Transfer files from S3 bucket to HDFS

Premise:

1. File transfer follow with CAP standard.
2. Contract file should just be just downloaded to SFTP folder.
3. File should be putted into HDFS folder with daily and whole files folder.

Operation

1. S3 Bucket

There are three different S3 bucket environment as development, stage and production environment, each of these are with different access key and secret key to access to S3 bucket.

For S3 bucket in production, there are 2 folders: Delta and TextFiles. New converted from ORC to .txt files will be stored in the Delta folder, files won’t be moved until EBI job run OK, so that we could ensure the files will never be missed.

For TextFiles folder, whole converted files are stored in the this folder, so that we could get any files that we need for later step use case in some case, also with the files kept for later use.

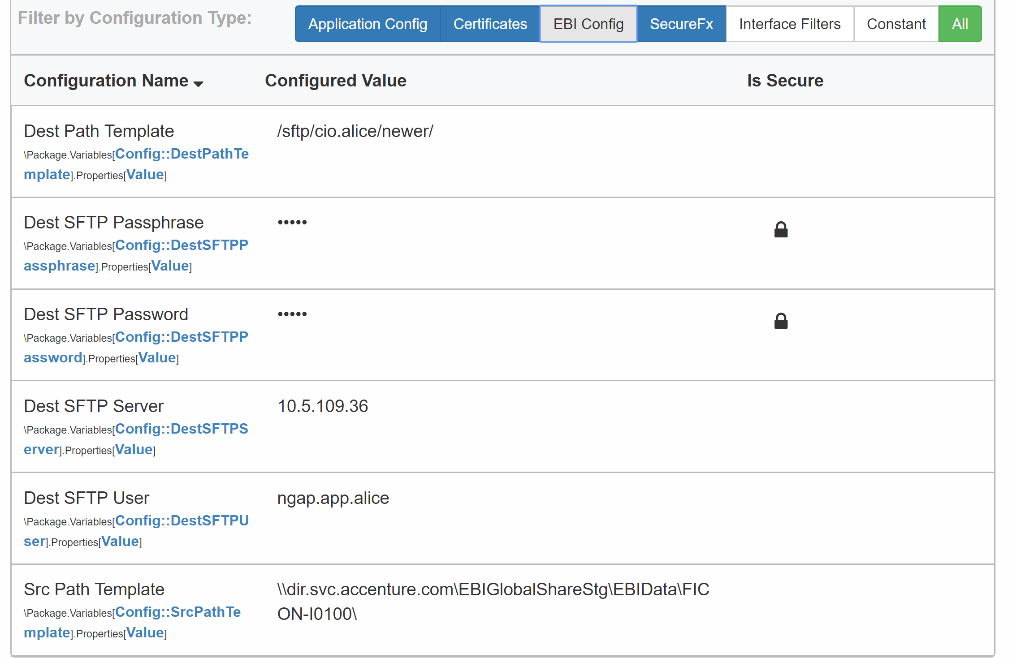
1. EBI job:

EBI job is used to download S3 bucket Delta folder files to SFTP folder.

