**Panorama general de la analítica**

Save note

TranscriptNotesDownloadsDiscuss

Interactive Transcript - Enable basic transcript mode by pressing the escape key

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at ::12 and follow transcript0:12

Hola nuevamente. En este video les presentaré algo de historia

relacionada con sistemas de analítica para entender mejor el contexto de estos sistemas.

En la evolución de los sistemas de información,

vemos que los sistemas transaccionales fueron los protagonistas durante largo tiempo;

podríamos decir, las últimas dos décadas de los 90,

y siguen vigentes en la actualidad.

Este protagonismo se dio,

ya que las organizaciones requerían de sistemas que les

permitieran realizar sus procesos misionales, por ejemplo,

procesos de compras, prestación de servicio o manejo de inventarios,

nómina para garantizar su operación 7 por 24 es decir,

un soporte a la operación 24 horas del día durante 7 días por semana.

Es así como el contar con bases de datos relacionales

que facilitaron el manejo de datos estructurados,

en particular, el almacenamiento de los datos en tablas,

caracterizados por tener filas y columnas acompañadas de una serie de

restricciones para garantizar unicidad de los registros, registros no vacíos,

relación entre registros, entre otros,

al igual que proveer un lenguaje de consulta denominado SQL,

que viene del inglés Structured Query Language,

fueron clave en este desarrollo.

Esto, por supuesto,

acompañado del software especializado para su manipulación,

denominado "Sistema manejador de bases de datos",

donde podemos listar los líderes a lo largo de la historia: Oracle,

SQL Server, PostgreSQL, entre otros.

Estos sistemas, como su nombre lo indica,

facilitan la manipulación de los datos,

gracias a la funcionalidad que brindan para

crear bases de datos y usuarios que puedan accederlas,

tablas al interior de ellas,

para el caso de los manejadores de bases de datos relacionales,

con toda su estructura y las restricciones que se requieren definir,

la facilidad para hacer consultas sobre ellas,

haciendo uso de SQL,

y toda una capa de funcionalidad para el manejo de

transacciones al interior de los procesos que genera la información.

Es así como ofrecen mecanismos para garantizar las propiedades de ACID,

que significan "atomicidad",

"consistencia", "aislamiento" y "durabilidad",

propiedades fundamentales en el comportamiento de una transacción.

Veamos un ejemplo para aclarar estos conceptos.

En el sector del entretenimiento,

queremos tener un sistema para compra de boletos en un cinema.

Este sistema debe permitir consultar la disponibilidad de sillas para

una función determinada y estar en capacidad de comprar una o más sillas,

garantizando que el pago se hace de forma correcta,

y al momento de llegar a la función,

se tendrá la disponibilidad de la silla o sillas compradas.

Si desglosamos el proceso de compra,

que denominamos "transacción",

identificamos las siguientes partes:

1. Consultar sillas disponibles para una función particular, por ejemplo,

ver la película de "Batman" el domingo 6 de marzo del

2022 a las 6 p.m. en una sala específica.

2. Seleccionar dos sillas contiguas para la compra.

3. Comprar las sillas utilizando una tarjeta de débito.

Y 4, obtener los boletos para entrar a la función.

Fíjense que dentro de ese proceso,

esperamos que se realice toda la transacción o se anule por completo,

en caso de ocurrir alguna falla.

Si ocurre esto, no queremos, por ejemplo,

pagar las boletas y al final,

que no nos las entreguen.

El garantizar que eso no ocurre se conoce con el nombre de "Atomicidad",

se realizan todos los pasos de la transacción o ninguno.

Adicionalmente, esperamos que si consultamos nuestro saldo de la

cuenta bancaria antes de la compra y luego de la compra,

sin que realicemos ninguna transacción adicional en el intermedio,

nuestro saldo sea el que consultamos originalmente,

menos el valor que pagamos por las dos boletas, no otro.

El garantizar ese estado correcto de la información al

final de la transacción, lo denominamos "Consistencia".

Por otra parte, no quisiéramos que una vez pagadas las boletas,

el sistema nos dijera que las sillas ya no están disponibles porque otra

persona las seleccionó y las compró mientras yo hacia mi transacción.

La propiedad que va a garantizarnos que esto no ocurra se denomina

"Aislamiento". ¿Qué significa?

que vamos a tener la sensación de que somos los únicos comprando,

aunque haya otras personas comprando y, quizás,

interesadas en las mismas sillas que yo seleccioné.

Finalmente, esperamos que quede registro de mi compra y no se

pierda la información terminada la transacción,

lo que denominamos "Durabilidad de la información".

Garantizar estas propiedades de ACID, y en particular,

el procesamiento transaccional que se denomina "OLTP",

del inglés, "Online Transaction Processing",

no es algo trivial, por lo cual,

el contar con la funcionalidad brindada por el sistema manejador de bases de datos,

es fundamental, por un lado,

al igual que el uso apropiado de esta

funcionalidad por parte de la aplicación desarrollada,

y no menos importante,

el tener un modelo apropiado de las tablas y sus restricciones,

es un requisito para garantizar las propiedades vistas previamente y,

en particular, asegurar que el tiempo de respuesta a las transacciones sea el apropiado.

No debería superar,

en el caso de los cinemas, los segundos.

Es en este último elemento donde hablamos de "calidad del modelo de datos",

y esperamos que no ocurran compras sobre sillas que no

existen o compras con valores negativos o sin valores.

Sin embargo, esto puede ocurrir,

sea por problemas en el diseño de la base de

datos o en el desarrollo de este tipo de aplicaciones.

Esta habituación, o esta situación,

va a ser importante entenderla cuando hablemos de sistemas de analítica.

Una vez se tuvo algo de control en la parte operativa de las organizaciones,

lo que denominamos "el mundo transaccional",

se pasó a querer analizar la información y a utilizarla para

apoyar decisiones relacionadas con su estrategia, como, por ejemplo,

decidir dónde abrir una sala de cines,

ya que se tenía información de direcciones de residencia de

los clientes o decisiones a nivel de su operación,

como saber qué oferta realizará un cliente que está comprando sus boletos.

Es para dar respuesta a estas necesidades que surge la analítica,

la cual la podemos definir como "procesos sobre datos",

con el fin de encontrar hallazgos que sean útiles.

Piensen lo amplio de esta definición y la gran oportunidad que

tienen las organizaciones o empresas de aprovechar su información,

típicamente proveniente de sus sistemas transaccionales,

para combinarla con datos externos que pueden venir de entes

regulatorios o de ambientes completamente diferentes,

como las redes sociales,

y todo esto, para comprender,

por ejemplo, comportamientos de clientes,

productos o servicios,

y mantenerse competitivas en un ambiente como el actual.

Veamos la diferencia entre Sistemas transaccionales y Sistemas de analítica.

En este último, nos interesa más los comportamientos de grupos de clientes

o productos y no tanto una transacción o consulta puntual por fecha, cliente, etcétera.

Eso nos puede dar una idea de

los retos que debemos afrontar en el mundo de la analítica,

y la diferencia para la concepción y el desarrollo de este estilo de proyectos.

A manera de ejemplo,

tendremos retos relacionados con integración de fuentes,

la calidad de sus datos,

que como vimos en el ejemplo de la compra de boletos,

no está resuelto;

la comprensión del significado de los datos,

su almacenamiento para garantizar eficiencia en el acceso a la información;

el tener información histórica y, por supuesto,

su visualización para aportar en los procesos de toma de decisiones.

Con esta claridad, pasemos a ver cómo ha sido la

evolución a nivel de sistemas de analítica en las organizaciones.

Iniciamos hablando de la Analítica 1.0.

Esta analítica inicia a finales de los años 90,

y se caracteriza por hacer análisis

de lo ocurrido en una organización con el fin de entender mejor cómo son,

por ejemplo, sus clientes.

En esta etapa de la analítica se dio origen al

concepto de "bodega de datos" o "data warehouse",

en inglés, donde la idea es almacenar la información histórica,

y poder hacer análisis de tendencias,

de comportamientos en un repositorio o base de datos

corporativa con la información dispuesta para realizar análisis y,

de esta forma, monitorear el logro de los objetivos de una

organización definidos en su planeación estratégica.

Esta etapa de la analítica se caracteriza

porque la información que se toma para los análisis,

como la información de clientes que se tiene en los sistemas transaccionales,

se organiza y se analiza para lograr reaccionar a lo que está ocurriendo.

Sin embargo, si hay varias personas o dependencias de la misma organización

interesadas en análisis similares y no se tiene

una estrategia apropiada para el desarrollo de estos sistemas,

se puede replicar el trabajo y tener múltiples versiones de la verdad,

lo cual se va a evidenciar en comités ejecutivos al tratar

de comparar indicadores importantes que fueron calculados de forma independiente.

Estas múltiples versiones pueden darse por problemas en la calidad de la

información utilizada para construir los análisis.

En este tipo de analítica,

la visualización basada típicamente en tableros de control,

es fundamental, al igual que la generación de

reportes que contienen los análisis realizados.

Esta etapa, al igual que las siguientes,

no es obsoleta en la actualidad.

Es una fuente valiosa de análisis destinado

a entender y contestar preguntas más del estilo,

¿qué pasó?, ¿por qué pasó?

La siguiente etapa, conocida como Analítica 2.0 o analítica predictiva,

se caracteriza por responder a la pregunta, ¿qué pasará?

Esta etapa inicia hacia los años 2006, 2007,

gracias a la evolución tecnológica para el

almacenamiento masivo de datos y procesamiento distribuida,

dada por la creación de Hadoop,

un "framework" para el procesamiento distribuido de grandes volúmenes de datos.

Es en esta etapa,

donde ya es posible almacenar grandes cantidades de datos,

y requirió por parte de los líderes en manejo de datos,

como Yahoo, Google,

entre otros, propuestas para el procesamiento y,

en particular, el análisis predictivo de datos.

En esta época se destaca la aparición del término "científico de datos",

rol que en la actualidad se caracteriza por ser el

responsable de analizar datos con el uso de modelos

matemáticos para encontrar relaciones entre los

datos que son poco evidentes y lograr resolver preguntas como,

¿qué clientes van a desertar de mi organización el próximo mes?, o,

¿cuál será el consumo de datos que tendré en mi organización en los próximos seis meses?

Pasamos ahora a la Analítica 3.0,

analítica prescriptiva o analítica operacional,

que es una combinación de las analíticas 1.0 y 2.0,

en la cual el reto es poder utilizar la analítica,

no sólo en el apoyo de decisiones estratégicas,

sino de decisiones operativas.

Una de las aplicaciones emblemáticas de la época corresponde a la de Monsanto,

utilizada en el sector agrícola para definir en qué momento sembrar una semilla,

en qué momento debe regarse,

qué cantidad de agua se requiere y en qué momento recolectar el grano.

En este ejemplo vemos,

que se trata de tener una respuesta a una decisión operativa,

y eso es lo que inspira esta etapa,

dar respuesta a la pregunta, ¿qué debo hacer?

Y utilizarla a nivel de grandes y pequeños datos,

en todo el ciclo de la vida de un cliente,

proceso o producto,

ya que se abre la puerta para que las

organizaciones puedan competir a nivel de analítica.

Finalmente, llegamos a la etapa 4.0,

que aún es un referente al que se quiere llegar.

Se trata de la analítica autónoma,

donde el sueño es tener modelos matemáticos autónomos basados en técnicas de "machine

learning" para extraer el conocimiento de los datos

y resolver preguntas similares a la analítica predictiva.

Esto significa que la analítica,

se espera que no la realicen las personas,

sino las máquinas,

gracias a la inteligencia artificial y, en particular,

al uso de técnicas como el aprendizaje automático,

representado por el "deep learning" o aprendizaje profundo.

Este tipo de aprendizaje,

se espera que se realice a partir del procesamiento del lenguaje natural,

y su efecto podría estar relacionado con el recorte de personal estadístico

y científicos de datos para realizar estas tareas de construcción de modelos analíticos.

Aunque se espera reducción,

el mensaje no es que reemplace los analistas humanos;

está más orientado a la posibilidad de tener un mayor número de modelos

correctos disponibles para aportar al analista en el momento de la toma de decisiones.

Para finalizar, es importante recordar que en

este video hablamos de los sistemas transaccionales,

que son previos a la aparición de la analítica,

pero siguen vigentes en la actualidad,

ya que su rol es aportar en la parte operativa de las organizaciones,

y son los generadores de las fuentes primarias de

datos que se utilizan en los sistemas de analítica.

Adicionalmente, hablamos de cuatro etapas en la

evolución de la analítica: analítica descriptiva,

predictiva, prescriptiva y autónoma;

que no son excluyentes en su desarrollo,

lo que significa que podemos encontrar organizaciones

donde pueden estar presentes al mismo tiempo diferentes estilos de analítica.

Para terminar, te invito a reflexionar sobre estos temas y pensar en

las organizaciones con las que has interactuado para responder,

¿qué tipos de analítica les puede interesar y cuáles están trabajando en la actualidad?

Hasta pronto.

**Arquitectura de componentes para aplicaciones de analítica descriptiva**

Save note

TranscriptNotesDownloadsDiscuss

Interactive Transcript - Enable basic transcript mode by pressing the escape key

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at :: and follow transcript0:00

[MUSIC]

Play video starting at ::14 and follow transcript0:14

Hola a todos nuevamente.

Hoy voy a hablar sobre generalidades de la analítica 1.0.

La analítica 1.0, conocida como analítica descriptiva o inteligencia de negocios,

que permite entender qué pasó y por qué pasó.

Y es la más antiguas de las analíticas,

pero no por eso es obsoleta o no debe utilizarse en las organizaciones.

Todo lo contrario, es fundamental para hacer monitoreo de la estrategia de una

organización por medio de sus datos.

Por ejemplo, el analizar el comportamiento de ventas gracias a la integración

de la información existente en diferentes dependencias de una organización.

Como las áreas de servicio al cliente, mercadeo y ventas,

para definir ajustes que permitan a las organizaciones lograr sus objetivos.

Play video starting at :1:1 and follow transcript1:01

Es así como un gerente de ventas quisiera saber,

por ejemplo, cuántos clientes tuvieron en el trimestre anterior.

Y compararlo con el número de clientes del mismo trimestre del año anterior,

lo que corresponde a una métrica normal.

Adicionalmente, quiere saber si el valor de ventas alcanzado por esos clientes

supera la meta establecida por el área, se mantiene o, al contrario, va en descenso.

En este caso, el valor de ventas es considerado un KPI,

un indicador clave de desempeño, you que es una métrica con una meta asociada.

De igual manera, el gerente puede querer relacionar este KPI con el mismo del año

anterior, el trimestre anterior o el de la competencia.

Ese conjunto de indicadores es lo que sería ideal graficar en lo que denominamos

un tablero de control o dashboard.

El tablero de control es un mecanismo de visualización

que busca llevar a un tomador de decisiones a explorar datos.

Y validar o generar nuevas hipótesis.

Usualmente, se espera que el analista interactúe con este tipo

de aplicaciones que lo llevarán de lo general a lo particular.

Y le permitirán centrarse en los indicadores o métricas que tienen un

valor, por ejemplo, por debajo de su expectativa.

Sin embargo, en la actualidad se espera la generación de

alertas por parte de los sistemas de analítica.

Que puedan generar acciones, como enviar un correo electrónico con el reporte.

O la sección del tablero que evidencia un posible problema que debe ser

rápidamente revisado para evitar complicaciones en el corto plazo.

Play video starting at :2:42 and follow transcript2:42

Este es el estilo de aplicaciones que se construyen asociadas a esta analítica.

En la cual es importante tener claridad de quién la utilizará y el impacto

o beneficio que traerá para esa persona o rol.

Y por supuesto, para la organización.

Siguiendo con nuestro ejemplo anterior,

el rol que la utilizaría es el gerente de ventas.

Quien dentro de sus funciones tiene la responsabilidad de

motivar e incentivar a su equipo de trabajo para

obtener mejores resultados representados en el incremento de las ventas.

Es así como debe hacer un seguimiento por vendedor para ver cómo capacitar o

acompañar a los vendedores que no logran sus metas.

O generar incentivos para aquellos que la superan.

De ahí viene la necesidad de tener mecanismos fáciles para encontrar

esta información o detectar posibles problemas.

Y poder explorar mejor la información para comprobar sus hipótesis o descartarlas.

En este caso,

el poder detectar que no se logró la meta de ventas establecidas en un período.

Y los vendedores con más bajos indicadores podría ayudarlo a revisar el

comportamiento de dichos vendedores en los últimos meses para definir la acción

a tomar para mejorar sus ventas.

Por ejemplo, cambiarle el segmento de clientes que atiende, el producto,

enviarlo a capacitación.

O definir un plan puntual para dicho vendedor o grupo de vendedores con un

comportamiento similar.

Play video starting at :4:8 and follow transcript4:08

Este tipo de hallazgo puede llevarlo a hacer ajustes a nivel de la estrategia de

su área.

O a sugerir ajustes a nivel de toda la organización.

you sea porque el potencial es mayor al establecido en la meta.

O porque para lograr esa meta se debe previamente invertir en procesos o

formación de sus empleados.

El conocer el usuario que interactuará con la aplicación permite adicionalmente

definir mejor la representación a utilizar para mostrar la información.

Se prefieren representaciones simples, con las cuales el usuario esté familiarizado.

Por ejemplo, gráficas de barra, tarjetas con valores de los indicadores y

colores para representar el nivel de logro del indicador.

Play video starting at :4:55 and follow transcript4:55

Veamos ahora qué partes o componentes se requieren para la construcción de estos

tableros de control.

Inicialmente, se revisan las fuentes de datos a utilizar.

Estas fuentes de datos, recordemos que vienen típicamente de sistemas

transaccionales aunque no son los únicos sistemas que generan la información.

En nuestro ejemplo, vemos que esa información puede venir de un ERP,

del inglés enterprise resource planning.

Y puede combinarse con información de quejas

y reclamos que haya realizado el cliente.

Estas fuentes corresponden a la primera capa de una

arquitectura de analítica 1.0 basada en componentes.

Las cuales pueden ser internas o externas a la organización.

O estructuradas o no estructuradas,

como serían tweets o mensajes en redes sociales.

Esa información debe ser analizada por la persona responsable de construir la

aplicación, con el fin de entender mejor el proceso misional a través de los datos.

E identificar problemas de calidad que pueden afectar el resultado final.

Llevando a tener valores de los indicadores que no corresponden con la

realidad de la organización.

Y que son fácilmente detectables por los expertos del negocio.

Quienes al evidenciar estas fallas dudarán de la confiabilidad de la información

reportada por la aplicación.

Llevándolos a retomar sus métodos y fuentes de análisis tradicionales.

Los cuales pueden ser incompletos y también poco confiables,

pero que han sido utilizados por largo tiempo.

De ahí la responsabilidad de hacer análisis detallados de

la calidad de los datos.

Y revisarlos con los expertos del negocio con el fin de definir estrategias

para su corrección o manejo en el momento de presentar la información.

Esta actividad es una de las realizadas en la capa de área de staging por un

proceso denominado ETL del inglés extract, transform, load.

Esta área es un área temporal que se utiliza para almacenar rápidamente el

contenido de las fuentes de datos en una estructura similar,

you sea en una base de datos relacional, o en archivos.

Y llevarlas a nuevas estructuras más cercanas a las utilizadas en la bodega de

datos.

En esta área, se realizan labores de limpieza, consolidación, estandarización.

Y, en general,

se transforman los datos para ser enviados a la bodega o almacén de datos.

Esta área está oculta para el usuario final y es persistente hasta afinar el

proceso ETL.

Es decir, cuando acaba el proceso, estos datos desaparecen.

Play video starting at :7:36 and follow transcript7:36

La información transformada en las estructuras que facilitan el análisis y

con el nivel de calidad acordado de forma conjunta con los expertos de negocio,

será almacenada típicamente en una bodega o almacén de datos.

La construcción de bodegas de datos, o data warehouse,

ha sido de estudio desde finales de los 90.

Y vemos que en este proceso metodologías como la propuesta por el

Grupo Kimball denominada ciclo de vida de bodegas de datos,

han aportado en el desarrollo de este tipo de aplicaciones.

El objetivo de estos repositorios históricos de bases de datos relacionales

es consolidar la información de la organización para tener una única versión

de la verdad de los datos.

Aunque la filosofía de estos repositorios es la adecuada,

su construcción puede llevar a soluciones que tomen mucho tiempo antes de

tener el repositorio soñado a nivel corporativo.

Y que dada la dinámica de las organizaciones, puede ser difícil,

no solo de construir, sino de mantener.

Estos repositorios pueden ser físicos o lógicos.

Lo que significa una gran base de datos o la

unión de varias bases de datos dadas por departamentos o temas de análisis,

conocidas como almacenes de datos o data marts.

En la actualidad, vemos que términos como data lakehouse empieza a tomar fuerza.

Los cuales acompañan las bodegas de datos tradicionales con la capacidad de

almacenar las fuentes de datos en data lakes,

o lagos de datos, para luego ser transformados.

Lo que apalanca el concepto de los ELT, extract load transform,

un concepto de las bodegas de datos modernas.

Play video starting at :9:19 and follow transcript9:19

La estructura de los datos es fundamental para agilizar y facilitar la labor de un

análisis.

Es ahí donde se habla de modelos multidimensionales para representar la

información.

Esos modelos fueron concebidos con la idea de cubos,

donde cada lado representa una perspectiva de análisis, denominada dimensión.

Y una celda es la intersección de varias dimensiones y muestra la información

de una medida o métrica de interés.

Play video starting at :9:48 and follow transcript9:48

Si retomamos nuestro ejemplo de ventas, las dimensiones serían fecha,

vendedor y producto vendido.

Mientras que las medidas serían la cantidad y dinero asociado a las ventas.

Play video starting at :10:1 and follow transcript10:01

La siguiente capa la denominamos capa de entrega de información.

Y su rol es pensar en el desempeño en el momento de consultar los datos.

Es así como hay ocasiones donde se deben crear estructuras adicionales a las de la

bodega o almacenes de datos para facilitar y agilizar la consulta de la información.

Play video starting at :10:21 and follow transcript10:21

Es en esta capa donde podemos hablar de la creación de cubos OLAP,

Online Analytical Processing.

Para algunos temas de análisis donde se evidencian claramente los

conceptos que previamente les presenté de dimensiones y medidas.

Play video starting at :10:37 and follow transcript10:37

Terminamos con la capa de aplicación,

que fue donde iniciamos la descripción de esta arquitectura.

Es así como la unión de estos elementos ilustrados será lo que consideramos

arquitectura de componentes de analítica 1.0 o inteligencia de negocios.

Esta arquitectura podemos formalizarla resaltando los nombres de los componentes.

Y colocándola de forma vertical, como la veremos en otros recursos del curso.

O podemos verla de forma horizontal con un toque más moderno, donde logramos ver

la convivencia de las bodegas de datos estructuradas con los lagos de datos.

Y de esta forma, el manejo de información,

tanto estructurada como semi estructurada y no estructurada.

Al igual que el desarrollo de aplicaciones de analítica 1.0 hasta

analítica 4.0 utilizando fuentes de datos de diversos tamaños.

Llegamos al final, y es importante recordar que en el

video hablamos de analítica 1.0, conocida como inteligencia de negocios.

En la cual comprendemos el pasado y en particular qué y por qué pasó.

Ese tipo de analítica ayuda en el seguimiento de los objetivos de una

organización.

Y requiere de una estrategia para su construcción en capas.

Esas capas son fuentes de datos, área temporal o de staging,

almacenamiento, entrega de la información y aplicación.

Las cuales se utilizarán en el proceso de ETL.

Play video starting at :12:6 and follow transcript12:06

Las aplicaciones, por lo general,

son tableros de control orientadas a un rol dentro de una organización.

Y el cual debe tener impacto, gracias a su aporte, en la validación o generación

de hipótesis que conllevan a acciones o toma de decisiones.

Play video starting at :12:22 and follow transcript12:22

Para terminar,

te invito a reflexionar sobre la utilidad que este tipo de aplicaciones puede tener

para ti en la organización en la que trabajas o te gustaría trabajar.

Hasta pronto.

[MUSIC]

