Professor : Dr. Haroon Malik
Group Name : The Fast & The Curious
Students : Kanimozhi Kalaichelvan, Hafsa Tahir
Date : 11 Dec 2018

Big Data - Google Play Store

Task 1:

Website crawled : Google Play Store
Link : https://play.google.com/store/apps

The designed WebCrawler takes in a Google Play Store link (it can be of any category) as argument and crawls through the given URL link to fetch all the Apps of that Category. It fetches information such as Docid, hreflink and Name of the Apps. It then Stores these information into corresponding Tables in the MySQL Database.

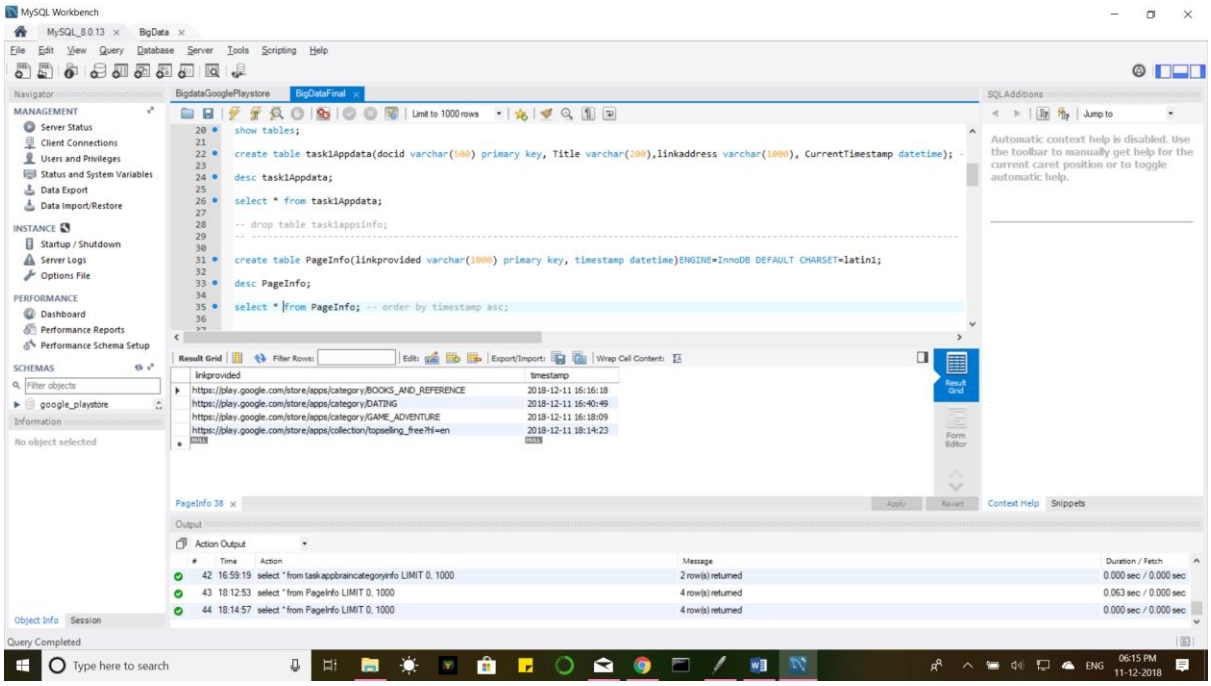
Task 2:

