Your Connection to ICT Research

Le management d'une immense banque de données

Stéphane Mouton

SST Department manager

16ème rencontre EVELINE M ARKIEWICZ

Brussels, 14/11/2015

FEDER

UNION EUROPEENNE

Wallonie

LE FONDS EUROPEEN DE DEVELOPPEMENT REGIONAL
ET LA WALLONIE INVESTISSENT DANS VOTRE AVENIR.

Centre d'Excellence en **Technologies** de l'**Information** et de la **Communication**

www.cetic.be

In the beginning …

First compare quality. Then compare cost.

Morrow Designs' 10 megabyte hard disk system: $3,695.

MORE MEMORY. LESS MONEY.

Compare Morrow Designs' DISCUS™ M26™ hard disk systems to any system available for S-100 or Cromemco machines. First, compare features. Then, compare cost per megabyte. The M26 works out to under $200 a megabyte. And the M10 is about half the cost of competing systems.

COMPLETE SUBSYSTEMS.

Both the M10 (8"), and the M26 (14"), are delivered complete with disk controller, cables, fan, power supply, cabinet and CP/M® operating system. It's your choice: 10 Mb 8" at $3,695 or 26 Mb 14" at $4,995. That's single unit. Quantity prices are available.

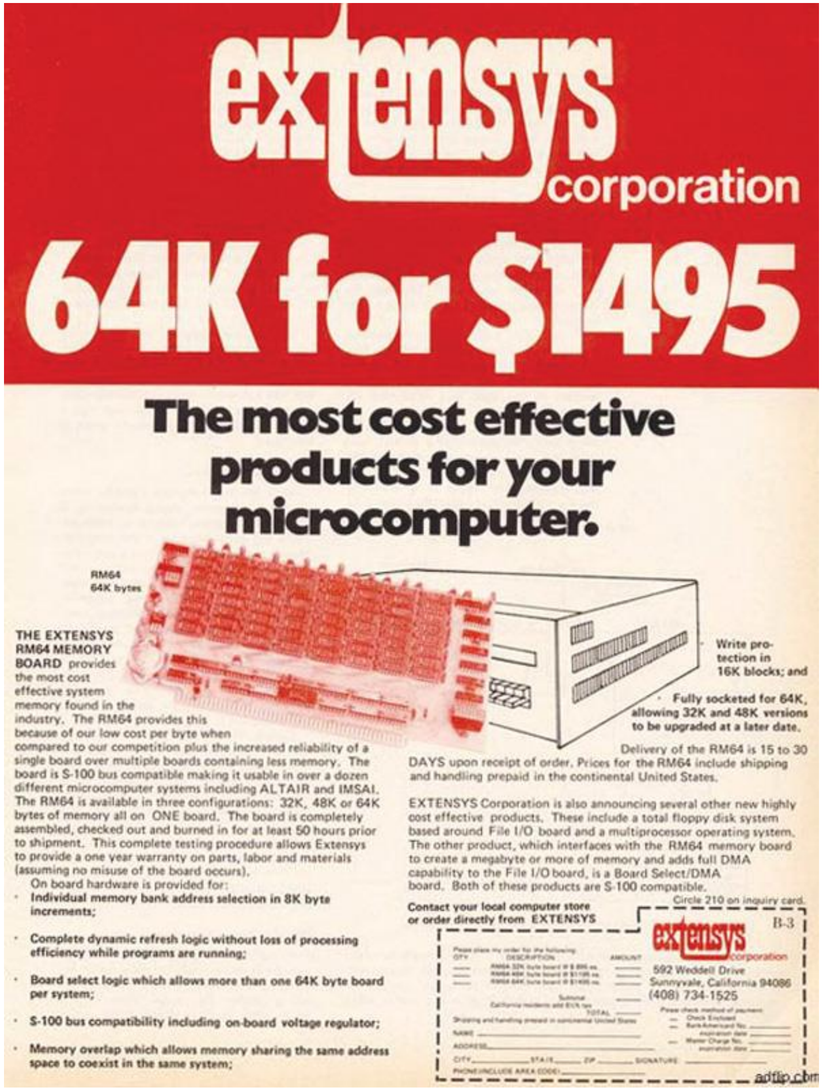
BUILD TO FOUR DRIVES.

104 Megabytes with the M26. 40+ megabytes with the M10. Formatted. Additional drives: M26: $4,495. M10: $3,195. Quantity discounts available.

Sausalito, CA, (415) 332-4443, offers a CP/M expanded to full Cromemco CDOS compatibility.

AND NOW, MULT-I/O.™

Mult-I/O is an I/O controller that allows multi-terminal and multi-purpose use of S-100 and Cromemco computers.



extensys corporation

64K for $1495

The most cost effective products for your microcomputer.

RM64
64K bytes

THE EXTENSYS RM64 MEMORY BOARD provides the most cost effective system memory found in the industry. The RM64 provides this because of our low cost per byte when compared to our competition plus the increased reliability of a single board over multiple boards containing less memory. The board is S-100 bus compatible making it usable in over a dozen different microcomputer systems including ALTAIR and IMSAI. The RM64 is available in three configurations: 32K, 48K or 64K bytes of memory all on ONE board. The board is completely assembled, checked out and burned in for at least 50 hours prior to shipment. This complete testing procedure allows Extensys to provide a one year warranty on parts, labor and materials (assuming no misuse of the board occurs).

On board hardware is provided for:

- **Individual memory bank address selection in 8K byte increments;**
- **Complete dynamic refresh logic without loss of processing efficiency while programs are running;**
- **Board select logic which allows more than one 64K byte board per system;**
- **S-100 bus compatibility including on-board voltage regulator;**
- **Memory overlap which allows memory sharing the same address space to coexist in the same system;**
- **Write protection in 16K blocks; and**
- **Fully socketed for 64K, allowing 32K and 48K versions to be upgraded at a later date.**

Delivery of the RM64 is 15 to 30 DAYS upon receipt of order. Prices for the RM64 include shipping and handling prepaid in the continental United States.

EXTENSYS Corporation is also announcing several other new highly cost effective products. These include a total floppy disk system based around File I/O board and a multiprocessor operating system. The other product, which interfaces with the RM64 memory board to create a megabyte or more of memory and adds full DMA capability to the File I/O board, is a Board Select/DMA board. Both of these products are S-100 compatible.

Circle 210 on inquiry card.

