Backlog Courrier - Anomalie #17437

Les fichiers volumineux ne sont pas capturés (Capture)

06/14/2021 10:44 AM - Etienne FAMERY

Status:	Intégré / Développé / Analysé	**Start date:**	06/14/2021
Priority:	0-Bloquant	**Due date:**	10/07/2021
Assignee:	Quentin RIBAC		
Category:			
Target version:	20.03 CD78		
Tags Courrier:	Branche TMA		

Description

Un fichier volumineux a été scanné par le service (380Mo/~500pages). Après paramétrage, la capture fonctionne, le LOT se trouve bien dans le dossier Done de Capture, les autres pdf sont consultables dans MaarchCourrier mais le pdf de 380Mo ne se trouve pas dans MaarchCourrier.

Related issues:

Related to Backlog Courrier - Anomalie #17806: Impossible d'enregistrer un do...	**Prêt à développer (S)**	07/19/2021

History

#2 - 06/14/2021 12:11 PM - Emmanuel DILLARD

- Status changed from A qualifier to A étudier

- Assignee changed from EDI PO to Florian AZIZIAN

#3 - 06/15/2021 10:41 AM - Florian AZIZIAN

- Assignee changed from Florian AZIZIAN to Etienne FAMERY

#5 - 06/15/2021 11:23 AM - Florian AZIZIAN

- Status changed from A étudier to Complément d'Informations

#6 - 06/23/2021 12:25 PM - Etienne FAMERY

- Assignee changed from Etienne FAMERY to Florian AZIZIAN

- Priority changed from 2-Sérieux to 0-Bloquant

Tests de l'import d'un fichier de 500Mo avec le module MaarchWSClient :

- Test du paramétrage avec un fichier de 1Mo => import dans MaarchCourrier fonctionnel
- Test avec le fichier de 500Mo => différentes erreurs liées à la config de php.ini retournées, après modification du post_max_size, upload_max_filesize et memory_size, aucune erreur retournée dans : logs apache, logs php, logs technique et php de MaarchCourrier et logs de MaarchCapture,

La seule information retournée est lors de l'exécution du script de Capture :

```
*******************************************************************************
**                                Maarch Capture                             **
**  (c) since 2008 Maarch SAS                                                **
*******************************************************************************
[...]
Workflow initialized with id 'WMAARCH_SCAN_TO_MC-1624442042-1369719840'
Get first workflow step name...
Next step name is 'ImportFiles'
MaarchCapture step inputs: Array
(
    [0] => Directory
    [1] => Target
    [2] => Action
    [3] => MoveDirectory
)
```

```
MaarchCapture step: Array
(
    [positional] => Array
        (
        )

    [executable] => MaarchCapture.php
    [command] => Array
        (
            [opts] => Array
                (
                    [positional] => Array
                        (
                        )

                    [ConfigName] => Capture
                    [executable] => init
                    [BatchName] => MAARCH_SCAN_TO_MC
                )

            [name] => init
        )

)

Capture::processStep(ImportFiles)
control of /opt/maarch/MaarchCapture/files/maarchBD//5_GB.pdf
Capture::processStep(SendToMaarch)
Killed
```

#7 - 06/23/2021 12:28 PM - Emmanuel DILLARD

- Project changed from Backlog Courrier to Backlog Capture

- Status changed from Complément d'Informations to A étudier

- Target version deleted (20.03 CD78)

#9 - 06/28/2021 10:17 AM - Emmanuel DILLARD

- Due date changed from 06/21/2021 to 06/28/2021

- Assignee changed from Florian AZIZIAN to Alex ORLUC

#10 - 06/28/2021 10:22 AM - Emmanuel DILLARD

- Target version set to 1.8 (Stable)

#11 - 06/28/2021 10:24 AM - Alex ORLUC

- Assignee changed from Alex ORLUC to Guillaume HEURTIER

#12 - 06/28/2021 11:05 AM - Emmanuel DILLARD

- Status changed from A étudier to En cours

#13 - 06/28/2021 04:14 PM - Emmanuel DILLARD

- Status changed from En cours to En cours (S)

#15 - 06/29/2021 11:17 AM - Guillaume HEURTIER

- Project changed from Backlog Capture to Backlog Courrier

- Status changed from En cours (S) to A tester

- Assignee changed from Guillaume HEURTIER to Etienne FAMERY

- Target version changed from 1.8 (Stable) to 20.03 CD78

#17 - 07/06/2021 10:17 AM - Ludovic ARAUJO

- Status changed from A tester to En cours

- Assignee changed from Etienne FAMERY to EDI PO

Nous avons constaté que la capture ce fait avec capturekofax,
La configuration du serveur a été édité : les ram php, le post_max_file_size et upload_max_size

Quentin aurai une idée de l'anomalie sur les fichier volumineux ?
Seule la pièce jointe volumineuse n'est pas ajoutée.

