ABIT 2012

Skúsenosti z implementácie a prevádzky systému anonymizácie citlivých údajov v komplexnom prostredí

Roman Pavčo

Zažime to spolu

Kategorizácia údajov v telco prostredí – citlivé údaje

Citlivé údaje sú všetky informácie, okrem tých, ktoré sú kategorizované z pohľadu dôvernosti ako „Verejné“.
Sú to najmä, no nie výlučne informácie chránené legislatívou SR.

• **Osobné údaje** – definované zákonom č. 428/2002 Z.z. o ochrane osobných údajov – údaje identifikujúce konkrétnu osobu alebo spojiteľné s konkrétnou osobou všobecného charakteru (zákaznícke dáta). Riziko pokuty!

• **Prevádzkové údaje** – definované zákonom č. 351/2011 Z.z. o elektronických komunikáciách
§ 57 - Prevádzkové údaje a lokalizačné údaje
(1) Prevádzkové údaje sú údaje vzťahujúce sa na užívateľa a na konkrétny prenos informácií v sieti a vznikajúce pri tomto prenose, ktoré sa spracúvajú na účely prenosu správy v sieti alebo na účely fakturácie.
(2) Lokalizačné údaje sú údaje spracúvané v sieti alebo prostredníctvom služby, ktoré označujú geografickú polohu koncového zariadenia užívateľa verejnej služby.
Do tejto kategórie patria aj údaje špecifické pre zákazníka – údaje spojiteľné s osobou súvisiace s poskytovanými službami, špecifickými pre konkrétneho zákazníka. Riziko pokuty!

• **Biznis údaje** – daňové tajomstvo, platové pomery zamestnancov, pohyby na bankových účtoch, obchodné informácie, atď.

Anonymizácia dát – požiadavky / problémy / potreby

Problémy

- veľký počet testovacích databáz
- rôzni dodávateľia
- fluktuácia
- neexistencia registra testovacích databáz
- testovacie databázy sú kópiou produkčných databáz
- „projektové“ databázy
- citlivé informácie sú mimo kontroly spoločnosti
- neexistencia jednotlivých procesov
- nejasné zodpovednosti
- databázové linky do produkčných databáz

Požiadavky

- zachovanie integrity medzi systémami
- zachovanie pohlavia fyzických osôb
- názov firmy (FOP) a priezvisko (FON) sa nachdázajú v rovnakom poli
- dátová kvalita (čistota dát)
- rýchlosť a flexibilita tvorby anonymizovaných databáz

Potreby

- Security roadmap (stratégia) schválená EMB – časť Data Security
- definovať jednotný proces v rámci anonymizácie
- zodpovednosť Info security manažéra za implementáciu procesu
- jasne definovaná podpora menežmentu a CIO

Anonymizácia dát – riešenie problému

Implementácia systému anonymizácie citlivých údajov

- Procesná
- Technická
- Organizačná

Anonymizácia dát – procesná časť 1

Pracovný postup pre anonymizáciu testovacej databázy

Definovanie techník anonymizácie

Technika anonymizácie	Popis	+	-
Nulovanie údajov (Nulling out)	Zmazanie dát alebo ich nahradenie nulovou hodnotou, prípadne inou hodnotou.	Jednoduché, efektívne	Ojedninelá využiteľnosť
Maskovanie údajov (Masking data)	Nahradenie vybraných polí maskovacím znakom (napr. 790818XXXX). Maskovací znak odstráni podstatnú časť záznamu.		
Substitúcia údajov (Substitution)	Náhodné nahradenie obsahu stĺpca údajov informáciou podobného významu. Napr. náhodné zamenenie priezvisk v databáze zákazníkov za priezviská z dodstatočne veľkého zoznamu náhodne vygenerovaných priezvisk.		
Premiešanie záznamov (Shuffling records)	Rovnaké ako substitúcia údajov, avšak náhradné údaje sú odvodzované priamo z príslušného anonymizovaného stĺpca.		
Číselný rozptyl (Number variance)	Algoritmus modifikuje jednotlivé údaje náhodným percentuálnym podielom jeho skutočnej hodnoty. Hodnoty môžu byť pozmeňované napr. s 10% podielom.	Len pre číselné údaje	Potreba použiť aj iný typ techniky
Vygenerovanie bezvýznamového textu (Gibberish Generation)	Substitúcia textu iným napr. náhodne vygenerovaným textom bez významu		
Šifrovanie / dešifrovanie (Encryption / Decryption)	Šifrovanie uložených dát pomocou vygenerovaných šifrovacích kľúčov		Pozor na únik dešifr. kľúča
Pseudosúčet (HASHing)	Jednosmerná transformácia skutočných údajov na modifikované bez pôvodného obsahu. Vhodné napr. na porovnanie dvoch rovnakých údajov z dvoch databáz bez možnosti oboznámenia sa s údajmi pri porovnávaní.		Pozor na reverziu

