Text Search and Text Analysis with SAP HANA

Philip Mugglestone, SAP

Agenda

- Overview and Key Challenges
- Text Search and Text Analysis with SAP HANA
- Customers Benefit from Text Analysis with SAP HANA
- Summary

Agenda

- Overview and Key Challenges
- Text Search and Text Analysis with SAP HANA
- Customers Benefit from Text Analysis with SAP HANA
- Summary

What vs. Why

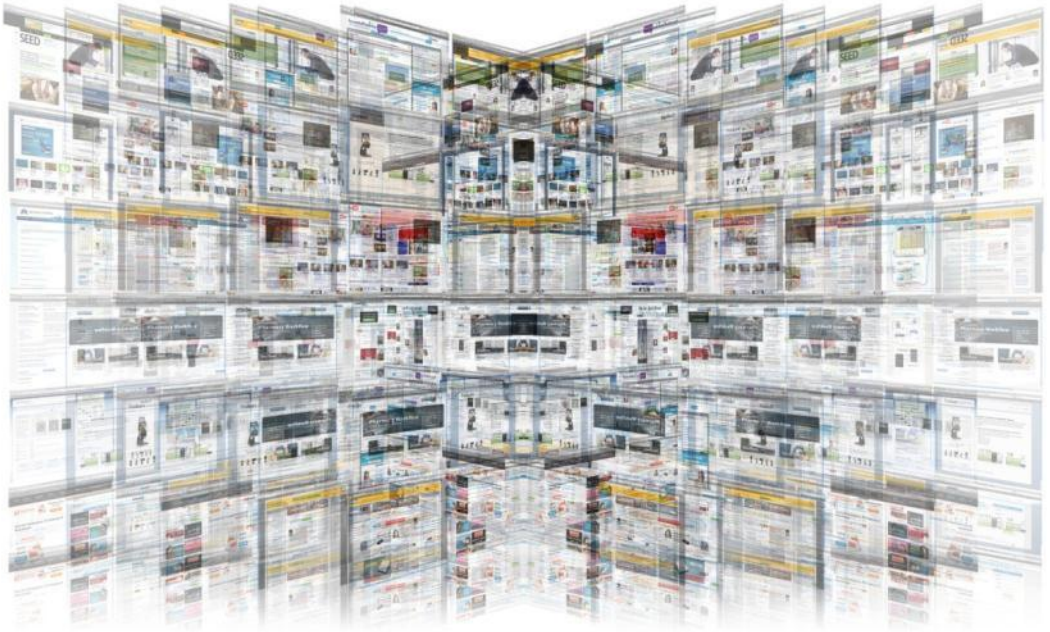
It’s often said that

- Structured data tells us “what”
- Unstructured data tells us “why”

