Big Data for Line-of-Business Content

Jeff Fried
CTO, BA Insight

Data Summit
May 2017
Big Data Is...

big data is
big data is like teenage sex
big data is better data
big data is not a monolith
big data is dead
big data is the future
big data is best processed in a platform that
big data is not about the data
big data is bad
big data is best defined as
big data is easy to define quizlet
About Jeff Fried

Focused on Search and SharePoint since 2004

Longtime Search Nerd
• CTO, BA Insight
• Senior PM, Microsoft
• VP, FAST
• SVP, LingoMotors

Passionate About
• Search
• SharePoint
• Search-driven applications
• Information Strategy

Blog:
BAinsight.com/blog

Technet Column
"A View from the Crawlspace"

jeff.fried@bainsight.com
About BA Insight

- Our software quickly connects SharePoint users to needed information
- Software portfolio includes:
  - **Connectivity** - Secure connectors and federation to over 50 content systems
  - **Applications** - Improve the user experience to boost productivity
  - **Classification** - Autotagging, metadata generation, and text analytics
  - **Analytics** - Insight into how SharePoint portals, search, & content are being used
- Hundreds of successful implementations including:
This session

LoB systems and Unstructured content

Data Discovery examples

Data Science examples

Handling LoB Content
LoB systems and Unstructured content

Process data
Human data
Machine data
Companies don’t use most of their data

SMBs: 9 TB
Unstructured 50 TB

0.6 TB
Semi-structured 2 TB

4 TB
Structured 12 TB

LEs: 75 TB

5 TB

50 TB

Average data volume per company

Only 12% used today

Source: Forrsights Strategy Spotlight: Business Intelligence And Big Data
Structured Data
High Degree of organization, such as a relational database

<table>
<thead>
<tr>
<th>Column</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>Joe Brown</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>02/13/1972</td>
</tr>
<tr>
<td>Date Admitted</td>
<td>02/05/2014</td>
</tr>
</tbody>
</table>

Unstructured Data
Information that is difficult to organize using traditional mechanisms

“The patient came in complaining of chest pain, shortness of breath, and lingering headaches...smokes 2 packs a day...family history of heart disease...has been experiencing similar symptoms for the past 12 hours....”
Connectors to Many Enterprise Systems

- Aderant
- Amazon S3
- Alfresco
- Box
- Confluence
- CuadraSTAR
- Elite / 3E
- EMC Documentum
- EMC eRoom
- Google Drive
- HP Consolidated Archive
- (EAS, aka Zantaz)
- HPE Records Manager/HP TRIM
- IBM Connections
- IBM Content Manager
- IBM DB2
- IBM FileNet P8
- IBM Lotus Notes
- IBM WebSphere
- iManage Work
- Jive
- LegalKEY
- LexisNexis Interaction
- Lotus Notes Databases
- Microsoft Dynamics CRM
- Microsoft Exchange
- Microsoft Exchange Public Folders
- Microsoft SQL Server
- MySQL
- NetDocuments
- Neudesic The Firm Directory
- Objective
- OpenText LiveLink/RM
- OpenText eDOCS DM
- Oracle Database
- Oracle WebCenter
- Oracle WebCenter Content (UCM/Stellent)
- PLC/Practical Law
- ProLaw
- Salesforce.com
- SAP ERP
- ServiceNow
- SharePoint Online
- SharePoint 2016
- SharePoint 2013
- SharePoint 2010
- SharePoint 2007
- Sitecore
- Any SQL-based CRM system
- Veeva Vault
- Veritas Enterprise Vault (Symantec eVault)
- West km
- Xerox DocuShare
- Yammer

Plus a proven architecture and process for creating new connectors to complex systems
The average $1 billion company maintains 48 disparate financial systems and uses 2.7 ERP systems

Source: The Hackett Group
Big Data on LoB Data: 

**Data Science and Data Discovery**

- **BIG DATA**
  - Volume, velocity, and variety of data
  - Potential business impact

- **DATA SCIENCE**
  - Complexity of analysis
  - Potential impact
  - Range of tools
  - Smart algorithms

- **DATA DISCOVERY**
  - Ease of use
  - Agility and flexibility
  - Time-to-results
  - Installed user base

- **DATA DISCOVERY**
  - Limited depth of information exploration
  - Low complexity of analysis
Data Science Examples

- Predictive Inventory Levels to Minimize Warehousing Costs
- Personalized Medicine Treatment Programs
- Smart Meter Monitoring for Customer Value Add
- Customer Churn Analysis for Increased Customer Lifetime Value
- Trade Options and Futures Pricing Platform

Source: PARC
Example: Optimizing Leaseholds & Mineral Rights in Oil & Gas Exploration

- Content:
  - Leasehold and well inventory documents
    - in multiple systems (FileNet, LiveLink)
    - Often scans of paper documents, not uniform format
  - Databases with exploration, operations, and project data
  - Sensor & seismic data
  - Public Market data, Industry data, and News

