REVINT
REVISTA INTERDISCIPLINAR DE ENSINO, PESQUISA E EXTENSÃO

COMPARAÇÃO DA EFICIÊNCIA DE BANCOS DE DADOS RELACIONAIS COMO *DATAWAREHOUSE* EM UM CONTEXTO DE *BUSINESS INTELLIGENCE*

NASCIMENTO, Jonathan da Silva[1]; PAULUS, Gustavo Bathu[1]; RUBERT, Diogo L. V. G.[2]; ANTONIAZZI, Rodrigo Luiz[2];

Resumo: Este estudo está voltado a comparação entre banco de dados relacionais, que devido à complexidade e custos de implementação de um projeto de BI em uma instituição, a escolha do banco torna-se parte importante do projeto. Os bancos utilizados foram o MySQL e Postgres, sendo que ambos são muito utiliz[1]ados pela sua facilidade de administração. Os testes foram baseados em um contexto simples de BI no qual foram extraídos dados de fontes diferentes visando comparar qual banco de dados é mais eficiente como data warehouse. Para realização deste processo de ETL no qual deu-se ênfase na eficiência dos bancos no contexto aplicado, foi utilizado a ferramenta Pentaho Data Integration da suíte Pentaho Community para realização do mesmo.

Palavras-Chave: Business Intelligence. Banco de Dados. Data Warehouse. Comparação de Eficiência.

Abstract: This study is aimed at comparing relational database, which due to the complexity and costs of implementing a BI project in an institution, the choice of the bank becomes an important part of the project. Banks used were MySQL and Postgres, both of which are widely used for their ease of administration. The tests were based on a simple context of BI in which different sources of data were extracted in order to compare which database is more efficient as data warehouse. To carry out this ETL process in which was given emphasis on efficiency of banks in applied context, we used the Pentaho Data Integration tool suite Pentaho Community to completion.

Keywords: Business Intelligence. Database. Data Warehouse. Efficiency Comparison.

INTRODUÇÃO

Devido à alta competitividade do mercado, instituições tem se adaptado à novas tecnologias. Essas tecnologias ajudam o processo de tomada de decisões, na qual relacionando informações coletadas internamente ou externamente à empresa, levam a compreensão de tendências do mercado e a identificação de particularidades

[1] Acadêmicos do Curso de Ciência da Computação, UNICRUZ, jonathanjsn@gmail.com, gustavo.bathu.paulus@hotmail.com
[2] Professores do Curso de Ciência da Computação, UNICRUZ, drubert@unicruz.edu.br, rantoniazzi@unicruz.edu.br

relacionadas ao seu negócio. Devido a essas exigências surge o Business Intelligence com o objetivo de auxiliar à tomada de decisão.

Devido à complexidade de um planejamento estratégico adequado, algumas empresas desistem já na fase de implementação, no qual o fracasso deve-se a vários fatores, mas os mais críticos são: falha na comunicação da estratégia, não conscientização dos funcionários, falta de uma mudança cultural na empresa, falta de um patrocinador, ou seja, falta de dedicação do presidente com o projeto.

Observando estes fatores, com o objetivo de minimizar gastos e tempo de implementação, foi realizado testes comparando dois bancos de dados muito utilizados, o MySQL e o postgres, no qual foram postos a uma comparação diante de um processo chamado de ETL (Extração, Transformação e Carga). Este processo gera muito processamento na máquina onde está sendo realizada, sendo assim um bom método para analisar os bancos utilizados.

CONCEITOS

Business Intelligence (BI)

De acordo com (PETRINI et al. 2006), Business Intelligence é um conjunto de tecnologias que tem como objetivo prover e oferecer suporte a um ambiente de informação. A necessidade de eficiência e agilidade no processo decisório nas instituições exige delas a utilização de soluções que geram informações consistentes e ao mesmo tempo sejam flexíveis de modo a se enquadrar nas suas necessidades e limitações. Dessa forma, é necessário efetuar uma análise dessas instituições e das ferramentas do mercado de modo a verificar quais delas são compatíveis.

Este conceito faz enxergar perspectivas novas acerca dos dados coletados, unindo-os e cruzando-os resultando em informações que antes talvez passassem despercebidas, ou seja, visto que as realocações de dados geram novas percepções, as informações podem ser extraídas em sua totalidade.

