**Details to Schedule Batch Processing**

In order to run the batch processing, we have to consider 3 code files that are as follows:

1. run\_code.py
2. option\_util.py
3. py file to be run in batch mode. E.g., Data\_EIN\_Network.py, Download\_Files.py or Humana\_Billing\_Code.py

Each python file serves a particular purpose that are as mentioned below:

1. run\_code.py File – It helps to create the bash files and divide the items into small chunks as specified in the input parameter. The chunks are created in the JSON file format and saved separately. In the case of Bash files, all the parameters are defined which are required by Slurm as defined by the HPC team. Jobs are submitted on the Slurm using the same file.

2. option\_util.py – This file consists of functions that are used to get the arguments from the input parameters and interpret it. The file path e.g., main, code, bash, input json directories are defined in this which are later used to execute the code.

3. Main .py file – These python files consist of the actual codes which are executed in the slurm in batch mode using the small chunks. The small chunks are used as input from the JSON file to process in the master for loop with the codes needed to execute the task.

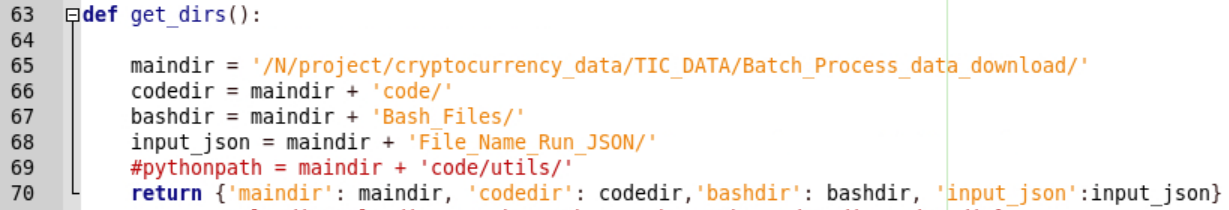
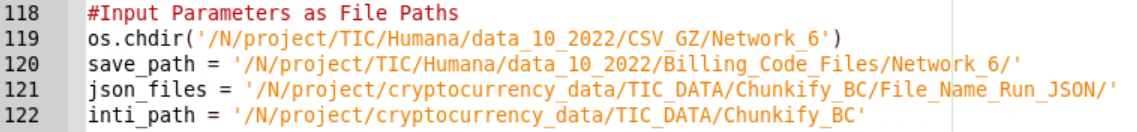
**NOTE:** Keep all the .py code files in a single folder and mention it as specified in step 3.

The input file used for the whole process must be divided into small parts or chunks. Generally, it will be a CSV file consisting of all CSV GZ file names from where the various billing codes has to be extracted, the name of the CSV GZ files which has to be downloaded, name of index JSON files with file paths to gain insights, etc.

The steps to set up the batch processing are as follows:

1. Create the CSV file that will be divided into small parts or chunks. Each batch process will consider the data and process it in separate jobs to expedite the process instead of one single loop. E.g., all the file paths of the CSV GZ files to be processed to extract the billing code.



1. Write the complete file path of the CSV file in the run\_code.py file consisting variable name as “ofilenames”.
2. Update the file paths of main, bash, code, etc. in the “get\_dirs” function mentioned in option\_util.py.
3. All the Main .py file consists of the file paths as mentioned:
   1. save\_path – To save the final output files. e.g., downloaded CSV\_GZ files, CSV file consisting separate billing codes, etc.
   2. json\_files – Saved small chunks of JSON files after running the run\_code.py file. Mention the same file path as in “get\_dirs” function in option\_util.py file.
   3. inti\_path – Mainly used to save the log/file tracking for the completed/processed data.
4. Save all the files and open “Terminal” with file path that consists of all the codes.
5. Write the below mentioned code and execute it in the terminal.



**NOTE:** “—chunks” parameter specifies the number of chunks to be created for the process. Furthermore, “—runcode” parameter will consist of the .py file which has to be executed and obtain the output.

**NOTE:** All the codes are saved in the location /N/project/TIC/Humana/code\_final.