IBM SolutionsConnect 2013

L'IBM TechSoftware nouvelle génération

28, 29 et 30 août

IBM Client Center Paris

#solconnect13

Transformez vos opportunités en succès

IBM

Anonymisation des données

James LEBAS
Avant Vente IBM Infosphere Optim

AGENDA

- Introduction
- Rappel de la Solution IBM : InfoSphere Optim
  - Gestion des données de test
  - Anonymisation des données sensibles
- Nouveautés Masking on Demand
- InfoSphere Discovery : Découverte des données sensibles
- Conclusion

AGENDA

- **Introduction**
- Rappel de la Solution IBM : InfoSphere Optim
  - Gestion des données de test
  - Anonymisation des données sensibles
- Nouveautés Masking on Demand
- InfoSphere Discovery : Découverte des données sensibles
- Conclusion

Qu'est-ce que l'anonymisation ?

Anglais : 'Data Masking'

-Un processus par lequel des données sont rendues anonymes, processus à l'issue duquel elles ne peuvent plus être affectées ou rattachées à une personne en particulier, à un individu.

-Données nominatives (nom, prénom, date de naissance,..) & indirectement (matricule, une adresse, téléphone, IP, etc...)

-Données sensibles (numéro compte bancaire, carte crédit, …)

Anonymiser : pourquoi ?

- A cause des réglementations
  - CNIL
  - European Personal Data Protection Directive
  - PCI-DSS
  - HIPAA
  - Sarbanes-Oxley (SOX)
  - Basel II
  - …
- Parce que les fuites de données ont des effets d'image et parfois financiers importants

La CNIL

- La loi CNIL du 7 août 2004 (cf annexe 1), dont les décrets d'application ont été publiés en octobre 2005, oblige à ***garantir l'anonymat des informations qui s'appliquent aux personnes physiques tout au long du cycle de vie des données au sein du système d'information***.
- Les amendes et sanctions pénales encourues sont détaillées dans les articles 226-16 à 22—24 du code pénal:
- jusqu'à 1,5 million d'euros d'amende pour l'entreprise,
  - Jusqu'à cinq ans d'emprisonnement et 300 000 euros d'amende pour leur dirigeant (DG et/ou DSI)
  - Interdiction d'exercer, ou fermeture d'un ou plusieurs établissements ayant servis à commettre l'infraction (art 131-88 et 131-39)
- Point d'attention particulière : l'accès à tout fichier contenant des données personnelles hors des frontières européennes doit être soumis à l'autorisation de la CNIL (réglementation européenne).

Le PCI DSS

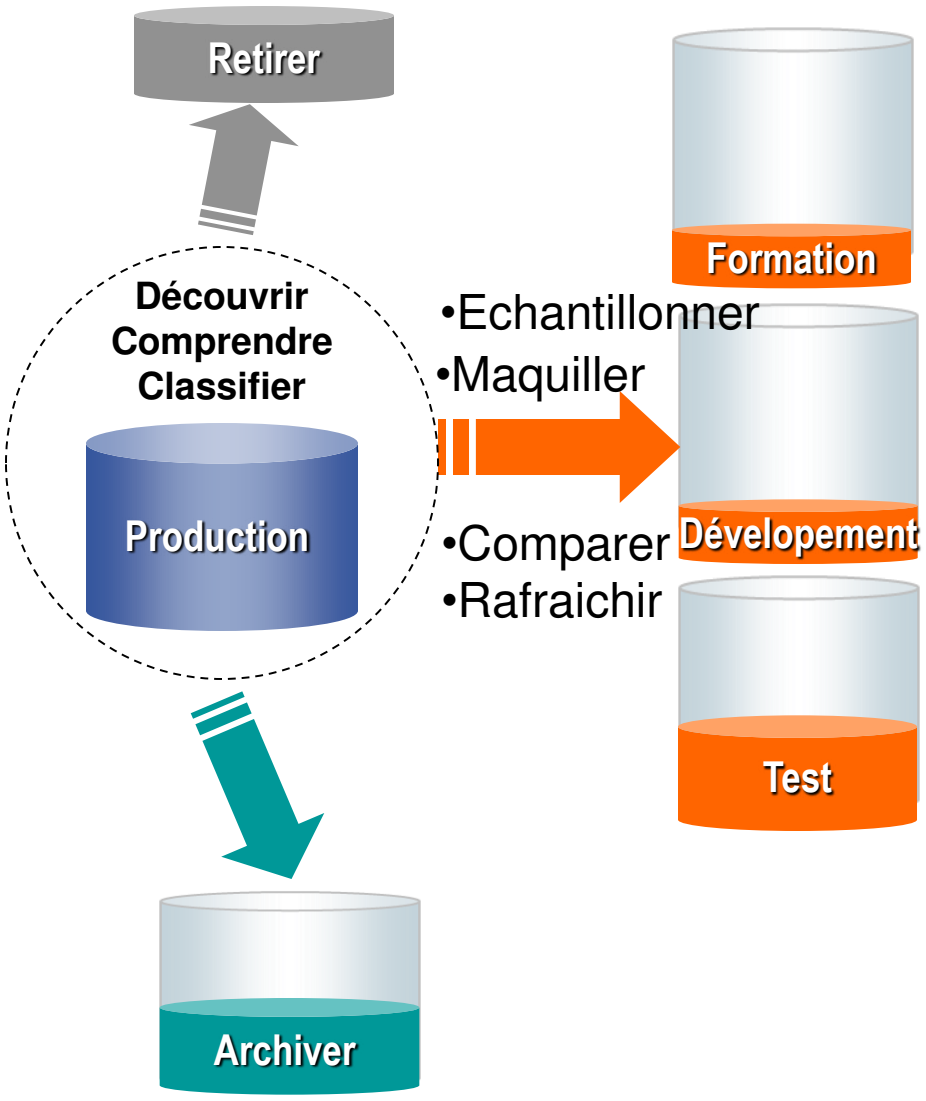
- Le Payment Card Industry Data Security Standard (PCI DSS), élaboré par MasterCard et Visa, et adopté également par American Express, JCB et Discover, interdit explicitement l'utilisation du numéro de compte réel pour le développement et les tests :
  - Cf requirement 6.3.4 du PCI DSS :
  - « **Production data (live PANs) are not used for testing or development** »
- Le PCI DSS prévoit des pénalités pouvant aller jusqu'à 500 000 dollars par incident.
- Le Responsable Sécurité des Systèmes d'Information d'une filiale française d'un grand groupe bancaire international estimait récemment que le risque encouru pour la filiale française sur ce seul aspect était de l'ordre de quelques centaines de millions d'euros.

AGENDA

- Introduction
- **Rappel de la Solution IBM : InfoSphere Optim**
  - Gestion des données de test
  - Anonymisation des données sensibles
- Nouveautés Masking on Demand
- InfoSphere Discovery : Découverte des données sensibles
- Conclusion

Les solutions InfoSphere Optim

Gérer le cycle de vie des données au travers d'environnements hétérogènes



Discover

- Accélérer les projets par la découverte des relations au travers de sources de données hétérogènes
- Découvrir les données sensibles à protéger et sécuriser

Test Data Management

- Rafraichir facilement et maintenir des environnements hors production à taille raisonnable
- Améliorer la qualité des applications et déployer de nouvelles fonctionnalités plus rapidement

Data Masking

- Protéger les informations sensibles d'une utilisation frauduleuse
- Eviter les fuites de données

