Datová kvalita

RNDr. Ondřej Zýka ondrej.zyka@profinit.eu

Datová kvalita

Jedna z kompetencí Data managementu

Cíl: Zajistit uživatelům data v „kvalitě“ potřebné k jejich činnosti

Kvalita dat:

Subjektivní pojem závislý na požadavcích a zkušenosti uživatelů, na způsobu použití dat

Kvalita dat není dána jejich strukturou nebo uložením.

Dimenze datové kvality

Dimenze	Popis
Dostupnost	Zda jsou informace k dispozici nebo snadno získatelné
Odpovídající velikost a granularita dat	Zda velikost dat a jejich granularita odpovídá vykonávaným úlohám
Věrohodnost	Zda jsou informace pokládány za pravdivé a důvěryhodné
Úplnost	Zda žádná data nechybí a zda jsou dostatečné rozsáhlá a detailní pro vykonávané úlohy
Výstižná reprezentace	Zda reprezentace dat má vhodnou strukturu
Konzistentní reprezentace	Zda jsou data reprezentována vždy ve stejném formátu
Snadnost zpracování	Zda jsou informace snadno zpracovatelné a použitelné pro rozdílné úlohy
Bezchybnost	Zda jsou informace a data přesné a hodnověrné
Interpretovatelnost	Zda je jasná definice informací, zda jsou v odpovídajícím jazyku, jednotkách a zda jsou označeny správnými symboly
Objektivita	Zda jsou informace nestranné a nepředpojaté
Relevantnost	Zda jsou informace použitelné a užitečné pro vykonávané úlohy
Reputace	Zda jsou informace považovány za spolehlivé v souvislosti s jejich zdrojem nebo obsahem
Bezpečnost	Zda omezení přístupu k datům a informacím odpovídá bezpečnostním pravidlům
Včasnost	Zda jsou pro vykonávané úlohy informace k dispozici včas
Srozumitelnost	Zda jsou informace snadno pochopitelné a srozumitelné
Přidaná hodnota	Zda a která data a informace jsou přínosné a jaké jsou výhody jejich použití

Základní otázky datové kvality

- Kdy jsou data kvalitní?
- Kdy jsou data nekvalitní?
- Jak prokázat, že jsou data kvalitní?
- Jak zvýšit kvalitu dat?
- Pozorování
  - Dodavatelé dat obecně nemají moc důvodů produkovat bezchybná data.
  - Nekvalitní data vytváří nesmírnou frustraci uživatelů dat.

- Kvalita dat se nedá dosáhnout pouze prostředky IT.
- Příklady
  - adresa@naznama.cz
  - Rodné číslo

Kdy jsou data kvalitní?

Kdy jsou data nekvalitní?

PROFINIT
new frontier group

Management a Finance

- Nutnost udržovat velké finanční nebo technické rezervy
- Nekonzistentní reporty napříč organizací
- Reporty s nedůvěryhodnými daty
- Rozhodnutí učiněná na základě špatných informací

Marketing

- Nepřesná segmentace zákazníků
- Drahé a neúčinné kampaně
- Nízká kvalita služeb pro zákazníky
- Nepořádek v zákaznických datech

Vlastníci systémů

- Duplicity v datech
- Nekonzistence mezi systémy
- Chybějící nebo nedohledatelné údaje
- Zastaralé informace

IT

- Vysoká náročnost nalezení požadovaných informací
- Nemožnost dohledání původu dat a zodpovědných pracovníků
- Nespokojenost uživatelů s dodávanými informacemi
- Neschopnost řešit konzistentně vady v datech

Příznaky nekvality v datech?

- Reporty nejdou porovnat
- Pracovníci si vedou soukromé agendy
- Pracovníci si nechávají výsledky kontrolovat

Proč se zabývat datovou kvalitou

- Výskyt chyb v datové kvalitě
- Nespokojenost uživatelů
- Legislativní požadavky, požadavky regulátorů
  - Solvency II
  - Basel II, Basel III

Jak prokázat kvalitu dat?

