

# CS 185C – Introduction to Big Data Analytics

**TripAdvisor Hotel Reviews**



**Submitted to : Prof. Peter Zadrozny**

**Submitted by : Team#2**

**Sailee Choudhary**

**Tin Hon Ng**

**Anumeha Umang Shah**

**Date: 11/19/2014**

Server : <http://216.121.58.230:8000/>

username : admin

Password for root : Noeasywayout203

## Contents

Introduction.....	5
Description.....	5
Objective .....	5
Cluster Description .....	5
Phase 1 - Loading and Verifying Data.....	7
Dataset Description.....	7
Where we obtained the data.....	7
How we obtained the data.....	7
Issues with obtaining the data .....	7
Structure and Key Fields of the data set(s) .....	8
Issues with the data set(s) .....	9
Pre-processing (ETL) of the data .....	9
Loading the data.....	13
1. Loading the reviews table .....	13
2. Loading the hotel information table.....	20
Data Verification.....	25
1. Verification of count of events.....	25
2. Verification of fields .....	27
3. Verification of data .....	30
Phase 2 - Building and Verifying Queries.....	37
1. Number of reviews for each year .....	37
Visualization .....	38
2. State wise count for reviews .....	38

Visualization .....	39
3. City wise count for reviews .....	39
Visualization .....	40
4. To find the number of cities for which reviews are recorded in each state.....	41
Visualization .....	43
5. The region that has the most number of hotel reviews according to postal code.....	43
Visualization .....	45
6. Month that has the most number of reviews in all the years .....	46
Visualization .....	47
7. Month wise count of reviews for years 2010 to 2012 .....	47
Visualization .....	50
8. Top 10 hotels that have the most reviews.....	50
Visualization .....	51
9. Top 10 hotels that have the most positive reviews(2.5 + on overall rating) .....	51
Visualization .....	53
10. Hotels that have the most negative reviews(2.5 - on overall rating).....	53
Visualization .....	54
11. Top 10 useful reviewers.....	55
Visualization .....	56
12. Hotel with most number of overall ratings as 5.0 .....	56
Visualization .....	57
13. Hotels with most number of overall ratings as 1.0.....	57
Visualization .....	58
14. Average of all types of ratings for hotels.....	59

Visualization .....	61
15. Top 5 hotels having negative reviews but the most number of review count in NYC.....	61
Visualization .....	62
16. Top 10 haunted hotels based on user's reviews in USA.....	63
Visualization .....	64
17. Top 10 haunted hotels based on user's reviews in NY .....	65
Visualization .....	67
18. The top 20 hotels with bed bugs based on user's reviews .....	67
Visualization .....	68

## Introduction

### Description

In today's fast paced digital world, customer reviews are so important. After all, when you look for dinner, place to stay, or entertainment, most likely you look for reviews to help you make your decision. Speaking of reviews, TripAdvisor is one of the most popular travel websites in the US providing people with reviews on travel-related content. Since the website services are free to the users, users can provide as many reviews as they want. Many people have experience staying in hotel while traveling, no surprise the number reviews on hotels are very high with rich contents.

Therefore, our team decided to do an in-depth analysis on the hotel review dataset provided by TripAdvisor. We will apply Splunk technology to perform this analysis. The major objective is to load data into Splunk and come up with appropriate queries to achieve our user stories and suitable conclusions that are useful to others.

### Objective

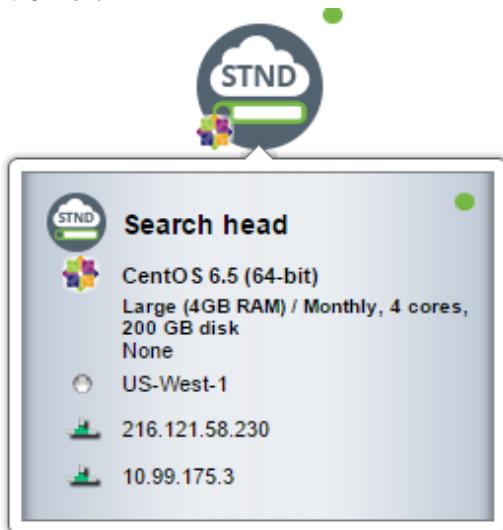
Using the hotel review dataset provided by TripAdvisor, our purpose is to speculate the correlation between different kinds of rating, author, devices they used, etc. Our aim is to discover the interesting facts about our dataset. We apply Splunk technology to perform this analysis. Again the main objective is to load data into Splunk and use it to trigger appropriate queries in order to arrive at suitable conclusions that are relevant and useful to others.

### Cluster Description

Server size : Large (4 GB, 4 cores)

OS : CentOS

Architecture : 64-bit



**Credentials :**

To login to Splunk : <http://216.121.58.230:8000/>

username : admin

Password for root : Noeasywayout203

## Phase 1 - Loading and Verifying Data

### Dataset Description

This dataset consists of 878561 reviews from 4333 hotels crawled from TripAdvisor. It consists of two files one with all the reviews for different hotels, and other that contains the information about the hotels. The data is spread over several span of years. According to the survey, 81% of travelers find user reviews important when determining which hotel to stay at during their trip. So, positive feedback from satisfied guests sharing their experiences with your Hotel is very precious. Negative reviews – in contrast – should be diminished and used to improve the quality of the Hotel. So, it is important to have reviews both for customers as well as hotel management. So, we decided to analyze this data.

### Where we obtained the data

We extracted our data from Carnegie Mellon University website.

URL : <http://www.cs.cmu.edu/~jiweil/html/hotel-review.html>

### How we obtained the data

Everyone likes to travel! And we always don't have someone who resides where we are travelling. We need to sort shelter in some or the other hotel. It is always better to live in good hotel where you can find all the necessary amenities. It is always irritating to pay loads of money and not get customer satisfaction. Being fond of travelling and exploring new places, we decided to dig deeper into reviews of different hotels. So, with the help of TripAdvisor hotel reviews that we obtained from <http://www.cs.cmu.edu/> we could do an in-depth analysis.

### Issues with obtaining the data

We used all the publicly available datasets. It required no registration or identification to access, and had no legal restrictions on their use.

## Structure and Key Fields of the data set(s)

File	Original Structure	New Structure
review.txt	json	json
offering.txt	json	csv

### 1. review.txt

{

**author:** {

**id:** unique id of each author  
**location:** the home location of the author  
**num\_reviews:** number of reviews posted by the author on TripAdvisor  
**username:** username specified by author on TripAdvisor

}

**date:** date when the review was posted (format : Month date, Year)

**date\_stayed:** date when author stayed at this hotel (format : Month Year)

**id:** unique review id for each review

**num\_helpful\_votes:** count of useful reviews provided by author

**offering\_id:** hotel id where the author stayed

**ratings:** { ratings provided by the author based on different criteria.

**cleanliness:** (0-5)

**location:** (0-5)

**overall:** (0-5)

**rooms:** (0-5)

**service:** (0-5)

**sleep\_quality:** (0-5)

**value:** (0-5)

}

**text:** any comments or suggestions that author wants to provide

**title:** title for the review provided by author

**via\_mobile:** tells if the review was posted via mobile or not

}

## 2. offering.csv

Field Name	Description
hotel_class	Hotel star rating information
region_id	Region id for the hotel depending on its location
url	The website link for the hotel
address_region	State where hotel is located
address_street-address	Street address for the hotel
address_postal-code	Address postal code
address_locality	City in which the hotel is located
type	Type for the hotel
hotel_id	Unique hotel id for each hotel
name	Name of the hotel

### Issues with the data set(s)

- Both the files, review.txt and offering.txt had a field 'id'.

For, review.txt, 'id' is for 'review id' and it is unique for each event across the whole file.

For, offering.txt, 'id' is for 'hotel id' and this field needs mapping with 'offering\_id' in the review.txt because for review.txt, the offering\_id specifies the 'hotel id'

So, while pre-processing the data, we changed the field 'id' in offering.txt to 'hotel\_id'

- We wanted to use the file offering.txt as lookup table. Splunk accepts only csv files as lookup tables. Therefore, we had to write a php script to change the json file to csv file.

### Pre-processing (ETL) of the data

1. review.txt - This file is in json format which is identified by Splunk. So, no pre-processing was required for this file.

2. offering.txt - This file contains information about hotels and so, we wanted to use this file as lookup table. The file was in json format and to be used as lookup table we had to convert it to csv. We wrote the following script for converting it to csv.

```

<?php

//config vars
$input_json = "offering.txt";
$output_csv = "offering.csv";
$csv_json_map = array(
    'hotel_class' => "hotel_class",
    'region_id' => "region_id",
    'url' => "url",
    'phone' => "phone",
    'details' => "details",
    'address_region' => "address->region",
    'address_street-address'=> "address->{'street-address'}",
    'address_postal-code' => "address->{'postal-code'}",
    'address_locality' => "address->locality",
    'type' => "type",
    'id' => "hotel_id",
    'name' => "name"
);

//processing starts
#open input json file
$handle = fopen($input_json, "r") or die("Couldn't get handle");

#open output csv file
$file = fopen($output_csv, 'w');

#write csv header row
$arr_header = array_keys($csv_json_map);
$header = implode(",",$arr_header);
fwrite($file, $header."\n");

if ($handle) {
    while (!feof($handle)) {
        $buffer = fgets($handle, 4096);

        #json_decode input row
        $json = json_decode($buffer);

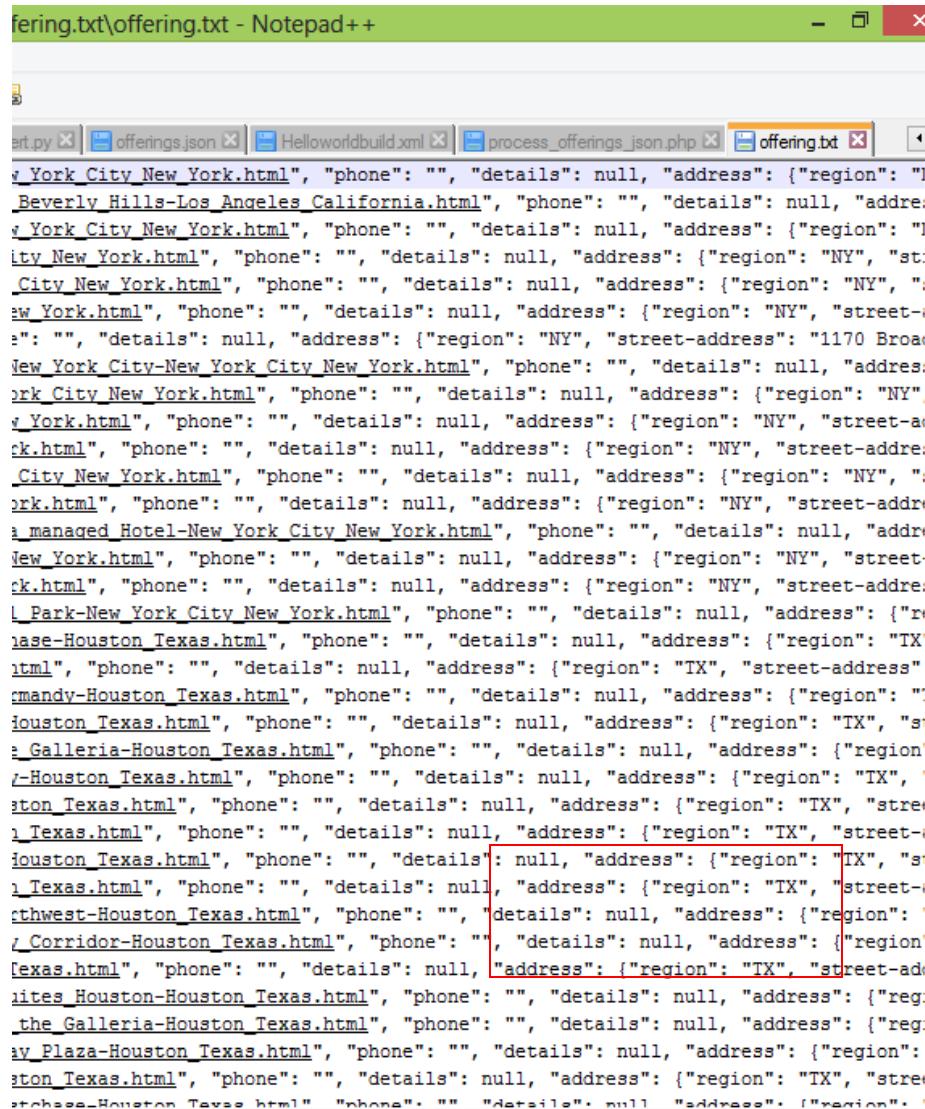
        #prepare row data
        $row = ",";
        foreach($csv_json_map as $csv_field => $json_field)
        {
            $value = "";
            eval("\$value = isset(\$json->$json_field)? \$json->$json_field : \"\";");
            $row .= "" . str_replace("\"", "", $value). ",";
        }
    }
}

```

```
    }  
    #write row  
    fwrite($file, $row."\n");  
}  
fclose($handle);  
}  
?>
```

Also, while doing this transformation, we changed the field 'id' to 'hotel\_id' as in the review.txt, there was field 'id' already used for review id. So, to avoid the confusion, we renamed this field.

In the original json file, the fields, 'details' and 'phone' were null. So, we removed these from our offerings.csv



Columns in final csv : offering.csv

hotel_class	region_id	url	address_region	address_street-address	address_postal-code	address_locality	type	hotel_id	name
2	2	31310 http://wAZ		2425 South 24th Street	85034	Phoenix	hotel	73706	BEST WESTERN Airport Inn
2	2	31310 http://wAZ		17211 North Black Canyon	85023	Phoenix	hotel	73712	Super 8 Phoenix
3	3	31310 http://wAZ		1100 N Central Ave	85004	Phoenix	hotel	73718	Lexington Hotel Central Phoenix
5	3	31310 http://wAZ		10831 South 51st Street	85044	Phoenix	hotel	73727	Grace Inn Phoenix

### Verification if the data was pre-processed correctly :

Check the first 5 records of original and csv files:

hotel_class	region_id	url	address_region	address_street-address	address_postal-code	address_locality	type	hotel_id	name
2	2	31310 http://wAZ		2425 South 24th Street	85034	Phoenix	hotel	73706	BEST WESTERN Airport Inn
2	2	31310 http://wAZ		17211 North Black Canyon	85023	Phoenix	hotel	73712	Super 8 Phoenix
3	3	31310 http://wAZ		1100 N Central Ave	85004	Phoenix	hotel	73718	Lexington Hotel Central Phoenix
5	3	31310 http://wAZ		10831 South 51st Street	85044	Phoenix	hotel	73727	Grace Inn Phoenix
2.5	2.5	31310 http://wAZ		1615 East Northern Ave.	85020	Phoenix	hotel	73739	BEST WESTERN PLUS InnSuites Phoenix Hotel & Suites
7	2	31310 http://wAZ		1711 W. Bell Road	85023	Phoenix	hotel	73743	A Victory Inn & Suites Phoenix North
8	3	31310 http://wAZ		9631 North Black Canyon	85021	Phoenix	hotel	73751	Courtyard by Marriott Phoenix North
9	3.5	31310 http://wAZ		50 East Adams Street	85004	Phoenix	hotel	73757	Renaissance Phoenix Downtown
10	2	31310 http://wAZ		502 W Camelback Rd	85013	Phoenix	hotel	73760	Days Inn Camelback Phoenix and Conference Center
11	2	31310 http://wAZ		2420 West Thomas Rd	85015	Phoenix	hotel	73768	Days Inn I-17 & Thomas
12		31310 http://wAZ		1550 North 52nd Drive	85043	Phoenix	hotel	73773	American Best Value Inn Phoenix/I-10 West
13		31310 http://wAZ		3210 Northwest Grand Av	85017	Phoenix	hotel	73782	Comfort Suites Conference Center
14	3	31310 http://wAZ		1515 North 44th Street	85008	Phoenix	hotel	73787	Holiday Inn & Suites Phoenix Airport North
15	3	31310 http://wAZ		2333 E Thomas Rd	85016	Phoenix	hotel	73792	Embassy Suites Phoenix Airport at 24th Street
16	3	31310 http://wAZ		2577 W Greenway Road	85023	Phoenix	hotel	73799	Embassy Suites Hotel Phoenix-North
17	2	31310 http://wAZ		1241 N 53rd Avenue	85043	Phoenix	hotel	73803	Days Inn Phoenix West
18	2.5	31310 http://wAZ		8101 North Black Canyon	85021	Phoenix	hotel	73805	BEST WESTERN Phoenix I-17 MetroCenter Inn
19	2.5	31310 http://wAZ		160 West Catalina Drive	85013	Phoenix	hotel	73810	Hampton Inn Phoenix Midtown (Downtown Area)
20		31310 http://wAZ		2990 West Thunderbird Rd	85053	Phoenix	hotel	73817	Windsor Palms Apartments
21	3.5	31310 http://wAZ		2435 South 47th Street	85034-6410	Phoenix	hotel	73821	Hilton Phoenix Airport
22	3	31310 http://wAZ		1500 N. 51st Ave.	85043	Phoenix	hotel	73825	Holiday Inn Phoenix-West
23	3	31310 http://wAZ		212 W Osborn	85013	Phoenix	hotel	73840	Holiday Inn Phoenix Downtown North
24	3.5	31310 http://wAZ		122 North 2nd Street	85004	Phoenix	hotel	73855	Hyatt Regency Phoenix
25	2	31310 http://wAZ		2725 North Black Canyon	85009	Phoenix	hotel	73859	La Quinta Inn Phoenix Thomas Road



The screenshot shows a Microsoft WordPad window with a green title bar labeled "offering - WordPad". The menu bar includes "File", "Home", and "View". The main content area displays a large block of JSON data representing hotel reviews. The data lists various hotels with their URLs, addresses, phone numbers, and other details. The JSON structure includes fields like "hotel\_class", "region\_id", "url", "name", "address", "phone", and "type". Some entries are highlighted with red boxes, such as the last two records which are identical.

```

[{"hotel_class": 4.0, "region_id": 60763, "url": "http://www.tripadvisor.com/Hotel_Review-g60763-d113317-Reviews-Casablanca_Hotel_Times_Square-New_York_City_New_York.html", "phone": "", "details": null, "address": {"region": "NY", "street-address": "147 West 43rd Street", "postal-code": "10036", "locality": "New York City"}, "type": "hotel", "id": 113317, "name": "Casablanca Hotel Times Square"}, {"hotel_class": 5.0, "region_id": 32655, "url": "http://www.tripadvisor.com/Hotel_Review-g32655-d76049-Reviews-Four_Seasons_Hotel_Los_Angeles_at_Beverly_Hills-Los_Angeles_California.html", "phone": "", "details": null, "address": {"region": "CA", "street-address": "300 S Doheny Dr", "postal-code": "90048", "locality": "Los Angeles"}, "type": "hotel", "id": 76049, "name": "Four Seasons Hotel Los Angeles at Beverly Hills"}, {"hotel_class": 3.5, "region_id": 60763, "url": "http://www.tripadvisor.com/Hotel_Review-g60763-d99352-Reviews-Hilton_Garden_Inn_Times_Square-New_York_City_New_York.html", "phone": "", "details": null, "address": {"region": "NY", "street-address": "790 Eighth Avenue", "postal-code": "10019", "locality": "New York City"}, "type": "hotel", "id": 99352, "name": "Hilton Garden Inn Times Square"}, {"hotel_class": 4.0, "region_id": 60763, "url": "http://www.tripadvisor.com/Hotel_Review-g60763-d93589-Reviews-The_Michelangelo_Hotel-New_York_City_New_York.html", "phone": "", "details": null, "address": {"region": "NY", "street-address": "152 West 51st Street", "postal-code": "10019", "locality": "New York City"}, "type": "hotel", "id": 93589, "name": "The Michelangelo Hotel"}, {"hotel_class": 4.0, "region_id": 60763, "url": "http://www.tripadvisor.com/Hotel_Review-g60763-d217616-Reviews-The_Muse_Hotel-New_York_City_New_York.html", "phone": "", "details": null, "address": {"region": "NY", "street-address": "130 West 46th Street", "postal-code": "10036", "locality": "New York City"}, "type": "hotel", "id": 217616, "name": "The Muse Hotel New York"}]

```

They are exactly same.

