Big Data Analysis Combining Website Visit Logs with User Segments and Website Content

By

Matic Kladnik, Blaž Fortuna, Pat Moore

Ljubljana, 10 October 2016

Motivation

- Visitors of websites leave traces:

  Web server logs, pixel tracking logs, query logs

- Website owners analyze traces to:
  - get knowledge of visitors behavior on their website
  - identify their interest
  - optimize advertisement

Motivation

- Every user has an ID stored in a cookie
- Third party sources log user interest on other sites (topics, products)
- Typical use-cases: Google Analytics, ComScore, … (end-user tools)
- Sophisticated use-cases (ad-hoc processing of data): MapReduce framework (Apache Hadoop)

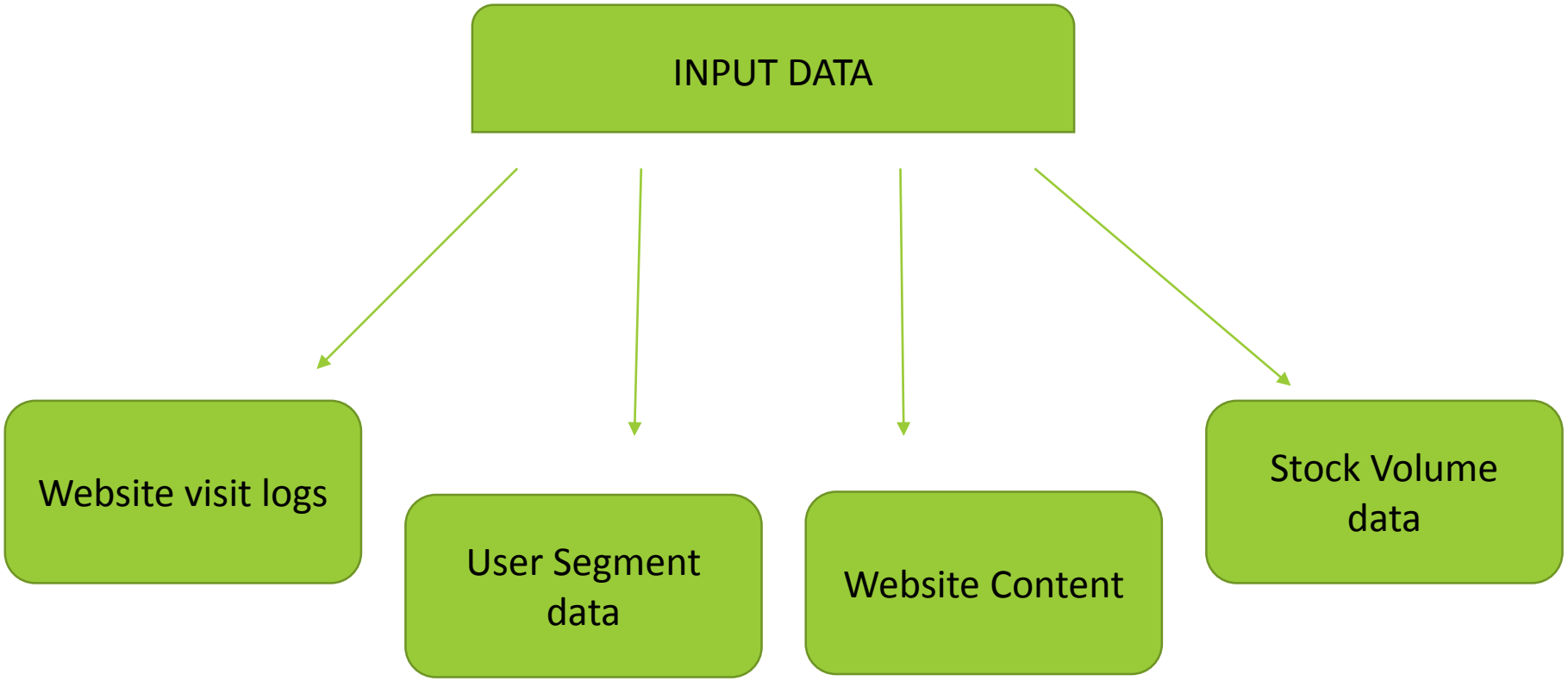
Motivation

- Data on different users can come from different sources
- Use all this data for:
  - Analysis
  - Multiple Use cases

