# Data Integration: Linking tourist events and attractions data to hotels and airlines information.

**Abstract**: I propose integrating data from viator.com, getyourguide.com, booking.com, kayak.com and https://www.tourradar.com/deals and giving customers one portal to view all travel related data.

Travelers who go on personal trips look for tourist attractions and special events that may occur during the time they wish to travel. Today a traveler books airlines tickets in one portal, after some research books a hotel. The job of planning vacation is not done STILL. He/she spends hours to find suitable places of interest.

I propose a solution to give customers a consolidated view of all information in these. Together they will save lots of time for travel planners.

When used together, they will be our recommendations to tourists.

A future proposal could be to get some information from users (emailed or ask them to create a short profile) and aim to send marketing emails to the users.

# 4 step process

The project will consist of 4 steps

1. **Data collection** – via apis and web scrapping from the above identified websites. Sample data from the portals are shared in this document.

**Tasks:**

1)identify websites to collect data

2)identify the types of data to collected

3)check if the data is extractable.

4)Document data to be extracted from each of the website.

5)Develop web scrapers using python and extract data from websites.

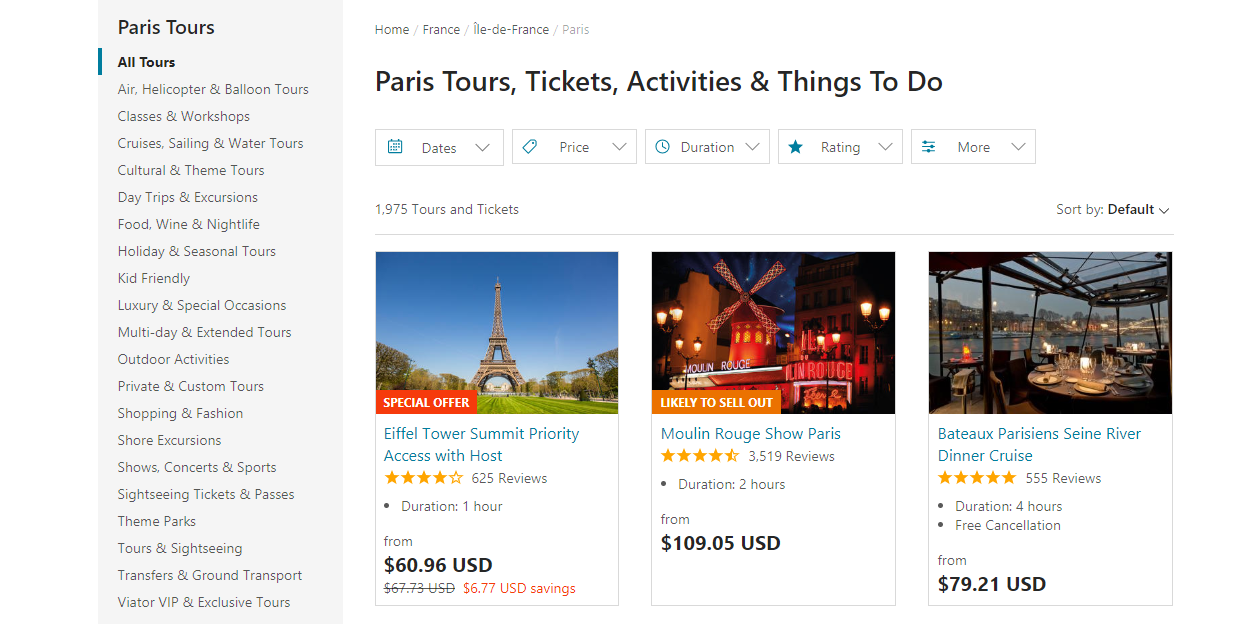
6)Using python client extract data using APIs where available.

7)Design a staging area in MySQL to host the raw response data.

