

Business Intelligence



Cruzeiro do Sul Virtual
Educação a distância

Material Teórico



Processos ETL

Responsável pelo Conteúdo:

Prof.^a Esp. Lucia Contente Mós

Revisão Textual:

Prof.^a Dr.^a Selma Aparecida Cesarin



- Definição de ETL ou ETC (Extração, Transformação e Carga);
- Processo de Extração;
- Processo de Transformação ou *Data Quality*;
- Processo de *Load/Carga*;
- Processo de Gerenciamento/*Management*.



OBJETIVOS DE APRENDIZADO

- Conhecer com detalhes os processos de Extração, Transformação, Carga e Gerenciamento;
- Conhecer e identificar os diversos tipos de fontes para a realização da extração;
- Criação e utilização da *Staging Area I*;
- Conhecer e identificar as tarefas da Etapa de Transformação;
- Montagem do *Data Quality*;
- Criação e utilização da *Staging Area II*;
- Conhecer e identificar as tarefas da etapa de Carga de Dados;
- Conhecer e aplicar os conceitos da etapa de Gerenciamento.



Orientações de estudo

Para que o conteúdo desta Disciplina seja bem aproveitado e haja maior aplicabilidade na sua formação acadêmica e atuação profissional, siga algumas recomendações básicas:



Assim:

- ✓ Organize seus estudos de maneira que passem a fazer parte da sua rotina. Por exemplo, você poderá determinar um dia e horário fixos como seu “momento do estudo”;
- ✓ Procure se alimentar e se hidratar quando for estudar; lembre-se de que uma alimentação saudável pode proporcionar melhor aproveitamento do estudo;
- ✓ No material de cada Unidade, há leituras indicadas e, entre elas, artigos científicos, livros, vídeos e sites para aprofundar os conhecimentos adquiridos ao longo da Unidade. Além disso, você também encontrará sugestões de conteúdo extra no item **Material Complementar**, que ampliarão sua interpretação e auxiliarão no pleno entendimento dos temas abordados;
- ✓ Após o contato com o conteúdo proposto, participe dos debates mediados em fóruns de discussão, pois irão auxiliar a verificar o quanto você absorveu de conhecimento, além de propiciar o contato com seus colegas e tutores, o que se apresenta como rico espaço de troca de ideias e de aprendizagem.

Definição de ETL ou ETC (Extração, Transformação e Carga)

Em um sistema de BI, os Dados são transformados em informações úteis por meio de ferramentas, como, por exemplo, *Extract, Transform, Load* (ETL).

Basicamente, uma ferramenta ETL é composta de três fases: **extração, transformação e carregamento** (PRIMAK, 2008):

- **E – Extração:** captura Dados brutos das diversas fontes de Dados da organização;
- **T – Transformação:** descarta Dados irrelevantes e agrupa Dados com base em categorias de negócio por meio de chaves e índices ágeis;
- **L/C – Load ou Carregamento:** disponibiliza as informações para os sistemas de destino.

Objetivos do ETL

Extrair os Dados das diversas fontes (Bancos de Dados relacionais, arquivos texto, planilhas eletrônicas, Dados provenientes das Redes Sociais), limpar, padronizar e transformar os Dados e realizar a carga dos Dados no *Data warehouse* e/ou *Data Mart*, permitindo que os Dados estejam prontos para execução de consultas e geração de relatórios estratégicos.