Anonymizácia dát – procesná časť 2

Samotný postup pozostáva z:
- Analýza možností anonymizácie
    - Identifikácia dáta, ich lokalizácia, vlastníctvo dát
    - Kategorizácia údajov
    - Ohodnotenie údajov
        - Dopad na použiteľnosť
        - Riziko prezradenia
    - Návrh techniky anonymizácie
    - Ohodnotenie zvyškového rizika
    - Dodatočné opatrenia
- Schválenie výsledkov analýzy

Samotný postup musí byť nalinkovaný na existujúce riadiace dokumenty:

- prevádzkový poriadok IT,
- interné postupy na restore databáz -
- iná bezpečnostne relevatná dokumentácia
- change management, atď.
-pozor na možný konflikt s inými dokumentami (napr. db linky)
- výnimky

Anonymizácia dát – procesná časť 2

- Ohodnotené atribúty citlivých údajov:

  Dopad na použiteľnosť údajov – biznis vlastníkmi a IT gestormi stanovený dopad na použiteľnosť anonymizovaných údajov pre testovacie účely (na stupnici Low/Medium/High)

  Riziko prezradenia – biznis vlastníkmi systému stanovená miera rizika prezradenia neanonymizovaného údaju (na stupnici Low/Medium/High)

- Pre každý citlivý údaj je stanovené:

  Posúdenie zvyškového rizika - na základe charakteru údaju, dopadu na použiteľnosť, rizika prezradenia a navrhovanej metódy anonymizácie (na stupnici Low/Medium/High);

  Dodatočné opatrenia - stanovené pre údaje s významným zvyškovým rizikom (stupeň Medium/High).

  Schválenie výsledkov ohodnotenia - po vykonaní analýzy možností anonymizácie citlivých údajov sú výstupy (ohodnotenie dopadu na použiteľnosť a rizika prezradenia, navrhnutá metóda anonymizácie, zvyškové riziko a dodatočné opatrenia) schválené biznis vlastníkmi

Microsoft Office
cel 97-2003 Worksh

Anonymizácia dát – organizačná časť 1

Role a zodpovednosti

Oddelenie bezpečnosti

definuje kategórie údajov,
definuje metódu hodnotenie dopadu na použiteľnosť údaju po anonymizácii,
definuje metódu stanovenia miery rizika prezradenia údaju bez anonymizácie,
definuje metódy anonymizácie,
navrhuje techniku anonymizácie,
navrhuje ohodnotenie zvyškové rizika,
navrhuje dodatočné opatrenia.

Biznis vlastník systému

identifikuje, kategorizuje a lokalizuje údaje,
určuje biznis účel použitia údaju,
stanovuje hodnotu dopadu anonymizácie na použiteľnosť údajov po anonymizácii,
stanovuje hodnotu miery rizika prezradenia údaju bez anonymizácie,
schvaľuje techniku anonymizácie,
schvaľuje návrh ohodnotenia zvyškového rizika a dodatočné opatrenia,
schvaľuje sumárne výstupy analýzy možností anonymizácie.

Anonymizácia dát – organizačná časť 2

Role a zodpovednosti

IT gestor aplikácie

Zodpovedá za výkonanie procesu anonymizácie dát v gestorovanej aplikácií,
Zodpovedá za používanie a poskytovanie výlučne len anonymizovanych dát interne ako aj tretím stranám aj pre účely vývoja a testovania.
IT gestor, hlavne menších aplikácií, môže anonymizáciu dát vykonať priamo a preto zodpovedá za vykonanie anonymizacie dát vo svojej aplikácii. Anonymizáciu môže vykonať "lokálne„ vo svojej databáze pomocou spoločných anonymizačných tabuliek.

Biznis vlastník systému

identifikuje, kategorizuje a lokalizuje údaje,
určuje biznis účel použitia údaju,
stanovuje hodnotu dopadu anonymizácie na použiteľnosť údajov po anonymizácii,
stanovuje hodnotu miery rizika prezradenia údaju bez anonymizácie,
schvaľuje techniku anonymizácie,
schvaľuje návrh ohodnotenia zvyškového rizika a dodatočné opatrenia,
schvaľuje sumárne výstupy analýzy možností anonymizácie.

