Business Intelligence over Text in the Cloud

Alexander Löser
Technische Universität Berlin

Database Systems and Information Management
Technische Universität Berlin
Agenda

► Why BI-over-Text?

► Cloud Technology for BI-over-Text

► Next Steps
# 2008 CIO Priorities

## 2008 CIO Technology Priorities

To what extent will each of the following technologies be a Top 5 priority for you in 2008?

<table>
<thead>
<tr>
<th>Technology</th>
<th>Rank 2008</th>
<th>Rank 2007</th>
<th>Rank 2006</th>
<th>2008 Increase*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Intelligence Applications</td>
<td>1</td>
<td>✔</td>
<td>✔</td>
<td>11.20%</td>
</tr>
<tr>
<td>Enterprise Applications (ERP, SCM, and CRM)</td>
<td>2</td>
<td>✔</td>
<td>**</td>
<td>8.02%</td>
</tr>
<tr>
<td>Server and Storage Technologies (Virtualization)</td>
<td>3</td>
<td>✔</td>
<td>9</td>
<td>8.45%</td>
</tr>
<tr>
<td>Legacy Application Modernization</td>
<td>4</td>
<td>✔</td>
<td>10</td>
<td>5.79%</td>
</tr>
<tr>
<td>Security Technologies</td>
<td>5</td>
<td>✔</td>
<td>2</td>
<td>8.53%</td>
</tr>
<tr>
<td>Technical Infrastructure</td>
<td>6</td>
<td>✔</td>
<td>12</td>
<td>4.67%</td>
</tr>
<tr>
<td>Networking, Voice, and Data Communications (VoIP)</td>
<td>7</td>
<td>✔</td>
<td>8</td>
<td>6.83%</td>
</tr>
<tr>
<td>Collaboration Technologies</td>
<td>8</td>
<td>✔</td>
<td>4</td>
<td>7.75%</td>
</tr>
<tr>
<td>Document Management</td>
<td>9</td>
<td>✔</td>
<td>**</td>
<td>7.91%</td>
</tr>
<tr>
<td>Service-Oriented Technologies (SOA and SOBA)</td>
<td>10</td>
<td>✔</td>
<td>6</td>
<td>6.71%</td>
</tr>
</tbody>
</table>

* Unweighted average budget change
** New question for 2007

Source: 2008 Gartner Executive Programs CIO Survey, January 10, 2008
What are CIOs missing?

Please give me an example of how your business intelligence solution could better meet your organizations main objective?

Source: Business Intelligence Survey, IDC, May, 2005
Example Scenario: List “linux” companies

Actual Query Intention

SELECT companies
FROM “The Web”
WHERE company.technology = “linux”
Example Scenario: List “linux” companies

Simple Human Strategy to solve Task

- Google [Linux companies]
- Read Top-10 Pages
- Identify companies
- Copy relevant companies into PPT
Example Scenario: List “linux” companies

{Occurrence, Company Name}

{ "cnt": 10, "company": "Novell"},
{"cnt": 9,"company": "Microsoft"},
{"cnt": 6, "company": "Google"},
{"cnt": 5,"company": "Wal-Mart"},
{"cnt": 5,"company": "IBM"},
{"cnt": 4,"company": "Chunghwa Telecom"},
{"cnt": 3,"company": "Intel"},
{"cnt": 3,"company": "Red Hat Inc."},
{"cnt": 3,"company": "Dell"},
{"cnt": 2,"company": "Xandros"},
{"cnt": 2,"company": "MontaVista"},
{"cnt": 2,"company": "Oracle"},
{"cnt": 2,"company": "Linus Torvalds"},
{"cnt": 2,"company": "Fujitsu"},
{"cnt": 2,"company": "Lead Engineering Embedded Alley"}
Example Scenario: Zoom into document