Outline

- Input data
- Preparing data -> dealing with scalability
- Use cases:
    - Comparing user segments
    - Identifying user segments by example article
    - Comparing company news visit logs with stock volume

Data


INPUT DATA
Website visit logs
User Segment data
Website Content
Stock Volume data

Data: Website Visit Logs

- Log of page views on a specific site
- Fields: user ID (cookie), time, page URL, …

- Example:

ns_vid=98837f72…425330&ns_utc=1462852828400&time=28400&type=hidden&name=politics.articles.2016-05-09.escaping-24-hour-coverage-of-donald-trump-is-easy-in-north-korea&ip=x.z.y.z&ns_vid_class=cookie&agent=Mozilla%2f5.0%20(Windows%20NT%206.1)…&kruxid=…

Data: User Segment

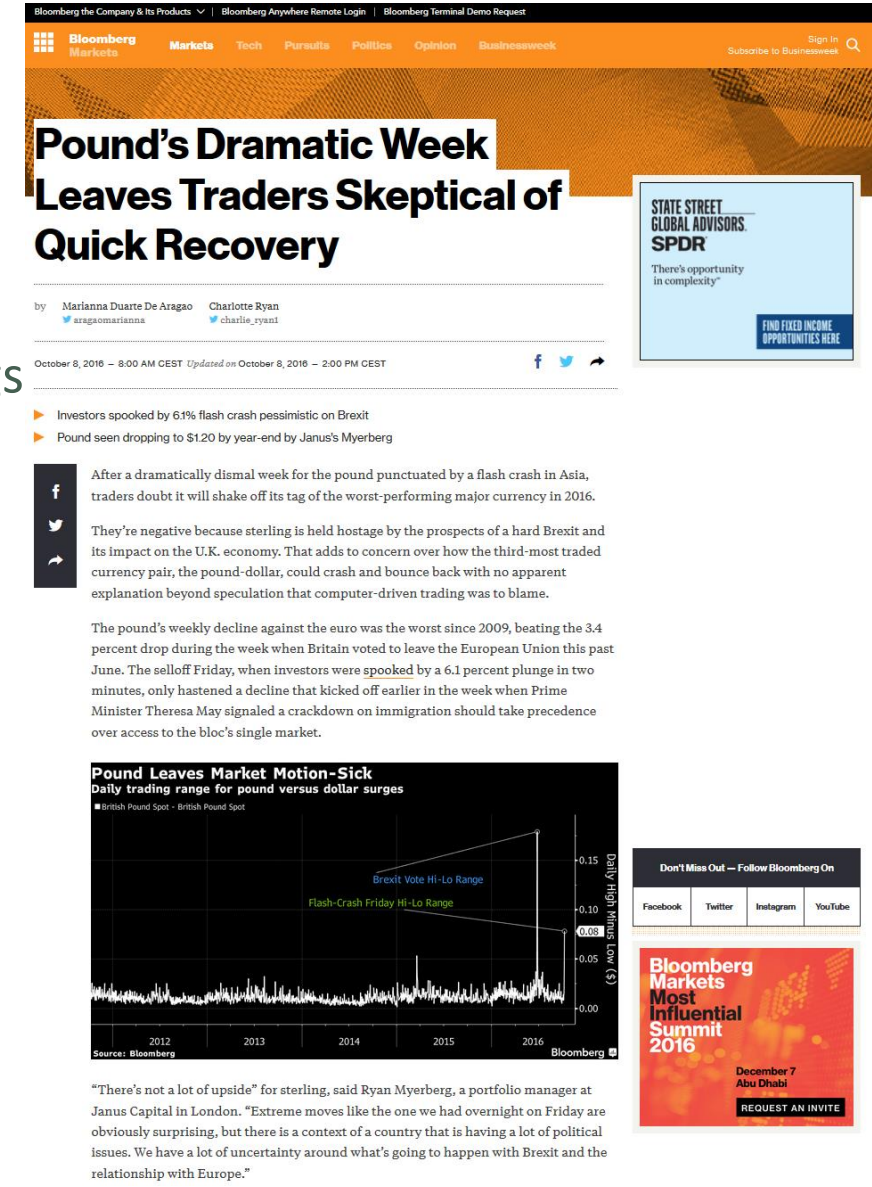
- Estimated Gender, Household Income, Occupation, Age, Country, Free Time Activities, etc.
- User's visit to a car dealership website noted by a third-party data provider (tracking)

```
data_provider;segment_category;segment_name;segment_id
DataLogix;Occupation (Minor);Chemist;HJ...EW
Acxiom;Income;$30,000 - $39,999;JM...EW
```

