L'anonymisation de données en masse chez Bouygues Telecom

Jean-Luc Lambert (REACTIV'Conseil)
Patrick Chambet (Bouygues Telecom)

JSSI 2011

ENTREPRISES EN CRÉATION
JEUNES ENTREPRISES INNOVANTES

Sommaire

- Constat: tester sur la donnée de production et devenu indispensable
- Problématiques rencontrées pour tester sur de la donnée de production
- Les solutions: méthodes d'anonymisation de données et mise en place d'une usine d'extraction de données
- Présentation de l'usine d'extraction de données de Bouygues Telecom
- Les grands principes de fonctionnement de l'usine
- Bilan

Constat (1/2)

- **La donnée production est indispensable à la qualification des produits car…**
  - **Les cas fonctionnels de production ont une complexité impossible à reproduire par création de données**
    - **Problématique de clients ayant vécu (10 ans, 20 ans, vieilles offres, …)**
    - **Problématique des incohérences liées à la multiplicité des points d'entrée (chargé de clientèle, internet, …)**
    - **Problématique des incohérences liées à des incidents en production ou à des corrections en production**
  - **La volumétrie de production ne peut être reproduite qu'à partir de données de production**

Constat (2/2)

- **La donnée production est indispensable à la qualification des produits car…**
  - **L'entretien de clients de test cohérents de bout-en-bout est un cauchemar sur un banc de test**
    - **Les KOs applicatifs cassent les données**
    - **On ne peut entretenir la donnée de test comme on entretient la donnée de production**
    - **Le test des systèmes transverses (obligations légales, Data Warehouse, fraude, …) nécessite des données cohérentes de bout-en-bout**

Problématiques (1/4)

- **Comment assurer une sortie de données conforme aux exigences ?**
  - **De la production**
    - **Contrôle des accès en production (respect de l'access management d'ITIL)**
    - **Contrôle de l'accès aux données (seules les personnes habilitées peuvent accéder aux données des clients)**
  - **De la sécurité**
    - **Les données personnelles ne doivent pas sortir de la zone sécurisée de production**
  - **Du business**
    - **Les données commerciales (clauses contractuelles, …) ne doivent pas être divulguées**

Problématiques (2/4)

- **Rappel des exigences concernant la protection des données personnelles**
  - Loi Informatique et Libertés (1978, modifiée en 2004)
    - **Art. 34: obligation de protection des données**
    - **Art. 36: obligation de purge des données**
    - **Aspect répressif: cf code pénal**
  - Art. 226-22 du code pénal
    - **" Le fait, par toute personne qui a recueilli (...) des données à caractère personnel (…), de porter, sans autorisation de l'intéressé, ces données à la connaissance d'un tiers qui n'a pas qualité pour les recevoir est puni de cinq ans d'emprisonnement et de 300 000 Euros d'amende.**
    - **La divulgation prévue à l'alinéa précédent est punie de trois ans d'emprisonnement et de 100 000 Euros d'amende lorsqu'elle a été commise par imprudence ou négligence."**
  - Art. 226-20 du code pénal
    - **"Le fait de conserver des données à caractère personnel au-delà de la durée prévue par la loi ou le règlement, par la demande d'autorisation ou d'avis, ou par la déclaration préalable adressée à la CNIL, est puni de cinq ans d'emprisonnement et de 300 000 Euros d'amende"**

Problématiques (3/4)

- **Rappel des exigences concernant la protection des données personnelles**
  - Directive européenne 95/46/CE (1995)
    - **"Le responsable du traitement doit mettre en œuvre les mesures appropriées pour protéger les données à caractère personnel contre la perte accidentelle, l'altération, la diffusion ou l'accès non autorisé"**
  - Norme PCI-DSS
    - **3.2 - Si des données d'authentification sensibles sont reçues et supprimées, obtenir et passer en revue les processus de suppression des données pour vérifier que ces dernières sont irrécupérables**
  - Norme ISO 27001
    - **A.10.7.3 - Procedures for the handling and storage of information shall be established to protect this information from unauthorized disclosure or misuse**
    - **A.12.4.2 - Test data shall be selected carefully, and protected and controlled**
    - **A.12.5.4 - Opportunities for information leakage shall be prevented**
    - **A.15.1.4 - Data protection and privacy shall be ensured as required in relevant legislation, regulations, and, if applicable, contractual clauses**

Problématiques (4/4)

