Social Media Analytics Using Greenplum’s Data Computing Appliance

Johann Schleier-Smith
Co-founder & CTO
Tagged Inc.

@jssmith

February 2, 2011
What is Tagged?

- 3rd largest social network in US (minutes spent)
- We enable anyone to meet and socialize with new people
- 100 million member accounts
- Longest sessions of any social network
Social Media Challenges

- Competing products are always a click away
- 1 or 2 years to obsolescence
- 50+ things on every page
- Collective behaviors dominate

**Must have world-class analytics for big data**
Analytics with Greenplum

Case Studies

Best Practices
Some Numbers

- Monthly Metrics
  - 5 billion page views
  - 20 million visitors
  - 3 PB downloaded
  - 10 billion ad impressions

- Monthly Logging
  - 50 billion log entries
  - 120 log streams
  - 10 TB log data
Tagged Infrastructure

Web browsers, mobile apps

Internet

PHP Web Tier

Java Applications

Image Storage

Oracle Operational Database

Greenplum Analytics Database

No ETL
Greenplum Data Computing Appliance

- Tagged deployment since 2008
- 2 cabinets
- 8 x data nodes
  - 8 CPU cores
  - 32G memory
  - 48 x 500G disk
- Compute machine
  - 64 CPU
  - 256G memory
  - 192TB disk
  - 384 spindles
  - 32 gbit network throughput
- Greenplum professional services determined appropriate sizing

Image courtesy of EMC
Greenplum’s compute and data architecture

- Same scalability principles used for big web services
- No data too big
- Always fast since CPU and storage scale together
- RAID + replication for data protection and availability

Diagram courtesy of EMC
Logging Approach

• Events for everything
  – Login
  – Page view
  – Friend request sent, confirmed, rejected
  – Message sent, read, deleted
  – Virtual gift given
  – Pets game trade
  – Farm game crop planted
  – Mafia player attacked

• No dependence on operational database

• Log more rather than less
  – Save data as best suited for analysis
  – Include extra “state of the world” information
  – Don’t think about space or cost constraints

• Keep data for 6-24 months
Analytics Approaches

- Dashboards / reports
- A/B testing
- Ad-hoc study
- Predictive modeling
Dashboards

- Java + JavaScript graphics libraries are our platforms
  - YUI tables and forms
  - Google visualization API
  - Our business is software so we build it ourselves

- Easy for everyone to check metrics, see what may have changed
- Visualizations quick to scroll through, over 100 graphs
- Working on automated anomaly detection
## A/B Testing

### Test Result: 59127

**new_home_whats_new**

| Test ID | 29211
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
</tr>
<tr>
<td>To</td>
</tr>
</tbody>
</table>

#### Test Dates

<table>
<thead>
<tr>
<th>Sun</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>28</td>
<td>29</td>
<td>30</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

### Versions and Summary Data:

<table>
<thead>
<tr>
<th>Version</th>
<th>Weight</th>
<th>Description</th>
<th>Friend Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>&quot;Default&quot;</td>
<td>--</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>&quot;MeetMe static box&quot;</td>
<td>+2.45% (+0.62%)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>&quot;No suggestion box&quot;</td>
<td>+1.67% (+0.6%)</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>&quot;No suggestion box, lower top ad&quot;</td>
<td>+1.88% (+0.6%)</td>
</tr>
</tbody>
</table>

### Profile Views Received

<table>
<thead>
<tr>
<th>Version</th>
<th>Num Users</th>
<th>Total Views Received</th>
<th>Avg Views Received</th>
<th>Err Views Received</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SQL for Ad-hoc Analysis

- Business and Product have questions, we have answers
- Complex, deep analysis for games
- Motivate new feature development
- Declarative language >10x more productive than Java
- Greenplum supports SQL 2003 OLAP extensions
Case Study: Business Operations

Q: why are page views tracking 3% lower than usual today?

- Review dashboards, per-feature and per-country metrics
- Common operational problems
  - Bugs introduced by daily release
  - Slow pages?
  - E-mail problems?
  - Regional holidays
- Want to achieve automated alerting & problem isolation
Case Study: Protests in Egypt

Graph showing unique logins per hour and traffic relative to the previous week.
Case Study: Protests in Egypt


Tagged Monitor

The Tagged International Traffic Monitor graph provides near real-time view of active users per minute by time of day and country.

View: Last 24 Hours  Country: Egypt  Updated: 2011-02-02 13:04:02.130

Active Users Per Minute vs Time of Day (PST)
Case Study: Mobile Usage

**Q: how should Tagged direct its mobile web development?**

- Parse logs
  - Identify User-Agents in HTTP headers that correspond to mobile platforms
  - Count number of users on mobile web browsers
  - Which features do mobile web users engage with most
    - Check messages
    - Browse profiles
    - Play meet me
    - Play pets
  - Now know what features to focus on in mobile web product

- Queries run in under one minute

- Quick iteration helps us develop rules for classifying headers

- Analysis done in one hour
Case Study: Tuning Mafia Game

- Find points of player confusion
- A/B test UI improvements
- Analyst must play game and understand its intricacies
- Log events with plenty of extra game state
- >3x revenue improvement
Case Study: Phishing from a Botnet

Q: *phishing spam is going on now, how do we stop it?*

- Forensic inquiry on phished accounts
  - Is there evidence of a brute-force password attack?
  - Have compromised users been active on Tagged recently?
  - Were credentials stolen from another web site?
  - Is spam highly distributed or coming from a few IP blocks?
  - Are there headers that identify the botnet?
  - What actions are common immediately after phishing takes place?
  - Were game players actively targeted?

- Bayesian classifier systems for spam identification

- Working on automated pattern extraction
Case Study: Matchmaking on 100 million accounts

- Meet Me is the dating feature on Tagged
- Connects people based on social patterns
- 10x increase in usage through A/B testing
Analytics with Greenplum
Case Studies
Best Practices
Best Practices in Social Media Analytics

- Application writes logs specifically for analytics
- Get plenty of space and log a lot
- Data can be queried in real time
- Use SQL for quick ad-hoc studies
- Automate reports with lots of graphs
- Compute statistical uncertainty in A/B tests
Data Computing Appliance Delivers

- Greenplum gobbles up all the data
- Analysts love the speed and SQL advantage
- DBAs find life is easy
- Business gets insightful and complete answers to questions, quickly
- Users enjoy the longest session times of any social network

Image courtesy of EMC