

FIAP GRADUAÇÃO

DATA SCIENCE

DATA GOVERNANCE & DATA SECURITY MANAGEMENT

Prof. Dr. Renê de Ávila Mendes

Objetivos da disciplina

DISCIPLINA: Data Governance & Data Security Management

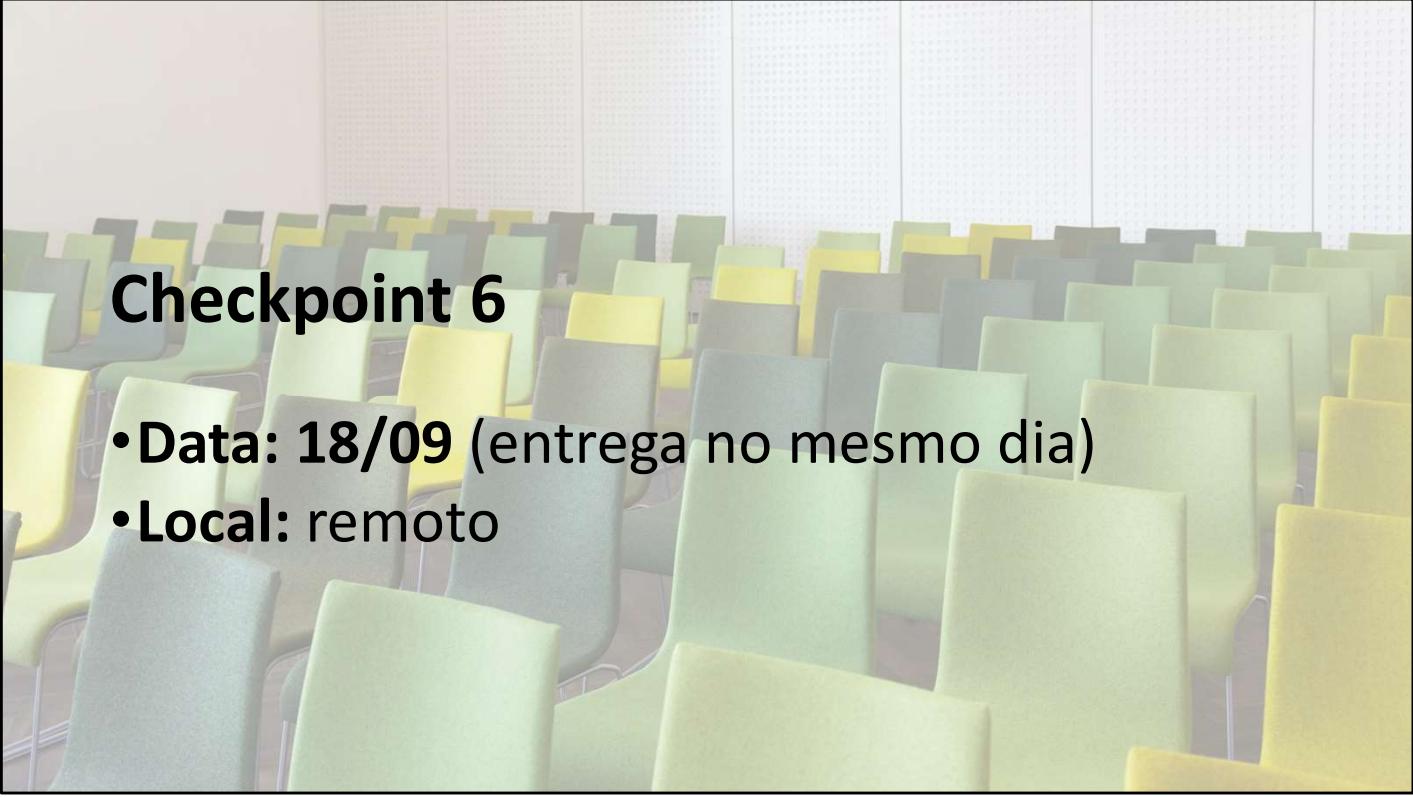
OBJETIVOS: Descubra como funciona um **projeto de banco de dados** dentro de um ambiente corporativo, aplicando **técnicas de levantamento e documentação de requisitos**, aderente aos projetos de bancos de dados e aprenda a representar esses requisitos em arquiteturas de solução tecnológica para Data distribution e Data integration, modelos de estruturas de dados e dicionários de dados buscando **Data quality**. Garanta a qualidade dos dados de uma empresa para prover os melhores subsídios à tomada de decisão de negócio, praticando **Data cleaning** para limpar, harmonizar, complementar e corrigir dados inconsistentes, incompletos ou incorretos. Compreenda como funciona o **ciclo de vida da informação** e as responsabilidades administrativas sobre os dados de negócio, buscando qualidade, segurança e compatibilidade com políticas de administração de informação corporativas auditáveis, aplicando práticas atuais de **Data profiling** e conhecendo os princípios de **Data auditing**, de forma a atender a **Lei Geral de Proteção de Dados (LGPD)**.

Assuntos – 2º Semestre

- Qualidade em metadados
- Arquiteturas de integração e distribuição física de banco de dados
- Master Data Management e Data Hub
- **Qualidade de dados**
- Enterprise Data Management
- LGPD

Próximas aulas

- **11/09 - Revisão de conteúdo**
- 18/09 - Checkpoint 6
- 25/09 - Aula para preparação para Challenge Sprint 4
- 02/10 - EDM aula 1
- 09/10 - EDM aula 2
- 16/10 - Segurança e Auditoria aula 1
- 23/10 - Segurança e Auditoria aula 2
- 30/10 - LGPD aula 1
- 06/11 - LGPD aula 2
- 13/11 - Kick-off Global Solutions
- 20/11 - Fériado
- 27/11 - Global Solutions
- 04/12 - Substitutiva



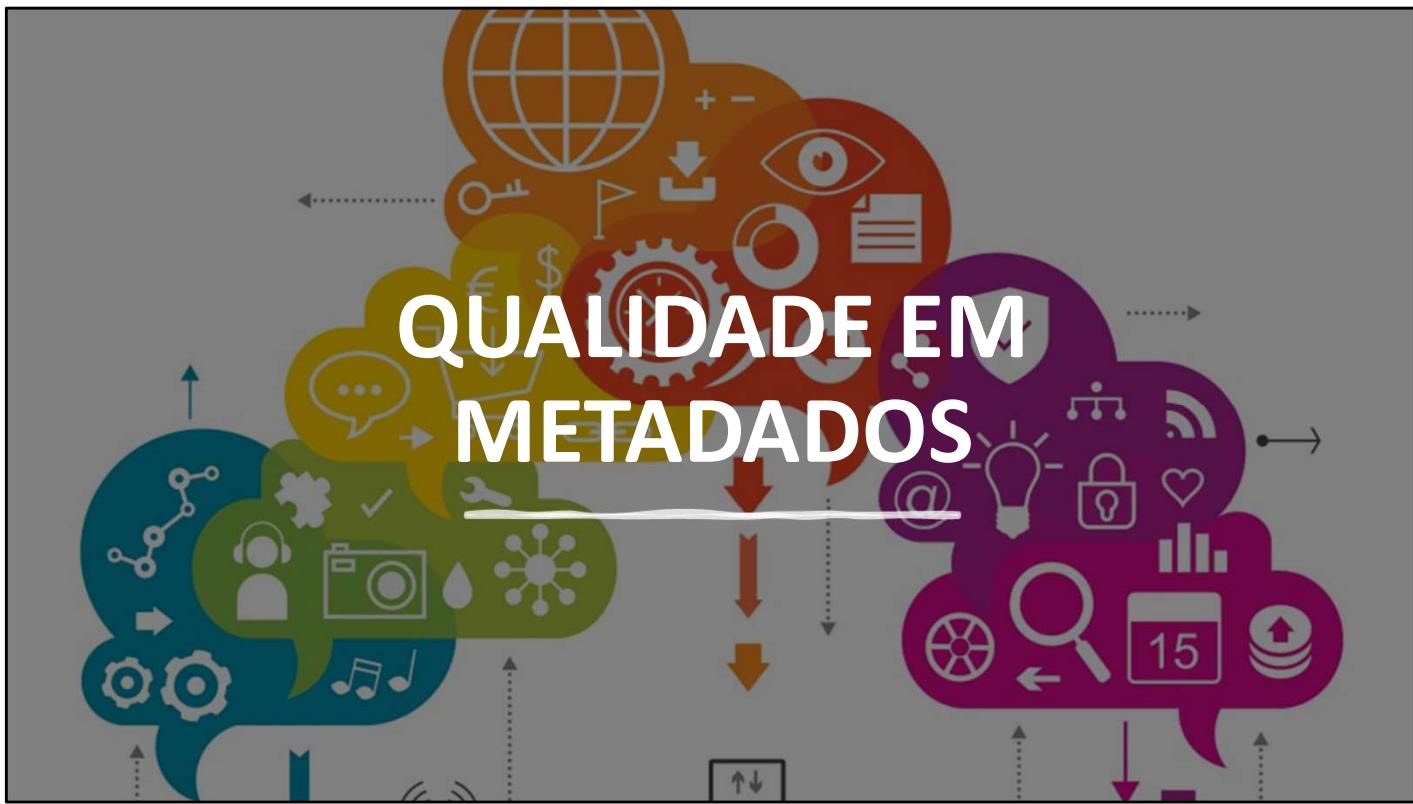
Checkpoint 6

- **Data:** 18/09 (entrega no mesmo dia)
- **Local:** remoto

REVISÃO DE CONCEITOS

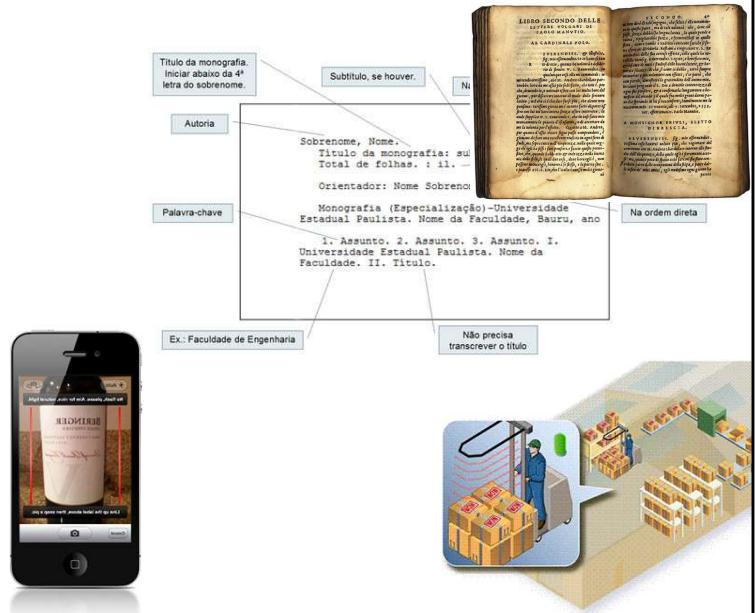


QUALIDADE EM METADADOS



Metadados – O que são

- Dados sobre os dados
- Etiquetas descritivas
- Descrevem estruturas de informação
- Os dados definem objetos reais assim como os metadados definem os dados

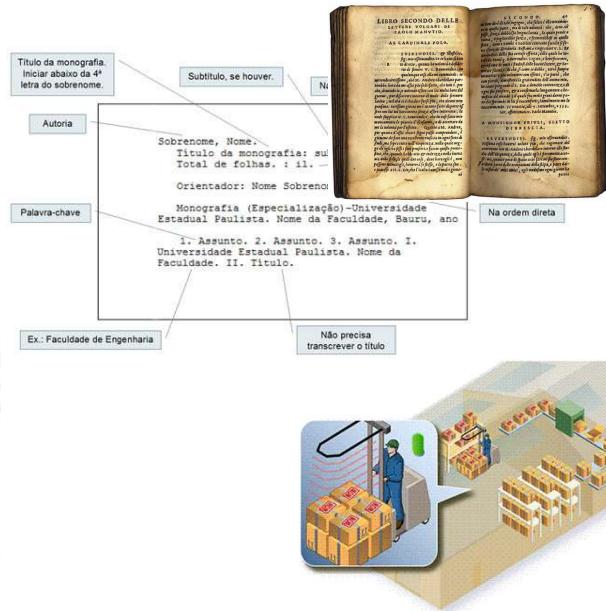


Enfatizar o ganho que há em utilizar uma ficha catalográfica quando um livro precisa ser localizado

Metadados – Outras definições

Metadados segundo a tecnologia:

“Tudo o que é necessário saber para executar”



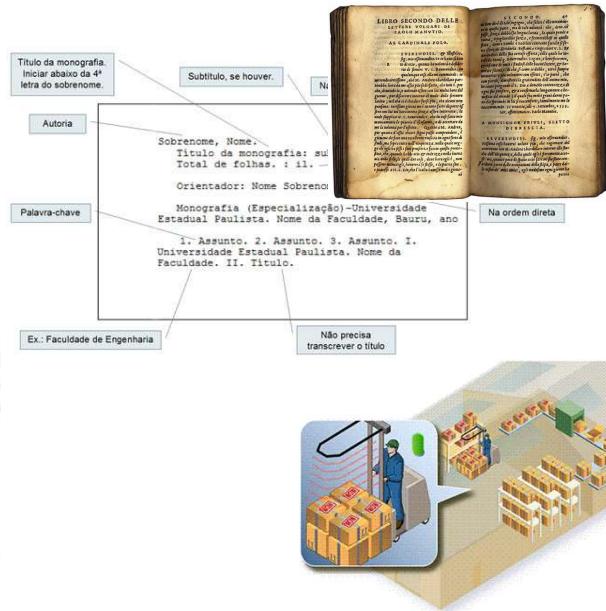
Fontes:

IMCue Solutions (<http://imcue.com>);

Metadados – Outras definições

Metadados definido pelo uso:

“O que é necessário saber para usar”



Fontes:

IMCue Solutions (<http://imcue.com>);

Metadados – Áreas potenciais

- Análise de negócios
- Arquitetura de negócios
- Definições de negócios
- Regras de negócio
- **Governança de dados**
- **Integração de dados**
- **Qualidade dos dados**
- Gestão de conteúdo de documentos
- **Infraestrutura de tecnologia de informações**

1. Análise de negócios: definições de dados, relatórios, os usuários, utilização, desempenho.
2. Arquitetura de negócios: papéis/funções e organizações, metas e objetivos.
3. Definições de negócios: os termos do negócio e explicações para um determinado conceito, fato ou outro item encontrado em uma organização.
4. Regras de negócio: padrão de cálculo e métodos de derivação.
5. Governança de dados: políticas, normas, procedimentos, programas, funções, organizações, atribuições de serviços.
6. Integração de dados: origens, destinos, transformações, linhagem, fluxos de trabalho ETL, EAI, EII, migração/conversão.
7. Qualidade dos dados: defeitos, métricas, classificações.
8. Gestão de conteúdo de documentos: dados não estruturados, documentos, taxonomias, ontologias, conjuntos de nome, descobertas/coberturas legais, índices de mecanismo de pesquisa.
9. Infraestrutura de tecnologia de informações: plataformas, redes, configurações de licenças.

Metadados – Áreas potenciais

- Modelos de dados lógicos
- Modelos de dados físicos
- Modelos de processo
- Portfólio de sistemas e governança de TI
- Informações da Arquitetura orientada a serviços (SOA)
- Design e desenvolvimento de Sistemas
- Gestão de sistemas

10. Modelos de dados lógicos: entidades, atributos, relacionamentos e regras, nomes comerciais e definições.
11. Modelos de dados físicos: arquivos, tabelas, colunas, exibições, definições de negócio, índices, uso, desempenho, gestão de alterações.
12. Modelos de processo: funções, atividades, funções/papéis, entradas/saídas, fluxo de trabalho, regras de negócio, calendários, armazenagem.
13. Portfólio de sistemas e governança de TI: bancos de dados, aplicações, projetos e programas, roteiro de integração, gestão de mudanças.
14. Informações da Arquitetura orientada a serviços (SOA): componentes, serviços, mensagens, dados mestres.
15. Design e desenvolvimento de Sistemas: requisitos, projetos e planos de teste, impacto.
16. Gestão de sistemas: segurança de dados, licenças, configuração, confiabilidade, níveis de serviço.

Metadados – Classificação quanto ao uso



Metadados de negócio - Documentam os elementos de negócio. **Exemplos** - Cálculos, algoritmos, dados corporativos, modelos de dados, valores de domínio, dados não estruturados etc.

Metadados técnicos e operacionais - Associados a elementos de desenvolvimento e implementação. **Exemplos** - Propriedades de modelos físicos de dados, armazenamento, comentários de código, agendamento de execução, regras de backup e recover etc.

