ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

Laboratoire ERIC

Équipe d'Accueil 3083

Self-Service Business Intelligence

Jérôme Darmont

eric.univ-lyon2.fr/~jdarmont/

IT4BI

Barcelona, May 2017

Lyon 1 · Université Lumière Lyon 2 · Université de Lyon · ISH

1

BIaaS · Security · Performance · BI4people

Outline

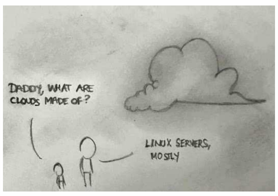
- Business intelligence as a service
- Data security in the cloud
- Data access performance
- BI4people research project

Self-service BI – J. Darmont 2017 · 2

ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

Part #1

Business intelligence as a service



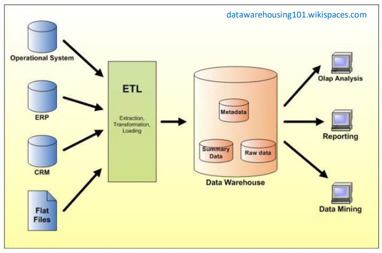
i.redd.it

Lyon 1 · Université Lumière Lyon 2 · Université de Lyon · ISH

3

BIaaS · Security · Performance · BI4people

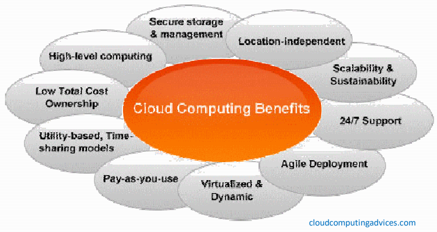
Classical data warehousing process



Self-service BI – J. Darmont 2017 · 4

BIaaS · Security · Performance · BI4people

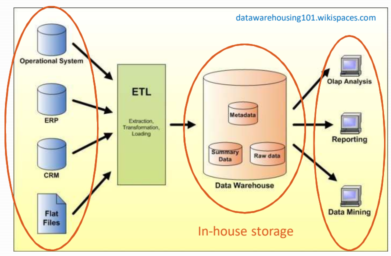
What a wonderful world



Self-service BI – J. Darmont 2017 · 5

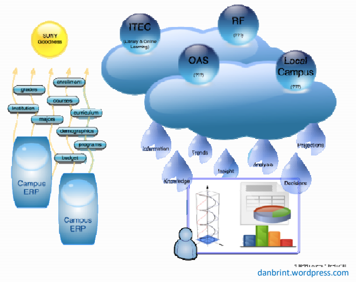
BIaaS · Security · Performance · BI4people

From grandpa's data warehouses...



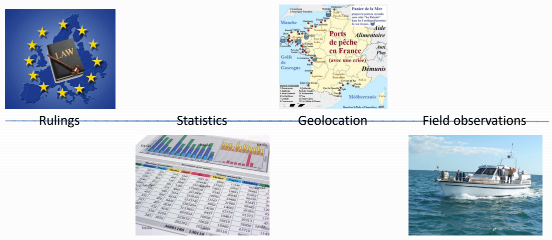
Self-service BI – J. Darmont 2017 · 6

...to cloud BI!

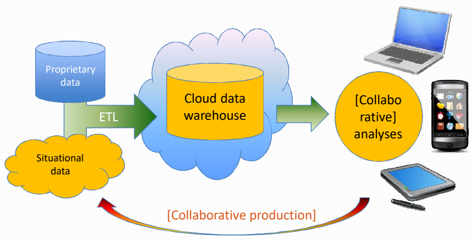


Watchdog scenario

NGO watching fishing activities

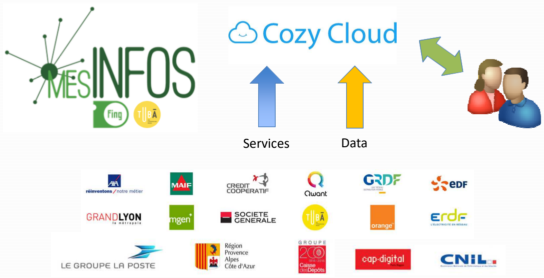


Toward self-service BI

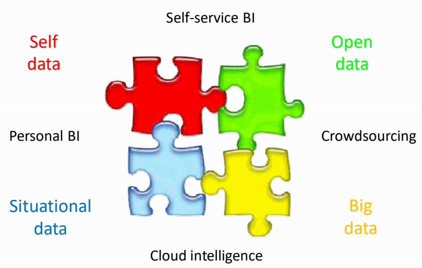


Elasticity + pay-as-you-go ⇒ gradual investment

Self data scenario



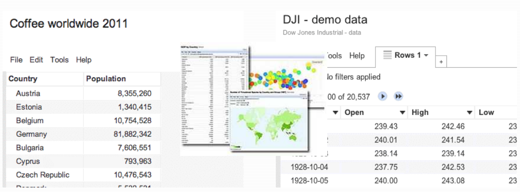
New data, new usages



New problems

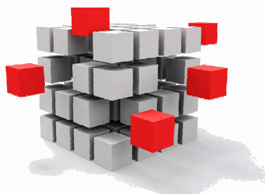
Let's play!

Google Fusion tables



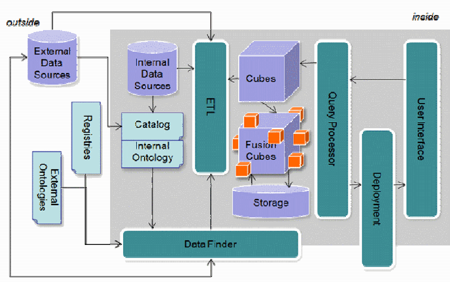
OLAP is spreadsheet on steroids

A. Abello, J. Darmont, L. Etcheverry, M. Golfarelli, J.N. Mazon, F. Naumann, **T.B. Pedersen**, S. Rizzi, J. Trujillo, P. Vassiliadis, G. Vossen, "Fusion Cubes: Towards Self-Service Business Intelligence", *International Journal of Data Warehousing and Mining*, Vol. 9, No. 2, 2013, 66-88

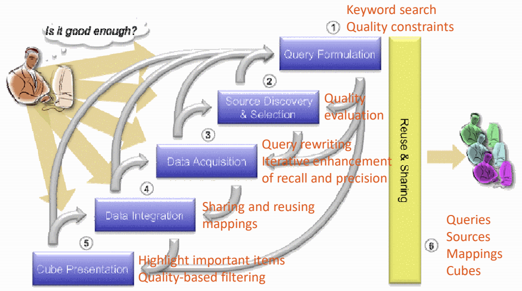


- Extensible schema
- Extensible instances
- Drill beyond (situational data)
- Near real-time

Services' architecture



Functional architecture



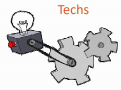
BI in your pocket

Non-expert

user

- On-demand analyses
- Search engine-like querying
- Drill-beyond
- Near real-time
- Collaborative