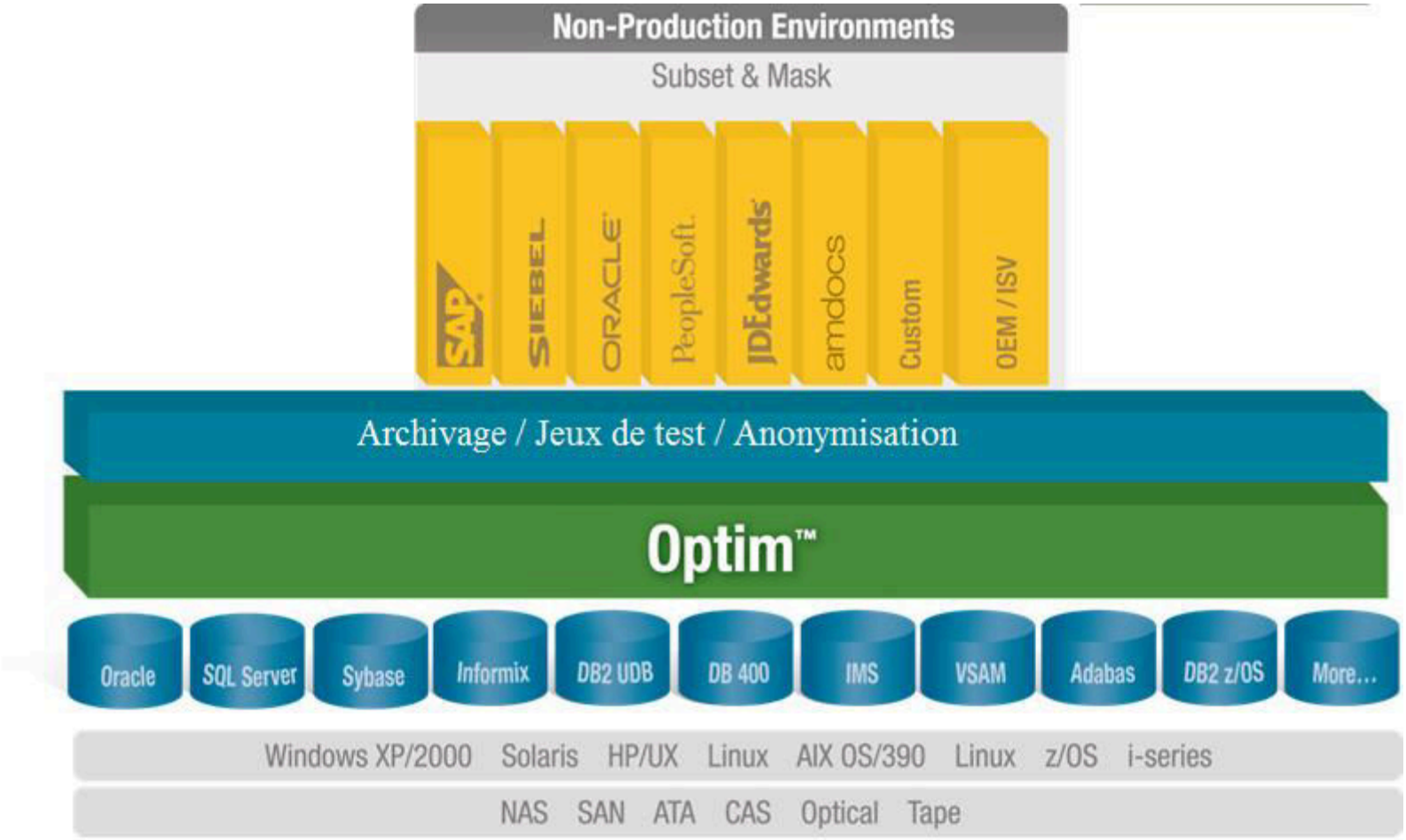
Data Growth Management

- Réduire le hardware, le stockage & les coûts de maintenance
- Rationaliser les upgrades applicatifs et améliorer la performance des applications

Application Retirement

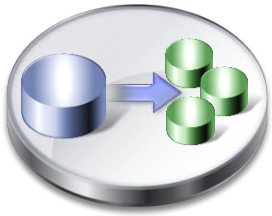
- Retirer en sécurité les systèmes historiques et applications redondantes en conservant les données
- Assurer l'accès aux données archivées indépendamment de l'application

OPTIM : une solution unique et multi-environnements



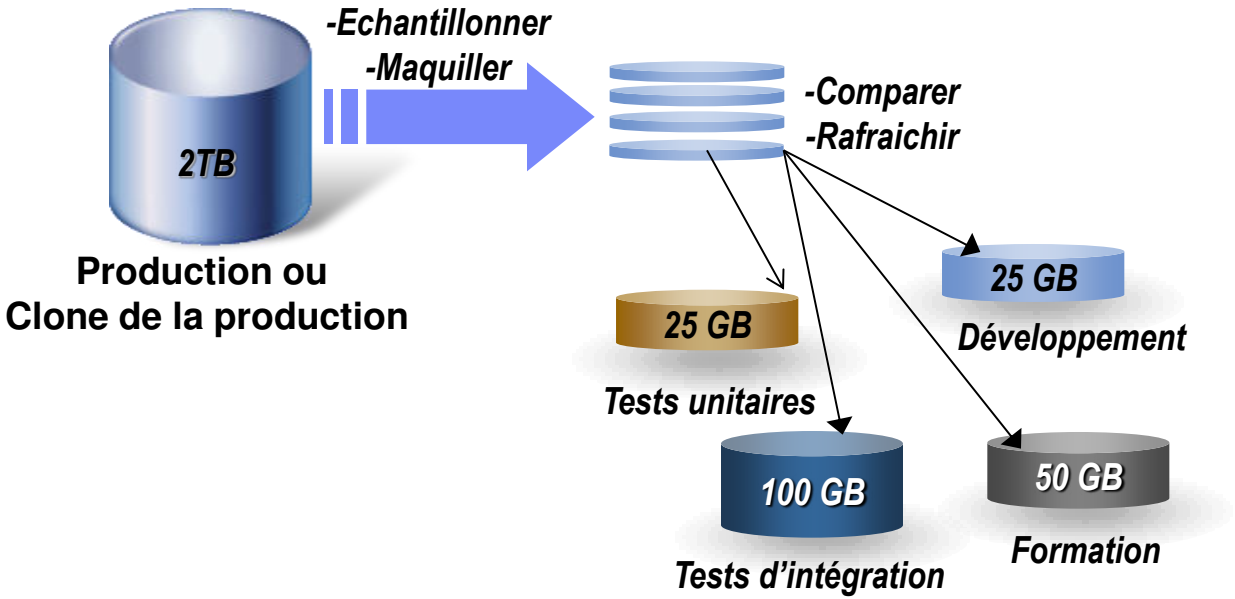
Optim fournit un **point de contrôle central** pour déployer des processus d'extraction, conservation, déplacement et protection des données de leur naissance à leur suppression.

InfoSphere Optim Test Data Management Solution



Test Data Management

Crée des jeux de données réduits à partir de la production pour le testing d'applications



Production ou Clone de la production

Besoins

- Créer des jeux de tests réduits et cohérents
- Automatiser la comparaison des résultats
- Protéger les données confidentielles utilisées hors production
- Raccourcir les cycles itératifs de testing et accélérer le "time to market"

Bénéfices

- Déploiement plus rapide et meilleure qualité
- Rafraichissement aisé des environnements
- Protège les données sensibles d'une utilisation frauduleuse

Optim Test Data Management supporte les environnements distribués et z/OS.