Arquitectura de componentes para aplicaciones de analítica descriptiva Save note Transcript Notes Downloads Discuss Interactive Transcript - Enable basic transcript mode by pressing the escape key You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key Play video starting at :: and follow transcript0:00 [MUSIC] Play video starting at ::14 and follow transcript0:14 Hola a todos nuevamente. Hoy voy a hablar sobre generalidades de la analítica 1.0. La analítica 1.0, conocida como analítica descriptiva o inteligencia de negocios, que permite entender qué pasó y por qué pasó. Y es la más antiguas de las analíticas, pero no por eso es obsoleta o no debe utilizarse en las organizaciones. Todo lo contrario, es fundamental para hacer monitoreo de la estrategia de una organización por medio de sus datos. Por ejemplo, el analizar el comportamiento de ventas gracias a la integración de la información existente en diferentes dependencias de una organización. Como las áreas de servicio al cliente, mercadeo y ventas, para definir ajustes que permitan a las organizaciones lograr sus objetivos. Play video starting at :1:1 and follow transcript1:01 Es así como un gerente de ventas quisiera saber, por ejemplo, cuántos clientes tuvieron en el trimestre anterior. Y compararlo con el número de clientes del mismo trimestre del año anterior, lo que corresponde a una métrica normal. Adicionalmente, quiere saber si el valor de ventas alcanzado por esos clientes supera la meta establecida por el área, se mantiene o, al contrario, va en descenso. En este caso, el valor de ventas es considerado un KPI, un indicador clave de desempeño, you que es una métrica con una meta asociada. De igual manera, el gerente puede querer relacionar este KPI con el mismo del año anterior, el trimestre anterior o el de la competencia. Ese conjunto de indicadores es lo que sería ideal graficar en lo que denominamos un tablero de control o dashboard. El tablero de control es un mecanismo de visualización que busca llevar a un tomador de decisiones a explorar datos. Y validar o generar nuevas hipótesis. Usualmente, se espera que el analista interactúe con este tipo de aplicaciones que lo llevarán de lo general a lo particular. Y le permitirán centrarse en los indicadores o métricas que tienen un valor, por ejemplo, por debajo de su expectativa. Sin embargo, en la actualidad se espera la generación de alertas por parte de los sistemas de analítica. Que puedan generar acciones, como enviar un correo electrónico con el reporte. O la sección del tablero que evidencia un posible problema que debe ser rápidamente revisado para evitar complicaciones en el corto plazo. Play video starting at :2:42 and follow transcript2:42 Este es el estilo de aplicaciones que se construyen asociadas a esta analítica. En la cual es importante tener claridad de quién la utilizará y el impacto o beneficio que traerá para esa persona o rol. Y por supuesto, para la organización. Siguiendo con nuestro ejemplo anterior, el rol que la utilizaría es el gerente de ventas. Quien dentro de sus funciones tiene la responsabilidad de motivar e incentivar a su equipo de trabajo para obtener mejores resultados representados en el incremento de las ventas. Es así como debe hacer un seguimiento por vendedor para ver cómo capacitar o acompañar a los vendedores que no logran sus metas. O generar incentivos para aquellos que la superan. De ahí viene la necesidad de tener mecanismos fáciles para encontrar esta información o detectar posibles problemas. Y poder explorar mejor la información para comprobar sus hipótesis o descartarlas. En este caso, el poder detectar que no se logró la meta de ventas establecidas en un período. Y los vendedores con más bajos indicadores podría ayudarlo a revisar el comportamiento de dichos vendedores en los últimos meses para definir la acción a tomar para mejorar sus ventas. Por ejemplo, cambiarle el segmento de clientes que atiende, el producto, enviarlo a capacitación. O definir un plan puntual para dicho vendedor o grupo de vendedores con un comportamiento similar. Play video starting at :4:8 and follow transcript4:08 Este tipo de hallazgo puede llevarlo a hacer ajustes a nivel de la estrategia de su área. O a sugerir ajustes a nivel de toda la organización. you sea porque el potencial es mayor al establecido en la meta. O porque para lograr esa meta se debe previamente invertir en procesos o formación de sus empleados. El conocer el usuario que interactuará con la aplicación permite adicionalmente definir mejor la representación a utilizar para mostrar la información. Se prefieren representaciones simples, con las cuales el usuario esté familiarizado. Por ejemplo, gráficas de barra, tarjetas con valores de los indicadores y colores para representar el nivel de logro del indicador. Play video starting at :4:55 and follow transcript4:55 Veamos ahora qué partes o componentes se requieren para la construcción de estos tableros de control. Inicialmente, se revisan las fuentes de datos a utilizar. Estas fuentes de datos, recordemos que vienen típicamente de sistemas transaccionales aunque no son los únicos sistemas que generan la información. En nuestro ejemplo, vemos que esa información puede venir de un ERP, del inglés enterprise resource planning. Y puede combinarse con información de quejas y reclamos que haya realizado el cliente. Estas fuentes corresponden a la primera capa de una arquitectura de analítica 1.0 basada en componentes. Las cuales pueden ser internas o externas a la organización. O estructuradas o no estructuradas, como serían tweets o mensajes en redes sociales. Esa información debe ser analizada por la persona responsable de construir la aplicación, con el fin de entender mejor el proceso misional a través de los datos. E identificar problemas de calidad que pueden afectar el resultado final. Llevando a tener valores de los indicadores que no corresponden con la realidad de la organización. Y que son fácilmente detectables por los expertos del negocio. Quienes al evidenciar estas fallas dudarán de la confiabilidad de la información reportada por la aplicación. Llevándolos a retomar sus métodos y fuentes de análisis tradicionales. Los cuales pueden ser incompletos y también poco confiables, pero que han sido utilizados por largo tiempo. De ahí la responsabilidad de hacer análisis detallados de la calidad de los datos. Y revisarlos con los expertos del negocio con el fin de definir estrategias para su corrección o manejo en el momento de presentar la información. Esta actividad es una de las realizadas en la capa de área de staging por un proceso denominado ETL del inglés extract, transform, load. Esta área es un área temporal que se utiliza para almacenar rápidamente el contenido de las fuentes de datos en una estructura similar, you sea en una base de datos relacional, o en archivos. Y llevarlas a nuevas estructuras más cercanas a las utilizadas en la bodega de datos. En esta área, se realizan labores de limpieza, consolidación, estandarización. Y, en general, se transforman los datos para ser enviados a la bodega o almacén de datos. Esta área está oculta para el usuario final y es persistente hasta afinar el proceso ETL. Es decir, cuando acaba el proceso, estos datos desaparecen. Play video starting at :7:36 and follow transcript7:36 La información transformada en las estructuras que facilitan el análisis y con el nivel de calidad acordado de forma conjunta con los expertos de negocio, será almacenada típicamente en una bodega o almacén de datos. La construcción de bodegas de datos, o data warehouse, ha sido de estudio desde finales de los 90. Y vemos que en este proceso metodologías como la propuesta por el Grupo Kimball denominada ciclo de vida de bodegas de datos, han aportado en el desarrollo de este tipo de aplicaciones. El objetivo de estos repositorios históricos de bases de datos relacionales es consolidar la información de la organización para tener una única versión de la verdad de los datos. Aunque la filosofía de estos repositorios es la adecuada, su construcción puede llevar a soluciones que tomen mucho tiempo antes de tener el repositorio soñado a nivel corporativo. Y que dada la dinámica de las organizaciones, puede ser difícil, no solo de construir, sino de mantener. Estos repositorios pueden ser físicos o lógicos. Lo que significa una gran base de datos o la unión de varias bases de datos dadas por departamentos o temas de análisis, conocidas como almacenes de datos o data marts. En la actualidad, vemos que términos como data lakehouse empieza a tomar fuerza. Los cuales acompañan las bodegas de datos tradicionales con la capacidad de almacenar las fuentes de datos en data lakes, o lagos de datos, para luego ser transformados. Lo que apalanca el concepto de los ELT, extract load transform, un concepto de las bodegas de datos modernas. Play video starting at :9:19 and follow transcript9:19 La estructura de los datos es fundamental para agilizar y facilitar la labor de un análisis. Es ahí donde se habla de modelos multidimensionales para representar la información. Esos modelos fueron concebidos con la idea de cubos, donde cada lado representa una perspectiva de análisis, denominada dimensión. Y una celda es la intersección de varias dimensiones y muestra la información de una medida o métrica de interés. Play video starting at :9:48 and follow transcript9:48 Si retomamos nuestro ejemplo de ventas, las dimensiones serían fecha, vendedor y producto vendido. Mientras que las medidas serían la cantidad y dinero asociado a las ventas. Play video starting at :10:1 and follow transcript10:01 La siguiente capa la denominamos capa de entrega de información. Y su rol es pensar en el desempeño en el momento de consultar los datos. Es así como hay ocasiones donde se deben crear estructuras adicionales a las de la bodega o almacenes de datos para facilitar y agilizar la consulta de la información. Play video starting at :10:21 and follow transcript10:21 Es en esta capa donde podemos hablar de la creación de cubos OLAP, Online Analytical Processing. Para algunos temas de análisis donde se evidencian claramente los conceptos que previamente les presenté de dimensiones y medidas. Play video starting at :10:37 and follow transcript10:37 Terminamos con la capa de aplicación, que fue donde iniciamos la descripción de esta arquitectura. Es así como la unión de estos elementos ilustrados será lo que consideramos arquitectura de componentes de analítica 1.0 o inteligencia de negocios. Esta arquitectura podemos formalizarla resaltando los nombres de los componentes. Y colocándola de forma vertical, como la veremos en otros recursos del curso. O podemos verla de forma horizontal con un toque más moderno, donde logramos ver la convivencia de las bodegas de datos estructuradas con los lagos de datos. Y de esta forma, el manejo de información, tanto estructurada como semi estructurada y no estructurada. Al igual que el desarrollo de aplicaciones de analítica 1.0 hasta analítica 4.0 utilizando fuentes de datos de diversos tamaños. Llegamos al final, y es importante recordar que en el video hablamos de analítica 1.0, conocida como inteligencia de negocios. En la cual comprendemos el pasado y en particular qué y por qué pasó. Ese tipo de analítica ayuda en el seguimiento de los objetivos de una organización. Y requiere de una estrategia para su construcción en capas. Esas capas son fuentes de datos, área temporal o de staging, almacenamiento, entrega de la información y aplicación. Las cuales se utilizarán en el proceso de ETL. Play video starting at :12:6 and follow transcript12:06 Las aplicaciones, por lo general, son tableros de control orientadas a un rol dentro de una organización. Y el cual debe tener impacto, gracias a su aporte, en la validación o generación de hipótesis que conllevan a acciones o toma de decisiones. Play video starting at :12:22 and follow transcript12:22 Para terminar, te invito a reflexionar sobre la utilidad que este tipo de aplicaciones puede tener para ti en la organización en la que trabajas o te gustaría trabajar. Hasta pronto. [MUSIC] Transcript language: Español : Added to Selection. Press [⌘ + S] to save as a note

Transcript language: Español

**Capa:** sinónimo de componente o parte de una arquitectura en este caso de solución de inteligencia de negocios.

**Data:** Materia prima de las bodegas de datos.

***Data marts*:** Almacenes de datos, típicamente relacionados con dependencias de la organización. La unión de estos *data mart* genera una bodega de datos.

***Data warehouse*:** Bodega de datos donde se acumulan datos para mantener su historia y de esta forma aportar en el análisis de información. Podemos hablar de un repositorio de datos el cual puede ser físico o lógico.

**ELT: Del inglés, *Extract, Load, and  Transform.*** Se refiere al proceso que extrae los datos de un Sistema primario y los carga en el sistema de bodega de datos para luego transformarlos y apoyar procesos de toma de decisión. Esta es una tendencia y está relacionada con el manejo de datos no estructurados.

**ETL: Del inglés, *Extract Transform and Load.*** Se refiere al proceso que extrae los datos de un Sistema primario, los transforma y los carga en el sistema de bodega de datos para apoyar procesos de toma de decisión.

***Hard business rules*:** reglas técnicas relacionadas con los dominios de datos. Afectan la forma de almacenar los datos, pero no su significado. Por ejemplo, el nombre del cliente debe ser una cadena de 80 caracteres. Se utilizan al extraer datos de una fuente o al llevarlas al área de *staging*.

**Modelo multidimensional*:*** Sinónimo de modelo dimensional, consiste en la representación de los datos en lo que se denomina un modelo conceptual, el cual facilita tareas de analítica.

***Soft business rules*:** reglas que afectan el dato mismo o su significado. Por ejemplo, la forma de manejar las edades en rangos o por ejemplo, al tener el mismo dato de un cliente como la dirección en diferentes sistemas, cuál es el que se debe tomar para almacenar en la bodega de datos**.**

***Staging area*:** área temporal donde se almacenan los datos en una estructura muy cercana a los datos de las fuentes originales.

**Repaso de conceptos de bases de datos relacionales y SQL**

Save note

TranscriptNotesDownloadsDiscuss

Interactive Transcript - Enable basic transcript mode by pressing the escape key

Save Note

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at :: and follow transcript0:00

[MUSIC]

Hola, en este recurso quiero hablarles sobre bases de datos relacionales.

Play video starting at ::15 and follow transcript0:15

Una base de datos relacional es un repositorio.

El cual utiliza como estructura de almacenamiento una serie de relaciones

que denominamos tablas.

Una tabla es un conjunto de filas y columnas que representan un concepto

de la vida real.

Por ejemplo, en las organizaciones podemos encontrar tablas de cliente,

productos, empleados.

Y cada una de las filas va a representar una instancia de la tabla.

Por ejemplo, un cliente, un producto, de acuerdo a la tabla que lo contenga.

Y va a tener como columnas las características del cliente, por ejemplo,

el nombre, el correo electrónico, el teléfono.

Play video starting at ::57 and follow transcript0:57

En este ejemplo vemos cómo tenemos la tabla de clientes.

Esta tabla tiene tres filas,

que representan tres clientes para la organización.

Y cada uno de ellos es descrito de acuerdo con su

identificador asignado a la organización.

Su nombre y su correo electrónico y su teléfono.

Play video starting at :1:15 and follow transcript1:15

Para manejar tablas y relaciones entre ellas se crearon los sistemas manejadores

de bases de datos.

En este caso, los relacionales como MySQL, Oracle, SQL Server, Postgre, entre otros.

Play video starting at :1:30 and follow transcript1:30

Un sistema manejador de bases de datos es un software que permite crear

diferentes tipos de objetos.

En nuestro caso, en un repositorio relacional tales como tablas,

relaciones entre tablas, restricciones, entre otros elementos.

Al igual que permiten la manipulación de dichos elementos.

Es así como se identifican como lenguajes,

el de definición o ddl, el de manipulación o dml.

Play video starting at :1:59 and follow transcript1:59

Empecemos con el de creación.

Play video starting at :2:2 and follow transcript2:02

Vemos acá en pantalla cómo tenemos una sentencia que nos permite crear una tabla.

Esa sentencia es el create table, acompañada del nombre de la tabla,

que en este caso es clientes.

Y damos la información de cada una de sus columnas, el ID, el nombre, etc.

Play video starting at :2:21 and follow transcript2:21

Adicionalmente, como también lo vemos en pantalla,

debemos asociar el tipo de dato de cada una de las columnas.

Play video starting at :2:30 and follow transcript2:30

A nivel de tipos de datos, en MySQL encontramos el INT,

que representa un número entero.

El VARCHAR, que representa una cadena de caracteres.

Donde especificamos la cantidad de caracteres máximos que queremos

almacenar en esa columna.

Play video starting at :2:48 and follow transcript2:48

Tenemos también el tipo fecha, que es DATE.

Si quisiéramos incluir a nivel de fecha la hora, tenemos el DATETIME.

Play video starting at :2:56 and follow transcript2:56

Si queremos manejar valores booleanos, tenemos BOOL.

O si queremos manejar números de punto flotante,

tenemos el FLLOT, entre los más utilizados.

Play video starting at :3:7 and follow transcript3:07

Por otro lado, dado que estas bases de datos fueron creadas principalmente para

soportar la operación de una organización..

Tenemos todo el manejo de restricciones de integridad.

En este caso, definir una llave primaria,

que se conocerá como la PRIMARY KEY de la relación.

Una llave foránea, que es una FOREIGN KEY de la relación.

Y podremos adicionalmente expresar cuando no queremos

permitir valores vacíos en nuestra columna con un not null.

O cuando no queremos tener valores repetidos en

una columna en las diferentes filas.

Por ejemplo, cuando definimos el identificador de un cliente que es

único al interior de la organización.

En este caso, podemos utilizar UNIQUE.

Play video starting at :3:53 and follow transcript3:53

En el ejemplo que vemos en pantalla vemos la sentencia para crear la tabla clientes.

La cual tiene un identificador, que es de tipo entero, y es la primary key.

Play video starting at :4:5 and follow transcript4:05

Adicionalmente, vemos que tanto nombre como correo electrónico y

teléfono no admiten valores nulos, por eso tenemos la palabra NOT NULL.

Y adicionalmente, el correo electrónico no debe ser el mismo entre

diferentes clientes, por lo cual le colocamos la restricción UNIQUE.

Play video starting at :4:25 and follow transcript4:25

Vemos también que asociado al ID del cliente,

que es la llave primaria, encontramos la palabra AUTO\_INCREMENT.

Esta opción nos indica la creación de una secuencia.

Un objeto en la base de datos que puede empezar en un valor 1 e irse

incrementando a medida de que lleguen nuevos clientes.

Play video starting at :4:48 and follow transcript4:48

Esta es una opción de creación de llaves primarias muy utilizada en el mundo de la

analítica.

Play video starting at :4:53 and follow transcript4:53

Sin embargo, recordemos que hay otra opción para crear llaves primarias,

que es la clásica del mundo transaccional.

Donde la organización le asigna el valor.

Play video starting at :5:3 and follow transcript5:03

Este es el caso de los identificadores a nivel de documentos de identidad para las

personas.

Fíjense que cada país tiene una regla de definición,

garantizando que dichos valores son únicos por persona.

Play video starting at :5:17 and follow transcript5:17

En este caso, en la creación de la sentencia,

se indicaría únicamente que el ID es primary key.

Y sabemos que al ingresar una nueva fila se debe indicar el valor en esa columna.

Play video starting at :5:31 and follow transcript5:31

En este ejemplo vemos el funcionamiento de las restricciones.

Si creamos una tabla sin restricciones, fíjense que podría tener nombre sin valor,

representados por NULL, al igual que teléfono sin valor.

Y vemos como en la parte derecha, dado que tenemos las restricciones,

ninguno de esos campos puede tener un valor.

Hablemos ahora de las llaves foráneas.

Play video starting at :5:56 and follow transcript5:56

Este concepto representa típicamente relaciones entre dos tablas,

con el fin de asegurar consistencia de los datos.

Es decir, como vemos en el ejemplo,

el poder garantizar que todo ID cliente registrado en la tabla de ventas.

Debe contener un identificador de cliente válido.

Play video starting at :6:16 and follow transcript6:16

Lo que significa un cliente que you fue registrado en la tabla de clientes.

Es así como en este ejemplo es posible tener ventas de los clientes 01 y 02,

porque existen en la tabla de clientes.

Play video starting at :6:32 and follow transcript6:32

Pasemos ahora a hablar del lenguaje DML.

Este lenguaje permite acceder y consultar los elementos que insertamos

previamente en las tablas creadas.

Y el resultado de una consulta SQL será una nueva tabla o relación.

Empecemos con el ejemplo más sencillo.

Queremos traer toda la información que ingresamos en la tabla de clientes.

Para ello hacemos un SELECT \* FROM clientes.

El asterisco representa las columnas que queremos consultar.

En este caso, todas las que tiene la tabla de clientes.

Es así como en el ejemplo vemos que la tabla tiene como atributos.

Un identificador, el nombre, correo electrónico,

teléfono y fecha de registro del cliente.

Y al hacer un SELECT \* FROM clientes, nos da como resultado una

nueva tabla con la misma información que la tabla original.

Play video starting at :7:30 and follow transcript7:30

Fíjate que ahora, si queremos traer solo las filas que cumplen una condición,

podemos expresarlo con el operador WHERE.

Así, en el ejemplo queremos traer únicamente el cliente cuyo

correo sea jp@ gmail.com.

De esta manera, cuando hacemos la consulta,

nos debe traer la fila relacionada con este cliente.

Play video starting at :7:54 and follow transcript7:54

Adicionalmente, vemos como a nivel de las columnas queremos traer solo el

nombre de este cliente.

Veamos el resultado.

Play video starting at :8:5 and follow transcript8:05

Fíjate que, a diferencia de la consulta anterior, el resultado es una

tabla con una sola fila y una sola columna, que corresponde al nombre.

Play video starting at :8:16 and follow transcript8:16

En general, lo que traemos después del SELECT vamos a denominarlo

una proyección sobre la totalidad de los atributos de la relación.

O relaciones que tenemos en el FROM.

Y lo que colocamos en el WHERE son condiciones que aplican a las

celdas para restringir el número de filas.

Lo que denominamos selecciones.

Play video starting at :8:38 and follow transcript8:38

Finalmente, después del FROM colocaremos los nombres de las tablas de las

cuales queremos traer la información.

Play video starting at :8:48 and follow transcript8:48

Las condiciones del WHERE podemos unirlas con operadores lógicos, como OR

o AND, como lo vemos en las consultas de estos ejemplos.

Play video starting at :8:58 and follow transcript8:58

Adicionalmente, podemos colocar comodines para representar el

contenido de una columna que no conocemos o que queremos generalizar.

En este caso, utilizamos el operador LIKE,

y como comodines podemos tener el porcentaje.

Que representa la existencia de 0 o más caracteres.

O el comodín Underscore, que representa la existencia de un carácter.

Play video starting at :9:24 and follow transcript9:24

Es así como en el ejemplo queremos consultar

los clientes que tienen identificador mayor que 1.

O cuyo nombre tiene 0 o más caracteres y termina en Pérez.

Play video starting at :9:39 and follow transcript9:39

Otro tipo de consulta común es sobre rangos de fechas,

ahí utilizaremos el operador BETWEEN.

Como lo vemos en este ejemplo.

Queremos traer todos los clientes que se registraron entre el primero de

junio del 2020 y el primero de junio del 2021.

Play video starting at :9:57 and follow transcript9:57

Es importante resaltar que dado que vamos a comparar fechas,

debemos convertir las cadenas de caracteres al tipo D.

Para lo cual utilizamos el formato correspondiente.

En este ejemplo, vemos que usamos %Y para representar cuatro dígitos del año.

Play video starting at :10:17 and follow transcript10:17

%n para representar dos dígitos del mes,

y %d para representar dos dígitos del día,

relacionado con la fecha.

Play video starting at :10:32 and follow transcript10:32

Otro concepto fundamental es el de JOIN.

Un operador que permite unir dos tablas, como lo veremos a continuación.

Play video starting at :10:40 and follow transcript10:40

Existen diferentes tipos de JOIN, el INNER JOIN, que es el más común.

El LEFT OUTER JOIN, que utilizaremos en el curso.

El RIGHT OUTER JOIN y el FULL OUTER JOIN.

El INNER JOIN representa la intersección que hay entre dos tablas.

Para esa intersección debemos especificar el campo que queremos que se intersecten.

Y el resultado va a ser todas las filas que tengan, en ese campo,

el mismo valor para la tabla, en este caso clientes, y para la tabla db.

[MUSIC]

En lenguaje natural, lo que quisiéramos expresar, por ejemplo, extraer

información de las ventas y el detalle de los clientes que las realizaron.

Play video starting at :11:26 and follow transcript11:26

Si tenemos las restricciones de llaves foráneas, esta consulta debería dar como

resultado todas las ventas con toda la información de los clientes.

Esto es, incluir el nombre, email, etc.

Play video starting at :11:41 and follow transcript11:41

Imaginemos ahora que queremos traer para los clientes registrados todas las compras

que realizaron el último mes.

Y en caso de no tener, queremos igual que muestren la información.

Con el fin, por ejemplo, de contactarlos y animarlos a realizar compras.

Play video starting at :11:58 and follow transcript11:58

Fíjate que en este caso la intersección arroja la primera parte

de lo que se quiere consultar.

Y faltaría incluir los clientes que no tienen ventas.

Es ahí donde tiene sentido utilizar un LEFT OUTEER JOIN.

El cual me da como resultado la intersección para las filas que están en

las dos tablas.

Y para los clientes que no están en la intersección, en este caso,

clientes representa la tabla izquierda de la consulta.

Play video starting at :12:28 and follow transcript12:28

Debo traerlos, y los datos relacionados con ventas vendrán vacíos,

es decir, con un valor de null.

Play video starting at :12:36 and follow transcript12:36

El comportamiento de un RIGHT OUTER JOIN es similar.

La gran diferencia es que la tabla que tomamos como referencia para

reportar todas sus filas es la derecha.

En este mismo ejemplo sería la de ventas.

Finalmente, el FULL OUTER JOIN me trae la intersección de las tablas.

Junto con las filas de la tabla de la izquierda y la derecha que no cumplen con

la intersección.

Play video starting at :13:4 and follow transcript13:04

Los demás detalles puedes consultarlos en la infografía de SQL que preparamos

para ti.

Play video starting at :13:9 and follow transcript13:09

Espero que esta introducción sobre bases de datos relacionales y SQL te permita

aclarar varias dudas.

Y consolidar las bases que requerimos para aprovechar al máximo el curso.

Hasta

pronto.

Arquitectura de componentes para proyectos de analítica 1.0 Elaborado por: María del Pilar Villamil y Daniel Galindo Ruiz ARQUITECTURA DE COMPONENTES STAGING AREA ALMACENAMIENTO ENTREGA DE INFORMACIÓN APLICACIÓN PROCESO ETL REPOSITORIO BASADO EN EL MODELO RELACIONAL Analítica 1.0 Inteligencia de negocios FUENTES DE DATOS Fuentes internas y externas Área temporal Procesos ETL - ELT Bodega de datos (física o lógica) Cubos OLAP (Online Analytical Processing) Tableros de control Análisis OLAP Modelos de analítica Extract - Transform- Load Conceptos de modelo relacional y SQL TENDENCIAS EN ARQUITECTURA Data LakeHouse INTERNAS Sistemas transaccionales u operativos de la organización. CRM "Customer Relationship Management" (Gestión de la Relación con el Cliente) Sensores. PQR (Peticiones, Quejas, Reclamos). FUENTES DE DATOS Existen diversas fuentes de donde se pueden obtener datos para realizar los análisis debidos. Particularmente nos interesan las fuentes internas y externas. Las fuentes externas son aquellas que no se hayan al directamente en las bases de datos de la organización. Las fuentes internas, por el contrario, son aquellas con las que ya cuenta la organización gracias al

