Big Data Expert Bar: Intelligente Analyse unstrukturierter Daten mit HP Idol

Arian van Hülsen
Big Data Solution Architect
Hewlett-Packard (Schweiz) GmbH
In **data**, there is **gold**
So... What is the market’s definition of Big Data?

*Gartner, Inc., “Big Data” Is Only the Beginning of Extreme Information Management*,
Mark A. Beyer, Anne Lapkin, Nicholas Gall, Donald Feinberg, Valentin T. Sribar, Published 7 April 2011

Big Data is no longer just a buzzword...
Big Data landscape

Machine Data

Business Data

Human Information

Vertica

HP Hardware & Services

Autonomy
Why is processing human information different?

Human Information is made up of ideas, is diverse, and has context. Ideas do not exactly match like data does; they have distance.
HP IDOL: the OS for human information

Meaning based computing

- Conceptual distance understanding
- Meaning extraction
- Continuous learning ability
- Language agnostic
- Supports over 1,000 file types and 400 content repositories
- Processing in real time
- Protected by 170 patents
Over 500 IDOL functions to augment your intelligence

**Inquire**
- Search your data
  - Automatic hyperlinking
  - Conceptual search
  - Keyword search
  - Fieldtext search
  - Phrase search
  - Phonetic search
  - Field modulation
  - Fuzzy matching

**Investigate**
- Analyze your data
  - Automatic query guidance
  - Highlighting
  - Parametric refinement
  - Summarization
  - Real-time predictive query
  - Metadata extraction
  - Automatic tagging
  - Faceted navigation

**Interact**
- Personalize your data
  - Implicit profiling
  - Explicit profiling
  - Community and expertise network
  - Agents
  - Intent-based ranking
  - Alerting
  - Social feedback

**Improve**
- Enhance your data
  - Eduction
  - Automatic clustering
  - Clustering 2D/3D
  - Auto classification
  - Auto language detection
  - Sentiment analysis
  - Automatic taxonomy generation
Search your data

• Conceptual, Keyword or Object
• Extensive Field combinations
• Full Meta Search
• Linearly Scalable
• Fault Tolerant
• Disaster Recovery Friendly
• All Information
• Real-Time Data
• Audio and Video
• Mapped Security
• Fully Extendable
• Leverages Existing Security
• Robust Architecture
Summarization

Quick summary
(N+ lines)

Context summary
(What is this doc about with relation to query terms?)

Concept summary
(What do I look for with regards to interest rates?)

Information Theory and Bayesian Inference
Clustering

Automatically partition the data so that similar information is clustered together
Topical sentiment analysis

Decomposition and classification within a sentence to pull out specific topics

“I stayed at the Marriott last week, and though the mattresses were very nice, the service was awful.”

Is this Positive? Negative? Neutral?
How much Positive? How much Negative?
Search video as easily as text
Transform rich media into intelligent assets

Automate

- Live video or playback from archived footage
- On-screen text recognition
- Face identification

Understand

- Automatically generated transcript using speech recognition
- Timecode synchronization
- Speaker identification
- Automatic keyframe generation

Act
Use Cases
Use Case 1: Warranty Analytics @VW
Impact of unexpected Warranty events

Record number of vehicles recalled in U.S. last year - regulator

Thu, Feb 12 2015

WASHINGTON, Feb 12 (Reuters) - A record 69.0 million vehicles were recalled in the United States last year, the U.S. auto safety regulator said on Thursday, led by high-profile recalls of General Motors Co ignition switches and Takata Corp air bags.

The National Highway Traffic Safety Administration said more than twice as many vehicles were recalled in 2014 than the previous record of 30.5 million in 2004.

GM’s recall of 2.6 million cars beginning in February 2014 to replace faulty ignition switches put the company under a harsh spotlight and may have prompted other automakers to be more vigilant about recalls.

Several investigations of GM, including the company’s own, concluded the automaker should have recalled the cars years before. At least 53 people have died in accidents after the ignition switch unexpectedly turned off the engine, disabling the air bags, power steering and power brakes.

Millions of vehicles were recalled last year because air bags made by Japanese supplier Takata can explode with too much force, spraying metal fragments inside the car. The air bags have been linked to five deaths in the United States, all in cars made by Honda Motor Co Ltd.

NHTSA itself has come under fire from members of Congress and safety advocates who say the agency has tried to spot auto makers and not done enough to crack down on automakers.

President Barack Obama, in his 2016 budget, asked Congress to boost NHTSA funding by 9 percent to $630 million, with much of the increase going towards hiring more investigators and developing better data mining tools to uncover defects.

“Congress has a chance to make this year’s budget funding an important step forward,” he wrote. 

Hyundai recalls more than 200,000 Elantras over steering problem

WASHINGTON (Reuters) - Hyundai Motor Co said on Thursday it was recalling 205,000 vehicles because of a power steering problem that could cause the car to lose control in an accident.

Hyundai said it will recall two models, the Elantra and a 2014 high-end Elantra, beginning Jan. 6, 2016.

UPDATE Fiat Chrysler N. America recall amount hits by recall costs

Tue, Feb 3 2015

DETROIT, Feb 3 (Reuters) - Extra costs for recalls last year out a full percentage point from Fiat Chrysler’s North American profit margins, the company said on Tuesday, with an unannounced recall bill of about $330 million.