Data: Website Content

- crawled from URLs in Website Visit Logs
- additional data to use in analysis
- words, entities, topics
- Each page is processed
  using a standard NLP pipeline:
  - list of topics
  - named entities
  - tickers (stocks from companies)

AAPL – Apple Inc.
AMZN - Amazon



Bloomberg the Company & Its Products | Bloomberg Anywhere Remote Login | Bloomberg Terminal Demo Request

Bloomberg Markets | Markets | Tech | Pursuits | Politics | Opinion | Businessweek | Sign In | Subscribe to Businessweek

Pound's Dramatic Week Leaves Traders Skeptical of Quick Recovery

by Marianna Duarte De Aragao (@aragaomarianna) Charlotte Ryan (@charlie_ryan)

October 8, 2016 – 8:00 AM CEST *Updated on* October 8, 2016 – 2:00 PM CEST

- Investors spooked by 6.1% flash crash pessimistic on Brexit
- Pound seen dropping to $1.20 by year-end by Janus's Myerberg

After a dramatically dismal week for the pound punctuated by a flash crash in Asia, traders doubt it will shake off its tag of the worst-performing major currency in 2016.

They're negative because sterling is held hostage by the prospects of a hard Brexit and its impact on the U.K. economy. That adds to concern over how the third-most traded currency pair, the pound-dollar, could crash and bounce back with no apparent explanation beyond speculation that computer-driven trading was to blame.

The pound's weekly decline against the euro was the worst since 2009, beating the 3.4 percent drop during the week when Britain voted to leave the European Union this past June. The selloff Friday, when investors were spooked by a 6.1 percent plunge in two minutes, only hastened a decline that kicked off earlier in the week when Prime Minister Theresa May signaled a crackdown on immigration should take precedence over access to the bloc's single market.

"There's not a lot of upside" for sterling, said Ryan Myerberg, a portfolio manager at Janus Capital in London. "Extreme moves like the one we had overnight on Friday are obviously surprising, but there is a context of a country that is having a lot of political issues. We have a lot of uncertainty around what's going to happen with Brexit and the relationship with Europe."

http://www.bloomberg.com/news/articles/2016-10-08/pound-s-dramatic-week-leaves-traders-skeptical-of-quick-recovery

Data: Stock Volume

- Trading data of some of the most traded stocks
- Example data (AAPL):

```
Date and Time,Open,High,Low,Close,Volume
2016-07-11 15:30,96.53,96.57,96.4,96.45,2646169
2016-07-11 16:00,96.45,96.7,96.37,96.39,1683786
2016-07-11 16:30,96.39,96.44,96.39,96.44,13403
2016-07-11 17:00,96.4,96.45,96.4,96.45,5135
2016-07-11 17:30,96.44,96.45,96.42,96.45,424976
```

Processing

- 1 month of Website Visit Logs consists of:
  - around 138.5 GB of (zipped) data
  - around 546 million rows in thousands of files

- 1 week of Website Visit Logs consists of:
  - around 32 GB of (zipped) data
  - 125,887,958 rows in hundreds of files

Properties of the Data

- Data is structured

ID, IP, value, …

- A lot of data

- Each row can be easily processed separately

Apache Hadoop and Hive

- Hadoop
  - Framework for distributed processing of large datasets
  - Computers in clusters
  - Scale from 1 to thousands of machines
  - Each machine used for local computation and storage

- Hive
  - Data Warehouse running on top of Hadoop
  - Distributed processing of structured data
  - Uses MapReduce methodology to split query into multiple tasks
  - Uses SQL-like language

Aggregating Visit Logs, Segment Data and Content

- Aggregating data by matching segment ID and user ID values
- Performance improvement not exactly linear

Instances	Duration	Improvement
2	3104 seconds	--
4	1763 seconds	x1.76
6	1427 seconds	x2.18