# y ETL 05-2022 María del Pilar Villamil Giraldo Modelado de datos y ETL 3 Metodología para desarrollar aplicaciones de analítica 1.0 Adaptación realizada por María del Pilar Villamil, a partir de algunos capítulos de la bibliografía que se encuentra al final, así como documentos construidos con otros expertos, en especial se hace un reconocimiento a José Abásolo, líder en esta temática, y la experiencia en proyectos de analítica. I. Introducción En esta lectura se describe una adaptación de la metodología ASUM-DM “Analytics Solutions Unified Method”, la cual está orientada al desarrollo de proyectos de analítica predictiva bajo el supuesto de que el entregable es un modelo analítico. Es así como se incluye la construcción de una aplicación basada en tableros de control que permite hacer un análisis descriptivo de un conjunto de datos y en particular el monitoreo del cumplimiento de metas establecidas en un plan estratégico de una organización. Se presenta una primera sección con la descripción de la adaptación de la metodología. Se resaltan en la segunda sección, los entregables de cada etapa de la metodología y se finaliza, en la sección III, dando unas ideas muy generales de cómo utilizarla en el desarrollo ágil de este tipo de proyectos. I. Descripción de la metodología El desarrollo de proyectos de analítica 1.0 implica cuatro componentes clave: 1. Componente de software 2. Componente de almacenamiento de la bodega 3. Componente de procesamiento 4. Componente de almacenamiento de fuentes de datos Modelado de datos y ETL 4 El componente de software está relacionado con la aplicación desarrollada para la interacción con usuarios finales o el API que puede utilizar alguna aplicación dentro del contexto de la organización, para aportar en los procesos de toma de decisiones. Este componente de software para nuestro caso corresponde a tableros de control. El componente de almacenamiento corresponde a una bodega de datos estructurada siguiendo un modelo ROLAP, con esquemas estrella representados en tablas. El componente de procesamiento es el responsable de realizar todo el proceso de ETL. Finalmente, el componente de almacenamiento de fuentes será un sandbox que corresponde a una copia de la base de datos transaccional que es la fuente para la tarea de analítica a realizar. Cada uno de los componentes debe ser desarrollado con unas pautas globales al proyecto completo de analítica y propias al componente, que permitan cumplir unos mínimos de calidad. Es así como para el desarrollo de este tipo de aplicación se propone como metodología una extensión de ASUM-DM Analytics Solutions Unified Method, como se muestra en la Figura 1, para incluir el desarrollo de una herramienta basada en tableros de control y la incorporación de analítica 1.0 utilizando elementos de la metodología The Business Dimensional Lifecycle, de desarrollo de Sistemas de Inteligencia de Negocios propuesta por el grupo Kimball [1],[2] . Modelado de datos y ETL 5 Figura 1. Resumen de etapas operativas de la adaptación de ASUM-DM La metodología ASUM-DM sigue una serie de etapas, no necesariamente de forma secuencial, en las cuales se resalta en este documento su inicio con el entendimiento del negocio (etapa 1 en la figura), pasando por el entendimiento y preparación de los datos (fases 5 y 7) hasta llegar a la construcción y evaluación del modelo analítico (etapas 8 y 9). Esta metodología se extiende al desarrollo de soluciones de analítica 1.0, creando entre las etapas de entendimiento de datos y preparación de datos la definición de un producto de software (etapa 6), donde se diseña el componente y la arquitectura de software. Adicionalmente, a nivel de entendimiento del negocio se agregan elementos para refinar requerimientos analíticos, en el enfoque analítico se incluye como opciones de analítica, la descriptiva; en los temas de construcción y evaluación del modelo se incluye el desarrollo de soluciones de analítica descriptiva. A continuación, se presenta una descripción general de las etapas, haciendo énfasis en el aporte al desarrollo de aplicaciones de analítica 1.0. Modelado de datos y ETL 6 Etapa 1 – Entendimiento del negocio El objetivo de esta etapa es entender los objetivos y requerimientos desde la perspectiva del negocio, para luego convertirlos en un problema que se pueda abordar desde la analítica. En particular se debe obtener una visión amplia de la empresa donde se hará el desarrollo de la aplicación, desde el punto de vista del negocio y donde se tendrá como entregables adicionales, la identificación de usuarios interesados – stakeholders-,y el levantamiento, análisis y estructuración de requerimientos analíticos. Con la participación de los patrocinadores del proyecto, se identifican las áreas y personas clave, de negocio y tecnología, que serán entrevistadas para hacer el levantamiento de requerimientos analíticos y evaluar su factibilidad. Igualmente, se identifican documentos clave a analizar antes de proceder a las entrevistas. Ejemplos de estos documentos son: planes estratégicos, informes anuales, etc. De igual manera, el grupo consultor debe revisar proyectos similares en empresas del mismo sector tanto de nivel nacional como internacional, con el fin de preparar propuestas de analítica que podrá utilizar para refinar los análisis que van a identificar en las entrevistas. En las entrevistas, se busca identificar los temas analíticos más relevantes para la organización y para cada uno de ellos los análisis requeridos para abordar el tema analítico. Un tema analítico es un problema que se debe resolver o un objetivo que se quiere lograr. Ejemplo de problema a resolver podría ser, en el caso de un gerente de ventas, elaborar el plan de ventas para el próximo año. En el caso de objetivo a lograr, un ejemplo podría ser aumentar la base de clientes rentables un 5% por año en los próximos 3 años. El plan estratégico, si se tiene, podrá ser una fuente de temas analíticos. Algunos temas analíticos pueden ser muy gruesos y se deben descomponer en subtemas más simples. Por ejemplo, en el caso del objetivo aumentar la base de clientes rentables, cada una de las estrategias que se piensa implantar para lograr el objetivo, puede convertirse en un subtema analítico, como es el caso de conocer el perfil de los clientes rentables actuales, para buscar más personas con esos perfiles. Una vez identificado un tema analítico, las entrevistas deben indagar por los análisis requeridos o requerimientos analíticos para la solución del problema. En el caso de elaborar un plan de ventas, un análisis requerido podría ser analizar ventas históricas por región y línea de producto en un periodo de tiempo. En el caso de conocer el perfil de los clientes rentables actuales, un análisis requerido podría ser segmentar los clientes por diferentes variables para identificar los perfiles interesantes. Cada análisis requerido necesita datos. Hay que identificar cuáles son esos datos, sus fuentes potenciales y los procesos de negocio que los generan. En principio, cada proceso de negocio de la cadena de valor de la organización genera datos cada vez que se ejecuta. Esos procesos Modelado de datos y ETL 7 de negocio son soportados por aplicaciones específicas o por algún ERP. Esas aplicaciones o ERP tienen sus bases de datos, que se constituyen en las fuentes de los datos que se requieren para los análisis requeridos. Identificar esas fuentes y esos datos es parte de la labor de levantamiento de requerimientos que se hace con las entrevistas. El resultado puede ser evidenciar la necesidad de tener otros proveedores o fuentes de datos para los requerimientos analíticos identificados. En diferentes entrevistas pueden surgir temas analíticos comunes. Por lo cual, se requiere una labor de revisión para unificar los conceptos y las necesidades analíticas y de forma posterior priorizarlos para su desarrollo. El resultado de esta tarea de revisión se presenta en un entregable con el siguiente formato, en el cual, a modo de ilustración, se muestra un tema analítico con sus análisis requeridos, datos requeridos y fuentes potenciales (procesos de negocio). Modelado de datos y ETL 8 Tema analítico Análisis requeridos o inferidos/requerimientos analíticos Procesos de negocio Datos requeridos Inteligencia de Mercadeo: Mercadeo Relacional: Conocimiento del cliente Análisis de frecuencia de compra de los clientes en un periodo de tiempo, días transcurridos desde la última fecha de compra, monto percibido y un indicador que muestre si está por encima, en el promedio o por debajo del promedio en esas métricas. Este análisis se puede resolver con un Tablero de control. Manejo de datos maestros: órdenes (compras de clientes); atención al cliente (consultas acerca de productos o servicios, quejas, solicitud de asistencia, etc.) \* Circunstancias del cliente (edad, sexo, estado civil, ocupación, hobbies, número de hijos, etc.) \* Comportamiento del cliente (compras, pagos, contactos, etc.) \* Información derivada (Categoría del cliente, categoría de sus quejas) Análisis de un grupo de clientes donde se muestre sus ventas en un rango de fechas, en términos de cantidad de ventas, dinero percibido, productos adquiridos. Este análisis se puede resolver con un Análisis OLAP. Segmentación de clientes por circunstancias, comportamiento, información derivada (lifetime value, propensión a desertar, categoría de quejas). Este análisis se puede resolver utilizando aprendizaje automático. A medida que se van identificando nuevos procesos de negocio que generan datos necesarios para satisfacer los requerimientos analíticos, Modelado de datos y ETL 9 se van detallando en otro formato, que relaciona procesos de negocio con entidades de negocio y que conforma un segundo entregable de esta fase como el que se muestra en la figura 2, relacionado con el tema analítico de conocer al cliente. Los conceptos o entidades de negocio que maneja la organización tales como clientes, productos, proveedores, sucursales, cuentas, etc., también se identifican en las entrevistas. Dichas entidades de negocio, denominados Dimensiones, se van añadiendo al mismo formato utilizado en la figura 2 (las filas son los procesos de negocio y las dimensiones las columnas). En la celda correspondiente a la intersección de un proceso de negocio y una Dimensión se coloca una X si esa Dimensión aparece referenciada o interviene cada vez que se ejecuta el proceso de negocio. Es así como en la figura 2 se tiene que, para el proceso de órdenes de compra de los clientes, las dimensiones de tiempo, cliente, canal entre otras están presentes. Figura 2. Matriz de bus: Dimensiones de negocio identificadas vs. Fuentes/Procesos de negocio. Modelado de datos y ETL 10 La matriz resultante, denominada Matriz de bus, sirve para varios propósitos: • Mostrar las fuentes de información que hay que incorporar a la bodega de datos a corto, mediano o largo plazo, si se quieren satisfacer todos los requerimientos analíticos identificados. • Mostrar las entidades de negocio o dimensiones, comunes a varios procesos de negocio. Para evitar réplicas no controladas de las dimensiones, hay que saber la información que se comparte entre los procesos. La matriz muestra esto a nivel de Dimensiones. Dimensiones compartidas por muchos procesos, tienen alcance corporativo y podrían ser candidatas a ser datos maestros (datos transversales a la organización que tienen un manejo particular). Dimensiones manejadas por uno o pocos procesos, son responsabilidad local del área. • Mostrar las Dimensiones comunes para desarrollar la bodega de datos, permite reutilizar lo existente, en proyectos independientes, mantener la consistencia de los datos y agilizar la construcción de la bodega. En el anexo 1, se presenta una guía de la forma de realizar las entrevistas citadas en esta etapa. Etapa 2. Enfoque analítico En esta etapa, es necesario traducir los objetivos del negocio en metas de analítica. Esto significa darle un enfoque descriptivo, de aprendizaje automático, o uso de optimización, entre otros, para lo cual se debe hablar en términos de tareas de analítica como describir, predecir, segmentar, clasificar, encontrar relaciones entre los datos, entre otros. Etapas 3 Requerimientos de datos En esta etapa se definen los datos que se van a utilizar en el proyecto, es así como se identifican proveedores de datos y fuentes de datos candidatas para luego a partir de esa información seleccionar las que se utilizarán en el proyecto. Etapas 4. Recolección de datos Esta etapa se encarga de acceder a los datos de las fuentes internas y externas a la organización y dejarlos disponibles para el proyecto. Etapa 5. Entendimiento de datos Esta etapa es clave ya que en ella se revisa detenidamente la información compartida para el proyecto, la cual será usada para realizar las tareas Modelado de datos y ETL 11 de analítica requeridas. En esta etapa se accede a los datos y se exploran antes de prepararlos. Un buen desarrollo de esta etapa ahorra problemas en el futuro. El principal objetivo es entender la organización a través de sus datos y familiarizarnos con ellos para la oportunidad analítica que tenemos en mente o para identificar nuevas oportunidades de analítica. De igual manera permite: • Determinar si los datos brindados son apropiados o no para la tarea de analítica que tengo como objetivo o si se requiere más información para entenderlos. • Identificar las transformaciones que requieren para realizar las tareas de analítica objetivo del proyecto. • Describir el proceso que generó los datos a partir de los datos mismos. • Entender la diferencia entre los datos esperados y la realidad de los mismos. • Identificar el nivel de la calidad de los datos gracias al análisis, interpretación y validación de reglas y formatos que deben cumplir los datos de una fuente específica. Es así como se utiliza el perfilamiento de datos para poder describir los datos y adicionalmente, se realiza un análisis de calidad de datos para entender la realidad de los mismos e identificar problemas y opciones para corregirlos. Estas opciones deben ser analizadas de forma conjunta con las personas del proyecto para llegar a acuerdos de transformaciones a realizar o cambios a nivel de la fuente original, que conllevan a nuevas entregas de datos o a la interacción con nuevos actores relacionados con los datos fuente que se tienen. Este entendimiento se apoya en los siguientes análisis: Nivel general • Consultar el número de registros y de atributos. • Determinar el atributo que identifica de forma única cada registro de la fuente de datos. • Visualizar una muestra de los datos. • Determinar el tipo de datos de los atributos, el formato y el dominio de los mismos. Modelado de datos y ETL 12 A nivel de datos numéricos • Describir distribuciones y patrones de los atributos con el uso de estadística descriptiva, (i.e., medidas de tendencia central y dispersión). • Identificar valores atípicos, utilizando estadísticos como el mínimo, el máximo, y la media. A nivel de datos categóricos • Identificar categorías de las variables y frecuencia de las mismas. • Relaciones entre atributos Dependencia entre ellos. Para identificar redundancia entre los atributos y posibles inconsistencias. A nivel de dimensiones de calidad de datos • Completitud • Unicidad • Consistencia: formato y semántica • Validez Etapa 6. Definición del producto de software En esta etapa se hace la definición y el diseño del componente de software que se usará como producto final para presentar a los usuarios. Estos componentes, típicamente tableros de control, se diseñan en función de los objetivos de analítica definidos en las etapas anteriores. En esta etapa, también se obtienen los requerimientos de la aplicación, se diseña la arquitectura de software que se utilizará en el proceso ETL y los mock-ups asociados al tablero o tableros de control. En este punto el contar con metodologías de analítica visual como el framework de Tamara Munzner [4] puede aportar en el desarrollo de este componente. La figura 3. muestra una posible arquitectura, con el fin de entender mejor el entregable de esta etapa. Modelado de datos y ETL 13 Figura 3. Arquitectura de proyectos de analítica 1.0 La arquitectura se define en términos de componentes, como se presenta a continuación. 1. El componente de fuentes de datos está integrado por aquellas fuentes que almacenan las transacciones y datos operacionales de organización, al igual que datos que provienen de entes de control. 2. El área temporal (staging area) contiene datos relacionados con procesos temporales relacionados con el ETL (Extracción, Transformación y Carga de datos) 3. Almacenamiento corresponde a la bodega de datos que tendrá un modelo multidimensional representado en tablas (ROLAP) para representar los datos estructurados, la cual puede estar acompañada de un lago de datos (data lake) para incluir datos semi-estructurados o no estructurados. 4. Entrega de información es un área adicional de almacenamiento que toma los datos de la bodega de datos y los almacena de forma cercana a los análisis realizados por usuarios finales, con el fin de mejorar el desempeño de las aplicaciones. 5. Aplicación capa compuesta por las aplicaciones que están a disposición de los usuarios finales, como los tableros de control. Los mock-ups o bocetos son una representación gráfica de la aplicación a desarrollar que permite al usuario final comprender el resultado de los análisis y de esta forma validarlo. A continuación, en la figura 4, se visualiza un ejemplo: Modelado de datos y ETL 14 Figura 4. Mock-up de un tablero de control- Figura tomada de: https://figmatemplate.com/insurance-dashboard-figma-template/ Etapa 7. Preparación de los datos Esta etapa es por lo general la más larga del proyecto de analítica ya que, en la mayoría de las organizaciones, los datos no tienen el nivel de calidad requerida para este objetivo. Además, es posible que se hayan pasado algunos problemas por alto en el entendimiento del negocio y entendimiento de datos, lo que puede ocasionar inconvenientes en esta etapa que afecten el desarrollo del proyecto. Adicional a los procesos de limpieza de datos relacionados con problemas de calidad de los mismos, se tiene la integración de fuentes cuando hay varias involucradas en la analítica a realizar, al igual que el diseño e implementación del proceso ETL (del inglés Extract Transform Load) para automatizar el proceso de analítica y la fase del diseño de la bodega de datos para analítica 1.0, en caso de que no exista, lo que corresponde a realizar el modelo multidimensional que representa los procesos de negocio a analizar y la creación de las tablas, si se trata de un modelo ROLAP. Modelado de datos y ETL 15 A nivel del proceso ETL esta etapa brinda claridad sobre las transformaciones requeridas a nivel de los datos (que ya se conocen), y una vez construido o conocido el modelo de la bodega de datos, se podrá terminar el diseño del ETL y proceder a implementar el proceso. A manera de ejemplo, a continuación, la figura 5 presenta un bloque que representa parte de un diseño de un proceso ETL. Figura 5. Diseño de un bloque de un proceso de ETL En este ejemplo, se toma como base la tabla empleados de la bodega de datos y se muestra las columnas de la fuente original, las transformaciones identificadas y las columnas de la tabla en la bodega de datos. A nivel de implementación podemos ver que los procesos de ETL, se transforman en sentencias SQL como se muestra en la figura 6. Modelado de datos y ETL 16 Figura 6. Implementación de un bloque de un proceso de ETL Los modelos multidimensionales corresponden a un modelo conceptual que permite responder a análisis requeridos. En estos modelos se identifican, los hechos que corresponden a la tabla central y las dimensiones que son las perspectivas de análisis. A manera de ejemplo, que se muestra en la figura 7, se toma el proceso de ventas de una empresa, donde el hecho es la venta con las medidas de unidades y valor de la venta y las dimensiones el cliente que hizo la compra, el producto que compró, la fecha y la sucursal en que la hizo, y el vendedor que atendió al cliente. Modelado de datos y ETL 17 Figura 7. Modelo multidimensional proceso de ventas Los entregables de esta fase incluyen diagramas de diseño del ETL ajustados, del modelo dimensional, al igual que la documentación del proceso de ETL y su implementación. Etapa 8. Construcción del modelo Esta etapa tiene como fin construir componentes analíticos que permitan cumplir los objetivos analíticos propuestos previamente. Esto se realiza de forma iterativa y puede requerir volver a la etapa de preparación de datos, para mejorar el resultado. Es en este punto donde se deben definir las técnicas de analítica como uso de estadística para el caso de analítica 1.0, o basadas en aprendizaje automático y relacionadas con los tipos de analítica predictiva o prescriptiva. De igual forma, incluye el diseño de pruebas de calidad, el cual para los modelos descriptivos basados en estadística descriptiva pueden ser validados utilizando técnicas de interacción hombre-máquina y trayendo elementos de la analítica visual. Mientras que otros modelos podrán ser validados adicionalmente desde el punto de vista cuantitativo, utilizando métricas como coeficiente de silueta para modelos basados en clústering particional, precision y recall, para modelos de clasificación y RMSE, MAPE para el caso de modelos de regresión. Finalmente, está la construcción del modelo analítico, que para el caso de analítica 1.0 consiste en la creación de los tableros de control previamente diseñados y validados por los usuarios finales, mientras que para los otros casos podrá relacionarse con el uso de algoritmos específicos asociados a las tareas de aprendizaje de segmentación, clasificación, regresión entre otros. Los entregables de esta fase incluyen la documentación e implementación del tablero de control. Modelado de datos y ETL 18 Etapa 9. Evaluación del modelo Esta etapa consiste en evaluar si los componentes analíticos generados y probados en la etapa anterior son apropiados para el negocio, teniendo en cuenta los criterios de éxito definidos por éste previamente, lo que incluye verificar la identificación de hallazgos de interés para el negocio. Etapa 10. Despliegue de la solución En esta etapa se lleve al ambiente de producción el resultado del ejercicio de analítica, con el fin de que el usuario final tenga acceso a él. Este acceso puede ser a parir de una aplicación Web, móvil o un API para que otras aplicaciones utilicen el resultado del ejercicio de analítica. Etapa 11. Retroalimentación En esta etapa se revisa con detalle lo realizado para entenderlo y si es necesario realizar ajustes. Aunque está en la etapa 11 no se sugiere que se haga únicamente al final del proyecto. La sugerencia en este punto es que esta retroalimentación sea continua y esté presente en las diferentes etapas. El realizarlo de esa manera minimiza riesgos de introducir errores en las actividades realizadas, tomar decisiones que no sean acordes al contexto de la organización y de tener un producto que no sea el esperado por el usuario final. II. Entregables de la metodología Como resumen de la metodología, a continuación, se presentan los entregables para proyectos de analítica 1.0 relacionados con las etapas presentadas: Etapa 1 – Entendimiento de negocio. E1.1 (Entregables-Etapa1): • Temas analíticos, análisis requeridos y fuentes de datos, consolidados. •Matriz de bus: Procesos de negocio vs Dimensiones de negocio identificadas • Resúmenes de las entrevistas realizadas. Etapa 6 – Definición del producto de software Modelado de datos y ETL 19 E6.1 Arquitectura: Componentes, descripción de fuentes de datos y de flujos de información. E6.2 Mock-ups: Representación visual de los tableros de control a desarrollar Etapa 7 – Preparación de datos E7.1 Diagrama del modelo de datos multidimensional E7.2 Bodega de datos inicial que responde a los requerimientos analíticos previamente seleccionados E7.3 Diseño e implementación de procesos de extracción, transformación y carga de datos Etapa 8 – Construcción del modelo E8.1 Implementación de los tableros de control III. Relación entre la metodología y el agilismo Uno de los retos de la metodología presentada es su uso para guiar el desarrollo de proyectos. Es en este punto donde ideas del agilismo [3] como el desarrollo por iteraciones para tener entregables desde las primeras iteraciones aportan. A manera de ejemplo, se pueden identificar iteraciones de dos semanas de duración con las etapas de la metodología, en las cuales se espera la interacción con los interesados en el proyecto para efectos de validación: Iteración 1 • Entendimiento del negocio – refinamiento de requerimientos analíticos • Entendimiento de datos – subconjunto de fuentes de datos de acuerdo con el orden seleccionado para desarrollar los requerimientos analíticos • Entendimiento del negocio – Comprensión de la infraestructura tecnológica Iteración 2 • Entendimiento de datos – subconjunto de fuentes de datos de acuerdo con el orden seleccionado para desarrollar los requerimientos analíticos • Diseño del componente de software para unos primeros análisis requeridos Modelado de datos y ETL 20 • Diseño del componente de software - Acceso a la infraestructura tecnológica Iteración 3 • Entendimiento del negocio– refinamiento de requerimientos analíticos • Preparación de datos con una propuesta de modelo de datos dimensional Iteración 4 • Entendimiento de los datos relacionados con nuevos análisis • Preparación de datos con una propuesta refinada de modelo de datos dimensional y el diseño del proceso ETL • Construcción de la base de datos que representa el modelo dimensional con la implementación de los primeros tableros Finalmente se debe tener en cuenta que para el éxito del proyecto se requiere: 1. Intervención de usuarios de nivel ejecutivo de la organización para aportar en las decisiones importantes del proyecto 2. Alta motivación y disponibilidad de los beneficiarios del Sistema a construir que van a participar en el proyecto. 3. Acceso oportuno a las fuentes de datos actuales y a sus responsables. 4. Conocimiento de la infraestructura tecnológica de la empresa para alinear el proyecto con su arquitectura de datos y de infraestructura. Anexo 1. Resumen de las entrevistas realizadas Como anexo de los entregables, se entrega el material base utilizado para su elaboración, principalmente los resúmenes de las entrevistas individuales realizadas. Dichos resúmenes deben ser aprobados por los entrevistados al final o poco tiempo después de la entrevista. A continuación, se muestra un ejemplo de libreto o cuestionario con estructura y temáticas ejemplo para el entrevistado que recomienda la metodología del grupo Kimball. Modelado de datos y ETL 21 A. Introducción • Presentación del entrevistador • Objetivo de la entrevista • Puntos a tratar • Confirma el tiempo disponible • Descripción de los pasos a seguir después de la entrevista B. Responsabilidad (Relación de las responsabilidades del entrevistado con el proyecto y esto con los objetivos de la empresa a la que pertenece) • Descripción de los objetivos de la empresa relacionados con su cargo • Descripción de la principal responsabilidad del entrevistado C. Objetivos del negocio y problemas (Relación del proyecto con las expectativas de la empresa y problemas que han tenido para para resolverlos) • Dentro de sus responsabilidades la que está relacionada con uno de los objetivos de la empresa • … • Análisis requeridos (Los tipos de análisis que hace actualmente y nuevos análisis que puedan proponer) • Datos utilizados en los análisis y sus características • ¿Qué oportunidades de mejora tiene la empresa desde su rol y los análisis que realiza? • Recapitulación y síntesis • Resumen de los hallazgos más importantes durante la entrevista • Cierre de la entrevista **Uso de la infografía entendiendo los datos**

Save note

TranscriptNotesDownloadsDiscuss

Interactive Transcript - Enable basic transcript mode by pressing the escape key

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at :: and follow transcript0:00

[MUSIC]

Hola nuevamente.

Quiero presentarte este recurso de entendimiento de datos.

Recordemos que esta actividad la hacemos con el objetivo de determinar si los

datos que nos dan son apropiados o no.

Para la tarea analítica que tenemos que desarrollar.

Play video starting at ::26 and follow transcript0:26

Adicionalmente, una vez sabemos que son los datos apropiados,

podemos determinar cuáles son las transformaciones.

Y toda la preparación que requerimos para que los datos que vamos a utilizar en esta

tarea de análítica sean los adecuados.

Play video starting at ::43 and follow transcript0:43

Una vez entendemos el objetivo, podremos venir acá a mirar algunos de sus

beneficios, de los beneficios más importantes.

Como puede ser el hecho de familiarizarnos con los datos,

que son la materia prima de nuestro proyecto.

O, por ejemplo, poder validar el proceso que están generando los datos y,

de esta manera, comunicárselo al negocio.

Play video starting at :1:4 and follow transcript1:04

También podremos encontrar un conjunto de actividades que pueden guiar esta tarea.

Y a nivel de actividades, por ejemplo,

podremos ver que esas actividades están relacionadas con preguntas, ¿no?

La primera de ellas sería, bueno,

¿qué significa un registro de la fuente de datos que nos dieron?

Para eso podríamos consultar el diccionario de datos compartidos,

comprender un poco mejor el contexto del negocio.

Vamos a ver, por ejemplo, cuáles son las fuentes de datos.

De igual manera, para cada una de estas preguntas vamos a encontrar una serie

de pasos que pueden ayudarnos a entender mejor este proceso.

[MUSIC]

Relacionado con actividades, tendremos otras adicionales.

Por ejemplo, podemos decir,

¿cuáles son esas características generales de los datos?

¿Cómo lo haríamos?

Consultemos el número de registros y de atributos.

Determinemos el atributo que identifica de forma única cada registro de la fuente

de datos.

En algunos conceptos lo conocíamos como una llave primaria.

Visualizar una muestra de datos.

Determinar el tipo de datos de los atributos,

el formato y el dominio de los mismos.

De igual manera que al final podríamos decir, por ejemplo,

qué oportunidades nuevas de analítica podremos identificar en futuros proyectos.

Buscar patrones o relaciones entre los atributos de los datos,

visualizar correlaciones, entre otros.

Acá podemos you regresar a la página principal.

Fíjense que podemos revisar adicionalmente algunas recomendaciones que les damos.

Estas recomendaciones las pueden ver antes de empezar.

¿Cómo nos preparamos?

Aquí, por ejemplo, es importantísimo saber para qué estamos haciendo esta actividad.

En el durante necesitamos entender bien cuál fue ese proceso de negocios que

generó los datos.

Y una cosa muy importante acá es evitar tomar decisiones teniendo

únicamente en cuenta la parte técnica.

Siempre tendremos que validarlo con los expertos.

O las personas que conocen más el negocio.

Al final, lograr crear una síntesis para hacer el resumen de los datos compartidos,

su contenido.

Y por supuesto,

poderlo comunicar y validar con los expertos los hallazgos encontrados.

Y de esta manera,

enriquecer este proceso que estamos haciendo de construcción en analítica.

Play video starting at :3:27 and follow transcript3:27

Finalmente, encontrarás una serie de conclusiones.

Donde en las conclusiones te invito a revisar cómo hacer la síntesis, ¿sí?

Entonces aquí, por ejemplo, empezamos a hablar de la síntesis que responde a la

pregunta número uno, como lo observan acá.

¿Qué significa un registro de la fuente de datos?

Nuestra respuesta es, nos dieron 700.000 clientes

de los cuales tenemos 10 columnas categóricas y 5 numéricas.

De esas dos columnas, 2 identifican de forma única a los clientes.

3 representan datos demográficos, 4 representan información financiera,

datos familiares.

Y por último, por ejemplo,

información del producto más antiguo que tiene el cliente con la empresa.

La mayoría de los clientes que nos compartieron son prepago,

con nivel económico 3, casados y con hijos.

Aquí valdría la pena si podemos colocar un porcentaje.

Por ejemplo, la mayoría de los clientes, representada por un 80%, son prepago.

Y así para cada una de las síntesis.

Está la síntesis y está la pregunta que está contestando.

Podemos navegar hacia la segunda página de conclusiones.

Y aquí tenemos, por ejemplo, a nivel de inconsistencias,

se identifican clientes con edad superior a 90 años.

Siendo el valor más crítico de 150 años, el cual está presente en 10 registros.

De igual manera, a nivel de estado civil,

se detectó como estado civil el valor de 6, del cual no se tiene información.

Finalmente, se tienen 10 valores de productos que superan los informados por

ustedes como valores posibles.

Como se dan cuenta, son ideas de cómo podemos cerrar el documento.

Y esta parte muy relacionada con la calidad de los datos.

Te invito a que lo revises con cuidado, con calma.

Play video starting at :5:15 and follow transcript5:15

Y si estuvieras interesado en mayor nivel de detalle,

en esta primera parte te dejo referencias donde podrás consultar.

Como te digo, con mayor detalle lo que acabamos de ver en este video tan corto.

Hasta

pronto.

[MUSIC]

# Modelado multidimensional de datos

Save note

TranscriptNotesDownloadsDiscuss

### **Interactive Transcript - Enable basic transcript mode by pressing the escape key**

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at :: and follow transcript0:00

[MUSIC]

Bienvenidos a este video introductorio sobre modelado multidimensional,

en el que hablaremos de lo qué es y cómo construirlo.

Play video starting at ::18 and follow transcript0:18

Un modelo multidimensional es la representación conceptual de un proceso de

negocio.

Como puede ser en el caso del sector retail, ventas, pedidos de clientes,

abastecimiento, entre otros.

Este modelo está caracterizado por tener una tabla principal denominada tabla de

hechos.

Que representa hechos relacionados con el proceso que se está modelando,

por ejemplo, la venta.

Play video starting at ::43 and follow transcript0:43

Este hecho está descrito por una serie de perspectivas que denominamos dimensiones.

En este caso, la fecha en que ocurre la venta de un producto,

el cliente que la hace, la sucursal donde la compra y el vendedor que lo atiende.

Finalmente, tenemos las medidas que son elementos que queremos observar

en los análisis.

Como por ejemplo, el número de unidades vendidas y valor de la venta.

Estos modelos deben facilitar al tomador de decisiones el análisis.

Esto implica que deben ser fáciles de comprender.

Fíjense lo que hicimos ahora,

íbamos hablando, describiendo el proceso y dibujando.

Deben también permitir resolver análisis de forma rápida y eficiente.

Y, en principio, no debe preocuparnos el incorporar redundancia de datos en ello.

Play video starting at :1:31 and follow transcript1:31

Este modelo es inspirado en los cubos Rubik de 3x3 y vemos cómo una de esas

aristas corresponde a una dimensión de análisis.

Este es el caso de esta arista que representa la fecha.

Vemos cómo para cada fecha tenemos los valores de la venta y la cantidad.

Acá, por ejemplo, tenemos la compra de un computador portátil realizada el

15 de marzo de 2020 por un costo de $1200.

Podemos observar que esa compra se realizó en la sucursal de Bogotá, Colombia.

Play video starting at :2:4 and follow transcript2:04

El siguiente ejemplo a revisar es la metodología de construcción.

Es así como debemos empezar por seleccionar el proceso de negocio

a modelar.

Como segundo paso debemos definir la granularidad de la tabla de hechos.

Que corresponde al significado de una fila en dicha tabla.

El paso 3 es la definición de dimensiones.

Y el cuarto, la identificación de las medidas.

En este punto es muy importante garantizar la consistencia entre el grano definido en

el paso 2 y las dimensiones declaradas.

Imaginemos que hablamos del total de ventas de una sucursal en una fecha dada.

En esa situación, la dimensión cliente no tiene sentido.

you que las ventas representadas corresponden a varios clientes y no

solo a uno.

Play video starting at :2:48 and follow transcript2:48

Veamos cómo construiremos el modelo inicial.

En el paso 1, seleccionamos el proceso de ventas.

Como paso 2,

definimos que queremos tener el detalle del producto comprado por el cliente.

Incluyendo la sucursal de compra y el vendedor.

Esto nos define las dimensiones.

Por último, llegamos a que de esas ventas queremos tener la cantidad y valor de

las mismas para hacer análisis.

Como conocer el valor total de las ventas de un día por vendedor,

sucursal o producto.

Todas ellas posibles de contestar con nuestro modelo.

Preparemos ahora lápiz y papel y sigamos la metodología propuesta.

Para la compra de boletas de un evento de entretenimiento,

como puede ser un partido de fútbol.

En este caso, ¿cuál es el proceso a modelar?

Play video starting at :3:33 and follow transcript3:33

Muy bien, compra de boleto.

El nivel de detalle es el detalle de la compra total realizada por un

cliente para un evento en una fecha que incluye todas las sillas compradas.

