

# Business Intelligence over Text in the Cloud

Alexander Löser

Technische Universität Berlin



Database Systems and Information Management  
Technische Universität Berlin

- ▶ Why BI-over-Text?
- ▶ Cloud Technology for BI-over-Text
- ▶ Next Steps

## 2008 CIO Technology Priorities

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

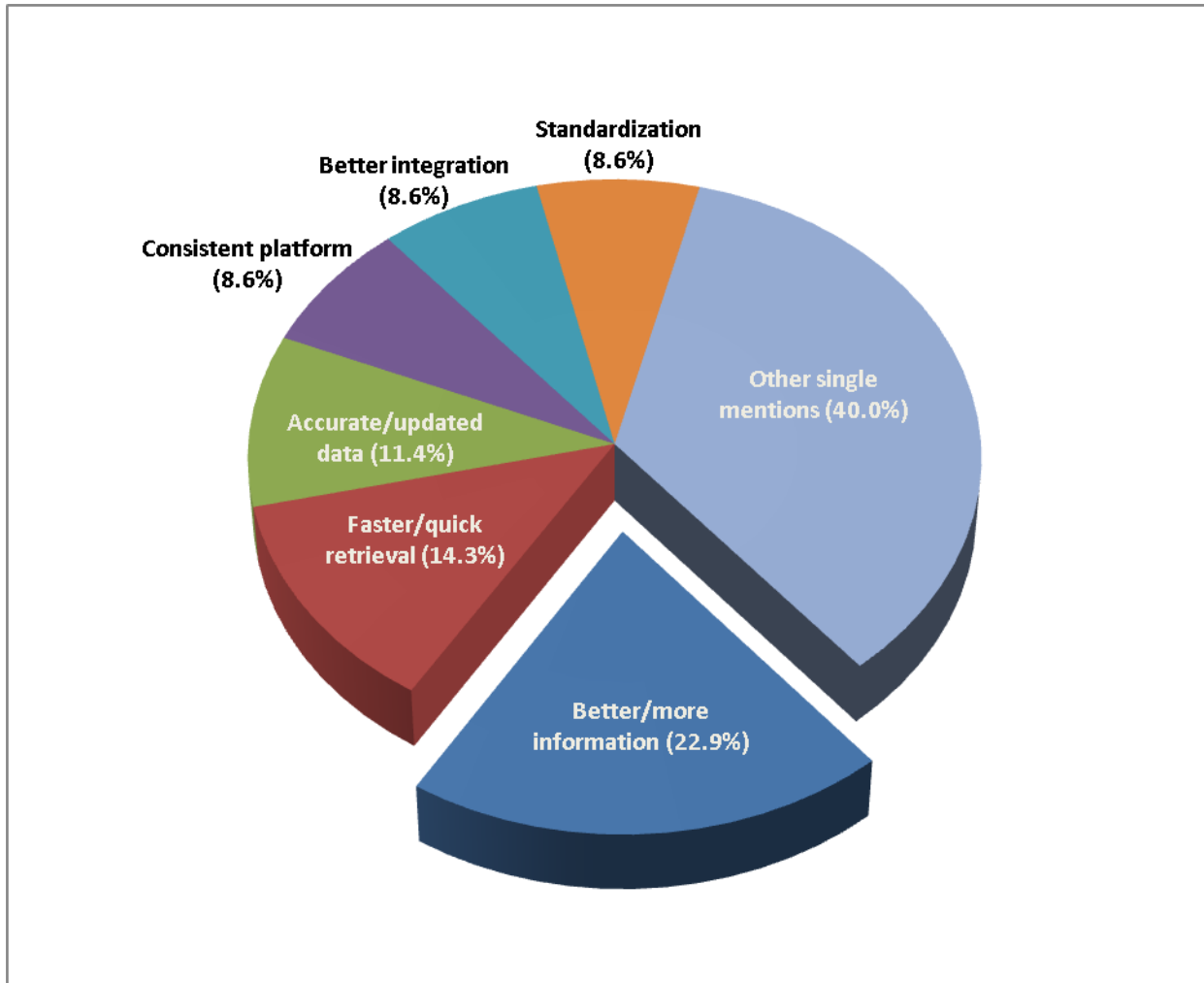
	Rank 2008		Rank 2007	Rank 2006	2008 Increase*
Business Intelligence Applications	1	←	1	1	11.20%
Enterprise Applications (ERP, SCM, and CRM)	2	←	2	**	8.02%
Server and Storage Technologies (Virtualization)	3	▲	5	9	8.45%
Legacy Application Modernization	4	▼	3	10	5.79%
Security Technologies	5	▲	6	2	8.53%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

Source: 2008 Gartner Executive Programs CIO Survey, January 10, 2008

\* Unweighted average budget change

\*\* New question for 2007

# 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



Google  Search [Advanced Search](#) [Preferences](#)

Web Results 1 - 10 of about 967,000 for [linux companies](#). (0.36 seconds)

**Category: Linux companies - Wikipedia, the free encyclopedia**  
24 Apr 2008 ... Linux-based companies are companies that have at the core a business model that involves the production of Linux distributions or developing ...  
[en.wikipedia.org/wiki/Category:Linux\\_companies](http://en.wikipedia.org/wiki/Category:Linux_companies) - 24k - [Cached](#) - [Similar pages](#)

**Some Companies Using Linux**  
A vast number of companies and organizations are deploying Linux for at least some applications. We list here a few of the ones that use Linux for major ...  
[www.aaxnet.com/design/linux2.html](http://www.aaxnet.com/design/linux2.html) - 18k - [Cached](#) - [Similar pages](#)

**Linux Magazine's Top 20 Companies to Watch in 2008 | Linux Magazine**  
10 Jan 2008 ... Microsoft is putting a lot of proprietary patented software into Linux through Novell. Most of the good guys at Novell have left the company ...  
[www.linux-mag.com/id/4766/](http://www.linux-mag.com/id/4766/) - 58k - [Cached](#) - [Similar pages](#)

**Google Directory - Computers > Software > Operating Systems ...**  
Directory Help Search only in Companies Search the Web ... Support services and software for Linux operating environment. ...  
[www.google.com/Top/Computers/Software/Operating\\_Systems/Linux/Companies/](http://www.google.com/Top/Computers/Software/Operating_Systems/Linux/Companies/) - 23k - [Cached](#) - [Similar pages](#)

