

1. Sistemas de Soporte de Decisión y almacenes de datos

1.1 Sistemas de Soporte de decisiones (DSS)

- 1.1.1 Características de los DSS
- 1.1.2 Taxonomía de los DSS
- 1.1.3 Arquitecturas y plataformas para los DSS

1.2 Almacenes de Datos

- 1.2.1 Características de un Almacén de Datos
- 1.2.2 Arquitecturas de Data Warehouses (almacenes de datos) y Data Mart (base de datos departamental)
- 1.2.3 Aplicaciones de Data Warehouse
- 1.2.4 Data Lake (lagos de datos)
- 1.2.5 Características de un Data Lake
- 1.2.6 Arquitecturas de Data Lake
- 1.2.7 Comparativa entre Data Lakes y Data Marts

1.1 Sistemas de Soporte de decisiones (DSS)

Inteligencia de Negocio

- La inteligencia de negocio (Business Intelligence, BI) puede definirse como un conjunto de modelos matemáticos y metodologías de análisis que explotan los datos disponibles para generar información y conocimiento útiles para procesos complejos de toma de decisiones.
- El propósito principal de los sistemas de BI es proporcionar a los trabajadores del conocimiento, las herramientas y metodologías que les permitan tomar decisiones efectivas y oportunas.
 - Si los responsables de la toma de decisiones pueden confiar en un sistema de inteligencia empresarial que facilite su actividad, podemos esperar que la calidad global del proceso de toma de decisiones mejore considerablemente.

Decisiones

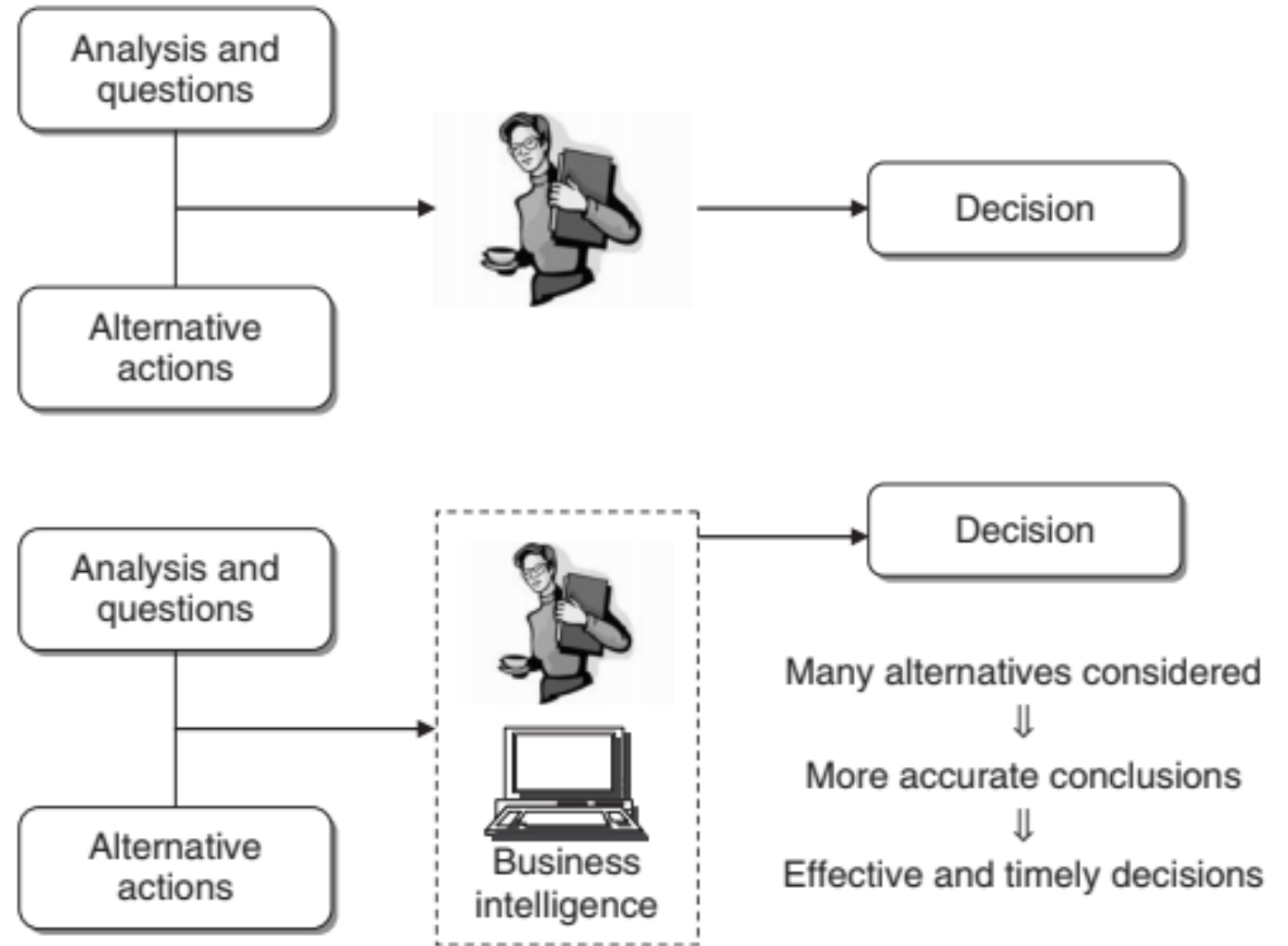
Decisiones efectivas. La aplicación de métodos analíticos rigurosos permite a los responsables de la toma de decisiones confiar en información y conocimientos que son más fiables.

- Como resultado, son capaces de tomar mejores decisiones y diseñar planes de acción que permiten alcanzar sus objetivos de una manera más eficaz.

Decisiones oportunas. Las empresas operan en entornos económicos caracterizados por niveles crecientes de competencia y un alto dinamismo.

- Como consecuencia, la capacidad de reaccionar rápidamente a las acciones de los competidores y a las nuevas condiciones del mercado es un factor crítico para el éxito o incluso la supervivencia de una empresa.

Beneficios de un sistema de BI



Datos, información y conocimiento

- La diferencia entre datos, información y conocimiento puede entenderse mejor a través de las siguientes observaciones.
 - **Datos.** En general, los datos representan una codificación estructurada de entidades primarias únicas, así como de transacciones que involucran a dos o más entidades primarias.
 - **Información.** La información es el resultado de las actividades de extracción y procesamiento realizadas en datos, y parece significativa para aquellos que la reciben en un dominio específico.
 - **Conocimiento.** La información se transforma en conocimiento cuando se utiliza para tomar decisiones y desarrollar las acciones correspondientes. El conocimiento consiste en información puesta a trabajar en un ámbito específico, reforzada por la experiencia y la competencia de los responsables de la toma de decisiones para abordar y resolver problemas complejos.