En este caso bajo el supuesto de que no nos interesa el nivel de detalle de la

ubicación de las sillas compradas.

En nuestro punto tres de la metodología identificamos nuevas dimensiones como son,

tipo de descuento, ubicación, forma de entrega.

Estas últimas dimensiones las identificamos comprendiendo mejor el

proceso de negocio.

Y llegamos al punto cuatro donde qué movidas proponemos, pensemos un segundo.

Perfecto, llegamos al total de sillas compradas, valor de la compra,

incluido descuento y valor del descuento.

Completemos ahora el modelo construido,

incluyendo características de las dimensiones, las que nombramos atributos.

Acá podemos imaginar que tenemos información del tipo de descuento como

descripción, fecha inicial del descuento y su vigencia.

Play video starting at :4:33 and follow transcript4:33

En este punto, se debe validar que esas son características de la dimensión que se

está analizando.

Y no dependen de más de una dimensión sin incluir la fecha.

Si no es así,

es posible que se trate de otra dimensión o de las medidas de la tabla de hechos.

Veamos ahora cómo relacionamos las dimensiones con las tablas de hecho.

Para comprender esto,

es importante hablar de las opciones que tenemos de persistir este modelo.

Play video starting at :5:1 and follow transcript5:01

Estas son, en una base de datos relacional o en una multidimensional.

En el primer caso hablamos de un modelo ROLAP,

en el cual se utiliza una base de datos relacional para su persistencia.

Y en el segundo caso hablamos de MOLAP,

que utiliza una base de datos multidimensional.

Estas últimas bases de datos almacenan ejes de intersección entre ellos en los

que se denominan jerarquía de atributo.

Un tema que se entenderá mejor al revisar el concepto de análisis OLAP que

veremos más adelante en el curso.

Para este ejemplo, trabajemos con una representación ROLAP en la cual

las dimensiones y las tablas de hechos son tablas.

Y las relaciones entre las tablas de hechos y las dimensiones se da por

medio de una llave foránea.

Para lo cual cada dimensión debe tener una llave primaria preferiblemente

o una llave única.

De esta manera, la tabla de hechos tendrá como columnas tantas como llaves foráneas

de las dimensiones con las que se relacionan.

Más el número de medidas que tenga.

Play video starting at :6:2 and follow transcript6:02

Con esto terminamos este video, espero que sea de utilidad.

Ahora a practicar para poder construir modelos multidimensionales.

Y aplicarlos a las tareas que realizarán.

[MUSIC]

Modelado multidimensional de datos Save note Transcript Notes Downloads Discuss Interactive Transcript - Enable basic transcript mode by pressing the escape key You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key Play video starting at :: and follow transcript0:00 [MUSIC] Bienvenidos a este video introductorio sobre modelado multidimensional, en el que hablaremos de lo qué es y cómo construirlo. Play video starting at ::18 and follow transcript0:18 Un modelo multidimensional es la representación conceptual de un proceso de negocio. Como puede ser en el caso del sector retail, ventas, pedidos de clientes, abastecimiento, entre otros. Este modelo está caracterizado por tener una tabla principal denominada tabla de hechos. Que representa hechos relacionados con el proceso que se está modelando, por ejemplo, la venta. Play video starting at ::43 and follow transcript0:43 Este hecho está descrito por una serie de perspectivas que denominamos dimensiones. En este caso, la fecha en que ocurre la venta de un producto, el cliente que la hace, la sucursal donde la compra y el vendedor que lo atiende. Finalmente, tenemos las medidas que son elementos que queremos observar en los análisis. Como por ejemplo, el número de unidades vendidas y valor de la venta. Estos modelos deben facilitar al tomador de decisiones el análisis. Esto implica que deben ser fáciles de comprender. Fíjense lo que hicimos ahora, íbamos hablando, describiendo el proceso y dibujando. Deben también permitir resolver análisis de forma rápida y eficiente. Y, en principio, no debe preocuparnos el incorporar redundancia de datos en ello. Play video starting at :1:31 and follow transcript1:31 Este modelo es inspirado en los cubos Rubik de 3x3 y vemos cómo una de esas aristas corresponde a una dimensión de análisis. Este es el caso de esta arista que representa la fecha. Vemos cómo para cada fecha tenemos los valores de la venta y la cantidad. Acá, por ejemplo, tenemos la compra de un computador portátil realizada el 15 de marzo de 2020 por un costo de $1200. Podemos observar que esa compra se realizó en la sucursal de Bogotá, Colombia. Play video starting at :2:4 and follow transcript2:04 El siguiente ejemplo a revisar es la metodología de construcción. Es así como debemos empezar por seleccionar el proceso de negocio a modelar. Como segundo paso debemos definir la granularidad de la tabla de hechos. Que corresponde al significado de una fila en dicha tabla. El paso 3 es la definición de dimensiones. Y el cuarto, la identificación de las medidas. En este punto es muy importante garantizar la consistencia entre el grano definido en el paso 2 y las dimensiones declaradas. Imaginemos que hablamos del total de ventas de una sucursal en una fecha dada. En esa situación, la dimensión cliente no tiene sentido. you que las ventas representadas corresponden a varios clientes y no solo a uno. Play video starting at :2:48 and follow transcript2:48 Veamos cómo construiremos el modelo inicial. En el paso 1, seleccionamos el proceso de ventas. Como paso 2, definimos que queremos tener el detalle del producto comprado por el cliente. Incluyendo la sucursal de compra y el vendedor. Esto nos define las dimensiones. Por último, llegamos a que de esas ventas queremos tener la cantidad y valor de las mismas para hacer análisis. Como conocer el valor total de las ventas de un día por vendedor, sucursal o producto. Todas ellas posibles de contestar con nuestro modelo. Preparemos ahora lápiz y papel y sigamos la metodología propuesta. Para la compra de boletas de un evento de entretenimiento, como puede ser un partido de fútbol. En este caso, ¿cuál es el proceso a modelar? Play video starting at :3:33 and follow transcript3:33 Muy bien, compra de boleto. El nivel de detalle es el detalle de la compra total realizada por un cliente para un evento en una fecha que incluye todas las sillas compradas. En este caso bajo el supuesto de que no nos interesa el nivel de detalle de la ubicación de las sillas compradas. En nuestro punto tres de la metodología identificamos nuevas dimensiones como son, tipo de descuento, ubicación, forma de entrega. Estas últimas dimensiones las identificamos comprendiendo mejor el proceso de negocio. Y llegamos al punto cuatro donde qué movidas proponemos, pensemos un segundo. Perfecto, llegamos al total de sillas compradas, valor de la compra, incluido descuento y valor del descuento. Completemos ahora el modelo construido, incluyendo características de las dimensiones, las que nombramos atributos. Acá podemos imaginar que tenemos información del tipo de descuento como descripción, fecha inicial del descuento y su vigencia. Play video starting at :4:33 and follow transcript4:33 En este punto, se debe validar que esas son características de la dimensión que se está analizando. Y no dependen de más de una dimensión sin incluir la fecha. Si no es así, es posible que se trate de otra dimensión o de las medidas de la tabla de hechos. Veamos ahora cómo relacionamos las dimensiones con las tablas de hecho. Para comprender esto, es importante hablar de las opciones que tenemos de persistir este modelo. Play video starting at :5:1 and follow transcript5:01 Estas son, en una base de datos relacional o en una multidimensional. En el primer caso hablamos de un modelo ROLAP, en el cual se utiliza una base de datos relacional para su persistencia. Y en el segundo caso hablamos de MOLAP, que utiliza una base de datos multidimensional. Estas últimas bases de datos almacenan ejes de intersección entre ellos en los que se denominan jerarquía de atributo. Un tema que se entenderá mejor al revisar el concepto de análisis OLAP que veremos más adelante en el curso. Para este ejemplo, trabajemos con una representación ROLAP en la cual las dimensiones y las tablas de hechos son tablas. Y las relaciones entre las tablas de hechos y las dimensiones se da por medio de una llave foránea. Para lo cual cada dimensión debe tener una llave primaria preferiblemente o una llave única. De esta manera, la tabla de hechos tendrá como columnas tantas como llaves foráneas de las dimensiones con las que se relacionan. Más el número de medidas que tenga. Play video starting at :6:2 and follow transcript6:02 Con esto terminamos este video, espero que sea de utilidad. Ahora a practicar para poder construir modelos multidimensionales. Y aplicarlos a las tareas que realizarán. [MUSIC] Transcript language: Español : Added to Selection. Press [⌘ + S] to save as a note

Transcript language: Español

# Errores modelado multidimensional

Save note

TranscriptNotesDownloadsDiscuss

### **Interactive Transcript - Enable basic transcript mode by pressing the escape key**

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at :: and follow transcript0:00

[MUSIC]

Play video starting at ::13 and follow transcript0:13

Hola, bienvenidos a este video sobre errores comunes en modelo

multidimensionales e ideas para mejorarlos.

Empecemos por recordar que los modelos multidimensionales tienen una

tabla central denominada tabla de hechos,

que está acompañada de una serie de tablas denominadas dimensiones.

El diseño de estos modelos requiere desarrollar una serie de competencias.

Y en este video nos concentraremos en la que corresponde a identificar errores en

los modelos y proponer soluciones.

Los casos que se presentan fueron inspirados en procesos reales.

Así que los invito a comprender el contexto y luego revisar el caso y la

solución propuesta.

Iniciemos con este caso denominado TuBoleta.

Y en el proceso que representa la compra de boletas para asistir a eventos,

como teatro, conciertos, entre otros.

En este caso, el gerente financiero de la entidad ha manifestado interés en resolver

algunas preguntas que conocemos como requerimientos analíticos.

Por ejemplo,

saber cuáles son los eventos que generan más ingresos por venta de boletas.

Es al tratar de resolver este caso que un estudiante nos

propone el siguiente modelo.

Play video starting at :1:26 and follow transcript1:26

Veámoslo con calma.

Identifiquemos la tabla de HechosCompraEvento y las diez dimensiones.

Play video starting at :1:34 and follow transcript1:34

Ahora pensemos en qué errores se evidencian en este modelo.

Play video starting at :1:39 and follow transcript1:39

El primer error está relacionado con la dimensión Hora.

Si se dan cuenta,

en el diagrama hay dos entidades que conceptualmente representan lo mismo.

¿Cuál es la solución?

Tengamos entonces una sola dimensión Hora y dos relaciones con esa dimensión

llamadas HoraCompra y HoraEvento.

Veamos con claridad la solución propuesta.

Fíjense que tenemos dos flechas de nombre,

que representan las relaciones HoraEvento y HoraCompra respectivamente.

Y en este escenario, eliminamos una de las dimensiones, por ejemplo, HoraEvento.

Play video starting at :2:14 and follow transcript2:14

El siguiente error a reforzar es el de esta dimensión NumSillas.

Fíjense que no es una dimensión sino una medida de la compra,

you que ese número depende del evento, del cliente, la fecha, y en general,

depende de más de una dimensión.

También pueden verlo como que es una característica de la compra que se hace.

Play video starting at :2:34 and follow transcript2:34

Esto nos lleva a corregirla, eliminando

la dimensión NumSillas y creando la medida NumeroSillas en la tabla de hechos.

Veamos otro de los errores clásicos en estos modelos.

Se trata del atributo Valor del descuento de la dimensión Descuento.

Fíjense que ese valor depende del cliente,

la fecha, el evento, entre otras dimensiones.

O pueden verlo como que corresponde a una característica de la compra de un evento.

Lo que nos lleva a concluir que se trata de una medida,

una reflexión similar al error anterior.

Play video starting at :3:7 and follow transcript3:07

Con esa claridad y de forma similar dada para el error dos, debemos eliminar

ese atributo de la dimensión y crear la medida ValorDcto en CompraEvento.

Play video starting at :3:18 and follow transcript3:18

Veamos entonces cómo queda el diagrama corregido integrando las diferentes

soluciones a los errores que presentamos previamente.

Play video starting at :3:26 and follow transcript3:26

Muy bien, you estamos un paso más adelante en temas de modelados multidimensional.

Play video starting at :3:32 and follow transcript3:32

Cambiemos de caso y vayamos ahora a uno de transporte público denominado

Transmilenio en el cual se presenta la entrada de un pasajero a una estación.

En este caso, de nuevo, vemos que hay elementos clave para identificar

dimensiones como estación, fecha, el tipo de tarjeta, entre otros.

De igual manera, identificamos a la subgerente de servicios.

A quien le gustaría saber con la información almacenada en el modelo

propuesto de qué localidades son los usuarios que utilizan Transmilenio en una

estación en una fecha y hora específica.

De nuevo,

partimos de un modelo multidimensional propuesto por un estudiante.

Play video starting at :4:11 and follow transcript4:11

Identifiquemos rápidamente la tabla de hechos.

Se trata de HechoViajes que está relacionada con cinco dimensiones.

Play video starting at :4:21 and follow transcript4:21

El primer error de este modelo está representado en la dimensión Ruta.

Si la miran en detalle, verán que dentro de sus atributos tenemos estaciones,

fecha, hora,

lo que nos está indicando que se trata de una tabla de hechos y no de una dimensión.

Play video starting at :4:37 and follow transcript4:37

La forma de corregirlo es dándole la categoría que tiene la entidad.

Por lo cual denominamos HechoRuta a la antigua dimensión y creamos las flechas

hacia las dimensiones con las que está relacionada.

Play video starting at :4:50 and follow transcript4:50

En esta corrección,

identificamos claramente numParadas como una medida de esa nueva tabla de hechos.

En esta solución vemos que, asociado a un caso de análisis, podemos tener más de una

tabla de hechos para representar la información que se requiere.

Y lo mismo ocurre en un proceso de negocio en el cual también podemos tener más de

una tabla de hechos.

Continuemos con la identificación de errores,

¿qué otros errores ven en esta segunda solución?

Play video starting at :5:20 and follow transcript5:20

Muy bien, se trata de la dimensión Tarifa.

Este error es el error tipo dos que vimos previamente,

así que te invito a proponer una solución para corregirlo.

Play video starting at :5:31 and follow transcript5:31

Llegamos al último caso Rhino1 que se trata de una agencia de viajes en África

y debemos modelarle el proceso de reserva de hotel.

Play video starting at :5:40 and follow transcript5:40

Veamos esta propuesta de modelo.

Tenemos la tabla de HechoReserva y siete dimensiones.

Play video starting at :5:48 and follow transcript5:48

Fíjense que todos los casos tratados a este momento tienen una dimensión fecha.

Y esto es una particularidad de estos modelos multidimensionales.

De nuevo, enfoquémonos en el modelo y detectemos errores.

Play video starting at :6:2 and follow transcript6:02

Este es un nuevo tipo de error, donde debemos preguntarnos,

¿qué valor colocamos en esos atributos?

you que ese cliente puede haber realizado varios safaris con Rhino, por lo tanto,

tenemos varias reservas.

Y es posible que en cada una de ellas haya utilizado una tarjeta distinta.

Así que tenemos dos opciones.

Una es colocarle como sufijo a esos campos Actual,

para saber que se trata de la última tarjeta que conocemos del cliente.

Sin embargo, si queremos saber los tipos de tarjeta que ha utilizado el cliente en

cada reserva, no podríamos hacerlo.

Así que la mejor opción es la segunda,

en la que retiran los campos de tarjeta de la dimensión Cliente.

Se crea una nueva dimensión llamada TipoTarjeta,

donde incluimos los tipos de tarjeta.

Y una relación entre la tabla de hechos y la nueva dimensión.

Una variación también válida es incluir en el cliente,

adicional a lo anterior, la información de la última tarjeta que se conoce de él.

Play video starting at :7:2 and follow transcript7:02

Ahora, quiero que pongan en práctica lo visto en el video para identificar otros

errores del modelo.

Play video starting at :7:9 and follow transcript7:09

Efectivamente hay varios y es posible que dentro de ellos hayas identificado los que

están en rojo.

Play video starting at :7:15 and follow transcript7:15

Trata de explicarlos.

¿Qué propuestas tienes para corregirlos?

Play video starting at :7:21 and follow transcript7:21

Imaginemos ahora que la reserva realizada incluye varias habitaciones para

hospedar el total de personas asociadas a la reserva.

¿Qué propones para manejar ese concepto?

Te voy a ayudar.

En este tipo de modelos, hay un concepto que denominamos grupos,

que es de mucha utilidad.

you te muestro en qué consiste.

Veamos el caso de Servicios.

En el enunciado se habla de que el hotel le informa al cliente los servicios

incluidos, desayunos, Internet, cena, etc.

Todos con el mismo costo de la habitación.

Así que ese concepto puede llevar a que seleccionemos varios de los servicios en

una reserva.

Esto es lo que denominamos grupo, en este caso grupo de servicios.

El modelo propuesto evidencia una forma valiosa de modelar este concepto.

Se trata de una dimensión, cuyo número de columnas corresponde al número de

servicios acompañado del identificador del servicio.

En este caso, fijémonos que la dimensión Grupos de Servicios tendrá seis columnas,

que incluyen los servicios de desayuno y cena, entre otros.

Esta es la solución ideal para grupos de atributos que se manejan de forma grupal

y no individual.

Por lo cual es perfecto para este caso en el que el valor de la habitación es el

mismo, independiente de los servicios seleccionados.

Play video starting at :8:39 and follow transcript8:39

Veamos una opción dos para el manejo de grupo.

En este caso, debemos crear 3 dimensiones, Servicio,

que contiene como fila a los posibles servicios.

GrupoServicio, que define los servicios de cada grupo.

Y Grupo, que contiene los identificadores y descripciones específicas de cada grupo.

En este caso, creamos una relación entre la dimensión Grupo y la tabla de hechos de

reserva para representar el grupo de servicios de dicha reserva y entre

Servicio, GrupoServicio y Grupo.

Play video starting at :9:10 and follow transcript9:10

Una variante de esta propuesta es cuando hay características de los servicios que

definen el grupo.

Es así como en este caso estas características representan el tipo de

desayuno y de cena que quiere el cliente

dentro de unas alternativas dadas como desayuno americano.

Por esta razón, se crea una columna adicional en la dimensión GrupoServicio

que permite manejar este concepto.

Los invito a revisar la actividad asociada en la plataforma a este video para

formalizar las preguntas que se deben resolver y entregar.

Recapitulemos lo visto en este video.

Identificamos tipos de errores con posibles soluciones y recomendaciones

para el manejo de grupos.

Para finalizar, incluye una forma de identificar los errores basada en

responder las preguntas que están en pantalla.

Como, por ejemplo,

cómo saber que la entidad candidata a ser una dimensión es en realidad una medida.

Play video starting at :10:7 and follow transcript10:07

Espero que esto sea útil y puedas aplicar los conceptos desarrollados en este nivel.

[MUSIC]

Errores modelado multidimensional Save note Transcript Notes Downloads Discuss Interactive Transcript - Enable basic transcript mode by pressing the escape key You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key Play video starting at :: and follow transcript0:00 [MUSIC] Play video starting at ::13 and follow transcript0:13 Hola, bienvenidos a este video sobre errores comunes en modelo multidimensionales e ideas para mejorarlos. Empecemos por recordar que los modelos multidimensionales tienen una tabla central denominada tabla de hechos, que está acompañada de una serie de tablas denominadas dimensiones. El diseño de estos modelos requiere desarrollar una serie de competencias. Y en este video nos concentraremos en la que corresponde a identificar errores en los modelos y proponer soluciones. Los casos que se presentan fueron inspirados en procesos reales. Así que los invito a comprender el contexto y luego revisar el caso y la solución propuesta. Iniciemos con este caso denominado TuBoleta. Y en el proceso que representa la compra de boletas para asistir a eventos, como teatro, conciertos, entre otros. En este caso, el gerente financiero de la entidad ha manifestado interés en resolver algunas preguntas que conocemos como requerimientos analíticos. Por ejemplo, saber cuáles son los eventos que generan más ingresos por venta de boletas. Es al tratar de resolver este caso que un estudiante nos propone el siguiente modelo. Play video starting at :1:26 and follow transcript1:26 Veámoslo con calma. Identifiquemos la tabla de HechosCompraEvento y las diez dimensiones. Play video starting at :1:34 and follow transcript1:34 Ahora pensemos en qué errores se evidencian en este modelo. Play video starting at :1:39 and follow transcript1:39 El primer error está relacionado con la dimensión Hora. Si se dan cuenta, en el diagrama hay dos entidades que conceptualmente representan lo mismo. ¿Cuál es la solución? Tengamos entonces una sola dimensión Hora y dos relaciones con esa dimensión llamadas HoraCompra y HoraEvento. Veamos con claridad la solución propuesta. Fíjense que tenemos dos flechas de nombre, que representan las relaciones HoraEvento y HoraCompra respectivamente. Y en este escenario, eliminamos una de las dimensiones, por ejemplo, HoraEvento. Play video starting at :2:14 and follow transcript2:14 El siguiente error a reforzar es el de esta dimensión NumSillas. Fíjense que no es una dimensión sino una medida de la compra, you que ese número depende del evento, del cliente, la fecha, y en general, depende de más de una dimensión. También pueden verlo como que es una característica de la compra que se hace. Play video starting at :2:34 and follow transcript2:34 Esto nos lleva a corregirla, eliminando la dimensión NumSillas y creando la medida NumeroSillas en la tabla de hechos. Veamos otro de los errores clásicos en estos modelos. Se trata del atributo Valor del descuento de la dimensión Descuento. Fíjense que ese valor depende del cliente, la fecha, el evento, entre otras dimensiones. O pueden verlo como que corresponde a una característica de la compra de un evento. Lo que nos lleva a concluir que se trata de una medida, una reflexión similar al error anterior. Play video starting at :3:7 and follow transcript3:07 Con esa claridad y de forma similar dada para el error dos, debemos eliminar ese atributo de la dimensión y crear la medida ValorDcto en CompraEvento. Play video starting at :3:18 and follow transcript3:18 Veamos entonces cómo queda el diagrama corregido integrando las diferentes soluciones a los errores que presentamos previamente. Play video starting at :3:26 and follow transcript3:26 Muy bien, you estamos un paso más adelante en temas de modelados multidimensional. Play video starting at :3:32 and follow transcript3:32 Cambiemos de caso y vayamos ahora a uno de transporte público denominado Transmilenio en el cual se presenta la entrada de un pasajero a una estación. En este caso, de nuevo, vemos que hay elementos clave para identificar dimensiones como estación, fecha, el tipo de tarjeta, entre otros. De igual manera, identificamos a la subgerente de servicios. A quien le gustaría saber con la información almacenada en el modelo propuesto de qué localidades son los usuarios que utilizan Transmilenio en una estación en una fecha y hora específica. De nuevo, partimos de un modelo multidimensional propuesto por un estudiante. Play video starting at :4:11 and follow transcript4:11 Identifiquemos rápidamente la tabla de hechos. Se trata de HechoViajes que está relacionada con cinco dimensiones. Play video starting at :4:21 and follow transcript4:21 El primer error de este modelo está representado en la dimensión Ruta. Si la miran en detalle, verán que dentro de sus atributos tenemos estaciones, fecha, hora, lo que nos está indicando que se trata de una tabla de hechos y no de una dimensión. Play video starting at :4:37 and follow transcript4:37 La forma de corregirlo es dándole la categoría que tiene la entidad. Por lo cual denominamos HechoRuta a la antigua dimensión y creamos las flechas hacia las dimensiones con las que está relacionada. Play video starting at :4:50 and follow transcript4:50 En esta corrección, identificamos claramente numParadas como una medida de esa nueva tabla de hechos. En esta solución vemos que, asociado a un caso de análisis, podemos tener más de una tabla de hechos para representar la información que se requiere. Y lo mismo ocurre en un proceso de negocio en el cual también podemos tener más de una tabla de hechos. Continuemos con la identificación de errores, ¿qué otros errores ven en esta segunda solución? Play video starting at :5:20 and follow transcript5:20 Muy bien, se trata de la dimensión Tarifa. Este error es el error tipo dos que vimos previamente, así que te invito a proponer una solución para corregirlo. Play video starting at :5:31 and follow transcript5:31 Llegamos al último caso Rhino1 que se trata de una agencia de viajes en África y debemos modelarle el proceso de reserva de hotel. Play video starting at :5:40 and follow transcript5:40 Veamos esta propuesta de modelo. Tenemos la tabla de HechoReserva y siete dimensiones. Play video starting at :5:48 and follow transcript5:48 Fíjense que todos los casos tratados a este momento tienen una dimensión fecha. Y esto es una particularidad de estos modelos multidimensionales. De nuevo, enfoquémonos en el modelo y detectemos errores. Play video starting at :6:2 and follow transcript6:02 Este es un nuevo tipo de error, donde debemos preguntarnos, ¿qué valor colocamos en esos atributos? you que ese cliente puede haber realizado varios safaris con Rhino, por lo tanto, tenemos varias reservas. Y es posible que en cada una de ellas haya utilizado una tarjeta distinta. Así que tenemos dos opciones. Una es colocarle como sufijo a esos campos Actual, para saber que se trata de la última tarjeta que conocemos del cliente. Sin embargo, si queremos saber los tipos de tarjeta que ha utilizado el cliente en cada reserva, no podríamos hacerlo. Así que la mejor opción es la segunda, en la que retiran los campos de tarjeta de la dimensión Cliente. Se crea una nueva dimensión llamada TipoTarjeta, donde incluimos los tipos de tarjeta. Y una relación entre la tabla de hechos y la nueva dimensión. Una variación también válida es incluir en el cliente, adicional a lo anterior, la información de la última tarjeta que se conoce de él. Play video starting at :7:2 and follow transcript7:02 Ahora, quiero que pongan en práctica lo visto en el video para identificar otros errores del modelo. Play video starting at :7:9 and follow transcript7:09 Efectivamente hay varios y es posible que dentro de ellos hayas identificado los que están en rojo. Play video starting at :7:15 and follow transcript7:15 Trata de explicarlos. ¿Qué propuestas tienes para corregirlos? Play video starting at :7:21 and follow transcript7:21 Imaginemos ahora que la reserva realizada incluye varias habitaciones para hospedar el total de personas asociadas a la reserva. ¿Qué propones para manejar ese concepto? Te voy a ayudar. En este tipo de modelos, hay un concepto que denominamos grupos, que es de mucha utilidad. you te muestro en qué consiste. Veamos el caso de Servicios. En el enunciado se habla de que el hotel le informa al cliente los servicios incluidos, desayunos, Internet, cena, etc. Todos con el mismo costo de la habitación. Así que ese concepto puede llevar a que seleccionemos varios de los servicios en una reserva. Esto es lo que denominamos grupo, en este caso grupo de servicios. El modelo propuesto evidencia una forma valiosa de modelar este concepto. Se trata de una dimensión, cuyo número de columnas corresponde al número de servicios acompañado del identificador del servicio. En este caso, fijémonos que la dimensión Grupos de Servicios tendrá seis columnas, que incluyen los servicios de desayuno y cena, entre otros. Esta es la solución ideal para grupos de atributos que se manejan de forma grupal y no individual. Por lo cual es perfecto para este caso en el que el valor de la habitación es el mismo, independiente de los servicios seleccionados. Play video starting at :8:39 and follow transcript8:39 Veamos una opción dos para el manejo de grupo. En este caso, debemos crear 3 dimensiones, Servicio, que contiene como fila a los posibles servicios. GrupoServicio, que define los servicios de cada grupo. Y Grupo, que contiene los identificadores y descripciones específicas de cada grupo. En este caso, creamos una relación entre la dimensión Grupo y la tabla de hechos de reserva para representar el grupo de servicios de dicha reserva y entre Servicio, GrupoServicio y Grupo. Play video starting at :9:10 and follow transcript9:10 Una variante de esta propuesta es cuando hay características de los servicios que definen el grupo. Es así como en este caso estas características representan el tipo de desayuno y de cena que quiere el cliente dentro de unas alternativas dadas como desayuno americano. Por esta razón, se crea una columna adicional en la dimensión GrupoServicio que permite manejar este concepto. Los invito a revisar la actividad asociada en la plataforma a este video para formalizar las preguntas que se deben resolver y entregar. Recapitulemos lo visto en este video. Identificamos tipos de errores con posibles soluciones y recomendaciones para el manejo de grupos. Para finalizar, incluye una forma de identificar los errores basada en responder las preguntas que están en pantalla. Como, por ejemplo, cómo saber que la entidad candidata a ser una dimensión es en realidad una medida. Play video starting at :10:7 and follow transcript10:07 Espero que esto sea útil y puedas aplicar los conceptos desarrollados en este nivel. [MUSIC] Transcript language: Español : Added to Selection. Press [⌘ + S] to save as a note

Transcript language: Español

# Proceso ETL/ELT

Save note

TranscriptNotesDownloadsDiscuss

### **Interactive Transcript - Enable basic transcript mode by pressing the escape key**

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at ::13 and follow transcript0:13

Hola, nuevamente. En este video les presentaré el proceso de extracción,

transformación y carga, un elemento fundamental en la automatización de la analítica.

