SitemapFeature

Sitemap processing in Nutch

What is Sitemap ?

Sitemaps are used by webmasters to tell search engines about the pages present on their website which are available for crawling. Sitemap generally also carry some additional metadata about those pages which helps a crawler to crawl the page effectively. This metadata includes:

- When was the page last updated ?
- How often it usually changes ?
- Priority of the page relative to other URLs in the site

For more information on Sitemaps, see the official page of Sitemap protocol

Steps to run

For Nutch 1.x:

```
bin/nutch sitemap <crawldb> [-hostdb <hostdb>] [-sitemapUrls <sitemapUrls>] [-threads <threads>] [-force] [-noFilter] [-noNormalize]
```

<crawldb> path to crawldb where the sitemap urls would be injected

Atleast one of these two must be provided:

- **-hostdb <hostdb>** path of a hostdb. Sitemap(s) from these hosts would be downloaded
- **-sitemapUrls <sitemapUrls>** path to sitemap urls directory

-threads <threads> Number of threads created per mapper to fetch sitemap urls

-force force update even if CrawlDb appears to be locked (CAUTION advised)

-noFilter Turn off URLFilters on urls (optional)

-noNormalize Turn off URLNormalizer on urls (optional)

For Nutch 2.x:

Please follow here.

How Nutch processes Sitemap ?

There are two use cases supported in Nutch's Sitemap processing:

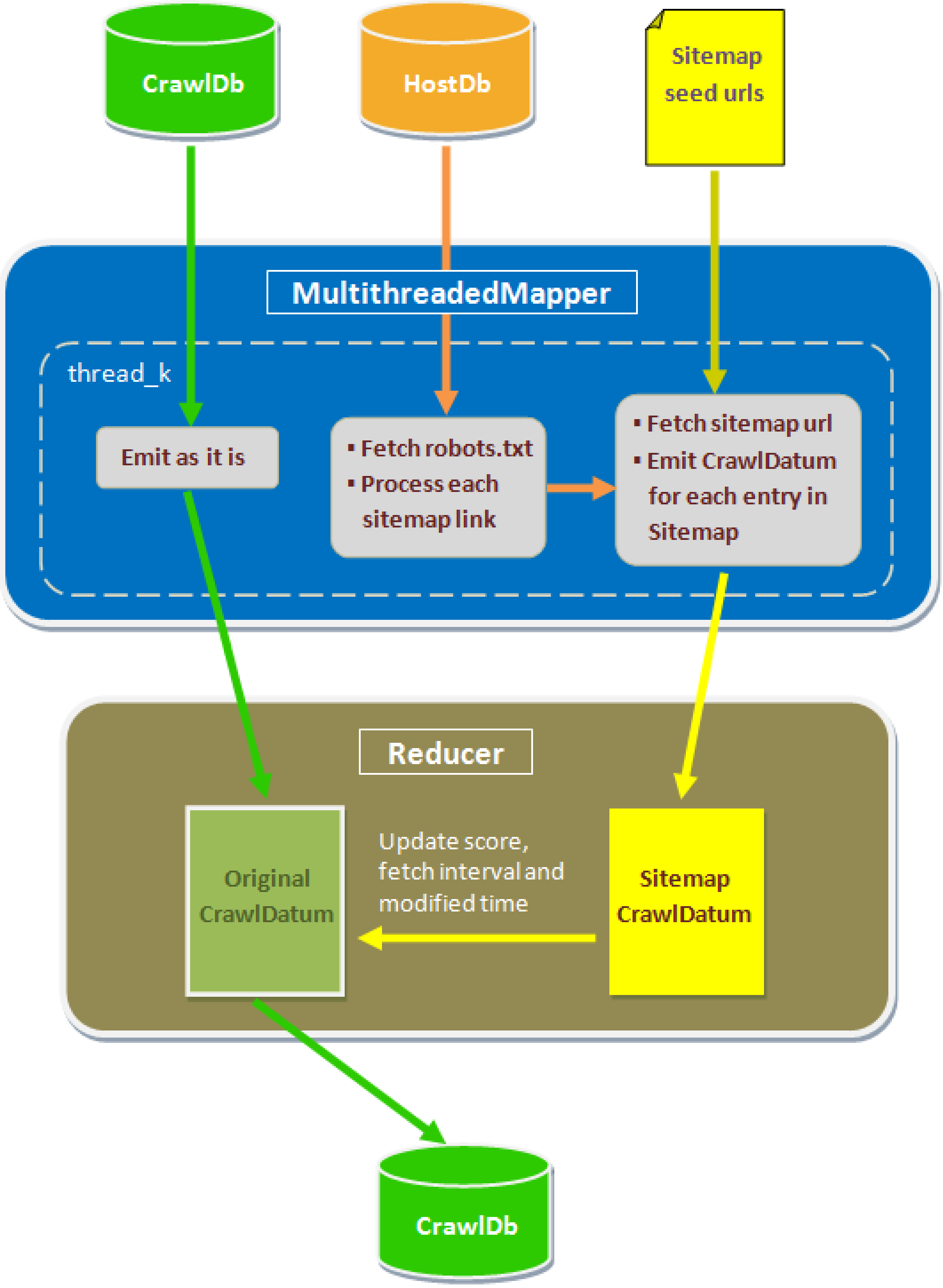
1. Sitemaps are considered as "remote seed lists". Crawl administrators can prepare a list of sitemap links and get only those sitemap pages. This suits well for targeted crawl of specific hosts. 2. For open web crawl, it is not possible to track each host and get the sitemap links manually. Nutch would automatically get the sitemaps for all the hosts seen in the crawls and inject the urls from sitemap to the crawldb.

For #1, the sitemap urls are directly fetched. Nutch uses Crawler Commons Project for parsing sitemaps. CrawlDatum objects are created for the urls extracted from sitemap along with their metadata.

For #2, we need a list of all hosts see throughout the duration of nutch crawl. Nutch's HostDb stores all the hosts that were seen in the long crawl. Link to the robots.txt of these hosts is generated by pre-pending "http://" or "https://" schemes to the hostname. Crawler Commons is used for robots.txt parsing and thus get the sitemap links. These sitemap links are then processed same as #1.

Fetching sitemaps over the web is a I/O bound activity. A MultithreadedMapper is used to parallely fetch multiple sitemaps and get better throughout. At the reducer side, if an existing record was present in the crawldb for the url found in on of the sitemaps, merging is done by copying the metadata to the existing crawldb record.

This figure describes the map-reduce job for Sitemap processing:



Information on the development of this feature is here. This is a new feature so there would be bugs and some corner cases when it would not work as expected. If you face any of those, we would be happy to discuss those over the user mailing list and get them corrected.