

# Hotel TULIP Web Server Data Analysis

**Assignment 2 - SIT742 Modern Data Science**

Alex Cummaudo <ca@deakin.edu.au>

Jake Renzella <jake.renzella@deakin.edu.au>

Deakin Software and Technology Innovation Laboratory

School of Information Technology

Deakin University, Australia

May 3, 2017

## Executive Summary

This report summarises findings from a data exploration on the Hotel TULIP web server logs, recorded between the periods of August 2014 and August 2015. Each log contains one *request*, or *hit*, that lists fourteen attributes as described in the attached Data Dictionary spreadsheet. Publicly known client IP addresses were extracted from the MaxMind GeoIP2<sup>1</sup> dataset to analyse the location of requests (narrowed down to city). Additionally, user agent strings were parsed to analyse device and browser statistics using the Python user-agents library<sup>2</sup>, thereby extrapolating demographics, usage trends, platform information, server performance, and security statistics from the raw logs provided in the dataset. Further details on the extraction of the data is provided in the source code attached in Appendix B, and an interactive version of this file is published on Databricks.

---

<sup>1</sup>See <http://dev.maxmind.com/geoip/geoip2/>.

<sup>2</sup>See <https://pypi.python.org/pypi/user-agents>.

# Contents

<b>1</b>	<b>Key Findings</b>	<b>4</b>
<b>2</b>	<b>Introduction</b>	<b>5</b>
<b>3</b>	<b>Dataset</b>	<b>5</b>
<b>4</b>	<b>Method</b>	<b>5</b>
4.1	Assumptions Made . . . . .	5
4.2	Extraction Process . . . . .	5
4.2.1	Internal/External IP Address Regular Expression . . . . .	5
4.3	Data Mining . . . . .	6
<b>5</b>	<b>Results</b>	<b>6</b>
5.1	Patterns Identified . . . . .	6
5.1.1	Internal Site Visitors . . . . .	6
5.1.2	External Site Visitors . . . . .	6
5.1.3	Hong Kong Visitors . . . . .	6
5.1.4	USA Visitors . . . . .	6
5.1.5	Australian Visitors . . . . .	6
5.1.6	PC Visitors . . . . .	6
5.1.7	Smartphone Visitors . . . . .	6
5.1.8	Tablet Visitors . . . . .	6
5.1.9	Bots Visitors . . . . .	6
<b>A</b>	<b>Additional Tables</b>	<b>7</b>
<b>B</b>	<b>Extrapolation Results</b>	<b>12</b>

## List of Figures

# 1 Key Findings

A list of key findings in the analysis are as thus:

- Foo

## **2 Introduction**

## **3 Dataset**

## **4 Method**

### **4.1 Assumptions Made**

### **4.2 Extraction Process**

#### **4.2.1 Internal/External IP Address Regular Expression**

To differentiate between private site visitors and external visitors, a regular expression was used to filter private/public IP address ranges. The regular expression is shown below:

```
c_ip regexp
      ' (^127\.)|(^10\.) |
      (^172\.1[6-9]\.) |
      (^172\.2[0-9]\.) |
      (^172\.3[0-1]\.) |
      (^192\.168\.) '
```

Using the not of the regular expression will select only public IP addresses.

## **4.3 Data Mining**

# **5 Results**

## **5.1 Patterns Identified**

### **5.1.1 Internal Site Visitors**

### **5.1.2 External Site Visitors**

### **5.1.3 Hong Kong Visitors**

### **5.1.4 USA Visitors**

### **5.1.5 Australian Visitors**

### **5.1.6 PC Visitors**

### **5.1.7 Smartphone Visitors**

### **5.1.8 Tablet Visitors**

### **5.1.9 Bots Visitors**

## A Additional Tables

Below are tables of frequency pattern results for each section identified in Section 5.

Table 1: Internal Request Frequency Patterns

From	To	Frequency
rooms	offers	370
dining	offers	333
about the hotel	offers	330
about the hotel	rooms	262
about the hotel	dining	248
facilities	offers	220
above and beyond	dining	208
facilities	rooms	205
facilities	about the hotel	202
home	offers	201
location and contacts	offers	189
rooms	dining	187
above and beyond	rooms	185
above and beyond	offers	180
facilities	dining	177
location and contacts	about the hotel	173
above and beyond	about the hotel	172
home	about the hotel	163
above and beyond	facilities	146
our city	about the hotel	141
location and contacts	rooms	128
location and contacts	dining	126
events	about the hotel	123

*Continued on next page...*



Table 1 (*continued from Page 7*): Internal Request Frequency Patterns

<b>From</b>	<b>To</b>	<b>Frequency</b>
about the hotel	rooms	123
rooms	offers	123
about the hotel	dining	122
dining	offers	122
home	dining	115
our city	offers	112
events	offers	110
home	rooms	107
our city	facilities	104
facilities	about the hotel	103
about the hotel	offers	103
rooms	dining	102
dining	offers	102
facilities	about the hotel	101
about the hotel	rooms	101
facilities	rooms	99
rooms	offers	99
facilities	dining	96
dining	offers	96
about the hotel	rooms	96
rooms	dining	96
above and beyond	dining	94
dining	offers	94
events	dining	92
our city	rooms	92

*Continued on next page...*

Table 1 (*continued from Page 7*): Internal Request Frequency Patterns

<b>From</b>	<b>To</b>	<b>Frequency</b>
above and beyond	rooms	89
rooms	offers	89
above and beyond	facilities	89
facilities	rooms	89
above and beyond	about the hotel	86
about the hotel	rooms	86
facilities	location and contacts	85
our city	dining	85
above and beyond	about the hotel	84
about the hotel	dining	84
facilities	about the hotel	82
about the hotel	dining	82
above and beyond	rooms	82
rooms	dining	82
our city	above and beyond	82
above and beyond	location and contacts	80
home	about the hotel	80
about the hotel	offers	80
location and contacts	about the hotel	80
about the hotel	offers	80
above and beyond	facilities	79
facilities	about the hotel	79
facilities	rooms	76
rooms	dining	76
above and beyond	about the hotel	76

*Continued on next page...*

Table 1 (*continued from Page 7*): Internal Request Frequency Patterns

<b>From</b>	<b>To</b>	<b>Frequency</b>
about the hotel	offers	76
above and beyond	facilities	76
facilities	offers	76
above and beyond	facilities	73
facilities	dining	73
location and contacts	home	72
facilities	home	71
our city	location and contacts	70
our city	about the hotel	70
about the hotel	offers	70
location and contacts	rooms	70
rooms	offers	70
events	rooms	69
events	facilities	69
home	rooms	68
rooms	offers	68
home	dining	68
dining	offers	68
location and contacts	about the hotel	68
about the hotel	dining	68
above and beyond	home	65
our city	about the hotel	65
about the hotel	rooms	65
our city	facilities	65
facilities	about the hotel	65

*Continued on next page...*

Table 1 (*continued from Page 7*): Internal Request Frequency Patterns

<b>From</b>	<b>To</b>	<b>Frequency</b>
location and contacts	about the hotel	64
about the hotel	rooms	64
home	about the hotel	63
about the hotel	dining	63
events	home	60
facilities	about the hotel	59
about the hotel	rooms	59
rooms	offers	59
our city	about the hotel	59
about the hotel	dining	59
location and contacts	dining	59
dining	offers	59
default	offers	58
our city	rooms	58
rooms	offers	58
about the hotel	rooms	58
rooms	dining	58
dining	offers	58

## **B   Extrapolation Results**

Attached on the following pages are the results from Databricks. You may also interact with this online on Databricks.