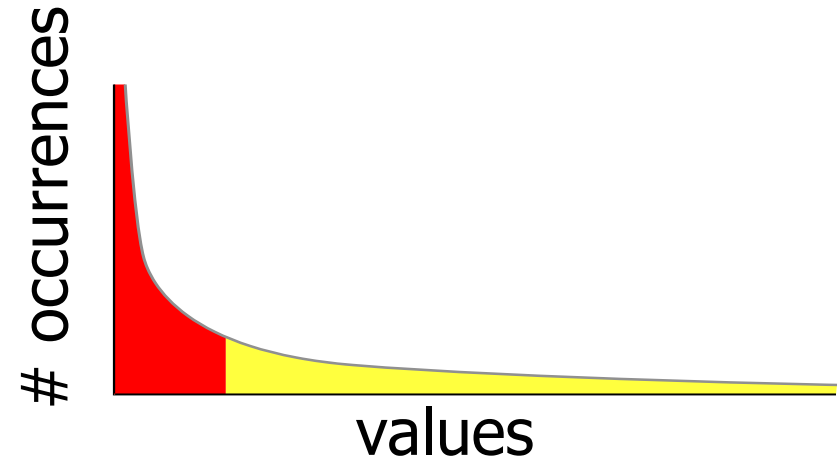**Relative weights?**

Estimating #records

- Easy for stored relations

- What about intermediate results?

project town

join

join

select desc=...

post_code   invoice   item
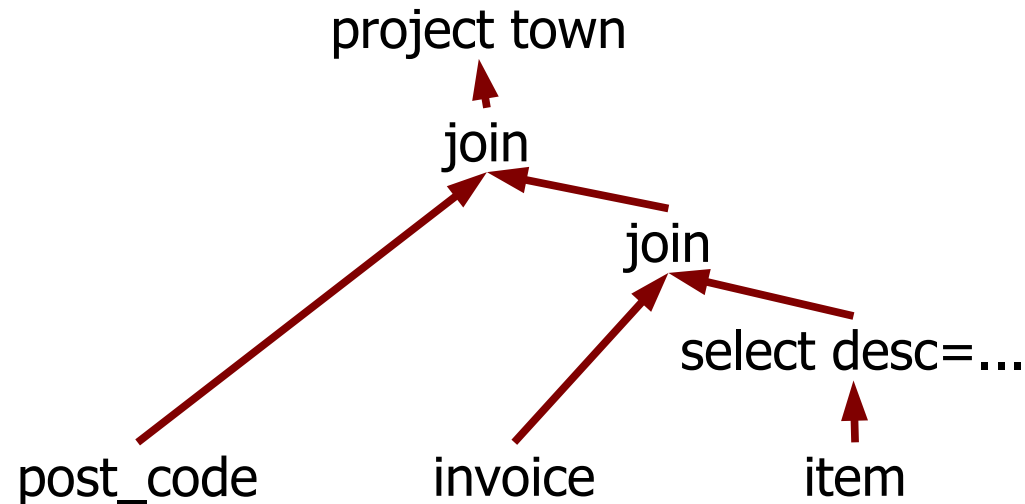
Estimating #records - Selection

- Know <u>#distinct</u>

  - Expected copies of each tuple:

    - #records / #distinct

- Real data are usually not evenly distributed:

  - 80/20 rule

  - Power law

Image source: Wikipedia

Estimating #records - Selection

- Know <u>most popular</u> and <u>#occurrences</u>

- Compute estimated #occurrences of others as uniformly distributed

Estimating #records - Selection

- Know <u>histogram</u>:

    - % of occurrences in each interval

        or

    - interval for fixed % of occurrences

- Compute #occurrences as uniformly distributed within each interval

Estimating #records - Join

- Product:
  - Multiply #records

- Join:
  - Select after product
  - For equi-join:
    - Predict matching records

Statistics

- Required statistics:
  - # distinct / # records
  - most popular
  - histogram
  - correlation

- Computing statistics:
  - Full computation is costly
  - Sampling provides good enough results

Conclusions

- Make sure that the best option is available:

  - Indexes and materialized views

  - Sufficient memory for all operators

- Make sure the best option is selected:

  - Tune relative weights

  - Provide current and sufficiently detailed statistics