TEST6 Robustness test

Prerequisites

This test requires restart of infrastructure components (database and network) - these steps must be coordinated with the other testers. Only the steps under "Database crash test" and "Network recovery test" below need to be coordinated.
Resubmit jobs after restart, restart of failed jobs, upload of old files at harvester restart, scheduler skips old jobs.

Uses heritrix3 templates default_order_xml

Install and Start System

On devel@kb-prod-udv-001.kb.dk, run the tests as specified under "Release Test Information" on 5.4 Release Test, and with TESTX set to TEST6

Check that the GUI is available and that the System Status does not show any startup problems.

Start a selective harvest

Start an **hourly** selective harvest for the 'netarkivet.dk' domain.

Create a new template:

- Under `DefinitionsEdit Harvest Templates`, download the template "default_orderxml", by choosing "Save to disk" in the pull-down menu, and clicking "Retrieve".
- Edit the template so that max-size-bytes is 5000 in the WARCWriterProcessor. Do this as follows.
  - Find `<bean id="simpleOverrides"` and then under `beanproperty value`, insert:

    ```
    metadata.jobName=default_orderxml_smallwarcs

    metadata.description=Default Profile generating small warc-files (5000 bytes)
    warcWriter.maxFileSizeBytes = 5000
    disposition.maxPerHostBandwidthUsageKbSec=30
    ```
  - Under `<bean id="metadata"`, modify the line to read:
    `<property name="jobName" value="default_orderxml_smallwarcs" />`
- Upload the template with the filename changed to "default_orderxml_smallwarcs.xml"

Modify domain templates

To do the following configures, go to DefinitionsFind Domain(s) and search for the domain-URL in question. The click on the given domain, and you'll find the defaultconfig under Configurations, which you can edit.

- Configure the defaultconfig for kum.dk to use template 'default_orderxml_smallwarcs'.
- Configure the defaultconfig for dbc.dk to use template 'default_orderxml', and max-hops=0
- Configure the defaultconfig for bs.dk to use template 'default_orderxml', and max-hops=10

Make a new snapshot harvest definition with a name you can remember

- Create a new snapshot harvest, with 'Max number of bytes per domain' set to 1,000,000 bytes (1 mbyte). Activate it.
- Go to 'Harvest status'->'All Jobs' in the left menu, and check that there are jobs for the harvest you just created. Check that the status of one of these changes from "New" to "Started", then go to "Systemstate""Overview of the system state", and check that no errors or warnings are present in the system overview.

  Now immediately proceed to the following test ("Stop the Test Automatically During Upload").

Stop the Test Automatically During Upload

1. Using the GUI, go to "Harvest status""All Jobs", and by clicking each Job ID for the snapshot harvest in turn, find the *job ID* for the job in which kum.dk is being harvested.
2. Go back to "Harvest status""All Jobs", and reload the page until the job you just identified has status "Started"... then immediately go to "Harvest status""H3 Remote Access", keep reloading the page until the job ID found above appears, and click the job ID (this may take several tries until it is ready) then immediately pause the job.
3. Go to "Harvest status""H3 Remote Access" and click the job ID you identified, then click "View/Search in cached Crawllog", then "Update cache". Go to "Harvest status""All Running Jobs" and search for "kum.dk" to find the job, then note down the name of the harvest machine (Host) for that job.
4. Download the attached script and modify it to point at the correct harvester and job number
5. Copy the script to `kb-prod-udv-001.kb.dk:/home/devel/` , give it a "`chmod 755`" then run it. It monitors the "warcs" directory and as soon as the first warcfile is uploaded it detects that uploading has started and shuts down the test instance.
6. Go to "Harvest status""H3 Remote Access"*identified Job ID,* and unpause the job (no explicit logout is necessary)
7. Wait for the job to complete, after which the TEST6 instance is stopped, starting with the apps on machine harvesting kum.dk

Save the Metadata Warcfile

- From `kb-prod-udv-001.kb.dk` , log into the harvester where kum.dk was being harvested (with user netarkdv if harvester is in Aarhus, and user `devel` if harvester is in Kbh).
- Find the crawldir in TEST6/harvester_low
- Find the metadata warcfile in the metadata subdirectory and copy it to TEST6/

