# ETL Pipeline

HI These are the below table which you need to run mandatory in the database .

Admin table (MYSQL only):

create table admin\_failed\_job(

id int auto\_increment primary key,

table\_name varchar(20),

dur\_time varchar(20),

arc\_date timestamp default now());

the above table is to store info if any process failed to include this duration also in the next run.

Tables to test which I have used you can use this or any other as you wish .

**CREATE TABLE if not exists events(**

**id int auto\_increment primary key,**

**event\_name varchar(255),**

**visitor varchar(255),**

**properties json,**

**browser json,**

**arc\_date timestamp default now()**

**);**

INSERT INTO events(event\_name, visitor,properties, browser,arc\_date)

VALUES (

'pageview',

'1',

'{ "page": "/" }',

'{ "name": "Safari", "os": "Mac", "resolution": { "x": 1920, "y": 1080 } }','2021-01-19 20:44:19'

),

('pageview',

'2',

'{ "page": "/contact" }',

'{ "name": "Firefox", "os": "Windows", "resolution": { "x": 2560, "y": 1600 } }','2021-01-20 20:44:19'

),

(

'pageview',

'1',

'{ "page": "/products" }',

'{ "name": "Safari", "os": "Mac", "resolution": { "x": 1920, "y": 1080 } }','2021-01-19 10:44:19'

),

(

'purchase',

'3',

'{ "page": "/gmail" }',

'{ "name": "Firefox", "os": "Windows", "resolution": { "x": 1600, "y": 900 } }','2021-01-19 02:44:19'

),

(

'purchase',

'4',

'{ "page": "/stores"}',

'{ "name": "Firefox", "os": "Windows", "resolution": { "x": 1280, "y": 800 } }','2021-01-20 02:44:19'

),

(

'purchase',

'4',

'{ "page": "/500" }',

'{ "name": "Chrome", "os": "Windows", "resolution": { "x": 1680, "y": 1050 } }','2021-01-21 02:44:19'

);

**create table if not exists employees(**

**id int auto\_increment primary key,**

**name json,**

**salary decimal(12,4),**

**mobile varchar(20),**

**arc\_date timestamp default now());**

insert into employees(name,salary,mobile,arc\_date)

values

('{"firstname":"syed","lastname":"mushtaq"}',35000.00,'97425xxxxx',now()),

('{"firstname":"stephen","lastname":"john"}',473460.230,'88678xxxxx',now()),

('{"firstname":"arjun","lastname":"kapoor"}',23415.0340,'984530xxxxx','2021-01-19 23:43:38');

Belwo is the config file which comes under **“./conf/etl\_config.ini”**

# Below is the db credentials

[db\_credentials]  
db\_type=mysql  
db=practice  
host=localhost  
port= 3306  
username= root  
password= root  
  
# tables from which data needs to be extracted . if not mentioned will not be

# extracted  
[tables]  
table\_list=events,employees  
  
  
*# please Specify format files for column names like Table\_Name  
# Path of Format files folder .*

*# these are the files where i have stored you can change based on your #system*

*# first format file is the columns from whcih data extracted file is #”table.fmt”*[support\_files]  
column\_path=G:/Support Files/FormatFile

# json columns is if there are any columns which have json data then give the #format of those json files   
json\_cols\_path=G:/Support Files/Json\_Columns

