Search Engine Marketing Analytics with Hadoop Ecosystem: Case Study for Red Engine Digital Agency

Mia Alowaish
Northwestern University
MaiAlowaish2012@u.northwestern.edu

Sunil Kakade
Northwestern University
Sunil.kakade@gmail.com

I. Introduction

As an integrated online marketing agency, RED thrives to make the right decisions in advertising to ensure high return on investment (ROI) for their clients advertising efforts. Insights about data are invaluable when it comes to post campaign analysis or future planning. RED is always looking for better ways to market to its audience based on learning from previous campaigns, or learning from other data sets that can be applied to their clients.

With all the technology available to collect data about digital marketing efforts and results, RED will definitely benefit from the use of Big Data in its daily operations, especially in the Search Engine Marketing (SEM) field. SEM or "Paid Search" refers to paid listings, such as Paid search advertising, PPC (pay-per-click), PPC (pay-per-call) and more. With the variety of SEM channels and the countless keywords visitors search for to see advertisements, marketers have a huge need to optimize their search bids and make the best out of their marketing budgets (better ROI). The data can be huge when trying to analyze search trends and come up with valuable insights to inform marketing and bidding decisions. This case study will use Big Data tools to analyze historical information of SEM and provide a working example that RED can use to utilize Big Data for its daily marketing optimization work.

II. Big Data Overview

According to Krishnan, Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity that cannot be processed using traditional technologies. RED marketing activities create volumes of data that can be classified into the following

- Advertising and online bid data: Search Engine Marketing (SEM) data including clicks, impressions, advertisement placement, keywords, cost, click-through rate, cost per click.
- Email data: email opens, views, click-through rate, conversion rate
- Web traffic data: Visitors' sessions on the website are all recorded to view engagement using pages visited, time on site, clicks and events, bounce rates.
- Transaction data: With online fundraising, data about donation amount, time, and conversion rate are recorded by the fundraising systems.

All of this data (and more) are collected at different rates, advertising clicks data will probably have more volume than transaction data, out of each 100 or 1000 clicks there is few donations (conversions). Email data are seasonal but could be of a big volume depending on the client base and targeted audience, but are less complex than advertising data. Web traffic is collected at a fast velocity and come with a lot of ambiguity. Integrating all this data together to make sense of the marketing efforts and optimize for the next campaign is definitely a good example to use Big Data as it needs more than basic excel sheets or database queries.

III. Big Data Technologies

The topic of Big Data is broad and encompasses many trends and new technology developments; the following list highlights some of the emerging technologies that are helping users cope with and handle Big Data in a cost-effective manner:

1. Hadoop

Hadoop is the most popular implementation of MapReduce, being an entirely open source platform to work with Big Data. It is flexible enough to handle multiple data sources, by aggregating multiple sources of data in order to do large scale processing, or reading database content to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes of constantly changing datasets. For example, weather or traffic data, web-based or social media data, or machine-to-machine transactional data.

2. MapReduce

MapReduce is the core and the programming paradigm that allows for massive job execution scalability against multiple servers and clusters of servers. When we say MapReduce implementation, we refer to two main jobs that it performs:

A. Map: the input dataset is converted into a different set of key/value pairs, or tuples

B. Reduce: several of the outputs of the "Map" task are combined to form a reduced set of tuples.

3. Hive

Hive is a "SQL-like" tool allowing conventional BI applications to run queries against Hadoop clusters. Hive was initially developed by Facebook to query the massive data they keep about their users; but then it was made open source for everyone's use. Hive has a higher-level abstraction of the Hadoop framework, allowing anyone to make queries against data stored in a Hadoop cluster just as if it's a regular data store; making it easy and familiar for BI users.

4. PIG

Apache Pig is a platform for analyzing large data sets; its infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist. Pig's language layer consists of a textual language called Pig Latin that provides reusable components for programming scripts. You can think of PIG as a bridge that tries to bring Hadoop closer to the realities of developers and business users. Unlike Hive and other tools, PIG has a "Perl-like" language does the query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

5. NoSQL Databases

NoSQL encompasses a variety of database technologies that were developed in response to the increased volume of data stored about users, objects and products, the frequency in which this data is accessed, and performance and processing needs. Relational databases were not designed to cope with the scale and agility challenges that face modern applications, nor were they built to take advantage of the cheap storage and processing power available today.

