Session 11413
Issues in Big Data: Analytics

Tom Deutsch, tdeutsch@us.ibm.com
Program Director, Big Data

Bob Foyle, bfoyle@us.ibm.com
Sr. Product Manager IBM Content Analytics with Enterprise Search
Abstract

- OK, so you’ve got all the data lying around and you’ve decided to make it work for you. Now, what you want to know is how to make sense of it. How do you do that? What do you get? Then, what do you do with it once you’ve gotten it? How do you know if the analysis is to be trusted?
- The speakers will address those issues, present IBM best practices, and discuss some real-world use cases.
Agenda

1. Big Data – Definitions & A Quick Story
2. Technology Primer
3. Dive on Content Analytics
4. Thoughts On Getting Started
Conventional Definition of “Big Data”

- Never before possible
  - Solely defined by large volumes
  - Only Unstructured Data
- Valuable insight, but difficult to extract
- Basically an ETL environment

This definition is wrong
Functional Big Data Definition

Dealing with information management challenges that don’t natively fit with traditional approaches to handling the problem
The technologies that deal with these problems are **broad and diverse, it is not just Hadoop**
Giving Rise to “Fit For Purpose Architectures”

The introduction of purpose-built technology focused on a specific computing problem that is compelling better than existing technologies.
Giving Rise to “Fit For Purpose Architectures”

Match the compute problem to the best way to handle it *rather than assuming SQL by default*
...Which In-turn Enables Paradigm Shifts*

- The Idea Of the Super-set
- Move from Sampling to “Absolute Knowledge"
- Combining Structured and Unstructured Analytics
- Importance of Augmented Decision Making
- Streaming Computing Paradigm
- Lowering The Cost of Experimentation
- Changing Role Of Archiving

(* but the Laws of Gravity Still Apply!)
Agenda

1. Big Data – Definitions & A Quick Story
2. Technology Primer
3. Dive on Content Analytics
4. Thoughts On Getting Started
The characteristics of big data

- Cost efficiently processing the growing **Volume**
  - Graph showing a 50x increase from 2010 to 2020
  - 35 ZB

- Responding to the increasing **Velocity**
  - 30 Billion RFID sensors and counting

- Collectively Analyzing the broadening **Variety**
  - Pie chart showing 80% of the world's data is unstructured

- Establishing the **Veracity** of big data sources
- **1 in 3** business leaders don’t trust the information they use to make decisions

© 2012 IBM Corporation
Hadoop is Well Suited for Handling Certain Types of Big Data Challenges

- Analyzing larger volumes may provide better results
- Deriving new insights from combinations of data types
- Larger data volumes are cost prohibitive with existing technology
- Exploring data – a sandbox for ad-hoc analytics
IBM Big Data Strategy: Move the Analytics Closer to the Data

New analytic applications drive the requirements for a big data platform

- Integrate and manage the full variety, velocity and volume of data
- Apply advanced analytics to information in its native form
- Visualize all available data for ad-hoc analysis
- Development environment for building new analytic applications
- Workload optimization and scheduling
- Security and Governance
InfoSphere BigInsights Brings Hadoop to the Enterprise

• Manages a wide variety and huge volume of data
• Augments open source Hadoop with enterprise capabilities
  – Performance Optimization
  – Development tooling
  – Enterprise integration
  – Analytic Accelerators
  – Application and industry accelerators
  – Visualization
  – Security
• Provides Enterprise Grade Hadoop analytics
IBM Significantly Enhances Hadoop

• **Scalable**
  – New nodes can be added on the fly.

• **Affordable**
  – Massively parallel computing on commodity servers

• **Flexible**
  – Hadoop is schema-less, and can absorb any type of data.

