EAS : Entity Analytic Solutions

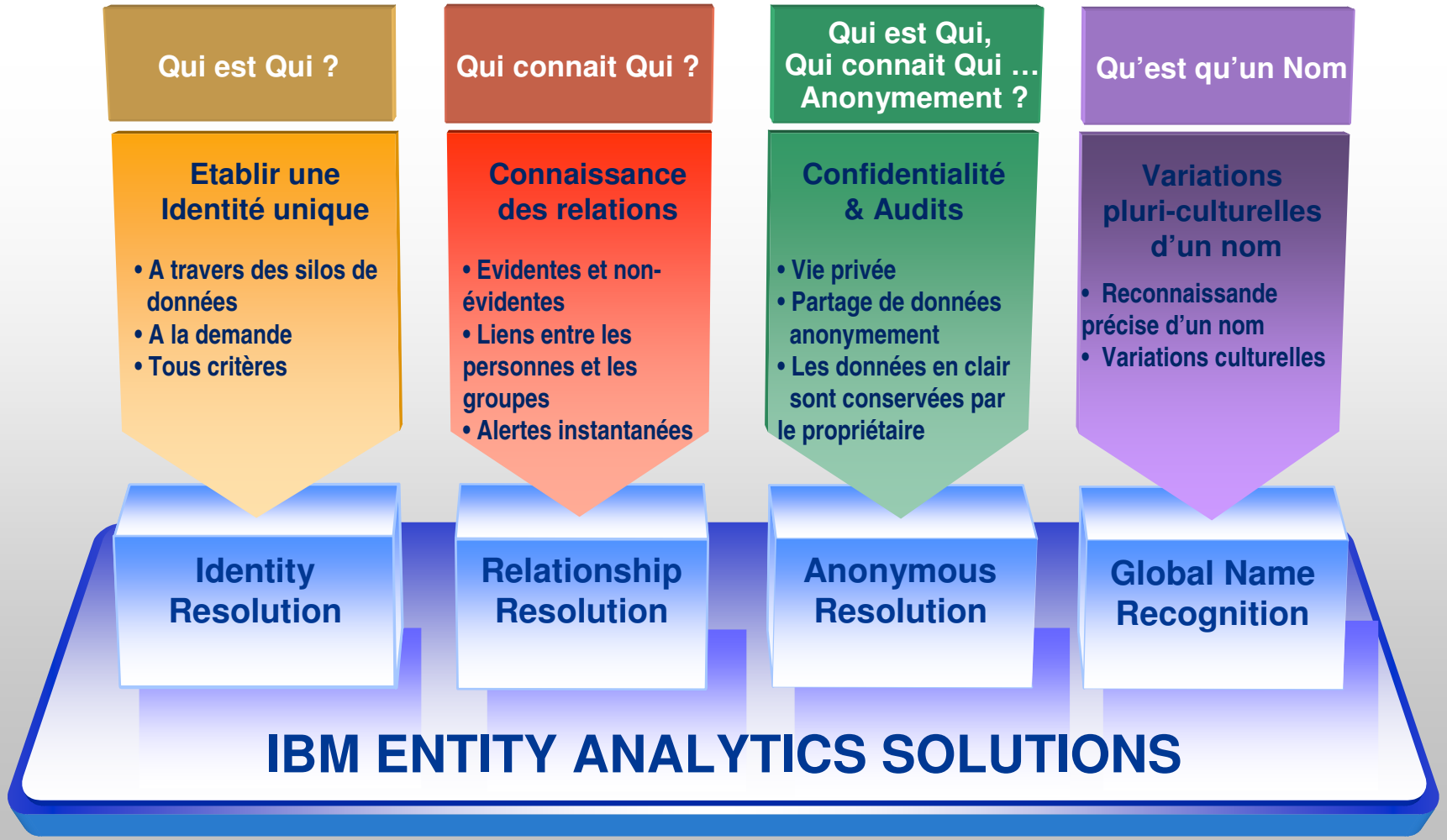
Hubert Pouderoux

Architecte Systèmes d'Information Décisionnels

hpouderoux@fr.ibm.com

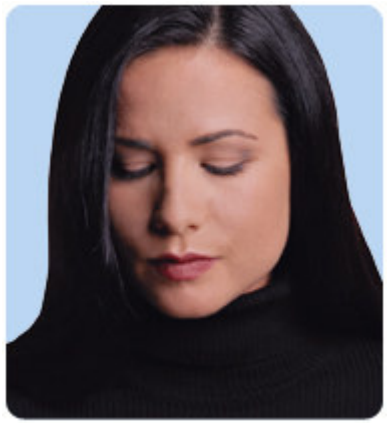
16 novembre 2006

Entity Analytics Solutions s'articule autour de 4 modules



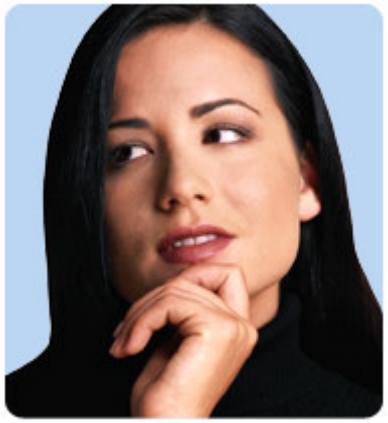
Qui est qui – Identity Resolution

Ce module permet de reconnaître une personne physique ou morale utilisant de multiples identités



Dr Katrin Dupont
1 Ave Dumesnil
78230 Saint-Cloud
Tel :01 49 05 50 84
CDF : 2640809737615
DDN :07/08/64
NPS : 068588345

Dr Cathy Dupond
10 rue Saint-Martin
Chatillon 51700
Tel : 03 49 05 50 84
NPS : 068588345



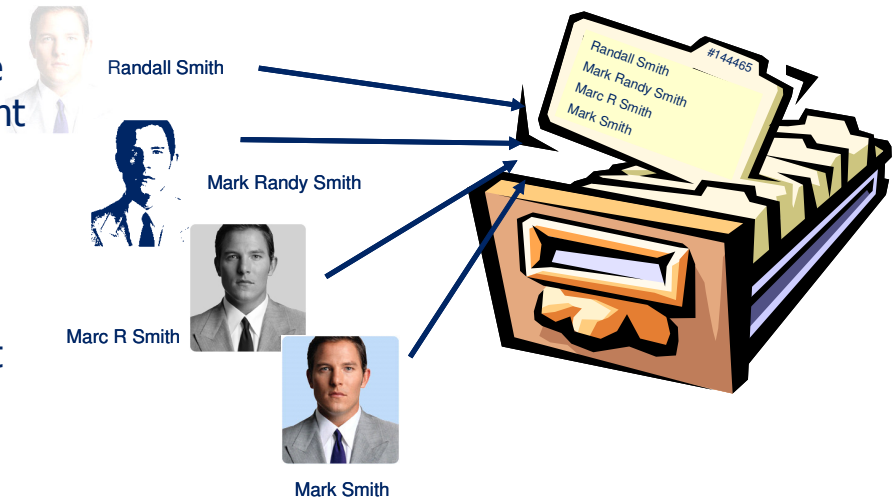
Mme Catherine Dupond
1 rue de Bourgogne
Chatillez 51700
Tel :03 49 05 60 55
DDN : 07/09/66
CDF : 2660973761563

Mme Katrin Dubois
Tour Europa Appt B24
La Défense, 92066
Tel : 01 45 65 45 40
DDN : 07/08/64
CDF : 2460809737615