#19 - 07/06/2021 10:53 AM - Emmanuel DILLARD

- Status changed from En cours to En cours (S)

#20 - 07/06/2021 03:15 PM - Emmanuel DILLARD

- Due date changed from 06/28/2021 to 08/03/2021

#21 - 07/06/2021 03:21 PM - Emmanuel DILLARD

- Status changed from En cours (S) to Complément d'Informations

- Assignee changed from EDI PO to Ludovic ARAUJO

Le contournement a-t-il débloqué la situation ?

Si oui, nous cherchons une solution plus aboutie mais cela va prendre un certain temps.

#22 - 07/12/2021 03:14 PM - Quentin RIBAC

- Assignee changed from Ludovic ARAUJO to Quentin RIBAC

#23 - 07/12/2021 03:42 PM - Quentin RIBAC

J’ai pu reproduire l’erreur avec un fichier volumineux (700Mo).

Le script de capture kofaxToMC génère des logs. Pour connaître leur emplacement :

- voir dans la crontab la ligne appelant kofax_capture.sh
- sur cette ligne, kofax_capture.sh doit être suivi d'un chemin d'un dossier
- dans ce dossier il doit y avoir un dossier LogsKofaxToMC, les logs sont à cet endroit

Le log de l'erreur reproduite est :

```
user@maarch-pc$ cat ../../LogsKofaxToMC/kofax_capture_20210712-151820.log
--- kofaxToMC ---

--- [1/1] /home/user/Documents/CD78/kofaxdocs/src/lot/17607_001.xml ---
Hash correct pour : /home/user/Documents/CD78/kofaxdocs/src/lot/17607_001_001.pdf
Hash correct pour : /home/user/Documents/CD78/kofaxdocs/src/lot/17607_001_002.pdf
injection du document principal ... (PDF)
Fatal error: Allowed memory size of 1073741824 bytes exhausted (tried to allocate 990436236 bytes) in /home/us
er/Documents/CD78/kofaxToMC/php/main.php on line 234
```

À cette endroit du code il y a l'appel à base64_encode(file_get_contents($filename)). Ce qui pose problème : il faut charger en mémoire à la fois le contenu de $filename par file_get_contents puis ce contenu encodé en base64 par base64_encode. C'est un double chargement, de plus un fichier en base64 est toujours ~30% plus lourd que l'original, donc pour être sûr, il faudrait allouer, pour un fichier PDF de 500Mo, environ 1.3Go *en plus* de ce qui est actuellement alloué.

À voir si on peut fournir à la fonction base64_encode autre chose qu'un texte brut, pour n'avoir à allouer que la mémoire pour la base64 et non pour le fichier original.

#24 - 07/13/2021 10:07 AM - Quentin RIBAC

- Status changed from Complément d'Informations to En cours (S)

#25 - 07/13/2021 12:25 PM - Quentin RIBAC

- Status changed from En cours (S) to A livrer

- Assignee changed from Quentin RIBAC to Ludovic ARAUJO

Un commit correctif du paquet kofaxToMC a été fait ici :

https://labs.maarch.org/deliveryteam/cd78/commit/13ca4d7d1e88b25eb36e9f677438683d7d914377

Au lieu de charger les fichiers en brut et en base64 puis en json, on prépare un stream du fichier avec un filtre d'encodage base64 qui n'est vraiment ouvert qu'au dernier moment, c'est-à-dire à la construction du json de la requête cURL.

Pour le déployer, se rendre dans le dossier d'installation actuel. Il s'agit normalement de /opt/maarch/modules/kofaxToMC, sinon voir le fichier

référencé dans la crontab.

Vérifier que le code est à jour avec le commit 04684fb1988944d18a582c6569fe1dae0f7f13a4 (avant-dernier commit sur le dépôt) à l'aide de git log puis récupérer les sources avec git pull. Attention, **il faut sauvegarder config.xml** avant de mettre à jour.

