Módulo 2: Conceptos de análisis y tecnología de Big Data

Introducción

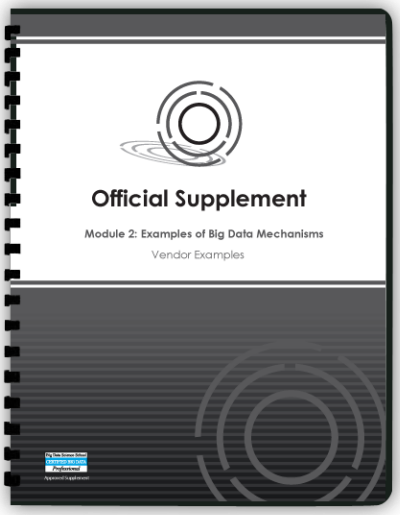
Este es el cuadernillo oficial del **Módulo 2: Conceptos de análisis y tecnología de Big Data** para el curso del BDSCP y su respectivo **Examen B90.02** de Pearson VUE.

Este documento consta de las siguientes tres partes:

- Parte I: ciclo de vida del análisis de Big Data
- Parte II: conceptos de análisis de Big Data
  - Técnicas de análisis estadístico
  - Técnicas de análisis semántico
  - Técnicas de aprendizaje automático (Machine Learning)
  - Técnicas de análisis visual
  - Relación de temas de análisis
- Parte III: conceptos tecnológicos de Big Data
  - Consideraciones tecnológicas de Big Data
  - Mecanismos tecnológicos de Big Data

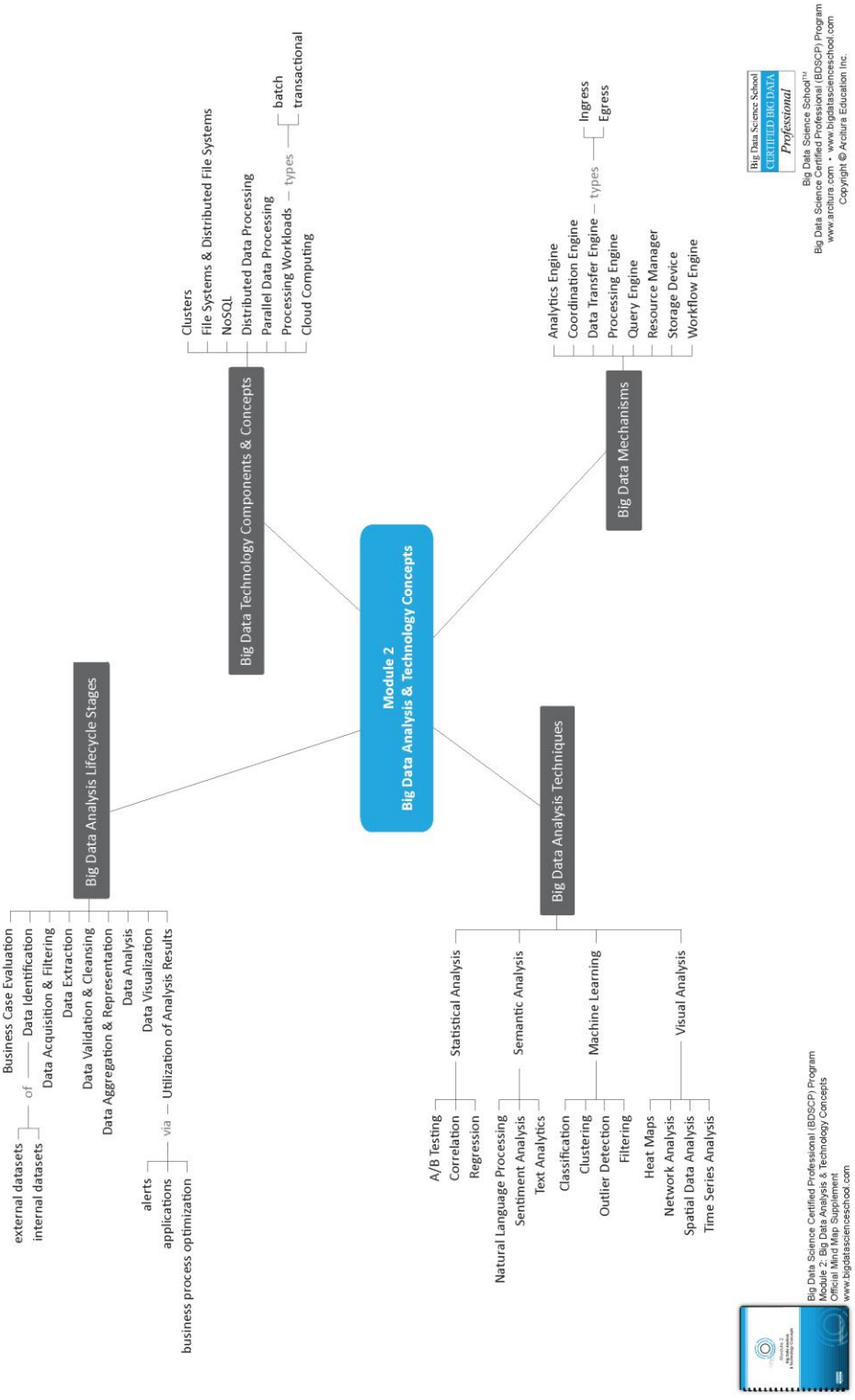
Complemento oficial: ejemplos de mecanismos de Big Data

Este complemento relaciona cada uno de los mecanismos de Big Data presentados en este módulo con uno o más productos o tecnologías de código abierto y/o de proveedores.



Póster del mapa mental

El *Póster del mapa mental del Módulo 2 del BDSCP* incluido en este cuadernillo ofrece una representación visual alternativa de los principales temas abordados en este curso.



Parte I: ciclo de vida del análisis de Big Data

El análisis de Big Data se diferencia del análisis de datos tradicional principalmente en lo relacionado a las características de volumen, velocidad y variedad de los datos procesados.

A fin de abordar las circunstancias y requisitos concretos que implica el análisis y la analítica de Big Data, se requiere un proceso fundamental, paso a paso, que permita organizar las tareas relacionadas con la recolección, procesamiento, producción y reutilización de los datos.

En las secciones siguientes se explora el ciclo de vida del análisis de Big Data que evidencia dichas tareas.

Etapas del ciclo de vida

El ciclo de vida de análisis de Big Data se puede dividir en las siguientes nueve etapas, como se muestra en la Figura 2.1:

1. Evaluación del caso empresarial
2. Identificación de datos
3. Adquisición y filtrado (filtering) de datos
4. Extracción de datos
5. Validación y limpieza (Cleansing) de datos
6. Agregación y representación de datos
7. Análisis de datos (Data Analysis)
8. Visualización de datos
9. Uso de los resultados del análisis

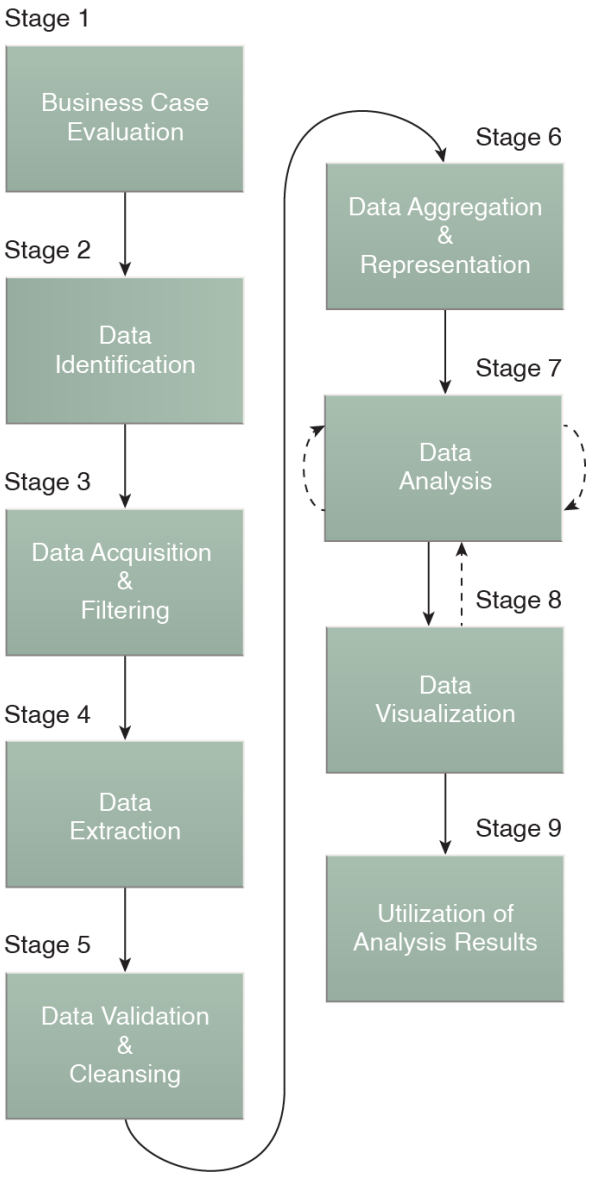
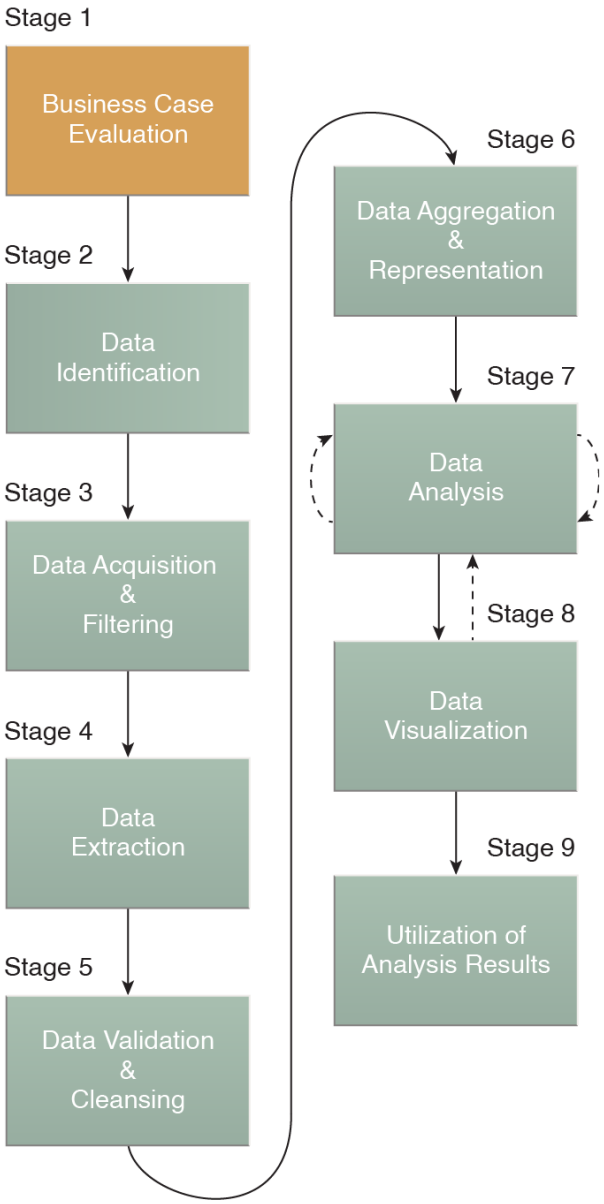


Figura 2.1 – Ciclo de vida de análisis de Big Data

Etapa 1: evaluación del caso empresarial

Todo ciclo de vida de análisis de Big Data debe comenzar con un alcance empresarial bien definido y un entendimiento claro de la justificación, motivación y metas de la ejecución del análisis. En la etapa de evaluación del caso empresarial es necesario crear, evaluar y aprobar un caso empresarial antes de proceder con las tareas reales y prácticas de análisis.



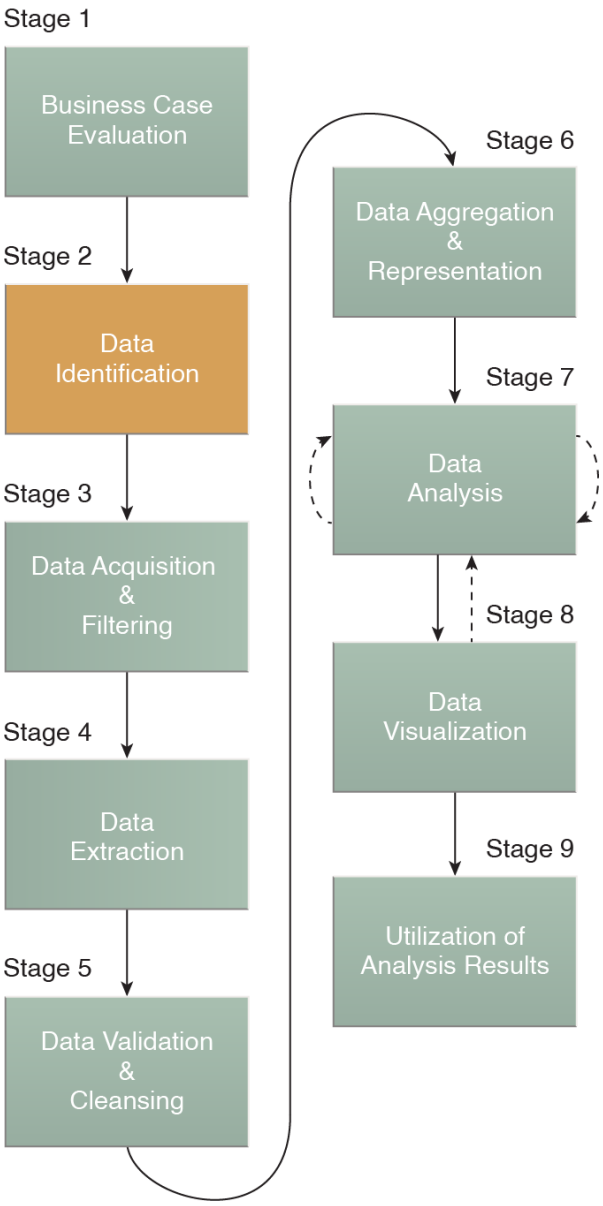
La evaluación de un caso empresarial de análisis de Big Data ayuda a las personas encargadas de tomar las decisiones a entender los recursos empresariales que se necesitarán y a saber qué retos empresariales se enfrentarán durante el análisis. La identificación más detallada de los KPI durante esta etapa ayuda a determinar qué tan cerca el resultado de los análisis de datos (Data Analysis) debe estar de alcanzar las metas y objetivos identificados.

Con base en los requisitos empresariales documentados en el caso empresarial se puede determinar si los problemas empresariales que se están abordando en realidad corresponden a problemas de Big Data. Para que un problema empresarial sea considerado un problema de Big Data, debe estar relacionado con una o más características de volumen, velocidad o variedad de Big Data.

Tenga en cuenta que otro resultado de esta etapa es la determinación del presupuesto subyacente necesario para llevar a cabo el proyecto de análisis. La adquisición necesaria de herramientas, hardware, capacitación, entre otros, debe ser conocida con anticipación para que la inversión prevista pueda ser sopesada frente a los beneficios esperados al alcanzar las metas. Para las iteraciones iniciales del ciclo de vida de análisis de Big Data será necesario realizar una mayor inversión inicial en tecnologías, productos y capacitación sobre Big Data, en comparación con las iteraciones realizadas después, en las cuales dicha inversión inicial puede ser aprovechada repetidamente.

Etapa 2: identificación de datos

La etapa de identificación de datos está orientada a identificar los datasets necesarios para el proyecto de análisis, así como las fuentes de los mismos.



Identificar una amplia variedad de fuentes de datos puede aumentar la probabilidad de encontrar patrones y correlaciones ocultos. Por ejemplo, puede ser útil identificar la mayor cantidad de tipos de fuentes de datos relacionados y de información posible, en especial cuando no hay certeza sobre qué se busca exactamente.

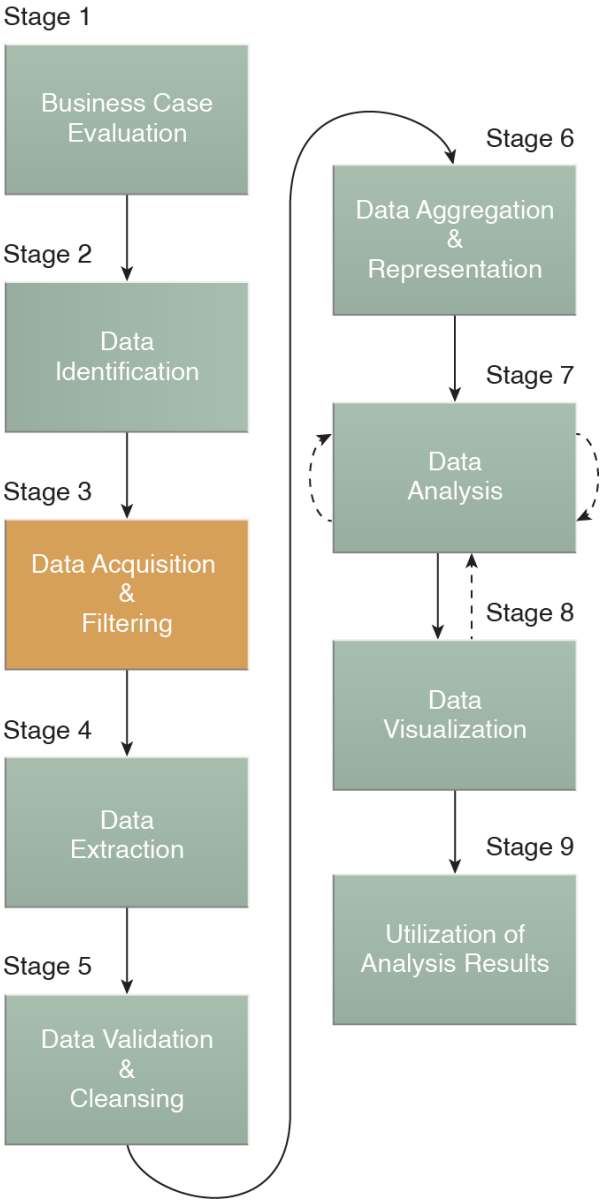
Dependiendo del alcance empresarial del proyecto de análisis y de la naturaleza de los problemas de la empresa que se estén tratando, los datasets requeridos y sus fuentes pueden ser internos o externos a la empresa.

En el caso de los datasets internos, por lo general, se compila y combina una lista de datasets disponibles desde fuentes internas, como data marts y sistemas operacionales, y se compara con una especificación predefinida del dataset.

En el caso de los datasets externos, por lo general, se compila y combina una lista de posibles proveedores externos de datos, como mercados de datos y datasets a disposición del público. Algunos tipos de datos externos pueden estar integrados en blogs u otras clases de sitios web basados en contenido, en cuyo caso sería necesario recolectar los datos utilizando herramientas automatizadas.

Etapa 3: adquisición y filtrado (filtering) de datos

Durante la etapa de adquisición y filtrado (filtering) de datos, se recopilan los datos de todas las fuentes de datos identificadas en la etapa anterior; luego, los datos son sometidos al filtrado (filtering) automático de datos corruptos o datos sin valor para los objetivos del análisis.



Dependiendo del tipo de fuente de datos, los datos pueden estar en forma de archivos de volcado de memoria; por ejemplo, los datos comprados a proveedores externos de datos; o también pueden requerir integración de API, por ejemplo, Twitter. En muchos casos, cuando se trata de datos externos sin estructurar, algunos o la mayoría de los datos adquiridos pueden ser no significativos (ruido) y se pueden descartar en el proceso de filtrado (filtering).

Los datos clasificados como "corruptos" pueden incluir registros de valores faltantes, sin sentido o tipos de datos inválidos. Es posible que los datos eliminados para cierto análisis sean de valor para un tipo de análisis diferente. Por consiguiente, se aconseja almacenar una copia exacta del dataset original antes de proceder con el filtrado (filtering). Para guardar la copia exacta en el espacio de almacenamiento necesario, esta se debe comprimir primero.

Tanto los datos internos como externos deben ser guardados una vez que se generan o entran en la empresa. Para los análisis por lotes, los datos se guardan en disco antes del análisis. En el caso de análisis en tiempo real, los datos primero son analizados, y luego guardados en disco.

Se puede agregar metadata por medio de la automatización de datos que provienen de fuentes de datos internas y externas, como se muestra en la Figura 2.2, con el fin de mejorar la clasificación y las consultas. Algunos ejemplos de metadata agregados son el tamaño y la estructura del dataset, información sobre la fuente, fecha y hora de creación o recopilación, e información específica de idioma. Es imprescindible que la metadata sea legible por máquina y sea transmitida a las etapas de análisis posteriores. Esto ayuda a mantener la procedencia de los datos a través del ciclo de vida de análisis de Big Data, lo cual contribuye a establecer y conservar la exactitud y calidad de los datos.