{Occurence, Company Name}

{ "cnt": 10, "company": "Novell"},
{ "cnt": 9, "company": "Microsoft"},
{ "cnt": 6, "company": "Google"},
{ "cnt": 5, "company": "Wal-Mart"},
{ "cnt": 5, "company": "IBM"},
{ "cnt": 4, "company": "Chunghwa Telecom"},
{ "cnt": 3, "company": "Intel"},
{ "cnt": 3, "company": "Red Hat Inc."},
{ "cnt": 3, "company": "Dell"},
{ "cnt": 2, "company": "Xandros"},
{ "cnt": 2, "company": "MontaVista"},
{ "cnt": 2, "company": "Oracle"},
{ "cnt": 2, "company": "Linus Torvalds"},
{ "cnt": 2, "company": "Fujitsu"},
{ "cnt": 2, "company": "Lead Engineering Embedded Alley"}
“BI-Over-Text” : More complex queries

Product combinations for grocery shop.
1. How many customers in the forum “BlogSpot.com combine wine from “Rioja” with a cheese of the quality “semi-curado”? 
2. List combinations by quality of wine and age.
3. Limit results to wines no older than 1996.

Self-information of a cancer-patient.
1. Search for cancer-types in the forum “www.krebsforum-fuer-angehoerige.de”.
2. Group cancer types by frequency for female patients between 50 and 60 years.
3. Filter forum contributions by region = “Europe, Germany, Saxony”
"BI-Over-Text" from 10.000 feet above

Ad-hoc Query Process

1. Type initial query
2. Browse through structured results and text documents
3. Refine query using additional operators, data sources, keywords
4. Until satisfied, go back to step 2
“BI-Over-Text” Analytics

► Simple keywords to start analysis
  (How to capture the intention of a BI-over-Text query from 2-5 words?)
► Extract object identifiers and relationships
► Identify “relevant” dimension, measurements and facts in extracted data
► Identify “relevant” documents for dimensions, measurements and facts
► Identify “relevant” OLAP-style query refinement operations
  (What if initial query does not provide intended results?)

Parallel Execution

► Massive amount of text data (Google provides 826,000 pages)
► User expects fast response (Ideally sub-seconds)
► Execute crawling, text analytics, query processing on a distributed system
  (How to bring up current systems for text analytics on a web scale?)
Agenda

► What is BI-Over-Text?

► Cloud Technology for BI-over-Text

► Next Steps
Towards a stack for BI-Over-Text in the cloud

3 master students and supervisors, 3 months time

► Can we build a prototype to answer simple BI-Over-Text queries?

► What research problems can we derive?

► Only based on open source/free available software?

► How stable is existing cloud technology?
Cloud Computing

What is Cloud Computing?

- Computing platform architecture
- Scales to any application
- High fault tolerance
- No generally accepted definition available
- Separation from Utility or Grid Computing is not obvious
System Architecture

- JAQL Query
  - Import
    - JAQL Operator
  - Information Extraction
    - JAQL Operator
  - Process Query
  - Hadoop Framework
  - CACHE
    (Crawled Pages/Extracted Entities)
System Architecture

- JAQL Query
- Import
  - JAQL Operator
- Information Extraction
  - JAQL Operator
- Process Query
- CACHE
  - (Crawled Pages/Extracted Entities)
- Hadoop Framework
What is Hadoop?

- Free software framework for data intensive applications
- Enables distributed processing of vast amounts of data on cloud computing architectures
- Supports clouds with 1000+ nodes
- Two components:
  1) Hadoop Distributed File System (HDFS)
  2) MapReduce Engine

Where can you get Hadoop?

Hadoop - HDFS

- Inspired by Google File System
- Distributed storage for large files
- Files are split up in multiple parts (default size 64MB)
- Parts are spread over the HDFS nodes
- Each part replicated (default 3 times)
Hadoop – MapReduce Engine

► Runs MapReduce programs
► Libraries for Java and C++
► Assigns Map and Reduce tasks to computing nodes
► Reduction of data transfer volume
  ▪ Tasks are assigned to nodes holding the data
► Node failures are transparently handled
  ▪ Tasks are restarted on node holding a replica of the data
Who uses Hadoop?