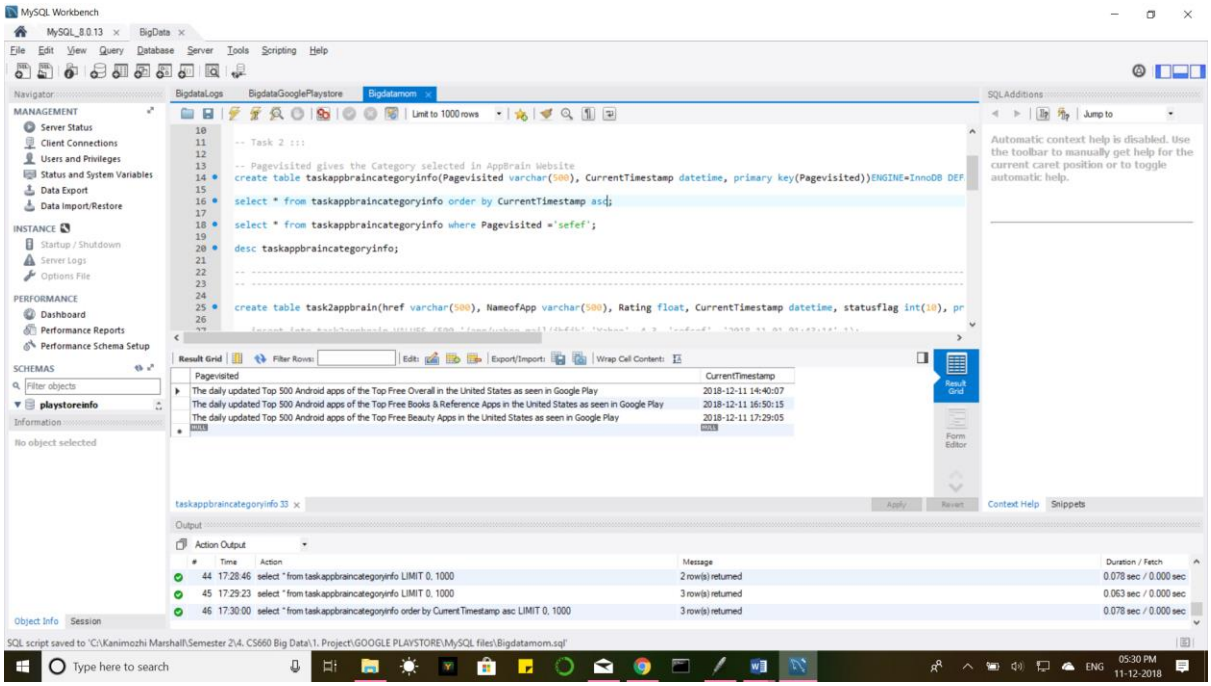
Link :
https://www.appbrain.com/stats/google-play-rankings/top_free/all/ushttps://www.appbrain.com/stats/google-play-rankings/top_free/all/us