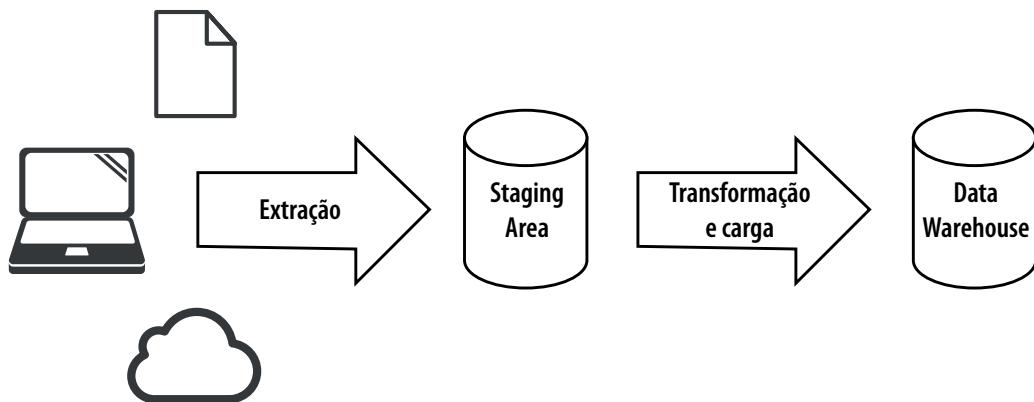


Figura 1 – Processo ETL

Trata-se do processo mais difícil no desenvolvimento de um ambiente de **BI**, pois consiste na extração dos Dados de diversas bases, com diferentes origens e tipos, na transformação, no ajuste e limpeza desses Dados, e na carga dos Dados na base de Dados do *Data Warehouse/Data Marts*.

Importância do ETL

Promove uma obtenção de visão consolidada dos Dados, por meio de análise a de relatórios que proporcionam decisões de negócios otimizadas, pois fornece o contexto histórico, de Dados, completo para os usuários da alta administração da empresa.

Também faz a união dos Dados em comum, por assunto, promove a precisão, integridade, consistência e acurácia dos Dados e permite uma etapa de auditoria fácil.

É por meio do processo de ETL que ocorre a integração entre os diversos Sistemas que uma empresa possui, pois faz um transporte de Dados planejado entre o OLTP, até chegar ao OLAP.

Por meio desse processo, é possível fazer a ligação entre as fontes de origem dos Dados e o modelo dimensional, item essencial para montagem do Projeto de BI.

Principais Componentes do ETL

Na Figura 2, é possível observar os componentes de todo o Processo ETL, que começa com as diversas fontes de Dados, que passam pela transformação e depois são inseridas no *Data Warehouse*.

Logo em seguida, os Dados estão prontos para os Sistemas de Apoio à tomada de decisão.