Contact your local computer store or order directly from EXTENSYS

B-3

Please place my order for the following:

QTY	DESCRIPTION	AMOUNT
____	RM64 32K byte board @ $ 895 ea.	____
____	RM64 48K byte board @ $1195 ea.	____
____	RM64 64K byte board @ $1495 ea.	____
	Subtotal	____
	California residents add 6½% tax	
	TOTAL	____

Shipping and handling prepaid in continental United States

NAME ____________________

ADDRESS ____________________

CITY __________ STATE ______ ZIP ______ SIGNATURE ____________

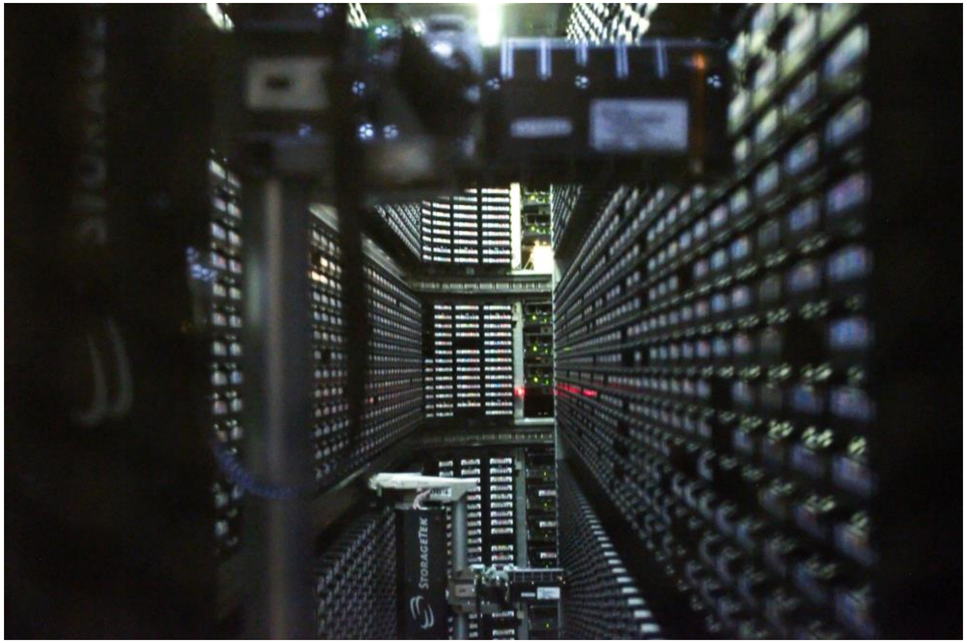
PHONE (INCLUDE AREA CODE) ____________

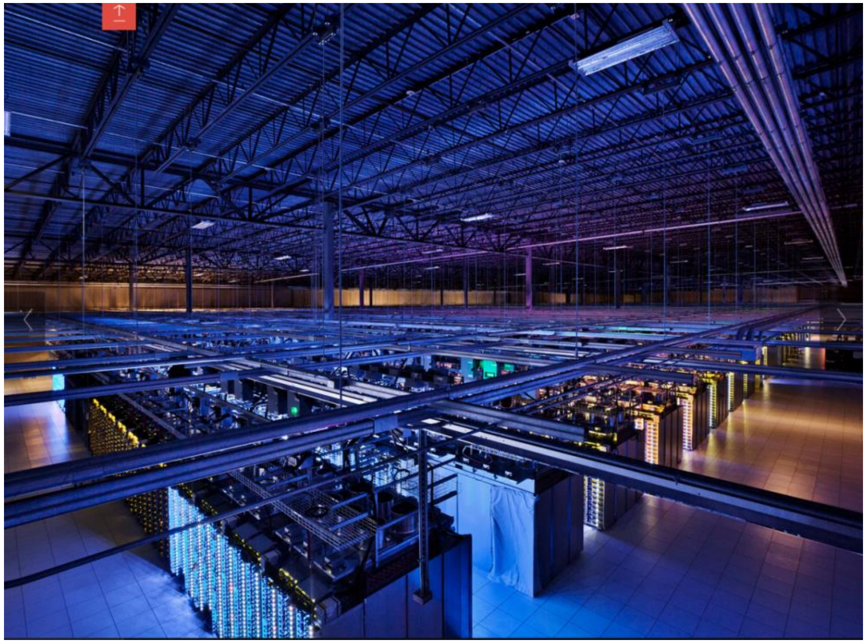
extensys corporation
592 Weddell Drive
Sunnyvale, California 94086
(408) 734-1525

Please check method of payment
— Check Enclosed
— BankAmericard No. ________
expiration date ________
— Master Charge No. ________
expiration date ________

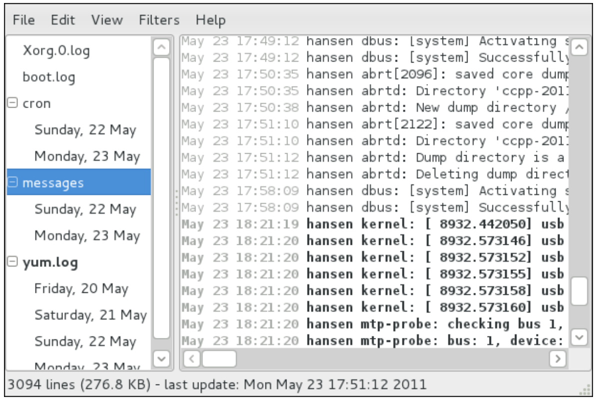
adflip.com

Then IT resource became affordable





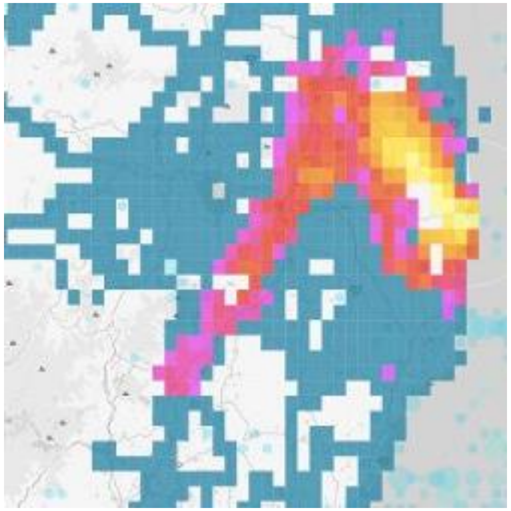
But data production explodes

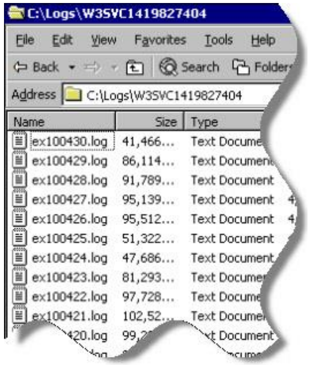


```
Xorg.0.log
boot.log
⊟ cron
    Sunday, 22 May
    Monday, 23 May
⊟ messages
    Sunday, 22 May
    Monday, 23 May
⊟ yum.log
    Friday, 20 May
    Saturday, 21 May
    Sunday, 22 May
    Monday, 23 May
```

```
May 23 17:49:12 hansen dbus: [system] Activating s
May 23 17:49:12 hansen dbus: [system] Successfully
May 23 17:50:35 hansen abrt[2096]: saved core dump
May 23 17:50:35 hansen abrtd: Directory 'ccpp-201
May 23 17:50:38 hansen abrtd: New dump directory /
May 23 17:51:10 hansen abrt[2122]: saved core dump
May 23 17:51:10 hansen abrtd: Directory 'ccpp-201
May 23 17:51:12 hansen abrtd: Dump directory is a
May 23 17:51:12 hansen abrtd: Deleting dump direct
May 23 17:58:09 hansen dbus: [system] Activating s
May 23 17:58:09 hansen dbus: [system] Successfully
May 23 18:21:19 hansen kernel: [ 8932.442050] usb
May 23 18:21:20 hansen kernel: [ 8932.573146] usb
May 23 18:21:20 hansen kernel: [ 8932.573152] usb
May 23 18:21:20 hansen kernel: [ 8932.573155] usb
May 23 18:21:20 hansen kernel: [ 8932.573158] usb
May 23 18:21:20 hansen kernel: [ 8932.573160] usb
May 23 18:21:20 hansen mtp-probe: checking bus 1,
May 23 18:21:20 hansen mtp-probe: bus: 1, device:
```