Metadados de processos - Definem e descrevem características de outros elementos do sistema. São uma especialidade dos metadados de negócio. **Exemplos** - Processos, regras de negócio, programas, empregos, ferramentas etc.

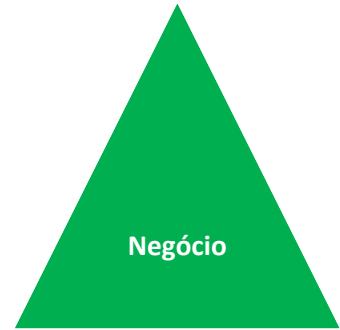
Metadados de administração – Dados sobre a gestão de dados. **Exemplos** – AD, regras, partilhas, CRUD, usuários de dados, estruturas e responsabilidades etc.

Metadados de negócio

Documentam os elementos de negócio.

EXEMPLOS:

- Descrições de processos de negócio
- Regras e algoritmos de negócio
- Análise da linhagem dos dados e impactos
- Modelos de dados conceitual e lógico
- Demonstrações de qualidade dos dados
- Informações sobre manejo dos dados
- Restrições regulatórias ou contratuais
- Restrições de valor
- Domínios de valor



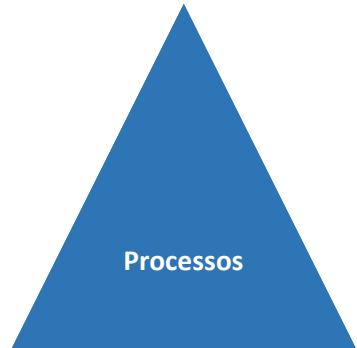
Metadados de negócio - Documentam os elementos de negócio. **Exemplos** - Cálculos, algoritmos, dados corporativos, modelos de dados, valores de domínio, dados não estruturados etc.

Metadados de processos

Descrevem outros elementos do sistema.
Especialização dos metadados de negócio.

EXEMPLOS:

- Armazém de dados e dados envolvidos
- Organismos regulatórios e governamentais
- Donos das organizações e partes interessadas
- Dependências entre processos
- Nome e ordem do processo e seu cronograma
- Papéis e responsabilidades
- Atividades da cadeia de valor



Metadados de processos - Definem e descrevem características de outros elementos do sistema. São uma especialidade dos metadados de negócio. **Exemplos** - Processos, regras de negócio, programas, empregos, ferramentas etc.

Metadados de administração

Descrevem a estrutura da gestão dos dados. Especialização dos metadados de negócio.

EXEMPLOS:

- Geradores de negócio/metas
- Regras de dados CRUD
- Definições de dados técnicos e de negócios
- Proprietários e usuários dos dados
- Regras de partilha de dados e contratos
- Administradores de dados, funções e responsabilidades
- Áreas de assunto dos dados
- Organização da governança



Metadados de administração – Dados sobre a gestão de dados. **Exemplos** – AD, regras, partilhas, CRUD, usuários de dados, estruturas e responsabilidades etc.

Metadados – Classificação quanto à função

DESCRITIVOS – descrevem um recurso com o propósito, por exemplo, de descoberta ou identificação. Isso pode incluir elementos como título, resumo, autor e palavras-chave.

ESTRUTURAIS – indicam como objetos compostos são colocados juntos, por exemplo, como é que páginas são ordenadas para formar capítulos.

ADMINISTRATIVOS – fornecem informações para auxiliar no gerenciamento de um recurso, como por exemplo, quando e como este foi criado, tipo de arquivo e outras informações técnicas, e sobre quem tem acesso a ele.

MD DESCRIPTIVOS – descrevem um recurso com o propósito, por exemplo, de descoberta ou identificação. Isso pode incluir elementos como título, resumo, autor e palavras-chave.

MD ESTRUTURAIS (áudio, vídeo, e-mail, XML) – indicam como objetos compostos são colocados juntos, por exemplo, como é que páginas são ordenadas para formar capítulos .Dublin Core, estruturas de campos, dicionário de rótulos e palavras-chave, esquemas XML

MD ADMINISTRATIVOS – fornecem informações para auxiliar no gerenciamento de um recurso, como por exemplo, quando e como o mesmo foi criado, tipo de arquivo e outras informações técnicas, e sobre quem tem acesso a ele. Existem vários subconjuntos de dados administrativos; dois deles, às vezes, são listados separadamente como tipos metadados:

- Metadados para gerenciamento de direitos, que tratam dos direitos de propriedade intelectual,
- e
- Metadados para preservação, que contêm informações necessárias ao arquivamento e à preservação de um determinado recurso.

EXEMPLOS CORPORATIVOS DE PADRÕES



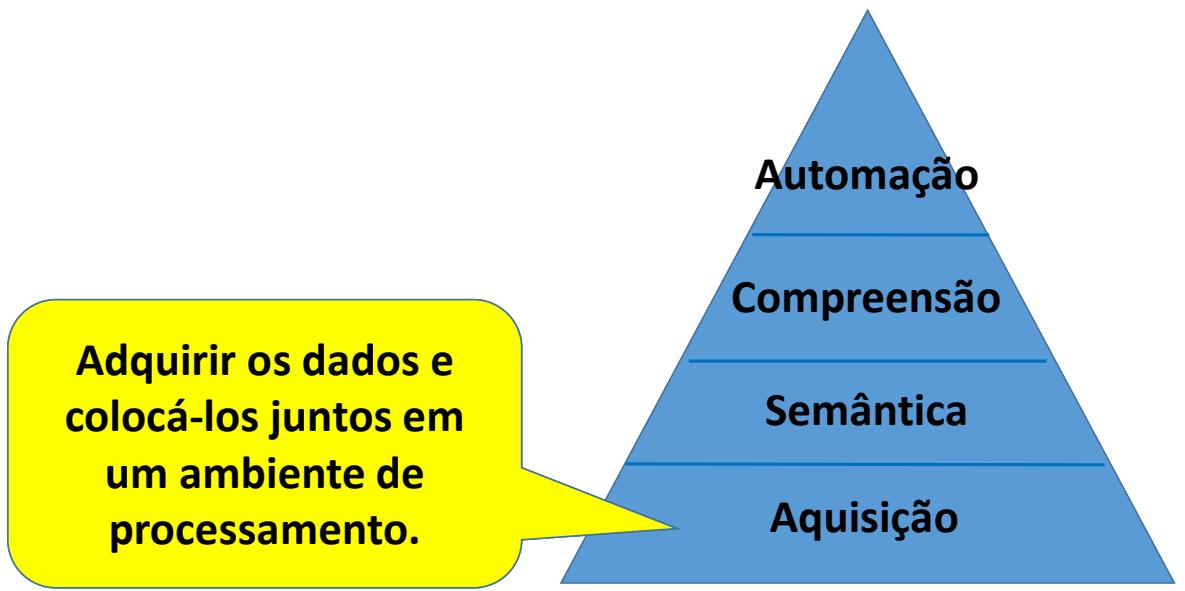
INTEGRAÇÃO DE DADOS ENTRE BASES



Técnicas de integração de dados

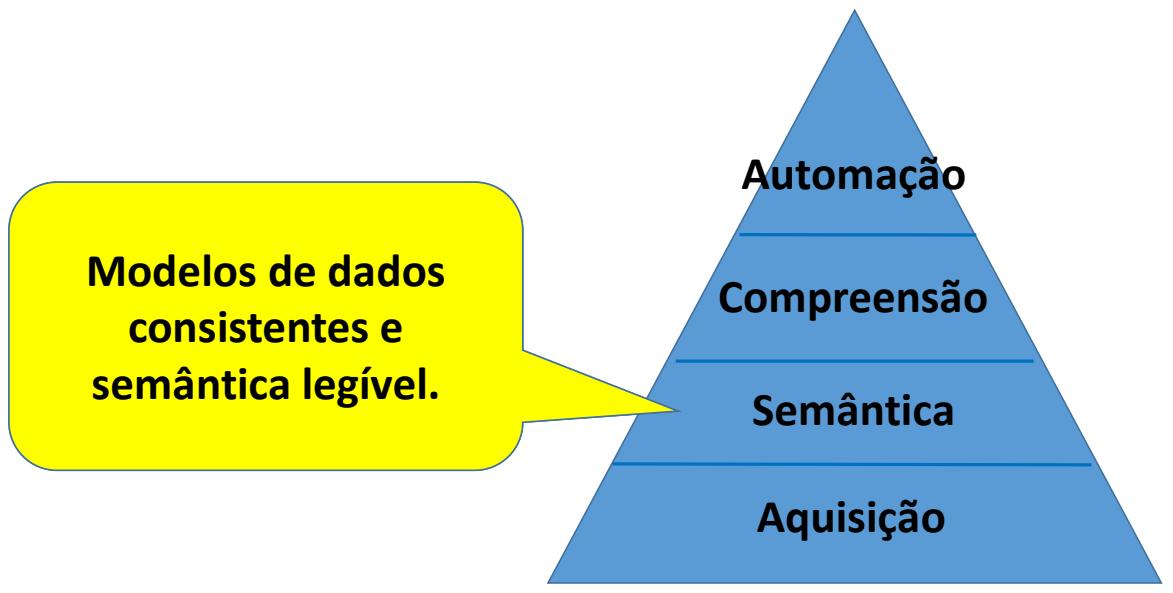
- Dados são considerados ativos corporativos
- Os ativos corporativos devem ser gerenciados
- O compartilhamento dos dados entre sistemas é necessário para que o dado seja:
 - Adquirido
 - Enriquecido
 - Corrigido
 - Consumido

Hierarquia de uso dos dados



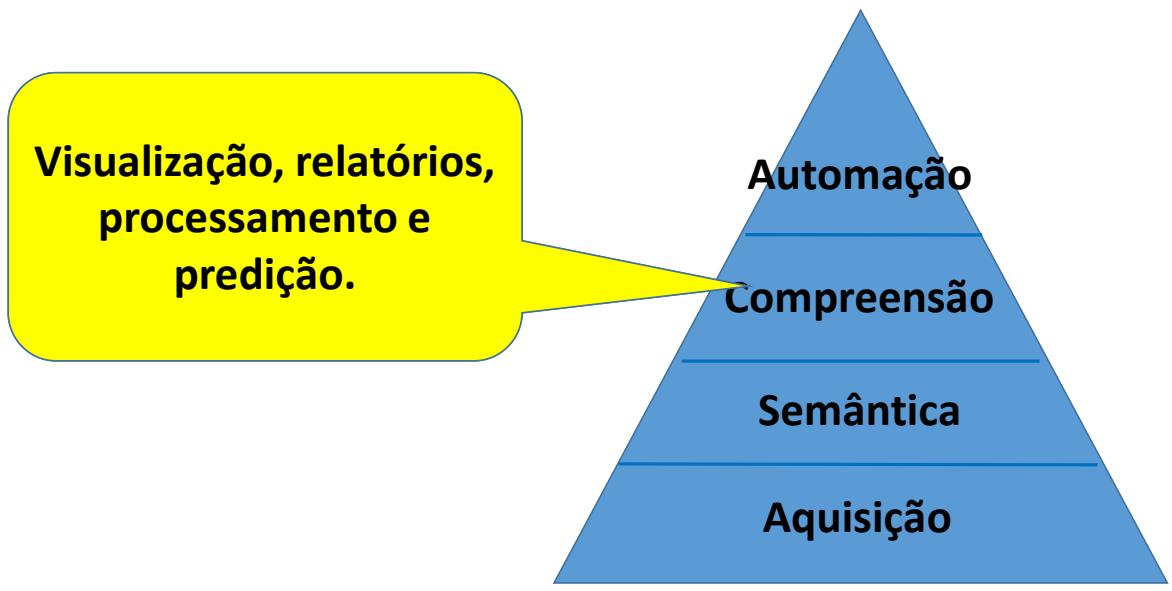
KREPS, Jay. *I Heart Logs: Event Data, Stream Processing, and Data Integration.* "O'Reilly Media, Inc.", 2014, p. 12.

Hierarquia de uso dos dados



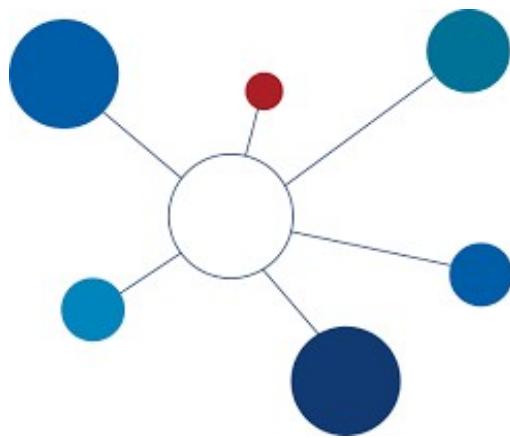
KREPS, Jay. *I Heart Logs: Event Data, Stream Processing, and Data Integration.* "O'Reilly Media, Inc.", 2014, p. 12.

Hierarquia de uso dos dados



KREPS, Jay. **I Heart Logs: Event Data, Stream Processing, and Data Integration.** "O'Reilly Media, Inc.", 2014, p. 12.

Integração de Dados - Definição



Tornar disponíveis todos os dados que uma organização possui a todos os serviços e sistemas que necessitem deles.

KREPS, Jay. *I Heart Logs: Event Data, Stream Processing, and Data Integration.* "O'Reilly Media, Inc.", 2014, p. 11.

Cópia em mídia externa

- Método útil para situações em que os dados precisam chegar ao destino mais rápido do que outros métodos podem resolver
- Casos de uso:
 - Migração de dados offline entre data centers (devido a mudança de fornecedor de cloud, por exemplo)
 - Importação de dados para cloud
 - Transporte de dados críticos (diminui o risco de interceptação ou violação)

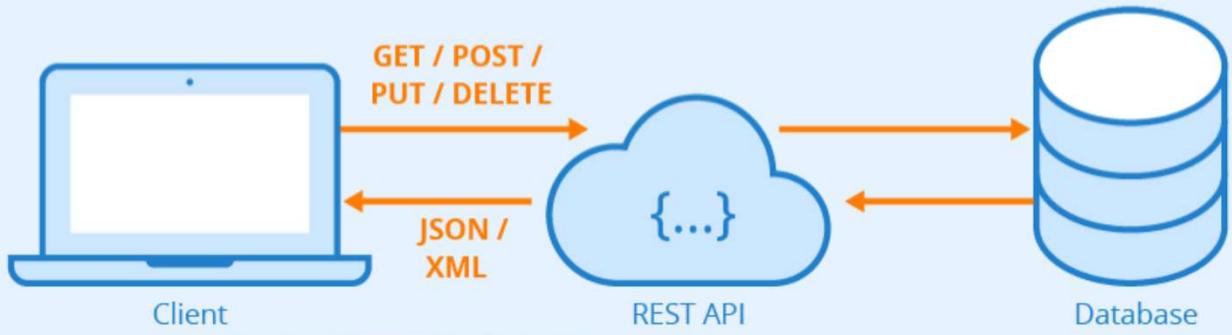
<https://docs.aws.amazon.com/snowball/latest/developer-guide/whatisedge.html>

Cópia em mídia externa - Exemplos

- AWS Snowball
 - 210 Terabytes
- AWS SnowMobile
 - 100 Petabytes
 - 26 anos a 10 Gbits



<https://docs.aws.amazon.com/snowball/latest/developer-guide/whatisedge.html>



File Transfer

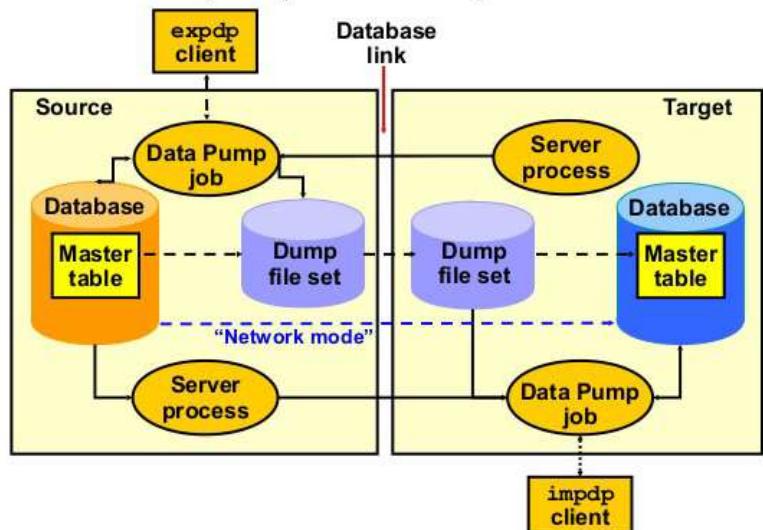
- Cópia de arquivos
- Normalmente entre servidores, mas pode ser utilizado por usuários finais
- Pode usar os
 - protocolos FTP e SFTP para arquivos
 - APIs REST para dados
- Mesmo conceito do EDI (Electronic Data Interchange)