Fonctions standards de support des applications ERP/CRM comme :

ORACLE e-businesssuite | PeopleSoft | SIEBEL | JDEdwards Enterprise Software | SAP | Other

InfoSphere Optim Data Masking

Anonymiser les informations sensibles avec des données réalistes mais fictives pour les environnements de test et développement

Les données personnelles identifiables sont masquées avec des données fictives mais fonctionnellement valides

Exigences

- Protéger les données confidentielles utilisées dans les systèmes hors production
- Utiliser des techniques de maquillage éprouvées
- Supporter les réglementations sur la privatisation des données
- Support des solutions ERP/CRM

Bénéfices

- Protège les informations sensibles de la fraude
- Empêche les pertes d'informations
- Permet une meilleure gouvernance des données

Maquillage cohérent

Optim comprend un set de techniques d'anonymisation des données incluant :

- String literal values
- Character substrings
- Random or sequential numbers
- Arithmetic expressions
- Concatenated expressions
- Date aging
- Lookup values

Exemple 1

Customer Information

Cust ID	123456	**SSN**	333-22-4444		
Name	Erica Schafer				
Address	12 Murray Court				
City	Austin	**State**	TX	**Zip**	78704

Les données sont maquillées en conservant leur contexte pour garantir l'intégrité de l'information

Exemple 2

Customers Table

CustID	FirstName	LastName
10000	Jeanne	Renoir
10001	Claude	Monet
10002	**Pablo**	**Picasso**

Intégrité référentielle maintenue grâce à la propagation des clés

Orders Table

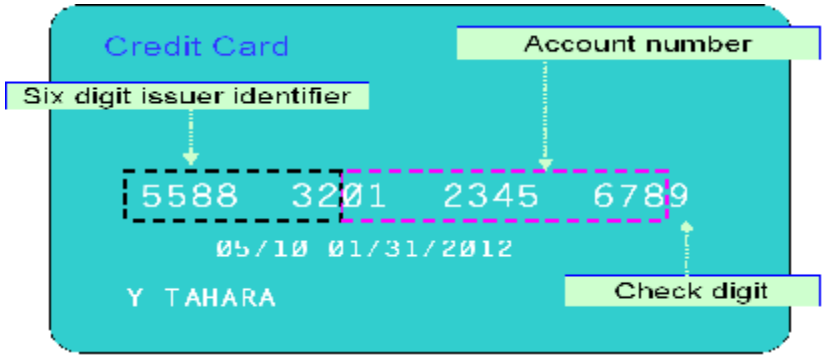
CustID	Item#	OrderDate
10002	86-5201	4-12-05
10002	94-3775	9-25-06

Les fonctions de maquillage

- Des algorithmes variés, paramétrables, non réversibles permettant de remplacer des valeurs réelles par des valeurs fictives réalistes et compréhensibles lors des tests :

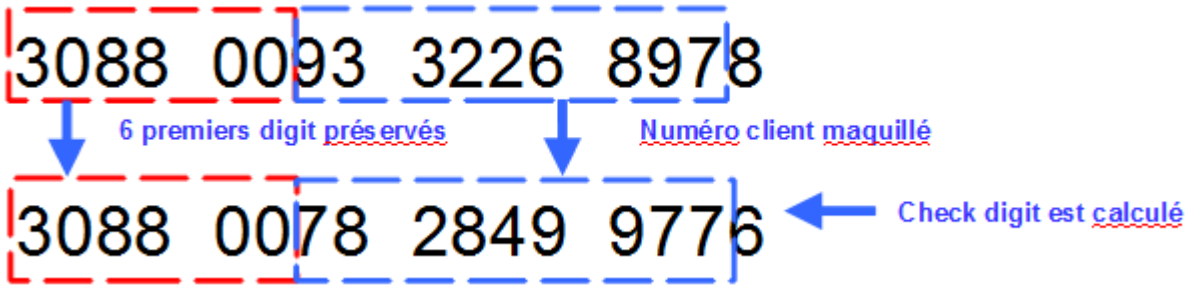
  - Hash_lookup mono ou multi-colonnes (ex : toutes les colonnes d'une adresse)
  - Random_lookup mono ou multi-colonnes : remplacement aléatoire
  - Lookup mono ou multi-colonnes : remplacement fixe
  - Shuffle : mélange et mode « battre les cartes »
  - Trans_CCN : génération de numéros de cartes de crédit (normes PCI)
  - Trans_email : modification d'adresse email (préfix, domaine…)
  - National ID : numéro de sécurité sociale (US, CA, FR, ES, IT, UK)
  - Age : calcul d'une nouvelle date quelque soit son format d'origine
  - Remplacement répétable/aléatoire de valeurs
  - Maintient de certaines valeurs d'origine de cas particuliers (Null, blanc…)
  - Etc…

Fonction : Credit Card Number

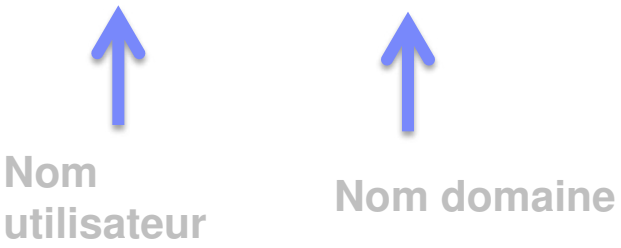


- **Sociétés supportées**
  - American Express, MasterCard, VISA and Discover
- **3 modes de maquillage supportés**
- Repeatable Masking mode
- Use 4 issuer digits mode
- Use 6 issuer digits mode:

Exemple – Use 6 issuer digits mode



Fonction : e-mail Address



- Maquillage de l'utilisateur, du domaine ou des 2
- Le Nom du domaine peut être généré aléatoirement ou récupéré d'une liste prédéfinie

- Exemples

James.lebas@fr.ibm.com	→	**hdcilkheuz@fr.ibm.com**	Maquillage user
James.lebas@fr.ibm.com	→	**James.lebas@yahoo.fr**	Maquillage domaine
James.lebas@fr.ibm.com	→	**jhaudfe@gdokza.com**	Génération des 2

Fonction : Hash Lookup

- **Ne génère pas de valeur aléatoire**
- **Récupère une colonne ou plusieurs à partir d'une table de référence**
- **Des fichiers d'exemples pour des tables de référence sont fournies :**
- Address – US, Canada, Germany, Spain, France, Italy, UK, Australia and Japan
- First name and Last name – US, Canada, Germany, Spain, France, Italy, UK, Australia and Japan
- Company name – US English
- **Vous pouvez créer vos propres tables de référence**

L'algorithme de hashing calcule une valeur par rapport à la table de référence

SEQ	CUST NO	FIRSTNAME	LASTNAME
1	49524	CHRISTINE	HAAS
2	01358	MICHAEL	THOMPSON
3	91841	SALLY	KWAN
4	69422	EVA	SPENSER
5	30137	VINCENZO	HENDERSON
6	59481	EILEEN	GEYER

Fonction : Repeatable Replacement

- **Renvoie toujours la même valeur de sortie par rapport à la valeur d'entrée**
- **Conserve le format d'entrée**
- **Les chiffres sont remplacés par des chiffres.**
- **Les caractères alphabétiques remplacés par des caractères alphabétiques (Minuscule & Majuscule)**
- **Les symboles ne sont pas remplacés**

- Exemples

9876#aBC → 0729#cUG

Fonction : Date Age Variance

- + ou - (jour, mois, année)
- Remplace seulement l'année par une année spécifique

2013/02/20	→	+ 1 An	→	**2014/02/20**
2013/02/20	→	- 1 mois	→	**2013/01/20**
2013/02/20	→	+ 10 jours	→	**2013/03/02**
2013/02/20	→	2015 Spécifique	→	**2015/02/20**