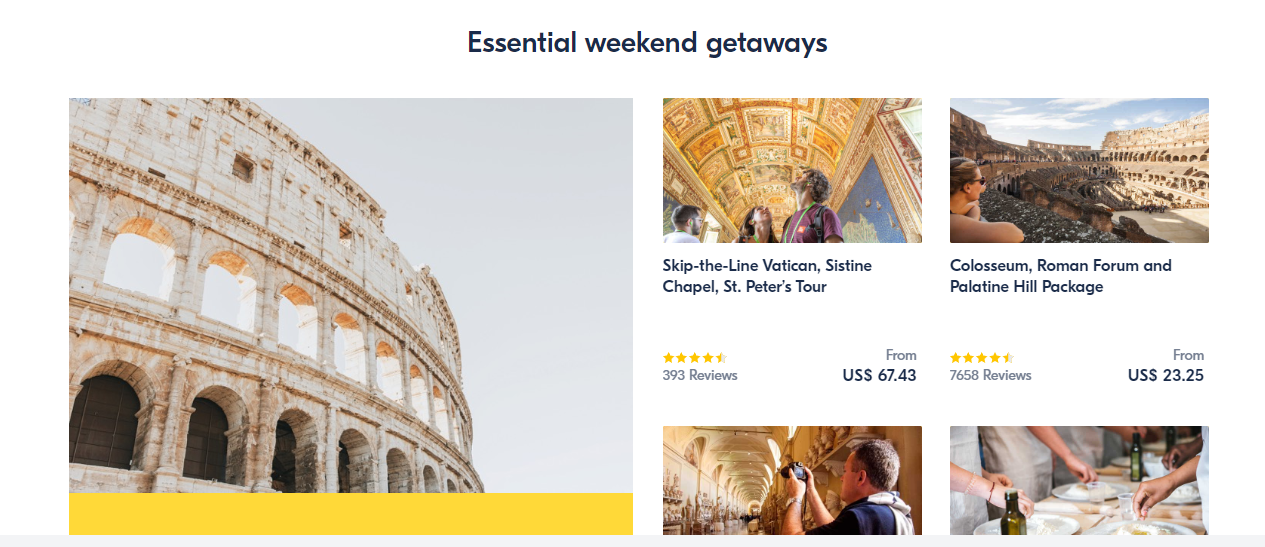
|  |  |
| --- | --- |
| Website | Method to extract data |
| viator.com | Web scrapping |
| getyourguide.com | Web scrapping (<https://stackoverflow.com/questions/33035202/python-beautiful-soup-crawling-with-json?rq=1>) |
| booking.com | API (<https://github.com/ZoranPandovski/BookingScraper>) |
| expedia | Web scraping (<https://www.scrapehero.com/scrape-flight-schedules-and-prices-from-expedia/>) |
| Tripadvisor | <https://stackoverflow.com/questions/50422110/basic-python-beautifulsoup-web-scraping-tripadvisor-reviews-and-data-cleaning>  <https://github.com/tomvannuenen/scrape-tripadvisor>  <https://www.scrapehero.com/how-to-scrape-tripadvisor/> |

Samples from websites:

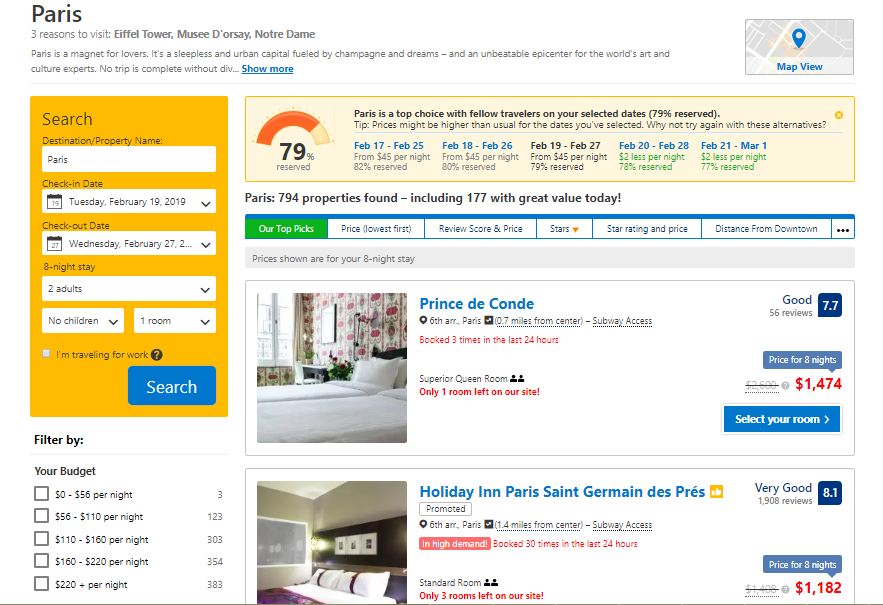
**viator.com**



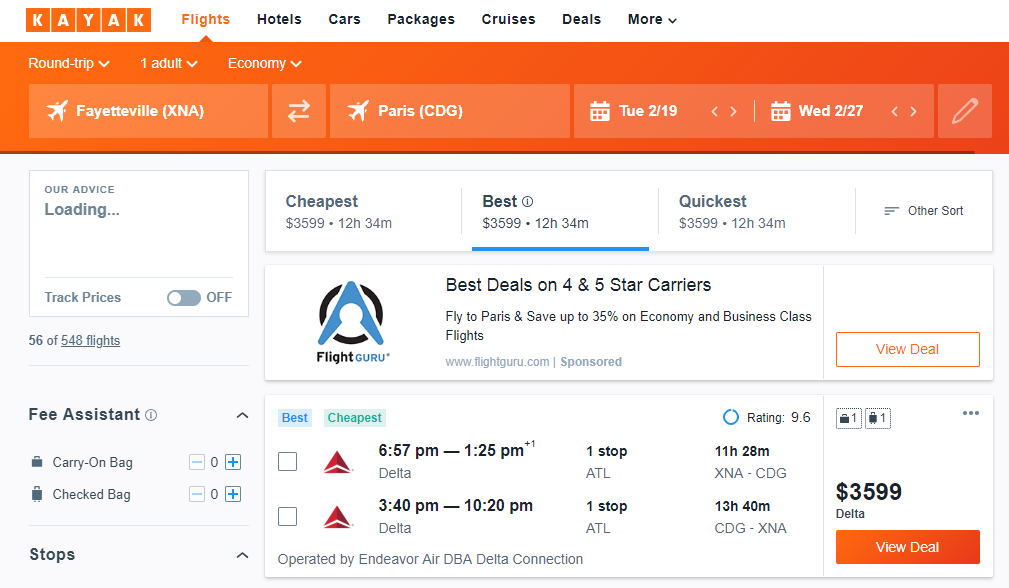
**getyourguide.com**



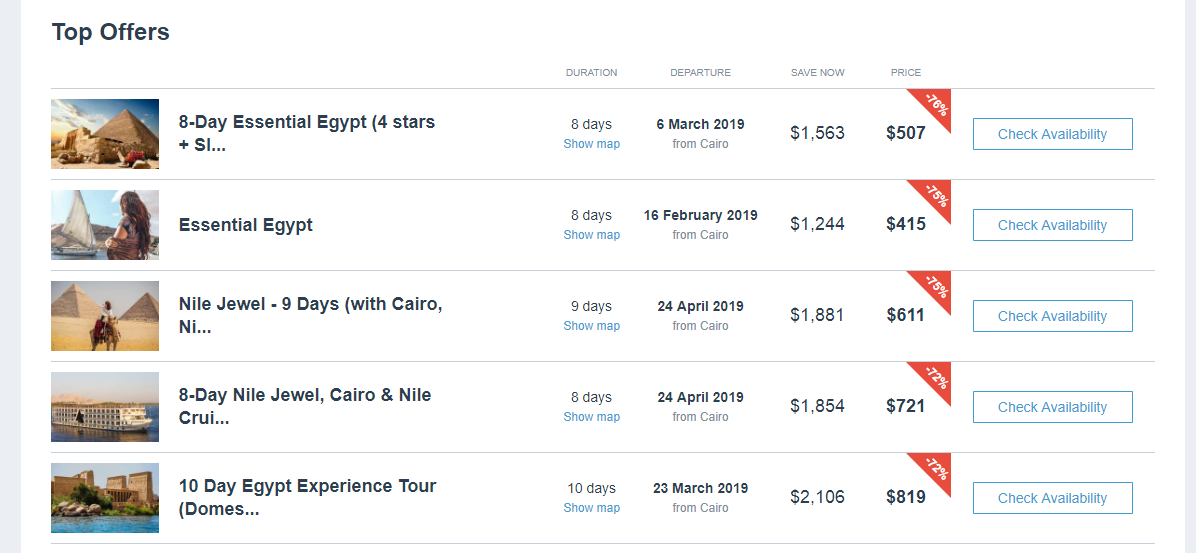
**booking.com**



**Kayak.com**



<https://www.tourradar.com/deals>



1. **Data aggregation and ETL for consumption**. Load data into an RDBMS for consumption

Design a warehouse for ETL process. Details will be worked out in coordination with ISYS 5833 (Professor Jeff Mullin’s class)

**Tasks:** <<warehousing tasks to be identified>>, design an ERD for RDBMS database.

1. **Build web portal which will read data made available in step2**.

Tech stack to use:

Mysql (db),

Python and React

Angular

**Tasks :** create wireframes, queries, web pages, test plan

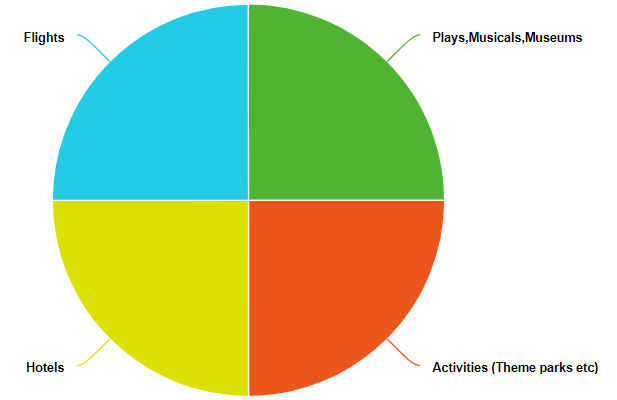
1. **Build services (rest/webservices) making the data from step2. Planning to use python to develop the webservice.**

**Tasks :** identify data/tables which will be exposed. Design the request, response, errors, tests and documentation for the services.

# Why Integrate These Datasets?

**Problem and challenges for vacation planning – today’s scenario.**

Vacation planning may be divided into 4 tasks (shown as 4 quadrants in the pie chart below). Today the problem is traveler might finish 50% of his planning (flights and hotel) with some effort but to complete the rest 50% (things to do) he has to rely on google or if luck friends! Plus this takes a lot of time.





I want to consolidate information from the 5 websites. By integrating the datasets and applying Information Integration techniques I wish to give travelers information they need in one place.

**Common data storage strategy:** I plan to store information parsed from all data sources into a local database. This consolidated data will be then opened up for use via portal/ other purposes.

# Challenges.

1)Not all websites provide a free API access to extract data. Need to do web scrapping.

2)Knowledge of web scrapping and ETL.

3)Knowledge of web development using python.