- Analysis: lease potential yield; likely trades & negotiations; asset utilization; trends
Example: Clinical Trial Management for Pharmaceutical Development

- **Content:**
  - Reports and Regulatory Filings
    - Two RMS’s: Documentum & Veeva Vault
      - structured and unstructured
  - Public research literature and prior trial data
  - Internal development and clinical trial data
  - Marketing & Industry data

- **Analysis:** compliance; risk areas; compound use opportunities
Example: Genetics Analysis for Clinical Research

Location
- Chromosome, Position
Location/Region Identifiers
- SNP (variant, rsid, HGVS, dbsnpid)
- Gene (Ensembl, Entrez, HGNC)

Additional Properties & Observations
- Gene Expression
- Epigenetics

Aggregate Observations
- Linkage Disequilibrium
- Population Allele Frequency
- Expression Quantitative Trait Loci (eQTL)
- Genome Wide Association Study (GWAS)
- Phenome Wide Association Studies (PheWAS)

Predictions/Insights
- Identify disease causal genes/effects
- Reveal unknown biological mechanisms
Example: Insurance Claim Analysis

- **Content:**
  - Databases with both structured and unstructured content
  - ERP process data and claim history
  - Internal actuarial and financial data
  - Partner/Channel data

- **Analysis:** per-claim supplier effectiveness

I am a software designer and sit at a desk all day. I could not sit comfortably for months. I was unable to work until the beginning of November. However, my company could not wait for me to recover and was not able to provide me with my job back. I did have six months of short term disability, which I have to repay. I was unable to get a new job until January 2, so I will be claiming lost earnings from April 4 until the end of December, nine months worth.
Example: (Pharma R&D) Unified View

1. Documentum Image
2. SharePoint Doc
3. Regulatory Record
4. MEDLINE article

Search: amgen 655

Relationships Discovered:
Antibodies: mAb
Receptors: DR5, IGF-1R
Labs: Oncology 1
People: David Chang
Examples: Financial Risk Management

Market Risk
(Risk Portal)

Fraud Risk
(Anti-Money Laundering)

Case Number: 834-2006-0000083

Excerpt: Procedures 1. Gather and verify all customer information. 2. Verify that the SSN/TIN/..account researched. 3. Determine if customer is assigned a TIN/SSN in Hogan. IF THEN The customer has a TIN/SSN assigned in Hogan or thi...

Click for full narrative text

Case Description: This customer wanted to withdraw $18,000 but decided to change the amount once he was informed of CTR. This customer said that he was going to go to another branch to collect to other needed amount. Customer did not mention what he needed this cash for.

- Total Amount: $18,000
- Branch Code: 06510
- Branch Office Code: 1500 10TH AVE
- Branch Address: Hillsboro, OR 97123
- Case Status: SAR-Closed
Example: Analyst Workbench

**Target Details**

**Ilich Ramírez Sánchez**

- **Alias:** Carlos “The Jackal”
- **Date of Birth:** 10/12/49
- **Country of Birth:** Venezuela
- **Last Contact:** 11/2/11
- **Chief Affiliation:** PFLP
- **Education:** Lumuka U., Moscow
- **Last Modified:** Feb. 14, 2012
- **Status:** Active

**Recent Analyst Reports**

- **General Intelligence Directorate - Camp Mantanas**
  
  Published 2 Days, 44 minutes ago

  A selection of articles related to general intelligence directorate - camp mantanas aka punto zero or point zero, training ground for Carlos and other known terrorists...

- **Bassam Abu Shrief – Testimony to Saudi Intelligence**
  
  Published 2 Days, 8 minutes ago

  Shrief claimed that Carlos received $20-50M for the safe release of the Arab hostages and had kept it for his personal use...

- **Carlos allegedly housed by Houari Boumedienne**
  
  Published 2 Days, 41 minutes ago

  Ex-Algerian President Boumedienne reportedly housed Carlos at the luxurious vacation villa overlooking Algeria paid for with oil money...

- **Negotiations ongoing with Sundanea over release of Carlos**
  
  Published 3 Days, 21 minutes ago

  French government officials are forced to deal with tyrannical Sundanea regime, offering trades of satellite equipment and other...

**Associated Global Attacks**

**Related Open Source**

- **Carlos the Jackal - Wikipedia, the free encyclopedia**
  
  en.wikipedia.org/wiki/Camilo_Castillo

- **Carlos the Jackal**
  
  Ilich Ramírez Sánchez (born October 12, 1949), also known as Carlos the Jackal, is a Venezuelan-carrying a life sentence...

- **Carlos – a Profile of Carlos the Jackal**
  
  terrorism.about.com/od/groupiskidsfs/pr/CarlosTheJackal.htm

  The terrorist known as Carlos the Jackal offered the image of a terrorist as a glamorous, ruthless criminal able to pull off amazing exploits...