**Linux Links - The Linux Portal: Companies**  
Part of Linux Links, comprehensive information and resources about the Linux Operating System.  
[www.linuxlinks.com/Companies/](http://www.linuxlinks.com/Companies/) - 58k - [Cached](#) - [Similar pages](#)

**Robert Half Technology - Linux/NOC Engineer - Growing Company!**  
Linux/NOC Engineer - Growing Company! - Find Engineering Jobs, Information Technology Jobs, Design Jobs at Robert Half Technology in San Jose, California.  
[www.careerbuilder.com/JobSeeker/Jobs/JobDetails.aspx?Job\\_DID=J3F66G77S17NKL11BRK-57k](http://www.careerbuilder.com/JobSeeker/Jobs/JobDetails.aspx?Job_DID=J3F66G77S17NKL11BRK-57k) - [Cached](#) - [Similar pages](#)

**Linux Hosting Company UK - Directory of UK Linux Hosting Companies**  
Directory of UK Linux Hosting Companies. Detailed profiles, customer reviews, location maps and photos of Linux Hosting Companies located within the UK.  
[www.freeindex.co.uk/categories/computers\\_and\\_internet/web\\_hosting/linux\\_hosting/](http://www.freeindex.co.uk/categories/computers_and_internet/web_hosting/linux_hosting/) - 35k - [Cached](#) - [Similar pages](#)

**Linux.com :: South African sister companies praise Linux-based ...**  
2 posts - Last post: 4 Aug  
Gospel Direct and Maranatha Record Co., sister companies based in South Africa, have exercised their faith in a Linux-based accounting ...  
[www.linux.com/feature/142777](http://www.linux.com/feature/142777) - 33k - [Cached](#) - [Similar pages](#)

**Why Microsoft and Linux companies are tying the knot**  
17 Jun 2007 ... Before answering that one, let's ask the other question: Why is Microsoft getting all lovey-dovey with Linux companies? Is it ...  
[www.linux-watch.com/news/NS6099316851.html](http://www.linux-watch.com/news/NS6099316851.html) - 37k - [Cached](#) - [Similar pages](#)

**LXer: Companies selling preinstalled Linux and no-OS: Linux**  
100+ posts - 63 authors - Last post: 28 Oct  
Or would it be better to simply drop the company from the DB if they are no longer offering GNU/Linux systems (or went out of business)? ...  
[l.xer.com/module/forums/t/23168/](http://l.xer.com/module/forums/t/23168/) - 206k - [Cached](#) - [Similar pages](#)

Google

## Actual Query Intention

```
SELECT companies
FROM "The Web"
WHERE company.technology = "linux"
```



# Example Scenario: List “linux” companies



Google  Search [Advanced Search](#) [Preferences](#)

Web Results 1 - 10 of about 967,000 for [linux companies](#). (0.36 seconds)

**Category: Linux companies - Wikipedia, the free encyclopedia**  
24 Apr 2008 ... Linux-based companies are companies that have at the core a business model that involves the production of Linux distributions or developing ...  
[en.wikipedia.org/wiki/Category:Linux\\_companies](http://en.wikipedia.org/wiki/Category:Linux_companies) - 24k - [Cached](#) - [Similar pages](#)

**Some Companies Using Linux**  
A vast number of companies and organizations are deploying Linux for at least some applications. We list here a few of the ones that use Linux for major ...  
[www.aaxnet.com/design/linux2.html](http://www.aaxnet.com/design/linux2.html) - 18k - [Cached](#) - [Similar pages](#)

**Linux Magazine's Top 20 Companies to Watch in 2008 | Linux Magazine**  
10 Jan 2008 ... Microsoft is putting a lot of proprietary patented software into Linux through Novell. Most of the good guys at Novell have left the company ...  
[www.linux-mag.com/id/4766/](http://www.linux-mag.com/id/4766/) - 58k - [Cached](#) - [Similar pages](#)

**Google Directory - Computers > Software > Operating Systems ...**  
Directory Help Search only in Companies Search the Web ... Support services and software for Linux operating environment. ...  
[www.google.com/Top/Computers/Software/Operating\\_Systems/Linux/Companies/](http://www.google.com/Top/Computers/Software/Operating_Systems/Linux/Companies/) - 23k - [Cached](#) - [Similar pages](#)

**Linux Links - The Linux Portal: Companies**  
Part of Linux Links, comprehensive information and resources about the Linux Operating System.  
[www.linuxlinks.com/Companies/](http://www.linuxlinks.com/Companies/) - 58k - [Cached](#) - [Similar pages](#)

**Robert Half Technology - Linux/NOC Engineer - Growing Company!**  
Linux/NOC Engineer - Growing Company! - Find Engineering Jobs, Information Technology Jobs, Design Jobs at Robert Half Technology in San Jose, California.  
[www.careerbuilder.com/JobSeeker/Jobs/JobDetails.aspx?Job\\_DID=J3F66G77S17NKL11BRK-57k](http://www.careerbuilder.com/JobSeeker/Jobs/JobDetails.aspx?Job_DID=J3F66G77S17NKL11BRK-57k) - [Cached](#) - [Similar pages](#)

**Linux Hosting Company UK - Directory of UK Linux Hosting Companies**  
Directory of UK Linux Hosting Companies. Detailed profiles, customer reviews, location maps and photos of Linux Hosting Companies located within the UK.  
[www.freeindex.co.uk/categories/computers\\_and\\_internet/web\\_hosting/linux\\_hosting/](http://www.freeindex.co.uk/categories/computers_and_internet/web_hosting/linux_hosting/) - 35k - [Cached](#) - [Similar pages](#)

**Linux.com :: South African sister companies praise Linux-based ...**  
2 posts - Last post: 4 Aug  
Gospel Direct and Maranatha Record Co., sister companies based in South Africa, have exercised their faith in a Linux-based accounting ...  
[www.linux.com/feature/142777](http://www.linux.com/feature/142777) - 33k - [Cached](#) - [Similar pages](#)

