Curso ETL Avanzado

Tabla de contenido

Introducción a Pentaho

Introducción a Pentaho

- Es una plataforma de BI “orientada a la solución” y “centrada en procesos” que incluye los componentes necesarios para implementar soluciones basadas en procesos.
- Pentaho posee principalmente una infraestructura/base de herramientas de análisis e informes integrado con un motor de workflow, con el fin de ejecutar las reglas de negocio necesarias que se encuentran expresadas en forma de procesos, actividades, además es capaz de generar la información de salida necesaria.
- Su modelo de ingresos parece estar orientado a los servicios (soporte, formación, consultoría y soporte a ISVs y distribuciones OEM) aunque posee algunas funcionalidades “Premium” que involucran un pago.

Introducción a Pentaho / Productos

- **Pentaho Analysis Services (Mondrian)**: es un servidor OLAP escrito en Java que permite analizar grandes cantidades de datos almacenados en bases de datos SQL de una forma interactiva sin necesidad de escribir las sentencias que serían necesarias para ello en SQL.
- Entre otras características destaca la compatibilidad con el MDX (expresiones multidimensionales) y el lenguaje de consulta XML(XMLA) para análisis y olap4j.

Introducción a Pentaho / Productos

- **Pentaho Reporting**: Consiste en un motor de presentación, capaz de generar informes programáticos definido en un archivo XML. Sobre esta solución se han desarrollado muchas herramientas, por ejemplo informes, diseñadores de interfaz gráfica de usuario, y asistentes tipo wizard. Un uso notable de esta herramienta es el Generador de informes para OpenOffice.org

Introducción a Pentaho / Productos

- **Pentaho Data Mining**: Es una envoltura alrededor del proyecto Weka. Es una suite de software que usa estrategias de aprendizaje de máquina, aprendizaje automático y minería de datos. Cuenta con series de clasificación, de regresión, de reglas de asociación, y de algoritmos de clustering, para así apoyar las tareas de análisis predictivo.

Introducción a Pentaho / Productos

- **Pentaho Dashboard**: Es una plataforma integrada que proporcionar información sobre sus datos, donde se pueden ver informes, gráficos interactivos y los cubos creados con las herramientas Pentaho Report Designer.

Introducción a Pentaho / Productos

- **Pentaho para Apache Hadoop**: Es un conector de bajo nivel para facilitar el acceso a grandes volúmenes de datos manejados en el proyecto Apache Hadoop.
- Apache Hadoop es un framework que soporta aplicaciones distribuidas bajo una licencia libre. Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Hadoop se inspiró en los documentos Google para MapReduce y Google File System (GFS).
- Hadoop es un proyecto de alto nivel Apache que está siendo construido y usado por una comunidad global de contribuyentes, mediante el lenguaje de programación Java. Yahoo! ha sido el mayor contribuyente al proyecto, y usa Hadoop extensivamente en su negocio.

- http://hadoop.apache.org/

Novedades en Pentaho 8.0

Novedades en Pentaho 8.0

- **AEL (Adaptive Execution Layer) mejorado y simplificado**

La característica de Capa de ejecución adaptable (AEL) incluye compatibilidad para las bibliotecas Apache Spark ™ empaquetadas con las distribuciones Cloudera, Hortonworks y Apache. Además, AEL ahora cuenta con un tiempo de inicio más rápido, características de seguridad mejoradas y un rendimiento mejorado al ejecutar archivos de transformación (.ktr).

https://spark.apache.org/

Novedades en Pentaho 8.0

- **Kafka y Streaming Ingestion en PDI**

PDI ahora admite la transmisión de Kafka (**Apache Kafka®**) con la ingesta de datos (paso de Kafka Consumer) y la publicación de datos (paso de Kafka Producer). Puede aprovechar la transmisión de Kafka para su uso en análisis, monitoreo y alertas casi en tiempo real. Puede conectarse a una fuente de datos de transmisión, como Kafka, y luego ingerir continuamente datos de transmisión. Por ejemplo, PDI con Kafka puede consumir eventos de flujo de clics de aplicaciones web, datos comerciales y registros de puntos de ventas. Estas características de ingestión de transmisiones en tiempo real están dirigidas a arquitectos e ingenieros de datos, desarrolladores de ETL y administradores de TI. PDI también admite la comunicación segura y la autenticación de usuarios con Kafka.