Processing query on 6 instances is done in less than half the time it took 2 instances to complete

- Group visit logs for each User ID

External Table Performance

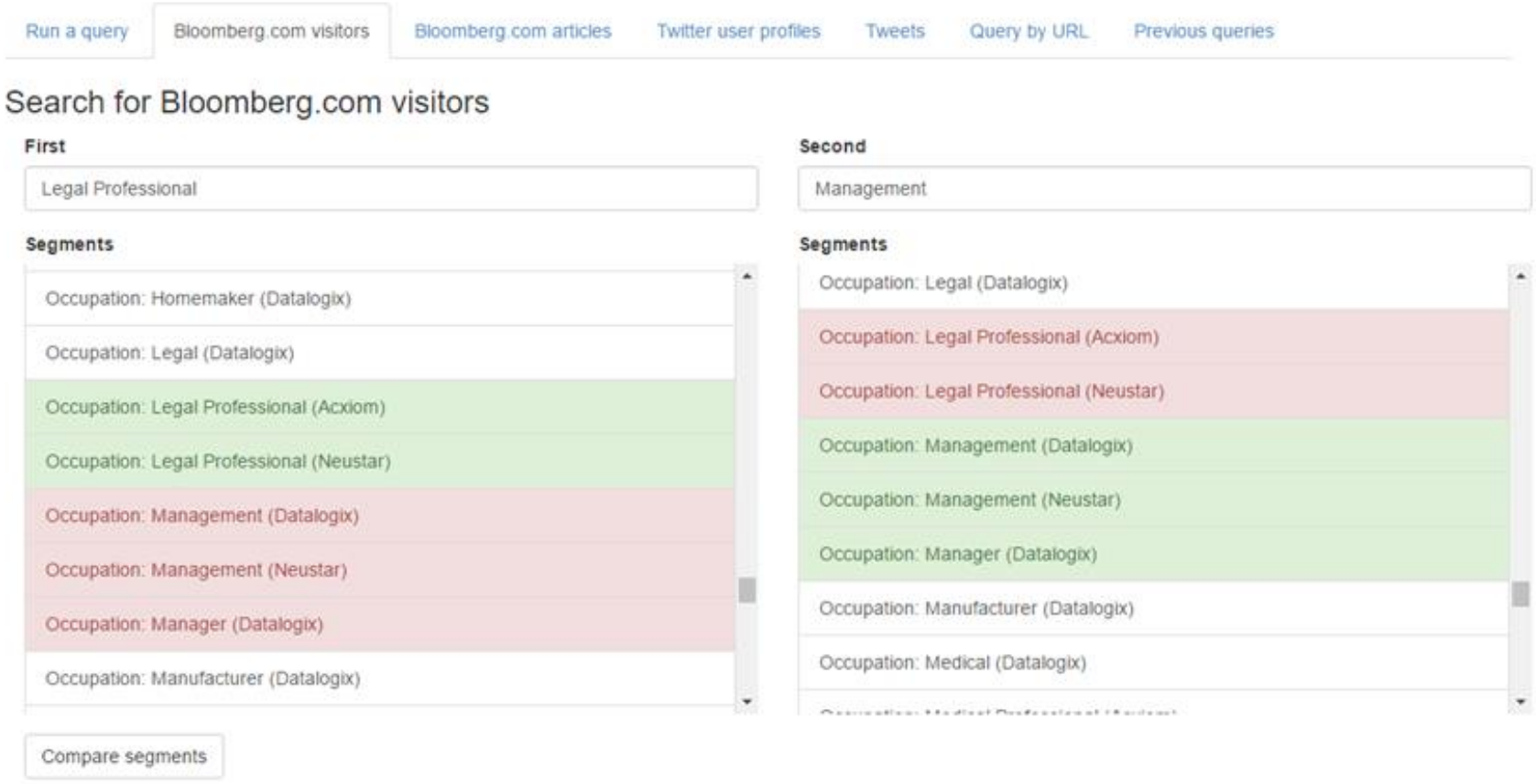
Instances	Duration	Improvement
2	713 seconds	--
4	446 seconds	x1.60
6	344 seconds	X2.07