Also, checking the number of records

4330	3	60878 http://w WA	300 Roy Street	98109 Seattle	hotel	1723585	The Maxwell Hotel Seattle
4331	3	60878 http://w WA	612 2nd Avenue	98104 Seattle	hotel	1724250	Courtyard Seattle Downtown
4332	3	60878 http://w WA	110 6th Avenue North	98109 Seattle	hotel	1755673	Hyatt Place Seattle/Downtow
4333		60878 http://w WA	9613 Aurora Ave N	98103 Seattle	hotel	2095177	Columbus Motor Inn
4334		60878 http://w WA	9501 Aurora Avenue Nort WA 98103	Seattle	hotel	3291041	Crown Inn
4335							
4336							
4337							

```
[root@search-head ~]# wc -l offering.txt
4333 offering.txt
[root@search-head ~]#
```

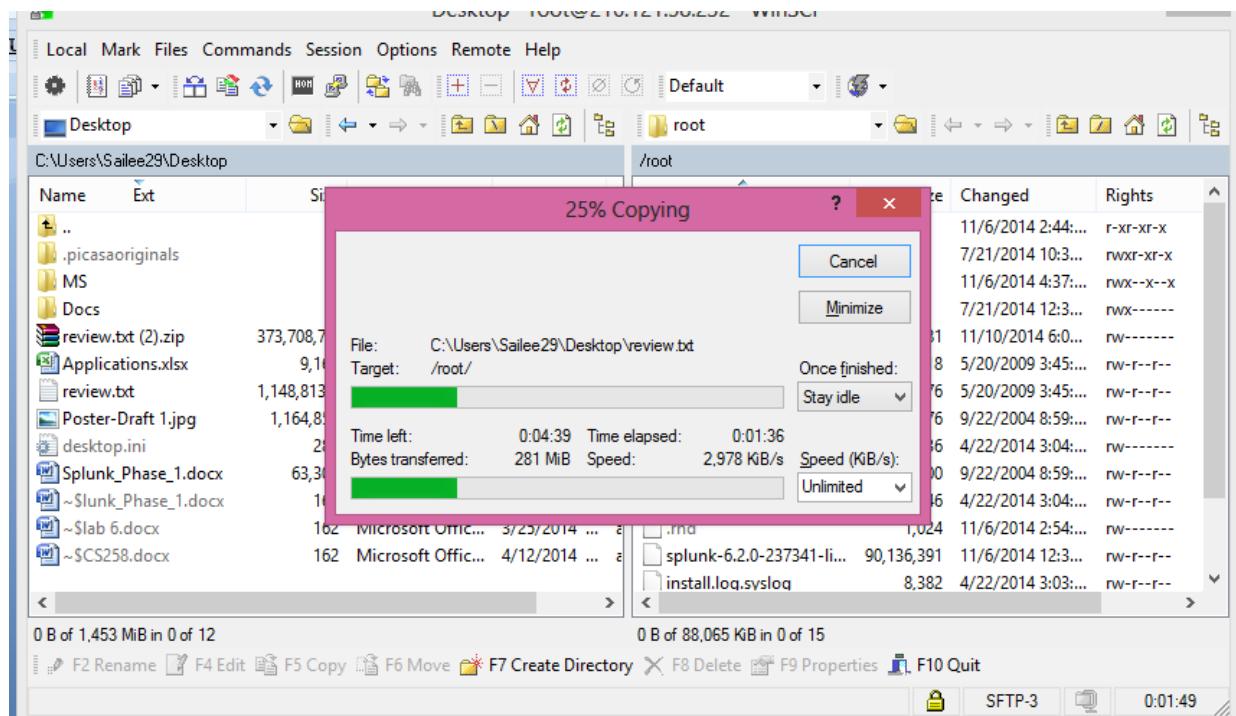
As in excel, the first line is for column name, so we have one extra record in csv file.

## Loading the data

### 1. Loading the reviews table

**Step 1 :** Transfer the data from local machine to server root directory.

Use WinSCP to transfer data from local machine to search head.



**Step 2:** Login to search head cluster and check if the data file is transferred.

```

root@search-head:~#
login as: root
root@216.121.58.234's password:
Last login: Wed Nov 19 13:25:29 2014 from c-24-4-154-19.hsd1.ca.comcast.net
[root@search-head ~]# ls
abc.json           install.log.syslog
anaconda-ks.cfg    review.csv
bcc.txt            review.json
hotel_reviews.json review.txt
install.log        splunk-6.2.0-237341-linux-2.6-x86_64.rpm
[root@search-head ~]#

```

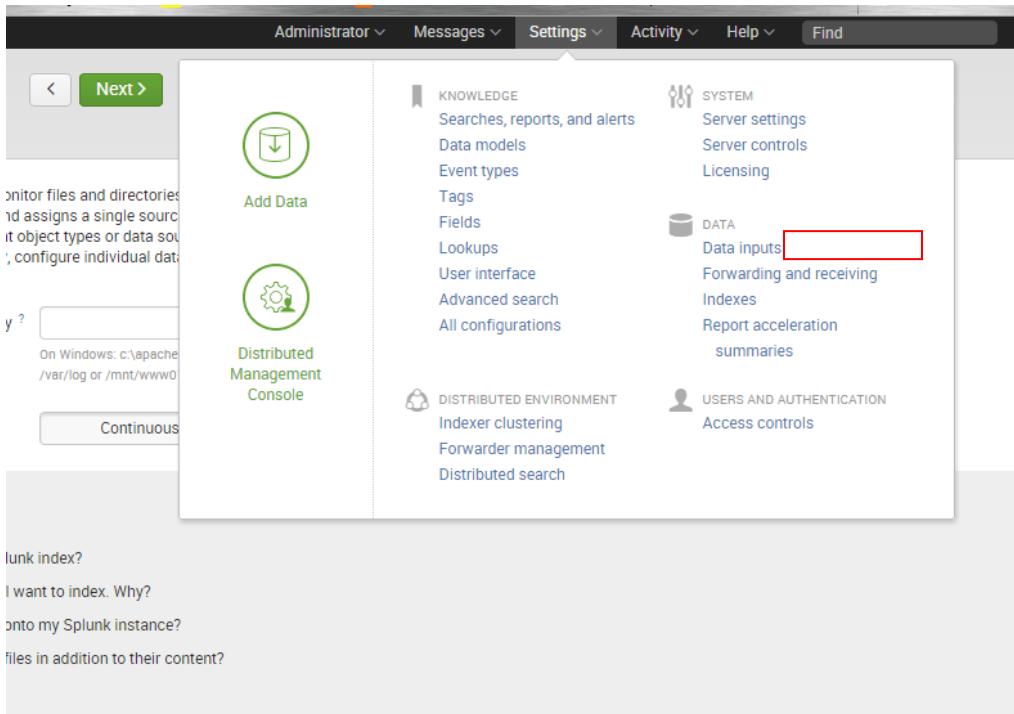
**Step 3 :** Login to the web interface of search head.

URL : <http://216.121.58.230:8000/>

username : admin

password : Noeasywayout203

**Step 4 :** Go to 'Settings' and click on 'Data inputs'.



**Step 5 :** Click on 'Add new' that resides besides 'Files & directories'.

The screenshot shows the 'Data inputs' configuration page. Under 'Local inputs', there is a table with three rows: 'Files & directories', 'TCP', and 'UDP'. The 'Actions' column for each row contains a 'Add new' button, which is highlighted with a red box.

Type	Inputs	Actions
Files & directories	5	Add new
TCP	0	Add new
UDP	0	Add new

**Step 6 :** Click on 'Browse' and on the next screen, go to the 'root' directory and select the file to be loaded.

File loaded : "review.txt"

Click on 'Index once' and click 'Next'

The screenshot shows the 'Add Data' configuration page in Splunk. The 'File or Directory' field contains '/root/review.txt'. The 'Continuously Monitor' button is highlighted with a red box. The 'Index Once' button is also highlighted with a red box.

**Select source**

- misc
- mnt
- net
- opt
- proc
- root**
  - abc.json
  - anaconda-ks.cfg
  - bcc.txt
  - hotel\_reviews.json
  - install.log
  - install.log.syslog
  - review.csv
  - review.json
  - review.txt**
  - splunk-6.2.0-237341-linux-2.6-x86\_64.rpm
- sbin
- selinux
- srv
- sys
- tmp

/root/review.txt

**Select**

### Step 7 : Set Sourcetype.

(1) Select sourcetype by first going to 'Structured' and then '**\_json**'

(2) Timestamp :

Select 'Advanced' and specify the timestamp format and fields.

Timestamp format : **%B %d, %Y**

Timestamp fields : date

Sourcetype: `_json`

**Timestamp**

- Extraction: Auto, Current time, Advanced... (highlighted)
- Time zone: Auto

Timestamp format: `%B %d, %Y`

Timestamp fields: `date`

	<code>_time</code>	<code>author.id</code>	<code>author.location</code>
1	12/17/12 12:00:00.000 AM	8C0B42FF3C0FA366A21CFD785302A032	Gold Coast

(3) Select 'Advanced' and click on 'New setting'

Add an extra field:

- Name : TRUNCATE

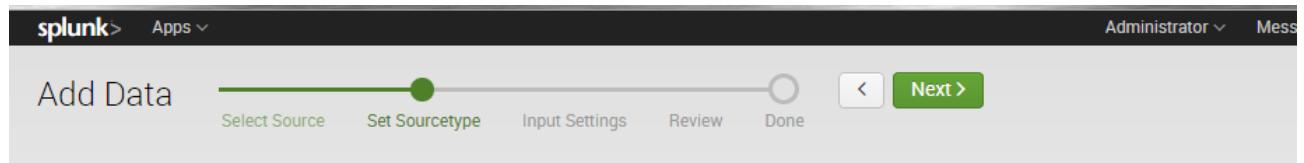
Value : 0

(Some of the events in our data were too long because of long comments in the 'text' field. So, few events were not loaded properly. In order to avoid this, we specified a setting to never truncate any event in props.conf)

- Name : MAX\_DAYS\_AGO

Value : 10951

(This field is used for specifying for how many days before the current date should Splunk index the data. Since, our data was starting from year 2001, we set this value to the maximum value(10951). Default value is 2000)



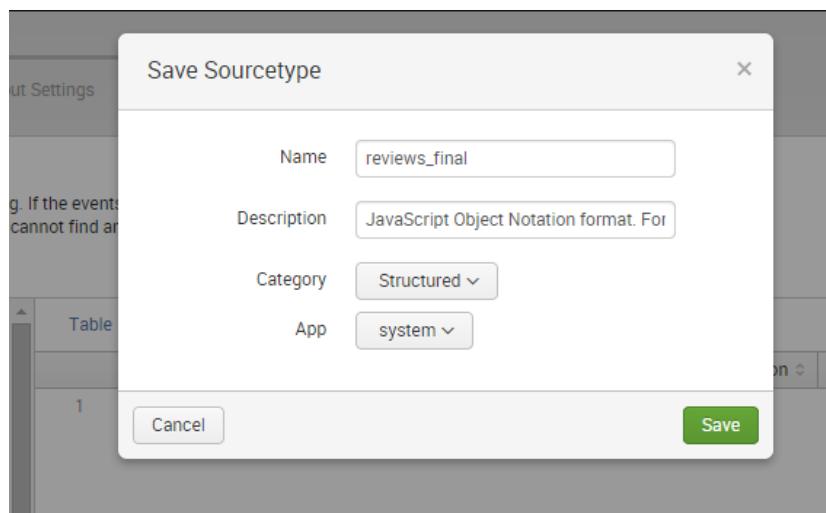
## Set Sourcetype

Data preview lets you see how Splunk sees your data before indexing. If the events look correct and have the right timestamps, click "Next" to proceed. If not, use the options below to define proper event breaks and timestamps. If you cannot find an appropriate source type for your data, create a new one by clicking "Save As".

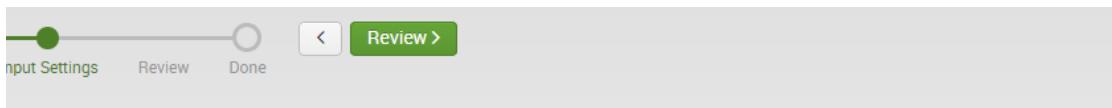
Source: /root/review.txt

The configuration screen shows various parameters for setting the sourcetype. The 'TRUNCATE' and 'MAX\_DAYS\_AGO' fields are highlighted with a red box. The 'TRUNCATE' field is set to 0 and the 'MAX\_DAYS\_AGO' field is set to 10951. Other fields include INDEXED\_EXTRACTIO (json), KV\_MODE (json), category (Structured), description (JavaScript Object Notation), disabled (false), pulldown\_type (true), TIME\_FORMAT (%B %d, %Y), and TIMESTAMP\_FIELDS (date).

**Step 8 :** Click on 'Save as' and save the sourcetype.



**Step 9 :** Specify the input settings and click on 'Review'



llows:

tain  
texts  
k loads

App Context Search & Reporting ▾

: host  
gitates,  
regular  
.learn

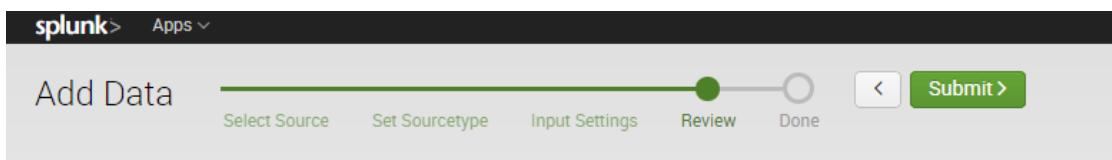
Host field value Constant value Regular expression on path Segment in path

search-head.splunk.com

ider  
ermining  
t your  
sys

Index Default ▾ Create a new index ▾  
Refresh

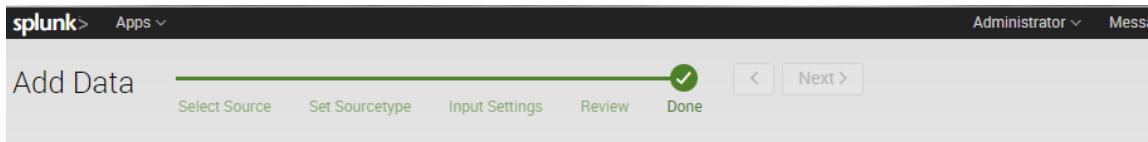
**Step 10:** If everything is correct, click on 'Submit' else go back and do the necessary changes.



### Review

Input Type	File Monitor
Source Path	/root/review.txt
Continuously Monitor	No, index once
Sourcetype	reviews_final
App Context	search
Host	search-head.splunk.com
Index	default

**Step 11:** Start Searching after you see the following screen.



✓ File input has been created successfully.

Configure your inputs by going to [Settings > Data Inputs](#)

[Start Searching](#)

Search your data now or see examples and tutorials. [\[link\]](#)

[Extract Fields](#)

Create search-time field extractions. Learn more about fields. [\[link\]](#)

[Add More Data](#)

Add more data inputs now or see examples and tutorials. [\[link\]](#)

[Download Apps](#)

Apps help you do more with your data. Learn more. [\[link\]](#)

[Build Dashboards](#)

Visualize your searches. [Learn more](#). [\[link\]](#)

## 2. Loading the hotel information table

The hotel information table provides information about the hotels such as address, url, name, etc. The hotel\_id field of this table connects with the

**Step 1 :** Login to the web interface of search head.

URL : <http://216.121.58.230:8000/>

username : admin

password : Noeasywayout203

**Step 2 :** Go to 'Settings' and click on 'Lookups'.

The screenshot shows the Splunk Settings interface. On the left, there's a sidebar with icons for 'Add Data' and 'Distributed Management Console'. The main area is titled 'Settings' and contains several categories: 'KNOWLEDGE' (Searches, reports, and alerts; Data models; Event types; Tags; Fields; Lookups), 'SYSTEM' (Server settings; Server controls; Licensing), 'DATA' (Data inputs; Forwarding and receiving; Indexes; Report acceleration summaries), 'DISTRIBUTED ENVIRONMENT' (Indexer clustering; Forwarder management; Distributed search), and 'USERS AND AUTHENTICATION' (Access controls). A red box highlights the 'Lookups' link under the Knowledge category.

**Step 3 :** Click on 'Add new' adjacent to Lookup table files.

The screenshot shows the 'Lookups' configuration page. At the top, it says 'Create and configure lookups.' Below that are three sections: 'Lookup table files' (List existing lookup tables or upload a new file.), 'Lookup definitions' (Edit existing lookup definitions or define a new file-based or external lookup.), and 'Automatic lookups' (Edit existing automatic lookups or configure a new lookup to run automatically.). To the right of each section is an 'Actions' button, with the 'Add new' button for 'Lookup table files' highlighted by a red box.

**Step 4 :** Fill out the details for lookup table files. Browse the file that you want to use as lookup table.

Destination app : search

Destination app \*

search

Upload a lookup file

Choose File offering.csv

Select either a plaintext CSV file or a gzipped CSV file.  
The maximum file size that can be uploaded through the browser is 500MB.

Destination filename \*

hotel\_info.csv

Enter the name this lookup table file will have on the Splunk server. If you are uploading a gzipped CSV file, enter a filename ending in ".gz". If you are uploading a plaintext CSV file, we recommend a filename ending in ".csv".

**Step 5 :** The new lookup table uploaded should be reflected in the lookup table files. Click on 'Permissions' for this file.

Lookup table files

Lookups » Lookup table files

New

Showing 1-1 of 1 item

Path	Owner	App	Sharing	Status	Actions
/opt/splunk/etc/users/admin/search/lookups/hotel_info.csv	admin	search	Private   Permissions	Enabled	Move   Delete

**Step 6 :** Give the appropriate permissions for the file and click on 'Save'

Here, we made the lookup table accessible to all the apps.

Object should appear in

Keep private  This app only (search)  All apps

Permissions

Roles	Read	Write
Everyone	<input type="checkbox"/>	<input type="checkbox"/>
admin	<input type="checkbox"/>	<input type="checkbox"/>
can_delete	<input type="checkbox"/>	<input type="checkbox"/>
power	<input type="checkbox"/>	<input type="checkbox"/>
splunk-system-role	<input type="checkbox"/>	<input type="checkbox"/>
user	<input type="checkbox"/>	<input type="checkbox"/>

**Step 7 :** On clicking on save, the file permissions should be reflected in the lookup table files

Lookup table files  
Lookups > Lookup table files

App context: Search & Reporting (search) Owner: Any

Show only objects created in this app context  Learn more

New

Showing 1-1 of 1 item

Path	Owner	App	Sharing	Status	Actions
/opt/splunk/etc/apps/search/lookups/hotel_info.csv	admin	search	Global   Permissions	Enabled	Move   Delete

**Step 8 :** Next, Go to 'Settings', then 'Lookups' and then click on 'Add New' besides 'Lookup definitions'

Destination app \*: search

Name \*: L\_hotel\_info

Type \*: File-based

Lookup file \*: hotel\_info.csv

Configure time-based lookup

Advanced options

Cancel Save

Save this Lookup definition. It should appear in 'Lookup definitions'. Click on 'Permissions' and select 'This app only (search)' and save.

