Participante

Edward Neftalí Liriano Gómez 2022-0437

Profesor

Francis Ramírez

Asignatura

Electiva 1 (Big Data)

Tema

Resumen de SQL Data Warehouse (Módulos 1 y 2)

Índice

Introducción a Data Warehousing

El data warehousing es una técnica bien establecida para centralizar los datos empresariales para informes y análisis. En la lección 1 se presentan los fundamentos de la data warehousing, incluyendo los problemas empresariales que aborda, la definición de un almacén de datos, arquitecturas comunes, componentes de una solución de data warehousing y roles involucrados en proyectos de data warehousing.

La Inteligencia de Negocios (BI, por sus siglas en inglés) es una técnica bien establecida para centralizar los datos empresariales para informes y análisis. Aunque los detalles específicos de las soluciones individuales pueden variar, hay elementos comunes en la mayoría de las implementaciones de inteligencia de negocios. La familiaridad con estos elementos facilitará la planificación y construcción de una solución efectiva de inteligencia de negocios.

El problema empresarial abordado incluye la dificultad para tomar decisiones efectivas debido a la distribución de datos clave en múltiples sistemas, lo que dificulta la recopilación y reconciliación de datos para la toma de decisiones. Un almacén de datos proporciona una solución centralizada para este problema, almacena grandes volúmenes de datos históricos de transacciones empresariales y está optimizado para operaciones de lectura que admiten consultas de datos.

Las arquitecturas de data warehousing incluyen la creación de un almacén de datos central, data marts departamentales o una arquitectura de cubo y radio que sincroniza un almacén de datos central con data marts departamentales. Los componentes de una solución de data warehousing incluyen fuentes de datos, proceso de extracción,

transformación y carga (ETL), áreas de preparación de datos, almacén de datos y modelos de datos.

Proyectos de Almacén de Datos:

- Los proyectos de almacén de datos tienen mucho en común con cualquier otra implementación de TI y pueden aplicar metodologías comúnmente utilizadas como Ágil, Cascada o Microsoft Solutions Framework (MSF).
- Sin embargo, a menudo requieren una comprensión más profunda de los objetivos comerciales clave y las métricas utilizadas para impulsar la toma de decisiones.

Roles en Proyectos de Almacén de Datos:

- Gerente de Proyecto.
- Arquitecto de Soluciones.
- Modelador de Datos.
- Administrador de Base de Datos.
- Especialista en Infraestructura.
- Desarrollador de ETL.
- Usuarios de Negocios.
- Gestores de Datos.

En la lección 2 se abordan consideraciones adicionales para una solución de data warehousing, como el diseño del esquema de la base de datos, índices de columna, fuentes de datos, proceso ETL, calidad de datos y gestión de datos maestros. Se discuten aspectos como el esquema de la base de datos, hardware, alta disponibilidad y recuperación ante desastres, seguridad, índices de columna, fuentes de datos y procesos ETL.

1. **Diseño de la Base de Datos del Almacén de Datos**:

- Se discute la importancia de diseñar tanto el esquema lógico como físico del almacén de datos. Se sugiere utilizar un esquema de estrella, donde las medidas numéricas se almacenan en tablas de hechos y se vinculan a múltiples tablas de dimensiones que contienen entidades comerciales relevantes.
- Se enfatiza la necesidad de comprender las dimensiones y medidas que serán utilizadas por los usuarios comerciales, así como la planificación cuidadosa de las claves para vincular hechos a dimensiones.

2. Hardware para el Almacén de Datos:

- Se aborda la selección de hardware, considerando los requisitos de procesamiento de consultas, almacenamiento, conectividad de red, redundancia de componentes y requerimientos de memoria en caso de uso de tecnologías en memoria.
- Se ofrecen opciones para construir la solución de almacén de datos, ya sea adquiriendo componentes individuales, utilizando

arquitecturas de referencia predefinidas o adquiriendo un dispositivo de hardware preconfigurado.