Evolución de los sistemas de información

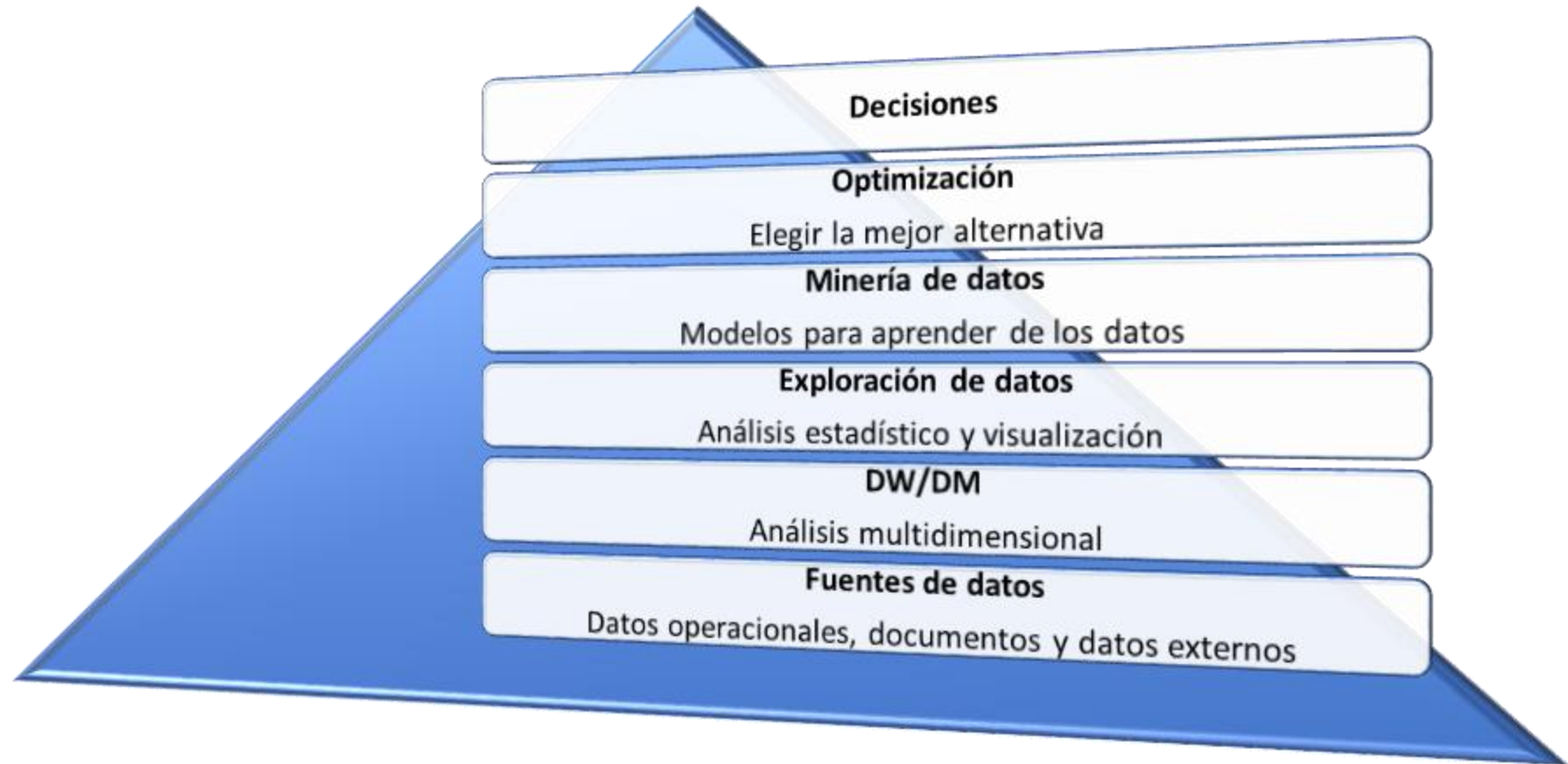
- Las computadoras digitales hicieron su aparición a finales de la década de 1940, y pronto comenzaron a aplicarse en el entorno empresarial.
 - Se caracterizaron por la difusión generalizada de aplicaciones que lograron un aumento de la eficiencia mediante la automatización de las operaciones rutinarias dentro de las empresas, especialmente en la administración, producción, investigación y desarrollo.
- En la década de 1970 comenzaron a surgir en las empresas la necesidad de diseñar aplicaciones informáticas, denominadas sistemas de gestión de información (MIS), con el fin de facilitar el acceso a información útil y oportuna para los responsables de la toma de decisiones.
 - Las computadoras centrales carecían de capacidades de visualización gráfica y se comunicaban con los usuarios a través de terminales de computadora basados en caracteres e impresoras de puntos.

-
- Desde finales de la década de 1980, la introducción de computadoras personales con sistemas operativos con interfaces gráficas y dispositivos señaladores, como un ratón, tuvo consecuencias.
 - Por un lado, se hizo posible implementar aplicaciones capaces de interacciones sofisticadas y de representación gráfica de los resultados.
 - Por otra parte, los trabajadores del conocimiento podían confiar en herramientas de procesamiento autónomos, evitando así el retraso en el acceso a los datos.
 - El aumento de las capacidades de procesamiento fue un factor crítico para futuros desarrollos.

-
- Se introdujo el concepto inicial de sistema de apoyo a la (toma de) decisión.
 - Los desarrollos posteriores generaron nuevos tipos de aplicaciones y arquitecturas: los sistemas de información ejecutiva (EIS) y los sistemas de información estratégica (SIS), que se introdujeron por primera vez a finales de la década de 1980 para apoyar a los ejecutivos en el proceso de toma de decisiones.
 - Estos sistemas estaban destinados a procesos de toma de decisiones no estructurados y, por lo tanto, representaban sistemas de apoyo pasivos orientados hacia un acceso oportuno y fácil a la información.

-
- Desde principios de la década de 1990, las arquitecturas de red y los sistemas de información distribuidos basados en modelos informáticos cliente-servidor comenzaron a ser ampliamente adoptados.
 - Surgió la necesidad de separar lógica y físicamente las bases de datos de los sistemas de información operativos. Esto dio lugar a los conceptos de almacenes de datos (DW) y data marts (DM).
 - Finalmente, hacia finales de la década de 1990, el término Business Intelligence (BI) comenzó a utilizarse para abordar en general la arquitectura que contiene a los DSSs, métodos analíticos y modelos utilizados para transformar los datos en información útil y conocimiento para los responsables de la toma de decisiones.

Componentes de un sistema de BI



-
- **Fuentes de Datos.** Los datos proceden en mayor parte de transacciones internas de carácter administrativo, logístico y comercial, y en otra parte de fuentes externas.
 - Sin embargo, aunque se hayan recopilado y almacenado de forma sistemática y estructurada, estos datos no pueden utilizarse directamente con fines de toma de decisiones.
 - Deben procesarse mediante herramientas de extracción adecuadas y métodos analíticos capaces de transformarlos en información y conocimientos que puedan ser posteriormente utilizados por los responsables de la toma de decisiones.
 - **DW/DM.** Permite, mediante el concepto de modelo multidimensional, estructurar los datos en un repositorio empresarial, para aplicar consultas OLAP
 - El DM mantiene los mismos conceptos, pero a una escala reducida de capacidad y funcionamiento

-
- **Exploración de datos.** Consta de las herramientas para realizar un análisis pasivo de BI, que consisten en sistemas de consulta e información, así como métodos estadísticos.
 - Se pide a los responsables de la toma de decisiones que generen preguntas previas o definan criterios de extracción de datos, y luego utilicen las herramientas de análisis para encontrar respuestas y confirmar su conocimiento original.
 - **Minería de datos.** Incluye metodologías activas de BI, cuya finalidad es la extracción de información y conocimiento a partir de los datos.
 - Se consideran modelos matemáticos para el reconocimiento de patrones, el aprendizaje automático y las técnicas de minería de datos

-
- **Optimización.** Los modelos de optimización permiten determinar la mejor solución a partir de un conjunto de acciones alternativas.
 - **Decisiones.** Corresponde a la elección y la adopción real de una decisión específica, y de alguna manera representa la conclusión natural del proceso de toma de decisiones.
 - La elección de una decisión es por parte de los responsables de la toma de decisiones, que también pueden aprovechar la información informal y no estructurada disponible para adaptar y modificar las recomendaciones y las conclusiones alcanzadas mediante el uso de modelos matemáticos.

Sistemas de apoyo a la toma de decisiones (Decision Support System, DSS)

- Un sistema de apoyo a la toma de decisiones (DSS) es una aplicación informática interactiva que combina datos y modelos matemáticos para ayudar a los responsables de la toma de decisiones a resolver problemas complejos que se enfrentan en la gestión de las empresas y organizaciones públicas y privadas.
- Las herramientas de análisis proporcionadas por una arquitectura de inteligencia empresarial pueden considerarse como DSS capaces de transformar los datos en información y conocimiento útiles para los responsables de la toma de decisiones.