Object should appear in  
 Keep private  This app only (search)  All apps

Permissions

Roles	Read	Write
Everyone	<input checked="" type="checkbox"/>	<input type="checkbox"/>
admin	<input type="checkbox"/>	<input checked="" type="checkbox"/>
can_delete	<input type="checkbox"/>	<input type="checkbox"/>
power	<input type="checkbox"/>	<input checked="" type="checkbox"/>
splunk-system-role	<input type="checkbox"/>	<input type="checkbox"/>
user	<input type="checkbox"/>	<input type="checkbox"/>

Cancel Save

**Step 9 :** Go to 'Settings', then 'Lookups' and then click on 'Add New' besides 'Automatic Lookups' and configure the fields and 'Save'.

The screenshot shows the 'Destination app' set to 'search', 'Name' to 'AL\_hotel\_info', and 'Lookup table' to 'L\_hotel\_info'. Under 'Apply to', 'sourceType' is selected and 'reviews\_final' is named. In the 'Lookup input fields' section, 'hotel\_id' is mapped to 'offering\_id'. Below it, under 'Add another field', there is a list of 'Lookup output fields' with mappings such as 'address\_locality' to 'hotel\_city', 'address\_postal-code' to 'hotel\_postal', etc. An 'Add another field' button is visible at the bottom.

**Step 10 :** Edit the Permissions and make this automatic lookup table available to Search App.

The screenshot shows the 'Permissions' dialog for an object. It has three radio button options: 'Keep private', 'This app only (search)' (which is selected), and 'All apps'. The 'Permissions' section lists roles and their access levels:

Roles	Read	Write
Everyone	<input checked="" type="checkbox"/>	<input type="checkbox"/>
admin	<input type="checkbox"/>	<input checked="" type="checkbox"/>
can_delete	<input type="checkbox"/>	<input type="checkbox"/>
power	<input type="checkbox"/>	<input checked="" type="checkbox"/>
splunk-system-role	<input type="checkbox"/>	<input type="checkbox"/>
user	<input type="checkbox"/>	<input type="checkbox"/>

At the bottom are 'Cancel' and 'Save' buttons.

**Step 11 :** The automatic lookup table should be reflected in 'Automatic lookups'

Automatic lookups

Lookups > Automatic lookups

App context Search & Reporting (search) Owner Any

Show only objects created in this app context Learn more

New

Showing 1-1 of 1 item

Results per page 25

Name	Lookup	Owner	App	Sharing	Status	Actions
reviews_final : LOOKUP: AL_hotel_info	L_hotel_info hotel_id AS offering_id OUTPUTNEW address_locality AS hotel_city "address_postal-code" AS hotel_postal address_region AS hotel_state "address_street-address" AS hotel_street_address hotel_class AS hotel_class name AS hotel_name region_id AS hotel_region_id type AS hotel_type url AS hotel_url	admin	search	App   Permissions	Enabled	Clone   Move   Delete

## Automatic lookup table :

Field Name	Field in Automatic lookup table
hotel_class	hotel_class
region_id	hotel_region_id
url	hotel_url
address_region	hotel_state
address_street-address	hotel_street_address
address_postal-code	hotel_postal
address_locality	hotel_city
type	hotel_type
hotel_id	hotel_id
name	hotel_name

## Data Verification

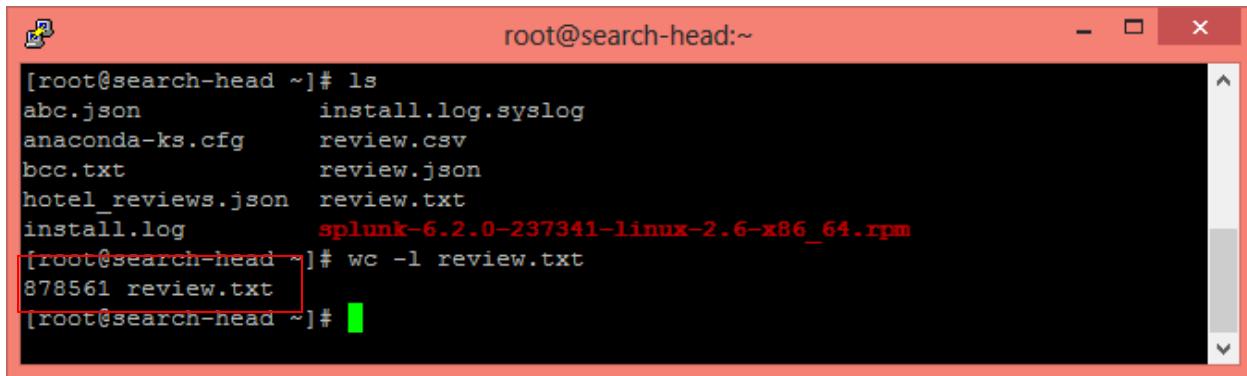
### 1. Verification of count of events

#### 1. Verification for count of table review.txt

review.txt contains all the hotel reviews from TripAdvisor. There are 878561 reviews in this file. We first verify if all the reviews have been uploaded correctly.

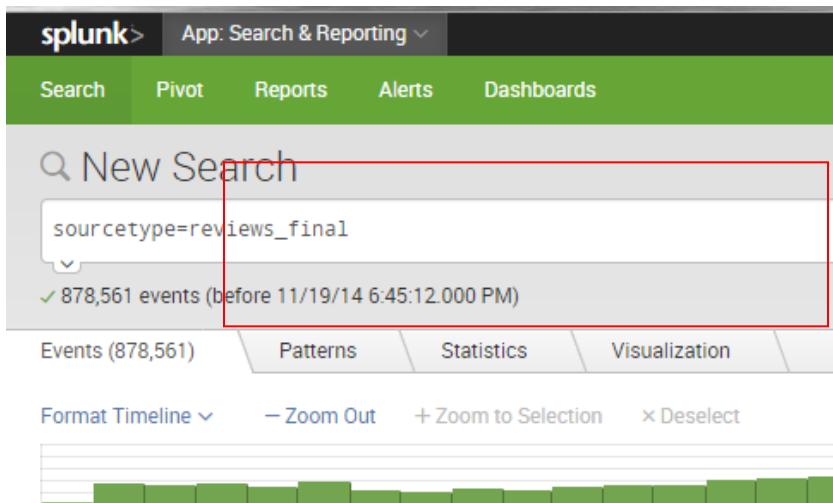
This is the snapshot of the count of records in the original file. (review.txt)

```
# wc -l review.txt
```



```
root@search-head:~# ls
abc.json           install.log.syslog
anaconda-ks.cfg    review.csv
bcc.txt            review.json
hotel_reviews.json review.txt
install.log        splunk-6.2.0-237341-linux-2.6-x86_64.rpm
[root@search-head ~]# wc -l review.txt
878561 review.txt
[root@search-head ~]#
```

We now check if there are same number of events in Splunk :



The screenshot shows the Splunk interface with the following details:

- Search Bar:** Contains the query `sourcetype=reviews_final`.
- Results Summary:** Shows `878,561 events (before 11/19/14 6:45:12.000 PM)`.
- Event Count:** The number `878,561` is highlighted with a red box.
- Navigation:** Includes tabs for `Events (878,561)`, `Patterns`, `Statistics`, and `Visualization`.
- Timeline:** A horizontal timeline bar at the bottom.

The total number of events in both the cases are same. So, we can conclude that all the events were loaded correctly

## 2. Verification for lookup table hotel\_info.csv

This is the lookup table and so it was converted to csv while pre-processing. We count the number of records in this file using the command:

```
# wc -l offering.csv
```

```

root@search-head:~# ls
abc.json          offering.csv
anaconda-ks.cfg   review.csv
bcc.txt           review.json
hotel_reviews.json review.txt
install.log       splunk-6.2.0-237341-linux-2.6-x86_64.rpm
install_log.syslog

[root@search-head ~]# wc -l offering.csv
4334 offering.csv
[root@search-head ~]#

```

Also, we can have screenshot from excel file.

4331	5	28970 http://www DC	2800 Penn	2000 / Washington hotel	84065 Four Seasons Washington D.C.
4332	4	28970 http://www DC	2121 P Str	20037 Washington hotel	84093 Palomar Washington DC, a Kimpton Hotel
4333	4.5	28970 http://www DC	806 15th S	20005 Washington hotel	235513 Sofitel Washington DC
4334	5	28970 http://www DC	16th & H St	20006 Washington hotel	84117 The Hay-Adams
4335					
4336					

Both of the above snapshots show one extra record as row for column names is also included in the file.

Let us verify if the lookup table was correctly uploaded to Splunk.

address_locality	address_postal_code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url
New York City	10036	NY	147 West 43rd Street	4	113317	Casablanca Hotel Times	60763	hotel	http://www.tripadv

The number of records match for both the files!

## 2. Verification of fields

### 1. review.txt

All the interesting fields collected from our data were properly loaded into splunk. The field used for timestamp is date. All the other fields are properly reflected.

```
> { [-]
  author: { [-]
    id: 631AAF84885D8AA48F4876100F92CEEB
    location: New York, New York, United States
    num_reviews: 1
    username: Danny K
  }
  date: December 20, 2012
  date_stayed: December 2012
  id: 147785077
  num_helpful_votes: 0
  offering_id: 120556
  ratings: { [-]
    cleanliness: 5
    location: 5
    overall: 5
    rooms: 5
    service: 5
    sleep_quality: 5
    value: 5
  }
  text: We arrived late at night and immediately were treated with everything we needed by Nikki and Adam. We were here for three nights and there wasn't anything we needed that the hotel wasn't able to help us with. They even provide a whole bunch of complimentary travel items you might forget (phone chargers, tooth brushes, contact solution etc.). They also have a very nice restaurant and bar in lobby which was very enjoyable to visit. I highly recommend the George hotel.
  title: "Unbelievable customer service"
  via_mobile: false
}
Show as raw text
```

**Interesting Fields**

```

a author.id 100+
a author.location 100+
a author.num_cities 100+
a author.num_helpful_votes 100+
a author.num_reviews 100+
a author.num_type_reviews 100+
a author.username 100+
a date 100+
# date_mday 31
a date_month 5
a date_stayed 100+
a date_wday 7
# date_year 1
a date_zone 1
a id 100+
a index 1
# linecount 1
a num_helpful_votes 56
a offering_id 100+
a punct 100+
a ratings.cleanliness 5
a ratings.location 5
a ratings.overall 5
a ratings.rooms 5
a ratings.service 5
a ratings.sleep_quality 5
a ratings.value 5
a splunk_server 1
a text 100+

```

**2. hotel\_info.csv**

To check if all the fields that are present in offering.csv are correctly loaded in Splunk, we first look at the fields in original file.

	A	B	C	D	E	F	G	H	I	J	K
1	hotel_class	region_id	url	address_region	address_street-address	address_postal-code	address_locality	type	hotel_id	name	
2		4	60763	http://w.NY	147 West 43rd Street		10036	New York City	hotel	113317	Casablanca Hotel 1

We also can have a look at the first line of this csv using command :

```
# head -n 1 offering.csv
```

```
root@search-head:~ [root@search-head ~]# head -n 1 offering.csv
hotel_class,region_id,url,address_region,address_street-address,address_postal-code,address_locality,type,hotel_id,name
[root@search-head ~]#
```

Let us check if all these fields are present in the uploaded lookup table in Splunk.

The screenshot shows a Splunk search interface. The search bar contains the query: | inputlookup hotel\_info.csv. Below the search bar, it says "0 events (before 11/19/14 7:20:51.000 PM)". Under the "Events" tab, there is one event displayed:

address_locality	address_postal-code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url
New York City	10036	NY	147 West 43rd Street	4	113317	Casablanca Hotel Times	60763	hotel	http://www.trin

All the fields are properly reflected in hotel\_info.csv.

### 3. Verification of data

We queried the data on some unique fields like 'id' and cross verified if the same data exists for the same file within the original file.

#### 1. review.txt

We checked the data for record with id = '147785077'

```
# sourcetype=reviews_final | search id=147785077
```

sourcetype=reviews\_final | search id=147785077

1 of 411,835 events matched

Events (1) Patterns Statistics Visualization

Format Timeline

Raw  20 Per Page

< Hide Fields	: All Fields	i Event
		<pre>&gt; { [-]     author: { [+]     }     date: December 20, 2012     date_stayed: December 2012     id: 147785077     num_helpful_votes: 0     offering_id: 120556     ratings: { [+]     }     text: We arrived late at night and immediately were treated with everything we needed by Nikki and Adam. We were here for three nights and there wasn't anything we needed that the hotel wasn't able to help us with. They even provide a whole bunch of complimentary travel items you might forget (phone chargers, tooth brushes, contact solution etc.). They also have a very nice restaurant and bar in lobby which was very enjoyable to visit. I highly recommend the George hotel.     title: "Unbelievable customer service"     via_mobile: false }</pre> <p>Show as raw text</p>

We compared this record with the record in original file.

```
# cat review.txt | grep '147785077'
```

```
[root@search-head ~]# cat review.txt | grep '147785077'
{"ratings": {"service": 5.0, "cleanliness": 5.0, "overall": 5.0, "value": 5.0, "location": 5.0, "sleep_quality": 5.0, "rooms": 5.0}, "title": "\u201cUnbelievable customer service\u201d", "text": "We arrived late at night and immediately were treated with everything we needed by Nikki and Adam. We were here for three nights and there wasn't anything we needed that the hotel wasn't able to help us with. They even provide a whole bunch of complimentary travel items you might forget (phone chargers, tooth brushes, contact solution etc.). They also have a very nice restaurant and bar in lobby which was very enjoyable to visit. I highly recommend the George hotel.", "author": {"username": "Danny K", "num_reviews": 1, "id": "631AAF84885D8AA48F4876100F92CEEB", "location": "New York, New York, United States"}, "date_stayed": "December 2012", "offering_id": 120556, "num_helpful_votes": 0, "date": "December 20, 2012", "id": 147785077, "via_mobile": false}
[root@search-head ~]#
```

Next, we verified the reviews posted by the author with username = 'illinitraveler714'. We found two matching events in Splunk.

```
# sourcetype=reviews_final | search "author.username"=illinitraveler714
```

The screenshot shows the Splunk interface with a search bar containing the query: `sourcetype=reviews_final | search "author.username"=illinitraveler714`. Below the search bar, it says "2 events (before 11/19/14 8:38:08.000 PM)". The main area displays two event logs:

```

{
  "-": {
    "author": "[+]",
    "date": "December 20, 2012",
    "date_stayed": "December 2012",
    "id": 147801697,
    "num_helpful_votes": 0,
    "offering_id": 223900,
    "ratings": "[+]",
    "text": "This hotel is close to the office, so it is a convenient location for work purposes. However, there are other Marriott's i that are closer to a more bustling night life. Location is 2-3 blocks north of the White House, which is nice. Bedding was clean and new, chairs were comfortable, bathroom was fine. Towels were a little old, which is always disappointing to n but they were clean. One item of note: the room contains a stove top and out of curiosity, I checked the drawers and was surprised they include a pretty nice pot, pan, some cooking utensils, plates, and glasses, etc. So for people who actually want to cook there pretty nice set up for a hotel."
  }
}

```

To cross-verify, we checked if these two records were the same in original file.

```
# cat review.txt | grep 'illinitraveler714'
```

```

root@search-head ~]# cat review.txt | grep 'illinitraveler714'
"ratings": {"service": 3.0, "cleanliness": 4.0, "overall": 3.0, "value": 3.0, "location": 3.0, "sleep_quality": 3.0, "rooms": 4.0}, "title": "\u201cClean hotel for extended stay travelers\u201d", "text": "This hotel is close to the office, so it is a convenient location for work purposes. However, there are other Marriott's in the area that are closer to a more bustling night life. Location is 2-3 blocks north of the White House, which is nice.\nBedding was clean and new, chairs were comfortable, bathroom was fine. Towels were a little old, which is always disappointing to me at hotels, but they were clean. One item of note: the room contains a stove top and out of curiosity, I checked the drawers and was surprised to find that they include a pretty nice pot, pan, some cooking utensils, plates, and glasses, etc. So for people who actually want to cook there, it's a pretty nice set up for a hotel.\nMy one complaint about this hotel is that I had to spend a lot of time in my room and the free wi-fi wasn't the greatest -- kept dropping my connection -- and immediately outside my window is another office, so I had to constantly be concerned with opening and closing drapes when changing or working. It was also pretty noisy in the early morning when the garbage trucks pick up dumpsters in the alley.", "author": {"username": "illinitraveler714", "num_cities": 3, "num_helpful_votes": 2, "num_reviews": 6, "num_type_reviews": 4, "id": "9246AOA32C3D55AD846390C8621F398B"}, "location": "Chicago, Illinois"}, "date_stayed": "December 2012", "offering_id": 223900, "num_helpful_votes": 0, "date": "December 20, 2012", "id": 147801697, "via_mobile": false}
{"ratings": {"service": 4.0, "cleanliness": 5.0, "overall": 4.0, "value": 4.0, "location": 5.0, "sleep_quality": 5.0, "rooms": 4.0}, "title": "\u201cFantastic, Clean Boutique Hotel in Central Location\u201d", "text": "I live in DC, and stayed here during a \"staycation\" for my anniversary with my husband a few weekends ago. We had an absolutely wonderful stay and I highly recommend this hotel to visitors! We were upgraded to a suite upon arrival, which was a standard room plus a nice chair off to the side and a desk area. The room was beautiful, the bed and linens were very fresh and clean, and the bathroom was beautiful and CLEAN. (Nothing is worse than dirt or mildew in bathrooms!!) I just cannot stay enough about how clean and beautiful the rooms were -- it knocked my socks off!\nThe following morning, we checked out and had breakfast at PostScript, which was very nice -- light, bright, tables at windows, with a great cup of tea and nice, refreshing menu options!\nFor area visitors, this hotel is in Downtown DC, 3 blocks from the McPherson Square Metro (White House/VA exit), which is on the Orange and Blue line -- connecting you to the airport, Virginia, GWU, Smithsonian, Capitol Hill, and Eastern Market. You're also 5 blocks from the White House in a nice neighborhood with plenty of restaurants in the area, or within a fast cab or metro ride!", "author": {"username": "illinitraveler714", "num_cities": 3, "num_helpful_votes": 2, "num_reviews": 6, "num_type_reviews": 4, "id": "9246AOA32C3D55AD846390C8621F398B"}, "location": "Chicago, Illinois"}, "date_stayed": "June 2012", "offering_id": 84122, "num_helpful_votes": 1, "date": "June 25, 2012", "id": 132748201, "via_mobile": false}
root@search-head ~]#

```

Next, we verified the reviews posted by the title = ' Clean hotel for extended stay travelers'. We found one matching event in Splunk.

```
# sourcetype=reviews_final | spath title | search title=" Clean hotel for extended stay travelers "
```

The screenshot shows the Splunk interface with the following details:

- Search Bar:** sourcetype=reviews\_final| spath title | search title=""Clean hotel for extended stay travelers""
- Results:** 1 event (before 11/19/14 8:57:39.000 PM)
- Event View:**
  - Selected Fields:** a\_host\_1, a\_source\_1, a\_sourcetype\_1
  - Interesting Fields:** a\_author\_id\_1, a\_author.location\_1, a\_author.num\_cities\_1, a\_author.num\_helpful\_votes\_1, a\_author.num\_reviews\_1, a\_author.num\_type\_reviews\_1, a\_author.username\_1, a\_date\_1, #date\_mday\_1, a\_date\_month\_1, a\_date\_stayed\_1, a\_date\_wday\_1
  - Event Content:**

```
i Event
> { [-]
  author: { [!]
    date: December 20, 2012
    date_stayed: December 2012
    id: 147801697
    num_helpful_votes: 0
    offering_id: 223900
    ratings: { [!]
      }
    text: This hotel is close to the office, so it is a convenient location for work purposes. However, there are other Marriott's in the area that are closer to a more bustling night life. Location is 2-3 blocks north of the White House, which is nice.\nBedding was clean and new, chairs were comfortable, bathroom was fine. Towels were a little old, which is always disappointing to me but they were clean. One item of note: the room contains a stove top and out of curiosity, I checked the drawers and was surprised they include a pretty nice pot, pan, some cooking utensils, plates, and glasses, etc. So for people who actually want to cook there pretty nice set up for a hotel.
    My one complaint about this hotel is that I had to spend a lot of time in my room and the free wi-fi wasn't the greatest -- kept dropping my connection -- and immediately outside my window is another office, so I had to constantly be concerned with opening and closing drapes when changing or working. It was also pretty noisy in the early morning when the garbage trucks pick up dumpsters in the alley.
    title: "Clean hotel for extended stay travelers"
    via_mobile: false
  }
}
```

To cross-verify, we checked if these two records were the same in original file.

```
[root@search-head ~]# cat review.txt | grep 'Clean hotel for extended stay travelers'
[{"ratings": {"service": 3.0, "cleanliness": 4.0, "overall": 3.0, "value": 3.0, "location": 3.0, "rooms": 4.0}, "title": "\u201cClean hotel for extended stay travelers\u201d", "text": "This hotel is close to the office, so it is a convenient location for work purposes. However, there are other Marriott's in the area that are closer to a more bustling night life. Location is 2-3 blocks north of the White House, which is nice.\nBedding was clean and new, chairs were comfortable, bathroom was fine. Towels were a little old, which is always disappointing to me but they were clean. One item of note: the room contains a stove top and out of curiosity, I checked the drawers and was surprised they include a pretty nice pot, pan, some cooking utensils, plates, and glasses, etc. So for people who actually want to cook there pretty nice set up for a hotel.\nMy one complaint about this hotel is that I had to spend a lot of time in my room and the free wi-fi wasn't the greatest -- kept dropping my connection -- and immediately outside my window is another office, so I had to constantly be concerned with opening and closing drapes when changing or working. It was also pretty noisy in the early morning when the garbage trucks pick up dumpsters in the alley.", "author": {"username": "slinatraveler714", "num_cities": 3, "num_helpful_votes": 2, "num_reviews": 6, "num_type_reviews": 4, "id": "9246AOA32C3D55AD846390C8621F398B", "location": "Chicago, Illinois"}, "date_stayed": "December 2012", "offering_id": 223900, "num_helpful_votes": 0, "date": "December 20, 2012", "id": 147801697, "via_mobile": false}
[root@search-head ~]#
```

## 2. hotel\_info.csv

We checked the record with hotel id = '113317' in Splunk :

# | inputlookup hotel\_info.csv | search hotel\_id = 113317

The screenshot shows the Splunk interface with the following details:

- Search Bar:** | inputlookup hotel\_info.csv | search hotel\_id = 113317
- Results:** 0 events (before 11/19/14 7:32:05.000 PM)
- Table View:**

address_locality	address_postal_code	address_region	address_street_address	hotel_class	hotel_id	name	region_id	type	url
New York City	10036	NY	147 West 43rd Street	4	113317	Casablanca Hotel Times Square	60763	hotel	<a href="http://www.tripadvisor.com/Hotel_Review-g60763-h113317-Casablanca_Hotel_Times_Square-New_York_City.html">http://www.tripadvisor.com/Hotel_Review-g60763-h113317-Casablanca_Hotel_Times_Square-New_York_City.html</a>

We verified if this was correct by checking it against the csv file.

	A	B	C	D	E	F	G	H	I	J
1	hotel_class	region_id	url	address_region	address_street-address	address_postal-code	address_locality	type	hotel_id	name
2	4	60763	http://w NY		147 West 43rd Street	10036	New York City	hotel	113317	Casablanca Hotel Times Square
3	5	32655	http://w CA		300 S Doheny Dr	90048	Los Angeles	hotel	76049	Four Seasons Hotel Los Angeles at Beverly Hills
4	3.5	60763	http://w NY		790 Eighth Avenue	10019	New York City	hotel	99352	Hilton Garden Inn Times Square
5	4	60763	http://w NY		152 West 51st Street	10019	New York City	hotel	93589	The Michelangelo Hotel

For record with hotel\_id = '72598'

```
# | inputlookup hotel_info.csv | search hotel_id = 72598
```

address_locality	address_postal-code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url
Seattle	98133	WA	13700 Aurora Ave N	2.5	72598	Comfort Inn & Suites Seattle	60878	hotel	http://www.tripadvisor.com/Hotel_Review-Comfort_Inn_Suites_Seattle-Seattle_Wash...

```
[root@search-head ~]# cat offering.csv | grep 72598
2.5,60878,http://www.tripadvisor.com/Hotel_Review-g60878-d72598-Reviews-Comfort_Inn_Suites_Seattle-Seattle_Washington.html,WA,13700 Aurora Ave N,98133,Seattle,hotel,72598,Comfort Inn & Suites Seattle
[root@search-head ~]#
```

### 3. Check if lookup table works correctly with the table review.txt.

We executed a query to see if the number of the events with the hotel name are same as the count corresponding to hotel name.

```
# *| stats count by hotel_name
```

The screenshot shows the Splunk Search & Reporting interface. A search query `\*| stats count by hotel\_name` has been run, resulting in 878,561 events. The results table lists various hotel names with their counts. The row for "Days Inn Columbus Airport" is highlighted with a red border.

hotel_name	count
Days Inn Central	36
Days Inn Charlotte / Sunset Road	31
Days Inn Charlotte North - Speedway - UNCC - Research Park	28
Days Inn Chicago	518
Days Inn Columbus Airport	4
Days Inn Columbus Fairgrounds	70
Days Inn Columbus North	59

## Find the offering

The screenshot shows the Splunk Search & Reporting interface again, this time searching for an offering ID. The results table displays event details, including the full event JSON. One event is shown in expanded view:

```

{
  "Time": "12/9/12 12:00:00.000 AM",
  "Event": {
    "author": {
      ...
    },
    "date": "December 9, 2012",
    "date_stayed": "December 2012",
    "id": 147083553,
    "num_helpful_votes": 0,
    "offering_id": 3574675,
    "ratings": {
      ...
    },
    "text": "We were looking for a place to stay close to the airport because we had an early flight the next morning. The price for this hotel was certainly right ($44/night!) and included breakfast and parking for the duration of your trip.\nWe had trouble finding the property because the sign was not lit. When we pulled into the parking lot the hotel looked abandoned! Very few cars and all windows blacked out. Almost didn't stay there but we were tired and it was late. This property is under new management and extensive renovation. It was a Holiday Inn (and from the inside, a very good one) and recently purchased by the current owner. We were warmly welcomed and our room was satisfactory. Breakfast began at 7 but we asked if we could get \"Something\" at 6 because of the flight. When we went downstairs, the variety of food was a nice surprise (including freshly brewed coffee). We were taken to the airport exactly as promised. When we came back into town, the hotel didn't have anyone to drive the van but told us to take a \"blue cab\" and they"
  }
}

```

On comparing it with the original file, the corresponding name matched with the hotel\_id

The screenshot shows a terminal window with the command `cat offering.csv | grep '3574675'`. The output shows a row of data, with the last part of the row for "Days Inn Columbus Airport" highlighted with a red box.

```

root@search-head:~#
login as: root
root@216.121.58.234's password:
Last login: Wed Nov 19 20:41:46 2014 from c-24-4-154-19.hsd1.ca.comcast.net
[root@search-head ~]# cat offering.csv | grep '3574675'
3,50226,http://www.tripadvisor.com/Hotel_Review-g50226-d3574675-Reviews-Days_Inn_Columbus_Airport-Columbus_Ohio.html,OH,750 Stelzer Road,43219,Columbus,hotel,3574675,
Days Inn Columbus Airport
[root@search-head ~]#

```

## Phase 2 - Building and Verifying Queries

### 1. Number of reviews for each year

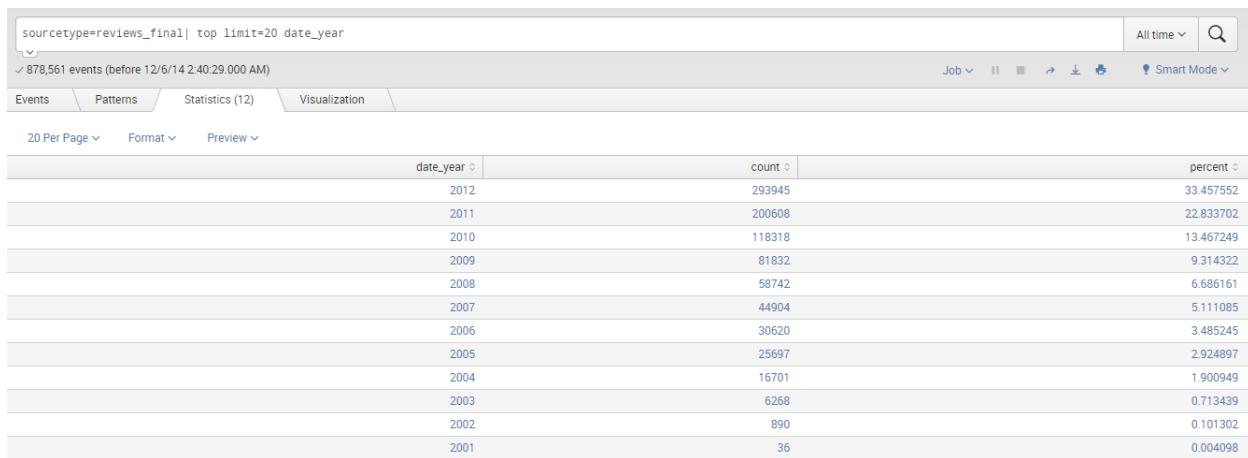
**Objective of the query:** To count the total number of reviews that were made in each year.

#### Description of the query :

Our dataset consists of reviews from years 2001 to 2012. We wanted to know the count of reviews in each year.

#### Query :

```
# sourcetype=reviews_final| top limit=20 date_year
```



The screenshot shows the Kibana interface with a search bar containing the query "# sourcetype=reviews\_final| top limit=20 date\_year". Below the search bar, it says "878,561 events (before 12/6/14 2:40:29 000 AM)". The interface includes tabs for Events, Patterns, Statistics (12), and Visualization. Under Events, there are filters for 20 Per Page, Format, and Preview. The main area displays a table with three columns: date\_year, count, and percent. The data shows the following distribution of reviews by year:

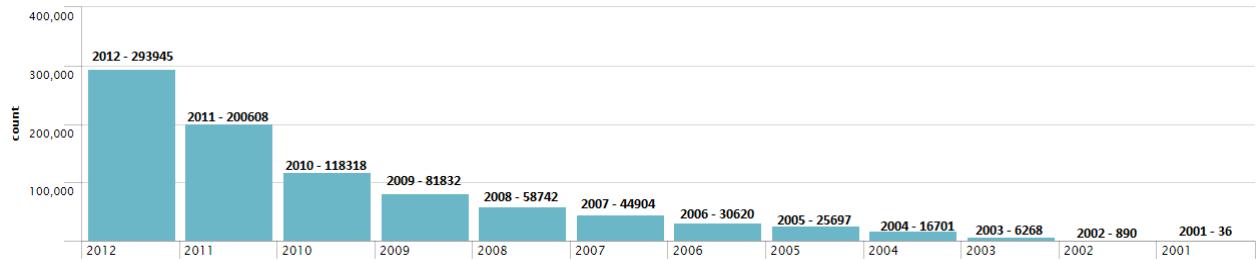
date_year	count	percent
2012	293945	33.457552
2011	200608	22.833702
2010	118318	13.467249
2009	81832	9.314322
2008	58742	6.686161
2007	44904	5.111085
2006	30620	3.485245
2005	25697	2.924897
2004	16701	1.900949
2003	6268	0.713439
2002	890	0.101302
2001	36	0.004098

**Verification:** We verified this data corresponding to year 2012.

```
# sourcetype=reviews_final | search date_year = 2012
```



## Visualization



## 2. State wise count for reviews

**Objective of the query :** To find top 5 states for with maximum number of reviews.

**Description of the query :**

The hotels in our dataset our spread over 17 states. We wanted to find the top five states with highest number of reviews that are recorded.

**Query :**

```
# sourcetype="reviews_final" | top limit=5 hotel_state
```

hotel_state	count	percent
NY	267057	30.397093
CA	210264	23.932772
TX	88559	10.080006
IL	64531	7.345079
DC	48337	5.501838

**Verification:** We verified the data for the state of NY.

```
# sourcetype="reviews_final" hotel_state=NY
```

## Visualization



### 3. City wise count for reviews

**Objective of the query :** To find top 5 cities for with maximum number of reviews.

**Description of the query :**

The hotels in our dataset our spread over multiple cities. We wanted to find the top five cities with highest number of reviews that are recorded.

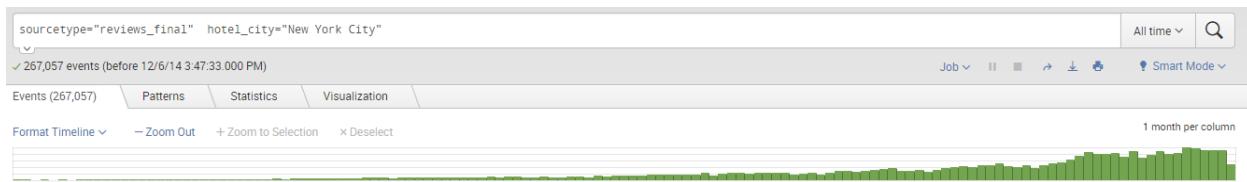
**Query :**

```
# sourcetype="reviews_final" | top limit=5 hotel_city
```

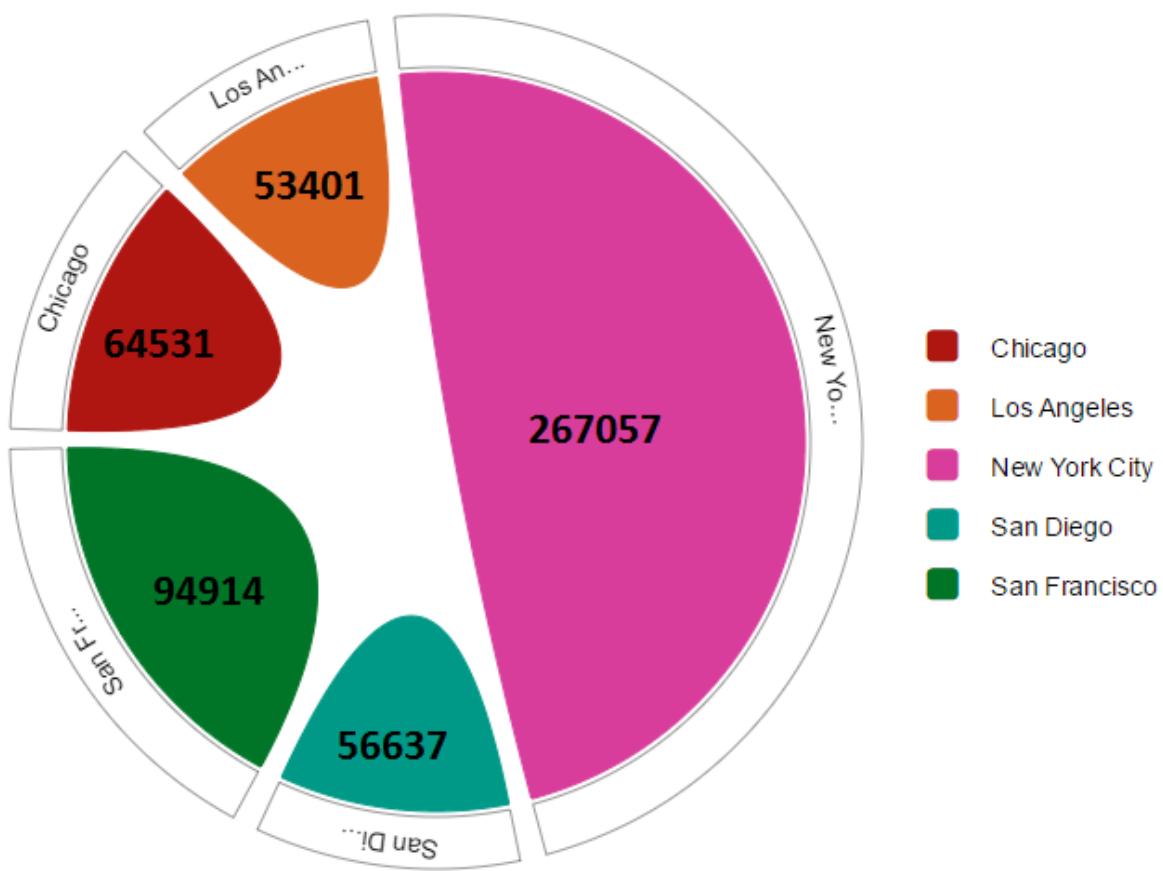
sourcetype="reviews_final"   top limit=5 hotel_city		All time
✓ 878,561 events (before 12/6/14 3:44:44.000 PM)		Smart Mode
Events	Patterns	Statistics (5)
20 Per Page	Format	Preview
hotel_city	count	percent
New York City	267057	30.397093
San Francisco	94914	10.803348
Chicago	64531	7.345079
San Diego	56637	6.446564
Los Angeles	53401	6.078235

**Verification:** We verified the data for the New York City.

```
# sourcetype="reviews_final" hotel_city="New York City"
```



## Visualization



From the above two queries, for top five states and top five cities, we can see that the statistics for NY state and New York City are same. The number of reviews for both of them are 267057. So, to be sure that this statistics are correct, we verified it using some queries which are as follows :

### Query 1 :

```
# | inputlookup hotel_info.csv | search address_locality="New York City"
```

### Result :

inputlookup hotel_info.csv   search address_locality="New York City"										All time	Smart Mode	
Events Patterns Statistics (435) Visualization												
20 Per Page Format Preview										< Prev 1 2 3 4 5 6 7 8 9 ... Next >		
address_locality	address_postal-code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url			
New York City	10036	NY	147 West 43rd Street	4	113317	Casablanca Hotel Times Square	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d113317-Reviews-Casablanca_Hotel_Times_Square-New_York_City_New_York.html			
New York City	10019	NY	790 Eighth Avenue	3.5	99352	Hilton Garden Inn Times Square	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d99352-Reviews-Hilton_Garden_Inn_Times_Square-New_York_City_New_York.html			
New York City	10019	NY	152 West 51st Street	4	93589	The Michelangelo Hotel	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d93589-Reviews-The_Michelangelo_Hotel-New_York_City_New_York.html			
New York City	10036	NY	130 West 46th Street	4	217616	The Muse Hotel New York	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d217616-Reviews-The_Muse_Hotel_New_York-New_York_City_New_York.html			
New York City	10036	NY	45 West 44th Street	4.5	208454	Sofitel New York	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d208454-Reviews-Sofitel_New_York-New_York_City_New_York.html			

## Query 2 :

```
# | inputlookup hotel_info.csv | search address_region="NY"
```

## Result :

inputlookup hotel_info.csv   search address_region="NY"										All time	Smart Mode	
Events Patterns Statistics (435) Visualization												
20 Per Page Format Preview										< Prev 1 2 3 4 5 6 7 8 9 ... Next >		
address_locality	address_postal-code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url			
New York City	10036	NY	147 West 43rd Street	4	113317	Casablanca Hotel Times Square	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d113317-Reviews-Casablanca_Hotel_Times_Square-New_York_City_New_York.html			
New York City	10019	NY	790 Eighth Avenue	3.5	99352	Hilton Garden Inn Times Square	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d99352-Reviews-Hilton_Garden_Inn_Times_Square-New_York_City_New_York.html			
New York City	10019	NY	152 West 51st Street	4	93589	The Michelangelo Hotel	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d93589-Reviews-The_Michelangelo_Hotel-New_York_City_New_York.html			
New York City	10036	NY	130 West 46th Street	4	217616	The Muse Hotel New York	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d217616-Reviews-The_Muse_Hotel_New_York-New_York_City_New_York.html			
New York City	10036	NY	45 West 44th Street	4.5	208454	Sofitel New York	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d208454-Reviews-Sofitel_New_York-New_York_City_New_York.html			

## Query 3 :

```
# sourcetype="reviews_final" | stats dc(hotel_city) by hotel_state | where hotel_state="NY"
```

## Result :

sourcetype="reviews_final"   stats dc(hotel_city) by hotel_state   where hotel_state="NY"										All time	Smart Mode	
Events Patterns Statistics (1) Visualization												
20 Per Page Format Preview										< Prev 1 2 3 4 5 6 7 8 9 ... Next >		
hotel_state	dc(hotel_city)											
NY	1											

## 4. To find the number of cities for which reviews are recorded in each state

**Objective of the query :** To find the number of cities in each state whose hotels are reviewed.

## Description of the query :

From the statistic of New York City and NY state, we were curious to find if there are any other states for whom hotel from only one city are reviewed.