- Co to je za číslo?
  - Jak vzniklo?
  - Kdo to kontroloval?
  - Byla použita všechna data?
  - Byla použita aktuální data?
- ?????

	Net solvency capital requirement (including the loss-absorbing capacity of technical provisions)	Gross solvency capital requirement (excluding the loss-absorbing capacity of technical provisions)
Market risk	A1	B1
Counterparty default risk	A2	B2
Life un	A3	B3
Health	A4	B4=A4
Non-lif	A5	B5=A5
Diversi	A6	B6
Intangible asset risk	A7	B7=A7

1059,56

Jsou nastaveny procesy a politiky

- Je definována politika datové kvality
- Je definována organizace DQ
  - Role
  - Job description
  - Accountability and responsibility assignment
- Jsou vytvořeny a udržovány slovníky DQ
  - Definice dat
  - Popis dat a datových toků
  - Stanovení metrik datové kvality pro jednotlivé prvky
- Jsou nastaveny procesy DQ
  - Nastaveno měření a reporting datové kvality
  - Nastaven proces řízení chyb v datové kvalitě
    - Identifikace, odhad dopadů, definice nápravy, ohodnocení nápravy, oprava dat, dokumentace opravy
  - Nastaven proces DQ operation

Slovníky datové kvality

PROFINIT
new frontier group

Level 3 type	Level 3 definition	Level 3 threshol	e	Level 4 ty	Level 4 - Indicator 1 definition	Level 4 - Indicator 1 thresh
List of values	A pre-defined list of values (1, 2,or I)	0%-2%		Uniqueness	The first 6 variables should uniquely define a record	0%-2%
Format	Numeric format	0%-2%		Uniqueness	The first 6 variables should uniquely define a record	0%-2%
Format	All the dates are yyyymmdd, plus the variable should be within a reasonable				stency of values across time - values from previous extract should match with those from the current extract	0%-2%
List of values	Sh				stency of values - for different records of the same contract these values should be consistent	0%-2%
List of values	Sh				stency of values - for different records of the same contract these values should be consistent	0%-2%
Format					stency of values across time - values from previous extract should match with those from the current extract	0%-2%
List of values					stency of values across time - values from previous extract should match with those from the current extract	0%-2%
List of values					stency of values across time - values from previous extract should match with those from the current extract	0%-2%
Format					stency of values - for different records of the same contract these values should be	0%-2%
Format					stency of values - for different records of the same contract these values should be	0%-2%
Format					stency of values across time - values from previous extract should match with those from the current extract	0%-2%
Format	All				stency of values - for different records of the same contract these values should be consistent	0%-2%
	range					
Format	Integer	0%-2%		Consistency	Consistency of values - for different records of the same contract these values should be	0%-2%
Format	Integer	0%-2%		Consistency	Consistency of values - for different records of the same contract these values should be	0%-2%
Format	Integer	0%-2%		Consistency	Consistency of values across time - values from previous extract should be consistent with those from the current extract unless there were changes in the contract	0%-2%
Format	Integer	0%-2%		Consistency	Consistency of values across time - values from previous extract should be consistent with those from the current extract unless there were changes in the contract	0%-2%
Format	Number	0%-2%		Consistency	Consistency of values across time - values from previous extract should match with those from the current extract	0%-2%
Format	Percentage (a number between 0 and 100)	0%-2%		Consistency	Consistency of values across time - values from previous extract should match with those from the current extract	0%-2%
List of values	A string of 2 characters	0%-2%		Reconciliation	Reconciliation of amounts against the data from other sources, e.g. operations	0%-2%
List of values	A string of 2 characters	0%-2%		Reconciliation	Reconciliation of amounts against the data from other sources, e.g. operations	0%-2%
Format	Number	0%-2%		Reconciliation	Reconciliation of amounts against the data from other sources, e.g. finance	0%-2%

Definice na obchodní úrovni
Definice na technické úrovni
Místo a formát uložení
Vlastník - Zodpovědná osoba
Parametry důležitosti, bezpečnosti, aktuálnosti, ...