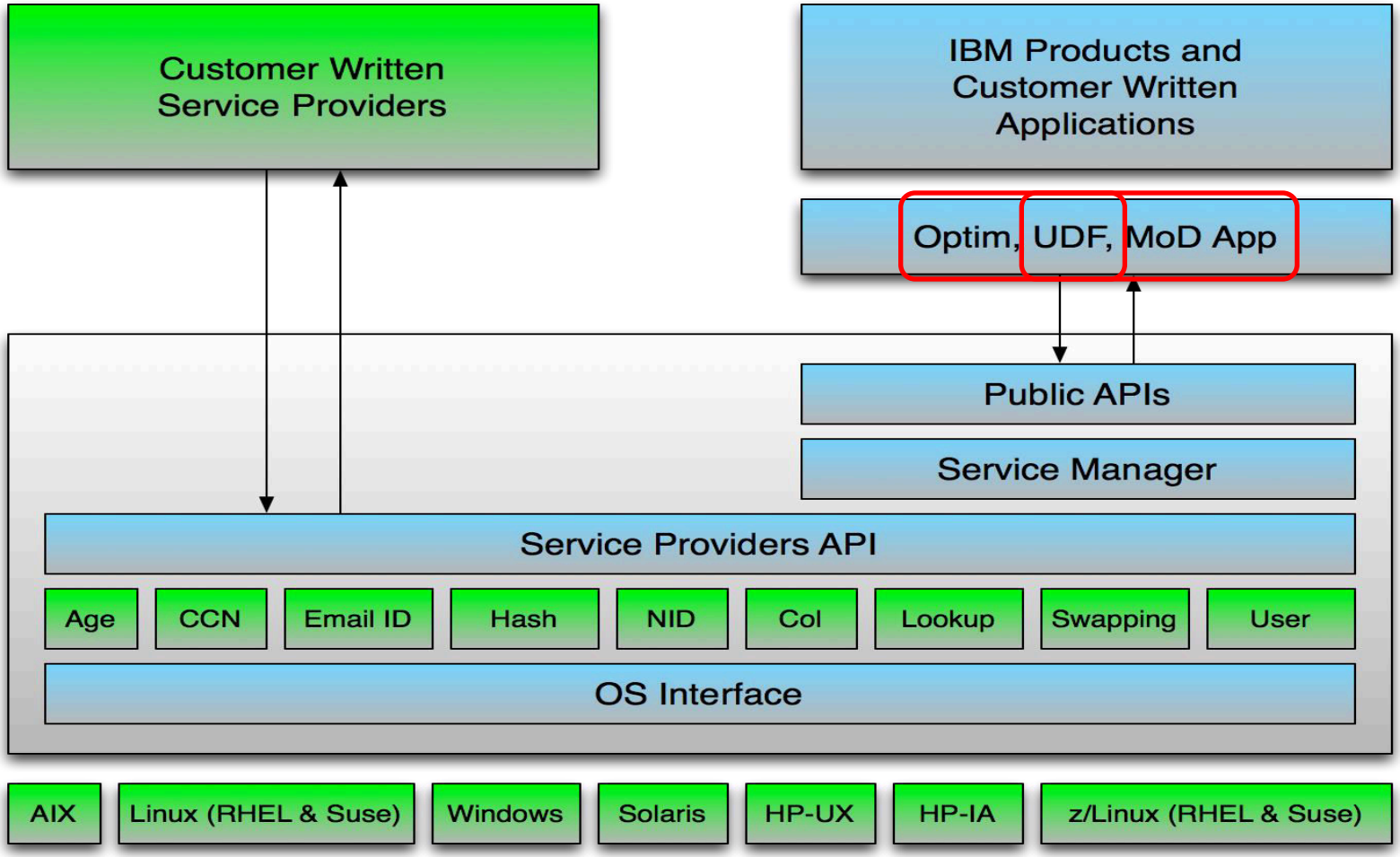
AGENDA

- Introduction
- Rappel de la Solution IBM : InfoSphere Optim
  - Gestion des données de test
  - Anonymisation des données sensibles
- **Nouveautés Masking on Demand**
- InfoSphere Discovery : Découverte des données sensibles
- Conclusion

Nouveautés : Masking on Demand

- **Jeux d'API fournissant des services de maquillage**
- **Librairies internes aux outils IBM & ouvertes aux applications tiers**
- **Multi Plateformes supportées**
- **Service de maquillage indépendant du type de source de données**

Architecture



User Defined Functions

- API directement intégré au langage SQL
- Permet de maquiller les données SGBDR avant que les données quittent la base.
- Maquillage sur place (sans besoin d'extraction)
- Même algorithmes utilisés par Optim quelque soit la plateforme ou le SGBDR.
- Optim 9.1.0.3 supporte DB2 z/OS, Teradata, Netezza, and Oracle

User Defined Functions

- **Create a new table by copying and masking from another**

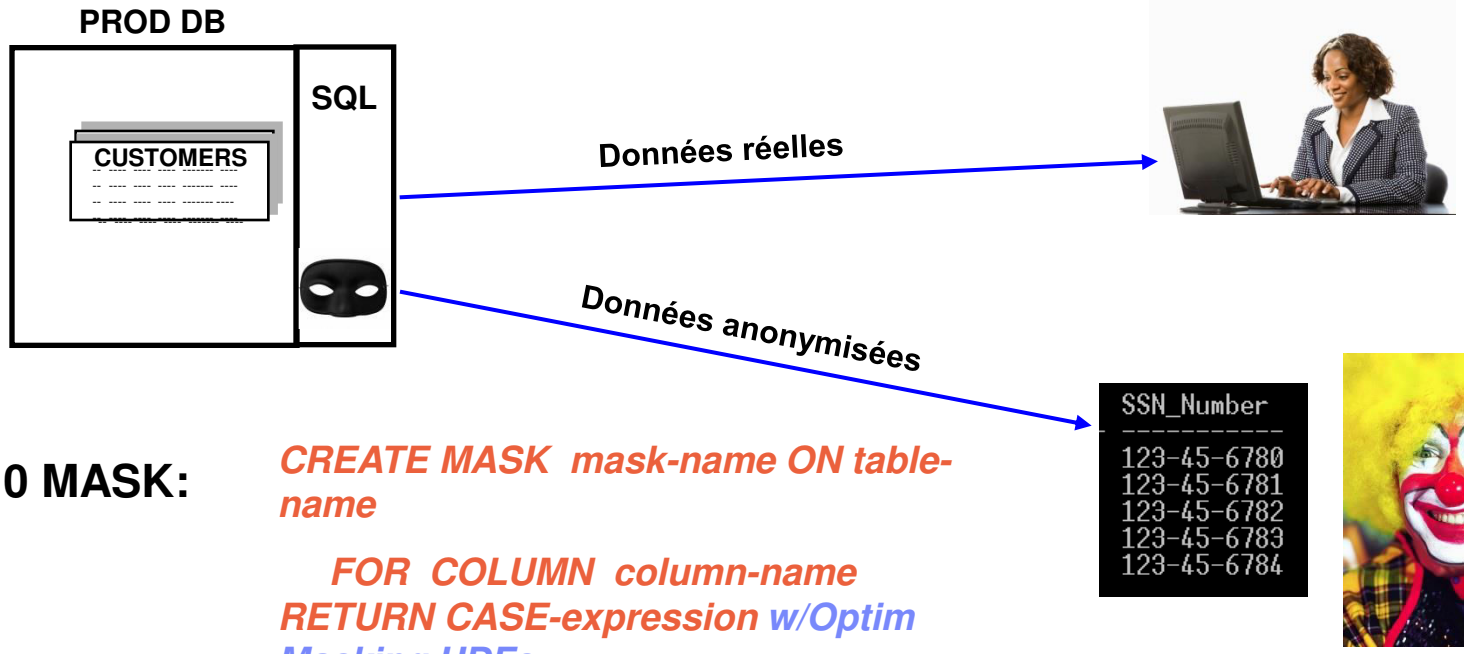
```
Create table oraccn_1m_mask (col1,col1mask) nologging as select visa  ,
OptimMaskStr_1(visa, 'pro=ccn,mtd=random, Flddef1=(name=col1,dt=char)')
from oraccn_1m;
```

- **Mask the existing credit card numbers in place**

```
update oraccn_1m
Set visa =  OptimMaskStr_1(visa, 'pro=ccn,mtd=random, Flddef1=(name=col1,dt=char)')
;
```