Les daré un contexto para entender mejor el porqué realizamos este proceso,

las competencias a nivel de aprendizaje que se requieren para su implementación,

el significado de hacer un ETL la primera vez y,

al final, retomaremos para hacer el cierre con los temas tratados en el video.

Lo que queremos con los procesos de analítica es poderlos

reutilizar cuando lleguen más datos de las fuentes de datos que se analizaron y,

de esta forma, mantener actualizados los análisis relacionados con dichas fuentes.

De igual manera, si hay nuevos análisis que requieren los mismos datos,

poder reutilizar lo previamente almacenado o hacer

una réplica siguiendo el mismo proceso para garantizar consistencia.

Es así como se habla de un proceso de automatización de la extracción de datos,

su transformación y carga,

conocido como proceso ETL,

del inglés "Extract",

"Transform", "Load", para el que se requiere conocer previamente las fuentes de datos,

los objetivos de la analítica a realizar,

el esquema de representación de los datos en repositorios donde

se tendrá la información para realizar las tareas de analítica,

la infraestructura donde se tendrán estos

servicios y la arquitectura de datos de la organización;

prerrequisitos logrados gracias a las tareas de

entender el negocio a nivel de sus objetivos estratégicos,

entender los datos y la infraestructura tecnológica para el desarrollo del ETL.

Recordemos que las fuentes de datos pueden ser diversas y,

en la mayoría de casos,

están relacionados con la operación de una organización.

Es en estas fuentes donde hablamos de fuentes de

datos estructuradas que vienen de los sistemas OLTP,

pero también encontramos fuentes semiestructuradas,

como son los datos de correos electrónicos;

fuentes no estructuradas,

como son las redes sociales como Facebook o WhatsApp,

estas últimas relacionadas con canales de comunicación hacia los clientes,

proveedores o empleados de una organización.

O tener acceso a flujos de datos como oxímetros

o monitores cardíacos utilizados en unidades de cuidados intensivos.

Tener acceso a dichas fuentes no es algo trivial.

Muchas veces requerimos de acuerdos de confidencialidad,

ajustes a contratos,

compra de datos o diseño de aplicaciones para recolectar los datos,

tareas que pueden tomar tiempo que afecte nuestro proyecto.

Una vez se tiene acceso a los datos,

el reto es poder tomar la información que nos brindan dichas fuentes,

responsabilidad del componente de extracción.

Este componente debe conocer el formato de la fuente de datos,

el lugar de almacenamiento y la frecuencia de llegada de nueva información.

Es así como, en ocasiones,

nos brindan acceso a copias de la base de datos transaccional que se denomina SandBox.

La extracción, en este caso,

corresponde a la ejecución de consultas, típicamente,

en SQL, para obtener la información que se requiere.

Otra forma es contar con copias en archivos,

comúnmente en formato CSV.

En este caso, debemos decidir dónde

almacenarlos y luego tener claridad de detalles de lectura,

como, por ejemplo, carácter de separación entre las columnas y tamaños esperados.

Otro mecanismo es el flujo de datos,

donde la responsabilidad del proceso no es extraerla,

sino permitir que sean depositados en algún repositorio.

Una estrategia "push" distinta a la de "pull",

que corresponde a los ejemplos previos.

En esta estrategia, cada vez que hay un nuevo dato,

la fuente debe colocarlo en el sitio destinado para ello,

sin necesidad de ser solicitado.

El segundo componente del proceso de ETL es el de transformación.

En este componente hablamos de dos tareas principales,

ambas relacionadas con los objetivos de analítica a realizar.

La primera, es la transformación para mejorar la calidad de los datos y la segunda,

para adaptar los datos a la tarea de analítica,

la cual se denomina "manipulación de datos" y en inglés "data wrangling".

En la primera tarea,

debemos determinar lo que ocurre cuando hay datos faltantes, datos duplicados,

valores fuera de rango o datos con distintos formatos,

caso típico del manejo de direcciones [inaudible].

En este punto, las decisiones giran en torno a realizar,

por ejemplo, tareas de imputación.

Dada una columna o variables sin valor,

podemos completarla con un valor mínimo,

máximo o, por ejemplo, la media.

Tareas de estandarización.

Llevar la fecha a un formato común,

por ejemplo, cuatro dígitos para el año,

dos para el mes y dos para el día; para direcciones,

típicamente, guiado por el país al cual pertenece la dirección,

o la decisión de manejar los textos siempre con mayúsculas o minúsculas,

o en formato de oración,

donde solo el primer carácter estará en mayúscula.

Tareas de eliminar registros cuando son duplicados o

cuando la mayoría de las columnas no aporta al análisis que se tiene,

o eliminar columnas cuando se tiene un valor

constante en ellas o un nivel muy alto de duplicación.

Estas tareas son típicas cuando se espera almacenarlas en bodegas o almacenes de datos.

En la segunda tarea,

debemos decidir, por ejemplo,

las variables a incluir en los análisis,

la creación de nuevas variables o la eliminación de algunas de ellas que son redundantes.

En esta segunda tarea es importante reflexionar si se trata de un

proceso muy particular para un tipo de analítica, y, si es así,

el realizarlo puede ser a partir de la información almacenada en un repositorio,

como la bodega de datos,

y no persistirlo o persistirlo en otro repositorio,

lo que denominamos la capa de entrega de información en la arquitectura que presentamos.

O el lago de datos,

para facilitar y agilizar analítica con datos ya preparados.

Esta tarea, podemos pensarla más claramente,

en un esquema de ELT,

donde los datos se cargan tal cual como son reportados por la fuente

en un lago de datos y luego se transforman para las tareas de analítica.

Llegamos al tercer componente, que es el de carga.

En este punto, hablamos de almacenar los datos en una bodega o almacén de

datos que maneja datos estructurados con un esquema apropiado para el análisis,

que en la mayoría de casos se identifica como una estructura...

En ROLAP, en la cual se tiene un modelo

multidimensional representado en una base de datos relacional.

En la actualidad encontramos representaciones en bases de

datos columnares que son repositorios no SQL o,

en su defecto, en archivos.

En este punto, es importante revisar las características del

repositorio para entender las actividades de preparación que requieren y,

de esta forma, realizar de forma ágil esta tarea.

En el caso de bases de datos relacionales, por ejemplo,

es importante desactivar el log transaccional que va

a causar incremento en el tiempo por cada registro a almacenar.

En este caso de almacenamiento,

en bodegas de datos,

no se requiere garantizar todas las propiedades así.

Se requieren, en particular,

mecanismos para saber cómo retomar el proceso en caso de falla y,

por supuesto, la durabilidad de los datos.

Estas dos funcionalidades se pueden lograr por medio de estrategias

complementarias a la funcionalidad del sistema

manejador de base de datos y de esta forma agilizarlos.

Como se observa a partir de la funcionalidad que brinda este proceso de ETL,

se requieren personas con competencias a nivel de manejo de datos en términos de SQL,

sistemas manejadores de bases de datos,

bodegas y almacenes de datos,

que comúnmente están concentradas en el rol de ingeniero de datos.

En lo relacionado con la programación,

se requieren conocimientos en lenguajes como Python,

Java, C o C++ los que manejan los ingenieros de software.

Y a nivel de analítica uno punto cero se requiere saber

de inteligencia de negocios y comprender, por ejemplo,

características generales de analítica predictiva y prescriptiva,

los cuales están relacionados con el rol de científico de datos.

Adicionalmente, el conocimiento de la arquitectura de datos y de infraestructura de una

organización son fundamentales para que el proceso

ETL cumpla con los estándares definidos por la organización.

Ahora, hablaremos del diseño e implementación de un proceso ETL por primera vez,

para realizar analítica uno punto cero y

almacenando los datos en una bodega o almacén de datos relacional.

Hablar de primera vez significa que no hay datos

previamente almacenados y por esto va a facilitar el proceso.

Con respecto al diseño; en este punto,

es importante definir el orden en el que se cargarán los datos,

tener claridad de las fuentes proveedoras de datos,

las transformaciones a realizar y el esquema en el que se van a almacenar los datos.

Sobre el orden, la recomendación es empezar de lo simple a lo complejo.

Es así como se inicia con las dimensiones que no tienen

relación con otras dimensiones y que cuya fuente es una única.

Aquí caen dimensiones como moneda, fecha.

Si no existen estas dimensiones,

en el mundo del OLTP,

se conocen como tablas de referencia o

dimensiones que para el caso de estudio son sencillas,

como empleados en nuestro ejemplo.

Luego, cargamos dimensiones que tienen

relaciones de jerarquía al interior de ellas o relaciones con otras dimensiones.

Este sería el caso de ciudad,

ya que pertenece a un país y se encuentra en una provincia.

Estos conceptos definen la jerarquía o relación de uno a muchos entre país y provincias.

Un país compuesto por muchas provincias y una provincia pertenece a un país;

y provincia y ciudades.

Finalmente, tenemos toda la información para cargar la tabla de hechos.

Veamos el tema de las tareas de extraer y transformar a nivel de diseño.

En el caso de empleados,

se trata de consultar la tabla de empleados del OLTP que tiene los campos numEmpleado,

Nombre, Apellido,

Fechanacimiento y cargo, y transformarlos utilizando, por ejemplo,

estas cinco transformaciones que resuelven principalmente los problemas

de calidad de datos para llevarlos a la tabla de empleados de la bodega de datos,

que a diferencia de la tabla transaccional,

tiene un idEmpleado que se genera en

la bodega de datos para identificar de forma única cada empleado;

el nombre, que es la unión del nombre y apellido de la tabla fuente y

la fecha que debe respetar el formato de dos dígitos para el día,

dos para el mes y cuatro para el año.

De forma similar, en el caso de ciudad,

debemos traer la información de tres fuentes distintas,

estandarizar la información a partir de las transformaciones propuestas

y generar su identificador en la bodega para almacenar

en una sola tabla las tres fuentes sin perder información.

En el caso de las tablas de hecho,

se deben cargar primero todas las dimensiones con las que está relacionada y luego,

realizando un join entre la información del

OLTP de las órdenes y las dimensiones ya cargadas en la bodega,

obtener los identificadores de cada dimensión y así cargarlos a la bodega.

El diseño nos permite entender mejor lo que debemos hacer antes de hacerlo.

Esta parte puede ser discutida entre los miembros del

proyecto y con personas de tecnología de la organización.

Todo esto con el fin de validarla antes de implementar.

Con la claridad del diseño del ETL estamos listos para realizar la implementación,

que se traduce en

el conjunto de instrucciones que permite realizar lo especificado en el diseño.

Es así,como siguiendo el diseño,

bajo el supuesto de que ya existen las tablas en la bodega de datos, aunque vacías,

debemos realizar el paso de consultar la fuente,

que en este caso, es un select asterisco from empleados por cargo igual a vendedor.

Luego, aplicar las transformaciones y llevarlas a una tabla temporal,

como lo vemos en la sentencia SQL,

para finalmente tomar la información de la tabla

temporal y llevarla a la tabla definitiva de la bodega de datos.

Como se observa en este ejemplo,

todo se hace de forma masiva para facilitar y agilizar el proceso.

Para finalizar, es importante recordar que en este video hablamos del proceso ETL,

que en la actualidad puede cambiar

a un proceso de ELT para cargar los datos y luego transformarlos,

aprovechando la capacidad computacional de la máquina destino.

Dentro de ese proceso vimos que es importante descomponerlo en el diseño y luego,

la implementación para validar la calidad y agilizar el proceso.

Para terminar, te invito

a reflexionar sobre tu experiencia en este tipo de procesos que a nivel del proyecto,

determina el esfuerzo que requieres para cargar

los datos al modelo definitivo en la bodega. Hasta pronto.

Proceso ETL/ELT Save note Transcript Notes Downloads Discuss Interactive Transcript - Enable basic transcript mode by pressing the escape key You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key Play video starting at ::13 and follow transcript0:13 Hola, nuevamente. En este video les presentaré el proceso de extracción, transformación y carga, un elemento fundamental en la automatización de la analítica. Les daré un contexto para entender mejor el porqué realizamos este proceso, las competencias a nivel de aprendizaje que se requieren para su implementación, el significado de hacer un ETL la primera vez y, al final, retomaremos para hacer el cierre con los temas tratados en el video. Lo que queremos con los procesos de analítica es poderlos reutilizar cuando lleguen más datos de las fuentes de datos que se analizaron y, de esta forma, mantener actualizados los análisis relacionados con dichas fuentes. De igual manera, si hay nuevos análisis que requieren los mismos datos, poder reutilizar lo previamente almacenado o hacer una réplica siguiendo el mismo proceso para garantizar consistencia. Es así como se habla de un proceso de automatización de la extracción de datos, su transformación y carga, conocido como proceso ETL, del inglés "Extract", "Transform", "Load", para el que se requiere conocer previamente las fuentes de datos, los objetivos de la analítica a realizar, el esquema de representación de los datos en repositorios donde se tendrá la información para realizar las tareas de analítica, la infraestructura donde se tendrán estos servicios y la arquitectura de datos de la organización; prerrequisitos logrados gracias a las tareas de entender el negocio a nivel de sus objetivos estratégicos, entender los datos y la infraestructura tecnológica para el desarrollo del ETL. Recordemos que las fuentes de datos pueden ser diversas y, en la mayoría de casos, están relacionados con la operación de una organización. Es en estas fuentes donde hablamos de fuentes de datos estructuradas que vienen de los sistemas OLTP, pero también encontramos fuentes semiestructuradas, como son los datos de correos electrónicos; fuentes no estructuradas, como son las redes sociales como Facebook o WhatsApp, estas últimas relacionadas con canales de comunicación hacia los clientes, proveedores o empleados de una organización. O tener acceso a flujos de datos como oxímetros o monitores cardíacos utilizados en unidades de cuidados intensivos. Tener acceso a dichas fuentes no es algo trivial. Muchas veces requerimos de acuerdos de confidencialidad, ajustes a contratos, compra de datos o diseño de aplicaciones para recolectar los datos, tareas que pueden tomar tiempo que afecte nuestro proyecto. Una vez se tiene acceso a los datos, el reto es poder tomar la información que nos brindan dichas fuentes, responsabilidad del componente de extracción. Este componente debe conocer el formato de la fuente de datos, el lugar de almacenamiento y la frecuencia de llegada de nueva información. Es así como, en ocasiones, nos brindan acceso a copias de la base de datos transaccional que se denomina SandBox. La extracción, en este caso, corresponde a la ejecución de consultas, típicamente, en SQL, para obtener la información que se requiere. Otra forma es contar con copias en archivos, comúnmente en formato CSV. En este caso, debemos decidir dónde almacenarlos y luego tener claridad de detalles de lectura, como, por ejemplo, carácter de separación entre las columnas y tamaños esperados. Otro mecanismo es el flujo de datos, donde la responsabilidad del proceso no es extraerla, sino permitir que sean depositados en algún repositorio. Una estrategia "push" distinta a la de "pull", que corresponde a los ejemplos previos. En esta estrategia, cada vez que hay un nuevo dato, la fuente debe colocarlo en el sitio destinado para ello, sin necesidad de ser solicitado. El segundo componente del proceso de ETL es el de transformación. En este componente hablamos de dos tareas principales, ambas relacionadas con los objetivos de analítica a realizar. La primera, es la transformación para mejorar la calidad de los datos y la segunda, para adaptar los datos a la tarea de analítica, la cual se denomina "manipulación de datos" y en inglés "data wrangling". En la primera tarea, debemos determinar lo que ocurre cuando hay datos faltantes, datos duplicados, valores fuera de rango o datos con distintos formatos, caso típico del manejo de direcciones [inaudible]. En este punto, las decisiones giran en torno a realizar, por ejemplo, tareas de imputación. Dada una columna o variables sin valor, podemos completarla con un valor mínimo, máximo o, por ejemplo, la media. Tareas de estandarización. Llevar la fecha a un formato común, por ejemplo, cuatro dígitos para el año, dos para el mes y dos para el día; para direcciones, típicamente, guiado por el país al cual pertenece la dirección, o la decisión de manejar los textos siempre con mayúsculas o minúsculas, o en formato de oración, donde solo el primer carácter estará en mayúscula. Tareas de eliminar registros cuando son duplicados o cuando la mayoría de las columnas no aporta al análisis que se tiene, o eliminar columnas cuando se tiene un valor constante en ellas o un nivel muy alto de duplicación. Estas tareas son típicas cuando se espera almacenarlas en bodegas o almacenes de datos. En la segunda tarea, debemos decidir, por ejemplo, las variables a incluir en los análisis, la creación de nuevas variables o la eliminación de algunas de ellas que son redundantes. En esta segunda tarea es importante reflexionar si se trata de un proceso muy particular para un tipo de analítica, y, si es así, el realizarlo puede ser a partir de la información almacenada en un repositorio, como la bodega de datos, y no persistirlo o persistirlo en otro repositorio, lo que denominamos la capa de entrega de información en la arquitectura que presentamos. O el lago de datos, para facilitar y agilizar analítica con datos ya preparados. Esta tarea, podemos pensarla más claramente, en un esquema de ELT, donde los datos se cargan tal cual como son reportados por la fuente en un lago de datos y luego se transforman para las tareas de analítica. Llegamos al tercer componente, que es el de carga. En este punto, hablamos de almacenar los datos en una bodega o almacén de datos que maneja datos estructurados con un esquema apropiado para el análisis, que en la mayoría de casos se identifica como una estructura... En ROLAP, en la cual se tiene un modelo multidimensional representado en una base de datos relacional. En la actualidad encontramos representaciones en bases de datos columnares que son repositorios no SQL o, en su defecto, en archivos. En este punto, es importante revisar las características del repositorio para entender las actividades de preparación que requieren y, de esta forma, realizar de forma ágil esta tarea. En el caso de bases de datos relacionales, por ejemplo, es importante desactivar el log transaccional que va a causar incremento en el tiempo por cada registro a almacenar. En este caso de almacenamiento, en bodegas de datos, no se requiere garantizar todas las propiedades así. Se requieren, en particular, mecanismos para saber cómo retomar el proceso en caso de falla y, por supuesto, la durabilidad de los datos. Estas dos funcionalidades se pueden lograr por medio de estrategias complementarias a la funcionalidad del sistema manejador de base de datos y de esta forma agilizarlos. Como se observa a partir de la funcionalidad que brinda este proceso de ETL, se requieren personas con competencias a nivel de manejo de datos en términos de SQL, sistemas manejadores de bases de datos, bodegas y almacenes de datos, que comúnmente están concentradas en el rol de ingeniero de datos. En lo relacionado con la programación, se requieren conocimientos en lenguajes como Python, Java, C o C++ los que manejan los ingenieros de software. Y a nivel de analítica uno punto cero se requiere saber de inteligencia de negocios y comprender, por ejemplo, características generales de analítica predictiva y prescriptiva, los cuales están relacionados con el rol de científico de datos. Adicionalmente, el conocimiento de la arquitectura de datos y de infraestructura de una organización son fundamentales para que el proceso ETL cumpla con los estándares definidos por la organización. Ahora, hablaremos del diseño e implementación de un proceso ETL por primera vez, para realizar analítica uno punto cero y almacenando los datos en una bodega o almacén de datos relacional. Hablar de primera vez significa que no hay datos previamente almacenados y por esto va a facilitar el proceso. Con respecto al diseño; en este punto, es importante definir el orden en el que se cargarán los datos, tener claridad de las fuentes proveedoras de datos, las transformaciones a realizar y el esquema en el que se van a almacenar los datos. Sobre el orden, la recomendación es empezar de lo simple a lo complejo. Es así como se inicia con las dimensiones que no tienen relación con otras dimensiones y que cuya fuente es una única. Aquí caen dimensiones como moneda, fecha. Si no existen estas dimensiones, en el mundo del OLTP, se conocen como tablas de referencia o dimensiones que para el caso de estudio son sencillas, como empleados en nuestro ejemplo. Luego, cargamos dimensiones que tienen relaciones de jerarquía al interior de ellas o relaciones con otras dimensiones. Este sería el caso de ciudad, ya que pertenece a un país y se encuentra en una provincia. Estos conceptos definen la jerarquía o relación de uno a muchos entre país y provincias. Un país compuesto por muchas provincias y una provincia pertenece a un país; y provincia y ciudades. Finalmente, tenemos toda la información para cargar la tabla de hechos. Veamos el tema de las tareas de extraer y transformar a nivel de diseño. En el caso de empleados, se trata de consultar la tabla de empleados del OLTP que tiene los campos numEmpleado, Nombre, Apellido, Fechanacimiento y cargo, y transformarlos utilizando, por ejemplo, estas cinco transformaciones que resuelven principalmente los problemas de calidad de datos para llevarlos a la tabla de empleados de la bodega de datos, que a diferencia de la tabla transaccional, tiene un idEmpleado que se genera en la bodega de datos para identificar de forma única cada empleado; el nombre, que es la unión del nombre y apellido de la tabla fuente y la fecha que debe respetar el formato de dos dígitos para el día, dos para el mes y cuatro para el año. De forma similar, en el caso de ciudad, debemos traer la información de tres fuentes distintas, estandarizar la información a partir de las transformaciones propuestas y generar su identificador en la bodega para almacenar en una sola tabla las tres fuentes sin perder información. En el caso de las tablas de hecho, se deben cargar primero todas las dimensiones con las que está relacionada y luego, realizando un join entre la información del OLTP de las órdenes y las dimensiones ya cargadas en la bodega, obtener los identificadores de cada dimensión y así cargarlos a la bodega. El diseño nos permite entender mejor lo que debemos hacer antes de hacerlo. Esta parte puede ser discutida entre los miembros del proyecto y con personas de tecnología de la organización. Todo esto con el fin de validarla antes de implementar. Con la claridad del diseño del ETL estamos listos para realizar la implementación, que se traduce en el conjunto de instrucciones que permite realizar lo especificado en el diseño. Es así,como siguiendo el diseño, bajo el supuesto de que ya existen las tablas en la bodega de datos, aunque vacías, debemos realizar el paso de consultar la fuente, que en este caso, es un select asterisco from empleados por cargo igual a vendedor. Luego, aplicar las transformaciones y llevarlas a una tabla temporal, como lo vemos en la sentencia SQL, para finalmente tomar la información de la tabla temporal y llevarla a la tabla definitiva de la bodega de datos. Como se observa en este ejemplo, todo se hace de forma masiva para facilitar y agilizar el proceso. Para finalizar, es importante recordar que en este video hablamos del proceso ETL, que en la actualidad puede cambiar a un proceso de ELT para cargar los datos y luego transformarlos, aprovechando la capacidad computacional de la máquina destino. Dentro de ese proceso vimos que es importante descomponerlo en el diseño y luego, la implementación para validar la calidad y agilizar el proceso. Para terminar, te invito a reflexionar sobre tu experiencia en este tipo de procesos que a nivel del proyecto, determina el esfuerzo que requieres para cargar los datos al modelo definitivo en la bodega. Hasta pronto. Transcript language: Español : Added to Selection. Press [⌘ + S] to save as a note

Transcript language: Español

# Modelado de dimensiones y recomendaciones de calidad de los modelos

Save note

TranscriptNotesDownloadsDiscuss

### **Interactive Transcript - Enable basic transcript mode by pressing the escape key**

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at ::9 and follow transcript0:09

Hola, nuevamente. En este video voy a hablarles de la forma de modelar dimensiones y los

elementos que debemos tener presentes para

garantizar cierto nivel de calidad del modelo propuesto.

Esto es importante para terminar de afinar la competencia de modelar

dimensionalmente y poder utilizarlo en contextos de organizaciones.

Recordemos que estamos trabajando en

el área de almacenamiento de datos para realizar análisis,

el cual, en esta aproximación,

corresponde a tener un modelo

dimensional que será luego implementado en una base de datos relacional,

a lo cual denominamos ROLAP.

Hasta este punto, hemos hablado de una serie de

criterios que caracterizan los modelos dimensionales

y que permiten validar su calidad,

es así como el modelo debe tener al menos una tabla central,

denominada tabla de hechos que representa un proceso de negocio.

Adicionalmente, siempre debemos hablar de una dimensión fecha,

que recordemos que en los análisis nos interesa lo que ha

pasado a lo largo del tiempo y no solo un punto en el tiempo.

El tipo de relaciones entre las dimensiones y la tabla de hechos, es 1-m;

un cliente hace varias compras y una compra está relacionada con un único cliente.

Las dimensiones que observamos son perspectivas de

análisis que usamos como filtro en el momento de hacer consultas,

por ejemplo, compras de la categoría de electrodomésticos,

entre el 1 de enero del 2022 y el 31 de marzo del 2022.

Los atributos de las dimensiones son características propias de la dimensión,

como lo es el nombre,

la categoría o subcategoría,

para el caso de producto.

No dependen de otras dimensiones adicionales al tiempo.

Ahora, abordemos el tema de cómo modelar una dimensión.

Algunos ejemplos de los clásicos de dimensiones son "cliente",

que en otros contextos puede ser el estudiante,

el proveedor o el paciente,

la geografía que puede representar la ubicación del cliente o de una sucursal,

el producto o servicio que corresponde a la

oferta realizada por una empresa a un cliente.

Estas dimensiones deben tener una llave primaria independiente del sistema transaccional,

vemos que sobre esta llave queremos control total y

no es posible si dependemos de la llave del sistema OLTP,

además hay varias formas de manejar la historia de los atributos de las dimensiones,

en las cuales no es posible utilizar la llave del OLTP,

ya que se tiene más de una fila representando la misma dimensión.

Esto último lo entenderemos mejor en

el recurso de historia de atributos de una dimensión.

Otros atributos comunes son los que vienen de las banderas o los indicadores,

en este caso particular,

de la dimensión "fecha",

observamos los dos últimos y hablamos de "festivo" y "día de la semana".

Entonces, fíjense que para una fecha en particular,

como 6 de Julio,

se podría decir que no es festivo,

mientras que para el 1 de mayo si es festivo.

Además, puedo entender por ejemplo,

si el 1 de mayo fue un día en la semana o de fin de semana,

para un año particular en el cual ocurrieron cierto número de ventas.

Vemos que también podemos tener múltiples campos de texto,

en este ejemplo tenemos la descripción de la fecha,

el nombre, el mes fiscal,

el día de la semana,

o como lo veíamos antes,

el saber si una fecha es o no un día festivo.

También pueden convivir múltiples jerarquías;

en nuestro ejemplo vemos en azul la jerarquía de fecha calendario y en

morado la jerarquía de fecha asociada al año fiscal de la organización.

Revisemos el concepto

de jerarquía para que todos tengamos el mismo lenguaje y recordemos que

una jerarquía es una relación de uno

a muchos al interior de los atributos de una dimensión,

es así como en nuestro ejemplo podemos observar que

varios días de la semana conforman un mes,

varios meses, un trimestre y cuatro trimestres un año.

Ahí estamos mostrando esa relación de uno a muchos,

un día pertenece a un único mes calendario,

pero un mes tiene múltiples días de la semana asociados.

Este concepto de jerarquía puede ser común en algunas de las dimensiones,

en particular aplica para el caso de la dimensión geografía, ¿ya sabes cuál es?

Adicionalmente, relacionado con el tema de jerarquías,

es posible que estés pensando en por qué todo está en la misma tabla de la

dimensión fecha y por qué no más bien tener tres tablas;

una de días, otra de meses y otra de años en un estilo normalizado,

lo que denominamos "Snow Flake" o copo de nieve.

En la parte de analítica se recomienda evitar la normalización,

principalmente porque puede ser tedioso el cargue y, en general,

el manejo puede ser más confuso en el momento de hacer consultas y en la actualidad,

aunque la cantidad de almacenamiento es importante,

no es grave manejar este tipo de recalificación controlada al interior de una dimensión,

es por eso que lo más simple y sencillo

es tener toda esa información en la misma dimensión.