O Business Intelligence é um coadjuvante na tomada de decisões. É basicamente um auxiliar em um ambiente empresarial onde os membros da equipe de tomada de decisões possam basear suas escolhas acerca dos resultados obtidos nas pesquisas. Visto que todas as empresas, independente do ramo, possuem objetivo de sempre melhorar e

evoluir, principalmente em relação aos seus lucros, é necessário que seus métodos para conseguir alcançar suas metas, também evoluam.

Pentaho

O sistema de pesquisa e coleta de dados é a principal forma de uma empresa ter conhecimento sobre demandas, opiniões públicas, o que falta, o que é preciso mudar, entre outros. Atualmente, a captação desse tipo de dados não é difícil, tendo em vista a rapidez em disseminação de informação existente. A grande questão está na tradução dos dados encontrados e o conceito do BI nos apresenta a solução para esse problema.

Segundo Batista (2004), constata que, "As ferramentas de *BI* podem fornecer uma visão sistêmica do negócio e ajudar na distribuição uniforme dos dados entre os usuários, sendo seu objetivo principal transformar grandes quantidades de dados em informações de qualidade para a tomada de decisões. Através delas, é possível cruzar dados, visualizar informações em várias dimensões e analisar os principais indicadores de desempenho empresarial".

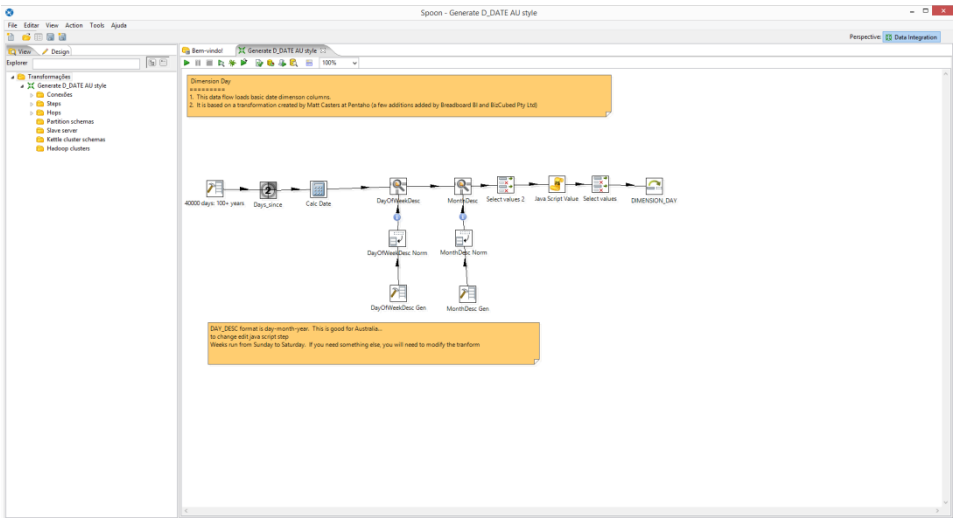
O Pentaho possui várias suites integradas que formam uma plataforma completa de BI, e sua distribuição é Open Source. Sua flexibilidade e robustez a torna uma ferramenta indispensável para uma solução eficaz além de sua simplicidade de uso. Todos os softwares da suite usam tecnologia JAVA e executam soluções como serviço, o que possibilita a conectar-se a qualquer banco de dados que possuam *drives Java Database Connectivity (JDBC)* além de prover acesso externo a ela, seja ele via Web Services ou por mecanismos baseados em SOAP/WSDL/UDDI.

Pentaho Data Integration (PDI)

Quando o projeto do data warehouse é estabilizado, um processo deve ser projetado para preencher o data warehouse com dados. Nós usamos o termo geral integração de dados para descrever o conjunto de atividades que resultem ou contribuir para o preenchimento de o data warehouse. Pentaho oferece uma coleção de ferramentas conhecidas coletivamente como Pentaho Data Integration que são projetados para suportar essa tarefa.

O PDI é um dos componentes mais poderosos responsável pelo processo de ETL, mas também possui uma infinidade de utilizações.

