FIAP

Sistemas de informação

enterprise analytics and data warehousing

Profº Fabiano J. Cury Marques

FIAP

Enterprise Analytics and Data Warehousing

Data quality

Profº Fabiano J. Cury Marques

FIAP

Agenda

- ✗ Processo de Data Quality
- ✗ Data Cleansing e Matching
- ✗ Cross-checking com fontes externas
- ✗ Regras de Data Quality
- ✗ Logging
- ✗ Exercícios
- ✗ Referências

Introdução

- ✘ Por que é importante?
- ✘ Esta questão é muito fácil de responder

- ✘ Suponha que não se possa confiar nos dados em seu DW por algum dos motivos:
  - dado é incorreto (correção)
  - dado está faltando (completeza)
  - dado não está sincronizado com outro (consistência)
  - dado é apenas uma estimativa (precisão)
  - dado está desatualizado (atualização temporal)

- ✘ Nestes casos, seria possível usar o DW?
- ✘ Se usar estes dados para tomar decisões e os dados não forem confiáveis, quanto isto custa para a empresa?
- ✘ Quanto custa para a carreira dos envolvidos?

Introdução

- Para que serve um DW se ele não é confiável?
- É importante pensar na qualidade dos dados o mais cedo possível, de preferência no começo do projeto
- A essência do DQ é prevenir dados ruins de entrarem no DW e consertar estes dados tão cedo quanto possível
- Para implementar isto, precisamos definir regras de qualidade que definem o que é um dado ruim e adicionar componentes na arquitetura que filtram estes dados, reportam, monitoram e permitem mecanismos de limpeza

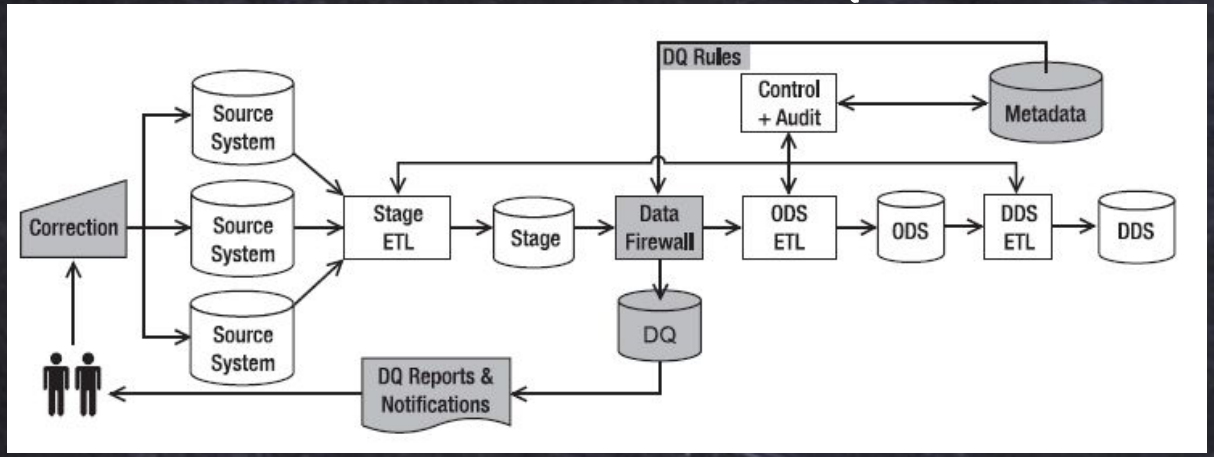
EXEMPLO

- No estudo de caso da Amadeus Entertainment, clientes podem comprar um produto ou assinar um serviço (pacote)
- A data em que um cliente assina um serviço pela primeira vez é conhecida como primeira data de assinatura e a data do cancelamento mais recente é conhecida como última data de cancelamento
- Suponha que um dia o processo de ETL extraia um registro de cliente com a última data de cancelamento menor que a primeira data de assinatura. Isto não é uma condição válida
- Ou a data de cancelamento ou a de assinatura está incorreta ou até ambas
- O processo de qualidade de dados detecta esta condição e o informa para as pessoas responsáveis pelos dados de assinatura. Eles então corrigem o dado no sistema fonte e por fim carrega o dado no DW