3094 lines (276.8 KB) - last update: Mon May 23 17:51:12 2011

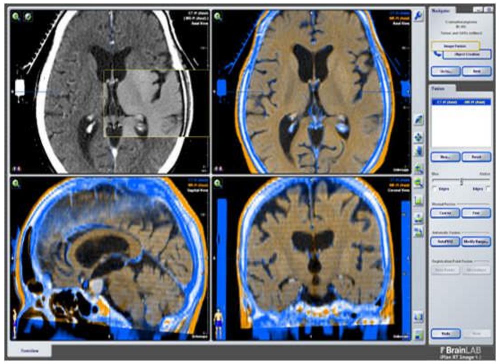


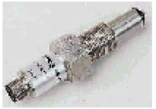


C:\Logs\W3SVC1419827404

Address C:\Logs\W3SVC1419827404

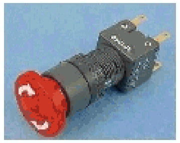
Name	Size	Type
ex100430.log	41,466...	Text Docume
ex100429.log	86,114...	Text Document
ex100428.log	91,789...	Text Document
ex100427.log	95,139...	Text Document
ex100426.log	95,512...	Text Document
ex100425.log	51,322...	Text Document
ex100424.log	47,686...	Text Document
ex100423.log	81,293...	Text Docume
ex100422.log	97,728...	Text Docume
ex100421.log	102,52...	Text Documen
420.log	99,2	xt Document





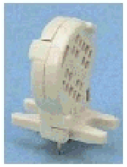
Capteur de niveau de liquide

Bouton poussoir

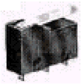


Bouton d'arrêt d'urgence

Détecteur de choc



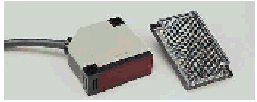
Capteur d'humidité



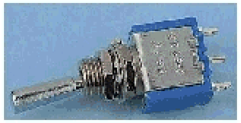
Capteur de fin de course

Capteur de proximité à ultrasons

Détecteur de gaz



Cellule photoélectrique



Interrupteur miniature

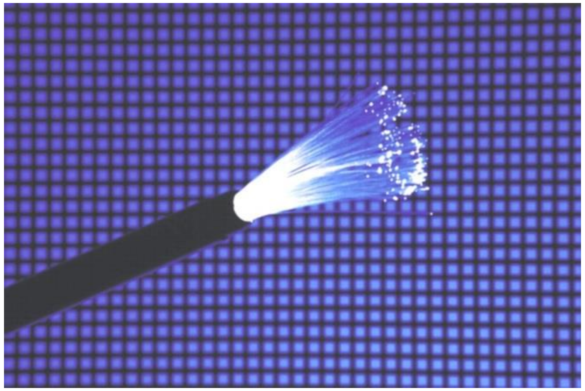
What is the status now?

- Network lags behind

“Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway.”

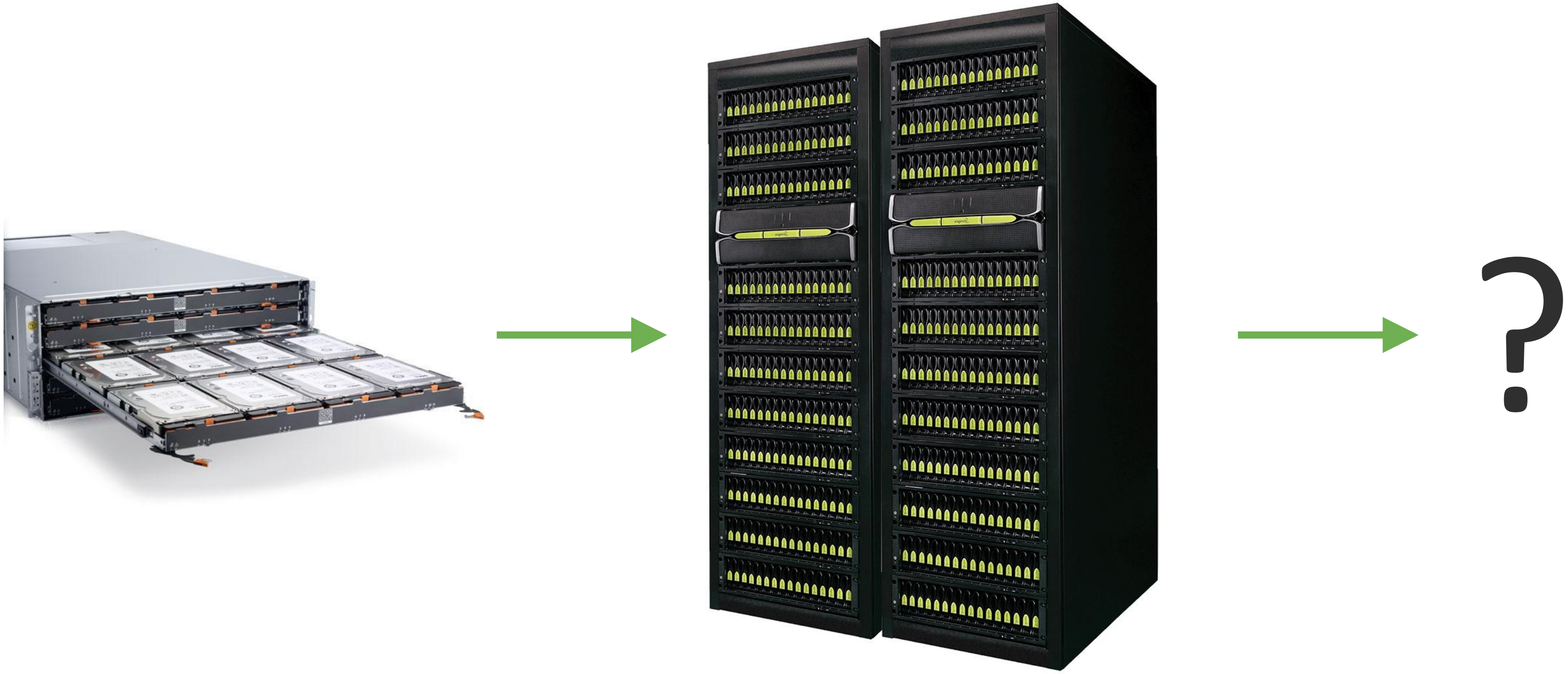
Tanenbaum, Andrew S. (1989). Computer Networks

