Hotel TULIP (a hypothetical organisation)

Sarah Sun

Deakin University

(None)

Overview

Background

Hotel TULIP (a hypothetical organisation)

Defn

Special purpose: Not only does it embody all the creative energy.

Spirit of TULIP-Lab, it's a "learning environment" on which the tourism and hospitality students are trained for future hoteliers.

- a hypothetical organisations
- A five star hotel that locates in Australia.

Analysis purposes

purposes

Improve their potential customers' online experience,and help their Market Promotion Division toidentify potential customers and their behaviour patterns

- In the past two decades.
- the Web server of Hotel TULIP has logged all the web traffic to the hotel website.
- Stored large amount of data related to the use of various web pages.

ETL

Extraction

- Existing Methods - Unzip and read all the network log files
- Count file number-120

```
Hotel: ex061126.log
Hotel: ex061127.log
Hotel: ex061128.log
Hotel: ex061129.log
Hotel: ex061130.log
Hotel: ex061201.log
Hotel: ex061202.log
Hotel: ex061203.log
Hotel: ex061204.log
Hotel: ex061205.log
Hotel: ex061206.log
Hotel: ex061207.log
Hotel: ex061208.log
Hotel: ex061209.log
Hotel: ex061210.log
Hotel: ex061211.log
Hotel: ex061212.log
Hotel: ex061213.log
Hotel: ex061214.log
```

Figure 1: loading

```
This zip have 120 documents.
```

Figure 2: numbers

Loading data

- Print the first five lines to view the data in general form

	date	time	s-sitename	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	cs-username	c-ip
0	2006-11-01	00:00:08	W3SVC1	127.0.0.1	GET	/Default.aspx	-	80.0	-	70.80.84.76
1	2006-11-01	00:00:08	W3SVC1	127.0.0.1	GET	/Tulip/home/en-us/home_index.aspx	-	80.0	-	70.80.84.76
2	2006-11-01	00:00:08	W3SVC1	127.0.0.1	GET	/Tulip/includes/js/CommonUtil.js	-	80.0	-	70.80.84.76
3	2006-11-01	00:00:09	W3SVC1	127.0.0.1	GET	/Tulip/common/common_style.aspx	lang=en-us	80.0	-	70.80.84.76
4	2006-11-01	00:00:09	W3SVC1	127.0.0.1	GET	/Tulip/common/en-us/images/top_img.jpg	-	80.0	-	70.80.84.76

Figure 3: head

c-ip	cs(User-Agent)	cs(Referer)	sc-status	sc-substatus	sc-win32-status
70.80.84.76	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...	http://www.google.com/search?sourceid=navclien...	200.0	0.0	0.0
70.80.84.76	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...	-	200.0	0.0	0.0
70.80.84.76	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...	http://www.hotelTulip.com.hk/Tulip/home/en-us/...	200.0	0.0	0.0
70.80.84.76	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...	http://www.hotelTulip.com.hk/Tulip/home/en-us/...	200.0	0.0	0.0
70.80.84.76	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...	http://www.hotelTulip.com.hk/Tulip/home/en-us/...	200.0	0.0	0.0

