Sistemas de Informação II

Arquiteturas e Componentes do Data Warehouse. Definição, comparação e enquadramento num caso de estudo

João Mortágua-

João Choupina Ferreira da Mota- 2020151878

Índice

1. Resumo

O nosso projeto baseia-se na investigação das diversas arquiteturas e componentes essenciais de um Data Warehouse (DW). Com isto, tentámos contextualizar a importância destes no ambiente tecnológico atual, bem como a necessidade de compreender e distinguir arquiteturas e enquadrá-las num caso de estudo específico.

Como referido anteriormente, a pesquisa explora minuciosamente os diferentes tipos de arquiteturas e destaca as suas vantagens e desvantagens. Além disso, são descritos os principais componentes entre os quais a base de dados central, ferramentas ETL, metadados e ferramentas de acesso.

O principal foco é a comparação de abordagens, evidenciando as suas aplicações práticas em contextos reais.

2. Introdução

A rápida expansão das operações digitais e a crescente geração de dados desafiam as organizações a adotarem estratégias eficazes para armazenamento, processamento e análise de informações. Neste contexto, os Data Warehouse (DW) emergem como elementos fundamentais para a gestão inteligente desses volumes massivos de dados. Este projeto irá explorar de forma abrangente as arquiteturas e os componentes essenciais dos Data Warehouse, visando fornecer insights significativos para a implementação prática e eficiente dessas estruturas.

No cenário empresarial atual, onde a informação é um recurso estratégico, a capacidade de extrair conhecimento a partir de grandes conjuntos de dados torna-se crucial para a tomada de decisões. Os Data Warehouse representam a resposta a esse desafio, proporcionando uma infraestrutura dedicada à consolidação, integração e análise de dados provenientes de diversas fontes.

Contudo, a complexidade crescente das arquiteturas e a diversidade dos componentes disponíveis podem apresentar desafios significativos para a implementação eficaz de Data Warehouse. Com este projeto visamos aprofundar a compreensão destas arquiteturas e componentes e analisar as suas aplicações.

Resumindo, este estudo visa uma análise crítica das diversas arquiteturas de Data Warehouse e uma descrição detalhada dos seus essenciais componentes.

3. Estado de arte genérica sobre Data Warehouse

3.1. O que é um Data Warehouse?

Um Data Warehouse é um repositório central de informações que podem ser analisadas para tomar decisões mais adequadas. Os dados fluem de sistemas transacionais, bancos de dados e de outras fontes para o Data Warehouse, normalmente com uma cadência regular. Analistas de negócios, engenheiros de dados, cientistas de dados e "decision makers" acessam os dados por meio de ferramentas de business intelligence (BI), clientes SQL e outras aplicações de análise.

Dados e análises tornaram-se indispensáveis para que as empresas se mantenham competitivas. Os usuários corporativos contam com relatórios, painéis e análises para extrair insights dos dados, monitorizar a performance dos negócios e apoiar a tomada de decisões. Os Data Warehouse alimentam esses relatórios, painéis e ferramentas de análise armazenando dados de maneira eficiente para minimizar a entrada e saída dos dados e fornecer resultados de consulta rapidamente para centenas e milhares de usuários simultaneamente.

3.2. Benefícios

- Tomada de decisões adequada
- Dados consolidados de diversas fontes
- Análise de dados históricos
- Qualidade, consistência e precisão dos dados
- Separação do processamento analítico dos bancos de dados tradicionais

3.3. Importância Estratégica

No contexto empresarial moderno, o Data Warehouse transcende a mera consolidação de dados. Tornou-se uma peça central na estratégia de análise de dados, capacitando empresas a transformarem informações brutas em inteligência acionável para impulsionar decisões estratégicas.

3.4. Desafios

A implementação e manutenção de Data Warehouse enfrentam desafios significativos. Questões relacionadas à segurança e qualidade dos dados destacam-se como obstáculos práticos. A garantia de segurança dos dados armazenados, a manutenção da integridade, a qualidade dos dados e a gestão eficiente de grandes volumes são desafios críticos que as organizações enfrentam.

Inovações tecnológicas, integração de fontes de dados não tradicionais e a evolução das demandas dos usuários finais representam também novos desafios e oportunidades à medida que as organizações procuram manter a relevância e a eficácia dos seus Data Warehouse.