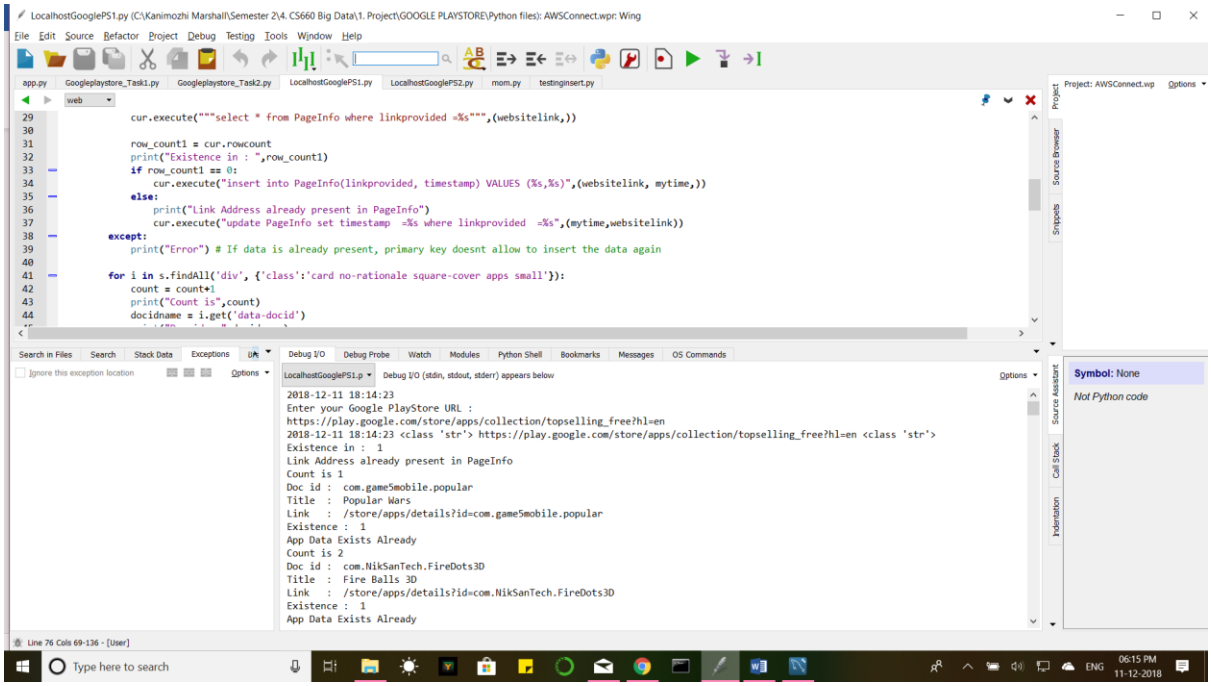
Website crawled : appbrain

The designed WebCrawler takes in a appbrain website link (it can be of any category) as argument and crawls through the given URL link to fetch all the Apps of that Category. Here the WebCrawler crawls through all the pages (usually 5 for the appbrain website) of the given category and fetches details of all the apps present in the 5 pages. It fetches the details such as Appname, hreflink, Rating of the Apps in Appbrain Website. It then Stores these information into corresponding Tables in the MySQL Database.

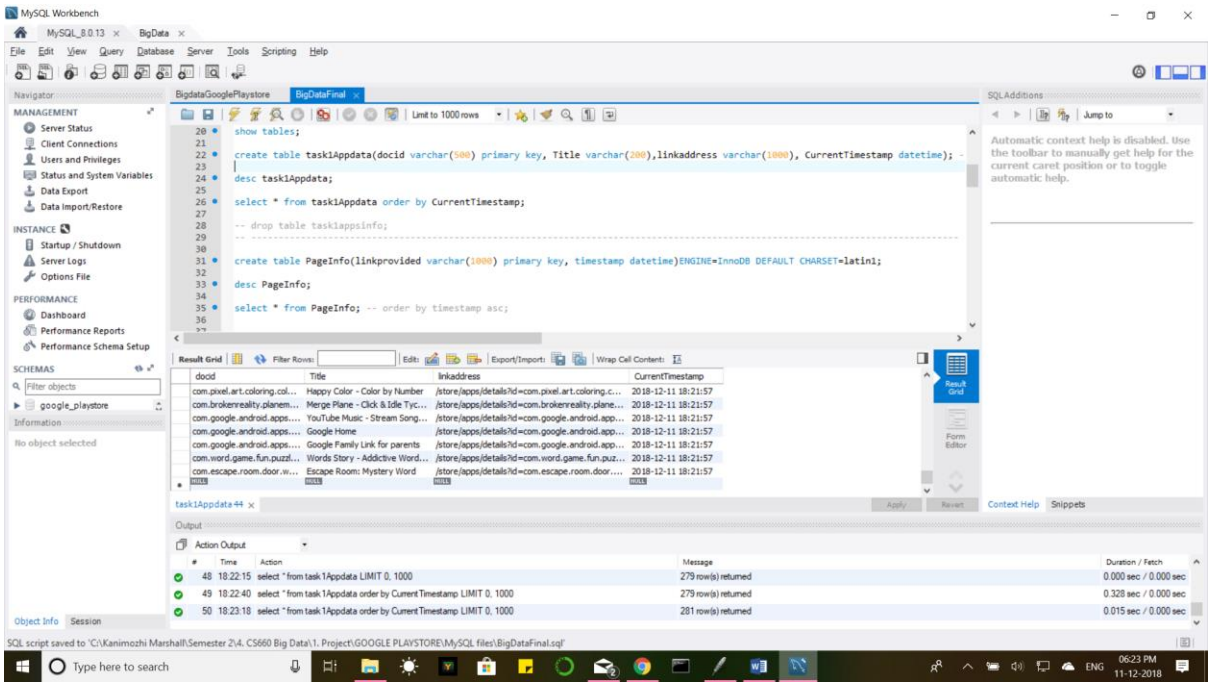
Results :