### Query:

```
# sourcetype="reviews_final" |stats dc(hotel_city) by hotel_state
```

The screenshot shows a Splunk search interface with the following command in the search bar:

```
sourcetype="reviews_final" |stats dc(hotel_city) by hotel_state
```

Below the search bar, it says "878,561 events (before 12/6/14 5:21:47.000 PM)". The interface includes tabs for Events, Patterns, Statistics (17), and Visualization. Under Statistics, the table is displayed:

hotel_state	dc(hotel_city)
TX	6
CA	4
AZ	1
CO	1
DC	1
FL	1
IL	1
IN	1
MA	1
MD	1
MI	1
NC	1
NY	1
OH	1
PA	1

### Verification Query 1:

```
# | inputlookup hotel_info.csv | dedup address_locality | where address_region="CA"
```

The screenshot shows a Splunk search interface with the following command in the search bar:

```
| inputlookup hotel_info.csv | dedup address_locality | where address_region="CA"
```

Below the search bar, it says "0 events (before 12/6/14 5:38:03.000 PM)". The interface includes tabs for Events, Patterns, Statistics (4), and Visualization. Under Statistics, the table is displayed:

address_locality	address_postal-code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url
Los Angeles	90048	CA	300 S Doheny Dr	5	76049	Four Seasons Hotel Los Angeles at Beverly Hills	32655	hotel	<a href="http://www.tripadvisor.com/Hotel_Review-g32655-d76049-Reviews-Four_Seasons_Hotel_Los_Angeles_at_Beverly_Hills-Los_Angeles_California.html">http://www.tripadvisor.com/Hotel_Review-g32655-d76049-Reviews-Four_Seasons_Hotel_Los_Angeles_at_Beverly_Hills-Los_Angeles_California.html</a>
San Diego	92108	CA	1865 Hotel Circle South	3	268157	Residence Inn San Diego Mission Valley	60750	hotel	<a href="http://www.tripadvisor.com/Hotel_Review-g60750-d268157-Reviews-Residence_Inn_San_Diego_Mission_Valley-San_Diego_California.html">http://www.tripadvisor.com/Hotel_Review-g60750-d268157-Reviews-Residence_Inn_San_Diego_Mission_Valley-San_Diego_California.html</a>
San Jose	95136	CA	200 Edenvale Avenue	4	123562	Dolce Hayes Mansion	33020	hotel	<a href="http://www.tripadvisor.com/Hotel_Review-g33020-d123562-Reviews-Dolce_Hayes_Mansion-San_Jose_California.html">http://www.tripadvisor.com/Hotel_Review-g33020-d123562-Reviews-Dolce_Hayes_Mansion-San_Jose_California.html</a>
San Francisco	94133	CA	1075 Columbus Ave		119657	Columbus Motor Inn	60713	hotel	<a href="http://www.tripadvisor.com/Hotel_Review-g60713-d119657-Reviews-Columbus_Motor_Inn-San_Francisco_California.html">http://www.tripadvisor.com/Hotel_Review-g60713-d119657-Reviews-Columbus_Motor_Inn-San_Francisco_California.html</a>

### Verification Query 2:

```
# | inputlookup hotel_info.csv | dedup address_locality | where address_region="TX"
```

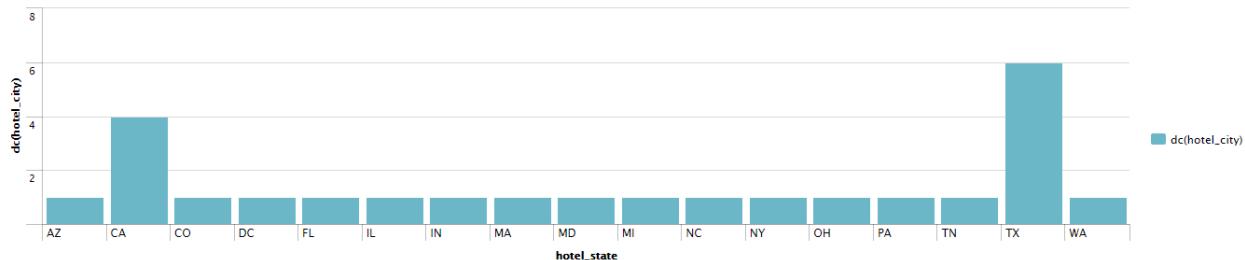
inputlookup hotel_info.csv   dedup address_locality   where address_region="TX"											All time	Smart Mode
Events Patterns Statistics (6) Visualization												
20 Per Page Format Preview												
address_locality	address_postal-code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url			
Houston	77042	TX	10609 Westpark Drive	3	1866401	Holiday Inn Hotel-Houston Westchase	56003	hotel	http://www.tripadvisor.com/Hotel_Review-g56003-d1866401-Reviews-Holiday_Inn_Hotel_Houston_Westchase-Houston_Texas.html			
San Antonio	78227	TX	2327 Pinn Road	2	2039194	Pinn Road Inn & Suites	60956	hotel	http://www.tripadvisor.com/Hotel_Review-g60956-d2039194-Reviews-Pinn_Road_Inn_Suites-San_Antonio_Texas.html			
Dallas	75245	TX	14925 Landmark Boulevard	2.5	109571	La Quinta Inn & Suites Dallas Addison Galleria	55711	hotel	http://www.tripadvisor.com/Hotel_Review-g55711-d109571-Reviews-La_Quinta_Inn_Suites_Dallas_Addison_Galleria-Dallas_Texas.html			
Austin	78735	TX	8212 Barton Club Drive	4	181851	Barton Creek Resort & Spa	30196	hotel	http://www.tripadvisor.com/Hotel_Review-g30196-d181851-Reviews-Barton_Creek_Resort_Spa-Austin_Texas.html			
Fort Worth	76102	TX	601 Main Street	3	223888	Courtyard by Marriott Fort Worth Downtown/Blackstone	55857	hotel	http://www.tripadvisor.com/Hotel_Review-g55857-d223888-Reviews-Courtyard_by_Marriott_Fort_Worth_Downtown_Blackstone-Fort_Worth_Texas.html			
El Paso	79925	TX	6645 Gateway West	2.5	2351689	Comfort Inn & Suites I-10 Airport	60768	hotel	http://www.tripadvisor.com/Hotel_Review-g60768-d2351689-Reviews-Comfort_Inn_Suites_I_10_Airport-El_Paso_Texas.html			

### Verification Query 3:

```
# | inputlookup hotel_info.csv | dedup address_locality | where address_region="AZ"
```

inputlookup hotel_info.csv   dedup address_locality   where address_region="AZ"											All time	Smart Mode
Events Patterns Statistics (1) Visualization												
20 Per Page Format Preview												
address_locality	address_postal-code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url			
Phoenix	85018	AZ	5200 East Camelback Road	5	73943	Royal Palms Resort and Spa	31310	hotel	http://www.tripadvisor.com/Hotel_Review-g31310-d73943-Reviews-Royal_Palms_Resort_and_Spa-Phoenix_Arizona.html			

### Visualization



### 5. The region that has the most number of hotel reviews according to postal code

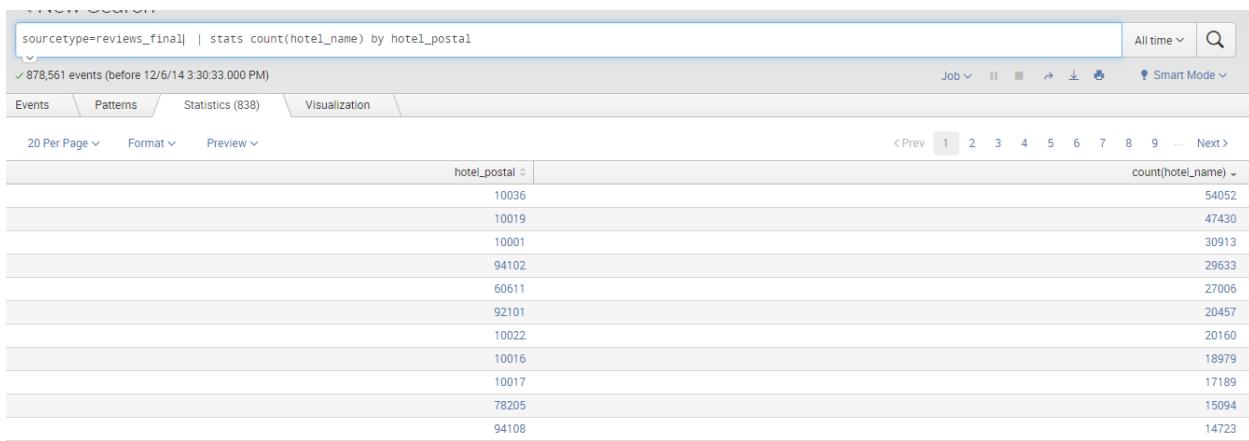
**Objective of the query :** To find the postal regions for which the highest number of reviews are recorded.

### Description of the query :

Every state is divided into multiple postal codes. The objective of this query is to find the postal region for which highest number of reviews are recorded.

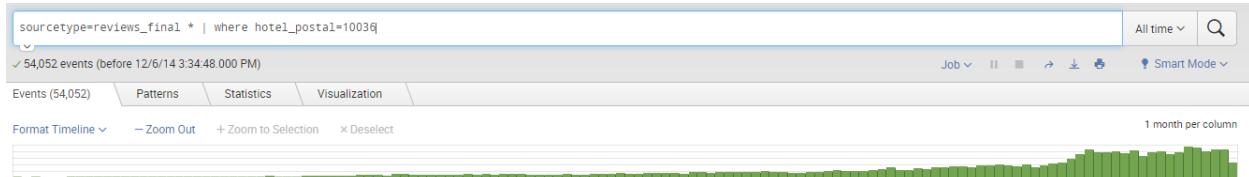
### Query :

```
# sourcetype=reviews_final | stats count(hotel_name) by hotel_postal
```

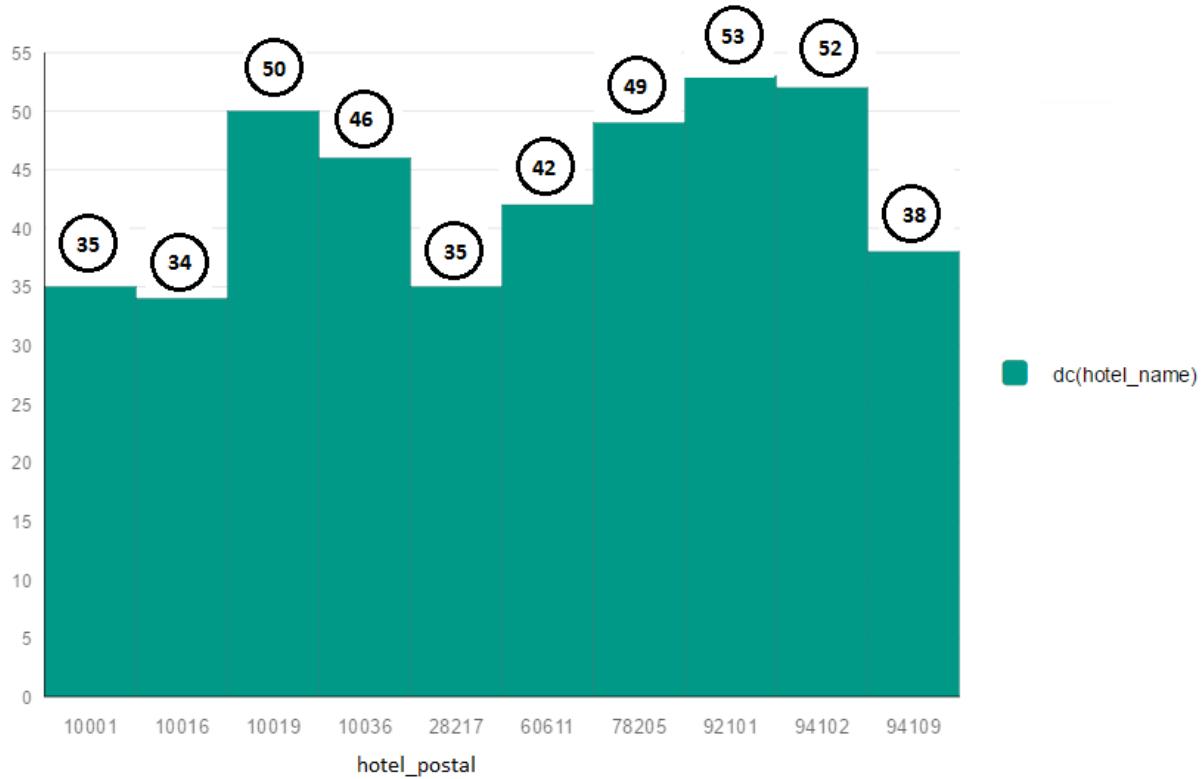


**Verification :** The total number of reviews for hotels having postal code as 10036 were verified.

```
# sourcetype=reviews_final * | where hotel_postal=10036
```



## Visualization

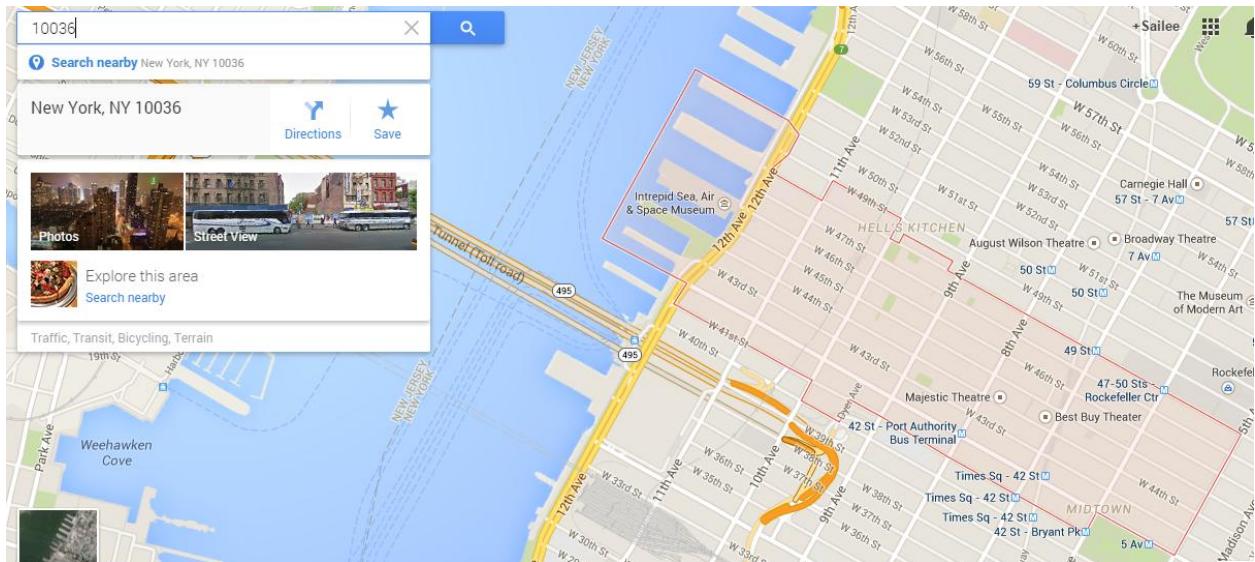


There are 46 hotels that have this postal code :

```
# |inputlookup hotel_info.csv | search "adress_postal-code"=10036
```

inputlookup hotel_info.csv   search "adress_postal-code"=10036										All time	Smart Mode					
Events		Patterns		Statistics (46)				Visualization								
20 Per Page		Format		Preview								< Prev	1	2	3	Next >
address_locality	address_postal-code	address_region	address_street-address	hotel_class	hotel_id	name	region_id	type	url							
New York City	10036	NY	147 West 43rd Street	4	113317	Casablanca Hotel Times Square	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d113317-Reviews-Casablanca_Hotel_Times_Square-New_York_City_New_York.html							
New York City	10036	NY	130 West 46th Street	4	217616	The Muse Hotel New York	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d217616-Reviews-The_Muse_Hotel_New_York_New_York_City_New_York.html							
New York City	10036	NY	45 West 44th Street	4.5	208454	Sofitel New York	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d208454-Reviews-Sofitel_New_York_New_York_City_New_York.html							
New York City	10036	NY	49 W 44th St	4	93396	The Iroquois	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d93396-Reviews-The_Iroquois-New_York_City_New_York.html							
New York City	10036	NY	130 West 44th Street	5	1641016	The Chatwal	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d1641016-Reviews-The_Chatwal-New_York_City_New_York.html							
New York City	10036	NY	300 W 44th Street	4.5	1646128	InterContinental New York Times Square	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d1646128-Reviews-InterContinental_New_York_Times_Square-New_York_City_New_York.html							
New York City	10036	NY	228 West 47th Street	3	93437	Edison Hotel Times Square	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d93437-Reviews-Edison_Hotel_Times_Square-New_York_City_New_York.html							
New York City	10036	NY	157 West 47 Street	4	93376	Stay.	60763	hotel	http://www.tripadvisor.com/Hotel_Review-g60763-d93376-Reviews-Stay-New_York_City_New_York.html							

This is the postal code for city of New York.



## 6. Month that has the most number of reviews in all the years

**Objective of the query :** To find the month for which the total number of reviews recorded are highest in all the years.

