Análise de Dados com Base em Processamento Massivo em Paralelo

Introdução

Profa. Dra. Cristina Dutra de Aguiar

Resumo:

Business intelligence pode ser definido como o processo de transformação de dados em informação e, posteriormente, em conhecimento. *Data warehousing* oferece suporte sólido para diversas funcionalidades requeridas por esse processo. Neste texto, são descritos conceitos relacionados a *business intelligence* e *data warehousing*. Também são detalhadas as diferenças entre o ambiente operacional, o qual é constituído por aplicações que oferecem suporte ao dia a dia do negócio, e o ambiente informacional, o qual é constituído por aplicações que analisam o negócio.

Conteúdo

1 BUSINESS INTELLIGENCE

Business Intelligence (BI) ou **Inteligência do Negócio** ou **Inteligência Empresarial** pode ser definido como o processo de transformação de dados em informação e, posteriormente, em conhecimento [3]. Ou seja, dados brutos em sua forma original, que considerados individualmente não possuem significado semântico, são organizados, estruturados, contextualizados frente a outros dados e processados de forma a gerar informação. O conhecimento advém da interpretação, da análise e do processamento da informação gerada. Ele possui o valor mais agregado e pode ser usado para orientar as ações das empresas, possibilitando a tomada de decisão estratégica. O conhecimento é normalmente obtido a partir das necessidades e das preferências dos clientes, dos processos de fabricação, da concorrência, das condições da indústria e das tendências econômicas, tecnológicas e culturais gerais. Exemplos de palavras-chave que são usualmente associadas ao conhecimento incluem: entendimento da informação, experiência, intuição, exploração, *know-how*, *insight*, dentre outras.

Inteligência do negócio vislumbra satisfazer às necessidades dos gerentes de analisar de forma eficiente e eficaz os dados corporativos, a fim de compreender melhor a situação do negócio e melhorar o processo de tomada de decisão estratégica. Ou seja, é voltado à análise de dados, visando à investigação de problemas estratégicos e organizacionais. Inteligência do negócio pode ser definida como um conjunto de processos que tem por objetivo produzir a informação certa para a pessoa certa na hora certa [4]. Portanto, inclui teorias, metodologias, processos, tecnologias, dentre outros, que vislubram transformar grandes quantidades de dados que, individualmente, não possuem muito significado, mas que, analisados e interpretados em conjunto, geram conhecimento essencial para uma boa gestão.

Como pensamento motivacional, pode-se destacar que a obtenção de informações estratégicas, relativas ao contexto de tomada de decisão, é de suma importância para o sucesso de uma empresa. Tais informações permitem à empresa um planejamento rápido frente às mudanças nas condições do negócio, essencial na atual conjuntura de um mercado globalizado.

Quando se pensa em inteligência de negócio, pensa-se em medidas, relatórios, análises, colaboração e gestão. Ou seja, existe a necessidade de:

- Criação de medidas (métricas) que indiquem o progresso da empresa com relação às suas metas. Um termo muito usado nesse sentido no meio empresarial é a sigla KPI (*key performance indicator*).
- Geração de relatórios que possibilitem análises complexas e que possuam visualização apropriada.
- Uso exploratório das informações com possibilidade de identificar tendências e realizar predições.

- Uso de ferramentas que possibilitem o trabalho colaborativo e que ofereçam suporte desde a obtenção dos dados até a geração do conhecimento.
- Gerenciamento do conhecimento para realizar a tomada de decisão estratégica bem fundamentada, resultando em ações bem-sucedidas que garantam um maior retorno sobre o investimento.

Data warehousing não desempenha todas as funcionalidades requeridas pelo processo de inteligência de negócio em sua amplitude. Porém, com certeza, oferece suporte sólido para diversas dessas atividades. Termos como processamento com foco em consultas analíticas, modelagem multidimensional, metodologias de projeto e outros, como técnicas de otimização e indexação, têm convergido na definição de arquiteturas modernas de sistemas de gerenciamento de dados [3].

2 DATA WAREHOUSING

Data warehousing engloba arquiteturas, algoritmos e ferramentas que possibilitam que dados selecionados de fontes de dados autônomas, heterogêneas e distribuídas sejam integradas em um único banco de dados, conhecido como *data warehouse* (DW) [5, 1]. É importante destacar que existe uma diferença entre os termos *data warehousing* e DW. DW se refere ao local aonde os dados estão fisicamente armazenados. Ou seja, o DW armazena dados, portanto, ele é o banco de dados do *data warehousing*. Já *data warehousing* é um conceito mais abrangente, que se refere ao ambiente como um todo. Ele engloba DW, *software*, *hardware* e *peopleware*.

