Unstructured Data: Taming the Textual Tsunami with HANA & Hadoop

Anthony Waite, Text Analysis Product Manager
Data & Analytic Engines, SAP Labs
Legal Disclaimer

The information in this presentation is confidential and proprietary to SAP and may not be disclosed without the permission of SAP. This presentation is not subject to your license agreement or any other service or subscription agreement with SAP. SAP has no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation and SAP’s strategy and possible future developments, products and/or platforms directions and functionality are all subject to change and may be changed by SAP at any time for any reason without notice. The information on this document is not a commitment, promise or legal obligation to deliver any material, code or functionality. This document is provided without a warranty of any kind, either express or implied, including but not limited to, the implied warranties of merchantability, fitness for a particular purpose, or non-infringement. This document is for informational purposes and may not be incorporated into a contract. SAP assumes no responsibility for errors or omissions in this document, and shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of this document. This limitation shall not apply in cases of intent or gross negligence.

All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.
Learning Points

- Differences between structured and unstructured data
- How to give structure to unstructured text
- How SAP HANA in real-time unifies enterprise information access
- How Hadoop can economically process textual batch jobs
- When to exploit the strengths of SAP HANA & Hadoop
Structured vs. Unstructured Data
Text Data Processing
SAP HANA
  • Search
  • Text analysis
Hadoop
• Structured vs. Unstructured Data
• Text Data Processing
• SAP HANA
  • Search
  • Text analysis
• Hadoop
80% of enterprise-relevant information originates in “unstructured” data:

- Email, contact-center notes
- Blogs, forum postings, social media
- Surveys, warranty claims
Data Categories

Structured
- Supports automated processing
  - Conforms with data models associated with databases and spreadsheets
  - Granular data stored in fields

Unstructured
- Generally does not support automated processing
  - No data model or not easily understood
  - Insufficient metadata
  - Noisy data communications such as an email message, blog or document
Common Structured Data

Employee Master Table

<table>
<thead>
<tr>
<th>EmployeeID</th>
<th>FirstName</th>
<th>AddressID</th>
<th>ShiftID</th>
<th>LastName</th>
<th>MiddleName</th>
<th>SEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Eric</td>
<td>1</td>
<td>1</td>
<td>Mariner</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Derrick</td>
<td>2</td>
<td>1</td>
<td>Wheezy</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Michael</td>
<td>3</td>
<td>1</td>
<td>Spence</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Maya</td>
<td>4</td>
<td>1</td>
<td>Gutierrez</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Roberta</td>
<td>5</td>
<td>1</td>
<td>Durbin</td>
<td>B</td>
<td></td>
</tr>
</tbody>
</table>

Department Table

<table>
<thead>
<tr>
<th>DepartmentID</th>
<th>Description</th>
<th>revguid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Engineering</td>
<td>3F9D2603-CB0E-4302-A48F-C045C5C1064026</td>
</tr>
<tr>
<td>2</td>
<td>Tool Design</td>
<td>AE940718-0489-4D66-BCCC-209F4AMC05211E</td>
</tr>
<tr>
<td>3</td>
<td>Sales</td>
<td>762COEE3-0554-4F99-5AB7-999F4F25FD1F3</td>
</tr>
<tr>
<td>4</td>
<td>Marketing</td>
<td>263C4A2B-8DCC-4C3C-8D1A-1E7A140A7164</td>
</tr>
<tr>
<td>5</td>
<td>Purchasing</td>
<td>D9C05611-9396-5F4A-AD88-34D58B965CA45</td>
</tr>
</tbody>
</table>

Spreadsheet

[Spreadsheet Image]
## Data Categories

<table>
<thead>
<tr>
<th>Structured</th>
<th>Supports automated processing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>– Conforms with data models associated with databases and spreadsheets</td>
</tr>
<tr>
<td></td>
<td>– Granular data stored in fields</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unstructured</th>
<th>Generally does not support automated processing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>– No data model or not easily understood</td>
</tr>
<tr>
<td></td>
<td>– Insufficient metadata</td>
</tr>
<tr>
<td></td>
<td>– Noisy data communications such as an email message, blog or document</td>
</tr>
</tbody>
</table>
Common Unstructured Data