• **Fault Tolerant**
  – Through MapReduce software framework

IBM Innovation

• **Performance & reliability**
  – Adaptive MapReduce, Compression, Indexing, Flexible Scheduler

• **Analytic Accelerators**

• **Productivity Accelerators**
  – Web-based UIs
  – Tools to leverage existing skills
  – End-user visualization

• **Enterprise Integration**
  – To extend & enrich your information supply chain.
InfoSphere BigInsights – A Closer Look

More Than Hadoop

- Performance & workload optimizations
- Unique text analytic engines
- Spreadsheet-style visualization for data discovery & exploration
- Built-in IDE & admin consoles
- Enterprise-class security
- High-speed connectors to integration with other systems
- Analytical accelerators
The Only Platform to Support Multiple Hadoop Distributions

- Provides a rich set of big data analytics and accelerators on top of open source
- Delivers a comprehensive big data platform on top of open source, that addresses all big data requirements.
Agenda

1. Big Data – Definitions & A Quick Story
2. Technology Primer
3. Dive on Content Analytics
4. Thoughts On Getting Started
Text Analytics is the basis for Content Analytics

**What is Text Analytics?**

*Text Analytics (NLP*) describes a set of linguistic, statistical, and machine learning techniques that allow text to be analyzed and key information extraction for business integration.*

**What is Content Analytics?**

*Content Analytics (Text Analytics + Mining) refers to the text analytics process plus the ability to visually identify and explore trends, patterns, and statistically relevant facts found in various types of content spread across internal and external content sources.*

* Natural Language Processing
Search and Analytics quick look

- 8 views for analysis, exploration and investigation
- Learn about different ways to discover Rapid Insights from content
- Easy to use to search and analyze
New in ICAwES v3.0 - Seamless Scale-out with BigInsights / Hadoop

- Select “Configure BigInsights Server”
- Enter the BigInsights Server Information
- Specify “Use IBM BigInsights” as a Collection setting

Admin user can confirm the setting on Topology View

...configuration files and ICA libraries, UIMA PEARs (including custom PEAR) and other required modules will be distributed to BigInsights servers automatically for that collection
New in ICAwES v3.0 - Seamless Scale-out with BigInsights / Hadoop
Documents View

**View the Documents based on current analysis criteria**

- Flagging of documents
- Auto highlighting of key entities based on current query
- Export documents to be used in another system.

**Document Analysis Dialog**

- Preview the whole document
- Display document metadata and annotated facets with highlighted texts
Facets View

Show the frequency and correlation indices of keywords belong to the selected facet

- Add selected keyword to current search condition

- Quick filter
- Multi-column sort
Deviation View and Trends View

- **Deviations View**
  - show how keyword deviates from average occurrence between other keywords.
  - Identify patterns that are cyclic.

- **Trends View**
  - show how keyword in each time period deviates from average occurrence (against the whole current matched documents).
  - Predict pattern behaviors in future.
Facet Pairs View

- Show the correlation between keywords belong to two different facets
- 3 view modes

Table view

Grid view

Bird’s eye view

Look at whole area

Quick filter and sort

See in detail
Connections View links highly correlated terms

• Show relationship between multiple facet values
• Connections between nodes represents correlation between two facet values
• Color of line represents the importance of correlation index (red is the highest)

Identify relations between “FORD”, “blow” and “FIRESTONE”
New in ICAwES v3.0 – Sentiment Analytics

Sentiment Expressions
• Lists positive or negative expressions for selected facet value with colors
• Color represents the rank of correlation
Create Dashboard Views for Executive Summaries

Click here
ICA integrates with Cognos BI reports

From ICA Text Miner, a user can:

- Issue a request to create a report
- List the created reports
- Open the created report
- Delete the created report
- Cognos reports can link to and from Text Miner
IBM Content Analytics adds value to Big Data...

**Healthcare Analytics**
- **Analyzing**: E-Medical records, hospital reports
- **For**: Clinical analysis; treatment protocol optimization
- **Benefits**: Better management of chronic diseases; optimized drug formularies; improved patient outcomes

**Crime Analytics**
- **Analyzing**: Case files, police records, 911 calls...
- **For**: Rapid crime solving & crime trend analysis
- **Benefits**: Safer communities & optimized force deployment

**Customer Care**
- **Analyzing**: Call center logs, emails, online media
- **For**: Buyer Behavior, Churn prediction
- **Benefits**: Improve Customer satisfaction and retention, marketing campaigns, find new revenue opportunities

**Insurance Fraud**
- **Analyzing**: Insurance claims
- **For**: Detecting Fraudulent activity & patterns
- **Benefits**: Reduced losses, faster detection, more efficient claims processes