O acesso aos dados integrados das diferentes fontes de dados é realizado, geralmente, em duas etapas. Na primeira delas, conhecida como ETL (*extract*, *transform*, *load*), traduzido como extração, transformação e carga, dados de interesse de cada fonte de dados são extraídos previamente, devendo ser traduzidos, filtrados, integrados aos dados relevantes de outras fontes e finalmente armazenados no DW. Na segunda etapa, conhecida como análise e consulta, as consultas analíticas são executadas diretamente sobre o DW, sem acessar as fontes de dados originais. Portanto, a informação integrada torna-se disponível para consulta ou análise imediata de usuários de sistemas de suporte à decisão, ou seja, usuários que podem utilizar as análises como suporte à tomada de decisão, denotados de usuários de SSD neste texto.

Considere uma empresa exemplo que será usada ao longo da disciplina. A empresa é chamada de **BI Solutions** e é apresentada a seguir.

Razão social: BI Solutions
Slogan: Desenvolvimento de soluções inteligentes para o seu negócio
Sobre a empresa: A BI Solutions é uma empresa de desenvolvimento de *software* totalmente brasileira e com alcance internacional, que implementa soluções inteligentes para atender os clientes dos mais diversos setores de negócio.

Uma aplicação de *data warehousing* referente aos salários dos funcionários que incorporam a folha de pagamento da **BI Solutions** pode integrar dados relativos aos funcionários, cargos ocupados por estes, filiais nas quais eles trabalham e datas nas quais recebem pagamento. Os assuntos de interesse dessa aplicação são os salários dos funcionários e a quantidade de lançamentos na folha de pagamento. Usuários de SSD desta aplicação podem realizar, utilizando as informações integradas:

- Análises de tendência simples. Por exemplo: Quais os gastos mensais em salários de um determinado funcionário no ano de 2019?
- Análises de tendência comparativas. Por exemplo: Quais os gastos mensais em salários dos funcionários de uma determinada filial nos últimos 3 anos?
- Análises de tendência múltiplas. Por exemplo: Quais os gastos mensais em salários dos funcionários de cada filial nos últimos 3 anos, de acordo com os diferentes cargos?

Surge, então, o questionamento se essas análises são possíveis de serem realizadas usando as aplicações de banco de dados já existentes. Muitas análises são plausíveis de serem realizadas considerando estas aplicações. Outras não, dependendo da complexidade das análises e dos dados manipulados. Mesmo que seja possível usar as aplicações de banco de dados já existentes, existem diversos desafios que precisam ser enfrentados, muitos dos quais se mostram extremamente custosos e, portanto, proibitivos para a produção da informação certa, na hora certa, para a pessoa certa.

Os seguintes desafios podem ser destacados:

- Dados de interesse de análise encontram-se espalhados em diferentes fontes de dados, assumem diferentes formatos e requerem processos de limpeza acurados. Por exemplo, dados detalhados de funcionários podem estar localizados em uma aplicação sob gerência do departamento de recursos humanos, enquanto que dados de cargos e salários podem estar detalhados em uma outra aplicação sob responsabilidade do departamento de finanças. Em adição à heterogeneidade dos dados, também encontram-se envolvidos aspectos de propriedade de cada uma das aplicações.
- Aplicações encontram-se projetadas com foco em normalização, visando diminuir e até mesmo eliminar a redundância, ou seja, a repetição dos mesmos dados.
- O foco em normalização impacta a complexidade de se especificar consultas analíticas, que passam a requerer uma grande quantidade de varreduras e junções das tabelas que representam as entidades e os relacionamentos. Neste contexto, torna-se necessária a existência de profissionais capacitados para especificar corretamente essas consultas.
- A complexidade das consultas impacta no desempenho das mesmas. Por exemplo, existem diversas situações enfrentadas pelas empresas nas quais as consultas analíticas muito

complexas executadas sobre aplicações já existentes precisam ser canceladas devido ao tempo gasto na sua execução. Pode-se citar o caso de consultas analíticas que, depois de três dias executando, devem ser interrompidas. Nesse caso, não há resposta. Mas, vamos supor que a resposta seja gerada depois de três dias de processamento, ou seja, depois de 72 horas. Qual a utilidade dessa resposta para o usuário de suporte à decisão? O tempo de resposta é viável?