https://kafka.apache.org

Novedades en Pentaho 8.0

- **Big Data Security: Clusters con nombre y soporte de Knox**

Pentaho ahora es compatible con Apache Knox en la distribución de Hadodows de Hadoop, lo que proporciona un punto de acceso único y seguro a los componentes de Hadoop en un clúster. Esta nueva integración permite a los clientes aprovechar PDI de forma segura y transparente cuando están utilizando clústeres de Hortonworks protegidos por Knox para el procesamiento de grandes datos. Cuando se utiliza junto con Apache Ranger™, puede controlar el acceso de nivel de usuario en su clúster. Además del control de acceso, Ranger le permite crear un registro de auditoría del acceso y las acciones del usuario.

https://knox.apache.org/

https://ranger.apache.org/

Novedades en Pentaho 8.0

- **Crear configuraciones de ejecución para servidor Pentaho**

Algunas actividades de ETL son ligeras, como cargar un archivo de texto pequeño para escribir en una base de datos. Para estas actividades, puede ejecutar su transformación localmente utilizando el motor Pentaho predeterminado. Algunas actividades de ETL son más exigentes, ya que contienen muchos pasos para llamar a otros pasos o una red de módulos de transformación. Para estas actividades, puede configurar una configuración de ejecución dedicada para ejecutar transformaciones en el servidor Pentaho.

Novedades en Pentaho 8.0

- **Nodos de trabajo**

Al usar los nodos de trabajo, ahora puede escalar de manera elástica las transformaciones y trabajos de PDI (es decir, elementos de trabajo) de manera fácil y segura, al mismo tiempo que los coordina y supervisa. Para monitorear el estado del elemento de trabajo, puede verificar una interfaz basada en web o usar la funcionalidad existente de Pentaho Operations Mart o los datos de registro de SNMP. Esta funcionalidad permite que las cargas de trabajo de PDI se ejecuten de manera efectiva a escala, coordinando y monitoreando los elementos enviados a los nodos de trabajo. Cuando se completa la ejecución, los elementos de trabajo completados se devuelven al servidor Pentaho.

Novedades en Pentaho 8.0

- **Filtros para la inspeccionar de datos**

Es posible explorar de mejor forma los datos de transformación aplicando filtros y visualizando los resultados. Podemos agregar filtros arrastrando campos al panel filtros o realizando acciones dentro de la visualización. También puede agregar múltiples filtros, con cada filtro individual refinando aún más los datos. La información sobre herramientas del panel filtros muestra un resumen de los filtros aplicados. Los filtros se pueden aplicar tanto en la vista de secuencia como en la de modelo, también a tablas planas, gráficos y tablas dinámicas.

Novedades en Pentaho 8.0

- **Formatos adicionales de Big Data**

Pentaho 8.0 agrega soporte para formatos de datos Avro y Parquet en PDI. Para los usuarios de big data, los pasos mejorados de transformación de entrada / salida de Avro y Parquet facilitan el proceso de recopilación de datos sin procesar de varias fuentes y la transferencia de esos datos al ecosistema de Hadoop para crear un conjunto de datos útil y resumido para el análisis. Las organizaciones ahora pueden diseñar tuberías de big data con un rendimiento de lectura y uso de almacenamiento óptimos aprovechando la interfaz de arrastrar y soltar PDI familiar y fácil de usar. Capa (AEL).

  - Avro Input
  - Avro Output
  - Parquet Input
  - Parquet Output

Novedades en Pentaho 8.1

Novedades en Pentaho 8.1

- **Steps de streaming mejorados en PDI**

Pentaho Data Integration (PDI) presenta diversas mejoras en su perfil de steps de streaming , incluyendo 2 pasos steps.

MQTT Consumer y MQTT Producer: El cliente PDI ahora puede extraer datos de transmisión desde un intermediario o clientes MQTT a través de una transformación MQTT. El paso **MQTT Consumer** ejecuta una transformación secundaria que se ejecuta de acuerdo con el tamaño o la duración del lote de mensajes, lo que le permite procesar un flujo continuo de registros casi en tiempo real. El paso **MQTT Producer** le permite publicar mensajes casi en tiempo real a un agente de MQTT.