<https://docs.aws.amazon.com/snowball/latest/developer-guide/whatisedge.html>

Data export/import

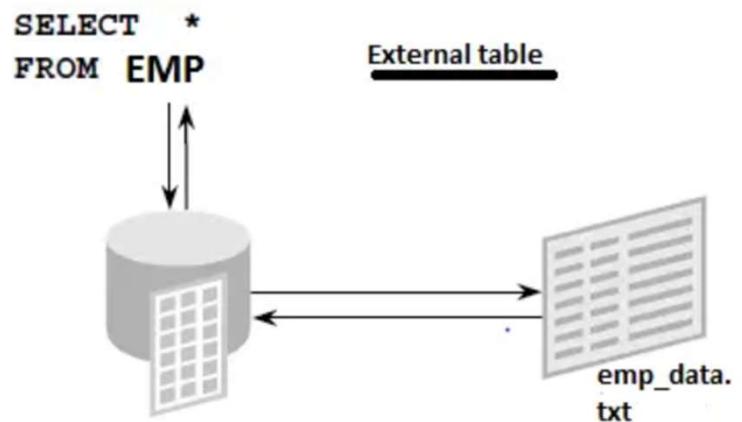
- Integração assíncrona de dados entre bases de dados
- Dados em formato binário proprietário ou SQL
- DDL em formato SQL

Data Pump Export and Import: Overview

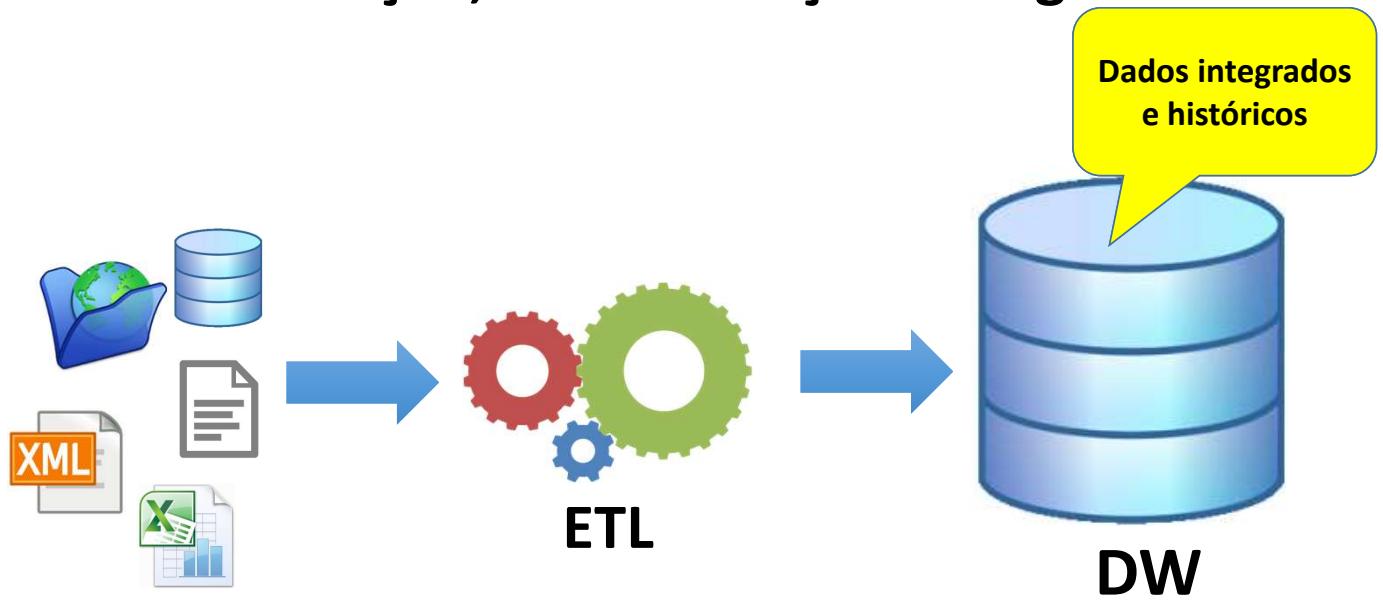


External table

- Integração offline de dados entre bases de dados
- Os dados são lidos do arquivo quando um SELECT é executado sobre a tabela externa



ETL – Extração, transformação e carga



Comparando ETL com ELT

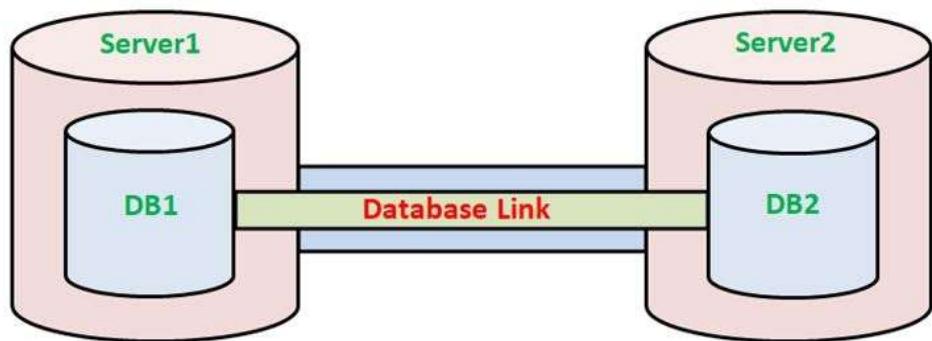


	ETL	ELT
Origem dos dados	Algumas	Todas
Transferência de dados	Lote	“Bulk” e streaming
Limpeza dos dados	Antes da carga	Prorrogada
Padronização de dados mestre	Antes da carga	Prorrogada

<https://www.lynda.com/Hadoop-tutorials/Comparing-big-data-ELT-traditional-ETL/385663/424483-4.html>

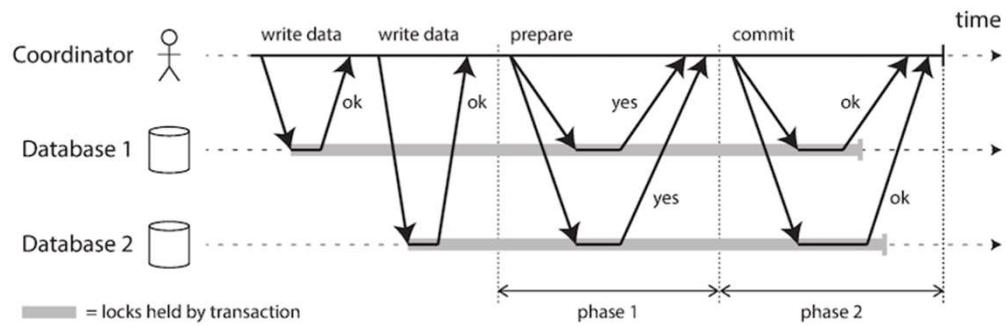
Database link

- Integração síncrona de dados entre bases de dados
 - Mesmo produto (Oracle X Oracle)
 - Produtos diferentes (Oracle x SQL Server)
- Requer uma conexão aberta entre os bancos



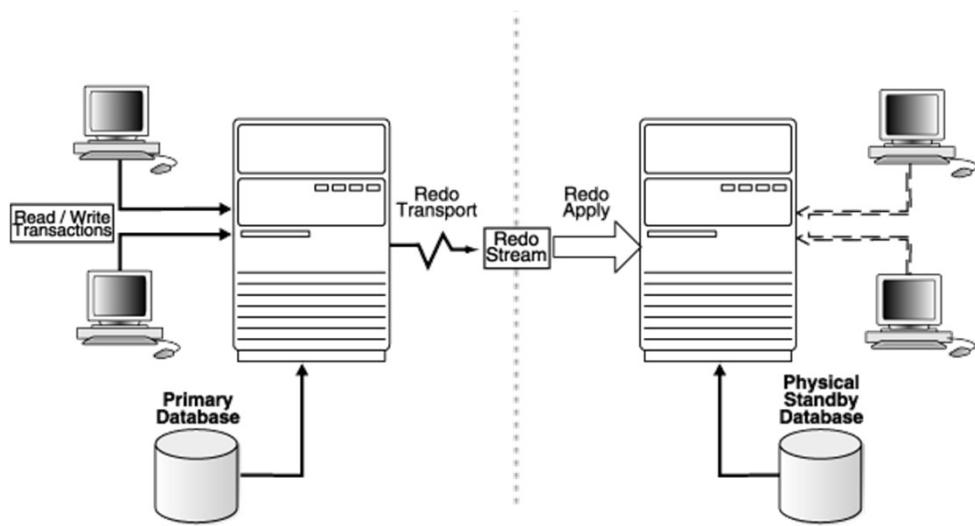
Replicação síncrona

- Síncrona (two-phase commit)



Replicação assíncrona

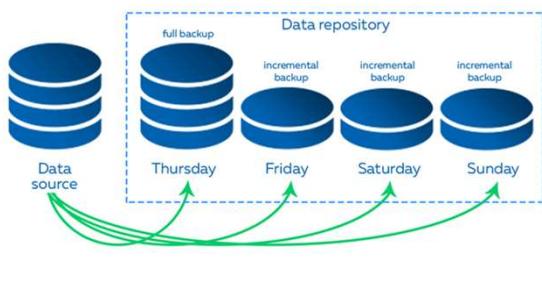
- Assíncrona (standby)



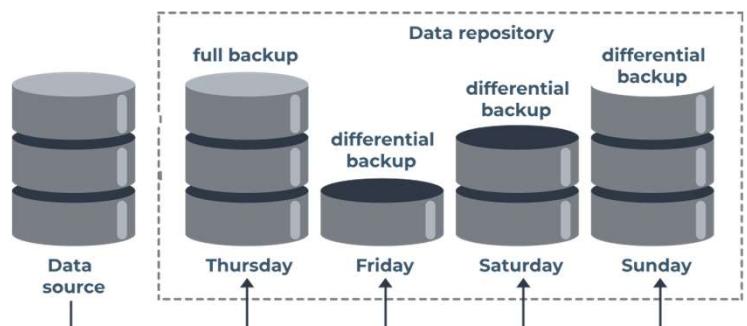
Backup e restore

- Cópia dos dados para uma mídia
- Permite a restauração dos dados
- Incremental e diferencial

Incremental



Differential

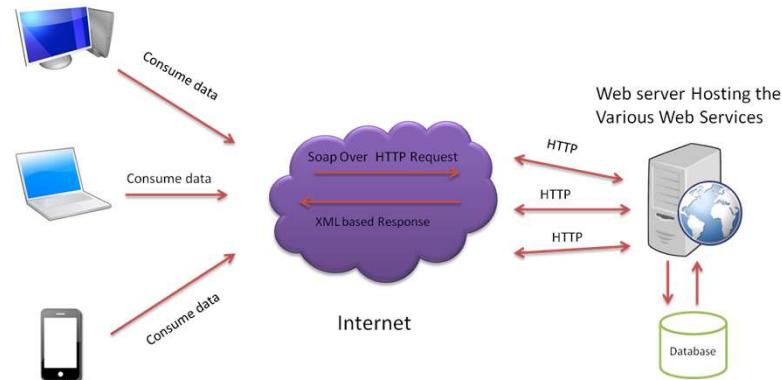


Snapshot

- Duplicação idêntica do banco de dados ou de um sistema inteiro
- Funcional: utilizado para duplicar um ambiente ou servidor ou uma base de dados

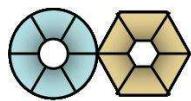
Web services

- Transferência de dados entre aplicações
- Utilizam protocolos como SOAP (Simple Object Access Protocol) e REST (Representationl Transfer Protocol)
- Trocam dados em XML, JSON, CSV etc.



MASTER DATA MANAGEMENT



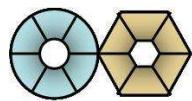


DMBOK – Gestão de dados Mestre e Referência



DEFINIÇÃO

“Planejar, implementar e controlar atividades para garantir consistência de dados Mestre e de Referência.”



MD/RDM – Alguns conceitos

Gestão de Dados
Mestre e
Referência

Planejar, implementar e
controlar atividades para
garantir a consistência dos
dados Mestre e de
Referência

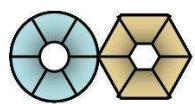
Gestão de Dados
de Referência

Elementos de
categorização de
outros dados

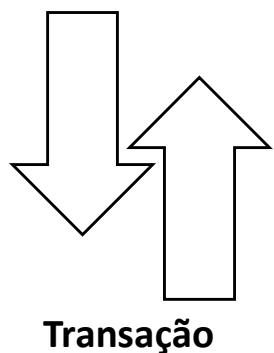
Gestão de Dados
Mestre

Dados
fundamentais do
negócio

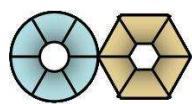
Dados de Referência são dados relacionados com códigos, tais como estado, país, status do pedido, tipo de movimentação do colaborador etc.



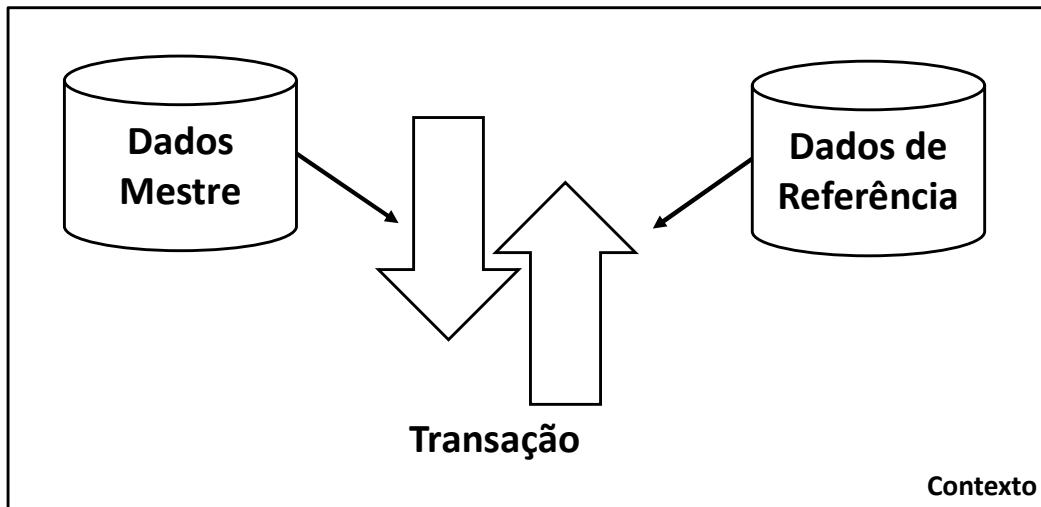
MD/RDM – Alguns conceitos



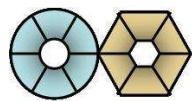
Dados mestres e de referência proveem o contexto para os dados transacionais.



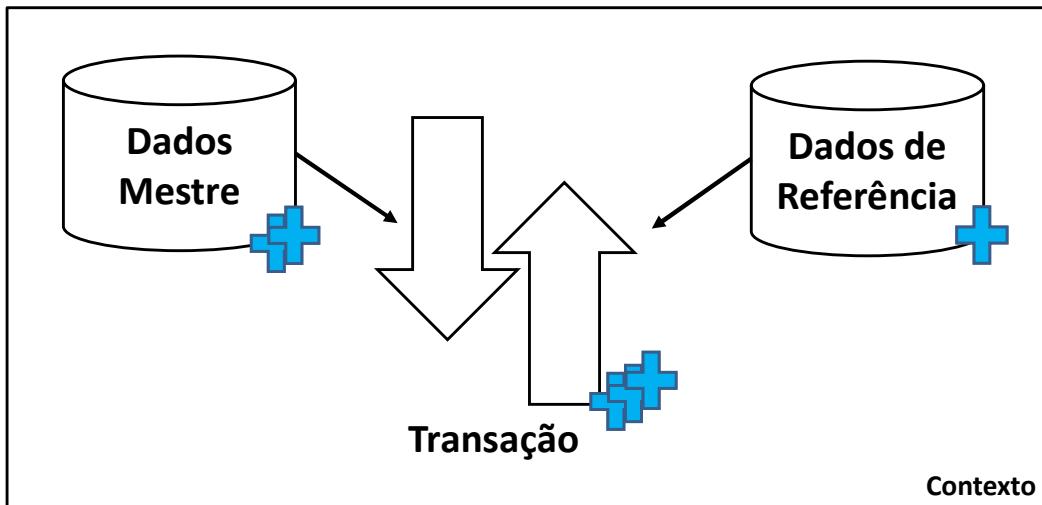
MD/RDM – Alguns conceitos



Dados mestres e de referência proveem o contexto para os dados transacionais.

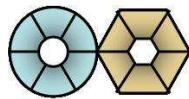


MD/RDM – Alguns conceitos



Volatilidade

Dados mestres e de referência proveem o contexto para os dados transacionais.