Techs



- Data warehouses/ OLAP
- Semantic Web
- Document management
- Data mining
- Social engineering

Open issues

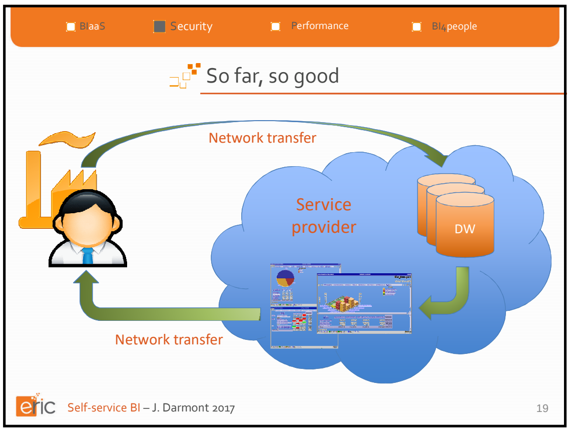
Advanced data model, query algebra
Tradeoff between data sharing and privacy
Formalization collective intelligence, interfaces
Economic model

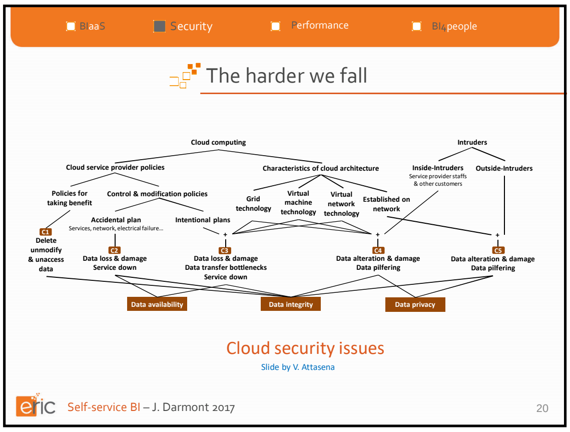
eric ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

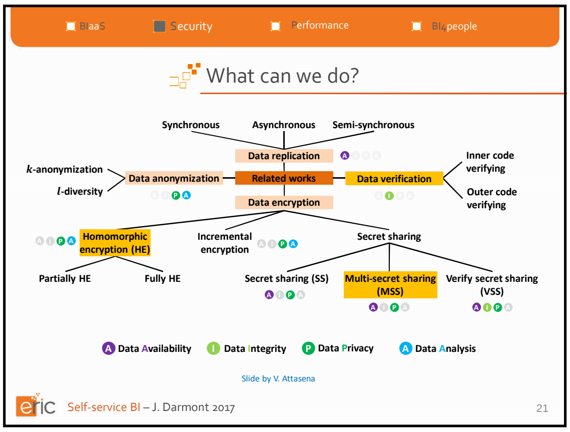
Data security in the cloud

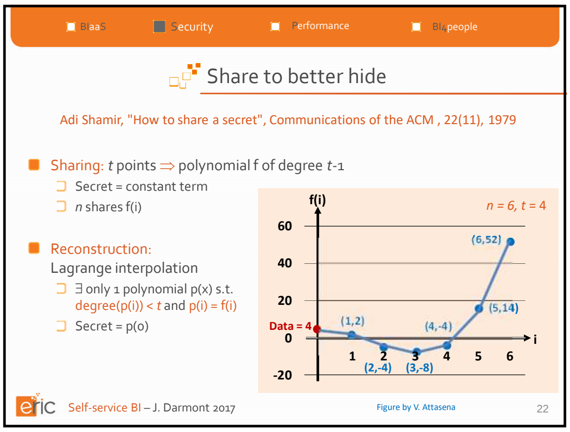
Part #2

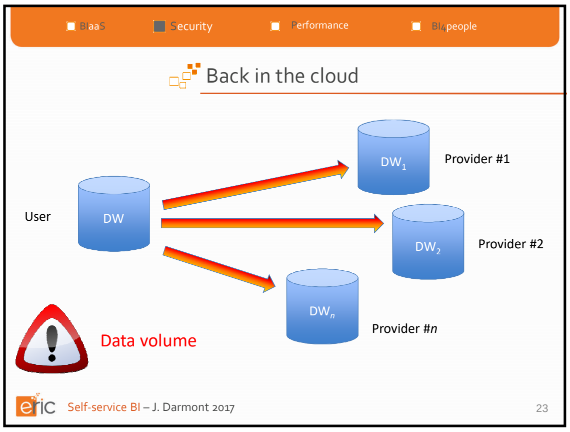
Lyon 1 UNIVERSITÉ LUMIÈRE LYON 2 UNIVERSITÉ DE LYON ish

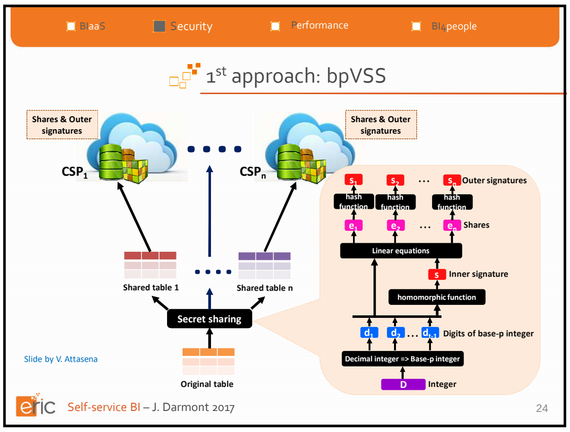

BIaaS Security Performance BI4people
So far, so good
Network transfer
Service provider
DW
Network transfer
Self-service BI – J. Darmont 2017
19

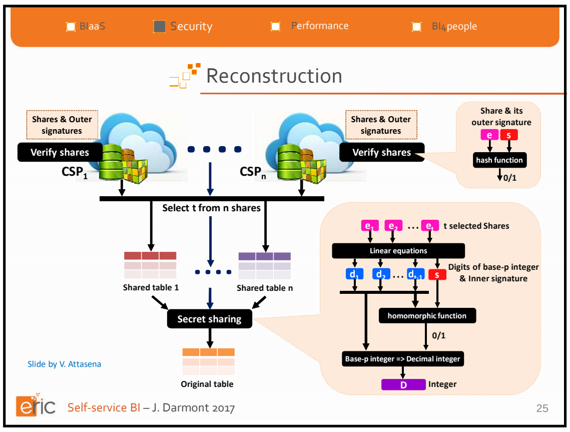

BIaaS Security Performance BI4people
The harder we fall
Cloud computing
Intruders
Cloud service provider policies
Characteristics of cloud architecture
Inside-Intruders
Outside-Intruders
Policies for taking benefit
Control & modification policies
Grid technology
Virtual machine technology
Virtual network technology
Established on network
Accidental plan
Intentional plans
Delete unmodify & unaccess data
Data loss & damage Service down
Data loss & damage Data transfer bottlenecks Service down
Data alteration & damage Data pilfering
Data alteration & damage Data pilfering
Data availability
Data integrity
Data privacy
Cloud security issues
Slide by V. Attasena
Self-service BI – J. Darmont 2017
20