User Defined Functions

Maquillage à l'aide d'une vue



DB2 10 MASK:

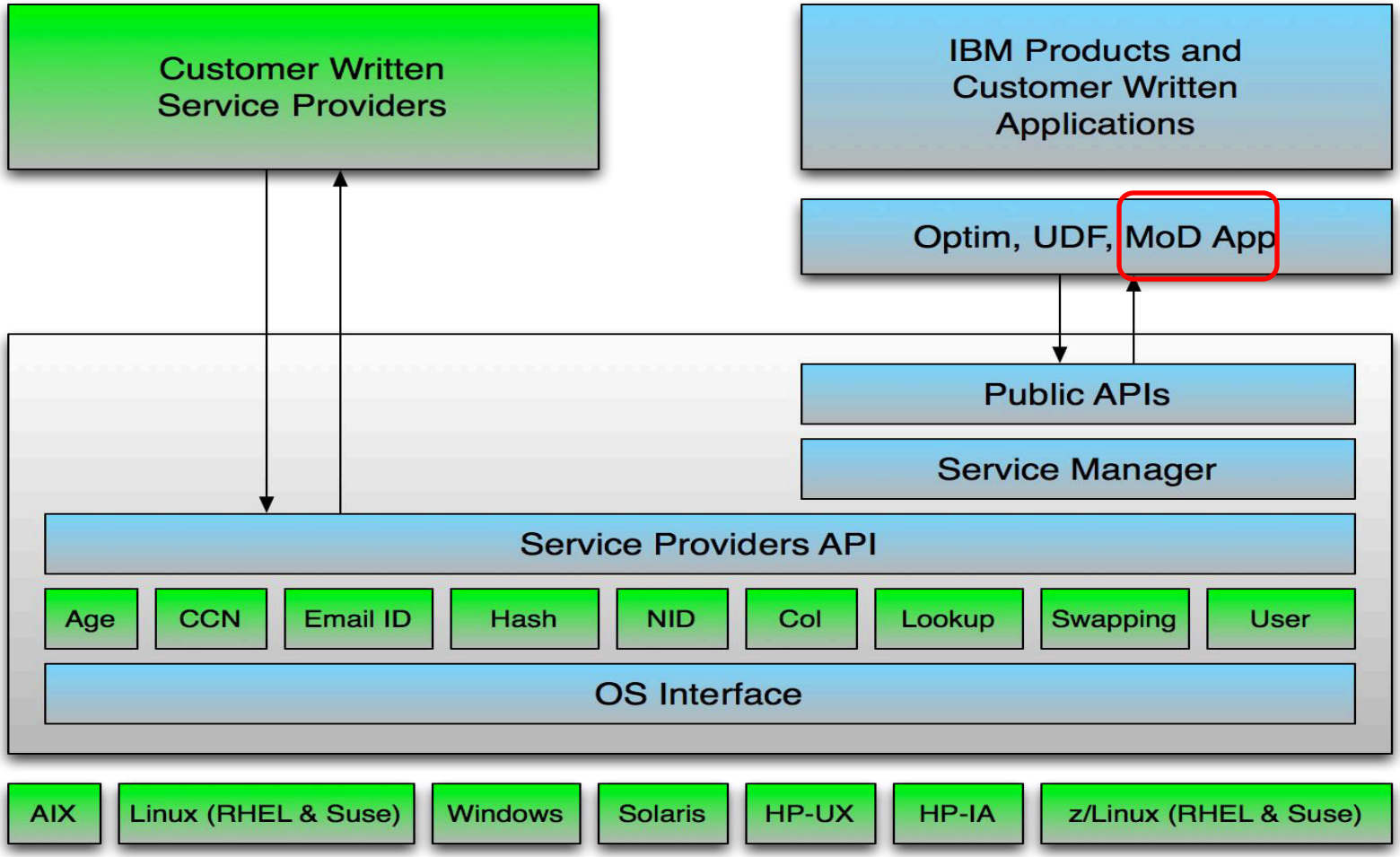
CREATE MASK mask-name ON table-name

FOR COLUMN column-name
RETURN CASE-expression w/Optim Masking UDFs

ENABLE;

DB2 MASK: A secure, Role-based column "view" containing consistent, repeatable dynamically masked data

Architecture



Masking on Demand Application

Sortie Juillet 2013

- Langage script de maquillage pour les sources suivantes :
  - Hadoop resources
  - Column Separated Value Files (CSV)
  - Fichier XML

- Fournira les mêmes algorithmes de maquillage (Optim & UDF)

- Syntaxe simple permettant de gérer :
  - File pattern
  - Alerte en cas d'anomalie

Masking on Demand Application -- Configuration File

Sortie Juillet 2013

```
{    "workitems": [ {
   "input": "file:///opt/IBM/MaskingOnDemand/TestData",
   "output": "file:///opt/IBM/MaskingOnDemand/results_new",
   "filepattern":"SingleCustomer.xml",
   "replace":true,
   "filetype":"XML",
   "delimiter":",",
   "parallelize":"NONE",
   "bulksize":0,
   "continueonerror":false,
   "nodes":[ {
         "path":"/customers/customer",
         "elements":[
            { "name":"EMAIL","path":"email_address",
               "mask":"PROVIDER=EML,FIELDS=[[EMAIL,WVARCHAR_SZ,70]],METHOD=MASK" },
            { "name":"CCN", "path":"ccn",
               "mask":"PROVIDER=CCN,FIELDS=[[*,WVARCHAR_SZ,70]],METHOD=MASK" },
            { "name":"STREET", "path":"address/street",
               "mask":"PROVIDER=AFF,FIELDS=[[*,WVARCHAR_SZ,70]],METHOD=MASK" }
         ] } ],
   "run":"MASKING"
```

Masking on Demand Application XML Masking Results

NEW

Sortie Juillet 2013

Before XML Document

```
<?xml version="1.0" encoding="utf-8"?>
<customers>
  <customer>
    <!-- All Valid and Present -->
    <first_name>Bobby</first_name>
    <middle_initial>J</middle_initial>
    <last_name>Fudge</last_name>
    <address>
      <street>100 Fifth Avenue</street>
      <city>New York</city>
      <state>NY</state>
      <zip>10014</zip>
    </address>
    <ccn>5411110000000017</ccn>
    <telephone>1-609-321-7654
    </telephone>
    <email_address> bobby1@yahoo.com
    </email_address>
  </customer>
</customers>
```