Internal Table Performance

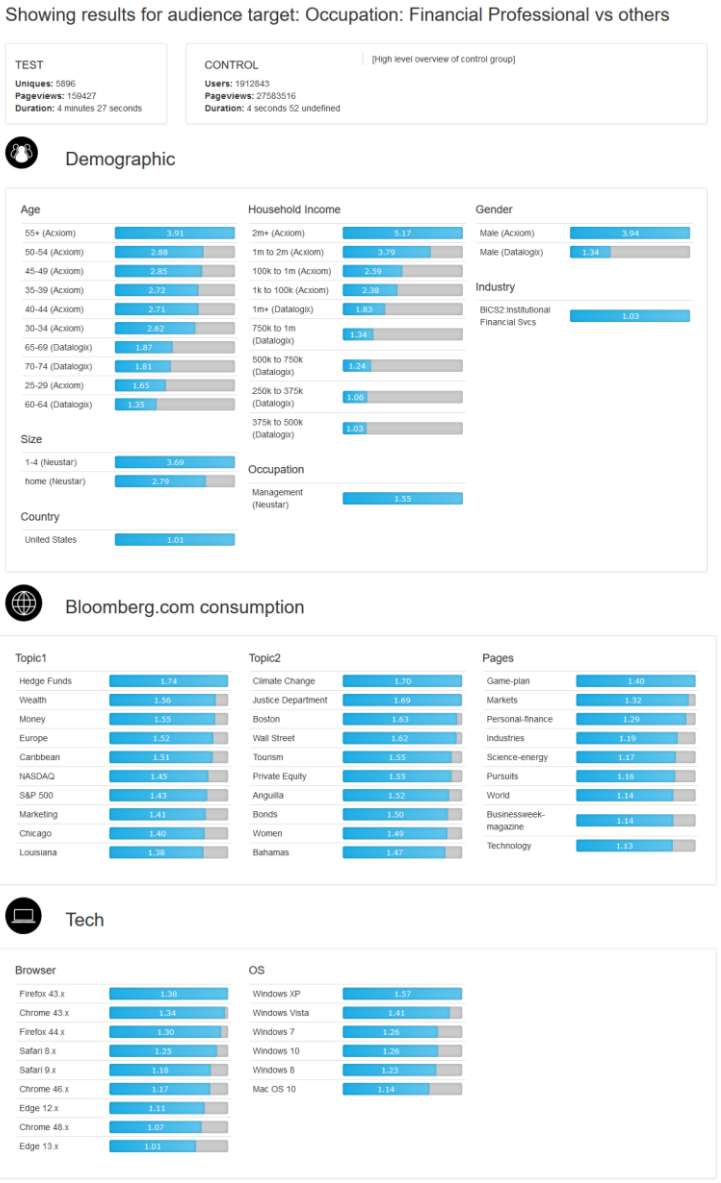
Instances	Duration	Improvement
2	275 seconds	--
4	265 seconds	x1.04
6	234 seconds	x1.18

Use Case: Comparing User Segments

- Compare behavior of different user segments
  - Optimize advertisement
  - Write articles they are interested in



Comparing User Segments



Comparing User Segments

Demographic

Age

55+ (Acxiom)	3.91
50-54 (Acxiom)	2.88
45-49 (Acxiom)	2.85
35-39 (Acxiom)	2.72
40-44 (Acxiom)	2.71
30-34 (Acxiom)	2.62
65-69 (Datalogix)	1.87
70-74 (Datalogix)	1.81
25-29 (Acxiom)	1.65
60-64 (Datalogix)	1.35

Size

1-4 (Neustar)	3.69
home (Neustar)	2.79

Country

United States	1.01

Household Income

2m+ (Acxiom)	5.17
1m to 2m (Acxiom)	3.79
100k to 1m (Acxiom)	2.59
1k to 100k (Acxiom)	2.38
1m+ (Datalogix)	1.83
750k to 1m (Datalogix)	1.34
500k to 750k (Datalogix)	1.24
250k to 375k (Datalogix)	1.06
375k to 500k (Datalogix)	1.03