Figure 4: head

Data attribute

Attribute Name	Data Type	Data Subtype	Description	Examples
date	MC	DATE - Date/time	Date of user access	1/11/2006
time	MC	DATE - Date/time	User access specific time	0:00:08
s-sitename	CN	STR	Name of the website visited by the user	W3SVC1
s-ip	CN	ID	It is the foundation that forms the basis of the Internet and provides information on various protocols to the transport layer	127 0 0 1
cs-method	CN	STR	Methods for obtaining data	GET
cs-uri-stem	CN	ADDR	The file or resource that the user is requesting access to when accessing the site	/Default aspx
s-port	MD	BIN	Like a door number, the client can find the corresponding server side by its ip address, but there are many ports on the server side and each application corresponds to a port number. In order to differentiate the ports, each port is numbered. The ports of an IP address are numbered by 16bit, and there can be up to 65536 ports . Ports are marked by port numbers, which are integers only and range from 0 to 65535	80
c-ip	CN	ID	It is the foundation that forms the basis of the Internet and provides information on various protocols to the transport layer	70 80 84 76
cs(User-Agent)	CN	URL	A user-agent can be a web crawler, a download manager or other application that can access the web. With each request sent to the server, the browser contains a User-Agent HTTP protocol header that identifies itself	Mozilla/4.0+(compatible;+MSIE+6.0;+ Windows+NT+5.1;+SV1)
sc-status	CN	ID	Server successfully returns web page	200
sc-substatus	MD	BIN	The number of specific items that caused the subtask to fail.	0
sc-win32-status	MD	BIN	IIS server status code: successfully completed	0

Figure 5: attribute

Data cleaning

- Calculates NaN for rows and columns in the data
- After cleaning up more than 15 percent of the columns, remove the rows with NaN

```
date              0.000024
time              0.000024
s-sitename        0.000024
s-ip              0.000024
cs-method         0.000024
cs-uri-stem       0.000024
cs-uri-query     93.454194
s-port            0.000024
cs-username     100.000000
c-ip              0.000024
cs(User-Agent)    0.041818
cs(Referer)      15.519242
sc-status         0.008982
sc-substatus      0.008982
sc-win32-status   0.008982
dtype: float64
```

Figure 6: Cleaning >=15% data

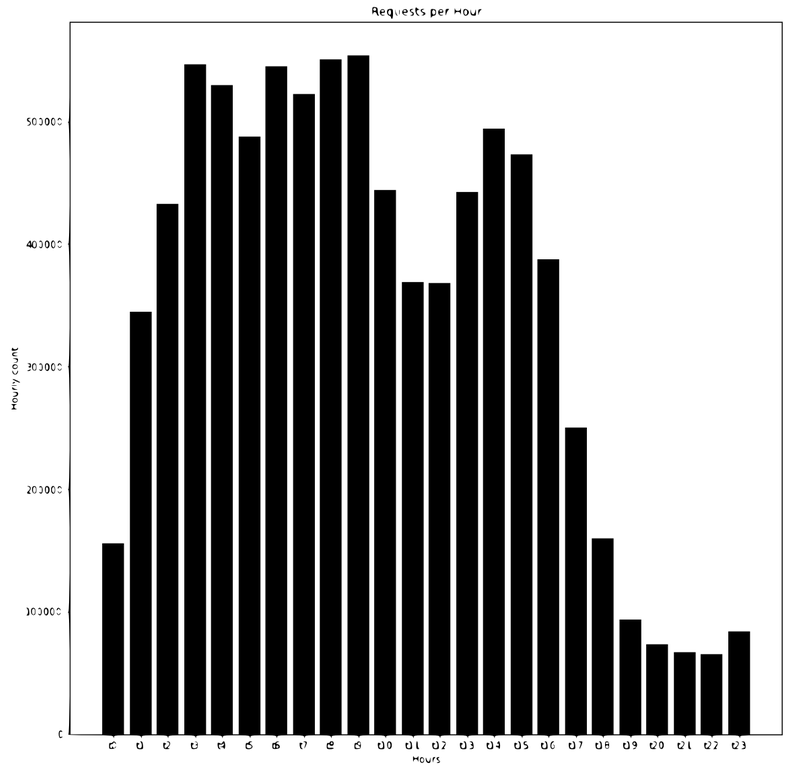
```
63742  74.110.32.161  Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...
63743  74.110.32.161  Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...
63744  74.110.32.161  Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...
63745  74.110.32.161  Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+...

       sc-status  sc-substatus  sc-win32-status
0          200.0           0.0              0.0
1          200.0           0.0              0.0
2          200.0           0.0              0.0
3          200.0           0.0              0.0
4          200.0           0.0              0.0
...          ...           ...              ...
63741      200.0           0.0              0.0
63742      200.0           0.0              0.0
63743      200.0           0.0              0.0
63744      200.0           0.0              0.0
63745      200.0           0.0              0.0

[8434645 rows x 12 columns]
```