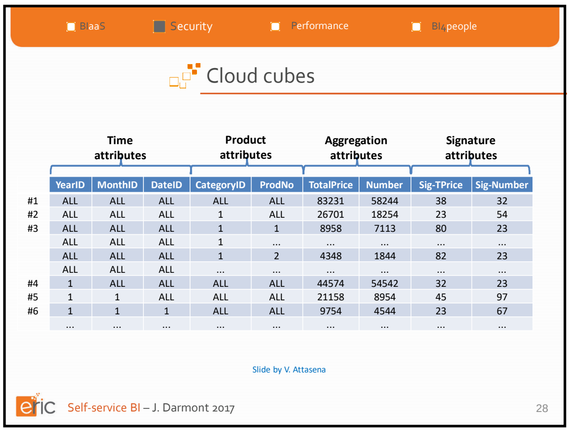

BIaaS Security Performance BI4people
What can we do?
Synchronous
Asynchronous
Semi-synchronous
Data replication
k-anonymization
Data anonymization
Related works
Data verification
Inner code verifying
l-diversity
Outer code verifying
Data encryption
Homomorphic encryption (HE)
Incremental encryption
Secret sharing
Partially HE
Fully HE
Secret sharing (SS)
Multi-secret sharing (MSS)
Verify secret sharing (VSS)
Data Availability
Data Integrity
Data Privacy
Data Analysis
Slide by V. Attasena
Self-service BI – J. Darmont 2017
21

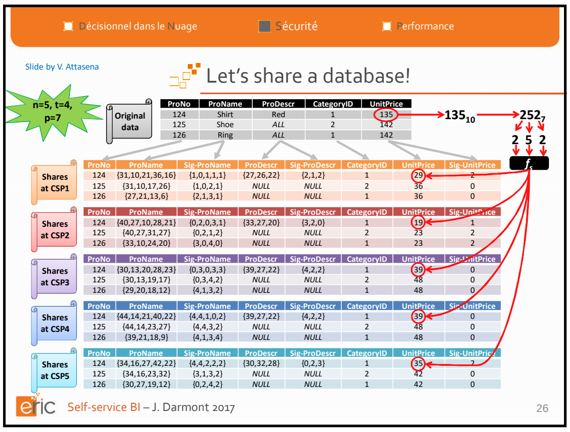

BIaaS Security Performance BI4people
Share to better hide
Adi Shamir, "How to share a secret", Communications of the ACM , 22(11), 1979
Sharing: t points ⇒ polynomial f of degree t-1
Secret = constant term
n shares f(i)
Reconstruction:
Lagrange interpolation
∃ only 1 polynomial p(x) s.t.
degree(p(i)) < t and p(i) = f(i)
Secret = p(0)
f(i)
n = 6, t = 4
60
40
20
0
-20
Data = 4
(6,52)
(5,14)
(1,2)
(4,-4)
(2,-4)
(3,-8)
1 2 3 4 5 6
i
Figure by V. Attasena
Self-service BI – J. Darmont 2017
22

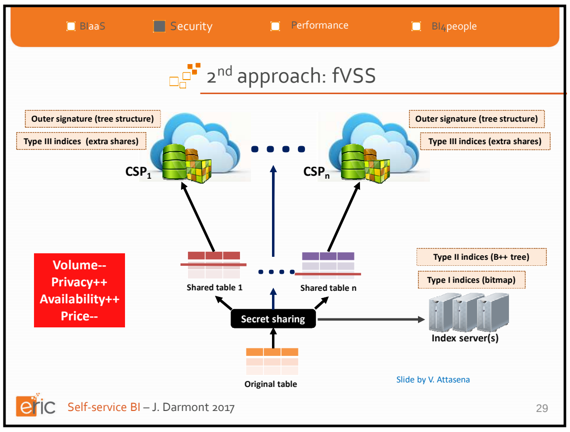

BIaaS Security Performance BI4people
Back in the cloud
User
DW
DW1
Provider #1
DW2
Provider #2
DWn
Provider #n
Data volume
Self-service BI – J. Darmont 2017
23

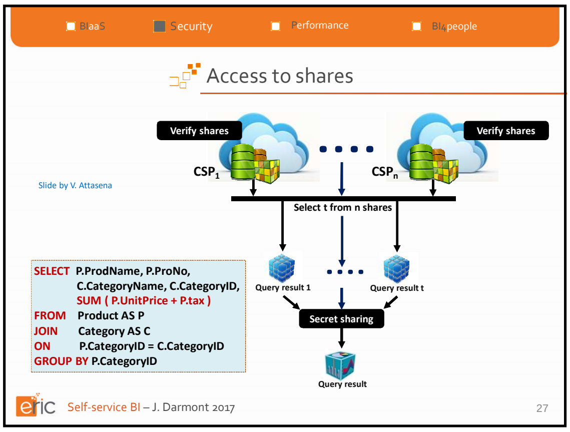

BIaaS Security Performance BI4people
1st approach: bpVSS
Shares & Outer signatures
Shares & Outer signatures
CSP1
CSPn
Outer signatures
hash function
Shares
Linear equations
Inner signature
homomorphic function
Digits of base-p integer
Decimal integer => Base-p integer
Integer
Shared table 1
Shared table n
Secret sharing
Original table
Slide by V. Attasena
Self-service BI – J. Darmont 2017
24

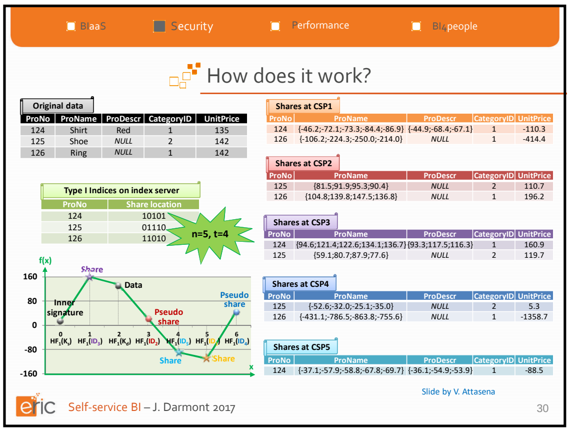

BIaaS Security Performance BI4people
Reconstruction
Shares & Outer signatures
Verify shares
CSP1
CSPn
Share & its outer signature
hash function
0/1
Select t from n shares
Shared table 1
Shared table n
Secret sharing
Original table
t selected Shares
Linear equations
Digits of base-p integer & Inner signature
homomorphic function
0/1
Base-p integer => Decimal integer
Integer
Slide by V. Attasena
Self-service BI – J. Darmont 2017
25