Dados mestre



Pessoas

Clientes, Funcionários,
Vendedores etc.

Locais

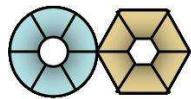
Imóveis, Instalações
etc.

Coisas

Produtos, Contas,
Contratos etc.

Os dados mestres são facilmente identificados em qualquer organização, pois são o foco dos processos de negócio. Uma vez que são categorizados por outros dados, tornam-se, então, os principais ativos de informação. Assim, um dado mestre é sempre crítico ao negócio. Eles são mais estáveis do que os dados que refletem as transações do dia a dia.

Fonte: FiapOn Fusion Master Data Management – Keylla Saes



Dados de referência



Dados de referência, por sua vez, são aqueles que permitem categorizar outros dados. As categorizações mais usuais são classificações ou agrupamentos desses dados, que podem se originar de processos internos da organização ou ter origem em dados externos. Habitualmente, são armazenados um código e sua descrição.

Fonte: FiapOn Fusion Master Data Management – Keylla Saes

PIERIM SUPERMERCADOS

Tradição e confiança!

Av. Mons Demosthenes Paraná Brasil Pontes, 1712
Jd Lavinia - Mococa – SP / CEP: 13737-632
Telefone 19 3665-3036 Email: supermercadopierim2@uol.com.br

O número a seguir deve constar de toda a correspondência, documentação de embarque e faturas pertinentes:
Nº DO PEDIDO: [100]

PARA:

Natacha Calazans do Nascimento
Edina Comércio de Bebidas
Cidade Universitária Zeferino Vaz
Campinas, s/número, cep 13084-971
19 3521 XXXX

PEDIDO DE COMPRA

Dados Mestre

Dados de Referência

Dados de Transacionais

ENVIAR PARA: (DADOS DA CENTRAL PIERIM)

Tiago Machado Flávio
Pierim Supermercados
Av Monsenhor Demosthenes Parana Brasil Pontes, 1712
Jd Lavinia – Mococa SP – CEP: 13737-632
19 3665-3036

DATA DO PEDIDO	REQUISITANTE	EMBARCADO POR	PONTO DE FOB	TERMOS

QUANTIDADE	UNIDADE	DESCRÍÇÃO	PREÇO UNITÁRIO	TOTAL
10cx	12 x 1 litro	Garapa pasteurizada em embalagem longa vida apresentação 1 litro	4,00	480,00

Dados de Referência são dados relacionados com códigos, tais como estado, país, status do pedido, tipo de movimentação do colaborador etc.

Os exemplos de Dados Mestre dependem do negócio: cliente, fornecedor, produto, colaboradores, contas, locais, alunos etc.

PIERIM SUPERMERCADOS*Tradição e confiança!*

Av. Mons Demosthenes Paraná Brasil Pontes, 1712
Jd Lavinia - Mococa – SP / CEP: 13737-632
Telefone 19 3665-3036 Email: supermercadopierim2@uol.com.br

O número a seguir deve constar de toda a correspondência, documentação de embarque e faturas pertinentes:
Nº DO PEDIDO: [100]

PARA:

Natacha Calazans do Nascimento
Edina Comércio de Bebidas
Cidade Universitária Zeferino Vaz
Campinas, s/número, cep 13084-971
19 3521 XXXX

PEDIDO DE COMPRA **Dados Mestre** **Dados de Referência** **Dados de Transacionais****ENVIAR PARA: (DADOS DA CENTRAL PIERIM)**

Tiago Machado Flávio
Pierim Supermercados
Av Monsenhor Demosthenes Parana Brasil Pontes, 1712
Jd Lavinia – Mococa SP – CEP: 13737-632
19 3665-3036

DATA DO PEDIDO	REQUISITANTE	EMBARCADO POR	PONTO DE FOB	TERMOS

QUANTIDADE	UNIDADE	DESCRÍÇÃO	PREÇO UNITÁRIO	TOTAL
10cx	12 x 1 litro	Garapa pasteurizada em embalagem longa vida apresentação 1 litro	4,00	480,00

Dados de Referência são dados relacionados com códigos, tais como estado, país, status do pedido, tipo de movimentação do colaborador etc.

Os exemplos de Dados Mestre dependem do negócio: cliente, fornecedor, produto, colaboradores, contas, locais, alunos etc.

PIERIM SUPERMERCADOS*Tradição e confiança!*

Av. Mons Demosthenes Paraná Brasil Pontes, 1712
Jd Lavinia - Mococa – SP / CEP: 13737-632
Telefone 19 3665-3036 Email: supermercadopierim2@uol.com.br

O número a seguir deve constar de toda a correspondência, documentação de embarque e faturas pertinentes:
Nº DO PEDIDO: [100]

PARA:

Natacha Calazans do Nascimento
Edina Comércio de Bebidas
Cidade Universitária Zeferino Vaz
Campinas, s/número, cep 13084-971
19 3521 XXXX

PEDIDO DE COMPRA **Dados Mestre** **Dados de Referência** **Dados de Transacionais****ENVIAR PARA: (DADOS DA CENTRAL PIERIM)**

Tiago Machado Flávio
Pierim Supermercados
Av Monsenhor Demosthenes Parana Brasil Pontes, 1712
Jd Lavinia – Mococa SP – CEP: 13737-632
19 3665-3036

DATA DO PEDIDO	REQUISITANTE	EMBARCADO POR	PONTO DE FOB	TERMOS

QUANTIDADE	UNIDADE	DESCRÍÇÃO	PREÇO UNITÁRIO	TOTAL
10cx	12 x 1 litro	Garapa pasteurizada em embalagem longa vida apresentação 1 litro	4,00	480,00

Dados de Referência são dados relacionados com códigos, tais como estado, país, status do pedido, tipo de movimentação do colaborador etc.

Os exemplos de Dados Mestre dependem do negócio: cliente, fornecedor, produto, colaboradores, contas, locais, alunos etc.

PIERIM SUPERMERCADOS*Tradição e confiança!*

Av. Mons Demosthenes Paraná Brasil Pontes, 1712
Jd Lavinia - Mococa - SP / CEP: 13737-632
Telefone 19 3665-3036 Email: supermercadopierim2@uol.com.br

O número a seguir deve constar de toda a correspondência, documentação de embarque e faturas pertinentes:
Nº DO PEDIDO: [100]

PARA:

Natacha Calazans do Nascimento
Edina Comércio de Bebidas
Cidade Universitária Zeferino Vaz
Campinas, s/número, cep 13084-971
19 3521 XXXX

PEDIDO DE COMPRA **Dados Mestre** **Dados de Referência** **Dados de Transacionais****ENVIAR PARA: (DADOS DA CENTRAL PIERIM)**

Tiago Machado Flávio
Pierim Supermercados
Av Monsenhor Demosthenes Parana Brasil Pontes, 1712
Jd Lavinia – Mococa SP – CEP: 13737-632
19 3665-3036

DATA DO PEDIDO	REQUISITANTE	EMBARCADO POR	PONTO DE FOB	TERMOS

QUANTIDADE	UNIDADE	DESCRIÇÃO	PREÇO UNITÁRIO	TOTAL
10cx	12 x 1 litro	Garapa pasteurizada em embalagem longa vida apresentação 1 litro	4,00	480,00

Dados de Referência são dados relacionados com códigos, tais como estado, país, status do pedido, tipo de movimentação do colaborador etc.

Os exemplos de Dados Mestre dependem do negócio: cliente, fornecedor, produto, colaboradores, contas, locais, alunos etc.

PIERIM SUPERMERCADOS*Tradição e confiança!*

Av. Mons Demosthenes Paraná Brasil Pontes, 1712
Jd Lavinia - Mococa - SP / CEP: 13737-632
Telefone 19 3665-3036 Email: supermercadopierim2@uol.com.br

O número a seguir deve constar de toda a correspondência, documentação de embarque e faturas pertinentes:
Nº DO PEDIDO: [100]

PARA:

Natacha Calazans do Nascimento
Edina Comércio de Bebidas
Cidade Universitária Zeferino Vaz
Campinas, s/número, cep 13084-971
19 3521 XXXX

PEDIDO DE COMPRA **Dados Mestre** **Dados de Referência** **Dados de Transacionais****ENVIAR PARA: (DADOS DA CENTRAL PIERIM)**

Tiago Machado Flávio
Pierim Supermercados
Av Monsenhor Demosthenes Parana Brasil Pontes, 1712
Jd Lavinia – Mococa SP – CEP: 13737-632
19 3665-3036

DATA DO PEDIDO	REQUISITANTE	EMBARCADO POR	PONTO DE FOB	TERMOS

QUANTIDADE	UNIDADE	DESCRÍÇÃO	PREÇO UNITÁRIO	TOTAL
10cx	12 x 1 litro	Garapa pasteurizada em embalagem longa vida apresentação 1 litro	4,00	480,00

Dados de Referência são dados relacionados com códigos, tais como estado, país, status do pedido, tipo de movimentação do colaborador etc.

Os exemplos de Dados Mestre dependem do negócio: cliente, fornecedor, produto, colaboradores, contas, locais, alunos etc.

PIERIM SUPERMERCADOS*Tradição e confiança!*

Av. Mons Demosthenes Paraná Brasil Pontes, 1712
 Jd Lavinia - Mococa - SP / CEP: 13737-632
 Telefone 19 3665-3036 Email: supermercadopierim2@uol.com.br

O número a seguir deve constar de toda a correspondência, documentação de embarque e faturas pertinentes:
Nº DO PEDIDO: [100]

PARA:

Natacha Calazans do Nascimento
 Edina Comércio de Bebidas
 Cidade Universitária Zeferino Vaz
 Campinas, s/número, cep 13084-971
 19 3521 XXXX

PEDIDO DE COMPRA **Dados Mestre** **Dados de Referência** **Dados de Transacionais****ENVIAR PARA: (DADOS DA CENTRAL PIERIM)**

Tiago Machado Flávio
 Pierim Supermercados
 Av Monsenhor Demosthenes Parana Brasil Pontes, 1712
 Jd Lavinia – Mococa SP – CEP: 13737-632
 19 3665-3036

DATA DO PEDIDO	REQUISITANTE	EMBARCADO POR	PONTO DE FOB	TERMOS

QUANTIDADE	UNIDADE	DESCRÍÇÃO	PREÇO UNITÁRIO	TOTAL
10cx	12 x 1 litro	Garapa pasteurizada em embalagem longa vida apresentação 1 litro	4,00	480,00

Dados de Referência são dados relacionados com códigos, tais como estado, país, status do pedido, tipo de movimentação do colaborador etc.

Os exemplos de Dados Mestre dependem do negócio: cliente, fornecedor, produto, colaboradores, contas, locais, alunos etc.

PIERIM SUPERMERCADOS*Tradição e confiança!*

Av. Mons Demosthenes Paraná Brasil Pontes, 1712
Jd Lavinia - Mococa - SP / CEP: 13737-632
Telefone 19 3665-3036 Email: supermercadopierim2@uol.com.br

O número a seguir deve constar de toda a correspondência, documentação de embarque e faturas pertinentes:
Nº DO PEDIDO: [100]

PARA:

Natacha Calazans do Nascimento
Edina Comércio de Bebidas
Cidade Universitária Zeferino Vaz
Campinas, s/número, cep 13084-971
19 3521 XXXX

PEDIDO DE COMPRA **Dados Mestre** **Dados de Referência** **Dados de Transacionais****ENVIAR PARA: (DADOS DA CENTRAL PIERIM)**

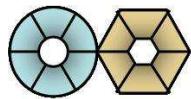
Tiago Machado Flávio
Pierim Supermercados
Av Monsenhor Demosthenes Parana Brasil Pontes, 1712
Jd Lavinia – Mococa SP – CEP: 13737-632
19 3665-3036

DATA DO PEDIDO	REQUISITANTE	EMBARCADO POR	PONTO DE FOB	TERMOS

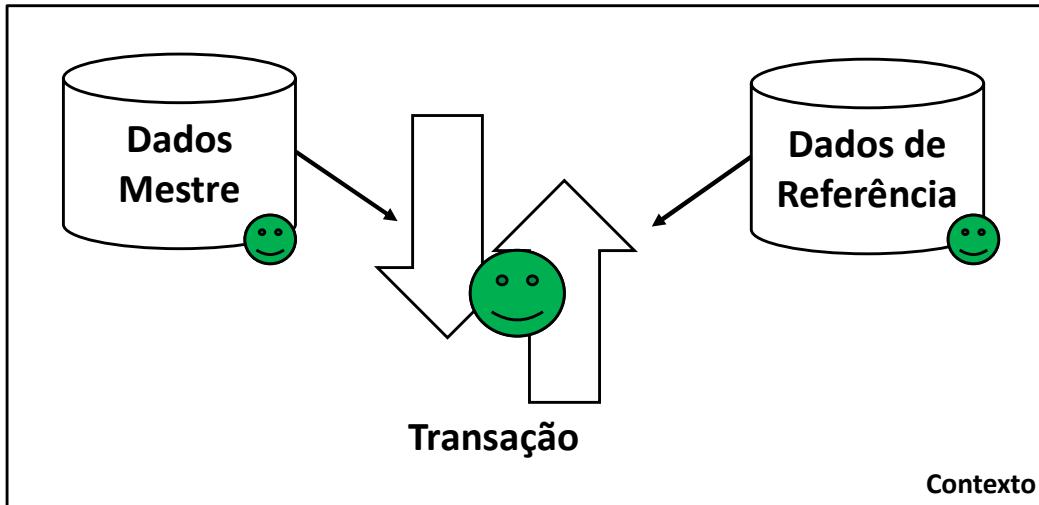
QUANTIDADE	UNIDADE	DESCRÍÇÃO	PREÇO UNITÁRIO	TOTAL
10cx	12 x 1 litro	Garapa pasteurizada em embalagem longa vida apresentação 1 litro	4,00	480,00

Dados de Referência são dados relacionados com códigos, tais como estado, país, status do pedido, tipo de movimentação do colaborador etc.

Os exemplos de Dados Mestre dependem do negócio: cliente, fornecedor, produto, colaboradores, contas, locais, alunos etc.

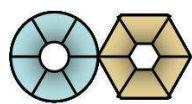


MD/RDM – Alguns conceitos



Qualidade

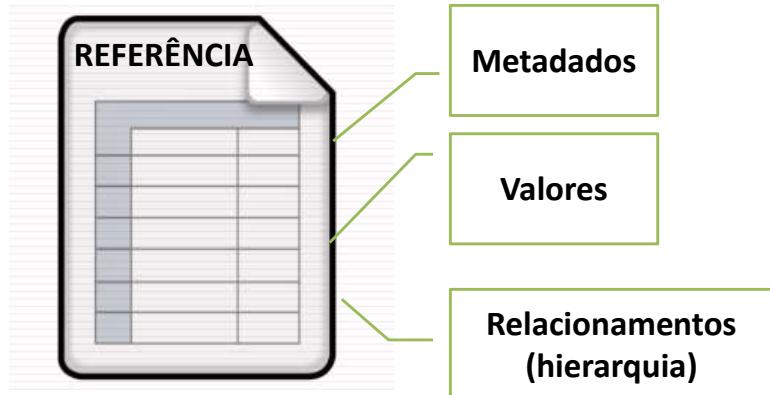
A qualidade do dado transacional é muito dependente da qualidade do dado de referência e do dado mestre. A melhoria da qualidade do dado mestre e do dado de referência melhora a qualidade de todos os dados e tem um grande impacto na confiança do negócio no seu próprio dado.



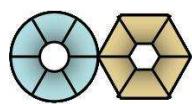
Gestão de dados de referência



Gestor de Dados
de Negócio



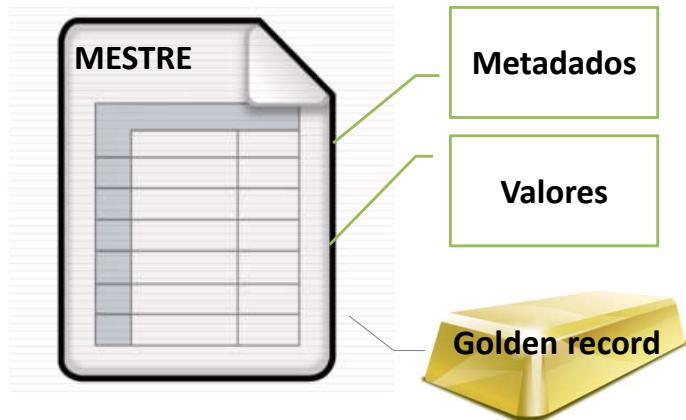
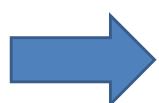
O gestor de dados de negócio mantém uma lista de dados válidos (códigos e etc) e seus significados para o negócio por meio de origem interna ou fontes externas. Ele também faz a gestão dos relacionamentos entre os valores dos dados de referência, particularmente em hierarquias.