- **Rappel des exigences concernant la protection des données personnelles**
  - A venir
    - **« Paquet Télécom » (2009)**
      - L'une de ses 4 directives (2002/58/CE) concerne la protection des données (obligation d'information des personnes en cas d'atteinte à leurs données)
      - Doit être transposé dans les droits nationaux des états membres d'ici mai 2011
      - Après plusieurs tentatives (cf Détraigne - Escoffier), devrait être transposé par ordonnance à l'initiative du Gouvernement (art. 34 I&L ?)
    - **Projet de loi Détraigne - Escoffier:**
      - Notification obligatoire à la CNIL et aux personnes des incidents de sécurité lorsqu'ils ont un impact sur des données personnelles
      - Votée en 1ère lecture par le Sénat début 2010 puis abandonnée

Les solutions (1/3)

- **Les méthodes d'anonymisation de données**
  - Suppression totale (d'une colonne)

    `Mme Jeanne DUPONT` ➔ `Mme Jeanne <null>`

    - **Si les données personnelles ne sont pas utiles dans le cadre de la prestation**
    - **Parfois complexe (contraintes d'intégrité)**
  - Mise à blanc (de champs)

    `Mme Gladys PARUE` ➔ `Mme _ _`

    - **Simple, mais perte de cohérence**
  - Masquage / troncature partielle
    - `M. Jethro MALODO` ➔ `M. Je MAL`
    - **Simple mais effets variables sur la cohérence**
    - **Dangereux, car information parfois reconstituable par recoupement**

Les solutions (2/3)

- **Les méthodes d'anonymisation de données (suite)**
  - Hashage, chiffrement

    `M. Jean BON` ➔ `M. 31337 AF76AF585H87AF23DEADBEEF`

    - **Clé à protéger ou à jeter**
    - **Cohérence possible, mais taille des champs modifiée**
    - **Modification des clés indexées: certaines recherches deviennent impossibles**
  - Substitution

    `M. Nicolas SYRKOZA` ➔ `M. Jean DUPONT`

    `M. Jean TANT-NALLU` ➔ `M. Jean DUPONT`

    - **Aléatoire ou non (ex: vieillissement)**
    - **Perte de cohérence et d'homogénéité en général**

Les solutions (3/3)

- **Les méthodes d'anonymisation de données (suite)**
  - **Renumérotation**

    `M. Jean-Kevin QUIDIE` ➔ `M. Abcd-Abcde ABCDEF`

    - **Cohérence conservée, conservation de la longueur des données**
    - **Perte de la qualité des données**
  - **Permutation**