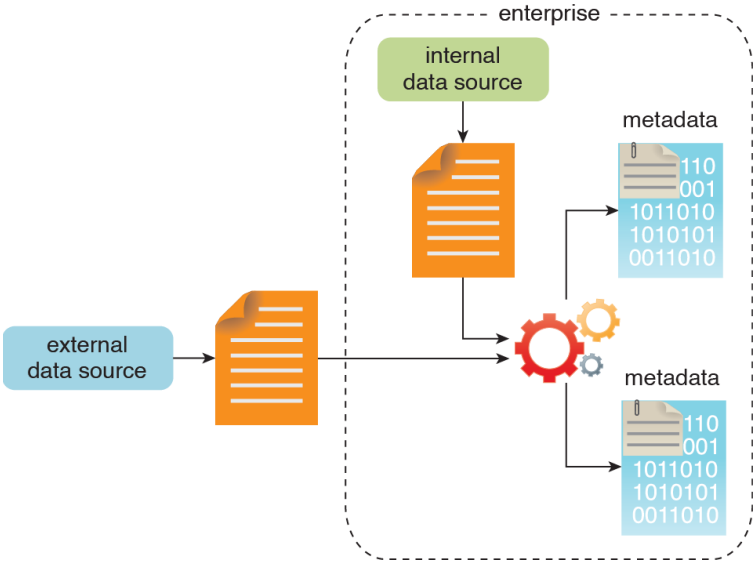


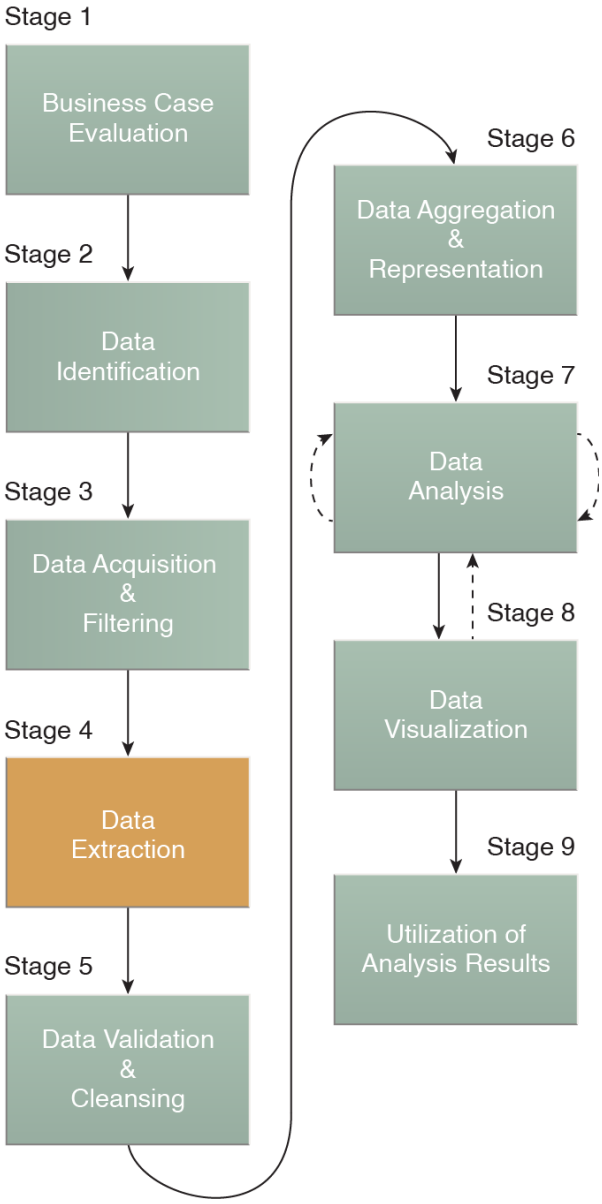
Figura 2.2 – Se agrega metadata a los datos provenientes de fuentes internas y externas.

Etapa 4: extracción de datos

Algunos datos que son identificados como entrada para el análisis pueden tener un formato incompatible con la solución de Big Data. Con los datos que provienen de fuentes externas, es más probable que se deban procesar distintos tipos de datos. La etapa del ciclo de vida que corresponde a la extracción de datos está orientada a extraer distintos datos y a convertirlos en un formato que la solución subyacente de Big Data pueda usar para el análisis de datos (Data Analysis).

El alcance de la extracción y la conversión requeridas dependerá de los tipos de analíticas y las capacidades de la solución de Big Data. Por ejemplo, tal vez no sea necesario extraer los campos requeridos de los datos de texto delimitados, si la solución subyacente de Big Data ya puede procesar directamente esos archivos, como en el caso de los archivos de registro (log files) de servidor web.

De manera similar, no será necesario extraer texto para la analítica de texto (text analytics) si la solución subyacente de Big Data ya puede leer el documento directamente en su formato original.



La Figura 2.3 ilustra la extracción de comentarios y la ID de un usuario integradas en un documento XML sin necesidad de una conversión adicional.

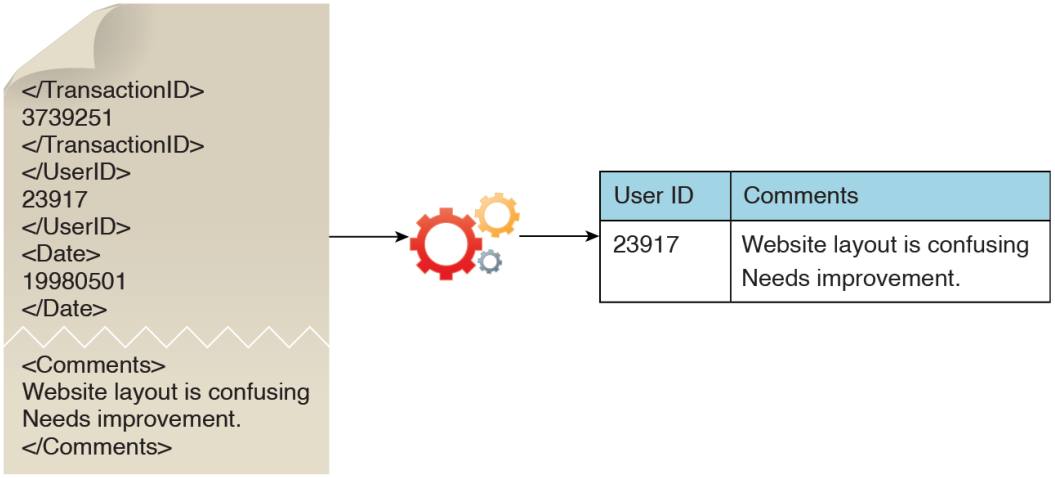


Figura 2.3 – Los comentarios e identificación del usuario son extraídos desde un documento XML.

La Figura 2.4 muestra la extracción de las coordenadas de latitud y longitud de un usuario a partir de un campo de JSON.

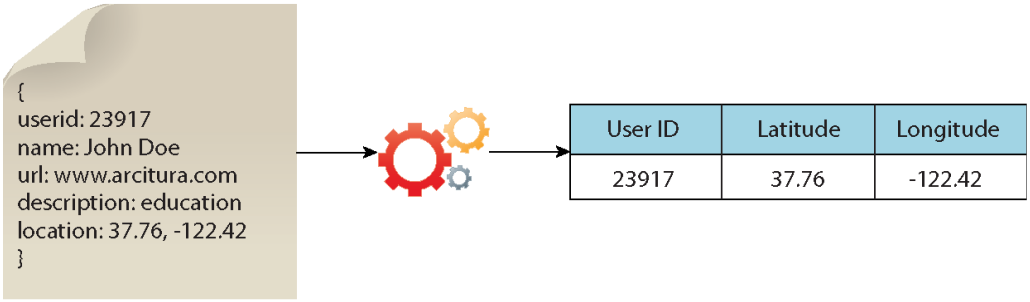


Figura 2.4 – La identificación y las coordenadas de un usuario son extraídas de un campo de JSON.

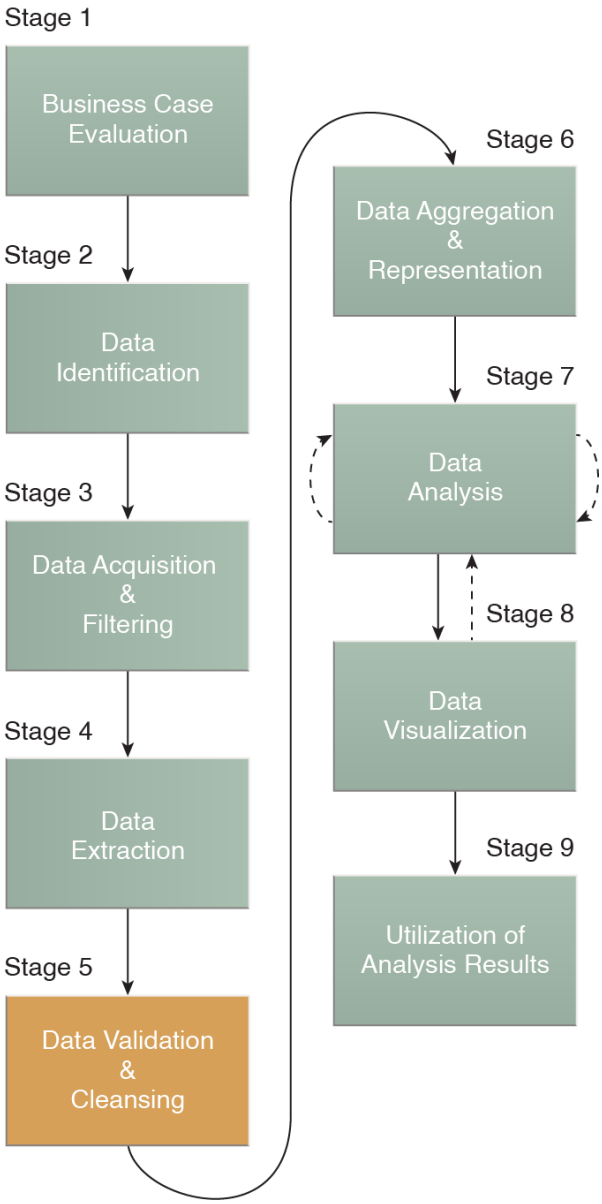
Se debe realizar una conversión adicional con el fin de separar los datos en dos campos distintos de acuerdo a lo requerido por la solución de Big Data.

Etapa 5: validación y limpieza (Cleansing) de datos

Los datos inválidos pueden producir sesgos y errores en los resultados de los análisis. A diferencia de los datos tradicionales de la empresa, en los cuales la estructura de los datos está predefinida y prevalidada, los datos de entrada del análisis de Big Data pueden no estar estructurados y no haber sido validados. Además, su complejidad puede hacer difícil determinar un conjunto de limitaciones adecuadas de validación.

La etapa de validación y limpieza (Cleansing) de datos está orientada a establecer normas de validación, a menudo complejas, y a eliminar cualquier dato inválido conocido.

Las soluciones de Big Data en ocasiones reciben datos redundantes en los diferentes datasets. Esta redundancia puede ser aprovechada para explorar datasets interconectados, con el fin de reunir parámetros de validación y completar datos válidos faltantes.



Por ejemplo, como se ilustra en la Figura 2.5:

1. El primer valor en el dataset B es validado frente a su valor correspondiente en el dataset A.
2. El segundo valor en el dataset B no es validado frente a su valor correspondiente en el dataset A.
3. Si falta un valor, este es introducido a partir del dataset A.

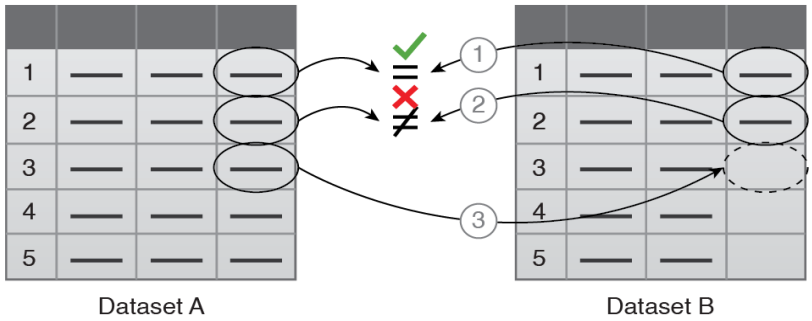


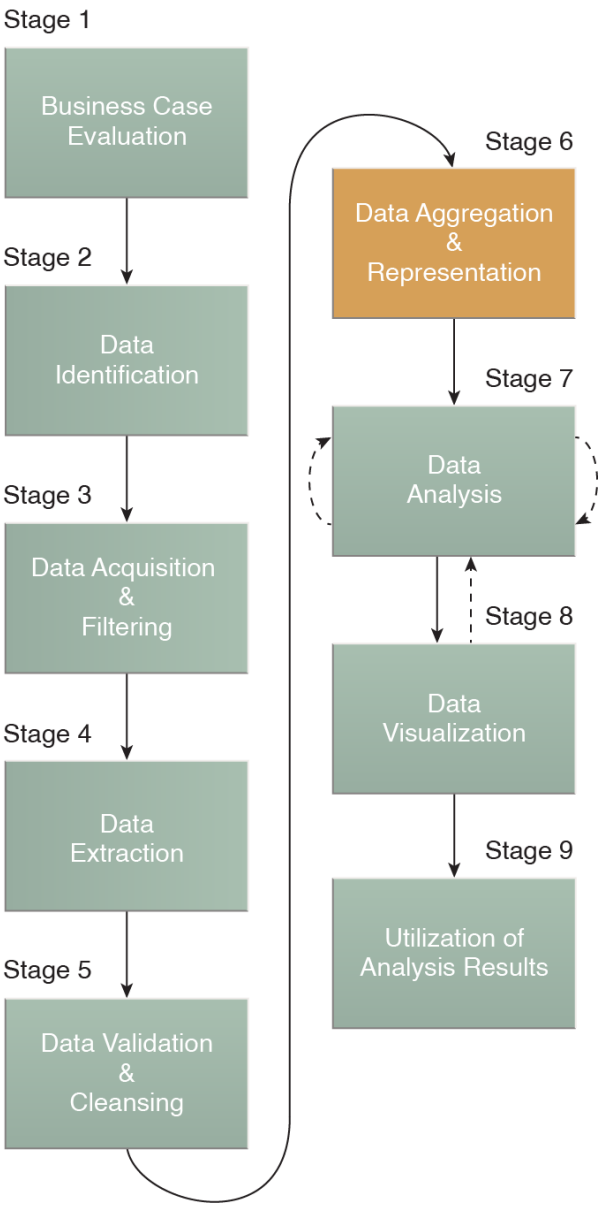
Figura 2.5 – La validación de datos puede ser usada para examinar datasets interconectados con el fin de completar datos válidos faltantes.

Para el análisis por lotes, se puede realizar la validación y limpieza (Cleansing) de datos por medio de una operación ETL offline. Para la analítica en tiempo real, se requiere un sistema basado en memoria más complejo para validar y limpiar los datos en la fuente. La procedencia puede desempeñar un papel fundamental a la hora de determinar la exactitud y calidad de los datos. Los datos que parecen ser inválidos aún pueden ser valiosos porque poseen patrones y tendencias ocultas, como se muestra en la Figura 2.6.

Figura 2.6 – La presencia de datos inválidos genera picos. Aunque los datos parecen anormales, pueden indicar un nuevo patrón.

Etapa 6: agregación y representación de datos

Los datos pueden estar distribuidos entre múltiples datasets, lo que implica que los datasets deben ser combinados por medio de campos comunes; por ejemplo, por fecha o ID. En otros casos, los mismos campos de datos pueden aparecer en múltiples datasets; por ejemplo, una fecha de nacimiento. De cualquier forma, se requiere un método de conciliación de datos o se debe determinar el dataset que representa el valor correcto.



La etapa de agregación y representación de datos está orientada a la integración de múltiples datasets para llegar a una vista unificada.

Puede ser complicado completar esta etapa debido a diferencias en:

- Estructura de los datos: aunque el formato de los datos puede ser el mismo, el modelo de datos puede ser diferente.
- Semántica: un valor etiquetado de formas diferentes en dos datasets diferentes puede significar lo mismo; por ejemplo, "*surname*" y "*last name*" (apellido).

Para resolver estas diferencias es necesaria una lógica compleja que se ejecute automáticamente sin la intervención humana. Los grandes volúmenes procesados por las soluciones de Big Data pueden hacer que la agregación de datos sea una operación que consuma mucho tiempo y esfuerzo.

En esta etapa, se deben considerar los requisitos futuros del análisis de datos (Data Analysis) para fomentar la reutilización de los datos. Ya sea que se requiera agregar datos o no, es importante entender que los mismos datos pueden ser almacenados de formas diferentes. Es posible que una forma sea más apropiada para un tipo de análisis particular que otra. Por ejemplo, los datos almacenados como BLOB se usarían muy poco si el análisis debe acceder a campos individuales de datos.

Una estructura de datos unificada por una solución de Big Data puede actuar como un denominador común que puede ser usado para una variedad de técnicas y proyectos de análisis. Para esto, puede ser necesario establecer un depósito central y estándar; por ejemplo, una base de datos NoSQL, como se muestra en la Figura 2.7.

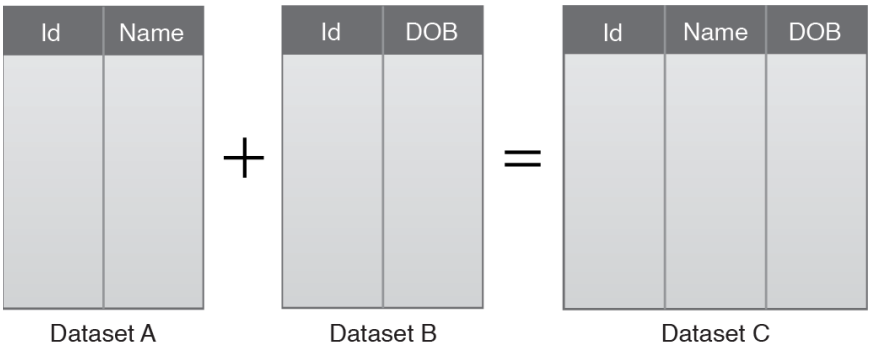


Figura 2.7 – Ejemplo sencillo de agregación de datos, en donde dos datasets son recopilados usando el campo Id.

La Figura 2.8 muestra los mismos datos almacenados en dos formatos diferentes. El dataset A contiene los datos deseados, pero hace parte de un BLOB al que no es fácil acceder para hacer consultas. El dataset B contiene los mismos datos organizados en un almacenamiento basado en columnas, lo cual facilita las consultas a cada campo.

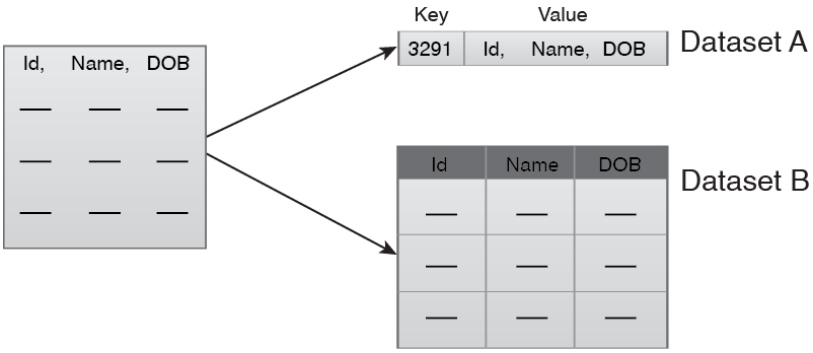
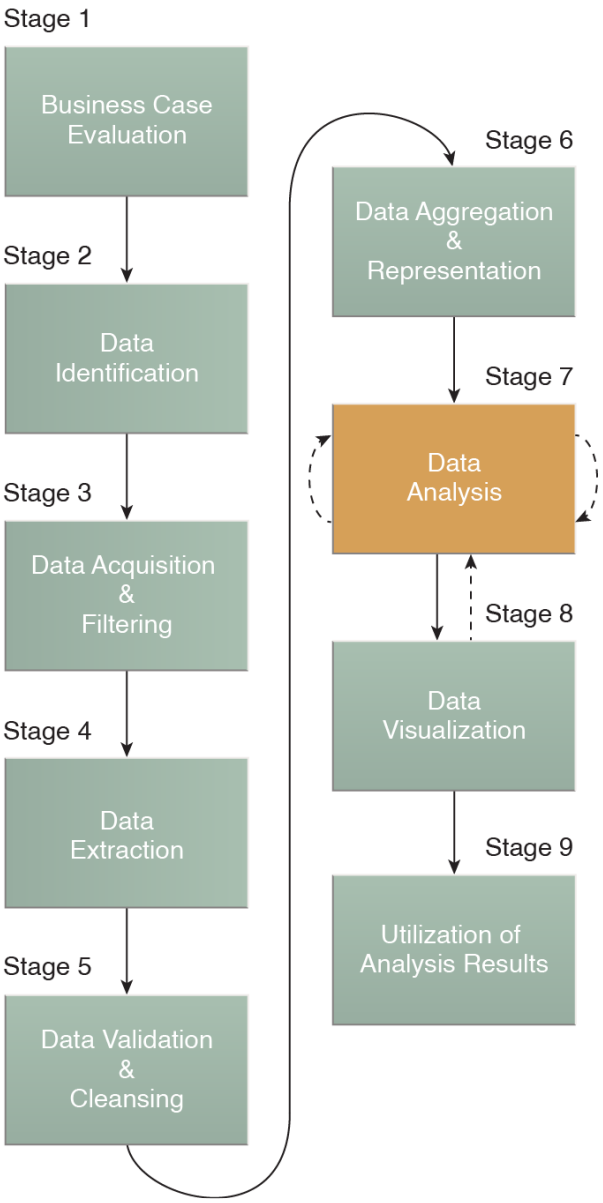


Figura 2.8 – Los datasets A y B se pueden combinar para crear una estructura de datos estandarizada con una solución de Big Data.

Etapa 7: análisis de datos (Data Analysis)

La etapa de análisis de datos (Data Analysis) está orientada a realizar la tarea real de análisis, lo cual normalmente involucra uno o más tipos de analíticas. Esta etapa puede ser de carácter iterativo, en particular si el análisis de datos (Data Analysis) es exploratorio, de manera que el análisis se repite hasta que se revela el patrón o correlación correspondiente. Se explicará brevemente el enfoque del análisis exploratorio, junto con el análisis confirmatorio de datos.

Según el tipo de analítica necesaria, esta etapa puede ser tan sencilla como solicitar que un dataset calcule datos agregados para que sean comparados, o por el contrario, puede ser tan exigente como combinar la minería de datos (Data Mining) con las técnicas complejas de análisis estadístico para encontrar patrones y anomalías o para generar un modelo estadístico o matemático que represente las relaciones entre las variables.



El enfoque adoptado para ejecutar esta etapa se puede clasificar como análisis confirmatorio o análisis exploratorio; este último está relacionado con la minería de datos (Data Mining), como se muestra en la Figura 2.9.

El Análisis Confirmatorio de Datos es un enfoque deductivo en el cual se propone previamente la causa del fenómeno investigado. La causa o suposición que se propone se conoce como hipótesis. Luego, los datos son analizados para comprobar o refutar la hipótesis y proporcionar respuestas definitivas a preguntas específicas. Normalmente se usan muestras de datos. Por lo general, los hallazgos o anomalías inesperados son ignorados debido a que se ha asumido una causa predeterminada.

El Análisis Exploratorio de Datos es un enfoque inductivo que está relacionado estrechamente con la minería de datos (Data Mining). No se generan hipótesis ni suposiciones predeterminadas. En lugar de eso, los datos son explorados mediante el análisis para lograr comprender la causa del fenómeno. Aunque quizá no aporte respuestas definitivas, este método ofrece una dirección general que puede ayudar a encontrar patrones o anomalías. Normalmente se usan grandes cantidades de datos y análisis visual.

