D4Science Infrastructure - Task #12227

Encoding issue on Dataminer proto 5

Jul 24, 2018 11:48 AM - Gianpaolo Coro

Status:	Closed	**Start date:**	Jul 24, 2018
Priority:	Normal	**Due date:**	
Assignee:	_InfraScience Systems Engineer	**% Done:**	100%
Category:	High-Throughput-Computing	**Estimated time:**	0.00 hour
Target version:	Data Processing		
Infrastructure:	Production		

Description

I have a difficult issue on one of the prototype Dataminers for which I need help:

There are several algorithms of text analysis that read input files in UTF-8 encoding and write json files in UTF-8. They run in R.

Only on dataminer5-proto, the UTF-8 file write crashes when there is a stressed character in the text (e.g. "oggi è una bella giornata"). The reported error is generic:

```
Warning message:
In writeLines(json, fileConn) : invalid char string in output conversion
```

writeLines is a native R function and is invoked correctly in the code:

```
fileConn<-file(outjsonfile,encoding = "UTF-8")
writeLines(json, fileConn)
close(fileConn)
```

Input files are plain text files read as UTF-8 files using bytes:

```
inputFile <- file(inputfile, encoding="UTF-8")
filetext<-readChar(inputFile, file.info(inputfile)$size, useBytes = T)
```

The only package used by the algorithms is "jsonlite".
I have checked the machine and R locales but they seem OK. Perhaps there is some other difference in the locales I cannot see.
From sample tests, this issue occurs only on dataminer5-proto.

History

#1 - Jul 24, 2018 12:10 PM - Gianpaolo Coro

For the time being, we are going to stop dataminer5-proto in order to make the NLP Hub work.

#2 - Jul 24, 2018 04:25 PM - Andrea Dell'Amico

- Status changed from New to In Progress

I just compared the relevant parts of dataminer4-proto and dataminer5-proto without founding any difference:

- environment variables
- R version
- version of the jsonlite R package
- tomcat options
- smartgears version

are the same on both servers. Is there a way to run a test from command line, so that I can trace the execution?

#3 - Jul 24, 2018 11:48 PM - Gianpaolo Coro

Yes, the issue is weird but this morning I have verified it is just on that machine. The process that highlighted the issue sends an XML file via POST to the DM, which contains the UTF-8 text. Indeed, I had thought there was something in the tomcat locale that saved the file in non-UTF format on the DM. The POST request is done by another DM algorithm and reproducing it as a standalone call could be a bit long.

A direct test using a file on the Workspace can be done using this link:

```
http://dataminer5-proto.d4science.org/wps/WebProcessingService?request=Execute&service=WPS&Version=1.0.0&gcube
```

-token=<token>&lang=en-US&Identifier=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.TAGME_ITALIAN_NER&DataInputs=inputfile=https%3A%2F%2Fdata.d4science.org%2FRkc1VUJ2ZDdMUGowTnkramdGcUpMcDBFV2JlODh4SEpHbWJQNStIS0N6Yz0;

but I'm not sure the issue will manifest. Perhaps it would be easier to reinstall the machine from scratch, otherwise I will need some time to setup a proper test.

#4 - Jul 25, 2018 06:06 PM - Andrea Dell'Amico

Gianpaolo Coro wrote:

> Yes, the issue is weird but this morning I have verified it is just on that machine. The process that highlighted the issue sends an XML file via POST to the DM, which contains the UTF-8 text. Indeed, I had thought there was something in the tomcat locale that saved the file in non-UTF format on the DM. The POST request is done by another DM algorithm and reproducing it as a standalone call could be a bit long.
>
> A direct test using a file on the Workspace can be done using this link:
>
> http://dataminer5-proto.d4science.org/wps/WebProcessingService?request=Execute&service=WPS&Version=1.0.0&gcube-token=<token>&lang=en-US&Identifier=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.TAGME_ITALIAN_NER&DataInputs=inputfile=https%3A%2F%2Fdata.d4science.org%2FRkc1VUJ2ZDdMUGowTnkramdGcUpMcDBFV2JlODh4SEpHbWJQNStIS0N6Yz0;
>
> but I'm not sure the issue will manifest.

It isn't reproducible regularly? I'm going to remove the host from the load balancer and start the tomcat instance again, so that I can run some tests.

> Perhaps it would be easier to reinstall the machine from scratch, otherwise I will need some time to setup a proper test.

Well, if we don't know why there's such a problem, reinstalling is not a guarantee. The VM was installed at the same time as dataminer4-proto and apparently there is no difference between the two.