NoSQL databases are schema-less databases, you do not need to build a schema like relational databases. NoSQL databases include key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They performance gain is achieved through removing the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

IV. Use Case: SEM Analytics with Big Data Tools

At RED, SEM staff is always looking for insights that helps them achieve better ROI and faster results with their audience. For SEM Management, they use Google Adwords to define the advertisement content, audience, price they want to bid on and placement details. Once the user (online audience) clicks on the advertisement and goes to the website, all the information are stored onto a separate platform, Google Analytics, which

contains the keyword on which the visitor came through along with his visit details, pages he visited, and actions he took. For SEM staff, they want to know quickly how the advertisement is doing, so they don't waste money bidding on keywords that has low conversion rates. Also, year over year data comparison is very important for them, since much of the non-profit organization work is seasonal (End of year tax deductible giving season), insights from the past season and trends of year over year changes can help them better achieve results.

With this in mind, I have chosen to extract data from both, Google Adwords and Google Analytics, and use Hadoop to connect the datasets together and draw insights for the SEM staff. The speed at which we can get results and report on information is very important in the SEM arena, where keywords are optimized automatically, and advertisers are competing for banner space, and in fundraising season, you want to adjust your efforts on a daily basis (based on campaign insights) so you do not waste your money on a campaign that's not showing any revenue. Getting the right insights to make quick and informed decisions will help maximize ROI for advertising efforts.

Using Hadoop, the extracted data from SEM and Web Analytics tools is merged and analyzed as a single large dataset to study how keywords and advertisement content affect the web visitor engagement and overall experience and conversion. Using joins, the data from both systems will be merged based on the "Ad content" field and analysis will be done on the resultant dataset.

V. Dataset Description

SEM tools provide reports on search keywords, advertisement content, cost per click data; Web Analytics software provide data on user interaction and engagement, as well as conversion and revenue data for visitors coming through advertisement and keywords. Merging this data together in a big data set will provide unique and valuable insights for the marketer to plan campaigns and gain better understanding of ROI.

The following datasets were extracted for the HELP relief organization (renamed for privacy). HELP works on crisis relief around the world, using online advertising to raise awareness and funds to help those in need in times of crisis. Two datasets are extracted from separate systems, below is a list of the fields extracted in the each file, extraction was done as CSV files and data was important into Hadoop file browser, and also as tables using the Hadoop interface for creating tables from files.

Google Adwords SEM Tool: Keyword-Ad content report - Fields:

Ad content
Ad group
Ad slot
Ad destination URL
Ad content | Ad group | Ad slot | Ad destination URL
Ad clicks
Ad cost
CPC
CPM
CTR
Ad Impressions
Number of instances: 2,335

Google Web Analytics: Keyword Visit & Engagement Report - Fields:

Keyword
Ad content
Keyword | Ad content
Visits
Bounce rate
New visits %
Return visits
Pageviews
Avg. pageviews per visit
Transactions
Transaction revenue
Avg. revenue per visit
Avg. order value
Total time on site (h:m)
Avg. visits per visitor
Number of instances: 10,005

VI. Results and Analysis

HIVE is suited to query the big data set and provide SEM teams with the high-level insights on the data that they are looking for. The massive number of keywords and their conversion related data can help them bid on the right keywords. Below are some queries that I got from the SEM team that are key to help them analyze and plan advertising campaigns. Using HIVE to get answers to these popular SEM questions, queries and answers are below: (Query results screenshots in appendix 2).

1. What are the best 10 keywords in bringing new visitors to the website, and what are the total visits for each?

```
SELECT keyword, returnvisits, visits
FROM keyword
WHERE keyword NOT RLIKE ('not provided|not set|test|control')
ORDER BY returnvisits DESC
LIMIT 10
```

This query excludes certain keywords that are known to their system for certain scenarios that shouldn't be included. The results (in Appendix 2) shows that the returning visitors mainly come on the "brand name" keyword by the name of the organization; this helps them better know how the visitors who know the organizations come back, and better attract new visitors.

2. What are the top 10 converting keywords for Ebola crisis in terms of transactions?

```
SELECT keyword, visits, transactions, transactionrevenue
FROM keyword
WHERE keyword RLIKE ('ebola|Ebola')
ORDER BY transactions DESC
LIMIT 3
```

This direct query from the table lists the keywords by the number of transactions they brought in, and mentions the related revenue to each.

3. During Typhoon Haiyan, what was the best performing ad groups in terms of clicks?

```
SELECT adgroup, sum(adclicks) as totalclicks
FROM adcontent
WHERE adcontent LIKE '%Typhoon%'
GROUP BY Adgroup
ORDER BY totalclicks DESC
LIMIT 3
```

For this query, we need to look into the adcontent table and sort by clicks, but since this is a flat table with ads in rows, grouping them by the group is necessary to get results at the ad-group level.