**Why Microsoft and Linux companies are tying the knot**  
17 Jun 2007 ... Before answering that one, let's ask the other question: Why is Microsoft getting all lovey-dovey with Linux companies? Is it ...  
[www.linux-watch.com/news/NS6099316851.html](http://www.linux-watch.com/news/NS6099316851.html) - 37k - [Cached](#) - [Similar pages](#)

**LXer: Companies selling preinstalled Linux and no-OS: Linux**  
100+ posts - 63 authors - Last post: 28 Oct  
Or would it be better to simply drop the company from the DB if they are no longer offering GNU/Linux systems (or went out of business)? ...  
[lxe.com/module/forums/t/23168/](http://lxe.com/module/forums/t/23168/) - 206k - [Cached](#) - [Similar pages](#)

Google

## Simple Human Strategy to solve Task

- Google [Linux companies]
- Read Top-10 Pages
- Identify companies
- Copy relevant companies into PPT



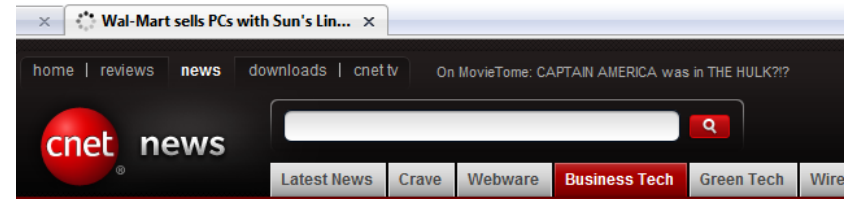
**{Occurrence , Company Name}**

```
{ "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"}
```



