"We are Drowning in (Big) Data but Starving for Knowledge"

Advanced Analytics to the Rescue

Shai Fine
Principal Engineer, Advanced Analytics

May 28, 2013
Thank You!
Big Data Explosion – *some big numbers*

**Over 24 Petabytes**
Data processed by Google every day in **2011**

**7 Exabytes**
Data traffic by mobile users worldwide in **2011**

**4 billion**
pieces of content shared on Facebook every day by **July 2011**

**250 Million**
Tweets per day in **Oct 2011**

**5.5 million**
Legitimate emails sent every second in **2011**

**158 products**
ordered per second on **Cyber Monday in 2010**

**1500+ blog posts**
Every minute in **2011**

**Internet devices: 1000 billion by 2013**
Up from 5 billion in 2010

**Internet traffic to increase 9x by 2013**
From 5 Exabyte a month to 56 Exabyte a month in 2013

**More video was uploaded to YouTube**
In last 2 months, than if ABC, NBC, and CBS had been airing new content since 1948

*Between the birth of the world and 2003, there were 5 Exabyte of information created. We now create 5 Exabyte every 2 days*

Eric Schmidt
Big data is growing exponentially ...

As per IDC, Big Data market is expected to grow from $3.2 billion in 2010 to $16.9 billion in 2015 with CAGR of 40%

7.9 Zettabytes of data will be created by 2015
Up from 1.8 zettabytes in 2011 at 40% CAGR
"BIG DATA IS LIKE TEENAGE SEX: EVERYONE TALKS ABOUT IT, NOBODY REALLY KNOWS HOW TO DO IT, EVERYONE THINKS EVERYONE ELSE IS DOING IT, SO EVERYONE CLAIMS THEY ARE DOING IT..."

PROF. DAN ARIELY. DUKE UNIVERSITY

We are drowning in (big) data, but starving for knowledge
Analytics Drives Innovation

• Most growth and innovation is focused at the analytics layer
• Customers are either paying for analytics consulting services or hiring data scientists

Analytics Examples

<table>
<thead>
<tr>
<th>CEP</th>
<th>Predictive</th>
</tr>
</thead>
<tbody>
<tr>
<td>StreamBase</td>
<td>Opera</td>
</tr>
<tr>
<td>Tibco</td>
<td>KXEN</td>
</tr>
<tr>
<td>MicroStrategy</td>
<td>Sense Networks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BI</th>
<th>Statistical Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qlik</td>
<td>Revolution</td>
</tr>
<tr>
<td>1010data</td>
<td>SPSS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Querying</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenplum</td>
<td></td>
</tr>
</tbody>
</table>

knowledge out

Data in

Visualization

Analytics

Data management

Next Gen

Standard
Analytics to the Rescue!?  

Big Data Will Generate 6 Million U.S. Jobs by 2015 😊😊😊  
(Gartner, 2012)

The United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data 😊😊😊  
(McKinsey 2011)

- Most growth and innovation is focused at the analytics layer
- Customers are either paying for analytics consulting services or hiring data scientists

### Analytics Examples

<table>
<thead>
<tr>
<th>CEP</th>
<th>Predictive</th>
<th>Statistical Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>StreamBase</td>
<td>Opera</td>
<td>Sense Networks</td>
</tr>
<tr>
<td>Tibco</td>
<td>KXEN</td>
<td>Revolution</td>
</tr>
<tr>
<td>MicroStrategy</td>
<td>Qlik</td>
<td>SPSS</td>
</tr>
<tr>
<td>1010data</td>
<td>Greenplum</td>
<td></td>
</tr>
</tbody>
</table>

• Most growth and innovation is focused at the analytics layer
• Customers are either paying for analytics consulting services or hiring data scientists

Knowledge flow diagram showing data in, analytics, data management, visualization, and knowledge out.
Enterprise Analytics in Action

A Sample of Predictive Analytics Solutions

“Your recent Intel purchases, tweet score and location history make you 63.1% welcome here...”
Manufacturing

Why is it a Big-Data Problem

• 100’s of parameters are tested per each die
• Between 100’s to 1,000’s die per each wafer
• 25 wafers in the same lot
  – Same flow through but not exactly the same fabrication process!
    ➢ Roughly ~2.5MB of data/record per product
• 10’s of different products produced in parallel in high-volume-manufacturing fab lines

• Overall, Terabytes of data are being produced every single day
  – BUT the “signal” is very week - defects are extremely rare
  – Still, the quality requirements are very strict
“Personalized Medicine”

- Testing is VERY expensive
- Equipment capital is significant
- Test Program development is complicated and require substantial amount of proficient resources
  - product engineering and coding/programming
- Throughput time is a function of the test duration