Hidden Value in Text

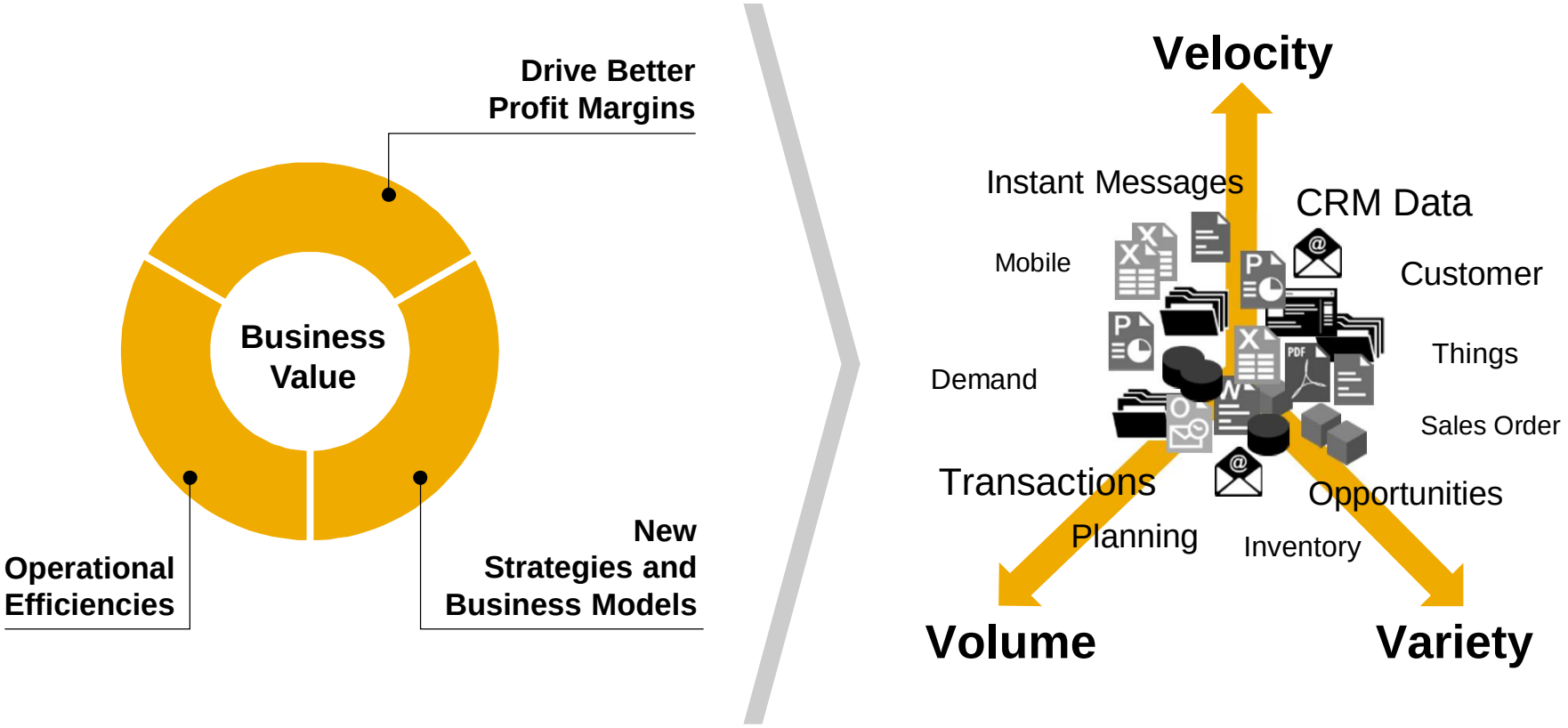
80% of enterprise-relevant information originates in “unstructured” data:

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



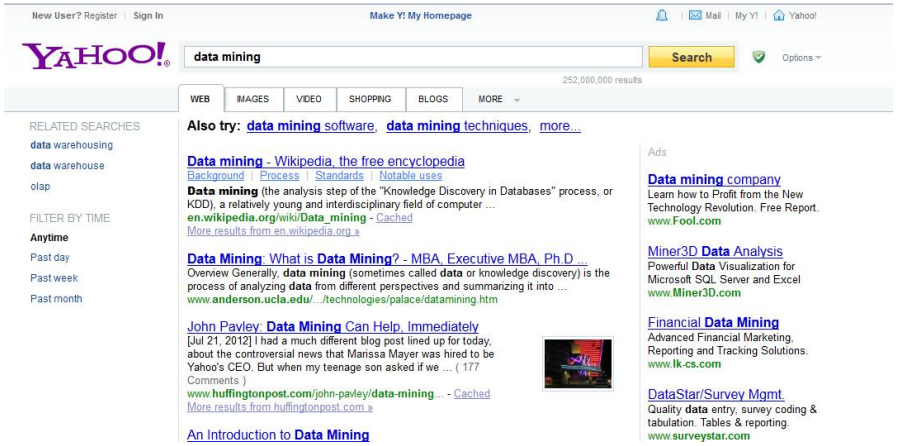
Big data matters

Transformational business value from data

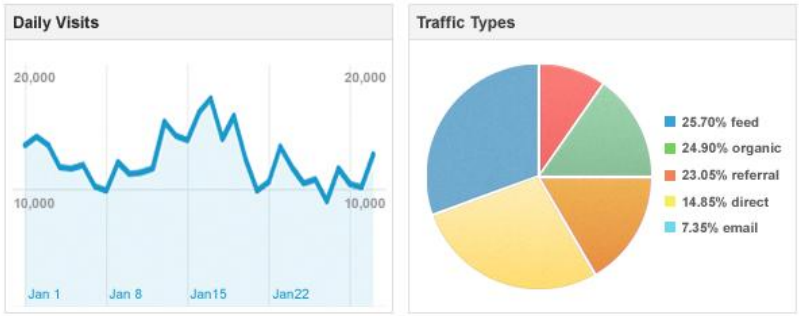


Search, Text Analysis, and Analytics

Search like….