Processo de data quality

Processo de Data quality

PROCESSO DE DATA QUALITY



- ✘ Data Firewall é um programa que verifica o dado que está entrando. Pode ser uma store procedure ou um pacote de ferramentas de ETL
- ✘ Metadata é a base de dados que armazena as regras de qualidade de dados, exemplo, "a última data de cancelamento deve ser maior que a primeira data de assinatura"
- ✘ Base de dados DQ armazena os dados ruins detectados pelo firewall
- ✘ Reports & Notification lêem a base DQ e informam as pessoas responsáveis pela qualidade dos dados
- ✘ Correction é o processo de corrigir os dados nos sistemas fontes

DETALHES DO PROCESSO

- Quando o dado é colocado na base de qualidade de dados, certas informações de auditoria são gravadas também, como:
  - sistema fonte de onde vem o dado
  - tabela de onde vem o dado
  - data e hora quando aconteceu
  - em qual regra o dado falhou
  - onde o dado seria carregado

- Na caixa Correction da figura anterior, as pessoas responsáveis pelas áreas de negócio relacionada ao dado consertarão os dados no sistema fonte, então, quando o dado for novamente carregado pelo processo de ETL, já estará correto e será carregado no DW

DETALHES DO PROCESSO

- ✘ Pode-se corrigir os dados automaticamente em seu caminho para o DW, mas geralmente isto não é uma boa prática pois o dado continuará errado nos sistemas fontes
  - Outras aplicações que podem usar os mesmos dados ou até usar os mesmos sistemas fontes irão publicar dados incorretos também

- ✘ Em vez de colocar o data firewall antes do ETL da ODS como na figura anterior, podemos colocar o data firewall antes do ETL da Stage.
  - O processo de extração fica mais lento, mas os dados ruins nem sequer tocam em qualquer área de nosso DW, nem mesmo na área de stage

- ✘ Pode-se ainda verificar os dados na fonte antes de extraí-los
  - Usando cláusula WHERE nas queries
  - Fica mais lento ainda e não pode-se reportar os dados ruins para tratá-los

DETALHES DO PROCESSO

- ✘ Muitas vezes o problema de qualidade não está em um registro específico em sim em uma coleção de registros

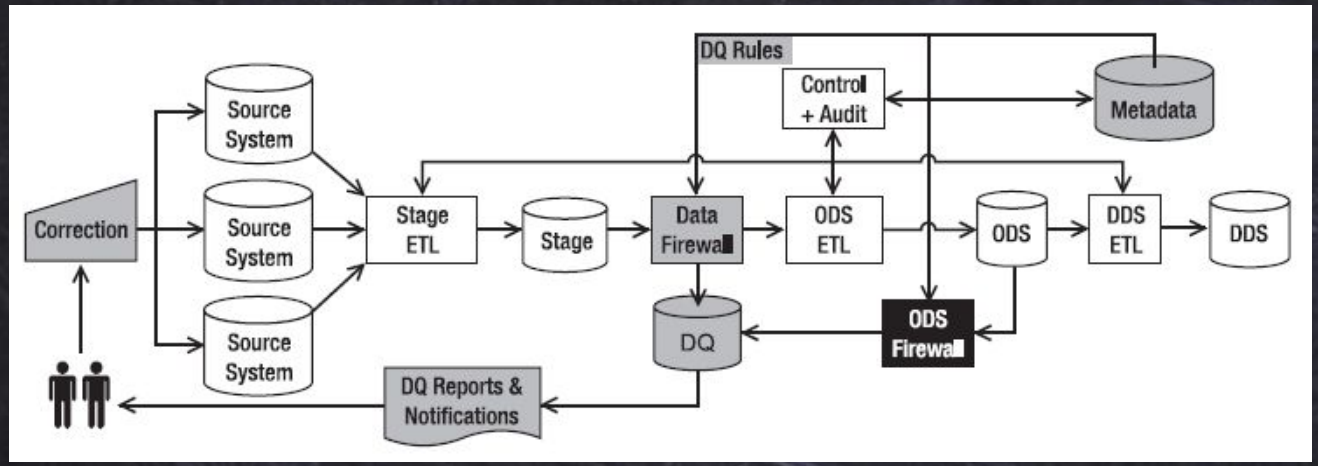
- ✘ Por exemplo, suponha que uma filial tenha de 20.000 a 60.000 transações por dia, com um total de venda entre $600.000 e $6 milhões