Anonymizácia dát – organizačná časť 3

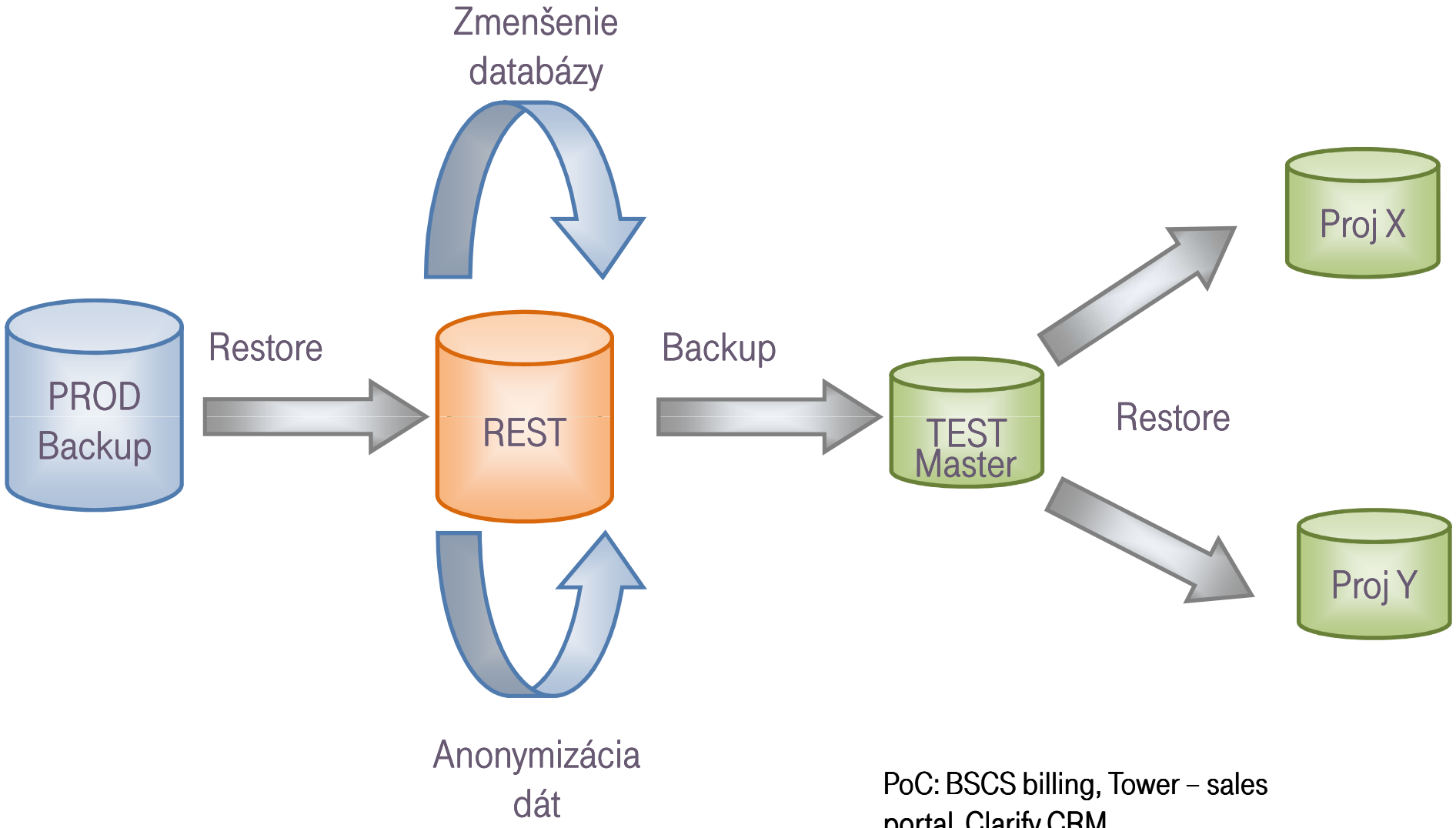
Role a zodpovednosti

Správca MasterTest prostredia :
Zodpovedný za správu a údržbu MasterTest prostredia
Iniciuje aktualizáciu MasterTest prostredia
Eviduje informácie o zmenách v databázovej štruktúre pre potreby anonymizácie.
Zabezpečuje aktualizáciu skriptov pre anonymizáciu pri zmenách v databáze
Udržiava a spravuje skripty pre anonymizáciu databáz.
Dáva pokyn databázovým administrátorom na vykonávanie databázových operácií (restore, backup,anonymizáciu).
Eviduje požiadavky na vytvorenie cieľových anonymizovaných prostredí.