- Amazon A9.com (Search Index Building, Analytics)
- Facebook (Logfile Analysis)
- Google & IBM (University Initiative to Address Internet-Scale Computing Challenges)
- Yahoo! (Crawling, Indexing, Searching)
  
  Yahoo! Hadoop Cluster runs Terabyte Sort Benchmark in 209 seconds

- And many others... (see http://wiki.apache.org/hadoop/PoweredBy)

Hadoop resembles Google‘s MapReduce Framework

- J. Dean, S. Ghemawat
  
  „MapReduce: Simplified Data Processing on Large Clusters“
JAQL 0.3

IBM Query Processing Language for Hadoop

- Pipe similar syntax (next version), not SQL-like
- Alternatives: Pig-Latin, Hive, Cascading, ...

- PRO: Works on semi-structured data (JSON)
- PRO: Extendible by user defined functions & operators
- PRO: Parallel execution model (map and reduce operations)
- (PRO: Close contact to developers)

- CON: Very early stage of development, e.g., syntax not stable

Our Job: Extend JAQL with operators for “BI-Over-Text”
Example Query Data Flow

Query = “Linux news”

$Urls = Function: Yahoo ($Query)

$Docs = Function: NUTCH ($URLS)

EACH $DOCS

$Parsed_text = Filter ($Doc)

$Parsed_text

EACH $DOC

$CalaisData = EXTRACT ($DOC)

$LinuxComp = FILTER (“LINUX”)
registerUDFs
registerFunction("fOpenCalais","de.tuberlin.dima.jaql.extensions.FullOpenCalais");
registerFunction("nutchCrawler","de.tuberlin.dima.jaql.extensions.NutchCrawl");
registerFunction("yahooBoss","de.tuberlin.dima.jaql.extensions.YahooBoss");

// find URLs
(urls = yahooBoss("linux news",100));

// crawl URLs
$crawledData = nutchCrawler(urls,"NUTCH",1,100);

// extract parsed text from crawled data
$parsedText = for($doc in $crawledData)
  if(exists($doc.ParseText)) [
    Text: $doc.ParseText.Text
  ];

// call openCalais for crawled parsed text
$calaisData = for ($doc in $parsedText)
  [fOpenCalais($doc.Text)];

// extract companies
$linuxComps = for ($cEntities in $calaisData[*].cEntities)
  for($cEntity in $cEntities)
    if($cEntity.entityType == 'CompanyTechnology') // filter for technology 'Linux'
      if($cEntity.properties.technology == 'Linux') // return company
        [company: $cEntity.properties.company];
$linuxCompCounts = group($comp in $linuxComps by $k = $comp.company into $bin)
  [{company: $k, cnt: count($bin)}];
$sortedLinuxCompCounts = sort($comp in $linuxCompCounts by $comp.cnt desc);
$sortedLinuxCompCounts;
Example JAQL Code incl. new functions

```java
register UDFS
registerFunction("fOpenCalais","de.tuberlin.dima.jaql.extensions.FullOpenCalais");
registerFunction("nutchCrawler","de.tuberlin.dima.jaql.extensions.NutchCrawl");
registerFunction("yahooBoss","de.tuberlin.dima.jaql.extensions.YahooBoss");

// find URLs
$urls = yahooBoss(" linux news",100);
// crawl URLs
$crawledData = nutchCrawler($urls,"NUTCH",1,100);
// extract parsed text from crawled data
$parsedText = for($doc in $crawledData)
    if(exists($doc.ParseText))
    {
        {Text: $doc.ParseText.Text};
    }
// call openCalais for crawled parsed text
$calaisData = for ($doc in $parsedText)
    [fOpenCalais($doc.Text)];
// extract companies
$linuxComps = for ($cEntities in $calaisData[*].cEntities)
// iterate over all cEntities
    for($cEntity in $cEntities) // filter for 'CompanyTechnology' type
        if($cEntity.entityType == 'CompanyTechnology') // filter for technology 'Linux'
            if($cEntity.properties.technology == 'Linux') // return company
                {{company: $cEntity.properties.company}};
$linuxCompCounts = group($comp in $linuxComps by $k = $comp.company into $bin)
    {{company: $k, cnt: count($bin)}};
$sortedLinuxCompCounts = sort($comp in $linuxCompCounts by $comp.cnt desc);
$sortedLinuxCompCounts;
```