#path where results to be stored  
output\_path=G:/Support Files/outputs  
  
*# for now it supports only hour or minutes like last 12 hours or last 30 mins*

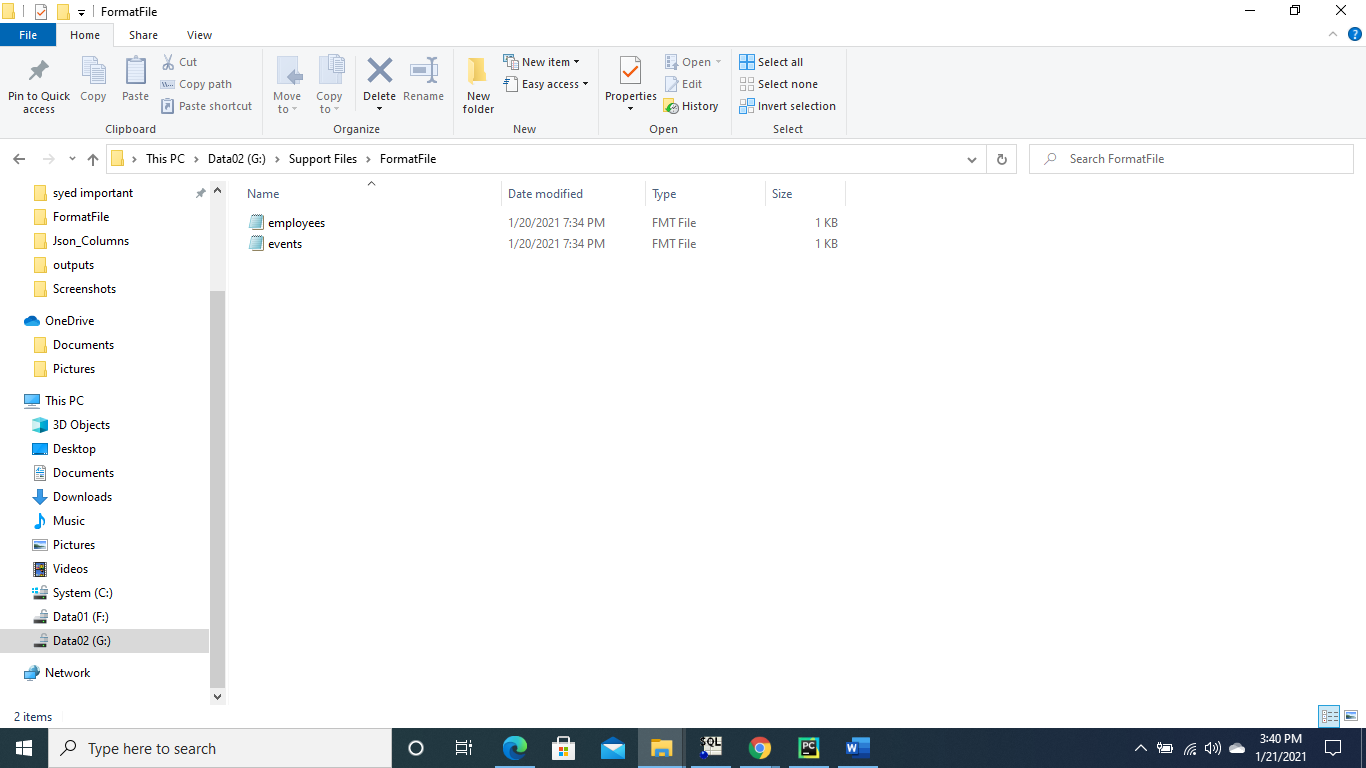
*# data like last 60 minutes or like last 12 hours should specify here*

*# for hours write “****hourly”*** *12*[data\_retriving\_durations]  
data\_interval=hourly   
data\_time\_dur=12  
  
*# below is the duration gap for window like for how many minutes should it schedule  
# write minutes,hourly,day .  
# time dur is not required for hourly , for minutes specify like 5,10,20.for day specify time like '12:30'*

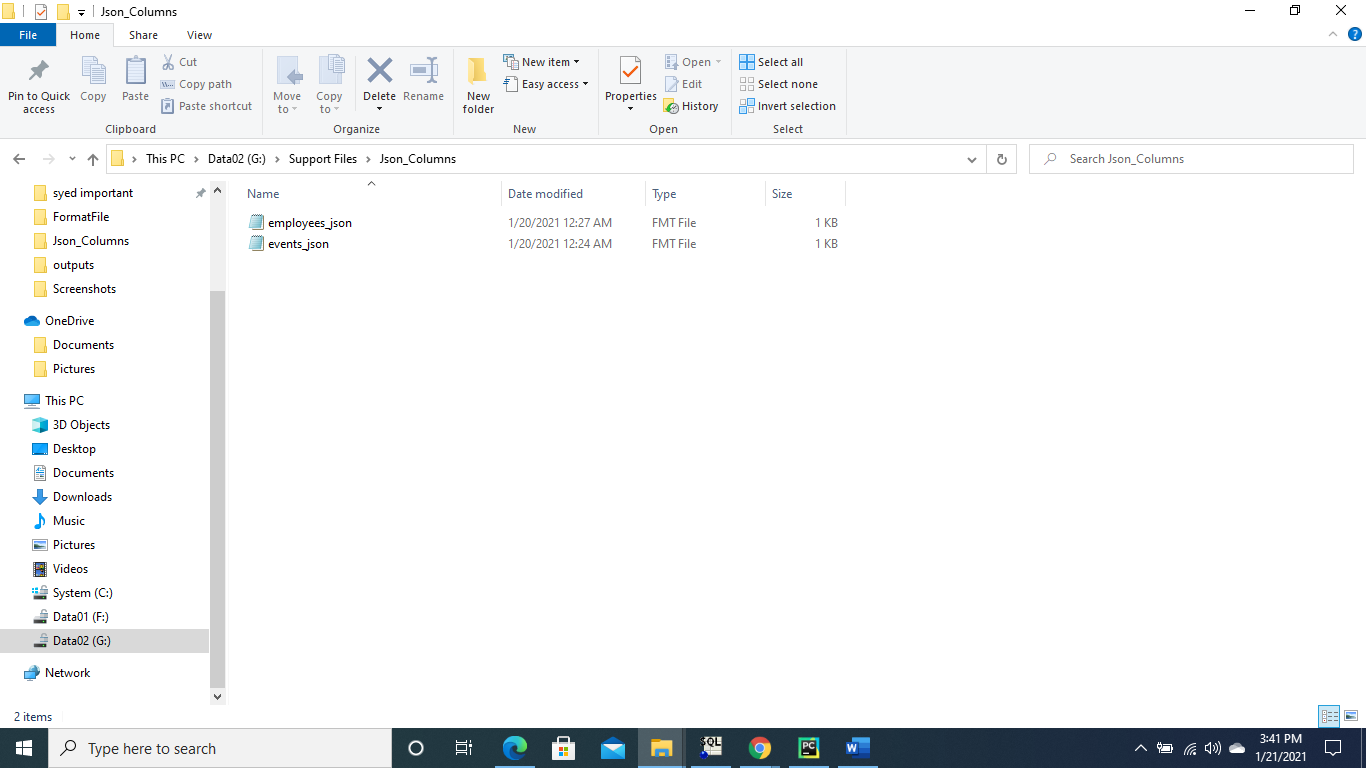
*#For the time when job should run like here for every hour time dur is not #required for hour it is for minutes*[scheduling]  
interval=minutes  
time\_dur=30  
  
  
*# This here is to specify json Columns and structure if any like Table\_Name\_json*

Below are the images like table name is column file name with format file only .