3. **Alta Disponibilidad y Recuperación ante Desastres**:

- Se discuten las técnicas de alta disponibilidad disponibles en SQL Server, como el mirroring de bases de datos y el clustering de servidores, y se proporcionan criterios para seleccionar la mejor opción para una solución específica.
- Se destaca la importancia de la redundancia a nivel de componentes individuales y se recomienda planificar una estrategia de recuperación ante desastres que incluya una copia de seguridad exhaustiva y tiempos de recuperación aceptables.

4. **Seguridad**:

- Se abordan los aspectos de seguridad relacionados con el acceso a los datos del almacén, incluyendo mecanismos de autenticación, permisos de usuarios y seguridad física de la base de datos y los medios de respaldo.
- Se presenta la opción de Always Encrypted en SQL Server 2016 como una forma de realizar operaciones en datos cifrados, manteniendo las claves en el cliente para proteger los datos sensibles.

5. **Índices de Almacén de Columnas**:

- Se introducen los índices de almacén de columnas, explicando cómo reducen el espacio en disco y las operaciones de E/S mediante la compresión agresiva de datos.
- Se describen las nuevas características de los índices de almacén de columnas en SQL Server 2016, como la compatibilidad con índices no agrupados y la habilitación del modo batch en funciones de agregación.

6. **Fuentes de Datos y Procesos ETL**:

- Se discuten los diferentes tipos de fuentes de datos y los métodos para conectar y extraer datos para el almacén de datos, incluyendo consideraciones sobre credenciales, formatos de datos y ventanas de adquisición de datos.
- Se detallan los procesos ETL, incluyendo la necesidad de etapas de almacenamiento intermedio para sincronizar, validar y transformar los datos antes de cargarlos en el almacén de datos.

7. **Calidad de Datos y Gestión de Datos Maestros**:

- Se explica la importancia de mantener la calidad de los datos en el almacén de datos y se sugiere la participación de usuarios comerciales como administradores de datos.
- Se presenta la gestión de datos maestros como una solución para garantizar la consistencia de los datos en múltiples aplicaciones comerciales, y se discute la opción de utilizar SQL Server Data Quality Services y SQL Server Master Data Services para este fin.

Planificación de la Infraestructura del Almacén de Datos

Descripción General del Módulo: El éxito de cualquier proyecto de inteligencia empresarial (BI) depende en gran medida de la infraestructura subyacente que sustenta la solución. Este módulo aborda cómo planificar y diseñar la infraestructura del almacén de datos, centrándose en optimizar el rendimiento, la escalabilidad y la disponibilidad mientras se gestionan los costos y la complejidad de implementación.

Lección 1: Consideraciones para la Infraestructura del Almacén de Datos

Esta lección comienza examinando las diversas consideraciones involucradas en la planificación de la infraestructura de un almacén de datos. Se destaca la importancia de comprender el tamaño y la complejidad de la solución de BI, así como el número de usuarios y los requisitos de disponibilidad. Además, se analizan los diferentes tipos de cargas de trabajo, como las tareas de Extracción, Transformación y Carga (ETL), los modelos de datos de SQL Server Analysis Services (SSAS) y las cargas de informes de SQL Server Reporting Services (SSRS).

El tamaño de un almacén de datos se clasifica típicamente en pequeño, mediano o grande, en función del volumen de datos, la complejidad del análisis y la cantidad de usuarios. Se discuten las características típicas de cada tamaño, así como los requisitos de disponibilidad asociados.

Además, se profundiza en las diferentes cargas de trabajo que afectan la infraestructura del almacén de datos, como las tareas de ETL, el procesamiento de modelos de datos y la generación de informes. Cada carga de trabajo tiene demandas específicas de recursos de hardware, como CPU, memoria, E/S de disco y ancho de banda de red, que deben considerarse al planificar la infraestructura.