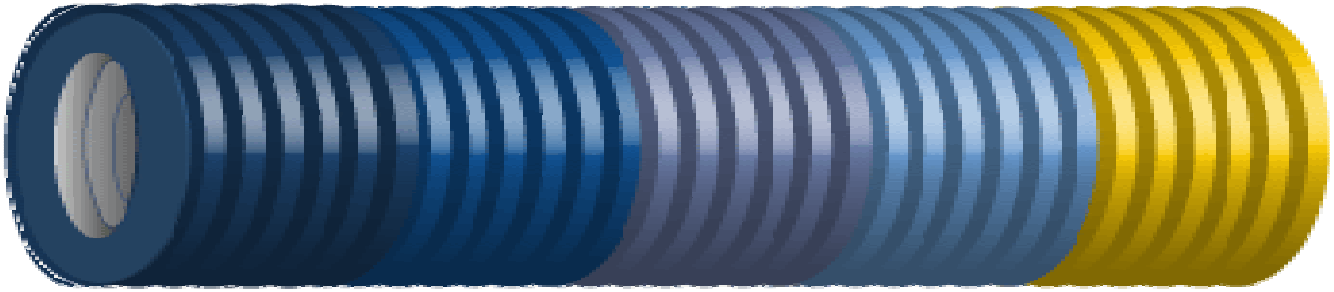
Qui est qui – Identity Resolution

- Identités exactes, mises à jour en permanence (en temps réel ou non)
  - Auto-correction
  - Pas besoin de rafraîchissement ou de rechargement
  - Pas de dégradation du temps de latence
- Tous les attributs des enregistrements sont conservés, maintenus et résolus
  - L'historique est conservé
  - Les identités peuvent être rapprochées, séparées ou enrichies en temps réel

- Exemples:
- Un souscripteur est déjà connu, présent dans les bases de données sous une autre identité (partielle ou totale).
- Une bande organisée de fraudeurs a cherché à souscrire un crédit en utilisant de fausses identités mais partageant une (ou plusieurs) information ; on ne sait pas a priori lesquelles.
- Un souscripteur est présent sans le savoir (client dormant).
- Le souscripteur est en fait un client indésirable (il n'avait pas remboursé totalement un précédent crédit à la consommation ou il est interdit de chéquier



Principe de fonctionnement



- Technologie non-intrusive prête à l'emploi
- Identités exactes, mises à jour en permanence en temps réel
  - **Auto-correction**
  - **Pas besoin de rafraîchissement ou de rechargement**
  - **Pas de dégradation du temps de latence**
- Toutes les caractéristiques des enregistrements sont conservées, maintenues et résolues
  - **Tous les attributs sont utilisés pour résoudre les identités, pas seulement nom + adresse**
  - **Les identités peuvent être rapprochées, séparées ou enrichies en temps réel**
- Support d'un nombre illimité de sources de données
  - **Sources internes et externes**

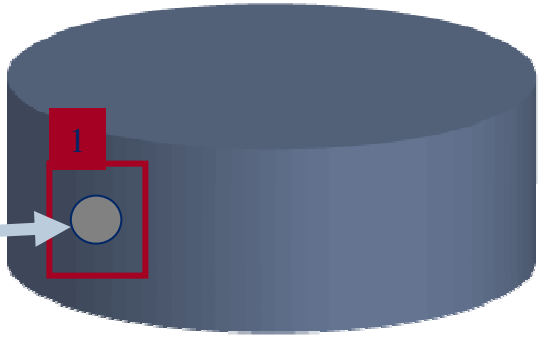
Résolution d'identité – attributs extensibles

Nom	Adresse	Identifiants	Attributs
– Nom – Prénom – Surnom – Nom d'usage – Nom d'organisation – Alias	– Adresse 1 – Adresse 2 – Adresse 3 – Ville – Etat/Province – Code Postal – Pays – Latitude/longitude	– Carte de crédit – Permis de conduire – Compte bancaire – Numéro CNI – Numéro Passeport – Carte de fidélité – Téléphone – Adresse Courriel – Adresse IP – **… defini à la demande**	– Date naissance – Nationalité – Lieu de Résidence – Lieu de naissance – Taille – Poids – Couleur yeux – Couleur cheveux – **… défini à la demande**

Résolution d'identité : premier exemple

Injection d'un premier enregistrement (2002)
Que se passe-t-il ?

Record A-70001
Marc R Smith
123 Main St
(713) 730 5769
537-27-6402
DL: 0001133107



EAS ID #144465

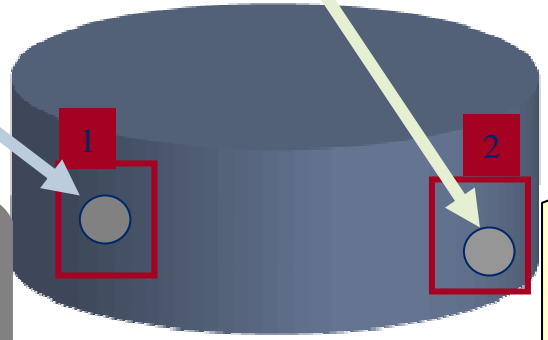
Noms	Marc R Smith	A-#70001
Adresse	123 Main St.	A-#70001
Tél	(713) 730-5769	A-#70001
CNI	537-27-6402	A-#70001
PC	1133107	A-#70001

Une nouvelle entité (#1) est créée dans la base

Injection d'un deuxième enregistrement (2003)

Record A-70001
Marc R Smith
123 Main St
(713) 730 5769
537-27-6402
PC: 0001133107

Record B-00009103
Randal M Smith
DDN: 17/06/1934
(713) 731 5577



EAS ID #144465

Noms	Marc R Smith	A-#70001
Adresse	123 Main St.	A-#70001
Tél	(713) 730-5769	A-#70001
CNI	537-27-6402	A-#70001
PC	1133107	A-#70001

EAS ID #321489

Noms	Randal M Smith	B-#9103
DDN	06/17/1974	B-#9103
Tél	(713) 731-5577	B-#9103

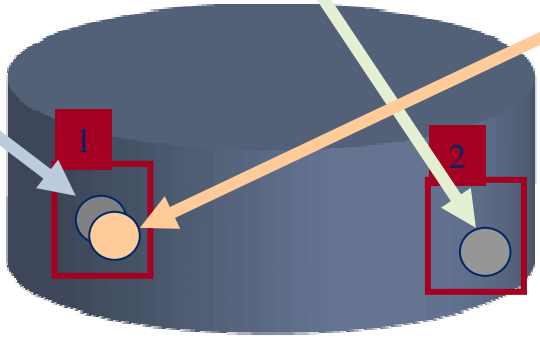
Pas de corrélation pertinente entre les identités – création d'une deuxième entité

Injection d'un troisième enregistrement (2004)

Record A-70001
Marc R Smith
123 Main St
(713) 730 5769
537-27-6402
PC: 0001133107

Record B-00009103
Randal M Smith
DDN: 17/06/1934
(713) 731 5577

Record C-6251
Mark Randy Smith
456 First Street
(713) 731 5577
PC:1133107



EAS ID #144465

Noms	Marc R Smith	A-#70001
	Mark Randy Smith	C-#6251
Adresses	123 Main St.	A-#70001
	456 First St	C-#6251
Tél	(713) 730-5769	A-#70001
	(713) 731-5577	C-#6251
CNI	537-27-6402	A-#70001
PC	1133107	C-#6251
	1133107	A-#70001

Des attributs sont rapprochés avec l'entité #1 = "**Résolution**"