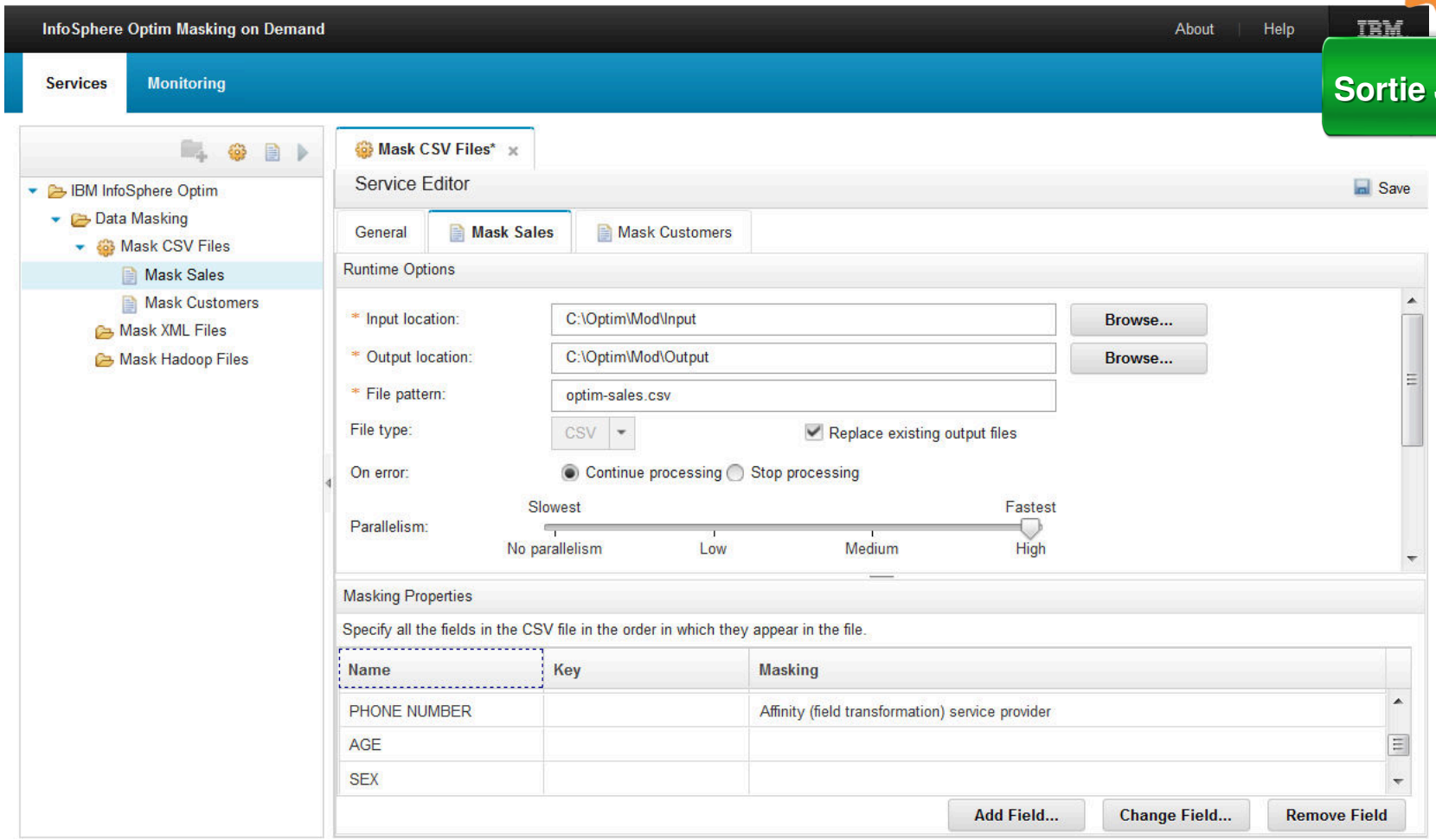
After XML Document

```
<?xml version="1.0" encoding="utf-8"?>
<customers>
  <customer>
    <!-- All Valid and Present -->
    <first_name>Bobby</first_name>
    <middle_initial>J</middle_initial>
    <last_name>Fudge</last_name>
    <address>
      <street>389 Ardhh Kymyla</street>
      <city>New York</city>
      <state>NY</state>
      <zip>10014</zip>
    </address>
    <ccn>5411116857029116</ccn>
    <telephone>1-609-321-7654
    </telephone>
    <email_address> email1@yahoo.com
    </email_address>
  </customer>
</customers>
```

Masking on Demand Application : User Interface

NEW

Sortie Juillet 2013



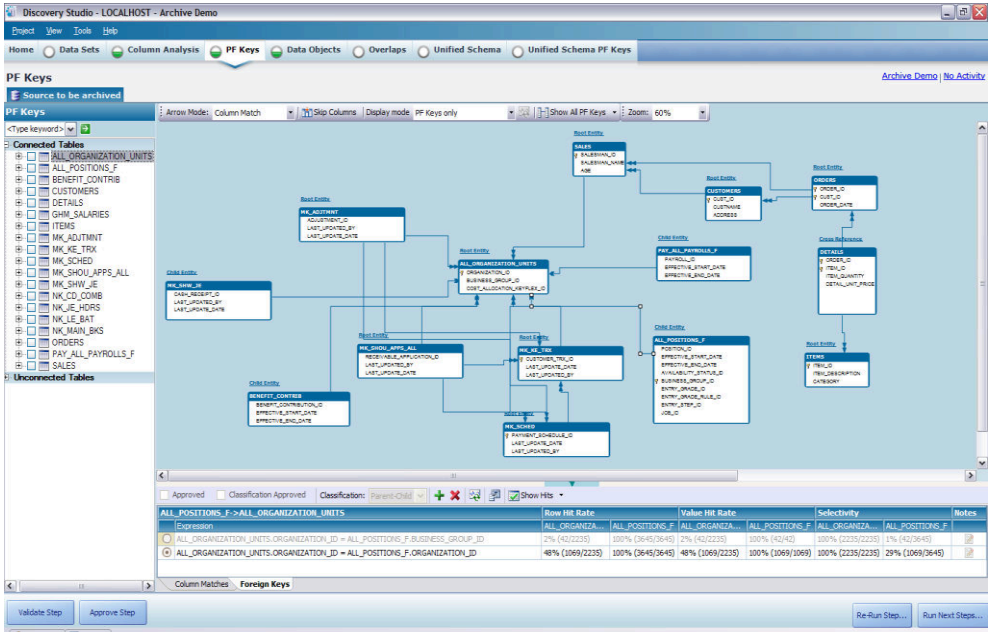
AGENDA

- Introduction
- Rappel de la Solution IBM : InfoSphere Optim
  - Gestion des données de test
  - Anonymisation des données sensibles
- Nouveautés Masking on Demand
- **InfoSphere Discovery : Découverte des données sensibles**
- Conclusion

InfoSphere Discovery

Découverte

Accélération du déploiement des projets grâce à l'automatisation de la reconnaissance de vos données



Exigences

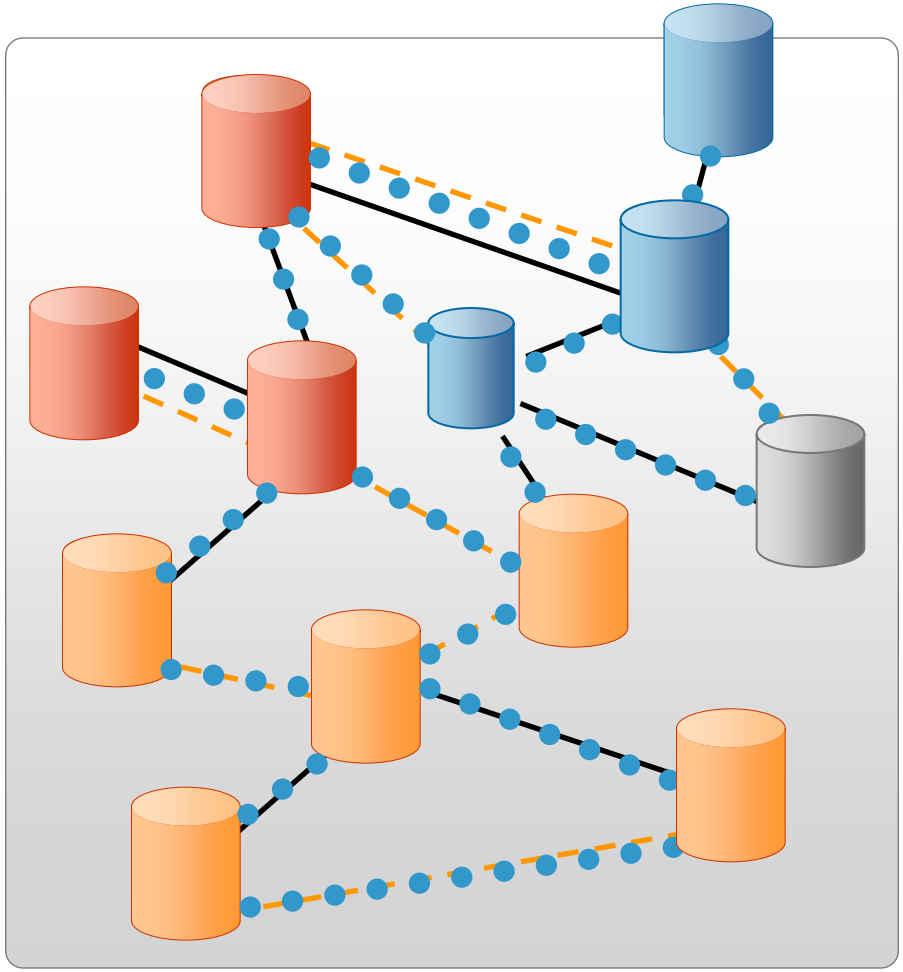
- Définition des objets métier pour les projets d'archivage et de test
- Reconnaissance des règles de transformation de données et des relations hétérogènes
- Identification des données sensibles à masquer à des fins de confidentialité