## {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"}
```



Home > News >  
March 30, 2004 4:50 PM PST  
**Wal-Mart sells PCs with Sun's Linux**  
By Stephen Shankland  
Staff Writer, CNET News

### Welcome Google user!

More headlines related to "Companies Linux %22Wal-Mart%22":

- Faces of the recession
- Andrew J. McKelvey, builder of Monster.com, dies
- Yahoo's search for a new CEO
- TV sales becoming litmus test for U.S. economy
- More matching headlines >

### Add CNET News to Google

Add CNET News headlines to your Google homepage or Google reader.



### Related Stories

Novell inks deals for IBM

MENLO PARK, Calif.--Wal-Mart Stores, the world's largest retailer, has begun selling Microtel PCs that come with Sun Microsystems' version of the Linux operating system.



## 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”

[Category Linux companies](#) - Wikipedia, the free encyclopedia  
28 Apr 2002 ... Linux-based companies are companies that have at the core a business model that involves the production of Linux distributions or developing ...  
[en.wikipedia.org/wiki/Category:Linux\\_companies](#) - 204 - [Cached](#) - [Similar pages](#)

[Some Companies Using Linux](#)  
A vast number of companies and organizations are deploying Linux for at least some applications. We list here a few of the ones that use Linux for major ...  
[www.sas.com/education/News/News\\_198](#) - [Cached](#) - [Similar pages](#)

[Linux Magazine's Top 20 Companies to Watch in 2008](#) | Linux Magazine  