Figura 1. Pentaho Data Integration



Extract, Transform and Load

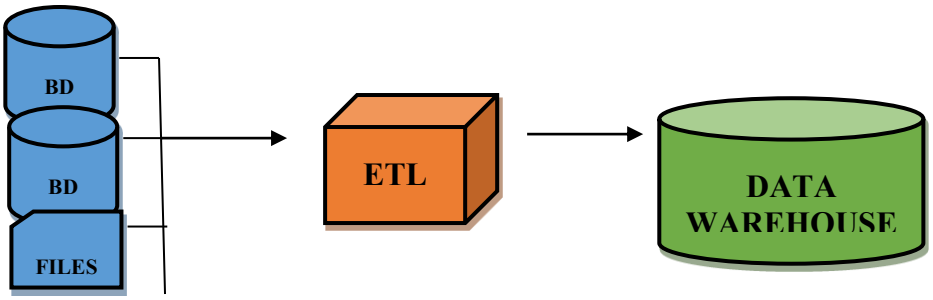
O processo de ETL (Extract, Transform and Load) é o processo mais crítico e demorado na construção de um Data Warehouse, pois consiste na extração dos dados de bases heterogêneas, na transformação e limpeza destes dados, e na carga dos dados na base do Data Warehouse.

As decisões gerenciais são tomadas com base nas informações geradas pelas ferramentas do tipo front-end. Estas informações são geradas através dos dados armazenados no Data Warehouse. Se estes dados não forem corretamente trabalhados no processo de extração, as informações geradas através deles farão com que decisões sejam tomadas erroneamente, podendo afetar diretamente os negócios da organização. Portanto, os dados devem representar a verdade, a mais pura verdade, nada mais que a verdade (KIMBALL, 1998 apud ABREU, 2007). A maior parte do esforço exigido no desenvolvimento de um DW é consumido neste momento e não é incomum que 80% de todo esforço seja empregado no processo de ETL, (INMON, 1997 apud ABREU, 2007).

Somente a extração dos dados leva mais ou menos 60 por cento das horas de desenvolvimento de um DW (KIMBALL, 1998 apud ABREU, 2007). Esta etapa do processo deve se basear na busca das informações mais importantes em sistemas fontes ou externos e que estejam em conformidade com a modelagem do DW. Tal busca de dados pode ser obstruída por problemas como a distribuição das origens dos dados, que podem estar em bases distintas com plataformas diferentes gerando a demanda de

utilização de formas de extração diferentes para cada local (ALMEIDA, 2006 apud ABREU, 2007).

Figura 2. Processo de ETL



PROCEDIMENTOS METODOLÓGICOS

Ambiente de Testes

O computador utilizado para realização dos testes foi adaptado para tal, utilizando uma instalação limpa do Windows 8.1 e minimizando os processos padrões do sistema operacional que não estavam sendo utilizados. A configuração do computador utilizado foi a seguinte:

Tabela 1. Configurações do Computador

Nome	Descrição
Sistema Operacional	Windows 8.1 64 Bits
Processador	AMD FX(tm) – 8320 3.50 Ghz
Memória RAM	16 GB

Definição dos Testes

Os testes foram analisados e comparados em situações adversas da fase de ETL, utilizando fontes diferentes e realizando o carregamento no data warehouse conforme mostrado na figura 2.

Cada inserção foi realizada 50 vezes, buscando encontrar a média de tempo de execução de cada consulta, coletando informações através do Monitor de Desempenho do Windows e da ferramenta Metrics do próprio PDI. As particularidades comparadas foram:

Monitor de Desempenho do Windows:

- % de utilização do processador;
- % de bytes confirmados em uso (RAM).

Metrics:

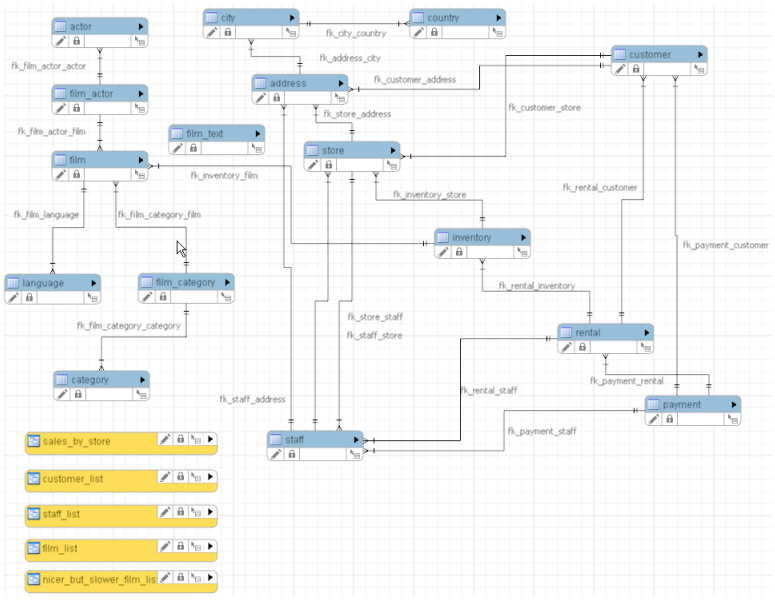
- Tempo de execução do componente Table Input e Dimension lookup/update;
- Tempo total do SQL (create statement, execute statement e open query);
- Tempo gasto para executar a transformação por completa.

