PREFÁCIO

Actualmente estamos confrontados com um mundo fascinante em que todos os dias surgem novidades no universo das novas tecnologias da informação, mas em que velhos problemas no que concerne à falta de usabilidade das aplicações que gerem o dia-a-dia das organizações se têm mantido como uma constante.

Os modernos sistemas de gestão de informação, utilizados pela maioria das empresas e organizações da nossa sociedade, são construídos com suporte em sistemas de bases de dados relacionais. Estes sistemas são, pela sua própria natureza estrutural, fundamentalmente vocacionadas para armazenar, com altos níveis de eficiência, os resultados das operações quotidianas das organizações.

O uso quase universal de bases de dados relacionais consolidou um conjunto de boas práticas que actualmente são respeitadas com um rigor quase religioso pela sua comunidade de utilizadores. O baixo, ou mesmo nulo, nível de redundância (repetição dos mesmos dados em mais do que um local da base de dados) é o exemplo paradigmático, e mais conhecido, de uma das regras basilares do desenho de bases de dados relacionais. Para atingir bons níveis de redundância e de integridade da informação uma base de dados é tipicamente composta por um grande número de tabelas que se encontram ligadas entre si através de campos comuns.

A exploração da informação armazenada numa base de dados é feita através da linguagem SQL (*Structured Query Language*). As técnicas de programação em SQL (vulgarmente designadas como *queries* ou pesquisas) são complexas e, consequentemente, fora do alcance dos utilizadores das bases de dados. Desse modo, a exploração dos dados está sempre limitada a um conjunto relativamente restrito de *queries* pré-programadas por especialistas na matéria. E, quando o utilizador tem uma necessidade diferente das que

foram previstas inicia-se um círculo vicioso em que é feito um pedido aos informáticos, esse pedido entra na «linha de montagem» das *queries*, e num momento no futuro o programa é, finalmente, disponibilizado ao utilizador que avaliará a sua eficácia e formatação, o que poderá provocar um *feedback* (neste caso negativo) que retomará todos as etapas do circuito.

Um gestor que pretenda analisar a informação de uma base de dados segundo uma perspectiva original, para tomar uma determinada decisão tem, além de ter conhecimentos de programação em SQL, a obrigação de conhecer em profundidade o esquema da base de dados, *i.e.*, em que tabelas estão armazenados os dados que procura e a forma como essas tabelas estão relacionadas entre si. Para um mero utilizador essas tarefas são um autêntico quebra-cabeças. Penso que nenhuma organização terá qualquer proveito em perder um bom gestor e ganhar um mau programador.

Num mundo em constante mutação os gestores não podem estar dependentes de que haja recursos disponíveis para que lhes sejam programados aplicativos que respondam às questões que nascem, por exemplo, em momentos de *brainstorming*, pois um atraso, por muito pequeno que seja, pode ter consequências dramáticas nos resultados de uma organização. Numa sociedade informada, e tecnologicamente multifacetada, não faz muito sentido que os elementos de topo nos processos de tomada de decisão tenham que depender da disponibilidade (ou da falta dela) de rotinas e procedimentos informáticos que analisem e agrupem os dados das operações da sua organização. Em muitas situações as necessidades dos gestores quase nunca estão cobertas pelos aplicativos existentes: o gestor escolhe sempre um caminho de exploração dos dados que ainda não foi desbravado pelos especialistas de informática.

Este livro contém uma descrição completa de alguns modelos e regras que permitem transformar a forma enigmática, e confusa, como os dados das operações diárias são apresentados aos utilizadores, numa nova infra-estrutura – o *Data Warehouse* – especialmente desenvolvida para promover a autonomia dos utilizadores na exploração e análise dos dados.

O *data warehousing* surge assim como uma alternativa à programação avulsa de *queries*. O seu objectivo é acabar com a dependência crónica em «quem sabe de informática». A estrutura complexa das bases de dados faz com que a atenção dos decisores esteja centrada nos obstáculos – as tabelas e a programação em SQL – e não naquilo que realmente importa: a Informa-

ção. A floresta do conhecimento está desse modo obstruída por um conjunto de mato que o gestor tem de desbravar inúmeras vezes.

O *data warehouse* tem essencialmente a ver com a autonomia do utilizador, ou seja, queremos que o utilizador seja um agente activo na exploração dos dados e não um mero recipiente passivo das *queries* programadas por outrem. Com esse novo papel o decisor poderá descobrir padrões de comportamento ou interacções entre os dados que doutra forma permaneceriam para sempre «soterrados» perante um monte de mato, muitas vezes coberto de espinhos.