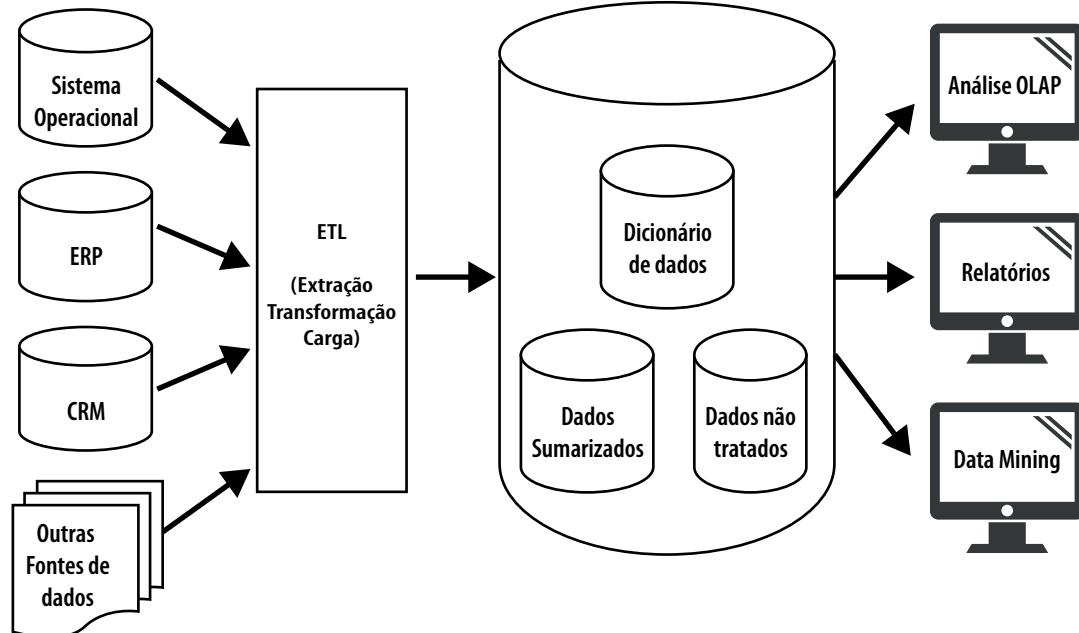


Figura 2 – Componentes ETL

Processo de Extração

Nesta etapa, os Dados são extraídos dos OLTPs (ambientes transacionais) e transportados para a *Staging Area* (área de transição ou área temporária), na qual são limpos, organizados e padronizados. É fundamental que todos os Dados estejam no mesmo formato, para que as consultas no **DW** sejam consistentes e integrais.

Na fase de extração do processo ETL, existem dois passos fundamentais, que são:

- **Definição das Fontes de Dados:** as origens dos Dados são diversas, como Bancos de Dados relacionais, planilhas eletrônicas, arquivos textos gerados pelo *Mainframe*, Dados das Redes Sociais etc. O importante, aqui, é determinar quais são as fontes de Dados que interessam para atender aos indicadores desenhados no planejamento estratégico da organização, conforme se observa na Figura 3;
- **Processo de Transporte dos Dados:** significa determinar quais serão as técnicas e as ferramentas utilizadas para mover os Dados das fontes originais e levá-los para a *Staging Area* (localização temporária dos Dados). Por exemplo, quando a fonte original se trata de um Banco de Dados, devem ser geradas *views* e Tabelas para abastecer o *Data Warehouse*. Outro item importante para se dar atenção ocorre quando a fonte original de Dados é um arquivo texto. É necessário ter muito cuidado com o formato dos Dados e das Máscaras, além das variações que um mesmo tipo de registro pode apresentar. As ferramentas ETL facilitam e automatizam esse processo, pois oferecem padronização do código, documentação, arquitetura de Integração e opções de componentes.