Analytics….



Text Analysis….

<PERSON>*Jim****</PERSON>*** *bought* ***<QUANTITY>****300****</QUANTITY>*** *shares of* ***<ORGANIZATION>****Acme Corp.****</ORGANIZATION>*** *in* ***<DATE>****2006****</DATE>***

Text Analysis Defined

Often referred to as Text Data Processing

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

Structured Database

Once structured it can be…

- Integrated
- Queried
- Analyzed
- Visualized
- Reported against

Unlocks Key Information from Text Sources to Drive Business Insight

Key Challenges

What do your customers really think about your brands, your products, and your services?

- How did customers respond to your last marketing campaign?
- What are the primary reasons for failure recorded in the maintenance records for a particular process, machine, or equipment?
- What are the key issues with the new product release based on call-center records?
- What are the serious side effects from the new drug therapy and how many are affected?
- What are the related engineering documents/sources for this particular topic of interest?

Massive amounts of unstructured data are being captured

- Operational, CRM
- Maintenance, Engineering
- R&D, Call Center
- Social media, blogs, forums,
- e-mails, documents, etc.

Companies are struggling to:

- Identify salient information from unstructured textual data
- Find, interpret, and analyze the content
- Combine unstructured with structured data
- Leverage the data in real-time to gauge and guide their business strategy and solve critical problems

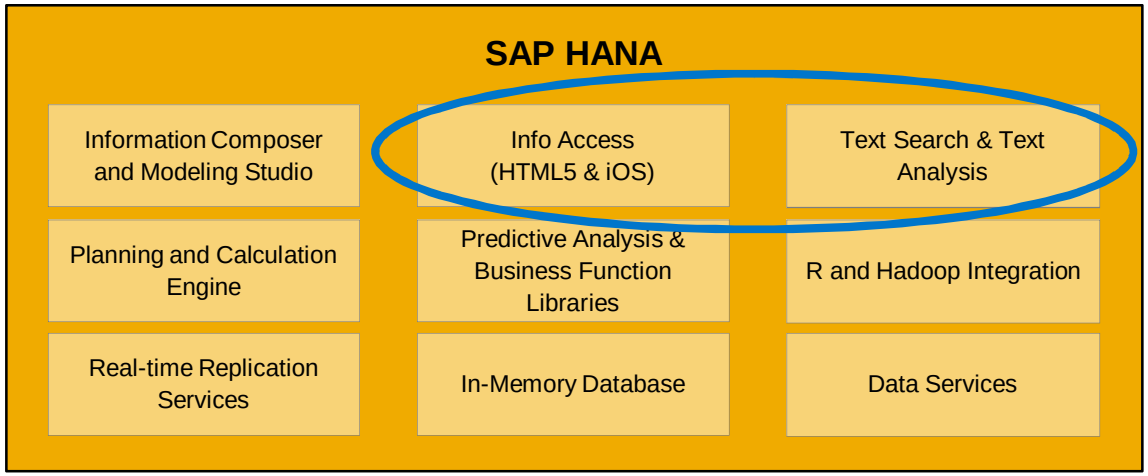
…and do all of this on one platform!

Agenda

- Overview and Key Challenges
- Text Search and Text Analysis with SAP HANA
- Customers Benefit from Text Analysis with SAP HANA
- Summary