Create a Fake Crawl Dir

From `kb-prod-udv-001.kb.dk` do:

```
ssh netarkdv@sb-test-har-001.statsbiblioteket.dk
cd TEST6/harvester_high
cp -r ~netarkdv/testdata-h3/TEST6/23-fakejobdir .
mkdir 23-fakejobdir/heritrix3/jobs/23-fakejobdir/logs
touch 23-fakejobdir/heritrix3/jobs/23-fakejobdir/logs/crawl.log
touch
23-fakejobdir/heritrix3/jobs/23-fakejobdir/logs/progress-statistics.log
```

Wait atleast 3 Hours then Restart the System

Wait atleast 3 Hours then Restart the System (by running `stop_test.sh` then `start_test.sh`)

Verify the restarted system. On devel@kb-test-adm-001

1. Check the log for warnings and errors.

```
cd /home/devel/$TESTX/log/
grep ERROR *.log | grep -v COMMON_ERROR
grep WARN *.log
```

When checking for warnings/errors, be sure to ignore any warnings/error that happened *before* the above restart. Also, the following kinds of entries are normal/known, and can be ignored:

```
arcrepositoryapplication0.log.0:WARNING: AdminDataFile (./admin.data)
was not found.

HarvestJobManagerApplication.log:13:12:05.567 WARN
d.n.h.s.jobgen.AbstractJobGenerator.generateJobs - Refusing to
schedule harvest definition 'TEST6-selective-harvest-HOURLY' in the
past. Skipped 71 events. Old nextDate was Fri Apr 13 13:59:19 CEST
2018 new nextDate is Mon Apr 16 13:59:19 CEST 2018

HarvestJobManagerApplication.log:13:12:20.959 WARN
d.n.h.s.HarvestSchedulerMonitorServer.processCrawlStatusMessage - Job
124 failed: HarvestErrors = dk.netarkivet.common.exceptions.IOFailure:
Crawl probably interrupted by shutdown of HarvestController

HarvestJobManagerApplication.log:13:13:17.710 WARN
d.n.h.s.HarvestSchedulerMonitorServer.processCrawlStatusMessage -
Received unexpected CrawlStatusMessage for job 23 with new status
FAILED, current state is DONE. Marking job as DONE. Reported
harvestErrors on job: dk.netarkivet.common.exceptions.IOFailure: Crawl
probably interrupted by shutdown of HarvestController

HarvestJobManagerApplication.2018-04-09.0.log:15:49:59.836 WARN
d.n.h.datamodel.H3HeritrixTemplate.insertAttributes - Placeholder
'%{MAX_HOPS}' not found in template. Therefore not substituted by '10'
in this template

HarvestJobManagerApplication.2018-04-09.0.log:15:49:59.837 WARN
d.n.h.datamodel.H3HeritrixTemplate.insertAttributes - Placeholder
'%{HONOR_ROBOTS_DOT_TXT}' not found in template. Therefore not
substituted by 'ignore' in this template

HarvestJobManagerApplication.2018-04-09.0.log:15:49:59.837 WARN
d.n.h.datamodel.H3HeritrixTemplate.insertAttributes - Placeholder
'%{EXTRACT_JAVASCRIPT}' not found in template. Therefore not
substituted by 'true' in this template

HarvestJobManagerApplication.2018-04-09.0.log:14:59:59.489 WARN
d.n.h.datamodel.HeritrixTemplate.editOrderXMLAddPerDomainCrawlerTraps
- Found empty trap for domain netarkivet.dk

ArcRepositoryApplication.log:13:11:49.119 WARN
```

```
d.n.a.arcrepository.ArcRepository.startUpload - Trying to upload file
'123-9-20180413105219139-00029-kb-test-har-004.kb.dk.warc.gz' that
already has state UPLOAD_COMPLETED for this replica

BitarchiveMonitorApplication_KBBM.2018-04-10.0.log:13:41:05.321 WARN
d.n.a.bitarchive.BitarchiveMonitor.updateWithBitarchiveReply -
Received batch reply with error: Batch job failed on 1 files. at BA
monitor from bitarchive 10.17.0.56_BitApp_1

BitarchiveMonitorApplication_KBBM.2018-04-11.0.log:11:09:47.037 WARN
d.n.c.distribute.JMSConnectionSunMQ.onException - JMSException with
errorcode 'C4056' encountered:

HarvestJobManagerApplication.2018-04-09.0.log:15:02:01.877 WARN
d.n.h.s.HarvestSchedulerMonitorServer.processCrawlStatusMessage - Job
2 failed: HarvestErrors = java.lang.RuntimeException: Exception during
crawl

GUIApplication.2018-04-10.0.log:11:05:28.412 WARN
dk.netarkivet.common.utils.DBUtils.setStringMaxLength - lastPeekUri of
dk.netarkivet.harvester.harvesting.frontier.FrontierReportLine@96f6d5
e3 is longer than the allowed 1000 characters. The contents is
truncated to length 1000. The untruncated contents was:
https://www.firstpost.com/%22data:image/jpeg;base64...

GUIApplication.2018-04-10.0.log:13:41:05.350 WARN
d.n.a.a.d.JMSArcRepositoryClient.batch - The batch job
'ID:59980-130.226.228.6(f0:ef:fc:a:6:4d)-40252-1523360465135: To
TEST6_COMMON_THE_REPOS ReplyTo
TEST6_COMMON_THIS_REPOS_CLIENT_130_226_228_6_GUIWS OK Job:
```

```
dk.netarkivet.viewerproxy.webinterface.CrawlLogLinesMatchingRegexp, on
filename-pattern: 31-metadata-[0-9]+\.(w)?arc(\.gz)?, for replica: KB'
resulted in the following error: Batch job failed on 1 files.
```