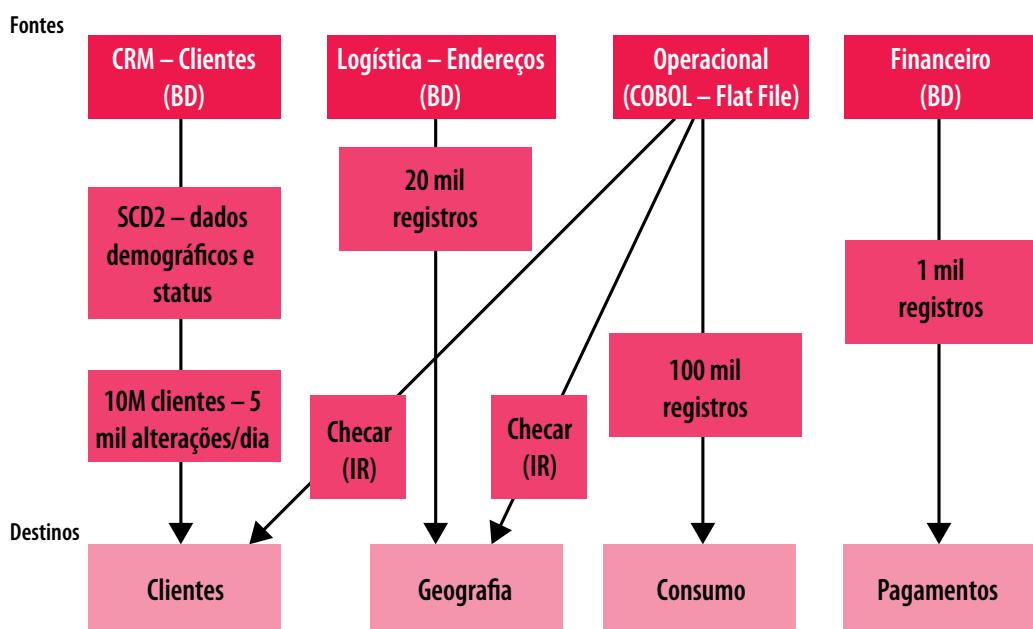


Figura 3 – Exemplos de Fontes de Dados

Fonte: Adaptado de TURBAN, 2009

Staging Area ou Data Staging

Trata-se de um local temporário, geralmente, um Banco de Dados específico para esse fim ou um *tablespace*, para armazenar os Dados que foram extraídos das Fontes de Dados.

Essa área também pode ser usada em outros procedimentos, como, por exemplo, processo de *backup* (pois contém Dados de todas as Fontes), *recover* (caso haja alguma falha de Dados no *Data Warehouse*, pode fazer a recuperação a partir dos Dados que se encontram na *Staging Area*), auditoria (é possível verificar as cargas de Dados que foram realizadas, por meio do histórico) e rastreabilidade (permite a investigação de problemas no processo ETL).

Por exercer todas essas funções, é fundamental que a *Staging Area* seja criada na execução do processo de ETL, conforme se observa na Figura 4.

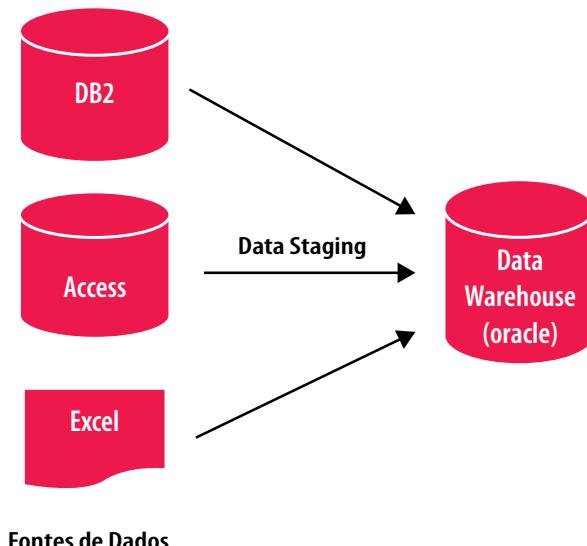


Figura 4 – Exemplo de *Data Staging*

Fonte: Adaptado de ARAÚJO, 2007

Processo de Transformação ou Data Quality

Para começar o Processo de Transformação, a *Staging Area* deve conter os Dados que foram extraídos das Fontes de Dados, que podem apresentar uma variedade imensa de erros.

São eles:

- Valores (nulos) presentes nos quais não deveriam estar (por exemplo, no campo nome do cliente, toda pessoa tem nome, não deve ter nulo);
- Uso incorreto de valores padrão (por exemplo, campos que apresentam algum tipo de *status*);