Metriky datové kvality

PROFINIT
new frontier group

Technické

- Data mají přípustné hodnoty, očekávaný formát, pohybují se v přípustném rozsahu, jsou jednoznačné – pokud je to požadováno, existují odpovídající záznamy v jiných systémech

Významové

- Hodnoty, počty a sumy jsou konzistentní v čase. Porovnání s historickými daty a benchmarky nevykazuje neodůvodněné odchylky.
- Existuje požadovaná konzistence mezi různými záznamy a hodnotami.

Požadavek

- Kontrakt musí mít definován Politiku zajištění

Metrika

- Procento kontraktů s vyplněným parametrem Politika zajištění

Tresholds

- OK > 99%
- Failed < 95 %

Baseline

- 96,2 %

Level 3 t				Level 4 ty		ndicator 1 thresh
List of va				Uniqueness		0%-2%
Forma				Uniqueness		0%-2%
Forma				Consistency		0%-2%
List of va				Consistency		0%-2%
List of va				Consistency		0%-2%
Forma				Consistency		0%-2%
List of va				Consistency		0%-2%
List of va				Consistency		0%-2%
Forma				Consistency		0%-2%
Forma				Consistency		0%-2%
Forma				Consistency		0%-2%
Forma				Consistency		0%-2%
Forma				Consistency		0%-2%
Forma				Consistency		0%-2%
Forma				Consistency		0%-2%
Forma				Consistency		0%-2%
Forma				Consistency		0%-2%
List of va				Reconciliation		0%-2%
List of values	A string of 2 characters	0%-2%		Reconciliation	Reconciliation of amounts against the data from other sources, e.g. operations	0%-2%
Format	Number	0%-2%		Reconciliation	Reconciliation of amounts against the data from other sources, e.g. finance	0%-2%