Databázový administrátor (per aplikáciu)
Vykonáva obnovenie databáz do MasterTest prostredia.
Spúšťa skripty pre rekonfiguráciu a dodatočné zmenšenie databáz.
Spúšťa skripty pre anaonymizaciu databáz v MasterTest prostred
Vykonáva zálohu anonymizovaných databáz z MasterTest prostredia

Aplikačný administrátor (per aplikáciu)
Zodpovedný za udržiavanie informácií o databázovej štruktúre aplikácie.
Informuje správcu MasterTest prostredia o zmenách v databáze, ktoré majú vplyv na anonymizáciu dát.

Anonymizácia dát – technická časť 1 (PoC)



PoC: BSCS billing, Tower – sales portal, Clarify CRM

Anonymizácia dát – technická časť 2 (PoC)

Zmenšenie databázy pred anonymizáciou (optimalizácia) – zákaznícke dáta

- Výraznejšie zmenšenie zákazníckych dát = zmenšenie objemu dát uložených vo veľkom množstve tabuliek = výrazný nárast doby vykonania operácie

- Optimalizácia bola na základe tzv. štatistickej vzorky, t.j. výber percentuálneho zastúpenia dát na základe kritérií (market, segment, služba, dátum deaktivácie, atď.)

- Len malá časť tabuliek so zákazníckymi dátami môže byť časovo efektívne zmenšená:
  - BSCS billing – zmenšenie objemu 11 tabuliek trvá 8 – 12 hodín a výsledný zisk predstavuje iba 40 GB, čo je cca 5% z celkového objemu databázy
  - Clarify CRM – zmenšenie objemu 6 tabuliek trvá 8 – 12 hodín a výsledný zisk predstavuje iba 55 GB, čo je cca 10% z celkového objemu databázy

Záver:
-Odstránenie zákazníckych dát je časovo náročné pri minimálnom zisku úložného priestoru
- Efektívnejší variant je odstránenie projektovo – irelevantných dát pri príprave cieľového prostredia, napr. logovacie tabuľky, tabuľky neovplyvňujúce funkčnosť aplikácií, historické prevádzkové tabuľky

Anonymizácia dát – technická časť 3 (PoC)

Alternatívy anonymizácie – zákaznícke dáta

	+	-
Integračná alternatíva	Využíva existujúcu funkcionalitu a business logiku Efektívne pre malú vzorku zákazníckych dát.	Procesne náročné. Časovo náročné pri veľkom množstve dát. Závislé na ostatných systémoch. Nutná dokonalá anonymizácia Neúplné. Nutná implementácia náhradného riešenia pre prepaid zákazníkov. Vyšší celkový čas anonymizácie. (10 000 zákazníkov – 4 hodiny)
Skriptová alternatíva	Procesne nenáročné. Nezávislé na ostatných systémoch. Anonymizovaná kompletná sada dát. Výkonnejšie.	Implementačne náročnejšie. Náročnejšie na údržbu. Citlivé na chyby v skriptoch.

Anonymizácia dát – technická časť 4 (PoC)

