

# bin/nutch mergedb

Mergedb is an alias for `org.apache.nutch.crawl.CrawlDbMerger`

This tool merges several crawl-db's into one, optionally filtering URLs through the current URLFilters, to skip prohibited pages. It is possible to use this tool just for filtering - in that case only one crawl-db should be specified in arguments. If more than one crawl-db contains information about the same URL, only the most recent version is retained, as determined by the value of `org.apache.nutch.crawl.CrawlDatum#getFetchTime()`. However, all metadata information from all versions is accumulated, with newer values taking precedence over older values.

Usage:

```
bin/nutch mergedb <output_crawl-db> <crawl-db1> [<crawl-db2> <crawl-db3> ...] [-normalize] [-filter]
```

**<output\_crawl-db>**: This allows us to specify a name for the new merged output crawl-db.

**<crawl-db1>**: Only one crawl-db parameter is used if we only wish to filter URLs through the current URLFilters. This enables us to filter unwanted pages from the crawl-db.

**[<crawl-db2> <crawl-db3> ...]**: Two or more crawl-db arguments can be passed if we wish to undertake a merging of crawl-db's. More information regarding duplication of URLs and URL metadata etc can be found above.

**[-normalize]**: If we know/think that URLs require to be normalized prior to being merged we pass this parameter. This uses the URLNormalizer on urls in the crawl-db(s), although this is usually not required.

**[-filter]**: Enables is to filter URLs through the current URLFilters. This can be used in conjunction with a single crawl-db argument.

[CommandLineOptions](#)