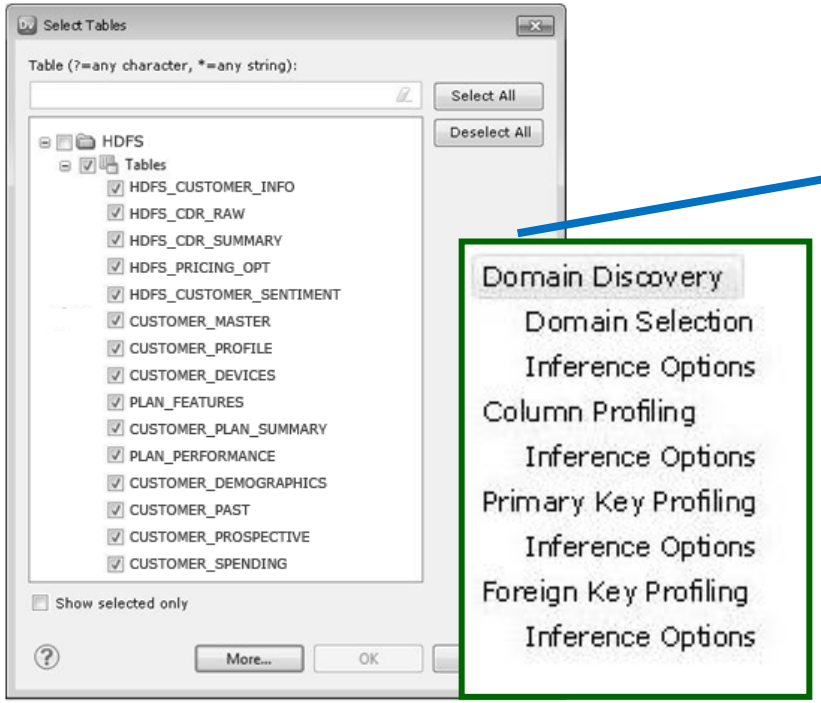
Metriky DQ

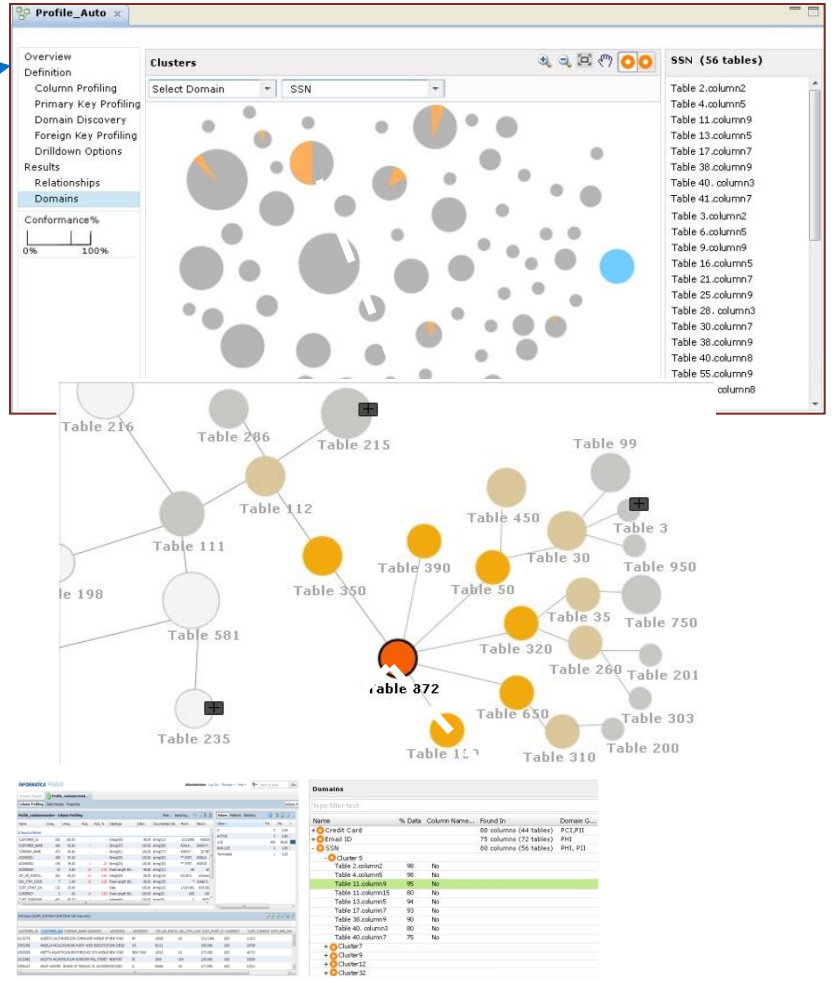
PROFINIT
new frontier group

- Počet not null hodnot
- Čísla
  - Rozsah
  - Histogram
  - Přesnost
  - Speciální hodnoty (0, 1, 100, 10, ..)
- Řetězce
  - Délka
  - Vzory, hodnoty extrémních vzorů
  - Minimum a maximum
- Vazby
  - Počet nepoužitých cizích klíčů
  - Histogram použití cizích klíčů
  - Počet neexistujících cizích klíčů