- O tratamento de dados temporais usualmente é incipiente, não existindo registro temporal para todas as análises possíveis de serem realizadas. Por exemplo, pode-se guardar historicamente os salários dos funcionários. Em contrapartida, isso não é considerado para todos os atributos de todas entidades armazenadas. Por exemplo, se o nome da filial é alterado e é feita uma atualização neste nome, o nome anterior da filial usualmente é perdido.

Data warehousing visa solucionar esses desafios. Por enquanto, destaca-se que, nesse ambiente, as análises podem ser realizadas eficientemente. Isto está relacionado ao fato de que o DW armazena dados integrados, cujas diferenças semânticas e de modelo já foram eliminadas. Adicionalmente, o DW é projetado com foco em assuntos de interesse, ou seja, os dados encontram-se organizados de forma a atender às análises dos usuários de SSD. No exemplo corrente, os assuntos de interesse são os salários dos funcionários e a quantidade de lançamentos na folha de pagamento. Ainda, o DW modela explicitamente e de forma simples o aspecto temporal. A disponibilidade dos dados é outra vantagem do *data warehousing*. Uma vez que as consultas e as análises são executadas diretamente no DW sem acessar as fontes de dados originais, o DW encontra-se disponível mesmo quando as fontes não estiverem disponíveis. Pode-se citar ainda como vantagem o fato de que o processamento local nas fontes de dados originais não é afetado por causa da participação destas no ambiente de *data warehousing*.

Como pode ser observado, a comparação sendo feita está levando em consideração características de aplicações de banco de dados já existentes e aplicações de *data warehousing*. É necessário, portanto, entender as diferenças existentes entre os ambientes operacional e informacional.

3 DIFERENÇAS ENTRE OS AMBIENTES OPERACIONAL E INFORMACIONAL

O ambiente operacional é constituído por aplicações que oferecem suporte ao dia a dia do negócio. O ambiente informacional, por sua vez, é constituído por aplicações que analisam o negócio. Como pode ser observado, o ambiente de dados para o suporte aos processos de gerência e tomada de decisão (ambiente informacional) é fundamentalmente diferente do ambiente convencional de processamento de transações (ambiente operacional).

Nesse sentido, a separação entre o ambiente operacional e o ambiente informacional tornou-se uma tendência. Desde que *data warehousing* provê a base para o ambiente informacional de uma empresa, o DW é mantido separadamente dos bancos de dados operacionais. Segundo Inmon [6], esta separação é motivada pela diferença presente nos dados, tecnologias, usuários e processamento de ambos ambientes, além da necessidade de segurança e de desempenho na execução das aplicações.

3.1 ASPECTOS GERAIS

Na Tabela 1 são contrastadas as principais diferenças existentes entre os ambientes operacional e informacional considerando os seguintes aspectos gerais: principal característica, tipos de operação mais frequentes e foco do desempenho. Estas diferenças são detalhadas a seguir.

Tabela 1: Comparativo entre os ambientes operacional e informacional considerando os aspectos gerais.

	Ambiente Operacional	Ambiente Informacional
Principal Característica	voltado ao processamento de transações (OLTP)	voltado ao processamento de consultas (OLAP)
Tipos de Operação mais Frequentes	inserção remoção atualização	leitura (consulta)
Foco do Desempenho	produtividade das transações	produtividade das consultas

O ambiente operacional é voltado ao processamento de transações, usualmente conhecido como processamento OLTP (*on-line transaction processing*). No ambiente OLTP, as operações mais frequentes são de inserção, remoção e atualização. No exemplo da folha de pagamento da empresa **BI Solutions**, o ambiente operacional OLTP está voltado a prover funcionalidades relacionadas: (i) à inserção de novos funcionários contratados pela empresa; (ii) à remoção de cargos que existiam anteriormente e que foram desativados; e (iii) à atualização dos salários dos funcionários de acordo com as promoções recebidas, as quais geraram mudanças de cargos ocupados por esses funcionários. Pelo fato do ambiente operacional ser voltado ao processamento de transações OLTP, a principal questão de desempenho é a produtividade das transações. Ou seja, visa-se o desempenho na manipulação dos dados operacionais, porém as análises gerenciais, quando plausíveis de serem realizadas, são usualmente ineficientes.

Por outro lado, o ambiente informacional é voltado ao processamento de consultas, conhe-

cido como processamento OLAP (*on-line analytical processing*[1]). Nesses ambientes, as operações mais frequentes são as de leitura, ou seja, consultas voltadas à analise dos dados para o suporte à tomada de decisão. Como exemplo, tem-se a geração de um relatório que mostra os gastos mensais em salários dos funcionários nos últimos 3 anos. Pelo fato do ambiente informacional ser voltado ao processamento de consultas OLAP, a principal questão de desempenho é a produtividade das consultas, as quais são otimizadas de forma a garantir eficiência na geração de análises gerenciais.