EAS ID #321489

Noms	Randal M Smith	B-#9103
DDN	06/17/1934	B-#9103
Tél	(713) 731-5577	B-#9103

Contexte à l'injection – 3ème enregistrement (2004)

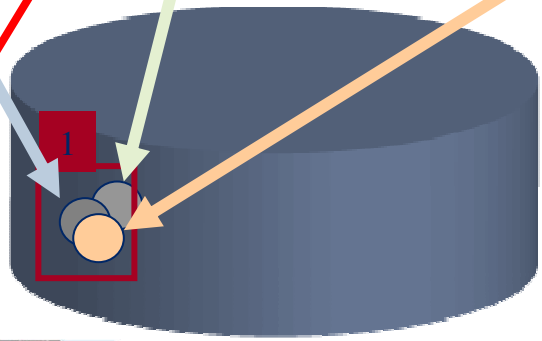
Record A-70001
Marc R Smith
123 Main St
(713) 730 5769
537-27-6402
PC: 0001133107

Record B-00009103
Randal M Smith
DDN: 17/06/1934
(713) 731 5577

Record C-6251
Mark Randy Smith
456 First Street
(713) 731 5577
PC:1133107

EAS ID #144465

Noms	Marc R Smith	A-#70001
	Mark Randy Smith	C-#6251
	Randal M Smith	B-#9103
Adresses	123 Main St.	A-#70001
	456 First St	C-#6251
Tél	(713) 730-5769	A-#70001
	(713) 731-5577	B-#9103
	(713) 731-5577	C-#6251
CNI	537-27-6402	A-#70001
PC	1133107	C-#6251
	1133107	A-#70001
DDN	06/17/1934	B-#9103



L'identité #2 correspond à l'identité #1 –
Les entités sont jointes

Injection d'un quatrième enregistrement (2005)

Record A-7000
Marc R Smith
123 Main St
(713) 730 5769
537-27-6402
PC: 0001133107

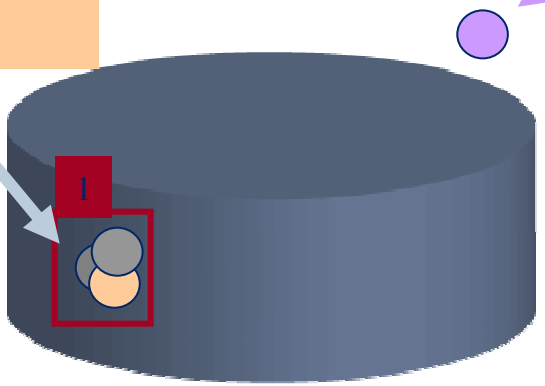
Record B-00009103
Randal M Smith
DDN: 17/06/1934
(713) 731 5577

Record C-6251
Mark Randy Smith
456 First Street
(713) 731 5577
PC:1133107

Record D-7700214
Randy Smith Sr.
DDN: 17/6/1934
(713) 731-5577
423-22-7027

EAS ID #144465

Noms	Marc R Smith	A-#70001
	Mark Randy Smith	C-#6251
	Randal M Smith	B-#9103
Adresses	123 Main St.	A-#70001
	456 First St	C-#6251
Tél	(713) 730-5769	A-#70001
	(713) 731-5577	B-#9103
	(713) 731-5577	C-#6251
CNI	537-27-6402	A-#70001
PC	1133107	C-#6251
	1133107	A-#70001
DDN	06/17/1934	B-#9103



L'enregistrement est rapproché de l'entité #1

Injection d'un quatrième enregistrement (2005)

Record A-70001
Marc
123
(713)
537-
PC: 0

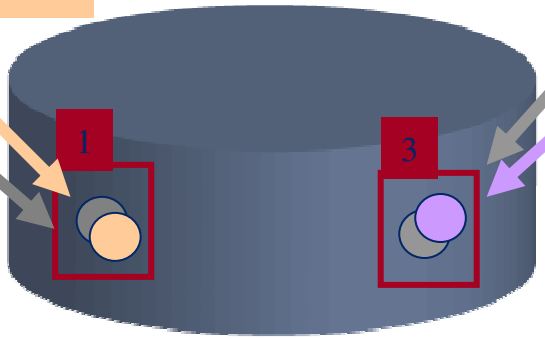
Record C-6251
Mark Randy Smith
456 First Street
(713) 731 5577
PC:1133107

← L'enfant

Record D-7700214
Randy Smith Sr.

Le père →

Record B-00009103
Randal M Smith
DDN: 17/06/1934
(713) 731 5577



Un conflit est identifié – le contexte est corrigé = "**Découplage**"

EAS ID #144465

Noms	Marc R Smith	A-#70001
	Mark Randy Smith	C-#6251
Adresses	123 Main St.	A-#70001
	456 First St	C-#6251
Tél	(713) 730-5769	A-#70001
	(713) 731-5577	C-#6251
CNI	537-27-6402	A-#70001
PC	1133107	C-#6251
	1133107	A-#70001

EAS ID #321463

Noms	Randy Smith Sr	D-#7700214
	Randal M Smith	B-#9103
Tél	(713) 731-5577	D-#7700214
	(713) 731-5577	B-#9103
CNI	423-22-7027	D-#7700214
DDN	06/17/1934	D-#7700214
	06/17/1934	B-#9103