The following kind of warning can be ignored, unless it appears repeatedly:

```
GUIApplication.2018-04-09.0.log:15:01:57.609 WARN
d.n.monitor.jmx.HostForwarding.registerRemoteMbeans - Failure
connecting to remote JMX MBeanserver (Host=kb-test-acs-001.kb.dk,
JMXport=8150, RMIport=8250, last seen live at Mon Apr 09 15:01:47 CEST
2018). Creating an error MBean
```

The following warning may occur after a while, and can be ignored as well:

```
WARNING: Error processing message '
Class:
com.sun.messaging.jmq.jmsclient.ObjectMessageImpl
getJMSMessageID():
ID:40-130.225.27.140(d2:1:3:b1:10:de)-46478-1197902260630
getJMSTimestamp():      1197902260630
getJMSCorrelationID():  null
JMSReplyTo:             null
JMSDestination:         TEST6_COMMON_THE_SCHED
getJMSDeliveryMode():   PERSISTENT
getJMSRedelivered():    false
getJMSType():           null
getJMSExpiration():     0
getJMSPriority():       4
Properties:             null'
dk.netarkivet.common.exceptions.UnknownID: Job id 23 is not known in
persistent storage
        at
dk.netarkivet.harvester.datamodel.JobDBDAO.read(JobDBDAO.java:294)
        at
dk.netarkivet.harvester.scheduler.HarvestSchedulerMonitorServer.proce
ssCrawlStatusMessage(HarvestSchedulerMonitorServer.java:103)
        at
dk.netarkivet.harvester.scheduler.HarvestSchedulerMonitorServer.visit
(HarvestSchedulerMonitorServer.java:285)
        at
dk.netarkivet.harvester.harvesting.distribute.CrawlStatusMessage.acce
pt(CrawlStatusMessage.java:133)
        at
dk.netarkivet.harvester.distribute.HarvesterMessageHandler.onMessage(
HarvesterMessageHandler.java:67)
        at
com.sun.messaging.jmq.jmsclient.MessageConsumerImpl.deliverAndAcknowl
edge(MessageConsumerImpl.java:330)
        at
com.sun.messaging.jmq.jmsclient.MessageConsumerImpl.onMessage(Message
ConsumerImpl.java:265)
        at
com.sun.messaging.jmq.jmsclient.SessionReader.deliver(SessionReader.j
ava:102)
        at
com.sun.messaging.jmq.jmsclient.ConsumerReader.run(ConsumerReader.jav
a:174)
        at java.lang.Thread.run(Thread.java:595)
```

Any other warning should be considered a release test failure.