Below are the websites, url pattern, sample data to be extracted and use of x-path and div tags to extract from the result of scrapping.

|  |
| --- |
| **Website:** <https://www.viator.com>  **Method**:webscrapping |
| **Tickets and activities**  **URL** <https://www.viator.com/Paris/d479-ttd> |
|  |
|  |
| Data to be collected for this:   * Image * Writeup * Title of the * Rating n reviews * Duration & other information like free cancellation etc. |
| **Attractions**  URL: https://www.viator.com/Paris-attractions/Moulin-Rouge/d479-a576 |
|  |
| Data to be collected for this:   * Image * Writeup * Title of the * Rating n reviews * Duration & other information like free cancellation etc. * From Price and currency. |

|  |
| --- |
| **Website:** https://www. getyourguide.com  **Method**:webscrapping |
| **Search url**  **URL** https://www.getyourguide.com/s/?q=Paris&date\_from=2019-06-05&date\_to=2019-06-19&customerSearch=1&searchSource=2 |
|  |
|  |
| Data to be collected for this:   * Image * Writeup * Title of the * Rating n reviews * Duration & other information like free cancellation,validity etc. * Price and currency |

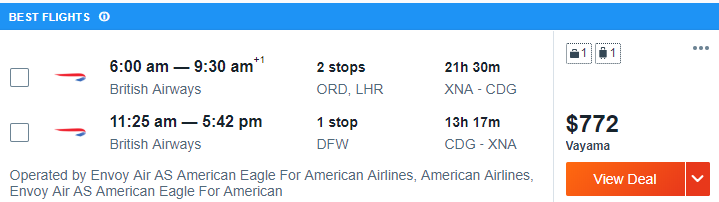
|  |
| --- |
| **Website:** https://www. booking.com  **Method**:webscrapping |
| **Search url**  **URL** https://www.booking.com/searchresults.en-us.html?label=gen173nr-1DCAEoggI46AdIM1gEaKkCiAEBmAExuAEIyAEM2AED6AEB-AECiAIBqAIDuAKrn-7jBQ&sid=cc3637ae1881963604ac8ff064e1ee65&sb=1&src=index&src\_elem=sb&error\_url=https%3A%2F%2Fwww.booking.com%2Findex.html%3Flabel%3Dgen173nr-1DCAEoggI46AdIM1gEaKkCiAEBmAExuAEIyAEM2AED6AEB-AECiAIBqAIDuAKrn-7jBQ%3Bsid%3Dcc3637ae1881963604ac8ff064e1ee65%3Bsb\_price\_type%3Dtotal%26%3B&ss=paris&is\_ski\_area=0&ssne=Apia&ssne\_untouched=Apia&dest\_id=-1307163&dest\_type=city&checkin\_year=2019&checkin\_month=4&checkin\_monthday=24&checkout\_year=2019&checkout\_month=4&checkout\_monthday=25&no\_rooms=1&group\_adults=2&group\_children=0&b\_h4u\_keep\_filters=&from\_sf=1 |
|  |
|  |
| Data to be collected for this:   * Image * ~~Writeup~~ * Address and miles from X info * name of the hotel * number of rooms * Rating n reviews * Duration & other information like free cancellation,validity etc. * Price and currency (including original price) |

Website: Kayak.com.

Type API.

Tool-kit to use:

<https://github.com/yasulab/kayak-api-lib4py>



Data to be collected for this:

* From and to
* airlines
* from and to time
* Number of stops
* Hours of flight
* Other information.
* Price and currency

|  |
| --- |
| **Website:** https://www. tourradar.com  **Method**:webscrapping |
| **Search url**  **URL** https://www.tourradar.com/deals/europe#month=march-2019 |
|  |
|  |
| Data to be collected for this:   * Image * title * duration * price per day * Rating n reviews * Destinations * Age range * Accomodation * Transport * Travel style * operator * other information. * Price and currency |

**Components and technical architecture:**

1)use scrapy for webscrapping. Scrapy returns in json format. (https://docs.scrapy.org/en/latest/intro/tutorial.html)

2)store the data into couchdb. (https://www.tutorialspoint.com/couchdb/)

3)ETL picks from here and gets the cleaned,aggregated data.

4)Use the flask,react,gulp, bower and Browserify to develop portal. (https://realpython.com/the-ultimate-flask-front-end/)

5)Use flask to develop services to expose the data as services. (https://blog.miguelgrinberg.com/post/designing-a-restful-api-with-python-and-flask)