En résumé sur la résolution d'identités

Injection continue et analyse en temps réel

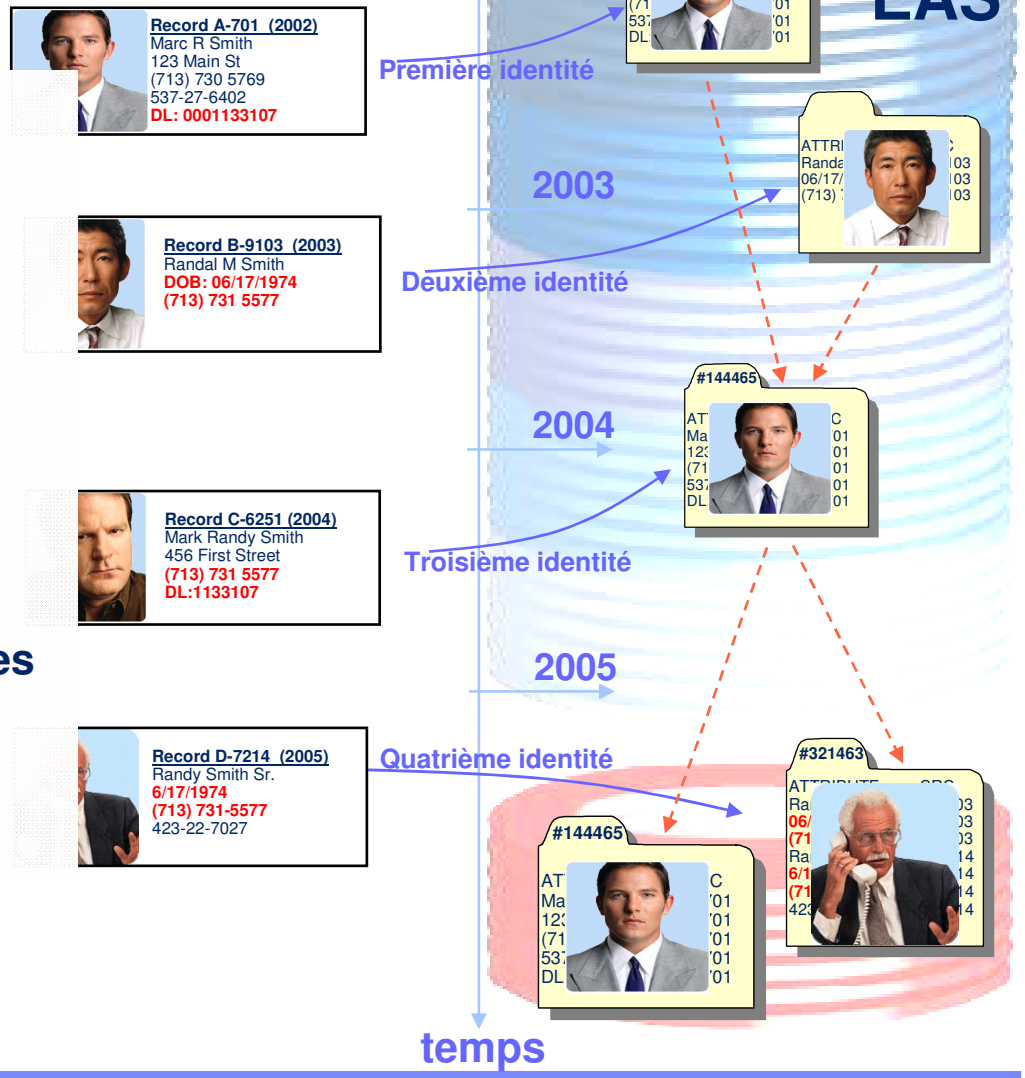
Tous les attributs sont utilisés

Contexte persistant

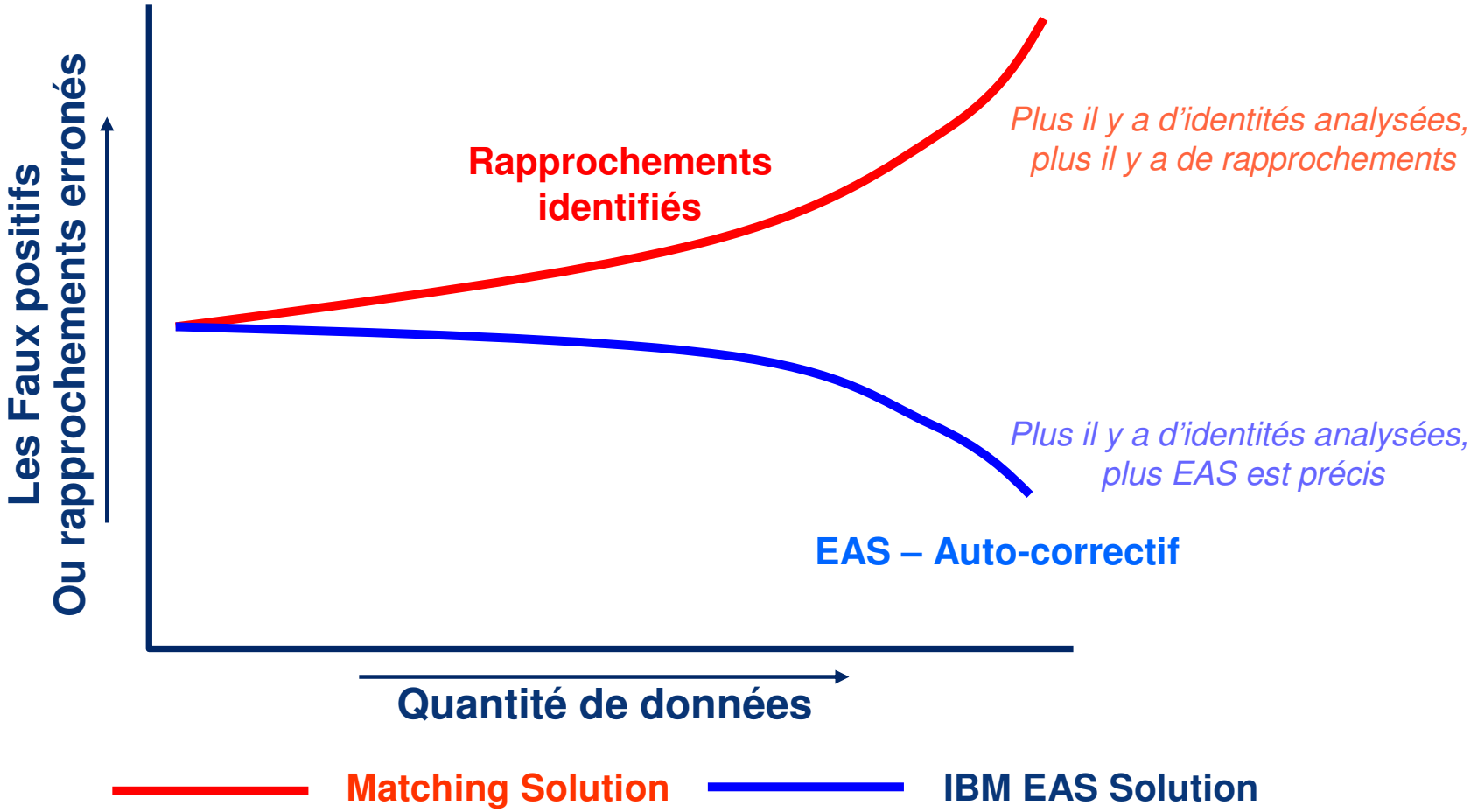
Pas d'apprentissage de données requis

Auto-correctif et indépendant des séquences