1.1.1

Características de los DSS

- Permite extraer y manipular información de una manera flexible.
- Ayuda en decisiones no estructuradas.
- Permite al usuario definir interactivamente qué información necesita y cómo combinarla.
- Suele incluir herramientas de simulación, modelado, etc.
- Puede combinar información de los sistemas transaccionales internos de la empresa con los de otra empresa externa.
- Capacidad de análisis multidimensional (OLAP)

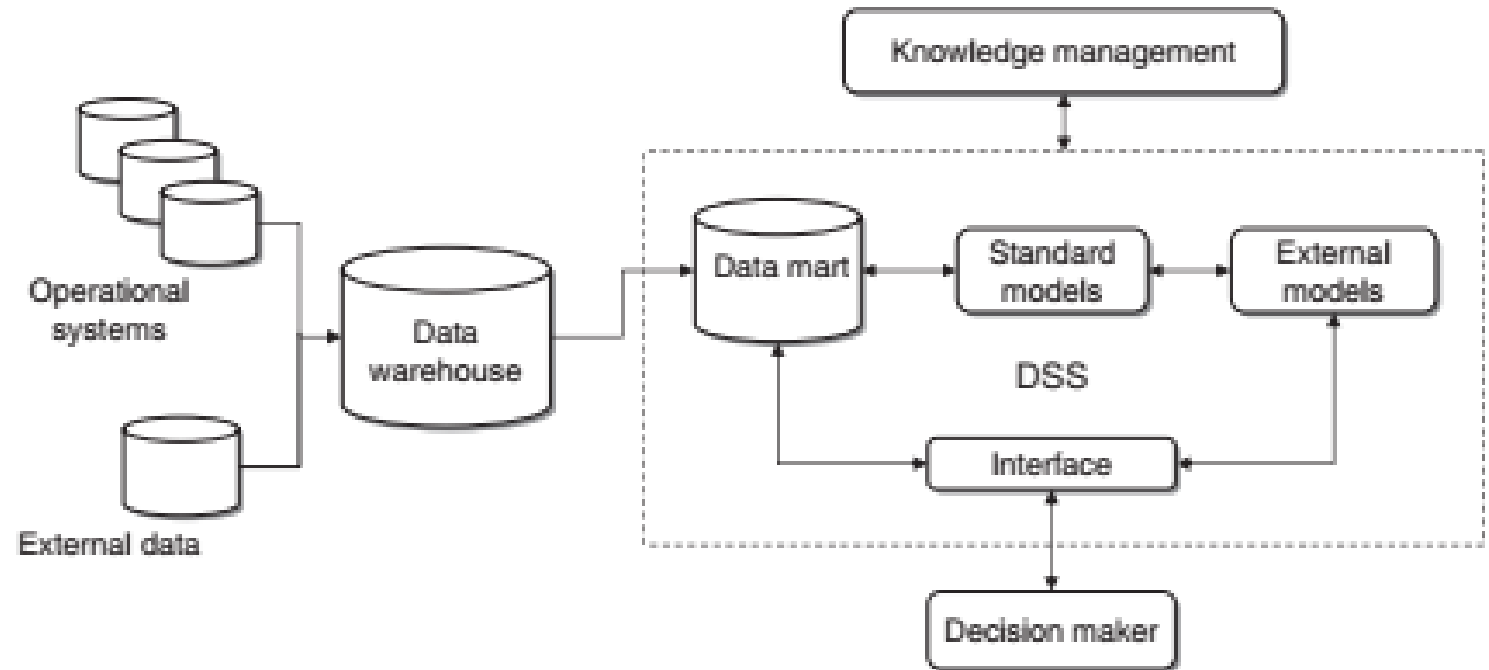
1.1.2 Taxonomía de los DSS

- Al igual que ocurre con la definición, no existe una taxonomía universalmente aceptada para los DSS. Diferentes autores proponen diferentes clasificaciones. Utilizando la relación con el usuario como criterio, Haettenschwiler distingue entre:
 - **DSS pasivo.**- Es un sistema de ayuda para el proceso de toma de decisiones, pero que no puede llevar a cabo una decisión explícita, sugerencias o soluciones.
 - **DSS activo.**- Puede llevar a cabo dicha decisión, sugerencias o soluciones.
 - **DSS cooperativo.**- Permite al encargado de la toma de decisiones (o a sus asesores) modificar, completar o perfeccionar las sugerencias de decisión proporcionadas por el sistema, antes de enviar de vuelta al sistema para su validación. El nuevo sistema mejora, completa y precisa las sugerencias del tomador de la decisión y las envía de vuelta a su estado para su validación. Entonces, todo el proceso comienza de nuevo, hasta que se genera una solución consolidada.

-
- Dependiendo del tipo de decisión, un DSS se pueden clasificar como estratégico, táctico u operativo.
 - **Decisiones estratégicas.** Las decisiones son estratégicas cuando afectan a toda la organización o al menos a una parte sustancial de ella durante un largo período de tiempo.
 - Influyen fuertemente en los objetivos y políticas generales de una empresa.
 - Se toman a un nivel organizativo superior, generalmente por parte de la alta dirección de la empresa.
 - **Decisiones tácticas.** Las decisiones tácticas sólo afectan a partes de una empresa y suelen estar restringidas a un único departamento.
 - El lapso se limita al horizonte a mediano plazo, típicamente hasta un año.
 - Se sitúan en el contexto determinado por las decisiones estratégicas.
 - En una jerarquía de la empresa, son tomadas por los mandos intermedios, como los jefes de los departamentos de la empresa.
 - **Decisiones operativas.** Las decisiones operacionales se refieren a actividades específicas que se llevan a cabo dentro de una organización y tienen un impacto modesto en el futuro.
 - Las decisiones operativas se enmarcan en los elementos y condiciones determinados por las decisiones estratégicas y tácticas.
 - Por lo general son realizadas a un nivel de organización inferior, por trabajadores responsables de una sola actividad o tarea, como jefes de oficina, jefes de taller o similares.

1.1.3

Arquitecturas y plataformas para los DSS



-
- **Gestión de datos.** El módulo de gestión de datos incluye una base de datos diseñada para contener los datos requeridos por los procesos de toma de decisiones.
 - En la mayoría de las aplicaciones la base de datos es un DM y mantiene una conexión con un DW empresarial, que representa el principal repositorio de los datos disponibles para desarrollar un análisis de BI.
 - **Gestión de modelos.** El módulo de gestión de modelos proporciona a los usuarios finales una colección de modelos matemáticos derivados de la investigación de operaciones, la estadística y el análisis financiero.
 - Por lo general, se trata de modelos relativamente sencillos que permiten llevar a cabo investigaciones analíticas que son muy útiles durante el proceso de toma de decisiones.

-
- **Interacciones.** Los trabajadores del conocimiento utilizan un DSS de forma interactiva para llevar a cabo sus análisis.
 - Se espera que el módulo responsable de estas interacciones reciba los datos de entrada de los usuarios de la manera más fácil e intuitiva, generalmente a través de la interfaz gráfica de un navegador web, y luego devuelva la información extraída y el conocimiento generado por el sistema en una forma gráfica apropiada.
 - **Gestión del conocimiento.** El módulo de gestión del conocimiento también está interconectado con el sistema integrado de gestión del conocimiento de la empresa.
 - Permite a los responsables de la toma de decisiones recurrir a las diversas formas de conocimiento colectivo, generalmente no estructurado, que representa la cultura corporativa.