10 Jan 2008 ... Microsoft is putting a lot of proprietary patented software into Linux through Novell. Most of the good guys at Novell have left the company ...  
[www.linux-mag.com/ENR/ENR\\_586](#) - [Cached](#) - [Similar pages](#)

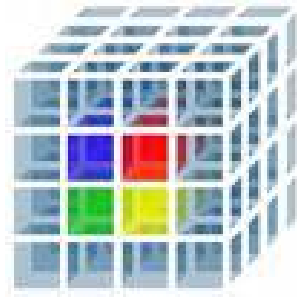
[Google Directory - Computers > Software > Operating Systems ...](#)  
Directory Help Search only in Companies Search the Web ... Support services and software for Linux operating environment ...  
[www.google.com/TopComputers/Software/Operating\\_Systems/Linux/Companies/](#) - 238 - [Cached](#) - [Similar pages](#)

[Linux Links - The Linux Portal: Companies](#)  
Part of Linux Links, comprehensive information and resources about the Linux Operating System.  
[www.linuxlinks.com/Companies/](#) - 586 - [Cached](#) - [Similar pages](#)

[Robert Half Technology - Linux/NOC Engineer - Growing Company!](#)  
Linux/NOC Engineer - Growing Company! Find Engineering Jobs, Information Technology Jobs, Design Jobs at Robert Half Technology in San Jose, California  
[www.careerbuilder.com/job/Search/jobDetail.aspx?job\\_ID=0F660775179K118R&C=876](#) - [Cached](#) - [Similar pages](#)

[Linux Hosting Company UK - Directory of UK Linux Hosting Companies](#)  
Directory of UK Linux Hosting Companies. Detailed profiles, customer reviews, location maps and photos of Linux Hosting Companies located within the UK.  
[www.hosted.co.uk/Categories/computer\\_and\\_internet/web\\_hosting/linux\\_hosting/](#) - 306 - [Cached](#) - [Similar pages](#)

+