Fiber network is crucial



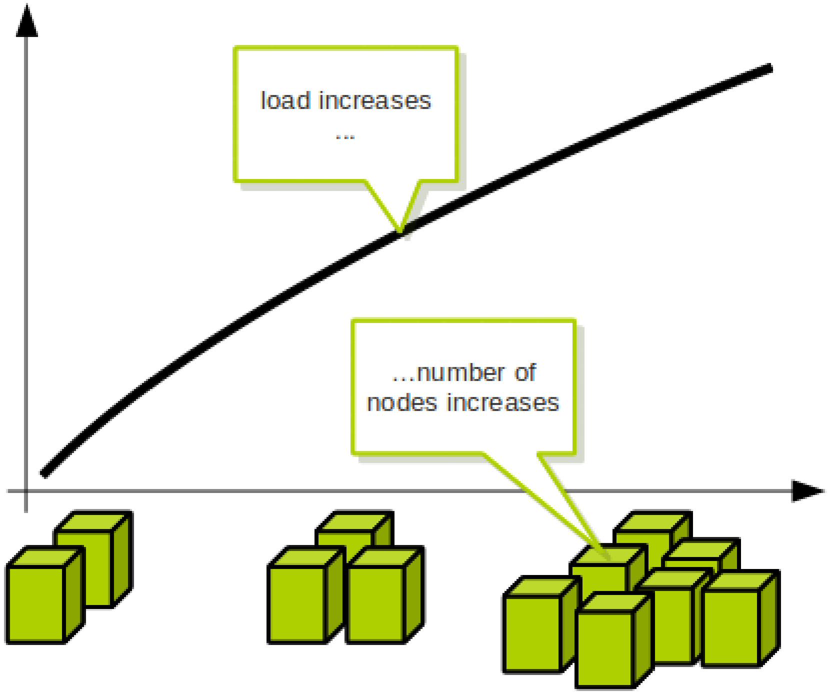
What is the status now?

- Most of “usual” commercial offering don’t stack up seamlessly



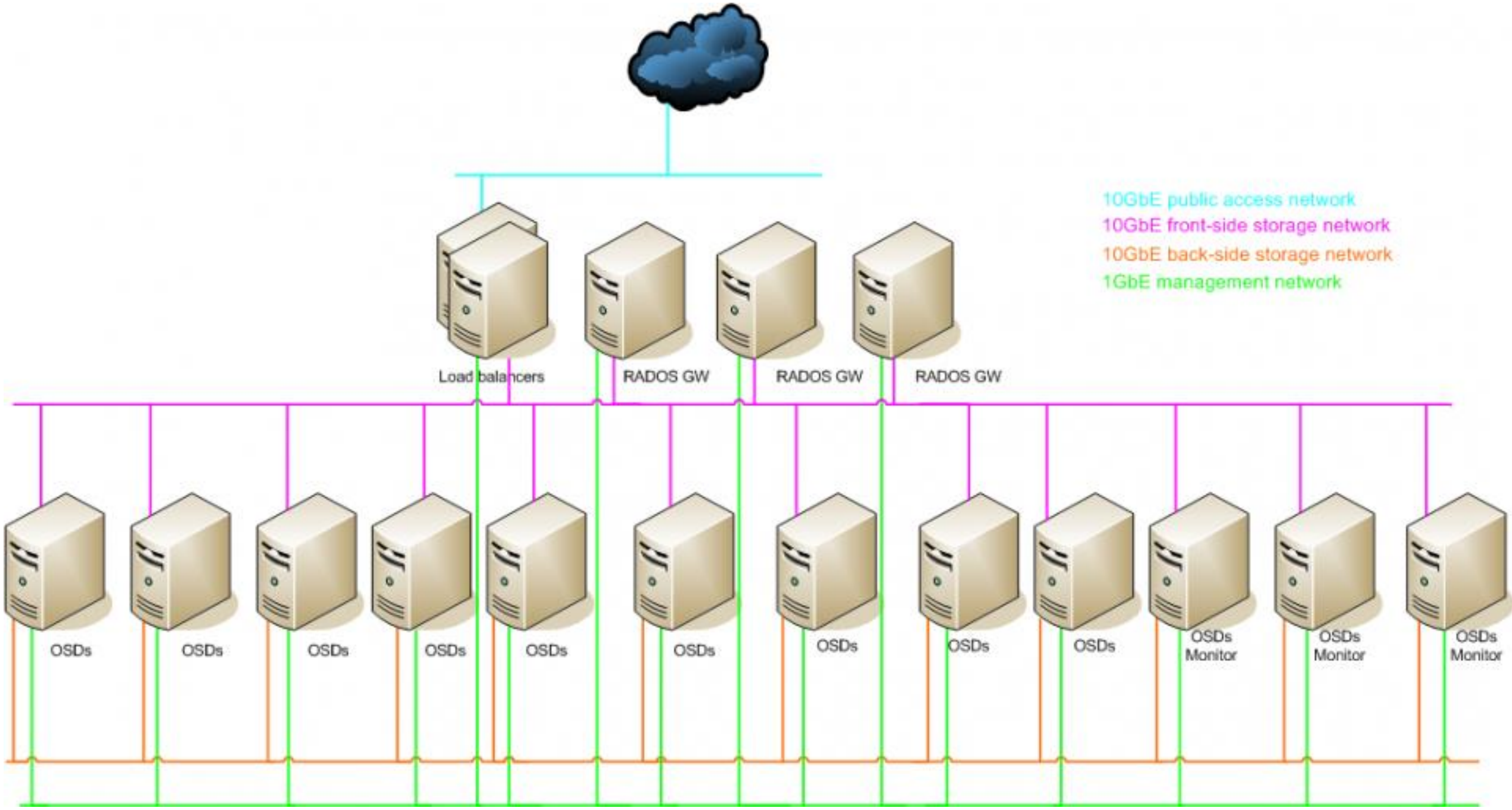
Need for scalability

- The capability of IT systems to adapt to system demand at a predictable, ideally linear, pace.



