Project Milestone 2

Shubham Garg, 2010ce10401
Sumeet Kumar Sinha, 2010ce10405
Tarun Kumar Singhal, 2010ce10411

Due date: Febraury 25, 2013, 11:55pm IST

1 Sources for Our data

- The data has been scrapped using Google APIs, online console by querying data for each city.
- The database of Hotels we have got online from a website with url= http://link.smartscreen.live.com/info/?a=infol=aHR0cDovL2FwaS5ob3RlbHNiYXNlLm9yZy9hcGlBY2Nlc3MucGhwP2VtYWlsPXNodWJoYW0xNDM2JTQwZ21haWwuY29tJnVzYWdlPWNyZWF0aW5nK2ErZGF0YWJhc2UrZm9yK2NvbGxlZ2UrcHJvamVjdA4807274365h=OhS
- But we would be using scrapping techniques for the hotels
- The link for the GoogleAPI is https://code.google.com/apis/console/

2 How We downloaded it (Readymade vs scraped)

- We have taken list of cities from Wikipedia
- Then have arranged in order and are then using programs written in curl php for data scrapping
- From the scrip we got a xml file which we converted into csv format which is ready to be inserted into the database.
- Here is an Example of the used curl php script

3 Code

¡?php
$query = array("Babaleshwar","Bangalore","Babiyal","Baddi","Badaun","Badagaon","Badepalle",...Burhanpur","Burla","Buxar","Byasanagar");
$proxy = 'proxy22.iitd.ernet.in : 3128';
fp = fopen("data.txt", "a");
$ch = curl_init();

foreach ($queryasvalue)
$url = 'https://maps.googleapis.com/maps/api/place/textsearch/xml?query = places+in+'.value.'sensor =truekey=AIzaSyCplAcR_aXPXqUlBnh3Fs7AzySTaun75 − I';
curl_setopt(ch, CURLOPT_URL,url);
curl_setopt(ch, CURLOPT_PROXY,proxy);
curl_setopt(ch, CURLOPT_FILE,fp);
curl_setopt(ch, CURLOPT_HEADER, 0);
curl_exec(ch);
curl_close(ch);

fclose($fp);

?¿

4 CLeanup Steps

- The data we have used is scrapped so in the process we have removed some redundancy.
- Statistics (Include a table which lists the name of the table, number of tuples in the table, time to load, size of raw dataset, size of raw dataset *after* cleanup.)

Table Name	No. of Tuples	Time to Load	Size of Raw Dataset	Size of raw database
city	619	523ms	3.4mb	132mb
picTable	11587	6234ms	3.4mb	132mb
hotel	13277	26454ms	3.4mb	132mb
touristAttractions	4563	3452ms	1.65mb	*Scraping from API
hotelType	40234	6327ms	1269kb	132mb
touristAttractionType	12598	1121ms	270kb	*Scraping from API
touristAttractionPictures	1261	1322ms	313kb	*Scraping from API

- Data for USER,TRIPHISTORY would be added accordingly after the completion of frontend.

5 Location of the raw dataset and the cleaned up dataset

- As the data has been scrapped so the raw as well as cleaned up dataset in on my Laptop only
- We are scrapping data from Google API @around 3000 raw tuples per day, so we having would be having around sufficient data till final submission
- We have done manual cleanup for dataset. Rest query has been defined in SQL file for loading the database.

6 Modified $E_R Diagram$