Pasemos ahora a la dimensión "producto" y

veamos que a diferencia de la dimensión "fecha",

tenemos una llave subrogada que corresponde

a un número consecutivo creado para

identificar de forma única un registro de la dimensión.

Sin embargo, para mantener la relación con el sistema OLTP,

incluimos adicionalmente como atributo la llave natural,

que en este caso del producto es el SKU.

Además, igual que en la dimensión fecha,

identificamos varios tipos de texto,

nombre, color, categoría, entre otros.

Otra de las recomendaciones es, por ejemplo,

incluir columnas adicionales si la llave

natural u otras columnas de la fuente de datos tienen semántica.

En este caso, vemos que el SKU tiene una semántica,

en la cual lo primero que viene es la subcategoría del producto,

que en este caso es deportiva,

dado que nuestro ejemplo es de ropa y lo

manejamos con una columna especial de subcategoría.

El segundo elemento que debemos representar es el concepto de medias tobilleras,

en el cual queda el nombre del producto.

El tercero es el color blanco y, por último,

la talla que representa personas que calzan 40.

Fíjate que todos esos elementos los podemos extraer y crear

nuevas columnas independientes para manejarlos en los análisis,

es por eso que a nivel analítico,

tendremos una nueva columna por cada concepto,

porque en este caso es de interés incluirlas en análisis.

Finalmente, y no menos importante,

es el hecho de que debemos evitar atributos sin valor,

recordemos que es más complejo a nivel de consultas SQL.

Siempre se debe diferenciar entre si hay valor o no,

con el operador IS NULL y recordemos que estos modelos se

caracterizan por su simplicidad para entenderlo y consultarlo.

Así que aquí, es preferible colocar un "no aplica" o "sin valor" para atributos,

y, en caso de ser necesario, crear,

por ejemplo, un producto no identificado para

casos con ventas sobre códigos de productos que no logramos saber cuál es.

En este contexto, es preferible manejarlo así y

evitar a toda costa el tener valores nuevos.

Con esto terminamos las ideas generales que debemos

tener presentes en el modelado de las dimensiones y espero que,

a pesar de que lo que hice aquí fue

aplicarlo únicamente para "fecha" y "producto", puedas generalizar.

En particular, te invito a pensar en la dimensión "cliente",

que es un caso clásico en modelos multidimensionales.

Hablemos ahora de calidad de modelos y cómo

podemos validarla para mejorarla si es necesario.

Recordemos que hay una serie de preguntas que

nos ayudan a determinar la calidad de un modelo;

esas preguntas van desde saber

si una dimensión en realidad es una dimensión o una medida,

hasta encontrar inconsistencias a nivel de detalle,

cuando tenemos ciertas dimensiones y

observamos que algunas medidas no coinciden con esas dimensiones.

Ahora quiero incluir nuevas preguntas que van a

complementar nuestro proceso de validar la calidad de un modelo.

Empecemos desde la pregunta más simple,

¿hay una tabla central?

Fíjate que debe existir al menos una,

que es la llamada tabla de hechos,

la cual relaciona las diferentes dimensiones con la

información de las medidas y representa un proceso de negocio.

Otro ejemplo clave es la existencia de la dimensión "fecha";

siempre debe existir dado que,

en la parte analítica,

lo que queremos hacer es análisis de qué ha pasado,

para lo cual necesitamos una trayectoria o una historia.

Esta dimensión "fecha" está conectada únicamente con las tablas de "hechos",

porque recuerda la estructura y toda la información que tiene dicha dimensión,

lo veíamos previamente;

información de días de la semana,

si es festivo o no y esto es útil cuando queremos hacer análisis sobre los hechos.

Por ejemplo, en este caso,

si quieres saber si los días que se compra más son días de semana o fines de semana o,

por ejemplo, si son festivos o no o si son los lunes,

los martes, etcétera.

Esta semántica no la requerimos si tenemos fechas como atributos de las dimensiones.

Imagínate que tenemos la fecha de nacimiento de un cliente,

¿crees que tenemos que saber qué día nació el cliente?

¿Será que nació un fin de semana o en primer trimestre del año?

Para el contexto particular de un "retail" no importa y,

en términos generales, no se requiere;

es así como las atributos de las dimensiones serán de tipo "Date"

y no debe existir una relación entre ellos y la dimensión fecha.

Otra pregunta fundamental es ver si toda la tabla de hechos tiene al menos una medida.

Siempre debe tener al menos una y hay casos que denominamos tablas de hechos Factless,

porque el proceso para este nivel de granularidad no tiene medidas.

Sin embargo, a pesar de ello,

nos inventamos una columna que tenga un

valor constante para todas las filas y juegue el rol de medida.

Esto lo veremos en otro recurso,

pero siempre necesitamos tener medidas

y esas medidas son números sobre las cuales podemos hacer

operaciones como sumar y que son medidas

aditivas o podemos traer el mínimo o el máximo valor,

hacer promedio, entre otras operaciones.

Otra pregunta es la relacionada con los atributos multivalor,

¿es permitido en estos modelos tener atributos multivalor?

Multivalor es tener, por ejemplo,

listas como de servicios,

productos de clientes en un solo atributo.

La respuesta es: todo atributo que tengamos en este modelo va a ser monovalor,

porque nos interesa su análisis y poderlo consultar de forma simple.

Por último, tenemos una ñapa que es,

¿hay relación directa por medio de las llaves foráneas entre las tablas de hecho?

Lo que quiere decir es que las tablas de hecho tendrían una

llave primaria y podrían tener llaves foráneas entre ellas.

En este nivel de modelado y,

por razones pedagógicas, es preferible pensar que no,

lo que significa que la tabla de hechos no tiene llave primaria,

a pesar de que para efectos de carga de datos

y facilidad en proceso puede existir entre otras razones.

Solo llaves foráneas de dimensiones y medidas vamos a manejar acá.

De esta forma, cuando estamos modelando es preferible

no manejar la llave primaria para evitar esa

tentación de conectar las

tablas de hechos directamente entre ellas y no por medio de las dimensiones,

que es como debe realizarse.

Llegamos al final y es importante recordar que en

este video hablamos de la forma de modelar dimensión,

ejemplificando el caso de la dimensión de fecha,

que es especial, y la de producto.

Adicionalmente, vimos algunas preguntas para validar la calidad del modelo.

Espero que estos puntos puedan ayudarte

a fortalecer la competencia de modelar multidimensionalmente,

validar la calidad de los modelos y tener como resultado mejores modelos. Hasta pronto.

Modelado de dimensiones y recomendaciones de calidad de los modelos Save note Transcript Notes Downloads Discuss Interactive Transcript - Enable basic transcript mode by pressing the escape key You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key Play video starting at ::9 and follow transcript0:09 Hola, nuevamente. En este video voy a hablarles de la forma de modelar dimensiones y los elementos que debemos tener presentes para garantizar cierto nivel de calidad del modelo propuesto. Esto es importante para terminar de afinar la competencia de modelar dimensionalmente y poder utilizarlo en contextos de organizaciones. Recordemos que estamos trabajando en el área de almacenamiento de datos para realizar análisis, el cual, en esta aproximación, corresponde a tener un modelo dimensional que será luego implementado en una base de datos relacional, a lo cual denominamos ROLAP. Hasta este punto, hemos hablado de una serie de criterios que caracterizan los modelos dimensionales y que permiten validar su calidad, es así como el modelo debe tener al menos una tabla central, denominada tabla de hechos que representa un proceso de negocio. Adicionalmente, siempre debemos hablar de una dimensión fecha, que recordemos que en los análisis nos interesa lo que ha pasado a lo largo del tiempo y no solo un punto en el tiempo. El tipo de relaciones entre las dimensiones y la tabla de hechos, es 1-m; un cliente hace varias compras y una compra está relacionada con un único cliente. Las dimensiones que observamos son perspectivas de análisis que usamos como filtro en el momento de hacer consultas, por ejemplo, compras de la categoría de electrodomésticos, entre el 1 de enero del 2022 y el 31 de marzo del 2022. Los atributos de las dimensiones son características propias de la dimensión, como lo es el nombre, la categoría o subcategoría, para el caso de producto. No dependen de otras dimensiones adicionales al tiempo. Ahora, abordemos el tema de cómo modelar una dimensión. Algunos ejemplos de los clásicos de dimensiones son "cliente", que en otros contextos puede ser el estudiante, el proveedor o el paciente, la geografía que puede representar la ubicación del cliente o de una sucursal, el producto o servicio que corresponde a la oferta realizada por una empresa a un cliente. Estas dimensiones deben tener una llave primaria independiente del sistema transaccional, vemos que sobre esta llave queremos control total y no es posible si dependemos de la llave del sistema OLTP, además hay varias formas de manejar la historia de los atributos de las dimensiones, en las cuales no es posible utilizar la llave del OLTP, ya que se tiene más de una fila representando la misma dimensión. Esto último lo entenderemos mejor en el recurso de historia de atributos de una dimensión. Otros atributos comunes son los que vienen de las banderas o los indicadores, en este caso particular, de la dimensión "fecha", observamos los dos últimos y hablamos de "festivo" y "día de la semana". Entonces, fíjense que para una fecha en particular, como 6 de Julio, se podría decir que no es festivo, mientras que para el 1 de mayo si es festivo. Además, puedo entender por ejemplo, si el 1 de mayo fue un día en la semana o de fin de semana, para un año particular en el cual ocurrieron cierto número de ventas. Vemos que también podemos tener múltiples campos de texto, en este ejemplo tenemos la descripción de la fecha, el nombre, el mes fiscal, el día de la semana, o como lo veíamos antes, el saber si una fecha es o no un día festivo. También pueden convivir múltiples jerarquías; en nuestro ejemplo vemos en azul la jerarquía de fecha calendario y en morado la jerarquía de fecha asociada al año fiscal de la organización. Revisemos el concepto de jerarquía para que todos tengamos el mismo lenguaje y recordemos que una jerarquía es una relación de uno a muchos al interior de los atributos de una dimensión, es así como en nuestro ejemplo podemos observar que varios días de la semana conforman un mes, varios meses, un trimestre y cuatro trimestres un año. Ahí estamos mostrando esa relación de uno a muchos, un día pertenece a un único mes calendario, pero un mes tiene múltiples días de la semana asociados. Este concepto de jerarquía puede ser común en algunas de las dimensiones, en particular aplica para el caso de la dimensión geografía, ¿ya sabes cuál es? Adicionalmente, relacionado con el tema de jerarquías, es posible que estés pensando en por qué todo está en la misma tabla de la dimensión fecha y por qué no más bien tener tres tablas; una de días, otra de meses y otra de años en un estilo normalizado, lo que denominamos "Snow Flake" o copo de nieve. En la parte de analítica se recomienda evitar la normalización, principalmente porque puede ser tedioso el cargue y, en general, el manejo puede ser más confuso en el momento de hacer consultas y en la actualidad, aunque la cantidad de almacenamiento es importante, no es grave manejar este tipo de recalificación controlada al interior de una dimensión, es por eso que lo más simple y sencillo es tener toda esa información en la misma dimensión. Pasemos ahora a la dimensión "producto" y veamos que a diferencia de la dimensión "fecha", tenemos una llave subrogada que corresponde a un número consecutivo creado para identificar de forma única un registro de la dimensión. Sin embargo, para mantener la relación con el sistema OLTP, incluimos adicionalmente como atributo la llave natural, que en este caso del producto es el SKU. Además, igual que en la dimensión fecha, identificamos varios tipos de texto, nombre, color, categoría, entre otros. Otra de las recomendaciones es, por ejemplo, incluir columnas adicionales si la llave natural u otras columnas de la fuente de datos tienen semántica. En este caso, vemos que el SKU tiene una semántica, en la cual lo primero que viene es la subcategoría del producto, que en este caso es deportiva, dado que nuestro ejemplo es de ropa y lo manejamos con una columna especial de subcategoría. El segundo elemento que debemos representar es el concepto de medias tobilleras, en el cual queda el nombre del producto. El tercero es el color blanco y, por último, la talla que representa personas que calzan 40. Fíjate que todos esos elementos los podemos extraer y crear nuevas columnas independientes para manejarlos en los análisis, es por eso que a nivel analítico, tendremos una nueva columna por cada concepto, porque en este caso es de interés incluirlas en análisis. Finalmente, y no menos importante, es el hecho de que debemos evitar atributos sin valor, recordemos que es más complejo a nivel de consultas SQL. Siempre se debe diferenciar entre si hay valor o no, con el operador IS NULL y recordemos que estos modelos se caracterizan por su simplicidad para entenderlo y consultarlo. Así que aquí, es preferible colocar un "no aplica" o "sin valor" para atributos, y, en caso de ser necesario, crear, por ejemplo, un producto no identificado para casos con ventas sobre códigos de productos que no logramos saber cuál es. En este contexto, es preferible manejarlo así y evitar a toda costa el tener valores nuevos. Con esto terminamos las ideas generales que debemos tener presentes en el modelado de las dimensiones y espero que, a pesar de que lo que hice aquí fue aplicarlo únicamente para "fecha" y "producto", puedas generalizar. En particular, te invito a pensar en la dimensión "cliente", que es un caso clásico en modelos multidimensionales. Hablemos ahora de calidad de modelos y cómo podemos validarla para mejorarla si es necesario. Recordemos que hay una serie de preguntas que nos ayudan a determinar la calidad de un modelo; esas preguntas van desde saber si una dimensión en realidad es una dimensión o una medida, hasta encontrar inconsistencias a nivel de detalle, cuando tenemos ciertas dimensiones y observamos que algunas medidas no coinciden con esas dimensiones. Ahora quiero incluir nuevas preguntas que van a complementar nuestro proceso de validar la calidad de un modelo. Empecemos desde la pregunta más simple, ¿hay una tabla central? Fíjate que debe existir al menos una, que es la llamada tabla de hechos, la cual relaciona las diferentes dimensiones con la información de las medidas y representa un proceso de negocio. Otro ejemplo clave es la existencia de la dimensión "fecha"; siempre debe existir dado que, en la parte analítica, lo que queremos hacer es análisis de qué ha pasado, para lo cual necesitamos una trayectoria o una historia. Esta dimensión "fecha" está conectada únicamente con las tablas de "hechos", porque recuerda la estructura y toda la información que tiene dicha dimensión, lo veíamos previamente; información de días de la semana, si es festivo o no y esto es útil cuando queremos hacer análisis sobre los hechos. Por ejemplo, en este caso, si quieres saber si los días que se compra más son días de semana o fines de semana o, por ejemplo, si son festivos o no o si son los lunes, los martes, etcétera. Esta semántica no la requerimos si tenemos fechas como atributos de las dimensiones. Imagínate que tenemos la fecha de nacimiento de un cliente, ¿crees que tenemos que saber qué día nació el cliente? ¿Será que nació un fin de semana o en primer trimestre del año? Para el contexto particular de un "retail" no importa y, en términos generales, no se requiere; es así como las atributos de las dimensiones serán de tipo "Date" y no debe existir una relación entre ellos y la dimensión fecha. Otra pregunta fundamental es ver si toda la tabla de hechos tiene al menos una medida. Siempre debe tener al menos una y hay casos que denominamos tablas de hechos Factless, porque el proceso para este nivel de granularidad no tiene medidas. Sin embargo, a pesar de ello, nos inventamos una columna que tenga un valor constante para todas las filas y juegue el rol de medida. Esto lo veremos en otro recurso, pero siempre necesitamos tener medidas y esas medidas son números sobre las cuales podemos hacer operaciones como sumar y que son medidas aditivas o podemos traer el mínimo o el máximo valor, hacer promedio, entre otras operaciones. Otra pregunta es la relacionada con los atributos multivalor, ¿es permitido en estos modelos tener atributos multivalor? Multivalor es tener, por ejemplo, listas como de servicios, productos de clientes en un solo atributo. La respuesta es: todo atributo que tengamos en este modelo va a ser monovalor, porque nos interesa su análisis y poderlo consultar de forma simple. Por último, tenemos una ñapa que es, ¿hay relación directa por medio de las llaves foráneas entre las tablas de hecho? Lo que quiere decir es que las tablas de hecho tendrían una llave primaria y podrían tener llaves foráneas entre ellas. En este nivel de modelado y, por razones pedagógicas, es preferible pensar que no, lo que significa que la tabla de hechos no tiene llave primaria, a pesar de que para efectos de carga de datos y facilidad en proceso puede existir entre otras razones. Solo llaves foráneas de dimensiones y medidas vamos a manejar acá. De esta forma, cuando estamos modelando es preferible no manejar la llave primaria para evitar esa tentación de conectar las tablas de hechos directamente entre ellas y no por medio de las dimensiones, que es como debe realizarse. Llegamos al final y es importante recordar que en este video hablamos de la forma de modelar dimensión, ejemplificando el caso de la dimensión de fecha, que es especial, y la de producto. Adicionalmente, vimos algunas preguntas para validar la calidad del modelo. Espero que estos puntos puedan ayudarte a fortalecer la competencia de modelar multidimensionalmente, validar la calidad de los modelos y tener como resultado mejores modelos. Hasta pronto. Transcript language: Español : Added to Selection. Press [⌘ + S] to save as a note