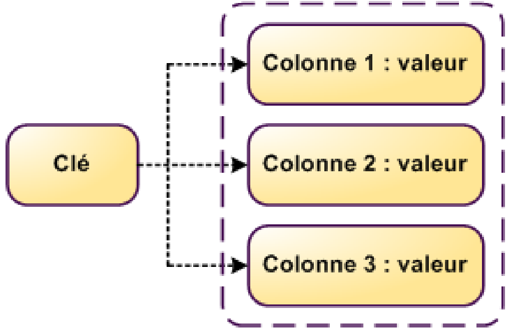
Scalable storage to cope with volume

- When a single IT resource can’t handle all the data anymore
  - *Distributed* solutions, based on many cloned (and possibly inexpensive) *nodes*

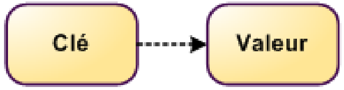


Scalable databases

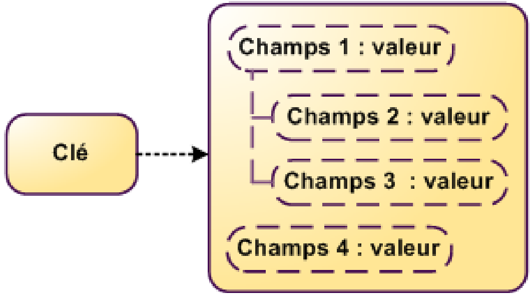
- New ways to structure data storage
  - NoSQL (Not Only SQL databases)



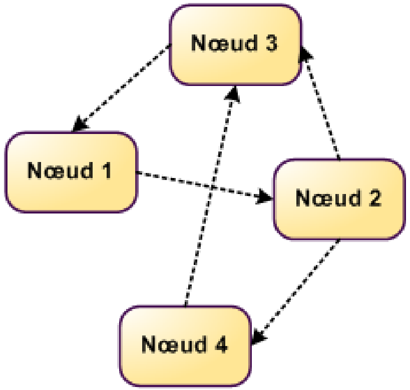
BDD Orientée colonnes



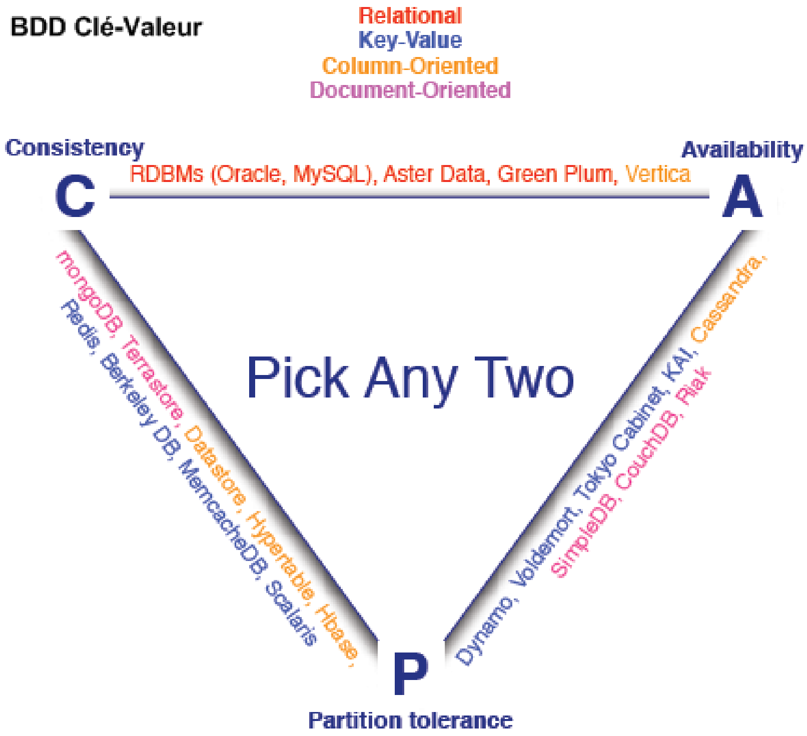
BDD Clé-Valeur



BDD Orientée document

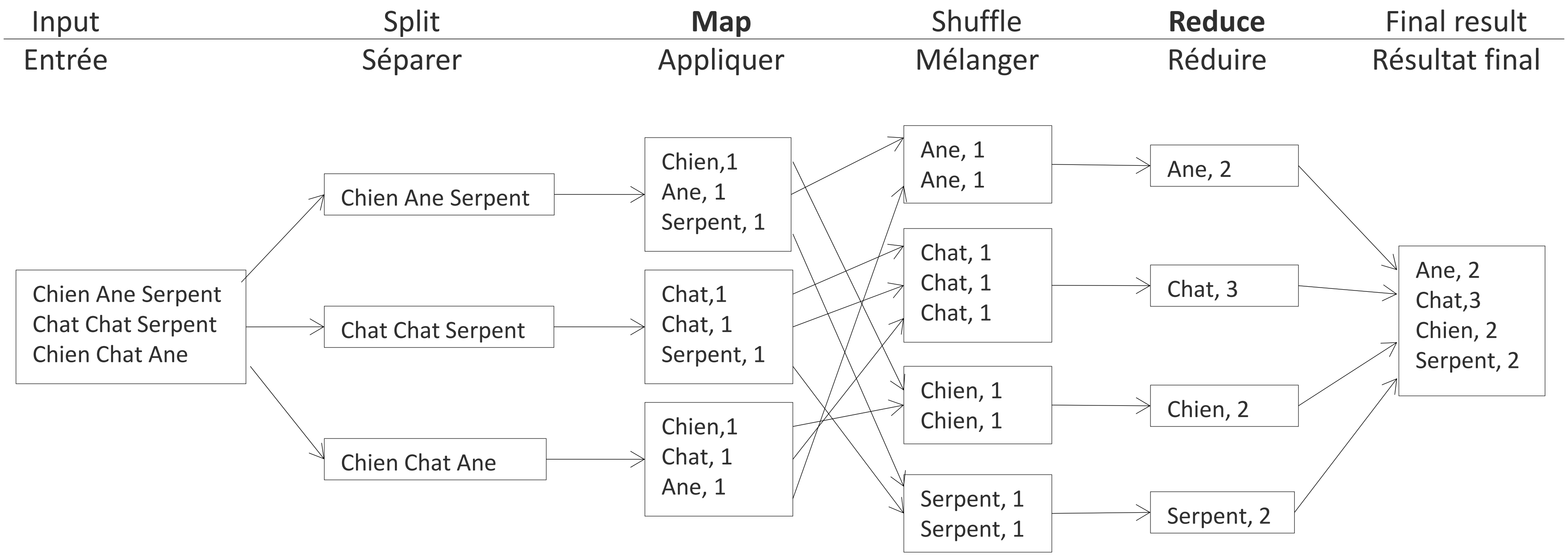


BDD Orientée graphe


Relational
Key-Value
Column-Oriented
Document-Oriented
Consistency
Availability
RDBMs (Oracle, MySQL), Aster Data, Green Plum, Vertica
C
A
mongoDB, Terrastore, Datastore, Hypertable, Hbase,
Redis, Berkeley DB, MemcacheDB, Scalaris
Pick Any Two
Dynamo, Voldemort, Tokyo Cabinet, KAI, Cassandra,
SimpleDB, CouchDB, Riak
P
Partition tolerance

Process and analyse data in a scalable way

- What if you have possibly an infinite number of computers to process your data?
  - Your data analysis doesn't change, the way you express it does.



