Project Work - Web Analytics

The project uses MySQL to perform analysis. The code for database connection is hidden for security reasons.

In [1]:
```
%load_ext sql
import mysql
```

In [2]:
```
# Connecting to the database
%sql mysql://root:            @localhost/mavenfuzzyfactory
```

Analysing Website Content

Website Content Analysis is conducted by first identifying the top pages and the top entry pages. This is done in order to generate an understanding of how traffic is interacting with the website. Based on this, bounce rate analysis and further testing is carried out.

Top Entry Pages

In [3]:
```
%%sql

SELECT *
FROM website_pageviews
LIMIT 10;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
10 rows affected.
```

Out[3]:

website_pageview_id	created_at	website_session_id	pageview_url
1	2012-03-19 08:04:16	1	/home
2	2012-03-19 08:16:49	2	/home
3	2012-03-19 08:26:55	3	/home
4	2012-03-19 08:37:33	4	/home
5	2012-03-19 09:00:55	5	/home
6	2012-03-19 09:05:46	6	/home
7	2012-03-19 09:06:27	7	/home
8	2012-03-19 09:10:08	6	/products
9	2012-03-19 09:10:52	6	/the-original-mr-fuzzy
10	2012-03-19 09:14:02	6	/cart

In [4]:
```
# website_pageviews overview, first 10 records.
# website_pageview_id is generated automatically regardless of session_id.
```

In [5]:
```
%%sql

SELECT *
FROM website_pageviews
WHERE website_session_id = 6;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
6 rows affected.
```

Out[5]:

website_pageview_id	created_at	website_session_id	pageview_url
6	2012-03-19 09:05:46	6	/home
8	2012-03-19 09:10:08	6	/products
9	2012-03-19 09:10:52	6	/the-original-mr-fuzzy
10	2012-03-19 09:14:02	6	/cart
11	2012-03-19 09:16:52	6	/shipping
13	2012-03-19 09:19:52	6	/billing

In [6]:
```
# Here, we can see all the pages visited for a given session (website session '6')
```

What are the **most viewed website pages**, ranked by session volume?
- The query is asked on 9 June, 2012

In [7]:
```
%%sql

SELECT
    pageview_url, COUNT(DISTINCT website_session_id) AS 'Visits'
FROM website_pageviews
WHERE created_at < '2012-06-09'
GROUP BY pageview_url
ORDER BY 2 DESC;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
7 rows affected.
```

Out[7]:

pageview_url	Visits
/home	10403
/products	4239
/the-original-mr-fuzzy	3037
/cart	1306
/shipping	869
/billing	716
/thank-you-for-your-order	306

In [8]:
```
# The bulk of traffic is being recieved by the home page, the products page and 'the original mr fuzzy' page.
```

What are the top **entry** pages?
- The request is made on 12 June, 2012

In [9]:
```
%%sql

SELECT
    DISTINCT website_session_id,
    MIN(website_pageview_id) AS 'entry_page_id'
FROM website_pageviews
WHERE created_at < '2012-06-12'
GROUP BY 1
LIMIT 10;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
10 rows affected.
```

Out[9]:

website_session_id	entry_page_id
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	12
9	14
10	15

In [10]:
```
# We have the first page viewed by each session in the query result.
# This result can be saved in a temporary table to access the first pages.
```

In [11]:
```
%%sql

CREATE TEMPORARY TABLE first_page_per_session
SELECT
    DISTINCT website_session_id,
    MIN(website_pageview_id) AS 'entry_page_id'
FROM website_pageviews
WHERE created_at < '2012-06-12'
GROUP BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
10714 rows affected.
```

Out[11]:
```
[]
```

In [12]:
```
%%sql

SELECT COUNT(fp.website_session_id) AS 'Sessions', pageview_url
FROM first_page_per_session fp
LEFT JOIN website_pageviews
ON fp.entry_page_id = website_pageview_id
GROUP BY 2;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
1 rows affected.
```

Out[12]:

Sessions	pageview_url
10714	/home

In [13]:
```
# Hence, entry page for all website sessions so far is the 'home' page.
# (All the traffic comes through the homepage).
```

Landing Page Performance & Testing

What is the **bounce rate** for the traffic landing on homepage?
- The query is asked on June 14, 2012

In [14]:
```
# In simpler terms, a customer is said to have 'bounced' if he/she does not
# visit a different page after landing on the entry page.
```

In [15]:
```
%%sql

SELECT
    website_session_id,
    COUNT(website_pageview_id) AS 'pages_viewed'
FROM website_pageviews
WHERE created_at < '2012-06-14'
GROUP BY 1
LIMIT 10;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
10 rows affected.
```

Out[15]:

website_session_id	pages_viewed
1	1
2	1
3	1
4	1
5	1
6	6
7	1
8	1
9	1
10	1

In [16]:
```
# The table above provides the number of pages viewed per website session.
```

In [17]:
```
%%sql

CREATE TEMPORARY TABLE pageviews
SELECT
    website_session_id,
    COUNT(website_pageview_id) AS 'pages_viewed'
FROM website_pageviews
WHERE created_at < '2012-06-14'
GROUP BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
11048 rows affected.
```