4. In terms of fundraising, which performed better? Typhoon Haiyen or Ebola Crisis?

```
SELECT SUM(transactionrevenue)
FROM keyword
WHERE keyword LIKE '%ebola';

SELECT SUM(transactionrevenue)
FROM keyword
WHERE keyword LIKE '%haiyan';
```

Running two queries in HIVE and looking at the results, Ebola raised $909 whereas Haiyan brought in $227.25 ; This helps the SEM staff direct their spending better in which they get better return on investment.

5. With what keywords did users browsed more pages when coming through search for the hunger crisis?

```
SELECT keyword, avgpageviewspervisit
FROM keyword
WHERE keyword LIKE '%hunger%'
ORDER BY avgpageviewspervisit DESC
```

This looks at keywords in general, and not only advertising keywords. Analyzing this helps SEM staff use the keywords that brings in engaged visitors who read and spend good time in the website.

6. What ads have the most transactions?

```
SELECT DISTINCT a.adcontent , transactions
FROM adcontent a JOIN keyword b
ON (a.adcontent = b.adcontent)
ORDER BY transactions DESC
```

Reviewing the results, it was noticed that advertisement with "Matching gift" had more conversions. This helps the team utilize the gift matching offers to bring in more revenue.

7. Which ad group had the best ROI in relation to South Sudan crisis?

```
SELECT adgroup , (sum(transactionrevenue)/sum(adcost))
as ROI
FROM adcontent a JOIN keyword b
ON (a.adcontent = b.adcontent)
WHERE transactionrevenue > 0 AND adcost >0
AND a.adcontent LIKE '%Sudan%'
GROUP by adgroup
```

This query requires joining the tables and calculating the ROI (revenue/spend). Grouping on the ad group level is needed since all data is on the lower ad level. Reviewing the results, there was an ad group with ROI less than 1, means the spend was more than the revenue generated. This is usually an alarm for SEM staff to stop this advertisement group, as it's not worth spending any more money.

Utilizing PIG:

In other cases, SEM staff want to look at reports, data extracted to look at insights, such as the top performing ads, they want to see a daily extract of their engagement and revenue data. These reports are not available for them in Google Adwords, as it only has advertising information, they want to know beyond the click, what happened on the website, and what dollars

were raised. The best way to automate this reports is use PIG, to extract results into files and load them onto the server. Below are two reports that SEM staff mentioned they would like to see as part of the post-campaign analysis.

8. Extract keyword and adcontent data report about advertisements that brought in revenue larger than $5,000 , and a separate report for advertisements generating lower revenue between $100 and $5,000

.

For this report, the best solution is to use a PIG script, to do the join between the two files, perform the data filters on the revenue figures and then split results into separate files for the team to review. The script and results are attached in appendix 3.

VII. Big Data Adoption at RED

With traditional technologies, RED now provides campaign insights through crunching email campaign revenue results to tell the average donation value and conversion rates, which does not help much if it wasn't integrated with a historical view of donation values across all marketing channels (email, website, SEM, social media and more). This becomes even more complex if RED will want to make decisions on whom to target for campaigns and know their audience. For example, revenue data and email conversions exist on the client email system or CRM (Customer Relationship Management), whereas their website visit details (pages visited, time on site) exist on the web traffic logs, and if customers clicked on an advertisement banner to come to the site, their click and keyword information are on the SEM bid management system. All of this data can be put together using Big Data technologies to be able to say something like "Customers using brand search keywords tend to engage more on the website and give a higher donation value, and usually come twice a year to donate, one of which is in the end of year tax-deductible season". Information like that in invaluable to fundraising decision making, and only Big Data technologies can integrate all the volumes of data from different sources to get these insights at a faster processing rate.

For a company like RED, big data is not currently an important or even central component of strategy and campaign planning. Taking advantage of big data will mean that RED will have a culture of methodical decision-making. RED analysts will spend their time analyzing data, converting data into actionable insights, generating forecasts through Big Data technologies. This will definitely be a faster and more effective process than what RED analysts currently now do to come up with campaign insights.

With this commitment, RED needs to make an investment on RED in training current analysts (who have coding backgrounds), or hire a developer to do initial builds for Hadoop and report configuration. RED data analysts can provide more analytics services to its clients now that it has the Big Data capabilities, extending the company service offerings.