		2015	2016	2017	2018	2019	2020	2021
Microsoft	Execution	67%	92%	99%	100%	100%	100%	100%
	Vision	60%	100%	100%	100%	100%	100%	100%
Tableau	Execution	100%	100%	100%	100%	100%	87%	74%
	Vision	63%	72%	74%	66%	66%	82%	80%
Qlik	Execution	76%	94%	71%	72%	73%	45%	53%
	Vision	60%	81%	57%	85%	85%	76%	82%
ThoughtSpot	Execution			31%	49%	49%	41%	47%
	Vision			29%	65%	65%	71%	87%
Sisense	Execution		35%	41%	58%	58%	28%	35%
	Vision		24%	65%	72%	72%	61%	71%
Oracle	Execution	56%		45%	34%	34%	32%	38%
	Vision	66%		14%	27%	26%	67%	67%
Looker	Execution				41%	41%	48%	60%
	Vision				17%	17%	32%	44%
TIBCO Software	Execution	49%	29%	51%	34%	34%	50%	44%
	Vision	63%	51%	44%	48%	48%	54%	59%
SAS	Execution	63%	77%	49%	46%	45%	25%	28%
	Vision	100%	68%	65%	55%	54%	66%	75%
Domo	Execution		58%	34%	42%	42%	18%	53%
	Vision		28%	16%	20%	20%	25%	41%
SAP	Execution	59%	69%	58%	37%	37%	22%	28%
	Vision	97%	68%	60%	55%	55%	70%	64%
Yellowfin	Execution	31%	19%	25%	14%	14%	19%	24%
	Vision	16%	15%	6%	15%	15%	63%	67%
IBM	Execution	59%	43%	49%	30%	29%	18%	31%
	Vision	93%	78%	68%	46%	46%	53%	51%
MicroStrategy	Execution	60%	61%	41%	65%	65%	54%	55%

Highlighting Tableau



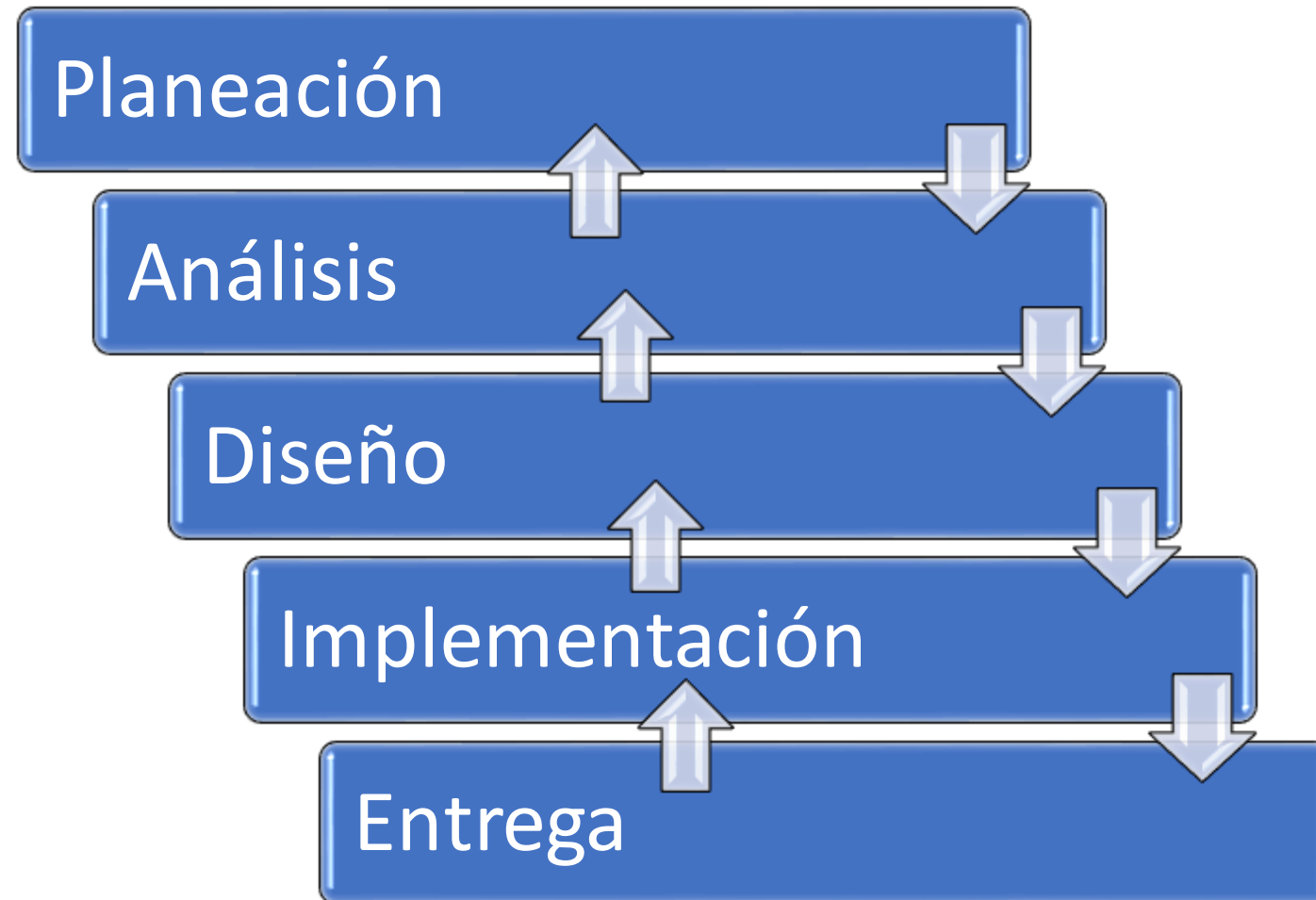
Resumen de las principales ventajas derivadas de la adopción de un DSS

- un aumento en el número de alternativas u opciones consideradas;
- un aumento del número de decisiones eficaces concebidas;
- una mayor conciencia y una comprensión más profunda del dominio analizado y de los problemas investigados;
- la posibilidad de ejecutar análisis hipotéticos variando las hipótesis y los parámetros de los modelos matemáticos;
- una capacidad mejorada para reaccionar con prontitud a eventos inesperados y situaciones imprevistas;
- una explotación de valor añadido de los datos disponibles;
- una mejor comunicación y coordinación entre los individuos y los departamentos de organización;
- un desarrollo más eficaz del trabajo en equipo;
- una mayor fiabilidad de los mecanismos de control, debido a la mayor inteligibilidad del proceso de decisión.

Desarrollo de un sistema de apoyo a la toma de decisiones

- A diferencia de otras aplicaciones de software, los DSS no suelen estar disponibles como programas estándar.
- Los entornos de análisis multidimensional han facilitado y estandarizado el acceso a las funciones pasivas de inteligencia empresarial.
- Sin embargo, para desarrollar la mayoría de los DSS todavía se requiere un proyecto específico.

Fases del desarrollo de un DSS



-
- **Planificación.** Se comprenden las necesidades y oportunidades para traducirlas en un proyecto y más tarde en un DSS exitoso.
 - La planificación suele implicar un estudio de viabilidad para abordar la cuestión: ¿Por qué se quiere desarrollar un DSS?
 - Durante el análisis de viabilidad, se establecen objetivos generales y específicos del sistema, destinatarios, posibles beneficios, tiempos de ejecución y costos. Si se decide continuar con el sistema, las fases de planificación deben ir seguidas de la definición de las actividades, tareas, responsabilidades y fases de desarrollo, para las que deben utilizarse metodologías clásicas de gestión de proyectos.

-
- **Análisis.** Es necesario definir a detalle las funciones del DSS, desarrollando y elaborando las conclusiones preliminares alcanzadas durante el estudio de viabilidad.
 - Por lo tanto, debe darse una respuesta a la siguiente pregunta: ¿Qué debe lograr el DSS y quién lo utilizará, cuándo y cómo?
 - Es necesario analizar los procesos de decisión a apoyar, para tratar de comprender a fondo todas las interrelaciones existentes entre los problemas abordados y el entorno circundante.
 - Por último, es necesario explorar los datos para comprender cuánta y qué tipo de información ya existe y qué información puede recuperarse de fuentes externas.

-
- **Diseño.** Durante la fase de diseño la pregunta principal es: ¿Cómo funcionará el DSS?
 - Se define toda la arquitectura del sistema, a través de la identificación de las plataformas de tecnología de hardware, la estructura de la red, las herramientas de software para desarrollar las aplicaciones y la base de datos específica que se utilizarán.
 - También es necesario definir en detalle las interacciones con los usuarios, mediante máscaras de entrada, visualizaciones gráficas en pantalla e informes impresos.
 - Otro aspecto que debe aclararse durante la fase de diseño es la elección de hacer o comprar, ya sea subcontratar la implementación del DSS a terceros, en su totalidad o en parte.