### Description of the query:

The intent of this query is to find which month has the highest total count for number of reviews recorded.

### Query :

```
# sourcetype=reviews_final | eval month=strftime(_time,"%m")| stats count as offering_id by month | sort date_count desc
```

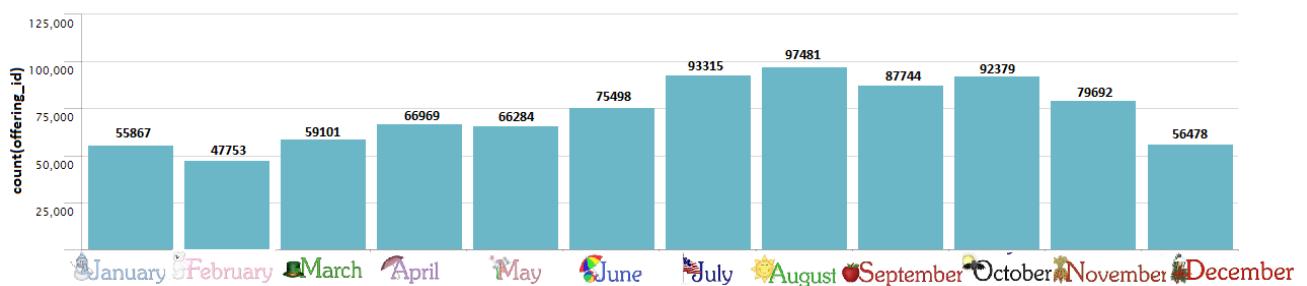
sourcetype=reviews_final   eval month=strftime(_time,"%m")  stats count as offering_id by month   sort date_count desc		All time ▾	Smart Mode ▾
878,561 events (before 12/6/14 7:49:37.000 PM)			
Events		Patterns	Statistics (12)
Visualization		Job ▾	II ■ ▶ ↻ ↺ ↻ ↺ Smart Mode ▾
20 Per Page ▾	Format ▾	Preview ▾	
month	offering_id		
08	97481		
07	93315		
10	92379		
09	87744		
11	79692		
06	75498		
04	66969		
05	66284		
03	59101		
12	56478		
01	55867		
02	47753		

**Verification:** We verified this for the month of August.

```
# sourcetype=reviews_final date_month=august
```



## Visualization



## 7. Month wise count of reviews for years 2010 to 2012

**Objective of the query :** To find the number of reviews recorded for each month for the years 2010, 2011 and 2012.

**Description of the query :**

To know if there was any specific month in each year where most reviews were recorded.

**For year 2010:**

```
# sourcetype=reviews_final | eval year=strftime(_time,"%y") | search year=10 | eval month=strftime(_time, "%m") | stats count as number_of_reviews by year,month
```

sourcetype=reviews\_final | eval year=strftime(\_time,"%y") | search year=10 | eval month=strftime(\_time, "%m") | stats count as number\_of\_reviews by year,month

All time  Smart Mode

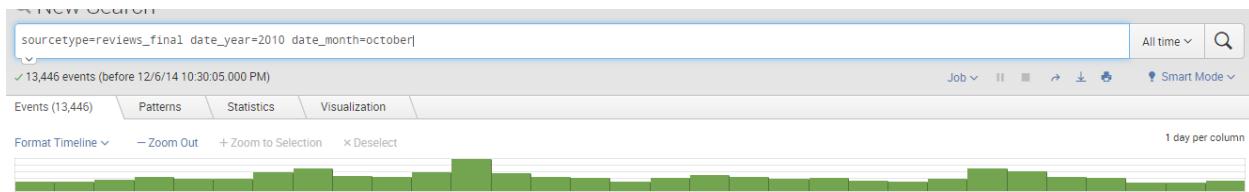
Events Patterns Statistics (12) Visualization

20 Per Page Format Preview

year	month	number_of_reviews
10	10	13446
10	08	13362
10	07	12029
10	09	11610
10	06	10824
10	11	10117
10	04	8736
10	05	8735
10	12	8276
10	03	7796
10	01	7026
10	02	6361

## Verification :

```
# sourcetype=reviews_final date_year=2010 date_month=october
```



## For year 2011:

```
# sourcetype=reviews_final | eval year=strftime(_time,"%y") | search year=11 | eval month=strftime(_time, "%m") | stats count as number_of_reviews by year,month
```

sourcetype=reviews\_final | eval year=strftime(\_time,"%y") | search year=11 | eval month=strftime(\_time, "%m") | stats count as number\_of\_reviews by year,month

All time  Smart Mode

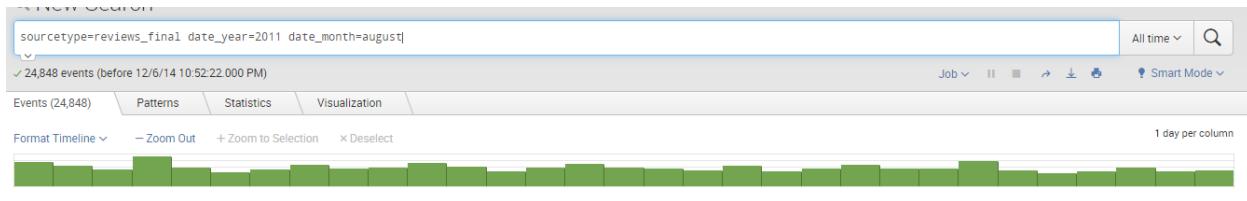
Events Patterns Statistics (12) Visualization

20 Per Page Format Preview

year	month	number_of_reviews
11	08	24848
11	09	23350
11	07	22054
11	10	21280
11	11	20488
11	06	17155
11	12	16281
11	05	12888
11	04	11670
11	03	11114
11	01	10365
11	02	9115

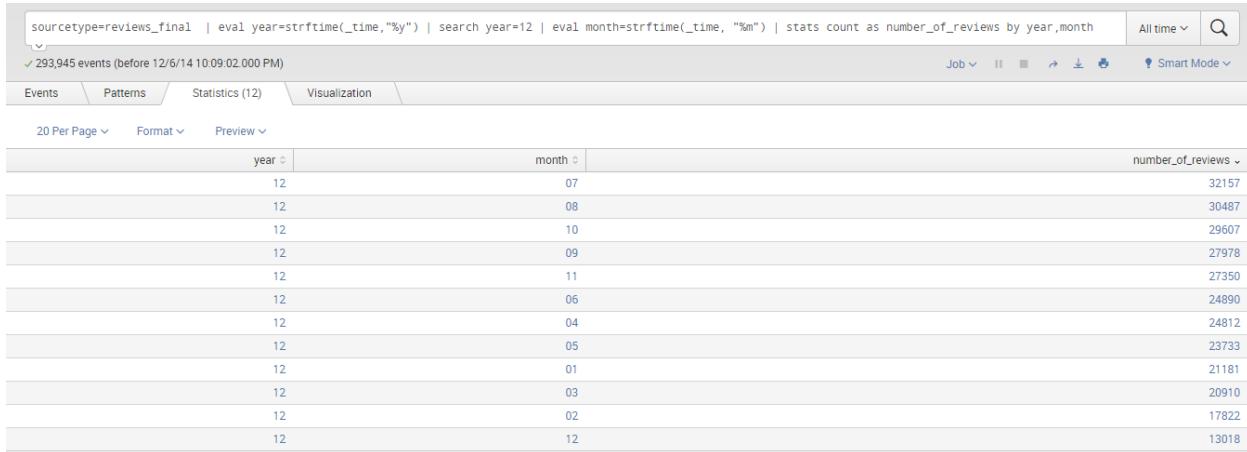
## Verification:

```
# sourcetype=reviews_final date_year=2011 date_month=august
```



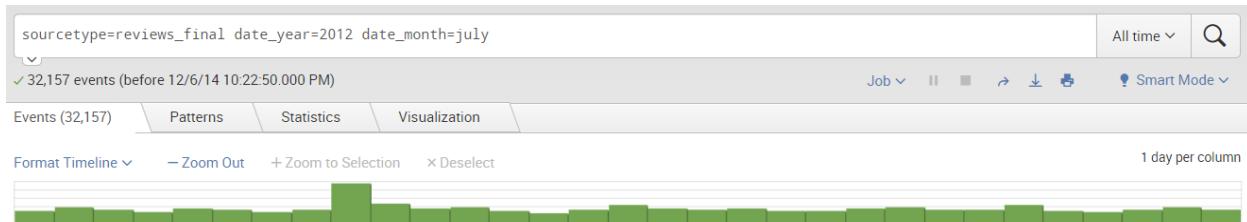
## For year 2012:

```
# sourcetype=reviews_final | eval year=strftime(_time,"%y") | search year=12 | eval month=strftime(_time, "%m") | stats count as number_of_reviews by year,month
```

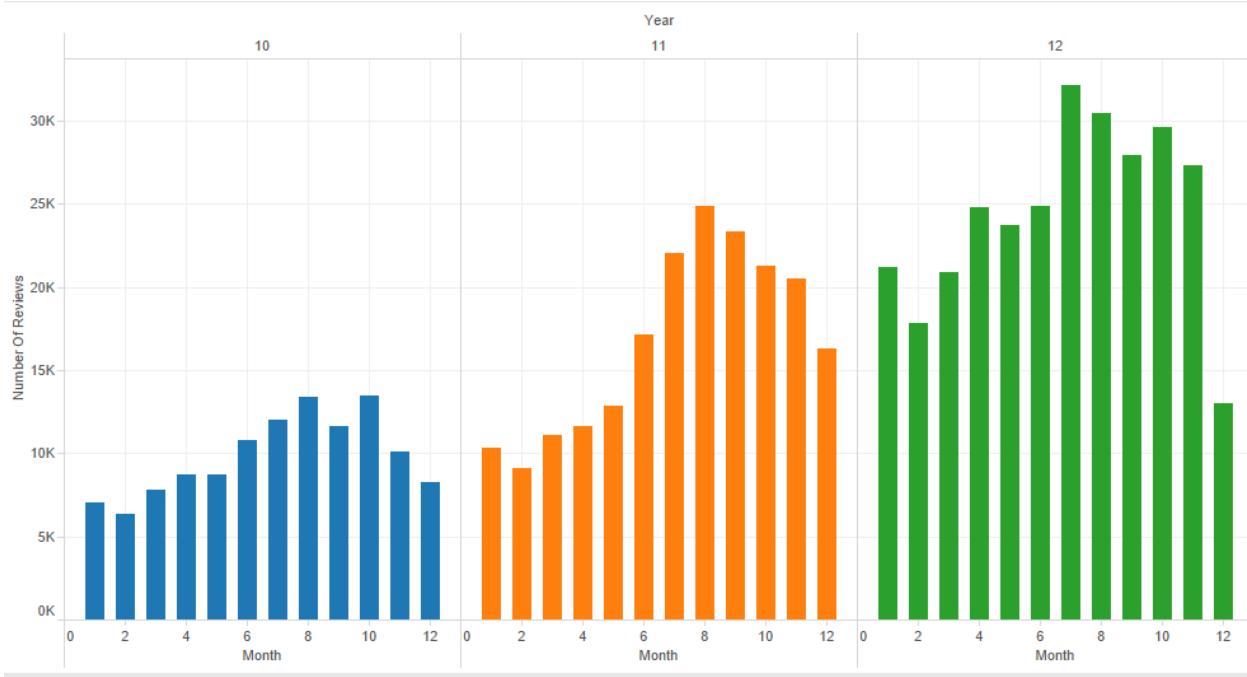


## Verification:

```
# sourcetype=reviews_final date_year=2012 date_month=july
```



## Visualization



## Result :

The most number of reviews in each year are recorded for the months of July, August, September and October.

## 8. Top 10 hotels that have the most reviews

**Objective of the query:** To find the hotels that have the most number of reviews.

**Description of the query:**

The intent of this query is to find the hotels that have the highest number of reviews from the customers.

**Query:**

```
# sourcetype=reviews_final* | stats count(id) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | sort limit=5 -count(id) | fields name, count(id) | rename count(id) as number_reviews
```

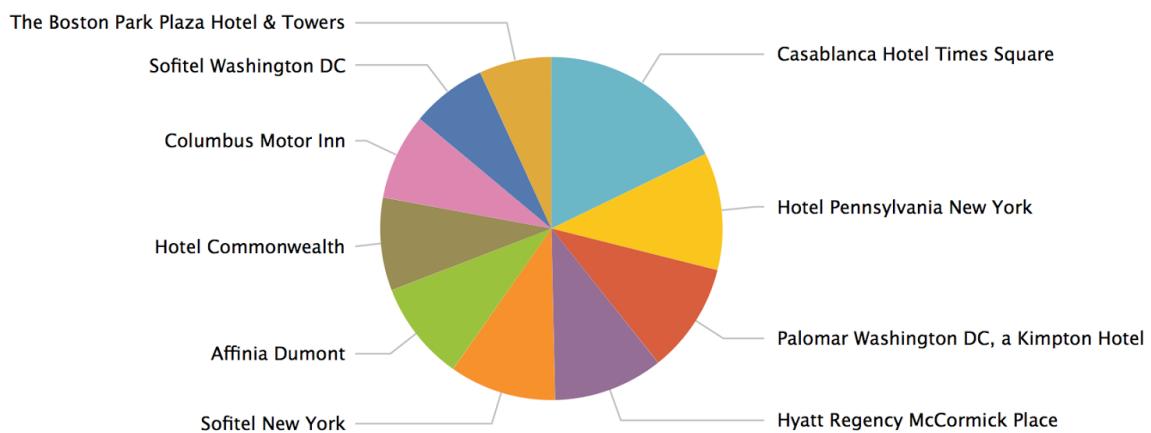
sourcetype=reviews_final*   stats count(id) by offering_id   lookup hotel_info.csv hotel_id as offering_id output name   sort limit=10 -count(id)   fields name, count(id)   rename count(id) as number_reviews		All time
✓ 878,561 events (before 12/7/14 12:19:36.000 AM)		Job Smart Mode
<a href="#">Events</a> <a href="#">Patterns</a> <a href="#">Statistics (10)</a> <a href="#">Visualization</a>		
20 Per Page <a href="#">Format</a> <a href="#">Preview</a>		
name	number_reviews	
Hotel Pennsylvania New York	21824	
Park Central	16036	
The New Yorker Hotel	14904	
Waldorf Astoria New York	14136	
Hudson New York	13540	
The Roosevelt Hotel	12872	
Affinia Manhattan	12680	
Edison Hotel Times Square	12136	
Hilton New York	12016	
Sofitel New York	11592	

## Verification:

```
# sourcetype=reviews_final* | stats count(id) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | search name = "Hotel Pennsylvania New York"
```

sourcetype=reviews_final*   stats count(id) by offering_id   lookup hotel_info.csv hotel_id as offering_id output name   search name = "Hotel Pennsylvania New York"		
✓ 878,561 events (before 12/7/14 12:22:23.000 AM)		
<a href="#">Events</a> <a href="#">Patterns</a> <a href="#">Statistics (1)</a> <a href="#">Visualization</a>		
20 Per Page <a href="#">Format</a> <a href="#">Preview</a>		
offering_id	count(id)	name
214197	21824	Hotel Pennsylvania New York

## Visualization



## 9. Top 10 hotels that have the most positive reviews(2.5 + on overall rating)

**Objective of the query:** To find the hotels that have the most number of positive reviews.

### Description of the query:

The intent of this query is to find the hotels that have the highest number of positive reviews from the customers.

### Query:

```
# sourcetype=reviews_final* | stats count(id) avg(ratings.overall) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | rename avg(ratings.overall) as average_rating | search average_rating > 2.5|sort limit=10 -count(id) | fields name, count(id) | rename count(id) as number_reviews
```

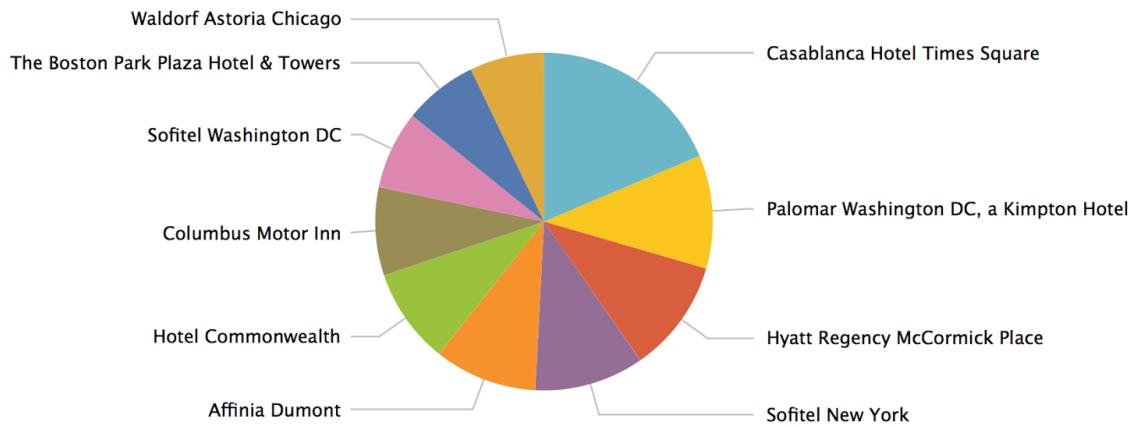
name	number_reviews
Park Central	16036
The New Yorker Hotel	14904
Waldorf Astoria New York	14136
Hudson New York	13540
The Roosevelt Hotel	12872
Affinia Manhattan	12680
Edison Hotel Times Square	12136
Hilton New York	12016
Sofitel New York	11592
The Belvedere	11544

### Verification:

```
# sourcetype=reviews_final* | stats count(id) avg(ratings.overall) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | search name="Park Central"
```

offering_id	count(id)	name
93520	16036	3.496383 Park Central

## Visualization



## 10. Hotels that have the most negative reviews(2.5 - on overall rating)

**Objective of the query:** To find the hotels that have the most number of reviews.

**Description of the query:**

The intent of this query is to find the hotels that have the highest number of reviews from the customers.