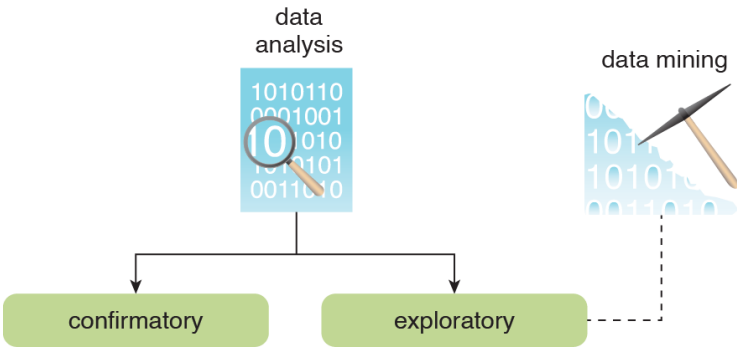


Figura 2.9 – El análisis de datos (Data Analysis) se puede ejecutar como análisis confirmatorio o exploratorio.

Etapa 8: visualización de datos

La capacidad de analizar cantidades masivas de datos y hallar información útil puede tener poco valor si los únicos que pueden interpretar los resultados son los analistas.

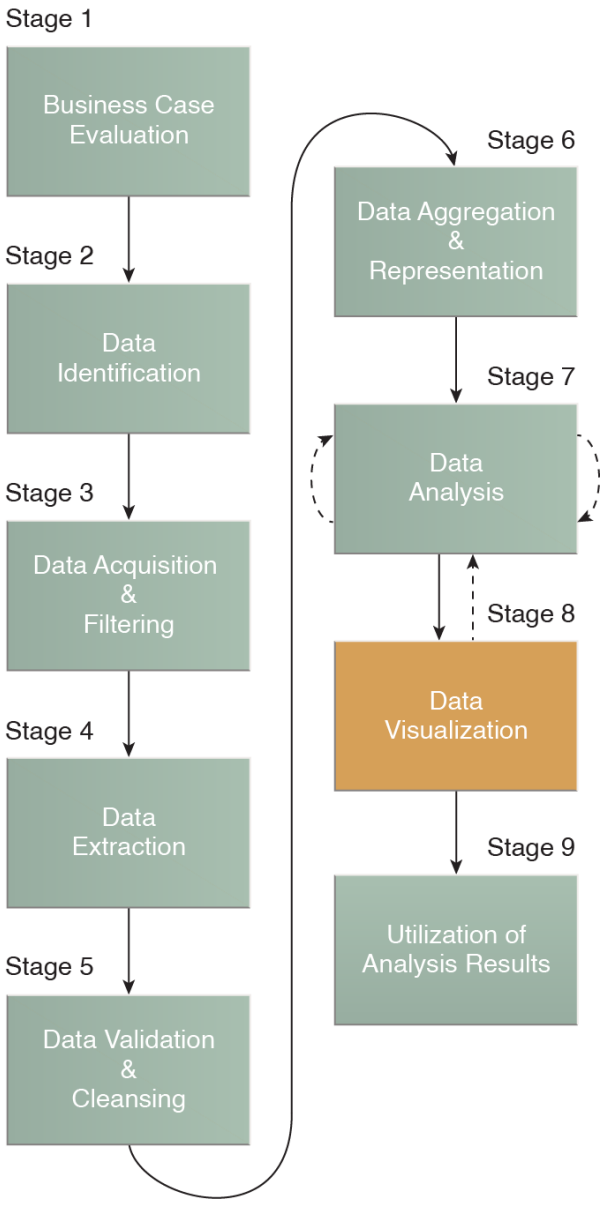
La etapa de visualización de datos está orientada a utilizar técnicas y herramientas de visualización de datos para comunicar gráficamente los resultados del análisis, de forma que los usuarios del negocio puedan interpretarlos efectivamente.

Los usuarios del negocio deben entender los resultados con el fin de obtener el valor del análisis y, como consecuencia, poder ofrecer retroalimentación, tal y como lo indican las líneas discontinuas que van de la etapa 8 a la etapa 7.

Los resultados de la etapa de visualización de datos les proporcionan a los usuarios la capacidad de realizar análisis visuales, facilitando respuestas a preguntas que los usuarios ni siquiera han planteado aún. El análisis visual se abordará más adelante en este cuaderno de trabajo.

Los mismos resultados pueden ser presentados de diversas formas, lo cual puede afectar su interpretación. Por esta razón es importante utilizar la técnica de visualización más adecuada, manteniendo el dominio empresarial en contexto.

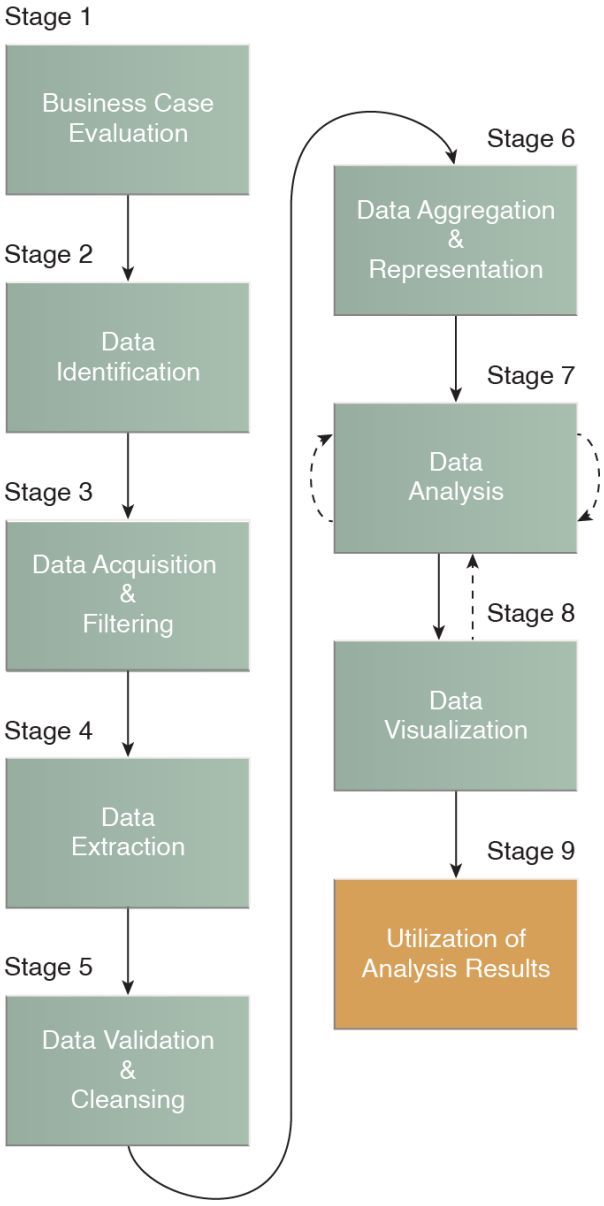
Asimismo, se debe tener en cuenta que es fundamental proporcionar un método que permita desglosar hasta estadísticas relativamente sencillas, con el objetivo de que los usuarios comprendan cómo estas fueron generadas.



Etapa 9: uso de los resultados del análisis

Luego de que los resultados del análisis sean puestos a disposición de los usuarios de la empresa para respaldar la toma de decisiones empresariales —por ejemplo, por medio de tableros de control (Dashboards)—, se pueden presentar más oportunidades para utilizar los resultados del análisis. La etapa de uso de los resultados del análisis está orientada a determinar cómo y cuándo se pueden aprovechar los datos procesados de análisis.

Dependiendo de la naturaleza de los problemas abordados por el análisis, es posible que los resultados de los análisis generen "modelos" que encapsulen nueva información sobre la naturaleza de los patrones y relaciones que existen en los datos que fueron analizados. Un modelo puede parecerse a una ecuación matemática o a un conjunto de normas. Los modelos pueden ser usados para mejorar la lógica del proceso empresarial y la lógica del sistema de aplicaciones, y pueden conformar la base de un nuevo sistema o software.



Algunas áreas exploradas durante esta etapa son las siguientes:

Datos de entrada para sistemas empresariales: los resultados del análisis de datos (Data Analysis) pueden ser ingresados automática o manualmente en los sistemas empresariales para mejorar y optimizar su comportamiento y desempeño. Por ejemplo, una tienda online puede ser alimentada con los resultados procesados de los análisis relacionados con los clientes, lo cual puede impactar la forma en que la tienda genera recomendaciones de productos. Los nuevos modelos pueden ser usados para mejorar la lógica de programación en sistemas empresariales existentes o pueden constituir la base de nuevos sistemas.

Optimización de procesos empresariales: los patrones, correlaciones y anomalías identificados y hallados durante el análisis de datos (Data Analysis) son utilizados para perfeccionar los procesos empresariales. Un ejemplo es la consolidación de rutas de transporte como parte de un proceso de la cadena de suministros. Los modelos también pueden brindar oportunidades de mejorar la lógica del proceso empresarial.

Alertas: los resultados del análisis de datos (Data Analysis) se pueden usar como entradas para alertas existentes o pueden formar la base de nuevas alertas. Por ejemplo, se pueden crear alertas para informar a los usuarios por medio de correo electrónico o mensajes de texto sobre un evento que requiere que los usuarios tomen una acción correctiva.

Lecturas opcionales

El libro *Analítica de Big Data*, incluido en este Módulo, presenta el ciclo de vida de análisis de Big Data con más detalle en el *Capítulo 10*: *Big Data Práctico*

Ejercicio 2.1: organice las etapas del ciclo de vida

Organice las siguientes etapas del ciclo de vida de Big Data en el orden correcto:

Visualización de datos	1. _____
Extracción de datos	2. _____
Identificación de datos	3. _____
Validación y limpieza (Cleansing) de datos	4. _____
Adquisición y filtrado (filtering) de datos	5. _____
Uso de los resultados del análisis	6. _____
Agregación y representación de datos	7. _____
Evaluación del caso empresarial	8. _____
Análisis de datos (Data Analysis)	9. _____

Las respuestas al ejercicio se encuentran al final de este cuadernillo.

Notas

Notas / Bocetos

Parte II: conceptos de análisis de Big Data

Cada uno de estos conceptos será analizado en esta sección:

- Análisis estadístico
- Análisis visual
- Aprendizaje automático (Machine Learning)
- Análisis semántico

Estas técnicas de análisis de datos que se pueden aplicar en la etapa de análisis de datos (Data Analysis) del ciclo de vida están agrupadas en las siguientes cuatro categorías principales:

Análisis estadístico

- Tests A/B
- Correlación
- Regresión

Análisis visual

- Mapas de calor
- Análisis de series temporales
- Análisis de redes
- Análisis de datos espaciales

Aprendizaje automático (Machine Learning)

- Clasificación
- Agrupamiento (Clustering)
- Detección de datos atípicos (outliers)
- Filtrado (filtering)

Análisis semántico

- Procesamiento de lenguaje natural
- Analítica de texto (text analytics)
- Análisis de sentimientos (Sentiment Analysis)

NOTA
Con excepción de las técnicas de aprendizaje automático (Machine Learning), cada una de las siguientes técnicas de análisis es complementada con un ejemplo sencillo basado en la situación de las ventas de helado que se presentó en el Módulo 1. Las técnicas de aprendizaje automático (Machine Learning) presentan una nueva situación basada en los requisitos de análisis de un banco.

Análisis estadístico

El análisis estadístico utiliza métodos estadísticos basados en fórmulas matemáticas como medio para analizar datos. Este tipo de análisis normalmente es utilizado para describir datasets por medio de resúmenes; por ejemplo, al indicar la media, mediana o la moda relacionadas con el dataset. También se puede usar para deducir patrones y relaciones dentro del dataset, como regresión y correlación.

En esta sección se describen los siguientes tipos de análisis estadístico:

- Tests A/B
- Correlación
- Regresión

Tests A/B

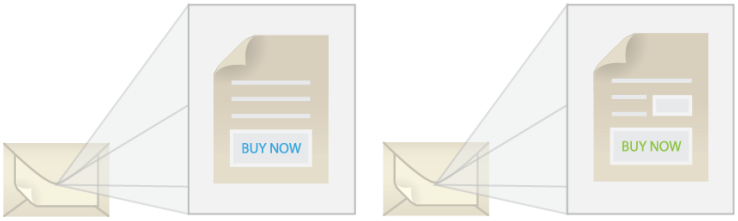
Los tests A/B, también conocidos como split o bucket testing, comparan dos versiones de un elemento para determinar qué versión es superior con base en métricas predefinidas. El elemento puede ser una variedad de cosas. Por ejemplo, puede ser contenido, como un sitio web; o una oferta para un producto o servicio, como ofertas en artículos electrónicos. La versión actual del elemento se conoce como versión de control, mientras que la versión modificada se conoce como tratamiento. Ambas versiones están sujetas simultáneamente a un experimento. Las observaciones son registradas para determinar qué versión es más exitosa.

Por ejemplo, para determinar el mejor diseño publicitario posible para el sitio web de la Empresa de helados A, se utilizan dos versiones publicitarias diferentes. La versión A corresponde a un anuncio que ya existe (el control), mientras que la versión B ha sido modificada ligeramente (el tratamiento) en comparación con el diseño de la versión A. Diferentes usuarios pueden ver simultáneamente ambas versiones:

- Versión A para el Grupo A
- Versión B para el Grupo B

El análisis de los resultados revela que la versión B del anuncio generó más ventas en comparación con la versión A. Si bien el test A/B se puede implementar en casi cualquier campo, es usado con mayor frecuencia en mercadeo. Generalmente, el objetivo es evaluar el comportamiento humano con la meta de aumentar las ventas, como en el ejemplo anterior.

En otras áreas, como en el ámbito científico, el objetivo puede ser simplemente observar qué versión funciona mejor para perfeccionar un proceso o producto. La Figura 2.10 ofrece un ejemplo de test A/B de dos versiones diferentes de correo electrónico enviadas simultáneamente.



E-Mail A E-Mail B

Figura 2.10 – Dos versiones diferentes de correo electrónico enviadas simultáneamente como parte de una campaña de mercadeo para que ver qué versión atrae más clientes potenciales.

Algunas preguntas de ejemplo pueden ser:

- *¿La nueva versión de un medicamento es mejor que la anterior?*
- *¿La nueva fórmula para un champú anticaspa será más efectiva que la anterior?*
- *¿La página de inicio del sitio web recientemente diseñada está generando más tráfico de usuarios?*

Correlación

La correlación es una técnica de análisis utilizada para determinar si dos variables están relacionadas entre sí. Si se encuentra que están relacionadas, el paso siguiente es determinar cuál es su relación. Por ejemplo, el valor de la Variable A aumenta a medida que el valor de la Variable B aumenta. Además, seguramente nos interesa encontrar qué tan relacionadas están las Variables A y B, lo cual significa que también queremos analizar en qué medida aumenta la Variable B en relación con la Variable A.

El uso de una correlación ayuda a entender un dataset y a encontrar las relaciones que pueden servir para explicar un fenómeno. De esta manera, la correlación se usa comúnmente en la minería de datos (Data Mining), en donde la identificación de relaciones entre variables de un dataset conduce al hallazgo de patrones y anomalías. Esto puede poner de manifiesto la naturaleza del dataset o la causa de un fenómeno.

Cuando se estima que dos variables están correlacionadas, se considera que se ajustan de acuerdo a relación lineal. Esto quiere decir que cuando una variable cambia, la otra variable también cambia de manera proporcional y constante.

Una correlación se expresa como un número decimal entre -1 y +1, lo cual se conoce como el coeficiente de correlación. El grado de relación pasa de fuerte a débil cuando cambia de -1 a 0 o de +1 a 0.

La Figura 2.11 muestra una correlación de +1, lo cual sugiere que hay una relación positiva fuerte entre las dos variables.

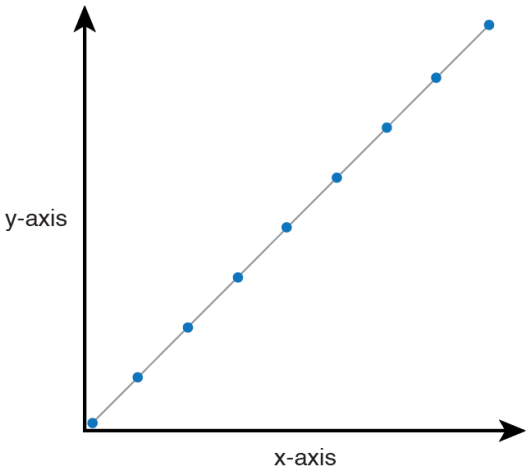


Figura 2.11 – Cuando una variable aumenta, la otra variable también aumenta, y viceversa.

La Figura 2.12 muestra una correlación de 0, lo cual sugiere que no hay ninguna relación entre las dos variables.

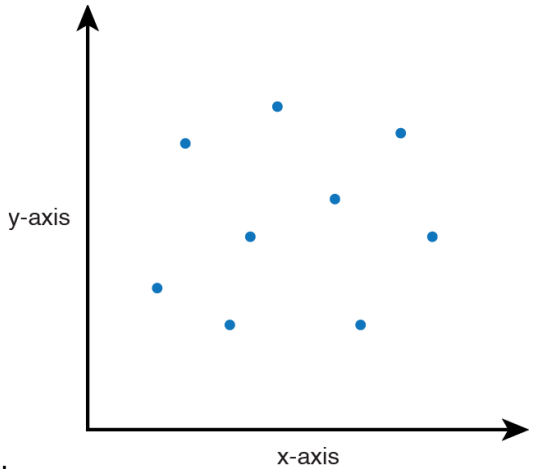


Figura 2.12 – Cuando una variable aumenta, la otra puede permanecer igual, aumentar o disminuir arbitrariamente.

En la Figura 2.13, una pendiente de -1 sugiere que hay una relación negativa fuerte entre las dos variables.

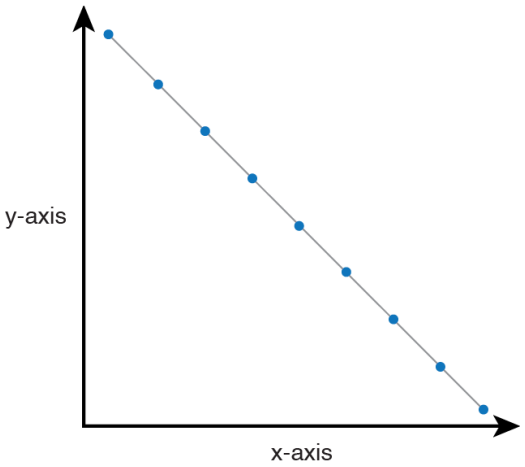


Figura 2.13 – Cuando una variable aumenta, la otra variable disminuye, y viceversa.

Por ejemplo, los gerentes piensan que las heladerías deben contar con un mayor inventario de helados durante los días calurosos, pero no saben qué tan grande debe ser la cantidad adicional. Para determinar si realmente existe una relación entre la temperatura y la venta de helados, los analistas primero aplican la correlación a la cantidad de helados vendidos y las lecturas registradas de la temperatura. Un valor de +0,75 sugiere que hay una relación fuerte entre las dos variables. Esta relación indica que, a medida que aumenta la temperatura, se venden más helados.

Otras preguntas de ejemplo relacionadas con la correlación pueden ser:

- *¿La distancia al mar afecta la temperatura de una ciudad?*
- *¿Los estudiantes que tienen un buen desempeño en primaria se desempeñan de igual forma en la secundaria?*
- *¿Hasta qué punto la obesidad se relaciona con el consumo excesivo de alimentos?*

Regresión

La técnica de análisis de regresión investiga cómo se relaciona una variable dependiente con una variable independiente dentro de un dataset. Como escenario de muestreo, la regresión podría ayudar a determinar el tipo de relación que existe entre la temperatura (es decir, la variable independiente) y el desempeño de los cultivos (es decir, la variable dependiente).

Aplicar esta técnica también sirve para determinar cómo cambia el valor de una variable dependiente con respecto a los cambios en el valor de la variable independiente. Por ejemplo, cuando la variable independiente aumenta, ¿la variable dependiente también aumenta? Si así es, ¿el aumento ocurre en una proporción lineal o no lineal?

Por ejemplo, con el fin de determinar la cantidad adicional de helado necesaria, los analistas aplican la regresión alimentando los valores de las lecturas de temperatura. Estos valores están basados en el pronóstico del clima como una variable independiente y, en la cantidad de helados como variable dependiente. Los analistas descubren que se necesita un 15% adicional de inventario por cada cinco grados de aumento en la temperatura. Se puede probar más de una variable independiente al mismo tiempo.

Sin embargo, en tales casos, solo puede cambiar una variable independiente; las demás permanecen constantes. La regresión puede ayudar a entender en qué consiste un fenómeno y porqué ocurrió. También se puede usar para predecir los valores de la variable dependiente cuando estos son desconocidos.

La regresión lineal representa un ritmo constante de cambio, como se muestra en la Figura 2.14.

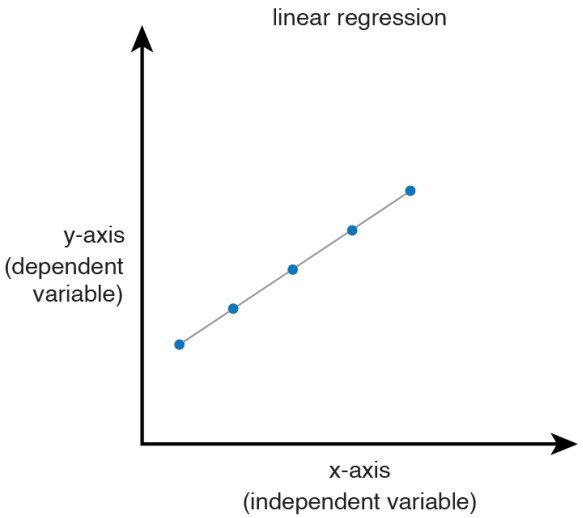


Figura 2.14 – Regresión lineal

La regresión no lineal representa un ritmo variable de cambio, como se muestra en la Figura 2.15.

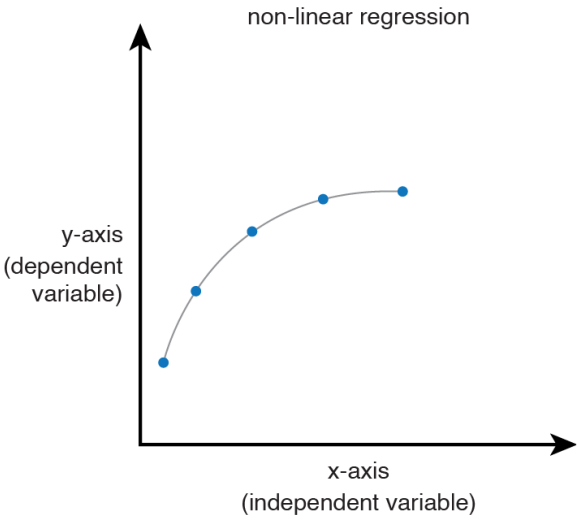


Figura 2.15 – Regresión no lineal

Algunas preguntas de ejemplo pueden ser:

- *¿Cuál será la temperatura de una ciudad que se encuentra a 250 millas de distancia del mar?*
- *¿Cuáles serán las calificaciones de un estudiante de secundaria, tomando como base sus calificaciones en primaria?*
- *¿Cuáles son las posibilidades de que una persona llegue a ser obesa con base en la cantidad de alimentos que consume?*

Regresión y correlación

La regresión y la correlación tienen varias diferencias importantes. La correlación no implica causalidad. El cambio en el valor de una variable puede no ser responsable del cambio en el valor de una segunda variable, aunque las dos sí pueden cambiar a un mismo ritmo. La correlación asume que ambas variables son independientes.