  - Se o número de transações é menor que 10.000 ou o total de vendas é de $500.000, provavelmente é um indicação de falhar na extração ou carga dos dados, assim queremos que está condição seja informada

  - Uma maneira de implementar isso é ter um firewall colocado após a carga dos dados (na ODS mas antes da DDS)

  - O firewall lê os dados carregado na ODS por exemplo e verifica se está de acordo com as regras de qualidade

DETALHES DO PROCESSO



- ✘ Se a coleção falhar nas regras, copia-se os dados para a base DQ
- ✘ Pode-se deixar os dados na ODS para serem carregados no DDS ou não (até que sejam arrumados)
- ✘ Depende da natureza do dado, se é importante estar presente no DW mesmo que não esteja completo, carrega-se, mas se dados incompletos causam problemas de análise e tomada de decisão, removemos os mesmos da ODS
- ✘ Se os dados ficarem, nós geralmente marcamos atributos de data quality nos registros: accuracy, precision, timeliness, completeness e consistency. Estes atributos são preenchidos com porcentagens

FIAP

DATA CLEANING E MATCHING

Data cleasing e matching

DATA CLEANSING

- Data Cleansing é o processo de identificar e corrigir dados "sujos"

- Por dados sujos entende-se: incompletos, errados, duplicados ou desatualizados

- Exemplos
  - Garantir que não existem registros de clientes duplicados
  - Que as tabelas de preço estejam corretas
  - Que produtos obsoletos estejam marcados de acordo
  - Que todas as assinatura referem-se a serviços existentes
  - Que todos os nomes de músicas e filmes são corretos

MATCHING

- No processo de data cleansing é importante ser capaz de determinar que um dado é corresponde a um outro dado
- Isto é chamado de data matching
- Usado para identificar registros duplicados quando fazendo uma pesquisa em dados de referência
- Particularmente relevante para tipos de dados caracter, como varchar pois para dados numéricos ou data/hora podemos simplesmente usar o operador igual (=)
- Para caracteres, isto pode não ser claro

EXEMPLO

- Ao carregar a tabela address (endereço) na NDS, pode-se encontrar que o nome da cidade é "Los Angles", que não existe na tabela city (cidade)
- Neste caso, precisamos bater "Los Angles" com "Los Angeles"
- Outro exemplo seria o nome do cliente "Robert Peterson". É necessário reconhecer que "Robert Peterson" é o mesmo que "Bob Peterson"
- Dados numéricos não é tão problemático pois podemos simplesmente usar o operador igual, exemplo 5 = 5. O único possível problema seria arredondamento
- 5,029 é igual a 5,03? Se a precisão for 2 decimais sim senão não.

FIAP

Matching nas ferramentas de ETL

- As ferramentas de ETL normalmente contém pelo menos três tipos de lógica de matching: exata, fuzzy (aproximada), baseada em regras
- Exata
  - Todos os caracteres são os mesmos. Ex: "Los Angeles" e "Los Angeles"
- Fuzzy
  - Quão similar um conjunto de dados é de outro conjunto de dados
  - Ex.: "You can´t go there" e "You cannot go there" tem uma pontuação de similaridade de 0.81666672
  - Você pode decidir que pontuação acima de 0,75 é um match
- Baseada em regras
  - Quando usa-se regras para identificar um match
  - Ex.: pode-se definir em uma tabela que para nomes "Bill" é o mesmo que "Willian" ou que em nomes de produtos "films" é igual a "movie" ou até omitir os espaços entre um código (KL 7923 M = KL7923M)

FIAP

Matching nas ferramentas de ETL

- As ferramentas de ETL normalmente contém pelo menos três tipos de lógica de matching: exata, fuzzy (aproximada), baseada em regras
- Exata
  - Todos os caracteres são os mesmos. Ex: "Los Angeles" e "Los Angeles"
- Fuzzy
  - Quão similar um conjunto de dados é de outro conjunto de dados
  - Ex.: "You can´t go there" e "You cannot go there" tem uma pontuação de similaridade de 0.81666672
  - Você pode decidir que pontuação acima de 0,75 é um match
- Baseada em regras
  - Quando usa-se regras para identificar um match
  - Ex.: pode-se definir em uma tabela que para nomes "Bill" é o mesmo que "Willian" ou que em nomes de produtos "films" é igual a "movie" ou até omitir os espaços entre um código (KL 7923 M = KL7923M)