Transcript language: Español

# de atributos de las dimensiones Modelado de datos y ETL 05-2022 María del Pilar Villamil Giraldo Modelado de datos y ETL 2 I. Comentarios generales Adaptación realizada por María del Pilar Villamil, del capítulo Capítulo 5 “Procurement” de la referencia [1], disponible en versión digital en la biblioteca de la universidad (biblioteca.uniandes.edu.co). Con el fin de resolver dudas que puedan quedar con esta lectura, puedes profundizar los conceptos vistos en esta lectura en la referencia [1]. En particular: • Tipos manejo de historia en dimensiones: “Slowly Changing Dimension Basics” que está organizado en los tipos individuales en las páginas 148- 159 • Combinación de tipos de manejo de historia: “Hybrid Slowly Changing Dimension Techniques” páginas 159 – 165. II. Detalle de la lectura Hasta este punto hemos trabajado pensando en que la información que cambia a lo largo del tiempo y que es de interés para ser analizada, está siempre almacenada en una tabla de hechos. Sin embargo, esto es una simplificación de la realidad. En muchas ocasiones queremos hacer análisis sobre las dimensiones y saber la versión exacta de la dimensión en el momento de la ocurrencia del hecho o en distintos momentos a lo largo del tiempo. Uno de los ejemplos clásicos está relacionado con la dimensión cliente. El comportamiento del cliente varía a lo largo del tiempo. No es lo mismo cuando hablamos de un cliente en su época de estudiante universitario, que cuando está pensionado. Esa información es fundamental para las organizaciones, ya que les permite comprender el comportamiento de un cliente en un momento de su vida para, por ejemplo, personalizar las ofertas que le hace y de esta manera aumentar la probabilidad de una respuesta positiva. Es así, como se han propuesto diferentes formas de manejar dicha historia y vamos a revisar en detalle cada una de ellas. Estas formas se Modelado de datos y ETL 3 conocen con los nombres de tipos para el manejo de cambios lentos de una dimensión (Slowly Changing Dimension Types), van desde el tipo 2 hasta tipo 7, y corresponden a propuestas realizadas por el grupo Kimball [1]. Aunque no todos los tipos son utilizados con la misma frecuencia, el conocerlos puede dar pautas para modelar datos siendo conscientes de la flexibilidad, eficiencia entre otros criterios. Se preguntarán ¿qué pasó con los tipos 0 y 1? En el tipo 0 la decisión es que los atributos no cambian, así que no se lleva historia. Mientras que en el tipo 1 la decisión es que, aunque cambie el atributo, no es importante para el negocio llevar la historia. Este es el caso de cambio de nombre de un cliente. Fíjense que puede cambiar, sin embargo, la empresa decide sobrescribir el nombre del cliente y no llevar historia del cambio. A continuación, como se muestra en la tabla cliente (Dimensión), se trabaja con el ejemplo de la dimensión cliente que representa una dimensión con su identificador para el repositorio analítico, tipo de identificación, número de identificación, nombre y la última versión de la categoría del cliente, sus ingresos y estado civil. Sin embargo, la empresa tiene la intención de manejar historia de su categoría, ingresos y estado civil. Es así como para ilustrar las características de los tipos del 2 al 7 se toma como ejemplo, la cliente María Castillo con dos versiones específicas: • Versión1: Categoría novato, con ingresos de 5’ (cinco millones), estado civil casado. Versión válida desde su afiliación 15-01-2007 hasta el 28-06-2017. • Versión2: Categoría VIP, con ingresos de 10’ (diez millones), estado civil casado, a partir del 29-06-2017. IdCliente TipoIdentificacion NumIdentificacion Nombre Categoría(\*) Ingresos(\*) EstadoCivil(\*) PK NN NN NN 11 TI 124567 Maria Castillo VIP 10’ Casado 24 CC 456789 Juan Perez Novato 5’ Soltero Cliente (Dimensión) Modelado de datos y ETL 4 Tipo 2: Basada en la creación de nuevas filas La idea detrás de este manejo es crear una nueva fila que represente una nueva versión de la dimensión a la cual se le lleva historia. Esta es la estrategia más sencilla y común en los modelos dimensionales. Sin embargo, puede llevar a tener un gran número de filas en una dimensión. Por lo tanto, es importante tener presente posibles problemas de desempeño para la carga de datos en estas dimensiones y en las consultas sobre ellas. Estos puntos deberán ser analizados para tener certeza de que esta es la mejor alternativa para el manejo de historia en la dimensión que están analizando. El manejo de la historia implica la creación de nuevas columnas que representan las fechas en las cuales estuvo vigente la versión representada por la fila. En el ejemplo, las vemos representadas por las columnas fechaVigenciaInicial (FeVigIni), y la fechaVigenciaFinal (FeVigFin). Adicionalmente, para encontrar fácilmente la versión actual de la dimensión, se crea la columna Actual (A) que es una bandera con valores de ‘S’ (Si) o ‘N’(No). Con respecto a estas columnas y como una regla de los modelos dimensionales, en términos generales, no se permiten celdas con valores nulos, ya que estos valores implican sentencias de consulta más complejas que afectan el tiempo para su desarrollo, mantenimiento y eficiencia en su ejecución. Es por ello, que se prefiere, por ejemplo, para el caso de la fecha de vigencia final, crear un valor en el futuro (i.e., 01-01-5000). Fíjense como comparado con la tabla de Cliente original, tenemos dos filas para representar a María Castillo y las tres nuevas columnas al final de la tabla. En este punto, es el momento de entender mejor la necesidad de tener llaves subrogadas. Como ven, la llave del sistema transaccional que en este caso es el tipo de identificación y el número de identificación, no nos sirve como llave de la dimensión, ya que debemos crear nuevas filas para representar las distintas versiones del cliente. Es así como se propone crear una nueva llave propia del sistema analítico que sea sencilla, la cual se denomina llave subrogada y corresponde a un número secuencial. Esta llave es la que se relaciona con la tabla de hechos y en este ejemplo, corresponde a la columna IdCliente. Modelado de datos y ETL 5 IdCliente TipoIdentificacion NumIdentificacion Nombre Categoría(\*) Ingresos(\*) EstadoCivil(\*) FeVigIni FeVigFin A PK NN NN NN NN NN NN 11 TI 124567 Maria Castillo Novato 5’ Casado 15-01-2007 28-06-2017 N 25 TI 124567 Maria Castillo VIP 10’ Casado 29-06-2017 01-01-5000 S 24 CC 456789 Juan Perez Novato 3’ Soltero 02-02-2021 01-01-5000 S Tipo 3: Basada en la creación de nuevas columnas El tipo 3 está basado en mantener como número de filas de una dimensión, las instancias propias de la dimensión. Lo que en este caso corresponde a tener tantas filas, como clientes tiene la empresa y, crear todo el manejo de historia como nuevas columnas con los valores históricos. Es así, como debemos conocer el número de cambios que queremos mantener, con el fin de crear ese número de columnas asociadas al atributo al que se le quiere llevar la historia. Para efectos del ejemplo, vamos a suponer que queremos manejar dos cambios de los atributos categoría e ingreso. Es por esta razón, que en la tabla de cliente-tipo3 vemos asociados a estos atributos dos columnas: CategoriaActl y CategoriaAnt, las cuales representan respectivamente la categoría actual y la anterior. De igual manera, para la columna ingresos, creamos las columnas IngresosActl e IngresosAnt, siguiendo la misma lógica que para el atributo categoría. Como observamos en el ejemplo, el número de filas se conserva igual que en el caso donde se manejaban los valores actuales del cliente y no se tenía historia. Este tipo 3 es poco frecuente en modelos multidimensionales. Sin embargo, su potencial está en combinarlo con otros tipos, como por ejemplo, el tipo 2. Cliente-tipo2 (Dimensión) Modelado de datos y ETL 6 IdCliente TipoIdentificacion NumIdentificacion Nombre CategoríaAnt CategoriaActl IngresosAntIngresosActlEstadoCivil(\*) PK NN NN NN 11 TI 124567 Maria Castillo Novato 1 (VIP) 5’ 10’ Casado 24 CC 456789 Juan Perez Novato 2 (Novato)3’ 3’ Soltero Tipo 4: Basada en la creación de nuevas tablas En el tipo 4, los supuestos que se tienen son que el número de cambios de un atributo es variable y desconocido y el tamaño de la dimensión sin historia, supera el millón de registros, así que crear nuevas filas y tener la opción tipo 2 puede resultar costoso en almacenamiento y llevar a ineficiencias en el momento de consultar la información. La propuesta entonces es tomar un subconjunto de columnas de la dimensión original y crear una nueva tabla con dichas columnas, la cual denominamos minidimensión. Las columnas a seleccionar para crear la nueva tabla, son columnas que cambian frecuentemente y típicamente en el mismo momento. En este caso vamos a crear una minidimensión denominada PerfilCliente como se muestra en la tabla Cliente-tipo4. Cliente-tipo3 (Dimensión) Modelado de datos y ETL 7 IdMini Categoría RangoIngresos EstadoCivil PK NN NN NN 1 Novato 2’-5’ Casado 2 VIP 5’-12’ Casado 3 Novato 2’-5’ Soltero IdCliente TipoIdentificacion NumIdentificacion Nombre PK NN NN NN 11 TI 124567 Maria Castillo 24 CC 456789 Juan Perez Fijémonos como en la minidimensión PerfilCliente, no está la identificación del cliente. Únicamente valores de las columnas que cambian en el tiempo y descritas en rangos. Esto se hace para aumentar la probabilidad de compartir una fila por varios clientes, en nuestro ejemplo y, de reducir el número de filas de dicha tabla. Lo que se espera es que esa tabla tenga menos registros que el número de instancias de la dimensión de la cual se originó la minidimensión. PerfilCliente (Minidimensión) Cliente-tipo4 (Dimensión) Modelado de datos y ETL 8 La pregunta que surge es ¿en qué momento relacionamos la minidimensión con la dimensión? Esto puede ocurrir en dos momentos. La primera opción, es en la tabla de hechos, por ejemplo, las ventas asociadas a dicho cliente. En ese caso tendremos la tabla HechoVentas con una columna para representar el cliente y otra el perfil del cliente en el momento del hecho, como se muestra a continuación. IdCliente IdPerfilCliente … ValorVenta FK FK, NN NN 11 1 … 500. 24 2 … 300. Esto permite conocer la versión del perfil del cliente que participó en la venta para nuestro ejemplo. Adicionalmente, hay ocasiones donde se desea llevar la historia de todos los cambios independiente de los hechos, en este caso de las ventas realizadas por el cliente. Es así como creamos la tabla HistoriaPerfilCliente y llevamos la historia con el tipo 2, como lo representa la tabla HistoriaPerfilCliente (TablaHecho). HechoVentas (Tabla de hechos) Modelado de datos y ETL 9 IdCliente IdPerfilCliente FeVigIni FeVigFin Cambio FK FK, NN NN NN 11 1 15-01-2007 28-06-2017 1 11 2 29-06-2017 01-01-5000 1 El tipo 4 es otro de los tipos frecuentemente utilizados en modelos dimensionales. Tipo 5: Basada en la creación de nuevas tablas y nuevas columnas Este tipo 5 es una modificación del tipo 4, donde mantenemos la minidimensión y la acompañamos de una extensión que en inglés se conoce como outrigger. Esta extensión para nuestro ejemplo contiene la versión actual de los perfiles de clientes y en la dimensión cliente, tenemos un atributo adicional, que contiene el valor de la versión actual del perfil. Esta extensión en realidad es una vista asociada a la consulta Q1: Select \* from PerfilCliente, que contiene las mismas columnas y filas de la minidimensión original, en este caso PerfilCliente. HistoriaPerfilCliente (TablaHecho tipo factless) Modelado de datos y ETL 10 IdMini Categoría RangoIngresos EstadoCivil PK NN NN NN 1 Novato 2’-5’ Casado 2 VIP 5’-12’ Casado 3 Novato 2’-5’ Soltero IdCliente TipoIdentificacion NumIdentificacion Nombre IdMiniPerfilActual PK NN NN NN FK 11 TI 124567 Maria Castillo 2 24 CC 456789 Juan Perez 3 Tipo 6: Basada en la creación de nuevas columnas y nuevas filas El tipo 6 contiene elementos del tipo 3, 2 y 1. Es así como adicionaremos asociadas al tipo 3, columnas para manejar la versión anterior y la actual o las versiones que se quieran de un atributo y adicionalmente, las tres columnas para representar el manejo de tipo 2. Como vemos en la tabla Cliente-tipo6 (Dimensión) para el caso de María Castillo, ella tiene dos filas asociadas y en ambas fechas vemos que en las Cliente-tipo5 (Dimensión) VPerfilActualCliente (Extensión) Modelado de datos y ETL 11 columnas CategoriaActl e IngresosActl tenemos el mismo valor para las dos filas. Es en este punto donde vemos el tipo 1. En términos generales podemos afirmar que para toda fila asociada a un mismo cliente en sus valoresActuales se tendrá el mismo valor. De igual manera observamos que en cada versión del cliente, siempre tenemos la versión de un momento particular y su versión actual. Este tipo tampoco es frecuente en los modelos dimensionales. Momento 1: IdCliente TipoIdentificacion NumIdentificacion Nombre CategoríaHist CategoriaActl IngresosHist IngresosActl EstadoCivil FeVigIni FeVigFin A PK NN NN NN 11 TI 124567 Maria Castillo Novato Novato 5’ 5’ Casado 15-01-2007 01-01-5000 S 24 CC 456789 Juan Perez Novato Novato 3’ 3’ Soltero 21-10-2020 01-01-5000 S Momento 2: IdCliente TipoIdentificacion NumIdentificacion Nombre CategoríaHist CategoriaActl IngresosHist IngresosActl EstadoCivil FeVigIni FeVigFin A PK NN NN NN 11 TI 124567 MariaCastillo Novato VIP 5’ 10’ Casado 1-01-2007 28-06-2017 N 24 CC 456789 Juan Perez Novato Novato 3’ 3’ Soltero 21-10-2020 01-01-5000 S 25 TI 124567 MariaCastillo VIP VIP 10’ 10’ Casado 29-06-2017 01-01-5000 S Cliente-tipo6 (Dimensión) Modelado de datos y ETL 12 Tipo 7: Basada en la creación de una nueva “dimensión” y de nuevas columnas para tener dos versiones de la dimensión en el mismo momento Finalmente, tenemos el tipo 7, el cual combina el tipo 2 para el manejo de historia, con la información de la última versión de la dimensión de análisis que se almacena en una nueva dimensión. La nueva dimensión, por ejemplo, VClienteActual, en realidad es una vista asociada a la consulta Q1 Select \* from Cliente Where actual=’S’, que contiene las mismas columnas de la dimensión original tipo 2 y como filas solo las versiones actuales. IdCliente TipoIdentificacion NumIdentificacion Nombre Categoría(\*) Ingresos(\*)EstadoCivil(\*) FeVigIni FeVigFin A PK NN NN NN NN NN NN 25 TI 124567 Maria Castillo VIP 10’ Casado 29-06-201701-01-5000 S 24 CC 456789 Juan Perez Novato 3’ Soltero 02-02-202101-01-5000 S Este tipo podemos verlo en el momento de registrar un hecho, ya que la tabla de hechos tendrá adicionalmente como columnas, la llave del cliente (dimensión involucrada) y la llave del cliente Actual (nueva dimensión). Este último contiene una llave foránea a la llave durable, que es una llave subrrogada, propia de la bodega de datos, creada para identificar instancias de una misma dimensión, en este caso un cliente. Esta llave garantiza que a pesar de que tenga varias filas en la dimensión y que su llave del sistema transaccional haya cambiado este valor se mantiene. Veamos como luce entonces una tabla de hechos que maneja la historia con este tipo 7. Este estilo de representación facilita análisis que involucren al mismo tiempo, el valor de la dimensión en el momento del hecho, para compararlo con el valor actual en esa dimensión. VClienteActual (Dimensión) Modelado de datos y ETL 13 IdCliente DurableIdClienteActual … ValorVenta FK FK, NN NN 11 38 … 500. 24 26 … 300. 25 38 … 600. III. Conclusión Al final de esta lectura la invitación es a pensar en las diferentes formas de modelar datos para el manejo adecuado de historia de atributos de una dimensión. Con respecto a esta lectura es importante tener claro dos grandes aprendizajes: 1. Hay cambios en las dimensiones que pueden ser de interés para los análisis, así que hay que determinar la forma de manejarlos. 2. La combinación del manejo por medio de columnas cuando se conoce el número de cambios a guardar y de filas cuando este número es desconocido es quizás una idea valiosa de modelar datos. HechoVentas (Tabla de hechos) Modelado de datos y ETL 14 3. El poder tener de forma simultánea la versión de la instancia de la dimensión (e.g., un cliente específico), en el momento del hecho y la versión actual facilitará los análisis. Estos aprendizajes con seguridad podrás aplicarlos también, en contextos distintos al ROLAP siones degeneradas Tipos de medidas Arquitectura – Data Warehouse Inspirado en: http://www.rhinoafrica.com/ Caso2 Rhino1 Director de mercadeo ¬ ¿Cuál fue el tipo de habitación preferida por los huéspedes en el último trimestre? ¬ ¿De qué países hacen mayor número de reservas? ¬ ¿Cuántas habitaciones y de qué tipo se reservan por un cliente? Identificación de errores Error 6: Inconsistencia entre el nivel de granularidad de la tabla de hechos y los atributos. ¿Qué pasa si la reserva realizada incluye varias habitaciones para hospedar el total de personas asociadas a la reserva? Solución 1: Manejar la habitación como un grupo de habitaciones. Explicado en el video de errores de modelos dimensionales. Solución Caso Rhino1 ¿Qué pasa si la reserva realizada incluye varias habitaciones para hospedar el total de personas asociadas a la reserva? Solución 2: Propuesta: Manejar una relación de 1 a m y adicionar una dimensión degenerada. Solución Caso Rhino1 ¿Qué pasa si la reserva realizada incluye varias habitaciones para hospedar el total de personas asociadas a la reserva? ¿Cada habitación puede ser pagada con diferentes tipos de tarjeta? Propuesta: Manejar una relación de 1 a m y adicionar una dimensión degenerada. Veamos el ajuste en el modelo. Solución Caso Rhino1 NumReserva(DD) ValorHabitacion FechaCheckIn FechaCheckOut NumReserva (DD) Actual IdPlan IdCliente IdFecha ¿Qué pasa si la reserva realizada incluye varias habitaciones para hospedar el total de personas asociadas a la reserva? ¿Cada habitación puede ser pagada con diferentes tipos de tarjeta? Propuesta: Manejar una relación de 1 a m y adicionar una dimensión degenerada. Veamos en un ejemplo cómo quedaría la tabla de hechos.. Solución Caso Rhino1 FechaCheckIn FechaCheckOut NumReserva (DD) Actual IdPlan IdCliente IdFecha Clien te​ Fecha CheckIn Fecha CheckOut Sucursal​ Num Reserva Habitacion​ … Numero​ Personas​ Valor​ Habitacion 11​ 20220320​ 20220323​ 1​ 101 1 (TV)​ 1​ 200.​ 11 20220320​ 20220323​ 1​ 101 2 (Doble)​ 2​ 150.​ NumReserva (DD) Hotel Cliente Plan Habitacion NumReserva (DD) NumeroPersonas ValorHabitacion TipoTarjeta Este número de reserva NO representa la llave primaria de la tabla HechoReserva!!! Aunque el nombre puede confundir. Agenda Grupos de atributos y tipos de medidas Dimensiones degeneradas Tipos de medida Tipos de medidas Hotel Cliente Plan Habitacion NumReserva (DD) NumPersonas ValorHabitacion TipoTarjeta FechaCheckIn FechaCheckOut Actual IdPlan IdCliente IdFecha Cliente​ Fecha CheckIn Fecha CheckOut Sucursal​ Num Reserva​ Habitacion …​ Num Personas​ Valor​ Habitacion 11​ 20220320​ 20220323​ ​202 101​ 1 (TV)​ 1​ 200.​ 11 20220320​ 20220323​ 202​ 101​ 2 (Doble)​ 2​ 150.​ ¿Es posible sumar las medidas en las diferentes dimensiones? Si: Medida aditiva No en ninguna: Medida no aditiva En algunas: Medida semi-aditiva Continuemos con el ejemplo de reservas y analicemos las medidas de HechoReserva. Tipos de medidas – Aditivas Cliente​ Fecha CheckIn Fecha CheckOut Sucursal​ Num Reserva​ Habitacion …​ Num Personas​ Valor​ Habitacion 11​ 20220320​ 20220323​ 202 101​ 1 (TV)​ 1​ 200.​ 11 20220320​ 20220323​ 202​ 101​ 2 (Doble)​ 2​ 150.​ ¿Es posible sumar las medidas en las diferentes dimensiones? ⎫Si: Medida aditiva. Aplica al caso de la medida NumPersonas- veamos X No en ninguna: Medida no aditiva X En algunas: Medida semi-aditiva ¿Cuántas personas de acuerdo con las reservas van a estar en el fin de semana del 19 de marzo del 2022 en la sucursal de Montreal? Select Sum(NumPersonas) From HechoReserva Where Fecha between ‘2022/03/19’ and ‘2022/03/20’ and Sucursal= ‘202’ /\*Montreal\*/ ¿Esta consulta SQL es correcta? Fíjate que sí, y mira que se utiliza la suma sobre la medida NumPersonas en las dimensiones Fecha y Sucursal, la cual se puede extender a otras dimensiones y sigue funcionando. Tipos de medidas – No aditivas ¿Es posible sumar las medidas en las diferentes dimensiones? X Si: Medida aditiva ⎫No en ninguna: Medida no aditiva. Aplica al caso de la medida %Dcto. Veamos. X En algunas: Medida semi-aditiva ¿Qué porcentaje de descuento en el pago de las habitaciones se aplicó a las reservas del fin de semana del 19 de marzo del 2022 en la sucursal de Montreal? Select Sum(%Dcto) From HechoReserva Where Fecha between ‘2022/03/19’ and ‘2022/03/20’ and Sucursal= ‘202’ /\*Montreal\*/ Cliente​ Fecha​ CheckIn Fecha​ CheckOut Sucursal​ NumReserva Habitacion …​ Numero​ Personas​ Valor​ Habitacio n %Dcto 11​ 20220320​ 20220323​ 202 101​ 1 (TV)​ 1​ 200.​ 5% 11 20220320​ 20220323​ 202 101​ 2 (Doble)​ 2​ 150.​ 10​% Esa suma de descuento fíjate que no tiene sentido. En el ejemplo hablaríamos qué, aplicamos un 15% de descuento y eso no es así. Podríamos decir el máximo o mínimo descuento aplicado, pero no sumarlos, por eso está tachado Sum(%Dcto). Esto aplica a las diferentes dimensiones. Tipos de medidas – Semi-aditivas ¿Es posible sumar las medidas en las diferentes dimensiones? X Si: Medida aditiva X No en ninguna: Medida no aditiva ⎫En algunas: Medida semi-aditiva. Aplica para la medida Ocupada de otro Hecho: HechoOcupación ¿Cuántas habitaciones estuvieron ocupadas durante el fin de semana del 19 de marzo del 2022 en la sucursal de Montreal? Nos daría 3 en el ejemplo y en realidad fueron dos: las habitaciones 1 y 2. Select Sum(Ocupada) From HechoReserva Where Fecha between ‘2022/03/19’ and ‘2022/03/20’ and Sucursal= ‘202’ /\*Montreal\*/ Fecha​ Sucursal​ Habitacion Ocupada 20220319​ 202 1 (TV)​ 1 20220320​ 202 1 (TV)​ 1 20220320​ 202 2 (Doble)​ 1 Pero si hacemos esta consulta: ¿Cuántas habitaciones fueron utilizadas durante el 19 de marzo del 2022 en la sucursal de Montreal? Nos daría 2 en el ejemplo, las habitaciones 1 y la 2 y es correcto. Select Sum(Ocupada) From HechoReserva Where Fecha = ‘2022/03/19’ and Sucursal= ‘202’ /\*Montreal\*/ HechoOcupación Para una fecha puntual , SI tiene sentido. Pero No para diferentes fechas. Tipo de tabla de hechos – Factless Factless Tabla de hechos sin medidas, dado que el hecho es la misma medida. En esta tabla NO estarán las habitaciones que estuvieron sin ocupar. Dado que toda tabla de hechos debe tener una medida, se propone adicionar una columna con un valor constante para todas las filas. En el ejemplo: ◊ Ocupada=1 2022 María del Pilar Villamil Giraldo Modelado de datos y ETL 2 I. Introducción En esta lectura se presenta una serie de conceptos complementarios a los vistos en recursos previos sobre el modelado de datos multidimensional. Se presenta una primera sección con varias definiciones complementarias sobre el modelado dimensional y al final una conclusión. II. Algunos conceptos Manejo de llaves foráneas: recordemos que en este mundo del modelado multidimensional las consultas a los datos deben ser fáciles de hacer. Es así como se recomienda evitar campos sin valor (en NULL), en particular, en aquellos que son llaves foráneas o utilizados regularmente en consultas como es el caso de las fechas. Por esta razón se sugiere por un lado crear en las dimensiones susceptibles de tener valores nulos, a nivel de la tabla de hechos, un registro que indique sin valor o valor desconocido. El identificador de ese registro será el que se utilice en el proceso de carga de los datos. Este sería el ejemplo para el caso de un cliente no identificado en una compra. Nótese como queda referenciado en la tabla de hechos en caso de ser requerido (El color rojo en la tabla de Cliente y en los campos IdCliente y DurableIdClienteActual representan esta situación). Cliente IdCliente TipoIdentificacion NumIdentificacion Nombre Categoría(\*) Ingresos(\*) EstadoCivil(\*) PK NN NN NN -1 CC 111111 No identificado SinCategoría SinEstadoCivil 11 TI 124567 Maria Castillo VIP 10’ Casado 24 CC 456789 Juan Perez Novato 5’ Soltero HechoVentas Modelado de datos y ETL 3 IdCliente DurableIdClienteActual … ValorVenta FK FK, NN NN 11 38 … 500. 24 26 … NULL -1 -1 … 600. Por otro lado, a nivel de manejo de fechas, es preferible tener un comodín para manejar posibles valores faltantes en dichos campos (e.g.,’ 01- 01-5000’). Es así, como para el manejo de historia de atributos de las dimensiones, se sugiere crear una fecha muy lejana y así representar que la versión de la dimensión que se quiere manejar es la vigente, razón por la cual su fecha fin de vigencia no se conoce. Esto se puede ver en la siguiente tabla, haciendo referencia a los registros de los clientes 24 y 25, para los cuales las características de Categoría, ingresos y estado civil son la versión actual conocida a nivel de la bodega de datos. Es en estos registros, donde el valor de FeVigFin tiene el valor del comodín (verlo en color rojo). ClienteManejoTipo2Historia IdCliente TipoIdentificacion NumIdentificacion Nombre Categoría(\*) Ingresos(\*) EstadoCivil(\*) FeVigIni FeVigFin A PK NN NN NN NN NN NN 25 TI 124567 Maria Castillo VIP 10’ Casado 29-06-2017 01-01-5000 S 24 CC 456789 Juan Perez Novato 3’ Soltero 02-02-2021 01-01-5000 S Por otro lado, las medidas de las tablas de hecho también pueden tener valores nulos. En ese caso se recomienda dejar NULL en dichos campos y dejar que sea la función utilizada para agregar los valores la que se encargue del manejo de estos valores. Esto sería el caso de sumar valor de venta en la tabla HechosVentas. En ese caso el operador SUM(ValorVenta) se encargaría de ese manejo. Este operador, ignora esos valores NULL y suma únicamente las filas con valor, que coincide con el comportamiento que se desea. Modelado de datos y ETL 4 Dimensión degenerada: Corresponden a dimensiones vacías que contienen únicamente un identificador y están típicamente relacionadas con números transacciones, referentes a relaciones de encabezados y detalles (1◊n) en el mundo transaccional. Esto sería el caso de encabezados de factura y detalles de factura. Esos números transaccionales que para el ejemplo sería el identificador de la factura, permite a nivel del detalle de la misma identificar todos los productos adquiridos bajo una misma factura. Este concepto lo encuentra detallada en la lectura de dimensiones degeneradas y tipos de medidas. Dimensión conforme: A nivel de un mismo proceso de negocio, pueden existir diferentes tablas de hecho para representar generalmente, diferentes niveles de granularidad. Por ejemplo, ventas a nivel del detalle de la transacción realizada por un cliente, en una fecha, una hora, sucursal y un conjunto de productos, o ventas mensuales por sucursal, donde se agrupan las ventas detalladas y en la cuales las dimensiones de hora y cliente no estarían en el hecho. Es así como dichas tablas de hecho en caso de compartir las dimensiones, como sería el caso de sucursal y fecha en nuestro ejemplo, tendrían que referenciar la misma dimensión, en nuestro caso representada por la misma tabla, ya sea la de Sucursal o la de Fecha. Veamos ahora otra situación, en la cual no se trata de diferentes tablas de hecho asociadas al mismo proceso de negocio, si no tablas de hecho para diferentes procesos de negocio. Este es el caso de ventas y envíos de productos. Estos dos procesos generan sus propios datos y comparten la dimensión Cliente. Sin embargo, para el caso de la venta del cliente se tiene un identificador para referenciarlo, mientras que a nivel de envíos de producto se tendrá la dirección del cliente. Fíjense que a pesar de que conceptualmente hablamos de cliente en los dos casos, puede ser tablas distintas y en casos específicos tener campos distintos. Lo importante en estas situaciones, es que si estas dimensiones comparten atributos estos sean conformes. Lo que quiere decir que deben tener el mismo nombre, la misma representación en nuestro caso a nivel del campo de la base de datos y el mismo contenido para los clientes compartidos. Si son idénticas conceptualmente, se recomienda copiar la tabla entre los diferentes procesos que la comparten, en caso por ejemplo de estar en repositorios diferentes. Este tipo de dimensiones se conocen también como dimensiones compartidas, dimensiones maestras o dimensiones comunes. Esquema de copo de nieve (Snowflake) Este concepto aplica a la normalización de las dimensiones. Es así como en cambio de tener una dimensión por ejemplo de Cliente (ver en la figura A. que se muestra a continuación de la entidad Cliente que dice RECOMENDADO), donde se tiene la descripción del tipo de identificación, Modelado de datos y ETL 5 la categoría del cliente y su estado civil, en la normalización que se muestra en la figura B, se tendrán adicional a la tabla de cliente, tablas para tipo de identificación, categoría y estado civil. Esta representación en copo de nieve, aunque es correcta, no se recomienda en estos sistemas ya que implica diferentes joins entre las dimensiones para reconstruir la información completa del concepto que representa la dimensión, en este caso Cliente. Esto es algo contrario a lo manejado en el mundo transaccional, por eso es importante mencionarlo. Los problemas relacionados con esta replicación de datos son responsabilidad de los procesos ETL. A. Esquema desnormalizado (Recomendado) B. Esquema copo de nieve (Normalizado) Modelado de datos y ETL 6 III. Conclusión Estos conceptos son importantes a nivel de manejo de datos. Recuerde que aun si está en modelos de datos distintos al relacional, el reflexionar sobre estas situaciones puede ayudarlo a tomar decisiones de forma previa a los problemas que se pueden presentar durante la consulta de los datos, debido a decisiones de modelar y almacenar los datos. Si estás interesado en complementar la lectura para mejor tu comprensión de los modelos multidimensionales, puedes revisar los capítulos 3,4 y 5 de la referencia [2 extracción, transformación y carga (ETL) incremental Modelado de datos y ETL 05-2022 María del Pilar Villamil Giraldo Modelado de datos y ETL 2 I. Detalle de la lectura Hasta este punto hemos hablado de procesos de ETL iniciales, en los cuales en la bodega o almacén de datos no hay información cargada. Sin embargo, esto ocurre muy pocas veces en los procesos de carga, típicamente cuando se inicia la carga de la bodega o parte de ella. El escenario más frecuente está relacionado con los procesos incrementales, en los cuales ya se tiene información en la bodega y donde es necesario adicionar o ajustar los datos existentes. Es así como una de las características del diseño del modelo de datos debe estar centrado en reducir el número de actualizaciones de datos, dado que es un proceso costoso en tiempo y recursos, y buscar opciones para que los cambios se conviertan en inserciones de datos. En los tipos de manejo de historia de los atributos de dimensiones vistos previamente, identificamos los tipos 4, 5 y 7 con esta filosofía de reducir actualizaciones en los datos previamente cargados. Sin embargo, el tipo 2, uno de los más utilizados, tiene explícito el hecho de la actualización, como se muestra en la figura 1, en particular en el paso 2.1.1, donde se realiza la actualización del registro anterior. Figura 1. Proceso para el manejo tipo 2 de atributos de dimensiones El diseño de este proceso de ETL, depende también de la forma como reportan las fuentes de datos. Con respecto a esto, se identifican tres formas diferentes de reportar los datos de una dimensión: Modelado de datos y ETL 3 1. Escenario 1- Envío de la última versión de los registros de las dimensiones: Esta opción es la más frecuente, si se trata de fuentes de sistemas OLTP, debido a que en pocas ocasiones se permite intervenir dichos sistemas para poder reportar solo los cambios. 2. Escenario 2- Envío de la versión anterior y la actual de los registros de las dimensiones: Esta opción facilita el proceso ETL, si efectivamente se reporta de esta manera en los casos en los cuales se presentaron cambios entre el último reporte de datos entregado a la bodega y el nuevo reporte. 3. Escenario 3- Envío de todos los cambios de las dimensiones: Esta opción si se reportan todos los cambios realizados entre el último reporte de datos entregado a la bodega y el nuevo reporte, aporta en el análisis histórico. Sin embargo, es el menos frecuente, cuando se habla de fuentes de sistemas OLTP. De acuerdo con estos elementos a considerar en procesos ETL es importante pensar en realizar acciones en masa y no por registro. De esta forma, el comienzo del proceso inicia con un cruce de todo el archivo que representa los datos del sistema OLTP con el supuesto del escenario 1 de envío de información y se compara con lo existente en la bodega de datos para generar una nueva columna que indica: 1. Etiqueta 1: Se trata de un registro nuevo (Paso 1.1 en la figura) 2. Etiqueta 2: Es un registro que ya existe y cambia (Pasos 2.1.1 y 2.1.2) 3. Etiqueta 3: Es un registro que ya existe y no cambia (Paso 2.2) Recordemos que este proceso ocurre en el área de staging. Una vez se tiene la nueva columna, ya es posible proceder a actualizar los registros con etiqueta 2 y luego se realizan las inserciones relacionadas con la etiqueta 1. Utilicemos el ejemplo de la dimensión Cliente, que se presenta en la figura 2, para ilustrar el proceso descrito previamente: Figura 2. Dimensión Cliente Modelado de datos y ETL 4 Supongamos, que tenemos información solo de dos clientes y centrémonos en "María Castillo”, como se muestra en la siguiente tabla. Archivo reportado del sistema OLTP Cédula nombre ingresos categoria 52345845 Maria Castillo 10' VIP 79388589 Juan Perez 3' VIP Fíjense que con respecto a lo que existe en la bodega de datos, que se muestra en la figura 2, los datos de “María Castillo” cambian, los ingresos pasan de 5’ a 10’ y la categoría de Novato a VIP. Es así como si seguimos el proceso planteado para el manejo de historia tipo 2, nos quedaría una nueva tabla con la siguiente información en la que se resaltan las nuevas columnas que se requieren y los datos que tendrían: TempCliente Cédula nombre ingresos categoria idCliente Etiqueta 52345845 Maria Castillo 10' VIP 11 2 79388589 Juan Perez 3' VIP 25 3 Cabe resaltar que el idCliente se requiere ya que no manejamos como llave primaria de la dimensión, la misma llave del sistema transaccional. Es así como utilizamos la cédula, que es el idTransaccional en la bodega, para identificar a los clientes, en particular a “María Castillo”, comparar su valor de ingresos y categoría, y así determinar si cambió o no. Una vez con estas etiquetas, se procede a insertar todos los registros de la nueva tabla, que tengan etiqueta 1 y 2, con una sentencia como esta: INSERT INTO dimCliente (idCliente, idTransaccional, nombre, ingresos, categoría, fechaInicio, fechaFin, actual) SELECT idClienteGenerado, cedula, nombre, ingresos, categoría, FECHA\_CARGA, FECHA\_MAX, ‘S’ FROM TempCliente WHERE etiqueta in (1,2) En la sentencia SQL de inserción previamente ilustrada, se evidencia el uso de idClienteGenerado, FECHA\_CARGA, FECHA\_MAX, ‘S’, valores que se generan en la sentencia o se tienen previamente calculados como es el caso de las fechas y la etiqueta, para ser utilizados en la sentencia y completar los campos requeridos en el proceso de carga. Es importante resaltar que este proceso no se hace Modelado de datos y ETL 5 cliente por cliente, si no de forma masiva, para todos los clientes que tienen el valor de etiqueta 1 o 2. Una vez se realiza la inserción se procede a hacer la sentencia de actualización para ajustar los campos fechafin y actual, también de forma masiva: UPDATE DimCliente C INNER JOIN TempCliente AS TC ON C.idCliente = TC.idCliente SET C.FechaFin = FECHA\_CARGA, C.Actual = ‘N’ WHERE TC.Etiqueta=2; II. Conclusión Al final de esta lectura la invitación es a recordar que estos procesos de carga incremental, en muchas ocasiones afectan registros previamente cargados en la bodega de datos. Es así que se hace necesario, crear tablas temporales para generar de forma masiva nuevas columnas que van a permitir realizar procesos de inserción y actualización de forma masiva. Estas actualizaciones se concentran en las tablas de dimensión, donde si bien en diferentes contextos pueden tener varias filas, no tienen la misma cantidad que las tablas de hecho, en las cuales no se recomienda estos procesos de actualización Generalidades de los tableros de control

Save note

TranscriptNotesDownloadsDiscuss

### **Interactive Transcript - Enable basic transcript mode by pressing the escape key**

You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key

Play video starting at ::12 and follow transcript0:12

Hola, nuevamente. Hoy voy a hablarles sobre las generalidades de los tableros de

control y algunas recomendaciones para que el resultado final sea de mayor calidad.

Llegamos a la capa de aplicación o consumo,

y como vimos en los videos iniciales,

nos vamos a concentrar en el desarrollo de tableros de

control relacionados con inteligencia de negocios.

Recordemos que un tablero de control es un

mecanismo de visualización cuyo principal objetivo es

mostrar de una forma organizada y consentido para

un tomador de decisiones experto en una temática,

datos con el fin de identificar hallazgos

útiles que faciliten sus funciones dentro de una organización.

Estos hallazgos, a nivel de analítica

1.0 están centrados en el seguimiento de la estrategia de una

organización que se traduce en analizar

sus metas para determinar el nivel de cumplimiento que tiene.

A nivel de elementos de los tableros tenemos

indicadores de incremento de ventas, métricas,

como el número de empleados y de quejas en una

fecha o en su evolución a lo largo del tiempo;

se puede incluir mapas para ver más fácil el

comportamiento a nivel nacional y poder comparar visualmente.

Incluye semáforos para la visualización de

indicadores o mapas de calor en tablas o de mapas,

como su nombre lo indica,

para ilustrar cambios en el tiempo, entre otros ejemplos.

Y pensar en la interacción,

se tendrán filtros para seleccionar, por ejemplo, fechas,

producto, rangos de ventas y qué tipo de navegación sobre estos datos se podrá ser.