SAP HANA

A platform for a new class of real-time analytics and applications



SAP HANA

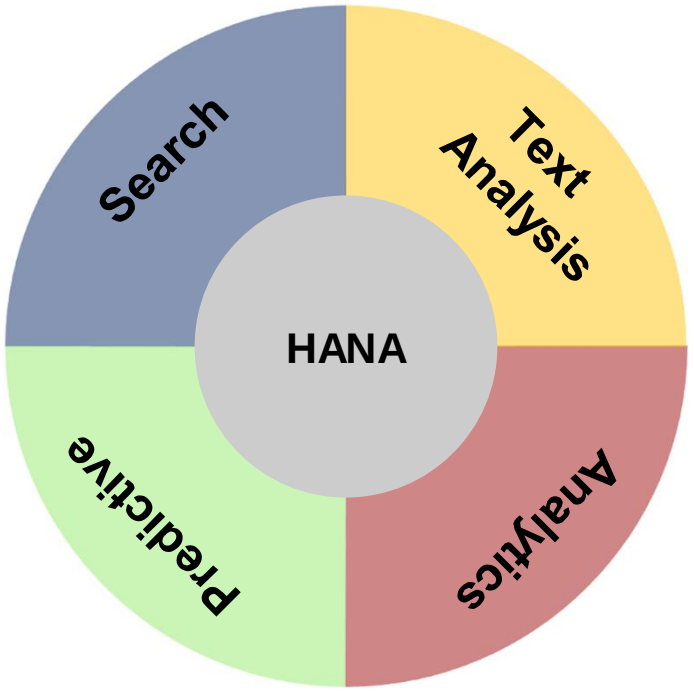
A unified platform for text search, text analysis, and analytics

Integrating complementary functionality on one platform to support a **unified analytics** strategy

- All data - structured and unstructured
- Hybrid models - analysis, search, analytics
- All in one engine

An in-memory computing platform provides

- Platform for data driven applications
- Unified access layer
- Flexible and seamless user interaction
- Low TCD - one model, different perspectives
- Low TCO – reduces redundant data persistency, engines, and data movement



SAP HANA

Structured and unstructured content capabilities

SAP HANA can store, search, and analyze various types of data

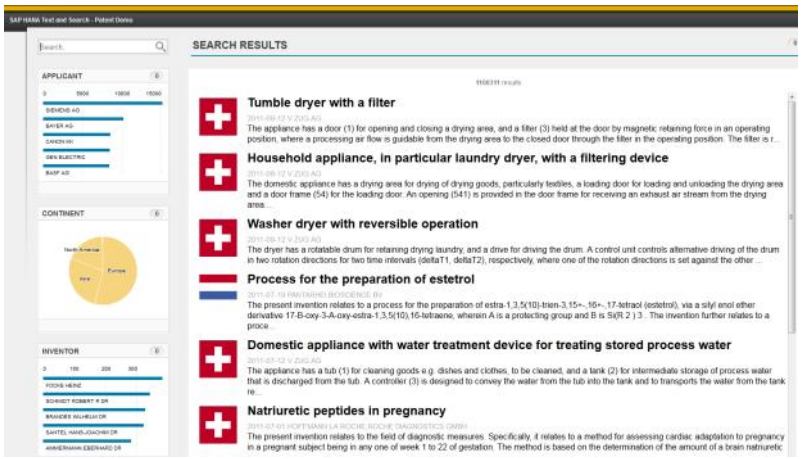
HANA Capability	Structured Data	Unstructured Text Data	Unstructured Multimedia Data
Store	✓	✓	✓
Extract	✓	✓	
Search	✓	✓	
Analyze	✓	✓	

SAP HANA Full-Text Searching

Natively process and search any structured or unstructured text content all in one flexible and robust database platform

Native full-text search

- **Exploit unstructured content** in SAP HANA
- Leverage one infrastructure for **analytical** and **text** search workloads in both OLAP and OLTP use cases
- Reduced **duplication**, **latency**, and operational **overhead**

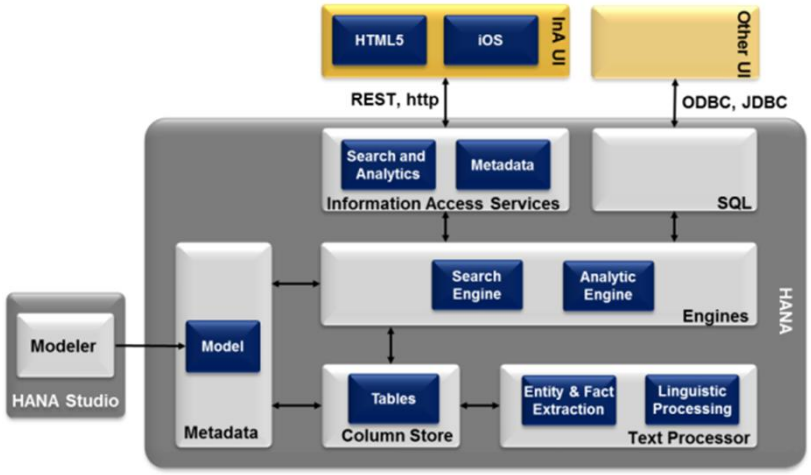


Graphical modeling

- **Easy to use** search definition
- **Built into existing** SAP HANA Modeling tools