Gestão de dados mestre

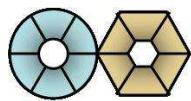


Gestor de Dados
de Negócio

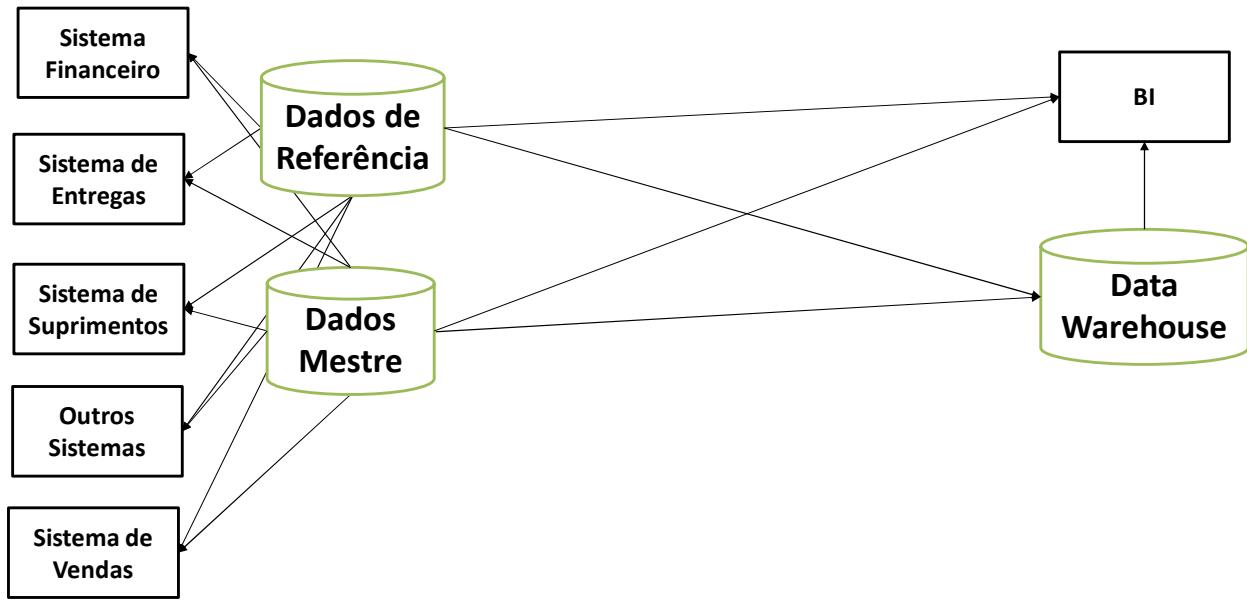


Gestão de dados mestres requer a identificação e / ou desenvolvimento do registro dourado da verdade para cada produto, região, pessoa ou organização.

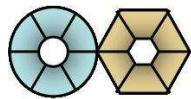
Em alguns casos, um "sistema de registros" (SOR) provê o dado definitivo de uma instância. Entretanto, até mesmo um sistema pode produzir accidentalmente mais de um registro para a mesma instância. Uma grande variedade de técnicas são usadas para determinar, da melhor forma possível, o dado mais preciso e atualizado da instância.



MD/RDM – Objetivo final

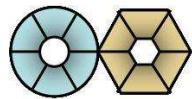


Uma vez que os valores mais precisos, correntes e relevantes são estabelecidos, dados mestres e de referência ficam disponíveis para compartilhamento entre sistemas de aplicações transacionais e data Warehouse / business intelligence. Alguns dados são replicados e propagados de uma base de dados mestres para uma ou mais bases de dados. Outras aplicações podem ler diretamente da base de dados mestres e de referência.



Dados de referência

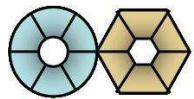
- Utilizados para classificar e categorizar outros dados.
- Domínio ditado por regras de negócio ou por órgãos governamentais.
- Variam de acordo com o contexto da aplicação.



Dados de referência



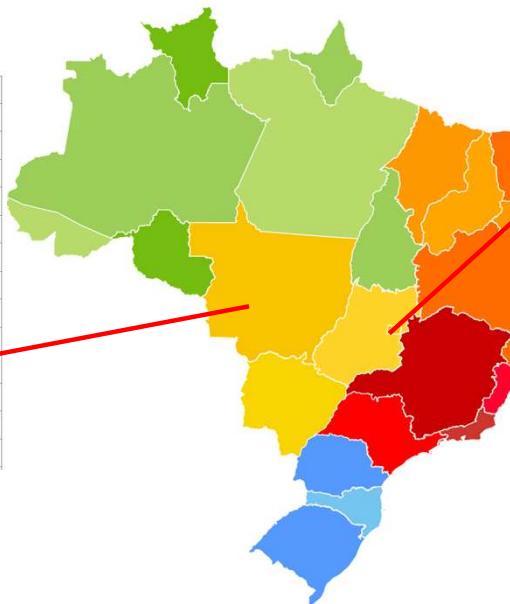
79	classTrabEstrang	trabEstrangeiro	E	N	1-1	002	-	estrangeiro. Classificação da condição do trabalhador estrangeiro no Brasil: 1 - Visto permanente; 2 - Visto temporário; 3 - Asilado; 4 - Refugiado; 5 - Solicitante de Refúgio; 6 - Residente em país fronteiriço ao Brasil; 7 - Deficiente físico e com mais de 51 anos; 8 - Com residência provisória e anistiado, em situação irregular; 9 - Permanência no Brasil em razão de filhos ou cônjuge brasileiros; 10 - Beneficiado pelo acordo entre países do Mercosul; 11 - Dependente de agente diplomático e/ou consular de países que mantém convênio de reciprocidade para o exercício de atividade remunerada no Brasil; 12 - Beneficiado pelo Tratado de Amizade, Cooperação e Consulta entre a República Federativa do Brasil e a República Portuguesa.
00	aceitador	funcEstrangeiro	E	C	1-1	001		Valores Válidos: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. <small>Correto com beneficiários.</small>



Dados de referência

Code	Subdivision name (pt)	Subdivision category
BR-DF	Distrito Federal	federal district
BR-AC	Acre	state
BR-AL	Alagoas	state
BR-AP	Amapá	state
BR-AM	Amazonas	state
BR-BA	Bahia	state
BR-CE	Ceará	state
BR-ES	Espírito Santo	state
BR-GO	Goiás	state
BR-MA	Maranhão	state
BR-MT	Mato Grosso	state
BR-MS	Mato Grosso do Sul	state
BR-MG	Minas Gerais	state

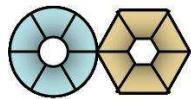
ISO 3166-2



```
01. package modelo;
02. public enum Estados {
03.     AC("Acre"),
04.     AL("Alagoas"),
05.     AM("Amazonas"),
06.     AP("Amapá"),
07.     BA("Bahia"),
08.     CE("Ceará"),
09.     DF("Distrito Federal"),
10.    ES("Espírito Santo"),
11.    GO("Goiás"),
12.    MA("Maranhão"),
13.    MG("Minas Gerais"),
14.    MS("Mato Grosso Sul"),
15.    MT("Mato Grosso"),
16.    PA("Pará"),
17.    PB("Paraíba"),
18.    PE("Pernambuco"),
19.    PI("Piauí"),
20.    PR("Paraná"),
21.    RJ("Rio de Janeiro"),
22.    RN("Rio Grande do Norte"),
23.    RO("Rondônia"),
24.    RR("Roraima"),
25.    RS("Rio Grande do Sul").
```

List box de UF em Java

Mais de um conjunto de valores de domínio de dados de referência podem se referir ao mesmo domínio conceitual. Cada valor é único dentro de seu domínio.

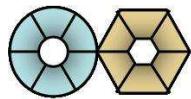


Dados de referência

English short name (upper/lower case)	Alpha-2 code	Alpha-3 code	Numeric code	ISO 3166-2 codes
Afghanistan	AF	AFG	004	ISO 3166-2:AF
Aland Islands	AX	ALA	248	ISO 3166-2:AX
Albania	AL	ALB	008	ISO 3166-2:AL
Algeria	DZ	DZA	012	ISO 3166-2:DZ
American Samoa	AS	ASM	016	ISO 3166-2:AS
Andorra	AD	AND	020	ISO 3166-2:AD
Angola	AO	AGO	024	ISO 3166-2:AO
Anguilla	AI	AIA	660	ISO 3166-2:AI
Antarctica	AQ	ATA	010	ISO 3166-2:AQ
Antigua and Barbuda	AG	ATG	028	ISO 3166-2:AG
Argentina	AR	ARG	032	ISO 3166-2:AR
Armenia	AM	ARM	051	ISO 3166-2:AM
Aruba	AW	ABW	533	ISO 3166-2:AW
Australia	AU	AUS	036	ISO 3166-2:AU
Austria	AT	AUT	040	ISO 3166-2:AT

Códigos múltiplos

Alguns conjuntos de dados de referência fazem referência cruzada entre valores de códigos múltiplos representando a mesma coisa. Diferentes aplicações de banco de dados podem utilizar conjuntos de códigos diferentes para representar o mesmo atributo conceitual



Dados de referência

English short name (upper/lower case)	Alpha-2 code	Alpha-3 code	Numeric code	ISO 3166-2 codes
Afghanistan	AF	AFG	004	ISO 3166-2:AF
Aland Islands	AX	ALA	248	ISO 3166-2:AX
Albania	AL	ALB	008	ISO 3166-2:AL
Algeria	DZ	DZA	012	ISO 3166-2:DZ
American Samoa	AS	ASM	016	ISO 3166-2:AS
Andorra	AD	AND	920	ISO 3166-2:AD
Angola	AO	AGO	024	ISO 3166-2:AO
Anguilla	AI	AIA	660	ISO 3166-2:AI
Antarctica	AQ	ATA	010	ISO 3166-2:AQ
Antigua and Barbuda	AG	ATG	028	ISO 3166-2:AG
Argentina	AR	ARG	032	ISO 3166-2:AR
Armenia	AM	ARM	051	ISO 3166-2:AM
Aruba	AW	ABW	533	ISO 3166-2:AW
Australia	AU	AUS	036	ISO 3166-2:AU
Austria	AT	AUT	040	ISO 3166-2:AT

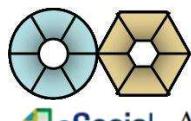
Aplicação A

Aplicação B

Aplicação C

Códigos múltiplos

Alguns conjuntos de dados de referência fazem referência cruzada entre valores de códigos múltiplos representando a mesma coisa. Diferentes aplicações de banco de dados podem utilizar conjuntos de códigos diferentes para representar o mesmo atributo conceitual



Dados de referência



Anexo III - Tabelas do eSocial - Manual de Orientação do eSocial – Versão 2.0

Tabela 3 – Tabela de Natureza das Rubricas da Folha de Pagamento

Código	Nome da Natureza da Rubrica	Descrição da Natureza da Rubrica
1000	Salário, vencimento, soldo ou subsídio.	Corresponde ao salário básico contratual do empregado contratado de acordo com a CLT e a remuneração mensal do servidor público, civil ou militar. Deve ser classificada nesse código também, a remuneração paga ao trabalhador afastado por motivo de doença ou acidente de trabalho, por período de até 15 dias.
1002	Descanso semanal remunerado - DSR	Valor correspondente a um dia de trabalho do empregado, incidente sobre as verbas de natureza variável, tais como: horas extras, adicional noturno, produção, comissão, etc.
1003	Horas extraordinárias	Valor correspondente a hora de trabalho do empregado, acrescido de percentual de no mínimo, 50%.

Alguns conjuntos de dados de referência também incluem definições de negócio para cada valor. Definições proveem informação diferenciada que um nome sozinho não provê. Definições raramente aparecem em relatórios ou em caixa de seleção, mas aparecem na função Ajuda para as aplicações, apoiando o uso apropriado dos códigos em contexto.



Dados de referência

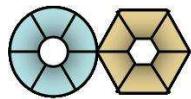
Anexo III - Tabelas do eSocial - Manual de Orientação do eSocial – Versão 2.0



Tabela 3 – Tabela de Natureza das Rubricas da Folha de Pagamento

Código	Nome da Natureza da Rubrica	Descrição da Natureza da Rubrica
1000	Salário, vencimento, soldo ou subsídio.	Corresponde ao salário básico contratual do empregado contratado de acordo com a CLT e a remuneração mensal do servidor público, civil ou militar. Deve ser classificada nesse código também, a remuneração paga ao trabalhador afastado por motivo de doença ou acidente de trabalho, por período de até 15 dias.
1002	Descanso semanal remunerado - DSR	Valor correspondente a um dia de trabalho do empregado, incidente sobre as verbas de natureza variável, tais como: horas extras, adicional noturno, produção, comissão, etc.
1003	Horas extraordinárias	Valor correspondente a hora de trabalho do empregado, acrescido de percentual de no mínimo, 50%.

Esse tipo de diferenciação é especialmente necessária para o direcionamento de classificações utilizadas em métricas de performance ou em outra análise de business intelligence.

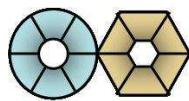


Dados de referência taxonômicos

CLASSIFICAÇÃO BRASILEIRA DE OCUPAÇÕES DOMICILIAR – CBO-DOMICILIAR				
Grande Grupo	Subgrupo principal	Subgrupo	Grupo de base	Titulação
1				MEMBROS SUPERIORES DO PODER PÚBLICO, DIRIGENTES DE ORGANIZAÇÕES DE INTERESSE PÚBLICO E DE EMPRESAS, GERENTES
	11			MEMBROS SUPERIORES E DIRIGENTES DO PODER PÚBLICO
		111		MEMBROS SUPERIORES DO PODER LEGISLATIVO, EXECUTIVO E JUDICIÁRIO
			1111	LEGISLADORES
			1112	DIRIGENTES GERAIS DA ADMINISTRAÇÃO PÚBLICA
			1113	MINISTROS DE TRIBUNAIS
		112		DIRIGENTES DE PRODUÇÃO, OPERAÇÕES E APOIO DA ADMINISTRAÇÃO PÚBLICA
			1122	DIRIGENTES DE PRODUÇÃO E OPERAÇÕES DA ADMINISTRAÇÃO PÚBLICA
			1123	DIRIGENTES DAS ÁREAS DE APOIO DA ADMINISTRAÇÃO PÚBLICA
		113		CHEFES DE PEQUENAS POPULAÇÕES
			1130	CHEFES DE PEQUENAS POPULAÇÕES
		114	-	DIRIGENTES E ADMINISTRADORES DE ORGANIZAÇÕES DE INTERESSE PÚBLICO
			1140	DIRIGENTES E ADMINISTRADORES DE ORGANIZAÇÕES DE INTERESSE PÚBLICO



Dados de referência taxonômicos podem ser importantes em vários contextos, mais显著地 para classificação de conteúdo, navegação multi-facetada, e business-intelligence. Em bases de dados relacionais tradicionais, dados de referência taxonômicos podem ser armazenados em uma relação recursiva. Ferramentas de gestão de taxonomia normalmente mantêm a informação hierárquica, além de outras coisas.

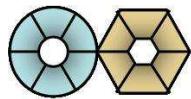


Dados de referência - Metadados



Meta-dados sobre conjunto de dados de referência devem documentar:

O significado e o propósito de cada domínio de valores de dados de referência.
As tabelas de referência e bases de dados onde o dado de referência aparece.
A origem do dado em cada tabela.
A versão corrente disponível
Quando foi a ultima atualização do dado
Como o dado é mantido em cada tabela
Quem é responsável pela qualidade do dado e do meta-dados



Dados mestre

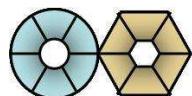


- Dados sobre as entidades de negócio vitais, que dão contexto às transações
- Domínio menos rígido, mas definido por regras de negócio, com variações semânticas
- Variam conforme o negócio

Variações semânticas: exemplo: Pessoa (física, jurídica)

Partes: indivíduos e organizações e seus papéis, tais como clientes, cidadãos, pacientes, vendedores, alunos, parceiros, concorrentes, empregados

Estruturas financeiras: contas contábeis, centros de custo, centros de lucro



Gestão de dados mestre - Desafios

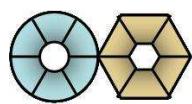
Determinar os dados dourados entre os potenciais valores conflitantes



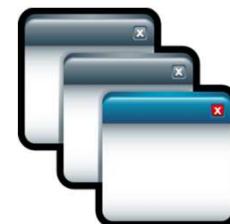
Utilizar os dados dourados ao invés dos menos precisos

Os desafios do MDM são:

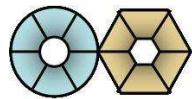
- 1) determinar os mais precisos valores dourados entre os os potenciais valores de dados conflitantes, e
- 2) utilizar valores dourados ao invés de outros dados menos precisos. O sistema de gestão de dados mestres tenta determinar os valores de dados dourados e fazer com que estes dados estejam disponíveis sempre que necessário.