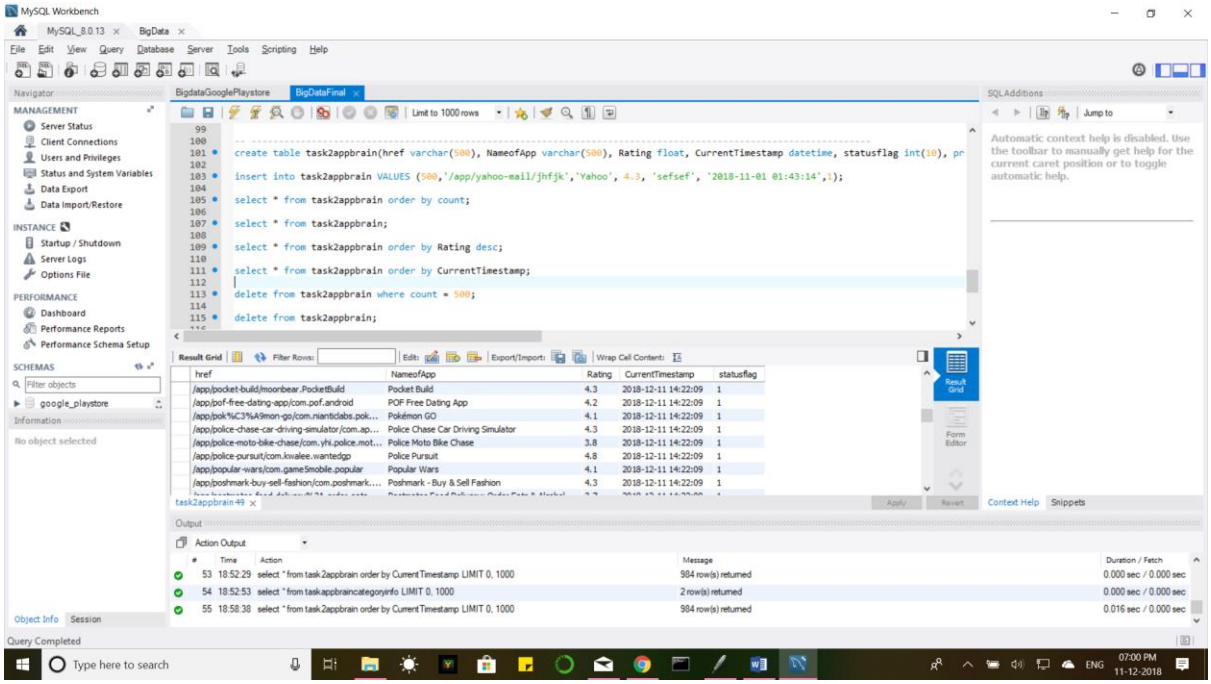

MySQL Workbench
show tables;
create table task1Appdata(docid varchar(500) primary key, Title varchar(200),linkaddress varchar(1000), CurrentTimestamp datetime);
desc task1Appdata;
select * from task1Appdata;
-- drop table task1appsinfo;
create table PageInfo(linkprovided varchar(1000) primary key, timestamp datetime)ENGINE=InnoDB DEFAULT CHARSET=latin1;
desc PageInfo;
select * from PageInfo; -- order by timestamp asc;
linkprovided
timestamp
https://play.google.com/store/apps/category/BOOKS_AND_REFERENCE
2018-12-11 16:16:18
https://play.google.com/store/apps/category/DATING
2018-12-11 16:40:49
https://play.google.com/store/apps/category/GAME_ADVENTURE
2018-12-11 16:18:09
https://play.google.com/store/apps/collection/topselling_free?hl=en
2018-12-11 18:14:23
Action Output
42 16:59:19 select * from taskappbraincategoryinfo LIMIT 0, 1000
2 row(s) returned
0.000 sec / 0.000 sec
43 18:12:53 select * from PageInfo LIMIT 0, 1000
4 row(s) returned
0.063 sec / 0.000 sec
44 18:14:57 select * from PageInfo LIMIT 0, 1000
4 row(s) returned
0.000 sec / 0.000 sec
Query Completed