Input Entrée

Split Séparer

Map Appliquer

Shuffle Mélanger

Reduce Réduire

Final result Résultat final

Chien Ane Serpent
Chat Chat Serpent
Chien Chat Ane

Chien Ane Serpent

Chat Chat Serpent

Chien Chat Ane

Chien,1
Ane, 1
Serpent, 1

Chat,1
Chat, 1
Serpent, 1

Chien,1
Chat, 1
Ane, 1

Ane, 1
Ane, 1

Chat, 1
Chat, 1
Chat, 1

Chien, 1
Chien, 1

Serpent, 1
Serpent, 1

Ane, 2

Chat, 3

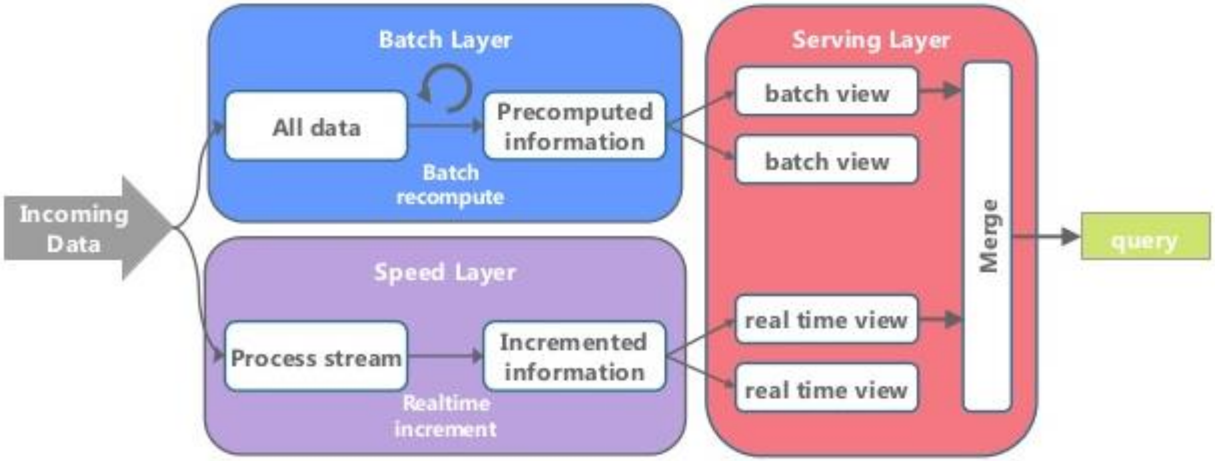
Chien, 2

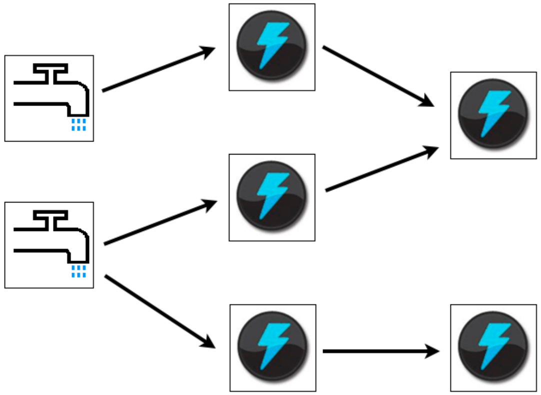
Serpent, 2

Ane, 2
Chat,3
Chien, 2
Serpent, 2

Process and analyse data in a scalable way

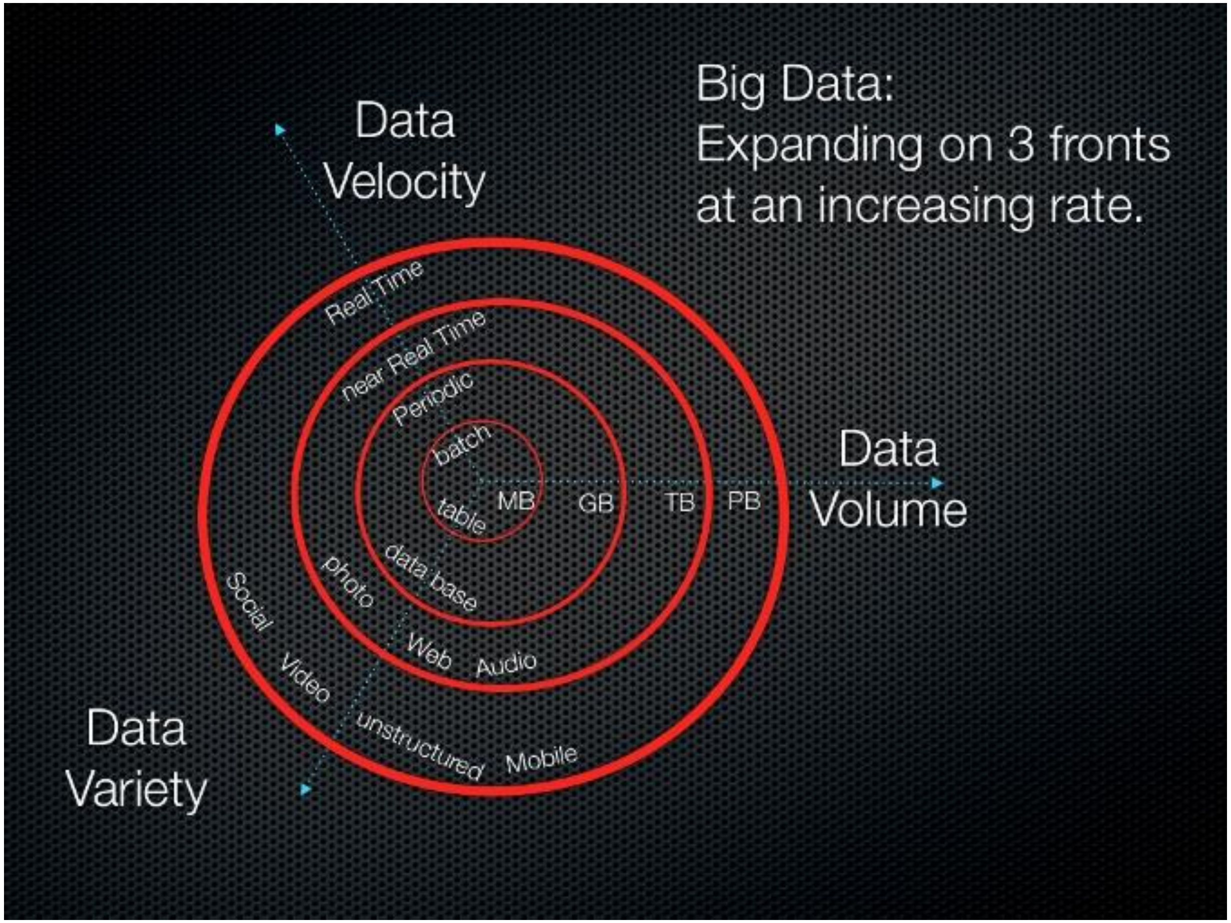
- New programming frameworks for data analysis





Big Data

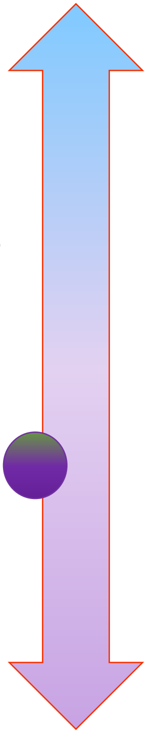
- A relative concept, linked to data
  - Volume
  - Variety
  - Velocity



Big Data:
Expanding on 3 fronts
at an increasing rate.