Exemplo de utilização de fuzzy para Match

- ✘ Imagine a leitura de uma tabela de artistas (artist2) do sistema fonte
- ✘ A tabela foi carregada na stage
- ✘ Agora deverá ser carregada na NDS
- ✘ A tabela contém 20 artistas
- ✘ Como não temos garantia de que no sistema fonte escreveram os nomes dos artistas de maneira idêntica, vamos usar fuzzy para fazer o lookup

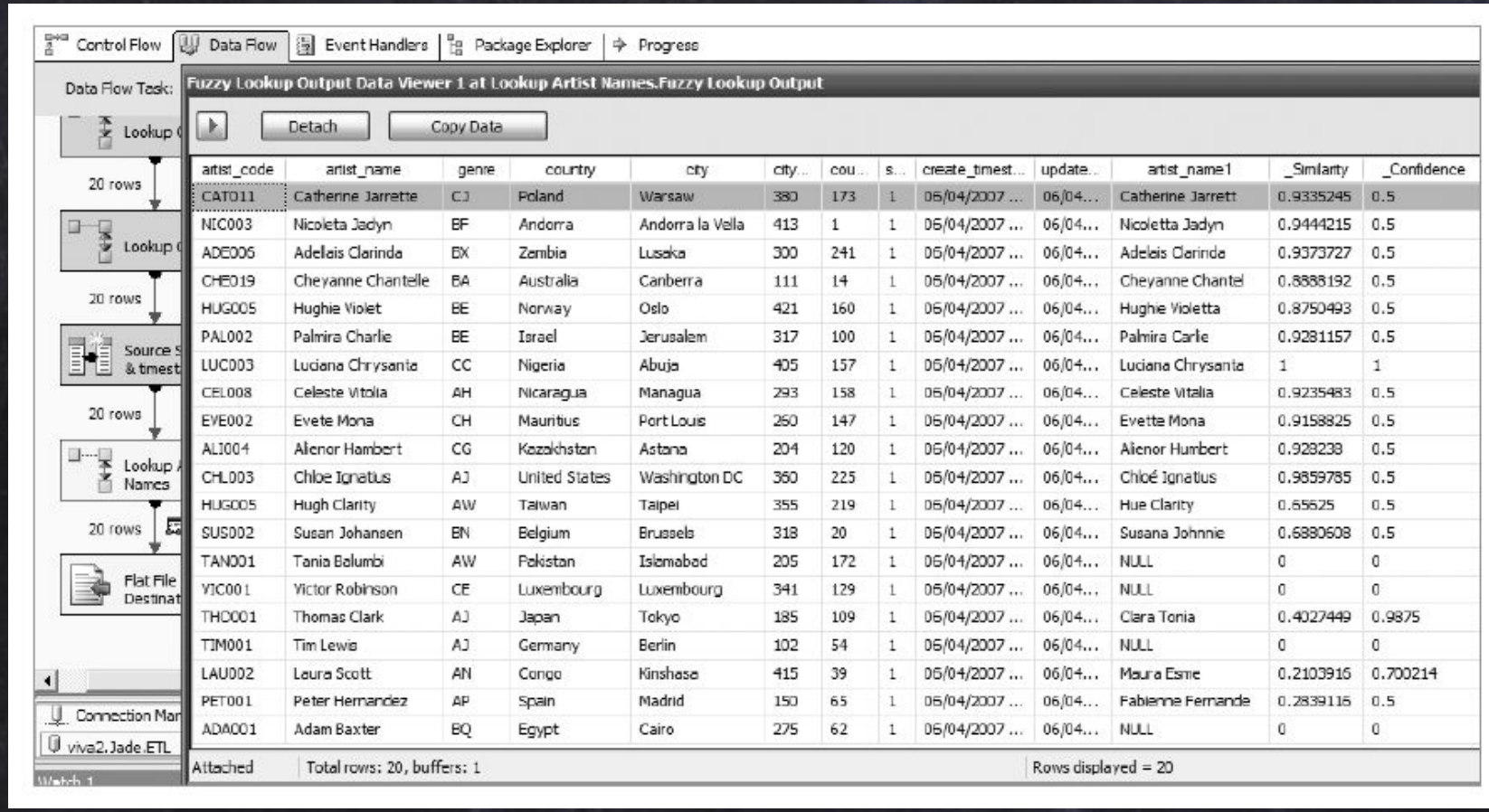
Exemplo de utilização de fuzzy para Match

