Implementation of Relational Operations (Part 2)

R&G - Chapters 12 and 14

An Alternative to Sorting: Hashing!

- **Idea:**
  - Many of the things we use sort for don't exploit the *order* of the sorted data
  - e.g.: removing duplicates in DISTINCT
  - e.g.: finding matches in JOIN
- **Often good enough to match all tuples with equal values**
- **Hashing does this!**
  - And may be cheaper than sorting! (Hmmm...!)
  - But how to do it for data sets bigger than memory??

General Idea

- **Two phases:**
  - Partition: use a hash function *h* to split tuples into partitions on disk.
    - Key property: all matches live in the same partition.
  - ReHash: for each partition on disk, build a main-memory hash table using a hash function *h2*