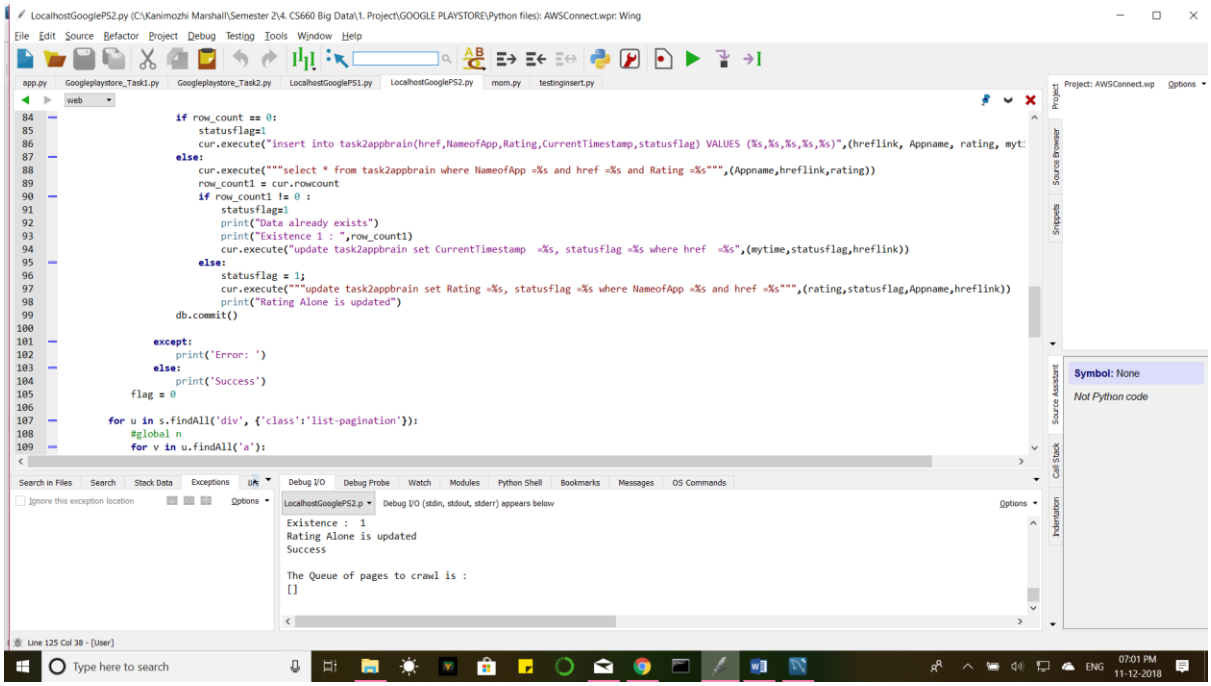

MySQL Workbench
-- Task 2 :::
-- Pagevisited gives the Category selected in AppBrain Website
create table taskappbraincategoryinfo(Pagevisited varchar(500), CurrentTimestamp datetime, primary key(Pagevisited))ENGINE=InnoDB DEF
select * from taskappbraincategoryinfo order by CurrentTimestamp asc;
select * from taskappbraincategoryinfo where Pagevisited ='sefef';
desc taskappbraincategoryinfo;
create table task2appbrain(href varchar(500), NameofApp varchar(500), Rating float, CurrentTimestamp datetime, statusflag int(10), pr
Pagevisited
CurrentTimestamp
The daily updated Top 500 Android apps of the Top Free Overall in the United States as seen in Google Play
2018-12-11 14:40:07
The daily updated Top 500 Android apps of the Top Free Books & Reference Apps in the United States as seen in Google Play
2018-12-11 16:50:15
The daily updated Top 500 Android apps of the Top Free Beauty Apps in the United States as seen in Google Play
2018-12-11 17:29:05
Action Output
44 17:28:46 select * from taskappbraincategoryinfo LIMIT 0, 1000
2 row(s) returned
0.078 sec / 0.000 sec
45 17:29:23 select * from taskappbraincategoryinfo LIMIT 0, 1000
3 row(s) returned
0.063 sec / 0.000 sec
46 17:30:00 select * from taskappbraincategoryinfo order by CurrentTimestamp asc LIMIT 0, 1000
3 row(s) returned
0.078 sec / 0.000 sec