Base de Dados

Foi utilizada uma base de dados de amostra, cujo nome do projeto é Sakila, no qual foi desenvolvido por Mike Hillyer, um ex- membro da equipe de documentação MySQL AB, e se destina a fornecer um esquema padrão que pode ser usado para exemplos em livros, tutoriais, artigos, amostras, e assim por diante.

Seu desenvolvimento começou no início de 2005. Os primeiros desenhos foram baseados no banco de dados utilizado no whitepaper da Dell, e foi projetado para representar uma loja de aluguel de DVD.

Figura 3. Estrutura da Base de Dados



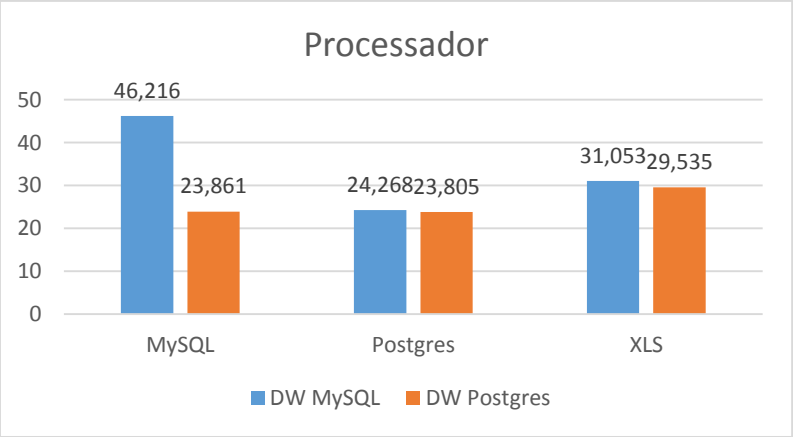
RESULTADOS E DISCUSSÕES

Os resultados dos testes foram demonstrados em gráficos para melhor visualização, cujo resultados foram:

Porcentagem de Utilização do Processador

Podemos perceber pelo gráfico que o BD Postgres utilizou menos processamento da máquina, um bom resultado tendo em vista que as vezes é necessário carregar dados muito grandes.

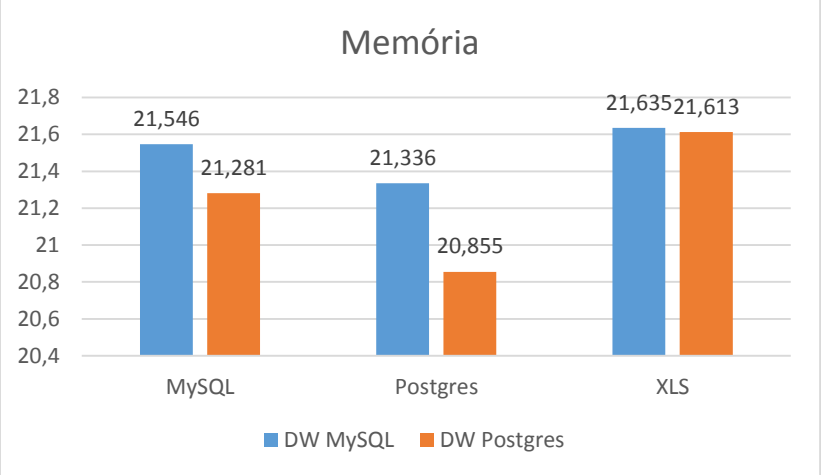
Gráfico 1. Utilização do Processador



Porcentagem de Utilização da Memória

Nesta comparação pouco se notou a diferença entre os BD, sendo que a diferença foi entre 1% a mais de utilização.