Figure 7: Finished cleaning 8434645 rows

Data Statistics Description

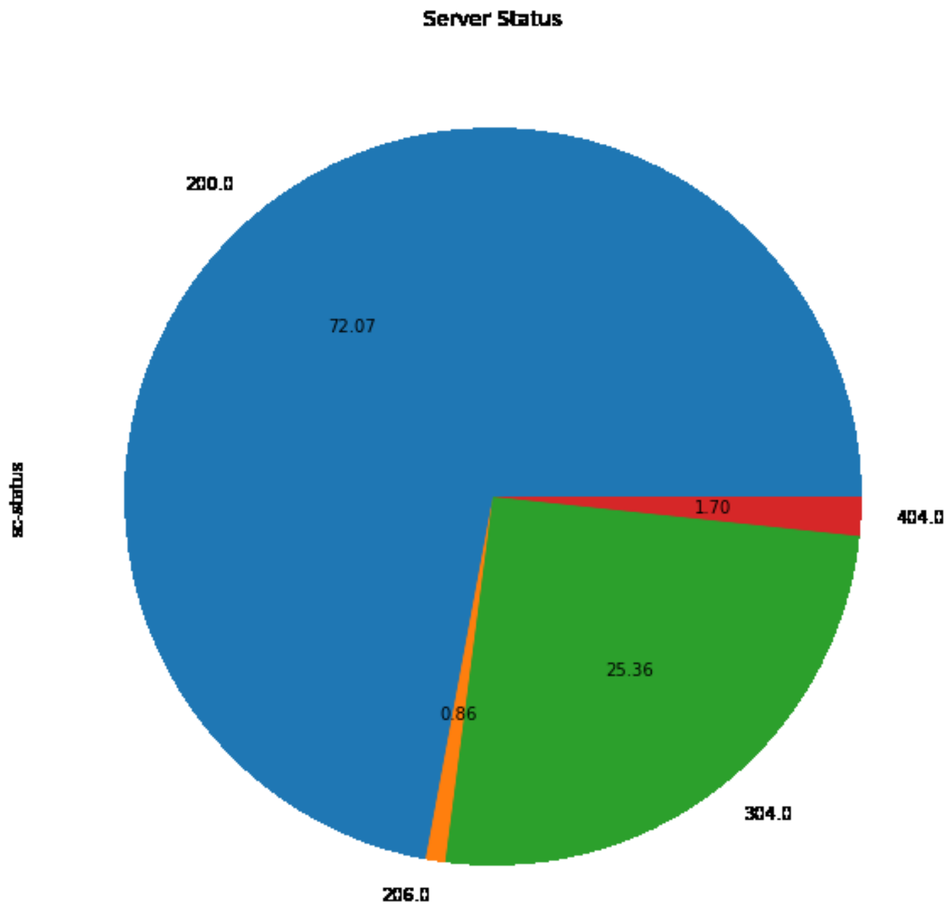
Step One - Traffic Analysis

- All data from 2006 to 2007 are broken down by hour, from 0:00 to 23:00, with traffic analysis for each hour.