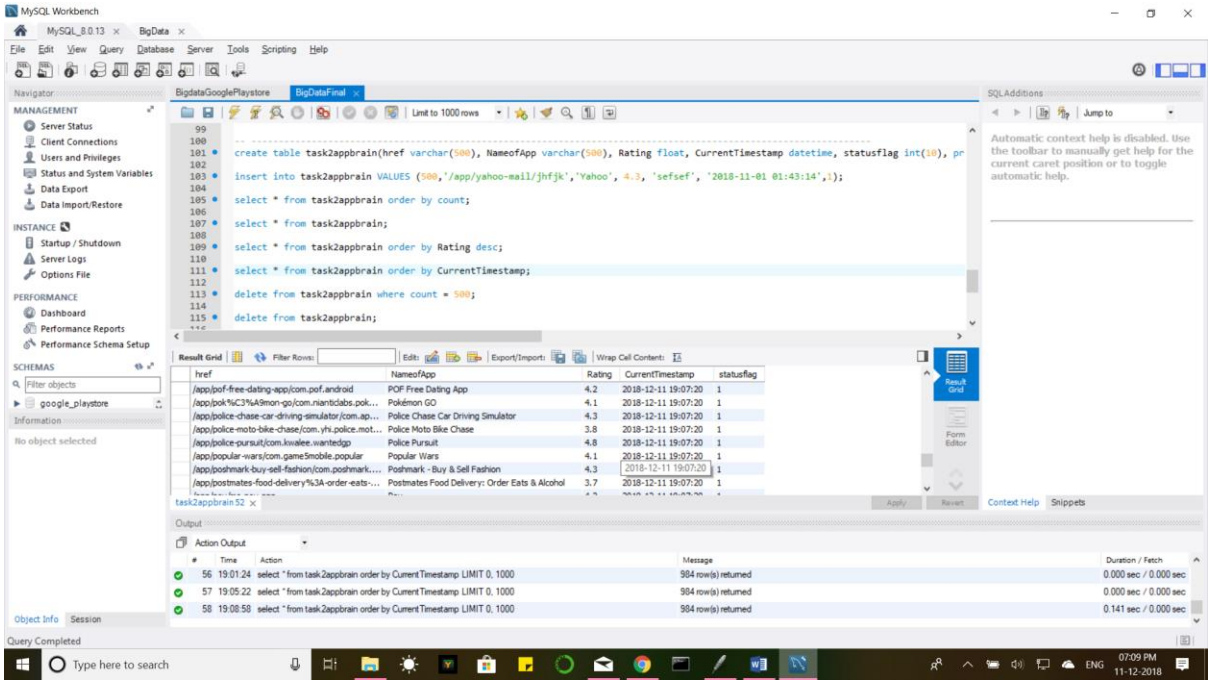


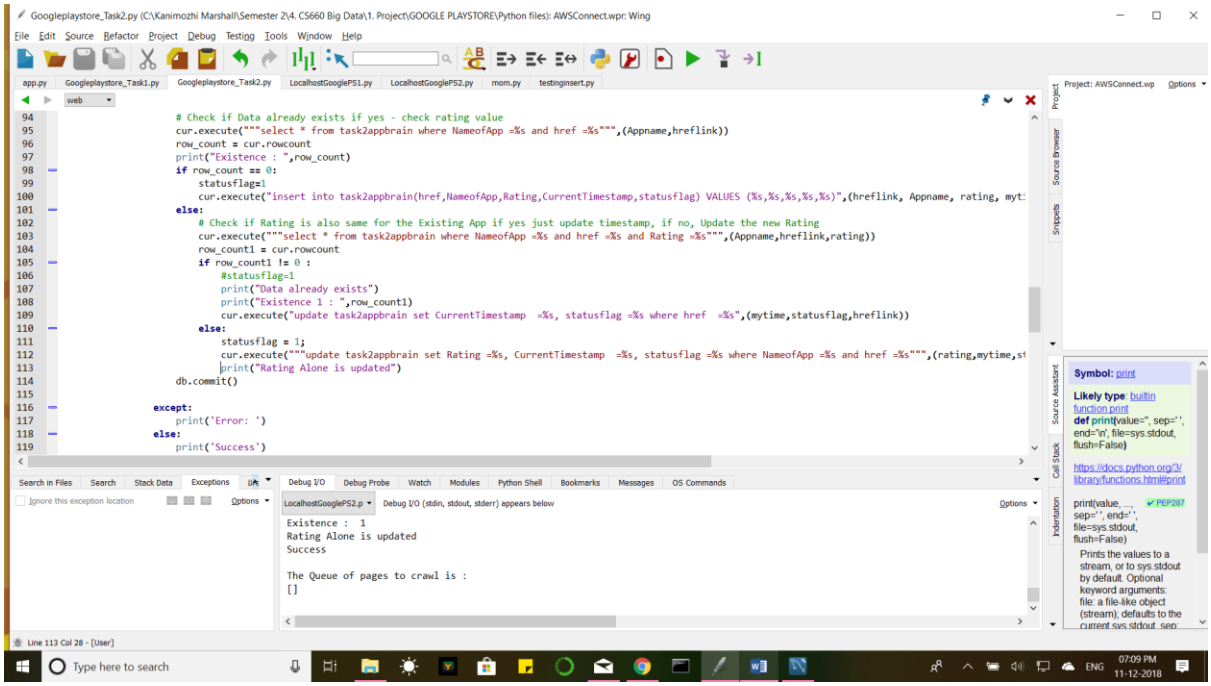
Updated App Data Timestamp:




MySQL Workbench
create table task2appbrain(href varchar(500), NameofApp varchar(500), Rating float, CurrentTimestamp datetime, statusflag int(10), pr
insert into task2appbrain VALUES (500,'/app/yahoo-mail/jhfjk','Yahoo', 4.3, 'sefsef', '2018-11-01 01:43:14',1);
select * from task2appbrain order by count;
select * from task2appbrain;
select * from task2appbrain order by Rating desc;
select * from task2appbrain order by CurrentTimestamp;
delete from task2appbrain where count = 500;
delete from task2appbrain;
Query Completed
Type here to search
07:00 PM
11-12-2018


LocalhostGooglePS2.py (C:\Kanimozhi Marshall\Semester 2\4. CS660 Big Data\1. Project\GOOGLE PLAYSTORE\Python files): AWSConnect.wpr: Wing
if row_count == 0:
statusflag=1
cur.execute("insert into task2appbrain(href,NameofApp,Rating,CurrentTimestamp,statusflag) VALUES (%s,%s,%s,%s,%s)",(hreflink, Appname, rating, myt
else:
cur.execute("""select * from task2appbrain where NameofApp =%s and href =%s and Rating =%s""",(Appname,hreflink,rating))
row_count1 = cur.rowcount
if row_count1 != 0 :
statusflag=1
print("Data already exists")
print("Existence 1 : ",row_count1)
cur.execute("update task2appbrain set CurrentTimestamp =%s, statusflag =%s where href =%s",(mytime,statusflag,hreflink))
else:
statusflag = 1;
cur.execute("""update task2appbrain set Rating =%s, statusflag =%s where NameofApp =%s and href =%s""",(rating,statusflag,Appname,hreflink))
print("Rating Alone is updated")
db.commit()
except:
print('Error: ')
else:
print('Success')
flag = 0
for u in s.findAll('div', {'class':'list-pagination'}):
#global n
for v in u.findAll('a'):
Existence : 1
Rating Alone is updated
Success
The Queue of pages to crawl is :
[]
Symbol: None
Not Python code
Type here to search
07:01 PM
11-12-2018