Starting with SEM as highlighted in this use case, RED and its clients will see the value of big data, and can slowly move to integration Email campaign data into Hadoop, Web Analytics and Social media campaigns for deeper insights and advanced reporting. The roadmap starts with SEM and ends with "integrated marketing data", as Hadoop is not only limited for online data, it can certainly take offline data and merge it with the online data for deeper insights.

RED can benefit greatly from implementing Big Data, this use case presented how SEM efforts can be optimized and report extraction can be streamlined using Big Data tools. As mentioned previously, it's not only SEM that will benefit from implementing Big Data, any marketing activity is measured using data, thus all teams and RED clients will benefit from the implementation and see better results and deeper insights into the campaigns; allowing RED to yield higher ROI and plan better campaigns.

Considering that RED clients are all not-for-profit organizations, costs associated with maintaining systems is an important factor. RED clients are always looking for open-source or free systems to collect their data, and avoid licensed software. The use of Hadoop is the perfect choice as it robust, efficient and also open-source. RED could also get the hosted service of Hadoop, and then pass the fees over to its client as part of campaign management and reporting services.

VIII. Conclusion

For some companies, Big Data may be a central component of their strategy. For others it could be an expensive distraction. RED definitely fits in the former description, and will reap the benefits of Big Data as soon as it is implemented. RED can model the Big Data system starting with itself (its website and advertising), building a workable prototype that can be shared with its client to get on the system and start seeing higher ROIs and making better decisions.

This use case presents how Search Engine Marketing

efforts can benefit from Big Data; If implemented correctly, the data will automatically feed into the central system, where a ready report or dashboard will show the staff how keywords are doing and how they did historically; analytics of users and traffic will be readily available and in no time, helping the staff fix problems with campaigns quickly, minimizing losses on keyword bids, and maximizing the chances for donation conversion. This makes the SEM channel marketing more effective, especially in the heavy campaigning season in end of year, where the last few days and hours of the year are so valuable to not-for-profits in fundraising.

To gain a 360 view of the customer, marketing data must be integrated to leverage internal and external data sources through the use of Big Data. RED can then have a complete understanding of customer behavior: what makes them donate, in what season, how they prefer to be reached, why they unsubscribe to mailing or advocacy lists, what factors lead them to connect to the organizations on social media. This is strategic for every decision made to maximize ROI and optimize marketing efforts overall.

References

Arthur, L. (2013). *Big data marketing: engage your customers more effectively and drive value*. Hoboken, NJ: Wiley.

Krishnan, K. (2013). *Data warehousing in the age of big data*. Waltham, MA: Morgan Kaufmann.

Lam, C. (2011). *Hadoop in action*. Greenwich, Conn.: Manning Publications.

Welcome to Apache Hadoop. (n.d.). *Apache*. Retrieved October 5, 2014, from http://hadoop.apache.org/

Appendices

Uploading files and loading tables
HIVE Query Results
PIG Script and Results

Appendix 1 : Uploading Files and loading tables
File Upload

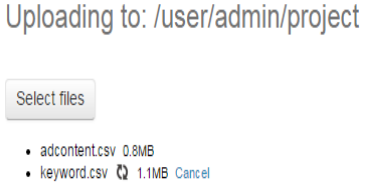
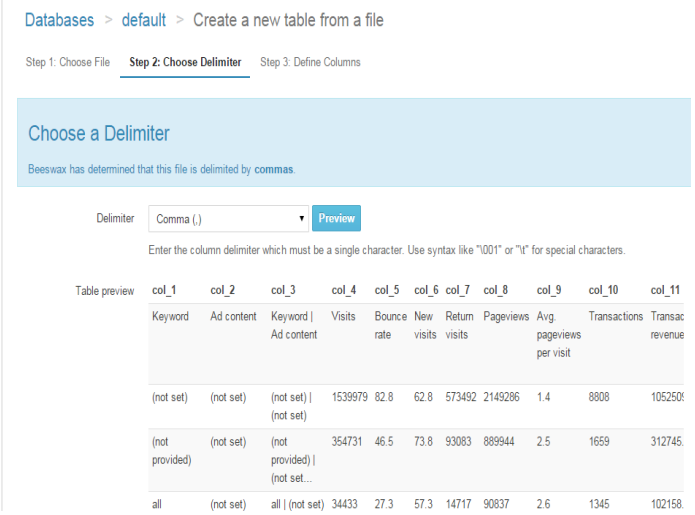
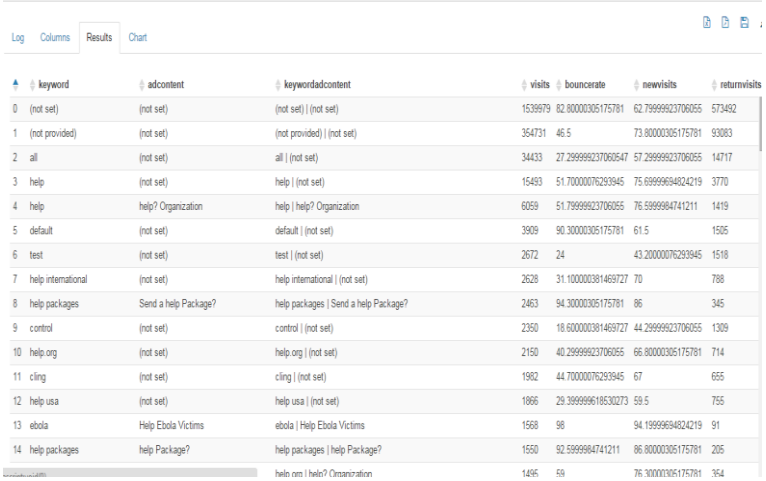