13

Profiling – měření DQ metrik





- Technický
- Uživatelský
- Speciální

Data Profiling

PROFINIT
new frontier group

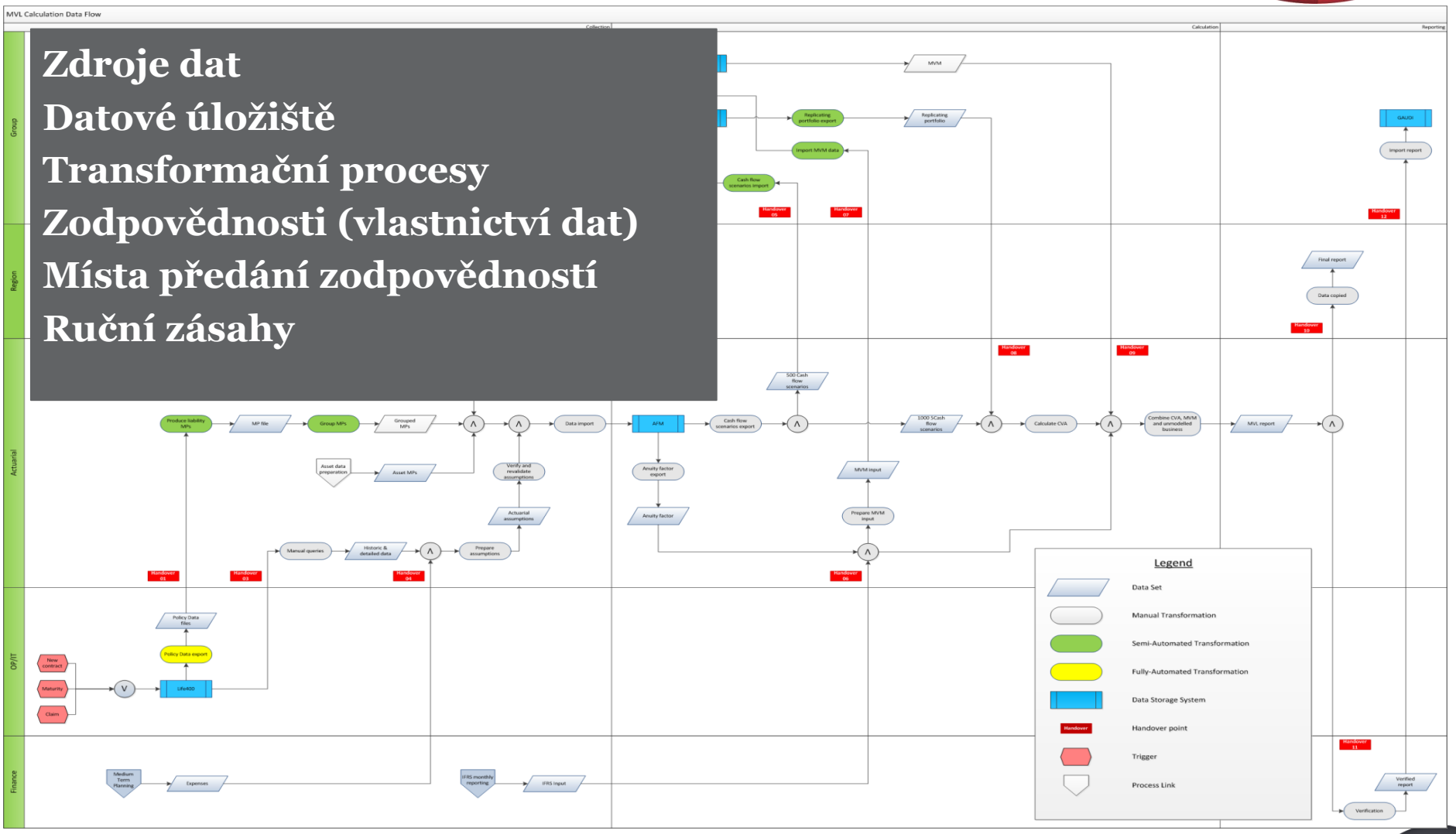
Profile Name: Profile_pmprpf

Name	Pattern	Frequency	Percent
CHDRCOY			
	9	907160	100.0
CHDRNUM			
	99999999	907160	100.0
POANUM			
	XX-X	70225	7.74
	X	836362	92.2
	NULL	321	0.04
	Others	252	0.03
BLABEL			
	X	907160	100.0
AGNTNUM01			
	99999	49070	5.41
	999999	857578	94.53
	NULL	443	0.05
	Others	69	0.01