A press release communication

SAP and Sybase Combine Forces to Unwire the Enterprise

Company Leaders Unveil Joint Strategy and Product Innovations for Customers in Enterprise Mobility, Business Analytics and Enterprise Information Management

MUNICH, BOSTON, AND FRANKFURT, GERMANY - AUGUST 15, 2010 - At a co-located event today in Boston and Frankfurt, company leaders from SAP AG (NYSE: SAP) and Sybase came together for the first time to share the strategy and joint product direction for the newly combined companies. A 3-way move that brings together established industry leaders and changes the game for enterprise mobility, SAP announced today to acquire Sybase on May 12, 2010. In the few months since that time, the two companies have moved quickly to create product road maps and innovations to transform the way inside and businesses work. The combined companies now represent the market-leading position in enterprise applications, business analytics and mobile infrastructure.

Building on the strengths of both companies’ existing business models, SAP confirmed that Sybase will be on as a separate, independent unit, and that customer investments in Sybase products will be preserved and supported. Setting the stage for the vision of an enabled enterprise and enabling the evolution of mobile as the new desktop, SAP and Sybase today also unveiled strategic product directions and planned co-innovations in three key areas: enterprise mobility, business analytics and enterprise information management (EIM).

Combined Companies Retain Core Strengths While Leveraging Synergies for Innovation

Outlining the first milestones since the close of the acquisition, SAP and Sybase today announced:

- Within the next nine months, the companies will bring together technologies to deliver a leading mobile platform for business that is based on open standards, runs on all major mobile operating systems, and manages and supports all major device types. With the platforms, customers and partners can build new mobile experiences on top of existing applications, such as SAP Business Suite software. In addition, SAP will introduce mobile experiences for all products, including both SAP Business Suite and SAP BusinessObjects, similar to what is available for mobile sales for customer relationship management (CRM) today.
- SAP and Sybase will implement the industry's broadest solution portfolio for EIM by pairing, certifying and offering SAP Business Suite and other solutions onto Sybase data management server — bringing customers a proven choice of database platforms for SAP applications.
- The companies will offer customers a complete and optimized high-performance business analytics infrastructure, boosting SAP BusinessObjects Business Intelligence (BI) solutions on data management servers to deliver end-to-end functionality from discovery (enterprise information management) to storage (calculation and storage engines for transactions, data marts, data warehouses) to consumption (query reporting and analysis, scorecarding, dashboarding, OLAP tools, mobile BI).
- The companies will incorporate SAP's in-memory computing technology across SAP and Sybase data management offerings, enabling customers to instantaneously access any type of data anywhere, any place in real-time.

SAP Co-CEO Bill McDermott and Jim Kemperman share their thoughts on today's announcement: "The combination of SAP and Sybase is a significant game changer in the industry. Less than three months after the bid announcement was made, we are highlighting the first joint products that will transform the way people and businesses work. The product roadmap and go-to-market strategy we are sharing today positions SAP to be the one company enabled to deliver a full suite of enterprise software and next-generation business intelligence to any device at any time. The coming together of SAP and Sybase is 100 percent a customer-focused move, delivering a value proposition for customers in mobility, analytics and information management that is unmatched and unique."

"At Sybase, we have been working on our "unleashed Enterprise" strategy over the past eight years to become the market leader in enterprise mobility software," said John Chen, CEO, Sybase. "With our unique portfolio, SAP will take mobility to the next level by authorizing analytics and running new technologies to now serves markets"
Common Unstructured Data

Forum postings

Hello All,
I have a situation in my department where I need to extract the ECC data to third party database (MS SQL). Initially we were thinking of using SAP BW extractors for extraction through SAP PI to push the data out but this doesn’t seem to be feasible. Is this a fair assessment of the situation?

What other methods can we use for data extraction? We are evaluating options for ABAP programs, SAP Queries and BAPIs. Is there any other way we can extract this data set. The data set is required for $d$,$f$,$t$,$r$,$p$,$m$,$p$.

We don’t need data for real time time basis, but on the same lines of a data warehouse extraction like SAP BW.