3.2 ASPECTOS RELACIONADOS AOS USUÁRIOS

Na Tabela 2 são contrastadas as principais diferenças existentes entre os ambientes operacional e informacional considerando os seguintes aspectos relacionados aos usuários: tipos de usuários, número de usuários concorrentes e interações com usuários. Estas diferenças são detalhadas a seguir.

Tabela 2: Comparativo entre os ambientes operacional e informacional considerando os aspectos relacionados aos usuários.

	Ambiente Operacional	Ambiente Informacional
Tipos de Usuários	administradores do sistema projetistas usuários finais	usuários de SSD (executivos, analistas, gerentes)
Número de Usuários Concorrentes	grande	relativamente pequeno
Interações com os Usuários	estáticas predefinidas	dinâmicas exploratórias

Os tipos e as características dos usuários também são bastante diferentes. Exemplos de usuários do ambiente operacional incluem administradores do sistema, projetistas e usuários finais. Nesse sentido, pode existir um grande número de usuários utilizando o sistema concorrentemente. A interação com esses usuários é usualmente estática, pré-definida. Ou seja, a aplicação é usualmente especificada por meio de um conjunto de janelas e oferece um conjunto de opções que podem ser escolhidas para que os usuários realizem suas atividades. Um exemplo de usuário final é o funcionário que cadastra novos contratados pela empresa. Nesse caso, o usuário escolhe a opção “cadastrar funcionário” e digita os dados solicitados até finalizar o cadastro.

[1] O termo OLAP foi introduzido em 1993 por Codd et al. [2] para definir a categoria de processamento analítico sobre um banco de dados histórico voltado para os processos de gerência e tomada de decisão.

No ambiente informacional, tem-se os usuários de SSD, os quais ocupam cargos em níveis estratégicos da organização. Ou seja, tem-se os executivos, analistas, gerentes. Claramente, existe um número muito menor de usuários de SSD quando comparado com o número de usuários finais que usam o sistema. Portanto, o número de usuários concorrentes no ambiente informacional é relativamente pequeno. Adicionalmente, as interações dos usuários são dinâmicas, exploratórias. Considere novamente o exemplo do relatório que mostra os gastos mensais em salários dos funcionários nos últimos 3 anos. Diferentes respostas podem ser geradas nesse relatório, caso ele seja executado em diferentes filiais. No caso de uma determinada filial, os usuários de SSD podem identificar que nos meses de dezembro e janeiro concentram-se os maiores valores de salários e, a partir disso, podem realizar outras análises com o objetivo de investigar o porquê deste resultado. Em outra filial, pode ser que não haja diferença significativa na variação dos salários e que, portanto, não existam outras investigações a serem realizadas. Isso significa que as interações dependem dos resultados que vão sendo obtidos em resposta às análises sendo realizadas.

3.3 ASPECTOS RELACIONADOS ÀS OPERAÇÕES

Na Tabela 3 são contrastadas as principais diferenças existentes entre os ambientes operacional e informacional considerando os seguintes aspectos relacionados às operações: volume e característica das operações. Estas diferenças são detalhadas a seguir.

Tabela 3: Comparativo entre os ambientes operacional e informacional considerando os aspectos relacionados às operações.

	Ambiente Operacional	Ambiente Informacional
Volume das Operações	relativamente alto	relativamente baixo
Característica das Operações	mais simples, acessando menos registros por vez	mais complexas, acessando muitos registros por vez

Considerando o volume e as características das operações, pode-se destacar que no ambiente operacional tem-se um volume relativamente alto de operações de inserção, remoção e atualização. Porém, essas operações usualmente acessam menos registros por vez, sendo caracterizadas comparativamente como mais simples.

No ambiente informacional, o volume de consultas submetido por usuários de SSD durante o processo exploratório de análise é relativamente mais baixo. Porém, essas consultas possuem maior complexidade quando comparadas àquelas executadas no ambiente operacional, porque usualmente requerem o acesso a vários registros por vez, demandam várias varreduras

e várias operações de junção. As características das operações, bem como sua complexidade, serão vistas em maior nível de detalhes em outro módulo do curso.