Out[17]:
```
[]
```

In [18]:
```
%%sql

SELECT
    COUNT(CASE WHEN website_session_id IS NOT NULL THEN website_session_id ELSE NULL END) AS 'sessions',
    SUM(CASE WHEN pages_viewed = 1 THEN 1 ELSE 0 END) AS 'bounced_sessions',
    (SUM(CASE WHEN pages_viewed = 1 THEN 1 ELSE 0 END))/
    (COUNT(CASE WHEN website_session_id IS NOT NULL THEN website_session_id ELSE NULL END))
    AS 'bounce_rate'
FROM pageviews;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
1 rows affected.
```

Out[18]:

sessions	bounced_sessions	bounce_rate
11048	6538	0.5918

In [19]:
```
# The bounce rate is almost 60%. This is poor for a paid marketing campaign.
# Creating a new lander page should be considered to reduce the bounce rate.
```

A new landing page **/lander-1** has been introduced against the **/home** page in a 50/50 test.

- Compare the bounce rate for the two groups

- The query is asked on 28 July, 2012

*- We are supposed to look at the time period where **/lander-1** began receiving traffic*

In [20]:
```
%%sql

SELECT MIN(created_at), pageview_url
FROM website_pageviews
WHERE pageview_url = '/lander-1';
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
1 rows affected.
```

Out[20]:

MIN(created_at)	pageview_url
2012-06-19 00:35:54	/lander-1

In [21]:
```
# /lander-1 began receiving traffic from 19 June 2012.
```

In [22]:
```
%%sql

CREATE TEMPORARY TABLE pageviews_2
SELECT
    ws.website_session_id,
    COUNT(wp.website_pageview_id) AS 'pages_viewed'
FROM website_sessions ws
LEFT JOIN website_pageviews wp
ON wp.website_session_id = ws.website_session_id
WHERE
    ws.created_at BETWEEN '2012-06-19' AND '2012-07-28' AND
    ws.utm_source = 'gsearch' AND
    ws.utm_campaign = 'nonbrand'
GROUP BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
4577 rows affected.
```

Out[22]: []

In [23]:
```
# The query saves the number of pages viewed per session
```

In [24]:
```
%%sql

CREATE TEMPORARY TABLE session_lander_id
SELECT
    ws.website_session_id,
    MIN(wp.website_pageview_id) AS 'lander_pg_id'
FROM website_sessions ws
LEFT JOIN website_pageviews wp
ON wp.website_session_id = ws.website_session_id
WHERE
    ws.created_at BETWEEN '2012-06-19' AND '2012-07-28' AND
    ws.utm_source = 'gsearch' AND
    ws.utm_campaign = 'nonbrand'
GROUP BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
4577 rows affected.
```

Out[24]: []

In [25]:
```
%%sql

CREATE TEMPORARY TABLE session_w_lander
SELECT
    sli.website_session_id,
    wp.pageview_url AS 'lander'
FROM session_lander_id sli
LEFT JOIN website_pageviews wp
ON sli.lander_pg_id = wp.website_pageview_id
ORDER BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
4577 rows affected.
```

Out[25]: []

In [26]:
```
# The query saves the lander page url for each relevant session.
```

In [27]:
```
%%sql

SELECT
    swl.website_session_id,
    swl.lander,
    pg2.pages_viewed
FROM session_w_lander swl
LEFT JOIN pageviews_2 pg2
ON swl.website_session_id = pg2.website_session_id
LIMIT 10;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
10 rows affected.
```

Out[27]:

website_session_id	lander	pages_viewed
11683	/lander-1	1
11684	/home	1
11685	/lander-1	1
11686	/lander-1	2
11687	/home	1
11688	/home	1
11689	/lander-1	3
11690	/home	1
11691	/lander-1	2
11692	/lander-1	1

In [28]:
```
# Using the 'sessions_w_lander' and 'pageviews_2' table we get the information
# for lander page as well as number of pages viewed for each relevant session.
```

In [29]:
```
%%sql

CREATE TEMPORARY TABLE ses_w_lander_pgs
SELECT
    swl.website_session_id,
    swl.lander,
    pg2.pages_viewed
FROM session_w_lander swl
LEFT JOIN pageviews_2 pg2
ON swl.website_session_id = pg2.website_session_id;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
4577 rows affected.
```

Out[29]: []

In [30]:
```
# This query result in the cell above is saved, in order to compare the bounce rates for '/home' and '/lander-1'.
```

In [31]:
```
%%sql

SELECT
    lander,
    COUNT(website_session_id) AS 'sessions',
    SUM(CASE WHEN pages_viewed =1 THEN 1 ELSE 0 END) AS 'bounced_sessions',
    SUM(CASE WHEN pages_viewed =1 THEN 1 ELSE 0 END)/COUNT(website_session_id) AS 'bounce_rate'
FROM ses_w_lander_pgs
GROUP BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
2 rows affected.
```

Out[31]:

lander	sessions	bounced_sessions	bounce_rate
/lander-1	2316	1233	0.5324
/home	2261	1319	0.5834

In [32]:
```
# The custom lander '/lander-1' has lower bounce rate.
# It would be beneficial to divert the 'gsearch nonbrand' traffic to this lander.
```

Is the traffic routed **correctly**?