En cambio, la regresión incluye variables dependientes e independientes que ya fueron identificadas, e implica que hay cierto grado de causalidad (que puede ser directo o indirecto) entre las variables dependientes e independientes.

En Big Data, la correlación puede ser aplicada en primer lugar para determinar si existe una relación. Luego, se puede aplicar la regresión para investigar la relación y predecir los valores de la variable dependiente, con base en los valores conocidos de la variable independiente.

Notas

Notas / Bocetos

Análisis visual

El análisis visual es una forma de análisis de datos (Data Analysis) que implica la representación gráfica de datos para facilitar o mejorar la percepción visual. Con base en la premisa de que los seres humanos pueden entender y extraer conclusiones más rápidamente a partir de gráficas que a partir de textos, el análisis visual actúa como una herramienta de detección en el campo de Big Data.

El objetivo es utilizar representaciones gráficas para entender mejor los datos analizados. Concretamente, ayuda a identificar y señalar patrones, correlaciones y anomalías ocultos. El análisis visual también está directamente relacionado con el Análisis Exploratorio de Datos, puesto que fomenta el planteamiento de preguntas desde diferentes perspectivas.

En esta sección se describen los siguientes tipos de análisis visual:

- Mapas de calor
- Análisis de series temporales
- Análisis de redes
- Análisis de datos espaciales

Mapas de calor

Los mapas de calor son una técnica de análisis visual efectiva para expresar patrones, composiciones de datos por medio de la relación de una parte con el todo, y distribuciones geográficas de datos. También facilitan la identificación de áreas de interés y el hallazgo de valores extremos (superiores/inferiores) dentro de un dataset.

Por ejemplo, con el fin de identificar la mejor y la peor elección de región para la venta de helados, se grafican los datos de las ventas de helados utilizando un mapa de calor. Se utiliza el color verde para resaltar las zonas con los mejores resultados, y se utiliza el color rojo para resaltar las zonas con los peores resultados.

Como tal, el mapa de calor es una representación visual de valores de datos codificada por colores. Se asigna un color a cada valor de acuerdo con el rango en el que se encuentre. Por ejemplo, un mapa de calor puede asignar los valores de 0 a 3 al color rojo, de 4 a 6 al color amarillo y 7 a 10 al color verde.

Un mapa de calor puede tener forma de una tabla o de un mapa. Una tabla representa una matriz de valores en la cual cada celda está codificada con un color de acuerdo con el valor, como se muestra en la Figura 2.16. También representa valores jerárquicos al usar rectángulos anidados codificados por color.

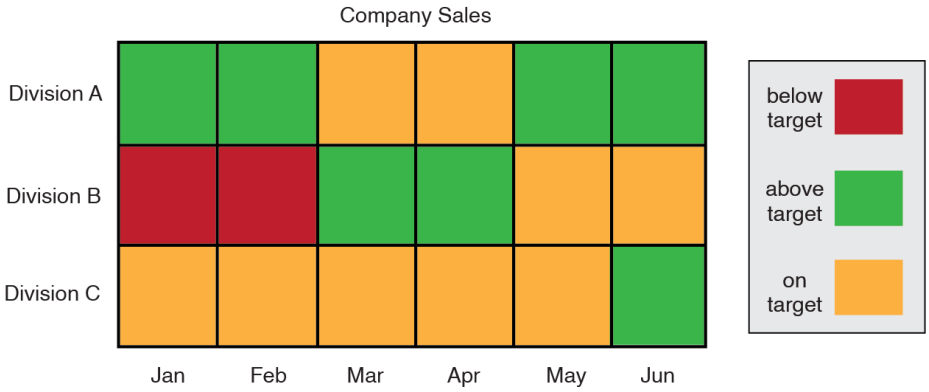


Figura 2.16 – La tabla del mapa de calor representa las ventas de tres departamentos de una empresa en un período de seis meses.

En la Figura 2.17 el mapa representa una medida geográfica en la que cada región está codificada con un color de acuerdo a cierta temática. En vez de colorear la región completa, el mapa tiene una capa compuesta por grupos de puntos coloreados relacionados con varias regiones o formas coloreadas que representan varias regiones.

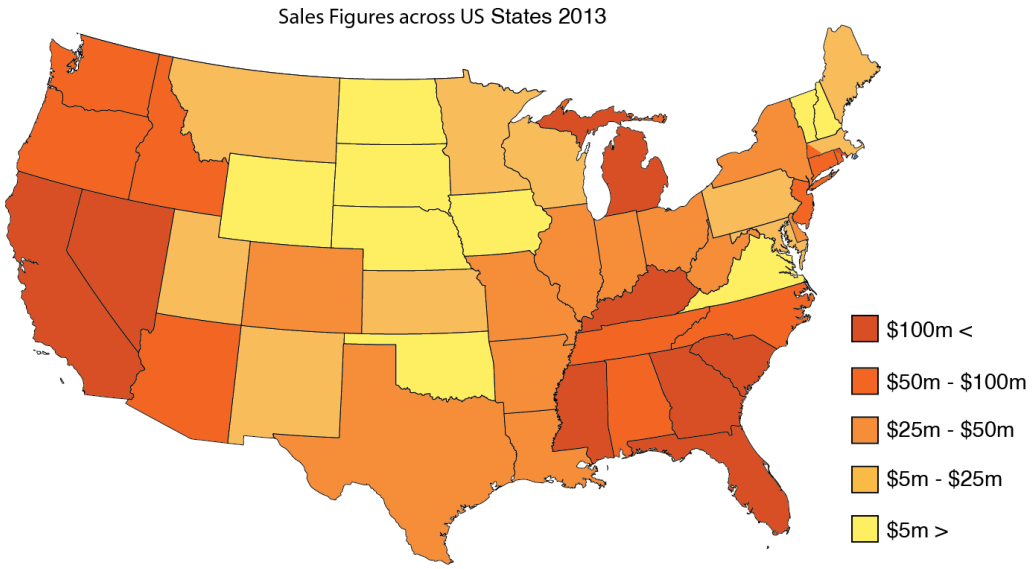


Figura 2.17 – Mapa de calor de las cifras de ventas en EE. UU. en el año 2013

Algunas preguntas de ejemplo pueden ser:

- *¿Cómo identificar visualmente un patrón relacionado con emisiones de carbón en un número importante de ciudades alrededor del mundo?*
- *¿Cómo puedo ver si existe algún patrón de los diferentes tipos de cáncer con respecto a distintos grupos étnicos?*
- *¿Cómo puedo analizar los jugadores de fútbol de acuerdo con sus fortalezas y debilidades?*

Análisis de series temporales

El análisis de series temporales es el análisis de datos (Data Analysis) que son registrados en intervalos periódicos de tiempo. Este tipo de análisis utiliza series temporales, es decir, una serie de valores ordenados por tiempo y registrados en intervalos regulares de tiempo. Un ejemplo puede ser una serie temporal representada por las cifras de ventas que son registradas al final de cada mes.

El análisis de series temporales permite encontrar patrones en los datos que dependen del tiempo. Una vez que es identificado, el modelo puede ser extrapolado para predicciones futuras. Por ejemplo, para identificar patrones de ventas estacionales, las cifras mensuales de ventas de helado son graficadas como series temporales que además ayudan a predecir las cifras de ventas de la siguiente temporada.

Por lo general, el análisis de series temporales es utilizado para realizar pronósticos al identificar tendencias a largo plazo, patrones estacionales periódicos y variaciones irregulares a corto plazo en el dataset. A diferencia de otros tipos de análisis, el análisis de series temporales siempre incluye el tiempo como una variable de comparación, y los datos recopilados siempre dependen del tiempo.

Por lo general, una serie temporal se expresa utilizando un diagrama de líneas, con el tiempo ubicado en el eje X y los valores de datos registrados en el eje Y.

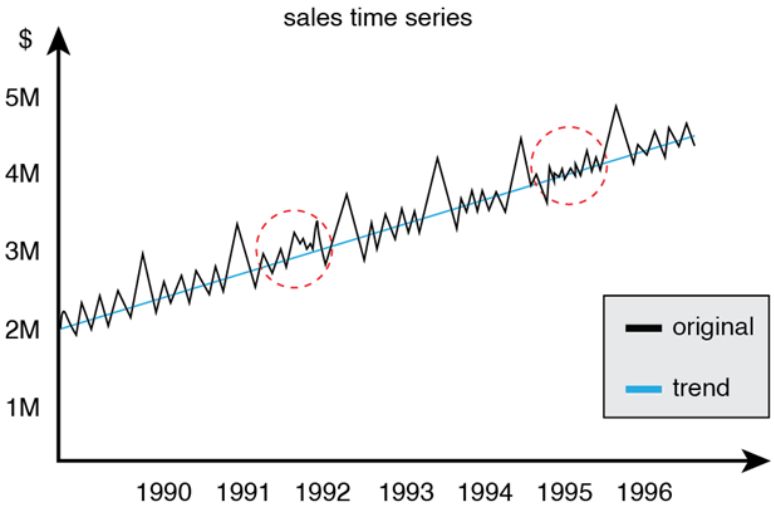


Figura 2.18 – Un diagrama de líneas representa una serie temporal de ventas desde 1990 hasta 1996.

La serie temporal que se presenta en la Figura 2.18 abarca siete años. Los picos espaciados uniformemente hacia el final de cada año muestran patrones estacionales periódicos, por ejemplo en las ventas para Navidad. Los círculos punteados de color rojo representan variaciones irregulares a corto plazo. La línea de color azul muestra una tendencia ascendente, lo cual indica un aumento en las ventas.

Algunas preguntas de ejemplo pueden ser:

- *¿Cuánta cosecha debería esperar el agricultor con base en los datos históricos de cosechas?*
- *¿Cuál es el aumento de la población que se espera en los próximos cinco años?*
- *¿La disminución actual de ventas es un hecho puntual u ocurre periódicamente?*

Análisis de redes

En el contexto del análisis visual, una red es una serie de entidades interconectadas. Una entidad puede ser una persona, grupo o un objeto de dominio empresarial como un producto. Las entidades pueden estar conectadas unas con otras de manera directa o indirecta. Algunas conexiones pueden ir en una sola dirección, de manera que no es posible hacer un recorrido en sentido contrario.

El análisis de redes es una técnica que se enfoca en analizar las relaciones entre entidades en la red, lo cual implica graficar las entidades como nodos y las conexiones como bordes entre los nodos. Existen variaciones especializadas de análisis de redes como:

- Optimización de rutas
- Análisis de redes sociales
- Pronóstico de propagación; por ejemplo, la propagación de una enfermedad contagiosa

El siguiente es un ejemplo sencillo de la aplicación del análisis de redes para la optimización de rutas, basado en las ventas de helado:

Algunos gerentes de heladerías están quejándose por el tiempo que los camiones de entregas tardan en llegar desde la bodega central y las tiendas en áreas remotas. Durante los días más calientes, los helados que entrega la bodega central a las tiendas remotas se derriten y no se pueden vender. El análisis de redes es utilizado para definir las rutas más cortas entre la bodega central y las tiendas remotas, con el fin de minimizar la duración de las entregas.

Considere la rede social de la Figura 2.19 como un ejemplo sencillo de análisis de redes sociales:

- John tiene muchos amigos, mientras que Alice solo tiene un amigo.
- Los resultados del análisis de redes sociales muestran que es más probable que Alice se haga amiga de John y Katie, pues tienen un amigo en común que se llama Oliver.

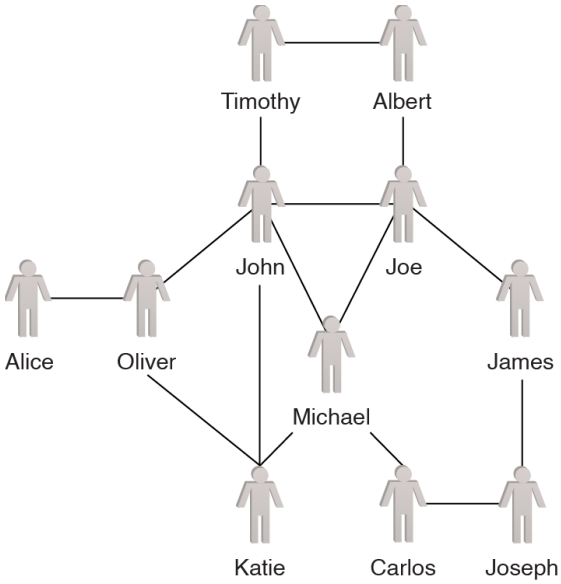


Figura 2.19 – Red social

Algunas preguntas de ejemplo pueden ser:

- *¿Cómo puedo identificar influenciadores dentro de un grupo grande de usuarios?*
- *¿Dos personas están relacionadas entre sí por una larga cadena de ancestros?*
- *¿Cómo puedo identificar patrones de interacción entre una gran cantidad de interacciones entre proteínas?*

Análisis de datos espaciales

El análisis de datos espaciales está orientado a analizar datos basados en la ubicación, con el fin de encontrar diferentes relaciones y patrones geográficos entre las entidades. Los datos espaciales o geoespaciales son usados normalmente para identificar la ubicación geográfica de entidades individuales.

Los datos espaciales son manipulados a través de un sistema de información geográfica (GIS) que traza datos espaciales en un mapa usando coordenadas de longitud y latitud. Gracias al aumento constante en la disponibilidad de datos basados en ubicación —como los datos de sensores y social media—, los datos espaciales pueden ser analizados para obtener información sobre la ubicación.

Por ejemplo, como parte de una expansión corporativa, se planea abrir más heladerías. Existe una norma que establece que dos tiendas no pueden estar a una distancia de 5 kilómetros la una de la otra, para evitar que las tiendas compitan entre sí. El análisis de datos espaciales es utilizado para graficar la ubicación de tiendas ya existentes, y para identificar ubicaciones óptimas para colocar tiendas nuevas al menos a 5 kilómetros de distancia de las tiendas ya existentes.

Las aplicaciones del análisis de datos espaciales incluyen optimización operativa y logística, ciencias ambientales y planeación de infraestructura. Los datos de entrada para el análisis de

datos espaciales pueden contener ubicaciones exactas, como longitud y latitud, o contener la información necesaria para calcular las ubicaciones, como códigos zip o direcciones IP. El análisis de datos espaciales proporciona características de análisis más sofisticadas que los mapas de calor.

Adicionalmente, el análisis de datos espaciales puede ser utilizado para determinar la cantidad de entidades que están dentro de un cierto radio de otra entidad. Por ejemplo, un supermercado utiliza el análisis de datos espaciales para realizar mercadeo dirigido, como se muestra en la Figura 2.20. Las ubicaciones son extraídas de los mensajes de social media de los usuarios. Se envían ofertas personalizadas en tiempo real con base en la proximidad del usuario.

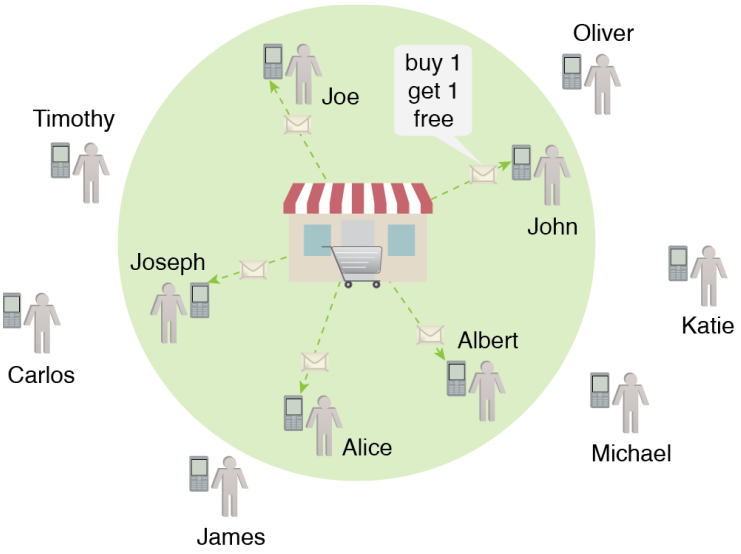


Figura 2.20 – El análisis de datos espaciales se puede usar para el mercadeo dirigido.

Algunas preguntas de ejemplo pueden ser:

- *¿Cuántas casas se verán afectadas por un proyecto de ampliación de vías?*
- *¿Qué tan lejos deben desplazarse los clientes para llegar a un supermercado?*
- *¿En dónde se encuentran las concentraciones altas y bajas de un mineral específico con base en las lecturas tomadas de varias ubicaciones de muestra en un área?*

Notas

Notas / Bocetos

Aprendizaje automático (Machine Learning)

Los seres humanos son buenos ubicando patrones y relaciones entre datos. Desafortunadamente no podemos procesar grandes cantidades de datos muy rápido. En comparación, las computadoras son expertas procesando grandes cantidades de datos rápidamente, pero solo si saben cómo hacerlo.

Si combinamos el conocimiento humano y el rápido procesamiento de las computadoras, entonces estas podrán procesar grandes cantidades de datos sin necesidad de la intervención humana. Ese es el concepto básico del aprendizaje automático (Machine Learning). El aprendizaje automático (Machine Learning) y su relación con la minería de datos (Data Mining) se presentaron en el Módulo 1. En esta sección se analizan estos temas con más detalle, abordando los siguientes tipos de técnicas de aprendizaje automático (Machine Learning):

- Clasificación
- Agrupamiento (Clustering)
- Detección de datos atípicos (outliers)
- Filtrado (filtering)

Antes de proceder con la explicación de estas técnicas, debemos establecer las dos leyes fundamentales que hacen parte del aprendizaje automático (Machine Learning):

- Ley de los Grandes Números, que se aplica a Big Data
- Ley de la Utilidad Marginal Decreciente, que no se aplica a Big Data

Ley de los Grandes Números

La Ley de los Grandes Números establece que la confianza con la cual se pueden realizar predicciones aumenta con el tamaño de los datos analizados. En otras palabras, la exactitud y aplicabilidad de los patrones y relaciones que se encuentran en un dataset grande serán mayores que los de un dataset pequeño. Esto quiere decir que cuanto mayor sea la cantidad de datos disponibles para análisis, más correctas serán las decisiones que podamos tomar.

Ley de la Utilidad Marginal Decreciente

En el contexto del análisis de datos (Data Analysis) tradicional, la ley de Utilidad Marginal Decreciente establece que, comenzando con un tamaño de muestra razonablemente grande, el valor obtenido a partir del análisis de datos adicionales disminuye a medida que se añaden más datos a la muestra original. Este es un principio del análisis de datos (Data Analysis) tradicional que afirma que los datos almacenados en un dataset de tamaño razonable proporcionan el valor máximo.

La ley de la Utilidad Marginal Decreciente no se aplica a Big Data. Cuanto más grande sea el volumen y la variedad de los datos que las soluciones de Big Data pueden procesar, mayor será la posibilidad de encontrar nuevos patrones y anomalías en cada lote de datos

adicionales. Por tanto, el valor de cada lote adicional no disminuye; por el contrario, se ofrece un valor mayor.

Clasificación

La clasificación es una técnica de aprendizaje supervisada en la cual los datos se clasifican en las categorías relevantes previamente aprendidas. Consta de dos pasos:

1. El sistema es alimentado con datos que ya están categorizados o etiquetados, de manera que puede entender las diferentes categorías.
2. El sistema es alimentado con datos desconocidos, pero similares, para clasificarlos de acuerdo con el entendimiento que ha desarrollado.

Una aplicación común de esta técnica es la del filtrado de correo no deseado (spam). Tenga en cuenta que la clasificación se puede ejecutar para dos o más categorías. En un proceso simplificado de clasificación, la computadora es alimentada con datos etiquetados durante el entrenamiento, lo cual le permite entender la clasificación, como se muestra en la Figura 2.21. Posteriormente, la computadora recibe datos sin etiquetar que ella misma clasifica.

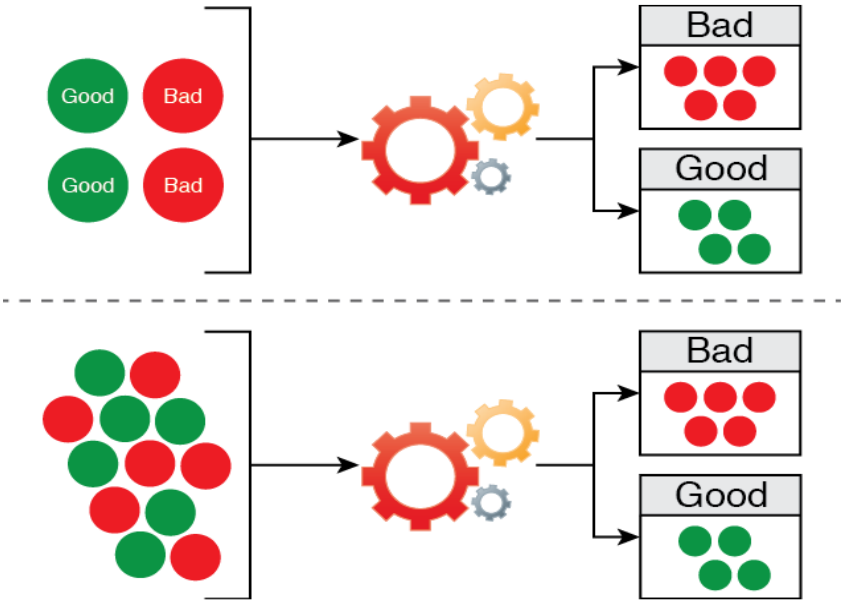


Figura 2.21 – El aprendizaje automático (Machine Learning) puede ser usado para clasificar datasets automáticamente.

Por ejemplo, un banco quiere averiguar cuáles de sus clientes probablemente están en mora en el pago de sus deudas. Con base en datos anteriores, se compila un dataset de entrenamiento que contiene ejemplos etiquetados de clientes que anteriormente han estado o que están en mora con sus pagos. Estos datos de entrenamiento alimentan un algoritmo de clasificación que es usado para entender quiénes son los clientes “buenos” y “malos”.

Finalmente, se ingresan datos nuevos sin etiquetar de los clientes, con el fin de averiguar si un determinado cliente pertenece a la categoría de morosidad.