**Query:**

```
# sourcetype=reviews_final* | stats count(id) avg(ratings.overall) by offering_id | lookup
hotel_info.csv hotel_id as offering_id output name | rename avg(ratings.overall) as
average_rating | search average_rating < 2.5|sort limit=10 -count(id) | fields name, count(id) |
rename count(id) as number_reviews
```

Splunk > App: Search & Reporting

Administrator Messages Settings Activity Help Find

Search & Reporting

New Search

sourcetype=reviews\_final\* | stats count(id) avg(ratings.overall) by offering\_id | lookup hotel\_info.csv hotel\_id as offering\_id output name | rename avg(ratings.overall) as average\_rating | search average\_rating < 2.5|sort limit=10 -count(id) | fields name, count(id) | rename count(id) as number\_reviews

878,561 events (before 12/1/14 11:40:11.000 PM)

Events (878,561) Patterns Statistics (10) Visualization

20 Per Page Format Preview

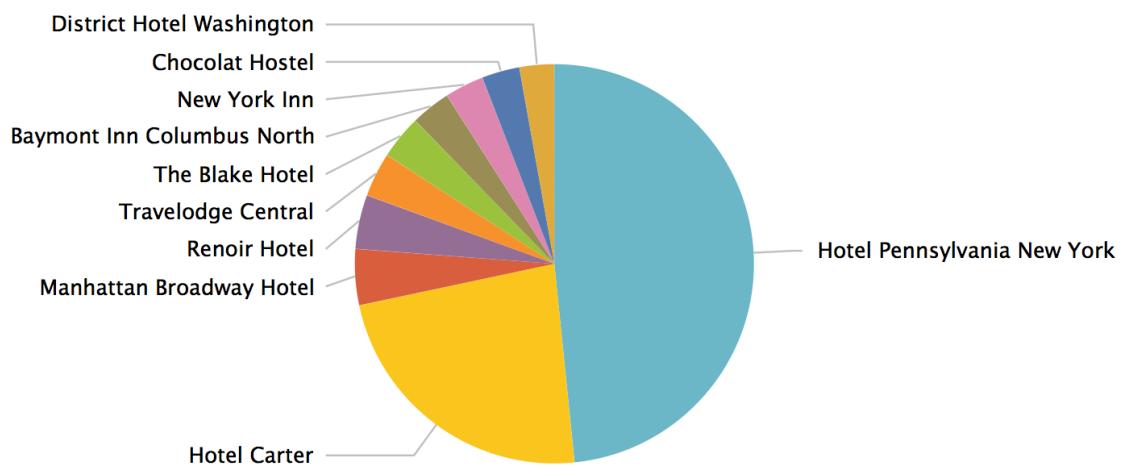
name	number_reviews
Hotel Pennsylvania New York	21824
Hotel Carter	9280
Manhattan Broadway Hotel	1696
Hotel Riverside Studios	1616
Renoir Hotel	1524
New York Inn	1256
District Hotel Washington	1180
Chocolat Hostel	1148
West Side Inn	976
Cecil Hotel	780

## Verification :

```
# sourcetype=reviews_final* | stats count(id) avg(ratings.overall) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | search name="Hotel Pennsylvania New York"
```

offering_id	count(id)	avg(ratings.overall)	name
214197	21824	2.469391	Hotel Pennsylvania New York

## Visualization



## 11. Top 10 useful reviewers

**Objective of the query:** To find the hotels that have the most number of reviews.

### Description of the query:

The intent of this query is to find the hotels that have the highest number of reviews from the customers.

### Query:

```
# sourcetype=reviews_final author.username != "" | stats sum(num_helpful_votes) by author.username | sort limit=10 -sum(num_helpful_votes)
```

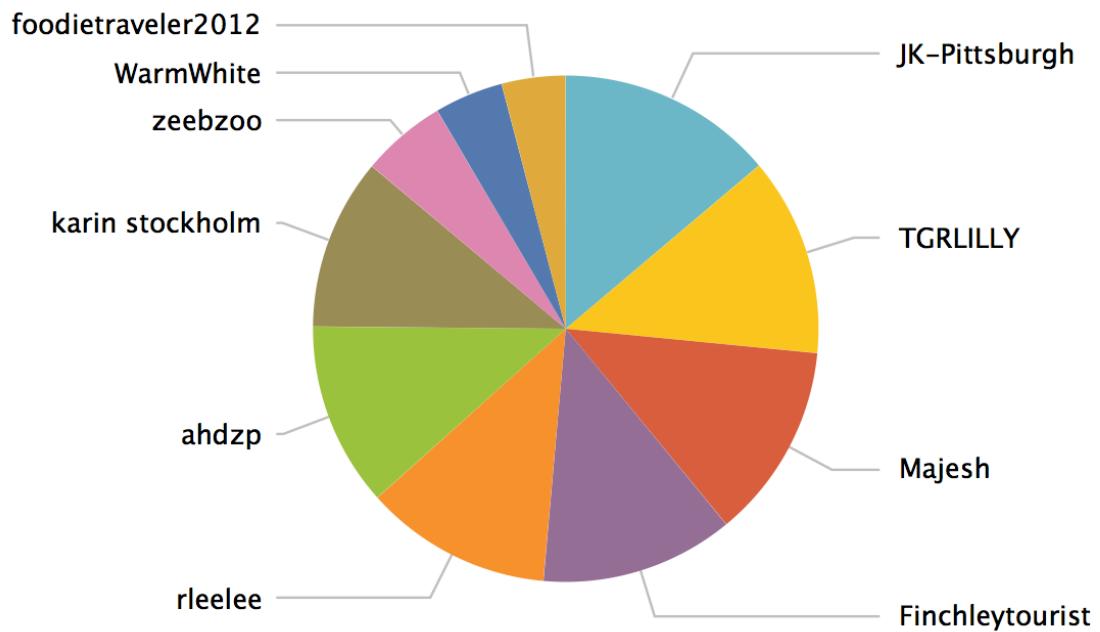
author.username	sum(num_helpful_votes)
JK-Pittsburgh	2060
Majesh	1860
rleeele	1764
foodietraveler2012	464
Michael S	432
zeebzoo	412
WillyJ	376
vudubabz	340
AIB1	324
BosPretzel123	316

### Verification :

```
# sourcetype=reviews_final author.username = "JK-Pittsburgh" | stats sum(num_helpful_votes) by author.username
```

author.username	sum(num_helpful_votes)
JK-Pittsburgh	2060

## Visualization



## 12. Hotel with most number of overall ratings as 5.0

**Objective of the query :** To find the hotel, and its location who has the most number of overall ratings as 5

**Description of the query :**

We wanted to find the hotel which has been rated the best in overall ratings by the maximum number of customers.

**Query:**

```
# sourcetype = reviews_final | search * ratings.overall=5.0 | top limit=10
hotel_name,hotel_city,hotel_street_address,hotel_postal
```

sourcetype = reviews\_final | search \* ratings.overall=5.0 | top limit=10 hotel\_name,hotel\_city,hotel\_street\_address,hotel\_postal

All time  Verbose Mode

Events (348,319) Patterns Statistics (10) Visualization

20 Per Page Format Preview

hotel_name	hotel_city	hotel_street_address	hotel_postal	count	percent
Sofitel New York	New York City	45 West 44th Street	10036	1968	0.566279
Casablanca Hotel Times Square	New York City	147 West 43rd Street	10036	1892	0.544410
Hilton Garden Inn Times Square	New York City	790 Eighth Avenue	10019	1582	0.455210
Argonaut Hotel - a Kimpton Hotel	San Francisco	495 Jefferson Street at Hyde	94109	1533	0.441110
Chancellor Hotel on Union Square	San Francisco	433 Powell Street	94102	1489	0.428450
Waldorf Astoria New York	New York City	301 Park Avenue	10022	1449	0.416940
Doubletree Guest Suites Times Square	New York City	1568 Broadway	10036-8201	1443	0.415214
The New York Palace Hotel	New York City	455 Madison Ave	10022	1399	0.402553
Affinia Manhattan	New York City	371 Seventh Ave	10001	1383	0.397949
Distrik Hotel	New York City	342 West 40th Street	10018	1373	0.395072

## Verification:

```
# sourcetype = reviews_final | search * ratings.overall=5.0 | where hotel_name=" Sofitel New York "
```

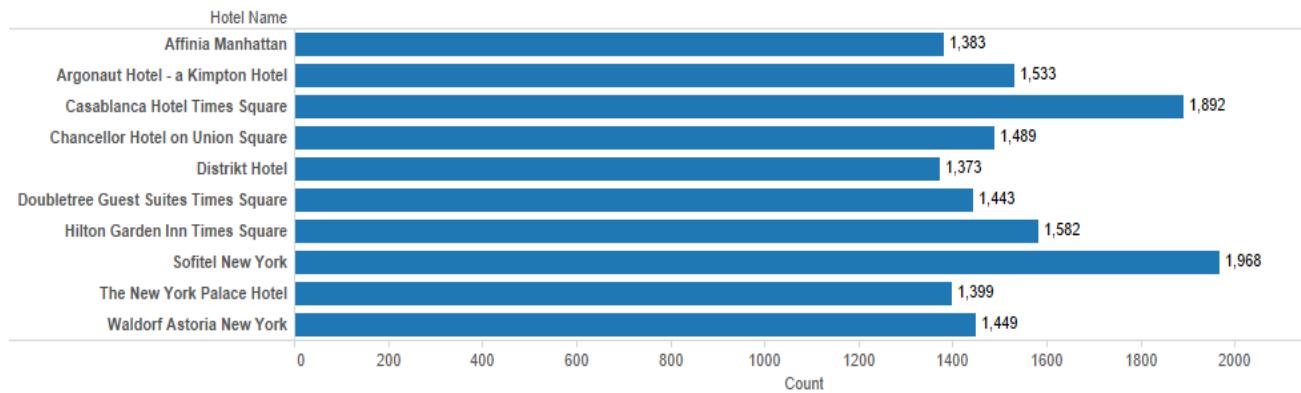
sourcetype = reviews\_final | search hotel\_name="Sofitel New York" "ratings.overall"="5.0"

All time  Verbose Mode

Events (1,968) Patterns Statistics Visualization

Format Timeline    1 month per column

## Visualization



## 13. Hotels with most number of overall ratings as 1.0

**Objective of the query :** To find the hotel, and its location who has the most number of overall ratings as 1.0

### Description of the query :

We wanted to find the hotel which has been rated the worst in overall ratings by the maximum number of customers.

**Query:**

```
# sourcetype = reviews_final | search ratings.overall=1.0 | top limit=10
hotel_name,hotel_city,hotel_street_address,hotel_postal
```

sourcetype = reviews\_final | search ratings.overall=1.0 | top limit=10 hotel\_name,hotel\_city,hotel\_street\_address,hotel\_postal

All time  Job        Verbose Mode

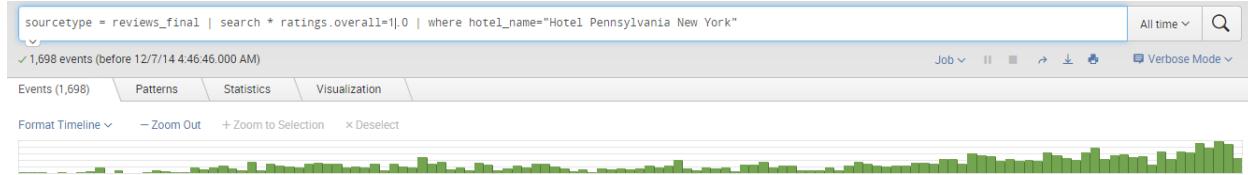
Events (53,848) Patterns Statistics (10) Visualization

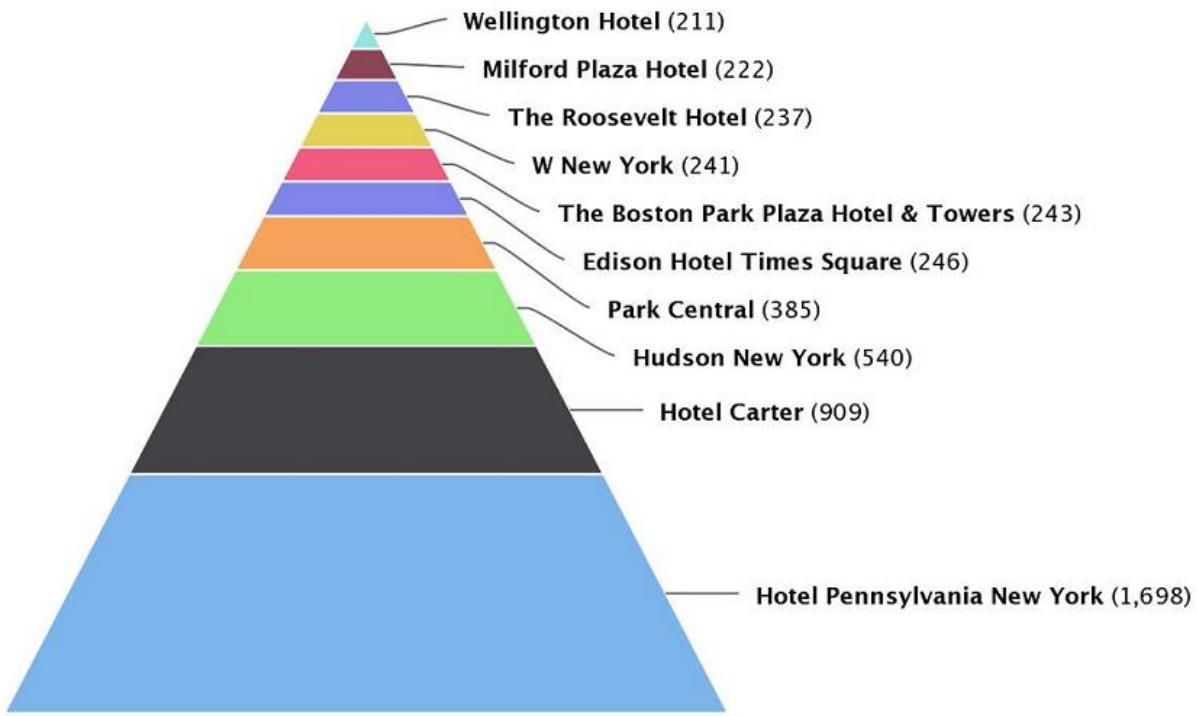
20 Per Page

hotel_name	hotel_city	hotel_street_address	hotel_postal	count	percent
Hotel Pennsylvania New York	New York City	401 Seventh Avenue at 33rd Street	10001	1698	3.161364
Hotel Carter	New York City	250 West 43rd Street	10036	909	1.692391
Hudson New York	New York City	356 West 58th Street	10019	540	1.005381
Park Central	New York City	870 Seventh Avenue at 56th Street	10019	385	0.716799
Edison Hotel Times Square	New York City	228 West 47th Street	10036	246	0.458007
The Boston Park Plaza Hotel & Towers	Boston	50 Park Plaza at Arlington Street	2116	243	0.452421
W New York	New York City	541 Lexington Avenue	10022	241	0.446698
The Roosevelt Hotel	New York City	45 East 45th Street at Madison Avenue	10017	237	0.441250
Milford Plaza Hotel	New York City	700 8th Avenue	10036	222	0.413323
Wellington Hotel	New York City	871 7th Ave at W 55th St	10019-3830	211	0.392843

**Verification :**

```
# sourcetype = reviews_final | search * ratings.overall=1.0 | where hotel_name="Hotel
Pennsylvania New York"
```

**Visualization**



#### 14. Average of all types of ratings for hotels

**Objective of the query :** To find the hotel, and its location who has the most number of overall ratings as 5

**Description of the query :**

We wanted to find the hotel which has been rated the best in overall ratings by the maximum number of customers.

**Query:**

```
# sourcetype=reviews_final* | stats count(id) , avg(ratings.overall) , avg(ratings.cleanliness), avg(ratings.location), avg(ratings.rooms), avg(ratings.sleep_quality), avg(ratings.service) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | rename avg(ratings.overall) as average_rating | search average_rating < 3.0|sort limit=5 -count(id) | fields name, hotel_city, count(id), average_rating, avg(ratings.cleanliness), avg(ratings.location), avg(ratings.rooms), avg(ratings.sleep_quality), avg(ratings.service) | rename count(id) as number_reviews
```

✓ 878,561 events (before 12/7/14 4:52:23.000 PM)

name	number_reviews	average_rating	avg(ratings.cleanliness)	avg(ratings.location)	avg(ratings.rooms)	avg(ratings.sleep_quality)	avg(ratings.service)
Hotel Pennsylvania New York	21824	2.469391	2.350402	4.484938	2.149820	2.671671	2.443849
Hotel Carter	9280	2.273276	1.912167	4.534577	1.931787	2.885827	2.143236
Congress Plaza Hotel	2776	2.899135	3.176271	4.370441	2.766917	3.095672	3.045531
Travelodge Chicago Downtown	1936	2.973140	3.163683	4.252841	2.687332	2.980695	3.278772
Radisson Plaza Lord Baltimore	1904	2.985294	3.159624	3.678378	3.032746	3.075099	3.428910

## Verification :

```
# sourcetype=reviews_final* | stats count(id) avg(ratings.overall), avg(ratings.location) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | search name="Hotel Pennsylvania New York"
```

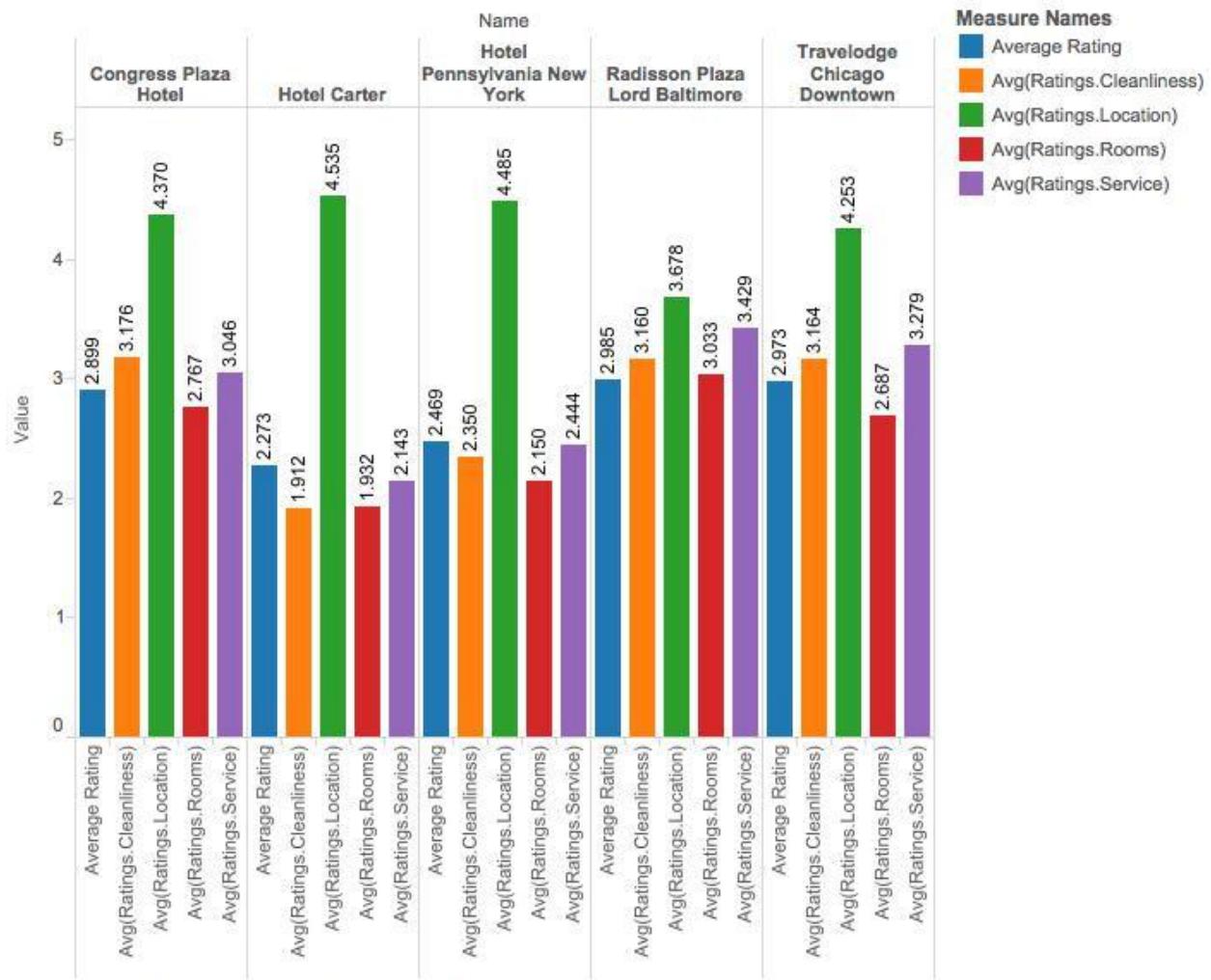
New Search

offering_id	count(id)	avg(ratings.overall)	avg(ratings.location)	name
214197	17864	2.550828	4.493188	Hotel Pennsylvania New York

- Here we see that the average ratings for location for the Hotel Pennsylvania New York is very high that is 4.48. This could be one possible reasons that people book this hotel the most even though the hotel has most negative reviews.
- We also observe that the negative rated hotels also have high review counts. What makes them a preferred choice even if the ratings are low. Is it because of location?
- To confirm this we found the average ratings for all the features such as overall, location, cleanliness, sleep\_qulaity etc for hotel which are reviewed the most as overall\_ratings < 2.5

## Visualization

## Sheet 1



Average Rating, Avg(Ratings.Cleanliness), Avg(Ratings.Location), Avg(Ratings.Rooms) and Avg(Ratings.Service) for each Name. Color shows details about Average Rating, Avg(Ratings.Cleanliness), Avg(Ratings.Location), Avg(Ratings.Rooms) and Avg(Ratings.Service).