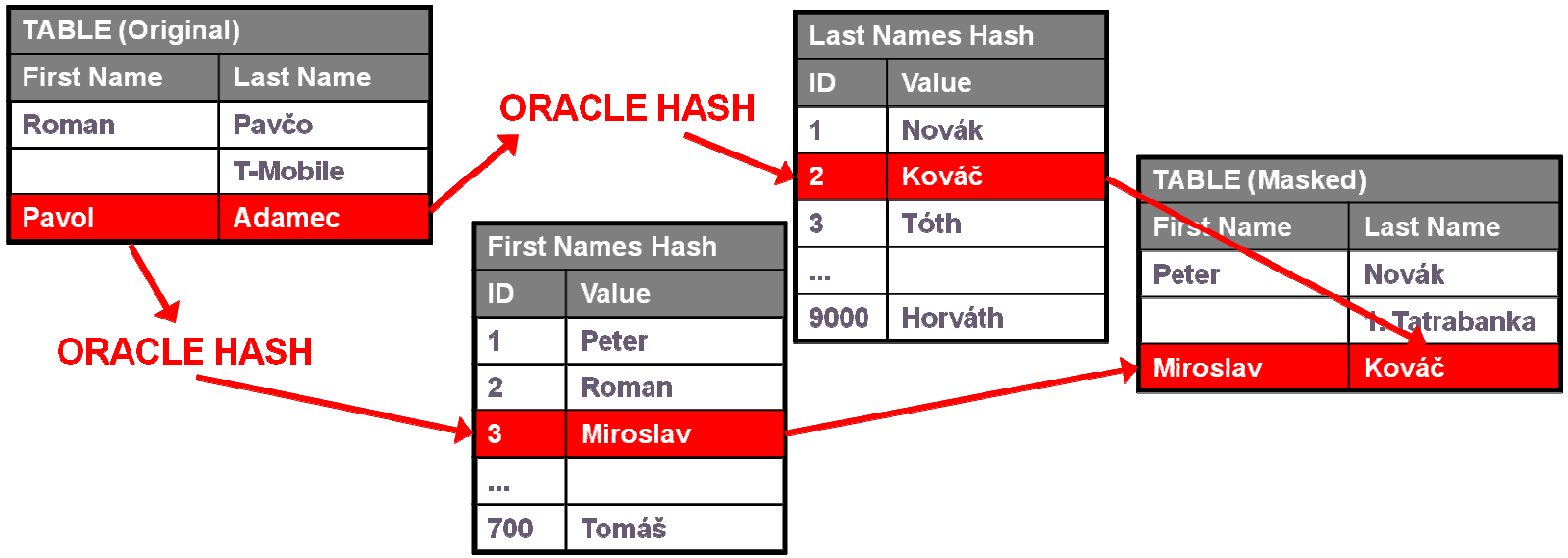
Porovnanie výkonu anonymizačných skriptov – zákaznícke dáta (substitúcia údajov)

Volanie anonymizačnej funkcie (štandardný nástroj): update table_site set name = lkp_hash(name, 'mt_hash_fname', 400)	1563 sekúnd
Priamy SQL príkaz (optimalizovaný): update table table_site set name = select value from mt_hash_fname where id = DBMS_UTILITY.GET_HASH_VALUE (upper(convert(trim(name),'us7ascii')), 1, 400)	384 sekúnd

Anonymizácia dát – technická časť 5 (PoC)

Výkon anonymizácie – zákaznícke dáta

- Samotná implementácia skriptov
  - Substitičné tabuľky
  - Anonymizačné funkcie
  - Anonymizačné skripty



Anonymizácia dát – technická časť 6 (PoC)

Štatistiky (substitúcia údajov)

- Cca 30 000 záznamov / sekunda

Systém	Tabuľka	Počet záznamov	Počet anonym. stĺpcov	Čas trvania operácie (s)
Clarify	TABLE_X_ACCOUNT_VIP	888	7	0,57
	TABLE_ADDRESS	9 309 746	14	2 700
	TABLE_BUS_ORG	1 893 173	9	2 000
	TABLE_X_ACCOUNT_CONT_HIST	2 348	6	1
	TABLE_X_ACCOUNT_CONTACT	898 550	6	2 100
	TABLE_X_EWH_CUSTOMER	847 942	9	75
	TABLE_CONTACT	5 775 738	18	1 400
	TABLE_LEAD	17 014	24	7
	TABLE_SITE_HISTORY	2 879 882	15	3 338
	TABLE_SITE	5 958 374	18	4 481
BSCS	CUSTOMER_ALL	2 325 732	7	3 100
	CCONTACT_ALL	3 258 098	27	3 350

Anonymizácia dát – technická časť 7 (PoC)

Konzistencia dát

First Name Original	First Name Masked	Last Name Original	Last Name Masked	ICO Original	ICO Masked	Personal Number Original	Personal Number Masked
JOZEF	Jakub	HLAVACEK	Poláková			741210/7032	7403034353
PETER	Andrea	ANTALIK	Gregor			701002/6078	7011288229
		EURÓPSKA KOMISIA - ZASTÚPENIE V SR	2. Ultrainvest, S.R.O.	31794335	34696641		
Martin	Slavomíra	Franciscy	1Pompova			7905316155	7905258878
MIROSLAV	Judita	GLADIŠ	Mitro			6912058736	6904192625
JAN	Lubomir	FERKO	Kubikova			761022/6162	7607067897
		SBS DYNASTY A.S.	6. Secret Service, S.R.O.	35762764	23741447		