Fonctionnement viable sur le long terme

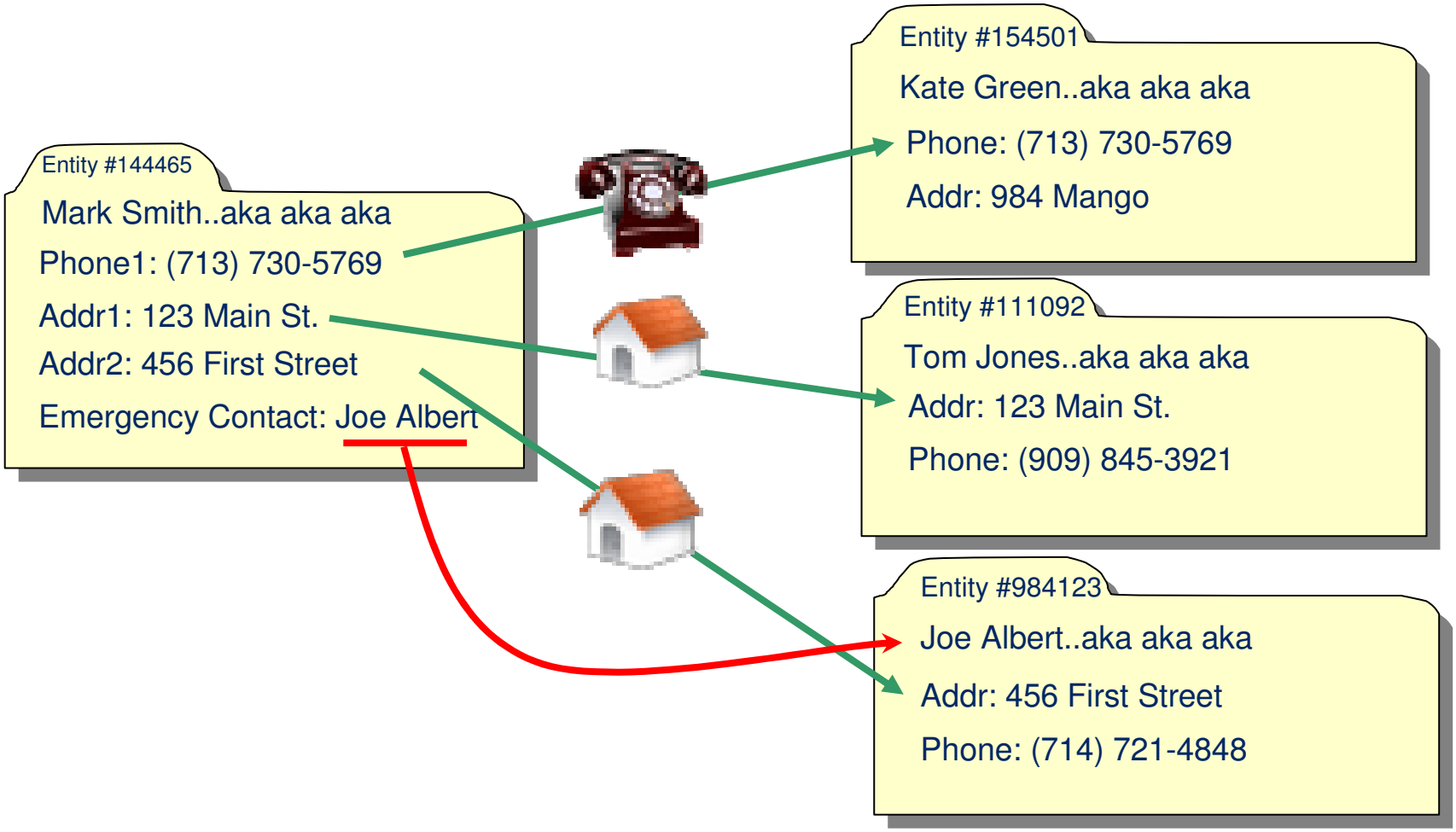


Entity Analytic Solutions n'est pas un "Matcher"

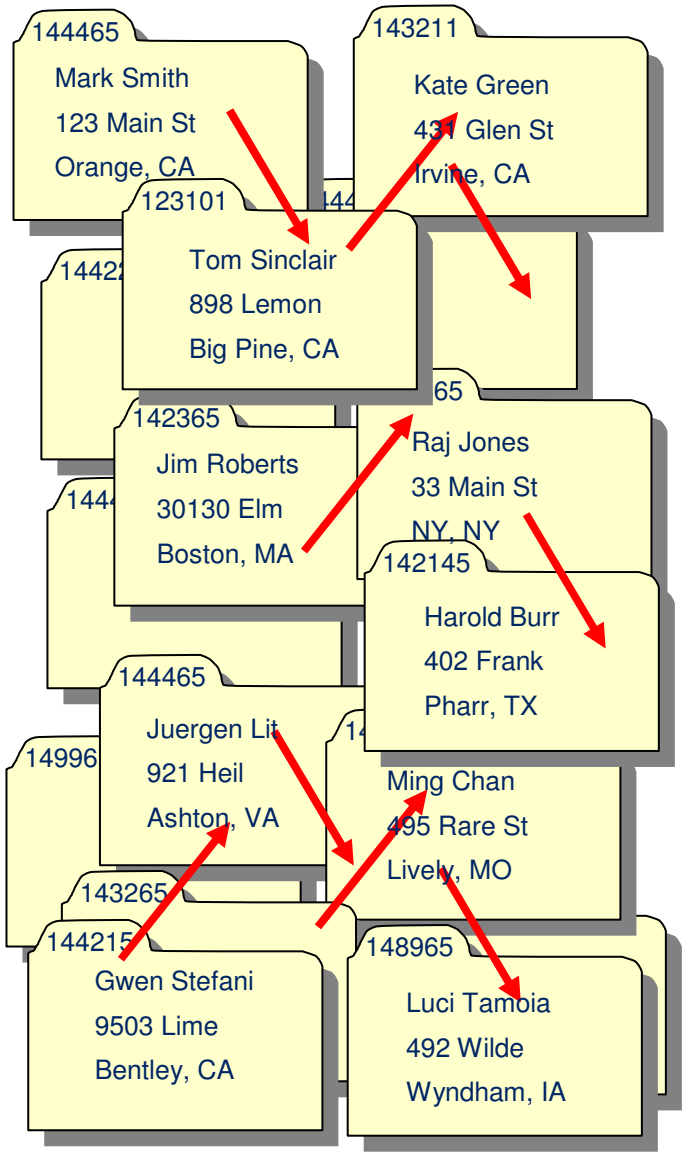


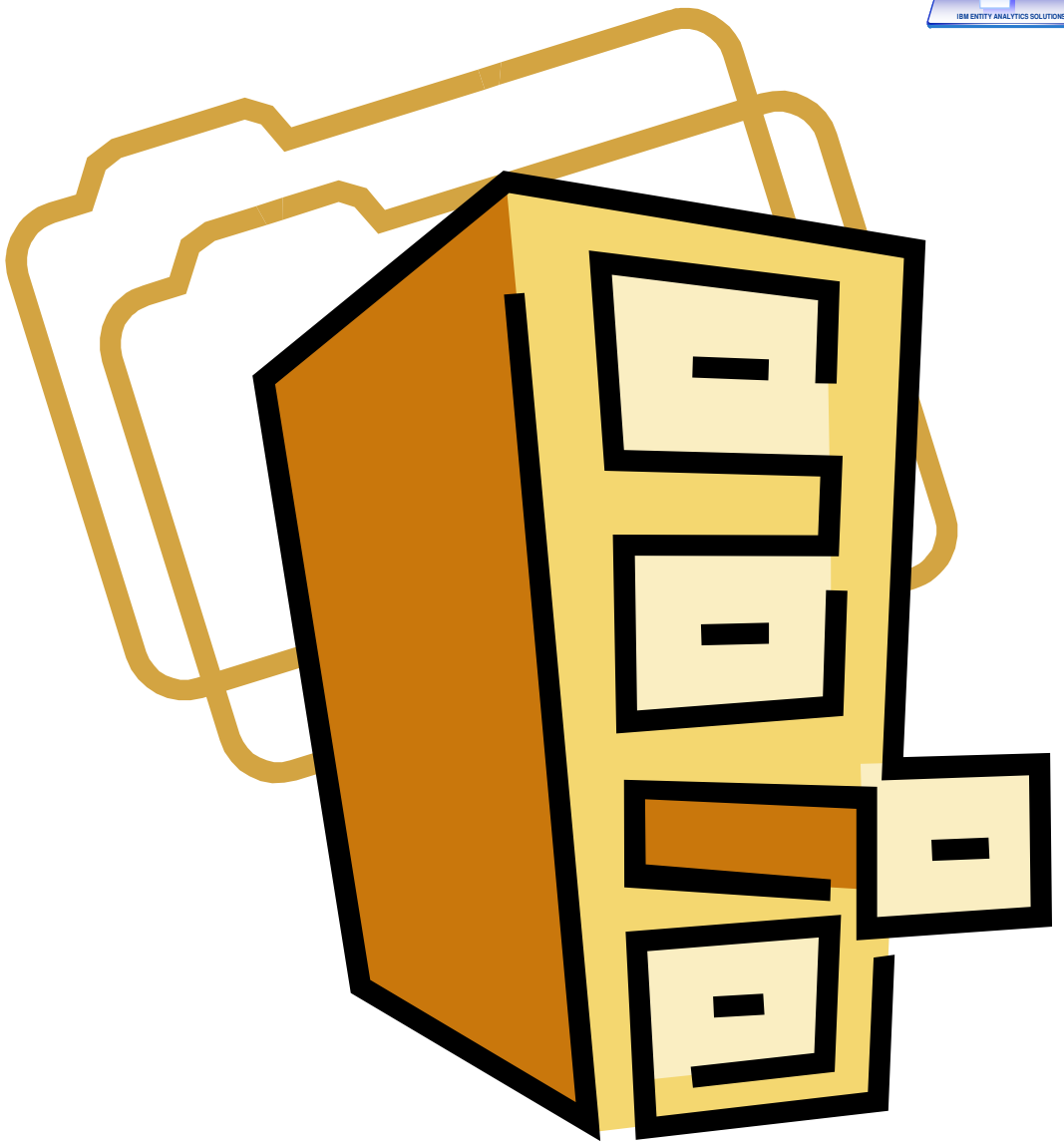
La capacité auto-corrective d'EAS limite les rapprochements erronés

