Big Data @ comScore
Turning Billions of Rows into Market Insight

Michael Brown
March 9th, 2012

comScore is a Global Leader in Measuring the Digital World

<table>
<thead>
<tr>
<th>NASDAQ</th>
<th>SCOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clients</td>
<td>1860+ worldwide</td>
</tr>
<tr>
<td>Employees</td>
<td>1000+</td>
</tr>
<tr>
<td>Headquarters</td>
<td>Reston, VA</td>
</tr>
<tr>
<td>Global Coverage</td>
<td>170+ countries under measurement; 43 markets reported</td>
</tr>
<tr>
<td>Local Presence</td>
<td>32 locations in 23 countries</td>
</tr>
</tbody>
</table>
Some of our Clients

<table>
<thead>
<tr>
<th>Media</th>
<th>Agencies</th>
<th>Telecom/Mobile</th>
<th>Financial</th>
<th>Retail</th>
<th>Travel</th>
<th>CPG</th>
<th>Pharma</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>WPP</td>
<td>AT&amp;T</td>
<td>Apple</td>
<td>Best Buy</td>
<td>Expedia</td>
<td>Procter &amp; Gamble</td>
<td>Microsoft</td>
<td>Dell</td>
</tr>
<tr>
<td>msn</td>
<td>OmnicomGroup</td>
<td>Verizon</td>
<td>Nabisco</td>
<td>Target</td>
<td>Travelocity</td>
<td>Travelocity</td>
<td>IBM</td>
<td>HP</td>
</tr>
<tr>
<td>yahoo</td>
<td>Starcom</td>
<td>Sprint</td>
<td>Nabisco</td>
<td>Target</td>
<td>Travelocity</td>
<td>Travelocity</td>
<td>eBay</td>
<td>Cisco</td>
</tr>
<tr>
<td>facebook</td>
<td>WPP</td>
<td>T-Mobile</td>
<td>Walmart</td>
<td>Target</td>
<td>Travelocity</td>
<td>Travelocity</td>
<td>eBay</td>
<td>Dell</td>
</tr>
<tr>
<td>AOL</td>
<td>OmnicomGroup</td>
<td>Liberty Media</td>
<td>Nike</td>
<td>Target</td>
<td>Travelocity</td>
<td>Travelocity</td>
<td>eBay</td>
<td>Dell</td>
</tr>
<tr>
<td>comScore</td>
<td>WPP</td>
<td>AT&amp;T</td>
<td>Nestle</td>
<td>Target</td>
<td>Travelocity</td>
<td>Travelocity</td>
<td>eBay</td>
<td>Dell</td>
</tr>
<tr>
<td>Media 47</td>
<td>WPP</td>
<td>T-Mobile</td>
<td>Nestle</td>
<td>Target</td>
<td>Travelocity</td>
<td>Travelocity</td>
<td>eBay</td>
<td>Dell</td>
</tr>
<tr>
<td>Media 47</td>
<td>WPP</td>
<td>T-Mobile</td>
<td>Nestle</td>
<td>Target</td>
<td>Travelocity</td>
<td>Travelocity</td>
<td>eBay</td>
<td>Dell</td>
</tr>
</tbody>
</table>

The Trusted Source for Digital Intelligence Across Vertical Markets

- 9 out of the top 10 INVESTMENT BANKS
- 4 out of the top 4 WIRELESS CARRIERS
- 47 out of the top 50 ONLINE PROPERTIES
- 45 out of the top 50 ADVERTISING AGENCIES
- 9 out of the top 10 MAJOR MEDIA COMPANIES
- 9 out of the top 10 AUTO INSURERS
- 11 out of the top 12 INTERNET SERVICE PROVIDERS
- 14 out of the top 15 PHARMACEUTICAL COMPANIES
- 11 out of the top 12 CONSUMER FINANCE COMPANIES
- 8 out of the top 10 CPG COMPANIES
Unified Digital Measurement™ (UDM) Establishes Platform For Panel + Census Data Integration

Unified Digital Measurement (UDM) Patent-Pending Methodology

Adopted by 90% of Top 100 U.S. Media Properties
Worldwide Tags per Month

High Level Data Flow

Panel → Tagging → ETL → SYBASE → Delivery
Product: Device Essentials

Smartphone and ISP market share from R&D with distributed SQL on structured data

Device Essentials helps you answer key questions, such as:

1. What types of devices are used to access the web across countries?
2. What is the relationship between content, device and country?
3. How much smartphones traffic is on WiFi vs. wireless carrier by OS?
4. Which wireless networks carry the largest share of traffic by device?
What types of devices are used to access the web across countries?