Gráfico 2. Utilização de Memória



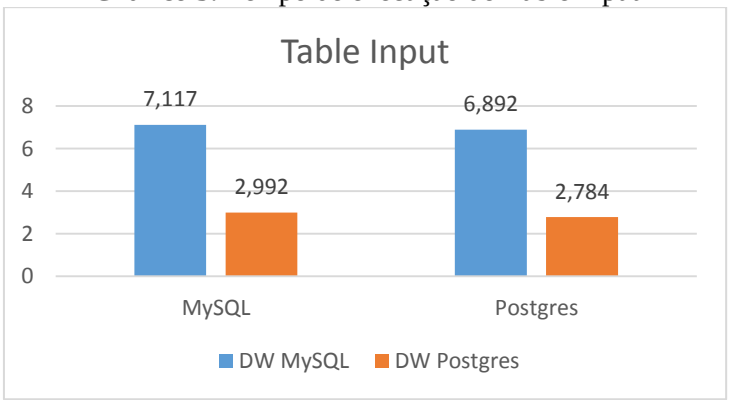
Tempo de Execução dos Componentes

Nesta etapa levou-se em consideração o tempo de execução dos componentes necessários para a transformação dos dados e o carregamento destes no data warehouse. O tempo de execução foi analisado em milissegundos e os componentes utilizados foram:

Table Input

Este componente é utilizado para a extração dos dados da base de origem, onde foi possível observar que o MySQL demorou mais que o dobro do tempo que o Postgres para selecionar os dados.

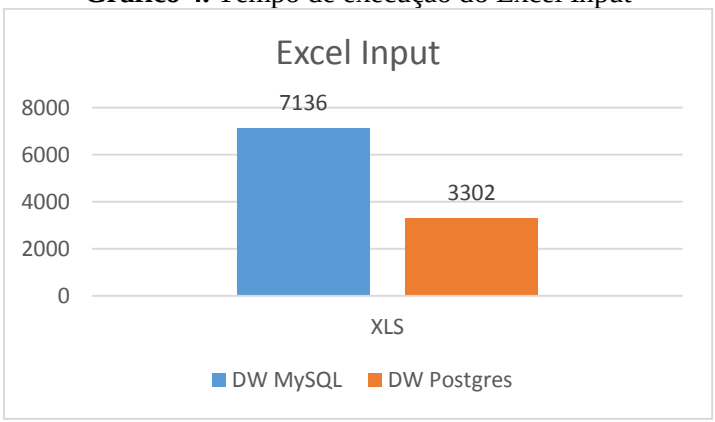
Gráfico 3. Tempo de execução do Table Input



Excel Input

Componente que carrega um arquivo com extensão xls. Nesta etapa levou-se em consideração não apenas o tempo que foi executado o componente, mas da preparação do mesmo para inserção no data warehouse, sendo só assim possível comparar este componente com os BD.

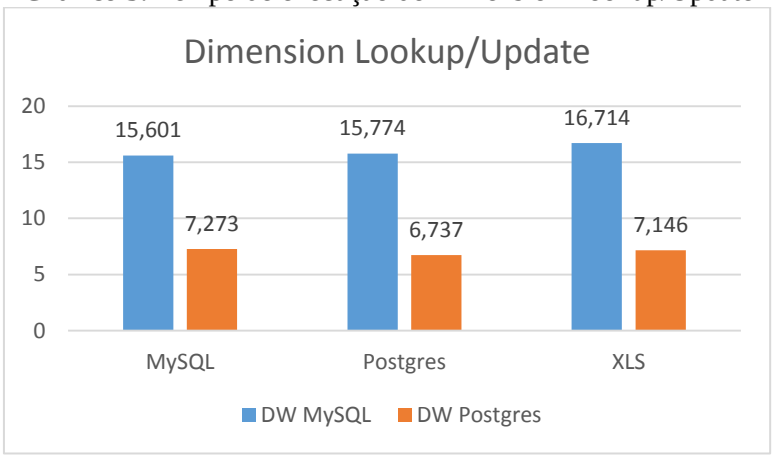
Gráfico 4. Tempo de execução do Excel Input



Dimension Lookup/Update

Este não só faz a seleção dos dados de origem do Table Input, mas também é responsável pela fase de dimensionamento das tabelas, que em seguida são carregadas para a inserção no data warehouse. Aqui pode-se ver com clareza que o Postgres utilizou menor tempo para a realização dessa tarefa.