Qui connait qui ? Relationship Resolution : la résolution de relations entre 'entités'



Association du contenu dans le référentiel des identités



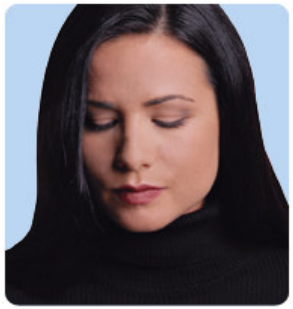


La résolution de relations – notion de "degrés de séparation"

(Règle associative: si A = B = C, alors A = C)

A: Mark Smith
Phone: (713) 730 5769

=



B: Kate Green
Phone: (713) 730 5769
Addr: 123 Main St

=

C: Tom Sinclair
Addr: 123 Main St

A: Mark Smith
Phone: (713) 730 5769

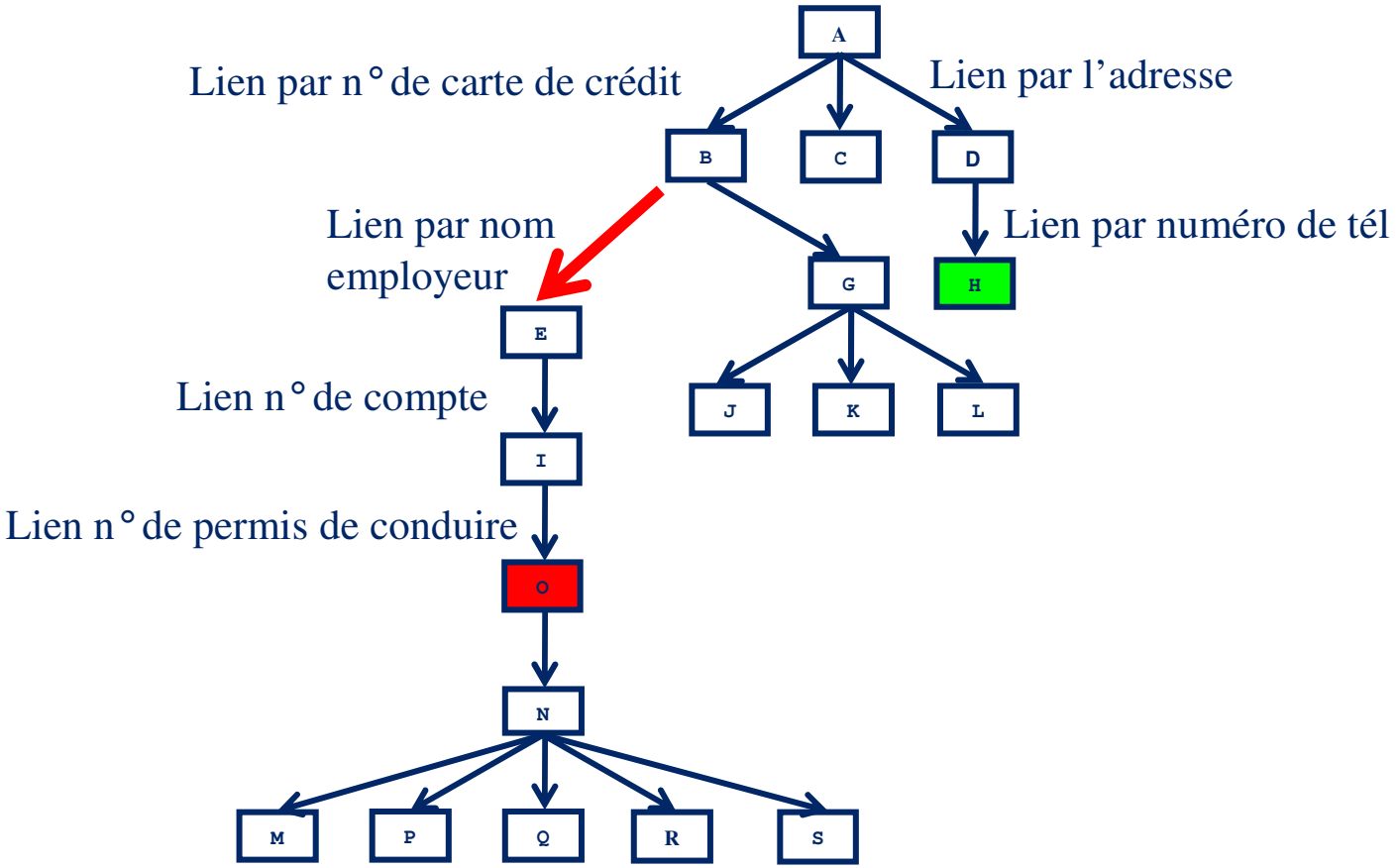
=

C: Tom Sinclair
Addr: 123 Main St

Mark est en relation avec Tom avec 2 degrés de séparation.

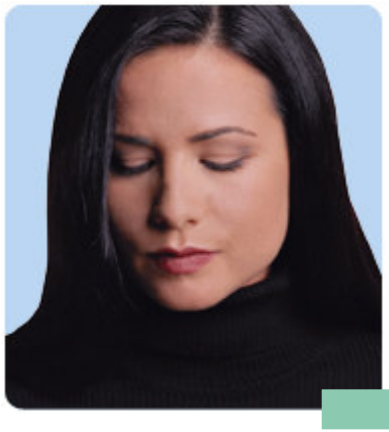
EAS gère jusqu'à 30 degrés de séparation!

Découverte de réseau : degré de séparation



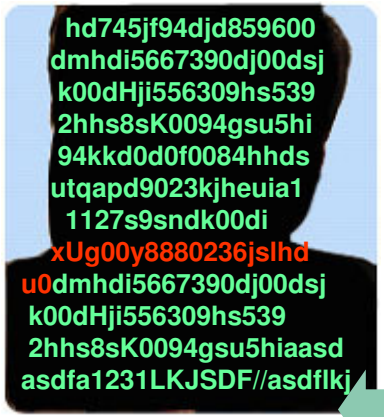
“Qui est qui et qui connait qui… en respectant l’anonymat”
IBM Anonymous Resolution

Permet à plusieurs intervenants de partager et de comparer des informations rendues anonymes.