Avantages

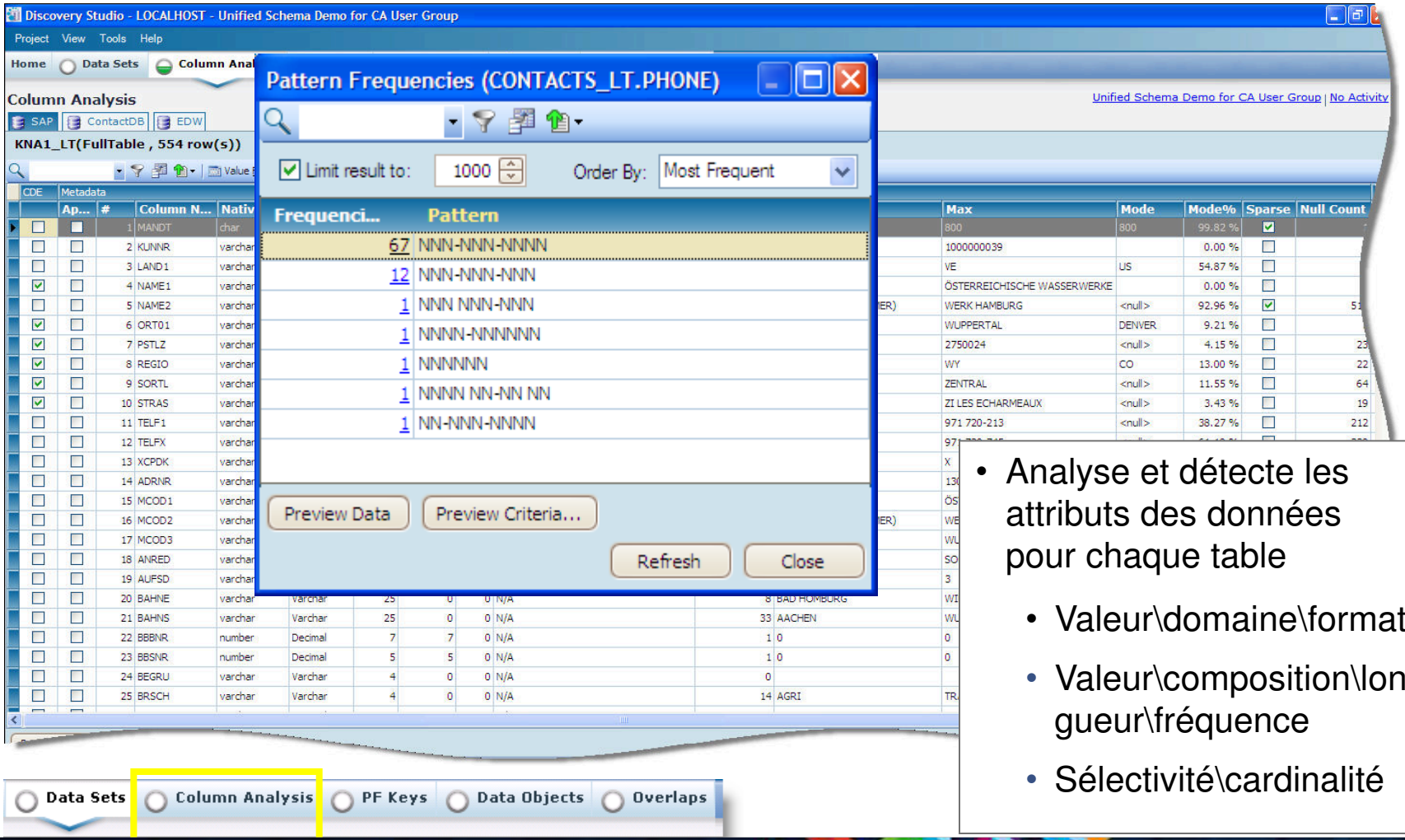
- Automatisation des activités manuelles permettant d'améliorer la durée de génération de valeur
- Découverte des relations entre les projets, permettant de diminuer les risques associés

Comprendre la distribution des données

- **IBM InfoSphere Discovery** automatise l'analyse des données et des relations pour une compréhension globale du capital de données
  - Identifie les ***relations*** qui relient les données individuelles en objets métiers
    - Clients, comptes, factures
  - Identifie la logique **complexe** qui relie les objets métier à travers plusieurs sources
  - Analyse basée sur les valeurs, formats

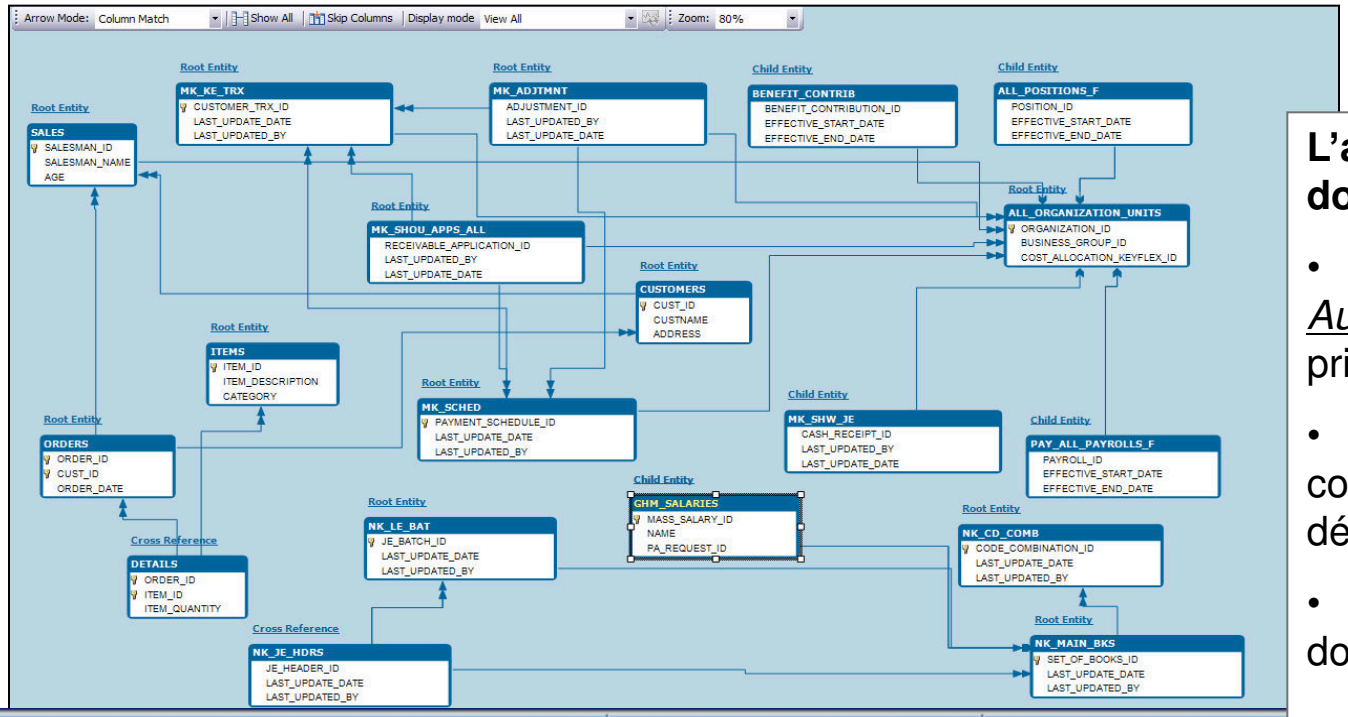


Analyse de colonnes



- Analyse et détecte les attributs des données pour chaque table
  - Valeur\domaine\format
  - Valeur\composition\longueur\fréquence
  - Sélectivité\cardinalité