Volume: data integration approaches

- 3 main models ensuring data bases autonomy
  - Full Data Integration (Data Warehousing)
    - Not suitable for data with a very short-term relevance
    - Can lead to improperly formed data warehouses, not focused on the analysis topic
  - Federated databases
    - One global schema
    - One main data format/model
    - Mostly static
  - Mediator/Wrapper
    - Support dynamic sets of data sources
    - High flexibility: schema can evolve at query time
    - Wrappers to handle heterogeneous data format



Tightly coupled

- Administrator builds a unique FDB schema

Loosely coupled

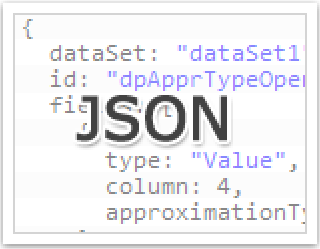
- Component DBs out of control
- Import schemas as view over data sources

Data integration leads to data variety

- Variety in content type

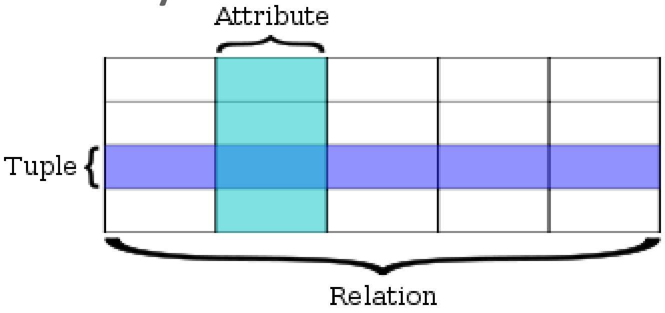
Data integration leads to data variety

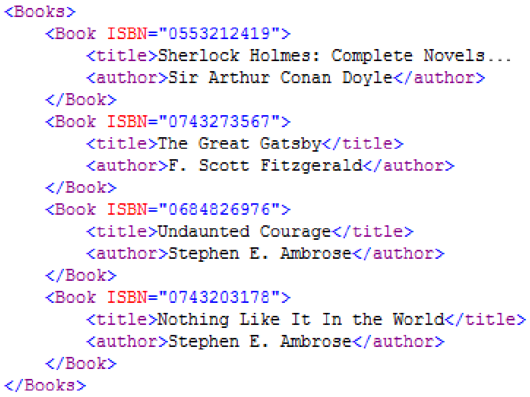
- Variety in data format



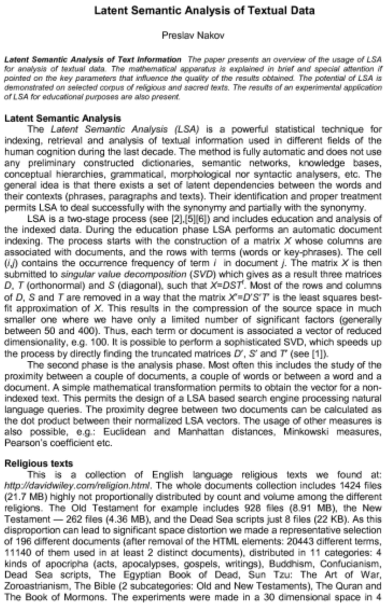
Data integration leads to data variety

- Variety in data model





```
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```



Latent Semantic Analysis of Textual Data

Preslav Nakov

Latent Semantic Analysis of Text Information *The paper presents an overview of the usage of LSA for analysis of textual data. The mathematical apparatus is explained in brief and special attention if pointed on the key parameters that influence the quality of the results obtained. The potential of LSA is demonstrated on selected corpus of religious and sacred texts. The results of an experimental application of LSA for educational purposes are also present.*

Latent Semantic Analysis

The *Latent Semantic Analysis (LSA)* is a powerful statistical technique for indexing, retrieval and analysis of textual information used in different fields of the human cognition during the last decade. The method is fully automatic and does not use any preliminary constructed dictionaries, semantic networks, knowledge bases, conceptual hierarchies, grammatical, morphological nor syntactic analysers, etc. The general idea is that there exists a set of latent dependencies between the words and their contexts (phrases, paragraphs and texts). Their identification and proper treatment permits LSA to deal successfully with the synonymy and partially with the synonymy.

LSA is a two-stage process (see [2],[5][6]) and includes education and analysis of the indexed data. During the education phase LSA performs an automatic document indexing. The process starts with the construction of a matrix X whose columns are associated with documents, and the rows with terms (words or key-phrases). The cell (i,j) contains the occurrence frequency of term i in document j. The matrix X is then submitted to *singular value decomposition* (*SVD*) which gives as a result three matrices D, T (orthonormal) and S (diagonal), such that $X=DST^t$. Most of the rows and columns of D, S and T are removed in a way that the matrix $X'=D'S'T'$ is the least squares best-fit approximation of X. This results in the compression of the source space in much smaller one where we have only a limited number of significant factors (generally between 50 and 400). Thus, each term or document is associated a vector of reduced dimensionality, e.g. 100. It is possible to perform a sophisticated SVD, which speeds up the process by directly finding the truncated matrices D', S' and T' (see [1]).

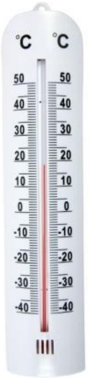
The second phase is the analysis phase. Most often this includes the study of the proximity between a couple of documents, a couple of words or between a word and a document. A simple mathematical transformation permits to obtain the vector for a non-indexed text. This permits the design of a LSA based search engine processing natural language queries. The proximity degree between two documents can be calculated as the dot product between their normalized LSA vectors. The usage of other measures is also possible, e.g.: Euclidean and Manhattan distances, Minkowski measures, Pearson's coefficient etc.

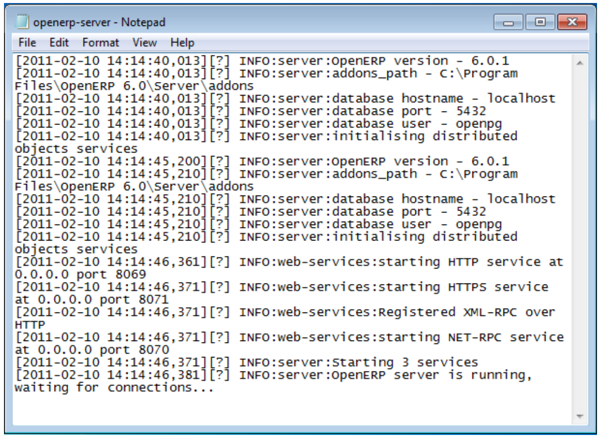
Religious texts

This is a collection of English language religious texts we found at: *http://davidwiley.com/religion.html*. The whole documents collection includes 1424 files (21.7 MB) highly not proportionally distributed by count and volume among the different religions. The Old Testament for example includes 928 files (8.91 MB), the New Testament — 262 files (4.36 MB), and the Dead Sea scripts just 8 files (22 KB). As this disproportion can lead to significant space distortion we made a representative selection of 196 different documents (after removal of the HTML elements: 20443 different terms, 11140 of them used in at least 2 distinct documents), distributed in 11 categories: 4 kinds of apocripha (acts, apocalypses, gospels, writings), Buddhism, Confucianism, Dead Sea scripts, The Egyptian Book of Dead, Sun Tzu: The Art of War, Zoroastrianism, The Bible (2 subcategories: Old and New Testaments), The Quran and The Book of Mormons. The experiments were made in a 30 dimensional space in 4

Data integration leads to data variety

- Variety in data source



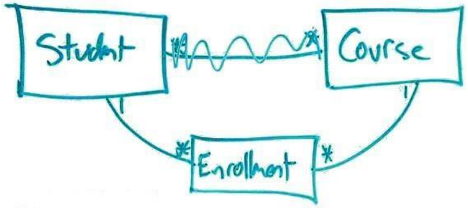