Novedades en Pentaho 8.1

- **Mejoras JMS Consumer y JMS Producer**

Los pasos de JMS Consumer y JMS Producer ahora son compatibles con el middleware de IBM MQ, lo que le permite crear flujos de datos con tales fuentes de datos heredadas (IBM MQ). Al igual que nuestros otros pasos de transmisión, el paso de JMS Consumer ahora funciona como una transformación principal que ejecuta una transformación secundaria que se ejecuta de acuerdo con el tamaño del lote de mensajes o la duración, procesando un flujo continuo de registros casi en tiempo real.

Novedades en Pentaho 8.1

- **Incremento de las capacidades de Spark en PDI**

Ahora puede ejecutar transformaciones en PDI con el motor Spark usando los siguientes pasos mejorados:

- Group By. Para conocer las diferencias al ejecutar este paso en Spark, consulte la sección en Use Group By with Spark.
- Unique Rows (Hashset). Utilice este paso con el motor de procesamiento Spark para ayudar a superar los problemas de restricción de memoria.
- Unique Rows.

Ejecutar sub-transformaciones con Spark. Ahora puede ejecutar sub-transformaciones con Spark en AEL usando el paso Transformation Executor, lo que le permite diseñar pipelines más complejas en PDI y ejecutarlas en Spark.

Spark History Server. Configure el registro de eventos de Spark para ser capturado y visualizado usando el servidor de historial de Spark. Consulte Configure Event Logging para obtener más información.

Novedades en Pentaho 8.1

- **Google Cloud Data Enhancements**

Ofrece la posibilidad de conectarse sin problemas a Google Cloud Storage mediante un navegador VFS para importar y exportar datos desde y hacia Google Drive. Con la adición de la nueva entrada de trabajo de Google BigQuery Loader, ahora puede usar BigQuery como fuente de datos con la consola de usuario Pentaho o el cliente PDI, configurar sus conexiones JDBC con un controlador Simba y crear canales ETL para acceder, enriquecer y almacenar datos con Google Cloud big data services.

Novedades en Pentaho 8.1

- **Mayor seguridad de AWS S3**

PDI ahora puede asumir permisos de rol de IAM y proporcionar acceso seguro de lectura / escritura a S3 sin la necesidad de proporcionar credenciales codificadas en cada paso. Esta flexibilidad adicional se adapta a los diferentes escenarios de seguridad de AWS para brindar una mejor experiencia de usuario debido a una menor carga de administración de credenciales, al tiempo que reduce el riesgo de seguridad resultante de las credenciales expuestas. Los pasos revisados de transformación S3 CSV Input y S3 File Output ahora permiten que PDI extraiga datos de Amazon Web Services con las mejoras de seguridad necesarias. Ambos pasos le permiten obtener sin problemas las claves de seguridad IAM de las variables de entorno, el directorio de inicio de su máquina o el perfil de instancia de EC2.

Novedades en Pentaho 8.1

- **Steps de Big Data nuevos y actualizados**

Se han agregado pasos de transformación de entrada y salida (ORC Input and Output) de registro optimizado (Optimized Record Columnal) para permitir que PDI realice el método de serialización de datos en columnas con indexación facilitando el desarrollo de pipelines que manejan estos formatos. El manejo nativo de los archivos ORC a través de los pasos de INPUT y OUTPUT está disponible desde cualquier sistema de almacenamiento estándar y también es accesible a través de los controladores del Sistema de archivos virtuales (VFS). Para mejorar el rendimiento, la ejecución nativa de los pasos puede ocurrir en el motor Pentaho o en Spark usando AEL

Novedades en Pentaho 8.1

- **Steps de Big Data nuevos y actualizados**

Nuevas opciones para ORC, Avro y Parquet. Se han agregado nuevas opciones de formato a los pasos de entrada y salida ORC, Avro y Parquet.

- Opción para agregar la fecha, la hora o una marca de tiempo para generar nombres de archivos.
- Sobrescribir archivos existentes.
- Conversión de tipo de datos, que permite cambiar los tipos de datos en cada uno de estos pasos.

Novedades en Pentaho 8.1

- **Actualizaciones adicionales de Big Data:**

Cassandra: Estos pasos se actualizan para admitir la versión 3.11 de Cassandra y la versión 5.1 de DataStax.