Catherine Dupond
1 rue de Bourgogne
Chatillez 51700
VIN# 585789543
Frequence Plus: 5678965
Tel: 03 49 05 60 55
Passeport : 995027890

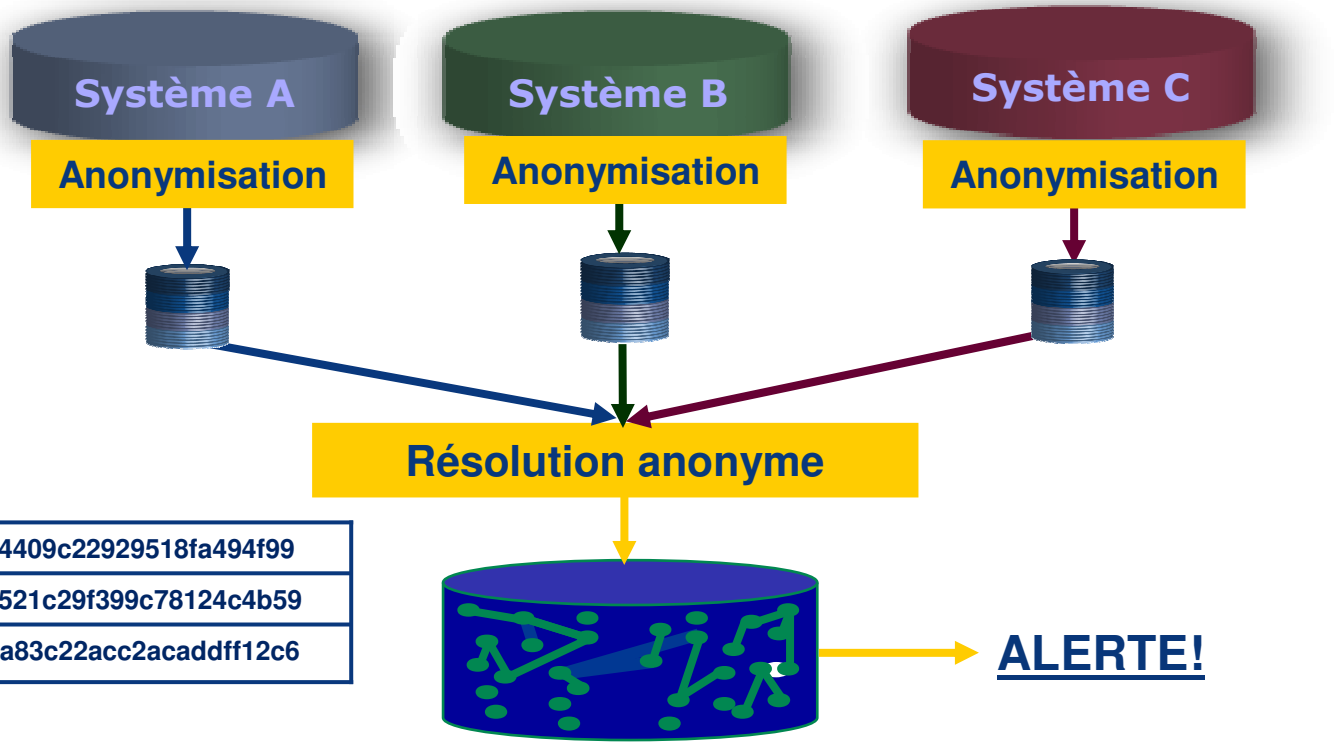
tr5y9hU000kdG563ks
Hjd55603jd98hj4jkf9jt
0089gehyud98kkdh00
Hydk8880h332jd78001
xug00y8880236jslhdu0
0012g6743kd85hf06h7
x084hdf75jc4539fhd89



hd745jf94djd859600dm
hdi5667390dj00dsjk00d
Hji556309hs5392hhs8s
K0094gsu5hi94kkd0d0f
0084hhdsutqapd9023kj
heuia11127s9sndk00di
xug00y8880236jslhdu0

Thomas Saint-Clair
49 Rue Basse
Louveciennes 78120
Compte n° : 97836553122
Compte n° : 00303450009
Tel : 01 50 16 03 82
Frequence Plus: 5678965

Fonctionnement et utilisation de l'anonymisation



A-100031	Prénom	cbd034409c22929518fa494f99
A-100031	Nom	b835b521c29f399c78124c4b59
A-100031	Date de naissance	40ddba83c22acc2acaddff12c6

- Pas de communication en clair de l'information
- Utilisation de l'information maîtrisée et sous le contrôle de son propriétaire
- Réduit les risques de violations des règles de protection de l'information

Reconnaissance des noms

Contrairement à d'autres éléments d'information, les noms revêtent différentes formes et particularités. Ce ne sont pas de simples chaînes de caractères, mais des objets flous susceptibles de beaucoup de variantes.

- Il n'existe pas de standard.
- Les noms peuvent contenir une variété d'informations optionnelles qui peut faire apparaître les noms de façon très différente :

Fautes de frappe,
- SMITH~SIMTH, JOHNSON~JPHNSON

Bruit
- THOMPSON~TH9MP2ON

Titres
- Dr., Rev, Haj, Sri., Col.,…

Préfixes
- Fitz, O', De La, Abdul, …

Surnoms
- Johnny, Betty, Alyosha, Paco, Drew

Phonétiques
- Leighton~Layton, Dena-Deane

Noms abrégés
- Fco, Ma, Mohd,…

Qualifieurs
- Jr., fils, neto, sobrinho, Ph.D.,

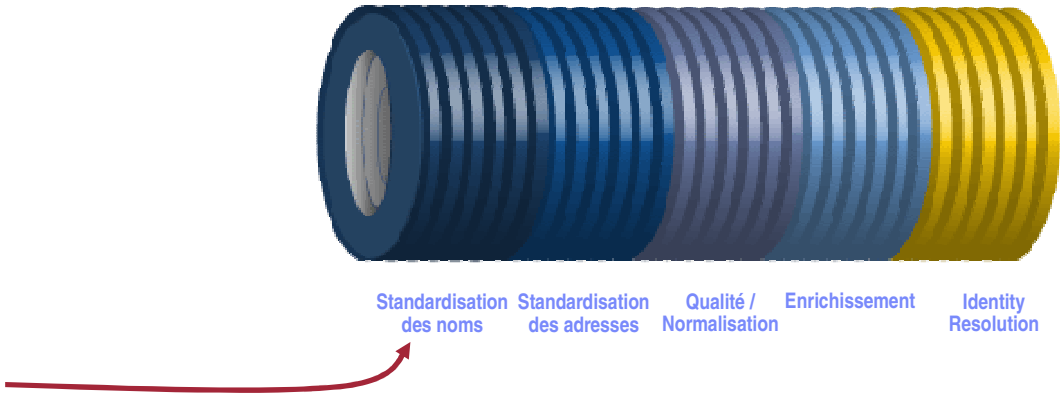
Séquences
- (Tung, Mao Tse)

Parties manquantes
- (Abdel ~ Abdel Rahman)

Variations
- (Mohammed, Imhemmed)