## 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 relation ships
- ▶ 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?)

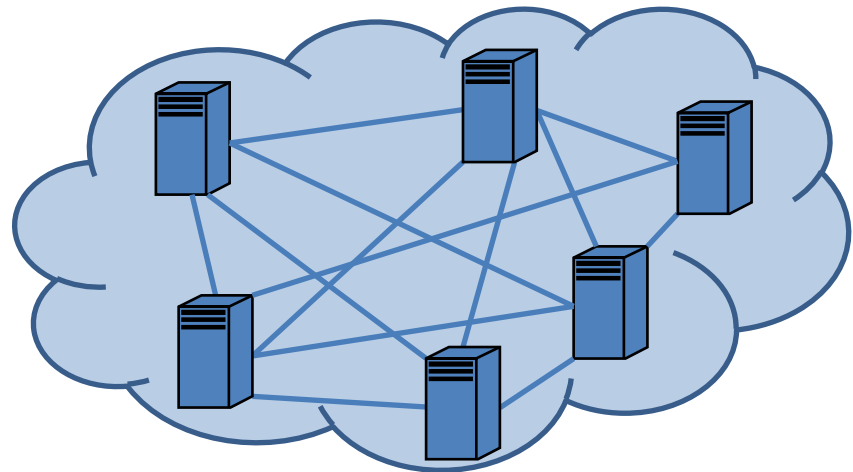
- ▶ What is BI-Over-Text?
- ▶ **Cloud Technology for BI-over-Text**
- ▶ Next Steps

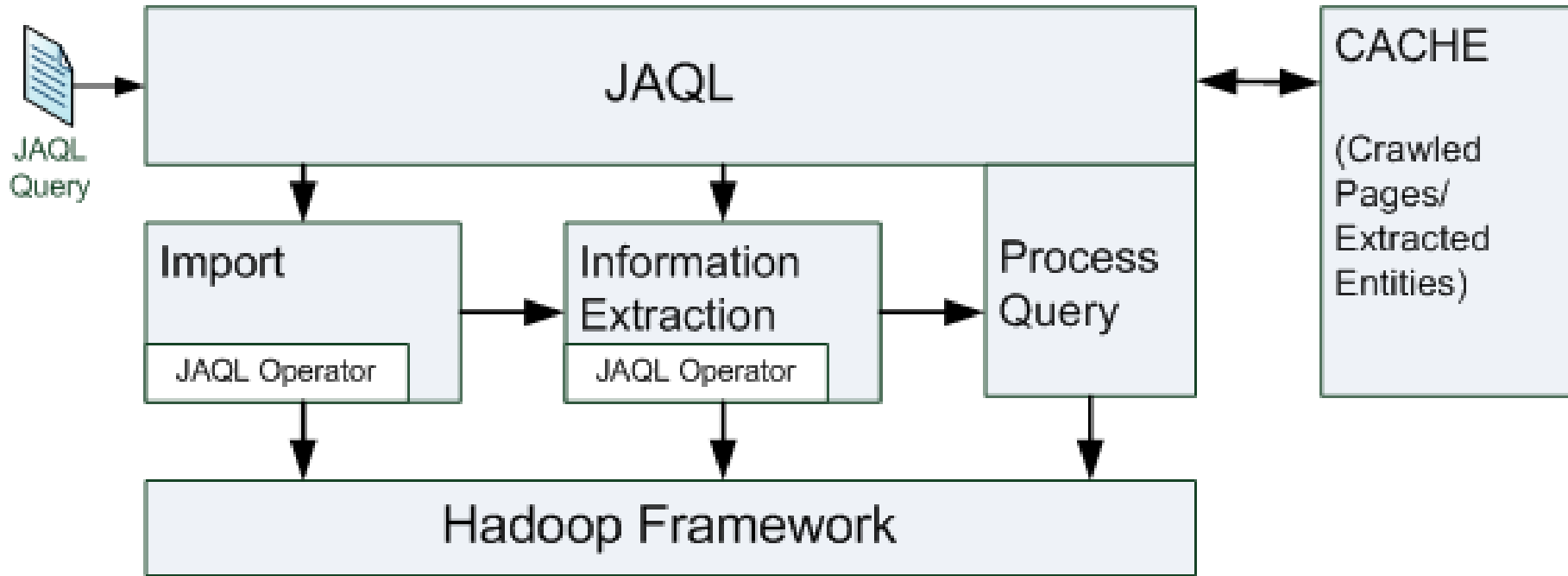
## 3 master students and supervisors, 3 months time

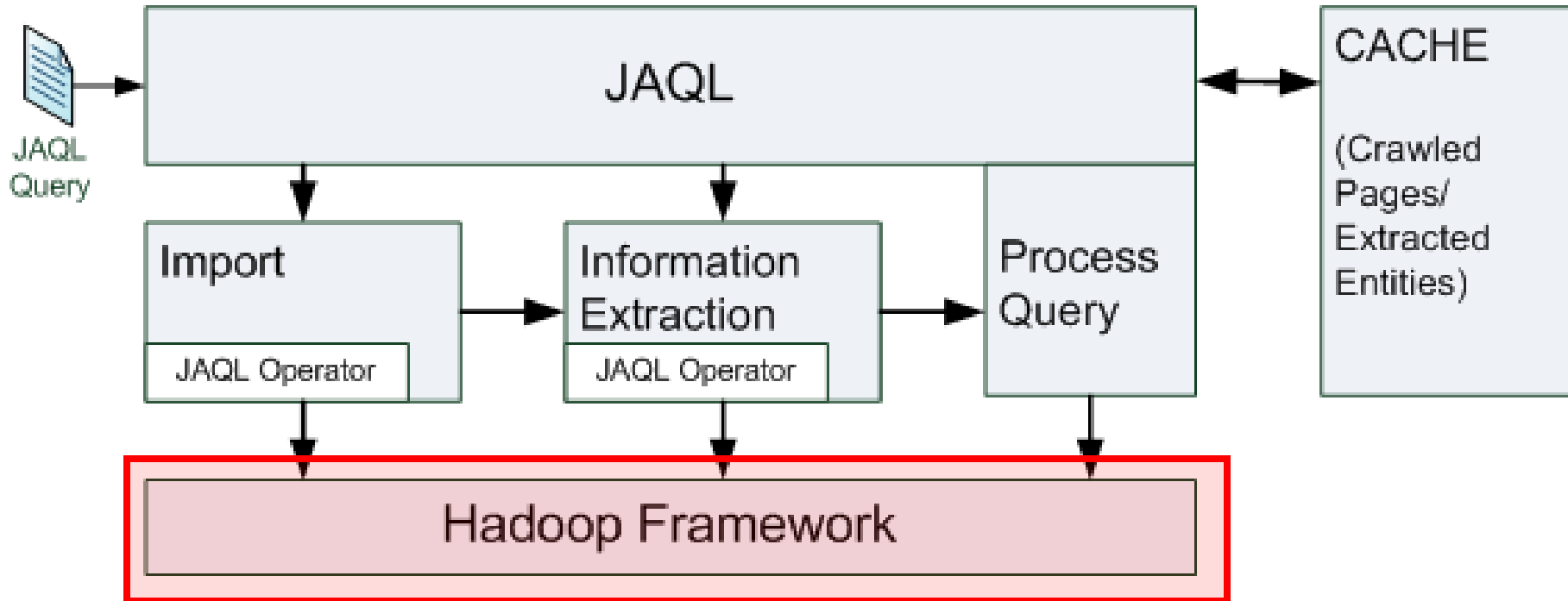
- ▶ Can we build a prototype to answer simple BI-Over-Text queries?
- ▶ What research problems can we derive?
- ▶ Only based on open source/free available software?
- ▶ How stable is existing cloud technology?

## ► 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







## ► What is Hadoop?

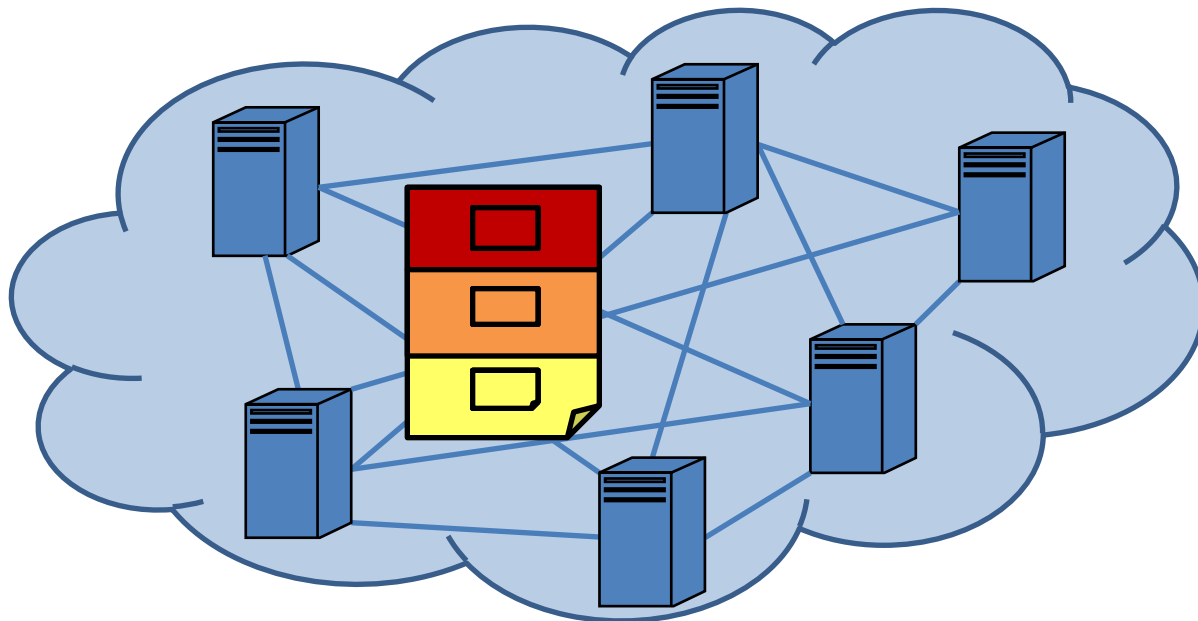
- Free software framework for data intensive applications
- Enables distributed processing of vast amounts of data on cloud computing architectures
- Supports clouds with 1000+ nodes
- Two components:
  - 1) Hadoop Distributed File System (HDFS)
  - 2) MapReduce Engine

## ► Where can you get Hadoop?

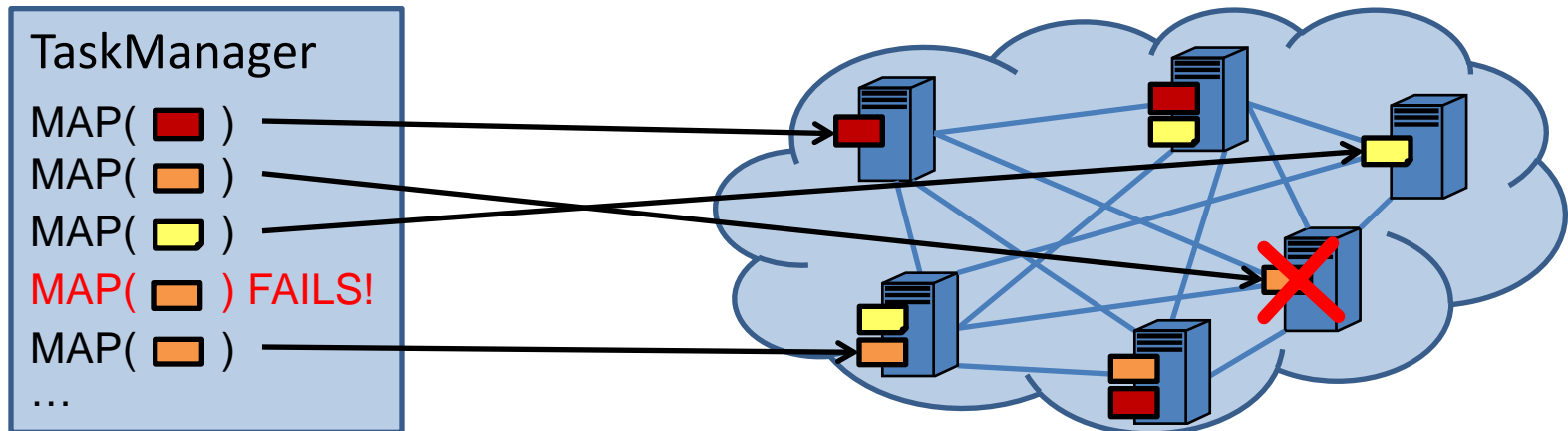
- Top-level Apache Project: <http://hadoop.apache.org/core/>



- ▶ 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)



- ▶ 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

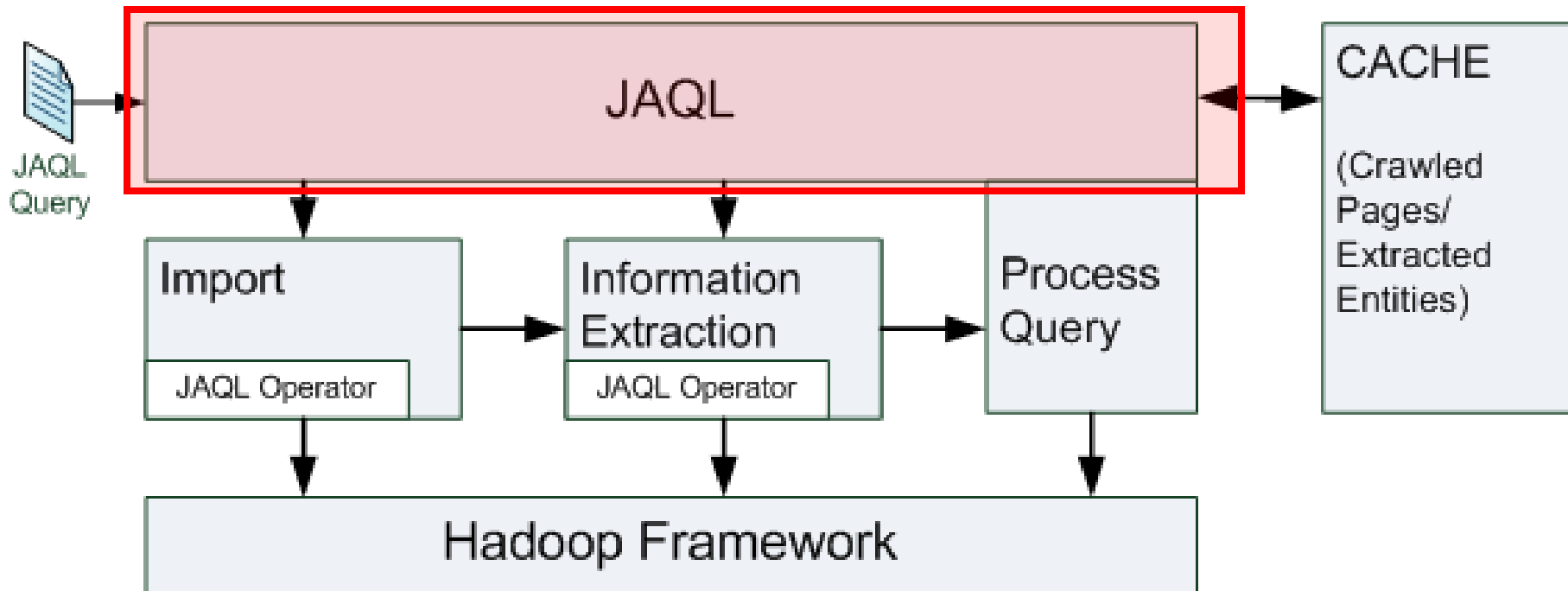


## ▶ Who uses Hadoop?

- 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“



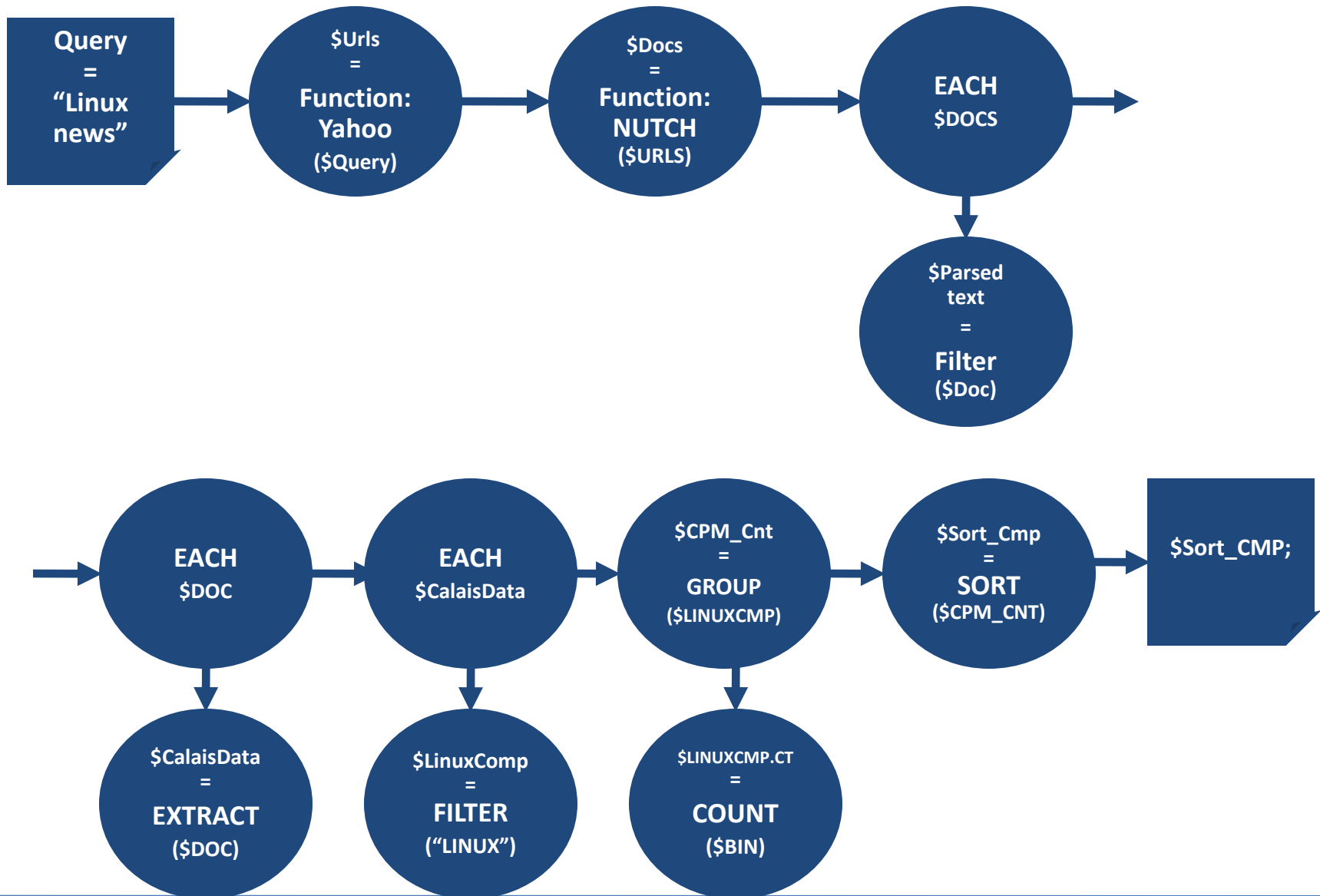
## IBM Query Processing Language for Hadoop

- ▶ 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



Source Selection

Data Crawling

Entity Extraction

Query Processing

