

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/us
https://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 :

The screenshot shows the MySQL Workbench interface with the following details:

- Navigator:** Schemas list includes `google_playstore`.
- SQL Editor:** Contains the following SQL script:

```
20 show tables;
21
22 create table task1Appdata(docid varchar(500) primary key, Title varchar(200), linkaddress varchar(1000), CurrentTimestamp datetime);
23
24 desc task1Appdata;
25
26 select * from task1Appdata;
27
28 -- drop table task1Appdata;
29
30
31 create table PageInfo(linkprovided varchar(1000) primary key, timestamp datetime)ENGINE=InnoDB DEFAULT CHARSET=latin1;
32
33 desc PageInfo;
34
35 select * from PageInfo; -- order by timestamp asc;
```
- Result Grid:** Displays the output of the `select * from PageInfo` query:

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
- Output:** Shows the execution of three queries:

#	Time	Action	Message	Duration / Fetch
42	16:59:19	select * from taskappbraincategoryinfo LIMIT 0, 1000	2 row(s) returned	0.000 sec / 0.000 sec
43	10:12:53	select * from PageInfo LIMIT 0, 1000	4 row(s) returned	0.063 sec / 0.000 sec
44	10:14:57	select * from PageInfo LIMIT 0, 1000	4 row(s) returned	0.000 sec / 0.000 sec

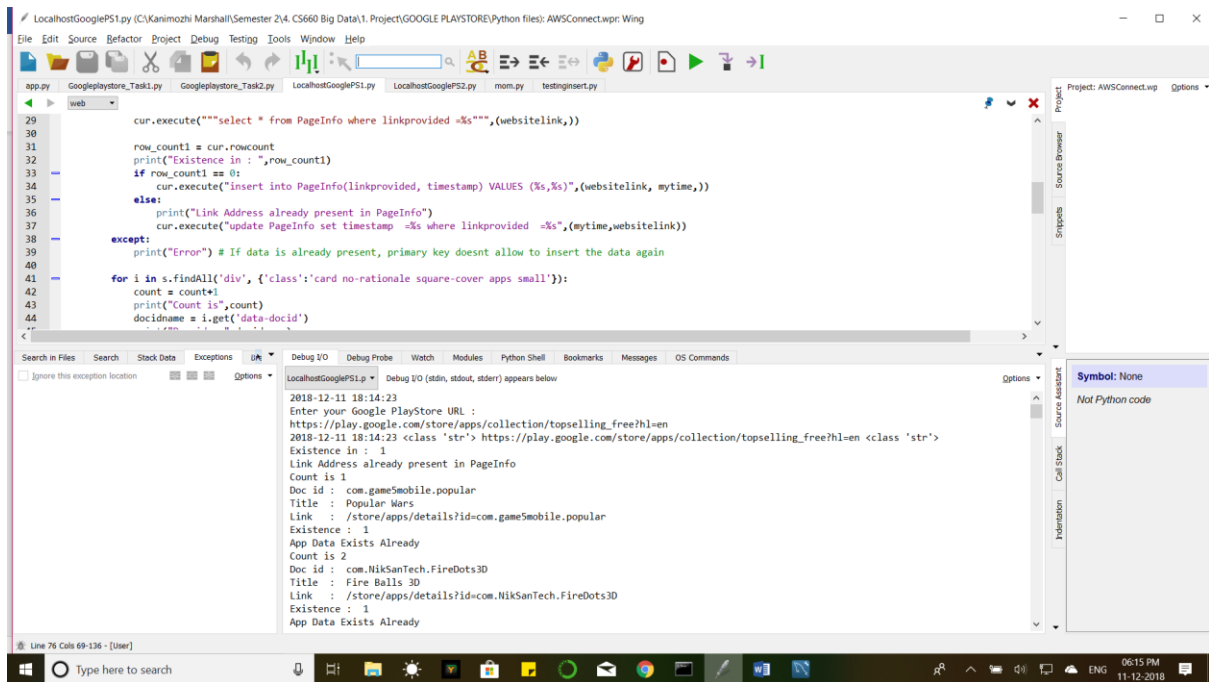
The screenshot shows the MySQL Workbench interface with the following details:

- Navigator:** Schemas list includes `playstoreinfo`.
- SQL Editor:** Contains the following SQL script:

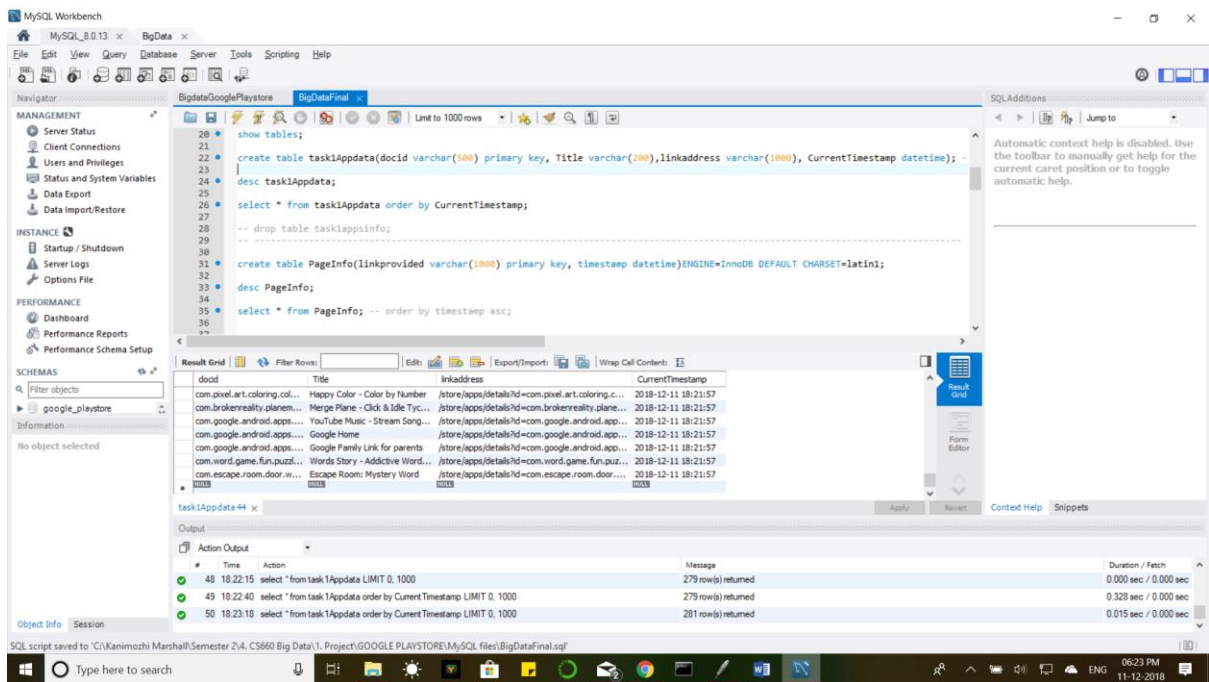
```
10 -- Task 2 : ::
11
12 -- Pagevisited gives the Category selected in Appbrain Website
13 create table taskappbraincategoryinfo(Pagevisited varchar(500), CurrentTimestamp datetime, primary key(Pagevisited))ENGINE=InnoDB DEF
14
15 select * from taskappbraincategoryinfo order by CurrentTimestamp asc;
16
17 select * from taskappbraincategoryinfo where Pagevisited = 'sefef';
18
19 desc taskappbraincategoryinfo;
20
21
22
23
24
25 create table task2appbrain(href varchar(500), Nameofapp varchar(500), Rating float, CurrentTimestamp datetime, statusflag int(10), pr
26
```
- Result Grid:** Displays the output of the `select * from taskappbraincategoryinfo` query:

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
- Output:** Shows the execution of three queries:

#	Time	Action	Message	Duration / Fetch
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

MySQL 8.0.13 x BgData x

File Edit View Query Database Server Tools Scripting Help

Navigator: BgDataGooglePlaystore BgDataFinal

MANAGEMENT

- Server Status
- Client Connections
- Users and Privileges
- Status and System Variables
- Data Export
- Data Import/Restore