Source Selection | Data Crawling | Entity Extraction | Query Processing

© Chair for Database Systems and Information Management
Example JAQL Code incl. new functions

```
register UDFs
registerFunction("fOpenCalais","de.tuberlin.dima.jaql.extensions.FullOpenCalais");
registerFunction("nutchCrawler","de.tuberlin.dima.jaql.extensions.NutchCrawl");
registerFunction("yahooBoss","de.tuberlin.dima.jaql.extensions.YahooBoss");

// find URLs
$urls = yahooBoss("linux news",100);

// crawl URLs
$crawledData = nutchCrawler($urls,"NUTCH",1,100);

// extract parsed text from crawled data
$parsedText = for($doc in $crawledData)
    if(exists($doc.ParseText))
        [{Text: $doc.ParseText.Text}];

// call openCalais for crawled parsed text
$calaisData = for ($doc in $parsedText)
    [fOpenCalais($doc.Text)];

// extract companies
$linuxComps = for ($cEntities in $calaisData[*].cEntities)
    // iterate over all cEntities
    for($cEntity in $cEntities)  // filter for 'CompanyTechnology' type
        if($cEntity.entityType == 'CompanyTechnology')  // filter for technology 'Linux'
            if($cEntity.properties.technology == 'Linux')  // return company
                [{company: $cEntity.properties.company}];

$linuxCompCounts = group($comp in $linuxComps by $k = $comp.company into $bin)
    [{company: $k, cnt: count($bin)}];

$sortedLinuxCompCounts = sort($comp in $linuxCompCounts by $comp.cnt desc);
$sortedLinuxCompCounts;
```
Example JAQL Code incl. new functions

```jaql
// find URLs
$url = yahooBoss("linux news",100);
// crawl URLs
$crawledData = nutchCrawler($url,"NUTCH",1,100);
// extract parsed text from crawled data
$parsedText = for($doc in $crawledData)
    if(exists($doc.ParseText)) [Text: $doc.ParseText.Text];
// call openCalais for crawled parsed text
$calaisData = for ($doc in $parsedText)
    [fOpenCalais($doc.Text)];
// extract companies
$linuxComps = for ($cEntities in $calaisData[*].cEntities)
// iterate over all cEntities
    for($cEntity in $cEntities) // filter for 'CompanyTechnology' type
        if($cEntity.entityType == 'CompanyTechnology')// filter for technology 'Linux'
            if($cEntity.properties.technology == 'Linux') // return company
                [{company: $cEntity.properties.company}];
$linuxCompCounts = group($comp in $linuxComps by $k = $comp.company into $bin)
    [{company: $k, cnt: count($bin)}];
$sortedLinuxCompCounts = sort($comp in $linuxCompCounts by $comp.cnt desc);
$sortedLinuxCompCounts;
```
Example JAQL Code incl. new functions

register UDFs
registerFunction("fOpenCalais","de.tuberlin.dima.jaql.extensions.FullOpenCalais");
registerFunction("nutchCrawler","de.tuberlin.dima.jaql.extensions.NutchCrawl");
registerFunction("yahooBoss","de.tuberlin.dima.jaql.extensions.YahooBoss");

// find URLs
$urls = yahooBoss("linux news",100);

// crawl URLs
$crawledData = nutchCrawler($urls,"NUTCH",1,100);

// extract parsed text from crawled data
$parsedText = for($doc in $crawledData)
  if(exists($doc.ParseText))   
    {Text: $doc.ParseText.Text};

// call openCalais for crawled parsed text
$calaisData = for ($doc in $parsedText)
  [fOpenCalais($doc.Text)];

// extract companies
$linuxComps = for ($cEntities in $calaisData[*].cEntities)
  // iterate over all cEntities
  for($cEntity in $cEntities)   // filter for 'CompanyTechnology' type
    if($cEntity.entityType == 'CompanyTechnology')   // filter for technology 'Linux'
      if($cEntity.properties.technology == 'Linux')               // return company

      [{company: $cEntity.properties.company}];

$linuxCompCounts = group($comp in $linuxComps by $k = $comp.company into $bin)
  [{company: $k, cnt: count($bin)} ];

$sortedLinuxCompCounts = sort($comp in $linuxCompCounts by $comp.cnt desc);
$sortedLinuxCompCounts;
Results of Yahoo Boss