Occupation

Management (Neustar)	1.55

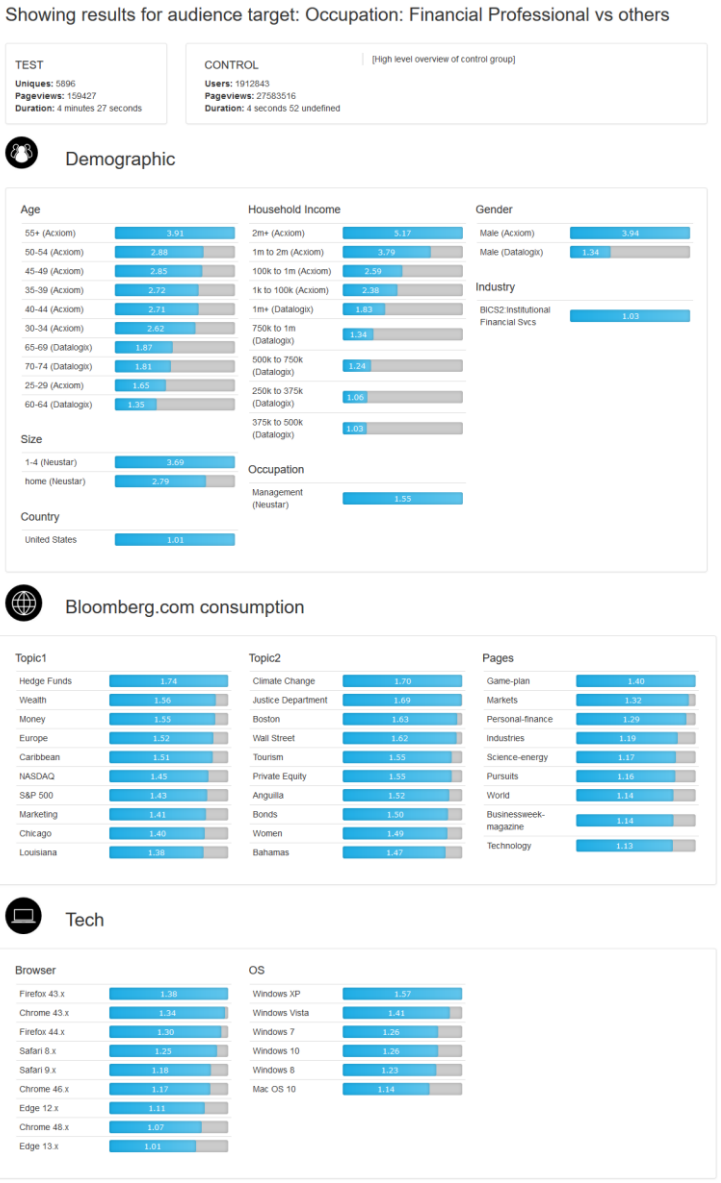
Gender

Male (Acxiom)	3.94
Male (Datalogix)	1.34

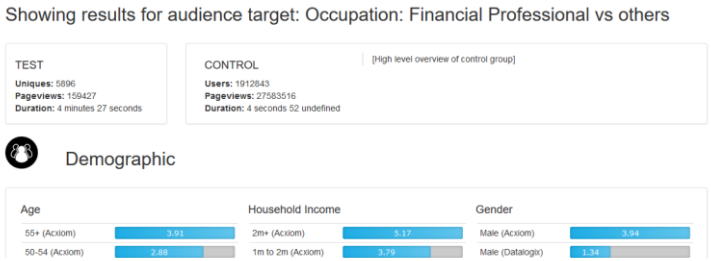
Industry

BICS2:Institutional Financial Svcs	1.03

Comparing User Segments



Comparing User Segments



Tech

Browser

Firefox 43.x	1.38
Chrome 43.x	1.34
Firefox 44.x	1.30
Safari 8.x	1.25
Safari 9.x	1.18
Chrome 46.x	1.17
Edge 12.x	1.11
Chrome 48.x	1.07
Edge 13.x	1.01

OS

Windows XP	1.57
Windows Vista	1.41
Windows 7	1.26
Windows 10	1.26
Windows 8	1.23
Mac OS 10	1.14

Use Case: Analyzing Interests for Topics

- Understand who is interested in specific topics
- Get insight into the article audience
- Topics selected from predefined list or defined via a document (Wikipedia page)
- Example: Advertisement

Number of Stories: 100
Total Pageviews: 2,231
Categories: Technology (81%), Industries (10%), World (7%)

