Data Mining Project assigment1

Shridhar Manvi, 1001096921

Design and Implementation:

1. The program imports each of the data files which are tab separated and stores them into dictionaries with the key of document being the Primary key of the imported dataset

EX: users.tsv is imported in users{} dictionary

key of users{} is the primary key of users.tsv

Note: For apps dictionary, dual key is followed since the apps.tsv file has (User_ID, Job_ID) as a composite key

Same process is followed for every other data file.

2. Once all the files are imported, dimension tables like location, jobs_dim etc. are built which will be used further in the program during rollup and drill operations in task2

3. There are two functions which perform each of the tasks provided in the question. The first task function accepts two variables users and apps file and constructs of cuboid by joining the two datasets on state_id and jobID keys.

firsttask(arg1,arg2) performs the first task and creates a cuboid to be used in the second step and computes the most applied job in each state. It also prints the top 5 jobs by state

secondtask(arg1) is passed with the cuboid created above and performs OLAP operations to get the final output in second task. It prints the top five most popular job title applied for in a given country

4. All operations have been performed by using inbuilt data types such as lists and dictionaries. All OLAP operations are performed by using looping mechanism. No 3rd party libraries are used.

5. The program accepts 5 arguments which are used in the program. The first one being the country to be used in slicing operation in second task. The second through 5th arguments are files used in the program.

6. The first task method creates a joined dataset called user_apps joined on user_id key from both the dictionaries (users and apps). user_apps has a composite primary key (user_id,job_id) since the combination is unique.

7. Based on the user_apps dataset obtained above, it counts the number of occurrences of each job Id for each state.

8. Top 5 most applied jobs are picked from the list and printed

9. The cuboid obtained in first task which has a list of [StateID,JobID, count] is passed to the second task method which then uses it to perform OLAP operations to obtain the most popular title in a given country.

10. The second task method slices the data in the cuboid for Country = Country entered through argument by user

11. After slicing, the data is rolled up from JobID to Title by joining the cuboid with Jobs data set using JobID

12. Next, the data is rolled up from StateID to CountryID and the top 5 most popular Job titles are presented on the output screen

How to execute the program:

1. The program requires Python 6 or later version to run.

2. Save the python file and the data sets in the same folder.

3.Open the terminal/command prompt and and execute the following command

python datamining.py US apps.tsv users.tsv jobs.tsv user_history.tsv

4. Navigate to the directory where the code and data sets are stored

5. The command accepts 5 arguments in the following order:

a. Country name: The country name has to be the same as what is present in the datasets. Ex: America is US, India is IN and so on

b. 2nd argument is the apps.tsv file

c. 3rd argument is the users.tsv file

d. 4th argument is the jobs.tsv file

e. 5th argument is the user_history.tsv file

DM project1

Execution screen shots of the program:

Full screenshot

```
shridhars-MacBook-Pro:tryagain shridharmanvi$ python datamining.py US apps.tsv users.tsv jobs.tsv user_history.tsv
StateID JobID           NumbOfApps
IL      601021          158
TX      98665           142
TX      187358          135
TX      10312           133
TX      741664          126



CountryID       JobID           NumbOfApps
US      Administrative Assistant        6123
US      Customer Service Representative 4626
US      Receptionist    2958
US      Executive Assistant     1924
US      Customer Service        1179
```