We observe that location ratings for all these hotels are very high. So location of the hotel is one of the compelling reason while making hotel choice besides the price and class.

## 15. Top 5 hotels having negative reviews but the most number of review count in NYC

**Objective of the query :** To find the hotels with most number of reviews negative in New York.

**Description of the query :**

We wanted to find the hotel which has been rated the best in overall ratings by the maximum number of customers.

**Query:**

```
# sourcetype=reviews_final |search * "hotel_city" = "New York City" | stats count(id) avg(ratings.overall), avg(ratings.cleanliness), avg(ratings.location), avg(ratings.rooms) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | rename avg(ratings.overall) as average_rating | search average_rating <= 2.5 | sort limit=5 -count(id) | fields name, count(id), average_rating, avg(ratings.cleanliness), avg(ratings.location), avg(ratings.rooms) | rename count(id) as number_reviews
```

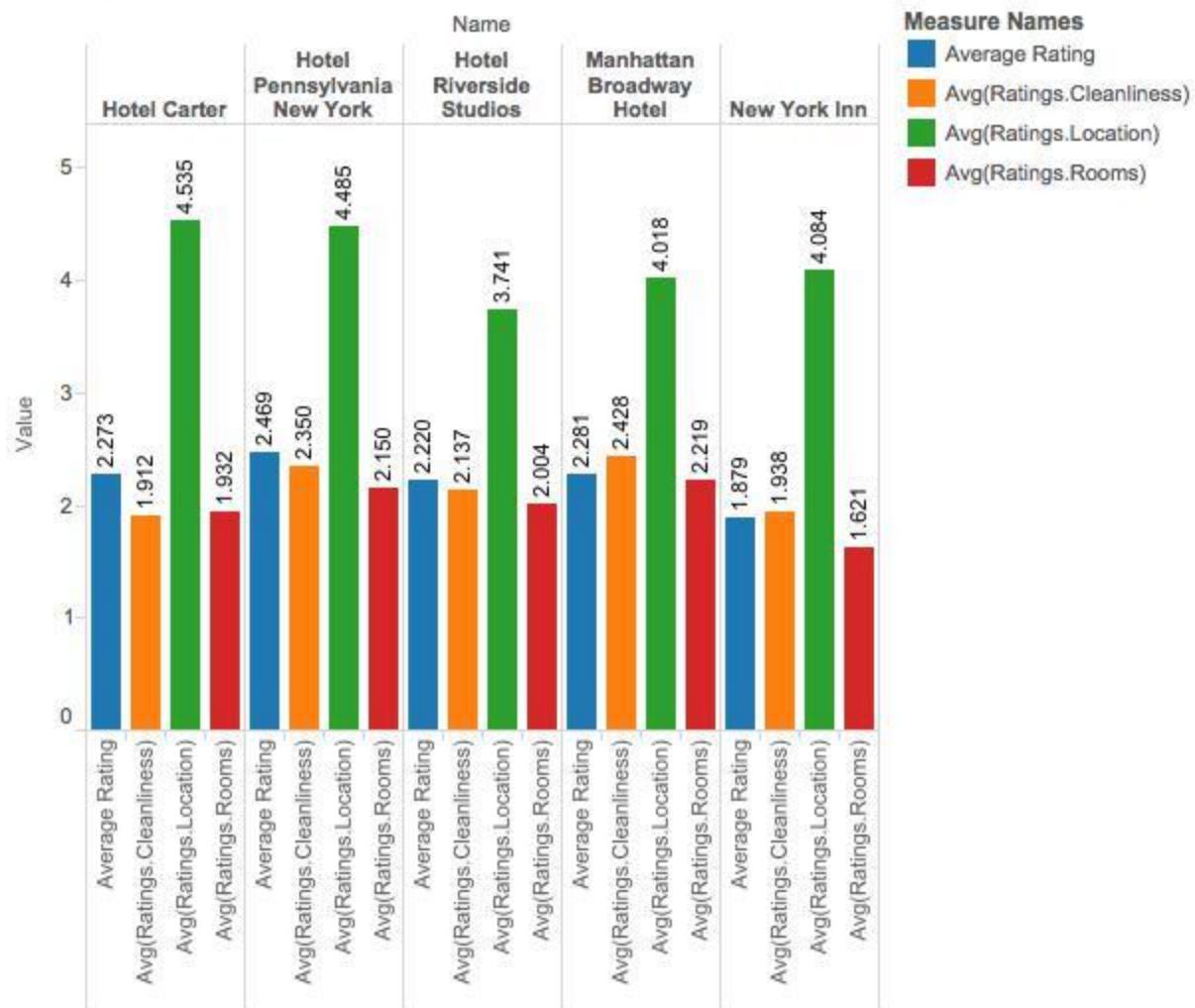
✓ 267,057 events (before 12//14 8:23:35.000 PM)						
Events (267,057)		Patterns	Statistics (5)	Visualization		
20 Per Page ▾ Format ▾ Preview ▾						
name	number_reviews	average_rating	avg(ratings.cleanliness)	avg(ratings.location)	avg(ratings.rooms)	
Hotel Pennsylvania New York	21824	2.469391	2.350402	4.484938	2.149820	
Hotel Carter	9280	2.273276	1.912167	4.534577	1.931787	
Manhattan Broadway Hotel	1696	2.280660	2.427562	4.018433	2.219331	
Hotel Riverside Studios	1616	2.220297	2.136691	3.741071	2.003676	
New York Inn	1256	1.878981	1.938224	4.084071	1.621094	

**Verification :**

```
# sourcetype=reviews_final* | stats count(id) avg(ratings.overall) by offering_id | lookup hotel_info.csv hotel_id as offering_id output name | search name="Hotel Pennsylvania New York"
```

**Visualization**

## Sheet 1



Average Rating, Avg(Ratings.Cleanliness), Avg(Ratings.Location) and Avg(Ratings.Rooms) for each Name. Color shows details about Average Rating, Avg(Ratings.Cleanliness), Avg(Ratings.Location) and Avg(Ratings.Rooms).

## 16. Top 10 haunted hotels based on user's reviews in USA

**Objective of the query :** To find the haunted hotels based on user reviews.

**Description of the query :**

We wanted to find the hotels which are reviewed as haunted by the maximum number of customers.

**Query:**

```
#sourcetype = reviews_final | search * "text" = "*haunted*" AND *ghost* | top limit=10
hotel_name
```

The screenshot shows the Kibana interface with a table visualization. The table has three columns: 'hotel\_name' (sorted by count), 'count' (sorted by count), and 'percent' (sorted by percent). The data is as follows:

hotel_name	count	percent
Omni Parker House	29	7.021792
The Menger Hotel	23	5.569007
Congress Plaza Hotel	20	4.842615
The Driskill	14	3.389831
Hotel Pennsylvania New York	14	3.389831
Hotel San Carlos	13	3.147700
Hollywood Roosevelt Hotel - A Thompson Hotel	11	2.663438
The St. Anthony Riverwalk, A Wyndham Hotel	10	2.421308
The Queen Anne Hotel	10	2.421308
The Horton Grand Hotel and Suites	9	2.179177

## Verification:

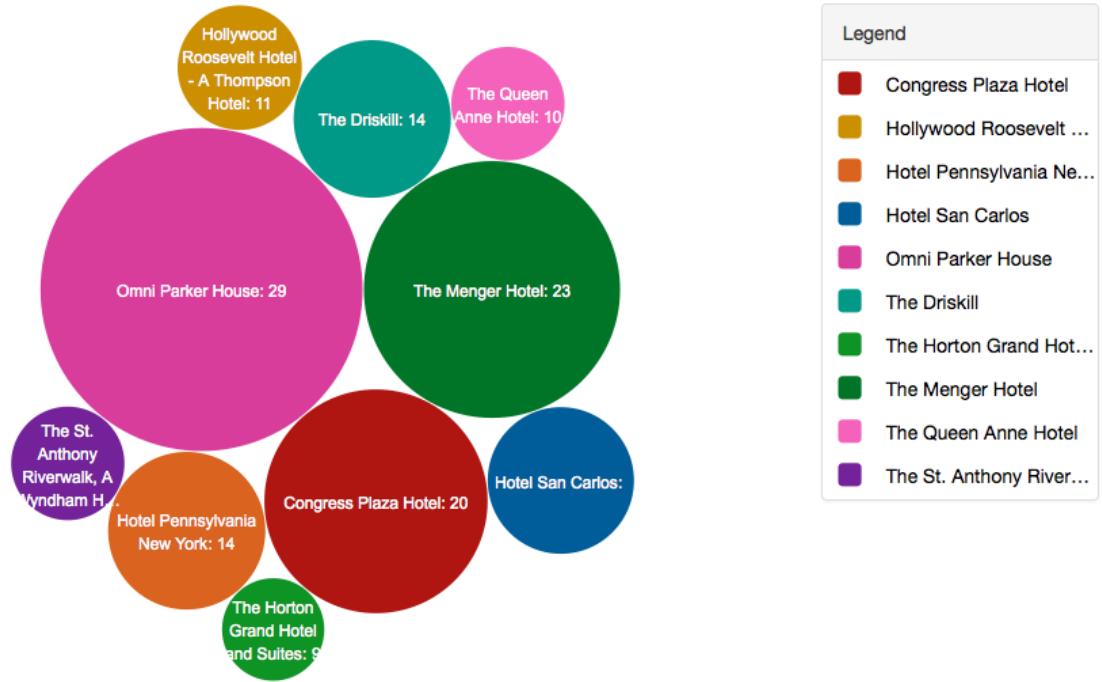
We verified it for the hotel Omni Parker House.

```
# sourcetype=reviews_final | search* "text" = "*haunted*" | where hotel_name = "Omni Parker
House"
```

The screenshot shows the Kibana interface with a table visualization. The table has three columns: 'hotel\_name' (sorted by count), 'count' (sorted by count), and 'percent' (sorted by percent). The data is as follows:

hotel_name	count	percent
Omni Parker House	29	7.021792

## Visualization



## 17. Top 10 haunted hotels based on user's reviews in NY

**Objective of the query :** To find the haunted hotels in NY based on user reviews.

**Description of the query :**

We wanted to find the hotels which are reviewed as haunted by the maximum number of customers.

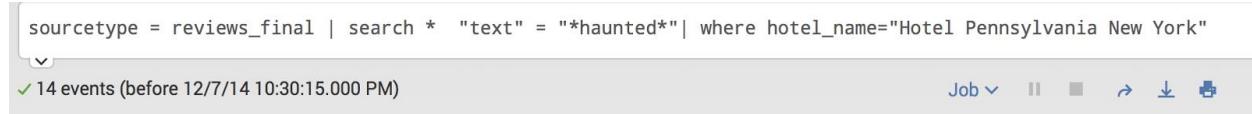
**Query:**

```
# sourcetype = reviews_final | search * "text" = "*haunted*" | where hotel_state = "NY" | top limit=10 hotel_name hotel_street_address hotel_state hotel_postal hotel_city
```

hotel_name	hotel_street_address	hotel_state	hotel_postal	hotel_city	count	percent
Hotel Pennsylvania New York	401 Seventh Avenue at 33rd Street	NY	10001	New York City	14	26.923077
Hotel Carter	250 West 43rd Street	NY	10036	New York City	4	7.692308
Chelsea Hotel	222 W 23rd St	NY	10011	New York City	4	7.692308
Wolcott Hotel	4 West 31st Street	NY	10001	New York City	2	3.846154
Hudson New York	356 West 58th Street	NY	10019	New York City	2	3.846154
Edison Hotel Times Square	228 West 47th Street	NY	10036	New York City	2	3.846154
Algonquin Hotel Times Square, Autograph Collection	59 W 44th St	NY	10036	New York City	2	3.846154
YOTEL New York at Times Square West	570 10th Ave	NY	10036	New York City	1	1.923077
Waldorf Astoria New York	301 Park Avenue	NY	10022	New York City	1	1.923077
The New York Palace Hotel	455 Madison Ave	NY	10022	New York City	1	1.923077

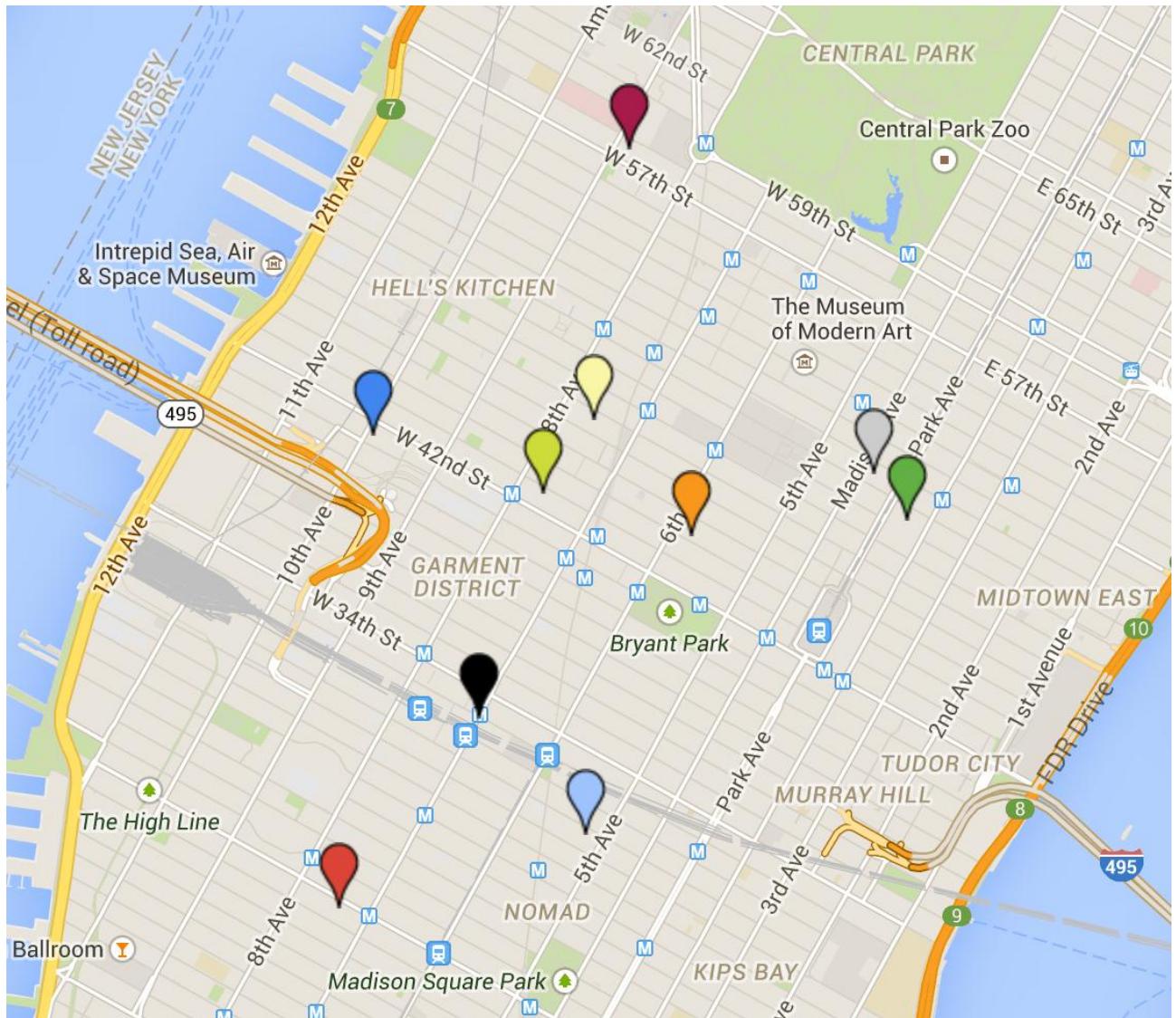
## Verification:

```
# sourcetype= reviews_final | search * "text" = "*haunted*" | where hotel_name = "Hotel Pennsylvania New York"
```



A screenshot of a search interface. The query entered is: `sourcetype = reviews_final | search * "text" = "*haunted*" | where hotel_name = "Hotel Pennsylvania New York"`. Below the query, it says `✓ 14 events (before 12/7/14 10:30:15.000 PM)`. To the right, there are several small icons: a downward arrow, a double arrow, a magnifying glass, a download icon, and a print icon.

## Visualization



## 18. The top 20 hotels with bed bugs based on user's reviews

**Objective of the query :** To find the hotels with bedbugs based on user reviews.

**Description of the query :**

We wanted to find the hotels which were having maximum bedbugs as reviewed by the customers.

**Query:**

```
# sourcetype = reviews_final | search * "text" = "*bed bugs*" AND "text" != "*no bed bugs*"
AND "text" != "*no bugs*" AND "text" != "*no sign of bed bugs*" AND "text" != "*any sign of
bed bugs*" | top limit=5 hotel_name,hotel_city,hotel_street_address
```

Q New Search

sourcetype = reviews\_final | search \* "text" = "\*bed bugs\*" AND "text" != "\*no bed bugs\*" AND "text" != "\*no sign of bed bugs\*" AND "text" != "\*any sign of bed bugs\*" | top limit=5 hotel\_name,hotel\_city,hotel\_street\_address

✓ 2,855 events (before 12/7/14 3:41:25.000 AM)

Events (2,855) Patterns Statistics (5) Visualization

20 Per Page ▾ Format ▾ Preview ▾

hotel_name	hotel_city	hotel_street_address	count	percent
Hotel Pennsylvania New York	New York City	401 Seventh Avenue at 33rd Street	90	3.152364
Hotel Carter	New York City	250 West 43rd Street	87	3.047285
Edison Hotel Times Square	New York City	228 West 47th Street	34	1.190893
Marrakech Hotel on Broadway	New York City	2688 Broadway	29	1.015762
Powell Hotel	San Francisco	28 Cyril Magnin Street	27	0.945709

## Verification :

```
# sourcetype = reviews_final | search * "text" = "*bed bugs*" AND "text" != "*no bed bugs*"
AND "text" != "*no bugs*" AND "text" != "*no sign of bed bugs*" AND "text" != "*any sign of
bed bugs*" hotel_name="Hotel Carter"
```

Search Pivot Reports Alerts Dashboards

Search & Reporting

Q New Search

sourcetype = reviews\_final | search \* "text" = "\*bed bugs\*" AND "text" != "\*no bed bugs\*" AND "text" != "\*no sign of bed bugs\*" AND "text" != "\*any sign of bed bugs\*" hotel\_name="Hotel Carter"

✓ 87 events (before 12/7/14 3:54:59.000 AM)

Save As ▾ Close All time ▾ Verbose Mode ▾

## Visualization