Non-PC Devices – Share of Total Traffic

<table>
<thead>
<tr>
<th>Country</th>
<th>Mobile</th>
<th>Tablet</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>0.6%</td>
<td>1.6%</td>
<td>6.2%</td>
</tr>
<tr>
<td>Singapore</td>
<td>0%</td>
<td>1.6%</td>
<td>5.9%</td>
</tr>
<tr>
<td>UK</td>
<td>1.0%</td>
<td>4.6%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Japan</td>
<td>2.3%</td>
<td>4.3%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Australia</td>
<td>3.5%</td>
<td>2.0%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Canada</td>
<td>1.5%</td>
<td>1.6%</td>
<td>3.4%</td>
</tr>
<tr>
<td>India</td>
<td>2.3%</td>
<td>1.5%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Spain</td>
<td>2.3%</td>
<td>1.5%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Chile</td>
<td>1.6%</td>
<td>2.3%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Germany</td>
<td>1.5%</td>
<td>1.6%</td>
<td>3.4%</td>
</tr>
<tr>
<td>France</td>
<td>1.5%</td>
<td>2.3%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Argentina</td>
<td>1.1%</td>
<td>2.3%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Brazil</td>
<td>0.6%</td>
<td>2.3%</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

% of country total page requests

What is the relationship between content, device and country?

Share of Total Traffic – Overall and Newspaper Category

- Other
- Tablet
- Mobile

Source: comScore Device Essentials TM – International Data May 2011
How much smartphones traffic is on WiFi vs. wireless carrier by OS?

**Smartphone Connection Type by OS - US**

- **iPhone**
  - Mobile: 53%
  - WiFi/LAN: 47%

- **Android**
  - Mobile: 78%
  - WiFi/LAN: 22%

Source: comScore Device Essentials TM – International Data May 2011

Which wireless networks carry the largest share of traffic by device?

**Provider Share of Traffic**

- **Mobile Total**
  - AT&T Wireless: 27%
  - Verizon Wireless: 21%
  - Sprint PCS: 10%
  - T-Mobile: 9%
  - Other (Wireline): 13%
  - Other (Wireless): 15%

- **iPhone**
  - AT&T Wireless: 27%
  - Verizon Wireless: 21%
  - Sprint PCS: 10%
  - T-Mobile: 9%
  - Other (Wireline): 13%
  - Other (Wireless): 15%

- **Android**
  - AT&T Wireless: 27%
  - Verizon Wireless: 21%
  - Sprint PCS: 10%
  - T-Mobile: 9%
  - Other (Wireline): 13%
  - Other (Wireless): 15%

Source: comScore Device Essentials TM – US Data June 2011
Device Essentials – What do we measure from the tag?

From each event on our network comScore receives...

<table>
<thead>
<tr>
<th>Event</th>
<th>IP Address</th>
<th>User Agent</th>
<th>Cookie</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Timestamp</td>
<td>• ISP</td>
<td>• OS</td>
<td>• 3rd party cookie</td>
</tr>
<tr>
<td>• URL</td>
<td>• Geography</td>
<td>• Device</td>
<td>• Persistent over time</td>
</tr>
<tr>
<td>• Event Type</td>
<td>• Country</td>
<td>• Browser Type</td>
<td>and sites</td>
</tr>
<tr>
<td>• Page view</td>
<td>• City</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Ad impression</td>
<td>• Local Market</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Video play</td>
<td>• Wi-Fi vs. Non-WiFi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Survey</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Campaign</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Device Essentials – Data Review

- **12 TBs per day** of new compressed data (372TB per month)
- **30 billion events** per day
- 1 million domains
- **150 million user agents** per month
- **15 billion cookies** per month
- Full page and referring URL as longest strings
- Distributed across cluster by cookie value to limit skew
- URL classification by company and category
Device Essentials – Methodology Design in Greenplum

- **SELECT** 1 week of data to work on a complete subset
  - 1 hour
- **GROUP BY** country, ISP, region, zip, device
  - 30 min
- **AGGREGATE** traffic across groups
  - 10 min
- **JOIN** on ISP and device lookups
  - 20 min
- **STORE** back to Greenplum cluster (distribute randomly give data skew)
  - 10 min
- **EXPORT** directly to corporate file systems and other systems
  - 10 min

<table>
<thead>
<tr>
<th>Action</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>2 hours</td>
</tr>
<tr>
<td>Processing</td>
<td>2 hours</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>4 hours</td>
</tr>
</tbody>
</table>

Greenplum SQL access raw column data. Rapid iteration!