**Automotive Quality Insight**
- **Analyzing**: Tech notes, call logs, online media
- **For**: Warranty Analysis, Quality Assurance
- **Benefits**: Reduce warranty costs, improve customer satisfaction, marketing campaigns

**Social Media for Marketing**
- **Analyzing**: Call center notes, SharePoint, multiple content repositories
- **For**: churn prediction, product/brand quality
- **Benefits**: Improve consumer satisfaction, marketing campaigns, find new revenue opportunities or product/brand quality issues
Interaction between Big Data and Content Analytics

- IBM Content Analytics running on top of BigInsights for full analytics and rapid insights
- BigSheets and IBM Content Analytics – Filter with BigSheets then analyze in ICA for graphical view of analytic results
- IBM Content Analytics as a NLP engine to add define concepts and add depth to Cognos Consumer Insights
Agenda

1. Big Data – Definitions & A Quick Story
2. Technology Primer
3. Dive on Content Analytics
4. Thoughts On Getting Started
Thoughts On Getting Started
Use Case Selection Criteria

- Well proven path
- Can be done off-line/non-disruptive to existing systems
- Intuitive that there is low hanging fruit for additional insights
- Data set is already stored, but under-instrumented or overly summarized
- Initial findings can be arrived at 3 weeks or less
- Initial use cases have to be accretive to next set
- Initial use cases have to leverage common technology for next set of use cases
One Example Of How To Think About It

Start here

Customer Email
Agent Notes
Web Log

Grow to

Predictive Cross Channel Modeling
Cross Channel Communications
Active Listening To Social Media @ Individual Level
Passive Listening To The Social Media Crowd

Value

Low
High

Complexity

Low
High

One Example Of How To Think About It
Use Case Starting Points

- **Mobile/Web Log**
  - Why: Reduce latency to under a day instead of weeks from 3rd parties
  - Why: Have the full data set for cross-correlation for other use cases

- **Agent Notes**
  - Why: Has proven to be an importance source of wisdom from the branches and Contact Centres
  - Why: An element of crowd sourcing
  - Why: Pulse of our internal teams whom we need to market to

- **Customer Email**
  - Why: They are often your best direct source of “pulse” of the business
  - Why: Lynchpin of crowd sourcing
  - Why: Six Sigma feedback component to cost take out

- **Passive Listening To The Social Media Crowd**
  - Why: Important early warning source of information
  - Why: You need the data – not summarization – for more advanced analytics

- **Cross Channel Communications**
- **Active Listening To Social Media @ Individual Level**
- **Predictive Cross Channel Modeling**
Examples Of Insight / Value Delivered

- **Web Log** *
  - Same day feedback on A|B testing new design (if sample is big enough)
  - Same day feedback on change in font/images/icons wording on behavior (if sample is big enough)
  - Input source into cross-channel behavior modeling

- **Agent Notes** *
  - Additional insight into customer “state” based on call handling
  - Feedback on features / policies not captured in formal system – nuance of customer opinion
  - Detecting of otherwise un-surfaced issues both at segment and individual customer level
  - Detection of staff dissatisfaction
  - Detection of otherwise hidden life events

- **Customer Email** *
  - Direct feedback on policy/fees
  - Notice of customers planning to attrition
  - Detecting of otherwise un-surfaced issues both at segment and individual customer level
  - (in private / brokerage situations early warning of law suits / advisor attrition)
  - Detection of otherwise hidden life events
  - Basis for cross-interaction behavioral / acceptance modeling

- **Passive Listening To The Social Media Crowd** *
  - Hour-by-hour feedback on announcements, offerings, policy changes
  - Taking the base reporting the rest of the org expects off the table so more important things can be focused on
  - Crowd feedback on competitor offerings
  - Path to SMARC

* All are expanded “feature vectors”
Case Study: Improvement Of Action Cluster Segmentation Approach

- Use of under-utilized data to expand modeled variables – “Feature Vectors” – The customers response to the firm’s value proposition
- Each feature vector is like a gene strand, which describes a facet, or set of customer behavior traits
Wait – How Do You Know If This Is Working?

- Quick – how many of you were part of an experiment this week?
- Back testing
- A|B real world testing
- “Test Fast, Fail Fast, Correct Fast”
- Concept of MVP (minimum viable product)
THINK BIG