Richard Palmer, the company’s chief financial officer, said that without the extra recall costs, a margin at FCA in North America would have been 9.2 percent for the year rather than the 4.2 percent the company achieved.

The company’s margin in North America was 7.9 percent in 2014, down from 12.1 percent in 2013. The margin was also down from 11.8 percent in 2012.

FCA US LLC, the company’s North American unit formerly known as Chrysler Group LLC, had net sales of $31.8 billion and net income of $336 million in 2014, up from $325 million in 2013.

Averaged across FCA US LLC’s shipments of 2.85 million vehicles in 2014, the additional recall and warranty costs, mostly for older models, reduced profit by an average of about $188 per vehicle sold.

The North American unit issued 44 separate memos offering about 6.1 million vehicles last year.

The company has said that its weighted per-vehicle costs are not available.

FCA Chief Executive Sergio Marchionne said last week when the parent company reported earnings that if the additional recall charges had been incurred individually, “they will be passed on to the consumer.”

Based on 2014 North American modified operating profit of $3.5 billion, about $570 FCA

Arizona is first U.S. state to sue GM over delayed recalls

Thu, Nov 19 2015

[Pages] - The first lawsuit brought by a state against General Motors Co over delayed recalls relating to a defective ignition switch. Arizona has accused General Motors of putting the public at risk by delaying the recall.

The state filed a complaint in Arizona Superior Court on Tuesday, accusing GM of putting the public at risk by delaying the recall.

The Arizona lawsuit is the first filed against GM by a state. Six states, including Arizona, have substantiated complaints that GM failed to act promptly to recall vehicles, and the Arizona lawsuit is the first filed against GM by a state.
‘Early’ Defect Detection in Warranty

- Structured Data
  - ‘Known’ Defects
    - Reporting Known Defects
      - Future Unknown Defects

- Unstructured Text

Linguistic Analysis

- Reporting Known & Unknown Defects
  - Early Defect Detection

© Copyright 2015 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.
The 'Early'

1- All Available Knowledge (service tickets)

2- Known Knowledge (historical logs)

3- Potentially New and Previously Unknown Knowledge

4- Relevant Knowledge

5- Difficult to Discover

Manual

Automated
Approach – Train a Taxonomy

Every time it rains I get water in the rear seat driver side

Leakage from sun roof. Changed seal.

Terms  | Weights
---|---
leak | 1300
rain | 900
water | 850
seal | 400
roof | 400

Customer defect description

Workshop defect description

30 % of Tickets 2012 – 2014 manually classified by After Sales

Existing knowledge could be used as training set for new taxonomy
HP Early Defect Detection - Technology

**Exploration**
- **HP Explore**
  Interactive exploration and discovery

**Reporting**
- **QlikView**
  Standard reporting and trending

**Language Translation**
- **SDL**
  Translate languages

**Unstructured Data**
- **HP IDOL**
  Manage/Index text data

**Structured Data**
- **HP Vertica**
  Manage structured data

**Data Integration**
- **HP IDOL Connectors & ETL**
  Extract, transform and load source system data

**Optional data sources**
- **Patient Records**
- **Blogs & News feeds**
- **Social Channels**
- **Call Center Text**
- **Enterprise Data**
Business Outcomes

Benefits
- Discovery of unknown defects
- Average time to identify defects reduced
- Proactive efforts resulted in overall reduced warranty cost
- Reduced risk to VW brand
- Industry excellence: Earlier defects detected in comparison to existing systems
Use Case 2: Healthcare Analytics @Universitätsspital Zürich
Optimize Revenue Management

ICD 10 Code Revision

Reduce miscoded Patient Records

Increase Revenue
HP ICD-10 Code Revision - Technology

Exploration
HP HealthCare
Interactive exploration and discovery

Reporting
QlikView
Standard reporting and trending

Language Translation
D2D
Translate languages

Unstructured Data
HP IDOL
Manage/Index text data

Structured Data
HP Vertica
Manage structured data

Data Integration
HP IDOL Connectors & ETL
Extract, transform and load source system data

Patient Records
Surgery Reports
Medication
Lab Results

External Topologies
Wide range of healthcare use cases including ...

Ultimate flexibility to scale

Abstraction

Reconciliation

ID discrepancies between diagnostic code and clinical notes

Monitor KPIs and Metrics

Reporting
Use Case 3: Social Media Analytics for Energy Industry
Be part of the industry conversations

How to get there?

Who are the influencers?

Where are they conversing?

Are they referring to our company in their conversations?

Has my influence as a brand gone up post engagement?

How do I engage with them?

Branding 4, 10%

Branding 3, 20%

Branding 2, 25%

Branding 1, 40%

Branding 5, 5%
Simplistic architecture

- HP Vertica
  Analytic DB
- HP IDOL
  Video, Audio, Text Encoding
- Hadoop
  Unstructured Data lake
- Spotfire
  Reporting
- R
  Advanced Analytics
- KPIs
- Reports
- Insights
- Interactions

Data Connectors

Social Media
News feeds
Video
Audio
Broadcast
Company Data
Hands-on IDOL
IDOL OnDemand APIs
Develop next-generation, insights-driven apps

- Search
- Sentiment Analysis
- Highlighting
- Language Detection
- Entity Extraction
- Text Extraction
- Convert to HTML

...and more
Danke für Ihr Interesse