```
[2011-02-10 14:14:40,013][?] INFO:server:OpenERP version - 6.0.1
[2011-02-10 14:14:40,013][?] INFO:server:addons_path - C:\Program
Files\OpenERP 6.0\Server\addons
[2011-02-10 14:14:40,013][?] INFO:server:database hostname - localhost
[2011-02-10 14:14:40,013][?] INFO:server:database port - 5432
[2011-02-10 14:14:40,013][?] INFO:server:database user - openpg
[2011-02-10 14:14:40,013][?] INFO:server:initialising distributed
objects services
[2011-02-10 14:14:45,200][?] INFO:server:OpenERP version - 6.0.1
[2011-02-10 14:14:45,210][?] INFO:server:addons_path - C:\Program
Files\OpenERP 6.0\Server\addons
[2011-02-10 14:14:45,210][?] INFO:server:database hostname - localhost
[2011-02-10 14:14:45,210][?] INFO:server:database port - 5432
[2011-02-10 14:14:45,210][?] INFO:server:database user - openpg
[2011-02-10 14:14:45,210][?] INFO:server:initialising distributed
objects services
[2011-02-10 14:14:46,361][?] INFO:web-services:starting HTTP service at
0.0.0.0 port 8069
[2011-02-10 14:14:46,371][?] INFO:web-services:starting HTTPS service
at 0.0.0.0 port 8071
[2011-02-10 14:14:46,371][?] INFO:web-services:Registered XML-RPC over
HTTP
[2011-02-10 14:14:46,371][?] INFO:web-services:starting NET-RPC service
at 0.0.0.0 port 8070
[2011-02-10 14:14:46,371][?] INFO:server:Starting 3 services
[2011-02-10 14:14:46,381][?] INFO:server:OpenERP server is running,
waiting for connections...
```

Challenges of data variety

- Identify relevant data sources

- Extract, clean and store data

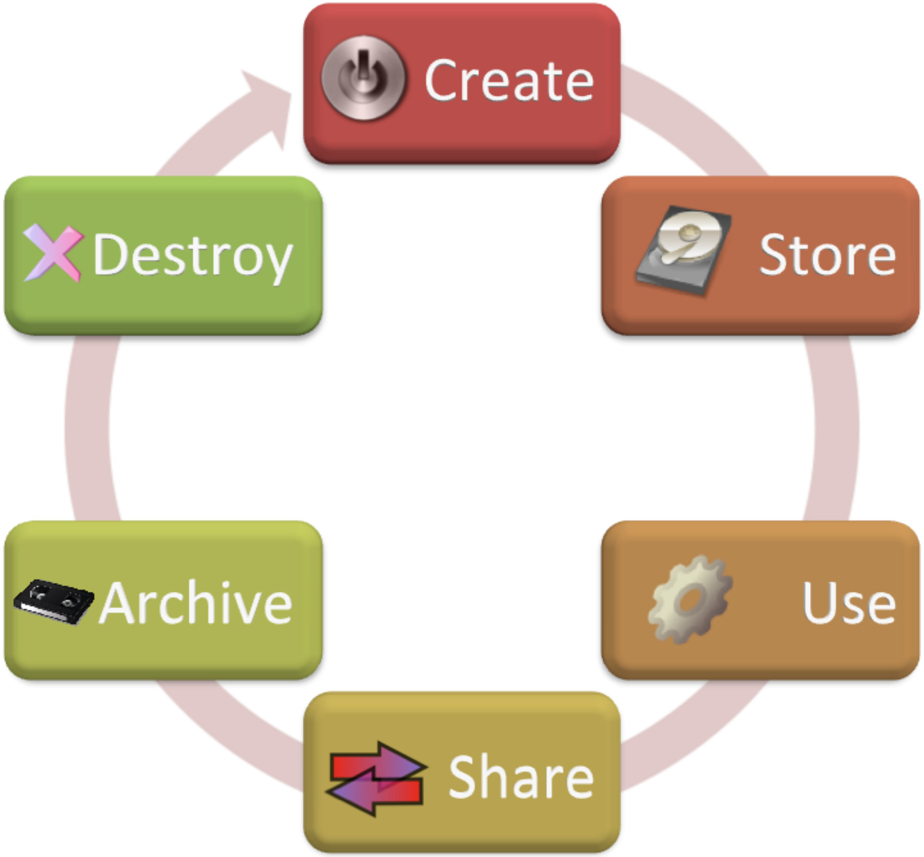


- Understand the data (semantics, meta-data)

- Deliver information

Data management also applies

- Not only in terms of cost



cetic

Your Connection to ICT Research

Aéropole de Charleroi-Gosselies
Bâtiment Éole
Rue des Frères Wright, 29/3
B-6041 Charleroi

Tel: +32.71.490.700
Fax: +32.71.490.799

www.cetic.be
info@cetic.be

linkedin.com/company/cetic

twitter.com/@CETIC

Questions?

Stéphane Mouton
SST Department manager

Tel : +32 71 490 726

Mob : +32 475 76 78 50

stephane.mouton@cetic.be