#5 - Jul 26, 2018 02:53 PM - Andrea Dell'Amico

- File Q2JqWHY2WlU1RWQ3eGEreHo2MmtPb3lnVUtsYXpLTUxHbWJQNStIS0N6Yz0-VLT added

- % Done changed from 0 to 20

The test you posted does not fail. The xml response is:

```
<wps:ExecuteResponse xmlns:wps="http://www.opengis.net/wps/1.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:ows="http://www.opengis.net/ows/1.1" xsi:schemaLocation="http://www.opengis.net/wps/1.0.0 http://schemas.opengis.net/wps/1.0.0/wpsExecute_response.xsd" serviceInstance="http://dataminer5-proto.d4science.org:80//wps/WebProcessingService" xml:lang="en-US" service="WPS" version="1.0.0">
<wps:Process wps:processVersion="1.1.0">
<ows:Identifier>
org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.TAGME_ITALIAN_NER
</ows:Identifier>
<ows:Title>TAGME_ITALIAN_NER</ows:Title>
</wps:Process>
<wps:Status creationTime="2018-07-26T14:46:53.740+02:00">
<wps:ProcessSucceeded>Process successful</wps:ProcessSucceeded>
</wps:Status>
<wps:ProcessOutputs>
<wps:Output>
<ows:Identifier>non_deterministic_output</ows:Identifier>
<ows:Title>NonDeterministicOutput</ows:Title>
<wps:Data>
<wps:ComplexData schema="http://schemas.opengis.net/gml/2.1.2/feature.xsd" mimeType="text/xml; subtype=gml/2.1.2">
<ogr:FeatureCollection xmlns:ogr="http://ogr.maptools.org/" xmlns:gml="http://www.opengis.net/gml" xmlns:d4science="http://www.d4science.org" xsi:schemaLocation="http://ogr.maptools.org/ result_8751.xsd">
<gml:featureMember>
<ogr:Result fid="F0">
<d4science:Data>
http://data.d4science.org/Q2JqWHY2WlU1RWQ3eGEreHo2MmtPb3lnVUtsYXpLTUxHbWJQNStIS0N6Yz0-VLT
</d4science:Data>
<d4science:Description>Log of the computation</d4science:Description>
<d4science:MimeType>text/csv</d4science:MimeType>
</ogr:Result>
<ogr:Result fid="F1">
<d4science:Data>
http://data.d4science.org/Q2JqWHY2WlU1RWQ3eGEreHo2MmtPa2ZZRXM4bmIvVEhHbWJQNStIS0N6Yz0-VLT
</d4science:Data>
```

```
<d4science:Description>outjsonfile</d4science:Description>
<d4science:MimeType>application/d4science</d4science:MimeType>
</ogr:Result>
</gml:featureMember>
</ogr:FeatureCollection>
</wps:ComplexData>
</wps:Data>
</wps:Output>
</wps:ProcessOutputs>
</wps:ExecuteResponse>
```

The result output file is attached.

#6 - Jul 26, 2018 03:01 PM - Andrea Dell'Amico

Run more than once, it never failed.

#7 - Jul 26, 2018 03:10 PM - Andrea Dell'Amico

@gianpaolo.coro@isti.cnr.it can you run some different test? I cannot explain the behaviour, the errors start on July 20th and last until tomcat was shutdown on July 24th. Did you try a tomcat restart in between?

#8 - Jul 26, 2018 04:41 PM - Gianpaolo Coro

Hi, the tomcat had been restarted. The fact that the test works enforces my guess that there is something at tomcat level, i.e. it occurs when an UTF-8 text is sent directly to the service via POST without passing from the Workspace.

I will be on vacation from next wee, thus I don't have time to assemble a more detailed test. Thus, either we wait after 21 August or you could reinstall the machine.

#9 - Jul 26, 2018 04:58 PM - Andrea Dell'Amico

Gianpaolo Coro wrote:

> Hi, the tomcat had been restarted. The fact that the test works enforces my guess that there is something at tomcat level, i.e. it occurs when an UTF-8 text is sent directly to the service via POST without passing from the Workspace.
>
> I will be on vacation from next wee, thus I don't have time to assemble a more detailed test. Thus, either we wait after 21 August or you could reinstall the machine.

I want to understand what's happening. If there's something broken at the tomcat level, the only possibility is a wrong manual intervention from someone.
Because, again, the VM have to be identical to dataminer4-proto (and to all the other tomcat installations, FYI).

#10 - Jul 26, 2018 05:01 PM - Andrea Dell'Amico

- Tracker changed from Incident to Task

#11 - Sep 03, 2018 08:25 PM - Andrea Dell'Amico

Can we restart this activity?

#12 - Sep 04, 2018 10:51 AM - Gianpaolo Coro

I'm going to build a test for Dataminer 5.

#13 - Sep 04, 2018 11:29 AM - Gianpaolo Coro

I cannot reproduce the issue systematically because it seems to be random. Is it possible to re-install the machine?

#14 - Sep 04, 2018 11:39 AM - Gianpaolo Coro

Is it possible to check that the following information is aligned on all dataminers?

```
environment variables
R version
R locale
tomcat locale
machine locale
version of the jsonlite R package
tomcat options
smartgears version
```

#15 - Sep 04, 2018 04:56 PM - Gianpaolo Coro

After @andrea.dellamico@isti.cnr.it has added this forcing in the Rprofile.site file:

```
readRenviron("/etc/default/locale")
LANG <- Sys.getenv("LANG")
if(nchar(LANG))
   Sys.setlocale("LC_ALL", LANG)
```

the problem does not occur anymore. Why this operation had to be forced?

#16 - Sep 04, 2018 05:26 PM - Andrea Dell'Amico

- % Done changed from 20 to 80

Gianpaolo Coro wrote:

> the problem does not occur anymore. Why this operation had to be forced?

You lamented a locale problem, so I looked for a way to explicitly set the R locale. The above commands set the R locale to be the same as the system one, that is en_US.UTF-8 on all our systems (we explicitly set that one too).

I'm going to add the commands to our Rprofile.site template.

#17 - Sep 04, 2018 05:37 PM - Andrea Dell'Amico

- Status changed from In Progress to Feedback

- % Done changed from 80 to 100

Done. As the problem was impacting dataminer5-proto.d4science.org only, the new version of Rprofile.site will be provisioned during the next infrastructure upgrade.

#18 - Sep 06, 2018 12:16 PM - Andrea Dell'Amico

- Status changed from Feedback to Closed

Files

Q2JqWHY2WlU1RWQ3eGEreHo2MmtPb3lnVUtsYXpLTUxHbWJQNSt6L3c2NkY...z0-VLT	Jul 26, 2018	Andrea Dell'Amico