Gestão de dados mestre - Implementação



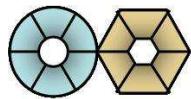
MDM pode ser implementado por meio de ferramentas de integração de dados (tais como ETL), ferramentas de higienização de dados, ODS, que servem como ponto central de dados mestres, ou por aplicações MDM especializadas



MDM – Foco primário

Identificar duplicidades nas fontes de dados e entre elas visando construir e manter IDs globais

- Identificação de registros duplicados dentro e através das fontes de dados para construir e manter IDs globais e referências cruzadas associadas para permitir a integração da informação.

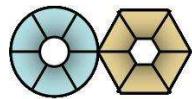


MDM – Foco primário

Identificar duplicidades nas fontes de dados e entre elas visando construir e manter IDs globais

Reconciliar fontes de dados para obter o Golden record

- Reconciliação entre as fontes de dados e fornecendo o “registro dourado” ou a melhor versão da verdade. Estes registros consolidados fornecem uma visão consolidada das informações pelos sistemas e buscam resolver inconsistências em nomes e endereços.



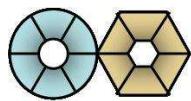
MDM – Foco primário

Identificar duplicidades nas fontes de dados e entre elas visando **construir e manter** IDs globais

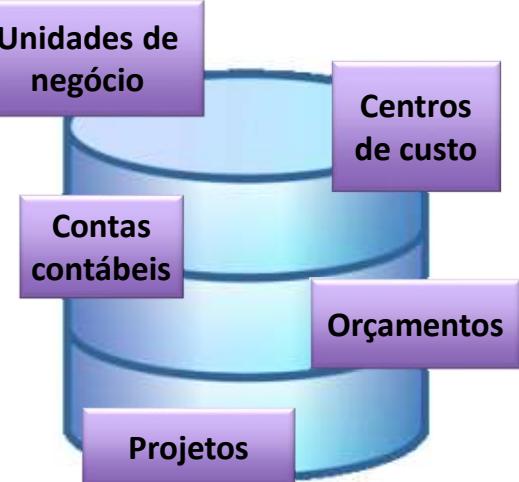
Reconciliar fontes de dados para obter o Golden record

Fornecer acesso ao Golden record às aplicações

- Fornecimento de acesso aos dados dourados entre as aplicações, seja por meio de leitura direta, ou pela replicação na alimentação de banco de dados OLTP e DW / business Intelligence.

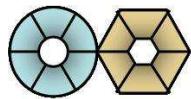


Dados mestre financeiros

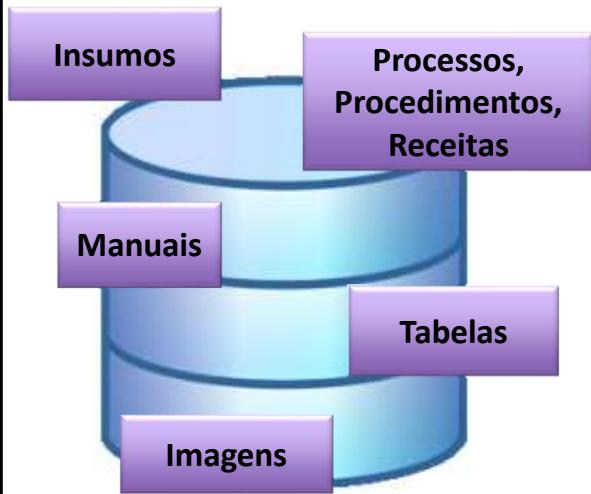


- Em geral mantidos por um software ERP
- ERP suporta back office distribuído
- Soluções de MDM financeiras gerenciam os dados e simulam cenários

ERP – Enterprise Resource Planning



Dados mestre de produtos

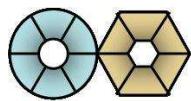


- **Produtos ou serviços, internos ou de concorrentes**
- **Estruturados ou não**
- **Sensíveis ao contexto**
- **MDM requer softwares especialistas (CAD, LMS, PLM)**

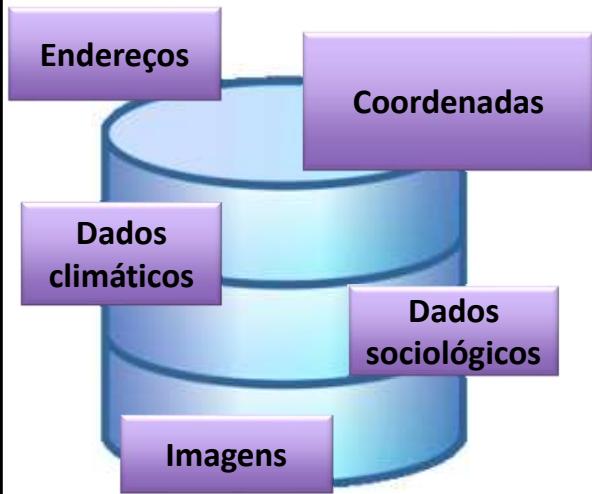
CAD – Computer Aided Design

LMS – Learning Management System

PLM – Product Lifecycle Management



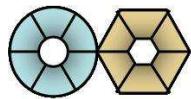
Dados mestre de localização



- **Produtos ou serviços, internos ou de concorrentes**
- **Estruturados ou não**
- **Sensíveis ao contexto**
- **MDM requer softwares especialistas**
- **Dados mestre diferem de dados de referência**

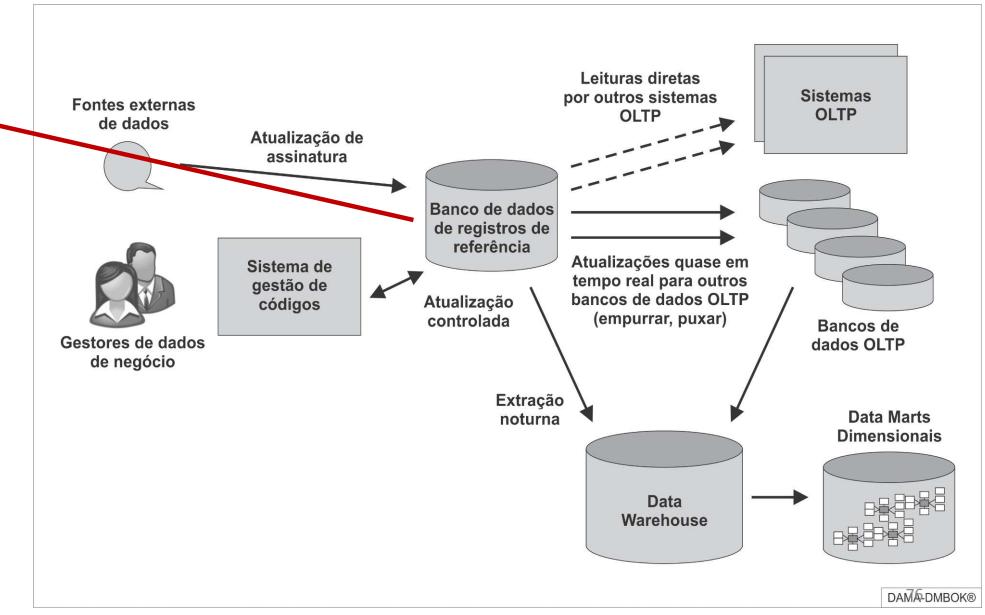
Dados de referência de localização geralmente incluem dados geopolíticos, como países, estados / províncias, região (county), cidades / vilas (towns), códigos postais, regiões geográficas, territórios de vendas, e assim por diante.

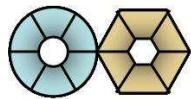
Dados mestres de localização incluem endereços comerciais de partes e localização da parte de negócio, e coordenadas de posicionamento geográfico, tais como latitude, longitude e altitude.



Exemplo de arquitetura - RD

Repositório central de dados de referência



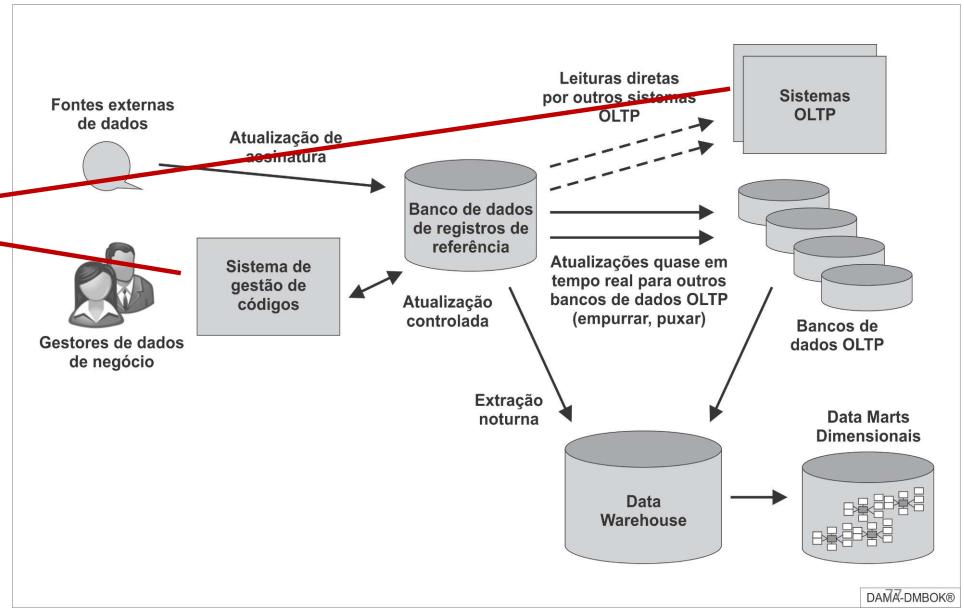


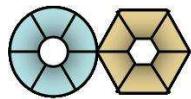
Exemplo de arquitetura - RD

Repositório central de dados de referência

Leitura direta

Integridade referencial garantida pelo código da aplicação.





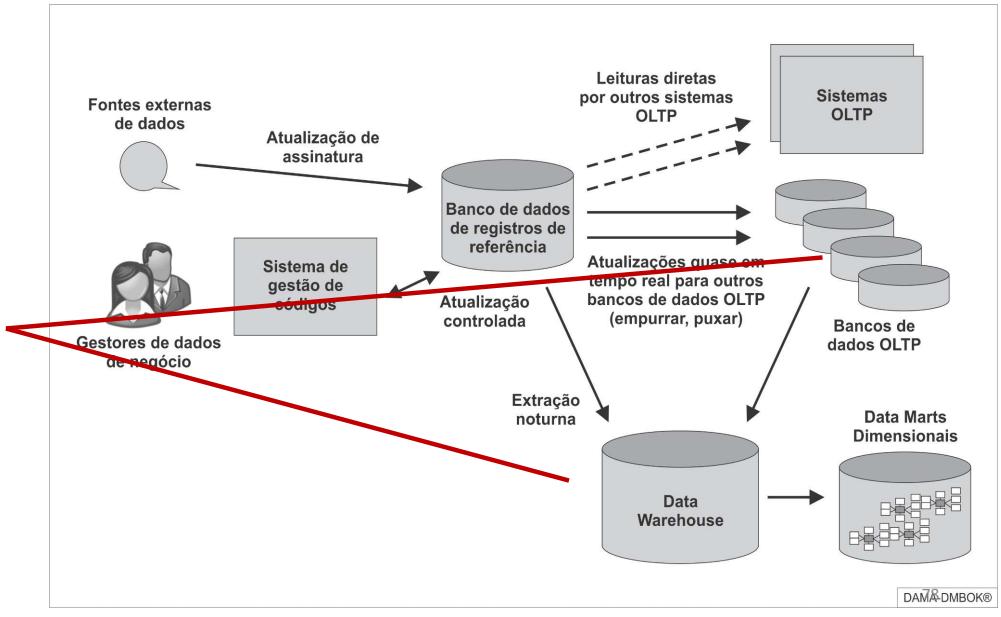
Exemplo de arquitetura - RD

Repositório central de dados de referência

Leitura direta

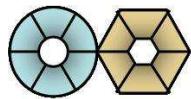
Replicação de dados

Integridade referencial garantida pelo banco de dados.



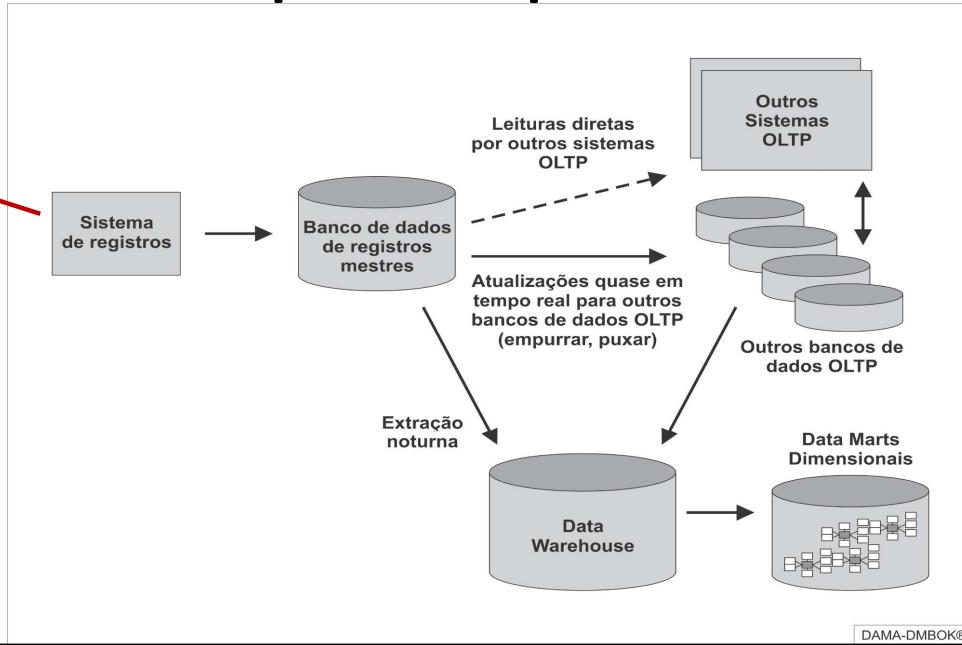
Tipos de replicação de dados:

- Replicação síncrona do repositório central para repositórios locais
- Subscrição (assíncrona)
- Atualização total



Exemplo de arquitetura - MD

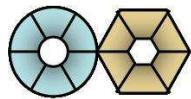
CRM
ERP
HR
Acadêmico



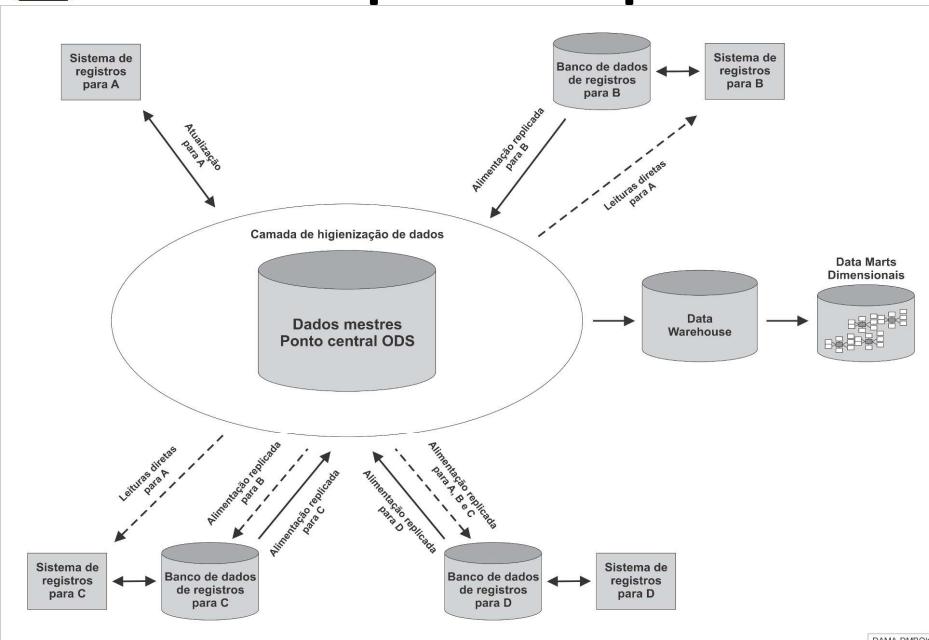
79

Áreas de interesse:

- RH – dados de funcionários
- CRM – dados de clientes
- ERP – dados financeiros e de produtos
- Sistema Acadêmico



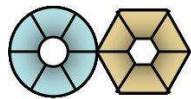
Exemplo de arquitetura - ODS



80

Uma implementação alternativa básica de projeto de ponto central e de aplicações que conversam é ter cada banco de dados de registro fornecendo o seu dado mestre ou de referência autorizável em um ODS que servirá como o ponto central do dado mestre e referência para todas as aplicações OLTP.