‘Info Access’ toolkit

- **Rapid development** of search enabled applications through reusable UI building blocks

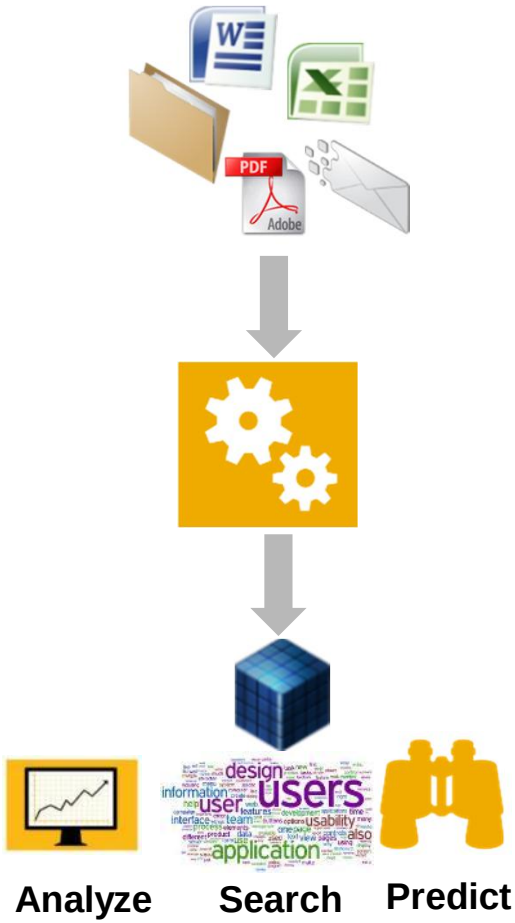


SAP HANA Text Analysis

Extract information from documents; Perform text analysis on unstructured data

- **File Filtering**
  - Unlock text from **binary documents**
  - Ability to **extract and process** unstructured text data from various file formats (txt, html, xml, pdf, doc, ppt, xls, rtf, msg)
  - Load binary, flat, and other documents directly into HANA for **native text search and analysis**

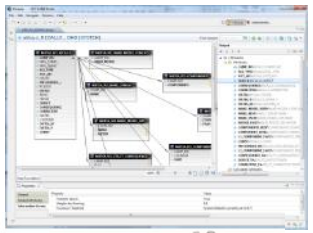
- **Native Text Analysis**
  - Give structure to unstructured textual content
  - **Expose linguistic markup** for text mining uses
  - **Classify entities** (people, companies, things, etc.)
  - **Identify domain facts** (sentiments, topics, requests, etc.)
  - **Supports up to 31 languages** for linguistic mark-up and extraction dictionary and **11 languages** for predefined core extractions



Building a Text Search & Text Analysis Based Application

Create Model

Use SAP HANA Studio to define the search data model and configure the search behavior

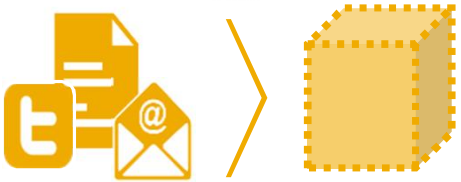


Create Full-text Index

Use SAP HANA Studio to create full-text indexes for search, file filtering, and optionally run Text Analysis

Run Text Analysis

Extract salient information from text (Linguistic Markup, Entity & Sentiment Extraction)