BIaaS Security Performance BI4people
Cloud cubes
Time attributes
Product attributes
Aggregation attributes
Signature attributes
YearID MonthID DateID CategoryID ProdNo TotalPrice Number Sig-TPrice Sig-Number
#1 ALL ALL ALL ALL ALL 83231 58244 38 32
#2 ALL ALL ALL 1 ALL 26701 18254 23 54
#3 ALL ALL ALL 1 1 8958 7113 80 23
ALL ALL ALL 1 ... ...
ALL ALL ALL 1 2 4348 1844 82 23
ALL ALL ALL ... ...
#4 1 ALL ALL ALL ALL 44574 54542 32 23
#5 1 1 ALL ALL ALL 21158 8954 45 97
#6 1 1 1 ALL ALL 9754 4544 23 67
... ...
Slide by V. Attasena
Self-service BI – J. Darmont 2017
28


Décisionnel dans le Nuage Sécurité Performance
Slide by V. Attasena
Let's share a database!
n=5, t=4, p=7
Original data
ProNo ProName ProDescr CategoryID UnitPrice
124 Shirt Red 1 135
125 Shoe ALL 2 142
126 Ring ALL 1 142
135₁₀
252₇
2 5 2
Shares at CSP1
Shares at CSP2
Shares at CSP3
Shares at CSP4
Shares at CSP5
Self-service BI – J. Darmont 2017
26


BIaaS Security Performance BI4people
2nd approach: fVSS
Outer signature (tree structure)
Type III indices (extra shares)
CSP1
CSPn
Volume--
Privacy++
Availability++
Price--
Shared table 1
Shared table n
Secret sharing
Original table
Type II indices (B++ tree)
Type I indices (bitmap)
Index server(s)
Slide by V. Attasena
Self-service BI – J. Darmont 2017
29


BIaaS Security Performance BI4people
Access to shares
Verify shares
CSP1
CSPn
Slide by V. Attasena
Select t from n shares
Query result 1
Query result t
Secret sharing
Query result
SELECT P.ProdName, P.ProNo,
C.CategoryName, C.CategoryID,
SUM (P.UnitPrice + P.tax)
FROM Product AS P
JOIN Category AS C
ON P.CategoryID = C.CategoryID
GROUP BY P.CategoryID
Self-service BI – J. Darmont 2017
27


BIaaS Security Performance BI4people
How does it work?
Original data
ProNo ProName ProDescr CategoryID UnitPrice
124 Shirt Red 1 135
125 Shoe NULL 2 142
126 Ring NULL 1 142
Type I indices on index server
ProNo Share location
124 10101
125 01110
126 11010
n=5, t=4
Shares at CSP1
Shares at CSP2
Shares at CSP3
Shares at CSP4
Shares at CSP5
Share
Data
Inner signature
Pseudo share
Slide by V. Attasena
Self-service BI – J. Darmont 2017
30

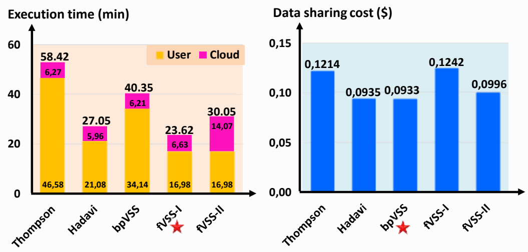
Us against the rest of the world

Data privacy			Data availability	Data integrity		Ability in case CSPs fail	
Data transferring	Data Storing	CSP group cheating		Inner code verification	Outer code verification	Querying shares	Updating shares
All approaches	All approaches	Only fVSS	All Approaches	Only [21][58][60] bpVSS, fVSS	Only bpVSS, fVSS	All approaches	Only fVSS

Features	[56]	[54]	[55][57]	[59]	[60] ★	[58] ★	[21]	bpVSS & fVSS ★
Data Types	Positive integers, Character, Strings,	Positive integers	Integers	Positive integers	Positive integers	Positive integers	Positive integers	Integer, Reals, Characters, Strings, Dates, Booleans
Shared data access								
Updates	-	-	Yes	Yes	Yes	Yes	Yes	Yes
Exact match queries	-	Yes	Yes	Yes	Yes	-	Yes	Yes
Range queries	-	Yes	Yes	Yes	Yes	-	Yes	Yes
Aggregation queries	Yes	Yes	Yes	Yes	Yes	Yes	-	Yes
Grouping queries	-	-	-	-	-	-	-	Yes

Slide by V. Attasena

Data sharing



Slide by V. Attasena

Slide by V. Attasena

Let's experiment!

Approaches:
- Thompson et al.'s [58]
- Hadavi et al.'s [60]
- bVSS
- fVSS-I (Balance share volume strategy)
- fVSS-II (Unbalance share volume strategy)

Benchmark:
Star Schema Benchmark (SSB)
SSB database: 757 MB

Measurements:
- Storage volume
- Data sharing time
- Data reconstruction time
- Data access time
- Data transfer volume

✚ Monetary costs

Parameter settings:
n = 5, t = 4, |signature| ≈ |data|/2

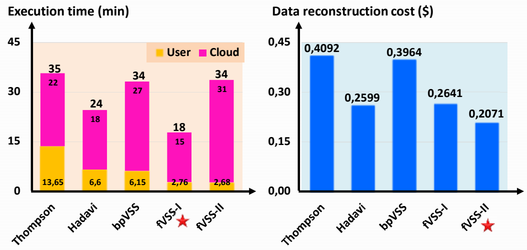
Virtual machines:

Approaches	CSP1	CSP2	CSP3	CSP4	CSP5	Index server
Thompson Hadavi fVSS-I	M3	M3	M3	M3	M3	M3
bpVSS	M3	M3	M3	M3	M3	-
fVSS-II	M3	M3	M2	M1	M1	M3

Share volume assignment:
Number of shared records (% of original)

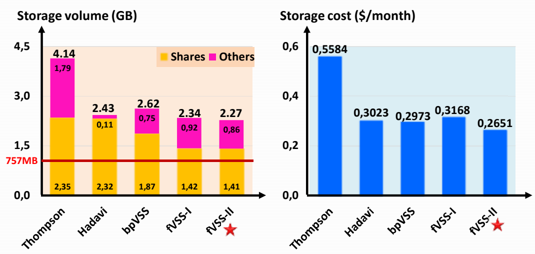
Approaches	CSP1	CSP2	CSP3	CSP4	CSP5
Thompson Hadavi bpVSS	100%	100%	100%	100%	100%
fVSS-I	60%	60%	60%	60%	60%
fVSS-II	81%	81%	60%	39%	39%