HBase. En los pasos Entrada de HBase y Salida de HBase, puede eliminar filas utilizando una clave de asignación. Una nueva opción le permite crear una plantilla de mapeo para extraer y escribir tuplas en y desde HBase.

MongoDB: Como una mejora de la seguridad, los pasos de MongoDB Input y MongoDB Output ahora admiten conexiones SSL. MongoDB también se ha actualizado al controlador 3.6.3, que admite las versiones 3.4 y 3.6.

Splunk: Actualizado a la versión 7.0 (PDI).

Novedades en Pentaho 8.1

- **Mejoras en la integración de datos**

New Data Lineage Analyzers. Data Lineage ahora cuenta con analizadores de entrada y salida JSON. (Contribute Additional Step and Job Entry Analyzers to the Pentaho Metaverse)

Compatibilidad con inyección de metadatos incorporada al paso **Table Input Step**. El campo de conexión en el paso **Table Input Step** ahora presenta la inyección de metadatos, es posible usar este paso para guardar las transformaciones inyectadas en el repositorio de Pentaho.

Generic Database Connection: Al configurar una base de datos genérica, puede usar la configuración Dialect para ayudarlo a definir un controlador JDBC personalizado y una URL para un dialecto de base de datos específico.

Nuevo filtro de selección en el Explorador de datos. Utilice el filtro Seleccionar para buscar una lista de valores para seleccionar como filtro mientras inspecciona sus datos dentro del cliente PDI.

Novedades en Pentaho 8.1

- **Mejoras en la integración de datos**

Borrar paso y búsqueda de entrada. En el panel Explorar del cliente PDI, ahora puede borrar su búsqueda de entrada de trabajo o paso de transformación actual haciendo clic en la "X" al lado del campo de búsqueda. Se agregó la capacidad de los administradores para eliminar el contenido de usuarios individuales.

Repositorio PDI mejorado. Ahora experimentará un rendimiento mejorado al abrir archivos, guardar archivos y explorar su repositorio Pentaho.

Pasos de transformación de Salesforce mejorados. PDI 8.1 utiliza la versión 41.0 de la API para la URL de WebServices en todos los pasos de Salesforce. Los siguientes pasos se actualizan ahora en PDI: Salesforce Input, Salesforce Insert, Salesforce Update, Salesforce Upsert, Salesforce Delete

Novedades en Pentaho 8.1

- **Mejoras en la integración de datos**

CSV File Step: Se mejoró el paso de transformación de la entrada de archivos CSV al agregar milisegundos en el campo de formato de fecha que permite al usuario controlar mejor el uso del manejo de archivos para controlar el número máximo de archivos abiertos simultáneamente y el tiempo entre descargas de archivos.

Mejoras de registro: Se incorpora el archivo PDI.log para capturar la ejecución de la transformación y los jobs. Además, ahora puede desplazarse a través de la salida del registro y copiar las secciones del texto del registro.

Novedades en Pentaho 8.1

- **Mejoras en la integración de datos**

Mejoras de permisos de administrador. Los administradores ahora pueden administrar mejor el contenido en Pentaho Repository Explorer. Cuando los usuarios individuales eliminan transformaciones, trabajos y conexiones de base de datos, los administradores pueden vaciar permanentemente sus carpetas de basura. Esta opción es útil cuando los usuarios dejan una organización y sus archivos eliminados deben borrarse permanentemente.

Novedades en Pentaho 8.1

- **Mejoras de Business Analytics**

Eje continuo para dimensiones de tiempo en visualizaciones. Las visualizaciones de gráficos de línea, área ahora usan una visualización continua de datos para la dimensión temporal. Los puntos de datos ahora son proporcionales a la duración del tiempo para una representación más precisa a nivel visual de las tendencias de datos. Anteriormente, el eje de tiempo utilizaba puntos de datos discretos espaciados por igual.

Timestamp en Reportes. Ahora puede agregar una marca de tiempo al contenido generado cuando ejecute un informe en segundo plano o programe un informe en la Consola de usuario. Los tipos de informes incluyen informes del diseñador de informes (.prpt), informes del analizador (.xanalyzer) e informes interactivos.