-
- **Implementación.** Una vez establecidas las especificaciones, es el momento de la ejecución, las pruebas y la instalación real, cuando el DSS se pone en marcha y se pone a trabajar.
 - Los problemas a los que se enfrenta esta última fase se remontan a los métodos de gestión de proyectos.
 - Otro aspecto de la fase de aplicación, que a menudo se pasa por alto, se refiere al impacto global en la organización determinado por el nuevo sistema.
 - Estos efectos deben controlarse mediante técnicas de gestión del cambio, asegurándose de que nadie se sienta excluido del proceso de innovación organizacional y rechace el DSS.

Prototipado del sistema

- El rápido desarrollo de prototipos ofrece claras ventajas.
 - Cada subsistema puede desarrollarse en realidad más rápidamente y, por lo tanto, está más fácilmente disponible.
 - Además, cuando un subsistema se libera a los usuarios, es posible verificar su conformidad con el propósito previsto y probar sus funciones, incluso si éstas aún no están completamente desarrolladas.
 - El desarrollo evolutivo de un DSS permite minimizar el riesgo de fracaso.
 - Además, las pruebas intermedias permiten corregir rápidamente la mayoría de los errores de diseño.
- Otros métodos de desarrollo pueden ser adoptados eficazmente con el fin de acelerar la implementación del software.
 - Estos incluyen técnicas de desarrollo ágil y técnicas de programación extremas.

Principales factores críticos que pueden afectar el grado de éxito de un DSS

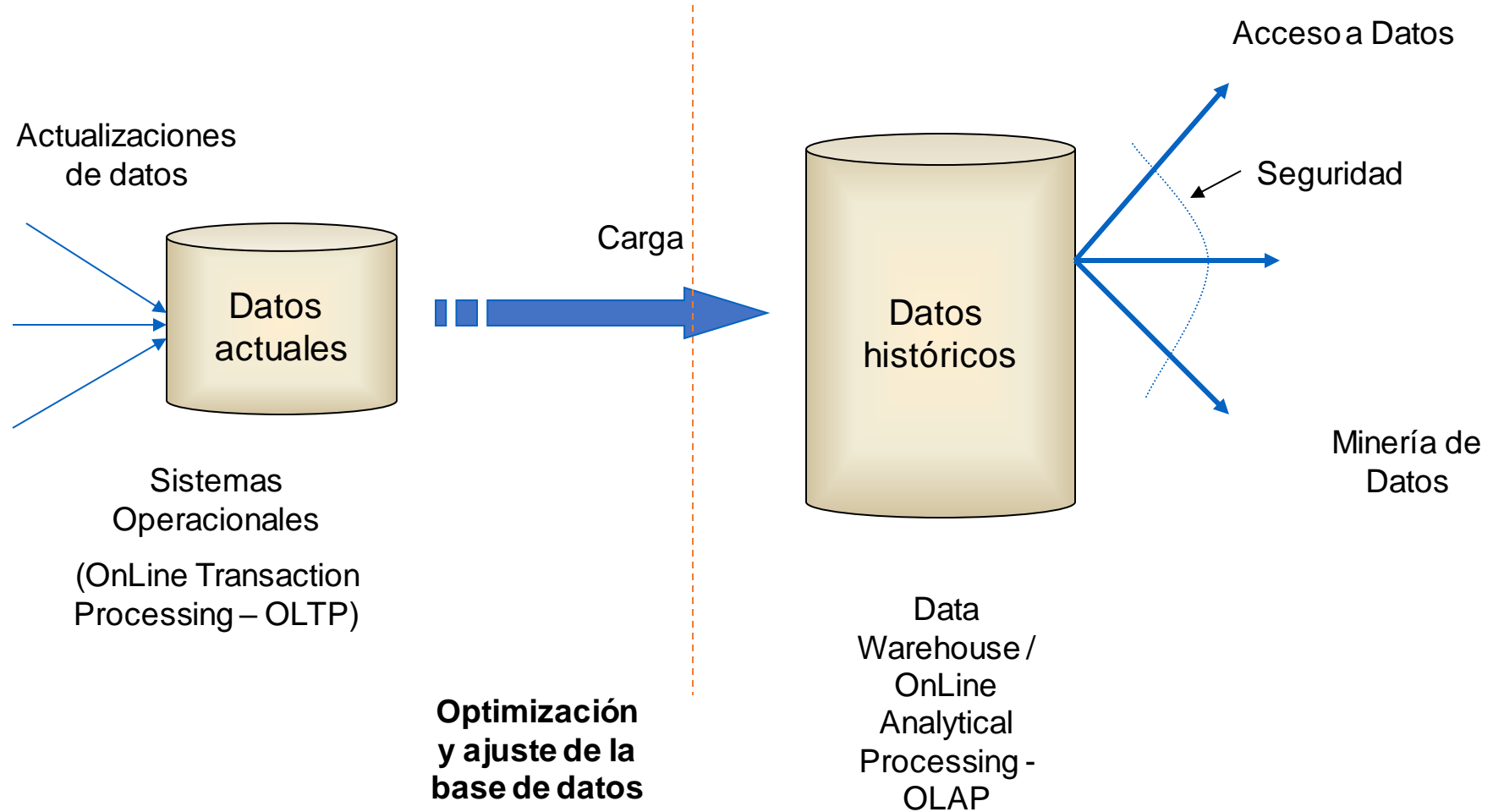
- **Integración.** Un integrador de sistemas es un experto en todos los aspectos involucrados en el desarrollo de un DSS, tales como arquitecturas de sistemas de información, procesos de toma de decisiones, modelos matemáticos y métodos de solución.
 - También puede ejercer una influencia positiva como agente de innovación capaz de superar la mayor parte de la resistencia al cambio que a menudo surge en cada organización.
- **Participación.** La exclusión o marginación del equipo del proyecto de los trabajadores del conocimiento es un error que a veces se realiza durante el diseño y desarrollo de un DSS.
 - La participación de los responsables de la toma de decisiones y de los usuarios durante el proceso de desarrollo es de vital importancia para evitar rechazar una herramienta que perciben como extraña.
- **Incertidumbre.** En general, los costos no son una preocupación importante en la aplicación de un DSS, y la ventaja de idear decisiones más eficaces compensa en gran medida los costos de desarrollo incurridos.
 - Es conveniente reducir la incertidumbre del proyecto mediante la creación de prototipos, la facilidad de uso, las pruebas del sistema durante las fases preliminares y una aplicación evolutiva.


1.2 Almacenes de Datos

Almacén de datos (Data Warehouse)

- Desde mediados de la década de 1990 se sintió la necesidad de una lógica y la separación de materiales entre las bases de datos que alimentan los datos de entrada en los DSS y arquitecturas de inteligencia empresarial, por un lado, y sistemas de información operacionales por otro.
- Definición de almacén de datos
 - William Inmon, *“El Data Warehouse es una colección de datos, orientados a un tema, integrados, no volátiles, variantes en el tiempo, organizados para el apoyo a toma de decisiones.”*
 - Ralph Kimball, *“Un Data Warehouse es una copia de los datos transaccionales, específicamente diseñada para realizar consultas y análisis.”*
 - El término almacén de datos (DW) indica todo el conjunto de actividades interrelacionadas con el diseño, la implementación y el uso de un almacén de datos.
- En este sentido, el principal objetivo del DW es facilitar la consulta de los datos por parte de los usuarios. Con ello, es posible generar reportes históricos, análisis estadísticos, predicciones, comparaciones, proyecciones y demás. Esto con la finalidad de poder tomar decisiones estratégicas cada vez más acertadas.

Concepto de DW





Un Data Warehouse se crea al extraer datos desde una o más bases de datos de aplicaciones operacionales.

Los datos extraídos son transformados para eliminar inconsistencias, si es necesario y luego ser cargados en el Data Warehouse.