INSTANCE

- Startup / Shutdown
- Server Logs
- Options File

PERFORMANCE

- Dashboard
- Performance Reports
- Performance Schema Setup

SCHEMAS

- Filter objects
- google_playstore
- Information
- No object selected

SQL Additions

Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.

```

99
100
101 create table task2appbrain(href varchar(500), NameofApp varchar(500), Rating float, CurrentTimestamp datetime, statusflag int(10), pr
102
103 insert into task2appbrain VALUES (500, '/app/yahoo-mail/jhfkj', 'Yahoo', 4.3, 'sefsef', '2018-11-01 01:43:14', 1);
104
105 select * from task2appbrain order by count;
106
107 select * from task2appbrain;
108
109 select * from task2appbrain order by Rating desc;
110
111 select * from task2appbrain order by CurrentTimestamp;
112
113 delete from task2appbrain where count = 500;
114
115 delete from task2appbrain;

```

Result Grid

href	NameofApp	Rating	CurrentTimestamp	statusflag
/app/pocket-build/incombrw/ PocketBuild	Pocket Build	4.3	2018-12-11 14:22:09	1
/app/pof-free-dating-app/com.pof.android	POF Free Dating App	4.2	2018-12-11 14:22:09	1
/app/pokfc3/%Amon-go/com.nianticlabs.pok...	Pokemon GO	4.1	2018-12-11 14:22:09	1
/app/police-chase-car-driving-simulator/com.ap...	Police Chase Car Driving Simulator	4.3	2018-12-11 14:22:09	1
/app/police-moto-bike-chase/com.yhi.police.mot...	Police Moto Bike Chase	3.8	2018-12-11 14:22:09	1
/app/police-pursuit/com.jwalee.wantedgp	Police Pursuit	4.8	2018-12-11 14:22:09	1
/app/popular-wars/com.gamestobles.popular	Popular Wars	4.1	2018-12-11 14:22:09	1
/app/postmark-buy-sell-fashion/com.postmark...	Postmark - Buy & Sell Fashion	4.3	2018-12-11 14:22:09	1

task2appbrain-49 x

Output

Action Output

#	Time	Action	Message	Duration / Fetch
53	18:52:29	select * from task2appbrain order by CurrentTimestamp LIMIT 0, 1000	984 row(s) returned	0.000 sec / 0.000 sec
54	18:52:53	select * from task2appbraincategoryinfo LIMIT 0, 1000	2 row(s) returned	0.000 sec / 0.000 sec
55	18:58:38	select * from task2appbrain order by CurrentTimestamp LIMIT 0, 1000	984 row(s) returned	0.016 sec / 0.000 sec

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: Wng

File Edit Source Refactor Project Debug Testrig Tools Window Help

app.py Googleplaystore_Task1.py localhostGooglePS2.py mom.py testinginsert.py

```

84
85 if row_count == 0:
86     statusflag=1
87     cur.execute("insert into task2appbrain(href,NameofApp,Rating,CurrentTimestamp,statusflag) VALUES (%s,%s,%s,%s,%s)",(hreflink, Appname, rating, myt
88
89 else:
90     cur.execute("select * from task2appbrain where NameofApp =%s and href =%s and Rating =%s", (Appname,hreflink,rating))
91     row_count1 = cur.rowcount
92     if row_count1 != 0:
93         statusflag=1
94         print("Data already exists")
95         print("Existence 1 : ",row_count1)
96         cur.execute("update task2appbrain set CurrentTimestamp =%s, statusflag =%s where href =%s", (mytime,statusflag,hreflink))
97     else:
98         statusflag = 1;
99         cur.execute("update task2appbrain set Rating =%s, statusflag =%s where NameofApp =%s and href =%s", (rating,statusflag,Appname,hreflink))
100         print("Rating Alone is updated")
101         db.commit()
102
103 except:
104     print('Error: ')
105 else:
106     print('Success')
107     flag = 0
108
109 for u in s.findAll('div', ('class':'list-pagination')):
110     #global n
111     for v in u.findAll('a'):

```

Search in Files Search Stack Data Exceptions

Debug I/O Debug Probe Watch Modules Python Shell Bookmarks Messages OS Commands

Ignore this exception location

LocalhostGooglePS2.p

Debug I/O (stdin, stdout, stderr) appears below

```

Existence : 1
Rating Alone is updated
Success

The Queue of pages to crawl is :
[]

```

Line 125 Col 38 - [user]

Type here to search

07:01 PM 11-12-2018

MySQL Workbench

MySQL 8.0.13 x BigData

File Edit View Query Database Server Tools Scripting Help

Navigation: BigDataGooglePlaystore BigDataFinal

MANAGEMENT: Server Status, Client Connections, Users and Privileges, Status and System Variables, Data Export, Data Import/Restore