- **Device Essentials - Methodology Design in Greenplum**
  - Initial R&D for proof of concept
    - ~100 lines of code written in PSQL
  - Generated 1 week WW and 1 month by country examples
    - Available on demand
  - Dynamic iteration for new features, QA and client requests
  - Identify mobile carrier ISPs and IP address ranges by
    - mobile device usage
    - mobile optimized site consumption
    - cookie behavior
  - User agent classification
  - Distributed MaxMind apply
  - Traffic Aggregations

No Engineering!
Greenplum cluster hardware

- 100 Node Cluster: 1 Master; 6 ETL; 108 Workers
- Using Dell R510 with 12 600GB 15K RAID, 64GB RAM, 24 cores (HT)
- 2,400 total CPUs; 6.4TB total memory; 540TB total disk space
- 10GB Ethernet
- Loads 400 billion rows in 8 hours
- PSQL, Perl, Python
- Compiled C libraries
- Dozens of users

Device Essentials – Production Solution

- User agent Dictionary
- URL Dictionary
- IP Dictionary
- Tags
- Production Processing
  - Distribute Tag Processing
  - Storage
  - Aggregation
- Sybase Data Warehouse
Introducing Social Essentials

Strategic challenges addressed by Social Essentials insights:

- Compare the reach and frequency of your brand's Facebook presence to other media channels.
- Determine whether you are reaching your target consumer.
- Tailor marketing campaigns and partnerships to your Facebook audience's interests and passions.
- Benchmark your brand's presence on Facebook to that of your competitors.
Segments are created by classifying the:

**Message** & **Location**

Users are classified as "exposed" when they see a message in the news feed or wall.

When fan information appears less prominently in areas the user may not typically see, we classify as "identified".

---

Social Essentials uses comScore’s PC based panel. Messages seen on a mobile phone are not included in this analysis.

---

**comScore Social Network Segments**

- **Brand’s Total Facebook Network**
  - **All Fans**
    - Exposed Fans
    - Newsfeed exposed
  - Identified Fans
    - Profile (wall) exposed
    - Profile (wall) identified

- **All Friends**
  - Exposed Friends
    - Newsfeed exposed
  - Identified Friends
    - Profile (wall) exposed
    - Profile (wall) identified
The vast majority of brand exposure happens in the Newsfeed and on Profile pages.

In May, Starbucks received 156 impressions for every 1 page view on their brand page.

![Bar chart showing the ratio of Facebook Impressions for Fans & Friends of Fans to Page Views on Brand Fan Pages.]

- **Starbucks**: 156 impressions
- **Southwest**: 42 impressions
- **Bing**: 45 impressions

Source: comScore Social Essentials, U.S., May 2011

Social Essentials provides full insight into a brand’s presence on Facebook.

**Key Measures**
- Topline data providing insight into the reach & frequency of a brand on Facebook
- Can include both unpaid and paid brand exposure
- Segments: fans, friends of fans, sponsored stories and social ads

**Profiling**
- Insight into each segments' demographics, interests, search and internet usage intensity

**Fan Value**
- Brand engagement: brand search and site visitation indices
- Target alignment: primary & secondary
- Spending propensity: total and category level

**Competitive Insight**
- 1 competitor (more can be added at additional cost)
- Key Measures comparison
- Fan base overlap
- Advertiser category engagement
Brands should focus on news feed and less on fan page

<table>
<thead>
<tr>
<th>Brand</th>
<th>Ratio of News Feed Exposures to Fan Page Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southwest Airlines</td>
<td>42 to 1</td>
</tr>
<tr>
<td>Bing</td>
<td>45 to 1</td>
</tr>
<tr>
<td>Starbucks</td>
<td>156 to 1</td>
</tr>
</tbody>
</table>

Fans and Friends visit brand’s site more frequently than average

<table>
<thead>
<tr>
<th>Brand</th>
<th>Percent of segment visiting the brand’s site Index in ()</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starbucks</td>
<td>Friends: (331), Fans: (522), Avg Internet User: (463)</td>
</tr>
<tr>
<td>Bing</td>
<td>Friends: (133), Fans: (155), Avg Internet User: (522)</td>
</tr>
<tr>
<td>Southwest</td>
<td>Friends: (266), Fans: (331), Avg Internet User: (522)</td>
</tr>
</tbody>
</table>