CRM First Name	BSCS First Name	CRM Last Name	BSCS Last Name	CRM ICO	BSCS ICO	CRM personal number	BSCS personal number
Katarína	Katarína	Marton	Marton			6612226038	6612226038
Michaela	Michaela	Mináriková	Mináriková			6407231435	6407231435
Michala	Michala	Polakovicova	Polakovicova			7611047114	7611047114
		1. Inbiz S.R.O.	1. Inbiz S.R.O.	51740516	51740516		
		2. Ad-Promo Consulting S.R.O.	2. Ad-Promo Consulting S.R.O.	56675600	56675600		
Alexander	Alexander	Moravcikova	Moravcikova			5806208804	5806208804
Slavomíra	Slavomíra	Antalova	Antalova			6111281440	6111281440

Záver:

- v tomto prípade je menej efektívne použiť komerčný nástroj
- dá sa efektívne anonymizovať pomocou na mieru implementovaných skriptov

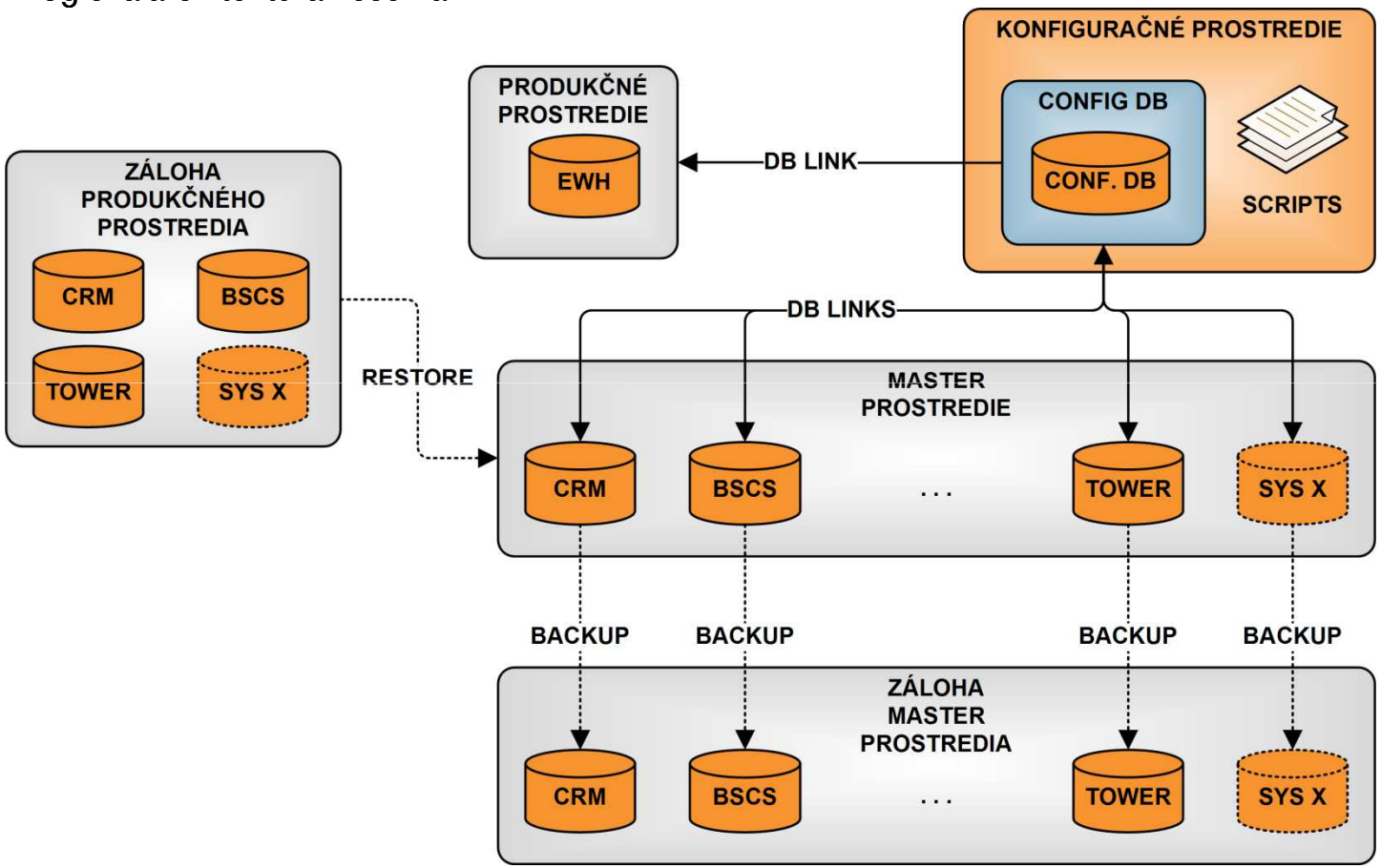
Anonymizácia dát – technická časť 8 (PoC)