Secret reconstruction



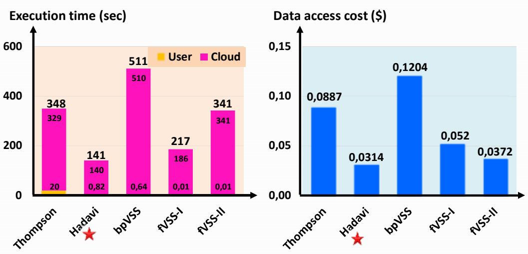
Slide by V. Attasena

Storage

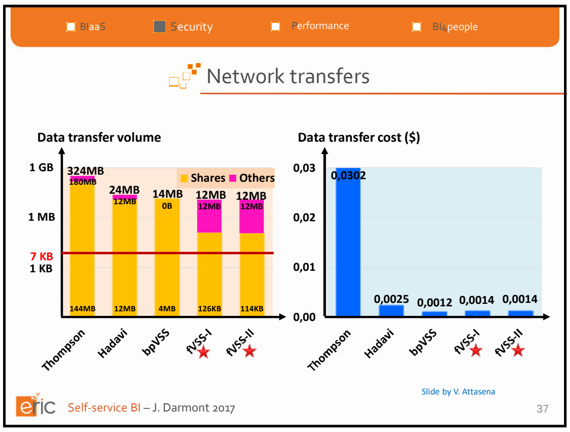


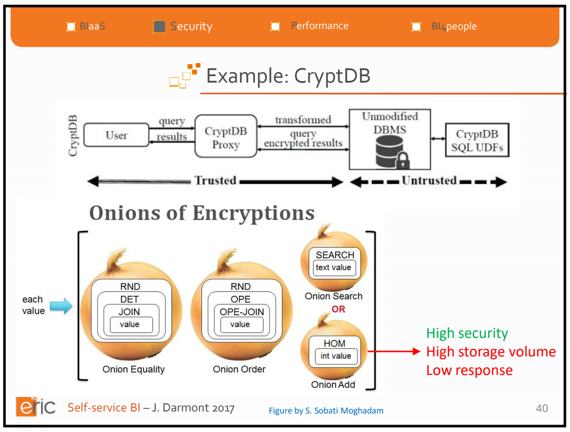
Slide by V. Attasena

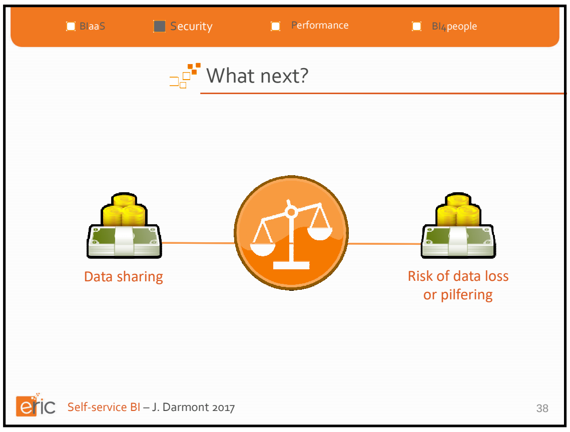
Query response

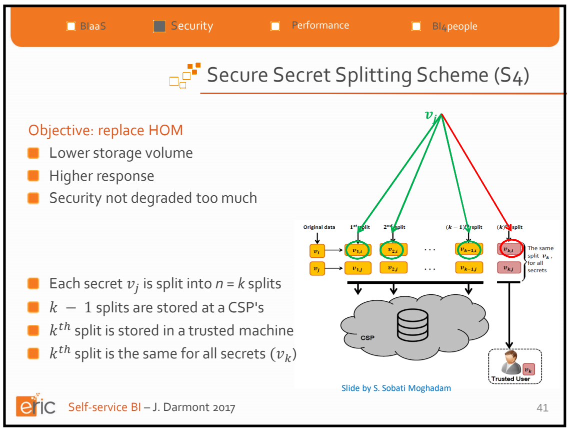


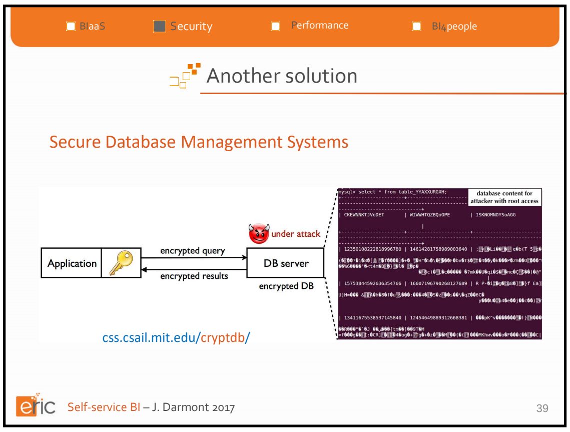
Slide by V. Attasena

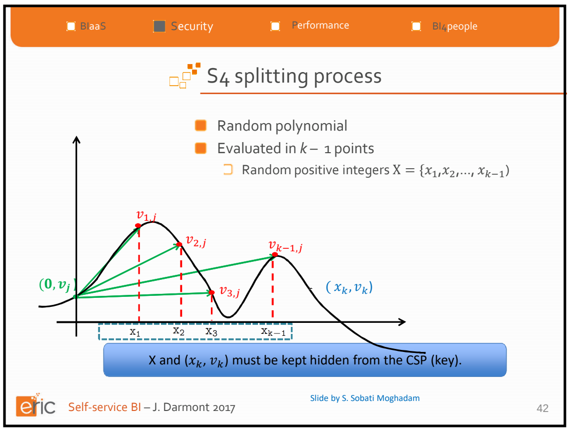

BIaaS
Security
Performance
BI4people
Network transfers
Data transfer volume
Data transfer cost ($)
1 GB
1 MB
7 KB
1 KB
324MB
180MB
24MB
12MB
14MB
0B
12MB
12MB
12MB
12MB
Shares
Others
144MB
12MB
4MB
126KB
114KB
0,03
0,02
0,01
0,00
0,0302
0,0025
0,0012
0,0014
0,0014
Thompson
Hadavi
bpVSS
fVSS-I
fVSS-II
Slide by V. Attasena
Self-service BI – J. Darmont 2017
37


BIaaS
Security
Performance
BI4people
Example: CryptDB
CryptDB
User
query
results
CryptDB Proxy
transformed query
encrypted results
Unmodified DBMS
CryptDB SQL UDFs
Trusted
Untrusted
Onions of Encryptions
each value
RND
DET
JOIN
value
Onion Equality
RND
OPE
OPE-JOIN
value
Onion Order
SEARCH
text value
Onion Search
OR
HOM
int value
Onion Add
High security
High storage volume
Low response
Self-service BI – J. Darmont 2017
Figure by S. Sobati Moghadam
40

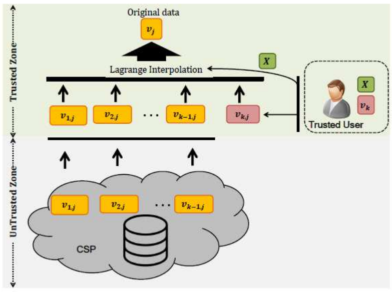

BIaaS
Security
Performance
BI4people
What next?
Data sharing
Risk of data loss or pilfering
Self-service BI – J. Darmont 2017
38