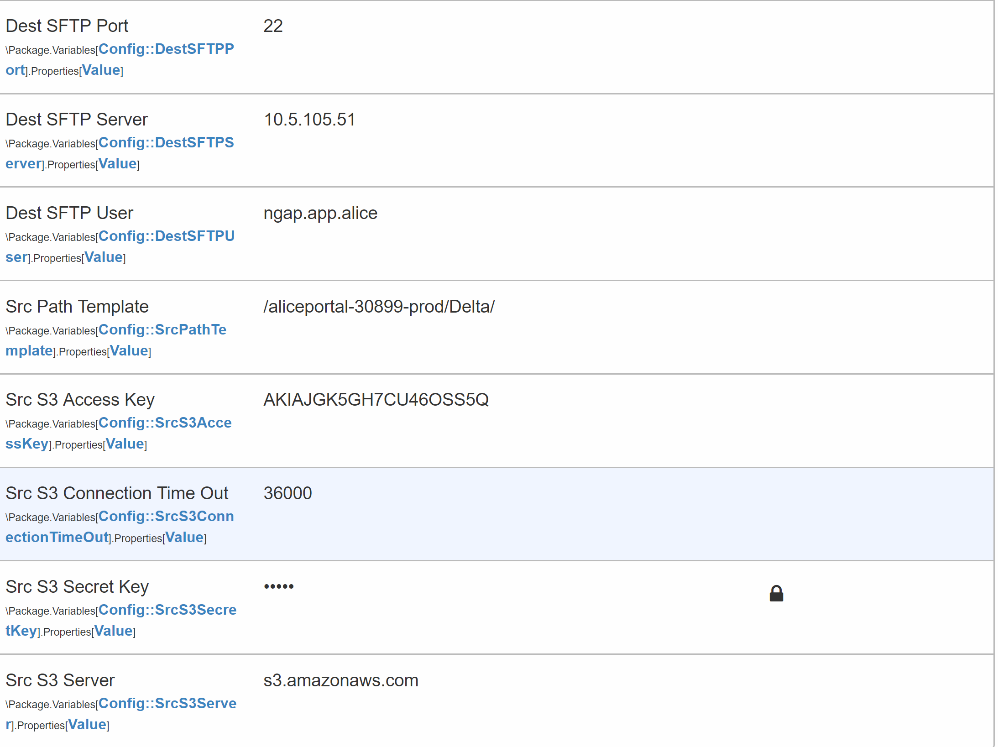
If and only if the in condition job with uploaded files to the S3 bucket folder runs successfully, then the EBI job could run to download files to SFTP.

There are two environment of EBI job, as staging and production, here is link with IMT for EBI: <https://imt.accenture.com/>

Here is staging environment deployed parameters:



Here is production environment deployed parameters:



We couldn’t run the EBI job manually so have to raise ticket to EBI team to help us to run the job to download file to SFTP.

