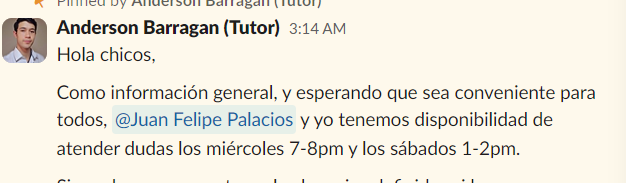
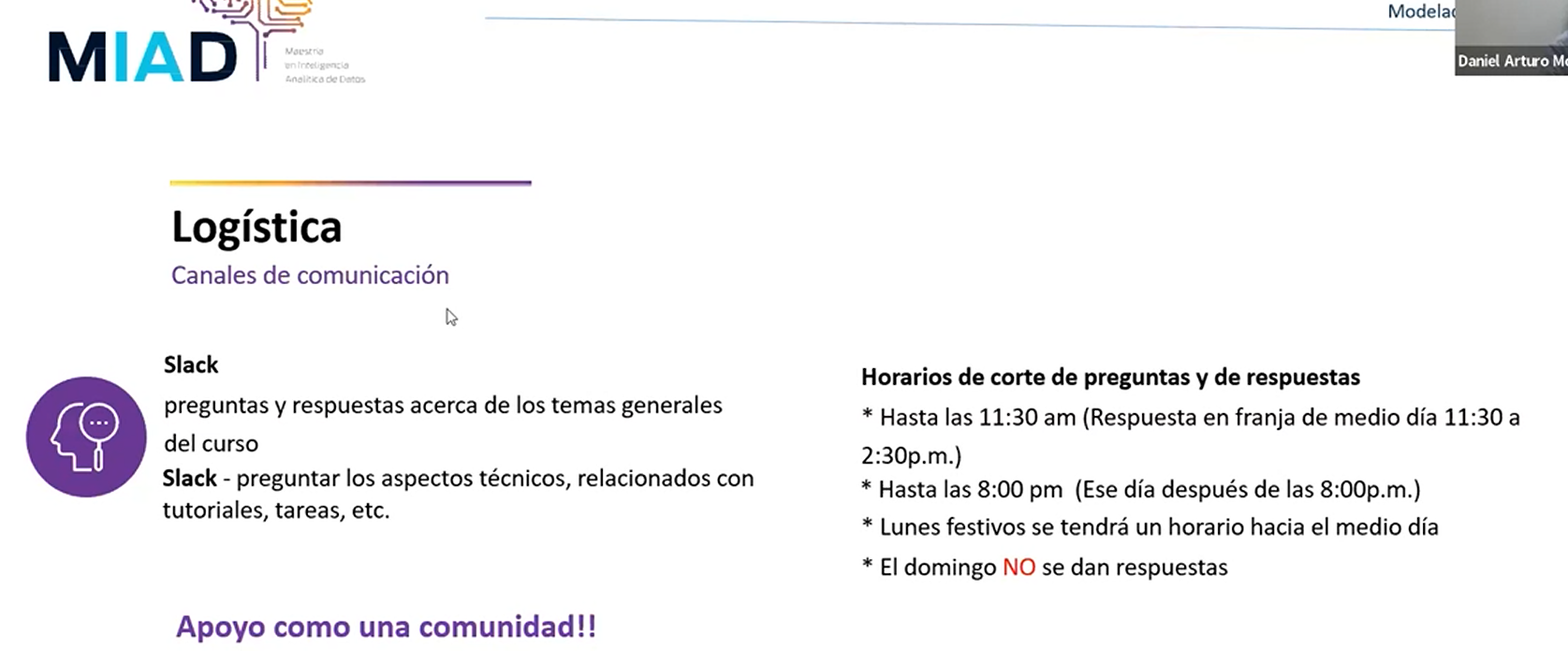