Stories

Show 10 entries CSV Excel PDF Search:

Similarity	Title	Pageview	Viral lift
10.9%	Private Equity's IHeart Radio Chokes on Debt Load It Can't Repay	1,670	3.18
13.2%	Sex Doesn't Sell After All, Study Says	235	3.00
23.5%	TV Ads Are About to Get Personal With New Targeting Tools	145	2.00
11.4%	Facebook's Second-Quarter Revenue, Profit Tops Estimates	120	1.00
12.0%	Why TV Networks Are Cutting Back on Commercials	98	1.00
11.5%	Apple Shifts TV Ads In-House as Chiat\Day Rift Widens	64	1.00
11.1%	The CDC's Anti-Smoking Ads Now Include E-Cigarettes	38	1.00
10.8%	Tesla's Best Celebrity Evangelist Emerges: It's Stephen Colbert	34	1.00
15.6%	Coming Soon to Your Smartwatch: Ads Targeting Captive Eyeballs	30	1.00
13.9%	Google Unveils Mobile News Reader to Compete With Facebook	19	1.00

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 … 10 Next

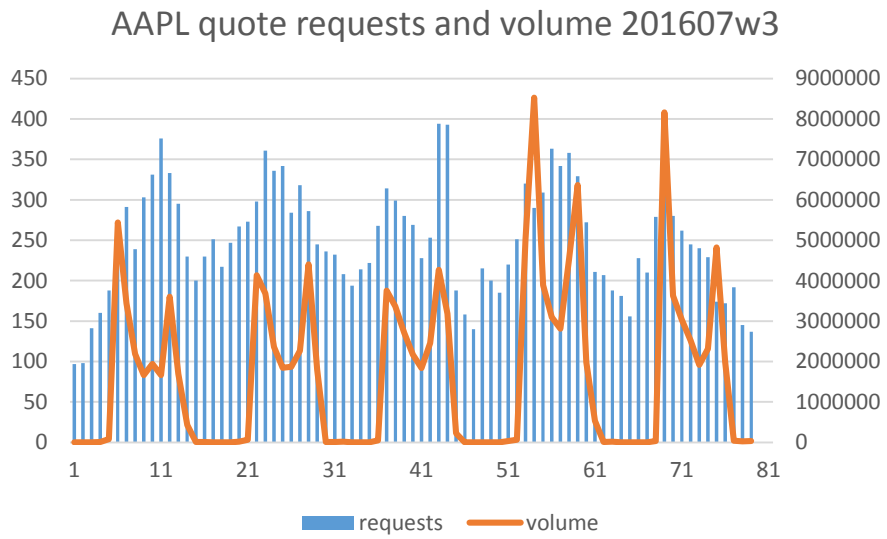
Top 10 topics 1:
Tech (39%), Advertising (21%), Markets (16%), TV Advertising (5%), Hedge Funds (2%), Mobile Phones (2%), Debt (2%)
Top 10 topics 2:
Advertising (28.39%), Markets (17.37%), Tech (11.44%), Media (5.08%), Marketing (5.08%), Social Network (2.12%), Mobile Phones (1.69%), Software (1.69%), Private Equity (1.27%), Internet (1.27%)