Novedades en Pentaho 8.1

- **Mejoras de Business Analytics**

Eje continuo para dimensiones de tiempo en visualizaciones. Las visualizaciones de gráficos de línea, área ahora usan una visualización continua de datos para la dimensión temporal. Los puntos de datos ahora son proporcionales a la duración del tiempo para una representación más precisa a nivel visual de las tendencias de datos. Anteriormente, el eje de tiempo utilizaba puntos de datos discretos espaciados por igual.

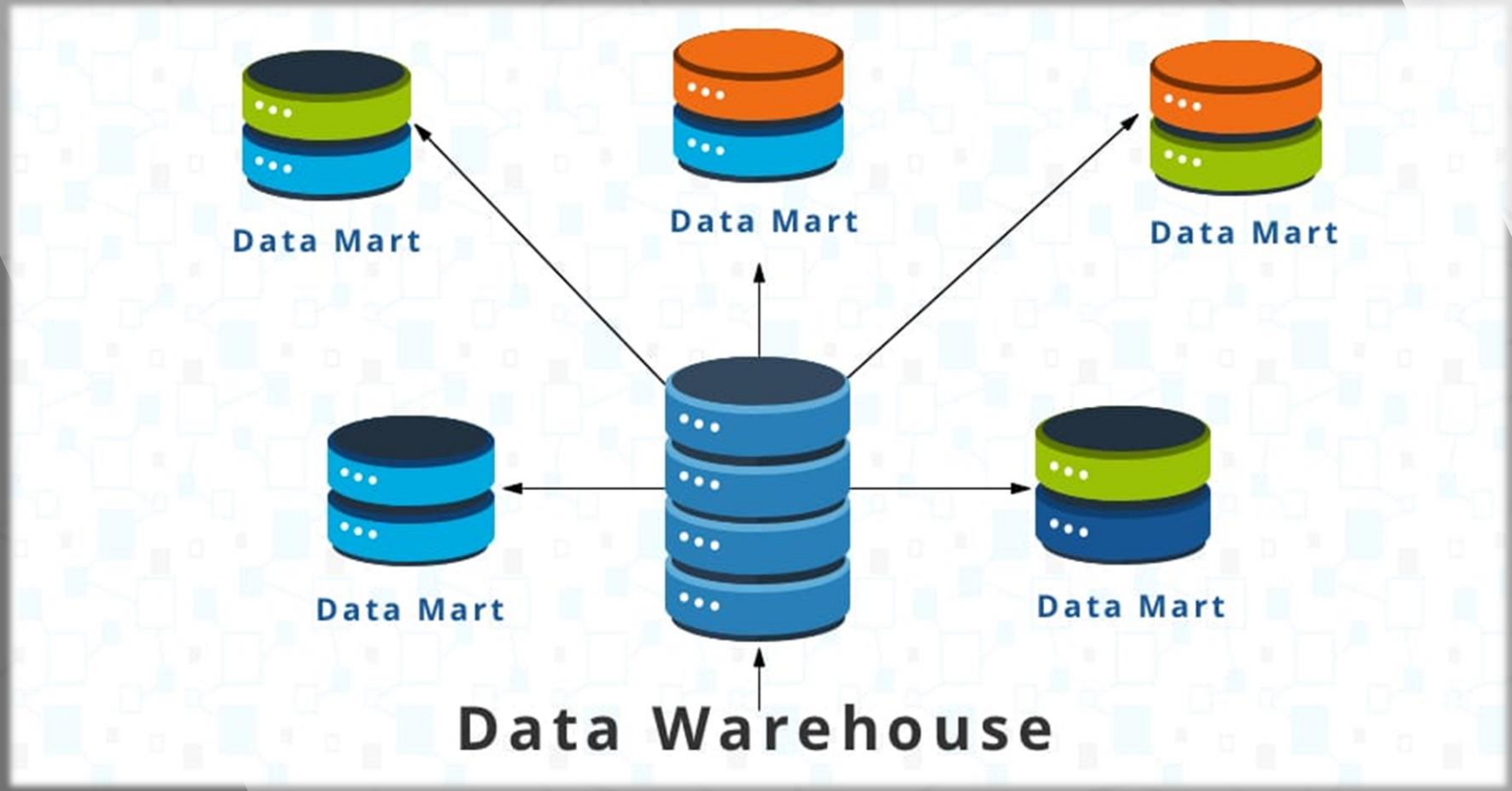
Timestamp en Reportes. Ahora puede agregar una marca de tiempo al contenido generado cuando ejecute un informe en segundo plano o programe un informe en la Consola de usuario. Los tipos de informes incluyen informes del diseñador de informes (.prpt), informes del analizador (.xanalyzer) e informes interactivos.