Tenemos, por ejemplo, ventas en un año y luego desglosarlo por trimestre.

Es así como, por ejemplo,

un gerente de servicios querrá ver la evolución en el tiempo del proceso de postventa,

con el fin de identificar empleados que tienen un bajo rendimiento,

dada la percepción de los clientes o tipos de quejas en ciertos procesos,

y con esta información podrá comparar empleados entre ellos

y diseñar estrategias de acompañamiento o de capacitación para entender

y fortalecer debilidades que pueden llevarlos a la desvinculación de clientes.

De igual manera, al identificar grupos

de quejas podrá orientar mejor

la capacitación del personal del call center y, en general,

podrá contar con información que le permitirá hacer ajustes a

sus procesos o definición estratégica de su área de la organización.

El manipular los datos para validar hipótesis es una de

las actividades clásicas relacionadas con tableros de control.

Esta validación puede darse en la interacción del experto con el tablero,

pero también, como una de las tendencias actuales,

puede ser lo contrario,

basarse en la generación de alertas al personal

responsable y su envío al correo celular con el fin de

darle información tan pronto se conoce para agilizar

su actuar y evitar complicaciones en sus áreas.

El éxito de estos proyectos de analítica consiste en

que los usuarios podrán generar nuevas hipótesis

y plantear nuevos análisis para responder a nuevas

preguntas que les surgen de su interacción con los datos.

Estos usuarios pueden ser diferentes a lo que originalmente se dirigió el tablero,

darles la oportunidad de manipular los datos de forma fácil y familiar

a su actividad podrá llevarlos

a solicitar nuevos análisis que darán origen, en muchas ocasiones,

a iniciar una nueva iteración del proceso completo de analítica desde la

integración de nuevas fuentes de datos hasta su

visualización e interacción en tableros de control.

Pasemos ahora a revisar algunas recomendaciones para que este "querer ser",

al usar los tableros de control,

pueda ser una realidad.

Construir un tablero de control,

en términos generales, no es una tarea compleja.

Sin embargo, hay que tener en cuenta varios aspectos antes de llegar a este punto.

Se debe entender el "para qué",

"por cuánto tiempo" y el "cómo",

¿será que se quieren para analizar un proceso concreto de la organización?

En el momento actual,

en el mediano plazo,

para uno o varios usuarios,

¿ya se tiene experiencia en este tipo de desarrollos?

¿Ya se tiene la tecnología requerida para su construcción y administración?

La respuesta a estas preguntas puede generar claridad en varias decisiones,

como el producto a utilizar,

la forma de darle acceso a las personas que van a utilizar los tableros,

los mecanismos para publicar,

mantenerlos actualizado, el orden en

su desarrollo o la identidad gráfica que se va a tener.

Es por ello que su desarrollo tiene mucho

sentido en el contexto de un proyecto centrado en datos,

como es el caso de los proyectos de analítica 1.0 que hemos trabajado.

Es al interior de ese proyecto donde una parte está asociada al

desarrollo de aplicaciones para publicar los resultados del proyecto,

donde se debe determinar el conjunto de

métricas o K-P-I's que se quieren poner a disposición.

Y una pregunta que pocas veces nos hacemos,

¿para qué nos va a servir?

Es importante desarrollar un proceso para la creación

de esos tableros y reflexionar sobre la respuesta a estas preguntas.

Pensemos en algunas de las respuestas y recomendaciones en este proceso de construcción.

La claridad en el público objetivo va

a permitir determinar con quién trabajan en el diseño del tablero,

quién va a validar qué tipos de gráficas a utilizar.

Debe ser clara la historia a contar,

que puede partir de las necesidades actuales del usuario,

quien va a interactuar con el tablero

e involucrar acciones o decisiones que él podrá hacer para apoyar su día a día.

El diseño gráfico es fundamental.

Recuerda que una imagen vale más que 1.000 palabras y el diseñar estos

tableros abre un mundo de posibilidades para los usuarios de los mismos.

Es en este punto donde contar con un mockup para ilustrar ideas es fundamental.

Toma de decisiones sobre los elementos gráficos a utilizar.

Piensa en representaciones efectivas para

los datos que tienes de acuerdo con el tipo de dato: numérico,

categórico; el análisis que deseas mostrar y el público objetivo.

Define el idioma a utilizar,

al igual que la información para acompañar cada gráfica.

Esto te permitirá estandarizar y facilitar el uso y desarrollo de estos tableros.

Determina o refina los filtros que vas a utilizar para la interacción.

Piensa en mecanismos para comparar los datos y dispón los datos en escalas comparables.

Recuerda que el contar con una definición de identidad gráfica que incluye logos,

colores, tipos de letras y tamaños,

ayuda a tener una unidad y facilitar la interacción con los usuarios.

Y lo más importante,

diseña y valida tu propuesta con el experto

de negocio para quien desarrollas el tablero de control.

Para tomar estas decisiones parte del usuario final,

conócelo bien y apóyate en un diseñador gráfico en esta parte.

Prefiere lo simple y visualizaciones sin saturación.

Ten un estándar para la información de las gráficas,

títulos cortos en el idioma seleccionado,

una forma de encontrar información sobre lo que se muestra.

No olvides nombrar los ejes e incluir unidades de los valores que visualices.

Estos pequeños detalles hacen la diferencia en el momento

de usar tableros de control y son los errores típicos de estas aplicaciones.

Llegamos al final del video en el cual revisamos generalidades de los

tableros de control y elementos a considerar a un alto nivel,

como es el caso de la selección de la terminología o el momento del diseño,

donde lo que se quiere es evitar tableros de control saturados que

no sean fáciles de entender y donde la interacción del experto sea difícil.

Si ya has participado en desarrollo de tableros de control,

piensa si tuviste en cuenta estos elementos o

si hay forma de mejorarlos y en futuras construcciones,

tener tableros de mejor calidad.

En este punto una última recomendación,

es preferible dedicar tiempo del proyecto para acompañar al experto en

el uso de estos tableros y así aumentar la posibilidad de que lo use,

y evitar que frente al primer tropiezo,

vuelva a sus esquemas iniciales de análisis y visualización,

que toman más tiempo en su construcción y pueden llevarlo a inconsistencias.

Hasta pronto.

Generalidades de los tableros de control Save note Transcript Notes Downloads Discuss Interactive Transcript - Enable basic transcript mode by pressing the escape key You may navigate through the transcript using tab. To save a note for a section of text press ⌘ + S. To expand your selection you may use ⌘ + arrow key. You may contract your selection using shift + ⌘ + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using ⌘ in conjunction with the alt key Play video starting at ::12 and follow transcript0:12 Hola, nuevamente. Hoy voy a hablarles sobre las generalidades de los tableros de control y algunas recomendaciones para que el resultado final sea de mayor calidad. Llegamos a la capa de aplicación o consumo, y como vimos en los videos iniciales, nos vamos a concentrar en el desarrollo de tableros de control relacionados con inteligencia de negocios. Recordemos que un tablero de control es un mecanismo de visualización cuyo principal objetivo es mostrar de una forma organizada y consentido para un tomador de decisiones experto en una temática, datos con el fin de identificar hallazgos útiles que faciliten sus funciones dentro de una organización. Estos hallazgos, a nivel de analítica 1.0 están centrados en el seguimiento de la estrategia de una organización que se traduce en analizar sus metas para determinar el nivel de cumplimiento que tiene. A nivel de elementos de los tableros tenemos indicadores de incremento de ventas, métricas, como el número de empleados y de quejas en una fecha o en su evolución a lo largo del tiempo; se puede incluir mapas para ver más fácil el comportamiento a nivel nacional y poder comparar visualmente. Incluye semáforos para la visualización de indicadores o mapas de calor en tablas o de mapas, como su nombre lo indica, para ilustrar cambios en el tiempo, entre otros ejemplos. Y pensar en la interacción, se tendrán filtros para seleccionar, por ejemplo, fechas, producto, rangos de ventas y qué tipo de navegación sobre estos datos se podrá ser. Tenemos, por ejemplo, ventas en un año y luego desglosarlo por trimestre. Es así como, por ejemplo, un gerente de servicios querrá ver la evolución en el tiempo del proceso de postventa, con el fin de identificar empleados que tienen un bajo rendimiento, dada la percepción de los clientes o tipos de quejas en ciertos procesos, y con esta información podrá comparar empleados entre ellos y diseñar estrategias de acompañamiento o de capacitación para entender y fortalecer debilidades que pueden llevarlos a la desvinculación de clientes. De igual manera, al identificar grupos de quejas podrá orientar mejor la capacitación del personal del call center y, en general, podrá contar con información que le permitirá hacer ajustes a sus procesos o definición estratégica de su área de la organización. El manipular los datos para validar hipótesis es una de las actividades clásicas relacionadas con tableros de control. Esta validación puede darse en la interacción del experto con el tablero, pero también, como una de las tendencias actuales, puede ser lo contrario, basarse en la generación de alertas al personal responsable y su envío al correo celular con el fin de darle información tan pronto se conoce para agilizar su actuar y evitar complicaciones en sus áreas. El éxito de estos proyectos de analítica consiste en que los usuarios podrán generar nuevas hipótesis y plantear nuevos análisis para responder a nuevas preguntas que les surgen de su interacción con los datos. Estos usuarios pueden ser diferentes a lo que originalmente se dirigió el tablero, darles la oportunidad de manipular los datos de forma fácil y familiar a su actividad podrá llevarlos a solicitar nuevos análisis que darán origen, en muchas ocasiones, a iniciar una nueva iteración del proceso completo de analítica desde la integración de nuevas fuentes de datos hasta su visualización e interacción en tableros de control. Pasemos ahora a revisar algunas recomendaciones para que este "querer ser", al usar los tableros de control, pueda ser una realidad. Construir un tablero de control, en términos generales, no es una tarea compleja. Sin embargo, hay que tener en cuenta varios aspectos antes de llegar a este punto. Se debe entender el "para qué", "por cuánto tiempo" y el "cómo", ¿será que se quieren para analizar un proceso concreto de la organización? En el momento actual, en el mediano plazo, para uno o varios usuarios, ¿ya se tiene experiencia en este tipo de desarrollos? ¿Ya se tiene la tecnología requerida para su construcción y administración? La respuesta a estas preguntas puede generar claridad en varias decisiones, como el producto a utilizar, la forma de darle acceso a las personas que van a utilizar los tableros, los mecanismos para publicar, mantenerlos actualizado, el orden en su desarrollo o la identidad gráfica que se va a tener. Es por ello que su desarrollo tiene mucho sentido en el contexto de un proyecto centrado en datos, como es el caso de los proyectos de analítica 1.0 que hemos trabajado. Es al interior de ese proyecto donde una parte está asociada al desarrollo de aplicaciones para publicar los resultados del proyecto, donde se debe determinar el conjunto de métricas o K-P-I's que se quieren poner a disposición. Y una pregunta que pocas veces nos hacemos, ¿para qué nos va a servir? Es importante desarrollar un proceso para la creación de esos tableros y reflexionar sobre la respuesta a estas preguntas. Pensemos en algunas de las respuestas y recomendaciones en este proceso de construcción. La claridad en el público objetivo va a permitir determinar con quién trabajan en el diseño del tablero, quién va a validar qué tipos de gráficas a utilizar. Debe ser clara la historia a contar, que puede partir de las necesidades actuales del usuario, quien va a interactuar con el tablero e involucrar acciones o decisiones que él podrá hacer para apoyar su día a día. El diseño gráfico es fundamental. Recuerda que una imagen vale más que 1.000 palabras y el diseñar estos tableros abre un mundo de posibilidades para los usuarios de los mismos. Es en este punto donde contar con un mockup para ilustrar ideas es fundamental. Toma de decisiones sobre los elementos gráficos a utilizar. Piensa en representaciones efectivas para los datos que tienes de acuerdo con el tipo de dato: numérico, categórico; el análisis que deseas mostrar y el público objetivo. Define el idioma a utilizar, al igual que la información para acompañar cada gráfica. Esto te permitirá estandarizar y facilitar el uso y desarrollo de estos tableros. Determina o refina los filtros que vas a utilizar para la interacción. Piensa en mecanismos para comparar los datos y dispón los datos en escalas comparables. Recuerda que el contar con una definición de identidad gráfica que incluye logos, colores, tipos de letras y tamaños, ayuda a tener una unidad y facilitar la interacción con los usuarios. Y lo más importante, diseña y valida tu propuesta con el experto de negocio para quien desarrollas el tablero de control. Para tomar estas decisiones parte del usuario final, conócelo bien y apóyate en un diseñador gráfico en esta parte. Prefiere lo simple y visualizaciones sin saturación. Ten un estándar para la información de las gráficas, títulos cortos en el idioma seleccionado, una forma de encontrar información sobre lo que se muestra. No olvides nombrar los ejes e incluir unidades de los valores que visualices. Estos pequeños detalles hacen la diferencia en el momento de usar tableros de control y son los errores típicos de estas aplicaciones. Llegamos al final del video en el cual revisamos generalidades de los tableros de control y elementos a considerar a un alto nivel, como es el caso de la selección de la terminología o el momento del diseño, donde lo que se quiere es evitar tableros de control saturados que no sean fáciles de entender y donde la interacción del experto sea difícil. Si ya has participado en desarrollo de tableros de control, piensa si tuviste en cuenta estos elementos o si hay forma de mejorarlos y en futuras construcciones, tener tableros de mejor calidad. En este punto una última recomendación, es preferible dedicar tiempo del proyecto para acompañar al experto en el uso de estos tableros y así aumentar la posibilidad de que lo use, y evitar que frente al primer tropiezo, vuelva a sus esquemas iniciales de análisis y visualización, que toman más tiempo en su construcción y pueden llevarlo a inconsistencias. Hasta pronto. Transcript language: Español : Added to Selection. Press [⌘ + S] to save as a note

Transcript language: Español

to de una organización Modelado de datos y ETL 05-2022 María del Pilar Villamil Giraldo Modelado de datos y ETL 2 I. Comentarios El siguiente texto fue elaborado por la profesora María del Pilar Villamil como material de estudio para el curso. Las referencias utilizadas se encuentran al final del mismo. II. Detalle de la lectura En ambientes de analítica descriptiva 1.0 que se caracterizan por la generación de reportes y esquemas de visualización sobre datos de una organización, y que es conocida como inteligencia de negocios, el uso de reportes o cubos OLAP (cubos en la figura 1) y tableros de control (dash board) es fundamental para presentar información a tomadores de decisión de una organización. A continuación, damos una descripción de estas aplicaciones. Reportes OLAP Los reportes o análisis OLAP (online analytical process), como su nombre lo indica permiten realizar análisis “al vuelo” de acuerdo con las necesidades del analista. Estos análisis se basan en la existencia de un cubo OLAP, estructura de datos que contiene como perspectivas de análisis lo que denominamos dimensiones, y como elemento de análisis las medidas. Las perspectivas en la figura 1 son la fecha, la ruta o vía, y la concesión, mientras que las medidas son el recaudo (dinero recaudado), y el tráfico de vehículos. La existencia del cubo OLAP permite al analista seleccionar diferentes dimensiones o medidas y construir su reporte a la medida. Es así, como en la figura 1 y 2, vemos que el analista seleccionó las dimensiones de ruta y año-mes y respectivamente las medidas de recaudo y tráfico, para generar dos reportes distintos. Figura 1. Cubo OLAP de recaudo y tráfico de peajes por año y mes específico Ruta/vi a Año-mes Recaudo: 6.196 Tráfico: 7.164 Modelado de datos y ETL 3 Figura 2. Análisis OLAP de recaudo de peajes por año y mes específico Figura 3. Análisis OLAP de tráfico de vehículos en peajes por año y mes específico Este tipo de visualización es ampliamente utilizado y puede ser enriquecido con colores, gráficos de barra, línea, etc. Con el fin de transmitir fácilmente el mensaje que se quiere. En este caso, dada la representación en tabla, es fácil entender el contenido y quizás es la forma de visualizar más sencilla, con la cual la mayoría de analistas están familiarizados. Tableros de control Los tableros de control son una herramienta de visualización que permite agrupar, ordenar y visualizar en una sola pantalla, un conjunto de métricas y de indicadores que son de interés para una organización. En el diseño de esta visualización se debe definir claramente el rol o actor que va a interactuar con la información, para guiar las métricas a visualizar, al igual que los elementos gráficos a utilizar, los colores, los mecanismos de interacción, los cuales deben ser cercanos a la persona a quién se dirige el tablero de control y respetar la imagen corporativa. Es así como en la figura 4 observamos un tablero de control que ilustra información relacionada con ventas. Vemos como el uso de los colores, de íconos para representar los accesorios que venden, al Modelado de datos y ETL 4 igual que el mapa, el diagrama de barras, las cajas de diálogo y las tablas son los elementos utilizados para la visualización y la interacción con los datos. Sin conocer mucho de la organización dueña de los datos es posible contar historias, conocer la organización e imaginar, roles que pueden interactuar con esa información, al igual que decisiones que podrían llegar a tomar. Figura 4. Tablero de control de ventas tomado de [https://www.logianalytics.com/dashboarddesignguide/wp-content/uploads/2015/09/Chap-1- dbg.png, Consultado el 28 de mayo 2021] ¿Qué podemos inferir de la empresa dueña de los datos que se muestran en la figura 5? Al parecer, se trata de una empresa que vende accesorios para hombre y mujer. La mayoría del dinero de transacciones de ventas (68%, £16,740,510), reportado en libras, corresponde a ventas a mujeres, mientras que el 31% a hombres y es posible que exista un 1% del dinero de las ventas que no se logre clasificar. Dentro de los accesorios que venden se encuentran, vestidos -el más vendido- y zapatos para dama, al igual que sacos y gafas para hombre -el menos vendido-. De igual manera, tiene Modelado de datos y ETL 5 sucursales en Norte América, América del Sur, Europa, Asia, África y Australia con una representación en dinero, liderada por el mercado europeo y con menores ingresos en África. Esta empresa, ha puesto a disposición de sus clientes diferentes Canales de adquisición, entre los que se encuentran los presenciales representados por su call center y, los en línea como el correo electrónico. Esta es una historia que podemos contar con la información de ese tablero de control. Pasemos ahora a imaginar un rol de la empresa que pueda interesarse por estos datos. Quizás se trata de un gerente de ventas que quiere mirar el comportamiento de sus ventas, no solo en cantidad, sino principalmente en dinero que ingresa a la empresa. Ese conocimiento, acompañado del género que más compra, quiere utilizarlo para trabajar con el área de mercadeo y fortalecer el mercadeo directo, guiándolo a un género y un tipo de accesorio particular. De igual manera, está diseñando estrategias para fortalecer las ventas en África, donde ve un gran potencial que no ha podido aprovecharse, para consolidarse en ese continente como líder en los accesorios que vende. Pensar y crear este tipo de historias es lo que debe guiar el diseño de un tablero de control. En particular, piense en: 1. El público objetivo: vimos en este punto que es importante identificar una persona específica o un cargo o rol dentro de una organización: ¿A quién está dirigido el tablero de control? 2. La historia a contar: ¿Qué tipo de historia quiere contar? ¿Para qué contarla? En este punto, piense en decisiones o acciones que podrán tomar el público objetivo. 3. Resuma en una frase el mensaje que quiere transmitir. 4. Cree su guion gráfico y defina el orden en que quiere presentar sus ideas, al igual que la interacción de los elementos. En este punto puede apoyarse en herramientas para realizar mock-ups1 que le permitirán desarrollar sus propios diseños. 5. Seleccione elementos gráficos efectivos para los datos que quiere representar. En este punto recuerde la gama de opciones que puede tener y aprenda a identificar los que son más apropiados, de acuerdo con el tipo de dato (numérico, categórico), el análisis que desea mostrar y el público objetivo. 6. Determine o refine los filtros que va a utilizar para la interacción, los cuales no deben ocupar mucho espacio y deben presentarse de la forma más sencilla y natural. Estos pueden estar relacionados con el tiempo, los accesorios, género o geografía si continuamos pensando en el tablero de control de la figura 4. 1 https://ninjamock.com/, https://moqups.com/ , ,https://pencil.evolus.vn/ , https://mockflow.com , https://wirefram e.cc/#about, https://balsamiq.com/, Modelado de datos y ETL 6 Finalmente, diseñe valide su propuesta con el experto de negocio para quien desarrolla el tablero de control. Si quieres profundizar en tableros de control puedes consultar estas referencias y dependiendo de la temática a visualizar revisa primero ideas para hacer la representación, antes de definir una forma de hacerlo. 1. Capítulo 2 de la referencia [1] 2. https://www.semetrical.com/google-data-studio-dashboard-tips/ 3. https://www2.insightsoftware.com/dashboard-design-guide/the-future-of-dashboarddesign/ 4. https://theymakedesign.com/best-web-design-inspiration-dashboards-d87015ffb711 Modelado de datos y ETL 7 Ahora te invito a seguir un proceso similar al que yo utilicé, para describir el siguiente tablero de control que está en un nivel avanzado. Puedes también, buscar tableros de control que te parezcan diferentes y valiosos para cuando debas crear los tuyos. Tablero de control tomado de: [Leidy Alexandra Lozano Jácome. Estrategia para desarrollar observatorios de Salud integrando inteligencia de negocios y Analítica visual. 2013] III. Conclusión Los tableros de control son un tipo de aplicación que es utilizada en diferentes disciplinas y con los años se ha consolidado como mecanismo para mostrar hallazgos y facilitar procesos de toma de decisiones. Aunque cada vez están más presente en el día a día de las personas, su diseño, Modelado de datos y ETL 8 integración con los datos y publicación aún son temas en los que es importante detenerse a reflexionar para que la calidad de los tableros sea mejor cada día  
  
Recomendaciones en la construcción de tableros de control en el contexto de aplicaciones de analítica 1.0 Modelado de datos y ETL 05-2022 María del Pilar Villamil Giraldo Modelado de datos y ETL 2 I. Comentarios El siguiente texto fue elaborado por la profesora María del Pilar Villamil como material de estudio para el curso. Las referencias utilizadas se encuentran al final del mismo. II. Detalle de la lectura Recordemos que nos encontramos en el contexto del desarrollo de un proyecto de analítica 1.0 en el cual, si seguimos la adaptación de la metodología ASUM-DM, tenemos una etapa de definición del producto de software, en la cual partimos de la identificación de requerimientos analíticos realizados en la etapa del entendimiento del negocio, al igual que el conocimiento de los datos y de la infraestructura de la organización, obtenidos en las etapas de entendimiento del negocio y de los datos, para diseñar el tablero de control, apoyados en el desarrollo de un boceto (mock-up). Es así como para el requerimiento analítico 1. de la tabla 1, vemos que se detalla el análisis requerido, el cual debe ser visualizado en un tablero de control dada la naturaleza del proyecto de analítica que se está trabajando. Tema analítico Análisis requeridos o inferidos/requerimientos analíticos Procesos de negocio Datos requeridos Inteligencia de Mercadeo: Conocimiento del cliente Análisis de frecuencia de compra de los clientes o tipos de cliente en un periodo de tiempo, días transcurridos desde la última fecha de compra, monto percibido y un indicador que muestre si está por encima, en el promedio o por debajo del promedio en esas métricas. Este análisis se puede resolver con un Tablero de control. Manejo de datos maestros: órdenes (compras de clientes); atención al cliente (consultas acerca de productos o servicios, quejas, solicitud de asistencia, etc.) \* Circunstancias del cliente (edad, sexo, estado civil, ocupación, hobbies, número de hijos, etc.) \* Comportamiento del cliente (compras, pagos, contactos, etc.) \* Información derivada (Categoría del cliente, categoría de sus quejas) Segmentación de clientes por circunstancias, comportamiento, información derivada (lifetime value, Modelado de datos y ETL 3 propensión a desertar, categoría de quejas). Este análisis se puede resolver utilizando aprendizaje automático. Tabla 1. Ejemplo de temas analíticos y requerimientos analíticos En este punto se debe resaltar que en dicho análisis se requiere ver como medidas, la frecuencia de compra de los clientes, los días transcurridos desde la última fecha de compra, el monto percibido, y esos valores dependen de los valores de las dimensiones tiempo y tipo de cliente sobre las cuales nos imaginamos se deben proporcionar filtros para su selección. Imagine que se selecciona un año específico y se muestra la información de las medidas que mencionamos previamente. La pregunta en este punto es: ¿es suficiente esa información o se debe dar mayor información que permita entender si ese dato está dentro del comportamiento esperado, o todo lo contrario? Es así, como en el requerimiento se cuenta adicionalmente, con un indicador que muestre si está por encima, en el promedio o por debajo del promedio en esas métricas. Este indicador rápidamente ubica al tomador de decisiones en la interpretación que puede darles a las métricas que se muestran y llevarlo a identificar el problema y guiar la solución a proponer. En este requerimiento se identifican otras formas de mostrar esa información para enriquecerla, ofrecer un contexto al tomador de decisión y así, facilitar su labor. Estas formas se resumen en los siguientes puntos: 1. Comparar el valor de una métrica con el de un KPI ya definido, posiblemente en otro contexto, por ejemplo, por parte de los entes reguladores del sector. 2. Colorear el ranking de los valores que se muestran y utilizar una propuesta como, por ejemplo, los menores valores correspondientes al primer cuartil serán de color rojo, mientras que los que están por encima del tercer cuartil tendrán un color verde y los que están en el intermedio, tendrán un color amarillo. 3. Comparar los indicadores con los de otros tipos de cliente en el mismo periodo de tiempo, para ver la diferencia. 4. Analizar el comportamiento a lo largo del tiempo para ver tendencias. Recuerde que estos análisis se proponen para un usuario o grupo de usuarios y esto debe guiar los elementos gráficos a presentar y la historia a comunicar como parte final de la consultoría, gracias a que se conoce el quién va a utilizarlos y el para qué. Las opciones de elementos gráficos las puede consultar de acuerdo con el producto de software seleccionado para construir los tableros de control. En este punto uno de los retos es ser consciente que pueden existir varios requerimientos analíticos similares, ya sea en el mismo tema analítico o en otro y si no se cuenta con una estrategia de Modelado de datos y ETL 4 desarrollo de tableros de control, al final se tendrán diferentes tableros con información similar. Estos tableros, desarrollados de forma independiente, podrán apoyar los mismos procesos de toma de decisión o similares, lo que lleva a generar múltiples versiones del mismo tablero de control, en las cuales el valor de los indicadores no necesariamente es el mismo. Esta situación puede ser originada por el cálculo directo de los indicadores en los tableros o por el uso de fuentes de datos que son transformadas de forma independiente para generar el tablero. Esta situación lleva a una dificultad al momento de escalar el desarrollo de tableros de control y es por ello que se debe prever esta situación y proponer mecanismos para mitigar este riesgo. El diseño de tableros de control usando mock-ups es una estrategia que facilita la comunicación entre los diseñadores y el tomador de decisiones. Se espera que, frente a estas propuestas, el tomador de decisiones pueda aportar y comentar sobre la forma de visualizar (gráficas, colores, logos institucionales, etc.), la completitud de la información que se muestra, el nivel de saturación que tiene, la facilidad para su interacción y uso cotidiano y de esta forma incorporar las sugerencias transmitidas y tener un producto de mejor calidad. Adicionalmente, es importante brindar mecanismos para descargar o compartir la información que se visualiza en el tablero y así generar reportes para diferentes usuarios. Este punto también debe ser validado con el usuario final del tablero de control. El diseño del tablero debe ser simple, se debe pensar en que menos información facilita la comprensión de lo que se tiene, además de simplificar gráficas, títulos, y esquemas de navegación. Prefiera tener botones para dar más información, o para llevar a otras páginas y detallar la visualización y no recargar una en particular. Tenga a mano un manual de identidad gráfica con esta información o constrúyalo para que tenga unidad en lo que se desarrolla y así, facilitar el desarrollo de los mismos y la comprensión por parte de tomadores de decisiones al momento de utilizar los tableros. Incluya botones de información para explicar el objetivo del tablero, la fuente de datos utilizada y en general aclarar la interacción con el tablero de control. III. Conclusión El desarrollo de tableros de control está presente en el día a día de las organizaciones. Sin embargo, es importante recordar lo que se debe definir previamente para guiar el desarrollo y evitar replicación de esfuerzos en esta actividad y confusión por parte de los usuarios finales

**FALTA DESPUES DE LA SEMANA &, ES DECIR DESDE. LA SEMNA 7**