[
  {
    "url": "http://linuxtoday.com/"
  },
  {
    "url": "http://www.linux.com/"
  },
  {
    "url": "http://www.linux.com/feature/c4201"
  },
  {
    "url": "http://lxer.com/"
  },
  {
    "url": "http://www.linuxworld.com/"
  },
  {
    "url": "http://news.softpedia.com/cat/Linux/"
  },
  ...
]
Results of NUTCH

[
{
"ID": "http://www.freshtechnews.com/linux.html",
"ParseData": {
.....
"ETag": "b3c0b9-67a9-c3698e40",
"Last-Modified": "Thu, 04 Dec 2008 08:58:25 GMT",
"OriginalCharEncoding": "windows-1252",
"OutLinks": {
"Count": 51,
"Urls": [
{
"Anchor": "",
"ToUrl": "http://www.loveme.com/go/89/"
},
...
{
"ToUrl": "http://www.freshtechnews.com/security.html"
}
]
},
"Server": "Apache/2.0.50 (Fedora)",
"Title": "Fresh Tech News - Linux News Headlines",
"URL": "http://www.freshtechnews.com/linux.html",
...,
"nutch.segment.name": "20081204103350"
},
"ParseText": {"Text": "Fresh Tech News - Linux News Headlines Fresh Tech News - Linux News Headlines ... "},
"RawContent": {"ContentTyp": "text/html", "Source": 
"<html><head><title>Fresh Tech News - Linux News Headlines</title><meta http-equiv="Content-Language" content="en">
"} }
]
Oracle, Emulex grant Linux data integrity Database maker Oracle and host bus adapter maker Emulex today announced that they have contributed code to eliminate silent data corruption to the open source Linux operating system. The two also said this code has been accepted.
Results: List “linux” companies

{Occurrence, Company Name}

{"cnt": 10, "company": "Novell"},
{"cnt": 9,"company": "Microsoft"},
{"cnt": 6," company": "Google"},
{"cnt": 5,"company": "Wal-Mart"},
{"cnt": 5,"company": "IBM"},
{"cnt": 4,"company": "Chunghwa Telecom"},
{"cnt": 3,"company": "Intel"},
{"cnt": 3,"company": "Red Hat Inc."},
{"cnt": 3,"company": "Dell"},
{"cnt": 2,"company": "Xandros"},
{"cnt": 2,"company": "MontaVista"},
{"cnt": 2,"company": "Oracle"},
{"cnt": 2,"company": "Linus Torvalds"},
{"cnt": 2,"company": "Fujitsu"},
{"cnt": 2,"company": "Lead Engineering Embedded Alley"}
Lessons Learned

I will backup my laptop every day.
I will backup my laptop every day.
I will backup my laptop every day.
I will backup my laptop every day.
Lessons Learned (Research Problems)

► Most UIMA jobs are on a single document (Local analysis)
  ▪ Operations on a single document can be abstracted as MAP operation.
  ▪ **Executing annotators** language detection, tokenization, POS tagging, named entity resolution, rule-based annotators, relationship annotators or sentiment analysis annotators

► Few operations are on a set of documents (Global Analysis)
  ▪ **Global Analysis operations require an REDUCE and MERGE operation.**
  ▪ **Training classifiers** for named entity or relationship extraction
  ▪ **Executing annotators** for corpus specific tasks, such as home-page annotation or object identification

► Goal
  ▪ Start with local analysis operations
  ▪ Execute them as a MAP operation in UIMA Wrapper
Lessons Learned (Research Problems)

► Current JAQL status
  ▪ Executes Map Reduce plans independent of
    • file size
    • operators in query plan
    • or execution time (costs) per operator

► Goal: Incorporate query plan into map/reduce execution
  ▪ Investigate new methods for
    • Re-order MR jobs
    • Monitor MR jobs
    • Modify MR jobs at query run time
MR Execution Manager

Map Reduce Query Plan

MapReduce Execution Manager

MR Monitor

Hadoop

start MR jobs
monitor running MR jobs
shift resources
Lessons Learned (Research Problems)