Data Mart vs Data Warehouse

Data Mart / Data Warehouse

Data Warehouse y Data Mart se usan como un repositorio de datos y sirven para el mismo propósito ¿Cuál es la herramienta más útil para las empresas? La respuesta a esta pregunta tal vez la tenga el 53 % de compañías que ya implementan una herramienta para el análisis de data desde el 2017. Entre estos dos términos hay que tener un especial cuidado, porque a veces se usan incorrectamente como sinónimos

Data Mart / Data Warehouse



¿Qué es Data Mart?

- Un **data mart** es una versión especial de almacén de datos (data warehouse). Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de múltiples formas para que diversos grupos de usuarios realicen la explotación de los mismos de la forma más conveniente según sus necesidades.
- El Data mart es un sistema orientado a la consulta, en el que se producen procesos batch de carga de datos (altas) con una frecuencia baja y conocida. Es consultado mediante herramientas **OLAP** (On line Analytical Processing - Procesamiento Analítico en Línea) que ofrecen una visión multidimensional de la información. Sobre estas bases de datos se pueden construir **EIS** (Executive Information Systems, Sistemas de Información para Directivos) y **DSS** (Decision Support Systems, Sistemas de Ayuda a la toma de Decisiones).
- En síntesis, se puede decir que los **data marts** son pequeños **data warehouse** centrados en un tema o un área de negocio específico dentro de una organización.

¿Qué es Data Mart?

En resumen, podríamos decir que los **Data Marts** son subconjuntos de datos de un **data warehouse** para áreas específicas. Entre las características de un **data mart** destacan:

- ✓ Usuarios limitados.
- ✓ Área específica.
- ✓ Tiene un propósito específico.
- ✓ Tiene una función de apoyo.

Data Mart

Al momento de crear un “Data Mart” de un área funcional de la empresa es preciso encontrar la estructura óptima para el análisis de su información, estructura que puede estar montada sobre una base de datos OLTP, como el propio datawarehouse, o sobre una base de datos OLAP. La designación de una u otra dependerá de los datos, los requisitos y las características específicas de cada departamento. De esta forma plantear plantear dos tipos de data marts:

- **Datamart OLAP**
- **Datamart OLTP**

Data Mart OLAP

Se basan en los populares cubos OLAP, que se construyen agregando, según los requisitos de cada área o departamento, las dimensiones y los indicadores necesarios de cada cubo relacional. El modo de creación, explotación y mantenimiento de los cubos OLAP es muy heterogéneo, en función de la herramienta final que se utilice.

Data Mart OLTP

Pueden basarse en un simple resumen del datawarehouse, no obstante, lo común es introducir mejoras en su rendimiento (las agregaciones y los filtrados suelen ser las operaciones más usuales) aprovechando las características particulares de cada área de la empresa. Las estructuras más comunes en este sentido son las tablas report, que vienen a ser *fact-tables* reducidas (que agregan las dimensiones oportunas), y las vistas materializadas, que se construyen con la misma estructura que las anteriores, pero con el objetivo de explotar la reescritura de queries (aunque sólo es posibles en algunos SGBD avanzados, como Oracle).

Data Mart OLTP

Los data marts que están dotados con estas estructuras óptimas de análisis presentan las siguientes ventajas:

- ✓ Poco volumen de datos
- ✓ Mayor rapidez de consulta
- ✓ Consultas SQL y/o MDX sencillas
- ✓ Validación directa de la información
- ✓ Facilidad para la historización de los datos

¿Qué es Data Warehouse?

- En el contexto de la informática, un **almacén de datos** (del inglés *data warehouse*) es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Se usa por reportajes y análisis de datos y se considera un componente fundamental de la inteligencia empresarial. Se trata, sobre todo, de un expediente completo de una organización, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos (especialmente OLAP, *procesamiento analítico en línea*). El almacenamiento de los datos no debe usarse con datos de uso actual. Los almacenes de datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas dependiendo del subsistema de la entidad del que procedan o para el que sean necesario.