Table creation from file



Browsing the table



Appendix 2 : HIVE Query Results
Q1 Query Results

	keyword	returnvisits	visits
0	all	14717	34433
1	help	3770	15493
2	default	1505	3909
3	help	1419	6059
4	help international	788	2628
5	help usa	755	1866
6	help.org	714	2150
7	cling	655	1982
8	nondonors	546	682
9	monthly	537	623

Q2 Query Results

	keyword	visits	transactions	transactionrevenu
0	ebola	418	9	227.25
1	ebola charity	77	5	163.6199951171875
2	help ebola	23	5	227.25

Q3 Query Results

	adgroup
0	Typhoon Haiyan
1	Typhoon
2	Philippines Typhoon

Q5 Query Results

	keyword	avgpageviewsp
0	http://www.help.org/emergency/horn-africa-hunger-crisis	17.20000076293
1	world hunger in africa	6.599999904632
2	child marriage child hunger child preventable disease	6.400000095367
3	chronic hunger in sahel region...what can be done to alleviate it	3.599999904632
4	"tote ""help"" hunger organization"	3.599999904632
5	poverty and hunger in madagascar	3.599999904632
6	world hunger statistics	3.599999904632
7	an organization that helps with poverty and hunger	3.400000095367
8	stop world hunger	3.200000047683
9	help fight hunger	3
10	help hunger	3

Q6 Query Results

	adcontent	transactions
0	help? Organization	209
1	help? Organization	105
2	help? Double Your Gift	36
3	help?: Your Gift Matched	36
4	help? Organization	36

Q7 Query Results

Recent queries | Query | Log | Columns | Results | Chart

	adgroup	roi
0	South Sudan News	1.1618246436957071
1	South Sudan	0.3717920569701719

Appendix 3 : PIG Script and Results

Q8 PIG Script

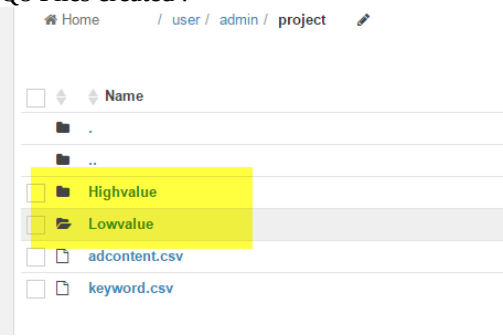
```
A = LOAD '/user/admin/project/adcontent.csv'
USING PigStorage(',')
AS (
 adcontent: CHARARRAY,
 adgroup: CHARARRAY,
 adslot: CHARARRAY,
 addestinationurl: CHARARRAY,
 adcontentdesturl: CHARARRAY,
 adclicks: INT,
 adcost:FLOAT,
 cpc:FLOAT,
 cpm:FLOAT,
 ctr:FLOAT,
 adimpressions: INT
);

B = LOAD '/user/admin/project/adcontent.csv'
USING PigStorage(',')
AS (
 keyword: CHARARRAY,
 adcontent: CHARARRAY,
 keywordadcontent: CHARARRAY,
 visits: INT,
 bouncerate:FLOAT,
 newvisits:FLOAT,
 returnvisits: INT,
 pageviews: INT,
 avgpageviewspervisit:FLOAT,
 transactions: INT,
 transactionrevenue:FLOAT,
 avgrevenuepervisit:FLOAT,
 avgordervalue:FLOAT,
 totaltimeonsitehm: CHARARRAY,
 avgvisitspervisitor:FLOAT
);

C = JOIN B by $1 LEFT OUTER, A BY $0;
D = FILTER C BY NOT(keyword MATCHES 'not provided|not set|test|control');
SPLIT D INTO X IF (transactionrevenue >5000 ), Y IF (transactionrevenue >100 AND transactionrevenue <5000 );
```

```
STORE X INTO '/user/admin/project/Highvalue' using PigStorage(',');
STORE Y INTO '/user/admin/project/Lowvalue' using PigStorage(',');
```