artist_code	artist_name	genre	country	city
CAT011	Catherine Jarrette	CJ	Poland	Warsaw
NIC003	Nicoleta Jadyn	BF	Andorra	Andorra la Vella
ADE006	Adellais Clarinda	BX	Zambia	Lusaka
CHE019	Cheyanne Chantelle	BA	Australia	Canberra
HUG005	Hughie Violet	BE	Norway	Oslo
PAL002	Palmira Charlie	BE	Israel	Jerusalem
LUC003	Luciana Chrysanta	CC	Nigeria	Abuja
CEL008	Celeste Vitolia	AH	Nicaragua	Managua
EVE002	Evete Mona	CH	Mauritius	Port Louis
ALI004	Alienor Hambert	CG	Kazakhstan	Astana
CHL003	Chloe Ignatius	AJ	United States	Washington DC
HUG005	Hugh Clarity	AW	Taiwan	Taipei
SUS002	Susan Johansen	BN	Belgium	Brussels
TAN001	Tania Balumbi	AW	Pakistan	Islamabad
VIC001	Victor Robinson	CE	Luxembourg	Luxembourg
THO001	Thomas Clark	AJ	Japan	Tokyo
TIM001	Tim Lewis	AJ	Germany	Berlin
LAU002	Laura Scott	AN	Congo	Kinshasa
PET001	Peter Hernandez	AP	Spain	Madrid
ADA001	Adam Baxter	BQ	Egypt	Cairo

Exemplo de utilização de fuzzy para Match

- ✘ Destes 20 artistas apresentados, nove são novos artistas. Onze deles já existem na NDS, mas seus nomes podem estar um pouco diferentes
- ✘ Por exemplo, um que já existe na NDS é "Catherine Jarrett", mas o dado que está entrando é "Catherine Jarrette"
- ✘ Usa-se uma transformação conhecida como Lookup Fuzzy com certos níveis de similaridade para determinar quando o artista já existe na NDS e assim será atualizado ou é um novo artista e será incluído
- ✘ Cidade e país serão também traduzidos em chaves da NDS, porém usando um lookup normal (exact match)

Exemplo de utilização de fuzzy para Match



artist_code	artist_name	genre	country	city	city...	cou...	s...	create_timest...	update...	artist_name1	_Similarity	_Confidence
CAT011	Catherine Jarrette	CJ	Poland	Warsaw	380	173	1	06/04/2007 ...	06/04...	Catherine Jarrett	0.9335245	0.5
NIC003	Nicoleta Jadyn	BF	Andorra	Andorra la Vella	413	1	1	06/04/2007 ...	06/04...	Nicoletta Jadyn	0.9444215	0.5
ADE005	Adelais Clarinda	BX	Zambia	Lusaka	300	241	1	06/04/2007 ...	06/04...	Adelais Clarinda	0.9373727	0.5
CHE019	Cheyanne Chantelle	BA	Australia	Canberra	111	14	1	06/04/2007 ...	06/04...	Cheyanne Chantel	0.8888192	0.5
HUG005	Hughie Violet	BE	Norway	Oslo	421	160	1	06/04/2007 ...	06/04...	Hughie Violetta	0.8750493	0.5
PAL002	Palmira Charlie	BE	Israel	Jerusalem	317	100	1	06/04/2007 ...	06/04...	Palmira Carle	0.9281157	0.5
LUC003	Luciana Chrysanta	CC	Nigeria	Abuja	405	157	1	06/04/2007 ...	06/04...	Luciana Chrysanta	1	1
CEL008	Celeste Vitalia	AH	Nicaragua	Managua	293	158	1	06/04/2007 ...	06/04...	Celeste Vitalia	0.9235483	0.5
EVE002	Evete Mona	CH	Mauritius	Port Louis	250	147	1	06/04/2007 ...	06/04...	Evette Mona	0.9158825	0.5
ALI004	Alienor Hambert	CG	Kazakhstan	Astana	204	120	1	06/04/2007 ...	06/04...	Alienor Humbert	0.928238	0.5
CHL003	Chloe Ignatius	AJ	United States	Washington DC	360	225	1	06/04/2007 ...	06/04...	Chloé Ignatius	0.9859785	0.5
HUG005	Hugh Clarity	AW	Taiwan	Taipei	355	219	1	06/04/2007 ...	06/04...	Hue Clarity	0.65625	0.5
SUS002	Susan Johansen	BN	Belgium	Brussels	318	20	1	06/04/2007 ...	06/04...	Susana Johnnie	0.6880608	0.5
TAN001	Tania Balumbi	AW	Pakistan	Islamabad	205	172	1	06/04/2007 ...	06/04...	NULL	0	0
VIC001	Victor Robinson	CE	Luxembourg	Luxembourg	341	129	1	06/04/2007 ...	06/04...	NULL	0	0
THO001	Thomas Clark	AJ	Japan	Tokyo	185	109	1	06/04/2007 ...	06/04...	Clara Tonia	0.4027449	0.9875
TIM001	Tim Lewis	AJ	Germany	Berlin	102	54	1	06/04/2007 ...	06/04...	NULL	0	0
LAU002	Laura Scott	AN	Congo	Kinshasa	415	39	1	06/04/2007 ...	06/04...	Maura Esme	0.2103916	0.700214
PET001	Peter Hernandez	AP	Spain	Madrid	150	65	1	06/04/2007 ...	06/04...	Fabienne Fernande	0.2839116	0.5
ADA001	Adam Baxter	BQ	Egypt	Cairo	275	62	1	06/04/2007 ...	06/04...	NULL	0	0