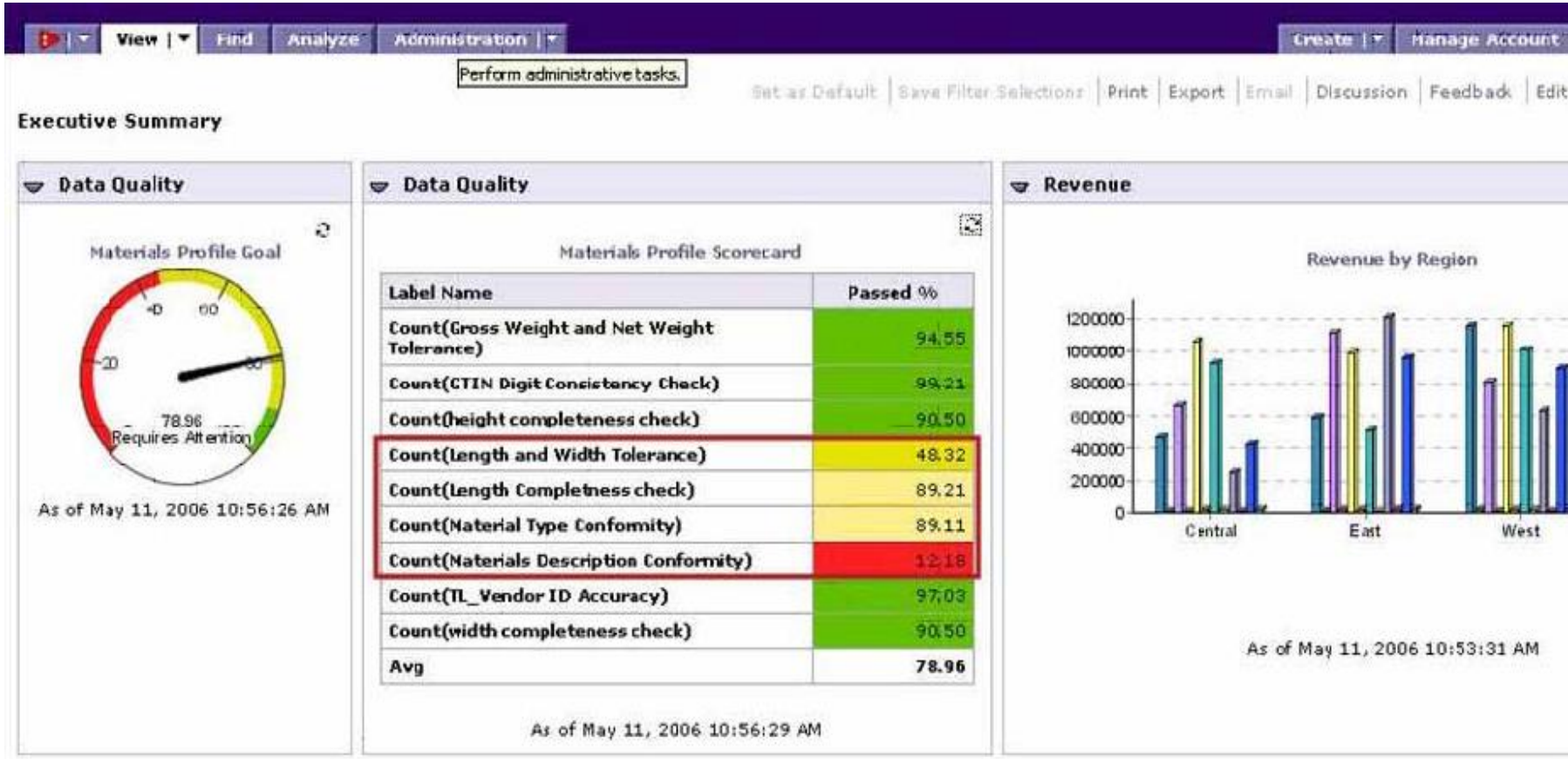
Popis datových toků

PROFINIT
new frontier group

- Zdroje dat
- Datové úložiště
- Transformační procesy
- Zodpovědnosti (vlastnictví dat)
- Místa předání zodpovědností
- Ruční zásahy