Q8 Files created :



Q8 : Browsing results


```
Int'l Day of the Girl,IDG,Top,,,168.0,279,1,53.88,3,5187.0,,,,,,,,,,,,,,,
International Women's Day,IWD,RHS,,,61.0,113,1,4.13,0,27447.0,,,,,,,,,,,,,,,
International Women's Day,IWD,Top,,,193.0,350,1,48.57,2,7223.0,,,,,,,,,,,,,,,
International Women's Day,IWD,RHS,,,44.0,81,1,3.74,0,21876.0,,,,,,,,,,,,,,,
International Women's Day,IWD,Top,,,162.0,279,1,51.93,3,5383.0,,,,,,,,,,,,,,,
International Women's Day,IWD,RHS,,,31.0,58,1,2.43,0,24237.0,,,,,,,,,,,,,,,
help? Global Organization,help,Top,,,298.0,529,1,44.39,2,11923.0,,,,,help,Brand Keywords,RHS,https
anding&s_src=redggbrandmicrosite6800&s_subsrc=brandempower&utm_source=googlegrants&utm_medium=cpc&
n=redmicrosite6800brand,help | help | help | help,,,,,,622
help? Global Organization,help,Top,,,298.0,529,1,44.39,2,11923.0,,,,,help,Brand Keywords,RHS,https
anding&s_src=redggbrandmicrosite6800&s_subsrc=brandpoverty&utm_source=googlegrants&utm_medium=cpc&
```


```
$30 Can Help Stop Ebola,CDC,Top,,,3.0,5,1,7.0,0,733.0,,,,,,,,,,,,,,,
$8 Can Help Stop Ebola,CDC,Top,,,5.0,14,2,24.43,0,612.0,,,,,,,,,,,,,,,
$8 Can Help Stop Ebola,CDC,Top,,,16.0,40,2,19.23,0,2083.0,,,,,,,,,,,,,,,
Help help Stop Ebola,CDC,RHS,,,1.0,4,4,11.08,0,388.0,,,,,,,,,,,,,,,
Help help Stop Ebola,CDC,Top,,,3.0,11,3,25.68,0,440.0,,,,,,,,,,,,,,,
$30 Can Help Stop Ebola,CDC,Top,,,19.0,44,2,18.21,0,2435.0,,,,,,,,,,,,,,,
Help help Stop Ebola,CDC,Top,,,21.0,53,2,24.6,0,2178.0,,,,,,,,,,,,,,,
Help help Stop Ebola,CDC,RHS,,,2.0,1,0,1.64,0,1032.0,,,,,,,,,,,,,,,
$30 Can Help Stop Ebola,CDC,RHS,,,1.0,5,5,2.15,0,2330.0,,,,,,,,,,,,,,,
$30 Can Help Stop Ebola,CDC,RHS,,,0.0,0,,0.0,0,700.0,,,,,,,,,,,,,,,
$8 Can Help Stop Ebola,CDC,RHS,,,0.0,0,,0.0,0,482.0,,,,,,,,,,,,,,,
$8 Can Help Stop Ebola,CDC,RHS,,,0.0,0,,0.0,0,1223.0,,,,,,,,,,,,,,,
help: {KeyWord:Tax Deduction},EOY,Top,,,0.0,0,,0.0,0,118.0,,,,,,,,,,,,,,,
help: {KeyWord:Tax Deduction},EOY,RHS,,,0.0,0,,0.0,0,246.0,,,,,,,,,,,,,,,
Int'l Day of the Girl,IDG,RHS,,,16.0,28,1,6.6,0,4351.0,,,,,,,,,,,,,,,
Day of the Girl / help,IDG,RHS,,,1.0,1,1,2.34,0,829.0,,,,,,,,,,,,,,,
IDG: Help Girls Worldwide,IDG,RHS,,,0.0,0,,0.0,0,151.0,,,,,,,,,,,,,,,
```