Ventajas de un Data Warehouse

Algunas de las principales ventajas son:

- Los almacenes de datos hacen más fácil el acceso a una gran variedad de datos a los usuarios finales
- Facilitan el funcionamiento de las aplicaciones de los sistemas de apoyo como informes de tendencia, por ejemplo: obtener los ítems con la mayoría de las ventas en un área en particular dentro de los últimos dos años; informes de excepción, informes que muestran los resultados reales frente a los objetivos planteados a priori.
- Los almacenes de datos pueden trabajar en conjunto y por lo tanto, aumentar el valor operacional de las aplicaciones empresariales, en especial la gestión de relaciones con clientes.

Desventajas de un Data Warehouse

Algunas de las principales desventajas podemos encontrar:

- A lo largo de su vida los almacenes de datos pueden suponer altos costos. El almacén de datos no suele ser estático. Los costos de mantenimiento son elevados.
- Los almacenes de datos se pueden quedar obsoletos relativamente pronto.
- A veces, ante una petición de información estos devuelven una información subóptima, que también supone una pérdida para la organización.
- A menudo existe una delgada línea entre los almacenes de datos y los sistemas operacionales. Hay que determinar qué funcionalidades de estos se pueden aprovechar y cuáles se deben implementar en el data warehouse, resultaría costoso implementar operaciones no necesarias o dejar de implementar alguna que sí vaya a necesitarse.

Big Data

¿Qué es Big Data?

Para obtener una definición verdaderamente acertada de lo que significa Big Data debemos abrir la mente y romper con el estereotipo de que en “Big” esta la clave, ya que las exigencias actuales no siempre están basadas en el volumen, sino que éste es sólo una parte del rompecabezas, siendo muy diversos los parámetros que se tienen en cuenta en cada ocasión.

¿Qué es Big Data?

Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.

¿Qué es Big Data?

Aunque el tamaño utilizado para determinar si un conjunto de datos determinado se considera Big Data no está firmemente definido y sigue cambiando con el tiempo, **la mayoría de los analistas y profesionales actualmente se refieren a conjuntos de datos que van desde 30-50 Terabytes a varios Petabytes**.

¿Por qué el Big Data es tan importante?

- Lo que hace que Big Data sea tan útil para muchas empresas es el hecho de que proporciona respuestas a muchas preguntas que las empresas ni siquiera sabían que tenían. En otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.
- La recopilación de grandes cantidades de datos y la búsqueda de tendencias dentro de los datos permiten que las empresas se muevan mucho más rápidamente, sin problemas y de manera eficiente. También les permite eliminar las áreas problemáticas antes de que los problemas acaben con sus beneficios o su reputación.

¿Por qué el Big Data es tan importante?

El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más felices. Las empresas con más éxito con Big Data consiguen valor de las siguientes formas:

- Reducción de coste.
- Más rápido, mejor toma de decisiones (Hadoop)
- Nuevos productos y servicios.

¿Desafíos de la calidad de datos en Big Data?

Las especiales características del Big Data hacen que su calidad de datos se enfrente a múltiples desafíos. Se trata de las conocidas como 5 Vs: Volumen, Velocidad, Variedad, Veracidad y Valor, que definen la problemática del Big Data.

Estas 5 características del big data provocan que las empresas tengan **problemas para extraer datos reales y de alta calidad, de conjuntos de datos tan masivos, cambiantes y complicados**.

Algunos desafíos a los que se enfrenta la calidad de datos de Big Data son:

¿Desafíos de la calidad de datos en Big Data?

Muchas fuentes y tipos de datos

Con tantas fuentes, tipos de datos y estructuras complejas, la dificultad de integración de datos aumenta. Las fuentes de datos de big data son muy amplias:

- **Datos de internet y móviles.**
- **Datos de Internet de las Cosas.**
- **Datos sectoriales recopilados por empresas especializadas.**
- **Datos experimentales.**

Y los tipos de datos también lo son:

- **Tipos de datos no estructurados**: documentos, vídeos, audios, etc.
- **Tipos de datos semi-estructurados**: software, hojas de cálculo, informes.
- **Tipos de datos estructurados**

Solo el 20% de información es estructurada y eso puede provocar muchos errores si no acometemos un proyecto de calidad de datos.