    `Mme Agathe THEBLUES` ➔ `M. Miles DAVIS`

    - **Cohérence possible, conservation du niveau de qualité des données**

➔ **Chez Bouygues Telecom, utilisation conjointe de plusieurs de ces méthodes**

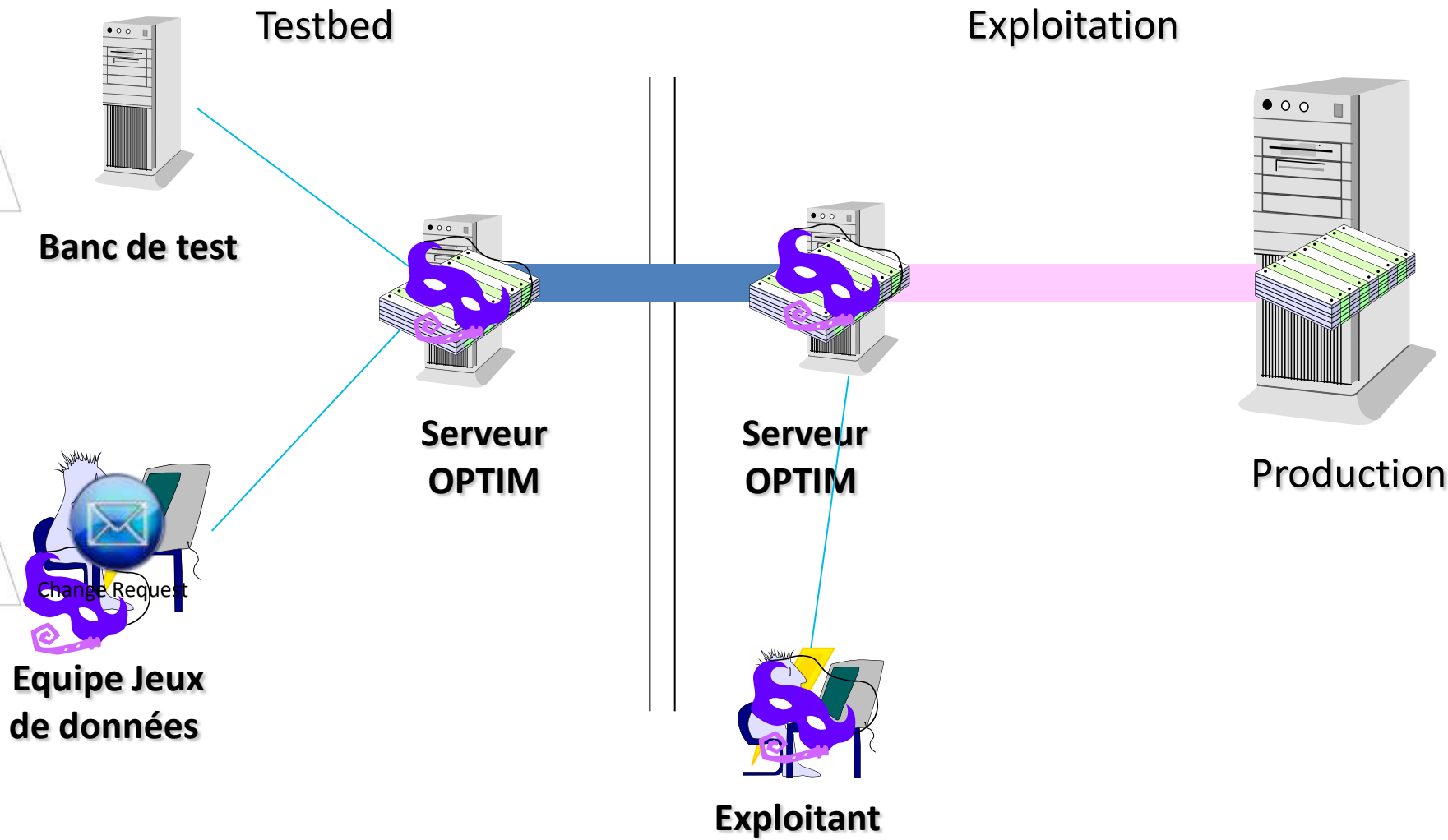
L'usine à jeux de données (1/2)

- **Mise en place d'une usine d'extraction de données de test permettant:**
  - De faire réaliser l'extraction des données de production par l'exploitant
    - **Conformité avec l'access management d'ITIL**
    - **L'exploitant est seul habilité à accéder à la production**
  - D'anonymiser les données
    - **Ce qui permet ensuite de les sortir du domaine de production, voire de les fournir à un tiers (intégrateur, éditeur pour debug, …)**
  - De charger les données extraites et anonymisées sur les bancs de test
    - **Ce qui sécurise le chargement et permet d'assurer la QoS du service d'anonymisation**

L'usine à jeux de données (2/2)

- **Chez Bouygues Telecom, l'usine à jeux de données comporte:**
  - Un outil d'extraction et d'anonymisation de données (OPTIM d'IBM) installé en production
  - Un outil de chargement de données sur les bancs de test (OPTIM aussi)
    - **Sur un autre serveur pour assurer l'étanchéité**
  - Un référentiel contenant
    - **La cartographie des attributs sensibles avec leur traitement d'anonymisation**
  - Une équipe qui
    - **Maintient à jour le référentiel**
    - **Pilote les extractions réalisées par l'exploitant**
    - **Charge les bancs de test avec les données**

Le fonctionnement de l'usine à jeux de données chez Bouygues Telecom



Les grands principes (1/3)

- **L'anonymisation conserve les fonctionnalités des données**
  - Cohérence des champs de jointure (réalisée par OPTIM)
  - Les valeurs anonymisées doivent être fonctionnelles (pas de 01 53 32 XX XX)
- **L'anonymisation est irréversible**
  - Il n'y a pas de règle de calcul qui puise être inversée: on ne peut pas retrouver la donnée initiale
- **Si l'anonymisation des données concerne plusieurs bases ou des fichiers de données l'anonymisation est cohérente sur tout le périmètre**
  - L'utilisation de tables de translation permet de mémoriser l'anonymisation
  - Les tables de translation sont conservées chiffrées en production

Les grands principes (2/3)