Appreciate your help.

Thanks
What vs. Why

It’s generally said that structured data tells us “what” and unstructured data tells us “why.”
“If you are not analyzing text – if you’re analyzing only transactional information – you’re missing opportunity or incurring risk.”

-- Seth Grimes, Alta Plana
"Organizations embracing text analytics all report having an epiphany moment when they suddenly knew more than before."

-- Phillip Russom, The Data Warehousing Institute
“The bulk of information value is perceived as coming from data in relational tables. The reason is that data that is structured is easy to mine and analyze.”

-- Prabhakar Raghavan, Yahoo Research
Business Intelligence Typically Runs Off Structured Data
How can you extend your BI investments to unstructured text data?
Structured vs. Unstructured Data

**Text Data Processing**
- SAP HANA
  - Search
  - Text analysis
- Hadoop
Text Data Processing Defined

1. Extract meaning
2. Transform into structured data for analysis

Once structured it can be...
- Integrated
- Queried
- Analyzed
- Visualized
- Reported against

Unlocks Key Information from Text Sources to Drive Business Insight
The **Entity Extraction** transform extracts:

- Predefined core entities (who, what, when, where, etc.)
- Predefined domains (voice of customer, public sector, enterprise)
Core and Domain Extraction

Text Data Processing gives ‘structure’ to 2 sorts of elements from text:

- **Core Entities:**

  
  Davey Jones was one of the Monkeys.

  \(<\text{PERSON}>Davey Jones</\text{PERSON}> was one of the Monkeys.\)

- **Domain Facts:**

  
  I love your product.

  I \(<\text{STRONGPOSITIVESENTIMENT}>\text{love}</\text{SPS}><\text{TOPIC}> your product </\text{TOPIC}>.\)
How Core Extraction Works

This is not a keyword search. Text Data Processing applies full linguistic and statistical techniques to make sure the entities which get returned are correct.

- Grammatical Parsing
  
  *Can we *bill* you?*
  *
  *Bill was the president.*

- Semantic Disambiguation

  *I talked to *Bill* yesterday.*
  *
  *The duck has a *bill.*
  *
  *The *bill* was signed into law.*
Supported Entities Types for Core Extraction

**Who:** people, job title, and national identification numbers

**What:** companies, organizations, financial indexes, and products

**When:** dates, days, holidays, months, years, times, and time periods

**Where:** addresses, cities, states, countries, facilities, internet addresses, and phone numbers

**How much:** currencies and units of measure

**Generic Concepts:** “text data”, “global piracy”, and so on

**Languages:** Arabic, English, Farsi, French, German, Italian, Japanese, Korean, Russian, Simplified Chinese, Spanish
Supported Entity Types Out-of-the-box

<table>
<thead>
<tr>
<th>NAME_DESIGNATOR</th>
<th>c/o, attn</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>President</td>
</tr>
<tr>
<td>PERSON</td>
<td>Barak Obama</td>
</tr>
<tr>
<td>PEOPLE</td>
<td>Greeks</td>
</tr>
<tr>
<td>LANGUAGE</td>
<td>Greek</td>
</tr>
</tbody>
</table>

| ADDRESS1              | 245 First Street Floor 16 |
| LOCALITY              | Cambridge |
| REGION@MINOR          | Napa County |
| REGION@MAJOR          | Connecticut |
| COUNTRY               | Brazil |
| CONTINENT             | South America |
| GEO_FEATURE           | Mount Fuji |
| GEO_AREA              | Scandinavia |

| ORGANIZATION@COMMERCIAL | AT&T |
| ORGANIZATION@EDUCATIONAL | University of Washington |
| ORGANIZATION@OTHER      | FBI |
| PRODUCT                | iPhone |
| TICKER                 | NYSE:SAP |

| SOCIAL_MEDIA@TWITTER_ID | @SAP |
| SOCIAL_MEDIA@TWITTER_TOPIC | #HANA |

| DATE                  | 2/14/2011 |
| DAY                   | Monday |
| MONTH                 | June |
| YEAR                  | 2011 |
| TIME                  | 3:47pm |
| TIME_PERIOD           | 3 days, from 9 to 5pm |
| HOLIDAY               | Memorial Day |