*- pull the volume of paid search nonbrand traffic landing on **/home** and **/lander-1, trended weekly** since June 1st?*

*- pull overall **bounce rate** as well*

- The query is asked on Aug 31, 2012

In [33]:
```
# We already saw how to pull pages_viewed for a website_session.
```

In [34]:
```
%%sql

CREATE TEMPORARY TABLE session_lander_id_2
SELECT
    ws.website_session_id,
    MIN(wp.website_pageview_id) AS 'lander_pg_id'
FROM website_sessions ws
LEFT JOIN website_pageviews wp
ON wp.website_session_id = ws.website_session_id
WHERE
    ws.created_at BETWEEN '2012-06-01' AND '2012-08-31' AND
    ws.utm_source = 'gsearch' AND
    ws.utm_campaign = 'nonbrand'
GROUP BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
11624 rows affected.
```

Out[34]:
```
[]
```

In [35]:
```
# This gives us the lander page id for each relevant session.
```

In [36]:
```
%%sql

CREATE TEMPORARY TABLE session_w_lander_2
SELECT
    sli2.website_session_id,
    wp.pageview_url AS 'lander'
FROM session_lander_id_2 sli2
LEFT JOIN website_pageviews wp
ON sli2.lander_pg_id = wp.website_pageview_id
ORDER BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
11624 rows affected.
```

Out[36]:
```
[]
```

In [37]:
```
# This gives us the lander page url for each relevant website session.
```

In [38]:
```
%%sql

CREATE TEMPORARY TABLE pageviews_3
SELECT
    ws.website_session_id,
    COUNT(wp.website_pageview_id) AS 'pages_viewed',
    ws.created_at
FROM website_sessions ws
LEFT JOIN website_pageviews wp
ON wp.website_session_id = ws.website_session_id
WHERE
    ws.created_at BETWEEN '2012-06-01' AND '2012-08-31' AND
    ws.utm_source = 'gsearch' AND
    ws.utm_campaign = 'nonbrand'
GROUP BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
11624 rows affected.
```

Out[38]: []

In [39]:
```
# The query gives us the number of pageviews per relevant session.
```

In [40]:
```
%%sql

SELECT
    pg3.website_session_id,
    pg3.pages_viewed,
    swl2.lander,
    pg3.created_at
FROM pageviews_3 pg3
LEFT JOIN session_w_lander_2 swl2
ON swl2.website_session_id = pg3.website_session_id
LIMIT 5;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
5 rows affected.
```

Out[40]:

website_session_id	pages_viewed	lander	created_at
9350	3	/home	2012-06-01 00:05:11
9351	3	/home	2012-06-01 00:06:39
9352	4	/home	2012-06-01 00:08:27
9354	1	/home	2012-06-01 01:08:43
9356	6	/home	2012-06-01 01:37:31

In [41]:
```
# The above result is contains all the information we need to study session volume
# for both lander pages.
```

In [42]:
```
%%sql

CREATE TEMPORARY TABLE vol_trending
SELECT
    pg3.website_session_id,
    pg3.pages_viewed,
    swl2.lander,
    pg3.created_at
FROM pageviews_3 pg3
LEFT JOIN session_w_lander_2 swl2
ON swl2.website_session_id = pg3.website_session_id;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
11624 rows affected.
```

Out[42]:
```
[]
```

In [43]:
```
%%sql

SELECT
    MIN(DATE(created_at)) AS 'week_start_date',
    SUM(CASE WHEN pages_viewed = 1 THEN 1 ELSE 0 END)/SUM(CASE WHEN website_session_id IS NOT NULL THEN 1 ELSE 0 END)
    AS 'bounce_rate',
    SUM(CASE WHEN lander = '/home' THEN 1 ELSE 0 END) AS 'home_sessions',
    SUM(CASE WHEN lander = '/lander-1' THEN 1 ELSE 0 END) AS 'lander_sessions'
FROM vol_trending
GROUP BY YEARWEEK(created_at)
ORDER BY 1;
```

```
 * mysql://root:***@localhost/mavenfuzzyfactory
14 rows affected.
```

Out[43]:

week_start_date	bounce_rate	home_sessions	lander_sessions
2012-06-01	0.6057	175	0
2012-06-03	0.5871	792	0
2012-06-10	0.6160	875	0
2012-06-17	0.5582	492	350
2012-06-24	0.5828	369	386
2012-07-01	0.5821	392	388
2012-07-08	0.5668	390	411
2012-07-15	0.5424	429	421
2012-07-22	0.5138	402	394
2012-07-29	0.4971	33	995
2012-08-05	0.5382	0	1087
2012-08-12	0.5140	0	998
2012-08-19	0.5010	0	1012
2012-08-26	0.5378	0	833

In [44]:
```
# From the table above it is cleary visible that the traffic has been successfully redirected.
```