Item 290

git_comments:

git_commits:

1. **summary:** Merge pull request #273 from smartive/NUTCH-2493 **message:** Merge pull request #273 from smartive/NUTCH-2493 NUTCH-2493: Add new parameter to crawler script to configure sitemap processing frequency

github_issues:

github_issues_comments:

github_pulls:

- 1. **title:** NUTCH-2493: Add new parameter to crawler script to configure sitemap processing frequency **body:** While using the crawler script with the sitemap processing feature introduced in #271 I encountered some performance issues when working with large sitemaps. Therefore one should be able to specify if sitemap processing based on HostDB should take place and if so how frequently it should be done
- 2. title: NUTCH-2493: Add new parameter to crawler script to configure sitemap processing frequency body: While using the crawler script with the sitemap processing feature introduced in #271 I encountered some performance issues when working with large sitemaps. Therefore one should be able to specify if sitemap processing based on HostDB should take place and if so how frequently it should be done.
- 3. **title:** NUTCH-2493: Add new parameter to crawler script to configure sitemap processing frequency **body:** While using the crawler script with the sitemap processing feature introduced in #271 I encountered some performance issues when working with large sitemaps. Therefore one should be able to specify if sitemap processing based on HostDB should take place and if so how frequently it should be done.

label: code-design

4. **title:** NUTCH-2493: Add new parameter to crawler script to configure sitemap processing frequency **body:** While using the crawler script with the sitemap processing feature introduced in #271 I encountered some performance issues when working with large sitemaps. Therefore one should be able to specify if sitemap processing based on HostDB should take place and if so how frequently it should be done.

github_pulls_comments:

- 1. I am +1 here, tested on local test crawl.
- 2. +1, patch looks good

github_pulls_reviews:

jira_issues:

- summary: Add configuration parameter for sitemap processing to crawler script
 description: While using the crawler script with the sitemap processing feature introduced in NUTCH2491 I encountered some performance issues when working with large sitemaps. Therefore one should be
 able to specify if sitemap processing based on HostDB should take place and if so how frequently it
 should be done.
- 2. summary: Add configuration parameter for sitemap processing to crawler script description: While using the crawler script with the sitemap processing feature introduced in NUTCH-2491 I encountered some performance issues when working with large sitemaps. Therefore one should be able to specify if sitemap processing based on HostDB should take place and if so how frequently it should be done.
- 3. **summary:** Add configuration parameter for sitemap processing to crawler script **description:** While using the crawler script with the sitemap processing feature introduced in NUTCH-2491 I encountered some performance issues when working with large sitemaps. Therefore one should be

able to specify if sitemap processing based on HostDB should take place and if so how frequently it should be done.

label: code-design

4. **summary:** Add configuration parameter for sitemap processing to crawler script **description:** While using the crawler script with the sitemap processing feature introduced in NUTCH-2491 I encountered some performance issues when working with large sitemaps. Therefore one should be able to specify if sitemap processing based on HostDB should take place and if so how frequently it should be done.

jira_issues_comments:

- 2. lewismc commented on issue #273: NUTCH-2493: Add new parameter to crawler script to configure sitemap processing frequency URL: https://github.com/apache/nutch/pull/273#issuecomment-356305137 I am +1 here, tested on local test crawl. ------- This is an automated message from the Apache Git Service. To respond to the message, please log on GitHub and use the URL above to go to the specific comment. For queries about this service, please contact Infrastructure at: users@infra.apache.org
- 4. lewismc closed pull request #273: NUTCH-2493: Add new parameter to crawler script to configure sitemap processing frequency URL: https://github.com/apache/nutch/pull/273 This is a PR merged from a forked repository. As GitHub hides the original diff on merge, it is displayed below for the sake of provenance: As this is a foreign pull request (from a fork), the diff is supplied below (as it won't show otherwise due to GitHub magic): diff --git a/src/bin/crawl b/src/bin/crawl index f34908740..7a32be207 100755 --- a/src/bin/crawl +++ b/src/bin/crawl @@ -63,6 +63,37 @@ function __to_seconds() { echo \$SECONDS \} +function __print_usage \{ + echo "Usage: crawl [options] <crawl_dir > <num_rounds > " + echo -e "" + echo -e "Arguments:" + echo -e " <crawl_dir>\t\t\tDirectory where the crawl/host/link/segments dirs are saved" + echo -e " <num_rounds>\t\t\tThe number of rounds to run this crawl for" + echo -e "" + echo -e "Options:" + echo -e " -i|--index\t\t\t\tIndexes crawl results into a configured indexer" + echo -e " -D\t\t\t\tA Java property to pass to Nutch calls" + echo -e " -w|--wait <NUMBER[SUFFIX]>\t\tTime to wait before generating a new segment when no URLs" + echo -e " \t\t\t\t\tare scheduled for fetching. Suffix can be: s for second," + echo -e " \t\t\t\tm for minute, h for hour and d for day. If no suffix is" + echo -e " \t\t\t\t\specified second is used by default. [default: -1]" + echo e " -s <seed_dir>\t\t\tPath to seeds file(s)" + echo -e " -sm <sitemap_dir>\t\t\tPath to sitemap URL file(s)" + echo -e " --num-slaves <num_slaves>\t\tNumber of slave nodes [default: 1]" + echo -e " \t\t\t\t\Note: This can only be set when running in distribution mode" + echo -e " --num-tasks <num_tasks>\t\tNumber of reducer tasks [default: 2]" + echo -e " --size-fetchlist <size_fetchlist>\tNumber of URLs to fetch in one iteration [default: 50000]" + echo -e " --time-limitfetch <time_limit_fetch>\tNumber of minutes allocated to the fetching [default: 180]" + echo -e " --numthreads <num_threads>\t\tNumber of threads for fetching / sitemap processing [default: 50]" + echo -e " --sitemaps-from-hostdb <frequency>\tWhether and how often to process sitemaps based on HostDB." + echo -e " \t\t\t\tSupported values are:" + echo -e " \t\t\t\t - never [default]" + echo -e " \t\t\t\t\t - always (processing takes place in every iteration)" + echo -e " \t\t\t\t\- once (processing only takes place in the first iteration)" + + exit 1 + } + # default values INDEXFLAG=false JAVA_PROPERTIES="" @@ -73,6" +104,7 @@ NUM_TASKS=2 # 2 x NUM_SLAVES SIZE_FETCHLIST=50000 # 25K x NUM_TASKS TIME LIMIT FETCH=180 NUM THREADS=50 +SITEMAPS FROM HOSTDB FREQUENCY=never while [[\$# > 0]] do @@ -117,36 +149,24 @@

```
do NUM_THREADS="${2}" shift 2 ;; + --sitemaps-from-hostdb) +
SITEMAPS_FROM_HOSTDB_FREQUENCY="${2}" + shift 2 + ;; *) break ;; esac done +if [[!
"$SITEMAPS_FROM_HOSTDB_FREQUENCY" =~ ^(never|always|once)$ ]]; then + echo "Error: --
sitemaps-from-hostdb <frequency> has to be one of never, always, once." + echo -e "" + __print_usage
+fi + if [[ $# != 2 ]]; then - echo "Usage: crawl [options] <crawl_dir> <num_rounds>" - echo -e "" - echo