Enquanto que numa base de dados o utilizador é confrontado com uma arquitectura complexa que abrange um grande número de tabelas e de associações entre elas, num armazém de dados os dados provenientes das bases de dados, e de outros sistemas operacionais, são transformados em diagramas mais pequenos – os denominados esquemas em estrela – organizados segundo as regras de negócio da organização. A última frase do parágrafo anterior é muito mais do que apenas uma pequena subtileza semântica, ela é de facto a pedra de toque, aquilo que diferencia de modo absoluto os sistemas orientados para o armazenamento de dados transaccionais, daqueles cuja tónica está posta de acordo com os processos de negócio, como é o caso do *data warehousing*.

No *data warehousing* o decisor visualiza os dados organizados de um modo semelhante à forma como ele os percebe quando executa as suas tarefas e, isso é uma grande vantagem sobre as aplicações em que a informação está estruturada segundo critérios abstractos que interessam apenas à forma como os dados têm que ser arrumados de modo a minimizar a redundância e a maximizar a sua integridade.

Um armazém de dados facilita a centralização de dados que podem ter múltiplas origens, *i.e.*, as fontes de informação que alimentam o armazém de dados podem ser, por exemplo, folhas de cálculo, ficheiros com listagens de códigos postais ou bases de dados. Estes dados podem ser produzidos internamente pela organização, como resultado das suas actividades, ou adquiridos no mercado. Todas estas fontes, independentemente da sua origem, designam-se como operacionais ou transaccionais, porque estão directamente associadas às operações básicas sobre as quais se estrutura o funcionamento das organizações.

Apesar da sua filosofia centralizadora o *data warehouse* permite que os dados sejam visualizados de acordo com uma grande plenitude de perspectivas. No *data warehouse* nunca se coloca o caso da visão única, estática e predeterminada por agentes externos, e estranhos, ao processo de tomada de decisão. Ou no mínimo não se deveria sequer equacionar essa hipótese.

Um *data warehouse*, ou armazém de dados, caracteriza-se por ter as seguintes propriedades:

a) É orientado para a publicação de dados;

b) O seu desenvolvimento é orientado exclusivamente para os utilizadores;

c) O *data warehouse* não tem nenhuma pretensão em ser uma aplicação para registar dados, nem, muito menos e ao contrário de uma base de dados, tem qualquer mecanismo de minimização da repetição de dados, pelo contrário, é vulgar, e normal, que nos armazéns de dados as mesmas peças de informação estejam repetidas em múltiplos locais do repositório de dados;

d) O armazém de dados é uma fotografia num dado momento do tempo do sistema transaccional de uma organização. Do mesmo modo que uma fotografia pode ser ampliada, cortada, ou sujeita a determinados filtros de cores e de contraste, também o *data warehouse* pode ser analisado sob múltiplos pontos de vista, e segundo uma linha temporal extensa e dinâmica;

e) Os únicos interesses que têm que ser sempre defendidos nas etapas de desenvolvimento do *data warehouse* são única e exclusivamente os dos utilizadores. Todas as considerações que tenham que ver com a optimização das estruturas de dados ou com problemas de equipamento são sempre secundárias. É como se fosse um combate de Boxe com resultado combinado: o utilizador ganha sempre no primeiro assalto por KO;

f) O *data warehouse* é independente de qualquer tecnologia e pode ser desenvolvido e utilizado em plataformas informáticas que variam desde sistemas de bases dados a sistemas proprietários especialmente concebidos para a gestão de armazéns de dados;

g) A informação num armazém de dados é sustentada por um repositório de dados construído de acordo com as regras de um modelo de dados especial denominado de modelo dimensional.

No fim, o objectivo fundamental do *Data Warehousing* é a análise da informação. É a utilização dos dados como uma forma de orientação das decisões de gestão de uma organização. Esta prática tem vindo a ter uma utilização crescente nos últimos anos e, presume-se que os indivíduos e as companhias com conhecimentos em análise terão uma grande procura no mercado.

Os negócios que negligenciarem o aspecto analítico dos seus sistemas de informação terão grandes dificuldades em competirem contra companhias com inteligência analítica. Com a contínua diminuição do custo da tecnologia a prática analítica tornar-se-á comum mesmo nas pequenas e médias empresas.

O livro *Data warehousing: conceitos e modelos* tem como objectivo fundamental apresentar os conceitos e os modelos fundamentais para o desenvolvimento de *data warehouse*s modernos e optimizados para serem facilmente utilizados. O modelo de dados dimensional, e os seus conceitos fundamentais, constituem a infra-estrutura de referência abordada detalhadamente neste livro.

Este livro destina-se não apenas aos profissionais e estudantes da área da Informática mas, a todos aqueles que no seu dia-a-dia, como é o caso dos gestores e outros decisores, sentem que as aplicações operacionais das suas organizações não se ajustam facilmente às tarefas mais específicas e especializadas inerentes aos processos de tomada de decisão.