IBM GLOBAL NAME RECOGNITION

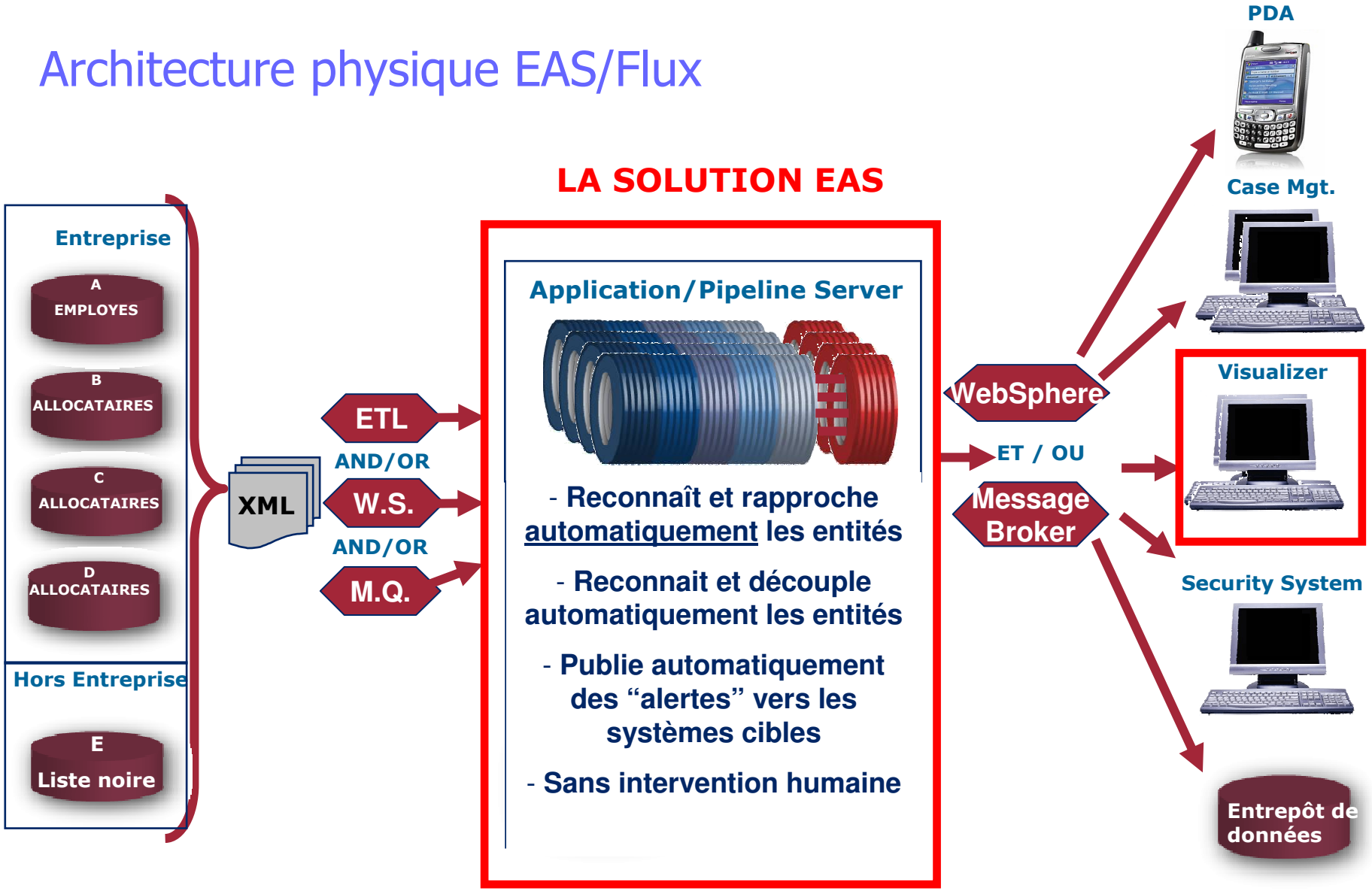
- Mono-attribut par opposition à multi-attribut
- Approche basée sur une base de connaissance de près d'1 milliard de noms appartenant à 200 pays
  - Automatise plus de 20 ans de recherches linguistiques
- Des règles appropriées à chaque langue pour effectuer les meilleures comparaisons/recherches /décompositions, avec des mécanismes de calcul de score
- S'intègre dans la solution EAS



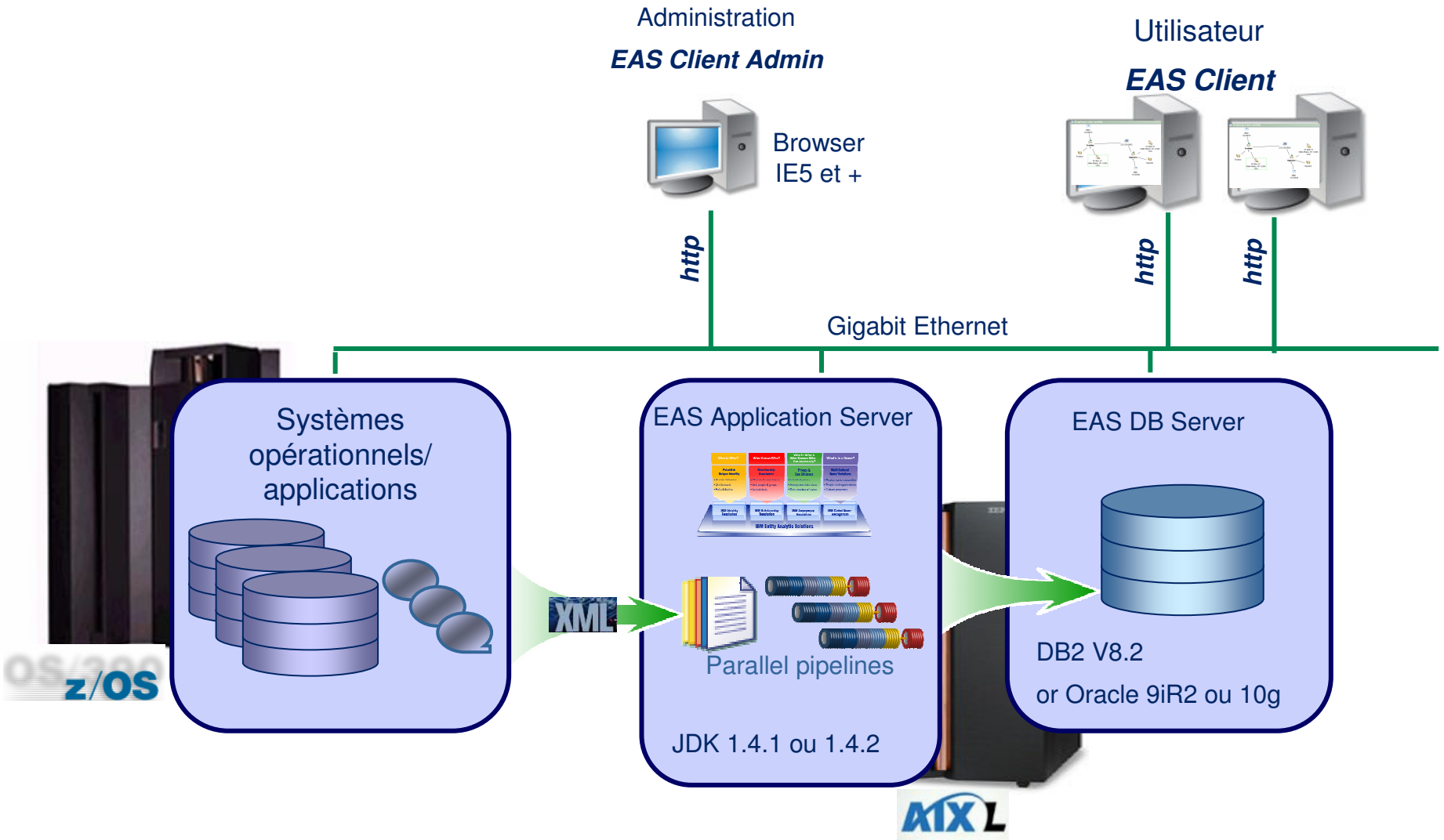
En conclusion : les points forts de la solution

- Utilisation de nombreux attributs
- Conservation du contexte historique
- Détection de relations même très éloignées
- Performances très élevées
- Fonctionnement en temps réel
- N'est pas focalisé sur un type de délinquance

Architecture physique EAS/Flux



Architecture technique : les composants



Merci