- Violação de integridade do domínio de Dados (por exemplo: o padrão seria [F para feminino e M para Masculino], mas se encontram os valores f, F, m, M, fem, masc, 1 e 2);
- Violação de integridade do Valor dos Dados (por exemplo: um Código que deve ser somente numérico, ter algum Dado que apresenta Letras);
- Violação de integridade Referencial de Dados (erros nos relacionamentos, por exemplo, o registro de pedido de um cliente não pode existir, a menos que o registro do cliente já exista);
- Armazenamento de campos calculados (por exemplo, não armazene o salário líquido e sim o salário bruto e os percentuais de desconto);
- Dados incorretos (por exemplo, o campo CEP preenchido com um valor errado, exemplo: 99999-999 ou ainda um número de cadastrado dessa forma CPF 000.000.000-00);
- Registros redundantes e/ou campos redundantes (por exemplo, ter o nome, o endereço e o telefone do cliente registrado em duas Tabelas diferentes). Vamos supor que o cliente também seja funcionário: se o Projeto de Banco de Dados for ruim, haverá o registro desses Dados na Tabela de cliente e também na Tabela de funcionários;
- Definição de máscaras e formatos para todos os Dados que necessitam (por exemplo, os campos que armazenam datas devem ter um formato definido e padronizado DD/MM/YYYY e todas as datas devem ficar armazenadas e visualizadas no formato estabelecido);
- Quando for um campo que armazena quantidades, deve ter a unidade estabelecida e padronizada (por exemplo, Kg, mt, KM, pc);
- Os tipos dos Dados devem ser respeitados conforme o especificado (por exemplo, o número da Nota Fiscal deve ser armazenado em um campo do tipo numérico e não em um campo alfanumérico);
- Garantir o armazenamento dos Dados corretos (por exemplo, todos os Dados devem ter nomes, descrições, devem ser únicos e completos).

Na Etapa de Transformação, todos esses erros devem ser corrigidos e, quando não for possível fazer a correção, os Dados que contiverem esses erros devem ser eliminados, ou seja, deve-se garantir a consistência, a integridade, respeitando as regras de negócio e qualidade dos Dados, por isso, essa etapa do processo ETL também pode ser chamada de *Data Quality*.

Somente com a qualidade de Dados, é possível realizar uma consulta no *Data warehouse* confiável e que realmente cumpra sua função de apoiar a tomada de decisão estratégica.

Para reconhecer um Dado com qualidade, ao analisar a massa de Dados, depois de passar por todo o processo de Transformação, não se deve encontrar Dados duplicados.

Outro apontador de qualidade é o conhecimento do número exato de clientes/produtos.

Mais um indicador de um Dado padronizado, na regra de negócio: suponha que na Empresa nenhum funcionário pode ganhar menos de 1 salário mínimo. Se os Dados de salário estiverem padronizados, não se deve encontrar nenhum salário nulo ou inferior ao salário mínimo estabelecido.

Deve-se fazer a preparação dos Dados históricos para realizar a transformação dos Dados e a realização de cálculos entre os Dados das colunas que sejam necessários, a geração e a utilização da *surrogate key*.

No *Data warehouse*, o objetivo é manter a integridade e criar as *Surrogate Keys* para não depender das chaves primárias que venham dos Sistemas legados.

Por exemplo: a chave do registro da Tabela de Clientes se chama *ch_cliente*. Para o registro do cliente ‘Gisele’, a chave é “gis123ch554”. É uma chave muito ruim de manipular, pois é alfanumérica.

O objetivo é que a *surrogate key* traga integridade e performance no acesso aos Dados do **DW**. Para tal, a **SK** sempre deve ser numérica e sequencial. As Sks são aplicadas nas dimensões para relacionar a Tabela fato.

Staging Area II

A criação dessa área não é uma obrigatoriedade, mas ajuda muito na organização do Processo, pois os Dados constantes nessa área encontram-se limpos, tratados e padronizados, isso significa que estão prontos para a próxima etapa que é a carga no *Data Warehouse*.

Essa área acaba servindo como um “porto seguro” dos Dados, antes de realizar a carga.

Processo de *Load/Carga*

Depois de realizada a transformação dos Dados, chegou o momento de fazer a carga, ou seja, a inserção dos Dados no *Data warehouse*. Analise o modelo dimensional e crie todas as estruturas físicas (servidores, infraestrutura) e lógicas (Banco de Dados, *tablespaces* e Tabelas) para a realização da carga dos Dados.

Para tanto, é necessário que sejam criadas as Tabelas destino e que se monte o Plano detalhado para cada carga para, então, fazer a carga inicial nas Tabelas fatos, conforme o modelo dimensional projetado.