Why caching?
- Speedup: Local Cache instead of Remote Web Page
- Availability: Local copy from Web
- Sample base: (Extraction) Operations, Query Optimization

Goal: Add cache for documents and extracted data
- Store crawled data and extracted entities
- Develop adaptors for JAQL

Available candidate systems
- HBASE: Open Source Big Table Version, now supported by Microsoft
- CASSANDRA: Structured P2P Network used for storage, developed by Face Book *(our choice)*
Example JAQL Code incl. new functions

```
registerFunction("nutchCrawler", "de.tuberlin.dima.jaql.extensions.NutchCrawl");
registerFunction("yahooBoss", "de.tuberlin.dima.jaql.extensions.YahooBoss");
registerFunction("JaqlFieldFromCache", "tub.dima.cassandraconnector.JaqlFieldFromCache");
registerFunction("InsertJaqlInCache", "tub.dima.cassandraconnector.InsertJaqlInCache");
registerFunction("JaqlMergeArrays", "tub.dima.cassandraconnector.JaqlMergeArrays");
$url = yahooBoss("linux news", 100);
// find urls in cache
$cacheURLS = JaqlFieldFromCache($url, true, "url");
// find urls not in cache
$nocacheURLS = JaqlFieldFromCache($url, false, "url");
// crawl data from urls not in cache
$crawledData = nutchCrawler($nocacheURLS, "NUTCH", 1, 5);
// insert new crawl data in cache
InsertJaqlInCache($crawledData);
// get parsed data from cache
$crawledCacheData = JaqlFieldFromCache($url, true, "ALL");
// merge cache und crawled data
$crawledDataAll = JaqlFieldFromCache($crawledData, $crawledCacheData);
// call query to crawled und cached data
....
```
Agenda

► Why BI-Over-Text?

► BI-Over-Text using state–of-the art technology

► Next steps
What is next?

► Towards an Algebra for BI-Over-Text
  ▪ How to identify ORDER BY clauses in keyword queries and corresponding measurements in text?
    • E.g., \[youngest chancellors germany\] denotes ORDER BY age
  ▪ Understand analysis-schema extracted from aggregated text
  ▪ Weight query interpretations using statistics from extracted data

► Extensions on Hadoop Level to speed up analysis
  ▪ What other operators beside Map and Reduce do we need?
  ▪ How to model costs for operations?
  ▪ Gather statistics for query interpretation efficiently on cloud

► Global Text-Analysis as Map Reduce Problem
  ▪ How to train classifiers (e.g., for unsupervised relationship extraction) effectively on a map-reduce platform?
What is next?: GOOLAP.INFO

Aggregating events, facts etc. about Persons, Companies, Organizations etc.
What is next?: GOOLAP.INFO

Gooolap

Search query: "Paris Hilton"

Results 1 - 3 from 3 results for your search query "Paris Hilton" and we guess you mean Person "Paris Hilton"