Exemplo de utilização de fuzzy para Match

- Neste exemplo de execução do SSIS, mostra-se que as 11 linhas que já existiam na NDS tiveram uma pontuação de similaridade acima de 85% e as outras não

- Assim pode-se definir este como um bom nível de similaridade para este caso e teremos então uma carga de informações por similaridade e não exata

- O que estiver abaixo do nível aceitável definido de similaridade será incluído como um novo registro por exemplo, em vez de ser apenas atualizado

Cross-checking

Cross-Checking

CROSS-CHECKING COM FONTES EXTERNAS

- Algumas atividades de data cleansing são feitas internamente, nos dados que existem no DW ou nos sistemas-fontes

- Em alguns caos, necessita-se verificar fontes externas de dados, por exemplo, garantir que um cep está correto usando a base dos correios, garantir que um CPF é válido de acordo com um web service da Receita Federal etc.

- Os dados externos são disponibilizados em vários formatos e maneiras. Isto afeta o jeito como integramos estes dados com nosso DW. Pode ser um CD ou DVD, pode ser uma aplicação que deve ser instalada, um EDI, um webservice etc.

CROSS-CHECKING COM FONTES EXTERNAS

- Baseado em como o dado ficará disponível para a integração com o DW, pode-se classificar os mecanismos em duas grandes abordagens:

- Trazemos os dados externos para dentro do DW
  - O fornecedor nos entrega os dados e nós os armazenamos no DW
  - Exemplo: quando carregamos a base dos correios em nosso DW

- Não trazemos os dados externos para o DW
  - O dado fica com o fornecedor e nós apenas buscamos o dado quando precisamos dele
  - Exemplo: quando usamos webservices da receita federal

Regras de DQ

Regras de data quality

Regras de data quality

- ✘ Cross-reference validation
  - Onde verificamos os dados que estão entrando contra os dados que já estão no DW
  - O objetivo é garantir que o valor do dado que está entrando esteja dentro de um intervalo que é calculado baseado nos dados que já estão no DW
  - Exemplo: espera-se que o dado que está entrando esteja dentro de um intervalo de 25% da média da coluna 1 de uma tabela qualquer
- ✘ DW internal validation
  - Onde verificamos o dado que já foi carregado no DW
  - O objetivo é verificar a qualidade dos dados em um nível agregado
  - Em outras palavras, os registros individuais podem estar corretos, porém, os totais estão incorretos
  - Você pode comparar o número com um valor padrão conhecido
  - Exemplo: o número de assinaturas da última semana é igual a 5.000 sendo

