###Run Instructions###

env.py contains all env varibles. Need to change Mysql connection parameter
queries.sql contains all table schema
main.py is main file to run the process

run main.py as follows for inital load :
python3 main.py 0

and as follows for one day load:
python3 main.py 1

CheckPoints:

- The pipeline should be able to do a full load to begin with and then fetch the incremental data every day.[Done]
- Modeling the data and choice of database is left to you.[Done]
- The pipeline should be able to deal with failures and should be designed to fetch and store data in a performance-efficient manner.[Done]
- Send us the link to a publicly accessible repository with the code and a high-level design of the tools used, flow, and data model.[Done]
- Send us a checklist of the tasks you have completed and what may have been missed due to time constraints or otherwise.[Done]

Go the Extra Mile!

- Try to get the initial load(fetch+DB insertion) to complete in less than 3 hours.[Done]
- Model the data and tune it such that read queries for a particular fund/multiple funds over a certain time period are performant.[Done]
- Make the project live(a free-tier cloud service DB, docker, etc.) and send us an executable/container.[No]

ToolsUsed:
python3 with Mysql Database.
Involved core Libraries are mysql.connector ,concurrent.futures,requests,logging

Modelling the data is done Based on following assumptions:

Here each type(Mutual Fund Type) is stored in type_table with excremental id,name columns (Ex: types are Close Ended Schemes (Balanced) ,Open Ended Schemes (ELSS) etc..)

And each group(Mutual Fund) is stored in group_table with excremental id,name columns(Ex: Aditya Birla Sun Life Mutual Fund,Axis Mutual Fund etc)

And each scheme is store is the nav_records table with the remaining appropriate detials.

In the above modeling. Type could have been further divided into sub-type based on values present in brackets. Was not sure whether to do or not and didn't stress on it enough because of time constraints.
Also the model is storing scheme name only once and not checking again in the future if scheme name gets checked in future. Should update in future. That's why maintained this field as text and new names would be appended at the end with comma separator but this feature is currently not implemented.

Schema for all 3 tables can be found below.

SchemaUsed is:(Same is present in queries.sql file)

```
CREATE TABLE `nav_records` (
  `id` int(11) unsigned NOT NULL AUTO_INCREMENT,
  `scheme_id`  int(11) unsigned,
  `isin_div_payout_or_growth` varchar(100) NOT NULL DEFAULT '',
  `isin_div_reinvestment` varchar(100) NOT NULL DEFAULT '',
  `nav` float(10,4) NOT NULL DEFAULT '0.0000',
  `repurchase_price` float(10,4) NOT NULL DEFAULT '0.0000',
  `sale_price` float(10,4) NOT NULL DEFAULT '0.0000',
  `record_date` DATE NOT NULL DEFAULT '0000-00-00',
  `group_id` int(11) unsigned,
  `type_id` int(11) unsigned,
   PRIMARY KEY (`id`),
   KEY `scheme_id`(`scheme_id`),
   KEY `group_id`(`group_id`),
   KEY `type_id`(`type_id`),
   KEY `record_date`(`record_date`)
);


CREATE TABLE `scheme_table`(
        `id` int (11) unsigned NOT NULL AUTO_INCREMENT,
```

```
	`scheme_id` varchar(250) NOT NULL,
	`name` text NOT NULL DEFAULT '',
	PRIMARY KEY(`id`),
	UNIQUE KEY`scheme_id`(`scheme_id`)
);


CREATE TABLE `group_table`(
	`id` int (11) unsigned NOT NULL AUTO_INCREMENT,
	`name` varchar(250) NOT NULL DEFAULT '',
	PRIMARY KEY(`id`),
	UNIQUE KEY`name`(`name`)
);

CREATE TABLE `type_table`(
	`id` int (11) unsigned NOT NULL AUTO_INCREMENT,
	`name` varchar(250) NOT NULL DEFAULT '',
	PRIMARY KEY(`id`),
	UNIQUE KEY`name`(`name`)
);
```

The initial insert time is as follows with 10 connectors. This could be further optimized if we are working with enterprise-level network throughputs and processing powers as we can clearly see the sys time is very less.

```
real  3m41.195s
user  1m54.352s
sys 0m2.228s
```

funds names, type names, scheme names are indexed and the corresponding ids are also indexed for faster querying. Dates were indexed.

Was unable to do containerized environment development. As i have no experience in it. I tried locally but was unable to set it up on the cloud due to time constraints.