RESUMO

Um ambiente de *data warehousing* consolida dados de interesse de provedores de informação autônomos, distribuídos e heterogêneos em uma única base de dados, chamada de *data warehouse*. Esse ambiente garante eficiência e flexibilidade na recuperação de informações estratégicas voltadas aos processos de gerência e tomada de decisão, e mantém os dados integrados no *data warehouse* com alta qualidade e confiabilidade. Os dados extraídos de cada provedor de informação são traduzidos, filtrados quando necessário e integrados com informações relevantes de outros provedores antes de serem armazenados no *data warehouse*. Este processo de carregamento dos dados é realizado de forma que consultas OLAP (*on-line analytical processing*) possam ser respondidas diretamente a partir do *data warehouse*, sem a necessidade de acesso aos provedores de informação originais. Em geral, o *data warehouse* representa uma única base de dados centralizada. Distribuir os dados armazenados nessa base de dados levando-se em consideração as características intrínsecas de aplicações de *data warehousing* e as necessidades dos usuários de sistemas de suporte a decisão apresenta várias vantagens, porém introduz novos desafios a ambientes de *data warehousing*. Dentro deste contexto, esta tese tem por objetivo propor o sistema WebD2W, enfocando um dos seus principais objetivos: a distribuição dos dados do *data warehouse*. O sistema WebD2W (*Web Distributed Data Warehousing*) é um ambiente de *data warehousing* distribuído clienteservidor que visa não somente a distribuição dos dados do *data warehouse*, mas também o acesso distribuído a esses dados usando a tecnologia Web como infra-estrutura. As suas metas genéricas consistem em: aumentar a disponibilidade dos dados do *data warehouse*, aumentar a disponibilidade de acesso a esses dados, manter a consistência dos dados distribuídos, proporcionar aumento no desempenho do processamento de consultas submetidas ao ambiente de *data warehousing*, garantir as transparências de fragmentação, de replicação e de localização na manipulação dos dados, e oferecer suporte a um grande número de usuários. Além de apresentar a arquitetura do sistema WebD2W, esta tese também propõe um conjunto de algoritmos voltados à fragmentação horizontal dos dados do *data warehouse*: algoritmo FHU–D, algoritmo FHU–DHA, algoritmo FHM–D, algoritmo FHM–DHA e algoritmo FH–MN. Esses algoritmos são baseados nos conceitos de grafo de derivação, de propagação das dimensões sendo fragmentadas aos vértices do grafo e de fragmentação ou reconstrução de agregações. Os algoritmos propostos constituem a fundamentação para o sistema WebD2W. Os diferenciais dos algoritmos propostos nesta tese referem-se ao fato de que esses algoritmos: (i) levam em consideração a organização dos dados do *data warehouse* em diferentes níveis de agregação; (ii) podem ser aplicados a diferentes cenários, de acordo com as características do grafo de derivação que representa a aplicação de *data warehousing* sendo fragmentada e com a dimensionalidade do processo de fragmentação; (iii) priorizam a execução de consultas *drill-down* e *roll-up* nos *sites* individuais, além de enfocarem consultas *slice and dice*; (iv) independem da forma de armazenamento dos dados multidimensionais em estruturas de dados relacionais (isto é, sistemas ROLAP) ou em estruturas de dados especializadas (ou seja, vi sistemas MOLAP); (v) podem ser aplicados tanto em situações nas quais todas as agregações que podem ser geradas a partir dos dados detalhados são armazenadas no *data warehouse* quanto em situações nas quais nem todas essas agregações são materializadas no *data warehouse*; e

(vi) oferecem um tratamento simétrico das dimensões e das medidas numéricas. Por fim, o sistema WebD2W foi contextualizado por meio de uma aplicação de *data warehousing* real voltada à análise de diagnósticos de saúde no Município de Belo Horizonte. Essa aplicação foi investigada em termos da importância da distribuição dos seus dados e do uso da Web como infra-estrutura de acesso. Em particular, um subconjunto dos dados da referida aplicação foi fragmentado horizontalmente pelo algoritmo FHU–D.

ABSTRACT

A data warehousing environment consolidates data of interest from distributed, autonomous and heterogeneous information sources into a single database, called as data warehouse. This environment guarantees efficiency and flexibility in the recovery of strategic information turned to management and decision-making processes, and maintains integrated data in the warehouse with high quality and reliability. The data extracted from each information source are translated, cleaned when needed and integrated with information from other sources before being stored into the data warehouse. This data loading process is performed *in advance*, so that OLAP (online analytical processing) queries can be answered directly from the data warehouse, without needing to access the original information sources. In general, the warehouse data are stored in a centralized database. Thus, the main reason for this work development is to distribute the data of such a database, taking into account both the characteristics of data warehousing applications and the needs of decision-making analysts. On the one hand, the data warehouse distribution introduces several advantages into a data warehousing environment. On the other hand, the distributed data warehousing environment must face additional challenges caused by such a distribution. In this thesis, we introduce the WebD2W system, focusing on one of its main objectives: the data warehouse distribution. The WebD2W (*Web Distributed Data Warehousing*) system is a distributed client-server data warehousing environment, which is aimed not only at the data warehouse distribution, but also at the distributed access to these data using the Web technology as an infrastructure. The generic objectives of the WebD2W system are: to increase the availability of the warehouse data, to increase the availability of access to such data, to maintain the distributed data consistency, to improve the OLAP query performance, to guarantee the fragmentation, replication and locati n transparencies in data manipulation, and to support a great number of users. Besides presenting the architecture of the WebD2W system, we also propose a set of algorithms for horizontally fragmenting the warehouse data: the FHU–D algorithm, the FHU–DHA algorithm, the FHM–D algorithm, the FHM–DHA algorithm and the FH–MN algorithm. These algorithms are based on the concepts of derivation graph, propagation of the fragmented dimensions and respective restrictions to the graph vertices, and fragmentation or reconstruction of aggregations. The proposed algorithms are used as a basis for the WebD2W system. The differentials of the proposed algorithms refer to the fact that these algorithms: (i) take into account the data warehouse organization in different levels of aggregation; (ii) can be applied to different scenarios, according both to the characteristics of the derivation graph which represents the data warehousing application being fragmented and to the dimensionality of the fragmentation process; (iii) focus on the execution of drill-down, roll-up and slice and dice queries in the individual sites; (iv) are independent of the multidimensional data storage in relational data structures (i.e., ROLAP systems) or in specialized data structures (i.e., MOLAP systems); (v) can be applied both when all aggregations that can be generated from the detailed data are stored in the warehouse and when not all aggregations are stored in such a database; and (vi) treat dimensions and numeric measures symmetrically. Finally, we contextualized the WebD2W system by means of a real data warehousing application. Such an application is aimed at analyzing the healthcare data of Belo Horizonte's Municipal district. The application was

investigated in terms of both the importance of its data distribution and the use of the Web technology as an infrastructure. A subset of the healthcare data was horizontally fragmented by the FHU–D algorithm.