¿Desafíos de la calidad de datos en Big Data?

Tremendo volumen de datos

- El volumen de datos es enorme, y eso complica la ejecución de un proceso de calidad de datos dentro de un tiempo razonable.
- Es difícil recolectar, limpiar, integrar y obtener datos de alta calidad de forma rápida. Se necesita mucho tiempo para transformar los tipos no estructurados en tipos estructurados y procesar esos datos.

¿Desafíos de la calidad de datos en Big Data?

Mucha volatilidad

- Los datos cambian rápidamente y eso hace que tengan una validez muy corta. Para solucionarlo necesitamos un poder de procesamiento muy alto.

- Si no lo hacemos bien, el procesamiento y análisis basado en estos datos puede producir conclusiones erróneas, que pueden llevar a cometer errores en la toma de decisiones.

¿Desafíos de la calidad de datos en Big Data?

No existen estándares de calidad de datos unificados

- En 1987 la Organización Internacional de Normalización (ISO) publicó las normas ISO 9000 para garantizar la calidad de productos y servicios. Sin embargo, el estudio de los estándares de calidad de los datos no comenzó hasta los años noventa, y **no fue hasta 2011 cuando ISO publicó las normas de calidad de datos ISO 8000**.
- Estas normas necesitan madurar y perfeccionarse. Además, **la investigación sobre la calidad de datos de big data ha comenzado hace poco** y no hay apenas resultados.
- La calidad de datos de big data es clave, no solo para poder obtener ventajas competitivas sino también impedir que **incurramos en graves errores estratégicos y operacionales basándonos en datos erróneos con consecuencias que pueden llegar a ser muy graves**.

Data Governance en Big data

Gobernabilidad significa asegurarse de que los datos estén autorizados, organizados y con los permisos de usuario necesarios en una base de datos, con el menor número posible de errores, manteniendo al mismo tiempo la privacidad y la seguridad.

Esto no parece un equilibrio fácil de conseguir, sobre todo cuando la realidad de dónde y cómo los datos se alojan y procesan está en constante movimiento.

A continuación veremos algunos pasos recomendados al crear un plan de Data Governance en Big Data.

Acceso y Autorización Granular a Datos

No se puede tener un gobierno de datos efectivo sin controles granulares.

Se pueden lograr estos controles granulares **a través de las expresiones de control de acceso**. Estas expresiones usan agrupación y lógica booleana para controlar el acceso y autorización de datos flexibles, con permisos basados en roles y configuraciones de visibilidad.

Seguridad perimetral, protección de datos y autenticación integrada

La gobernabilidad no ocurre sin una seguridad en el punto final de la cadena. Es importante **construir un buen perímetro y colocar un cortafuegos alrededor de los datos, integrados con los sistemas y estándares de autenticación existentes**. Cuando se trata de autenticación, es importante que las empresas se sincronicen con sistemas probados.

Encriptación y Tokenización de Datos

El siguiente paso después de proteger el perímetro y autenticar todo el acceso granular de datos que se está otorgando es **asegúrese de que los archivos y la información personalmente identificable (PII) estén encriptados y tokenizados de extremo a extremo del pipeline de datos**.

Constante Auditoría y Análisis

La estrategia no funciona sin una auditoría. Ese nivel de visibilidad y responsabilidad en cada paso del proceso es lo que permite a la TI "gobernar" los datos en lugar de simplemente establecer políticas y controles de acceso y esperar lo mejor. También es cómo las empresas pueden mantener sus estrategias actualizadas en un entorno en el que la forma en que vemos los datos y las tecnologías que utilizamos para administrarlos y analizarlos están cambiando cada día.

Una arquitectura de datos unificada

En última instancia, el responsable de TI que supervisar la estrategia de administración de datos empresariales, debe pensar en los detalles del acceso granular, la autenticación, la seguridad, el cifrado y la auditoría. Pero no debe detenerse ahí. Más bien debe pensar en cómo cada uno de estos componentes se integra en su arquitectura de datos global. También debe pensar en cómo esa infraestructura va a necesitar ser escalable y segura, desde la recolección de datos y almacenamiento hasta BI, analítica y otros servicios de terceros. La gobernanza de los datos es tanto acerca de repensar la estrategia y la ejecución como sobre la propia tecnología.

Consultas

EIDOS GROUP