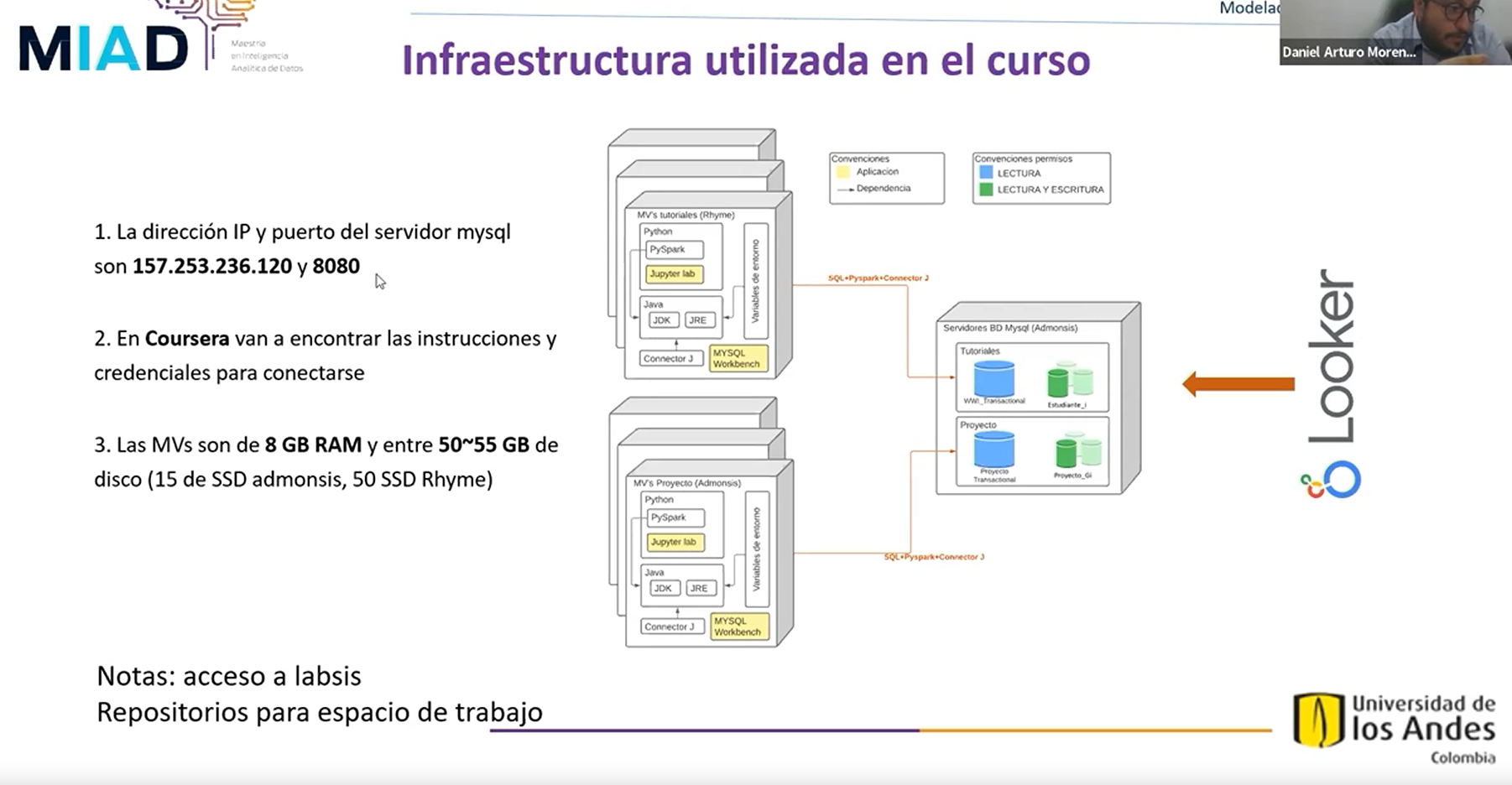
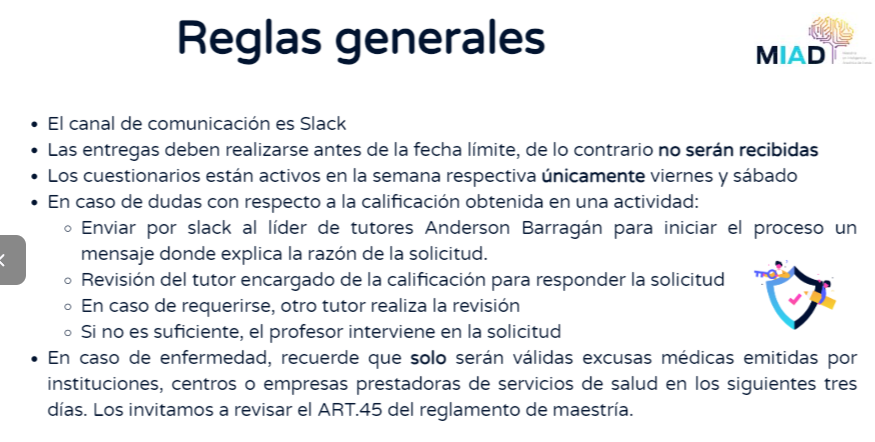
**Modelado de Datos y ETL**