And also look at the location



Belwo is for json columns tablename\_json.fmt file only



**Just specify column keys syntax which is fixed with no data for values**

Since employee contains json name column

{"name":

{"firstname":"",

"lastname":""}

}

The syntax i need to write in table\_json.fmt file

Events

{

"properties":{"page":""},

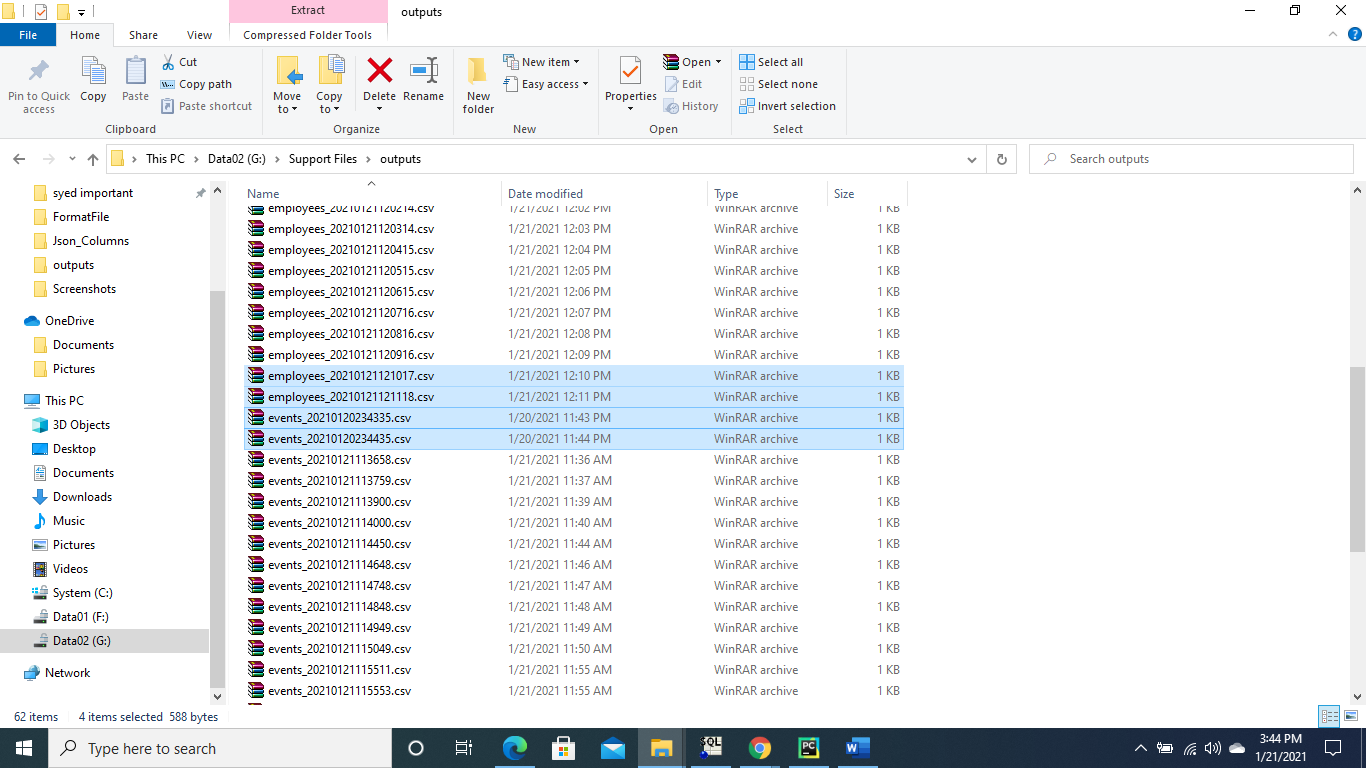
"browser":

{"name":"","os":"","resolution":{"x":"","y":""}}

}

**Just specify column keys syntax which is fixed with no data for values**

**Below is the output files check the path and files are like tablename and timestamp**



To run program as service

Go to that path where program is there and run

**Python File name.py start |stop|restart**

**Python MyDaemon.py start -- in my case**

Note : Use linux system only it does not work in windows

Also I couldn’t create service clean

Below are the files i have attached to be placed **if using my tables**. to test if not use your Own

Under G:/Support Files/FormatFile

Place these files

Under json\_cols\_path=G:/Support Files/Json\_Columns

Place these files