BIaaS
Security
Performance
BI4people
Secure Secret Splitting Scheme (S4)
Objective: replace HOM
Lower storage volume
Higher response
Security not degraded too much
Each secret v_j is split into $n = k$ splits
$k - 1$ splits are stored at a CSP's
k^{th} split is stored in a trusted machine
k^{th} split is the same for all secrets (v_k)
CSP
Trusted User
Slide by S. Sobati Moghadam
Self-service BI – J. Darmont 2017
41

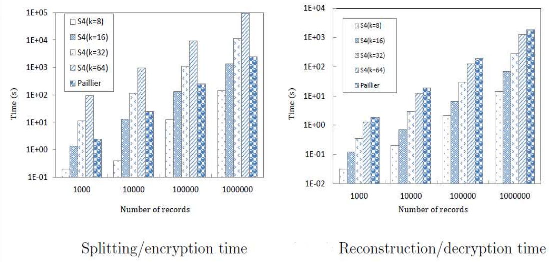

BIaaS
Security
Performance
BI4people
Another solution
Secure Database Management Systems
Application
encrypted query
encrypted results
DB server
encrypted DB
under attack
database content for attacker with root access
css.csail.mit.edu/cryptdb/
Self-service BI – J. Darmont 2017
39


BIaaS
Security
Performance
BI4people
S4 splitting process
Random polynomial
Evaluated in $k - 1$ points
Random positive integers $X = \{x_1, x_2, \ldots, x_{k-1}\}$
$(0, v_j)$
$v_{1,j}$
$v_{2,j}$
$v_{3,j}$
$v_{k-1,j}$
(x_k, v_k)
x_1
x_2
x_3
x_{k-1}
X and (x_k, v_k) must be kept hidden from the CSP (key).
Slide by S. Sobati Moghadam
Self-service BI – J. Darmont 2017
42

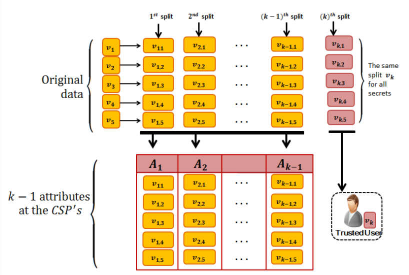
S4 reconstruction process



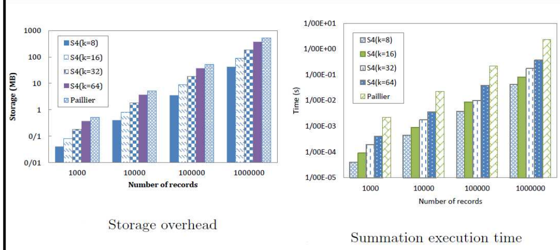
S4 vs. Paillier (HOM)



Aggregation queries with S4



S4 vs. Paillier (HOM)



S4 security assessment

- Secure against honest but curious provider
  - Private key needed to recontruct the splits
- S4 reveals linear combinations of splits
- Security is not guaranteed any more if an adversary knows:
  - Some secrets
  - Their splits
  - Adding noise shall make this much more difficult

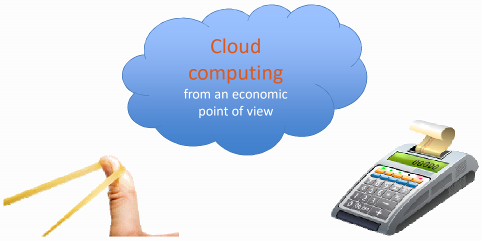
ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

Data access performance

Part #3

media.bestofmicro.com

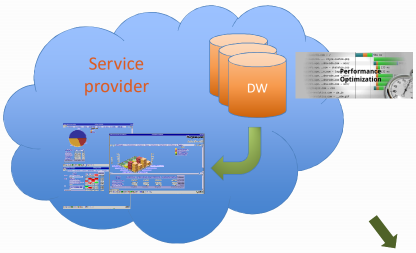
Pay more for better... performance



Elasticity

Pay as you go

Le'ts be pragmatic!

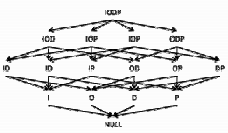


- Indices
- Caches
- Materialized views
- Partitioning
- ...

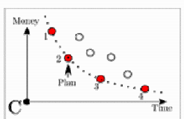
$\text{Cost}_{\text{global}} = \text{Cost}_{\text{transfer}} + \text{Cost}_{\text{computation}} + \text{Cost}_{\text{storage}}$

Problems & contributions

Select views to materialize



Multicriteria optimization



- Flexible pricing models
- Cost models for view materialization
- Optimization process detailed model

Transfer cost

$$C_t(D,Q,A) = C_t^-(D,Q) + C_t^+(A)$$

Ascending Descending

$$\approx C_t^+(A)$$

EC2

Volume	Cost
0 GB – 1 GB	0
1 GB – 10 TB	\$0.12 / GB
10 TB – 40 TB	\$0.09 / GB

- D : Dataset
- Q : Query workload
- A : Query answer

Computation cost

Processing time

$$C_c(Q,IC) = \sum_{i=1}^{n_Q}\sum_{j=1}^{n_{IC}} t(Q_i,IC_j) \times c_c(IC_j)$$

Rental cost

- $Q = \{Q_i\}$ / $i = 1..n_Q$: Query workload
- $IC = \{IC_j\}$ / $j = 1..n_{IC}$: Configuration of computation instances

Storage cost

S3

Volume	Cost
0 TB – 1 To	\$0.140 / GB
1 TB – 450 To	\$0.125 / GB

Storage cost Storage duration

$$C_s(D) = \sum_{k=1}^{n_D} c_s\big(s(D_k)\big) \times t(D_k)$$

Data size

- $D = \{D_k\}$ / $k = 1..n_D$: Données stockées par périodes de temps

Computation cost with MVs

Processing time

$$C_c(Q,V,IC) = T(Q,V) \times c_c(IC_0) \times n_{IC}$$

Rental cost

$$T(Q,V) = T_{proc}(Q,V) + T_{mat}(V) + T_{maint}(V)$$

Query execution Materialization Maintenance

- *Q* : Query workload
- *V* : Set of materialized views
- *IC* : Configuration of computation instances