INSTANCE: Startup / Shutdown, Server Logs, Options File

PERFORMANCE: Dashboard, Performance Reports, Performance Schema Setup

SCHEMAS: Filter objects, google_playstore, No object selected

SQL Editor:

```

99
100
101 create table task2appbrain(href varchar(500), NameofApp varchar(500), Rating float, CurrentTimestamp datetime, statusflag int(10), pr
102
103 insert into task2appbrain VALUES (500,"/app/yahoo-mail/jhfjk","Yahoo", 4.3, "sefset", "2018-11-01 01:43:14",1);
104
105 select * from task2appbrain order by count;
106
107 select * from task2appbrain;
108
109 select * from task2appbrain order by Rating desc;
110
111 select * from task2appbrain order by CurrentTimestamp;
112
113 delete from task2appbrain where count = 500;
114
115 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%A4mon-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.yh.police.mot...	Police Moto Bike Chase	3.8	2018-12-11 19:07:20	1
/app/police-pursuit/com.kwalee.wantedgo	Police Pursuit	4.8	2018-12-11 19:07:20	1
/app/popular-wars/com.game5moble.popular	Popular Wars	4.1	2018-12-11 19:07:20	1
/app/postmark-buy-sell-fashion/com.postmark...	Postmark - 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

Output:

Action Output

Time	Action	Message	Duration / Fresh
56 19:01:24	select * from task2appbrain order by CurrentTimestamp LIMIT 0, 1000	964 row(s) returned	0.000 sec / 0.000 sec
57 19:05:22	select * from task2appbrain order by CurrentTimestamp LIMIT 0, 1000	964 row(s) returned	0.000 sec / 0.000 sec
58 19:08:58	select * from task2appbrain order by CurrentTimestamp LIMIT 0, 1000	964 row(s) returned	0.141 sec / 0.000 sec

Query Completed

Type here to search

Googleplaystore_Task2.py (C:\Kinimozhi Marshall\Semester 2\4. CS660 Big Data\1. Project\GOOGLE PLAYSTORE\Python files: AWSConnect\wpr: Wing

File Edit Source Refactor Project Debug Testing Tools Window Help

app.py Googleplaystore_Task1.py Googleplaystore_Task2.py LocalhostGoogleFS1.py LocalhostGoogleFS2.py mom.py testinginsert.py

Project: AWSConnect.wpr Options

Source Browser: Snippets

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=" ", 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: a string separator to use between values, defaults to ' '.

```

94
95 # Check if data already exists if yes - check rating value
96 cur.execute("""select * from task2appbrain where NameofApp =%s and href =%s""",(Appname,hreflink))
97 row_count = cur.rowcount
98 print("Existence : ",row_count)
99 if row_count == 0:
100     statusflag=1
101     cur.execute("insert into task2appbrain(href,NameofApp,Rating,CurrentTimestamp,statusflag) VALUES (%s,%s,%s,%s,%s)",(hreflink, Appname, rating, myt:
102
103 else:
104     # Check if Rating is also same for the Existing App if yes just update timestamp, if no, Update the new Rating
105     cur.execute("""select * from task2appbrain where NameofApp =%s and href =%s and Rating =%s""",(Appname,hreflink,rating))
106     row_count1 = cur.rowcount
107     if row_count1 != 0:
108         #statusflag=1
109         print("Data already exists")
110         print("Existence 1 : ",row_count1)
111         cur.execute("update task2appbrain set CurrentTimestamp =%s, statusflag =%s where href =%s",(mytime,statusflag,hreflink))
112     else:
113         statusflag = 1;
114         cur.execute("update task2appbrain set Rating =%s, CurrentTimestamp =%s, statusflag =%s where NameofApp =%s and href =%s", (rating,mytime,s
115         print("Rating Alone is updated")
116
117 db.commit()
118
119 except:
120     print('Error: ')
121 else:
122     print('Success')

```

Search in Files Search Stack Data Exceptions UN

Debug I/O Debug Probe Watch Modules Python Shell Bookmarks Messages OS Commands

Ignore this exception location Options

LocalhostGoogleFS2.p Debug I/O (stdin, stdout, stderr) appears below

Existence : 1
Rating Alone is updated
Success

The Queue of pages to crawl is :
[]

Line 113 Col 28 - [User]

Type here to search