4. Análise/Descrição do tema escolhido

4.1. Arquiteturas

4.1.1. Centralizada

A arquitetura centralizada concentra todos os dados num único repositório central, simplificando a gestão e garantindo consistência. Este modelo destaca-se pela sua simplicidade, facilidade de implementação e manutenção centralizada. Todos os dados são armazenados num local dedicado, facilitando o controlo e a garantia de integridade.

Algumas vantagens incluem a simplicidade de implementação, manutenção centralizada e garantia de consistência dos dados. No entanto, desvantagens podem surgir em ambientes com grande volume de dados distribuídos, onde a centralização pode resultar em problemas de desempenho.

4.1.2. Distribuída

A arquitetura distribuída dispersa os dados em vários locais, proporcionando um maior desempenho em ambientes complexos. A descentralização oferece flexibilidade, permitindo que diferentes unidades de uma organização gerenciem e acessem os dados localmente.

A principal vantagem é a escalabilidade, pois a carga de dados é distribuída, evitando desleixos. No entanto, a gestão descentralizada pode introduzir complexidades operacionais e de segurança, requerendo soluções robustas.

4.1.3. Hub and Spoke

A arquitetura Hub and Spoke centraliza os dados cruciais num *hub*, conectando-se a satélites especializados. Isto equilibra a centralização e descentralização para otimizar o desempenho e a flexibilidade. O *hub* serve como ponto central de controlo e consolidação, enquanto os satélites mantêm a flexibilidade local.

Essa arquitetura procura maximizar a eficiência operacional, permitindo a centralização do controlo enquanto mantém a flexibilidade em locais específicos. No entanto, requer uma gestão cuidadosa da arquitetura para garantir coesão e bom desempenho.

4.1.4. Virtual

A arquitetura Data Warehouse Virtual permite o acesso a dados distribuídos sem a necessidade de consolidação física. Esta cria uma camada virtual que integra dados de fontes heterogêneas, proporcionando flexibilidade na análise sem a sobrecarga de consolidar fisicamente os dados.

Esta arquitetura oferece flexibilidade, permitindo a análise de dados sem a necessidade de os mover fisicamente. No entanto, a virtualização pode introduzir atrasos devido à necessidade de acessar a dados distribuídos e desafios de integração devem ser gerenciados de forma eficaz.

4.2. Componentes

4.2.1. Área de Staging

A área de staging é a zona intermediária onde os dados brutos são extraídos de fontes diversas, passam por transformações necessárias e, finalmente, são carregados no Data Warehouse. O processo ETL desempenha um papel vital na limpeza, enriquecimento e integração dos dados.

4.2.2. Database

A base de dados do Data Warehouse armazena os dados consolidados e transformados. Os esquemas dimensionais (como estrela e floco de neve) são comuns para otimizar consultas analíticas, enquanto esquemas normalizados garantem a integridade e eficiência do armazenamento.

4.2.3. Servidor OLAP (Online Analytical Processing)

O servidor OLAP permite análises multidimensionais eficientes. Modelos MOLAP (Multidimensional OLAP) armazenam dados pré-agregados para uma rápida resposta, enquanto modelos ROLAP (Relational OLAP) consultam dados diretamente no armazenamento relacional para uma maior flexibilidade.

4.2.4. Ferramentas de ETL (Extração, Transformação e Carga)

Ferramentas ETL automatizam o fluxo de dados desde a extração até ao carregamento, simplificando e agilizando o processo. Essas ferramentas desempenham um papel fundamental na integridade e atualização contínua dos dados no Data Warehouse.

A escolha da correta ferramenta afetará

- O tempo gasto na extração de dados
- Abordagens para extrair dados
- Tipo de transformações aplicadas

4.2.5. Metadados

O repositório de metadados armazena informações sobre os dados no Data Warehouse, incluindo origens, transformações aplicadas, estruturas de dados e relacionamentos. Isso proporciona transparência para os usuários finais e respetivos administradores.

4.2.6. Ferramentas de Relatório e Análise

Ferramentas de relatório e análise capacitam os usuários a extraírem insights dos dados armazenados. Estas oferecem interfaces intuitivas para a criação de consultas, relatórios e painéis, facilitando a tomada de decisões informadas.