Časové nároky (v hodinách)

	BSCS	Clarify	Tower	Config DB
Restore produkčnej databázy	6	5	3	
Dodatočná úprava	2	2	8	
Dodatočné zmenšenie	5	2		
Príprava na anonymizáciu				2
Anonymizácia	8	24	3	
Záloha MasterTest databáz	4	5	3	
CELKOM	**25**	**38**	**17**	**2**

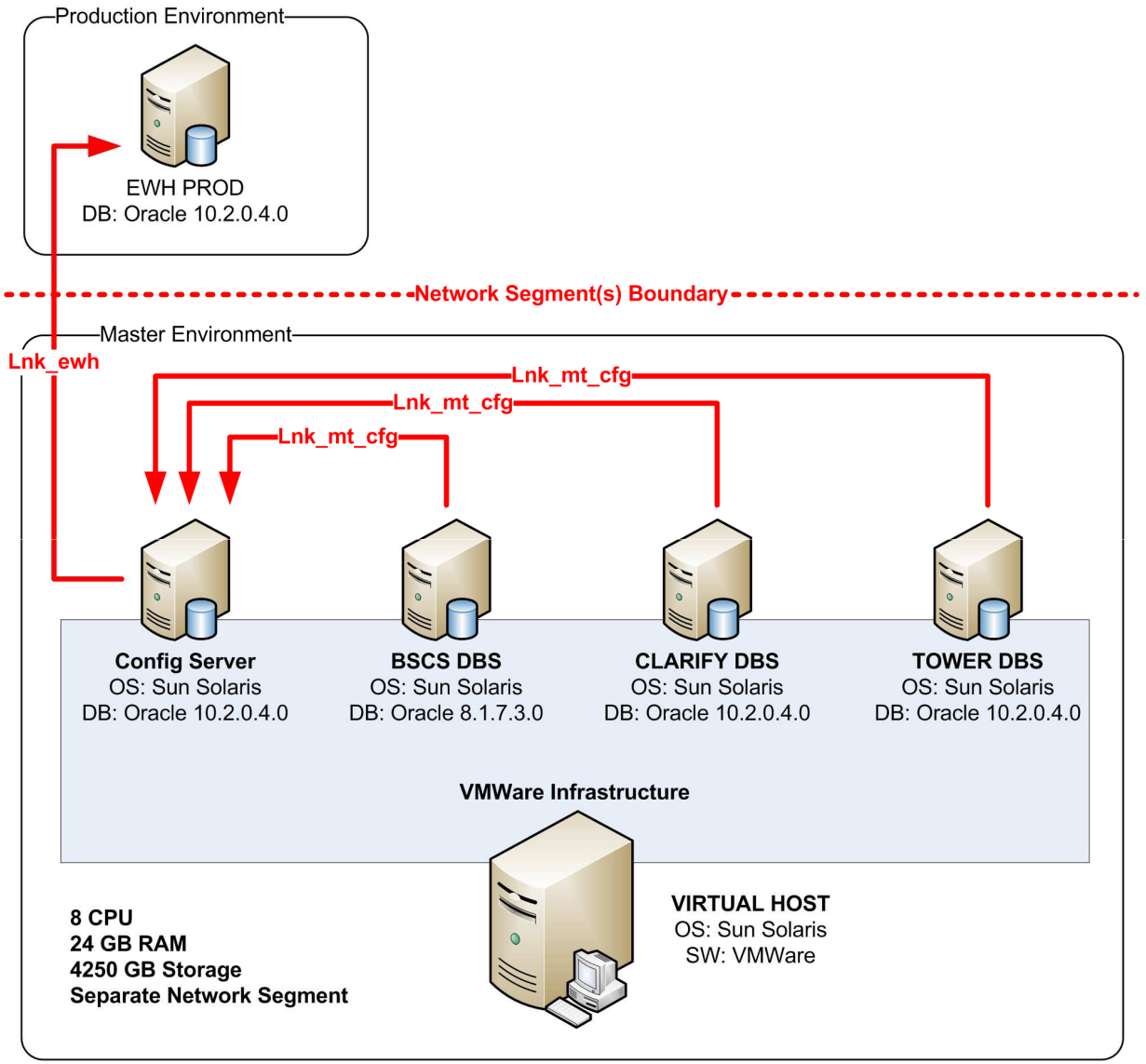
Anonymizácia dát – technická časť 9 (PoC)