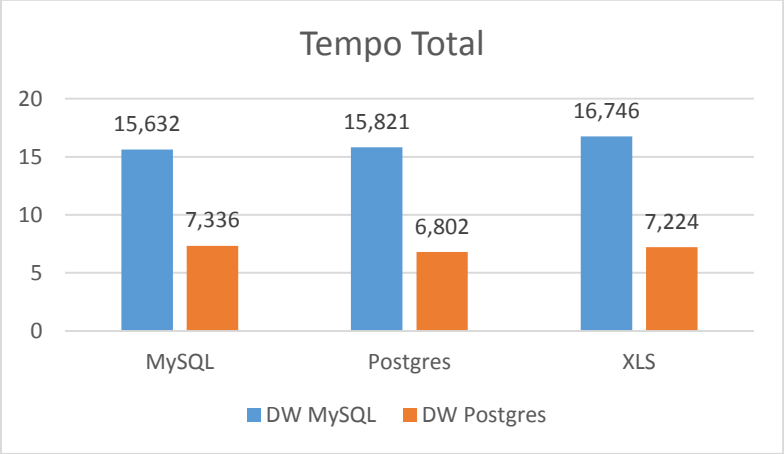
Gráfico 5. Tempo de execução do Dimension Lookup/Update



Tempo Total de Execução

O tempo total inclui a execução de todos os componentes necessários para uma ETL, donde foram obtidos os resultados finais da comparação.

Gráfico 6. Tempo total de execução



CONCLUSÃO

Ao analisar os resultados demonstrados nos gráficos nos quais foram obtidos através dos testes realizados em ambos os bancos, podemos destacar algumas conclusões sobre a eficiência deles como data warehouse.

Ambos foram de fácil implementação e uso, e garantiram a consistência dos dados nos testes realizados. Porém quando comparado a tempo de execução e processamento, o Postgres obteve melhores resultados em todos os testes, sendo possível destacar que executou os procedimentos de ETL com menos da metade do tempo que o MySQL. Na comparação de memória utilizada por cada um não foi possível distingui-los, pois a diferença entre a utilização foi em média de 1%.

Podemos então dizer que o Postgres seria a melhor opção para uso como data warehouse em um projeto de BI, sendo que garantiu melhor desempenho e utilizou menos recursos do computador, o que seria de grande valia visando que uma instituição de médio à grande porte processaria uma quantidade significativa de dados em quantidade, variedade e dimensões diferentes, podendo assim garantir a eficiência do projeto.

Referencias

FERREIRA, M.; PIAUHY, C.; CARVALHO, J.; SILVA, R.; VIEIRA, V. **Um estudo de caso com análise comparativa entre ferramentas de BI livre e proprietária**. In: Escola Regional de Banco de Dados (ERBD), 2010, Joinville. Anais... Porto Alegre: SBC, 2010.

PETRINI, M.; FREITAS, M. T.; POZZEBON, M. (2006). **Inteligência de negócios ou inteligência competitiva: noivo neurótico, noiva nervosa**. Encontro da Associação Na cional de Pós-Graduacão e Pesquisa em Administração (EnANPAD).

BATISTA E. O. SISTEMA DE INFORMAÇÃO: o uso consciente da tecnologia para o gerenciamento, São Paulo, Ed. Saraiva, 2004.

ABREU, Fábio Silva Gomes da Gama e. **Estudo de usabilidade do software Talend open Studio como ferramenta padrão para ETL dos sistemas-clientes da aplicação PostGeoOlap**. 2007. Monografia (Graduação em Sistemas de Informação) – Faculdade Salesiana Maria Auxiliadora, Macaé, 2007.

KIMBALL, Ralph. **Data Warehouse Toolkit**. Tradução Mônica Rosemberg; Revisão Técnica Ronal Stevis Cassiolato. São Paulo: Makron Books, 1998.

INMON, W. H.; HACKATHORN, Richard D. **Como Usar o Data Warehouse**. Tradução: Olávio Faria. Rio de Janeiro: Infobook, 1997.

ALMEIDA, Alexandre Marques de. **Proposição de indicadores para avaliação técnica de projetos de Data Warehouse: um estudo de caso no Data Warehouse da plataforma Lattes**. 2006. Monografia (Pós-Graduação em Engenharia de Produção) – Universidade Federal de Santa Catarina, Florianópolis, 2006.