Algunas preguntas de ejemplo pueden ser:

- *¿Se debería aceptar o rechazar una solicitud de tarjeta de crédito con base en otras solicitudes aceptadas o rechazadas?*
- *Con base en los ejemplos conocidos de frutas y vegetales, ¿se puede determinar si un tomate es una fruta o una verdura?*
- *Con base en el registro de una huella digital anterior, ¿se puede decir si una huella digital pertenece a un sospechoso?*

Agrupamiento (Clustering)

El agrupamiento (Clustering) es una técnica de aprendizaje no supervisada en la cual los datos son divididos en diferentes grupos, de modo que los datos en cada grupo tienen propiedades similares. No requiere un aprendizaje previo de las categorías; en su lugar, las categorías son generadas implícitamente de acuerdo con las agrupaciones de datos. La forma en que los datos son agrupados depende del tipo de algoritmo usado. Cada algoritmo usa una técnica diferente para identificar clusters.

Generalmente, el agrupamiento (Clustering) es utilizado en la minería de datos (Data Mining) para entender las propiedades de un dataset determinado. Después de desarrollar este entendimiento, la clasificación puede ser utilizada para hacer mejores predicciones sobre datos similares, pero nuevos, no mostrados.

El agrupamiento (Clustering) se puede aplicar a la categorización de documentos desconocidos y a campañas de marketing personalizadas, al agrupar clientes con comportamiento similar. El diagrama de dispersión de la Figura 2.22 ofrece una representación visual del funcionamiento de la técnica de agrupamiento (Clustering).

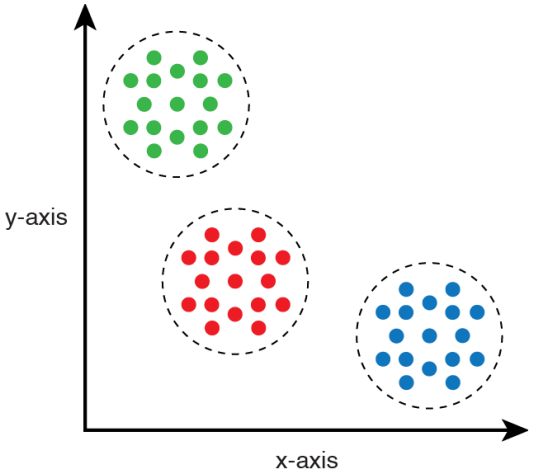


Figura 2.22 – Un diagrama de dispersión que resume los resultados del agrupamiento (Clustering).

Por ejemplo, el banco quiere que sus clientes conozcan una variedad de productos financieros nuevos, con base en los perfiles de clientes que tiene en sus registros. Los analistas clasifican a los clientes en varios grupos usando la técnica de agrupamiento (Clustering). Luego, se le da a conocer al grupo uno o más productos financieros que mejor encajen con las características de su perfil general.

Algunas preguntas de ejemplo pueden ser:

- *¿Cuántas especies diferentes de árboles existen, con base en las semejanzas entre ellos?*
- *¿Cuántas categorías diferentes de elementos hay en la tabla periódica?*
- *¿Cuáles son los diferentes grupos de virus, con base en sus características?*

Detección de datos atípicos (outliers)

La detección de datos atípicos (outliers) es el proceso de búsqueda de datos que son significativamente diferentes o inconsistentes con el resto de los datos dentro de un dataset determinado. Esta técnica de aprendizaje automático (Machine Learning) es usada para identificar anomalías, anormalidades y desviaciones que pueden ser favorables (como oportunidades) o no (como riesgos).

La detección de datos atípicos (outliers) está estrechamente relacionada con los conceptos de clasificación y agrupamiento (Clustering), a pesar de que sus algoritmos se centran en encontrar valores anormales. Puede estar basada en el aprendizaje supervisado o en el aprendizaje no supervisado. Algunas aplicaciones de la detección de datos atípicos (outliers) incluyen la detección de fraudes, los diagnósticos médicos, el análisis de datos (Data Analysis) de redes y el análisis de datos (Data Analysis) de sensores. Un diagrama de dispersión puede servir para identificar visualmente datos atípicos (outliers), como se muestra en la Figura 2.23.

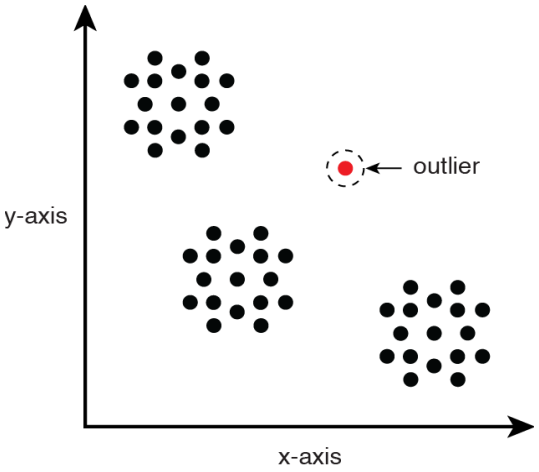


Figura 2.23 – Un diagrama de dispersión que resalta un dato atípico (outlier).

Por ejemplo, a fin de averiguar si una transacción es fraudulenta o no, el equipo de TI desarrolla un sistema empleando una técnica de detección de datos atípicos (outliers) basada en aprendizaje supervisado. Un conjunto de transacciones fraudulentas desconocidas son ingresadas primero en el algoritmo de detección de datos atípicos (outliers). Luego de entrenar el sistema, las transacciones son ingresadas en el algoritmo de detección de datos atípicos (outliers) para predecir si serán fraudulentas o no.

Algunas preguntas de ejemplo pueden ser:

- *¿Un jugador está consumiendo drogas que mejoran su desempeño?*
- *¿Existe alguna fruta o vegetal identificado erróneamente en el dataset de entrenamiento utilizado para la tarea de clasificación?*
- *¿Existe una cepa del virus que no responda a los medicamentos?*

Filtrado (filtering)

El filtrado (filtering) es el proceso automatizado de búsqueda de elementos desde un conjunto de elementos. Los elementos pueden ser filtrados con base en el comportamiento del usuario o comparando el comportamiento de múltiples usuarios. El filtrado (filtering) generalmente se aplica por medio de los dos siguientes enfoques:

- Filtrado (filtering) colaborativo
- Filtrado (filtering) basado en contenido

Un medio común por el cual se implementa el filtrado (filtering) es el sistema de recomendación. El filtrado (filtering) colaborativo es una técnica de filtrado de elementos basada en la colaboración, o fusión, del comportamiento pasado de un usuario. El comportamiento pasado de un usuario objetivo, incluyendo sus gustos, puntuaciones, historial de compras, etc., es comparado con el comportamiento de usuarios parecidos. De acuerdo con las semejanzas entre los comportamientos de los usuarios, los elementos del usuario objetivo son filtrados.

El filtrado (filtering) se basa exclusivamente en las semejanzas entre el comportamiento de los usuarios, y requiere una gran cantidad de datos de comportamiento de los usuarios, con el fin de filtrar los datos con exactitud. Ese es un ejemplo de la aplicación de la ley de los Grandes Números.

El filtrado (filtering) basado en contenido es una técnica de filtrado (filtering) de datos enfocada en las semejanzas entre usuarios y elementos. Un perfil de usuario se crea con base en el comportamiento pasado de un usuario, por ejemplo sus gustos, puntuaciones, historial de compras, etc. Las semejanzas identificadas entre el perfil del usuario y los atributos de varios elementos facilitan que se filtren los elementos del usuario. Al contrario del filtrado (filtering) colaborativo, el filtrado (filtering) basado en contenido se centra exclusivamente en las preferencias individuales del usuario y no requiere datos sobre otros usuarios.

Un sistema de recomendación predice las preferencias de un usuario y genera sugerencias para el usuario, según corresponda. Usualmente, las sugerencias se relacionan con elementos recomendados, como películas, libros, páginas web, personas, etc. Un sistema de recomendación usa la técnica de filtrado (filtering) colaborativo o la técnica de filtrado (filtering) basado en contenido para generar sugerencias. También puede estar basado en un híbrido de las dos técnicas mencionadas, con el objetivo de ajustar la exactitud y efectividad de las sugerencias generadas.

Por ejemplo, con el fin de hallar oportunidades de ventas cruzadas, el banco crea un sistema de recomendación que utiliza filtrado (filtering) basado en contenido. Con base en las coincidencias encontradas entre los productos financieros adquiridos por los clientes y las propiedades de productos financieros similares, el sistema de recomendación automatiza sugerencias para posibles productos financieros que también podrían ser de interés de los clientes.

Algunas preguntas de ejemplo pueden ser:

- *¿Cómo se podrían mostrar únicamente los artículos de las noticias que le interesan a un usuario?*
- *¿Cuáles son los destinos vacacionales recomendados con base en el historial de viajes de un turista?*
- *¿Qué otros nuevos usuarios se pueden sugerir como amigos con base en el perfil actual de una persona?*

Notas

Notas / Bocetos

Análisis semántico

Un fragmento de texto o los datos de reconocimiento de voz pueden tener distintos significados en diferentes contextos, mientras que una oración completa puede conservar su significado incluso si está estructurada de formas distintas. Con el fin de extraer información valiosa, los datos de textos y de reconocimiento de voz deben ser comprendidos por las máquinas de la misma forma que los humanos los comprendemos. El análisis semántico representa las prácticas que buscan extraer información importante de los datos de texto y de reconocimiento de voz.

En esta sección se describen los siguientes tipos de análisis semántico:

- Procesamiento de lenguaje natural (NLP)
- Analítica de texto (text analytics)
- Análisis de sentimientos (Sentiment Analysis)

Procesamiento de lenguaje natural

El procesamiento de lenguaje natural es la capacidad que tiene una computadora de comprender el discurso y el texto humano de forma natural, tal como lo hacen las personas. Esto permite que las computadoras ejecuten una variedad de tareas útiles, como búsquedas de textos completos.

A fin de mejorar la calidad del servicio al cliente, la empresa de helados utiliza el procesamiento de lenguaje natural para transcribir las llamadas de los clientes en datos textuales que posteriormente son analizados para conocer las razones más comunes de insatisfacción de los clientes.

En lugar de codificar en sí las reglas de lenguaje necesarias, se aplica el aprendizaje automático (Machine Learning) —supervisado o sin supervisar— con el fin de desarrollar en la computadora el entendimiento del lenguaje natural. En general, cuantos más datos de aprendizaje tenga la computadora, mayor será la precisión con la que podrá descifrar el discurso y el texto producidos por los seres humanos.

El procesamiento de lenguaje natural incluye reconocimiento tanto de texto como de voz. En el caso del reconocimiento de voz, el sistema intenta comprender lo que se dice y luego realiza una acción; por ejemplo, transcribir el texto.

Algunas preguntas de ejemplo pueden ser:

- *¿Cómo se puede desarrollar un sistema de telefonía que reconozca la extensión del departamento correcto cuando es pronunciada por el interlocutor?*
- *¿Cómo se pueden identificar automáticamente los errores gramaticales?*
- *¿Cómo se puede diseñar un sistema que entienda correctamente los distintos acentos del inglés?*

Analítica de texto (text analytics)

Comparado con el texto estructurado, el texto sin estructurar es generalmente mucho más difícil de analizar y de buscar. La analítica de texto (text analytics) es el análisis especializado de texto mediante la aplicación de técnicas de minería de datos (Data Mining), aprendizaje automático (Machine Learning) y procesamiento de lenguaje natural, con el fin de obtener valor del texto sin estructurar. En esencia, la analítica de texto (text analytics) proporciona la capacidad de descubrir el texto en lugar de simplemente buscarlo.

Se puede obtener información útil a partir de datos basados en textos ayudando a las empresas a comprender la información contenida en una cantidad considerable de texto. Como continuación del ejemplo anterior del NLP, los datos textuales transcritos son analizados aún más usando la analítica de texto (text analytics), con el fin de extraer información importante sobre las principales razones de la insatisfacción de los clientes.

El principio básico de la analítica de texto (text analytics) es convertir el texto sin estructurar en datos que pueden ser buscados y analizados. A medida que aumenta la cantidad de documentos digitalizados, correos electrónicos, publicaciones de social media y archivos de registro (log files), aumenta la necesidad de las empresas de aprovechar cualquier clase de valor que puedan obtener de estas formas de datos semiestructurados y sin estructurar. Si las empresas se dedican exclusivamente al análisis de datos operativos (estructurados), podrían pasar por alto oportunidades de ahorrar costos o expandir sus negocios, especialmente aquellas empresas enfocadas en los clientes.

Algunas aplicaciones incluyen la clasificación y búsqueda de documentos, así como la generación de un panorama completo de los clientes mediante la extracción de información de un sistema de Relación con los Clientes (CRM).

La analítica de texto (text analytics) por lo general incluye dos pasos:

1. Análisis del texto contenido en los documentos, con el fin de extraer:
    - Entidades con nombre: personas, grupos, lugares, empresas
    - Entidades basadas en patrones: números de seguro social, códigos postales
    - Conceptos: una representación abstracta de una entidad
    - Hechos: relaciones entre entidades
2. Categorización de los documentos usando las entidades y hechos extraídos.

Se puede usar la información extraída para realizar una búsqueda (dependiendo del contexto) de entidades, con base en el tipo de relaciones entre las mismas. La Figura 2.24 muestra una representación simplificada del análisis de texto.

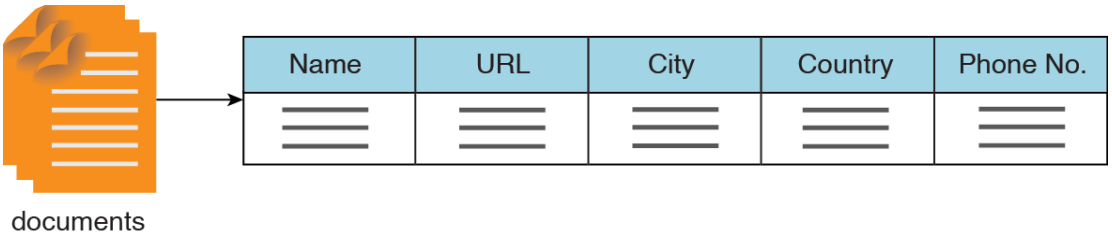


Figura 2.24 – Entidades extraídas de los archivos de texto usando reglas semánticas, estructuradas de forma que puedan ser buscadas.

Algunas preguntas de ejemplo pueden ser:

- *¿Cómo puedo categorizar los sitios web con base en el contenido de sus páginas?*
- *¿Cómo puedo encontrar libros con contenido relevante para el tema que estoy estudiando?*
- *¿Cómo puedo identificar los contratos que tienen información confidencial de la empresa?*

Análisis de sentimientos (Sentiment Analysis)

El análisis de sentimientos (Sentiment Analysis) es una forma especializada de análisis de texto, enfocada en determinar el sesgo o las emociones de las personas. Este tipo de análisis determina la actitud del autor de un texto, analizando el texto dentro del contexto del lenguaje natural. El análisis de sentimientos (Sentiment Analysis) no solo proporciona información sobre los sentimientos de las personas, sino además sobre la intensidad de dichos sentimientos. Esta información puede ser integrada posteriormente en el proceso de toma de decisiones. Entre algunas aplicaciones comunes del análisis de sentimientos (Sentiment Analysis) están la identificación temprana de los niveles de satisfacción de los clientes, la medición del éxito de un producto y la detección de tendencias nuevas.

Por ejemplo, la empresa de helados quiere saber qué sabores de helados son los que más les gustan a los niños. Los datos de ventas por sí solos no proporcionan esta información, ya que los niños que consumen los helados no son necesariamente quienes los compran. Aquí se aplica el análisis de sentimientos (Sentiment Analysis) a la retroalimentación de los clientes que está archivada en el sitio web de la empresa, con el fin de extraer información específicamente relacionada con la preferencia de los niños por ciertos sabores de helado.

Algunas preguntas de ejemplo pueden ser:

- *¿Cómo podemos medir las reacciones de los clientes ante el nuevo empaque del producto?*
- *¿Cuál de los participantes de un concurso de canto es el probable ganador?*
- *¿Los clientes se están cambiando a la competencia?*

Relación de temas de análisis

Los siguientes temas de análisis y analítica fueron examinados en el Módulo 1:

- Análisis cuantitativo

- Análisis cualitativo
- Minería de datos (Data Mining)
- Analítica descriptiva
- Analítica diagnóstica
- Analítica predictiva
- Analítica prescriptiva
- Aprendizaje supervisado
- Aprendizaje no supervisado

La mayor parte de estas prácticas relacionadas con el análisis pueden ser aplicadas por —o de alguna forma están relacionadas con— algunas de las técnicas de análisis precedentes. La siguiente lista describe brevemente cómo pueden estar relacionadas estas áreas temáticas.

- Análisis cuantitativo: algunos ejemplos de análisis cuantitativo son la correlación y la regresión. Los tests A/B pueden utilizar las técnicas de análisis cuantitativo para comparar los resultados.
- Análisis cualitativo: el NLP, la analítica de texto (text analytics) y el análisis de sentimientos (Sentiment Analysis) pueden ser utilizados para respaldar el análisis cualitativo.
- Minería de datos (Data Mining): la minería de datos (Data Mining) puede ser ejecutada o respaldada por medio de correlaciones, mapas de calor, análisis de series temporales, análisis de redes, análisis de datos espaciales, agrupamiento (Clustering), detección de datos atípicos (outliers), procesamiento de lenguaje natural y analítica de texto (text analytics).
- Analítica descriptiva: los tests A/B, mapas de calor y análisis de datos espaciales son considerados formas de analítica descriptiva.
- Analítica diagnóstica: las correlaciones, la regresión, los análisis de series temporales, análisis de redes y análisis de datos espaciales son considerados formas de analítica diagnóstica.
- Analítica predictiva: las correlaciones, la regresión, los análisis de series temporales, la clasificación, el agrupamiento (Clustering), la detección de datos atípicos (outliers), el filtrado (filtering), procesamiento de lenguaje natural, la analítica de texto (text analytics) y el análisis de sentimientos (Sentiment Analysis) son considerados formas de analítica predictiva.
- Analítica prescriptiva: la analítica prescriptiva está basada en técnicas de analítica predictiva, y por lo tanto está relacionada con las mismas técnicas de análisis que la analítica predictiva. Asimismo, la analítica prescriptiva puede emplear mapas de calor, análisis de redes y análisis de datos espaciales para ilustrar los resultados de forma gráfica.
- Aprendizaje supervisado: la clasificación, detección de datos atípicos (outliers), filtrado (filtering), procesamiento de lenguaje natural, analítica de texto (text analytics) y el análisis de sentimientos (Sentiment Analysis) pueden utilizar el aprendizaje supervisado.

- Aprendizaje no supervisado: el agrupamiento (Clustering), la detección de datos atípicos (outliers), filtrado (filtering), procesamiento de lenguaje natural, analítica de texto (text analytics) y el análisis de sentimientos (Sentiment Analysis) pueden hacer uso del aprendizaje no supervisado.

Ejercicio 2.2: relacione los enunciados de los problemas con las técnicas de análisis

Identifique correctamente qué técnica de análisis se aplica mejor a cada uno de los siguientes enunciados de problemas. *Las respuestas al ejercicio se encuentran al final de este cuadernillo.*

Analítica de texto (text analytics)

Análisis de series temporales

Agrupamiento (Clustering)

Filtrado (filtering)

Análisis de redes

Regresión

Clasificación

Análisis espacial de datos

NLP

Análisis de sentimientos (Sentiment Analysis)

Correlación

Tests A/B

Detección de datos atípicos (outliers)

Mapas de calor

1. Una nueva empresa startup planea lanzar una aplicación para teléfonos inteligentes que recomienda productos a sus usuarios con base en comentarios subjetivos recibidos en múltiples sitios web de social media y blogs. ¿Qué técnica de análisis semántico puede usar la empresa para interpretar los comentarios textuales, con el fin de comprender las razones por las cuales un usuario recomienda un producto?

   _____

2. Una agencia de empleo adquirió otras dos agencias como parte de una expansión corporativa. Con el fin de minimizar los costos operativos, la alta gerencia planea disminuir la cantidad de reclutadores. Sin embargo, algunos gerentes señalaron que esto puede tener como resultado un aumento en los tiempos de atención, en lo que se refiere al escaneo de documentos y ubicación de los candidatos correctos. Uno de los gerentes con más conocimientos tecnológicos considera que esta tarea puede ser automatizada. ¿Qué técnica de análisis semántico puede utilizar el equipo de TI para escanear una gran cantidad de solicitudes de candidatos de forma automática?

   _____

3. Jane es una bióloga que cree que el índice de fotosíntesis está relacionado de alguna manera con la intensidad de la luz. Ella ha recopilado algunos datos sobre la intensidad de la luz y el índice de fotosíntesis. ¿Qué técnica de análisis estadístico necesita Jane para probar o refutar su hipótesis?

_____

4. Jane ha comprobado que el índice de fotosíntesis está relacionado con la intensidad de la luz. Ahora, ella quiere determinar con precisión la intensidad de luz óptima para obtener el índice más alto de fotosíntesis. ¿Qué técnica de análisis estadístico debe usar Jane?

_____

5. Una agencia ambiental desea cuantificar el daño causado por el vertimiento de líquidos peligrosos, el cual a su vez es el resultado de accidentes de camiones comerciales. La agencia obtuvo datos relacionados con la configuración de las tuberías interconectadas de drenaje al costado de las carreteras. ¿Qué técnica de análisis visual debe ser usada para analizar el flujo de líquido en las tuberías de drenaje?

_____

6. Un sitio web de social media les proporciona a sus usuarios la capacidad de personalizar el contenido visualizado. Actualmente, los usuarios solo pueden personalizar las categorías de contenido. Sin embargo, el contenido en sí, tal como noticias o juegos, es el mismo para todos los usuarios que seleccionaron una categoría específica. A fin de mejorar aún más la experiencia online de los usuarios, los desarrolladores web desean personalizar el contenido hasta el nivel de cada usuario. ¿Qué técnica de aprendizaje automático (Machine Learning) se debe usar en este escenario?

_____

7. Roger trabaja como analista en una gran cadena hotelera. Se le ha solicitado que elabore un informe sobre los niveles diarios de ocupación de las habitaciones de cada uno de los hoteles en el país. Roger obtiene los datos diarios de cada hotel. A fin de hacer su informe más sencillo, Roger planea usar valores descriptivos como bajo, medio y alto para representar los niveles de ocupación de las habitaciones. ¿Qué técnica de análisis visual debe usar Roger para que la gerencia pueda determinar con facilidad los hoteles cuyas habitaciones tienen un nivel de ocupación bajo?