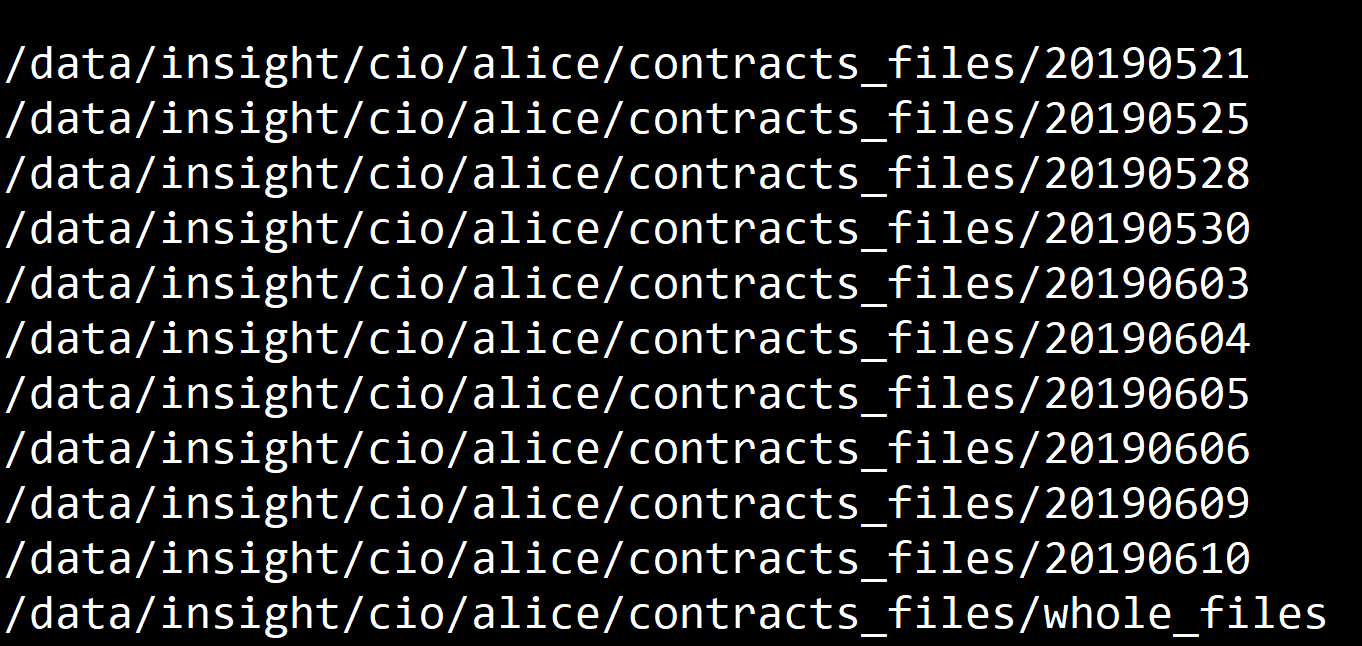
1. Put files to HDFS

For SFTP folder, there will separate with different date to upload the file in the HDFS folder, also with one whole files folder to store with the whole contract files for later use case.

Here for the step that using the HDFS commend: create date folder in SFTP, move whole downloaded file from EBI job to this folder, create one date folder in HDFS, put the whole date folder files to HDFS date folder as daily uploaded files, put the date folder files to the whole file folder.

So we will get date folder in HDFS representing the daily uploaded contract files and one whole files folder to representing whole files converted by OCR in HDFS.

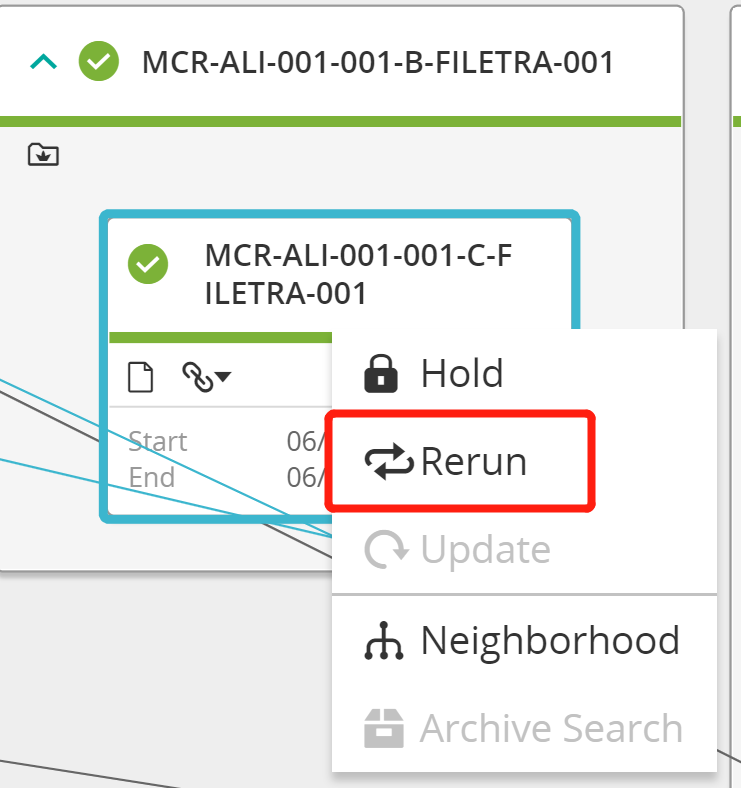
Folder structure as bellow:



1. Move Delta folder files to TextFiles folder in S3 bucket

After the EBI job runs OK, then we will move the whole files in the Delta folder of S3 bucket to the whole files folder as TextFiles, so that we could ensure that the files are already downloaded with EBI without missing files, also we could get the files in the TextFiles in S3 bucket for later use case.

In fact, the job is conditioned with other job, if the conditioned job runs OK, then our job will be triggered automatically, If we want to manually run the job , we could use the Control-M website(with: <https://cloud-ctm.scheduling.accenture.com/SelfService/> ) to find our job: folder name: MCR-ALI-001-001-B-FILETRA-001 with job name: MCR-ALI-001-001-C-FILETRA-001, then rerun the job, it will run in background.



After the whole process finished, then we could find the whole .txt files in HDFS folder both in the date folder and the whole\_files folder for later step to process.