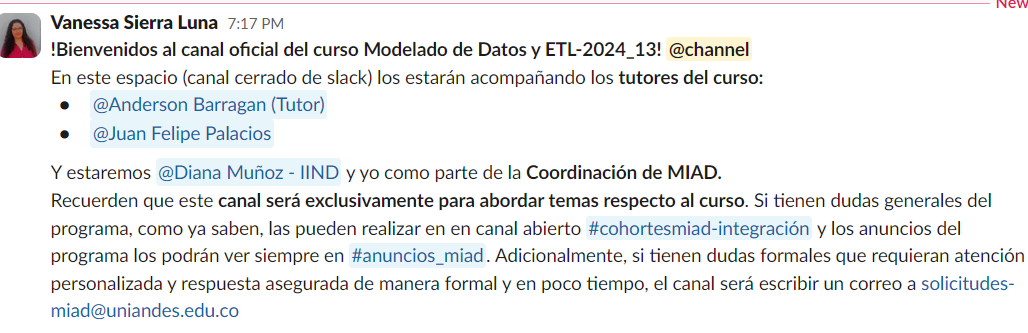












**Panorama general de la analítica**

[Panorama general de la analítica | Coursera](https://www.coursera.org/learn/modelado-de-datos-y-etl/lecture/uoqQn/panorama-general-de-la-analitica)

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**

[Arquitectura de componentes para aplicaciones de analítica descriptiva | Coursera](https://www.coursera.org/learn/modelado-de-datos-y-etl/lecture/xEfKA/arquitectura-de-componentes-para-aplicaciones-de-analitica-descriptiva)

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.

Reproduce el video desde :1:1 y sigue la transcripción1:01

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.

Reproduce el video desde :2:42 y sigue la transcripción2:42

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.

Reproduce el video desde :4:8 y sigue la transcripción4:08

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.

Reproduce el video desde :4:55 y sigue la transcripción4:55

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.

Reproduce el video desde :7:36 y sigue la transcripción7:36

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.

Reproduce el video desde :9:19 y sigue la transcripción9:19

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.

Reproduce el video desde :9:48 y sigue la transcripción9:48

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.

Reproduce el video desde :10:1 y sigue la transcripción10:01

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.

Reproduce el video desde :10:21 y sigue la transcripción10:21

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.

Reproduce el video desde :10:37 y sigue la transcripción10:37

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.

Reproduce el video desde :12:6 y sigue la transcripción12:06

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.

Reproduce el video desde :12:22 y sigue la transcripción12:22

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.

[

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

[Repaso de conceptos de bases de datos relacionales y SQL | Coursera](https://www.coursera.org/learn/modelado-de-datos-y-etl/lecture/0noKT/repaso-de-conceptos-de-bases-de-datos-relacionales-y-sql)

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

Reproduce el video desde ::15 y sigue la transcripción0:15

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.

Reproduce el video desde ::57 y sigue la transcripción0:57

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.

Reproduce el video desde :1:15 y sigue la transcripción1:15

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.

Reproduce el video desde :1:30 y sigue la transcripción1:30

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.

Reproduce el video desde :1:59 y sigue la transcripción1:59

Empecemos con el de creación.

Reproduce el video desde :2:2 y sigue la transcripción2:02

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.

Reproduce el video desde :2:21 y sigue la transcripción2:21

Adicionalmente, como también lo vemos en pantalla,

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

Reproduce el video desde :2:30 y sigue la transcripción2:30

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.

Reproduce el video desde :2:48 y sigue la transcripción2:48

Tenemos también el tipo fecha, que es DATE.

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

Reproduce el video desde :2:56 y sigue la transcripción2:56

Si queremos manejar valores booleanos, tenemos BOOL.

O si queremos manejar números de punto flotante,

tenemos el FLLOT, entre los más utilizados.

Reproduce el video desde :3:7 y sigue la transcripción3:07

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.