_____

8. Joe trabaja para una agencia policial y ha recopilado cifras relacionadas con crímenes durante los últimos 10 años. Mientras buscaba entre los datos, Joe identificó un posible patrón. Sin embargo, las cifras en cuestión se remontan a unos pocos meses atrás, y él no está seguro del tipo de patrón que identificó. Joe considera que la identificación correcta de tendencias puede ayudar a reducir los crímenes de forma proactiva. ¿Qué técnica de análisis visual debe emplear Joe?

_____

9. Alice es una estudiante de investigación médica que actualmente está estudiando la relación entre la demográfica de los pacientes y sus enfermedades. Ella ha recopilado varios casos conocidos de muestra, donde se identifican factores fundamentales cuya presencia tiene como resultado la posible infección con ciertas enfermedades. ¿Qué técnica de aprendizaje automático (Machine Learning) debe usar Alice para predecir si un nuevo paciente contraerá una enfermedad específica?

_____

10. Henry es un botánico que ha estado recopilando datos sobre especies de plantas durante su investigación en el Amazonas. Algunas de estas especies de plantas tienen características comunes, mientras que otras son muy distintas. Sin embargo, Henry no

puede identificar las características comunes. En su intento inicial de agrupar las especies de plantas semejantes, Henry no tuvo éxito. ¿Qué técnica de aprendizaje automático (Machine Learning) debería usar Henry para agrupar las especies de plantas semejantes?

_____

11. En las elecciones nacionales, el partido de la oposición hizo críticas de fraude, motivo por el cual se estableció una comisión investigativa. Los analistas que trabajaron con la comisión tuvieron acceso a las cifras de votos en todas las urnas de votación. ¿Qué técnica de aprendizaje automático (Machine Learning) usaron los analistas para identificar las urnas en las que se presentaron patrones anormales de votación?

_____

12. John ha estado llevando a cabo una investigación sobre personas que sufren de enfermedades relacionadas con el humo. La investigación inicial muestra que hay una conexión directa entre las fábricas que emiten humo y la distancia a la que se encuentran de las residencias donde habitan las personas afectadas. Como parte de la preparación de un informe, John quiere determinar cuál puede ser recomendada como la distancia segura entre las fábricas y los barrios residenciales. ¿Qué técnica de análisis visual debe aplicar si dispone de los datos sobre la ubicación de las personas afectadas y las fábricas?

_____

13. Adam es un diseñador web que está haciendo cambios a la página de inicio de un sitio web de ventas minoristas, con el fin de mejorar la navegación y de que sea más fácil de usar. Sin embargo, Adam no está seguro de cuál es el diseño más adecuado. ¿Qué técnica de análisis estadístico debe usar Adam para determinar el diseño más adecuado?

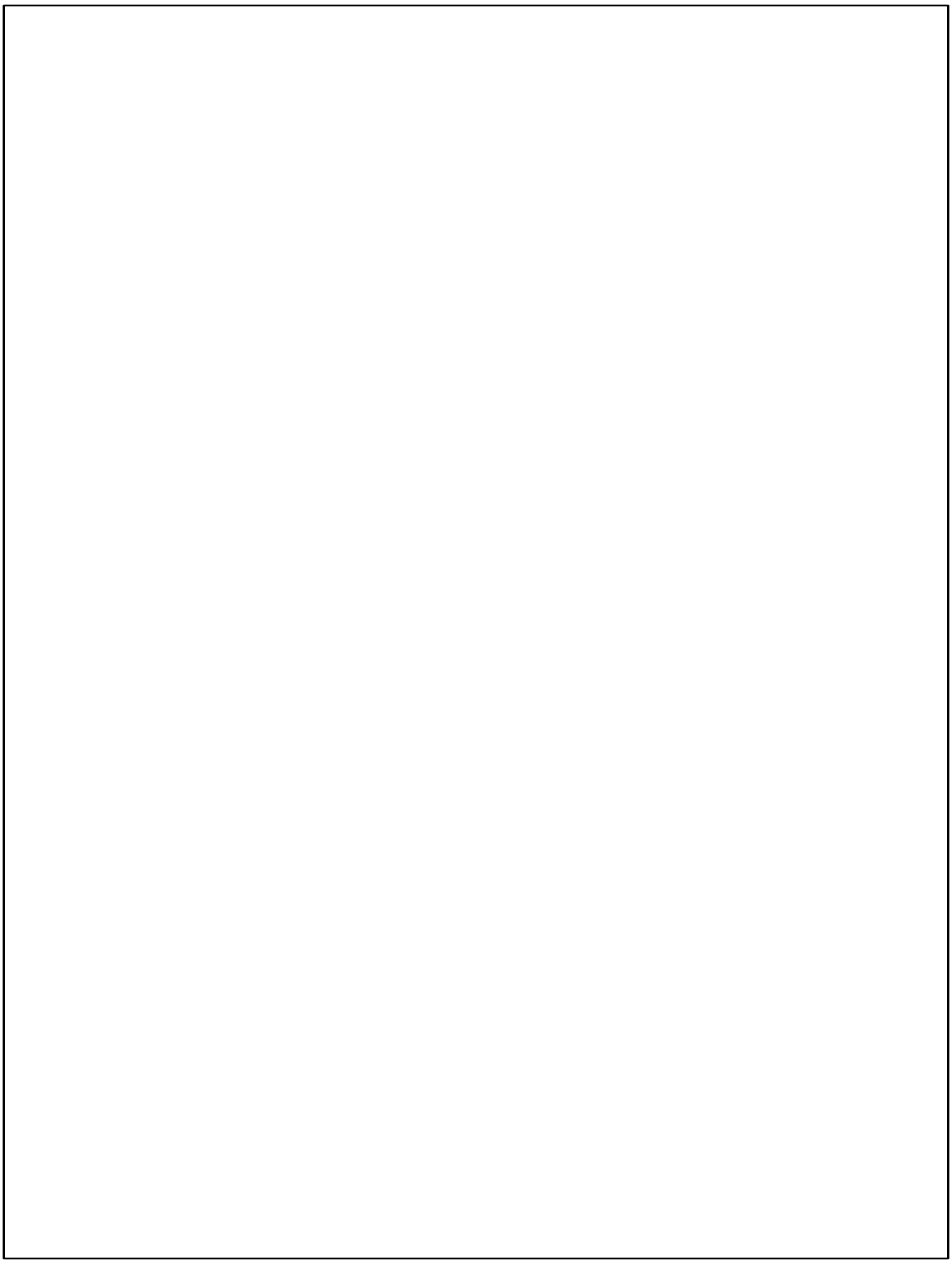
_____

14. Recientemente, una biblioteca digitalizó la mayor parte de sus libros a fin de que sus afiliados puedan acceder a los libros en formato electrónico mediante la internet. Sin embargo, algunos de sus afiliados con más edad se quejaron de que no pueden leer en una pantalla por tiempo prolongado debido a problemas de visión. La biblioteca entonces decidió convertir los libros digitalizados en audiolibros. ¿Qué técnica de análisis semántico se debe usar para convertir los libros electrónicos en audiolibros?

_____

Notas

Notas / Bocetos



Parte III: conceptos tecnológicos de Big Data

Esta parte del cuaderno de trabajo está dividida en las siguientes secciones:

- Consideraciones tecnológicas de Big Data
- Mecanismos tecnológicos de Big Data

Consideraciones tecnológicas de Big Data

Esta sección presenta los siguientes componentes y conceptos tecnológicos clave que son relevantes para las soluciones y los mecanismos de Big Data:

- Clusters
- Sistemas de archivos y sistemas de archivos distribuidos
- NoSQL
- Procesamiento distribuido de datos
- Procesamiento de datos en paralelo
- Cargas de trabajo de procesamiento
- Cloud Computing

Clusters

En informática, un cluster es un conjunto perfectamente acoplado de servidores o nodos. Por lo general, estos servidores tienen las mismas especificaciones de hardware y están conectados por medio de una red para trabajar como una sola unidad, como se muestra en la Figura 2.25. Cada nodo en el cluster cuenta con sus propios recursos dedicados —como memoria y disco duro— y ejecuta su propio sistema operativo, igual que un computador de escritorio. Se puede utilizar un cluster para ejecutar una tarea con base en frameworks de procesamiento distribuido/paralelo.

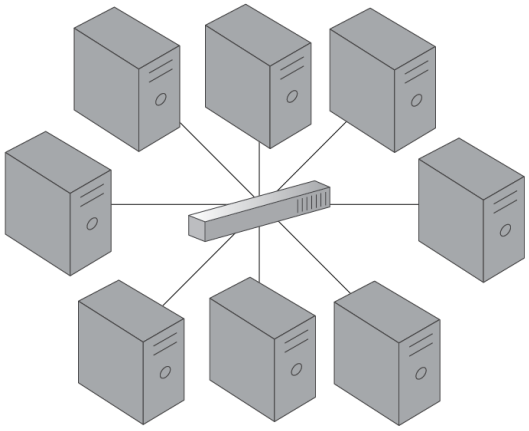
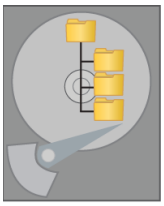


Figura 2.25 – Símbolo utilizado para representar un cluster.

Sistemas de archivos

Un sistema de archivos es un método de almacenamiento y organización de datos en un medio de almacenamiento; por ejemplo unidades flash, DVD y discos duros, representados en la Figura 2.26. Un archivo es una unidad atómica de almacenamiento usada por el sistema de archivos para almacenar datos. Los archivos son organizados dentro de directorios.

Un sistema de archivos ofrece una vista lógica de los datos almacenados en un medio de almacenamiento en forma de una estructura de árbol de archivos y directorios. Los sistemas operativos hacen uso de los sistemas de archivos para almacenar datos. Cada sistema operativo soporta uno o más sistemas de archivos, como NTFS en Windows y ext en Linux.



hard drive

Figura 2.26 – Símbolo utilizado para representar un disco duro.

Sistemas de archivos distribuidos

En informática, un cluster es un conjunto perfectamente acoplado de servidores o nodos. Estos servidores se encuentran conectados por medio de una red para trabajar como una sola unidad. Un sistema de archivos distribuido es un sistema de archivos que puede almacenar una gran cantidad de archivos distribuidos en un cluster, como se ilustra en la Figura 2.27.

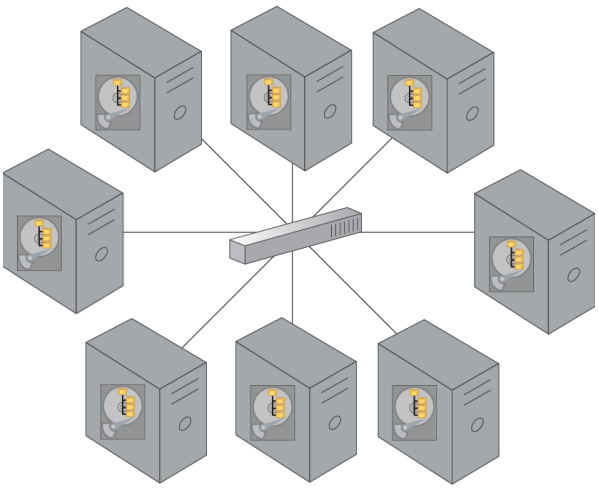


Figura 2.27 – Símbolo utilizado para representar los sistemas de archivos distribuidos.

El cliente ve los archivos como si fueran locales, y puede acceder a ellos a través de múltiples ubicaciones. Algunos ejemplos son Google File System (GFS) y Hadoop Distributed File System (HDFS).

NoSQL

Una base de datos NoSQL es una base de datos no relacional altamente escalable, tolerante a errores y que fue específicamente diseñada para albergar datos sin estructurar. Una base de datos NoSQL generalmente proporciona una interfaz de consulta basada en API, en vez de una interfaz SQL. Sin embargo, algunas bases de datos NoSQL también proporcionan una interfaz de consulta basada en SQL, como se muestra en la Figura 2.28.

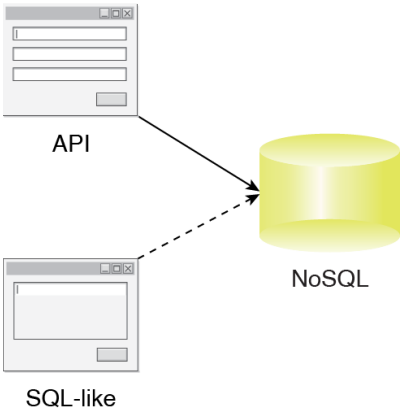


Figura 2.28 – Una base de datos NoSQL puede proporcionar una interfaz de consulta basada en API o en SQL.

Procesamiento de datos en paralelo

El procesamiento de datos en paralelo implica la ejecución simultánea de múltiples subtareas que, en conjunto, componen una tarea mayor. La idea es reducir el tiempo de ejecución al dividir una tarea grande en múltiples tareas más pequeñas.

A pesar de que el procesamiento de datos en paralelo se puede lograr por medio de múltiples máquinas conectadas en red, lo más común es utilizar una sola máquina con múltiples procesadores o núcleos, como se muestra en la Figura 2.29.

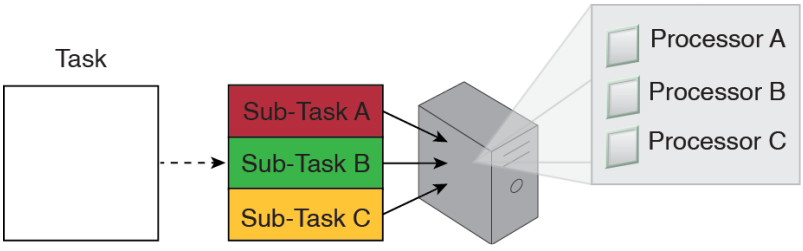


Figura 2.29 – Una tarea se puede dividir en tres subtareas que son ejecutadas en paralelo por tres procesadores distintos de la misma máquina.

Procesamiento distribuido de datos

La relación entre el procesamiento distribuido de datos y el procesamiento de datos en paralelo es que se aplica el principio de dividir un problema difícil en tantas partes como sea necesario. Sin embargo, el procesamiento distribuido de datos se logra por medio de máquinas que se encuentran separadas físicamente, pero que están unidas en red como un cluster. En la Figura 2.30, una tarea ha sido dividida en tres subtareas que son ejecutadas en tres máquinas diferentes que comparten un mismo switch.

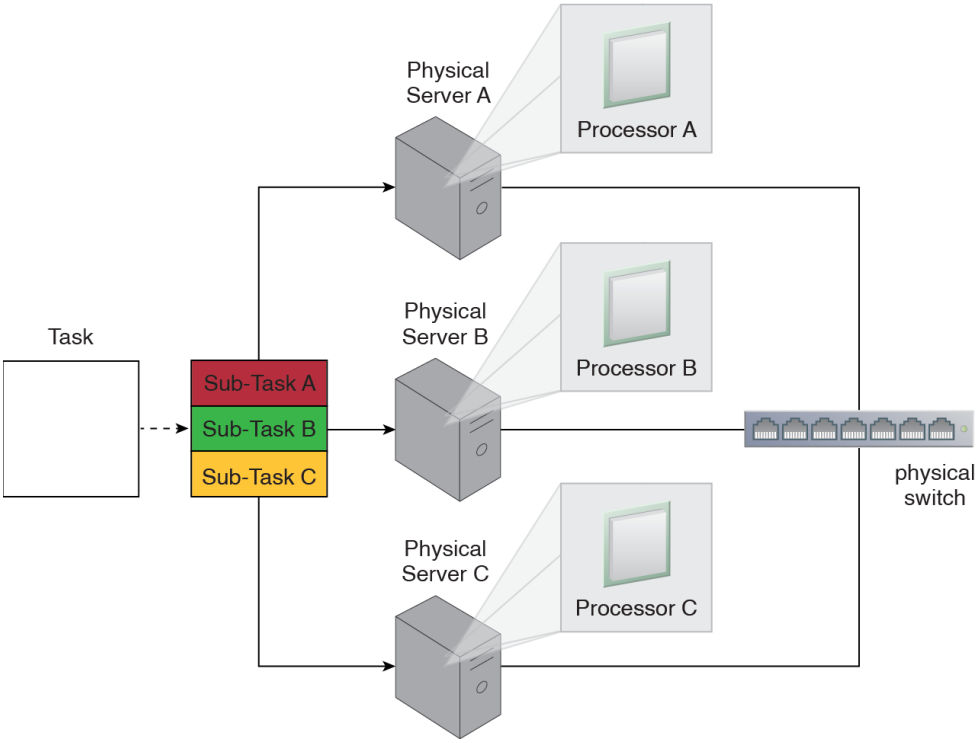


Figura 2.30 – Ejemplo de procesamiento distribuido de datos.

Cargas de trabajo de procesamiento

En Big Data, una carga de trabajo de procesamiento se define como la cantidad y tipo de datos que son procesados en un determinado lapso de tiempo. Por lo general, las cargas de trabajo se dividen en dos tipos:

- Por lotes
- Transaccionales

Procesamiento de cargas de trabajo: por lotes

El procesamiento de cargas de trabajo por lotes, también conocido como procesamiento fuera de línea, implica el procesamiento de datos por lotes y usualmente genera retrasos (que tienen como resultado respuestas de alta latencia). Por lo general, las cargas de trabajo por lotes

suponen grandes cantidades de datos con operaciones secuenciales de lectura o escritura, y constituyen grupos de peticiones de lectura o escritura.

Estas peticiones pueden ser complejas e involucrar múltiples joins. Los sistemas de OLAP generalmente procesan las cargas de trabajo por lotes. La Inteligencia de negocios (BI) y el análisis estratégico pertenecen a esta categoría, ya que son tareas con un nivel de lectura altamente intensivo que implican grandes volúmenes de datos. Como se muestra en la Figura 2.31, una carga de trabajo por lotes comprende operaciones agrupadas de lectura/escritura con un volumen de datos más grande, que consisten en joins complejos y respuestas con mucha latencia.

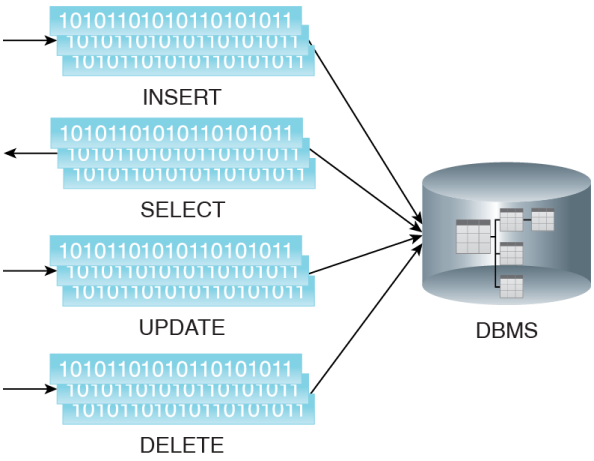


Figura 2.31 – Una carga de trabajo por lotes puede incluir operaciones agrupadas de lectura/escritura para INSERTAR, SELECCIONAR, ACTUALIZAR y ELIMINAR.

Procesamiento de cargas de trabajo: transaccionales

El procesamiento de cargas de trabajo transaccionales, también conocido como procesamiento online, sigue un enfoque en el cual los datos son procesados de forma interactiva y sin retrasos, lo que tiene como resultado respuestas con poca latencia. Las cargas de trabajo transaccionales incluyen cantidades pequeñas de datos con operaciones aleatorias de lectura/escritura.

Dentro de esta categoría se encuentran el OLTP y los sistemas operacionales —los cuales tienen un alto nivel de operaciones de escritura—, así como la Inteligencia de negocios (BI) y la analítica operacional (operational analytics), que tienen niveles altos de operaciones de lectura. A pesar de que estas cargas de trabajo contienen una mezcla de consultas de lectura y escritura, por lo general tienen más operaciones de escritura que de lectura.

Las cargas de trabajo transaccionales están compuestas por operaciones aleatorias de lectura/escritura que incluyen menos joins y requieren respuestas con menos latencia y un volumen de datos menor, como se muestra en la Figura 2.32.

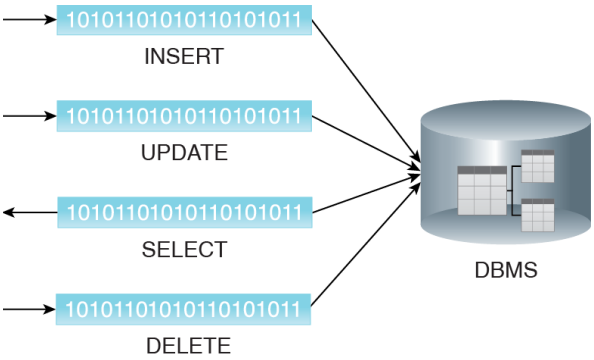


Figura 2.32 – Las cargas de trabajo transaccionales tienen menos joins y una latencia menor que las cargas de trabajo por lotes.

Cloud Computing

Cloud Computing consiste en una forma especializada de informática distribuida que introduce modelos de utilización para recursos de TI escalables y cuantificables de suministro remoto. Las soluciones de datos masivos pueden ser implementadas parcial o totalmente en la nube, con el fin de aprovechar los recursos de almacenamiento y cómputo de los que dispone el proveedor de la nube.

Los servicios de procesamiento en clúster que requieren las soluciones de datos masivos pueden beneficiarse de los recursos de TI altamente escalables y flexibles que se encuentran disponibles en entornos basados en la nube. El procesamiento de datos basado en lotes de Hadoop es ideal para el modelo de informática en la nube de “pago por uso”, el cual puede reducir los costos operativos, ya que el tamaño de un clúster de Hadoop puede variar de algunos a unos cuantos miles de nodos.