Découverte des clés primaires/étrangères



HQ_EMPPERS->HQ_EMP	Row Hit Rate		Value Hit Rate	
Expression	HQ_EMP	HQ_EMPPERS	HQ_EMP	HQ_EMPPER
HQ_EMP.EMPLOYEE_ID = HQ_EMPPERS.EMPID	89% (223/250)	97% (223/230)	89% (223/250)	100% (223/

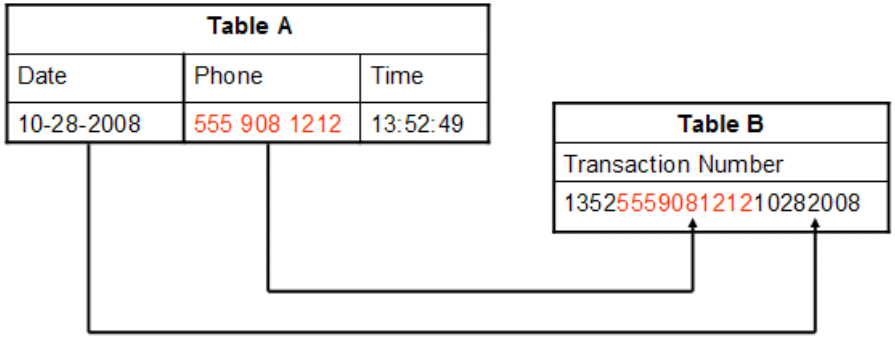
L'analyse du contenu des données permet :

- La découverte *Automatique* des clés primaires/étrangères
- Des statistiques complètes sur les clés découvertes
- La visualisation des données
- Les lignes manquantes, les clés orphelines

Data Sets | Column Analysis | PF Keys | Data Objects | Overlaps

Identifier les données sensibles

- Certains types de données sensibles sont faciles à reconnaitre mais d'autres sont cachés
  - Agrégés avec d'autres éléments dans une ligne
  - Décomposés et répartis sur plusieurs colonnes
  - Insérés dans un commentaire ou dans des zones de texte

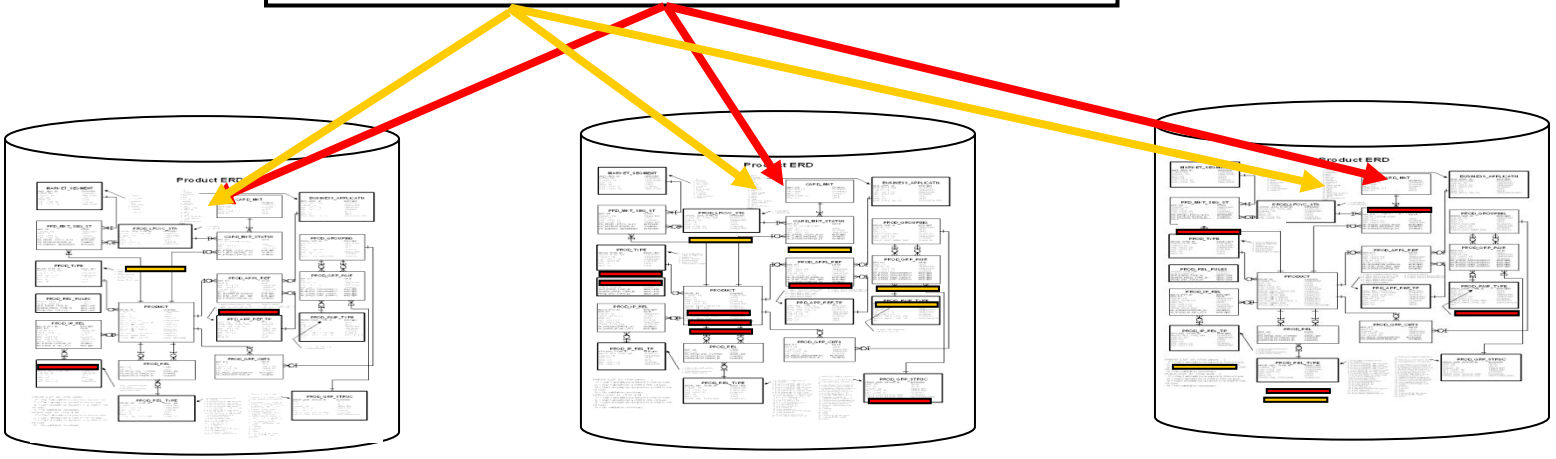


Les instances de données sensibles cachées peuvent présenter un risque de conformité

Découverte des données sensibles

Dictionnaire des données sensibles

Row	Member	SS #	Age	Phone	Sex
1	595846226	123-45-6789	15	(123) 456-7890	M
2	567472596	138-27-1604	8	(138) 271-6037	F
3	540450091	154-86-4196	22	(154) 864-1961	M
4	514714372	173-44-7900	55	(173) 447-8996	F
5	490204164	194-26-1648	4	(194) 261-6476	F
6	466861109	217-57-3046	66	(217) 573-0453	M
987,623	444629628	243-68-1812	25	(243) 681-8107	F
987,624	423456789	272-92-3629	87	(272) 923-6280	M



- Rechercher les données sensibles sur chaque système est long et complexe
- Tout ou partie de ces données se retrouvent dans des centaines de tables et colonnes

AGENDA

- Introduction
- Rappel de la Solution IBM : InfoSphere Optim
  - Gestion des données de test
  - Anonymisation des données sensibles
- Nouveautés Masking on Demand
- InfoSphere Discovery : Découverte des données sensibles
- **Conclusion**

Bénéfices de la solution IBM InfoSphere Optim

- Suite complète d'anonymisation et Test Data Management

- Nombreuses fonctions de **maquillage** des données sensibles

permettant de **réduire le risque** de fuite de données et **être en conformité** avec les législation dans ce domaine.

- Conserve la cohérence des données pour des tests de qualité

- Accélérateurs pour les phases de découverte des données sensibles (Option InfoSphere Discovery) inclus dans le package Enterprise

IBM SolutionsConnect 2013

L'IBM TechSoftware nouvelle génération

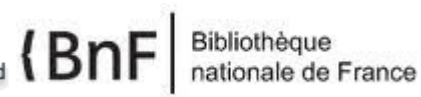
Quelques clients Infosphere Optim en France

Crédit du Nord

Children Worldwide Fashion

BNP PARIBAS
La banque d'un monde qui change

l'Assurance Maladie



Annexe – Quelques sites web utiles

- Information Management France : http://www-01.ibm.com/software/fr/data/
- Evénements IBM France : http://www-05.ibm.com/fr/events/
- IBM TEC France : http://www-05.ibm.com/fr/events/tec/
- InfoSphere Optim : http://www-01.ibm.com/software/data/optim/protect-data-privacy/
- Suivez toute l'actualité sur :
  - Facebook : IBM Information Management France
  - LinkedIn : IBM Information Management France
  - Viadéo : IBM Information Management France

IBM SolutionsConnect 2013

L'IBM TechSoftware nouvelle génération

MERCI