- **Les cartographies d'anonymisation sont maintenues hors de la production**
  - Les bancs de tests du testbed fournissent les informations sur les évolutions fonctionnelles des bases de données qui vont arriver en Production
- **Les fonctions d'anonymisation sont mises au point et testées en pré-prod**
  - L'équipe a les mêmes accès en pré-prod qu'un analyste de test de pré-prod
  - Le même serveur permet d'extraire de Production ou de pré-prod
- **Les méthodes et cartographies d'anonymisation sont publiées et disponibles pour les acteurs de test**

Les grands principes (3/3)

Attribut	Exemple de transformation
AdresseDeclaree.boitePostale	Préfixe à 'BP ' + renumérotation sur 7 digits
AdresseDeclaree.complementAdr1	Préfixe à 'APT ' + renumérotation sur 7 digits
AdresseDeclaree.complementAdr2	Préfixe à 'APT ' + renumérotation sur 7 digits
AdresseDeclaree.cp	Sans anonymisation
AdresseDeclaree.num	renumérotation
AdresseDeclaree.rue	Préfixe à 'RUE ANONYME' + renumérotation sur 7 digits
AdresseDeclaree.ville	Sans anonymisation
AdresseElectronique.eMail	Préfixe "email" + renumérotation sur 6 digits + suffixe "@bouyguestelecom.fr"
AdresseNormalisee.ligne1	Deduit des translations sur les champs métiers élémentaires
AdresseNormalisee.ligne2	Deduit des translations sur les champs métiers élémentaires
CarteBancaire.noCarte	renumérotation sur 16 digits. Pas de recherche de compatibilité bancaire
ClientPayeur.identifiantFonctionnelClient	Conservation du préfixe [123456NK]. + renumérotation respectueuse des règles métiers (8 digits pour le prefixe '1', 2 digits pour le prefixe '2', …)
CoordonneesTelephoniques.noTel	Conservation du préfixe (0033|+33|33|0|aucun), des 5 digits suivants + renumérotation des 4 derniers digits. Inchangé pour les numéros courts.
Entreprise.noSiren	renumérotation sur 9 digits
Entreprise.raisonSociale	Préfixe à 'raisonSociale' + renumérotation sur 6 digits
Individu.civilite	Sans anonymisation
Individu.nom	Préfixe à 'NOM' + renumérotation sur 6 digits
Individu.prenom	Préfixe à 'PRE' + renumérotation sur 6 digits
JusticatifIdentite.numero	Préfixe à 'PID' + renumérotation sur 7 digits
LigneGsm.msisdn	Conservation du préfixe (0033|+33|33|0|aucun), des 5 digits suivants + renumérotation des 4 derniers digits. Inchangé pour les numéros courts.
SimLogique.imsi	Conservation des 11 premiers digits + renumérotation des 4 derniers digits
SimPhysique.iccid	Sur forme '893320[0-9]{13}', conservation des 15 premiers digits + renumérotation des 4 derniers digits
TerminalMobile.imei	Conservation des 10 premiers digits, renumérotation des 4 digits suivants, recalcul du digit de contrôle pour compatibilité avec l'algorithme de Luhn

Bilan

- **La mise en place de l'usine d'extraction a permis de rassurer tous les acteurs**
  - La Sécurité SI qui a l'assurance que les données ne sont pas diffusées à l'extérieur
  - La Production qui maîtrise le processus d'extraction et peut garantir le non-impact
  - Les MOEs qui n'ont pas à se préoccuper de l'anonymisation et la voient comme un surcoût raisonnable (quelques j.h)
- **L'usine est devenue centre d'expertise et référent pour les sorties de données de production**
  - Elle est consultée sur chaque sortie de données et préconise l'anonymisation à réaliser
- **L'usine a un coût de fonctionnement raisonnable**
  - 3 personnes financées aux 2/3 par des projets

Questions

Pour aller plus loin

- **Les bonnes pratiques Informatique et Libertés de protection des données**
  - http://www.cnil.fr/vos-responsabilites/vos-obligations/
- **Anonymisation des données personnelles : l'AFCDP propose un référentiel**
  - http://www.securityvibes.com/community/fr/blog/2008/05/23/anonymisation-des-donn%C3%A9es-personnelles-lafcdp-propose-un-r%C3%A9f%C3%A9rentiel
- **Législation européenne: protection des données**
  - http://europa.eu/legislation_summaries/information_society/l24120_fr.htm