Tiene sentido que las empresas ya estén implementando Cloud Computing con el objetivo de reutilizar la nube para sus iniciativas de Big Data por las siguientes razones:

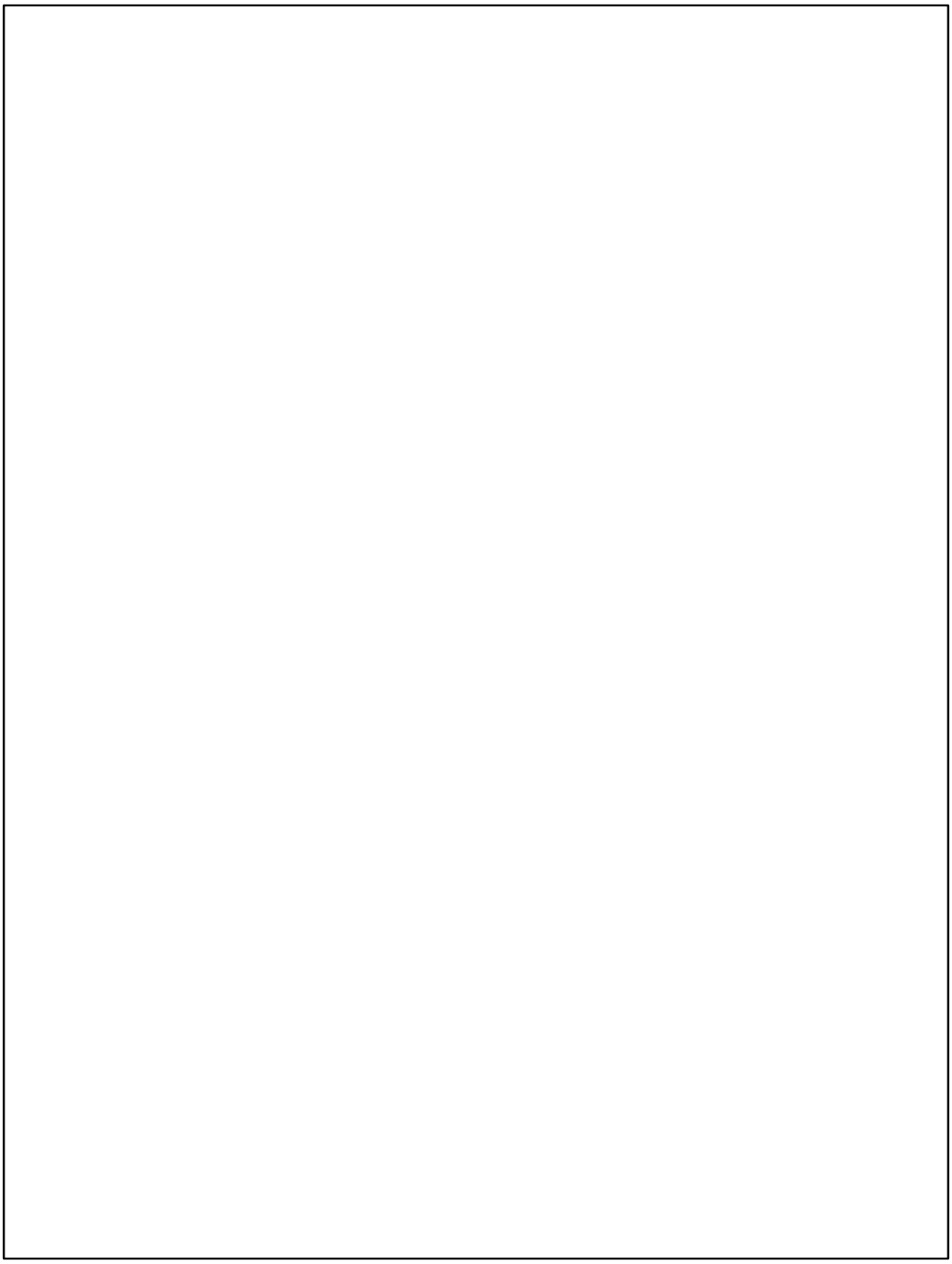
- el equipo de TI ya cuenta con las capacidades necesarias de Cloud Computing
- los datos de entrada ya existen en la nube

Es lógico que, aquellas empresas que tienen pensado ejecutar analítica en datasets que están disponibles en el mercado de datos, migren a la nube, ya que la mayoría de los mercados almacenan sus datos en la nube; por ejemplo, Amazon S3.

En resumen, Cloud Computing ofrece tres componentes requeridos por la solución de Big Data: datos de entrada, cálculo y almacenamiento.

Notas

Notas / Bocetos



Mecanismos tecnológicos de Big Data

Las soluciones de Big Data requieren un entorno de procesamiento distribuido que pueda hospedar datos veloces, variados y en grandes volúmenes. Este tipo de entorno es proporcionado por una plataforma compuesta por un conjunto de almacenamiento distribuido y tecnologías de procesamiento.

Los mecanismos de Big Data representan los principales componentes comunes de las soluciones de Big Data, independientemente de los productos de código abierto o de proveedores usados durante la implementación. Esta sección aborda estos mecanismos y su funcionamiento como factores activos en las soluciones de Big Data, a fin de proporcionar las características y funcionalidades requeridas para ejecutar el ciclo de vida de análisis de Big Data.

Se abordan los siguientes mecanismos fundamentales de Big Data:

- Dispositivo de almacenamiento
- Motor de procesamiento
- Gestor de recursos
- Motor de transferencia de datos
- Motor de consultas (Query Engine)
- Motor analítico (Analytics Engine)
- Motor de flujo de trabajo (Workflow)
- Motor de coordinación

Como mínimo, cualquier solución de Big Data debe incluir mecanismos de motor de procesamiento, dispositivo de almacenamiento y gestor de recursos, a fin de procesar efectivamente grandes datasets y respaldar el ciclo de vida de análisis de Big Data.

Complemento de Ejemplos de mecanismos de Big Data

Los mecanismos que se describen en esta sección se presentan intencionalmente como componentes abstractos con conjuntos de características comunes y fundamentales. El cuadernillo complementario *Ejemplos de mecanismos de Big Data* incluido en este curso muestra ejemplos reales de productos de código abierto para cada uno de los mecanismos.

Cada implementación de producto para un mecanismo será diferente y puede incluir características y tecnologías propietarias. Como se explica en el complemento, el framework de código abierto de Hadoop, presentado brevemente en el módulo 1, establece un entorno de solución compuesto por varios mecanismos.

Dispositivo de almacenamiento

Los dispositivos de almacenamiento proporcionan un entorno de almacenamiento de datos subyacente para guardar los datasets que posteriormente serán procesados por las soluciones de Big Data. Un dispositivo de almacenamiento puede tener un sistema de archivos distribuido o una base de datos.

Figura 2.33 – Símbolo utilizado para representar los dispositivos de almacenamiento.

Los sistemas de archivos distribuidos se pueden usar para guardar datos inmutables que serán utilizados con fines de acceso de transmisión y procesamiento por lotes (Batch Processing). Las bases de datos, tales como los repositorios NoSQL, pueden ser usadas para almacenar datos estructurados y sin estructurar, y para acceder a datos de lectura/escritura, como se muestra en la figura 2.34. Tenga en cuenta que los sistemas de archivos distribuidos y las bases de datos se encuentran en dispositivos de almacenamiento en disco.

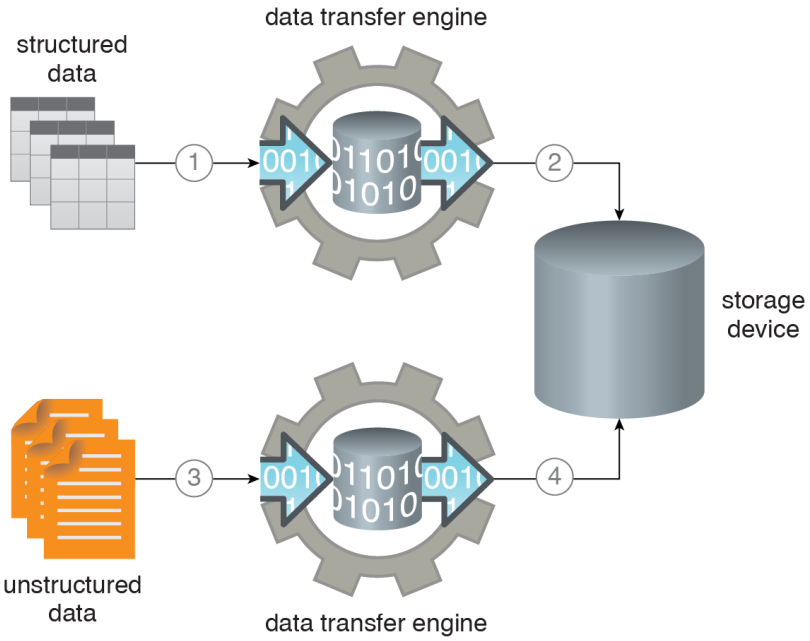


Figura 2.34 – Los datos estructurados son importados al dispositivo de almacenamiento (1) mediante un motor de transferencia de datos (2). Los datos sin estructurar (3) son importados usando otro tipo de motor de transferencia de datos (4).

Motor de procesamiento

El motor de procesamiento es responsable del procesamiento de los datos recuperados generalmente desde los dispositivos de almacenamiento, basados en una lógica predefinida,

con el fin de producir un resultado. Cualquier procesamiento de datos requerido por la solución de Big Data se ejecuta mediante el motor de procesamiento.

Figura 2.35 – Símbolo utilizado para representar un motor de procesamiento.

Un motor de procesamiento de Big Data utiliza un framework distribuido de programación paralela que le permite procesar grandes cantidades de datos distribuidos entre múltiples nodos. Para hacerlo, necesita recursos de procesamiento del gestor de recursos.

Los motores de procesamiento se dividen en dos categorías:

- Un motor de procesamiento por lotes (Batch Processing) que soporta el procesamiento de datos por lotes, en el que las tareas de procesamiento pueden tardar minutos o incluso horas para completarse. Este tipo de motor de procesamiento presenta una latencia alta.
- Un motor de procesamiento en tiempo real que soporta el procesamiento de datos en tiempo real, incorporando tiempos de respuesta menores a un segundo. Este tipo de motor de procesamiento presenta una latencia baja.

Los requisitos de procesamiento de la solución de Big Data determinan el tipo de motor de procesamiento que se utilizará. La Figura 2.36 muestra un ejemplo en el cual un trabajo de procesamiento es reenviado a un motor de procesamiento mediante el gestor de recursos.

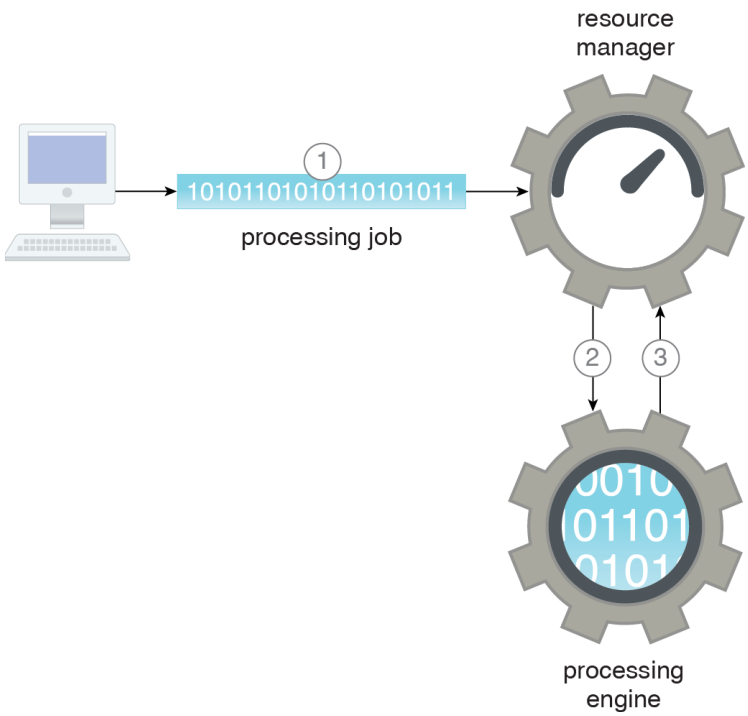


Figura 2.36 – Un trabajo de procesamiento es enviado al gestor de recursos (1). El gestor de recursos posteriormente asigna un conjunto inicial de recursos y envía el trabajo al motor de procesamiento (2), en el cual se solicitan más recursos del gestor (3).

Gestor de recursos

Una solución de Big Data puede procesar los datos almacenados en varias formas, y todas las solicitudes de procesamiento de datos requieren la asignación de recursos de procesamiento. Los usuarios de las soluciones de Big Data pueden hacer varias solicitudes de procesamiento de datos, cada una de las cuales puede tener diferentes requisitos en términos de cargas de trabajo de procesamiento.

Figura 2.37 – Símbolo utilizado para representar el gestor de recursos.

Un gestor de recursos opera como un planificador que prioriza y coordina las solicitudes de procesamiento, de acuerdo con los requisitos individuales de cargas de trabajo de procesamiento. El gestor de recursos actúa esencialmente como un árbitro de recursos, el cual gestiona y asigna los recursos disponibles, como se muestra en el ejemplo de la figura 2.38.

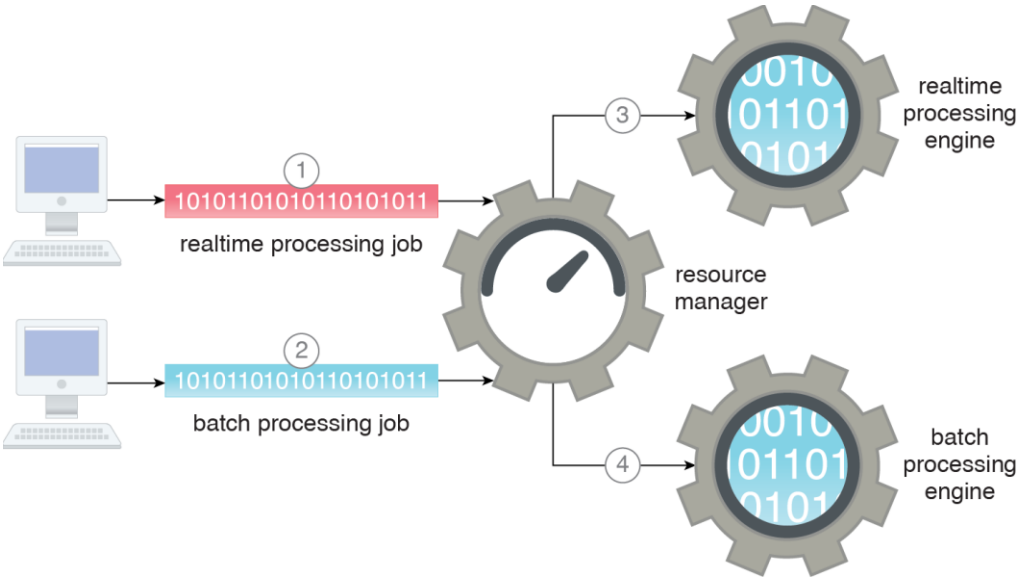


Figura 2.38 – Un trabajo de procesamiento en tiempo real (1) y un trabajo de procesamiento por lotes (Batch Processing) (2) son enviados para su ejecución. El gestor de recursos asigna los recursos de acuerdo con los requisitos de carga de trabajo y, a continuación, planifica los trabajos en un motor de procesamiento en tiempo real (3) y un motor de procesamiento por lotes (Batch Processing) (4), respectivamente.

Motor de transferencia de datos

Los datos deben ser importados antes de que puedan ser procesados por la solución de Big Data. De forma similar, los datos procesados deben ser exportados a otros sistemas antes de que puedan ser usados por un agente externo a la solución de Big Data.

Figura 2.39 – Símbolo utilizado para representar el motor de transferencia de datos.

Un motor de transferencia de datos permite que los datos sean transferidos dentro o fuera de los dispositivos de almacenamiento de la solución de Big Data. A diferencia de otros sistemas de procesamiento de datos, donde los datos de entrada se ajustan a un esquema y están estructurados en su mayoría, las fuentes de datos de una solución de Big Data tienden a incluir una combinación de datos estructurados y datos sin estructurar.

Un motor de transferencia de datos es compatible con las funciones de ingreso y de egreso de datos, en cuyo caso puede calificarse de la siguiente manera:

- motor de ingreso de transferencia de datos
- motor de egreso de transferencia de datos

Las funcionalidades de ingreso y egreso de transferencia de datos pueden agruparse bajo las siguientes categorías:

- eventos (ingreso solamente)
- archivos (ingreso y egreso)
- relacional (ingreso y egreso)

Generalmente, un motor de transferencia de datos ofrece solamente una de las funcionalidades enumeradas. Es común que una solución de Big Data esté compuesta por varios y distintos motores de transferencia de datos, a fin de facilitar variados requisitos de importación y exportación de los diferentes tipos de datos.

Los motores ingreso de transferencia de datos basados en eventos utilizan un modelo de publicación-subscripción, basado en el uso de colas, para asegurar una alta fiabilidad y disponibilidad. Estos motores facilitan el procesamiento de los datos en movimiento en función del agente, lo que permite ejecutar varias actividades de transformación y limpieza (Cleansing) de datos en tiempo real.

Los motores de transferencia de datos permiten sustituir los datos distribuidos en varias fuentes hospedadas en múltiples sistemas por fuera de la solución de Big Data. Un motor de transferencia de datos puede usar internamente un motor de procesamiento para procesar en paralelo múltiples datasets de gran tamaño. Esto permite que se puedan importar o exportar grandes cantidades de datos en poco tiempo. Un motor de flujo de trabajo (Workflow) puede integrarse con un motor de transferencia de datos, a fin de permitir una importación y exportación automatizadas de los datos. En la figura 2.40 se muestra un ejemplo de un motor de transferencia de datos.

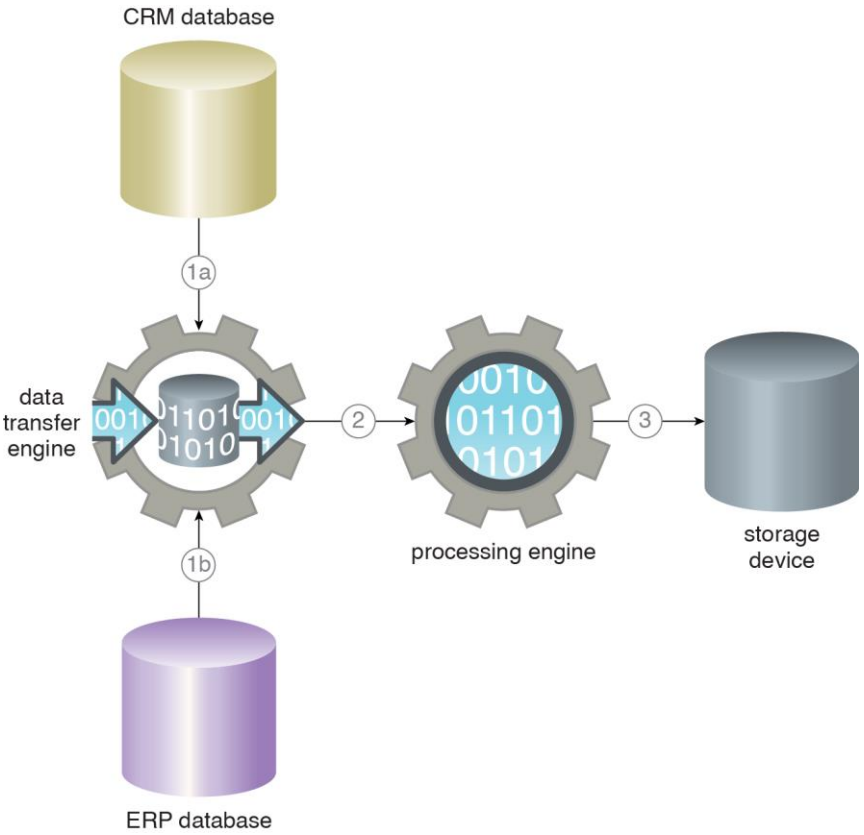


Figure 2.40 – Un motor de transferencia de datos importa los datos desde dos bases de datos diferentes (1a, 1b). Sin embargo, el trabajo real de importación es ejecutado por el motor de procesamiento (2), el cual ejecuta los trabajos de importación y, a continuación, guarda los datos importados en el dispositivo de almacenamiento (3).

Motor de consultas (Query Engine)

El motor de procesamiento permite que los datos sean consultados y manipulados de otras formas, pero se requiere una programación personalizada para implementar este tipo de funcionalidad. No se espera que los analistas que trabajan con las soluciones de Big Data sepan cómo programar los motores de procesamiento.

Un motor de consultas (Query Engine) abstrae el motor de procesamiento de los usuarios finales, proporcionando una interfaz de usuario que puede usarse para consultar datos subyacentes y cuenta con características para crear planes de ejecución de consultas.

Figura 2.41 – Símbolo utilizado para representar un motor de consultas (Query Engine).

Los lenguajes que son más familiares y fáciles de trabajar, como SQL, pueden ser utilizados por usuarios sin conocimientos técnicos para realizar tareas de ETL y ejecutar consultas

especializadas para actividades de análisis de datos (Data Analysis). Las funciones de procesamiento comunes que ejecuta un motor de consultas (Query Engine) incluyen sumas, promedios, agrupaciones, uniones y clasificaciones.

El motor de consulta, el cual se ejecuta de forma subyacente, transforma de forma homogénea las consultas de los usuarios en el código de bajo nivel correspondiente que puede ser utilizado por el motor de procesamiento. El uso de motores de consultas (Query Engines) puede reducir el tiempo de desarrollo y permitir el manejo de datasets de gran tamaño, sin necesidad de escribir una complicada lógica de programación. En la Figura 2.42 se muestra un ejemplo de un motor de consultas (Query Engine).

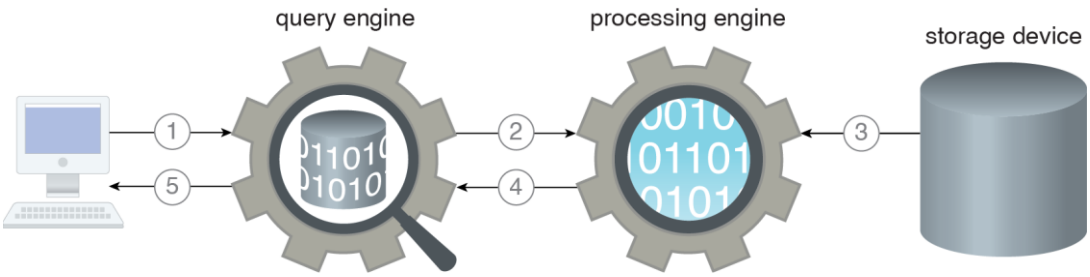


Figura 2.42 – Un cliente realiza una consulta simple de agregación de los datos guardados en el dispositivo de almacenamiento (1). El motor de consultas (Query Engine) genera un plan de ejecución de consultas, y crea trabajos que deben ser ejecutados por el motor de procesamiento (2). El motor de procesamiento obtiene los datos necesarios del dispositivo de almacenamiento (3) y posteriormente ejecuta los trabajos necesarios. Los resultados son reenviados al motor de consultas (Query Engine) (4), el cual los reenvía al cliente luego de ser procesados (5).

Motor analítico (Analytics Engine)

Como soporte de los requisitos de procesamiento analítico, un motor analítico (Analytics Engine) puede procesar avanzados algoritmos estadísticos y de aprendizaje automático (Machine Learning), incluyendo la identificación de patrones y correlaciones. Generalmente utiliza un motor de procesamiento para ejecutar los algoritmos en los datasets de gran tamaño.