Source: comScore Social Essentials, US, May 2011
Bing Fans outperform the average Bing visitor in Search

Bing : Fans and Friends Search Activity vs. Average Bing Searcher

- Searches Per Searcher: 24.3 vs. 40.9
- Searches Per Session: 2.6 vs. 2.8
- Searches Per Day: 4.6 vs. 5.8
- Search Usage Days: 5.3 vs. 7.1

Starbucks Fans & Friends spend more money and spend more often at Starbucks than the typical Starbucks consumer

Starbucks: Fans & Friends Purchase Activity vs. Average Internet User

- Spending per Buyer: 8% vs. 11%
- Transactions per Buyer: 11%
Social Essentials – Fans are heavy Facebookers

Avg sessions per person

- 10 20 30 40 50 60 70 80 90

Avg Facebook user
Southwest
Bing
Starbucks


Daily Volume of Extracted Facebook Items

Daily Volume of Extracted Entities (MM)

Ads
Brand-News
Brand-Profile

Social Essentials – Data Review

- Social Network page content collected as panelists browse
  - Fewer events than tagging (~2M panelists, ~160 raw FB pages/panelist/day)
  - Average page-size: ~15 KB
- Hadoop (with MapR DFS)
  - Data is too large (~5 TB / day) to analyze without massively parallel processing
    Processing day of data sequentially on a single machine could take ~19 days!
  - Engineers analyze text in Pig, Perl, Java, Python

Social Essentials – Methodology Design

- Hadoop streaming + Perl's regex = Fast text processing
- Page content data has the raw, un-rendered HTML
- Use regular expressions to identify:
  - News feed, ads, fan pages, and other pages
  - Brand exposure
  - Method of exposure
- Building more extensible parser
- Rapid testing in Hadoop of new processing code
- Social Dictionary maps page content to brands
Challenges in Extracting Social Networking Entities

- ~320 million social pages per day
  - 5 TB per day
- Need to process ~3,700 pages/sec just to keep up
  - Avg page-size ~15KB
  - + time to load to HDFS, buffer time for errors, etc...
- Hadoop cluster used to extract entities from streaming data
  - 300 machines together process ~44K pages per second
    - Each machine processes ~145 pages/sec
  - Daily Facebook entity extraction completed in ~3 hours

Hadoop cluster hardware and processing overview

Production Hadoop Cluster

- 75 nodes: Mix of Dell R710 and R510
- 1744 total CPUs; 4.7 TB total memory; 1.1PB total disk space, 1GB Ethernet

Daily: 2 Hr / 70G 1 Hr / 15Gx 0.5 Hr / 15Gx 0.5 Hr / 15Gx
Some Tips & Tricks

Compression w/Sorting

- Compress Log Files when processing large volumes of log data

- Several advantages to Sorting Data First:
  - Reduces the size of the data
  - Improves application performance

- Examples:
  - 1 Hour of our data (313 GB raw, 815 million rows)
  - Standard compression of time ordered data is 93GB (30% of original)
  - Standard compression on a 2 key sorted set is 56GB (18% of original)
  - For one day it saves 800GB
  - For one month it saves 25 TB
  - For 90 days it saves 75TB
Splitting of Data

- We took a sample of our census log data for one hour
  - Row count was 49,404,267 which consumed 28,451,294,113 bytes raw.
  - When GZIPed the size was 3,446,775,819; 12.11% of original
  - We broke the file into smaller files to test parallelism

- Test results of a simple line count program written in Pig

<table>
<thead>
<tr>
<th></th>
<th>Time (seconds)</th>
<th>Rows/Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>One raw file</td>
<td>49</td>
<td>1,008,250</td>
</tr>
<tr>
<td>One compressed file</td>
<td>451</td>
<td>109,453</td>
</tr>
<tr>
<td>Split Compressed (4MM) x13</td>
<td>61</td>
<td>809,906</td>
</tr>
<tr>
<td>Split Compressed (2MM) x25</td>
<td>36</td>
<td>1,372,340</td>
</tr>
</tbody>
</table>

Useful Factoids

Colorful, bite-sized graphical representations of the best discoveries we unearth.

Visit www.comscoredatamine.com or follow @datagems for the latest gems.
Thank You!

Michael Brown  
CTO  
comScore, Inc.

mbrown@comscore.com

"Ha! Webster’s blown his cerebral cortex."