Para depois implementar as rotinas de identificação das dimensões e da carga de Dados dessas dimensões. Depois disso, deve-se realizar a carga final de Dados na Tabela fato e testar todos os processos de carga.

Então, começa o Processo de dar continuidade ao processo de cargas incrementais das dimensões, a inserção pode ser horária, diária, semanal, quinzenal, mensal etc.

Geralmente, a carga das dimensões é diária e incremental, ou seja, a carga será só dos novos registros que surgirem nos Sistemas fontes, sendo que os mesmos procedimentos devem ser executados para as cargas das Tabelas fato.

Depois de realizada a carga no *Data warehouse*, é promovido o acesso dos Dados pelos usuários finais com o auxílio de ferramentas de visualização de Dados.

Também devem ser criados e mantidos os metadados que descrevem os Dados e a organização do Sistema. Podem ser, ainda, fórmulas utilizadas para cálculo, descrições das Tabelas disponíveis aos usuários, descrições dos campos das Tabelas, permissões de acesso e informações sobre os administradores do sistema, entre outras.

Processo de Gerenciamento/*Management*

Essa última etapa, mas não menos importante, é composta por serviços para auxiliar no Gerenciamento de todo o Processo ETL, criação e automatização de tarefas por meio do uso de jobs, planos de *backup*, verificação de itens de segurança e *compliance* (regras e políticas).

Aqui se faz o controle, a segurança, a estabilidade e a *performance* de todas as cargas e para todo o ambiente de ETL.

Também é feita a definição da periodicidade das cargas de Dados, rotinas de execução, encontrar problemas nos processos implantados e buscar otimização constante em todos os processos ETL.

- **Gerenciamento de Processos:** faz o controle das tarefas que mantêm o Sistema atualizado e consistente, gerenciando as diversas tarefas que são realizadas durante a construção e a manutenção dos componentes de um sistema de *Data warehouse*;
- **Gerenciamento de Replicação:** serve para selecionar, editar, resumir, combinar e carregar no *Data warehouse* as informações a partir das bases operacionais e das fontes externas, envolvendo programação bastante complexa, sendo que existem ferramentas poderosas que permitem que esses processos sejam gerenciados de forma mais amigável, além do controle da qualidade dos Dados que serão carregados.

Passos de Implementação de ETL (em resumo)

A seguir, observam-se os principais passos sequenciais para a criação e o estabelecimento de todo o processo ETL:

- **1º passo:** localizar todas as origens de dados;
- **2º passo:** planejar o processo de extração de dados;

- **3º passo:** projeto do Modelo Dimensional;
- **4º passo:** criar a *Starging Area I* e fazer o transporte de dados;
- **5º passo:** realizar a limpeza, padronização e transformação dos dados;
- **6º passo:** criar a *Staging Area II* – Dados limpos e prontos;
- **7º passo:** criar todas as estruturas físicas e lógicas;
- **8º passo:** realizar as Cargas Iniciais de dados de Tabelas de dimensões e da Tabela fato;
- **9º passo:** Gerenciamento das Cargas Incrementais, auditoria do processo e otimização constante.

Como se pode observar, são muitos passos de complexidade alta. Por esse motivo, cada passo deve ser executado com muito cuidado, respeitando todas as normas, as diretrizes e as políticas estabelecidas para que as irregularidades sejam minimizadas e, de preferência, que os erros sejam eliminados.

Implementação do ETL com o uso de Ferramentas

As ferramentas escolhidas deverão possibilitar a definição de aplicativos com interfaces gráficas amigáveis, geradores de relatórios, possibilidades de visualização de Dados em diversas formas e a importação dos Dados obtidos para ferramentas do usuário final, como planilhas eletrônicas e softwares editores de textos.

O conjunto de ferramentas dedicadas ao desenvolvimento de aplicações são chamadas **OLAP** (*On Line Analytical Processing*).