Configure App

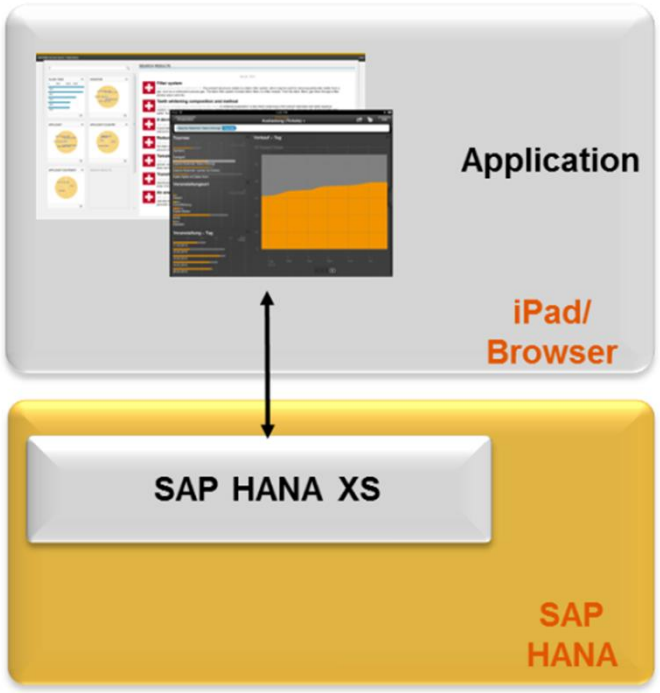
Use SAP HANA Info Access toolkit to define layout and data for the App

Consume Data

Search on Text and/or filter, analyze, and perform advanced analytics on Text Analysis table output

SAP HANA Info Access

Configuration toolkit for quickly developing and deploying
SAP HANA search and visualization based browser and iOS apps



Capabilities:

- SAP HANA Info Access includes: **Services, UI & Client Library Toolkit, and App**
- Allows for **visualization and interaction** of data in SAP HANA (full-text searching, filtering, drill-down, charts)
- Supports **iPad** (download from Apple App Store) and **Browser** (HTML5) deployments
- **Consumes SAP HANA models**
- **NOT a general purpose BI tool!**

Benefits:

- Quick **development** and **deployment** time
- Low TCO and **fast response times** with 2-tier architecture
- **Included with SAP HANA license**

NB: iPad application to be released shortly after initial SAP HANA SPS5

Text Search and Text Analysis Use Cases / Scenarios

MANUFACTURING I RETAIL I HEALTHCARE I BANKING I UTILITIES I TELCO | PUBLIC SECTOR | FINANCIAL SERVICES

OPERATIONS I HR I FINANCE | IT | SALES I MARKETING

Customer Sentiment Analysis

Maintenance Record Analysis and Equipment Monitoring

Document Searching

Engineering Requirements Tracing

Scientific and Medical Research and Pattern Detection

Brand Monitoring and Customer Complaint Analysis

SAP HANA Text Search & Text Analysis

Benefits

For the Business

Exploit Unstructured Data

Ability to extract and analyze information from unstructured content

Faster Time to Analysis

Achieve faster search and analysis results by leveraging a high-performing in-memory platform

Flexibility

Perform text search, text analysis, and analytics all in one unified platform

For IT

Landscape Simplification

Reduces redundant data persistency, engines, and data movement

Total Cost of Development

One unified platform and model for text search, text analysis, and analytics

Unified Access Layer

Quickly develop and connect applications to search and explore data in SAP HANA

Agenda

- Overview and Key Challenges
- Text Search and Text Analysis with SAP HANA
- Customers Benefit from Text Analysis with SAP HANA
- Summary

Mantis Technology Group – Internet Industry

Software solution provider specializing in enterprise custom services for online retailers & high transaction volume provision systems

99% reduced
ETL times

6x faster
Text analysis processing

Significant Simplification of Data Architecture
Moved from 23 servers to 1 Hana One server

Product: Pulse Analytics – Social Media Analytics By SAP HANA One (Cloud)

Business Challenges

- Offer rapid analysis of social media channels to track consumers and influencers and measure brand against industry metrics
- Scale social media analytics service offering to handle ever increasing volumes of data cost-effectively

Technical Challenges

- Reduce the ETL load times to deliver real-time analysis
- Analyze large volumes of social media data – more than 1M documents daily
- Lower cost of managing cluster of 18 Text Analysis XI and 3 MySQL servers

Benefits

- New real-time analytical capabilities allow for visual presentation of data that is free from previous performance-based constraints
- Faster natural-language-based sentiment analysis with topic identification
- Data Architecture simplification by replacing 20+ separate servers with 1 instance of SAP HANA One

"We can get close to an order of magnitude improvement in performance, additional headroom, access to new practical capabilities (as a result of the performance improvements) AND… still save money!"