El proceso de transformar, crear el detalle de tiempo variante, resumir y combinar las extracciones de datos, ayudan a crear el ambiente para el acceso a la información Institucional.

Finalidad del DW

La finalidad de un DW consiste en auxiliar a la administración a comprender el pasado y planear el futuro. La administración busca respuestas a preguntas como:

- ¿Qué están comprando nuestros clientes? ¿Qué no están comprando? ¿Qué incentivos han funcionado antes con los mismos clientes en esta época del año?.
- ¿Cuántos de nuestros vendedores visitan a un mismo cliente?.
- ¿Qué están haciendo nuestros competidores?
- ¿Cómo se comparan nuestros costos para cada línea de producto durante los últimos tres años?.

La promesa del DW es “sacar datos” de los sistemas operacionales para ayudar a las empresas a tomar mejores decisiones.

1.2.1 Características de un Almacén de Datos

Orientado a entidades. Los datos contenidos en un DW se extraen de las principales entidades de interés para el análisis, tales como productos, clientes, pedidos y ventas. Por otro lado, los sistemas transaccionales están más orientados hacia las actividades operativas y se basan en cada transacción única registrada. La orientación hacia las entidades permite que el desempeño de una empresa sea evaluado más fácilmente.

Integrado. Los datos procedentes de las diferentes fuentes están integrados y homogeneizados a medida que se cargan en un almacén de datos.

Variante de tiempo. Todos los datos introducidos en un almacén de datos se etiquetan con la hora y período al que se refieren. Como consecuencia, la dimensión de tiempo en cualquier almacén de datos es un elemento crítico. De esta manera, las aplicaciones DSS pueden desarrollar una tendencia de análisis histórico.

Persistente. Una vez que se han cargado datos en un DW, suelen ser no se modificables y se mantienen de forma permanente. Esta característica facilita organizar el acceso de solo lectura por parte de los usuarios y simplifica el proceso de actualización, evitando concurrencia de escritura que es de importancia crítica para los sistemas operacionales.

Fusionado. Los datos almacenados en un DW se obtienen de los datos primarios que pertenecen a los sistemas operacionales. Por un lado, la reducción del espacio necesario para almacenar en el DW los datos acumulados a lo largo de los años; por otro lado, la información consolidada puede ser capaz de satisfacer mejor las necesidades del negocio en el sistema de inteligencia.

Desnormalizado. A diferencia de las bases de datos operacionales, los datos almacenados en un DW no están estructurados en forma normal, sino que pueden prever redundancias, para permitir un tiempo de respuesta más corto a consultas complejas.

- Es posible identificar tres categorías principales de datos que alimentan un DW: datos internos, datos externos y datos personales.
 - **Datos internos.** Los datos internos se recopilan a través de transacciones que presiden rutinariamente las operaciones de una empresa, como administración, contabilidad, producción y logística. Estos datos suelen proceder de diferentes componentes del sistema de información:
 - sistemas de back-office, que recopilan registros transaccionales básicos, como pedidos, facturas, inventarios, datos de producción y logística;
 - sistemas front-office, que contienen datos procedentes de actividades de call-center, asistencia al cliente, ejecución de campañas de marketing;
 - sistemas basados en web, que recopilan transacciones de ventas en sitios web de comercio electrónico, visitas a sitios web, datos disponibles en formularios rellenados por sitios web existentes y clientes potenciales.

- **Datos externos.** Hay varias fuentes de datos externos que se pueden utilizar para ampliar la riqueza de la información almacenada en las bases de datos internas. Una fuente importante de datos externos la proporciona sistemas de información geográfica (SIG), que representan un conjunto de aplicaciones para adquirir, organizar, almacenar y presentar datos territoriales. Estos contienen información en relación con las entidades que tienen una posición geográfica específica.
- **Datos personales.** En la mayoría de los casos, los responsables de la toma de decisiones que realizan un análisis de inteligencia empresarial también se basan en la información y las evaluaciones personales almacenadas en hojas de cálculo o bases de datos locales ubicadas en sus equipos. La recuperación de tal información y su integración con datos estructurados de origen interno y externo fuentes es uno de los objetivos de los sistemas de gestión del conocimiento.

Ventajas de DW

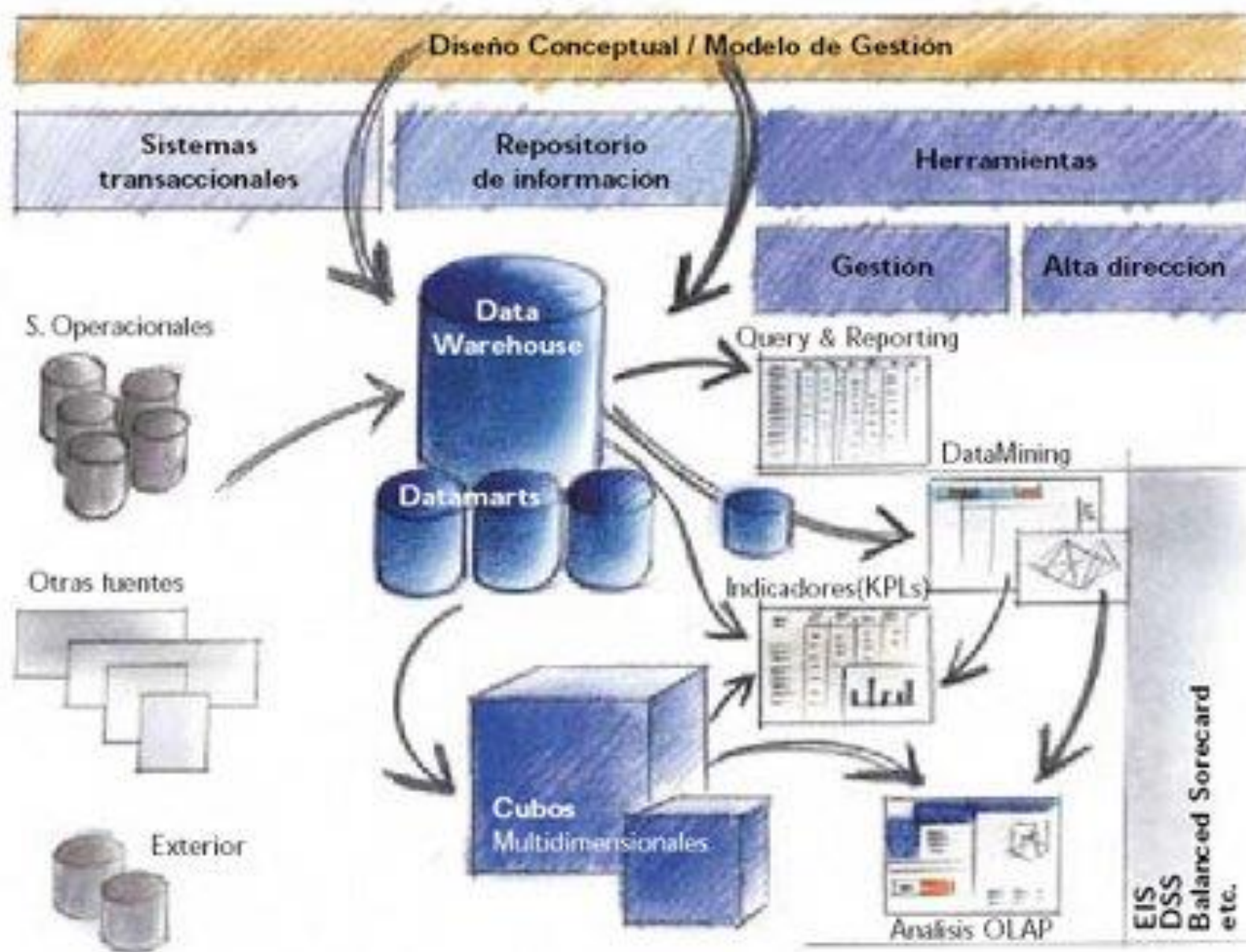
- **Integración.** En muchos casos, los DSS deben acceder a información procedente de varias fuentes de datos, distribuidas en diferentes partes de una organización o derivada de fuentes externas. La integración de datos puede lograrse mediante diferentes técnicas, por ejemplo, mediante el uso de métodos de codificación uniformes, la conversión a unidades de medida estándar y el logro de una homogeneidad semántica de información.
- **Calidad.** Los datos transferidos desde los sistemas operacionales al DW se examinan y corrigen con el fin de obtener información fiable y libre de errores, en la medida de lo posible.
- **Eficacia.** Si una consulta se dirigiera a los sistemas transaccionales, se correría el riesgo de comprometer gravemente la eficiencia requerida por las aplicaciones. Una mejor solución es dirigir estas consultas para análisis OLAP al almacén de datos, separadas físicamente de los sistemas operacionales.
- **Capacidad de ampliación.** Los datos almacenados en sistemas transaccionales se extienden a lo largo de un limitado lapso en el pasado, y se archivan permanentemente en dispositivos de almacenamiento masivo fuera de línea. Los sistemas de BI necesitan acceder a todos los datos pasados disponibles para poder captar tendencias y detectar patrones recurrentes. Esto es posible debido a la capacidad de los almacenes de datos para retener información histórica.