Um armazenamento de dados operacionais (ODS) é usado para relatórios operacionais e como fonte de dados para o armazém de dados corporativos (EDW).

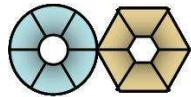


Regras de consolidação

1. **Match-merge:** correspondência das várias origens e produção de um novo registro representativo

81

Regras de correspondência (batimento) são complexas devido à necessidade de identificar tantas quanto possíveis circunstâncias, com diferentes níveis de confidencialidade e de confiança depositada em valores de dados em diferentes campos de diferentes fontes. Os desafios com as regras para correspondência e fusão são: 1) a complexidade operacional de conciliar os dados, e 2) o custo de reverter a operação se houver uma falsa fusão (fusão de registros por falha na identificação e escolha das regras).

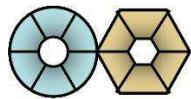


Regras de consolidação

ORIGEM_SISTEMA	ORIGEM_PK	PESSOA_NOME	PESSOA_NASCIMENTO_DATA	PESSOA_CPF	PESSOA_SEXO	PESSOA_ENDERECHO
4	123	D. Silveira	NULL	123456789	NULL	R. Quit., 53
2	456	Darci Silveira	15/12/1954	NULL	NULL	Avenida Principal, 1.283 – ap 12
6	1254	Darcy S. da Cunha	15/12/1954	123456789	F	Rua da Quitanda, 53 – casa A
3	10	Darci Silveira da Cunha	NULL	NULL	M	NULL
10	12	Darci Silveira da Cunha	15/12/1954	123456789	F	Rua da Quitanda, 53 – casa A

82

Regra de ligação-correspondência, por outro lado, é uma operação simples, já que atua sobre a tabela de referência cruzada e não sobre os campos individuais do registro de dados mestre fundido, embora isto possa ser mais difícil de apresentar informações completas a partir de vários registros.

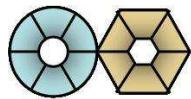


Regras de consolidação

2. **Match-link:** correspondência das várias origens através de conexões, sem produção de um novo registro

83

Regra de ligação-correspondência, por outro lado, é uma operação simples, já que atua sobre a tabela de referência cruzada e não sobre os campos individuais do registro de dados mestre fundido, embora isto possa ser mais difícil de apresentar informações completas a partir de vários registros.



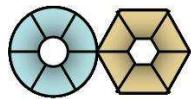
Regras de consolidação

ORIGEM_SISTEMA	ORIGEM_PK	PESSOA_NOME	PESSOA_NASCIMENTO_DATA	PESSOA_CPF	PESSOA_SEXO	PESSOA_ENDERECO
4	123	D. Silveira	NULL	123456789	NULL	R. Quit., 53
2	456	Darci Silveira	15/12/1954	NULL	NULL	Avenida Principal, 1.283 – ap 12
6	1254	Darcy S. da Cunha	15/12/1954	123456789	F	Rua da Quitanda, 53 – casa A
3	10	Darci Silveira da Cunha	NULL	NULL	M	NULL

ORIGEM_SISTEMA	ORIGEM_PK	CORRESPONDENTE_ORIGEM_SISTEMA	CORRESPONDENTE_ORIGEM_PK
4	123	2	456
4	123	6	1254
4	123	3	10

84

Regra de ligação-correspondência, por outro lado, é uma operação simples, já que atua sobre a tabela de referência cruzada e não sobre os campos individuais do registro de dados mestre fundido, embora isto possa ser mais difícil de apresentar informações completas a partir de vários registros.



Regras de consolidação

ORIGEM_SISTEMA	ORIGEM_PK	PESSOA_NOME	PESSOA_NASCIMENTO_DATA	PESSOA_CPF	PESSOA_SEXO	PESSOA_ENDERECHO
4	123	D. Silveira	NULL	123456789	NULL	R. Quit., 53
2	456	Darci Silveira	15/12/1954	NULL	NULL	Avenida Principal, 1.283 – ap 12
6	1254	Darcy S. da Cunha	15/12/1954	123456789	F	Rua da Quitanda, 53 – casa A
3	10	Darci Silveira da Cunha	NULL	NULL	M	NULL

ATRIBUTO	MANDATARIO_ORIGEM_SISTEMA	MANDATARIO_ORIGEM_PK	CONFIANCA_VALOR
PESSOA_NOME	3	10	90
PESSOA_NASCIMENTO_DATA	2	456	50
PESSOA_CPF	4	123	50
PESSOA_SEXO	6	1254	25
PESSOA_ENDERECHO	6	1254	50

Regra de ligação-correspondência, por outro lado, é uma operação simples, já que atua sobre a tabela de referência cruzada e não sobre os campos individuais do registro de dados mestre fundido, embora isto possa ser mais difícil de apresentar informações completas a partir de vários registros.

KDD - DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS



O QUE KDD É

O KDD é o **processo não trivial** de identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

Por não trivial, queremos dizer que alguma **pesquisa ou inferência** está envolvida; isto é, não é um cálculo direto de quantidades predefinidas, como calcular o valor médio de um conjunto de números.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

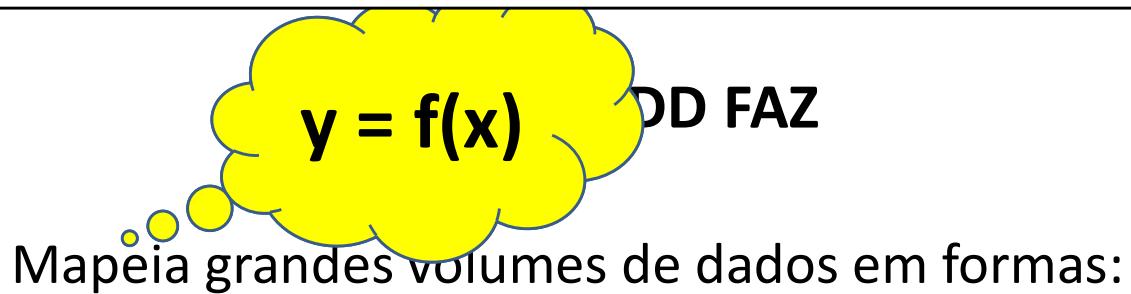
O QUE KDD É

O KDD é o processo não trivial de identificar **padrões** válidos, novos, potencialmente úteis e compreensíveis nos dados.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

Aqui, dados são um conjunto de fatos (por exemplo, casos em um banco de dados) e padrão é uma **expressão em alguma linguagem** que descreve um subconjunto dos dados ou um modelo aplicável ao subconjunto.

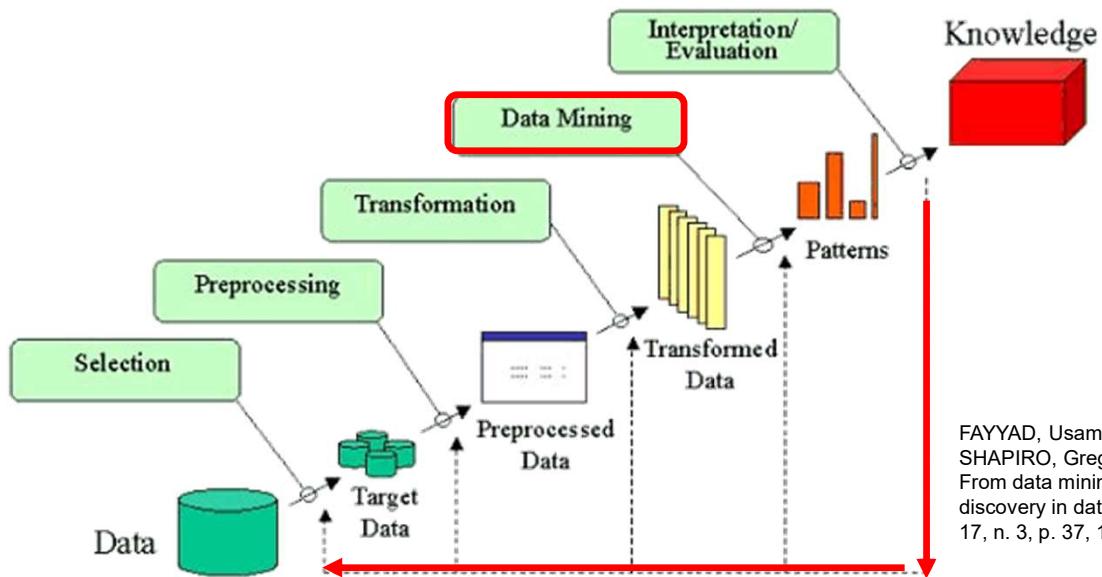
FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.



- Mais compactas (análise descritiva)
- Mais abstratas (padrão/processo de geração dos dados)
- Mais úteis (modelos preditivos)

Quando se pensa na palavra “mapear” deve-se pensar em uma função de mapeamento.

Descoberta de conhecimento em bases de dados



FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic.
From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996.

90

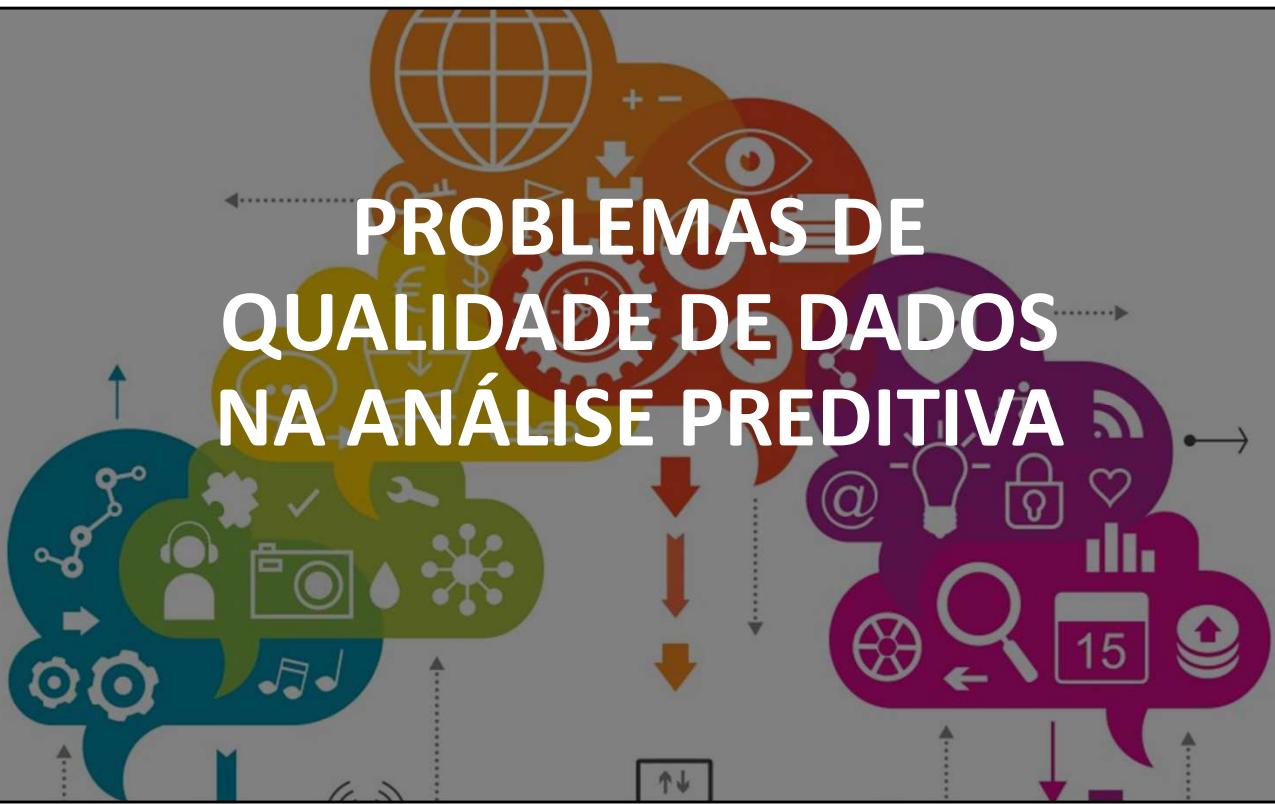
- Data Mining é **parte de um processo** de descoberta de conhecimento a partir dos dados
- Entre as tarefas de mineração está a classificação de dados
- Os passos de **preparação, seleção e limpeza dos dados** e a incorporação de conhecimento prévio são essenciais para garantir a utilidade dos resultados

MAS O DESEMPENHO DA CLASSIFICAÇÃO É CONHECIDO APENAS DEPOIS DA EXECUÇÃO

Processo custoso

Se a classificação for ruim, será necessário reiniciar o processo

PROBLEMAS DE QUALIDADE DE DADOS NA ANÁLISE PREDITIVA



MD – Algumas convenções

- **Dataset** – conjunto de dados usado no processo de mineração;
- **Base de conhecimento** – outro nome para os dados utilizados na mineração, quando destes se obtém um modelo que os explique;
- **Instância, objeto, exemplar, unidade observacional** – cada linha de um conjunto de dados, formalmente representada pelo vetor \vec{x}_i

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R.** Elsevier Brasil, 2017, p9)

Dataset e objeto

\vec{x}_i	ID	x_{ij}						y_i
		<i>i1</i>	<i>i2</i>	<i>i3</i>	<i>i4</i>	<i>i5</i>	<i>i6</i>	
\vec{x}_i	1	Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial
	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
	2	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM

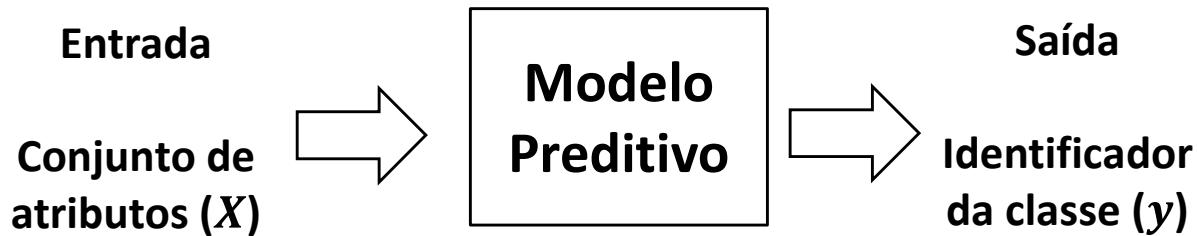
	<i>n</i>	Pedro	M	casado	30	Entregador	R\$ 700	NÃO

Dataset: $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_i, \dots, \vec{x}_n\}$

Objeto: $\vec{x}_i = \{x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{id}, \textcolor{red}{y}\}$

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R.** Elsevier Brasil, 2017, p10)

Análise Preditiva



Processo que permite descobrir o relacionamento entre exemplares de um dataset (objetos) e os rótulos a eles associados.

Processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributos de classe).

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Análise Preditiva

\vec{x}_i	ID	x_{ij}						y_i
		<i>i1</i>	<i>i2</i>	<i>i3</i>	<i>i4</i>	<i>i5</i>	<i>i6</i>	
1	Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial	Experiência
1	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
2	02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM
...
<i>n</i>	<i>n</i>	Pedro	M	casado	30	Entregador	R\$ 700	NÃO



Processo que permite descobrir o relacionamento entre exemplares de um dataset (objetos) e os rótulos a eles associados.

Processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributos de classe).

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Modelo preditivo

Relacionamento descoberto entre exemplares e rótulos, podendo ser descrito na forma de funções ou organizado em estruturas de dados.

O algoritmo adotado para a construção do modelo preditivo faz o **ajuste** dos parâmetros do modelo.

Uma vez determinado, o modelo preditivo pode ser usado para **predizer** o rótulo de exemplares desconhecidos.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Modelo preditivo **treinamento**

Relacionamento descoberto entre exemplares e rótulos, podendo ser descrito na forma de funções ou organizado em estruturas de dados.

O algoritmo adotado para a construção do modelo preditivo faz o **ajuste** dos parâmetros do modelo.

Uma vez determinado, o modelo preditivo pode ser usado para **predizer** o rótulo de exemplares desconhecidos.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Modelo preditivo

Relacionamento descoberto entre exemplares e rótulos, podendo ser descrito na forma de funções ou organizado em estruturas de dados.