Doug Turner, CEO of Mantis Technology Group

Technical Implementation

1. Key SAP HANA features

- SAP HANA One: In-memory processing in the cloud
- Native text analysis functionality in SAP HANA for full-text indexing, fuzzy search and sentiment analysis
- Automatic textual data Indexing capabilities
- Join Data on all dimensions as it is created
- Fuzzy Search for advanced clustering of similar mentions
- On-premise capabilities

2. Technical KPIs

- Significantly reduced Extract, Transform and Load times
- 6x increase in text analysis processing
- Simplification of data architecture – single Unified Information Access Platform

3. Implemented by Mantis Technology Group

4. Partners

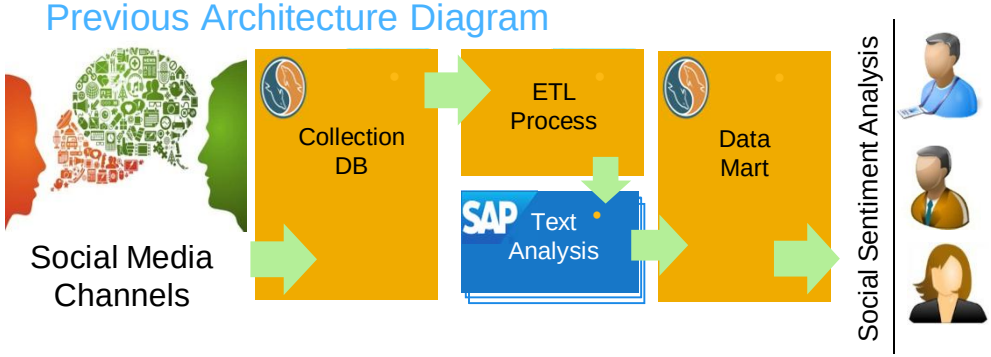
- Data Center provided by Amazon EC2

5. Next Steps

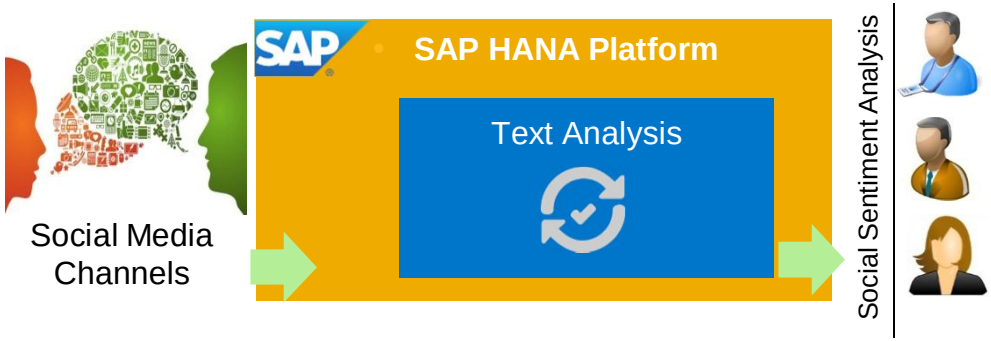
Go-live on SAP HANA One by SAPPHIRE

Architecture Diagram

Previous Architecture Diagram



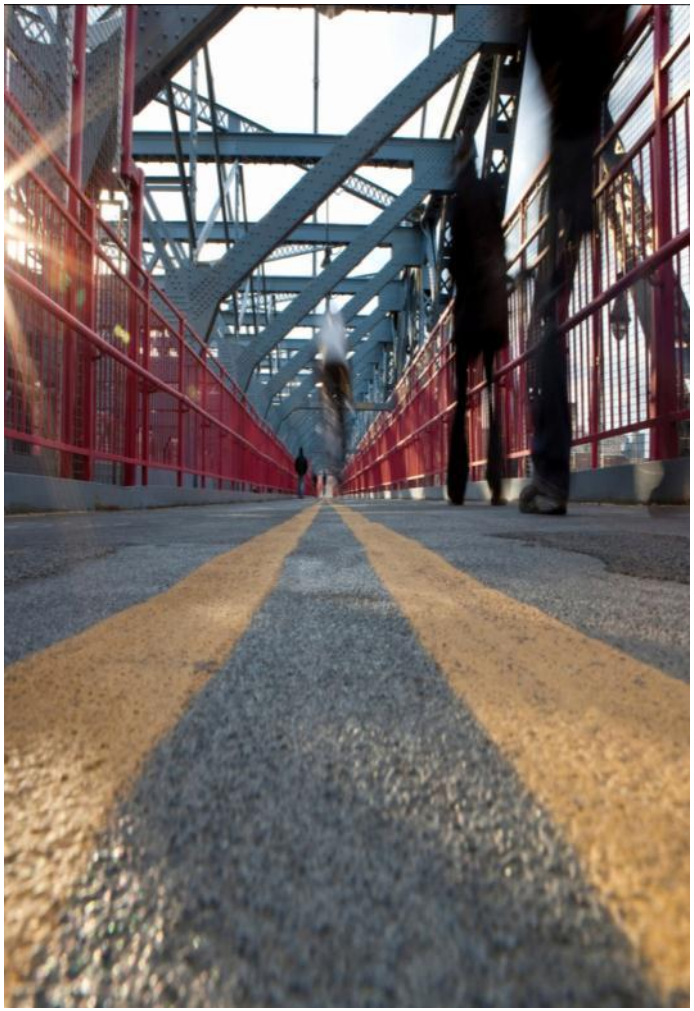
HANA Architecture Diagram



Agenda

- Overview and Key Challenges
- Text Search and Text Analysis with SAP HANA
- Customers Benefit from Text Analysis with SAP HANA
- Summary

The Bottom Line



SAP HANA provides a powerful unified platform for all your analytic and application development needs:

- Analytics
- Predictive
- Text Search
- Text Analysis

- Customers can exploit unstructured and structured data in one platform

- IT Simplification – Reduces TCO and TCD by eliminating data redundancy, movement, and hardware servers/engines

- Provides a unified access layer for fast search application development and deployment

Thank You

Contact information:

Ashish Sahu, Solution Marketing, SAP

ashish.sahu@sap.com

Matthew Zenus, Solution Management, SAP

matthew.zenus@sap.com

Appendix

Text Analysis

Core and Domain Extraction

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

Core Entities:

Davey Jones was one of the Monkeys.

<PERSON>Davey Jones</PERSON> was one of the Monkeys.

Domain Facts:

I love your product.

I <STRONGPOSITIVESENTIMENT>love</SPS><TOPIC> your product </TOPIC>.

Text Analysis

How Core Extraction Works

This is not a keyword search!