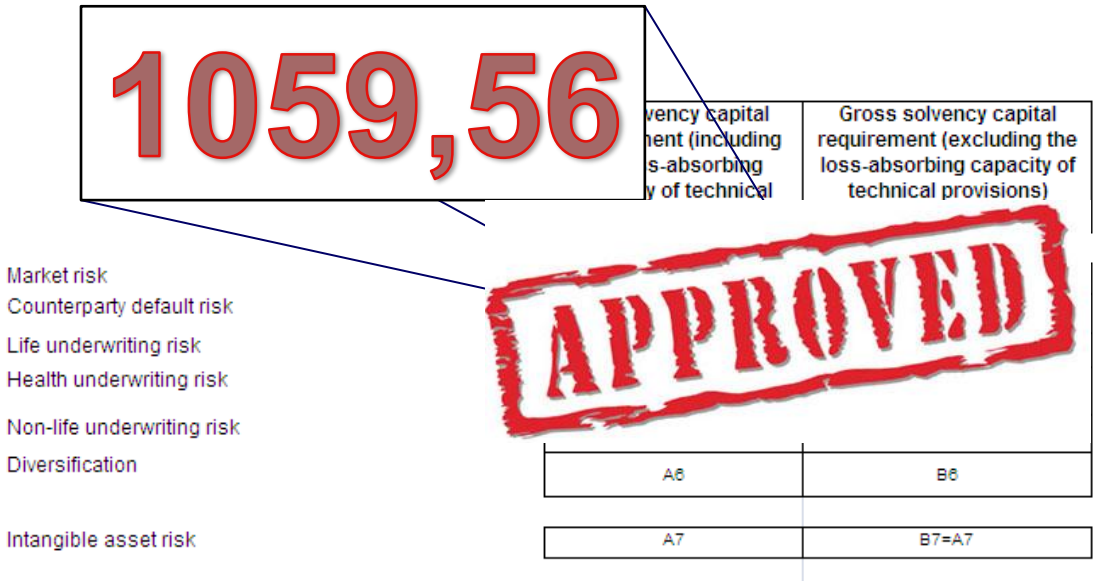


DQ měření a reportování



Jak prokázat kvalitu dat?

- Udělal jsem všechno pro to, abych číslu mohl věřit.



Důvody nekvalitních dat

- Pouze dva zdroje znečištění dat

- Na vstupu
  - Uživatelé
  - Zdrojové systémy
- Zastarávání dat
  - Deset let starý telefonní seznam neobsahuje kvalitní data

Kdy a jak čistit data

- Vždy je možné „zlepšit“ kvalitu dat
- Pokud si nikdo nestěžuje, nemá smysl investovat zvyšování kvality dat
- Pokud se objeví problém s datovou kvalitou, je nutné porovnávat přínosy a náklady čištění

Jak zvýšit kvalitu dat?

- Čištění dat
  - Neexistuje jedno správné řešení
  - Obecně data nejdou vyčistit
- Čtyři základní metody
  - Nechat kvalitu dat na uživateli – nečistit
  - Jednorázové čištění
  - Čistění příchozích dat
  - Čištění používaných dat
  - Nalezení a úprava znečišťovatele
- Vzdělávání uživatelů a původců dat
- Příklad (voda v jezeře)

Co si zapamatovat

PROFINIT
new frontier group

- Co to je datová kvalita
- Jak se pozná, že jsou data kvalitní
- Kdo a jak pozná, že jsou data nekvalitní
- Jaké metody se používají pro čištění dat
- Kde a jak vzniká nekvalita dat
- Co to jsou dimenze datové kvality
- Co to je profiling dat
- Jak se dá prokázat, že jsou informace získaná z dat kvalitní

PROFINIT
new frontier group

Diskuse