LOGGING

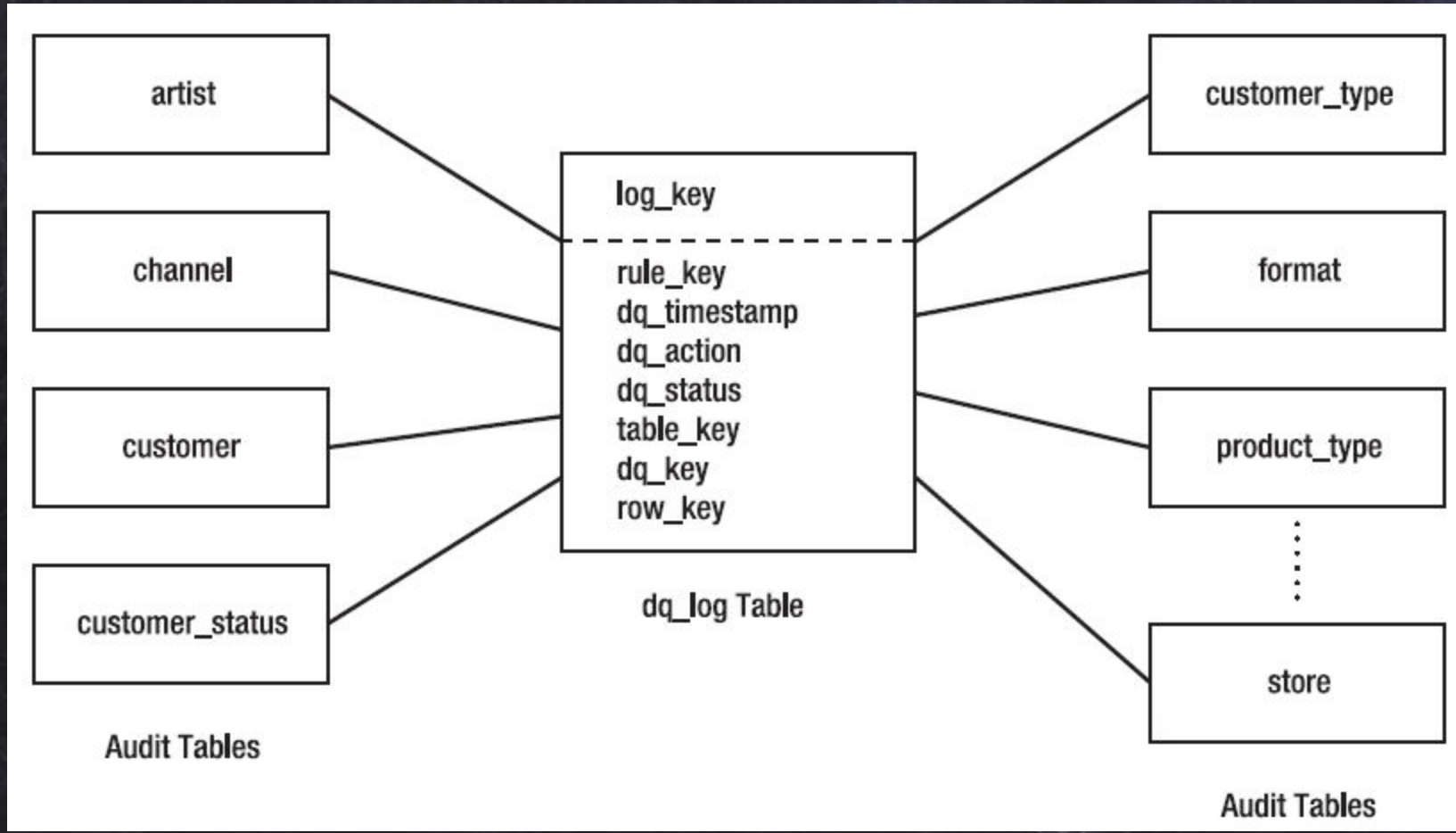
- Quando uma regra de data quality é violada, armazena-se o evento em uma base de dados de data quality
- Armazena-se a hora em que ocorreu a violação, qual regra foi violada, qual ação foi tomada e o status de correção em uma tabela chamada DQ log
- Também é necessário armazenar as linhas que violaram as regras de qualidade.
- Este processo de registrar os eventos de violação de regras, assim como as linhas que violaram as regras é conhecido como **data quality logging**

