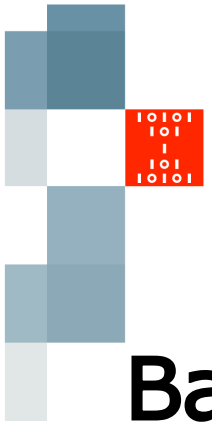

netarchive.dk

Batching in the bitarchive

Lars Clausen
Netarkivet

Batch overview

- Runs on one bitarchive
- Runs on all machines on matching files
- Returns concatenated strings
- Per-file or per-record
- Exceptions & failed files are collected
- #processed is counted
- Code must exist on bitarchives

Anatomy of a batch job

- `FileBatchJob`
  - `initialize(OutputStream out)`
  - `boolean processFile(file, out)`
  - `finish(out)`
- `ARCBatchJob extends FileBatchJob`
  - `boolean processRecord(record, out)`

Filters

- Allows skipping files in two ways:
  - processOnlyFilesNamed(fileNames)
    - Picks single files
  - processOnlyFilesMatching(String regexp)
    - Tries to match all files
- ARCBatchJob can also filter records:
  - override getFilter() method
    - Has a few default implementations

Future developments

- Lock down bitarchives
- Make code moveable
- Make more ways to combine results
- Turn into map-reduce?