**About "Paris Hilton"**

Quotation: 2 results found

```
... tearfully to the Judge handling her case in court yesterday...  

It's not right...```

Person: 2 results found

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datestring</td>
<td>on Sunday</td>
</tr>
<tr>
<td>Status</td>
<td>past</td>
</tr>
<tr>
<td>TravelDestination</td>
<td>Los Angeles</td>
</tr>
<tr>
<td>Date</td>
<td>2008-05-14</td>
</tr>
</tbody>
</table>

Family relation: 2 results found

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person_relative</td>
<td>Kathy Hilton</td>
</tr>
<tr>
<td>Familyrelationtype</td>
<td>parent</td>
</tr>
<tr>
<td>Person_relative</td>
<td>Rick Hilton</td>
</tr>
</tbody>
</table>

**Articles to "Paris Hilton"**

- **DocID: 10726**
  - **Title:** Paris Hilton is out of jail, after serving three weeks in the same. She was imprisoned due to a violation of her probation.
  - **Details:** Paris Hilton expressed joy at her release as she left the Century Regional Detention Facility in Lynwood, Ca... The document was created on 2008-09-17.

- **DocID: 10501**
  - **Title:** With photographers and media gathering for an appearance by Paris Hilton, who is sentenced to serve jail time, the fashion line of prison-striped lingerie embroidered with the words 'Free Paris' in support of Hilton. A... The document was created on 2008-09-17.

- **DocID: 10550**
  - **Title:** A day after being released to house arrest, Paris Hilton was brought back into court today in handcuffs, and ordered to report in jail. Hilton was escorted from the courtroom screaming and crying. It's not right... The document was created on 2008-09-17.
Now
► Prototype executes list-type queries on a cloud.
► Works with state-of-the-art cloud technology.

Next
► Query refinement and elementary Hadoop operators.
► Web-query-interface for community testing
► Extension to larger infrastructure planned for 2009.
Acknowledgements

► Stephan Ewen
► Roland Hager
► Fabian Hueske
► Alexander Löser
► Volker Markl
► Bernd Rabe
► Ronny Schwierzinski
Select Literature

Information Extraction

► **Self-Supervised Learning**: Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, Oren Etzioni: Open Information Extraction from the Web. *International Joint Conferences on Artificial Intelligence (IJCAI) 2007: 2670-2676*

► **Algebraic**: Frederick Reiss, Shivakumar Vaithyanathan, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu: An Algebraic Approach to Rule-Based Information Extraction. *International Conference on data engineering (ICDE) 2008: 933-942*

Schema generation from extracted uncertain data


► Marcos Antonio Vaz Salles, Jens-Peter Dittrich, Shant Kirakos Karakashian, Olivier René Girard, Lukas Blunschi: iTrails: Pay-as-you-go Information Integration in Dataspaces. *International Conference on Very Large Databases (VLDB) 2007: 663-674*

Optimization in Text Databases

► Alpa Jain, AnHai Doan, Luis Gravano: Optimizing SQL Queries over Text Databases. *International Conference on Data Engineering (ICDE) 2008: 636-645*

BI-Over-Text

► Alpa Jain, AnHai Doan, Luis Gravano: Optimizing SQL Queries over Text Databases. *International Conference on Data Engineering (ICDE) 2008: 636-645*


► Web 2.0 Business Analytics. Alexander Löser, Gregor Hackenbroich, Hong-Hai Do, Henrike Berthold. *Datenbank Spektrum 25/2008*


► R-Cubes: OLAP Cubes Contextualized with Documents. Juan Manuel Perez et.al. *ICDE 2007*
Lessons Learned (Engineering Level)

Hurdles to overcome

► Code of cloud components usually 0.XXX version (XXX<3) 😞
► Often zero documentation available
► Lots of RAM per core required
► Most of the time went into configuration problems, missing class path entries, security permissions

Implications

► VM-ware image for distributing our BI-Over-Text Cloud stack
► Simplified access for community by web-based query interface
Text Analytics

Out-of-the-box data
► Web Services for complex, atomic and named entities

Significant additional effort
► Infrastructures for extracting, managing and scalable storage of named entities
► Web Services for extracting named entities

High additional effort
► Screen scraper
OpenCalais.ORG

- PRO: Fast web service for 30+ facts + events (crucial for BI)
- PRO: Web-wide object identification (crucial for BI)
- PRO: Encapsulates complexity of Text Analytics
- PRO: Free up to 4 requests per second
- PRO: No real alternative currently

- CON: Domain mostly limited to news
- CON: Not extensible (at least not in the free web version)
- CON: RDF format difficult to parse (e.g., JENA Toolkit)

Non-Free alternative: TEXTRUNNER Project  Etzioni et.al.
<table>
<thead>
<tr>
<th>Extractor</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balie</td>
<td>0.6374</td>
<td>0.6726</td>
</tr>
<tr>
<td>Calais</td>
<td>0.9395</td>
<td>0.7408</td>
</tr>
<tr>
<td>LingPipe</td>
<td>0.6418</td>
<td>0.5608</td>
</tr>
<tr>
<td>OpenNLP</td>
<td>0.7772</td>
<td>0.3844</td>
</tr>
<tr>
<td>Stanford</td>
<td>0.9645</td>
<td>0.8843</td>
</tr>
</tbody>
</table>

Source: Own experiment on Reuters Corpus CONLL 2002 Entity Extraction Goldstandard (Person)