Figura 2.43 – Símbolo utilizado para representar el motor analítico (Analytics Engine).

Se utiliza un motor analítico (Analytics Engine) en casos en que las funciones relativamente simples de manipulación de datos de un motor de consultas (Query Engine) no son suficientes. Algunos motores analíticos propietarios también proporcionan características especializadas de análisis de datos (Data Analysis), tales como procesamiento de análisis de texto (text analytics) y de análisis del log de máquinas. La Figura 2.43 muestra un ejemplo de un motor analítico (Analytics Engine).

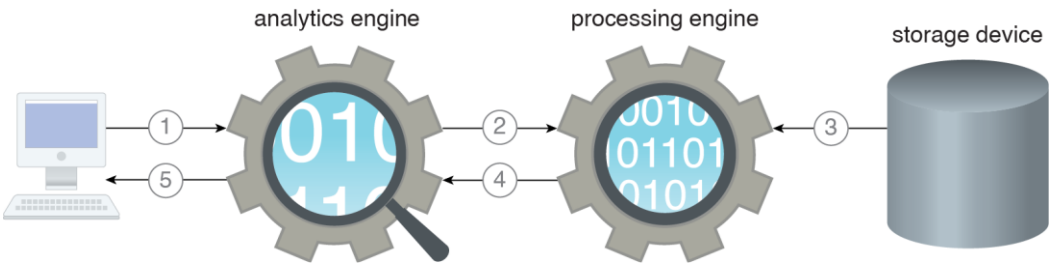


Figura 2.43 – Un cliente realiza una operación estadística avanzada sobre los datos guardados en el dispositivo de almacenamiento (1). El motor analítico (Analytics Engine) crea trabajos que se deben ejecutar en un motor de procesamiento (2). El motor de procesamiento obtiene los datos necesarios del dispositivo de almacenamiento (3) y posteriormente ejecuta los trabajos necesarios. Luego, los resultados son enviados al motor analítico (Analytics Engine) (4), el cual los reenvía al cliente después del procesamiento posterior.

Motor de flujo de trabajo (Workflow)

La capacidad de consultar datos y realizar operaciones de ETL a través del motor de consultas (Query Engine) es útil para los análisis especializados de datos (Data Analysis). Sin embargo, a menudo es necesario realizar repetidamente el mismo conjunto de operaciones en un determinado orden para obtener resultados actualizados basados en los datos más recientes. Un motor de flujo de trabajo (Workflow) permite diseñar y procesar una secuencia compleja de operaciones que puede ser activada periódicamente o cuando los datos estén disponibles.

Figura 2.44 – Símbolo utilizado para representar un motor de flujo de trabajo (Workflow).

La lógica de un flujo de trabajo (Workflow) procesado por un motor de flujo de trabajo (Workflow) puede implicar la participación de otros mecanismos de Big Data, como se muestra en la Figura 2.45. Por ejemplo, un motor de flujo de trabajo (Workflow) puede ejecutar una lógica que recopile datos relacionales a partir de múltiples bases de datos, en intervalos periódicos y a través del motor de transferencia de datos, y finalmente guardar los resultados en un dispositivo de almacenamiento NoSQL.

Los flujos de trabajo (Workflows) definidos son similares a diagramas de flujo lógico, incluyendo bifurcaciones, uniones y decisiones, y generalmente dependen de un motor de procesamiento por lotes (Batch Processing) para ser ejecutados. Los datos de salida de un flujo de trabajo (Workflow) pueden convertirse en los datos de entrada de otro.

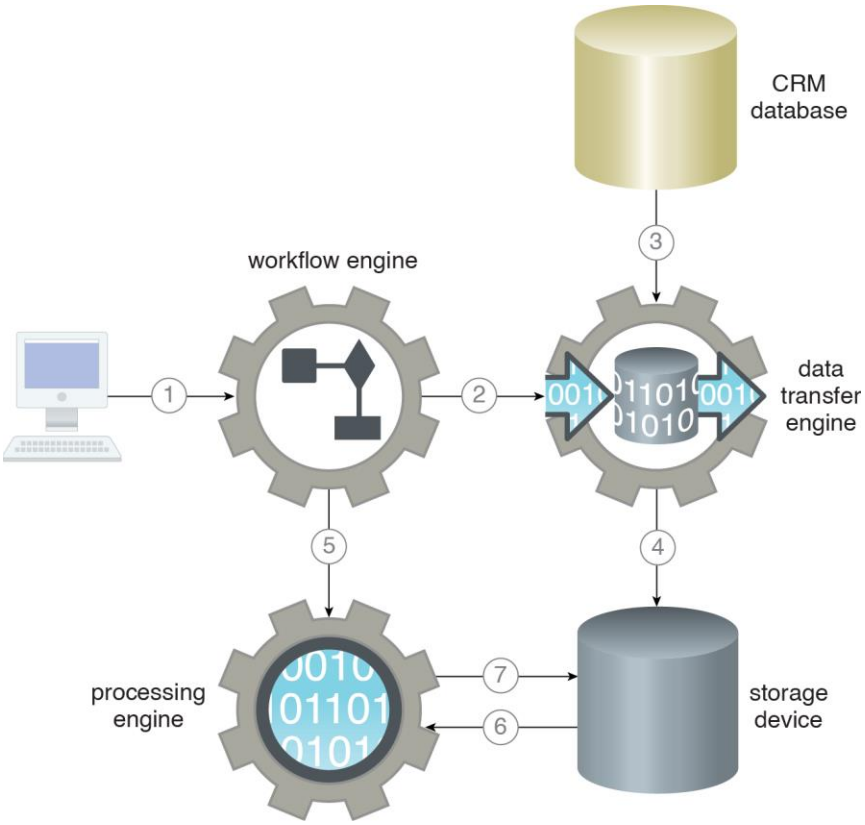


Figura 2.45 – Primero, un cliente crea un flujo de trabajo (Workflow) mediante el motor de flujo de trabajo (Workflow) (1). Como primer paso del trabajo de configuración, el motor de flujo de trabajo (Workflow) activa un trabajo de ingreso de datos (2), el cual es ejecutado por el motor de transferencia de datos bajo la modalidad de importación de datos desde una base de datos CRM (3). Luego, los datos importados son guardados en el dispositivo de almacenamiento (4). Como parte del segundo paso del trabajo de configuración, el motor de flujo de trabajo (Workflow) activa el motor de procesamiento para ejecutar el trabajo de procesamiento de datos (5). En respuesta, el motor de procesamiento recupera los datos necesarios desde el dispositivo de almacenamiento (6), ejecuta el trabajo de procesamiento de datos y, a continuación, guarda los resultados nuevamente en el dispositivo de almacenamiento (7).

Motor de coordinación

Una solución distribuida de Big Data que deba ejecutarse en varios servidores depende de un motor de coordinación, a fin de garantizar la consistencia operativa en todos los servidores involucrados. El motor de coordinación hace posible desarrollar soluciones distribuidas de Big Data altamente confiables y disponibles que puedan ser implementadas en un cluster.

Figura 2.46 – Símbolo utilizado para representar un motor de coordinación.

En ocasiones, el motor de procesamiento utilizará el motor de coordinación para coordinar el procesamiento de datos en un gran número de servidores. Así, no es necesario que el motor de procesamiento tenga una lógica de coordinación propia.

El motor de coordinación también se puede usar para las siguientes tareas, como se muestra en la Figura 2.47:

- gestionar los bloqueos distribuidos
- gestionar las colas distribuidas
- establecer un registro de disponibilidad para obtener información de la configuración
- asegurar una comunicación asíncrona entre los procesos que se están ejecutando en diferentes servidores

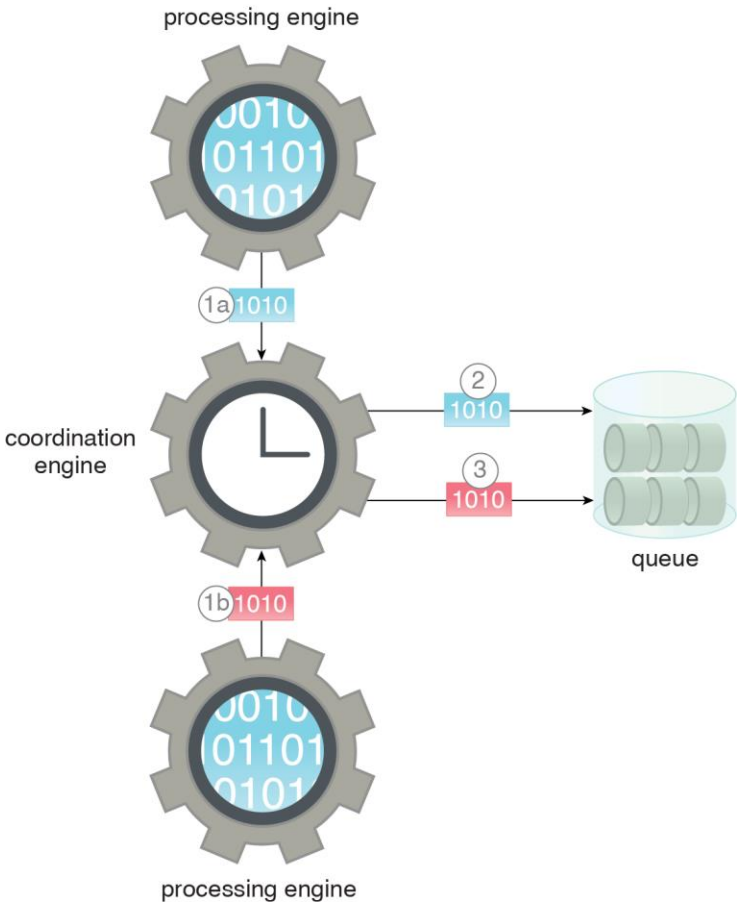


Figura 2.47 – Dos nodos de un cluster deben escribir en una cola compartida como parte de la ejecución de un trabajo, y ambos envían una solicitud de escritura al mismo tiempo (1a, 1b). El motor de coordinación supervisa la solicitud de escritura. Se envía una solicitud a la cola (2) antes de que la otra solicitud se envíe de forma serializada (3).

Ejercicio 2.3: complete los espacios en blanco

1. El _____ está limitado debido a que, por lo general, solo puede realizar operaciones simples de datos, como promedios, uniones y clasificaciones. El _____ es más avanzado y puede ejecutar algoritmos estadísticos y de aprendizaje automático (Machine Learning) complejos.

2. Las funcionalidades de los motores de _____ y los motores de _____ de transferencia de datos se pueden clasificar según eventos, archivos y funcionalidades relacionales.

3. Los motores de procesamiento por lotes (Batch Processing) presentan una _____, lo que significa que se pueden demorar algunas horas para terminar una tarea de procesamiento de datos. Por otro lado, los _____proporcionan tiempos de respuesta menores a un segundo.

4. El mecanismo de un _____ se puede usar para procesar secuencias complejas de operaciones de forma periódica o cuando los datos estén disponibles.

5. Un _____ programa las solicitudes de procesamiento de datos y actúa como un árbitro de recursos, que gestiona y asigna los recursos disponibles a las diferentes aplicaciones. Este mecanismo, en conjunto con los mecanismos del _____ y del _____, es de carácter obligatorio en las plataformas de Big Data para lograr una interoperabilidad entre datasets de gran tamaño.

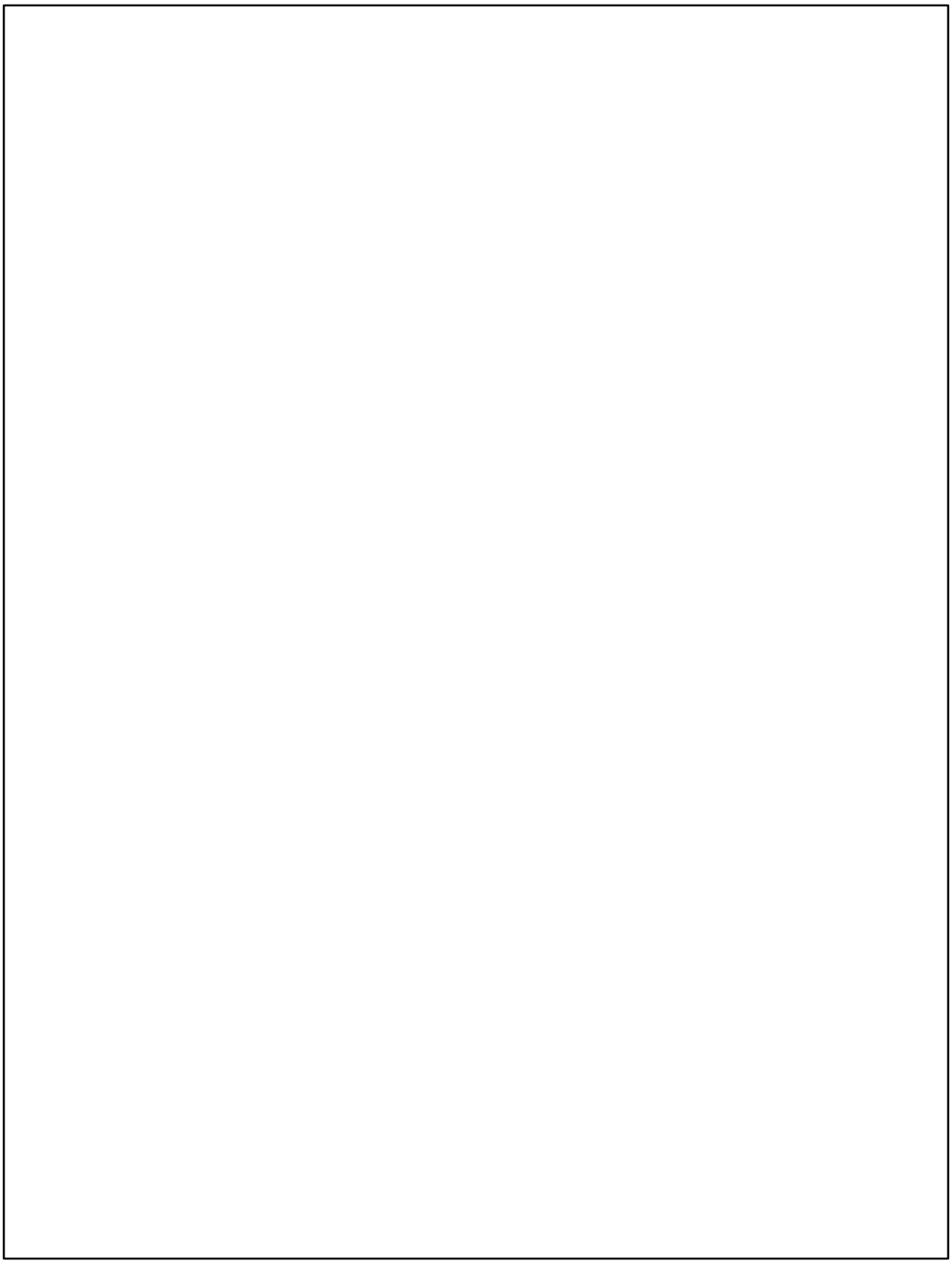
6. Los _____ y las _____ son ejemplos de mecanismos de dispositivos de almacenamiento.

7. El mecanismo del _____ se usa para mantener la consistencia operativa entre los múltiples _____ en una solución distribuida de Big Data. Este mecanismo es usado generalmente por el mecanismo del _____ para un procesamiento distribuido y uniforme de los datos.

Las respuestas al ejercicio se encuentran al final de este cuadernillo.

Notas

Notas / Bocetos



Respuestas a los ejercicios

Respuestas al ejercicio 2.1

1. Evaluación del caso empresarial
2. Identificación de datos
3. Adquisición y filtrado (filtering) de datos
4. Extracción de datos
5. Validación y limpieza (Cleansing) de datos
6. Agregación y representación de datos
7. Análisis de datos (Data Analysis)
8. Visualización de datos
9. Uso de los resultados del análisis

Respuestas al ejercicio 2.2

1. Análisis de sentimientos (Sentiment Analysis)
2. Analítica de texto (text analytics)
3. Correlación
4. Regresión
5. Análisis de redes
6. Filtrado (filtering)
7. Análisis de series temporales
8. Mapas de calor
9. Clasificación
10. Agrupamiento (Clustering)
11. Detección de datos atípicos (outliers)
12. Análisis espacial de datos
13. Tests A/B

14. Procesamiento de lenguaje natural (NLP)

Respuestas al ejercicio 2.3

1. El **motor de consultas (Query Engine)** está limitado debido a que, por lo general, solo puede realizar operaciones simples de datos, como promedios, uniones y clasificaciones. El **motor analítico (Analytics Engine)** es más avanzado y puede ejecutar algoritmos estadísticos y de aprendizaje automático (Machine Learning) complejos.

2. Las funcionalidades de los motores de **ingreso** y los motores de **egreso** de transferencia de datos se pueden clasificar según eventos, archivos y funcionalidades relacionales.

3. Los motores de procesamiento por lotes (Batch Processing) presentan una **alta latencia**, lo que significa que se pueden demorar algunas horas para terminar una tarea de procesamiento de datos. Por otro lado, los **motores de procesamiento en tiempo real** proporcionan tiempos de respuesta menores a un segundo.

4. El mecanismo de un **motor de flujo de trabajo (Workflow)** se puede usar para procesar secuencias complejas de operaciones de forma periódica o cuando los datos estén disponibles.

5. Un **gestor de recursos** programa las solicitudes de procesamiento de datos y actúa como un árbitro de recursos, que gestiona y asigna los recursos disponibles a las diferentes aplicaciones. Este mecanismo, en conjunto con los mecanismos del **dispositivo de almacenamiento** y del **motor de procesamiento**, es de carácter obligatorio en las plataformas de Big Data para lograr una interoperabilidad entre datasets de gran tamaño.

6. Los **sistemas de archivos distribuidos** y las **bases de datos** son ejemplos de mecanismos de dispositivos de almacenamiento.

7. El mecanismo del **motor de coordinación** se usa para mantener la consistencia operativa entre los múltiples **servidores** en una solución distribuida de Big Data. Este mecanismo es usado generalmente por el mecanismo del **motor de procesamiento** para un procesamiento distribuido y uniforme de los datos.

Examen B90.02

El curso que acaba de finalizar corresponde al Examen B90.02, que es un examen oficial del Programa Profesional Certificado de Ciencias de Big Data (BDSCP).

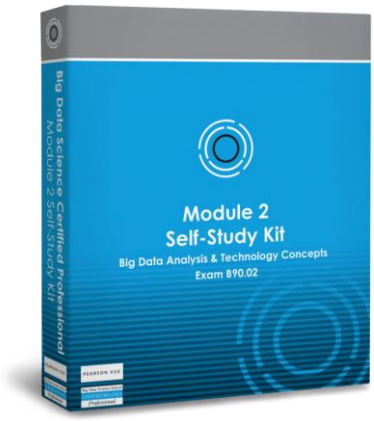
Este examen se puede tomar en los Centros de Examinación de Pearson VUE en todo el mundo, o a través de Pearson VUE Proctoring Online, que le permite tomar exámenes desde su casa o estación de trabajo y contar con supervisión en vivo. Si desea obtener más información, visite las siguientes páginas web:

www.bigdatascienceschool.com/exams/

www.pearsonvue.com/arcitura/

www.pearsonvue.com/arcitura/op/ (Online Proctoring)

Kit de autoaprendizaje del Módulo 2



Para este Módulo se encuentra disponible un kit oficial de autoaprendizaje de BDSCP, el cual le ofrece materiales y recursos de estudio adicionales, incluyendo una guía separada de autoaprendizaje, CD de audiotutoría y tarjetas de memorización.

Tenga en cuenta que las versiones de este kit de autoaprendizaje están disponibles con y sin un cupón para el Examen B90.02 de Pearson VUE.

Si desea obtener más información, visite la siguiente página web:
www.bigdataselfstudy.com

Información y recursos de contacto

Comunidad de AITCP

Únase a la creciente comunidad internacional de Profesionales Certificados en TI de Arcitura (Arcitura IT Certified Professional, AITCP), conectándose mediante las plataformas oficiales de social media: LinkedIn, Twitter, Facebook y YouTube.

Los links de social media y de la comunidad están disponibles en:

- www.arcitura.com/community
- www.servicetechbooks.com/community

Información general del programa

Si desea conocer información general acerca del programa del BDSCP y los requisitos de certificación, visite las siguientes páginas web:

www.bigdatascienceschool.com y www.bigdatascienceschool.com/matrix/

Información general acerca de los módulos del curso y los kits de autoaprendizaje

Si desea conocer información general acerca de los módulos del curso del BDSCP y los kits de autoaprendizaje, visite las siguientes páginas web:

www.bigdatascienceschool.com y www.bigdataselfstudy.com

Inquietudes acerca del examen de Pearson VUE

Si desea conocer información relacionada con la presentación de los exámenes del BDSCP en los centros de examinación de Pearson VUE o mediante la supervisión online de Pearson VUE, visite las siguientes páginas web:

www.pearsonvue.com/arcitura/
www.pearsonvue.com/arcitura/op/ (Online Proctoring)

Programación de talleres dirigidos al público y guiados por instructores

Si desea conocer la más reciente programación de los talleres del BDSCP guiados por instructores que están abiertos al público, visite la siguiente página web:

www.bigdatascienceschool.com/workshops

Talleres privados guiados por instructores

Los entrenadores certificados pueden realizar los talleres directamente en sus instalaciones, con la opción de supervisión de exámenes en el sitio. Si desea saber más acerca de las opciones y tarifas, envíe un correo electrónico a la siguiente dirección:

info@arcitura.com

o llame a la línea

1-800-579-6582

Convertirse en un entrenador certificado

Si usted está interesado en alcanzar el rango de Entrenador Certificado para este o cualquier otro curso o programa de Arcitura, puede obtener más información visitando la siguiente página web:

www.arcitura.com/trainerdevelopment/

Inquietudes generales sobre BDSCP

En caso de que tenga otras preguntas relacionadas con este Curso, o cualquier otro Módulo, Examen o Certificación que haga parte del programa BDSCP, envíe un correo electrónico a la siguiente dirección:

info@arcitura.com

o llame a la línea

1-800-579-6582

Notificaciones automáticas

Si desea que se le notifique automáticamente sobre cambios o actualizaciones al programa BDSCP y a los sitios de recursos relacionados, envíe un mensaje de correo electrónico en blanco a la siguiente dirección:

notify@arcitura.com

Retroalimentación y comentarios

Ayúdenos a mejorar este curso. Envíe su retroalimentación o comentarios a la dirección de correo electrónico:

info@arcitura.com