MySQL Workbench
create table task2appbrain(href varchar(500), NameofApp varchar(500), Rating float, CurrentTimestamp datetime, statusflag int(10), pr
insert into task2appbrain VALUES (500,'/app/yahoo-mail/jhfjk','Yahoo', 4.3, 'sefsef', '2018-11-01 01:43:14',1);
select * from task2appbrain order by count;
select * from task2appbrain;
select * from task2appbrain order by Rating desc;
select * from task2appbrain order by CurrentTimestamp;
delete from task2appbrain where count = 500;
delete from task2appbrain;
Result Grid
href
NameofApp
Rating
CurrentTimestamp
statusflag
/app/pof-free-dating-app/com.pof.android
POF Free Dating App
4.2
2018-12-11 19:07:20
1
/app/pok%C3%A9mon-go/com.nianticlabs.pok...
Pokémon GO
4.1
2018-12-11 19:07:20
1
/app/police-chase-car-driving-simulator/com.ap...
Police Chase Car Driving Simulator
4.3
2018-12-11 19:07:20
1
/app/police-moto-bike-chase/com.yhi.police.mot...
Police Moto Bike Chase
3.8
2018-12-11 19:07:20
1
/app/police-pursuit/com.kwalee.wantedgp
Police Pursuit
4.8
2018-12-11 19:07:20
1
/app/popular-wars/com.game5mobile.popular
Popular Wars
4.1
2018-12-11 19:07:20
1
/app/poshmark-buy-sell-fashion/com.poshmark....
Poshmark - Buy & Sell Fashion
4.3
2018-12-11 19:07:20
1
/app/postmates-food-delivery%3A-order-eats-...
Postmates Food Delivery: Order Eats & Alcohol
3.7
2018-12-11 19:07:20
1
Action Output
56 19:01:24 select * from task2appbrain order by CurrentTimestamp LIMIT 0, 1000 984 row(s) returned 0.000 sec / 0.000 sec
57 19:05:22 select * from task2appbrain order by CurrentTimestamp LIMIT 0, 1000 984 row(s) returned 0.000 sec / 0.000 sec
58 19:08:58 select * from task2appbrain order by CurrentTimestamp LIMIT 0, 1000 984 row(s) returned 0.141 sec / 0.000 sec
Query Completed
Type here to search
07:09 PM
11-12-2018


Googleplaystore_Task2.py (C:\Kanimozhi Marshall\Semester 2\4. CS660 Big Data\1. Project\GOOGLE PLAYSTORE\Python files): AWSConnect.wpr: Wing
# Check if Data already exists if yes - check rating value
cur.execute("""select * from task2appbrain where NameofApp =%s and href =%s""",(Appname,hreflink))
row_count = cur.rowcount
print("Existence : ",row_count)
if row_count == 0:
statusflag=1
cur.execute("insert into task2appbrain(href,NameofApp,Rating,CurrentTimestamp,statusflag) VALUES (%s,%s,%s,%s,%s)",(hreflink, Appname, rating, myt
else:
# Check if Rating is also same for the Existing App if yes just update timestamp, if no, Update the new Rating
cur.execute("""select * from task2appbrain where NameofApp =%s and href =%s and Rating =%s""",(Appname,hreflink,rating))
row_count1 = cur.rowcount
if row_count1 != 0 :
#statusflag=1
print("Data already exists")
print("Existence 1 : ",row_count1)
cur.execute("update task2appbrain set CurrentTimestamp =%s, statusflag =%s where href =%s",(mytime,statusflag,hreflink))
else:
statusflag = 1;
cur.execute("""update task2appbrain set Rating =%s, CurrentTimestamp =%s, statusflag =%s where NameofApp =%s and href =%s""",(rating,mytime,st
print("Rating Alone is updated")
db.commit()
except:
print('Error: ')
else:
print('Success')
Existence : 1
Rating Alone is updated
Success
The Queue of pages to crawl is :
[]
Symbol: print
Likely type: builtin function print
def print(value=', sep=' ', end='\n', file=sys.stdout, flush=False)
https://docs.python.org/3/library/functions.html#print
print(value, ..., sep=' ', end='\n', file=sys.stdout, flush=False)
Prints the values to a stream, or to sys.stdout by default. Optional keyword arguments:
file: a file-like object (stream); defaults to the current sys.stdout. sep:
Type here to search
07:09 PM
11-12-2018