```
register UDFsregisterFunction("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
$urls = yahooBoss(„ linux news",100);
// crawlURLs$
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

```
registerUDF(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);
```

```
// crawlURLs$
```

```
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

```
registerUDF(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
$urls = yahooBoss(„ linux news",100);
// crawlURLs$
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

```
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
$urls = yahooBoss(„ linux news",100);
// crawlURLs$
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

```
registerUDF(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
$urls = yahooBoss(„linux news“,100);
// crawlURLs$
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;
```

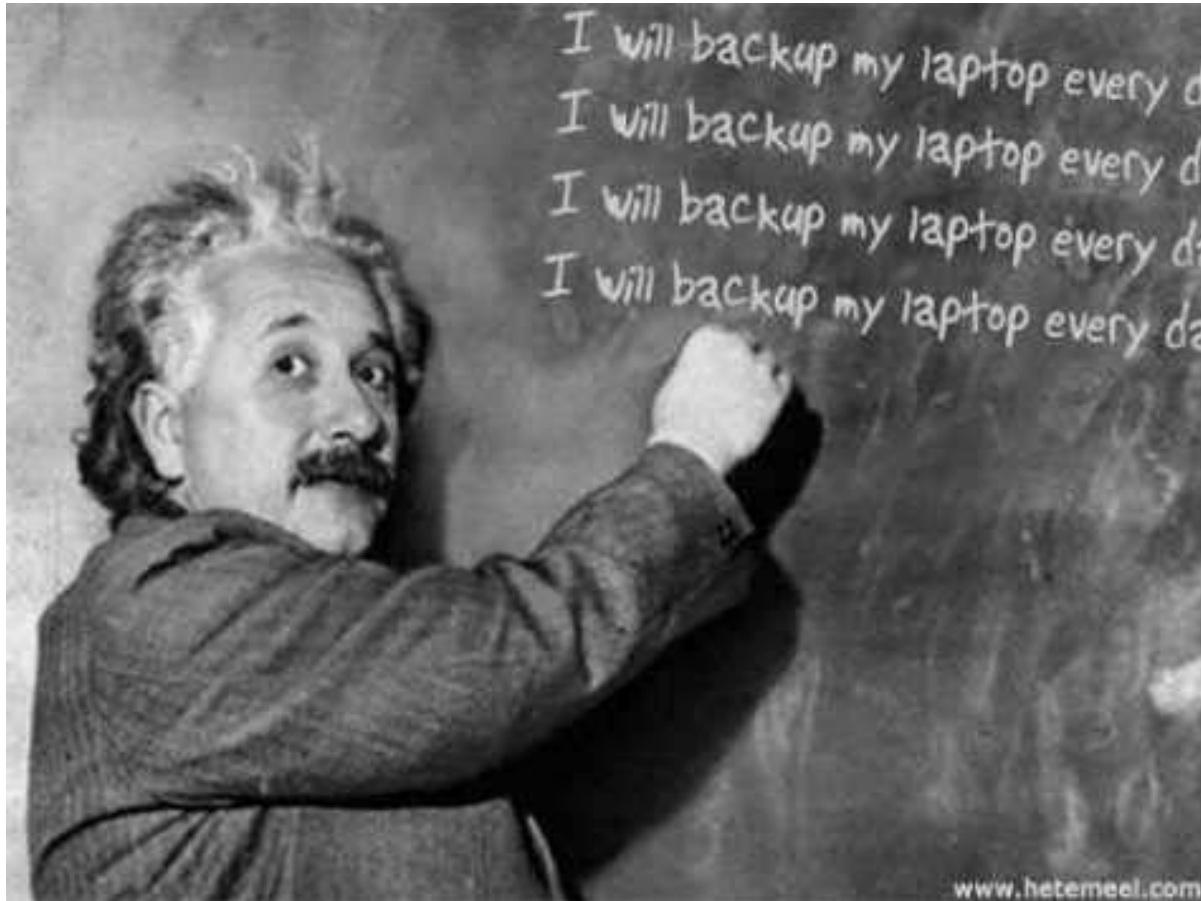
```
[
  {
    "url": "http://linxtoday.com/"
  },
  {
    "url": "http://www.linux.com/"
  },
  {
    "url": "http://www.linux.com/feature/c4201"
  },
  {
    "url": "http://lxxer.com/"
  },
  {
    "url": "http://www.linuxworld.com/"
  },
  {
    "url": "http://news.softpedia.com/cat/Linux/"
  },
  ...
]
```

```
[
{
  "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/"
        },
        ...
        {
          "Anchor": "Security?News",
          "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": {"ContentType": "text/html", "Source": "<html>\r\n<head>\r\n<title>Fresh Tech News - Linux News Headlines</title>\n<META HTTP-EQU"}
},
...
]
```