Reproduce el video desde :3:53 y sigue la transcripción3:53

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.

Reproduce el video desde :4:5 y sigue la transcripción4:05

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.

Reproduce el video desde :4:25 y sigue la transcripción4:25

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.

Reproduce el video desde :4:48 y sigue la transcripción4:48

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

analítica.

Reproduce el video desde :4:53 y sigue la transcripción4:53

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.

Reproduce el video desde :5:3 y sigue la transcripción5:03

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.

Reproduce el video desde :5:17 y sigue la transcripción5:17

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.

Reproduce el video desde :5:31 y sigue la transcripción5:31

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.

Reproduce el video desde :5:56 y sigue la transcripción5:56

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.

Reproduce el video desde :6:16 y sigue la transcripción6:16

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.

Reproduce el video desde :6:32 y sigue la transcripción6:32

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.

Reproduce el video desde :7:30 y sigue la transcripción7:30

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.

Reproduce el video desde :7:54 y sigue la transcripción7:54

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

nombre de este cliente.

Veamos el resultado.

Reproduce el video desde :8:5 y sigue la transcripción8:05

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.

Reproduce el video desde :8:16 y sigue la transcripción8:16

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.

Reproduce el video desde :8:38 y sigue la transcripción8:38

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

cuales queremos traer la información.

Reproduce el video desde :8:48 y sigue la transcripción8:48

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

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

Reproduce el video desde :8:58 y sigue la transcripción8:58

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.

Reproduce el video desde :9:24 y sigue la transcripción9:24

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.

Reproduce el video desde :9:39 y sigue la transcripción9:39

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.

Reproduce el video desde :9:57 y sigue la transcripción9:57

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.

Reproduce el video desde :10:17 y sigue la transcripción10:17

%n para representar dos dígitos del mes,

y %d para representar dos dígitos del día,

relacionado con la fecha.

Reproduce el video desde :10:32 y sigue la transcripción10:32

Otro concepto fundamental es el de JOIN.

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

Reproduce el video desde :10:40 y sigue la transcripción10:40

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.

Reproduce el video desde :11:26 y sigue la transcripción11:26

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.

Reproduce el video desde :11:41 y sigue la transcripción11:41

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.

Reproduce el video desde :11:58 y sigue la transcripción11:58

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.

Reproduce el video desde :12:28 y sigue la transcripción12:28

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

es decir, con un valor de null.

Reproduce el video desde :12:36 y sigue la transcripción12:36

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.

Reproduce el video desde :13:4 y sigue la transcripción13:04

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

para ti.

Reproduce el video desde :13:9 y sigue la transcripción13:09

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.

Reproduce el video desde :13:39 y sigue la transcripción13:39

[MUSIC]

**Semana 2**

**Uso de la infografía entendiendo los datos**

[Uso de la infografía entendiendo los datos | Coursera](https://www.coursera.org/learn/modelado-de-datos-y-etl/lecture/7GPzW/uso-de-la-infografia-entendiendo-los-datos)

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.

Reproduce el video desde ::26 y sigue la transcripción0:26

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.

Reproduce el video desde ::43 y sigue la transcripción0:43

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.

Reproduce el video desde :1:4 y sigue la transcripción1:04

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.

Reproduce el video desde :3:27 y sigue la transcripción3:27

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.

Reproduce el video desde :5:15 y sigue la transcripción5:15

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]

SEMANA 3

**Modelado multidimensional de datos**

Guardar nota

[Modelado multidimensional de datos | Coursera](https://www.coursera.org/learn/modelado-de-datos-y-etl/lecture/Xv8ja/modelado-multidimensional-de-datos)

Bienvenidos a este video introductorio sobre modelado multidimensional,

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

Reproduce el video desde ::18 y sigue la transcripción0:18

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.

Reproduce el video desde ::43 y sigue la transcripción0:43

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.

Reproduce el video desde :1:31 y sigue la transcripción1:31

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.

Reproduce el video desde :2:4 y sigue la transcripción2:04

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.

Reproduce el video desde :2:48 y sigue la transcripción2:48

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?

Reproduce el video desde :3:33 y sigue la transcripción3:33

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.

Reproduce el video desde :4:33 y sigue la transcripción4:33

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.

Reproduce el video desde :5:1 y sigue la transcripción5:01

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.

Reproduce el video desde :6:2 y sigue la transcripción6:02

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]