Storage cost with MVs

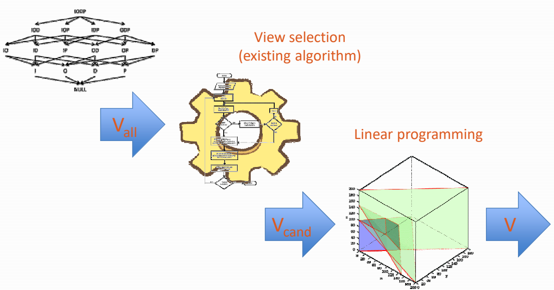
Storage cost Storage duration

$$C_s(D,V) = c_s(s(D) + s(V)) \times t$$

Data size

- *D* : Dataset
- *V* : Set of materialized views

Optimization process

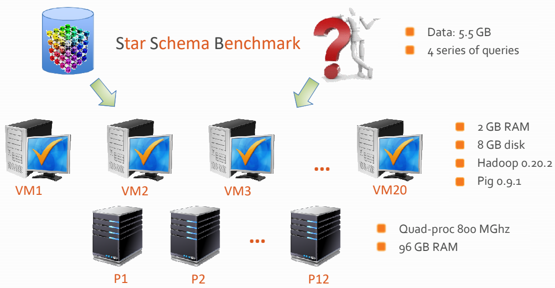


Optimization problems

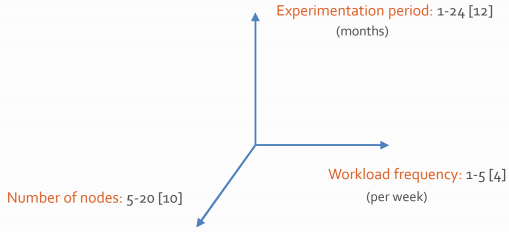
Find a set of materialized views $V \subseteq V_{cand}$

- *MV1*
  - Minimize T_{proc}
  - Constraint: $C \leq C_{max}$
- *MV2*
  - Minimize C
  - Constraint: $T_{proc} \leq T_{max}$
- *MV3*
  - Minimize $\alpha \times T_{proc} + (1 - \alpha) \times C$

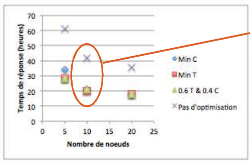
Environnement expérimental



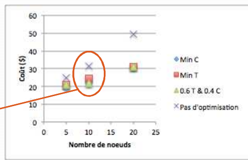
Parameters



Experimental results

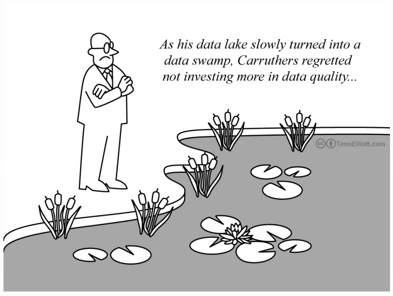


Performance gain: 110 %



Cost gain: 30 %

...or Data swamps?



timoelliott.com

What (should have we done) next?

More pricing models

amazon webservices

Google IaaS

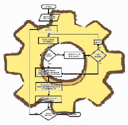
hp Cloud Services

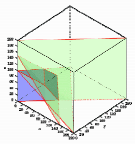
IBM

vmware vSphere

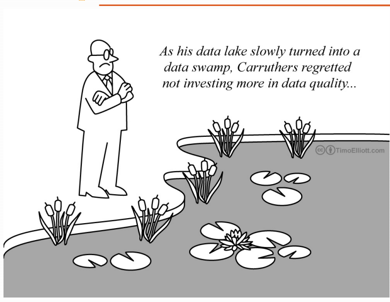
ORACLE

View materialization
Cost optimization



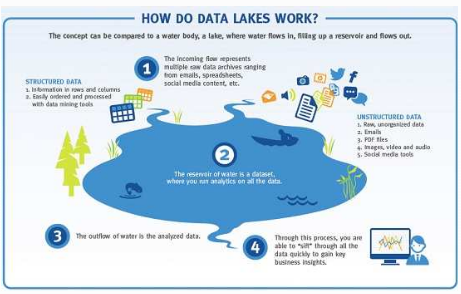


...or Data swamps?



timoelliott.com

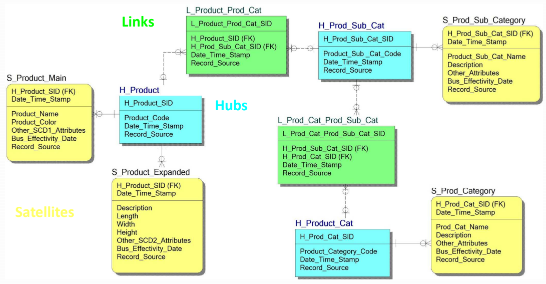
Something newer: Data lakes...



www.digitalnewsasia.com

Data vaults

(see also anchor modeling)



hanshultgren.files.wordpress.com

Filling data vaults with data lakes?

Ongoing research

- Data vaults as a storage solution for data lake metadata

Metadata
Descriptive | Administrative | Structural
s3.amazonaws.com

- Performance comparison
  - Metadata stored in a relational data vault
  - Metata stored in a NoSQL DBMS

eric ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

BI4people research project

Part #4

Lyon 1 UNIVERSITÉ LUMIÈRE LYON 2 UNIVERSITÉ DE LYON ish

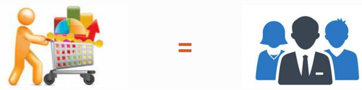
Motivation, action!

The term "data-driven innovation" (DDI) refers to the capacity of businesses and public sector bodies to make use of information from improved data analytics to develop improved services and goods that facilitate everyday life of individuals and of organizations, including SMEs.

AGENCE NATIONALE DE LA RECHERCHE ANR

- 2016
  - Preproposal
  - Proposal
    …but good reviews
- 2017
  - Preproposal
  - Proposal

