Task 3

Objective: understand end-to-end pipeline for an ETL project

Step 1: Data validation

- There will be one/many data files (gz format) in along with just one control file (text format) which has number of records
- Validate the record count in data file with what we have in control file. If both are matching then move further with ETL processing else just abort the job.

Step 2: ETL

- Calculate #users for each agent and agreement status provided negotiations haven't started at all should be excluded from calculation
- Save the result for each agent at separate location

Step 3: Deliver to downstream

- Deliver the extract at given NAS location along with control file as mentioned in Step 1

Step 4: Think of all the aspects of job failure/restartability and prepare a shell script which can be triggered from a job scheduler

Ref:

- https://stackoverflow.com/questions/42761912/how-to-read-gz-compressed-file-by-pyspark
- https://kb.iu.edu/d/afar