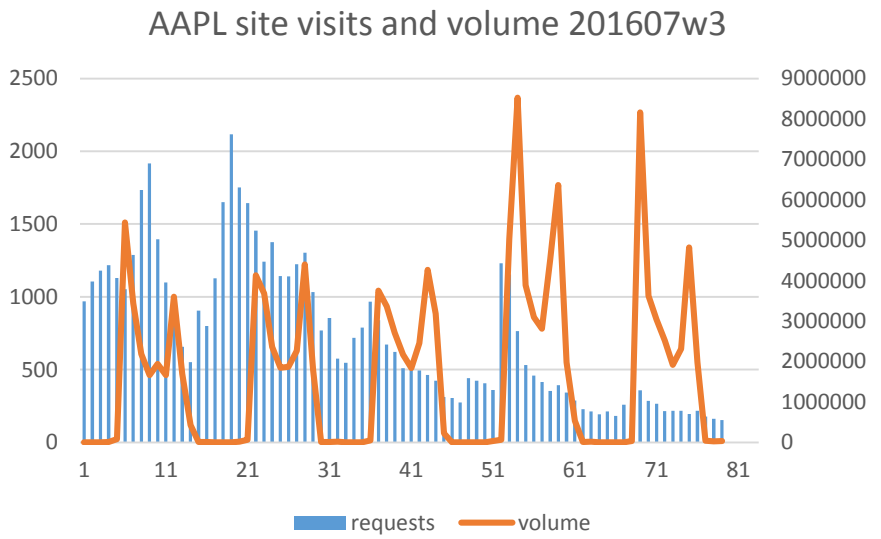
Keywords

Top 10 people:
Mark Elliot Zuckerberg (5.56%), Sheryl Kara Sandberg (3.97%), Douglas MacMillan (3.17%), Arianna S Huffington (2.38%), Andy Pollack (1.98%), James R Callan (1.98%), Martin Stuart Sorrell (1.59%), Barack H Obama (1.59%), Tim Cook (1.59%), Marissa A Mayer (1.19%)

Use Case: Comparing traffic and stock volume data

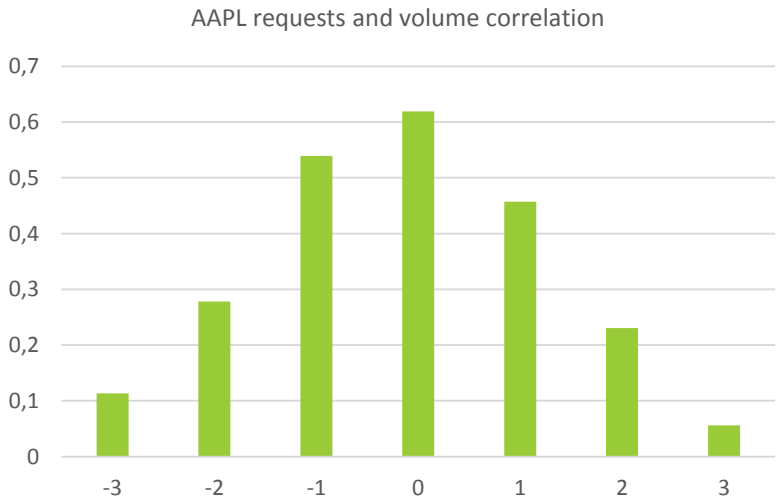
- Comparing correlation between:
  - stock volume data and quote webpage requests and
  - stock volume data and visits to relevant article sites
- Using Apple Inc. (AAPL) as target company

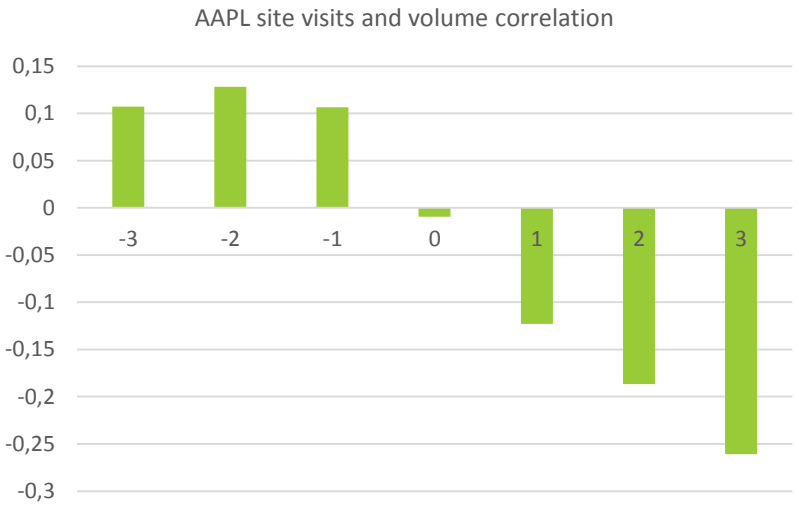




Use Case: Comparing traffic and stock volume data

- Comparing correlation between stock volume values with values for quote requests and visits to relevant sites
- Higher correlation with AAPL quote requests
- We look at information about the stock before we trade it





Conclusion

- Certain level of processing scalability -> many use cases and make additional analysis on data
- Improvement not exactly linear when adding instances to cluster
- Significantly high correlation between stock volume and quote requests