This notebook contains no code but instructions only

Move data pack to local drive to speed things up

* It's highly recommended to copy the the data pack from shared drive to your local C drive to significantly speed up the read/write process
  + To do so you can copy the folder "ChinaMutualFundProd" from Z:\dave\data to your local path. E.g. C:\Users\hyin1\temp\_data\
  + Then change the path variable in Z:\dave\python\webscraping\eastmoney.py accordingly (row 53: path="Z:\dave\data\ChinaMutualFundProd\fund\_details\%s.csv" changed to path="C:\Users\hyin1\temp\_data\fund\_details\%s.csv")
  + Not all the underlying data are important. Fund\_basics and fund\_holdings are the ones we must update. Others such as manager bio, NAV etc are not essential
  + We typically update after each disclosure month (month 1(Q4),3(A),4(Q1),7(Q2),8(S),10(Q3))

Update raw data

use CMD --> python.exe path to run. If encounter error, try to open <https://www.eastmoney.com/> in chrome as a way to trigger the re-connection, then re-run the script

* 1\_get\_static\_list: Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\1\_get\_static\_list.py
  + this one gives the fund universe
  + the underlying informaiton updates daily
* 2\_get\_fund\_basics: Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\2\_get\_fund\_basics.py
  + this one gives the fund tagging (type, issue date etc.)
  + the underlying information udptes daily
  + before running, delete files with basics\_n.csv under \ChinaMutualFundProd\fund\_details
  + run the multi-processing enabled script
  + when one sub process passed through completely, you will receive notification email
  + the process is highly likely to break due to some unknown internet connection issue. When this happens, open eastmoney website (as a trigger), and re-run the script (the scirpt will start from the last break point). You can also see new basics\_n.csv dumped to \ChinaMutualFundProd\fund\_details
  + in the script there is finished=[], you can input process code to skip the finished batch. e.g. finished=[0,3,4] will skip the 0, 3, 4 subprocess (we run 30 parallel downloaders). DO NOT FORGET to empty the list once you finish all the batches
* 3\_ update holdings (top up updating on the existing data block)
  + If you are updating for annual and semi-annual disclosure, first run Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\3\_1\_2\_clean\_existing\_hlds\_for\_annual\_and\_semi\_topup.py . Otherwise, ignore this step
  + Run Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\3\_2\_get\_fund\_hlds\_topup.py untill you get everything (Follow the same procedures in 2\_get\_fund\_bacis, e.g. delete all hlds\_topup.csv files)
  + No need to worry about creating initial hlds data block, please refer to 3\_0\_get\_fund\_hlds.py and 3\_1\_1\_combine\_hlds.py just in case
  + Run Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\3\_3\_merge\_fund\_hlds\_topup.py
* 3.1\_ check downloading quality (IMPORTANT): for unknown reason there maybe some fund positioning skipped during the downloading process. We need to find these out and do some "make up"
  + Run Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\3\_4\_check\_hlds\_quality\_1.py, follow the instructions written in the script
  + Run Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\3\_4\_check\_hlds\_quality\_2.py, follow the instructions written in the script
  + Run Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\3\_4\_check\_hlds\_quality\_3.py, follow the instructions written in the script
    - If you see err msg 'Need run step 2 until no -1 in the is\_incomplete\_download column' then you need to re-run step 3\_4\_check\_hlds\_quality\_2
    - If you see "There are some funds with incomplete holdings download!" (this is likely the results) then you can go to 3\_5\_get\_fund\_hlds\_makeup.py and 3\_6\_merge\_hlds\_with\_makeup.py
      * after running 3\_6\_merge\_hlds\_with\_makeup.py, repeat the above process from 3\_4\_check\_hlds\_quality\_1.py
    - If you see "All fund holdings are clean now!" you can stop the holdings update loop
* 4\_ update nav
  + No need to worry about creating initial nav data block, please refer to 4\_2\_combine\_nav.py and 3\_1\_1\_combine\_hlds.py just in case
  + Run Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\4\_3\_get\_fund\_nav\_topup.py
* 5\_ run scripts from 5\_1 to 8\_1 for non-essential data including manager track record, manager bio, fund filing date, fund investor type split, and fund share outstanding
  + Just remember to delete the corresponding csv dump in the fund\_details folder before the batch run
  + All the data from 5\_1 to 8\_1 are non-topup update
* 6\_ Finally run Z:\dave\python\china\_mutual\_fund\A\_update\_raw\_data\9\_tidyup\_all\_data.py

Process raw data

use CMD to run this Z:\dave\python\china\_mutual\_fund\B\_process\_raw\_data\process\_raw\_data\_all\_in\_one.py

* Note that this process takes some time to run and will consume a lot of memory and CPU. If you see memory error (likely to happen on B\_a\_dump\_network\_stats.py) then try to reduce job\_N and re-run