bin/nutch fetch

Fetch is an alias for org.apache.nutch.fetcher.Fetcher

This fetcher uses a well-known model of one producer (a QueueFeeder) and many consumers (FetcherThread-s).

QueueFeeder reads input fetchlists and populates a set of FetchItemQueue-s, which hold FetchItem-s that describe the items to be fetched. There are as many queues as there are unique hosts, but at any given time the total number of fetch items in all queues is less than a fixed number (currently set to a multiple of the number of threads).

As items are consumed from the queues, the QueueFeeder continues to add new input items, so that their total count stays fixed (FetcherThread-s may also add new items to the queues e.g. as a results of redirection) - until all input items are exhausted, at which point the number of items in the queues begins to decrease. When this number reaches 0 fetcher will finish.

This fetcher implementation handles per-host blocking itself, instead of delegating this work to protocol-specific plugins. Each per-host queue handles its own "politeness" settings, such as the maximum number of concurrent requests and crawl delay between consecutive requests - and also a list of requests in progress, and the time the last request was finished. As FetcherThread-s ask for new items to be fetched, queues may return eligible items or null if for "politeness" reasons this host's queue is not yet ready.

If there are still unfetched items in the queues, but none of the items are ready, FetcherThread-s will spin-wait until either some items become available, or a timeout is reached (at which point the Fetcher will abort, assuming the task is hung).

Nutch 1.x

```
Usage: bin/nutch fetch [-D...] <segment> [-threads n]
```

****: This is the path to the previously generated segment directory we wish to fetch.

[-threads n]: This argument invokes the number of threads we wish to work concurrently on fetching URLs in the desired segment e.g. the number of fetcher threads the fetcher should use. This is also determines the maximum number of requests that are made at once (each fetcher thread handles one connection).

[-D...]: overwrite a Nutch/Hadoop property from command-line, e.g.

[-Dfetcher.parse=true]: Make fetcher parse documents, overwriting the default value defined in nutch-default.xml or the setting in nutch-site.xml.

Nutch 2.x

```
Usage: FetcherJob (<batchId> | -all) [-crawlId <id>] [-threads N] [-resume] [-numTasks N]
       <batchId>     - crawl identifier returned by Generator, or -all for all
                     generated batchId-s
       -crawlId <id> - the id to prefix the schemas to operate on,
                     (default: storage.crawl.id)
       -threads N    - number of fetching threads per task
       -resume       - resume interrupted job
       -numTasks N   - if N > 0 then use this many reduce tasks for fetching
                     (default: mapred.map.tasks)
```

CommandLineOptions