| CURRENCY              | 17 euros |
| MEASURE               | 217 meters |
| PERCENT               | 4% |

| PHONE                 | 617-677-2030 |
| NIN@US_SSN            | 522-89-2255 |
| NIN@FR_INSEE          | xxx |
| NIN@CA_SIN            | xxx |
| URI@EMAIL             | john.smith@sap.com |
| URI@IP                | 165.14.2.0 |
| URI@URL               | http://sap.com |

| Syntactic Entities:  |
| NOUN_GROUP           |
| PROP_MISC            |

| big umbrella |
| Cup o’ Soup |
Domain Fact Extraction

Domain extraction is built on top of Core Extraction and realized through rules which looks for facts -- relations between entities or states involving an entity.

• Relations:

  *Manuel read the book.*

• States:

  *Dina’s birthday is June 12th.*
Supported Domain Extraction Out-of-the-box

Voice of Customer

**Sentiments:** strong positive, weak positive, neutral, weak negative, strong negative and problems

**Requests:** customer requests

Other Specialized Content

**Public Sector:** such as person-organization, person-alias, travel events and security

**Enterprise:** mergers and acquisitions, as well as executive job changes

Language Support: English, French, German and Spanish

Language Support: English; Arabic and Simplified Chinese (security only)

These are starter packs that can be built upon for a specific deployment
Agenda

• Structured vs. Unstructured Data
• Text Data Processing
• **SAP HANA**
  • Search
  • Text analysis
• Hadoop
Real Real-Time Analytics

Traditional

- DB indexes
- Materialized views
- Summarization tables

SAP In-Memory

- Scan speed: 2 M/ms*
- Virtual views
- Aggregation speed: 50 M/s*

* per core; speed varies depending on the cardinalities and data types
Old School Applications

Application

Database

Mass data
New Age – Extreme Applications

Application

Database

Mass data
The Enterprise Landscape Challenge

UI
Search
Database

UI
Data Mining
Database

UI
Text Analysis
Database

UI
Graph
Database

Transactional System/Data Source

UI
Analytical System

SAP HANA
SAP HANA for Holistic Information Access

Integrating complementary functionality on one platform:

- all data – structured and unstructured
- multiple workloads – analysis, search, mining

In-memory computing platform provides:

- data driven applications
- low TCD – one model, different perspectives
- low TCO – avoid redundant data persistency and engines
Search in SAP HANA SPS04

Capabilities

- Native full-text search
- Graphical search modeling in SAP HANA Studio
- Configure HTML5 applications with SAP HANA UI toolkit

Benefits

- Exploit unstructured content in SAP HANA without additional costs
- Less data duplication and movement – leverage one infrastructure for analytical and search workloads
- Easy-to-use modeling tools – use HANA Studio Modeler to create search models
- Build Search Applications quickly – UI building blocks provided
Search Architecture in SAP HANA SPS04

Metadata Text Processor

Linguistic Processing

Search Engine

Fuzzy

Ranking

Snippets

Information Access Services

Search

Modeler

HANA Studio

Search Model

Metadata

Column Store

Tables

Linguistic Processing

Text Processor

SAP HANA

SAP HANA UI

UI components
Search Application Example in SAP HANA SPS04

RESULTS

America's Job Bank: A Corporate Solution
06/26/2007 in career
The Internet career site America's Job Bank, which currently posts more than 2 million jobs online, will cease to exist as of midnight, June 30, 2007. About 18 months ago, the Labor Dept. decided not to fund the site (www.jobsearch.org) past the end of...

GM Foresees Fewer Vauxhall Job Cuts
11/19/2009
General Motors raised union hopes that there could be fewer redundancies at Vauxhall now the U.S. parent company has decided not to sell the historic car marque. Meetings between GM executives, the Government and trade unions yesterday were described...

Energy Innovation Should Trump Job Creation
06/22/2008
Across the country, it is hard to find an elected official who hasn't jumped on the green economy bandwagon. So many stump speeches marry the policy objectives of saving the environment and creating jobs. It starts at the top. President Barack Obama has...