As vantagens de se usar as ferramentas ETL são: a programação gráfica é baseada em parâmetros, a lógica é transparente e de alto nível, a documentação é gerada de forma automática, a geração e o suporte aos dicionários de Dados é automático, há procedimentos de agendamento de tarefas, biblioteca de conexões com Banco de Dados e arquivos, há procedimentos para balanceamento de carga e paralelismo, controle de versão de mudanças.

O Mercado possui vários Cursos a respeito e profissionais para trabalhar com essas ferramentas.

As desvantagens são: o custo muito elevado, não são ferramentas simples de manusear, portanto, necessitam de grande dedicação para aprendizagem.

Principais *players* de Mercado de ferramentas ETL

- Informatica PowerCenter;
- Oracle Warehouse Builder;
- BO Data Integrator;
- Power BI da Microsoft;

- *Cognos Decision Stream*;
- *IBM Data Stage*;
- *Microsoft SQL Server DTS*;
- *SAS Enterprise ETL Server* (mais utilizada nos Projetos que utilizam o *SAS Studio*);
- *Oracle Data Integrator*, poderosa ferramenta de integração adquirida da *Sunopsis*.

Implementação do ETL Com o uso de Programação Manual

As vantagens dessa modalidade são que a implementação inicial é mais barata e mais rápida, pois os ETLs mais simples podem ser codificados facilmente.

No entanto, existem muitas desvantagens, pois os *scripts* e os programas devem ser documentados e mantidos, além dos Dicionários de Dados que devem ser atualizados.

Não há suporte para agendamento de tarefas, também não há balanceamento de carga e nem controle de mudança de versão.

Todas as conexões com Banco de Dados e arquivos também devem ser programadas e montadas manualmente.

Você deve analisar as características e as condições do seu negócio, para decidir se fará a implantação com o uso de ferramentas ETL ou de forma manual.

Geralmente, a escolha mais acertada envolve um Projeto piloto de forma manual e depois de realizar todos os acertos se faz a implantação definitiva com o auxílio das ferramentas ETL.

Seja qual for a sua escolha, é importante observar que o processo ETL é bem complexo e fundamental para o sucesso da implantação do ambiente de *Business Intelligence*.

Material Complementar

Indicações para saber mais sobre os assuntos abordados nesta Unidade:



Livros

Banco de Dados: Projeto e Implementação

MACHADO, F. N. R. Banco de Dados: projeto e implementação. São Paulo: Érica, 2004. 398 p.

Projeto de Banco de Dados: uma Visão Prática

MACHADO, F. N. R.; ABREU, M. P. de. Projeto de Banco de Dados: uma visão prática. 15.ed. São Paulo: Érica, 2007. 300 p.

Oca Oracle Database 11G – Administração I

WATSON, J. Oca Oracle Database 11G – Administração I. São Paulo: Bookman Companhia, 2009.

OCP Oracle Database 11G – Administração II

BRYLA, B. OCP Oracle Database 11G – Administração II. São Paulo: Bookman Companhia, 2009.

OCA Oracle Database 11G – Fundamentos I AO SQL

RAMKLASS, R.; WATSON, J. OCA Oracle Database 11G – Fundamentos I AO SQL. São Paulo: Bookman Companhia, 2008.

Projetando e Administrando Banco de Dados SQL Server 2000 .net: como Servidor Enterprise

PATTON, R.; OGLE, J. Projetando e administrando Banco de Dados SQL Server 2000 .net: como servidor enterprise. Tradução de Andréa Barbosa Bento; Cláudia Reali; Lineu Carneiro de Castro. Rio de Janeiro: Alta Books, 2002. 792 p.

Perspectivas em Ciência da Informação

JAMIL, G. L. Aspectos do ambiente gerencial e seus impactos no uso dos sistemas de inteligência competitiva para processos decisórios. Perspectivas em Ciência da Informação, Belo Horizonte, v. 6, n. 2, p. 261-274, jul./dez. 2001.