Departamento de Datos (Data Mart)

- Los departamentos de datos (Data Marts DM) son sistemas que recopilan todos los datos requeridos por un departamento específico de la empresa, como marketing o logística, con el fin de realizar análisis de BI y ejecutar aplicaciones de apoyo a la toma de decisiones específicas para la propia función.
- Por lo tanto, un DM puede considerarse como un almacén de datos funcional o departamental de un tamaño menor y un tipo más específico que el DW general de la empresa.
- Con el fin de implementar aplicaciones de BI, algunas empresas se someten a diseñar y desarrollar de forma incremental una serie de DM integrados en lugar de un DW central, con el fin de reducir el tiempo de implementación y las incertidumbres relacionadas con el proyecto.

1.2.2 Arquitecturas de Data Warehouses (almacenes de datos) y Data Mart (base de datos departamental)

- La arquitectura de referencia de un DW incluye los siguientes componentes funcionales principales.
 - El propio DW, junto con los DMs adicionales, que contienen los datos y las funciones que permiten acceder a los datos, visualizarse y quizás modificarse.
 - Aplicaciones de adquisición de datos, también conocidas como herramientas de extracción, transformación y carga (ETL) o back-end, que permiten extraer, transformar y cargar los datos en el DW.
 - Aplicaciones de inteligencia empresarial y de apoyo a la toma de decisiones, que representan el front-end y permiten a los trabajadores del conocimiento llevar a cabo los análisis y visualizar los resultados.



- A nivel empresarial se consideran:
 - El nivel de las fuentes de datos y las herramientas ETL relacionadas que normalmente se instalan en uno o más servidores.
 - El DW y cualquier DM, posiblemente disponible en uno o más servidores, y separado de los que contienen las fuentes de datos. Este segundo nivel también incluye los metadatos que documentan el origen y el significado de los registros almacenados en el DW.
 - El nivel de los análisis que aumentan el valor de la información contenida en un almacén de datos a través de herramientas de consulta, presentación de informes y, posiblemente, de apoyo a la toma de decisiones sofisticadas.

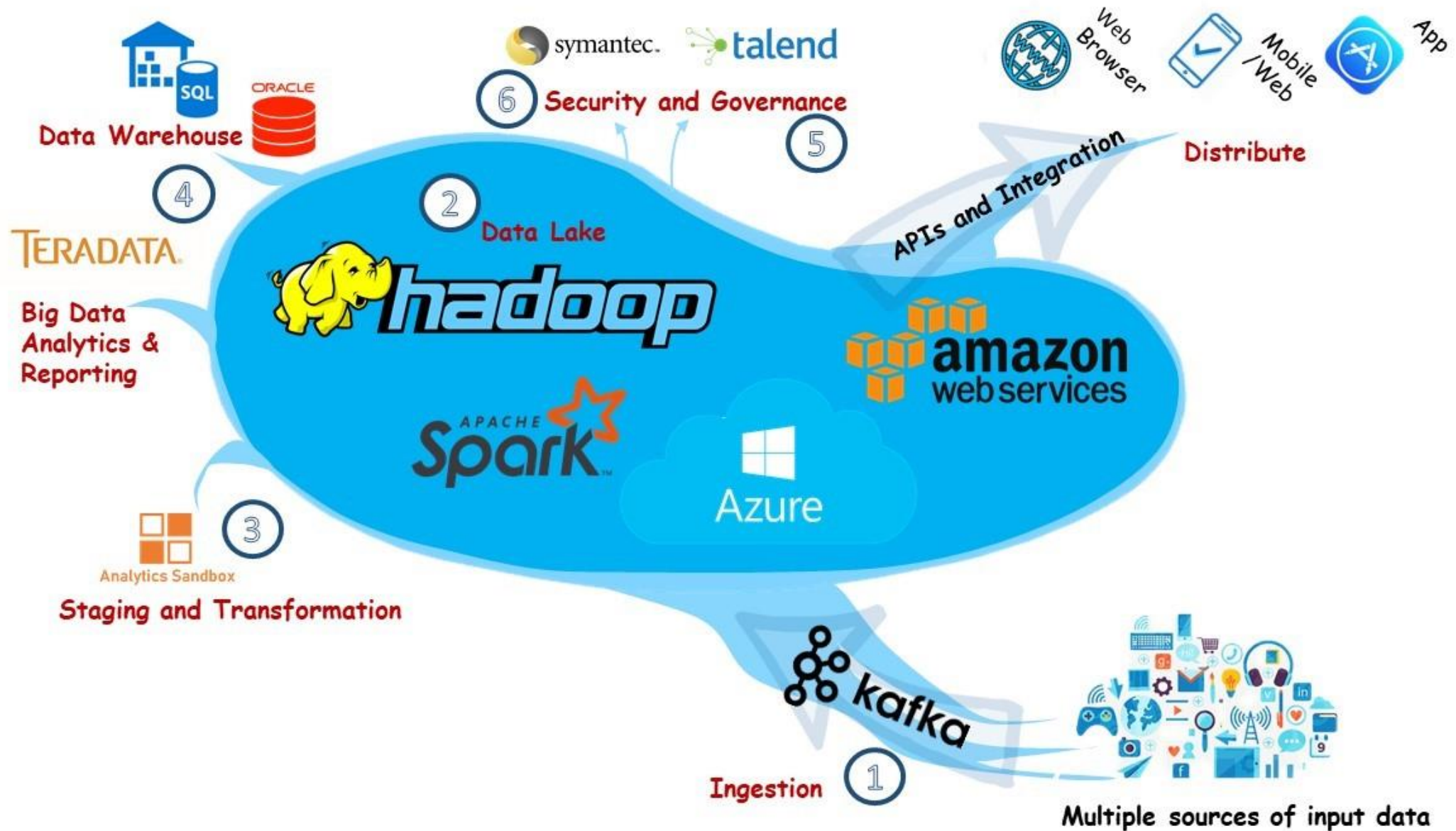
1.2.3

Aplicaciones de DW

- **DW en el sector de telecomunicaciones**
- Las empresas de telecomunicaciones utilizan DW para almacenar los datos de millones de clientes. Esto implica el respaldo de facturas, servicios utilizados, registros de llamadas realizadas, equipos vendidos, entre otros. Toda esta información es de gran utilidad para actividades como:
 - El diseño de estrategias de marketing
 - Las auditorías en el área de operaciones
 - Los análisis sobre la prestación de los servicios
 - Las previsiones de riesgos de fuga de clientes y demás
- **DW en el sector de consumo masivo**
- Las empresas de consumo masivo implementan el DW para mantener su competitividad en el mercado. De esta forma pueden predecir, por ejemplo, la cantidad de producción que necesitarán para satisfacer la demanda en un rango de tiempo determinado. Las cadenas minoristas también pueden compartir ciertos accesos de sus DW con sus proveedores. Esto les dará a los fabricantes información relacionada con el suministro de productos y la venta de los mismos al consumidor final, además de acceder a datos determinantes para la elaboración de campañas de marketing.
- **DW en el sector de transporte**
- Tanto en el sector de viajes como en el de distribución, el uso del DWH es para almacenar información de los clientes, destinos más frecuentados, administración de transportes de carga, seguimiento del equipaje, así como datos como las reservaciones de viaje a un determinado destino, o los tiempos de entrega de los pedidos, permitirán desarrollar análisis para la creación de promociones o para diagnósticos de los procesos logísticos de la organización.