LOGGING

- ✘ As linhas que violaram as regras são armazenadas também na base de dados de qualidade, porém não na tabela de log
- ✘ Elas são armazenadas em tabelas chamadas de auditoria
- ✘ A estrutura destas tabelas é a mesma da tabela de destino no DW, mesmo nome, colunas etc.
- ✘ A diferença entre a estrutura da tabela de auditoria e a tabela destino do dados está em uma coluna adicional – a chave primária . A tabela de auditoria tem como chave primária uma SK. Esta chave primária é então armazenada na tabela DQ log como chave estrangeira

FIAP

LOGGING



AUDITORIA

- A auditoria da qualidade de dados refere-se a fazer consultas nas tabelas de auditoria e de log com o objetivo de encontrar quando uma regra foi violada, quais e quantas linhas foram impactadas, as tabelas fonte e destino envolvidas etc.

- Por fim, o propósito é descobrir a qualidade dos dados no DW

- Isto pode ser feito avaliando quantas regras foram violadas nos meses recentes, quantas vezes cada regra, se isso é impactante para o negócio etc.

- As tabelas de auditoria e log devem ser destruídas e arquivadas regularmente para não ocupar muito espaço do DW

FIAP

AUDITORIA

Data Quality Reports

- Unregistered Payments → Music Films Books
- Monthly Registrations
- Cost Comparison
- Unrecorded Campaigns

Customer Data

- Double Counting
- Blank Email Address
- Supplier Performance
- Comm. Subscription
- Campaign Results

Campaign Management

- Invalid Bounce Dates
- Negative Delivery
- Cross Rates Predictions
- Unknown Customers
- Duplicate Addressee

Amadeus Entertainment Group

Data Quality Dashboard

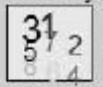
Overall Quality

What Is Data Quality?

Why is data quality important? It is not difficult to answer this simple question. Imagine if you could not trust the data in your data warehouse. Perhaps because some data is incorrect (accuracy), or because some data is missing (completeness),

Data Quality Process

Accuracy vs Integrity



This week we will focus our attention on the data integrity and accuracy within the campaign response and supplier performance area taking into account the decree-

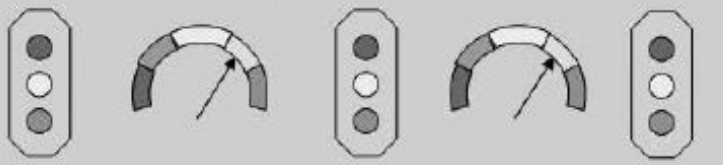
Documentation

Reports Process Guide

Helpdesk

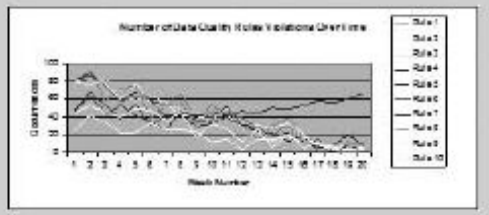
FAQ Number New Issue

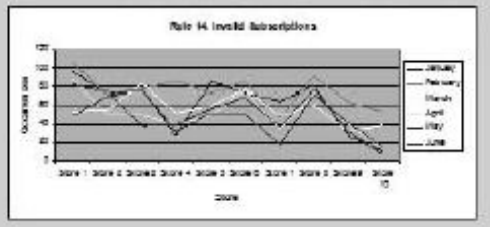
Key Data Quality Indicators

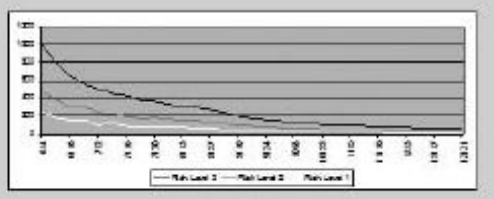


Sales Finance Supplier CRM Product

Critical Data Quality Charts







Referências

- KIMBALL, R., ROSS, M. The Data Warehouse Toolkit. 2ª ed., John Wiley Professional, 2002.
- MACHADO, F. N. R. Tecnologia e Projeto de Data Warehouse. 1ª ed., São Paulo: Ed. Érica, 2004.

FIAP

OBRIGADO!