4.2.7. Segurança e Gerenciamento

Mecanismos de segurança garantem que apenas usuários autorizados tenham acesso a determinados dados. O gerenciamento de usuários controla permissões, garantindo a integridade e a confidencialidade de dados sensíveis.

4.2.8. Backup

Procedimentos de backup e recuperação são essenciais na proteção contra perda de dados. Estratégias eficazes garantem a disponibilidade contínua do Data Warehouse, mesmo em caso de falhas inesperadas.

4.3. Caso de Estudo

4.3.1. Contextualização

Para aplicar este conhecimento num caso de estudo prático decidimos optar pela escolha de um banco internacional para o fazer. Este atua em diversos segmentos como investimentos e serviços corporativos tendo ele uma presença global e que enfrenta desafios complexos.

A escolha deste caso baseia-se na sua representatividade no setor financeiro e na necessidade crítica de implementação de um Data Warehouse robusto que aprimore as suas operações. A sua complexidade de operações e os rígidos regulamentos oferecem um cenário ideal para analisar a aplicação prática de arquiteturas e componentes do Data Warehouse.

4.3.2. Arquitetura escolhida

Neste banco, optou-se por uma abordagem de arquitetura distribuída para o Data Warehouse. Esta escolha foi motivada pela necessidade de lidar com grandes volumes de dados transacionais, distribuídos em diferentes unidades de negócios ao redor do mundo. A descentralização permite escalabilidade e agilidade na análise de dados em ambientes complexos.

Por outras palavras, esta arquitetura foi escolhida devido à sua capacidade de escalar horizontalmente, adaptando-se às constantes evoluções do setor financeiro. Isto permite que cada unidade de negócios mantenha a sua autonomia operacional, enquanto os dados relevantes são centralizados para análises mais abrangentes.

4.3.3. Componentes escolhidos

A implementação envolve uma área de staging robusta, onde os dados brutos de diversas fontes são extraídos. O processo ETL é altamente automatizado, garantindo a consistência e qualidade dos dados durante as fases de transformação e carga. Isto é crucial para a integridade das análises realizadas posteriormente.

A base de dados do Data Warehouse utiliza uma combinação de esquemas dimensionais e normalizados. Os esquemas dimensionais são aplicados para áreas que exigem consultas analíticas frequentes, enquanto os esquemas normalizados são utilizados para garantir a integridade e a eficiência no armazenamento de grandes conjuntos de dados.

Para análises multidimensionais, o banco optou por uma abordagem híbrida. O uso de MOLAP é preferido para agregações prévias, proporcionando respostas rápidas a consultas frequentes, enquanto o ROLAP é adotado para consultas mais flexíveis.

O processo ETL é executado por ferramentas especializadas, garantindo eficiência e consistência.

Os metadados são gerenciados num repositório central, proporcionando uma visão abrangente das origens dos dados, transformações aplicadas e relacionamentos entre conjuntos de dados. Isto é fundamental para acompanhar o histórico dos dados e garantir que eles sejam usados de forma responsável.

5. Conclusões

Ao implementar um Data Warehouse numa instituição financeira, a escolha de uma arquitetura distribuída proporcionou uma análise abrangente e autonomia operacional. Com isto, a instituição obteve uma redução significativa de perdas financeiras.

A implementação impactou positivamente os resultados operacionais, catalisando uma mudança na abordagem estratégica. A capacidade de respostas rápidas a consultas analíticas e a deteção proativa de padrões anômalos redefiniram o cenário operacional.

Em última análise, este projeto destaca não apenas a eficácia do Data Warehouse na mitigação de desafios específicos do setor financeiro, mas também ressalta o papel crítico da tecnologia na transformação positiva das operações e decisões estratégicas em organizações complexas e dinâmicas.

6. Referências

https://www.astera.com/pt/knowledge-center/data-warehouse-architecture/

https://dspace.uevora.pt/rdpc/bitstream/10174/22072/1/Mestrado%20-%20Engenharia%20Inform%C3%A1tica%20-%20Ad%C3%A3o%20Baptista%20Pereira%20Lopes%20-%20Aplica%C3%A7%C3%A3o%20de%20t%C3%A9cnicas%20de%20business%20intelligence....pdf

https://aws.amazon.com/pt/what-is/data-warehouse/

https://www.astera.com/pt/knowledge-center/data-warehouse-architecture/

https://www.oracle.com/pt/database/what-is-a-data-warehouse/#link3