- **BBC News | World | Carlos the Jackal - three decades of crime**
  
  news.bbc.co.uk/2/hi/42346.stm

  Carlos the Jackal is linked to a series of bombings, kidnappings and hijackings in the 1970s and early 1980s.

- **Vargas the Jackal is sentenced to life again - CNN**
  
  articles.cnn.com/2011/09/02/world/europe_france-carlos-jackal-1-ilich-ramir...

  Dec-16, 2011 – “Carlos the Jackal,” once among the world’s most wanted fugitives, has been sentenced to life in prison for his role in a series of fatal bombings...

**Subject Matter Experts**

<table>
<thead>
<tr>
<th>RATE</th>
<th>NAME</th>
<th>EXT</th>
<th>MAIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Callaway</td>
<td>33432</td>
<td>E-Link</td>
</tr>
<tr>
<td>8</td>
<td>Trimmer</td>
<td>33435</td>
<td>E-Link</td>
</tr>
<tr>
<td>6</td>
<td>Stafford</td>
<td>32334</td>
<td>E-Link</td>
</tr>
<tr>
<td>5</td>
<td>Robinson</td>
<td>21458</td>
<td>E-Link</td>
</tr>
<tr>
<td>4</td>
<td>St. John</td>
<td>21345</td>
<td>E-Link</td>
</tr>
<tr>
<td>4</td>
<td>Peters</td>
<td>29677</td>
<td>E-Link</td>
</tr>
</tbody>
</table>
Data Discovery Applications - Patterns

Research Portal
- Analyst’s workbench
- Management Adviser
- Innovation Center

Unified View
- Voice of the Customer
- Logistics Center
- Consolidated Dashboard

Customer Service
- Call Center
- Online Service
- Sales Dashboard

Compliance
- Fraud Center
- E-Discovery
- Info Governance
Handling LoB Content
A “Recipe” for harnessing LoB data

- **Start with a Target Application**
  Incorporate your business drivers

- **Connect to Authoritative Sources**
  Develop a list, prioritize, and iterate

- **Create Structure from Human Language**
  Using text analytics techniques

- **Deploy a polished, flexible UX**
  Focus on users and use cases, don’t over-constrain it
Selecting Sources

Content Sources for Onboarding

- R&D projects - reports (a)
- R&D projects - research notebooks (b)
- Historical projects (OCR) (c)
- Prototype data (d)
- Lab notes (e)
- Patent prep library (f)
- CAD drawings (g)
- Testing/Stress Data (h)
- Design Patterns (i)
- Technical Data Sheets (j)
- Expert Profiles (k)
- Regulation Database (l)
- Subscription (OneSource, Lexis) (m)
- Industry database (n)
- Competitor Web Crawling (o)
- Industry patents (p)
- Newswires (q)
Text Analytics Techniques

- Tokenization/POS Tagging: Language identification, parts of speech, phrases, alternative word forms
- Entity Extraction: People, places, things, metadata markup for search accuracy
- Facets/Categories: Identify/label concepts, enable browse & navigation strategies, concept search
- Time/Date Extraction: Time references, associations with entities, events
- Event/Relationship Extractions: Subject/Action/Object relationships, patterns
- Sentiment Extraction: Evaluate tone, positive/negative/neutral sentiment
- More Analytic Options: e.g., Geo-tags, social distance, many other options

Documents

BA INSIGHT
Entity Extraction

- Well Established
- Often essential to faceted navigation
- Driven by lexical resources (Taxonomies)

Politics Of Ukraine
[In July 1994, [Leonid Kuchma] was elected as [Ukraine]'s second president in free and fair elections.] [Kuchma was reelected in [November 1999] to another five-year term, with 56 % of the vote.] [International observers criticized aspects of the election, especially slanted media coverage; however, the outcome of the vote was not called into question.] [In [March 2002], [Ukraine] held its most recent parliamentary elections, which were characterized by the Organization of Security and Cooperation in Europe (OSCE) as flawed but an improvement over the 1998 elections.] [The pro-presidential "For a United [Ukraine]" bloc won the largest number of seats, followed by the reformist "Our [Ukraine]" bloc of former Prime Minister Viktor Yushchenko, and the Communist Party.] [There are 450 seats in parliament, with half chosen from party lists by proportional vote and half from individual constituencies.]
The Red Valley property lies within the Qilian fold belt which is host to gold deposits.

Extracted Fact: Substances x Locations
User Experience is a Discipline
This session

LoB systems and Unstructured content

Data Science examples

Data Discovery examples

Handling LoB Content
Bridging two worlds

Traditional IM
- Requirements based
- Top-down design
- Integration and re-use
- Technology Consolidation
- World of EDW, CRM, ERP, ECM
- Competence Centers
- Commercial Software

“Big Data” Style
- Opportunity Oriented
- Bottom-up Experimentation
- Immediate use and gratification
- Tool proliferation
- “World of Hadoop”
- Hackathons
- Open Source
Questions

Jeff.Fried@BAinsight.com
www.BAinsight.com