EMC: Armazenamento e Gerenciamento de Informações

EMC: Armazenamento e Gerenciamento de Informações. São Paulo: EMC2, 2011.

Arquitetura da Informação

CAMARGO, L. S. de A.; VIDOTTI, S. A. B. G. Arquitetura Da Informação. Rio de Janeiro: LTC,2011.

Administração de Sistemas de Informação

O'BRIEN, J. A.; MARAKAS, G. M.; Administração de Sistemas de Informação. Porto Alegre: Mc Graw Hill, 2013.

 Leitura**Data Warehouse – Modelagem Dimensional**

PITON, R.; *Data Warehouse – Modelagem Dimensional*.

<http://bit.ly/34sGm6>

Qualidade na Modelagem dos Dados de um Data Warehouse

ARAÚJO, E. M. T.; BATISTA, M. de L. S. Qualidade na Modelagem dos Dados de um *Data Warehouse*.

<http://bit.ly/35Xs0Az>

Referências

- BARBIERI, C. **BI: business intelligence**: modelagem ‘&’ tecnologia. Rio de Janeiro: Axcel Books do Brasil, 2001. 424 p.
- BECKER, J. L. **Estatística básica**: transformando Dados em informação. Porto Alegre: Bookman, 2015.
- BUSINESS Intelligence**: um enfoque gerencial. Porto Alegre: Bookman, 2009.
- CASTRO, L. N. de. **Introdução à mineração de Dados**: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.
- COUGO, P. **Modelagem conceitual e Projeto de Bancos de Dados**. Rio de Janeiro: Campus, 1997.
- DATE, C. J. **Introdução a sistemas de bancos de Dados**. Tradução de Daniel Vieira. 8.ed. Rio de Janeiro: Elsevier, 2003. 865 p.
- ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados**. Tradução de Marília Guimarães Pinheiro *et al.* 4. ed. São Paulo: Pearson Addison Wesley, 2005. 724 p.
- ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados**. 6.ed. São Paulo: Pearson, 2011.
- GILLENSON, M. L. **Fundamentos de sistemas de gerência de banco de Dados**. Tradução de Acauan Fernandes; Elvira Maria Antunes Uchoa. Rio de Janeiro: LTC, 2006. 304 p.
- INMON, W. H. **Como construir o Data Warehouse**. Rio de Janeiro: Campus, 1997.
- KIMBALL, R. **Data Warehouse Toolkit**. Rio de Janeiro: Campus, 1998.
- KWECKO, V. *et al.* Ciência de Dados aplicada na análise de processos cognitivos em grupos sociais: um estudo de caso. In: **Brazilian Symposium on Computers in Education** (Simpósio Brasileiro de Informática na Educação – SBIE). 2018. p.1543.
- LEBLANC, P. **Microsoft SQL Server 2012**. Porto Alegre: Bookman, 2014.
- PRIMAK, F. V. **Decisões com BI (Business Intelligence)**. São Paulo: Ciência Moderna, 2008.
- REZENDE, D. A. **Inteligência organizacional como Modelo de Gestão em organizações privadas e públicas**: guia para projetos de *Organizational Business Intelligence* – OBI. São Paulo: Atlas, 2015.
- ROSINI, A. M.; PALMISANO, A. **Administração de sistemas de informação e a gestão do conhecimento**. São Paulo: Thomson, 2003. 219 p.

SILBERSCHATZ, A., KORTH, H. F.; SUDARSHAN, S. **Sistema de Banco de Dados.** Tradução de Daniel Vieira. 3.ed. São Paulo: Pearson Makron Books, 2007. 778 p.

TURBAN, E. **Business intelligence:** um enfoque gerencial para a inteligencia do negócio. Porto Alegre: Bookman, 2009. 256 p.



Cruzeiro do Sul
Educacional