```
{
  "entityType": "CompanyTechnology",
  "id": "http://d.opencalais.com/genericHasher-1/c006792a-8d72-36d2-a685-19ec5c24828f",
  "instances": [
    {
      "length": 223,
      "offset": 6529,
      "text": "[month. Click to learn more. on the console.]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]"
    },
  ],
  "properties": {
    "company": "Oracle",
    "technology": "Linux"
  }
},
```

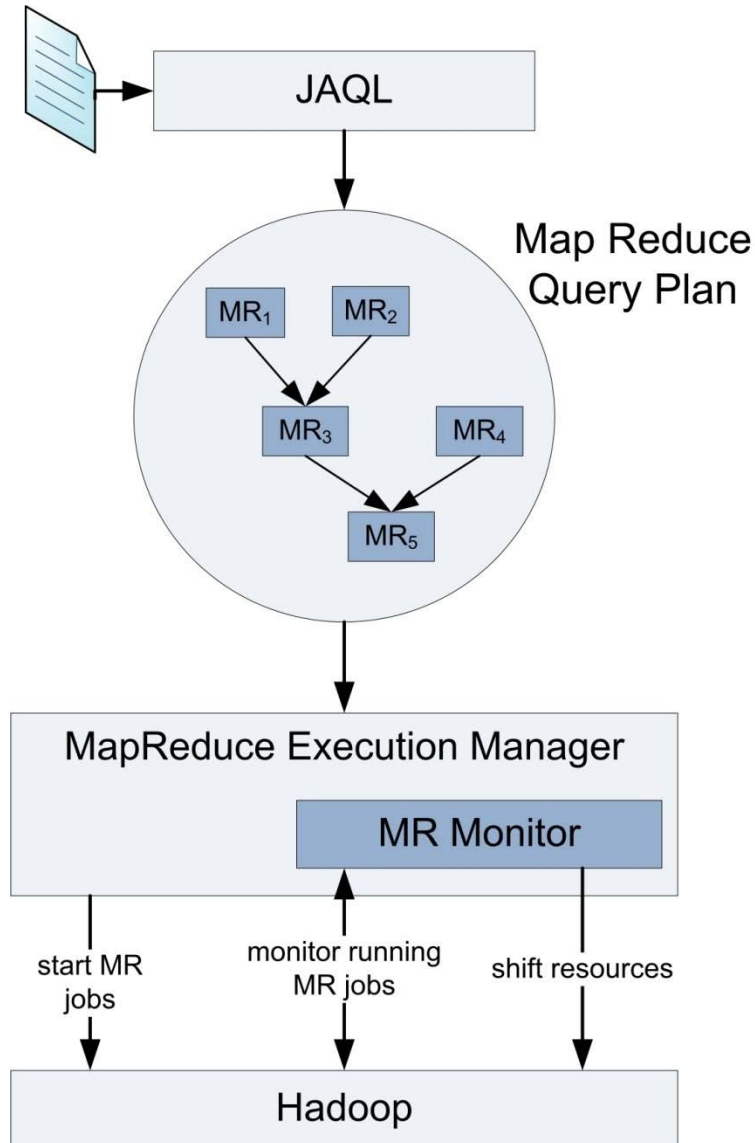
## {Occurrence , Company Name}

```
{ "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"}
```



- ▶ 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

- ▶ 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



## 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)**

Source Selection

Data Crawling

Entity extraction

Query Processing

Caching

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

- ▶ Why BI-Over-Text?
- ▶ BI-Over-Text using state-of-the art technology
- ▶ **Next steps**

## ► 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?

Go  lap  
Search. Analyse. Decide.

## Featured

- [Where are most companies located for "Search Engine Technology"?](#)
- [What is the average age of CTOs for Search Engine Technology?](#)
- [How many acquisitions were made by Google between 2004 and 2008?](#)
- [Who predicted the OIL price?](#)
- [Who is more popular: Putin or Clinton?](#)
- [How does the Dax correlate to positive and negative sentiments from analysts?](#)

Aggregating events, facts etc. about Persons, Companies, Organizations etc.

# What is next?: GOOLAP.INFO



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

▼ 1

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

▼ 2

"It's not right! ..."

Person travel: 2 results found

Name	Value
------	-------

▼ 1

**Datestring** on Sunday

**Status** past

**Traveldestination** [Los Angeles](#)

**Date** 2008-09-14

▶ 2

Family relation: 2 results found

Name	Value
------	-------

▼ 1

**Person\_relative** [Kathy Hilton](#)

**Familyrelationtype** parent

▼ 2

**Person\_relative** [Rick Hilton](#)

## Articles to "Paris Hilton"

1221687889818-85FDAB4B-2559220

Paris Hilton is out of jail, after serving three weeks in the same. She was imprisoned due to a violation of her probation 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: 10726)

1221687515445-1AB16E99-3705756

With photographers and media gathering for an appearance by Paris Hilton , who is sentenced to serve jail time, the fa 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: 10501)

1221687615122-746EE973-2444443

A day after being released to house arrest, Paris Hilton was brought back into court today in handcuffs, and ordered to sentence in jail, Hilton was escorted from the courtroom screaming and crying. 'It's not right...

The document was created on 2008-09-17 (DocID: 10550)



## 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 planed for 2009.

- ▶ Stephan Ewen
- ▶ Roland Hager
- ▶ Fabian Hueske
- ▶ Alexander Löser
- ▶ Volker Markl
- ▶ Bernd Rabe
- ▶ Ronny Schwierzinski

## 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

- ▶ Xin Dong, Alon Y. Halevy: Malleable Schemas: A Preliminary Report. *WebDB 2005*: 139-144
- ▶ Marcos Antonio Vaz Salles, Jens-Peter Dittrich, Shant Kirakos Karakashian, Olivier René Girard, Lukas Blunzsch: iTrails: Pay-as-you-go Information Integration in Dataspaces. *International Conference on Very Large Databases (VLDB) 2007*: 663-674

## Optimization in Text Data Bases

- ▶ 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
- ▶ Raghu Ramakrishnan and Andrew Tomkins: Towards a PeopleWeb. *IEEE Computer* 40(8): 63-72.
- ▶ Web 2.0 Business Analytics. Alexander Löser, Gregor Hackenbroich, Hong-Hai Do, Henrike Berthold. *Datenbank Spektrum* 25/2008
- ▶ T. S. Jayram, Andrew McGregor, S. Muthukrishnan, Erik Vee: Estimating Statistical Aggregates on Probabilistic Data Streams. *PODS 07*
- ▶ R-Cubes: OLAP Cubes Contextualized with Documents. Juan Manuel Perez et.al. *ICDE 2007*



## 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

Low  
↑  
Additional Effort Extraction/Cleansing  
↓  
High

## 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



Unstructured Information Management Architecture

*An Apache Incubator Project.*



YOOName  
Named Entity Recognition Software



dapper

lixto  
THE WEB INTELLIGENCE COMPANY

kapow  
TECHNOLOGIES



## 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.

# OpenCalais vrs. Other Extractors



Extractor	Precision	Recall
Balie	0,6374	0,6726
Calais	0,9395	0,7408
LingPipe	0,6418	0,5608
OpenNLP	0,7772	0,3844
Stanford	0,9645	0,8843

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