1.2.4 Data Lake (Lagos de Datos, DL)

- Un data lake es un repositorio de almacenamiento centralizado que contiene Big Data de varias fuentes en un formato granular y sin procesar.
 - Puede guardar datos estructurados, semiestructurados o no estructurados, lo que significa que los datos pueden conservarse en un formato más flexible para usarlos en un futuro.
 - Al guardar datos, un DL los asocia con identificadores y etiquetas de metadatos para poder extraerlos más rápidamente.
- El término «data lake» fue acuñado por James Dixon, director tecnológico de Pentaho, y hace referencia a la naturaleza particular de los datos de un DL, a diferencia de los datos limpios y procesados guardados en los sistemas tradicionales de almacenes de datos.
- Los DL suelen estar configurados sobre un clúster de hardware de consumo económico y escalable. Esto permite volcar los datos al DL por si se necesitan más adelante sin tener que preocuparse por la capacidad de almacenamiento. Los clústeres pueden existir localmente o en la cloud.



1.2.5

Características de un Data Lake

- Un DL funciona a partir de un principio llamado *schema-on-read* o esquema contra escritura. Esto significa que no existe un esquema predefinido en el que deban encajarse los datos antes de almacenarlos.
- Un único repositorio compartido de datos, normalmente almacenado en el Sistema de archivos distribuido (DFS).
- Incluye funcionalidades de orquestación y programación de trabajos (por ejemplo, a través de YARN).
- Contiene un conjunto de aplicaciones o flujos de trabajo para consumir, procesar o actuar sobre los datos.
- El fácil acceso de los usuarios, ya que los propietarios de datos pueden entonces consolidar datos de clientes, proveedores y operaciones, eliminando barreras técnicas e incluso políticas para compartir datos.

1.2.6 Arquitecturas de Data Lake

- Un clúster de Hadoop de servidores distribuidos sirve de almacenamiento de Big Data.
 - En el núcleo de Hadoop es su capa de almacenamiento, el HDFS (Sistema de archivos distribuidos de Hadoop), que almacena y replica los datos en múltiples servidores.
 - YARN es el gestor de recursos que decide cómo programarlos en cada nodo.
 - MapReduce es el modelo de programación que utiliza Hadoop para dividir los datos en subconjuntos más pequeños y procesarlos en sus clústeres de servidores.

- Más allá de estos tres componentes básicos, el ecosistema Hadoop engloba varias herramientas suplementarias, como Hive, Pig, Flume, Sqoop y Kafka que ayudan con la ingesta, la preparación y la extracción de datos.
- Los DL de Hadoop pueden configurarse localmente o en la cloud mediante plataformas de empresa como Cloudera o HortonWorks.
 - Otros DL cloud, como Azure, envuelven sus funcionalidades alrededor de la arquitectura Hadoop.

- Amazon Simple Storage Service (Amazon S3) que proporciona la función de almacenamiento.
 - Kinesis Streams, Kinesis Firehose, Snowball y Direct Connect son herramientas de ingesta de datos que permiten a los usuarios transferir cantidades de datos a S3.
- Además de S3, existe DynamoDB, una base de datos No-SQL de baja latencia, y Elastic Search, un servicio que ofrece un mecanismo simplificado para realizar consultas en el DL.
 - Los Cognito User Pools definen la autenticación de usuarios y el acceso al DL.
 - Los servicios como Security Token Service, Key Management Service, CloudWatch y CloudTrail garantizan la seguridad de los datos.
 - Para el procesamiento y la analítica existen herramientas como RedShift, QuickSight, EMR y ML.
- La oferta de productos de AWS presenta una curva de aprendizaje inicial muy empinada. No obstante, sus completas funciones tienen un uso muy extendido en las aplicaciones de BI.

Data lakes en AWS

Data lakes en Azure

- La capa de almacenamiento se denomina Azure Data Lake Store (ADLS) y la analítica contiene dos componentes: Azure Data Lake Analytics y HDInsight.
- ADLS está creada siguiendo la norma HDFS y tiene una capacidad de almacenamiento ilimitada.
 - Es capaz de almacenar billones de archivos en un único archivo de más de un petabyte.
 - ADLS permite guardar datos en cualquier formato y es seguro y escalable.
 - Es compatible con cualquier aplicación que utilice la norma HDFS.
- HDInsight es un servicio analítico de DL en cloud. Creado sobre Hadoop YARN, permite acceder a los datos mediante herramientas como Spark, Hive, Kafka y Storm.
 - Es compatible con la seguridad de nivel empresarial debido a su integración con Azure Active Directory.
- Azure Data Lake Analytics es otro servicio analítico, pero su enfoque es distinto.
 - En lugar de emplear herramientas como Hive, utiliza un lenguaje llamado U-SQL, que es una combinación de SQL y C#, para acceder a los datos.
 - Es idóneo para procesamientos de Big Data por lotes, dado que ofrece mayor velocidad a menor coste (tan solo se paga por las tareas que se utilicen).

1.2.7 Comparativa entre Data Lakes y Data Marts

Características	Almacén de datos	Data Lake
Datos	Relacional de sistemas transaccionales, bases de datos operacionales y aplicaciones de línea de negocios	No relacional y relacional de dispositivos IoT, sitios web, aplicaciones móviles, redes sociales y aplicaciones corporativas
Esquema	Diseñado antes de la implementación de DW (esquema en escritura)	Escrito en el momento del análisis (schema-on-read)
Precio / rendimiento	Resultados de consulta más rápidos con un mayor costo de almacenamiento	Los resultados de las consultas se vuelven más rápidos usando almacenamiento de bajo costo
Calidad de datos	Datos altamente curados que sirven como la versión central de los reales	Cualquier información que pueda o no ser curada (es decir, datos sin procesar)
Usuarios	Analistas comerciales	Científicos de datos, desarrolladores de datos y analistas de negocios (usando datos curados)
Analítica	Informe por lotes, BI y visualizaciones	Aprendizaje automático, análisis predictivo, descubrimiento de datos y creación de perfiles

DIFERENCIAS	
Data Lake	Data Warehouse
Se almacena toda la información independientemente de la fuente y su estructura.	Se almacena información, proveniente de sistemas transaccionales, que pasa por un proceso de normalización y transformación.
Funciona como un "repositorio" de datos estructurados, semi-estructurados y no-estructurados.	Se combinan tecnologías para recopilar y gestionar datos procedentes de distintas fuentes (suelen almacenarse de manera estructurada).
Se define el esquema de la data después de su almacenamiento, al momento de ser usada.	Se define el esquema antes del almacenamiento y antes de su uso.
Usa el proceso extracción-carga-transformación.	Usa el proceso extracción-transformación-carga.
Ideal para hacer todo tipo de análisis en profundidad.	Ideal para usuarios operacionales.

Ventajas de DL

- Capacidad de obtener valor a partir de tipos ilimitados de datos
- Posibilidad de almacenar todo tipo de datos estructurados y no estructurados, desde datos de CRM hasta publicaciones en redes sociales
- Mayor flexibilidad: no tiene que tener todas las respuestas por adelantado
- Posibilidad de almacenar datos en bruto: puede refinarlo a medida que su comprensión mejore
- Formas ilimitadas de consultar los datos
- Aplicación de una variedad de herramientas para obtener una idea de lo que significan los datos
- Acceso democratizado a los datos a través de una única vista unificada de datos en toda la organización cuando se utiliza una plataforma de gestión de datos efectiva