Text Analysis 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.

SAP HANA Search, Text Analysis, and Info Access

Key capabilities by release

Capability	SAP HANA SPS4	SAP HANA SPS5	Details
Searching	✓	✓	SPS4 – Full-text Indexing, Modeling, Search (Full-text, Fuzzy, Freestyle), and Term Mapping
Processing	✓	✓	Linguistic Analysis
Text Analysis		✓	Expose Linguistic Markup; Entity and Sentiment Extraction
File Filtering		✓	Extraction of text from binary documents such as pdf, Office files, email, compressed archives, etc.
Info Access Toolkit	✓	✓	Info Access Toolkit (client library, UI components) SPS4 – HTML5 (search) SPS5 – HTML5 + iOS (search, text analysis, and analytics); SPS5 HTML5 capabilities include new UI layout, additional chart type, enhanced filter capabilities, and simplified deployment.

Scenario Considerations for HANA Text Analysis

SAP HANA Text Analysis & SAP Data Services Text Data Processing

If you want to…..	SAP HANA Text Analysis	SAP Data Services Text Data Processing
Load data into SAP HANA using SAP SLT or a 3rd party ETL tool; then analyze text data using text analysis capabilities in SAP HANA	✓	
Leverage native search capabilities in SAP HANA in conjunction with text analytics (e.g. search-based and text mining applications for investigative discovery)	✓	
Have SAP HANA automatically update frequent changes to text analysis processes (without having to re-load the data)	✓	
Access linguistic markup generated when text is processed, which is persisted in SAP HANA (e.g. tokenization, uninflected forms / stemming, part of speech)	✓	
Have high-performing text analysis in SAP HANA	✓	
Not load, store, or process the unstructured text data or documents in SAP HANA (because of cost / space concerns)		✓
Perform text analytics at the source (e.g. push text data processing down into Hadoop) to uncover relevant nuggets of info that can be loaded into SAP HANA		✓
Perform transformations before loading data into SAP HANA (e.g. cleanse, match / de-duplicate and enrich text data)		✓
Create you own custom text analysis rules		✓
Support continuous text analysis workloads that are submitted regularly		✓

© 2012 SAP AG. All rights reserved.