Planificación del Hardware del Almacén de Datos

Esta lección se enfoca en la planificación del hardware específico necesario para admitir la solución de almacén de datos. Se examinan las opciones de hardware disponibles y cómo seleccionar la configuración adecuada en función de los requisitos de la solución. Se discuten aspectos como el procesador, la memoria, el almacenamiento y la red, y cómo cada uno afecta el rendimiento y la escalabilidad del almacén de datos.

Se presentan diferentes topologías de servidores, desde arquitecturas de un solo servidor hasta distribuciones más complejas que separan las cargas de trabajo de ETL, análisis y generación de informes en servidores dedicados. También se abordan las estrategias para lograr alta disponibilidad, como el uso de clústeres de conmutación por error y grupos de disponibilidad.

Revisión del Módulo y Conclusiones

En esta sección, se revisan los conceptos clave discutidos en el módulo y se resumen las mejores prácticas para planificar y diseñar la infraestructura del almacén de datos. Se enfatiza la importancia de considerar cuidadosamente los requisitos de la solución de BI y

seleccionar la configuración de hardware adecuada para garantizar el rendimiento, la escalabilidad y la disponibilidad óptimos. Además, se destacan los próximos pasos para continuar aprendiendo sobre el tema y aplicar los conocimientos adquiridos en proyectos reales de BI.

Se profundiza en la planificación del hardware para un almacén de datos (data warehouse), un componente fundamental en las soluciones de inteligencia empresarial (BI). Se destaca la necesidad de considerar cuidadosamente las recomendaciones proporcionadas por Microsoft y sus socios de hardware, ya que el diseño de un almacén de datos difiere significativamente de otras cargas de trabajo de bases de datos.

Importancia del Hardware en el Almacén de Datos:

- Se subraya que el almacén de datos es la base de una solución de BI, lo que implica que su diseño y configuración tienen un impacto crucial en el rendimiento y la eficacia de toda la solución.

Arquitecturas de Referencia de Fast Track Data Warehouse:

- Se introducen las arquitecturas de referencia de Fast Track Data Warehouse de Microsoft SQL Server, diseñadas por especialistas en SQL Server y consultores de socios de hardware.
- Estas arquitecturas proporcionan especificaciones predefinidas de hardware y configuración que han sido ampliamente probadas con cargas de trabajo reales de almacenes de datos.

Soporte Multivendor y Pretestado:

- Microsoft se ha asociado con múltiples proveedores de hardware para ofrecer diseños de sistemas pretestados y certificados que pueden implementarse más fácilmente.

Herramientas de Especificación del Sistema:

- Se menciona la disponibilidad de herramientas proporcionadas por Microsoft y sus socios de hardware para ayudar en la determinación de la configuración de hardware necesaria.

Arquitectura de Sistema Equilibrado por Núcleos:

- Se explica el concepto de arquitectura equilibrada por núcleos, que se basa en el entendimiento de que las cargas de trabajo de los almacenes de datos tienden a requerir transferencias de datos masivas a través de múltiples componentes del sistema.

Determinación de Requisitos de Procesador y Memoria:

- Se detalla el proceso para determinar los requisitos de procesador y memoria, centrándose en el cálculo del MCR (Tasa de Consumo Máximo) por núcleo de CPU y su aplicación en la estimación del número de núcleos necesarios.

Determinación de Requisitos de Almacenamiento:

- Se aborda el proceso de estimación de los requisitos de almacenamiento, que implica considerar el volumen de datos que el sistema debe manejar, así como el crecimiento futuro y la compresión de datos.

Consideraciones para el Hardware de Almacenamiento:

- Se discuten varios factores a tener en cuenta al seleccionar el hardware de almacenamiento, incluido el tamaño y la velocidad del disco, la configuración RAID, la elección entre DAS o SAN, entre otros.

Almacenes de Datos de SQL Server:

- Se destaca la disponibilidad de aparatos preconfigurados basados en las arquitecturas de referencia de Fast Track Data Warehouse, así como la existencia de soluciones específicas para almacenes de datos a gran escala, como SQL Server Parallel Data Warehouse.