Ceci fonctionne sur ma machine pour un pdf de 700Mo en document principal ou en pièce-jointe, cependant il a fallu mettre memory_limit=-1 dans le php.ini de apache2, et le chargement est très long à l'ouverture du courrier dans l'application, de l'ordre de la minute.

#26 - 07/16/2021 09:23 AM - Emmanuel DILLARD

- Status changed from A livrer to A tester

#27 - 08/03/2021 10:12 AM - Emmanuel DILLARD

- Status changed from A tester to Complément d'Informations

#28 - 08/03/2021 05:34 PM - Emmanuel DILLARD

- Status changed from Complément d'Informations to Intégré / Développé / Analysé

#29 - 08/03/2021 05:39 PM - Ludovic ARAUJO

Les correctifs ne permettent la résolution de l'anomalie.
A voir avec nous sur le serveur clients si vous le souhaitez

#30 - 08/03/2021 05:39 PM - Ludovic ARAUJO

- Status changed from Intégré / Développé / Analysé to A traiter

- Assignee deleted (Ludovic ARAUJO)

#31 - 08/03/2021 05:58 PM - Emmanuel DILLARD

- Due date changed from 08/03/2021 to 08/16/2021

- Status changed from A traiter to A étudier

- Assignee set to Guillaume HEURTIER

#32 - 08/03/2021 06:01 PM - Emmanuel DILLARD

- Status changed from A étudier to Etude planifiée

#34 - 08/13/2021 02:34 PM - Emmanuel DILLARD

- Subject changed from Courrier capturé n'arrive pas dans MaarchCourrier to Les fichiers volumineux ne sont pas capturés (Capture)

#35 - 08/13/2021 02:43 PM - Emmanuel DILLARD

- Related to Anomalie #17806: Impossible d'enregistrer un document très volumineux added

#36 - 08/16/2021 11:17 AM - Emmanuel DILLARD

- Status changed from Etude planifiée to Prêt à développer (S)

- Assignee deleted (Guillaume HEURTIER)

#37 - 08/17/2021 03:27 PM - Emmanuel DILLARD

- Status changed from Prêt à développer (S) to En cours (S)

- Assignee set to Quentin RIBAC

#39 - 08/17/2021 05:37 PM - Emmanuel DILLARD

- Due date changed from 08/16/2021 to 08/20/2021

#40 - 09/06/2021 02:34 PM - GIT LAB

Commit ajouté sur la branche **fix/17437/develop** de **MaarchCourrier**
FIX #17437 TIME 1:30 revert to finfo::buffer where the new method was useless, added error handling for new method, removed “resource” mode for getMimeTypeAndFileSize()
https://labs.maarch.org/maarch/MaarchCourrier/commit/0f7fa47d2b55497ce8a011fc9a35f2554b1fe0f9

#41 - 09/06/2021 02:37 PM - Quentin RIBAC

- Status changed from En cours (S) to A tester

#42 - 09/06/2021 03:12 PM - GIT LAB

Commit ajouté sur la branche **fix/17437/develop** de **MaarchCourrier**
FIX #17437 TIME 0 better error messages
https://labs.maarch.org/maarch/MaarchCourrier/commit/2b8b7681454afef5951f4b65aa5450ac651748b5

#43 - 09/06/2021 05:55 PM - GIT LAB

Commit ajouté sur la branche **fix/17437/develop** de **MaarchCourrier**
FIX #17437 TIME 0:10 added missing error handling
https://labs.maarch.org/maarch/MaarchCourrier/commit/eeb22be0405b5f7bec2c56ba953f2fb71f2ffd6d

#45 - 09/09/2021 03:02 PM - Emmanuel DILLARD

- Due date changed from 08/20/2021 to 09/21/2021

#46 - 09/15/2021 09:37 AM - Emmanuel DILLARD

- Target version changed from 20.03 CD78 to 20.03 (Restreint)

- Tags Courrier Branche TMA added

#47 - 09/17/2021 04:20 PM - Emmanuel DILLARD

- Target version changed from 20.03 (Restreint) to 20.03 CD78

#48 - 09/21/2021 10:40 AM - Emmanuel DILLARD

- Due date changed from 09/21/2021 to 10/05/2021

#49 - 09/21/2021 05:50 PM - Emmanuel DILLARD

- Due date changed from 10/05/2021 to 10/07/2021

#50 - 09/28/2021 05:05 PM - Quentin RIBAC

- Status changed from A tester to Intégré / Développé / Analysé