3.4 ASPECTOS RELACIONADOS AOS DADOS

Na Tabela 4 são contrastadas as principais diferenças existentes entre os ambientes operacional e informacional considerando os seguintes aspectos relacionados aos dados: volume, projeto do banco de dados e granularidade. Estas diferenças são detalhadas a seguir.

Tabela 4: Comparativo entre os ambientes operacional e informacional considerando os aspectos relacionados aos dados.

	Ambiente Operacional	Ambiente Informacional
Volume de Dados	*megabytes* a *gigabytes*	*gigabytes* a *terabytes* a *petabytes*
Projeto do Banco de Dados	normalizado	multidimensional
Granularidade dos Dados	nível de detalhe específico	diferentes níveis de detalhe

O projeto do banco de dados operacional é normalizado. Portanto, ele foca em entidades e relacionamentos entre essas entidades. Por exemplo, na aplicação da **BI Solutions**, armazenam-se funcionários que ocupam cargos. Os substantivos *funcionário* e *cargo* representam conjuntos de entidades, enquanto o verbo *ocupa* representa conjuntos de relacionamentos entre essas entidades. Ou seja, o foco principal do projeto não se refere aos salários ou à quantidade de lançamentos. Esses dados até podem estar armazenados, porém não como foco principal de interesse. Outra característica da normalização é que ela visa diminuir ou até mesmo eliminar a redundância. No projeto do banco de dados, considera-se um único nível de detalhe, que é o nível de detalhe requerido pela aplicação. Como resultado, o ambiente operacional gerencia um volume de dados que usualmente varia de *megabytes* a *gigabytes* e que, quando comparado com o ambiente informacional, é relativamente menor.

Em contrapartida, no ambiente informacional, o projeto do banco de dados é feito de forma multidimensional, considerando as diferentes perspectivas de análise dos usuários de SSD [7]. No exemplo da folha de pagamento da **BI Solutions**, o projeto multidimensional é feito considerando os assuntos de interesse salário dos funcionários e quantidade de lançamentos na folha de pagamento. Outra característica do projeto multidimensional é que ele não elimina redundância. Adicionalmente, no exemplo corrente, a granularidade considerada é dia. Mas,

também é possível armazenar os dados considerando-se meses, bimestres, trimestres, quadrimestres, semestres e assim por diante. Como resultado, o ambiente informacional gerencia um volume de dados que usualmente varia de *gigabytes* a *terabytes* a *petabytes* e que, quando comparado com o ambiente operacional, é relativamente maior.

3.5 EXEMPLOS DE APLICAÇÃO

Na Tabela 5 são ilustrados exemplos de aplicações do ambiente operacional e do ambiente informacional.

Tabela 5: Exemplos de aplicação do ambiente operacional e do ambiente informacional.

	Ambiente Operacional	Ambiente Informacional
Exemplos de Aplicação	transações bancárias empréstimos de livros contas a pagar matrículas em cursos	planejamento de *marketing* análise financeira tomada de decisão planejamento estratégico

4 CONCLUSÃO

Neste texto, foram descritos os seguintes conceitos e aspectos relacionados:

- *Business Intelligence*: definição, objetivos, pensamento motivacional e tarefas.
- *Data warehousing*: definição, acesso às informações, exemplos de análise e de conhecimento gerado por essas análises e desvantagens do uso de sistemas existentes para a tomada de decisão em contraposto às vantagens do uso do *data warehousing*.
- Diferenças entre os ambientes operacional e informacional: definição de ambiente operacional e do ambiente informacional, separação entre os ambientes, diferenças baseadas em diferentes aspectos (gerais, usuários, operações e dados) e exemplos de aplicação.

Os conceitos apresentados nesse texto serão adicionalmente investigados ao longo da disciplina.

Referências

[1] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *SIGMOD Record*, 26(1):65–74, 1997.

[2] E.F. Codd, S.B. Codd, and C.T. Salley. Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. *White paper*, 1993.

[3] M. Golfarelli, S. Rizzi, and I. Cella. Beyond data warehousing: what's next in business intelligence? In *Proc. 7^{th} DOLAP*, pages 1–6, 2004.

[4] W. Grossmann and S. Rinderle-Ma. *Fundamentals of Business Intelligence*. Springer, 2015.

[5] J. Hammer, H. Garcia-Molina, J. Widom, W. Labio, and Y. Zhuge. The Stanford data warehousing project. *IEEE Data Engineering Bulletin*, 18(2):41–48, 1995.

[6] W. H. Inmon. *Building the Data Warehouse*. Wiley, 4th edition, 2005.

[7] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley, 2nd edition, 2002.