BP's Ex-CEO Gets British Government Job
07/02/2010 in globabiz
Lord Browne, the former head of BP (BP) who was once reputed to be Tony Blair's favourite businessman, has been brought back into the heart of Government to help find the £6.2bn in public spending cuts promised in the Budget. One of his first tasks...

New Job Performance Requirement: Don't Worry, Be Happy!
01/06/2008
Should attitude matter when you are on the job? I suppose its important to be cheerful at a time of crisis around our kids, and even our spouse, but what about our colleagues? According to the Chicago Tribune, yes. The...

America's Favorite Pickup
04/10/2006 in author Slide Show
I’ve got to think that when Bill Ford goes to bed at night he says a little prayer of thanks for the F-150 pickup truck. That's because Ford's (F) chairman and chief executive knows that for the past 30 years the F-150 has...
Support in-database text data processing to extract information from unstructured content.

- **Text Processor**
  - File filtering, automatic language identification
  - Configure and control text analysis with native database tools
  - Store the results of Natural Language Processing (NLP) in HANA tables...
    - Linguistic processing – tokenization, stemming, POS-tagging
    - Semantic processing – named entity extraction, sentiment analysis
  - ... for downstream consumption via search, analytics, or data/text mining

- **Leverage the REST API and client libraries from Search**
  - Interface and building blocks to quickly develop hybrid search/text mining applications
Text Analysis in SAP HANA SPS05 & Beyond

<table>
<thead>
<tr>
<th>App</th>
<th>Information Retrieval</th>
<th>Information / Concept Extraction</th>
<th>Clustering</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search + Explore</td>
<td>Full-text search fuzzy, phrase etc.</td>
<td>Explore drill-down, navigate</td>
<td>similar documents, related terms, document features</td>
<td>Categorization</td>
</tr>
<tr>
<td>Full-text index</td>
<td>Table</td>
<td>Text Mining index</td>
<td>Table / optimized store</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Use case/features</th>
<th>Pre-process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explore + Organize</td>
<td>NLP (tokenization, stemming, POS)</td>
</tr>
</tbody>
</table>

- Full-text search
- Explore drill-down, navigate
- Similar documents, related terms, document features
- Categorization
- Term frequencies, assoc., co-occurrence
- Phrases, entities, facts, sentiments

ASPUG
Agenda

- Structured vs. Unstructured Data
- Text Data Processing
- SAP HANA
  - Search
  - Text analysis
- Hadoop
What is Hadoop?

Addresses challenges:
- **Open-source** project administered by the Apache Software Foundation
- Allows for **scalable storage of massive structured and unstructured data** on commodity hardware clusters
- Designed for **non-real time analysis**

Key Services:
- Reliable data storage using the Hadoop Distributed File System (HDFS)
- Parallel data processing and query execution using MapReduce

Enterprises starting to adopt:
- Originally employed by Yahoo and Facebook
- Today used in finance, technology, telecom, media and entertainment, government, research institutions and other markets with significant big data
SAP Data Services 4.1

Enterprise Applications

Open  Scalable  Reusable

Legacy

Business Intelligence

Unstructured Data

hadoop

asug
How does SAP Data Services 4.1 play with Hadoop?

Data Services **identifies, extracts, structures and transforms meaningful information** from Hadoop and can provision the data to SAP HANA.

This **allows for reliable, efficient and real-time analysis across all enterprise information assets** – structured or unstructured.

- Load only the meaningful/relevant information to SAP HANA
- Make data normalized
- Understand the quality of the result and react during load – avoiding the load of bad data
- Relate to structured data from other sources
- Easy and flexible configuration
- Optimized execution – using sources and targets native functionalities
Text Data Processing pushed down as a Map Reduce Job
Providing structure to unstructured data allows for automated business processing and discovery

Text Data Processing gives structure to unstructured data

SAP HANA unifies enterprise information access for real-time analytics

Data Services pushes down Text Data Processing batch jobs into Hadoop clusters for economical parallel processing
Thank you for participating.

Please provide feedback on this session by completing a short survey via the event mobile application.

SESSION CODE: 0211

Learn more year-round at www.asug.com