-e "Arguments:" - echo -e " <crawl_dir>\t\t\tDirectory where the crawl/host/link/segments dirs are
saved" - echo -e " <num_rounds>\t\t\t\tThe number of rounds to run this crawl for" - echo -e "" - echo -e
"Options:" - echo -e " -i|--index\t\t\tIndexes crawl results into a configured indexer" - echo -e " -
D\t\t\t\tA Java property to pass to Nutch calls" - echo -e " -w|--wait <NUMBER[SUFFIX]>\t\tTime to
wait before generating a new segment when no URLs" - echo -e " \t\t\t\tare scheduled for fetching.
Suffix can be: s for second," - echo -e " \t\t\t\tm for minute, h for hour and d for day. If no suffix is" -
echo -e " \t\t\t\tspecified second is used by default. [default: -1]" - echo -e " -s <seed_dir>\t\t\t\tPath to
seeds file(s)" - echo -e " -sm <sitemap_dir>\t\t\tPath to sitemap URL file(s)" - echo -e " --num-slaves
<num_slaves>\t\tNumber of slave nodes [default: 1]" - echo -e " \t\t\t\t\tNote: This can only be set when
running in distribution mode" - echo -e " --num-tasks <num_tasks>\t\tNumber of reducer tasks [default:
2]" - echo -e " --size-fetchlist <size_fetchlist>\tNumber of URLs to fetch in one iteration [default:
50000]" - echo -e " --time-limit-fetch <time_limit_fetch>\tNumber of minutes allocated to the fetching
[default: 180] "-echo-e"--num-threads < num\_threads > \t \t Number of threads for fetching / sitemap in the latest and the l
processing [default: 50]" - - exit 1 + __print_usage fi CRAWL_PATH="$1" @@ -164,7 +184,7 @@
bin="`cd "$bin"; pwd`" # determines whether mode based on presence of job file mode=local if [ -f
"${bin}"/../*nutch*.job]; then - mode=distributed + mode=distributed fi if [[ "$mode" = "local" ]]; then
NUM_SLAVES=1 @@ -199,6 +219,13 @@ function __bin_nutch { fi } +function __update_hostdb { +
if [[ -d "$CRAWL_PATH"/crawldb ]]; then + echo "Updating HostDB" + __bin_nutch updatehostdb -
crawldb "$CRAWL_PATH"/crawldb -hostdb "$CRAWL_PATH"/hostdb + fi +} + # initial injection if [[!
-z $SEEDDIR ]]; then echo "Injecting seed URLs" @@ -229,16 +256,15 @@ do echo `date` ": Iteration
$a" fi - # create / update HostDB - if [[ -d "$CRAWL_PATH"/crawldb ]]; then - echo "Updating HostDB"
  __bin_nutch updatehostdb -crawldb "$CRAWL_PATH"/crawldb -hostdb "$CRAWL_PATH"/hostdb - fi
+ if [[ "$SITEMAPS_FROM_HOSTDB_FREQUENCY" = "always" ||
("$SITEMAPS_FROM_HOSTDB_FREQUENCY" = "once" && $a -eq 1) ]]; then + # create / update
HostDB on first run + [[ $a -eq 1 ]] && __update_hostdb - # sitemap processing based on HostDB - if [[ -
d "$CRAWL_PATH"/hostdb ]]; then - echo "Processing sitemaps based on hosts in HostDB" -
   bin_nutch sitemap "$CRAWL_PATH"/crawldb -hostdb "$CRAWL_PATH"/hostdb -threads
$NUM_THREADS + # sitemap processing based on HostDB + if [[ -d "$CRAWL_PATH"/hostdb ]]; then
+ echo "Processing sitemaps based on hosts in HostDB" + __bin_nutch sitemap
"$CRAWL PATH"/crawldb -hostdb "$CRAWL PATH"/hostdb -threads $NUM THREADS + fi fi echo
"Generating a new segment" @@ -311,6 +337,8 @@ do echo "Skipping indexing ..." fi +
 commands fall into WebGraph territory # and should be uncommented based on your requirements ---
----- This is an automated message from the Apache Git
Service. To respond to the message, please log on GitHub and use the URL above to go to the specific
comment. For queries about this service, please contact Infrastructure at: users@infra.apache.org
```

- 5. Thank you [~mfeltscher]
- 6. SUCCESS: Integrated in Jenkins build Nutch-trunk #3494 (See [https://builds.apache.org/job/Nutch-trunk/3494/]) NUTCH-2493: Add new parameter to crawler script to configure sitemap (moreno: [https://github.com/apache/nutch/commit/7085b265e7743789af4b9c54f3914c5826c8fe6f]) * (edit) src/bin/crawl
- 7. Bulk close of issues resolved for 1.15