2. Go to the system overview page and check that all the expected applications are listening and are up without warnings or errors.
   If there is a warning of this kind:

   ```
   Remote JMX bean generated exception:
   javax.management.InstanceNotFoundException: dk.netarkivet.common.logging:applicationinstanceid=,name
   ```

```
=error_host_kb-test-har-004.kb.dk_8150,httpport=8076,machine=kb-test-adm-001.kb.dk,applicationname=dk.netarkivet.common.webinterface.GUIWebServer,index=0,channel=,replicaname=KBN,hostname=kb-test-har-004.kb.dk,location=K
```

then refresh the system state overview page. The warning should disappear.

3. Check that the scheduler schedules only one job for the hourly selective harvest.

Check that a job can be resubmitted

1. Go to "Harvest status""All Jobs", select job status "Failed", and press "Show". Check that you can reject a job for resubmission using the "Reject?" button so that it is no longer visible when you list failed jobs.
2. Check that you can see the rejected job when you now list all jobs.
3. Click on one or more "Genstart"/"Restart?" buttons to resubmit. Note that you only can resubmit jobs failed due to harvesting errors, not due to upload errors.
4. Check that the job-status changes to "resubmitted" and that a new Job is made from the same harvestdefinition with the same configurations.
5. Check that resubmitted jobs contain information about which job they were resubmitted (NAS-1466)

Check Report Generation

Use a browser set up as a viewerproxy connection for this test (see https://sbprojects.statsbiblioteket.dk/pages/viewpage.action?pageId=37597440#TheNetarkivetDistributedTest/DevelEnvironment-ViewerProxyUsage). Select any completed job and click on the "Browse reports for jobs" link.

You should see a list like

```
metadata://netarkivet.dk/crawl/setup/duplicatereductionjobs?majorversion=1&minorversion=0&harvestid=1&harvestnum=10&jobid=14
metadata://netarkivet.dk/crawl/setup/crawler-beans.cxml?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/setup/harvestInfo.xml?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/setup/seeds.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/archivefiles-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/crawl-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/frontier-summary-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/hosts-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/mimetype-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/processors-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/responsecode-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/seeds-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/source-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/reports/threads-report.txt?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/alerts.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/crawl.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
```

metadata://netarkivet.dk/crawl/logs/heritrix3_err.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/heritrix3_out.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/heritrix_out.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/job.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/nonfatal-errors.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/progress-statistics.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/runtime-errors.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/scope.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14
metadata://netarkivet.dk/crawl/logs/uri-errors.log?heritrixVersion=3.3.0-LBS-2016-02&harvestid=1&jobid=14

```
metadata://netarkivet.dk/crawl/index/cdx?majorversion=2&minorversion=0&harvestid=1&jobid=14&filename=14-1-20161101215537865-00000-ciblee_2015_sb-test-har-001.statsbiblioteket.dk.warc
```

Check that a few (like, 3) of the entries are present and browse each in turn. (Note that the HeritrixVersion, harvestIf, and jobId will differ). Some of the entries might be empty.

The following two tests ("Database crash test" and "Network recovery test") must be coordinated with the other testers.

Database crash test

Tests that the system can survive a database crash/stop and resume operation after the database is restarted

1. Log in as root on kb-test-adm-001

   ```
   ssh test@kb-test-adm-001
   su
   ```

2. Stop the postgresdb and wait a couple of minutes.

   ```
   /etc/init.d/postgresql stop
   ```

3. Verify that the GUI has lost the connection to the database by listing domains or harvest definitions.
4. Restart the database

   ```
   /etc/init.d/postgresql start
   ```

5. Check that the different GUI pages works as usual.
6. Create a new active selecive and verify the a job is created and started.

Network recovery test

Tests that the system can survive a network crash/stop and resume operation after the becomes available.

Disable the network/switch/DC for some minutes and see that all batch processes reconnects and continue after restart

1. login on to kb-test-adm-001 as root and stop the networkinterface by installing a cron-job that does this for you:
   Install script restartNetworkWithWait.sh as root cronjob (Add 0 17 * * * (/root/restartNetworkWithWait.sh) to restart network at 5 PM)

   ```
   #!/bin/bash

   # stopping network
   /etc/init.d/network stop
   # waiting 3 minutes
   /bin/sleep 3m
   # starting network
   /etc/init.d/network start
   ```

2. Check that the connection to the GUI is lost.
3. After 5 minutes verify the system comes back online
   a. Verify that the GUI pages are working properly.
   b. Create a new active selective harvest definition and verify that a new job is created and started.
   c. Run a batch job or two and verify these work correctly.

Shutdown the system