Logická architektúra riešenia



Anonymizácia dát – technická časť 10 (PoC)

Master test prostredie



Anonymizácia dát – technická časť 11 (PoC)

Substitúcia údajov – substitičná tabuľka

Prvé meno – 750 najpoužívanejších mien z EWH
Priezvisko – 10 000 najpoužívanejších priezvisk z EWH
Názov firmy – 30 000 názvov firiem z EWH
Ulica – 1 000 ulíc z EWH
Mesto – 500 obcí z EWH
Dni v roku – 365 dní náhodne usporiadaných
Zástupné znaky – pre substitúciu znakov

Iné anonymizačné funkcie

RČ, dátum narodenia, DIČ, atď.

Anonymizácia dát – záver a odporúčania

- Systém anonymizácie test / development databáz s citlivými údajmi v telco prostredí je nutnosťou – legislatívne potreby, bezpečnostné potreby, prevádzkové potreby atď.
- Systém anonymizácie sa dá prirovnať k PDCA cyklu, t.j. ide o kontinuálny proces
- Procesná časť bez organizačnej časti alebo bez technickej časti nie je funkčná
- Jasné definovanie povinností a ich vymožiteľnosť je v tomto procese mandatórna
- Technická časť musí byť dostatočne flexibilná a rýchla, ľahko udržiavateľná
- Existencia plánu pokrývania databáz, revízie existujúcich databáz
- Pozor na výnimky!
- € nutné investície
- Brzdí rozvoj spoločnosti
- Podpora menežmentu, (p)odpora prevádzky

BACKUP

Zoznam skratiek

Analýza rizík	systematický proces analýzy prostredia a vzťahov medzi jednotlivými atribútmi (aktíva, hrozby, zraniteľnosti, dopady) vo vzťahu k rizikám, ktorým môže byť Spoločnosť vystavená. Je to ohodnotenie pravdepodobnosti nastania rizika a jeho dopadu na Spoločnosť.
Dodatočné opatrenie	existujúci postup, politika, prostriedok, spôsob alebo iná činnosť slúžiaca k minimalizácii alebo eliminácii možnosti vzniku, pôsobenia a následkov rizika.
Dopad na použiteľnosť	predstavuje mieru vplyvu anonymizácie údaju na jeho následné využitie v testovacom scenári.
Dostupnosť	požiadavka, aby aktívum bolo na požiadavku autorizovanej entity prístupné a schopné použitia.
Dôvernosť	požiadavka, aby informačné aktívum nebolo sprístupnené neautorizovaným entitám
Informačná bezpečnosť	súbor aspektov týkajúcich sa dosiahnutia a udržiavania dôvernosti, integrity, dostupnosti informačných aktív
Informačné aktívum	objekt (aktívum) súvisiace so spracúvaním informácií, ktoré pre Spoločnosť priamo predstavuje hodnotu alebo narušenie jeho bezpečnosti môže mať pre Spoločnosť negatívny dopad.
Informačný systém (IS, aj „IT/ITC service“)	informačné a komunikačné technológie, hardvérové a softvérové prvky (server, IT služba, aplikácia alebo aplikačný modul, atď.), ktoré sú prevádzkované v Spoločnosti.
Integrita	požiadavka, aby informačné aktíva neboli stratené, zničené alebo zmenené neautorizovaným alebo náhodným spôsobom
Ohodnotenie rizík	celkový proces identifikácie, analýzy a zhodnotenia rizika
Riziko	predstavuje ohrozenie obchodných cieľov a implementácie obchodnej stratégie vznikom potenciálnych udalostí, činností a zlyhaní pôsobiacich v rámci Spoločnosti alebo mimo nej. Riziko je merateľná možnosť, že budúcnosť môže byť iná ako predpokladáme.
Master testovacia databáza	(MasterTest prostredie) upravená verzia databáz extrahovaných z produkčných systémov