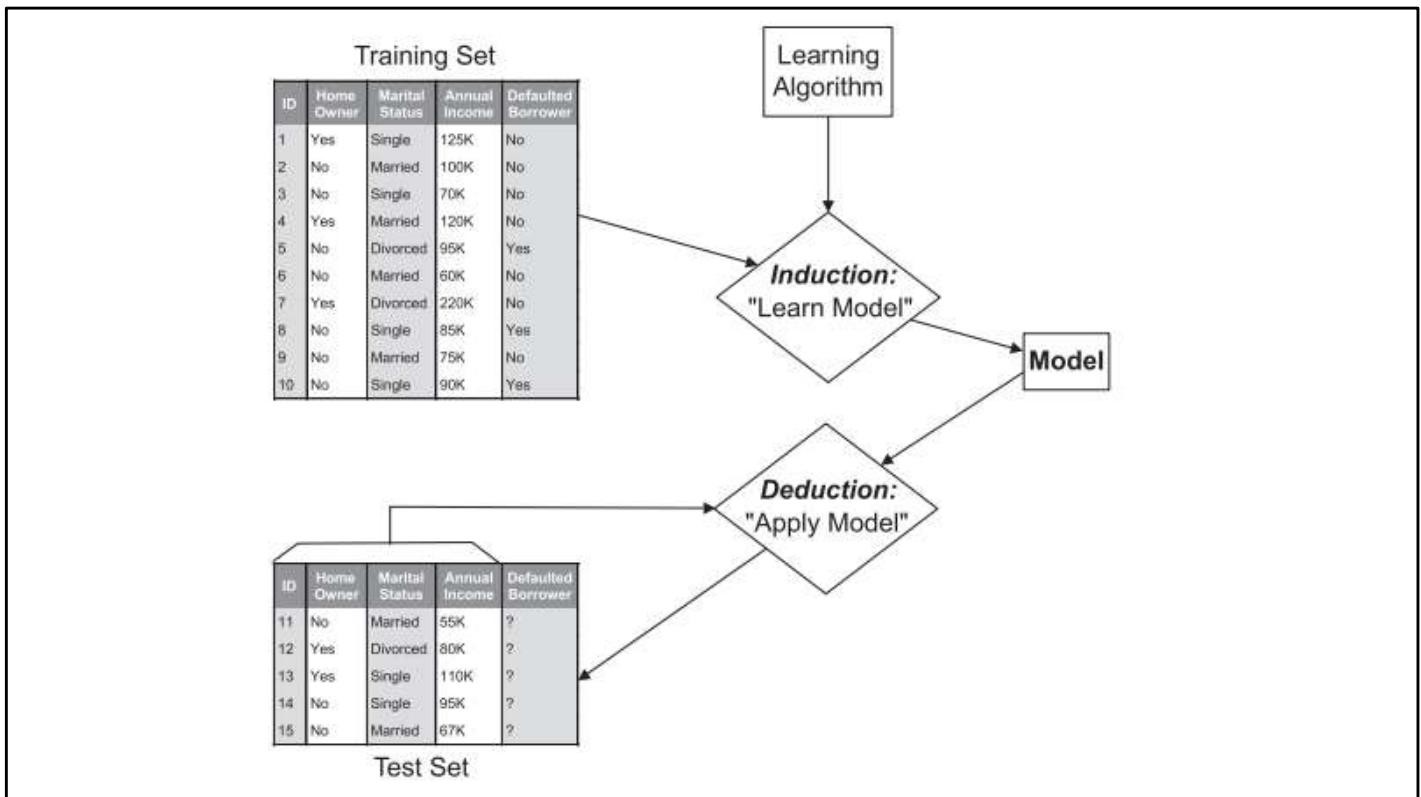
O algoritmo adotado para a construção do modelo preditivo faz o **ajuste** dos parâmetros do modelo.

teste

Uma vez determinado, o modelo preditivo pode ser usado para **predizer** o rótulo de exemplares desconhecidos.

O processo de aplicação do modelo é popularmente conhecido como teste e consiste em apresentar um novo exemplar para o modelo, que lhe fornecerá um rótulo de acordo com o mapeamento previamente descoberto.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)



TAN, Pang-Ning et al. **Introduction to data mining**. Pearson Education India, 2007.

Borrower: mutuário, pessoa que toma empréstimo

Qualidade do dataset

Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Vetorização Salarial	Experiência
01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM
...
n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO

Uma amostra (dataset) de **baixa qualidade** (atributos pouco representativos, poucos atributos, poucos exemplares, inconsistências) não será tão informativa a ponto de produzir um modelo de boa qualidade.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p78)

Qualidade do dataset

Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Retenção Salarial	Experiência
01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM
...
n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO



Qualidade



Erros de predição

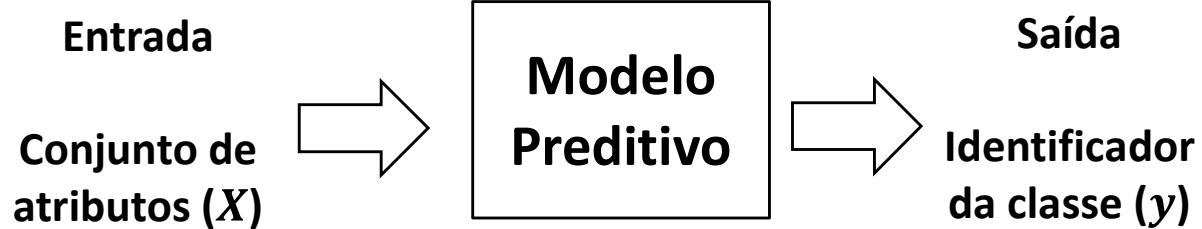
(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R.** Elsevier Brasil, 2017, p78)

O PROCESSO DE ANÁLISE DE DADOS



O processo de análise de dados

Data Mining: Análise Preditiva



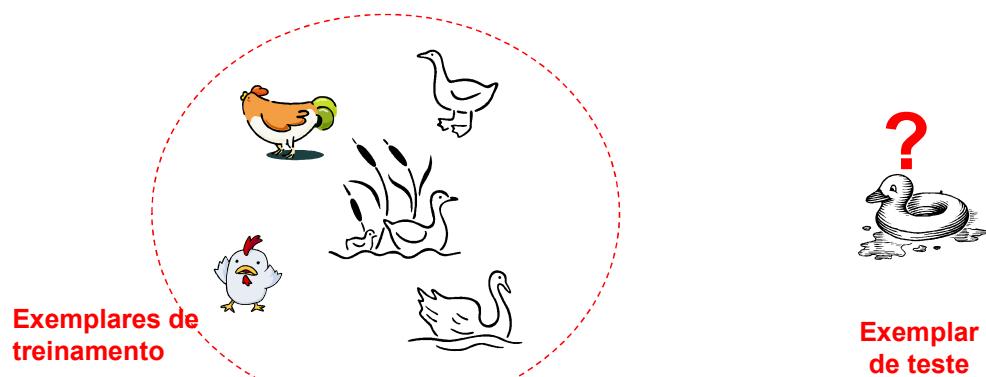
DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil, 2017, p77

103

Processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributos de classe).

O processo de análise de dados

Análise Preditiva: Exemplos



TAN, Pang-Ning et al. **Introduction to data mining**. Pearson Education India, 2007.

O processo de análise de dados

Análise Preditiva: Exemplos



Fonte da imagem: <https://www.mundoecologia.com.br>

O processo de análise de dados

Análise Preditiva: Como o *dataset* é estruturado

\vec{x}_i	ID	x_{ij}						y_i	
		<i>i1</i>	<i>i2</i>	<i>i3</i>	<i>i4</i>	<i>i5</i>	<i>i6</i>		
\vec{x}_i	1	Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial	Experiência
	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM	
	02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM	
	
	<i>n</i>	Pedro	M	casado	30	Entregador	R\$ 700	NÃO	

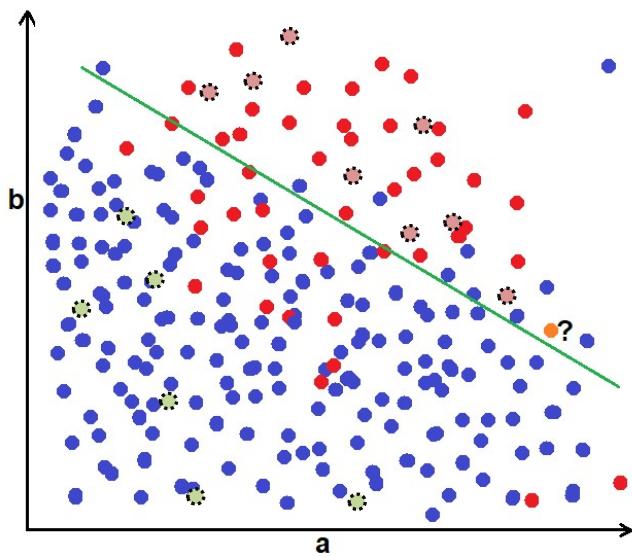
DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil, 2017, p77

106

Processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributos de classe).

O processo de análise de dados

Análise Preditiva: Classificação de dados



107

A classificação de dados é uma tarefa de mineração de dados que consiste em ajustar parâmetros de um algoritmo por um conjunto de dados de treinamento usado com a finalidade de inferir a classe de um objeto (não classificado) em análise.

O processo de análise de dados

Definindo qualidade e complexidade de dados

QUALIDADE DE DADOS – **Fidelidade** com que os dados representam pessoas, objetos, eventos ou conceitos. Quanto maior a qualidade, maior a proximidade entre a representação e o objeto ou fato representado.

COMPLEXIDADE DE DADOS – **Esforço** necessário para descrever um conjunto de dados. Quanto maior o esforço, mais complexos são os dados.

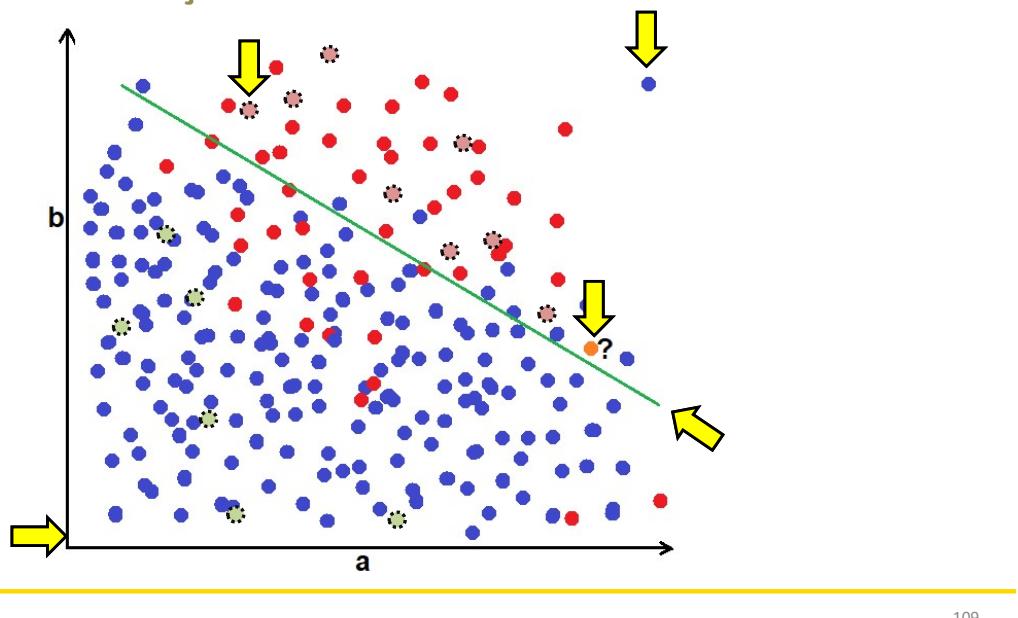
DE ÁVILA MENDES, Renê; DA SILVA, Leandro Augusto. Modeling the combined influence of complexity and quality in supervised learning. *Intelligent Data Analysis*, v. 26, n. 5, p. 1247-1274, 2022.

O processo de análise de dados

Análise Preditiva: Classificação de dados

Problemas intrínsecos nos dados:

- dados ausentes (missing values)
- dados discrepantes (outliers)
- sobreposição de atributos
- dimensionalidade do conjunto de dados



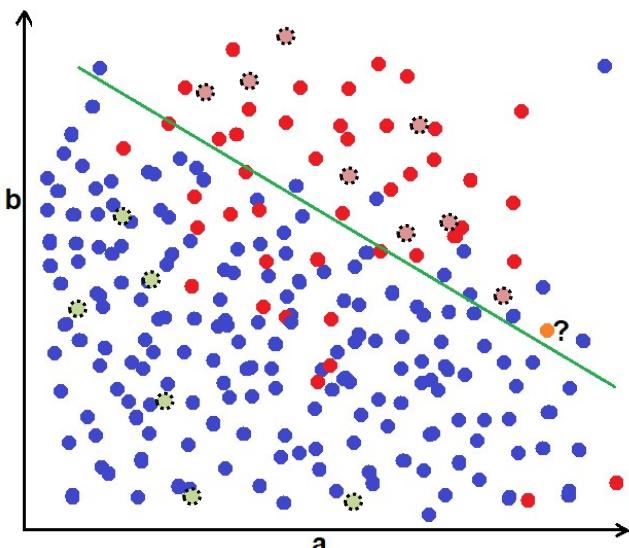
109

Problemas intrínsecos nos dados:

- dados ausentes (missing values)
- dados discrepantes (outliers)
- sobreposição de atributos
- dimensionalidade do conjunto de dados

O processo de análise de dados

O que afeta o desempenho da classificação



QUALIDADE DOS DADOS

- Valores ausentes (*missing values*)
- Valores discrepantes (*outliers*)

COMPLEXIDADE DOS DADOS

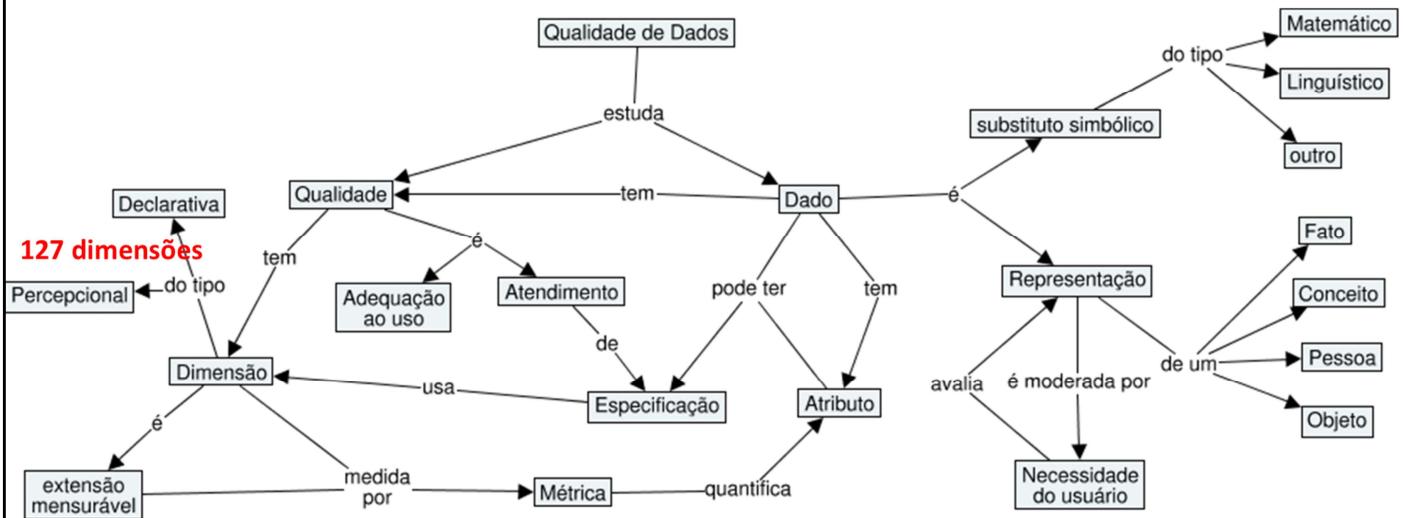
- Dimensionalidade
- Sobreposição de atributos
- Separabilidade das classes

LORENA, Ana C. et al. How Complex is your classification problem? A survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, v. 52, n. 5, p. 1-34, 2019.

110

Uma parte significativa do desempenho do algoritmo de classificação depende da complexidade e da qualidade do conjunto de dados. A Complexidade dos Dados envolve a investigação dos efeitos da **dimensionalidade**, da **sobreposição de atributos descritivos** e da **separabilidade das classes**.

Qualidade de Dados



JAYAWARDENE, Vimuthki; SADIQ, Shazia; INDULSKA, Marta. *An analysis of data quality dimensions*. 2015.

111

As literaturas acadêmica e técnica apresentam dimensões da qualidade de dados, que podem ser agrupadas em duas categorias: declarativas e de uso;

DECLARATIVAS – intrínsecas aos dados; que em si mesmas explicam os dados
 DE USO – avaliações quanto à adequação ao uso

Mapa conceitual de qualidade de dados.