On-demand BI, yes but…

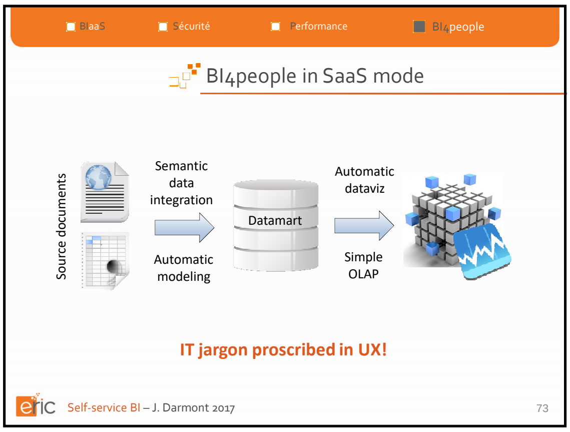


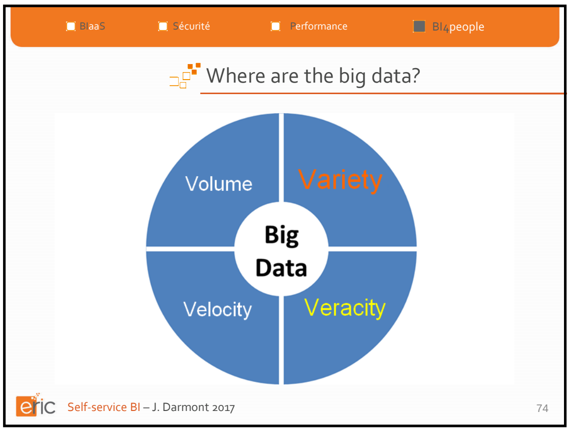
Self-service BI = Expert business

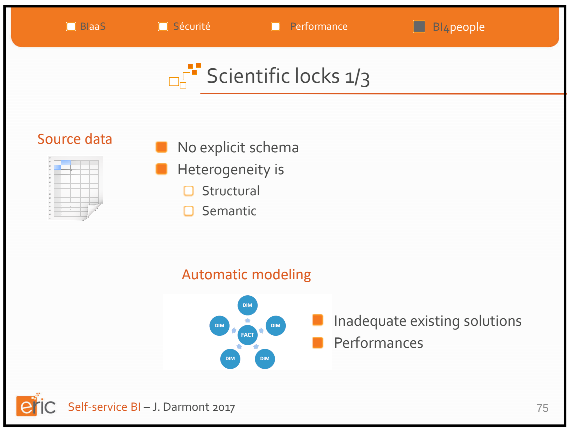
BI4people = SMEs, Organizations, Independents, Active citizens

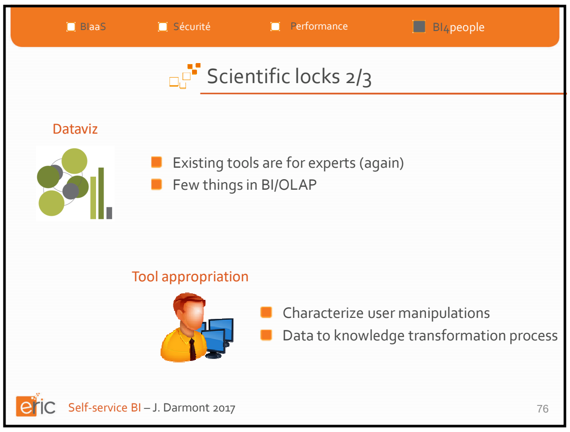
Plus a collaborative layer!

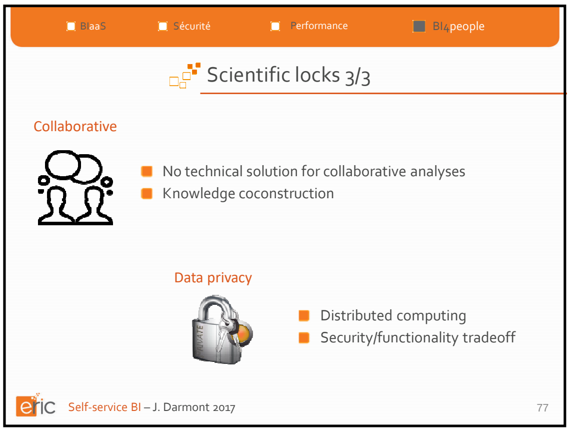
www.businessmarches.com

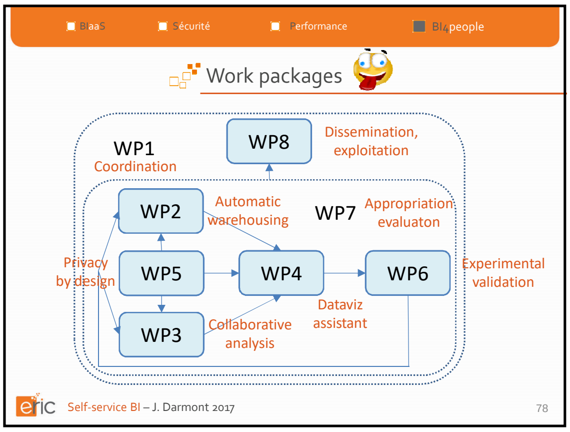

BIaaS
Sécurité
Performance
BI4people
BI4people in SaaS mode
Source documents
Semantic data integration
Automatic modeling
Datamart
Automatic dataviz
Simple OLAP
IT jargon proscribed in UX!
Self-service BI – J. Darmont 2017
73


BIaaS
Sécurité
Performance
BI4people
Where are the big data?
Volume
Variety
Velocity
Veracity
Big Data
Self-service BI – J. Darmont 2017
74


BIaaS
Sécurité
Performance
BI4people
Scientific locks 1/3
Source data
No explicit schema
Heterogeneity is
Structural
Semantic
Automatic modeling
Inadequate existing solutions
Performances
Self-service BI – J. Darmont 2017
75


BIaaS
Sécurité
Performance
BI4people
Scientific locks 2/3
Dataviz
Existing tools are for experts (again)
Few things in BI/OLAP
Tool appropriation
Characterize user manipulations
Data to knowledge transformation process
Self-service BI – J. Darmont 2017
76

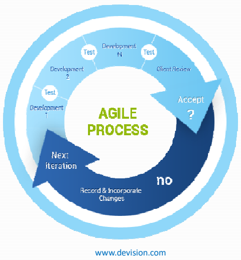

BIaaS
Sécurité
Performance
BI4people
Scientific locks 3/3
Collaborative
No technical solution for collaborative analyses
Knowledge coconstruction
Data privacy
Distributed computing
Security/functionality tradeoff
Self-service BI – J. Darmont 2017
77


BIaaS
Sécurité
Performance
BI4people
Work packages
WP1
Coordination
WP8
Dissemination, exploitation
WP2
Automatic warehousing
WP7
Appropriation evaluaton
Privacy by design
WP5
WP4
WP6
Experimental validation
WP3
Collaborative analysis
Dataviz assistant
Self-service BI – J. Darmont 2017
78

A little agility

Face au monde qui change, il vaut mieux penser le changement que changer le pansement.
Pierre Dac & Francis Blanche

Program testing can be used to show the presence of bugs, but never to show their absence.
Edsger Dijkstra



WP2 Automatic data warehousing

- Tabular data typology
- Generic model for representic tabular data
- Automatic process for warehousing tabular data

WP3 Collaborative analysis

- Typology of collaborative approaches in BI
- Design and implementation of a collaborative analysis framework

WP4 Dataviz assistant

- User context definition
- Design of elementary OLAP visualizations
- Dashboard representation
- Dataviz recommendation algorithms
- Dataviz user appropriation study

WP5 Privacy by design

- Transversal design of data privacy
- Cryptographic protocol that allows shared analyses without disclosing fine-grained data

WP6 Experimental validation

- Unit testing, integration testing
- Field testing

WP7 Appropriation evaluation

- Protocol definition
- User observation
- Post-observation interviews
- Cross analysis of transcriptions w.r.t. transferability

ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

Simplicity is the ultimate sophistication.
Leonardo da Vinci

Objective – Deliverable

- Preindustrial prototype
- Open source
- Valorization
  - Service in MesInfos
  - Maybe more…

Plan B

La Région Auvergne-Rhône-Alpes

HORIZON 2020

ANR Labcom

FUNDBYU Découvrir, Croire, Soutenir