(a) traffic analysis

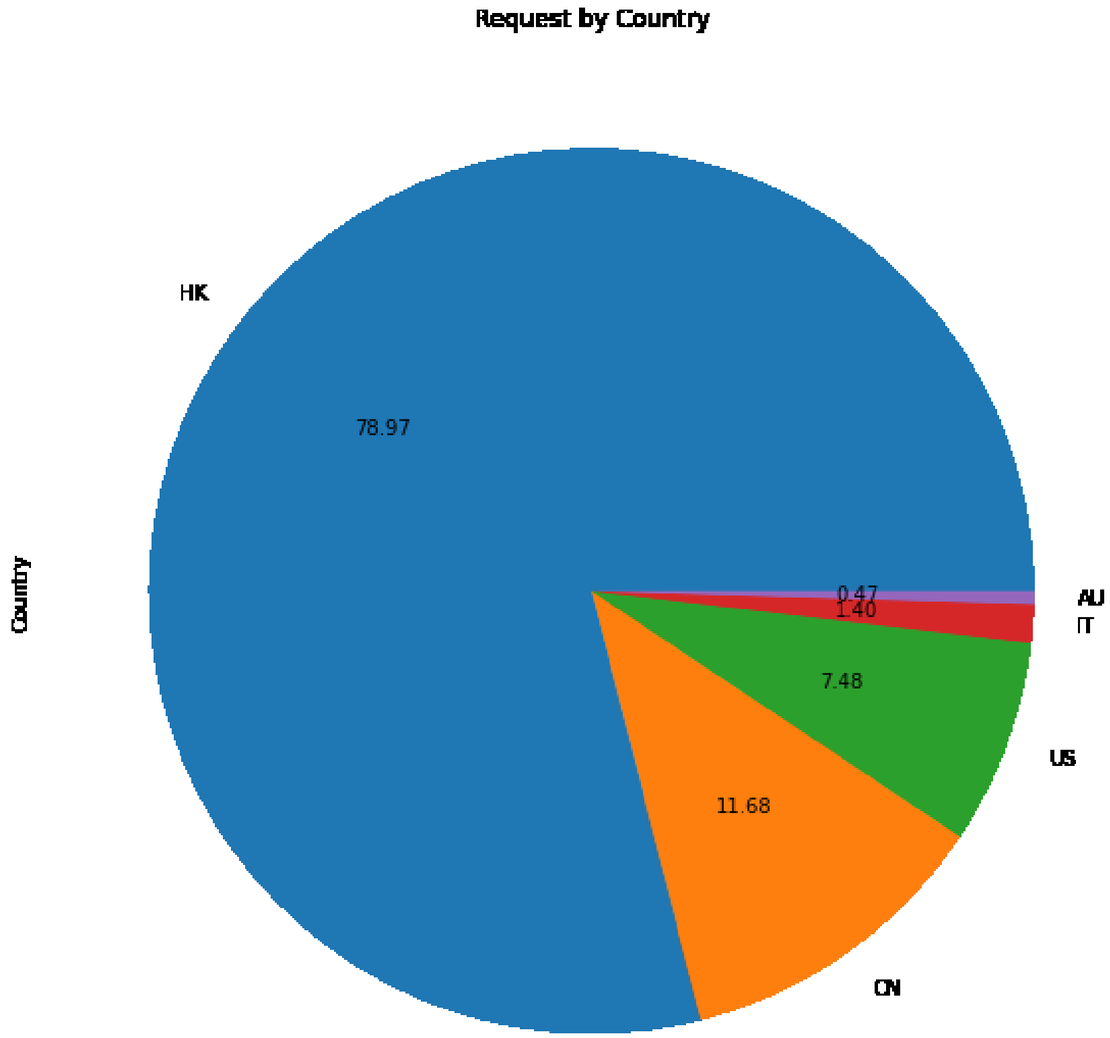
Step two: Server Analysis



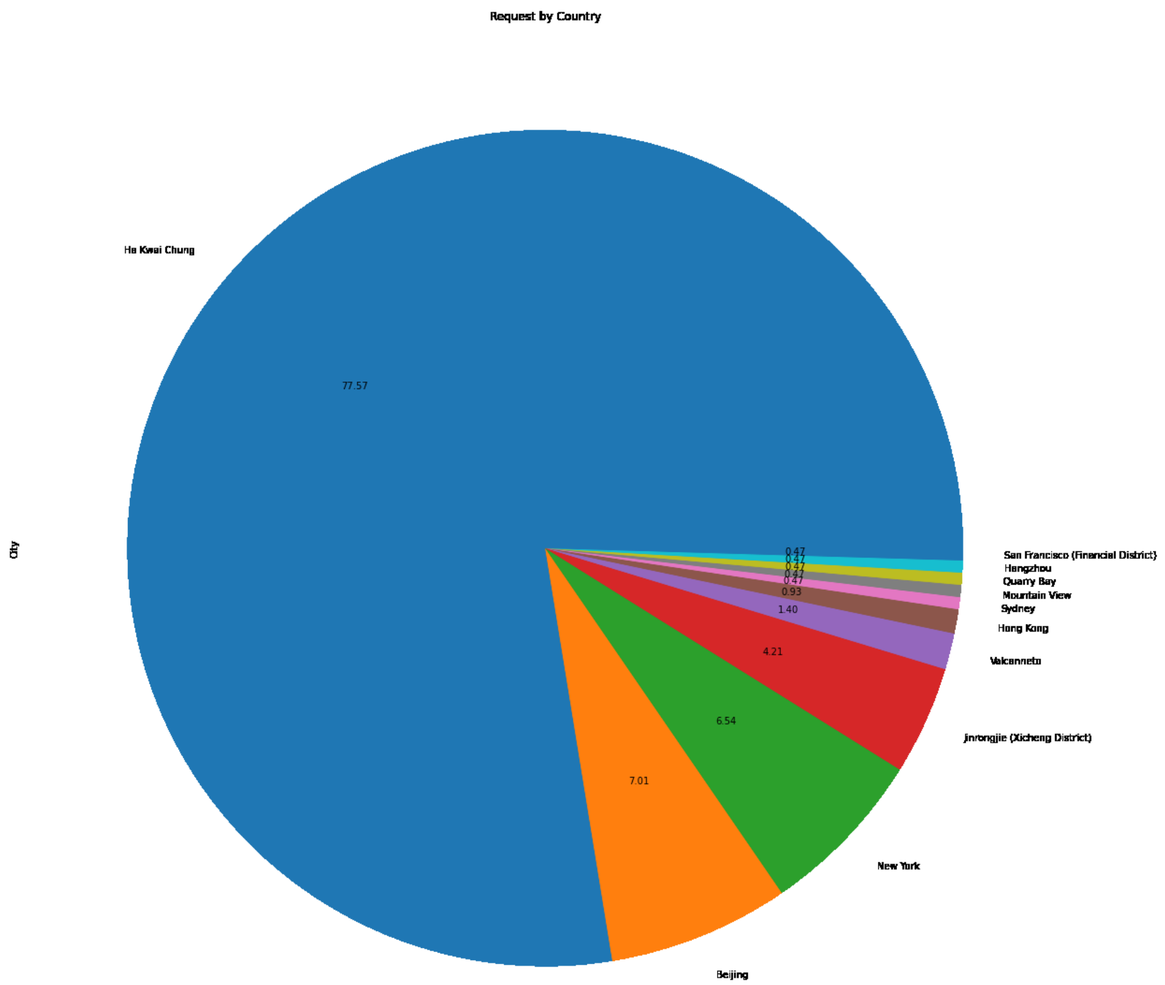
(b) traffic analysis

Geographic Analysis

Country



City



(d) traffic analysis

Conclusion

Conclusion

- The overall idea: decompress the dataset, load it, read it. And clean up any unnecessary data in it, such as nan, - or spaces.
- Three aspects of the data were visualised for the data.

  By using bar charts and pie charts to analyse when web traffic is concentrated and the corresponding situation on the website, and finally by city and country, to compare the countries and cities with the highest number of visitors.

Questions?

Contact Information

very very Fresh Fish: Sarah Sun

Deakin University, Australia

SUNHAN@DEAKIN.EDU.AU

GROUP 00