- Can the capital and test time duration be reduced via reduction in the testing?
- Personalized testing program
  - Every unit is treated according to its unique historical data utilizing predictive analytics technologies
About Telmap
- Telmap is a mobile app location-based service provider
- Delivers premium navigation service through various cellular service providers
- Intel’s subsidiary since 2011

Personal, Context Aware Location Based Recommendations System
- Provide the right offer, to the right user in a given context to maximize user experience and value
Solution – A two layer Hybrid architecture

Offline

Crunch raw data into meaningful patterns which do not tend to change dramatically

- Raw data
- Recommendation algorithm
- Underline patterns

Run on a scalable platform (Hadoop), Gain scalability

Online

Use latest user data, context and underline patterns to compute user recommendation on demand

- Compute recommendation using computed model and latest data
- Context filtering
- Recommendation

Use latest feedback for real time recommendation

Use context filtering to apply Context layer
Key Data Sources

Internal Sources
- Navigations Logs
- Content Searches

Analytics Engine

External Sources
- User Demographics
- Weather
- Calendar: holidays, weekend/weekday
- Social Media*

*Social Media data includes external sources such as Facebook, YouTube, and LinkedIn.
Recommendation Engine

- **Modeling layer consist of**
  - Collaborative Filtering (Matrix Factorization)
  - Content Based (Logistic Regression)
  - Off-line training is implemented on top of Intel’s Hadoop distribution

- **Context is incorporated in real-time**
  - Pre-Filtering
    - Mainly “time” oriented
    - Filter out offerings
  - Post-Filtering
    - Mainly “location” oriented
    - Effect ranking

- **Personalization is achieved via an interplay between the Modeling and the Filtering layers**

- **Essentially this is a Big Data problem, but**
  - Sparsity level user/items ~0.003%
    - Due to the incorporation of personalized context
  - Interactive system with a continuous flow of useful and transient information
Pilot Overview

• Location and Timeline
  – Conducted in the UK for 9 weeks at 1H 2013

• Offerings are Coupons
  – Only positive feedback is recorded
    • No history for rejected offerings
    • No indication about actual sells

• Training
  – Initial data set
    • ~200K customers; ~400K items; sparsity ~0.003%
  – Adaptation
    • Models are retrained every week using the accumulated data

• Pilot Participants: ~1K enlisted customers
  – Half receiving highly targeted offers
  – Half receiving random location based offers (control group)

• Goal: Increase Convergence Rate
  – Higher CTR on participants with targeted ads Vs. control group
  – Increase CTR over time (first month Vs. second month)
“Wisdom” of the Crowd

- The team placed bets before the pilot
- The stakes are high
  - “A dinner for the winner”

<table>
<thead>
<tr>
<th>Voter:</th>
<th>First month</th>
<th>Second month</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>2</td>
<td>100%</td>
<td>120%</td>
<td>20%</td>
</tr>
<tr>
<td>3</td>
<td>100%</td>
<td>160%</td>
<td>60%</td>
</tr>
<tr>
<td>4</td>
<td>140%</td>
<td>180%</td>
<td>40%</td>
</tr>
<tr>
<td>5</td>
<td>160%</td>
<td>140%</td>
<td>-20%</td>
</tr>
<tr>
<td>6</td>
<td>120%</td>
<td>160%</td>
<td>40%</td>
</tr>
<tr>
<td>7</td>
<td>140%</td>
<td>200%</td>
<td>60%</td>
</tr>
<tr>
<td>8</td>
<td>100%</td>
<td>120%</td>
<td>20%</td>
</tr>
<tr>
<td>9</td>
<td>120%</td>
<td>160%</td>
<td>40%</td>
</tr>
<tr>
<td>10</td>
<td>60%</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>11</td>
<td>180%</td>
<td>240%</td>
<td>60%</td>
</tr>
<tr>
<td>12</td>
<td>20%</td>
<td>20%</td>
<td>0%</td>
</tr>
<tr>
<td>13</td>
<td>120%</td>
<td>140%</td>
<td>20%</td>
</tr>
<tr>
<td>14</td>
<td>140%</td>
<td>160%</td>
<td>20%</td>
</tr>
<tr>
<td>Avg</td>
<td>114%</td>
<td>141%</td>
<td>27%</td>
</tr>
</tbody>
</table>
Wrap-up

• The distinction between Data, Information, and Knowledge
  – Data is available, the challenge is in transforming it to a significant business value

• Big Data & Advanced Analytics
  – The platform perspective – No Silver Bullet
    • Big Data platform will contain several solutions
  – The computation perspective
    • Distributed Learning, Stream Analytics